Week8Assignment

John Navarro

November 19, 2016

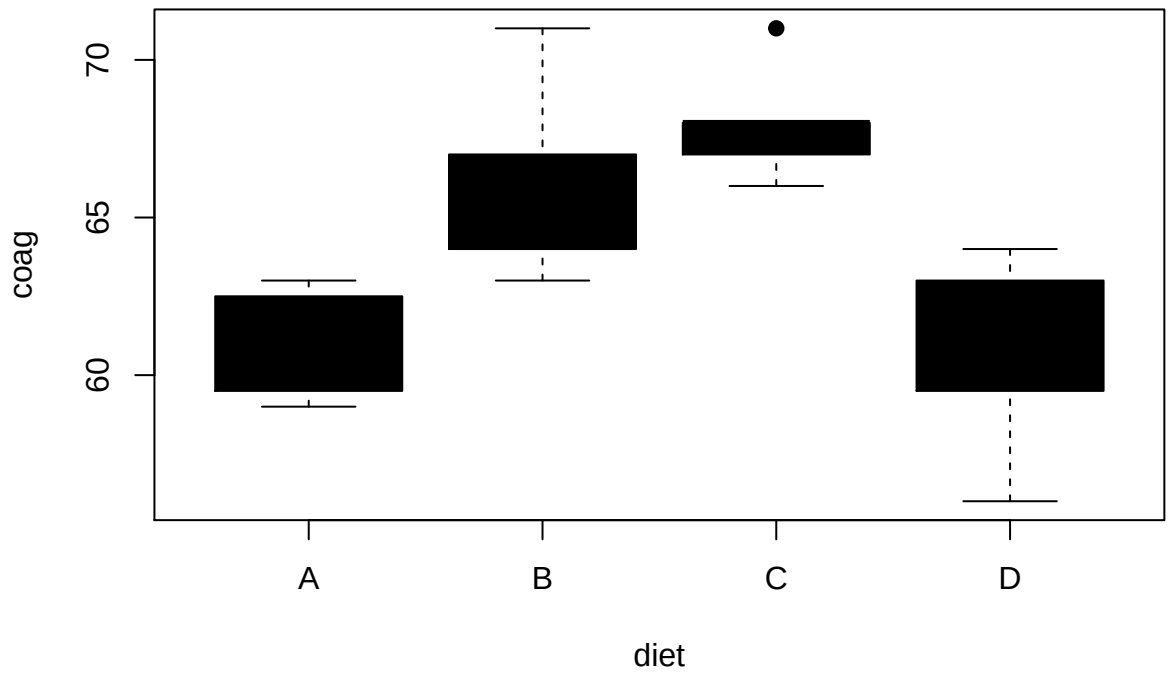
```
#load faraway package, plot coagulation data
library(faraway)
```

```
## Warning: package 'faraway' was built under R version 3.3.2
```

```
coagulation
```

```
##    coag diet
## 1    62    A
## 2    60    A
## 3    63    A
## 4    59    A
## 5    63    B
## 6    67    B
## 7    71    B
## 8    64    B
## 9    65    B
## 10   66    B
## 11   68    C
## 12   66    C
## 13   71    C
## 14   67    C
## 15   68    C
## 16   68    C
## 17   56    D
## 18   62    D
## 19   60    D
## 20   61    D
## 21   63    D
## 22   64    D
## 23   63    D
## 24   59    D
```

```
plot(coag~diet, data=coagulation, pch=19,col="black")
```



Evaluate visually mean values of each group and whether the differences between them are significant to check your intuition later. It seems like the mean values of A and D are close and the mean values of B and C are relatively close. With the wide range of values in B and D, I would say that the differences between them are significant.

```
#Show group means and sizes. Also can separate by diet group
summaryByGroup<-aggregate(coag~diet,data=coagulation,FUN=summary)
means<-cbind(Means=summaryByGroup$coag[,4],Sizes=c(4,6,6,8))
rownames(means)<-as.character(summaryByGroup$diet)
means
```

```
##   Means Sizes
## A    61     4
## B    66     6
## C    68     6
## D    61     8
```

```
Group1.dietA<-subset(coagulation,coagulation$diet=="A")
Group1.dietA
```

```
##   coag diet
## 1   62    A
## 2   60    A
## 3   63    A
## 4   59    A
```

```r
summary(Group1.dietA)
```

```
##       coag       diet
##  Min.   :59.00   A:4
##  1st Qu.:59.75   B:0
##  Median :61.00   C:0
##  Mean   :61.00   D:0
##  3rd Qu.:62.25
##  Max.   :63.00
```

```r
mean(Group1.dietA[,1])
```

```
## [1] 61
```

1.1 ANOVA for the data

```r
#Fit linear model, summary and anova
model <- lm(coag~diet, data=coagulation)
modelSummary <- summary(model)
modelANOVA <- anova(model)
```

Observe the summary, look and interpret results of fitted linear model and regression ANOVA. In the summary, the Pr(t) for the slopes of dietB and diet C are statistically significant from 0. The Rsquared is good at 0.67, and Adjusted R squared is a little lower at 0.62. The fstatistic and pvalue show that there is a relationship with the predictors and the response variable. The p-value is significantly different from 0. The Anova shows the same F statistic/p-values. We can reject the null hypothesis that all slope offsets are equal to zero.

```r
modelSummary$coefficients
```

```
##                 Estimate Std. Error      t value     Pr(>|t|)
## (Intercept) 6.100000e+01   1.183216 5.155441e+01 9.547815e-23
## dietB       5.000000e+00   1.527525 3.273268e+00 3.802505e-03
## dietC       7.000000e+00   1.527525 4.582576e+00 1.805132e-04
## dietD       2.991428e-15   1.449138 2.064281e-15 1.000000e+00
```

```r
modelSummary$df
```

```
## [1]  4 20  4
```

```r
modelSummary$fstatistic
```

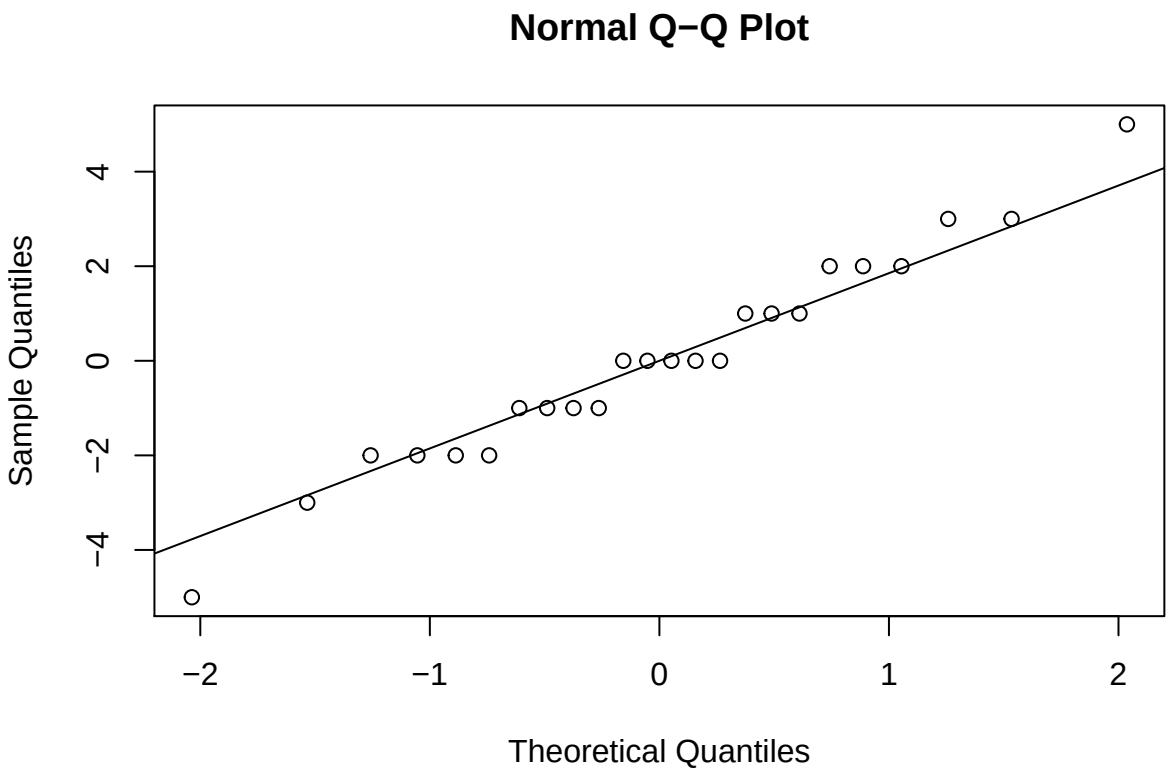
```
##    value    numdf    dendf
## 13.57143  3.00000 20.00000
```

```r
modelANOVA
```

```
## Analysis of Variance Table
##
## Response: coag
##           Df Sum Sq Mean Sq F value    Pr(>F)
## diet       3    228    76.0  13.571 4.658e-05 ***
## Residuals 20    112     5.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. If the formula is coag~diet then why are we getting estimates of the parameters dietB, dietC, and dietD? Because dietA is considered the Intercept, and the slopes of dietB,C and D, will be the difference between that diet group and diet A. *2. Analize statistical significance of all parameters based on p-values and standard errors.* The Intercept has a SE of 1.18 and a very small P value, this tells us that it is significantly different from zero. But it is an intercept at 61, so that seems obvious. DietB has a a slope of 5 and SE of 1.53 and a P value of 0.0038 so it is statistically significat at the 99% level and the difference between the diet B mean and the diet A mean value is different from zero. DietC has a slope of 7 and standard error of 1.53, it's pvalue is 0.00018, which is statistically significant at 99.9% level. The difference bewteen the mean value of dietC is differnt from the mean value of dietA. DietD has a slope of 0 and a SE of 1.45, its Pvalue is 1, and the mean value of diet D is not different from the mean value of dietA. *3. Analize the values of parameters and interpret them. For example, What does the value of coefficient for dietC tell you?* The intercept has a slope of 61 which means that its mean is 61. dietB has a slope of 5, which means that the mean value of dietB(66) is 5 more than the mean value of group A. dietC has a slope of 7, which means that the mean value of dietC(68) is 7 more than the mean value of group A. dietD has a slope of 0, which means that the mean value of dietD(61) is equal to the mean value of group A. *4. Analize the goodness of fit based on the determination coefficient, F-statistic.* Rsquared is 0.67 which is close to 0.70 so it shows a good correlation. The F statistic is 13.57 and its pvalue is small which tells us that it is statistically significant and there is a goodness of fit. *5. Analyze the residuals.*

```
#see if the residuals are distributed normally
qqnorm(model$residuals)
qqline(model$residuals)
```



The residuals look normally distributed in the center, close to zero, but at the extreme edges it looks like they differ from the line. This could be cloudy because of the low levels of residuals(20).

Create matrix with dummy variable inputs for ANOVA. *Why are we creating 3 input variables if we are given 4 groups?* We are looking at the differences in means of the 3 groups(B,C & D) with the first group(A)

```
#Creat data frame coag that has T/F columns for data groups
coag<-coagulation
coag$x1<-coag$diet=="B"
coag$x2<-coag$diet=="C"
coag$x3<-coag$diet=="D"
coag
```

```
##    coag diet    x1    x2    x3
## 1    62    A FALSE FALSE FALSE
## 2    60    A FALSE FALSE FALSE
## 3    63    A FALSE FALSE FALSE
## 4    59    A FALSE FALSE FALSE
## 5    63    B  TRUE FALSE FALSE
## 6    67    B  TRUE FALSE FALSE
## 7    71    B  TRUE FALSE FALSE
## 8    64    B  TRUE FALSE FALSE
## 9    65    B  TRUE FALSE FALSE
## 10   66    B  TRUE FALSE FALSE
## 11   68    C FALSE  TRUE FALSE
## 12   66    C FALSE  TRUE FALSE
## 13   71    C FALSE  TRUE FALSE
```

```
## 14   67    C FALSE  TRUE FALSE
## 15   68    C FALSE  TRUE FALSE
## 16   68    C FALSE  TRUE FALSE
## 17   56    D FALSE FALSE  TRUE
## 18   62    D FALSE FALSE  TRUE
## 19   60    D FALSE FALSE  TRUE
## 20   61    D FALSE FALSE  TRUE
## 21   63    D FALSE FALSE  TRUE
## 22   64    D FALSE FALSE  TRUE
## 23   63    D FALSE FALSE  TRUE
## 24   59    D FALSE FALSE  TRUE
```

```
#Fit full and null linear models for coag~x1+x2+x3 (all inputs) and coag~1 (intercept only). Compare th
coag.model.full<-lm(coag~x1+x2+x3, data=coag)
coag.model.null<-lm(coag~1,data=coag)
anova(coag.model.null,coag.model.full)
```

```
## Analysis of Variance Table
##
## Model 1: coag ~ 1
## Model 2: coag ~ x1 + x2 + x3
##   Res.Df RSS Df Sum of Sq      F    Pr(>F)
## 1     23 340
## 2     20 112  3       228 13.571 4.658e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coag.model <- lm(coag~diet, data=coag)
summary(coag.model)
```

```
##
## Call:
## lm(formula = coag ~ diet, data = coag)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5.00  -1.25   0.00   1.25   5.00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.100e+01  1.183e+00  51.554  < 2e-16 ***
## dietB       5.000e+00  1.528e+00   3.273 0.003803 **
## dietC       7.000e+00  1.528e+00   4.583 0.000181 ***
## dietD       2.991e-15  1.449e+00   0.000 1.000000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.366 on 20 degrees of freedom
## Multiple R-squared:  0.6706, Adjusted R-squared:  0.6212
## F-statistic: 13.57 on 3 and 20 DF,  p-value: 4.658e-05
```

```r
anova(coag.model)
```

```
## Analysis of Variance Table
##
## Response: coag
##           Df Sum Sq Mean Sq F value    Pr(>F)
## diet       3    228    76.0  13.571 4.658e-05 ***
## Residuals 20    112     5.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Compare and explain the output of anova(coag.model.null,coag.model.full) with the outputs of summary(coag.model) and anova(coag.model).

1. Compare degrees of freedom. anova of the null and full linear models have 3 degrees of freedom in the full lm, for the 3 predictors(X1, X2, X3). summary(coag.model) gives 3df for the 3 predictors and 20 for the number in the sample minus the 4 coefficents in the model including the intercept. anova(coag.model) gives 3df for the 3 predictors and 20 for the number in the sample minus the 4 coefficents in the model including the intercept. *2. Compare the sums of squares.* Anova of the null/full gives no sum of squares (null) since the fitted values are all on the horizontal line which goes through the intercept. Model Full has sum of squares = 228 anova(coag.model) gives sum of squares of 228 for the diet predictors and sum of squares of 112 for the residuals *3. compare the p-values.* All the p-values are the same 4.658e-05 (all zero)

Calculate manually the sum of squares shown in ANOVA table.

In order to do that we need grand mean and the vector of full length of group means.

```r
#Calculate sum of square
grand.mean <-  mean(coagulation$coag)
group.mean <- as.vector(coag$coag)
#Calculate SST,SSE and SSM. Observe decomposition of the variance representation SST.
SST <- sum((coagulation$coag - grand.mean)^2)
SSM=0
for(i in 1:4){
    SSM <- SSM+((means[i,1]-grand.mean)^2)*means[i,2]
}
SSE <- SST-SSM
c(SST=SST,SSE=SSE,SSM=SSM)
```

```
## SST SSE SSM
## 340 112 228
```

```r
anova(coag.model)
```

```
## Analysis of Variance Table
##
## Response: coag
##           Df Sum Sq Mean Sq F value    Pr(>F)
## diet       3    228    76.0  13.571 4.658e-05 ***
## Residuals 20    112     5.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(coag.model.null,coag.model.full)
```

```
## Analysis of Variance Table
##
## Model 1: coag ~ 1
## Model 2: coag ~ x1 + x2 + x3
##   Res.Df RSS Df Sum of Sq      F    Pr(>F)
## 1     23 340
## 2     20 112  3       228 13.571 4.658e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2 Experiment plan

Check what experiment plan (basis) R uses in lm:

```r
model.matrix(coag.model)
```

```
##    (Intercept) dietB dietC dietD
## 1            1     0     0     0
## 2            1     0     0     0
## 3            1     0     0     0
## 4            1     0     0     0
## 5            1     1     0     0
## 6            1     1     0     0
## 7            1     1     0     0
## 8            1     1     0     0
## 9            1     1     0     0
## 10           1     1     0     0
## 11           1     0     1     0
## 12           1     0     1     0
## 13           1     0     1     0
## 14           1     0     1     0
## 15           1     0     1     0
## 16           1     0     1     0
## 17           1     0     0     1
## 18           1     0     0     1
## 19           1     0     0     1
## 20           1     0     0     1
## 21           1     0     0     1
## 22           1     0     0     1
## 23           1     0     0     1
## 24           1     0     0     1
## attr(,"assign")
## [1] 0 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$diet
## [1] "contr.treatment"
```

Explain the meaning of this matrix. This is a matrix that represents the basis that R uses in ANOVA. The first column is all 1's and the other 3 have 1's corresponding to their groups. For example, if the data point is in dietB, that row will have a 1, if not, it will have a 0.

```r
#Fit alternative model without intercept.
#Check its experiment plan (basis).

coag.altmodel<-lm(coag~diet-1,data=coagulation)
summary(coag.model)
```

```
##
## Call:
## lm(formula = coag ~ diet, data = coag)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5.00  -1.25   0.00   1.25   5.00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.100e+01  1.183e+00  51.554  < 2e-16 ***
## dietB       5.000e+00  1.528e+00   3.273 0.003803 **
## dietC       7.000e+00  1.528e+00   4.583 0.000181 ***
## dietD       2.991e-15  1.449e+00   0.000 1.000000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.366 on 20 degrees of freedom
## Multiple R-squared:  0.6706, Adjusted R-squared:  0.6212
## F-statistic: 13.57 on 3 and 20 DF,  p-value: 4.658e-05
```

```r
anova(coag.altmodel)
```

```
## Analysis of Variance Table
##
## Response: coag
##           Df Sum Sq Mean Sq F value    Pr(>F)
## diet       4  98532 24633.0  4398.8 < 2.2e-16 ***
## Residuals 20    112     5.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
model.matrix(coag.altmodel)
```

```
##    dietA dietB dietC dietD
## 1      1     0     0     0
## 2      1     0     0     0
## 3      1     0     0     0
## 4      1     0     0     0
## 5      0     1     0     0
## 6      0     1     0     0
## 7      0     1     0     0
## 8      0     1     0     0
## 9      0     1     0     0
## 10     0     1     0     0
## 11     0     0     1     0
```

```
## 12     0     0     1     0
## 13     0     0     1     0
## 14     0     0     1     0
## 15     0     0     1     0
## 16     0     0     1     0
## 17     0     0     0     1
## 18     0     0     0     1
## 19     0     0     0     1
## 20     0     0     0     1
## 21     0     0     0     1
## 22     0     0     0     1
## 23     0     0     0     1
## 24     0     0     0     1
## attr(,"assign")
## [1] 1 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$diet
## [1] "contr.treatment"
```

Compare ANOVA tables for coag.model and coag.altmodel. Explain difference between number of degrees of freedom for diet. diet has one extra df in the anova for coag.altmodel, this is because it has 4 predictors and no intercept. The anova for coag.model has 3 degrees of freedom, because it has an intercept and 3 predictors. *Explain difference between sums of squares of diet* The total sum of squares in anova(coag.model) is showing the difference between the means of diet groups and the grand mean(64). The Sum ofSquares of Diet is the portion explained by the predictors. The total sum of squares in anova(coag.atlmodel) is a much larger number. This is describing the diffrence between the means and zero. *Explain difference between F-statistics.* The F value is the ratio of the mean squared errors and since the SSM is so large in the coag.altmodel, it gives us a large F value for diet in anova(coag.altmodel) *Explain the differences between the two bases. Explain difference between null hypotheses tested by the two F-tests.* The basis for the coag.model has 1's in the first column and 0s and 1s in the 2nd 3rd and 4th column, corresponding to the groups. The basis for the coag.altmodel has 1s and 0s in all the columns corresponding to the groups.

The Null hypothesis of anova(coag.model) is looking to see if all the offsets of the slopes are zero. The null hypothesis of anova(coag.altmodel) is looking to see if all the means for the different groups are equal to zero.