Universidad de Cantabria
Curso 2011–2012

Práctica 2: Regresión lineal

R también permite trabajar conjuntamente con más de una variable. En particular esta sección se centra en el caso de 2 variables con el fin de estudiar el comportamiento de una de ellas a partir de otra.

1. Análisis descriptivo de variables bidimensionales

En primer lugar construimos la tabla de doble entrada con dos variables que seleccionamos del fichero *Pulso.rda* como son *Actividad* y *Fumar*. Esta tabla se construye a partir del menú Estadísticos ⇒ Tablas de Contingencia ⇒ Tabla de doble entrada. Se abre una ventana de dialogo en la que podemos seleccionar dichas variables en filas o columnas. Entre otras opciones, esta ventana también nos permite elegir si queremos que en la tabla aparezcan las frecuencas absolutas (*Sin porcentajes*) o bien las frecuencias relativas en porcentajes (*Porcentajes totales*). Del resultado de la tabla podemos ver que la mayoría de las personas que realizan una actividad *media* o *alta* son no fumadores.

Con el fin de analizar la relación entre variables también resulta interesante la realización de distintos tipos de gráficos como es el diagrama de dispersión analizado en clase. En este caso vamos a analizar la relación que existe entre las variables *Peso* y *Altura*. En el menú Gráficos encontramos la opción Diagrama de dispersión. Por defecto aparece marcada la opción `línea suavizada`, que ofrece una regresión a los puntos bajo el criterio de mínimos cuadrados. La orden que genera el gráfico de la figura 1 es la siguiente:

```
> scatterplot(Peso~Altura, reg.line=lm, smooth=FALSE, labels=FALSE, boxplots='xy',

span=0.5, data=Pulsaciones)
```

A la vista de la figura, se deduce que existe una relación lineal directa entre ambas variables. Además se observa que la figura muestra los diagramas de cajas correspondientes a las dos variables. Para confirmar la existencia de una alta correlación entre las variables vamos a calcular el coeficiente de correlación de Pearson seleccionando en el menú Estadísticos ⇒ Resúmenes ⇒ Test de correlación. En el cuadro de diálogo se deben elegir las variables correspondientes (*Peso* y *Altura* en este ejemplo).

```
> cor.test(Altura, Peso, alternative=''two.sided'', method=''pearson'')
```

Como ya esperábamos el coeficiente de correlación es relativamente alto y positivo (0.7848664) lo que

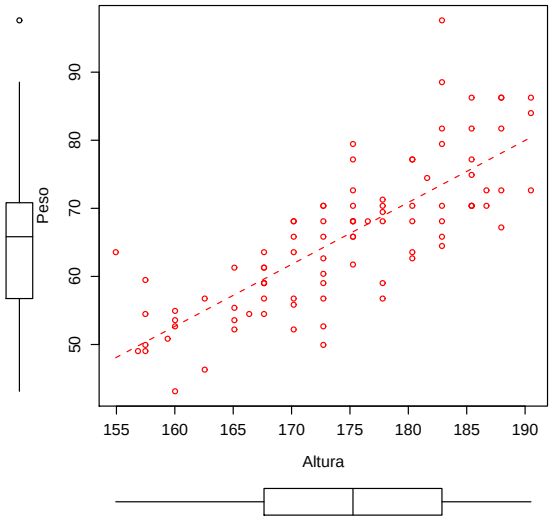

Figura 1: Diagrama de dispersión de las variables *Peso* y *Altura* del fichero *Pulso.rda*.

confirma la relación directa entre las variables.

Practica tú mismo

---

1) Hacer un diagrama de dispersión de la *Altura* frente al *Peso* separando por sexos y responder a las siguientes preguntas:

1. ¿Aumenta en la misma proporción el peso con la altura en los hombres y en las mujeres?
2. ¿Cuál de los dos grupos presenta un valor de la correlación mayor?[a] Indicar el valor en cada caso.

[a]En este caso se necesita filtrar los datos por *Sexo* antes de calcular el coeficiente de correlación.

Practica tú mismo

---

2) Representar un diagrama de dispersión de *Pulso1* frente al *Pulso2*.

1. Añadir un título apropiado.
2. ¿Qué sugiere el gráfico sobre la relación entre estas dos variables?
3. Añadir *Correr* como variable de agrupamiento.
4. ¿Difiere mucho la relación *Pulso1-Pulso2* entre correr o no?

2. Regresión lineal

En el menú Estadísticos ⇒ Ajuste de modelos ⇒ Regresión lineal, podemos obtener los parámetros de la recta de ajuste ($Y = a + bX$) del *Peso* en función de la *Altura* que se muestra en la figura 1. En la ventana de diálogo tendremos que seleccionar la variable explicada o dependiente y la variable explicativa o independiente que en este ejemplo corresponden al *Peso* y la *Altura*, respectivamente.

```
> RegModel.1 <- lm(Peso~Altura, data=Pulsaciones)
> summary(RegModel.1)
```

A partir de la tabla de los *Coeficientes* podemos obtener los parámetros de la ecuación de la recta de ajuste que resulta ser: ($Y = -92{,}95231 + 0{,}9101X$). Si únicamente se quieren obtener los coeficientes de la recta bastaría con escribir:

```
> RegModel.1 <- lm(Peso~Altura, data=Pulsaciones)
> coef(RegModel.1)
```

El dato del coeficiente de determinación *R-squared*, dado por $R^2 = \frac{S^2(\hat{y})}{S^2(y)}$, indica la proporción de la variación total de la variable dependiente Y que es explicada por el modelo de regresión o que se debe a la variación en la variable independiente X. Su valor está confinado en el intervalo [0,1] y da información de la bondad del ajuste. Cuanto más próximo esté el valor a 1 mejor se ajusta el modelo a los datos. Normalmente este valor se da en tanto por ciento.

Practica tú mismo

---

3) Calcular la ecuaciones de la recta de ajuste del *Peso* y la *Altura* agrupados por *Sexo* y contestar a las siguientes preguntas:

1. Escribir la ecuación resultante para las mujeres.
2. ¿Cuál es la pendiente de la recta obtenida en el caso de los hombres?
3. ¿Para que *Sexo* el modelo se ajusta mejor? ¿Por qué?

Practica tú mismo

4) Un ciclista se desplaza en línea recta con un movimiento uniforme para el cual según las leyes de la mecánica su posición x en un instante t vendrá dada por la ecuación $x = x_0 + vt$ donde x_0 es la posición inicial y v la velocidad.
Se han tomado los siguientes valores de su posición x en metros y el tiempo t en segundos:

x (metros)	14	26.2	37.7	51	61.8	76	84.2
t (segundos)	2	4	6	8	10	12	14

A partir de estos datos estimar:

1. el coeficiente de correlación
2. los valores de la posición inicial y la velocidad del ciclista por medio de una regresión lineal.
3. el espacio que recorrido por el ciclista transcurridos 9 segundos.
4. el error estandar de la estimación y la fracción de varianza explicada por el modelo.

2.1. Transformaciones de modelos no lineales

En algunos casos es posible, mediante la aplicación de determinadas transformaciones, expresar la relación entre dos variables mediante un modelo lineal. Un ejemplo, son los modelos exponencial ($Y = ae^{bX}$) y potencial ($Y = aX^b$) que se linealizan mediante la aplicación de logaritmos permitiendo así su resolución mediante el método de mínimos cuadrados que acabamos de ver. Como ejemplo, programaremos la resolución del siguiente ejercicio:

En la siguiente tabla se recogen datos de un experimento en el que se ha analizado el tiempo de secado de un tipo de pintura dependiendo de la cantidad añadida de un compuesto que reduce el tiempo de secado.

Tiempo de secado (horas)	12	10.5	10	9	8.5	8	7.5	7.2	7
Aditivo (gramos)	1	2	3	4	5	6	7	8	9

Sabiendo que la relación entre la variable tiempo (T) y la cantidad de aditivo (C) se expresa mediante un modelo del tipo $T = \beta C^{\alpha}$ siendo β y α dos constantes, responder a las siguientes cuestiones:

1. dibuje un diagrama de dispersión para verificar que es razonable suponer que la relación es de tipo potencial.
2. los valores de β y α.
3. el tiempo de secado de la pintura cuando se mezclan 9.5 gramos de aditivo.

En primer lugar introducimos los datos:

```
t<-c(12,10.5,10, 9, 8.5, 8, 7.5, 7.2, 7)
c<-c(1,2,3,4,5,6,7,8,9)
```

Identificamos la variable dependiente (x) y la independiente (y) y calculamos los logaritmos de cada una para transformar el modelo potencial a lineal ($log(T) = log(\beta) + b \cdot \log(C)$).

```
x<-log(c)
y<-log(t)
```

Estimamos los parámetros del modelo de regresión lineal para las variables calculadas x e y.

```
fit <- lm(y~x)
a <- fit$coeff[1];a
beta<-exp(a);beta
b <- fit$coeff[2];b
```

Construimos las funciones de ambos modelos para hacer la representacion

```
fx <- function(x) exp(a) * x^b
fxrecta <- function(x) a+b*x
```

Representamos los valores de los datos originales y transformados y ajustamos el modelo correspondiente a cada caso a partir de los parámetros calculados.

```
pdf("pintura.pdf", width=7, height=3)
  par(mfrow=c(1,2), mar=c(4,4,1,1))
  plot(c,t, xlab="c", ylab="t", type="n")
    curve(fx, col="blue",lwd=4,add=TRUE)
    points(c,t, pch=19, col="red")
  plot(x,y, xlab="log(c)", ylab="log(t)", type="n")
    curve(fxrecta, col="blue",lwd=4,add=TRUE)
    points(x,y , pch=19, col="red")
dev.off()
```

Practica tú mismo

---

5) En un experimento se han recogido los siguientes datos de presión de un gas para varios valores de su volumen, mantiendo constante la temperatura.

V (mm^3)	54.3	61.8	72.4	88.7	118.6	194.0
P (mm de Hg)	61.2	49.5	37.6	28.4	19.2	10.1

Los principios de la termodinámica relacionan los valores de volumen (V) y presión (P) de un proceso adiabático mediante la expresión $PV^\gamma = C$ siendo γ y C dos constantes. Calcular:

1. Los valores de γ y C expresando la presión en función del volumen.
2. Estimar el valor de la presión para un volumen de 90 mm^3
3. Representar los valores de los datos originales y transformados y las líneas de ajuste obtenidas en cada caso.