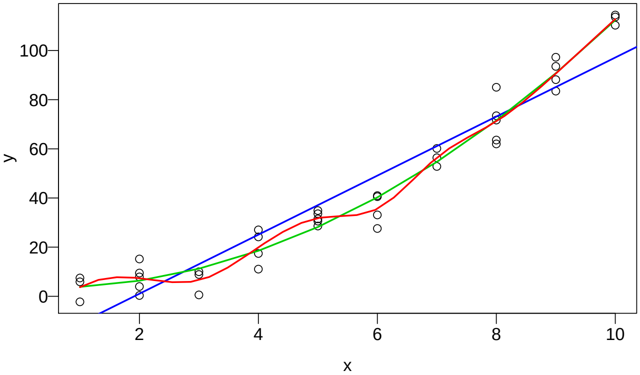
8 Selección del Modelo



−ebz

Contenido

−ebz

Selección de variables

En la mayoría de los casos prácticos el analista tiene una base de posibles regresores candidatos para explicar/ajustar la variable respuesta bajo estudio.

Elegir el subconjunto de regresores que finalmente modelen la respuesta es el *problema de selección de variables (modelos)*.

En el proceso de elegir el modelo más adecuado uno considera que, a mayor sea el número de regresores, se tendrá una *“mejor explicación”* de la respuesta. Por otro lado, mientras menos regresores haya en el modelo, menor será la variabilidad en la respuesta. El compromiso entre ambos es lo que se conoce como seleccionar el *mejor modelo*.

El problema de selección de variables se discute idealmente suponiendo que los regresores están en las escalas adecuadas, que no hay datos atípicos ni observaciones de influencia.

Los procedimientos que se verán a continuación son de gran ayuda pero no garantizan que se llegue al mejor modelo, que quizás no es único.

−ebz

Consecuencias de la incorrecta especificación del modelo de RLM

Subespecificación:

Suponga que

$$\mathbb{E}[y] = X_1\beta_1 + X_2\beta_2 = X\beta$$

con $X_{1_{n\times(1+p)}}$ y $X_{2_{n\times(r)}}$, pero que en realidad se ajusta el modelo $y = X_1\beta_1 + \epsilon$. Entonces,

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_1] &= (X_1'X_1)^{-1}X_1'\mathbb{E}[y] \\
&= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2) \\
&= \beta_1 + \underbrace{(X_1'X_1)^{-1}X_1'X_2\beta_2}_{\text{sesgo de } \beta_1} \\
&= \beta_1 + \Delta\beta_2
\end{aligned}$$

y donde $\Delta = (X_1'X_1)^{-1}X_1'X_2$ se conoce como *matriz alias*.

Notas:

- En esta situación $\hat{\beta}_1$ es *insesgado* sólo si $X_1'X_2 = 0$. Esto es, solamente si los regresores (columnas) de X_1 son *ortogonales* a los regresores de X_2 (caso del *diseño de experimentos*).
- Los estimadores $\hat{\beta}_1$ tienen la misma varianza: $\text{var}(\hat{\beta}_1) = \sigma^2(X_1'X_1)^{-1}$.

Consecuencias de la incorrecta especificación del modelo de RLM

Subespecificación:

- s^2 *sobreestima* σ^2

$$s^2 = \frac{1}{n-q}\hat{e}'\hat{e} \quad \text{entonces} \quad \mathbb{E}[s^2] = \sigma^2 + \frac{\beta_2' X_2' (I-H) X_2 \beta_2}{n-q} > \sigma^2$$

  Por lo tanto se detectan menos regresores significativos de los debidos.

- Para el caso de la respuesta,

$$\begin{aligned} \hat{y} &= X_1 (X_1' X_1)^{-1} X_1' y \\ &= X_1 (X_1' X_1)^{-1} X_1' (X_1 \beta_1 + X_2 \beta_2 + \epsilon) \\ &= X_1 \beta_1 + X_1 (X_1' X_1)^{-1} X_1' X_2 \beta_2 + \eta \end{aligned}$$

  donde $\eta = X_1 (X_1' X_1)^{-1} X_1' \epsilon$. Entonces, puesto que $\mathbb{E}[\eta] = 0$, se tiene

$$\mathbb{E}[\hat{y}] = X_1 \beta_1 + X_1 (X_1' X_1)^{-1} X_1' X_2 \beta_2 \quad \text{y} \quad \text{var}(\hat{y}) = \sigma^2 X_1 (X_1' X_1)^{-1} X_1' = \sigma^2 H$$

  Esto es, la respuesta ajustada es sesgada y para el residual $\hat{e} = y - \hat{y}$, se tiene que

$$\mathbb{E}[\hat{e}] = (I-H) X_2 \beta_2 \quad \text{y} \quad \text{var}(\hat{e}) = \text{var}[(I-H)y] = \sigma^2 (I-H)$$

  la subestimación sesga los residuales $\hat{e}$ pero no afecta la varianza.

−ebz

Consecuencias de la incorrecta especificación del modelo de RLM

Sobre-especificación:

Por otro lado, supongamos ahora que ajustamos el modelo

$$y = X_1\beta_1 + X_2\beta_2 = X\beta + \epsilon$$

pero que en realidad $\mathbb{E}[y] = X_1\beta_1$. Entonces,

$$\mathbb{E}[\hat{\beta}] = \begin{bmatrix} \beta_1 \\ 0 \end{bmatrix} ; \quad \text{y} \quad \mathbb{E}[\hat{y}] = X_1\beta_1$$

Esto es, el estimador $\hat{\beta}$ es insesgado, al igual que $\hat{y}$, la respuesta ajustada. Sin embargo, se tiene que

$$(X'X)^{-1} = \begin{pmatrix} (X_1'X_1)^{-1} + LML' & -LM \\ -ML' & M \end{pmatrix}$$

con $M = I - H$ y $L = (X_1'X_1)^{-1}X_1'X_2$. Luego,

$$\text{varianza “}\textit{aparente}\text{”}(\hat{\beta}_i) = \text{varianza “}\textit{real}\text{”}(\hat{\beta}_i) + (LML)_{ii}$$

Esto es, la varianza de $\hat{\beta}_1$ esta *“inflada”* ($LML' > 0$), y por lo tanto se notarían menos regresores significativos de los debidos.

Criterios para evaluar modelos con subconjuntos de variables

Supóngase que hay una base de K posibles variables independientes o regresores para ser incluidos en el modelo de regresión lineal

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i; \qquad i = 1, \ldots, n$$

1. Coeficiente de determinación múltiple

$$R^2_{(q)} = \frac{\mathrm{SC}_{\mathrm{Reg}}(\beta_{(q)})}{\mathrm{SC}_{\mathrm{Tot}}}$$

- Hay $\binom{K}{p}$ distintas combinaciones que incluyen $p = q - 1$ regresores.
- $R^2_{(q)}$ necesariamente crece con q.
- R^2 indica el porcentaje de la variación explicado por la regresión (siempre que incluya el término constante β_0).

2. Coeficiente de determinación múltiple ajustado

$$\bar{R}^2_{(q)} = 1 - \frac{n-1}{n-q}(1 - R^2_{(q)})$$

- No necesariamente aumenta al aumentar q.
- En casos extremos $\bar{R}^2$ puede ser negativo.

Criterios para evaluar modelos con subconjuntos de variables

3. Error cuadrático medio (ECM)

$$\text{ECM}_{(q)} = \frac{\text{SC}_{\text{Res}}(q)}{n-q} = s_q^2$$

- Aquel subconjunto de p regresores que minimiza $\text{ECM}_{(q)}$, maximiza $R^2_{(q)}$.

4. C_p de Mallows

$$C_p = \frac{\text{SC}_{\text{Res}}(p)}{\hat{\sigma}^2} - n + 2(p+1)$$

- Si el modelo de p regresores tiene un sesgo despreciable, entonces

$$\mathbb{E}[C_p|\text{Sesgo}=0] = \frac{(n-(p+1))\sigma^2}{\sigma^2} - n + 2(p+1) = p+1,$$

  donde $\hat{\sigma}^2$ es CM_{Res} del modelo completo.
- Elija aquel modelo de p regresores *"más cercano a la recta"*.
- Elija aquel modelo con mínimo C_p.

Criterios para evaluar modelos con subconjuntos de variables

5. Criterio de Información de Akaike (AIC)

Si la varianza es desconocida,

$$\text{AIC} = n \log\left(\frac{\text{SC}_{\text{Res}}}{n}\right) + 2p + \text{cte.}$$

Si la varianza es conocida, AIC es equivalente a C_p de Mallows.

- Elija aquel modelo con menor AIC.

6. Suma de cuadrados de los errores de predicción (PRESS)

Si el modelo se desea para predicción, elija aquel que minimice el PRESS:

$$\text{PRESS}_{(p)} = \sum_{i=1}^{n}(y_i - \hat{y}_{(i)})^2 = \sum_{i=1}^{n}\left(\frac{\hat{e}_i}{1 - h_{ii}}\right)^2$$

Procedimientos computacionales

1. Todas las posibles regresiones
Si hay K regresores potenciales, entonces hay 2^K distintos modelos de regresión lineal.

$$\begin{aligned}
y &= \beta_0 + \epsilon \\
y &= \beta_0 + \beta_1 x_1 + \epsilon \\
\cdot & \quad \cdots \\
y &= \beta_0 + \beta_1 x_K + \epsilon \\
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \\
\cdot & \quad \cdots \\
y &= \beta_0 + \beta_1 x_{K-1} + \beta_2 x_K + \epsilon \\
\cdot & \quad \cdots \\
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + \epsilon
\end{aligned}$$

2. Búsqueda directa en t Ajuste el modelo

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + \epsilon$$

y quédese con aquellos regresores con más grande

$$t_j = \frac{\hat{\beta}_j}{\text{ee}(\hat{\beta}_j)}, \quad j = 1, \ldots, K$$

–ebz

3. Regresión por pasos - Criterio F-parcial

1. *Selección hacia adelante*. Se selecciona aquel regresor x_j con máxima correlación con la respuesta y. Después, se elige aquel regresor con máxima correlación con el residual de la primer regresión, $r_1 = y - \hat{\alpha}_0 - \hat{\alpha}_1 x_j$. Se continua así hasta que el regresor por entrar no tenga un valor mínimo de F–parcial (F_{in}).
2. *Selección hacia atrás*. Se incluyen los K regresores y se eliminan aquellos con mínima $|t_j|$ hasta que el siguiente candidato tenga un valor de F–parcial (ó t) mínimo. (F_{out}).
3. *Selección a pasos.* A cada paso se revisan todos los regresores. Puede suceder que un regresor que ya haya sido incluido en el modelo termine excluido del mismo.

4. Regresión por pasos - Criterio de Akaike (AIC)

Similar a los procedimientos anteriores, se busca eficientemente el modelo con el menor AIC.

Los distintos criterios no terminan necesariamente con los mismos modelos.

Ejemplo: Datos sobre cemento de Hald[1]

El siguiente juego de datos es sobre el endurecimiento de cemento Portland, famoso por su nada fácil modelación.

variable	concepto
x_1	Cantidad de tricalcio de aluminiato, 3 $CaO \cdot Al_2O_3$.
x_2	Cantidad de tricalcio de silicato, 3 $CaO \cdot SiO_2$.
x_3	Cantidad de tricalcio de aluminio ferrito, 4 $CaO \cdot Al_2O_3 \cdot Fe_2O_2$.
x_4	Cantidad de dicalcio de silicato, 2 $CaO \cdot SiO_2$.
y	Calor en calorías por gramo de cemento.

Los regresores, x_1, x_2, x_3, x_4 son medidos como porcentaje del peso de las *ollas* donde se hace el cemento.

obs	x_1	x_2	x_3	x_4	y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

[1] Draper & Smith (1998).

Ejemplo: Datos sobre cemento de Hald (cont.)

Criterios para la selección de modelos

			Estadísticos						Coeficientes				
k	p	q	s	S^2	R^2	$\bar{R}^2$	C_p	AIC	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
1	0	1	15.04	226.31	·	·	442.92	·	95.42	·	·	·	·
2	1	2	10.73	115.06	0.53	0.49	202.55	102.41	81.48	1.869	·	·	·
3	1	2	9.08	82.39	0.67	0.64	142.49	98.07	57.42	·	0.789	·	·
4	1	2	13.28	176.31	0.29	0.22	315.15	107.96	110.20	·	·	−1.256	·
5	1	2	8.96	80.35	0.67	0.64	138.73	97.74	117.57	·	·	·	−0.738
6	2	3	2.41	5.79	0.98	0.97	2.68	64.31	52.58	1.468	0.662	·	·
7	2	3	11.08	122.71	0.55	0.46	198.09	104.01	72.34	2.312	·	0.494	·
8	2	3	2.73	7.48	0.97	0.97	5.50	67.63	103.10	1.400	·	·	−0.614
9	2	3	6.45	41.54	0.85	0.82	62.44	89.93	72.08	·	0.731	−1.008	·
10	2	3	9.32	86.89	0.68	0.62	138.23	99.52	94.16	·	0.311	·	−0.457
11	2	3	4.19	17.57	0.94	0.92	22.37	78.74	131.28	·	·	−1.200	−0.724
12	3	4	2.31	5.35	0.98	0.98	3.04	63.90	48.19	1.696	0.657	0.250	·
13	3	4	2.31	5.33	0.98	0.98	3.02	63.87	71.65	1.452	0.416	·	−0.237
14	3	4	2.38	5.65	0.98	0.98	3.50	64.62	203.64	·	−0.923	−1.448	−1.557
15	3	4	2.86	8.20	0.97	0.96	7.34	69.47	111.68	1.052	·	−0.410	−0.643
16	4	5	2.45	5.98	0.98	0.97	5.00	65.84	62.41	1.551	0.510	0.102	−0.144

Ejemplo: Datos sobre cemento de Hald (cont.)

Regresión por pasos - criterio Akaike (AIC)

Función `stepAIC` del paquete MASS

El paquete *MASS* de R ofrece la búsqueda de "el mejor modelo" lineal (y otras familias) con base en el *criterio de Akaike*.
La función `stepAIC` es la utilizada para la selección del modelo.

```
R > dat <- read.table('../data/Hald.dat', header=TRUE)
R > names(dat) <- c("C1","C2","C3","C4","Y")
R > modAIC <- stepAIC(lm(Y~1,data=dat), scope=Y~C1+C2+C3+C4,
                    direction=c("both","forward","backward")[2], trace=3)
```

Ejemplo: Datos sobre cemento de Hald (cont.)

Función `stepAIC`

```
Start:  AIC=71.44
Y ~ 1

       Df Sum of Sq     RSS    AIC
+ C4    1   1831.90  883.87 58.852
+ C2    1   1809.43  906.34 59.178
+ C1    1   1450.08 1265.69 63.519
+ C3    1    776.36 1939.40 69.067
<none>              2715.76 71.444

Step:  AIC=58.85
Y ~ C4

       Df Sum of Sq     RSS    AIC
+ C1    1    809.10   74.76 28.742
+ C3    1    708.13  175.74 39.853
<none>               883.87 58.852
+ C2    1     14.99  868.88 60.629
```

```
Step:  AIC=28.74
Y ~ C4 + C1

       Df Sum of Sq    RSS    AIC
+ C2    1    26.789 47.973 24.974
+ C3    1    23.926 50.836 25.728
<none>              74.762 28.742

Step:  AIC=24.97
Y ~ C4 + C1 + C2

       Df Sum of Sq    RSS    AIC
<none>              47.973 24.974
+ C3    1   0.10909 47.864 26.944
```

–ebz

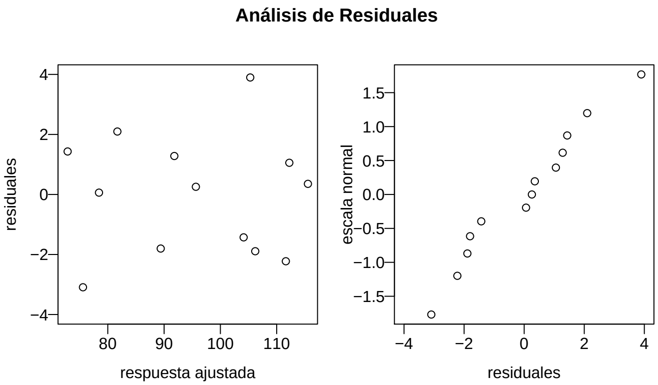
Ejemplo: Datos sobre cemento de Hald (cont.)

Modelo seleccionado mediante criterio AIC

```
R > print(summary(modAIC))
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  71.6483    14.1424   5.066 0.000675
x4           -0.2365     0.1733  -1.365 0.205395
x1            1.4519     0.1170  12.410 5.78e-07
x2            0.4161     0.1856   2.242 0.051687

Residual standard error: 2.309 on 9 degrees of freedom
Multiple R-squared:  0.9823,Adjusted R-squared:  0.9764
F-statistic: 166.8 on 3 and 9 DF,  p-value: 3.323e-08

R > print(anova(modAIC))
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq  F value    Pr(>F)
x4         1 1831.90 1831.90 343.6758 1.771e-08
x1         1  809.10  809.10 151.7934 6.150e-07
x2         1   26.79   26.79   5.0259   0.05169
Residuals  9   47.97    5.33
```

−ebz

Ejemplo: Datos sobre cemento de Hald (cont.)

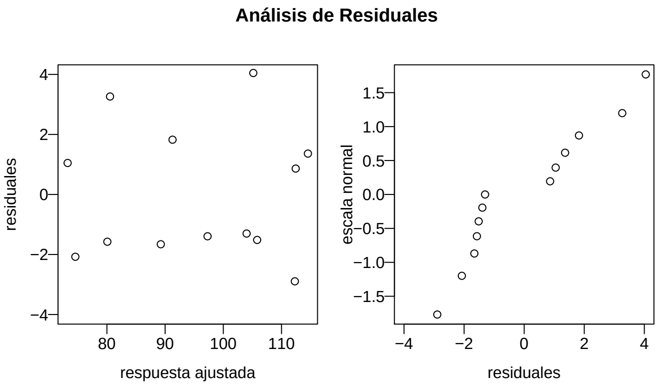


Ejemplo: Datos sobre cemento de Hald (cont.)

Modelo seleccionado mediante criterio C_p de Mallows

```
R > print(summary(mod2Reg))
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 52.57735    2.28617   23.00 5.46e-10
x1           1.46831    0.12130   12.11 2.69e-07
x2           0.66225    0.04585   14.44 5.03e-08

Residual standard error: 2.406 on 10 degrees of freedom
Multiple R-squared:  0.9787,Adjusted R-squared:  0.9744
F-statistic: 229.5 on 2 and 10 DF,  p-value: 4.407e-09

R > print(anova(mod2Reg))
Analysis of Variance Table

Response: y
Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 1450.1 1450.08  250.43 2.088e-08
x2         1 1207.8 1207.78  208.58 5.029e-08
Residuals 10   57.9    5.79
```

Ejemplo: Datos sobre cemento de Hald (cont.)



Coeficientes estandarizados - *Coeficientes Beta*

En el modelo de regresión

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i; \qquad i = 1, \ldots, n$$

las unidades de los coeficientes β_j son las de la respuesta y sobre las del correspondiente regresor x_j. Luego, la *importancia relativa* de los regresores comparando la magnitud de los coeficientes es válida *solamente* si las unidades y los niveles de los regresores es la misma. Una manera de facilitar la comparación es entonces mediante la estandarización de la variable respuesta y los regresores.
Considere las variables *estandarizadas*, para $i = 1, \ldots, n$,

$$Y_i = \frac{y_i - \bar{y}}{s_y}; \qquad X_{ij} = \frac{x_{ij} - \bar{x}_{\cdot j}}{s_{x_j}}, \quad j = 1, \ldots, p$$

donde $s_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2$ y $s_{x_j}^2 = \sum_{i=1}^n (x_{ij} - \bar{x}_{\cdot j})^2$.
Entonces, se tiene que $\bar{Y} = 0$, $\bar{X}_j = 0$, $j = 1, \ldots, p$, y

$$\sum_{i=1}(Y_i - \bar{Y})^2 = 1 = \sum_{i=1}^{n}(X_{ij} - \bar{X}_{\cdot j})^2, \quad j = 1, \ldots, p$$

Coeficientes estandarizados - *Coeficientes Beta*

Así el modelo de regresión se puede escribir ahora como $Y = \underline{X}B + \epsilon$, donde

$$Y_i = B_1 X_{i1} + \cdots + B_p X_{ip} + \epsilon_i; \qquad i = 1, \ldots, n$$

y la importancia de los regresores se puede valorar comparando la magnitud del coeficiente estimado (MCO), donde

$$\hat{B} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T Y$$

Note que el modelo no tiene término constante, pues al estar la respuesta centrada $\hat{B}_0 = \bar{Y} = 0$.
Finalmente, para recuperar los coeficientes $\hat{\beta}$ a partir de $\hat{B}$,

$$\hat{\beta}_j = \hat{B}_j \left(\frac{s_y}{s_{x_j}} \right)^{1/2}, \qquad j = 1, \ldots, p$$

y

$$\hat{\beta}_0 = \bar{y} - \sum_{j=1}^{p} \hat{\beta}_j \bar{x}_{\cdot j}$$

−ebz

Ejemplo: Televisión por cable (cont.)

Coeficientes estandarizados - *Coeficientes Beta*

```
dat <- read.table("cableTV.dat",header=1)
dat$valCat  <-  dat$valor/1000

mod1 <- lm(renta ~ ninos + adultos + tvtot + valCat, data=dat)
print(summary(mod1)$coefficients)

cat("\n>>> Coeficientes Beta <<<\n")
library(lsr)
print(standardCoefs(mod1))
```

```
              Estimate  Std. Error    t value     Pr(>|t|)
(Intercept)  9.8056592 10.32633150  0.9495782 0.3488378682
ninos       -4.9143208  2.73477110 -1.7969770 0.0809733319
adultos      2.6400570  2.44211099  1.0810553 0.2870653625
tvtot        0.4505260  0.11444988  3.9364482 0.0003750645
valCat       0.1298918  0.03140986  4.1353819 0.0002106499

>>> Coeficientes Beta <<<
                b       beta
ninos   -4.9143208 -0.2935141
adultos  2.6400570  0.1331888
tvtot    0.4505260  0.6115323
valCat   0.1298918  0.5646786
```

−ebz