Multiple linear regression with R: the case with continuous explanatory variables

Nirian Martín

19/5/2020

Importing data: from computer

```
getwd()
```

```
## [1] "C:/Users/nimar/OneDrive - Universidad Complutense de Madrid (UCM)/UCMCurso20192010/Da
tosCategoricos/regresionLogistica"
```

```
setwd("C://Users/nimar/OneDrive - Universidad Complutense de Madrid (UCM)/UCMCurso20192010/Da
tosCategoricos")
fichero1="butler.txt"
misDatos1 <- read.table(file=fichero1,header=TRUE, sep = "\t", dec = ".")
head(misDatos1)  # first few rows
```

```
##   Assignment Miles Deliveries Time
## 1          1   100          4  9.3
## 2          2    50          3  4.8
## 3          3   100          4  8.9
## 4          4   100          2  6.5
## 5          5    50          2  4.2
## 6          6    80          2  6.2
```

Importing data: from internet

```
fichero2 <- "https://raw.githubusercontent.com/NMANMA/classRoomFiles/master/butler.txt"
misDatos2 <- read.delim(file=fichero2,header=TRUE, sep = "\t", dec = ".")
head(misDatos2)  # first few rows
```

```
##   Assignment Miles Deliveries Time
## 1          1   100          4  9.3
## 2          2    50          3  4.8
## 3          3   100          4  8.9
## 4          4   100          2  6.5
## 5          5    50          2  4.2
## 6          6    80          2  6.2
```

Most simple linear regression: intercept model

```
model0 <- lm(Time ~ 1, data=misDatos1)
```

Simple linear regression (1 explanatory variable)

```
model1 <- lm(Time ~ Miles, data=misDatos1)
summary(model1)
```

```
##
## Call:
## lm(formula = Time ~ Miles, data = misDatos1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5565 -0.4913  0.1783  0.7120  1.2435
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.27391    1.40074   0.909  0.38969
## Miles        0.06783    0.01706   3.977  0.00408 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.002 on 8 degrees of freedom
## Multiple R-squared:  0.6641, Adjusted R-squared:  0.6221
## F-statistic: 15.81 on 1 and 8 DF,  p-value: 0.00408
```

```
anova(model1) # test if Miles coefficient=0
```

```
## Analysis of Variance Table
##
## Response: Time
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## Miles      1 15.8713 15.8713  15.815 0.00408 **
## Residuals  8  8.0287  1.0036
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model0,model1)  # test if Miles coefficient=0
```

```
## Analysis of Variance Table
##
## Model 1: Time ~ 1
## Model 2: Time ~ Miles
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1      9 23.9000
## 2      8  8.0287  1    15.871 15.815 0.00408 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple linear regression (2 explanatory variables)

```
model2 <- lm(Time ~ Miles + Deliveries, data=misDatos1)
summary(model2)
```

```
##
## Call:
## lm(formula = Time ~ Miles + Deliveries, data = misDatos1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79875 -0.32477  0.06333  0.29739  0.91333
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.868701   0.951548  -0.913 0.391634
## Miles        0.061135   0.009888   6.182 0.000453 ***
## Deliveries   0.923425   0.221113   4.176 0.004157 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5731 on 7 degrees of freedom
## Multiple R-squared:  0.9038, Adjusted R-squared:  0.8763
## F-statistic: 32.88 on 2 and 7 DF,  p-value: 0.0002762
```

```
anova(model2) # test if Miles coefficient=0 OR if Deliveries cofficient=0 (2 tests)
```

```
## Analysis of Variance Table
##
## Response: Time
##            Df  Sum Sq Mean Sq F value   Pr(>F)
## Miles       1 15.8713 15.8713  48.316 0.000221 ***
## Deliveries  1  5.7293  5.7293  17.441 0.004157 **
## Residuals   7  2.2994  0.3285
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model0,model2)  # test if Miles coefficient=0 AND if Deliveries cofficient=0 (1 test)
```

```
## Analysis of Variance Table
##
## Model 1: Time ~ 1
## Model 2: Time ~ Miles + Deliveries
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1      9 23.9000
## 2      7  2.2994  2    21.601 32.878 0.0002762 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Linear Hypothesis in Multiple linear regression (2 explanatory variables)

```
install.packages("car", repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/nimar/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)
```

```
## package 'car' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\nimar\AppData\Local\Temp\Rtmp27q42k\downloaded_packages
```

```
library(car)
```

```
## Loading required package: carData
```

```
HMat <- matrix(c(0,1,0,0,0,1), nrow=2, ncol=3,byrow=TRUE) # build the hypothesis matrix
HVec <- c(0,0) # build the hypothesis vector
linearHypothesis(model2, HMat, HVec) # test if Miles coefficient=0 AND if Deliveries cofficie
nt=0 (1 test)
```

```
## Linear hypothesis test
##
## Hypothesis:
## Miles = 0
## Deliveries = 0
##
## Model 1: restricted model
## Model 2: Time ~ Miles + Deliveries
##
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1      9 23.9000
## 2      7  2.2994  2    21.601 32.878 0.0002762 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Nested test (incremental F-test) in Multiple linear regression (2 explanatory variables)

```
anova(model1, model2)  # test if Deliveries coefficient = 0
```

```
## Analysis of Variance Table
##
## Model 1: Time ~ Miles
## Model 2: Time ~ Miles + Deliveries
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1      8 8.0287
## 2      7 2.2994  1    5.7293 17.441 0.004157 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
HMat <- matrix(c(0, 0, 1),nro=1,ncol=3,byrow=TRUE) # build the hypothesis matrix
HVec <- c(0) # build the hypothesis vector
linearHypothesis(model2, HMat, HVec) # test if Deliveries coefficient = 0
```

```
## Linear hypothesis test
##
## Hypothesis:
## Deliveries = 0
##
## Model 1: restricted model
## Model 2: Time ~ Miles + Deliveries
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1      8 8.0287
## 2      7 2.2994  1    5.7293 17.441 0.004157 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(model2, "Deliveries=0") # test if Deliveries coefficient = 0
```

```
## Linear hypothesis test
##
## Hypothesis:
## Deliveries = 0
##
## Model 1: restricted model
## Model 2: Time ~ Miles + Deliveries
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1      8 8.0287
## 2      7 2.2994  1    5.7293 17.441 0.004157 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Other nested test (incremental F-test) Multiple linear regression (2 explanatory variables)

```
HMat <- matrix(c(0, 1, -1),nro=1,ncol=3,byrow=TRUE) # build the hypothesis matrix
HVec <- c(0) # build the hypothesis vector
linearHypothesis(model2, HMat, HVec) # test if Miles and Deliveries coefficient are equals
```

```
## Linear hypothesis test
##
## Hypothesis:
## Miles - Deliveries = 0
##
## Model 1: restricted model
## Model 2: Time ~ Miles + Deliveries
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1      8 7.2141
## 2      7 2.2994  1    4.9147 14.961 0.006148 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(model2, "Miles=Deliveries") # test if Miles and Deliveries coefficient are e
quals
```

```
## Linear hypothesis test
##
## Hypothesis:
## Miles - Deliveries = 0
##
## Model 1: restricted model
## Model 2: Time ~ Miles + Deliveries
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1      8 7.2141
## 2      7 2.2994  1    4.9147 14.961 0.006148 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```