HW3, Econ 217

Swati Sharma

2/18/2018

Problem 1

Part a:

```
library("foreign")
library("dplyr")

vetData <- read.csv("https://people.ucsc.edu/~aspearot/Econ_217/veteran.csv")

subData <- subset(vetData, TIME!=0, na.rm=TRUE)
summary(subData)
```

```
##        ID            TIME             Y                trt
##  Min.   :  1   Min.   :  1.0   Min.   :0.0000   standard:69
##  1st Qu.: 35   1st Qu.: 25.0   1st Qu.:1.0000   test    :68
##  Median : 69   Median : 80.0   Median :1.0000
##  Mean   : 69   Mean   :121.6   Mean   :0.9343
##  3rd Qu.:103   3rd Qu.:144.0   3rd Qu.:1.0000
##  Max.   :137   Max.   :999.0   Max.   :1.0000
##       celltype      karno          diagtime          age
##  adeno    :27   Min.   :10.00   Min.   : 1.000   Min.   :34.00
##  large    :27   1st Qu.:40.00   1st Qu.: 3.000   1st Qu.:51.00
##  smallcell:48   Median :60.00   Median : 5.000   Median :62.00
##  squamous :35   Mean   :58.57   Mean   : 8.774   Mean   :58.31
##                 3rd Qu.:75.00   3rd Qu.:11.000   3rd Qu.:66.00
##                 Max.   :99.00   Max.   :87.000   Max.   :81.00
##  priortherapy
##  no :97
##  yes:40
##
##
##
##
```

```
tapply(subData$TIME, subData$trt, summary)
```

```
## $standard
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     3.0    25.0    97.0   115.1   153.0   553.0
##
## $test
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   24.75   52.50  128.21  117.25  999.00
```

```
tapply(subData$celltype, subData$trt, summary)
```

```
## $standard
##     adeno     large smallcell  squamous
##         9        15        30        15
```

```
##
## $test
##     adeno     large smallcell  squamous
##        18        12        18        20
```

```
subData <- subset(subData, TIME<900)
```

The variable "TIME" represents time to death in the instance that someone is recorded with a Y value of 1 (death); generally, it time since the first observation of that individual. It has a min value of 1 with a mean of 121.6 and a max of 999. However, the value of the 75th percentile is 144, which means that there are outlier points in this data. I drop those outliers.

The "Y" variable indicates censoring or death. It takes the value of either 0 or 1. The 25th and 75th percentile values are both 1, while the mean is 0.9343. This means that most people in this study die at some point.

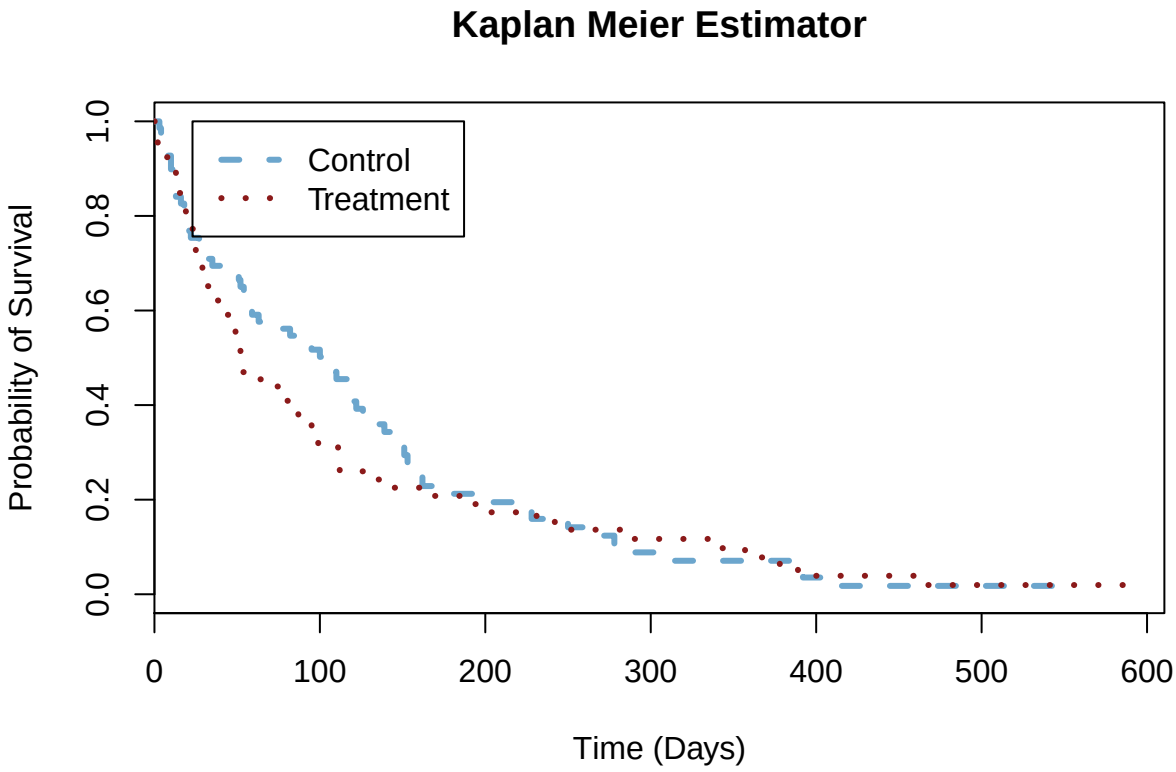
The "trt" variable indicates whether someone is in the treatment or control group. There are 69 individuals in the control group while there are 68 in the treated. This is a good balance, so I don't remove anything here.

Finally, I look at the "celltype" data. It has 27 instances of adeno and large cells, 48 of small cells, and 35 of squamous. After testing for balance in the of these cell types in the treatment and control group, I find that the spread of cell types differs quite a lot between the two groups. For this assignment, I don't change anything but if I were to do it for a different purpose, I'd use the ROSE package to randomly over or under sample and balance the data. Alternatively, I could synthetically generate data for "missing" observations of the minority group.

Part b:

```
library(survival)

fit <- survfit(Surv(TIME,Y)~trt, data = subData)
plot(fit,lty = 2:3, lwd = 3, col = c("skyblue3", "firebrick4"),
     main = "Kaplan Meier Estimator", xlab = "Time (Days)", ylab = "Probability of Survival" )
legend(23, 1, c("Control", "Treatment"), lty = 2:3, lwd = 3, col = c("skyblue3", "firebrick4"))
```



The graph of the Kaplan-Meier Estimates shows that individuals in the treatment group are estimated to survive for a longer period of time than indivduals in the control group. Initially, (up until around day 240) the probability of survival of individuals in the control group is higher, relative to that of the treatment group. After, the probability of survival of the treatment group is higher. By approximately day 575, the entire control population is estimated to be dead but a there is still a probability of survival for someone in the treatment group beyond that point. They are estimated to die by day 600.

Part c:

```
form<-as.formula(Y~trt+celltype+age+priortherapy+offset(log(TIME)))
hazard_glm<-glm(form,family=poisson("log"),data=subData)
summary(hazard_glm)
```

```
##
## Call:
## glm(formula = form, family = poisson("log"), data = subData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0327  -0.4148   0.3452   0.9836   2.8881
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -4.468746   0.562208  -7.949 1.89e-15 ***
## trttest            0.244161   0.194945   1.252 0.210401
## celltypelarge     -0.942356   0.286710  -3.287 0.001013 **
## celltypesmallcell -0.071636   0.261005  -0.274 0.783729
## celltypesquamous  -0.988132   0.271546  -3.639 0.000274 ***
## age                0.001367   0.009345   0.146 0.883717
## priortherapyyes    0.230683   0.202358   1.140 0.254298
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 187.67  on 134  degrees of freedom
## Residual deviance: 161.06  on 128  degrees of freedom
## AIC: 427.06
##
## Number of Fisher Scoring iterations: 6
```

0.1401 represents the change in the expected log of the hazard ratio relative to a one unit change in trttest, holding all other predictors constant. There is an increase of 1.15 - 1 = 0.15 in the hazard of death relative to the control. This means that people in the treatment group (compared to the control group and holding everything else constant) are 15% more likely to likely to die in a certain period of time given that they've survived up until then.

Because the covariates on large, small, and squamous celltypes are all negative, the hazard of having andenocarinoma (omitted group) is highest. Squamous has the lowest (most negative) covariate, so people with squamous cell carconoma have the highest chance of survival. People with the large cell cancer have the second highest chance of survival.

The covariate on age is positive 0.006, so as people in the study grow older, they have a higher hazard, or lower probability of survival. Furthermore, a value of 0.0107 in front of the priortherapyyes predictor indicates that if a person have had prior therapy, their probability of survival declines. This could be because if they've had prior therapy, they've been fighting it longer so their cancer is advanced or that it has been a reoccuring issue for them.

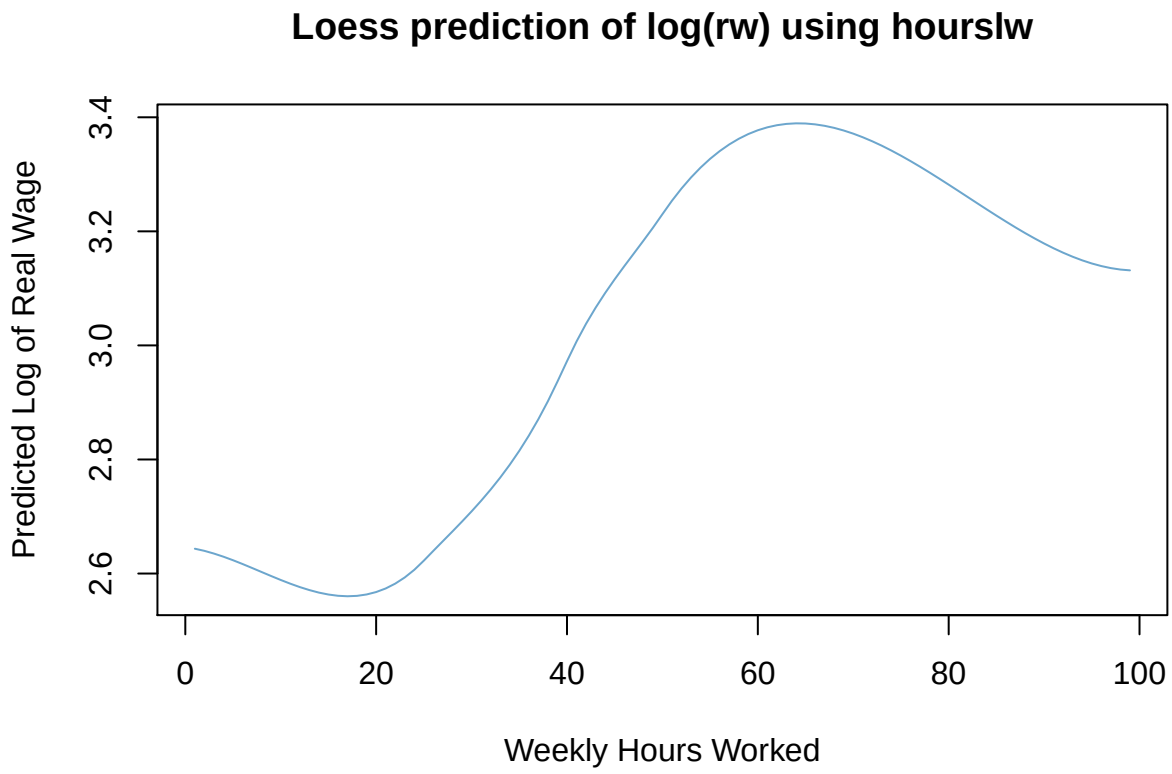
Problem 2

Part a:

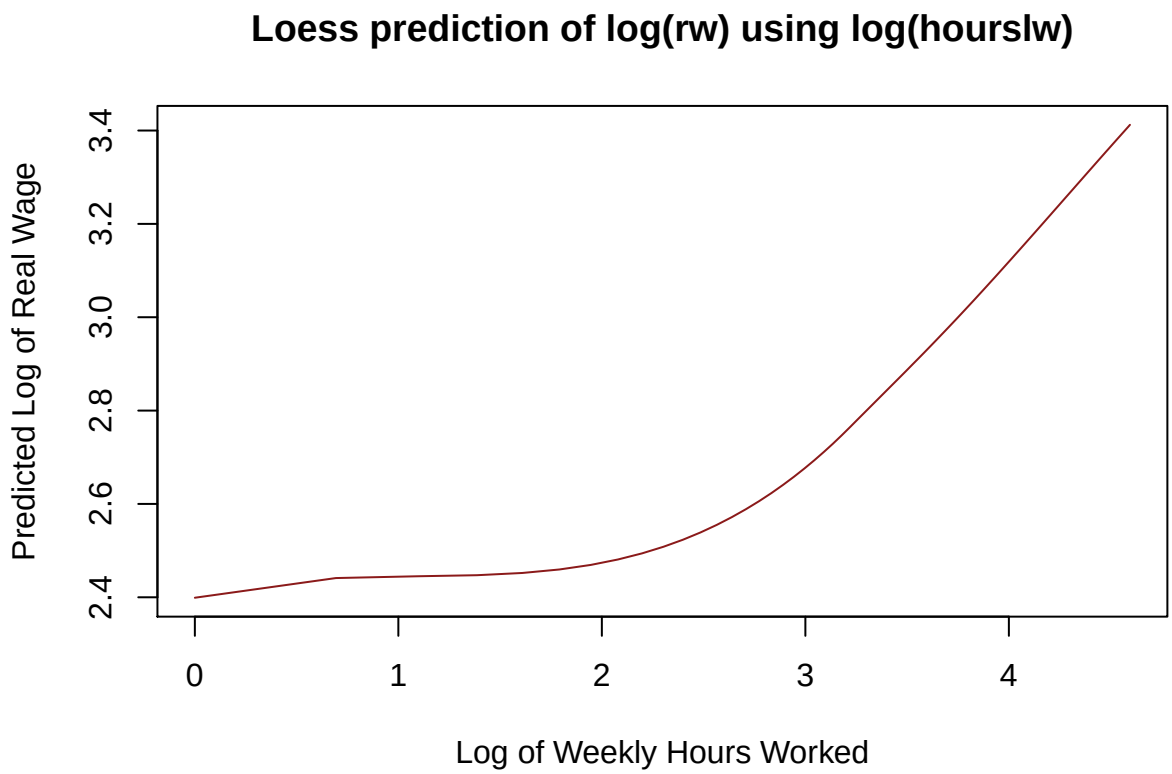
```
orgData <- read.dta("https://people.ucsc.edu/~aspearot/Econ_217/org_example.dta")
subData<- subset(orgData, state=="CA" & year==2013 & (!is.na(orgData[,21])) & (!is.na(orgData[,30])))

hourslw <- seq(min(subData$hourslw):max(subData$hourslw))
orderedData <- as.data.frame(hourslw)

loess1 <- loess(log(rw)~hourslw, subData, span = 0.8,degree = 1)
plot(hourslw,predict(loess1, orderedData),type="l", col="skyblue3",
     main = "Loess prediction of log(rw) using hourslw",
     xlab = "Weekly Hours Worked", ylab = "Predicted Log of Real Wage")
```




```
loess2 <- loess(log(rw)~log(hourslw), subData, span = 1,degree = 1)
plot(log(hourslw),predict(loess2, orderedData),type="l", col="firebrick4",
     main = "Loess prediction of log(rw) using log(hourslw)" ,
     xlab = "Log of Weekly Hours Worked", ylab = "Predicted Log of Real Wage")
```



The graph of the first loess procedure shows that log of real wage is increasing for weekly hours worked between

approximately 20 and 60 hours. Anywhere else, and it appears to be decreasing. The highest earners work just over 60 hours a week, and would earn about exp(3.37)=29 dollars in real wage per hour. The lowest earners work just under 20 hours a week, and earn about exp(2.55)=12.80 dollars in real wage per hour. This could be because anyone who works under 20 hours a week is working-part time. These types of people tend (eg students) tend to work minimum wage jobs so it makes sense that they earn the lowest wage. Perhaps people who work upwards of 20 hours will continue to get raises until they work 60 hours but beyond that, they are either undervalued or represent a different population. It could be true that laborers who work upwards of 60 hours a week provide cheap, blue-collar labor so their wage decreases. As hours worked increases above this interval, the percent change in real wages is negative.

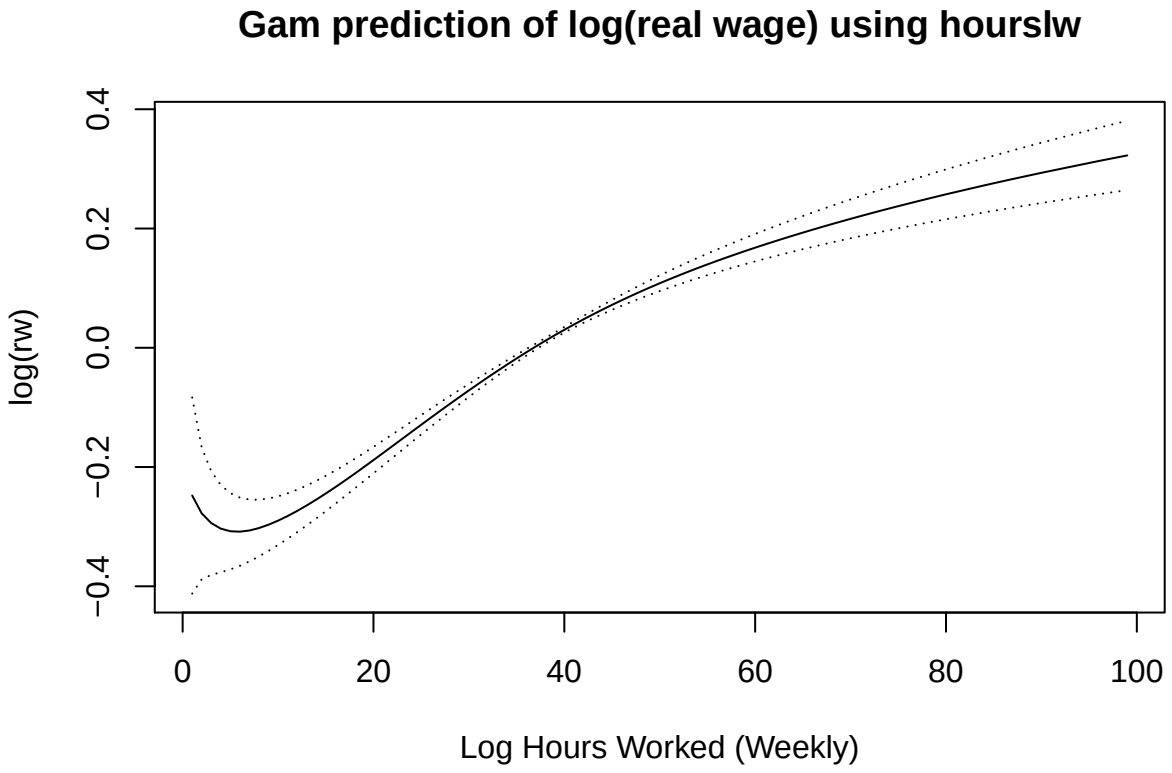
The graph of the second loess procedure demonstrates the estimated relationship between log real wage and log hours worked weekly. The entire graph is increasing, so as we increase weekly hours worked by 1%, we'd expect percent of real wages to increase by the slope of the regression. The slope of the regression increases more sharply (relative to the other points) at around 2.5 log(hourslw). Between 3 and 4, the slope does not seem to change. In that interval, the slope is approximately (3.1-2.7)= 0.4 so if we increased weekly hours worked from 3 to 4 percent, we'd expect to see an increase in real wages by 0.4%.

Part b:

```
library("gam")
gam2 <- gam(log(rw)~s(log(hourslw),2)+educ+age,data=subData)
summary(gam2)
```

```
##
## Call: gam(formula = log(rw) ~ s(log(hourslw), 2) + educ + age, data = subData)
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.80926 -0.33046 -0.03218  0.32495  2.79393
##
## (Dispersion Parameter for gaussian family taken to be 0.2748)
##
##     Null Deviance: 2951.307 on 6631 degrees of freedom
## Residual Deviance: 1820.207 on 6624 degrees of freedom
## AIC: 10263.91
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##                       Df  Sum Sq Mean Sq F value    Pr(>F)
## s(log(hourslw), 2)     1  144.39 144.386  525.44 < 2.2e-16 ***
## educ                   4  791.51 197.878  720.11 < 2.2e-16 ***
## age                    1  130.25 130.251  474.00 < 2.2e-16 ***
## Residuals           6624 1820.21   0.275
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                     Npar Df Npar F     Pr(F)
## (Intercept)
## s(log(hourslw), 2)       1 123.84 < 2.2e-16 ***
## educ
## age
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(gam2,se=TRUE, rug=FALSE,terms="s", xlab = "Log Hours Worked (Weekly)",
     ylab = "log(rw)", main = "Gam prediction of log(real wage) using hourslw")
```



The gam estimation of the relationship between log real wage and log hours worked weekly is positive after approximately 7 units of log(hourslw). Compared to the second graph of the last question, the graph seems smoother. This is probably because we have a wider range of the log(hourslw) compared to before. The interpretation is similar though, for a one percent increase in hours worked weekly, we expect to see a percent change in real wage by the slope of the line. So any percent increase in hours worked weekly past 7% yields an percent increase in real wage. The confidence interval is the tightest around 40 log(hourslw) and widest at 0. This is because we have the most observations for individuals who work 40 hours a week so we can be more confident in our model estimates.

Part c:

```
for(h in 1:20){
  for(i in 1:nrow(subData)){
    datadrop<-subData[i,]
    datakeep<-subData[-i,]
    fit<-loess(log(rw)~log(hourslw),datakeep,family="gaussian",span=(h/20), degree=1)
    dropfit<-predict(fit,datadrop,se=FALSE)
    sqrerr<-(log(datadrop$rw)-as.numeric(dropfit))^2
    ifelse(i*h==1,results<-data.frame(h,i,sqrerr), results<-rbind(results,data.frame(h,i,sqrerr)))
  }
}

spanVal <- results[which.min(results$sqrerr),1]

loess3 <- loess(log(rw)~(hourslw),subData,span = (spanVal/20),degree = 1)
plot((hourslw),predict(loess3, orderedData), type="l", main = "Loess prediction using
     cross-validation", xlab = "Weekly Hours Worked", ylab = "Predicted Log of Real Wage" )
```

I chose a span value of 14/20 because that value minimized the sum of squared errors of the loess estimator. Because span is the optimal bandwidth value, a span which is too small reduces bias but eases noise and one that is too large will oversmooth and increase bias while reducing variance. Using this span value does not means that loess does not overfit the data but also does not have a lot of bias.

The plot for this part is attached to the back. The graph shows that in order to maximize log(real wage) the optimal number of hours to work in a week is 60. The graph is increasing between about 20 to 60 hours a week, which means that as hours worked increase between this interval, so should the percentage of real wage. The graph under 20 and over 60 hours a week is decreasing, which shows that an increase in hours worked in those areas decreases the percentage increase of real wage.

Problem 3

Part a:

```
subData<- subset(orgData, state=="CA" & year==2013)

# simple linear regression of log(rw) on educ and age
linModel <- lm(log(rw)~educ+age,data=subData)
summary(linModel)
```

```
##
## Call:
## lm(formula = log(rw) ~ educ + age, data = subData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.70626 -0.35978 -0.03157  0.34157  2.49244
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.9066343  0.0280951   67.86   <2e-16 ***
## educHS           0.3028645  0.0238486   12.70   <2e-16 ***
## educSome college 0.4313257  0.0229303   18.81   <2e-16 ***
## educCollege      0.8646217  0.0238054   36.32   <2e-16 ***
## educAdvanced     1.1622200  0.0268586   43.27   <2e-16 ***
## age              0.0115285  0.0004829   23.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.538 on 6835 degrees of freedom
##   (6948 observations deleted due to missingness)
## Multiple R-squared:  0.3523, Adjusted R-squared:  0.3518
## F-statistic: 743.4 on 5 and 6835 DF,  p-value: < 2.2e-16
```

```
# creating confidence interval
se <- summary(linModel)$coefficients["educCollege","Std. Error"]
B_college <- summary(linModel)$coefficients["educCollege","Estimate"]
error <- se*qt(.975,linModel$df)
CI <- c(B_college-error,B_college+error)
CI
```

```
## [1] 0.8179558 0.9112876
```

Holding all else constant, a person with a college education earns 86.4% more than someone with a less than high school education. The 95% confidence interval is (0.8179558 0.9112876).

Part b:

```
# Bootstrapping
for(b in 1:1000){
  tempData <- subData[sample(nrow(subData),nrow(subData),replace=TRUE),]
  tempLM <- lm(log(rw)~educ+age,data=tempData)
  tempBeta <- coef(summary(tempLM))[4,1]
  tempSE <- coef(summary(tempLM))[4,2]
  ifelse(b==1, bootCoef <- data.frame(tempBeta, tempSE), bootCoef <-rbind(bootCoef,data.frame(tempBeta,
}

quantile(bootCoef$tempBeta,prob=c(0.025,0.975),CI.level = 0.95,na.rm=TRUE)
```

The 95% confidence interval is smaller than part a. It equals (0.8232835, 0.9071164). The reason it has gotten smaller is because we resampled data so many times that we can be more confident in our estimates.

Part c:

```
orgData <- read.dta("https://people.ucsc.edu/~aspearot/Econ_217/org_example.dta")
```

```
## Warning in read.dta("https://people.ucsc.edu/~aspearot/Econ_217/
## org_example.dta"): value labels ('fipscounty') for 'fipscounty' are missing
```

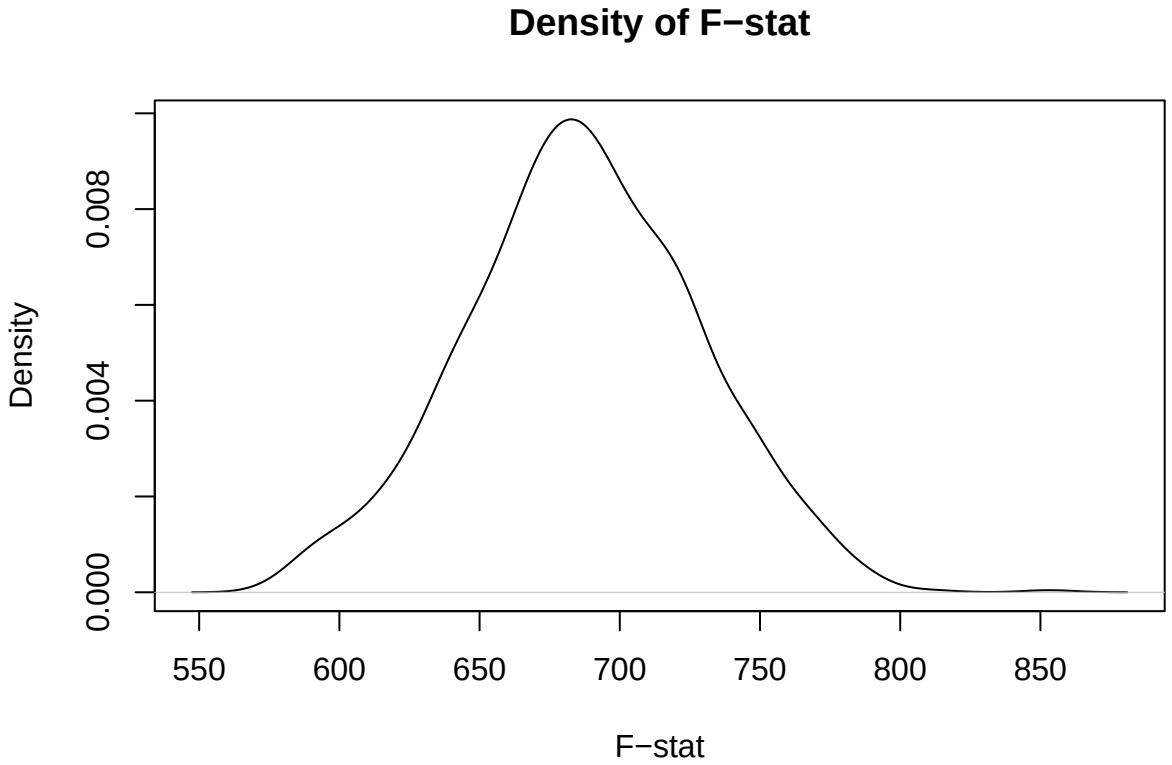
```
## Warning in read.dta("https://people.ucsc.edu/~aspearot/Econ_217/
## org_example.dta"): value labels ('ind12') for 'ind12' are missing
```

```
subData<- subset(orgData, state=="CA" & year==2013 & (!is.na(orgData[,21])) & (!is.na(orgData[,30])))

urLM <- lm(log(rw)~educ+age,data=subData)
subData <- cbind(subData, resid = resid(urLM), fitted = fitted(urLM))
subData$resid <- as.numeric(subData$resid)
subData$fitted <- as.numeric(subData$fitted)

for(b in 1:1000){
   subData$randResid<-sample(subData$resid, nrow(subData),replace=TRUE)
   subData$rw_boot <- subData$randResid + subData$fitted

   rLM <- lm(rw_boot~age,data=subData)
   SSR_r <- sum(resid(rLM)^2)
   SSR_ur <- sum(subData$resid^2)
   F_stat <-  ((SSR_r - SSR_ur)/4)/(SSR_ur/(nrow(subData)-6-1))
  ifelse(b==1, residuals <- data.frame(SSR_r, SSR_ur, F_stat), residuals <-rbind(residuals,data.frame(SS
}
plot(density(residuals$F_stat), main = "Density of F-stat", xlab="F-stat")
```



Because, Pr(F-stat<700) = 0.0085 (approximately) is the highest density value, we can say that most sampled F-statistics have a value of about 700. The F-stat values are pretty symmetrically distributed on either side of the plot. The F-stat has a mean of 685.5864, so resampling demonstrates convergence to the mean because we have the highest probability of drawing that value according to the graph.