Data 621 Final Project

Omar Pineda, Jeff Littlejohn, Sergio Ortega Cruz, Chester Poon, Simon Ustoyev

Contents

Data 621 Final Project

Abstract

This project examines which demographic and geographic factors are associated with educational and social mobility. Using data included in the AER R package from a 1980 survey of 1100 high school seniors that included a follow-up six years later to gauge the level of education that had been obtained, we constructed a regression model that predicts educational attainment based on a selection of attributes included in the dataset. To start, we performed exploratory data analysis to understand the data and detect issues problematic with regression such as multicollinearity. Next, we constructed a simple multiple linear regression model using an achievement test score, average tuition for a four-year college in the student's state, a student's home distance from a four-year college, and whether their high school was in an urban area. We checked the model's assumptions, including testing for influential outliers, before reviewing model diagnostics. Next, we created a new model that included an interaction term between the distance and urban variables. This model yielded similar results. In the end, we found that an achievement test score had the highest positive correlation with educational attainment, while average state school tuition had the largest negative association. Distance or whether a high school was in an urban area displayed little correlation.

Keywords

Demographic, Education, Attainment, Scores, Equality.

Introduction

Despite the national narrative that America affords its inhabitants an unprecedented land of opportunity,intergenerational social mobility, defined as the likelihood that a child born to parents in the bottom fifth of the income distribution reaches the top fifth, is higher in many other advanced countries.Fewer than eight percent of Americans born in the bottom 20% of the income distribution reach the top 20%, whereas more than 13 percent of Canadians do. However, as Harvard Economist Raj Chetty has demonstrated, significant differences in upward mobility rates exist across the United States (Chetty el al. 2016). "In this country, of all countries, a person's zip code shouldn't decide their destiny," President Barack Obama said in 2015(Kaufman). We wish to understand which demographic and geographic factors are associated with social mobility. While educational attainment is not a perfect proxy for income, we will use it as our outcome variable, our metric of interest. Although our data set captures a survey from more than 30 years ago, we think the construction of a model that predicts such a metric is nonetheless a worthwhile endeavor.

Literature Review

We reviewed three papers for our review that look at social mobility where each study had a slightly different approach at analyzing the different independent variables that affect intergenerational social and economic outcome. Broadly, the first paper reviewed observed economic outcomes children in families that moved from a poorer to a better income neighborhood. The second paper reviewed researched the interplay of a child's gender from poorer neighborhoods versus higher income neighborhoods and those children's adult economic outcome. While these two papers studies were focused on communities in the United States, our third paper reviewed looked at differences in Denmark that draw heavily from the theories of sociologist Pierre Bourdieu on the concept of "cultural capital", which is the theory that individuals can possess a form of "capital" quantified by the amount of knowledge and behavior (culture) that promote social mobility in a socially stratified society (Mollegaard & Jager 2015).

The first paper we looked at, The Impacts of Neighborhoods on Intergenerational Mobility I:Childhood Exposure Effects found that families that moved from a lower to a better income area had children that had greater social mobility than those that stayed. There was a positive improvement in the child's adult income, which rose at a rate of about 4% that was proportional the number of years spent growing up in the higher income area (Chetty & Hendren 2016). The dataset composed of federal income tax records from 1996 through 2012 and focused on families with children born between 1980 and 1988 and moved across neighborhoods between 1997 and 2010. This paper reconciles conflicting papers by introducing the number of years growing up in the better neighborhood. Interestingly, a symmetrical finding was found for families moving from higher to lower income neighborhoods in which the children of those families had adult incomes that reduced 4% per years lived in the lower income neighborhood (Chetty & Hendren 2016). The key variables are fairly different from the ones we are using from the dataset in AER. For this paper, those variables are parents income, parent location, child's adult earned income, teenage birth, marriage, educational attainment, and the childs employment at adulthood. While the primary independent variable is the childs adult earned income, ours is educational attainment. Our analysis is also more focused on key variables that are characteristic of a neighborhood, whereas this study is more focused on those associated with the family unit. Our dataset also does not contain any interventional events with each family as this study has. A way to take our study further would be to complete the dataset in AER with family income as well as the childs adult earned income. The second paper we reviewed, Childhood Environment and Gender Gaps in Adulthood takes a deeper look into the effects of gender and social mobility. Like the first paper we reviewed, the data was drawn from tax records for families with children born in the 1980s. The major finding of this study found that the traditional gender gap in employment and income earnings are reversed in poor families, particularly in high-poverty neighborhoods. Boys from these families are less likely to work in adulthood and the issue is compounded with single-parent families. The authors write: Low-income boys who grow up in high-poverty, high-minority areas work less than girls. These areas also have higher rates of crime, consistent with a model in which boys with lower latent earnings potential who grow up in environments of concentrated poverty switch from the formal labor market to crime or other illicit activities (Chetty et al. 2016).

Though the paper does a comprehensive analysis on the impact of environment based on economic class and the interplay of gender, a deeper analysis that would include race as a more prominent factor would strengthen this paper significantly. This is especially true if we approach the analysis of these factors (race, class, gender, etc.) as observations that are not mutually exclusive. The last two papers we reviewed included clear and concrete independent variables for their analyses, the third paper we reviewed is different in its approach in many ways. The data quantifies three different forms of capital as measured in three major categories: economic, social and cultural. (Mollegaard & Jager 2015). Economic capital is self-explanatory in that it's a measurement of income and assets. Cultural capital is measured in less concrete terms such as years of educational attainment of the parents and grandparents, subscriptions to news outlets, and other things. Social capital is interesting in that it is best understood as a kind of network analysis manifesting as professional networks, friendships, familial connections, etc. The paper also addresses the intergenerational effects on educational attainment as far as two-generations, whereas most studies review a single generation prior. The findings of the paper found that in Denmark, educational attainment or more academically driven educational attainment depended more on cultural capital than of social/economic capital (Mollegaard & Jager 2015). While these measures are fascinating and the results of the study show that cultural capital measures are certainly valid and useful independent variables, it's difficult to apply an appropriate weight to some of the cultural capital examples. Generally, applying the appropriate weight of the impact of various aspects of cultural capital are difficult to accomplish and to justify. Further research is needed to properly assess the impact of specific variables associated with cultural capital to apply a stronger model in predicting educational outcomes. For our project, measuring social and cultural capital appropriately will be unfeasible. Overall, a common measure of all reviewed papers focuses on parental income or economic means to assess the probability of social upward mobility of the childs adulthood income. This "intergenerational mobility" is the common theme as a measure of social mobility. For our project, while we are not focused on income as a measure, our project is focused primarily on educational attainment to serve as a proxy for social mobility. We are also looking to identify the strongest predictors for upward social mobility as measured in educational attainment.

Methodology

Experimentation and Results:

The research question

Are the numbers of years of education beyond highschool related to the relationship of elements such as exam scores, tuition paid, distance they had to travel to get to school and/or if their highschool was or not situated in and urban/rural environment?

Understanding the Data Set

The first data set we will use for this R Guide is CollegeDistance from the R package AER. This is a cross-sectional data set from a survey conducted by the Department of Education in 1980, with a follow-up in 1986, containing 14 variables relating to the characteristics of the students surveyed for this data set, their families and the area in which they live.

Though the data set contains 14 different variables, we will only use score, the achievement test score obtained during the student's senior year of high school, urban, whether the student's high school is located in an urban area, distance, the distance the student lives from a 4-year college (in 10's of miles), tuition, the average 4-year college tuition in the student's state (in 1000's of dollars), and education, the number of years of education attained 6 years after high school graduation, in any of the models we create. Of these chosen variables, we can see that score,distance, tuition and education are quantitative and urban is categorical.

Table 1: Data summary

Name	CollegeDistance
Number of rows	4739
Number of columns	14
Column type frequency:	
factor	8
numeric	6
Group variables	None

Variable type: factor

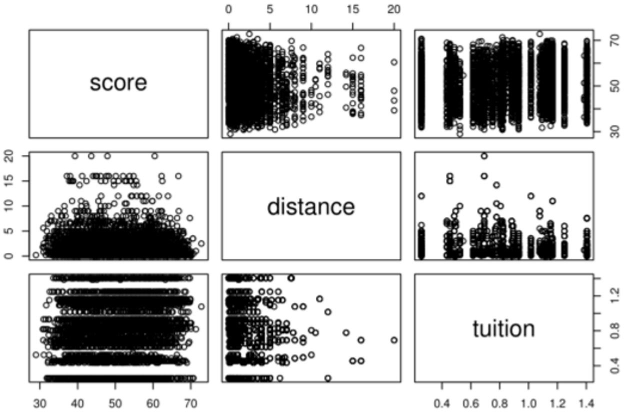
skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
gender	0	1	FALSE	2	fem: 2600, mal: 2139
ethnicity	0	1	FALSE	3	oth: 3050, his: 903, afa: 786
fcollege	0	1	FALSE	2	no: 3753, yes: 986
mcollege	0	1	FALSE	2	no: 4088, yes: 651
home	0	1	FALSE	2	yes: 3887, no: 852
urban	0	1	FALSE	2	no: 3635, yes: 1104
income	0	1	FALSE	2	low: 3374, hig: 1365
region	0	1	FALSE	2	oth: 3796, wes: 943

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
score	0	1	50.89	8.70	28.95	43.92	51.19	57.77	72.81	â â â â â
unemp	0	1	7.60	2.76	1.40	5.90	7.10	8.90	24.90	â â â â â
wage	0	1	9.50	1.34	6.59	8.85	9.68	10.15	12.96	â â â â â
distance	0	1	1.80	2.30	0.00	0.40	1.00	2.50	20.00	â â â â â
tuition	0	1	0.81	0.34	0.26	0.48	0.82	1.13	1.40	â â â â â
education	0	1	13.81	1.79	12.00	12.00	13.00	16.00	18.00	â â â â â

Checking Multicollinearity

One of the very first things we want to do before making any model is check for multicollinearity between our quantitative predictor variables. Multicollinearity is a problem because it will inflate the standard error of a model as well as make the parameter estimates inconsistent.



Looking at our plot, it does not appear that any of our quantitative predictor variables are highly correlated, or have a strong linear relationship with one another.

Correlation Matrices

Just to be certain there is no strong correlation, let's calculate the correlation matrix of the three variables using cor() which takes the variables of interest as its arguments, similar to plot()

```
##                score    distance    tuition
## score     1.00000000 -0.06797927  0.1298585
## distance -0.06797927  1.00000000 -0.1009806
## tuition   0.12985848 -0.10098058  1.0000000
```

We see that the pairwise correlations between our quantitative predictor variables are very low. We only consider multicollinearity to be a problem when the absolute value of the correlation is greater than .9, so we would say that there is no multicollinearity issue in this model.

Variance Inflation Factor

I'll use a third way to detect multicollinearity in a model. We consider the variance inflation factor (VIF) of a variable by first making a model using that variable as the response and all other predictor variables in the original model as the predictors. For example, to calculate the VIF of achievement test score we need to create this linear model:

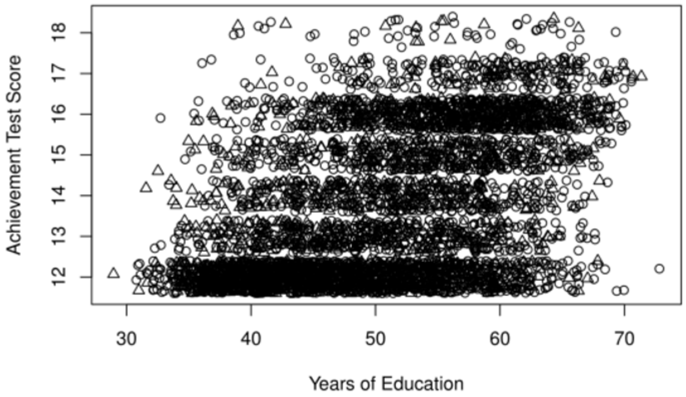
Then, we need to plug the R-squared value of that model into our VIF equation, $VIF = \frac{11}{1-r^2}$

```
## [1] 1.020309
```

So the VIF of achievement test score for this model is around 1.02. This is great news since we only consider multicollinearity to be an issue when VIF>10.

Visualizing the Data

Now that we know there is no evidence of multicollinearity between our predictor variables, we can move on to visualizing some of the data. We can plot the student's achievement test score, the quantitative response, versus the number of years of education they have attained, the quantitative predictor, varying the shape of their plotted point by whether or not their high school is located in an urban area, the categorical predictor, because this satisfies our variable requirements. The circles indicate that the student represented went to a high school in an urban area. The triangles indicate that the student represented went to a high school in a non-urban area



There doesn't appear to be a distinct separation of the data points. This might lead us to believe that there is not a significant difference between the educational attainment of students at urban and non-urban schools, but we can confirm whether or not this is true using a test later on.

Multiple Linear Regression

Building the Model

Criteria for Comparing all Kinds of Models

There are many types of criteria that we can use to determine whether or not one model of a set of data is better than another model of that same set of data. First, when comparing any two models we can use the Akaike Information Criteria (AIC) or Mallow's Cp.

For Comparing Nested Models

If we are comparing nested models, i.e. one model's predictors are a subset of the other model's predictors, we have additional comparison options. For comparing nested models, we can use the p-values from t-tests or partial F-tests, only choosing the more complex model if the p-value of the t-test or F-test is below .05. We can also use adjusted R-squared as a comparison tool; however, I would discourage anyone from using adjusted R-squared as their comparison criteria because the other methods of comparison are more rigorous.

Creating the model

Stepwise Selection

Now, on to the topic of actually building the model. First, we can discuss the idea of building a model in a stepwise fashion. We will be using the backwards elimination method to create a model that predicts educational attainment 6 years after graduation using all or a subset of the variables discussed earlier and AIC as my model comparison criteria. I will specify the simplest model as a model using only the intercept to model educational attainment 6 years after graduation.

```
## Warning: package 'MASS' was built under R version 3.5.3
```

```
## Start:  AIC=4339.19
## education ~ score + urban + distance + tuition
## 
##            Df Sum of Sq   RSS    AIC
## - urban     1       0.5 11815 4337.4
## <none>                  11815 4339.2
## - tuition   1      10.8 11826 4341.5
## - distance  1      53.3 11868 4358.5
## - score     1    3178.2 14993 5466.2
## 
## Step:  AIC=4337.39
## education ~ score + distance + tuition
## 
##            Df Sum of Sq   RSS    AIC
## <none>                  11815 4337.4
## - tuition   1        11 11826 4339.8
## - distance  1        62 11877 4360.2
## - score     1      3205 15020 5472.8
```

```
## 
## Call:
## lm(formula = education ~ score + distance + tuition, data = CollegeDistance)
## 
## Coefficients:
## (Intercept)        score     distance      tuition  
##     9.15680      0.09547     -0.05013     -0.14369
```

We can see from the output that our final model actually contains all of the chosen variables, except for whether or not a student's high school is located in an urban area, as the optimal way of predicting that student's educational attainment 6 years after graduation. One thing to note about this process is that, although the two models' AIC differ by less than 10, the chosen model is the model with fewer predictor variables because of the necessary balance between accuracy and complexity that AIC uses.

All subsets

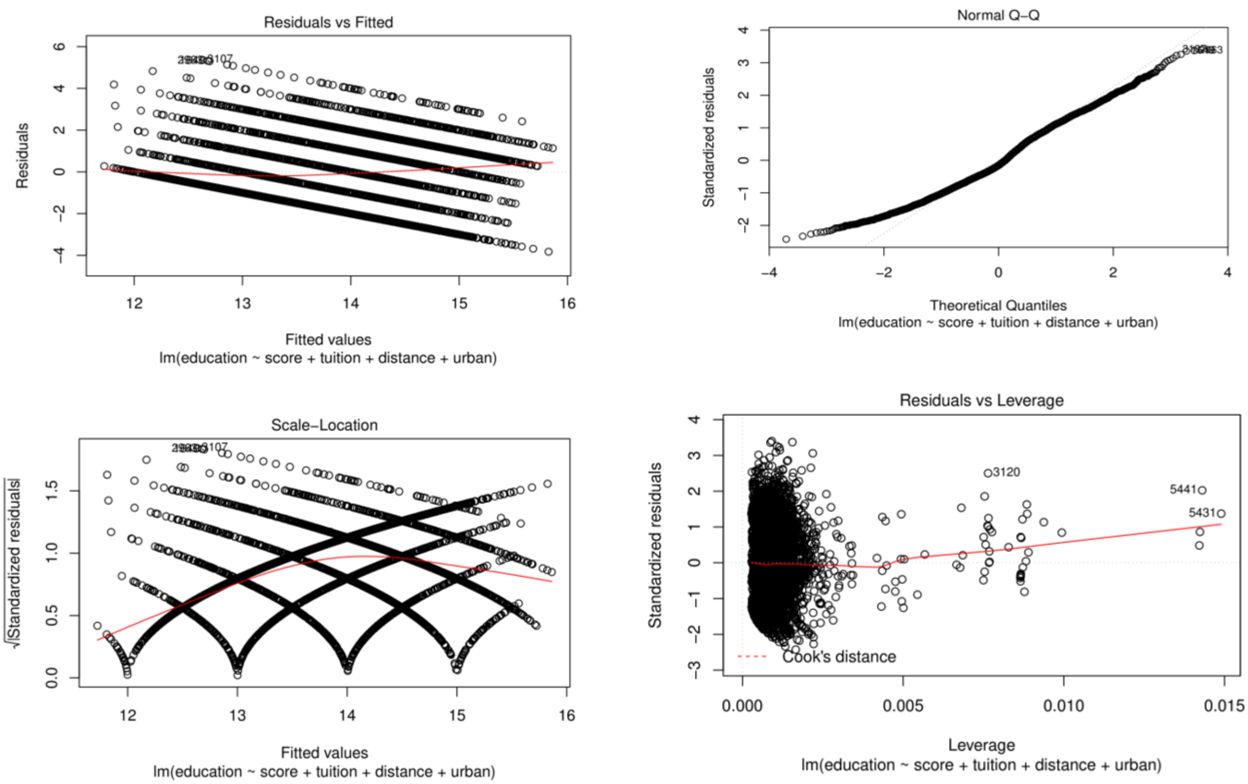
Lastly, we have the all subsets method. This method requires that you calculate all of the AIC values, or Mallow's Cp's, for all possible models that can be created from any subset of the variables you are considering for the model.

Creating a New Model

We've looked at our data set, gone through a short example of how one might visualize a relatively simple multiple linear regression situation and discussed some sophisticated techniques of model building as well as how to compare models. Now, we can create a multiple linear regression model manually. For this model, We will use achievement test score, average tuition for a four-year college in the student's state, distance from a four year college and whether or not the student's high school was located in an urban area to predict the number of years of education the student had attained six years after graduation.

Checking Model Assumptions

At this point, we need to check the assumptions of our model to be sure that the model provides accurate results.



The first plot is the residual plot, a comparison of the residuals of our model against the fitted values produced by our model, Although the residuals do not seem to be evenly spread around 0 for all fitted values, the range of the residuals at each fitted value appears to be roughly the same, so we can conclude there is no evidence of heteroskedasticity. The next plot is the QQ-plot. Though most of the points seem to fall on the line which indicates that our residuals come from a normal distribution, there are some points that stray from the line in the lower and upper quantiles of the plot. It is possible that these points do not come from a normal distribution, but most of our points seem to come from a normal distribution so there is not a lot to worry about here.The third plot created is the scale-location plot, from this plot, we can see that there is likely a U-shaped trend in our residuals.

Outliers

This plot graphs the standardized residuals against their leverage. It also includes the Cook's distance boundaries. Any point outside of those boundaries would be an outlier in the x direction. Since we cannot even see the boundaries on our plot, we can conclude that we have no outliers.

Indicator Variables

```
##
## Call:
## lm(formula = education ~ score + tuition + distance + urban,
##     data = CollegeDistance)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.826 -1.181 -0.246  1.219  5.367
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.141015   0.148905  61.388  < 2e-16 ***
## score        0.095596   0.002679  35.686  < 2e-16 ***
## tuition     -0.142627   0.068517  -2.082   0.0374 *
## distance    -0.048723   0.010539  -4.623 3.88e-06 ***
## urbanyes     0.025619   0.057090   0.449   0.6536
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.58 on 4734 degrees of freedom
## Multiple R-squared:  0.221,  Adjusted R-squared:  0.2203
## F-statistic: 335.7 on 4 and 4734 DF,  p-value: < 2.2e-16
```

We should first take note of how our categorical variable urban is written in our summary of the model. In the model, this variable is treated as an indicator variable meaning that it can take one of two values, either 0 or 1. In the summary, this variable is labeled as urbanyes.The variable would need the number of categories minus one indicator functions.

Discussion and Conclusions:

The Regression Equation

From our summary, we can also get the multiple linear regression equations:

$\widehat{EducationalAttainment} = 9.141 + .0956 \cdot AchievementScore - .1426 \cdot Tuition - .0487 \cdot Distance + .0256 \cdot I(urban = yes)$

Since we have an indicator variable, urban, in our equation, this equation can actually be broken down into two separate equations, one for if the student's school is located in an urban area and another for if the student's school is in a non-urban area. For students going to school in urban areas:

$\widehat{EducationalAttainment} = (9.141 + .0256) + .0956 \cdot AchievementScore - .1426 \cdot Tuition - .0487 \cdot Distance$

For students going to school in non-urban areas:

$\widehat{EducationalAttainment} = 9.141 + .0956 \cdot AchievmentScore - .1426 \cdot Tuition - .0487 \cdot CollegeDistance$

Notice that the only change in the two equations is the intercept.

Interpreting Regression Coefficients

For every increase in achievement test score by 1 point, the predicted years of education of a student increases by .0956 years when that student's state's average tuition for a 4-year college, distance from a 4-year college and location of school (urban or non-urban) all stay the same. You might notice that the only change in the interpretation that we made from the interpretation for simple linear regression is that you must state that all of the other predictors in the model are being held constant.

We can also interpret the coefficient of our indicator function in a similar way. For example, the predicted years of education of a student increases by .0256 when their high school is located in an urban location as opposed to a non-urban location when their achievement test score, average state tuition for a four-year college and distance from a four-year college all remain the same.

Finally, thinking about the intercept in context, we can see that the intercept wouldn't really make sense to interpret because even a student that scored a zero on their achievement test, had an average state tuition for a 4-year college of zero dollars, lived zero miles from a 4-year college and went to a non-urban high school would still have at least 12 years of education since they had to have graduated in order to be part of this data set and our intercept is less than 12. We cannot use this model to determine anything about students who did not graduate.

Confidence Intervals for Regression Coefficients

We recognize that these regression coefficient values given by our summary are simply point estimates so we'll need to create confidence intervals for them as well.

```
##                   2.5 %       97.5 %
## (Intercept)  8.84909261  9.432937106
## score        0.09034461  0.100848185
## tuition     -0.27695296 -0.008301033
## distance    -0.06938312 -0.028062221
## urbanyes    -0.08630305  0.137541361
```

We interpret the confidence interval for achievement test score, a quantitative variable, like this: we are 95% confident that an increase by 1 point in achievement test score will increase the number of years of education a student achieves 6 years after high school graduation by between .09 and .1 years on average, all other variables held constant. We interpret the confidence interval for our indicator variable, a categorical variable, like this: we are 95% confident that a student going to high school in an urban area will achieve between -.086 and .138 years of education more than a student going to a high school in a non-urban area on average, all other variables held constant. Again, we cannot interpret our intercept because it does not make sense within the limits of our data set.

T-Tests

From our summary, we can also determine the significance of each of our predictor variables through the results of a t-test. The results of the t-tests tell us that all of the variables in our model are significant predictors of a student's number of years of education except whether their high school is in an urban or non-urban area since that is the only p-value greater than .05. This means we can remove our urban variable from the model because it has no significant relationship with number of years of education that a student achieves 6 years after high school graduation. This is a similar conclusion to the one that was drawn from our automated model building procedure earlier.

Making Predictions

Next, we can make some predictions using our model. We can use our regression equation to predict the years of education a student would have attained if they had a test score of 52, the average college tuition for their state was $900, they lived five miles from a four-year college and they attended an urban high school.

```
## [1] 13.98511
```

From our equation, we can see that this student would have likely pursued just shy of 2 years of education after graduating high school. Now, let's compare that value with a similar student who only differed by where they went to high school. This student did not go to high school in an urban area.

```
## [1] 13.95951
```

From our equation, it seems that this student would have also pursued just shy of 2 years of education after graduating high school. However, take note of the difference between the two predicted years of education.

```
## [1] 0.0256
```

The difference between the two educational attainment levels is the value of the regression coefficient estimated for the urban indicator variable. Since this coefficient is small, the difference between the two students is also small.We know that providing a point estimate is usually not adequate when providing a prediction based on a regression model because it gives an answer that is very specific and likely not exactly correct. Therefore, we might want to make a prediction interval.

The only difference for the multiple linear regression models is that the new data frame must contain specified values for all of the variables in your regression model. So, our new data frame, corresponding to the example student that attended a high school in an urban area, would look like this:

```
##   score tuition distance urban
## 1    52     0.9      0.5   yes
```

Now, we can use this data frame to create a 95% confidence interval for the student with a high school in an urban area that we made a point prediction for previously.

```
##        fit      lwr      upr
## 1 13.98492 10.88636 17.08349
```

We can see that the point estimate given by this prediction interval is slightly different than the value we calculated, however, this can be attributed to a rounding error. We see that our 95% prediction interval is [10.886, 17.083], thus, I am 95% confident that an individual student with an achievement score of 52, an average state tuition for four-year colleges of $900, living five miles from a four-year college and going to a high school in an urban area will have achieved a total of between 10.886 and 17.083 years of education 6 years after graduating from high school.

Confidence Intervals

In addition to prediction intervals, we can also create confidence intervals for the mean years of education students with specific characteristics might attain 6 years after high school graduation.

```
##        fit      lwr      upr
## 1 13.98492 13.89019 14.07965
```

From this interval we can say we are 95% confident that the mean number of years of education a student would attain 6 years after high school graduation, given that they had an achievement test score of 52, an average state tuition for four-year colleges of $900, live five miles from the nearest four-year college and went to a high school in an urban area, is between 13.89 and 14.08 years.

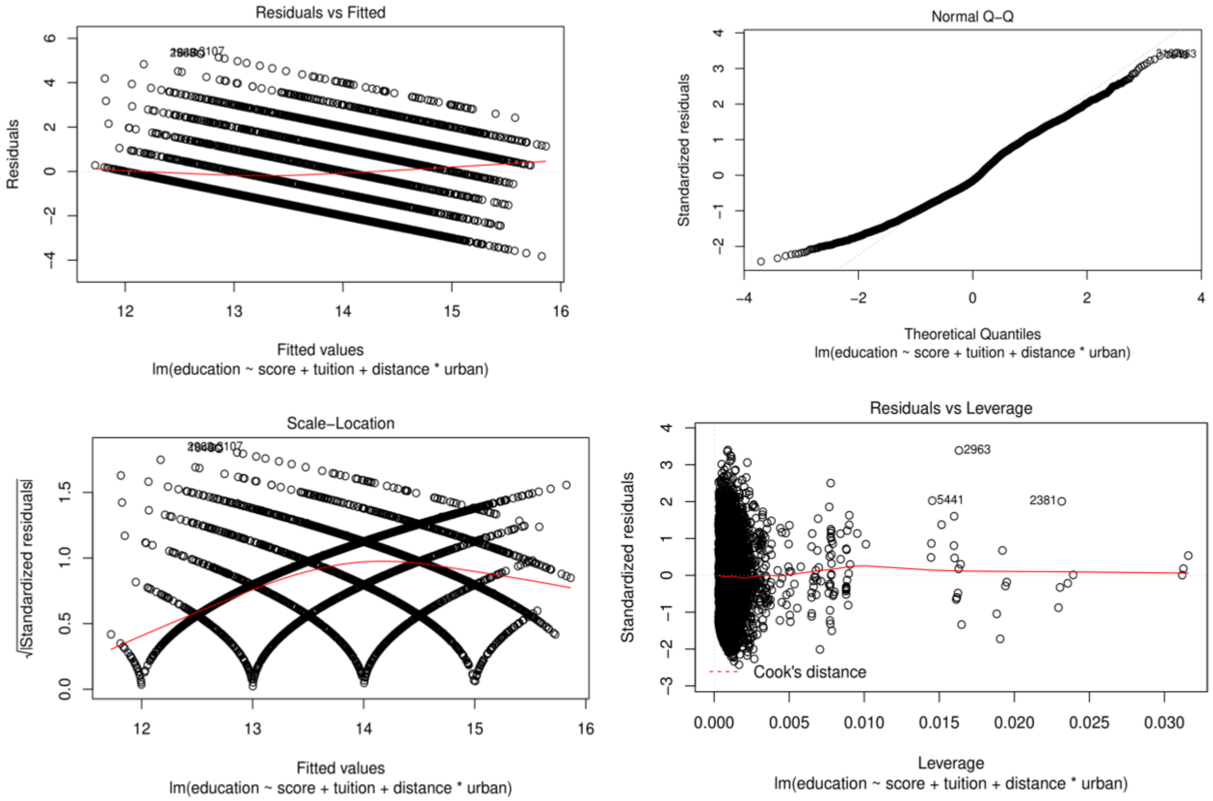
Adding Interaction Terms

Creating a New Model

Another attribute we can add to our models is interaction terms. For this model we will use an interaction between our distance distance and urban variables. By that we mean that the relationship between the education level that a student achieves and the distance they are from a four-year college depends on whether the student's high school is located in an urban area or not.

Checking Model Assumptions

Since we have created a second model, we need to check our model assumptions again.



We see a similar residual plot to our first residual plot, with the same over-/underprediction problem. Again, there appears to be no strong evidence for heteroskedasticity. Finally, we see no outliers on this plot. Additionally, the other 3 diagnostic plots look very similar to the diagnostic plots generated for our first model.

The Regression Equation

```
##
## Call:
## lm(formula = education ~ score + tuition + distance * urban,
##     data = CollegeDistance)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8258 -1.1814 -0.2471  1.2193  5.3674
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          9.141319   0.149012  61.346  < 2e-16 ***
## score                0.095592   0.002680  35.665  < 2e-16 ***
## tuition           -0.142507   0.068555  -2.079   0.0377 *
## distance          -0.048802   0.010627  -4.592  4.5e-06 ***
## urbanyes           0.022658   0.076409   0.297   0.7668
## distance:urbanyes  0.004735   0.081204   0.058   0.9535
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.58 on 4733 degrees of freedom
## Multiple R-squared:  0.221,  Adjusted R-squared:  0.2201
## F-statistic: 268.5 on 5 and 4733 DF,  p-value: < 2.2e-16
```

We see from our summary output that our new regression equation looks like this:

$\widehat{EducationalAttainment} = 9.141 + .096 \cdot AchievementScore - .143 \cdot Tuition - .049 \cdot Distance + .026 \cdot I(urban = yes) + .005 \cdot Distance \cdot I(urban = yes)$

Again, we can split our regression equation into two equations, one for a student going to a high school in an urban area and one for a student going to a high school in a non-urban area. For student's going to school in urban areas:

$\widehat{EducationalAttainment} = (9.141 + .026) + .096 \cdot AchievementScore - .143 \cdot Tuition - (.049 - .005) \cdot Distance$

For student's not going to school in urban areas:

$\widehat{EducationalAttainment} = 9.141 + .096 \cdot AchievementScore - .143 \cdot Tuition - .049 \cdot Distance$

This time we see that the difference between the two equations is not just the intercept, but also the coefficient for the distance a student lives from a four-year college since our urban variable effects both of those locations in the equation.

The T-test Interpretation

Our summary also tells us the results of the t-test performed for our interaction term (distance:urbanyes). We can see in the table that the t-statistic is .058 and the corresponding p-value is .9535. This means the relationship between a student's number of years of education and distance from a four-year college does not significantly depend on whether that student's high school is located in an urban area.

F-tests

Another test that our summary output displays is the results of an F-test for the model. We can see from our summary output that the F-statistic for this model is 268.5 and this is compared to an F-table with 5 and 4733 degrees of freedom to obtain a p-value of less than $2.2 \cdot 10^{-16}$. This means that at least one of the variables in our model has a significant relationship with our predictor.

Partial F-tests

A final test we can look at to understand the significance of the variables in our model is the partial F-test. The partial F-test test the null hypothesis that $\beta_m = \ldots = \beta_p = 0$, a subset of your regression coefficients is equal to 0, against the alternative hypothesis that at least one β from the subset of βs is not equal to 0.To run the test, we will use the command anova(complete.model, reduced.model).

```
## Analysis of Variance Table
##
## Model 1: education ~ score + tuition + distance + urban
## Model 2: education ~ score + tuition + distance * urban
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1   4734 11815
## 2   4733 11815  1  0.008489 0.0034 0.9535
```

Our output tells us that our partial F-statistic is .0034 and, from that, we get a p-value of .9535. This p-value is not significant at even a high significance level of .1 so we can conclude that the relationship between the education level that a student achieves and the distance they are from a four-year college does not depend on whether the student's high school is located in an urban area or not.

APPENDIX

References

Chetty, R., & Hendren, N. (2016). The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects. doi: 10.3386/w23001

Chetty, R., Hendren, N., Lin, F., Majerovitz, J., & Scuderi, B. (2016). Childhood Environment and Gender Gaps in Adulthood. doi: 10.3386/w21936

MÃ¸llegaard, S., & JÃ¦ger, M. M. (2015). The effect of grandparentsâ economic, cultural, and social capital on grandchildrens educational success. Research in Social Stratification and Mobility, 42, 11â “19. doi: 10.1016/j.rssm.2015.06.004

Kaufman, Greg (2015, December 17). Why Achieving the American Dream Depends on Your Zip Code. Talk Poverty. Retrieved from https://talkpoverty.org/2015/12/17/american-dream-zip-codes-affordable-housing/

Kleiber C, Zeileis A (2008). Applied Econometrics with R. Springer-Verlag, New York. ISBN 978-0-387-77316-2, https://CRAN.R-project.org/package=AER.

Code

```
knitr::opts_chunk$set(echo = TRUE)
library (AER)
library (reshape2)
library (dplyr)
library(ggplot2)
library(psych)
library(skimr)
data(CollegeDistance)
#psych::describe(CollegeDistance)
skimr::skim(CollegeDistance)
#Page Limit
#plot(CollegeDistance[c(3,10, 11)])
knitr::include_graphics("final3.png")
cor(CollegeDistance[c(3,10, 11)])

score <- lm(score ~ distance + tuition, data = CollegeDistance)
1/(1-summary(score)$r.squared)
#newscore <- jitter(CollegeDistance$score, factor = 2)
#newed <- jitter(CollegeDistance$education, factor = 2)
#plot(newscore, newed, xlab = "Years of Education", ylab = "Achievement Test Score", pch = as.numeric(C
knitr::include_graphics("final4.png")
library(MASS)
starting.model <- lm(education ~ score + urban + distance + tuition, data = CollegeDistance)
simple.model <- lm(education ~ 1, data = CollegeDistance)
stepAIC(starting.model, scope = list(upper = starting.model, lower = simple.model), direction = "backwa

schoolmod <- lm(education ~ score + tuition + distance + urban, data = CollegeDistance)
#plot(schoolmod)
knitr::include_graphics("final1.png")
summary(schoolmod)
confint(schoolmod)
u <- (9.141+.0256) + .0956*52 - .1426*.900 - .0487*.5
u
```

```
nu <- 9.141 + .0956*52 - .1426*.900 - .0487*.5
nu
abs(u-nu)

newdata <- data.frame(score = 52, tuition = .9, distance = .5, urban = 'yes')
newdata
predict(schoolmod, newdata, interval = "predict")
predict(schoolmod, newdata, interval = "confidence")

schoolmod2 <- lm(education ~ score + tuition + distance*urban, data = CollegeDistance)

#plot(schoolmod2)taking out to adjust to page limit
knitr::include_graphics("final2.png")
summary(schoolmod2)
anova(schoolmod, schoolmod2)
```