Regressió lineal múltiple

Regressió lineal múltiple

Tenim ara k variables (no necessàriament aleatòries) independents $X_1, \ldots, X_k$ i una variable dependent Y

Suposam el model

$$\mu_{Y|x_1,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

Els paràmetres β_i són desconeguts i els estimam a partir d'una mostra:

$$(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)_{i=1,\ldots,n}$$

amb $n > k$ (el nombre d'observacions ha de ser més gran que el nombre de variables)

Escriurem $\underline{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$

Regressió lineal múltiple

Traduïm aquest model en

$$Y|x_1,\dots,x_k = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + E_{x_1,\dots,x_k}$$

on

- $Y|x_1,\dots,x_k$ és la v.a. que dóna el valor de Y quan cada $X_i = x_i$
- $E_{x_1,\dots,x_k}$ són les v.a. error, o residuals, i representen l'error aleatori del model associat a $(x_1,\dots,x_k)$

A partir d'una mostra

$$(\underline{x}_i, y_i)_{i=1,2,\dots,n}$$

obtendrem estimacions $b_0, b_1, \dots, b_k$ dels paràmetres $\beta_0, \beta_1, \dots, \beta_k$

Regressió lineal múltiple

Diguem

$$\widehat{y}_i = b_0 + b_1 x_{i1} + \cdots + b_k x_{ik}$$
$$y_i = b_0 + b_1 x_{i1} + \cdots + b_k x_{ik} + e_i$$

Aleshores

- $\widehat{y}_i$ és el valor predit de y_i a partir de $\underline{x}_i$ i els estimadors $b_0, b_1, \ldots, b_k$ dels paràmetres
- e_i estima l'error $E_{\underline{x}_i}$
- $e_i = y_i - \widehat{y}_i$

Regressió lineal múltiple

Escrivim-ho en forma matricial. Diguem

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix}, \ \widehat{\mathbf{y}} = \begin{pmatrix} \widehat{y}_1 \\ \widehat{y}_2 \\ \vdots \\ \widehat{y}_n \end{pmatrix}, \ \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$$

Regressió lineal múltiple

Les equacions

$$\begin{aligned} \widehat{y}_i &= b_0 + b_1 x_{i1} + \cdots + b_k x_{ik} \\ y_i &= b_0 + b_1 x_{i1} + \cdots + b_k x_{ik} + e_i \end{aligned}$$

corresponen a

$$\begin{aligned} \widehat{\mathbf{y}} &= \mathbf{X}\mathbf{b} \\ \mathbf{y} &= \mathbf{X}\mathbf{b} + \mathbf{e} \end{aligned}$$

Mètode dels mínims quadrats

Definim l'error quadràtic SS_E com:

$$
\begin{aligned}
SS_E = & \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \\
= & \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik})^2.
\end{aligned}
$$

Els estimadors de $\beta_0, \beta_1, \ldots, \beta_k$ pel mètode de mínims quadrats seran els valors $b_0, b_1, \ldots, b_k$ que minimitzin SS_E

Mètode dels mínims quadrats

Per calcular-los, calculam les derivades parcials de SS_E respecte de cada b_i, les igualam a 0, les resolem, i comprovam que la solució $(b_0, \ldots, b_k)$ trobada dóna un mínim...

Teorema

Els estimadors per mínims quadrats de $\beta_0, \beta_1, \ldots, \beta_k$ a partir de la mostra $(\underline{x}_i, y_i)_{i=1,2,\ldots,n}$ són donats per l'equació següent:

$$\mathbf{b} = \left(\mathbf{X}^t \cdot \mathbf{X}\right)^{-1} \cdot \left(\mathbf{X}^t \cdot \mathbf{y}\right).$$

Exemple 3

Es postula que l'alçada d'un nadó (y) té una relació lineal amb la seva edat en dies (x_1), la seva alçada en néixer en cm (x_2), el seu pes en kg en néixer (x_3) i l'augment en tant per cent del seu pes actual respecte del seu pes en néixer (x_4)

El model és

$$\mu_{Y|x_1,x_2,x_3,x_4} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

En una mostra de $n = 9$ nins, els resultats varen ser els de la taula següent:

Exemple 3

y	x_1	x_2	x_3	x_4
57.5	78	48.2	2.75	29.5
52.8	69	45.5	2.15	26.3
61.3	77	46.3	4.41	32.2
67	88	49	5.52	36.5
53.5	67	43	3.21	27.2
62.7	80	48	4.32	27.7
56.2	74	48	2.31	28.3
68.5	94	53	4.3	30.3
69.2	102	58	3.71	28.7

Exemple 3

$$\mathbf{X} = \begin{pmatrix} 1 & 78 & 48.2 & 2.75 & 29.5 \\ 1 & 69 & 45.5 & 2.15 & 26.3 \\ 1 & 77 & 46.3 & 4.41 & 32.2 \\ 1 & 88 & 49 & 5.52 & 36.5 \\ 1 & 67 & 43 & 3.21 & 27.2 \\ 1 & 80 & 48 & 4.32 & 27.7 \\ 1 & 74 & 48 & 2.31 & 28.3 \\ 1 & 94 & 53 & 4.3 & 30.3 \\ 1 & 102 & 58 & 3.71 & 28.7 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 57.5 \\ 52.8 \\ 61.3 \\ 67 \\ 53.5 \\ 62.7 \\ 56.2 \\ 68.5 \\ 69.2 \end{pmatrix}$$

$\mathbf{b}$ serà $(\mathbf{X}^t \cdot \mathbf{X})^{-1} \cdot (\mathbf{X}^t \cdot \mathbf{y})$

Exemple 3

```
> X=
 matrix(c(1,78,48.2,2.75,29.5,1,69,45.5,2.15,26.3,
 1,77,46.3,4.41,32.2,1,88,49,5.52,36.5,
 1,67,43,3.21,27.2,1,80,48,4.32,27.7,
 1,74,48,2.31,28.3,1,94,53,4.3,30.3,
 1,102,58,3.71,28.7),nrow=9,byrow=TRUE)
> y=cbind(c(57.5,52.8,61.3,67,53.5,62.7,56.2,68.5,
  69.2))
```

Exemple 3

```
> t(X)%*%X
        [,1]     [,2]      [,3]      [,4]      [,5]
[1,]   9.00   729.00   439.000   32.6800   266.700
[2,] 729.00 60123.00 35947.200 2702.4100 21715.300
[3,] 439.00 35947.20 21568.180 1604.3880 13026.010
[4,]  32.68  2702.41  1604.388  128.6602   990.268
[5,] 266.70 21715.30 13026.010  990.2680  7980.830
> t(X)%*%y
          [,1]
[1,]   548.700
[2,] 45001.000
[3,] 26946.890
[4,]  2035.521
[5,] 16348.290
```

Exemple 3

El producte $\mathbf{X}^t\mathbf{X}$ és:

$$\begin{pmatrix} 9 & 729 & 439 & 32.68 & 266.7 \\ 729 & 60123 & 35947.2 & 2702.41 & 21715.3 \\ 439 & 35947.2 & 21568.18 & 1604.388 & 13026.01 \\ 66.07 & 6108.19 & 3541.008 & 128.66 & 1948.561 \\ 266.7 & 21715.3 & 13026.01 & 990.27 & 7980.83 \end{pmatrix}$$

El producte $\mathbf{X}^t\mathbf{y}$ és

$$\mathbf{X}^t\mathbf{y} = \begin{pmatrix} 548.7 \\ 45001 \\ 26946.89 \\ 2035.52 \\ 16348.29 \end{pmatrix}$$

Exemple 3

El vector d'estimadors dels coeficients $\beta_0, \beta_1, \ldots, \beta_4$ és

$$\mathbf{b} = \begin{pmatrix} 9 & 729 & 439 & 32.68 & 266.7 \\ 729 & 60123 & 35947.2 & 2702.41 & 21715.3 \\ 439 & 35947.2 & 21568.18 & 1604.388 & 13026.01 \\ 66.07 & 6108.19 & 3541.008 & 128.66 & 1948.561 \\ 266.7 & 21715.3 & 13026.01 & 990.27 & 7980.83 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 548.7 \\ 45001 \\ 26946.89 \\ 2035.52 \\ 16348.29 \end{pmatrix}$$

Exemple 3

```
> round(solve(t(X)%*%X)%*%(t(X)%*%y),4)
        [,1]
[1,]  7.1475
[2,]  0.1001
[3,]  0.7264
[4,]  3.0758
[5,] -0.0300
```

Exemple 3

Obtenim

$$\mathbf{b} = \begin{pmatrix} 7.1475 \\ 0.1001 \\ 0.7264 \\ 3.0758 \\ -0.03 \end{pmatrix}$$

La funció lineal de regressió cercada és:

$$\widehat{y} = 7.1475 + 0.1001x_1 + 0.7264x_2 + 3.0758x_3 - 0.03x_4.$$

Exemple 3

```
> Xd=X[,c(2:5)]
> lm(y~Xd)
Call:
lm(formula = y ~ Xd)
Coefficients:
(Intercept)        Xd1        Xd2        Xd3         Xd4
    7.14753    0.10009    0.72642    3.07584    -0.03004
```

Propietats

- La recta de regressió passa pel vector mitjà $(\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_k, \overline{y})$:

$$\overline{y} = b_0 + b_1\overline{x}_1 + \cdots + b_k\overline{x}_k$$

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

$$\overline{\widehat{y}} = \overline{y}$$

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

$$s_e^2 = \frac{SS_E}{n}$$

Sumes de quadrats

Siguin

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

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

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

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

Teorema

Si la regressió és per mínims quadrats,

$$SS_T = SS_R + SS_E \text{ o, equivalentment, } s_y^2 = s_{\widehat{y}}^2 + s_e^2$$

Coeficient de determinació

El coeficient de determinació d'una regressió lineal és

$$R^2 = \frac{SS_R}{SS_T} = \frac{s^2_{\widehat{y}}}{s^2_y}$$

Representa la fracció de la variabilitat de y que és explicada per la variabilitat del model de regressió lineal

El coeficient de correlació múltiple de y respecte de $x_1, \ldots, x_k$ és

$$R = \sqrt{R^2}$$

Coeficient de determinació

R^2 tendeix a créixer amb k, fins i tot si les variables que afegim són redundants

Per tenir-ho en compte, en lloc d'emprar

$$R^2 = \frac{SS_R}{SS_T} = \frac{SS_T - SS_E}{SS_T}$$

s'empra el coeficient de determinació ajustat

$$R^2_{adj} = \frac{MS_T - MS_E}{MS_T}$$

on

$$MS_T = \frac{SS_T}{n-1}, MS_E = \frac{SS_E}{n-k-1}$$

Queda

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

Exemple 3

Coeficients de determinació del nostre exemple

```
> # X i y ja definits
> b=solve(t(X)%*%X)%*%(t(X)%*%y)
> y.cap=X%*%b
> SS.T=sum((y-mean(y))^2)
> SS.R=sum((y.cap-mean(y))^2)
> SS.E=sum((y.cap-y)^2)
> round(c(SS.T,SS.R,SS.E),3)
[1] 321.240 318.274   2.966
```

$$R^2 = \frac{SS_R}{SS_T} = \frac{318.274}{321.24} = 0.991$$

$$R^2_{adj} = 1 - (1 - R^2)\left(\frac{9-1}{9-4-1}\right) = 0.982$$

Exemple 3

```
> Xd=X[,c(2:5)]
> summary(lm(y~Xd))
...

Residual standard error: 0.861 on 4 degrees of freedom
Multiple R-squared: 0.9908, Adjusted R-squared: 0.9815
F-statistic: 107.3 on 4 and 4 DF,  p-value: 0.0002541
> summary(lm(y~Xd))$r.squared
[1] 0.9907683
> summary(lm(y~Xd))$adj.r.squared
[1] 0.9815367
```

Comparació de models

Sovint ens interessarà comparar dos models lineals per a una mateixa variable dependent (per exemple, si afegim o llevam una variable, millora el model?)

Aquesta comparació se sol fer comparant els R^2_{adj}: qui el tengui més gran, guanya

```
> Xd=X[,c(2:5)]
> summary(lm(y~Xd))$adj.r.squared
[1] 0.9815367
> Xd1=X[,c(2:4)]
> summary(lm(y~Xd1))$adj.r.squared
[1] 0.9851091
```

El model és millor si no tenim en compte X_4 (l'augment de pes en %)

Comparació de models

Altres índexs que darrerament es fan servir per comparar models:

- AIC (Akaike's Information Criterion)

$$AIC = n \ln(SS_E/n) + 2k$$

AIC quantifica quanta informació de Y es perd amb el model i quantes variables hi empram: el millor model és el que té un valor de AIC més petit

```
> AIC(lm(y~Xd))
[1] 27.54953
> AIC(lm(y~Xd1))
[1] 25.62252
```

Comparació de models

Altres índexs que darrerament es fan servir per comparar models:

- BIC (Bayesian Information Criterion)

$$BIC = n \ln(SS_E/n) + k \ln(n)$$

BIC quantifica quanta informació de Y es perd amb el model i quantes variables i dades hi empram: el millor model és el que té un valor de BIC més petit

```
> BIC(lm(y~Xd))
[1] 28.73288
> BIC(lm(y~Xd1))
[1] 26.60864
```

Solen donar la mateixa conclusió, i si donen diferent és convenient dir-ho

Supòsits del model

Suposarem d'ara endavant que les variables aleatòries error $E_i = E_{\underline{x}_i}$ són incorrelades, i totes normals de mitjana totes 0 i de variància totes σ_E^2

Teorema

Sota aquestes hipòtesis, els estimadors $b_0, \ldots, b_k$ de $\beta_0, \ldots, \beta_k$ són màxim versemblants i a més no esbiaixats.

Supòsits del model

Teorema

Sota aquestes hipòtesis,

$$Cov(\beta_0, \beta_1, \ldots, \beta_k) = \sigma_E^2 \cdot (X^t \cdot X)^{-1}$$

i un estimador no esbiaixat de σ_E^2 és

$$S^2 = \frac{SS_E}{n-k-1}$$

Fa una estona a S^2 li hem dit MS_E

Exemple 3

En el nostre exemple, una estimació de la variància comuna dels errors σ_E^2 és

$$S^2 = \frac{2.9656}{9-4-1} = 0.7414$$

i una estimació de la matriu de covariàncies de $\beta_0, \ldots, \beta_4$ és

$$S^2 \cdot (X^t \cdot X)^{-1}$$

$$= \begin{pmatrix} 270.919 & 5.325 & -12.521 & -13.743 & -1.4 \\ 5.325 & 0.115 & -0.266 & -0.326 & -0.0176 \\ -12.521 & -0.266 & 0.618 & 0.742 & 0.0416 \\ -13.743 & -0.326 & 0.742 & 1.122 & -0.00598 \\ -1.4 & -0.0176 & 0.0416 & -0.00598 & 0.0277 \end{pmatrix}$$

Intervals de confiança

Teorema

Sota aquestes hipòtesis,

- *L'error estàndard de cada estimador b_i és*

$$\sqrt{(\sigma_E^2 \cdot (X^t X)^{-1})_{ii}}$$

(l'arrel quadrada de la i-èsima entrada de la diagonal de $\sigma_E^2 \cdot (X^t X)^{-1}$, començant per $i = 0$)

Intervals de confiança

Teorema

Sota aquestes hipòtesis,

- *Cada fracció*

$$\frac{\beta_i - b_i}{\sqrt{(S^2 \cdot (X^t X)^{-1})_{ii}}}$$

  segueix un llei t de Student amb $n-k-1$ graus de llibertat

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

$$b_i \pm t_{n-k-1,1-\frac{\alpha}{2}} \cdot \sqrt{(S^2 \cdot (X^t X)^{-1})_{ii}}$$

Exemple 3

Al nostre exemple, cerquem un interval de confiança del 95% per β_2.

Recordem els b_i i calculem la diagonal de $S^2 \cdot (X^t X)^{-1}$:

```
> round(t(b),4)
       [,1]   [,2]   [,3]   [,4]  [,5]
[1,] 7.1475 0.1001 0.7264 3.0758 -0.03
> S2=SS.E/(9-4-1)
> round(diag(S2*solve(t(X)%*%X)),4)
[1] 270.9188   0.1154   0.6176   1.1219   0.0277
```

Exemple 3

Per tant, serà

$$\begin{aligned}\beta_2 &= 0.7264 \pm t_{4,0.975}\sqrt{0.6176} \\ &= 0.7264 \pm 2.776\sqrt{0.6176} = 0.7265 \pm 2.1816\end{aligned}$$

Obtenim $]-1.455, 2.908[$

Exemple 3

Calculem l'interval de confiança al 95% per β_0:

Exemple 3

```
> round(confint(lm(y~Xd)),3)
              2.5 % 97.5 %
(Intercept) -38.552 52.847
Xd1          -0.843  1.043
Xd2          -1.456  2.908
Xd3           0.135  6.017
Xd4          -0.492  0.432
```

Intervals de confiança

Teorema

Siguin $\underline{x}_0 = (x_{01}, \ldots, x_{0k})$ *una observació de* $X_1, \ldots, X_k$ *i* $\mathbf{x}_0 = (1, x_{01}, \ldots, x_{0k})$. *Sota les nostres hipòtesis,*

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

$$S\sqrt{\mathbf{x}_0 \cdot (X^t \cdot X)^{-1} \cdot \mathbf{x}_0^t}$$

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

$$S\sqrt{1 + \mathbf{x}_0 \cdot (X^t \cdot X)^{-1} \cdot \mathbf{x}_0^t}$$

Intervals de confiança

Teorema

Siguin $\underline{x}_0 = (x_{01}, \ldots, x_{0k})$ *una observació de* $X_1, \ldots, X_k$ *i* $\mathbf{x}_0 = (1, x_{01}, \ldots, x_{0k})$. *Sota les nostres hipòtesis,*

- *Les fraccions*

$$\frac{\mu_{Y|\underline{x}_0} - \widehat{y}_0}{S\sqrt{\mathbf{x}_0 \cdot (X^t \cdot X)^{-1} \cdot \mathbf{x}_0^t}}$$

$$\frac{y_0 - \widehat{y}_0}{S\sqrt{1 + \mathbf{x}_0 \cdot (X^t \cdot X)^{-1} \cdot \mathbf{x}_0^t}}$$

segueixen lleis t *de Student amb* $n-k-1$ *graus de llibertat*

Intervals de confiança

Teorema

Siguin $\underline{x}_0 = (x_{01}, \ldots, x_{0k})$ *una observació de* $X_1, \ldots, X_k$ *i* $\mathbf{x}_0 = (1, x_{01}, \ldots, x_{0k})$. *Sota les nostres hipòtesis,*

- *Un interval de confiança del* $(1-\alpha)\cdot 100\%$ *per* $\mu_{Y|\underline{x}_0}$ *és*

$$\widehat{y}_0 \pm t_{n-k-1,1-\frac{\alpha}{2}} \cdot S\sqrt{\mathbf{x}_0 \cdot (X^t \cdot X)^{-1} \cdot \mathbf{x}_0^t}$$

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

$$\widehat{y}_0 \pm t_{n-k-1,1-\frac{\alpha}{2}} \cdot S\sqrt{1 + \mathbf{x}_0 \cdot (X^t \cdot X)^{-1} \cdot \mathbf{x}_0^t}$$

Exemple 3

Al nostre exemple, volem trobar intervals de confiança del 95% per $\mu_{Y|\underline{x}_0}$ i y_0 per a $\underline{x}_0 = (69, 45.5, 2.15, 26.3)$.

$$\begin{aligned}\widehat{y}_0 &= b_0 + b_1 x_{01} + b_2 x_{02} + b_3 x_{03} + b_4 x_{04} \\ &= 7.1475 + 0.1001 \cdot 69 + 0.7264 \cdot 45.5 \\ &\qquad + 3.0758 \cdot 2.15 - 0.03 \cdot 26.3 = 52.929\end{aligned}$$

Calculem

$$\begin{aligned}&\mathbf{x}_0 (X^t X)^{-1} \mathbf{x}_0^t \\ &\quad = (1, 69, 45.5, 2.15, 26.3) \cdot (X^t X)^{-1} \cdot \begin{pmatrix} 1 \\ 69 \\ 45.5 \\ 2.15 \\ 26.3 \end{pmatrix}\end{aligned}$$

```
> xvec=rbind(c(1,69,45.5,2.15,26.3))
> xvec%*%solve(t(X)%*%X)%*%t(xvec)
          [,1]
[1,] 0.3614889
```

Exemple 3

L'interval de confiança per $\mu_{Y|\underline{x}_0}$ és

$$\begin{aligned}\mu_{Y|\underline{x}_0} &= \widehat{y}_0 \pm t_{9-4-1,0.975} \cdot S\sqrt{\mathbf{x}_0 \cdot (X^t \cdot X)^{-1} \cdot \mathbf{x}_0^t} \\ &= 52.929 \pm 2.776 \cdot \sqrt{0.7414} \cdot \sqrt{0.3615} \\ &= 52.929 \pm 1.437\end{aligned}$$

Dóna $]51.492, 54.366[$

L'interval de confiança per y_0 és

$$\begin{aligned}y_0 &= \widehat{y}_0 \pm t_{9-4-1,0.975} \cdot S\sqrt{1 + \mathbf{x}_0 \cdot (X^t \cdot X)^{-1} \cdot \mathbf{x}_0^t} \\ &= 52.929 \pm 2.776 \cdot \sqrt{0.7414} \cdot \sqrt{1 + 0.3615} \\ &= 52.929 \pm 2.789\end{aligned}$$

Dóna $]50.14, 55.718[$

Exemple 3

Per calcular-ho amb R, convé organitzar les observacions en un data frame (ja ho hauríem d'haver fet abans!)

```
> X.df=as.data.frame(cbind(y,Xd))
> names(X.df)=c("y","x1","x2","x3","x4")
> str(X.df)
'data.frame': 9 obs. of  5 variables:
$ y : num  57.5 52.8 61.3 67 53.5 62.7 56.2 68.5 69.2
$ x1: num  78 69 77 88 67 80 74 94 102
$ x2: num  48.2 45.5 46.3 49 43 48 48 53 58
$ x3: num  2.75 2.15 4.41 5.52 3.21 4.32 2.31 4.3 3.71
$ x4: num  29.5 26.3 32.2 36.5 27.2 27.7 28.3 30.3 28.7
```

Exemple 3

```
> regressio=lm(y~x1+x2+x3+x4,data=X.df)
> regressio
Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = X.df)
Coefficients:
(Intercept)           x1           x2           x3           x4
    7.14753      0.10009      0.72642      3.07584     -0.03004
> newdata=data.frame(x1=69,x2=45.5,x3=2.15,x4=26.3)
```

Exemple 3

```
> predict.lm(regressio,newdata,
   interval="prediction",level=0.95)
       fit      lwr      upr
1 52.92898 50.13952 55.71845
> predict(regressio,newdata,
   interval="confidence",level=0.95)
       fit      lwr      upr
1 52.92898 51.49164 54.36633
```

Té sentit una regressió lineal?

Com en el cas simple, ens interessa el contrast

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1 : \text{hi ha qualque } \beta_i \neq 0 \end{cases}$$

Si acceptam la hipòtesi nul·la, l'estimació donada per la regressió és constant i el model lineal no és adequat

Té sentit una regressió lineal?

Això es pot fer amb k contrastos

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

emprant l'estadístic corresponent

$$\frac{\beta_i - b_i}{\sqrt{(S^2 \cdot (X^t X)^{-1})_{ii}}}$$

que segueix una llei t de Student amb $n - k - 1$ graus de llibertat

Però són k contrastos, i no independents, per tant mantenir el nivell de significació global és complicat

ANOVA en la regressió lineal

Una altra possibilitat és emprar un ANOVA:

Si

$$\beta_1 = \beta_2 = \cdots = \beta_k = 0,$$

aleshores

$$\mu_{Y|\underline{x}_1} = \cdots = \mu_{Y|\underline{x}_n} (= \beta_0)$$

Per tant, si al contrast

$$\left\{ \begin{array}{l} H_0 : \mu_{Y|\underline{x}_1} = \cdots = \mu_{Y|\underline{x}_n} \\ H_1 : \text{no és veritat que\ldots} \end{array} \right.$$

rebutjam la hipòtesi nul·la, implica que podem rebutjar que $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ i el model tendrà sentit

ANOVA en la regressió lineal

La taula és

Font de variació	Graus de llibertat	Suma de quadrats	Quadrats mitjans	F	p-valor
Regressió	k	SS_R	MS_R	MS_R/MS_E	p-valor
Error	$n-k-1$	SS_E	MS_E		

on

$$MS_R = \frac{SS_R}{k}, \quad MS_E = \frac{SS_E}{n-k-1}, \quad F = \frac{MS_R}{MS_E}$$

i si la hipòtesi nul·la és vertadera (i els errors són normals), F segueix una llei F de Fisher amb k i $n-k-1$ graus de llibertat:

$$\text{p-valor} = P(F_{k,n-k-1} \geqslant F)$$

Exemple 3

La taula en el nostre exemple és

Font de variació	Graus de llibertat	Suma de quadrats	Quadrats mitjans	F	p-valor
Regressió	4	318.274	79.569	107.323	≈ 0
Error	4	2.9656	0.7414		

Concloem que el model lineal és adequat segons aquesta anàlisi

Exemple 3

Amb R

```
> anova(lm(y ~ Xd))
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
Xd         4 318.27  79.569  107.32 0.0002541 ***
Residuals  4   2.97   0.741
```

Exemple 3

R també fa tots els contrastos

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

dins el `lm`

```
> summary(lm(y~Xd))
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.14753   16.45961   0.434   0.6865
Xd1          0.10009    0.33971   0.295   0.7829
Xd2          0.72642    0.78590   0.924   0.4076
Xd3          3.07584    1.05918   2.904   0.0439 *
Xd4         -0.03004    0.16646  -0.180   0.8656
...
```