Data Frames

Code

Vamos a descargar de la pagina de UCI Machine Learning repository la base de datos de Adults.

Lo primero que hay que verificar es nuestro directorio actual de trabajo y modificarlo si necesario.

Hide

```
getwd() # directorio actual
```

```
[1] "C:/Users/mbtec/Documents/GitHub/PropedeuticoDataScience2017/CuadernosR/2_realdat
a"
```

Con el comando `setwd` pondemos cambiar el directorio de trabajo. Por ejemplo

Hide

```
# setwd("C:/Users/mbtec/Documents/GitHub/PropedeuticoDataScience2017/CuadernosR/Cuader
no2")
```

En RStudio la maneta facil tambien es dar click en los 3 puntos `...` que aparecen en la ventana de Files, navegar al destino, y luego dar click en `More` y seleccionar la opcion `Set As Working Directory`. Alternativamente (mas facil) en el menu principal en `Session`. Otra opcion es trabajar siempre con proyectos que "fijan" el working directory. El working directory lo pueden cambiar cuantas veces quieran.

Bajar los datos

Hide

```
download.file("https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.d
ata", "adultsdata.csv")
```

```
trying URL 'https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.dat
a'
Content type 'text/plain; charset=UTF-8' length 3974305 bytes (3.8 MB)
downloaded 3.8 MB
```

Alternativamente se puede descargar manualmente. Ahora hay que leerlo en R.

Hide

```
adults <- read.csv("adultsdata.csv", header = FALSE)
head(adults) # imprime las primeras lineas
```

V	V2	V3	V4	V	V6	V7
<int>	<fctr>	<int>	<fctr>	<int>	<fctr>	<fctr>
1	39 State-gov	77516	Bachelors	13	Never-married	Adm-clerical
2	50 Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial
3	38 Private	215646	HS-grad	9	Divorced	Handlers-cleaners
4	53 Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners
5	28 Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty
6	37 Private	284582	Masters	14	Married-civ-spouse	Exec-managerial

6 rows | 1-9 of 15 columns

Podemos accesar a las variables por numero de columna o por nombre. En este caso como usamos `header = FALSE` automaticamente R eligio los nombres V1, V2, etc.

Manualmente podemos elegir nombres

Hide

```
names(adults) <-  c("age", "workclass", "fnlwgt", "education", "education_num", "marit
al_status", "occupation", "relationship", "race", "sex", "capital_gain", "capital_los
s", "hours_per_week", "native_country","uji")
head(adults)
```

a..	workclass	fnlwgt	education	education_num	marital_status	occu
<int>	<fctr>	<int>	<fctr>	<int>	<fctr>	<fctr>
1	39 State-gov	77516	Bachelors	13	Never-married	Adm-
2	50 Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec
3	38 Private	215646	HS-grad	9	Divorced	Hand
4	53 Private	234721	11th	7	Married-civ-spouse	Hand
5	28 Private	338409	Bachelors	13	Married-civ-spouse	Prof-
6	37 Private	284582	Masters	14	Married-civ-spouse	Exec

6 rows | 1-8 of 15 columns

Para accesar los datos pueden usar `$` o `[[]]` como en listas.

Hide

```
table(adults$native_country)
```

```
                         ?                    Cambodia
                       583                          19
                    Canada                       China
                       121                          75
                  Columbia                        Cuba
                        59                          95
         Dominican-Republic                     Ecuador
                        70                          28
                El-Salvador                     England
                       106                          90
                    France                      Germany
                        29                         137
                    Greece                    Guatemala
                        29                          64
                     Haiti           Holand-Netherlands
                        44                           1
                  Honduras                        Hong
                        13                          20
                   Hungary                       India
                        13                         100
                      Iran                      Ireland
                        43                          24
                     Italy                      Jamaica
                        73                          81
                     Japan                        Laos
                        62                          18
                    Mexico                     Nicaragua
                       643                          34
Outlying-US(Guam-USVI-etc)                        Peru
                        14                          31
               Philippines                      Poland
                       198                          60
                  Portugal                   Puerto-Rico
                        37                         114
                  Scotland                       South
                        12                          80
                    Taiwan                    Thailand
                        51                          18
            Trinadad&Tobago                United-States
                        19                        29170
                   Vietnam                   Yugoslavia
                        67                          16
```

En general pueden saber mucho de un data frame con la funcion summary (en teoria, pero nunca le he encontrado practica…)

Hide

```
summary(adults)
```

```
     age                     workclass         fnlwgt
Min.   :17.00    Private         :22696   Min.   :  12285
1st Qu.:28.00    Self-emp-not-inc: 2541   1st Qu.: 117827
Median :37.00    Local-gov       : 2093   Median : 178356
Mean   :38.58    ?               : 1836   Mean   : 189778
3rd Qu.:48.00    State-gov       : 1298   3rd Qu.: 237051
Max.   :90.00    Self-emp-inc    : 1116   Max.   :1484705
                 (Other)         :  981
        education     education_num
 HS-grad     :10501   Min.   : 1.00
 Some-college: 7291   1st Qu.: 9.00
 Bachelors   : 5355   Median :10.00
 Masters     : 1723   Mean   :10.08
 Assoc-voc   : 1382   3rd Qu.:12.00
 11th        : 1175   Max.   :16.00
(Other)      : 5134
               marital_status             occupation
 Divorced             : 4443    Prof-specialty :4140
 Married-AF-spouse    :   23    Craft-repair   :4099
 Married-civ-spouse   :14976    Exec-managerial:4066
 Married-spouse-absent:  418    Adm-clerical   :3770
 Never-married        :10683    Sales          :3650
 Separated            : 1025    Other-service  :3295
 Widowed              :  993   (Other)         :9541
         relationship                    race
 Husband       :13193    Amer-Indian-Eskimo:  311
 Not-in-family : 8305    Asian-Pac-Islander: 1039
 Other-relative:  981    Black             : 3124
 Own-child     : 5068    Other             :  271
 Unmarried     : 3446    White             :27816
 Wife          : 1568

     sex          capital_gain    capital_loss    hours_per_week
 Female:10771   Min.   :    0   Min.   :   0.0   Min.   : 1.00
 Male  :21790   1st Qu.:    0   1st Qu.:   0.0   1st Qu.:40.00
                Median :    0   Median :   0.0   Median :40.00
                Mean   : 1078   Mean   :  87.3   Mean   :40.44
                3rd Qu.:    0   3rd Qu.:   0.0   3rd Qu.:45.00
                Max.   :99999   Max.   :4356.0   Max.   :99.00

       native_country      uji
 United-States:29170    <=50K:24720
 Mexico       :  643    >50K : 7841
 ?            :  583
 Philippines  :  198
 Germany      :  137
 Canada       :  121
(Other)       : 1709
```

Analisis de covarianzas

Vamos a elegir una submatriz de datos con solo tres variables para ejemplificar conceptos

	age <int>	education_num <int>	hours_per_week <int>	uji_numeric <dbl>
1	39	13	40	1
2	50	13	13	1
3	38	9	40	1
4	53	7	40	1
5	28	13	40	1
6	37	14	40	1

6 rows

Vamos a ver la matriz de covarianzas:

Hide

```
cov(adults2)
```

```
                    age education_num hours_per_week uji_numeric
age            186.061400     1.2818493      11.580130   1.3649972
education_num    1.281849     6.6188899       4.705338   0.3686853
hours_per_week  11.580130     4.7053379     152.458995   1.2126508
uji_numeric      1.364997     0.3686853       1.212651   0.1828259
```

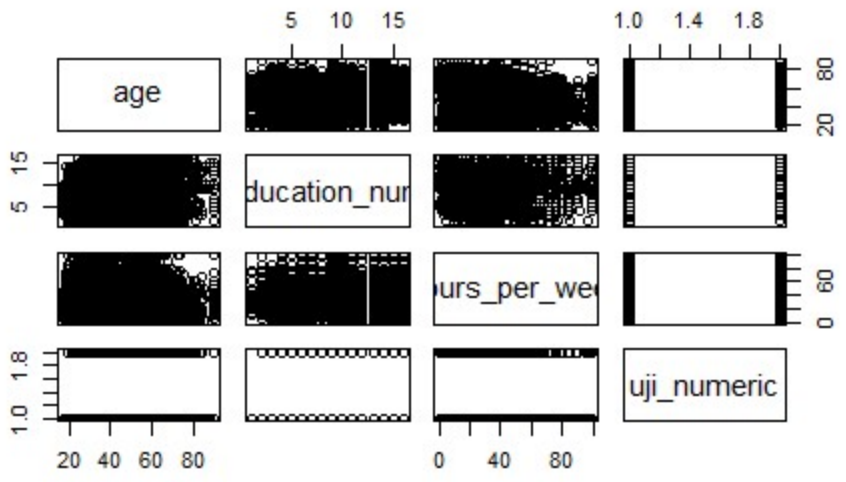
La matriz de covarianzas no es muy util…. Es mas util ver correlaciones

Hide

```
cor(adults2)
```

```
                    age education_num hours_per_week uji_numeric
age            1.00000000    0.03652719     0.06875571   0.2340371
education_num  0.03652719    1.00000000     0.14812273   0.3351540
hours_per_week 0.06875571    0.14812273     1.00000000   0.2296891
uji_numeric    0.23403710    0.33515395     0.22968907   1.0000000
```

Otra manera de visualizarlo



Podrian predecir el ingreso con la edad, educacion y horas trabajas?

Para esto se usan las regresiones lineales (mas detalles manana y jueves)

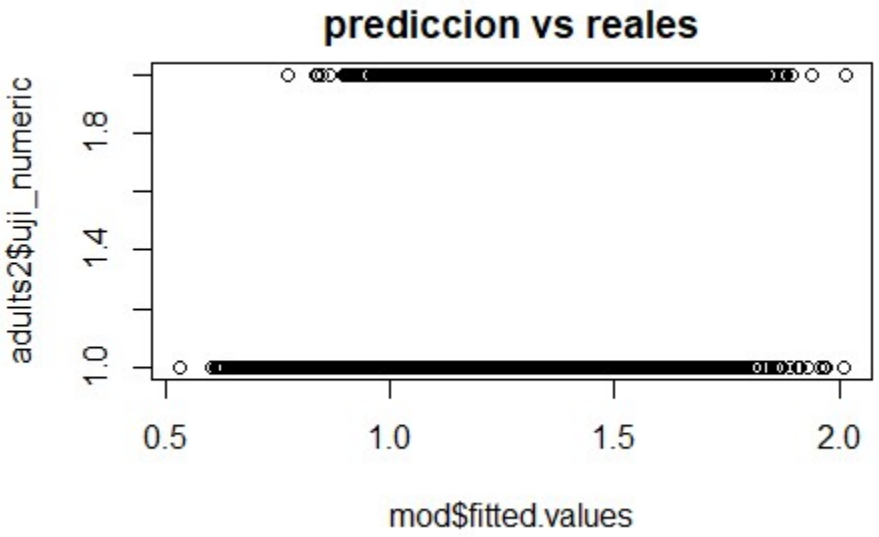
Hide

```
summary(mod)
```

```
Call:
lm(formula = uji_numeric ~ ., data = adults2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0123 -0.2703 -0.1310  0.2109  1.2272

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.2403725  0.0117049   20.54   <2e-16 ***
age            0.0066230  0.0001568   42.24   <2e-16 ***
education_num  0.0502245  0.0008386   59.89   <2e-16 ***
hours_per_week 0.0059008  0.0001750   33.71   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3849 on 32557 degrees of freedom
Multiple R-squared:  0.1899,    Adjusted R-squared:  0.1898
F-statistic:  2543 on 3 and 32557 DF,  p-value: < 2.2e-16
```

Hide

```
plot(
  mod$fitted.values,
  adults2$uji_numeric,
  main = "prediccion vs reales"
)
```



A mano

Hide

```
X <- as.matrix(adults2[ ,1:3])
Y <- adults2$uji_numeric
X$colones <- 1
```

```
Coercing LHS to a list
```

Hide

```
beta <- solve(t(X)%*%X, t(X)%*%Y)
```

```
Error in t(X) %*% X : requires numeric/complex matrix/vector arguments
```