Final Project Sangjin Lee

Part II

The report will look at economic mobility across generations in the contemporary USA.

Hypothesis

Educational factors are more strongly associated with economic mobility than social factors.

Method

Compare the relationship of each variable with mobility - two from educational factors and two from social factors.

Educational factors Test_scores: Residuals from a linear regression of mean math and English test scores on household income per capita. **Graduation**: Residuals from a linear regression of the actual college graduation rate on household income per capita.

Social factors Single_mothers: Number of single female households with children divided by the total number of households with children. **Social_capital**: Index combining voter turnout, participation in the census, and participation in community organizations.

To figure out each relationship we will conduct linear regression for the two variables. In this case, comparing P-values is not observed to be a strong analysis in clarifying the correlation between the variables because the P-values for all variables are all small. Therefore, we will assess P-values, however, further investigate taking following steps:

1. Plot each variable into a scatter graph. Also fit a simple parametric model to render enhanced visual understanding of changes in the variable data.
2. Genereate a fitted linear regression model for each variable.
3. Obtain summary for each fitted linear model. This is to understand the distribution of our variables.
4. Compare P-values to understand the statistical significance of each variable or, in other words, how strong the association is for each variable to ‘economic mobility’ - therefore, testing whether to reject the null hypothesis.
5. If the P-value analysis does not yield a significant result, compare the slopes of linear regression lines. If we find that the slope of the regression line is significantly different from zero, we will conclude that there is a significant relationship between the independent and dependent variables.
    a. Get the linear regression model
    b. Get the slope
    c. Compare all slopes for the variables
6. Compare and investigate which group of variables show greater slopes in absolute value.
7. Input x=Q1 and x=Q3 in the regression model for each variable. Get the differences betwen y=b0+Q1x and y=b0+Q3x to observe the predicted mobility change and compare by variables.
8. Understand all the previous tests and conclude on the factor group with the strongest association to economic mobility.

Reading Mobility CSV file to load dataset

```
dat <- read.csv("mobility.csv")
```

Plotting scatter graphs of ‘Mobility~(each variable)’

```
# Plotting each variable
testscore <- ggplot(dat,aes(x=Test_scores,y=Mobility))+geom_point(size=0.6,color="#F7
CF08")+geom_smooth()+theme(panel.grid.major = element_line(size = 0.3, linetype = 'so
lid',colour = "grey"),panel.grid.minor = element_blank())

graduation <- ggplot(dat,aes(x=Graduation,y=Mobility))+geom_point(size=0.6,color="#F7
CF08")+geom_smooth()+theme(axis.title.y=element_blank(),axis.text.y=element_blank(),a
xis.ticks.y = element_blank(),panel.grid.major = element_line(size = 0.3, linetype =
'solid',colour = "grey"),panel.grid.minor = element_blank())

singlemothers <- ggplot(dat,aes(x=Single_mothers,y=Mobility))+geom_point(size=0.6,col
or="#08C8F7")+geom_smooth()+theme(panel.grid.major = element_line(size = 0.3, linetyp
e = 'solid',colour = "grey"),panel.grid.minor = element_blank())

socialcapital <- ggplot(dat,aes(x=Social_capital,y=Mobility))+geom_point(size=0.6,col
or="#08C8F7")+geom_smooth()+theme(axis.title.y=element_blank(),axis.text.y=element_bl
ank(),axis.ticks.y = element_blank(),panel.grid.major = element_line(size = 0.3, line
type = 'solid',colour = "grey"),panel.grid.minor = element_blank())

# Arranging all the plots into one grid
grid.arrange(testscore,graduation,singlemothers,socialcapital,top = textGrob("Relatio
nship between Economic Mobility and Each Variable",gp=gpar(fontsize=16,font=2)),nrow
 = 2)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 36 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 36 rows containing missing values (geom_point).
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 160 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 160 rows containing missing values (geom_point).
```

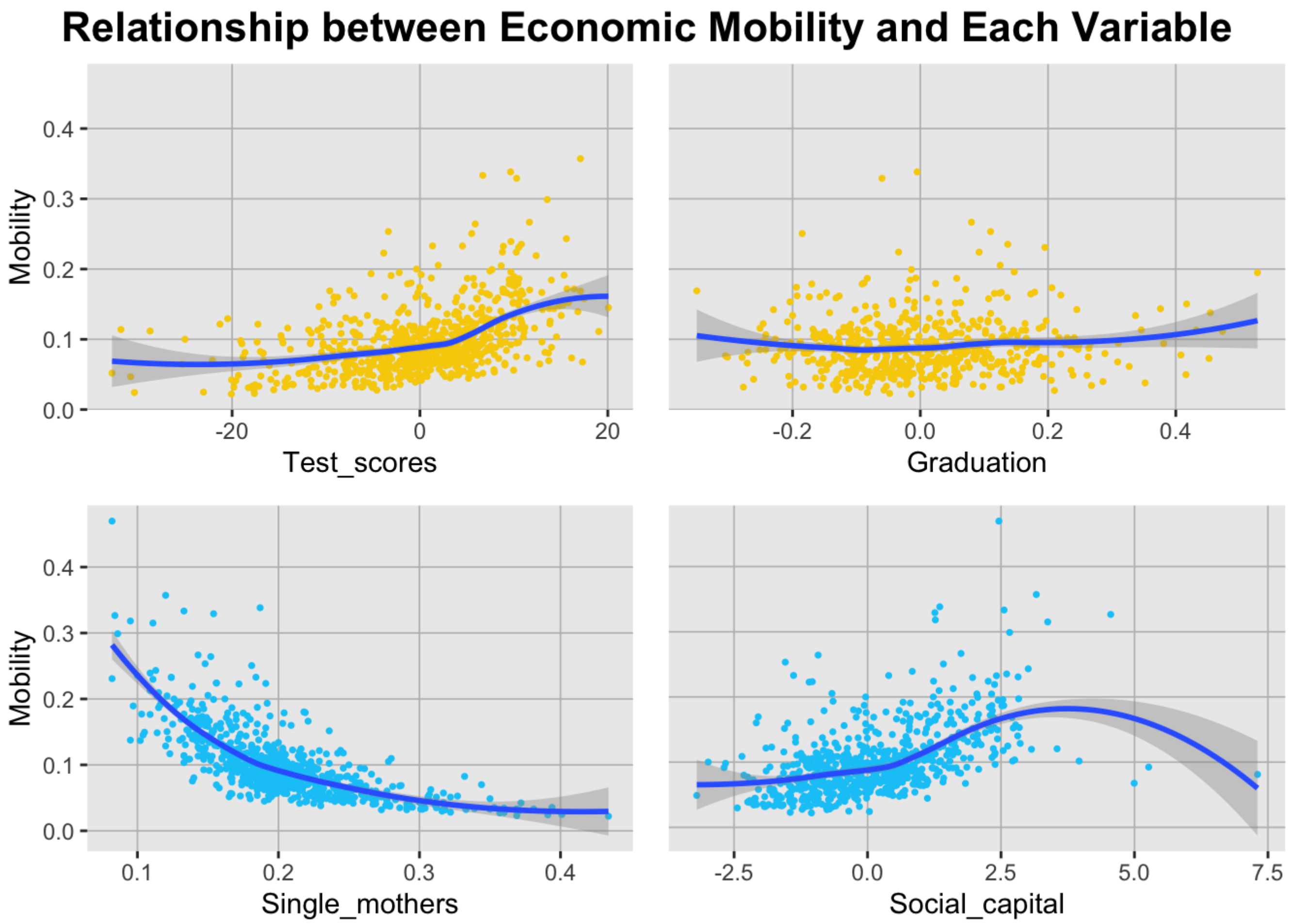
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 12 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 12 rows containing missing values (geom_point).
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 29 rows containing non-finite values (stat_smooth).
## Warning: Removed 29 rows containing missing values (geom_point).
```



The four scatterplots representing four different relationships betwen Test_scores, Graduation, Single_mothers, and Social_capital are represented as above. The four graphs are althogether arranged into one grid for a more convenient visual representation. Test_score + Graduation and Single_mothers + Social_capital each has the same qualitative color for their points to categorize the educational and social factors. These are all represented on a cartesian plane with scatterplots to better show the numerical position of a variable in relation to other variables. Also the numeric quantities of these variables are represented on a linear numeric scale. The y-axis title and text are removed in order to make the data graphic more concise, as the mobility is the response variable for all cases. The x-axis is different for all graphs because exploratory variables differ. The trendline is added to each graph; a parametric model with shaded area representing 95% confidence interval. The minor gridlines are removed to make the data graphic look more simple.

P-values of each variable

```
ls <- list()
# Getting P-value from each regression summary. Then storing into the list.
ls$graduation_pvalue <- summary(lm(Mobility~Graduation, data=dat))$coefficients[2, 4]
ls$testscore_pvalue <- summary(lm(Mobility~Test_scores, data=dat))$coefficients[2, 4]
ls$single_pvalue <- summary(lm(Mobility~Single_mothers, data=dat))$coefficients[2, 4]
ls$social_pvalue <- summary(lm(Mobility~Social_capital, data=dat))$coefficients[2, 4]

ls
```

```
## $graduation_pvalue
## [1] 0.1333785
##
## $testscore_pvalue
## [1] 1.050781e-36
##
## $single_pvalue
## [1] 1.895341e-102
##
## $social_pvalue
## [1] 3.287087e-48
```

You can observe that the P-value for each variable is as follows:

Educational factors Graduation : 0.1333785 Test_scores : 1.050781e-36

Social factors Single_mothers : 1.895341e-102 Social_capital : 3.287087e-48

Social factors seem to have lower P-values and Graduation rate renders a P-value that is higher than 0.05, which means it is statistically insignificant. However, 3 of the 4 P-values are **all statistically significant**, thus we cannot conclude on a distinguishable association. Because the P-value for the Graduation rate is high, it seems that rejecting the null hypothesis is right and therefore the social factors are more stongly associated with economic mobility. Yet, a further exploration on the variables needs to be taken into account.

Linear Regression Model

Proceed to Step.5 of the method procedure that is: *If the P-value analysis does not yield a significant result, compare the slopes of linear regression lines.*

To re-emphasize, we need to determine whether there is a significant linear relationship betwen economic mobility and the other variables by conducting a hypothesis test that is comparing regression slopes. The slopes of linear regression lines are obtained as follows:

```
# linear regression model assigned to a variable
graduation_m <- lm(Mobility~Graduation, data=dat)
testscore_m <- lm(Mobility~Test_scores, data=dat)
single_m <- lm(Mobility~Single_mothers, data=dat)
social_m <- lm(Mobility~Social_capital, data=dat)

# Table to show the slope for each variable
table_slope <- matrix(c(coef(graduation_m)[2],coef(testscore_m)[2],coef(single_m)[2],
coef(social_m)[2]),ncol=4,byrow=TRUE)
colnames(table_slope) <- c("Graduation", "Test_scores","Single_mothers","Social_capit
al")
rownames(table_slope) <- c("Slope")
table_slope <- as.table(table_slope)
table_slope
```

```
##         Graduation  Test_scores Single_mothers Social_capital
## Slope  0.018856565  0.002613196   -0.688638476    0.021018588
```

The linear regression models fitted onto the graphs of variables look like this:

```
# Plotting each variable
testscore <- ggplot(dat,aes(x=Test_scores,y=Mobility))+geom_point(size=0.6,color="#F7CF08")+geom_smooth(method=lm)+theme(panel.grid.major = element_line(size = 0.3, linetype = 'solid',colour = "grey"),panel.grid.minor = element_blank())

graduation <- ggplot(dat,aes(x=Graduation,y=Mobility))+geom_point(size=0.6,color="#F7CF08")+geom_smooth(method=lm)+theme(axis.title.y=element_blank(),axis.text.y=element_blank(),axis.ticks.y = element_blank(),panel.grid.major = element_line(size = 0.3, linetype = 'solid',colour = "grey"),panel.grid.minor = element_blank())

singlemothers <- ggplot(dat,aes(x=Single_mothers,y=Mobility))+geom_point(size=0.6,color="#08C8F7")+geom_smooth(method=lm)+theme(panel.grid.major = element_line(size = 0.3, linetype = 'solid',colour = "grey"),panel.grid.minor = element_blank())

socialcapital <- ggplot(dat,aes(x=Social_capital,y=Mobility))+geom_point(size=0.6,color="#08C8F7")+geom_smooth(method=lm)+theme(axis.title.y=element_blank(),axis.text.y=element_blank(),axis.ticks.y = element_blank(),panel.grid.major = element_line(size = 0.3, linetype = 'solid',colour = "grey"),panel.grid.minor = element_blank())

# Arranging all the plots into one grid
grid.arrange(testscore,graduation,singlemothers,socialcapital,top = textGrob("Economic Mobility vs Each Variable (linear regression model)",gp=gpar(fontsize=14,font=2)),nrow = 2)
```

```
## Warning: Removed 36 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 36 rows containing missing values (geom_point).
```

```
## Warning: Removed 160 rows containing non-finite values (stat_smooth).
```

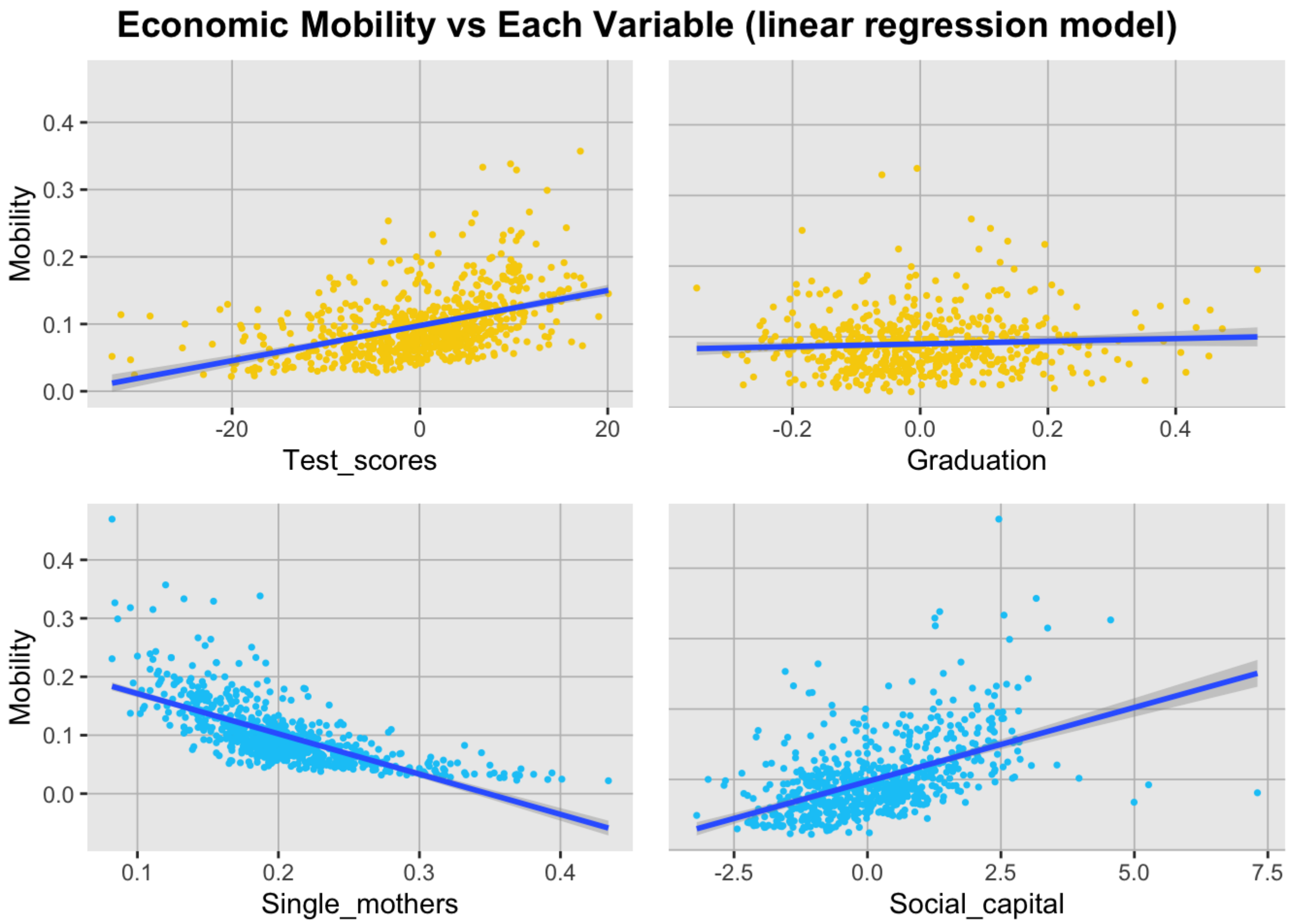
```
## Warning: Removed 160 rows containing missing values (geom_point).
```

```
## Warning: Removed 12 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 12 rows containing missing values (geom_point).
```

```
## Warning: Removed 29 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 29 rows containing missing values (geom_point).
```



Input x=Q1 and x=Q3 in the regression model for each variable. Get the differences betwen y=b0+Q1x and y=b0+Q3x to observe the predicted mobility change and compare by variables.

Get 1st and 3rd quartiles

```
# Making a table to show quartile values for x-variables
table_var <- matrix(c(summary(dat$Graduation)[2],summary(dat$Graduation)[5],summary(d
at$Test_scores)[2],summary(dat$Test_scores)[5],summary(dat$Single_mothers)[2],summary
(dat$Single_mothers)[5],summary(dat$Social_capital)[2],summary(dat$Social_capital)[5
]),ncol=2,byrow=TRUE)

colnames(table_var) <- c("Quartile 1", "Quartile 3")
rownames(table_var) <- c("Graduation","Test_scores","Single_mothers","Social_capital"
)
table_var <- as.table(table_var)
table_var
```

```
##                Quartile 1 Quartile 3
## Graduation       -0.09700    0.08300
## Test_scores      -4.29300    5.55400
## Single_mothers    0.17100    0.22600
## Social_capital   -0.76550    0.96525
```

Input x=Q1 and x=Q3 to each regression model and predict the values for Y1 and Y2, which are values of mobility in our case.

```
l <- list(
predict(graduation_m, data.frame(Graduation = c(-0.097,0.083))),
predict(testscore_m, data.frame(Test_scores = c(-4.293,5.554))),
predict(single_m, data.frame(Single_mothers = c(0.171,0.226))),
predict(social_m, data.frame(Social_capital = c(-0.7655,0.9653))))

head(l)
```

```
## [[1]]
##          1          2 
## 0.08807152 0.09146570 
## 
## [[2]]
##          1          2 
## 0.08654371 0.11227585 
## 
## [[3]]
##         1         2 
## 0.1222889 0.0844138 
## 
## [[4]]
##          1          2 
## 0.08108441 0.11746338
```

Calculate the difference between each pair of output values.

This is to look at the **predicted change of mobility between the range of x=Q3 and x=Q1.**

```
# Predicted change of mobility for each variable
lst <- list((0.09146570 - 0.0880715),
(0.11227585 - 0.08654371),
(0.0844138- 0.1222889),
(0.11746338 - 0.08108441))
head(lst)
```

```
## [[1]]
## [1] 0.0033942
## 
## [[2]]
## [1] 0.02573214
## 
## [[3]]
## [1] -0.0378751
## 
## [[4]]
## [1] 0.03637897
```

Conclusion

The P-value test has shown that P-values for 3 out of 4 variables (Test_score, Single_mothers, Social_capital) are statistically significant. This was problematic because we couldn't determine which variables had higher association with mobility. This led us to conduct another test, which was the regression slope test.

As mentioned before, if there is a significant linear relationship between the independent variable X and the dependent variable Y, the slope will not equal zero. Given that simple regression line model is **Y=B0+B1X** :

Ho: B1 = 0

Ha: B1 ≠ 0

The null hypothesis states that the slope is equal to zero, and the alternative hypothesis states that the slope is not equal to zero. **If we find that the slope of the regression line is significantly different from zero, we will conclude that there is a significant relationship between the independent and dependent variables.**

Therefore, to see how different the slope of the regression line is from zero, we need to check the absolute value of the slope. Now, the slope of 'Single_mothers' is the only negative one, so we will just take *abs(-0.6886385)* and compare.

Sorted by descending order, the slopes are arranged as such: **Single_mother > Social_capital > Graduation > Test_scores**

By factor groups, we can see that it is: **Social factors > Educational factors**

When we compare the *predicted change of mobility between the range of x=Q3 and x=Q1*, the values sorted for each variable were as follows: **Single_mother > Social_capital > Test_scores > Graduation**

Again, by factor groups, we can still see that it is: **Social factors > Educational factors**

This is in line with what we have previously seen in P-value test and we can observe that social factors showed higher association in all analysis. All in all, we can reject the null hypothesis that is *"Educational factors more strongly associated with higher economic mobility than social factors"* and accept the alternative hypothesis.

The main limitation to this exploration was the number of variables tested for each factor. For this exploration, we had two educational factors and two social factors each and determined what factors had stronger association with, in other words, played greater role in economic mobility. However, I believe that more variables could be taken into account to test the hypothesis because there many more variables than just two in the aspect of education and society. The limitation could have occurred due to many possible reasons such as the time limitation, computational power, limited capacity to include on a project , and so on. For future improvement, higher number of variables could be taken from a bigger dataset to yield a more accurate result about the relationship between factors.

Social factors are more strongly associated with economic mobility than educational factors