Μάθημα : ΑΝΑΛΥΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ - ΕΞΕΤΑΣΕΙΣ ΧΕΙΜΕΡΙΝΟΥ ΕΞΑΜΗΝΟΥ 2018-19
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
******* Διάρκεια Εξέτασης : 2.30 ώρες *******

!! Επιλέξτε 4 Ζητήματα από τα 7 !!

ΖΗΤΗΜΑ 1 (Βαθμ. 2.5)

(Α) Δώστε τον ορισμό των υπολοίπων $\mathbf{e}$ σε ένα γενικό γραμμικό μοντέλο $\mathbf{y}=\mathbf{X\beta}+\boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon}\sim N_n(0,\sigma^2\mathbf{I})$. Στη συνέχεια βρείτε την κατανομή των $\mathbf{e}$ και δείξτε ότι $E(\mathbf{e})=\mathbf{0}$, $V(\mathbf{e})=\sigma^2(\mathbf{I}-\mathbf{H})$ και $\text{cov}(\mathbf{e},\hat{\mathbf{y}})=\mathbf{0}$, όπου $\mathbf{H}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

(Β) Με βάση τη σ.π.π. της $\mathbf{y}$, $f(\mathbf{y})=\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left[\frac{-(\mathbf{y}-\mathbf{X\beta})'(\mathbf{y}-\mathbf{X\beta})}{2\sigma^2}\right]$, $\left\{\frac{-(\underline{Y}-\underline{X}\underline{\beta})'(\underline{Y}-\underline{X}\underline{\beta})}{2\sigma^2}\right\}$

(i) Βρείτε τη μεγιστοποιημένη λογαριθμοποιημένη συνάρτηση πιθανοφάνειας $\hat{\ell}$ για το γενικό γραμμικό μοντέλο, δεδομένου ότι η ε.μ.π. της σ^2 είναι η $\hat{\sigma}^2=\frac{\text{SSE}}{n}$.

(ii) Στη συνέχεια δείξτε ότι το κριτήριο AIC είναι $\text{AIC}=n\left[\ln(2\pi)+\ln(\text{SSE}/n)+1\right]+2(p+1)$ για το μοντέλο αυτό.

ΖΗΤΗΜΑ 2 (Βαθμ. 2.5)

(Α) Δείξτε ότι η ελεγχοσυνάρτηση για τον έλεγχο H_0: $\beta_1=\beta_2=\ldots=\beta_k=0$, με εναλλακτική την H_1: τουλάχιστον ένα $\beta_j\neq 0$, γράφεται και ως $F=\frac{R^2/k}{(1-R^2)/(n-k-1)}$, όπου R^2 ο συντελεστής προσδιορισμού.

(Β) Μεσιτικό γραφείο θέλει να εξετάσει τη σχέση μεταξύ της τιμής αγοράς οικίας (Y), με τον αντίστοιχο φόρο (X_1) και το μέγεθος της οικίας (X_2).

(i) Βρείτε το VIF για τις δύο επεξηγηματικές μεταβλητές και εξετάστε αν υπάρχει ένδειξη πολυσυγγραμμικότητας.

(ii) Ελέγξτε αν η μεταβλητή X_1 χρειάζεται στο μοντέλο $E(y)=\beta_0+\beta_1X_1$ και εξετάστε αν η παρατήρηση 6 είναι σημείο επιρροής.

(iii) Δεδομένου ότι η X_1 είναι απαραίτητη στο μοντέλο, χρειάζεται να προστεθεί και η μεταβλητή X_2 ;

(iv) Να κατασκευαστεί ένα 0.95 διάστημα εμπιστοσύνης της παραμέτρου β_1 από το μοντέλο που έχει προκύψει.

[**Δίνονται:** Η απόσταση Cook $D_i=\frac{e_i^2h_{ii}}{pS^2(1-h_{ii})^2}$ (p ο αριθμός παραμέτρων στο μοντέλο, i=1,...,24), $r_{X_1,X_2}=0.65$, $\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2=57.52$, $\sum_{i=1}^{n}y_i^2=29611.81$, $\bar{y}=34.63$. $n=24$

Για το μοντέλο **με τη** X_1 **μόνο** : $h_{66}=0.15$, $e_6=3.62$, $R^2=0.764$, $\hat{\beta}_1=3.32$, $S^2=8.93$

και για το μοντέλο **με τις** X_1 **και** X_2: $\text{SSR}=663.60$, $R^2=0.798$, $\hat{\beta}_2=6.10$, $\sqrt{c_{22}}=.876$].

ΖΗΤΗΜΑ 3 (Βαθμ. 2.5) $se(\hat{\beta}_1)=1.2$ Δίνονται: $\hat{\beta}_1=3.1$

(Α) Περιγράψτε πώς μέσω μιας ψευδομεταβλητής Z (=0, αν τα δεδομένα ανήκουν στην ομάδα I και =1, αν ανήκουν στην II) στο μοντέλο $E(Y)=\beta_0+\beta_1X+\beta_2Z+\beta_3XZ$, μπορούμε να εξετάσουμε αν (α) δύο διαφορετικές ευθείες ή (β) δύο παράλληλες ευθείες ή (γ) μια ευθεία ταιριάζουν στα δεδομένα μας.

(Β) Εξετάζεται ο βαθμός επίδοσης (Y), n=14 υπαλλήλων εταιρείας ένα μήνα μετά την πρόσληψή τους σε σχέση με ένα τεστ ικανότητας (X). Ορίζεται μεταβλητή Z =0, αν γυναίκα και Z=1, αν άντρας. Να εφαρμοστούν οι έλεγχοι του Ζητήματος **3(Α)** στα δεδομένα αυτά.

[**Δίνονται:** SSE(α)= 20.736, SSE(β)= 24.043, SSE(γ)= 29.484].

ΖΗΤΗΜΑ 4 (Βαθμ. 2.5)

(Α) Δώστε τον ορισμό του κριτηρίου Cp-Mallows. Ποιά είναι η χρήση του ;

(Β) Εξετάζεται η γραμμική παλινδρόμηση της Y σε σχέση με 5 επεξηγηματικές μεταβλητές X_1, X_2, X_3, X_4, X_5 σε δείγμα μεγέθους n=60. Με βάση τον παρακάτω πίνακα επιλέξτε τα δύο καλύτερα μοντέλα. Στη συνέχεια αξιοποιώντας τον έλεγχο F για τη σύγκριση δύο **εμφωλευμένων** μοντέλων και το κριτήριο AIC (χωρίς τους κοινούς για τα δύο μοντέλα όρους) να βρεθεί το βέλτιστο από τα δύο. [**Δίνεται:** $S=\left(SSE/(n-k-1)\right)^{1/2}$]

	R^2	$\bar{R}^2$	C_p	S	X_1	X_2	X_3	X_4	X_5
1	75.5	75.1	90.4	66.316			X		
1	58.7	58.0	190.9	86.113					X
2	90.1	89.7	5.4	42.610		X	X		
2	82.3	81.7	51.9	56.892			X	X	
3	90.7	90.2	3.4	41.497	X	X	X		
3	90.1	89.6	7.0	42.851		X	X	X	
4	91.0	90.3	4.0	41.359	X	X	X	X	
4	90.8	90.2	4.9	41.702	X	X	X		X
5	91.0	90.1	6.0	41.740	X	X	X	X	X

ΖΗΤΗΜΑ 5 (Βαθμ. 2.5)

Ο ακόλουθος πίνακας δίνει την επίδραση του χρόνου αντίδρασης (Α) και της θερμοκρασίας αντίδρασης (Β) στη συγκέντρωση ενός χημικού προϊόντος

Χρόνος (ώρες)	Θερμοκρασία (°C)		
	50	75	100
2	4.7	5.5	4.0
4	6.4	5.9	6.3
6	7.9	9.0	11.4

Θεωρώντας τις δείκτριες μεταβλητές $x_1=\begin{cases}1, & \text{αν θερμοκρασία } 50°C\\ 0, & \text{αλλιώς}\end{cases}$, $x_2=\begin{cases}1, & \text{αν θερμοκρασία } 75°C\\ 0, & \text{αλλιώς}\end{cases}$

και $z_1=\begin{cases}1, & \text{αν 2 ώρες}\\ 0, & \text{αλλιώς}\end{cases}$ $z_2=\begin{cases}1, & \text{αν 4 ώρες}\\ 0, & \text{αλλιώς}\end{cases}$, προσαρμόζεται στα δεδομένα το μοντέλο παλινδρόμησης $E(y)=\beta_0+\beta_1x_1+\beta_2x_2+\gamma_1z_1+\gamma_2z_2$.

(i) Εξετάστε αν οι παράγοντες Θερμοκρασία και Χρόνος επιδρούν σημαντικά στη συγκέντρωση του χημικού προϊόντος.
(Δίνεται: SSE= 6.458)

(ii) Να συμπληρωθεί ο παρακάτω πίνακας. Να δοθούν τελικές ερμηνείες και συμπεράσματα.

Μεταβλητές	$\hat{\beta}$	se($\hat{\beta}$)	t	p-τιμή
Σταθερά	9.8778	0.9471	10.430	<0.001
x1	-0.900	1.037		
x2	-0.433	1.037		
Z1	-4.700	1.037		
Z2	-3.233	1.037		

ΖΗΤΗΜΑ 6 (Βαθμ. 2.5)

(Α) Έστω μοντέλο παλινδρόμησης Poisson $f(y)=\frac{\exp(-\mu_x)\,\mu_x^y}{y!}$, y=0,1,2, ..., με συνάρτηση σύνδεσης $g(\mu_x)=\ln\mu_x=\beta'x$. Γράψτε τη λογαριθμοποιημένη συνάρτηση πιθανοφάνειας $\ell=L(\beta)$ αυτού του μοντέλου.

(Β) Έστω n=72 οδικές διασταυρώσεις με σιδηροδρομικές γραμμές. Με βάση την παλινδρόμηση Poisson, εξετάζεται η σχέση του αριθμού (Y) ατυχημάτων/διασταύρωση με τις συμμεταβλητές X_1 (χρονική περίοδος), X_2 (=0, αν πεζός, =1, αν όχι πεζός) .

(i) Να συμπληρωθούν οι παρακάτω πίνακες.

(ii) Συγκρίνετε τα δύο μοντέλα με την ελεγχοσυνάρτηση Deviance καθώς και με το κριτήριο AIC= $-2\hat{\ell}+2d$ (d ο συνολικός αριθμός παραμέτρων στο μοντέλο) και γράψτε το προσαρμοσμένο **τελικό μοντέλο.**

(iii) Υπολογίστε και ερμηνεύστε τις εκτιμημένες ποσότητες $\exp(\hat{\beta}_j)$ του **τελικού μοντέλου**.

ΜΟΝΤΕΛΟ: 2 **Μεταβλητές**	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	**p-τιμή**	$\exp(\hat{\beta}_j)$
Σταθερά	-5.10261	0.09060	-56.320	<0.001	
X_1	0.00648	0.00838			
X_2	-0.51614	0.06218			
Ελεγχοσυνάρτηση deviance δίνεται ως D_2=**531.60** και η τιμή του κριτηρίου AIC_2=**808.95**					

ΜΟΝΤΕΛΟ: 1 **Μεταβλητές**	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	**p-τιμή**	$\exp(\hat{\beta}_j)$
Σταθερά	-5.10079	0.09056	-56.323	<0.001	
X_2	-0.51614	0.06218			
Ελεγχοσυνάρτηση deviance δίνεται ως D_1=**532.19** με αντίστοιχη τιμή $\hat{\ell}_1 = -401.775$ και τιμή του κριτηρίου **AIC_1=_____**					

ΖΗΤΗΜΑ 7 (Βαθμ. 2.5)

(Α) Έστω το μοντέλο της λογιστικής παλινδρόμησης $\ln\left(\frac{p_i}{1-p_i}\right)=\boldsymbol{\beta}'\mathbf{x}_i$, με p_i την πιθανότητα επιτυχίας στις n_i δοκιμές. Δώστε τον ορισμό των υπολοίπων Pearson για το μοντέλο αυτό.

(Β) Σε μελέτη εξετάζεται ο αριθμός θανάτων Y_i σε ομάδα n_i ασθενών, i=1, … , 10 ομάδες. Με βάση τη λογιστική παλινδρόμηση, εξετάζεται η επίδραση των συμμεταβλητών X_1 (ηλικία) και X_2 (=1, αν καπνιστές, =0, αν μη-καπνιστές) στη σχετική πιθανότητα επιτυχίας $\frac{p_i}{1-p_i}$.

(i) Να συμπληρωθούν οι παρακάτω πίνακες.

(ii) Συγκρίνετε τα δύο μοντέλα με την ελεγχοσυνάρτηση Deviance και με το κριτήριο AIC και γράψτε το προσαρμοσμένο **τελικό μοντέλο**.

(iii) Υπολογίστε και ερμηνεύστε τις εκτιμημένες ποσότητες $\exp(\hat{\beta}_j)$ του **τελικού μοντέλου**.

ΜΟΝΤΕΛΟ: 1 **Μεταβλητές**	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	**p-τιμή**	$\exp(\hat{\beta}_j)$
Σταθερά	-7.798	0.106	-73.566	<0.001	
X_1	0.845	0.029			
Ελεγχοσυνάρτηση deviance δίνεται ως D_1= **84.079** και η τιμή του κριτηρίου AIC_1=**143.06**					

ΜΟΝΤΕΛΟ: 2 **Μεταβλητές**	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	**p-τιμή**	$\exp(\hat{\beta}_j)$
Σταθερά	-8.134	0.140	-58.100	<0.001	
X_1	0.843	0.029			
X_2	0.410	0.108			
Ελεγχοσυνάρτηση deviance δίνεται ως D_2=**68.138** και η τιμή του κριτηρίου AIC_2=**129.12**					