Regresión múltiple

AJUSTE DE MODELOS CON MÁS DE UNA VARIABLE PREDICTORA

Ideas básicas

Explorar qué variables (cuantitativas) pueden explicar los valores de una variable (cuantitativa) de interés.

Ajustar un modelo lineal con varias variables predictoras.

Utilizaremos los datos del fichero fat.R disponible en recursos.

```
fat <- readRDS('fat.R')
```

Variables:

```
> head(fat)
  case body.fat body.fat.siri density age weight height  BMI ffweight neck chest abdomen   hip thigh knee
1    1     12.6          12.3  1.0708  23 154.25  67.75 23.7    134.9 36.2  93.1    85.2  94.5  59.0 37.3
2    2      6.9           6.1  1.0853  22 173.25  72.25 23.4    161.3 38.5  93.6    83.0  98.7  58.7 37.3
3    3     24.6          25.3  1.0414  22 154.00  66.25 24.7    116.0 34.0  95.8    87.9  99.2  59.6 38.9
4    4     10.9          10.4  1.0751  26 184.75  72.25 24.9    164.7 37.4 101.8    86.4 101.2  60.1 37.3
5    5     27.8          28.7  1.0340  24 184.25  71.25 25.6    133.1 34.4  97.3   100.0 101.9  63.2 42.2
6    6     20.6          20.9  1.0502  24 210.25  74.75 26.5    167.0 39.0 104.5    94.4 107.8  66.0 42.0
  ankle bicep forearm wrist ratio    gBMI
1  21.9  32.0    27.4  17.1  1.11  (0,25]
2  23.4  30.5    28.9  18.2  1.19  (0,25]
3  24.0  28.8    25.2  16.6  1.13  (0,25]
4  22.8  32.4    29.4  18.2  1.17  (0,25]
5  24.0  32.2    27.7  17.7  1.02 (25,30]
6  25.6  35.7    30.6  18.8  1.14 (25,30]
```

Objetivo

Construir un modelo para predecir el porcentaje de grasa corporal (body.fat)

```
> res.1 <- lm(body.fat~age,data=fat)
> summary(res.1)

Call:
lm(formula = body.fat ~ age, data = fat)

Residuals:
     Min       1Q   Median       3Q      Max
-18.0697  -5.7025   0.2846   4.8301  25.0739

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.95546    1.73576   6.312 1.25e-09 ***
age          0.17786    0.03724   4.776 3.04e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.435 on 250 degrees of freedom
Multiple R-squared:  0.08362,   Adjusted R-squared:  0.07996
F-statistic: 22.81 on 1 and 250 DF,  p-value: 3.045e-06
```

El ajuste es muy pobre. La variable ***age*** tiene cierta relación con ***body.fat*** pero no puede predecir bien sus valores.

Modelo 2:
Edad y peso como variables predictores de body.fat

```
> res.2 <- lm(body.fat~age+weight,data=fat)
> summary(res.2)

Call:
lm(formula = body.fat ~ age + weight, data = fat)

Residuals:
     Min       1Q   Median       3Q      Max
-15.3171  -4.3293   0.2917   3.9898  18.5237

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -18.37392    2.57545  -7.134 1.06e-11 ***
age           0.18269    0.02853   6.403 7.54e-10 ***
weight        0.16271    0.01224  13.298  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.696 on 249 degrees of freedom
Multiple R-squared:  0.4642,    Adjusted R-squared:  0.4599
F-statistic: 107.9 on 2 and 249 DF,  p-value: < 2.2e-16
```

Al considerar edad y peso como predictores, el ajuste es mejor.

Comparación de modelos

```
> anova(res.1,res.2)
Analysis of Variance Table

Model 1: body.fat ~ age
Model 2: body.fat ~ age + weight
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    250 13818.1
2    249  8079.7  1    5738.4 176.85 < 2.2e-16 ***
```

Al incorporar una nueva variable predictora, la suma de cuadrados residual pasa de 13818 a 8079.

Esta reducción, teniendo en cuenta que aumentamos la complejidad del modelo, es significactiva (p=2.2e-16).

Podemos concluir que el modelo 2 es más adecuado.

Sin embargo, el modelo no es mut bueno (r^2 vale solo 0.46)

Selección de variables predictores para body.fat

```
res.cor <- cor(fat %>%
                 dplyr::select(-c(gBMI,case,body.fat.siri))) %>%
           round(2)
as.data.frame(res.cor) %>% formattable()
```

	body.fat	density	age	weight	height	BMI	ffweight	neck	chest	abdomen	hip	thigh	knee	ankle	bicep	forearm	wrist	ratio
body.fat	1.00	-0.99	0.29	0.61	-0.09	0.73	0.02	0.49	0.70	0.81	0.63	0.56	0.51	0.27	0.49	0.36	0.35	-0.77
density	-0.99	1.00	-0.28	-0.59	0.10	-0.71	-0.01	-0.47	-0.68	-0.80	-0.61	-0.55	-0.50	-0.26	-0.49	-0.35	-0.33	0.76
age	0.29	-0.28	1.00	-0.01	-0.17	0.12	-0.24	0.11	0.18	0.23	-0.05	-0.20	0.02	-0.11	-0.04	-0.09	0.21	-0.47
weight	0.61	-0.59	-0.01	1.00	0.31	0.89	0.79	0.83	0.89	0.89	0.94	0.87	0.85	0.61	0.80	0.63	0.73	-0.54
height	-0.09	0.10	-0.17	0.31	1.00	-0.02	0.49	0.25	0.13	0.09	0.17	0.15	0.29	0.26	0.21	0.23	0.32	0.02
BMI	0.73	-0.71	0.12	0.89	-0.02	1.00	0.55	0.78	0.91	0.92	0.88	0.81	0.71	0.50	0.75	0.56	0.63	-0.66
ffweight	0.02	-0.01	-0.24	0.79	0.49	0.55	1.00	0.68	0.59	0.50	0.70	0.68	0.70	0.58	0.65	0.55	0.67	-0.11
neck	0.49	-0.47	0.11	0.83	0.25	0.78	0.68	1.00	0.78	0.75	0.73	0.70	0.67	0.48	0.73	0.62	0.74	-0.54
chest	0.70	-0.68	0.18	0.89	0.13	0.91	0.59	0.78	1.00	0.92	0.83	0.73	0.72	0.48	0.73	0.58	0.66	-0.72
abdomen	0.81	-0.80	0.23	0.89	0.09	0.92	0.50	0.75	0.92	1.00	0.87	0.77	0.74	0.45	0.68	0.50	0.62	-0.82
hip	0.63	-0.61	-0.05	0.94	0.17	0.88	0.70	0.73	0.83	0.87	1.00	0.90	0.82	0.56	0.74	0.55	0.63	-0.45
thigh	0.56	-0.55	-0.20	0.87	0.15	0.81	0.68	0.70	0.73	0.77	0.90	1.00	0.80	0.54	0.76	0.57	0.56	-0.38
knee	0.51	-0.50	0.02	0.85	0.29	0.71	0.70	0.67	0.72	0.74	0.82	0.80	1.00	0.61	0.68	0.56	0.66	-0.42
ankle	0.27	-0.26	-0.11	0.61	0.26	0.50	0.58	0.48	0.48	0.45	0.56	0.54	0.61	1.00	0.48	0.42	0.57	-0.19
bicep	0.49	-0.49	-0.04	0.80	0.21	0.75	0.65	0.73	0.73	0.68	0.74	0.76	0.68	0.48	1.00	0.68	0.63	-0.42
forearm	0.36	-0.35	-0.09	0.63	0.23	0.56	0.55	0.62	0.58	0.50	0.55	0.57	0.56	0.42	0.68	1.00	0.59	-0.31
wrist	0.35	-0.33	0.21	0.73	0.32	0.63	0.67	0.74	0.66	0.62	0.63	0.56	0.66	0.57	0.63	0.59	1.00	-0.42
ratio	-0.77	0.76	-0.47	-0.54	0.02	-0.66	-0.11	-0.54	-0.72	-0.82	-0.45	-0.38	-0.42	-0.19	-0.42	-0.31	-0.42	1.00

Selección de variables predictores para body.fat

body.fat	1.00
density	-0.99
age	0.29
weight	0.61
height	-0.09
BMI	0.73
ffweight	0.02
neck	0.49
chest	0.70
abdomen	0.81
hip	0.63
thigh	0.56
knee	0.51
ankle	0.27
bicep	0.49
forearm	0.36
wrist	0.35
ratio	-0.77

```
res.cor[1,] %>% as.data.frame %>% formattable()
```

Como primera opción, escogeremos como posibles variables candidatas a buenos predictores de body.fat a las que tengan una correlación superior a 0.4 en valor absoluto. Añadimos también la edad.

Nos interesan variable que sean fáciles de medir. Por ello, prescindimos de ***density***.

Selección de variables

```
res.3 <- lm(body.fat~age+weight+BMI+
            neck+chest+
            abdomen+hip+knee+
            bicep+forearm+
            wrist+ratio,data=fat)
summary(res.3)
```

```
Call:
lm(formula = body.fat ~ age + weight + BMI + neck + chest + abdom
    hip + knee + bicep + forearm + wrist + ratio, data = fat)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8810 -2.7109 -0.1901  2.8502  8.6986

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.31265   35.80198   0.595  0.55221
age           0.04028    0.02782   1.448  0.14898
weight       -0.07940    0.04451  -1.784  0.07572 .
BMI           0.33357    0.24081   1.385  0.16728
neck         -0.46577    0.21694  -2.147  0.03280 *
chest        -0.07343    0.09583  -0.766  0.44429
abdomen       0.48827    0.35265   1.385  0.16747
hip           0.21395    0.32793   0.652  0.51475
knee          0.11803    0.21717   0.543  0.58730
bicep         0.18066    0.15686   1.152  0.25056
forearm       0.41408    0.18548   2.233  0.02651 *
wrist        -1.49920    0.47891  -3.130  0.00196 **
ratio       -32.09399   28.26027  -1.136  0.25724
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4 on 239 degrees of freedom
Multiple R-squared:  0.7464,    Adjusted R-squared:  0.7337
F-statistic: 58.63 on 12 and 239 DF,  p-value: < 2.2e-16
```

Selección de variables

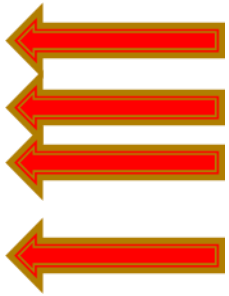
```
res.3.1 <- lm(body.fat~age+weight+BMI+
                neck+chest+
                abdomen+
                bicep+forearm+
                wrist+ratio,data=fat)
summary(res.3.1)
```

```
Call:
lm(formula = body.fat ~ age + weight + BMI + neck + chest + abdomen +
    bicep + forearm + wrist + ratio, data = fat)

Residuals:
    Min      1Q  Median      3Q     Max
-9.9146 -2.7962 -0.1916  2.8930  8.8884

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.28710   23.32986   0.269  0.78779
age           0.04055    0.02734   1.483  0.13925
weight       -0.06532    0.04089  -1.598  0.11145
BMI           0.29222    0.23628   1.237  0.21737
neck         -0.48345    0.21443  -2.255  0.02506 *
chest        -0.08576    0.09470  -0.906  0.36605
abdomen       0.70963    0.14254   4.978 1.22e-06 ***
bicep         0.19652    0.15508   1.267  0.20631
forearm       0.42950    0.18418   2.332  0.02053 *
wrist        -1.46478    0.47524  -3.082  0.00229 **
ratio       -14.30197   10.99897  -1.300  0.19474
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.991 on 241 degrees of freedom
Multiple R-squared:  0.7455,    Adjusted R-squared:  0.7349
F-statistic: 70.58 on 10 and 241 DF,  p-value: < 2.2e-16
```



Selección de variables

```
res.3.1 <- lm(body.fat~age+weight+BMI+
              neck+chest+
              abdomen+
              bicep+forearm+
              wrist+ratio,data=fat)
summary(res.3.1)
```

```
> anova(res.1,res.2,res.3.1,res.3)
Analysis of Variance Table

Model 1: body.fat ~ age
Model 2: body.fat ~ age + weight
Model 3: body.fat ~ age + weight + BMI + neck + chest + abdomen + bicep +
    forearm + wrist + ratio
Model 4: body.fat ~ age + weight + BMI + neck + chest + abdomen + hip +
    knee + bicep + forearm + wrist + ratio
  Res.Df     RSS Df Sum of Sq        F Pr(>F)
1    250 13818.1
2    249  8079.7  1    5738.4 358.6859 <2e-16 ***
3    241  3838.0  8    4241.6  33.1411 <2e-16 ***
4    239  3823.6  2      14.4   0.4506 0.6378
```

De momento, el modelo res.3.1 sería el más adecuado. Pero todavía se puede simplificar más, eliminando variables que no tienen una influencia significativa.

Eliminamos

chest, ratio, BMI, bíceps, age y weight

```
res.3.2 <- lm(body.fat~weight+
                neck+
                abdomen+
                forearm+
                wrist,data=fat)
summary(res.3.2)
```

```
Call:
lm(formula = body.fat ~ weight + neck + abdomen + forearm + wris
    data = fat)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3460 -2.9636 -0.1398  2.9166  8.9272

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -27.41180    7.08847  -3.867 0.000141 ***
weight       -0.11370    0.02394  -4.748 3.48e-06 ***
neck         -0.33822    0.20507  -1.649 0.100370
abdomen       0.93256    0.05218  17.872  < 2e-16 ***
forearm       0.49642    0.17036   2.914 0.003897 **
wrist        -1.15196    0.43359  -2.657 0.008405 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.007 on 246 degrees of freedom
Multiple R-squared:  0.738,     Adjusted R-squared:  0.7327
F-statistic: 138.6 on 5 and 246 DF,  p-value: < 2.2e-16
```

Eliminamos
neck

```
res.3.3 <- lm(body.fat~weight+
                abdomen+
                forearm+
                wrist,data=fat)
summary(res.3.3)
```

```
Call:
lm(formula = body.fat ~ weight + abdomen + forearm + wrist, data = fat)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8002 -2.8728 -0.1545  2.8980  8.3845

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -31.29679    6.70886  -4.665 5.06e-06 ***
weight       -0.12557    0.02292  -5.479 1.05e-07 ***
abdomen       0.92137    0.05192  17.747  < 2e-16 ***
forearm       0.44638    0.16822   2.654 0.008480 **
wrist        -1.39177    0.40991  -3.395 0.000799 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.021 on 247 degrees of freedom
Multiple R-squared:  0.7351,    Adjusted R-squared:  0.7308
F-statistic: 171.4 on 4 and 247 DF,  p-value: < 2.2e-16
```

En este modelo, todas las variables son significativas y la correlación prácticamente no ha cambiado respecto al modelo con todas las variables.

$$body.fat = -31.3 - 0.13 \times weight + 0.92 \times abdomen + 0.45 \times forearm - 1.39 \times wrist$$

Comparación de modelos y decisión final

```
> anova(res.1,res.2,res.3.3,res.3.2,res.3.1,res.3)
Analysis of Variance Table

Model 1: body.fat ~ age
Model 2: body.fat ~ age + weight
Model 3: body.fat ~ weight + abdomen + forearm + wrist
Model 4: body.fat ~ weight + neck + abdomen + forearm + wrist
Model 5: body.fat ~ age + weight + BMI + neck + chest + abdomen + bicep +
    forearm + wrist + ratio
Model 6: body.fat ~ age + weight + BMI + neck + chest + abdomen + hip +
    knee + bicep + forearm + wrist + ratio
  Res.Df     RSS Df Sum of Sq        F  Pr(>F)
1    250 13818.1
2    249  8079.7  1    5738.4 358.6859 < 2e-16 ***
3    247  3994.3  2    4085.4 127.6805 < 2e-16 ***
4    246  3950.6  1      43.7   2.7305 0.09976 .
5    241  3838.0  5     112.6   1.4075 0.22219
6    239  3823.6  2      14.4   0.4506 0.63780
```

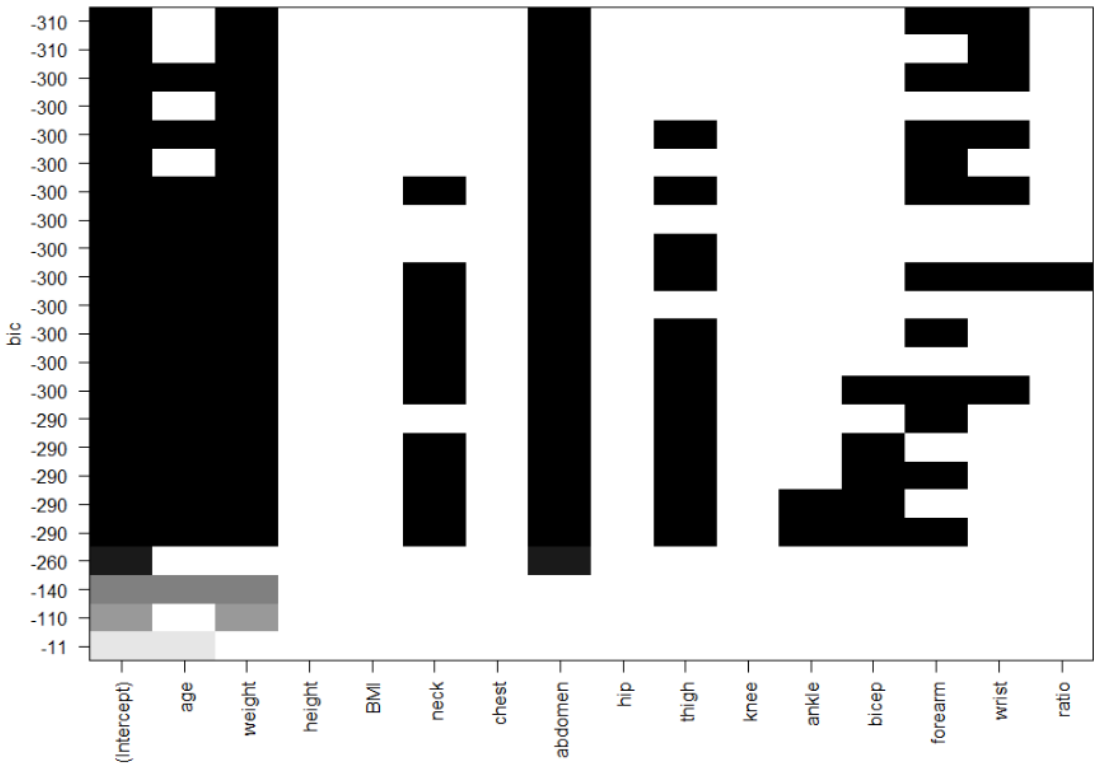
El modelo 3 mejora a los modelos más simples. La adición de nuevas variables en los modelos 4-6 no introducen una reducción significativa en la SCE (RSS: residual sum of squares). Por lo tanto, escogemos el modelo 3, que incluye weight, abdomen, forearm y wrist como predictora de body.fat con r_2=0.74.

Selección automàtica de modelos

```
ressub <- regsubsets(body.fat~.-body.fat.siri-density-
                       ffweight-case-gBMI, data=fat,nbest=3,
                     method='backward')

plot(ressub)
```

BIC: Critero de información Bayesiano (el modelo es mejor cuando BIC es menor).



Según este criterio, el mejor modelo seria con weight, abdomen, wrist i *forearm.*

Si prescindimos de *forearm*, el resultado es casi equivalente.

El modelo sugerido es el que hemos encontrado anteriormente haciendo la selección manual.

Resumen

En el modelo de regresión lineal múltiple utilizamos varias variables predictoras.

Debemos evaluar el modelo seleccionando aquellas variables que tienen una contribución significativa.

r^2 indica el % de variabilidad de la variable dependiente que se explica por su relación con las variables explicativas.

Podemos comparar modelos jerárquicos para determinar el modelo más adecuado.