Homework Week 7

John Navarro

November 14, 2016

1 Fitting linear models

1.1 Read the data

```
#Read the data
datapath <- "C:/Users/JohntheGreat/Documents/MSCA/StatisticalAnalysis/Week7/Assignments"
Regression.ANOVA.Data<-as.matrix(read.csv(file=paste(datapath,"DataForRegressionANOVA.csv",sep="/"),head
Regression.ANOVA.Data<-read.csv(file=paste(datapath,"DataForRegressionANOVA.csv",sep="/"),header=TRUE,se
head(Regression.ANOVA.Data)
```

```
##     Output    Input1      Input2
## 1 1.990483 0.2592977  0.08728194
## 2 1.773109 0.5542947 -0.51937150
## 3 2.217273 1.3126170  1.47849923
## 4 2.060437 0.4878214  1.55278763
## 5 1.489216 0.3880355  2.00101613
## 6 2.933162 1.5341078  0.99548754
```

1.2 Fit linear models using: no inputs, only Input1, only Input2, both Input1 and Input2.

```
#Fit four linear models
#no inputs
fit.1<-lm(Output~1,data=Regression.ANOVA.Data)
#only input 1
fit.1.2<-lm(Output~1+Input1,data=Regression.ANOVA.Data)
#Only input 2
fit.1.3<-lm(Output~1+Input2,data=Regression.ANOVA.Data)
# both input 1 and 2
fit.1.2.3<-lm(Output~.,data=Regression.ANOVA.Data)
```

2 Compare ANOVA table of each fit with the summary

2.1 Outputs of anova().

```
anova(fit.1.2)
```

```
## Analysis of Variance Table
##
## Response: Output
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Input1      1 302.89 302.893  813.48 < 2.2e-16 ***
## Residuals 498 185.43   0.372
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Degrees of freedom of anova, 1 for Input 1, 498 for the Residuals
anova(fit.1.2)$Df
```

```
## [1]   1 498
```

```
#dftot = n-1 is total degrees of freedom
dftot <-  length(Regression.ANOVA.Data$Output)-1
#dfI = k-1 is the degrees of freedom from the Input
dfI <- 2-1
#dfR = dftot - dfI is the degrees of freedom of the residuals
dfR <- dftot - dfI
print(c(dfI, dfR))
```

```
## [1]   1 498
```

```
#Sum of squares is equal to the sum of the square of each value subtracted from the mean of
#the sample
anova(fit.1.2)$"Sum Sq"
```

```
## [1] 302.8933 185.4275
```

```
#Total sum of squares is given by subtracting each data Output value, by the sample mean of
#data output, squaring that value and summing them all
SST <- sum((Regression.ANOVA.Data$Output - mean(Regression.ANOVA.Data$Output))^2)
#Sum of squares of residuals, is found in the linear model fit.1.2
SSR <- sum(fit.1.2$residuals^2)
#SSI = SST - SSR gives us the Sum of Squares of Input
SSI <- SST - SSR
print(c(SSI, SSR))
```

```
## [1] 302.8933 185.4275
```

```
#F value is the variation between Sample Means divided by the variation within the samples
anova(fit.1.2)$"F value"[1]
```

```
## [1] 813.4763
```

```
#Mean square estimate of inputs
MSI <- SSI/dfI
#Mean square estimate of the residuals
MSR <- SSR/dfR
#F value is the ratio of these two Mean Square values
Fval <- MSI/MSR
Fval
```

```
## [1] 813.4763
```

```
#Pr(>F) gives us the probability that 2 samples have means this different, tells us if they
#are statistically significant if <0.5
anova(fit.1.2)$"Pr(>F)"[1]
```

```
## [1] 8.790997e-107
```

```
pf(813.4763,1,498, lower.tail = F)
```

```
## [1] 8.791037e-107
```

What does "<2.2e-16" mean in the output of anova()? Since our p-value is smaller than the significance level, we reject the null hypothesis and conclude that the Input 1 describes the variance in the data.

2.2 Compare summary(fit.1) and anova(fit.1)

```
summary(fit.1)
```

```
##
## Call:
## lm(formula = Output ~ 1, data = Regression.ANOVA.Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4078 -0.6506  0.0208  0.6267  3.4581
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.08296    0.04424   47.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9892 on 499 degrees of freedom
```

```
anova(fit.1)
```

```
## Analysis of Variance Table
##
## Response: Output
##            Df Sum Sq Mean Sq F value Pr(>F)
## Residuals 499 488.32  0.9786
```

```
c(anova(fit.1)$"Sum Sq",sum(fit.1$residuals^2))
```

```
## [1] 488.3208 488.3208
```

```r
c(anova(fit.1)$Df,fit.1$df.residual,summary(fit.1)$df[2])
```

```
## [1] 499 499 499
```

Why anova table does not show fields F value and Pr(>F)? In this case, the table does not show F value and Pr(>F) because we are only looking at the Intercept, not at any of the Inputs. Fvalue compares the variances of means of two samples. Here we are only looking at one. Similarly, there can be no Pr(>F) in this case.

2.3 Compare summary(fit.1.2) and anova(fit.1.2)

```r
summary(fit.1.2)
```

```
## 
## Call:
## lm(formula = Output ~ 1 + Input1, data = Regression.ANOVA.Data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.76930 -0.39403 -0.01415  0.41282  1.92275 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1.22598    0.04059   30.20   <2e-16 ***
## Input1       0.80032    0.02806   28.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.6102 on 498 degrees of freedom
## Multiple R-squared:  0.6203, Adjusted R-squared:  0.6195 
## F-statistic: 813.5 on 1 and 498 DF,  p-value: < 2.2e-16
```

```r
anova(fit.1.2)
```

```
## Analysis of Variance Table
## 
## Response: Output
##            Df Sum Sq Mean Sq F value    Pr(>F)    
## Input1      1 302.89 302.893  813.48 < 2.2e-16 ***
## Residuals 498 185.43   0.372                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(fit.1.2)$fstatistic
```

```
##    value    numdf    dendf 
## 813.4763   1.0000 498.0000 
```

```r
c(F.value=anova(fit.1.2)$"F value"[1],Df=anova(fit.1.2)$Df,P.value=anova(fit.1.2)$"Pr(>F)"[1])
```

```
##       F.value           Df1           Df2       P.value 
##  8.134763e+02  1.000000e+00  4.980000e+02 8.790997e-107
```

What is H0 for F value in anova(fit.1.2) and for F-ststistic in summary(fit.1.2)? The null hypothesis in the anova is that the Input1 model is the same as the intercept only model. The null hypothesis in the summary(fit.1.2) is that the slopes of the intercept only model and the linear model are equal.

```r
summary(fit.1.2)$r.squared
```

```
## [1] 0.6202753
```

```r
#Obtain r-squared from anova, and calculate it manually
anova(fit.1.2)$"Sum Sq"[1]/sum(anova(fit.1.2)$"Sum Sq")
```

```
## [1] 0.6202753
```

2.4 Compare summary(fit.1.3) and anova(fit.1.3)

```r
summary(fit.1.3)
```

```
## 
## Call:
## lm(formula = Output ~ 1 + Input2, data = Regression.ANOVA.Data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.2544 -0.6262  0.0086  0.6350  3.4838 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  2.14399    0.06501   32.98   <2e-16 ***
## Input2      -0.05754    0.04494   -1.28    0.201    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.9886 on 498 degrees of freedom
## Multiple R-squared:  0.003281,   Adjusted R-squared:  0.00128 
## F-statistic: 1.639 on 1 and 498 DF,  p-value: 0.201
```

```r
anova(fit.1.3)
```

```
## Analysis of Variance Table
## 
## Response: Output
##            Df Sum Sq Mean Sq F value Pr(>F)
## Input2      1   1.60 1.60219  1.6393  0.201
## Residuals 498 486.72 0.97735
```

What do you conclude from the anova table? Since the Pr(F) is higher than the critical value. We can reject the null hypothesis and say that the variance from Input 2 is not significantly different from the variance explained by the Intercept model.

```
c(F.value=anova(fit.1.3)$"F value"[1],Df=anova(fit.1.3)$Df,P.value=anova(fit.1.3)$"Pr(>F)"[1])
```

```
##     F.value         Df1         Df2     P.value
##   1.6393254   1.0000000 498.0000000   0.2010141
```

```
summary(fit.1.3)$fstatistic
```

```
##      value      numdf      dendf
##   1.639325   1.000000 498.000000
```

What do you conclude from the F-statistic and its p-value? Since the F Value is 1.63, we can say that the ratio of the 2 Mean Square values are similar. Since the P value is high, we conclude that the variance from Input 2 is not significantly different from the variance from the intercept model. The linear model 2 input does not add new information. *What is the minimum level for which you reject the null hypothesis?* We would reject the null hypothesis at a Pvalue of 0.05 *What is the null hypothesis of this F-test?* The null hypothesis of the F-test is stating that the variance from Linear model using input 2 is equal to the variance of the Intercept only linear model.

2.5 Compare summary(fit.1.2.3) and anova(fit.1.2.3)

```
summary(fit.1.2.3)
```

```
##
## Call:
## lm(formula = Output ~ ., data = Regression.ANOVA.Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.76631 -0.39358 -0.01411  0.40432  1.91861
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.21480    0.05179  23.458   <2e-16 ***
## Input1       0.80116    0.02819  28.423   <2e-16 ***
## Input2       0.00970    0.02787   0.348    0.728
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6107 on 497 degrees of freedom
## Multiple R-squared:  0.6204, Adjusted R-squared:  0.6188
## F-statistic: 406.1 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
anova(fit.1.2.3)
```

```
## Analysis of Variance Table
##
## Response: Output
##            Df  Sum Sq Mean Sq  F value Pr(>F)
## Input1      1 302.893 302.893 812.0408 <2e-16 ***
## Input2      1   0.045   0.045   0.1212 0.7279
## Residuals 497 185.382   0.373
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What do you conclude from the anova table? That Input 1 explains much more of the total variance than Input 2 *Compare F-statistic and R2 values.*

The F-statistic is 406.1 The R squared is 0.6204 and the Adjusted is almost the same at 0.6188, this tells us that Input 2 does not add much information

3 Use anova() to compare nested linear models

```
anova(fit.1.2,fit.1.2.3)
```

```
## Analysis of Variance Table
##
## Model 1: Output ~ 1 + Input1
## Model 2: Output ~ Input1 + Input2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    498 185.43
## 2    497 185.38  1    0.0452 0.1212 0.7279
```

```
summary(fit.1.2.3)
```

```
##
## Call:
## lm(formula = Output ~ ., data = Regression.ANOVA.Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.76631 -0.39358 -0.01411  0.40432  1.91861
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.21480    0.05179  23.458   <2e-16 ***
## Input1       0.80116    0.02819  28.423   <2e-16 ***
## Input2       0.00970    0.02787   0.348    0.728
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6107 on 497 degrees of freedom
## Multiple R-squared:  0.6204, Adjusted R-squared:  0.6188
## F-statistic: 406.1 on 2 and 497 DF,  p-value: < 2.2e-16
```

Did adding Input.2 change RSS in the anova table? It decreased it by a very small amount the RSS went from 185.43 to 185.38 *What do you conclude from Pr(>|t|) in summary(fit.1.2.3) and Pr(>F) in* *anova(fit.1.3,fit.1.2.3)?* From the Pr(>t) we conclude that the Intercept and Input1 can explain the variance in the data, while Input 2 does not. The Pr(>F) is 0.7279, so we conclude that Model 2 is not statistically significant from Model 1

Why anova(fit.1.3,fit.1.2.3) returns P-value of F-statistic, but summary(fit.1.2.3) returns Pr(>|t|) of t-statisic? In anova, we use the F-test as a ratio of variances of two samples. In Linear model, the t-test is used to compare slopes.

```
anova(fit.1,fit.1.2.3)
```

```
## Analysis of Variance Table
##
## Model 1: Output ~ 1
## Model 2: Output ~ Input1 + Input2
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    499 488.32
## 2    497 185.38  2    302.94 406.08 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(fit.1.2.3)
```

```
##
## Call:
## lm(formula = Output ~ ., data = Regression.ANOVA.Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.76631 -0.39358 -0.01411  0.40432  1.91861
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.21480    0.05179  23.458   <2e-16 ***
## Input1       0.80116    0.02819  28.423   <2e-16 ***
## Input2       0.00970    0.02787   0.348    0.728
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6107 on 497 degrees of freedom
## Multiple R-squared:  0.6204, Adjusted R-squared:  0.6188
## F-statistic: 406.1 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
c(anova(fit.1,fit.1.2.3)$F[2],summary(fit.1.2.3)$fstatistic[1])
```

```
##           value
## 406.081 406.081
```

Explain what is H0 for F-test in summary(fit.1.2.3) The null Hypothesis states that slopes of Input 1 and Input 2 are equal to zero.