Sam Farré Baltà (1605576)

Anàlisi de Dades Complexes: Segon Lliurament

Exercici 1.

(a) Núvol de punts mitjançant polinomi de segon grau.

Ajustem el model següent:

$$y_i = \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \,,$$

Per tant la matriu de disseny i el vector de paràmetres son:

$$X = \begin{pmatrix} x_1 & x_1^2 \\ x_2 & x_2^2 \\ \vdots & \vdots \\ x_n & x_n^2 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

(b) Dues rectes de pendent diferent.

Considerant el vector indicador:

$$z_i = \begin{cases} 1 & \text{si } i \in \text{grup 1} \\ 0 & \text{si } i \in \text{grup 2} \end{cases}$$

Tenim el model

$$y_i = \beta_0 + \beta_1 \, x_i \, z_i + \beta_2 \, x_i \, (1 - z_i) \,.$$

Per tant la matriu de disseny i el vector de paràmetres seran:

$$X = \begin{pmatrix} 1 & x_1 \, z_1 & x_1 \, (1 - z_1) \\ 1 & x_2 \, z_2 & x_2 \, (1 - z_2) \\ \vdots & \vdots & \vdots \\ 1 & x_n \, z_n & x_n \, (1 - z_n) \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

Exercici 2.

(a) Càlcul de $\mathbb{E}(\widehat{\beta})$ i $\mathbb{V}(\widehat{\beta})$.

La solució dels mínims quadrats és

$$\widehat{\beta} = (X^T X)^{-1} X^T y \,,$$

amb $y = X\beta + \varepsilon$, per tant:

$$\begin{aligned} \widehat{\beta} &= (X^T X)^{-1} X^T (X\beta + \varepsilon) = \\ &= (X^T X)^{-1} X^T X \beta + (X^T X)^{-1} X^T \varepsilon = \\ &= \beta + (X^T X)^{-1} X^T \varepsilon \end{aligned}$$

Calculem l'esperança de $\widehat{\beta}$:

$$\begin{aligned} \mathbb{E}(\widehat{\beta}) &= \mathbb{E}(\beta + (X^T X)^{-1} X^T \varepsilon) = \\ &= \beta + (X^T X)^{-1} X^T \, \mathbb{E}(\varepsilon) = \boxed{\beta} \end{aligned}$$

Calculem la variància de $\widehat{\beta}$:

$$\begin{aligned}\mathbb{V}(\widehat{\beta}) &= \mathbb{V}(\beta + (X^TX)^{-1}X^T\varepsilon) = \\ &= (X^TX)^{-1}X^T\,\mathbb{V}(\varepsilon)\,X(X^TX)^{-1} = \\ &= (X^TX)^{-1}X^T\,\sigma^2 I\,X(X^TX)^{-1} = \\ &= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} = \\ &= \boxed{\sigma^2(X^TX)^{-1}}\end{aligned}$$

(b) Variància de $\widehat{y}$.

Tenim el següent:

$$\hat{y} = X\widehat{\beta} = X(X^TX)^{-1}X^Ty$$

Per tant la variància de $\hat{y}$ és

$$\begin{aligned}\mathbb{V}(\hat{y}) &= \mathbb{V}(X(X^TX)^{-1}X^Ty) = \\ &= X(X^TX)^{-1}X^T\,\mathbb{V}(y)\,(X(X^TX)^{-1}X^T)^T,\end{aligned}$$

on $\mathbb{V}(y) = \mathbb{V}(X\beta + \varepsilon) = \mathbb{V}(\varepsilon) = \sigma^2 I$, i per tant:

$$\begin{aligned}\mathbb{V}(\hat{y}) &= \sigma^2 X(X^TX)^{-1}X^T(X(X^TX)^{-1}X^T)^T = \\ &= \boxed{\sigma^2 X(X^TX)^{-1}X^T}\end{aligned}$$

(c) Variància de ε.

El vector d'errors és:

$$\begin{aligned}\varepsilon = y - \hat{y} &= y - X\widehat{\beta} = \\ &= y - X(X^TX)^{-1}X^Ty = \\ &= (I - X(X^TX)^{-1}X^T)\,y\end{aligned}$$

Per tant, la seva variància serà:

$$\begin{aligned}\mathbb{V}(\varepsilon) &= \mathbb{V}((I - X(X^TX)^{-1}X^T)\,y) = \\ &= (I - X(X^TX)^{-1}X^T)\,\mathbb{V}(y)\,(I - X(X^TX)^{-1}X^T)^T \\ &= \sigma^2\,(I - X(X^TX)^{-1}X^T)\,(I - X(X^TX)^{-1}X^T)^T \\ &= \boxed{\sigma^2\,(I - X(X^TX)^{-1}X^T)}\end{aligned}$$

Exercici 3.

Aquest exercici està fet mitjançant un *script* d'R (veure `lliur2.r`), excepte l'últim apartat.

(a) Funció *link*.

Escollim l'enllaç d'identitat, ja que s'especifica que segueix una distribució amb mitjana $\beta_1 x_i$, és a dir, una relació directa. Provant el model amb `link = "identity"` obtenim els mateixos resultats que la sortida proporcionada.

(b) IC del 95% per β_1.

El verdader valor de β_1 es troba en l'interval $(1.719, 3.913)$ amb un 95% de confiança.

(c) Aberracions cromosòmiques per $x = 4$.

Per a una dosi $x = 4$, s'esperen 10.5 aberracions cromosòmiques en mitjana. El valor real es troba en l'interval $(6.009, 14.991)$ amb un 95% de confiança.

(d) Funció log-versemblança i estimador de β_1.

La funció log-versemblança és

$$l_i(\beta_1 \mid y) = \sum_{i=1}^{n} (y_i \log(\mu_i(\beta_1)) - \mu_i(\beta_1) - \log(y_i!))$$
$$= \boxed{\sum_{i=1}^{n} (y_i \log(\beta_1 x_i) - \beta_1 x_i - \log(y_i!))}\,.$$

Anem a maximitzar-la. Si la derivem respecte β_1, obtenim:

$$\frac{\partial l}{\partial \beta_1} = \sum_{i=1}^{n} \left(\frac{y_i}{\beta_1} - x_i \right),$$

que igualant a zero i aïllant ens permet obtenir el màxim, és a dir, l'estimador de β_1 :

$$\sum_{i=1}^{n} \left(\frac{y_i}{\beta_1} - x_i \right) = 0 \implies \frac{1}{\beta_1} \sum_i y_i = \sum_i x_i$$
$$\widehat{\beta}_1 = \frac{\sum y_i}{\sum x_i} = \frac{\frac{1}{n}\sum y_i}{\frac{1}{n}\sum x_i} = \frac{\overline{y}}{\overline{x}}$$

Exercici 4.

(a) Observacions influents.

Les observacions influents son punts de dades que afecten desproporcionadament els resultats del model de regressió. Si es treuen, el model canvia significativament (canvien coeficients, prediccions, etc.). Es poden detectar de les següents maneres:

- ***Leverage*:** El valor de l'element h_{ii} de la matriu

  $$H = X(X^T X)^{-1} X^T$$

  es pot interpretar com el *leverage* de la observació i-èssima del model. Si aquest valor és molt alt,

  $$h_{ii} > \frac{2m}{n}\,,$$

  amb m el nombre de predictors i n el nombre d'observacions, pot haver-hi una potencial influència excessiva.
- **Distància de Cook:** La distància de Cook mesura directament la influència de cada observació:

  $$D_i = \frac{\widehat{(\beta} - \widehat{\beta}_{(i)})^T X^T X\,(\widehat{\beta} - \widehat{\beta}_{(i)})}{m\,\widehat{\sigma}^2}\,,$$

  amb X de mida $n \times m$ i $\widehat{\beta}_{(i)}$ l'estimador de β després d'eliminar la i-èssima observació. Com amb el *leverage*, si té un valor massa elevat,

  $$D_i > \frac{4}{n}\,,$$

  indica una possible influència excessiva.

(b) Càlcul d'un IC del 95%.

Tenim el següent model de regressió lineal:

$$\text{preu} = \beta_0 + \beta_1 \cdot \text{km} + \varepsilon\,.$$

Primer calculem les estimacions dels paràmetres:

$$\widehat{\beta}_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}\,,$$
$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\,\overline{x}\,,$$

amb $\overline{x}, \overline{y}$ les mitjanes de les observacions x_i i y_i respectivament. Ara calculem la predicció per un quilometratge de 50000:

$$\hat{y}_0 = \widehat{\beta}_0 + 50000\,\widehat{\beta}_1$$

I el seu error estàndard:

$$SE_0 = \widehat{\sigma}^2 \sqrt{\frac{1}{n} + \frac{(50000 - \overline{x})^2}{\sum (x_i - \overline{x})^2}} \quad \text{amb} \quad \widehat{\sigma}^2 = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

A partir d'això ja podem calcular l'interval de confiança:

$$IC_{0.95} = \left(\hat{y}_0 \pm t_{0.975,\,4} \cdot SE_0\right)$$

amb $t_{0.975,4}$ el quantil t de Student de 0.975 i amb 4 graus de llibertat.

(c) *Null deviance* i *Residual deviance*.

La *null deviance* és la desviació del model amb només l'*intercept* (és a dir, el model "buit") respecte la desviació del model saturat (amb un paràmetre per cada observació; és a dir, el model que s'ajusta perfectament). És la desviació total a explicar pels predictors. La *null deviance* es calcula fent

$$2(l_{\text{sat}} - l_{\text{null}})\,,$$

amb $l_{\text{sat}}, l_{\text{null}}$ les log-versemblança del model saturat i del model nul respectivament.

La *residual deviance* és la desviació del model amb predictors (el model ajustat) respecte el model saturat. Quantifica la variabilitat no explicada pel model ajustat. Es calcula de la següent manera:

$$2(l_{\text{sat}} - l_{\text{mod}})\,,$$

amb $l_{\text{sat}}, l_{\text{mod}}$ les log-versemblança dels models saturat i ajustat respectivament.