Estadística:
Regressió Lineal II

Adrià Marín Salvador (1456429)
Albert Rodas Barceló (1418925)
Marc Seguí Coll (1465613)
Universitat Autònoma de Barcelona

Maig 2020

Introducció

Sigui $\{(x_i, y_i)\}_{i=1}^n$ un conjunt de n parelles de punts. Podem interpretar aquestes observacions com simplement punts del pla, i trobar la recta que millor els ajusta. Aquest mètode es coneix com a mètode dels mínims quadrats. [1]

Una alternativa a aquest plantejament és entendre que les x són variables deterministes, i que les y són variables aleatòries. El model fonamental per entendre la distribució de les y serà suposar que $y_i = \alpha + \beta x_i + \varepsilon_i$, on els errors ε_i segueixen una distribució $N(0, \sigma^2)$. Podem trobar els estimadors de α i β a partir del mètode de màxima versemblança.

Finalment, podem també considerar que les observacions segueixen una distribució normal bidimensional. Llavors, (x, y) constitueix un vector aleatori. Aquest cas es pot reduir a l'anterior assumint que x i ε són independents.

Així, considerarem en tot l'informe que x és determinista, i que y és una variable aleatòria que segueix $y = \alpha + \beta x_i + \varepsilon_i$, amb ε_i seguint $N(0, \sigma^2)$ i independent de x_i.

Lema previ

Per tal de facilitar els càlculs posteriors, enunciem el següent lema.

Lema. *Siguin $Y_1, Y_2, \ldots, Y_n$ i $X_1, X_2, \ldots, X_m$ variables aleatòries. Siguin $U_1 = \sum_{i=1}^{n} a_i Y_i$ i $U_2 = \sum_{j=1}^{n} b_j X_j$ per a constants $a_1, a_1, \ldots, a_n$ i $b_1, b_2, \ldots, b_m$. Llavors*

$$\mathrm{Cov}(U_1, U_2) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \mathrm{Cov}(Y_i, X_j)$$

La demostració es pot trobar a la pàgina 272 de [2] i no la reproduïm aquí.

Moments de la variable aleatòria $y = \alpha + \beta x + \varepsilon$

Suposem que tenim una mostra de n parells de punts $\{(x_i, y_i)\}_{i=1}^n$. Considerem que les x_i són deterministes i que les y_i són observacions d'una variable aleatòria de la forma $y_i = \alpha + \beta x_i + \varepsilon_i$, amb $\alpha, \beta \in \mathbb{R}$ i on els errors ε_i són independents amb distribució $N(0, \sigma^2)$. Denotarem per y_x la variable y que correspon a al parell (x, y), de manera que escriurem $y_x = \alpha + \beta x + \varepsilon$, amb $\varepsilon \approx N(0, \sigma^2)$.

Llavors, podem calcular els moments següents de y_x,

$$\begin{aligned} E[y_x] &= E[\alpha + \beta x + \varepsilon] = E[\alpha] + E[\beta x] + E[\varepsilon] = \alpha + \beta x \\ V[y_x] &= V[\alpha + \beta x + \varepsilon] = V[\varepsilon] = \sigma^2 \end{aligned} \tag{1}$$

on usem fortament que $\alpha, \beta \in \mathbb{R}$ i que x és determinista, pel que els tres valors són constants reals.

L'estimador $\hat{\beta}$

En l'entrega anterior [3] vam argumentar com l'estimador de β obtingut pel mètode dels mínims quadrats era [4]

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n}\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n}\Big(\sum_{i=1}^n x_i\Big)^2} \tag{2}$$

Veiem que, si definim $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$, podem desenvolupar el següent:

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i^2 + \bar{x}^2 - 2x_i\bar{x}) = \sum_{i=1}^n x_i^2 + n\bar{x}^2 - 2\bar{x}\sum_{i=1}^n x_i = \sum_{i=1}^n x_i^2 - n\bar{x}^2 = \sum_{i=1}^n x_i^2 - \frac{1}{n}\Big(\sum_{i=1}^n x_i\Big)^2$$

que és exactament el denominador de l'expressió 2. També, definint $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$,

$$\begin{aligned} \sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})] &= \sum_{i=1}^n (x_i y_i - \bar{x}y_i - x_i\bar{y} + \bar{x}\bar{y}) = \sum_{i=1}^n x_i y_i - \bar{x}\sum_{i=1}^n y_i - \bar{y}\sum_{i=1}^n x_i + n\bar{x}\bar{y} = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} = \\ &= \sum_{i=1}^n x_i y_i - \frac{1}{n}\sum_{i=1}^n x_i \sum_{i=1}^n y_i \end{aligned}$$

Fixem-nos doncs que es tracta del numerador de l'expressió 2. Així, si usem el fet que $\sum_{i=1}^n (x_i - \bar{x}) = n\bar{x} - n\bar{x} = 0$, la podem reescriure com

$$\hat{\beta} = \frac{\sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})]}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i - \bar{y}\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{3}$$

D'aquesta manera, si definim $\mathrm{SSX} = \sum_{i=1}^n (x_i - \bar{x})^2$ i $c_i = \dfrac{x_i - \bar{x}}{\mathrm{SSX}}$, tenim que

$$\hat{\beta} = \sum_{i=1}^n c_i y_i \tag{4}$$

com volíem veure.

La distribució de l'estimador $\hat{\beta}$

A partir de l'expressió anterior podem trobar ara l'esperança i la variància de $\hat{\beta}$. Calculem primer

$$E[\hat{\beta}] = E\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\text{SSX}}\right] = \frac{1}{\text{SSX}}\sum_{i=1}^{n}E\Big[(x_i-\bar{x})y_i\Big] = \frac{1}{\text{SSX}}\sum_{i=1}^{n}(x_i-\bar{x})E[y_i] =$$
$$= \frac{1}{\text{SSX}}\sum_{i=1}^{n}(x_i-\bar{x})(\alpha+\beta x_i) = \frac{\alpha\sum_{i=1}^{n}(x_i-\bar{x}) + \beta\sum_{i=1}^{n}(x_i-\bar{x})x_i}{\text{SSX}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})x_i}{\text{SSX}}\beta = \beta$$

on usem l'expresió 1 i el fet que

$$\text{SSX} = \sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x}) = \sum_{i=1}^{n}(x_i-\bar{x})x_i - \bar{x}\sum_{i=1}^{n}(x_i-\bar{x}) = \sum_{i=1}^{n}(x_i-\bar{x})x_i$$

Això demostra que $\hat{\beta}$ és un estimador no esbiaixat de β. A continuació, com que $y_1,\ldots,y_n$ són independents, $(x_i-\bar{x})y_1,\ldots,(x_i-\bar{x})y_n$ són independents, podem calcular

$$V[\hat{\beta}] = V\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\text{SSX}}\right] = \frac{1}{\text{SSX}^2}V\Big[\sum_{i=1}^{n}(x_i-\bar{x})y_i\Big] = \frac{1}{\text{SSX}^2}\sum_{i=1}^{n}V[(x_i-\bar{x})y_i] =$$
$$= \frac{1}{\text{SSX}^2}\sum_{i=1}^{n}(x_i-\bar{x})^2V[y_i] = \frac{1}{\text{SSX}^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\sigma^2 = \frac{\sigma^2\text{SSX}}{\text{SSX}^2} = \frac{\sigma^2}{\text{SSX}}$$

Definim $\sigma_\beta^2 \equiv V[\hat{\beta}] = \frac{\sigma^2}{\text{SSX}}$.

Veiem, doncs, que $\hat{\beta}$ és més precís com major sigui SSX. D'aquesta manera, per tal de minimitzar la variància de $\hat{\beta}$ podem augmentar el número d'observacions n o, amb n fixat, triar les x_i el més separades possible, per tal de maximitzar la variància del conjunt $\{x_1,\ldots,x_n\}$. A partir de l'exercici anterior, com que $\hat{\beta}$ és combinació lineal de les y_i, que són normals, es distribueix de manera normal. Havent calculat l'esperança i la variancia, tenim $\hat{\beta} \approx N(\beta,\sigma_\beta^2)$.

L'estimador $\hat{\alpha}$

Ens disposem ara a trobar l'esperança i la variància de $\hat{\alpha}$. Calculem abans

$$\text{Cov}(\hat{\beta},\bar{y}) = \text{Cov}\left(\sum_{i=1}^{n}\frac{y_i}{n},\sum_{i=1}^{n}c_iy_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{c_j}{n}\text{Cov}(y_i,y_j) = \sum_{i=1}^{n}\frac{c_i}{n}\sigma_\beta^2 = \frac{\sigma_\beta^2}{n}\sum_{i=1}^{n}c_i =$$
$$= \frac{\sigma_\beta^2}{n}\sum_{i=1}^{n}\frac{x_i-\bar{x}}{\text{SSX}} = \frac{\sigma_\beta^2}{n\text{SSX}}\sum_{i=1}^{n}(x_i-\bar{x}) = 0$$

on usem que les y_i són independents i el Lema previ.

Ara, havíem vist en l'entrega anterior [3] que $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$. Així, podem calcular

$$E[\hat{\alpha}] = E[\bar{y} - \hat{\beta}\bar{x}] = E[\bar{y}] - \bar{x}E[\hat{\beta}] = E[\frac{1}{n}\sum_{i=1}^{n}(\alpha + \beta x_i + \varepsilon_i)] - \bar{x}\beta = \alpha + \beta\bar{x} - \bar{x}\beta = \alpha$$

pel que $\hat{\alpha}$ és un estimador no esbiaixat de α.

També podem calcular

$$V[\hat{\alpha}] = V[\bar{y} - \hat{\beta}\bar{x}] = V[\bar{y}] + V[\hat{\beta}\bar{x}] - 2\text{Cov}(\bar{y}, \hat{\beta}\bar{x}) = \frac{\sigma^2}{n} + \bar{x}^2 V[\hat{\beta}] - 2\bar{x}\text{Cov}(\bar{y}, \hat{\beta}) = \frac{\sigma^2}{n} + \bar{x}^2\sigma_\beta^2 = \quad (5)$$
$$= \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{\text{SSX}} = \sigma^2\Big(\frac{1}{n} + \frac{\bar{x}^2}{\text{SSX}}\Big)$$

on usem que, com que $\sum_{i=1}^{n}(\alpha + \beta x_i)$ és una constant, les ε_i són independents, i que

$$V[\bar{y}] = V[\frac{1}{n}\sum_{i=1}^{n}(\alpha + \beta x_i + \varepsilon_i)] = \frac{1}{n^2}V[\sum_{i=1}^{n}(\alpha + \beta x_i) + \sum_{i=1}^{n}\varepsilon_i] = \frac{1}{n^2}V[\sum_{i=1}^{n}\varepsilon_i] = \frac{n}{n^2}V[\varepsilon_1] = \frac{\sigma^2}{n}$$

Definim $\sigma_\alpha^2 \equiv V[\hat{\alpha}] = \sigma^2\Big(\frac{1}{n} + \frac{\bar{x}^2}{\text{SSX}}\Big)$. Veiem que, novament, incrementant el número d'observacions i prenent els valors de x més separats, aconseguim millors estimadors per α. Observem també el següent. Amb un número donat de punts n, la variància de $\hat{\alpha}$ serà el més petita possible quan $\bar{x} = 0$. D'aquesta manera, si ens és possible, agafarem punts amb coordenades x positives i coordenades x negatives de manera que $\bar{x} = 0$. Tal i com s'ha comentat anteriorment, per tal de reduir també la variància de $\hat{\beta}$ prendrem aquestes x_i el més separades possible entre elles.

L'estimador $\hat{y_x} = \hat{\alpha} + \hat{\beta}x$

Definim ara $\hat{y_x} = \hat{\alpha} + \hat{\beta}x$. Així, calculem $E[\hat{y_x}] = E[\hat{\alpha} + \hat{\beta}x] = E[\hat{\alpha}] + E[\hat{\beta}x] = E[\hat{\alpha}] + xE[\hat{\beta}] = \alpha + x\beta$.

A continuació, calculem

$$\text{Cov}(\hat{\alpha}, \hat{\beta}) = \text{Cov}(\bar{y} - \bar{x}\hat{\beta}, \hat{\beta}) = \text{Cov}(\bar{y}, \hat{\beta}) - \bar{x}\text{Cov}(\hat{\beta}, \hat{\beta}) = -\bar{x}\sigma_\beta^2$$

usant també el Lema previ.

Amb aquesta informació podem ara calcular

$$V[\hat{y_x}] = V[\hat{\alpha} + \hat{\beta}x] = V[\hat{\alpha}] + V[\hat{\beta}x] + 2\text{Cov}(\hat{\alpha}, \hat{\beta}x) = \sigma_\alpha^2 + x^2\sigma_\beta^2 - 2x\bar{x}\sigma_\beta^2 = \sigma_\alpha^2 + x\sigma_\beta^2(x - 2\bar{x}) =$$
$$= \sigma^2\Big(\frac{1}{n} + \frac{\bar{x}}{\text{SSX}}\Big) + \frac{x\sigma^2}{\text{SSX}}(x - 2\bar{x}) = \sigma^2\Big(\frac{1}{n} + \frac{\bar{x} + x^2 - 2x\bar{x}}{\text{SSX}}\Big) = \sigma^2\Big(\frac{1}{n} + \frac{(x - \bar{x})^2}{\text{SSX}}\Big)$$

Definint $n_x = \Big(\frac{1}{n} + \frac{(x-\bar{x})^2}{\text{SSX}}\Big)^{-1}$, obtenim que $V[\hat{y_x}] = \frac{\sigma^2}{n_x}$. Veiem que els valors y que podem estimar amb més precisió són els corresponents a les x més properes a $\bar{x}$. Així, els valors predits per la regressió lineal són més fiables per les x centrals (d'entre les nostres dades) que pels extrems o les extrapolacions.

El valor de $V[\hat{y_x}]$ també el podríem haver trobat amb el procediment següent. Adonem-nos que $\hat{\alpha}$ no és res més que $\hat{y_0}$. Movent l'eix d'ordenades podem aconseguir que qualsevol valor x sigui 0 sense afectar aquesta translació al valor SSX. Així doncs, donat un valor x_k movem els eixos de manera que $x_k = 0$, i utilitzant l'Expressió 5, deduïm que $V[\hat{y_x}] = \sigma^2(\frac{1}{n} + \frac{\bar{x}'^2}{\text{SXX}})$, on $\bar{x}'^2$ és la mitjana dels valors que pren x respecte els nous eixos, desplaçats x_k unitats. Per la linealitat de la mitjana tenim que $\bar{x}' = \bar{x} - x_k$, recuperant la mateixa expressió que amb el procediment anterior.

Ampliació. Un embolcall de confiança

Fem notar el següent fet, que ens sembla una continuació natural del que s'ha fet fins aquí. Fixem x qualsevol tal que $x \notin \{x_1, \ldots, x_n\}$ i considerem que volem estimar el valor real y_x a partir de l'estimador $\hat{y}_x$. Definim $e = y_x - \hat{y}_x$. Llavors, tenim $E[e] = E[y_x] - E[\hat{y}_x] = \alpha + \beta x - \alpha - \beta x = 0$. També, com que y_x és un valor futur que no és utilitzat en l'estimador $\hat{y}_x$, es té que y_x i $\hat{y}_x$ són independents, i $V[e] = V[y_x] + V[\hat{y}_x] = \sigma^2(1 + \frac{1}{n} + \frac{\bar{x}'^2}{\mathrm{SXX}})$. D'aquesta manera,

$$Z = \frac{y_x - \hat{y}_x}{\sigma\sqrt{1 + \frac{1}{n} + \frac{\bar{x}^2}{\mathrm{SXX}}}}$$

es distribueix com una $N(0,1)$. Utilitzant l'estimador de σ^2 $S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, es pot demostrar (pàgina 582 de [2]) que $\frac{(n-2)S^2}{\sigma^2} \approx \chi^2_{n-2}$. Per tant,

$$T = \frac{y_x - \hat{y}_x}{S\sqrt{1 + \frac{1}{n} + \frac{\bar{x}^2}{\mathrm{SXX}}}}$$

segueix una distribució $t-$Student amb $n-2$ graus de llibertat.

D'aquesta manera, el següent interval de predicció conté y_x amb una probabilitat de $1-\delta$

$$\left(\hat{\alpha} + \hat{\beta}x - t_{\delta/2}\sqrt{1 + \frac{1}{n} + \frac{\bar{x}^2}{\mathrm{SXX}}},\ \hat{\alpha} + \hat{\beta}x + t_{\delta/2}\sqrt{1 + \frac{1}{n} + \frac{\bar{x}^2}{\mathrm{SXX}}}\right)$$

Per tant, fixant δ, podem representar al pla la recta $y = \hat{\alpha} + \hat{\beta}x$ i l'embolcall determinat entre les funcions $\hat{\alpha} + \hat{\beta}x + t_{\delta/2}\sqrt{1 + \frac{1}{n} + \frac{\bar{x}^2}{\mathrm{SXX}}}$ i $\hat{\alpha} + \hat{\beta}x - t_{\delta/2}\sqrt{1 + \frac{1}{n} + \frac{\bar{x}^2}{\mathrm{SXX}}}$. Si fixem x i mesurem el seu valor corresponent y_x, aquest es trobarà dins de l'embolcall definit amb una probabilitat de $1-\delta$. Tal i com s'ha argumentat anteriorment, aquest embolcall serà el màxim d'estret a $\bar{x}$ i s'anirà eixamplant a mesura que ens n'allunyem.

Referències

[1] J. del Castillo, *Regressió tres models en un* (2020)

[2] D. Wackerly, W. Mendenhall, R. Scheaffer, *Mathematical Statistics with Applications. 7th edition* ISBN-13: 978-0-495-38508-0

[3] J. del Castillo, *Regressió* (2020)

[4] A. Marín, A. Rodas, M. Seguí, *Estadística: Regressió lineal* (2020)