Minería de Datos, 2008-2009
Exploración con R Commander

R

1. Ojear el contenido de *http://www.r-project.org/*.
2. Instalar R.
   (a) En la sección Download pinchar en *CRAN* y seleccionar un CRAN mirror.
   (b) En el mirror elegido, en la sección *Download and Install R*, seleccionar el sistema operativo en el que se quiere instalar R.
   (c) Si es Windows,
      1. Pulsar *Windows (95 and later)* y pulsar en *base* (*http://cran.au.r-project.org/bin/windows/*)
      2. Pulsar en *README.R.-2.8.0* y ojear las instrucciones.
      3. Bajar el fichero R-2.8.0-win32.exe y ejecutar para instalar R.
      4. Para añadir paquetes a la instalación, ejecutar R y utilizar el menú Paquetes-Instalar paquete(s) siguiendo las instrucciones.
         Otro método es bajarse todos los paquetes utilizando el enlace *contrib* (debajo de *base* en *http://cran.au.r-project.org/bin/windows/*) y utilizar el menú Paquetes-Instalar paquete(s) a partir de archivos zip locales...
3. Explorar en el menú Ayuda el contenido de las funciones de ayuda de R. Documentación y manuales de interés sobre R pueden encontrarse en: *http://cran.r-project.org/doc/contrib/*.. Buenos manuales con distinto grado de complejidad son:
   (a) *http://www.math.csi.cuny.edu/Statistics/R/simpleR/printable/simpleR.pdf*
   (b) *http://cran.r-project.org/doc/contrib/usingR.pdf*
   (c) *http://cran.es.r-project.org/doc/contrib/Kuhnert+Venables-R_Course_Notes.zip*

R Commander

Instalación y visión preliminar

1. Ver el contenido de la página *http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/*. Bajarse un *introductory manual* en *http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/Getting-Started-with-the-Rcmdr.pdf*. y un artículo de Journal of Statistical Software en http://www.jstatsoft.org/v14/i09/paper.
2. Ejecutar R y mediante *Paquetes-Instalar paquete(s)* instalar el paquete *Rcmdr*.
3. Mediante *Paquetes-Cargar paquete...* cargar el paquete Rcmdr. Esto equivale al comando *library(Rcmdr)* en consola. Probablmente pida la instalación de más paquetes necesarios: aceptar la petición.

4. Entender las funciones de menú Ayuda.

5. Explorar las funciones de R Commander. Las funciones del menú *Fichero* son autoexplicativas (los ficheros de instrucciones son ficheros de programas –tienen la extensión .R) y lo mismo sucede con el menú *Editar*.

Carga de datos y análisis exploratorio

1. Abrir el fichero *ordenadores.xls* mediante *Datos-importar datos-From Excel, Access or Dbase data set...* dando como nombre a los datos por ejemplo *Ordenadores* y eligiendo la hoja en la que se encuentran los datos.

   Comprobar que los datos se han cargado correctamente pulsando en el botón *Editar conjunto de datos* o *Visualizar conjunto de datos*.

2. Mediante *Estadísticos-Resúmenes-Datos activos* obtener un resumen descriptivo de los datos.

3. Repasar el resto de funciones del menú *Estadísticos-Resúmenes* entendiendo sus acciones sobre los datos (consultar o repasar en alguna fuente los detalles de los procedimientos estadísticos utilizados en cada función y, alternativamente, consultar la ayuda en el diálogo de cada procedimiento). Por ejemplo, (suponemos que estamos ante una muestra aleatoria simple):

   (a) Mediante *Estadísticos-Resúmenes numéricos* determinar qué valoración mínima recibe el 25% de los ordenadores mejor valorados.

   (b) Mediante *Estadísticos-Correlation test...* comparar los coeficientes de correlación de Pearson, de Spearman y de Kendall entre las variables $tiempo2$ y $valoracion$.

   Dadas dos variables ordinales definidas sobre un mismo conjunto de individuos,

   - El coeficiente de correlación por rangos de Spearman entre dos variables continuas medidas en un conjunto de individuos es el coeficiente de correlación de Pearson aplicado a los órdenes que los individuos presentan según las dos variables.
   - El coeficiente de correlación τ de Kendall entre dichas variables se define como:

$$\tau = \frac{S_t}{\binom{n}{2}} \in [-1, 1]$$

   donde $S_t = \sum_{j=1}^{n(n-1)/2} \delta_i$ siendo $\delta_i = +1$ si el par i-ésimo de individuos presenta el mismo orden según las dos variables, y $\delta_i = -1$ en caso contrario. Recuérdese que el número posible de pares que se pueden formar con n individuos es $\binom{n}{2} = n(n-1)/2$.

4. Mediante el menú *Estadísticos-Tablas de contingencia* realizar una prueba de independencia Chi-cuadrado entre las variables *origen* y *tipo*, seleccionando también *Components of the chi square statistic* e *Imprimir las frecuencias esperadas*. Las hipótesis nula y alternativa son:

$$\begin{cases} H_0 : origen \text{ y } tipo \text{ son independientes} \\ H_1 : \text{lo contrario} \end{cases}$$

   y suponiendo cierta H_0 el estadístico χ^2 visto en los apuntes de repaso de Descriptiva, sigue una $\chi^2_{(n-1)(m-1)}$ donde n y m son el número de categorías de las dos variables.

   Evaluando el *p-value* decidir sobre la veracidad de H_0. Si la muestra fuese representativa del mercado español, ¿podría decirse que los fabricantes españoles se han especializado en algún tipo de ordenador?.

5. Evaluar la normalidad de la variable $tiempo1$ mediante *Estadísicos-Resúmenes-Test de normalidad de Shapiro-Wilks.*

6. Mediante *Estadísticos-Medias-prueba t para una muestra*, evaluar si la media poblacional de la variable $tiempo1$ puede considerarse igual que su mediana en la muestra con una significación del 5%. Responded a la misma pregunta observando el intervalo de confianza obtenido. Recordar que el estadístico utilizado bajo H_0 es:

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

   ¿Por qué podemos suponer que el estadístico sigue esta distribución?.

7. Mediante *Estadísticos-Medias-Anova de un factor*, determinar si el *origen* del ordenador afecta significativamente a la *valoración*, suponiendo que la distribución de la variable $valoracion$ puede considerarse normal.

   - ¿Qué significación presenta la respuesta a tal pregunta?. Por tanto, ¿puede considerarse que existe asociación entre las variables *tipo* y *valoración*?.
   - Relacionar el ratio de correlación η^2 calculado en la práctica anterior con las sumas de cuadrados de la tabla del análisis de la varianza.

8. Si la muestra de ordenadores fuese representativa del mercado español de ordenadores personales, ¿podríamos afirmar que la mitad del mercado corresponde a ordenadores portátiles?. Utilizar *Estadísticos-Proporciones-prueba de proporciones para una muestra*.

9. Mediante *Estadísticos-Varianzas* determinar si el comportamiento de los ordenadores portátiles según $tiempo1$ es más errático que el de los ordenadores de sobremesa. Realizar los tres contrastes consultando sus detalles en el libro electrónico *Engineering Statistics Handbook*, bajado la Semana 2. Por ejemplo, para la prueba de Levene el estadístico bajo H_0 (varianzas iguales) es:

$$\frac{S_B^2/m}{S_I^2/n - m} \sim F_{m,n-m}$$

   donde S_B^2 la suma de cuadrados intergrupos y S_I^2 la suma de cuadrados intragrupos ambas referidas a la variable D con valores $d_{ij} = |x_{ij} - \bar{x}_j|$, $i = 1, ..., n_j$; $j = 1, ..., m$, donde m es el número de grupos y n_j el número de individuos del grupo j.

10. Por si acaso la variable *valoración* no siguiese una distribución normal,

    - Responder a la misma pregunta que en 7, utilizando el test no paramétrico de Kruskal-Wallis mediante *Estadísticos-Pruebas no paramétricas-Test de Kruskal-Wallis*. Consultar p. ej. en el libro *Engineering Statistics Handbook* o en Wikipedia el contenido de este test.
    - Evaluar si la variable *tipo* posee alguna influencia en la valoración de los ordenadores utilizando el test de Wilcoxon mediante *Estadísticos-Pruebas no paramétricas-Test de Wilcoxon para dos muestras*. El test de sumas de rangos de Wilcoxon también recibe el nombre de test U de Mann-Whitney-Wilcoxon. Consultar los detalles de este test en el libro *Engineering Statistics Handbook* bajo la denominación *Mann-Whitney U-Test* o por ejemplo en http://en.wikipedia.org/wiki/Mann-Whitney_U.

Gráficos multipropósito

1. Realizar una visión general de las funciones del menú *Gráficas* obteniendo para algunas variables del fichero diversos gráficos tratados en el repaso de estadística descriptiva: Gráfico secuencial, histograma, gráfico de tallo y hojas (*stem & leaf plot*), diagrama de caja (*box plot*), gráfico de barras y gráfico de sectores.

2. Realizar los siguientes análisis gráficos:

   - Evaluar gráficamente la normalidad de las variables ram, $tiempo1$, $tiempo2$ y $valoración$, comparando sus cuantiles con los de la distribución gausiana o normal. Entender el gráfico y concluir sobre la posible normalidad de cada variable. Comparar el gráfico con los resultados del test de normalidad de Shapiro-Wilks.
   - Explorar todas las posibles relaciones entre las variables del fichero mediante una matriz de diagramas de dispersión con rectas por mínimos cuadrados pero sin líneas suavizadas e incluyendo histogramas en la diagonal.
   - Comparar mediante un diagrama de caja la variable $tiempo1$ según el origen del ordenador. ¿Qué puede decirse sobre los valores centrales y la heterogeneidad de dicha variable según dicho origen?. Identificar algún valor atípico.
   - Evaluar gráficamente la relación entre la variable ram y $valoración$ para cada tipo de ordenador, mediante un diagrama de dispersión que incluya las rectas de regresión para los dos tipos (portátiles y sobremesa).
   - Estudiar la evolución conjunta de $valoración$ vs $tiempo1$, $tiempo2$ y $clock$ mediante una gráfica lineal, de la siguiente forma:
     – Para ello, ordenar antes el fichero tecleando en la consola de R (o en la Ventana de instrucciones de R Commander pulsando después en el botón Ejecutar):
     $>$ `indicessorted = order(Ordenadores$valoracion)`
     $>$ `Ordenadores = Ordenadores[indicessorted,]`
     (La primera instrucción guarda los índices de los ordenadores ordenados según valoración de manera descendente, y la segunda los dispone y guarda según dicho orden).
     – Mediante *Gráficas-Gráfica lineal*, elegir valoración como variable x y las demás variables como variables y (pulsando Ctrl al elegirlas).
   - Mediante *Gráficas-XY* realizar un gráfico de dispersión de $valoracion$ (explicada) frente a ram y $clock$ (explicativas) según el $origen$ del ordenador agrupando según $tipo$ de ordenador.

3. Revisar la utilidad de las siguientes funciones del menú *Gráficas.*

Modelo Lineal y Lineal Generalizado

1. Mediante *Estadísticos-Ajuste de modelos-regresión lineal...*

   (a) Realizar una regresión lineal de la variable *valoración* sobre las variables $clock, ram, tiempo1$ y $tiempo2$.¿Qué variables resultan significativas y a qué nivel?. ¿Puede considerarse que el modelo explica suficientemente la valoración que los usuarios realizan sobre sus ordenadores?. ¿Qué porcentaje de información contenida en $valoración$ explica linealmente el modelo?.

   (b) Realizar sucesivamente regresiones sin las variables que resultaron significativas en las regresiones anteriores y analizar los resultados.

(c) Por cada unidad de tiempo menos que un ordenador tarde en la variable $tiempo1$ ¿cuánto aumenta su valoración por término medio?.

2. Mediante *Modelos-Selecciona el modelo activo...* seleccionamos el modelo lineal estimado en primer lugar (con las cuatro variables regresoras) como el modelo activo para el resto de las operaciones de este punto. Comprobar la selección mediante *Modelos-Resumir el modelo.*

   (a) Obtener intervalos de confianza para cada coeficiente de la regresión mediante *Modelos-Intervalos de confianza...*

   (b) Mediante *Modelos-Test de Hipótesis-Hipótesis lineal...*, determinarsi el modelo $valoración = a_1 \cdot clock + a_2 \cdot ram + a_3 \cdot tiempo1 + a_4 \cdot tiempo2 + a_0$ verifica:

      1. Que $a_4 = -0.5$.
      2. Que $a_0 = 10$, $a_1 = 0$, $a_2 = 0$, $a_2 = -0.1$, $a_3 = -0.5$. Para ello, añadir filas con unos en la diagonal y especificar el valor correspondiente en el *Right-hand side*.

   (c) Mediante *Modelos-Gráficas-Gráficas básicas de diagnóstico...* obtener gráficos básicos de diagnóstico del modelo lineal estimado. Estudiar los gráficos de la primera fila evaluando la bondad del modelo lineal y la hipótesis de normalidad respectivamente. En la segunda fila, investigar la heterogeneidad de varianzas y la influencia de las observaciones individuales en el modelo (ver el significado de la distancia de Cook p. ej. en Wikipedia).

   (d) Mediante *Modelos-Gráficas-Componentes + resíduos* y *Modelos-Gráficas-Gráfica de los efectos* obtener gráficos de la inuencia de cada variable explicativa.Véase por ejemplo http://en.wikipedia.org/wiki/Partial_residual_plot.

3. Vamos a comparar un modelo lineal y un modelo logístico para resolver el problema de clasificación planteado en la práctica anterior: determinar el $tipo$ de ordenador a partir de su valor en $tiempo1$.

   (a) Si no está ya creada, crear una nueva variable $tipon$ con valor 1 si portátil y valor 0 si sobremesa, mediante *Datos-modificar variables del conjunto de datos activo-Calcular una nueva variable*. Como exprsión a calcular, utilizar: *as.numeric(tipo == "portátil")* que convertirá el valor TRUE de la prueba lógica $(tipo ==$"portatil"$)$ en el valor 1 y el valor FALSE en el valor 0.

   (b) Modelo lineal:

      1. Mediante *Estadisticos-Ajuste de modelos-Modelo lineal*, estimar un modelo lineal utiliando $tipon$ como variable respuesta y $tiempo1$ como regresora.
      2. Mediante *Modelos-añadir las estadísticas de las observaciones a los datos* insertar en el fichero Ordenadores los valores ajustados (asegurarse de que el modelo activo es el modelo recién ajustado). Consultar estas predicciones en el fichero interpretándolas a la luz del documento Clasificación con regresión visto en la práctica anterior.
      3. Mediante *Datos-modificar variables del conjunto de datos activo-Calcular una nueva variable* crear en el fichero una nueva variable denominada $yhatLM$ mediante la expresión *as.numeric(fit.XXXX >= 0.5)* indicando que la variable $yhatLM$ valdrá 1 si la variable con los valores ajustados es ≥ 0.5.

   (c) Modelo logístico:

      1. Mediante *Estadísticos-Ajuste de modelos-Modelo lineal generalizado*, estimar un modelo logístico utilizando las mismas variables $tipon$ y $tiempo1$ (consultar los detalles y la estimación de este modelo en cualquier referencia de la bibliografía o Wikipedia).

2. Añadir al fichero Ordenadores los valores ajustados asegurándose de que el modelo activo es el modelo recien estimado. Interpretar dichas predicciones y compararlas someramente con las del modelo lineal.
3. Al igual que en el caso anterior, crear en el fichero Ordenadores una nueva variable denominada $yhatGLM$ para estimar el tipo de ordenador a partir de los valores ajustados por el modelo logístico.

(d) Cálculo de la tasa de error de clasificación para cada uno de los modelos. En la consola de R (o en la ventana de instrucciones de R Comander y pulsando en Ejecutar), calcular la tasa de error de clasificación mediante, por ejemplo:

```
> ErrorRateLM = sum(yhatLM != tipon) / dim(Ordenadores)[1]
> ErrorRateGLM = sum(yhatGLM != tipon) / dim(Ordenadores)[1]
```

Cerrar R Commander

1. Al cerrar R Commander es posible guardar el fichero de comandos (instrucciones) de R generados al ejecutar los procesos anteriores desde el interface de usuario de R Commander, así como el fichero de salida (contenido de la ventana de resultados) y el fichero de datos cuando éstos han resultado modificados.