Regression Analysis Project on

MEDIAN HOUSEHOLD INCOMES OF NEW YORK CITY NEIGHBORHOODS

Anustha Shrestha
Baruch College
December 21, 2018

TABLE OF CONTENTS

*Photo Credits: Anustha Shrestha, New York City, 2015

Chapter 1. Introduction

1.1.Topic

Studying the distribution of income across various geographies has always been a subject of interest to the field of economics. If income is taken as one of the indicators of well-being of the society, then by studying the characteristics of geographical area and how it relates to the general household income of the area, we can understand the various social, demographic and economic influences on well-being of the population residing in the geographical area. Furthermore, understanding the various factors that relates to the income distribution of neighborhoods helps influence policies and programs targeted towards the well-being of neighborhoods.

New York City is one of the largest and the most diverse metropolitan cities in the world. The diversity and the richness of cultures in its neighborhoods doesn't come without the cost of disparate income between them. New York also has income inequality spread across its neighborhoods. Since New York neighborhoods are so diverse and demonstrate such a high range of1. income distribution ranging from neighborhoods with median household income of $21,000 to neighborhoods with median household income of $155,000, the study will be very impactful.

Therefore, in this project, I try to understand the median household income of the 195 neighborhoods in New York City and understand how household incomes relate to various demographic and socio-economic characteristics of those neighborhoods, while exploring various methods and topics related to simple regression and multiple regression analysis.

1.2 Data Source

The data used for the project is the American Community Survey 5-Years Estimate for New York City Neighborhood Tabulation Areas (NTA). These ACS tables (Economic and Social) have been retrieved from New York City Department of Planning website below:

https://www1.nyc.gov/site/planning/data-maps/nyc-population/american-community-survey.page

1.3 Variables

There are 195 Neighborhood Tabulation Areas in New York City (including the 5 boroughs). Each of these NTAs is as an observation. However, five neighborhoods, *Riker's Island, Airport, park-cemetery-etc-Queens, park-cemetery-etc-Staten Island, and park-cemetery-etc-Manhattan* did not have complete data; therefore, they were removed from the list. It should also be noted that these neighborhoods, as classified by New York City, aren't exactly residential neighborhoods; they include parks, cemeteries, correctional facilities and airports. Data from these neighborhoods are not going to be of much help for the scope of our project. The final number of observations used for the study, n = 190.

Out of all the measures available in the American Community Survey report, the following variables have been selected for the model:

a. **Income:** Median Household Income (in dollars). Since we are trying to understand household income of neighborhoods, *Income* is going to be the dependent variable in our model.

b. **AvgHHSzE:** Average Household Size. The average household size provides a meaningful control to our model. The size of the household may impact the household income i.e. if there are more household members who are working adults in a single household, the income may be high. In an ideal situation we could have computed per capita household income (Income / Size of Household), however, we have a

variable that gives us the number of working adults in a household. Thus, a middle path has been taken and average household size included as an important control variable in our model.

The following variables relates to share of the population of the neighborhood. The share variables are expressed as whole numbers in the scale of 0-100.

c. **UnempShare:** Share of civilian labor force population of 16 years and over who are unemployed. Income goes hand in hand with employment, so it is natural to pick unemployment share as one of the variables.

d. **WomenShare:** Share of the civilian labor force population of 16 years and over who are women. It can be argued that more women in the labor force would mean that the household income might be higher as there might be more people in the household who are earning.

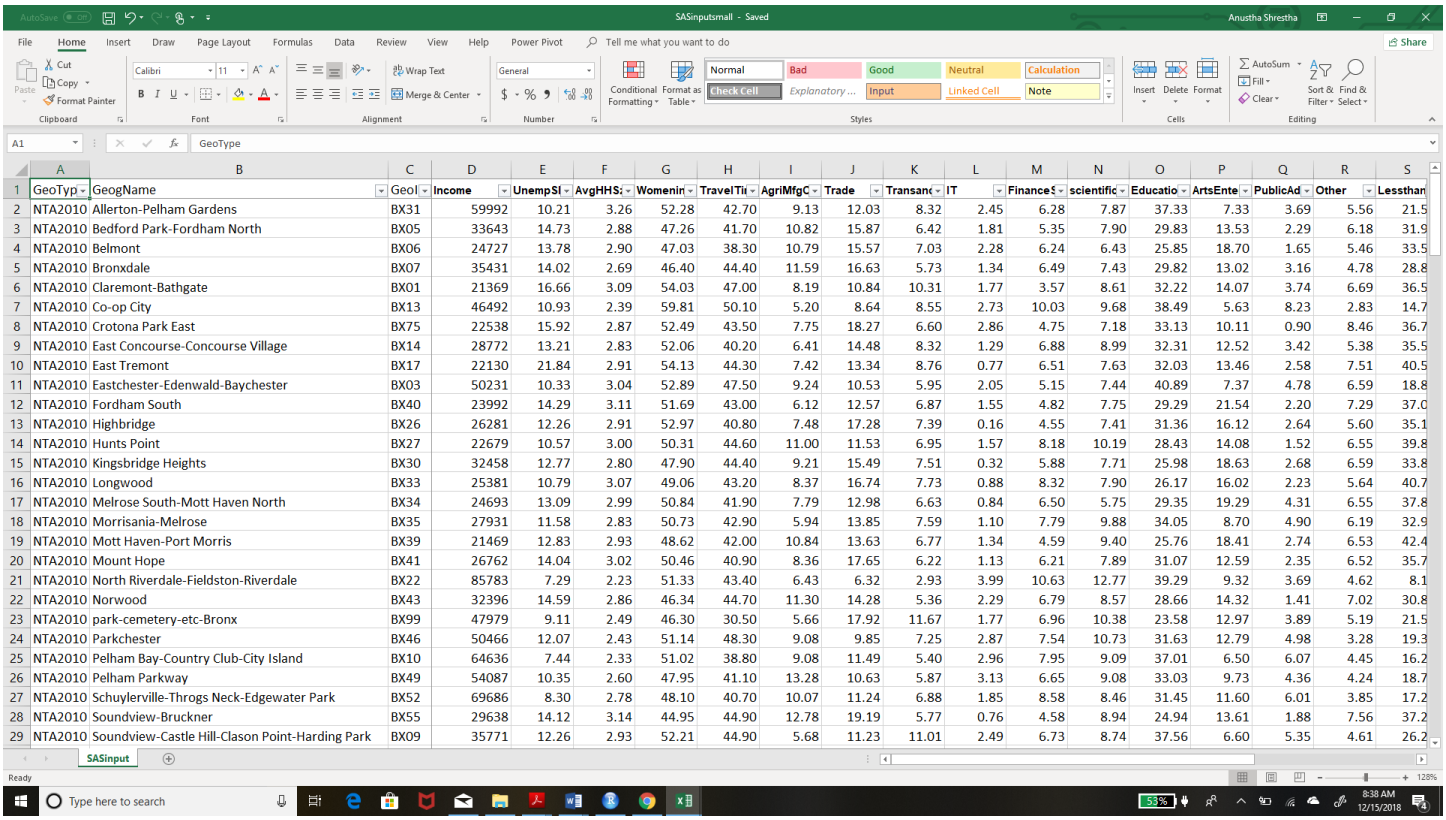
e. **LessthanHS**: Share of population 25 years and over who are less than High School graduate. Income on an individual level is related to education level. It is expected that college graduates will have more income that high school graduates. It is interesting to see if education plays a similar role in neighborhood household income as well. Therefore, the variable share of population who are less than High School graduate was selected for the model.

f. **BachelorandH**: Share of population 25 years and over with Bachelor's degree or higher. Share of population with Bachelor's degree or higher is the next variable that shows that education level of the neighborhood.

g. **Foreign born**: Share of population who were born outside of the U.S and its territories. Since New York is known for its diversity and has a significant population of people who have immigrated from another country, it is noteworthy to see if the household income of neighborhoods differs at different levels of share of population who were born outside the U.S. increases.

h. **FinanceShare**: Share of the civilian labor force population of 16 years and over who are engaged in Finance and insurance, and real estate and rental and leasing industry. New York is the finance center of the capitalist world. Naturally, there are a lot of people who are engaged in the finance or financial services industry. Therefore, share of people who are engaged in finance and related industry has been chosen.

1.4 Data View



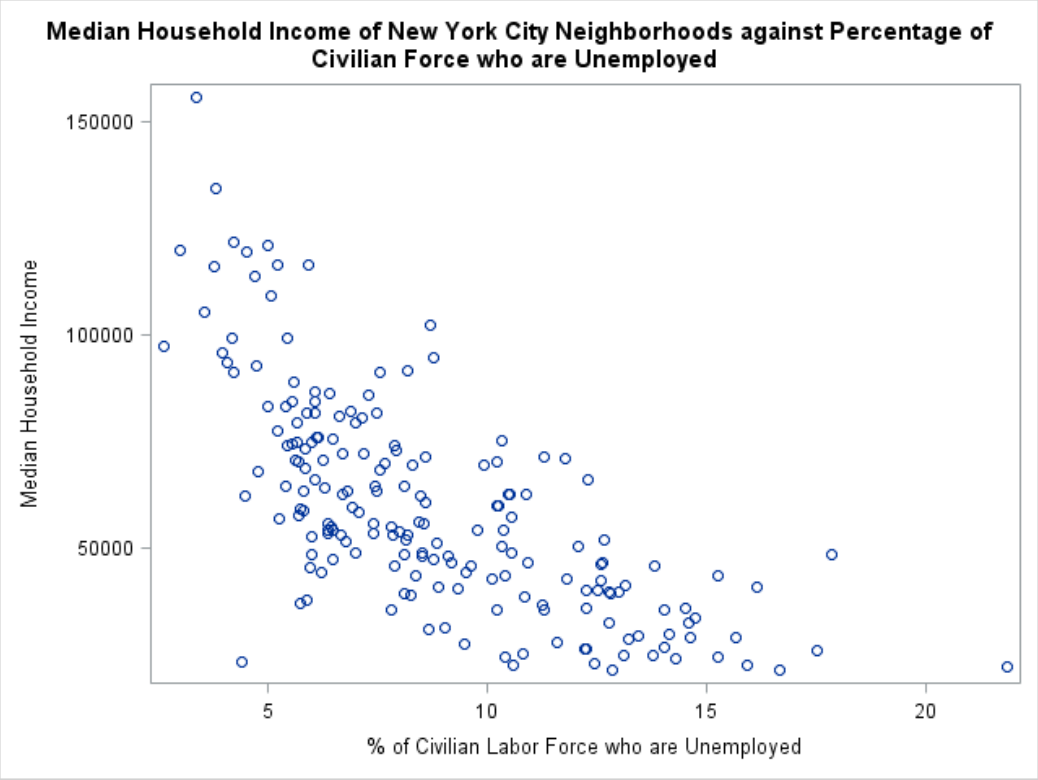
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	GeoTyp	GeogName	GeoI	Income	UnempSI	AvgHHS:	Womenin	TravelTi	AgriMfgC	Trade	Transanc	IT	Finance$	scientific	Educatio	ArtsEnte	PublicAd	Other	Lessthan
2	NTA2010	Allerton-Pelham Gardens	BX31	59992	10.21	3.26	52.28	42.70	9.13	12.03	8.32	2.45	6.28	7.87	37.33	7.33	3.69	5.56	21.5
3	NTA2010	Bedford Park-Fordham North	BX05	33643	14.73	2.88	47.26	41.70	10.82	15.87	6.42	1.81	5.35	7.90	29.83	13.53	2.29	6.18	31.9
4	NTA2010	Belmont	BX06	24727	13.78	2.90	47.03	38.30	10.79	15.57	7.03	2.28	6.24	6.43	25.85	18.70	1.65	5.46	33.5
5	NTA2010	Bronxdale	BX07	35431	14.02	2.69	46.40	44.40	11.59	16.63	5.73	1.34	6.49	7.43	29.82	13.02	3.16	4.78	28.8
6	NTA2010	Claremont-Bathgate	BX01	21369	16.66	3.09	54.03	47.00	8.19	10.84	10.31	1.77	3.57	8.61	32.22	14.07	3.74	6.69	36.5
7	NTA2010	Co-op City	BX13	46492	10.93	2.39	59.81	50.10	5.20	8.64	8.55	2.73	10.03	9.68	38.49	5.63	8.23	2.83	14.7
8	NTA2010	Crotona Park East	BX75	22538	15.92	2.87	52.49	43.50	7.75	18.27	6.60	2.86	4.75	7.18	33.13	10.11	0.90	8.46	36.7
9	NTA2010	East Concourse-Concourse Village	BX14	28772	13.21	2.83	52.06	40.20	6.41	14.48	8.32	1.29	6.88	8.99	32.31	12.52	3.42	5.38	35.5
10	NTA2010	East Tremont	BX17	22130	21.84	2.91	54.13	44.30	7.42	13.34	8.76	0.77	6.51	7.63	32.03	13.46	2.58	7.51	40.5
11	NTA2010	Eastchester-Edenwald-Baychester	BX03	50231	10.33	3.04	52.89	47.50	9.24	10.53	5.95	2.05	5.15	7.44	40.89	7.37	4.78	6.59	18.8
12	NTA2010	Fordham South	BX40	23992	14.29	3.11	51.69	43.00	6.12	12.57	6.87	1.55	4.82	7.75	29.29	21.54	2.20	7.29	37.0
13	NTA2010	Highbridge	BX26	26281	12.26	2.91	52.97	40.80	7.48	17.28	7.39	0.16	4.55	7.41	31.36	16.12	2.64	5.60	35.1
14	NTA2010	Hunts Point	BX27	22679	10.57	3.00	50.31	44.60	11.00	11.53	6.95	1.57	8.18	10.19	28.43	14.08	1.52	6.55	39.8
15	NTA2010	Kingsbridge Heights	BX30	32458	12.77	2.80	47.90	44.40	9.21	15.49	7.51	0.32	5.88	7.71	25.98	18.63	2.68	6.59	33.8
16	NTA2010	Longwood	BX33	25381	10.79	3.07	49.06	43.20	8.37	16.74	7.73	0.88	8.32	7.90	26.17	16.02	2.23	5.64	40.7
17	NTA2010	Melrose South-Mott Haven North	BX34	24693	13.09	2.99	50.84	41.90	7.79	12.98	6.63	0.84	6.50	5.75	29.35	19.29	4.31	6.55	37.8
18	NTA2010	Morrisania-Melrose	BX35	27931	11.58	2.83	50.73	42.90	5.94	13.85	7.59	1.10	7.79	9.88	34.05	8.70	4.90	6.19	32.9
19	NTA2010	Mott Haven-Port Morris	BX39	21469	12.83	2.93	48.62	42.00	10.84	13.63	6.77	1.34	4.59	9.40	25.76	18.41	2.74	6.53	42.4
20	NTA2010	Mount Hope	BX41	26762	14.04	3.02	50.46	40.90	8.36	17.65	6.22	1.13	6.21	7.89	31.07	12.59	2.35	6.52	35.7
21	NTA2010	North Riverdale-Fieldston-Riverdale	BX22	85783	7.29	2.23	51.33	43.40	6.43	6.32	2.93	3.99	10.63	12.77	39.29	9.32	3.69	4.62	8.1
22	NTA2010	Norwood	BX43	32396	14.59	2.86	46.34	44.70	11.30	14.28	5.36	2.29	6.79	8.57	28.66	14.32	1.41	7.02	30.8
23	NTA2010	park-cemetery-etc-Bronx	BX99	47979	9.11	2.49	46.30	30.50	5.66	17.92	11.67	1.77	6.96	10.38	23.58	12.97	3.89	5.19	21.5
24	NTA2010	Parkchester	BX46	50466	12.07	2.43	51.14	48.30	9.08	9.85	7.25	2.87	7.54	10.73	31.63	12.79	4.98	3.28	19.3
25	NTA2010	Pelham Bay-Country Club-City Island	BX10	64636	7.44	2.33	51.02	38.80	9.08	11.49	5.40	2.96	7.95	9.09	37.01	6.50	6.07	4.45	16.2
26	NTA2010	Pelham Parkway	BX49	54087	10.35	2.60	47.95	41.10	13.28	10.63	5.87	3.13	6.65	9.08	33.03	9.73	4.36	4.24	18.7
27	NTA2010	Schuylerville-Throgs Neck-Edgewater Park	BX52	69686	8.30	2.78	48.10	40.70	10.07	11.24	6.88	1.85	8.58	8.46	31.45	11.60	6.01	3.85	17.2
28	NTA2010	Soundview-Bruckner	BX55	29638	14.12	3.14	44.95	44.90	12.78	19.19	5.77	0.76	4.58	8.94	24.94	13.61	1.88	7.56	37.2
29	NTA2010	Soundview-Castle Hill-Clason Point-Harding Park	BX09	35771	12.26	2.93	52.21	44.90	5.68	11.23	11.01	2.49	6.73	8.74	37.56	6.60	5.35	4.61	26.2

Chapter 2. A Simple Regression Model

2.1 Scatterplot

Since we are interested in learning about the Income of various neighborhoods in New York City, our y-variable is going to be the Median Household Income of neighborhoods, *Income*. We can speculate that one of the variables that should strongly be related to *Income* would be the rate of unemployment in the neighborhood. Therefore, we calculate the share of Civilian Labor Force who are Unemployed as a measure of unemployment in the neighborhoods. Our x-variable in this case is the share of Labor Force who are Unemployed, *UnempShare*.

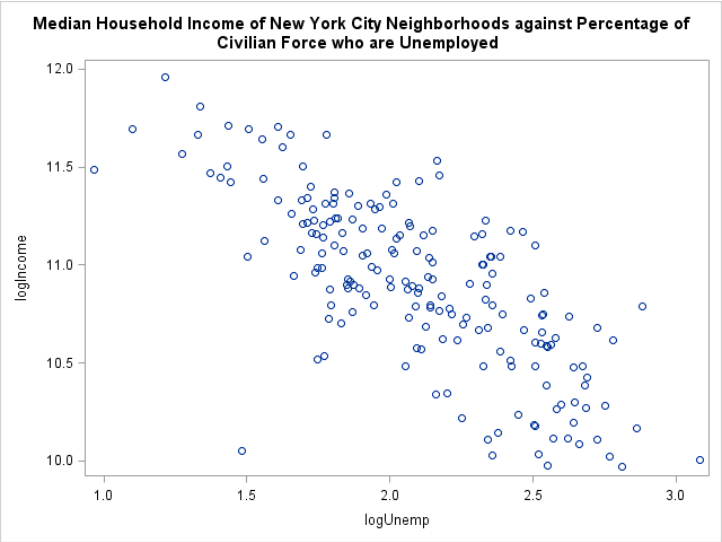
Since these are both numerical variables, the analysis of the scatterplot will help add some insight into the relationship between the two variables.



2.2 Analysis of the Scatterplot

The scatterplot of Median Household Income (*Income*) against percentage of Civilian Labor Force who are Unemployed *(UnempShare*) shows a downward sloping trend as one would expect, i.e. at higher unemployment rates, the household income is lower. However, the scatterplot is slightly curved, which violates the assumption of linearity. In order to get rid of the curvature, we need to transform either x or y variable or both before we fit a linear model.

In addition to the curvature, there seems to be one high leverage point and one outlier in our graph. A point is said to have high leverage if it has a high x- value. In our case, the high leverage point belongs to the data point that has approximately 22% unemployment rate, which is the Bronx neighborhood, East Tremont (BX17). Outliers are points that have big residuals. In our case, the Brooklyn neighborhood, Williamsburg (BK72), is the outlier as it has low unemployment rate of approximately 4% and a low median income of around $25,000.



A few transformations were conducted, including log, square root and 1/x. The best results were seen using natural log transformation on both the x and y variables (logIncome and logUnemp); the transformation eliminated the curvature that was prominent in the previous scatterplot. Since, log of 0 is not defined, any data point with 0 income or unemployment rate, would be eliminated during the transformation process. In our case, we had already recognized the neighborhoods with the missing values and removed them from our data set. However, it should be be noted that these data points were removed not only because they had 0 values, but the deleted neighborhoods included parks, cemetery and airport areas; these are not exactly residential neighborhoods, so it was reasonable for us to delete these observations. The outlier that we saw earlier is still visible, but the transformation has made the scatterplot look more linear.

2.3 The Linear Regression Model

After we transformed our dependent and independent variables in the scatterplot in the previous section. We can now conduct a simple regression using our dependent and independent variables. In our linear regression model, the dependent variable is the natural log of Median Household Income (logIncome) and the independent variable is the natural log of share of Civilian Labor Force who are Unemployed (logUnemp). The model can be written as follows:

$$\text{logIncome}_{(\text{logUnemp})} = \beta_0 + \beta_1 \text{ logUnemp} + \varepsilon$$

The simple linear model assumes that at each level of the x-variable, there is a subpopulation of y-variables given x. We assume that these subpopulations of y given x is normally distributed with the parameters μ and σ, which are the subpopulation mean and subpopulation standard deviation.

a) In our model, y given x or Y_x term represents the subpopulation distribution of logIncome at the given level of logUnemp. (Remember LogIncome is the natural log of Median Household Income and logUnemp is the natural log of Unemployment Rate).

b) β_0 and β_1 are the population intercept and population slope. In our model, ($\beta_0 + \beta_1$ logUnemp) represents the subpopulation average of the distribution of logIncome values given the level of logUnemp. For instance, at logUnemp = 1, the subpopulation average of logIncome is $\beta_0 + \beta_1$. At logUnemp = 2, the mean of subpopulation of log of Income is $\beta_0 + 2\beta_1$.

c) The term ε, is the bell curve template of our model and provides the variance of our model, or the width of the subpopulation distribution of Y given x, Y_x. The variance of ε, is the estimates the variance of Y_x, or log of Income given log of Unemployment.

2.4.SAS output for the Fitted

In simple regression, the next step is to obtain the estimate of these population parameters. Using least squares methodology, we can calculate the sample slope (b_1) and sample intercept (b_0), which are the unbiased estimators of population slope (β_1) and population intercept (β_0) [the meaning of unbiased estimator is discussed later in the section]. By estimating sample slope and sample intercept, we determine the $\hat{y}$ – equation which is written as follows:

$$\hat{y} = b_0 + b_1 x$$

Since we know that b_0 is the estimate of the population intercept, and b_1 is an estimate of the population slope, $\hat{y}$ consequentially, is the estimate of the population average of the y given x.

For our project, we use SAS to fit our model for our y-variable, logIncome using our x-variable, logUnemp. The resulting output is below:

The REG Procedure
Model: MODEL1
Dependent Variable: logIncome

Number of Observations Read	**190**
Number of Observations Used	190

Analysis of Variance

Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	19.93556	19.93556	235.77	<.0001
Error	188	15.89670	0.08456		
Corrected Total	189	35.83226			

Root MSE	**0.29079**	R-Square	**0.5564**
Dependent Mean	10.90520	**Adj R-Sq**	0.5540
Coeff Var	2.66650		

Parameter Estimates

Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	12.62716	0.11411	110.65	<.0001
logUnemp	**1**	-0.82319	0.05361	-15.35	<.0001

The SAS output computes our sample slope and sample intercept and our y-hat line could be written as follows:

$$\widehat{logIncome} = 12.63 - 0.82\ logUnemp$$

The intercept 12.63 is our baseline. The coefficient of logUnemp, the log of unemployment share is -0.82. This tells us as the unemployment share increases by 1%, then household income decreases by 0.82%. However, this interpretation is not valid until we determine if the slope parameter is statistically significant.

2.5 Analysis of Output

a) Once we run our simple regression analysis, it is important to see if the independent variable provides us any information about the dependent variable. Since population slope, β_1 shows how the independent variable of the model relate to the independent variable, it is important conduct a hypothesis test on β_1. If β_1 were equal to 0 then we would know that the variable unemployment share does not tell us anything about the value of Median Household Income.

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

In order to test our hypothesis, we use sample slope (b_1) since it is an unbiased estimator of β_1. We calculate test statistics using sample slope (b_1) and the standard error of the slope (s_{b1}), which we can obtain from the output table.

Test statistic: $\text{t-stat} = \frac{b_1 - 0}{s/\sqrt{SSx}} = \frac{b_1}{s_{b1}} = \frac{-0.823}{0.054} = -15.35$

Rejection region: $|t - stat| > t - critical\ value, \alpha = 0.05$
t-critical value with d.f = 188 is 1.97

We reject the null hypothesis that population slope β_1, is equal to 0 because the |t-stat| = 15.35 is greater than the t-critical value of 1.97.

Since the population parameter β_1, is not known to us, the one way we can test to see if β_1, is equal to 0 is by using the unbiased estimator of population slope, b_1. The sample slope b_1 is just one of many slopes of samples of size n= 190 that could be drawn from the parent population of neighborhoods. Since there are many ways to draw a sample of size n=190 from the original population, the daughter population will have a very high population size since each sample has their own sample slope and intercept. The average of these sample slopes of daughter population is equal to β_1. Thus, we can say that the sample slope, b_1 is an unbiased estimator of population slope β_1.

As b_1 is the unbiased estimator of β_1, we first test to see if b_1 is close to 0. If the null hypothesis were true, then b_1 would in fact be close to 0. Then we have to ask the question: how close is close? In order to determine the closeness, we need to calculate the standard deviation of the daughter population of sample slopes, s_{b1}. Since we need to measure distance in standard units, we need calculate z statistic by dividing ordinary distance from expected value by the standard deviation. We calculate t-stat at $\beta_1 = 0$, i.e. t-stat = $(b_1 - E(b_1))/ s_{b1}$, which is equal to b_1/ s_{b1}. By comparing the t-stat with the t-critical value, we will be able to determine if we can reject or if we fail to reject the null hypothesis.

b) General $\hat{y}$ equation can be written as:

$$\hat{y} = b_0 + b_1 x$$

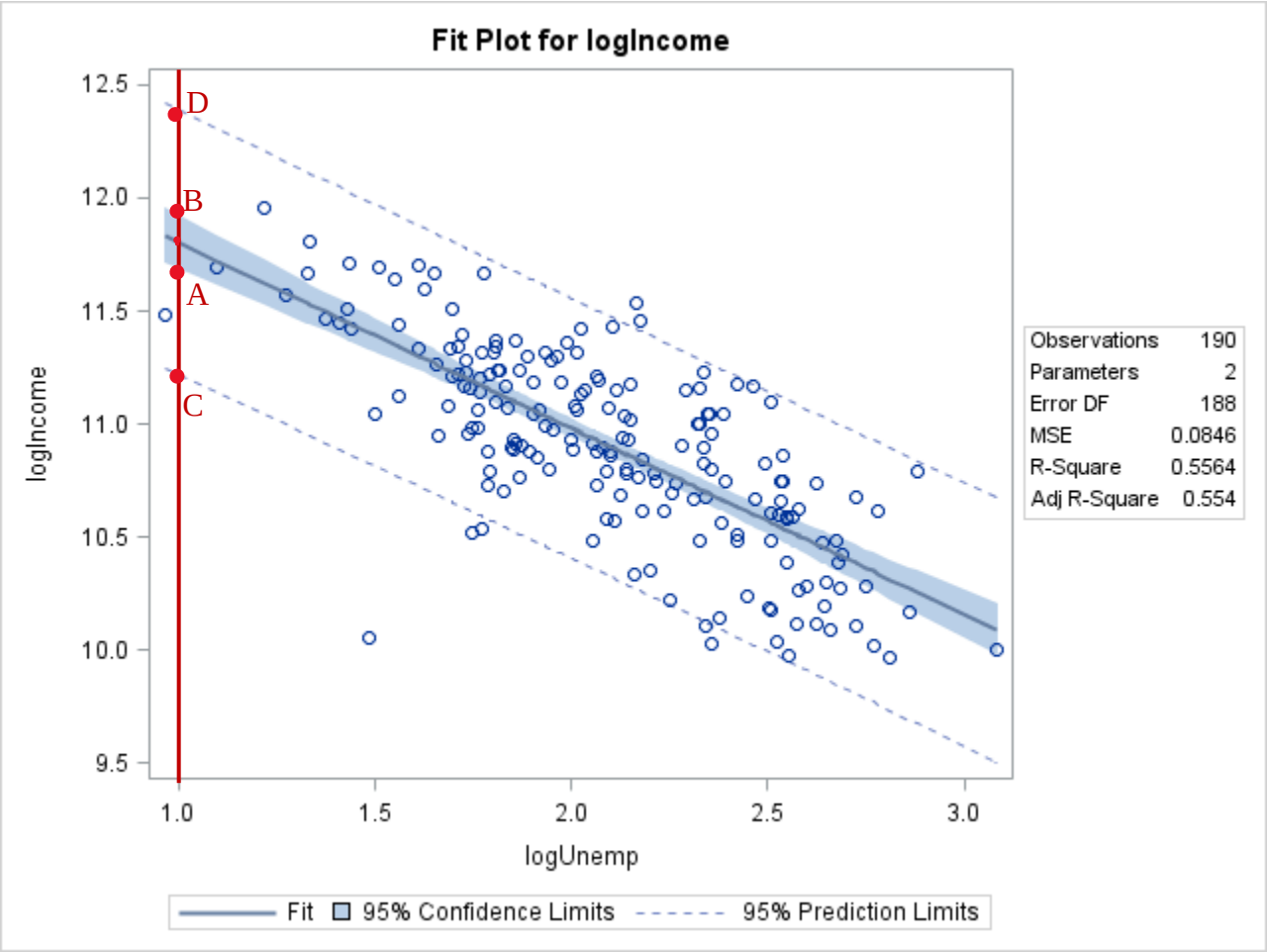
In the context of our model it is: $\widehat{logIncome} = 12.63 - 0.82\ logUnemp$

$\hat{y}$ values are calculated using sample intercepts and sample slopes at given level of x. In our case, $\hat{y}$ represents the natural log of Median Household Income and it is calculated given x value, i.e. the log of Unemployment rate. $\hat{y}$ is a point estimate of $E(Y_x)$ or in simpler terms, it is the point estimate of the mean of the bell curve distribution of subpopulation of Y given x.

The above stated sample slope (b_1 = -0.82) and sample intercept (b_0 = 12.63) comes from just one of many samples of size n=190 that can be drawn from the parent population. We know that there are many samples of size n= 190 that could be drawn from the population of neighborhoods. For each different sample, we can compute their own sample slopes and intercepts, and calculate different $\hat{y}$ given x. Therefore, the $\hat{y}$ values at a given x has its own daughter population. The grand average of all these $\hat{y}$ values given x, is equal to the average of the mean of the bell curve distribution of Y given x ($\mu_{Yx)}$, which can also be written as $E(Y_x)$ in a random variable notation. $\hat{y}$, therefore, is said to be an unbiased estimator of $E(Y_x)$.

In our example, $\widehat{logIncome}_{logUnemp}$ is an unbiased estimator of the $logIncome_{logUnemp}$.

c)



The 95% confidence band shows the region that includes 95% confidence intervals for $E(Y_x)$ at different levels of x. In this case, at each level of logUnemp, we are 95% confident that the population average of logIncome given the logUnemp falls within the confidence interval (the blue region). In the Fit Plot for logIncome, we can see the vertical cut on 1.0 logUnemp. The vertical line cuts the blue region at 2 points, namely A and B, which are approximately 11.75 and 11.9 respectively. These are the lower and the upper limits of the 95% Confidence Interval for logIncome at logUnemp= 1. We are 95% confident that the average of log of Median Household Income given log of Unemployement Rate =1 lies in the interval (11.75, 11.9). In order to find the confidence interval of the original variable, the Median Household Income, we need to transform these log values back to the original scale by taking the exponential of upper and lower limit of the confidence intervals.

The 95% prediction band shows the region that includes 95% prediction intervals for the next observation, given the x value. In this case, at each level of logUnemp, we are 95% confident that logIncome is going to lie in the prediction interval (the interval between the dotted blue lines). The vertical line cuts the dotted blue lines at 2 points, namely C and D, which are approximately 11.3 and 12.45 respectively. These are the lower and the upper limits of the 95% Prediction Interval for logIncome at logUnemp = 1. We are 95% confident that the next value of log of Median Household Income given log of Unemployement Rate =1 lies in the interval (11.3, 12.45). In order to find the prediction interval of the original variable, the Median Household Income, we need to transform these log values back to the original scale by taking the exponential of upper and lower limit of the prediction intervals.

Chapter 3. Matrix Approach to Regression

3.1 Simple Linear Regression in Matrix Terms

For the regression of y on x using a matrix approach, I have used a subset of my data, n=8 observations. Here are the 8 rows that were chosen:

GeogName	GeoID	LogIncome	LogUnemp
University Heights-Morris Heights	BX36	10.03	2.52
Bath Beach	BK27	10.97	1.95
Erasmus	BK95	10.67	2.31
Central Harlem North-Polo Grounds	MN03	10.48	2.67
SoHo-TriBeCa-Civic Center-Little Italy	MN24	11.71	1.44
Hunters Point-Sunnyside-West Maspeth	QN31	11.12	1.56
Queens Village	QN34	11.17	2.46
Great Kills	SI54	11.42	2.02

a) **X** matrix with logUnemp as the x-variables is as follows:

$$X = \begin{bmatrix} 1 & 2.52 \\ 1 & 1.95 \\ 1 & 2.31 \\ 1 & 2.67 \\ 1 & 1.44 \\ 1 & 1.56 \\ 1 & 2.46 \\ 1 & 2.02 \end{bmatrix}$$

b) The **y**-vector is composed of observed y-variable values, LogIncome and the matrix is as follows:

$$\overrightarrow{Y_x} = \begin{bmatrix} 10.3 \\ 10.97 \\ 10.76 \\ 10.48 \\ 11.71 \\ 11.12 \\ 11.17 \\ 11.42 \end{bmatrix}$$

c) The following Excel output shows the b-vector, Hat matrix, and the **y-hat** vector:

H= X ((X'X)^-1)X							
0.24	0.08	0.18	0.28	-0.07	-0.03	0.22	0.10
0.08	0.14	0.10	0.06	0.20	0.19	0.09	0.14
0.18	0.10	0.15	0.20	0.03	0.05	0.17	0.11
0.28	0.06	0.20	0.34	-0.14	-0.09	0.26	0.09
-0.07	0.20	0.03	-0.14	0.45	0.39	-0.04	0.17
-0.03	0.19	0.05	-0.09	0.39	0.34	-0.01	0.16
0.22	0.09	0.17	0.26	-0.04	-0.01	0.21	0.10
0.10	0.14	0.11	0.09	0.17	0.16	0.10	0.13

y hat
10.59
11.09
10.78
10.46
11.55
11.44
10.64
11.03

b-vector
12.82
-0.88

The excel calculation is shown below:

X		y
1	2.52	10.03
1	1.95	10.97
1	2.31	10.67
1	2.67	10.48
1	1.44	11.71
1	1.56	11.12
1	2.46	11.17
1	2.02	11.42

X'X	
8.00	16.95
16.95	37.34

(X'X)^-1	
3.24	-1.47
-1.47	0.70

X'							
1	1	1	1	1	1	1	1
2.520665	1.953966	2.31143533	2.674669	1.436602	1.562104	2.464562697	2.024006

H= X ((X'X)^-1)X							
0.24	0.08	0.18	0.28	-0.07	-0.03	0.22	0.10
0.08	0.14	0.10	0.06	0.20	0.19	0.09	0.14
0.18	0.10	0.15	0.20	0.03	0.05	0.17	0.11
0.28	0.06	0.20	0.34	-0.14	-0.09	0.26	0.09
-0.07	0.20	0.03	-0.14	0.45	0.39	-0.04	0.17
-0.03	0.19	0.05	-0.09	0.39	0.34	-0.01	0.16
0.22	0.09	0.17	0.26	-0.04	-0.01	0.21	0.10
0.10	0.14	0.11	0.09	0.17	0.16	0.10	0.13

y hat = Hy
10.59
11.09
10.78
10.46
11.55
11.44
10.64
11.03

b-vector
12.82
-0.88

y-hat = Xb
10.59
11.09
10.78
10.46
11.55
11.44
10.64
11.03

$$\vec{b} = (X'X)^{-1} X' \vec{y}$$

d) A regression procedure was conducted in SAS to fit the model and check the values. From the regression output below, b_0 and b_1 are 12.82 and -0.88 which ties to the **b-vector** we calculated in Excel.

The REG Procedure
Model: MODEL1
Dependent Variable: LogIncome

Number of Observations Read	**8**
Number of Observations Used	8

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.12032	1.12032	7.52	0.0336
Error	6	0.89334	0.14889		
Corrected Total	7	2.01366			

Root MSE	**0.38586**	R-Square	**0.5564**
Dependent Mean	10.94729	**Adj R-Sq**	0.4824
Coeff Var	3.52473		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	**1**	12.81680	0.69506	18.44	<.0001
LogUnemp	**1**	-0.88247	0.32171	-2.74	0.0336

e)
For simple regression we can compute the entries of the hat matrix directly from the data. For instance, we can compute the value of $h_{2,3}$

We know,
$h_{ij} = \left(\frac{1}{n}\right) + (x_i - \bar{x})k_j$, where $k_j = \frac{x_2 - \bar{x}}{SSx}$
$\bar{x} = 2.12$
$SSx = \sum(x_i - \bar{x})^2 = 1.44$
$x_2 = 1.95$
$x_3 = 2.31$

$h_{2,3} = \left(\frac{1}{8}\right) + (1.95 - 2.12)(\frac{2.31 - 2.12}{1.44})$
$= \left(\frac{1}{8}\right) + (-0.17)(0.13)$
$= \left(\frac{1}{8}\right) + (-0.0221)$
$= 0.1029$

x bar	xi - xbar	(xi - xbar)^2
2.1185	0.4022	0.1617
	-0.1645	0.0271
	0.1929	0.0372
	0.5562	0.3093
	-0.6819	0.4650
	-0.5564	0.3096
	0.3461	0.1198
	-0.0945	0.0089
Total	0.0000	1.4386

This is equal to the point in the second row and third column (highlighted in yellow) in the Hat matrix, H.

3.2 Multiple Linear Regression in Matrix Terms

A second x-variable, Average Household Size (AvgHHSize) was added to the selected dataset for this exercise.

GeogName	GeoID	LogIncome	LogUnemp	AvgHHSize
University Heights-Morris Heights	BX36	10.03	2.52	2.85
Bath Beach	BK27	10.97	1.95	2.88
Erasmus	BK95	10.67	2.31	2.85
Central Harlem North-Polo Grounds	MN03	10.48	2.67	2.42
SoHo-TriBeCa-Civic Center-Little Italy	MN24	11.71	1.44	2.09
Hunters Point-Sunnyside-West Maspeth	QN31	11.12	1.56	2.35
Queens Village	QN34	11.17	2.46	3.51
Great Kills	SI54	11.42	2.02	2.81

a) The new **X** matrix is as follows:

$$X = \begin{bmatrix} 1 & 2.52 & 2.85 \\ 1 & 1.95 & 2.88 \\ 1 & 2.31 & 2.85 \\ 1 & 2.67 & 2.42 \\ 1 & 1.44 & 2.09 \\ 1 & 1.56 & 2.35 \\ 1 & 2.46 & 3.51 \\ 1 & 2.02 & 2.81 \end{bmatrix}$$

b) Proc IML procedure in SAS was used to fit the model and calculate the b-vector. The output of the program, b-vector is as follows:

b
12.196266
-1.098745
0.396587

c) A regression procedure was conducted in SAS to fit the model and check the values in the PROC IML b-vector. The output matches and b_0, b_1 and b_2 which are 12.196, -1.099 and 0.397 respectively.

The REG Procedure
Model: MODEL1
Dependent Variable: LogIncome

Number of Observations Read	**8**
Number of Observations Used	8

Analysis of Variance

Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	1.25991	0.62996	4.18	0.0857
Error	5	0.75374	0.15075		
Corrected Total	7	2.01366			

Root MSE	**0.38826**	R-Square	**0.6257**
Dependent Mean	10.94729	**Adj R-Sq**	0.4760
Coeff Var	3.54666		

Parameter Estimates

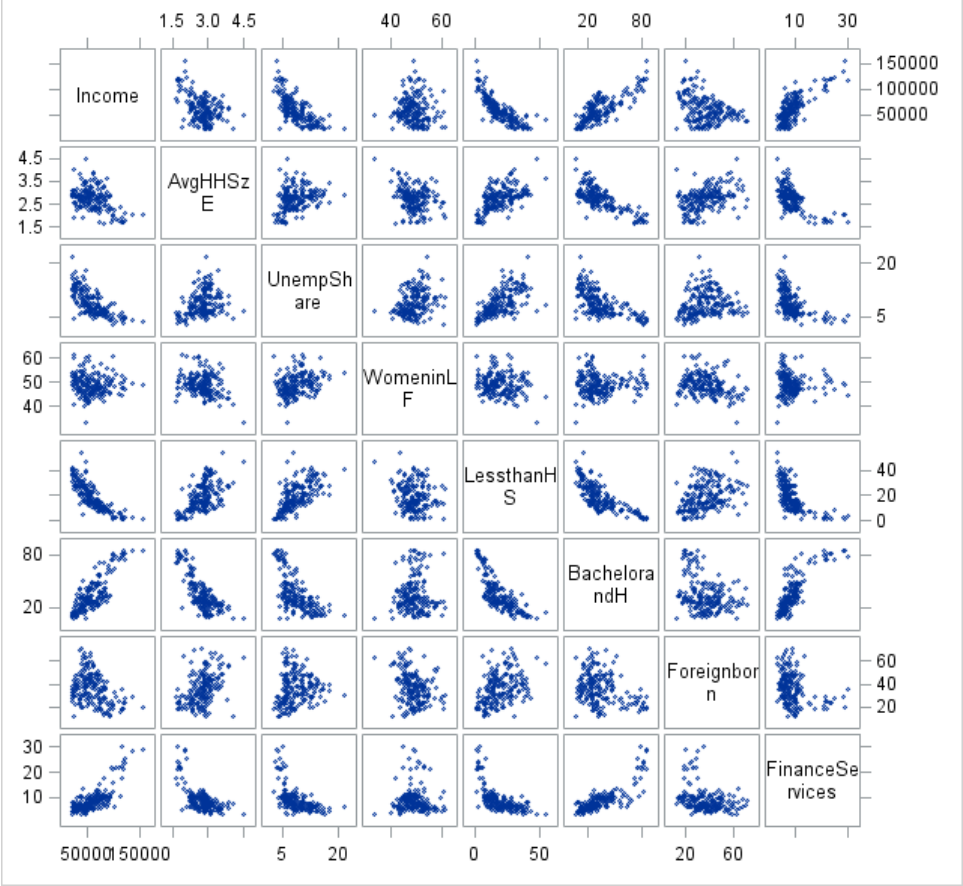
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	12.19627	0.95129	12.82	<.0001
LogUnemp	**1**	-1.09875	0.39408	-2.79	0.0385
AvgHHSize	**1**	0.39659	0.41212	0.96	0.3801

Chapter 4. Model Selection

4.1 Best Subsets Model Selection

a)
Before we proceed fitting the multiple regression model, it is beneficial to explore the relationship between our variables. Below is the scatterplot matrix that shows the relationship between the variables in our dataset. Most importantly, we need to understand the relationship between our dependent variable, median household income (Income), with other independent variables such as average household size (AvgHHSzE), unemployment share (UnempShare), share of women in labor force (WomeninLF), share of population who are less than high school graduate (LessthanHS), share of population who have bachelor's degree and higher (BachelorandH), share of population who are foreign born (Foreignborn) and share of population who are engaged in financial services industry (FinanceServices).

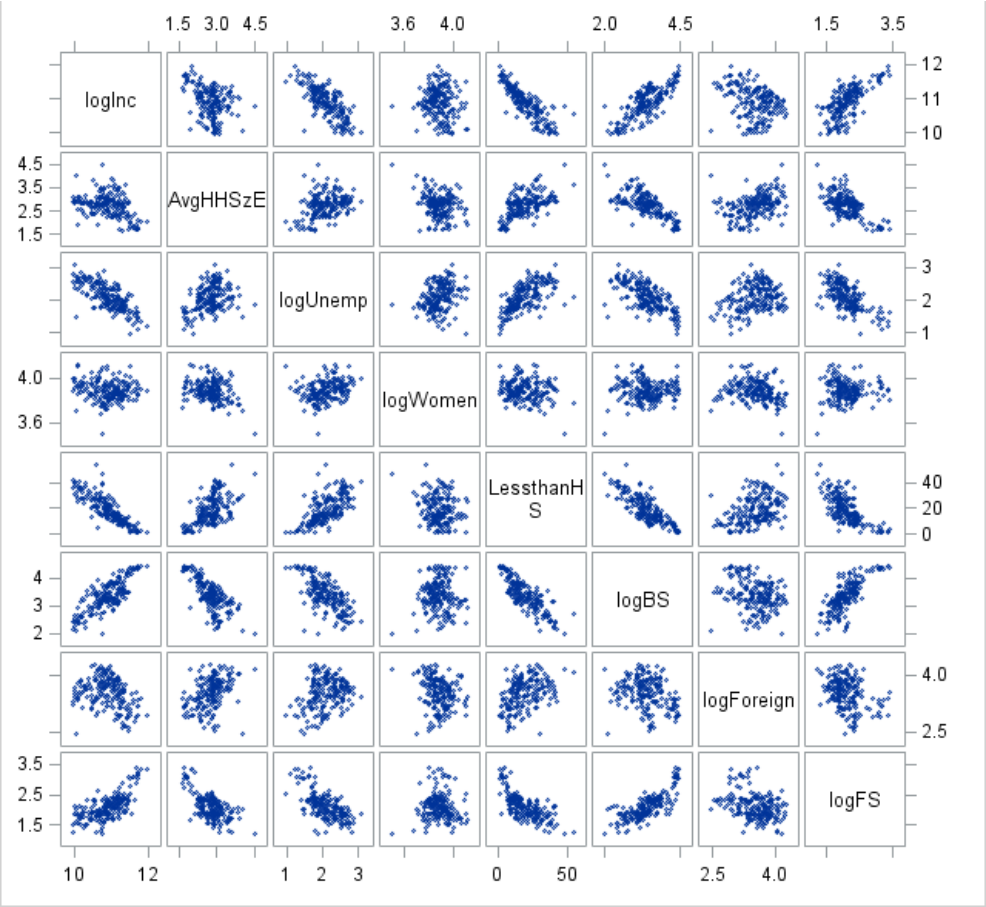
Scatterplot Matrix Showing Relationship of Income against Other Independent Variables



b)
The scatterplot of Income against AvgHHSz, UnempShare, LessthanHS and ForeignBorn show downward sloping trends, but have curvature. In order to reduce the curvature, these variables were transformed using natural log. Furthermore, the variable, Foreignborn, also seems to be heteroscedastic, i.e. the data is less spread out in some areas, whereas it is more spread out in other areas. Similarly, scatterplot of Income against variables BachelorandH and FinancialServices are slightly curved as well as demonstrate heteroscedasticity. These variables were also transformed using natural log.

Various transformations were tried to eliminate curvature and heteroscedasticity. The best transformation resulted in transformation of both the dependent as well as independent variables using natural log; the variables, average household size (AvgHHSzE) and percentage of population who are less than high school graduate (LessthanHS) were not transformed. The final scatterplot matrix with all the desired transformation is shown below.

Scatterplot Matrix Showing Relationship of Income against Other Independent Variables after Transformation



Transformations:
logInc = ln (Income)
logUnemp = ln (UnempShare)
logWomen = ln (WomenShare)
logBS = ln (BachelorandH)
logForeign = log (Foreignborn)
logFS = log (FinanceServices)

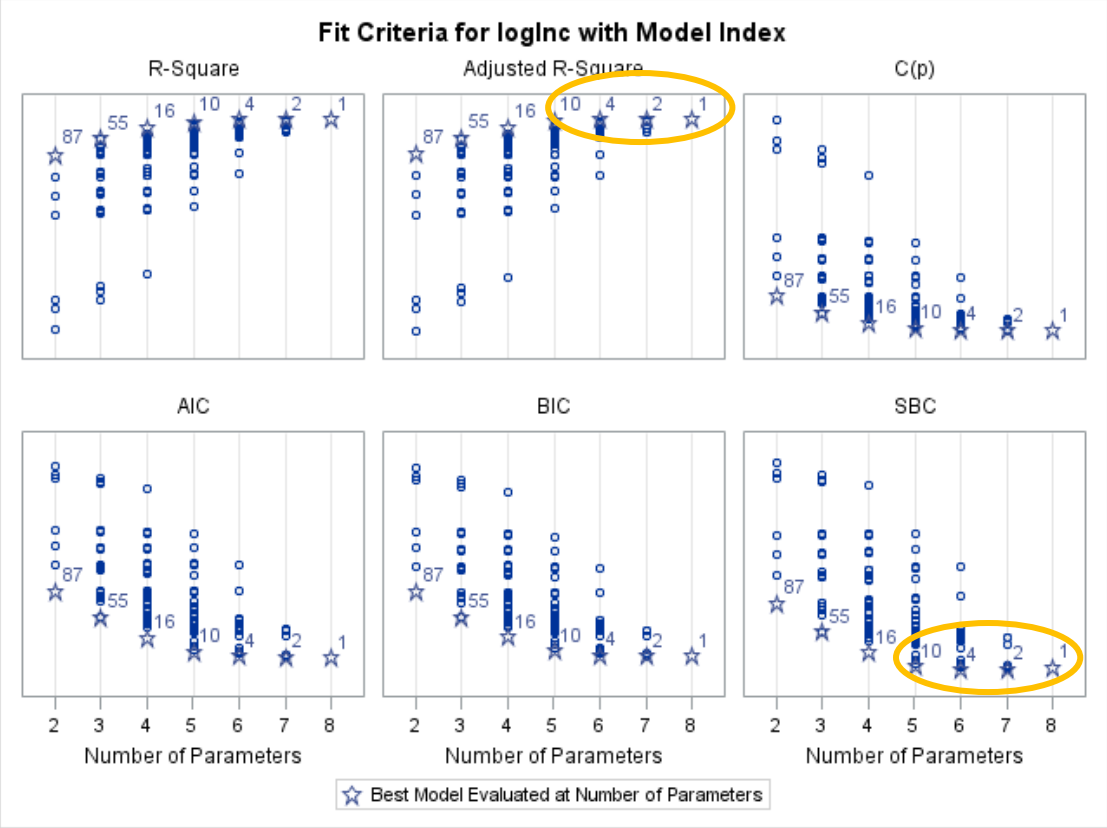
AvgHHSze, the average household size and LessthanHS, the share of population under 25 years who are not high school graduates, are the two variables that were not transformed.

c)

Best Subset Model Selection Routine from SAS was conducted to identify the best model for our dataset. The summary table is given below followed by the criteria plot. The summary table lists all the subsets of all the variables that could be used in our model and provides the Adjusted R^2, R^2, C(p), AIC and SBC measures for each model. Some of the best models are highlighted in yellow below. For instance, model index #10 shows the Adjusted R^2 of 0.8459. The variables included in the model are average household size (AvgHHSSzE), log of unemployment share (logUnemp), share of population who are less than high school graduate (LessthanHS) and log of share of population with bachelor's degree and higher (logBS). Similarly, other models with various combinations of independent variables are listed in the descending order of the Adjusted R^2.

Model Index	Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
1	7	**0.8569**	0.8622	8.0000	-677.4886	-651.51243	AvgHHSzE logUnemp logWomen LessthanHS logBS logForeign logFS
2	6	**0.8556**	0.8602	8.6157	-676.7774	-654.04826	AvgHHSzE logUnemp LessthanHS logBS logForeign logFS
3	6	**0.8544**	0.8591	10.1078	-675.2480	-652.51882	AvgHHSzE logUnemp logWomen LessthanHS logBS logFS
4	5	**0.8535**	0.8574	10.3026	-675.0203	-655.53819	AvgHHSzE logUnemp LessthanHS logBS logFS
5	6	**0.8513**	0.8560	14.1311	-671.1842	-648.45506	AvgHHSzE logUnemp logWomen LessthanHS logBS logForeign
9	5	**0.8462**	0.8503	19.6800	-665.7865	-646.30436	AvgHHSzE logWomen LessthanHS logBS logFS
10	4	**0.8459**	0.8491	19.2024	-666.3288	-650.09372	AvgHHSzE logUnemp LessthanHS logBS
11	5	**0.8436**	0.8478	23.0367	-662.5873	-643.10514	AvgHHSzE LessthanHS logBS logForeign logFS
16	3	**0.8222**	0.8250	49.0537	-640.1463	-627.15824	AvgHHSzE LessthanHS logBS
17	5	**0.8203**	0.8250	53.0676	-636.1349	-616.65275	AvgHHSzE logUnemp logWomen LessthanHS logFS
54	4	**0.7865**	0.7910	95.9332	-604.4197	-588.18460	AvgHHSzE logWomen LessthanHS logForeign
55	2	**0.7825**	0.7848	100.2266	-602.7932	-593.05215	logWomen LessthanHS
56	3	**0.7814**	0.7848	102.1263	-600.8603	-587.87224	logWomen LessthanHS logForeign
86	3	**0.7160**	0.7205	187.0467	-551.1745	-538.18641	logUnemp logWomen logBS
87	1	**0.7149**	0.7164	188.4329	-552.4215	-545.92743	LessthanHS
88	2	**0.7145**	0.7176	188.9672	-551.1667	-541.42561	LessthanHS logForeign
124	1	**0.1293**	0.1339	957.6448	-340.2745	-333.78047	AvgHHSzE
125	2	**0.1273**	0.1366	956.1596	-338.8544	-329.11335	logWomen logForeign
126	1	**0.0951**	0.0999	1002.561	-332.9551	-326.46110	logForeign
127	1	**0.0093**	0.0145	1115.334	-315.7323	-309.23825	logWomen

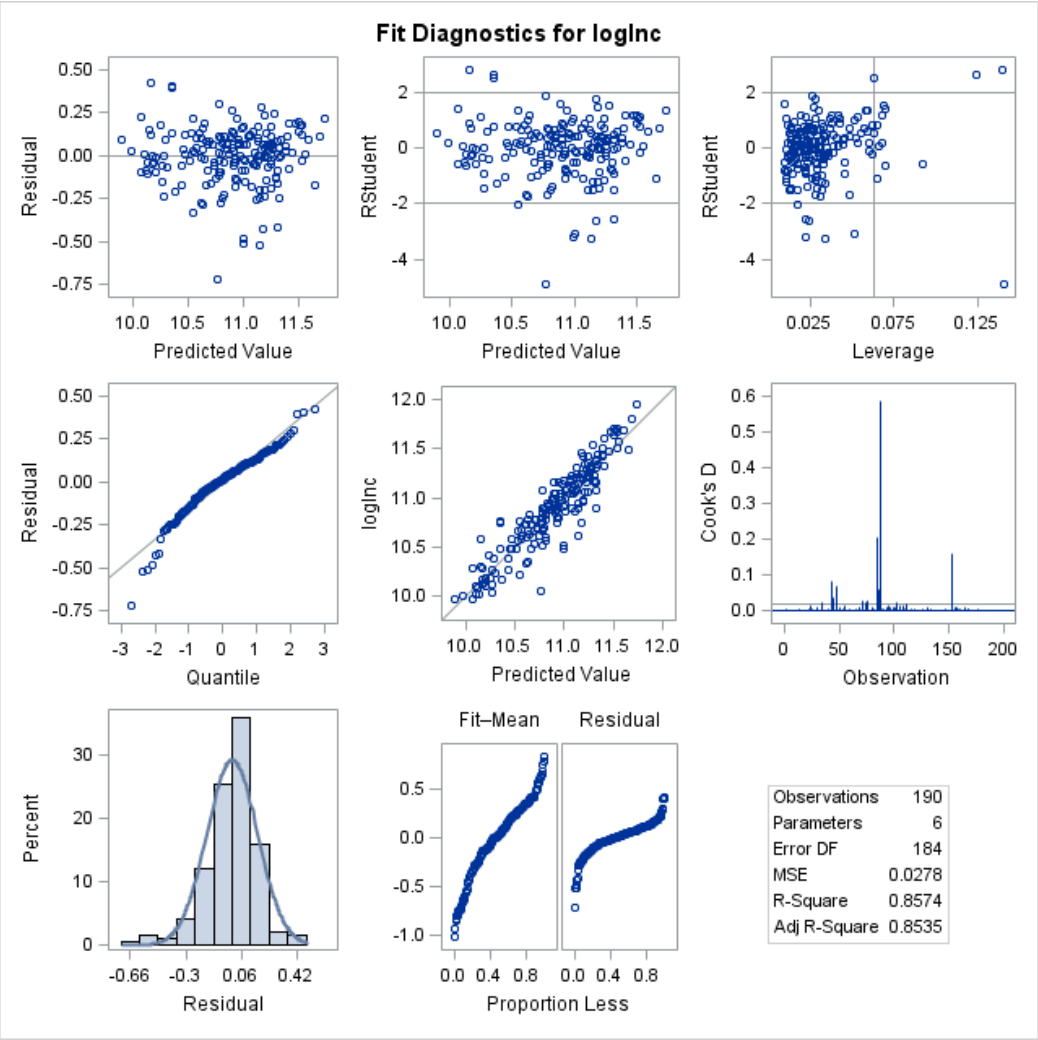
*Please note that there were altogether 127 subsets, however, keeping the limited space in mind, a few subset rows that were not as important for the discussion were deleted.



d)
From the Criteria plot, there were four contending models – the models indexed as #1, #2, #4, and #10. In the criteria plot above, I have circled these four models within the SBC box. Similarly, if we look at plots for AIC, BIC and C(p) criteria, these four models keep emerging as the victors with the lowest values. Again, these four models, also have the highest R^2 and Adjusted R^2. In order to look up for the variables that were included in these top four contending models, we need to look up the summary table. I have highlighted these four models in yellow.

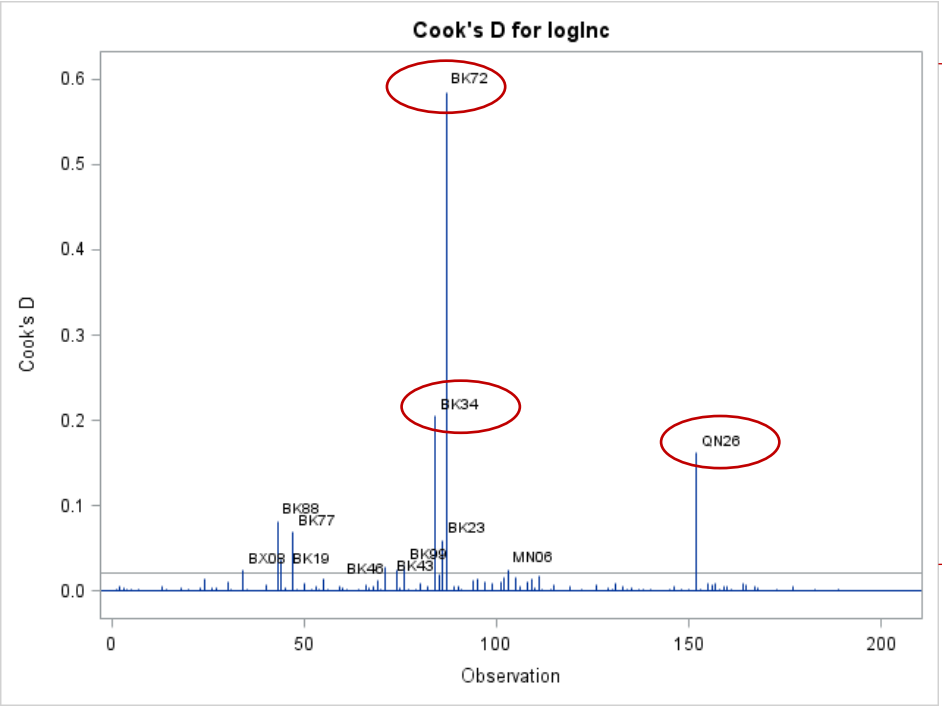
There is not much difference in the Adjusted R^2 values for model #1, #2 and #4, keeping simplicity in mind, I picked the model with the least number of variables, model #4. The variables included in the model were: average household size (AvgHHSzE), log of unemployment rate (logUnemp), percentage of population who are less than high school graduate (LessthanHS), log of percentage of population who had bachelor's degree or higher (logBS) and percentage of population who were engaged in financial services industry (logFS).

e) A regression procedure was conducted to fit our model, and the diagnostic plots were run.

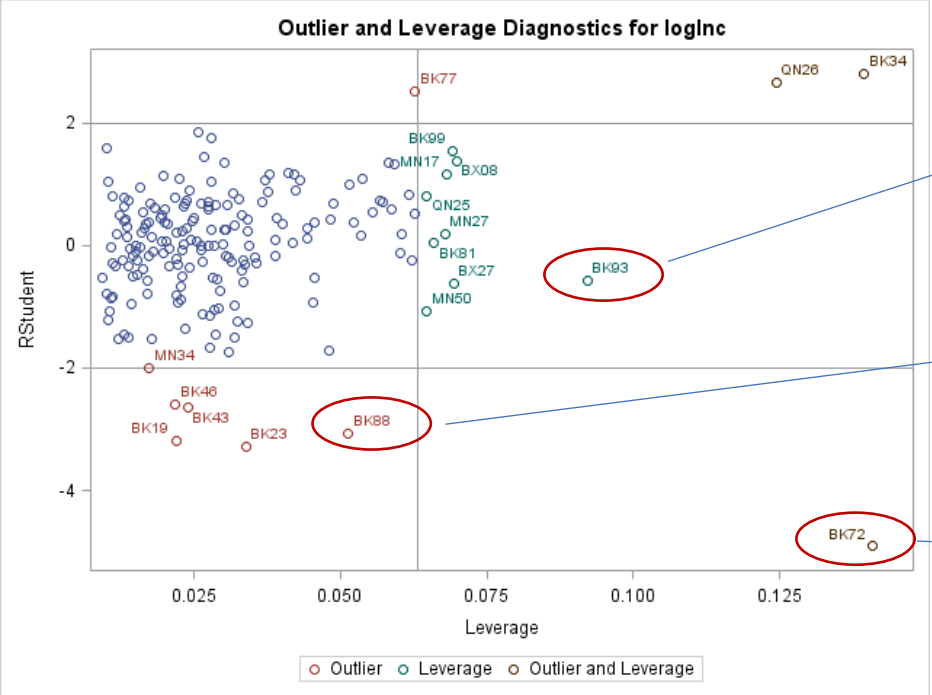


In the diagnostic plot above, we can see that the residuals are mostly evenly spread around 0. Most of the data lies within the RStudent cut-off of 2. We can see a few leverage points and a few observations flagged as influence point in the Cook's D table. The quantile table shows that the data is mostly normally distributed, except for the few observations in towards the left.

To further identify the influential points and the outliers, labeled diagnostics plots were created, which are on the following page.



SAS uses the 4/n rule as a cut-off, so any point above the gray line has been flagged as an influential point. Some of the examples are highlighted.



Leverage point

Outlier

Leverage point as well as outlier

Altogether 12 neighborhoods are flagged as influential points by the 4/n Cook's D cutoff, with the Brooklyn neighborhood BK72, Williamsburg, being the most influential observation with Cook's D 0.6. The RStudent plot also flags many neighborhoods as high leverage points and outliers. For instance, neighborhood BK72, Williamsburg, is both a high leverage point as well as an outlier (hence an influential point). Similarly, neighborhoods BK93, Starrett City, is a high leverage point, whereas, neighborhoods BK88, Borough Park, and BK23, West Brighton, are outliers.

4.2 Forward Stepwise Model Selection

Summary of Stepwise Selection									
Step	**Variable Entered**	**Variable Removed**	**Label**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	LessthanHS		LessthanHS	1	0.7164	0.7164	188.433	475.02	<.0001
2	logWomen			2	0.0683	0.7848	100.227	59.35	<.0001
3	logUnemp			3	0.0206	0.8054	75.0062	19.70	<.0001
4	AvgHHSzE		AvgHHSzE	4	0.0086	0.8140	65.6138	8.58	0.0038
5	logBS			5	0.0382	0.8522	17.1637	47.56	<.0001
6	logFS			6	0.0069	0.8591	10.1078	8.90	0.0032
7	logForeign			7	0.0031	0.8622	8.0000	4.11	0.0441

a)
A forward stepwise procedure was conducted, which is another method of model selection. Using the default stepwise selection method, SAS first brings LessthanHS, share of population who are less than high school graduate, as the first variable for inclusion. Then it added logWomen, log of share of women in the labor force, as the next variable and did not remove LessthanHS. Then logUnemp, log of unemployment share, AvgHHSzE, average household size, logBS, log of share of population with bachelor's degree or higher, logFS, log of share of population who are engaged in the finance industry and logForeign, log of share of population who were foreign born, were included respectively and none of the prior variables were deleted. All the variables that we had selected on our data set was included by SAS forward stepwise procedure.

b)
From the The Best Subset Model, the variables we had selected were logUnemp, AvgHHSze, LessthanHS, logBS and logFS. The forward stepwise selection, however, includes logWomen as well as logForeign as regressors. This is not a surprising result because when we used the best subset model selection routine, we ended up with four contending models and one of them included all the above listed variables.

4.3 Variance Inflation

a)
When we talk about the process of model selection, the issue of bias-variance trade-off becomes very pertinent. On one hand, increasing the number of regressors tends to decrease the bias with respect to estimating the expected mean; on the other, when we add regressors whose true slope is zero, we tend to increase the variance of the estimator of expected mean, especially if the variables are highly related.

In multiple regression, when we add regressors that are highly correlated to each other, we tend to inflate the variance of the estimated regression coefficients. In such instances where regressor variables are correlated among themselves, multicollinearity is said to exist. When the variables are multicollinear, we end up facing "variance inflation". Multicollinearity inflates the standard error of the coefficients of the correlated regressors. As a result, the t-statistics are artificially driven towards the direction of the null hypothesis (that the slope parameter is equal to 0). Furthermore, due to the resulting large sampling variability of the coefficients, the estimated regression coefficients could be very different when we take another sample. As a result, the estimated regression coefficients might not provide correct information about true regression slope.

Variance Inflation Factor (VIF) is a widely used statistic to identify how much the variance is inflated due to the presence of highly correlated regressor variables in the model as opposed to when regressor variables are

not linearly related. When the VIF of a variable is close to 1, we know that the correlation between that variable and other regressor variables is low, when the VIF of a variable is high, the correlation between the variable and other regressor variables are also high. Presence of highly correlated variable does not prevent us from fitting a good model. However, due to increased variance of the estimated regression coefficients, these estimated slopes give us insufficient information about the true stope of the regressor variables.

b)
VIF for variables AvgHHSze, logUnemp, LessthanHS, logBS and logFS were 2.54, 2.24, 3.42, 5.37 and 3.05 respectively. The VIFs for these variables are relatively low, but for a model with this size, VIFs of 5.37 and 3.42 could be problematic. The variables LogBS and LessthanHS are collinear.

The REG Procedure
Model: MODEL1
Dependent Variable: logInc

Number of Observations Read	**190**
Number of Observations Used	190

Analysis of Variance

Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	5	30.72262	6.14452	221.27	<.0001
Error	184	5.10963	0.02777		
Corrected Total	189	35.83226			

Root MSE	**0.16664**	R-Square	**0.8574**
Dependent Mean	10.90520	**Adj R-Sq**	0.8535
Coeff Var	1.52810		

Parameter Estimates

Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	Intercept	**1**	9.03312	0.34029	26.55	<.0001	0
AvgHHSzE	AvgHHSzE	**1**	0.36116	0.03957	9.13	<.0001	2.53678
logUnemp		**1**	-0.19261	0.04599	-4.19	<.0001	2.24051
LessthanHS	LessthanHS	**1**	-0.01888	0.00213	-8.86	<.0001	3.42839
logBS		**1**	0.38158	0.05174	7.37	<.0001	5.37665
logFS		**1**	0.16577	0.05079	3.26	0.0013	3.05286

4.4 Analysis of Output

From the table above, we can also show the $\hat{y}$ line:

$$\widehat{logIncome} = 9.03 + 0.36\ AvgHHSzE - 0.19\ logUnemp - 0.02\ LessthanHS + 0.38\ logBS + 0.17\ logFS$$

From the equation above, we can see that the intercept is 9.03, which is our baseline for logIncome. With everything else being constant, if average household size increases by 1 unit, then we expect Household Income to increase by 36 %. The outcome is not surprising, if we assume that increasing household size implies that there are more working adults in the household, then the household income will also be higher, thus pushing the median household income to be high. On the other hand, if the share of population who are less than high school graduate increases by 1 point, then the Household income will decrease by -2% given everything else is

constant. It can be assumed that people who have less than a high school degree will find jobs that are less paid and it makes sense that at higher shares of population without a high school diploma, the household income will be lower. In these cases, since our dependent variable has been transformed using natural log, a unit change in the independent variable implies that the dependent variable increases by ($\beta_1 * 100$) percent.

When both the dependent and independent variables are transformed using natural logs, we end up measuring elasticity i.e. a percentage change in independent variable causes β_1 percentage change in the dependent variable. In our model, as unemployment share increases by 1%, then Household income decreases by 0.19%. The negative slope estimated by the model is as we had expected (we also briefly talked about it in the simple regression model). It makes sense that high unemployment shares imply that people do not have a source of income, thus the median household income is low. Similarly, as share of population with bachelor's degree increases by 1%, then the Household income increases by 0.38% and when the share of population who are engaged in finance increases by 1%, then the Household income increases by 0.17%. This shows that both variables, share of population who have bachelor's degree and higher, and share of population engaged in finance industry have positive relation with median household income. At higher shares of these variables, the household income is also higher. It should be noted that this does not determine causality. We are only merely interpreting when at higher shares of certain characteristics of the neighborhood, the median household income is more likely to be higher or lower.

The Adjusted R^2 for the model is 0.8535 which tells us that 85.35% variance in the dependent variable, log of Household Incomes of neighborhoods is explained by the independent variables selected in our mode.

4.5 Cook's D

a)

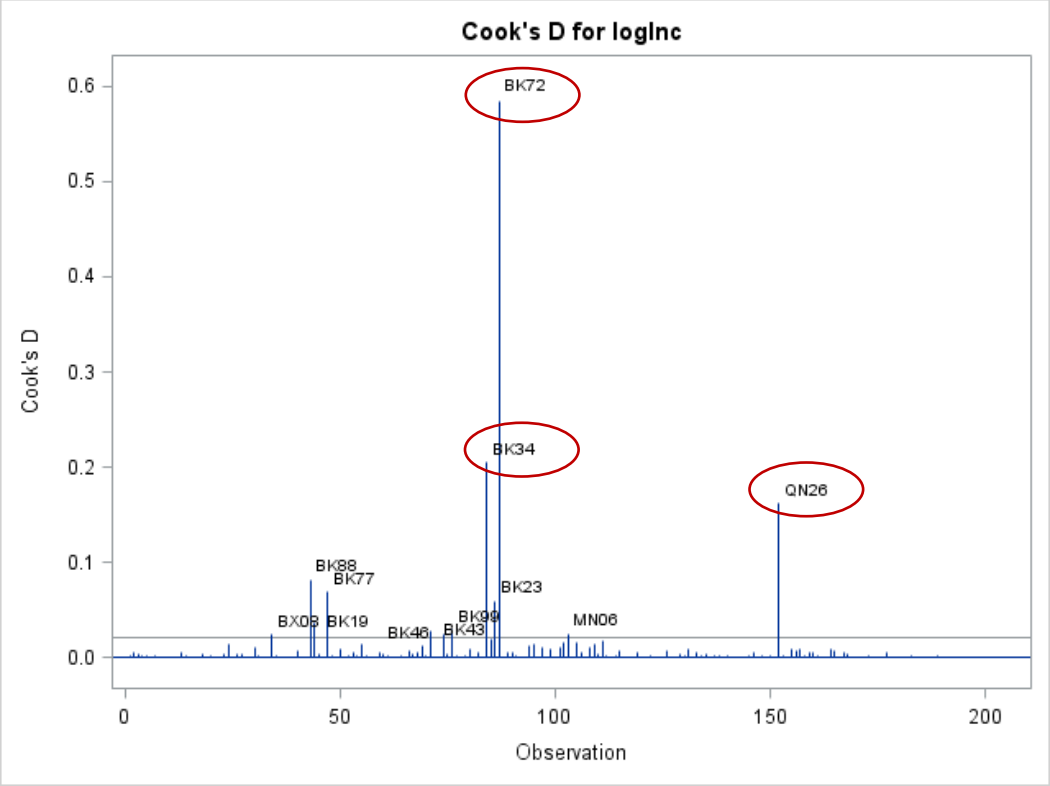
In the previous section, we briefly talked about our model and how certain characteristics of the neighborhoods is related to the median household income. For instance, lower unemployment share implies higher median household income. However, there are always the cases of outliers and leverage points in the dataset. Furthermore, when an observation is both an outlier as well as a leverage point, it is called an influential point.

Cook's D measures the influence of a single observation, on all the fitted values (y-hat values). It shows an aggregate impact of an observation on all the fitted values as opposed to a localized impact of i^{th} observation on the fitted value, $\hat{y}_i$.

To measure Cook's D for the i^{th} observation, we first calculate $\hat{y}$ values for all the n observation. Then we calculate the y-hat values without the i^{th} case in the fitting. Then we take the sum of all the residuals ($\hat{y}$ less the $\hat{y}$ obtained by excluding i^{th} case), square them and divide by MSE times p. MSE stands for Mean Squared Error and p = k +1, where k is the number of predictor variables. This computation is done for every i^{th} observation to measure Cook's D for each observation. The computation can be summarized as below:

$$D_i = \frac{\sum_{j=1}^{n}(\widehat{y_i} - \widehat{y_{j(i)}})^2}{pMSE}$$

b)
Using the 4/n cutoff, there are 12 observations that are flagged as highly influential points by SAS. BK72, Williamsburg, has the highest Cook's D of approximately 0.6, followed by neighborhoods, BK34, Sunset Park East, and QN26, North Corona. All the other high influence points are labeled in the plot below.



Chapter 5. Special Topic - Ridge Regression

1. Understanding Ridge Regression

In Chapter 4, Section 3, we discussed about multicollinearity and variance inflation. In multiple regression, when we add more regressors that are highly correlated to each other, we increase the variance of the regression of coefficients. As the standard error of the slope increases, we artificially drive the t-statistic toward the direction of the null hypothesis that the true slope parameter is equal to 0. As a result, in case of high collinearity, the estimated sample slopes might not provide correct information about the true regression slope. Variance Inflation Factor (VIF) is one of the ways to identify how much the variance is inflated due to the presence of highly correlated regressor variables.

In our model, it was seen that coefficients LessthanHS, share of population who are less than high school graduate, and LogBS, share of population who have bachelor's degree and higher have VIFs of 3.42 and 5.37 each, showing that these variables demonstrate some degree of collinearity in the model. In such cases, where the VIF is high, we can either drop one of the collinear variables or we can decide to keep all the variables, risking high standard error of the slope. In some cases, the collinear variables might be too important to be dropped. One of the ways we can keep all the variables in the model, without having to worry much about variance inflation is by using Ridge Regression.

The concept of Ridge regression can be better explained through the matrix method. As we demonstrated in Chapter 3, the b-vector was calculated as follows:

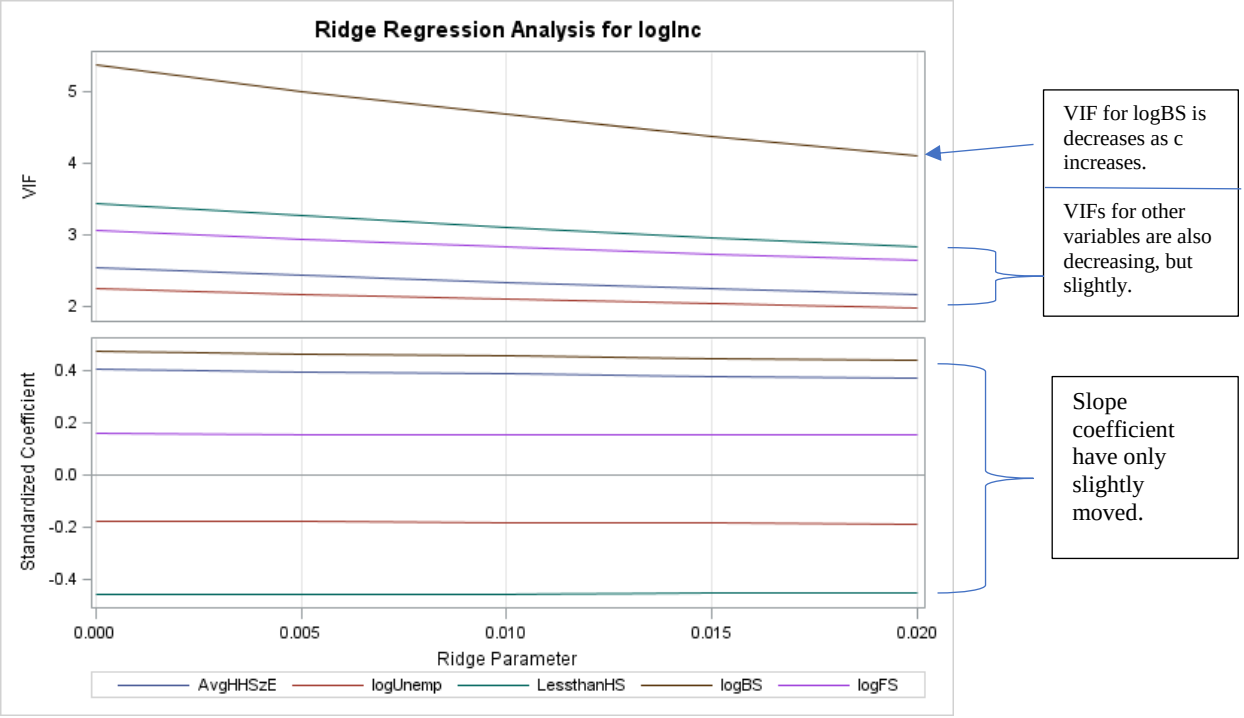
$$\vec{b} = (X'X)^{-1} X' \vec{y}$$

In the case of multicollinearity, the matrix inversion, $(X'X)^{-1}$, is computationally unstable. Therefore, in Ridge Regression, we add a ridge to the X'X matrix to make the computation of its inverse more stable. A ridge is a small positive value added to the diagonal element of the matrix. By doing so, we make the calculation of the matrix inverse more stable, thus allowing us to find an estimate of b-vector. It has been shown that the sample stopes calculated using the ridge method, are better estimator of true population slope, albeit being "biased". The process of added a ridge is shown below:

$$\overrightarrow{b_r(c)} = (X'X + cI)^{-1} X' \vec{y}$$

2. Ridge Trace and VIF

I performed Ridge Regression using SAS on our model that includes 5 variables – Average Household Size, log of Unemployment share, share of population less than high school graduate, log of share of population who hold bachelor's degree and higher and log of share of population engaged in financial services. The value of the constant I provided for the ridge regression is in the range 0.005 to 0.020, which I increased by the increments of 0.005. Below are the graphs and the output for our procedure.



In the graph above, the bottom box is called the ridge trace and it shows the slope coefficients at different values of c. The box above traces the VIFs at different values of c. We can see that as the constant increase, VIF of all the variables consistently decreases. The most dramatic change in the VIF can be seen for the variable, log BS. It should be noticed that the slope estimates for our variables do not change much as the value of c changes; only a minor decrease in slope can be seen in the variables.

3. Output

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_PCOMIT_	_RMSE_	Intercept	AvgHHSzE	logUnemp	LessthanHS	logBS	logFS	logInc
1	MODEL1	PARMS	logInc	.	.	0.16664	9.03312	0.36116	-0.19261	-0.01888	0.38158	0.16577	-1
2	MODEL1	RIDGEVIF	logInc	0.000	.	.	.	2.53678	2.24051	3.42839	5.37665	3.05286	-1
3	MODEL1	RIDGE	logInc	0.000	.	0.16664	9.03312	0.36116	-0.19261	-0.01888	0.38158	0.16577	-1
4	MODEL1	RIDGEVIF	logInc	0.005	.	.	.	2.42968	2.16802	3.26078	5.00887	2.93967	-1
5	MODEL1	RIDGE	logInc	0.005	.	0.16667	9.08927	0.35331	-0.19649	-0.01883	0.37397	0.16502	-1
6	MODEL1	RIDGEVIF	logInc	0.010	.	.	.	2.33120	2.09983	3.10728	4.67863	2.83286	-1
7	MODEL1	RIDGE	logInc	0.010	.	0.16673	9.14194	0.34581	-0.20011	-0.01877	0.36686	0.16433	-1
8	MODEL1	RIDGEVIF	logInc	0.015	.	.	.	2.24033	2.03555	2.96619	4.38093	2.73194	-1
9	MODEL1	RIDGE	logInc	0.015	.	0.16683	9.19146	0.33864	-0.20349	-0.01871	0.36019	0.16369	-1
10	MODEL1	RIDGEVIF	logInc	0.020	.	.	.	2.15620	1.97484	2.83608	4.11161	2.63648	-1
11	MODEL1	RIDGE	logInc	0.020	.	0.16697	9.23812	0.33177	-0.20666	-0.01864	0.35394	0.16310	-1

The table above is the SAS output that shows us the estimated parameters for our model after applying Ridge Regression. We can see that by adding a constant c = 0.005, there is only a slight change in the coefficients of our variables. The highlighted row above shows that the coefficient of variable logBS has decreased slightly from 0.38 to 0.37 (highlighted in yellow in the column titled logBS). At the same time, the VIF has also fallen from 5.38 to 5.01 (highlighted in blue in the column title logBS). At c = 0.020, the coefficient of logBS is 0.35 (green highlight), and the VIF has reached to 4.11(gray highlight). Since the VIF was already relatively low for our variables, adding a small ridge does not change the slope estimate or VIF by much.