Pràctica 6

Regressió lineal simple

Aquesta pràctica té com a objectiu estudiar i analitzar els models de regressió lineal en `R`. Per consolidar de manera millor tots els conceptes, es treballarà amb les dades recollides pel físic escocès James Forbes al segle 19 que trobeu al fitxer `Forbes.csv` al campus virtual.
La base de dades conté la temperatura d'ebullició de l'aigua (en graus Fahrenheit), la pressió atmosfèrica (en mil·libars) i l'altitud (en metres) de 17 localitats dels Alps i d'Escòcia. L'objectiu de James Forbes era estimar l'altitud sobre el nivell del mar a partir de la temperatura d'ebullició de l'aigua.
Aquestes dades són les mateixes que s'han utilitzat a les diapositives de teoria per introduir els conceptes de regressió lineal.

1 Equació de regressió

Considerem dues variables aleatòries X i Y i suposem que Y depèn linealment de X. Considerem una mostra $\{(x_i, y_i)\}_{i=1}^n$, corresponent a valors de la parella de variables (X, Y).
La variable X s'anomena variable independent, predictora o explicativa. En el cas que estudiarem, no és aleatòria i els seus valors són fixats per l'experiment. Donada la variable X, la variable Y és aleatòria i s'anomena dependent o de resposta.
La relació entre aquestes dues variables es pot expressar mitjançant la recta d'equació

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

on β_0 i β_1 s'anomenen coeficients de regressió i són els paràmetres que hauran de ser estimats a partir de les dades de la mostra, i els errors ε_i són variables aleatòries independents amb distribució normal $N(0, \sigma^2)$.

En aquesta pràctica seguirem l'objectiu de J. Forbes i veurem com estimar l'altitud d'una localitat (variable resposta) a partir de la temperatura d'ebullició de l'aigua (variable explicativa).

2 Estimació dels paràmetres

Els estimadors de mínims quadrats de β_0 i β_1 (que coincideixen amb els de màxima versemblança quan es suposa que els errors són normals) s'obtenen minimitzant la suma dels errors quadràtics

$$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

De l'optimització obtenim els valors

$$\hat{\beta}_1 = \frac{S_{xy}}{S_x^2}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

on $S_{xy} = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})$ és la covariància empírica i $S_x^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2$ la variància empírica.

Els valors ajustats a la recta de regressió són

$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

Les estimacions dels errors (o residus) venen donades per

$$\hat{\varepsilon}_i = y_i - \hat{y_i} = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i).$$

Aquestes fórmules ens permeten trobar un estimador de σ^2, conegut com a variància dels residus,

$$\hat{\sigma}^2 = \frac{SSE}{n-2},$$

on SSE és la suma de quadrats dels residus

$$SSE = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y_i})^2.$$

La quantitat $n-2$ representa els graus de llibertat dels residus i correspon al nombre de dades menys el nombre de paràmetres estimats.

Estudiem la distribució d'aquests estimadors. Calculant-ne l'esperança obtenim

$$\mathbb{E}(\hat{\beta}_0) = \beta_0, \qquad \mathbb{E}(\hat{\beta}_1) = \beta_1 \qquad \text{i} \qquad \mathbb{E}(\hat{\sigma}^2) = \sigma^2,$$

per tant, són estimadors no esbiaixats.
Es pot demostrar que els estimadors de β_0 i β_1 no són independents i les seves variàncies i covariància venen donades per

$$\begin{aligned} Var(\hat{\beta}_0) &= \frac{\sigma^2}{n}\left(1 + \frac{\overline{x}^2}{S_x^2}\right) \\ Var(\hat{\beta}_1) &= \frac{\sigma^2}{nS_x^2} \\ Cov(\hat{\beta}_0, \hat{\beta}_1) &= -\frac{\sigma^2 \overline{x}}{nS_x^2}. \end{aligned}$$

Si suposem que els errors tenen distribució normal, els estimadors de de β_0 i β_1 tindran distribució normal amb les mitjanes i variàncies indicades.
D'altra banda,

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{SSE}{\sigma^2} \sim \chi^2_{n-2}.$$

Aquestes propietats són útils per fer inferència sobre els paràmetres. A l'hora de normalitzar, haurem de reemplaçar σ^2 pel seu estimador $\hat{\sigma}^2$. D'aquesta manera, els estadístics que obtindrem tindran distribució t de Student.

Veiem-ne un exemple. Considerem les dades de J. Forbes i estimem els paràmetres de la recta de regressió:

```
forbes <- read.csv("~/.../Forbes.csv")
x <- forbes$Boiling
y <- forbes$Altitude

# Estadístics de la mostra
n <- length(x)
mx <- mean(x)
my <- mean(y)
sx <- sd(x)
```

```
sy <- sd(y)
sxy <- cov(x,y)

# Coeficients
beta1 <- sxy/sx^2
beta0 <- my-beta1*mx
SSE <- (n-1)*(sy^2-sxy^2/sx^2)
sigma2 <- SSE/(n-2)

# Valors ajustats i errors
yhat <- beta0+beta1*x
errors <- y-yhat

# Representació (diagrama de dispersió de les dades)
plot(x, y, xlab="Temperatura d'ebullició", ylab="Altitud",
      pch=19, col="deepskyblue3")
abline(beta0, beta1)     # s'afegeix la recta de regressió al plot
```

Obtenim els valors

$$\hat{\beta}_0 = 37492.53 \qquad \hat{\beta}_1 = -176.9546 \qquad \hat{\sigma}^2 = 5612.074$$

Per tant, la recta de regressió és

$$y = 37492.53 - 176.9546x$$

i es veu representada a la Figura 1.

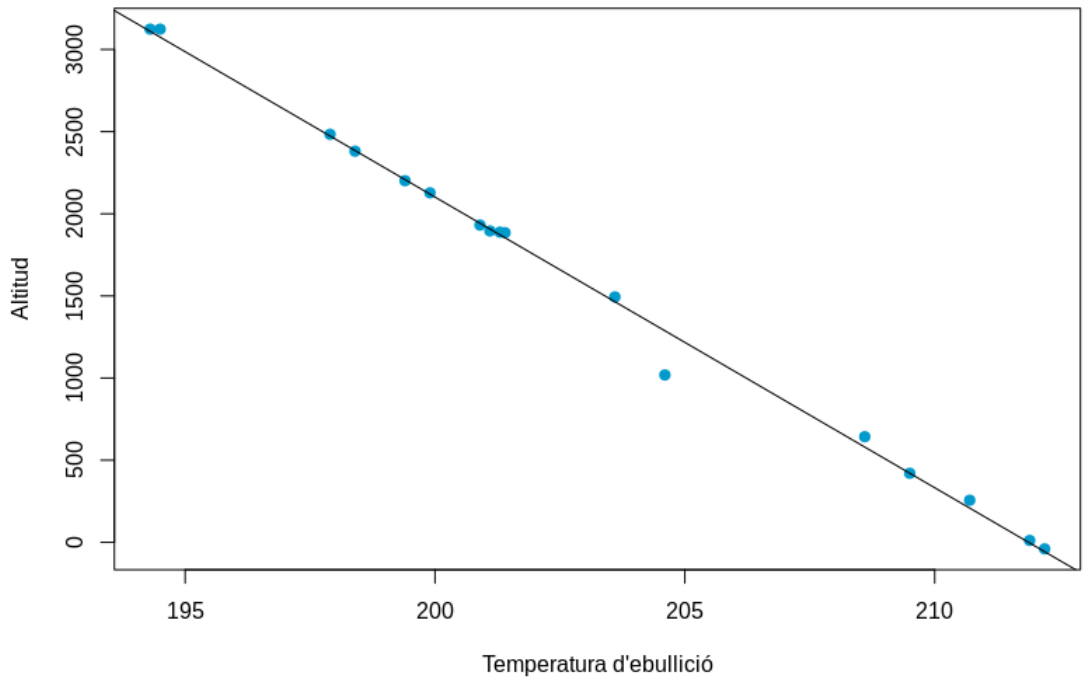

Figura 1: Diagrama de dispersió de les dades de J. Forbes i recta de regressió.

3 Regressió lineal

La funció `lm` s'utilitza per tractar amb models lineals i, en particular, per realitzar models de regressió. La manera més clara d'entendre com funciona i quin resultats ens proporciona és amb un exemple.

Considerem les dades de J. Forbes. Recordeu que la variable `x` representa la temperatura d'ebullició i la variable `y` l'altitud. Si executem les instruccions

```
recta <- lm(y~x)
reg <- summary(recta)
reg
```

obtenim el següent output:

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-268.696   -3.705   13.611   29.059   63.172

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 37492.526    660.181   56.79   <2e-16 ***
x            -176.955      3.252  -54.42   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 74.91 on 15 degrees of freedom
Multiple R-squared:  0.995, Adjusted R-squared:  0.9946
F-statistic:  2962 on 1 and 15 DF,  p-value: < 2.2e-16
```

Com es pot veure, s'obté molta més informació que els estimadors dels coeficients de la recta de regressió. En aquest apartat i en els següents l'analitzarem en detall.

Les instruccions

```
plot(x, y, xlab="Temperatura d'ebullició", ylab="Altitud",
      pch=19, col="deepskyblue3")
abline(recta)
```

produeixen el mateix plot representat per la Figura 1.

Les instruccions

```
coef(recta)
sigma(recta)
```

o, en alternativa,

```
reg$coefficients[,"Estimate"]
reg$sigma
```

ens donen els estimadors $\hat{\beta}_0$, $\hat{\beta}_1$ i $\hat{\sigma}$ (no $\hat{\sigma}^2$), respectivament. Trobem aquests valors també al `summary` del model ajustat: a la columna `Estimate` dels coeficients i a `Residual standard error`. Observeu que coincideixen amb els càlculs obtinguts a partir de les fórmules que hem fet a l'apartat anterior.

4 Inferència sobre el model

4.1 Intervals de confiança per als paràmetres

A la Secció 2 hem estudiat la distribució dels paràmetres del model. Normalitzant els estimadors de β_0 i β_1, obtenim

$$\frac{\hat{\beta}_i - \beta_i}{se(\hat{\beta}_i)} \sim t_{n-2}, \qquad i = 0, 1, \tag{1}$$

on $se(\hat{\beta}_i)$ representa l'error estàndard estimat. Aquests estadístics són

$$\begin{aligned} se(\hat{\beta}_0) &= \hat{\sigma}\sqrt{\frac{1}{n}\left(1 + \frac{\overline{x}^2}{S_x^2}\right)} \\ se(\hat{\beta}_1) &= \frac{\hat{\sigma}}{\sqrt{n}S_x}. \end{aligned}$$

i es calculen en `R` amb l'instrucció

```
reg$coefficients[,"Std. Error"]
```

A partir de (1) podem construir els intervals de confiança:

$$\left[\hat{\beta}_i - t_{n-2,1-\frac{\alpha}{2}}\, se(\hat{\beta}_i),\ \hat{\beta}_i + t_{n-2,1-\frac{\alpha}{2}}\, se(\hat{\beta}_i)\right].$$

En `R` es poden calcular directament amb la funció `confint`:

```
confint(recta)
```

Per defecte l'interval és amb 95% de confiança i es pot canviar agregant l'opció `conf.level=1-alpha`.

A la Seccio 2 hem vist que

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{SSE}{\sigma^2} \sim \chi^2_{n-2}.$$

A partir d'aquesta distribució obtenim l'interval de confiança per σ^2:

$$\left[\frac{(n-2)\hat{\sigma}^2}{\chi^2_{n-2,1-\frac{\alpha}{2}}},\ \frac{(n-2)\hat{\sigma}^2}{\chi^2_{n-2,\frac{\alpha}{2}}}\right].$$

4.2 Contrasts d'hipòtesis per a $\hat{\beta}_0$ i $\hat{\beta}_1$

Per $i = 0, 1$, considerem el contrast d'hipòtesis

$$\begin{cases} H_0 : \beta_i = \beta_{i_0} \\ H_1 : \beta_i \neq \beta_{i_0} \end{cases}$$

on β_{i_0} és un valor fixat. Intuïtivament, l'estadístic de contrast que s'utilitza per resoldre un test d'aquest tipus és

$$\frac{\hat{\beta}_i - \beta_{i_0}}{se(\hat{\beta}_i)} \sim t_{n-2}.$$

Els programes com `R`, treballen per defecte amb $\beta_{i_0} = 0$. L'última columna del `summary` del model ajustat conté els p-valors d'aquests tests (no són tests simultanis!) i la penúltima columna conté els valors dels estadístics de contrast.
El test d'hipòtesis

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

és d'especial interès. Si acceptem la hipòtesis nul·la, aleshores implícitament estem dient que la variable `x` no serveix per explicar la variabilitat de `y`. La recta de regressió tendria pendent 0, és a dir, seria horitzontal.

Tornem al nostre exemple. Al `summary` del model ajustat o executant la comanda

```
reg$coefficients
```

veiem la següent informació:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 37492.526    660.181   56.79   <2e-16 ***
x            -176.955      3.252  -54.42   <2e-16 ***
```

L'última columna ens indica que els p-valors dels contrasts d'hipòtesis

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0 \end{cases} \qquad i = 0, 1$$

són pràcticament zero, per tant, en els dos casos rebutgem la hipòtesis nul·la. En particular, la recta de regressió no és horitzontal i podem afirmar que hi ha relació lineal entre la temperatura d'ebullició i l'altitud.

5 Prova de significació de la regressió

Volem saber si la variable X aporta informació rellevant per explicar la variabilitat de Y. A l'apartat anterior, hem vist que una forma de fer-ho és realitzar el contrast

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

En aquest apartat veurem un'altra forma d'estudiar-ho que resulta equivalent en el cas de la regressió lineal simple, però que permet respondre a preguntes més generals quan es treballa amb models més complexos. Aquest mètode es basa en les sumes de quadrats següents:

- **Suma de quadrats total**:

$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2.$$

  Mesura la variabilitat total de la variable Y.

- **Suma de quadrats de la regressió**:

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2.$$

  Representa la part de variabilitat de Y que ha sigut explicada per la regressió sobre X.

- **Suma de quadrats dels errors**:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

  Indica la part de variabilitat de Y que queda sense ser explicada després de la regressió.

Es compleix la relació

$$SST = SSR + SSE.$$

Les mides de SSR i SSE relatives al total SST serveixen per valorar quant significatiu és el model de regressió. Observem que $SSE = 0$ equival a un ajust perfecte, mentre $SSR = 0$ implica que

la X no aporta res a la explicació de la Y. Evidentment, les situacions habituals són $SSR > 0$ i $SSE < SST$.

Es pot veure que $SSR/1$ i $SSE/(n-2)$ són estimadors no esbiaixats de σ^2. Considerem el quocient entre aquestes dues quantitats. Si SSR és petit, el quocient serà pròxim a 1. Si SSR és gran, el quocient també ho serà.
Sota la hipòtesi nul·la $H_0 : \beta_1 = 0$,

$$F_{obs} = \frac{SSR/1}{SSE/(n-2)} \sim F_{1,n-2}.$$

Per tant, es pot fer un contrast d'hipòtesis amb regió crítica

$$\{F_{obs} > F_{1-\alpha,1,n-2}\}.$$

Amb les sumes de quadrats definides abans, es pot definir el **coeficient de determinació** R^2, que correspon al quocient entre la variabilitat explicada per la regressió i la variabilitat total:

$$R^2 = \frac{SSR}{SST}.$$

Aquest coeficient coincideix amb el quadrat del coeficient de correlació r^2.

El valor observat F_{obs} i el seu corresponent p-valor i el coeficient de determinació R^2 apareixen a les últimes dues línies del `summary` del model ajustat. Tornant al nostre exemple, això correspon a les línies

```
Multiple R-squared:  0.995, Adjusted R-squared:  0.9946
F-statistic:  2962 on 1 and 15 DF,  p-value: < 2.2e-16
```

El valor de R^2 és el que correspon a `Multiple R-squared`. L'R^2 ajustat s'utilitza en regressió múltiple i té en compte el nombre de variables independents preses en consideració.
Observem que el valor observat F_{obs} és 2962 i correspon al quadrat (arrodonit) de -54.42, que és el valor de l'estadístic de contrast del test d'hipòtesis amb $H_0 : \beta_1 = 0$. Aquest fet no és una casualitat. En el cas particular de la regressió lineal simple, l'estadístic de contrast de la regressió és el quadrat de l'estadístic de contrast sobre la pendent.

6 Estimació de la resposta mitjana i predicció

6.1 Interval de confiança per a la mitjana de la variable resposta

Volem donar un interval de confiança per a la mitjana de y condicionada a que la variable independent agafi un valor x_0. Tenim

$$\mu_{y|x=x_0} = \mathbb{E}(Y|x = x_0) = \beta_0 + \beta_1 x.$$

Per tant, normalitzant adequadament, obtenim que $\mu_{y|x=x_0}$ està continguda a l'interval

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{n-2,1-\frac{\alpha}{2}} \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{nS_x^2}}.$$

A `R` la funció `predict` calcula aquest interval.

Considerem les dades de J. Forbes. Suposem que volem un interval per la altitud mitjana quan les temperatures d'ebullició de l'aigua són 190 ºF, 200 ºF i 210 ºF. Les instruccions

```
data <- data.frame(x=c(190,200,210))
predict(recta, newdata=data, interval="confidence")
```

donen com a resultat

```
          fit        lwr        upr
1 3871.1442 3773.3741 3968.9144
2 2101.5978 2057.7957 2145.4000
3  332.0514  269.7198  394.3831
```

La primera columna és el valor ajustat. Les altres corresponen als límits inferior i superior dels intervals, respectivament. Per defecte l'interval és amb 95% de confiança i es pot canviar agregant l'opció `level=1-alpha`.

6.2 Interval de predicció de nou valors

Donat un valor x_0 de la variable independent, volem calcular un interval pel valor y_0 de la variable resposta. Aquest interval serà més amplio que l'interval de confiança presentat a l'apartat anterior per la major variabilitat induïda pel terme d'error amb variància σ^2.
Es pot demostrar que

$$Var(y_0 - \hat{y}_0) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{nS_x^2}\right)$$

Estimant σ^2 amb $\hat{\sigma}^2$ i normalitzant adequadament, per un valor donat x_0, y_0 està contingut a l'interval

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{n-2,1-\frac{\alpha}{2}} \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{nS_x^2}}.$$

A `R` la funció `predict` s'utilitza també per calcular aquest interval.

Considerem novament les dades de J. Forbes. Suposem que volem un interval per la altitud quan les temperatures d'ebullició de l'aigua són 190 °F, 200 °F i 210 °F. Les instruccions

```
data <- data.frame(x=c(190,200,210))
predict(recta, newdata=data, interval="prediction")
```

donen com a resultat

```
          fit        lwr        upr
1 3871.1442 3683.9142 4058.3743
2 2101.5978 1936.0239 2267.1717
3  332.0514  160.6416  503.4612
```

Com en el cas anterior, la primera columna és el valor ajustat i les altres corresponen als límits inferior i superior dels intervals, respectivament. Per defecte l'interval és amb 95% de confiança i es pot canviar agregant l'opció `level=1-alpha`.