Estadística:
Regressió Lineal

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

Introducció

Si realitzem una sèrie d'observacions sobre dues variables aleatòries X i Y, ens podem preguntar si estan relacionades d'alguna manera observant la mostra resultant que hem obtingut. Podem anar més enllà i tractar de determinar explícitament la relació que existeix entre aquestes variables.

Suposarem doncs que tenim una mostra $\{(x_i, y_i)\}_{i=1}^n$ obtinguda a partir d'observacions de les variables X i Y. Com a primer intent, la forma més senzilla de relacionar les dades de la mostra és de forma lineal, és a dir, suposar que $y_i = \alpha + \beta x_i$ per a certs $\alpha, \beta \in \mathbb{R}$. Notem que si realitzem una altra observació i obtenim x_{n+1}, aleshores si coneixem α i β podrem predir el valor d'y_{n+1}.

Tot i això, a la pràctica quasi es donarà la igualtat degut que estem tractant amb variables aleatòries. És per això que l'expressió ens definim l'error en l'observació i-èssima com seria $\varepsilon_i = y_i - \alpha - \beta x_i$ de forma que $y_i = \alpha + \beta x_i + \varepsilon_i$. Llavors la qüestió està en trobar quins α i β fan que la nostra predicció tingui un error el més petit possible.

Com que intentem determinar la correlació lineal de $\{(x_i, y_i)\}_{i=1}^n$, considerarem només casos on no totes les x_i siguin iguals. En cas contrari, l'estudi perd el sentit.

Minimització de la suma dels errors quadràtics

Una forma d'atacar el problema es tractant de minimitzar la suma dels errors quadràtics. Definim doncs:

$$\mathrm{SS}(\alpha, \beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \tag{1}$$

Definim també, per comoditat, els següents estadístics[1]:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i, \quad \overline{y^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2$$

[1] Tot i que sembli natural emprar la mitjana aritmètica, el seu ús ha anat canviant al llarg de la història. Ja al segle V a.C. els grecs la manipulaven com si fos una eina més per visualitzar les dades. Fins al segle XVI la mitjana aritmètica es pensava com el punt intermig entre dos nombres. Durant aquest segle, es va reconèixer l'extensió de la mitjana aritmètica a n nombres i els astrònoms van començar a tractar amb ella per tal de reduir errors en les mesures. Fins al 1800 la mitjana aritmètica era un medi per arribar a un cert objectiu. Quetelet (1796-1874) va ser un dels primers científics en utilitzar la mitjana aritmètica com un valor representatiu [1].

Volem solucionar el sistema

$$\begin{cases} \dfrac{\partial \mathrm{SS}}{\partial \alpha}(\hat{\alpha},\hat{\beta}) = -2\displaystyle\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i) = 0 \\ \dfrac{\partial \mathrm{SS}}{\partial \beta}(\hat{\alpha},\hat{\beta}) = -2\displaystyle\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)x_i = 0 \end{cases}$$

De la primera eqaució obtenim que $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$. Si a la segona substituim $\hat{\alpha}$ pel que acabem de trobar ens queda que $\hat{\beta} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}$. D'aquesta manera, els valors que satisfan el sistema són

$$\hat{\alpha} = \frac{\bar{y}\overline{x^2} - \bar{x}\overline{xy}}{\overline{x^2} - \bar{x}^2}, \quad \hat{\beta} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} \tag{2}$$

Fem notar que, com que almenys un dels x_i és diferent dels altres, els denominadors de les expressions anteriors són no nuls.

Per demostrar que la solució que hem trobat és vertaderament un mínim de SS, calculem la matriu Hessiana de SS en el punt $(\hat{\alpha},\hat{\beta})$

$$\mathrm{H}_{\mathrm{SS}}(\hat{\alpha},\hat{\beta}) = \begin{pmatrix} \dfrac{\partial^2 \mathrm{SS}}{\partial \alpha^2}(\hat{\alpha},\hat{\beta}) & \dfrac{\partial^2 \mathrm{SS}}{\partial \alpha \partial \beta}(\hat{\alpha},\hat{\beta}) \\ \dfrac{\partial^2 \mathrm{SS}}{\partial \beta \partial \alpha}(\hat{\alpha},\hat{\beta}) & \dfrac{\partial^2 \mathrm{SS}}{\partial \beta^2}(\hat{\alpha},\hat{\beta}) \end{pmatrix} = \begin{pmatrix} 2n & 2n\bar{x} \\ 2n\bar{x} & 2n\overline{x^2} \end{pmatrix}$$

que té determinant $4n^2\left(\overline{x^2} - \bar{x}^2\right) = 4n\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2 \geq 0$. Notem que per tal que la igualtat es compleixi, necessàriament $x_i - \bar{x} = 0$ per a tot $i \in \{1,\ldots,n\}$. En aquest cas doncs $x_i = x_j$ per a tots $i,j \in \{1,\ldots,n\}$, situació que no considerem. D'aquesta manera en ser el determinant estrictament positiu, el punt $(\hat{\alpha},\hat{\beta})$ és un mínim de SS.

Fem notar que tant $\hat{\alpha}$ com $\hat{\beta}$ són estadístics de la nostra mostra, és a dir, $\hat{\alpha} = \hat{\alpha}(x_1,\ldots,x_n,y_1,\ldots,y_n)$ i $\hat{\beta} = \hat{\beta}(x_1,\ldots,x_n,y_1,\ldots,y_n)$. Aquesta situació és natural degut que $\mathrm{SS}(\alpha,\beta)$ està definida en termes de la mostra. Si haguéssim obtingut que el mínim de SS és independent de la mostra, aquest fet implicaria que els valors $\hat{\alpha}$ i $\hat{\beta}$ que minimitzen la suma dels errors quadràtics són els mateixos per a cada experiment. Però això no té cap sentit ja que podem considerar dos experiments tals que al primer obtenim una mostra $\{(0,0),(1,1)\}$ i al segon una mostra $\{(0,0),(1,2)\}$. Clarament la recta que ajusta millor el primer experiment és la recta que passa per les dues observacions de la mostra, és a dir, $y = x$. Per al segon experiment la recta és també la que passa pels punts de la mostra, $y = 2x$. Evidentment les rectes són diferents i per tant la recta que millor ajusta cada experiment depèn de la mostra obtinguda.

A partir d'ara, per comoditat, denotarem $\alpha \equiv \hat{\alpha}$ i $\beta \equiv \hat{\beta}$.

Propietats de la recta de regressió lineal

Si considerem el pla cartesià, podem representar la nostra mostra si tractem la observació (x_i,y_i) com un punt al pla. Abusant de notació, anomenarem recta de regressió lineal a la recta $y = \alpha + \beta x$. Notem que minimitzant SS hem obtingut que $\alpha = \bar{y} - \beta\bar{x}$. Per tant ens queda trivialment que $\bar{y} = \alpha + \beta\bar{x}$ fet que demostra que la recta passa, efectivament, pel punt $(\bar{x},\bar{y})$.

Això suggereix que d'alguna manera podem quantificar l'error a partir de $\bar{y}$. Si calculem la suma de les desviacions respecte la mitjana al quadrat:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \tag{3}$$

Notem que aquestes desviacions estan relacionades amb l'error quadràtic i la desviació de les prediccions respecte la mitjana. Fixem-nos que $\overline{\hat{y}} \equiv \frac{1}{n}\sum_{i=1}^{n}\hat{y_i} = \frac{1}{n}\sum_{i=1}^{n}(\alpha+\beta x_i) = \alpha+\beta\bar{x} = \bar{y}-\beta\bar{x}+\beta\bar{x} = \bar{y}$. Si calculem el darrer sumand de l'expressió 3:

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\hat{y_i})(\hat{y_i}-\bar{y}) &= \sum_{i=1}^{n}(y_i\hat{y_i}-\bar{y}y_i-\hat{y_i}^2+\bar{y}\hat{y_i})\\ &= \sum_{i=1}^{n}y_i(\alpha+\beta x_i)-n\bar{y}^2-\sum_{i=1}^{n}(\alpha+\beta x_i)^2+n\bar{y}^2\\ &= n\alpha\bar{y}+n\beta\overline{xy}-n\alpha^2-n\beta^2\overline{x^2}-2n\alpha\beta\bar{x}\\ &= n(\bar{y}^2-\beta\bar{x}\bar{y}+\beta\overline{xy}-\bar{y}^2-\beta^2\bar{x}^2+2\beta\bar{x}\bar{y}-\beta^2\overline{x^2}-2\beta\bar{x}\bar{y}+2\beta^2\bar{x}^2)\\ &= n\beta(\overline{xy}-\bar{x}\bar{y}-\beta[\overline{x^2}-\bar{x}^2])\\ &= 0\end{aligned}$$

Per tant és clar que

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-\hat{y_i})^2+\sum_{i=1}^{n}(\hat{y_i}-\bar{y})^2 \tag{4}$$

Per comoditat denotarem el següent:

$$\text{SST} = \sum_{i=1}^{n}(y_i-\bar{y})^2, \quad \text{SSR} = \sum_{i=1}^{n}(y_i-\hat{y_i})^2, \quad \text{SSE} = \sum_{i=1}^{n}(\hat{y_i}-\bar{y})^2$$

Veiem que SST és la norma al quadrat del vector $(y_1-\bar{y},\dots,y_n-\bar{y})$. SSR és la norma al quadrat del vector $(\hat{y_1}-\bar{y},\dots,\hat{y_n}-\bar{y})$. SSE és la norma al quadrat del vector $(y_1-\hat{y_1},\dots,y_n-\hat{y_n})$. Clarament

$$(y_1-\hat{y_1},\dots,y_n-\hat{y_n})+(\hat{y_1}-\bar{y},\dots,\hat{y_n}-\bar{y}) = (y_1-\bar{y},\dots,y_n-\bar{y})$$

.

Estimació màxim versemblant dels paràmetres de la recta

Podem suposar ara que les variables X són deterministes, i que les variables Y_i són aleatories amb $Y_i = \alpha+\beta x_i+\varepsilon_i$, on els errors ε_i són independents i segueixen una distribució normal $N(0,\sigma^2)$. Llavors, les variables Y_i segueixen una distribució $N(\mu_i,\sigma^2)$, amb $\mu_i = \alpha+\beta x_i$ per a $1\le i\le n$.

Llavors, la funció de densitat de Y_i és $f_{Y_i}(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu_i)^2}{2\sigma^2}}$.

D'aquesta manera, la funció log-versemblança és

$$\hat{l} = -\frac{n}{2}\log(2\pi\sigma^2)-\sum_{i=1}^{n}\frac{(y_i-\mu_i)^2}{2\sigma^2} = -\frac{n}{2}\log(2\pi)-n\log\sigma-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2$$

Volem minimitzar aquesta funció i per tant imposem el següent

$$\begin{cases}\dfrac{\partial\hat{l}}{\partial\alpha}(\hat{\alpha},\hat{\beta},\hat{\sigma}) = \dfrac{1}{\hat{\sigma}^2}\displaystyle\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i) = 0\\ \dfrac{\partial\hat{l}}{\partial\beta}(\hat{\alpha},\hat{\beta},\hat{\sigma}) = \dfrac{1}{\hat{\sigma}^2}\displaystyle\sum_{i=1}^{n}x_i(y_i-\hat{\alpha}-\hat{\beta}x_i) = \dfrac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(y_i-\bar{y}-\hat{\beta}\bar{x}-\hat{\beta}x_i) = 0\\ \dfrac{\partial\hat{l}}{\partial\sigma}(\hat{\alpha},\hat{\beta},\hat{\sigma}) = -\dfrac{n}{\hat{\sigma}}+\dfrac{1}{\hat{\sigma}^3}\displaystyle\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)^2 = 0\end{cases}$$

Del sistema anterior es dedueix que $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ i $\hat{\beta} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}$ ja que les dues primeres equacions són equivalents a les usades en la minimització de SS. Si substituim aquests valors a la tercera equació ens queda finalment que

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2}{n}}$$

Veiem, doncs, que els estimadors obtinguts per a α i β pel mètode de màxima versemblança assumint un comportament normal dels errors són els mateixos que obtenim amb el mètode dels mínims quadrats. Això implica que, en efecte, si assumim que els errors ε_i són independents i idènticament distribuïts seguint una normal de mitjana 0, els valors més probables de α i β són els que minimitzen la funció suma de quadrats SS, que no assumeix cap hipòtesi de normalitat.

Podem pensar també en per què assumim que els errors ε_i segueixen una distribució normal. Sembla lògic pensar que la variable aleatòria que ens genera els errors té esperança nul·la, que és simètrica (ja que no esperem cap tenèndcia a sobreestimar Y o a subestimar-la) i que serà regular. Amb aquestes hipòtesis, en virtud del prinipi de Gauss, podem assegurar que els errors seguiran una distribució normal, la qual cosa ens justifica aquesta elecció.

Predicció d'intervals de confiança

Sigui $\underline{x} = \{x_1, \ldots, x_n\}$ una mostra d'una variable aleatòria $X \approx N(\mu, \sigma^2)$. Sigui ara x_{n+1} una nova observació de la variable X.

Per hipòtesi, $x_{n+1} \approx N(\mu, \sigma^2)$. Considerem ara l'estimador de μ, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Per ser suma de normals, tenim que $n\bar{x} \approx N(n\mu, n\sigma^2)$ i $\bar{x} \approx N(\mu, \frac{\sigma^2}{n})$. D'aquesta manera, $x_{n+1} - \bar{x} \approx N(0, \sigma^2 + \frac{\sigma^2}{n})$. Per tant, normalitzant, $\frac{x_{n+1} - \bar{x}}{\sigma\sqrt{1+\frac{1}{n}}} \approx N(0,1)$.

Considerem ara l'estimador de la variàcia $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Pel Teorema de Fisher, sabem que $\frac{(n-1)s^2}{\sigma^2} \approx \chi^2_{n-1}$. D'aquesta manera, $\frac{s}{\sigma} \approx \sqrt{\frac{\chi^2_{n-1}}{n-1}}$.

Per tant, observem que $\frac{\frac{x_{n+1}-\bar{x}}{\sigma\sqrt{1+\frac{1}{n}}}}{\frac{s}{\sigma}} \approx \frac{N(0,1)}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}$. És a dir, per definició, $\frac{x_{n+1}-\bar{x}}{s\sqrt{1+\frac{1}{n}}} \approx t_{n-1}$.

Sigui ara $t_{n-1,0.025}$ el valor que deixa a la seva dreta una àrea de 0.025 sota una $t-$Student de $n-1$ graus de llibertat. L'interval de predictiu de x_{n+1} amb un 95% de confiança és

$$\left(\bar{x} - s \cdot t_{n-1,0.025}\sqrt{1+\frac{1}{n}}, \bar{x} + s \cdot t_{n-1,0.025}\sqrt{1+\frac{1}{n}}\right)$$

Comentaris i futurs punts de recerca

Arribats a aquest punt, creiem que és directe que ens aparegui la pregunta següent, i en particular per als estimadors obtinguts a partir del mètode de mínims quadrats. Si volem usar aquests estimadors per a fer inferències estadístiques, necessitem conèixer les seves propietats estadístiques. En particular, cal conèixer si són estimadors esbiaixats o no, i si són consistents o no. Ens interessa també saber com es distribueixen. L'objectiu d'obtenir aquesta informació és, entre d'altres, poder donar intervals de confiança per a aquests estimadors, o poder, per exemple, realitzar tests d'hipòtesi per determinar si la recta de regressió obtinguda té una ordenada a l'origen consistent amb zero.

Creiem també que és necessari un mètode per determinar, amb una certa confiança, si les dades $\{(x_i, y_i)\}_{i=1}^n$ estan vertaderament relacionades de forma lineal o no. Segons hem pogut investigar, els valors α i β estan estretament relacionats amb el coeficient de correlació de la mostra. Aquest es un estimador del coeficient de correlació de Pearson entre les variables X i Y. En cas que $Y = A + BX$ es té que el coeficient de correlació de Pearson és igual a 1. Per tant podem interpretar que quan més proper a 1 sigui el coeficient de correlació de la mostra, més bo serà l'ajust lineal.

Molts models en què $Y = F(X)$ amb F una funció polinòmica o exponencial es poden estudiar amb aquest mètode aplicant logaritmes a cada banda de la igualtat, obtenint una relació lineal. Ens plantegem llavors si mitjançant manipulacions algebraiques podem arribar a aplicar la regressió lineal d'alguna manera en casos menys elementals, com per exemple si $Y = \sin(X)$.

Notem que hem obtingut la recta de regressió lineal a partir de la minimització de la suma dels errors quadràtics, que coincideix amb la recta obtinguda en l'estimació màxim versemblant. En els textos llegits es parla d'altres mètodes d'ajust en els quals no es minimitza la suma dels errors quadràtics si no que tracta de minimitzar altres quantitats com la suma dels valors absoluts dels errors. Un altre exemple és l'anomenada regressió de Deming. Seria interessant investigar quina relació hi ha entre els mètodes i en quines situacions s'utilitza cadascun.

Ens preguntem també, i pensant ja en aplicacions pràctiques d'aquest mètode estadístic, com es poden tenir en compte en el mòdel incerteses en les dades $\{(x_i, y_i)\}_{i=1}^n$. Per exemple, quan hom mesura magnituds en un laboratori, aquestes sempre porten associada una incertesa, que pot ser de tipus instrumental o estadístic, o d'ambdós tipus alhora. En cas que, després, es vulgui analitzar la correlació lineal d'aquestes dades i obtenir estimadors per als paràmetres α i β, com es poden introduir aquestes incerteses en el model?

Referències

[1] A. Bakker, Freudenthal Institute, Utrecht University. *The Early History of Average Values and Implications for Education*, Journal of Statistics Education Volume 11, Number 1 (2003) http://jse.amstat.org/v11n1/bakker.html

[2] J. del Castillo, *Estadística. Diapositives de la primera part del curs*, Campus Virtual de la Universitat Autònoma de Barcelona (2020)

[3] J. del Castillo, *Estadística. Diapositives de la segona part del curs*, Campus Virtual de la Universitat Autònoma de Barcelona (2020)

[4] J. del Castillo, *La Campana de Gauss*, MATerials MATemàtics Volum 2011, treball no. 4, 8 pp. ISSN: 1887-1097 Publicació electrònica de divulgació del Departament de Matemàtiques de la Universitat Autònoma de Barcelona (2011) http://mat.uab.cat/matmat/PDFv2011/v2011n04.pdf

[5] L. Orellana, *Regresión Lineal Simple*, Departamento de Matemática. Facultad de ciencias exactas y naturales. Universidad de Buenos Aires (2011)