Αρ. Μητρώου: **Ονοματεπώνυμο:**

Εξέταση στο Μεταπτυχιακό Μάθημα: Στατιστική Μοντελοποίηση/ΣΤΑΤΙΣΤΙΚΑ ΠΡΟΤΥΠΑ/GLMS (13/2/2024)

******* Διάρκεια Εξέτασης : 1.30 ώρες *******

ΖΗΤΗΜΑ 1 (Υποχρεωτικό. Επιλέξτε 1 από τα 3 ακόλουθα ερωτήματα) (Βαθμ. 1.5)

(1Α) Έστω γενικό γραμμικό μοντέλο Ε(**y**)=**Xβ**. Δείξτε ότι η ελεγχοσυνάρτηση $F=\frac{SSR/k}{SSE/(n-k-1)}$ 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 $F=\frac{R^2/k}{(1-R^2)/(n-k-1)}$, όπου R^2 ο συντελεστής προσδιορισμού.

(1Β) 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;

(1Γ) Περιγράψτε σύντομα τους δείκτες R^2, $\bar{R}^2$, $R^2_{\text{πρόβλεψη}}$, καθώς και τα κριτήρια Cp-Mallows και AIC. Πώς μπορούν να μας βοηθήσουν στην αξιολόγηση ενός γενικού γραμμικού μοντέλου Ε(**y**)=**Xβ.**

ΖΗΤΗΜΑ 2 (Υποχρεωτικό) (Βαθμ. 4.5)

Εξετάζεται η γραμμική παλινδρόμηση μιας μεταβλητής y, σε σχέση με 4 επεξηγηματικές μεταβλητές x_1, x_2, x_3, x_4. 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

```
The regression equation is
y = 62.4 + 1.55 x1 + 0.510 x2 + 0.102 x3 - 0.144 x4


Predictor      Coef  SE Coef      T      P    VIF
Constant      62.41    70.07   0.89  0.399
x1           1.5511   0.7448   2.08  0.071   38.5
x2           0.5102   0.7238                254.4
x3           0.1019   0.7547   0.14  0.896   46.9
x4          -0.1441   0.7091                282.5

S = 2.44601   R-Sq = 98.2%   R-Sq(adj) = 97.4%

PRESS = 110.347   R-Sq(pred) =

Analysis of Variance

Source           DF       SS      MS       F      P
Regression        4  2667.90  666.97
Residual Error    8    47.86    5.98
Total            12  2715.76
```

Β' ανάλυση:

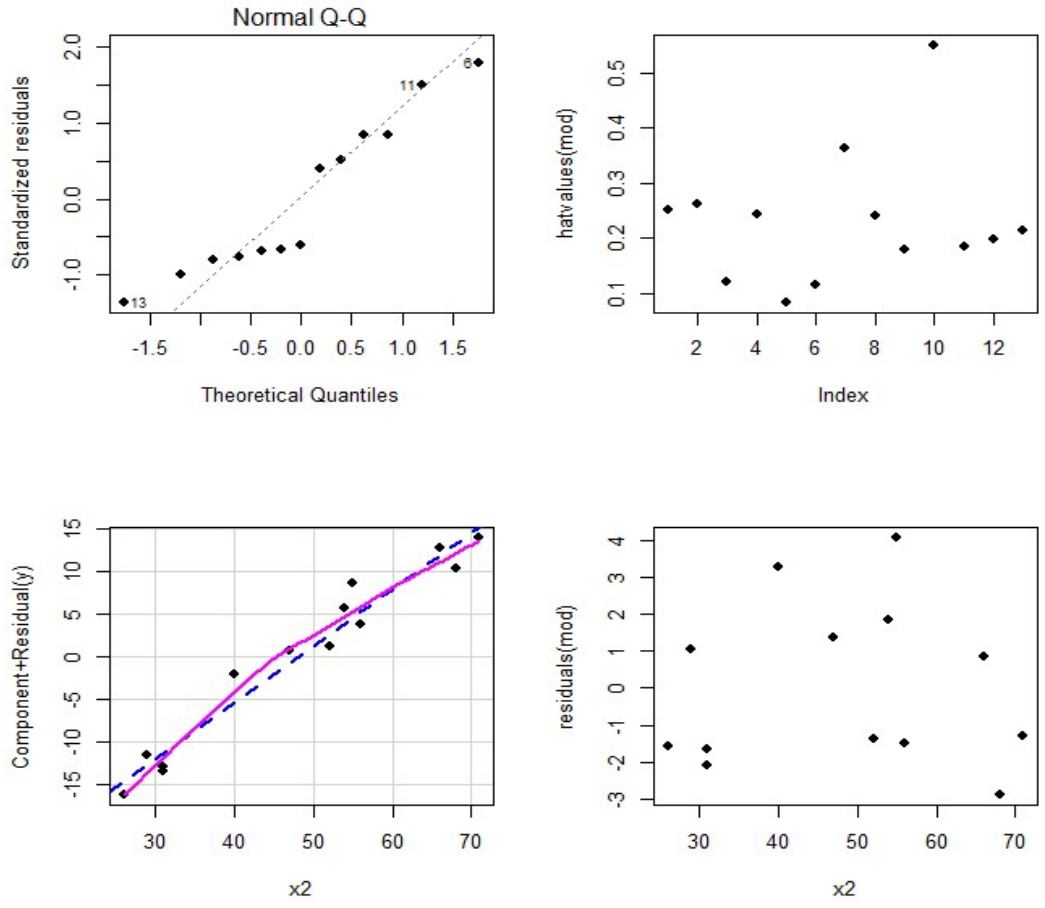
Δίνονται αποτελέσματα προσαρμογών διαφόρων μοντέλων με επιλεγμένες μεταβλητές. Ο παρακάτω πίνακας παρουσιάζει μερικούς δείκτες για την προσαρμογή των μοντέλων αυτών.

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

(ii) Γράψτε την ελεγχοσυνάρτηση **F** για τη σύγκριση **δύο** εμφωλευμένων μοντέλων. Στη συνέχεια αξιοποιώντας **τον έλεγχο F**, καθώς και το **δείκτη** $\bar{R}^2$ να βρεθεί το βέλτιστο μοντέλο από τα παραπάνω δύο. $S=\left(SSE/(n-k-1)\right)^{1/2}$

Μοντέλο	**Μεταβλητές**	**Y με**	R^2 **(x100%)**	$R^2_{\text{πρόβλεψη}}$ **(x100%)**	C_p	S	**AIC**
1	1	x4	67.5	56.0	138.7	8.9639	97.744
2	1	x2	66.6	55.7	142.5	9.0771	98.070
3	1	x1	53.4	37.4	202.5	10.7270	102.412
4	2	x1 x2	97.9	96.5	2.7	2.4063	64.312
5	2	x1 x4	97.2	95.5	5.5	2.7343	67.634
6	2	x3 x4	93.5	89.2	22.4	4.1921	78.745
7	3	x1 x2 x4	98.2	96.9	3.0	2.3087	63.866
8	3	x1 x2 x3	98.2	96.7	3.0	2.3121	63.904
9	3	x1 x3 x4	98.1	96.5	3.5	2.3766	64.620
10	4	x1 x2 x3 x4	98.2	95.9	5.0	2.4460	65.837

(iii) Σχολιάστε σύντομα τις παρακάτω γραφικές παραστάσεις των τυποποιημένων υπολοίπων, των h_{ii}, των μερικών υπολοίπων για τη μεταβλητή X_2 και τα υπόλοιπα σε σχέση με τη X_2 του **τελικού μοντέλου**.



(iv) Ελέγξτε αν χρειάζεται να εισαχθεί η μεταβλητή $W=X_2^2$ στο **τελικό μοντέλο**: R^2 = 98.5%, $\bar{R}^2$ = _____

Δίνονται: SSE=40.19 , $R^2_{\text{πρόβλεψη}}$ = 97.47%, AIC=61.57, $\hat{\beta}_W$=-0.006624 , $\sqrt{C_{ww}}$ = 0.0015738

Επιλέξτε ΕΝΑ από τα επόμενα 3 Ζητήματα (Βαθμ. 4.0)

ΖΗΤΗΜΑ 3

Εξετάζεται ο βαθμός επίδοσης (Y), n=34 υπαλλήλων εταιρείας, ένα μήνα μετά την πρόσληψή τους, σε σχέση με ένα αρχικό τεστ ικανότητας (X_1). Ορίζεται η δείκτρια μεταβλητή X_2 =0, αν τριτοβάθμια εκπαίδευση και X_2=1, αν

δευτεροβάθμια. Εξηγήστε σύντομα 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 για τα δύο επίπεδα εκπαίδευσης, όπου $x_3 = x_1 x_2$, η μεταβλητή που εκφράζει την αλληλεπίδραση μεταξύ των μεταβλητών x_1 και x_2.

Συμπληρώσετε τα κενά στα ακόλουθα αποτελέσματα και κατασκευάστε ένα 95% διάστημα εμπιστοσύνης για το συντελεστή β_1 της x_1 του **τελικού μοντέλου.** 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 = 0.917 + 2.95 x1 + 1.11 x2 - 1.36 x3

Predictor      Coef  SE Coef      T      P
Constant     0.9174   0.6442   1.42  0.165
x1           2.9452   0.4008   7.35  <0.001
x2                     1.054
x3          -1.3625   0.6373

R-Sq =        R-Sq(adj) = 66.8% R-Sq(pred) = 60.67%

Analysis of Variance

Source           DF      SS      MS      F      P
Regression
Residual Error   30  27.279   0.909
Total            33  90.400
```

Regression Analysis: y versus x1, x2

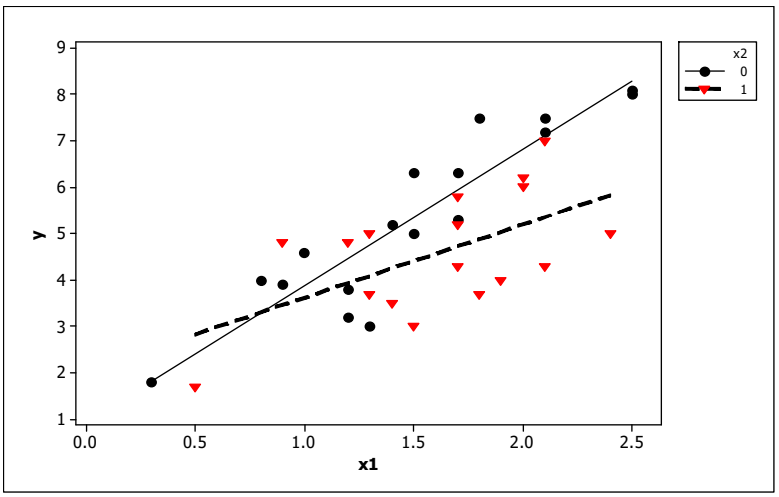
```
The regression equation is
y = 1.73 + 2.41 x1 - 1.03 x2

Predictor      Coef  SE Coef      T      P
Constant     1.7260   0.5507   3.13   0.004
x1           2.4062   0.3291   7.31  <0.001
x2                    0.347

S = 1.00701   R-Sq = 65.2%   R-Sq(adj) =

PRESS = 37.3605   R-Sq(pred) =

Analysis of Variance

Source           DF      SS      MS      F      P
Regression        2  58.964  29.482  29.07  <0.001
Residual Error   31  31.436   1.014
Total            33  90.400
```

Regression Analysis: y versus x1

```
The regression equation is
y = 1.38 + 2.30 x1

Predictor      Coef  SE Coef      T      P
Constant     1.3802   0.6001   2.30  0.028
x1           2.2976
```

$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = 9.48235$

```
S = 1.12283
R-Sq = 55.4%   R-Sq(adj) = 54.0%  R-Sq(pred) = 49.88%

Analysis of Variance

Source           DF      SS      MS      F      P
Regression        1  50.056  50.056
Residual Error
```



ΖΗΤΗΜΑ 4

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 $D_M(\hat{\boldsymbol{\beta}})=-2\left(\hat{\ell}_M-\hat{\ell}_{\text{κορ}}\right)$, όπου $\hat{\ell}_M$ η μεγιστοποιημένη λογαριθμοποιημένη συνάρτηση πιθανοφάνειας του μοντέλου M που μας ενδιαφέρει και $\hat{\ell}_{\text{κορ}}$ 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 τα κενά στους παρακάτω πίνακες. (Τα $\exp(\hat{\beta}_j)$ υπολογίζονται μόνο για το **τελικό μοντέλο**)

(ii) Με βάση τον έλεγχο Wald, τη διαφορά των ελεγχοσυναρτήσεων Deviance, και λαμβάνοντας υπόψη το δείκτη Ψευδο-R^2_D Deviance (βλ. πινακάκι πιο κάτω), καθώς και το κριτήριο AIC, επιλέξτε το καλύτερο από τα τρία μοντέλα **M0, M1, M2**. Γράψτε το προσαρμοσμένο **τελικό μοντέλο**.

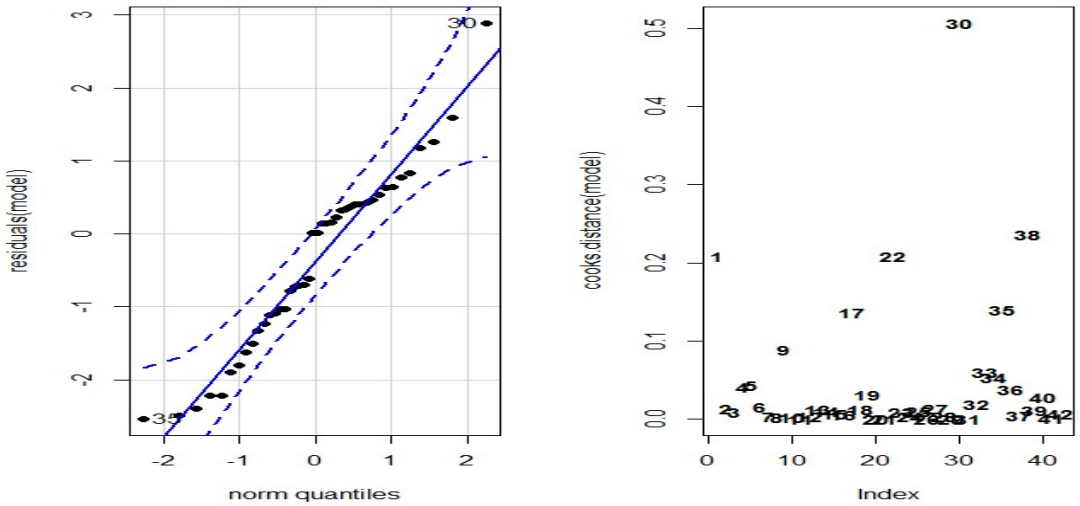
(iii) Κατασκευάστε 0.95-διαστήματα εμπιστοσύνης για τα $\exp(\beta_j)$ και με βάση αυτών ερμηνεύστε τις εκτιμημένες ποσότητες $\exp(\hat{\beta}_j)$ του **τελικού μοντέλου**.

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

ΜΟΝΤΕΛΟ: M2 **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	_____	
X_2	-0.013193	0.0629	_____	_____	
AIC_2=245.78					
ΜΟΝΤΕΛΟ: M1 **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	_____	_____	
$\hat{\ell}_1$ = _____ και η τιμή του κριτηρίου **AIC_1=** _____					
ΜΟΝΤΕΛΟ: M0 Για το μοντέλο χωρίς συμμεταβλητές (Null model) $\hat{\ell}_0$ =-222.219 και η τιμή του κριτηρίου **AIC_0=** _____					

Μοντέλο	Deviance β.ε.	Deviance	Διαφορά στους β.ε.	Διαφορά Deviance	Pr(>Chi)	Deviance Ψευδο- R^2_D $R^2_D=1-\frac{D(\hat{\boldsymbol{\beta}})}{D_0}\ (\times 100\%)$
M0	**41**	**272.305**				
M1	**40**	**67.695**				
M2	**39**	**67.652**	**1**	**0.043**		**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 κατανομής Bernoulli $f(y)=p^y(1-p)^{1-y}$, $y=0, 1$ με παράμετρο p.
Γράψτε το μοντέλο της λογιστικής παλινδρόμησης για 2 συμμεταβλητές .

(5Β) Σε μελέτη n ασθενών, ερευνητής θέλει να εξετάσει αν Y (αγγειοσυστολή ναι=1, όχι=0), σχετίζεται με τον εισπνεόμενο όγκο αέρα X_1, και με τον παρατηρούμενο ρυθμό εισπνοής X_2. Με βάση τη λογιστική παλινδρόμηση, εξετάζεται η επίδραση των συμμεταβλητών αυτών στη σχετική πιθανότητα επιτυχίας (odds) $\frac{p_x}{1-p_x}$.

(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_1} του **τελικού μοντέλου**.
(iii) Υπολογίστε τις εκτιμημένες ποσότητες $\exp(\hat{\beta}_j)$ του **τελικού μοντέλου.**

Με τη βοήθεια της ποσότητας $e^{\hat{\beta}_1}$ (odds ratio), εκφράστε κατά πόσο αύξηση στον εισπνεόμενο όγκο αέρα επιδρά στη σχετική πιθανότητα εμφάνισης αγγειοσυστολής $\frac{p_x}{1-p_x}$ για το **τελικό μοντέλο**.

ΜΟΝΤΕΛΟ: M0	Για το μοντέλο χωρίς συμμεταβλητές (Null model)

Ελεγχοσυνάρτηση deviance δίνεται ως **D_0=54.04**
με αντίστοιχη τιμή $\hat{\ell}_0$ = -27.01992 και τιμή του κριτηρίου **AIC_0=** _____

ΜΟΝΤΕΛΟ: M1 **Μεταβλητές**	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	**p-τιμή**	$\exp(\hat{\beta}_j)$
Σταθερά				<0.001	XXXX
X_1	1.33568	0.61621	_____	_____	_____

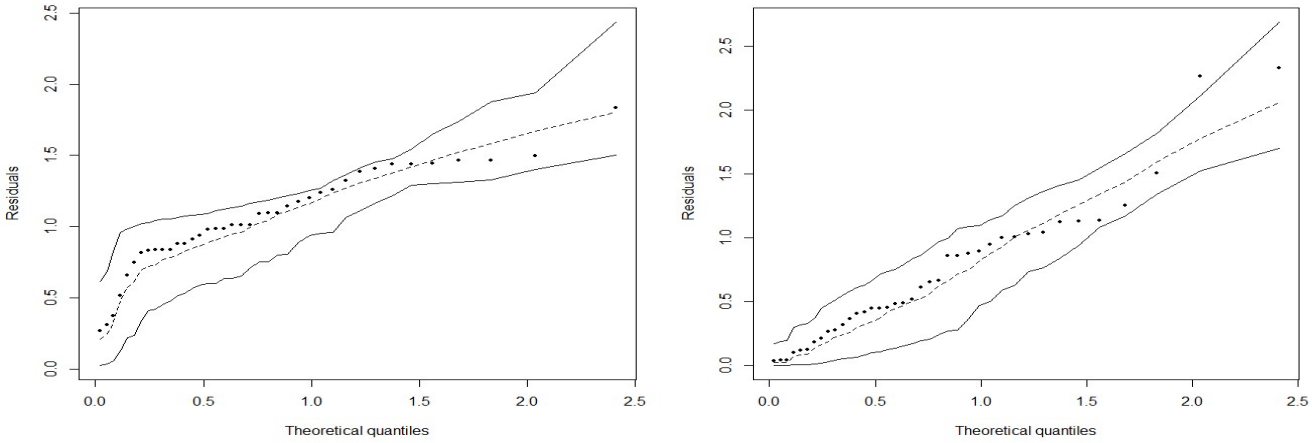
Ελεγχοσυνάρτηση deviance δίνεται ως **D_1= 46.989**
με αντίστοιχη τιμή $\hat{\ell}_1 = -23.49469$ και η τιμή του κριτηρίου **AIC_1=** _____
McFadden ψευδο-R^2 =0.1305

ΜΟΝΤΕΛΟ: M2 **Μεταβλητές**	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	z_j	**p-τιμή**	$\exp(\hat{\beta}_j)$
Σταθερά	-9.5296	3.2332	-2.947	0.00320	XXXX
X_1	3.8822	1.4286	_____	_____	_____
X_2	2.6491	0.9142	_____	_____	_____

$\hat{\ell}_2$ = _____ και η τιμή του κριτηρίου **AIC_2= 35.772**

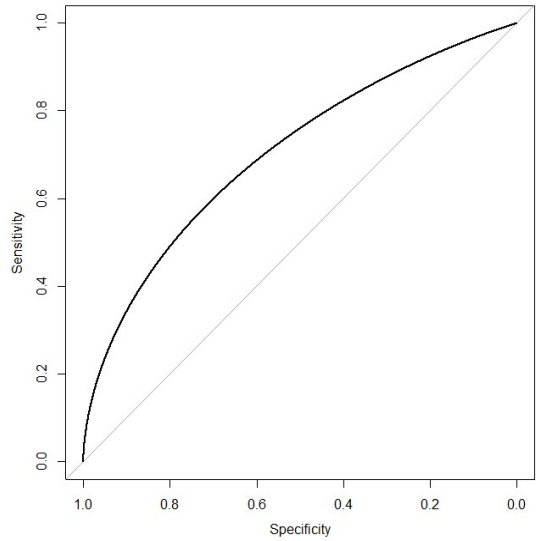
McFadden ψευδο-R^2 =0.4491

(iv) Ενισχύστε τα συμπεράσματά σας με τις ακόλουθες γραφικές παραστάσεις των υπολοίπων deviance και τις καμπύλες ROC για τα Μοντέλα 1 και 2 αντίστοιχα



AUC =Area under the curve

ΜΟΝΤΕΛΟ 1 **AUC=**0.7037



ΜΟΝΤΕΛΟ 2 **AUC=**0.9011