Μάθημα : ΑΝΑΛΥΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ - ΕΞΕΤΑΣΕΙΣ ΧΕΙΜΕΡΙΝΟΥ ΕΞΑΜΗΝΟΥ 2015-16
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
***** Διάρκεια Εξέτασης : 2.30 ώρες *****

!! Επιλέξτε 4 Ζητήματα από τα 7 !!

ΖΗΤΗΜΑ 1 (Βαθμ. 2.5)

(α) Στο γενικό γραμμικό μοντέλο $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ ισχύει ότι η ε.ε.τ. $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ $\sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$, και $\frac{SSE}{\sigma^2} \sim \chi^2_{n-p}$, όπου X ο πίνακας σχεδιασμού και SSE το άθροισμα τετραγώνων λόγω σφάλματος. Δεδομένου ότι $\hat{\beta}_j$ και $\frac{SSE}{\sigma^2}$ είναι ανεξάρτητα, βρείτε ένα γ-διάστημα εμπιστοσύνης για τον άγνωστο συντελεστή β_j μιας μεταβλητής x_j του μοντέλου.

(β) Έστω $y = \beta_0 + \varepsilon \sim N(\beta_0, \sigma^2)$, χωρίς επεξηγηματικές μεταβλητές, τότε με βάση τ.δ. $y_1, y_2, \ldots, y_n$, ανεξαρτήτων παρατηρήσεων από την κατανομή αυτή, δείξτε με τη μεθοδο μέγιστης πιθανοφάνειας ότι $\hat{\beta}_0 = \bar{y}$ και κατά συνέπεια SST=SSE.

ΖΗΤΗΜΑ 2 (Βαθμ. 2.5)

Έστω το γενικό γραμμικό μοντέλο $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

(i) Δείξτε ότι το άθροισμα τετραγώνων λόγω παλινδρόμησης είναι $SSR = \mathbf{y}'(\mathbf{H} - \frac{1}{n}\mathbf{J}_n)\mathbf{y}$, όπου $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ ο πίνακας προβολής και $\mathbf{J}_n$ ο πίνακας του οποίου όλα τα στοιχεία είναι **1**.

(ii) Δώστε τον ορισμό του συντελεστή προσδιορισμού R^2. Τι εκφράζει;

(iii) Γράψτε το δειγματικό συντελεστή συσχέτισης $r_{y,\hat{y}}$ μεταξύ των παρατηρήσεων $\mathbf{y}$ και $\hat{\mathbf{y}}$. Όταν η σταθερά β_0 περιλαμβάνεται στο μοντέλο παλινδρόμησης, ισχύει $\sum_{i=1}^{n} e_i = 0$. Στην περίπτωση αυτή δείξτε ότι $r^2_{y,\hat{y}} = R^2$.

(iv) Έστω k ο αριθμός των επεξηγηματικών μεταβλητών. Δείξτε ότι η ελεγχοσυνάρτηση για τον έλεγχο H_0: $\beta_1=\beta_2=\ldots=\beta_k=0$ έναντι H_1: τουλ. ένα $\beta_j \neq 0$, γράφεται ως $F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$.

ΖΗΤΗΜΑ 3 (Βαθμ. 2.5)

(α) Έστω το γενικό γραμμικό μοντέλο $E(\mathbf{y}) = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2$ (ο πρώτος όρος περιλαμβάνει τη σταθερά β_0 και q_1 μεταβλητές, ο δεύτερος q_2 μεταβλητές). Δείξτε πότε η ε.ε.τ. $\hat{\boldsymbol{\beta}}_1$ είναι αμερόληπτη εκτιμήτρια της $\boldsymbol{\beta}_1$ στην περίπτωση που ο δεύτερος όρος $\mathbf{X}_2\boldsymbol{\beta}_2$ παραλείπεται από το μοντέλο, ενώ χρειάζεται.

(β) Έστω υπόλοιπα $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} \sim N_n(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H}))$ ενός γενικού γραμμικού μοντέλου. Δώστε τον ορισμό δύο περιπτώσεων τυποποιημένων υπολοίπων. Πώς μας χρησιμεύουν;

ΖΗΤΗΜΑ 4 (Βαθμ. 2.5)

(α) Ερευνάται η σχέση μεταξύ y και x_1 και έστω δείκτρια μεταβλητή x_2 ($x_2 = 0$ αν τα δεδομένα είναι της κατηγορίας Α και $x_2 = 1$, αν είναι της κατηγορίας Β). Περιγράψτε πώς μέσω αυτής της x_2 στο μοντέλο $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, μπορούμε να ελέγξουμε αν χρειάζεται να προσαρμοστούν (Ι) δύο διαφορετικές ευθείες (ΙΙ) δύο παράλληλες ευθείες ή (ΙΙΙ) μια κοινή ευθεία και για τις δύο κατηγορίες, όπου $x_3 = x_1 x_2$, η μεταβλητή που εκφράζει την αλληλεπίδραση μεταξύ των μεταβλητών x_1 και x_2.

(β) Να γίνουν αυτοί οι έλεγχοι στην περίπτωση που y = ρυθμός φωτοσύνθεσης, x_1 = ηλιακή ακτινοβολία και x_2 = διαθεσιμότητα του νερού, x_2 =0 (αν χαμηλή), ενώ x_2 =1 (αν υψηλή), με βάση τα ακόλουθα αποτελέσματα :

Regression Analysis: y με x_1, x_2, x_3

The regression equation is

y = 114 + 43.5 x_1 - 25.9 x_2 - 20.6 x_3

Predictor	Coef	SE Coef	T	P
Constant	113.88	29.47	3.86	0.003
x_1	43.480	3.213		
x_2	-25.94	44.620		
x_3		4.188		

S = 33.4748 R-Sq = 96.8% R-Sq(adj) = 95.8%

Analysis of Variance

Source	DF	SS
Regression		338736
Residual Error		11206
Total	13	349942

Regression Analysis: y με x_1, x_2

The regression equation is

$y = 214 + 31.3x_1 - 224\ x_2$

Predictor	Coef	SE Coef	T	P
Constant	214.40	37.49	5.72	<0.001
x_1		3.636		
x_2	-224.39	33.71		

R-Sq = 89.0%   R-Sq(adj) = 87.0%

Analysis of Variance

Source	DF	SS
Regression		
Residual Error	11	38358
Total	13	349942

Regression Analysis: y με x_1

The regression equation is

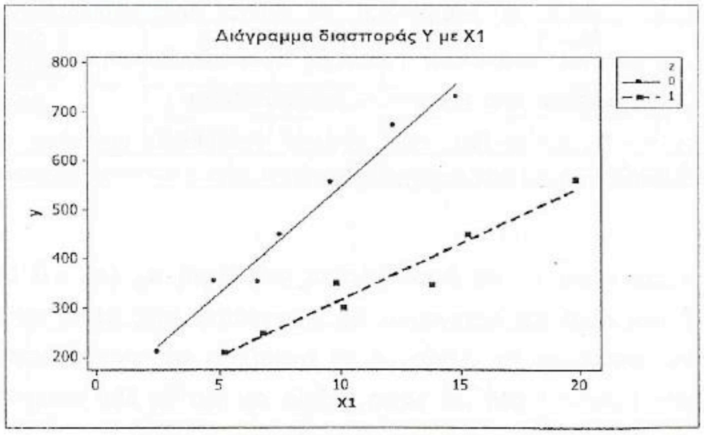
$y = 186 + 22.8\,x_1$

Predictor	Coef	SE Coef	T	P
Constant	186.48	79.98	2.33	0.038
x_1		7.308		

S = 126.769   R-Sq = 44.9%   R-Sq(adj) = 40.3%

Analysis of Variance

Source	DF	SS
Regression	1	157098
Residual Error		
Total	13	349942



ΖΗΤΗΜΑ 5 (Βαθμ. 2.5)

Κατασκευαστής λάστιχων αυτοκινήτου θέλει να εξετάσει αν υπάρχουν διαφορές μεταξύ των τεσσάρων θέσεων τροχού ενός οχήματος, ως προς τη φθορά λάστιχων, για την ίδια απόσταση. Σε κάθε θέση χρησιμοποιήθηκαν 5 τυχαία επιλεγμένα λάστιχα αυτοκινήτου από συνολικά 20 κομμάτια, επίσης τυχαία επιλεγμένα.

θέση 1	θέση 2	θέση 3	θέση 4
20.94	18.28	28.54	20.18
19.01	21.20	27.99	18.79
20.33	19.39	30.07	19.20
17.12	14.82	37.23	34.34
15.92	11.28	38.85	34.71

Υιοθετώντας την κωδικοποίηση

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

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 $\sum_{i=1}^{c}\sum_{j=1}^{n_i} y_{ij}^2 = 12176.7$, S= 5.368, $S=\left(SSE/(n\text{-}k\text{-}1)\right)^{1/2}$].

ii) Να συμπληρωθεί και να ερμηνευτεί ο παρακάτω πίνακας

Μεταβλητές	$\hat{\beta}$	se($\hat{\beta}$)	t	p-τιμή
Σταθερά	25.444			
X_1	-6.78	3.4		
X_2	-8.45	3.4		
X_3	7.09	3.4		

ΖΗΤΗΜΑ 6 (Βαθμ. 2.5)

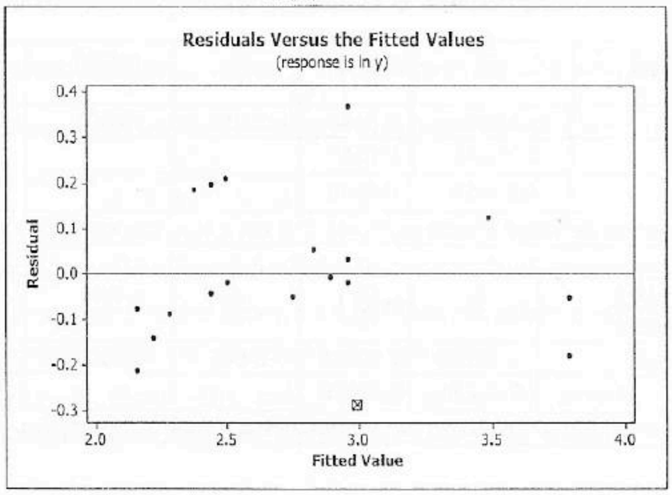
Για τη λειτουργία μιας μονάδας παραγωγής επί 21 ημέρες, εξετάζεται η γραμμική εξάρτηση της διαρροής αμμωνίας Υ (σε log), από τις μεταβλητές X_1 (ταχύτητα λειτουργίας της μονάδας) και X_2 (θερμοκρασία νερού, ^oC).

(i) Συμπληρώστε τον παρακάτω πίνακα. Σχολιάστε τα αποτελέσματά σας.

[Δίνεται: S =0.172, $r_{X_1X_2} = 0.782$, SST=5.482, $AIC=-2\hat{\ell}+2d=n\ln(2\pi)+n+2+n\ln(SSE/n)+2p$]

Μεταβλητές	$\hat{\beta}$	se($\hat{\beta}$)	t	p-τιμή	VIF
Σταθερά	-0.752	0.273	-2.75	0.013	
X_1	0.035	0.007			
X_2	0.063	0.020			

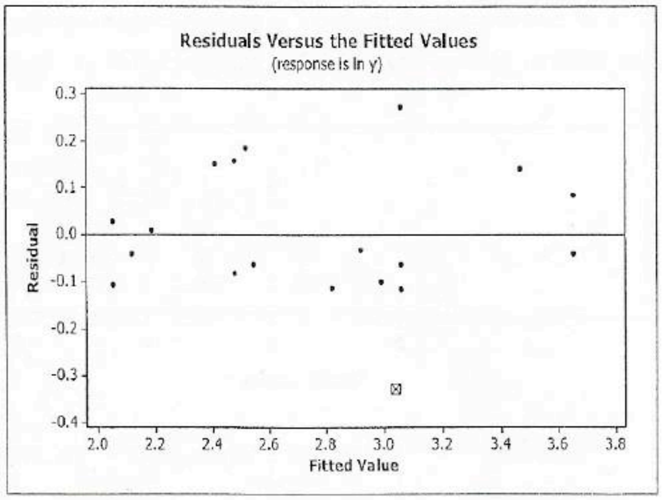
$R^2 =$ ___ %, $C_p=\dfrac{SSE(p)}{SSE_{\pi\lambda\acute{\eta}\rho\varepsilon\varsigma}/(n\text{-}p')}+2p\text{-}n=$ ___ , AIC= ___



(ii) Για το παραπάνω μοντέλο δίνεται ότι $h_{21,21}=0.276$. Αποτελεί η παρατήρηση 21 σημείο επιρροής του μοντέλου;

(iii) Δεδομένου ότι στο μοντέλο υπάρχουν οι μεταβλητές X_1 και X_2 θεωρείται ότι το μοντέλο βελτιώνεται με την προσθήκη της X_1^2; $SSE_{\pi\lambda\acute{\eta}\rho\varepsilon\varsigma}=0.3858$, $R^2=$ ___ %, $C_p=$ ___ , AIC= ___

(iv) Εξετάστε εκ νέου αν η παρατήρηση 21 αποτελεί σημείο επιρροής για το μοντέλο (iii) ($h_{21,21}=0.288$).



ΖΗΤΗΜΑ 7 (Βαθμ. 2.5)

Έστω μοντέλο παλινδρόμησης Poisson $f(y)=\dfrac{\exp(-\mu)\,\mu^{y}}{y!}$, $y=0,1,2,\ldots$, με συνάρτηση σύνδεσης $g(\mu_x)=\ln\mu_x=\boldsymbol{\beta}'\mathbf{x}$ και με ελεγχοσυνάρτηση Deviance$=-2\left(\hat{\ell}_0-\hat{\ell}_{\kappa o\rho}\right)=2\sum_{i=1}^{n}\left[y_i\ln(y_i/\hat{\mu}_i)-(y_i-\hat{\mu}_i)\right]$.

(α) Δώστε τον ορισμό των υπολοίπων Pearson και Deviance για το μοντέλο αυτό. Πώς τα χρησιμοποιούμε ;

(β) Προσαρμόζοντας μοντέλα της παλινδρόμησης Poisson σε δεδομένα 44 ορυχείων μιας περιοχής, εξετάζεται η σχέση του αριθμού ρωγμών σε οροφή ορυχείου (Y), με τις συμμεταβλητές X_1 και X_2 (χαρακτηριστικά των στρωμάτων του ορυκτού) καθώς και με τη X_3 (έτη λειτουργίας του ορυχείου).

Αφού συμπληρωθούν οι παρακάτω πίνακες να ερμηνευτούν οι εκτιμημένες ποσότητες $\exp(\hat{\beta}_j)$ και η γραφική παράσταση των υπολοίπων Deviance του **τελικού μοντέλου**. Συμφωνούν οι έλεγχοι Wald, Deviance και τα κριτήρια AIC;

ΜΟΝΤΕΛΟ: 3 **Μεταβλητές**	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	**p-τιμή**	$\exp(\hat{\beta}_j)$
Σταθερά	-3.39	0.9842	-3.445	<0.001	-
X_1	0.05860	0.0117			
X_2	-0.00376	0.0049			
X_3	-0.03408	0.0147			
Ελεγχοσυνάρτηση deviance δίνεται ως D_3=41.329 και η τιμή του κριτηρίου AIC_3=145.6					

ΜΟΝΤΕΛΟ: 2 **Μεταβλητές**	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	**p-τιμή**	$\exp(\hat{\beta}_j)$
Σταθερά	-3.599	0.9440	-3.813	<0.001	-
X_1	0.05874	0.0117			
X_3	-0.03563	0.0148			
Ελεγχοσυνάρτηση deviance δίνεται ως D_2=41.952 και η τιμή του κριτηρίου AIC_2=144.22					

ΜΟΝΤΕΛΟ: 1 **Μεταβλητές**	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	**p-τιμή**	$\exp(\hat{\beta}_j)$
Σταθερά	-3.32859	0.90886	-3.662	<0.001	-
X_1	0.05234	0.01109			
Ελεγχοσυνάρτηση deviance δίνεται ως D_1=48.620 και η τιμή του κριτηρίου AIC_1=148.89