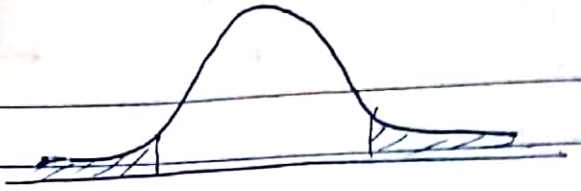
Θέμα 2021



Ζήτημα 1 y = ρυθμός φωτοσύνθεσης, x_1: πλευρική ακτινοβολία, x_2: διδιάστατα νερού (δείκτρια)
$x_2 = 0$: χαμηλή
$x_2 = 1$: υψηλή

$$\hat{y} = 114 + 43.5\,x_1 - 25.9\,x_2 - 20.6\,x_3$$
($\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$)

$$t_i^* = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)} \sim t_{n-p} \qquad t_{n-p} = t_{14-4} = t_{10}$$

SST έχει 13 df το οποίο είναι $n-1$ με βάση τη θεωρία, άρα $n-1=13 \Leftrightarrow n=14$ παρατηρήσεις.
$k=3$ μεταβλητές, $p = k+1 = 4$ παράμετροι

$$\bar{R}^2 = 1 - \frac{n-1}{n-p}\left(1-R^2\right) \simeq 95.84\%$$

	T	P
x_1		
x_2	−0.58	0.574
x_3	−4.92	0.0006 < 0.001

$SSE = SST - SSR = 11,206$

Μοντέλο $y \sim x_1, x_2$ άρα $k=2$, $p=k+1=3$, $df = n-p = 11$ → για t-test
t-test με $df = 11$

	T	P
x_1	8.62	0.000
x_2	−6.66	$3.58 \cdot 10^{-5} < 0.001$

Άλλον για το μοντέλο $y \sim x_1, x_2$
H_0: $\beta_0 = \beta_1 = \beta_2 = 0$ $\quad H_1$: έστω ένας συντελεστής $\neq 0$.
$df_1 = 2$ → αριθμός μεταβλητών, $df_2 = 11$.

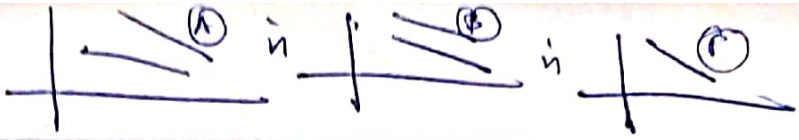
$$F^* = \frac{MSR}{MSE} = 44.68 \qquad p\text{-value} = 5.238 \cdot 10^{-6} < 0.001$$
(Ftest df_1, df_2) $= F_{2,11}$.

Μοντέλο $y \sim x_1$: $\hat{y} = 186 + 22.8\,x_1$ ($\hat{\beta}_1$)

$$R^2_{pred} = 1 - \frac{PRESS}{SST} = 1 - \frac{255010}{349942} \simeq 0.27$$

Έχουμε 1 μεταβλητή: $df_1 = 1$, $df_2 = n-p = 14-(1+1) = 12$
$F^* = 9.78 \rightarrow p\text{-value} = 0.008$

Τέλος εφωτήσεων συμπλήρωσης.

Εξετάζουμε αν  ή ή

• Αν $\beta_3 \neq 0$ τότε έχουμε (A).

Από την ανάλυση του μοντέλου $y \sim x1, x2, x3$ έχουμε για τη $x3$ ότι το p-value < 0.001 άρα θεωρητικά σημαντικός ο έλεγχος και $\beta_3 \neq 0$.

Άρα θα προσαρμώσουμε 2 διαφορετικές ευθείες. $F_{1,10}$

Εναλλακτικός τρόπος:

$H_0: \beta_3 = 0 \;:\; M_0: y \sim x1, x2$ $\qquad H_1: \beta_3 \neq 0 \;:\; M_1: y \sim x1, x2, x3$

$$F^* = \frac{(SSE_{M_0} - SSE_{M_1})/1}{SSE_{M_1}/(14-3-1)} = \frac{38358 - 11,206}{11,206/10} = 24.22988$$

$F\_test_{1,10} = 6.02 \cdot 10^{-4} < 0.001$ άρα θεωρητικά σημαντικός ο έλεγχος επ. απορρίπτω H_0 και δεχομαι $H_1: \beta_3 \neq 0$.

Άρα πρέπει να προσαρμοστούν δύο ευθείες:

$E(y_x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

Προσαρμοσμένες ευθείες:

(1) $E(y_x) = \beta_0 + \beta_1 \cdot x_1$, $x_2 = 0$

(2) $E(y_x) = \beta_0 + \beta_1 x_1 + \beta_2 + \beta_3 x_1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1$, $x_2 = 1$

Ερμηνεία:

Αν η διαθεσιμότητα νερού είναι χαμηλή, τότε ο ρυθμός με τον οποίο αυξάνει ο ρυθμός φωτοσύνθεσης ως προς την πλιακή ακτινοβολία είναι μεγαλύτερος σε σχέση με το αν η διαθεσιμότητα νερού ήταν υψηλή. ✓

Ζήτημα 2 $y \sim x_1, x_2, x_3, x_4, x_5$ $\quad E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$

Α ανάλυση:

Πλήρες μοντέλο Συμπλήρωση κενών.

$n = 51$ (df of SST $+1 = 50+1 = 51$) $\rightarrow$ παρατηρήσεις

$k = 5$ μεταβλητές, $p = k+1 = 6$ παράμετροι

$df = n - p = 45$

x_3: $t^* = 1.48$, p-value $= 0.145$

x_5: $t^* = -1.98$, p-value $= 0.054$

$$R^2 = \frac{SSR}{SST} = 0.9724 \simeq 97.24\%$$

$$F^* = 317.16 \sim F_{\underset{\text{μεταβλητές}}{5}, \underset{n-p=45}{df}} = F_{5,45}(F^*), \quad p\text{-value} = 6.928 \cdot 10^{-34}$$

Επιτυγχάνεται μεγάλο R^2 που σημαίνει ότι το μοντέλο επεξηγεί ικανοποιητικά τα δεδομένα μας. Ωστόσο, από τα t-tests και τα αντίστοιχα p-values φαίνεται πως κάποιες μεταβλητές ίσως να μην είναι σημαντικές για το μοντέλο και άρα θα πρέπει να εξεταστεί η περίπτωση μοντέλων με λιγότερες μεταβλητές.

Επίσης, θα πρέπει να ελεγχθεί πολυσυγγραμμικότητα των μεταβλητών x_1, x_2, x_3, x_4, x_5 με βάση το κριτήριο VIF (αρά VIF values < 5)

Β ανάλυση

(i) Επιλέγουμε 5 και 7 μοντέλα λόγω των αρών τους.

(ii) Σύγκριση M_5 και M_7 $\rightarrow k=3, p=4$ $\qquad \rightarrow k=4, p=5$

$H_0: M_5: x_2, x_4, x_5$ (Reduced) $\qquad H_1: M_7: x_2, x_3, x_4, x_5$ (Full model)

$df_R - df_F = 1$, $\quad df_F = n - p = 51 - 5 = 46$

$$F^* = \frac{(SSE_R - SSE_F)/1}{SSE_F / df_F} \sim F_{(df_R - df_F), df_F} = F_{1,46}$$

$$S = \frac{\sqrt{SSE}}{\sqrt{n-k-1}} \Rightarrow SSE = (n-k-1)S^2, \quad SSE_R = (51-3-1)\cdot 473969^2$$

$SSE_F = (51-4-1)\cdot 467719^2$, $F^* \simeq 2.26$ $\rightarrow$ p-value $\simeq 0.139$

άρα έλεγχος δεν σημαντικός που σημαίνει ότι δεν μπορώ να απορρίψω την H_0 οπότε κρατάω το $M_5: y \sim x_2, x_4, x_5$

(iii) residuals vs norm quantiles

Με βάση το QQ plot, τα υπόλοιπα φαίνεται να είναι κοντά σε ευθεία ώστε να ακολουθούν ικανοποιητικά κανονική κατανομή με εξαίρεση το σημείο 33 που μοιάζει για outlier.

residuals vs fitted

Φαίνεται να μην υπάρχει ένα trend στα residuals, για τιμές μεγάλες των fitted values. Επίσης για μικρές τιμές φαίνεται να μη είναι εντός band όπως θα θέλαμε.

hat values vs index

$\frac{2p}{n} = \frac{2 \cdot 3}{51} = 0.1176$

Για λίγο $> \frac{2p}{n}$ σημεία επιρροής → έχω κάποια.

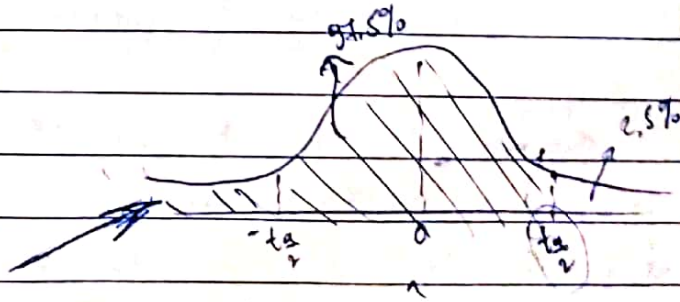
dffits vs index.

$2\sqrt{\frac{p}{n}} = 2\sqrt{\frac{3}{51}} \simeq 0.485$. και $|dffits_i| > 2\sqrt{\frac{p}{n}}$ επ. : παρατήρηση ίσως σημείο επιρροής

παρατηρήσεις 5 και 44 σημεία επιρροής.

partial residuals plot (vs x2 and x4)

Βλέπουμε ότι και στα 2 μεταβλητές x2 και x4 τα σημεία σχηματίζουν καλά μια ευθεία. Ειδικά για την x4, ωραίο και για x2 υπάρχει trend. Καταλήξουμε ότι χρειαζόμαστε και τις δύο μεταβλητές στο μοντέλο.



(iv) Μοντέλο 3: $y \sim x_2, x_4$

Δ. 95%: $(-920.677, \dots)$ των πρόβλεψης μιας νέας παρατήρησης Y_{x_0}, όταν $\hat{Y}_{x_0} = 110.976$ και $x_0'(X'X)^{-1}x_0 = 0.0377$

Από σημειώσεις βάσεων General Linear Model: $\hat{Y}_{x_0} - Y_{x_0} \sim N(0, \sigma^2(x_0'(X'X)^{-1}x_0 + 1))$ και

$Y_{x_0} \in \hat{Y}_{x_0} \pm t_{\frac{\alpha}{2}} \cdot S \cdot (x_0'(X'X)^{-1}x_0 + 1)^{1/2}$, $S = 499.000$ από πίνακα.

άρα $-920.677 = 110.976 - t_{\frac{\alpha}{2}} \cdot 499.000 \cdot \sqrt{1.0377}$ $\Rightarrow$ $t_{\frac{\alpha}{2}} \simeq 2.02$ Το ίδιο βρίσκω για $qt(0.975, 48) \simeq 2.01$ (48 = n − p) και μετά υπολογίζω.

Ζήτημα 3 Poisson: $f(y) = \frac{e^{-\mu x} \cdot \mu x^y}{y!}$, $y = 0, 1, 2, \dots$; canonical link: $g(\mu x) = \ln \mu x = \beta' \cdot x$

επαγγελματική deviance: $D_M(\hat{\beta}) = -2(\hat{\ell}_M - \hat{\ell}_{sat}) = 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right)\right]$, $AIC = -2\hat{\ell}_M + 2(d)$ → αν πολλές αριθμοί παραμέτρων

$Y \to$ πλήθη ατυχημάτων ανά νησί σε $n = 30$ νησιά , X_1 έκθεση νησιού , X_2 απόσταση από κεντρικότερο νησί , X_3 έτερό του κοντινότερου νησιού

(i) Συμπλήρωση:

ΜΟΝΤΕΛΟ 3 x_3: $\frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \to$ z-test p-value $= 2 \cdot$ pnorm(abs(−2.887), mean=0, sd=1, lower.tail=FALSE) $= 0.0039$

ΜΟΝΤΕΛΟ 2: x_2: z-test οριακής p-value = 0.4526

ΜΟΝΤΕΛΟ 1: x_2: p-value = 0.3462.

ΜΟΝΤΕΛΟ 1: $\hat{\ell}_1 = ;$, $AIC = ;$

για M_0: $D_0(\hat{\beta}) = -2(\hat{\ell}_0 - \hat{\ell}_{sat}) \to D_0(\hat{\beta}) = -2\hat{\ell}_0 + 2\hat{\ell}_{sat} \Rightarrow \hat{\ell}_{sat} = \frac{D_0(\hat{\beta})}{2} + \hat{\ell}_0$

$\to \hat{\ell}_{sat} = -80.415$

για M_1: $D_1(\hat{\beta}) = -2(\hat{\ell}_1 - \hat{\ell}_{sat}) \Rightarrow \hat{\ell}_1 = \hat{\ell}_{sat} - \frac{D_1(\hat{\beta})}{2} = -1835.315$

$AIC_1 = -2\hat{\ell}_1 + 2 \cdot d \to^{d=2} AIC_1 = -2 \cdot (-1835.315) + 2 \cdot 2 = 3674.63$

για M_0: $AIC_0 = -2\hat{\ell}_0 + 2 \cdot d \to^{d=1} -2 \cdot (-1835.78) + 2 = 3673.56$

$D(\hat{\beta}) \sim \chi^2_{n-p}$

M_1: $H_0: M_0$, $H_1: M_1$ (ΣΕ ΣΧΕΣΗ ΜΕ M_0) $D_1(\hat{\beta}) - D_0(\hat{\beta}) \sim \chi^2_{q=1} \to$ pchisq(0.93, df=1, lower.tail=FALSE) = (p-value) = 0.3349

Διαφορά σε αριθμό παραμέτρων για τα 2 μοντέλα

Άρα μη στατ. σημαντικός έλεγχος → διαλέγω το M_0 (μικρότερο μοντέλο)

M_2 (ΣΕ ΣΧΕΣΗ ΜΕ M_1): $H_0: M_1$, $H_1: M_2$ $(M_1 \subseteq M_2)$

$D_2(\hat{\beta}) - D_1(\hat{\beta}) \sim \chi^2_{q=1} = \chi^2_{1\text{ διαφορά σε παραμ.}} \to$ p-value = 0.1138

άρα πάλι μη στατ. σημαντικός έλεγχος ⇒ δε μπορώ να απορρίψω H_0 άρα επιλέγω το M_1.

$$R^2_D = \left(1 - \frac{D_2(\hat{\beta})}{D_0}\right) 100\% = \left(1 - \frac{3507.3}{3510.73}\right) \cdot 100\% = 0.098\%$$

(ii) Βρίσκατε M_0 και M_3 πιθανά καλύτερα μοντέλα με βάση Deviance.

$H_0: M_0 \to$ 1 παράμετρος (σταθ) , $H_1: M_3 \to$ 4 παράμετροι

$D_3(\hat{\beta}) - D_0(\hat{\beta}) \sim \chi^2_{q=3} \to$ p-value $\simeq 6.96 \cdot 10^{-200} << 0.001 \to$ άρα στατ. σημαντικός έλεγχος και απορρίπτω $H_0 \Rightarrow$ επιλέγω M_3.

Επίσης $AIC_3 < AIC_0 \to M_3$ καλύτερο

Wald: για τα z-tests p-values του M_3 όλες είναι σημαντικές 5

(iii) $e^{\hat{\beta}_2}$ του M_3: $e^{\hat{\beta}_2} = e^{5.889e-03} = 1.0059$ άρα για μια μονάδα αύξηση X_2, το γ αναμένεται να αυξηθεί κατά 0.59%.

Ζήτημα 4

(A) $Y \sim$ διωνυμική, $f(y) = \binom{n}{y} p^y (1-p)^{n-y}$, $y = 0, 1, 2, \dots, n$ παράμετροι: p, n

$y_x \sim b(n_x, p_x)$ ($n_x > 1$ διωνυμική δεδομένα)

$\eta(x) = g(\mu_x) = \ln \frac{\mu_x}{n_x - \mu_x} = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k = x'\beta$ (1)

↓ link function.

$\mu_x = E(y) = n_x p_x$, $V(y) = n_x p_x (1-p_x)$

Λύνοντας την (1) ως προς p_x: $p_{x_i} = \frac{e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}}} = \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}}$

άρα $E(y_i) = n_i p_i = n_i \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}}$: μοντέλο

(B) m : πελάτες τράπεζας, αποπληρωμή πιστωτικής κάρτας $Y = \begin{cases} 1, \text{ναι} \\ 0, \text{όχι} \end{cases}$,

X_1: μηνιαίο υπόλοιπο πιστωτικής κάρτας, X_2: ετήσιο εισόδημα, $X_3 = \begin{cases} 1, \text{φοιτητής} \\ 0, \text{όχι φοιτητής} \end{cases}$

(i) $Z = \frac{\hat{\beta}_j - \beta_j(0)}{se(\hat{\beta}_j)} \sim N(0,1)$ ασυμπτωτικά

$H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$

ΜΟΝΤΕΛΟ 1

άρα $z_2 = \frac{1.181 \cdot 10^{-5}}{9.397 \cdot 10^{-6}}$ → p-value = $2 \cdot$ pnorm($|z_2|$, mean=0, sd=1, lower.tail=FALSE) = 0.209

άρα έλεγχος μη στατ. σημαντικός → δεν μπορούμε να απορρίψουμε H_0. Αποδεχόμαστε $\beta_2 = 0$.

$z_3 \rightarrow$ p-value = 0.136

ΜΟΝΤΕΛΟ 2:

$z_3 = -4.009 \rightarrow$ p-value = $6.098 \cdot 10^{-5}$

ΜΟΝΤΕΛΟ 3:

$z_3 = \frac{\hat{\beta}_3}{se(\hat{\beta}_3)} = 3.48^2 \rightarrow$ p-value = 0.000497

ΜΟΝΤΕΛΟ 1: $AIC_1 = 1179.2$

$AIC = -2\hat{l}_M + 2d$ (d: αριθμός παραμέτρων)

άρα $\hat{l}_1 = d - \frac{AIC_1}{2} = 4 - \frac{1179.2}{2} = -586.6$.

ΜΟΝΤΕΛΟ 2: $\hat{l}_2 = d - \frac{AIC_2}{2} = 3 - \frac{1178.8}{2} = -586.4$

ΜΟΝΤΕΛΟ 3: $AIC_3 = -2\hat{l}_3 + 2d = -2 \cdot (-1127.042) + 2 \cdot 2 = 2258.084$

Τα Μοντέλα 1 και 2 έχουν χαμηλότερο AIC σε σχέση με Μοντέλο 3 (συμπληρωτικοί).
Άρα θα τα συγκρίνω M_1 vs M_2.

H_0: $M_2 \sim x_1, x_3$ c H_1: $M_1 \sim x_1, x_2, x_3$

$D_2 - D_1 \sim \chi^2_{\text{διαφορά παραμέτρων}} = \chi^2_1$

$D_2 - D_1 = 1172.8 - 1171.2 = 1.6 \sim \chi^2_1$

p-value = 0.2059 άρα μη σημ. συμπληρωτικός έλεγχος, οπότε δεν μπορώ να απορρίψω την H_0 ⇒ επιλέγω $M_2 \sim x_1, x_3$ → το καλύτερο μοντέλο

(ii) 95% ΔΕ για e^{β_3} του τελικού μοντέλου M_2

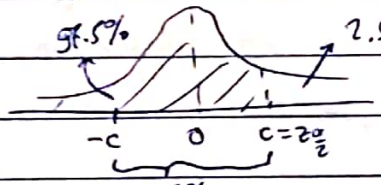
$Z = \frac{\hat{\beta}_j - \beta_j(0)}{se(\hat{\beta}_j)} \rightarrow Z_3 = \frac{\hat{\beta}_3 - \beta_3}{se(\hat{\beta}_3)} \sim N(0,1)$ ασυμπτωτικά

$P[-c \le Z_3 \le c] = 0.95 \rightarrow c = Z^{-1}(0.975) = z_{\frac{\alpha}{2}}$ του ΔΕ

$-c \le \frac{\hat{\beta}_3 - \beta_3}{se(\hat{\beta}_3)} \le c \Rightarrow -c\,se(\hat{\beta}_3) \le \hat{\beta}_3 - \beta_3 \le c\,se(\hat{\beta}_3) \rightarrow \hat{\beta}_3 + c\,se(\hat{\beta}_3) \ge \beta_3 \ge \hat{\beta}_3 - c\,se(\hat{\beta}_3)$

$\Rightarrow \beta_3 \in \hat{\beta}_3 \pm c \cdot se(\hat{\beta}_3) \rightarrow e^{\beta_3} \in e^{\hat{\beta}_3 \pm c \cdot se(\hat{\beta}_3)}$.

$c \simeq 1.96$ από qnorm(0.975)



$\hat{\beta}_3 = -0.6814$, $se(\hat{\beta}_3) = 0.17$.

άρα $e^{\hat{\beta}_3 \pm c \cdot se(\hat{\beta}_3)}$ → $e^{-0.6814 - 1.96 \cdot 0.17} \simeq 0.363$

→ $e^{-0.6814 + 1.96 \cdot 0.17} = 0.701$.

άρα 95% ΔΕ για e^{β_3}: $0.363 \le e^{\beta_3} \le 0.701$.

(iii) $e^{\hat{\beta}_i}$ για M_2 -

M_2: $\hat{\beta}_1 = 5.982 \cdot 10^{-3}$, $\hat{\beta}_3 = -6.814 \cdot 10^{-1}$

$e^{\hat{\beta}_1} \simeq 1.006$, $e^{\hat{\beta}_3} \simeq 0.506$

$$\mu_x = E(Y_x) = n_x p_x = n_x \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}} \rightarrow p_x$$

$$\frac{p_x}{1-p_x} = \frac{\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}}{1-\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}} = \frac{\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}}{\frac{1+e^{x_i'\beta}-e^{x_i'\beta}}{1+e^{x_i'\beta}}} = e^{x_i'\beta} = e^{\beta_0+\beta_1 x_1+\beta_3 x_3}$$

Αν φοιτητής $(x_3=1)$: $\frac{p_x}{1-p_x} = e^{\beta_0+\beta_1 x_1+\beta_3} = e^{\beta_3} \cdot e^{\beta_0+\beta_1 x_1}$

Αν όχι φοιτητής $(x_3=0)$: $\frac{p_x}{1-p_x} = e^{\beta_0+\beta_1 x_1}$ ●

$$\frac{\left(\frac{\hat{p}_x}{1-\hat{p}_x}\right)_{x_3=1}}{\left(\frac{\hat{p}_x}{1-\hat{p}_x}\right)_{x_3=0}} = e^{\hat{\beta}_3} \simeq 0.506 \simeq 50.6\% \rightarrow$$ δηλ. για φοιτητή η πιθανότητα σχετική πολλαπλασιάζεται με περίπου 0.506

Άρα αν κάποιος είναι φοιτητής (εν σχέση με έναν μη φοιτητή) συνεπάγεται μείωση κατά 49.4% στη σχετική πιθανότητα αποπληρωμής της πιστωτικής κάρτας. Ανεκμετάλλευτο αποτέλεσμα για έναν φοιτητή.

(iv) Το εμβαδόν κάτω από τη ROC curve είναι το AUC. Καλή/μεγάλη επιτυχία πρόβλεψης ενός μοντέλου σημαίνει πως υπάρχουν τιμές του ορίου για με υψηλή ευαισθησία και ταυτόχρονα υψηλή εξειδίκευση. Τότε η ROC πλησιάζει την άνω αριστερή γωνία του τετραγώνου του σχήματος. Ένας δείκτης που μετρά κατά πόσο πλησιάζει αυτή τη γωνία, είναι το εμβαδόν κάτω από τη καμπύλη (area under curve AUC) με μέγιστη τιμή το 1.
Εδώ βλέπουμε ότι M_1, M_2 έχουν την καλύτερη ROC και ίσα βαθιά. Όμως M_2 έχει λιγότερες μεταβλητές ⇒ επιλογή του καλύτερου.