Αρ. Μητρώου: 03400131 Ονοματεπώνυμο: Δεπάσας Βασίλειος

Εξέταση στο Μεταπτυχιακό Μάθημα: Στατιστική Μοντελοποίηση (10/2/2022)

Επιλέξτε ΔΥΟ από τα 5 Ζητήματα ***** Διάρκεια Εξέτασης: 1.30 ώρες *****

ΖΗΤΗΜΑ 1

Ερευνάται η σχέση μεταξύ y (ποσότητα μετάλλου) και x_1 (ταχύτητα παραγωγής) για δύο παραγωγικές διαδικασίες 1 και 2. Έστω δείκτρια μεταβλητή x_2 (x_2=1 - αν διαδικασία=1, x_2=0 - αν διαδικασία=2).

Αφού συμπληρώσετε τα κενά στα ακόλουθα αποτελέσματα, εξηγήστε πώς μέσω του μοντέλου $E(y_x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, μπορούμε να ελέγξουμε αν χρειάζεται να προσαρμοστούν

(I) δύο διαφορετικές ευθείες, **(II)** δύο παράλληλες ευθείες, ή **(III)** μια κοινή ευθεία και για τις δύο παραγωγικές διαδικασίες, όπου $x_3 = x_1 x_2$, η μεταβλητή που εκφράζει την αλληλεπίδραση μεταξύ των μεταβλητών x_1 και x_2.

Να δοθούν ερμηνείες για το τελικό μοντέλο (βλ. και σχετικό διάγραμμα πιο κάτω).

Regression Analysis: y versus x1; x2; x3

The regression equation is
y = 7.6 + 1.32 x1 + 90.4 x2 - 0.177 x3

Predictor	Coef	SE Coef	T	P
Constant	7.57	20.87	0.36	0.720
x1	1.32205	0.09262	14.27	<0.001
x2	90.39	28.35	3.19	0.004
x3	-0.1767	0.1288	-1.37	0.183

R-Sq = 94.5% R-Sq(adj) = 93.78% R-Sq(pred) = 92.62%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	169165	56388	130.95	<0.001
Residual Error	23	9904	431		
Total	26	179069			

Regression Analysis: y versus x1; x2

The regression equation is
y = 27.3 + 1.23 x1 + 53.1 x2

Predictor	Coef	SE Coef	T	P
Constant	27.28	15.41	1.77	0.089
x1	1.23074	0.06555	18.77	<0.001
x2	53.129	8.210	6.47	$1.08 \cdot 10^{-6}$

R-Sq = 94% R-Sq(adj) = 93.5% R-Sq(pred) = 92.52%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	168355	84178	188.57	$2.10 \cdot 10^{-15}$
Residual Error	24	10714	446		
Total	26	179069			

Regression Analysis: y versus x1

The regression equation is
y = 64.0 + 1.20 x1

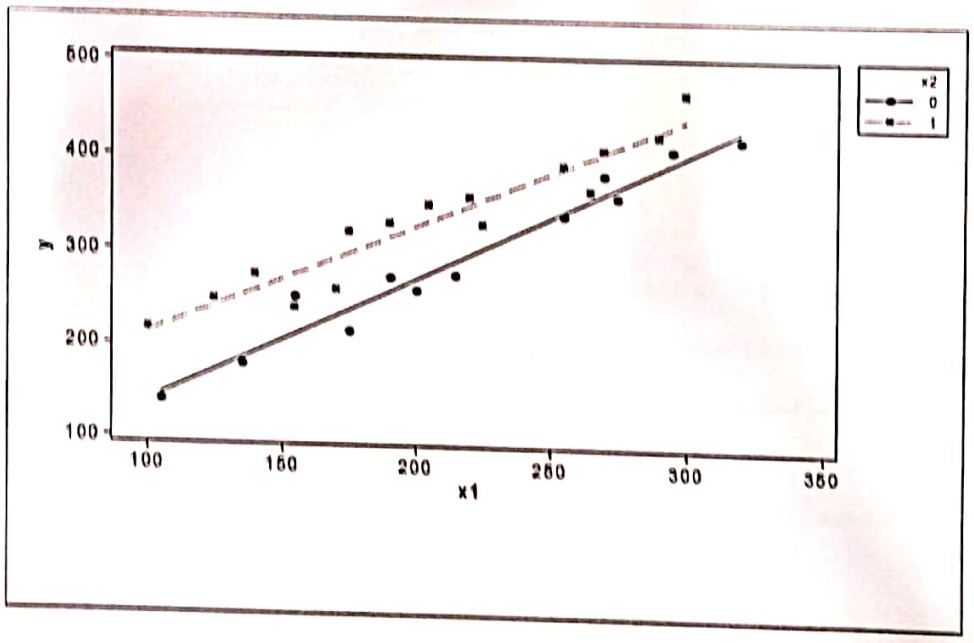
Predictor	Coef	SE Coef	T	P
Constant	64.04	23.25	2.75	0.011
x1	1.1963	0.1061	11.28	<0.001

PRESS = 34546.9 R-Sq = 83.6%

R-Sq(adj) = 82.9% R-Sq(pred) = 80.86%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	149661	149661	121.20	$1.15 \cdot 10^{-11}$
Residual Error	25	30811	1234.84		
Total	26	180,532			



ΖΗΤΗΜΑ 2

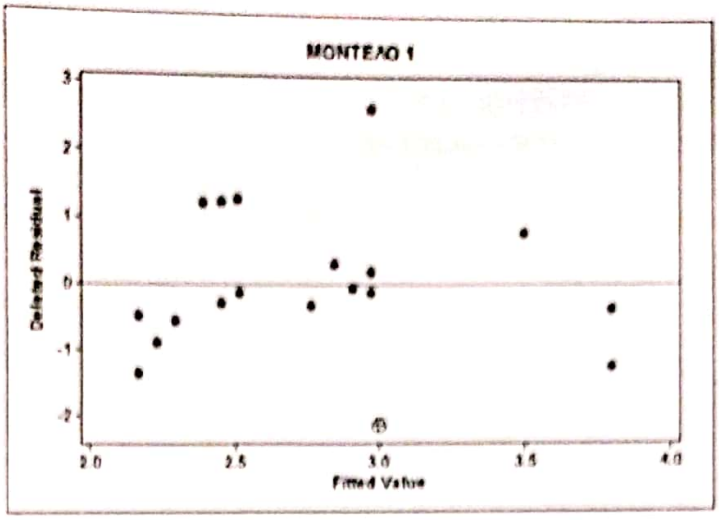
Α) Δώστε τον ορισμό ενός τυποποιημένου (standardized) υπολοίπου r_i και ενός deleted υπολοίπου r_i'. Πώς μας βοηθούν;

Β) 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, °C).

(Ι) Συμπληρώστε τον παρακάτω πίνακα και βρείτε το διορθωμένο δείκτη $\bar{R}^2$ = _____ %. Σχολιάστε τα αποτελέσματά σας.

[Δίνονται: S =0.172, $r_{X_1X_2}$ = 0.782, R^2 = 90.3 %, $R^2_{\text{πρόβλεψη}}$ = 85.9 %]

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



Για το παραπάνω μοντέλο δίνεται ότι $e_{20} = -0.29$, $h_{20,20} = 0.28$ και απόσταση Cook $D_{20} = \frac{r_{20}^2 h_{20,20}}{p(1-h_{20,20})}$.

(ii) Αποτελεί η παρατήρηση 20 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	_____	165.3
x2	0.06817	0.01719	3.97	0.001	2.6
X_1^2	-0.0009398	0.0003682	-2.55	_____	167.2

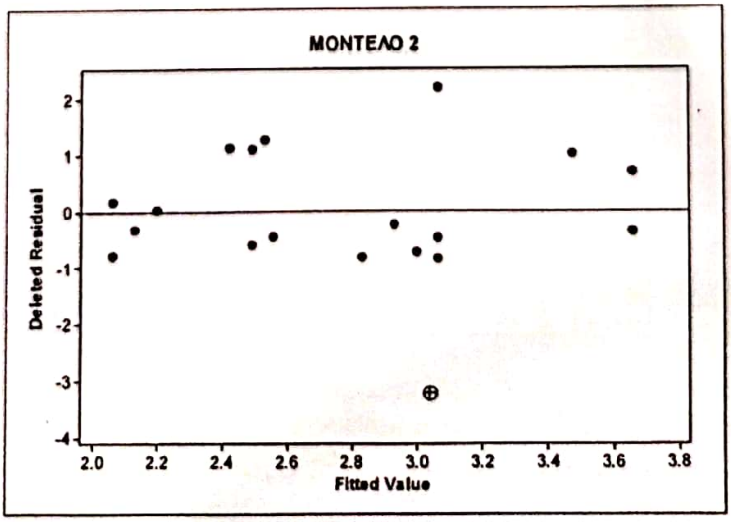
R-Sq = _____ R-Sq(adj) = _____

PRESS = 0.621669 R-Sq(pred) = _____

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			

(iv) Εξετάστε εκ νέου, αν η παρατήρηση 20 αποτελεί σημείο επιρροής για το νέο μοντέλο ($r_{20} = -2.59$, $h_{20,20} = 0.29$).



ΖΗΤΗΜΑ 3

Εξετάζεται η γραμμική παλινδρόμηση μιας μεταβλητής y, σε σχέση με 5 επεξηγηματικές μεταβλητές $x_1, x_2, \ldots, x_5$. Ακολουθούν τα βασικά σημεία της ανάλυσης.

Α ανάλυση: περιλαμβάνει όλες τις επεξηγηματικές μεταβλητές. Συμπληρώστε τον παρακάτω πίνακα και σχολιάστε σύντομα τα αποτελέσματα της ανάλυσης αυτής.

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

The regression equation is
y = 4.4 - 0.0003 x1 + 0.0016 x2 + 2.60 x3 + 0.219 x4 - 0.00953 x5

Predictor	Coef	SE Coef	T	P
Constant	4.39	15.24	0.29	0.776
x1	-0.00031	0.03193	-0.01	0.992
x2	0.00161	0.04080	0.04	_____
x3	2.603	1.754	1.48	0.151
x4	0.2190	0.1161	1.89	0.071
x5	-0.009534	0.004527	-2.11	_____

R-Sq = _____ R-Sq(adj) = 76.3% R-Sq(pred) = _____

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	913.90	182.78	19.62	_____
Residual Error	24	223.53	9.31		
Total	29	1137.43			

Β ανάλυση:

Δίνονται αποτελέσματα προσαρμογών διαφόρων μοντέλων με επιλεγμένες μεταβλητές. Ο παρακάτω πίνακας παρουσιάζει μερικούς δείκτες για την προσαρμογή των μοντέλων αυτών.

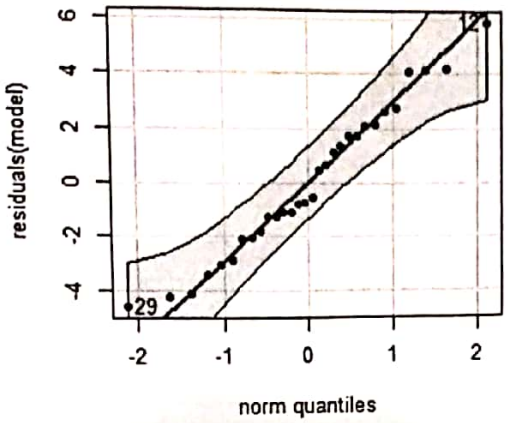
(i) Επιλέξτε δύο εμφωλευμένα μοντέλα που με βάση τα κριτήρια θεωρείτε ότι είναι τα καλύτερα.

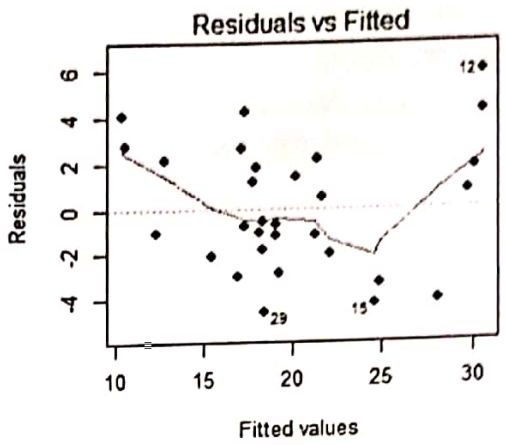
(ii) Στη συνέχεια αξιοποιώντας τον έλεγχο F για τη σύγκριση **δύο εμφωλευμένων** μοντέλων, καθώς και το δείκτη R^2 να βρεθεί το βέλτιστο μοντέλο από τα παραπάνω δύο.

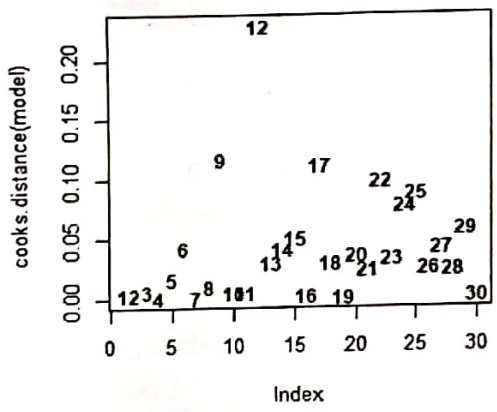
Δίνεται: $S=\left(SSE/(n-k-1)\right)^{1/2}$

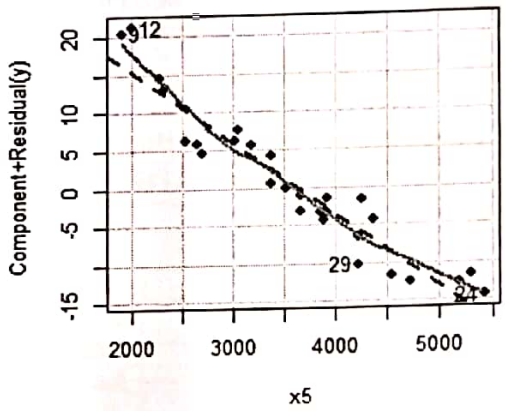
Μοντέλο	Μεταβλητές	Y με	$\bar{R}^2$ (x100%) (διορθ.)	$R^2_{\text{πρόβλεψη}}$ (x100%)	C_p	S	AIC
1	1	x1	0.751	0.707	3.3	3.1220	157.374
2	1	x5	0.717	0.668	7.4	3.3309	161.261
3	2	x4 x5	0.754	0.707	3.9	3.1040	157.938
4	2	x1 x5	0.748	0.697	4.7	3.1453	158.729
5	3	x3 x4 x5	0.781	0.736	2.0	2.9323	155.390
6	3	x1 x4 x5	0.757	0.705	4.7	3.0902	158.538
7	4	x2 x3 x4 x5	0.772	0.709	4.0	2.9902	157.387
8	4	x1 x3 x4 x5	0.772	0.716	4.0	2.9903	157.389
9	5	x1 x2 x3 x4 x5	0.763	0.681	6.0	3.0519	159.387

(iii) Σχολιάστε σύντομα τις παρακάτω γραφικές παραστάσεις των υπολοίπων, τις αποστάσεις Cook , καθώς και των μερικών υπολοίπων για τη μεταβλητή X_5 του **τελικού μοντέλου**.









(iv) <u>Αν θεωρήσουμε</u> ότι το **Μοντέλο 3** είναι το καλύτερο, να βρεθεί το πάνω άκρο του 0.95 – Δ.Ε. (**19.073**, _____) της πρόβλεψης ***μιας νέας παρατήρησης*** Y_{x_0}, όταν η σημειακή πρόβλεψη είναι $\hat{Y}_{x_0} = 25.73$ και $x_0'(X'X)^{-1}x_0 = 0.092488$.

<u>ΖΗΤΗΜΑ 4</u>

Έστω μοντέλο παλινδρόμησης Poisson $f(y)=\frac{\exp(-\mu_x)\,\mu_x^y}{y!}$, y=0,1,2, ..., με συνάρτηση σύνδεσης $g(\mu_x)=\ln\mu_x=\beta'x$ και ελεχγοσυνάρτηση Deviance $D_M(\hat{\beta}) = -2\left(\hat{\ell}_M - \hat{\ell}_{κορ}\right) = 2\sum_{i=1}^{n}[y_i \ln(y_i/\hat{\mu}_i)]$, όπου $\hat{\ell}_M$ η μεγιστοποιημένη λογαριθμοποιημένη συνάρτηση πιθανοφάνειας του μοντέλου M που μας ενδιαφέρει και $\hat{\ell}_{κορ}$ η αντίστοιχη του κορεσμένου μοντέλου και κριτήριο $AIC=-2\hat{\ell}_M+2d$, όπου d ο συνολικός αριθμός παραμέτρων στο μοντέλο.

Σε n=42 ομάδες ασθενών με κοινά χαρακτηριστικά εξετάζεται αν ο αριθμός (Y) ασθενών με θετική ανταπόκριση θεραπείας/ομάδα εξαρτάται από τη δοσολογία συγκεκριμένου φαρμάκου (X_1) και από το φύλο (X_2=1 αν γυναίκα, και X_2=0 αν άντρας).

(i) Να συμπληρωθούν οι p-τιμές του ελέγχου Wald στον παρακάτω πίνακα, καθώς και οι τιμές του κριτηρίου AIC.
(ii) Με βάση τον έλεγχο Wald, τη διαφορά των ελεγχοσυναρτήσεων Deviance, και λαμβάνοντας υπόψη το δείκτη Ψευδο-R_D^2 Deviance (βλ. πινακάκι πιο κάτω), καθώς και το κριτήριο AIC, επιλέξτε το καλύτερο από τα τρία μοντέλα **M0, M1, M2**. Γράψτε το προσαρμοσμένο <u>**τελικό μοντέλο**</u>.

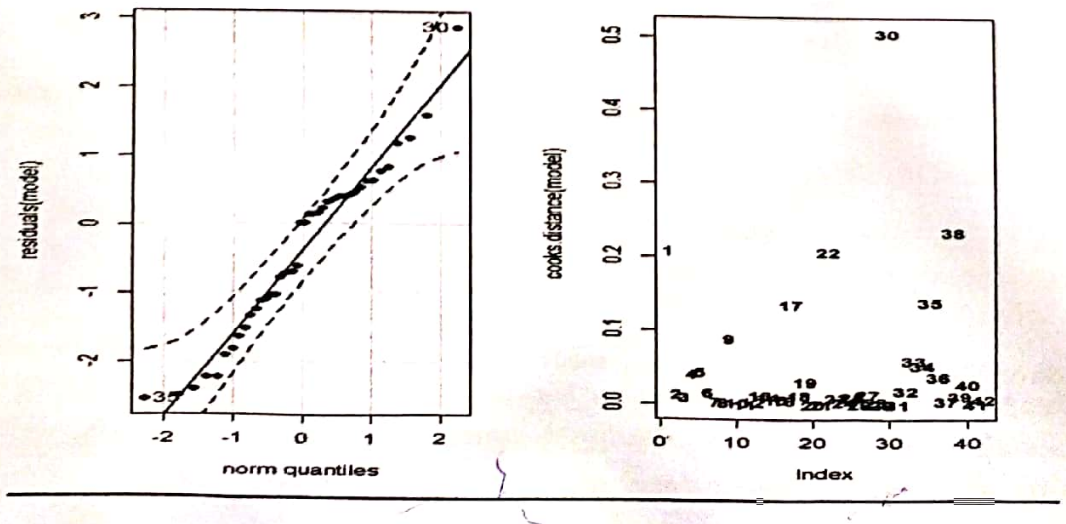
(iii) Κατασκευάστε 0.95-διαστήματα εμπιστοσύνης για τα $\exp(\beta_j)$ και με βάση αυτών ερμηνεύστε τις εκτιμημένες ποσότητες $\exp(\hat{\beta}_j)$ του **τελικού μοντέλου**.

(iv) Σχολιάστε σύντομα το γραφικό έλεγχο των υπολοίπων Deviance και τη γραφική παράσταση (index plot) της απόστασης Cook του **τελικού μοντέλου.**

ΜΟΝΤΕΛΟ: 2 Μεταβλητές	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	p-τιμή	Διαστήματα εμπιστοσύνης
Σταθερά	0.383749	0.1003	3.826	0.00013	XXXXXX
X_1	-0.129716	0.0087	-14.835	$8.1 \cdot 10^{-50}$	
X_2	-0.013193	0.0629	-0.210	0.834	
AIC_2=245.78					
ΜΟΝΤΕΛΟ: 1 Μεταβλητές	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	p-τιμή	
Σταθερά	0.37677	0.09462	3.982	<0.001	XXXXXX
X_1	-0.12962	0.00873	-14.848	$7.17 \cdot 10^{-50}$	[0.864, 0.894]
$\hat{\ell}_1$ = -120.115 και η τιμή του κριτηρίου AIC_1= 244.83					
ΜΟΝΤΕΛΟ: 0 Για το μοντέλο χωρίς συμμεταβλητές (Null model) $\hat{\ell}_0$ =-222.219 και η τιμή του κριτηρίου AIC_0= 446.44					

Μοντέλο	Deviance β.ε.	Deviance	Διαφορά στους β.ε.	Διαφορά Deviance	Pr(>Chi)	Deviance Ψευδο-R^2_D $R^2_D = 1 - \frac{D(\hat{\beta})}{D_0}$ $(\times 100\%)$
M0	41	272.305				
M1	40	67.695	1	204.61	$2.06 \cdot 10^{-46}$	75.14%
M2	39	67.652	1	0.043	0.836	75.16 %

Γραφικός έλεγχος των υπολοίπων Deviance και γράφημα δείκτη (index plot) της απόστασης Cook για το **τελικό μοντέλο**



ΖΗΤΗΜΑ 5

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

Γράψτε το μοντέλο της λογιστικής παλινδρόμησης για k συμμεταβλητές.

(5Β) Σε μελέτη 900 νεογνών, ερευνητής θέλει να εξετάσει αν ο αριθμός λιποβαρών νεογνών Y_i ανά ομάδα n_i με κοινά χαρακτηριστικά, σχετίζεται με την κοινωνικοοικονομική τάξη X_1, 0 (ανώτερη), 1 (μεσαία), 2 (χαμηλή) της μητέρας, με την κατανάλωση οινοπνευματωδών ποτών X_2, 0 (μεγάλη), 1 (μέτρια), 2 (χαμηλή) και με το αν η μητέρα καπνίζει (X_3=1 αν ναι και X_3=0 αν όχι). Για τη μεταβλητή X_1 κατασκευάζονται 2 δείκτριες μεταβλητές, με κατηγορία αναφοράς την ανώτερη (0). Για τη μεταβλητή X_2 κατασκευάζονται 2 δείκτριες μεταβλητές, με κατηγορία αναφοράς τη μεγάλη (0).

(i) Να συμπληρωθεί ο παρακάτω πίνακας (τα $\exp(\hat{\beta}_j)$ υπολογίζονται **μόνο** για το **τελικό μοντέλο**).

Κάνοντας χρήση του ελέγχου Wald, των ελέγχων deviance και του κριτηρίου AIC, επιλέξτε το καλύτερο μοντέλο.

(ii) Να κατασκευαστεί ένα 95% διάστημα εμπιστοσύνης για την ποσότητα του e^{β_3} του **τελικού μοντέλου**.

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

Με τη βοήθεια της ποσότητας $e^{\hat{\beta}_3}$ (odds ratio), εκφράστε κατά πόσο το κάπνισμα επιδρά στη σχετική πιθανότητα λιποβαρούς νεογνού $\frac{p_x}{1-p_x}$ για το **τελικό μοντέλο**.

ΜΟΝΤΕΛΟ: 1 Μεταβλητές	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	p-τιμή	$\exp(\hat{\beta}_j)$
Σταθερά	-1.93076	0.237765	-8.12	<0.001	XXXX
X_1 (1)	0.33955	0.268221	1.27	0.20554	
X_1 (2)	0.34905	0.282507	1.24	_____	
X_2 (1)	-0.59317	0.331750	-1.79	_____	
X_2 (2)	-0.78627	0.256567	-3.06	0.002	
X_3	0.56712	0.235797	_____	_____	
Ελεγχοσυνάρτηση deviance δίνεται ως D_1= 13.8104 και η τιμή του κριτηρίου AIC_1= 80.404					
ΜΟΝΤΕΛΟ: 2 Μεταβλητές	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	p-τιμή	$\exp(\hat{\beta}_j)$
Σταθερά	-1.81488	0.21994	-8.25	<0.001	XXXX
X_2 (1)	-0.568408	0.33017	-1.720	_____	
X_2 (2)	-0.689248	0.24706	-2.790	0.005	
X_3	0.647925	0.22440	2.890	0.004	
Ελεγχοσυνάρτηση deviance δίνεται ως D_2= 15.970 με αντίστοιχη τιμή $\hat{\ell}_2$ = _____ και η τιμή του κριτηρίου AIC_2= 78.563					
ΜΟΝΤΕΛΟ: 3 Μεταβλητές	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	p-τιμή	$\exp(\hat{\beta}_j)$
Σταθερά	-2.32769	0.136389	-17.07	<0.001	XXXX
X_3	0.70805	0.222079	_____	_____	
$\hat{\ell}_3$ = −39.04761 και η τιμή του κριτηρίου AIC_3= _____					

(iv) Ενισχύστε τα συμπεράσματά σας με τις ακόλουθες καμπύλες ROC για τα Μοντέλα 1, 2 και 3

AUC =Area under the curve

ΜΟΝΤΕΛΟ 1 **AUC=0.6239**

ΜΟΝΤΕΛΟ 2 **AUC=0.6347**

ΜΟΝΤΕΛΟ 3 **AUC=0.5762**