Homework 2

Samuel Jones

February 16, 2018

Reproduction

1. Conduct the median (LAD) regressions. (1 point)

```
library(AER)
library(quantreg)
library(xtable)
library(knitr)
data(CPS1988)
# Table 1A Regression
reg1.A <- rq(log(wage) ~ ethnicity + education +experience + I(experience^2),
             data = CPS1988)
table1.A <- xtable(summary(reg1.A)$coef, auto = TRUE)
myvars <- c("Value", "t value")
table1.A <- table1.A[myvars]
colnames(table1.A) <- c("estimates", "t-values")
# Table 2A regression
reg2.A <- rq(log(wage) ~ ethnicity + education + experience + I(experience^2)
             + I(experience^3) + I(experience^4), data = CPS1988)
table2.A <- xtable(summary(reg2.A)$coef, auto = TRUE)
table2.A <- table2.A[myvars]
colnames(table2.A) <- c("estimates", "t-values")
```

Table 1: Table 1.A LAD Reproduction

	estimates	t-values
(Intercept)	4.2792303	206.43447
ethnicityafam	-0.2511647	-16.47547
education	0.0934622	71.77056
experience	0.0762888	68.94154
I(experience^2)	-0.0012739	-50.71619

Table 2: Tabel 2.A LAD Reproduction

	estimates	t-values
(Intercept)	4.0054037	172.08584
ethnicityafam	-0.2456088	-17.37182
education	0.0954809	74.06656
experience	0.1663440	39.23523
I(experience^2)	-0.0085624	-22.66241
I(experience^3)	0.0002011	16.20437
I(experience^4)	-0.0000018	-13.23514

2. Compare your results to those reported in the paper. (1 point)

- The estimates are the same for both tables, but the t-values do not match.

3. Briefly explain what you think the regression summaries mean (give it your best). (1 point for effort)

- Because this is a log-level regression, for every unit of change (times 100) in the regressor there is a % change in the wage. For example in Table 1.A, for every year of additional experience, the individual's wage increases ~ 7%.

Extension

Note that if we want to conduct a regression of y on x; x^2, we use the following R command: $lm(y\ x + I(x^2))$ After attempting to reproduce Tables 1.A and 2.A, write a short report (in the same document) that extends the replication, telling each story of the data. Apply the 20/60/20 rule. Make sure your report includes reproducible R code, either as an appendix or through OSF/Harvard's Dataverse/Github (or some other site of your choice). Here's what I expect to be included in the extension part of the report:

1. Create histograms of the wages, log wages, education, and experience variables, with a summary of what you're seeing. (1 point)

```
# Wages Histogram
png(filename = "HW2Q1wages.png", width = 2000, height = 2000, res = 200)
par(cex = 1.5, mar = c(5, 4, 1, 1), oma = c(0, 0, 0, 0))
hist(CPS1988$wage, prob = FALSE,
     breaks = "FD",
     ylim = c(0, 2000),
     xlim = c(0, 20000),
     xlab = "Weekly Wage",
     main = "Histogram of Weekly Wages",
     col = "grey")
rug(CPS1988$wage, col = rgb(1, 0, 0, .3))
rug(CPS1988$wage[CPS1988$wage > 2400], col = rgb(0, 0, 1, 1))
dev.off()
```

```
## pdf
##   2
```

```
# Log Wages Histogram
png(filename = "HW2Q1logwages.png", width = 2000, height = 2000, res = 200)
par(cex = 1.5, mar = c(5, 4, 1, 1), oma = c(0, 0, 0, 0))
hist(log(CPS1988$wage),
     prob = FALSE,
     breaks = "Scott",
     ylim = c(0, 2000),
     xlab = "Log Weekly Wage",
     main = "Histogram of Log Weekly Wages",
     col = "grey")
rug(log(CPS1988$wage), col = rgb(1, 0, 0, .3))
rug(log(CPS1988$wage[CPS1988$wage > 2400]), col = rgb(0, 0, 1, 1))
dev.off()
```

```
## pdf
##   2
```

```
# Education Histogram
png(filename = "HW2Q1education.png", width = 2000, height = 2000, res = 200)
par(cex = 1.5, mar = c(5, 4, 1, 1), oma = c(0, 0, 0, 0))
```

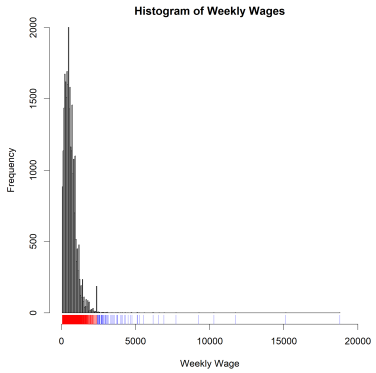


Figure 1: Wages Histogram

```
hist(CPS1988$education, prob = FALSE,
     breaks = "FD",
     xlab = "Years of Education",
     main = "Histogram of Years of Education",
     xlim = c(0, 20),
     col = "grey")
dev.off()
```

```
## pdf
##   2
```

```
# Experience Histogram
png(filename = "HW2Q1experience.png", width = 2000, height = 2000, res = 200)
par(cex = 1.5, mar = c(5, 4, 1, 1), oma = c(0, 0, 0, 0))
hist(CPS1988$experience, prob = FALSE,
     breaks = "FD",
     xlab = "Years of Potential Experience",
     ylim = c(0, 1000),
     xlim = c(-10, 70),
     main = "Histogram of Years of Potential Experience",
     col = "grey")
axis(side = 1, at = seq(-10, 70, by = 10))
dev.off()
```

```
## pdf
##   2
```

- The majority of the weekly wages are below $2,000. There are a few outliers above $5,000.
- The log wages histogram look much more like a normal distribution.
- There's a huge spike corresponding to high school graduation. Those with more education than high school greatly outnumbered those with less than a high school education.
- The data peaks somewhere between ten and twenty years. The data also shows people with negative experience. This is okay because the measure is "potential experience" with a formula that makes it possible for people to have negative experience.

2. Create scatterplots of wages and log wages against education and experience, with fitted linear or median regression lines, again with a summary. (1 point)

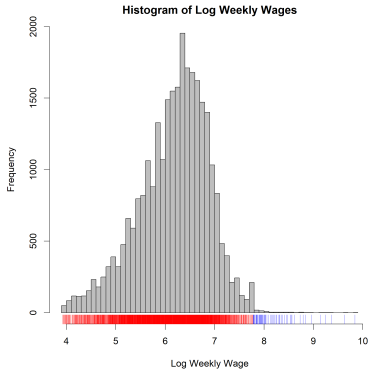


Figure 2: Log Wages Histogram

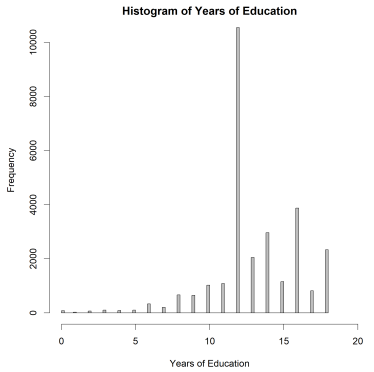


Figure 3: Education Histogram

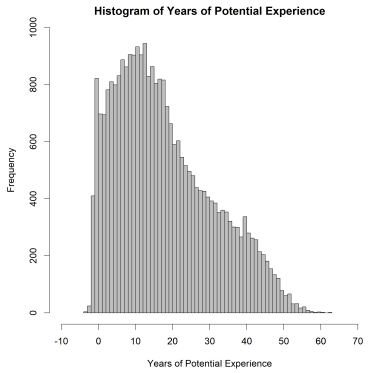


Figure 4: Experience Histogram

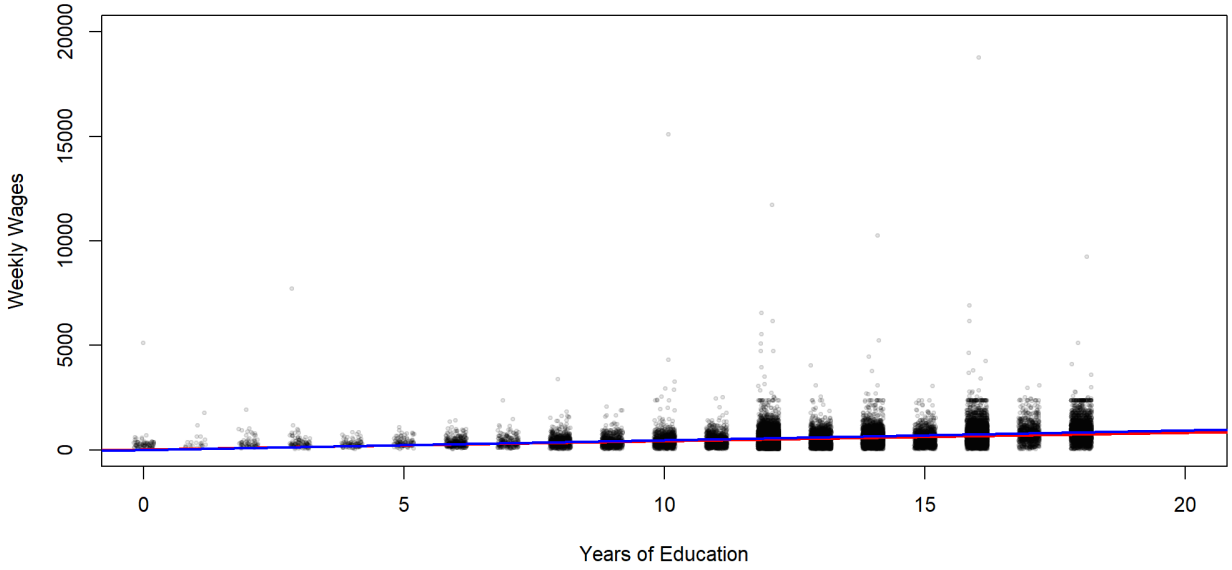
```r
# Wage vs Education
rq1 <- rq(wage ~ education, data = CPS1988)
lm1 <- lm(wage ~ education, data = CPS1988)
rq2 <- rq(log(wage) ~ education, data = CPS1988)
lm2 <- lm(log(wage) ~ education, data = CPS1988)
png(filename = "HW2Q2ScatterEd.png", width = 2000, height = 2000, res = 200)
par(cex = 1.5, mar = c(5, 4, 2, 1), mfrow = c(2,1))
plot(jitter(CPS1988$education),
     jitter(CPS1988$wage),
     ylab = "Weekly Wages",
     xlab = "Years of Education",
     pch = 19,
     cex = .5,
     col = rgb(0, 0, 0, .1),
     xlim = c(0, 20),
     ylim = c(0, 20000))
abline(rq1, col = "red", lwd = 2)
abline(lm1, col = "blue", lwd = 2)
plot(jitter(CPS1988$education),
     jitter(log(CPS1988$wage)),
     ylab = "Log Weekly Wages",
     xlab = "Years of Education",
     cex = .5,
     pch = 19,
     col = rgb(0, 0, 0, .1),
     xlim = c(0, 20))
abline(rq2, col = "red", lwd = 2)
abline(lm2, col = "blue", lwd = 2)
dev.off()
```

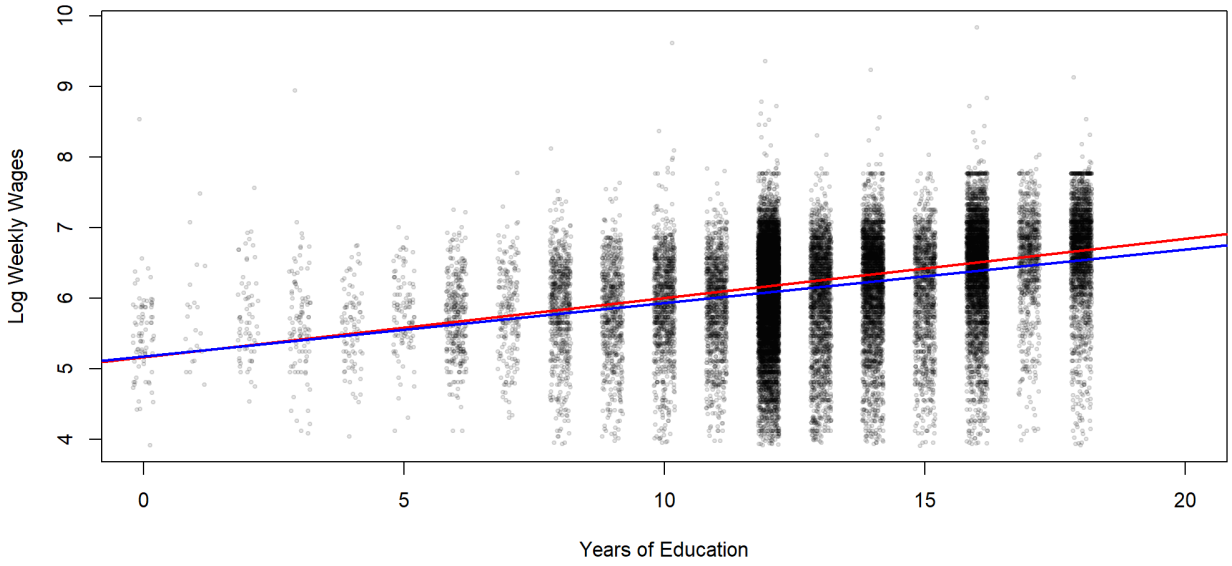
```
## pdf
##   2
```

```r
# Wage vs Experience
rq1 <- rq(wage ~ experience, data = CPS1988)
lm1 <- lm(wage ~ experience, data = CPS1988)
rq2 <- rq(log(wage) ~ experience, data = CPS1988)
lm2 <- lm(log(wage) ~ experience, data = CPS1988)
png(filename = "HW2Q2ScatterExp.png", width = 2000, height = 2000, res = 200)
par(cex = 1.5, mar = c(5, 4, 2, 1), mfrow = c(2,1))
plot(jitter(CPS1988$experience),
     jitter(CPS1988$wage),
     ylab = "Weekly Wages",
     xlab = "Potential Years of Experience",
     pch = 19,
     cex = .5,
     col = rgb(0, 0, 0, .1),
     ylim = c(0, 20000))
abline(rq1, col = "red", lwd = 2)
abline(lm1, col = "blue", lwd = 2)
plot(jitter(CPS1988$experience),
     jitter(log(CPS1988$wage)),
     ylab = "Log Weekly Wages",
     xlab = "Potential Years of Experience",
```

```
     cex = .5,
     pch = 19,
     col = rgb(0, 0, 0, .1))
abline(rq2, col = "red", lwd = 2)
abline(lm2, col = "blue", lwd = 2)
dev.off()
```

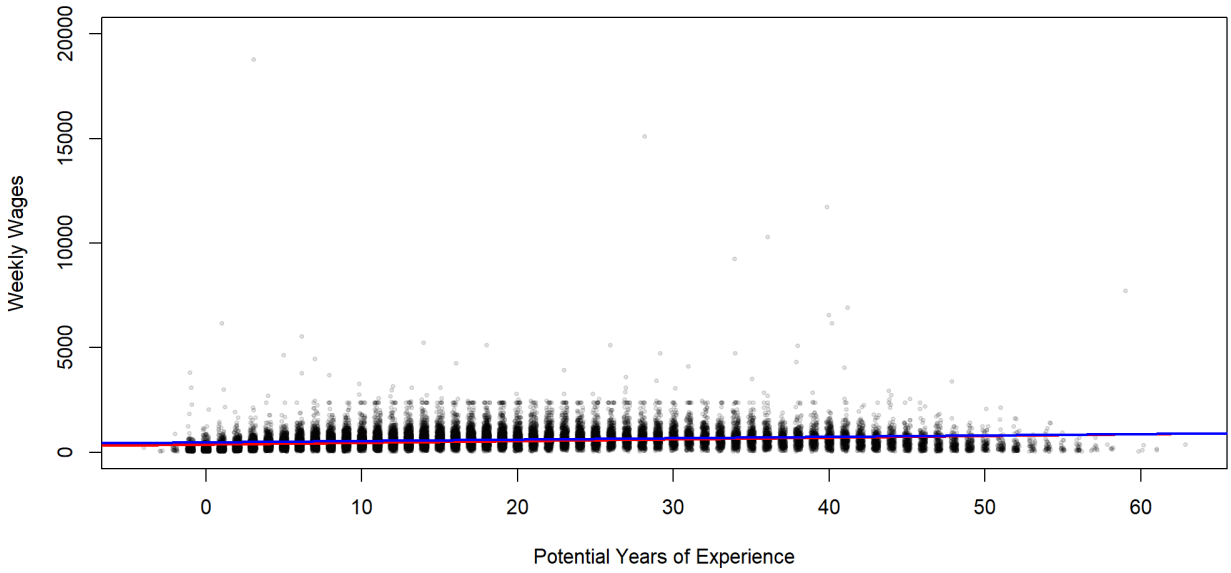
```
## pdf
##   2
```

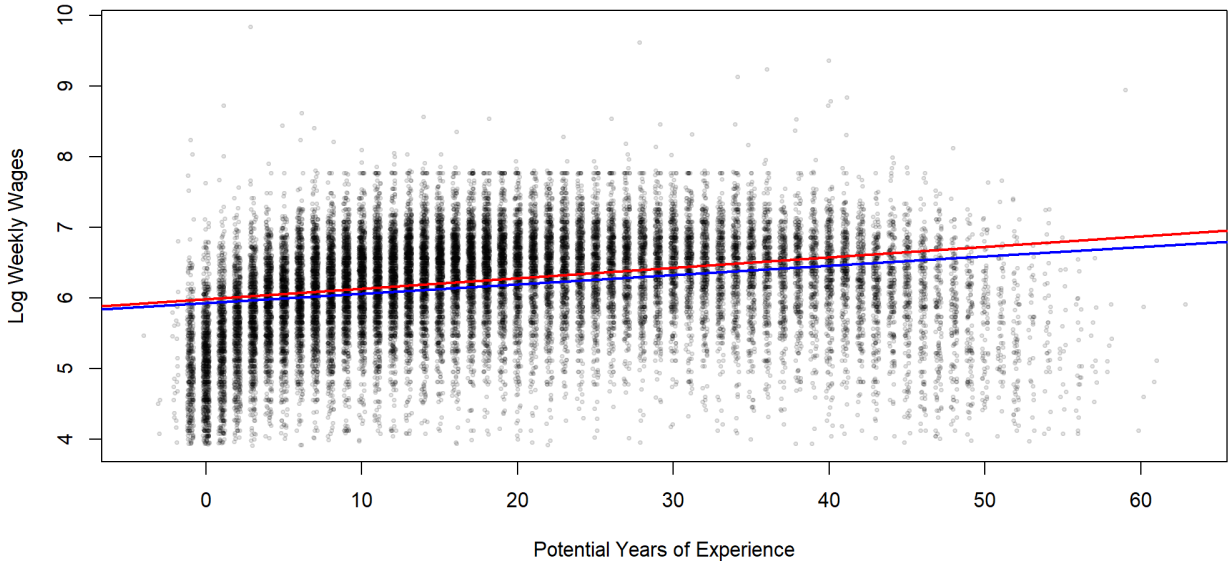




{width = 30%}

- The plots seem to make a good case for increased wages with increasing years of education. The second plot shows that there is a lot of variation as wages increase as well.





{width = 30%}

- There apears to be acurvilinear relationship (an upside down U) to the data. This relationshipo is only visible in the plot using log wages.

3. Compare wages and log wages to the normal distribution, with a summary of what you think is going on. (1 point)

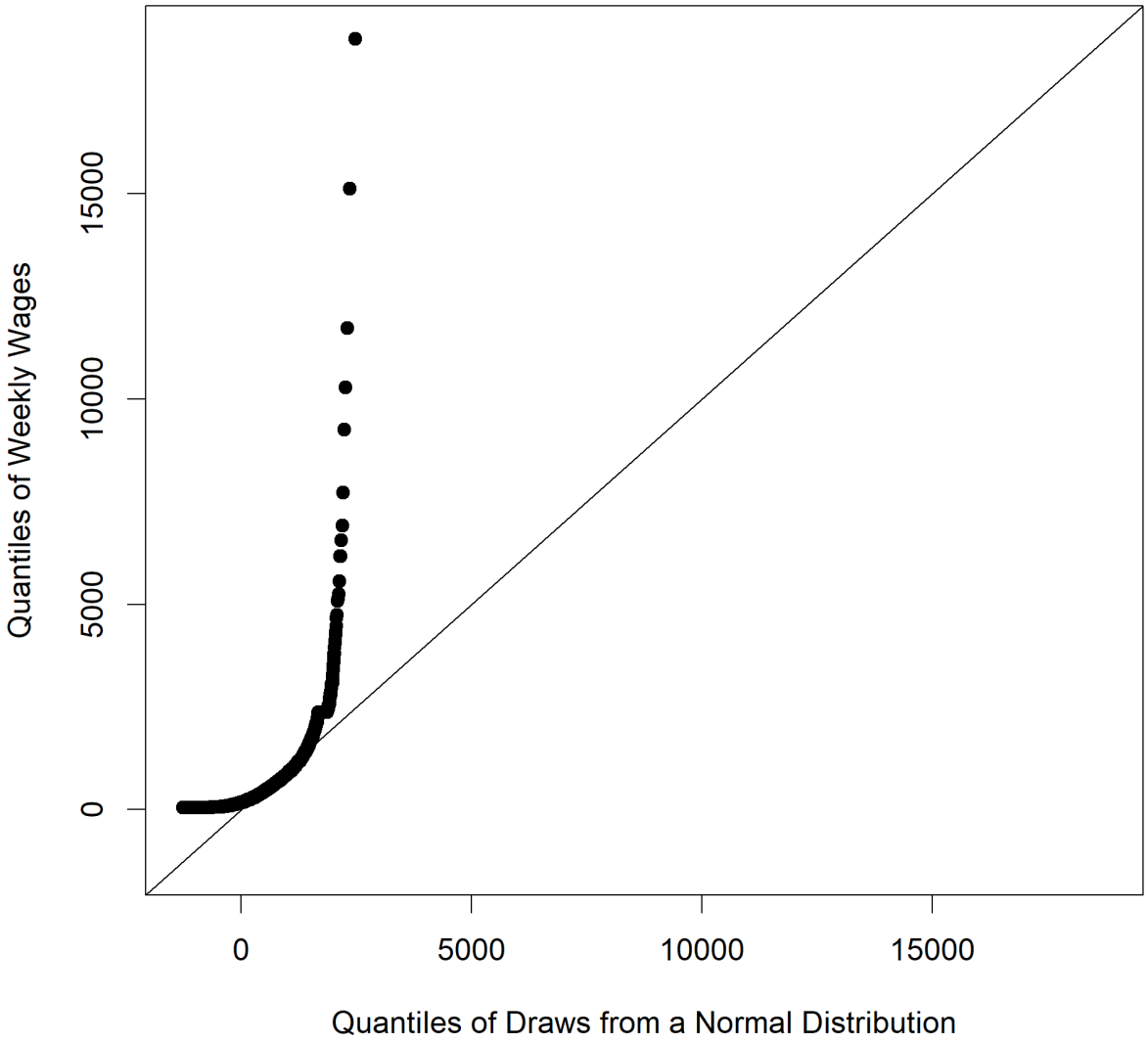
```
# Wage QQ
k <- 1:length(CPS1988$wage)
p <- (k - 0.5)/length(CPS1988$wage)
norm.q <- qnorm(p, mean(CPS1988$wage), sd(CPS1988$wage))
wages.q <- quantile(CPS1988$wage, probs = p, type = 5)
png(filename = "HW2Q3wageqq.png", width = 2000, height = 2000, res = 200)
par(cex = 1.5)
min <- min(floor(min(wages.q, norm.q)))
```

```r
max <- max(ceiling(max(wages.q, norm.q)))
plot(norm.q, wages.q,
     xlim = c(min, max),
     ylim = c(min, max),
     ylab = "Quantiles of Weekly Wages",
     xlab = "Quantiles of Draws from a Normal Distribution",
     pch = 19)
abline(0, 1)
dev.off()
```

```
## pdf
##   2
```

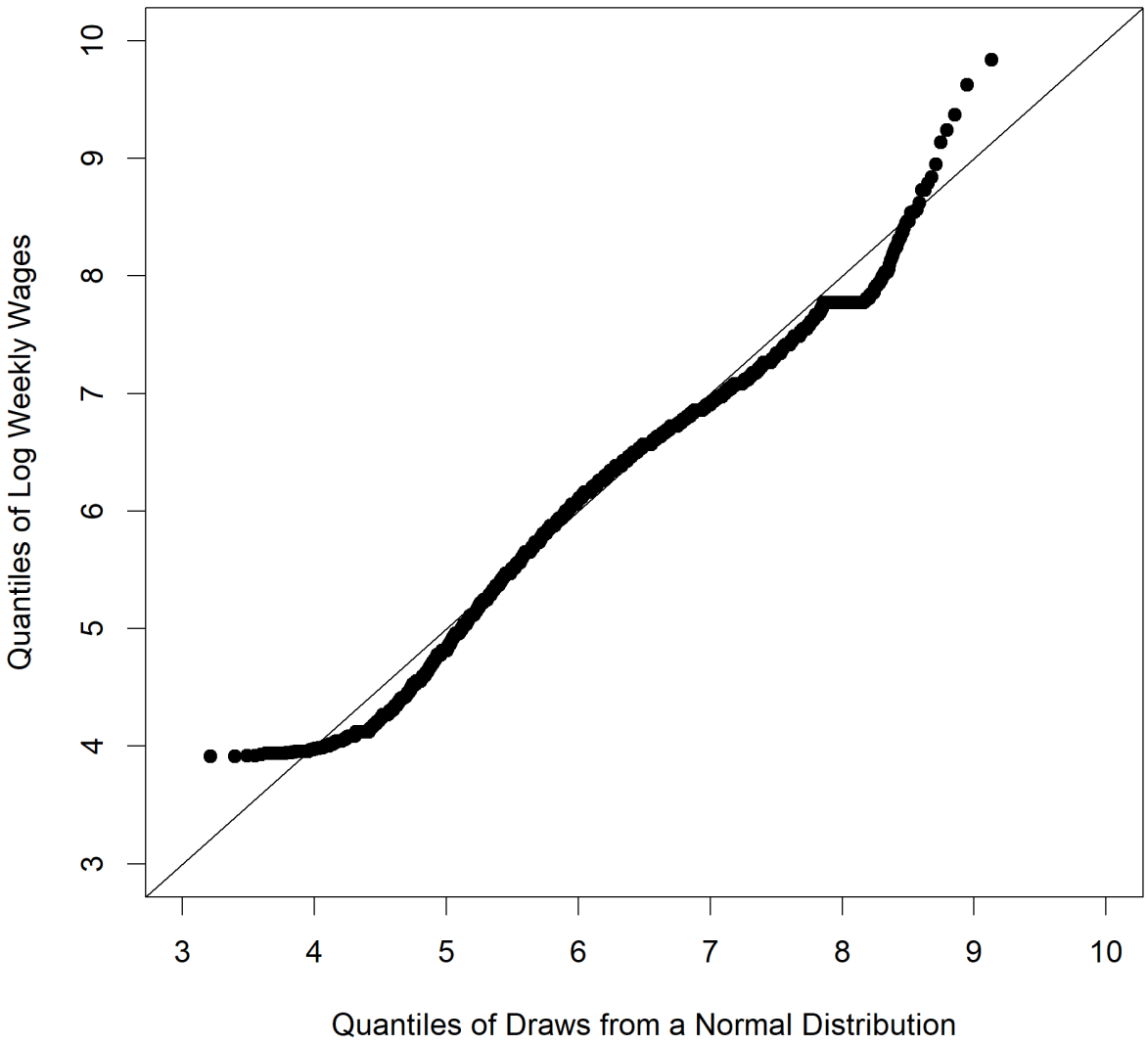
```r
#Wage Log QQ
k <- 1:length(CPS1988$wage)
p <- (k - 0.5)/length(CPS1988$wage)
norm.q <- qnorm(p, mean(log(CPS1988$wage)), sd(log(CPS1988$wage)))
wages.q <- quantile(log(CPS1988$wage), probs = p, type = 5)
png(filename = "HW2Q3logwageqq.png", width = 2000, height = 2000, res = 200)
par(cex = 1.5)
min <- min(floor(min(wages.q, norm.q)))
max <- max(ceiling(max(wages.q, norm.q)))
plot(norm.q, wages.q,
     xlim = c(min, max),
     ylim = c(min, max),
     ylab = "Quantiles of Log Weekly Wages",
     xlab = "Quantiles of Draws from a Normal Distribution",
     pch = 19)
abline(0, 1)
dev.off()
```

```
## pdf
##   2
```



{width = 30%}

- The QQ plot shows a very skewed distribution. The lowest quantile is compressed.



{width = 30%}

- This QQ plot is much closer to a normal distribution. The lowest and highest quantiles are both not normal. In the middle of the distribution wages are slightly higher than a normal distribution would predict.

4. Plot the regression residuals for the regressions in Tables 1.A and 2.A, using ordinary linear regression rather than median regression. Are there outliers? If so, what type are they? Would median regression or ordinary linear regression be more appropriate for these data? (1 point)

```
reg1.A.lm <- lm(log(wage) ~ ethnicity + education + experience + I(experience^2),
                data = CPS1988)
reg2.A.lm <- lm(log(wage) ~ ethnicity + education + experience + I(experience^2)
                + I(experience^3) + I(experience^4), data = CPS1988)
library(car)
png(filename = "HW2Q4qqresiduals.png", width = 2000, height = 2000, res = 200)
```

```r
par(mfrow = c(1, 2), cex = 1.3, mar = c(5, 4, 2, 1))
qqPlot(reg1.A.lm,
       main = "Table 1.A Regression Residuals",
       id.n = 3,
       pch = 19,
       cex = .75,
       col = rgb(0, 0, 0, .1),
       ylab = "Jackknife Residuals")
```

```
##  8345 15959 15387
## 28153 28154 28155
```

```r
abline(0, 1)
qqPlot(reg2.A.lm,
       main = "Table 2.A Regression Residuals",
       id.n = 3,
       pch = 19,
       cex = .75,
       col = rgb(0, 0, 0, .1),
       ylab = "Jackknife Residuals")
```

```
##  8345 15959 15387
## 28153 28154 28155
```

```r
abline(0, 1)
dev.off()
```

```
## pdf
##   2
```

```r
# Influence Index Plots
png(filename = "HW2Q4Influence1A.png", width = 2000, height = 2000, res = 200)
par(cex = 1.3, mar = c(5, 4, 2, 1))
influenceIndexPlot(reg1.A.lm, id.n = 10)
dev.off()
```

```
## pdf
##   2
```

```r
#Median Cook's D
median(cooks.distance(reg1.A.lm))
```

```
## [1] 8.917623e-06
```

```r
max(cooks.distance(reg1.A.lm))
```

```
## [1] 0.02122913
```

```r
max(cooks.distance(reg1.A.lm))/median(cooks.distance(reg1.A.lm))
```

```
## [1] 2380.582
```

```r
influence(reg1.A.lm)$hat[15387]/median(influence(reg1.A.lm)$hat)
```

```
##    15387
## 16.80369
```

```r
cooks.distance(reg1.A.lm)[15387]/median(cooks.distance(reg1.A.lm))
```

```
##    15387
```

```
## 2380.582
```

```
rstudent(reg1.A.lm)[15387]
```

```
##   15387
## 7.52034
```

```
CPS1988[15387, ]
```

```
##          wage education experience ethnicity smsa region parttime
## 15387 7716.05         3         59      cauc  yes  south      yes
```

```
CPS1988[17228, ]
```

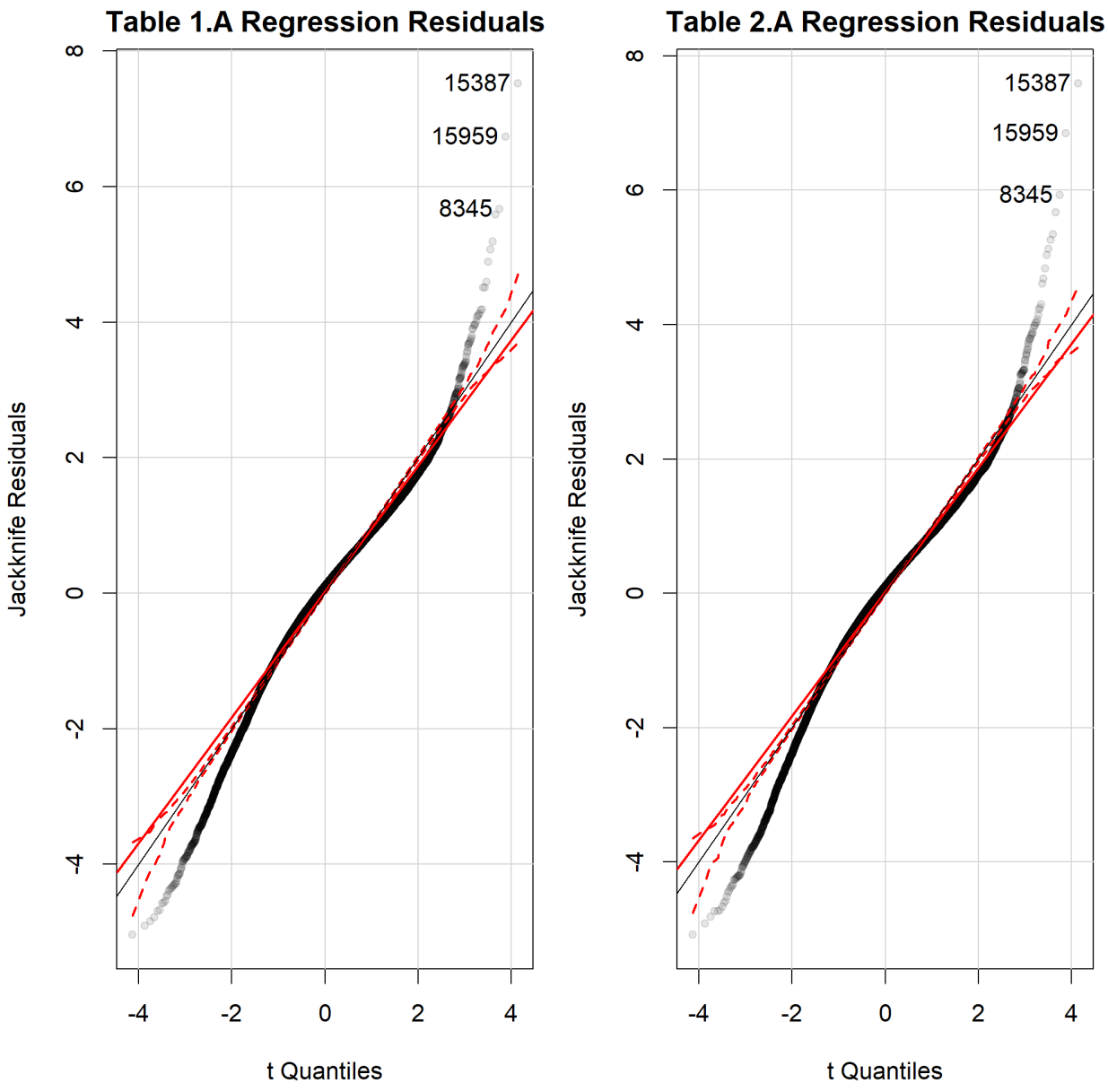
```
##         wage education experience ethnicity smsa region parttime
## 17228 370.37         0         63      cauc  yes  south       no
```

```
# Remove outliers sensiticity analysis
reg1.A.lm.drop <- lm(log(wage) ~ ethnicity + education + experience
                     + I(experience^2), data = CPS1988[-c(15387, 17228),])
reg2.A.lm.drop <- lm(log(wage) ~ ethnicity + education + experience
                     + I(experience^2) + I(experience^3) + I(experience^4),
                     data = CPS1988[-c(15387, 17228),])
coefficients(reg1.A.lm)/coefficients(reg1.A.lm.drop)
```

```
##     (Intercept)   ethnicityafam       education      experience
##       1.0009284       1.0019153       0.9980907       0.9960944
## I(experience^2)
##       0.9940984
```

```
coefficients(reg2.A.lm)/coefficients(reg2.A.lm.drop)
```

```
##     (Intercept)   ethnicityafam       education      experience
##       1.0010048       1.0051805       0.9997754       0.9839811
## I(experience^2) I(experience^3) I(experience^4)
##       0.9617555       0.9359569       0.9144426
```



{width = 30%}

- There are many observations with jackknife residuals greater than +/- 4. There are outliers. 2 of them appear to be very experienced persons with very high and very low wages. Median regression appears to be better for these data because a few cases with high leverage are "throwing off" the rest of the data.

5. Using median or ordinary linear regression, evaluate the backcasting quality of the Mincer type model (Table 1.A). Use 5-fold cross-validation to test whether this model is too complex. (1 point)

```
png(filename = "HW2Q5backcast.png", width = 2000, height = 2000, res = 200)
par(cex = 1.3, mar = c(5, 4, 2, 1))
plot(predict(reg1.A.lm), log(CPS1988$wage),
     xlab = "Predicted Log Wages",
     ylab = "Actual Log Wages",
     xlim = c(3, 10),
     ylim = c(3, 10),
```

```
     pch = 19,
     col = rgb(0, 0, 0, .1),
     cex = 0.5)
abline(0, 1)
dev.off()
```

```
## pdf
##   2
```

```
# Cross Validation
cv.fun <- function(K = 5){
  folds <- K
  ## generate vector of fold numbers
  fold.num <- rep(1:folds,
                  length.out = length(CPS1988$wage))
  fold.ran <- sample(fold.num)
  # Create data matrix
  CPS1988$exp.sq <- CPS1988$experience^2
  CPS1988$logwage <- log(CPS1988$wage)
  data1 <- CPS1988
  rmse.simple <- c()
  rmse.complex <- c()
  for(i in 1:folds){ # loop through the different folds
    # training data are those that are not in the current fold
    train <- data1[fold.ran != i, ]
    # test data are those that are in the current fold
    test <- data1[fold.ran == i, ]
    # Estimate LAD models on training data
    train1 <- rq(logwage ~ education + ethnicity + experience,
                 data = train)
    train2 <- rq(logwage ~ education + ethnicity
                 + experience + exp.sq, data = train)
    # Make predictions on test data
    test1 <- predict(train1, newdata = test)
    test2 <- predict(train2, newdata = test)
    # Calculate rMSE
    rmse.simple[i] <- sqrt(sum((test$logwage - test1)^2)/length(test$logwage))
    rmse.complex[i] <- sqrt(sum((test$logwage - test2)^2)/length(test$logwage))
  }
  return(list(rmse.simple, rmse.complex))
}
xval <- replicate(10, cv.fun())
rmse.simple <- sapply(xval[1, ], mean)
rmse.complex <- sapply(xval[2, ], mean)

mean(rmse.simple)
```

```
## [1] 0.6381292
```

```
mean(rmse.complex)
```

```
## [1] 0.5873354
```

```
# Quadratic vs. Quartic
cv.fun <- function(K = 5){
  folds <- K
```

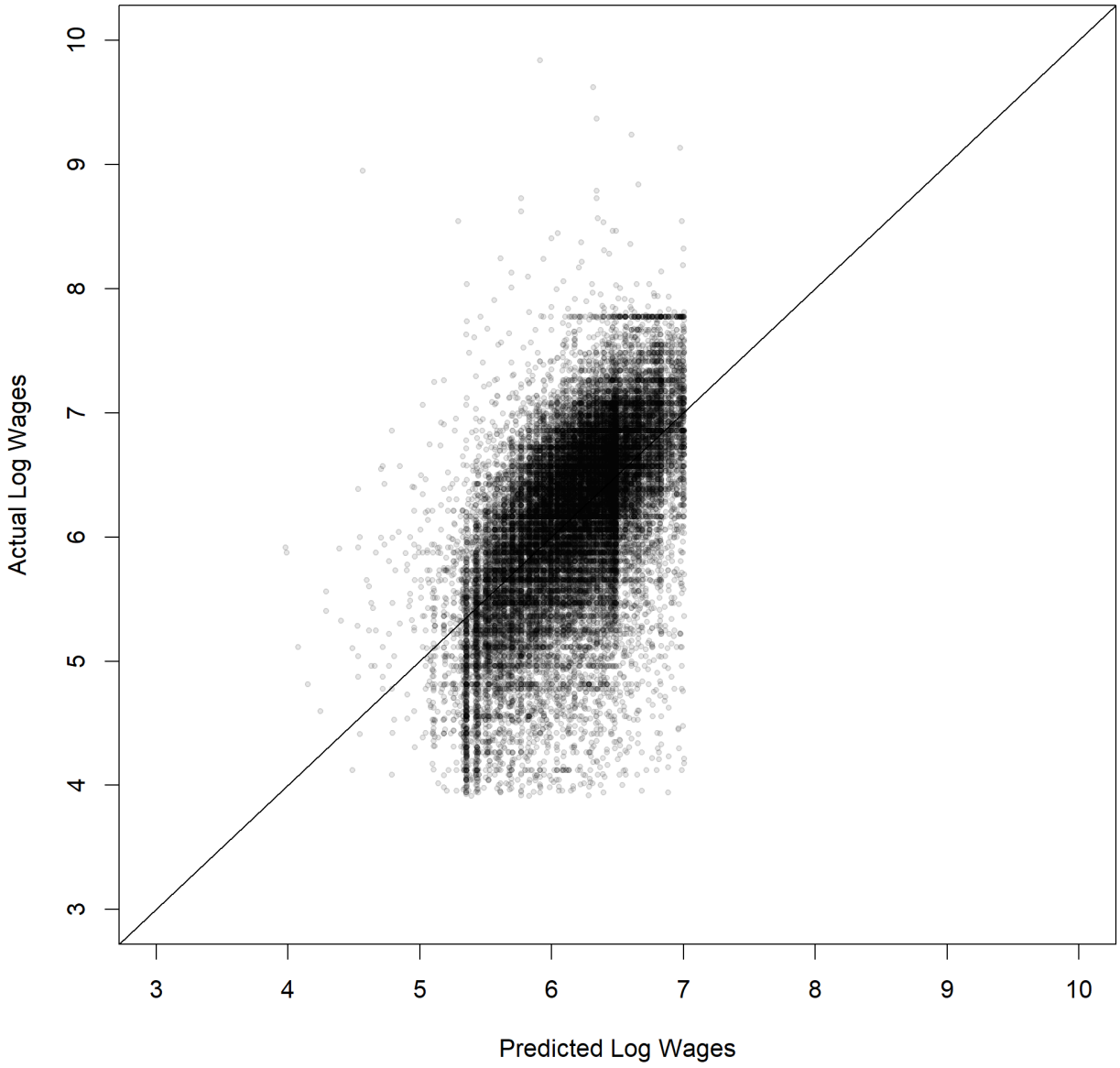
```
  ## generate vector of fold numbers
  fold.num <- rep(1:folds,
                  length.out = length(CPS1988$wage))
  fold.ran <- sample(fold.num)
  # Create data matrix
  CPS1988$exp.sq <- CPS1988$experience^2
  CPS1988$exp.cube <- CPS1988$experience^3
  CPS1988$exp.quad <- CPS1988$experience^4
  CPS1988$logwage <- log(CPS1988$wage)
  data1 <- CPS1988
  rmse.simple <- c()
  rmse.complex <- c()
  for(i in 1:folds){ # loop through the different folds
    # training data are those that are not in the current fold
    train <- data1[fold.ran != i, ]
    # test data are those that are in the current fold
    test <- data1[fold.ran == i, ]
    # Estimate LAD models on training data
    train1 <- rq(logwage ~ education + ethnicity
                 + experience + exp.sq, data = train)
    train2 <- rq(logwage ~ education + ethnicity
                 + experience + exp.sq
                 + exp.cube + exp.quad, data = train)
    # Make predictions on test data
    test1 <- predict(train1, newdata = test)
    test2 <- predict(train2, newdata = test)
    # Calculate rMSE
    rmse.simple[i] <- sqrt(sum((test$logwage - test1)^2)/length(test$logwage))
    rmse.complex[i] <- sqrt(sum((test$logwage - test2)^2)/length(test$logwage))
  }
  return(list(rmse.simple, rmse.complex))
}
xval <- replicate(10, cv.fun())
rmse.simple <- sapply(xval[1, ], mean)
rmse.complex <- sapply(xval[2, ], mean)

mean(rmse.simple)
```

```
## [1] 0.5873295
```

```
mean(rmse.complex)
```

```
## [1] 0.5808887
```



{width = 30%}

- There are modest gains when using the more complicated model for cross-validation. The quadratic and quartic models are not very different. So, lacking an theoretical justification, its probably best to stick with the simpler model.

6. Discuss the statistical inference story. Explain why using confidence intervals would be valid or invalid for these data. (1 point)

```
CPS1988$logwage <- log(CPS1988$wage)
CPS1988$exp.sq <- CPS1988$experience^2
# Estimate quadratic OLS
quad.lm <- lm(logwage ~ education + ethnicity
              + experience + exp.sq, data = CPS1988)
summary(quad.lm)
```

```
##
## Call:
## lm(formula = logwage ~ education + ethnicity + experience + exp.sq,
##     data = CPS1988)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9428 -0.3162  0.0580  0.3756  4.3830
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.321e+00  1.917e-02  225.38   <2e-16 ***
## education      8.567e-02  1.272e-03   67.34   <2e-16 ***
## ethnicityafam -2.434e-01  1.292e-02  -18.84   <2e-16 ***
## experience     7.747e-02  8.800e-04   88.03   <2e-16 ***
## exp.sq        -1.316e-03  1.899e-05  -69.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5839 on 28150 degrees of freedom
## Multiple R-squared:  0.3347, Adjusted R-squared:  0.3346
## F-statistic:  3541 on 4 and 28150 DF,  p-value: < 2.2e-16
```

```
vcov1 <- vcov(quad.lm)
vars <- diag(vcov1)
beta.hat <- coefficients(quad.lm)
lower <- beta.hat - 1.96*sqrt(vars)
upper <- beta.hat + 1.96*sqrt(vars)
lower
```

```
##   (Intercept)     education ethnicityafam    experience        exp.sq
##   4.283813536   0.083179333  -0.268683820   0.075748339  -0.001353282
```

```
upper
```

```
##   (Intercept)     education ethnicityafam    experience        exp.sq
##   4.358976456   0.088166304  -0.218044772   0.079198122  -0.001278851
```

- The data were collected using sampling techniques from the US as long as the subjects fell within some fairly specific characteristics. There appears to also be measurement error and bias, some intentional and someunintentional. The data are never claimed to be a random sample. The data are also restricted to a single year. There could be factors not included in the data that drive or influence the data.

7. Discuss the causality story. To what degree do you think we can make causal claims about the regression models used? Explain your reasoning. (1 point)

- I don't think we have a sufficient foundation for claims on causality. My advisor has often said that he doesn't feel justified making claims for causality if he doesn't have experimental data to back up the claim. This seems like a very prudent approach to data analysis in hte sciences. Specifically, the data did not result from random sampling, there was no blinding and we don't know if there were non-responses (self selection bias) or atttrition factors in this study. It's also possible there exists omitted variable bias and there are probably fixed effects that are also biasing the data.