Μάθημα : ΑΝΑΛΥΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ - ΕΞΕΤΑΣΕΙΣ ΧΕΙΜΕΡΙΝΟΥ ΕΞΑΜΗΝΟΥ 2014-15
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
******* Διάρκεια Εξέτασης : 2.30 ώρες *******

!! Επιλέξτε 4 Ζητήματα από τα 7 !!

ΖΗΤΗΜΑ 1

(α) Αφού διαιρεθούν όλοι οι όροι του προσαρμοσμένου απλού γραμμικού μοντέλου $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$, με τη δειγματική τυπική απόκλιση των y -παρατηρήσεων $\sqrt{S_{yy}/(n-1)}$ και μετασχηματιστεί το x ως $x'=x/\sqrt{S_{xx}/(n-1)}$, δείξτε ότι ο συντελεστής του νέου x' είναι απλά ο συντελεστής συσχέτισης r_{xy} του Pearson (S_{xx} και S_{yy} είναι τα αθροίσματα τετραγώνων των x και y παρατηρήσεων αντίστοιχα).

$\cdot\sqrt{S_{xx}/(n-1)}$

(β) Δείξτε ότι στο γενικό γραμμικό μοντέλο με τη σταθερά $E(\mathbf{y})=\mathbf{X\beta}$, ισχύουν ότι

(i) $\operatorname{cov}(y_i,\hat{y}_i)=\sigma^2 h_{ii}$, όπου $h_{ii}=\mathbf{x}_i'(\mathbf{X'X})^{-1}\mathbf{x}_i$, τα διαγώνια στοιχεία του πίνακα $\mathbf{H}=\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X}'$,

(ii) $\mathbf{X'e}=\mathbf{0}$ και (iii) $\sum_{i=1}^{n} e_i=0$. **(Βαθμ. 2.5)**

ΖΗΤΗΜΑ 2

(i) Δείξτε ότι η εκτιμήτρια μέγιστης πιθανοφάνειας της παραμέτρου $\boldsymbol{\beta}$ στο γενικό γραμμικό μοντέλο $\mathbf{y}=\mathbf{X\beta}+\boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon}\sim N_n(\mathbf{0},\sigma^2\mathbf{I}_n)$ είναι η $\hat{\boldsymbol{\beta}}=(\mathbf{X'X})^{-1}\mathbf{X'y}$, και η αντίστοιχη της σ^2 είναι η $\hat{\sigma}^2=SSE/n$, όπου **X** ο πίνακας σχεδιασμού και SSE το άθροισμα τετραγώνων λόγω σφάλματος.

(ii) Δείξτε ότι η μεγιστοποιημένη λογαριθμοποιημένη συνάρτηση πιθανοφάνειας για το γενικό γραμμικό μοντέλο είναι $\hat{\ell}(\hat{\boldsymbol{\beta}},\hat{\sigma}^2)=-\frac{n}{2}[\ln(2\pi\, SSE/n)+1]$ και στη συνέχεια βρείτε το κριτήριο AIC$=-2\hat{\ell}(\hat{\boldsymbol{\beta}},\hat{\sigma}^2)+2d$ για το μοντέλο αυτό, όπου d ο συνολικός αριθμός παραμέτρων στο μοντέλο.

(Βαθμ. 2.5)

ΖΗΤΗΜΑ 3

Έστω δύο γενικά γραμμικά μοντέλα παλινδρόμησης M_0 και M_1, όπου $M_0 \subset M_1$ και το μοντέλο M_0 με μια μεταβλητή λιγότερο.

(α) Εξηγείστε πώς θα επιλέξτε το καλύτερο μοντέλο από τα δύο, με βάση

(i) έναν έλεγχο F και (ii) το κριτήριο AIC.

(β) Δείξτε ότι $\bar{R}_1^2>\bar{R}_0^2 \Leftrightarrow F=\dfrac{SSE_0-SSE_1}{SSE_1/(n-p)}>1$, όπου $\bar{R}_1^2=1-\dfrac{SSE_1/(n-p)}{SST/(n-1)}$, ο διορθωμένος συντελεστής προσδιορισμού. **(Βαθμ. 2.5)**

ΖΗΤΗΜΑ 4

Εξετάζεται η γραμμική παλινδρόμηση της Y σε σχέση με τις επεξηγηματικές μεταβλητές X_1, X_2, X_3. Με βάση τον παρακάτω πίνακα και για μέγεθος δείγματος n=10, να βρεθεί το καταλληλότερο μοντέλο, κάνοντας χρήση ενός ελέγχου F, καθώς και του κριτηρίου AIC, όπου κρίνετε ότι είναι απαραίτητο.

Πλήθος μεταβλητών στο μοντέλο	R^2	$\bar{R}^2$	C_p (Mallows)	S	X_1	X_2	X_3
1	87.7	86.1	19.7	3.0242			X
1	64.7	60.3	67.5	5.1207		X	
2	96.3	95.2	3.7	1.7765		X	X
2	90.3	87.5	16.2	2.8699	X		X
3	97.1	95.7	4.0	1.6888	X	X	X

$\left(S=\left(\mathbf{e'e}/_{n-k-1}\right)^{1/2}\right)$ **(Βαθμ. 2.5)**

ΖΗΤΗΜΑ 5

Κατασκευαστής θέλει να εξετάσει αν υπάρχουν διαφορές μεταξύ τεσσάρων τύπων λίπους, ως προς την απορροφητικότητά τους κατά τη διάρκεια παρασκευής συγκεκριμένης τροφής. Κάθε τύπος χρησιμοποιήθηκε σε 5 τυχαία επιλεγμένα δείγματα τροφής από συνολικά 20 δείγματα, επίσης τυχαία επιλεγμένα.

1	2	3	4
64	78	75	55
72	91	93	66
68	97	78	49
56	85	63	57
95	77	76	45

Υιοθετώντας την κωδικοποίηση

$$x_1 = \begin{cases} 1, & \text{αν τύπος 1} \\ 0, & \text{αλλιώς} \end{cases} \quad x_2 = \begin{cases} 1, & \text{αν τύπος 2} \\ 0, & \text{αλλιώς} \end{cases} \quad x_3 = \begin{cases} 1, & \text{αν τύπος 3} \\ 0, & \text{αλλιώς} \end{cases}$$

προσαρμόζεται στα δεδομένα το μοντέλο παλινδρόμησης $E(y)=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3$.

(i) Να γίνει ο έλεγχος $H_0:\beta_1=\beta_2=\beta_3=0$ με εναλλακτική H_1: τουλάχιστον ένα $\beta_j \neq 0$.

[Δίνονται $\sum_{i=1}^{c}\sum_{j=1}^{n_i} y_{ij}^2 = 108152$, SSR=2603.6].

(ii) Να συμπληρωθεί και να ερμηνευτεί ο παρακάτω πίνακας

Μεταβλητές	$\hat{\beta}$	$se(\hat{\beta})$	**t**	**p-τιμή**
Σταθερά	54.400	4.833	11.26	<0.001
X_1	16.600	6.834		
X_2	31.200	6.834		
X_3	22.600	6.834		

(Βαθμ. 2.5)

ΖΗΤΗΜΑ 6 Εξετάζεται η σχέση μεταξύ της παραγωγής ζαχαρότευτλου Y (x 1000 τόνους), του έτους παραγωγής X_2 και της αντίστοιχης για το έτος, μέσης θερμοκρασίας X_1 (^{o}C). (i) Συμπληρώστε τον παρακάτω πίνακα. Σχολιάστε τα αποτελέσματά σας.

Το προσαρμοσμένο μοντέλο παλινδρόμησης είναι:

$\hat{y} = 342 + 6.23\,x_1 + 21.0\,x_2$

Μεταβλητές	$\hat{\beta}$	$se(\hat{\beta})$	**t**	**p-τιμή**	**VIF**
Σταθερά	341.950	85.280	4.01	0.005	
X_1	6.232	4.244			
X_2	20.977	3.996			

S = 33.1958, R^2 =___ %, $\bar{R}^2$ = ___ %, $r_{x_1x_2}$ = -0.404

Ανάλυση Διασποράς

Πηγή	**β.ε.**	**ΑΤ**	**ΜΑΤ**	**F**	**p-τιμή**
Παλινδρόμηση		30936			
Σφάλμα					
Σύνολο	9	38650			

(ii) Για το παραπάνω μοντέλο δίνεται ότι $h_{33}=0.925$. Αποτελεί η τρίτη παρατήρηση σημείο επιρροής;

(iii) Να κατασκευαστεί ένα 0.99-διάστημα εμπιστοσύνης για το συντελεστή β_2 του έτους παραγωγής.

(Βαθμ. 2.5)

ΖΗΤΗΜΑ 7

Έστω μοντέλο της κατανομής Poisson $f(y)=\dfrac{\exp(-\mu)\,\mu^y}{y!}$, $y=0,1,2,\ldots,$.

(i) Πώς μπορούμε να εισάγουμε συμμεταβλητές $\mathbf{x}$ σε αυτό το μοντέλο, ώστε στη συνέχεια να μπορεί να εξεταστεί πιθανή εξάρτηση της y από τις x.

(ii) Τι είναι μια συνάρτηση σύνδεσης και να δωθεί αυτή για το μοντέλο της παλινδρόμησης Poisson.

(iii) Αναφέρετε δύο τεχνικές με τις οποίες μπορούμε να ελέγξουμε τη συμβολή μιας συμμεταβλητής x στο μοντέλο παλινδρόμησης Poisson.

(iv) Εξετάζεται η εξάρτηση του αριθμού Y αποζημιώσεων λόγω τροχαίων ατυχημάτων (ανά τύπο συμβολαίου), από την ηλικία του οδηγού και την κατηγορία οχήματος (0 ή 1). Να συμπληρωθεί ο παρακάτω πίνακας μετά από προσαρμογή ενός μοντέλου παλινδρόμησης Poisson στα δεδομένα και να ερμηνευτούν οι εκτιμημένοι συντελεστές $\hat{\beta}$.

Μεταβλητές	$\hat{\beta}$	$se(\hat{\beta})$	**Z**	**p-τιμή**
Σταθερά	2.49828	0.08104	23.66	<0.001
Τύπος οχήματος (X_1)	-0.53749	0.03692		
Ηλικία (X_2)	0.83664	0.02067		

(v) Η τιμή της ελεγχοσυνάρτησης deviance για το μοντέλο M_0 που περιέχει μόνο τη συμμεταβλητή X_1 δίνεται ως D_0=3535.4 και η αντίστοιχη τιμή για το μοντέλο M_1 με την επιπλέον συμμεταβλητή X_2, ως D_1=3315.8. Ελέγξτε τη σημαντικότητα της διαφοράς των τιμών αυτών και αν ο έλεγχος αυτός συμφωνεί με τα αποτελέσματα του παραπάνω πίνακα για την ηλικία.

(Βαθμ. 2.5)