ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ

ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ

ΕΡΓΑΣΤΗΡΙΟ ΒΙΟΜΑΘΗΜΑΤΙΚΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ (ΠΜΣ):

«ΜΕΘΟΔΟΛΟΓΙΑ ΒΙΟΪΑΤΡΙΚΗΣ ΕΡΕΥΝΑΣ, ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΚΛΙΝΙΚΗ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ»

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΔΙΑΤΡΙΒΗ

“Εφαρμογές του μοντέλου Vuong ως μεθόδου επιλογής”

“Applications of Vuong's model selection method”

Η Τριμελής Επιτροπή,

Μπατσίδης Απόστολος, Επίκουρος Καθηγητής του Πανεπιστημίου Ιωαννίνων (Επιβλέπων).

Στεφανίδης Ιωάννης, Καθηγητής του Τμήματος Ιατρικής του Πανεπιστημίου Θεσσαλίας.

Δοξάνη Χρυσούλα, Επιστημονικός Συνεργάτης του Τμήματος Ιατρικής του Πανεπιστημίου Θεσσαλίας

Χρήστος Ι. Κοκκότης

ΛΑΡΙΣΑ, 2017

Περιεχόμενα

Κεφάλαιο 1 Περίληψη

Σκοπός της παρούσας Μεταπτυχιακής Διπλωματικής Διατριβής είναι αρχικά να περιγραφεί, παρουσιασθεί η διαδικασία ελέγχου που έχει προταθεί στη βιβλιογραφία από τον Vuong (1989) για την επιλογή κατάλληλου μοντέλου μεταξύ δύο πιθανών και στη συνέχεια να εφαρμοστεί σε ιατρικά σύνολα δεδομένων. Η διαδικασία του Vuong (1989) είναι από τις πλέον διαδομένες μεθοδολογίες επιλογής μοντέλου έχοντας λάβει περισσότερες από 4000 αναφορές, ενώ έχει επεκταθεί από διάφορους ερευνητές (βλέπε, μεταξύ άλλων, Jimenez-Gamero et al. (2016), Jimenez-Gamero and Batsidis (2017)). Σε αυτό το πλαίσιο η διάρθρωση της διπλωματικής διατριβής είναι η ακόλουθη. Στο Κεφάλαιο 2 Εισαγωγή παρουσιάζεται το υπό μελέτη πρόβλημα και αναφέρονται βασικές έννοιες για την κατανόησή του. Στο Κεφάλαιο 3 Κανόνες Απόφασης παρατίθεται η μέθοδος του Vuong (1989) και οι κανόνες απόφασης που προτάθηκαν από αυτόν στην ειδική περίπτωση της επιλογής μεταξύ δύο μη εμφωλευμένων μοντέλων. Για την καλύτερη κατανόηση της μεθόδου δίνεται η θεωρητική του εφαρμογή στην ειδική περίπτωση που τα ανταγωνιστικά μοντέλα είναι η λογαριθμοκανονική και η εκθετική κατανομή. Στο Κεφάλαιο 4 Αριθμητικές Εφαρμογές εφαρμόζεται η διαδικασία ελέγχου του Vuong (1989) σε ιατρικά σύνολα δεδομένων μέσω του λογισμικού Spyder σε γλώσσα προγραμματισμού Python 2.7. Η διατριβή ολοκληρώνεται με το Κεφάλαιο 5 Επίλογος, όπου αναφέρονται περιπτώσεις κακής χρήσης της μεθοδολογίας καθώς και ειδικές περιπτώσεις όπου η μεθοδολογία δε μπορεί να εφαρμοστεί. Επιπλέον, δίνεται μια σύντομη αναφορά σε εναλλακτικές προσεγγίσεις, που έχουν εμφανιστεί στη βιβλιογραφία και σε θέματα για περαιτέρω έρευνα.

Chapter 1 Abstract

The purpose of this postgraduate dissertation is initially to describe the hypothesis testing procedure which proposed by Vuong (1989) for the selection of a suitable model between two possible. Afterwards the procedure is illustrated via medical data sets. Vuong's process (1989) is one of the most widely used modeling methodologies having received more than 4.000 citations and has been extended by several researchers (see Jimenez-Gamero et al. (2016), Jimenez-Gamero and Batsidis (2017)). In this context the structure of the dissertation is as follows. Chapter 2 Introduction presents the problem under study and sets out basic concepts for understanding. Chapter 3 Decision Rules describes Vuong's (1989) method and the decision rules proposed by him in the special case of selecting non-nested models. For a

better understanding of the method, the theoretical application is given in the specific case where the competitive models are the log-normal and the exponential distribution. In Chapter 4 Numerical Applications, Vuong's (1989) hypothesis testing procedure is applied to medical data sets through the Spyder software in Python 2.7. The dissertation ends with Chapter 5 Epilogue, which mentions cases of misuse of the methodology and special cases where the methodology cannot be applied. In addition, a brief reference is made to alternative approaches, which have been appeared in the literature and to some open problems for future research.

Κεφάλαιο 2 Εισαγωγή

Σε πολλούς τομείς της καθημερινότητας προκύπτει η ανάγκη της εξήγησης ενός τυχαίου φαινομένου. Η εξήγηση αυτή πολλές φορές δεν είναι καθόλου εύκολη, διότι το φαινόμενο μπορεί να εξαρτάται από πολλές παραμέτρους, οι οποίες συνδυάζονται σε κάποιο πολύπλοκο μοντέλο με συνέπεια να μην είναι εφικτή η επιβεβαίωση. Για το λόγο αυτό οι ερευνητές υποθέτουν ή υιοθετούν πιθανά μοντέλα και μέσα από διαδικασίες ελέγχου προσπαθούν να επιλέξουν αυτό που είναι πιο "κοντά" στο άγνωστο πραγματικό μοντέλο, έτσι ώστε να επιλέξουν το καλύτερο δυνατό μοντέλο. Η αναγκαιότητα αυτή είχε ως συνέπεια να προταθούν στη βιβλιογραφία πολλά κριτήρια με αντικειμενικό στόχο την καλύτερη επιλογή του μοντέλου που εξηγεί καλύτερα τις παρατηρήσεις μας.

Η παρούσα μεταπτυχιακή διατριβή εστιάζει στην παρουσίαση και στην εφαρμογή της πιο γνωστής μεθόδου επιλογής μοντέλου που εντάσσεται στους στατιστικούς ελέγχους για την επιλογή του καλύτερου μοντέλου. Ειδικότερα, θα παρουσιαστεί η μεθοδολογία που προτάθηκε από τον Vuong (1989) για την αντιμετώπιση του ακόλουθου προβλήματος. Έστω $X_1,\ldots,X_n$ ένα τυχαίο δείγμα (τ.δ.) από κάποιο πληθυσμό με συνάρτηση πυκνότητας πιθανότητας (σ.π.π.) ή συνάρτηση πιθανότητας (σ.π.) $h(\cdot)$, η συναρτησιακή μορφή της οποίας μας είναι άγνωστη. Θέλουμε να εξετάσουμε αν μπορεί να θεωρηθεί ότι το τ.δ. προέρχεται είτε από την οικογένεια κατανομών $H_f=\{f(x,\boldsymbol{\theta}),\boldsymbol{\theta}\in\Theta\}$ είτε από την οικογένεια κατανομών $H_g=\{g(x,\boldsymbol{\gamma}),\boldsymbol{\gamma}\in\Gamma\}$, όπου οι συναρτησιακές μορφές των f και g μας είναι γνωστές, αλλά εξαρτώνται από έναν πεπερασμένο αριθμό άγνωστων παραμέτρων, που συμβολίζονται με τα διανύσματα $\boldsymbol{\theta}$ και $\boldsymbol{\gamma}$, αντίστοιχα. Τα σύνολα Θ και Γ απαρτίζουν τον παραμετρικό χώρο, όπου οι αντίστοιχες πυκνότητες $f(x,\boldsymbol{\theta})$ και $g(x,\boldsymbol{\gamma})$

είναι καλά ορισμένες. Στο σημείο αυτό θα πρέπει να αναφερθεί ότι οι δύο οικογένειες κατανομών F και G μπορεί να είναι είτε μη εμφωλευμένες ή ξεχωριστές (non-nested or separated) είτε επικαλυπτόμενες είτε η μία από τις δύο οικογένειες εμφωλευμένη στην άλλη. Στη συνέχεια δίνεται η μη αυστηρή επεξήγηση των όρων αυτών, καθώς θέλουμε να αποφύγουμε τους αυστηρούς μαθηματικούς ορισμούς που έχουν δοθεί από τον Pesaran (1987).

Λέμε ότι δύο οικογένειες κατανομών είναι μη εμφωλευμένες όταν κανένα μοντέλο της μιας οικογένειας δεν μπορεί να προκύψει από την άλλη, είτε με περιορισμούς στις παραμέτρους είτε μέσω μια οριακής διαδικασίας, δηλαδή όταν $F \cap G = \emptyset$. Λέμε ότι δύο οικογένειες κατανομών είναι επικαλυπτόμενες αν $F \cap G \neq \emptyset$ και επιπλέον καμία δεν είναι υποσύνολο της άλλης. Τέλος, λέμε ότι η F είναι εμφωλευμένη στην G αν $F \subseteq G$.

Οι κανόνες απόφασης που προτάθηκαν από τον Vuong (1989) αντιμετωπίζουν και τις τρεις περιπτώσεις. Η παρούσα μεταπτυχιακή διατριβή εστιάζει στη μεθοδολογία του Vuong (1989) για την επιλογή αυστηρά μη εμφωλευμένων μοντέλων (strictly non-nested models) και συγκεκριμένα στην περιγραφή και ανάλυση της μεθοδολογίας και τυχόν αδυναμίες εφαρμογής. Οι κανόνες απόφασης εφαρμόζονται τόσο σε θεωρητικό υπόβαθρο μέσω της μελέτης μίας ειδικής περίπτωσης όσο και μέσω της εφαρμογής τους σε σύνολα ιατρικών δεδομένων με τη βοήθεια της Python 2.7 στο λογισμικό Spyder.

Σε αυτό το στατιστικό πλαίσιο, ο Vuong (1989) πρότεινε κανόνες απόφασης για τον έλεγχο της υπόθεσης ότι τα δύο μοντέλα είναι ισοδύναμα έναντι της εναλλακτικής ότι ένα από τα δύο μοντέλα είναι καλύτερο από το άλλο για τη μοντελοποίηση των δεδομένων μας. Η μεθοδολογία του βασίζεται στην ακόλουθη ιδέα. Τα δύο μοντέλα F και G θεωρούνται ισοδύναμα αν είναι εξίσου "κοντά" στον άγνωστο πληθυσμό με συνάρτηση πυκνότητας πιθανότητας (σ.π.π.) ή συνάρτηση πιθανότητας (σ.π.) $h(\cdot)$, η συναρτησιακή μορφή της οποίας μας είναι άγνωστη, ενώ η οικογένεια κατανομών F (αντίστοιχα G) μοντελοποιεί καλύτερα το φυσικό πρόβλημα αν είναι "πιο κοντά" στον άγνωστο πληθυσμό από την G (αντίστοιχα από την F). Από τα παραπάνω γίνεται αντιληπτό ότι το πρώτο πρόβλημα που προκύπτει είναι πως θα μετρηθεί η εγγύτητα δύο πληθυσμών. Η αναφερόμενη εγγύτητα των δύο κατανομών είναι δυνατό να μετρηθεί με διάφορους τρόπους, υιοθετώντας διαφορετικά μέτρα της απόστασής τους ή της απόκλισής τους. Ο Vuong (1989) χρησιμοποίησε ως μέτρο της εγγύτητας δύο κατανομών το μέτρο απόκλισης των Kullback and Leibler (1951). Στη συνέχεια για λόγους πληρότητας παρατίθεται ο ορισμός του.

Ορισμός 1 Έστω F και G δύο κατανομές, με σ.π. ή σ.π.π. f(x,θ) και g(x,γ), αντίστοιχα. Η Kullback-Leibler απόκλιση (βλέπε και Kullback, 1959) της G ως προς την F ορίζεται ως

$$I_{fg}(\theta,\gamma) = E_\theta\left[\ln\frac{f(X,\theta)}{g(X,\gamma)}\right],$$

όπου με $E_\theta(\cdot)$ συμβολίζεται η αναμενόμενη τιμή υπό την οικογένεια κατανομών F.

Συνεπώς, στην περίπτωση συνεχών δεδομένων, το μέτρο απόκλισης των Kullback-Leibler της G ως προς την F γράφεται, υπό την προϋπόθεση ότι το ολοκλήρωμα υπάρχει, ως

$$I_{fg}(\theta,\gamma) = \int_{R_f} \ln\left[\frac{f(x,\theta)}{g(x,\gamma)}\right] f(x,\theta)\,dx,$$

όπου R_f είναι το πεδίο ορισμού της τυχαίας μεταβλητής X υπό την οικογένεια κατανομών F.

Από τη συζήτηση που προηγήθηκε και λαμβάνοντας υπόψη τον Ορισμό 1 η οικογένεια κατανομών H_f θεωρείται ότι περιγράφει καλύτερα το τ.δ. αν και μόνο αν η ελάχιστη πιθανή «απόσταση» μεταξύ του μοντέλου F και της αληθινής κατανομής είναι μικρότερη από την αντίστοιχη ελάχιστη πιθανή «απόσταση» μεταξύ του μοντέλου G και της αληθινής κατανομής. Σε μαθηματικούς όρους αυτό διατυπώνεται

η οικογένεια κατανομών H_f θεωρείται ότι περιγράφει καλύτερα το τ.δ. αν και μόνο

$$\text{αν } \inf_\theta I_{fh} < \inf_\gamma I_{gh},$$

ή ισοδύναμα αν και μόνο αν

$$\sup_\theta E_h\left[\ln f\left(X,\boldsymbol{\theta}\right)\right] > \sup_\gamma E_h\left[\ln g\left(X,\boldsymbol{\gamma}\right)\right].$$

Στο σημείο αυτό υποθέτουμε ότι υπάρχει και είναι μοναδική μια τιμή της παραμέτρου θ (αντίστοιχα γ) για την οποία ελαχιστοποιείται η απόσταση μεταξύ της F (αντίστοιχα G) και της αληθινής κατανομής. Ουσιαστικά η υπόθεση αυτή σημαίνει ότι υπάρχει ένα και μόνο ένα μέλος της οικογένειας F (της οικογένειας G αντίστοιχα) που βρίσκεται πλησιέστερα στην άγνωστη h. Οι τιμές αυτές συμβολίζονται με $\boldsymbol{\theta}_*$ και με $\boldsymbol{\gamma}_*$, αντίστοιχα, δηλαδή είναι

$$\boldsymbol{\theta}_* = \arg\max_{\boldsymbol{\theta}\in\Theta} E_h\left[\ln f\left(X,\boldsymbol{\theta}\right)\right],$$

και

$$\boldsymbol{\gamma}_* = \arg\max_{\boldsymbol{\gamma}\in\Gamma} E_h\left[\ln g\left(X,\boldsymbol{\gamma}\right)\right],$$

και είναι γνωστές ως οι ψευδοαληθείς τιμές των $\boldsymbol{\theta}$ και $\boldsymbol{\gamma}$, αντίστοιχα.

Από τα παραπάνω γίνεται αντιληπτό ότι το τ.δ. προέρχεται από την οικογένεια κατανομών για την οποία η ποσότητα $E_h\left[\ln f\left(X,\boldsymbol{\theta}_*\right)\right]$ ή $E_h\left[\ln g\left(X,\boldsymbol{\gamma}_*\right)\right]$ είναι μέγιστη. Επομένως,

i) αν $E_h\left[\ln f\left(X,\boldsymbol{\theta}_*\right)\right] = E_h\left[\ln g\left(X,\boldsymbol{\gamma}_*\right)\right]$ ή $E_h\left[\ln \frac{f(X,\boldsymbol{\theta}_*)}{g(X,\boldsymbol{\gamma}_*)}\right] = 0$, τότε οι δύο κατανομές περιγράφουν ισοδύναμα το τ.δ.,

ii) αν $E_h\left[\ln f\left(X,\boldsymbol{\theta}_*\right)\right] > E_h\left[\ln g\left(X,\boldsymbol{\gamma}_*\right)\right]$ ή $E_h\left[\ln \frac{f(X,\boldsymbol{\theta}_*)}{g(X,\boldsymbol{\gamma}_*)}\right] > 0$, τότε το τ.δ περιγράφεται καλύτερα από την κατανομή F, και τέλος,

iii) αν $E_h\left[\ln f\left(X,\boldsymbol{\theta}_*\right)\right] < E_h\left[\ln g\left(X,\boldsymbol{\gamma}_*\right)\right]$ ή $E_h\left[\ln \frac{f(X,\boldsymbol{\theta}_*)}{g(X,\boldsymbol{\gamma}_*)}\right] < 0$, τότε το τ.δ περιγράφεται καλύτερα από την κατανομή G.

Με βάση το παραπάνω σκεπτικό, ο Vuong (1989) ανήγαγε το υπό μελέτη πρόβλημα στον έλεγχο της $H_0 : E_h\left[\ln \frac{f\left(X,\boldsymbol{\theta}_*\right)}{g\left(X,\boldsymbol{\gamma}_*\right)}\right] = 0$, έναντι της αντίστοιχης κάθε φορά εναλλακτικής υπόθεσης. Η στατιστική συνάρτηση που προτάθηκε από τον Vuong (1989) για τον έλεγχο της παραπάνω υπόθεσης καθώς και οι κανόνες απόφασης όταν οι οικογένειες κατανομών είναι μη εμφωλευμένες αποτελούν αντικείμενο μελέτης του επόμενου κεφαλαίου.

Κεφάλαιο 3 Κανόνες Απόφασης

Ο στόχος του κεφαλαίου αυτού είναι διττός. Από τη μια μεριά θα παρουσιαστεί η στατιστική συνάρτηση που προτάθηκε από τον Vuong (1989) για τον έλεγχο της υπόθεσης ότι τα δύο ανταγωνιστικά μοντέλα είναι ισοδύναμα καθώς και ο κανόνας απόφασης. Στη συνέχεια, για την καλύτερη κατανόηση του κανόνα θα παρουσιαστεί η ειδική περίπτωση που οι δύο ανταγωνιστικές οικογένειες είναι η λογαριθμοκανονική και η εκθετική κατανομή.

Από όσα προηγήθηκαν προέκυψε ότι το υπό μελέτη πρόβλημα έχει αναχθεί στον έλεγχο της $H_0 : E_h\left[\ln \frac{f\left(X,\boldsymbol{\theta}_*\right)}{g\left(X,\boldsymbol{\gamma}_*\right)}\right] = 0$. Είναι εύκολα αντιληπτό ότι η ποσότητα

$E_h\left[\ln\frac{f\left(X,\boldsymbol{\theta}_*\right)}{g\left(X,\boldsymbol{\gamma}_*\right)}\right]$ είναι άγνωστη και για τη διενέργεια του υπό μελέτη ελέγχου θα πρέπει να χρησιμοποιηθεί ένας εκτιμητής της, ο οποίος θα ήταν επιθυμητό να έχει κάποιες καλές στατιστικές ιδιότητες π.χ. να είναι συνεπής και επιπλέον να μπορεί να προσδιοριστεί η κατανομή του υπό τη μηδενική υπόθεση. Υπό την προϋπόθεση ότι πληρούνται οι συνθήκες [1] που παρατίθενται από τον Vuong (1989), η άγνωστη αυτή ποσότητα μπορεί να εκτιμηθεί με συνέπεια από τη στατιστική συνάρτηση

$$\frac{1}{n}LR_n\left(\widehat{\boldsymbol{\theta}}_n,\widehat{\boldsymbol{\gamma}}_n\right)=\frac{1}{n}\sum_{i=1}^{n}\ln\frac{f\left(X_i,\widehat{\boldsymbol{\theta}}_n\right)}{g\left(X_i,\widehat{\boldsymbol{\gamma}}_n\right)}$$

όπου $\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\gamma}}_n$ οι Ε.Μ.Π. των $\boldsymbol{\theta}_*$ και $\boldsymbol{\gamma}_*$, αντίστοιχα. Επομένως, η άγνωστη ποσότητα εκτιμάται με συνέπεια από τη δειγματική μέση τιμή του πηλίκου μέγιστων πιθανοφανειών, ενώ η ασυμπτωτική της κατανομή της προσδιορίζεται στο επόμενο θεώρημα.

Θεώρημα (Vuong (1989)) Υπό τις υποθέσεις Α1-Α7 που δίδονται από τον Vuong (1989)

$$n^{-\frac{1}{2}}LR_n\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\gamma}}_n\right)-n^{\frac{1}{2}}E_h\left[\ln\frac{f\left(X,\boldsymbol{\theta}_*\right)}{g\left(X,\boldsymbol{\gamma}_*\right)}\right]\xrightarrow{\kappa.\kappa.}N\left(0,\omega_*^2\right),$$

όπου

$$\omega_*^2=Var_h\left[\ln\frac{f\left(X,\boldsymbol{\theta}_*\right)}{g\left(X,\boldsymbol{\gamma}_*\right)}\right]=E_h\left[\ln\frac{f\left(X,\boldsymbol{\theta}_*\right)}{g\left(X,\boldsymbol{\gamma}_*\right)}\right]^2-\left\{E_h\left[\ln\frac{f\left(X,\boldsymbol{\theta}_*\right)}{g\left(X,\boldsymbol{\gamma}_*\right)}\right]\right\}^2,$$

με $\omega_*^2<\infty$

Είναι άμεσα αντιληπτό ότι η διακύμανση αυτή δεν μπορεί να υπολογιστεί στην πράξη, καθώς εξαρτάται από την άγνωστη αληθινή κατανομή $h\left(\cdot\right)$. Επομένως, θα πρέπει η άγνωστη διακύμανση ω_*^2 να εκτιμηθεί κατάλληλα, έτσι ώστε η δειγματική μέση τιμή του πηλίκου μέγιστων πιθανοφανειών να αποτελέσει τη βάση ενός στατιστικού ελέγχου. Ο Vuong (1989) πρότεινε δύο πιθανούς τρόπους συνεπούς εκτίμησης, αλλά στη συνέχεια παρουσιάζεται αυτός που χρησιμοποιείται συνηθέστερα σε πρακτικές εφαρμογές. Ειδικότερα,

[1] Οι συνθήκες αυτές δεν παρατίθενται καθώς είναι τεχνικές μαθηματικές και παραπέμπουμε σχετικά στην εργασία του Vuong (1989). Δυστυχώς οι συνθήκες αυτές δεν πληρούνται πάντοτε και η μη επαλήθευσή τους οδηγεί σε λανθασμένες αποφάσεις σε πρακτικές εφαρμογές. Περισσότερα θα αναφερθούν στον επίλογο αυτής της διατριβής.

η άγνωστη διακύμανση εκτιμάται από την αντίστοιχη δειγματική ποσότητα, που ορίζεται στη σχέση:

$$\widehat{\omega}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left[ln\frac{f(X_i,\hat{\theta}_n)}{g(X_i,\hat{\gamma}_n)}\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n} ln\frac{f(X_i,\hat{\theta}_n)}{g(X_i,\hat{\gamma}_n)}\right]^2.$$

Επομένως, κατά τον συνδυασμό των παραπάνω αποτελεσμάτων για τον έλεγχο της υπό μελέτης μηδενικής υπόθεσης, η στατιστική συνάρτηση που προκύπτει είναι η

$$T_\nu = n^{-\frac{1}{2}}\frac{LR_n(\hat{\theta}_n,\hat{\gamma}_n)}{\widehat{\omega}_n},$$

για την οποία ισχύουν τα ακόλουθα:

- Υπό την μηδενική υπόθεση ότι τα δύο μοντέλα είναι ισοδύναμα:

$$T_\nu = n^{-\frac{1}{2}}\frac{LR_n(\hat{\theta}_n,\hat{\gamma}_n)}{\widehat{\omega}_n} \xrightarrow{\kappa.\kappa} N(0,1)$$

- Υπό την υπόθεση ότι το τυχαίο δείγμα περιγράφεται πιο ικανοποιητικά από την F

$T_\nu \xrightarrow{\sigma.\beta.} +\infty$, δηλαδή λαμβάνει μεγάλες τιμές

- Υπό την υπόθεση ότι το τυχαίο δείγμα περιγράφεται πιο ικανοποιητικά από την G

$T_\nu \xrightarrow{\sigma.\beta.} -\infty$, δηλαδή λαμβάνει μικρές τιμές

Κανόνας Απόφασης

Έστω T_ν^{obs} η παρατηρούμενη τιμή της στατιστικής συνάρτησης. Με επίπεδο σημαντικότητας α και συμβολίζοντας με Z_α, όπου Z_α το σημείο εκείνο όπου

$$P(Z > Z_\alpha) = \alpha$$

και συνοψίζοντας τα παραπάνω προκύπτουν τα ακόλουθα:

- Αν $T_v^{obs} > Z_\alpha$, απορρίπτεται η μηδενική υπόθεση ότι οι δύο οικογένειες κατανομών περιγράφουν ισοδύναμα το τυχαίο δείγμα και το τ.δ. προέρχεται από την κατανομή F.
- Αν $T_v^{obs} < -Z_\alpha$, απορρίπτεται η μηδενική υπόθεση ότι οι δύο οικογένειες κατανομών περιγράφουν ισοδύναμα το τυχαίο δείγμα και το τ.δ. προέρχεται από την κατανομή G.
- Αν $\left|T_v^{obs}\right| \leq Z_{\alpha/2}$, δε δύναται να διακρίνουμε ποιά από τις δύο οικογένειες κατανομών μοντελοποιεί καλύτερα τις δειγματικές παρατηρήσεις.

Λογαριθμοκανονική-Εκθετική Κατανομή

Στην ενότητα αυτή αποσαφηνίζεται η μεθοδολογία του Vuong (1989) στην περίπτωση που τα δύο «ανταγωνιστικά» μοντέλα είναι αυτό της λογαριθμοκανονικής και της εκθετικής κατανομής (βλέπε σχετικά και Μπαρδάκας, 2013) .

Ορισμός 2 Η τ.μ. X ακολουθεί λογαριθμοκανονική κατανομή με παράμετρο $\boldsymbol{\theta}=\left(\theta_1,\theta_2\right)$, $\theta_1 \in R, \theta_2 > 0$, αν οι δυνατές τιμές της είναι $x > 0$ και η σ.π.π. της δίνεται από τη σχέση

$$f(x,\boldsymbol{\theta}) = \frac{\exp\left[-\frac{\left(\ln x-\theta_1\right)^2}{2\theta_2}\right]}{x\left(2\pi\theta_2\right)^{1/2}},\ x>0,\ \boldsymbol{\theta}=\left(\theta_1,\theta_2\right),\ \theta_1 \in R, \theta_2 > 0.$$

Στην περίπτωση αυτή γράφουμε $X \sim \ln N\left(\theta_1,\theta_2\right).$

Ο λογάριθμος της σ.π.π. της λογαριθμοκανονικής κατανομής δίνεται από τη σχέση

$$\ln f(\mathrm{x},\boldsymbol{\theta}) = -\ln \mathrm{x} - \frac{1}{2}\ln(2\pi\theta_2) - \frac{(\ln \mathrm{x}-\theta_1)^2}{2\theta_2}$$

Ορισμός 3 Η τ.μ. X ακολουθεί εκθετική κατανομή με παράμετρο γ, $\gamma > 0$, αν οι δυνατές τιμές της είναι $x \geq 0$ και η σ.π.π. της δίνεται από τη σχέση:

$$g\left(x,\gamma\right)=\gamma^{-1}e^{-x/\gamma},\ x \geq 0,\ \gamma > 0.$$

Στην περίπτωση αυτή γράφουμε $X \sim Exp\left(1/\gamma\right).$

Ο λογάριθμος της σ.π.π της $X \sim Exp\left(1/\gamma\right)$ δίνεται από τη σχέση

$$\ln g\left(X,\gamma\right)=-\ln\gamma-\frac{X}{\gamma}.$$

Για τη θεωρητική εφαρμογή της μεθοδολογίας του Vuong (1989) σε αυτήν την ειδική περίπτωση ουσιαστικά απαιτείται η υλοποίηση των ακόλουθων βημάτων.

Βήμα 1ο: Εύρεση των Ε.Μ.Π. $\hat{\boldsymbol{\theta}}_n$ και $\hat{\gamma}_n$. Οι Ε.Μ.Π. $\widehat{\boldsymbol{\theta}}_n = (\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n})$ και $\hat{\gamma}_n$, αντίστοιχα, των παραμέτρων $\boldsymbol{\theta}=\left(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\right)$ και γ είναι

$$\hat{\theta}_{1n}=\frac{1}{n}\sum_{i=1}^{n}\ln X_i\,,\qquad \hat{\theta}_{2n}=\frac{1}{n}\sum_{i=1}^{n}\left(\ln X_i-\hat{\theta}_{1n}\right)^2$$

και

$$\hat{\gamma}_n=\frac{1}{n}\sum_{i=1}^{n}X_i$$

Βήμα 2ο : Υπολογισμός της $LR_n\left(\hat{\boldsymbol{\theta}}_n,\hat{\gamma}_n\right)=\sum_{i=1}^{n}\ln\frac{f\left(X_i,\hat{\boldsymbol{\theta}}_n\right)}{g\left(X_i,\hat{\gamma}_n\right)}$. Είναι ύστερα από αλγεβρικές πράξεις :

$$LR_n\left(\hat{\boldsymbol{\theta}}_n,\hat{\gamma}_n\right)=-n\hat{\theta}_{1n}-\frac{1}{2}n\ln\left(2\pi\hat{\theta}_{2n}\right)+\frac{n}{2}+n\ln\hat{\gamma}_n.$$

Βήμα 3ο : Υπολογισμός της

$$\widehat{\omega}_n^2=\frac{1}{n}\sum_{i=1}^{n}\left[-lnX_i-\frac{1}{2}ln\left(2\pi\hat{\theta}_{2n}\right)-\frac{\left(lnX_i-\hat{\theta}_{1n}\right)^2}{2\hat{\theta}_{2n}}+ln\hat{\gamma}_n+\frac{X_i}{\hat{\gamma}_n}\right]^2-\frac{1}{n^2}\left[LR_n(\widehat{\boldsymbol{\theta}}_n,\hat{\gamma}_n)\right]^2$$

Βήμα 4ο : Υπολογισμός της $T_\nu=n^{-\frac{1}{2}}\frac{LR_n(\widehat{\boldsymbol{\theta}}_n,\hat{\gamma}_n)}{\widehat{\omega}_n}$

Βήμα 5ο Κανόνας απόφασης: όπως αυτός αναφέρθηκε πρωτύτερα

Κεφάλαιο 4 Αριθμητικές Εφαρμογές

Στο Κεφάλαιο αυτό θα γίνει εφαρμογή του κανόνα απόφασης που παρουσιάσθηκε νωρίτερα σε δύο σύνολα δεδομένων που θα περιγραφούν στη συνέχεια και παρατίθενται από τον Tai (2003). Ειδικότερα, στην πρώτη στήλη του πίνακα που ακολουθεί δίνεται ο χρόνος επιβίωσης 35 ασθενών που είχαν περιορισμένο στάδιο μικροκυτταρικού καρκίνου του πνεύμονα (LC), ενώ στην δεύτερη στήλη δίνεται ο χρόνος επιβίωσης για ασθενείς που πέθαναν από καρκίνο του τραχήλου της μήτρας (CC). Το ερώτημα που μπορεί να τεθεί από έναν ερευνητή είναι ποια κατανομή περιγράφει, μοντελοποιεί, καλύτερα το χρόνο επιβίωσης αυτών των δύο τύπων καρκίνου. Δύο πιθανά τέτοια μοντέλα είναι αυτό της εκθετικής και της λογαριθμοκανονικής κατανομής. Σε όσα ακολουθούν F είναι η λογαριθμοκανονική κατανομή και G η εκθετική κατανομή (βλέπε Ορισμός 2 και 3 πρωτύτερα).

Επομένως, στην ενότητα αυτή θα εφαρμοστεί όπως αναφέρθηκε παραπάνω η μεθοδολογία του Vuong (1989) στην περίπτωση που τα δύο «ανταγωνιστικά» μοντέλα είναι αυτό της λογαριθμοκανονικής και της εκθετικής κατανομής (βλέπε και Μπραρδάκας 2013). Για το λόγο αυτό θα γίνει χρήση της γλώσσας προγραμματισμού Python έκδοσης 2.7 στο λογισμικό Spyder, όπου θα συνταχθεί ο κώδικας που θα κάνει τους απαραίτητους υπολογισμούς και θα τυπώνει τα ανάλογα αποτελέσματα για τη σύγκριση των δύο ανταγωνιστικών μοντέλων που μας ενδιαφέρουν.

LC	CC
4.04	5.26
4.70	6.64
5.82	8.38
6.15	9.80
7.07	11.08
7.36	11.18
7.56	12.56
7.76	12.66
7.82	13.45
7.86	14.14
7.86	17.46
7.89	17.52
8.15	20.91
8.19	21.67
8.84	23.18
9.04	25.74
9.17	25.78
9.24	32.55
9.47	34.13
9.67	37.55
10.03	38.07
10.06	38.70
10.13	39.85
10.26	41.88
10.32	50.83
10.36	51.16
10.42	53.98
10.42	55.96
10.45	57.11
10.52	62.50
10.52	66.08
10.72	67.82
10.75	67.86
10.75	70.55
11.15	78.05

Πίνακας, Χρόνοι επιβίωσης ασθενών με LC και CC.

Σύντομη επεξήγηση του κώδικα

Για την υλοποίηση όσων αναφέρθηκαν προηγουμένως είναι απαραίτητη η σύνταξη ενός κώδικα, με τη βοήθεια του οποίου θα γίνει η σύγκριση των δύο ανταγωνιστικών μοντέλων που μας ενδιαφέρουν. Επομένως, για να επιτευχθεί αυτό πραγματοποιούμε τα εξής βήματα, τα οποία είναι όμοια για τις εφαρμογές 1 και 2, χρησιμοποιώντας διαφορετικά σύνολα ιατρικών δεδομένων σε κάθε εφαρμογή. Αρχικά δηλώνουμε τις απαραίτητες βιβλιοθήκες, οι οποίες είναι η math και από scipy.stats η norm. Έπειτα σε μορφή πίνακα εισάγουμε τα δεδομένα που επιθυμούμε να αναλύσουμε (γραμμές 4-6 της Εφαρμογής 1). Στη συνέχεια, υπολογίζουμε τον αριθμό των όρων του πίνακα, δηλαδή το μέγεθος του δείγματος (γραμμή 8 της Εφαρμογής 1). Έπειτα, υπολογίζουμε το λογάριθμο του κάθε στοιχείου του πίνακα των δεδομένων για την διευκόλυνση στις πράξεις (γραμμές 10-12 της Εφαρμογής 1). Αφότου γίνουν τα βήματα αυτά υπολογίζουμε το $\hat{\theta}_{1n}, \hat{\theta}_{2n}$ και $\hat{\gamma}_n$ (γραμμές 14, 16-20 και 22, αντίστοιχα, της Εφαρμογής 1). Επιπρόσθετα, με βάση τους τύπους που αναφέρθηκαν παραπάνω υπολογίζουμε την ποσότητα $LR_n(\hat{\theta}_n, \hat{\gamma}_n)$, $\widehat{\omega}_n^2$ και T_ν (γραμμές 24, 26-30 και 32, αντίστοιχα της Εφαρμογής 1). Επίσης, εκτυπώνεται η τιμή της T_ν (γραμμή 34 της Εφαρμογής 1). Τέλος, ακολουθώντας τον κανόνα απόφασης όπως έχει αναφερθεί προηγουμένως γίνεται ο ανάλογος έλεγχος και εκτυπώνονται τα επιθυμητά αποτελέσματα. Ειδικότερα θα τυπωθεί l=1 όταν επιλέγεται η λογαριθμοκανονική κατανομή, e=1 όταν επιλέγεται η εκθετική και equalv=1 όταν τα μοντέλα είναι ισοδύναμα.

Παρακάτω παρατίθεται ο κώδικας (Εφαρμογή 1) και το τι τυπώνει στην περίπτωση που τα δεδομένα μας είναι για τους χρόνους επιβίωσης για ασθενείς που είχαν περιορισμένο στάδιο μικροκυτταρικού καρκίνου του πνεύμονα (LC).

Εφαρμογή 1:

```
import math
from scipy.stats import norm

LC =[4.04, 4.70, 5.82, 6.15, 7.07, 7.36, 7.56, 7.76, 7.82, 7.86, 7.86, 7.89,
8.15, 8.19, 8.84, 9.04, 9.17, 9.24, 9.47, 9.67, 10.03, 10.06, 10.13, 10.26,
10.32, 10.36, 10.42, 10.42, 10.45, 10.52, 10.52, 10.72, 10.75, 10.75, 11.15]

n= len(LC)

lnLC = []
for i in range(len(LC)):
    lnLC.append(math.log(LC[i]))

eth1=sum(lnLC)/n

s= 0
for i in lnLC:
    s+= pow(i-eth1, 2)

eth2=s/n

egama=sum(LC)/n

LR=((-n)*eth1)-(0.5*n*math.log(2*math.pi*eth2))+(0.5*n)+(n*math.log(egama))

s1=0
for i in range(len(LC)):
    s1+= pow((-lnLC[i]-(0.5*math.log(2*math.pi*eth2))-(pow((lnLC[i]-eth1), 2)/(2*eth2))+ math.log(egama) + (LC[i]/egama)), 2)

ew=(s1/n)-(pow(LR, 2))/(pow(n, 2))

TVL=(pow(n,-0.5)*LR)/pow(ew, 0.5)

print TVL

if TVL > norm.ppf(0.95):
    print ("l=1")
elif TVL <- norm.ppf(0.95):
    print ("e=1")
elif abs(TVL) < norm.ppf(0.975):
    print ("equalv=1")

```

Έπειτα, από την εκτέλεση του παραπάνω κώδικα για τα συγκεκριμένα ιατρικά σύνολα δεδομένων η τιμή του T_v είναι 6.5353 και συνεπώς με βάση τη τιμή αυτή και τους κανόνες απόφασης προκύπτει ότι l=1, δηλαδή ότι το μοντέλο της λογαριθμοκανονικής κατανομής είναι καλύτερο.

Στη συνέχεια, παρατίθεται ο κώδικας (Εφαρμογή 2) και το τι τυπώνει στην περίπτωση που τα δεδομένα μας είναι οι χρόνοι επιβίωσης για ασθενείς που πέθαναν από καρκίνο του τραχήλου της μήτρας (CC).

Εφαρμογή 2:

```
import math
from scipy.stats import norm

CC =[5.26, 6.64, 8.38, 9.80, 11.08, 11.18, 12.56, 12.66, 13.45, 14.14,
17.46, 17.52, 20.91, 21.67, 23.18, 25.74, 25.78, 32.55, 34.13, 37.55,
38.07, 38.70, 39.85, 41.88, 50.83, 51.16, 53.98, 55.96, 57.11, 62.50,
66.08, 67.82, 67.86, 70.55, 78.05]

n= len(CC)

lnCC = []
for i in range(len(CC)):
    lnCC.append(math.log(CC[i]))

eth1=sum(lnCC)/n

s= 0
for i in lnCC:
    s+= pow(i-eth1, 2)

eth2=s/n

egama=sum(CC)/n

LR=((-n)*eth1)-(0.5*n*math.log(2*math.pi*eth2))+(0.5*n)+(n*math.log(egama))

s1=0
for i in range(len(CC)):
    s1+= pow((-lnCC[i]-(0.5*math.log(2*math.pi*eth2))-(pow((lnCC[i]-eth1), 2)/(2*eth2))+math.log(egama)+ (CC[i]/egama)), 2)

ew=(s1/n)-pow(LR, 2)/pow(n, 2)

TVL=(pow(n,-0.5)*LR)/pow(ew, 0.5)

print TVL

if TVL > norm.ppf(0.95):
    print ("l=1")
elif TVL <- norm.ppf(0.95):
    print ("e=1")
elif abs(TVL) < norm.ppf(0.975):
    print ("equalv=1")
```

Έπειτα, από την εκτέλεση του παραπάνω κώδικα για τα συγκεκριμένα ιατρικά σύνολα δεδομένων η τιμή του T_v είναι 2.3508 και συνεπώς με βάση τη τιμή αυτή και τους κανόνες απόφασης προκύπτει ότι l=1, δηλαδή ότι το μοντέλο της λογαριθμοκανονικής κατανομής είναι καλύτερο.

Παρατήρηση Οι λογάριθμοι των πιθανοφανειών που χρησιμοποιούνται στο στατιστικό του Vuong (1989) επηρεάζονται αν ο αριθμός των άγνωστων παραμέτρων των μοντέλων F και G που εκτιμούμε είναι διαφορετικός. Έτσι, ο Vuong (1989) προτείνει τη θεώρηση της διορθωμένης στατιστικής συνάρτησης της μορφής:

$$L\tilde{R}_n\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\gamma}}_n\right)=LR_n\left(\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\gamma}}_n\right)-K_n\left(\boldsymbol{\theta},\boldsymbol{\gamma}\right),$$

όπου $K_n\left(\boldsymbol{\theta},\boldsymbol{\gamma}\right)$ ένας παράγοντας διόρθωσης. Στο πλαίσιο αυτό, ο Vuong (1989) πρότεινε δύο διαφορετικούς παράγοντες διόρθωσης. Ο πρώτος βασίζεται στο κριτήριο πληροφορίας του Akaike (1973) και είναι:

$$K_n^1(\boldsymbol{\theta},\boldsymbol{\gamma}) = p - q,$$

ενώ ο δεύτερος βασίζεται στο κριτήριο πληροφορίας του Schwarz (1978) και είναι ο:

$$K_n^2(\boldsymbol{\theta},\boldsymbol{\gamma}) = \frac{p}{2}\ln n - \frac{q}{2}\ln n,$$

όπου p και q ο αριθμός των εκτιμώμενων παραμέτρων των F και G, αντίστοιχα, ενώ η κατανομή της στατιστικής συνάρτησης παραμένει η ίδια. Για τα δύο παραδείγματα που αναφέρθηκαν πρωτύτερα προκύπτει ότι για το πρώτο σύνολο δεδομένων η απόφαση δεν αλλάζει χρησιμοποιώντας είτε τον πρώτο είτε τον δεύτερο παράγοντα διόρθωσης. Για το δεύτερο σύνολο δεδομένων όμως η απόφαση είναι διαφορετική καθώς προκύπτει ότι τα δύο μοντέλα είναι ισοδύναμα. Το γεγονός αυτό φανερώνει τη χρησιμότητα του παράγοντα διόρθωσης.

Κεφάλαιο 5 Επίλογος

Αντικείμενο της παρούσας διατριβής ήταν η παρουσίαση και η εφαρμογή της διαδικασίας ελέγχου που προτάθηκε από τον Vuong (1989) για την επιλογή ενός εκ των δύο ανταγωνιστικών μη εμφωλευμένων μοντέλων. Η μεθοδολογία αποσαφηνίστηκε μέσω της μελέτης της ειδικής περίπτωσης επιλογής μεταξύ της λογαριθμοκανονικής και της εκθετικής κατανομής. Στη συνέχεια, κάνοντας χρήση του κώδικα που αναπτύχθηκε κατά το κεφάλαιο των Αριθμητικών Εφαρμογών, ο οποίος αποτελεί εργαλείο ελέγχου των δύο «ανταγωνιστικών» μοντέλων για τα συγκεκριμένα σύνολα ιατρικών δεδομένων καταλήξαμε στο συμπέρασμα ότι στις δύο εφαρμογές το μοντέλο της λογαριθμοκανονικής είναι καλύτερο. Με παρόμοιο τρόπο η μέθοδος μπορεί να εφαρμοστεί για οποιαδήποτε άλλα δύο ανταγωνιστικά μοντέλα, εφόσον πληρούνται οι υποθέσεις εφαρμογής της μεθόδου του Vuong (1989). Στο σημείο αυτό θα πρέπει να αναφερθεί ότι είναι δυνατή η θεώρηση περισσοτέρων από δύο ανταγωνιστικών μοντέλων χρησιμοποιώντας μεθοδολογίες πολλαπλών συγκρίσεων όπως αυτές περιγράφονται και εφαρμόζονται στην Shimodaira (1998).

Μετά την πρωτοπόρα εργασία του Vuong (1989) πλήθος ερευνητικών εργασιών έχουν εμφανιστεί στη βιβλιογραφία με σκοπό την εφαρμογή ή την επέκταση της μεθόδου. Ειδικότερα, ο Greene (1994) υιοθέτησε τη μεθοδολογία για τις περιπτώσεις όπου τα ανταγωνιστικά μοντέλα είναι τα ZIP vs. Poisson και τα zero-inflated negative binomial vs. negative binomial. Οι έλεγχοι αυτοί λόγω της σπουδαιότητάς τους σε πρακτικές εφαρμογές

έχουν ενσωματωθεί στο λογισμικό Stata. Επίσης, η μέθοδος έχει εφαρμοστεί για την επιλογή κατάλληλου μοντέλου παλινδρόμησης.

Παρότι όμως η μέθοδος του Vuong (1989) είναι αρκετά γενική κάποιες φορές δεν μπορεί να εφαρμοστεί λόγω μη ικανοποίησης των υποθέσεών της. Για παράδειγμα όταν τα δεδομένα δίνονται ως αριθμός παρατηρήσεων σε συγκεκριμένα διαστήματα τότε η μεθοδολογία του Vuong (1989) δεν μπορεί να εφαρμοστεί. Σε μία τέτοια περίπτωση έχουν προταθεί εναλλακτικές μεθοδολογίες από τους Vuong and Wang (1993) και Jimenez-Gamero et al. (2011, 2014). Επιπλέον δεν εφαρμόζεται η μεθοδολογία του Vuong (1989) όταν η πιθανοφάνεια ενός μοντέλου δε μπορεί να υπολογιστεί (ευσταθείς κατανομές) και όταν δεν ικανοποιούνται οι υποθέσεις εφαρμογής, βλέπε μεταξύ άλλων αποκομμένη (truncated) Laplace κατανομή με παραμέτρους θέσης και κλίμακας. Στις δύο τελευταίες περιπτώσεις είναι δυνατή η εφαρμογή της μεθόδου που προτάθηκε από τους Jimenez-Gamero et al. (2016) και η οποία βασίζεται σε ένα μέτρο εγγύτητας μεταξύ χαρακτηριστικών συναρτήσεων. Πρόσφατα, οι Jimenez-Gamero and Batsidis (2017) πρότειναν αντίστοιχη μεθοδολογία με αυτή του Vuong (1989) για την περίπτωση διακριτών δεδομένων, χρησιμοποιώντας ένα μέτρο εγγύτητας μεταξύ πιθανογεννητριών συναρτήσεων. Η μέθοδός τους είναι ιδιαίτερα χρήσιμη όταν η πιθανοφάνεια δεν μπορεί να υπολογιστεί (διακριτές ευσταθείς) είτε όταν η συνάρτηση πιθανότητας δεν έχει κλειστή μορφή.

Παρότι έχουν εμφανιστεί κάποιες επεκτάσεις της μεθόδου του Vuong (1989) παραμένουν ανοικτά κάποια θέματα προς διερεύνηση. Για παράδειγμα η περίπτωση που τα δεδομένα δεν αποτελούν τυχαίο δείγμα ή που τα υπό θεώρηση μοντέλα είναι ημιπαραμετρικά και όχι παραμετρικά. Επίσης η δημιουργία βιβλιοθήκης στην R για διάφορες δυνατές επιλογές μοντέλων με τον ταυτόχρονο έλεγχο αν πρόκειται για εμφωλευμένα, επικαλυπτόμενα, ή μη εμφωλευμένα και την κατάλληλη επιλογή στατιστικού ελέγχου.

Τέλος, θα ήταν παράλειψη να μην επισημάνουμε ότι στη στατιστική βιβλιογραφία έχει εμφανιστεί πληθώρα εργασιών που προτείνεται η επιλογή εκείνου του μοντέλου που μεγιστοποιεί ή ελαχιστοποιεί ένα συγκεκριμένο κριτήριο (για παράδειγμα Akaike (1974)), αντί για την επιλογή μοντέλου μέσω του ελέγχου υποθέσεων. Επιχειρήματα υπέρ του ελέγχου υποθέσεων παρατίθενται μεταξύ άλλων από τους Vuong (1989), Vuong and Wang (1993) και Genius and Strazzera (2002), ενώ επιχειρήματα υπέρ των κριτηρίων παρατίθενται από τους Granger et al. (1995), Preminger and Wettstein (2005). Η παράθεση των επιχειρημάτων αυτών ξεφεύγει από τους σκοπούς αυτής της διπλωματικής διατριβής.

Αναφορές

Akaike, H. (1973). Information Theory and an Extension of the Likelihood Ratio Principle. Proceedings of the Second International Symposium of Information Theory, ed. by B. N. Petrov and F. Csaki. Budapest: Akademiai Kiado, 257-281.

Akaike, H. (1974). A New Look at the Statistical Model Identification. IEEE Transactions on Automatic Control, AC-19, 716-723.

Genius M. and Strazzera E. (2002). A note about model selection and tests for non-nested contingent valuation models. Economics Letters 74, 363-370.

Granger, C.W.J., King, M.L., White, H. (1995). Comments on testing economic theories and the use of model selection criteria. J. Econometrics 67, 173-187.

Greene, W. H. (1994). Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models. Working paper, Stern School of Business, NYU EC 94-10.

Jimenez-Gamero, M. D., Pino Mejias, R., Alba Fernández, M. V., Moreno Rebollo, J. L., (2011). Minimum Phi-Divergence Estimation in Misspecified Multinomial Models. Computational Statistics and Data Analysis. 2011. Vol. 55, 3365-3378.

Jimenez-Gamero, M. D., Alba Fernández, M. V., Barranco-Chamorro, I., Muñoz-Garcia, J., (2014). Two classes of divergence statistics for testing uniform association. Statistics. 2014. Vol. 48. Núm. 2. Pag. 367-387.

Jiménez-Gamero, M. D., Batsidis, A., Alba-Fernández, M. V. (2016). Fourier methods for model selection. Ann Inst Stat Math (2016) 68:105–133

Jimenez-Gamero, MD and Batsidis, A. (2017). Minimum distance estimators for count data based on the probability generating function with applications. Metrika, 80,503-545.

Kullback, S. and Leibler, R. A. (1951). On information and Sufficiency. Annals of Mathematical Statistics, 22, 79-86.

Kullback, S. (1959). Information Theory and Statistics. New York: Wiley.

Pesaran, M. H. (1987). Global and Partial Non-nested Hypotheses and Asymptotic Local Power. Econometric Theory, 3, 69-97.

Preminger, A. and Wettstein, D. (2005). Using the Penalized Likelihood Method for Model Selection with Nuisance Parameters Presents only under the Alternative: An Application to Switching Regression Models. Journal of Times Series Analysis, 26: 715-741.

Schwarz, G. (1978). Estimating the Dimension of a Model. Ann Statist. 6, no.2, 461-464.

Shimodaira, H. (1998). An application of multiple comparison techniques to model selection. Annals of the Institute of Mathematical Statistics, 50, 1–13.

Tai, P. (2003). Twenty-year follow-up study of long-term survival of limited-stage small cell lung cancer and overview of prognostic and treatment factors. Int. Journal of Radiation Oncology Biol. Phys, 56(3) 626-633.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica,57, 257–306.

Vuong, Q. and Wang, W. (1993). Minimum chi-square estimation and tests for model selection. Journal of Econometrics, 56, 141-168.

Μπαρδάκας, Κ. (2013). Έλεγχος μη εμφωλευμένων μοντέλων, Μια κριτική ανασκόπηση (Μεταπτυχιακή Διατριβή), Πανεπιστήμιο Ιωαννίνων.