Γραμμικά Μοντέλα
Το Απλό Γραμμικό Μοντέλο

Διδάσκουσα: Λουκία Μελιγκοτσίδου
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
Τμήμα Μαθηματικών

March 29, 2020

Γραμμική Παλινδρόμηση

Έστω δύο μεταβλητές X και Y. Μια συναρτησιακή σχέση μεταξύ των μεταβλητών της μορφής

$$Y = f(X)$$

είναι μια ντετερμινιστική σχέση (*deterministic relationship*), που σημαίνει ότι η τιμή της X καθορίζει πλήρως την τιμή της Y.
Για παράδειγμα, $Y = \beta_0 + \beta_1 X$, γραμμική σχέση.

Η στατιστική σχέση μεταξύ δυο μεταβλητών είναι της μορφής

$$Y = f(X) + \varepsilon,$$

όπου ε τυχαίος (στοχαστικός) όρος. Η σχέση αυτή είναι στοχαστική (*stochastic relationship*). Η τυχαία μεταβλητή Y εξαρτάται από την μεταβλητή X (η οποία έχει προκαθορισμένες τιμές), αλλά και από κάποιους μη μετρήσιμους παράγοντες που συνοψίζονται στον στοχαστικό όρο ε.

Για παράδειγμα, $Y = \beta_0 + \beta_1 X + \varepsilon$, απλή γραμμική παλινδρόμηση(*regression*) η απλό γραμμικό μοντέλο (*simple linear model*).

Έχοντας παρατηρήσει δείγμα ζευγών $(X_i, Y_i), i = 1, \ldots, n$, για τα οποία υποθέτουμε ότι ακολουθούν το μοντέλο γραμμικής παλινδρόμησης,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

σκοπός μας είναι η εκτίμηση και γενικά η στατιστική συμπερασματολογία για τις άγνωστες παραμέτρους β_0, β_1 χρησιμοποιώντας το δείγμα (X_i, Y_i).

Στο απλό γραμμικό μοντέλο η Y είναι η εξαρτημένη μεταβλητή (*dependent or response*) και η X είναι η ανεξάρτητη μεταβλητή (*independent or predictor*)
Τα ε_i ονομάζονται τυχαία σφάλματα.
Σύμφωνα με τη μοντελοποίηση, η Y είναι τ.μ. ενώ η X όχι.

Υποθέσεις για τα τυχαία σφάλματα:

- $E(\varepsilon_i) = 0$, σφάλματα με μηδενική μέση τιμή.
- $V(\varepsilon_i) = \sigma^2$, ομοσκεδαστικότητα (ίση διασπορά).
- $Cov(\varepsilon_i, \varepsilon_j) = 0$, ασυσχέτιστα τυχαία σφάλματα (το σφάλμα σε οποιαδήποτε δοκιμή δεν επηρεάζει τα σφάλματα άλλων δοκιμών).

Οι υποθέσεις για τους τυχαίους όρους οδηγούν σε υποθέσεις για τα Y_i. Έχουμε, λοιπόν,

- $E(Y_i) = \beta_0 + \beta_1 X_i$, $V(Y_i) = \sigma^2$, $Cov(Y_i, Y_j) = 0$

Η γραμμή παλινδρόμησης δίνει την αναμενόμενη τιμή της Y για κάθε τιμή της X.

Απλό γραμμικό μοντέλο

Απλό σημαίνει ότι υπάρχει μια μόνο ανεξάρτητη μεταβλητή.

Γραμμικό σημαίνει γραμμικό ως προς τις παραμέτρους.

Το υπόδειγμα $Y_i = \beta_0 + \beta_1 X_i^2 + \varepsilon_i$ είναι γραμμικό, ενώ το $Y_i = \beta_1^{\beta_1} X_i + \varepsilon_i$ όχι.

Ερμηνεία των Παραμέτρων της Παλινδρόμησης

β_0: είναι το σημείο όπου η ευθεία τέμνει τον άξονα των Y, δηλαδή αντιστοιχεί στην αναμενόμενη τιμή του Y για $X = 0$

β_1: είναι η κλίση της ευθείας και αντιπροσωπεύει την μεταβολή (αύξηση ή μείωση) στην αναμενόμενη τιμή του Y που αντιστοιχεί σε αύξηση του X κατά μια μονάδα.

Εκτίμηση παραμέτρων με τη Μέθοδο Ελαχίστων Τετραγώνω

Η Μ.Ε.Τ. στοχεύει στον προσδιορισμό της γραμμής παλινδρόμησης έτσι ώστε να ελαχιστοποιηθούν συνολικά οι αποκλίσεις των σημείων (που αντιστοιχούν στα ζεύγη (X_i, Y_i)) από την ευθεία (ελαχιστοποίηση των σφαλμάτων).

Έχουμε $\varepsilon_i = Y_i - E(Y_i) = Y_i - (\beta_0 + \beta_1 X_i)$.
Επειδή $E(\varepsilon_i) = 0$ δεν εξετάζουμε την ποσότητα $\sum_{i=1}^n \varepsilon_i$ (η οποία θα είναι ίση με 0), αλλά παίρνουμε το άθροισμα των τετραγώνων

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

Οι εκτιμήτριες των β_0, β_1 προκύπτουν από την ελαχιστοποίηση του Q.

$$\begin{cases} \dfrac{dQ}{d\beta_0} = -2\displaystyle\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) = 0 \\ \dfrac{dQ}{d\beta_1} = -2\displaystyle\sum_{i=1}^{n}X_i(Y_i - \beta_0 - \beta_1 X_i) = 0 \end{cases} \Rightarrow$$

$$\begin{cases} \displaystyle\sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i \\ \displaystyle\sum_{i=1}^{n} X_i Y_i = \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2 \end{cases} \quad \text{Κανονικές Εξισώσεις}$$

Λύνοντας ως προς β_0 και β_1 έχουμε

$$\hat{\beta}_1 = \frac{\displaystyle\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}} = \frac{\displaystyle\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \frac{1}{n}\left[\sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i\right] = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Εναλλακτική μορφή του απλού γραμμικού μοντέλου

$$Y_i = \beta_0^* + \beta_1(X_i - \bar{X}) + \varepsilon_i, \text{ όπου } \beta_0^* = \beta_0 + \beta_1 \bar{X}$$
$$\text{ή } Y_i = \beta_0^* + \beta_1 \widetilde{X_i} + \varepsilon_i \text{ όπου } \widetilde{X_i} = X_i - \bar{X}$$

Η εκτιμήτρια του β_1 είναι η ίδια με αυτή της αρχικής εκδοχής του απλού γραμμικού μοντέλου.
Για το β_0^* είναι: $\hat{\beta}_0{}^* = \hat{\beta}_0 + \hat{\beta}_1 \bar{X} = \bar{Y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 \bar{X} = \bar{Y}$

Θεώρημα. Τα $\hat{\beta}_0$ και $\hat{\beta}_1$ είναι γραμμικοί συνδυασμοί των Y_i.

Απόδειξη. Θα δείξουμε ότι η $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$ μπορεί να γραφτεί ως $\hat{\beta}_1 = \sum k_i Y_i$, όπου $k_i = \dfrac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$.

Και επειδή τα X_i είναι γνωστές σταθερές και τα k_i θα είναι γνωστές σταθερές και άρα το $\hat{\beta}_1$ είναι γραμμικός συνδυασμός των Y_i.

Έχουμε

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n}(X_i - \bar{X})Y_i - \sum_{i=1}^{n}(X_i - \bar{X})\bar{Y}$$

$$= \sum_{i=1}^{n}(X_i - \bar{X})Y_i - \bar{Y}\sum_{i=1}^{n}(X_i - \bar{X}) = \sum_{i=1}^{n}(X_i - \bar{X})Y_i$$

Άρα $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \sum k_i Y_i$

Προφανώς και $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \frac{1}{n}\sum Y_i - \bar{X}\sum k_i Y_i$ είναι γραμμικός συνδυασμός των Y_i.

Ιδιότητες των ποσοτήτων k_i

- $\sum k_i = 0$, γιατί $\dfrac{\sum_{i=1}^{n}(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \dfrac{0}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = 0$
- $\sum k_i X_i = 1$, γιατί $\sum k_i X_i = \sum \dfrac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} X_i =$
  $$\sum \frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} X_i - \sum \frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \bar{X} = \frac{\sum (X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = 1$$
- $\sum k_i^2 = \dfrac{1}{\sum (X_i - \bar{X})^2}$,

  γιατί $\sum k_i^2 = \sum \left[\dfrac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2}\right]^2 = \dfrac{\sum (X_i - \bar{X})^2}{(\sum (X_i - \bar{X})^2)^2} = \dfrac{1}{\sum (X_i - \bar{X})^2}$

Θεώρημα των $Gauss - Markov$. Για το απλό γραμμικό μοντέλο οι εκτιμήτριες ελαχίστων τετραγώνων $\hat{\beta}_0$ $\hat{\beta}_1$
1) είναι αμερόληπτες
2) έχουν ελάχιστη διασπορά μεταξύ των αμερόληπτων εκτιμητριών που είναι γραμμικές συναρτήσεις των Y_i.

Απόδειξη

- Αμεροληψία της $\hat{\beta}_1$: Θέλουμε να δείξουμε ότι $E(\hat{\beta}_1) = \beta_1$.

  $$E(\hat{\beta}_1) = E(\sum k_i Y_i) = \sum k_i E(Y_i) = \sum k_i(\beta_0+\beta_1 X_i) = \beta_0 \underbrace{\sum k_i}_{0} + \beta_1 \underbrace{\sum k_i X_i}_{1} = \beta_1$$

- Έστω ότι όλες οι αμερόληπτες εκτιμήτριες του β_1 που είναι γραμμικές συναρτήσεις των Y_i είναι της μορφής
  $$b_1 = \sum c_i Y_i,$$
  όπου c_i αυθαίρετες σταθερές. Επειδή έχουμε αμεροληψία:

  $$E(b_i) = \beta_i \Rightarrow E(\sum c_i Y_i) = \sum c_i(\beta_0 + \beta_1 X_i) = \beta_0 \sum c_i + \beta_1 \sum c_i X_i = \beta_1.$$

  Άρα πρέπει $\sum c_i = 0$ και $\sum c_i X_i = 1$.

  Η διασπορά του b_1 είναι
  $$V(b_1) = V(\sum c_i Y_i) = \sum c_i^2 V(Y_i) = \sum c_i^2 \sigma^2 = \sigma^2 \sum c_i^2,$$
  αφού $Cov(Y_i, Y_j) = 0$.

  Έστω ότι τα c_i έχουν τη μορφή $c_i = k_i + d_i$ όπου τα k_i είναι όπως ορίστηκαν στην εκτιμήτρια $\hat{\beta}_1 = \sum k_i Y_i$ και τα d_i είναι αυθαίρετες σταθερές.

  Συνεπώς
  $$V(b_1) = \sigma^2 \sum c_i^2 = \sigma^2 \sum (k_i + d_i)^2$$
  $$= \sigma^2 \Big[\sum k_i^2 + \sum d_i^2 + 2 \sum k_i d_i \Big]$$

$$= \underbrace{\sigma^2 \sum k_i^2}_{V(\hat{\beta}_1)} + \sigma^2 \sum d_i^2 + 2\sigma^2 \sum k_i d_i$$

Έχουμε $\sum k_i = 0$ και $\sum c_i = \sum(k_i + d_i) = 0 \Rightarrow \sum d_i = 0$
$\sum k_i X_i = 1$ και $\sum c_i X_i = \sum(k_i + d_i)X_i = 1 \Rightarrow \sum d_i X_i = 0$.

Είναι $\sum k_i d_i = \dfrac{\sum(X_i - \bar{X})d_i}{\sum_{i=1}^n (X_i - \bar{X})^2} = \dfrac{\sum X_i d_i}{\sum_{i=1}^n (X_i - \bar{X})^2} - \bar{X}\dfrac{\sum d_i}{\sum_{i=1}^n (X_i - \bar{X})^2} = 0$,
οπότε $V(b_1) = V(\hat{\beta}_1) + \sigma^2 \sum d_i^2$.

Η ποσότητα $\sigma^2 \sum d_i^2$ ελαχιστοποιείται για $\sum d_i^2 = 0$. Άρα η διασπορά του b_1 είναι ελάχιστη όταν $\sum d_i^2 = 0. \Leftrightarrow d_i = 0\ \forall i$, δηλαδή $c_i = k_i, \forall i$.
Συνεπώς η εκτιμήτρια των ελαχίστων τετραγώνων (ε.ε.τ.), $\hat{\beta}_1$, έχει την ελάχιστη διασπορά μεταξύ των αμερόληπτων γραμμικών εκτιμητριών.

- Αμεροληψία της $\hat{\beta}_0$:

$$E(\hat{\beta}_0) = E(\bar{Y} - \hat{\beta}_1\bar{X}) = E\Big[\frac{1}{n}(\sum_{i=1}^n Y_i - \hat{\beta}_1 \sum_{i=1}^n X_i)\Big] =$$

$$= \frac{1}{n}\Big[\sum_{i=1}^n E(Y_i) - (\sum_{i=1}^n X_i)E(\hat{\beta}_1)\Big] =$$

$$= \frac{1}{n}\Big[\sum(\beta_0 + \beta_1 X_i) - \beta_1 \sum X_i\Big] =$$

$$= \frac{1}{n}\Big[n\beta_0 + \beta_1 \sum X_i - \beta_1 \sum X_i\Big] = \beta_0.$$

Παρατήρηση: Τα X_i δεν είναι τυχαίες μεταβλητές.
Τα Y_i είναι τυχαίες μεταβλητές, ανεξάρτητες αλλά <u>όχι ισόνομες</u> (έχουν διαφορετικές αναμενόμενες τιμές και κοινή διακύμανση).

Εκτίμηση του σ^2

- Αν $Y_1, Y_2, \ldots, Y_n$ τυχαίο δείγμα από κατανομή με γνωστό μέσο μ και διασπορά σ^2, τότε η εκτιμήτρια του σ^2 είναι η $\hat{\sigma}^2 = \frac{1}{n}\sum(Y_i - \mu)^2$. Αν ο μέσος μ είναι άγνωστος θα εκτιμηθεί από το $\bar{Y}$ και τότε το σ^2 εκτιμάται από το άθροισμα των τετραγωνικών αποκλίσεων των Y_i από τον κοινό τους μέσο, $S^2 = \frac{1}{n-1}\sum(Y_i - \bar{Y})^2$. Είναι $E(S^2) = \sigma^2$ [αμερόληπτη εκτιμήτρια].

- Στο γραμμικό μοντέλο τα Y_i έχουν διαφορετικές κατανομές που εξαρτώνται από τα X_i. Επομένως, η απόκλιση κάθε παρατήρησης πρέπει να υπολογιστεί από το μέσο της: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$.
  Άρα, αν συμβολίσουμε με $\hat{\varepsilon}_i$ τις εκτιμήσεις των σφαλμάτων (κατάλοιπα-*residuals*), υπολογίζουμε το άθροισμα

  $$\sum \hat{\varepsilon}_i{}^2 = \sum(Y_i - \hat{Y}_i)^2 = \sum(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

  error sum of squares or residuals sum of squares
  άθροισμα τετραγώνων καταλοίπων

  Μια αμερόληπτη εκτιμήτρια του σ^2 είναι η

  $$\hat{\sigma}^2 = \frac{1}{n-2}\sum(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

  mean square error
  μέσο τετραγωνικό σφάλμα

  Διαιρούμε με $n-2$ (β.ε.) καθώς έχουν εκτιμηθεί 2 παράμετροι.