Multiple linear regression

P. Economou

Biomedical Engineering

2019–2020

Outline

Part A

- Multiple linear regression
  - Definition
  - Least Squares Estimation
  - Analysis of Variance
  - Coefficient of determination
  - Examples

Part B

- Statistical tests
  - t–test
  - F–test
- Prediction and Confidence Intervals
- Examples

Part C

- Regression Diagnostics
- Examples

Part D

- Model selection
  - Criteria to compare models
  - Stepwise regression
- Examples

Πολλαπλή Γραμμική Παλινδρόμηση

Συχνά συναντάμε προβλήματα, για τα οποία υπάρχουν υποψίες ή και ενδείξεις ότι οι τιμές κάποιας μεταβλητής που μας ενδιαφέρει επηρεάζεται από περισσότερες από μία επεξηγηματικές μεταβλητές.

Η ανάλυση τέτοιων προβλημάτων απαιτεί μια επέκταση του απλού γραμμικού μοντέλου, η οποία θα λαμβάνει υπόψη της την ταυτόχρονη ύπαρξη $k \geq 2$ ανεξάρτητων ή επεξηγηματικών μεταβλητών.

Το **γενικό γραμμικό μοντέλο** δίνεται από τη σχέση

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \,,$$

όπου

- y_i, $i = 1, 2, \ldots, n$, οι τιμές των παρατηρήσεων της εξαρτημένης μεταβλητής y
- x_{ij}, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, k$, οι τιμές για την i–οστή παρατήρηση των ανεξάρτητων ή επεξηγηματικών μεταβλητών x_j
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$, οι άγνωστες παράμετροι του μοντέλου και
- ε_i, $i = 1, 2, \ldots, n$ τα τυχαία σφάλματα

Τα τυχαία σφάλματα

Τα τυχαία σφάλματα υποθέτουμε ότι ικανοποιούν τις παρακάτω υποθέσεις (ανάλογες του απλού γραμμικού μοντέλου)

- $E(\varepsilon_i) = 0$, για κάθε i
- $V(\varepsilon_i) = \sigma^2$, για κάθε i, δηλαδή τα τυχαία σφάλματα ικανοποιούν την υπόθεση της ομοσκεδαστικότητας
- $cov(\varepsilon_i, \varepsilon_j) = 0$, για $i \neq j$, δηλαδή τα ε_i είναι ασυσχέτιστα μεταξύ τους.

Diagnostic tests

Πολλαπλή Γραμμική Παλινδρόμηση – Πίνακες

Η σχέση

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \,,$$

μπορεί να γραφτεί και υπό τη μορφή πινάκων ως εξής:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \,,$$

όπου

- $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$ είναι το $n \times 1$ διάνυσμα τιμών της μεταβλητής απόκρισης y,
- $\mathbf{X}$ ένας $n \times p$ πίνακας με $p = k + 1$,
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ το $p \times 1$ διάνυσμα των παραμέτρων και
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ το $n \times 1$ διάνυσμα των τυχαίων σφαλμάτων.

Ο πίνακας $\mathbf{X}$ ονομάζεται **πίνακας σχεδιασμού** και έχει την ακόλουθη μορφή

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix} .$$

Ιδιότητες

Υπό την επιπλέον υπόθεση της κανονικότητας των τυχαίων σφαλμάτων μπορούμε να γράψουμε ότι το n–διάστατο διάνυσμα ε ακολουθεί την πολυμεταβλητή ή πολυδιάστατη Κανονική κατανομή

$$\varepsilon \sim N_n\left(\mathbf{0}, \sigma^2 \mathbf{I}\right)$$

- $E(\varepsilon) =$
- $V(\varepsilon) =$
- $E(\boldsymbol{y}) = E(\boldsymbol{y}) = E(\mathbf{X}\boldsymbol{\beta} + \varepsilon) =$
- $V(\boldsymbol{y}) = V(\boldsymbol{y}) = V(\mathbf{X}\boldsymbol{\beta} + \varepsilon) =$

Μέθοδος ελαχίστων τετραγώνων

Η μέθοδος ελαχίστων τετραγώνων για την εκτίμηση των παραμέτρων β βασίζεται, όπως και στο απλό γραμμικό μοντέλο, στην ελαχιστοποίηση της παράστασης

$$S(\beta)\,(\mathbf{y} - E(\mathbf{y}))'\,(\mathbf{y} - E(\mathbf{y})) = (\mathbf{y} - \mathbf{X}\beta)'\,(\mathbf{y} - \mathbf{X}\beta)\ .$$

Οι **εκτιμήτριες ελαχίστων τετραγώνων** λαμβάνονται

- παραγωγίζοντας την $S(\beta)$ ως προς β και
- θέτοντας την παραπάνω σχέση ίση με το μηδέν

Από την επίλυση του συστήματος λαμβάνουμε τις εκτιμήτριες ελαχίστων τετραγώνων, οι οποίες δίνονται από τη σχέση

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y}\,.$$

Κατανομή Εκτιμητριών Ελαχίστων Τετραγώνων: $\hat{\beta} \sim N_p\left(\beta, \sigma^2\,(\mathbf{X}'\mathbf{X})^{-1}\right)$

Θεώρημα Gauss–Markov

Στο γενικό γραμμικό μοντέλο η εκτιμήτρια ελαχίστων τετραγώνων $\hat{\beta}$ είναι η αμερόληπτη γραμμική εκτιμήτρια του β με την ελάχιστη δυνατή διασπορά.

Εκτίμηση μοντέλου – Υπόλοιπα

- Το μοντέλο που εκτιμούμε δίνεται από τη σχέση

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{H}\mathbf{y}, \end{aligned}$$

  όπου $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$, ο **πίνακας προβολής** (καλείται και **hat matrix**, ο οποίος είναι συμμετρικός και ταυτοδύναμος.
- Τα **υπόλοιπα** (residuals) υπολογίζονται ως

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}.$$

Στατιστικός έλεγχος F – Συντελεστής Προσδιορισμού R^2

- Η **στατιστική συνάρτηση** F μπορεί να χρησιμοποιηθεί για τον έλεγχο της σημαντικότητας της παλινδρόμησης, δηλαδή της εξέτασης της υπόθεσης όλων των συντελεστών των μεταβλητών ταυτόχρονα

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

έναντι της

$$H_1 : \text{τουλάχιστον ένα } \beta_j \neq 0$$

Υπό την H_0 η στατιστική συνάρτηση F ακολουθεί την κατανομή $F_{k,(n-p)}$.

- Ο **συντελεστής προσδιορισμού** R^2 ορίζεται με τον ίδιο ακριβώς τρόπο και εκφράζει ακριβώς τα ίδια πράγματα με την περίπτωση του απλού γραμμικού μοντέλου. Δηλαδή, ο συντελεστής προσδιορισμού ορίζεται ως

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST},$$

όπου

$$SST = SSR + SSE.$$

Ο δείκτης προσδιορισμού ικανοποιεί τη σχέση $0 \leq R^2 \leq 1$.

Ανάλυση διασποράς

Πίνακας ανάλυσης διασποράς για ένα μοντέλο γραμμικής παλινδρόμησης με k επεξηγηματικές μεταβλητές

Πηγή μεταβλητότητας	Άθροισμα τετραγώνων	Βαθμοί ελευθερίας	Μέσο άθροισμα τετραγώνων	Έλεγχος F
Παλινδρόμηση (Regression)	$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	k	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$
Υπόλοιπα (Error)	$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$n - k - 1 = n - p$	$MSE = S^2 = \frac{SSE}{n-k-1} = \frac{SSE}{n-p}$	
Σύνολο (Total)	$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$n - 1$		

Χρήση Υπολογιστή

↪ Στο αρχείο STRENGTHCEMENTS.sav παρουσιάζονται δεδομένα για την αντοχή y διάφορων τύπων τσιμέντων σε σχέση με τις ποσότητες των διάφορων συστατικών τους (μεταβλητές x_1 έως x_4). Να προσαρμοστεί ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης, το οποίο να σχετίζει την αντοχή του τσιμέντου με τις ποσότητες των συστατικών του.

↪ Στο αρχείο CHEMICAL.sav παρουσιάζονται τα δεδομένα από ένα πείραμα για τη μελέτη της απόδοσης y μιας χημικής αντίδρασης συναρτήσει 7 ελεγχόμενων μεταβλητών x_1 έως x_7 της αντίδρασης.

(i) Να προσαρμοστεί ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης, που να συνδέει την y με τις 7 ελεγχόμενες μεταβλητές της αντίδρασης (x_1 έως x_7).

(ii) Να εκτελεστεί ο μετασχηματισμός $y \to \sqrt{y}$ και να προσαρμοστεί για τη νέα μεταβλητή ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης. Να συγκρίνετε τα δυο μοντέλα ως προς τη σημαντικότητά τους.

Part B

- Statistical tests
  - t–test
  - F–test
- Prediction and Confidence Intervals
- Examples

Έλεγχος t – Υποθέσεις

Με τον έλεγχο t μπορούμε να εξετάσουμε τις υποθέσεις

$$H_0 : \beta_j = \beta_{j(0)}, \quad (\text{αν } \beta_{j(0)} = 0 \text{ η } x_j \text{ δε συμβάλλει στο μοντέλο})$$
$$H_1 : \beta_j \neq \beta_{j(0)}$$

για τους συντελεστές β_j, $j = 1, 2, \ldots, k$, των μεταβλητών στη συνάρτηση παλινδρόμησης

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \, .$$

Έλεγχος t – Στατιστική συνάρτηση ελέγχου

Η στατιστική συνάρτηση ελέγχου δίνεται από την σχέση

$$t = \frac{\hat{\beta}_j - \beta_{j(0)}}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_{j(0)}}{S\sqrt{c_{jj}}},$$

η οποία ακολουθεί την κατανομή t_{n-k-1}, με n το πλήθος των παρατηρήσεων και k το πλήθος των επεξηγηματικών μεταβλητών

- c_{jj} το j–οστό διαγώνιο στοιχείο του πίνακα $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$.
- $S^2 = SSE/(n-k-1)$ η αμερόληπτη εκτιμήτρια της άγνωστης διασποράς σ^2 των τυχαίων σφαλμάτων.

Επομένως, με βάση την ελεγχοσυνάρτηση αυτή μπορούν να εξεταστούν για κάθε j οι παραπάνω υποθέσεις, καθώς και να κατασκευαστούν διαστήματα εμπιστοσύνης για τα β_j.

Έλεγχος t – διάστημα εμπιστοσύνης

Πιο συγκεκριμένα ένα $100(1-\alpha)\%$ **διάστημα εμπιστοσύνης** για το συντελεστή $\beta_j,\ j = 0, 1, \ldots, k$, του μοντέλου δίνεται ως

$$\left[\hat{\beta}_j - t_{a/2} S\sqrt{c_{jj}}\ ,\ \hat{\beta}_j + t_{a/2} S\sqrt{c_{jj}}\right],$$

όπου $t_{\alpha/2}$ είναι το άνω $100(\alpha/2)$ ποσοστιαίο σημείο της κατανομής t με $(n-k-1)$ βαθμούς ελευθερίας.

Έλεγχος t – Σχόλια

Οι έλεγχοι t εξετάζουν κάθε ένα συντελεστή χωριστά, αλλά υπό τη συνθήκη ότι οι υπόλοιπες $k-1$ μεταβλητές συμπεριλαμβάνονται στο μοντέλο. Αυτό έχει σαν συνέπεια τα αποτελέσματα των k ελέγχων να μην αποτελούν αξιόπιστο οδηγό για το ποιες μεταβλητές τελικά χρειάζονται στο μοντέλο.

Έλεγχος F

Με τον έλεγχο F μπορούμε να εξετάσουμε ταυτόχρονα τη στατιστική σημασία q μεταβλητών $x_1, \ldots, x_q$, δηλαδή να εξετάσουμε τις υποθέσεις

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_q = 0, \ \beta_i \ (i \neq 1, 2, \ldots, q) \text{ χωρίς περιορισμούς}$$

$$H_1 : \text{ όλα τα } \beta_i \text{ χωρίς περιορισμούς.}$$

Η στατιστική συνάρτηση ελέγχου δίνεται από την σχέση

$$\frac{(SSE_0 - SSE_1)/q}{SSE_1/(n-p)} \sim F_{q,(n-p)} .$$

όπου

- SSE_0 άθροισμα τετραγ. των υπολοίπων υπό την H_0
- SSE_1 άθροισμα τετραγ. των υπολοίπων υπό την H_1

▸ Stepwise

Έλεγχος F – Σχόλια

$\hookrightarrow$ Τονίζεται ότι ο έλεγχος F μπορεί να εφαρμοστεί μόνο στην περίπτωση που τα δύο μοντέλα που συγκρίνονται είναι εμφωλευμένα, δηλαδή όταν $M_0 \subset M_1$.

$\hookrightarrow$ Ο έλεγχος F της ανάλυσης διασποράς, που παρουσιάζεται στα αποτελέσματα της προσαρμογής ενός μοντέλου παλινδρόμησης ελέγχει την ειδική περίπτωση της υπόθεσης ότι όλοι οι συντελεστές είναι μηδέν

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

έναντι της

$$H_1 : \text{ τουλάχιστον ένα } \beta_j \neq 0 .$$

Διαστήματα εμπιστοσύνης και πρόβλεψης

Όπως και στο απλό γραμμικό μοντέλο μπορούμε να κατασκευάσουμε

- διαστήματα εμπιστοσύνης για την μέση τιμή του y και
- διαστήματα πρόβλεψης για μια νέα παρατήρηση του y

όταν οι επεξηγηματικές μεταβλητές πάρουν κάποιες συγκεκριμένες τιμές και
Ο τρόπος κατασκευής καθώς και η ερμηνεία των διαστημάτων αυτών είναι ανάλογη με το απλό γραμμικό μοντέλο.

Χρήση Υπολογιστή

$\hookrightarrow$ Να εφαρμόσετε τους ελέγχους t και F στα δεδομένα των αρχείων STRENGTHCEMENTS.sav και CHEMICAL.sav.

Part C

- Regression Diagnostics
- Examples

Τυχαία σφάλματα

Τα τυχαία σφάλματα ε παίζουν κεντρικό ρόλο στο γενικό γραμμικό μοντέλο και οφείλουν να ικανοποιούν τις ▸ συνθήκες που αναφέρθηκαν νωρίτερα.
Οι συνθήκες αυτές μπορούν να γραφούν συνοπτικά ως

$$\varepsilon \sim N_n\left(\mathbf{0}, \sigma^2 \mathbf{I}\right),$$

όπου με $\mathbf{I}$ συμβολίζουμε τον n–διάστατο μοναδιαίο πίνακα.

Υπόλοιπα (συνήθη)

Παραβιάσεις αυτών των προϋποθέσεων ενδέχεται να έχουν σοβαρές επιπτώσεις στην εκτίμηση του μοντέλου και κατά συνέπεια επιβάλλεται η εξέταση των ισχυρισμών αυτών. Η εξέταση των παραπάνω προϋποθέσεων πραγματοποιείται κυρίως μέσω γραφικών παραστάσεων.

Επειδή οι τιμές των ε είναι άγνωστες, η εξέταση των υποθέσεων για τα τυχαία σφάλματα γίνεται υποχρεωτικά μέσω των υπολοίπων, τα οποία υπολογίζονται από τη σχέση

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$$

Στη συνέχεια θα δούμε πώς μπορούν οι προϋποθέσεις για τα τυχαία σφάλματα να διατυπωθούν για τα υπόλοιπα $\mathbf{e}$.

Υπόλοιπα (συνήθη)

Μπορούμε να δείξουμε ότι τα υπόλοιπα $\mathbf{e}$ ακολουθούν την κατανομή

$$\mathbf{e} \sim N_n\left(\mathbf{0}, \sigma^2\left(\mathbf{I} - \mathbf{H}\right)\right)$$

και άρα

$$e_i \sim N\left(0, \sigma^2\left(1 - h_{ii}\right)\right), \quad i = 1, 2, \ldots, n\,.$$

Υπόλοιπα (συνήθη)

$\hookrightarrow V(e_i) = \sigma^2 (1 - h_{ii})$

$\hookrightarrow$ Η σ^2 των τυχαίων σφαλμάτων εκτιμάται γενικώς από την αμερόληπτη $S^2 = \frac{\mathbf{e}'\mathbf{e}}{n-p}$.

$\hookrightarrow$ Επειδή ο πίνακας $\mathbf{I} - \mathbf{H}$ δεν είναι διαγώνιος, τα υπόλοιπα e_i και e_j εκτός του ότι δεν έχουν την ίδια διασπορά, δεν είναι και ασυσχέτιστα μεταξύ τους, αφού

$$cov(e_i, e_j) = -\sigma^2 h_{ij} \neq 0, \quad i \neq j\,,$$

όπου $h_{ij} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_j = \mathbf{x}_j'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i = h_{ji}$, τα μη διαγώνια στοιχεία του πίνακα $\mathbf{H}$.

$\hookrightarrow$ Το κάθε υπόλοιπο μπορεί να εκφραστεί ως

$$e_i = y_i - \sum_{j=1}^{n} h_{ij} y_j = \varepsilon_i - \sum_{j=1}^{n} h_{ij}\, \varepsilon_j$$

και ως εκ τούτου είναι φανερό ότι κάθε υπόλοιπο e_i εξαρτάται όχι μόνο από το αντίστοιχο τυχαίο σφάλμα ε_i, αλλά και από τα άλλα σφάλματα ε_j.

Τυποποιημένα υπόλοιπα

Τα **τυποποιημένα υπόλοιπα** (standardized residuals ή internally studentized residuals) μέσω των σχέσεων

$$r_i = \frac{e_i}{S\sqrt{1-h_{ii}}}, \ i = 1, 2, \ldots, n\,.$$

Αυτή η τυποποίηση αφαιρεί την ετεροσκεδαστικότητα των συνήθων υπολοίπων, που αναφέραμε παραπάνω. Κατά συνέπεια μπορούμε να προβούμε σε ασφαλέστερες αξιολογήσεις για τις υποθέσεις των τυχαίων σφαλμάτων.

Να σημειώσουμε ότι η ορολογία μπορεί να διαφέρει μεταξύ συγγραφέων και στατιστικών πακέτων Η/Υ. Για παράδειγμα, το SPSS αποκαλεί τα παραπάνω r_i υπόλοιπα "studentized", ενώ ως "standardized" υπόλοιπα θεωρεί τις ποσότητες e_i/S.

Deleted υπόλοιπα

Τα **deleted υπόλοιπα** αποτελούν μια τροποποίηση των τυποποιημένων υπολοίπων, σύμφωνα με την οποία χρησιμοποιείται μια εναλλακτική εκτιμήτρια της σ^2, την οποία αποκτούμε, αφού προσαρμοστεί το μοντέλο χρησιμοποιώντας όλες τις παρατηρήσεις **πλην της i–οστής**. Πιο συγκεκριμένα, τα deleted υπόλοιπα ορίζονται ως

$$r_i' = \frac{e_i}{S_{(i)}\sqrt{1-h_{ii}}}, \quad i = 1, \ldots, n,$$

όπου $S^2_{(i)}$ είναι η εκτιμήτρια της σ^2 που προκύπτει από την προσαρμογή του μοντέλου στις $n-1$ παρατηρήσεις, αφού έχει εξαιρεθεί η i–οστή, αφαιρώντας έτσι την όποια σημαντική επίδραση ασκεί η παρατήρηση αυτή στη συνάρτηση παλινδρόμησης. Δηλαδή

$$S^2_{(i)} = \frac{\sum_{j\neq i}(y_j - \mathbf{x}_j'\hat{\boldsymbol{\beta}}_{(i)})^2}{n-1-p} = \frac{SSE_{(i)}}{n-1-p},$$

όπου $SSE_{(i)}$ το άθροισμα τετραγώνων λόγω σφάλματος που προκύπτει από την παλινδρόμηση χωρίς το σημείο i.
Γενικά, μπορούμε να πούμε ότι τα τυποποιημένα και τα deleted υπόλοιπα προτιμούνται των συνήθων. Τα deleted υπόλοιπα r_i' συχνά αναφέρονται και ως "externally studentized residuals".

Κατανομή των τυποποιημένων και των deleted υπολοίπων

$\hookrightarrow$ Ούτε τα r_i ούτε τα r_i' ακολουθούν την Κανονική κατανομή. Παρόλα αυτά η εξέταση των υπολοίπων αυτών πραγματοποιείται με γραφικές παραστάσεις καταλληλότητας της Κανονικής κατανομής και αυτό γιατί η εμπειρία πράγματι υποδεικνύει τη χρησιμότητα αυτής της διαδικασίας

$\hookrightarrow$ Επίσης, γνωρίζουμε ότι η ανεξαρτησία των ε_i δεν ισχύει για τα e_i καθώς και για τα άλλα υπόλοιπα που βασίζονται σε αυτά.

$\hookrightarrow$ Ένα συγκεκριμένο πρόβλημα που προκύπτει με τα τυποποιημένα υπόλοιπα r_i είναι ότι για μικρά δείγματα οι τιμές τους έχουν τον περιορισμό

$$|r_i| < \sqrt{n-p}\,.$$

Κατά συνέπεια ο οδηγός $|r_i| > 3$ που καμια φορά προτείνεται για τον εντοπισμό άτυπων τιμών, είναι άχρηστος για μικρά δείγματα, αφού η ανισότητα $|r_i| < 3$ είναι δεδομένη όταν $n - p < 10$.

Εξέταση υπολοίπων – Γραφικοί έλεγχοι υπολοίπων

1. **Γραφική εξέταση της Κανονικής κατανομής** η οποία χρησιμοποιείται και για τον εντοπισμό άτυπων (ή απόμακρων ή έκτροπων ή απομονωμένων) σημείων (outliers), που ενδέχεται να εμφανίζονται και να θέτουν ερωτήματα για την καταλληλότητα του μοντέλου. Αν η εικόνα δεν είναι ικανοποιητική, τότε ίσως χρειάζεται να γίνει κάποιος μετασχηματισμός π.χ. της y ή ίσως χρειάζονται επιπλέον ανεξάρτητες μεταβλητές στο μοντέλο. Ίσως πρέπει να αφαιρεθούν κάποιες παρατηρήσεις που μπορεί να προέκυψαν από άλλο πληθυσμό ή από λανθασμένες μετρήσεις.

2. **Γραφική παράσταση των e_i επί των $\hat{y}_i$**, που εξετάζει κυρίως τη σταθερότητα της διασποράς (ομοσκεδαστικότητα), καθώς επίσης και σημεία που ενδέχεται να ξεφεύγουν. Αν η εικόνα δεν είναι καλή, ίσως πάλι να χρειάζεται να γίνει κάποιος μετασχηματισμός, π.χ. από y σε $\log y$.

Εξέταση υπολοίπων – Γραφικοί έλεγχοι υπολοίπων

3 **Γραφικές παραστάσεις για τον έλεγχο της συσχέτισης** σε περιπτώσεις που τα δεδομένα έχουν μια ορισμένη χρονική ή άλλη σειρά:

1. διάγραμμα των υπολοίπων ως προς τη σειρά, χρόνο ή θέση των παρατηρήσεων και
2. διάγραμμα των e_i έναντι των e_{i-1}, για την εξέταση της συσχέτισης μεταξύ των υπολοίπων.

Συνήθως η εξάρτηση μεταξύ των υπολοίπων (λόγω της σχέσης $\mathbf{X}'\mathbf{e} = \mathbf{0}$) δεν έχει σημαντική επίπτωση στην εμφάνιση των γραφημάτων. Συνεπώς θεωρούμε ότι σημεία που κατανέμονται με τυχαιότητα σε αυτά τα γραφήματα δηλώνουν ανεξαρτησία μεταξύ των σφαλμάτων και άρα δεν παραβιάζεται η σχετική υπόθεση του μοντέλου. Σε διαφορετική περίπτωση, θα πρέπει η αυτοσυσχέτιση να ληφθεί υπόψη στο μοντέλο.

Εξέταση υπολοίπων – Γραφικοί έλεγχοι υπολοίπων

4 **Γραφική παράσταση των e_i επί των x_i** τιμών μιας επεξηγηματικής μεταβλητής x. Τυχαία διάσπαρτα σημεία μας ικανοποιούν ειδάλλως δεν ισχύει η υπόθεση της ανεξαρτησίας των ε_i και x_i. Σε αυτή την περίπτωση ίσως η μεταβλητή x να χρειάζεται κάποιο μετασχηματισμό.

Γενικώς, αν κάποιο γράφημα ή κάποια γραφήματα δεν ικανοποιούν τις απαιτούμενες συνθήκες, αυτό σημαίνει ότι το μοντέλο δεν περιγράφει επαρκώς τα δεδομένα.

Η υπόθεση $E(\varepsilon) = \mathbf{0}$ δεν εξετάζεται στις παραπάνω γραφικές παραστάσεις, επειδή η αντίστοιχη ιδιότητα $\bar{e} = \frac{\sum_{i=\mathbf{1}}^{n} e_i}{n} = 0$ ικανοποιείται αυτομάτως

Χρήση Υπολογιστή

↪ Μέσω διαγνωστικών ελέγχων να εξεταστεί η καταλληλότητα του απλού γραμμικού μοντέλου για τα 4 σύνολα δεδομένα που βρίσκονται στο αρχείο Anscombe.sav.

↪ Μέσω διαγνωστικών ελέγχων να εξεταστεί η καταλληλότητα του μοντέλου με όλες τις διαθέσιμες επεξηγηματικές μεταβλητές για τα δεδομένα του αρχείου STRENGTHCEMENTS.sav.

↪ Μέσω διαγνωστικών ελέγχων να εξεταστεί η καταλληλότητα του μοντέλου και με τις 7 διαθέσιμες επεξηγηματικές μεταβλητές για τα δεδομένα του αρχείου CHEMICAL.sav.

Part D

- Model selection
  - Criteria to compare models
  - Stepwise regression
- Examples

Μέτρα καταλληλότητας

συντελεστή προσδιορισμού

- Η πρόσθεση επιπλέον μεταβλητών σε ένα μοντέλο οδηγεί πάντα στην αύξηση του συντελεστή προσδιορισμού R^2

  Άρα μόνο μια αξιοσημείωτη αύξηση του συντελεστή προσδιορισμού από την πρόσθεση μιας μεταβλητής ή μια δραστική μείωσή του από την αφαίρεση μιας μεταβλητής μπορεί να εκληφθεί θετικά για τη σημαντικότητα της μεταβλητής αυτής.

Μέτρα καταλληλότητας

Διορθωμένος συντελεστή προσδιορισμού

- Ο **διορθωμένος** ή **τροποποιημένος συντελεστής προσδιορισμού** $\bar{R}^2$ (adjusted–R^2) ορίζεται ως

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

όπου $p = k + 1$ το πλήθος παραμέτρων του μοντέλου.

Ο διορθωμένος συντελεστής προσδιορισμού προτιμάται έναντι του συνηθισμένου συντελεστή προσδιορισμού R^2

- για τις περιπτώσεις που ο συντελεστής προσδιορισμού R^2 είναι ακατάλληλος ($n \simeq p$), αλλά και
- για την αρχική αξιολόγηση κατά την επιλογή του βέλτιστου μοντέλου. Αυτό που τον κάνει ελκυστικό είναι η ιδιότητα του να μην αυξάνεται η τιμή του όταν προστίθεται στο μοντέλο μια οποιαδήποτε επεξηγηματική μεταβλητή, αλλά όταν αυτή η μεταβλητή πράγματι βελτιώνει το μοντέλο.

Στατιστική συνάρτηση C_p–Mallows

Η στατιστική συνάρτηση C_p–Mallows βασίζεται στην ακρίβεια πρόβλεψης του μοντέλου και πιο συγκεκριμένα στο Μέσο Τετραγωνικό Σφάλμα, το οποίο ορίζεται γενικώς για την εκτιμήτρια $\hat{\theta}$ μιας παραμέτρου θ από τη σχέση

$$\begin{aligned} MT\Sigma(\hat{\theta}) &= E(\hat{\theta}-\theta)^2 = \ldots \\ &= V(\hat{\theta}) + \left[E(\hat{\theta}) - \theta\right]^2 \\ &= V(\hat{\theta}) + bias^2 \,. \end{aligned}$$

Η παραπάνω σχέση μας δίνει και το λόγο που στρέφουμε την προσοχή μας στο ΜΤΣ αφού συνδέεται με τη μεροληψία (bias) εκτιμητριών παραμέτρων ενός μοντέλου που δεν έχει οριστεί σωστά και δίνει μεροληπτικές εκτιμήτριες.

Στατιστική συνάρτηση C_p–Mallows

Αν η μεροληψία στο μοντέλο με τις p μεταβλητές είναι αμελητέα, τότε **ως βέλτιστο μοντέλο** θωρείται εκείνο για το οποίο ισχύει

$$C_p \simeq p\,.$$

Ειδικότερα, αν υπάρχουν περισσότερα από ένα μοντέλο με $C_p \simeq p$, προτιμότερο είναι εκείνο με το μικρότερο p, δηλαδή αυτό που περιέχει λιγότερες μεταβλητές.

Akaike's information criterion (AIC)
Bayesian information criterion (BIC)

Τα *AIC* και *BIC* αποτελούν κριτήρια επιλογής του βέλτιστου μοντέλου με το όσο το δυνατόν μικρότερο αριθμό παραμέτρων. Στη γενική περίπτωση ορίζονται από τις σχέσεις

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

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

Το προτιμητέο μοντέλο με βάση αυτό τα κριτήρια αυτά είναι εκείνο με το μικρότερο *AIC* **ή/και** *BIC*.

Η πρόσθεση επιπλέον μεταβλητών (στατιστικά σημαντικών ή όχι) αυξάνει το $\ln L$, άρα ο δεύτερος όρος των *AIC* και *BIC* μειώνεται. Από την άλλη, αυξάνονται οι πρώτοι όροι.

Τελικά η εισαγωγή επιπλέον παραμέτρων στο μοντέλο μειώνει την τιμή των *AIC* και *BIC* μόνο αν αυτές βελτιώνουν την προσαρμογή του μοντέλου σε βαθμό που υπερβαίνει το αυξημένο αντίβαρο του πρώτου όρου τους.

Akaike's information criterion (AIC)
Bayesian information criterion (BIC)

Για το γενικό γραμμικό μοντέλο τα AIC και BIC λαμβάνουν την ακόλουθη μορφή

$$AIC = n\left[\ln\left(2\pi SSE/n\right) + 1\right] + 2\left(p+1\right)$$
$$BIC = n\left[\ln\left(2\pi SSE/n\right) + 1\right] + \left(p+1\right)\ln n$$

Χρήση Υπολογιστή

↪ Να υπολογίσετε τα μέτρα καταλληλότητας για τα δεδομένα του αρχείου STRENGTHCEMENTS.sav. Επιλέξτε το βέλτιστο μοντέλο με τα μέτρα αυτά.

↪ Να υπολογίσετε τα μέτρα καταλληλότητας για τα δεδομένα του αρχείου CHEMICAL.sav. Επιλέξτε το βέλτιστο μοντέλο με τα μέτρα αυτά.

Διαδικασίες επιλογής μοντέλου με βήματα

Η μεταβολή του SSE οδηγεί στον έλεγχο F

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

για την πρόσθεση ή αφαίρεση q όρων ενός μοντέλου (βλ. προηγούμενες διαφάνειες ▸ F-test).

↪ Αν εξετάσουμε την πρόσθεση μιας μεταβλητής (q=1) στο μοντέλο και η μείωση του SSE θεωρείται μικρή, τότε η μεταβλητή δε χρειάζεται στο μοντέλο. Αντιθέτως, αν η μείωση είναι μεγάλη, τότε η μεταβλητή πρέπει να εισαχθεί στο μοντέλο.

↪ Στην περίπτωση αφαίρεσης μιας μεταβλητής από το μοντέλο, τότε μια μικρή αύξηση του SSE σημαίνει ότι η μεταβλητή πράγματι δε χρειάζεται στο μοντέλο, αλλιώς αν η αύξηση είναι μεγάλη δεν μπορεί να αφαιρεθεί.

↪ Το τι αποτελεί «μεγάλη» ή «μικρή» μεταβολή του SSE κρίνεται συνήθως από τη στατιστική σημαντικότητα του ελέγχου F.

Διαδικασίες επιλογής μοντέλου με βήματα

Η εξεύρεση του ενδεχομένως βέλτιστου μοντέλου γίνεται με τη διαδοχική εφαρμογή του ελέγχου F σε διάφορα μοντέλα

Η εξέταση όλων των δυνατών υποσυνόλων δεν είναι πάντα εφικτή, διότι υπάρχουν $\binom{k}{r}$ υποσύνολα μεγέθους r και $\sum_{r=1}^{k} \binom{k}{r} = 2^k - 1$ υποσύνολα διαφόρων μεγεθών.

Για το λόγο αυτό έχουν αναπτυχθεί διάφορες τεχνικές για την επιλογή του καλύτερου υποσυνόλου μεταβλητών από ένα αρχικό σύνολο. Αυτές οι τεχνικές βασίζονται στον έλεγχο F για την αφαίρεση και την πρόσθεση μεταβλητών σε ένα μοντέλο.

Διαδικασίες επιλογής μοντέλου με βήματα

- **Διαδικασία της διαδοχικής αφαίρεσης ή απαλοιφής** (Backward Elimination
- **Διαδικασία της διαδοχικής πρόσθεσης ή της προς τα εμπρός επιλογής** (Forward Selection)
- **Κατά βήματα εμπρός πίσω επιλογή** (Stepwise selection).

Διαδικασία της διαδοχικής αφαίρεσης

1. Εισάγουμε όλες τις διαθέσιμες μεταβλητές στο μοντέλο.
2. Αφαιρούμε τη λιγότερο σημαντική μεταβλητή, αν η ενέργεια αυτή δε μεταβάλλει σημαντικά την τιμή της διαφοράς $SSE_0 - SSE_1$ (σύμφωνα με τον έλεγχο F).
3. Προσαρμόζουμε ξανά ένα μοντέλο παλινδρόμησης στα δεδομένα, παραλείποντας τη μεταβλητή που αφαιρέσαμε.
4. Επαναλαμβάνουμε τα βήματα 2 και 3 μέχρι η αφαίρεση μιας οποιασδήποτε μεταβλητής να είναι στατιστικά σημαντική, οπότε και σταματάμε.

Διαδικασία της διαδοχικής πρόσθεσης

1. Ξεκινάμε με το μοντέλο $y = \beta_0$, όπου β_0 σταθερός όρος και μάλιστα $\hat{\beta}_0 = \bar{y}$.
2. Προσθέτουμε τη μεταβλητή που δίνει τη μεγαλύτερη σημαντική μείωση του SSE, δηλαδή δίνει τη μεγαλύτερη στατιστικά σημαντική τιμή του ελέγχου F και επομένως θα είναι η μεταβλητή που συμβάλλει περισσότερο στην επεξήγηση της y.
3. Προσαρμόζουμε ξανά ένα μοντέλο παλινδρόμησης στα δεδομένα συμπεριλαμβάνοντας τη μεταβλητή που προσθέσαμε και διερευνούμε ποια θα είναι η επόμενη μεταβλητή που θα δώσει στατιστικά σημαντική μείωση στο SSE σε σύγκριση με το μοντέλο που περιέχει μόνο την πρώτη μεταβλητή και έτσι, να εισαχθεί στο μοντέλο.
4. Επαναλαμβάνουμε τα προηγούμενα δύο βήματα μέχρι η τιμή του ελέγχου F για την πρόσθεση οποιασδήποτε από τις εναπομείνασες μεταβλητές να μην είναι στατιστικά σημαντική, οπότε και σταματάμε.

Διαδικασία κατά βήματα

Η μέθοδος αυτή αφορά μια «διόρθωση» της μεθόδου της διαδοχικής πρόσθεσης.

Η «διόρθωση» αυτή αφορά την παρεμβολή ενός επιπλέον ελέγχου σε κάθε επανάληψη της διαδικασίας διαδοχικής πρόσθεσης μίας μεταβλητής.

Ο έλεγχος αυτός εξετάζει την περίπτωση, όπου η πρόσθεση μιας νέας μεταβλητής στο μοντέλο οδηγεί στην εξασθένηση της σημαντικότητας κάποιας άλλης, που είχε εισαχθεί νωρίτερα, με συνέπεια να πρέπει να εξεταστεί η αξία παραμονής της στο μοντέλο.

Παρατηρήσεις

$\hookrightarrow$Οι διαδικασίες βασίζονται σε στατιστικούς ελέγχους υποθέσεων, δεν είναι όμως η μόνη βάση σύγκρισης μεταξύ μοντέλων. Ιδιαίτερα σε μεγάλα μεγέθη δειγμάτων ενδέχεται οι έλεγχοι αυτοί να μην είναι πολύ χρήσιμοι, διότι μπορεί πολύ μικρά β_j, χωρίς ουσιαστική συμβολή, να βγουν στατιστικά σημαντικά.

$\hookrightarrow$Για την ολοκλήρωση της στατιστικής ανάλυσης, μετά από την επιλογή του μοντέλου, επιβάλλεται πάλι η εξέταση των προϋποθέσεων του μοντέλου.

Χρήση Υπολογιστή

↪ Να προσδιορίσετε το βέλτιστο μοντέλο για τα δεδομένα του αρχείου STRENGTHCEMENTS.sav. Εξετάστε τις προϋποθέσεις του γενικού γραμμικού μοντέλου για το μοντέλο αυτό.

↪ Να προσδιορίσετε το βέλτιστο μοντέλο για τα δεδομένα του αρχείου CHEMICAL.sav. Εξετάστε τις προϋποθέσεις του γενικού γραμμικού μοντέλου για το μοντέλο αυτό.