Αρ. Μητρώου: **Ονοματεπώνυμο:**

Εξέταση στο Μεταπτυχιακό Μάθημα: Στατιστική Μοντελοποίηση (11/2/2021)

Επιλέξτε ΔΥΟ από τα 4 Ζητήματα *** Διάρκεια Εξέτασης : 1.30 ώρες *******

ΖΗΤΗΜΑ 1

Ερευνάται η σχέση μεταξύ y (ρυθμός φωτοσύνθεσης) και x_1 (ηλιακή ακτινοβολία) και έστω δείκτρια μεταβλητή x_2 (διαθεσιμότητα του νερού, x_2=0 - αν χαμηλή, x_2=1 - αν υψηλή). Αφού συμπληρώσετε τα κενά στα ακόλουθα αποτελέσματα, εξηγήστε πώς μέσω του μοντέλου $E(y_x) = \beta_0 + \beta x_1 + \beta_2 x_2 + \beta_3 x_3$, μπορούμε να ελέγξουμε αν χρειάζεται να προσαρμοστούν **(I)** δύο διαφορετικές ευθείες, **(II)** δύο παράλληλες ευθείες, ή **(III)** μια κοινή ευθεία και για τις δύο κατηγορίες διαθεσιμότητας του νερού, 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 = 114 + 43.5 x1 - 25.9 x2 - 20.6 x3

Predictor	Coef	SE Coef	T	P
Constant	113.88	29.47	3.86	0.003
x1	43.480	3.213	13.53	<0.001
x2	-25.94	44.62	_____	
x3	-20.616	4.188	-4.92	_____

R-Sq = 96.8% R-Sq(adj) = _____ R-Sq(pred) = 93.93%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	338736	112912	100.76	0.000
Residual Error	10	_____			
Total	13	349942			

Regression Analysis: y versus x1; x2

The regression equation is
y = 214 + 31.3 x1 - 224 x2

Predictor	Coef	SE Coef	T	P
Constant	214.40	37.49	5.72	0.000
x1	31.348	3.636	8.62	0.000
x2	-224.39	33.71	_____	

R-Sq = 89.0% R-Sq(adj) = 87.0% R-Sq(pred) = 81.17%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	311584	155792	44.68	_____
Residual Error	11	38358	3487		
Total	13	349942			

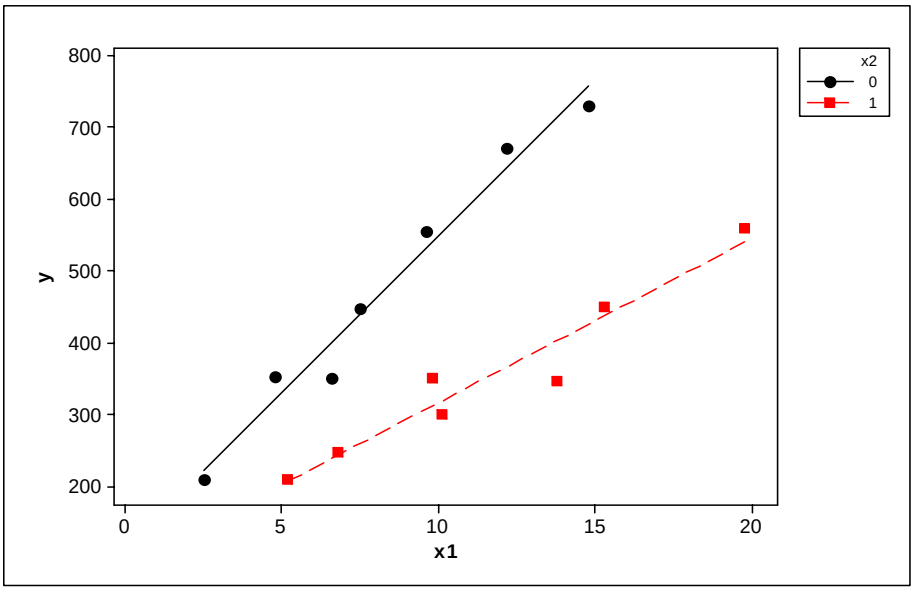
Regression Analysis: y versus x1

The regression equation is
y = 186 + 22.8 x1

PRESS = 255010 R-Sq(pred) = _____

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	157098	157098	9.78	_____
Residual Error	12	192844	16070		
Total	13	349942			



ΖΗΤΗΜΑ 2

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 = - 324751 + 6.4 x1 + 5.64 x2 + 62.3 x3 + 0.520 x4 - 30614 x5

Predictor	Coef	SE Coef	T	P	VIF
Constant	-324751	970973	-0.33	0.740	
X1	6.44	19.34	0.33	0.741	1.7
X2	5.637	1.921	2.93	0.005	2.3
X3	62.33	42.02	1.48	_____	1.6
X4	0.52006	.02291	22.70	<0.001	2.6
X5	-30614	15475	-1.98	_____	1.1

R-Sq = _____ R-Sq(adj) = 96.9% R-Sq(pred) = 95.51%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	3.53751E+14	7.07503E+13	317.16	_____
Residual Error	45	1.00382E+13	2.23072E+11		
Total	50	3.63790E+14			

Β ανάλυση:

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

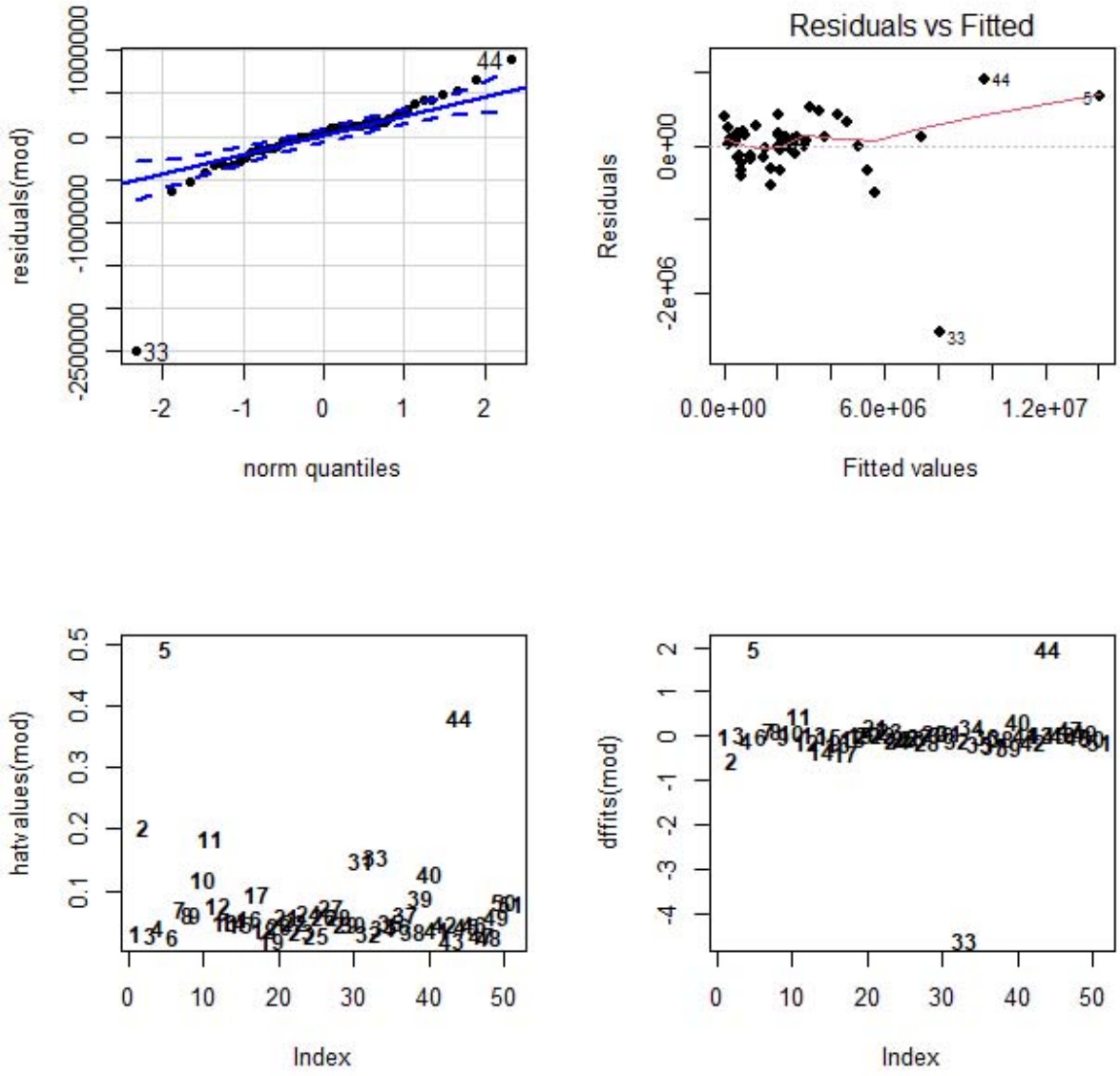
(i) Επιλέξτε δύο εμφωλευμένα μοντέλα που με βάση τα κριτήρια θεωρείτε ότι είναι τα καλύτερα.

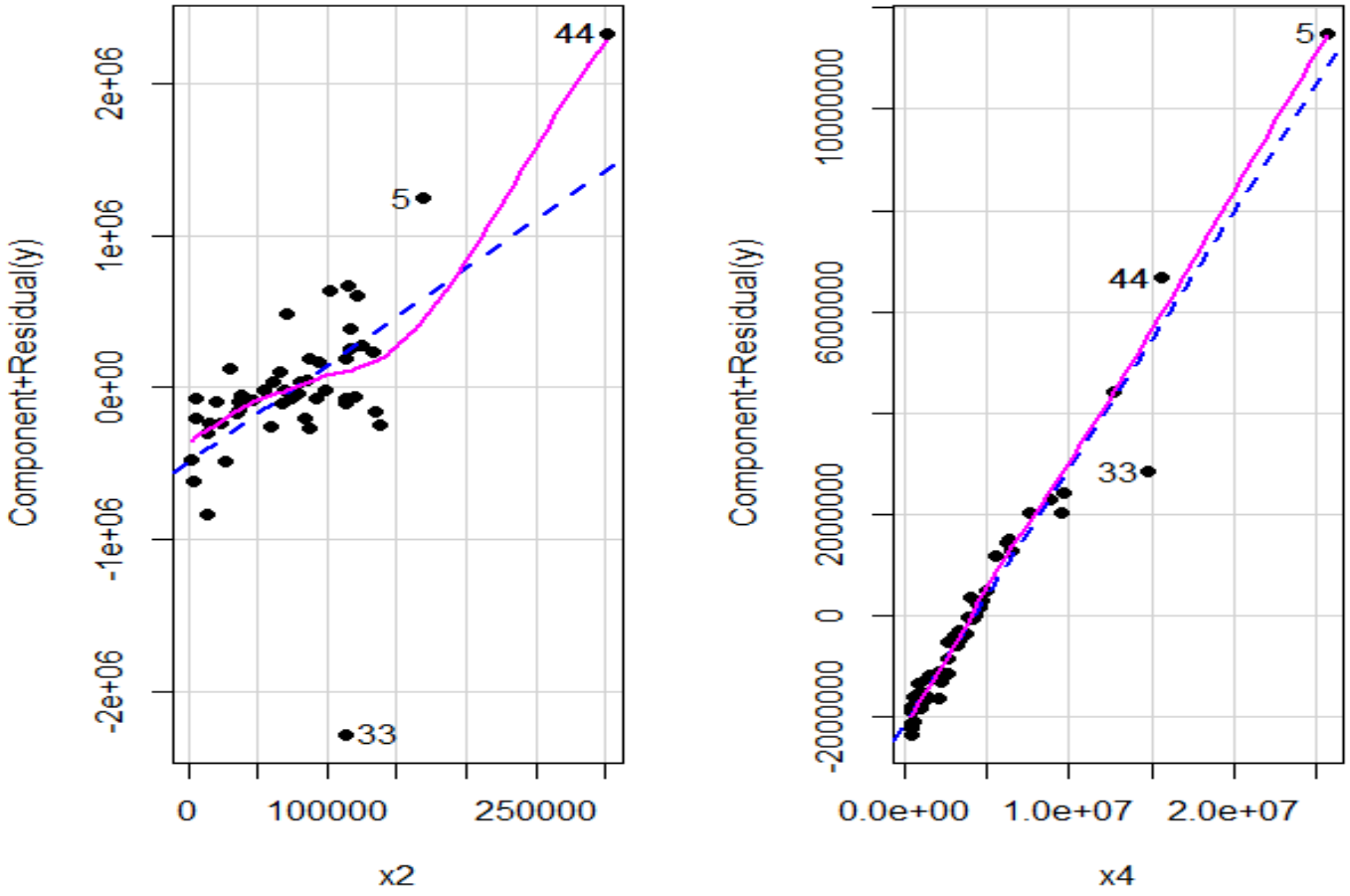
(ii) Στη συνέχεια αξιοποιώντας τον έλεγχο F για τη σύγκριση **δύο** εμφωλευμένων μοντέλων, να βρεθεί το βέλτιστο μοντέλο από τα παραπάνω δύο.

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

Μοντέλο	Μεταβλητές	Υ με	$\bar{R}^2$ (x100%) (διορθ.)	$R^2_{\text{πρόβλεψη}}$ (x100%)	C_p	S	AIC
1	1	X_4	95.8	94.9	19.9	551858	1497.238
2	1	X_2	51.5	47.0	728.3	1878743	1622.195
3	2	$X_2\ X_4$	96.6	95.3	8.6	499000	1487.907
4	2	$X_3\ X_4$	96.4	95.3	11.6	512956	1490.730
5	3	$X_2\ X_4\ X_5$	96.9	95.5	4.3	473969	1483.594
6	3	$X_2\ X_3\ X_4$	96.8	95.5	6.1	482953	1485.509
7	4	$X_2\ X_3\ X_4\ X_5$	97.0	95.6	4.1	467719	1483.143
8	4	$X_1\ X_2\ X_4\ X_5$	96.9	95.4	6.2	478429	1485.452
9	5	$X_1\ X_2\ X_3\ X_4\ X_5$	96.9	95.5	6.0	472305	1485.017

(iii) Σχολιάστε τις παρακάτω γραφικές παραστάσεις των υπολοίπων, των h_{ii} , των DFFITS, καθώς και των μερικών υπολοίπων για τις μεταβλητές X_2 και X_4 του **<u>τελικού μοντέλου</u>**.





(iv) **<u>Αν υποθέσουμε</u>** ότι το Μοντέλο 3 είναι το καλύτερο, να βρεθεί το πάνω άκρο του 0.95 – Δ.Ε. (-920677, _____) της πρόβλεψης ***μιας νέας παρατήρησης*** $Y_{\mathbf{x}_0}$, όταν η σημειακή πρόβλεψη είναι $\hat{Y}_{\mathbf{x}_0} = 110976$ και $\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 = 0.0377$.

<u>ΖΗΤΗΜΑ 3</u>

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

Μέσω μοντέλων παλινδρόμησης Poisson εξετάζεται αν ο αριθμός ειδών φυτών (Υ) ανά νησί, σε n=30 νησιά, σχετίζεται με τις συμμεταβλητές X_1 (εμβαδόν του νησιού) και X_2 (απόσταση από το πιο κοντινό νησί), καθώς και με τη X_3 (εμβαδόν του πιο κοντινού νησιού).

(i) Να συμπληρωθούν οι παρακάτω πίνακες .

(ii) Επιλέξτε το καλύτερο μοντέλο με βάση τους ελέγχους Deviance και Wald, καθώς και με το κριτήριο AIC. Γράψτε το προσαρμοσμένο **τελικό μοντέλο**.

(iii) Βρείτε και ερμηνεύστε την εκτιμημένη ποσότητα $\exp(\hat{\beta}_2)$ του **τελικού μοντέλου** .

(iv) Αξιολογήστε τις πιο κάτω γραφικές παραστάσεις του **τελικού μοντέλου**.

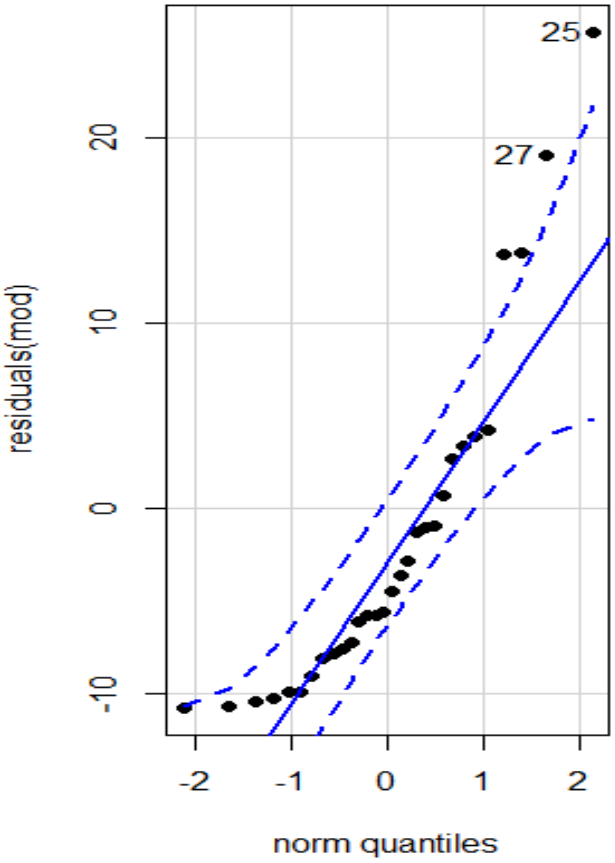
ΜΟΝΤΕΛΟ: 3 Μεταβλητές	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	**p-τιμή**
Σταθερά	4.180e+00	2.918e-02	143.283	<0.001
X_1	4.302e-04	1.199e-05	35.879	<0.001
X_2	5.889e-03	1.434e-03	4.106	<0.001
X_3	-8.064e-05	2.793e-05	-2.887	_____
AIC_3= 2756.1				

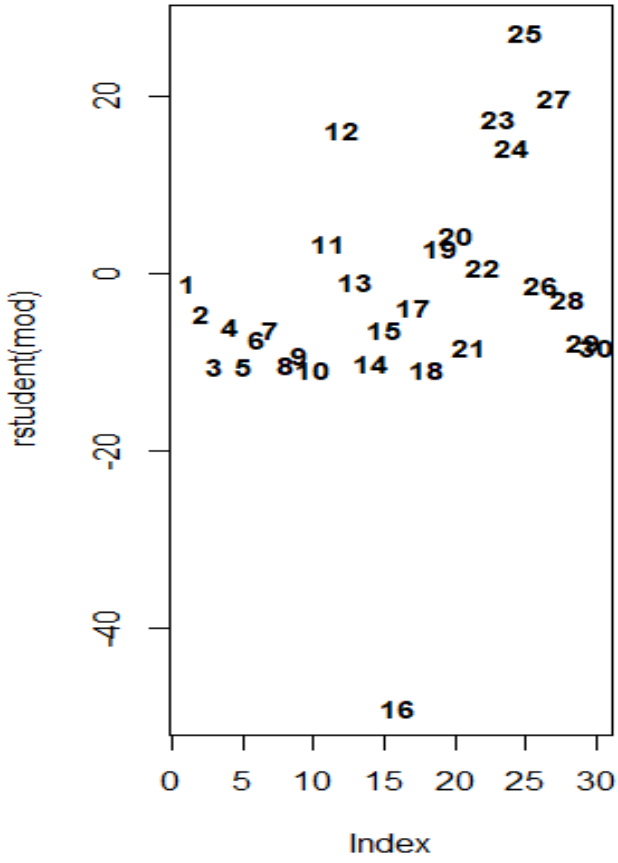
ΜΟΝΤΕΛΟ: 2 Μεταβλητές	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	**p-τιμή**
Σταθερά	4.44600	0.0255200	174.219	<2e-16
X_2	-0.00108	0.0014400	-0.751	_____
X_3	0.00004	0.0000218	1.644	0.100
AIC_2= 3674.1				

ΜΟΝΤΕΛΟ: 1 Μεταβλητές	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	**p-τιμή**
Σταθερά	4.45877	0.024234	183.987	<2e-16
X_2	-0.00135	0.001430	-0.942	_____
$\hat{\ell}_1 =$ _____ και τιμή του κριτηρίου **AIC_1=** _____				
Για το μοντέλο χωρίς καμμία συμμεταβλητή (**Μοντέλο 0**), μόνο με το σταθερό όρο, $\hat{\ell}_0$ =-1835.78 και τιμή του κριτηρίου **AIC_0=** _____				

Μοντέλο	**Deviance β.ε.**	**Deviance**	**Διαφορά στους β.ε.**	**Διαφορά Deviance**	**Pr(>Chi)**	**Deviance Ψευδο-** R_D^2 $R_D^2 = 1 - \frac{D(\hat{\boldsymbol{\beta}})}{D_0}$ $(\times 100\%)$
0	29	3510.73				
1	28	3509.8	1	0.93	_____	0.03 %
2	27	3507.3	1	2.5	_____	_____
3	26	2587.2	1	920.1	<0.001	26.3 %

(iv) Γραφικές παραστάσεις των υπολοίπων deviance και των υπολοίπων πιθανοφάνειας του τελικού μοντέλου





ΖΗΤΗΜΑ 4

(4Α) Έστω Υ 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 .

(4B) Σε μελέτη m πελατών, τμήμα τράπεζας θέλει να εξετάσει αν πελάτης θα αποπληρώσει την πιστωτική του κάρτα Y (ναι=1, όχι=0), σε σχέση με τη X_1 (μηνιαίο υπόλοιπο πιστωτικής κάρτας του), με το ετήσιο εισόδημά του X_2 και με το αν είναι ο πελάτης φοιτητής (X_3=1 αν ναι και X_3=0 αν όχι) .

(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 $e^{\hat{\beta}_3}$, εκφράστε κατά πόσο η ιδιότητα του φοιτητή επιδρά στη σχετική πιθανότητα αποπληρωμής της πιστωτικής κάρτας $\frac{p_x}{1-p_x}$ για το **τελικό μοντέλο**.

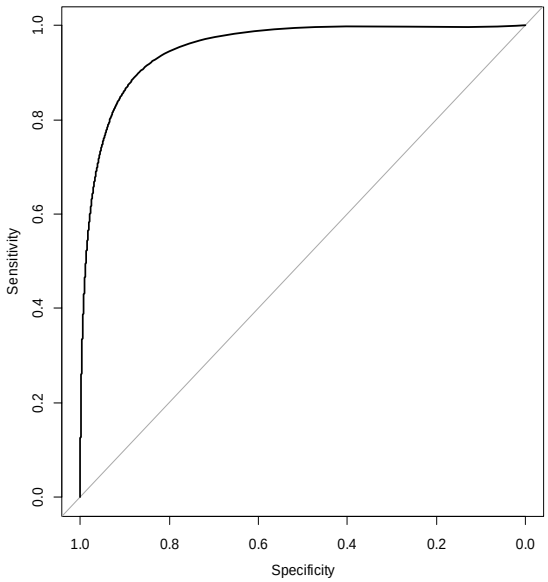
ΜΟΝΤΕΛΟ: 1 **Μεταβλητές**	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	**p-τιμή**	$\exp(\hat{\beta}_j)$
Σταθερά	-11.5800	0.583900	-19.831	<0.001	XXXX
X_1	5.973e-03	2.760e-04	21.641	<0.001	
X_2	1.181e-05	9.397e-06	1.257	_____	
X_3	-4.106e-01	2.758e-01	-1.489	_____	
Ελεγχοσυνάρτηση deviance δίνεται ως **D_1=1171.2** και η τιμή του κριτηρίου **AIC_1=1179.2**					
ΜΟΝΤΕΛΟ: 2 **Μεταβλητές**	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	**p-τιμή**	$\exp(\hat{\beta}_j)$
Σταθερά	-1.111e+01	4.413e-01	-25.183	<0.001	XXXX
X_1	5.982e-03	2.759e-04	21.681	<0.001	
X_3	-6.814e-01	1.700e-01	-4.009	_____	
Ελεγχοσυνάρτηση deviance δίνεται ως **D_2= 1172.8** με αντίστοιχη τιμή $\hat{\ell}_2 =$ _____ και τιμή του κριτηρίου **AIC_2= 1178.8**					
ΜΟΝΤΕΛΟ: 3 **Μεταβλητές**	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	**p-τιμή**	$\exp(\hat{\beta}_j)$
Σταθερά	-3.47429	0.08027	-43.281	<0.001	XXXX
X_3	0.45170	0.12972	_____	_____	
$\hat{\ell}_3 =$ -1127.042 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=0.9538

ΜΟΝΤΕΛΟ 2 AUC=0.9538



ΜΟΝΤΕΛΟ 3 AUC=0.5502