Θέμα

Α1) Επειδή οι τιμές των ε είναι αγνωστες, η εξέταση των υποθέσεων για τα σφαλματα γινεται υποχρεωτικα μεσω των υπολοίπων

- $e = y - \hat{y} = y - X\hat{\beta} = y - X(X'X)^{-1}X'y = y - Hy = (I-H)y$
$= (I-H)(X\beta+\varepsilon) = X\beta - HX\beta + \varepsilon - H\varepsilon \Rightarrow$

$$e = (I-H)\varepsilon$$

- $E(e) = E[(I-H)\varepsilon'] = (I-H)E(\varepsilon) = 0 \Rightarrow E(e) = 0$
- $V(e) = V[(I-H)\varepsilon] = (I-H)V(\varepsilon)(I-H)' = \sigma^2(I-H)$
- $Cov(e_i, e_j) = \sigma^2 h_{ij} \neq 0$

Α2) 1) $f(y_i) = \frac{1}{(2\Pi\sigma^2)^{1/2}} exp\left[-\frac{(y_i - \beta' x_i)^2}{2\sigma^2}\right] \Rightarrow$

$$f(y_i) = \frac{1}{(2\Pi\sigma^2)^{1/2}} exp\left[-\frac{(y_i - x_i'\beta)^2}{2\sigma^2}\right]$$

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\Pi\sigma^2}} exp\left[-\frac{(y_i - x_i'\beta)^2}{2\sigma^2}\right] \Rightarrow$$

$$L(\beta, \sigma^2) = L = \frac{1}{(2\Pi\sigma^2)^{n/2}} exp\left[-\frac{(y - \beta' x)^2}{2\sigma^2}\right] \Rightarrow$$

$$\ln L = -\frac{n}{2}\ln(2\Pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)$$

$$\frac{\partial \ln L}{\partial \beta} = -\frac{1}{\sigma^2}(X'X\beta - X'y) \Rightarrow \hat{\beta} = (X'X)^{-1}X'y$$

$$\hat{\sigma}^2 = \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{n} = \frac{SSE}{n}$$

Άρα η μεγιστοποιημένη συνάρτηση πιθανοφάνειας θα είναι

$$\hat{l} = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}^2) - \frac{n\hat{\sigma}^2}{2\hat{\sigma}^2} \Rightarrow$$

$$\hat{l} = -\frac{n}{2}\left[\ln(2\pi) + \ln\hat{\sigma}^2 + 1\right] \Rightarrow$$

$$\hat{l} = -\frac{n}{2}\left[\ln(2\pi\, SSE/n) + 1\right]$$

ii) Το AIC αποτελεί ένα κριτήριο επιλογής των βέλτιστων μοντέλων με το σύν το δυνατόν μικρότερο αριθμό παραμέτρων. Στην γενική περίπτωση ορίζεται από την σχέση

$$AIC = 2d - 2\ln L$$

όπου d το πλήθος των παραμέτρων του μοντέλου και L η μεγιστοποιημένη τιμή της συνάρτησης πιθανοφάνειας.

$$AIC = 2d - 2\ln L = 2d - 2\left(-n\ln\hat{\sigma} \cdot \frac{1}{2}n - \frac{n}{2}\ln(2\pi)\right)$$

$$= 2(P+1) + \left(2n\ln\hat{\sigma} + n + n\ln(2\pi)\right) \Rightarrow$$

$$AIC = 2(P+1) + n\left[\ln(2\pi) + 1 + \ln\left(\frac{SSE}{n}\right)\right]$$

A3) Οι έλεγχοι και τα κριτήρια που αφορούν την επιλογή των καταλληλότερων μεταβλητών για την ανάπτυξη του μοντέλου είναι τα μέτρα καταλληλότητας, οι στατιστικοί έλεγχοι και οι γραφικές παραστάσεις.

Για την αφαίρεση μια επεξηγηματικής μεταβλητές είναι αναγκαίο να γίνει ο έλεγχος F για την σημαντικότητα του συντελεστή της μεταβλητής αυτής. Έπειτα εξετάζονται και οι τιμές των R^2, $\bar{R}^2$ και AIC των μοντέλων με και χωρίς την μεταβλητή αυτή.

A4) Τα τυποποιημένα υπόλοιπα ορίζονται ως:

$$r_i = \frac{e_i}{s\sqrt{1-h_{ii}}} \qquad i = 1, 2, \dots, n$$

Η τυποποίηση των υπολοίπων αφαιρεί την ετεροσκεδαστικότητα των συνήθων υπολοίπων και έτσι μπορούμε να προβούμε σε ασφαλέστερες αξιολογήσεις για τις υποθέσεις των τυχαίων σφαλμάτων.

Τα deleted υπόλοιπα αποτελούν μια τροποποίηση των τυποποιημένων υπολοίπων σύμφωνα με την οποία χρησιμοποιείται μια εναλλακτική εκτίμητρια της σ^2, την οποία αποκτάμε αφού προσαρμοστεί το μοντέλο χρησιμοποιώντας όλες τις παρατηρήσεις πλην της i-στης.

$$r_i^* = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}} \qquad i = 1, 2, \dots, n$$

Τα διαγώνια στοιχεία h_{ii} ονομάζονται μοχλευση. Οι εν λόγω τιμές μετρούν την επίδραση της παρατηρούμενης τιμής της μεταβλητής αποκρισης πάνω στις προβλεπόμενες τιμές.

Οι τιμές των h_ii κυμαίνονται από 0 έως 1. Η αναμενόμενη τιμής τους είναι (p+1)/n. Αν υπήρχαν τιμές h_ii 3 φορές μεγαλύτερες από την μέση τιμή τότε έχαμε εκτρωπές τιμές στην i-γραμμή του πίνακα σχεδιασμού, οι οποίες πιθανότατα διαφοροποιούν έντονα την προσαρμογή του μοντέλου.

Οι τιμές μοχλεύσης χρησιμοποιούνται επίσης και για τον υπολογισμό των αποστάσεων Cook για να μετρούν την συνολική επιρροή της κάθε παρατήρησης στο μοντέλο. Τιμές των D_i άνω της μονάδας πρέπει να μας προβληματίσουν.

Α5) Ελέγχαμε πρώτα την υπόθεση της απουσίας αλληλεπιδράσεων μεταξύ της μεταβλητής, που κατασκευάσαμε, δηλαδή την ιδιότητα των κλίσεων

$H_0: B_3 = 0$ [κατάσταση Β] έναντι $H_1: B_3 \neq 0$ [κατάσταση Α]

Με τον έλεγχο F:

$$F = \frac{(SSE_B - SSE_A)/1}{SSE_A/(n-4)}$$

Στην περίπτωση που δεν απορριφθεί η μηδενική υπόθεση στο πρώτο βήμα, δηλαδή δεν διαπιστώσουμε καμία αλληλεπίδραση μεταξύ δυο μεταβλητών, προχωράμε στην εξέταση της τιμής του συντελεστή B_2.

$H_0: B_2 = 0$ [κατάσταση Γ] έναντι $H_1: B_2 \neq 0$ [κατάσταση Β]
(αφού αποφασίσαμε ότι $B_3 = 0$)

$$F = \frac{(SSE_\Gamma - SSE_B)/1}{SSE_B/(n-3)}$$

Στην περίπτωση που δεν απορρίψουμε την μηδενική υπόθεση μπορούμε να πούμε ότι τα δεδομένα προέρχονται από τον ίδιο ακριβώς πληθυσμό και μπορούν να θεωρηθούν ως μια ομάδα.