36-402 Homework 4

James “Morgan” Hawkins

Friday Feb 17, 2023

```r
ucb = data.frame(UCBAdmissions)
library(dplyr)
```

Problem 1

Problem 1 (a)

```r
depA = UCBAdmissions[,,"A"]
pA = sum(depA) / sum(UCBAdmissions)
depB = UCBAdmissions[,,"B"]
pB = sum(depB) / sum(UCBAdmissions)
depC = UCBAdmissions[,,"C"]
pC = sum(depC) / sum(UCBAdmissions)
depD = UCBAdmissions[,,"D"]
pD = sum(depD) / sum(UCBAdmissions)
depE = UCBAdmissions[,,"E"]
pE = sum(depE) / sum(UCBAdmissions)
depF = UCBAdmissions[,,"F"]
pF = sum(depF) / sum(UCBAdmissions)

p_maleA = depA[1]/(depA[1] + depA[2])
p_femaleA = depA[3]/(depA[3] + depA[4])
p_maleB = depB[1]/(depB[1] + depB[2])
p_femaleB = depB[3]/(depB[3] + depB[4])
p_maleC = depC[1]/(depC[1] + depC[2])
p_femaleC = depC[3]/(depC[3] + depC[4])
p_maleD = depD[1]/(depD[1] + depD[2])
p_femaleD = depD[3]/(depD[3] + depD[4])
p_maleE = depE[1]/(depE[1] + depE[2])
p_femaleE = depE[3]/(depE[3] + depE[4])
p_maleF = depF[1]/(depF[1] + depF[2])
p_femaleF = depF[3]/(depF[3] + depF[4])
```

```
ate_male = (pA * p_maleA) + (pB * p_maleB) + (pC * p_maleC) + (pD * p_maleD) + (pE * p_maleE) + (pF * p
ate_female = (pA * p_femaleA) + (pB * p_femaleB) + (pC * p_femaleC) + (pD * p_femaleD) + (pE * p_female
ate_male
```

```
## [1] 0.3873186
```

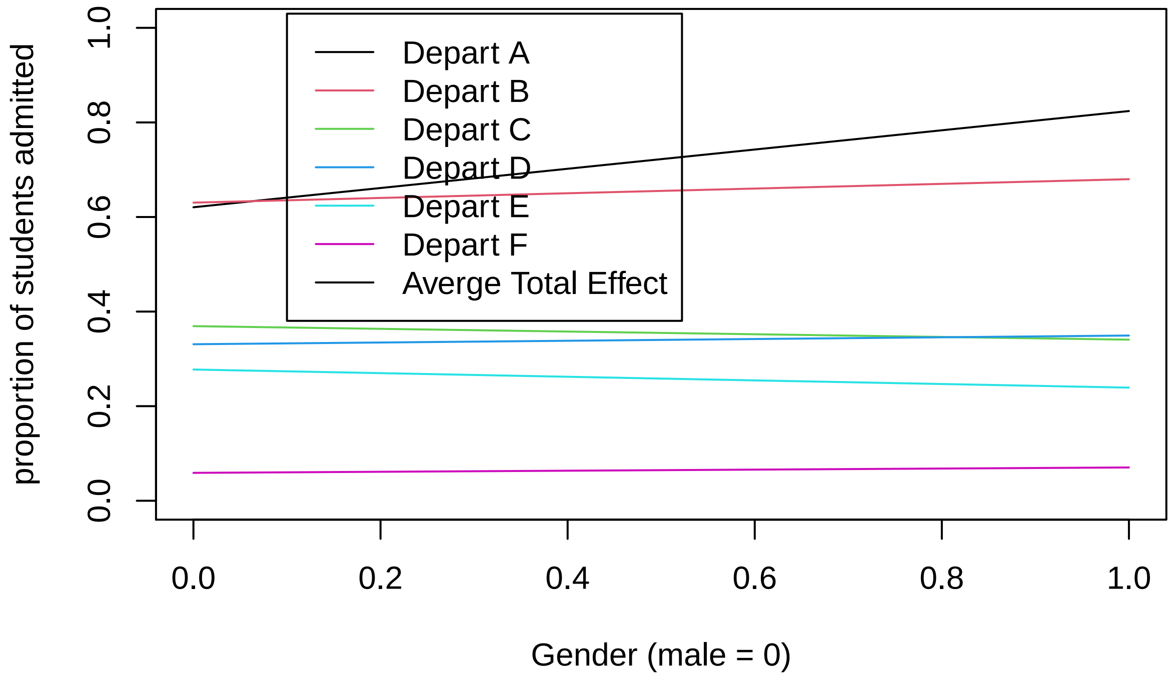
```
ate_female
```

```
## [1] 0.4299554
```

The adjusted treatment effect is .387 and .430 for males and females respectively.Adjusting for department did make a difference because the treatment effects were .445 and .3035 for males and females previously so the group with a higher admission rate switched from males to females.

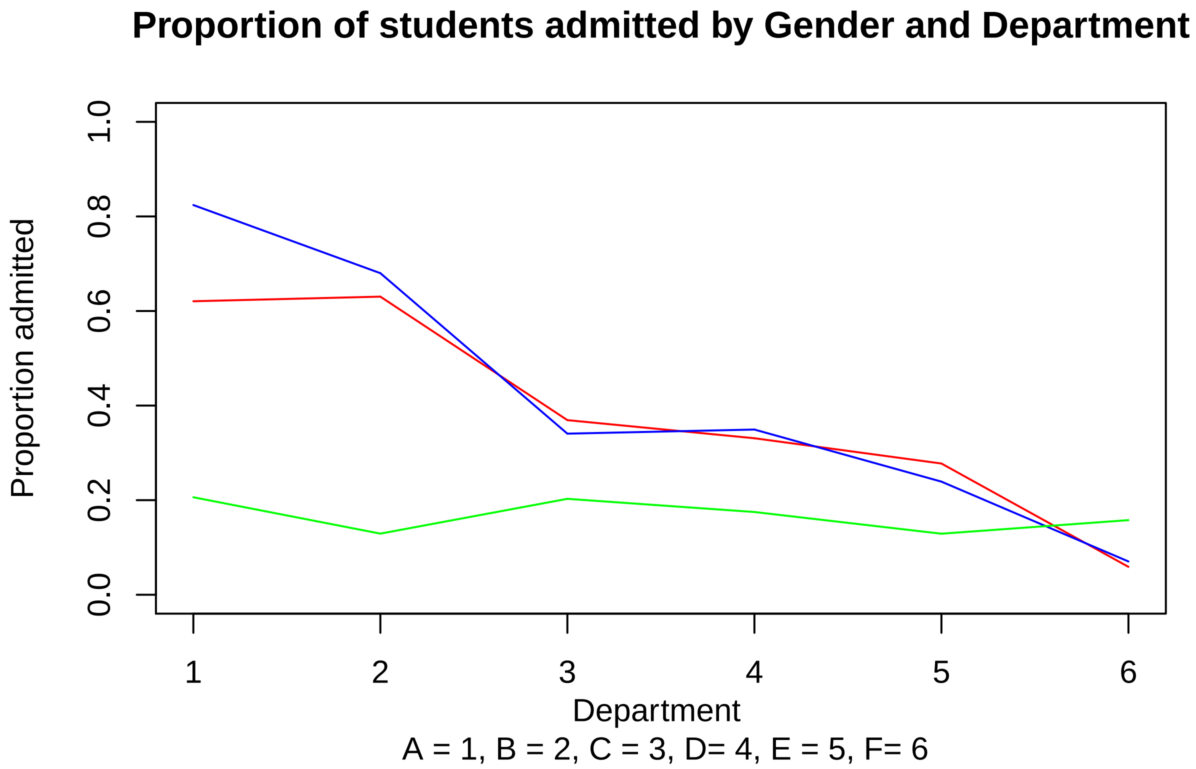
Problem 1 (b)

```
male.acceptances = c(p_maleA, p_maleB, p_maleC, p_maleD,
                     p_maleE, p_maleF, ate_male
                     )
female.acceptances = c(p_femaleA, p_femaleB, p_femaleC, p_femaleD,
                      p_femaleE, p_femaleF, ate_female)

plot(0,0,cex = 0, xlim = c(0,1), ylim = c(0,1), xlab = "Gender (male = 0)",ylab = "proportion of studen

for(i in 1:6){
  segments(0,male.acceptances[i], 1, female.acceptances[i],col = i)

}
legend(x = .1, y = 1.03, legend=c("Depart A", "Depart B", "Depart C", "Depart D", "Depart E", "Depart F
```



Problem 1 (c)

```
plot(1, type="n", main = "Proportion of students admitted by Gender and Department",
     xlim = c(1,6), ylim = c(0,1), ylab = "Proportion admitted",
     xlab = "Department \n A = 1, B = 2, C = 3, D= 4, E = 5, F= 6")
segments(1:5, male.acceptances[1:5], 2:6, male.acceptances[2:6], col = "red")
segments(1:5, female.acceptances[1:5], 2:6, female.acceptances[2:6], col = "blue")
segments(1:5, c(pA, pB, pC, pD, pE), 2:6, c(pB, pC, pD, pE, pF), col = "green")
```



The estimate of for the adjusted treatment effect is different from the regression estimate from regressing Y on X because the effect that gender has on the proportion of students admitted is not constant. Different departments have different effects that gender has on acceptance.

Problem 2

Problem 2 (a)

```
sat <- read.table("/Users/morganhawkins/Downloads/CASE1201.ASC", header = TRUE)
head(sat,10)
```

```
##          state  sat takers income years public expend rank
## 1         Iowa 1088      3    326 16.79   87.8  25.60 89.7
## 2  SouthDakota 1075      2    264 16.07   86.2  19.95 90.6
## 3  NorthDakota 1068      3    317 16.57   88.3  20.62 89.8
## 4       Kansas 1045      5    338 16.30   83.9  27.14 86.3
## 5     Nebraska 1045      5    293 17.25   83.6  21.05 88.5
## 6      Montana 1033      8    263 15.91   93.7  29.48 86.4
## 7    Minnesota 1028      7    343 17.41   78.3  24.84 83.4
## 8         Utah 1022      4    333 16.57   75.2  17.42 85.9
## 9      Wyoming 1017      5    328 16.01   97.0  25.96 87.5
## 10   Wisconsin 1011     10    304 16.85   77.3  27.69 84.2
```

```r
#sat
```

```r
sat[order(sat$sat, decreasing = T),c(1,2)] %>% head
```

```
##         state  sat
## 1        Iowa 1088
## 2 SouthDakota 1075
## 3 NorthDakota 1068
## 4      Kansas 1045
## 5    Nebraska 1045
## 6     Montana 1033
```

```r
mod = lm(sat ~ takers + rank, data=sat)
residual.df = data.frame(state = sat$state, residual = residuals(mod))
residual.df[order(residual.df$residual,decreasing = T),] %>% head
```

```
##            state residual
## 35   Connecticut 53.89280
## 1           Iowa 53.51935
## 28  NewHampshire 45.83384
## 41 Massachusetts 41.92908
## 36       NewYork 40.85268
## 7      Minnesota 40.61878
```

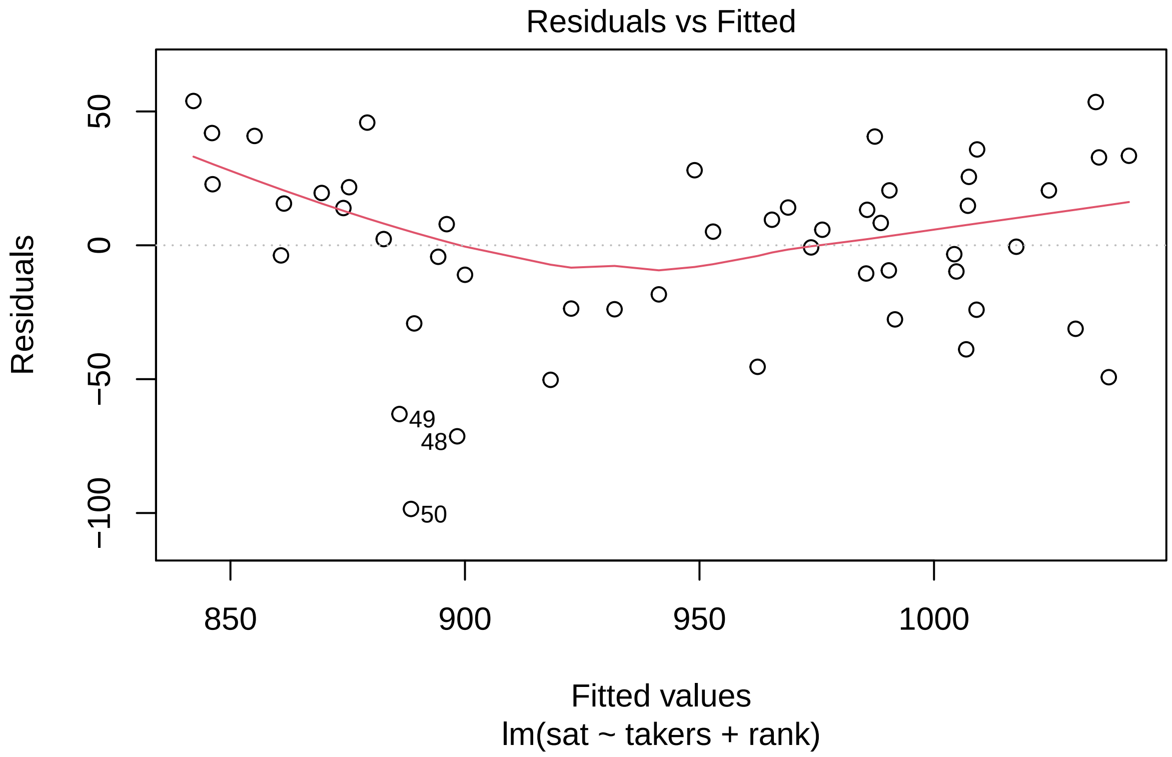
```r
mod %>% summary
```

```
##
## Call:
## lm(formula = sat ~ takers + rank, data = sat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -98.48 -22.31   5.46  21.40  53.89
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 412.8554   194.2227   2.126   0.0388 *
## takers       -0.8170     0.6584  -1.241   0.2208
## rank          6.9574     2.2229   3.130   0.0030 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 33.83 on 47 degrees of freedom
## Multiple R-squared:  0.7814, Adjusted R-squared:  0.7721
## F-statistic:    84 on 2 and 47 DF,  p-value: 3.032e-16
```

From this output I see that controlling for takers and rank shifts the ranking significantly. In the top 10 best ranked states when we control for takers and rank, there are 3 states that were previously in the bottom 15 states. Iowa remained with a high ranking, moving from 1 to 2. Many of the top states are now stakes on the east coast whereas previously there were states closer to the center of the country. The state that rose the most was Massachusetts This is likely because Massachusetts had a relatively low rank but higher number of takers, so its performance was under predicted because our model estimates a negative relationship between takers and sat and a positive relationship between rank and sat. The inverse is true for Arkansas Arkansas has a relatively low number of test takers, but a high rank in average score, so our model over predicted its performance.

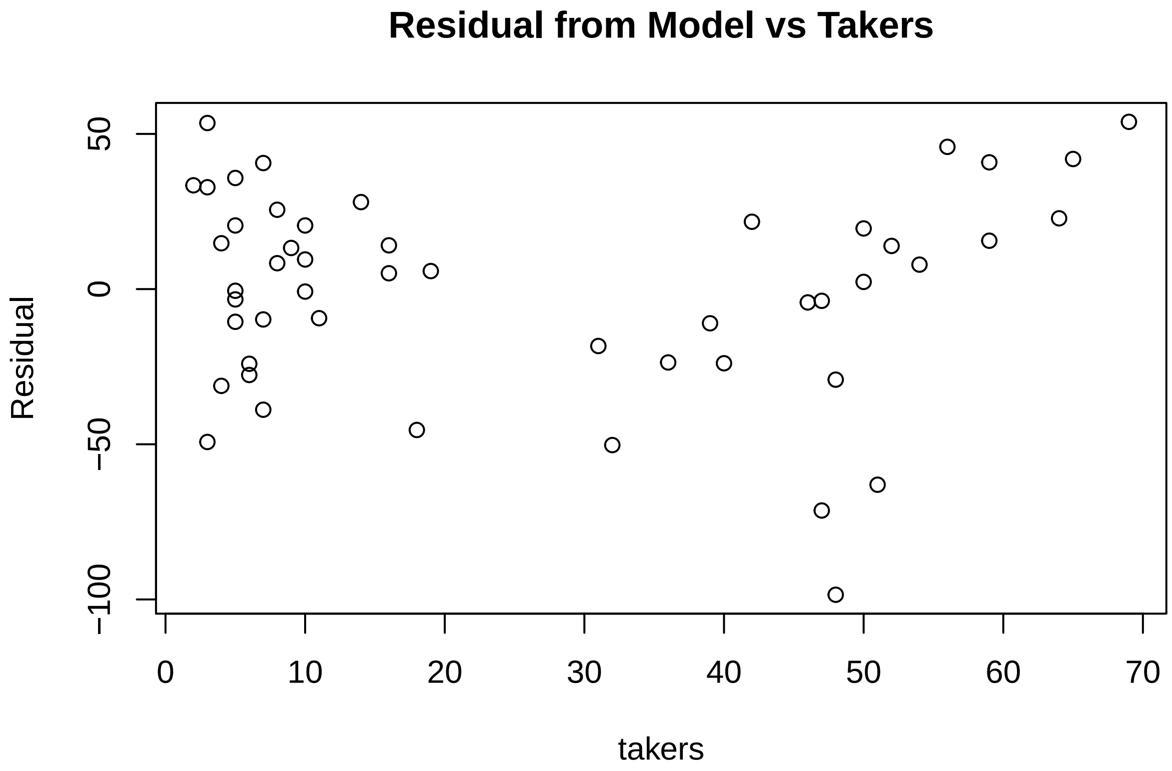
Problem 2 (b)

```
plot(mod,1)
```



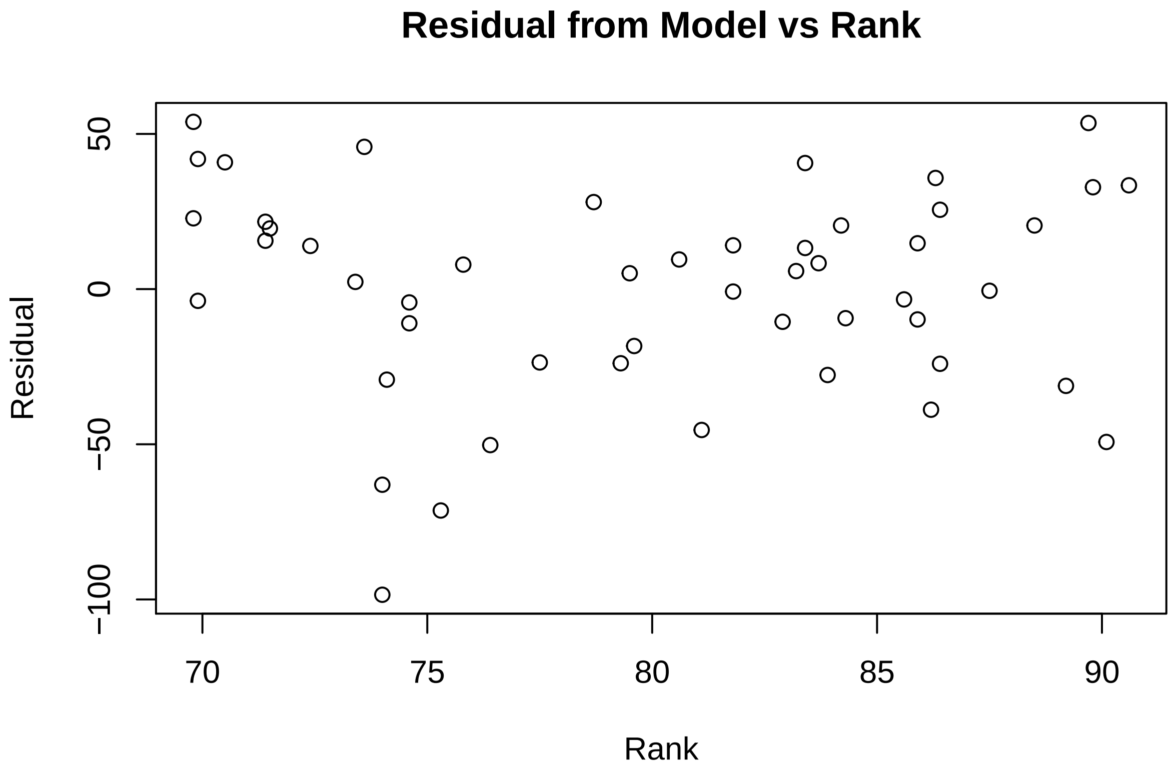
The residuals vs fitted plot seems to show a weak relationship between the fiutted values and residuals. It appears that our model over predicts states that have low and high predicted sat scores (<900 and >975), but under predicts states with sat scores towards to middle (900-975)

```
plot(sat$takers, residuals(mod), xlab = "takers", ylab = "Residual", main = "Residual from Model vs Tak
```



looking at the plot above, there does not appear to be any strong relationship between residuals and the takers variable. However, our residuals may follow a weak parabolic pattern indicating that our model may be over predicting states with low and high percentages of takers(<20, >50) and under predicting states that have takers values closer to the mean (20-50).

```r
plot(sat$rank, residuals(mod), ylab = "Residual", xlab = "Rank", main = "Residual from Model vs Rank")
```



Similar to the previous plot, the plot above also does not appear to show any relationship between residuals and the rank variables.

Problem 2 (c)

```
mod.log = lm(sat ~ log(takers) + rank, data=sat)
summary(mod.log)
```

```
##
## Call:
## lm(formula = sat ~ log(takers) + rank, data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -94.458 -17.305   5.322  22.825  48.467
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  882.081    224.126   3.936 0.000273 ***
## log(takers)  -45.192     14.059  -3.214 0.002364 **
## rank           2.396      2.330   1.028 0.308982
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

```
## Residual standard error: 31.12 on 47 degrees of freedom
## Multiple R-squared:  0.8149, Adjusted R-squared:  0.8071
## F-statistic: 103.5 on 2 and 47 DF,  p-value: < 2.2e-16
```

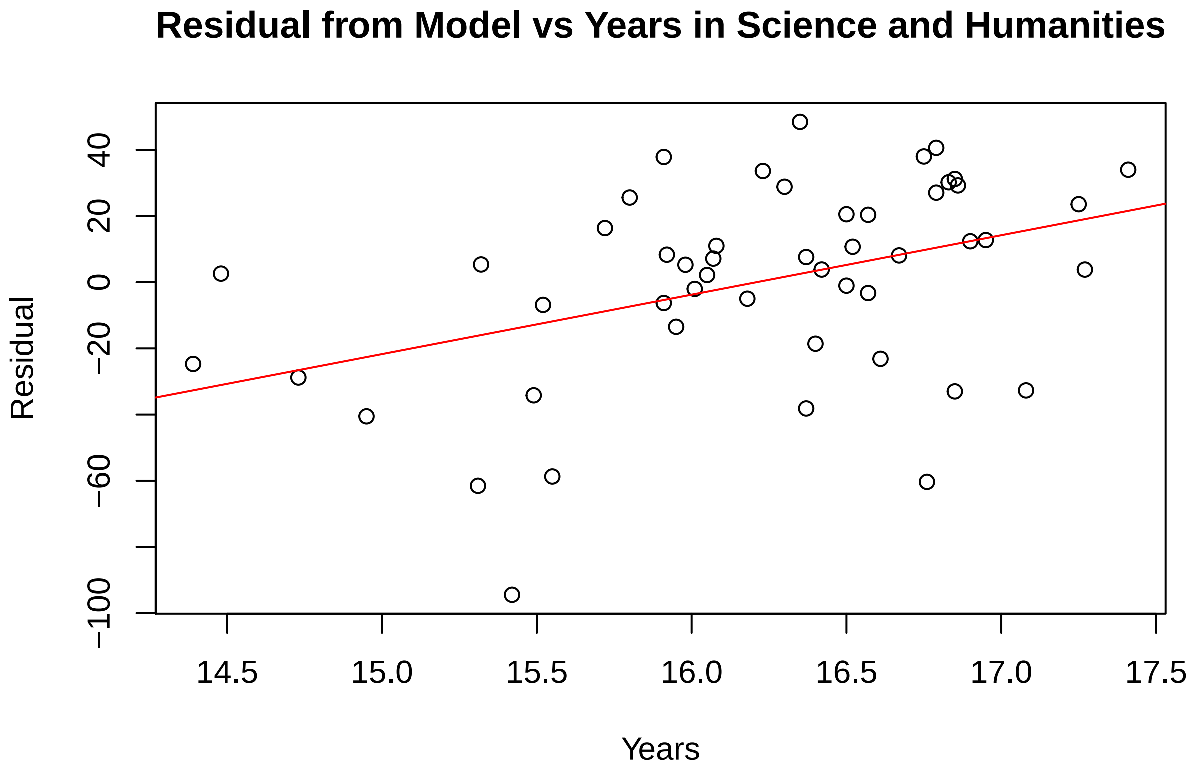
looking at the residual plots above, it may be appropriate to transform the takers variable with the log function. Applying this transformation improves our Rˆ2 to .8149 from .7814

```
#sat[order(residuals(mod), decreasing = T),1:2]
#sat[order(residuals(mod.log), decreasing = T),1:2]

#rank(residuals(mod.log))-1:50
#sat[48,]
```

Applying this transformation does change the ranks of states a little. New Hampshire and Connecticut switched places. Massachusetts is now ranked 11th and is no longer the stater with the largest increase in position. The bottom of the rankings remain similar.

Problem 2 (d)

```
plot(sat$year, residuals(mod.log), xlab = "Years", ylab = "Residual", main = "Residual from Model vs Yea
lm(residuals(mod.log) ~ sat$years) %>% abline(col = "red")
```



The plot above shows a positive approximately linear relationship between years and the residual from our transformed model. Also, the slope coefficient is statistically significant as shown in the model summary below (p value = 0.00291).

```r
summary(lm(residuals(mod.log) ~ sat$years))
```

```
##
## Call:
## lm(formula = residuals(mod.log) ~ sat$years)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -80.282  -9.517   4.964  17.332  45.938
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -291.159     92.897  -3.134  0.00294 **
## sat$years     17.963      5.726   3.137  0.00291 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 28.06 on 48 degrees of freedom
## Multiple R-squared:  0.1701, Adjusted R-squared:  0.1529
## F-statistic: 9.841 on 1 and 48 DF,  p-value: 0.002914
```

Problem 2 (e)

```r
m.1 = lm(years ~ sqrt(takers) + rank, data=sat)
m.2 = lm(sat ~ sqrt(takers) + rank, data=sat)
df.res = data.frame(m1.res = residuals(m.1), m2.res = residuals(m.2))
m.3 = lm(m2.res ~ m1.res, data=df.res)
m.4 = lm(sat ~ sqrt(takers) + rank + years, data=sat)
summary(m.3)
```

```
##
## Call:
## lm(formula = m2.res ~ m1.res, data = df.res)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -79.242  -7.384   2.043  16.091  40.617
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.802e-15  3.871e+00   0.000 1.000000
## m1.res       2.463e+01  5.898e+00   4.175 0.000125 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 27.37 on 48 degrees of freedom
## Multiple R-squared:  0.2664, Adjusted R-squared:  0.2512
## F-statistic: 17.43 on 1 and 48 DF,  p-value: 0.0001245
```

```
summary(m.4)
```

```
##
## Call:
## lm(formula = sat ~ sqrt(takers) + rank + years, data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -79.242  -7.384   2.043  16.091  40.617
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    96.317    254.444   0.379 0.706772
## sqrt(takers)   -9.508      6.515  -1.459 0.151258
## rank            6.204      2.243   2.766 0.008135 **
## years          24.625      6.024   4.088 0.000173 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.96 on 46 degrees of freedom
## Multiple R-squared:  0.8538, Adjusted R-squared:  0.8443
## F-statistic: 89.56 on 3 and 46 DF,  p-value: < 2.2e-16
```

The coefficient for m1.res is the same as years.

Problem 3

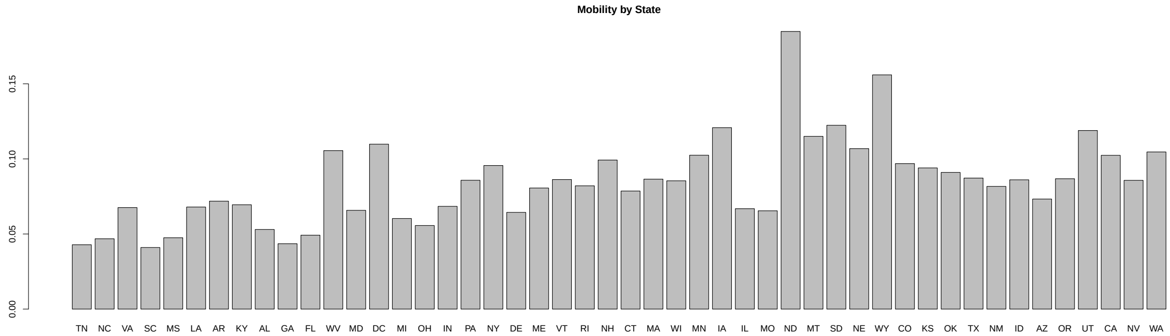
Problem 3 (a)

```
mobility = read.csv("/Users/morganhawkins/Downloads/mobility.csv")
mobility = na.omit(mobility)
```

I will omit rows from the data set that have missing values for my entire analyses. The upside to doing this is that I am using the same data for all my analysis. This means I am using a more clearly defined dataset. The downside is that I am unnecessarily omitting data point for some questions. Some columns have no missing values so if we were to analyze just those two columns in a question, we could have used the whole dataset for that question. Another potential downside is that missing values may not be randomly present creating a bias in our data set. For example, maybe cities with a population that is too small don't have a school so they would have a missing value for school_spending. This would cause our estimated mean population in communities to be artificially large since we are omitting communities below a certain population from our dataset.

Problem 3 (b)

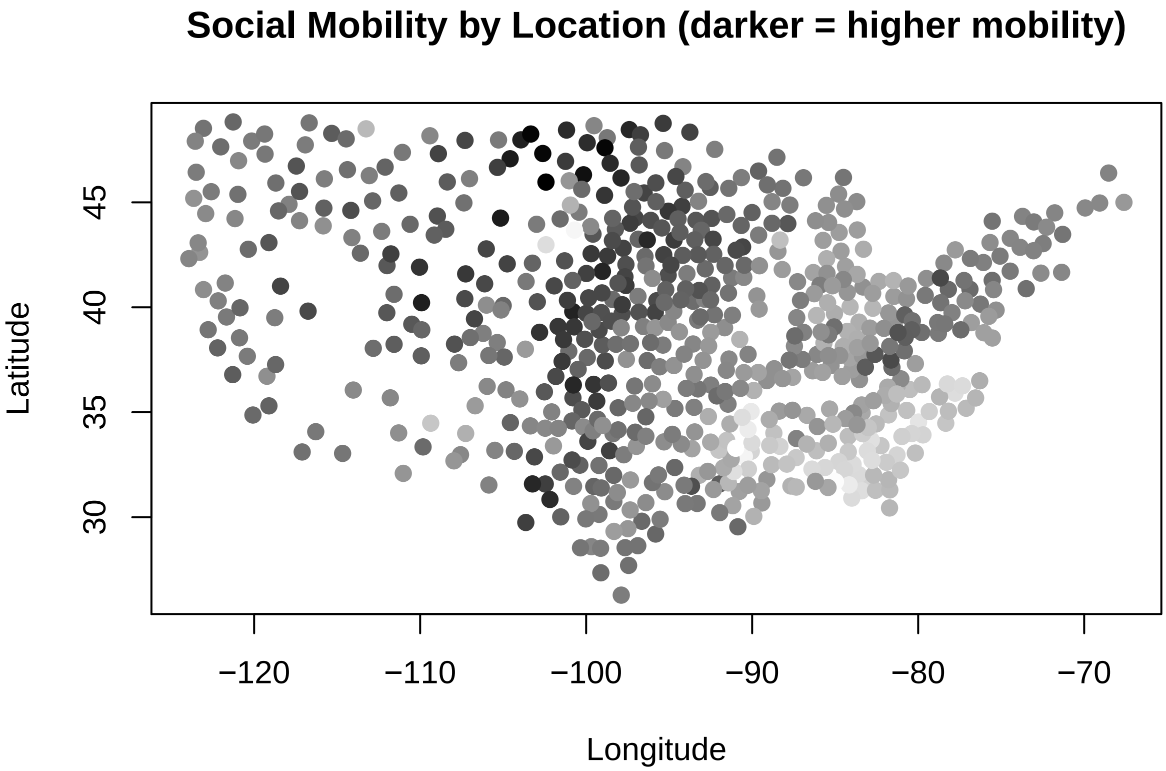
```
state_mobilities = c()
for(s in unique(mobility$State)){
  temp.mobility = subset(mobility, State == s)
  total.pop = sum(temp.mobility$Population)
  state_mobilities[s] = sum(temp.mobility$Mobility*temp.mobility$Population/total.pop)

}
barplot(state_mobilities, main = "Mobility by State")
```



Looking at the bar plot above, I notice that ND has exceptionally high mobility along with Wyoming while states like Illinois, Missouri, South Carolina, Tennessee, North Carolina, Mississippi, Georgia, and Florida have exceptionally low mobility. It appears that economic mobility does vary by state.

Problem 3 (c)

```
#logging before normalization values to be between 0 and 1 to make distribution of normalized values mo
#hist(log(mobility$Mobility))
mobility.range = max(log(mobility$Mobility)) - min(log(mobility$Mobility))
mobility.min = min(log(mobility$Mobility))

mobility$Mobility.Scaled = (log(mobility$Mobility) - mobility.min)/mobility.range


plot( mobility$Longitude, mobility$Latitude, cex = 1.2, pch = 16,
      col = rgb(1-mobility$Mobility.Scaled,
                1-mobility$Mobility.Scaled,
                1-mobility$Mobility.Scaled),
      xlab = "Longitude",
      ylab = "Latitude",
      main = "Social Mobility by Location (darker = higher mobility)"
      )
```



It appears that communities near the east and west coasts have lower social mobility than states closer to the middle of the country. The southeastern United States seems to have the lowest mobility while the great plains seem to have the highest.

Problem 3 (d)

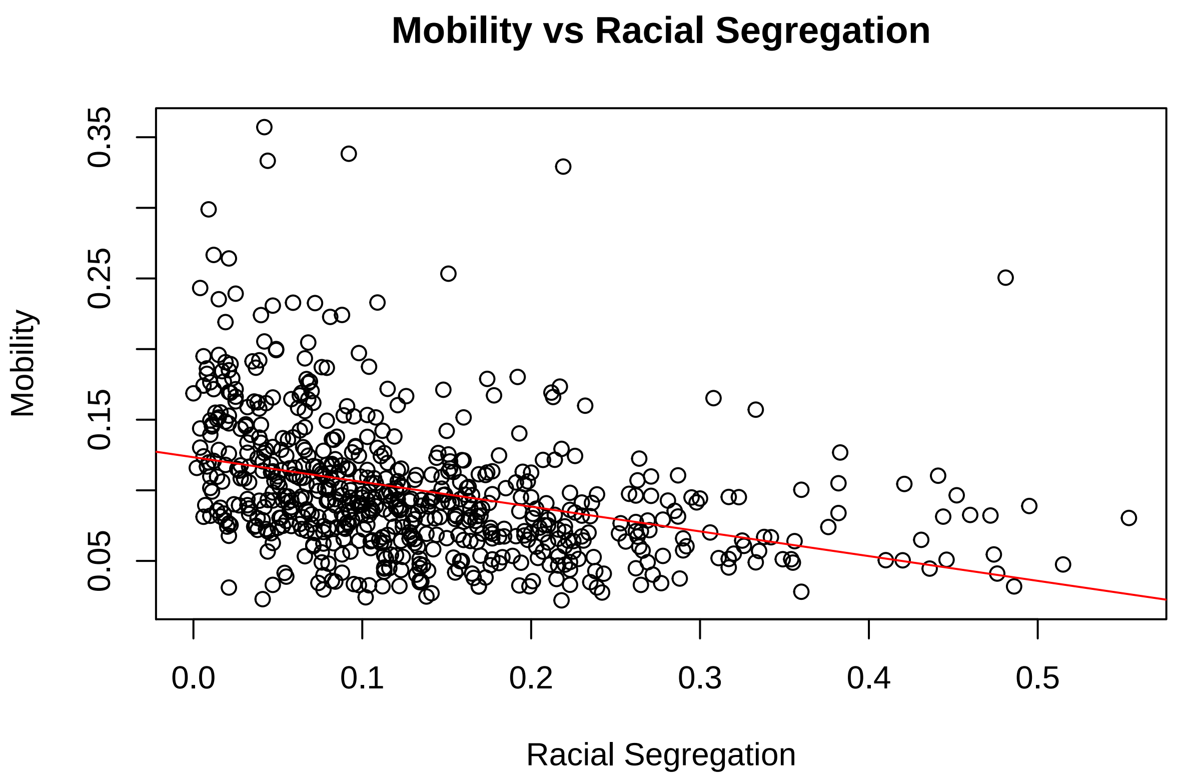
We would be interested in residual test score after regressing our household income because this shows us how the community performs on standardized testing when we remove the effect that household income has on standardized test scores.

Problem 3 (e)

```r
colnames(mobility)
```

```
##  [1] "ID"                   "Name"                 "Mobility"
##  [4] "State"                "Population"           "Urban"
##  [7] "Black"                "Seg_racial"           "Seg_income"
## [10] "Income"               "Gini"                 "Share01"
## [13] "Middle_class"         "School_spending"      "Student_teacher_ratio"
## [16] "Test_scores"          "Social_capital"       "Religious"
## [19] "Violent_crime"        "Single_mothers"       "Divorced"
## [22] "Married"              "Longitude"            "Latitude"
## [25] "Mobility.Scaled"
```

```
#i)
racial.mod = lm(Mobility ~ Seg_racial, data = mobility)
plot(mobility$Seg_racial,mobility$Mobility, xlab = "Racial Segregation", ylab = "Mobility", main = "Mob
abline(racial.mod, col = "red")
```



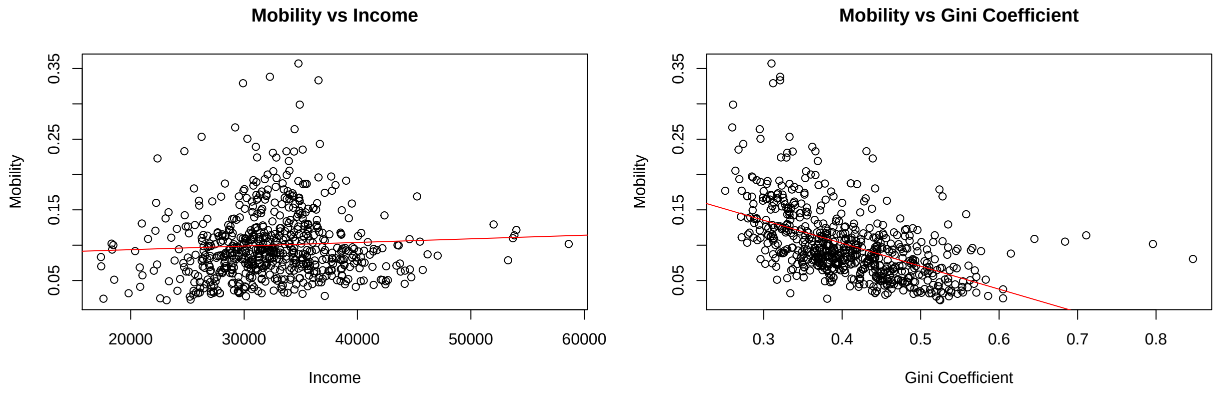
The plot above looks approximately linear and racial segregation appears to be predictive of mobility.

```
summary(racial.mod)
```

```
##
## Call:
## lm(formula = Mobility ~ Seg_racial, data = mobility)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.093308 -0.030167 -0.007078  0.020000  0.244108
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.123441   0.003036  40.663   <2e-16 ***
## Seg_racial  -0.175072   0.018136  -9.653   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04606 on 633 degrees of freedom
## Multiple R-squared:  0.1283, Adjusted R-squared:  0.1269
## F-statistic: 93.19 on 1 and 633 DF,  p-value: < 2.2e-16
```

After fitting a model, we see in the output summary above that racial segregation is able to explain 12.83% of the variance in mobility between communities. As shown in the model summary above, there is a significant relationship between Mobility and racial segregation with p value < 10e-16.

```
#ii)
par(mfrow = c(1,2))
income.mod = lm(Mobility ~ Income, data = mobility)
plot(mobility$Income,mobility$Mobility, xlab = "Income", ylab = "Mobility", main = "Mobility vs Income"
abline(income.mod, col = "red")

gini.mod = lm(Mobility ~ Gini, data = mobility)
plot(mobility$Gini,mobility$Mobility, xlab = "Gini Coefficient", ylab = "Mobility", main = "Mobility vs
abline(gini.mod, col = "red")
```



Both plots above look approximately linear. Gini goes seem to be predictive of mobility, but it is unclear whether income is

```
summary(income.mod)
```

```
##
## Call:
## lm(formula = Mobility ~ Income, data = mobility)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.074212 -0.032125 -0.009868  0.019317  0.255937
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.339e-02  1.191e-02   7.001 6.51e-12 ***
## Income      5.120e-07  3.613e-07   1.417    0.157
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04925 on 633 degrees of freedom
## Multiple R-squared:  0.003163,   Adjusted R-squared:  0.001588
## F-statistic: 2.008 on 1 and 633 DF,  p-value: 0.1569
```

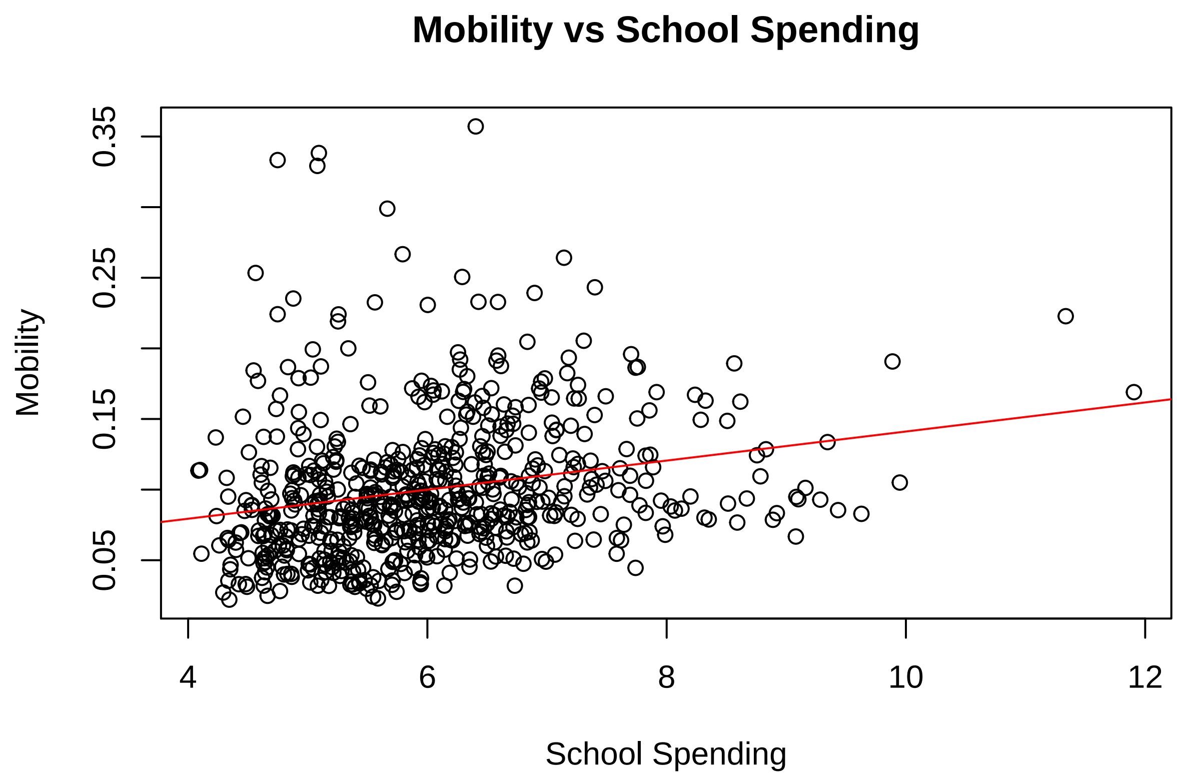
After fitting a model, we see in the output summary above that income does not have a significant relationship with with mobility (p value = .157).

```
summary(gini.mod)
```

```
##
## Call:
## lm(formula = Mobility ~ Gini, data = mobility)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.092243 -0.028512 -0.008483  0.019252  0.225149
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.232773   0.008733   26.66   <2e-16 ***
## Gini        -0.325093   0.020996  -15.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04201 on 633 degrees of freedom
## Multiple R-squared:  0.2747, Adjusted R-squared:  0.2736
## F-statistic: 239.8 on 1 and 633 DF,  p-value: < 2.2e-16
```

Fitting a linear model with mobility as the response and gini as the predictor, we see that gini has a strong relationship with income (p values < 2e-16). It is able to explain 27.47$ of the variance in mobility.

```
#iii)
school.mod = lm(Mobility ~ School_spending, data = mobility)
plot(mobility$School_spending,mobility$Mobility, xlab = "School Spending", ylab = "Mobility", main = "M
abline(school.mod, col = "red")
```



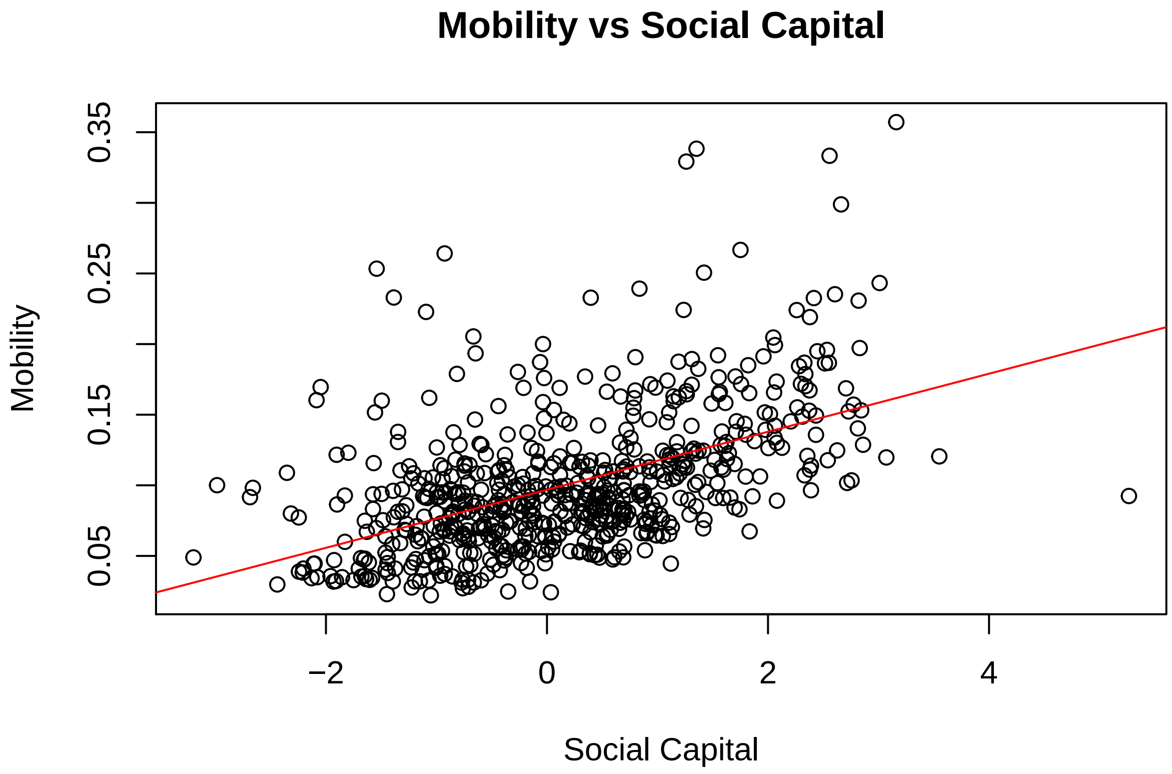
The relationship between school_spending and Mobility looks approximately linear and positive.

```r
summary(school.mod)
```

```
## 
## Call:
## lm(formula = Mobility ~ School_spending, data = mobility)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.075532 -0.032818 -0.008637  0.021683  0.253027 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     0.038211   0.010468   3.650 0.000284 ***
## School_spending 0.010291   0.001713   6.007 3.19e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 0.04798 on 633 degrees of freedom
## Multiple R-squared:  0.05393,    Adjusted R-squared:  0.05244 
## F-statistic: 36.08 on 1 and 633 DF,  p-value: 3.189e-09
```

After fitting a model, we see in the output summary above that school is able to explain 5.393% of the variance in mobility. As shown in the model summary above, there is a significant relationship between school spending and mobility (p values = 3.19e-09).

```
#iv)
social.mod = lm(Mobility ~ Social_capital, data = mobility)
plot(mobility$Social_capital,mobility$Mobility, xlab = "Social Capital", ylab = "Mobility", main = "Mob
abline(social.mod, col = "red")
```



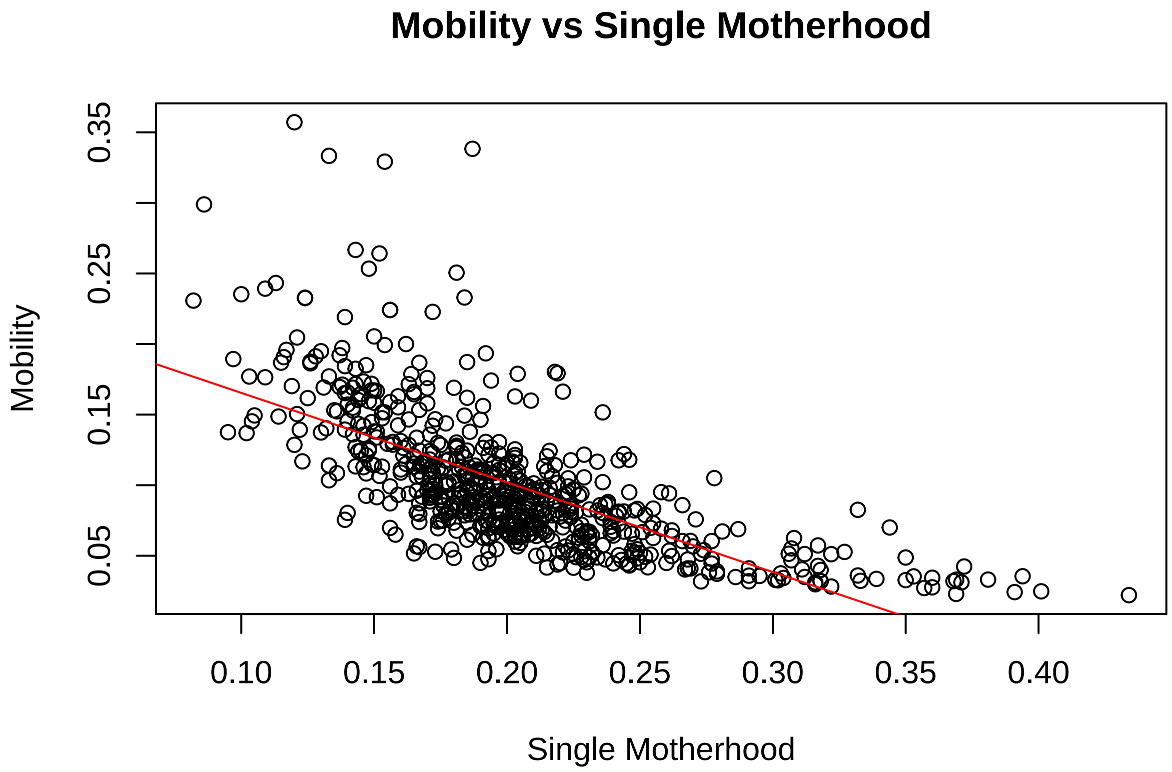
The relationship between social capital and mobility looks approximately linear and positive.

```
summary(social.mod)
```

```
##
## Call:
## lm(formula = Mobility ~ Social_capital, data = mobility)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.112600 -0.027996 -0.008579  0.017564  0.213682
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.096823   0.001696   57.08   <2e-16 ***
## Social_capital 0.020550   0.001374   14.96   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0424 on 633 degrees of freedom
## Multiple R-squared:  0.2612, Adjusted R-squared:   0.26
## F-statistic: 223.8 on 1 and 633 DF,  p-value: < 2.2e-16
```

After fitting a model, we see that social capital is able to explain 26.12% of the variance in mobility. As shown in the model summary above, the relationship between social capital and mobility is significant with p value < 2e-16

```
#v)
single.mod = lm(Mobility ~ Single_mothers, data = mobility)
plot(mobility$Single_mothers,mobility$Mobility, xlab = "Single Motherhood", ylab = "Mobility", main = "
abline(single.mod, col = "red")
```



The relationship between single motherhood and mobility looks very strong and approximately linear

```
summary(single.mod)
```

```
##
## Call:
## lm(formula = Mobility ~ Single_mothers, data = mobility)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.072460 -0.023689 -0.004915  0.014956  0.228149
##
## Coefficients:
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.228851   0.005669   40.37   <2e-16 ***
## Single_mothers -0.634718   0.027030  -23.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03606 on 633 degrees of freedom
## Multiple R-squared:  0.4656, Adjusted R-squared:  0.4647
## F-statistic: 551.4 on 1 and 633 DF,  p-value: < 2.2e-16
```

After fitting a model, we see that single motherhood is able to explain almost half the variance in mobility between communities with an adj R^2 of .4647. The relationship between single motherhood and social mobility is very strong with a p value < 2e-16

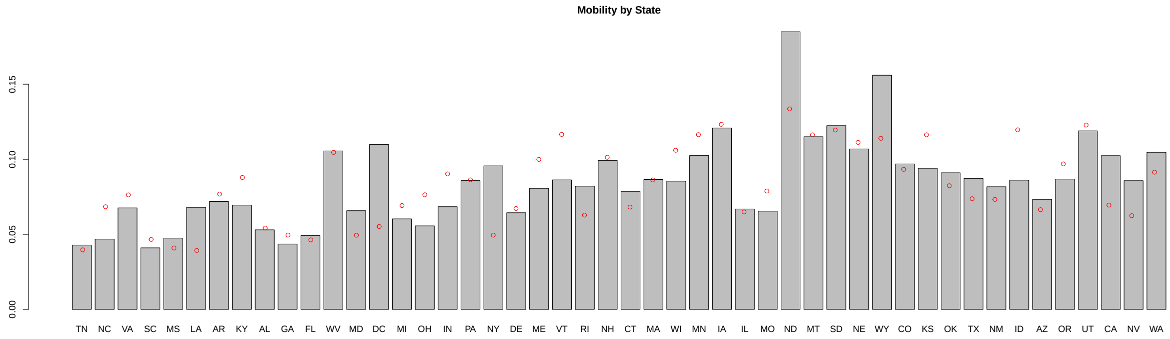
```
full.mod <- lm(Mobility ~ Seg_racial + Gini + Income + School_spending + Social_capital + Single_mothers
summary(full.mod)
```

```
##
## Call:
## lm(formula = Mobility ~ Seg_racial + Gini + Income + School_spending +
##     Social_capital + Single_mothers, data = mobility)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.066575 -0.021770 -0.005223  0.013607  0.217628
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.427e-01  1.398e-02  17.357  < 2e-16 ***
## Seg_racial      -4.001e-02  1.496e-02  -2.674  0.00769 **
## Gini            -3.289e-02  2.452e-02  -1.341  0.18042
## Income          -1.353e-06  3.013e-07  -4.492 8.40e-06 ***
## School_spending  2.824e-03  1.315e-03   2.147  0.03218 *
## Social_capital   8.892e-03  1.549e-03   5.740 1.48e-08 ***
## Single_mothers  -4.841e-01  3.553e-02 -13.623  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03389 on 628 degrees of freedom
## Multiple R-squared:  0.5317, Adjusted R-squared:  0.5272
## F-statistic: 118.8 on 6 and 628 DF,  p-value: < 2.2e-16
```

Fitting the full model together we get some changes. Gini become insignificant, income becomes significant.

Problem 3 (f)

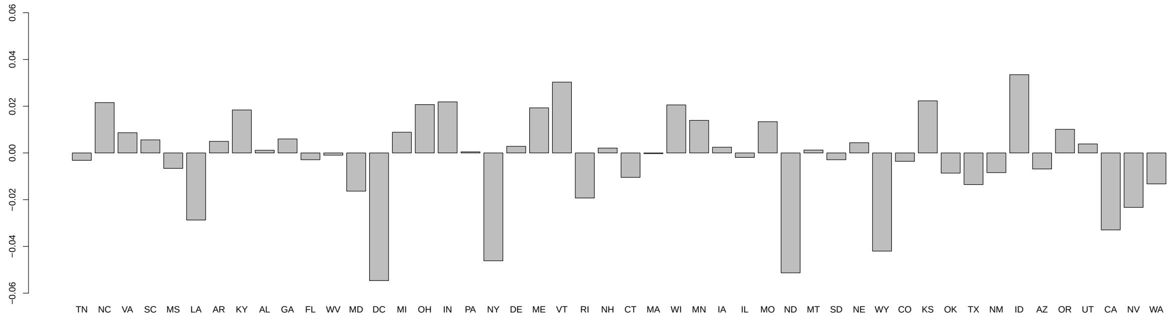
```
state_mobilities = c()
pred_mobilities = c()
for(s in unique(mobility$State)){
  temp.mobility = subset(mobility, State == s)
  total.pop = sum(temp.mobility$Population)
  temp.pred.mobil = predict(full.mod, temp.mobility)
  temp.pred.mobil = sum(temp.pred.mobil*temp.mobility$Population/total.pop)
  pred_mobilities = c(pred_mobilities, temp.pred.mobil)

  state_mobilities[s] = sum(temp.mobility$Mobility*temp.mobility$Population/total.pop)


}
#pred_mobilities
#length(seq(.75,50,(end-.75)/47))
barplot(state_mobilities, main = "Mobility by State")
end = 57
points(seq(.75,end,(end-.75)/47),pred_mobilities, col = "red")
```



The plot is showing how well our model is capturing the differences in mobility between states without explicitly including state in our model.

Problem 3 (g)

```
barplot(pred_mobilities - state_mobilities, ylim = c(-.06,.06))
```



I think location should be included in our model because most of the states seem to have small residuals from their predicted, but other states such as DC, ND, and WY have very large residuals. Much larger than any others.

Problem 3 (h)

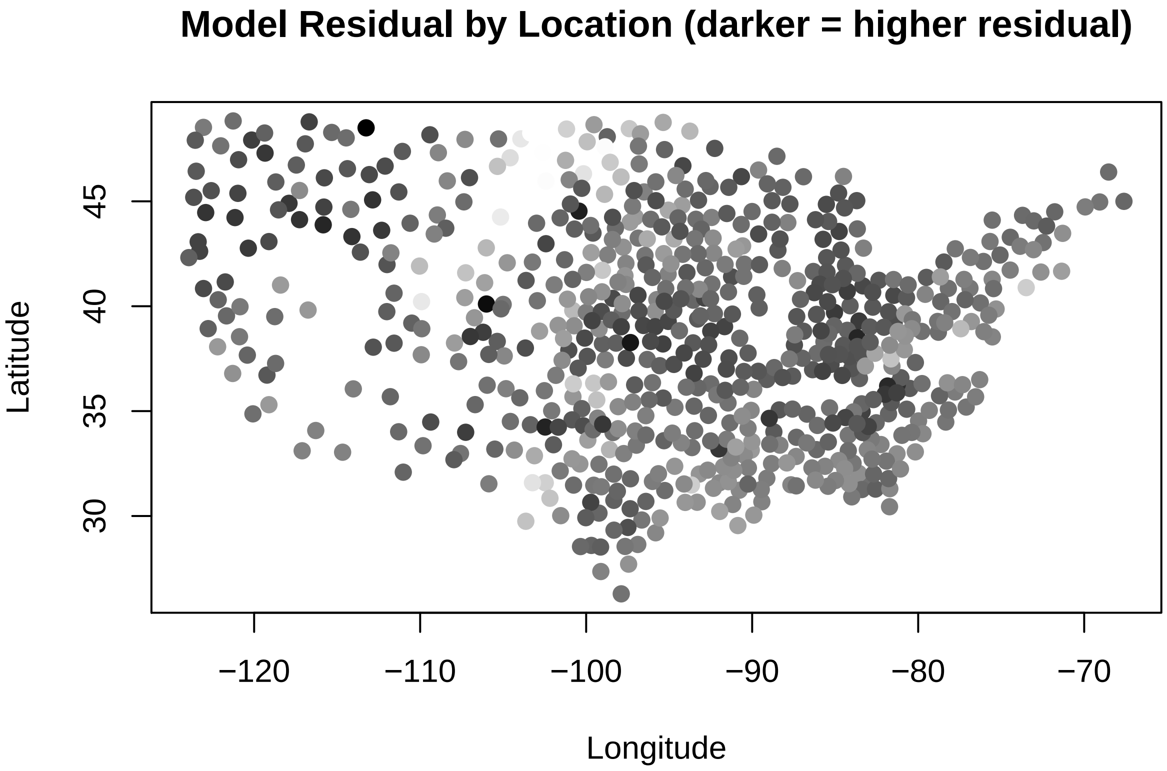
```
full.mod.loc <- lm(Mobility ~ Seg_racial + Gini + Income + School_spending + Social_capital + Single_mo
summary(full.mod.loc)
```

```
##
## Call:
## lm(formula = Mobility ~ Seg_racial + Gini + Income + School_spending +
##     Social_capital + Single_mothers + Longitude + Latitude, data = mobility)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.076217 -0.020070 -0.005259  0.012150  0.215189
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.527e-01  2.253e-02   6.776 2.85e-11 ***
## Seg_racial      -2.840e-02  1.450e-02  -1.959 0.050607 .
## Gini            -3.957e-02  2.460e-02  -1.609 0.108199
## Income          -1.276e-06  2.907e-07  -4.391 1.33e-05 ***
## School_spending  4.514e-03  1.307e-03   3.453 0.000591 ***
## Social_capital   9.342e-03  1.633e-03   5.723 1.63e-08 ***
## Single_mothers  -4.355e-01  3.464e-02 -12.571  < 2e-16 ***
## Longitude       -8.776e-04  1.154e-04  -7.605 1.05e-13 ***
## Latitude        -3.709e-04  3.816e-04  -0.972 0.331458
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03247 on 626 degrees of freedom
## Multiple R-squared:  0.5715, Adjusted R-squared:  0.566
## F-statistic: 104.3 on 8 and 626 DF,  p-value: < 2.2e-16
```

```
#summary(full.mod)
```

Some estimates did change. Seg_racial increased by just under a standard error and school spending decreased by a little over one standard error. Latitude does no appear to be significant predictor of mobility. However, Longitude has a significant relationship with mobility (p value 1.05e-13).

```
#logging before normalization values to be between 0 and 1 to make distribution of normalized values mo
#hist(log(mobility$Mobility))

mob = predict(full.mod.loc, mobility) - mobility$Mobility
mob = mob + min(mob) + 1
mobility.range = max((mob)) - min((mob))
mobility.min = min((mob))

mobility$Mobility.Scaled = ((mob) - mobility.min)/mobility.range


plot( mobility$Longitude, mobility$Latitude, cex = 1.2, pch = 16,
      col = rgb(1-mobility$Mobility.Scaled^2,
                1-mobility$Mobility.Scaled^2,
                1-mobility$Mobility.Scaled^2),
      xlab = "Longitude",
      ylab = "Latitude",
      main = "Model Residual by Location (darker = higher residual)"
      )
```



There appears to be a little bit of information still left in location. The East and west seem to be slightly over predicted by our model. Additionally, states in the great plains seem to be under predicted, especially states in the northern great plains.