Απλό γραμμικό Μοντέλο

Για x μη στοχαστική και Y στοχαστική εξαρτημένη μεταβλητή

$$E(Y|x) = E(Y_x) = \beta_0 + \beta_1 x = \mu_x$$

⇒ Εκτίμηση : $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Πραγματικότητα : $Y_i = E(Y_{xi}) + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ (τυχαίο σφάλμα)

Για τα $\hat{\beta}_0$ και $\hat{\beta}_1$: ελαχιστοποίηση του συνολικού τυχαίου σφάλματος

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \underbrace{(Y_i - E(Y_{xi}))}_{\varepsilon_i}^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

⇒ Προκύπτει για $(\hat{\beta}_0, \hat{\beta}_1)$ ότι : $\left.\frac{\partial S(\beta_0,\beta_1)}{\partial \beta_0}\right|_{(\hat{\beta}_0,\hat{\beta}_1)} = 0$

και $\left.\frac{\partial S(\beta_0,\beta_1)}{\partial \beta_1}\right|_{(\hat{\beta}_0,\hat{\beta}_1)} = 0$

όπου $\hat{\beta}_0 = -\hat{\beta}_1 \bar{x} + \bar{y}$

και $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$

$S_{xy} = \sum_{i=1}^{n} y_i (x_i - \bar{x})$

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

Υποθέσεις για τα τυχαία σφάλματα ε_i :

1) $E(\varepsilon_i) = 0$
2) $V(\varepsilon_i) = \sigma^2$ (ομοσκεδαστικότητα)
3) $cov(\varepsilon_i, \varepsilon_j) = 0$ $\forall i,j$ ασυσχέτιστα - ανεξάρτητα.
4) $\varepsilon_i \sim N(0, \sigma^2)$

Κατανομές τ.μ.

1) $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \Rightarrow E(Y_i) = \mu_i$
$\Rightarrow V(Y_i) = V(\varepsilon_i) = \sigma$
$\Rightarrow cov(Y_i, Y_j) = cov(\varepsilon_i, \varepsilon_j) = 0$
$\Big\} Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

2) $\bar{Y} \sim N(\beta_0 + \beta_1 x_i, \frac{\sigma^2}{n})$

3) $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \Rightarrow E(\hat{\beta}_1) \overset{\text{αποδ.}}{=} \frac{1}{S_{xx}} E(S_{xy}) = \beta_1$

$\Rightarrow V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$

$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$ $\Rightarrow$ Τυποποίηση $\frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{S_{xx}}} \sim N(0,1)$

✱ Όμως σ^2 άγνωστο

Άρα θα χρησιμοποιήσουμε τον S^2, ο οποίος δεν είναι αμερόληπτος εκτιμήτρια: $\boxed{S^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2} \rightarrow e_i^2$

$= \frac{1}{n-2} SSE = \boxed{\frac{1}{n-2}\left(S_{YY} - \frac{S_{xy}^2}{S_{xx}}\right)}$

✱ Ισχύει ότι:

$$S_{YY} = SST = SSE + SSR$$
$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\bar{Y} - \hat{Y}_i)^2$$

** $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 0$

Άρα έχουμε ότι: $\frac{\sigma^2}{S_{xx}} \rightarrow \frac{S^2}{S_{xx}} = \hat{V}(\hat{\beta}_1)$ ← εκτίμηση

✱ Τυπικό σφάλμα $se(\hat{\beta}_1) = \sqrt{\hat{V}(\hat{\beta}_1)}$

και ισχύει ότι $t = \frac{\hat{\beta}_1 - \beta_1}{S / \sqrt{S_{xx}}} \sim t_{n-2}$

✱ γιατί $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-2}$

(t-test με $H_0: \beta_1 = \beta_1(0)$ και $H_1: \beta_1 \neq \beta_1(0)$

• p-value: $P(|t_{n-2}| > t)$

• για a δεδομένο επιλέγουμε εσάν t αυτόν την περιοχή όπου $P(|t| > t_{a/2}) = a$)

και ΔΕ με $(100)(1-a)\%$ βαθμό εμπιστοσύνης

$$\left[\hat{\beta}_1 - t_{n-2, a/2} \frac{S}{S_{xx}}, \hat{\beta}_1 + t_{n-2, a/2} \frac{S}{S_{xx}}\right]$$

4) Για την $\hat{b}_0 \sim N\left(b_0, \sigma^2 \frac{\sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$ → μη ομοσκεδαστική

* προσοχή: οι εκτιμήτριες $\hat{b}_0$ και $\hat{b}_1$ δεν είναι ασυσχέτιστες: $\text{cov}(\hat{b}_0, \hat{b}_1) = -\frac{\bar{x}\sigma^2}{S_{xx}}$

5) Για την $\hat{Y}_i \sim N\left(\mu_i, \sigma^2\left(\frac{1}{n} + \frac{(x_i - \bar{x})}{S_{xx}}\right)\right)$

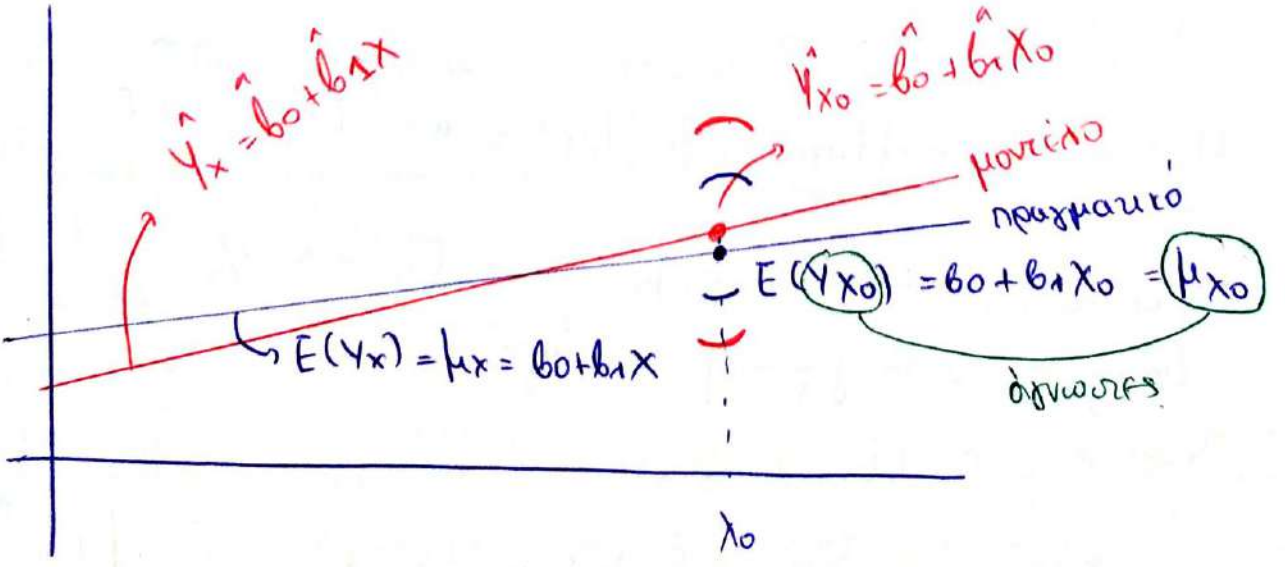
μη ομοσκεδαστική

Συντελεστής προσδιορισμού R^2: $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
ποσοστό μεταβλητότητας της Y που εξηγείται από την X. Όταν $R^2 \to 1$, τότε η εξάρτηση είναι ισχυρή.

Συντελεστής συσχέτισης Pearson: $r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \in [-1, 1]$

↳ Για το απλό γραμμικό μοντέλο: $r_{xy}^2 = R^2$

Πρόβλεψη σημειακής τιμής $\hat{Y}_i$ και μέση τιμής



Διαστήματα Εμπιστοσύνης πρόβλεψης και μέσης τιμής πρόβλεψης

$$E(\hat{Y}_{x_0}) = \mu_{x_0}$$

$$V(\hat{Y}_{x_0}) = \sigma^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)$$

} με κανονική κατανομή $\sim N(0,1)$ και αν αντί για $\sigma \longrightarrow S$

τότε $t = \dfrac{\hat{Y}_{x_0} - \mu_{x_0}}{S\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}} \sim t_{n-2}$

με Δ.Ε: $\left[\hat{Y}_{x_0} \pm t_{\alpha/2} S\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}\right]$

Αντίστοιχα για το Y_{x_0}: $\hat{Y}_{x_0} - Y_{x_0} \sim N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)\right)$

και αντικαθιστώντας το $\sigma \longrightarrow S$: $t = \dfrac{\hat{Y}_{x_0} - Y_{x_0}}{S\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}} \sim t_{n-2}$

με Δ.Ε: $\left(\hat{Y}_{x_0} \pm t_{\alpha/2} S\sqrt{1 + \frac{1}{n}\frac{(x_0-\bar{x})^2}{S_{xx}}}\right)$

Παρατήρηση: → το διάστημα εμπιστοσύνης της πρόβλεψης Y_{x_0} είναι μεγαλύτερο από της μέσης τιμής μ_{x_0}

Παρατηρήσεις:

1) **Μετατροπή μη γραμμικών σε γραμμικό:** π.χ. $Y_i = ae^{bx_i} + \varepsilon_i$
$\Rightarrow \ln Y_i = b_0 + b_1 x_i + \varepsilon_i^*$
④ $1/Y \sim 1/x$

} θα πρέπει να το φέρουμε σε κατάλληλη μορφή ώστε να έχουμε ανοίχηση.

Θέτοντας κάθε τι που υπολογίζουμε θα πρέπει να το γυρίσουμε στο αρχικό

⊗ όμως $E(Y^*) = E(1/Y) \neq 1/E(Y)$
και $E(Y^*) = E(\ln Y) \neq \ln E(Y)$

} μόνο προσεγγιστικά μπορούμε να δώσουμε τα Δ.Ε.

Ερμηνεία Συντελεστών

$$\hat{Y}_0 = \hat{b}_0 + \hat{b}_1 x_0$$

1) $\hat{b}_1$: αν αυξηθεί η X κατά μια μονάδα, αναμένεται η Y να αυξηθεί κατά b_1 (δεν υπάρχει εξάρτηση από το X)

2) Αν υπάρχει συντελεστής από μεταβλητή x^2 :

$$E(Y_x) = b_0 + b_1 x + b_2 x^2 + \text{άλλες μεταβλητές}$$

αν X αυξηθεί κατά μια μονάδα, τότε η διαφορά των αναμενόμενων τιμή θα είναι : $E(Y_{x+1}) - E(Y_x) = b_1 + b_2 + 2 x b_2$

εξάρτηση από το X

Έλεγχος υποθέσεων απλού γραμμικού μοντέλου

1) Residuals e_i vs Fitted values : θέλουμε να παρατηρήσουμε τυχαιότητα

2) Normal Q-Q plot μεταξύ e_i και s_i

a) βρίσκουμε τα ποσοστιαία σημεία του δείγματος (για τα e_i)

b) συγκρίνουμε τα ποσοστιαία σημεία του s_i

Θέλουμε να ακολουθούν μια ευθεία.

Πολλαπλό Γραμμικό Μοντέλο

ΒΑΣΙΚΑ ΣΤΟΙΧΕΙΑ

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \quad \text{με} \quad \underline{\varepsilon} \sim N_n(0, \sigma^2 I_n)$$

↳ ελαχιστοποιούμε $\underline{\varepsilon}$

- $E(\varepsilon_i) = 0$
- $Var(\varepsilon_i) = \sigma^2$ ομοσκεδαστικότητα
- $Cov(\varepsilon_i, \varepsilon_j) = 0$

$$S(\underline{\beta}) = \dots = \sum_{i=1}^{n} \varepsilon_i^2$$

$$\frac{\partial S(\underline{\beta})}{\partial \underline{\beta}} = 0 \Rightarrow \boxed{\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{Y}} \Rightarrow \hat{\underline{Y}} = X\hat{\underline{\beta}} = X(X'X)^{-1}X'\underline{Y}$$

$$\Rightarrow \boxed{\hat{\underline{Y}} = H\underline{Y}}$$

πίνακας προβολής

(*) υπόλοιπα: $e = Y - \hat{Y} = (I-H)Y$
$= (I-H)(X\beta+\varepsilon) = (I-H)\varepsilon$

Τι καταναμές ακολουθούν οι βασικές μεταβλητές;

1) Υπόθεση: $\boxed{\underline{\varepsilon} \sim N_n(\underline{0}, \sigma^2 I)}$, $\varepsilon_i \sim N(0, \sigma^2)$ κ.λ.π.

2) $\boxed{\underline{Y} \sim N_n(X\underline{\beta}, \sigma^2 I_n)}$ (*) $V(\underline{Y}) = E((\underline{Y}-X\underline{\beta})(\underline{Y}-X\underline{\beta})') = \dots = E(\underline{\varepsilon}\underline{\varepsilon}')$
$= E((\underline{\varepsilon}-E(\underline{\varepsilon}))(\underline{\varepsilon}-E(\underline{\varepsilon}))') = V(\underline{\varepsilon})$

3) για την $\hat{\underline{\beta}} = \underbrace{(X'X)^{-1}X'}_{A}\underline{Y} \rightarrow E(\hat{\underline{\beta}}) = AE(\underline{Y}) = AX\underline{\beta} = \underline{\beta}$

$\rightarrow V(\hat{\underline{\beta}}) = V(AY) \overset{\text{τύπος}}{=} AV(Y)A' = \sigma^2 AA' = \sigma^2 \underbrace{(X'X)^{-1}}_{C}$

άρα $\boxed{\hat{\underline{\beta}} \sim N_p(\underline{\beta}, \sigma^2(X'X)^{-1})} \rightsquigarrow \hat{\beta}_j \sim N(\beta_j, \sigma^2 C_{jj})$ με $cov(\hat{\beta}_j, \hat{\beta}_\ell) = \sigma^2 C_{j\ell}$,

↳ $p = k+1$ → bias = αριθμός παραμέτρων
↳ μεταβλητές

όμως σ^2 άγνωστο
↳ αντικαθίσταται με $S^2 = \dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(p)} = \dfrac{SSE}{n-p} = \dfrac{\sum e_i^2}{n-p}$ $\left(E(S^2) = \sigma^2\right)$

↳ στο απλό γραμμικό ήταν $p=2$

άρα όταν θα κάνουμε έλεγχος, αυτός θα γίνεται με την

μεταβλητή: $t_j = \frac{\hat{\beta}_j - \beta_j}{s\sqrt{c_{jj}}} \sim t_{n-p}$

και $se(\hat{\beta}_j) = s\sqrt{c_{jj}}$

4) $\underline{\hat{Y}} = H\underline{Y} = X\hat{\underline{\beta}} \sim N_n(X\underline{\beta}, \sigma^2 H)$

(*) $V(\hat{Y}) = V(HY) = E\{(HY - E(HY))(HY - E(HY))'\} = HV(Y)H' = \sigma^2 HH' = \sigma^2 H$ γιατί H συμμετρικοί ταυτοδύναμος

Προσοχή: η $\hat{Y}_i$ με τις $\hat{Y}_j$ δεν είναι ανεξάρτητες

$cov(\hat{Y}_i, \hat{Y}_j) = cov(\hat{Y}_i, Y_i) = \sigma^2 h_{ij}$

$\hat{Y}_i = \underline{\beta}' x_i = \sum_{j=1}^{n} h_{ij} Y_j$ και άρα $\hat{Y}_i \sim N(\underline{\beta}' x_i, \sigma^2 h_{ii})$

5) Για τα υπόλοιπα $\underline{e} \sim N_n(\underline{0}, \sigma^2(I-H))$

$e_i \sim N(0, \sigma^2(1-h_{ii}))$ Δεν υπάρχει πια η ομοσκεδαστικότητα

και πλέον δεν υπάρχει ανεξαρτησία

$cov(e_i, e_j) = -\sigma^2 h_{ij}$

(*) SOS: δεν χρησιμοποιούμε ως αμερόληπτη εκτιμήτρια του σ^2 την $\hat{\sigma}^2 = \frac{e'e}{n}$

γιατί $\frac{E(e_i^2)}{n} = \frac{V(e_i) + \cancel{E(e_i)}^{=0}}{n} = \frac{\sigma^2(1-h_{ii})}{n} = \frac{\sigma^2}{n}\sum(1-h_{ii}) \overset{(*)}{=} \frac{\sigma^2}{n}(n-p) = \frac{\sum e_i^2}{n}$

(*) $tr(H) = tr(X(X'X)^{-1}X') \underset{tr(AB)=tr(BA)}{=} tr((X'X)(X'X)^{-1}) = tr(I) = p \neq \sigma^2$

άρα $\sum_{i=1}^{n} h_{ii} = p$

(*) τα αποτελέσματα αυτά προκύπτουν από την ανάλυση ελαχίστων τετραγώνων.

Εκτίμηση Μέγιστης Πιθανοφάνειας

$$\ell = \ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\underline{Y}-\underline{X}\underline{b})'(\underline{Y}-\underline{X}\underline{b})$$

$$\Rightarrow \frac{\partial \ell}{\partial \underline{b}} = 0 \Rightarrow \hat{b} = (X'X)^{-1}X'Y$$

$$\frac{\partial \ell}{\partial \sigma^2} = 0 \Rightarrow \hat{\sigma}^2 = \frac{\underline{e}'\underline{e}}{n} = \frac{SSE}{n}$$

(δεν είναι αμερόληπτη ενώ η $S^2 = \frac{SSE}{n-p}$ είναι)

Μέγιστη πιθανοφάνεια μοντέλου

$$\hat{\ell} = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\frac{SSE}{n}\right) - \frac{n}{2}$$

∴ εξαρτάται από το σφάλμα SSE

Έλεγχοι

1) Έλεγχος του λόγου των πιθανοφανειών

$H_0: M_0$, $H_1: M_1$ με $M_0 \subset M_1$

$$-2(\hat{\ell}_0 - \hat{\ell}_1) = n\ln\left(\frac{SSE_0}{SSE_1}\right) \sim \chi^2_d$$

↳ διαφορά παραμέτρων.

2) Στατιστικός έλεγχος F για $H_0: M_0$, $H_1: M_1$ με $M_0 \subset M_1$

$$F^* = \frac{(SSE_0 - SSE_1)/d}{SSE_1/(n-p)} \sim F_{d,(n-p)}$$

(d → διαφορά αριθμού παραμέτρων; $(n-p)$ → παράμετροι M_1)

απορρίπτουμε την H_0 για μικρά p-values: $p = P[F > F^*]$

3) Στατιστικός έλεγχος $t \equiv F$ έλεγχος για μοντέλο με διαφορά μιας μεταβλητής

$H_0: M_0 (b_j = 0)$, $H_1: M_1 (b_j \neq 0) \Rightarrow F = \frac{(SSE_0 - SSE_1)/1}{SSE_1/(n-p)} = t^2_{n-p} \sim F_{1,n-p}$

↳ όσο πιο μεγάλη F_1 τόσο μικρότερη η p-value και τόσο μεγαλώνει το σφάλμα SSE_0

$$\text{p-value} = P(F > F^*) = F(|t_i| > t^*)$$

Diagnostics with residuals

1) Κανονικά υπόλοιπα $\underline{e} \sim N_n(0, \sigma^2(I-H))$, $e_i \sim N(0, \sigma^2(1-h_{ii}))$

Παρατηρηση 1: δεν υπαρχει ομοσκεδαστικότητα καθως h_{ii} διαφορετικό $\forall i$. (σε αντίθεση με τα ε_i που εχαν ομοσκεδαστικότητα με σ^2)

(*) $e_i = Y_i - \sum_{j=1}^{n} h_{ij} Y_j$

$= \varepsilon_i - \sum_{j=1}^{n} h_{ij} \varepsilon_j$

(*) ισχύει ότι $\frac{1}{n} \le h_{ii} \le \frac{1}{c_i}$

2) Staudarized residuals

$r_i = \frac{e_i}{S\sqrt{1-h_{ii}}}$ έχει variance 1 αλλά δεν είναι normally distributed

$S^2 = \frac{SSE}{n-p}$

3) Deleted residuals ≡ staudrized residuals όπου κάνουμε fit το μοντέλο σε όλες τις παρατηρήσεις εκτός από την i.

$r_i' = \frac{e_i}{S_{(i)}\sqrt{1-h_{ii}}} \sim t_{n-p-1}$

$S_i^2 = \frac{SSE(i)}{n-p-1}$

4) Press residuals: Εαν αφαιρέσουμε την i-οστή παρατήρηση προκύπτει μια εναλλακτική εκτίμηση $\hat{Y}_{(i)}$

$\text{press\_residual} = Y_i - \hat{Y}_{(i)} = \frac{e_i}{1-h_{ii}}$

(*) Έλεγχοι: Δεν πρέπει να παρατηρούμε patterns

Θα → e_i vs e_{i-1}

→ e_i vs i

→ e_i vs X_i

→ e_i vs $\hat{Y}_i$ → δεν ισχύει ότι αναμένουμε ομοσκεδαστικότητα

$COV(\underline{e}, \hat{Y}) = 0$

Όσο πιο μικρό το press residual, τόσο πιο μεγάλη είναι η ικανότητα του μοντέλου να δίνει σωστές προβλέψεις για άγνωστες παρατηρήσεις.

Κριτήρια Επιλογής Μοντέλου

1) Δείκτης $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \leq 1$

Όσο περισσότερες μεταβλητές προσθέτουμε, τόσο θα μικραίνει το SSE. Μας ενδιαφέρουν οι υψηλές αυξομειώσεις για να καταλάβουμε ότι η μεταβλητή που προστίθεται ή αφαιρείται έχει μεγάλη γραμμική εξάρτηση με την Y

2) Adjusted R^2 : $\bar{R}^2 = 1 - \frac{(SSE)/(n-p)}{SST/(n-1)}$

M_1: μοντέλο με p μεταβλητές

M_2: μοντέλο με p+q μεταβλητές

$SSE_2 < SSE_1 \Rightarrow R_2^2 > R_1^2$

αυξάνεται μόνο αν προστεθεί μια μεταβλητή που βελτιώνει πολύ το μοντέλο

ανισχύει $\bar{R}_2 > \bar{R}_1 \Rightarrow \frac{SSE_2}{n-p-q} < \frac{SSE_1}{n-p} \dots \Rightarrow \boxed{\frac{(SSE_1 - SSE_2)/q}{SSE_2/(n-p-q)} > 1}$

↳ Εάν η τιμή της ελεγχοσυνάρτησης F είναι >1 τότε η προσθήκη q μεταβλητών είναι σημαντική για το μοντέλο.

3) $R^2_{prediction}$: βασίζεται στα υπόλοιπα PRESS

$$R^2_{pred} = 1 - \frac{PRESS}{SST} \quad \text{όπου} \quad PRESS = \sum_{i=1}^{n} \left(\frac{e_i}{1-h_{ii}}\right)^2$$

Όσο πιο μικρή είναι η τιμή PRESS, τόσο λιγότερο επηρεάζει το μοντέλο η αφαίρεση μιας παρατήρησης i $\Rightarrow$ επομένως το μοντέλο έχει υψηλή δυνατότητα πρόβλεψης και άρα όσο πιο μεγάλο το R^2_{pred}

4) C_p - Mallows

$$C_p = \frac{SSE(p)}{SSE(p')}(n-p') + 2p - n$$

$SSE(p)$ → υπο εξέταση μοντέλο

$SSE(p')$ → πλήρη μοντέλο : p' παράμετροι

Θέλουμε $\boxed{C_p \leq p}$ ⊛ $C_p = p$ για το πλήρη μοντέλο.

5) $AIC = -2\hat{\ell} + 2q$ → αριθμός παραμέτρων μοντέλου

$$AIC = n\ln\left(\frac{SSE}{n}\right) + 2p + n\ln(2\pi) + n + 2$$

(normal $\hat{\epsilon}$)

→ Θέλουμε να είναι μικρό και παράλληλα

Κριτήριο Πολυσυγγραμμικότητας

⊛ ο έλεγχος F συμφωνεί με το AIC λόγω SSE

οι μεταβλητές να είναι σχετικά σημαντικές

$VIF_j = \frac{1}{1-R_j^2}$ αν $VIF > 5 \Rightarrow$ πολυσυγγραμμικότητα

Κριτήρια για Σημεία Επιρροής

$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_j - \bar{x})^2}$

όταν x_i μακριά από $\bar{x}$ τότε h_{ii} υψηλό

1) $h_{ii} = x_i'(x'x)^{-1}x_i$ αν $h_{ii} > 2p/n$ η παρατήρηση i (hat values) θεωρείται σημείο επιρροής.

2) Απόσταση Cook $D_i = \frac{e_i^2 h_{ii}}{pS^2(1-h_{ii})^2} = \frac{r_i^2 h_{ii}}{p(1-h_{ii})}$ $>> 1$ τότε θεωρείται σημείο επιρροής

απόσταση $\hat{b}$ από $\hat{b}_{(i)}$

→ έχει αφαιρεθεί η i-οστή παρατήρηση.

3) $DFBETAS_{ji} = \dfrac{\hat{b}_j - \hat{b}_{j(i)}}{\sqrt{s^2_{(i)} c_{jj}}}$

αν $|DFBETAS_{ji}| > 2/\sqrt{n}$ τότε η i-οστή παρατήρηση έχει επιρροή στον-συντελεστή b_j

4) $DFFITS_i = \dfrac{\hat{Y}_i - \hat{Y}_{(i)}}{\sqrt{s^2_{(i)} h_{ii}}}$, $|DFFITS_i| > 2\sqrt{p/n}$ τότε η i-οστή παρατήρηση ίσως είναι influential point.

Διαγνωστικά γραφήματα για τα αναγκαίοτητα των μεταβλητών

1) Added variable plots $\quad \underline{Y} = X\underline{b} + \underline{\varepsilon}$

extra μεταβλητή

$\underline{Y} = X\underline{b} + \gamma W + \varepsilon^*$

$\curvearrowright \cdot(I-H)$

$\underline{e} = \hat{\gamma}(I-H)W + (I-H)\varepsilon^*$

Αν $\underline{e} \sim [I-H]W$ σχηματίζω ευθεία γραμμή, τότε η W είναι αναγκαία στο μοντέλο $\Leftarrow$

υπόλοιπα όταν κάνουμε regression του Y ως προς X

υπόλοιπα όταν κάνουμε regression του W ως προς X

2) Partial residual plots

$E(Y) = X^+\underline{b}^+ = X\underline{b} + \gamma W$

residual: $\underline{e}^+ = \underline{y} - X^+\underline{\hat{b}}^+$

$= (\underline{y} - X\underline{\hat{b}}) - \hat{\gamma}W$

$= \underline{\tilde{e}} - \hat{\gamma}W$

$\underline{\tilde{e}} = e^+ + \hat{\gamma}W$

plot: $\underline{\tilde{e}}$ και W

$\Rightarrow$ αναμένω με γραμμική σχέση

$\Rightarrow$ αν δεν είναι, τότε η μεταβλητή χρειάζεται μετασχηματισμό.

Παρατηρήσεις

Αν έχουμε μοντέλο που περιέχει και indicator μεταβλητές - κατηγορικές

π.χ. $Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3$ (1)

όπου $X_2 = \begin{cases} 0 \\ 1 \end{cases}$ με βάση κάποια αυθαίρετη που διαχωρίζει τα δεδομένα σε δυο ομάδες.

και $X_3 = X_1 \cdot X_2$ μεταβλητή αλληλεπίδρασης.

↝ Μπορούμε να προσαρμόσουμε το μοντέλο απευθείας

↝ ή μπορούμε να κάνουμε διαφορετικά, μπορεί η κλίση της ευθείας παλινδρόμησης, ή η τομή με τους άξονες αλλάζει για μια από τις δυο ομάδες.

Για $X_2 = 0$: $E(Y_I) = b_0 + b_1 X_1$ (2)

Για $X_1 = 1$: $E(Y_{II}) = (b_0 + b_2) + (b_1 + b_3) X_1$ (3)

Με ελέγχους F εξετάζουμε: (ανάλυση ανάλογα των R).

1) $H_0: b_3 = 0$, $H_1: b_3 \neq 0$ στο μοντέλο (1)

2) Αν $b_3 \neq 0$ τότε υπάρχει αλληλεπίδραση και έτσι δείχνουμε δυο διαφορετικές ευθείες

3) Αν $b_3 = 0$ (μπαίνουμε σημαντική η H_0), ελέγχουμε τον $H_0: b_2 = 0$ και $H_1: b_2 \neq 0$ στο $Y = b_0 + b_1 X_1 + b_2 X_2$

4) Αν $b_2 \neq 0$ τότε έχουμε δυο παράλληλες ευθείες
Αν $b_2 = 0$ τότε έχουμε μια κοινή ευθεία.

Poisson Παλινδρόμηση

- Η τ.μ Y ακολουθεί κατανομή Poisson με αναίρεση πυκνότητα πιθανότητας : $f(Y) = \frac{e^{-\mu}\mu^Y}{Y!}$, $Y = 0,1,2,\dots$

$$E(Y) = \mu$$
$$V(Y) = \mu$$

- Για την παράμετρο μ: (αναμενόμενη τιμή)

$$\mu \to \mu_x = e^{x'\underline{\beta}} > 0$$

$$\Rightarrow \ln \mu_x = x'\underline{\beta} = g(\mu_x) \quad \text{με} \quad \boxed{g(x) = \ln x}$$

log link function

(*) στο γραμμικό μοντέλο είχαμε: $Y \sim N(\mu_x, \sigma^2)$

$$g(\mu_x) = \mu_x = x'\beta$$

- Οι παράμετροι $\hat{\beta}$ προκύπτουν από την μεγιστοποίηση της πιθανοφάνειας : $L = \prod_{i=1}^{n} f(y_i, \mu_i)$ όπου $\mu_i = \mu_{xi} = e^{x_i'\underline{\beta}}$

$$l = \ln L = \sum_{i=1}^{n} \left[-e^{x_i'\underline{\beta}} + y_i x_i'\beta - \ln(y_i!) \right]$$

$\frac{\partial}{\partial \beta_j} \Rightarrow$ $X'(Y - \hat{\mu}) = 0$ (μη γραμμικό ως προς X) — επαναληπτική μέθοδος

όχι όπως γραμμικό μοντέλο $X'(Y - X\hat{\beta}) = 0$

- Ερμηνεία Συντελεστών : e^{β_j} πολλαπλασιαστική αλλαγή στην αναμενόμενη τιμή της Y (στο μ δηλαδή), όταν το X_j αυξάνεται κατά 1 μονάδα. (πρέπει να πάρω το εκθετικό της τιμής που δίνει η R για τους συντελεστές)

Στατιστικοί Έλεγχοι Συμπερασμάτων Μεταβλητών

1) Στατιστικός έλεγχος Wald $\hat{\beta}_j \sim N(\beta_j, \hat{V}(\hat{\beta}_j))$

προκύπτει από τον Hessian μήτρα των δεύτερων παραγώγων $\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_m}$

$$\frac{\hat{\beta}_j - \beta_j}{(\hat{V}(\hat{\beta}_j))^{1/2}} \sim N(0,1)$$

με $H_0: \beta_j = 0$, $H_1: \beta_j \neq 0$

2) Likelihood ratio test: $\hat{\ell}_1 = \max_{\underline{\beta}} \ell(\beta_0, \dots, \beta_k)$

$\hat{\ell}_0 = \max_{\underline{\beta}} \ell(\beta_0, \dots, \beta_k)$ ↳ nested στο M_1 με d λιγότερες μεταβλητές.

$-2(\hat{\ell}_0 - \hat{\ell}_1) \sim \chi^2_d$

3) Deviance log-likelihood ratio statistic

H_0: our model ⇝ restriction $g(\mu_i) = x_i' \underline{\beta}$

H_1: saturated model ⇝ δεν επιβάλλονται περιορισμοί ⇒ αριθμός παραμετρικών = αριθμός παρατηρήσεων.

Υπολογίζεται η πιθανοφάνεια $\left\{\tilde{\mu}_i = y_i \;\; \frac{\partial \ell}{\partial \mu_i} = 0\right\}$

για H_0: $\hat{\mu}_i = \hat{y}_i = e^{x_i \hat{\beta}}$

για H_1: $\tilde{\mu}_i = y_i$

και υπολογίζουμε $D(\hat{\beta}) = -2\{\ell(\hat{\beta}) - \ell\}$

$$2\Big\{\sum_{i=1}^{n} y_i \ln(y_i/\hat{\mu}_i) - \sum_{i=1}^{n}(y_i - \hat{\mu}_i)\Big\} \quad (\text{το } \sum(y_i - \hat{\mu}_i) = 0)$$

$$D(\hat{\beta}) = 2\sum_{i=1}^{n} y_i \ln y_i/\hat{\mu}_i \sim \chi^2_{n-p}$$

Γενικά για H_0: M^* και H_1: M $(M^* \subset M)$: $-2(\ell(\hat{\beta}^*) - \ell(\hat{\beta})) \sim \chi^2_q$

Σχεδόν επιβάλλονται περιορισμοί όπου $\beta_j = 0$ $\quad D(\hat{\beta}^*) - D(\hat{\beta}) \sim \chi^2_q$

Κριτήρια Επιλογής Μοντέλων

1) $AIC = -2 \sum_{i=1}^{n} [-e^{x_i'\hat{\beta}} + y_i x_i'\hat{\beta} - \ln(y_i!)] + 2q$

2) Υπάρχαν τάσεις R^2 αλλά δεν είναι αξιόπιστοι

3) Έλεγχος Deviance μοντέλου σε σχέση με null deviance ⇒ θέλουμε να είναι όσο το δυνατόν μικρότερη.

Residuals

1) Rearson $r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$

2) Standardised $r_i^{PS} = \frac{r_i^P}{\sqrt{1-h_{ii}}}$

όπου h_{ii} διαγώνιο του $\hat{H} = \hat{W}^{1/2} X (X'\hat{W}X)^{-1} X' \hat{W}^{1/2}$

3) Deviance residual

$$r_i^D = \text{sgn}(y_i - \hat{\mu}_i) \left[2 \{ y_i \ln(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i) \} \right]^{1/2}$$

⇒ $D(\hat{\beta}) = \sum_{i=1}^{n} (r_i^D)^2$ ⇒ Standardize Deviance residual $r_i^{DS} = \frac{r_i^D}{\sqrt{1-h_{ii}}}$

4) Likelihood residuals (student των R) εξαρτώνται από (r_i^{PS}) και (r_i^{DS})

⇒ Partial residual plots : $r_{ji}^T = \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} + \hat{\beta}_j x_{ij}$

- αν είναι ευθεία η γραμμή τότε η x_j μένει ανεπηρέαστη
- αν δεν υπάρχει pattern, χρησιμοποιούμε μεταμορφώσεις.

⇒ Added variable plots (ίδιο με παλαιότερο.)

Offset στην Πολλωνυμική Poisson

Πρέπει να λάβουμε υπόψη το μέγεθος του πληθυσμού στον οποίο συμβαίνουν τα γεγονότα Y_i

$\mu = E(Y)$ = αναμενόμενος αριθμός γεγονότων/ανά μονάδα (κάτι)

π.χ για 7 αεροπορικές εταιρείες με κοινά χαρακτηριστικά, ο αναμενόμενος αριθμός αεροπλάνων θα είναι $\mu_1^* = 7\mu_1$

- Αν η αναμενόμενη συμπεριφορά του αντικειμένου, όπως το n_i δηλώνει τον αριθμό των αντικειμένων με τα κοινά χαρακτηριστικά.

άρα $\mu_i^* = n_i \mu_i = n_i \exp(x_i' \underline{b})$

$$\ln(\mu_i^*) = \ln(n_i) + \underline{b}' x_i$$

$$\Rightarrow \ln(\mu_i^*/n_i) = b' x_i = n_i$$

Logistic Regression

***Link function** : μας δίνει την γραμμική πρόβλεψη

- $Y \sim$ Normal : $\eta_x = g(\mu_x) = X'b$ (identity link function)
- $Y \sim$ Poisson : $\eta_x = g(\mu_x) = \ln \mu_x = X'b$ (log link function)
- $Y \sim$ Bernoulli ($n_x = 1$) ή $Y \sim$ Binomial ($n_x > 1$)

$$\eta_x = g(\mu_x) = \ln\left(\frac{\mu_x}{n_x - \mu_x}\right) = X'b$$

Στην περίπτωση της λογιστικής παλινδρόμησης, η εξαρτημένη μεταβλητή Y είναι διακριτή και ακολουθεί διωνυμική κατανομή. Δηλαδή η συναρτηση πιθανότητας για Y επιτυχίες σε n ανεξάρτητες δοκιμές bernoulli είναι :

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y} = \exp\left\{ y \ln\left[\frac{p}{1-p}\right] + n \ln(1-p) + \ln\binom{n}{y} \right\}$$

(odds)

(*) p: πιθανότητα επιτυχίας

μορφή exponential family όπου $\theta = \ln\left[\frac{p}{1-p}\right]$ δείχνει την link function

Logistic Regression Model :

$$\ln\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1 x_{i1} + \dots + b_k x_{ik} \quad , i = 1, \dots, n$$

όπου $p_i = p_{x_i} = \dfrac{\exp(b_0 + \dots)}{1 + \exp(b_0 + \dots)}$

και $\mu_i = E(Y_i) = n_i p_i = n_i \dfrac{e^{x_i' \underline{b}}}{1 + e^{x_i' \underline{b}}}$

Η αναρτιον πιθανοφάνεια είναι :

$$\ell = \ln L = \sum_{i=1}^{u} \left\{ y_i x_i' \underline{b} - n_i \ln(1 + e^{x_i' \underline{b}}) + \ln\binom{n_i}{y_i} \right\}$$

όπου $\dfrac{\partial \ell}{\partial b_j} = 0$ μας δίνει επαναληπτικές εξισώσεις για τους συντελεστές.

Ερμηνεία Συντελεστών

$$\frac{\hat{p}}{1-\hat{p}} = e^{x' \underline{b}} \quad \text{καθώς} \quad \hat{p} = \frac{e^{x' \hat{\underline{b}}}}{1 + e^{x' \hat{\underline{b}}}}$$

- αν $b_j < 0$ τότε ο λόγος odds θα υποδιπλασιάζεται κατά $e^{\hat{b}_j}$
- αν $b_j > 0$ τότε ο λόγος odds θα πολλαπλασιάζεται κατά $e^{\hat{b}_j}$

όταν η μεταβλητή X_j αυξηθεί κατά μία μονάδα.

* Αν μια μεταβλητή είναι η $\ln x$ και όχι η x, η ερμηνεία αλλάζει.

Αν η η X αυξηθεί κατά ένα ποσοστό π.χ. 10% τότε :

$\Delta(\text{log odds}) = b[\ln(x + 0.1x) - \ln x] = b \ln(1.1x/x) = b \cdot \ln(1.1)$

άρα η πολλαπλασιαστική μεταβολή των odds θα είναι $\exp(b \ln(1.1))$

Odds ratio

$$\frac{\frac{\hat{P}_1}{1-\hat{P}_1}}{\frac{\hat{P}_2}{1-\hat{P}_2}} = \exp(x_1 - x_2)'\underline{b}$$

και εάν υποθέσαμε ότι είχαμε μια ποσοτική μεταβλητή X_2 και μια κατηγορική X_1 τότε:

$$= \frac{\exp(b_0 + \hat{b}_2 x_2)}{\exp(b_0 + \hat{b}_1 + \hat{b}_2 x_2)} \rightsquigarrow$$

για οποιοδήποτε X_2 το odds ratio είναι $e^{\hat{b}_1}$

Έλεγχοι

1) Wald statistic : $\hat{\underline{b}} \sim N_p(\underline{b}, \hat{V}(\hat{\underline{b}}))$ → καταλληλη προσεγγιση

$$Z = \frac{\hat{b}_j - b_j(0)}{(\hat{V}(\hat{b}))^{1/2}} \sim N(0,1) \text{ ασυμπτωτικά}$$

$se(\hat{b}) =$

Όπου $H_0: b_j = 0$ vs $H_1: b_j \neq 0$ με

$100(1-\alpha)\%$ Δ.Ε.

$\hat{b}_j \pm z_{\alpha/2} se(\hat{b}_j)$

→ και για το odds ratio : $\exp(\hat{b}_j \pm z_{\alpha/2} se(\hat{b}_j))$

2) Deviance

M_0 (μοντέλο με περιορισμούς) $\subset M_1$

M_1 πλήρες μοντέλο

saturated model

Διαφορά Deviance : $D_0 - D_1 = -2(\hat{\ell}_0 - \hat{\ell}_s) + 2(\hat{\ell}_1 - \hat{\ell}_s)$

$= -2(\hat{\ell}_0 - \hat{\ell}_1) \sim \chi^2_{p_1 - p_2}$ → διαφορά παραμέτρων

* πάντα θετικό γιατί M_0 έχει λιγότερες παραμέτρους από M_1 και άρα (deviates) απολύτως περισσότερο από το saturated

* όσο πιο μεγάλη η διαφορά, τόσο πιο πιθανό να απορρίψουμε το M_0.

Deviance

Binomial Data

$Y_i \sim b(n_i, p)$

$\mu_i = E(Y_i) = n_i p_i$

$$D(\hat{b}) = -2 \sum_{i=1}^{u} y_i \ln\left(\frac{\hat{p}_i}{\tilde{p}_i}\right) + (n_i - y_i)\ln\left(\frac{1-\hat{p}_i}{1-\tilde{p}_i}\right)$$

για $\tilde{p}_i = \frac{y_i}{n_i}$ και $\hat{p}_i = \frac{\hat{\mu}_i}{n_i}$

Παρατήρηση: υπάρχει εξάρτηση της D από την προβλεπόμενη τιμή $\hat{\mu}_i$ και την παρατηρούμενη τιμή Y_i. Όποτε η deviance ενός μοντέλου για τα binomial data έχει νόημα να την εξετάσουμε και μόνη της. Όσο πιο χαμηλή, τόσο πιο κοντά είναι η παρατήρηση Y_i στην προβλεπόμενη $\hat{\mu}_i$

Binary Data

$Y \sim B(p_i)$

$\mu_i = E(Y_i) = p_i$

$$D(\hat{b}) = -2 \sum_{i=1}^{u} \{ \hat{\mu}_i \, \text{logit}(\hat{\mu}_i) + \ln(1-\hat{\mu}_i) \}$$

Παρατήρηση: η D εξαρτάται μόνο από την προβλεπόμενη τιμή $\hat{\mu}_i$. Άρα δεν έχει νόημα να την εξετάσουμε μόνη της. Μόνο με σύγκριση των deviance δύο μοντέλων.

Επιλογή Μοντέλου

$$AIC = 2 \sum_{i=1}^{u} \left[n_i \ln(1 + e^{x_i' \hat{\beta}}) - y_i x_i' \hat{\beta} - \ln\binom{n_i}{y_i} \right] + 2p$$

Diagnostics with residuals: ισχύουν τα αντίστοιχα.

(*) χρήση half normal QQ plot για ανίχνευση outliers

Likelihood residuals: συνδυασμός των pearson και deviance residuals. Προσεγγίζει την μεταβολή της deviance αν παραλειφθεί η i-οστή παρατήρηση. Όσο πιο μεγάλη μεταβολή, τόσο πιο πιθανό να είναι outlier.