

Test of fit and model selection based on likelihood function

Abdolreza Sayyareh

▶ **To cite this version:**

Abdolreza Sayyareh. Test of fit and model selection based on likelihood function. Mathematics [math]. AgroParisTech, 2007. English. NNT : 2007AGPT0020 . pastel-00003400

HAL Id: pastel-00003400
https://pastel.hal.science/pastel-00003400

Submitted on 12 Feb 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N°/__/__/__/__/__/__/__/__/__/__/__/

THÈSE

pour obtenir le grade de

Docteur

de

l'Institut des Sciences et Industries du Vivant et de l'Environnement (Agro Paris Tech)

Spécialité : STATISTIQUE

Présentée et soutenue publiquement
par

ABDOLREZA SAYYAREH

22 JUIN

TEST OF FIT AND MODEL SELECTION BASED ON LIKELIHOOD FUNCTION

Directeur de thèse : ***AVNER BAR-HEN***

Codirecteur de thèse : ***DANIEL COMMENGES***

Travail réalisé : Organisme, ISPED, unité 875 BIOSTATISTIQUE, université Bordeaux 2, F-33076 Bordeaux cedex

Devant le jury :

M. Christophe Biernacki, Professeur, **LILLE 1**……..**Rapporteur**

M. Jérôme Saracco, Professeur, **Bordeaux 4**……**Rapporteur**

M. Jean Jacques Daudin, Professeur **INA-Paris-Grignon**……**Examinateur**

M. Daniel Commenges, Docteur, Directeur de recherche, **INSERM**.....**Codirecteur de thèse**

M. Avner Bar-Hen, Professeur, **INA-Paris-Grignon**, **Paris 13**……..**Directeur de thèse**

L'Institut des Sciences et Industries du Vivant et de l'Environnement (Agro Paris Tech) est un Grand Etablissement dépendant du Ministère de l'Agriculture et de la Pêche, composé de l'INA PG, de l'ENGREF et de l'ENSIA (décret n° 2006-1592 du 13 décembre 2006)

A mes parents.

A Saba, ma fille.

A mon èpouse.

A Mohammad mon petit fils qui parti trop vite et rest en France pour toujours.

Remerciements

A Monsieur le Professeur Jean Jacques Daudin:

Je vous remercie d'avoire acceptè de présider ce jury. Veuillez recevoir mes plus vifs remerciements et l'expression de ma sincére considèration.

A Monsieur le Professeur Christophe Biernacki:

Vous m'avez fait un grand honneur en acceptant de juger ce travail. Votre grand expèience en estimateur veraisemblance, modèlisation statistique et statistiques asymptotique m'apporteront beaucoup. Je vous remercie sincèrement d'avoir acceptè d'être le rapporteur de ma thèse.

A Monsieur le Professeur Jèrome Saracco:

Vous m'avez fait un grand honneur en acceptant de juger ce travail. Votre grand expèience en modèlisation statistique et statistiques asymptotique m'apporteront beaucoup. Je vous remercie sincèrement d'avoir acceptè d'être le rapporteur de ma thèse.

A Monsieur le Professeur Avner Bar-Hen et

A Monsieur le Docteur Daniel Commenges:

Je ne pourrai jamais assez vous remercier pour tout ce que vous m'avez apportè. Vous m'avez prenez sous votre ailes. Je vous remercie de m'avoir confiè ces sujets de recherche. Votre disponibilitè et les conseils prècieux que vous m'avez donnè m'ont permis de bien travailler et d'amèiorer mes connaissances. Une grande merci pour la formation scientifique que vous m'as transmise.Votre attention. Ce fut un rèel bonheur de travailler avec vous. Merci infiniment pour tout et le rest. J'espère pouvoir continuer à travailler avec vous. Et Daniel, un petit mot pour toi, toujours tu'était avec moi, parfois comme mon frère et parfois comme mon ami mais toujours comme mon pro-

fesseur, et aussi toi Avner, merci pour la confiance que tu m'as accordée au début de ma demande et pendant ce période.

Un immense merci à tous:

A Ecole Doctoral ABIES et Agro. Paris Tech. A UMR INAPG/ENGREF/INRA, Dep. Mathématiques et Informatique appliquée. A Mme. Françoise Launay, Mme. Corinne Fiers et Mme. Alice François. A l'èquipe biostat dans son ensemble, à Hélèn, Virginie, Alioum, Pierre, Rodolphe et Luc. Merci a Guillaume pour sauvetages d'ordi. à la rescousse. A professeur Nikolin, Professeur Salamon et Professeur Salmi. Aux doctorans qui se sont succédé : Jérémie, Cécile, Julia, Etienne, et bien-sûr Cécile (Sommen) pour votre conseils durant cette période, et ceux que j'oublie.

Introduction en français

Un problème important de la statistique concernant un échantillon i.i.d, de taille n est de tester si ses observations viennent d'une distribution spécifiée. Cela signifie qu'il y a une incertitude et nous devons prendre une décision. Un processus décisionnel en situation d'incertitude est en grande partie basé sur l'application d'analyse de données statistiques pour l'évalution des risques probabilistes de notre décision. Dans une situation réaliste nous avons seulement un ensemble de données actuelles et nous devons établir 'la connaissance'. La connaissance est ce que nous savons et la communication de la connaissance est 'l'information'. Les données sont seulement l'information brute et non la connaissance de celles-ci. Les données deviennent des informations quand cela devient pertinent à notre problème de décision. L'information devient 'le fait' quand les données peuvent le soutenir. Enfin le fait devient la connaissance quand il est utilisé dans l'achèvement réussi du probléme de décision. Le processus réfléchi statistique basé sur les données, construira les modèles statistiques pour la prise de décision en situation d'incertitude. La statistique résulte du besoin de placer la connaissance sur une base systématique d'évidence. Ceci a exigé une étude des lois de probabilité. Ainsi, la fonction de densité est un concept fondamental dans la statistique. La vraie fonction de densité, que nous dénotons $f(.)$ est inconnue. Nous appelons cette distribution la 'vrai distribution'. Un modèle est une famille des distributions et est appelé 'bien- specifié' s'il contient la vrai distribution; on peut également parler du 'vrai modéle' mais cela peut être fallacieux (induire en erreur). Les données sont insuffisantes pour reconstruire chaque détail de $f(.)$. Alors parfois nous l'estimons et parfois nous l'approximons. Le secteur d'estimation de densité peut être paramétrique ou nonparamétrique. Le cas nonparamétrique est la construction d'une estimation de la fonction densité des données observées où le nombre de paramètres est considéré comme infini. Dans ce cas-ci l'estimation de la densité, $f(.)$ pour tous les points dans son support impliquerait l'estimation

d'un nombre infini de paramètres. Historiquement on peut dire que l'estimateur (nonparamétrique) de densité le plus ancien utilisé est l'histogramme qui a été raffiné pour obtenir les estimateurs lisses par l'approche d'estimation à noyaux, voir, Fix and Hodges (1951), Devroye (1985) and Silverman (1986). Voir la figure 1. L'autre cas est le cas paramétrique, où nous supposons connue la forme de la fonction de densité et nous voulons estimer seulement les paramètres. Dans le cas paramétrique nous supposons que les données sont générée à partir des familles paramètriques de distributions connues. L'approche la plus employée est basée sur les estimateurs de maximum de vraisemblance l'(EMV) et certaines de ses modifications. Généralement ce secteur est lié au problème de test d'hypothése. En tant que propriété naturelle, nous voulons considérer les méthodes pour construire les procédures qui sont efficaces, c'est-à-dire, asymptotiquement optimales. La théorie derrière le EMV garantit cette optimalité. Dans le problème de test d'hypothése, nous définissons formellement les hypothèse nulle et alternative au sujet des paramètres de la densité fondamentale. Les quantités de base dont nous avons besoin dans le test d'hypothèse sont la valeur critique qui fournit le niveau du test, la puissance du test et la dimension de l'échantillon requise pour obtenir une puissance donnée. D'autre part nous pouvons comparer deux modèles en concurrence, par exemple une densité normale contre une densité double exponentielle. Puisque nous pouvons imaginer plusieurs modèles pour l'approche de $f(.)$, la question du "choix du modèle" surgit. Par la choix de modèle nous nous rappelons Ockham (1282-1347) qui a déclaré que des 'entités ne doivent pas être multiplisitées au de la nécessité, qui est connue sous le nom de rasoir d'Ockham. Simplement un modèle est un ensemble d'équations, ou de fonctions avec quelques paramètres ajustables, nous pouvons définir un modèle en tant qu'ensemble de probabilité, ou d'hypothèses statistiques. Le choix de modèle consiste à sélectionner un modèle mathématique parmi un jeu de modèles potentiels, celui qui se collera au mieux à notre série d'observations. Nous considérons une famille de densités comme

modèle, dans la quelle les membres diffèrent par la valeur des paramètres. Notre recherche est de trouver le vrai modèle. Nous disons qu'un modéle est vrai si et seulement si une des densités qu'il contient est la vraie. Il est nécessaire de choisir l'ensemble de modèles avant de commencer. Dans le modèle de régression linéaire par exemple, le choix de modèle est difficile car nous avons 2^p modéle potentiels où p est le nombre de variables explicatives qui sont candidates à l'explication de la variable réponse. Le problème est que des termes supplémentaires ajoutent des paramètres ajustables supplémentaires, et ceux-ci amélioreront l'ajustement. Pour prouver la sensibilité du choix de modèle, et l'importance de ce concept dans le secteur de recherche, nous considérons deux modèles pour un ensemble de données actuel, les comme $\{Y = \beta_0 + \beta_1 X_1 + \varepsilon, \quad \beta_0, \beta_1 \in \mathcal{R}; \quad \varepsilon \sim \mathcal{N}(0,1)\}$ et $\{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon, \quad \beta_0, \beta_1, \beta_2 \in \mathcal{R}; \quad \varepsilon \sim \mathcal{N}(0,1)\}$. Nous considérons une situation de simulation dans laquelle nous savons que le premier modèle est correct, c.-à-d. nous acceptons que $\beta 2 = 0$. La figure 2 montre que l'inclusion d'une variable explicative avec coefficient zéro dans les modéle changent la distribution de l'estimateur de β_1. De ce fait l'utilisation du mauvaise modèle (deuxième modèle) nous mène à une mauvaise inférence pour β_1, le paramètre qui doit être dans le modèle. Dans la littérature, les méthodes classiques de choix de modèle sont connues sous forme de test d'ajustement, Pearson (1900), et le test d'hypothèse classique, Neyman-Pearson (1933-1938) pour des modèles à un paramètre, et sa prolongation qui emploie le paradigme de Neyman-Pearson avec l'estimateur de maximum de vraisemblance pour nous donner une méthode de test universel, le test de rapport de vraisemblance. D'autre part quelques méthodes récentes dans les critères de choix de modèle sont le critère d'information d'Akaike (AIC), (Akaike, 1973), le critére Bayésien d'information (BIC) (Schwarz, 1978), la technique de validation croisée qui est asymptotiquement équivalent à *AIC* dans le cas paramètrique, et le critère minimum de longueur de description, Bozdogan (2000) qui est asymptotiquement équivalent au *BIC*. En fait nous savons que le test d'hypothèse

classique avec sa théorie étendue optimise la qualité de l'ajustement. Ainsi pourquoi y a-t-il besoin d'autres méthodes de sélect de modèle? La réponse est que cette méthode ne se prolonge pas simplement à l'hypothése non emboîtée et puis avec cette méthode nous ne pouvons pas faire une analyse profonde du problème du sélection de modèle dans de situations réelle. Un autre point important est que la conclusion des critéres comme *AIC* ne sont jamais au sujet de la vérité ou de la fausseté d'une hypothèse, mais au sujet de sa proximité à la vérité. D'autre part dans le test d'hypothèses classique on cherche à minimiser les erreurs des types *I* et *II* qui ne sont pas compatibles. Il y a une autre objection au raisonnement du test d'hypothèses classiques. Il peut être difficile de trouver un modèle bien specifié. Il peut encore être approprié de choisir le meilleur modèle parmi un ensemble (non specifié) de modèles. Notre travail porte sur la méthode de maximum de vraisemblance et particulièrment sur l'*AIC*. Ceci parce que l'*AIC* peut être employé pour les modèles emboîtés et non emboîtés. L'*AIC* adopte le critère de Kullback-Leibler en tant que sa fonction de divergence. Fisher dans son introduction originale du critère de suffisance, a exigé que la statistique devrait résumer la totalité de l'information appropriée fournie par l'échantillon, et le probléme de discriminer de l'idée de Kullback-Leibler est de considérer une mesure de la distance ou de la divergence entre les distributions statistiques en termes de leur mesure d'information. Nous pouvons également considérer la distance d'Hellinger ou la distance de Matusita de l'affinité, voir Bar-Hen et Daudin (1998). En fait ils ont défini le rapport de log-vraisemblance comme l'information d'une observation y à distinguer entre deux hypothéses liées à la sélection du modèle. Il y a beaucoup de manières de définir la divergence, mais dans tout le manque d'ajustement désigné sous le nom de divergence. Dans la littérature il y a quelques autres versions du critère d'Akaike. Dans le modèle de régression linéaire, la statistique la plus populaire pour le choix de modèle est le C_p de Mallows (1973). D'autres critères sont des critère l' AIC_c corrigé par Hurvith et Tsai (1989), le critère

prolongé de l'information EIC par Ishiguro et al (1997), cette approche a été prolongée au choix de l'estimateur semi-paramétrique par Commenges et al. (2007) et ICOMP par Bozdogan (2000). Ils sont à la recherche du choix du poids de la pénalité du critère, ce qui est lié à la parcimonie du modéle. D'autre part classiquement nous pouvons considérer les modèles formulés comme des distributions de probabilité. En fait la sélection des modéles se fait en deux étapes. Dans la premiére étape nous devons choisir l'ensemble des modèles. La deuxième étape de sélection de modéle est bien connue comme l'évaluation des paramètres, c.-à-d. une fois que l'ensemble des modèles possibles sont choisis, l'analyse mathématique nous permet de déterminer le meilleur de ces modèles. Mais que signifie le meilleur? Une bonne technique de choix de modèle équilibrera l'ajustement et la complexité. L'ajustement est généralement déterminé par la divergence minimum ou au sens de la vraisemblance, et la complexité est généralement mesurée en comptant le nombre de paramètres libres dans le modèle. Pour choisir parmi les modèles en concurrence, nous devons décider quel critère doit être employé pour évaluer les modèles, et puis pour faire la meilleure inférence quant à laquelle modèle est préférable. Comme nous avons dit, nous pouvons considérer la divergence entre les modèles comme critère de choix de modèle. Alors notre recherche sera de trouver le modèle avec la divergence minimum par rapport à la vrai densité qui est parfois complétement inconnue et parfois inconnue dans le paramètre. Un travail intéressant est effectué par Vuong (1989) qui emploi le critère de Kullback-Leibler pour mesurer la proximité d'un modéle au vrai. Il considère la limite de la pénalité dans l'*AIC* comme une quantité négligeable quand la dimension de l'échantillon devient grande. Il y a une période importante pour les test de sélection de modèle, de Cox (1961-1962) à Vuong (1989). Le test de Vuong comme un test pour choix de modèle est différent de test de Cox. Avec le test de Cox chaque modèle est évalué contre les données, c.à-d. le modéle alternatif fournit la puissance. En fait le test de Cox est une modification du test de rapport de

vraisemblance de Neyman-Pearson. D'autre part le test de Vuong est un test d'hypothèse relatif. Dans ce cas les tests de modèles sont évalués contre les données et l'un contre l'autre. La différence entre les deux test est importante. Le test de Cox est valable pour des hypothèses non-emboîtées tandis que le test de Vuong s'utilise pour sélectionner des modèles non-emboîtées. Il est nécessaire de souligner qu'à l'origine le test de rapport de vraisemblance est un test statistique d'ajustement entre deux modèles emboîtées. Par ce test un modèle relativement plus complexe est comparé à un modèle plus simple. D'autre part les tests classiques d'ajustement sont fréquenement employés par tous les chercheurs qui ont besoin de l'interprétation statistique de leur données. Historiquement Pearson (1900) a proposé le premier test d'ajustement qui est connu comme test de χ^2. Cet test de base est devenu une source importante pour le développement des secteurs principaux en probabilité et statistique. Fisher (1922) a présenté la vraisemblance dans le contexte de l'estimation au point pour un paramètre d'intérêt, mais au commencement la vraisemblance est un outil pour traiter l'incertitude due à la quantité d'information limitée continue dans les données. C'est la fonction entière de vraisemblance qui saisit toute l'information dans les données. Alors pour chercher un test d'ajustement la fonction de vraisemblance est un premier candidat.

Notre Objectif

Nous nous concentrons sur la théorie asymptotique pour la sélection de modèle. Nous étudions la situation sous laquelle les procèdures de sélection de modèles sont asymptotiquement optimales pour choisir un modèle. Notre travail port sur l'inférence au sujet de l'AIC (un cas de vraisemblance pènalisée) d'Akaike (1973), où comme estimateur de divergence de Kullback-Leibler est intimement reliée à l'estimateur de maximum de vraisemblance. Comme une partie de la statistique inférentielle, dans le contexte de test d'hypothèse, la divergence de Kullback-Leibler et le lemme de Neyman-Pearson sont deux concepts fondamentaux. Tous les deux sont au sujet du rapports de

vraisemblance. Neyman-Pearson est au sujet du taux d'erreur du test du rapport de vraisemblance et la divergence de Kullback-Leibler est l'espérance du rapport de log-vraisemblance. Ce raccordement présente une autre interprétation de la divergence de Kullback-Leibler dans la limite de la perte de puissance du test du rapport de veraisemblance quand la distribution fausse est employée pour une de l'hypothèse, c.-à-d. la divergence de Kullback-Leibler de deux fonctions de distribution P et Q mesure combien de puissance nous perdons avec la test du rapport de vraisemblance si nous mouvais spécifiée l'hypothèse alternative P comme Q. Nous voulons encore confirmer que l'estimateur de la divergence de Kullback-Leibler qui est la fonction maximisée (et normalisée) de vraisemblance, assymptotiquement pourrait être une bonne statistique pour le choix de modèle. Par ceci nous éliminons le partie normalisée du test du rapport de vraisemblance qui est une cause qui à l'incapacité de l'étude classique de puissance. En fait nous voulons développons une approche pour le test d'ajustement basé sur des fonctions vraisemblance normalisées (par nombre d'observation) et de l'AIC normalisée quand la dimension de l'échantillon devient grande.

Notre approche est basée sur l'AIC et la différence de l'AIC pour deux modèles de concurrence en utilisant l'intervalle de confiance au lieu du test de hypothèse comme son double, c'est parce que l'intervalle de confiance est un ensemble de toutes les hypothèses acceptables avec la confiance pré assignée. L'évaluation d'un intervalle de confiance pour deux modèles emboîtées ou non-emboîtées en concurrence est concentrée dessus, que l'intervalle de confiance contiant zéro ou pas. En bref nous considérons les AIC car une statistique qui nous laissent présenter une statistique de test pour selection de modèle. Cette idée est différente de l'idée originale au sujet de l'AIC qui considére AIC comme critère qui ordonne modèles. Nous voulons souligner que le choix de modèle pourrait impliquer une différence entre la simplicité et l'ajustement. Il y a beaucoup de manières de faire cette différence. Essentiellement cependant, il n'y a aucune méthode qui est meilleure que toutes

les autres dans toutes les conditions, c.-à-d. pour toutes les méthode $m1$ et $m2$, il y a sont des circonstances dans lesquelles $m1$ est meilleur que le $m2$, et d'autres circonstances dans lesquelles $m2$ est meiuex que $m1$. Il semble qu'il est difficile de comparer les méthodes, parce qu'il parfois nous guidera à une conclusion non-admissible. Au lieu de choisir une méthode nous pouvons analyser notre problème et préciser notre but et les moyens de réaliser notre but et d'expliquer finalement comment un critère fonctionne en réalisant notre but. Le domaine du choix de modèle est très grand. Une catégorisation du problème de choix de modèle peut être considérée selon que les modèles sont emboîtés, en chevauchement ou non emboîtés. Généralement deux modèles seraient non emboîté s'il n'est pas possible de conduire chacun d'eux par les autres l'un ou l'autre au moyen d'un ensemble exact de restriction paramétrique ou en raison d'un processus limiteur. La littérature sur test d'hypothése non emboîtée a été imitiée par Cox (1961), Cox (1962) et Atkinson (1970), ce sujet appliqué par Pesaran (1974) et Pesaran et Deaton (1978). L'analyse des modèles régression non emboîtés a été considéré par Davidson et Mackinnon (1981), Fisher et McAller (1981) et Dastoor (1983). D'autre part Vuong (1989) a considéré le test d'hypothèse quand deux modèles en concurrence sont emboîtés, chevauchement ou non emboîtée. Son approche est basée sur la distribution asymptotique de la différence des fonctions de log-vraisemblance pour deux modèles en concurrence. Shimodaira (1998) et Shimodaira (2001) a considéré l'‘erreur d'echantillonnage de l'*AIC* dans des comparaisons multiples et a construit un ensemble avec de bons modéle plutôt que de choisir un modéle simple. Récemment la distribution asymptotique de l'*AIC* dans des modèles de régression linéaire et la correction de biais du ces statistiques sont discutées par Yanagihara et Ohomoto (2005).

Contents

1 Introduction **19**

1.1 Our Objective 27

1.2 Plan of Thesis 29

2 Reminders about models

and some asymptotic results **30**

2.1 Models 30

2.2 Model Selection 32

2.3 Goal of Model Selection and its means 34

2.4 Nested and Non-Nested Models 35

2.5 Probability Metrics 36

2.6 Akaike framework and his Theorem 37

2.7 Complexity in model selection 38

2.8 Asymptotic theory 42

2.9 Goodness of Fit Test and

Classical Hypothesis Testing 43

2.10 Reminder on Theorems and Lemmas 45

3 Reminder on Goodness of Fit Tests **47**

3.1 Testing fit to a fixed distribution 47

3.1.1 Basic Goodness of Fit Test 48

3.1.2 Tests on the basis of Functional Distance 49

3.2 Adaptation of tests coming from the fixed-distribution 51

3.3 Tests on the basis of Correlation and Regression 52

3.4 Tests on the basis of Likelihood Functions 54

3.4.1 Berk-Jones's statistics 54

3.4.2 Generalized Linear Models (GLMs) and Deviance 55

4 Motivation to Model Selection Tests **61**

4.1 Introduction 61

4.2 Assumptions 63

4.3 Likelihood Function and

Maximum Likelihood Estimator 65

4.3.1 Correctly Specified and Mis-Specified models 68

4.4 Metrics on spaces of probability 69

4.4.1 Kullback-Leibler Discrepancy (divergence) 73

4.5 Consistency of Maximum Likelihood

Estimator 76

4.6 Akaike Information Criterion (AIC) 77

4.7 Distribution of Maximum Likelihood Estimator 79

5 Proposed test for Goodness of Fit Test: A test based on empirical likelihood ratio 82

5.1 Introduction 82

5.2 Our objective 84

5.3 Union-Intersection Test 85

5.4 Proposed test based on empirical likelihood ratio 86

5.4.1 Level of test 89

5.4.2 Comparison with Berk-Jones's test 89

5.4.3 Bahadur efficiency of proposed test 89

6 Proposed Model selection tests based on likelihood and AIC 92

6.1 Introduction 92

6.2 Known parameters case 98

6.3 Unknown parameters case 106

6.4 Test function 110

6.5 Variance estimation 111

6.6 Distribution of T_n under $\mathcal{H}_1$ and Power of Test 123

6.6.1 Distribution of Test Statistic T_n under $\mathcal{H}_1$ 123

6.6.2 Power of Test 127

6.7 Consistency of Test 130

6.7.1 Power computation 131

6.7.2 Invariance 132

7 Test For Model Selection based on difference of AIC's:

application to tracking interval for $\Delta\mathcal{E}KL$ 136

7.1 Introduction 136

7.2 Objective 138

7.3 Non-Nested Models comparison 140

7.3.1 Motivation to Confidence Interval construction 142

7.3.2 Confidence Interval for $\Delta\mathcal{E}KL$ 144

7.4 Logistic Regression: 149

8 Conclusion and perspective 156

9 Bibliography 161

10 Appendix 165

I APPENDIX A 166

10.1 Introduction 1

10.2 Expected Kullback-Leibler Criteria and AIC 5

10.3 Hypothesis Testing 7

10.4 Simulation 10

10.4.1 exploration of our result 10

10.4.2 Application to The Multiple Regression Model. 11

II APPENDIX B 20

10.5 Introduction 22

10.6 Theory about inference of differences of AIC criteria 24

10.6.1 Estimating a difference of Kullback-Leibler divergences 24

10.6.2 Tracking interval for a difference of Kullback-Leibler divergences 27

10.6.3 Extension to regression models 29

10.7 Application to logistic regression: a simulation study 30

10.8 Choice of the best coding of age in a study of depression 32

10.8.1 The Paquid study 32

10.9 Discussion 34

Chapter 1

Introduction

An important problem in statistics concerning a sample of n independent and identically distributed observations is to test whether these observations come from a specified distribution. It means that there is a uncertainty and we have to make a decision. Decision making process under uncertainty is largely based on application of statistical data analysis for probabilistic risk assessment of our decision. In realistic situation we have only a set of data at hand and we need to build knowledge from it. Knowledge is what we know and the communication of knowledge is information. The data are only crude information and not knowledge by themselves. The data becomes information when it becomes relevant to our decision problem. The information becomes fact when the data can support it. Finally the fact becomes knowledge when it is used in the successful completion of decision problem. Then

$$\text{Data} \rightarrow \text{Information} \rightarrow \text{Facts} \rightarrow \text{Knowledge}$$

The statistical thinking process based on data will construct statistical models for decision making under uncertainty. Statistics arise from the need to place knowledge on a systematic evidence base. This required a study of the laws of probability. The level of exactness of statistical models increases when level of improvements on decision making increases. Thus the probability density function is a fundamental concept in statistics. The true probability density function, that we denote $f(.)$ is unknown. The model that we can think of as having given rise to the observation is usually very complex. A convenient framework is to consider that the observations are realizations of independent and identical random variables; then the whole model is specified by their common probability density function, $f(.)$. We call this distribution the true distribution or data generating distribution. A model is family of distribution and is called well-specified if it contains the true distribution; one may also speak of "true model" but this may be misleading. The data are insufficient to reconstruct every detail of $f(.)$. Then sometimes we estimate and sometimes we approximate this density. The density estimation area may be nonparametric or parametric. The nonparametric case is the construction of an estimate of the density function from the observed data where the number of parameters is considered as infinite. In this case, estimation of the density $f(.)$ over all points in its support would involve estimation of an infinite number of parameters. Historically we can say that the oldest used (nonparametric) density estimator is the histogram which has been refined for obtaining smooth estimators by the kernel approach, see, Fix and Hodges (1951), Devroye (1985) and Silverman (1986).See Figure 1.

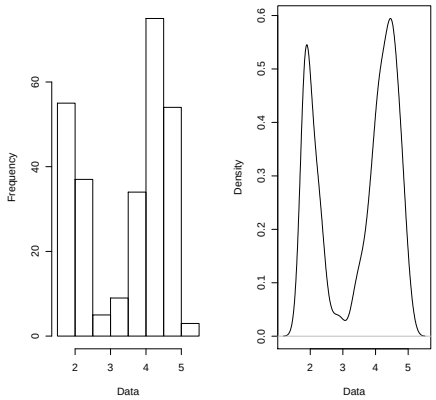


Figure 1: Histogram and kernel estimator of the same data. As we see these two approachs give us relatively the same information about the data generating probability.

The other case is the parametric case, where we assume the shape of the density function and we want only to estimate the parameters. In the parametric case we assume that the data are drawn from one of the known parametric families of distributions. The most widely used approach for such a construction is based on Maximum Likelihood Estimators (MLE) and some of its modifications. Generally this area is related to hypothesis testing problem. As natural property, we want to consider the methods for constructing procedures which are efficient, that is, asymptotically optimal. The theory behind the *MLE* guaranties this optimality. In hypothesis testing problem formally we define the null and alternative hypotheses about the parameters of the underlying density. The basic quantities that we need in hypothesis testing are the critical value that provides the desired level α, the power of test and the sample size required to achieve a given power. On the other hand we may compare two competing models, for example a normal density against a double exponential density.

The density approximation methodology is an alternative to kernel density estimation, but computationally as simple as parametric methods. It is based on the mode finding algorithm. Since we may imagine several models for approaching $f(.)$, the "Model Selection" issue arises. By model selection we remember Ockham (1282-1347) who stated that "Entities are not to be multiplied beyond necessity", which is known as Ockham's razor (Occam's razor). Simply a model is a set of equations, or functions with some adjustable parameters, or we may define a model as a sets of probabilistic, or statistical hypotheses. Model selection is the task of selecting a mathematical model from a set of potential models, i.e. determining the principle behind a series of observations. Some people however consider it as an intermediate step in model selection, and say that the model selection is to select a particular density from a model. We consider a family of densities as a model, where its members differ by the value of the parameters.

Our search is for the true model. We say a model is true if and only if one of the densities it contains is true. It is necessary to choose the set of models before beginning. In the linear regression model for instance, the model choice is difficult because we have 2^p potential model where p is the number of the explicative variables which are candidate to explanation of the response variable. The problem is that extra terms add extra adjustable parameters, and these will improve fit; the question however is "does an extra term added to an equation count as beyond necessity" if the gain in fit is too small?" If so, what counts as too small? How do we make this trade off between the addition of new parameters and gain in fit? And what is gained by the trade off? These are some questions in model selection. To show that the sensitivity of model selection, and the importance of this concept in research area consider two models for a set of data at hand, as $\{Y = \beta_0 + \beta_1 X_1 + \varepsilon, \quad \beta_0, \beta_1 \in \mathcal{R}; \quad \varepsilon \sim \mathcal{N}(0,1)\}$ and $\{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon, \quad \beta_0, \beta_1, \beta_2 \in \mathcal{R}; \quad \varepsilon \sim \mathcal{N}(0,1)\}$. We consider a simulation situation in which we know the first model is correct model, i.e we accept that $\beta_2 = 0$.

Figure 2 shows that including an explanatory variable that have zero coefficient in the model, changes the distribution of the estimator of β_1. Thus using the wrong model (second model) guides us to wrong inference about β_1, the parameter which must be in model.

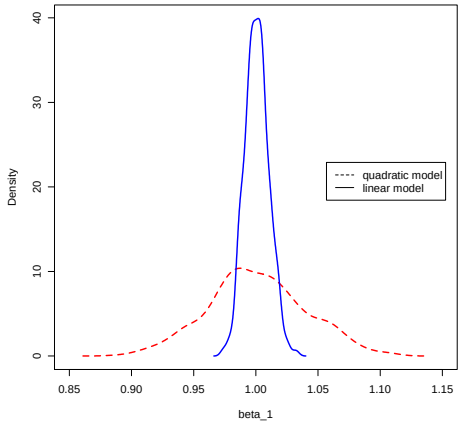


Figure 2: The solid density is much narraower than the dashed density. It shows that including predictors that have zero coefficient in the model will change the distribution of the estimate of β_1.

In the literature the classical method of model selection is known as goodness of fit test, Pearson (1900), and classical hypothesis testing, Neyman-Pearson (1933-1938), for one parameter models, and its extension which uses the Neyman-Pearson paradigm along with maximum likelihood estimator to give us a general-purpose testing procedure, the likelihood ratio test. On the other hand some recent methods in model selection criteria are Akaike information Criterion (AIC), (Akaike, 1973), the Bayesian information criterion (BIC) (Schwarz, 1978), Cross Validation technique, which is asymptotically equivalent to the *AIC* in the parametric case, and Minimum Description Length criterion, Bozdogan (2000) which is asymptotically equivalent to the *BIC*.

As a matter of fact we know that classical hypothesis testing with its extensive theory succeeds in goodness of fit. So why is there a need for other method of model selection? The answer is that this method does not extend straightforwardly to non-nested hypothesis and then with this method we can not make a deep analysis into the problem of model selection in real situations. Another important point is that the conclusion of the criteria like *AIC* are never about the truth or falsity of a hypothesis, but about its closeness to the truth. On the other hand it seems that the rational behind the classical hypothesis testing is minimization of the type I error and the type II error which are incompatible. But the actual practice is a trade off between these two errors. There is another objection to the rationale of classical hypothesis testing. It may be difficult to find a well-specified model (all models are wrong...). It may still be relevant to choose the best model among a set of (misspecified) models.

Our focus in this work is on maximum likelihood method and especially on *AIC*. This because *AIC* can be used for nested and non-nested models. The rationale of model choice is different from the classical testing approach. *AIC* adopts the Kullback-Leibler measures as its discrepancy function. In fact this statistic is an estimator of the relevant part of Kullback-Leibler (1951) discrepancy. Fisher, in his original introduction of the criterion of sufficiency, required "that the statistic chosen should summarize the whole of the relevant information supplied by the sample", and the Kullback-Leibler idea problem of discrimination is by considering a measure of the distance or discrepancy between statistical distributions in terms of their measure of information. We may also consider the Hellinger or Matusita distance of affinity, see Bar-Hen and Daudin (1998). In fact they defined the loglikelihood ratio as the information in observation for discriminating between two hypotheses related to the model selection. There are many ways of defining discrepancy, but in all of them the lack of fit is referred to as discrepancy. In the literature there are some other versions of Akaike's

criterion. In the linear regression model the most popular statistic for model choice is Mallows's C_p (Mallows, 1973). Other criteria are the corrected Akaike's criterion AIC_c proposed by Hurvith and Tsai (1989), the extended information criterion EIC by Ishiguro et al (1997); this approach has been extended to the choice of semi-parametric estimator by Commenges et al (2007) and ICOMP by Bozdogan (2000). They are in search of put enough weight on the quality of penalty term of the criterion which is related to the parsimony of the model.

On the other hand classically we may consider the models formulated as probability distribution. In fact model selection will be done in two steps. In the first step we must choose the set of models. The second step of model selection is well known as the estimation of parameters, i.e. once the set of possible models are selected, the mathematical analysis allows us to determine the 'best' of these models. Here, what means that the best? A good model selection technique will balance goodness-of-fit and complexity. Goodness of fit is generally determined in the minimum discrepancy (like Chi-square) or likelihood sense and the complexity is generally measured by counting the number of free parameters in the model. To select among competing models, one must decide which criterion to use to evaluate the models, and then make the best inference as to which model is preferable. As we said we may consider the discrepancy between the models as the criterion for model selection. Then our search will be find a model with minimum discrepancy from the true density which is sometimes completely unknown and sometimes unknown in parameter.

A kind of search is formulated as the hypothesis testing for model selection. An interesting work is done by Vuong (1989) who uses the Kullback-Leibler criterion to measure the closeness of a model to the true one. He considers the penalty term in *AIC* as a negligible quantity when the sample size gets large. Any way the *AIC* evaluation of models must agree with the likelihood choice or ordering of these models when the models have the same numbers of adjustable parameters. There

is an important period for model selection tests, from Cox to Vuong. The Vuong's test (1989) as a model selection test is different of Cox (1961) and Cox (1962) type test. By Cox test each model is evaluated against the data, i.e. the alternative model provides the power. In fact Cox test is a modification of Neyman-Pearson maximum likelihood ratio test. On the other hand the Vuong's test is a relative hypothesis test. In this kind of test the models are evaluated against the data and each other. Separation between the Cox's test and Vuong's test is important. The Cox test is for non-nested hypotheses and the Vuong's test is for non-nested model selection. It is necessary we emphasize that originally the likelihood ratio test is a statistical test of the goodness of fit test between two nested models. By this test a relatively more complex model is compared to a simpler model to see if it fits a particular dataset significantly better. Sometimes we refer to any test for model selection as goodness of fit test. But the goodness of fit tests as the approaches to model selection have their area and they are known as a category of model selection approaches. The goodness of fit tests frequently used by any researcher who need to statistical interpretation of their data and model selection. Historically for it was in 1900 when Pearson proposed the first test of goodness-of-fit, the χ^2 test to solve this problem. This basic test became a major source for the development of key areas in probability and statistics. There is no such method for unbind data. Fisher (1922) introduced the likelihood in the context of estimation. Although the obvious role of the likelihood function is to provide a point estimate for a parameter of interest, initially the likelihood is a tool for dealing with the uncertainty due to the limited amount of information contained in the data. It is the entire likelihood function that captures all the information in the data. Then in searching for an unbind goodness-of-fit test the likelihood function is a first candidate.

1.1 Our Objective

Our focus is on asymptotic theory for model selection. We study the situation under which model selection procedures are asymptotically optimal for selecting a model. We can say, all of things in our work is inference about the *AIC* (a kind of penalized likelihood), Akaike (1973), to model selection, where as an estimator for Kullback-Leibler discrepancy is intimately connected with maximum likelihood estimator. As a part of statistical inference, in the hypothesis testing context, the Kullback-Leibler divergence and the Neyman-Pearson lemma are two fundamental concepts. Both are about likelihood ratios. The Neyman-Pearson is about error rate of likelihood ratio tests and Kullback-Leibler divergence is the expected log-likelihood ratio. This connection introduces another interpretation of the Kullback-Leibler divergence in term of the loss of power of the likelihood ratio test when the wrong distribution is used for one of the hypothesis, i.e. the Kullback-Leibler divergence from two distribution functions P to Q measures how much power we lose with the likelihood ratio test if we mis-specify the alternative hypothesis P as Q. We want again to confirm that the Kullback-Leibler divergence estimator which is the (normalized) maximized likelihood function, asymptotically could be a good statistic for model selection. By this we eliminate the normalized part of likelihood ratio test, which is a cause that to inability the classical power study. In fact we want develop an approach to goodness-of-fit test based on normalized likelihood functions and normalized AIC's when the sample size gets large.

Our approach is based on *AIC* and difference of *AIC*'s for two competing models using confidence interval instead of hypothesis testing as its dual; it is because the confidence interval is a set of all acceptable hypotheses with pre-assigned confidence. The evaluation of a confidence interval for two competing nested or non-nested model is concentrated on whether the confidence interval has

contained zero or not. In brief we consider the *AIC* as a statistic which let us introduce a test statistic to model selection. This idea is different from the original idea about the *AIC* which considers the *AIC* as a criterion which allows to order the models. We want to emphasize that model selection could involve a trade-off between simplicity and fit. However there are many ways of making this trade-off. Essentially however, there is no method that is better than all the others under all conditions, i.e. for any methods $m1$ and $m2$, there are circumstances in which $m1$ is better than $m2$, and there are other circumstances in which $m2$ will do better than $m1$.

It seems that it is difficult to compare the methods, because it sometimes will guide us to an invalid conclusion. Instead to choose a method we can analyze our problem and precise our aim and the means to achieve our aim and finally to explain how a criterion works in achieving our aim. The area of model selection is very wide. A categorization of model selection problem can be considered according to whether the models are nested, overlap our non-nested. Generally two models are said to be non-nested if it is not possible to drive each of them from the other one either by means of an exact set of parametric restriction or as a result of a limiting process. The literature on non-nested hypothesis testing in statistics was pioneered by Cox (1961), Cox (1962) and Atkinson (1970), this subject applied by Pesaran (1974) and Pesaran and Deaton (1978). The analysis of non-nested regression models considered by Davidson and MacKinnon (1981), Fisher and McAleer (1981) and Dastoor (1983). Vuong (1989) considered the hypothesis testing when two competing models are nested, overlap and non-nested. His approach is based on the asymptotic distribution of difference of log-likelihood functions for two competing models. Shimodaira (1998) and Shimodaira (2001) has considered the sampling error of *AIC* in multiple comparisons and has constructed a set of good models rather than choosing a single model. Recently the asymptotic distribution of *AIC* in linear regression models and the bias correction of this statistics are discussed by Yanagihara and Ohomoto

(2005).

1.2 Plan of Thesis

In the remainder of this chapter we will bring some definitions, theorems and lemmas which will be frequently used. Chapter 2 is about theory of models. In chapter 3 we recall the goodness-of-fit tests as a base to introduce late a new approach and test statistics in goodness of fit test. Chapter 4 contains the assumptions and necessary instruments to develop our ideas in subsequent chapters. In chapter 5 we will propose a new test based on the likelihood ratio test for an empirical distribution function and we verify some aspects of this test. Chapter 6 concerns our proposed test when we want to test whether the unknown true density could be a member of a parametric family. This chapter is largely related to maximized likelihood function (and then *AIC*) and its asymptotic distribution, where we are interested in finding a criterion to achieve a reasonable model in multiple regression models. A simulation study is done which confirms our idea, see, appendix A. In chapter 7 we will introduce the difference of expected Kullback-Leibler divergence related to competing models to verify and a normalized difference of *AIC* as an estimator of it. The confidence interval as a dual of hypothesis testing is constructed to assess which model is better in Kullback-Leibler sense. The simulation study for logistic regression models confirms our idea in this chapter. We use our idea about real data when the variable under study is dichotomous. See appendix B.

Chapter 2

Reminders about models

and some asymptotic results

2.1 Models

The question of choosing a model is of course central in statistics. Usually we are not in the situation without any knowledge. We have a menu of rival models which could be used to describe the data. Let $\mathcal{M}$ denote a class of these candidate models. Each model $\mathcal{G} \in \mathcal{M}$ is considered as a set of probability distribution functions for our data, i.e.

$$\mathcal{G} = \{g(.,\beta) : \mathcal{R} \to \mathcal{R}^+; \beta \in B \subseteq \mathcal{R}^d\} = (g^\beta(.))_{\beta \in B}$$

where $g(.,\beta)$ denotes a probability distribution for observation Y and B represents the parameter space which can be different across different models $\mathcal{G}$. We note that in this framework that it may or may not be the case that one of the candidate models $\mathcal{G}$ in $\mathcal{M}$ is a correct model.

For example in a simple case may be we know that our observation has a χ^2 density but the true parameter of density is unknown. In the Figure 3 some of the members of χ^2 density is shown. The question is which member of this family is the data generating density.

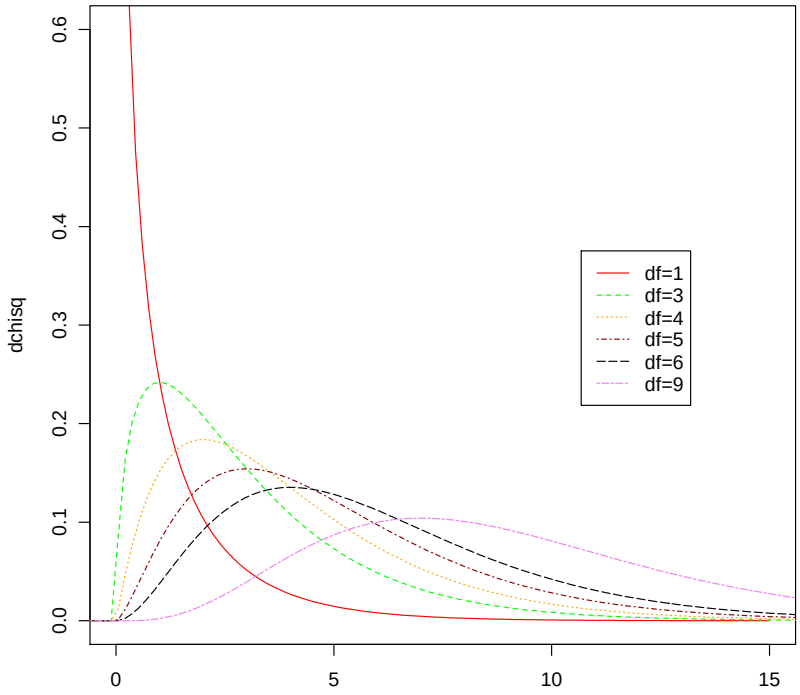


Figure 3: Example of some members of Chi squared family. As a step of model selection sometimes we must select a member of the family of densities.

As another example consider the normal linear model as $Y = X\beta + \varepsilon$ with usual hypotheses of normality and independence of the ε. We say this model is the full model. It is a suspected that some regressors i.e. some columns of X are not necessary to explaining Y, which means that the true values of the coefficients of the coefficient for these regressors are equal to zero, but which ones? Then the appropriate candidate models are all sub-models of the full model given by zero restrictions on the parameter vector.

2.2 Model Selection

As a starting point consider observations $\bar{Y} = (Y_1, Y_2, ...Y_n)$ from a scale and regression model of the form $Y = X\beta + \sigma\varepsilon$, where X is a fixed $n \times k$ matrix, $\beta \in \mathcal{R}^k$ is a vector of unknown regression coefficients, σ is a scale parameter, and ε is a vector of errors such that $(\varepsilon_1, ..., \varepsilon_n)$ is a random sample from a density $f(.)$. Popular choices for $f(.)$ include the normal, Student's t, logistic and Cauchy distrributions. On the other hand distributions on the positive real line include the exponential, gamma and so on. As a simple class of models consider the class $\mathcal{M}$ with two members as $\mathcal{G}_1 = \{\mathcal{N}(\mu, \sigma^2); \mu \in \mathcal{R}, \quad \sigma^2 \in \mathcal{R}^+\}$ and $\mathcal{G}_2 = \{\mathcal{C}(a,b); a \in \mathcal{R}, \quad b \in \mathcal{R}^+\}$ where $\mathcal{C}$ stands for Cauchy density. The model selection in the first step is choose between $\mathcal{G}_1$ and $\mathcal{G}_2$ and in the second step is choosing a member of the selected family in the first step. This two families for some of its members are shown in Figure 4. Model selection is a classical topic in statistics which concerns a vector of observation $\bar{Y} = (Y_1, Y_2, ..., Y_n)$ with the unknown density $f(.)$. The ultimate goal of model selection is to approach $f(.)$. As we said in the last section there are many possible models, that is sets of densities indexed by parameters. We denote a model as $\mathcal{G} = \mathcal{G} = \{g(., \beta) : \mathcal{R} \to \mathcal{R}^+; \beta \in B \subseteq \mathcal{R}^d\} = (g^\beta(.))_{\beta \in B}$. If we set $f(.) = g(., \hat{\beta}_n)$ where $\hat{\beta}_n$ a function of $\bar{Y}$ is the estimator of β, clearly there is

a risk which is known as the approximation risk. On the other hand if there is a member of $\mathcal{G}$, say $g(y;\beta_0)$ which is equal (or near) to $f(.)$ using the $g(.,\hat{\beta}_n)$ will introduce the other type of risk as the estimation risk.

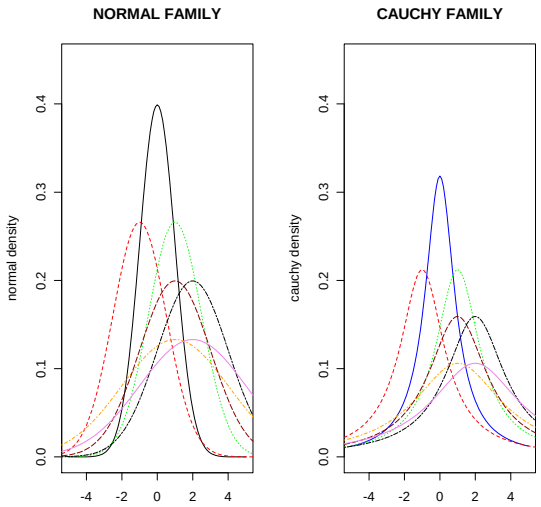


Figure 4: Two possible densities for data at hand. In a simple case of two candidate models, the model selection in the first step is choosing between two models and then choosing a member of the selected model.

The discrepancy between $f(.)$ and $g(.,\hat{\beta}_n)$ is known as bias term, which is in fact the mis-specification risk, and the discrepancy between $g(y;\beta_0)$ and $g(.,\hat{\beta}_n)$, is known as the variance term, which is a statistical risk, i.e.

$$Overall \quad Risk = Risk \quad of \quad Modeling + Risk \quad of \quad estimation$$

or

$$Overall \quad discrepancy = Bias \quad + \quad Variance$$

How we can minimize these two types of risks, is the model selection object. In fact model selection is the compromise between these two types of risks. Now if we had K models as $\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_K$ the model selection is in search of a model $\mathcal{G}_j, \quad j = 1, 2, ..., K$ which minimizes the discrepancy between $f(.)$ and $g(., \hat{\beta}_n)$. In model selection like the classical statistic we want to minimize the bias and the variance to find the optimal model which has the minimum risk.

2.3 Goal of Model Selection and its means

The goal of model selection depends on the research area. But as a common goal in model selection we are interested using the selected model in the prediction of the unobserved data. In the Akaike (1973) framework, a basic assumption is that the domain of unobserved data is the same as the domain in which the data are sampled, in other words we could think about new data as the data which re-sampled from observed data. Then there is a connection between model selection and predictive accuracy which is the expected fit of the unobserved data. But a point about the predictive accuracy is that its value for observed data is larger than its value for unobserved data. The fit can be assessed by the method of least squares or by the likelihood function. But the method of least squares have limitations. The question which arises is whether the likelihood approach applies to all cases? If the hypothesis is probabilistic, our hypothesis has a likelihood associated and we can choose a reasonable function of the likelihood as the model selection criterion. This function in the literature is known as the log-likelihood function. The only problem with the (log)-likelihood function is that this function depends on the sample size. To solve it, we normalize this function by the sample size. When we have the K competing models, in each model there is a vector of parameters. When we estimate the parameters of each models in fact in each K models we find a member which is

the best fitting under model. Now we can say that the aim of model selection is to maximize the predictive accuracy of the best fitting from the competing models. It is clear that when the estimation of the parameter(s) in likelihood sense in the model is a random variable the normalized maximized likelihood also is a random variable. This value minus the number of parameter in model divided by the sample size is a unbiased estimator of the predictive accuracy. This is the Akaike information criterion, *AIC*, for model selection, which states that we should choose the model with the lowest value of this criterion. But this criterion is used as if it were deterministic; we wish to change emphasize its statistical nature.

2.4 Nested and Non-Nested Models

We will bring the mathematical definition of nested and non-nested models in the next chapters, but simply we can say that two models are nested if one model can be reduced to the other model by imposing restriction on certain parameters. Two models are non-nested or completely separated if one model cannot be reduced to the other model by imposing restrictions on certain parameters. Also two models can be non-nested in terms of their functional forms and error structures. For example

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

and

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u$$

are two nested models. Discriminating between these two models, can be based on a t-test under ordinary least squares or a likelihood ratio test under either maximum likelihood or least squares. On the other hand

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

and

$$Y = \beta_7 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + u$$

are two non-nested models. Testing such models can be based on model selection tests using different concepts as for nested models.

2.5 Probability Metrics

The model selection is related to the distance between probability measures or densities. Determining whether a sequence of probability measures converges is a task for a statistician or a probabilist. The quantify that convergence in terms of some probability metric is all of things which we expect from a probability metric. In the literature there are a host of metrics to quantify the distance between probability measures. We should notice that some of them are not even metrics in the strict sense. Selecting a metric depends on our problem. Fortunately we can define a wide range of metrics. We set $(\Omega, \mathcal{F})$ as a measurable space and $\mathcal{M}$ be a space of all probability measures on $(\Omega, \mathcal{F})$. Then we consider convergence in $\mathcal{M}$, with P and Q as two probability measures on Ω and two density function with respect to σ-finite dominating measure which could be $(P+Q)/2$. By setting $\Omega = \mathcal{R}$. we can consider two distribution functions corresponding to the densities. By this assumption some measure of distance could be defined on probability measures, on densities or on distribution functions. Some of more important metrics in statistics are 1) **Discrepancy** metric, this metrics is in $[0,1]$ and is scale-invariant. 2) **Hellinger** distance, which define between two densities function, its value is in $[0,\sqrt{2}]$, see, Lecam (1986). 3) **Kullback-Leibler divergence (Relative entropy)**, Kullback-Leibler (1951), this criteria is defined on two densities and its value is in $[0,\infty]$. The relative entropy is not a metric, because it is not symmetric and does not satisfy the triangle inequality,

but it has many useful properties, including additivity over marginals of product measures, Cover and Thomas (1991). 4) **Kolmogorov (or Uniform)** metric, Kolmogorov (1933), this metric, is a distance between two distribution functions with value in $[0,1]$. This metric is invariant under all increasing one-to-one transformation on the line. 5) **Total variation** distance, its value is in $[0,1]$. 6) **Lèvy metric**, is a distance between two distribution functions and takes value in $[0,1]$, this measure is shift invariant but not scale invariant. 7) **Prohkorov (or Lèvy-Prokhorov)** metric, Prokhorov (1956), this metric is theoretically important because it metricizes weak convergence on any separable metric space, it assumes value in $[0,1]$. 8) **Separation** distance, this distance was advocated by Aldous and Diaconis (1987) to study Markov chains. However, it is not a metric and is in $[0,1]$ 9) **Wasserstein** metric, and 10) χ^2 **distance**, is defined on two densities and its value is on $[0,\infty]$, see Pearson (1900). This distance is not symmetric in its arguments and therefore not a metric. There are many inequalities between these metrics, but for our object one of the most important relation is related to the Kullback-Leibler divergence and other metrics. In section 3.4 we will talk about some of these inequalities.

2.6 Akaike framework and his Theorem

An inferential framework was developed by Hirotugu Akaike (1973) for thinking about how models are used to make prediction. But the prediction is for future data not for the data at hand (the old data). Prediction is of fundamental importance in all the science. Prediction accuracy is of obvious importance. Akaike not only introduced a framework in which predictive accuracy is the goal of inference, indeed provided a methodology for estimating a model predictive accuracy. Akaike introduced a criterion as Akaike information criterion (*AIC*) for model selection which is expressed

by a theorem. In fact he answered to the question as Given the data at hand, how is one to estimate how well a model will do in predicting new data, data that one does not yet have? The Akaike's theorem imposes a penalty term for complexity to the likelihood of the old data (goodness of fit) to describe how much of gain in likelihood there must be to off-set a given loss in simplicity. Naturally the Akaike's theorem has assumptions. First he defines the distance between a fitted model and the truth by using the Kullback-Leibler discrepancy. Second, he assumes that the new data will be drawn from the same underlying reality that generated the data at hand which has two parts: that the true function that connects independent to dependent variables is the same across data sets, and that the distribution that determines how the values of independent variables are selected is also the same. The Akaike's criterion is an unbiased estimator for Kullack-Leibler discrepancy, up to additive and multiplicative constants. This criterion allows to compare both nested and non-nested models as two important varieties in model selection. An other interpretation of *AIC* is that when this criterion is applied to the model selection, the number of the parameters of the model that it leads us to choose, can be viewed as an estimate of the number of parameters of the smallest correct model.

2.7 Complexity in model selection

Complexity is due to the number of parameters and functional form of the model, where the latter refers to the way in which the parameters are combined in the model equation. Many people believe that model selection should be based not only on goodness of fit, but must also consider model complexity. It seems clear that the goodness of fit is a necessary but not sufficient condition in model selection. An important consideration in model selection is to avoid choosing unnecessarily complex models because a simple model is more tractable, the stability of parameter estimates is greater and it

will generalize better to new data sets than a complex model which increase the predictive accuracy of the model. The complexity of model can be illustrated by considering the range of probability distributions of observations specified by the model equation that a model can occupy in model space. As an example consider 4 points as A, B, C and D in Figure 5, where they are modeled by a constant, a linear, a quadratic and cubic models. This is clear that the more complex model, the better the fit between the model and the points. This result is about the observed data, but consider a new data with $x = 5$. For this observation we have four prediction dependent on our model as $3.25, 6, 4.75$ and 17.04 respectively. By inspection it seems that a value about 5 is reasonable. This example shows that the better fit does not necessarily produce better inference.We show this result by Figure 5 which indicates that the complexity in mathematical form does not help to model selection. This is the meaning of Ockham's razor, see Chapter 1.

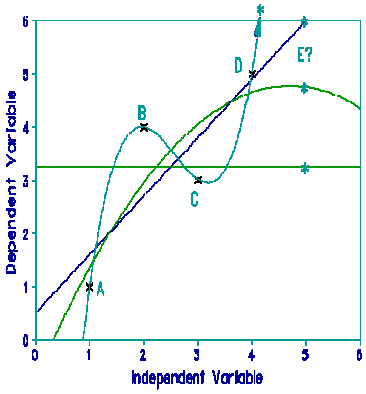


Figure 5: Four models are fitted to the data points, It shows it is not a case that the more complex model fits better than the simpler models to prediction.

All models occupy a section of model space. Then a simple model is a small section of this space and a complex model will occupy a large section of the model space. Myung (2000) in an example shows that the true model (quadratic model) but also more complex models (model of degree four and a non-linear model) can fit data well, which is why goodness of fit is not a sufficient condition for model selection. For example *AIC*, *BIC* and root mean squared deviation *RMSD* are some criteria for model selection which proposed that adjust for variation in the number of parameters as the complexity among models which was developed by Akaike (1973), Schwarz (1978) and Friedman et al (1995) respectively. Recently Information-theoretic measure of complexity (ICOMP) was developed by Bozdogan (1990, 2000). He considered two penalty terms as model complexity which are related to the covariance matrix of parameter estimates for the model. To show the role of complexity in model selection, consider a simulation study on regression model as follows. For nested models compute *AIC*, *BIC* and log-likelihood function. The result of simulation shows by Figure 6. We see that *AIC* and *BIC* have a minimum when we take three good explanatory variables, but the log-likelihood increases when the number of unuseful explanatory variables increases. To compare with usual criterion we also draw the *R* squared and adjusted *R* squared.

By these two simple examples we see that the model selection should be based not solely on goodness of fit, but must also consider model complexity. It is shown that model selection based only on the fit to observed data will result in the choice of an unnecessarily complex model that overfits the data. The effect of over fitting must be properly offset by model selection methods.

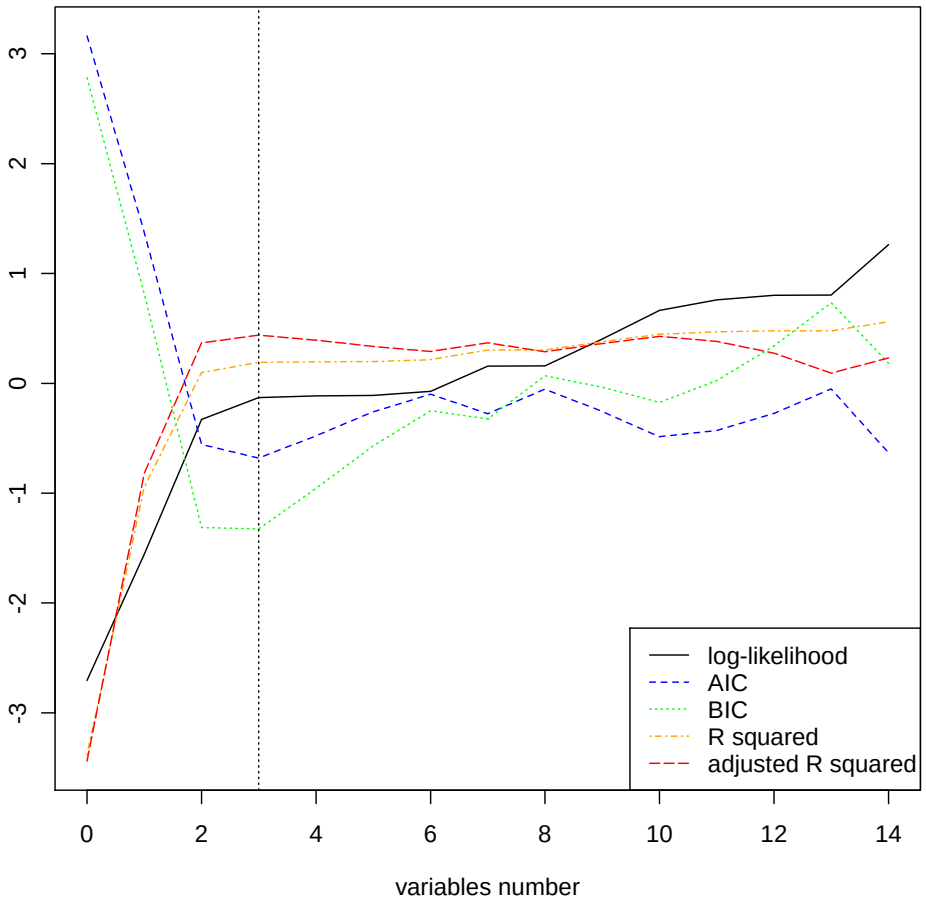


Figure 6: Comparison of some criteria to model selection.

2.8 Asymptotic theory

Asymptotic theory is a branch of statistics which have developed because of some theorems and relations in theory of probability. In fact asymptotic theory is concerned with the situation where the sample size is large or could be large. The most important of these theorems are the Weak and Strong Law of large numbers and Central limit theorems. By these theorems many problems in statistics are solvable. But there is a question, whether asymptotic results are useful, i.e. when using a sample with finite size, are we close enough to the asymptotic results ? Answer to this hard question involves the solution of the more difficult finite sample problem. If we want to defend the asymptotic theory we can say that this idea give insight into what constitutes a reasonable approach for the finite sample case. For example by this theory the maximum likelihood estimator becomes extremely popular, and in any area of science all of people use it without anxiety even for small sample.

Example 2.1 (A simple example of asymptotic distribution in hypothesis testing)

The asymptotic theory is a set of mathematical results useful in approximating the distribution of random elements. This random elements in general could be any statistics. To illustrate why this approximations are useful tools in hypothesis testing we can consider a known and simple case where $Y_i \sim \mathcal{N}(\mu, \sigma^2)$ and we are interested in testing the hypothesis $\mathcal{H}_0 : \mu = \mu_0$ for some specified value of μ_0. One way to test is to form the statistic $H_n = \frac{\sqrt{n}(\overline{Y}-\mu_0)}{\sigma}$ when $\overline{Y}$ is a simple average of n i.i.d random variables $Y_i; i = 1,2,...,n$. If the true variance σ^2 is known then $H_n \sim \mathcal{N}(0,1)$. By this we could construct a rejection region and make a decision about $\mathcal{H}_0$. When σ^2 is an unknown parameter we form the statistic $\overline{H}_n = \frac{\sqrt{n}(\overline{Y}-\mu_0)}{\hat{\sigma}}$ where $\hat{\sigma}$ is an estimate for σ, in this case $\overline{H}_n \sim t_{n-1}$ and could thus again construct a rejection region.

We note however, that when n gets large, the t-student distribution approaches the standard normal distribution. This suggest that in large samples we would not make a big mistake by ignoring the fact that σ is estimated rather than known a priori. Now consider the situation where Y_i is not known to be from a normal distribution.

Note that it still holds that $\mathcal{E}_{\mathcal{H}_0}(\overline{Y}) = \mu_0$. When $\mathcal{H}_0$ is true we would still expect H_n or $\overline{H}_n$ to be close to zero. It then seems reasonable to continue to use H_n or $\overline{H}_n$ as test statistics. The problem however is that we no longer know the distribution of these two statistics and thus are unable to construct a test. This is a situation where an approximation to the distribution of H_n is useful. It is known that by central limit theorem H_n will have a limit distribution which is very close to standard normal distribution $\mathcal{N}(0,1)$, see Figure 7.

On the other hand $\overline{H}_n = H_n \frac{\sigma}{\hat{\sigma}}$, now if $\hat{\sigma}$ be the maximum likelihood estimator for σ by the weak law of large numbers we have that $\frac{\sigma}{\hat{\sigma}} \xrightarrow{\mathcal{P}} 1$ then by Slutsky's theorem $\overline{H}_n$ asymptotically is $\mathcal{N}(0,1)$ and we can construct a rejection region.

2.9 Goodness of Fit Test and Classical Hypothesis Testing

Hypothesis testing is generally formulated in terms of null and alternative hypotheses, type one and type two errors and the power of test. If we ask which test is better, the answer is that the test which has a highest power among all possible tests (for fixed type I error), i.e. an ideal test is uniformly most powerful test. If we are not able to find a uniformly most powerful test, we turn to the search of a test with an acceptable power function. In all-purpose goodness of fit tests framework there can be no optimal test, because there is no specific alternative hypothesis, so it is impossible to define

the power of the test simply. In goodness of fit test we do not have any clear criteria for choosing one goodness of fit test procedure over another, i.e. one can propose a goodness of fit test and a computational method(s). To verify a proposed test we are restricted to verify the power of our test against a few alternatives. We must notice that these alternatives must be carefully chosen.

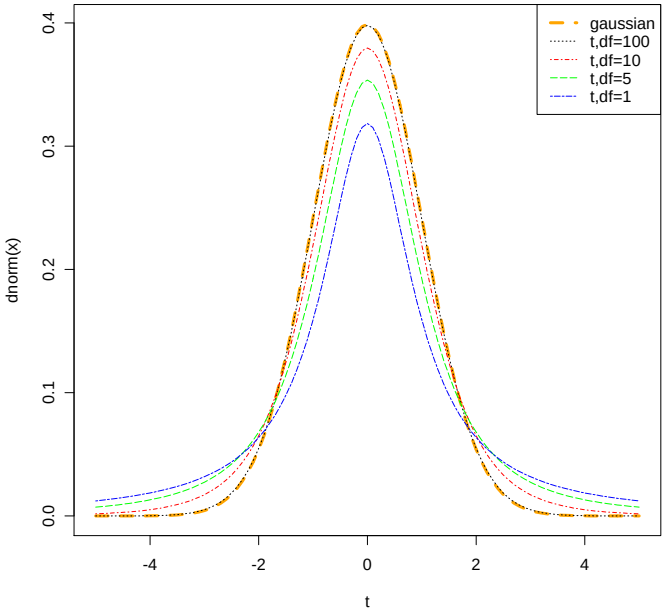


Figure 7: Some members of t-student model with standard normal density as its limit distribution.

2.10 Reminder on Theorems and Lemmas

The following lemmas and theorems will be used in this work.

Lemma 2.1 (A1) *If $Y_n \xrightarrow{\mathcal{L}} Y$, and a_1, a_2 are constants with $a_2 \neq 0$, then $a_2 Y_n + a_1 \xrightarrow{\mathcal{L}} a_2 Y + a_1$.*

Theorem 2.1 *(B1) [Central limit theorem(CLT), i.i.d. case (Lindeberg-Lévy)] Let $Y_i, i = 1, 2, ..., n$ be i.i.d. with mean μ and finite variance σ^2. Then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Y_i - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

The Lindeberg-Lévy CLT is a special case of Lindeberg-Feller or Lyapunov CLT for not necessarily identically distributed independent random variables. CLT for dependent variables is also established, see Lehmann (1998).

Theorem 2.2 *(B2) [Weak law of large numbers]. Let $Y_i, i = 1, 2, ..., n$ be i.i.d. with mean μ and finite variance σ^2. Then*

$$\frac{1}{n} \sum_{i=1}^{n} Y_i \xrightarrow{\mathcal{P}} \mu$$

Theorem 2.3 *(B3) Suppose $Y_1, Y_2, ..., Y_n$ i.i.d. with density $f(., \theta)$ where θ is fixed at some arbitrary value in the parameter space Θ. Let φ is a function of $f(., \theta)$ and $W(Y; \theta) = \varphi(Y, \theta) - \mathcal{E}_f\{\varphi(Y; \theta)\}$ be a measurable function of y for all θ, and a continuous function of θ for almost all y. Suppose that $(i) \Theta$ is compact, and that $(ii) \frac{1}{n} \sum_{i=1}^{n} w(Y_i; \theta)$ converges to zero in probability on Θ. Then if $(iii) |\varphi(y; \theta)| < g(y)$ for some function g satisfying $\mathcal{E}_f\{g(Y)\} < \infty$ then we have*

$$\sup_{\theta \in \Theta} |\frac{1}{n} \sum_{i=1}^{n} W(Y_i; \theta)| \xrightarrow{\mathcal{P}} 0$$

The history of this theorem come back to definition of stochastically equicontinuous functions for example Billingsley (p. 55 1968), Billingsley (p. 355 1995) and Andrews (1992). In fact this

theorem is a combination of two theorems. The first one says that under the assumptions (i) and (iii) $\frac{1}{n}\sum_{i=1}^{n}\varphi(Y_i;\theta)$ is stochastically equicontinuous and that $\mathcal{E}_f\{\varphi(Y;\theta)\}$ is (equi) continuous. The second one says that if $\frac{1}{n}\sum_{i=1}^{n}W(Y_i;\theta)\xrightarrow{P}0$ theorem B3 is right.

Theorem 2.4 *(B4)[Slutsky's theorem] Let Y_n, Y, W_n be random vectors or variables. If $Y_n \xrightarrow{\mathcal{L}} Y$ and $W_n \xrightarrow{\mathcal{L}} c$, for a constant c, then*

(i) $Y_n + W_n \xrightarrow{\mathcal{L}} X + c$

(ii) $W_n Y_n \xrightarrow{\mathcal{L}} cY$

(iii) $W_n^{-1} Y_n \xrightarrow{\mathcal{L}} c^{-1}Y$ provided $c \neq 0$

Where sometimes we have to consider c as a scaler and sometimes as a vector.

Note that no restrictions are imposed on the possible dependence among the random variables involved.

Theorem 2.5 *(B5) If $Y_n \xrightarrow{\mathcal{P}} Y$, then also $Y_n \xrightarrow{\mathcal{L}} Y$.*

Theorem 2.6 *(B6) $Y_n \xrightarrow{\mathcal{P}} c$, for a constant c if and only if $Y_n \xrightarrow{\mathcal{L}} c$.*

Theorem 2.7 *(B7) [Continuous mapping]. Let $g : \mathcal{R}^k \to \mathcal{R}^m$ be continuous at every point of a set $S \subset \mathcal{R}^k$ such that $p(Y \in S) = 1$.*

(i) If $Y_n \xrightarrow{\mathcal{L}} Y$, then $g(Y_n) \xrightarrow{\mathcal{L}} g(Y)$.

(ii) If $Y_n \xrightarrow{\mathcal{P}} Y$, then $g(Y_n) \xrightarrow{\mathcal{P}} g(Y)$.

The continuous mapping theorem has many important applications that are based on the following simple convergence theorem. Assume that $Y_n \xrightarrow{\mathcal{P}} c$ where c is a constant and $W_n \xrightarrow{\mathcal{L}} W$, then we have $\begin{pmatrix} Y_n \\ W_n \end{pmatrix} \xrightarrow{\mathcal{L}} \begin{pmatrix} c \\ W \end{pmatrix}$ jointly. Now the Slutsky's theorem is a simple application of the continuous mapping theorem.

Chapter 3

Reminder on Goodness of Fit Tests

3.1 Testing fit to a fixed distribution

The goodness-of-fit (gof) are used for verifying whether or not the experimental data come from the postulated model. In this direction one must decide if theoretical and experimental distributions are the same. Then gof is a hypothesis testing problem and the problem is concerned with the choice of one of these two alternative hypothesis

$$\mathcal{H}_0 : F(y) = F_0(y) \quad \forall y$$

$$\mathcal{H}_1 : F(y) \neq F_0(y)$$

for a fixed distribution function F_0.

In fact we can put gof tests into two classes. The first class divides the range of the data into disjoint cells and compares the observed numbers to the expected number under the hypothesized distribution. Naturally they are useful for discrete case but we can use them in continuous case also.

The second class of tests are used for continuous distributions. For these types, we compare an empirical distribution function of the data with a distribution function under $\mathcal{H}_0$. The test statistic for these tests is based on a measure of correlation between the distributions or based on some measure of distance between the two distribution functions. A good reference for gof tests is D'Agostino and Stephens (1986).

The most popular goodness of fit test is due to Pearson (1900). As a new look to this statistic, for any generalized linear model, the Pearson goodness of fit test is the score test statistic for testing the postulated model against the saturated model. The relationship between the Pearson statistic and the residual deviance is therefore the relationship between the score test and the likelihood ratio test statistics.

3.1.1 Basic Goodness of Fit Test

The most important goodness of fit test goes back at least to Pearson's Chi-squared test (1900). He establishes the asymptotic χ^2 distribution for a goodness of fit statistic for the multinomial distribution. It can be useful in both discrete and continuous cases when the data be grouped into classes (or cells). This statistic is given by

$$X^2 = \sum_{j=1}^{k} \frac{(O_j - np_j)^2}{np_j} = D_{k-1}^T \Sigma_{k-1}^{-1} D_{k-1}$$

where O_j is the number of observations in cell C_j, $p_j = P_{\mathcal{H}_0}(Y \in C_j)$ then by this definition we have:

$$D_m = n^{-1/2}(O_1 - np_1, ..., O_m - np_m)^T \xrightarrow{\mathcal{L}} N \sim (0, \Sigma_m) \quad for \quad m \leq k$$

$$\Sigma_m = \left(\sigma_{ij}\right)$$

$$\sigma_{ij} = \begin{cases} -p_i p_j & \text{if } i \neq j \\ p_i(1-p_i) & \text{if } i = j \end{cases}$$

$$\Sigma_{k-1}^{-1} = \left(\delta_{ij}\right)$$

$$\delta_{ij} = \begin{cases} p_k^{-1} & \text{if } i \neq j \\ p_i^{-1} + p_k^{-1} & \text{if } i = j \end{cases}$$

A well known result in the asymptotic theory of tests of fit says that under $\mathcal{H}_0$

$$\chi^2 \xrightarrow{\mathcal{L}} \chi^2_{k-1}$$

In the continuous case the $\mathcal{X}^2$ statistic will not distinguish two different distributions sharing the same cell probabilities. It is because we look only to the cell frequency which produces a loss of information that results in lack of power.

3.1.2 Tests on the basis of Functional Distance

A proposed way to improve the Pearson's statistics is by employing a functional distance to measure the discrepancy between hypothesized distribution F_0 and the empirical distribution function F_n where for i.i.d. random variables $Y_1, Y_2, ..., Y_n$ is defined as

$$F_n(y) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,y]}(Y_i).$$

The first one of this type is the test statistics which to known as the Cramér-Von Mises type statistics. Here we reintroduce them and others in brief.

Cramér (1928) and in a more general form Von-Mises(1931) proposed

$$\omega_n^2 = n\int_{-\infty}^{\infty} (F_n(y) - F_0(y))^2 \zeta(y)\, dy$$

for some weight function ζ as an adequate measure of discrepancy.

The Kolmogorov test (1933) is the easiest and also most natural non-parametric test. It is based on the L_∞ norm and computes the distance between an empirical and the theoretical distribution function under the null hypothesis. Under $\mathcal{H}_1$ the difference between the empirical and theoretical distribution functions will be noticeable. This statistic is given by

$$D_n = \sqrt{n} \sup_{y \in \mathcal{R}} |F_n(y) - F_0(y)|$$

A problem mathematically similar to Kolmogorov's was studied by Smirnov (1939,1941) he has considered D_n^+ and D_n^- where

$$D_n^+ = \sqrt{n} \sup_{y \in \mathcal{R}} (F_n(y) - F_0(y))$$

$$D_n^- = \sqrt{n} \sup_{y \in \mathcal{R}} (F_0(y) - F_n(y))$$

The statistics D_n, D_n^+ and D_n^- are known as Kolmogorov-Smirnov statistics. They have the advantage of being distribution free. Thus the same p-values can be used to obtain the significance level when testing it to any continuous distribution.

In search of this property for ω_n^2 has introduced a simple modification. A modification for Cramér-Von Mises distance is

$$W_n^2(\psi) = n \int_{-\infty}^{\infty} \psi(F_0(y)) \{(F_n(y) - F_0(y))^2\} \, dF_0(y)$$

which was proposed by Smirnov (1936-1937). All the statistics which can be obtained by varying ψ as we said are usually refereed to as statistics of Cramér-Von Mises type, two of them are as follows. The Cramér-Von Mises's statistic obtained by W_n^2 for $\psi(.) = 1$,

$$W_n^2 = n \int_{-\infty}^{\infty} (F_n(y) - F_0(y))^2 \, dF_0(y)$$

and the Anderson-Darling's statistic (1955) for $\psi(t) = (t(1-t))^{-1}$,

$$A_n^2 = n \int_{-\infty}^{\infty} \frac{(F_n(y) - F_0(x))^2}{F_0(y)(1-F_0(y))} \, dF_0(y)$$

Consideration of different weight functions ψ allows the statistician to put special emphasis on the detection of particular sets of alternatives. Some people prefer employing Cramér-Von Mises statistics instead of Kolmogorov-Smirnov statistics; it is because Kolmogorov-Smirnov statistics accounts only for the largest deviation between $F_n(t)$ and $F(t)$, while the other one is a weighted average of all the deviations between $F_n(t)$ and $F(t)$. Anyway we reject $\mathcal{H}_0$ if in each case the value of the statistic is large.

3.2 Adaptation of tests coming from the fixed-distribution

All the procedures in the last section were based on a distribution obtained from a sample and fixed distribution. A way to adapt this idea for the parametric case is replacing the fixed distribution by $F(.,\theta)$, that is by a model. Historically it was Pearson who suggested

$$\hat{\chi}^2 = \sum_{j=1}^{k} \frac{(O_j - np_j(\hat{\theta}_n))^2}{np_j(\hat{\theta}_n)}$$

where $p_j(\hat{\theta}_n)$ denotes the probability under $F(.,\theta)$, that Y_1 falls into cell j. At these times, Pearson did not realize that the estimation of parameters changes the asymptotic distribution of $\hat{\chi}^2$. It was Fisher who pointed out that if $\hat{\theta}_n$ is the maximum likelihood estimator $\hat{\chi}^2$ has an asymptotic χ^2 distribution. He also pointed out that the estimating parameter from the grouped data instead of the complete data will cause a loss of information resulting in lack of power. Chernoff and Lehmann (1954) is a good reference for the parametric case. The choice of cells is an important part of the search for asymptotic distribution of Pearson's statistic. Because the distribution of Pearson statistic

is a consequence of the asymptotic normality of the cell frequencies, then it will be sensitive to the magnitude of these frequencies. Hence , combining neighboring cells with few observations is suggested by Cochran (1952).

Adaptation for $W_n^2(\psi)$ and $\hat{D}_n$ are

$$\hat{W}_n^2(\psi) = n\int_{-\infty}^{\infty} \psi(F(y,\hat{\theta}_n))\{(F_n(y) - F(y,\hat{\theta}_n))^2\}\, dF(y,\hat{\theta}_n)$$

and

$$D_n = \sqrt{n}\sup_{y\in\mathcal{R}} |F_n(y) - F(y,\hat{\theta}_n)|$$

respectively. Unfortunately in general the nice property exhibited by $W_n^2(\psi)$ and D_n of being distribution free does not carry over to the parametric case. The asymptotic distribution of these two statistics is due to Darling (1955). He showed that these asymptotic distributions are a function of a Gaussian process.

3.3 Tests on the basis of Correlation and Regression

Goodness-of-fit tests in this subsection focus on the analysis of the probability plot. We consider $\mathcal{F}^{ls}$ as a location scale family of distribution functions i.e. given a probability measure K_0, we will assume that $\mathcal{F}^{ls}$ is the family of distribution functions obtained from K_0 by location or scale changes. Assume that K_0 is standardized and suppose that $Y_1, Y_2, ...Y_n$ is an i.i.d. sample whose common distribution belongs to $\mathcal{F}^{ls}$ and has mean μ and variance σ^2 . In fact we want to test that

$$\mathcal{H}_0 : F(y) = K_0(\frac{y-\mu}{\sigma})$$

Let $Y_r = (Y_{(1)}, Y_{(2)}, ..., Y_{(n)})$ be the corresponding ordered statistics and $W_r = (W_{(1)}, W_{(2)}, ..., W_{(n)})$ be an ordered sample with underlying distribution function K_0 and let $m' = (m_1, m_2, ..., m_n)$ and

$V = (v_{ij})$ be, respectively, the mean vector and the covariance matrix of W_r. If $\mathcal{H}_0$ is true

$$W_{(i)} = \frac{Y_{(i)} - \mu}{\sigma}, \quad \text{in distribution}, \quad i = 1, 2, ..., n$$

Then the plot of $Y_{(1)}, Y_{(2)}, ..., Y_{(n)}$ against the points $m_1, m_2, ..., m_n$ should be approximately linear. Lack of linearity in this plot suggests that the distribution of Y_i does not belong to the family of distribution in $\mathcal{H}_0$ and then we would expect to see some curvature. Checking this linearity is often done by eye. However, some analytical approaches have been devised to test it. On the other hand we know that

$$S_n^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

is a consistent estimator for σ^2 on the other hand

$$\hat{\mu}_{BLUE} = \overline{Y}_n$$

and

$$\hat{\sigma}_{BLUE} = \frac{m' V^{-1} Y_r}{m' V^{-1} m}$$

are the best linear unbiased estimator of μ and σ. Hence, under the null hypothesis, $\frac{\hat{\sigma}^2_{BLUE}}{S_n^2}$ should be near to 1.The Shapiro-Wilk (1965) or W-test statistic is a normalized version of $\frac{\hat{\sigma}^2_{BLUE}}{S_n^2}$,

$$W_n = \frac{(m' V^{-1} Y_r)^2}{(m' V^{-1} V^{-1} m) \sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

It is clear that $W_n \in [0, 1]$ and the small value of this statistic would lead to rejection of the null hypothesis. According to simulation by Shapiro et al.(1968) it seems that the W-test is one of the most powerful normality tests against a wide range of alternatives. A weakness of the W-test is that the procedure may be not consistent for testing fit to non-normal families of distributions and also computation of this test requires previous computation of m and V^{-1}.

The Shapiro-Francia test is based on replacing matrix V^{-1} by the identity I which defined as

$$W_n' = \frac{(m'Y_r)^2}{(m'm)\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

the computation of which is easier than W_n. A further simplification of the W_n' was proposed by Weisberg et al (1975) by replacing m by the vector $M = \Phi^{-1}(\frac{i-3/8}{n+1/4}), \quad i = 1,2,...,n$, and Φ denotes the standard Gaussian distribution function. This statistic is easier to compute than W_n'

3.4 Tests on the basis of Likelihood Functions

3.4.1 Berk-Jones's statistics

Berk and Jones (1979) have defined a test statistic on the basis of hypothesized and empirical distribution function in a fixed point y. Then for fixed y we have

$$nF_n(y) \sim Bin(n, F(y))$$

$$\lambda_n(y) = \frac{\sup_{F(y)} \mathcal{L}_n(F(y))}{\mathcal{L}_n(F_0(y))} = \frac{\mathcal{L}_n(F_n(y))}{\mathcal{L}_n(F_0(y))} =$$
$$\{\frac{F_n(y)}{F_0(y)}\}^{nF_n(y)} \{\frac{1-F_n(y)}{1-F_0(y)}\}^{n(1-F_n(y))}$$

by defining

$$\log \lambda_n(y) = nK(F_n(y), F_0(y))$$

where

$$K(F_n(y), F_0(y)) = F_n(y)\,(\log(\frac{F_n(y)}{F_0(y)})) + (1-F_n(y))\,\log\frac{1-F_n(y)}{1-F_0(y)}$$

the Berk-Jones's statistics is given by

$$R_n = \sup_{y\in\mathcal{R}} n^{-1} \log \lambda_n(y) = \sup_{y\in\mathcal{R}} K(F_n(y), F_0(y)).$$

We reject $\mathcal{H}_0$ for large value of R_n. Under $\mathcal{H}_0$

$$nR_n - \log\log(n) + \frac{1}{2}\log\log\log(n) - \frac{1}{2}\log(4\pi) \xrightarrow{\mathcal{L}} W$$

where

$$F(w) = e^{-4e^{-y}}$$

Einmahl and McKeague (2003) propose an integral statistic T_n defined by

$$T_n = 2\int_0^1 K(F_n(y), F_0(y))\, dF_0(y)$$

Jager et al (2005) introduced a related statistic, the "reversed Berk-Jones statistic" which differs from the Berk-Jones statistic.

3.4.2 Generalized Linear Models (GLMs) and Deviance

The generalized linear models expresses the means of the response variables as some function of a linear combination of the explanatory variables

$$\mathcal{E}\{Y|X\} = \Upsilon(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k)$$

where the form of the function $\Upsilon(.)$ is known and the parameters of the model $\beta_0, \beta_1, ..., \beta_k$ are not known. If the function $\Upsilon(.)$ is the identity function and Y has the normal distribution this model is the simple linear model.

An issue is to evaluate the relevance of our model for our data and how well it fits the data (gof). The strategy is finding a simple model but with a good fit (the principle of parsimony). GLMs, McCullagh and Nelder (1989), provide a fairly simple, but widely useful extension of the usual Normal linear model. Start with the standard linear model meeting the Gauss-Markov conditions

with $p = k+1$

$$\underset{(n\times 1)}{Y} = \underset{(n\times p)(p\times 1)}{X\beta} + \underset{(n\times 1)}{\varepsilon}$$

$$\underset{(n\times 1)}{\mathcal{E}(Y)} = \underset{(n\times 1)}{\theta} = \underset{(n\times p)(p\times 1)}{X\beta}$$

X: Matrix of observed data values.

$X\beta$: Linear structure vector.

ε: Error terms .

Y: A variable which is distributed as i.i.d normal random with mean θ and constant variance σ^2.

Generalization

We generalize this with a new "linear predictor" based on the mean of the outcome variable Y which is no longer required to be normally distributed or even continuous.

$$\underset{(n\times 1)}{\xi(\mu)} = \underset{(n\times 1)}{\eta} = \underset{(n\times p)(p\times 1)}{X\beta}$$

where $\xi(.)$ be an invertible, smooth function of the mean vector $\mu = \mathcal{E}(Y)$.

The effect of the explanatory variables is now expressed in the model only through the link from the linear structure, $X\beta$, to the linear predictor, $\eta = \xi(\mu)$, controlled by the form of the link function, $\xi(.)$. This link function connects the linear predictor to the mean of the outcome variable not directly to the expression of the outcome variable itself, so the outcome variable can now take on a variety of non-normal forms. The link function connects the stochastic component describes some response variable from a wide variety of forms to all of the standard normal theory supporting the linear systematic component through the mean function

$$\xi(\mu) = \eta = X\beta$$

$$\xi^{-1}(\xi(\mu)) = \xi^{-1}(\eta) = \xi^{-1}(X\beta) = \mu = \mathcal{E}(Y)$$

In general we suppose that the stochastic component Y is distributed according to a member of exponential family with mean μ as follows

$$g(y;\theta_i,\varphi) = \exp\{\frac{(y\theta_i - b(\theta_i))}{\varphi/w_i} - c(y,\varphi)\}$$

where the weight w_i is a known constant and $\varphi > 0$ is a scale parameter (often it is considered as a nuisance parameter). The stochastic and systematic components are linked by a function of η which is taken from the inverse the of the canonical link, $b(\theta)$. Given b the function c is determined by the requirement the g integrates to one. The GLMs are free of the assumption that the residuals have mean zero and constant variance, but there are more complex stochastic structures. We may consider the residual as

$$R = Y - \xi^{-1}(X\beta)$$

but this does not provide the nice distribution theory we get from the standard linear model.

Deviance

The deviance function is a residual function for generalized linear models. This is built in a similar fashion as the likelihood ratio statistics, comparing the log-likelihood from a proposed model specification to the maximum log-likelihood possible through the saturated (full or maximal) model. The saturated model is loosely defined as the model where the number of parameters equals the number of data points. The resulting difference is multiplied by two and called the summed deviance. In fact in saturated model there is basically one parameter per observation then we can think of this as the most general model possible with the maximum number of parameters that can be estimated.

The deviance assesses the goodness of fit for the model by looking at the difference between two log-likelihood functions. The resulting difference is multiplied by two and called the summed Deviance. The goodness of fit intuition is derived from the idea that this sum constitutes the summed

contrast of individual likelihood contributions with the native data contributions to the saturated model. As we say the point here is to compare the log-likelihood for the proposed (current) model

$$l(\hat{\theta},\varphi|\underline{y}) = \sum_{i=1}^{n}\{\frac{y_i\hat{\theta}_i - b(\hat{\theta}_i)}{\varphi/w_i} + c(y_i,\varphi)\}$$

to the same log-likelihood function with identical data and the same link function, expect that it now with n coefficients for the n data points, i.e. the saturated model log-likelihood function

$$l(\tilde{\theta},\varphi|\underline{y}) = \sum_{i=1}^{n}\{\frac{y_i\tilde{\theta}_i - b(\tilde{\theta}_i)}{\varphi/w_i} + c(y_i,\varphi)\}.$$

The latter is the highest possible value for the log-likelihood function achievable with the given data, then

$$l(\tilde{\theta},\varphi|\underline{y}) \geq l(\hat{\theta},\varphi|\underline{y})$$

The deviance function is then given by

$$\mathcal{D}(\theta,\underline{y}) = 2\sum_{i=1}^{n}[l(\tilde{\theta}_i,\varphi|\underline{y}) - l(\hat{\theta}_i,\varphi|\underline{y})] = 2\sum_{i=1}^{n}[y_i(\tilde{\theta}_i - \hat{\theta}_i) - (b(\tilde{\theta}_i) - b(\hat{\theta}_i)](\varphi/w_i)^{-1}.$$

This statistic under some conditions is asymptotically χ^2_{n-k}. The conditions will be discussed for each type of response data individually. In fact the distribution of the deviance is approximately $\chi^2_{n-k,\nu}$, where ν is the non-centrality parameter. When the Y_i's are normal and the link is identity function and the variance is known the deviance has a exact χ^2 distribution. Otherwise we will consider the ratio of mean deviances, which does not involve the scale parameter in the exponential dispersion family. In general, we use the deviance in goodness-of-fit tests for Poisson and Binomial GLMs where we can calculate the deviance from the data and there is no unknown parameter. On the other hand it is noticeable that sometimes the deviance is not informative. For example for Bernoulli observations the Deviance depends on the sufficient statistics not the individual observation and so is, of little use for measuring goodness of fit.

Dependence of degree of freedom to n when we talk about asymptotic density seems irrelevant. In fact when we consider θ as a known vector with fixed length k

$$\mathcal{D}(\theta,\underline{y}) = 2\sum_{i=1}^{n}[l(\widetilde{\theta}_i,\varphi|\underline{y}) - l(\theta_i,\varphi|\underline{y})] = 2\sum_{i=1}^{n}[y_i(\widetilde{\theta}_i-\theta_i) - (b(\widetilde{\theta}_i)-b(\theta_i)](\varphi/w_i)^{-1} \sim \chi^2_k$$

if $\mathcal{E}(\hat{\theta}) = \theta$ and under mild regularity conditions. The proof is given in any statistical standard book as example Lehmann (1986), but this proof does not generally hold for saturated models because the length of θ is not fixed and grows with sample size. Clearly a small deviance implies a good fit.

The deviance function depends on φ then simply the unscaled deviance function is defined as

$$\varphi\mathcal{D}(\theta,\underline{y}) = 2\sum_{i=1}^{n} w_i[y_i(\widetilde{\theta}_i-\theta_i) - (b(\widetilde{\theta}_i)-b(\theta_i)]$$

Example 3.1 (Normal (linear) model)

If Y is distributed according the normal model we have

$$g_Y(y;\theta,\varphi) = (2\pi\sigma^2)^{-1/2}\exp\{-(y-\mu)^2/2\sigma^2\} =$$

$$\exp\{(y\mu-\mu^2/2)/\sigma^2 - \frac{1}{2}(y^2/\sigma^2+\log(2\pi\sigma^2)\}$$

Now for $Y_1, Y_2, ..., Y_n$ from the $\mathcal{N}(\mu_i,\sigma^2)$ where μ_i's are distinct we have $\theta_i = \mu_i$, $\varphi = \sigma^2$, and

$a(\varphi) = \varphi$, $b(\theta_i) = \theta_i^2/2$, $c(y,\varphi) = -\frac{1}{2}\{y^2/\sigma^2+\log(2\pi\sigma^2)\})$

for a sample with tail n $\widetilde{\theta}_i = y_i$ and $\hat{\theta}_i = \hat{\mu}_i$ then ,

$$\mathcal{D}(\underline{y};\hat{\mu}) = \sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{\sigma^2}$$

the residual sum of squares. This deviance is a function of unknown parameter σ^2.

The unscaled deviance function is given by $\sigma^2\mathcal{D}(\underline{y};\hat{\mu}) = \sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2$ which is residual sum of squares for the proposed (current) model.

Example 3.2 (Logistic model)

Suppose $Y_i \sim Bin(n_i, \pi_i)$. Then

$$\log \prod_{i=1}^{n} g_Y(y_i; \pi_i) = \sum_{i=1}^{n} \{y_i \log \pi_i + (n_i - y_i) \log(1 - \pi_i)\}$$

and

$$\mathcal{D}(\underline{y}; \hat{\mu}) = 2 \log \frac{\prod_{i=1}^{n} g_Y(y_i; y_i)}{\prod_{i=1}^{n} g_Y(y_i; \hat{\pi}_i)} = 2 \sum_{i=1}^{n} \{y_i \log \frac{y_i}{n_i \hat{\pi}_i} + (n_i - y_i) \log \frac{n_i - y_i}{n_i - n_i \hat{\pi}_i}\}$$

If we perform a logistic regression, y_i is a $0-1$ outcome then for $n_i = 1$, $0 \log 0 = 1$ and $\varphi = 1$, we have

$$\mathcal{D}(\underline{y}; \hat{\mu}) = -2 \sum_{i=1}^{n} \{y_i \log \hat{\pi}_i + (1 - y_i) \log(1 - \hat{\pi}_i)\}$$

Chapter 4

Motivation to Model Selection Tests

4.1 Introduction

This chapter is a motivation to the next three chapters, especially to chapters 6 and 7. Model selection goes through estimating the performance of different models in order to choose the best one. On the other hand we know that the statistical models are typically merely approximations to reality and so most often are wrong; however they may be useful. First because a little of knowledge is better than nothing, second an assumed parametric model may be close to the true unknown model, so that very little is lost by the assumed model and we can use the rich literature of parametric statistics, and third in some applications the parameters for an assumed parametric model can often be interpreted usefully. Then selection and evaluation of a model is an important step. To do it we turn to the hypothesis testing for model selection or using some criteria.

Sometimes we consider only the parameter β as unknown, that is, we have assumed the shape of the distribution up to the value of an unknown parameter, which allows us to focus on inference for

this parameter but in many situations one may not have enough confidence that this is so. It means that sometimes we need to test the shape of the density for instance. Generally in this direction we require formulation of null and alternative hypothesis. An ideal hypothesis test is the test which gives the highest power among all possible tests at the fixed level of test (UMP tests). But this type of tests do not work for non-nested parametric families.

In other situations, a problem concerned with n i.i.d observations is to test whether the observations have a particular distribution, in other words we want to test whether a particular distribution fits our data. In some cases these tests are informal. Procedures of this kind are called goodness of fit tests.

There is a controversy about the connection between hypothesis testing and goodness of fit tests, because the alternative hypothesis is not very clear for the goodness of fit test. Then one wants to know how well this method will perform in a decision situation. That is, how do we assess the performance of the test? For answering this essential question we must be able to study the power of the proposed method against some alternatives. It seems that there is not an overall approach to define the alternative hypothesis in goodness of fit tests and it depends on the situation. In model selection we can consider each of the postulated models as the null model. It leads us to consider the null likelihood function or a function of it as our criterion to define a test. Thus we need to use a metric to verify the proximity of the postulated model to the true one. This chapter is essentially related to assumptions, properties of maximum likelihood estimators, maximized likelihood function, *AIC* and some metrics which are useful to model selection.

4.2 Assumptions

Suppose that the random variable Y is a measurable real valued function from a probability measure space $(\Omega, \mathcal{F}, P_0)$ into $(\mathcal{R}, \mathcal{B}_{\mathcal{R}})$ where for all $I \in \mathcal{B}_{\mathcal{R}}$ we have $\mu(I) = P_0(X \in I) = \int_I dP_0$ as the probability law of X, which admits a density $f \equiv \frac{dP_0}{d\nu}$ where ν is a σ-finite measure on $(\mathcal{R}, \mathcal{B}_{\mathcal{R}})$ and P_0 is absolutely continuous w.r.t. ν with regard to the Lebesgue's measure on $\mathcal{R}$. Suppose that f is unknown and $Y_1, Y_2, ..., Y_n$ be an i.i.d sample, independent of Y where $Y \sim f(.)$ and with the same distribution as Y. We know that the observations can not be infinite but we assume that n becomes large.

Consider P_β as a member of the family of all parametric probability measures on $(\Omega, \mathcal{F})$ which is absolutely continuous w.r.t. η a σ-finite measure on $(\mathcal{R}, \mathcal{B}_{\mathcal{R}})$ and admits the density $g(.;\beta) = \frac{dP_\beta}{d\eta}$ with regard to the Lebesgue's measure on $\mathcal{R}$ which are measurable in y for every $\beta \in B$ (compact) and continuous in β for every $y \in \mathcal{R}$. Then $P_\beta(Y \in I) = \int_I g(y;\beta)\, d\eta$. Always we can choice $\eta = \nu$ if for all β, P_β is absolutely continuous w.r.t. any measure $\eta \neq \nu$ it follows P_β and P_0 are absolutely continuous w.r.t. $\frac{1}{2}(\eta + \nu)$ then we can replace η and ν by $\frac{1}{2}(\eta + \nu)$. The notation $g(Y;\beta)$ asserts that $g(.,.) : \Omega \times B \to \mathcal{R}^+$ then $\log f : \Omega \times B \to \mathcal{R}$. Here after sometimes we show $g(.;\beta)$ by f_β.

If g belongs to a parametric family of densities this family (assumed or postulated family) could be considered as $\mathcal{G} = \{g(.,\beta) : \mathcal{R} \to \mathcal{R}^+; \beta \in B \subseteq \mathcal{R}^d\} = (g^\beta(.))_{\beta \in B}$. When we consider a parametric model we assume that the parameter uniquely determines the probability law related to a member of $\mathcal{G}$ i.e. if we know the parameter then we know the underlying probability law.

Related to $f(.)$ and $g(.,\beta)$ suppose that the following conditions are satisfied.

(C0) i) $f(y)$ is measurable in y, $g(y;\beta)$ is measurable in y for each $\beta \in B$, ii) $g(.,.)$ under $\mathcal{G}$ are distinct (β is identifiable), iii) and also $g(.,.)$ is continuous in β for each $y \in \mathcal{R}$.

(C1) B is compact.

(C2) i) $f(y)$ and $g(y;\beta)$, $\forall y \in \mathcal{R}$ are greater than zero and ii) $\mathcal{S}_\beta = \{y; g(y;\beta) > 0\}$ the common support of $g(.,\beta)$ does not depend on β .

(C3) i) $g(y,\beta)$ is twice continuously differentiable as a function of β and also ii) $\int_R \log g(y;\beta)\, dy$ is twice differentiable under the integral sign with respect to β for all $y \in \mathcal{S}_\beta$.

(C3)$'$ $g(.,\beta)$ is three times differentiable with respect to β and the third derivative is continuous with possibility of differentiating under the integral sign.

(C3)$''$ If $\beta_\star$ denotes the true value of β there exists a positive number $c(\beta_\star)$ and a function $M_{\beta_\star}(y)$ such that

$$|\frac{\partial^3}{\partial\beta^3}\log g(y;\beta)| \leq M_{\beta_\star}(y) \quad \forall y \in \mathcal{S}_\beta, |\beta - \beta_\star| < c(\beta_\star)$$

and

$$\mathcal{E}_{\beta_\star}\{M_{\beta_\star}(Y)\} < \infty.$$

(C4) There is a function ϑ which does not depend on β, such that $|\log g(y;\beta)| \leq \vartheta(y)$ $\forall \beta \in B$ and $\mathcal{E}_f(\vartheta(Y)) < \infty$.

If $f(.)$ can be zero then $\log f(.)$ can be $-\infty$. Then we consider the extended $\mathcal{R}$ i.e $\overline{\mathcal{R}}$ and assume that $\{\omega \in \Omega : X(\omega) = \infty\}$ and $\{\omega \in \Omega : Y(\omega) = -\infty\}$ both lie in $\mathcal{F}$ and the random variable $Y : \Omega \to \overline{\mathcal{R}}$ defined on $(\Omega, \mathcal{F}, .)$ is measurable and hence the log likelihood function is a measurable extended real valued function.

4.3 Likelihood Function and Maximum Likelihood Estimator

Fisher (1912, 1922) introduced the likelihood in the context of estimation via the method of maximum likelihood. The likelihood is a tool for dealing with the uncertainty due to the limited amount of information contained in the data. The purpose of the likelihood function is to convey information about unknown quantities (a parameter or unobserved random values or a mixed of both of them). The information is incomplete, and the likelihood function will express the degree of incompleteness. Officially the likelihood function is defined as below. When a parametric model is available, we ask what is the best estimate by data at hand. Here the uncertainty is in a way a nuisance.

Definition 4.1 *Assuming a statistical model parametrized by a fixed and unknown* β*, the likelihood* $L(\beta)$ *is the probability of the observed data z considered as a function of* β*.*

The data *z* could include any set of observations. Fisher (1922) noticed that it is the entire likelihood function that captures all the information contained in the data about a certain parameter not just its maximizer but in the context of point estimation we are looking for the maximum likelihood estimator. The likelihood function for an i.i.d. sample with size *n* and with density $f(.) = g(.,\beta), \beta \in B$ is defined as

$$\prod_{i=1}^{n} g(y_i;\beta)$$

and the log-likelihood function is

$$\sum_{i=1}^{n} \log g(y_i;\beta).$$

Our interest is in the weighted or normalized log-likelihood function which is defined as:

$$\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta).$$

When we write "big Y" the log-likelihood becomes a random function, and every things about it, including its derivatives, is also random. Each different value of $\beta \in B$ as a specified point in log-likelihood function and its derivatives gives a different random variable. Like every other random variable, they have probability distributions.

An important aspect of likelihood functions in asymptotic theory is finding a root for this kind of functions which is consistent for the true value of β. Under (C0)i,ii) and (C2)ii) we have:

$$P_{\beta_\star}\{\prod_{i=1}^{n}g(Y_i;\beta_\star) > \prod_{i=1}^{n}g(Y_i;\beta)\} \to 1 \quad \text{as} \quad n \to \infty$$

for any fixed $\beta \neq \beta_\star$ where $\beta_\star = \arg\max_{\beta\in B}\mathcal{E}_f\{\log g(Y;\beta)\}$ is the true value of parameter and we remember that we set $f(.) = g(.,.)$, Lehmann (1983). It is because

$$P_{\beta_\star}\{\prod_{i=1}^{n}g(Y_i;\beta_\star) > \prod_{i=1}^{n}g(Y_i;\beta)\} = P_{\beta_\star}\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_\star) > \frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta)\} =$$

$$P_{\beta_\star}\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta) - \frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_\star) < 0\}$$

By the law of large numbers, the left side tends in probability to

$$\mathcal{E}_{\beta_\star}\{\log\frac{g(Y_i;\beta)}{g(Y_i;\beta_\star)}\}$$

the log function is strictly convex then by Jensen's inequality it is less than

$$\log\mathcal{E}_{\beta_\star}\{\frac{g(Y_i;\beta)}{g(Y_i;\beta_\star)}\} < 0$$

thus

$$P_{\beta_\star}\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta) - \frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_\star) < 0\} \to 1 \quad \text{as} \quad n \to \infty$$

If we set $\beta = \beta_\star + \delta$ where δ is a positive arbitrary value near to 0, it follows that the likelihood function has a local maximum at $\beta_\star$. By this result, the density of random sample at true $\beta_\star$ exceeds that at any other fixed β with high probability when n is large. We do not know $\beta_\star$ but we can determine the value $\hat{\beta}_n$ of β which maximizes the density of the random sample. If this value exists and is unique, it is the Maximum Likelihood Estimator (MLE). Then ,

$$\hat{\beta}_n = \arg\max_{\beta \in B} \{\prod_{i=1}^{n} g(Y_i;\beta)\}.$$

The *MLE* has many large sample properties which make it popular and attractive for all researcher. It is asymptotically consistent, efficient and unbiased and the estimates themselves are normally distributed. Generally, a single number is not enough to represent a function. If the log likelihood is well approximated by a quadratic function, we need at least the location of its maximum and the curvature at the maximum. When the size of sample gets large these two quantities become more acceptable.

We usually find the maximum likelihood estimator as a solution of the score function

$$\nabla \log g(\overline{Y};\beta) = \nabla \sum_{i=1}^{n} \log g(Y_i;\beta) = 0,$$

where ∇^j is the j-th derivative of $g(.,\beta)$ with respect to β and $\overline{Y} = (Y_1, Y_2, ..., Y_n)$. If the solution is denoted by $\hat{\beta}_n$ we have

$$\nabla \sum_{i=1}^{n} \log g(Y_i;\hat{\beta}_n) = 0.$$

Note that this does not imply that $\nabla \sum_{i=1}^{n} \log g(Y_i;\beta_\star) = 0$. Just the opposite. In fact $\nabla \sum_{i=1}^{n} \log g(Y_i;\beta_\star)$ is a random variable and hence doesn't have a constant value. The second derivative of the log-likelihood is negative, so if define

$$I(\beta) = -\mathcal{E}_g\{\nabla^2 \sum_{i=1}^{n} \log g(Y_i;\beta)\} = -\mathcal{E}_g\{\frac{\partial^2 \log g(\overline{Y};\beta)}{\partial\beta\partial\beta'}\}$$

or in other form the information matrix is given by

$$J(\beta) = \mathcal{E}_g\{\frac{\partial \log g(\overline{Y};\beta)}{\partial \beta}\frac{\partial \log g(\overline{Y};\beta)}{\partial \beta'}\}$$

then a stronger consistency result for $\hat{\beta}_n$ is

$$\hat{\beta}_n \xrightarrow{a.s} \beta_\star.$$

For almost every sequence of sample where $\mathcal{E}_g\{\log g(Y;\beta_\star)\}$ exists (see White, 1982), the curvature at $\hat{\beta}_n$ is $I(\hat{\beta}_n)$. A large curvature $I(\hat{\beta}_n)$ is associated with strong peak which indicates less uncertainty about β. The quantity $I(\hat{\beta}_n)$is called the observed Fisher information. An important asymptotic property of normalized Likelihood function is that according to the weak law of large numbers, for each $\beta \in B$ we have,

$$\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta)\} \xrightarrow{\mathcal{P}} \mathcal{E}_g\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta)\} \qquad (1)$$

The compactness of B confident that the supremes on β exists and also $\sup_{\beta\in B}$ of a measurable function is measurable. This property for parameter space is discussed in White (1994). We note that the compactness of parameter space is not necessary for consistency of MLE but we need to this condition for (1). Anyway we set B as a compact set because for use of ULLN also we need this kind of parameter space.

4.3.1 Correctly Specified and Mis-Specified models

If the data generating density f was known, then we would know everything. The estimation, inference and especially the hypothesis testing arise because f is unknown. Then we postulate a model, $g(.;\beta) \in \mathcal{G}$, and the question which arises is whether the $Y_1, Y_2, ..., Y_n$ is an i.i.d. sample of $g(.;\beta) \in \mathcal{G}$. A fundamental assumption in classical hypothesis testing is that f belongs to a parametric family

of densities i.e. $f \in \mathcal{G}$. If so, there exists $\beta_\star$ which implies $f(.) = g(.;\beta_\star)$. In this case we say the model is correctly (or well) specified. On the other hand if $\nexists \beta \in B$ which implies $f(.) = g(.;\beta)$ we say the model is mis-specified. Fortunately as we will see in the next section in this case there exists a $\beta_0 \in B$ which minimizes the discrepancy between $f(.)$ and $g(.,.)$.

When the model is correctly specified the statistical inference and specially the asymptotic inference is straightforward; see Wald (1949) for strongly consistency of *MLE* and Cramer (1945) and Hajek (1970) for asymptotic variance. In the mis-specified case it is hard to decide whom to give credits for the asymptotic behavior of *MLE*. Huber (1967) proved consistency of *MLE* under some regularity conditions. Akaike (1973) recognized it but provided only heuristics. White (1982) provided an exact proof..

4.4 Metrics on spaces of probability

Metrization of probability measures i.e. defining a notation of distance is important, since in statistics one is often concerned about convergence of estimates based on finite samples to the true parameter which is often a probability measure and a definition of convergence is the notation of a distance.

Two usual metrics are: Total variation (TV) and Hellinger distance (HD). For the probability space $(\Omega, \mathcal{F})$ the TV distance between probability measures P and Q is defined as:

$$D^{TV}(P,Q) = \sup_{E\in\mathcal{F}} |P(E) - Q(E)| = \frac{1}{2}\int |f - q|\, d\mu$$

and the HD is defined as:

$$H^2(P,Q) = \frac{1}{2}\int [(\sqrt{f} - \sqrt{q}]^2\, d\mu$$

where μ is any measure that dominates both P and Q and f and q are the densities of P and Q respectively with respect to measure μ i.e. $f = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$.

This is easy to see that the $H^2(P,Q)$ is independent of μ and also f and q. To show it, let $\mu_1 = P_1 + Q_1$ and define:

$$f_1 = \frac{dP_1}{d\mu_1} \quad \text{and} \quad q_1 = \frac{dQ_1}{d\mu_1}.$$

It is clear that μ_1 dominates both P_1 and Q_1, so the derivatives exist. On the other hand μ dominate μ_1. Now

$$\int [\sqrt{f} - \sqrt{q}]^2 \, d\mu = \int \left[\sqrt{\frac{dP}{d\mu_1}\frac{d\mu_1}{d\mu}} - \sqrt{\frac{dQ}{d\mu_1}\frac{d\mu_1}{d\mu}} \right]^2 d\mu =$$

$$\int \left[\sqrt{\frac{dP}{d\mu_1}} - \sqrt{\frac{dQ}{d\mu_1}} \right]^2 d\mu_1$$

This shows that the invariance of the Hellinger distance to the choice of the dominating measure μ. All of these measures are known as D-divergences or Ali-Silvey distances which defined as bellow.

Definition 4.2 *Definition Given any continuous convex function $D : [0, +\infty] \to \mathcal{R} \cup \{\infty\}$, the D-divergence between f and q is given by $I_D(f,q) = \int_z q(z) D(\frac{f(z)}{q(z)})$. The TV, HD and KL divergence are given by choosing $D(u) = \frac{1}{2}|u-1|$, $D(u) = \frac{1}{2}(\sqrt{u}-1)^2$ and $D(u) = u\log(u)$ respectively.*

It is clear that the Kullback-Leibler discrepancy is a convex function. This convexity could be easily verify for the nested models.

Example 4.1 (Convexity of KL for Bernoulli distribution)

Consider the Bernoulli model $Bin(1,\pi)$ a member of this family is $Bin(1,\pi_0)$ where π_0 is known. Then we have $g(Y;\pi) = \pi^Y(1-\pi)^{1-Y}, \quad Y = 0,1$. The respective measure for this model is given by $KL\{g(.;\pi);g(.;\pi_0)\} = \pi_0 \log\frac{\pi_0}{\pi} + (1-\pi_0)\log\frac{1-\pi_0}{1-\pi}$ if we set $\pi_0 = 0.1, 0.2, ...0.9$

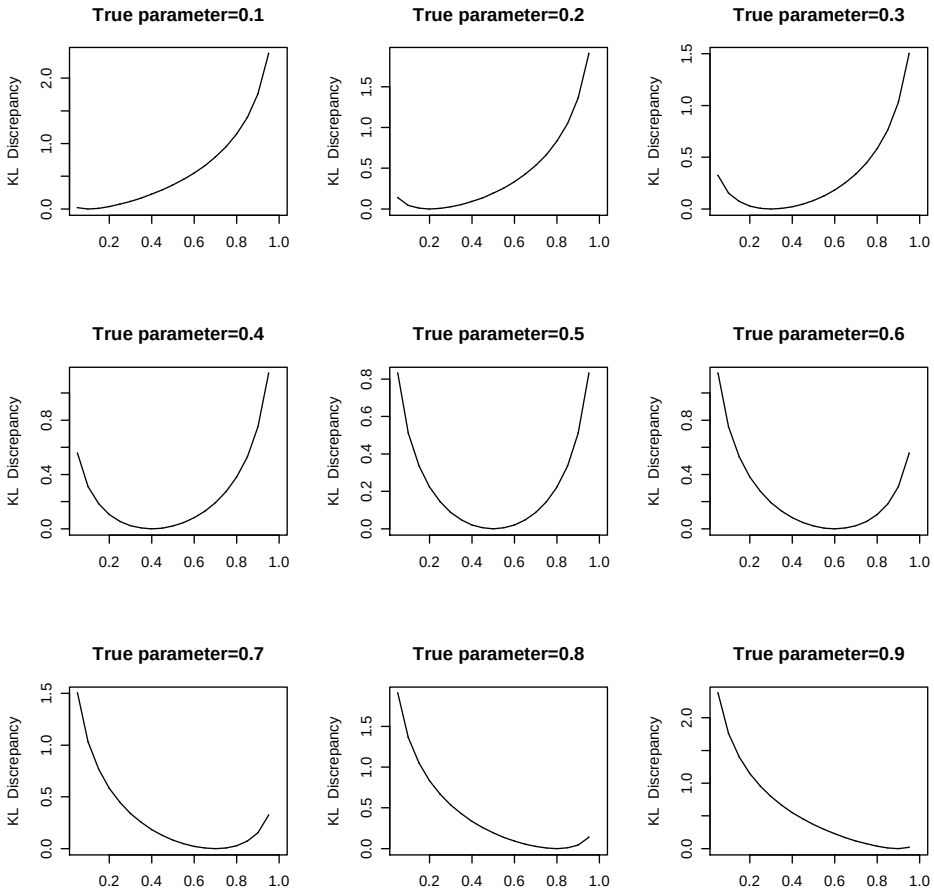


Figure 8: An example which shows that the convexity of Kulback-Leibler criterion for Bernoulli family and its minimum which happens at π_0.

Example 4.2 (Convexity of KL for Normal distribution)

The KL discrepancy for the Normal model as $\{\mathcal{N}(\mu,1), \quad \mu \in \mathcal{R}\}$ is given by $\frac{(\mu-\mu_0)^2}{2}$. For $\mu_0 = 1$ the KL discrepancy shows in Figure 8.

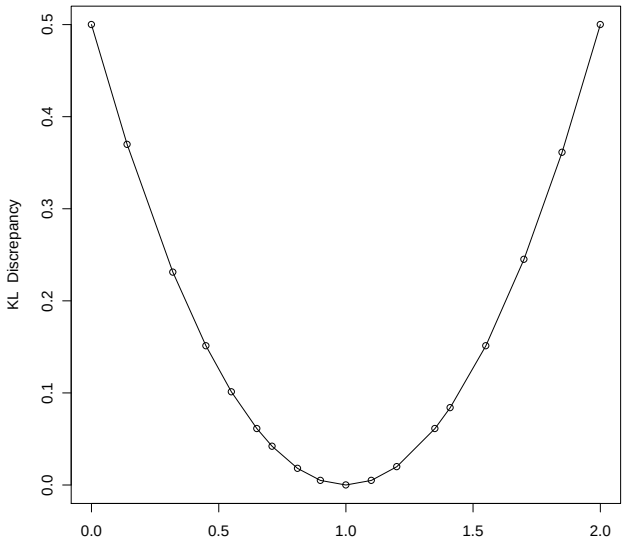


Figure 9: Kullback-Leibler discrepancy for Normal density.

These two examples suggest that in realistic situations when we do not have any knowledge about

the true density which generates the data we should search for a minimization procedure for the KL discrepancy. This idea agrees with the distance concept of KL criterion.

4.4.1 Kullback-Leibler Discrepancy (divergence)

An important discrepancy measure in statistics is known as the Kullback-Leibler discrepancy (Kullback-Leibler distance, although the term "distance" needs to be interpreted because this criterion does not satisfy some properties of a usual distances). By notation of section 4.4 the Kullback-Leibler discrepancy is defined as:

$$KL(Q;P) = \int f \log \frac{f}{q}\, d\mu$$

There is a relation between three distances as follows

$$[D^{TV}(P,Q)]^2 \leq 2H^2(P,Q) \leq KL(Q;P)$$

It follows that if the KL discrepancy between a sequence of probabilities $\{P_n\}$ and a fixed probability P goes to zero, then this convergence should happen for Hellinger and total variation sense.

We set a sample of i.i.d. random variables as $Y_1, Y_2, ..., Y_n$ having pdf $f = f(.)$ and a parametric model: $\mathcal{G} = (g(.,\beta)) = (g^{\beta}(.))_{\beta \in B}$ The Kullback-Leibler discrepancy (KL criterion) for the data generating density f and $g(.;\beta) \in \mathcal{G}$ is defined as

$$KL\{g(.;\beta); f(.)\} = \int \log \frac{f(y)}{g(y;\beta)} f(y)\, dy$$

which is a non-negative quantity. By definition, the more $g(.;\beta)$ agrees with $f(.)$ the smaller $KL\{g(.;\beta); f(.)\}$ is. Then the closest member in $\mathcal{G}$ to the f is $g(.,\beta_0)$ where $\beta_0 \in B$ is the minimizer of $KL\{g(.;\beta); f(.)\}$ as defined in 4.3. Under this divergence, $g(.;\beta_0)$ is the best approximation to f under model $\mathcal{G}$. It is important we notice that when the model is correctly specified we have:

$$\beta_0 = \beta_\star.$$

Note that according to the weak law of large numbers, for each $\beta \in B$ we have,

$$\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta)\} \xrightarrow{\mathcal{P}} \mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta)\} \qquad (1)$$

A natural estimator of β_0, Knight (1999), is the minimizer $\hat{\beta}_n$ of the

$$KL_n\{g(.;\beta);f(.)\} = \frac{1}{n}\sum_{i=1}^{n}\log\frac{f(Y_i)}{g(Y_i;\beta)}$$

since it can be written as

$$\frac{1}{n}\sum_{i=1}^{n}\log f(Y_i) - \frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta)$$

$\hat{\beta}_n$ minimizes the second term. We note that this term is the negative normalized log-likelihood function, then the minimizer of $KL_n\{g(.;\beta);f(.)\}$ is simply the maximum likelihood estimator. Hence we can compare $f(.)$ with $g(.;\hat{\beta}_n)$ an estimate of $g(.;\beta_0)$ the best approximation to f under $\mathcal{G}$.

Now consider a random variable whose distribution comes from $\mathcal{G}$ (the correctly specified case). Let $\beta_\star$ denote the data generating parameter and $Y_1, Y_2, ..., Y_n$ i.i.d from the underlying distribution. The Kullback-Leibler divergence is defined (under (C0)ii) as

$$KL\{g(.;\beta);g(.,\beta_\star)\} = \mathcal{E}_{\beta_\star}\{\log\frac{g(Y;\beta_\star)}{g(Y;\beta)}\}.$$

As we see, the KL measure the distance of the model from the true density and is not observable because $g(Y;\beta_\star)$ is unknown. Then an essential question in this case is that how can we use it? This number is nonnegative because

$$\mathcal{E}_{\beta_\star}\{\log\frac{g(Y;\beta_\star)}{g(Y;\beta)}\} = \mathcal{E}_{\beta_\star}\{-\log\frac{g(Y;\beta)}{g(Y;\beta_\star)}\} \geq -\log\mathcal{E}_{\beta_\star}\{\frac{g(Y;\beta)}{g(Y;\beta_\star)}\} > 0$$

with equality if and only if $\beta = \beta_\star$.

Now the Kullback-Leibler divergence is connected with maximum likelihood estimation as below.

$$\frac{1}{n}\sum_{i=1}^{n}\log\frac{g(Y_i;\beta_\star)}{g(Y_i;\hat{\beta}_n)} = \frac{1}{n}\sum_{i=1}^{n}\log\frac{g(Y_i;\beta_\star)}{g(Y_i;\hat{\beta}_n)} - KL\{g(.;\hat{\beta}_n);g(.;\beta_\star)\} + KL\{g(.;\hat{\beta}_n);g(.;\beta_\star)\} \leq 0$$

Then

$$0 \leq KL\{g(.;\hat{\beta}_n);g(.;\beta_\star)\} \leq |\frac{1}{n}\sum_{i=1}^{n}\log\frac{g(Y_i;\beta_\star)}{g(Y_i;\hat{\beta}_n)} - KL\{g(.;\hat{\beta}_n);g(.;\beta_\star)\}| \leq$$

$$|\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_\star) - \mathcal{E}_{\beta_\star}\{\log g(Y;\beta_\star)\}| + |\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n) - \mathcal{E}_{\beta_\star}\{\log g(Y;\hat{\beta}_n)\}| \leq$$

$$|\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_\star) - \mathcal{E}_{\beta_\star}\{\log g(Y;\beta_\star)\}| + \sup_{\beta\in B}|\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta) - \mathcal{E}_{\beta_\star}\{\log g(Y;\beta)\}|$$

the first term by (1) and the second term by (1) and theorem $B3$ converge to zero (a.e.) in probability.

Then

$$KL\{g(Y;\hat{\beta}_n);g(Y;\beta_\star)\} \to 0 \quad \text{as} \quad n \to \infty$$

and a.e.

$$\frac{g(Y;\hat{\beta}_n)}{g(Y;\beta_\star)} \to 1$$

which implies that

$$\hat{\beta}_n \xrightarrow{\mathcal{P}} \beta_\star.$$

On the other hand assume that $f \notin \mathcal{G}$ and $\beta_0 = \beta_0(f)$ denote the minimizer of the $KL\{g(.;\beta); f(.)\}$.

Now

$$\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta) \xrightarrow{a.s.} \mathcal{E}_f\{\log g(Y;\beta)\} = \mathcal{E}_f\{\log f(Y)\} - \mathcal{E}_f\{\log\frac{f(Y)}{g(Y;\beta)}\}$$

then under (C0) i,ii) and (C4)

$$argmax_\beta \frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta) \xrightarrow{\mathcal{P}} argmax_\beta\{\mathcal{E}_f\{\log f(Y)\} - \mathcal{E}_f\{\log\frac{f(Y)}{g(Y;\beta)}\}\}$$

which means that

$$\hat{\beta}_n \xrightarrow{\mathcal{P}} \beta_0.$$

In the next section we will talk about consistency of maximum likelihood estimator.

4.5 Consistency of Maximum Likelihood Estimator

The mean theorem about the consistency of maximum likelihood estimator is given in theorem 2.2 of White (1982) which says that under (C0),(C1) and (C2) the maximum likelihood estimator $\hat{\beta}_n$ almost surely converges to β_0 the unique minimizer of KL divergence between data generating density and postulated model. As we saw if the model is correctly specified then the $\hat{\beta}_n$ is consistent for true parameter $\beta_\star$ which is unique. This later result is the classical consistency of *MLE*. Biernacki (2004), proved that under (C1), (C2), (C3) and this hypothesis that the

$$\max_{j=1,2} \mathcal{E}_g \sup_{\beta\in B} |\varphi_j(Y;\beta)| < \infty, \quad \mathcal{E}_g \sup_{\beta\in B} |\nabla\varphi_j(Y;\beta)| < \infty \quad j = 1,2$$

where $\varphi_1(Y;\beta) = \nabla \log g(Y;\beta)$, and $\varphi_2(Y;\beta)$ is the function $W(Y;\beta)$ defined in theorem B3 for $\varphi(.) = \log(.)$, $\hat{\beta}_n$ is consistent for β_0 iff $\frac{1}{n}\sum_{i=1}^n W(Y_i,\hat{\beta}_n) \xrightarrow{\mathcal{P}} 0$. The consistency of MLE, Wald (1949), is strongly related to choice of the parameter space B. In general we say that the parameter space B (an open interval) contains an open interval L of which the true parameter value $\beta_\star$ is an interior point. May be we set B as a finite set. The compactness (closed and bounded) is nearest property to finiteness. It is often said that compactness is the next best thing to finiteness, because the more modern definition of compact space says that a space is compact if each of its open covers has a finite sub covers.

Indeed under (C0)i,ii) and (C2) and finiteness of B, $\hat{\beta}_n$ exists, it is unique and consistent w.p.1. Under (C0), (C2)ii), (C3) and the condition which consider B as a open interval contain an open interval L of which the true parameter value $\beta_\star$ is an interior point, Lehmann (1999), we conclude

that $\nabla \sum_{i=1}^{n} \log g(Y_i;\beta) = 0$ has a unique root as $\hat{\beta}_n$ which is consistent for $\beta_\star$, that is

$$\hat{\beta}_n \xrightarrow{\mathcal{P}} \beta_\star.$$

Which means that $\hat{\beta}_n$ is a consistent estimator for $\beta_\star$. As a counterexample of the inconsistent *MLE* consider Ferguson's example, see , Lehmann (1998).

Example 4.3

Suppose that

$$g(y;\beta) = \beta \frac{1}{2} 1_{[-1,1](y)} + \frac{1-\beta}{\delta(\beta)} \left(1 - \frac{|y-\beta|}{\delta(\beta)}\right) 1_{c(\beta)}(y)$$

where $\beta \in B = [-1,1]$ $\delta(.)$ is decreasing and continuous with $\delta(0) = 1$, $0 < \delta(\beta) \leq 1-\beta$ for $0 < \beta < 1$, and $c(\beta) \equiv (\beta - \delta(\beta), \beta + \delta(\beta))$. Note that $g(y;\beta)$ is continuous in β for all y, and $g(y;0) = (1-|y|)1_{[-1,1]}(y)$ is the triangular density, while $g(y,1) = \frac{1}{2}1_{[-1,1]}(y)$ is the uniform density.Since a continuous function on compact set $[0,1]$ achieves its maximum on the set, and regularity conditions is satisfied for this example thus a *MLE* exists. Now if $\delta(\beta) \to 0$ rapidly enough as $\beta \to 1$ then $\hat{\beta}_n \xrightarrow{a.s.} 1$ for every $\beta \in [0,1]$ no matter what the true value of β.

4.6 Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) initially was proposed as an estimate of minus twice the expected log-likelihood. We notice that the important part of the KL divergence is $\mathcal{E}_f\{\log g(Y;\beta)\}$ which has an estimator as

$$\frac{1}{n}\sum_{i=1}^{n} \log g(Y_i;\hat{\beta}_n).$$

It can be considered as an estimator of the distance between the true density and the model. Now the stress is on $\hat{\beta}_n$ because $\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)$ provides an overestimate and then the maximized likelihood function has a positive bias as an estimator of the expected log-likelihood. Since $\hat{\beta}_n$ corresponds to the empirical distribution, say, F_n which introduces the estimator. In fact both of them depend on the same sample.

Unfortunately when $f \notin \mathcal{G}$

$$\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n) \nrightarrow \mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)\}$$

and introduce the bias, according to Konishi and Kitagawa (1996) and Bozdogan(2000) we have,

$$bias = \mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n) - \int_{\mathcal{R}}\log g(y;\hat{\beta}_n)\,f(y)\,dy\} = \frac{1}{n}tr(I^{-1}J) + O(n^{-2})$$

where as before I is the inverse Fisher information matrix in inner product (Hessian) form, and J is the outer product form of the Fisher information matrix for vector β

$$I = -\mathcal{E}_f\{\frac{\partial^2 \log g(Y;\beta)}{\partial\beta\partial\beta'}\}$$

and

$$J = \mathcal{E}_f\{\frac{\partial \log g(Y;\beta)}{\partial\beta}\frac{\partial \log g(Y;\beta)}{\partial\beta'}\}.$$

An estimate for these two information matrices on the base of any estimator $\bar{\beta}_n$ and empirical distribution function is given by

$$\hat{I} = \left\{-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log g(Y_i;\beta)}{\partial\beta_r\partial\beta_s}|\beta=\hat{\beta},\quad r,s=1,2,...,p\right\}$$

and

$$\hat{J} = \left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log g(Y_i;\beta)}{\partial\beta_r}\frac{\partial \log g(Y_i;\beta)}{\partial\beta_s}|\beta=\hat{\beta},\quad r,s=1,2,...,p\right\}.$$

If $f \in \mathcal{G}$, $tr(I^{-1}J) = tr(I_p) = p$, where $p = dim(B)$. (The Information Matrix Equivalence Test, White (1982)). Asymptotically we have:

$$bias = \frac{p}{n} + O(n^{-2})$$

which gives $\hat{bias} = \frac{p}{n}$. Now the criterion based on the bias corrected normalized log-likelihood is given by

$$\frac{1}{n}\sum_{i=1}^{n} \log g(Y_i; \hat{\beta}_n) - \hat{bias}$$

Akaike (1973) introduced a criterion as

$$AIC = -2n\{\frac{1}{n}\sum_{i=1}^{n} \log g(Y_i; \hat{\beta}_n) - \hat{bias}\} = -2\sum_{i=1}^{n} \log g(Y_i; \hat{\beta}_n) + 2p,$$

or

$$\frac{1}{2n}AIC = -\frac{1}{n}\sum_{i=1}^{n} \log g(Y_i; \hat{\beta}_n) + \frac{p}{n}.$$

When there are several competing models, the values of AIC's are computed. The model with minimum AIC value is chosen as the best model to fit the data. When n gets large the fixed penalty term $2p$ does not change and we expect that $\frac{p}{n} \to 0$. However for finite n *AIC* is a way of expressing the parsimony principle.

4.7 Distribution of Maximum Likelihood Estimator

Here we review the convergence in distribution of maximum likelihood estimators and allow β to be a vector. In fact under some regularity conditions

$$\sqrt{n}(\hat{\beta}_n - \beta_\star) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I^{-1}(\beta_\star))$$

here $I_1(\beta_\star)$ is the Fisher information from a single observation with true density $f(.) = g(.;\beta_\star)$, and the regularity conditions are $(C0)-(C3)$,$(C3)'$and $(C3)''$.

If $f \notin \mathcal{G}$ under same regularity conditions as above

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I^{-1}(\beta_0)J(\beta_0)I^{-1}(\beta_0))$$

where I and J evaluated at β_0 are defined as before. The $I^{-1}(\beta_0)J(\beta_0)I^{-1}(\beta_0)$ is a robust variance, since it is correct regardless whether f the true density is correctly specified or not. A proof of this asymptotic distribution is as follow.

Based on $Y_1, Y_2, ..., Y_n$ an i.i.d. sample we have the likelihood function as $\log\prod_{i=1}^n g(Y_i;\beta) = \sum_{i=1}^n \log g(Y_i;\beta)$, expanding the normalized derivative of this likelihood function about β_0, It follows that

$$\frac{1}{n}\frac{\partial\sum_{i=1}^n \log g(Y_i;\beta)}{\partial\beta} = \frac{1}{n}\frac{\partial\sum_{i=1}^n \log g(Y_i;\beta)}{\partial\beta}|_{\beta=\beta_0} + \frac{1}{n}\frac{\partial^2\sum_{i=1}^n \log g(Y_i;\beta^\dagger)}{\partial\beta\partial\beta'}(\beta-\beta_0)$$

at $\hat{\beta}_n$ we have

$$\frac{1}{n}\frac{\partial\sum_{i=1}^n \log g(Y_i;\hat{\beta}_n)}{\partial\beta} = \frac{1}{n}\frac{\partial\sum_{i=1}^n \log g(Y_i;\beta)}{\partial\beta}|_{\beta=\beta_0} + \frac{1}{n}\frac{\partial^2\sum_{i=1}^n \log g(Y_i;\beta^\dagger)}{\partial\beta\partial\beta'}(\hat{\beta}_n-\beta_0)$$

or

$$0 = \frac{1}{n}\frac{\partial\sum_{i=1}^n \log g(Y_i;\beta)}{\partial\beta}|_{\beta=\beta_0} + \frac{1}{n}\frac{\partial^2\sum_{i=1}^n \log g(Y_i;\beta^\dagger)}{\partial\beta\partial\beta'}(\hat{\beta}_n-\beta_0)$$

where $|\beta^\dagger-\beta| \leq |\beta-\hat{\beta}_n|$. It is clear that

$$\mathcal{E}_f\{\frac{\partial\log g(Y;\beta)}{\partial\beta}\}|_{\beta=\beta_0} = 0$$

and

$$\mathcal{V}ar_f\{\frac{\partial\log g(Y;\beta)}{\partial\beta}\}|_{\beta=\beta_0} = J(\beta_0) = \mathcal{E}_f\{\frac{\partial\log g(\overline{Y};\beta)}{\partial\beta}\frac{\partial\log g(\overline{Y};\beta)}{\partial\beta'}\}|_{\beta=\beta_0}.$$

Now by the Central Limit Theorem, at $\beta=\beta_0$ we have

$$n^{-1/2}\sum_{i=1}^n \frac{\partial\log g(Y_i;\beta)}{\partial\beta} \xrightarrow{L} \mathcal{N}(0, J(\beta_0)).$$

On the other hand

$$\frac{1}{n}\frac{\partial^2 \sum_{i=1}^n \log g(Y_i;\beta^\dagger)}{\partial\beta\partial\beta'} = \frac{1}{n}\sum_{i=1}^n \frac{\partial^2 \log g(Y_i;\beta^\dagger)}{\partial\beta\partial\beta'} \xrightarrow{\mathcal{P}} \mathcal{E}_f\{\frac{\partial^2 \log g(Y_i;\beta^\dagger)}{\partial\beta\partial\beta'}\}|_{\beta=\beta_0} = -I(\beta_0).$$

Thus from the Taylor expansion we have

$$n^{-1/2}\frac{\partial \sum_{i=1}^n \log g(Y_i;\beta_0)}{\partial\theta} = -\frac{1}{n}\frac{\partial^2 \log g(Y_i;\beta^\dagger)}{\partial\beta\partial\beta'}\sqrt{n}(\hat{\beta}_n - \beta_0)$$

and by Slutsky's theorem

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{L} \mathcal{N}(0, I^{-1}(\beta_0)J(\beta_0)I^{-1}(\beta_0))$$

if I is invertible. Now if $f \in \mathcal{G}$, $\beta_0 = \beta_\star$ and $I(\beta_0) = J(\beta_0)$

Chapter 5

Proposed test for Goodness of Fit Test:

A test based on empirical likelihood ratio

5.1 Introduction

The method which we want to discuss in this chapter may be viewed as an application of a goodness of fit measure extended to the likelihood ratio test. When we are in goodness of fit test situation we

have a null hypothesis which is completely clear and an alternative which is completely vague as

$$\mathcal{H}_0 : F(y) = F_0(y) \quad \forall y \in \mathcal{Y}$$

against

$$\mathcal{H}_1 : F(y) \neq F_0(y) \quad \text{for some} \quad y \in \mathcal{Y}$$

where $F_0(.)$ is a known distribution function. Chapter 3 has introduced some goodness of fit test to this problem. Here we want to introduces an approach which introduce a test statistic using known likelihood ratio test. This idea is not in favor or against the known goodness of fit tests, but is an approach which helps us to solve a problem with a different method which works for binned and unbind data. The likelihood ratio approach has an extensive theory which is a guaranty for this method of test. In fact this idea is based on the Berk-Jones statistics (1979), see, 3.4.1. More precisely the Berk-Jones statistics could be defined as a supreme of

$$K(F_n(y), F_0(y))) = \begin{cases} F_n(y)\left(\log(\frac{F_n(y)}{F_0(y)})\right) + (1 - F_n(y)) \log \frac{1-F_n(y)}{1-F_0(y)} & \text{if } 0 < F_0(y) < F_n(y) < 1 \\ 0 & \text{if } 0 \leq F_n(y) \leq F_0(y) \leq 1 \\ \infty & \text{otherwise.} \end{cases}$$

This is the Kullback-Leibler discrepancy for two Bernoulli distributions. It is known that $K(F_n(y), F_0(y))$ behaves as $\frac{1}{2} \frac{(F_n(y)-F_0(y))^2}{F_0(y)(1-F_n(y))}$. This last term is half of the Pearson statistics for $F_n(y)$ which is distributed as $Bin(1, F_0(y))$ for a fixed y. The theorem 9.1 Knight (1999) shows that when we consider the goodness of fit test for multinomial distribution, the Pearson χ^2 statistic is asymptotically equivalent to the likelihood ratio statistic. Berk-Jones proposed that we can fix y and construct a test statistic by likelihood ratio test for goodness of fit test problem as above.

5.2 Our objective

In this chapter we are in search of a goodness of fit test for the simple situation where $F_0(.)$ is a known distribution function, if not, as a common approach for goodness of fit test we have to estimate the unknown parameter(s) at first and then apply the test. Parametric case will change our situation for model selection from testing for a specified distribution belong to the model to the more general situation which is testing for a family of distributions (model). Our idea is as follows, consider a random sample as $(Y_1, Y_2, ..., Y_n)$ and a goodness of fit test procedure which introduces a likelihood ratio test for each fixed z which could be between any of two Y's. Here we must emphasis that $F(y)$ is an unknown distribution function, whereas $F(z)$ with fixed z is an unknown parameter. If we separate the null hypothesis $\mathcal{H}_0 : F(y) = F_0(y) \quad \forall y \in \mathcal{Y}$ to several null hypotheses as $\mathcal{H}_{0z} : F(z) = F_0(z) \quad \forall z \in \mathcal{Z}$ we can construct a likelihood ratio for each one of the $\mathcal{H}'_{0z}s$ for each fixed z, and then construct a test for our essential hypothesis testing problem. Fortunately this concept is known in statistics. The Union-Intersection test (UIT) is our proposal to solve this problem. As a test statistic we generalized the logic of the likelihood ratio test. In fact we defined a weight function as $w(z)$. This weight function permit us to construct different tests. As a simple choice we consider $w(z) = C_n F_n(z)$ or more complex choice as a complex function of $F_n(z)$. For $C_n = n^{-1}$ we verify the level and efficiency of our test. If we set $C_n = n^{-2}$ the type I error of our test is less than type I error of Berk-Jones's test. After a brief review of UIT, we will construct the likelihood ratio test statistic by UIT. The level of test and efficiency for this test will be verified. It seems that our statistic is comparable by Berk-Jones statistic.

5.3 Union-Intersection Test

The likelihood ratio test (LRT) method is a commonly used method of hypothesis test construction. Another method, which is appropriate when the null hypothesis is expressed as an intersection, is the union-intersection test (UIT). In classical statistics we may write

$$\mathcal{H}_0 : \theta \in \bigcap_{\gamma \in \Gamma} \Theta_\gamma$$

where Γ is an arbitrary index set that may be finite or infinite, depending on the problem. By this notation we have

$$\mathcal{H}_1 : \theta \in \bigcup_{\gamma \in \Gamma} \Theta_\gamma^c$$

Suppose that for each of the testing $\mathcal{H}_{0\gamma} : \theta \in \Theta_\gamma$ against the alternative hypothesis $\mathcal{H}_{1\gamma} : \theta \in \Theta_\gamma^c$. We know that the rejection region for the test of $\mathcal{H}_{0\gamma}$ is $\{y : T_\gamma(y) \in R_\gamma\}$ where $T_\gamma(.)$ is the test statistic. Thus if any of the $\mathcal{H}_{0\gamma}$ is rejected, then $\mathcal{H}_0$ must also be rejected, it offers a rejection region for UIT as

$$\bigcup_{\gamma \in \Gamma} \{y : T_\gamma(y) \in R_\gamma\}$$

As a simple example for UIT we consider a known hypothesis test in elementary statistics.

Example 5.1

Let $Y_1, Y_2, ..., Y_n$ be a i.i.d. random sample from $\mathcal{N}(\mu, \sigma^2)$, where μ and σ^2 are unknown parameters. We want to test that $\mathcal{H}_0 : \mu = \mu_0$ against $\mathcal{H}_1 : \mu \neq \mu_0$, where μ_0 is a specified number. As a UIT we can write

$$\mathcal{H}_0 : \{\mu : \mu \leq \mu_0\} \cap \{\mu : \mu \geq \mu_0\}$$

This null hypothesis could be write as intersection of two new null hypotheses as $\mathcal{H}_{0Lower} : \{\mu : \mu \leq \mu_0\}$ and $\mathcal{H}_{0Upper} : \{\mu : \mu \geq \mu_0\}$. Now as the classical approach we will test

$$\mathcal{H}_{0Lower} : \mu \leq \mu_0 \quad \text{against} \quad \mathcal{H}_{1Lower} : \mu > \mu_0$$

with rejection region $\frac{1/n\sum_{i=1}^{n} Y_i - \mu_0}{S/\sqrt{n}} \geq t_{Lower}$ and

$$\mathcal{H}_{0Upper} : \mu \geq \mu_0 \quad \text{against} \quad \mathcal{H}_{1Upper} : \mu < \mu_0$$

with rejection region $\frac{1/n\sum_{i=1}^{n} Y_i - \mu_0}{S/\sqrt{n}} \leq t_{Upper}$. Then the rejection region of the UIT of

$$\mathcal{H}_0 : \{\mu : \mu \leq \mu_0\} \cap \{\mu : \mu \geq \mu_0\}$$

against

$$\mathcal{H}_1 : \{\mu : \mu \geq \mu_0\} \cup \{\mu : \mu \leq \mu_0\}$$

for $t_{Lower} = -t_{Upper}$ will be express as $|\frac{1/n\sum_{i=1}^{n} Y_i - \mu_0}{S/\sqrt{n}}| \geq t_{Lower}$ which is the two sided test.

5.4 Proposed test based on empirical likelihood ratio

Consider $\overline{Y} = (Y_1, Y_2, ..., Y_n)$ as an i.i.d. random sample with unknown distribution function $F(.)$. We set $F_0(.)$ as a known distribution function. The official goodness of fit test is contain testing

$$\mathcal{H}_0 : F(y) = F_0(y) \quad \forall y \in \mathcal{Y}$$

against

$$\mathcal{H}_1 : F(y) \neq F_0(y) \quad \text{for some} \quad y \in \mathcal{Y}$$

A key for proposing a goodness of fit test is that the distribution function $F(z)$ for a fixed z is an unknown parameter. It reduces the goodness of fit test to a LRT test as

$$\mathcal{H}_{0z} : F(z) = F_0(z) \quad \forall z \in \mathcal{Z}$$

against

$$\mathcal{H}_{1z} : F(z) \neq F_0(z) \quad \text{for some} \quad z \in \mathcal{Z}.$$

Our idea is to rewrite this hypothesis testing as the UIT, thus we have

$$\mathcal{H}_0 : \bigcap_{z \in \mathcal{Z}} H_{0z}$$

against

$$\mathcal{H}_1 : \bigcup_{z \in \mathcal{Z}} H_{1z}$$

For each z we can define a new random variable, see, Berk and Jones (1978), thus we have

$$Y_{iz} = 1\{Y_i \leq z\} = \begin{cases} 1 & \text{if } Y_i \leq z \\ 0 & \text{if } Y_i > z \end{cases}$$

for $i = 1, 2, ..., n$

Now we have a parametric test with a binary variable with value in $\{0,1\}^n$, i.e.

$$Y_{iz} \sim Bin(1, F(z))$$

and

$$\sum_{i=1}^{n} Y_{iz} = nF_n(z) \sim Bin(n, F(z)).$$

The likelihood function is given by

$$\mathcal{L}(F(z)) = \mathcal{L}(F(z); \overline{Y}_{iz}) = (F(z))^{nF_n(z)}(1 - F(z))^{n(1-F_n(z))}$$

The likelihood ratio test is given by

$$\lambda_n(z) = \frac{\sup_{F(z)} \mathcal{L}_n(F(z))}{\mathcal{L}_n(F_0(z))}$$

$$\lambda_n(z) = \mathcal{L}^{F_n(z)/F_0(z)} \equiv \frac{\mathcal{L}_n(F_n(z))}{\mathcal{L}_n(F_0(z))}$$

for the large value of $\lambda_n(z)$ we reject the null hypothesis. The log likelihood function is given by

$$\log \lambda_n(z) = \log \mathcal{L}^{F_n(z)/F_0(z)} = nF_n(z)\log(\frac{F_n(z)}{F_0(z)}) + n(1-F_n(z))\log(\frac{1-F_n(z)}{1-F_0(z)}).$$

The propose test statistics for testing $\mathcal{H}_0$ against $\mathcal{H}_1$ is

$$T_n = \int_{\mathcal{R}} \log \mathcal{L}^{F_n(z)/F_0(z)} d(w(z))$$

The reasonable choose of $w(z)$ will give us a reasonable test statistic. A choice could be $w(z) = C_n F_n(z)$, and a simple one is given by $C_n = n^{-1}$ which defines our statistics as

$$T_n = \int_{\mathcal{R}} [nF_n(z)\log(\frac{F_n(z)}{F_0(z)}) + n(1-F_n(z))\log(\frac{1-F_n(z)}{1-F_0(z)})] d(F_n(z)) =$$

$$\frac{1}{n}\sum_{y_i \in \mathcal{A}_i} \log \mathcal{L}^{F_n(y_i)/F_0(y_i)}$$

where $\mathcal{A}_i = [Y_i, Y_{i+1}]$ or $\mathcal{A}_i = [Y_{(i)}, Y_{(i+1)}]$. By this we have

$$T_n = \frac{1}{n}\sum_{i=1}^{n} \{nF_n(Y_i)\log(\frac{F_n(Y_i)}{F_0(Y_i)}) + n(1-F_n(Y_i))\log(\frac{1-F_n(Y_i)}{1-F_0(Y_i)})\}$$

or

$$T_n = \frac{1}{n}\sum_{i=1}^{n} \{nF_n(Y_{(i)})\log(\frac{F_n(Y_{(i)})}{F_0(Y_{(i)})}) + n(1-F_n(Y_{(i)}))\log(\frac{1-F_n(Y_{(i)})}{1-F_0(Y_{(i)})})\}$$

where $Y_{(i)}$ is the ith ordered statistics and also they are the discontinuity points of $F_n(.)$ when $z_i \in \mathcal{A}_i$.

It is common used in statistics which consider

$$F_n(Y_{(i)}) = \frac{i-\varepsilon}{n-2\varepsilon+1} \qquad \varepsilon \in [0,1].$$

5.4.1 Level of test

Two important aspects of any test is the level and the power of test. The rejection region for the UIT

is given by

$$\bigcup_{z\in\mathcal{Z}}\{\log\mathcal{L}^{F_n(z)/F_0(z)}\in R_z\}$$

which defines the level of UIT as

$$\alpha_{UIT}=P_{\mathcal{H}_0}\left(\bigcup_{z\in\mathcal{Z}}\{\log\mathcal{L}^{F_n(z)/F_0(z)}\geq c\right)=P_{\mathcal{H}_0}\left(\sup_{z\in\mathcal{Z}}\log\mathcal{L}^{F_n(z)/F_0(z)}\geq c\right).$$

Now we have

$$T_n\leq\frac{1}{n}\sum_{i=1}^{n}\sup_{z_i}(\log\mathcal{L}^{F_n(z_i)/F_0(z_i)})=\sup_{z}(\log\mathcal{L}^{F_n(z)/F_0(z)})$$

then

$$P_{\mathcal{H}_0}(T_n\geq c)\leq P_{\mathcal{H}_0}(\sup_{z}(\log\mathcal{L}^{F_n(z)/F_0(z)})\geq c)=\alpha_{UIT}.$$

5.4.2 Comparison with Berk-Jones's test

For $C_n=n^{-2}$ the level of our test is less than level of Berk-Jones's test. It is because

$$\frac{1}{n}\frac{1}{n}\sum_{i=1}^{n}\sup_{z_i}(\log\mathcal{L}^{F_n(z_i)/F_0(z_i)})\leq\sup_{z\in\mathcal{Z}}(\frac{1}{n}\log\mathcal{L}^{F_n(z)/F_0(z)})$$

5.4.3 Bahadur efficiency of proposed test

We defined the test statistic as

$$T_n=\frac{1}{n}\sum_{z_i\in\mathcal{A}_i}\log\mathcal{L}^{F_n(z_i)/F_0(z_i)}$$

then

$$P_{\mathcal{H}_0}(\frac{1}{n}T_n\geq t)=P_{\cap H_{0z}}\{\frac{1}{n}(\frac{1}{n}\sum_{z}\log\mathcal{L}^{F_n(z)/F_0(z)})\geq t\}$$

$$\leq P_{\mathcal{H}_{0z_{max}}}\{\frac{1}{n}\frac{1}{n}n(\max_{z}\log\mathcal{L}^{F_n(z)/F_0(z)})\geq t\}=$$

$$P_{\mathcal{H}_{0z_{max}}}\{\frac{1}{n}\log \mathcal{L}_{z_{max}}^{F_n(z)/F_0(z)} \geq t\} \leq \sum_{\mathcal{H}_{1z_{max}}} P_{\mathcal{H}_{0z_{max}}}\{\frac{1}{n}\sum_{i=1}^{n}\log\frac{Bin(1,F(z_{max}))}{Bin(1,F_0(z_{max}))} \geq t\} \quad (*)$$

$$\leq |\mathcal{H}_{1z_{max}}| \max_{F(z_{max})} P_{\mathcal{H}_{0z_{max}}}\{\exp\sum_{i=1}^{n}\log\frac{Bin(1,F(z_{max}))}{Bin(1,F_0(z_{max}))} \geq \exp(nt)\} =$$

$$|\mathcal{H}_1| \max_{F(z_{max})} P_{\mathcal{H}_{0z_{max}}}\{\exp\sum_{i=1}^{n}\log\frac{Bin(1,F(z_{max}))}{Bin(1,F_0(z_{max}))} \geq \exp(nt)\} \leq$$

$$|\mathcal{H}_1| \max_{F(z_{max})} \mathcal{E}_{\mathcal{H}_{0z_{max}}}\{\frac{\exp\sum_{i=1}^{n}\log\frac{Bin(1,F(z_{max}))}{Bin(1,F_0(z_{max}))}}{\exp(nt)}\} \quad \text{(by Markov inequality)} =$$

$$|\mathcal{H}_1| \max_{F(z_{max})} \exp(-nt)\mathcal{E}_{\mathcal{H}_{0z_{max}}}\{\prod_{i=1}^{n}\frac{Bin(1,F(z_{max}))}{Bin(1,F_0(z_{max}))}\} =$$

$$|\mathcal{H}_1| \max_{F(z_{max})} \exp(-nt)\{\mathcal{E}_{\mathcal{H}_{0z_{max}}}\frac{Bin(1,F(z_{max}))}{Bin(1,F_0(z_{max}))}\}^n =$$

$$|\mathcal{H}_1| \max_{F(z_{max})} \exp(-nt)\{\sum Bin(1,F_0(z_{max}))\}^n \leq |\mathcal{H}_1|\exp(-nt)$$

then

$$-\frac{2}{n}\log P_{\mathcal{H}_0}(\frac{1}{n}T_n \geq t) \geq 2t - \frac{2\log|\mathcal{H}_1|}{n}$$

we know that

$$1/n\log\mathcal{L}^{F_n(z)/F_0(z)} = \frac{1}{n}(\log\mathcal{L}^{F_n(z)} - \log\mathcal{L}^{F_0(z)}) =$$

$$\frac{1}{n}(\log\mathcal{L}^{F(z)} - \log\mathcal{L}^{F_0(z)}) + \frac{1}{n}(\log\mathcal{L}^{F_n(z)} - \log\mathcal{L}^{F(z)}) \qquad \text{(iff)} \quad F(z) \neq F_0(z) \quad \text{(under} \quad H_{1z})$$

$$= \frac{1}{n}\sum_{i=1}^{n}\log\frac{Bin(1,F(z))}{Bin(1,F_0(z))} + \frac{1}{n}(\log\mathcal{L}^{F_n(z)} - (\log\mathcal{L}^{F(z)}) \xrightarrow{\mathcal{P}} \mathcal{E}_{H_{1z}}\log\frac{Bin(1,F(z))}{Bin(1,F_0(z))} + (0)\chi^2 =$$

$$KL\{Bin(1,F(z)), Bin(1,F_0(z))\} \quad \text{a.s} \quad \text{under} \quad H_{1z}.$$

Thus

$$\frac{1}{n}\sum_{z}\log\mathcal{L}^{F_n(z)/F_0(z)} \xrightarrow{\mathcal{P}} \mathcal{E}_{H1}KL(Bin(1,F(Y)), Bin(1,F_0(Y)))$$

$$-\frac{2}{n}\log P_{\mathcal{H}_0}(\frac{1}{n}T_n \geq t) \geq \inf_{\mathcal{H}_0}\mathcal{E}_{\mathcal{H}1}KL(Bin(1,F(Y)), Bin(1,F_0(Y)))$$

Bahadur (1967) showed that the other part of inequality for all of tests is right, then

$$-\frac{2}{n}\log P_{\mathcal{H}_0}(\frac{1}{n}T_n \geq t) = 2\inf_{\mathcal{H}_0}\mathcal{E}_{\mathcal{H}1}KL(Bin(1,F(Y)), Bin(1,F_0(Y))).$$

(*) Because

$$\frac{1}{n}\log \mathcal{L}_{z_{max}}^{F_n(z)/F_0(z)} \leq \sup_{F\in\mathcal{H}_1} \frac{1}{n}\sum_{i=1}^{n}\log\frac{Bin(1,F(z_{max}))}{Bin(1,F_0(z_{max}))}.$$

Chapter 6

Proposed Model selection tests based on likelihood and AIC

6.1 Introduction

The major goal of this chapter is to introduce and develop a methodology of model selection. This chapter in theory and method is strongly related to the next chapter. In real situation for any inference about a data set at hand we are interested in selecting a model among a lot of parametric models. In usual hypothesis testing we suppose that $\overline{Y} = (Y_1, Y_2, ..., Y_n)$ is a random sample with density $g(.;\beta)$ for some $\beta \in B$. In simple case we set $B = B_0 \cup B_1$ for two disjoint sets B_0 and B_1. We would like to decide if $\beta \in B_0$ or $\beta \in B_1$, where usually $dim(B_0) < dim(B_1)$. To use the likelihood for hypothesis testing, we may use the standard ratio as the supreme of the likelihood under alternative hypothesis divided on the likelihood under null hypothesis and rejecting the null hypothesis for large

value of this ratio. On the other hand we may use this principle for testing between two elements of two different families. The other approach to make a decision as above is likelihood-based interval inference as the dual of the hypothesis testing. Fortunately both of them work under asymptotic theory for likelihood function.

The restriction by this approach is clear. This approach works very well for nested models, but in other cases it does not work. Our idea is somewhat different, not in principle, but in applying the likelihood function. In fact when we have a data set all hypotheses about the distribution of data are null hypotheses.

We want to report the likelihood under the null hypothesis as the normalized likelihood under indicated parameters in the null hypothesis. We can reject the null hypothesis if its likelihood is too small, then we conclude that there are other hypotheses which are better than our hypothesis. We will use both of hypothesis testing and confidence interval with other interpretation. As we know the hypothesis testing is an absolute discrimination. We may consider the relative discrimination as R^2, *AIC* and *BIC* depending on the problem. This is a fact that this criteria always choose a model. If two competing models are very bad, we would like to be able to reject both of them. The confidence interval inference about the model selection criterion is our idea. Because by the confidence intervals there is an opportunity to select a set of good choice. This interval will be constructed for expectation of the log-likelihood. This confidence interval let us take in order the models. This approach has an interpretation for a confidence interval for Kullback-Leibler discrepancy. The best model will be a model with greatest lower and upper limit for expectation. A model with this kind of upper and lower limits is a model with minimum Kullback-Leibler discrepancy. In the real situation, when we have n observations at hand, before collecting this data we consider some hypotheses like independence and identically distribution about data, the first question is about the true distribution of this sample.

This question is on entire population. This question arises because this is the first step to decision making. This is a question in model selection area. We can assume whether the true density belongs to the postulated model or not. Whether true model belongs to the specified parametric model, $\mathcal{G}$, or not there is a member of this family as $g(.,\beta)$ which is equal or nearly equal to $f(.)$. The difficulty exists yet. To search this member of parametric family we estimate its parameters by the maximum likelihood approach. Now we have a member in the family as $g(y;\hat{\beta}_n)$. This is the best choice in the model, based on the data. Now an estimation of $f(.)$ is $g(y;\hat{\beta}_n)$. If our choice about family was not very bad the likelihood statistic under observed data must be large. Then the normalized maximized (log)likelihood seems to be a good choice as a criterion to model selection. Akaike (1973) has said, "assume that the true distribution does not belong to the specified parametric family", and introduced his criterion. It means that the normalized maximized log-likelihood for observed data does not converge to the expected log-likelihood, where expectation is taken under true density, i.e there is a bias. This last result guides us toward *AIC*. We must emphasize that the hypothesis testing is an absolute discrimination, and the *AIC* is a relative discrimination. Then the classical hypothesis testing is a method and a criterion like *AIC* will answer a different question than hypothesis testing. Anyway we can consider the normalized maximum log-likelihood and *AIC* as the random variables which has a distribution. Before anything we need to formulate the null hypothesis and some comments on our approach.

The aim of this formulation is to decide whether or not $\mathcal{G}$ does contain the true density f? If the statistical model is correctly specified, we have,

$$f(y) = g(y;\beta_\star) \qquad \forall y \in \mathcal{R} \quad \text{and some} \quad \beta_\star \in B$$

then the question which arises could be formulated as a null hypothesis

$$\mathcal{H}_0 : f(y) = g(y;\beta_\star) \qquad \forall y \in \mathcal{R} \quad \text{and some} \quad \beta_\star \in B.$$

In general, $\beta_\star$ is an unknown parameter, in this case a natural way to testing $\mathcal{H}_0$ will be to find an estimator for β which is near to $\beta_\star$ and then to build a procedure for testing. In a simple case $\beta_\star$ could be known. When $\beta_\star$ is unknown the measurability of $g(y;\beta)$ in y for every $\beta \in B$ and continuity of $g(y;\beta)$ in $\beta \in B$ for every $y \in \mathcal{R}$ ensures us that for all n there exists a measurable likelihood estimator (MLE). When the true distribution belongs to $\mathcal{G}$ for some $\beta_\star \in B$ (under $\mathcal{H}_0$), the *MLE* is consistent for $\beta_\star$ under Wald (1949) conditions, see, 4.5 . Now a question is what would happen to the maximum likelihood estimator when the model under consideration is not correct (the model is misspecified)? As we saw in 4.5 it is clear that the maximum likelihood estimator would not converge to $\beta_\star$, because it does not longer makes sense. When a statistical model is misspecified, as we saw in 4.4 the maximum likelihood estimator converges to the minimizer β_0 of the Kullback- Leibler criterion, instead of the parameter which we consider under null hypothesis (the true parameter). Then any difference between the postulated model $g(.,\beta_\star)$ and the true density $f(.)$ is error due to model misspecification.

The theorems 2.1, 2.2, and 3.2 of White (1982) and in more detail in White (1994) are good references for study the asymptotic distribution behavior of $\sqrt{n}(\hat{\beta}_n - \beta_0)$. To evaluate $\mathcal{H}_0$ as our immediate goal, we note that if $\frac{1}{n}\sum_{i=1}^{n} \log g(Y_i;\beta_\star)$ has a large value we will conclude that the postulated model in some sense is near to the true distribution. But in realistic cases $\beta_\star$ is unknown, thus we are going to testing $\mathcal{H}_0$ by a reasonable test statistic which converges to a constant function of $\frac{1}{n}\sum_{i=1}^{n} \log g(Y_i;\beta_\star)$.

When we evaluate the likelihood function at its maximizer we can say that the smaller the likeli-

hood, the worse the goodness of fit. A problem for this kind of fits is that because of transformation properties of likelihood functions w.r.t. change of variable, in general it is not invariant as a goodness of fit test. By definition of $\mathcal{H}_0$ as above, we can consider $\mathcal{H}_1 : f \notin \mathcal{G}$ as the alternative hypothesis, which shows that there is no $\beta \in B$ which permits us to consider $f(x) = g(y;\beta)$. It is clear that this type of alternative is completely vague.

In the most general way we have

$$\begin{cases} \mathcal{H}_0 : f(y) = g(y;\beta_\star) \quad \forall y \in \mathcal{R} \quad \text{and} \quad \text{some} \quad \beta_\star \in B \\ \mathcal{H}_1 : f(y) = g(y) \qquad \text{with} \quad g(.) \neq g(.;.) \end{cases}$$

In other words we want to test the postulated density against different shapes. It is clear that the alternative hypothesis like that is unuseful. In this chapter we consider only the hypothesis like the goodness of fit test as

$$\begin{cases} \mathcal{H}_0 : f(y) = g(y;\beta_\star) \quad \forall y \in \mathcal{R} \quad \text{and} \quad \text{some} \quad \beta_\star \in B \\ \mathcal{H}_1 : f(y) \neq g(y;\beta_\star) \qquad \text{for at least a} \quad y \in \mathcal{R} \end{cases}$$

To evaluate the above hypothesis problem because $\beta_\star$ is unknown we estimate it by maximum likelihood approach and define the normalized maximized likelihood function as the test statistic. We verify the asymptotic distribution of this statistic where its expectation is consider under different situation. The normalized log-likelihood minus its expectation helps us solving the invariant problem under some kind of transformation. This statistic is a part of the AIC criterion, then our theorems in this chapter, asymptotically are valid about AIC. We begin with a simple case when the $\beta_\star$ the parameters involved in the postulated density are known. In this case the asymptotic distribution of our statistic follows by simple application of the Central limit theorem. It helps us to verify

whether the data at hand follow a known density. A normal example is considered and the power of the test is studied by simulation. The result of the simulation shows that our test has a reasonable power although its power is less than the Kolmogorov-Smirnov test. But I think that this is not a disagreement, because the model selection involves a trade-off between simplicity and fit. I want emphasize that in the literature there is no method that is better than all the others under all conditions; on the other hand for any two methods, there are circumstances in which one of them is better than the other one. It means that every method has a some risk even in well behaved situations. The important things is that a method must have a reasonable result.

The other part in this chapter is concerned with the realistic situation when we want to know whether the true density belongs to a parametric family with unknown parameters. Our approach to point estimation is again the maximum likelihood method. Here we assume that for the parameters under study the maximum likelihood estimator exists and it is unique. Biernacki (2004) has made a test which compares the log-likelihood evaluated at $\overline{\beta}$ one of the relative maxima of the log-likelihood function and its expected value, which is calculated as if $\overline{\beta}$ is the true parameter. In fact he proposed a test for testing whether the maximum of his function is a global maximum. On the other hand he detected if a given solution to likelihood equation is consistent.

To searching the asymptotic distribution of our statistic we consider three situations for $\mathcal{E}\{\log g(Y;\beta)\}$ as $\mathcal{E}_{\beta_\star}\{\log g(Y;\beta_\star)\}$, $\mathcal{E}_{\hat{\beta}_n}\{\log g(Y;\hat{\beta}_n)\}$ and $\mathcal{E}_{\beta_\star}\{\log g(Y;\hat{\beta}_n)\}$. The second expectation is the estimator for the first one and the third one is related to risk of estimation when the true density belongs to the $\mathcal{G}$ and also is the relevant part of the Kullback-Leibler criterion. On the other hand the first and third one are the limiting values for our statistic the normalized maximized likelihood in some sense and the difference of the our statistic from the second one converges to zero. For each situation we established a theorem which each one shows that the standard value of normalized maximized

likelihood function asymptotically is distributed according to the normal distribution with mean zero and certain variance. According to Biernacki (2004) we talk about the variance estimation. For each theorem we bring an example to show that it works. We verify our theorem under alternative hypothesis. We define the power of test and showed that our test is consistent. It is shown that our statistics minus its expected normalized loglikelihood under $\beta_\star$ is invariant under orthogonal linear transformation. Our focus in this chapter is on theorem 6.5. This theorem admits us to propose a new method to Multiple Regression model selection. The simulation study shows that this theorem has a good result. The result is given in appendix A.

6.2 Known parameters case

In the first step we consider an i.i.d. random sample of size n and density $f(y)$. We want to test whether the parametric density $g(y;\beta_\star)$ where $\beta_\star$ is known, is well-specified. In fact this test is a goodness of fit test. For doing it we need to evaluate the distribution of the test statistic.
In Theorem 1 we will find the asymptotic density of this statistic. This approach to testing for model selection has the advantage of simplicity. A simulation study to evaluate the performance of our statistic is done and we compare it with the Kolmogorov-Smirnov statistic.

Theorem 6.1 *Let $Y_1,...,Y_n$ i.i.d. random sample with unknown density $f(y)$. Suppose that $\mathcal{E}_{\beta_\star}\{\log g(Y;\beta_\star)\}$ exists, and $\mathcal{E}_{\beta_\star}\{(\log g(Y;\beta_\star))^2\} < \infty$, then*

$$n^{-1/2}\sum_{i=1}^{n}[\log g(Y_i;\beta_\star) - \mathcal{E}_{\beta_\star}\{\log g(Y_i;\beta_\star)\}] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{V}ar_{\beta_\star}\{\log g(Y;\beta_\star)\})$$

Proof : The proof is a direct usage of central limit theorem. We note that

$$\mathcal{E}_{\beta_\star}\{\frac{1}{n}\sum_{1}^{n}\log g(Y;\beta_\star)\} = \mathcal{E}_{\beta_\star}\{\log g(Y;\beta_\star)\}$$

and

$$\mathcal{V}ar_{\beta_\star}\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_\star)\} = \frac{1}{n}\mathcal{V}ar_{\beta_\star}\{\log g(Y;\beta_\star)\}$$

now by CLT

$$n^{-1/2}\sum_{i=1}^{n}[\log g(Y_i;\beta_\star) - \mathcal{E}_{\beta_\star}\{\log g(Y_i;\beta_\star)\}] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{V}ar_{\beta_\star}\{\log g(Y;\beta_\star)\}) \qquad \blacksquare$$

Example 6.1

Suppose that $Y_1, Y_2, ..., Y_n$ i.i.d. with density $f(y)$ we want to test whether or not $Y_1, Y_2, ..., Y_n$ is a i.i.d. random sample with normal density and parameter vector as $\beta_\star = (\mu_\star, \sigma_\star^2)$, where $\beta_\star$ is given, it means that under null hypothesis $Y_i \sim \mathcal{N}(\mu_\star, \sigma_\star^2)$. Then

$$g(Y_1, ..., Y_n; \beta_\star) = \prod_{i=1}^{n} g(Y_i, ; \beta_\star) = (2\pi\sigma_\star^2)^{-n/2}\exp\{-\frac{1}{2}\sum_{i=1}^{n}(\frac{Y_i - \mu_\star}{\sigma_\star})^2\}$$

by taking logarithm we have

$$\sum_{i=1}^{n}\log g(Y_i;\beta_\star) = -\frac{n}{2}\log(2\pi\sigma_\star^2) - \frac{1}{2}\sum_{i=1}^{n}(\frac{Y_i - \mu_\star}{\sigma_\star})^2 = nb - \frac{1}{2}\sum_{i=1}^{n}(\frac{Y_i - \mu_\star}{\sigma_\star})^2$$

where $b = -\frac{1}{2}\log 2\pi\sigma_\star^2$. We know that

$$X_n(\overline{Y}) = \sum_{1}^{n}(\frac{Y_i - \mu_\star}{\sigma_\star})^2 \sim \chi_n^2.$$

Thus $\mathcal{E}_{\beta_\star}\{X_n(\overline{Y})\} = n$ and $\mathcal{V}ar_{\beta_\star}\{X_n(\overline{Y})\} = 2n$. Now by straight application of CLT we have:

$$\frac{n^{-1/2}(X_n(\overline{Y}) - n)}{\sqrt{2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Now as an application of Theorem 6.1, $\mathcal{E}_{\beta_\star}\{\log g(Y,\beta_\star)\} = -\frac{1}{2}\log 2\pi\sigma_0^2 - 1/2$ and $\mathcal{V}ar_{\beta_\star}\{\log g(Y;\beta_\star)\} = \frac{1}{2}$ which gives us the same result as above

$$\frac{n^{-1/2}\sum_{i=1}^{n}[b-\frac{1}{2}\{(\frac{Y_i-\mu_\star}{\sigma_\star})^2\}-b+\frac{1}{2}]}{\sqrt{\frac{1}{2}}} = \frac{n^{-1/2}(n-X_n(\overline{Y}))}{\sqrt{2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

By symmetry of standard normal distribution the straight CLT and our theorem are the same.

Also

$$T_n(\overline{Y};\beta_\star) = \frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_\star) = b - \frac{1}{2n}X_n(\overline{Y}).$$

In consequence, weighted log-likelihood function has a biased χ^2 distribution with expectation and variance as follows

$$\mathcal{E}_{\beta_\star}\{T_n(\overline{Y};\beta_\star)\} = b - \frac{1}{2}$$

and

$$\mathcal{V}ar_{\beta_\star}\{T_n(\overline{Y};\beta_\star)\} = \frac{1}{2n}$$

The distribution function of T_n is given by,

$$P(T_n\{\overline{Y};\beta_\star\} \le t) = P(X_n(\overline{Y}) > 2n(b-t)) = 1 - \frac{1}{\Gamma(n/2)2^{n/2}}\int_0^{2n(b-t)} x_n^{\frac{n}{2}-1}e^{x_n/2}\,dx_n \qquad b > t$$

which is the survival function for a χ^2 distribution. The integral is an incomplete gamma function.

We reject $\mathcal{H}_0$ if

$$T_n(\overline{Y};\beta_\star) < t$$

then the test function is given by

$$\phi(\overline{Y}) = \begin{cases} 1 & \text{if } X_n(\overline{Y}) > C \\ 0 & \text{if } X_n(\overline{Y}) < C \end{cases}$$

where $C = 2n(b-t)$. The level of test is,

$$\alpha = \mathcal{E}_{\mathcal{H}_0}(\phi(\overline{Y})) = P_{\mathcal{H}_0}(X_n(\overline{Y}) > C)$$

where $C = \chi^2_{\alpha,n}$. The power of the test is given by

$$\gamma(\mu) = \mathcal{E}_{\mathcal{H}_1}(\phi(\overline{Y})) = P_{\mathcal{H}_1}(X_n(\overline{Y}) > \chi^2_{\alpha,n})$$

Power Simulation:

It is known that if $Y_1, Y_2, ..., Y_n$ are i.i.d. $\mathcal{N}(\mu_\star, \sigma^2_\star)$ and if μ_0 as the possible value of $\mu_\star$ is given the statistic $\sum_{i=1}^n (Y_i - \mu_0)^2$ has the non-central χ^2 distribution with n degrees of freedom and $n(\mu_\star - \mu_0)^2$ as its non-centrality parameter. By this fact at $\sigma^2_\star = 1$ we have;

$$X_n(\overline{Y}) \sim \chi^2(n, n(\mu_\star - \mu_0)^2).$$

For power computation we use the software "R". We compare the power of our test to the power of the Kolmogorov-Smirnov test. We consider the data generating density as the $\mathcal{N}(0,1)$, each time we generate $n = 5, 30, 50$ observations of this density.

The size of the simulation is $m = 10000$. We want to test whether or not $Y_i \sim \mathcal{N}((0.1)t, 1), \quad i = 1, 2, ..., n$ and $t = 1, 2, ..., 30$. (as the alternative hypothesis) We set the pre-assigned levels as $\alpha = 0.2, 0.1, 0.05$. The result of the simulation is given in Tables 1.4-3.4 and Figure 9. For $n = 5$ at any level two tests are nearly equivalent. At $n = 30$ and $n = 50$ the power of Kolmogorov-Smirnov test is better than our test. For large value of $(0.1)t$ i.e. when we are far from of true density the two tests have almost the same power. As we see the Kolmogorov-Smirnov is more powerful than our test. But we emphasize that our approach has a reasonable power and on the other hand the likelihood function is a simple function which any recearcher know it. When we are one unit far from the true mean the power is about one.

Table 6.1- Power comparison of the Kolmogorov-Smirnov's test (K-S) and proposed test (LL) based on likelihood function for n=5

Normal Mean	$\alpha = 0.2$		$\alpha = 0.1$		$\alpha = 0.05$	
	K-S	LL	K-S	LL	K-S	LL
0.1	0.2077	0.2050	0.1087	0.1034	0.0557	0.0522
0.2	0.2279	0.2200	0.1315	0.1138	0.0656	0.0588
0.3	0.2718	0.2500	0.1595	0.1317	0.087	0.0705
0.4	0.3269	0.2799	0.1951	0.1572	0.1191	0.0880
0.5	0.3971	0.3245	0.2571	0.1920	0.1594	0.1124
0.6	0.4623	0.3778	0.3145	0.2356	0.2176	0.1447
0.7	0.5354	0.4385	0.3892	0.2882	0.2643	0.1857
0.8	0.6129	0.5047	0.4614	0.3492	0.3339	0.2360
0.9	0.6856	0.5738	0.5418	0.4172	0.4076	0.2953
1.0	0.7425	0.6430	0.6125	0.4900	0.4819	0.3627
1.1	0.8051	0.7096	0.6822	0.5650	0.5525	0.4363
1.2	0.8538	0.7710	0.7567	0.6392	0.6282	0.5137
1.3	0.8957	08252	0.8054	0.7096	0.6941	0.5917
1.4	0.9243	0.8711	0.8574	0.7736	0.7659	0.6671
1.5	0.9446	0.9083	0.8949	0.8293	0.8056	0.7370
1.6	0.9694	0.9372	0.9273	0.8758	0.8544	0.7990
1.7	0.9790	0.9585	0.9456	0.9128	0.8901	0.8517
1.8	0.9879	0.9737	0.9644	0.9411	0.9249	0.8945
1.9	0.9910	0.9840	0.9770	0.9618	0.9495	0.9277
2.0	0.9951	0.9907	0.9853	0.9761	0.9667	0.9523
2.1	0.9969	0.9948	0.9921	0.9857	0.9791	0.9698
2.2	0.9984	0.9972	0.9952	0.9918	0.9850	0.9817
2.3	0.9995	0.9986	0.9972	0.9955	0.9908	0.9893
2.4	0.9997	0.9993	0.9980	0.9976	0.9955	0.9940
2.5	0.9998	0.9997	0.9995	0.9988	0.9959	0.9968
2.6	0.9998	0.9998	0.9999	0.9994	0.9980	0.9984
2.7	0.9999	0.9999	0.9997	0.9997	0.9991	0.9992
2.8	1.0000	1.0000	1.0000	1.0000	0.9993	0.9996
2.9	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998
3.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 6.2- Power comparison of the Kolmogorov-Smirnov's test (K-S) and proposed test (LL) based on likelihood function for n=30.

Normal Mean	$\alpha = 0.2$		$\alpha = 0.1$		$\alpha = 0.05$	
	K-S	LL	K-S	LL	K-S	LL
0.1	0.2224	0.2117	0.1254	0.1076	0.0666	0.0547
0.2	0.3561	0.2478	0.2261	0.1232	0.1397	0.0702
0.3	0.5420	0.3113	0.3849	0.1785	0.2655	0.1012
0.4	0.7116	0.4034	0.5717	0.2524	0.4495	0.15469
0.5	0.8512	0.5202	0.7510	0.3574	0.6190	0.2391
0.6	0.9361	0.6495	0.8715	0.4896	0.7846	0.3576
0.7	0.9780	0.7727	0.9481	0.6343	0.9012	0.5035
0.8	0.9943	0.8720	0.9818	0.7693	0.9573	0.6577
0.9	0.9982	0.9387	0.9950	0.8748	0.9874	0.7945
1.0	0.9997	0.9755	0.9993	0.9427	0.9970	0.8951
1.1	1.0000	0.9920	1.0000	0.9783	0.9993	0.9553
1.2	1.0000	0.9978	1.000	0.9933	0.9998	0.9844
1.3	1.0000	1.0000	1.0000	0.9983	1.0000	0.9956
1.4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 6.3- Power comparison of the Kolmogorov-Smirnov's test (K-S) and proposed test (LL) based on likelihood function for n=50

Normal Mean	$\alpha = 0.2$		$\alpha = 0.1$		$\alpha = 0.05$	
	K-S	LL	K-S	LL	K-S	LL
0.1	0.2843	0.2150	0.1649	0.1097	0.0846	0.0559
0.2	0.4970	0.2620	0.3298	0.1419	0.2071	0.0763
0.3	0.7151	0.3461	0.57730	0.2048	0.4388	0.1193
0.4	0.8891	0.4684	0.7960	0.3079	0.6570	0.1975
0.5	0.9674	0.6172	0.9236	0.4531	0.8483	0.3222
0.6	0.9934	0.7657	0.9824	0.6236	0.9502	0.4905
0.7	0.9994	0.8832	0.9968	0.7843	0.9884	0.6745
0.8	0.9998	0.9544	0.9993	0.9014	0.9980	0.8313
0.9	1.0000	0.9865	0.9999	0.9655	0.9999	0.9321
1.0	1.0000	0.9971	1.0000	0.9910	1.0000	0.9800
1.1	1.0000	1.0000	1.0000	0.9983	1.0000	0.9960
1.2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

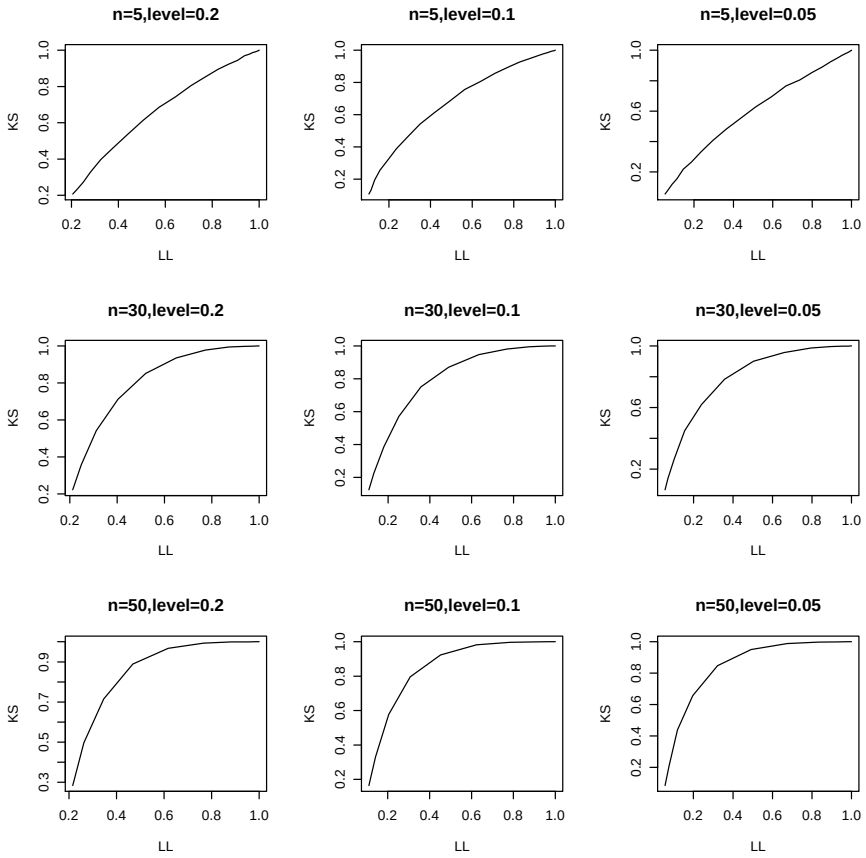


Figure 10: Power comparison of Kolmogorov-Smirnov test and normalized log-likelihood test.

Example 6.2

Suppose $Y_1, Y_2, ..., Y_n$ is an i.i.d.sample with unknown density $f(.)$ we want to test whether the exponential density is a good fit to Y? Then

$$\mathcal{H}_0 : f(y) = g(y; \lambda_\star).$$

with

$$g(y; \lambda_\star) = \lambda_\star^{-1} e^{-\lambda_\star^{-1} y}$$

we have:

$$\frac{1}{n} \sum_{i=1}^{n} \log g(Y_i; \lambda_\star) = -\log \lambda_\star - \frac{\sum_{i=1}^{n} Y_i}{n\lambda_\star}$$

It is known that

$$\sum_{i=1}^{n} Y_i \sim \Gamma(n, \lambda_\star)$$

by CLT,

$$n^{-1/2}(\frac{\sum_{i=1}^{n} Y_i}{n} - \lambda_\star) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \lambda_\star^2)$$

now we have

$$\frac{1}{n} \sum_{i=1}^{n} \log g(Y_i; \lambda_\star) = -\log \lambda_\star - \frac{\sum_{i=1}^{n} Y_i}{n\lambda_\star}$$

then

$$\frac{\frac{1}{n}\sum_{i=1}^{n} \log g(Y_i; \lambda_\star) - (-\log \lambda_\star - 1)}{\sqrt{\frac{1}{n}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

On the other hand by Theorem 6.1, for $i = 1, 2, ..., n$, we have,

$$\mathcal{E}_{\lambda_\star}\{\log g(Y_i; \lambda_\star)\} = -\log \lambda_\star - 1$$

and

$$\frac{1}{n} \mathcal{V}ar_{\lambda_\star}\{\log g(Y_i; \lambda_\star)\} = \frac{1}{n}$$

which gives the same asymptotic density for normalized log-likelihood function as above.

$$P(\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\lambda_\star)\leq t)=1-\frac{1}{\Gamma(n)\lambda_\star^n}\int_0^{-n\lambda_\star(t+\log\lambda_\star)}x_n^{n-1}e^{x_n/\lambda_\star}\,dx_n \qquad t<-\log\lambda_\star$$

Where $X_n=\sum_{i=1}^{n}Y_i$, this distribution is the survival function for a gamma distribution. Again we can consider the integral as an incomplete gamma function. In this case

$$\phi(\overline{Y})=\begin{cases}1 & \text{if } \frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\lambda_\star)<t\\ 0 & \text{if } \frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\lambda_\star)>t\end{cases}$$

now

$$\alpha=\mathcal{E}_{\mathcal{H}_0}(\phi(\overline{Y}))=P_{\mathcal{H}_0}(X_n(\overline{Y})>-n\lambda_\star(t+\log\lambda_\star))=$$

$$1-P_{\mathcal{H}_0}(X_n(\overline{Y})<-n\lambda_\star(t+\log\lambda_\star)),\quad t<-\log\lambda_\star$$

We may compare this test with a Kolmogorov-Smirnov test or other suitable tests.

6.3 Unknown parameters case

To find a test for model selection we consider a more realistic case, when the parameters of the postulated model are unknown. We established some theorems under specified and mis-specified hypotheses to find the test for model selection in different situations. The main part of them are the differences between normalized maximized log-likelihood and $\mathcal{E}\{\log g(.,\beta)\}$ where β could be $\beta_\star$ or β_0 dependent on specified or mis-specified case respectively, see 4.4.1, and expectation is considered under different situations. All the theorems apply to *AIC*, because in asymptotic situation the distribution of normalized maximized log-likelihood and *AIC* is not different. Here we emphasize that our focus is on *AIC*.

As a starting point we want to test the null hypothesis $\mathcal{H}_0 : f(y) = g(y;\beta_\star)$ for all $y \in \mathcal{R}$ and some $\beta_\star \in B$, if we reduce it to $\mathcal{H}_0 : f(y) = g(y;\beta_\star)$ *a.e.* in possible range of y for some $\beta_\star \in B$, this null hypothesis is equivalent to testing for:

$$\mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n}\log f(Y_i)\} = \mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_\star)\}$$

or

$$\mathcal{E}_{\beta_\star}\{\frac{1}{n}\sum_{i=1}^{n}\log f(Y_i)\} = \mathcal{E}_{\beta_\star}\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_\star)\}$$

where $\mathcal{E}_\beta$ stands for expectation on $g(.,\beta)$.

We saw that when $\beta_\star$ is known, we reject $\mathcal{H}_0$ for a small value of $\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_\star)$. If $\beta_\star$ is unknown we propose the test statistic as

$$T_n(\overline{Y},\hat{\beta}_n) = \frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)$$

where $\overline{Y} = (Y_1,...,Y_n)$. As we saw in 4.4.1 this is the bias estimator for the important part of KL divergence and then an estimator for discrepancy (distance) between the true density and the postulated model. The test function for this type hypothesis is given by

$$\phi(\overline{Y}) = \begin{cases} 1 & \text{if } T_n(\overline{Y},\hat{\beta}_n) < K_n \\ \\ 0 & \text{if } T_n(\overline{Y},\hat{\beta}_n) > K_n \end{cases}$$

Note that according to the weak law of large numbers, for each $\beta \in B$ we have,

$$\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta)\} \xrightarrow{\mathcal{P}} \mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta)\} \qquad (1)$$

In theorem 6.2, by (1), convergence of $\hat{\beta}_n$ to $\beta_\star$ and conditions (C0)-(C4), we will show that $T_n(\overline{Y},\hat{\beta}_n)$ is a consistent estimator for $\mathcal{E}_{\beta_\star}\{(\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_\star)\}$. It is noticeable that all the theorems in this chapter are useful to construct the confidence interval for expectation part that is $\mathcal{E}_\beta(\log g(Y_i;\beta))$ where in the theorems we replace the β by $\beta_\star$ either $\hat{\beta}_n$ in argument or in expectation.

Theorem 6.2 *Suppose that $Y_1,...,Y_n$ i.i.d with unknown density $f(.)$. Let $\mathcal{G} = \{g(.,\beta);\beta \in B \subseteq \mathcal{R}\}$ is a parametric family of assumed densities for $Y_i's$. If $\mathcal{H}_0$ holds, under conditions (C0)-(C4) and (1) we have:*

$$T_n(\overline{Y},\hat{\beta}_n) \xrightarrow{\mathcal{P}} \mathcal{E}_{\beta_\star}\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_\star)\}.$$

Proof :

$$|\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n) - \mathcal{E}_{\beta_\star}(\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_\star))| =$$

$$|\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n) - \mathcal{E}_{\beta_\star}(\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)) + \mathcal{E}_{\beta_\star}(\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)) - \mathcal{E}_{\beta_\star}(\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_\star))|$$

$$\leq |\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n) - \mathcal{E}_{\beta_\star}(\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n))| + |\mathcal{E}_{\beta_\star}(\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)) - \mathcal{E}_{\beta_\star}(\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_\star))|$$

$$\leq \sup_{\beta\in B}|\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta) - \mathcal{E}_{\beta_\star}(\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta))| + |\mathcal{E}_{\beta_\star}(\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)) - \mathcal{E}_{\beta_\star}(\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_\star))| \qquad (2)$$

By (C4) $|\log g(y;\beta)| \leq M(y)$ On the other hand by (1) we have

$$\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta) - \mathcal{E}_{\beta_\star}(\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta)) \xrightarrow{\mathcal{P}} 0$$

Now under conditions (C0) and (C1) by theorem $B3$ the first term in (2) converges to zero. For the second term in (2)

$$\mathcal{E}_{\beta_\star}\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n) - \frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_\star)\} = [\frac{1}{n}\sum_{i=1}^{n}\mathcal{E}_{\beta_\star}(\nabla\log g(Y_i;\beta_\star))](\hat{\beta}_n - \beta_\star) + o_p(\hat{\beta}_n - \beta_\star)$$

by (C3) $\mathcal{E}_{\beta_\star}(\nabla\log g(Y_i;\beta_\star))$ exists and is equal to zero, this completes the proof. ■.

Theorem 6.3 *Under conditions $(C0)-(C3)$ suppose that $0 < \mathcal{V}ar_{\beta_\star}\{\log g(Y,\beta_\star)\} < \infty$, we have:*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n) - \mathcal{E}_{\beta_\star}(\log g(Y_i;\beta_\star))] \xrightarrow{\mathcal{L}} \mathcal{N}(0,\mathcal{V}ar_{\beta_\star}\{\log g(Y,\beta_\star)\}).$$

Proof :

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n) - \mathcal{E}_{\beta_\star}(\log g(Y_i;\beta_\star))] =$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n)-\log g(Y_i;\beta_\star)+\log g(Y_i;\beta_\star)-\mathcal{E}_{\beta_\star}(\log g(Y_i;\beta_\star))]=$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n)-\log g(Y_i;\beta_\star)]+\frac{1}{\sqrt{n}}\sum_{1}^{n}[\log g(Y_i;\beta_\star)-\mathcal{E}_{\beta_\star}(\log g(Y_i;\beta_\star)]$$

The second term on the right, by direct usage of the central limit theorem (CLT) asymptotically has a normal distribution with average and variance equal to zero and $\mathcal{V}ar_{\beta_\star}\{\log g(Y,\beta_\star)\}$ respectively.

For the first term, by Taylor's expansion we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n)-\log g(Y_i;\beta_\star)]=$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[(\hat{\beta}_n-\beta_\star)\nabla\log g(Y_i;\beta_\star)+o_p(\hat{\beta}_n-\beta_\star)]=$$

$$\sqrt{n}(\hat{\beta}_n-\beta_\star)\frac{1}{n}\sum_{i=1}^{n}\nabla\log g(Y_i;\beta_\star)+\sqrt{n}o_p(\hat{\beta}_n-\beta_\star)\qquad(3)$$

we know that under regularity conditions $(C0)-(C3)''$

$$\sqrt{n}(\hat{\beta}_n-\beta_\star)\xrightarrow{\mathcal{L}}\mathcal{N}(0,I_1^{-1}(\beta_\star)).\qquad(4)$$

and by WLLN

$$\frac{1}{n}\sum_{i=1}^{n}\nabla\log g(Y_i;\beta_\star)\xrightarrow{\mathcal{P}}\mathcal{E}_{\beta_\star}(\nabla\log g(Y;\beta_\star))=0$$

by Slutsky's theorem

$$\sqrt{n}(\hat{\beta}_n-\beta_\star)\frac{1}{n}\sum_{i=1}^{n}\nabla\log g(Y_i;\beta_\star)\xrightarrow{\mathcal{L}}0$$

thus

$$\sqrt{n}(\hat{\beta}_n-\beta_\star)\frac{1}{n}\sum_{i=1}^{n}\nabla\log g(Y_i;\beta_\star)\xrightarrow{\mathcal{P}}0.\qquad(5)$$

On the other hand

$$\sqrt{n}o_p(\hat{\beta}_n-\beta_\star)=o_p(\sqrt{n}(\hat{\beta}_n-\beta_\star))$$

by (4) $\sqrt{n}(\hat{\beta}_n - \beta_\star) = O_p(1)$ then

$$\sqrt{n}o_p(\hat{\beta}_n - \beta_\star) = o_p(O_p(1)) = o_p(1) \qquad (6)$$

by (5) and (6)

$$\sqrt{n}(\hat{\beta}_n - \beta_\star)\frac{1}{n}\sum_{i=1}^{n}\nabla \log g(Y_i;\beta_\star) + \sqrt{n}o_p(\hat{\beta}_n - \beta_\star) \xrightarrow{\mathcal{P}} 0$$

Now applying the Slutsky's theorem. ■.

6.4 Test function

As we saw we reject $\mathcal{H}_0$ if

$$T_n(\overline{Y}, \hat{\beta}_n) = \frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n) < K_n$$

then the test function is given by

$$\phi(\underline{Y}) = \begin{cases} 1 & \text{if} \quad T_n(\overline{Y}, \hat{\beta}_n) < K_n \\ \\ 0 & \text{if} \quad T_n(\overline{Y}, \hat{\beta}_n) > K_n \end{cases}$$

The level of test is defined by

$$\alpha_n = \mathcal{E}_{\mathcal{H}_0}(\phi(\overline{Y})) = P_{\mathcal{H}_0}(T_n(\overline{Y}, \hat{\beta}_n) < K_n) \to \alpha_\infty$$

Now we have

$$\frac{\sqrt{n}(K_n - \mathcal{E}_{\beta_\star}(\log g(Y;\beta_\star)))}{\sqrt{\mathcal{V}ar_{\beta_\star}(\log g(.;\beta_\star)}} \to Z_{\alpha_\infty} < 0$$

where the standard value Z_{α_∞} is related to asymptotic distribution of $\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)$ under Theorem 6.3, and this is the α-quantile of the standard normal distribution. Then

$$K_n = \mathcal{E}_{\beta_\star}\{\log g(Y;\beta_\star)\} + Z_{\alpha_\infty}\sqrt{\mathcal{V}ar_{\beta_\star}(\log g(Y;\beta_\star))/n}$$

We consider the alternative hypothesis as

$$\mathcal{H}_1 : \nexists \quad \beta \in B \quad \text{such that} \quad f(y) = g(y;\beta)$$

which means that $\mathcal{H}_1 : f(.) \notin \mathcal{G}$. It is a statistical agnostic which does not change our results considerably. In this case we saw that there exists

$$\beta_0 = \arg\max_{\beta \in B} \mathcal{E}_{\mathcal{H}_1}\{\log g(Y;\beta)\} \quad \text{such that} \quad \hat{\beta}_n \xrightarrow{\mathcal{P}} \beta_0.$$

6.5 Variance estimation

We need to estimate $\mathcal{V}ar_{\beta_\star}\{\log g(Y,\beta_\star)\}$ By theorem 6.3 we have,

$$\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n) - \mathcal{E}_{\beta_\star}(\log g(Y_i;\beta_\star))]}{\sqrt{\mathcal{V}ar_{\beta_\star}\log g(Y,\beta_\star)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Biernacki (2004) has introduced two natural consistent estimators for $\mathcal{V}ar_{\beta_\star}\log g(Y,\beta_\star)$ which are

$$v(\hat{\beta}_n) = \mathcal{V}ar_{\hat{\beta}_n}\{\log g(Y;\hat{\beta}_n)\}$$

and

$$V_n(\hat{\beta}_n) = \frac{1}{n}\sum_{i=1}^{n}(\log g(Y_i;\hat{\beta}_n)^2 - (\frac{1}{n}\sum_{i=1}^{n}(\log g(Y_i;\hat{\beta}_n))^2$$

In proposition 1 he has shown that under consistency of $\hat{\beta}_n$ both of $v(\hat{\beta}_n)$ and $V_n(\hat{\beta}_n)$ converge to $\mathcal{V}ar_{\beta_0}\{\log g(Y,\beta_0)\}$ in probability.

Now we consider these two estimators. It is clear that

$$\frac{v(\hat{\beta}_n)}{\mathcal{V}ar_{\beta_*}\{\log g(Y,\beta_*)\}} \xrightarrow{\mathcal{P}} 1$$

and also

$$\frac{V_n(\hat{\beta}_n)}{\mathcal{V}ar_{\beta_*}\{\log g(Y,\beta_*)\}} \xrightarrow{\mathcal{P}} 1.$$

By Slutsky's theorem we have:

$$\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n) - \mathcal{E}_{\beta_\star}\{\log g(Y_i;\beta_\star)\}]/\sqrt{\mathcal{V}ar_{\beta_\star}\{\log g(Y,\beta_\star)\}}}{\sqrt{v(\hat{\beta}_n)/\mathcal{V}ar_{\beta_\star}\{\log g(Y,\beta_\star)\}}} =$$

$$\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n) - \mathcal{E}_{\beta_\star}\{\log g(Y_i;\beta_\star)\}]}{\sqrt{v(\hat{\beta}_n)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

We have the same result for $V_n(\hat{\beta}_n)$ as,

$$\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n) - \mathcal{E}_{\beta_\star}\{\log g(Y_i;\beta_\star)\}]}{\sqrt{V_n(\hat{\beta}_n)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Example 6.3

Suppose $Y_1, Y_2, ..., Y_n$ is an i.i.d. sample with unknown density $f(.)$. We want to test whether $f(.)$ is a member of the normal family. Formally we want to test that $\mathcal{H}_0 : Y \sim \mathcal{N}(\mu_\star, \sigma_0^2)$ where $\mu_\star$ is unknown and σ_0^2 is known. The MLE of $\mu_\star$ is $\hat{\mu}_n = \tilde{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Now

$$g(Y_i;\sigma_0^2,\mu_\star) = (2\pi)^{-1/2}(\sigma_0^2)^{-1/2}\exp\{-\frac{1}{2}(\frac{Y_i-\mu_\star}{\sigma_0})^2\}$$

and

$$\log g(Y_i;\sigma_0^2,\hat{\mu}_n) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma_0^2 - \frac{1}{2\sigma_0^2}(Y_i-\hat{\mu}_n)^2$$

the weighted log-likelihood function is,

$$\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\sigma_0^2,\hat{\mu}_n) = b - \frac{\hat{\sigma}_n^2}{2\sigma_0^2}$$

where $b = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma_0^2$ and $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{\mu}_n)^2$ the *MLE* of the population variance (which has assumed as a known). Now

$$\mathcal{X} = n\frac{\hat{\sigma}_n^2}{\sigma_0^2} \sim \chi_{n-1}^2$$

by CLT we know that

$$\frac{n\frac{\hat{\sigma}_n^2}{\sigma_0^2}-(n-1)}{\sqrt{2(n-1)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

and then

$$\frac{\frac{\hat{\sigma}^2}{2\sigma_0^2}-\frac{n-1}{2n}}{\sqrt{\frac{n-1}{2n^2}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

As we see $\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\sigma_0^2,\hat{\mu}_n)$ only depends on $\hat{\sigma}_n^2$ which is an ancillary statistic for $\mu_\star$ the unknown parameter. By direct usage of the CLT, and the Lemma A1 we have,

$$\frac{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\sigma_0^2,\hat{\mu}_n)-(b-\frac{n-1}{2n})}{\sqrt{\frac{n-1}{2n^2}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

On the other hand, for $i=1,2,...,n$

$$\mathcal{E}_{\mu_\star}(\log g(Y_i;\sigma_0^2,\mu_\star)) = b-\frac{1}{2}$$

and

$$\frac{1}{n}\mathcal{V}ar_{\mu_\star}\{\log g(Y_i;\sigma_0^2,\mu_\star)\} = \frac{1}{2n}$$

which indicate we do not need to variance estimation. Now by Theorem 6.3,

$$\frac{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\sigma_0^2,\hat{\mu}_n)-(b-\frac{1}{2})}{\sqrt{\frac{1}{2n}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

But for a large n we consider $n \simeq n-1$, and these two results are the same.

Here

$$\frac{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\sigma_0^2,\hat{\mu}_n)-\mathcal{E}_{\mu_\star}\{\log g(Y_i;\sigma_0^2,\mu_\star)\}}{\sqrt{\frac{1}{n}\mathcal{V}ar_{\mu_\star}\{\log g(Y_i;\sigma_0^2,\mu_\star)\}}} = \frac{\frac{1}{2}-\frac{\hat{\sigma}_n^2}{2\sigma_0^2}}{\sqrt{\frac{1}{2n}}}$$

Then the gof in this example reduce to a comparison between the sample and population variance.

With the preassigned level test we have

$$\mathcal{E}_{\mathcal{H}_0}(\phi(\underline{Y})) = P_{\mathcal{H}_0}\{\frac{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\sigma_0^2,\hat{\mu}_n)-\mathcal{E}_{\mu_\star}\{\log g(Y_i;\sigma_0^2,\mu_\star)}{\sqrt{\frac{1}{n}\mathcal{V}ar_{\mu_\star}\{\log g(Y_i;\sigma_0^2,\mu_\star)\}}} < \frac{K_n-b+\frac{1}{2}}{\sqrt{\frac{1}{2n}}}\}$$

thus

$$K_n = -Z_{\frac{\alpha}{2}}\sqrt{\frac{1}{2n}} + b - \frac{1}{2}$$

In this example we are able to easily compute the power function. It is because under any density for Y_i's $\hat{\sigma}_n^2$ has a asymptotic normal density. In fact

$$\sqrt{n}(\hat{\sigma}_n^2 - \sigma_1^2) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mu_1'^4 - \sigma_1^4)$$

where $\mu_1 = \mathcal{E}_{\mathcal{H}_1}(Y)$, $\sigma_1^2 = \mathcal{V}ar_{\mathcal{H}_1}(Y)$, $0 < \mu_1'^4 = \mathcal{E}_{\mathcal{H}_1}(Y-\mu_1)^4 < \infty$ and $\mu_1'^4 > \sigma_1^4$.

Now

$$\gamma_n = P\{Z > \frac{\sqrt{n}[(b-K_n)2\sigma_0^2 - \sigma_1^2]}{\sqrt{\mu_1'^4 - \sigma_1^4}}\} = P\{Z > \frac{\sqrt{n}[(\frac{1}{2} + Z_{\alpha/2}\sqrt{\frac{1}{2n}})2\sigma_0^2 - \sigma_1^2]}{\sqrt{\mu_1'^4 - \sigma_1^4}}\}$$

Theorem 6.4 *Under Theorem 6.3*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n) - \mathcal{E}_{\hat{\beta}_n}(\log g(Y_i;\hat{\beta}_n))] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{V}ar_{\beta_\star}\{\log g(Y,\beta_\star)\}).$$

Proof :

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n) - \mathcal{E}_{\hat{\beta}_n}(\log g(Y_i;\hat{\beta}_n))] =$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n) - \mathcal{E}_{\beta_\star}(\log g(Y_i;\beta_\star))] - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\mathcal{E}_{\hat{\beta}_n}(\log g(Y_i;\hat{\beta}_n)) - \mathcal{E}_{\beta_\star}(\log g(Y_i;\beta_\star))]$$

By Theorem 6.3

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n) - \mathcal{E}_{\beta_\star}(\log g(Y_i;\beta_\star))] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{V}ar_{\beta_\star}\{\log g(Y,\beta_\star)\}). \qquad (7)$$

and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\mathcal{E}_{\hat{\beta}_n}(\log g(Y_i;\hat{\beta}_n)) - \mathcal{E}_{\beta_\star}(\log g(Y_i;\beta_\star)] =$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[(\hat{\beta}_n - \beta_\star)\mathcal{E}_{\beta_\star}(\nabla \log g(Y_i,\beta_\star)) + o_p(\hat{\beta}_n - \beta_\star)] =$$

$$\sqrt{n}(\hat{\beta}_n - \beta_\star)\frac{1}{n}\sum_{i=1}^{n}\mathcal{E}_{\beta_\star}(\nabla \log g(Y_i,\beta_\star)) + o_p(\sqrt{n}((\hat{\beta}_n - \beta_\star)) \qquad (8)$$

The first term clearly is zero and the second term by (6) converges to zero; now by Slutsky's theorem for (7) and (8) the result holds. ■

Example 6.4

Suppose $Y_1, Y_2, ..., Y_n$ is an i.i.d. sample with unknown density $f(.)$. We want to check that whether $f(.)$ is a member of the normal family. Formally we want to test that $\mathcal{H}_0 : Y \sim \mathcal{N}(\mu_\star, \sigma_0^2)$ where $\mu_\star$ is unknown and σ_0^2 is known. The *MLE* of $\mu_\star$ is given by $\hat{\mu}_n = \tilde{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$

$$g(Y_i; \sigma_0^2, \mu_\star) = (2\pi)^{-1/2}(\sigma_0^2)^{-1/2}\exp\{-\frac{1}{2}(\frac{Y_i - \mu_\star}{\sigma_0})^2\}$$

and

$$\log g(Y_i; \sigma_0^2, \hat{\mu}_n) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma_0^2 - \frac{1}{2\sigma_0^2}(Y_i - \hat{\mu}_n)^2\}.$$

Now

$$\sum_{i=1}^n \log g(Y_i; \sigma_0^2, \hat{\mu}_n) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma_0^2 - \frac{1}{2\sigma_0^2}\sum_{i=1}^n (Y_i - \hat{\mu}_n)^2$$

and

$$\mathcal{E}_{\mu_\star}\{\sum_{i=1}^n \log g(Y_i; \sigma_0^2, \hat{\mu}_n)\} = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma_0^2 - \frac{1}{2\sigma_0^2}\sum_{i=1}^n (Y_i - \hat{\mu}_n)^2 = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma_0^2 - \frac{n-1}{2}$$

and also

$$\mathcal{E}_{\hat{\mu}_n}\{\sum_{i=1}^n \log g(Y_i; \sigma_0^2, \hat{\mu}_n)\} = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma_0^2 - \frac{1}{2\sigma_0^2}\sum_{i=1}^n (Y_i - \hat{\mu}_n)^2\} = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma_0^2 - \frac{n-1}{2}$$

now by Theorem 6.4 we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n [\log g(Y_i; \hat{\mu}_n) - \mathcal{E}_{\hat{\mu}_n}(\log g(Y_i; \hat{\mu}_n))] = -\frac{1}{2\sigma_0^2}\sum_{i=1}^n (Y_i - \hat{\mu}_n)^2 + \frac{n-1}{2} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{V}ar_{\mu_\star}\{\log g(Y, \sigma_0^2, \mu_\star)\})$$

or

$$\frac{-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (Y_i - \hat{\mu}_n)^2 + \frac{n-1}{2}}{\sqrt{\frac{n}{2}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

After simplification we have

$$-\frac{\frac{\hat{\sigma}^2}{2\sigma_0^2}-\frac{n-1}{2n}}{\sqrt{\frac{1}{2n}}}$$

which by symmetry property of normal density has the same result as the example 6.3 using Theorem 6.3.

In the next example without any information on the variance of the log-likelihood function we are able to compute the critical value of the test.

Example 6.5

Let $Y_1, Y_2, ..., Y_n$ i.i.d. according to the normal distribution $\mathcal{N}(\mu_\star, \sigma_\star^2)$. Suppose $\mu_\star = a\sigma_\star, \sigma_\star > 0$ and $a = \frac{\mu_\star}{\sigma_\star}$ is known. We wish to test that whether $Y_i \sim \mathcal{N}(a\sigma_\star, \sigma_\star^2)$. Then we have

$$g(Y_i;\sigma_\star) = (2\pi)^{-1/2}(\sigma_\star^2)^{-1/2}\exp\{-\frac{1}{2\sigma_\star^2}(Y_i - a\sigma_\star)^2\}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\sigma_\star) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma_\star^2 - \frac{1}{2n\sigma_\star^2}\sum_{i=1}^{n}(Y_i - a\sigma_\star)^2\}$$

In this case

$$\hat{\sigma}_n = \frac{1}{2}[-a\tilde{Y} + \sqrt{(a\tilde{Y})^2 + 4\tilde{Y^2}}]$$

and

$$\hat{\hat{\sigma}}_n = \frac{1}{2}[-a\tilde{Y} - \sqrt{(a\tilde{Y})^2 + 4\tilde{Y^2}}]$$

where $\tilde{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $\tilde{Y^2} = \frac{1}{n}\sum_{i=1}^{n} Y_i{}^2$. We have

$$(a\tilde{Y})^2 + 4\tilde{Y^2} \xrightarrow{\mathcal{P}} \sigma_\star^2(a^2+2)^2.$$

Then, $\hat{\hat{\sigma}}_n \xrightarrow{\mathcal{P}} \frac{1}{2}[-a^2\sigma_\star - \sqrt{\sigma_\star^2(a^2+2)^2}] < 0$, then this estimator is not consistent.

On the other hand $\hat{\sigma}_n \xrightarrow{P} \sigma_\star$ which is a consistent estimator for $\sigma_\star$. Here the Theorem 6.3 is not directly applicable, because,

$$\mathcal{E}_\sigma\{\log g(Y;\sigma_\star)\} = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma_\star^2 - \frac{1}{2}$$

which is depending on $\sigma_\star$.

On the other hand we can write

$$\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\sigma_\star) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma_\star^2 - \frac{\tilde{Y^2}}{2\sigma_\star^2} + \frac{a\tilde{Y}}{\sigma_\star} - \frac{a^2}{2}$$

and then

$$\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\sigma}_n) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\hat{\sigma}_n^2 - \frac{\tilde{Y^2}}{2\hat{\sigma}_n^2} + \frac{a\tilde{Y}}{\hat{\sigma}_n} - \frac{a^2}{2}$$

and

$$\mathcal{E}_{\hat{\sigma}_n}\{\log g(Y;\hat{\sigma}_\star)\} = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\hat{\sigma}_n^2 - \frac{1}{2}$$

We have,

$$\mathcal{V}ar_{\sigma_\star}\{\log g(Y;\sigma_\star)\} = \frac{1}{4}\mathcal{V}ar_{\sigma_\star}\{(\frac{Y-a\sigma_\star}{\sigma_\star})^2\} = \frac{1}{2}$$

Now by theorem 6.4,

$$\frac{\frac{1}{n}\sum_{i=1}^{n}\{\log g(Y_i;\hat{\sigma}_n) - \mathcal{E}_{\hat{\sigma}_n}\{\log g(Y;\hat{\sigma}_n)\}}{\sqrt{\frac{1}{2n}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

where

$$\frac{1}{n}\sum_{i=1}^{n}\{\log g(Y_i;\hat{\sigma}_n) - \mathcal{E}_{\hat{\sigma}_n}\{\log g(Y;\hat{\sigma}_n)\} = -\frac{\tilde{Y^2}}{2\hat{\sigma}_n^2} + \frac{a\tilde{Y}}{\hat{\sigma}_n} - \frac{a^2}{2} + \frac{1}{2}$$

The level of test is given by,

$$\mathcal{E}_{\mathcal{H}_0}(\phi(Y)) = P\{Z < \frac{C - \mathcal{E}_{\hat{\sigma}_n}\{\log g(Y;\hat{\sigma}_n)\}}{\sqrt{1/2n}}\}$$

Then

$$C = -Z_{\alpha/2}\sqrt{1/2n} - \frac{1}{2}\log 2\pi - \frac{1}{2}\log\hat{\sigma}_n^2 - \frac{1}{2}$$

Theorem 6.5 *Under Theorem 6.3*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n)-\mathcal{E}_{\beta_\star}(\log g(Y_i;\hat{\beta}_n))] \xrightarrow{\mathcal{L}} \mathcal{N}(0,\mathcal{V}ar_{\beta_\star}\{\log g(Y,\beta_\star)\}).$$

Proof : By Taylor expansion we have

$$\sum_{i=1}^{n}\log g(Y_i;\beta_\star)=\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)+(\hat{\beta}_n-\beta_\star)\sum_{i=1}^{n}\nabla\log g(Y_i,\hat{\beta}_n)+\frac{1}{2}(\hat{\beta}_n-\beta_\star)^2\sum_{i=1}^{n}\nabla^2\log g(Y_i,\hat{\beta}_n)+o_p(1)=$$

$$\sum_{i=1}^{n}\log g(Y_i;\beta_\star)=\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)+\frac{1}{2}(\hat{\beta}_n-\beta_\star)^2\sum_{i=1}^{n}\nabla^2\log g(Y_i,\hat{\beta}_n)+o_p(1)$$

it is known that under $\mathcal{H}_0$

$$\mathcal{V}ar_f\{\nabla\log g(Y,\beta)=\mathcal{E}_f\{(\nabla\log g(Y,\beta))^2\}\}=-\mathcal{E}_f\{\nabla^2\log g(Y,\beta)\}.$$

Now we replace the second sum in the right hand side by its population analogue which is

$$n\mathcal{V}ar_f\{\nabla\log g(Y,\beta)$$

then

$$\sum_{i=1}^{n}\log g(Y_i;\beta_\star)=\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)-\frac{1}{2}n(\hat{\beta}_n-\beta_\star)^2\mathcal{V}ar_f\{\nabla\log g(Y,\beta)\}+o_p(1)$$

$$\mathcal{E}_f\{\sum_{i=1}^{n}\log g(Y_i;\beta_\star)\}=\mathcal{E}_f\{\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)\}-\frac{1}{2}\mathcal{E}_f\{n(\hat{\beta}_n-\beta_\star)^2\mathcal{V}ar_f\{\nabla\log g(Y,\beta)\}\}+o_p(1)=$$

$$\mathcal{E}_f\{\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)\}-\frac{1}{2}+o_p(1) \quad (9)$$

Under $\mathcal{H}_0$ we have

$$\mathcal{E}_{\beta_\star}\{\sum_{i=1}^{n}\log g(Y_i;\beta_\star)\}=\mathcal{E}_{\beta_\star}\{\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)\}-\frac{1}{2}+o_p(1).$$

or

$$\mathcal{E}_{\beta_\star}\{\sum_{i=1}^{n}\log g(Y_i;\beta_\star)\}-\mathcal{E}_{\beta_\star}\{\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)\}=-\frac{1}{2}+o_p(1).$$

In Theorem 6.3 we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n)-\mathcal{E}_{\beta_\star}(\log g(Y_i;\beta_\star))] \xrightarrow{\mathcal{L}} \mathcal{N}(0,\mathcal{V}ar_{\beta_\star}\{\log g(Y,\beta_\star)\}).$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n)-\mathcal{E}_{\beta_\star}\{\log g(Y_i;\hat{\beta}_n)\}-\{\mathcal{E}_{\beta_\star}\{\log g(Y_i;\beta_\star)\}-\mathcal{E}_{\beta_\star}(\log g(Y_i;\hat{\beta}_n))\}] \xrightarrow{\mathcal{L}} \mathcal{N}(0,\mathcal{V}ar_{\beta_\star}\{\log g(Y,\beta_\star)\}).$$

or

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n)-\mathcal{E}_{\beta_\star}\{\log g(Y_i;\hat{\beta}_n)\}+\frac{1}{2n}+o_p(\frac{1}{n})] \xrightarrow{\mathcal{L}} \mathcal{N}(0,\mathcal{V}ar_{\beta_\star}\{\log g(Y,\beta_\star)\}).$$

Then the theorem holds ■

Example 6.6

Suppose $Y_1,Y_2,...,Y_n$ is an i.i.d. sample with unknown density $f(.)$. We want to test that $\mathcal{H}_0 : Y \sim \mathcal{N}(\mu_\star,\sigma_\star^2)$ where $\mu_\star$ and $\sigma_\star^2$ are unknown. The *MLE* of the parameters are given by $\hat{\mu}_n = \tilde{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\tilde{Y})^2$ respectively.

$$g(Y_i;\sigma_\star^2,\mu_\star) = (2\pi)^{-1/2}(\sigma_\star^2)^{-1/2}\exp\{-\frac{1}{2}(\frac{Y_i-\mu_\star}{\sigma_\star})^2\}$$

and

$$\log g(Y_i;\hat{\sigma}_n^2,\hat{\mu}_n) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\hat{\sigma}_n^2 - \frac{1}{2\hat{\sigma}_n^2}(Y_i-\hat{\mu}_n)^2$$

then

$$\sum_{i=1}^{n}\log g(Y_i;\hat{\sigma}_n^2,\hat{\mu}_n) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\hat{\sigma}_n^2 - \frac{n}{2}$$

now

$$\mathcal{E}_{\mu_\star,\sigma_\star^2}\{\sum_{i=1}^{n}\log g(Y_i;\hat{\sigma}_n^2,\hat{\mu}_n)\} = \mathcal{E}_{\mu_\star,\sigma_\star^2}\{-\frac{n}{2}\log 2\pi - \frac{n}{2}\log\hat{\sigma}_n^2 - \frac{n}{2}\} =$$

$$-\frac{n}{2}\log 2\pi - \frac{n}{2}\mathcal{E}_{\mu_\star,\sigma_\star^2}\{\log\hat{\sigma}_n^2\} - \frac{n}{2}.$$

But

$$\mathcal{E}_{\mu_\star,\sigma_\star^2}\{\log\hat{\sigma}_n^2\} = \log 2 + \Psi(\frac{n-1}{2}) + \log\frac{\sigma_\star^2}{n}$$

where Ψ is the digamma function,see, Hurvich and Tsai (1989).

$$\sum_{i=1}^{n}\log g(Y_i;\hat{\sigma}_n^2,\hat{\mu}_n) - \mathcal{E}_{\mu_\star,\sigma_\star^2}\{\sum_{i=1}^{n}\log g(Y_i;\hat{\sigma}_n^2,\hat{\mu}_n)\} = -\frac{n}{2}\log\hat{\sigma}_n^2 + \frac{n}{2}\mathcal{E}_{\mu_\star,\sigma_\star^2}\{\log\hat{\sigma}_n^2\} =$$

$$\frac{n}{2}\{\log\frac{2}{n} + \Psi(\frac{n-1}{2}) + \log\sigma_\star^2 - \log\hat{\sigma}_n^2\}.$$

by Theorem 6.5 we have

$$\frac{\frac{1}{\sqrt{n}}\{\frac{n}{2}\{\log\frac{2}{n} + \Psi(\frac{n-1}{2}) + \log\sigma_\star^2 - \log\hat{\sigma}_n^2\}\}}{\sqrt{\frac{n}{2}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

We can use this result to construct a confidence interval.

Example 6.7

Consider the linear model as $Y = X\beta_\star + \varepsilon$ as usual, where $\varepsilon \sim \mathcal{N}(0,\sigma_\star^2 I)$. We have:

$$\log\prod_{i=1}^{n} g(Y_i;\beta_\star,\sigma_\star^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma_\star^2 - \frac{1}{2\sigma_\star^2}(Y - X\beta_\star)^T(Y - X\beta_\star).$$

The *MLE* of the parameters are given by $\hat{\beta}_n = (X^TX)^{-1}X^TY$ and $\hat{\sigma}_n^2 = \frac{(Y-X\hat{\beta}_n)^T(Y-X\hat{\beta}_n)}{n}$ respectively.

And

$$\log\prod_{i=1}^{n} g(Y_i;\hat{\beta}_n,\hat{\sigma}_n^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\hat{\sigma}_n^2 - \frac{1}{2\hat{\sigma}_n^2}(Y - X\hat{\beta}_n)^T(Y - X\hat{\beta}_n).$$

Then

$$\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n,\hat{\sigma}_n^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\hat{\sigma}_n^2 - \frac{n}{2}$$

now

$$\mathcal{E}_{\beta_\star,\sigma_\star^2}\{\sum_{i=1}^{n}\log g(Y_i;\hat{\sigma}_n^2,\hat{\beta}_n)\} = \mathcal{E}_{\beta_\star,\sigma_\star^2}\{-\frac{n}{2}\log 2\pi - \frac{n}{2}\log\hat{\sigma}_n^2 - \frac{n}{2}\} =$$

$$-\frac{n}{2}\log 2\pi - \frac{n}{2}\mathcal{E}_{\beta_\star,\sigma_\star^2}\{\log\hat{\sigma}_n^2\} - \frac{n}{2}.$$

But

$$\mathcal{E}_{\beta_\star,\sigma^2_\star}\{\log\hat{\sigma}^2_n\} = \log 2 + \Psi(\frac{n-p}{2}) + \log\frac{\sigma^2_\star}{n}$$

where Ψ is the digamma function,see, Hurvich and Tsai (1989).

$$\sum_{i=1}^{n}\log g(Y_i;\hat{\sigma}^2_n,\hat{\beta}_n) - \mathcal{E}_{\beta_\star,\sigma^2}\{\sum_{i=1}^{n}\log g(Y_i;\hat{\sigma}^2_n,\hat{\beta}_n)\} = -\frac{n}{2}\log\hat{\sigma}^2_n + \frac{n}{2}\mathcal{E}_{\beta_\star,\sigma^2_\star}\{\log\hat{\sigma}^2_n\} =$$

$$\frac{n}{2}\{\log\frac{2}{n} + \Psi(\frac{n-p}{2}) + \log\sigma^2_\star - \log\hat{\sigma}^2_n\}.$$

by Theorem 6.5 we have

$$\frac{\frac{1}{\sqrt{n}}\{\frac{n}{2}\{\log\frac{2}{n} + \Psi(\frac{n-p}{2}) + \log\sigma^2_\star - \log\hat{\sigma}^2_n\}\}}{\sqrt{\frac{n}{2}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

We can use this result to construct a confidence interval.

$$1-\alpha = P(L < \sigma^2_\star < U) = p(\log L < \log\sigma^2_\star < \log U) =$$

$$p(\frac{\frac{1}{\sqrt{n}}\{\frac{n}{2}\{\log\frac{2}{n} + \Psi(\frac{n-p}{2}) + \log L - \log\hat{\sigma}^2_n\}\}}{\sqrt{\frac{n}{2}}} <$$

$$\frac{\frac{1}{\sqrt{n}}\{\frac{n}{2}\{\log\frac{2}{n} + \Psi(\frac{n-p}{2}) + \log\sigma^2_\star - \log\hat{\sigma}^2_n\}\}}{\sqrt{\frac{n}{2}}} < \frac{\frac{1}{\sqrt{n}}\{\frac{n}{2}\{\log\frac{2}{n} + \Psi(\frac{n-p}{2}) + \log U - \log\hat{\sigma}^2_n\}\}}{\sqrt{\frac{n}{2}}})$$

Then

$$\frac{\frac{1}{\sqrt{n}}\{\frac{n}{2}\{\log\frac{2}{n} + \Psi(\frac{n-p}{2}) + \log L - \log\hat{\sigma}^2_n\}\}}{\sqrt{\frac{n}{2}}} = -Z_{\frac{\alpha}{2}}$$

and

$$\frac{\frac{1}{\sqrt{n}}\{\frac{n}{2}\{\log\frac{2}{n} + \Psi(\frac{n-p}{2}) + \log U - \log\hat{\sigma}^2_n\}\}}{\sqrt{\frac{n}{2}}} = Z_{\frac{\alpha}{2}}$$

now we have

$$L = \exp\{-Z_{\frac{\alpha}{2}}\sqrt{2} - \log\frac{2}{n} + \Psi(\frac{n-p}{2}) + \log\hat{\sigma}^2_n\}$$

and

$$U = \exp\{Z_{\frac{\alpha}{2}}\sqrt{2} - \log\frac{2}{n} + \Psi(\frac{n-p}{2}) + \log\hat{\sigma}^2_n\}$$

Example 6.8

Consider the candidate linear model as $Y = X\beta + \varepsilon$ as usual, where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. Assume that the true model is $Y = X_t\beta_t + \varepsilon_t$ where $\varepsilon_t \sim \mathcal{N}(0, \sigma_t^2 I)$, X_t $(n \times p_t)$ is the design matrix and $(\beta_t^T, \sigma_t^2)^T$ is the parameter vector. Clearly

$$\log \prod_{i=1}^{n} g(Y_i; \beta, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)$$

Then

$$\sum_{i=1}^{n} \log g(Y_i; \hat{\beta}_n, \hat{\sigma}_n^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\hat{\sigma}_n^2 - \frac{1}{2}\frac{(Y - X\hat{\beta}_n)^T(Y - X\hat{\beta}_n)}{\hat{\sigma}_n^2}$$

where $\hat{\beta}_n = (X^T X)^{-1} X^T Y$ and $\hat{\sigma}_n^2 = \frac{(Y - X\hat{\beta}_n)^T(Y - X\hat{\beta}_n)}{n}$ are the *ML* estimator of the parameters in the candidate model. We notice that here Y has the covariance matrix as $\sigma_t^2 I$ and $\hat{\sigma}_n^2$ is an estimator for $\mathcal{V}ar\{\varepsilon\}$.

On the other hand

$$\mathcal{E}_f\{\sum_{i=1}^{n} \log g(Y_i; \hat{\beta}_n, \hat{\sigma}_n^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\mathcal{E}_f\{\log\hat{\sigma}_n^2\} - \frac{1}{2}\{\mathcal{E}_f\{\frac{(Y - X\hat{\beta}_n)^T(Y - X\hat{\beta}_n)}{\hat{\sigma}_n^2}\}\} =$$

$$\mathcal{E}_f\{\sum_{i=1}^{n} \log g(Y_i; \hat{\beta}_n, \hat{\sigma}_n^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\mathcal{E}_f\{\log\hat{\sigma}_n^2\} - \frac{1}{2}\{\mathcal{E}_f\{\frac{(Y - X\hat{\beta}_n)^T(Y - X\hat{\beta}_n)}{\sigma_t^2}\frac{\sigma_t^2}{\hat{\sigma}_n^2}\}\}.$$

We know that

$$\frac{n\hat{\sigma}_n^2}{\sigma_t^2} \sim \chi^2_{(n-p)}$$

then

$$\mathcal{E}_f\{\frac{n\sigma_t^2}{\hat{\sigma}_n^2}\} = \frac{n^2}{(n - p - 2)}.$$

Also

$$\mathcal{E}_f\{\frac{(Y - X\hat{\beta}_n)^T(Y - X\hat{\beta}_n)}{\hat{\sigma}_n^2}\} = \mathcal{E}_f\{\frac{(Y - X\hat{\beta}_n)^T(Y - X\hat{\beta}_n)}{\sigma_t^2}\frac{\sigma_t^2}{\hat{\sigma}_n^2}\}$$

but $\hat{\sigma}^2$ and $\hat{\beta}_n$ are independent, thus

$$\mathcal{E}_f\{\frac{(Y-X\hat{\beta}_n)^T(Y-X\hat{\beta}_n)}{\hat{\sigma}_n^2}\} = p\frac{n}{(n-p-2)} = \frac{np}{(n-p-2)}$$

and

$$\mathcal{E}_f\{\log\prod_{i=1}^{n} g(Y_i;\hat{\beta}_n,\hat{\sigma}^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\mathcal{E}_f\{\log\hat{\sigma}_n^2\} - \frac{1}{2}\frac{np}{(n-p-2)}.$$

Now

$$\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n,\hat{\sigma}^2) - \mathcal{E}_f\{\log\prod_{i=1}^{n} g(Y_i;\hat{\beta}_n,\hat{\sigma}^2) = -\frac{n}{2}\log\hat{\sigma}_n^2 - \frac{n}{2} + -\frac{n}{2}\mathcal{E}_f\{\log\hat{\sigma}_n^2\} + \frac{1}{2}\frac{np}{(n-p-2)}.$$

By Theorem 6.5 we have

$$\frac{\frac{1}{\sqrt{n}}\{-\frac{n}{2}\log\hat{\sigma}_n^2 - \frac{n}{2} + \frac{n}{2}\mathcal{E}_f\{\log\hat{\sigma}_n^2\} + \frac{1}{2}\frac{np}{(n-p-2)}\}}{\sqrt{\frac{1}{2}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

As we see in last example

$$\mathcal{E}_f\{\log\hat{\sigma}^2\} = \log 2 + \Psi(\frac{n-p}{2}) + \log\frac{\sigma^2}{n}$$

then

$$\frac{\frac{1}{\sqrt{n}}\{-\frac{n}{2}\log\hat{\sigma}_n^2 - \frac{n}{2} + \frac{n}{2}\{\log 2 + \Psi(\frac{n-p}{2}) + \log\frac{\sigma^2}{n}\} + \frac{1}{2}\frac{np}{(n-p-2)}\}}{\sqrt{\frac{1}{2}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

6.6 Distribution of T_n under $\mathcal{H}_1$

and Power of Test

6.6.1 Distribution of Test Statistic T_n under $\mathcal{H}_1$

For power computations we need to know the asymptotic distribution of $T_n(Y;\hat{\beta}_n) = \frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)$ under $\mathcal{H}_1$, where the generating density $f \notin \mathcal{G}$ as we saw in 3.5.1 in this case $T_n(Y;\hat{\beta}_n)$ does not

converge to $\mathcal{E}_f\{\log g(Y,\hat{\beta}_n)\}$. This alternative hypothesis is a "figurative alternative" because it is completely vague. Following sections 4.3 and 4.4.1 the maximum likelihood estimator in this case estimates the value β_0 that makes $g(Y;\beta_0)$ as close (in KL sense) to $f(.)$ as an $g(Y;\beta)$ can get. As we saw before

$$\hat{\beta}_n \xrightarrow{P} \beta_0, \quad \sqrt{n}(\hat{\beta}_n-\beta_0) \quad \text{is} \quad O_p(1) \quad \text{and} \quad 0=\nabla\mathcal{E}_f\{\log g(Y;\beta)\}_{|_{\beta=\beta_0}}=\mathcal{E}_f\{\frac{\nabla g(Y;\beta_0)}{g(Y;\beta_0)}\}$$

indeed (1) is right and we have the same theorems as theorems 6.2, 6.3, 6.4 at β_0 under $\mathcal{H}_1$.

Theorem 6.6 (Theorem 2') *Suppose that $Y_1,Y_2...,Y_n$ i.i.d with unknown density $f(.)$. Let $\mathcal{G}=\{g(.,\beta);\beta\in B\subseteq\mathcal{R}\}$ is a parametric family of assumed densities for $Y_i's$. If $\mathcal{H}_1$ holds, under conditions (C0)-(C4) and (1) we have:*

$$T_n(\overline{Y},\hat{\beta}_n)\xrightarrow{\mathcal{P}}\mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_0)\}=\mathcal{E}_f\{\log g(Y;\beta_0)\}.$$

Proof :

$$|\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)-\mathcal{E}_f(\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_0))|=$$

$$|\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)-\mathcal{E}_f(\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n))+\mathcal{E}_f(\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n))-\mathcal{E}_f(\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_0))|$$

$$\leq|\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)-\mathcal{E}_f(\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n))|+|\mathcal{E}_f(\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n))-\mathcal{E}_f(\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_0))|$$

$$\leq\sup_{\beta\in B}|\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta)-\mathcal{E}_f(\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta))|+|\mathcal{E}_f(\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n))-\mathcal{E}_f(\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_0))| \qquad (2')$$

By (C4) $|\log g(y;\beta)|\leq\vartheta(y)$ On the other hand by (1) we have

$$\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta)-\mathcal{E}_f(\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta))\xrightarrow{\mathcal{P}}0.$$

now under conditions (C0) and (C1) by Theorem *B*3 the first term in (2') converges to zero.

For second term in (2')

$$\mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)-\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_0)\}=[\frac{1}{n}\sum_{i=1}^{n}\mathcal{E}_f(\nabla\log g(Y_i;\beta_0))](\hat{\beta}_n-\beta_0)+o_p(\hat{\beta}_n-\beta_0)$$

by (C3) $\mathcal{E}_f(\nabla \log g(Y_i;\beta_0))$ exists and is equal to zero, this complete the proof. ■

Theorem 6.7 (Theorem 3') *Under conditions (C0)-(C3) suppose that* $0 < \mathcal{V}ar_f\{\log f(X,\theta^\star)\} < \infty$, *we have:*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n) - \mathcal{E}_f(\log g(Y_i;\beta_0))] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{V}ar_f\{\log g(Y,\beta_0)\}).$$

Proof :

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n) - \mathcal{E}_f(\log g(Y_i;\beta_0))] =$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n) - \log g(Y_i;\beta_0) + \log g(Y_i;\beta_0) - \mathcal{E}_f(\log g(Y_i;\beta_0))] =$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n) - \log g(Y_i;\beta_0)] + \frac{1}{\sqrt{n}}\sum_{1}^{n}[\log g(Y_i;\beta_0) - \mathcal{E}_f(\log g(Y_i;\beta_0)]$$

The second term on the right, by direct usage of the central limit theorem (CLT) asymptotically has a normal distribution with average and variance equal to zero and $\mathcal{V}ar_f\{\log g(Y,\beta_0)\}$ respectively.

For the first term, by Taylor's expansion we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n) - \log g(Y_i;\beta_0)] =$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[(\hat{\beta}_n - \beta_0)\nabla \log g(Y_i;\beta_0) + o_p(\hat{\beta}_n - \beta_0)] =$$

$$\sqrt{n}(\hat{\beta}_n - \beta_0)\frac{1}{n}\sum_{i=1}^{n}\nabla \log g(Y_i;\beta_0) + \sqrt{n}o_p(\hat{\beta}_n - \beta_0) \qquad (3')$$

we know that under regularity conditions (C0)-(C3)"

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I^{-1}(\beta_0)J(\beta_0)I^{-1}(\beta_0)). \qquad (4')$$

By WLLN

$$\frac{1}{n}\sum_{i=1}^{n}\nabla \log g(Y_i;\beta_0) \xrightarrow{\mathcal{P}} \mathcal{E}_f(\nabla \log g(Y;\beta_0)) = 0$$

from these two last convergences by Slutsky's theorem we have

$$\sqrt{n}(\hat{\beta}_n - \beta_0)\frac{1}{n}\sum_{i=1}^{n}\nabla \log g(Y_i;\beta_0) \xrightarrow{\mathcal{L}} 0$$

which also implies

$$\sqrt{n}(\hat{\beta}_n - \beta_0)\frac{1}{n}\sum_{i=1}^{n}\nabla \log g(Y_i;\beta_0) \xrightarrow{\mathcal{P}} 0. \qquad (5')$$

On the other hand

$$\sqrt{n}o_p(\hat{\beta}_n - \beta_0) = o_p(\sqrt{n}(\hat{\beta}_n - \beta_0))$$

by (4') $\sqrt{n}(\hat{\beta}_n - \beta_0) = O_p(1)$ then

$$\sqrt{n}o_p(\hat{\beta}_n - \beta_0) = o_p(O_p(1)) = o_p(1) \qquad (6')$$

by (5') and (6')

$$\sqrt{n}(\hat{\beta}_n - \beta_0)\frac{1}{n}\sum_{i=1}^{n}\nabla \log g(Y_i;\beta_0) + \sqrt{n}o_p(\hat{\beta}_n - \beta_0) \xrightarrow{\mathcal{P}} 0$$

Now applying the Slutsky's theorem. ■

Theorem 6.8 (Theorem 4') *:*

Under theorem $3'$ *we have*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n) - \mathcal{E}_f(\log g(Y_i;\hat{\beta}_n))] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{V}ar_f\{\log g(Y,\beta_0)\}).$$

Proof :

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n) - \mathcal{E}_{\hat{\beta}_n}(\log g(Y_i;\hat{\beta}_n))] =$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n) - \mathcal{E}_f(\log g(Y_i;\beta_0))] - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\mathcal{E}_{\hat{\beta}_n}(\log g(Y_i;\hat{\beta}_n)) - \mathcal{E}_f(\log g(Y_i;\beta_0))]$$

By theorem 3'

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n) - \mathcal{E}_f(\log g(Y_i;\beta_0))] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{V}ar_f\{\log g(Y,\beta_0)\}). \qquad (7')$$

and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\mathcal{E}_{\hat{\beta}_n}(\log g(Y_i;\hat{\beta}_n)) - \mathcal{E}_f(\log g(Y_i;\beta_0)] =$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[(\hat{\beta}_n-\beta_0)\mathcal{E}_f(\nabla \log g(Y_i,\beta_0)) + o_p(\hat{\beta}_n-\beta_0)] =$$

$$\sqrt{n}(\hat{\beta}_n-\beta_0)\frac{1}{n}\sum_{i=1}^{n}\mathcal{E}_f(\nabla \log g(Y_i,\beta_0)) + o_p(\sqrt{n}((\hat{\beta}_n-\beta_0)) \qquad (8')$$

the first term clearly is zero and the second term by (6') converges to zero, now by Slutsky's theorem for (7') and (8') the result holds. ■

6.6.2 Power of Test

Before talking about power we proof a useful lemma. If $d(x,y) = ||x-y|| = (\sum_{i=1}^{k}(x_i-y_i)^2)^{1/2}$ denotes the Euclidean distance function on $\mathcal{R}^k$, a sequence of the random variables Y_n is said to converge in probability to the Y if for every $\varepsilon' > 0$, $P_r(||Y_n - Y|| > \varepsilon') \to 0$, Van der Vaart (1998) and Lehmann (1998). This is denoted by $||Y_n - Y|| \xrightarrow{\mathcal{P}} 0$ or as before $Y_n \xrightarrow{\mathcal{P}} Y$.

Lemma 6.1 *Let* $U_n = \begin{pmatrix} X \\ Y_n \end{pmatrix}$ *and* $U_0 = \begin{pmatrix} X \\ Y_0 \end{pmatrix}$ *where* X *is a random variable and* $Y_n \xrightarrow{\mathcal{P}} Y_0$ *then for every* $\varepsilon > 0$

$$U_n \xrightarrow{\mathcal{P}} U_0$$

Proof : The convergence in probability is equivalent to individual convergence in probability of the vector elements. Trivially $X \xrightarrow{\mathcal{P}} X$ then

$$p\{||U_n' - U_0'|| > \varepsilon^{1/2}\} = p\{(X-X)^2 + (Y_n - Y_0)^2 > \varepsilon\} = p\{(Y_n - Y_0)^2 > \varepsilon\}$$

now by the fact that the quadratic function is continuous the right-hand side converges to zero by the continuous mapping theorem. Then

$$U_n \xrightarrow{\mathcal{P}} U_0. \qquad ■$$

Now by Lemma 1 and continuous mapping theorem $\log g(Y;\hat{\beta}_n) \xrightarrow{\mathcal{P}} \log g(Y;\beta_0)$ which implies that $\log g(Y;\hat{\beta}_n) \xrightarrow{\mathcal{L}} \log g(Y;\beta_0)$. In general we can split the random variable $\log g(Y;.)$ into its positive and negative parts which means that $\log g(Y;.) = \{\log g(Y;;.)\}^+ - \{\log g(Y;.)\}^-$ where $\{\log g(Y;.)\}^+ = max\{\log g(Y;.),0\}$, $\{\log g(Y;.)\}^- = max\{-\log g(Y;.),0\}$ and

$$\mathcal{E}_f\{\log g(Y;.)\} = \mathcal{E}_f\{\log g(Y;.)\}^+ - \mathcal{E}_f\{\log g(Y;.)\}^-.$$

Assume that $\log g(Y;\hat{\beta}_n)$ is non-negative for all n,

$$|\mathcal{E}_f\{\log g(Y;\hat{\beta}_n)\} - \mathcal{E}_f\{\log g(Y;\beta_0)\}| =$$

$$|\int_0^\infty \{P(\log g(Y;\hat{\beta}_n)) > y - P(\log g(Y;\beta_0)) > y\}\, dy| =$$

$$|\int_0^\vartheta \{P(\log g(Y;\hat{\beta}_n)) > y - P(\log g(Y;\beta_0)) > y\}\, dy| \leq$$

$$\int_0^\vartheta |P\{(\log g(Y;\hat{\beta}_n)) > y\} - P\{(\log g(Y;\beta_0)) > y\}|\, dy \to 0$$

because the interval of integration is bounded. Then by convergence in mean

$$\mathcal{E}_f\{\log g(Y;\hat{\beta}_n)\} \to \mathcal{E}_f\{\log g(Y;\beta_0)\}.$$

for general result we can consider the positive and negative part of $\log g(Y;\hat{\beta}_n)$. On the other hand by continuous mapping theorem $\{\log g(Y;\hat{\beta}_n)\}^2 \xrightarrow{\mathcal{L}} \{\log g(Y;\beta_0)\}^2$. And again

$$\mathcal{E}_f\{(\log g(Y;\hat{\beta}_n))^2\} \to \mathcal{E}_f\{(\log g(Y;\beta_0))^2\}$$

now

$$\mathcal{V}ar_f\{\log g(Y;\hat{\beta}_n)\} \to \mathcal{V}ar_f\{\log g(Y_i;\beta_0)\}.$$

thus we can approximate

$$\mathcal{V}ar_f\{\frac{1}{n}\sum_{i=1}^{n} \log g(Y_i;\beta_0)\}$$

by

$$\mathcal{V}ar_f\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)\}.$$

By the Portmanteau theorem for identity function

$$\mathcal{E}_{\mathcal{H}_1}\{\log g(Y;\hat{\beta}_n)\} \to \mathcal{E}_{\mathcal{H}_1}\{\log g(Y;\beta_0)\}$$

where expectation under $\mathcal{H}_1$ indicates that expectation is taken under the density which is specified in $\mathcal{H}_1$. On the other hand by the continuous mapping theorem $\{\log g(Y;\hat{\beta}_n)\}^2 \xrightarrow{\mathcal{L}} \{\log g(Y;\beta_0)\}^2$. and again

$$\mathcal{V}ar_{\mathcal{H}_1}\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)\} \to \mathcal{V}ar_{\mathcal{H}_1}\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_0)\}.$$

Now by (C4) the uniform integrability of $\log g(g;\beta)$ we have

$$\mathcal{E}_{\mathcal{H}_1}\{\log g(Y;\hat{\beta}_n)\} \xrightarrow{\mathcal{P}} \mathcal{E}_{\mathcal{H}_1}\{\log g(Y;\beta_0)\}$$

and

$$\mathcal{V}ar_{\mathcal{H}_1}\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)\} \xrightarrow{\mathcal{P}} \mathcal{V}ar_{\mathcal{H}_1}\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_0)\}.$$

Then the asymptotic density of interest could be changed to asymptotic density of

$$\frac{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n) - \mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_\star)\}}{\sqrt{\mathcal{V}ar_f\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_\star)\}}}$$

which is more realistic in theory.

The power of the test for level α_n is defined by

$$\gamma_n = \mathcal{E}_f(\phi(\underline{Y})) = P_f(T_n(\overline{Y},\hat{\beta}_n) < K_n).$$

Now

$$\gamma_n = P\left(\frac{1/n\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n) - \mathcal{E}_f(\log g(Y_i;\beta_0))]}{\sqrt{\mathcal{V}ar_f(\log g(Y;\beta_0))/n}} < \frac{K_n - \mathcal{E}_f(\log g(Y;\beta_0))}{\sqrt{\mathcal{V}ar_f(\log g(Y;\beta_0))/n}}\right) =$$

$$P\left(\frac{\sqrt{n}(1/n\sum_{i=1}^{n}[\log g(Y_i;\hat{\beta}_n)-\mathcal{E}_f(\log g(Y_i;\beta_0))]}{\sqrt{\mathcal{V}ar_f(\log g(Y;\beta_0))}} < WZ_{\alpha_n}+\frac{\sqrt{n}\{\mathcal{E}_{\beta_\star}(\log g(Y;\beta_\star))-\mathcal{E}_f(\log g(Y;\beta_0))\}}{\sqrt{\mathcal{V}ar_f(\log g(Y;\beta_0))}}\right) \quad (9)$$

where $W=\sqrt{\frac{\mathcal{V}ar_{\beta_\star}(\log g(Y;\beta_\star))}{\mathcal{V}ar_f(\log g(Y;\beta_0))}}$, then

$$\gamma_n=\Phi\left\{WZ_{\alpha_n}+\frac{\sqrt{n}\{\mathcal{E}_{\beta_\star}(\log g(Y;\beta_\star))-\mathcal{E}_f(\log g(Y;\beta_0))\}}{\sqrt{\mathcal{V}ar_f(\log g(Y;\beta_0))}}\right\}.$$

From the last equality we can see that the power of the test mainly depends on the difference in expectations of the log- density functions under $\mathcal{H}_0$ and $\mathcal{H}_1$.

It seems that in practice we need to compute the power function as below

$$\gamma_n=\Phi\left\{WZ_{\alpha_n}+\frac{\sqrt{n}\{\mathcal{E}_{\beta_\star}(\log g(Y;\beta_\star))-\mathcal{E}_f(\log g(Y;\hat{\beta}_n))\}}{\sqrt{\mathcal{V}ar_f(\log g(Y;\hat{\beta}_n))}}\right\}.$$

or

$$\gamma_n=\Phi\left\{WZ_{\alpha_n}+\frac{\sqrt{n}\{\mathcal{E}_{\hat{\beta}_n}(\log g(Y;\hat{\beta}_n))-\mathcal{E}_f(\log g(Y;\hat{\beta}_n))\}}{\sqrt{\mathcal{V}ar_f(\log g(Y;\hat{\beta}_n))}}\right\}.$$

6.7 Consistency of Test

By the definition of *MLE* we know that

$$\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)\geq\sup_{\beta\in B}\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta)-o_p(1)$$

then

$$\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)\geq\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_0)-o_p(1)\rightarrow\mathcal{E}_f\{\log g(Y;\beta_0)\}-o_p(1)$$

and

$$\mathcal{E}_f\{\log g(Y;\beta_0)\}-\mathcal{E}_{\beta_\star}\{\log g(Y;\beta_\star)\}\leq\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n)-\mathcal{E}_{\beta_\star}\{\log g(Y;\beta_\star)\}+o_p(1)\leq$$

$$\sup_{\beta\in B}\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta)-\mathcal{E}_\beta\{\log g(Y;\beta)\}\}+o_p(1)$$

The right side of the last inequality under conditions (C0) and (C1) and (1) by Theorem $B3$ converges to zero. By this we conclude that

$$\mathcal{E}_{\beta_\star}\{\log g(Y;\beta_\star)\} - \mathcal{E}_f\{\log g(Y;\beta_0)\} > 0$$

Then the right side of (9) goes to ∞ when $n \to \infty$ and its left side converges in distribution; it follows that $\gamma_n \to 1$ as $n \to \infty$. Thus this test is consistent. On the other hand from the power function and the last inequality we see that when the difference gets large which means that when the hypothesized density under $\mathcal{H}_1$ is far from the hypothesized density under $\mathcal{H}_0$ in expectation the power function naturally gets large.

6.7.1 Power computation

For power computation we need the bootstrap estimation of $\mathcal{E}_{\beta_\star}\{\log g(Y;\beta_\star)\}$, $\mathcal{E}_f(\log g(Y;\beta_0))\}$, $\mathcal{V}ar_f\{\log g(Y;\beta_0))\}$, $\mathcal{V}ar_{\beta_\star}\{\log g(Y;\beta_\star))\}$ and also $\mathcal{V}ar_f\{\log g(Y;\beta_0))\}$.

Algorithm 1 Bootstrap estimation of $\mathcal{E}_{\beta_\star}\{\log g(Y;\beta_\star)\}$ and $\mathcal{V}ar_{\beta_\star}\{\log g(Y;\beta_\star))\}$.

Select the p.d.f. $g(.,\beta) \in \mathcal{G}$, and the sample size n.

Estimate β by the maximum likelihood approach, say, $\hat{\beta}_n$

Generate a sequence of b random sample $Y_1^{(j)}, Y_2^{(j)}, ..., Y_n^{(j)}, j = 1, 2, ..., b$ from a distribution with p.d.f. $g(y;\hat{\beta}_n)$

Estimate $\hat{\beta}_n^{(j)Boot}$ by maximizing $\sum_{i=1}^n \log g(y_i^{(j)}, \hat{\beta}_n)$, $j = 1, 2, ..., b$

Compute

$$\mathcal{E}_{\beta_\star}\{\log g(Y;\beta_\star)\} \simeq \mathcal{E}_{Boot}\{\log g(Y^{(j)};\hat{\beta}(j)Boot)\} = \frac{1}{b}\sum_{j=1}^{b}\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i^{(j)};\hat{\beta}(j)Boot_n)$$

and

$$\mathcal{V}ar_{\beta_\star}\{\log g(Y;\beta_\star))\} \simeq \frac{1}{b-1}\sum_{j=1}^{b}[\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i^{(j)};\hat{\beta}_n^{(j)Boot}) - \frac{1}{B}\sum_{j=1}^{b}\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i^{(j)};\hat{\beta}_n^{(j)Boot})]^2$$

Algorithm 2 Bootstrap estimation of $\mathcal{E}_f\{\log g(Y;\beta_0)\}$ and $\mathcal{V}ar_f\{\log g(Y;\beta_0))\}$.

Generate a sequence of b random sample $Y_1^{(j*)}, Y_2^{(j*)}, ..., Y_n^{(j*)}$, $j = 1, 2, ..., b$ from original sample.

Estimate $\hat{\beta}_n^{(j*)B}$ by maximizing $\sum_{i=1}^{n}\log g(y_i^{(j*)},\beta)$

Compute

$$\mathcal{E}_f\{\log g(Y;\beta_0)\} \simeq \mathcal{E}_{Boot}\{\log g(Y^{(j*)};\beta^{(j*)Boot})\} = \frac{1}{b}\sum_{j=1}^{b}\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i^{(j*)};\hat{\beta}_n^{(j*)Boot})$$

and

$$\mathcal{V}ar_f\{\log g(Y;\beta_0))\} \simeq \frac{1}{b-1}\sum_{j=1}^{b}[\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i^{(j*)};\hat{\beta}_n^{(j*)Boot}) - \frac{1}{b}\sum_{j=1}^{b}\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i^{(j*)};\hat{\beta}_n^{(j*)Boot})]^2$$

6.7.2 Invariance

Our test statistic in general is not invariant under transformation because a single distribution is being used to compute this gof test statistic. But it is not necessarily a defect. Here we consider an example for our test statistic $T_n(Y;\hat{\beta}_n)$ and then verify the general case.

Example 6.9

Suppose $Y_1, Y_2, ..., Y_n$ is an i.i.d. sample with exponential density, $Y_i \sim \exp(\lambda_\star)$ then

$$g(y;\lambda_\star) = {\lambda_\star}^{-1}e^{-{\lambda_\star}^{-1}y}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\lambda_\star) = -\log\lambda_\star - \frac{\sum_{i=1}^{n}Y_i}{n\lambda_\star}$$

The *MLE* for $\lambda_\star$ is $\hat{\lambda}_\star = \tilde{Y}$ then

$$\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\lambda}_\star) - \mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\lambda_\star) = \log\frac{\lambda_\star}{\hat{\lambda}_\star}$$

We consider a transformation as $Y = T^2$. Then $J = |2T|$ and

$$w(t;\lambda_\star) = {\lambda_\star}^{-1}e^{-{\lambda_\star}^{-1}t^2}2t$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\log w(T_i;\hat{\lambda}_T) = -\log\hat{\lambda}_T - \frac{\sum_{i=1}^{n}T_i^2}{n\hat{\lambda}_T} + \frac{1}{n}\sum_{i=1}^{n}\log 2T_i \qquad \text{where} \quad \hat{\lambda}_T = \frac{1}{n}\sum_{i=1}^{n}T_i^2$$

this shows that our test statistic is not invariant.

$$\frac{1}{n}\sum_{i=1}^{n}\log w(T_i;\hat{\lambda}_T) - \mathcal{E}_w\{\frac{1}{n}\sum_{i=1}^{n}\log w(T_i;\lambda_\star)\} = \log\frac{\lambda_\star}{\hat{\lambda}_T} + \frac{1}{n}\sum_{i=1}^{n}(\log 2T_i - \mathcal{E}_w(\log 2T_i))$$

when n gets large the second term on the right is negligible by WLLN. We conclude that

$$\frac{1}{n}\sum_{i=1}^{n}\log f(Y_i;\hat{\lambda}_\star) - \mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\lambda_\star)\} \simeq \frac{1}{n}\sum_{i=1}^{n}\log w(T_i;\hat{\lambda}_T) - \mathcal{E}_w\{\frac{1}{n}\sum_{i=1}^{n}\log w(T_i;\lambda_\star)\}$$

Then the term $T_n(\overline{Y},\hat{\beta}_n) - \mathcal{E}_w\{\frac{1}{n}\sum_{i=1}^{n}\log g(T_i;\lambda_\star)\}$ is asymptotically invariant under the one to one transformation.

Now if $Y_1, Y_2, ...Y_n$ be i.i.d. with common density $f(.)$ and assumed density $g(.;\beta)$ and $W_i = k_i(Y_1,...,Y_n), i = 1,2,...,n$ the joint density of $W_1, W_2, ..., W_n$ is given by $g(k^{-1};\beta)|\det(J)|$ where J is

Jacobian matrix of

$$H(Y_1,...,Y_n) = \begin{pmatrix} k_1(Y_1,...,Y_n) \\ k_2(Y_1,...,Y_n) \\ . \\ . \\ . \\ k_n(Y_1,...,Y_n) \end{pmatrix}$$

where k_i's are real-valued continuous functions If H is a one-to-one function with inverse H^{-1} the Jacobian matrix of H is $n \times n$ matrix $\left(\left(\frac{\partial k_i^{-1}(w_1,...,w_n)}{\partial w_j} \right)_{n\times n} = \left(\frac{\partial k_i(y_1,...,y_n)}{\partial y_j} \right)_{n\times n}^{-1} \right.$ and Jacobian of H defined to be the determinant of this matrix. Then log-likelihood function for $W_1,...,W_n$ is given by

$$h(\overline{W};\beta) = h(H^{-1}(\overline{W});\beta)|\det(J)|$$

and

$$\frac{1}{n}h(\overline{W};\beta) = \frac{1}{n}\log h(W_1,...,W_n;\beta) = \frac{1}{n}\log h(k_1^{-1}(W_1,...,W_n),...,k_n^{-1}(W_1,...,W_n);\beta) + \frac{1}{n}\log|\det(J)|$$

$$= \frac{1}{n}\log h(W_1,...,W_n;\beta) + \frac{1}{n}\log|\det(J)|$$

J is not depending on β then

$$\sup_{\beta\in B}\frac{1}{n}\log h(W_1,...,W_n;\beta) = \sup_{\beta\in B}\frac{1}{n}\log g(Y_1,...,Y_n;\beta)$$

and

$$\frac{1}{n}\log h(W_1,...,W_n;\bar{\beta}_n) = \frac{1}{n}\log g(k_1^{-1}(W_1,...,W_n),...,k_n^{-1}(W_1,...,W_n);\hat{\beta}_n) + \frac{1}{n}\log|J|$$

$$= \frac{1}{n}\log g(Y_1,...,Y_n;\hat{\beta}_n) + \frac{1}{n}\log|\det(J)|$$

where $\bar{\beta}_n = t(\hat{\beta}_n)$ is *MLE* w.r.t. transformed data.

$$\mathcal{E}_h\{\log h(W_1,...,W_n;\beta)\} = \mathcal{E}_h\{\log g(Y_1,...,Y_n;\beta) + \log|\det(J)|\} =$$

$$\int\int ...\int \{\sum_{i=1}^{n} \log g(Y_i;\beta)\} |\det(J)| \prod_{i=1}^{n} g(y_i;\beta) dy_1 ... dy_n + \mathcal{E}_h\{\log|\det(J)|\}$$

now

$$\frac{1}{n}\log h(W_1,...,W_n;\bar{\beta}_n) - \frac{1}{n}\mathcal{E}_h\{\log h(W_1,...,W_n;\beta)\} =$$

$$\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n) - \int\int ...\int \{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta)\}|J|\prod_{i=1}^{n} g(y_i;\beta)dy_1...dy_n + \frac{1}{n}\log|\det(J)| - \frac{1}{n}\mathcal{E}_h\{\log|\det(J)|\}.$$

For a linear transformation like $Y_j = \sum_{i=1}^{n} a_{ij}X_i, \quad j = 1,...,n$, then the Jacobian of the transformation is the determinant of the matrix $\left(a_{ij}\right)$ and two last terms in the right of equality vanish. If this transformation is also orthogonal we have $|\det(J)| = 1$ then for $\beta_\star$

$$\frac{1}{n}\log h(W_1,...,W_n;\bar{\beta}_n) - \frac{1}{n}\mathcal{E}_h\{\log h(W_1,...,W_n;\beta_\star)\} = \frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n) - \mathcal{E}_{\beta_\star}\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_\star)\}$$

then

$$\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\hat{\beta}_n) - \mathcal{E}_{\beta_\star}\{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i;\beta_\star)\}$$

is invariant under orthogonal linear transformation.

Chapter 7

Test For Model Selection based on difference of AIC's: application to tracking interval for $\Delta \mathcal{E}KL$

7.1 Introduction

Usually a statistical process is to drive a model from theory and then use statistical methods to estimate its parameter(s). In regression models for instance, the goal is to determine whether or not an "independent" variable or a set of "independent" variables, has a statistically significant effect

upon a dependent variable. In this circumstance the questions which arise are what are the methods and how they work for model selection. The analysis of models has followed two approaches in the literature; the hypothesis testing and the model selection criteria. If we find by a method, a positive effect of independent variable(s) on a dependent variable, we will confirm our model. Sometimes we choosee a model which is at least not falsified. Clearly this approach is different of classical hypothesis testing. Two models may be nested or non-nested, and in the latter case they may overlap or not. The nested models are frequently studied in both theoretical and applied statistics. On the other hand the non-nested models are less discussed. Historically a serious studies on non-nested models can be found in a period from Cox (1961), Cox (1962) to Vuong (1989). In search of similarities and differences between Cox's test and Vuong's test we may say that the Vuong's test is a development of the Cox test. As a classical usage of these two tests the Cox's test is a test about non-nested hypothesis where the emphasis of Vuong's test is on non-nested models. Both tests are a generalization of the likelihood ratio tests (LRT) under different sense. In Cox's test the difference between the log-likelihood ratio and its expected value under the null hypothesis is considered. The Cox's test says that a true model must be able to predict the performance of the specific alternatives , i.e. a true null should not distort the actual performance of the alternative model. The idea is to compare the true performance of the alternative model with the expected performance of the alternative model under the null hypothesis. We may make any decision about two competing models. The important points is that when we reject a hypothesis, there is no means that it is rejected in favor of the specific alternative. For example the rejection of both models implies that neither model could predict the results of the other model. Then we conclude that both models are mis-specified. May be a solution to this difficulty is to use a model selection approach which chooses the model which is closest to the true model. We must notice that the other difficulty with

Cox's test is calculating the expected value of the log-likelihood ratio under the null hypothesis. Another candidate in a similar situation is Vuong's test. In Vuong viewpoint, the best model is the model which maximizes the relevant part of KL risk. The null hypothesis of Vuong's test is the expectation under the true model of the log-likelihood ratio of the two candidate models are equal to zero, which means that two candidate models are equivalent. This expectation however is unknown. But Vuong's test works, because the decision making procedure by Vuong's test does not depend on this unknown quantity.

7.2 Objective

The problem of model selection by model selection criterion is that it produces a deterministic outcome, defined by the ranking of the values of the criterion, and it does not take account the probabilistic nature of the result. On the other hand the differences in the criterion values may not be statistically sufficient because the deterministic model selection criterion approach would consider a model better than another model while in fact they may be considered as statistically equivalent. This is a reason why Vuong (1989) considered a probabilistic framework. On the other hand the log-likelihoods used in the Vuong's test are affected if the number of coefficients in the two models is different and therefore the test must be corrected for the degrees of freedom. For a relative solution to these two problems we focus on interval estimation for normalized difference of a model selection criteria of two competing models as the dual of the hypotheses testing problem when the models are non-nested. Our attention is on Akaike Information Criterion (AIC), see 4.5.1, and expected Kullback-Leibler $(\mathcal{E}KL)$, see 4.4.1. The Akaike criterion is often used as the measure of model accuracy. In fact this statistic considers the lack of fit measure and the parsimony as a principal of

model selection. From decision theory we realize that comparison could be based on some function of the likelihood ratio of nested or non-nested models. Thus it makes sense that we consider some condition on this kind of function. A possible function of this ratio could be the log function, the expectation of this function define the especial loss function which is known as the Kullback-Leibler. It was Akaike (1973) who introduced the expected value over the data of the Kullback-Leibler loss as the risk function on which model selection can be made. In searching for the estimator for this risk we notice that the difference of *AIC*'s for two models detect the changes when we must choice the best model. It is noticeable that the normalized difference of *AIC*'s is an estimator for the difference of $\mathcal{E}KL$'s for two models. By these means we want to construct a confidence interval about the expected Kullback-Leibler risks difference, where the expectation is taken under the unknown true density. After tracking this interval we are in a decision making situation. If this interval contains zero, we will conclude that the two models are equivalent in Kullback-Leibler sense related to the true density.

In this chapter we will bring some necessary definition and by two theorems and corollary we will argue that we can achieve a Vuong-like test under other considerations which are useful for tracking an interval. A simulation study shows that the confidence interval has a good interpretation in model selection where the models are logistic models in regression context. We use this approach of model selection for real data to verify the relation between body-mass index and depression in elderly people; see appendix *B.*

7.3 Non-Nested Models comparison

Many models comparisons are performed among models that are not nested. In the literature a method for comparing the non-nested regression models come back to Hotelling (1940), Kendall and Stuart (1967) and Pesaran (1978). Consider two families of parametric densities as $\mathcal{G} = (g^{\beta}(.))_{\beta\in B} = \{g(y;\beta);\beta\in B)\}$, $\mathcal{K} = (h^{\gamma}(.))_{\gamma\in\Gamma} = \{h(y;\gamma);\gamma\in\Gamma\}$ and an i.i.d. random sample from the true density $f(.)$.

Definition 7.1 (Non-Nested models) *Two models $\mathcal{G}$ and $\mathcal{K}$ are strictly non-nested iff $\mathcal{G}\cap\mathcal{K} = \varnothing$.*

This definition may be generalized by Kullback-Leibler divergence term between two models. Following 4.3 and in mis-specified case we set

$$\beta_0 = argmax_{\beta}\mathcal{E}_f\{\log g(Y;\beta)\} \quad \text{and} \quad \gamma_0 = argmax_{\gamma}\mathcal{E}_f\{\log h(Y;\gamma)\}$$

such that if $f\notin\mathcal{G}$, $\hat{\beta}_n \xrightarrow{\mathcal{P}} \beta_0$ and if $f\notin\mathcal{K}$, $\hat{\gamma}_n \xrightarrow{\mathcal{P}} \gamma_0$ where $\hat{\beta}_n$ and $\hat{\gamma}_n$ are their maximum likelihood estimators under $g(.,.)$ and $h(.,.)$ respectively, and

$$\beta_\star = argmax_{\beta}\mathcal{E}_g\{\log g(Y;\beta)\} \quad \text{and} \quad \gamma_\star = argmax_{\gamma}\mathcal{E}_h\{\log h(Y;\gamma)\}$$

which are the true values of β and γ under $\mathcal{G}$ and $\mathcal{K}$ (when one at time they are the correctly specified models) respectively. As we saw if the true density $f(.)$ belongs to the $\mathcal{G}$ the *MLE* of β converges to $\beta_\star$ and if the true density $f(.)$ belongs to the $\mathcal{K}$ the *MLE* of γ converges to $\gamma_\star$. Define

$$\beta_{0h} = argmax_{\beta}\mathcal{E}_h\{\log g(Y;\beta)\} \quad \text{and} \quad \gamma_{0g} = argmax_{\gamma}\mathcal{E}_g\{\log h(Y;\gamma)\}$$

If $h(.,.)$ be the true density $\beta_{0h} = \beta_0(\gamma_\star)$ and if $g(.,.)$ is the true density $\gamma_{0g} = \gamma_0(\beta_\star)$.

Definition 7.2 (Non-Nested models in KL sense) *We say two models are non-nested if and only if*

$$KL\{h(Y,\gamma_\star(\beta_0));g(Y,\beta_0)\}\neq 0 \quad and \quad KL\{g(Y,\beta_\star(\gamma_0));h(Y,\gamma_0)\}\neq 0. \quad \forall\beta_\star\in B \quad and \quad \forall\gamma_\star\in\Gamma$$

The KL distance from the true density $f(.)$ for densities $g(.,\beta_0)$ and $h(.,\gamma_0)$ are given by

$$KL\{g(.,\beta_0);f(.)\}=\mathcal{E}_f\{\log f(Y)\}-\mathcal{E}_f\{\log g(Y,\beta_0)\}$$

and

$$KL\{h(.,\gamma_0),f(.)\}=\mathcal{E}_f\{\log\{f(Y)\}-\mathcal{E}_f\{\log h(Y,\gamma_0)\}.$$

Since the first term in both of KL(.,.)'s is unknown the KL(.,.) can not be estimated directly, but it can be noticed that when two models are compared, the first term of KL(.,.) remains constant, so that minimization of the criterion only depends on the second terms. To compare these two models we notice that $KL\{g(.,\beta_0);f(.)\}=KL\{h(.,\gamma_0);f(.)\}$ if and only

$$\mathcal{E}_f\{\log g(Y;\beta_0)\}=\mathcal{E}_f\{\log h(Y;\gamma_0)\}.$$

Then two models are equally close in KL sense to the true density $f(.)$ if the last equality is true. This lead us to model selection criterion in a hypothesis testing framework, see, Vuong (1989). The null hypothesis is given by;

$$\mathcal{H}_0:\mathcal{E}_f\{\log\frac{g(Y;\beta_0)}{h(Y;\gamma_0)}\}=0$$

which meaning that two models are equivalent. The alternatives could be

$$\mathcal{H}_g=\mathcal{E}_f\{\log\frac{g(Y;\beta_0)}{h(Y;\gamma_0)}\}>0$$

or

$$\mathcal{H}_h=\mathcal{E}_f\{\log\frac{g(Y;\beta_0)}{h(Y;\gamma_0)}\}<0$$

The first alternative hypothesis meaning that $\mathcal{G}$ is better then $\mathcal{K}$ and the second alternative hypothesis meaning that $\mathcal{K}$ is better then $\mathcal{G}$. Consequently when we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_g$, we say, $\mathcal{G}$ is less misspecified than $\mathcal{K}$ and when we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_h$, we say, $\mathcal{K}$ is less misspecified than $\mathcal{G}$.

7.3.1 Motivation to Confidence Interval construction

In model selection context, selection the null hypothesis is not easy and on the other hand we faced with many alternatives and sometimes with infinite number of alternatives. Generally in hypothesis testing when we decide about null hypothesis we do not add more and more alternative hypothesis, in fact in hypothesis testing we select the one best alternative to compare against. Confidence intervals are equivalent to encapsulating the results of many hypothesis tests. They explicitly show the region where we are likely to find the true answer. In this section we want to show how we can construct a pivot to building a confidence interval for difference of expected Kullbak-Leibler risks for two models related to the true density. We do it in two parts. In the first part we consider the statistic T_n as in chapter 6 but here for standardized ratio of two non-nested models g and h,say, S_n. Taylor expansion of numerator of S_n guides us to Vuong's theorem (1989). In a second part our focus is on regression context in the spirit of conditional Kullback-Leibler criterion for reduced models, (for reduced models, see, Commenges et al. (2007)). In Theorem 7.1 and in the spirit of Vuong's theorem we find the asymptotic distribution of a statistic which is a little different from S_n by considering the expected Kullback-Leibler in a regression context instead of simple average of T_n in numerator of S_n. In Theorem 7.2 using Theorem 7.1, we find the asymptotic distribution of a difference of normalized AIC criterion, say, $D_n(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n})$. The last result is a basis to construct a confidence interval for difference of expected Kullback-Leibler of two models related to true density, say, $\Delta_n(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n})$,

which helps us to choose between g and h.

Under Theorem 6.6 we have

$$\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i,\hat{\beta}_n)\xrightarrow{\mathcal{P}}\mathcal{E}_f\{\log g(Y,\beta_0)\}$$

and similarly

$$\frac{1}{n}\sum_{i=1}^{n}\log h(Y_i,\hat{\gamma}_n)\xrightarrow{\mathcal{P}}\mathcal{E}_f\{\log h(Y,\gamma_0)\}$$

then we expect that

$$\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i,\hat{\beta}_n)-\frac{1}{n}\sum_{i=1}^{n}\log h(Y_i,\hat{\gamma}_n)=\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i,\hat{\beta}_n)}{h(Y_i,\hat{\gamma}_n)}\}\xrightarrow{\mathcal{P}}\mathcal{E}_f\{\log\frac{g(Y,\beta_0)}{h(Y,\gamma_0)}\}.$$

By this result we chois the left-hand side of the last relation as the test statistic. To do a test we need to know the distribution of this test statistic. As the classical approach we consider

$$S_n=\frac{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i,\hat{\beta}_n)}{h(Y_i,\hat{\gamma}_n)}\}-\mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i,\hat{\beta}_n)}{h(Y_i,\hat{\gamma}_n)}\}\}}{\sqrt{\mathcal{V}ar_f\{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i,\hat{\beta}_n)}{h(Y_i,\hat{\gamma}_n)}\}\}}}$$

or

$$S_n=\frac{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i,\hat{\beta}_n)}{h(Y_i,\hat{\gamma}_n)}\}-\left[\mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i,\beta_0)}{h(Y_i,\gamma_0)}\}\}-\mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i,\hat{\beta}_n)}{h(Y_i,\hat{\gamma}_n)}\}\}-\mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i,\beta_0)}{h(Y_i,\gamma_0)}\}\}\right]}{\sqrt{\mathcal{V}ar_f\{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i,\hat{\beta}_n)}{h(Y_i,\hat{\gamma}_n)}\}\}}}.$$

By Taylor expansion the last two terms in the numerator are negligible because

$$\mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i,\hat{\beta}_n)}{h(Y_i,\hat{\gamma}_n)}\}-\mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i,\beta_0)}{h(Y_i,\gamma_0)}\}\}=$$

$$(\hat{\beta}_n-\beta_0)^T\mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n}\nabla\log g(Y_i,\beta_0)\}-(\hat{\gamma}_n-\gamma_0)^T\mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n}\nabla\log h(Y_i,\gamma_0)\}+o_p(1)\xrightarrow{\mathcal{P}}0.$$

Thus

$$S_n=\frac{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i,\hat{\beta}_n)}{h(Y_i,\hat{\gamma}_n)}\}-\mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i,\beta_0)}{h(Y_i,\gamma_0)}\}\}}{\sqrt{\mathcal{V}ar_f\{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i,\hat{\beta}_n)}{h(Y_i,\hat{\gamma}_n)}\}\}}}.$$

On the other hand

$$\mathcal{V}ar\{\sum_{i=1}^{n}\log\frac{g(Y_i,\hat{\beta}_n)}{h(Y_i,\hat{\gamma}_n)}\}=\sum_{i=1}^{n}\mathcal{V}ar_f\{\log\frac{g(Y_i,\hat{\beta}_n)}{h(Y_i,\hat{\gamma}_n)}\}+2\sum\sum_{i<j}\mathcal{C}ov_f\{\log\frac{g(Y_i,\hat{\beta}_n)}{h(Y_i,\hat{\gamma}_n)},\log\frac{g(Y_j,\hat{\beta}_n)}{h(Y_j,\hat{\gamma}_n)}\}.$$

Now

$$\hat{\mathcal{V}ar}_f\{\log\frac{g(Y_i,\hat{\beta}_n)}{h(Y_i,\hat{\gamma}_n)}\}=\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i,\hat{\beta}_n)}{h(Y_i,\hat{\gamma}_n)}\}^2-\{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i,\hat{\beta})}{h(Y_i,\hat{\gamma})}\}\}^2.$$

Then

$$S_n=\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\log\frac{g(Y_i,\hat{\beta}_n)}{h(Y_i,\hat{\gamma}_n)}-\mathcal{E}_f\{\log\frac{g(Y_i,\beta_0)}{h(Y_i,\gamma_0)}\}\}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i,\hat{\beta}_n)}{h(Y_i,\hat{\gamma}_n)}\}^2-\{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i,\hat{\beta})}{h(Y_i,\hat{\gamma})}\}\}^2+\frac{2}{n}\sum\sum_{i<j}\mathcal{C}ov_f\{\log\frac{g(Y_i,\hat{\beta}_n)}{h(Y_i,\hat{\gamma}_n)},\log\frac{g(Y_j,\hat{\beta}_n)}{h(Y_j,\hat{\gamma}_n)}\}}}.$$

The covariance term is a part of S_n and needs to computation, but it is reasonable if we expect that

$$\mathcal{C}ov_f\{\log\frac{g(Y_i,\hat{\beta}_n)}{h(Y_i,\hat{\gamma}_n)},\log\frac{g(Y_j,\hat{\beta}_n)}{h(Y_j,\hat{\gamma}_n)}\}\rightarrow\mathcal{C}ov_f\{\log\frac{g(Y_i,\beta_0)}{h(Y_i,\gamma_0)},\log\frac{g(Y_j,\beta_0)}{h(Y_j,\gamma_0)}\}$$

and use this fact that $\frac{g(Y_i,\beta_0)}{h(Y_i,\gamma_0)}$ and $\frac{g(Y_j,\beta_0)}{h(Y_j,\gamma_0)}$ are independent. Now if we consider the covariance term as negligible (which in Vuong's theorem (1989) disappears) by Vuong's theorem (1989), V_n has asymptotically the standard normal density. Thus

$$S_n=\frac{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i,\hat{\beta}_n)}{h(Y_i,\hat{\gamma}_n)}\}-\mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i,\hat{\beta}_n)}{h(Y_i,\hat{\gamma}_n)}\}\}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i,\hat{\beta}_n)}{h(Y_i,\hat{\gamma}_n)}\}^2-\{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i,\hat{\beta})}{h(Y_i,\hat{\gamma})}\}\}^2}}\xrightarrow{\mathcal{L}}\mathcal{N}(0,1)$$

S_n is different from Vuong's statistics in expectation term.

7.3.2 Confidence Interval for $\Delta\mathcal{E}KL$

Let $\overline{Z}_n=(Z_1,Z_2,...,Z_n)$, with $Z_i=(Y_i,X_i);i=1,2,...,n$; $(Y_i\in\mathcal{R}^d,X_i\in\mathcal{R}^m)$ is a sample of independent identically distributed random variables with common true unknown density (generating or true model) $f_{Y,X}(.,.)=f(.,.)$, and with conditional density of Y given X as $f^t_{Y|X}(.,.)$. Consider

$g_{Y|X}(Y|X,\beta)$ and $h_{Y|X}(Y|X,\gamma)$ as two non-nested models (postulated or candidate models). Following 4.4.1 we consider β_0 and γ_0 as the minimizer of KL criterion. It is known that the maximum likelihood estimators $\hat{\beta}_n$ and $\hat{\gamma}_n$ are consistent for β_0 and γ_0 respectively. For the reduced model, the KL criterion for these two models is given by

$$KL\{g_{Y|X}(.|.,\beta_0); f_{Y|X}(.,.)\} = \mathcal{E}_f\{\log f_{Y|X}(Y|X)\} - \mathcal{E}_f\{\log g_{Y|X}(Y|X,\beta_0)\}$$

and

$$KL\{h_{Y|X}(.|.,\gamma_0); f_{Y|X}(.,.)\} = \mathcal{E}_f\{\log\{ f_{Y|X}(Y|X)\} - \mathcal{E}_f\{\log h_{Y|X}(Y|X,\gamma_0)\}.$$

where in both of them the first part is irrelevant (because for all postulated models this term is fixed) and the second part is the relevant part for our goal. In both of above KL criteria the relevant parts are the quantity of interest, but can not be estimated, because they depend on unknown f. Akaike (1973) found that the expectation of the relevant part can be estimated. Denote the fitted models by $g_{Y|X}(Y|X,\hat{\beta}_n)$ and $h_{Y|X}(Y|X,\hat{\gamma}_n)$. The conditional KL criterion for the relevant parts, say, CKLs are

$$CKL_{g,n} = \mathcal{E}_f\{\log g_{Y|X}(Y|X,\hat{\beta}_n)|\overline{\mathcal{Z}}_n\}.$$

and

$$CKL_{h,n} = \mathcal{E}_f\{\log h_{Y|X}(Y|X,\hat{\gamma}_n)|\overline{\mathcal{Z}}_n\}.$$

The expected CKL, say, $\mathcal{E}KL_{g,n}$ is given by

$$\mathcal{E}_f\{CKL_{g,n}\} = \mathcal{E}KL_{g,n} = \mathcal{E}_f\{\log g_{Y|X}(Y|X,\hat{\beta}_n)\}$$

and similarly for $CKL_{h,n}$

$$\mathcal{E}_f\{CKL_{h,n}\} = \mathcal{E}KL_{h,n} = \mathcal{E}_f\{\log h_{Y|X}(Y|X,\hat{\gamma}_n)\}$$

Following sections 4.4.1 and 4.5.1 there is a asymptotic relation about $\mathcal{E}KL$ and its estimator. In fact based on Taylor's expansion of $-n\mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^n \log g_{Y|X}(Y_i|X_i,\hat{\beta}_n)\}$ about β_0 if $n(\hat{\beta}_n-\beta_0)(\hat{\beta}_n-\beta_0)^T$ is uniformly integrable, we have:

$$-\mathcal{E}_f\{\log g_{Y|X}(Y|X,\hat{\beta}_n)\} = -\mathcal{E}_f\{\log g_{Y|X}(Y|X,\beta_0)\} + \frac{1}{2n}tr(I^{-1}J) + o(n^{-1})$$

and

$$-\mathcal{E}_f\{\log g_{Y|X}(Y|X,\hat{\beta}_n)\} = -\mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^n \log g_{Y|X}(Y_i|X_i,\hat{\beta}_n)\} + \frac{1}{n}tr(I^{-1}J) + o(n^{-1}).$$

By these we conclude that

$$-\mathcal{E}_f\{\log g_{Y|X}(Y|X,\beta_0)\} = -\mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^n \log g_{Y|X}(Y_i|X_i,\hat{\beta}_n)\} + \frac{1}{2n}tr(I^{-1}J) + o(n^{-1}).$$

Theorem 7.1 *Under assumption A6, Vuong (1989), (For F-almost all (y,x), the function $|\log g(y|x)|^2$) and $|\log g(y|x)|^2$ are dominated by true (distribution function) $F_{Y,X}$-integrable functions independent of parameters in postulated models) we have*

$$\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n\{\log\frac{g(Y_i|X_i,\hat{\beta}_n)}{h(Y_i|X_i,\hat{\gamma}_n)} - \{\mathcal{E}KL_{g,n} - \mathcal{E}KL_{h,n}\}\}}{\sqrt{\frac{1}{n}\sum_{i=1}^n\{\log\frac{g(Y_i|X_i,\hat{\beta}_n)}{h(Y_i|X_i,\hat{\gamma}_n)}\}^2 - \{\frac{1}{n}\sum_{i=1}^n\{\log\frac{g(Y_i|X_i,\hat{\beta}_n)}{h(Y_i|X_i,\hat{\gamma}_n)}\}\}^2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Proof : Vuong (1989) has shown that

$$\frac{\frac{1}{\sqrt{n}}\{\sum_{i=1}^n \log\frac{g(Y_i|X_i,\hat{\beta}_n)}{h(Y_i|X_i,\hat{\gamma}_n)} - \mathcal{E}_f\{\log\frac{g_{Y|X}(Y|X,\beta_0)}{h_{Y|X}(Y|X,\gamma_0)}\}\}}{\sqrt{\mathcal{V}ar_f\{\log\frac{g(Y|X,\beta_0)}{h(Y|X,\gamma_0)}\}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Then

$$\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n\{\log\frac{h(Y_i|X_i,\hat{\gamma}_n)}{g(Y_i|X_i,\hat{\beta}_n)} - \mathcal{E}_f\{\log\frac{g_{Y|X}(Y_i|X_i,\hat{\beta}_n)}{h_{Y|X}(Y_i|X_i,\hat{\gamma}_n)}\} + \mathcal{E}_f\{\log\frac{g_{Y|X}(Y_i|X_i,\hat{\beta}_n)}{h_{Y|X}(Y_i|X_i,\hat{\gamma}_n)}\} - \mathcal{E}_f\{\log\frac{g_{Y|X}(Y_i|X_i,\beta_0)}{h_{Y|X}(Y|X,\gamma_0)}\}\}}{\sqrt{\mathcal{V}ar_f\{\log\frac{g(Y|X,\beta_0)}{h(Y|X,\gamma_0)}\}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Now by lemma 6.1 (see 6.6.2), if we set $U_n = \begin{pmatrix} Y_i \\ X_i \\ \hat{\beta}_n \end{pmatrix}$ and $U_0 = \begin{pmatrix} Y_i \\ X_i \\ \beta_0 \end{pmatrix}$ then $U_n \xrightarrow{\mathcal{P}} U_0$ and

similarly for $V_n = \begin{pmatrix} X_i \\ \hat{\beta}_n \end{pmatrix}$ and $V_0 = \begin{pmatrix} X_i \\ \beta_0 \end{pmatrix}$ where $V_n \xrightarrow{\mathcal{P}} V_0$. They implies the convergence in distribution for U_n and V_n. Now by continuous mapping theorem $\log g(Y_i, X_i, \hat{\beta}_n) \xrightarrow{\mathcal{P}} \log g(Y_i, X_i, \beta_0)$ and $\log g(X_i, \hat{\beta}_n) \xrightarrow{\mathcal{P}} \log g(X_i, \beta_0)$ thus $\log g(Y_i|X_i, \hat{\beta}_n) \xrightarrow{\mathcal{P}} \log g(Y_i|X_i, \beta_0)$ which implies that $\log g(Y_i|X_i, \hat{\beta}_n) \xrightarrow{\mathcal{L}} \log g(Y_i|X_i, \beta_0)$. In general for $i = 1, 2, ..., n$; we can split the random variable $\log g(Y_i|X, .)$ into its positive and negative parts which means that $\log g(Y_i|X_i, .) = \{\log g(Y_i|X_i, .)\}^+ - \{\log g(Y_i|X_i, .)\}^-$ where $\{\log g(Y_i|X_i, .)\}^+ = max\{\log g(Y_i|X_i, .), 0\}$, $\{\log g(Y_i|X_i, .)\}^- = max\{-\log g(Y_i|X_i, .), 0\}$ and

$$\mathcal{E}_f\{\log g(Y_i|X_i, .)\} = \mathcal{E}_f\{\log g(Y_i|X_i, .)\}^+ - \mathcal{E}_f\{\log g(Y_i|X_i, .)\}^-.$$

Assume that $\log g(Y_i|X_i, \hat{\beta}_n)$ is non-negative for all n and using (C4) for conditional density,

$$|\mathcal{E}_f\{\log g(Y_i|X_i, \hat{\beta}_n)\} - \mathcal{E}_f\{\log g(Y_i|X_i, \beta^0)\}| =$$

$$|\int_0^\infty \{P(\log f(Y_i|X_i, \hat{\beta}_n)) > \eta - P(\log f(Y_i|X_i, \beta_0)) > \eta\}\, d\eta| =$$

$$|\int_0^g \{P(\log f(Y_i|X_i, \hat{\beta}_n)) > \eta - P(\log f(Y_i|X_i, \beta_0)) > \eta\}\, d\eta| \leq$$

$$\int_0^g |P\{(\log f(Y_i|X_i, \hat{\beta}_n)) > \eta\} - P\{(\log f(Y_i|X_i, \beta_0)) > \eta\}|\, d\eta \to 0$$

because the interval of integration is bounded. Then by convergence in mean

$$\mathcal{E}_f\{\log f(Y_i|X_i, \hat{\beta}_n)\} \to \mathcal{E}_f\{\log f(Y_i|X_i, \beta_0)\}.$$

Similarly

$$\mathcal{E}_f\{\log g(Y_i|X_i, \hat{\gamma}_n)\} \to \mathcal{E}_f\{\log g(Y_i|X_i, \gamma_0)\}.$$

which implies that

$$\mathcal{E}_f\{\log \frac{g(Y_i|X_i, \hat{\beta}_n)}{h(Y_i|X_i, \hat{\gamma}_n)}\} \to \mathcal{E}_f\{\log \frac{g_{Y|X}(Y|X, \beta_0)}{h_{Y|X}(Y|X, \gamma_0)}\}$$

Now we have

$$\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\log\frac{g(Y_i|X_i,\hat{\beta}_n)}{h(Y_i|X_i,\hat{\gamma}_n)}-\mathcal{E}_f\{\log\frac{g_{Y|X}(Y_i|X_i,\hat{\beta}_n)}{h_{Y|X}(Y_i|X_i,\hat{\gamma}_n)}\}\}}{\sqrt{\mathcal{V}ar_f\{\log\frac{g(Y|X,\beta_0)}{h(Y|X,\gamma_0)}\}}}\xrightarrow{\mathcal{L}}\mathcal{N}(0,1).$$

If we use the variance estimator we have

$$\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\log\frac{g(Y_i|X_i,\hat{\beta}_n)}{h(Y_i|X_i,\hat{\gamma}_n)}-\mathcal{E}_f\{\log\frac{g_{Y|X}(Y_i|X_i,\hat{\beta}_n)}{h_{Y|X}(Y_i|X_i,\hat{\gamma}_n)}\}\}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i|X_i,\hat{\beta}_n)}{h(Y_i|X_i,\hat{\gamma}_n)}\}^2-\{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i|X_i,\hat{\beta}_n)}{h(Y_i|X_i,\hat{\gamma}_n)}\}\}^2}}\xrightarrow{\mathcal{L}}\mathcal{N}(0,1)$$

Then

$$\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\log\frac{g(Y_i|X_i,\hat{\beta}_n)}{h(Y_i|X_i,\hat{\gamma}_n)}-\{\mathcal{E}_f\{CKL_{g,n}\}-\mathcal{E}_f\{CKL_{h,n}\}\}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i|X_i,\hat{\beta}_n)}{h(Y_i|X_i,\hat{\gamma}_n)}\}^2-\{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i|X_i,\hat{\beta}_n)}{h(Y_i|X_i,\hat{\gamma}_n)}\}\}^2}}\xrightarrow{\mathcal{L}}\mathcal{N}(0,1)$$

and thus

$$\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\log\frac{g(Y_i|X_i,\hat{\beta}_n)}{h(Y_i|X_i,\hat{\gamma}_n)}-\{\mathcal{E}KL_{g,n}-\mathcal{E}KL_{h,n}\}\}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i|X_i,\hat{\beta}_n)}{h(Y_i|X_i,\hat{\gamma}_n)}\}^2-\{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i|X_i,\hat{\beta}_n)}{h(Y_i|X_i,\hat{\gamma}_n)}\}\}^2}}\xrightarrow{\mathcal{L}}\mathcal{N}(0,1).\qquad\blacksquare$$

Theorem 7.2 *Under Theorem 7.1 let*

$$D_n(g^{\hat{\beta}_n},h^{\hat{\gamma}_n})=n^{-1}\left[\frac{1}{2}AIC_{g,n}-\frac{1}{2}AIC_{h,n}\right]$$

and

$$\Delta_n(g^{\hat{\beta}_n},h^{\hat{\gamma}_n})=\mathcal{E}KL_n(g^{\hat{\beta}_n};f)-\mathcal{E}KL_n(h^{\hat{\gamma}_n};f)$$

where $AIC_{g,n}=-\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i|X_i,\hat{\beta}_n)+\frac{tr(I_g^{-1}J_g)}{n}$, $AIC_{h,n}=-\frac{1}{n}\sum_{i=1}^{n}\log h(Y_i|X_i,\hat{\gamma}_n)+\frac{tr(I_h^{-1}J_h)}{n}$ *et* $\mathcal{E}KL_n(g^{\hat{\beta}_n},f)=\mathcal{E}_f\{\log\frac{f(Y)}{g(Y;\hat{\beta}_n)}\}$ *and* $\mathcal{E}KL_n(h^{\hat{\gamma}_n},f)=\mathcal{E}_f\{\log\frac{f(Y)}{h(Y;\hat{\gamma}_n)}\}$, *under Theorem 7.1 we have*

$$n^{1/2}\left[D_n(g^{\hat{\beta}_n},h^{\hat{\gamma}_n})-\Delta_n(g^{\hat{\beta}_n},h^{\hat{\gamma}_n})\right]\xrightarrow{\mathcal{L}}\mathcal{N}(0,\omega_\star^2)$$

where $\omega_\star^2$ *is* $\mathcal{V}ar_f\{\log\frac{g(Y|X,\beta_0)}{h(Y|X,\gamma_0)}\}$.

proof : We know that $D_n(g^{\hat{\beta}_n},h^{\hat{\gamma}_n})=-n^{-1}\left[\sum_{i=1}^{n}\{\log\frac{g(Y_i|X_i,\hat{\beta}_n)}{h(Y_i|X_i,\hat{\gamma}_n)}\}\right]+\frac{p-q}{n}$ thus

$$D_n(g^{\hat{\beta}_n},h^{\hat{\gamma}_n})-\Delta_n(g^{\hat{\beta}_n},h^{\hat{\gamma}_n})=-\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i|X_i,\hat{\beta}_n)}{h(Y_i|X_i,\hat{\gamma}_n)}\}-\mathcal{E}KL_n(g^{\hat{\beta}_n};f)+\mathcal{E}KL_n(h^{\hat{\gamma}_n};f)+\frac{p-q}{n}$$

where

$$p = tr(I_g^{-1} J_g) + O(n^{-1})$$

and

$$q = tr(I_h^{-1} J_h) + O(n^{-1}),$$

see, 4.5.1. When the model is well specified p and q are the number of parameters in densities g and h respectively. Then

$$n^{1/2}\left[D_n(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n}) - \Delta_n(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n})\right] = -\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\log\frac{g(Y_i|X_i,\hat{\beta}_n)}{h(Y_i|X_i,\hat{\gamma}_n)} - \{\mathcal{E}KL_{g,n} - \mathcal{E}KL_{h,n}\}\}\right] + \frac{p-q}{\sqrt{n}}.$$

Now using Theorem 7.1, this fact that $\frac{p-q}{\sqrt{n}}$ for large n is negligible and symmetric property of normal distribution we have

$$n^{1/2}\left[D_n(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n}) - \Delta_n(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n})\right] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \omega_\star^2). \qquad \blacksquare$$

Corollary 1: Under Theorem 7.2, a $(1-\alpha)\%$ confidence interval for $\Delta_n(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n})$ is given by

$$\left[D_n(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n}) - n^{-1/2} z_{\alpha/2}\hat{\omega}_n \quad , \quad D_n(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n}) + n^{-1/2} z_{\alpha/2}\hat{\omega}_n\right]$$

where as before

$$\hat{\omega}_n^2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i|X_i,\hat{\beta}_n)}{h(Y_i|X_i,\hat{\gamma}_n)}\}^2 - \{\frac{1}{n}\sum_{i=1}^{n}\{\log\frac{g(Y_i|X_i,\hat{\beta}_n)}{h(Y_i|X_i,\hat{\gamma}_n)}\}\}^2}. \qquad \blacksquare$$

7.4 Logistic Regression:

The Logistic regression model, Cox (1970), has become a widely accepted method of analysis of binary (dichotomous) data. There are similarities and differences between linear and logistic regression. As we saw in 3.4.2 by generalizing the linear models we achieve a wide range of models to

describe the data. This generalization exactly introduces a new linear predictor based on the mean of the outcome variable Y which does no longer have to be normally distributed or even continuous. In fact in logistic models we have $Y_i \sim Bin(1,\pi_i)$ and $\xi(\mu) = \eta = X\beta$ where ξ be an invertible, smooth function of the mean vector $\mu = \mathcal{E}(Y)$. The explanatory variable X is in model through the logit link function $\eta = \log(\frac{\pi}{1-\pi})$, which is known as the log-odds transformation or *logit*. A model for the log-odds is called a logit or logistic regression model. It is seen that the logit transformation yields a linear relationship for the logit model. In this case the logit link is commonly used but the other link like probit and the complementary $\log - \log$ is available. For multiple logistic regression we have

$$\log(\frac{\pi}{1-\pi}) = X\beta$$

and then

$$\pi = \frac{\exp(X\beta)}{1+\exp(X\beta)}.$$

It is clear that the derivatives of likelihood function with respect to the parameters are not linear in parameters then maximum likelihood estimator for β is given by the iterative procedure like Newton-Raphson algorithm which gives

$$\beta^{(t+1)} = \beta^{(t)} + \{X^T diag\{\pi_i^{(t)}(1-\pi_i^{(t)})\}X\}^{-1}X^T(Y-\pi^{(t)})$$

with start OLS solution for β at iteration $t = 0$ as $\beta^{(0)}$.

In the model selection context usually the measures of goodness of fit are based on the residuals. In fact we determine whether the fitted model's residual variation is small, displays no systematic tendency and follows the variability postulated by the model. In logistic regression

$$\hat{\pi} = \frac{\exp(X\hat{\beta})}{1+\exp(X\hat{\beta})}.$$

Two usual measures of goodness of fit test for logistic regression are Pearson Chi-square and the likelihood ratio (Deviance). These statistics have both the χ^2 distribution and lack of fit occurs when the values of these statistics are large. Hosmer and Lemesho (1989) discuss two methods of grouping based on the ranked estimated probabilities that form groups of equal numbers of subjects (deciles of risk) or use fixed cut points on the $[0,1]$ interval. Tsiatis (1980) proposed an approach based on fixed groups in the covariate space that yields a score test for fit.

Sometimes it is interesting that we categorize some explanatory variables in the regression models. For categorizing the cutpoint must be meaningful in the research area. This introduces some regressors in our model. We consider the Body-Mass Index (*BMI*) as an important explanatory variable which effects the depression; some people consider three categories for *BMI* as poor (desirable), average and high (morbidly obese). Introducing a logistic model for modeling binary response as depression (Y) acording to $BMI(X_1)$, age (X_2) and gender (X_3). The logit is given by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_2 * X_3 + \varepsilon$$

A categorization of *BMI* could be done by terciles. We note that in the logistic scale the linear model is not nested in the tercile model. The terciles will introduce two regressors in the model as X_{11} and X_{12}. See table 7.1.

Table 7.1- Introduced regressors by terciles.

Category for BMI	X_{11}	X_{12}
BMI $\in$ Tersile 1	1	0
BMI $\in$ Tersile 2	0	1
BMI $\in$ Tersile 3	0	0

Now the model is

$$Y = \beta_0 + \beta_{11}X_{11} + \beta_{12}X_{12} + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_2 * X_3 + \varepsilon$$

where X_{11} and X_{12} are the $0-1$ regressors. This model describes three parallel regression planes which can differ in their intercept. See table 7.2.

Table 7.2- Three parallel regression models generated by terciles.

Category for BMI	*Regression Model*
BMI $\in$ Tersile 1	$Y = (\beta_0 + \beta_{11}) + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_2 * X_3 + \varepsilon$
BMI $\in$ Tersile 2	$Y = (\beta_0 + \beta_{12}) + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_2 * X_3 + \varepsilon$
BMI $\in$ Tersile 3	$Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_2 * X_3 + \varepsilon$

β_0 is an intercept for person with BMI in tercile 3. Here a BMI in tercile 3 serves as a baseline category or reference group with which the other depression categories are compared. If age and gender distributions are the same for the three groups, we could compare the mean of the three groups.

We also consider the quadratic model as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_3 + \beta_5 X_2 * X_3 + \varepsilon.$$

It is clear that the linear model in logistic scale is nested in quadratic model. The logistic curve for this three models are shown in figure 11. A simple analysis of linear, tercile and quadratic models in logistic scale are given at the end of this section. As the likelihood and *AIC* comparison of models as we talked in 1.9 the likelihood function increases when the number of parameters in the model increases. The linear case has five parameters while both the tercile and quadratic models have six parameters. But for *AIC* is a little different, the *AIC*'s are ordered according to where the models are nested or non-nested. The results for this three models is given in table 7.3.

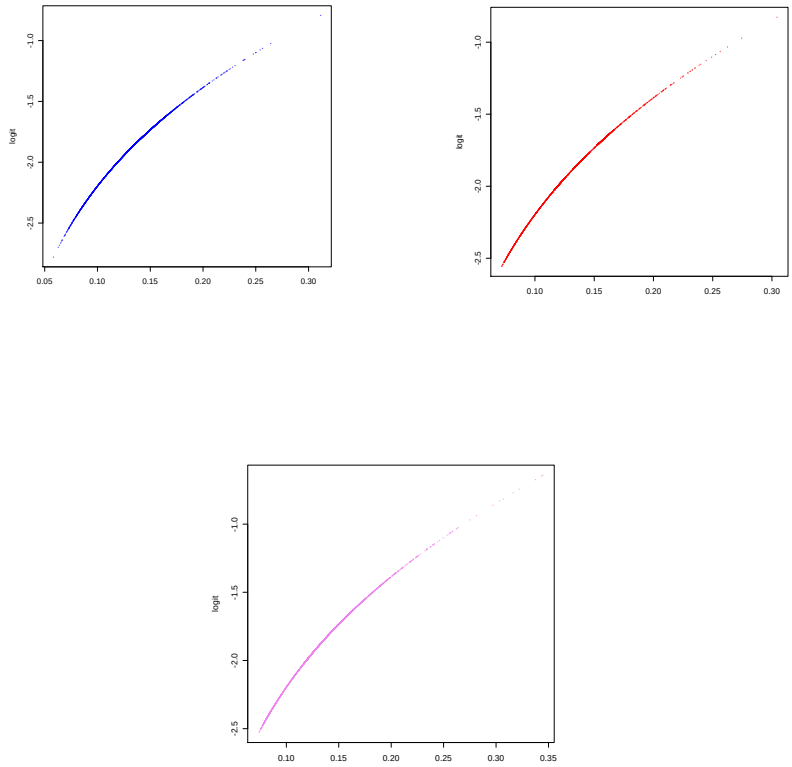


Figure 11: Logistic curves for linear, tercile and quadratic models in logistic scale.

Table 7.3- Maximized likelihood values and AIC's for linear, tercile and quadratic models in logistic scale

Model	Likelihood	AIC
Linear	-1346.25	2702.5
Tercile	-1345.60	2703.2
Quadratic	-1342.93	2697.9

A simple analysis of our real data is given in table 7.4-7.6. A deeper analysis of this section is pre-

sented in appendix B.

Table 7.4- Estimated coefficients for linear, tercile and quadratic models.

Linear model in Logistic scale				
Coefficients	*Estimate*	*Std.Error*	$Z-value$	$P(Z>\|z\|)$
β_0	-4.48569	1.02396	-4.381	1.18e-05
β_1	-0.02952	0.01362	-2.167	0.030232
β_2	0.04303	0.01212	3.551	0.000384
β_3	3.09190	1.16200	2.661	0.007794
β_4	-0.03922	0.01526	-2.269	0.010189

Tercile model in Logistic scale				
Coefficients	*Estimate*	*Std.Error*	$Z-value$	$P(Z>\|z\|)$
β_0	-5.32560	0.91920	-5.794	6.88e-09
β_{11}	0.31106	0.12699	2.450	0.01430
β_{12}	0.14375	0.12889	1.115	0.26473
β_2	0.04255	0.01211	3.515	0.00044
β_3	3.04914	1.16188	2.624	0.00868
β_4	-0.03875	0.01526	-2.539	0.01110

Quadratic model in Logistic scale				
$Coefficients$	$Estimate$	$Std.Error$	$Z-value$	$P(Z>\|z\|)$
β_0	-1.360874	1.546192	-0.880	0.378780
β_1	-0.269327	0.089172	-3.020	0.002525
β_2	0.004672	0.001701	2.746	0.006031
β_3	0.041546	0.012136	3.423	0.000619
β_4	3.051574	1.161999	2.626	0.008636
β_5	-0.039072	0.015263	-2.560	0.010472

Chapter 8

Conclusion and perspective

The purpose of this research is to clarify some facts and provide a simple test to model selection which is a relatively new branch of mathematical statistics. The aim of statistical modeling is to identify the model that most closely approximates the underlying process. As a part of model selection, in chapter 5 we are in search of a goodness of fit test for the simple situation where $F_0(.)$ is a known distribution function; when there are unknown parameters, we have to first to estimate and then plug-in it into the test statistic. For example in a simple normal case with mean and variance as unknown parameters, we can estimate these parameters by their known estimators as sample mean and sample variance respectively and obtain a goodness of fit test for normality. Our idea is considering a random sample of size n and a goodness of fit test procedure which introduces a likelihood ratio test for each fixed value in the variable space. The known Union-Intersection test (UIT) is our proposal to solve this problem. The level of test and efficiency for this test has been verified. It seems that our statistic is comparable to the Berk-Jones statistic. As a further work we may consider more complex weight function in the definition of the proposed statistic, and compare

the new statistics with other goodness of fit tests. On the other hand from a statistical standpoint, the observed data are tainted with sampling error. Consequently, when we fit a model to the data, the model performance reflects the population pattern and also the patterns due to sampling error. Such patterns will be specific to the particular sample and will not repeat themselves in other samples. A complex model with many parameters tends to capture these sample patterns more easily than a simple model with few parameters. Then, the complex model yields a better fit to the data, but it may not be because of its ability to more accurately approximate the underlying process but rather because of its ability to capitalize on sampling error. Therefore, choosing a model based solely on its fit, without appropriately filtering out the effects due to sampling error, will result in choosing an overly complex model that generalizes poorly to other data from the same underlying process. Consequently model selection should not be based on a model's ability to fit particular sample data but instead should be based on its ability to capture the characteristics of the population. There are actually some different tests for model selection and consequently some different questions can be asked about them. Each of the tests has advantages and disadvantages in their domain of usage. In almost all of the tests and criteria for model selection the maximum likelihood estimator and maximized likelihood function have an essential role. With a careful attention there are two separate functions over the parameter space. The first is the probability density for maximum likelihood estimator over the parameter space, and the second one is the likelihood function, which defines the probability of the data in any particular point in the parameter space. As we see both are defined on the parameter space but each has a different meaning. The i.i.d. assumption allows to obtain normal asymptotic distributions for both the maximum likelihood estimator and the log-likelihood of the observed data. This knowledge is a starting point to define a simple model selection criterion as the normalized maximized likelihood function. This works for some known cases when the distribution

of data is normal. But its disadvantage is that using the data at hand for estimation and evaluation. On the other hand the maximized log-likelihood increases when the number of useless parameters in a model increases. This leads to select the more complex model. Akaike solved these two difficulties by his assumption about domain of data and by parsimony quantity. The *AIC* introduced by Akaike considers the parsimony principle to reduce the bias term in model selection. The main goal of this statistic is to estimate twice the relevant part of the Kullback-Leibler divergence. But this is an unusual quantity to estimate, because it depends on the number of observations. In fact we encounter the same difficulty as for estimating a sum in a population by a sum in the sample; in this case we know that the error of estimation grows when the sample size increases. The normalization idea is useful to solve this problem, and allows us to define a criterion for model selection. In fact in chapter 6 and related appendix (appendix A) we want to show that the normality of the *AIC*'s and that the constructed confidence interval by normalized *AIC* reflects the fact about models. When we do not know the correct model the average of limits of these types of confidence intervals give us an idea about the number of parameters in the model. On the other hand these types of confidence intervals show us that the *AIC* is not an increasing function of the number of parameters in the model. In fact complexity in model is good for reduction of bias, while simplicity of model reduces the tendency to over-fit. On the other hand the best trade-off between unknown bias and unknown variance is the aim of model selection. But how to achieve this trade off? This is the main question in model selection. In chapter 6 and appendix A we are about the reduction of bias. With the assumption that the future observations are in the same domain as the observed data it seems that the bias is generally more important than the variance. In chapter 6 and appendix A we are about the reduction of bias. With the assumption that the future observations are in the same domain as the observed data it seems that the bias is generally more important than the variance. As the first step in model

selection we are in search of a admissible bond for the average of confidence interval limits to select the best model under parsimony. This admissible set of models must be contain all models which are near to the selected model by *AIC*. Consider some models with $k, k+1, ...k_*, ..., k+l$ explanatory variables. Assume that the model with k_* explanatory variables is the selected model by *AIC*. On the other hand by our simulation we see that for intermediate sample size there is an intersection between some of the confidence intervals for models with $j \neq k_*$ explanatory variables and that for selected model. We say two models are near to each other if their average confidence interval limits have intersection. We set these kind of models in the admissible set. Now our search for the best model will be in this set. This chapter needs to be developed by further work with other models for finding the admissible line which enables us to select the set of candidate models. After it, we may use a classical variable selection approach to select the best model between the condidate models or we may use the approach developed in chapter 7 to compare the models . Anyway the result of this chapter is a basis for chapter 7.

In chapter 7 and appendix B we improve our idea by constructing a tracking confidence interval for a difference of expected Kullback-Leibler risks for two candidate models. The proposed confidence interval contains the difference of Kullbak-Leibler risks with a fixed probability. This interval has another interpretation for the use of *AIC*'s. In fact we are not in a situation to detect the best model but we are in search for a model which has the relatively less risk compared to other models. It is because all the models are mis-specified. For constructing the confidence interval we need to estimate the variance of a normalized difference of *AIC*'s; a good estimation would take into account the covariance between two maximized log-likelihoods, but it seems that finding this covariance is difficult and is an open problem. Another open problem arises in a situation where we have many competing models. It is because in a real situation we have a sample of size *n* and many competing

models to fit to the data at hand. We may propose a two-stage approach where in the first stage we choose the best two models by means of maximized likelihood function and then return to the proposed approach to choose the best one. But a good search could be done by a generalized approach. On the other hand we assumed that our sample are independent and identically distributed, a nice generalization would relax this assumption to extend this approach. In this work we have applied our results to normal regression models and logistic regression. But the theory is general and could be applied to the other types of regression models like Poisson regression for counting response variable and log-normal model as a standard approach to the analysis of skewed response variable, see Finney(1941) and Bradu and Mundlak (1970) may be of interest. Here we consider the model selection for one dimensional random variables a generalization could be done for p dimensional random variables.

Chapter 9

Bibliography

Akaike, H. (1973) *Information theory and an extension of maximum likelihood principle.* Second International Symposium on Information Theory, Akademia Kiado, 267-281.

Atkinson, A.C.(1970) *A method for discriminating between models* Journal of the Royal Statistical Society B **32**, 323-344

Berk, R.H. and Jones, D.H. (1979) *Goodness-of-Fit Test that dominate the Kolmogorov statistics.* Zeitschrift fur Wahrsheinlichkeitstheorie und Verwandte Gebiete, **47**, 47-59.

Biernacki, C. (2004) *Testing for a Global Maximum of the Likelihood.* Journal of Computational and Graphical Statistics, **14**, 3, 657-674.

Bozdogan, H. (2000) *Akaik's information criterion and recent developments in information complexity.* Journal of Mathematical Psychology,**44**, 62-91.

Chernoff, H. and Lehmann, E.L. (1954) *The use of maximum likelihood estimates in* χ^2 *tests of goodness of fit.* Ann. Math. statist. **25**, 579-586.

Cochran, W.G. (1952) *The* χ^2 *test of goodness of fit.* Ann. Math. statist. **23**, 315-345.

Commenges, D. Joly, P. Gegout-Petit, A. and Liquet, B. (2007)*Choice between semi-parametric estimators of Markov and non-Markov multi-stat models from generally onservations.* Scandinavian Journal of statistics, in press.

Cox, D.R.(1961) *Test of separate families of hypothesis* proceeding of the 4th Berkeley symposium, Vol. **1**(University of California Press,Berkeley), 105-123.

Cox, D.R.(1962) *Further result on tests of separate families of hypeses*Journal of the Royal Statistical Society B **24**, 406-424.

Dastoor, N.K. (1983) *Some aspects of testing non-nested hypothesis* Journal of Econometrics **21**, 213-228.

Davidson, R. and MacKinnon, (1981)*Several tests for model specification in the presence of alternative hypotheses* Econometrica **49**, 781-793.

Fisher, G.R., and McAleer, M. (1981) *Alternative procedures and associated tests of significance for non-nested hypotheses* Journal of Econometrics, **16**, 103-119.

Hurvich, C.M. and Tsai,C.L. (1989)*Regression and time series model selection criterion* Biometrika **76**, 297-307

Ishiguro, M., Sakamoto,Y. and Kitagawa, G. (1997) *Bootstrapping log likelihood and EIC, an extension of AIC*Annals of the institue of Statistical Mathematics, **49**, 411-434

Jager, L. and Wellner, J.A. (2005)*A new goodness of fit test: the reversed Berk-Jones statistic* http://bayes.stat.washington.edu/www/research/reports/2004/tr443.pdf.

Knight, K. (1999) *Mathematical Statistics* Chapman and Hall.

Konishi, S. and Kitagawa, G. (1996) *Generalized Information Criteria in Model Selection.* Biometrika **83**, 4, 575-590.

Lehmann, E.L. (1998) *Elements of Large-Sample Theory.* Springer-Verlag, New York.

Lehmann, E.L. (1986) *Testing Statistical Hypothesis.* Wiley, New York.

Linhart, H. and Zucchini, W. (1986) *Model Selection.* Wiley, New York.

Mallows, C.L. (1973) *Some comments on C_p* Technometrics, **15**, 661-675.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models* CHAPMAN AND Hall.

Myung, I.J. (2000) *The importance of complexity in model selection* Journal of Mathematic **44**, 190-204.

Pesaran, M.H. (1974) *On the general problem of model selection* Review of Economic Studies **41**, 153-171.

Pesaran, M.H. andDeaton, A.S. (1978) *Testing non-nested nonlinear regression models* Econometrica **46**, 667-694.

Shapiro, S.S. and Wilk, M.B. (1965) *An analysis of variance test for normality.* Biometrika **52**, 591-611.

Shapiro, S.S., Wilk, M.B. and Chen, H.J. (1968) *A comparative study of various tests for normality.* J. Amer. Statist. Ass. **63**, 1343-1372.

Schwarz, G. (1978) *Estimating the dimension of a model* Annals of Statistics, **6**, 461-464

Shimodaira, H. (1998) *An application of multiple comparison techniques to model selection* Annals of Ins. statistical mathemathics **50**, No. 1, 1-13.

Shimodaira, H. (2001) *Multiple comparisons of log-likelihoods and combining non-nested models with application to phylogenetic tree selection* Communication in statistics**30**, 1751-1772.

Stephens, M.A. (1986) editors. *Goodness-of-Fit Techniques.* Marcel Dekker, New York.

Van der Varrt, A.W. (1998) *Asymptotic Statistics.* Cambridge University Press.

Vuong, Quang H. (1989) *Likelihood ratio tests for model selection and non-nested hypotheses The level of test and efficiency for this test will be verified. It seems that our statistic is comparable by*

Berk-Jones statistic. . Econometrica, **57,** No. 2, 307-333.

White, H. (1982) *Maximum Likelihood Estimation of Misspecified Models.* Econometrica, **50**(1):1-26, jan.

White, H. (1994) *Estimation Inference and Specification Analysis.* Cambridge University Press.

Weisberg, S. and Bingham, C. (1975) *An approximate analysis of variance test for non-normality suitable for machine calculation.* Technometrics **17**, 133-134.

Yanagihara, H. and Ohomoto, C. (2005) *On distribution of AIC in linear regression models* Journal of Statistical Planning and Inference **133**, 417-433.

Chapter 10

Appendix

Part I

APPENDIX A

Model selection: Application to the Multiple Regression Model

A.Sayyareh, D. Commenges and A. Bar-Hen

29th June 2007

Model selection: Application to the Multiple Regression Model

SUMMARY

Some key words : Akaike criterion, Confidence interval, Kullback-Leibler, Model selection, Multiple regression, Variable selection.

10.1 Introduction

Model selection is estimating the performance of different models in order to choose the best one. It proceeds in two steps. The first step is to select a model (as the family of hypotheses or family of densities) between competing models and second step is to select a particular hypothesis or density from the model. The first step is sometimes a hard step, it needs to some background. For example in regression survey may be we start with linear models and then complicate (if necessary) the model by allowing to the facts about population under study. In literature the selection problem is where the rival models come from a nested hierarchy of k-degree polynomials. If there is no background, that is required is that the models share the common goal of predicting the same data. The second step is estimation of parameters from the observed data. On the other hand the statistical models are typically merely approximations to reality and so sometimes a wrong model is fit to the observations, but in practice we do it for some reasons. First because a little of knowledge is better than nothing, second an assumed parametric model may be close to the true unknown model, so that very little is lost by assumed model and we can use the rich literature of parametric statistics, and third

in some statistical subjects the estimated parameters for an assumed parametric model can often be interpreted usefully.

In practic when we collect the data, there are many unobserved data from population under study and also future observatons. As a aim of model selection may be we are in search of the model to find the functional value of the unobserved response or to use the model to prediction of future. How we can confident that the postulated model is accurate? Thus selection and evaluation of a model is an important step in any research. For example in fitting curve context adding new terms add extra adjustable coefficients (parameters) and these will improve fit to some degree. The problem is when we add new term we gain in fit, but if this gain is small how do we make this trade-off between addition a new term and gain in fit? And which value of gain is small or too small? So we turn to the hypothesis testing or ordering the models by model selection criteria. The first one introduce an absolute discrimination and second one is a relative discrimination.

Instead of the classical hypothesis testing approach to cover the analysis of the non-nested models may be we consider the hypothesis testing to model selection. But the hypothesis testing is a deterministic approach. Generally in curve fitting area in which the dependent random variable Y is a function of the explanatory variable(s) the means for detect the fit is least square or likelihood approaches. The least square approach has a limitation when the error term is not normal. Then it is reasonable if we take the likelihood or equivalently the log-likelihood function as a measure of fit. As the goal of the fitting curve we want use the fitted model to predict the future. In the Akaike framework, the base assumption is that the new data are the re-sampled from the past (the data at hand). This is an advantage for Akaike (1973) Information criterion, *AIC*, as the estimator for relevant part of the Kullback-Leibler discrepancy. In this direction fit is defined in terms of discrepancy from the true density, or the closeness to the true density. When we are in search of the

best model there is not a reason for separate the hypotheses as the null and alternative hypotheses, i.e. all of hypotheses are the null hypothesis. This is a point which indicates that may be in Akaike framework we consider a function of the maximized log-likelihood as the test statistic. On the other hand because the conclusion of *AIC* is not never about the truth or falsity of a hypothesis, but about its closeness to the truth, we take this logic for our idea and use the confidence interval as a set of acceptable hypotheses. Consider a sample of i.i.d. random variables $\overline{Y} = (Y_1, Y_2, ..., Y_n)$ which follows a linear regression model. It means that $Y_i = \sum_{j=1}^{p_\star} \beta_j X_{ij} + \varepsilon_i; \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_\star^2); \quad i = 1, 2, ..., n.$ The vectors $X_i = (X_{i1}, X_{i2}, ...X_{ip_\star})^T$ of covariate values, and the vector $\beta_\star = (\beta_{1_\star}, \beta_{2_\star}, ..., \beta_{p_\star})^T$ of regression coefficients is to be estimated. In this case our parameter vector is $\theta_\star = (\beta_\star^T, \sigma_\star^2)^T$. Then we have,

$$Y = X_\star \beta_\star + \varepsilon_\star, \quad \varepsilon_\star \sim \mathcal{N}(0, \sigma_\star^2 I). \qquad (1)$$

We refer to (1) as the true model. Consider the postulated models as

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I). \qquad (2)$$

The postulated models are defer in ranks of design matrices $X_{n\times p}$. The standard approach in model selection is fitting postulated models to the observations and then determine which of them is the best approximation to the true model.

Formally by notation in classical statistics we want to test the null hypothesis $\mathcal{H}_0 : f(y) = g(y; \beta_\star)$ for all $y \in \mathcal{R}$ and some $\beta_\star \in B$, if we reduce it to $\mathcal{H}_0 : f(y) = g(y; \beta_\star) \quad a.e.$ in possible range of y for some $\beta_\star \in B$ this null hypothesis is equivalent to testing for,

$$\mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n} \log f(Y_i)\} = \mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n} \log g(Y_i|X_i, \beta_\star).\} \qquad (3)$$

A known measure of discrepancy between the true and the postulated models is the Kullback-Leibler

criterion. In literature Shimodaira (1998,2001) has extended the Linhart's test (1988) with different concept as Cox (1962) by definition its confidence set at a given significance level. He consider a set of postulated models as $\{M(j)|j \in \mathcal{M}\}$ and for each $j \in \mathcal{M}$ consider a statistical test. We propose an other criterion which takes in order the models with property as minimum average confidence interval for expected Kullback-Leibler criterion. In fact in this approach we consider the null hypothesis as $\mathcal{H}_0^j : Y = X_{n\times j}^{(j)}\beta + \varepsilon^{(j)}, \quad \varepsilon^{(j)} \sim \mathcal{N}(0, \sigma^2 I), \quad j \in Z$ (an integer set) and construct a confidence interval for negative expected *AIC* and we decide for which postulated model the limits of intervals are minimum. For simplicity consider two non-nested postulated models as

$$\mathcal{H}_0^k : Y = X_{n\times k}^{(k)}\beta + \varepsilon^{(k)}, \quad \varepsilon^{(k)} \sim \mathcal{N}(0, \sigma^2 I),$$

and

$$\mathcal{H}_0^l : Y = X_{n\times l}^{(l)}\gamma + \varepsilon^{(l)}, \quad \varepsilon^{(l)} \sim \mathcal{N}(0, \sigma^2 I)$$

We noted that by these hypotheses we are not in the situation to decide which model is the correct model, but we want to know which model is better. Now the search for the best model, in the first step will be the search between all of non-nested models with k and l explanatory variable, separately, and then comparing the average of interval limits for two postulated models.

As the decision we choice the model with minimum of the average of interval limits. By notation in literature this process is the variable subset selection of the multiple regression. All of investigation is to selecting a best subset of predictors. Many different definition of best can be found in the literature. The forward selection method for subset selection is common in statistic, it checked for improvement in the partial F-values and R^2. The usual statistics to verify that whether or not the proposed model is significant are R^2 (adjusted), the residual mean square, and Mallow's C_p.
For example, the forward selection includes additional variables in the model based on maximizing

the increment to R^2 from step to step, but in a conditional sense. By these criteria, a best model can be identified for fixed value of explanatory variables (a specified subset with k element) but there is no general method for selecting an overall best model. As an other investigation to model choice may be we consider the information estimator for true density and postulated model with some explanatory variables as Akaike (1973). Akaike has used his criterion for selecting among the competing models. In fact he select a model with minimum lack of fit in care of parsimony. In our approach and in information context we want to consider all of models with k explanatory variables by constructing a confidence interval for respected sub-class of models. In search of best model we consider the minimum average of the confidence interval limits, where the minimization procedure is taken on classes of all sub sets for $k = 1,2,...,K$ explanatory variables. After making decision about the number of the explanatory variable in the model we can investigate the best model in the interest subset of explanatory variables. In other word we want to check that a model with $k = j$, explanatory variables is enough or not. To answer to this question we consider a measure of goodness as average of confidence intervals for subset with $k = j$ variables. This kind of model selection is a overall type search.

10.2 Expected Kullback-Leibler Criteria and AIC

More generally, let $\overline{Z}_n = (Z_1, Z_2, ..., Z_n)$, with $Z_i = (Y_i, X_i); i = 1,2,...,n$; $(Y_i \in \mathcal{R}^d, X_i \in \mathcal{R}^m)$ be a sample of independent identically distributed random variables with common true unknown density (generating model) $f^t_{Y,X}(.,.) = f^t(.,.)$ and with conditional density of Y given X as $f_{Y|X}(.,.)$. Consider $g^{\beta}(.) = g_{Y|X}(Y|X,\beta)$ as postulated model and set β_0 as the minimizer of KL criterion. It

is known that the maximum likelihood estimator $\hat{\beta}_n$ is consistent for β_0. For reduced model, (see, Commenges et al(2007)) the KL criterion is given by

$$KL\{g_{Y|X}(.|.,\beta_0); f_{Y|X}(.,.)\} = \mathcal{E}_f\{\log f_{Y|X}(Y|X)\} - \mathcal{E}_f\{\log g_{Y|X}(Y|X,\beta_0)\}$$

where the first part is irrelevant and second part is relevant part for our goal.

As the more complicated distance may be we consider the Hellinger or Matusita distance of affinity, see, Bar-Hen and Daudin (1998) for asymptotic distribution of this statistic.

In regression context we have the variance as a parameter to estimate, but our focus is on the regression coefficients and for simplicity we eliminate the variance estimator in notation. Fortunately the variance and coefficients estimators are independent and there is not difficulty to search for the asymptotic distribution of statistics which contain both of them at the same time. In above *KL* criterion the relevant part is quantity of interest, but can not be estimated, because they depend on unknown f.

Denote the fitted models by $g_{Y|X}(Y|X,\hat{\beta}_n)$. The conditional KL criterion for relevant part,'say' CKLs is

$$CKL_{g,n} = \mathcal{E}_f\{\log g_{Y|X}(Y|X,\hat{\beta}_n)|\overline{Z}_n\}.$$

The expected CKL, say,'$\mathcal{E}KL_{g,n}$' is given by

$$\mathcal{E}_f\{CKL_{g,n}\} = \mathcal{E}KL_{g,n} = \mathcal{E}_f\{\log g_{Y|X}(Y|X,\hat{\beta}_n)\}$$

$\mathcal{E}_f\{CKL_{g,n}\}$ is a consistent estimator for $\mathcal{E}_f\{\log g_{Y|X}(Y|X,\beta_0)$. Using the empirical distribution function for expected $CKL_{g,n}$, then its sample analogue is $\frac{1}{n}\sum_{i=1}^{n}\log g_{Y|X}(Y_i|X_i,\hat{\beta}_n)$ which minimizes an estimator of $KL\{g_{Y|X}(.|.,\beta); f_{Y|X}(.,.)\}$

Model selection based on Kullback-Leibler discrepancy (*KL*), is developed by inference about relevant part of the *KL* divergence. It was Akaike (1973) which introduced an estimator for relevant part

as Akaike Information Criteria, (AIC). Originally The AIC is defined as

$$AIC = -2L^{g^{\hat{\beta}_n}} + 2p$$

where $L^{g^{\hat{\beta}_n}}$ is the maximized log-likelihood function for postulated model. As noted by Hurvich and Tsai (1989) when the dimension of the postulated model, increase in comparison to n, the sample size, AIC becomes strongly biased which leads to over fitting problem. They have proposed a biased corrected estimator of AIC in linear regression context. In fact they shown that in this case the corrected AIC is

$$CAIC = n\log\hat{\sigma}^2 + \frac{n(n+p)}{n-p-2}.$$

AIC is the unbiased estimator for $-\mathcal{E}KL_{.,n}$. Now constructing a confidence interval for $-\mathcal{E}KL_{.,n}$. make sense, because this confidence interval will be a confidence interval for $\mathcal{E}(AIC)$. We saw that the postulated models are different in design matrices, then they have the different CAIC. By construction the confidence interval for $\mathcal{E}(AIC)$ we will be able to sort the postulated models.

10.3 Hypothesis Testing

If we write the null hypothesis $\mathcal{H}_0$ by notation in (3) this hypothesis is equivalent to $\mathcal{H}_0 : KL(g^{\beta_\star}; f) = 0$ we propose the test statistic , as $\hat{KL}(g^{\beta_\star}; f)$ then we reject $\mathcal{H}_0$ if $\hat{KL}(g^{\beta_\star}; f) > C$ which is equivalent to $T_n(\overline{Y}, \hat{\beta}_n) = \frac{1}{n}\sum_{i=1}^{n} \log g(Y_i|X_i, \hat{\beta}_n) < K_n$. This is the bias estimator for the KL (relevant part of KL) divergence and then a biased estimator for distance between the true and the postulated model. The biased term is given in Konishi and Kitagawa (1996) and Bozdogan (2000) as follows,

$$bias = \mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n} \log g(Y_i|X_i, \hat{\beta}_n) - \int_{\mathcal{R}} \log g(y|X_i, \hat{\beta}_n) f(y)\, dy\} = \frac{1}{n} tr(I^{-1}J) + O(n^{-2})$$

where I is the inverse Fisher information matrix in inner product (Hessian) form, and J is the outer product form of the Fisher information matrix for vector β.

In specified case $tr(I^{-1}J) = p$ the number of parameter in postulated model. The test function for this type hypothesis is given by

$$\phi(\overline{Y}) = \begin{cases} 1 & \text{if } T_n(\overline{Y}, \hat{\beta}_n) < K_n \\ 0 & \text{if } T_n(\overline{Y}, \hat{\beta}_n) > K_n \end{cases}$$

Under some regularity conditions and this fact that

$$\frac{1}{n}\sum_{i=1}^{n} \log g(Y_i;\beta)\} \xrightarrow{\mathcal{P}} \mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n} \log g(Y_i;\beta)\} \qquad (4)$$

this test statistic is consistent or asymptotically unbiased for $\mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n} \log g(Y_i;\beta_\star)\}$

Theorem 1 : Suppose that $Y_1, ..., Y_n$ i.i.d with unknown density $f(.)$. Let $\mathcal{G} = \{g(.,\beta); \beta \in B \subseteq \mathcal{R}\}$ is a parametric family of assumed densities for $Y_i's$. If $\mathcal{H}_0$ holds, under conditions (C0)-(C4) and (4) we have:

$$T_n(\overline{Y}, \hat{\beta}_n) \xrightarrow{\mathcal{P}} \mathcal{E}_f\{\frac{1}{n}\sum_{i=1}^{n} \log g(Y_i;\beta_\star)\}.$$

To make a decision about $\mathcal{H}_0$ we need to know the distribution of the test statistic under null hypothesis. In theorem 2 we handle an asymptotic density of our statistics.

Theorem 2: Under regularity conditions

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} [\log g(Y_i|X_i, \hat{\beta}_n) - \mathcal{E}_f\{\log g(Y_i|X_i, \hat{\beta}_n)\}] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{V}ar_f\{\log g(Y|X_i, \beta_\star)\}).$$

proofs are given in Chapter 6, see Theorems 6.2 and 6.5.

Corollary 1:

$$\mathcal{Z}_n = \frac{-\frac{AIC}{2n} - \mathcal{E}KL_{g,n}}{\sqrt{\frac{1}{n}\mathcal{V}ar_f\{\log g(Y_i|X_i,\beta_\star)\}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

It is because, by theorem 2

$$\mathcal{Z}_n = \frac{\frac{1}{n}(p-\frac{1}{2}AIC) - \mathcal{E}KL_{g,n}}{\sqrt{\frac{1}{n}\mathcal{V}ar_f\{\log g(Y_i|X_i,\beta_\star)\}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

which implys that

$$\mathcal{Z}_n = \frac{-\frac{AIC}{2n} - \mathcal{E}KL_{g,n}}{\sqrt{\frac{1}{n}\mathcal{V}ar_f\{\log g(Y_i|X_i,\beta_\star)\}}} + \frac{\frac{p}{n}}{\sqrt{\frac{1}{n}\mathcal{V}ar_f\{\log g(Y_i|X_i,\beta_\star)\}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

We assumed that $\mathcal{V}ar_f\{\log g(Y_i|X_i,\beta_\star)\} < \infty$, using Slutsky's theorem for $\mathcal{Z}_n$ and $\frac{\frac{p}{n}}{\sqrt{\frac{1}{n}\mathcal{V}ar_f\{\log g(Y_i|X_i,\beta_\star)\}}} \rightarrow 0$ show that the corollary is true. As an estimator for $\mathcal{V}ar_f\{\log g(Y_i|X_i,\beta_\star)\}$ we use the estimator as

$$\frac{1}{n}\sum_{i=1}^{n}\{\log g(Y_i|X_i,\hat{\beta}_n)\}^2 - \{\frac{1}{n}\sum_{i=1}^{n}\log g(Y_i|X_i,\hat{\beta}_n)\}^2.$$

See Biernacki (2004)and using Slutsky's theorem.

Using theorem 2 we can achieve a confidence interval for $\mathcal{E}KL_{g,n}$ or $\mathcal{E}_f(AIC)$ as follows

$$p\{-z_{\alpha/2} < \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\log g(Y_i|X_i,\hat{\beta}_n) - \mathcal{E}KL_{g,n}\}}{\sqrt{\mathcal{V}ar_f\{\log g(Y|X,\beta_\star)\})}} < z_{\alpha/2}\} = 1-\alpha$$

which give us a $(1-\alpha)\%$ confidence interval for $\mathcal{E}KL_{g,n}$ as

$$\left[\left(T_n(\overline{Y},\hat{\beta}_n) - n^{-1/2}z_{\alpha/2}\sqrt{\mathcal{V}ar_f\{\log g(Y|X,\beta_\star)\}}\right), \left(T_n(\overline{Y},\hat{\beta}_n) + n^{-1/2}z_{\alpha/2}\sqrt{\mathcal{V}ar_f\{\log g(Y|X,\beta_\star)\}}\right)\right]$$

or for $-\mathcal{E}KL_{g,n}$ as

$$\left[\left(\frac{AIC-2p}{2n} - n^{-1/2}z_{\alpha/2}\sqrt{\mathcal{V}ar_f\{\log g(Y|X,\beta_\star)\}}\right), \left(\frac{AIC-2p}{2n} + n^{-1/2}z_{\alpha/2}\sqrt{\mathcal{V}ar_f\{\log g(Y|X,\beta_\star)\}}\right)\right]$$

Corollary 1 help us to construct a confidence interval for $-\mathcal{E}KL_{g,n}$ as

$$\left[\left(\frac{AIC}{2n} - n^{-1/2}z_{\alpha/2}\sqrt{\mathcal{V}ar_f\{\log g(Y|X,\beta_\star)\}}\right), \left(\frac{AIC}{2n} + n^{-1/2}z_{\alpha/2}\sqrt{\mathcal{V}ar_f\{\log g(Y|X,\beta_\star)\}}\right)\right]$$

10.4 Simulation

10.4.1 exploration of our result

To explore and apply the corollary 1, we consider two simulation studies. Figure 1 shows the result of simulation study of normality for standardized AIC. We generate 10^5 observations from a bivariate uniform density each one on $[-\sqrt{3},\sqrt{3}]$. We consider the logistic linear regression and find the precisely estimate of $\mathcal{E}KL_{g,n}$ and $\mathcal{V}ar_f\{\log g(Y_i|X_i,\beta_\star)\}$ which are respectively $\check{\mathcal{E}}KL_{g,n}=-0.40879$ and $\check{\mathcal{V}}ar_f\{\log g(Y_i|X_i,\beta_\star)\}=0.31518$. For sample size n=1000 and b=1000 iterations, we achieve 1000 values for AIC in logistic regression. To confirm that our quantity Z_n is asymptotically standard normal we draw the histogram of observed AIC's and its cumulative distribution function to comparison with standard normal density. These figures are agreement with normality of *AIC*.

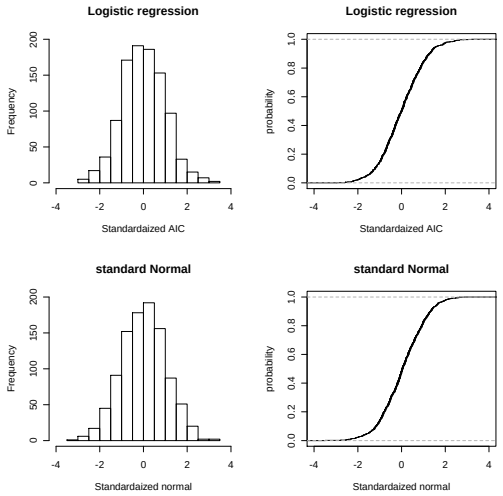


Figure 1: Comparison of histograms and cumulative distribution functions of observed AIC's and standard normal density

10.4.2 Application to The Multiple Regression Model.

As an illustration of this approach we consider the model choice in multiple regression. Consider the regression model as (2), i.e.

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1}$$

Suppose there is a suspects that some regressors are unusefull for explaning Y, i.e. the true value of the coefficients of these regressors are zero but we do not know which of the coefficients. Then the appropriate candidate models are all submodels of the regression equation. To formulate this, let $u \in \{0,1\}^p$, that is u is a $p \times 1$ vecor of ones and zeros. Then we can define the submodels as

$$\{\beta : \beta_j = 0 \quad \text{if} \quad u_j = 0; j = 1,...,p\}.$$

By yhis notation the full model is corresponding to $u = (1,...,1)$ and the set of all candidate models is given by

$$\mathcal{M} = \{M_u : u \in \{0,1\}^p\}.$$

To illustrate our approach we considered i.i.d sample of size n of (Y, X_1, X_2, X_3). As a true model we set $Y = 0.5 + X_1 + 1.25X_3$. By this knowledge we want to construct a confidence interval for $-\mathcal{E}KL$. In fact we expect that average of the uppers and lowers limits of confidence interval of $-\mathcal{E}KL$ for the models with two explanatory variable be less than same things for the models with one explanatory variable. Our simulation study for n=10000 observations shows that the average confidence interval for models with $\{X_1, X_2\}$ and $\{X_2, X_3\}$ is in the left of the average confidence interval for models with $\{X_1\}$, $\{X_2\}$ and $\{X_3\}$ as explanatory variables. This average interval for models with one and two explanatory variables were $(1.63852, 1.66645)$ and $(1.51282, 1.54030)$ respectively. The length of these intervals are 0.027936 and 0.02748. For intercept model the $-\mathcal{E}KL$ was 1.935429. This result confirms that best model to apply is the model with two explanatory

variables, see table 1. For another true model as $Y = 0.5 + X_1 + 1.25X_2 + 2.5X_4$ we consider the subclasses as $\left\{\{X_1,X_2,X_3\},\{X_1,X_3,X_4\},\{X_2,X_3,X_4\}\right\}$, $\left\{\{X_1,X_2\},\{X_1,X_3\},\{X_1,X_4\},\{X_2,X_3\},\{X_2,X_4\},\{X_3,X_4\}\right\}$ and $\left\{\{X_1\},\{X_2\},\{X_3\},\{X_4\}\right\}$ and construct the average intervals for each one, the result was $(1.789956, 1.817359)$, $(2.014563, 2.042036)$ and $(2.283159, 2.351173)$ respectively. The lengths of these intervals are 0.027403, 0.027473 and 0.068014. For intercept model the $-\mathcal{E}KL$ was 2.522105. This result again confirms that the model must be a model with three explanatory variables, see table 2. The result for relatively small sample size is a little different. For example for $n = 100$ observations the result for regression model with four explanatory variables is given in table 3. The intervales are overlap and length of average of interval limits are increased.

Table 1- The average of interval limits for AIC's and its length for regression model.

(case with three explanatory variables, n=10000)

True Model: $Y = 0.5 + X_1 + 1.25X_3$		
class of explanatory variables	*average of interval limits*	length of interval
$\left\{\{X_1\},\{X_2\}\{X_3\}\right\}$	(1.63851, 1.66645)	0.02794
$\left\{\{X_1,X_2\}\{X_2,X_3\}\right\}$	(1.51282, 1.54030)	0.02748

Table 2- The average of interval limits for AIC's and its length for regression model.

(case with four explanatory variables, n=10000)

True Model: $Y = 0.5 + X_1 + 1.25X_2 + 2.5X_4$		
class of explanatory variables	*average of interval limits*	length of interval
$\left\{\{X_1\},\{X_2\},\{X_3\},\{X_4\}\right\}$	(2.283159, 2.351173)	0.068014
$\left\{\{X_1,X_2\},\{X_1,X_3\},\{X_1,X_4\},\{X_2,X_3\},\{X_2,X_4\},\{X_3,X_4\}\right\}$	(2.014563, 2.042036)	0.027473
$\left\{\{X_1,X_2,X_3\},\{X_1,X_3,X_4\},\{X_2,X_3,X_4\}\right\}$	(1.789956, 1.817359)	0.027403

Table 3- The average of interval limits for AIC's and its length for regression model.

(case with four explanatory variables, n=100)

True Model: $Y = 0.5 + X_1 + 1.25X_2 + 2.5X_4$		
class of explanatory variables	*average of interval limits*	length of interval
$\left\{\{X_1\},\{X_2\},\{X_3\},\{X_4\}\right\}$	(1.973091, 2.642405)	0.669314
$\left\{\{X_1,X_2\},\{X_1,X_3\},\{X_1,X_4\},\{X_2,X_3\},\{X_2,X_4\},\{X_3,X_4\}\right\}$	(1.865973, 2.153741)	0.287766
$\left\{\{X_1,X_2,X_3\},\{X_1,X_3,X_4\},\{X_2,X_3,X_4\}\right\}$	(1.641730, 1.923548)	0.281818

Example 1: As a consequence of theorem 2 consider the linear model described in (2). The log-likelihood for this model is given by

$$\log\prod_{i=1}^{n} g(Y_i|X_i\beta,\sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta).$$

The *MLE* of the parameters β and σ are given by $\hat{\beta}_n = (X^TX)^{-1}X^TY$ and $\hat{\sigma}_n^2 = \frac{(Y-X\hat{\beta}_n)^T(Y-X\hat{\beta}_n)}{n}$ respectively. Under model (1) we have

$$\mathcal{E}_f\{\log\prod_{i=1}^{n} g(Y_i|X_i\beta,\sigma^2)\} = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(X_\star\beta_\star - X\beta)^T(X_\star\beta_\star - X\beta).$$

The expected maximized log-likelihood is

$$-\frac{n}{2}\log 2\pi - \frac{n}{2}\log\hat{\sigma}_n^2 - \frac{n\sigma_\star^2}{2\hat{\sigma}_n^2} - \frac{1}{2\hat{\sigma}_n^2}(X_\star\beta_\star - X\hat{\beta}_n)^T(X_\star\beta_\star - X\hat{\beta}_n).$$

It is known that

$$\mathcal{E}_f\{\frac{n\sigma_\star^2}{2\hat{\sigma}_n^2}\} = \frac{n^2}{2}\mathcal{E}_f\{(\frac{n\hat{\sigma}_n^2}{\sigma_\star^2})^{-1}\} = \frac{n^2}{2(n-p-2)}$$

and

$$\mathcal{E}_f\{\frac{1}{2\hat{\sigma}_n^2}(X_\star\beta_\star - X\hat{\beta}_n)^T(X_\star\beta_\star - X\hat{\beta}_n)\} = \frac{1}{2}\mathcal{E}_f\{\frac{\sigma_\star^2}{\hat{\sigma}_n^2}\frac{(X_\star\beta_\star - X\hat{\beta}_n)^T(X_\star\beta_\star - X\hat{\beta}_n)}{\sigma_\star^2}\} = \frac{np}{2(n-p-2)}.$$

Now

$$\sum_{i=1}^{n}\mathcal{E}KL_{g,n} = -\frac{n}{2}\log 2\pi - \frac{n}{2}\mathcal{E}_f\{\log\hat{\sigma}_n^2\} - \frac{n(n+p)}{2(n-p-2)}.$$

On the other hand

$$\mathcal{E}_f\{\log\hat{\sigma}_n^2\} = \Psi(\frac{n-p}{2}) + \log\frac{2\sigma_\star^2}{n}$$

where Ψ is the digamma function,see, Hurvich and Tsai (1989)

By theorem 2 we have

$$\frac{\frac{1}{\sqrt{n}}\{\frac{n}{2}\log\frac{2\sigma_\star^2}{n} + \frac{n}{2}\Psi(\frac{n-p}{2}) - \frac{n}{2}\log\hat{\sigma}_n^2 - \frac{n}{2} + \frac{n(n+p)}{2(n-p-2)}\}}{\sqrt{\frac{1}{2}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

We can use this result to construct a confidence interval.

$$1-\alpha = P(L<\sigma_\star^2<U) = p(\log L<\log\sigma_\star^2<\log U) = p\{\frac{\log\frac{2}{n}+\Psi(\frac{n-p}{2})+\log L-\log\hat{\sigma}_n^2-1+\frac{n+p}{n-p-2}\}}{\sqrt{\frac{2}{n}}}<$$

$$\frac{\log\frac{2}{n}+\Psi(\frac{n-p}{2})+\log\sigma_\star^2-\log\hat{\sigma}_n^2-1+\frac{n+p}{n-p-2}}{\sqrt{\frac{2}{n}}}<\frac{\log\frac{2}{n}+\Psi(\frac{n-p}{2})+\log U-\log\hat{\sigma}_n^2-1+\frac{n+p}{n-p-2}}{\sqrt{\frac{2}{n}}}\}$$

Then

$$\frac{\log\frac{2}{n}+\Psi(\frac{n-p}{2})+\log L-\log\hat{\sigma}_n^2-1+\frac{n+p}{n-p-2}}{\sqrt{\frac{2}{n}}} = -Z_{\frac{\alpha}{2}}$$

and

$$\frac{\log\frac{2}{n}+\Psi(\frac{n-p}{2})+\log U-\log\hat{\sigma}_n^2-1+\frac{n+p}{n-p-2}}{\sqrt{\frac{2}{n}}} = Z_{\frac{\alpha}{2}}$$

now we have

$$L = \exp\{-Z_{\frac{\alpha}{2}}\sqrt{\frac{2}{n}}-\log\frac{2}{n}-\Psi(\frac{n-p}{2})+\log\hat{\sigma}_n^2+1-\frac{n+p}{n-p-2}\}$$

and

$$U = \exp\{Z_{\frac{\alpha}{2}}\sqrt{\frac{2}{n}}-\log\frac{2}{n}-\Psi(\frac{n-p}{2})+\log\hat{\sigma}_n^2+1-\frac{n+p}{n-p-2}\}$$

Conclusion

The purpose of this research is to clarify some facts and provide a simple test to model selection which is relatively new branch of mathematical statistics. The aim of statistical modeling is to identify the model that most closely approximates the underlying process. On the other hand from a statistical standpoint, observed data are tainted with sampling error. Consequently, when we fit a model to the data, the model's performance reflects population pattern and also the patterns due to sampling error. Such patterns will be specific to the particular sample and will not repeat themselves in other samples. A complex model with many parameters tends to capture these sample patterns more easily than a simple model with few parameters. Then, the complex model yields a better fit to the data, it is not because of its ability to more accurately approximate the underlying process but rather because of its ability to capitalize on sampling error. Therefore, choosing a model based solely on its fit, without appropriately filtering out the effects due to sampling error, will result in choosing an overly complex model that generalizes poorly to other data from the same underlying process. Consequently model selection should not be based on a model's ability to fit particular sample data but instead should be based on its ability to capture the characteristics of the population. There are actually some different tests to model selection and consequently some different questions can be asked about them. Each of tests have advantages and disadvantage in their domain of usage. In almost all of the tests and criteria to model selection the maximum likelihood estimator and maximized likelihood function have a essential role. With a careful attention there are two separate functions over parameter space. The first is the probability density for maximum likelihood estimator over the parameter space, and the second one is the likelihood function, which defined the probability of the data in any particular point in parameter space. As we see both are defined on parameter space but each has a different meaning. They are related by normality assumption which also determines

the stochastic behavior of the log-likelihood of the observed data. This knowledge is a starting point to define a simple model selection criterion as normalized maximized likelihood function. This works for some known case when the distribution of data is normal. But its disadvantage is that using the data at hand for estimation and evaluation. On the other hand increases when the number of useless parameters in a model increases. This leads to select the more complex model. Akaike solved these two difficulties by his assumption about domain of data and by parsimony quantity. The *AIC* introduced by Akaike consider the parsimony principal to reduce the bias term in model selection. The main goal of this statistics is to estimate two times the relevant part of the Kullback-Leibler divergence. But this is an unusual quantity to estimate, because it depends to the number of observations. In fact we encounter the same difficulty as for estimating a sum in a population by a sum in the sample, in this case we know that the error of estimation grows when the sample size increases. The normalization idea is useful to solve this problem. This is the criterion which we use to clear the fundamental problem in model selection. In fact we want to show that the normality of the *AIC*'s and that the constructed confidence interval by normalized *AIC* reflects the fact about models. When we do not know the correct model the average of limits of these types of confidence interval give us an idea about the number of parameters in the model. On the other hand these types of confidence intervals show us the *AIC* is not a increasing function of the number of parameters in the model. Actually we are in search of a distinguished line between the values of the average of confidence interval limits to select the best model under parsimony. In fact complexity in model is good for reduction of bias, and that simplicity of model reduces the tendency to over fit. On the other hand the best trade off between unknown bias and unknown variance is the model selection criterion aims. But how to do it trade off? This is the main question in model selection. Here we are about the reduction of bias. With this hypothesis that the future observations

are in the same domain as observed data it seems that the bias is more important than variance. Actually as the first step in model selection we are in search of a admissible bond for the average of confidence interval limits to select the best model under parsimony. This admissible set of models must be contain all models which are near to the selected model by *AIC*. Consider some models with $k, k+1, \ldots k*, \ldots k+l$ explanatory variables. Assume that the model with $k*$ explanatory variables is the selected model by *AIC*. On the other hand by our simulation it seems that for intermediate sample size there is the intersection between some of the confidence intervales for models with $j \neq k*$ explanatory variables and that for selected model. We say two models are neer to each other if there average confidence interval limits has intersection. We set these kind of models in the admissible set. Now our search will be in this set. This chapter needs to be developed by further work with other models for finding the admissible line which enable us to select the set of candidate models. This work needsbe developed by further work with other models for finding the distinguished line which enable us to select the set of simpler candidate models.

REFERENCES

Akaike, H. (1973) *Information theory and an extension of maximum likelihood principle.* Second International Symposium on Information Theory, Akademia Kiado,267-281.

Bar-Hen, A. and Daudin, J.J. (1998) *Asymptotic distribution of Matusita's distance: Application to the location model* Biometrika, **85**, 2, 477-481

Commenges, D. Joly, P. Gegout-Petit, A. and Liquet, B. (2007)*Choice between semi-parametric estimators of Markov and non-Markov multi-stat models from generally onservations.* Scandinavian Journal of statistics, in press.

Konishi, S. and Kitagawa, G. (1996) *Generalised Information Criteria in Model Selection.* Biometrika **83,** 4, 875-590.

Linhart, H. and Zucchini, W. (1986) *Model Selection.* Wiley, New York.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models* Chapman and Hall

White, H. (1982) *Maximum Likelihood Estimation of Misspecified Models.* Econometrica, **50**(1):1-26, jan.

White, H. (1994) *Estimation Inference and Specification Analysis.* Cambridge University Press.

Part II

APPENDIX B

Inference about differences of AIC: application to the choice of variable coding in logistic regression

D. Commenges [1,2], A. Sayyareh [1,3], L. Letenneur [2,4] and A. Bar-Hen [5]

1 INSERM, U875, Bordeaux, F-33076, France

2 Univ Bordeaux 2, Bordeaux, F33076, France

3 Univ Razi, Kermanshah, Iran

4 INSERM, U593, Bordeaux, F33076, France

5 INRA, U518, Paris, F-75231, France

E-mail: daniel.commenges@isped.u-bordeaux2.fr

Inference about differences of AIC: application to the choice of variable coding in logistic regression

SUMMARY

We propose a better use of Akaike information criterion (AIC), focussing on two issues: inference (one must not forget that AIC is a statistic) and interpretation (the exact value of AIC has no direct interpretation while we are interested in quantifying the risks when using particular models). We propose a statistic, a normalisation of a difference of Akaike criteria, which estimates the difference of expected Kullback-Leibler risks between maximum likelihood estimators of the distribution in two different models. The variability of this statistic can be estimated so that an interval can be constructed which contains the true difference of expected Kullback-Leibler risks with a pre-specified probability. A simulation study shows that the method works and it is illustrated using a study of the relationship between body-mass index and depression in elderly people.

Some key words : Akaike criterion, depression, Kullback-Leibler, logistic regression, model choice.

10.5 Introduction

Since its proposal by Akaike (1973), Akaike information criterion (AIC) has had a huge impact on so-called “model choice”, in particular in the application of statistical methods; see the presentation of deLeuwe (1992). It is often used in its original simple form, precisely because of its simplicity. Many variants of the criterion have been proposed. We may cite in particular the EIC (Konishi and Kitagawa, 1996; Shibata, 1997) which makes use of the bootstrap; the approach has been extended

to the choice of semiparametric estimators by Liquet, Sakarovitch and Commenges (2004) and Commenges et al. (2007). One concern with AIC is that it is felt that it does not put enough weight on the quality of parsimony of the model, and other criteria have been proposed such as the BIC (Schwartz, 1978) or approaches based on complexity (Bozdogan, 2000).

Our aim is to propose a better use of AIC, keeping here in the framework of parametric models. We will focus essentially on two issues which have been rather neglected in theoretical developments. The first is that of inference: it is generally forgotten that AIC is a statistic and as such has a distribution. AIC is commonly used to select the "best" model on the basis of a sample; however if another sample of the same size was available we might find that another model has a smaller AIC. So we should pay attention to the differences of AIC between different models and be able to estimate the variability of these differences. Such a study can be based on the results of Vuong (1989). However Vuong (1989) placed himself in an asymptotic context in which the Akaike correction is negligible.

The other issue is the interpretation of differences of AIC. Indeed, the value of AIC has no intrinsic meaning; in particular AIC is not invariant to a one-to-one transformation of the random variables. Investigators commonly display big numbers, only the last digit of which are used to decide which is the smallest. We recall that a normalized difference of AIC is an estimate of a difference of Kullback-Leibler risks and thus, is interpretable. We give some examples of values of such differences to help develop an intuition of what a large or a small difference is.

In section 2 we present the relevant Kullback-Leibler risk and we show that the normalized difference of AIC is an estimate of the difference of risks; moreover we propose a so-called "tracking interval" which should contain the difference of risks with a given probability; we also give insight in the interpretation of the differences of risks. This general approach may change the use of AIC since

we do not pretend to detect the "best" model but identify which estimators are acceptable on the basis of the available data. For illustrating this general approach we apply it to the problem of choosing between different codings of an explanatory variable in logistic regression. Section 3 presents a simulation study which allows in particular to assess the properties of the proposed tracking interval. In section 4 we present an illustration on real data: this is a study of the effect of body-mass index (BMI) on depression using data from the Paquid study. Section 5 is a short conclusion.

10.6 Theory about inference of differences of AIC criteria

10.6.1 Estimating a difference of Kullback-Leibler divergences

Consider a sample of independently identically distributed (iid) random variables $\overline{Y}_n = (Y_i, i = 1\ldots, n)$ having probability density function (pdf) $f = f(.)$. Let us consider two models : $(g) = (g^{\beta}(.))_{\beta \in B}, B \subset \Re^p$ and $(h) = (h^{\gamma}(.))_{\gamma \in \Gamma}, \Gamma \subset \Re^q$.

Definition 10.1 *(i)* (g) *and* (h) *are non-overlapping if* $(g) \cap (h) = \emptyset$*; (ii)* (g) *is nested in* (h) *if* $(g) \subset (h)$*; (iii)* (g) *is well specified if there is a value* $\beta_* \in B$ *such that* $g^{\beta_*} = f$*; otherwise it is mis-specified.*

The loglikelihood loss of g^{β} relatively to f for observation Y is $\log \frac{f(Y)}{g^{\beta}(Y)}$. Akaike (1973) grounds this choice of a loss function by arguing that all information for discriminating between distributions is contained in the likelihood ratio (Blackwell, 1953) so that the loss should be a function of it, and showing that the logarithm is the best function to choose. The expectation of this loss under f, or risk, is the Kullback-Leibler divergence (Kullback, 1968) between g^{β} and f: $\mathrm{KL}(g^{\beta}, f) = \mathrm{E}_f[\log \frac{f(Y)}{g^{\beta}(Y)}]$. We have $\mathrm{KL}(g^{\beta}, f) \geq 0$ and $\mathrm{KL}(g^{\beta}, f) = 0$ implies that $g^{\beta} = f$, that is $\beta = \beta_*$. The Kullback-Leibler divergence is often intuitively interpreted as a distance between the two pdf (or more generally between the two probability measures) but this is not mathematically a distance; in particular the

Kullback-Leibler divergence is not symmetric. It may be felt that this is a drawback, and in particular it makes any graphical representation perilous. However this feature may also have a deep meaning in our particular problem: there is no symmetry between f, the true pdf, and g^{β}, a possible pdf. So we shall take on the fact that the Kullback-Leibler divergence is an expected loss (with respect to f) and not a distance. We assume that there is a value $\beta_0 \in B$ which minimizes $\mathrm{KL}(g^{\beta}, f)$. If the model is well specified $\beta_0 = \beta_*$; if the model is mis-specified $\mathrm{KL}(g^{\beta_0}, f) > 0$.

Since the main interest of a model is to approach f, it is of obvious interest to estimate β_0. We have that

$$\mathrm{KL}(g^{\beta}, f) = \mathrm{E}_f[\log f(Y)] - \mathrm{E}_f[\log g^{\beta}(Y)].$$

The first term on the right-hand side $H(f) = \mathrm{E}_f[\log f(Y)]$ is the entropy of f and cannot be estimated directly since f is unknown; however, it does not depend on the parameters β nor on (g). The second term on the right-hand can not be directly computed because of the expectation under f; however, replacing f by its empirical estimate we obtain the estimator

$$-n^{-1}\sum_{i=1}^{n} \log g^{\beta}(Y_i) = -n^{-1} L_{\overline{Y}_n}^{g^{\beta}},$$

where $L_{\overline{Y}_n}^{g^{\beta}}$ is the loglikelihood based on the sample $\overline{Y}_n$. Thus, the maximum likelihood estimator $\hat{\beta}_n$ minimizes a natural estimator of $\mathrm{KL}(g^{\beta}, f)$. Moreover it can be shown that $\hat{\beta}_n$ is a consistent estimator of β_0.

Now if we consider two or more models, there is the problem of choosing between them. A natural way is to choose (g) against (h) if $\mathrm{KL}(g^{\beta_0}, f) < \mathrm{KL}(h^{\gamma_0}, f)$; we shall say in that case that (g) is *closer* to f than (h) (avoiding to qualify (g) as"better" which may be misleading in this context). There are two problems: (i) we can not estimate $\mathrm{KL}(g^{\beta_0}, f)$ because the entropy of f can not be correctly estimated; (ii) β_0 and γ_0 are unknown. The two problems are solved by noting

that we can estimate the difference of Kullback-Leibler divergences $\mathrm{KL}(g^{\beta_0}, f) - \mathrm{KL}(h^{\gamma_0}, f)$ by $-n^{-1}(L_{Y_n}^{\hat{\beta}_n} - L_{Y_n}^{\hat{\gamma}_n})$.

This result may not be completely satisfactory in practice if n is not very large because the distribution we will use is $g^{\hat{\beta}_n}$ rather than g^{β_0}. In consequence a more relevant criterion for model choice is $\mathrm{E}_f[\log \frac{f(Y)}{g^{\hat{\beta}_n}(Y)}]$ that we call the expected Kullback-Leibler risk (or simply Kullback-Leibler risk) and that we denote by $\mathrm{EKL}(g^{\hat{\beta}_n}, f)$. This is the point of view introduced by Akaike (1973). He also showed that $n^{-1}L_{Y_n}^{\hat{\beta}_n}$ overestimated $\mathrm{E}_f[\log g^{\hat{\beta}_n}(Y)]$ (because of the maximisation procedure) and proposed a criterion correcting for the number of parameters of the model:

$$\mathrm{AIC}(g^{\hat{\beta}_n}) = -2L_{Y_n}^{\hat{\beta}_n} + 2p.$$

Akaike's approach was revisited by Linhart and Zucchini (1986) who showed that:

$$\mathrm{EKL}(g^{\hat{\beta}_n}, f) = \mathrm{KL}(g^{\beta_0}, f) + \frac{1}{2}n^{-1}\mathrm{Tr}(I_g^{-1}J_g) + o(n^{-1}), \tag{10.1}$$

where $I_g = -\mathrm{E}_f[\frac{\partial^2 \log g^{\beta}(Y)}{\partial \beta^2}|_{\beta_0}]$ and $J_g = \mathrm{E}_f\{[\frac{\partial \log g^{\beta}(Y)}{\partial \beta}|_{\beta_0}][\frac{\partial \log g^{\beta}(Y)}{\partial \beta}|_{\beta_0}]^T\}$. This can be nicely interpreted by saying that the risk $\mathrm{EKL}(g^{\hat{\beta}_n}, f)$ is the sum of the mis-specification risk $\mathrm{KL}(g^{\beta_0}, f)$ plus the statistical risk $\frac{1}{2}n^{-1}\mathrm{Tr}(I_g^{-1}J_g)$. Note in passing that if (g) is well specified we have $\mathrm{KL}(g^{\beta_0}, f) = 0$ and $I_g = J_g$, and thus $\mathrm{EKL}(g^{\hat{\beta}_n}, f) = \frac{p}{2n} + o(n^{-1})$.

We also have:

$$\mathrm{EKL}(g^{\hat{\beta}_n}, f) = -n^{-1}L_{Y_n}^{\hat{\beta}_n} + H(f) + \frac{1}{n}\mathrm{Tr}(I_g^{-1}J_g) + o_p(n^{-1}). \tag{10.2}$$

Here we have essentially estimated $\mathrm{E}_f[\log g^{\beta_0}(Y)]$ by $n^{-1}L^{\hat{\beta}_n}$ but because of the overestimation bias, the factor $\frac{1}{2}$ in the last term disappears; thus the term $\frac{1}{n}\mathrm{Tr}(I_g^{-1}J_g)$ is the sum of two equal terms, the statistical error and the estimation bias of the mis-specification risk (of course the mis-specification risk is estimated up to the constant $H(f)$). Akaike criterion follows from (10.2) by

multiplying by $2n$, deleting the constant term $H(f)$ and replacing $\mathrm{Tr}(I_g^{-1}J_g)$ by p; in fact the correction p arises only if the model is well-specified (in which case $I_g = J_g$) but Linhart and Zucchini (1986) argue that it can be used even if the model is not well-specified. Using (10.2) we obtain:

$$-n^{-1}\{L_{Y_n}^{g^{\hat{\beta}_n}} - L_{Y_n}^{h^{\hat{\gamma}_n}} - [\mathrm{Tr}(I_g^{-1}J_g) - \mathrm{Tr}(I_h^{-1}J_h)]\} = \Delta(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n}) + o_p(n^{-1}),$$

where $\Delta(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n}) = \mathrm{EKL}(g^{\hat{\beta}_n}, f) - \mathrm{EKL}(h^{\hat{\gamma}_n}, f)$. It is possible to estimate the matrices I_g, J_g, I_h and J_h by plugging the estimators $\hat{\beta}_n$ and $\hat{\gamma}_n$ into the expression of these matrices, and thus an estimator of $\Delta(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n})$ is obtained. A simpler estimator of $\Delta(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n})$ is obtained by using the Akaike approximation $\mathrm{Tr}(I_g^{-1}J_g) \approx p$:

$$D(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n}) = \frac{1}{2}n^{-1}[\mathrm{AIC}(g^{\hat{\beta}_n}) - \mathrm{AIC}(h^{\hat{\gamma}_n})] = -n^{-1}[L_{Y_n}^{g^{\hat{\beta}_n}} - L_{Y_n}^{h^{\hat{\gamma}_n}} - (p-q)].$$

We will prefer estimator $g^{\hat{\beta}_n}$ to $h^{\hat{\gamma}_n}$ if this "estimate" is negative, meaning that the estimate of the expected loss incurred in using $g^{\hat{\beta}_n}$ in place of f is less than that incurred in using $h^{\hat{\gamma}_n}$.

Thus, in contrast with AIC, $D(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n})$ has an interpretation since it tracks the quantity of main interest $\Delta(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n})$ with pretty good accuracy. Moreover it has important invariance properties.

Lemma 1 (Invariance properties) *Both $\Delta(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n})$ and $D(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n})$ are invariant under reparametrization, one-to-one transformation of the observed variables and change of the reference probability.*

The proof is straightforward. It can be noted that AIC itself is invariant under reparametrization but neither under one-to-one transformation of the observed variables nor change of the reference probability.

10.6.2 Tracking interval for a difference of Kullback-Leibler divergences

In practice the epidemiologists or biostatisticians choose the model which has the best AIC. However it is important to know with which confidence we can infer the sign of the difference of the EKL from

the difference of the AIC. Moreover the statistic $D(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n})$ estimates the difference of the expected losses which is of interest *per se* and should be interpreted. We are in a context of model choice or rather of estimator choice. The question is not to find the true model, because all the models are more or less mis-specified; it is not even to choose the closest model, but the best estimator based on the available sample. The good choice does not depend only on the models but also on the quantity of information (here essentially n) available in the sample.

We focus on the case where $g^{\beta_0} \neq h^{\gamma_0}$. This is necessarily the case if the models do not overlap and may also be often the case even if the models overlap or are nested. Using Theorem 3.3 of Vuong (1989), which is valid under conditions clearly stated by this author, we obtain that in that case:

$$n^{1/2}[D(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n}) - \Delta(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n})] \xrightarrow{D} \quad \mathcal{N}(0, \omega_*^2),$$

where $\omega_*^2 = \text{var}\left[\log \frac{g^{\beta_0}(Y)}{h^{\gamma_0}(Y)}\right]$. A natural estimator of ω_*^2 is

$$\hat{\omega}_n^2 = n^{-1} \sum_{i=1}^{n} \left[\log \frac{g^{\hat{\beta}_n}(Y_i)}{h^{\hat{\gamma}_n}(Y_i)}\right]^2 - \left[n^{-1} \sum_{i=1}^{n} \log \frac{g^{\hat{\beta}_n}(Y_i)}{h^{\hat{\gamma}_n}(Y_i)}\right]^2.$$

From this we can compute the tracking interval (A_n, B_n), where $A_n = D(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n}) - z_{\alpha/2} n^{-1/2} \hat{\omega}_n$ and $B_n = D(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n}) + z_{\alpha/2} n^{-1/2} \hat{\omega}_n$, where $1 - \Phi(z_{\alpha/2}) = \alpha/2$ and Φ is the cdf of the standard normal variable. This interval has the property:

$$P_f[A_n < \Delta(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n}) < B_n] \longrightarrow 1 - \alpha,$$

where P_f represents the probability with density f.

We can also judge whether the values within the intervals correspond to large or small expected losses, according to the hint given by Commenges et al. (2007). This paper established a link between the value of $\text{KL}(g, f)$ and the relative error made in evaluating the typical set whose probability is underestimated using g rather than f, and used this to qualify KL values of 10^{-1}, 10^{-2},

10^{-3}, 10^{-4} as "large", "moderate", "small" and "negligible" respectively. As an example the KL divergence of a double exponential relative to a normal distribution with same mean and variance is of order 10^{-1} what may be called a "large" value. We may also measure on this scale the magnitude of the Akaike correction of $(p-q)/n$: for instance if we compare two models with $p-q=1$ and we have $n=100$ or with $p-q=5$ and we have $n=500$ the Akaike correction is 10^{-2} in both cases, a value qualified as "moderate"; as a matter of fact Akaike correction is rarely negligible in epidemiological studies. As already noted we can give an interpretation of EKL from (10.1) as the sum of the mis-specification risk $\mathrm{KL}(g^{\beta_0}, f)$ and the estimation risk, approximated by $p/2n$. For a well specified model the risk is about $p/2n$; for instance it is 10^{-2} if $p=10$ and $n=500$.

10.6.3 Extension to regression models

All that has been said can be extended to regression models $(g_{Y|X}) = (g^{\beta}_{Y|X}(.|.))_{\beta\in B}$ and $(h_{Y|X}) = (h^{\gamma}_{Y|X}(.|.))_{\gamma\in\Gamma}$. This can be done as in Vuong (1989) by directly defining the Kullback-Leibler divergence in term of conditional densities: $\mathrm{KL}(g^{\beta}_{Y|X}, f_{Y|X}) = \mathrm{E}_f[\log \frac{f_{Y|X}(Y|X)}{g^{\beta}_{Y|X}(Y|X)}]$, where the expectation is taken for the true distribution of the couple Y,X. However this approach has the drawback of requiring a new definition of the Kullback-Leibler divergence . The so-called reduced model approach (Commenges et al., 2007) is more satisfactory. Consider a sample of iid couples of variables $(Y_i,X_i), i=1,\ldots,n$ having joint pdf f, $f(y,x) = f_{Y|X}(y|x)\,f_X(x)$. Consider the model $(g) = (g^{\beta}(.,.))_{\beta\in B}$ such that $g^{\beta}(y,x) = g^{\beta}_{Y|X}(y|x)\,f_X(x)$; the model is called "reduced" because $f_X(.)$ is assumed known. The Kullback-Leibler divergence is:

$$\mathrm{KL}(g^{\beta}, f) = \mathrm{E}_f[\log f_{Y|X}(Y|X)] - \mathrm{E}_f[\log g^{\beta}_{Y|X}(Y,X)],$$

that is the term in $f_X(.)$ disappears (so that we do not need to know it in fact) and we get the same definition as in Vuong (1989) using only the conventional Kullback-Leibler divergence .

10.7 Application to logistic regression: a simulation study

As an illustration of this general procedure we will apply it to the problem of choice of the coding of an explanatory variable in logistic regression. We considered iid samples of size n of triples $(Y_i, x_1^i, x_2^i), i = 1, \ldots, n$ from the following distribution (which plays the role of the true distribution f). The conditional distribution of Y_i given (x_1^i, x_2^i) was logistic with $\text{logit}[f_{Y|X}(1|x_1^i, x_2^i)] = 0.5 + x_1^i + 2x_2^i$, where $f_{Y|X}(1|x_1^i, x_2^i) = P_*(Y_i = 1|x_1^i, x_2^i)$, P_* stands for the true probability; the marginal distributions of (x_1^i, x_2^i) were bivariate normal with zero expectation and variance equal to the identity matrix. We considered model (g) specified by $\text{logit}[g^{\beta}_{Y|X}(1|x_1^i, x_2^i)] = \beta_0 + \beta_1 x_1^i + \beta_2 x_2^i$, which was well specified and the (mis)-specified model (h) defined as $\text{logit}[h^{\gamma}_{Y|X}(1|x_1^i, x_2^i)] = \gamma_0 + \sum_{l=1}^{2} \gamma_l x_{1l}^i + \gamma_3 x_2^i$, where x_{1l}^i were dummy variables indicating in which categories x_1^i fell; the categories were defined using terciles of the observed distribution of x_1, and this was represented by two dummy variables: x_{11}^i indicating whether x_1^i fell in the first tercile or not, x_{12}^i indicating whether x_1^i fell in the second tercile or not.

Since model (g) is well specified we know that $g^{\beta_0} = f$, that the mis-specification error $\text{KL}(g^{\beta_0}, f)$ is zero and that $\text{Tr}(I_g^{-1} J_g) = p$. As for model (h) we must compute the quantities of interest by simulation. We can compute that in the logistic regression the l,k term of the matrix J_h is $\text{E}_f[x_l (Y - \frac{e^{x\gamma_0}}{1+e^{x\gamma_0}})^2 x_k]$, and that the l,k term of the matrix I_h is $\text{E}_f[x_l \frac{e^{x\gamma_0}}{(1+e^{x\gamma_0})^2} x_k]$. We estimated γ_0 by fitting model (h) on a simulated data set with $n = 10^5$. Our precise estimate $\check{\gamma}_0$ was thus $\hat{\gamma}_n$ for $n = 10^5$. We used it to precisely estimate J_h and I_h as $\check{I}_h = 10^{-5} \sum_{i=1}^{10^5} [x_l^i \frac{e^{x^i \check{\gamma}_0}}{(1+e^{x^i \check{\gamma}_0})^2} x_k^i]$ and $\check{J}_h = 10^{-5} \sum_{i=1}^{10^5} [x_l^i (Y_i - \frac{e^{x^i \check{\gamma}_0}}{1+e^{x^i \check{\gamma}_0}})^2 x_k^i]$. We estimated $\text{KL}(h^{\gamma_0}, f)$ by $10^{-5} \sum_{i=1}^{10^5} \log \frac{f_{Y|X}(Y_i|x_1^i, x_2^i)}{h^{\check{\gamma}_0}_{Y|X}(Y_i|x_1^i, x_2^i)}$. We also computed a precise estimate of ω_*^2, $\check{\omega}_*^2$, by the empirical variance of $\log \frac{f_{Y|X}(Y_i|x_1^i, x_2^i)}{h^{\check{\gamma}_0}_{Y|X}(Y_i|x_1^i, x_2^i)}$ computed on 10^5 replicas. Thus we can compute a precise estimate of $\text{EKL}(h^{\hat{\gamma}_n}, f)$ and $\text{EKL}(g^{\hat{\beta}_n}, f)$ by replac-

ing the terms on right-hand of (10.1) by their estimates. Because (g) is well specified we obtain immediately $\mathrm{EKL}(g^{\hat{\beta}_n}, f) \approx \frac{3}{2n}$; a precise estimate of $\mathrm{EKL}(g^{\hat{\beta}_n}, f) - \mathrm{EKL}(h^{\hat{\gamma}_n}, f)$ is thus given by $\check{\Delta} = \frac{3}{2n} - \mathrm{KL}(h^{\check{\gamma}_0}, f) - \frac{1}{2n}\mathrm{Tr}(\check{I}_h^{-1}\check{J}_h)$. We find first that $\mathrm{KL}(h^{\check{\gamma}_0}, f) \approx 7.28\ 10^{-3}$, a value approaching the "moderate magnitude". We found 3.998 and 3.999 for the values of $\mathrm{Tr}(\check{I}_h^{-1}\check{J}_h)$ for $n = 250$ and $n = 1000$ respectively. These values are very close to $q = 4$ (that would obtain if (h) was well-specified) so, in the following we will use this approximation. Using this approximation we can compute $\check{\Delta} = -\frac{1}{2n} - \mathrm{KL}(h^{\check{\gamma}_0}, f)$ and obtain $\check{\Delta} = -9.28\ 10^{-3}$ for $n = 250$ and $\check{\Delta} = -7.78\ 10^{-3}$ for $n = 1000$. We also find $\check{\omega}_*^2 = 1.44\ 10^{-2}$. We can then compute the standard error of D as $n^{-1/2}\check{\omega}_*$ and find $7.59\ 10^{-3}$ and $3.79\ 10^{-3}$ for $n = 250$ and $n = 1000$ respectively. We see at once that there is more chance that the tracking interval does not contain zero for $n = 1000$ than for $n = 250$.

We generated 1000 replications from the above model for $n = 250$ and $n = 1000$. For each replication we computed the maximum likelihood estimates and the AIC. We computed the histogram of $D(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n})$ (see Figure 1): its shape is approximately in accordance with the asymptotic normal distribution for both sample sizes; the empirical mean was $-9.50\ 10^{-3}$ and $-7.67\ 10^{-3}$ for $n = 250$ and $n = 1000$ respectively, close to the values of $\check{\Delta}$. The empirical variance of D (not shown) was in agreement with the theoretical variance computed from $\check{\omega}_*^2$. The mean of the estimated variances $\hat{\omega}_*^2$ was $1.88\ 10^{-2}$ and $1.54\ 10^{-2}$ for $n = 250$ and $n = 1000$ respectively, also reasonably close to the $\check{\omega}_*^2$. The proportion of replicas for which $\check{\Delta}$ was outside the .95 tracking interval was 0.045 and 0.053 for $n = 250$ and $n = 1000$ respectively. The proportion of replicas for which zero was outside of the tracking interval was 0.197 and 0.514 for $n = 250$ and $n = 1000$ respectively, and in all cases (g) was prefered to (h). These results are summarized in Table 1.

10.8 Choice of the best coding of age in a study of depression

10.8.1 The Paquid study

The application is based on the Paquid research programme (Letenneur et al., 1999), a prospective cohort study of mental and physical aging that evaluates social environment and health status. The target population consists of subjects aged 65 years and older living at home in southwestern France. The baseline variables registered included socio-demographic factors, medical history and psychometric tests. In particular the CESD scale for depression was completed. Here we illustrate the method of the paper by examining possible models of association of depression and BMI. As is conventional, depression was considered as a binary trait coded by a dichotomized version of the CESD (using the thresholds 17 and 23 for men and women respectively). We worked with the sample of the first visit of the Paquid study and we excluded the subjects who were diagnosed demented at that visit: the sample size was 3484. We fitted logistic regression models for explaining depression from BMI, age and gender. We entered age, gender and their interaction as explanatory variables. As for BMI which was the factor of main interest, we tried a linear (in the logistic scale) model and then we challenged the linear model by trying a categorization of BMI in terciles and a quadratic model. Both the tercile and the quadratic models have six parameters while the linear model has five. Note that the linear model is not nested in the tercile model while it is in the quadratic model.

The values of AIC, and the D statistic and tracking intervals (taking as reference the linear model) are given in Table 2. The tercile model had a larger AIC than the linear model but the point estimate (D) of the difference of risks was lower than 10^{-4} a level that we have qualified “negligible”, and zero was well inside the tracking interval. So from the point of view of Kullback-Leibler risk there was no evidence than one model is better than the other. When it comes to comparing the linear

and the quadratic model, because the first is nested in the second, we can use the likelihood ratio test: the null hypothesis is that the best distribution is in the linear sub-model. The hypothesis was strongly rejected ($p < 0.01$). We tend to conclude that the shape of the effect is not linear and that we may approach it better with a quadratic term. The point estimate of the difference of risks was 0.0007, a value which approaches the 10^{-3} level that we qualified to be a small (but not negligible) difference. The tracking interval was $[-0.0001; 0.0022]$ which includes zero, so we are not really sure to incur a smaller risk with the quadratic model. However we can correct the lower bound of the interval by the following argument. If $(g) \subset (h)$ we have that $\mathrm{KL}(g^{\beta_0}, f) \geq \mathrm{KL}(h^{\gamma_0}, f)$. Thus from equation (10.1), using the approximation $\mathrm{Tr}(I_g^{-1} J_g) \approx p$ we obtain $\Delta(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n}) \geq -\frac{1}{2n}(p-q)$. In our case we obtain $\Delta(g^{\hat{\beta}_n}, h^{\hat{\gamma}_n}) \geq -1.4\ 10^{-4}$. Thus the maximum increased risk in using the quadratic model is negligible. It may seem paradoxical (in view of the likelihood ratio test) that we can not assert with high probability that the estimator based on the quadratic model is better than that based on the linear model, but we must remember that the asymptotic law of the likelihood ratio we use is not the same as in the likelihood ratio test. The likelihood ratio test tells us that the quadratic model is *closer* than the linear model from the true distribution but it is still possible that we incur a larger risk when using the quadratic model estimator because of the increased statistical risk; however from the tracking interval we see that we are exposed to a negligible additional expected Kullback-Leibler risk when using the quadratic model while it is likely that it is in fact smaller. In conclusion, in this application there is no reason to prefer the tercile model to the linear model but there are some reasons to prefer the quadratic model to the linear model. Figure 1 shows the shape of the effect of BMI with the quadratic model, taking as reference the median BMI (equal to 24.2). This is a U-shaped curve yielding the lower risks of depression for medium values of the BMI, somewhat shifted however toward large BMI. Of course the epidemiological interpretation of this result is delicate and

the apparent effect that we have detected is the consequence of complex biological and psychological mechanisms that we do not attempt to explore here. Several other studies have found links between BMI and depression; see for instance Rantanen et al. (2000).

10.9 Discussion

We have proposed a statistic which tracks the difference of expected Kullback-Leibler risks between maximum likelihood estimators in two different models. Moreover we have an estimator of the variance of this statistic and we can construct a "tracking interval". *In fine* we can do more than simply choosing the estimator which has the lowest AIC. We can estimate the difference of risks. This difference of risk has the same meaning in different problems and we may become accustomed to considering differences of 10^{-2}, 10^{-3}, 10^{-4} as moderate, small and negligible respectively, as we are accustomed to interpret correlation coefficients or odds-ratios for instance.

A more complex and related problem occurs if we try a large number of models. In that case we have a family of estimators $g_1^{\hat{\beta}_{1n}}, \ldots, g_K^{\hat{\beta}_{Kn}}$. We may first compute the AIC of the $g_k^{\hat{\beta}_{kn}}$; let us call $g_{k_0}^{\hat{\beta}_{k_0 n}}$ the estimator with the smallest AIC. For the other estimators we may compute $D(g_k^{\hat{\beta}_{kn}}, g_{k_0}^{\hat{\beta}_{k_0 n}})$. Of course the $D(g_k^{\hat{\beta}_{kn}}, g_{k_0}^{\hat{\beta}_{k_0 n}})$ are correlated and a confidence interval has to take into account this correlation as well as the multiple testing issue (Edwards and Hsu, 1983; Hsu, 1984). Shimodaira (2001) has proposed an interesting approach to this problem, leading to define a set of admissible models.

REFERENCES

Akaike, H. (1973). Information theory and an extension of maximum likelihood principle, Second International Symposium on Information Theory, Akademia Kiado, 267-281.

Bengtsson, T. and Cavanaugh, J.E. (2006). An improved Akaike information criterion for state-space model selection. *Computational Statistics and Data Analysis* **50**, 2635-2654.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *J. Math. Psych.* **44**, 62-91.

Commenges, D., Joly, P, Gégout-Petit, A. and Liquet, B. (2007). Choice between semi-parametric estimators of Markov and non-Markov multi-state models from generally coarsened observations. *Scandinavian Journal of Statistics*, **34**, 33-52.

deLeuwe, J. (1992). Introduction to Akaike (1973) Information theory and an extension of the maximum likelihood principle, in *Breakthroughs in Statistics* (Kotz, S. and Johnson, N.L., Eds), New-York: Springer.

Edwards, D.G. and Hsu, J.C. (1983). Multiple Comparisons with the best treatment, *Journal of the American Statistical Association* **78**, 965-971

Hsu, J.C. (1984). Constrained simultaneous confidence intervals for multiple comparisons with the best. *Annals of Statistics* **12**, 1136-1144.

Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika* **83**, 875-890

Kullback, S. (1968). *Information Theory and Statistics*, New York: Dover.

Letenneur, L., Gilleron, V., Commenges, D., Helmer, C., Orgogozo, JM. and Dartigues, JF. (1999). Are sex and educational level independent predictors of dementia and Alzheimer's disease ? Incidence data from the PAQUID project. *Journal of Neurology Neurosurgery and Psychiatry* **66**,

177-183.

Linhart, H. and Zucchini,W. (1986). *Model Selection*, New York: Wiley.

Liquet, B., Sakarovitch, C. and Commenges, D. (2003). Bootstrap choice of estimators in parametric and semi-parametric families: an extension of EIC *Biometrics* **59**, 172-178.

Rantanen T, Penninx BW, Masaki K, Lintunen T, Foley D, Guralnik JM. (2000). Depressed mood and body mass index as predictors of muscle strength decline in old men. *J Am Geriatr Soc.* **6**, 613-617.

Schwarz, G. (1978). Estimating the dimension of a model, *Ann. Statist.* **6**, 461-464.

Shibata, R. (1997). Bootstrap estimate of Kullback-Leibler information for model selection, *Statist. Sin.* **7**, 375-394.

Shimodaira, H. (2001). Multiple Comparisons Of Log-Likelihoods And Combining Nonnested Models With Applications To Phylogenetic Tree Selection. *Commun. Statist.* **30**, 1751, 1772.

Vuong, Q.H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica* **57**, 307-333

Table 8.1: Simulation study: choice between tercile and linear model for the explanatory variable in a logistic regression model

n	$\check{\Delta}$	$\bar{D}$	$\bar{\hat{\omega}}^2$	Coverage rate	Power
250	$-9.28\ 10^{-3}$	$-9.50\ 10^{-3}$	$1.88\ 10^{-2}$	0.955	0.197
1000	$-7.78\ 10^{-3}$	$-7.67\ 10^{-3}$	$1.54\ 10^{-2}$	0.947	0.514

Table 8.2: Application: comparison of the linear, tercile and quadratic models for the effect of BMI on depression: D and the tracking interval for the difference of Kullback-Leibler risks are with respect to the linear model.

Model	# parameters	Likelihood	AIC	D	Tracking interval
Linear	5	-1346.25	2702.5	-	-
Tercile	6	-1345.60	2703.2	-0.0001	$[-0.0009; 0.0007]$
quadratic	6	-1342.93	2697.9	0.0007	$[-0.0001; 0.0022]$

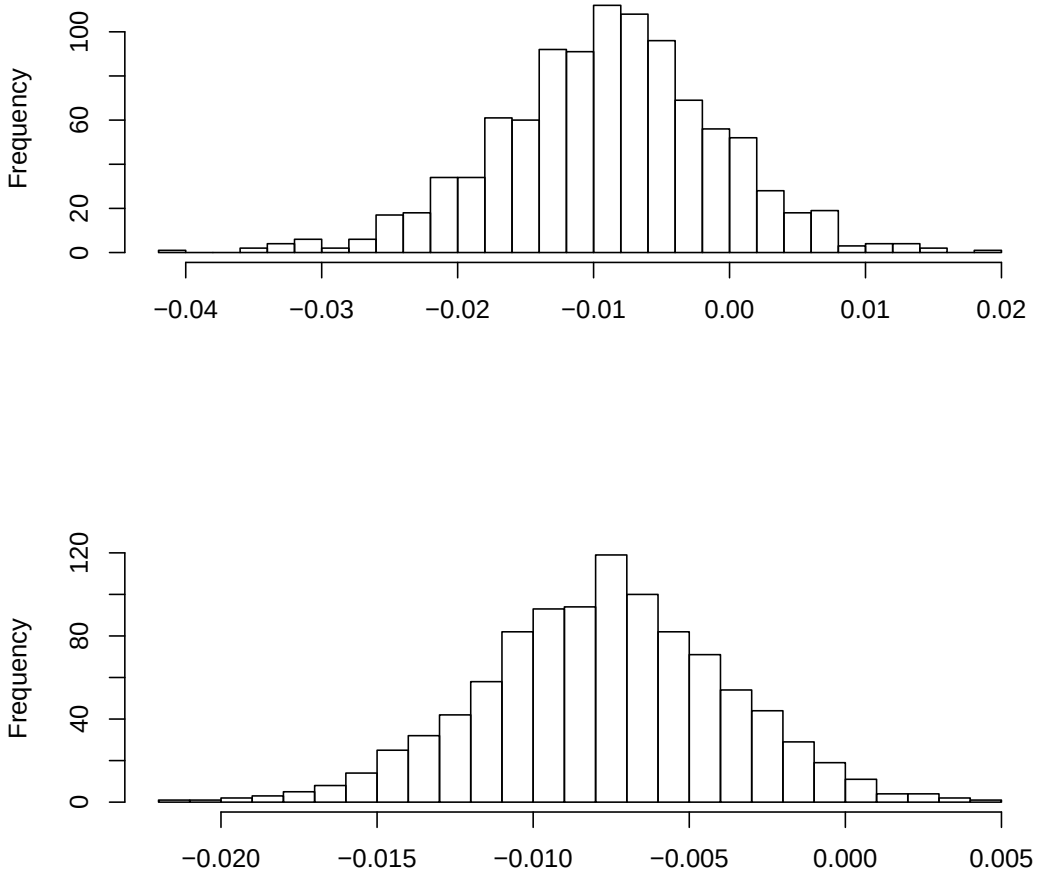


Figure 8.1: Histogram of the values of D (which estimates the difference of Kullback-Leibler risks between the tercile and the linear models) in the simulation: upper figure, $n = 250$, lower figure, $n = 1000$.

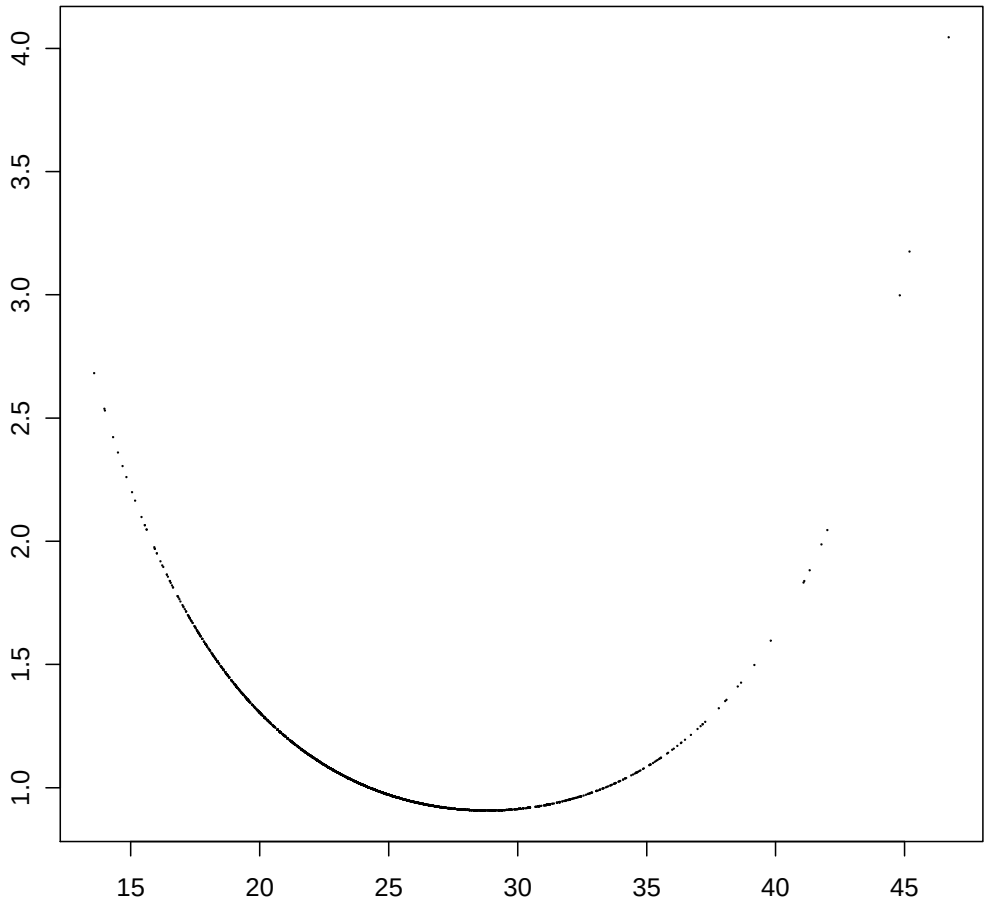


Figure 8.2: Estimated “effect” of the BMI on depression in the quadratic model: odds-ratios with respect to the probability at the median of BMI (24.2); the dots have for abscissas the observed BMI values.