R for Regression
Dolores Romero Morales

The aim of this workshop is to perform Regression with R using a brief guide to R for Regression and three datasets (tupelo.txt, blood.txt, housing.txt). All files can be found in the course directory. The tupelo and the blood datasets were used during the lecture. More information on the housing dataset can be found in https://archive.ics.uci.edu/ml/machine-learning-databases/housing/.

Once you are ready with the guide, use the datasets to perform the following steps using R:

Task 1
Tupelo dataset: Linear Regression

Step 1. Read the tupelo.txt file into a data frame.
Answer:
mytupelo <- read.table(file.choose(),header=TRUE)

Note:
With file.choose() a new window will pop up and you will need to find the file.

Step 2. Regress the Cost variable against Capacity and Year.
Answer:
model1 <- lm(mytupelo$Cost ~ .,mytupelo)

Note:
You can use attach() to shorten the names of the variables of a data frame. For instance, by calling attach(mytupelo), you can use Capacity instead of mytupelo$Capacity. Use detach(mytupelo) when you want to end the attachment. Be careful with attach() in case you are working with several data frames.

Step 3. Give a summary of the model. Check the fit of the model and the significance of the explanatory variables.
Answer:
summary(model1)

Note1: This is the output we get with summary()

```
Call:
lm(formula = mytupelo$Cost ~ ., data = mytupelo)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4966 -1.4692 -0.2149  1.2054  3.2877

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.5488     3.1186   6.589 0.000307 ***
Year          0.6187     0.3521   1.757 0.122312
Capacity     -3.2121     1.1368  -2.826 0.025567 *
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.079 on 7 degrees of freedom
Multiple R-squared: 0.7903, Adjusted R-squared: 0.7304
F-statistic: 13.19 on 2 and 7 DF, p-value: 0.004223

Note2:
Fit of a linear regression model is given by the Adjusted R Squared, while the significance of explanatory variables by their p-value. We can see that Year has a p-value of 0.122312. If we choose a significance of 0.05, this variable is not significant to the model.

Note3:
At this point, it would be good to detach the data frame, i.e.,
detach(mytupelo)

Task 2
Blood dataset: Logistic Regression

Step 1. Read the blood.txt file into a data frame.
Answer:
```
myblood <- read.table(file.choose(),header=TRUE)
```

Step 2. Regress the Blood variable against Smoke and Alcohol.
Answer:
```
model2 <- glm(Blood ~ ., family='binomial', myblood)
```

Step 3. Give a summary of the model, and check the fit of the model and the significance of the explanatory variables.
Answer:
```
summary(model2)
logLik(model2)
```

Note1:
The output of the summary() function is

```
Call:
glm(formula = Blood ~ ., family = "binomial", data = myblood)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5205  -0.6426  -0.5371   0.4434   1.9358

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.39440    0.37295  -3.739 0.000185 ***
Smoke        2.26416    0.51287   4.415 1.01e-05 ***
Alcohol     -0.07816    0.08500  -0.920 0.357792
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 137.186 on 99 degrees of freedom
Residual deviance: 77.999 on 97 degrees of freedom
AIC: 83.999

Number of Fisher Scoring iterations: 7

Note2:
The significance of explanatory variables is again measured by their p-value. We can see that Alcohol has a p-value of 0.357792. If we choose a significance of 0.05, this variable is not significant to the model.

Note3:
In logistic regression, when used for probability estimation, has not got a natural way to measure fit. Both, Cox-Snell and Nagelkerke R Square, are a try to simulate what R Square does in Linear Regression.
An alternative way to measure fit is the log likelihood. This is always a negative number. When comparing two models in terms of fit, we would like to choose the one with the smallest absolute value of the log likelihood.

Task 3
Housing dataset: Multiple Regression

Step 1. Read the housing.txt file into a data frame.
Answer:
myhousing <- read.table(file.choose(),header=TRUE)

Note:
Although not required, it is advisable to call the attach function and the dimension one.

Step 2. Regress the last variable against the rest.
Answer:
modelwith13 <-lm(myhousing$MEDV~.,myhousing)

Step 3. Give a summary of the model, and check the fit of the model and the significance of all explanatory variables.
Answer:
summary(modelwith13)

Note1: See the output below.

Call:
lm(formula = MEDV ~ ., data = myhousing)

Residuals:
    Min   1Q Median   3Q   Max
-15.595 -2.730 -0.518   1.777 26.199

Coefficients:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
CRIM        -1.080e-01  3.286e-02  -3.287 0.001087 **
ZN           4.642e-02  1.373e-02   3.382 0.000778 ***
INDUS        2.056e-02  6.150e-02   0.334 0.738288
CHAS         2.687e+00  8.616e-01   3.118 0.001925 **
NOX         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
RM           3.810e+00  4.179e-01   9.116  < 2e-16 ***
AGE          6.922e-04  1.321e-02   0.052 0.958229
DIS         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
RAD          3.060e-01  6.635e-02   4.613 5.07e-06 ***
TAX         -1.233e-02  3.760e-03  -3.280 0.001112 **
PTRATIO     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
B            9.312e-03  2.686e-03   3.467 0.000573 ***
LSTAT       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,  Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

Note2: The question is what set of variables we should build the model on. The step() function performs model selection in a stepwise fashion based on the Akaike Information Criterion (AIC). This criterion is a tradeoff between the number of variables used and the fit of the model obtained with those variables. We can see below that we would end up with 11 out the 13 explanatory variables.

```
step(modelwith13)

Start:  AIC=1589.64
myhousing$MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE +
    DIS + RAD + TAX + PTRATIO + B + LSTAT

          Df Sum of Sq   RSS    AIC
- AGE      1      0.06 11079 1587.7
- INDUS    1      2.52 11081 1587.8
<none>                 11079 1589.6
- CHAS     1    218.97 11298 1597.5
- TAX      1    242.26 11321 1598.6
- CRIM     1    243.22 11322 1598.6
- ZN       1    257.49 11336 1599.3
- B        1    270.63 11349 1599.8
- RAD      1    479.15 11558 1609.1
- NOX      1    487.16 11566 1609.4
- PTRATIO  1   1194.23 12273 1639.4
- DIS      1   1232.41 12311 1641.0
- RM       1   1871.32 12950 1666.6
- LSTAT    1   2410.84 13490 1687.3

Step:  AIC=1587.65
```

```
myhousing$MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + DIS +
    RAD + TAX + PTRATIO + B + LSTAT

          Df Sum of Sq   RSS    AIC
- INDUS    1      2.52 11081 1585.8
<none>                 11079 1587.7
- CHAS     1    219.91 11299 1595.6
- TAX      1    242.24 11321 1596.6
- CRIM     1    243.20 11322 1596.6
- ZN       1    260.32 11339 1597.4
- B        1    272.26 11351 1597.9
- RAD      1    481.09 11560 1607.2
- NOX      1    520.87 11600 1608.9
- PTRATIO  1   1200.23 12279 1637.7
- DIS      1   1352.26 12431 1643.9
- RM       1   1959.55 13038 1668.0
- LSTAT    1   2718.88 13798 1696.7

Step:  AIC=1585.76
myhousing$MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD + TAX +
    PTRATIO + B + LSTAT

          Df Sum of Sq   RSS    AIC
<none>                 11081 1585.8
- CHAS     1    227.21 11309 1594.0
- CRIM     1    245.37 11327 1594.8
- ZN       1    257.82 11339 1595.4
- B        1    270.82 11352 1596.0
- TAX      1    273.62 11355 1596.1
- RAD      1    500.92 11582 1606.1
- NOX      1    541.91 11623 1607.9
- PTRATIO  1   1206.45 12288 1636.0
- DIS      1   1448.94 12530 1645.9
- RM       1   1963.66 13045 1666.3
- LSTAT    1   2723.48 13805 1695.0

Call:
lm(formula = myhousing$MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS +
    RAD + TAX + PTRATIO + B + LSTAT, data = myhousing)

Coefficients:
(Intercept)         CRIM           ZN         CHAS          NOX           RM          DIS          RAD
TAX      PTRATIO
  36.341145    -0.108413     0.045845     2.718716   -17.376023     3.801579    -1.492711
0.299608    -0.011778    -0.946525
          B        LSTAT
   0.009291    -0.522553
```