Multiple linear regression with R: the case with categorical (dummy) explanatory variables

Nirian Martín

20/5/2020

Importing data: from computer (categorical through FACTORS)

```
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="Johnson2.txt"
misDatos1 <- read.table(file=fichero1,header=TRUE, sep = "\t", dec = ".")
head(misDatos1)  # first few rows
```

```
##   Months.Since.Last.Service Type.of.Repair Repair.Time..hours.
## 1                         2              E                 2.9
## 2                         6              M                 3.0
## 3                         8              E                 4.8
## 4                         3              M                 1.8
## 5                         2              E                 2.9
## 6                         7              E                 4.9
```

Importing data: from internet (categorical through NUMBERS)

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

```
##   Months.Since.Last.Service Type.of.Repair Repair.Time..hours.
## 1                         2              1                 2.9
## 2                         6              0                 3.0
## 3                         8              1                 4.8
## 4                         3              0                 1.8
## 5                         2              1                 2.9
## 6                         7              1                 4.9
```

Most simple linear regression: intercept model

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

Simple linear regression (1 explanatory CONTINUOUS variable, for both data sets)

```
model1 <- lm(Repair.Time..hours. ~ Months.Since.Last.Service, data=misDatos1)
summary(model1)
```

```
##
## Call:
## lm(formula = Repair.Time..hours. ~ Months.Since.Last.Service,
##     data = misDatos1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2597 -0.4772  0.1821  0.4509  1.0362
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.1473     0.6050   3.549  0.00752 **
## Months.Since.Last.Service    0.3041     0.1004   3.029  0.01634 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.781 on 8 degrees of freedom
## Multiple R-squared:  0.5342, Adjusted R-squared:  0.4759
## F-statistic: 9.174 on 1 and 8 DF,  p-value: 0.01634
```

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

```
## Analysis of Variance Table
##
## Response: Repair.Time..hours.
##                            Df Sum Sq Mean Sq F value  Pr(>F)
## Months.Since.Last.Service   1  5.596   5.596  9.1739 0.01634 *
## Residuals                   8  4.880   0.610
## ---
## 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: Repair.Time..hours. ~ 1
## Model 2: Repair.Time..hours. ~ Months.Since.Last.Service
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1      9 10.476
## 2      8  4.880  1     5.596 9.1739 0.01634 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple linear regression (2 explanatory variables, 1 CONTINUOUS, 1 CATEGORICAL): case 2 CATEGORICAL THROUGH NUMBERS (misDatos2)

```
class(misDatos2$Type.of.Repair)
```

```
## [1] "integer"
```

```
table(misDatos2$Type.of.Repair)
```

```
##
## 0 1
## 4 6
```

```
attach(misDatos2)
table(Type.of.Repair)
```

```
## Type.of.Repair
## 0 1
## 4 6
```

```
Type.of.Repair
```

```
##  [1] 1 0 1 0 1 1 0 0 1 1
```

```
model4 <- lm(Repair.Time..hours. ~ Months.Since.Last.Service + Type.of.Repair, data=misDatos
2)
summary(model4)
```

```
##
## Call:
## lm(formula = Repair.Time..hours. ~ Months.Since.Last.Service +
##     Type.of.Repair, data = misDatos2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49412 -0.24690 -0.06842 -0.00960  0.76858
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 0.93050    0.46697   1.993 0.086558 .
## Months.Since.Last.Service  0.38762    0.06257   6.195 0.000447 ***
## Type.of.Repair              1.26269    0.31413   4.020 0.005062 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.459 on 7 degrees of freedom
## Multiple R-squared:  0.8592, Adjusted R-squared:  0.819
## F-statistic: 21.36 on 2 and 7 DF,  p-value: 0.001048
```

```
anova(model4) # test if Months.Since.Last.Service coefficient=0 OR if Type.of.Repair cofficie
nt=0 (2 tests)
```

```
## Analysis of Variance Table
##
## Response: Repair.Time..hours.
##                            Df Sum Sq Mean Sq F value   Pr(>F)
## Months.Since.Last.Service  1 5.5960  5.5960  26.556 0.001319 **
## Type.of.Repair              1 3.4049  3.4049  16.158 0.005062 **
## Residuals                   7 1.4751  0.2107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model0,model4)  # test if Months.Since.Last.Service coefficient=0 AND if Type.of.Repair
cofficient=0 (1 test)
```

```
## Analysis of Variance Table
##
## Model 1: Repair.Time..hours. ~ 1
## Model 2: Repair.Time..hours. ~ Months.Since.Last.Service + Type.of.Repair
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1      9 10.4760
## 2      7  1.4751  2    9.0009 21.357 0.001048 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
detach(misDatos2)
```

Multiple linear regression (2 explanatory variables, 1 CONTINUOUS, 1 CATEGORICAL): case 1 CATEGORICAL THROUGH FACTORS (misDatos1)

```
class(misDatos1$Type.of.Repair)
```

```
## [1] "factor"
```

```
table(misDatos1$Type.of.Repair)
```

```
##
## E M
## 6 4
```

```
attach(misDatos1)
table(Type.of.Repair)
```

```
## Type.of.Repair
## E M
## 6 4
```

```
??contrasts
```

```
## starting httpd help server ... done
```

```
contrasts(Type.of.Repair)  # categorical regressors for Type.of.Repair
```

```
##   M
## E 0
## M 1
```

```
model2 <- lm(Repair.Time..hours. ~ Months.Since.Last.Service + Type.of.Repair, data=misDatos
1)
summary(model2)
```

```
##
## Call:
## lm(formula = Repair.Time..hours. ~ Months.Since.Last.Service +
##     Type.of.Repair, data = misDatos1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49412 -0.24690 -0.06842 -0.00960  0.76858
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                2.19319    0.35576   6.165 0.000461 ***
## Months.Since.Last.Service  0.38762    0.06257   6.195 0.000447 ***
## Type.of.RepairM           -1.26269    0.31413  -4.020 0.005062 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.459 on 7 degrees of freedom
## Multiple R-squared:  0.8592, Adjusted R-squared:  0.819
## F-statistic: 21.36 on 2 and 7 DF,  p-value: 0.001048
```

```
anova(model2) # test if Months.Since.Last.Service coefficient=0 OR if Type.of.Repair cofficie
nt=0 (2 tests)
```

```
## Analysis of Variance Table
##
## Response: Repair.Time..hours.
##                           Df Sum Sq Mean Sq F value   Pr(>F)
## Months.Since.Last.Service  1 5.5960  5.5960  26.556 0.001319 **
## Type.of.Repair             1 3.4049  3.4049  16.158 0.005062 **
## Residuals                  7 1.4751  0.2107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model0,model2)  # test if Months.Since.Last.Service coefficient=0 AND if Type.of.Repair
cofficient=0 (1 test)
```

```
## Analysis of Variance Table
##
## Model 1: Repair.Time..hours. ~ 1
## Model 2: Repair.Time..hours. ~ Months.Since.Last.Service + Type.of.Repair
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1      9 10.4760
## 2      7  1.4751  2    9.0009 21.357 0.001048 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Type.of.Repair2 <- relevel(Type.of.Repair, ref = "M")
table(Type.of.Repair2)
```

```
## Type.of.Repair2
## M E
## 4 6
```

```
contrasts(Type.of.Repair2)  # categorical regressors for Type.of.Repair
```

```
##   E
## M 0
## E 1
```

```
model3 <- lm(Repair.Time..hours. ~ Months.Since.Last.Service + Type.of.Repair2, data=misDatos
1)
summary(model3)
```

```
##
## Call:
## lm(formula = Repair.Time..hours. ~ Months.Since.Last.Service +
##     Type.of.Repair2, data = misDatos1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49412 -0.24690 -0.06842 -0.00960  0.76858
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 0.93050    0.46697   1.993 0.086558 .
## Months.Since.Last.Service  0.38762    0.06257   6.195 0.000447 ***
## Type.of.Repair2E           1.26269    0.31413   4.020 0.005062 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.459 on 7 degrees of freedom
## Multiple R-squared:  0.8592, Adjusted R-squared:  0.819
## F-statistic: 21.36 on 2 and 7 DF,  p-value: 0.001048
```

```
anova(model3) # test if Months.Since.Last.Service coefficient=0 OR if Type.of.Repair cofficie
nt=0 (2 tests)
```

```
## Analysis of Variance Table
##
## Response: Repair.Time..hours.
##                            Df Sum Sq Mean Sq F value   Pr(>F)
## Months.Since.Last.Service  1 5.5960  5.5960  26.556 0.001319 **
## Type.of.Repair2            1 3.4049  3.4049  16.158 0.005062 **
## Residuals                  7 1.4751  0.2107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model0,model3)  # test if Months.Since.Last.Service coefficient=0 AND if Type.of.Repair
cofficient=0 (1 test)
```

```
## Analysis of Variance Table
##
## Model 1: Repair.Time..hours. ~ 1
## Model 2: Repair.Time..hours. ~ Months.Since.Last.Service + Type.of.Repair2
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1      9 10.4760
## 2      7  1.4751  2    9.0009 21.357 0.001048 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
detach(misDatos1)
```