<u>ΖΗΤΗΜΑ 1</u> (Επιλέξτε 1 από 5) (Βαθμ. 1.5)

(1Α) Έστω γενικό γραμμικό μοντέλο Ε(**y**)=**Xβ**. Δείξτε ότι η ελεγχοσυνάρτηση για την υπόθεση H_0: $\beta_1=\beta_2=...=\beta_k=0$ με εναλλακτική την H_1: τουλάχιστον ένα $\beta_j \neq 0$, γράφεται και ως $F=\dfrac{R^2/k}{(1-R^2)/(n-k-1)}$, όπου R^2 ο συντελεστής προσδιορισμού.

(1Β) Εστω μοντέλο $E(Y)=\beta_0+\beta_1 X_1$. Περιγράψτε τα βήματα που θα ακολουθούσατε όταν εξετάζετε την **<u>πρόσθεση</u>** μιας επεξηγηματικής μεταβλητής X_2 στο μοντέλο αυτό.

(1Γ) Περιγράψτε σύντομα πώς μπορούν να μας βοηθήσουν οι συντελεστές R^2, $\overline{R}^2$, $R^2_{\text{πρόβλεψη}}$, καθώς και οι δείκτες Cp-Mallows και AIC, στην αξιολόγηση ενός μοντέλου Ε(**y**)=**Xβ**.

(1Δ) Πώς κατασκευάζουμε τις γραφικές παραστάσεις «μερικών υπολοίπων» και «πρόσθετων μεταβλητών» και πώς μας βοηθούν στην αξιολόγηση ενός μοντέλου Ε(**y**)=**Xβ**;

(1Ε)
(i) Περιγράψτε πώς μέσω μιας ψευδομεταβλητής Z (=0, αν τα δεδομένα ανήκουν στην ομάδα I και =1, αν ανήκουν στην ομάδα II) στο μοντέλο $E(Y)=\beta_0+\beta_1 X+\beta_2 Z+\beta_3 XZ$, μπορούμε να εξετάσουμε αν στα δεδομένα μας ταιριάζουν (α) δύο διαφορετικές ευθείες , ή (β) δύο παράλληλες ευθείες, ή (γ) μια ευθεία.
Στη συνέχεια εφαρμόστε αυτές τις μεθόδους στα παρακάτω δεδομένα.

(ii) Εξετάζεται ο βαθμός επίδοσης (Υ), n=14 υπαλλήλων εταιρείας, ένα μήνα μετά την πρόσληψή τους, σε σχέση με ένα τεστ ικανότητας (Χ). Ορίζεται μεταβλητή Z =0, αν γυναίκα και Z=1, αν άντρας.

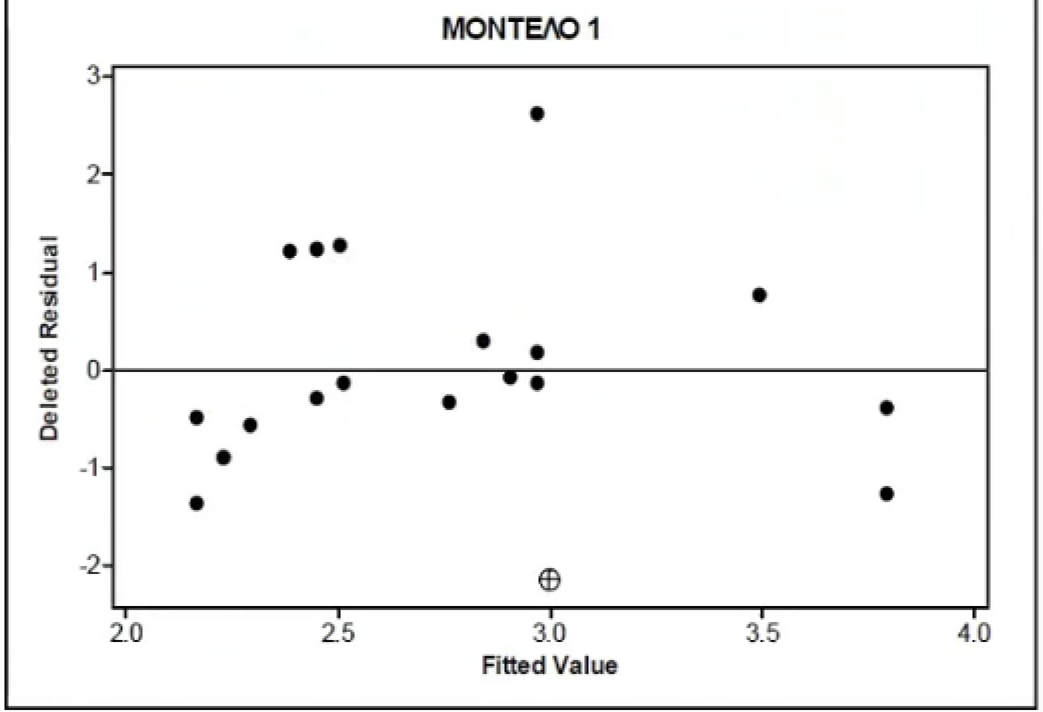
[**Δίνονται**: SSE(α)= 20.736, SSE(β)= 24.043, SSE(γ)= 29.484].

ΖΗΤΗΜΑ 2 (Βαθμ. 3) **Υποχρεωτικό**

Για τη λειτουργία μιας μονάδας παραγωγής επί 21 ημέρες, εξετάζεται η γραμμική εξάρτηση της διαρροής αμμωνίας Y (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, $r_{X_1X_2} = 0.782$, R^2 = 90.3 %]

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			



Για το παραπάνω μοντέλο δίνεται ότι $e_{21} = -0.287$, $h_{21,21}=0.276$ και απόσταση Cook $D_{21}=\dfrac{r_{21}^2 h_{21,21}}{p(1-h_{21,21})}$.

(ii) 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;

The regression equation is

$y = -4.58 + 0.155\,x1 + 0.0682\,x2 - 0.000940\,x_1^2$

Predictor	Coef	SE Coef	T	P	VIF
Constant	-4.575	1.517	-3.02	0.008	
x1	0.15506	0.04724	3.28	0.004	165.3
x2	0.06817	0.01719	3.97	-----	2.6
x_1^2	-0.0009398	0.0003682	-2.55	-----	167.2

R-Sq = ------   R-Sq(adj) = -----

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	5.0959	1.6986	74.85	-------
Residual Error	17	0.3858	0.0227		
Total	20	5.4817			

<u>ΖΗΤΗΜΑ 3</u>: (Βαθμ. 2.0) <u>Υποχρεωτικό</u>

Εξετάζεται η γραμμική παλινδρόμηση μιας μεταβλητής y σε σχέση με 6 επεξηγηματικές μεταβλητές $x_1, x_2, \ldots, x_6$, σε δείγμα μεγέθους n=20. Δίνονται αποτελέσματα προσαρμογών διαφόρων μοντέλων με επιλεγμένες μεταβλητές. Ο παρακάτω πίνακας παρουσιάζει μερικούς δείκτες για την προσαρμογή των μοντέλων αυτών.

(i) Επιλέξτε <u>δύο εμφωλευμένα μοντέλα</u> που με βάση τα κριτήρια θεωρείτε ότι είναι τα καλύτερα.

(ii) Στη συνέχεια αξιοποιώντας τον έλεγχο F για τη σύγκριση δύο <u>εμφωλευμένων</u> μοντέλων, καθώς και το κριτήριο AIC (χωρίς να υπολογίσετε τους κοινούς όρους), να βρεθεί το βέλτιστο μοντέλο από τα παραπάνω δύο.

[**Δίνονται:** $S=\left(SSE/(n-k-1)\right)^{1/2}$, $AIC=n\left[\ln(2\pi)+\ln(SSE/n)+1\right]+2(p+1)$]

Μοντέλο	Μεταβλητές	Υ με	R^2 (x100%)	$R^2_{πρόβλεψη}$ (x100%)	C_p	S
1	1	X_1	75.0	67.54	17.8	0.62838
2	1	X_6	40.2	21.23	64.8	0.97184
3	2	$X_1\ X_2$	85.4	80.55	5.7	0.49390
4	2	$X_1\ X_4$	82.6	75.27	9.5	0.53883
5	3	$X_1\ X_2\ X_4$	89.9	83.75	1.6	0.42265
6	3	$X_1\ X_2\ X_5$	86.0	79.00	6.9	0.49841
7	4	$X_1\ X_2\ X_4\ X_5$	90.3	81.97	3.1	0.42842
8	4	$X_1\ X_2\ X_4\ X_6$	90.0	81.95	3.5	0.43520
9	5	$X_1\ X_2\ X_4\ X_5\ X_6$	90.3	79.68	5.0	0.44279
10	5	$X_1\ X_2\ X_3\ X_4\ X_5$	90.3	80.01	5.1	0.44335
11	6	$X_1\ X_2\ X_3\ X_4\ X_5\ X_6$	90.4	78.10	7.0	0.45864

ΖΗΤΗΜΑ 4

(4Α) Έστω μοντέλο παλινδρόμησης Poisson $f(y)=\dfrac{\exp(-\mu_x)\,\mu_x^y}{y!}$, $y=0,1,2,\ldots$, με συνάρτηση σύνδεσης $g(\mu_x)=\ln\mu_x=\boldsymbol{\beta}'\mathbf{x}$ και

ελεγχοσυνάρτηση Deviance $=-2\left(\hat{\ell}_M-\hat{\ell}_{\kappa o \rho}\right)=2\sum_{i=1}^{n}\left[y_i\ln(y_i/\hat{\mu}_i)\right]$, όπου $\hat{\ell}_M$ η μεγιστοποιημένη λογαριθμοποιημένη συνάρτηση

πιθανοφάνειας του μοντέλου Μ που μας ενδιαφέρει και κριτήριο $\text{AIC}=-2\hat{\ell}_M+2d$, όπου d ο συνολικός αριθμός παραμέτρων στο μοντέλο.

(4Β) Προσαρμόζονται μοντέλα της παλινδρόμησης Poisson σε n=30 αεροσκάφη δύο τύπων Α και Β και εξετάζεται η σχέση του αριθμού ζημιών (Υ) ανά αεροσκάφος, με τις συμμεταβλητές X_1 (=1, τύπος Α και =0, τύπος Β) και X_2 (βάρος βομβών σε τόνους), καθώς και με τη X_3 (μήνες εμπειρίας του πληρώματος).

(i) Να συμπληρωθούν οι παρακάτω πίνακες.

(ii) Συγκρίνετε τα τρία μοντέλα με την ελεγχοσυνάρτηση Deviance (με διαδοχική αφαίρεση) και με το κριτήριο AIC και γράψτε το προσαρμοσμένο τελικό μοντέλο.

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

(iv) Ενισχύστε τα συμπεράσματά σας με τις πιο κάτω γραφικές παραστάσεις του **<u>τελικού μοντέλου</u>**.

ΜΟΝΤΕΛΟ: 3 Μεταβλητές	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	p-τιμή	$\exp(\hat{\beta}_j)$
Σταθερά	-0.406	0.877	-0.463	0.644	
X_1	0.569	0.504	1.128	0.2595	
X_2	0.165	0.068			
X_3	-0.014	0.008			

Ελεγχοσυνάρτηση deviance δίνεται ως D_3= 25.95 και η τιμή του κριτηρίου AIC_3=87.65

ΜΟΝΤΕΛΟ: 2 Μεταβλητές	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	p-τιμή	$\exp(\hat{\beta}_j)$
Σταθερά	-0.699	0.853	-0.819	0.413	
X_2	0.222	0.046			
X_3	-0.012	0.008			

Ελεγχοσυνάρτηση deviance δίνεται ως D_2= 27.22 και η τιμή του κριτηρίου AIC_2=86.92

ΜΟΝΤΕΛΟ: 1 Μεταβλητές	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	p-τιμή	$\exp(\hat{\beta}_j)$
Σταθερά	-1.70	0.507	-3.356	<0.001	
X_2	0.231	0.047			

Ελεγχοσυνάρτηση deviance δίνεται ως D_1=29.21
με αντίστοιχη τιμή $\hat{\ell}_1 =$_____ και τιμή του κριτηρίου **AIC_1=_______**

Για το μοντέλο χωρίς καμμία συμμεταβλητή, μόνο με το σταθερό όρο, **D_0= 53.88**
με αντίστοιχη τιμή $\hat{\ell}_0$ =-53.79 και **AIC_0=_________**

ΖΗΤΗΜΑ 5

(5Α) Εστω Υ τ.μ. της Διωνυμικής κατανομής $f(y)=\binom{n}{y}p^{y}(1-p)^{n-y}$, $y=0,1,2,...,n$, με παραμέτρους p και n.

(i) Γράψτε το μοντέλο της λογιστικής παλινδρόμησης για k συμμεταβλητές .

(ii) Δίνεται η ελεγχοσυνάρτηση deviance ως $D(\hat{\boldsymbol{\beta}})=2\sum_{i=1}^{m}\left\{y_i\ln\left(\frac{y_i}{\hat{\mu}_i}\right)+(n_i-y_i)\ln\left(\frac{n_i-y_i}{n_i-\hat{\mu}_i}\right)\right\}$, $\hat{\mu}_i=n_i\hat{p}_i$.

Δώστε τον ορισμό υπολοίπων Deviance για το μοντέλο της λογιστικής παλινδρόμησης.

(5Β) Σε μελέτη $m=165$ ασθενών, γιατρός θέλει να εξετάσει αν ασθενής πάσχει από καρδιοπάθεια Υ (ναι=1, όχι=0), σε σχέση με το φύλο (X_1, 1=άνδρας, 0=γυναίκα), με πόνο στο στήθος (X_2 , 1=μεγάλος, 2=μέτριος, 3=μικρός) και με το ζάχαρο (X_3, 1=ναι, 0=όχι).

Για τις τρεις κατηγορίες της X_2 κατασκευάζονται δύο δείκτριες μεταβλητές ως ακολούθως $X_2(2)=\begin{cases}1, & \text{αν τύπου 2}\\ 0, & \text{αλλιώς}\end{cases}$, $X_2(3)=\begin{cases}1, & \text{αν τύπου 3}\\ 0, & \text{αλλιώς}\end{cases}$, με κατηγορία 1 ως κατηγορία αναφοράς.

(i) Να συμπληρωθεί ο παρακάτω πίνακας. Κάνοντας χρήση του ελέγχου Wald και των ελεγχοσυναρτήσεων deviance, εξετάστε αν οι συμμεταβλητές X_j συμβάλλουν στα μοντέλα αυτά.

(ii) Να κατασκευαστεί ένα 95% διάστημα εμπιστοσύνης για την ποσότητα του e^{β_1} του **τελικού μοντέλου**.

(iii) Με τη βοήθεια της ποσότητας $e^{\hat{\beta}_1}$, εκφράστε κατά πόσο το φύλο επιδρά στη σχετική πιθανότητα ύπαρξης καρδιοπάθειας ενός ατόμου $\frac{p_x}{1-p_x}$ για το **τελικό μοντέλο**.

<u>ΜΟΝΤΕΛΟ: 1</u> **Μεταβλητές**	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	**p-τιμή**	$\exp(\hat{\beta}_j)$
Σταθερά	0.175	0.354	0.49	0.622	
X_1	1.347	0.421	3.20		
$X_2(2)$	-2.519	0.461	-5.47		
$X_2(3)$	-2.089	0.578	-3.61	<0.001	
X_3	0.392	0.614			

Ελεγχοσυνάρτηση deviance δίνεται ως **D_1= 172.96** και η τιμή του κριτηρίου **AIC_1=182.96**

<u>ΜΟΝΤΕΛΟ: 2</u> **Μεταβλητές**	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	**p-τιμή**	$\exp(\hat{\beta}_j)$
Σταθερά	0.223	0.345	0.65	0.518	
X_1	1.329	0.419	3.18	0.0015	
$X_2(2)$	-2.504	0.459			
$X_2(3)$	-2.072	0.576			

Ελεγχοσυνάρτηση deviance δίνεται ως **D_2= 173.37**

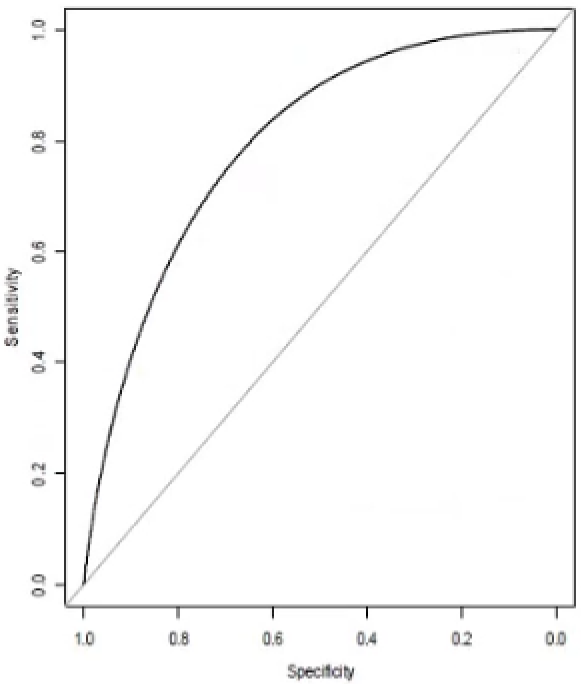
με αντίστοιχη τιμή $\hat{\ell}_2 = -86.685$ και τιμή του κριτηρίου **AIC_2=** _______

<u>ΜΟΝΤΕΛΟ: 3</u> **Μεταβλητές**	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	**p-τιμή**	$\exp(\hat{\beta}_j)$
Σταθερά	-0.511	0.298	-1.71	0.087	
X_1	1.091	0.355	3.07		

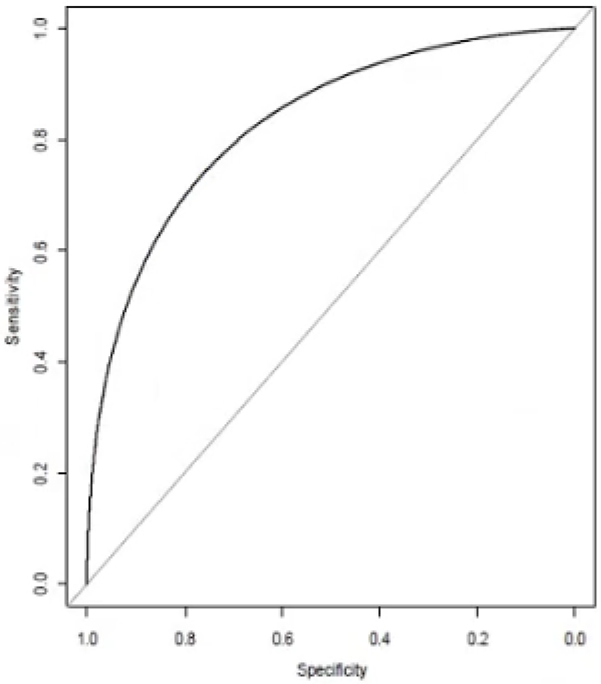
Ελεγχοσυνάρτηση deviance δίνεται ως **D_3= 216.27** και η τιμή του κριτηρίου **AIC_3= 220.27**

(iv) Για τα **Μοντέλα 1, 2 και 3** κατασκευάζονται οι ακόλουθες καμπύλες ROC. Πώς ερμηνεύονται τα παρακάτω αποτελέσματα; <u>AUC =Area under the curve</u>

<u>ΜΟΝΤΕΛΟ</u> 1 AUC= 0.7963



<u>ΜΟΝΤΕΛΟ 2</u> AUC= 0.8302



<u>ΜΟΝΤΕΛΟ 3</u> AUC= 0.6116