Regressió lineal simple

Regressió lineal

La taula següent dóna l'alçada mitjana (en cm) dels nins a determinades edats (en anys):

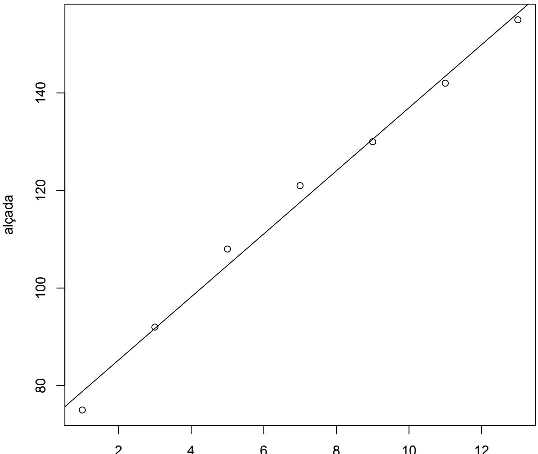
edat	1	3	5	7	9	11	13
alçada	75	92	108	121	130	142	155

A la lliçó 2 de R-I calculàvem amb R la millor relació lineal

$$\text{alçada} \approx b_0 + b_1 \cdot \text{edat}$$

Regressió lineal

```
> edat=c(1,3,5,7,9,11,13)
> alçada=c(75,92,108,121,130,142,155)
> plot(edat,alçada)
> abline(lm(alçada~edat))
```



Regressió lineal simple

Tenim parelles d'observacions de dues variables X, Y:

$$(x_i, y_i)_{i=1,2,\ldots,n}$$

i volem estudiar com depèn el valor de Y del de X:

- La variable aleatòria Y és la variable dependent o de resposta
- La variable (no necessàriament aleatòria) X és la variable de control, independent o de regressió

Volem trobar la millor relació funcional que expliqui la variable Y conegudes les observacions de la variable X. Per ara, cercam una relació lineal que expliqui Y en funció de X.

Regressió lineal simple

Suposam que

$$\mu_{Y|x} = \beta_0 + \beta_1 x$$

on $\mu_{Y|x}$ és el valor esperat de Y quan X val x, i β_0 (terme independent) i β_1 (pendent) són dos paràmetres que volem estimar

Amb una mostra $(x_i, y_i)_{i=1,2,\ldots,n}$, calcularem estimacions b_0 i b_1 de β_0 i de β_1

Això ens donarà la recta de regressió per a la nostra mostra:

$$\widehat{y} = b_0 + b_1 x$$

que donat un valor x_0 de X ens estimarà el valor $\widehat{y}_0 = b_0 + b_1 x_0$ de Y sobre el mateix individu

Regressió lineal simple

El model anterior el reescrivim com a

$$\begin{aligned} Y|x &= \mu_{Y|x} + E_x \\ &= \beta_0 + \beta_1 x + E_x, \end{aligned}$$

on

- $Y|x$ és la variable aleatòria "valor de Y quan X val x"
- E_x és la variable aleatòria error o residu, que dóna la diferència entre el valor de Y i el valor "esperat" $\mu_{Y|x}$, és a dir, $\beta_0 + \beta_1 x$
- Com que suposam que $\mu_{Y|x} = \beta_0 + \beta_1 x$, suposam que $\mu_{E_x} = 0$ per a cada x

Mínims quadrats

Per a cada observació (x_i, y_i), tendrem

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \Rightarrow \varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$$

Diguem l'error quadràtic teòric d'aquest model a

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

A la regressió lineal per mínims quadrats, els estimadors b_0 i b_1 de β_0 i β_1 que cercam són els valors de "les incògnites" β_0 i β_1 que minimitzen aquest SS_ε

Mínims quadrats

Anem a minimitzar SS_ε. El mínim (b_0, b_1) de

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

anul·larà les derivades respecte de β_0 i β_1.

Derivem:

$$\begin{aligned} \frac{\partial SS_\varepsilon}{\partial \beta_0} &= -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \\ \frac{\partial SS_\varepsilon}{\partial \beta_1} &= -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i \end{aligned}$$

Mínims quadrats

El (b_0, b_1) que cercam satisfà

$$2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0$$
$$2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)x_i = 0$$

Reescrivim:

$$\begin{aligned} nb_0 + \Big(\sum_{i=1}^{n} x_i\Big) b_1 &= \sum_{i=1}^{n} y_i \\ \Big(\sum_{i=1}^{n} x_i\Big) b_0 + \Big(\sum_{i=1}^{n} x_i^2\Big) b_1 &= \sum_{i=1}^{n} x_i y_i \end{aligned}$$

Mínims quadrats

Les solucions són

$$
\begin{aligned}
b_1 &= \frac{n\sum\limits_{i=1}^{n} x_i y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{n\sum\limits_{i=1}^{n} x_i^2 - (\sum\limits_{i=1}^{n} x_i)^2} \\
b_0 &= \frac{\sum\limits_{i=1}^{n} y_i - b_1 \sum\limits_{i=1}^{n} x_i}{n}
\end{aligned}
$$

i donen el mínim de SS_ε

Mínims quadrats

Considerem les mitjanes

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

i les variàncies i covariància

$$\begin{aligned}
s_x^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\Big(\sum_{i=1}^{n} x_i^2\Big) - \overline{x}^2 \\
s_y^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n}\Big(\sum_{i=1}^{n} y_i^2\Big) - \overline{y}^2 \\
s_{xy} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\Big(\sum_{i=1}^{n} x_i y_i\Big) - \overline{x}\cdot\overline{y}
\end{aligned}$$

Mínims quadrats

Els valors de b_0 i b_1 trobats abans es poden reescriure de la forma següent:

Teorema

Els estimadors b_0 i b_1 per mínims quadrats de β_0 i β_1 són

$$b_1 = \frac{s_{xy}}{s_x^2}, \quad b_0 = \overline{y} - b_1\overline{x}.$$

Escriurem

$$\widehat{y} = b_0 + b_1 x$$

Direm a $\widehat{y}$ el valor estimat de Y quan $X = x$

Per a cada observació (x_i, y_i), direm l'error a

$$e_i = y_i - \widehat{y}_i = y_i - b_0 - b_1 x_i$$

Exemple 1

Volíem calcular la recta de regressió per mínims quadrats de

edat (x)	1	3	5	7	9	11	13
alçada (y)	75	92	108	121	130	142	155

```
> x=c(1,3,5,7,9,11,13)
> y=c(75,92,108,121,130,142,155)
> x.b=mean(x)
> y.b=mean(y)
> s2.x=var(x)*6/7
> s2.y=var(y)*6/7
> s.xy=cov(x,y)*6/7
> round(c(x.b,y.b,s2.x,s2.y,s.xy),3)
[1]   7.000 117.571  16.000 674.531 103.429
```

Exemple 1

$\overline{x}$	$\overline{y}$	s_x^2	s_y^2	s_{xy}
7.000	117.571	16	674.531	103.429

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{103.429}{16} = 6.4643$$

$$b_0 = \overline{y} - b_1\overline{x} = 117.571 - 6.4643 \cdot 7 = 72.3209$$

Obtenim

$$\widehat{y} = 72.3209 + 6.4643x$$

```
> lm(y~x)$coefficients
(Intercept)           x
  72.321429    6.464286
```

Alerta!

Els càlculs involucrats en la regressió lineal són molt poc robusts: els arrodoniments poden influir molt en el resultat final

A la Wikipedia (`http://en.wikipedia.org/wiki/Simple_linear_regression`) hi trobareu un exemple detallat d'una regressió de pes en funció d'alçada. Calculada en metres dóna:

$$\widehat{y} = 61.272x - 39.062$$

Si es passen les alçades a polzades, s'arrodoneixen, es calcula la recta de regressió, i es torna a passar el resultat a metres, dóna

$$\widehat{y} = 61.675x - 39.746$$

Exemple 2

En un experiment on es volia estudiar l'associació entre consum de sal i pressió arterial, a alguns individus se'ls assignà aleatòriament una quantitat diària constant de sal en la seva dieta, i al cap d'un mes se'ls mesurà la tensió mitjana. Alguns resultats varen ser els següents

X (sal, en g)	Y (Pressió, en mm de Hg)
1.8	100
2.2	98
3.5	110
4.0	110
4.3	112
5.0	120

Trobau la recta de regressió lineal per mínims quadrats de Y en funció de X

Exemple 2

$\overline{x}$	$\overline{y}$	s_x^2	s_y^2	s_{xy}
3.467	108.333	1.2856	55.2222	8.1444

$b_1 =$

$b_0 =$

Obtenim la recta $\widehat{y} =$

Podeu comprovar que amb R dóna el mateix

```
> sal=c(1.8,2.2,3.5,4,4.3,5)
> ten=c(100,98,110,110,112,120)
> lm(ten~sal)$coefficients
```

Propietats

- La recta de regressió passa pel vector mitjà $(\overline{x}, \overline{y})$:

$$b_0 + b_1\overline{x} = \overline{y}$$

- La mitjana dels valors estimats és igual a la mitjana dels observats:

$$\overline{\widehat{y}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{y}_i = \frac{1}{n}\sum_{i=1}^{n}(b_0 + b_1 x_i) = b_0 + b_1\overline{x} = \overline{y}$$

- Els errors $(e_i)_{i=1,\dots,n}$ tenen mitjana 0:

$$\overline{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - b_0 - b_1 x) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i) = 0$$

Propietats

Direm suma de quadrats dels errors a

$$SS_E = \sum_{i=1}^{n} e_i^2$$

Els errors $(e_i)_{i=1,\ldots,n}$ tenen variància

$$s_e^2 = \frac{1}{n}\Big(\sum_{i=1}^{n} e_i^2\Big) - \overline{e}^2 = \frac{SS_E}{n} - 0 = \frac{SS_E}{n}$$

Propietats

Teorema

Si les variables aleatòries error E_{x_i} tenen totes mitjana 0 i la mateixa variància σ_E^2 i, dues a dues, tenen covariància 0, aleshores

- *b_0 i b_1 són els estimadors lineals no esbiaixats òptims (més eficients) de β_0 i β_1*
- *Un estimador no esbiaixat de σ_E^2 és $S^2 = \dfrac{SS_E}{n-2}$*

Teorema

Si a més les variables aleatòries error E_{x_i} són normals, aleshores b_0 i b_1 són els estimadors màxim versemblants de β_0 i β_1 (i no esbiaixats)

Exemple 1

Si suposam que al nostre exemple d'edats i alçades els errors tenen la mateixa variància i són incorrelats, podem estimar aquesta variància:

```
> x=c(1,3,5,7,9,11,13)
> y=c(75,92,108,121,130,142,155)
> y.cap=72.321+6.464*x
> errors=y-y.cap
> SSE=sum(errors^2)
> S2=SSE/(length(x)-2)
> S2
[1] 8.314296
```

Tenim que $S^2 = 8.3143$, i estimam que σ_E^2 val això

Exemple 2

Si suposam que al nostre exemple de sal i tensió arterial els errors tenen la mateixa variància i són incorrelats, podem estimar aquesta variància:

```
> sal=c(1.8,2.2,3.5,4,4.3,5)
> ten=c(100,98,110,110,112,120)
> ten.cap=86.371+6.335*sal
> errors.ten=ten-ten.cap
> SSE=sum(errors.ten^2)
> S2=SSE/(length(sal)-2)
> S2
[1] 5.436475
```

Tenim que $S^2 = 5.4365$, i estimam que σ_E^2 val això

Això és tot?

Hem estimat els coeficients β_0 i β_1 i la variable $Y|x$, per a cada x, al model

$$\mu_{Y|x} = \beta_0 + \beta_1 x$$

però ens pot interessar més:

- Com és de significativa l'estimació obtinguda?
- Error estàndard d'aquests estimadors
- Intervals de confiança

Amb R obtenim molt més. . .

```
> summary(lm(alçada~edat))
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  72.3214     2.1966   32.92 4.86e-07 ***
edat          6.4643     0.2725   23.73 2.48e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
   ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.883 on 5 degrees of freedom
Multiple R-squared: 0.9912, Adjusted R-squared: 0.9894
F-statistic: 562.9 on 1 and 5 DF,  p-value: 2.477e-06
```

Com és de significativa la regressió?

Entenem que la recta $\widehat{y} = b_0 + b_1 x$ és una bona aproximació de y com a funció lineal de x quan aquesta recta explica molta part de la variabilitat de y

Es quantifica amb el coeficient de determinació R^2

```
> summary(lm(alçada~edat))$r.squared
[1] 0.9911957
```

Sumes de quadrats

Siguin:

- $SS_T = \sum_{i=1}^{n}(y_i - \overline{y})^2$: suma total de quadrats

$$SS_T = n \cdot s_y^2$$

- $SS_R = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$: suma de quadrats de la regressió

$$SS_R = n \cdot s_{\widehat{y}}^2$$

- $SS_E = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$: suma de quadrats dels errors

$$SS_E = n \cdot s_e^2$$

Sumes de quadrats

Teorema

En una regressió lineal pel mètode de mínims quadrats, es té que

$$SS_T = SS_R + SS_E$$

o equivalentment,

$$s_y^2 = s_{\widehat{y}}^2 + s_e^2$$

El coeficient de determinació R^2

El coeficient de determinació d'una regressió lineal és

$$R^2 = \frac{SS_R}{SS_T} = \frac{s_{\widehat{y}}^2}{s_y^2}$$

Per tant, R^2 és la fracció de la variabilitat de y que queda explicada per la variabilitat de $\widehat{y}$

Si la regressió lineal és per mínims quadrats,

$$R^2 = \frac{SS_T - SS_E}{SS_T} = 1 - \frac{SS_E}{SS_T} = 1 - \frac{s_e^2}{s_y^2}$$

El coeficient de determinació R^2

A més, $R^2 = r_{xy}^2$, el coeficient de correlació al quadrat

$$
\begin{aligned}
R^2 &= \frac{SS_R}{SS_T} = \frac{\sum_{i=1}^{n}(b_1 x_i + b_0 - \overline{y})^2}{n s_y^2} \\
&= \frac{\sum_{i=1}^{n}\left(\frac{s_{xy}}{s_x^2} x_i - \frac{s_{xy}}{s_x^2}\overline{x}\right)^2}{n s_y^2} \\
&= \frac{\frac{s_{xy}^2}{s_x^4}\sum_{i=1}^{n}(x_i - \overline{x})^2}{n s_y^2} \\
&= \frac{s_{xy}^2}{s_x^4} \cdot \frac{s_x^2}{s_y^2} = \frac{s_{xy}^2}{s_x^2 \cdot s_y^2} = r_{xy}^2
\end{aligned}
$$

Exemple 1

```
> x=c(1,3,5,7,9,11,13)
> y=c(75,92,108,121,130,142,155)
> y.cap=72.321+6.464*x
> SST=sum((y-mean(y))^2)
> SSR=sum((y.cap-mean(y))^2)
> SSE=sum((y-y.cap)^2)
> round(c(SST,SSR,SSE),3)
[1] 4721.714 4679.729  41.571
```

$$R^2 = \frac{4679.729}{4721.714} = 0.9912$$

```
> cor(x,y)^2
[1] 0.9911957
```

Exemple 2

```
> sal=c(1.8,2.2,3.5,4,4.3,5)
> ten=c(100,98,110,110,112,120)
> ten.cap=86.371+6.335*sal
> SST=sum((ten-mean(ten))^2)
> SSR=sum((ten.cap-mean(ten))^2)
> SSE=sum((ten-ten.cap)^2)
> round(c(SST,SSR,SSE),3)
[1] 331.333 309.553  21.746
```

$$R^2 =$$

El valor de R^2 no és suficient!

No és possible valorar la bondat del model només basant-se amb el valor de R^2. Vegem quatre conjunts de parells (x_i, y_i), generats específicament amb aquest objectiu, continguts en el data frame `anscombe` de R:

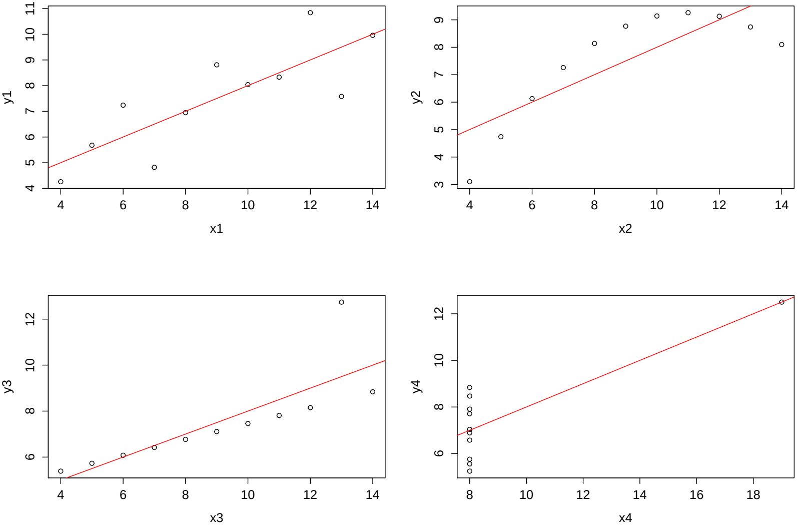
```
> data(anscombe)
> str(anscombe)
'data.frame': 11 obs. of  8 variables:
 $ x1: num  10 8 13 9 11 14 6 4 12 7 ...
 $ x2: num  10 8 13 9 11 14 6 4 12 7 ...
 $ x3: num  10 8 13 9 11 14 6 4 12 7 ...
 $ x4: num  8 8 8 8 8 8 8 19 8 8 ...
 $ y1: num  8.04 6.95 7.58 8.81 8.33 ...
 $ y2: num  9.14 8.14 8.74 8.77 9.26 8.1 6.13 3.1 ...
 $ y3: num  7.46 6.77 12.74 7.11 7.81 ...
 $ y4: num  6.58 5.76 7.71 8.84 8.47 7.04 5.25 12.5 ...
```

Anem a fer-ne les regressions i a mostrar-ne els R^2 respectius i l'ajustament gràfic de les rectes.

El valor de R^2 no és suficient!

```
> summary(lm(y1~x1,data=anscombe))$r.squared
[1] 0.6665425
> summary(lm(y2~x2,data=anscombe))$r.squared
[1] 0.6665425
> summary(lm(y3~x3,data=anscombe))$r.squared
[1] 0.6665425
> summary(lm(y4~x4,data=anscombe))$r.squared
[1] 0.6665425
> #Anem a representar els resultats
> par(mfrow=c(2,2))
> plot(y1~x1,data=anscombe)
> abline(lm(y1~x1,data=anscombe),col=2)
> plot(y2~x2,data=anscombe)
> abline(lm(y2~x2,data=anscombe),col=2)
> plot(y3~x3,data=anscombe)
> abline(lm(y3~x3,data=anscombe),col=2)
> plot(y4~x4,data=anscombe)
> abline(lm(y4~x4,data=anscombe),col=2)
```

El valor de R^2 no és suficient!



Supòsits del model

Suposam d'ara endavant que cada E_{x_i} segueix una distribució normal amb mitjana $\mu_{E_{x_i}} = 0$, la mateixa variància σ_E^2, i $\sigma(E_{x_i}, E_{x_j}) = 0$ per a cada parella i, j

Si només tenim molt pocs y per a cada x, això no es pot contrastar, però implica que els $(e_i)_{i=1,\ldots,n}$ provenen d'una $N(0, \sigma_E^2)$, amb σ_E^2 estimada per S^2, i això sí que ho podem contrastar

Exemple 1

A l'exemple de les alçades i les edats

```
> x=c(1,3,5,7,9,11,13)
> y=c(75,92,108,121,130,142,155)
> y.cap=72.321+6.464*x
> errors=y-y.cap
> SSE=sum(errors^2)
> S2=SSE/5
> ks.test(errors,"pnorm",0,sqrt(S2))
One-sample Kolmogorov-Smirnov test
data:  errors
D = 0.1746, p-value = 0.9583
alternative hypothesis: two-sided
```

Exemple 2

A l'exemple de la sal i la tensió

```
> sal=c(1.8,2.2,3.5,4,4.3,5)
> ten=c(100,98,110,110,112,120)
> ten.cap=86.371+6.335*sal
> errors=ten-ten.cap
> SSE=sum(errors^2)
> S2=SSE/4
> ks.test(errors,"pnorm",0,sqrt(S2))
One-sample Kolmogorov-Smirnov test
data:  errors
D = 0.2553, p-value = 0.7479
alternative hypothesis: two-sided
```

Intervals de confiança

Teorema

Sota aquestes hipòtesis,

- *Els errors estàndard dels estimadors b_1 i b_0 són, respectivament,*

$$\frac{\sigma_E}{s_x\sqrt{n}} \quad i \quad \frac{\sigma_E\sqrt{s_x^2+\overline{x}^2}}{s_x\sqrt{n}}$$

En aquests errors estàndard (i tots els que segueixen), estimam σ_E per mitjà de $S=\sqrt{S^2}$

Intervals de confiança

Teorema

Sota aquestes hipòtesis,

- *Les fraccions*

$$\frac{b_1 - \beta_1}{\frac{S}{s_x\sqrt{n}}} \quad i \quad \frac{b_0 - \beta_0}{\frac{S\sqrt{s_x^2+\overline{x}^2}}{s_x\sqrt{n}}}$$

  segueixen lleis t de Student amb $n-2$ graus de llibertat.

Intervals de confiança

Per tant, sota aquestes hipòtesis,

- Un interval de confiança del $(1-\alpha)\cdot 100\%$ per β_1 és

$$\left]b_1 - t_{n-2,1-\frac{\alpha}{2}}\frac{S}{s_x\sqrt{n}},\, b_1 + t_{n-2,1-\frac{\alpha}{2}}\frac{S}{s_x\sqrt{n}}\right[$$

  Ho escriurem

$$\beta_1 = b_1 \pm t_{n-2,1-\frac{\alpha}{2}}\frac{S}{s_x\sqrt{n}}$$

- Un interval de confiança del $(1-\alpha)\cdot 100\%$ per β_0 és

$$\beta_0 = b_0 \pm t_{n-2,1-\frac{\alpha}{2}}\frac{S\sqrt{s_x^2+\overline{x}^2}}{s_x\sqrt{n}}$$

Exemple 1

A l'exemple de les alçades en funció de l'edat, havíem obtingut la recta

$$\widehat{y} = 72.321 + 6.464x$$

i $\overline{x} = 7$, $s_x^2 = 16$, $n = 7$, $S^2 = 8.314$

Un interval de confiança al 95% per β_1 és

$$\begin{aligned}\beta_1 &= b_1 \pm t_{n-2,1-\frac{\alpha}{2}} \frac{S}{s_x\sqrt{n}} \\ &= 6.464 \pm t_{5,0.975} \frac{\sqrt{8.314}}{4\sqrt{7}} \\ &= 6.464 \pm 2.5706 \cdot 0.2724 = 6.464 \pm 0.7\end{aligned}$$

És l'interval $]5.764, 7.164[$

Exemple 1

A l'exemple de les alçades en funció de l'edat, havíem obtingut la recta

$$\widehat{y} = 72.321 + 6.464x$$

i $\overline{x} = 7$, $s_x^2 = 16$, $n = 7$, $S^2 = 8.314$

Un interval de confiança al 95% per β_0 és

$$\begin{aligned}
\beta_0 &= b_0 \pm t_{n-2,1-\frac{\alpha}{2}} \frac{S\sqrt{s_x^2 + \overline{x}^2}}{s_x\sqrt{n}} \\
&= 72.321 \pm t_{5,0.975} \frac{\sqrt{8.314} \cdot \sqrt{16 + 7^2}}{4\sqrt{7}} \\
&= 72.321 \pm 2.5706 \cdot 2.1966 = 72.321 \pm 5.647
\end{aligned}$$

És l'interval $]66.674, 77.968[$

Exemple 1

Obtenim

- Interval del 95% per a β_1: $]5.764, 7.164[$
- Interval del 95% per a β_0: $]66.674, 77.968[$

```
> confint(lm(y~x),level=0.95)
                2.5 %    97.5 %
(Intercept) 66.674769 77.968088
x            5.763904  7.164668
```

Exemple 2

A l'exemple de la tensió en funció de la sal, havíem obtingut la recta

$$\widehat{y} = 86.371 + 6.335x$$

i $\overline{x} = 3.467$, $s_x^2 = 1.2856$, $n = 6$, $S^2 = 5.4365$, $t_{4,0.975} = 2.7764$

L'interval de confiança al 95% per β_1 és

Exemple 2

A l'exemple de la tensió en funció de la sal, havíem obtingut la recta

$$\widehat{y} = 86.371 + 6.335x$$

i $\overline{x} = 3.467$, $s_x^2 = 1.2856$, $n = 6$, $S^2 = 5.4365$, $t_{4,0.975} = 2.7764$

L'interval de confiança al 95% per β_0 és

Intervals de confiança

Teorema

Sota aquestes hipòtesis, i si x_0 és un possible valor de X

- *L'error estàndard de $\widehat{y}_0$ com a estimador de $\mu_{Y|x_0}$ és*

$$\sigma_E \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{n s_x^2}}$$

- *La fracció*

$$\frac{\widehat{y}_0 - \mu_{Y/x_0}}{S\sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{n s_x^2}}}$$

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

Intervals de confiança

Teorema

Sota aquestes hipòtesis, i si x_0 *és un possible valor de* X

- *L'error estàndard de* $\widehat{y}_0$ *com a estimador de* y_0 *és*

$$\sigma_E \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{n s_x^2}}$$

- *La fracció*

$$\frac{\widehat{y}_0 - y_0}{S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{n s_x^2}}}$$

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

Intervals de confiança

Per tant, sota aquestes hipòtesis,

- Un interval de confiança del $(1-\alpha)\cdot 100\%$ per $\mu_{Y|x_0}$ és

$$\mu_{Y|x_0} = \widehat{y}_0 \pm t_{n-2,1-\frac{\alpha}{2}} S \sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{ns_x^2}}$$

- Un interval de confiança del $(1-\alpha)\cdot 100\%$ per y_0 és

$$y_0 = \widehat{y}_0 \pm t_{n-2,1-\frac{\alpha}{2}} S \sqrt{1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{ns_x^2}}$$

Exemple 1

A l'exemple de les alçades en funció de l'edat, havíem obtingut la recta

$$\widehat{y} = 72.321 + 6.464x$$

i $\overline{x} = 7$, $s_x^2 = 16$, $n = 7$, $S^2 = 8.314$

Suposem que volem estimar l'alçada y_0 d'un nin de $x_0 = 10$ anys

$$\widehat{y}_0 = 72.321 + 6.464 \cdot 10 = 136.961$$

Interval de confiança al 95% per aquest valor? Interval de confiança al 95% per al valor esperat?

Exemple 1

Un interval de confiança al 95% per y_0 és

$$\begin{aligned} y_0 &= \widehat{y}_0 \pm t_{n-2,1-\frac{\alpha}{2}} S \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{n s_x^2}} \\ &= 136.961 \pm t_{5,0.975} \sqrt{8.314} \cdot \sqrt{1 + \frac{1}{7} + \frac{(10-7)^2}{7 \cdot 16}} \\ &= 136.961 \pm 2.5706 \cdot 3.189 = 136.961 \pm 8.198 \end{aligned}$$

És l'interval $]128.8, 145.2[$

Exemple 1

Un interval de confiança al 95% per $\mu_{Y|x_0}$ és

$$
\begin{aligned}
\mu_{Y|x_0} &= \widehat{y}_0 \pm t_{n-2,1-\frac{\alpha}{2}} S \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{n s_x^2}} \\
&= 136.961 \pm t_{5,0.975} \sqrt{8.314} \cdot \sqrt{\frac{1}{7} + \frac{(10-7)^2}{7 \cdot 16}} \\
&= 136.961 \pm 2.5706 \cdot 1.362 = 136.961 \pm 3.501
\end{aligned}
$$

És l'interval $]133.5, 140.5[$

Exemple 1

```
> regressio=lm(y~x)
> newdata=data.frame(x=10)
> predict.lm(lm(y~x),newdata,
  interval="prediction",level=0.95)
       fit      lwr     upr
1 136.9643 128.7665 145.162
> predict(lm(y~x),newdata,
  interval="confidence",level=0.95)
       fit      lwr      upr
1 136.9643 133.4624 140.4662
```

Té sentit una regressió lineal?

Si $\beta_1 = 0$, el model de regressió lineal no té sentit:

$$Y = \beta_0 + E$$

i les variacions en els valors de Y són totes degudes a l'error.

El contrast

$$\left\{ \begin{array}{l} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{array} \right.$$

el podem realitzar amb l'interval de confiança per a β_1: si 0 no hi pertany, rebutjam la hipòtesi nul·la

Exemples

Hem obtingut:

- A l'exemple 1, un interval del 95% per a β_1 és $]5.764, 7.164[$
- A l'exemple 2, un interval del 95% per a β_1: $]4.004, 8.666[$

Als dos casos concloem que $\beta_1 \neq 0$ i que per tant tenia sentit fer la regressió lineal

Amb R

```
> summary(lm(alçada~edat))
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  72.3214     2.1966   32.92 4.86e-07 ***
edat          6.4643     0.2725   23.73 2.48e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
   ‘.’ 0.1 ‘ ’ 1
...
```

Els `t value` són els dels contrastos amb H_0 : “coeficient $= 0$”, i els p-valors són els d'aquests contrastos. Podem rebutjar que $\beta_1 = 0$ (i que $\beta_0 = 0$)