Εξέταση στο μεταπτυχιακό μάθημα: Στατιστική Μοντελοποίηση

******* Διάρκεια Εξέτασης : 2.30 ώρες *******

ΖΗΤΗΜΑ Α (Επιλέξτε 1 από 5) (Βαθμ. 2.5)

(Α1) Έστω γενικό γραμμικό μοντέλο **y=Xβ+ε**, όπου **X** ο πίνακας σχεδιασμού, $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ και εκτιμήτρια ελαχίστων τεταραγώνων $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Να βρεθεί η κατανομή των υπολοίπων $\mathbf{e}$, η μέση τιμή $E(\mathbf{e})$ και ο πίνακας διασποράς-συνδιασποράς των, $V(\mathbf{e})$. Στη συνέχεια βρείτε τη διασπορά του υπολοίπου e_i και τη συνδιασπορά μεταξύ των e_i και e_j, $i \neq j$, όπου $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

(Α2) Έστω γενικό γραμμικό μοντέλο **y=Xβ+ε**.

(i) Με βάση τη σ.π.π. της παρατήρησης y_i, $f(y_i)=\frac{1}{\left(2\pi\sigma^2\right)^{1/2}}\exp\left[\frac{-\left(y_i-\boldsymbol{\beta}'\mathbf{x}_i\right)^2}{2\sigma^2}\right]$, $y_i \in \mathbb{R}$, $i=1,2,...,n$, δείξτε ότι η μεγιστοποιημένη λογαριθμοποιημένη συνάρτηση πιθανοφάνειας για το γενικό γραμμικό μοντέλο είναι η $\hat{\ell} = -\frac{n}{2}[\ln(2\pi\, SSE/n)+1]$ δεδομένου ότι η ε.μ.π. της σ^2 είναι η $\hat{\sigma}^2=\sum_{i=1}^{n}\left(y_i-\hat{\boldsymbol{\beta}}'\mathbf{x}_i\right)^2/n=\frac{SSE}{n}$.

(ii) Δώστε τον ορισμό του κριτηρίου AIC. Στη συνέχεια δείξτε ότι το κριτήριο AIC είναι $AIC=n\left[\ln(2\pi)+\ln(SSE/n)+1\right]+2(p+1)$ για το μοντέλο αυτό. Πώς χρησιμοποιείται ;

(Α3) Εστω μοντέλο $E(Y)=\beta_0+\beta_1X_1+\beta_2X_2$. Περιγράψτε τα βήματα που θα ακολουθούσατε όταν εξετάζεται η <u>αφαίρεση</u> μιας επεξηγηματικής μεταβλητής από το μοντέλο αυτό.

(Α4) Δώστε τον ορισμό ενός τυποποιημένου (standardized) υπολοίπου r_i και ενός deleted υπολοίπου r_i'. Πώς μας βοηθούν τα υπόλοιπα r_i και r_i', τα διαγώνια στοιχεία h_{ii} του πίνακα $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, καθώς και η απόσταση Cook $D_i=\frac{r_i^2 h_{ii}}{p(1-h_{ii})}$ στη διάγνωση ενός γενικού γραμμικού μοντέλου;

(Α5) Περιγράψτε πώς μέσω μιας ψευδομεταβλητής Z (=1, αν τα δεδομένα ανήκουν στην κατηγορία I και =0, 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.

ΖΗΤΗΜΑ Β (Υποχρεωτικό) (Βαθμ. 4.0)

Για τη διερεύνηση της εξάρτησης μιας μεταβλητής y από 5 επεξηγηματικές μεταβλητές $x_1, x_2, \ldots, x_5$, προσαρμόστηκε ένα γενικό γραμμικό μοντέλο σε δείγμα μεγέθους n=60. Ακολουθούν τα βασικά σημεία της ανάλυσης (**Β1, Β2, Β3**).

<u>Β1 ανάλυση</u>: περιλαμβάνει όλες τις επεξηγηματικές μεταβλητές. Τα παρακάτω περιέχουν μερικά από τα αποτελέσματα της ανάλυσης αυτής. Συμπληρώστε και σχολιάστε τα αποτελέσματα.

Regression Analysis: y versus x1; x2; x3; x4; x5

```
The regression equation is
y = 99.4 - 2.02 x1 - 33.8 x2 + 0.456 x3 - 3.39 x4 - 0.0012 x5

Predictor       Coef  SE Coef      T      P  VIF
Constant       99.43    19.68   5.05  0.000
x1           -2.0175   0.9007  -2.24  0.029  1.9
x2           -33.831    5.243  -6.45 <0.001  3.6
x3           0.45620  0.03926  11.62 <0.001  5.9
x4            -3.390    3.574  -0.95  0.347  3.7
x5          -0.00116  0.03180  -0.04  0.971  6.2

R-Sq = 91.0%    R-Sq(adj) = 90.1%

Analysis of Variance

Source           DF       SS      MS       F      P
Regression        5   947930  189586  108.82 <0.001
Residual Error   54    94079    1742
Total            59  1042009
```

<u>B2 ανάλυση</u>: δίνονται αποτελέσματα προσαρμογών διαφόρων μοντέλων με επιλεγμένες μεταβλητές. Ο παρακάτω πίνακας παρουσιάζει μερικούς δείκτες για την προσαρμογή των μοντέλων αυτών.

(i) Δώστε τον ορισμό του διορθωμένου συντελεστή $\bar{R}^2$.

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

(iii) Στη συνέχεια αξιοποιώντας τον έλεγχο F για τη σύγκριση δύο <u>εμφωλευμένων</u> μοντέλων, να βρεθεί το βέλτιστο από τα παραπάνω δύο.

[**Δίνεται:** $S=\left(SSE/(n-k-1)\right)^{1/2}$]

Μοντέλο	Μεταβλητές	Y με	R^2 (x100%)	$\bar{R}^2$ (x100%)	C_p	S	AIC
1	1	X_3	75.5	75.1	90.4	66.316	505.30
2	1	X_5	58.7	58.0	190.9	86.113	536.65
3	2	$X_2\ X_3$	90.1	89.7	5.4	42.610	453.17
4	2	$X_3\ X_4$	82.3	81.7	51.9	56.892	487.86
5	3	$X_1\ X_2\ X_3$	90.7	90.2	3.4	41.497	450.93
6	3	$X_2\ X_3\ X_4$	90.1	89.6	7.0	42.851	454.79
7	4	$X_1\ X_2\ X_3\ X_4$	91.0	90.3	4.0	41.359	451.45
8	4	$X_1\ X_2\ X_3\ X_5$	90.8	90.2	4.9	41.702	452.44
9	5	$X_1\ X_2\ X_3\ X_4\ X_5$	91.0	90.1	6.0	41.740	453.45

Τα δύο καλύτερα δείχνουν να είναι το M5 και το M7, συγκρίνουμε

$$F=\frac{SSE_5\text{-}SSE_7}{SSE_7/(n-p)}=\frac{96431-94081}{94081/55}=1.36\simeq(-1.17)^2$$, p-value=P(F>1.36)=0.246, επιλέγω το μοντέλο 5

Συγκρίνοντας και το M3 με M5

$$F=\frac{SSE_3\text{-}SSE_5}{SSE_5/(n-p)}=\frac{103489-96431}{96431/56}=4.09876\simeq(-2.02)^2$$, p-value=P(F>)=**0.048**, επιλέγω το μοντέλο 5

<u>B3 ανάλυση</u>: περιλαμβάνει τις επεξηγηματικές μεταβλητές X_1, X_2, X_3. Τα παρακάτω περιέχουν μερικά από τα αποτελέσματα της ανάλυσης αυτής. Συμπληρώστε τον παρακάτω πίνακα και σχολιάστε τα αποτελέσματα καθώς και τις γραφικές παραστάσεις που ακολουθούν για το μοντέλο αυτό.

<u>(B3) (α):</u>

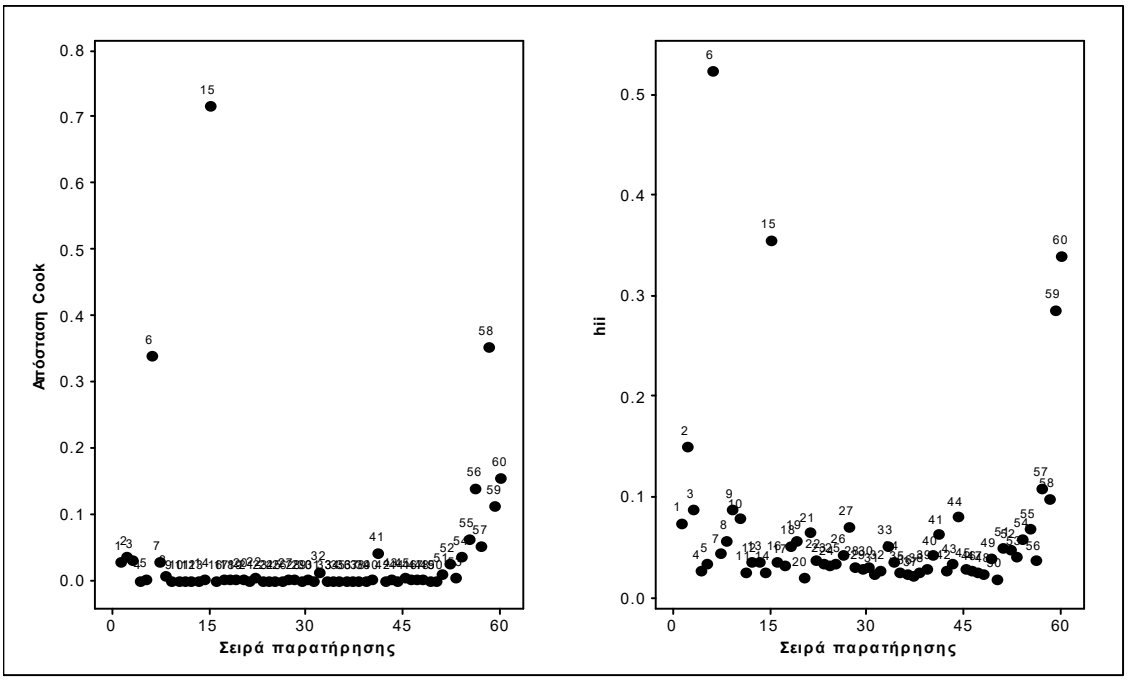
Regression Analysis: y versus x1; x2; x3 (Μοντέλο 5)

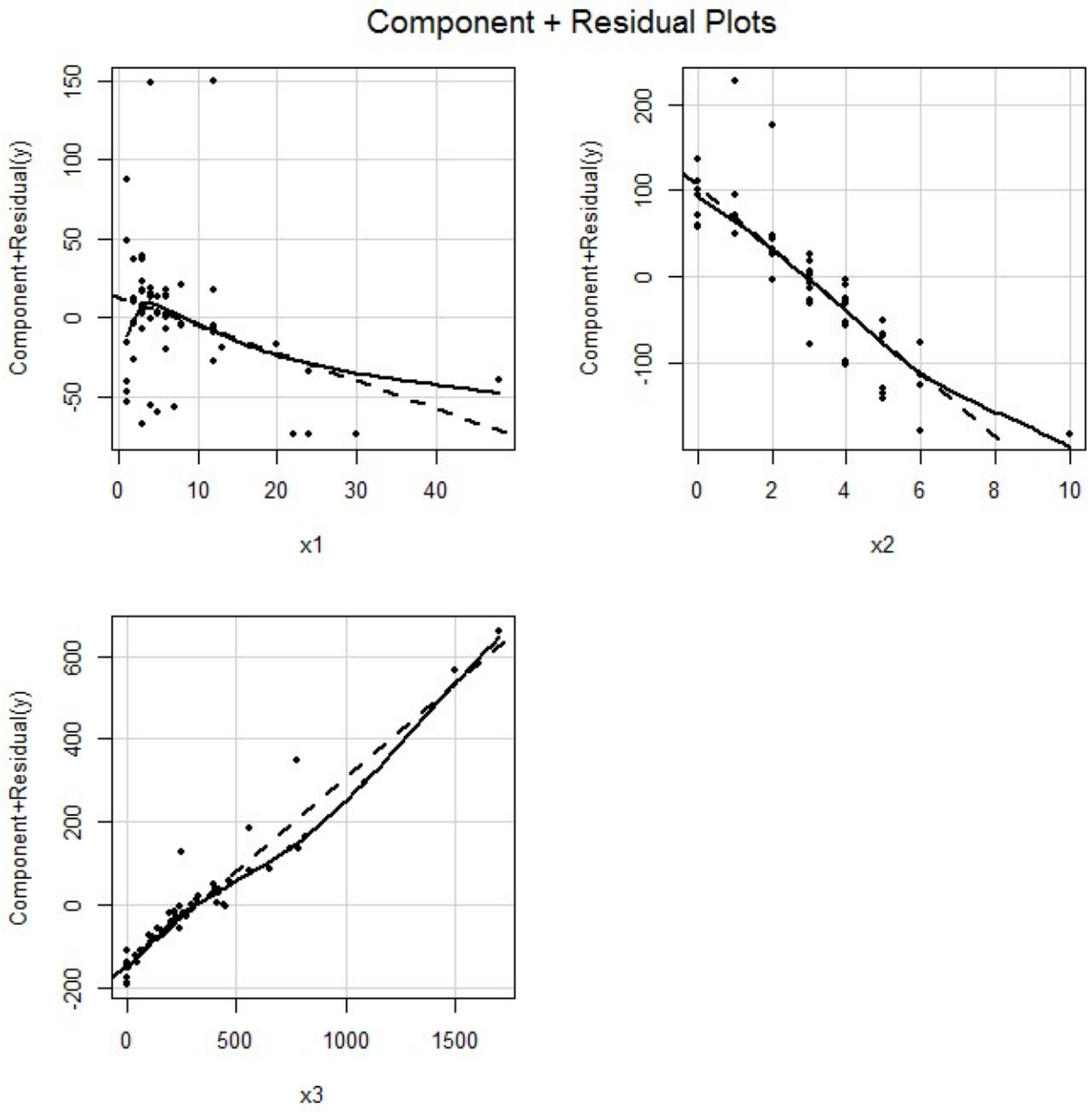
```
The regression equation is
y = 87.2 - 1.75 x1 - 36.5 x2 + 0.451 x3

Predictor      Coef  SE Coef      T      P   VIF
Constant      87.20    16.26   5.36  <0.001
x1         -1.7528   0.8658  -2.02  0.048   1.8
x2         -36.494    4.019  -9.08  <0.001  2.1
x3         0.45083  0.02027  22.25  <0.001  1.6

R-Sq = 90.7%   R-Sq(adj) = 90.2%
```

<u>(B3) (β)</u> <u>Γραφικές παραστάσεις</u>:

(i) Απόσταση Cook και διαγώνια στοιχεία h_{ii} του πίνακα $\mathbf{H}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, $2p/n=0.133\dot{3}$

(ii) Γραφικές παραστάσεις των μερικών υπολοίπων για τις μεταβλητές X_1, X_2, X_3.


Απόσταση Cook
Σειρά παρατήρησης
hii
Σειρά παρατήρησης


Component + Residual Plots
Component+Residual(y)
x1
Component+Residual(y)
x2
Component+Residual(y)
x3

Επιλέξτε <u>ΕΝΑ</u> από τα ακόλουθα 3 Ζητήματα (Γ ή Δ) <u>(Βαθμ. 3.5)</u>

<u>ΖΗΤΗΜΑ Γ:</u> (Γ1) Έστω το μοντέλο της κατανομής Poisson $f(y)=\frac{e^{-\mu}\mu^{y}}{y!}$, $y=0, 1, 2, \ldots..$

(i) Δείξτε ότι ανήκει στην εκθετική οικογένεια κατανομών και πώς μέσα από αυτήν αναγνωρίζουμε τη συνάρτηση σύνδεσης .

(ii) Γράψτε το μοντέλο παλινδρόμησης Poisson για τρεις συμμεταβλητές X_1 , X_2 και X_3 .

(Γ2) Προσαρμόζοντας μοντέλα της παλινδρόμησης Poisson σε δεδομένα 44 ορυχείων μιας περιοχής, εξετάζεται η σχέση του αριθμού ρωγμών σε οροφή ορυχείου (Υ), με τις συμμεταβλητές X_1 και X_2 (χαρακτηριστικά του ορυχείου) καθώς και με την X_3 (έτη λειτουργίας του ορυχείου).

(i) Αφού συμπληρωθούν οι παρακάτω πίνακες, να ερμηνευτούν οι εκτιμημένες ποσότητες $\exp(\hat{\beta}_j)$ καθώς και οι γραφικές παραστάσεις των υπολοίπων Pearson και Deviance που ακολουθούν, του **<u>τελικού μοντέλου</u>**.

Απ:

αν αυξηθεί το Χ1 κατά μία μονάδα θα πολλαπλασιαστεί η αναμενόμενος αριθμός των ρωγμών στην οροφή ενός ορυχείου με 1.0605, δηλαδή θα αυξηθεί κατά 6,1%.

αν αυξηθεί το Χ3 κατά μία μονάδα θα πολλαπλασιαστεί η αναμενόμενος αριθμός των ρωγμών στην οροφή ενός ορυχείου με 0.965, δηλαδή θα μειωθεί κατά 3.5%.

(ii) Συμφωνούν οι έλεγχοι Wald, Deviance και τα κριτήρια AIC;

<u>ΜΟΝΤΕΛΟ: 1</u> 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	5.008	<0.001	
X_2	-0.00376	0.0049	-0.761	0.4464	
X_3	-0.03408	0.0147	-2.326	0.02	
Ελεγχοσυνάρτηση deviance δίνεται ως **D_1=41.329** και η τιμή του κριτηρίου **AIC_1=145.6**					

<u>ΜΟΝΤΕΛΟ: 2</u> 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	5.030	<0.001	1.0605
X_3	-0.03563	0.0148	-2.405	0.0162	0.965
Ελεγχοσυνάρτηση deviance δίνεται ως **D_2=41.952** και η τιμή του κριτηρίου **AIC_2=144.22**					

<u>ΜΟΝΤΕΛΟ: 3</u> 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	**4.721**	<0.001	
Ελεγχοσυνάρτηση deviance δίνεται ως **D_3=48.620** και η τιμή του κριτηρίου **AIC_3=148.89**					

<u>Επιλέγουμε το Μοντέλο 2</u>

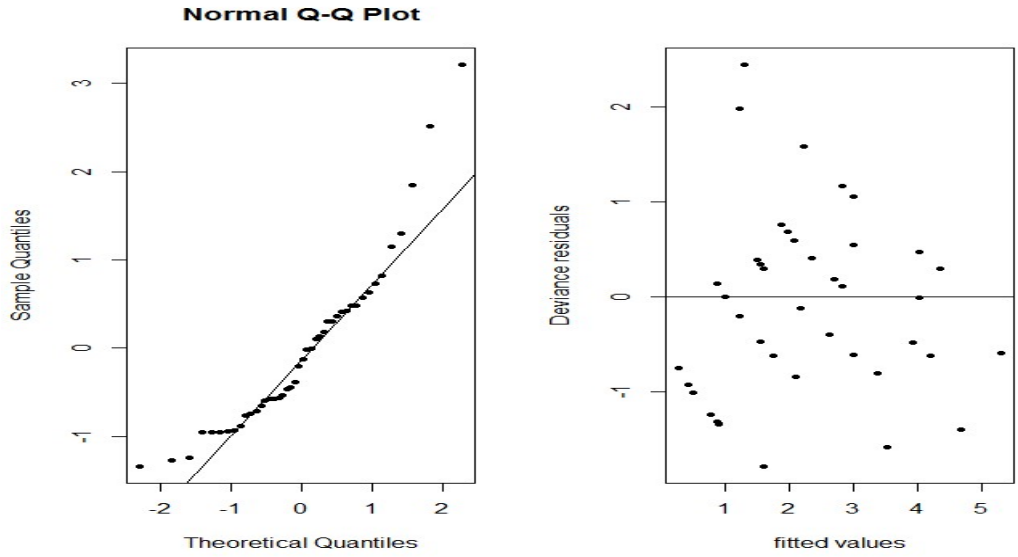
$D_2 - D_1 = 41.952 - 41.329 = 0.623$,

p-value= pchisq(0.623,1,lower.tail=FALSE)= 0.4299

$D_3 - D_2 = 48.620 - 41.952 = 6.668$,

p-value=pchisq(6.668,1,lower.tail=FALSE)= 0.00982

<u>Γραφικές παραστάσεις του τελικού μοντέλου</u> (η πρώτη είναι για τα υπόλοιπα Pearson)



<u>ΖΗΤΗΜΑ Δ</u>

(Δ1) Εστω Y 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 τέσσερις συμμεταβλητές X_1, X_2, X_3 και X_4.

(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. Αναφέρετε δύο γραφικές παραστάσεις που βοηθούν στην αξιολόγηση ενός τέτοιου μοντέλου.

(Δ2) Σε μελέτη $n=200$ ασθενών, γιατρός θέλει να εξετάσει την επιβίωση ασθενή Y (ναι=1, όχι=0), σε σχέση με την ηλικία (X_1), την πίεση (X_2), το σφυγμό (X_3) και την κατεπείγουσα εισαγωγή (X_4) (ναι=1, όχι=0).

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

(ii) Να κατασκευαστεί ένα 95% διάστημα εμπιστοσύνης για την ποσότητα του e^{β_4} του <u>τελικού μοντέλου</u>.

(iii) Με τη βοήθεια της ποσότητας $e^{\hat{\beta}_1}$, εκφράστε κατά πόσο η αύξηση της ηλικίας κατά ένα έτος, επιδρά στη σχετική πιθανότητα επιβίωσης ενός ατόμου $\frac{p_x}{1-p_x}$ για το <u>τελικό μοντέλο</u>.

Απ: αν αυξηθεί η ηλικία κατά ένα έτος θα πολλαπλασιαστεί το odds ενός ατόμου με 0.96662, δηλαδή θα μειωθεί η σχετική πιθανότητα επιβίωσής του κατά 3.3%.

<u>ΜΟΝΤΕΛΟ: 1</u> Μεταβλητές	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	**p-τιμή**	$\exp(\hat{\beta}_j)$
Σταθερά	3.16121	1.43286	2.206	0.02737	
X_1	-0.03522	0.01097	-3.211	0.00132	
X_2	0.01377	0.00608	2.266	0.02342	
X_3	0.00639	0.00734	0.870	0.38403	
X_4	-2.41432	0.77599	-3.111	0.00186	
Ελεγχοσυνάρτηση deviance δίνεται ως **D_1= 167.05** και η τιμή του κριτηρίου **AIC_1= 177.05**					

ΜΟΝΤΕΛΟ: 2 **Μεταβλητές**	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	**p-τιμή**	$\exp(\hat{\beta}_j)$
Σταθερά	3.67855	1.30652	2.816	0.00487	
X_1	-0.03395	0.01087	-3.124	0.00178	0.96662
X_2	0.01323	0.00599	2.210	0.02709	1.01332
X_4	-2.28763	0.75817	-3.017	0.00255	0.101507
Ελεγχοσυνάρτηση deviance δίνεται ως **D_2=167.82** με αντίστοιχη τιμή $\hat{\ell}_2 = -83.91$ και τιμή του κριτηρίου **AIC_2= 175.82**					
ΜΟΝΤΕΛΟ: 3 **Μεταβλητές**	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	**p-τιμή**	$\exp(\hat{\beta}_j)$
Σταθερά	5.50876	1.03351	5.330	<0.001	
X_1	-0.03402	0.01069	-3.181	0.00147	
X_4	-2.45354	0.75257	-3.260	0.00111	
Ελεγχοσυνάρτηση deviance δίνεται ως **D_3=173.08** και η τιμή του κριτηρίου **AIC_3=179.08**					

Επιλέγουμε το Μοντέλο 2

$D_2 - D_1 = 167.82 - 167.05 = 0.77$,

p-value=P(X^2>0.77)=pchisq(0.77,1,lower.tail=FALSE)= 0.380

$D_3 - D_2 = 173.08 - 167.82 = 5.26$,

p-value=P(X^2>5.26)= pchisq(5.26,1,lower.tail=FALSE)= 0.0218

95% for β_4

$-3.773606719 < \beta_4 < -0.80164269$

95% for e^{β_4}

$0.02296907 < e^{\beta_4} < 0.4485915$