Laboratorio 2: Análisis de datos bivariantes

Tablas de Contingencia, Diagramas de Barras

1. Se introducen los siguientes datos cualitativos correspondientes a $n = 10$ individuos en los que se estudia el `sexo` y `color de ojos`.

sexo	ojo.col
mujer	negro
hombre	negro
hombre	azul
hombre	verde
hombre	verde
mujer	verde
mujer	negro
hombre	verde
mujer	azul
mujer	azul

Para introducir los datos, hay que:

Datos → Nuevo conjunto de datos...

En la ventana de diálogo **Introducir el nombre del conjunto de datos** se puede poner, por ejemplo, `Ojosexo`. Pulsar OK.

Aparece una hoja de datos donde se puede cambiar los nombres (por defecto) de las dos primeras columnas `var1` y `var2` por los nombres que aparecen en la tabla anterior,

haciendo click en las casillas de `var1` y `var2`. No olvidar especificar los tipos de datos como `character` cuando se metan los nombres. Una vez hecho todo esto, cerrar la hoja de datos. Por defecto `Ojosexo`, será el conjunto de datos *activo*.

Tareas:

a) Obtener la tabla de contingencia (con las frecuencias absolutas y relativas) de los datos:

En la *ventana de instrucciones*, escribe

```
table(Ojosexo)
table(Ojosexo)/10
```

¿Cuántas personas son mujeres con ojos negros?

¿Cuál es la proporción de personas que son hombres con ojos azules?

b) Obtener las distribuciones relativas de frecuencias marginales:

En la *ventana de instrucciones*, escribe

```
totPercents(table(Ojosexo))/100
```

(dividimos entre 100 de modo que los valores quedan entre 0 y 1 y no en porcentajes).

¿Qué proporción de las personas son hombres? ¿Qué proporción de las personas tienen los ojos verdes?

c) Crear un diagrama de barras:

En la *ventana de instrucciones*, cargar la librería `lattice` y realizar un diagrama de barras más *cool*:

```
library(lattice)
barchart(Ojosexo)
```

o incluso mejor

```
barchart(Ojosexo, groups=Ojosexo$sexo)
```

¿Hay más mujeres u hombres con los ojos azules? ¿Cuál es el color de ojos más común entre los hombres?

Diagramas de puntos, covarianza y correlación

2. Cargar los datos de *Anscombe*:

Datos → Conjunto de datos en paquetes → Leer conjunto de datos desde paquete adjunto...

En la ventana de diálogo, en el panel derecho, elegir el paquete `datasets` (haciendo click dos veces). Buscar a lo largo del panel derecho hasta encontrar `anscombe` y seleccionarlo (haciendo click dos veces). Pulsar OK.

El fichero `anscombe` aparecerá en azul entre el menú principal y la *ventana de instrucciones*, por defecto será el fichero de datos activo.

Observar el conjunto de datos haciendo click en `Visualizar conjunto de datos` (se debería ver un conjunto de datos con 11 filas y 8 columnas).

a) Hacer un diagrama de puntos de `y1` frente a `x1`:

Graficas → Diagrama de dispersion...

En la ventana de diálogo, elegir `x1` como la `variable x` en el panel izquierdo, e `y1` como la `variable y` en el panel derecho.

Desmarcar los campos `línea suavizada` y `Cajas de dispersion marginales` y marcar `Linea de minimos cuadrados` (se pueden marcar también otras gráficas). Aumentando el tamaño de los ejes y de los puntos (ej. de 1 a 1.5) se hará más claro el gráfico.

Se pueden guardar los gráficos mediante:

Graficos → Guardar grafico en fichero → como PDF/Postscript/EPS...

En la ventana de diálogo marcar el formato que se quiera. Por ejemplo, con un ancho y alto de 10×10 pulgadas en formato `Postscript`, sale un gráfico de calidad que ocupa media página en Microsoft Word.

b) Obtener la matriz de gráfico de dispersión de todo el conjunto de datos:

Graficas → Matriz de diagramas de dispersion...

En la ventana de diálogo, en el campo `Select variable` seleccionar todas las variables. Desmarcar `Lineas suavizadas` y marcar `Lineas de minimos cuadrados`. Pulsar OK.

¿Son los gráficos simétricos alrededor de la diagonal principal?

c) Obtener diagramas de cajas (un diagrama de cajas para cada variable), en una figura. Usar diferentes colores para `(x1,y1)`, `(x2,y2)`, etc. (1=black, 2=red, 3=green, 4=blue):

En la *ventana de instrucciones*, escribir

```
boxplot(anscombe,col=c(1,2,3,4,1,2,3,4))
```

Identifica los diagramas de puntos de: $(x1, y1)$, $(x2, y2)$, $(x3, y3)$, $(x4, y4)$ y describe los patrones de comportamiento.

NOTA: para consultar sobre qué representa un diagrama de cajas, ver:

`http://es.wikipedia.org/wiki/Diagrama_de_caja`

d) Obtener la matriz de correlaciones del conjunto completo de datos:

Estadisticos → Resumenes → Matriz de Correlaciones...

En la ventana de diálogo seleccionar todas las variables.

Calcular las correlaciones siguientes: $Cor(x_1, y_1)$, $Cor(x_2, y_2)$, $Cor(x_3, y_3)$, $Cor(x_4, y_4)$

Basándose en los correspondientes diagramas de dispersión de la parte b), explica cuál de estas correlaciones deberían usarse y cuáles no.

e) Obtener la covarianza entre $x1$ e $y1$.

En la *ventana de instrucciones*, escribir

```
cov(anscombe$x1,anscombe$y1)
```

Regresión Lineal

3. Se usan los mismos datos que en el problema 2.

   a) Ajustar un modelo de regresión lineal de `y1` respecto de `x1` e interpretarlo. Indicar cuál es el porcentaje de variabilidad de `y1` explicada por su dependencia lineal sobre `x1`. Es decir, se calcula R^2 mediante:

   Estadisticos → Ajuste de modelos → Regresion Lineal...

   En la ventana de diálogo, especificar `y1` como la `Variable explicada` en el panel de la izquierda y `x1` como la `Variable explicativa` en el panel derecho. El modelo generado (objeto) se graba como `RegModel.1` También se puede cambiar el nombre del modelo en `Introducir un nombre para el modelo:`.

   El nombre `RegModel.1` aparece en azul y, por defecto, se convierte en el *modelo activo*.

   Se observa el resumen del ajuste en la *ventana de resultados*.

   El coeficiente de determinación, R^2, aparece como `Multiple R-Squared`, el `Residual standard error` se calcula dividiendo entre $n-2$ (se pierden 2 grados de libertad al estimar la pendiente y la ordenada en el origen).

      i) Obtener el gráfico de residuos, el gráfico de residuos respecto a las prediccio-

nes. Observar la existencia o no de algún patrón en el gráfico de residuos, ¿qué implicación se deduce respecto a la validez del modelo?

Modelos → Graficas → Graficas basicas de diagnostico

Con esta secuencia se obtiene una figura con 4 sub-figuras, donde el gráfico de residuos está en la parte superior izquierda. El resto de gráficos no se estudian en el presente curso de Estadística I.

Para obtener una figura con sólo un gráfico de residuos, se añaden primero los residuos al conjunto de datos:

Modelos → Añadir las estadisticas de las observaciones a los datos...

En el cuadro de diálogo, marcar `Valores ajustados` (es decir las predicciones) y los `Residuos`, y desmarcar el resto, salvo que se quiera obtener más información.

El conjunto de datos activos aumenta adjuntando dos columnas nuevas que se denominan `fitted.RegModel.1` y `residuals.RegModel.1` (si se cambia el nombre del modelo en la parte b, entonces los nombres se ajustan de modo coherente).

Para hacer el gráfico de residuos de `Residuos` frente a `Ajustados` se tienen que seguir los mismos pasos que en la parte a) donde `x1` se reemplaza por `Fitted` e `y1` por `Residuals`.

b) Repetir el apartado a) para `x1` siendo la variable dependiente e `y1`, siendo la variable independiente. Comparar los parámetros estimados en el apartado b) con los

obtenidos en el apartado a) ¿son los mismos o son diferentes? ¿por qué?)

Programación en R

Para facilitar el trabajo y hacer reproducible cualquier análisis, se puede programar todos los procedimientos anteriores mediante un *script* directamente en `R` (sin la ayuda de `RCommander`).

Introducir en el *prompt* principal del programa (`>`):

```
Ojosexo <- edit(as.data.frame(NULL))

table(Ojosexo)

table(Ojosexo)/10


totPercents(table(Ojosexo))/100


library(lattice)

X11()

barchart(Ojosexo)

X11()

barchart(Ojosexo, groups=Ojosexo$sexo)


data(anscombe, package="datasets")

X11()
```

```
scatterplot(y1~x1, reg.line=lm, smooth=FALSE, labels=FALSE, boxplots=FALSE,
  span=0.5, cex.axis=1.5, cex.lab=1.5, data=anscombe)
X11()
scatterplot.matrix(~x1+x2+x3+x4+y1+y2+y3+y4, reg.line=lm, smooth=FALSE,
  span=0.5, diagonal = 'density', data=anscombe)

X11()
boxplot(anscombe,col=c(1,2,3,4,1,2,3,4))
cor(anscombe[,c("x1","x2","x3","x4","y1","y2","y3","y4")],
  use="complete.obs")
cov(anscombe$x1,anscombe$y1)



RegModel.1 <- lm(y1~x1, data=anscombe)
summary(RegModel.1)
oldpar <- par(oma=c(0,0,3,0), mfrow=c(2,2))
plot(RegModel.1)
par(oldpar)

anscombe$fitted.RegModel.1 <- fitted(RegModel.1)
anscombe$residuals.RegModel.1 <- residuals(RegModel.1)
```

```
RegModel.2 <- lm(y1~fitted.RegModel.1, data=anscombe)

summary(RegModel.2)

oldpar <- par(oma=c(0,0,3,0), mfrow=c(2,2))

plot(RegModel.2)

par(oldpar)
```