525 Lab 1

Anya Conti

April 8, 2017

```r
data <- read.csv("/Users/Anya/Documents/JuniorYear/Spring/Stat 525/newproduct.csv")
library(car)

head(data)
```

```
##     time     cost     FTEs difficulty
## 1 135.37 6.556668 83.21995        1.5
## 2  87.44 3.194270 58.52738        0.8
## 3 127.85 6.166437 83.47341        1.3
## 4  93.35 5.220994 70.56098       -0.1
## 5  99.89 5.308939 67.39017        1.2
## 6 134.80 6.819050 61.41208       -0.2
```

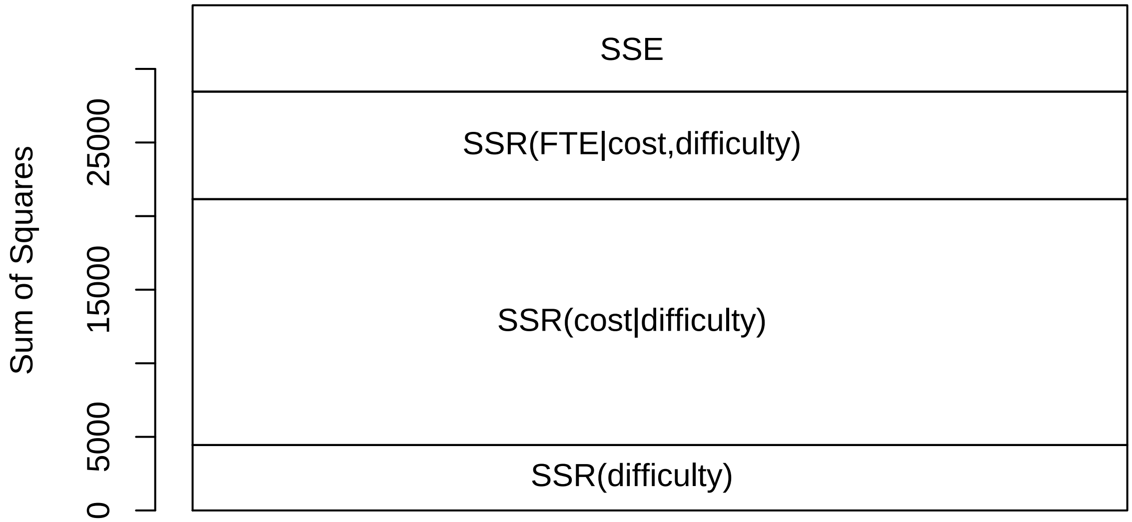
```r
tail(data)
```

```
##      time     cost     FTEs difficulty
## 45 107.67 5.682397 53.45378        0.5
## 46  87.84 4.470095 78.87357        0.4
## 47  82.63 4.213845 28.13952        0.6
## 48 100.10 4.528173 75.26041        0.6
## 49  92.12 5.777649 78.29541        0.6
## 50  97.75 4.283243 62.16149        1.4
```

```r
fit.cFd <- lm(time~cost+FTEs+difficulty,data=data)
fit.Fcd <- lm(time~FTEs+cost+difficulty,data=data)
fit.dcF <- lm(time~difficulty+cost+FTEs,data=data)
fit.cdF <- lm(time~cost+difficulty+FTEs,data=data)
fit.Fdc <- lm(time~FTEs+difficulty+cost,data=data)
fit.dFc <- lm(time~difficulty+FTEs+cost,data=data)

anova.result <- anova(fit.dcF)
labels <- c("SSR(difficulty)","SSR(cost|difficulty)","SSR(FTE|cost,difficulty)")

barplot(as.matrix(anova.result[,2]),density=0,ylab="Sum of Squares")
text(.67,anova.result[1,2]/2,labels[1])
text(.67,anova.result[1,2]+anova.result[2,2]/2,labels[2])
text(.67,anova.result[1,2]+anova.result[2,2]+anova.result[3,2]/2,labels[3])
text(.67,anova.result[1,2]+anova.result[2,2]+anova.result[3,2]+anova.result[4,2]/2,"SSE")
```



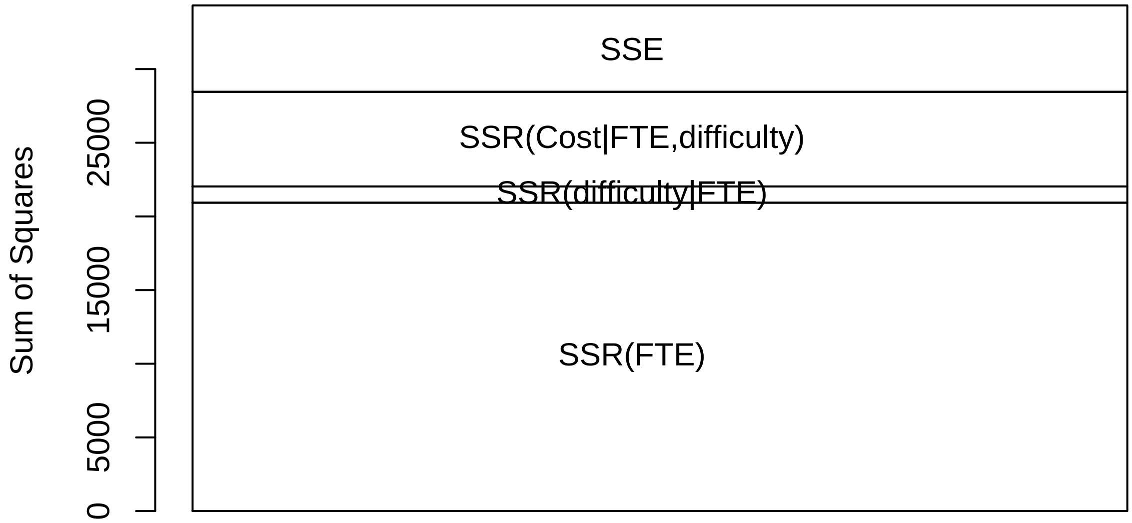
Note: *Throughout these questions, ESS = Explained Sum of Squares (what you call SSR), RSS = Residual Sum of Squares (what you call SSE), and TSS = Total Sum of Squares*
Also, Degrees of freedom of model = n-k-1 where k = number of independent variables.

1. The total height of the stacked bars represents TSS, the total sum of squares of time (total variation in time), which is ESS, the explained sum of squares (the variation in time explained by the model), plus RSS, the residual sum of sqares (the variation in time not explained by the model)

2. See code and figure below.

```
anova.result2 <- anova(fit.Fdc)
labels <- c("SSR(FTE)","SSR(difficulty|FTE)","SSR(Cost|FTE,difficulty)")

barplot(as.matrix(anova.result2[,2]),density=0,ylab="Sum of Squares")
text(.67,anova.result2[1,2]/2,labels[1])
text(.67,anova.result2[1,2]+anova.result2[2,2]/2,labels[2])
text(.67,anova.result2[1,2]+anova.result2[2,2]+anova.result2[3,2]/2,labels[3])
text(.67,anova.result2[1,2]+anova.result2[2,2]+anova.result2[3,2]+anova.result2[4,2]/2,"SSE")
```



3. var(y) in the model is calculated as an estimate for the true population variance using the formula below

$$var(y) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$n = 50$$

$$var(y) = \frac{1}{50-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$49 * var(y) = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Thus, these two lines of code are eqivalent.

```
var(data$time)*(49)
```

```
## [1] 34324.75
```

```
sum(anova(fit.dcF)[,2])
```

```
## [1] 34324.75
```

4. FTEs explains highest amount of variance in time, but cost is not far behind. This can be found by looking at the simple linear regression of each, and how much of the variance in time each model explains. The anova tables of these models are shown below.

```
anova(lm(time~cost, data=data))
```

```
## Analysis of Variance Table
##
## Response: time
##           Df Sum Sq Mean Sq F value   Pr(>F)
## cost       1  20542 20542.3  71.542 4.51e-11 ***
## Residuals 48  13782   287.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(time~FTEs, data=data))
```

```
## Analysis of Variance Table
##
## Response: time
##           Df Sum Sq Mean Sq F value    Pr(>F)
## FTEs       1  20930 20930.3  75.006 2.253e-11 ***
## Residuals 48  13394   279.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(time~difficulty, data=data))
```

```
## Analysis of Variance Table
##
## Response: time
##            Df  Sum Sq Mean Sq F value  Pr(>F)
## difficulty  1  4445.6  4445.6  7.1417 0.01026 *
## Residuals  48 29879.2   622.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5. On its own, cost explains 59.85% of the variation in time ($R^2 = \frac{ESS}{TSS}$) or has an ESS of 20542. ESS of cost given that difficulty and FTE are already in the model is only 6421.5. Thus the ESS definitely does depend on what else is in the model. Each of the variables is correlated to some extent with the other variables in the model (see table of correlation coefficients below). Thus, some part of the variation in time that is explained by one of the variables can also be explained by the others. Essentially there is a bit of overlap in what they can explain because of their correlation with each other. The 6421.5 represents the variance that cost can explain alone, and that difficulty and FTE cannot explain despite the overlap. The amount that cost can explain does not truly change however, just some of that has already been explained.

```
summary(lm(time~cost, data=data))
```

```
##
## Call:
## lm(formula = time ~ cost, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.308 -10.603   0.844   8.915  43.895
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.666     11.160   0.687    0.495
## cost          18.455      2.182   8.458 4.51e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.95 on 48 degrees of freedom
## Multiple R-squared:  0.5985, Adjusted R-squared:  0.5901
## F-statistic: 71.54 on 1 and 48 DF,  p-value: 4.51e-11
```

```
anova(lm(time~cost, data=data))
```

```
## Analysis of Variance Table
##
## Response: time
##           Df Sum Sq Mean Sq F value   Pr(>F)
## cost       1  20542 20542.3  71.542 4.51e-11 ***
## Residuals 48  13782   287.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit.dFc)
```

```
## Analysis of Variance Table
##
## Response: time
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## difficulty  1  4445.6  4445.6  34.843 4.049e-07 ***
## FTEs        1 17588.6 17588.6 137.855 1.942e-15 ***
## cost        1  6421.5  6421.5  50.330 6.579e-09 ***
## Residuals  46  5869.0   127.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cor(data)
```

```
##                 time      cost      FTEs difficulty
```

```
## time       1.0000000 0.7736073 0.7808801  0.3598820
## cost       0.7736073 1.0000000 0.4686515  0.3010203
## FTEs       0.7808801 0.4686515 1.0000000  0.2378011
## difficulty 0.3598820 0.3010203 0.2378011  1.0000000
```

6. On its own, FTE explains 60.98% of the variation in time ($R^2 = \frac{ESS}{TSS}$) or has an ESS of 20930. ESS of FTE given that difficulty and cost are already in the model is only 7304.6. Thus the ESS definitely does depend on what else is in the model. Each of the variables is correlated to some extent with the other variables in the model (see table of correlation coefficients below). Thus, some part of the variation in time that is explained by one of the variables can also be explained by the others. Essentially there is a bit of overlap in what they can explain because of their correlation with each other. The 7304.6 represents the variance that FTE can explain alone, and that difficulty and cost cannot explain despite the overlap. The amount that FTE can explain does not truly change however, just some of that has already been explained.

```r
summary(lm(time~FTEs, data=data))
```

```
##
## Call:
## lm(formula = time ~ FTEs, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.913  -9.665  -2.222  11.607  36.464
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 49.05910    6.32362   7.758 5.12e-10 ***
## FTEs         0.80240    0.09265   8.661 2.25e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.7 on 48 degrees of freedom
## Multiple R-squared:  0.6098, Adjusted R-squared:  0.6016
## F-statistic: 75.01 on 1 and 48 DF,  p-value: 2.253e-11
```

```r
anova(lm(time~FTEs, data=data))
```

```
## Analysis of Variance Table
##
## Response: time
##           Df Sum Sq Mean Sq F value    Pr(>F)
## FTEs       1  20930 20930.3  75.006 2.253e-11 ***
## Residuals 48  13394   279.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(fit.cdF)
```

```
## Analysis of Variance Table
##
## Response: time
##            Df  Sum Sq Mean Sq  F value    Pr(>F)
## cost        1 20542.3 20542.3 161.0050 < 2.2e-16 ***
## difficulty  1   608.9   608.9   4.7723   0.03405 *
## FTEs        1  7304.6  7304.6  57.2512 1.297e-09 ***
## Residuals  46  5869.0   127.6
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
cor(data)
```

```
##                 time      cost      FTEs difficulty
## time       1.0000000 0.7736073 0.7808801  0.3598820
## cost       0.7736073 1.0000000 0.4686515  0.3010203
## FTEs       0.7808801 0.4686515 1.0000000  0.2378011
## difficulty 0.3598820 0.3010203 0.2378011  1.0000000
```

7. On its own, difficulty explains 12.95% of the variation in time ($R^2 = \frac{ESS}{TSS}$) or has an ESS of 4445.6. ESS of difficulty given that FTE and cost are already in the model is only 216.1. Thus the ESS definitely does depend on what else is in the model. Each of the variables is correlated to some extent with the other variables in the model (see table of correlation coefficients below). Thus, some part of the variation in time that is explained by one of the variables can also be explained by the others. Essentially there is a bit of overlap in what they can explain because of their correlation with each other. The 216.1 represents the variance that difficulty can explain alone, and that FTE and cost cannot explain despite the overlap. The amount that difficulty can explain does not truly change however, just some of that has already been explained.

```r
summary(lm(time~difficulty, data=data))
```

```
##
## Call:
## lm(formula = time ~ difficulty, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.074 -13.861   0.274  17.641  51.962
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   86.202      6.211  13.880   <2e-16 ***
## difficulty    16.820      6.294   2.672   0.0103 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.95 on 48 degrees of freedom
## Multiple R-squared:  0.1295, Adjusted R-squared:  0.1114
## F-statistic: 7.142 on 1 and 48 DF,  p-value: 0.01026
```

```r
anova(lm(time~difficulty, data=data))
```

```
## Analysis of Variance Table
##
## Response: time
##            Df  Sum Sq Mean Sq F value  Pr(>F)
## difficulty  1  4445.6  4445.6  7.1417 0.01026 *
## Residuals  48 29879.2   622.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(fit.cFd)
```

```
## Analysis of Variance Table
##
## Response: time
##            Df  Sum Sq Mean Sq  F value    Pr(>F)
## cost        1 20542.3 20542.3 161.0050 < 2.2e-16 ***
```

```
## FTEs        1  7697.4  7697.4  60.3300 6.531e-10 ***
## difficulty  1   216.1   216.1   1.6934    0.1996
## Residuals  46  5869.0   127.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
cor(data)
```

```
##                 time      cost      FTEs difficulty
## time       1.0000000 0.7736073 0.7808801  0.3598820
## cost       0.7736073 1.0000000 0.4686515  0.3010203
## FTEs       0.7808801 0.4686515 1.0000000  0.2378011
## difficulty 0.3598820 0.3010203 0.2378011  1.0000000
```

7. The estimated model is as follows:

$$\hat{time} = 2.65327 + 11.97624 * cost + 3.91443 * difficulty + 0.54017 * FTEs$$

8. Estmated variance of the erros is calculated with $\sqrt{\frac{RSS}{n-k-1}}$ where k is number of explanatory variables so n-k-1 = degrees of freedom. This can be found on a summary table of the regression as Residual standard error: 11.3.

```r
sqrt(5869/(46))
```

```
## [1] 11.29544
```

9. 82.9% of the variation in time can be explained by the model. This is the R^2 value which can be found using $R^2 = \frac{ESS}{TSS}$ This can be found on the summary table as Multiple R-squared: 0.829.

```r
sum(anova(fit.dcF)[1:3,2])/sum(anova(fit.dcF)[,2])
```

```
## [1] 0.8290144
```

10. Let k be number of explanatory variables so n-k-1 = degrees of freedom, then the F-statistic is calculated as follows:

$$\frac{ESS/k}{RSS/(n-k-1)}$$

```r
(sum(anova(fit.dcF)[1:3,2])/3)/(anova(fit.dcF)[4,2]/46)
```

```
## [1] 74.34281
```

The F-statistic for two models 1 and 2 where all variables in model 1 are included in model 2, plus at least one additional explanatory variable in model 2 is given by

$$\frac{(RSS_1 - RSS_2)/(k_2 - k_1)}{(RSS_2)/(n - k_2 - 1)}$$

Note: Both models use some of the same variables and data so $n_1 = n_2 = n$ Let model 1 be a model with no explanatory variables (just intercept which must go through $\bar{y}$ so $\hat{y} = \bar{y}$) so $RSS_1 = \sum_{i=1}^{n}(\hat{y} - y_i)^2 = \sum_{i=1}^{n}(\bar{y} - y_i)^2 = TSS_1$, and $k = 0$ Let model 2 be a model with at least 1 variable in it, and same y. Note: $TSS_2 = \sum_{i=1}^{n}(\bar{y} - y_i)^2 = TSS_1 = RSS_1$ F-statistic to compare the two models is

$$\frac{(RSS_1 - RSS_2)/(k_2 - k_1)}{(RSS_2)/(n - k_2 - 1)}$$

$$\frac{(TSS_1 - RSS_2)/(k_2 - 0)}{(RSS_2)/(n - k_2 - 1)}$$

$$\frac{(TSS_2 - RSS_2)/(k_2)}{(RSS_2)/(n - k_2 - 1)}$$

$$\frac{(ESS_2)/(k_2)}{(RSS_2)/(n-k_2-1)}$$

This is the same as the formula used above for the one model. This is essentially comparing if some model (with any variables) is a better predictor of y compared to a model that just uses an intercept (set as the mean of y), which is the model with no variables in it Since the p-value for this F-test is less than 2.2e-16, we can reject the null hypothesis that this model (model 2) does not provide a significantly better fit compared to the model (model 1) with no variables, and thus use the model with variables as a result

Note: This is the same as a test of restrictions, where $H_0 : R\beta = r$ and $H_A : R\beta \neq r$, where r is a vertical vertor of length 3 filled with 0's, β is a vertical vector with all the model parameters, and R is the following matrix:

```r
R <- cbind(c(0,0,0),diag(1,3))
r <- c(0,0,0)

R
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    0    1    0    0
## [2,]    0    0    1    0
## [3,]    0    0    0    1
```

```r
linearHypothesis(fit.dcF, R, r)
```

```
## Linear hypothesis test
##
## Hypothesis:
## difficulty = 0
## cost = 0
## FTEs = 0
##
## Model 1: restricted model
## Model 2: time ~ difficulty + cost + FTEs
##
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     49 34325
## 2     46  5869  3     28456 74.343 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```