ΑΚΡΑΙΕΣ ΤΙΜΕΣ ΣΤΗΝ ΠΑΛΙΝΔΡΟΜΗΣΗ

- ΤΡΟΠΟΙ ΕΛΕΓΧΟΥ
- ΤΡΟΠΟΙ ΑΝΤΙΜΕΤΩΠΙΣΗΣ
- ΣΥΝΕΠΕΙΕΣ ΠΡΟΒΛΗΜΑΤΟΣ
- ΠΑΡΑΔΕΙΓΜΑΤΑ

Ακραίες παρατηρήσεις

Γνώρισμά τους η μη προσαρμογή σε κάποιο μοντέλο που φαίνεται να προσαρμόζεται σωστά το κύριο σώμα των παρατηρήσεων.

ΤΡΟΠΟΙ ΕΛΕΓΧΟΥ

- Παρατηρήσεις με απόλυτες τιμές των τυποποιημένων υπολοίπων μεγαλύτερες του 3 θεωρούνται ακραίες *.
- Παρατηρήσεις με απόλυτες τιμές των τυποποιημένων υπολοίπων μεταξύ του 2 και 3 θεωρούνται πιθανές ακραίες και χρήζουν εξέτασης.
- Στατιστικός τρόπος ελέγχου.

Στατιστικός τρόπος ελέγχου.

- Επιτυγχάνεται με τα Μαθητικοποιημένα Διαγραφόμενα Υπόλοιπα.
- Απόλυτες τιμές των μαθητικοποιημένων διαγραφόμενων υπολοίπων για μία παρατήρηση μεγαλύτερες του IDF.T(1-a/2,n-p-2), δηλαδή του

$$t_{a/2,n-p-2}$$

υποδεικνύουν τη συγκεκριμένη παρατήρηση ως ακραία.

ΑΝΕΞΑΡΤΗΤΕΣ- ΑΚΡΑΙΕΣ

- Η εξέταση της ύπαρξης ακραίων τιμών λόγω μίας ή περισσότερων ανεξάρτητων μεταβλητών επιτυγχάνεται μέσω των ποσοτήτων Leverage , που δίνονται από τη σχέση:

$$h_{ii} = x_i^{'}\left(X^{'}X\right)^{-1}x_i$$

- Επισημαίνονται παρατηρήσεις με τιμές αυτού του δείκτη μεγαλύτερες του

$$\frac{2\ (\ p\ +\ 1\ )}{n}$$

Leverage

- Το λογισμικό μας εφοδιάζει με τις centered leverage δηλ. τις

$$h_{ii}^{*} = h_{ii} - \frac{1}{n}$$

οπότε συγκρίνεται με την τιμή

$$\frac{2p}{n}$$

- Διαφοροποίηση cut off point

$$\frac{3p}{n}$$

Multivariate outliers

- Cook Distance: Μας καθορίζει πόσο οι τιμές των υπολοίπων όλων των περιπτώσεων θα μεταβληθούν αν η συγκεκριμένη τιμή δε ληφθεί υπόψη στους υπολογισμούς των συντελεστών του μοντέλου
- Cut-off point: 1 ή 4/(n-p) ή 4/[n-p-1]]

IDF.F(1-a,p+1,n-p-1)= $F_{a,p+1,n-p-1}$

Cook's Distance

$$C_i = \frac{e_i^2}{p+1} \frac{h_{ii}}{1-h_{ii}}$$

Multivariate Outliers

- Mahalanobis Distance

  Παρατηρήσεις με μεγάλες τιμές της απόστασης Mahalanobis θα πρέπει να εξετάζονται.

ΔΙΟΡΘΩΣΗ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ

- Εξέταση μήπως πρόκειται για λάθος καταγραφή ή δεν ορίστηκε η κωδική ονομασία των ελλιπών τιμών
- Απόρριψη ακραίων τιμών?
- Μετασχηματισμός για διόρθωση του προβλήματος.
- Χρήση ανθεκτικών μεθόδων σε ακραίες τιμές (Huber (1973)

ΣΥΝΕΠΕΙΕΣ

- Οι εκτιμητές και οι διακυμάνσεις αυτών δεν έχουν τις γνωστές ιδιότητες
- Ίσως προκαλείται και πρόβλημα κανονικότητας
- Μη εγκυρότητα των ελέγχων υποθέσεων και Δ.Ε.

Επηρεάζουσες παρατηρήσεις

- Πειραματικές μονάδες που επιδρούν σημαντικά στο μοντέλο παλινδρόμησης. Π.χ. οι συντελεστές των παραμέτρων του μοντέλου αλλάζουν αρκετά όταν οι τιμές των συγκεκριμένων πειραματικών μονάδων εξαιρούνται από τον υπολογισμό τους. Μία τέτοια κατάσταση είναι ανεπιθύμητη καθώς θέλουμε ένα μοντέλο παλινδρόμησης που να μην εξαρτάται από τις τιμές ενός μικρού αριθμού πειραματικών μονάδων.

Επηρεάζουσες- Τρόποι ελέγχου Linear Regression Save

- **DfBeta(s)**: Η διαφορά στις τιμές των συντελεστών της παλινδρόμησης αν δεν ληφθεί υπόψη η συγκεκριμένη πειραματική μονάδα. Υπολογίζεται και για τον σταθερό όρο. Οι τυποποιημένες τιμές παρατίθενται στη στήλη Standardized DfBeta. Απόλυτες τιμές αυτών μεγαλύτερες από

$$2 / \sqrt{n}$$

μας υποδεικνύουν παρατήρηση που επιδρά στην εκτίμηση των συντελεστών της παλινδρόμησης.

Επηρεάζουσες- Τρόποι ελέγχου Linear Regression Save

- **DfFit**:

$$t_i \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{1/2}$$

- Μετρά την διαφορά στην προσαρμογή, δηλαδή στην εκτιμώμενη τιμή, αν δεν συμπεριληφθεί η συγκεκριμένη παρατήρηση στους υπολογισμούς. Δίνονται και οι αντίστοιχες τυποποιημένες τιμές Standardized DfFit. Απόλυτες τιμές αυτών μεγαλύτερες του

$$2\sqrt{\frac{p+1}{n}}$$

υποδεικνύουν επηρεάζουσες παρατηρήσεις.

Επηρεάζουσες- Τρόποι ελέγχου Linear Regression Save

- **Covariance ratio**: Το πηλίκο της ορίζουσας του πίνακα διακυμάνσεων συνδιακυμάνσεων χωρίς η συγκεκριμένη παρατήρηση να λαμβάνεται υπόψη στους υπολογισμούς προς την αντίστοιχα ορίζουσα όταν λαμβάνεται υπόψη. Τιμές μεγαλύτερες (μικρότερες αντίστοιχα) του

$$1 + 3\frac{p+1}{n} \qquad 1 - 3\frac{p+1}{n}$$

υποδεικνύουν επηρεάζουσα παρατήρηση.

Υλοποίηση μεθόδων με cut-off point

- Υπολογισμός του cut-off point.
- Δημιουργία στήλης με αύξοντες αριθμούς παρατήρησης. Συνάρτηση $CASENUM.
- Γραφική παράσταση των διαγνωστικών μέτρων ως προς τη νέα μεταβλητή.

  Graphs→ Legacy Dialog→ Scatter/Dot και Simple Scatter.

Υλοποίηση μεθόδων με cut-off point

- Κάνοντας διπλό κλικ στο γράφημα που προκύπτει και έπειτα δεξί κλικ ζητούμε να Add→ Y Axis Reference Line και στο πλαίσιο Position δηλώνουμε την τιμή του cut-off point. Αν υπάρχουν σημεία πάνω από αυτή την γραμμή με δεξί κλικ και επιλογή του Show Data Labels μας υποδεικνύεται ποια παρατήρηση είναι (εναλλακτικά αφού το επιλέξουμε με δεξί κλικ και επιλογή του Go to Case) .

Παράδειγμα
Chatterjee and Price (1980), p. 21

- Στο αρχείο δεδομένων chatterjee21.sav καταγράφονται 30 παρατηρήσεις και 2 μεταβλητές που αφορούν την ακροαματικότητα πριν το δελτίο ειδήσεων (lead in) και την ακροαματικότητα του δελτίου ειδήσεων (newsrate). Θέλουμε να εξετάσουμε αν το πρόγραμμα πριν τις ειδήσεις επηρεάζει την ακροαματικότητα των ειδήσεων.