Introducción a R.
Mathieu Kessler
Departamento de Matemática Aplicada y Estadística
Universidad Politécnica de Cartagena

Manejo básico de la línea de comandos en R

Se utilizará como apoyo para esta práctica las transparencias "Introducción a R", descargables en el aula virtual.

1. Primer paso: descargar el fichero datosprimerasesion.Rdata

Este fichero se puede descargar desde el aula virtual, en el Area de prácticas, en el directorio "Datos necesarios para las prácticas". Se recomienda descargarlo en la carpeta "Mis Documentos" del ordenador

2. Segundo paso: ejecutar R

3. Tercer paso: cargar el área de trabajo llamado datosprimerasesion.Rdata

Para ello tenemos dos opciones:

1. En el menú Archivo, escogemos la entrada "Cargar área de trabajo", recorremos la estructura de archivos para hasta llegar al fichero que descargamos en el primer paso.

2. En la línea de comandos, introducimos:
`load`("C:/Documents and Settings/Mathieu/Mis documentos/datosprimerasesion.Rdata")
donde se cambia el camino "C:/Documents and Settings/Mathieu/Mis documentos/" por el adecuado a vuestro ordenador.

4. Unos pocos comandos útiles de R

4.1. La instrucción `ls()`

La instrucción `ls()` permite obtener una lista de los objetos definidos en la sesión actual.

```
> ls()
```

```
[1] "anscombe"           "cemento"            "escombrerashoras09"
[4] "geyser"             "newcomb"            "prueba_acceso"
```

Observamos como tenemos 6 objetos definidos en nuestra sesión. Son todos conjuntos de datos asociados a distintos ejemplos que iremos viendo.
Nota: *Si como resultado de la instrucción ls(), obtenemos más objetos, es porque R, al ejecutarse ha rescatado los objetos que se definieron en la última sesión que se llevó a cabo. Para empezar con un entorno "limpio", podemos hacer lo siguiente antes del tercer paso descrito arriba*:

```
> rm(list = ls())
```

4.2. La instrucción `names()`

La instrucción `names`) aplicada a un conjunto de datos (dataframe) concreto, nos permite conocer los nombres de sus columnas. Escogemos por ejemplo el conjunto de datos llamado escombrerashoras09:

```
> names(escombrerashoras09)
```

```
 [1] "NO2"   "SO2"   "DD"    "VV"    "TMP"   "HR"    "PRB"   "RS"    "fecha"
[10] "año"   "mes"   "día"   "hora"
```

Se trata de un conjunto de datos asociados a los niveles de calidad del aire durante parte del año 2009, en el Valle de Escombreras, un valle industrial cerca de Cartagena. Los datos se obtuvieron en la página de la Consejería de Agricultura y Agua de la Comunidad Autónoma de la Región de Murcia (http://www.carm.es/cmaot/calidadaire/portal/). Aparte de los contaminantes y los instantes de medición, aparecen mediciones de:

- DD: dirección del viento (grados)
- VV: Velocidad del viento (m/s)
- TMP: temperatura (°C)
- HR: humedad relativa (% H.R.)
- PRB: Presión atmósférica (mb)
- RS: Radiación solar (W/m^2)"

4.3. La instrucción `dim`

La instrucción `dim` aplicada a un conjunto de datos concretos nos permite conocer el número de filas y columnas que lo constituyen.

```
> dim(escombrerashoras09)
```

```
[1] 6071   13
```

Deducimos que el conjunto tiene 6071 instantes de medición...

4.4. La instrucción head

La instrucción head aplicada a un conjunto de datos concretos nos permite visualizar las prímeras lineas del conjunto:

```
> head(escombrerashoras09)
```

```
   NO2 SO2    DD  VV  TMP   HR    PRB  RS               fecha  año mes día hora
1 10.6 3.2 151.0 1.5 11.7 83.8 1019.0 0.1 2009-01-01 00:00:00 2009   1   1    0
2 16.7 4.1  52.2 0.9 11.7 81.7 1019.0 0.1 2009-01-01 01:00:00 2009   1   1    1
3 14.6 3.8 154.2 0.4 11.6 80.0 1018.7 0.1 2009-01-01 02:00:00 2009   1   1    2
4 13.2 4.6 107.5 0.5 11.3 80.5 1018.2 0.1 2009-01-01 03:00:00 2009   1   1    3
5 11.7 5.9  62.3 0.6 10.8 82.2 1018.0 0.2 2009-01-01 04:00:00 2009   1   1    4
6 14.7 5.9  66.7 0.5 10.9 81.5 1018.0 0.2 2009-01-01 05:00:00 2009   1   1    5
```

5. Ejercicios

1. Para cada uno de los conjuntos de datos (ver descripción de cada uno de los conjuntos en el apéndice) contestar a la siguiente pregunta:

   a) ¿Qué individuos describe el conjunto? ¿Cuántos individuos son? ¿Cuántas variables contiene el conjunto?

2. Obtener:

   a) Para el conjunto `escombrerashoras09`, los valores del individuo número 5503. ¿A qué fecha corresponde la medición?

   b) Para el conjunto `cemento` todos los valores de la variable resistencia.

   c) Para el conjunto `prueba_acceso`, la media del expediente del alumno número 84934. Guardar este valor en un nuevo objeto llamado u.

3. Añadir al conjunto de datos `prueba_acceso` una columna que se llame `MEDIA_PARTES12` que sea igual a la media de las variables `MEDIA_PARTE1` y `MEDIA_PARTE2`.

4. Construir un nuevo conjunto de datos llamado `misdatos`, que contenga dos variables `x` e `y` que tomen los siguientes valores:

`x`	1.2	4.3	3.2	-3
`y`	0.042	-0.98	0	1..02

5. Guardar el área de trabajo en un nuevo fichero llamado "miprimerasesionenR.Rdata"

Apendice: descripción de los conjuntos de datos usados

1. `anscombe`: Anscombe (1973), "Graphs in statistical analysis", *American Statistician*, **27**, pp 17-21, construyó cuatro conjuntos de datos artificiales para ilustrar los peligros de llevar a cabo ajustes sin visualizar los datos primero.

2. `cemento`: Se estudia la relación entre la composición de un cemento tipo Portland y el calor desprendido durante la fase de fraguado. Los datos se pueden encontrar en el fichero hald.txt. La variable Y es la cantidad de calor desprendido en calorías por gramos de cemento, mientras que las variables X1, X2 X3 y X4 representan el contenido en porcentaje de cuatro productos A, B, C y D.

3. `escombrerashoras09`: Se trata de un conjunto de datos asociados a los niveles de calidad del aire durante parte del año 2009, en el Valle de Escombreras, un valle industrial cerca de Cartagena. Los datos se obtuvieron en la página de la Consejería de Agricultura y Agua de la Comunidad Autónoma de la Región de Murcia (http://www.carm.es/cmaot/calidadaire/portal/). Aparte de los contaminantes y los instantes de medición, aparecen mediciones de:

   - DD: dirección del viento (grados)
   - VV: Velocidad del viento (m/s)
   - TMP: temperatura (°C)
   - HR: humedad relativa (% H.R.)
   - PRB: Presión atmósférica (mb)
   - RS: Radiación solar (W/m^2)"

4. `geyser`: Un geyser es un nacimiento de agua hirviente que de vez en cuando se vuelve inestable y expulsa agua y vapor. El geyser .Old Faithful.en el parque de Yellowstone en Wyoming es probablemente el más famoso del mundo. Los visitantes del parque se acercan al emplazamiento del geyser intentando no tener que esperar demasiado para verlo estallar. Los servicios del Parque colocan un cartel donde se anuncia la próxima erupción. Es por lo tanto de interés estudiar los intervalos

de tiempo entre dos erupciones conjuntamente con la duración de cada erupción. En este conjunto de datos están los datos correspondientes a la duración de 222 erupciones asi como el intervalo de tiempo hasta la siguiente erupción, durante los meses de agosto 1978 y agosto 1979. Las unidades de medición son mn.

5. `newcomb`: Newcomb fue el primero en conseguir ¡en 1882! una estimación bastante precisa de la velocidad de la luz. Las mediciones recogidas a continuación corresponden a los tiempos codificados que tardó un rayo de luz en recorrer el camino de ida y vuelta desde el laboratorio de Simon Newcomb situado en el Río Potomac hasta un espejo situado en la base del "Washington Monument", en total una distancia de 7400m. Para obtener los tiempos en nano segundos ($10^{-9}s$) no codificados, hay que añadir 24800 a cada dato.[1]

6. `prueba_acceso`: En este conjunto están recogidos los resultados de las pruebas de acceso del Distrito Único de la Región de Murcia, obtenidos con la autorización de la Comisión de las PAU de la Región de Murcia. Se ha eliminado cualquier campo que permita relacionar los datos con un alumno o un centro concreto. Además se ha limitado a los alumnos que han superado las pruebas de acceso: para ellos, tienen que ser aptos (haber obtenido más de 4 en el promedio de las dos partes de la prueba de acceso), y que su nota final (40 % promedio dos partes, 60 % nota expediente) sea superior a 5.

   Las variables tienen nombres autoexplicativos, excepto a lo mejor:

   - CONV: tiene dos valores posibles: E: "Extraordinaria", y O: "Ordinaria".
   - TIPO_CL: ¿se trata de un alumno COU ("C"), o LOGSE("L")?

[1] Fuente: Moore, David S. and McCabe, George P. (1989). Introduction to the Practice of Statistics, W. H. Freeman and Company: New York, NY, pp 3-16.