Project 1 – Linear Regression Analysis
Modeling Education's Impact on Income

Gunica Sharma

I. Introduction

The main goal of this project is to analyze the relationship between educational level and total personal income across different geographic regions in the United States. I will be using the CDI dataset, which provides data on total personal income, the percentage of the adult population with a bachelor's degree, and the geographical region (North East, North Central, South, and West). In order to do this, I will make use of a Simple Normal Linear Regression Model to explore how the percentage of individuals with a bachelor's degree (predictor variable) influences the per capita income (response variable) in each of the four regions. This analysis aims to identify regional differences in income levels based on education level and judge the strength of the relationship within each region. By doing such it can be better understood how this relationship varies geographically which is interesting for determining if education plays a role in income.

II. Summary

On the data I ran a few analyses, the first one being a histogram of the plots. The first plot I graphed was income distribution by region (Figure 1.1) this analysis showed that North East and South had the highest income distribution overall. Then I ran an analysis on the degree percentage distribution by region (Figure 1.2). Through this I found out that the lowest number of degrees are found in the South were on average 15-20% of people had degrees. Next, I plotted a boxplot of Income by Region (Figure 2.1). Through this I found that the North East on average tended to have the highest income on average. Next, I plotted the Degree Percentage vs Region which suggested that the North East had the highest average of degree percentage (Figure 2.2). Additionally, I plotted a scatterplot of Income vs Degree for all regions on one plot (Figure 3.1). Through this plot it also confirmed that in general the highest income compared to degree percentage was North East. The last analysis I did was calculating the mean income by region, standard deviation by region, mean degree percentage by region and the standard deviation by region. Through this I found that the average income for each region was 18301.09 for north central, 20598.77 for north east, 17486.99 for south, and 18322.58 for the west. This data confirms the idea that the north east has the highest level of income. For the mean degrees, I obtained 19.77500 for north central, 21.78447 for north east, 21.04342 for south, and 22.04675 for the west. From this, it suggests that on average the highest percentage of degrees are found in the west. I also plotted each of the scatterplots and the corresponding regression lines (Figures 4.1-4.4).

III. Model Fitting

I regressed my data to fit onto each of the explanatory variables. The first region I analyzed was the North East, I obtained a variance of 7335008 and an R^2 value of 0.6619. The R^2 value in this case indicates that about 66.19% of the variability of income is explained by the percent of individuals with a degree. This suggests that this is relatively a good fit for the model. For the North Central region, the variance is 4,411,341 and the R^2 is 0.4202. Through this the R^2 value indicates that about 42.02% of the variability in income is explained by the percentage of individuals with a degree, the variance is lower in this region, but the model explains an even

smaller proportion of the variation compared to the North East region. The next region is south which had a variance of 7,474,349 and an R^2 of 0.4975 which suggests that there is 49.75% variability in income that is explained by the percentage of individuals with a degree. In this region there is higher variance suggesting that there is more variability in income that is not explained by the percentage of individuals with a degree. Lastly the West, had a variance of 8,214,318 and 0.5567 for the R^2. This indicates that about 55.67% of the variability in the income may be due to the percentage of individuals with a degree, the variance is the highest in all of the regions which means there is a greater spread of unexplained variability. In conclusion, the best region with the highest R^2 value is the North East, this means that the percentage of adults with a degree is a good predictor of income in this region. I will be choosing this model as the best model.

IV. Diagnostics

I performed the model diagnostics next. I did many tests the first one was plot of residuals vs Xi. The residuals were mostly scattered around the zero-mark indicating that there is linearity and constant variance, however, there were a few outliers labelled 357, 310, and 339 (Figure 5.1). Next, I plotted a histogram of the residuals, for this I got a roughly normal distribution, however, it was very slightly skewed to the right (Figure 5.2). Next, I created a normal QQ plot which checks for the normality of the residuals by comparing observed vs theoretical, for this I got majority of the points were falling where they were supposed to be, however, there were some outliers towards the top and bottom as seen by the (Figure 5.3). I also performed a Shapiro Wilk Test and obtained a p-value of 0.0001605 which suggests that this model does not follow the normality assumption. Then I performed a Durbin-Watson Test, which had a p-value of 0.354 suggesting there is independence. Lastly, I performed the Breusch-Pagan test, for this I got a p-value of 5.769e-05 suggesting the variance of residuals is not constant. This is also a violation of the constant variance. Therefore, the residuals are likely not normally distributed, the residuals are independent, and there is potential non-constant variance.

V. Analysis

I performed the analysis next, for this I created my final model, my confidence intervals, test statistics, p-values, nulls and alternatives, and the R^2. I created the following table for my results.

Statistics	Value
Intercept (β0)	9,223.82
Slope (β1)	522.16
95% CI for β0	[7,534.13, 10,913.50]
95% CI for β1	[448.50, 595.82]
t-statistic for β0	10.83
t-statistic for β1	14.06
p-value for β1	$<2\times10^{-16}$
R^2	0.6619
Residual Standard Error	2,708

F-Statistic	197.8
p-value for Model	<2.2×10^−16

The null hypothesis for was H0: β1 = 0 meaning degree has no effect on income and my alternative was that H1 = β1 does not equal 0 meaning that degree does have an effect on income. The final model I obtained was the following, Income=9223.82+522.16×Degree, this suggests that for each percentage increase in degree the average income increases by $522.16. The 95% confidence interval for the intercept is [7,534.13, 10,913.50], and for the slope it is [448.50, 595.82]. This means that with 95% confidence, each percentage increase in the adult population with a degree is associated with an increase in total income between approximately $448.50 and $595.82. The t-statistic for the slope is 14.06 with a corresponding p-value of <2×10^16 suggesting that the percentage of adults with a degree has a statistically significant effect and we reject the null hypothesis showing that degree does have an effect on income. The R^2 value of 0.6619 indicates that approximately 66.19% of the variation in income is explained by the percentage of adults with a degree in the North East region. The residual standard error of the model is 2,708, and the model is statistically significant as indicated by the F-statistic of 197.8.

VI. Inference

The analysis of the North East region shows a statistically significant positive relationship between the percentage of adults with a degree and total personal income. For each percentage increase income rises by approximately $522. The confidence interval [448.50, 595.82] reflects this effect, indicating that higher education leads to higher income. With an R^2 of 0.66, the model explains 66% of the income variation, highlighting education as a key factor influencing income. However, violations of normality and constant variance assumptions suggest some limitations, indicating that other factors not included in the model may also impact income levels. These findings support the idea that education plays a crucial role in economic wellbeing.

VII. Predictions

Through my model I ran the following predictions

Income=9223.82+522.16×15=9223.82+7832.40=17,056.22 this suggests when 15% of the population has a bachelors degree the average annual income is $17,056.

Income=9223.82+522.16×21=9223.82+10,965.36=20,189.18 this suggests that when 21% of the population holds a bachelors degree the predicted income is $20,189.

Income=9223.82+522.16×25=9223.82+13,054.00=22,277.82 this suggests that when 25% of the population holds a bachelors degree the predicted income is $22,278.

VIII. Conclusion

In this analysis, I found a statistically significant positive relationship between the percentage of adults with a bachelor's degree and total personal income in the North East region. Higher education was associated with increased average income, indicating the importance of education and its role in economic well-being. However, one limitation of the model is the observed

violations of normality and constant variance assumptions, which suggest that the model may not fully account for all of the different factors that can influence income. These factors may be the industry someone is in, cost of living, and socioeconomic status. These were not included in the model and could impact the results. Future analysis could address these limitations by incorporating additional variables for an even better analysis.

Figures

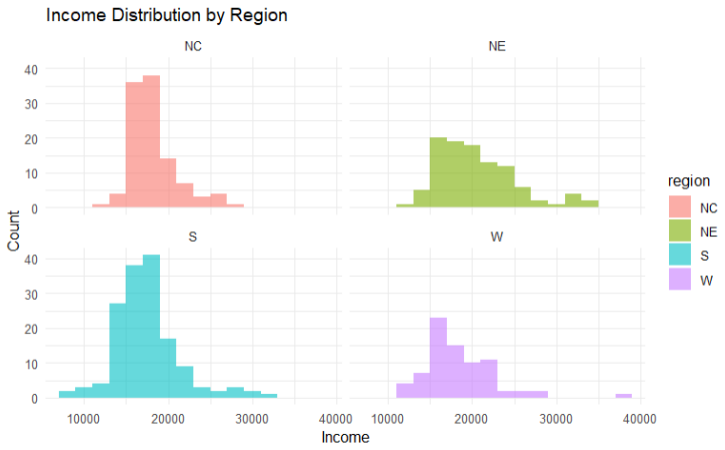

Figure 1.1

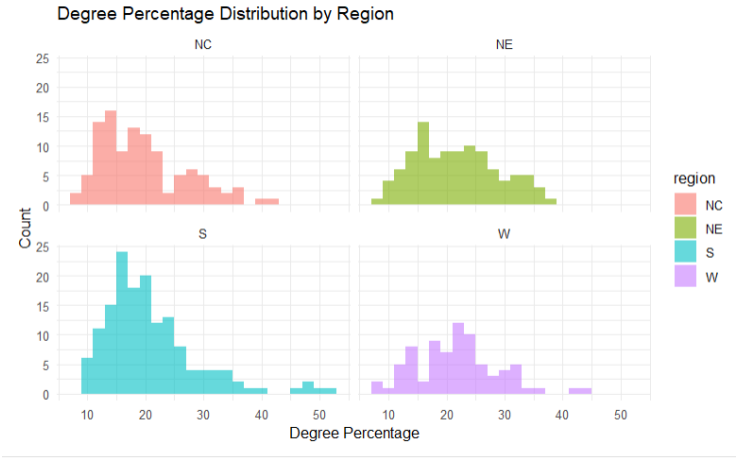

Figure 1.2

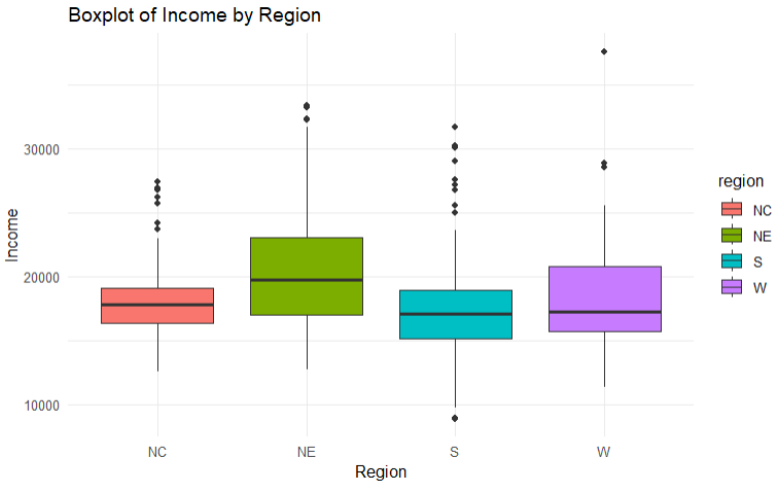

Figure 2.1

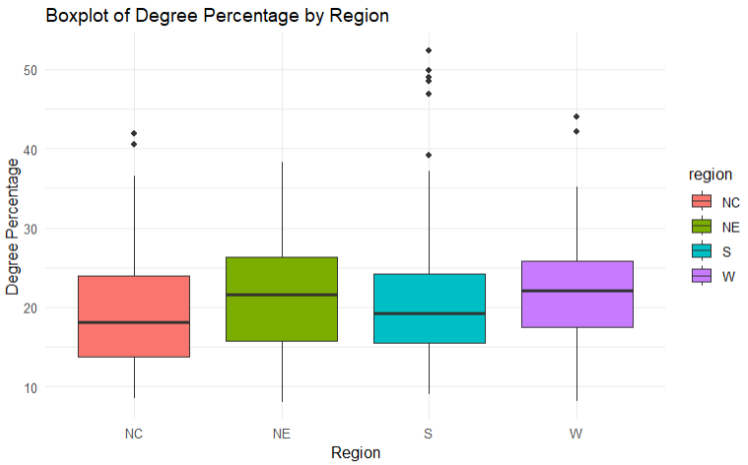

Figure 2.1

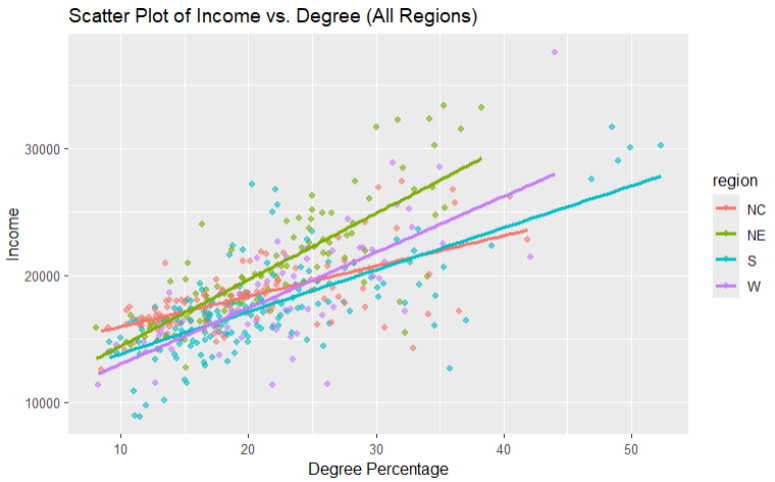

Figure 3.1

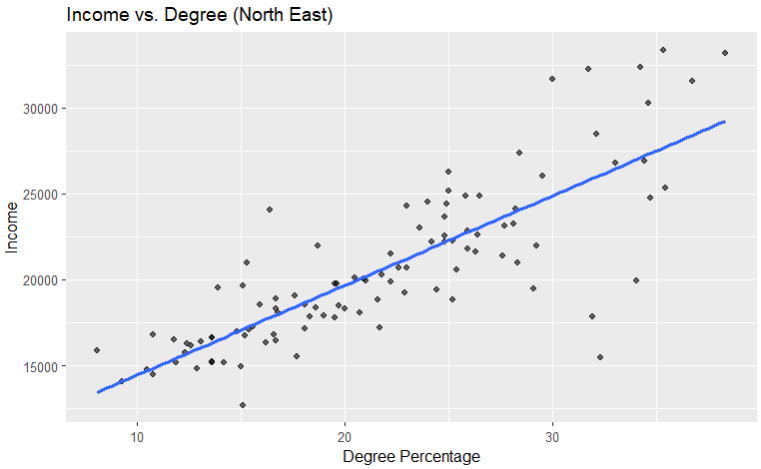

Figure 4.1

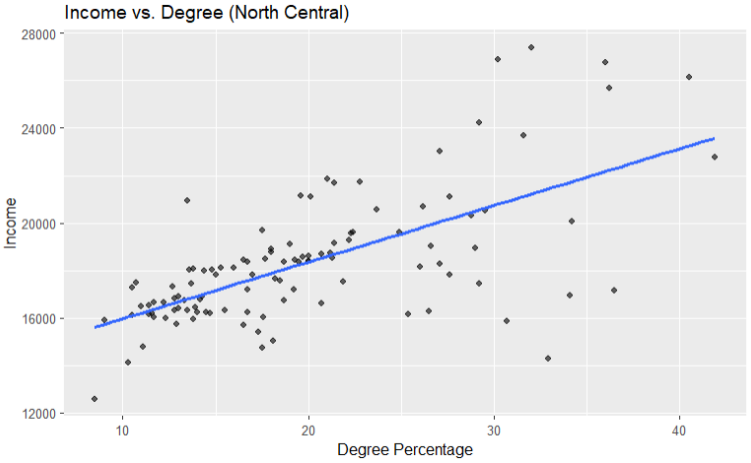

Figure 4.2

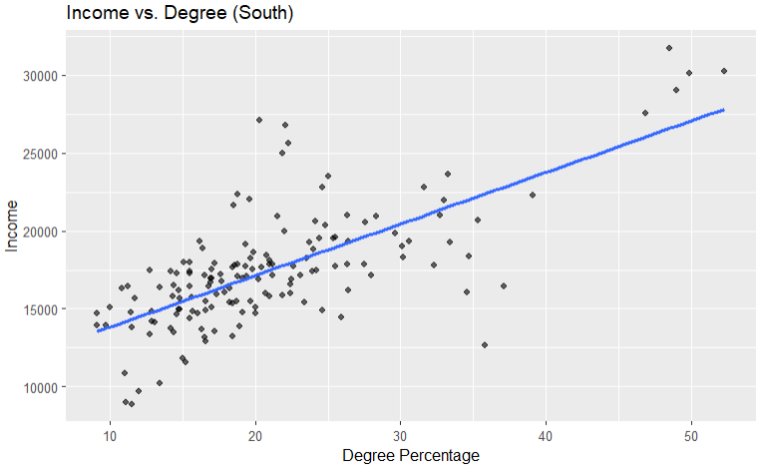

Figure 4.3

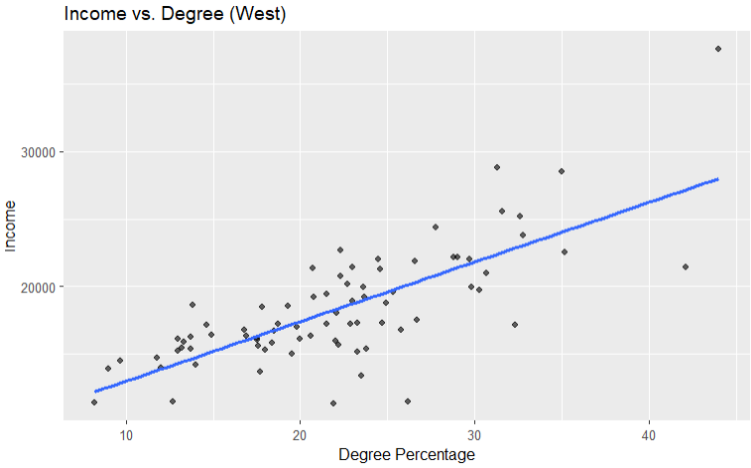

Figure 4.4

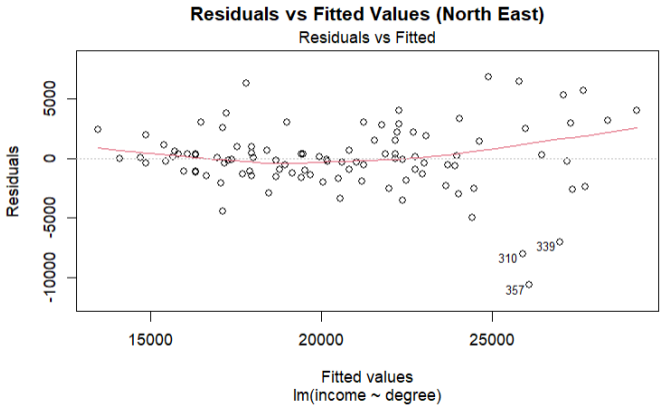

Figure 5.1

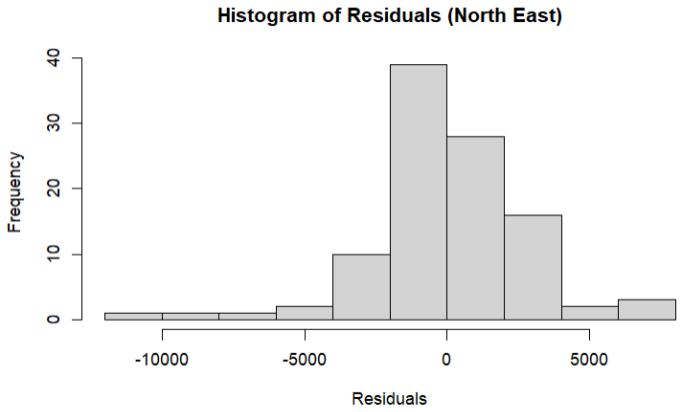

Figure 5.2

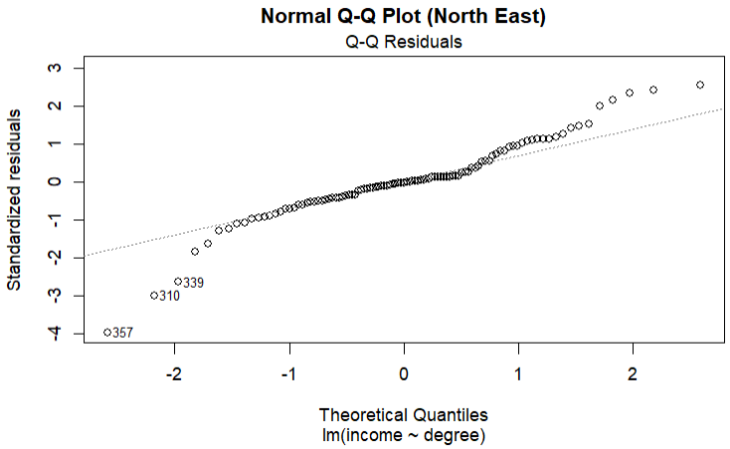

Figure 5.3