SLR and ANOVA Analysis of UCLA STEM Students

Noah Jones

10/13/2021

An exploratory data analysis of UCLA STEM student data. Of the 55 variables collected, we focus on students' Socio-Economic Struggle (SES), Mother and Father's Education,

```r
library(car)
```

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

```r
library(effects)
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```r
library(lsmeans)
```

```
## Loading required package: emmeans
```

```
## The 'lsmeans' package is now basically a front end for 'emmeans'.
## Users are encouraged to switch the rest of the way.
## See help('transition') for more information, including how to
## convert old 'lsmeans' objects and scripts to work with 'emmeans'.
```

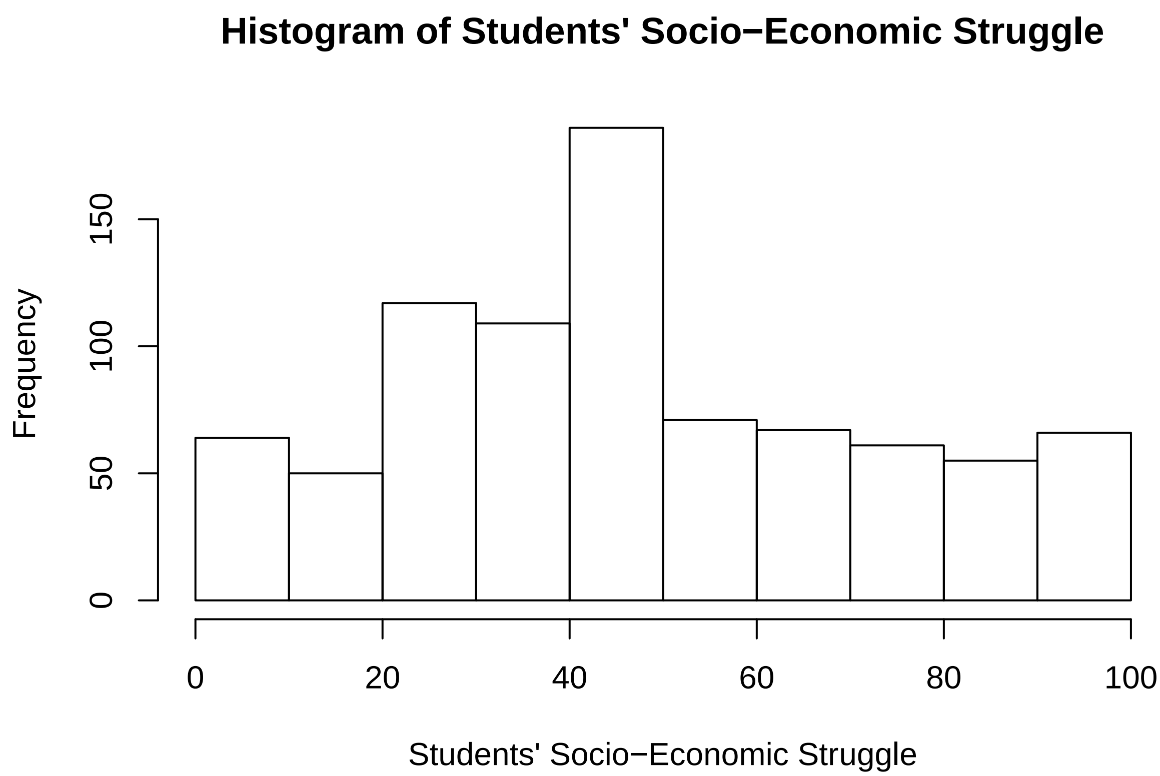
```r
library(mvtnorm)
library(survival)
library(MASS)
library(multcomp)
```

```
## Loading required package: TH.data
```

```
##
## Attaching package: 'TH.data'
```

```
## The following object is masked from 'package:MASS':
##
##     geyser
```

```
stem <- read.csv("stemjune20.csv")
hist(stem$SES, xlab = "Students' Socio-Economic Struggle",
     main = "Histogram of Students' Socio-Economic Struggle")
```



The Socio-Economic Struggle data appears to be approximately normal, and slightly right skewed.

Computing summary statistics for SES by Father's Education

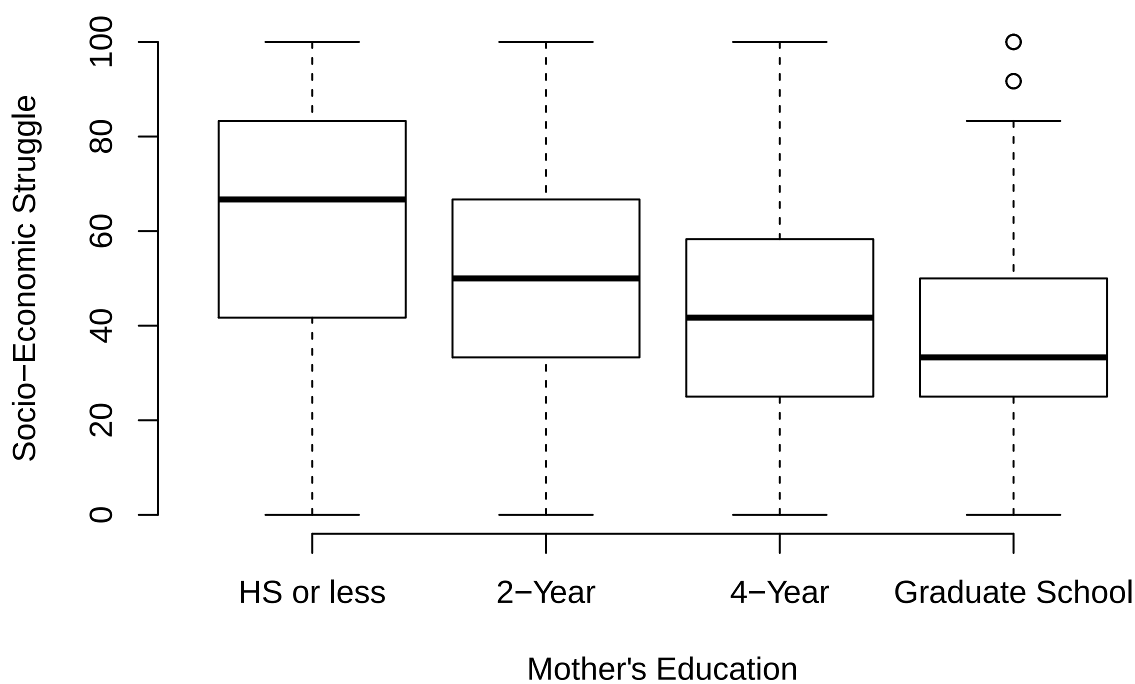
```
tapply(stem$SES, stem$FatherEdu, summary)
```

```
## [[1]]
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   41.70   50.00   58.30   64.29   75.00  100.00      11
##
## $`Four-year College`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   25.00   41.70   44.16   58.30  100.00       3
##
## $`Graduate or Professional Degree`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0    25.0    33.3    37.6    50.0   100.0       3
##
## $`High school or less`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
```

```
##    0.00   50.00   66.70   64.86   83.30  100.00       2
##
## $`Two-year College`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   33.30   50.00   54.47   75.00  100.00       1
```

Computing boxplot of SES by Mother's Education

```
boxplot(SES~factor(MotherEdu, levels = levels(stem$MotherEdu)[c(4,5,2,3)]),
        data = stem, axes = FALSE,xlab = "Mother's Education", ylab = "Socio-Economic Struggle")
axis(side = 2)
axis(side = 1, at = c(1,2,3,4),
     labels = c("HS or less", "2-Year", "4-Year", "Graduate School"))
```



Based on this side by side boxplot, it would appear that the higher a student's mother's education is, the lower their socio-economic struggle tends to be.

Since MotherEdu is a categorical variable with 4 levels, we could run an F test with ANOVA, which would test whether any of the 4 means are significantly different from the total mean. See below code for this analysis.

```
m1 <- aov(stem$SES~stem$MotherEdu)
summary(m1)
```

```
##                   Df Sum Sq Mean Sq F value Pr(>F)
```

```
## stem$MotherEdu   4  91415   22854   40.63 <2e-16 ***
## Residuals      841 473071     563
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 20 observations deleted due to missingness
```

```r
TukeyHSD(m1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = stem$SES ~ stem$MotherEdu)
##
## $`stem$MotherEdu`
##                                                           diff        lwr
## Four-year College-                                  -2.8421384 -40.451424
## Graduate or Professional Degree-                   -11.0373512 -48.720315
## High school or less-                                16.7425758 -20.944915
## Two-year College-                                    0.8111111 -37.308922
## Graduate or Professional Degree-Four-year College   -8.1952128 -13.850815
## High school or less-Four-year College               19.5847141  13.899027
## Two-year College-Four-year College                   3.6532495  -4.416260
## High school or less-Graduate or Professional Degree  27.7799269  21.625709
## Two-year College-Graduate or Professional Degree     11.8484623   3.442254
## Two-year College-High school or less               -15.9314646 -24.357944
##                                                          upr     p adj
## Four-year College-                                 34.767147 0.9995945
## Graduate or Professional Degree-                   26.645612 0.9303886
## High school or less-                               54.430066 0.7430054
## Two-year College-                                  38.931144 0.9999974
## Graduate or Professional Degree-Four-year College  -2.539611 0.0007689
## High school or less-Four-year College              25.270402 0.0000000
## Two-year College-Four-year College                 11.722759 0.7293064
## High school or less-Graduate or Professional Degree 33.934145 0.0000000
## Two-year College-Graduate or Professional Degree   20.254671 0.0011831
## Two-year College-High school or less               -7.504985 0.0000029
```

Our F test in the ANOVA table has a very low p value $<$2e-16, so we reject the null hypothesis that the 4 population means are equal, concluding that at least one of the pairs is statistically different. In our Post-Hocs, which test significance of each pair, We see that the differences in Mother's Education between Graduate and 4 year college, Graduate and 2 year college, Graduate and High school, 4 year college and high school, and 2 year college and high school are all significant in predicting a student's Socio-Economic Struggle.

Below we compute the 95% confidence interval for the population mean of SES using SLR.

```r
m1 <- lm(stem$SES~1)
confint(m1)
```

```
##                2.5 %   97.5 %
## (Intercept) 45.92996 49.41826
```

We are 95% confident that the population mean for students' socio-economic struggle lies between 45.92996 and 49.41826

Computing the 95% confidence interval by hand.

Since we have a large sample size, we can approximate using the 95% Z score, 1.96, for our confidence interval calculation: $\bar{X} \pm 1.96 \cdot S_{\bar{X}}$

```
S_xbar <- sqrt(var(stem$SES, na.rm = TRUE))/sqrt(length(stem$SES))
upper <- mean(stem$SES, na.rm = TRUE) + S_xbar*1.96
lower <- mean(stem$SES, na.rm = TRUE) - S_xbar*1.96
c(lower, upper)
```

```
## [1] 45.95266 49.39557
```

Here, we are dealing with students' ability to cope with academic stress as a predictor for students' perception on the quality of UCLA academics.

We first check that all assumptions of ANOVA are met, namely Normality, Independence, and Homogeneity of Residuals. We follow that by running an ANOVA, with the above listed predictor and outcome variable.

```
attach(stem)
copeacadstress<-recode(Q3.13,"'Agree'='always';'Disagree'='rarely';'Not Sure'='sometimes';'Strongly Agr
table(copeacadstress)
```

```
## copeacadstress
##    rarely sometimes    always
##       155       242       452
```

```
tapply(Academic,copeacadstress,var,na.rm=1)
```

```
##    rarely sometimes    always
## 175.33402  96.51183 132.98628
```

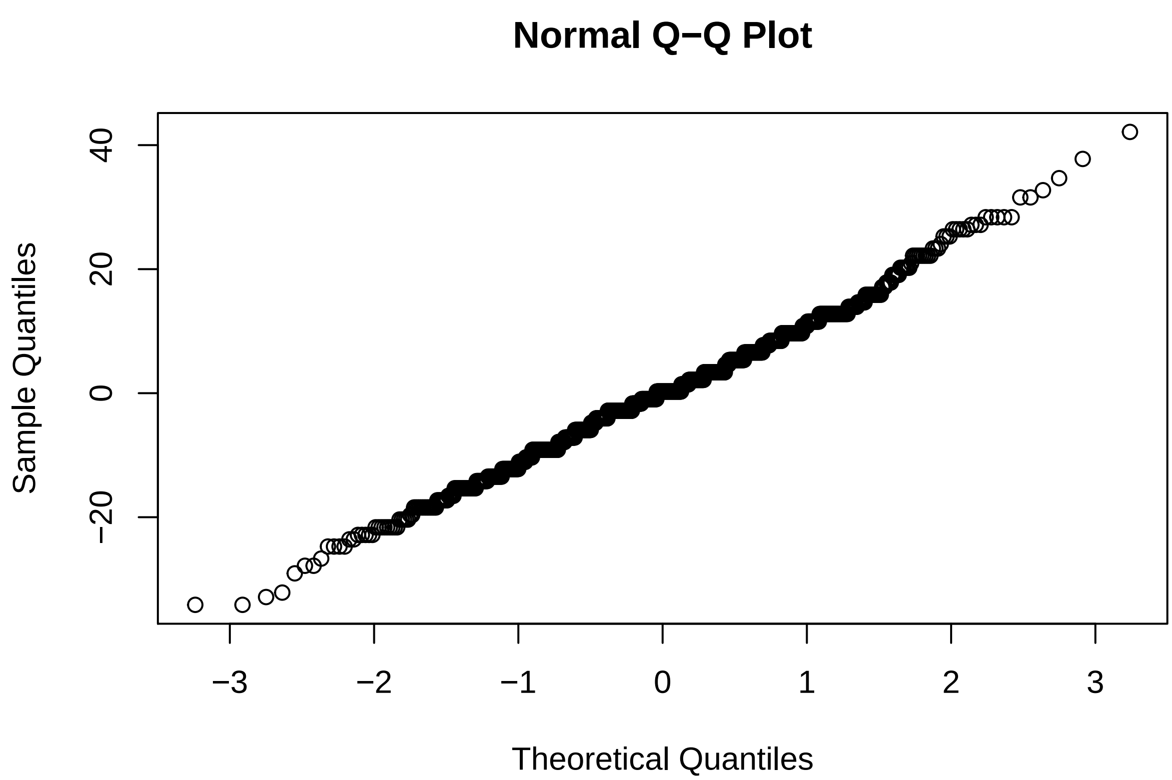
```
F_max <- 175.33402/96.51183
F_crit <- 1.85
F_max < F_crit # So we fail to reject H0
```

```
## [1] TRUE
```

```
m2 <- aov(Academic~copeacadstress)
shapiro.test(resid(m2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(m2)
## W = 0.99471, p-value = 0.005251
```

```r
qqnorm(resid(m2))
```




```r
summary(m2)
```

```
##                  Df Sum Sq Mean Sq F value Pr(>F)    
## copeacadstress    2  65624   32812   251.7 <2e-16 ***
## Residuals       833 108612     130                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 30 observations deleted due to missingness
```

```r
tapply(Academic,copeacadstress,mean,na.rm=1)
```

```
##    rarely sometimes    always 
##  39.17273  50.95672  62.22838
```

```r
TukeyHSD(m2)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## 
## Fit: aov(formula = Academic ~ copeacadstress)
## 
```

```
## $copeacadstress
##                     diff       lwr      upr p adj
## sometimes-rarely 11.78400  9.011366 14.55662     0
## always-rarely    23.05565 20.548409 25.56289     0
## always-sometimes 11.27166  9.117834 13.42548     0
```

Here, we run a very similar analysis but under an ANCOVA model. The ANCOVA, or Analysis of Covariance, model includes the effect of a potential covariate on our outcome variable, UCLA students' Sense of Belonging. We check our ANCOVA assumptions, calculate means after adjusted for the covariate, calculate Regression beta estimates for beta 1, 2, and 3, run necessary Post-Hocs on our ANCOVA model, and lastly compute the achieved power, or practical significance, of our findings.

```r
m3 <- aov(Academic~copeacadstress*Belonging)
summary(m3)
```

```
##                           Df Sum Sq Mean Sq F value  Pr(>F)
## copeacadstress             2  65881   32941 315.047 < 2e-16 ***
## Belonging                  1  21170   21170 202.469 < 2e-16 ***
## copeacadstress:Belonging   2   1010     505   4.829 0.00822 **
## Residuals                821  85842     105
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 39 observations deleted due to missingness
```

```r
shapiro.test(resid(m3))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(m3)
## W = 0.9972, p-value = 0.1656
```

```r
m4 <- aov(Academic~copeacadstress+Belonging)
summary(m4)
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## copeacadstress   2  65881   32941   312.1 <2e-16 ***
## Belonging        1  21170   21170   200.6 <2e-16 ***
## Residuals      823  86852     106
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 39 observations deleted due to missingness
```

```r
cor(Academic,Belonging,use="complete.obs")
```

```
## [1] 0.5298407
```

```r
lsmeans(m4, "copeacadstress")
```

```
##  copeacadstress lsmean    SE  df lower.CL upper.CL
##  rarely           42.1 0.857 823     40.4     43.8
##  sometimes        51.7 0.676 823     50.4     53.0
##  always           60.8 0.500 823     59.8     61.8
##
## Confidence level used: 0.95
```

```r
tapply(Belonging,copeacadstress,var, na.rm=TRUE)
```

```
##    rarely sometimes    always
##  181.2108  164.9978  177.7214
```

```r
cov(Academic[copeacadstress=="rarely"],Belonging[copeacadstress=="rarely"],use="complete.obs")
```

```
## [1] 64.49562
```

```r
cov(Academic[copeacadstress=="sometimes"],Belonging[copeacadstress=="sometimes"],use="complete.obs")
```

```
## [1] 43.08701
```

```r
cov(Academic[copeacadstress=="always"],Belonging[copeacadstress=="always"],use="complete.obs")
```

```
## [1] 81.09166
```

```r
beta_1 <- 64.49562/181.2108
beta_2 <- 43.08701/164.9978
beta_3 <- 81.09166/177.7214
c(beta_1,beta_2,beta_3)
```

```
## [1] 0.3559149 0.2611369 0.4562853
```

```r
summary(lm(Academic~Belonging, data = stem[copeacadstress=="always",]))
```

```
##
## Call:
## lm(formula = Academic ~ Belonging, data = stem[copeacadstress ==
##     "always", ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.503  -5.944   0.270   5.903  31.570
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.33314    2.29248   14.54   <2e-16 ***
## Belonging    0.45222    0.03508   12.89   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.853 on 439 degrees of freedom
##   (28 observations deleted due to missingness)
## Multiple R-squared:  0.2746, Adjusted R-squared:  0.273
## F-statistic: 166.2 on 1 and 439 DF,  p-value: < 2.2e-16
```

```
posthoc <- glht(m4, linfct = mcp(copeacadstress = "Tukey"))
summary(posthoc)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = Academic ~ copeacadstress + Belonging)
##
## Linear Hypotheses:
##                          Estimate Std. Error t value Pr(>|t|)
## sometimes - rarely == 0    9.5971     1.0798   8.888   <2e-16 ***
## always - rarely == 0      18.6885     1.0143  18.426   <2e-16 ***
## always - sometimes == 0    9.0914     0.8479  10.722   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

```
eta_group <- 65881/(65881+21170+86852)
eta_covariate <- 21170/(65881+21170+86852)
c(eta_group,eta_covariate)
```

```
## [1] 0.3788376 0.1217345
```

```
computed_power <- 1.0000000
computed_sample_size <- 149
```

Analysis of results, and comparison between ANOVA and ANCOVA models.

In both the ANOVA and the ANCOVA, we see that our grouping variable, copeacadstress, is significant. The posthocs on the unadjusted means from the ANOVA show that all differences are significant. The posthocs on the means adjusted for Belonging in the ANCOVA also show that all differences in adjusted means are significant.

Aside from the differences in theoretical underpinnings of ANOVA and ANCOVA, we can see that the ANCOVA result not only provides a SS and F test for the grouping variable, copeacadstress, but also for the covariate, Belonging, which is found to be significant. We notice that our ANCOVA result has a lower SSresidual than the ANOVA result, 86852 vs. 108612, implying that this model has less error. Finally, as mentioned in earlier parts, we notice differences in the calculations of the means. The means for the rarely and sometimes groups in the ANOVA model are lower than their ANCOVA counterparts, and the always group has a higher mean in the ANOVA model than in the ANCOVA model.

The ANCOVA model has less error because we are accounting for the covariate, Belonging, in the model. Since Belonging is correlated with our outcome variable, Academic, with a correlation value of 0.53, it is beneficial to use ANCOVA, which will account for the linear relationship between Belonging and Academic when creating the ANOVA based on our grouping variable, copeacadstress. The reason for the differences in the means between the ANOVA and the ANCOVA models is due to the fact that the ANCOVA means are adjusted with respect to Belonging.