Regression Analysis Project

By: Ethan Hershman

Report:

I. Introduction

Our goal is to find the statistically best model for the percentage of individuals in a county that have at least a bachelor's degree compared to per capita income for each geographic region (West, South, Northern Central, North Eastern). To reach this goal, we hope to compare four simple linear models using the CDI data based on each region. The data seems interesting as it may show a relationship to how higher education can impact income across different geographic regions. Personally, as students, it is interesting to see how furthering our education can impact our future income as well. Furthermore, the analysis can provide insight into economic inequality, such as disparities in income and access to education, and can be shown in our report.

II. Summary

Our summary plots are scatterplots of the percentage of individuals in a county that have at least a bachelor's degree to their income per capita. In all four regions (West, South, Northern Central, and North Eastern), the scatterplots have a positive correlation. However, for each scatterplot, there appears to be outliers and some variance between the variables.

As for numerical summaries, we decided to extract the mean and standard deviation values for each region's income. To compare by regions, the North East has the highest mean income of 20598.77 dollars and the South has the lowest mean income of 17846.99 dollars. When looking at the standard deviation of income for each region, the North East had the highest standard deviation of 4635.07 dollars, while the Northern Central region had the lowest standard deviation of 2745.45 dollars. Using these statistics, we can infer that the North East has a greater spread of income values, with less values clustered around its mean income of 20598.77 dollars. Looking at the Northern Central region, we can infer that this region's values have a smaller spread and have more values clustered around the mean income of 18301.09 dollars. Based on the summary models and the numerical summaries, we are able to see that there is a positive, and possibly linear, trend between the percentage of individuals in a county that have at least a bachelor's degree (X) to their income per capita (Y).

III. Model Fitting

We can further test in our model fitting whether a linear regression model is the best fit for each of the scatterplots.

For the West, the estimated linear regression line was $Y = 440.32X + 8615.05$, with an estimated variance of 8214318 dollars. The South had an estimated linear regression line of $Y = 330.61X + 10529.79$, and an estimated variance of 7474349 dollars. As for Northern Central, the estimated linear regression line was $Y = 238.67X + 13581.41$, and estimated variance as 4411341 dollars. Lastly, the North East had an estimated linear regression line of $Y = 552.16X + 9223.82$, with estimated standard deviation of 7335008 dollars. The West has the highest variance of 8214318 dollars, and the Northern Central has the lowest variance of 4411341 dollars.

The decision for the best model can be backed by the R^2 values: 0.5567 for Western region, 0.4975 for Southern region, 0.4202 for Northern Central, and 0.6619 for the North Eastern region. We can see that the best region for predicting income per capita by percent of individuals with at least a bachelor's degree is the North Eastern region. We will move forward with the **model for the North East** as this has the highest R^2 value of 0.6619, showing that this model explains the variation in the Income per capita (Y) based on the Percent of the adult population with at least a bachelor's degree (X) the best.

IV. Diagnostics

The diagnostic plots performed were a residual plot, histogram of residuals, and a normal probability/Q-Q plot. From these plots, we also tested for normality and constant variance.

From the residual graph, there appears to be a few outliers, with greater variance in the shape of a "cone" as the x-value increases. This would not be ideal because the errors may suggest non-constant variance, non-normality, or a possible non-linear pattern. Based on the histogram of residuals, we can see that this histogram is left skewed with a couple outliers. Although the histogram appears normally distributed, we must also consider the outliers, which makes a judgment call that the deviation may be too severe and violates the diagnostic. Based on the Q-Q plot, the plot appears approximately normal, but has several outliers, as seen by the data above and below on the tail ends. By considering the outliers in the Q-Q plot, this would suggest non-normality instead. Overall, if there were no outliers present in the data, then our assumptions would not be violated, but without the removal of outliers, our assumptions of linearity, normality, and constant variance are violated.

V. Analysis

The income per capita by percentage of individuals with at least a bachelor's degree in the North Eastern region is best represented by the line: $Y = 552.16X + 9223.82$. We decided this was the best model to use as the R^2 value for the North Eastern region was the highest amongst all regions within the dataset. So, for every percent increase in individuals with at least a bachelor's degree, the income per capita will increase by 552.16 dollars on average. The 95% confidence interval for an alpha = 0.05 for the slope of the income is: (448.5001, 595.8176). We also performed several hypothesis tests to analyze the fit of our model.

For the Shapiro Wilks Test, the null hypothesis was the population is normally distributed, and alternative hypothesis is population is not normally distributed. The p-value for the Shapiro Wilks Test was 0.0001605, and t-statistic of 0.94034.

For the Fligner-Killeen Test, the null hypothesis is that the variances of the two groups are equal, and the alternative hypothesis is that the variances of the two groups are not equal. The p-value for the Fligner-Killeen Test was 0.2577, and t-statistic of 94.11065.

For the Linear Relationship Test, the null hypothesis is there is not a significant linear relationship between X and Y, and the alternative hypothesis is there is a significant linear relationship between X and Y. The t-statistic for the linear relationship test was 14.06246, and p-value of $1.589 * 10^{-25}$.
Lastly, it is important to note that the R^2 value for the North Eastern region (best model) was 0.6619.

VI. Interpretation

Assuming an alpha level of 0.05 and p-value of 0.0001605, which is less than alpha = 0.05, from the Shapiro Wilks Test, we can conclude that the residuals are not normally distributed. There is sufficient evidence to reject H0 and suggest that the population is not normally distributed.

Assuming an alpha level of 0.05 and a p-value of 0.2577, which is greater than alpha = 0.05, from the Fligner-Killeen Test, we fail to reject H0, and we can conclude that there is constant variance.

Assuming an alpha level of 0.05 and a p-value of 1.59 * 10^-25, which is less than alpha = 0.05, from the linear relationship test, we can reject H0 and conclude that there is a significant linear relationship between the percent of adults who have at least a bachelor's degree and per capita income.

We are 95% confident that when the percentage of the adult population with at least a bachelor's degree is changed by a percent, the estimated change in income per capita is between 448.50 and 595.82 dollars on average.

VII. Prediction Results

We predict the average total personal income when the percent of the adult population is 15 is 17056.20 dollars. We are 95% confident that the income when the percent of the adult population with a bachelor’s degree is 15 is between 16328.20 and 17784.19 dollars.

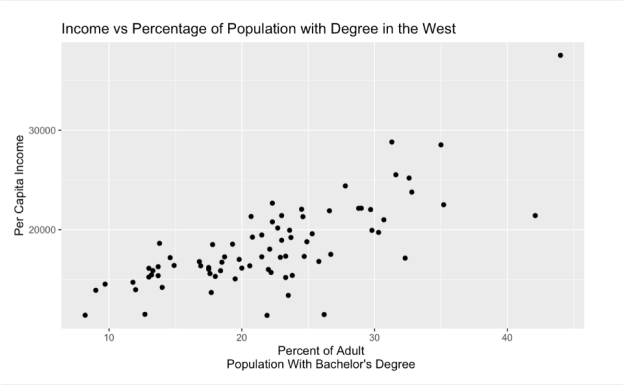
We predict the total personal income of 21 percent of the adult population is 20189.15 dollars. We are 95% confident that the total personal income when the percent of adult population is 21 is between 14790.24 and 25588.06 dollars. The width of the prediction interval is wider than the confidence interval above because the prediction is of an individual response, which takes account of more randomness. So, the prediction interval takes into account the uncertainty and variability of individual outcomes.

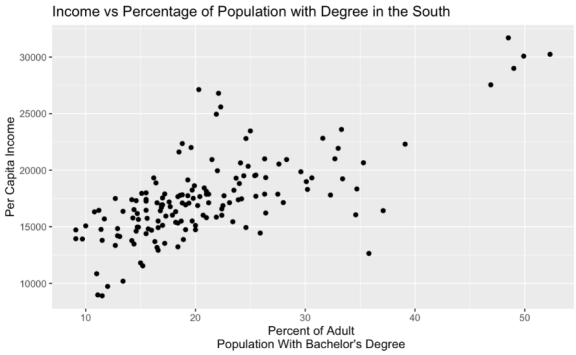
We predict the average total personal income when the percent of the adult population is 25 is 22277.79 dollars. We are 95% confident that the income when the percent of the adult population with a bachelor’s degree is 25 is between 21697.84 and 22857.73 dollars.

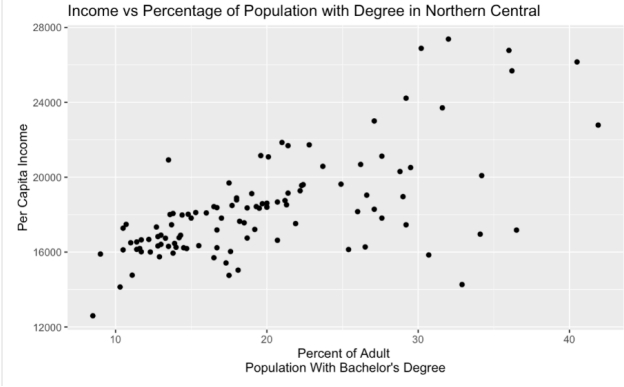
VIII. Conclusion

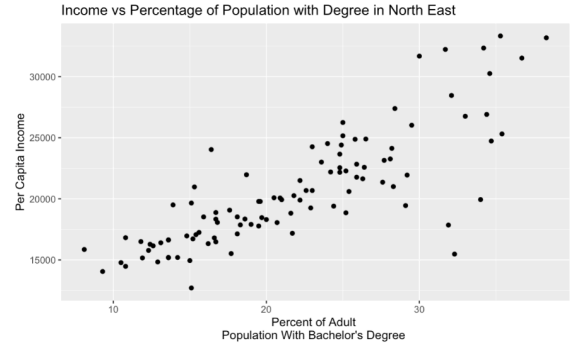
From our findings, the North East model is the best fit model to show a positive linear relationship between the percentage of adults who have at least a bachelor's degree and their income per capita. While all the models showed a positive linear relationship, there is sufficient evidence to conclude that the North East Model has the strongest linear relationship due to its highest R^2 value. After performing the model diagnostics and checking assumptions on the North East model, we concluded that all the assumptions were fairly violated except that the errors were independent, as shown by non-linearity, non-normality, and non-constant variance in the residual plot, histogram of errors, and Q-Q plot. Through three different kinds of hypothesis testing, we conclude that there is a significant linear relationship between per capita income and the percentage of individuals who have at least a bachelor’s degree. However, we should note a limitation is that due to the increased variance observed at the higher percentages, this model would not be effective at predicting the Income per Capita for populations with high percentages of adults with at least a bachelor's degree.

II. Summary Plots and Numerical Summaries



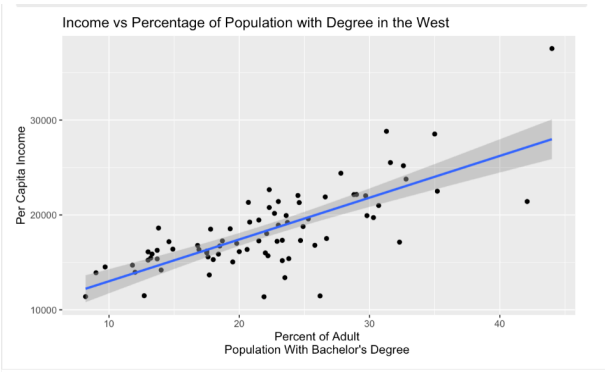


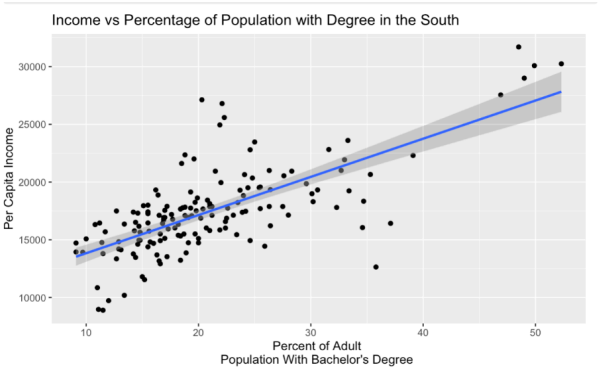


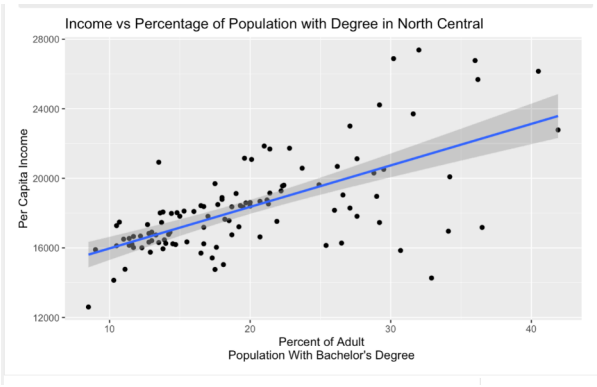


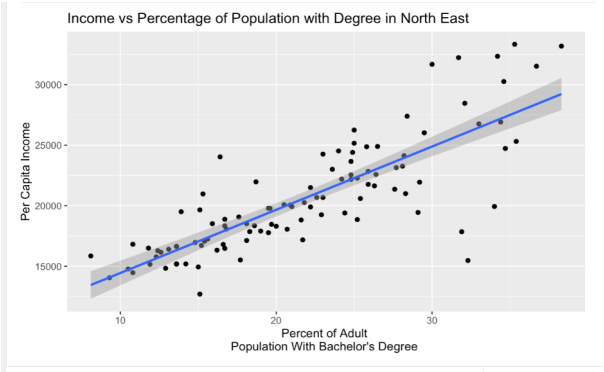
Sample Means:	Standard Deviations:
West: 18322.58	West: 4276.342
South: 17486.99	South: 3843.827
North Central: 18301.09	North Central: 2745.446
North East: 20598.77	North East: 4635.068

III. Model Fitting









Summary Table for West Region

```
Coefficients:
(Intercept)       degree
     8615.1        440.3


Call:
lm(formula = income ~ degree, data = west)

Residuals:
    Min      1Q  Median      3Q     Max
-8684.3 -1477.3   191.7  1557.8  9552.1

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept)  8615.05    1052.20   8.188 0.00000000000524227 ***
degree        440.32      45.37   9.705 0.00000000000000686 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2866 on 75 degrees of freedom
Multiple R-squared:  0.5567,    Adjusted R-squared:  0.5508
F-statistic: 94.19 on 1 and 75 DF,  p-value: 0.000000000000006856
```

Summary Table for South Region

```
Coefficients:
(Intercept)       degree
    10529.8        330.6


Call:
lm(formula = income ~ degree, data = south)

Residuals:
    Min      1Q  Median      3Q     Max
-9724.7 -1362.8   114.9  1255.6  9883.8

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept) 10529.79     612.48   17.19 <0.0000000000000002 ***
degree        330.61      27.13   12.19 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2734 on 150 degrees of freedom
Multiple R-squared:  0.4975,    Adjusted R-squared:  0.4941
F-statistic: 148.5 on 1 and 150 DF,  p-value: < 0.00000000000000022
```

Summary Table for Northern Central Region

```
Coefficients:
(Intercept)       degree
    13581.4        238.7


Call:
lm(formula = income ~ degree, data = norcen)

Residuals:
    Min      1Q  Median      3Q     Max
-7167.6  -915.4   105.6   886.6  6159.2

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept) 13581.41     575.14  23.614 < 0.0000000000000002 ***
degree        238.67      27.23   8.765   0.0000000000000334 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2100 on 106 degrees of freedom
Multiple R-squared:  0.4202,    Adjusted R-squared:  0.4147
F-statistic: 76.83 on 1 and 106 DF,  p-value: 0.00000000000003344
```

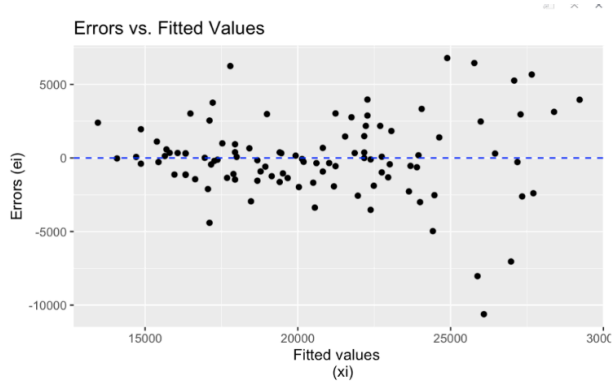
Summary Table for North East Region

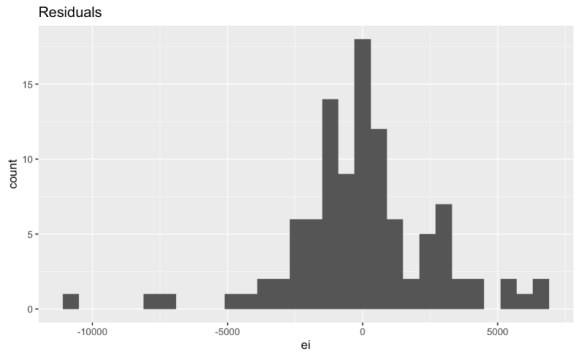
```
Call:
lm(formula = income ~ degree, data = noreast)

Residuals:
     Min       1Q   Median       3Q      Max
-10613.5  -1276.2    -68.9   1256.6   6790.4

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept)  9223.82     851.77   10.83 <0.0000000000000002 ***
degree        522.16      37.13   14.06 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2708 on 101 degrees of freedom
Multiple R-squared:  0.6619,    Adjusted R-squared:  0.6586
F-statistic: 197.8 on 1 and 101 DF,  p-value: < 0.00000000000000022
```

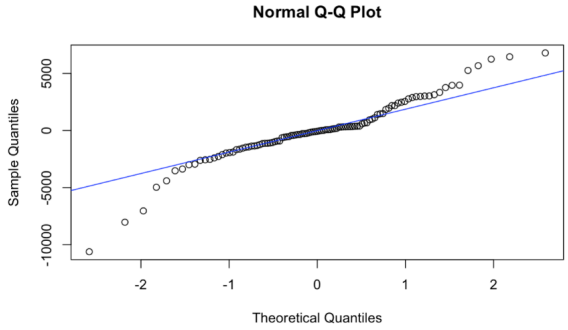
Model:	Estimated Variances:
Model 1 → West	8214318 dollars
Model 2 → South	7474349 dollars
Model 3 → North Central	4411341 dollars
Model 4 → North East	7335008 dollars

Model:	R^2 Values:
Model 1 → West	0.5567
Model 2 → South	0.4975
Model 3 → North Central	0.4202
Model 4 → North East	0.6619

IV. Diagnostics







V. Analysis

Hypothesis Tests:	
Shapiro-Wilks Test	H0: The population is normally distributed. HA: The population is not normally distributed. p-value = 0.0001605 t-statistic: 0.94034 Conclusion: Reject null hypothesis
Fligner-Killeen Test	H0: The variances of the two groups are equal. HA: The alternative hypothesis is that the variances of the two groups are not equal. p-value = 0.2577 t-statistic: 94.11065 Conclusion: Fail to reject null hypothesis
Linear Relationship Test	H0: There is not a significant linear relationship between X and Y. HA: There is a significant linear relationship between X and Y. p-value = 1.589 * 10^-25 t-statistic = 14.06246 Conclusion: Reject null hypothesis

R^2	0.6619
Confidence Interval for Slope	(448.50 , 595.82)

R Appendix

```
library(tidyverse)
library(ggplot2)
cdi <- read.csv("~/Downloads/CDI.csv")
west <- cdi %>% filter(region == "W")
south <- cdi %>% filter(region == "S")
norcen <- cdi %>% filter(region == "NC")
noreast <- cdi %>% filter(region == "NE")
#II. SUMMARY
qplot(degree, income, data = west) + labs(title =
"Income vs Percentage of Population with Degree in the West", x = "Percent of Adult
Population With Bachelor's Degree ", y = "Per Capita Income")
qplot(degree, income, data = south) + labs(title =
"Income vs Percentage of Population with Degree in the South", x = "Percent of Adult
Population With Bachelor's Degree ", y = "Per Capita Income")
qplot(degree, income, data = norcen) + labs(title =
"Income vs Percentage of Population with Degree in Northern Central", x = "Percent of Adult
Population With Bachelor's Degree", y = "Per Capita Income")
qplot(degree, income, data = noreast) + labs(title =
"Income vs Percentage of Population with Degree in North East", x = "Percent of Adult
Population With Bachelor's Degree ", y = "Per Capita Income")
#Numerical Summaries
#Sample Means
mean(west$income)
mean(south$income)
mean(norcen$income)
mean(noreast$income)
```

```r
#Standard Deviation
sd(west$income)
sd(south$income)
sd(norcen$income)
sd(noreast$income)
#III. MODEL FITTING
qplot(degree, income, data = west) + geom_smooth(method = 'lm') + labs(title =
"Income vs Percentage of Population with Degree in the West", x = "Percent of Adult
Population With Bachelor's Degree ", y = "Per Capita Income")
model = lm(income ~ degree, data = west)
model
summary(model)
sd_squared <- summary(model)$sigma^2
sd_squared
qplot(degree, income, data = south) + geom_smooth(method = 'lm') + labs(title =
"Income vs Percentage of Population with Degree in the South", x = "Percent of Adult
Population With Bachelor's Degree ", y = "Per Capita Income")
model2 = lm(income ~ degree, data = south)
model2
summary(model2)
sd_squared2 <- (summary(model2)$sigma)^2
sd_squared2
qplot(degree, income, data = norcen) + geom_smooth(method = 'lm') + labs(title =
"Income vs Percentage of Population with Degree in North Central", x = "Percent of Adult
Population With Bachelor's Degree ", y = "Per Capita Income")
model3 = lm(income ~ degree, data = norcen)
model3
summary(model3)
sd_squared3 <- (summary(model3)$sigma)^2
sd_squared3
qplot(degree, income, data = noreast) + geom_smooth(method = 'lm') + labs(title =
"Income vs Percentage of Population with Degree in North East", x = "Percent of Adult
Population With Bachelor's Degree ", y = "Per Capita Income")
model4 = lm(income ~ degree, data = noreast)
model4
summary(model4)
sd_squared4 <- summary(model4)$sigma^2
sd_squared4
#IV. DIAGNOSTICS
#Plot ei vs. Xi
#Get list of residuals for each model
ei = resid(model4)
xi = fitted(model4)
ggplot(noreast, mapping = aes(x = xi, y = ei)) + geom_point() + geom_hline(
  yintercept = 0, linetype = "dashed", color = "blue") + labs(x = "Fitted values
  (xi)", y = "Errors (ei)") + ggtitle("Errors vs. Fitted")
#Histogram of ei
ggplot(model4, mapping = aes(x = ei)) + geom_histogram() + ggtitle("Residuals")
#QQ Plot
qqnorm(ei)
qqline(ei, col = "blue")
#Test for Normality
y = shapiro.test(ei)
```

```r
y$statistic
#Test for Constant Variance
x = fligner.test(ei ~ xi)
x$statistic
#Test for Linear Relationship
coef(summary(model4))[2,4]
#Test-Statistics
coef(summary(model4))[2,3]
summary(model4)
#Confidence Intervals
CIs = confint(model4, level = 0.95)
CIs[2,]
# VII. PREDICTION RESULTS
intercept <- coef(model4)[1]
slope <- coef(model4)[2]
#percent = 15
intercept + slope * 15
predict(model4, data.frame(degree = 15), interval = "confidence", level = 0.95)
#percent = 21
intercept + slope * 21
predict(model4, data.frame(degree = 21), interval = "prediction", level = 0.95)
#percent = 25
intercept + slope * 25
predict(model4, data.frame(degree = 25), interval = "confidence", level = 0.95)
```