Γραμμικά Μοντέλα
Κατάλοιπα Παλινδρόμησης και Ανάλυση Διασποράς

Διδάσκουσα: Λουκία Μελιγκοτσίδου
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
Τμήμα Μαθηματικών

March 28, 2020

Κατάλοιπα Παλινδρόμησης

Έστω το απλό γραμμικό μοντέλο

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ i = 1, \ldots, n.$$
$$E(Y_i) = \beta_0 + \beta_1 X_i,\ \varepsilon_i = Y_i - E(Y_i).$$

Η παλινδρόμηση εκτιμάται από την ευθεία $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ (λέμε ότι προσαρμόζουμε αυτή την ευθεία στα δεδομένα μας). Η τιμή $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ καλείται προσαρμοσμένη τιμή ($fitted\ value$) ενώ η Y_i καλείται παρατηρούμενη τιμή ($observed\ value$).

Το i κατάλοιπο, $\hat{\varepsilon}_i$, είναι η διαφορά μεταξύ της παρατηρούμενης και της προσαρμοσμένης τιμής:

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i.$$

Τα κατάλοιπα είναι οι εκτιμήσεις των τυχαίων όρων. Η μοντελοποίηση στηρίχθηκε σε συγκεκριμένες υποθέσεις για τους τυχαίους όρους (κλασικές υποθέσεις). Τα κατάλοιπα χρησιμοποιούνται για να ελεγχθεί αν ισχύουν οι υποθέσεις αυτές.

Ιδιότητες καταλοίπων

1) $\sum \hat{\varepsilon}_i = 0$ $(*)$

Είναι $\sum \hat{\varepsilon}_i = \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = \sum Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum X_i = 0$
(από την πρώτη κανονική εξίσωση)
$(*) \Rightarrow \sum (Y_i - \hat{Y}_i) = 0 \Rightarrow \sum Y_i = \sum \hat{Y}_i$ Οι παρατηρούμενες τιμές και οι προσαρμοσμένες τιμές έχουν ίδιο μέσο.

2) $\sum \hat{\varepsilon}_i^2$ είναι ελάχιστο (απαίτηση στη μέθοδο ελαχίστων τετραγώνων).

3) $\sum X_i \hat{\varepsilon}_i = 0$

Είναι $\sum X_i \hat{\varepsilon}_i = \sum X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = \sum X_i Y_i - \hat{\beta}_0 \sum X_i - \hat{\beta}_1 \sum X_i^2 = 0$
(από την δεύτερη κανονική εξίσωση)

4) $\sum \hat{Y}_i \hat{\varepsilon}_i = 0$

Είναι $\sum \hat{Y}_i \hat{\varepsilon}_i = \sum (\hat{\beta}_0 + \hat{\beta}_1 X_i) \hat{\varepsilon}_i = \hat{\beta}_0 \hat{\varepsilon}_i + \hat{\beta}_1 \sum X_i \hat{\varepsilon}_i = 0$

Ας μην ξεχνάμε ότι τα κατάλοιπα $\hat{\varepsilon}_i$ είναι οι εκτιμήσεις των τυχαίων σφαλμάτων, ε_i. Κάτω από τις υποθέσεις του γραμμικού μοντέλου $E(\varepsilon_i) = 0, V(\varepsilon_i) = \sigma^2$, $Cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j$. Τα κατάλοιπα, λοιπόν, πρέπει να είναι τυχαία κατανεμημένα γύρω από το 0, ασυσχέτιστα και ομοσκεδαστικά. Αν αυτά δεν ισχύουν, τότε το μοντέλο που έχουμε προσαρμόσει δεν είναι κατάλληλο για τα δεδομένα μας.

Το αν ισχύουν οι υποθέσεις για τους τυχαίους όρους ελέγχεται με γραφικό έλεγχο καταλοίπων, δηλαδή με τα γραφήματα των καταλοίπων ως προς τα X_i ή τα i.

Ένα γράφημα καταλοίπων που επιβεβαιώνει τις κλασικές υποθέσεις παρουσιάζει την εικόνα σύννεφου τυχαίων σημείων γύρω από τη γραμμή του 0 και δεν έχει τίποτα συστηματικό. Ό, τι συστηματικό ανιχνευθεί στο γράφημα καταλοίπων αντιστοιχεί σε απόκλιση από τις κλασικές υποθέσεις και πρέπει να μοντελοποιηθεί.

Παλινδρόμηση και Ανάλυση Διασποράς

Έστω το μοντέλο $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$.
Η διασπορά των Y_i μετριέται από το άθροισμα των τετραγώνων των διαφορών $Y_i - \bar{Y}$, δηλαδή από το συνολικό άθροισμα τετραγώνων

$$\sum (Y_i - \bar{Y})^2 \; Total \; sum \; of \; squares \; (SST)$$

Το άθροισμα των τετραγώνων των καταλοίπων είναι

$$\sum \hat{\varepsilon}_i = \sum (Y_i - \hat{Y})^2 \; Error \; sum \; of \; squares \; (SSE)$$

Ισχύει ότι

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

Θα δείξουμε ότι

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2$$

Απόδειξη

$$\sum(Y_i-\bar{Y})^2=\sum\Big[(Y_i-\hat{Y}_i)+(\hat{Y}_i-\bar{Y})\Big]^2=$$
$$=\sum(\Big[(Y_i-\hat{Y}_i)^2+(\hat{Y}_i-\bar{Y})^2+2(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y})\Big]=$$
$$=\sum(Y_i-\hat{Y}_i)^2+\sum(\hat{Y}_i-\bar{Y})^2$$

Γιατί

$$2\sum(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y})=2\sum\hat{Y}_i(Y_i-\hat{Y}_i)-2\bar{Y}\sum(Y_i-\hat{Y})=$$
$$=2\sum\hat{Y}_i\hat{\varepsilon}_i-2\bar{Y}\sum\hat{\varepsilon}_i=0\quad\text{από ιδιότητες καταλοίπων.}$$

Άρα

$\sum(Y_i-\bar{Y})^2$	$=$	$\sum(Y_i-\hat{Y}_i)^2$	$+$	$\sum(\hat{Y}_i-\bar{Y})^2$
SST		SSE		SSR ($Regression\ sum\ of\ squares$)
Συνολική μεταβλητότητα των Y_i		Μεταβλητότητα που που αποδίδεται στα σφάλματα		Μεταβλητότητα που εξηγείται από την παλινδρόμηση
$n-1$ β.ε.		$n-2$ β.ε.		1 β.ε.

Πίνακας Ανάλυσης Διασποράς (*Analysis of Variance Table - ANOVA*)

Πηγή Μεταβλητότητας *SourceofVariation*	Άθροισμα Τετραγώνων *SS*	*B.E.* *d.f.*	*MeanSquare* *MS*
Παλινδρόμηση	$SSR=\sum(\hat{Y}_i-\bar{Y})^2$	1	$MSR=\dfrac{SSR}{1}$
Σφάλματα	$SSE=\sum(Y_i-\hat{Y})^2$	$n-2$	$MSE=\dfrac{SSE}{n-2}$
Σύνολο	$SST=\sum(Y_i-\bar{Y})^2$	$n-1$	

Θεώρημα *Cochran.* Έστω n παρατηρήσεις $Y_i\sim N(\mu,\sigma^2)$ και $\sum(Y_i-\bar{Y})^2=Q_1+Q_2+\ldots+Q_k$ όπου τα Q_i είναι τετραγωνικές μορφές ως προς τα $Y_1,Y_2,\ldots,Y_n$ με j, $j=1,\ldots,k$ β.ε. αντίστοιχα. Τότε οι τ.μ. $Q_1,Q_2,\ldots,Q_k$ είναι ανεξάρτητες και $\dfrac{Q_j}{\sigma^2}\sim X^2_{(r_j)}$, $j=1,\ldots,k$ αν και μόνο αν $\sum_{j=1}^k r_j=n-1$

Το θεώρημα $Cochran$ μπορεί να χρησιμοποιηθεί για τον έλεγχο υποθέσεων

$$\begin{array}{ll} H_0: & \beta_1 = 0 \quad (\text{δεν υπάρχει γραμμική σχέση}) \\ H_1: & \beta_1 \neq 0 \end{array}$$

με τη χρήση του κριτηρίου F.

Κάτω από την $H_0 : Y_i \sim N(\beta_0, \sigma^2)$, δηλαδή τα Y_i είναι ταυτοτικά κατανεμημένα.

Έχουμε $\sum(Y_i - \bar{Y})^2 = \underbrace{\sum(Y_i - \hat{Y}_i)^2}_{Q_1} + \underbrace{\sum(\hat{Y}_i - \bar{Y})^2}_{Q_2}$

Οι υποθέσεις του θεωρήματος $Cochran$ πληρούνται. Άρα οι τ.μ. $\dfrac{\sum(Y_i - \hat{Y}_i)^2}{\sigma^2}$και $\dfrac{\sum(\hat{Y}_i - \bar{Y}_i)^2}{\sigma^2}$ είναι ανεξάρτητες με $n-2$ και 1 β.ε. αντίστοιχα.

Παρατήρηση

Αν $U \sim X^2_{(\nu_1)}$, $V \sim X^2_{(\nu_2)}$ και U, V ανεξάρτητες τότε $F = \dfrac{U/\nu_1}{V/\nu_2} \sim F(\nu_1, \nu_2)$

Επομένως, η τ.μ.

$$\frac{\dfrac{\sum(\hat{Y}_i - \bar{Y})^2}{\sigma^2}}{\dfrac{\sum(Y_i - \hat{Y}_i)^2}{\sigma^2(n-2)}} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\hat{\sigma}^2} = \frac{MSR}{MSE} \sim F_{(1,n-2)}$$

Αλλά

$$\sum(\hat{Y}_i - \bar{Y})^2 = \sum(\hat{\beta}_0 - \hat{\beta}_1 X_i - \bar{Y})^2 = \sum(\bar{Y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 X_i - \bar{Y})^2$$

$$= \hat{\beta}_1^2 \sum(X_i - \bar{X})^2$$

Δηλαδή

$$F = \frac{\hat{\beta}_1^2 \sum(X_i - \bar{X})^2}{\hat{\sigma}^2} \sim F_{(1,n-2)}$$

Η $H_0 : \beta_1 = 0$ απορρίπτεται για μεγάλες παρατηρούμενες τιμές της στατιστικής συνάρτησης

$F = \dfrac{\hat{\beta}_1^2 \sum (X_i - \bar{X})^2}{\hat{\sigma}^2}$, δηλ. η $H_0 : \beta_1 = 0$ απορρίπτεται σε επίπεδο στατιστικής σημαντικότητας α αν $F^* > F_{\alpha(1,n-2)}$

Ακραία τιμή της ελεγχοσυνάρτησης (δηλαδή παρατηρούμενη τιμή στην μακριά δεξιά ουρά της F κατανομής) σημαίνει ότι μεγάλο μέρος της συνολικής μεταβλητότητας της απαντητικής μεταβλητής εξηγείται από τη γραμμική παλινδρόμηση και αντίστοιχα μικρό μέρος μένει ανερμήνευτο και αποδίδεται στα τυχαία σφάλματα. Αυτό, φυσικά, αποτελεί ένδειξη ότι υφίσταται η γραμμική σχέση ανάμεσα στην απαντητική και την επεξηγηματική μεταβλητή και άρα πρέπει να απορριφθεί η H_0.

Σχέση ανάμεσα στον έλεγχο t και στον έλεγχο F

Ο έλεγχος υποθέσεων

$$\begin{aligned} H_0 : &\quad \beta_1 = 0 \\ H_1 : &\quad \beta_1 \neq 0 \end{aligned}$$

μπορεί να βασιστεί στη στατιστική συνάρτηση $T = \dfrac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)} \sim t_{(n-2)}$, όπου $s(\hat{\beta}_1)^2 = \dfrac{\hat{\sigma}^2}{\sum (X_i - \bar{X})^2}$.

Κάτω από την $H_0 : \dfrac{\hat{\beta}_1}{s(\hat{\beta}_1)} \sim t_{(n-2)}$. Απορρίπτουμε την H_0 : σε ε.σ.σ. α αν $\left| \dfrac{\hat{\beta}_1}{s(\hat{\beta}_1)} \right| > t_{a/2(n-2)}$

Επίσης, κάτω από την H_0 είναι

$$T^2 = \frac{\hat{\beta}_1^2}{s(\hat{\beta}_1)^2} = \frac{\hat{\beta}_1^2}{\dfrac{\hat{\sigma}^2}{\sum (X_i - \bar{X})^2}} = \frac{\hat{\beta}_1^2 \sum (X_i - \bar{X})^2}{\hat{\sigma}^2} = F$$

Δηλαδή, στο απλό γραμμικό μοντέλο, ο έλεγχος ANOVA (F-test) είναι ισοδύναμος με τον έλεγχο στατιστικής σημαντικότητας του συντελεστή κλίσης (t-test). Το F-test, ωστόσο γενικεύεται και μπορεί να χρησιμοποιηθεί και στην πολλαπλή παλινδρόμηση για τον συνολικό έλεγχο γραμμικής σχέσης ανάμεσα στην απαντητική μεταβλητή και σε έναν αριθμό από επεξηγηματικές μεταβλητές. Από την άλλη πλευρά, το t-test μπορεί να χρησιμοποιηθεί γενικά ως αμφίπλευρος ή μονόπλευρος έλεγχος για τους επι μέρους συντελεστές ενός γραμμικού μοντέλου.

Πίνακας $ANOVA$ για την Πολλαπλή Παλινδρόμηση

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$, $p+1$ παράμετροι.

Πηγή Μεταβλητότητας	SS	$d.f.$	MS	F	$p-value$
Παλινδρόμηση	SSR	p	SSR/p	MSR/MSE	
Σφάλματα	SSE	$n-p-1$	$SSE/n-p-1$		
Σύνολο	SST	$n-1$			

$p-value$: η πιθανότητα μια τ.μ.$\sim F_{(p,n-p-1)}$ να πάρει τιμή τόσο ακραία ή περισσότερο ακραία από F^*.

Έλεγχος: $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$, Δεν υπάρχει γραμμική σχέση.

$F = \dfrac{MSR}{MSE} \sim F(p, n-p-1)$. Απορρίπτουμε την H_0 σε επίπεδο στατιστικής σημαντικότητας α αν $F^* > F_\alpha(p, n-p-1)$ (ή ισοδύναμα αν $p-value < \alpha$).

Συντελεστής Προσδιορισμού

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

Το R^2 εκφράζει το ποσοστό της συνολικής μεταβλητότητας των Y_i που οφείλεται στην παλινδρόμηση. Παίρνει τιμές στο (0,1).
Είναι ένα μέτρο καλής προσαρμογής του μοντέλου (measure of goodness of *model fit*).

Προσαρμοσμένος Συντελεστής Προσδιορισμού

$$R^2_{adj} = 1 - \Big(\frac{n-1}{n-p-1}\Big)\frac{SSE}{SST} < R^2$$

Ο R^2_{adj} λαμβάνει υπόψη του και το πλήθος των άγνωστων παραμέτρων του μοντέλου σε συνδυασμό με το πλήθος των παρατηρήσεων (διόρθωση σε σχέση με το R^2).

!Αν προσθέσουμε ανεξάρτητες μεταβλητές στο μοντέλο το R^2 πάντα αυξάνει ενώ το R^2_{adj} όχι απαραίτητα.

Πολυσυγγραμικότητα

Το πρόβλημα της πολυσυγγραμμικότητας παρουσιάζεται όταν οι ερμηνευτικές μεταβλητές δεν είναι γραμμικώς ανεξάρτητες.

Έστω το γραμμικό μοντέλο

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \ (1)$$

όπου για κάποια $\lambda_1 \neq \lambda_2 \neq 0$ ισχύει

$$\lambda_1 X_{i1} + \lambda_2 X_{i2} = 0 \Rightarrow X_{i1} = cX_{i2}, c = -\frac{\lambda_2}{\lambda_1}$$

Τότε το μοντέλο (1) μπορεί να γραφτεί ως

$$Y_i = \beta_0 + (\beta_1 c + \beta_2) X_{i2} + \varepsilon_i$$

και δεν μπορούμε να ξεχωρίσουμε τη συμβολή κάθε μεταβλητής X_{ij} στην ερμηνεία του Y_i.

Στην πράξη, το πρόβλημα της πολυσυγγραμμικότητας παρουσιάζεται όταν η δύο (ή περισσότερες) επεξηγηματικές μεταβλητές είναι πολύ συσχετισμένες (δηλαδή όταν η δειγματική τους συσχέτιση είναι, κατά απόλυτη τιμή, πολύ κοντά στο 1). Σε αυτή την περίπτωση, αν εκτιμήσουμε το μοντέλο (1), οι εκτιμήσεις δεν θα είναι στατιστικά σημαντικές και θα έχουν μεγάλη διασπορά.

Τι κάνουμε?

1)Αφαιρούμε 1 από τις συσχετισμένες μεταβλητές από την ανάλυση

2)Αν έχουμε πάρα πολλές ανεξάρτητες μεταβλητές αρκετές από τις οποίες συσχετισμένες $\longrightarrow PCA$