Θέματα 2020

Ζήτημα 1

(1Α) ΓΓΜ: $E(y)=X\beta$

$H_0: \beta_1=\beta_2=\cdots=\beta_k=0$, H_1: τουλάχιστον ένα $\beta_j \neq 0$.

SSE_0 $\qquad$ SSE_1.

$$F=\frac{(SSE_0-SSE_1)/k}{SSE_1/(n-k-1)} \sim F_{k,n-k-1}$$

~~άρα για νδο:~~ $F=\dfrac{R^2/k}{(1-R^2)/(n-k-1)}$ ~~αρκεί νδο~~ $\dfrac{R_1^2}{1-R_1^2}=\dfrac{SSE_0-SSE_1}{SSE_1}$ (1)

~~$SSR=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$~~ για το ~~$H_0: E(y)=\beta_0$ → από το fit $\hat{\beta}_0=\bar{y}$~~

$\hat{y}_i=\hat{\beta}_0$. άρα $SSR_0=\sum(\hat{\beta}_0-\hat{\beta}_0)^2=0$. → $R_0^2=\frac{SSR}{SST}=0$.

$SSE_0=SST_0-\overset{=0}{SSR_0}=SST$ $\qquad$ $\boxed{SST=SST_0=SST_1}$ → εξαρτάται μόνο από δεδομένα

$SSE_1=SST_1-SSR_1=SST-SSR_1$

~~άρα~~ ~~(1)~~ $\dfrac{SSE_0-SSE_1}{SSE_1}=\dfrac{\cancel{SST}-\cancel{SST}+SSR_1}{SST-SSR_1}=\dfrac{\frac{SSR_1}{SST}}{1-\frac{SSR_1}{SST}}=\dfrac{R_1^2}{1-R_1^2}$ ✓

(1Β) $E(Y)=\beta_0+\beta_1X_1$

• H_0: M_0: $E(Y)=\beta_0+\beta_1X_1$ $\qquad$ H_1: M_1: $E(Y)=\beta_0+\beta_1X_1+\beta_2X_2$

~~έλεγχος~~ $F: \dfrac{(SSE_0-SSE_1)/1}{SSE_1/(n-2-1)} \sim F_{1,n-2-1}=F_{1,n-3}$

μεταβλητές M_1

Αν το p-value του F-test προκύψει στατιστικά σημαντικό (πχ. <0.05) τότε απορρίπτω την H_0 (M_0) και επιλέγω H_1 (μοντέλο M_1).

- Βρίσκω R^2, $\bar{R}^2$ (adjusted), R^2 prediction
- Ελέγχω τιμές AIC
- Ελέγχω τιμές C_p

(• Ίσως σύγκριση υπολοίπων)

$\boxed{\text{ΖΗΤΗΜΑ 2}}$

$n = 21$, $Y \sim X_1, X_2$

(i) $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

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

Δίνονται $S = 0.172$, $r_{x_1 x_2} = 0.782$, $R^2 = 90.3\%$

Μεταβλ.	$\hat{\beta}$	$se(\hat{\beta})$	t	p-τιμή	VIF
Σταθερά	-0.752	0.273	-2.75	0.013	
X_1	0.035	0.007			
X_2	0.063	0.020			

$$VIF_j = \frac{1}{1 - R^2_{j \to R^2}}$$

j μεταβλητή ~~[illegible]~~ όπου η j μεταβλητή γίνεται regress σε σχέση με υπόλοιπες $x_j \sim x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$

$s = \sqrt{MSE} = \sqrt{\frac{SSE}{n-p}} \to SSE = (n-p) \cdot s^2$ όπου S γνωστό και $p = 3$ (αριθμός παραμέτρων)

$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \to SST = \frac{SSE}{1-R^2}$ όπου SSE γνωστό από πριν και R^2 ομοίως

$R^2_j = \frac{SSR_j}{SST \to \text{σταθερό}}$ άρα ψάχνω το $SSR_j \to$ SSE αφαιρώντας τη j μεταβλητή από μοντέλο

Βγάζοντας 1 μεταβλητή, προκύπτει απλό γραμμικό μοντέλο και ισχύει τότε:

~~$SSR_j = \hat{\beta}_i$~~ $SSR = \hat{\beta}_i \cdot S_{x_i x_i}$

μεταβλητή που μένει

Άρα βγάζοντας τη X_1 θέλω να υπολογίσω το $R^2_1 \neq \frac{SSR_1}{SST}$ και έχω:

$SSR_1 = \hat{\beta}_2 \cdot S_{x_2 x_2}$?

γνωστό

$$r_{x_1, x_2} = \frac{S_{x_1 x_2}}{\sqrt{S_{x_1 x_1} \cdot S_{x_2 x_2}}} = \frac{\sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\left(\sum (x_{1i} - \bar{x}_1) \cdot \sum (x_{2i} - \bar{x}_2)\right)^{1/2}} \quad (1)$$

από $se(\hat{\beta}_1) = \frac{s}{\sqrt{S_{x_1 x_1}}}$ και ομοίως $se(\hat{\beta}_2) = \frac{s}{\sqrt{S_{x_2 x_2}}}$ (2)

από (1) και (2) βρίσκω $S_{x_1 x_2}$ (και τα $S_{x_1 x_1}$, $S_{x_2 x_2}$)

Για το VIF: $x_2 \sim x_1$ δηλ. κάνει regress την x_2 με τις υπόλοιπες μεταβλητές (δηλ. x_1)

άρα $r^2_{x_1, x_2} = R^2$ αυτού του regression δηλ. $R^2_j = r^2_{x_1, x_2}$ και $VIF_{x_1} = VIF_{x_2} = \frac{1}{1 - r^2_{x_1, x_2}}$

$$= \frac{1}{1 - 0.782^2} \approx 2.574$$

(ii) $e_{21} = -0.287$, $h_{21,21} = 0.276$, $D_{21} = \dfrac{r_{21}^2 \, h_{21,21}}{p(1-h_{21,21})}$

Αν $D_{21} >> 1$ τότε θεωρείται σημείο επιρροής.

παράμετροι

$$r_i^2 = \frac{e_i^2 \rightarrow \text{γνωστό}}{s^2(1-h_{ii})}$$

γνωστό γνωστό για σημείο $i=21$

$$r_{21}^2 = \frac{e_{21}^2}{s^2(1-h_{21,21})} \simeq 3.846 \quad , \quad D_{21} = \frac{3.846 \cdot 0.276}{3\cdot(1-0.276)} \simeq 0.489.$$

μικρότερο του 1 άρα με βάση Cook Distance το σημείο $i=21$ δεν είναι σημείο επιρροής.