ΘΕΜΑΤΑ 2021

ΖΗΤΗΜΑ 1

Ισχύει $x_2 = 0 \rightarrow$ χαμηλή, $x_2 = 1 \rightarrow$ υψηλή, οπότε
$x_3 = 0 \rightarrow$ χαμηλή, $x_3 = x_1 \rightarrow$ υψηλή

Το μοντέλο $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ ισοδυναμεί με

$$E(y_\chi) = \beta_0 + \beta_1 x_1 \rightarrow \text{χαμηλή} \quad \text{ή με}$$

$$E(y_\upsilon) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 \rightarrow \text{υψηλή}.$$

Συγκρίνοντας τις σχέσεις αυτές, γίνεται αντιληπτό πως (I) εάν $\beta_3 \neq 0$, τότε οι συντελεστές του x_1 διαφέρουν στις δύο περιπτώσεις και ως εκ τούτου απαιτούνται δύο ευθείες διαφορετικών κλίσεων για την προσαρμογή στα δεδομένα. (II) Εάν $\beta_3 = 0$, αλλά $\beta_2 \neq 0$, τότε ξανά δύο ευθείες θα πρέπει να προσαρμοστούν στα δεδομένα, μία για κάθε είδος (υψηλή/χαμηλή), μόνο που στην περίπτωση αυτή η κλίση τους θα είναι κοινή και ίση με β_1. (παράλληλες ευθείες). Τέλος, (III) εάν $\beta_3 = 0$ και $\beta_2 = 0$, τότε οι άνω σχέσεις ταυτίζονται, οπότε μία μόνο ευθεία αρκεί για την προσαρμογή στα δεδομένα.

Οι έλεγχοι που προτείνονται στα πλαίσια αυτά είναι οι ακόλουθοι: αρχικά, ο έλεγχος της υπόθεσης $H_0: \beta_3 = 0$ έναντι της $H_1: \beta_3 \neq 0$, η οποία ισοδυναμεί με την υπόθεση απουσίας αλληλεπίδρασης μεταξύ των x_1 και x_2. Εάν η υπόθεση αυτή απορριφθεί, τότε είναι βέβαιο πως ισχύει η περίπτωση (I). Διαφορετικά, ακολουθεί ο έλεγχος της υπόθεσης $H_0: \beta_2 = 0$ έναντι της $H_1: \beta_2 \neq 0$. Απόρριψη της υπόθεσης αυτής ισοδυναμεί με

την περίπτωση (II), εκ' αποδοχή της με την περίπτωση (III).

$$\bar{R}^2 = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right] = 1 - \left[\frac{0.032 \cdot 13}{10}\right] = 0.958,$$

άρα $\bar{R}^2 = 95.8\%$

$SSE = SST - SSR = 349442 - 338736 \Rightarrow SSE = 11206$

Στο y versus x1: $R^2_{pred.} = 1 - \frac{PRESS}{SST} \Rightarrow$

$$R^2_{pred.} = 1 - \frac{255010}{349442} \Rightarrow R^2_{pred} = 27.13\%$$

Όπως υποδεικνύει και το διάγραμμα, για την περιγραφή των δεδομένων απαιτούνται δύο διαφορετικές ευθείες. Αυτό μπορεί να φανεί από τα ήδη υπάρχοντα δεδομένα, αλλά προκύπτει και από τους ελέγχους που πραγματοποιήσαμε:

$H_0: \beta_3 = 0$, $F = \frac{SSE_2 - SSE_3}{SSE_3/(n-p)} = \frac{38358 - 11206}{11206/10} \Rightarrow$

$\Rightarrow F = 24.229$ με p-value $= 0.0006$, οπότε η H_0 απορρίπτεται και άρα 2 ευθείες διαφορετικής κλίσης απαιτούνται.

ΖΗΤΗΜΑ 2

A] $R^2 = \frac{SSR}{SST} = \frac{3.53751 \cdot 10^{14}}{3.6379 \cdot 10^{14}} = 97,24\%$

Φαίνεται οι X_5 και X_1 να μην έχουν καλή προσαρμογή βάσει του F-test. Το κατά πόσο είναι απαραίτητες για το μοντέλο απαιτεί περισσότερη διερεύνηση.

B] Βάσει του κριτηρίου C_p διαλέγουμε όσο το δυνατόν μικρό C_p και αντίστοιχα S και AIC. Αντίθετα, οι τιμές των R^2 και R^2_{pred} επιθυμούμε να είναι όσο το δυνατό υψηλότερες. Βάσει αυτών, 2 μοντέλα που σίγουρα ξεχωρίζουν είναι τα M5 και M7.

Ισχύει $SSE = S^2(n-k-1)$, άρα

$SSE_5 = 1.0558 \cdot 10^{13}$ και $SSE_7 = 1.0063 \cdot 10^{13}$

Τα μοντέλα διαφέρουν κατά 1 μεταβλητή, άρα ισχύει

$$F = \frac{SSE_5 - SSE_7}{SSE_7/(n-p)} = \frac{(1.0558 - 1.0063) \cdot 10^{13}}{1.0063 \cdot 10^{13}/46} = 2.26,$$

με p-value = 0.139, το οποίο είναι αρκετά υψηλό, άρα επιλέγουμε το μοντέλο M5, βάσει του κριτηρίου F.

Σε ό,τι αφορά το Μοντέλο 3, το Δ.Ε για τη διμεταβλητή πρόβλεψη είναι το $[\hat{\mu} - \underbrace{t_{0.025,48} \cdot S \cdot (\dot{x}'(x'x)^{-1}\dot{x} + 1)^{1/2}}_{\delta}, \dots]$

Άρα $110976 - \delta = -920677 \Rightarrow \delta = 1031653$

Άρα το άλλο άκρο είναι το $110976 + \delta = 1142629$

ΖΗΤΗΜΑ 3

$AIC_0 = -2\hat{l}_0 + 2 \cdot 1 \Rightarrow AIC_0 = 3673.56$

$D_0 - D_1 = 0.93 \Rightarrow -2(\hat{l}_0 - \hat{l}_1) = 0.93 \Rightarrow \hat{l}_1 = \hat{l}_0 - \frac{0.93}{2} \Rightarrow$

$\Rightarrow \hat{l}_1 = -1836.245$

$AIC_1 = -2\hat{l}_1 + 2 \cdot 2 \Rightarrow AIC_1 = 3676.49$

$D_2 - D_3 = 920.1$, pval $< 0.001 \rightarrow$ απόρριψη M_2

$R_D^2 = 1 - D(\hat{\beta}) / D_0$

Για το μοντέλο 2: $R_D^2 = 1 - 3507.3 / 3510.73 \Rightarrow$

$$R_D^2 = 0.09\%$$

Confint: $z_{\alpha/2} = 1.96$, άρα είναι το

$[5.889\,10^{-3} - 1.96 \cdot 1.434 \cdot 10^{-3},\ 5.889 \cdot 10^{-3} + 1.96 \cdot 1.434 \cdot 10^{-3}] =$

$[3.07836 \cdot 10^{-3},\ 8.69964 \cdot 10^{-7}]$ για τον $\hat{\beta}_2$, άρα

$1.00308 < e^{\hat{\beta}_2} < 1.0087$, με μέση τιμή

$e^{\hat{\beta}_2} = 1.0059$

Ερμηνεία: Εάν η απόσταση από το πιο κοντινό νηί αυξηθεί κατά 1 μονάδα, τότε ο αριθμός αδών φυτών αυξάνεται κατά 0.59% (αφού πολ/ται επί 1.0059)

ΖΗΤΗΜΑ 4

$\ln \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_1^i + \dots + \beta_k x_k^i$, για τον i-οστό παρατηρητή

x_1: μηνιαίο υπόλοιπο x_2: εισόδημα $x_3 = 1$: φοιτητής

(i) Ισχύει $AIC_2 = -2\hat{\ell}_2 + 2\cdot 3 \Rightarrow 1178.8 - 6 = -2\hat{\ell}_2 \Rightarrow \hat{\ell}_2 = -586.4$

$AIC_3 = -2\hat{\ell}_3 + 2\cdot 2 \Rightarrow AIC_3 = 2258.084$

$D_2 - D_1 = 1.6$, p-value = 0.206 μικρη διαφορα deviance, οχεσια υψηλό p-value → το M_1 απορρίπτεται

$D_3 - D_2 = -2(\hat{\ell}_3 - \hat{\ell}_2) = 1081.284$, p-value < 0.001, όρα λόγω της πολύ υψηλής διαφοράς των Deviance και του πολύ μικρού p-value, το M_3 απορρίπτεται

Δεχόμαστε, άρα, το M_2, μοντέλο που έρχεται σε συμφωνία και με το κριτήριο AIC αλλά και με τους ελέγχους Wald (για όλα τα β ισχύει p-value < 0.001).

(ii) $z_{\alpha/2} = 1.96$, ισχύει $\beta_3 \in [\hat{\beta}_3 - z_{\alpha/2}\, se(\hat{\beta}_3),\ \hat{\beta}_3 + z_{\alpha/2}\, se(\hat{\beta}_3)]$

$= [-0.6814 - 1.96\cdot 0.17,\ -0.6814 + 1.96\cdot 0.17] = [-1.0146,\ -0.3482]$

Άρα $e^{\beta_3} \in [0.3625,\ 0.706]$

(iii) Δεδομένου ότι $e^{\hat{\beta}_2} = 0.506$, φαίνεται πως ένας φοιτητής έχει 49.4% μειωμένη πιθανότητα να αποπληρώσει την πιστωτική σε σχέση με κάποιον που δεν είναι φοιτητής.

(iv) Οι καμπύλες ROC αποτελούν ένα δείκτη καλής προσαρμογής μιας δυαδικής λογιστικής παλινδρόμησης και βασίζεται στο λόγο της ευαισθησίας προς την ειδικότητα του μοντέλου. Συγκεκριμένα όσο μεγαλύτερο είναι το AUC (area under the curve) της ROC, τόσο μεγαλύτερη είναι η προβλεπτική ικανότητα του μοντέλου.

Από τις δοθείσες καμπύλες φαίνεται ξεκάθαρα ότι η προβλεπτική ικανότητα του Μοντέλου 3 δεν είναι καθόλου καλή (το AUC του τείνει εγγύς στο 0,5, δηλαδή αυτό του random classifier), ενώ αντίθετα είναι πολύ καλή για τα Μοντέλα 1 και 2, αφού το AUC τείνει στη μονάδα.

⊕ αναφορά στο C.I. του προηγούμενου ερωτήματος
δηλ. κυμαίνεται από 30% έως 61% περίπου