ΠΡΟΒΛΗΜΑ ΣΥΓΓΡΑΜΜΙΚΟΤΗΤΑΣ

- 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.

ΠΡΟΒΛΗΜΑ ΣΥΓΓΡΑΜΜΙΚΟΤΗΤΑΣ

- ΕΝΔΕΙΞΕΙΣ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ
- ΤΡΟΠΟΙ ΕΛΕΓΧΟΥ
- ΔΙΟΡΘΩΣΗ ΠΡΟΒΛΗΜΑΤΟΣ
- ΣΥΝΕΠΕΙΕΣ ΠΡΟΒΛΗΜΑΤΟΣ
- ΠΑΡΑΔΕΙΓΜΑΤΑ

ΕΝΔΕΙΞΕΙΣ

- Προσθήκη ή αφαίρεση μίας ανεξάρτητης μεταβλητής (ή ακόμα και παρατήρησης) επιφέρει μεγάλες αλλαγές στους εκτιμητές των συντελεστών του μοντέλου της γραμμικής παλινδρόμησης.
- Το F- τεστ του πίνακα ΑΝΑΔΙΑ να απορρίπτει τη μηδενική υπόθεση, ενώ τα επιμέρους t-τεστ για τους συντελεστές της παλινδρόμησης να μην απορρίπτονται.
- Μεγάλοι συντελεστές συσχέτισης μεταξύ ζευγαριών ανεξάρτητων μεταβλητών.

ΕΝΔΕΙΞΕΙΣ

- Πρόσημα των εκτιμητών των παραμέτρων μη αναμενόμενα σε σχέση με την εκ των προτέρων γνώση, εμπειρία ή από τη γνωστή θεωρία.
- Μεγάλα διαστήματα εμπιστοσύνης για τους συντελεστές παλινδρόμησης σημαντικών μεταβλητών.

ΤΡΟΠΟΙ ΕΛΕΓΧΟΥ

Collinearity Diagnostics: Linear Regression Statistics

- Variance Inflation Factor
- Tolerance
- Condition number
- Condition index
- Variance-decomposition proportion

Variance Inflation Factor

Δίνεται από τη σχέση:

$$VIF_j = \frac{1}{1 - R_j^2}$$

όπου

$$R_j^2$$

ο συντελεστής προσδιορισμού της j ανεξάρτητης μεταβλητής ως προς τις άλλες p-2 το πλήθος ανεξάρτητες μεταβλητές.

Variance Inflation Factor

- Πρόβλημα συγγραμμικότητας αν

$$\max_j VIF_j > 10$$

- Επίσης αν

$$\frac{\sum_{j=1}^{p} VIF_j}{p} > 1$$

Tolerance

- Δίνεται από τη σχέση

$$1/VIF_j = 1\text{-}\mathrm{R}_{\mathrm{j}}^2$$

Τιμές αυτού του δείκτη μικρότερες από 0.1 υποδεικνύουν πρόβλημα συγγραμμικότητας.

Condition number

- Το πηλίκο της μέγιστης προς ελάχιστης ιδιοτιμής του πίνακα Χ.
- Τιμές αυτού του δείκτη μεγαλύτερες από 1000 υποδεικνύουν πρόβλημα συγγραμικότητας.

Condition index

- Η τετραγωνική ρίζα του πηλίκου της μέγιστης ιδιοτιμής του πίνακα X ως προς την j ιδιοτιμή.

- Τιμές αυτού του δείκτη μεταξύ 30 και 100 υποδεικνύουν μέτριο πρόβλημα, ενώ τιμές αυτού του δείκτη μεγαλύτερες του 100 υποδεικνύουν σοβαρό πρόβλημα.

Variance-decomposition proportion

- Χρησιμοποιώντας το γεγονός ότι

$$X = UDV'$$

$$U'U = V'V = I_p$$

$$D = diag(\lambda_0, \lambda_1, ..., \lambda_p)$$

Variance-decomposition proportion

Είναι:

$$Var\left(\hat{\beta}\right)=\sigma^2\left(X'X\right)^{-1}=\sigma^2VD^2V'$$

$$Var\left(\hat{\beta}_j\right)=\sigma^2\sum\frac{v_{jk}^2}{\lambda_j^2}$$

Variance-decomposition proportion

- Ορίζεται ως Variance-decomposition proportion το πηλίκο

$$\pi_{jk} = \frac{\phi_{jk}}{\phi_k}$$

$$\phi_{kj} = \frac{\nu^2_{jk}}{\lambda^2_j}$$

$$\phi_k = \sum_{k=0}^{p} \phi_{kj}$$

ΚΑΝΟΝΑΣ

- Ο αριθμός των condition index με τιμές μεγαλύτερες του 30 προσδιορίζει το πλήθος των εξαρτήσεων στις στήλες του Χ. Έπειτα ο καθορισμός των μεταβλητών που λαμβάνουν μέρος σε αυτές γίνεται εντοπίζοντας τις υψηλές τιμές του Variance-decomposition proportion (τιμές μεγαλύτερες του 50%) σε τουλάχιστον δύο συντελεστές της παλινδρόμησης.

ΔΙΟΡΘΩΣΗ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ

- Διεξαγωγή παλινδρόμησης με κύριες συνιστώσες.
- Εφαρμογή της μεθόδου της Ridge Regression.
- Απαλοιφή ανεξάρτητων μεταβλητών ή προσθήκη περιορισμών για τις παραμέτρους ή προσθήκη μερικών παρατηρήσεων.

ΣΥΝΕΠΕΙΕΣ

- Οι εκτιμητές των παραμέτρων του μοντέλου της παλινδρόμησης τείνουν να έχουν μεγάλη δειγματική διακύμανση όταν οι ανεξάρτητες μεταβλητές είναι υψηλά συσχετισμένες. Επομένως, οι εκτιμητές αυτοί τείνουν να διαφέρουν εξαιρετικά από ένα δείγμα σε άλλο. Αυτό ίσως έχει ως αποτέλεσμα ανακριβείς πληροφορίες για τους συντελεστές παλινδρόμησης (λάθος συμπερασματολογία).

ΣΥΝΕΠΕΙΕΣ

- Επιπλέον, η ερμηνεία των συντελεστών παλινδρόμησης ότι μετρούν τη μεταβολή στην αναμενόμενη τιμή της εξαρτημένης μεταβλητής όταν η αντίστοιχη ανεξάρτητη αυξηθεί κατά μία μονάδα, ενώ όλες οι υπόλοιπες ανεξάρτητες μεταβλητές παραμένουν ίδιες δεν μπορεί να χρησιμοποιηθεί.

Παλινδρόμηση με κύριες συνιστώσες

ΑΔΥΝΑΜΙΑ ΤΟΥ ΛΟΓΙΣΜΙΚΟΥ

Ridge Regression

1. Κανονικοποιούμαι τις μεταβλητές.

2. Οι εκτιμητές ελαχίστων τετραγώνων με τη μέθοδο Ridge υπολογίζονται από τη σχέση:

$$b_R = \left(Z'Z + kI\right)^{-1} Z'Y^*$$

3. Η επιλογή του k γίνεται μεταξύ άλλων με την Ridge Trace μέθοδο.

Ridge trace μέθοδος

- Επιλέγουμε εκείνο το k στο διάστημα (0,1) για το οποίο παρατηρούμε ότι οι εκτιμητές των συντελεστών της παλινδρόμησης γίνονται πιο σταθεροί.
- Η γραφική παράσταση των εκτιμητών ως προς το k ίσως βοηθά σε αυτή την απόφαση.
- Επιστροφή στο αρχικό μοντέλο?

Επιστροφή στο αρχικό μοντέλο

Δίνονται με βάση τις σχέσεις:

$$b_i = \frac{b_{iR} S_Y}{S_{X_i}}$$

$$b_o = b_{oR} - \sum_{i=1}^{p} \frac{b_{iR} \bar{X}_i S_Y}{S_i} = \bar{Y} - \sum_{i=1}^{p} \frac{b_{iR} \bar{X}_i S_Y}{S_i}$$

Παραδείγματα

1. Chatterjee p. 146
2. Chatterjee p. 152
3. Chatterjee p. 158