Regresión lineal simple

Los datos

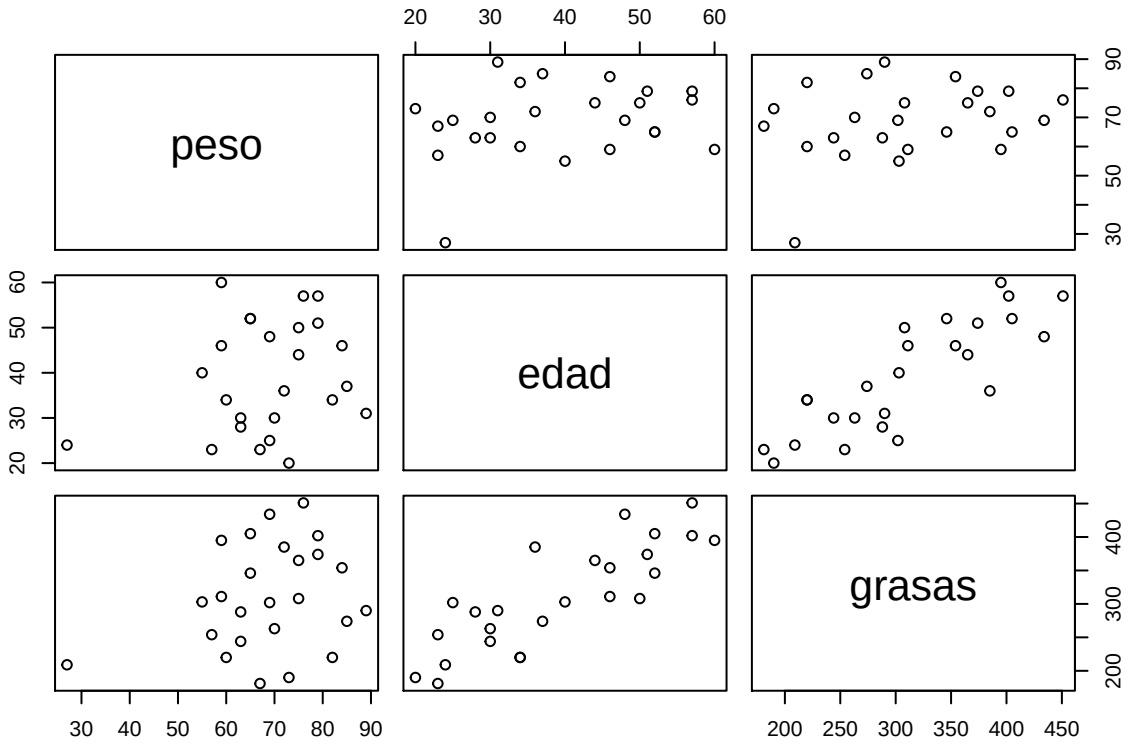
Los datos del fichero `EdadPesoGrasas.txt` corresponden a tres variables medidas en 25 individuos: edad, peso y cantidad de grasas en sangre. Para leer el fichero de datos y saber los nombres de las variables:

```
grasas <- read.table('http://www.uam.es/joser.berrendero/datos/EdadPesoGrasas.txt', header = TRUE)
names(grasas)
```

```
## [1] "peso"   "edad"   "grasas"
```

Con el fin de conocer las relaciones existentes entre cada par de variables podemos representar una matriz de diagramas de dispersión. Al parecer existe una relación lineal bastante clara entre la edad y las grasas, pero no entre los otros dos pares de variables. Por otra parte el fichero contiene un dato atípico.

```
pairs(grasas)
```



Para cuantificar el grado de relación lineal, calculamos la matriz de coeficientes de correlación:

```
cor(grasas)
```

```
##          peso   edad grasas
## peso   1.0000 0.2400 0.2653
## edad   0.2400 1.0000 0.8374
## grasas 0.2653 0.8374 1.0000
```

Cálculo y representacion de la recta de mínimos cuadrados

El comando básico es `lm` (*linear models*). El primer argumento de este comando es una fórmula `y ~ x` en la que se especifica cuál es la variable respuesta o dependiente (y) y cuál es la variable regresora o independiente (x). El segundo argumento, llamado `data` especifica cuál es el fichero en el que se encuentran las variables. El resultado lo guardamos en un objeto llamado `regresion`. Este objeto es una lista que contiene toda la información relevante sobre el análisis. Mediante el comando `summary` obtenemos un resumen de los principales resultados:

```
regresion <- lm(grasas ~ edad, data = grasas)
summary(regresion)
```

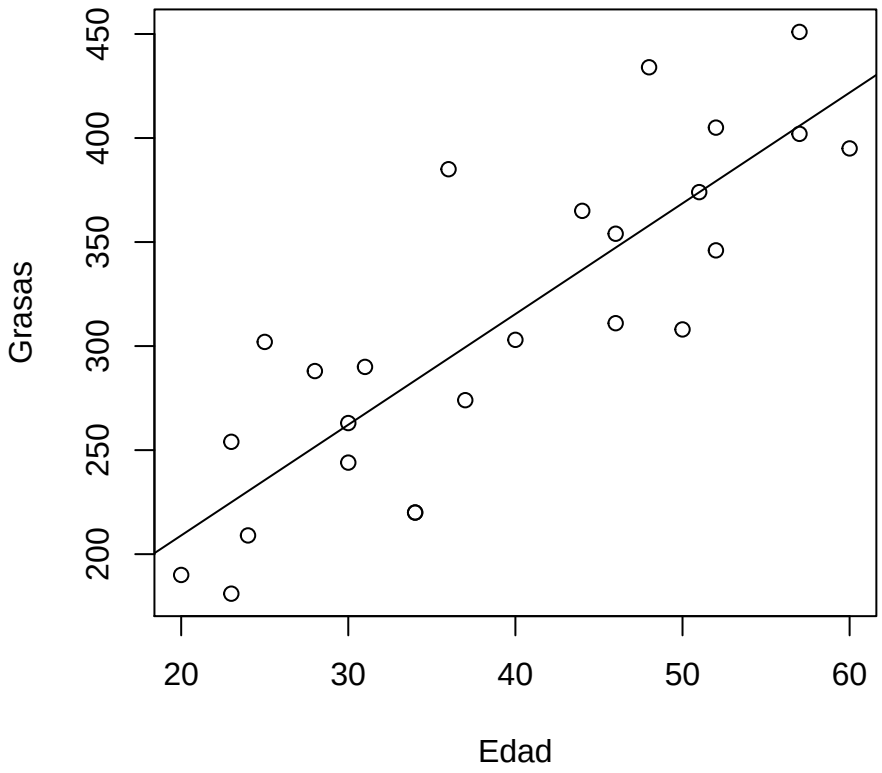
```
##
## Call:
## lm(formula = grasas ~ edad, data = grasas)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -63.48 -26.82  -3.85  28.32  90.88
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  102.575     29.638    3.46   0.0021 **
## edad           5.321      0.724    7.35  1.8e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.5 on 23 degrees of freedom
## Multiple R-squared:  0.701,  Adjusted R-squared:  0.688
## F-statistic:   54 on 1 and 23 DF,  p-value: 1.79e-07
```

Los parámetros de la ecuación de la recta de mínimos cuadrados que relaciona la cantidad de grasas en la sangre en función del peso vienen dados por la columna ´Estimate´ de la tabla ´Coefficients´ de la salida anterior. Por lo tanto, en este ejemplo la ecuación de la recta de mínimos cuadrados es:

$$y = 102.575 + 5.321x$$

Los siguientes comandos representan la nube de puntos (comando `plot`) y añaden la representación gráfica de la recta de mínimos cuadrados (comando `abline` aplicado al objeto generado por `lm`):

```
plot(grasas$edad, grasas$grasas, xlab='Edad', ylab='Grasas')
abline(regresion)
```



El **coeficiente de determinación** (es decir, el coeficiente de correlación al cuadrado) mide la bondad del ajuste de la recta a los datos. A partir de la salida anterior, vemos que su valor en este caso es `Multiple R-squared:  0.701`.

Cálculo de predicciones

Supongamos que queremos utilizar la recta de mínimos cuadrados para predecir la cantidad de grasas para individuos de edades $31, 31, 32, \ldots, 50$. Basta crear un fichero de datos que contenga las nuevas variables regresoras y usar el comando `predict`:

```
nuevas.edades <- data.frame(edad = seq(30, 50))
predict(regresion, nuevas.edades)
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12
## 262.2 267.5 272.8 278.2 283.5 288.8 294.1 299.4 304.8 310.1 315.4 320.7
##    13    14    15    16    17    18    19    20    21
## 326.0 331.4 336.7 342.0 347.3 352.6 358.0 363.3 368.6
```

Por ejemplo, para un individuo de 30 años, predecimos una cantidad de grasas de 262.2

Inferencia en el modelo de regresión simple

Suponemos ahora que los datos proceden de un modelo de regresión simple de la forma:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, n,$$

donde los errores aleatorios ϵ_i son independientes con distribución normal de media 0 y varianza σ^2.

Bajo este modelo,

- Los **errores típicos** de los estimadores de los parámetros β_0 y β_1 se encuentran en la columna `Std Error` de la salida anterior. En el ejemplo, sus valores son 29.638 y 0.724 respectivamente.
- La columna `t value` contiene el **estadístico t**, es decir, cociente entre cada estimador y su error típico. Estos cocientes son la base para llevar a cabo los contrastes $H_0 : \beta_0 = 0$ y $H_0 : \beta_1 = 0$. Los correspondientes **p-valores** aparecen en la columna `Pr(>|t|)`. En este caso son muy pequeños por lo que se rechazan ambas hipótesis para los niveles de significación habituales.
- El estimador de la **desviación típica de los errores** σ aparece como `Residual standard error` y su valor en el ejemplo es 43.5
- Los **intervalos de confianza para los parámetros** se obtienen con el comando `confint`. El parámetro `level` permite elegir el nivel de confianza (por defecto es 0.95):

```
confint(regresion)
```

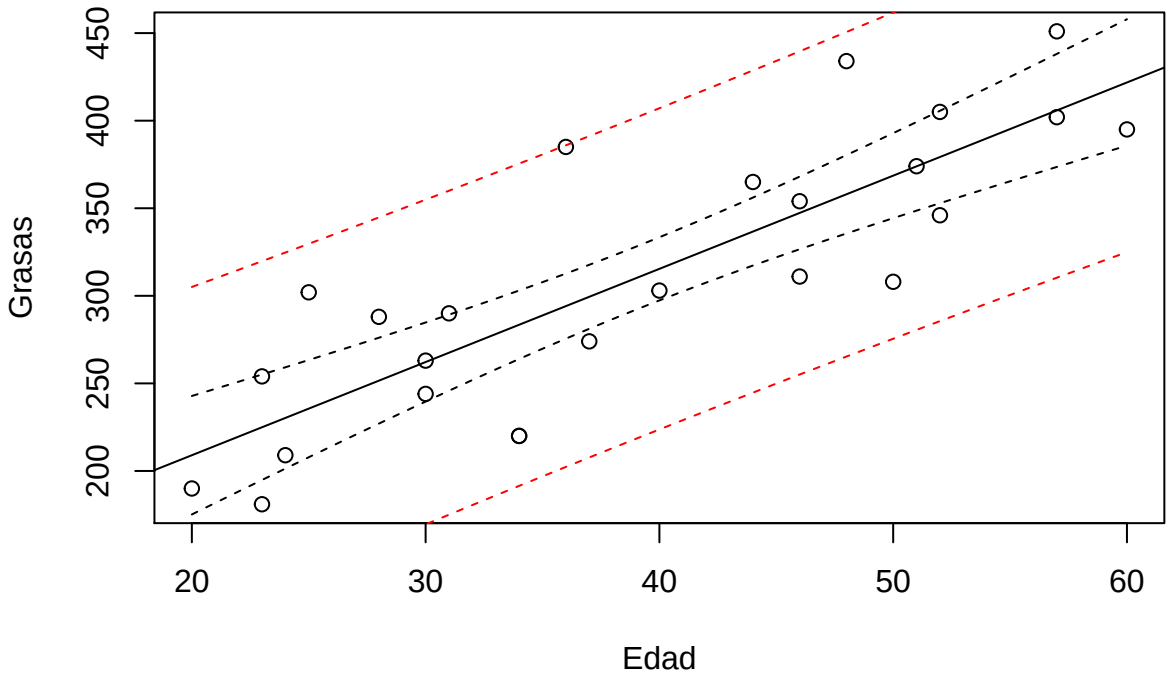
```
##               2.5 %  97.5 %
## (Intercept) 41.265 163.885
## edad         3.822   6.819
```

```
confint(regresion, level = 0.90)
```

```
##                 5 %    95 %
## (Intercept) 51.780 153.370
## edad         4.079   6.562
```

- Los **intervalos de confianza para la respuesta media** y los **intervalos de predicción para la respuesta** se pueden obtener usando el comando `predict`. Por ejemplo, el siguiente código calcula y representa los dos tipos de intervalos para el rango de edades que va de 20 a 60 años (los de predicción en rojo):

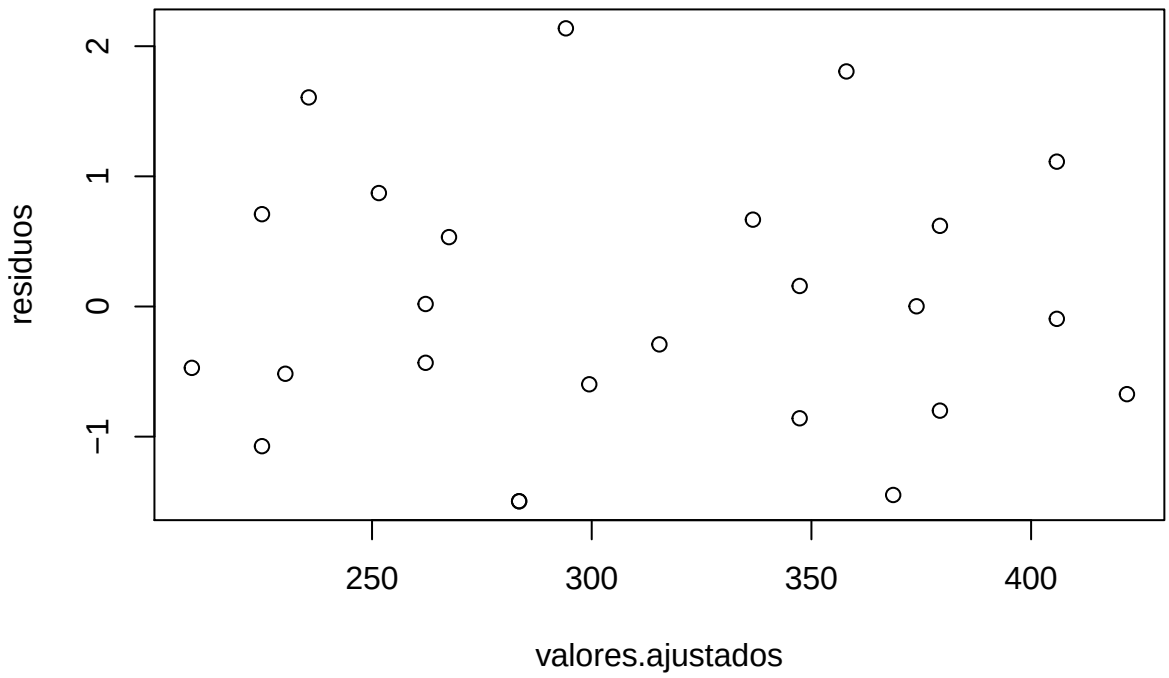
```
nuevas.edades <- data.frame(edad = seq(20, 60))
# Grafico de dispersion y recta
plot(grasas$edad, grasas$grasas, xlab='Edad', ylab='Grasas')
abline(regresion)

# Intervalos de confianza de la respuesta media:
# ic es una matriz con tres columnas: la primera es la prediccion, las otras dos son los extremos del i
ic <- predict(regresion, nuevas.edades, interval = 'confidence')
lines(nuevas.edades$edad, ic[, 2], lty = 2)
lines(nuevas.edades$edad, ic[, 3], lty = 2)

# Intervalos de prediccion
ic <- predict(regresion, nuevas.edades, interval = 'prediction')
lines(nuevas.edades$edad, ic[, 2], lty = 2, col = 'red')
lines(nuevas.edades$edad, ic[, 3], lty = 2, col = 'red')
```



Diagnóstico del modelo

Los **valores ajustados** $\hat{y}_i$ y los **residuos** $e_i = \hat{y}_i - y_i$ se pueden obtener con los comandos `fitted` y `residuals` respectivamente. Los residuos estandarizados se obtienen con `rstandard`. Por ejemplo, el siguiente código obtiene una representación de los residuos estandarizados frente a los valores ajustados, que resulta útil al llevar a cabo el diagnóstico del modelo:

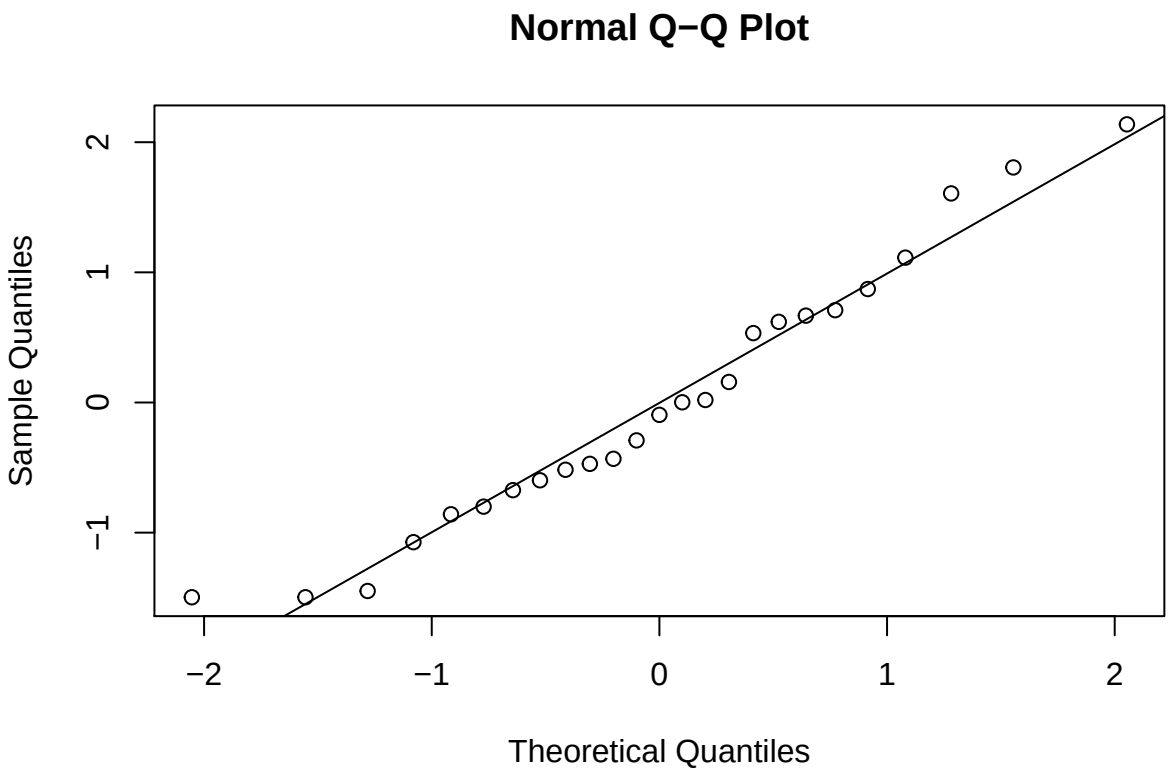
```
residuos <- rstandard(regresion)
valores.ajustados <- fitted(regresion)
plot(valores.ajustados, residuos)
```



No se observa ningún patrón especial, por lo que tanto la homocedasticidad como la linealidad resultan hipótesis razonables.

La hipótesis de normalidad se suele comprobar mediante un *QQ plot* de los residuos. El siguiente código sirve para obtenerlo:

```
qqnorm(residuos)
qqline(residuos)
```



Dado que los puntos están bastante alineados, la normalidad también parece aceptable.

Ejercicios

1. Ajusta el modelo que explica la cantidad de grasas en función del peso.

- Calcula y representa gráficamente la recta de regresión, junto con la correspondiente nube de puntos.
- ¿Cuánto vale el coeficiente de correlación al cuadrado en este caso?
- ¿Cuánto valen los estimadores de todos los parámetros del modelo?
- Contrasta la hipótesis de que la pendiente de la recta es cero a nivel 0.05.
- Calula un intervalo de confianza para la pendiente de la recta de nivel 90%.
- Calcula y representa los intervalos de confianza al 95% de la cantidad de grasas media para los individuos entre 30 y 90 kg.
- Lleva a cabo el diagnóstico del modelo.

2. Supongamos que la variable regresora toma los valores $x = 1, 2, \ldots, 10$. El siguiente código de R genera una muestra que sigue el modelo de regresión lineal (cuando $\beta_0 = 0$, $\beta_1 = 1$ y $\sigma = 0.3$), extrae el valor de la pendiente estimada $\hat{\beta}_1$ y resume los principales resultados.

- Anota el valor del estimador de la pendiente que has obtenido y su error típico. ¿Has obtenido una buena estimación?
- Repite el procedimiento anterior 1000 veces y haz un estudio descriptivo de las 1000 pendientes estimadas resultantes. Estudia si se corresponden los resultados de la simulación con las propiedades teóricas del

estimador de la pendiente. El error típico obtenido en el problema anterior, ¿refleja adecuadamente la variabilidad del estimador observada en la simulación?

- Repite la simulación, pero generando los datos de manera que las variables de error ϵ_i proceden de una distribución t de Student con 5 grados de libertad en lugar de una distribución normal. Describe las propiedades de la distribución del estimador que se deducen de la simulación.

```r
# Variable regresora (dieño fijo) y parámetros
x = seq(1,10)
beta0 <- 0
beta1 <- 1
sigma <- 0.3

# Genera la variable respuesta
y <- beta0 + beta1*x + rnorm(length(x), sd=sigma)

# Ajusta el modelo
reg <- lm(y~x)

# Extrae el valor de la pendiente estimada
coefficients(reg)[2]

# Resume el ajuste
summary(reg)
```