ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

- ΠΟΤΕ ΚΑΙ ΓΙΑΤΙ ΧΡΗΣΙΜΟΠΟΙΕΙΤΑΙ
- ΜΟΝΤΕΛΟ
- ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ
- ΕΡΜΗΝΕΙΑ ΤΩΝ ΕΚΤΙΜΗΤΩΝ ΤΩΝ ΠΑΡΑΜΕΤΡΩΝ ΤΩΝ ΣΥΝΤΕΛΕΣΤΩΝ ΠΑΛΙΝΔΡΟΜΗΣΗΣ
- ΥΠΟΘΕΣΕΙΣ
- ΠΙΝΑΚΑΣ ΑΝΑΔΙΑ
- ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑ ΓΙΑ ΤΙΣ ΠΑΡΑΜΕΤΡΟΥΣ

ΜΟΝΤΕΛΟ

Είναι

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} + \varepsilon_i$$

ή

$$Y = X\ \beta + \varepsilon\,,$$

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Εκτίμηση παραμέτρων

- Πολλαπλή παλινδρόμηση

$$\hat{\beta} = \left(X^{'} X \right)^{-1} X^{'} Y$$

- Απλή παλινδρόμηση

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

και

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n \sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i \right)^2}$$

Εκτιμώμενες τιμές-Υπόλοιπα Ερμηνεία των εκτιμητών των παραμέτρων

- Εκτιμώμενη τιμή

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + ... + \hat{\beta}_p X_{ip},$$

- Υπόλοιπο

$$e_i = Y_i - \hat{Y}_i$$

Υποθέσεις

- Τα σφάλματα ακολουθούν α) κανονική κατανομή με μέση τιμή 0 και β) σταθερή διακύμανση
- Τα σφάλματα είναι ασυσχέτιστα μεταξύ τους ανά δύο
- Το μοντέλο είναι ορθό.
- Υπάρχει ο πίνακας $\left(X^{'} X \right)^{-1}$
- Δεν υπάρχουν ακραίες ή επηρεάζουσες παρατηρήσεις.

ΠΙΝΑΚΑΣ ΑΝΑΔΙΑ

ΠΗΓΗ	Β.Ε	ΑΘΡ. ΤΕΤΡ.	ΜΕΣΑ ΤΕΤΡ.	F TEST
ΠΑΛ/ΣΗ	p	SSreg	MSreg= SSreg/p	F= MSreg/MSres
Υπόλοιπα	n-p-1	SSres	MSres= SSres/(n-p-1)	
Συνολικά	n-1	SStot		

$$SSreg = \hat{\beta}' X'Y - n\bar{Y}^2$$

$$SStot = Y'Y - n\bar{Y}^2$$

$$SSres = SStot - SSreg$$

- Το F στατιστικό ποια υπόθεση ελέγχει και υπό ποιες προϋποθέσεις? Ποια η κρίσιμη περιοχή του ελέγχου?

Συμπερασματολογία για τις παραμέτρους

$$\hat{\beta} \sim N_p\left(\beta, \sigma^2\left(X'X\right)^{-1}\right)$$

$$\hat{\sigma} = \sqrt{MS_{res}}$$

$$\left(\hat{\beta}_j - t_{a/2,n-p-1}\sqrt{\hat{Var}(\hat{\beta}_j)}, \hat{\beta}_j + t_{a/2,n-p-1}\sqrt{\hat{Var}(\hat{\beta}_j)}\right)$$

$$t = \frac{\hat{\beta}_j - b_j}{\sqrt{\hat{Var}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j - b_j}{\hat{\sigma}\sqrt{c_{i+1,i+1}}}$$

$c_{i+1,i+1}$ το (i+1,i+1) στοιχείο του πίνακα $\left(X'X\right)^{-1}$

Παράδειγμα προσαρμογής μοντέλου (έχοντας τις προεπιλογές)

Διαδικασία

Regression Linear

- Αρχείο karakostasp.22

Linear Regression Statistics

$$R^2 = \frac{SS_{reg}}{SS_{tot}}$$

$$\text{Adjusted } R^2 = R^2 - \frac{(1-R^2)(k-1)}{(n-k)}$$

k το πλήθος των ανεξάρτητων μεταβλητών συμπεριλαμβανομένου και του σταθερού όρου

ΜΕΘΟΔΟΙ ΕΠΙΛΟΓΗΣ ΜΕΤΑΒΛΗΤΩΝ

- Enter: όλες οι ανεξάρτητες μεταβλητές εισάγονται στο προσαρμοζόμενο μοντέλο με μιας
- Remove: θα βγουν μονομιάς όλες οι μεταβλητές
- Forward: μπαίνει μία κάθε φορά με κριτήριο αν η F-τιμή είναι μεγαλύτερη μίας καθορισμένης
- Backward: όλες οι μεταβλητές εισάγονται αρχικά σε ένα βήμα και έπειτα «βγαίνει» μία κάθε φορά με κάποιο κριτήριο στη βάση κάποιου F τεστ και αν η τιμή του είναι μικρότερη μίας προκαθορισμένης
- Stepwise: συνδυασμός των Backward και Forward.

Linear regression Statistics

- Estimates (προεπιλογή)

  Εμφανίζει τους εκτιμητές ελαχίστων τετραγώνων των παραμέτρων του μοντέλου παλινδρόμησης, τα αντίστοιχα τυπικά σφάλματά τους και τέλος τα t στατιστικά για τον έλεγχο της υπόθεσης ότι οι παράμετροι αυτές διαφέρουν σημαντικά από το 0.

Linear regression Statistics

- Confindence intervals

  Υπολογίζει το 95% Δ.Ε. για τους συντελεστές της παλινδρόμησης.

- Covariance matrix

  Μας δίνεται ο πίνακας διακυμάνσεων-συνδιακυμάνσεων των συντελεστών της παλινδρόμησης. Επιπλέον ο πίνακας συσχέτισης δίνεται.

Linear regression Statistics

- Model Fit (προεπιλογή)

  Αναφέρονται οι μεταβλητές που εισέρχονται ή φεύγουν από το μοντέλο. Επιπλέον υπολογίζει το R στατιστικό, το συντελεστή προσδιορισμού, καθώς και τον διορθωμένο συντελεστή προσδιορισμού. Επιπλέον εμφανίζει τον πίνακα ΑΝΑΔΙΑ (ANOVA).

Linear regression Statistics

- R squared change (έχει νόημα για πολλαπλή παλινδρόμηση)

  Η μεταβολή στο συντελεστή παλινδρόμησης που προκαλείται από την προσθήκη ή αφαίρεση μιας ανεξάρτητης μεταβλητής. Αν για κάποια μεταβλητή είναι μεγάλη αυτό σημαίνει ότι η μεταβλητή αυτή είναι καλή για την πρόβλεψη των τιμών της εξαρτημένης μεταβλητής.

Linear regression Statistics

- Descriptives

  Υπολογίζονται οι διαθέσιμες παρατηρήσεις, οι μέσες τιμές και τυπικές αποκλίσεις για όλες τις μεταβλητές. Επιπλέον, δίνεται ο πίνακας συσχέτισης των μεταβλητών και η p-τιμή για το μονόπλευρο έλεγχο υποθέσεων.

Linear regression Statistics

- *Partial Correlation*: Η συσχέτιση μεταξύ της εξαρτημένης και μίας ανεξάρτητης μεταβλητής όταν αφαιρείται η γραμμική επίδραση και από τις δύο, η οποία οφείλεται στην αμοιβαία σχέση τους με άλλες. Δηλαδή πρόκειται για το μερικό συντελεστή συσχέτισης.
- *Part Correlation*: Η συσχέτιση μεταξύ της εξαρτημένης και μίας ανεξάρτητης μεταβλητής όταν αφαιρείται, μόνο από την ανεξάρτητη μεταβλητή η γραμμική επίδραση η οποία οφείλεται στην αμοιβαία σχέση τους με άλλες. Συνήθως ονομάζεται semipartial correlation.

Linear regression Statistics

Collinearity diagnostics

Η συγγραμμικότητα (Collinearity) ή πολυσυγγραμμικότητα (multicollinearity) είναι εκείνη η ανεπιθύμητη κατάσταση (εμφανίζεται στην πολυμεταβλητή παλινδρόμηση) όπου μία ανεξάρτητη μεταβλητή είναι γραμμική συνάρτηση των υπόλοιπων ή κάποιων ανεξάρτητων μεταβλητών. Για την εξέταση αυτού του προβλήματος δίνονται διάφορα διαγνωστικά μέτρα.

Linear regression Statistics

Durbin Watson

Στατιστικός έλεγχος για αυτοσυσχέτιση πρώτου βαθμού των υπολοίπων.

Casewise diagnostics

Μας δίνονται οι τιμές των μεταβλητών για περιπτώσεις που έχουν τυποποιημένα υπόλοιπα με απόλυτη τιμή μεγαλύτερη του καθορισμένου κριτηρίου. Το συνηθέστερο κριτήριο είναι 3 φορές η τυπική απόκλιση. Είναι χρήσιμος έλεγχος για την ύπαρξη ακραίων τιμών.

Linear regression PLOTS

- Γραφικές παραστάσεις (διαγράμματα διασποράς) για τον έλεγχο των απαιτούμενων υποθέσεων για την εφαρμογή του μοντέλου της γραμμικής παλινδρόμησης όπως είναι υπόθεση της κανονικότητας, της γραμμικότητας και της ίσης διασποράς. Επιπλέον, μπορούν να γίνουν γραφικές παραστάσεις για τον εντοπισμό ακραίων και επηρεάζουσων παρατηρήσεων

Linear regression PLOTS

- Εξαρτημένη μεταβλητή (DEPENDNT)
- Τυποποιημένες και Adjusted εκτιμώμενες τιμές (*ZPRED και *ADJPRED)
- Τυποποιημένα, διαγραφόμενα, μαθητικοποιημένα και μαθητικοποιημένα διαγραφόμενα υπόλοιπα

(*ZRESID, *DRESID, *SRESID,*SDRESID).

Linear regression PLOTS

- Normal Probability plot και Histogram το λογισμικό μας δίνει στο Output το P-P γράφημα για τον έλεγχο της κανονικότητας των τυποποιημένων υπολοίπων καθώς επίσης και το ιστόγραμμά τους για να διαπιστώσουμε πιθανές αποκλίσεις από την κανονικότητα.
- Produce all partial plots το λογισμικό μας εφοδιάζει με διαγράμματα διασποράς πολύ χρήσιμα για τον έλεγχο των υποθέσεων του μοντέλου της πολλαπλής παλινδρόμησης

Linear regression SAVE PREDICTED VALUES

- Unstandardized
- Standardized
- Αφαιρείται από κάθε εκτιμώμενη τιμή η μέση τιμή των εκτιμώμενων τιμών και το αποτέλεσμα διαιρείται με την τυπική απόκλιση των εκτιμώμενων τιμών

PREDICTED VALUES

- Adjusted. Η εκτιμώμενη τιμή για την πειραματική μονάδα όταν αυτή δε χρησιμοποιηθεί για τον υπολογισμό των συντελεστών παλινδρόμησης
- S.E. of mean predictions. Το τυπικό σφάλμα των εκτιμώμενων τιμών. Ένας εκτιμητής της μέσης τιμής της εξαρτημένης μεταβλητής για πειραματικές μονάδες που έχουν ίδιες τιμές εξαρτημένων μεταβλητών.

RESIDUALS

- Unstandardized

$$e_i = Y_i - \hat{Y}_i$$

- Standardized

$$e_{si} = \frac{e_i}{\sqrt{MSres}}$$

- Studentized

$$t_i = \frac{e_i}{\sqrt{(1 - p_{ii})\,MSres}}$$

$$P = X(X'X)^{-1}X'$$

RESIDUALS

- Deleted
- Είναι τα υπόλοιπα που προκύπτουν αν η συγκεκριμένη πειραματική μονάδα δεν ληφθεί υπόψη για τον υπολογισμό των συντελεστών παλινδρόμησης.
- Studentized deleted

  Τα λεγόμενα μαθητικοποιημένα διαγραφόμενα υπόλοιπα που προκύπτουν αν διαιρέσουμε την τιμή των διαγραφόμενων υπολοίπων με το τυπικό σφάλμα τους

DISTANCES

Mahalanobis

- Η απόσταση αυτή μας δείχνει πόσο απέχει μία τιμή μίας ανεξάρτητης μεταβλητής από το μέσο των περιπτώσεων. Μεγάλες τιμές υποδεικνύουν ότι η συγκεκριμένη πειραματική μονάδα έχει extreme values σε μία ή περισσότερες ανεξάρτητες μεταβλητές.

DISTANCES

- Cook's: Η απόσταση αυτή μετρά πόσο οι τιμές των υπολοίπων όλων των περιπτώσεων θα μεταβληθούν αν η συγκεκριμένη τιμή δε ληφθεί υπόψη στους υπολογισμούς των συντελεστών του μοντέλου. Μεγάλες τιμές αυτής της απόστασης υποδεικνύουν ότι η εξαίρεση της συγκεκριμένης πειραματικής μονάδας επιφέρει σημαντικές και ουσιαστικές αλλαγές.

DISTANCES

- Leverage values: Μετρούν την επίδραση μίας πειραματικής μονάδας στην προσαρμογή του μοντέλου της παλινδρόμησης. Παίρνει τιμές από 0 (όχι ενδείξεις επίδρασης) έως (N-1)/N.

PREDICTION INTERVALS

- Mean:100 (1-α)% διάστημα εμπιστοσύνης για την

$$E\left(Y_i\right)$$

$$\left(\hat{Y}_i - S_{\hat{Y}_i} t_{a/2,n-p-1}, \hat{Y}_i + S_{\hat{Y}_i} t_{a/2,n-p-1}\right)$$

$$S_{\hat{Y}_i} = \sqrt{MSres}\sqrt{x_i^{'}\left(X^{'}X\right)^{-1}x_i}$$

PREDICTION INTERVALS

- Individual υπολογίζεται για κάθε πειραματική μονάδα το κάτω και άνω όριο του 100 (1-α)% διαστήματος εμπιστοσύνης για την ατομική τιμή Y_i

$$\left(\hat{Y}_i - S_i t_{a/2,n-p-1}, \hat{Y}_i + S_i t_{a/2,n-p-1}\right)$$

$$S_i = \sqrt{MSres}\sqrt{\left(1 + x_i'\left(X'X\right)^{-1}x_i\right)}$$

Linear Regression Save

- Influence Statistics: Στατιστικά για την εξέταση επηρεάζουσων παρατηρήσεων
- Coefficient Statistics έχουμε την δυνατότητα της αποθήκευσης των συντελεστών της παλινδρόμησης σε ένα σύνολο δεδομένων ή σε νέο αρχείο.

Linear Regression Options

- ΧΕΙΡΙΣΜΟΣ ΕΛΛΙΠΩΝ ΤΙΜΩΝ
- ΔΗΛΩΝΟΥΜΕ ΤΗΝ ΠΙΘΑΝΟΤΗΤΑ- ΤΙΜΗ ΤΟΥ F ΚΡΙΤΗΡΙΟΥ ΓΙΑ ΕΙΣΑΓΩΓΗ Η ΕΞΑΓΩΓΗ ΜΙΑΣ ΜΕΤΑΒΛΗΤΗΣ ΑΠΟ ΤΟ ΜΟΝΤΕΛΟ
- Η τιμή στο Entry θα πρέπει να είναι μικρότερη από την τιμή στο Removal. Αυξάνοντας την τιμή στο Entry υπεισέρχονται περισσότερες μεταβλητές, ενώ μειώνοντας την τιμή στο Removal βγαίνουν από το μοντέλο περισσότερες μεταβλητές.
- ΚΑΘΟΡΙΖΟΥΜΕ ΑΝ ΥΠΕΙΣΕΡΧΕΤΑΙ ΣΤΟ ΜΟΝΤΕΛΟ ΣΤΑΘΕΡΟΣ ΟΡΟΣ (include constant in equation)