$df_{SST}+1=26+1$ ή από μέτρηση σημείων στο διάγραμμα.

ΖΗΤΗΜΑ 1

$E(y_x)=b_0+b_1x_1+b_2x_2+b_3x_3$, $n=27$ παρατηρήσεις , $x_2=\begin{cases}1, & \text{αν διαδικασία}=1\\ 0, & \text{αν διαδικασία}=2\end{cases}$

y: ποσότητα μετάλλου , x_1: ταχύτητα παραγωγής , $x_3=x_1 \cdot x_2$

Στα αποτελέσματα του regression analysis για το Μοντέλο 1 : y vs x_1, x_2, x_3
φαίνονται οι έλεγχοι t που πραγματοποιούνται για τις 3 συμμεταβλητές :

x_j μεταβλητή : $t^*=\dfrac{\hat{b}_j-b_j}{se(b_j)}$

$k=3$ μεταβλητές $\Rightarrow p=k+1=4$ και $n-p=27-4=23$

Έχοντας $H_0: b_j=0$, $H_1: b_j\neq 0$ ως στατιστικό έλεγχο υποθέσεων.

$$t^*=\frac{\hat{b}_j-b_j^{\,\to 0}}{se(\hat{b}_j)}=\frac{\hat{b}_j}{se(\hat{b}_j)}\sim t_{n-p}=t_{23}$$

άρα για x_2 : $t^*=\dfrac{\hat{b}_2}{se(\hat{b}_2)}$ ($\hat{b}_2\to$ coef, $se(\hat{b}_2)\to$ SE coef) δίνεται $t^*=3.19$

άρα p-value $\simeq 0.004$

Για x_3: $t^*=\dfrac{\hat{b}_3}{se(\hat{b}_3)}=\dfrac{-0.1767}{0.1288}\simeq -1.37\sim t_{23}$

p-value $\simeq 0.183$

$$\bar{R}^2=R\text{-}Sq(adj)=1-\frac{n-1}{n-p}(1-R^2)=1-\frac{26}{23}(1-0.945)=0.9378\simeq 93.78\%$$

Για την ANOVA του μοντέλου 1.

Έχουμε $SST=SSR+SSE \Rightarrow SSR=SST-SSE=179,069-9,904\Rightarrow$
$SSR=169,165$

Για το Μοντέλο 2 : y vs x_1, x_2. , $n=27$, $k=2$ μεταβλητές, $p=k+1=3$, $n-p=27-3=24$

$$\bar{R}^2=1-\frac{n-1}{n-p}(1-R^2)\Rightarrow 1-R^2=(1-\bar{R}^2)\frac{n-p}{n-1}\Rightarrow R^2=1-\frac{n-p}{n-1}(1-\bar{R}^2)=$$

$$=1-\frac{24}{26}(1-0.935)=0.94\simeq 94\%$$

Scanned with CamScanner

Στην ανάλυση ANOVA έχουμε ότι:

$F^* = \frac{MSR}{MSE} = 188.57$ και είναι έλεγχος F που αφορά τον στατιστικά έλεγχο υπόθεσεων.

$H_0: \beta_0 = \beta_1 = \beta_2 = 0$ H_1: τουλάχιστον ένας συντελεστής $\beta_i \neq 0$, $i = 0, 1, 2$.

Τότε: $F^* = 189.57 \sim F_{d, n-p} = F_{2, 24}$

(d: αριθμός ανεξάρτητων μεταβλητών μοντέλου)

άρα p-value $= 2.10 \cdot 10^{-15}$

Αλλά για τη μεταβλητή x_2: t-test με $t^* = \frac{\hat{\beta}_2}{se(\beta_2)} = \frac{53.129}{8.210} = 6.47$

και $t^* = 6.47 \sim t_{n-p} = t_{24}$ $\rightarrow$ p-value $= 1.08 \cdot 10^{-6}$

Για το μοντέλο 3: y vs x_1: $n = 27$, $k = 1$, $p = k+1 = 2$, $n-p = 25$

$R^2_{pred} = 1 - \frac{PRESS}{SST}$ (1)

Όμως $R^2 = \frac{SSR}{SST} \Rightarrow SST = \frac{SSR}{R^2} = \frac{149,661}{0.829} = 180,532$

απ' όπου $SSE = SST - SSR = 180,532 - 149,661 = 30,871$

και από (1): $R^2_{pred} = 1 - \frac{PRESS}{SST} = 1 - \frac{34,546.9}{180,532} \simeq 0.8086 = 80.86\%$

$MSE = \frac{SSE}{n-p} = \frac{30,871}{25} = 1,234.84$

και $F^* = \frac{MSR}{MSE} = \frac{149,661}{1,234.84} = 121.20 \sim F_{d, n-p} = F_{1, 25}$

(d: αριθμός ανεξάρτητων μεταβλητών μοντέλου)

άρα p-value $= 4.15 \cdot 10^{-11}$

Μέσω του μοντέλου $E(y_x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ μπορούμε να ελέγξουμε αν χρείζεται να προσαρμοστούν δύο διαφορ. ευθείες, δύο παράλληλες ή μια κοινή για τις δύο παραμετρικές διακλάδωσεις ($x_2 = 0$, $x_2 = 1$).

Scanned with CamScanner

Αυτό γιατί

- Αν $\beta_3 \neq 0$:

  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$

  - Όταν $x_2 = 0$: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$
  - Όταν $x_2 = 1$: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 + \hat{\beta}_3 x_1 = (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3) x_1$

  Δηλαδή δύο γραμμικές σχέσεις.

- Αν $\beta_3 = 0$
  - Αν $\beta_1 = 0$: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2$ 1 κοινή ευθεία για τις 2 διαφορετικές
  - Αν $\beta_1 \neq 0$:
    - Όταν $x_2 = 0$: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$
    - Όταν $x_2 = 1$: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 = (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 x_1$

    Δηλαδή δύο παράλληλες ευθείες.

Το μοντέλο 3 βάζουμε ότι έχει αρκετά χαμηλότερους βέλτιστους προσδιορισμούς σε σχέση με τα 2 πρώτα μοντέλα που φάνηκαν να έχουν καλύτερα τα δεδομένα μας.

Θα συγκρίνουμε τα 2 μοντέλα (1 & 2) μέσω ενός στατιστικού ελέγχου:

$H_0: \beta_3 = 0$: $M_2 : y \sim x_1, x_2$, $H_1: \beta_3 \neq 0$: $M_1 : y \sim x_1, x_2, x_3$

$$F^* = \frac{(SSE_{M_2} - SSE_{M_1})/1}{SSE_{M_1}/(97-3-1)} = \frac{10714 - 9904}{9904/23} = 1.88 \sim F_{d,23} = F_{1,23}$$

(d: διαφορά παραμέτρων για τα 2 μοντέλα)

αρα p-value = 0.184 (πχ για επίπεδο σημαντικότητας 5%)

Φαίνεται, λοιπόν, ότι ο έλεγχος είναι στατιστικά μη σημαντικός που σημαίνει ότι δεν μπορούμε να απορρίψουμε την H_0, δηλαδή το μοντέλο με x_1, x_2 (M2) το οποίο είναι και το μοντέλο που θα επιλέγαμε ως βέλτιστο δηλαδή $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

δηλ. 2 παράλληλες ευθείες θα ήταν η προσαρμογή με βάση τα παραπάνω.

Ωστόσο, είναι σαν σκεφτόμαστε ένα σήμα που φαίνεται να εξετάζει την

Scanned with CamScanner

προσαρμογή του μοντέλου 1 (δύο μη παράλληλες ευθείες). Παρ' όλα αυτά βλέπουμε ότι και αυτό δεν απέχει πολύ από το προηγούμενο σενάριο καθώς οι ευθείες δείχνουν να έχουν παρόμοια κλίση.

Η ερμηνεία που δίνουμε στο διάγραμμα είναι ότι όταν για δεδομένη τιμή ταχύτητας παραγωγής, μεταβαίνουμε από τη διαδικασία 2 ($x_2=0$) στη διαδικασία 1 ($x_2=1$) τότε έχουμε αύξηση στην ποσότητα μετάλλου. Παράλληλα με βάση το διάγραμμα ο ρυθμός αύξησης της ποσότητας μετάλλου για τη διαδικασία 1 ($x_2=1$) είναι μικρότερος (κλίση ευθείας) από τον αντίστοιχο για τη διαδικασία 2 ($x_2=0$).

Για το Μοντέλο 2 που βρήκαμε ως βέλτιστο θα ίσχυαν τα ίδια, όχι όμως ότι οι ρυθμοί αύξησης της ποσότητας μετάλλου είναι διαφορετικοί.

ΖΗΤΗΜΑ 4

$f(y)=\frac{e^{-\mu_x}\cdot \mu_x^{y}}{y!}$, $y=0,1,2.$ $\quad \eta(x)=g(\mu_x)=\ln\mu_x=\beta' x \Rightarrow \mu_x=e^{\beta' x}$

$D_M(\hat{\beta})=-2(\hat{l}_M-\hat{l}_{\kappa o \rho})=2\sum_{i=1}^{n}\left[y_i \ln\frac{y_i}{\hat{\mu}_i}\right]$, $\quad AIC=-2\hat{l}_M+2d$

$n=72$, $Y \sim X_1$ και X_2 (Y: αριθμός ασθενών, X_1: δοσολογία φαρμάκου, X_2: φύλο), $\quad x_2=\begin{cases}0, & \text{άντρας}\\ 1, & \text{γυναίκας}\end{cases}$

Για τις p-values όσον αφορά τους ελέγχους Wald έχουμε:

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

$z_j=\frac{\hat{\beta}_j}{se(\hat{\beta}_j)}\sim N(0,1)$

άρα για Μοντέλο 1: $z_1=-14.835 \rightarrow$ p-value $= 2\cdot$ pnorm(abs(-14.835), mean=0, sd=1, lower.tail=FALSE) (εντολή R.)

$z_2=\frac{\hat{\beta}_2}{se(\hat{\beta}_2)}=\frac{-0.013193}{0.0629}=-0.210 \rightarrow$ p-value $=0.834$.

Ομοίως για Μοντέλο 1:

$$z_1 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{-0.12962}{0.00873} = -14.848 \rightarrow \text{p-value} = 7.17 \cdot 10^{-50}$$

για Μοντέλο 0 ισχύει:

$$D_0(\hat{\beta}) = -2(\hat{\ell}_0 - \hat{\ell}_{κορ}) \Rightarrow \hat{\ell}_{κορ} = \frac{D_0(\hat{\beta})}{2} + \hat{\ell}_0 = \frac{272.305}{2} + (-222.219)$$

$$\Rightarrow \hat{\ell}_{κορ} = -85.567$$

Για Μοντέλο 1: $D_1(\hat{\beta}) = -2(\hat{\ell}_1 - \hat{\ell}_{κορ}) \Rightarrow \hat{\ell}_1 = \hat{\ell}_{κορ} - \frac{D_1(\hat{\beta})}{2} =$

$$= -85.567 - \frac{67.695}{2} \Rightarrow \hat{\ell}_1 = -120.415$$

$$AIC_1 = -2\hat{\ell}_1 + 2 \cdot d \quad (d: \text{όλοι οι παράμετροι}) = -2 \cdot (-120.415) + 2 \cdot 2 = 244.83$$

$$AIC_0 = -2\hat{\ell}_0 + 2 \cdot d = -2 \cdot (-222.219) + 2 \cdot 1 = 446.438$$

Το M_1 σε σχέση με M_0 έχουν 1 μεταβλητή άρα και παράμετρο διαφορά, δηλ. η διαφορά σε β.ε είναι 1.

$$D_0(\hat{\beta}) - D_1(\hat{\beta}) = 272.305 - 67.695 = 204.61$$

$H_0: M_0$, $H_1: M_1$

$D_0(\hat{\beta}) - D_1(\hat{\beta}) \sim \chi^2_q = \chi^2_1$ (q: διαφορά σε παραμέτρους 2 μοντέλων)

άρα για $204.61 \sim \chi^2_1 \rightarrow$ p-value $= 2.06 \cdot 10^{-46}$.

άρα βλ. σημαντικός έλεγχος και απορρίπτω $H_0 \Rightarrow$ επιλέγω M_1 over M_0

Ομοίως $D_1(\hat{\beta}) - D_2(\hat{\beta}) = 0.043 \sim \chi^2_1 \rightarrow$ p-value $= 0.836$

μη βλ. σημαντικός έλεγχος και δεν απορρίπτω $H_0 \Rightarrow$ επιλέγω M_1 over M_2

Μοντέλο 1:

$$R^2_D = \left(1 - \frac{D_1(\hat{\beta})}{D_0}\right) \cdot 100\% = \left(1 - \frac{67.695}{272.305}\right) \cdot 100\% = 75.14\%$$

Scanned with CamScanner

(ii)

Στο M_2 με βάση έλεγχο Wald η x_2 ήταν μη στατιστ. σημαντική επίσης M_1 έχει το χαμηλότερο AIC και deviance ελάχιστα διαφορετικό από M_2 ενώ ο χ^2 έλεγχος της διαφοράς των deviance μεταξύ M_1 και M_2 έδειξε ότι επιλέγω το M_1 ως το βέλτιστο μοντέλο. Επίσης και ο δείκτης ψευδο R^2_D - deviance έδειξε μεγάλη αύξηση περιγράφοντας από το M_0 στο M_1 και ελάχιστη αύξηση από το M_1 στο M_2.

Άρα: καλύτερο μοντέλο:

M_1: $E(y) = \mu_x = e^{\beta' x} = e^{\beta_0 + \beta_1 \cdot x_1}$

και το προσαρμοσμένο μοντέλο: $\hat{\mu}_x = e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot x_1} = e^{0.376777 - 0.12962 \cdot x_1}$.

δηλαδή προέκυψε ανεξαρτησία από το φύλο (x_2 μεταβλητή).

(iii) Θέλουμε ένα 0.95-ΔΕ για e^{β_1} του M_1.

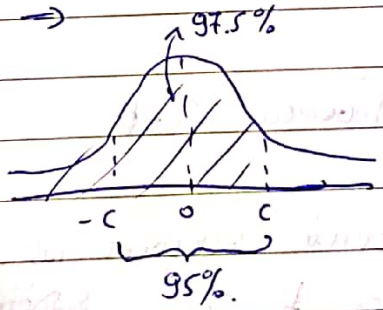
$z_1 = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \sim N(0,1)$ ασυμπτωτικά

Άρα βαθμό c τέτοια ώστε: $P[-c \le z_3 \le c] = 0.95 \rightarrow c = Z^{-1}(0.975)$ και ΔΕ:

$$-c \le \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \le c \rightarrow -c \cdot se(\hat{\beta}_1) \le \hat{\beta}_1 - \beta_1 \le c \cdot se(\hat{\beta}_1) \rightarrow$$

$$\rightarrow \hat{\beta}_1 - c \cdot se(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + c \cdot se(\hat{\beta}_1) \text{ και}$$

$$e^{\hat{\beta}_1 - c \cdot se(\hat{\beta}_1)} \le e^{\beta_1} \le e^{\hat{\beta}_1 + c \cdot se(\hat{\beta}_1)}$$



$c \simeq 1.96$ από qnorm(0.975) (R εντολή)

$\hat{\beta}_1 = -0.12962$, $se(\hat{\beta}_1) = 0.00873$ άρα

$\hat{\beta}_1 - c \cdot se(\hat{\beta}_1) = -0.1467$, $\hat{\beta}_1 + c \cdot se(\hat{\beta}_1) = -0.1125$

και $e^{\hat{\beta}_1 - c \cdot se(\hat{\beta}_1)} = e^{-0.1467} = 0.864$, $e^{\hat{\beta}_1 + c \cdot se(\hat{\beta}_1)} = e^{-0.1125} = 0.894$.

άρα 0.95 ΔΕ για e^{β_1}: $0.864 \le e^{\beta_1} \le 0.894$.

$e^{\hat{\beta}_1} = e^{-0.12962} = 0.878$

Παρατηρούμε ότι η τιμή του $e^{\hat{\beta}_1}$ είναι εντός του 95% ΔΕ του e^{β_1} όπως αναμέναμε.

Scanned with CamScanner

Μ1: $\hat{\mu}_x = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1}$

Για αύξηση κατά 1 μονάδα της x_1, δηλ. $x_1' = x_1 + 1$:

$$\hat{\mu}_{x'} = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1'} = e^{\hat{\beta}_0 + \hat{\beta}_1 (x_1+1)} = e^{\hat{\beta}_1} \cdot e^{\hat{\beta}_0 + \hat{\beta}_1 x_1} = e^{\hat{\beta}_1} \cdot \hat{\mu}_x$$

δηλ. πολ/σιάζει η τιμή του $\hat{\mu}_x$ με $e^{\hat{\beta}_1} = 0.878$ ή ισοδύναμα μειώθηκε κατά $1 - 0.878 = 0.122 = 12.2\%$

άρα αντίστοιχα η δοσολογία φαρμάκου κατά 1 μονάδα προκύπτει μείωση 12.2% στον αριθμό ασθενών με θετική απάντηση.

(iv) Για τα υπόλοιπα deviance παρατηρούμε ότι γενικώς κινούνται όπως θα περιμέναμε, δηλ πάνω στην ευθεία γραμμή με εξαίρεση τα σημεία 30 και 35 που ίσως είναι σημεία επιρροής και κάποια σημεία κατά στη μέση του άξονα x των normal quantiles. [illegible] Να σημειωθεί επίσης ότι τα υπόλοιπα deviance δεν κατανέμονται σύμφωνα με την κανονική κατανομή και χρησιμεύουν στον εντοπισμό άτυπων ή απόμακρων παρατηρήσεων.

Απόσταση Cook:

Γενικά θεωρούμε ως σημεία επιρροής εκείνα με Cook Distance μεγαλύτερο του 1. Στο δοθέν διάγραμμα, φαίνεται ότι η παρατήρηση 30 να έχει σχετικά υψηλή Cook Distance αλλά [illegible] δεν μπορεί να χαρακτηριστεί σημείο επιρροής αφού $D_{30} < 1$ και κοντά στο 0.5. Τελικά, δεν μπορούμε να χαρακτηρίσουμε κάποιο σημείο ως σημείο επιρροής με βάση το Cook Distance και χρειάζεται ίσως κάποιο άλλο κριτήριο εύρεσης σημείων επιρροής για να εξετάσουμε περαιτέρω το σημείο 30 καθώς και το σημείο 35.

Scanned with CamScanner