Hypothesis verification for Lineal Model

Ramón Daniel Regueiro Espiño

The aim of this document is to create a R code that automatises hypothesis verification for linear model regression up to 8 explanatory variables. To illustrate the usage of the code, `leafburn` data are considered.

```
rm(list=ls())
library(faraway)
datos1<-leafburn
n<-dim(datos1)[2]
datos<-matrix(,nrow=dim(datos1)[1],)
for (i in 1:n){
  if (is.numeric(datos1[,i])) datos<-cbind(datos,datos1[,i])
}
```

For our case, we'll only be interested in develop the model:

$$burntime = \beta_0 + \beta_1\, nitrogen + \beta_2\, chlorine + \beta_3\, potassium\,,$$

although other models can be build.

```
datos<-datos[,-1]
datos<-datos1[,c("burntime","nitrogen","chlorine","potassium")]
datos<-datos1
datos
```

```
##    nitrogen chlorine potassium burntime
## 1      3.05     1.45      5.67      2.2
## 2      4.22     1.35      4.86      1.3
## 3      3.34     0.26      4.19      2.4
## 4      3.77     0.23      4.42      4.8
## 5      3.52     1.10      3.17      1.5
## 6      3.54     0.76      2.76      1.0
## 7      3.74     1.59      3.81      1.2
## 8      3.78     0.39      3.23      1.3
## 9      2.92     0.39      5.44     33.9
## 10     3.10     0.64      6.16      5.9
## 11     2.86     0.82      5.48     14.8
## 12     2.78     0.64      4.62     10.2
## 13     2.22     0.85      4.49      7.8
## 14     2.67     0.90      5.59     25.1
## 15     3.12     0.92      5.86     11.2
## 16     3.03     0.97      6.60     14.1
## 17     2.45     0.18      4.51     30.9
## 18     4.12     0.62      5.31      3.2
## 19     4.61     0.51      5.16      1.5
## 20     3.94     0.45      4.45      2.2
## 21     4.12     1.79      6.17      2.3
## 22     2.93     0.25      3.38      7.8
```

```
## 23     2.66     0.31      3.51      8.1
## 24     3.17     0.20      3.08      8.3
## 25     2.79     0.24      3.98     22.4
## 26     2.61     0.20      3.64     21.4
## 27     3.74     2.27      6.50      1.7
## 28     3.13     1.48      4.28      1.8
## 29     3.49     0.25      4.71      5.4
## 30     2.94     2.22      4.58      1.7
```

```r
n<-dim(datos)[1]
p<-dim(datos)[2]
if (p==1) print("Only one variable, no model")
if (p==2) modelo<-lm(datos[,1]~datos[,2])
if (p==3) modelo<-lm(datos[,1]~datos[,2]+datos[,3])
if (p==4) modelo<-lm(datos[,1]~datos[,2]+datos[,3]+datos[,4])
if (p==5) modelo<-lm(datos[,1]~datos[,2]+datos[,3]+datos[,4]+datos[,5])
if (p==6) modelo<-lm(datos[,1]~datos[,2]+datos[,3]+datos[,4]+datos[,5]+datos[,6])
if (p==7) modelo<-lm(datos[,1]~datos[,2]+datos[,3]+datos[,4]+datos[,5]+datos[,6]+datos[,7])
if (p==8) modelo<-lm(datos[,1]~datos[,2]+datos[,3]+datos[,4]+datos[,5]+datos[,6]+datos[,7]+datos[,8])
if (p>8) print("More than 8 explanatory variables")
summary(modelo)
```

```
##
## Call:
## lm(formula = datos[, 1] ~ datos[, 2] + datos[, 3] + datos[, 4])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.06174 -0.27178  0.01693  0.27907  0.86446
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.13487    0.38103   8.227 1.04e-08 ***
## datos[, 2]  -0.20049    0.17843  -1.124 0.271442
## datos[, 3]   0.15173    0.09319   1.628 0.115553
## datos[, 4]  -0.04666    0.01066  -4.376 0.000174 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4566 on 26 degrees of freedom
## Multiple R-squared:  0.4494, Adjusted R-squared:  0.3859
## F-statistic: 7.074 on 3 and 26 DF,  p-value: 0.001249
```

Collinearity induces misinterpretation of explanatory variables in our target variable which leads to poor estimations. We study collinearity and variance inflation factor.

```r
cor.datos<-cor(datos[,-1])
for (i in 1:(length(cor.datos[1,])-1)){
  for (j in (i+1):length(cor.datos[1,])){
    if (abs(cor.datos[i,j])>0.6) print ("Worrying correlation")
  }
}

vif.modelo<-vif(modelo)
for (i in 1:length(vif.modelo)) {
  if (vif(modelo)[i]>5) print("Worryinf VIF")
```

```
}
```

We can introduce a variable selection method to choose the number of explanatory variables that maximizes the likelihood using the AIC criteria.

```
modelo2<-step(modelo)
```

```
## Start:  AIC=-43.33
## datos[, 1] ~ datos[, 2] + datos[, 3] + datos[, 4]
##
##              Df Sum of Sq    RSS     AIC
## - datos[, 2]  1    0.2632 5.6841 -43.906
## <none>                     5.4208 -43.328
## - datos[, 3]  1    0.5527 5.9735 -42.416
## - datos[, 4]  1    3.9924 9.4132 -28.772
##
## Step:  AIC=-43.91
## datos[, 1] ~ datos[, 3] + datos[, 4]
##
##              Df Sum of Sq    RSS     AIC
## - datos[, 3]  1    0.3135 5.9976 -44.295
## <none>                     5.6841 -43.906
## - datos[, 4]  1    4.0771 9.7612 -29.683
##
## Step:  AIC=-44.3
## datos[, 1] ~ datos[, 4]
##
##              Df Sum of Sq    RSS     AIC
## <none>                     5.9976 -44.295
## - datos[, 4]  1     3.848 9.8455 -31.425
```

```
vif.modelo<-vif(modelo2)
for (i in 1:length(vif.modelo)) {
  if (vif(modelo)[i]>5) print("Worrying VIF in step model")
}
```

We study now linearity, homocedasticity and normality hypothesis including outlier data.

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
r1<-rstandard(modelo2)
r2<-rstudent(modelo2)

if (harvtest(modelo2)$p.value <0.05) print("Linearity with outliers is not verified")
if (bptest(modelo2)$p.value <0.05) print("Homocedasticity is not verified with atypicals")
if (shapiro.test(r2)$p.value <0.05) print("Normality of residuals with outliers is not verified")
```

Now, we study the data without outliers, which are selected using the student test.

```r
leverage<-hat(model.matrix(modelo2))
print("Number of outlier data:")
```

```
## [1] "Number of outlier data:"
```

```r
print(length(which(leverage>2*p/n)))
```

```
## [1] 1
```

```r
maiores<-as.vector(which(abs(r2)>1.96))
datos3<-datos[maiores*(-1),]

if (p==1) print("Only one variable, no model")
if (p==2) modelo3<-lm(datos3[,1]~datos3[,2])
if (p==3) modelo3<-lm(datos3[,1]~datos3[,2]+datos3[,3])
if (p==4) modelo3<-lm(datos3[,1]~datos3[,2]+datos3[,3]+datos3[,4])
if (p==5) modelo3<-lm(datos3[,1]~datos3[,2]+datos3[,3]+datos3[,4]+datos3[,5])
if (p==6) modelo3<-lm(datos3[,1]~datos3[,2]+datos3[,3]+datos3[,4]+datos3[,5]+datos3[,6])
if (p==7) modelo3<-lm(datos3[,1]~datos3[,2]+datos3[,3]+datos3[,4]+datos3[,5]+datos3[,6]+datos3[,7])
if (p==8) modelo3<-lm(datos3[,1]~datos3[,2]+datos3[,3]+datos3[,4]+datos3[,5]+datos3[,6]+datos3[,7]+datos
if (p>9) print("More than 8 explanatory variables")

modelo4<-step(modelo3)
```

```
## Start:  AIC=-50.62
## datos3[, 1] ~ datos3[, 2] + datos3[, 3] + datos3[, 4]
##
##               Df Sum of Sq    RSS     AIC
## - datos3[, 2]  1   0.05524 3.5062 -52.175
## - datos3[, 3]  1   0.22896 3.6799 -50.821
## <none>                     3.4509 -50.620
## - datos3[, 4]  1   2.81976 6.2707 -35.897
##
## Step:  AIC=-52.18
## datos3[, 1] ~ datos3[, 3] + datos3[, 4]
##
##               Df Sum of Sq    RSS     AIC
## - datos3[, 3]  1    0.1745 3.6807 -52.815
## <none>                     3.5062 -52.175
## - datos3[, 4]  1    3.4230 6.9292 -35.101
##
## Step:  AIC=-52.82
## datos3[, 1] ~ datos3[, 4]
##
##               Df Sum of Sq    RSS     AIC
## <none>                     3.6807 -52.815
## - datos3[, 4]  1     3.269 6.9497 -37.018
```

```r
summary(modelo4)
```

```
##
## Call:
## lm(formula = datos3[, 1] ~ datos3[, 4])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -0.6366 -0.2214 -0.0171  0.2171  0.6762
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.591074   0.097724  36.747  < 2e-16 ***
## datos3[, 4] -0.036357   0.007566  -4.805 5.62e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3762 on 26 degrees of freedom
## Multiple R-squared:  0.4704, Adjusted R-squared:   0.45
## F-statistic: 23.09 on 1 and 26 DF,  p-value: 5.617e-05
```

```
r3<-rstudent(modelo4)
pvalor<-harvtest(datos[,1]~modelo2$fitted.values)$p.value

if (pvalor <0.05) print("Linearity with outliers is not verified")
if (bptest(modelo4)$p.value <0.05) print("Homocedasticity is not verified with atypicals")
if (shapiro.test(r3)$p.value <0.05) print("Normality of residuals with outliers is not verified")
```