Práctica 3

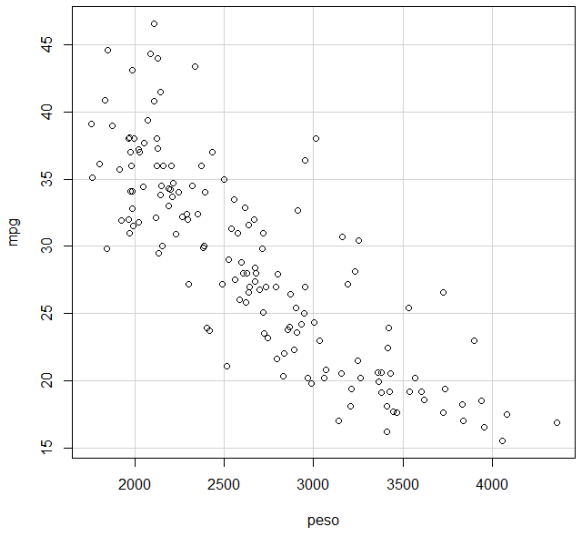
Regresión lineal

Descarga del Campus Virtual los conjuntos de datos: *Datos31*, *Datos32.*

1. Diagrama de dispersión

Importa el fichero Excel *Datos31*, en este conjunto figuran los datos de consumo y otras características de 153 automóviles.

Ejemplo 3.1. Dibuja en un diagramas de dispersión los pares de datos correspondientes a *X* = *peso* e *Y* = *mpg*.
Solución: Selecciona *Graficas* y a continuación *Diagrama de dispersión*
en la pestaña *Datos* selecciona las variables *X* e *Y*, en las pestaña *Opciones* desactiva todo y acepta.



En la nube de puntos observamos una posible relación lineal negativa entre *X* e *Y*. Debemos señalar que en Europa el consumo de mide en litros por 100 kilómetros y en USA en mpg, de ahí que salga una relación lineal negativa y no positiva como cabría esperar.

2. Coeficiente de correlación lineal de Pearson.

Ejemplo 3.2. Calcula la covarianza y el coeficiente de correlación lineal de Pearson de las variables *X* = *peso* e *Y* = mpg (consumo).

Solución: Escribe en la ventana de instrucciones

```
cov(Datos31$peso,Datos31$mpg)
cor(Datos31$peso,Datos31$mpg)
```

y ejecuta.

En la ventana de resultados aparece

```
> cov(Datos31$peso,Datos31$mpg)
[1] -3698.403
```

```
> cor(Datos31$peso,Datos31$mpg)
[1] -0.8293171
```

El valor del coeficiente de correlación nos indica una correlación negativa alta.

También puedes obtener el coeficiente de correlación seleccionando las variables *mpg* y *peso* en la ventana que se abre al seguir la secuencia

Estadísticos → *Resúmenes* → *Matriz de correlaciones*

La salida es:

```
> cor(Datos31[,c("mpg","peso")], use="complete")
           mpg        peso
mpg   1.0000000 -0.8293171
peso -0.8293171  1.0000000
```

3. Regresión lineal

En esta sección vamos a ver como se obtienen los coeficientes de correlación de *Y* sobre *X* y el valor pronosticado para *Y* a partir de un valor de *X*.

3.1. Recta de regresión

Para estimar los parámetros β_0 y β_1 del modelo de regresión lineal simple debes seguir la secuencia

Estadísticos → *Ajuste de modelos* → *Regresión lineal*

Ejemplo 3.3. Obtén la recta de regresión de *Y* = *mpg* sobre *X* = *peso.*

Solución: *Estadísticos* → *Ajuste de modelos* → *Regresión lineal*

En la ventana Regresión lineal selecciona *mpg* como variable explicada, *peso* como variable explicativa, asígnale un nombre al modelo (por defecto RegModel.1) y acepta.

Los resultados que aparecen en la ventana salida son:

```
> RegModel.1 <- lm(mpg~peso, data=Datos31)

> summary(RegModel.1)

Call:
lm(formula = mpg ~ peso, data = Datos31)

Residuals:
   Min     1Q Median     3Q    Max
-9.3100 -2.8428 -0.6126  2.1259 12.6761

Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept) 55.9929463  1.5298214    36.60   <2e-16 ***
peso        -0.0101722  0.0005578   -18.24   <2e-16 ***
```

Estimación de β_0

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Estimación de β_1

Residual standard error: 4.146 on 151 degrees of freedom
Multiple R-squared: 0.6878, Adjusted R-squared: 0.6857
F-statistic: 332.6 on 1 and 151 DF, p-value: < 2.2e-16

La columna *Estimate* proporciona los valores de las estimaciones de β_0 y β_1, con lo que el modelo de regresión lineal simple que mejor se ajusta a estos datos es:

$$mpg = 55.9929463 - 0.0101722 \cdot peso$$

La salida también nos proporciona el coeficiente de determinación $R^2 = r^2 = 0.6878$ (Multiple R-squared)

3.2. Realización de pronósticos

Los valores que proporciona la recta de regresión para un valor dado de la variable explicativa pueden interpretarse como predicciones del valor de la variable explicada.

Ejemplo 3.4. Utiliza el modelo de regresión lineal simple obtenido en el *ejemplo 3.3.* para predecir el consumo en mpg de un automóvil que pesa 4025 libras y el de otro que pesa 1780 libras.
Solución: Podemos hacerlo de dos maneras.
1) Escribe en la ventana de instrucciones y ejecuta:

55.9929463 - 0.0101722*4025; 55.9929463 - 0.0101722*1780

La salida es,

```
> 55.9929463 - 0.0101722*4025;  55.9929463 - 0.0101722*1780
[1] 15.04984
[1] 37.88643
```

2) Escribe en la ventana de instrucciones:

```
predict(RegModel.1,data.frame(peso=c(4025,1780)))
```

donde RegModel.1 es el nombre del modelo. La salida es:

```
> predict(RegModel.1,data.frame(peso=c(4025,1780)))
     1      2
15.05001 37.88650
```

A partir del modelo de regresión tenemos que para un peso de 4025 libras se estima un consumo de 15.05 mpg y para un peso de 1780 libras se estima un consumo de 37.89 mpg

4. Ejercicios propuestos

Ejercicio 1. Carga fichero de texto *Datos32 (Anscombe 1973)*. Obtén los coeficientes de correlación lineal y las rectas de regresión de Y1, Y2 e Y3 sobre X y de Y4 sobre X4. Dibuja las gráficas de dispersión de X con Y1, X con Y2, X con Y3 y X4 con Y4 ¿a qué conclusiones llegas?

Ejercicio 2. Lee el conjunto de datos *mtcars* del paquete *datasets* de *R*. Supón que estamos interesados en conocer una aproximación al consumo en *mpg* conocido el valor de una de las variables *hp, qsec* o *wt* ¿cuál de las 3 variables anteriores es la mejor variable explicativa? Una vez seleccionada la variable explicativa, obtén la recta de regresión lineal correspondiente.

Soluciones

Ejercicio 1.

```
> cor(Datos32[,c("X","X4","Y1","Y2","Y3","Y4")], use="complete")
```

	X	X4	Y1	Y2	Y3	Y4
X	1.000000000	-0.4000000	0.81642052	0.81623651	0.81628674	0.002969709
X4	-0.400000000	1.0000000	-0.29727146	-0.45071096	-0.28912321	0.816521437
Y1	0.816420516	-0.2972715	1.00000000	0.75000540	0.46871668	0.064982372
Y2	0.816236506	-0.4507110	0.75000540	1.00000000	0.58791933	-0.014442321
Y3	0.816286739	-0.2891232	0.46871668	0.58791933	1.00000000	0.022624662
Y4	0.002969709	0.8165214	0.06498237	-0.01444232	0.02262466	1.000000000

```
> RegModel.1 <- lm(Y1~X, data=Datos32)
> summary(RegModel.1)

Call:
lm(formula = Y1 ~ X, data = Datos32)

Residuals:
    Min      1Q  Median      3Q     Max
-1.92127 -0.45577 -0.04136  0.70941  1.83882

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0001     1.1247   2.667  0.02573 *
X             0.5001     0.1179   4.241  0.00217 **
---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared:  0.6665,    Adjusted R-squared:  0.6295
F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

Recta de regresión de Y1 sobre X: Y1=3 + 0.5X

```
> RegModel.2 <- lm(Y2~X, data=Datos32)
> summary(RegModel.2)

Call:
lm(formula = Y2 ~ X, data = Datos32)

Residuals:
   Min     1Q Median     3Q    Max
-1.9009 -0.7609  0.1291  0.9491  1.2691

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.001      1.125   2.667  0.02576 *
X              0.500      0.118   4.239  0.00218 **
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared:  0.6662,     Adjusted R-squared:  0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002179
```

Recta de regresión de Y2 sobre X: Y2=3 + 0.5X

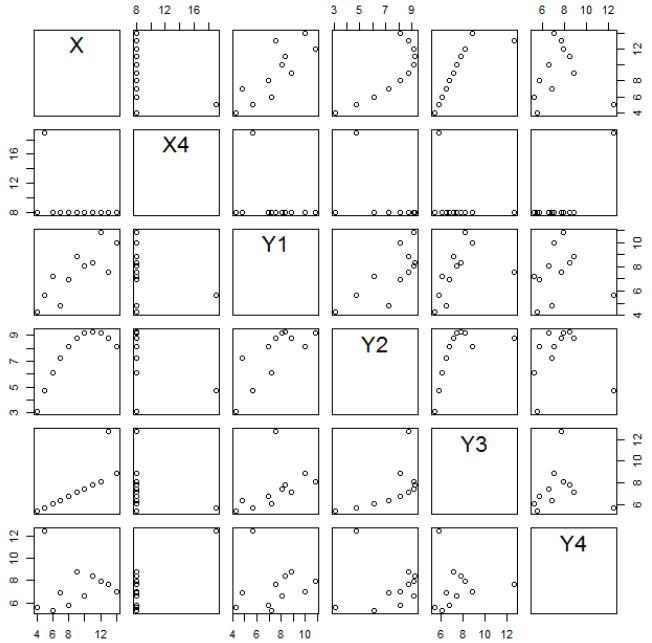
```
> RegModel.3 <- lm(Y3~X, data=Datos32)
> summary(RegModel.3)

Call:
lm(formula = Y3 ~ X, data = Datos32)
Residuals:
   Min     1Q  Median     3Q    Max
-1.1586 -0.6146 -0.2303  0.1540  3.2411

Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)   3.0025     1.1245     2.670    0.02562 *
X             0.4997     0.1179     4.239    0.00218 **
---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared:  0.6663,     Adjusted R-squared:  0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176
```

Recta de regresión de Y3 sobre X: Y3=3 + 0.5X

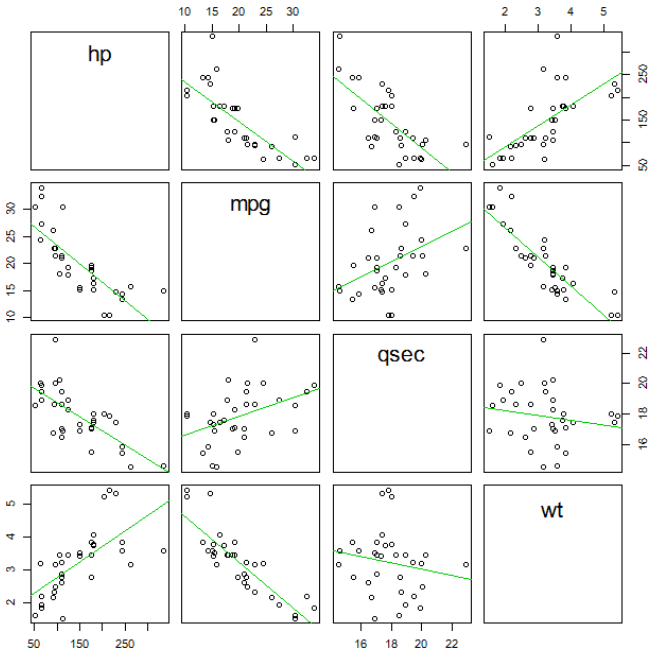
```
> RegModel.4 <- lm(Y4~X4, data=Datos32)
> summary(RegModel.4)

Call:
lm(formula = Y4 ~ X4, data = Datos32)
Residuals:
   Min     1Q Median     3Q    Max
-1.751 -0.831  0.000  0.809  1.839

Coefficients:

             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)  3.0017     1.1239      2.671    0.02559 *
X4            0.4999     0.1178     4.243     0.00216 **
---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared:  0.6667,     Adjusted R-squared:  0.6297
F-statistic:    18 on 1 and 9 DF,  p-value: 0.002165
```

Recta de regresión de Y4 sobre X4: Y4=3 + 0.5X4

Diagramas de dispersión. En lugar de dibujar cada uno por separado vamos a representarlos en una matriz de diagramas de dispersión, para ello en *Gráficas* seleccionamos *Matriz de diagramas de dispersión*, así obtenemos:



Ejercicio 2. En *Graficas* seleccionamos *matriz de diagramas de dispersión*, en la ventana emergente elegimos las variables indicadas en el problema, la matriz de diagramas de dispersión es:



Parece que la variable que mejor explica *mpg* es *wt*. Vamos a ver cuánto valen los coeficientes de correlación lineal, la matriz de correlaciones (*Estadísticos*, *resúmenes*, *matriz de correlaciones*) es:

```
> cor(mtcars[,c("hp","mpg","qsec","wt")], use="complete")
```

	hp	*mpg*	*qsec*	*wt*
hp	*1.0000000*	*-0.7761684*	*-0.7082234*	*0.6587479*
mpg	*-0.7761684*	*1.0000000*	*0.4186840*	*-0.8676594*
qsec	*-0.7082234*	*0.4186840*	*1.0000000*	*-0.1747159*
wt	*0.6587479*	*-0.8676594*	*-0.1747159*	*1.0000000*

Vemos que máximo {| *-0.7761684*|, |*0.4186840*|, |*-0.8676594*|} = |*-0.8676594*| = | cor(*mpg*,*wt*) |, por lo tanto, de las tres posible variables explicativas, la variable que mejor explica a *mpg* es *wt*.

```
> RegModel.1 <- lm(mpg~wt, data=mtcars)

> summary(RegModel.1)

Call:
lm(formula = mpg ~ wt, data = mtcars)

Residuals:
   Min     1Q  Median     3Q    Max
-4.5432 -2.3647 -0.1252  1.4096  6.8727

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt           -5.3445     0.5591  -9.559  1.29e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,    Adjusted R-squared:  0.7446
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

Recta de regresión de *mpg* sobre *wt*: *mpg*= 37.2851 - 5.3445 *wt*

Redondeando: *mpg*= 37.29 - 5.34 *wt*