

Práctica 3

Regresión y correlación lineal – Tablas de contingencia

Contenido

1. Introducción a la Correlación y Regresión.....	1
1.1. Diagrama de dispersión.....	2
1.2. Coeficiente de correlación lineal de Pearson	3
1.3. Regresión lineal simple.....	3
a) Recta de regresión.....	4
b) Realización de pronósticos	5
2. Tablas de contingencia	6
3. Ejercicios propuestos.....	7

El estudio de la relación entre dos variables depende del tipo de variables que se considera. Habrá que distinguir los siguientes tipos:

- **Cuantitativa-Cuantitativa.** El fin es establecer un modelo de regresión fiable para predecir una v.a. cuantitativa Y , a partir de la observación de una v.a. cuantitativa X .
- **Cualitativa-Cualitativa.** Se busca determinar si existe relación estadística entre los valores de dos v.a. cualitativas o de tipo factor, X e Y .
- **Cualitativa-Cuantitativa.** Este tipo de relación ya ha sido considerado en la práctica anterior. El objetivo es comparar el comportamiento de la v.a. cuantitativa, Y , en los grupos definidos por una v.a. cualitativa o factor, X .

1. Introducción a la Correlación y Regresión

Cuando se trata de estudiar si existe relación entre los valores de dos variables cuantitativas observadas en un mismo grupo de individuos, se pueden diferenciar dos tipos de estudio:

- La **correlación**, que establece el grado de relación o asociación entre dos variables cuantitativas;
- la **regresión**, que pretende expresar las relaciones entre variables cuantitativas a través de un modelo funcional y utilizar dicho modelo para predecir o aproximar el comportamiento de una variable, Y , llamada **variable** de respuesta o **dependiente** a partir del conocimiento de otra variable, X , que se denomina **variable** predictora, explicativa o **independiente**.

El tipo de relación que se puede establecer entre dos variables aleatorias continuas es una relación de tipo funcional, es decir, una función matemática que conecte la variable respuesta y la variable explicativa:

$$Y=f(X) + \varepsilon$$

El tipo de función, f , en la mayoría de los casos se desconoce, pero puede estar sugerido por cuestiones teóricas o indicado por los datos muestrales y el investigador debe elegir una función adecuada para aproximarla. No obstante, además de buscar un modelo que se ajuste bien, también suele buscarse un modelo que sea lo más sencillo posible.

La forma más sencilla para detectar si existe algún tipo de relación o dependencia entre dos variables continuas es mediante un diagrama de dispersión o nube de puntos.

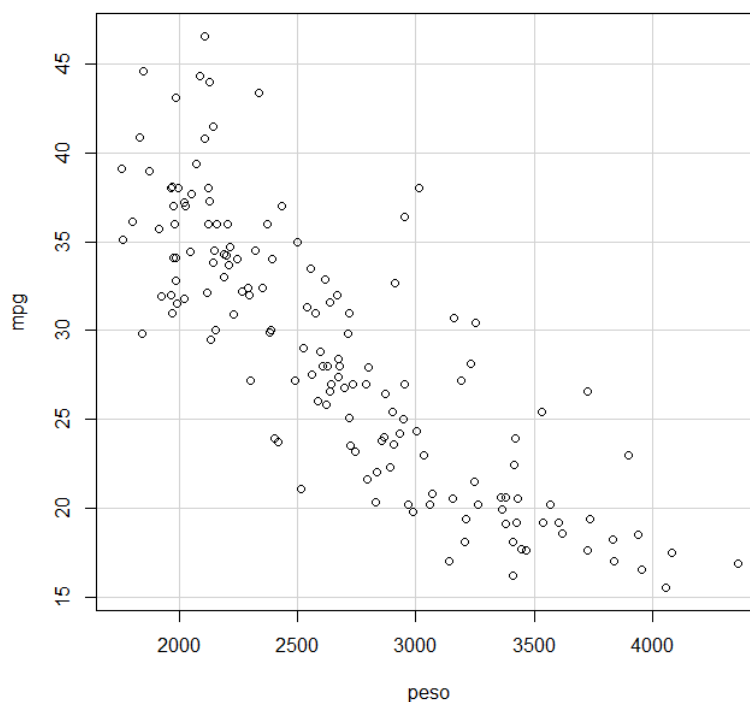
1.1. Diagrama de dispersión

Se representan ambas variables en un sistema de coordenadas. Por cada individuo i se representa el valor (x_i, y_i) y los puntos generan una nube que gráficamente puede revelar algún tipo de asociación como la lineal (si los puntos tienden a posicionarse alrededor de una línea recta), o de otro tipo.

Ejemplo 1. Importa el fichero Excel *Datos31*, en este conjunto figuran los datos de consumo y otras características de 153 automóviles.

Dibuja en un diagramas de dispersión los pares de datos correspondientes a $X = \text{peso}$ e $Y = \text{mpg}$.

Solución: Selecciona **Gráficas** 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 se 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.

Cuando la gráfica de la nube de puntos nos indica que es factible buscar la recta que mejor se “ajuste” a dicha nube, estamos ante un problema de regresión lineal que se denomina **regresión lineal simple** si se utiliza

una única variable independiente. En el caso de que la nube de puntos se parezca a una curva pero no a una recta, buscaremos un modelo de **regresión no lineal**.

Cuando se tienen varias variables que pueden utilizarse como variables explicativas de una misma variable respuesta, se denominaría modelo de **regresión múltiple** y se pueden obtener todas las gráficas de dispersión al mismo tiempo con la opción **Gráficas → Matriz de Diagramas de Dispersión**.

1.2. Coeficiente de correlación lineal de Pearson

Si consideramos dos v.a. continuas X e Y observadas conjuntamente sobre los mismos individuos. A partir de los datos muestrales que consisten en n pares observados $(x_1, y_1), \dots, (x_n, y_n)$, se puede estimar numéricamente la posible relación lineal entre las variables con el **coeficiente de correlación lineal de Pearson**.

Ejemplo 2. Calcula la covarianza y el coeficiente de correlación lineal de Pearson de las variables $X = \text{peso}$ e $Y = \text{mpg}$ (consumo).

Solución: Escribe en la ventana de instrucciones y ejecuta.

```
attach(datos31) # R recuerda la base de datos, por lo que solo hay que dar el nombre de la variable
cov(peso, mpg)
cor(peso, mpg)
```

En la ventana de resultados aparece

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

De hecho se pueden obtener los correspondientes coeficientes de correlación lineal de Pearson dos a dos de la variable explicada y todas las posibles variables explicativas de interés. De esta forma podemos determinar cuál es la “mejor” variable explicativa con el fin de estudiar el modelo de regresión lineal simple para estos datos.

1.3. Regresión lineal simple

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 .

a) Recta de regresión

El modelo más sencillo de regresión es el modelo lineal de regresión simple, es decir, con una sola variable explicativa y con relación lineal entre la variable dependiente y la independiente, que se expresa como

$$Y = a + bX + \varepsilon$$

donde a , b son los denominados **coeficientes del modelo de regresión lineal** y donde ε es una variable aleatoria (error aleatorio).

Para estimar los parámetros a y b del modelo de regresión lineal simple debes seguir la secuencia
Estadísticos → Ajuste de modelos → Regresión lineal

Ejemplo 3. Obtén la recta de regresión de $Y = mpg$ sobre $X = peso$.

Solución: En la ventana Regresión lineal selecciona *mpg* como variable explicada, *peso* como variable explicativa, asígnele 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 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 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
```

Estimación de a

Estimación de b

En la salida que proporciona el R Commander, debemos fijarnos en la estimación de los coeficientes del modelo que nos proporciona la columna *Estimate*.

En el ejemplo, con las estimaciones de a y b , 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** (Multiple R-squared), una medida que suele utilizarse para juzgar la bondad de ajuste de un modelo de regresión (lineal o no).

En el caso de que el modelo de regresión ajustado al diagrama de dispersión de X e Y sea el definido por la recta de regresión, se verifica que $R^2 = r^2$, es decir, el cuadrado del coeficiente de correlación lineal coincide con el coeficiente de determinación.

En el ejemplo que estamos considerando, $R^2 = r^2 = 0.6878$ (Multiple R-squared).

b) 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 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: En primer lugar debes comprobar que los valores 1780 y 4025 se encuentran entre los valores mínimo y máximo de la variable peso en la muestra. Para verificar esa condición puedes utilizar “*Resúmenes numéricos*” para “*peso*” o bien escribir en la ventana de instrucciones:

```
min(datos31$peso); max(datos31$peso)
```

La salida es

```
> min(datos31$peso); max(datos31$peso)
```

```
[1] 1755
```

```
[1] 4360
```

Observamos que los dos valores se encuentran entre los valores mínimo y máximo de la muestra así que podemos ir al paso siguiente. Para obtener las predicciones podemos proceder 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.

2. Tablas de contingencia

En el caso de interesar la relación entre dos variables cualitativas o factores, R Commander proporciona en el menú de Estadísticos, las **tablas de contingencia** o tablas de doble entrada para analizar los datos de un fichero.

Estadísticos → Tablas de contingencia → Tabla de doble entrada

También se puede Introducir directamente los datos de una tabla de doble entrada.

Ejemplo 2.1. Construye la tabla de contingencia entre las variables ncil (número de cilindros) y origen (1= EE.UU, 2 = Europa, 3=Japón).

Solución: Después de convertir dichas variables numéricas en variables de tipo factor. Una de ellas se introduce en Filas y la otra se introduce en Columnas.

Con la tabla de contingencia más sencilla (sin porcentajes y desactivando la opción de Test de independencia Chi-cuadrado en la pestaña de **Estadísticos**) se tiene el recuento de individuos para cada pareja de valores de las variables.

Frequency table:					
	ncil				
origen	3	4	5	6	8
EE.UU	0	44	0	24	17
Europa	0	19	3	3	0
Japón	1	39	0	3	0

En la pestaña de **Estadísticos** se pueden obtener además de las frecuencias observadas, los porcentajes totales, por filas y por columnas.

Total percentages:						
	3	4	5	6	8	Total
EE.UU	0.0	28.8	0	15.7	11.1	55.6
Europa	0.0	12.4	2	2.0	0.0	16.3
Japón	0.7	25.5	0	2.0	0.0	28.1
Total	0.7	66.7	2	19.6	11.1	100.0

Con **Porcentajes totales** se pueden observar las **distribuciones marginales** de cada una de las variables estudiadas. Así, el 55.6% de los coches proceden de EE.UU, el 16.3% de Europa y el 28.1% son coches japoneses. También se observa que el 66.7% de los coches son de 4 cilindros y el 19.6% son de 6 cilindros.

Solo fabrican coches de 3 cilindros en Japón, de 5 cilindros en Europa y de 8 cilindros en EE.UU.

Si en lugar de marcar la opción de porcentajes totales, se solicitan **Porcentajes por filas** y/o **Porcentajes por columnas** entonces se pueden deducir las **distribuciones condicionadas** de una de las variables respecto a cada uno la otra.

Row percentages:

origen	3	4	5	6	8	Total	Count
EE.UU	0.0	51.8	0	28.2	20	100	85
Europa	0.0	76.0	12	12.0	0	100	25
Japón	2.3	90.7	0	7.0	0	100	43

Column percentages:

origen	3	4	5	6	8
EE.UU	0	43.1	0	80	100
Europa	0	18.6	100	10	0
Japón	100	38.2	0	10	0
Total	100	99.9	100	100	100
Count	1	102.0	3	30	17

Por ejemplo, de los 85 coches procedentes de EE.UU, más de la mitad (51.8%) son de 4 cilindros, el 28.2% son de 6 y el 20% de los coches de EE.UU tienen 8 cilindros.

También podemos observar que entre los 102 coches con 4 cilindros, el 43.1% proceden de EE.UU, mientras que el 38.2% proceden de Japón y el 18.6% son europeos.

Nota: Las tablas de contingencia sólo tienen sentido para variables con pocas modalidades (nominales, ordinales o discretas). Si se desea representar la distribución conjunta de dos variables cuantitativas con muchas modalidades es necesario agrupar previamente los valores de cada una de las variables en pocas clases con el procedimiento **Datos / Modificar variables del conjunto de datos activo / Recodificar variables...**

3. Ejercicios propuestos

Ejercicio 1. 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.

Ejercicio 2. Con los datos del fichero *acero2.rda*,

- ¿Qué se puede decir sobre si existe o no relación lineal entre el *consumo* y la emisión de *CO*?
- Representa gráficamente el *consumo* frente a *CO* y comenta dicho gráfico.
- Ajusta un modelo de regresión lineal simple para explicar el *consumo* en función de *CO*. Calcula e interpreta el coeficiente de determinación para dicho modelo.
- Con el modelo de regresión lineal construido para estimar el *consumo* a partir de las emisiones de *CO*, estima el consumo en las horas en las que se emiten 5 t/h de monóxido de carbono. ¿y la estimación para 8 t/h?
- Si se quiere predecir el *consumo* a través de un modelo lineal simple, y las posibles variables explicativas son: la emisión de *N2O*, la emisión de *SO2* y la emisión de *NOx*, ¿cuál es la variable que debería considerarse?
- Representa gráficamente el *consumo* frente a la variable seleccionada en el apartado anterior y comenta dicho gráfico.

- g) Ajusta un modelo de regresión lineal simple para explicar el consumo en función la variable elegida en el apartado b).
- h) Calcula e interpreta el coeficiente de determinación para dicho modelo.
- i) ¿Cuánto se estima que aumenta el consumo por cada unidad de N2O emitida?
- j) Realiza una predicción del consumo energético en una hora en la que la emisión de N2O ha sido de 6 t/h.

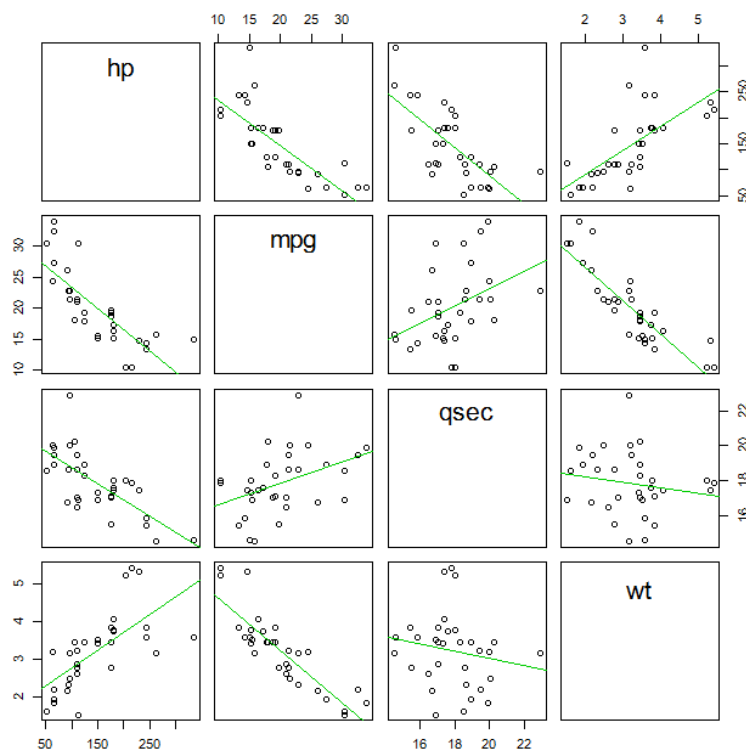
Ejercicio 3. Lee el conjunto de datos *anscombe* del paquete *datasets* de R. Obtén los coeficientes de correlación lineal y las rectas de regresión de y_i sobre x_i , $i=1, 2, 3, 4$. Dibuja las gráficas de dispersión de x_i con y_i , ¿a qué conclusiones llegas?

Ejercicio 4. Con los datos del fichero *acero2.rda*,

- a) Construye la tabla de contingencia entre las variables *averias* y *temperatura*. Deduce las distribuciones marginales de *averias* y *temperatura*.
- b) Dentro de la muestra, ¿cuántas veces ha habido temperatura alta y el sistema ha estado apagado?, ¿cuántas ha estado encendido y había temperatura media?
- c) Dentro de las horas en las que se ha trabajado con temperatura media, ¿qué porcentaje de ellas el sistema de detección de sobrecalentamiento estaba encendido?


Soluciones

Ejercicio 1. 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 $\{| -0.7761684 |, | 0.4186840 |, | -0.8676594 |\} = | -0.8676594 | = | \text{cor}(mpg, wt) |$, por lo tanto, de las tres posibles 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*:

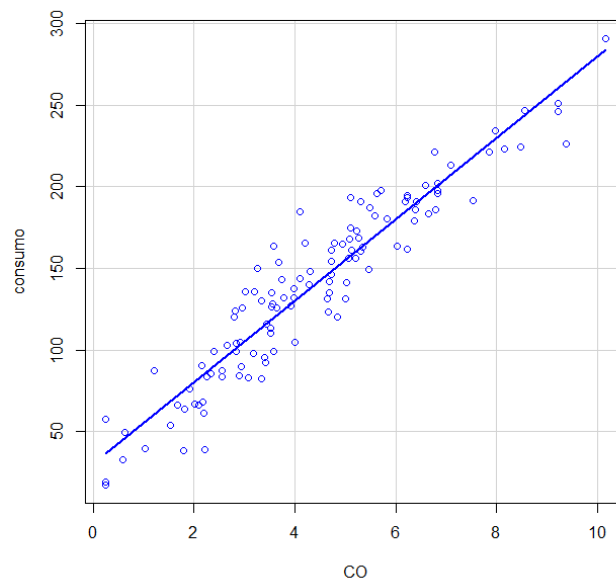
$$mpg = 37.2851 - 5.3445 \cdot wt$$

Coefficiente de determinación: 0.7528. Con el modelo de regresión lineal se explica el 75.28% de la variabilidad del consumo (*mpg*).

Ejercicio 2. Con los datos del fichero **acero2.rda**,

```
> cor(acero2[,c("CO", "consumo")], use="complete")
```

	CO	consumo
CO	1.0000000	0.9451725
consumo	0.9451725	1.0000000



```
> RegModel.1 <- lm(consumo~CO, data=acero2)
> summary(RegModel.1)
```

Call:

```
lm(formula = consumo ~ CO, data = acero2)
```

Residuals:

```
Min 1Q Median 3Q Max
-46.428 -13.530 -0.449 10.225 51.962
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	30.4748	3.8897	7.835	2.58e-12 ***
CO	24.8967	0.8022	31.037	< 2e-16 ***

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

Residual standard error: 18.1 on 115 degrees of freedom

Multiple R-squared: 0.8934, Adjusted R-squared: 0.8924

F-statistic: 963.3 on 1 and 115 DF, p-value: < 2.2e-16

Recta de regresión de *consumo* sobre *CO*: $\text{consumo} = 30.4748 + 24.8967 \text{ CO}$

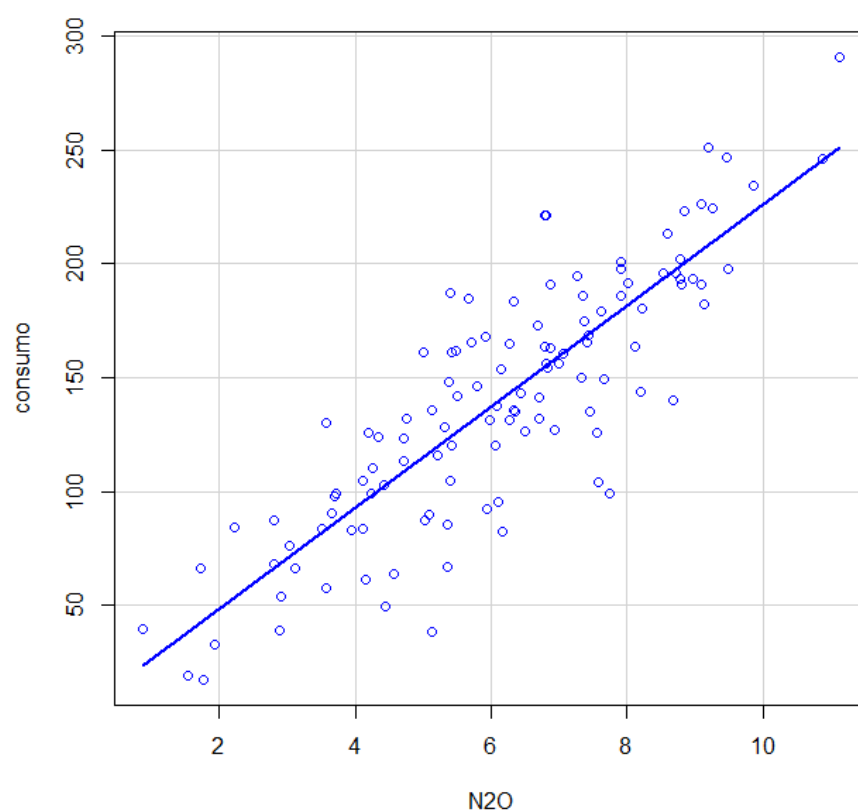
Coefficiente de determinación: 0.8934. Con el modelo de regresión lineal se explica el 89.34% de la variabilidad del consumo.

```
> numSummary(acero2[, "CO", drop=FALSE], statistics=c("quantiles"), quantiles=c(0,1))
0% 100%
0.25 10.17

> predict(RegModel.1, data.frame(CO=c(5,8)))
1 2
154.9584 229.6486
```

```
> cor(acero2[,c("consumo", "N2O", "NOx", "SO2")], use="complete")
```

	consumo	N2O	NOx	SO2
consumo	1.00000000	0.85256929	0.5582371	-0.028557917
N2O	0.85256929	1.00000000	0.5317281	0.007062938
NOx	0.55823709	0.531728148	1.0000000	-0.126162900
SO2	-0.02855792	0.007062938	-0.1261629	1.000000000



```
> RegModel.2 <- lm(consumo~N2O, data=acero2)
```

```
> summary(RegModel.2)
```

Call:

```
lm(formula = consumo ~ N2O, data = acero2)
```

Residuals:

```
Min 1Q Median 3Q Max
-79.044 -16.629 0.961 18.755 66.334
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.086	8.188	0.499	0.619
N2O	22.138	1.265	17.494	<2e-16 ***

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

Residual standard error: 28.97 on 115 degrees of freedom
 Multiple R-squared: 0.7269, Adjusted R-squared: 0.7245
 F-statistic: 306.1 on 1 and 115 DF, p-value: < 2.2e-16

Recta de regresión de *consumo* sobre *N2O*: $\text{consumo} = 4.086 + 22.138 \text{ CO}$.

Coefficiente de determinación: 0.7269. Con el modelo de regresión lineal se explica el 72.69% de la variabilidad del consumo.

```
> numSummary(acero2[, "N2O", drop=FALSE], statistics=c("quantiles"), quantiles=c(0,1))
0% 100%
0.87 11.14

> predict(RegModel.2, data.frame(N2O=c(6)))
1
136.9153
```

Ejercicio 3.

```
> cor(anscombe[,c("x1", "x2", "x3", "x4", "y1", "y2", "y3", "y4")], use="complete")
```

	x1	x2	x3	x4	y1	y2	y3	y4
x1	1.0000000	1.0000000	1.0000000	-0.5000000	0.8164205	0.8162365	0.8162867	-0.3140467
x2	1.0000000	1.0000000	1.0000000	-0.5000000	0.8164205	0.8162365	0.8162867	-0.3140467
x3	1.0000000	1.0000000	1.0000000	-0.5000000	0.8164205	0.8162365	0.8162867	-0.3140467
x4	-0.5000000	-0.5000000	-0.5000000	1.0000000	-0.5290927	-0.7184365	-0.3446610	0.8165214
y1	0.8164205	0.8164205	0.8164205	-0.5290927	1.0000000	0.7500054	0.4687167	-0.4891162
y2	0.8162365	0.8162365	0.8162365	-0.7184365	0.7500054	1.0000000	0.5879193	-0.4780949
y3	0.8162867	0.8162867	0.8162867	-0.3446610	0.4687167	0.5879193	1.0000000	-0.1554718
y4	-0.3140467	-0.3140467	-0.3140467	0.8165214	-0.4891162	-0.4780949	-0.1554718	1.0000000

```
> RegModel.1 <- lm(y1~x1, data=anscombe)
> summary(RegModel.1)
```

Call:

lm(formula = y1 ~ x1, data = anscombe)

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 *
x1	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 X1: $Y1 = 3 + 0.5X1$

```
> RegModel.2 <- lm(y2~x2, data=anscombe)
> summary(RegModel.2)
```

Call:
 lm(formula = y2 ~ x2, data = anscombe)

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 *
x2	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 X2: $Y2 = 3 + 0.5X2$

```
> RegModel.3 <- lm(y3~x3, data=anscombe)
> summary(RegModel.3)
```

Call:
 lm(formula = y3 ~ x3, data = anscombe)

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 *
x3	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 X3: $Y3=3 + 0.5X3$

```
> RegModel.4 <- lm(y4~x4, data=anscombe)
> summary(RegModel.4)
```

Call:

lm(formula = y4 ~ x4, data = anscombe)

Residuals:

Min	1Q	Median	3Q	Max
-1.751	-0.831	0.000	0.809	1.839

Coefficients:

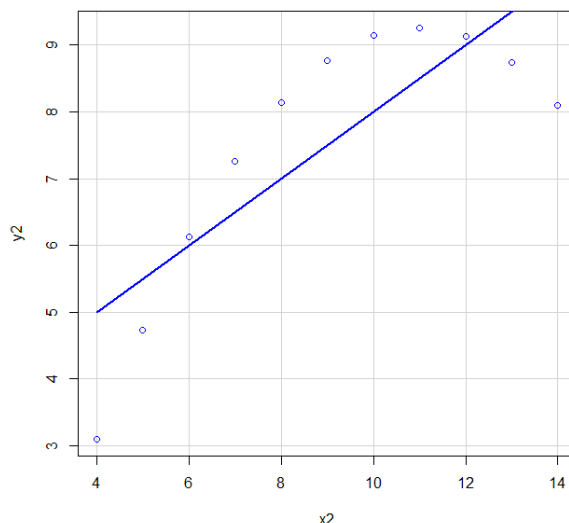
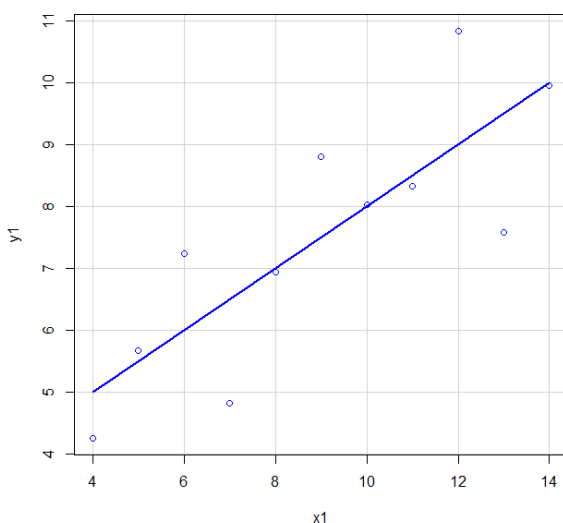
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.0017	1.1239	2.671	0.02559 *
x4	0.4999	0.1178	4.243	0.00216 **

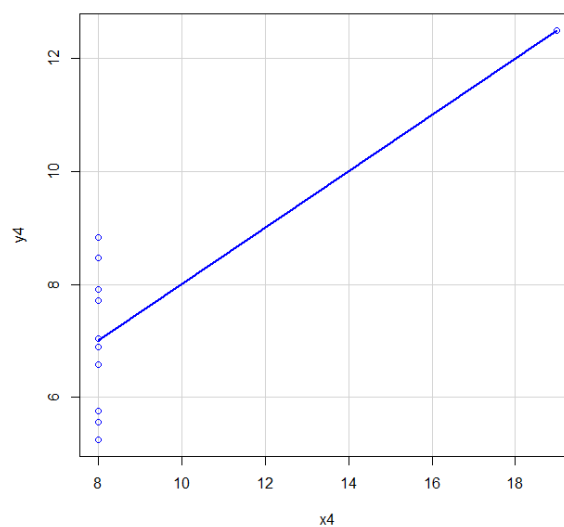
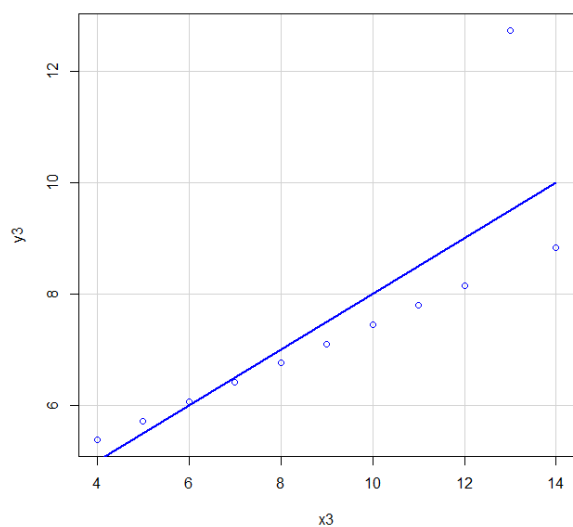
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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: $Y4=3 + 0.5X4$

En lugar de dibujar todos en una *Matriz de diagramas de dispersión*, vamos a representarlos cada uno por separado.





Ejercicio 4. Con los datos del fichero **acero2.rda**,

Frequency table:

temperatura

averias Baja Media Alta

No 24 27 38

Si 14 6 8

Total percentages:

	Baja	Media	Alta	Total
No	20.5	23.1	32.5	76.1
Si	12.0	5.1	6.8	23.9
Total	32.5	28.2	39.3	100.0

Frequency table:

sistema

temperatura OFF ON

Baja 19 19

Media 16 17

Alta 24 22

Row percentages:

sistema

temperatura	OFF	ON	Total	Count
Baja	50.0	50.0	100	38
Media	48.5	51.5	100	33
Alta	52.2	47.8	100	46