STA 108 Project 1

Abstract:

In this project, we are interested in analyzing the relationship (or lack thereof) between the number of active physicians against three predictor variables. These are the total population, the total number of beds and the total income. The data set has been collected from 440 different areas split into 4 regions, with each area having independent data surrounding these variables. Regression model (1.1) will be firstly used to test the three relationships and reviewed to see whether this type of model was a good fit. After going through the models, we will then evaluate each relationship's linear association by calculating the Coefficient of Determination. This will allow us to determine how strong each model is for each relationship. We will then make inferences about regression parameters surrounding the relationships between per capita income against the percentage of individuals in a county having at least a bachelor's degree. This will be done by finding confidence intervals for each of the 4 regions, running an Analysis of Variance and using F-Tests. Having assumed many assumptions in order to regress these relationships using the linear regression model (1.1), it is important to test whether the data supports these assumptions. Thus we shall also conduct tests to determine whether the linear regression model (2.1) is a more appropriate model to use for each of the first 3 fitted regression models.

Part 1: Fitting Regression Models

Having assumed that the first-order regression model (1.1) is appropriate, we regressed the total number of active physicians against the three predictor variables. The three variables are the total population, number of hospital beds, and total personal income.

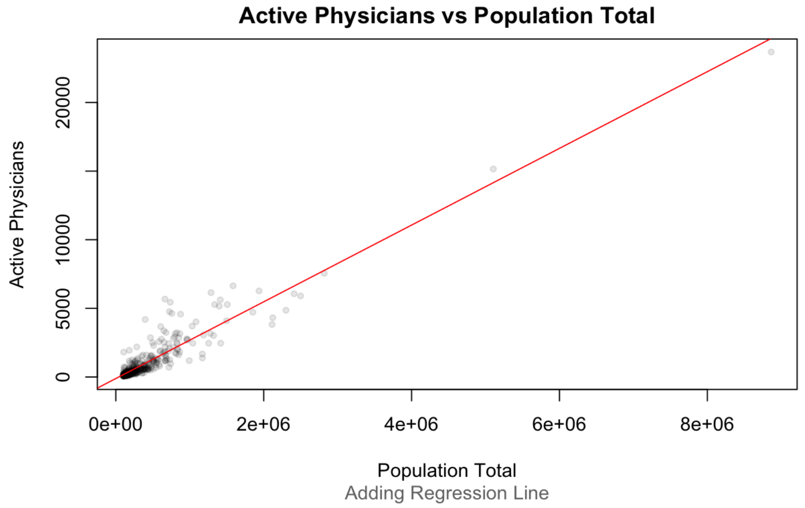
Model 1:
Active physicians (Y) regressed against the population total(X).
$\hat{Y} = -110.6 + 0.002795X$

Model 2:
Active physicians (Y) regressed against the number of hospital beds(X).
$\hat{Y} = -95.9322 + 0.7431X$

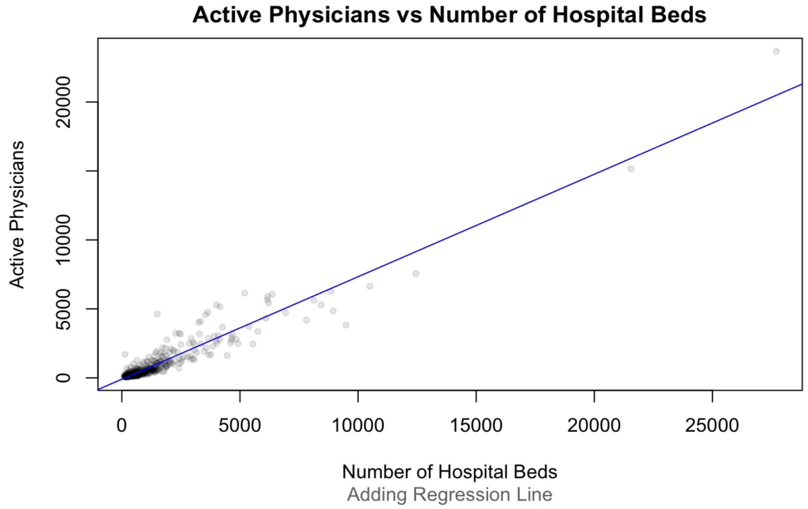
Model 3:
Active physicians (Y) regressed against total personal income(X).
$\hat{Y} = -48.3948 + 0.1317X$

Each of these models were plotted on separate graphs and reviewed to determine whether a linear regression relation appeared to provide a good fit for each of the three predictor variables.

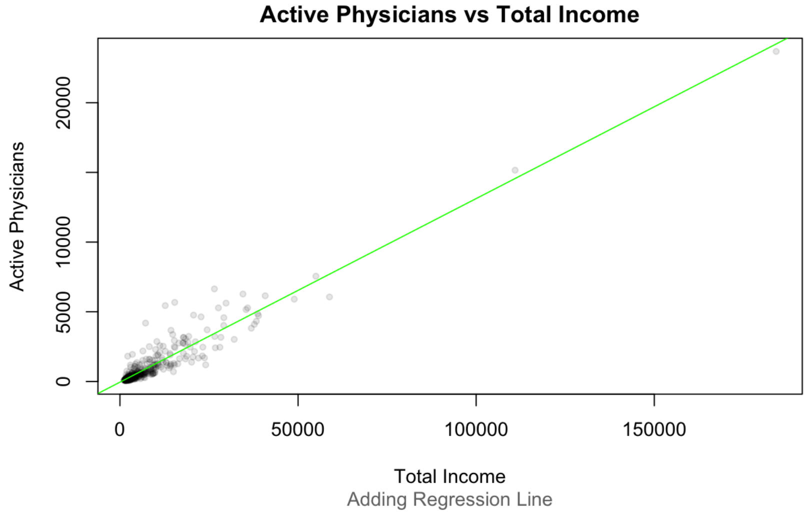
Model 1:



Model 2:



Model 3:



Based on the graphs and estimated regression models, a linear regression relation does appear to provide a good fit for each of the three predictor variables. The regression lines appear to go through all the data in an even split.

The Mean Squared Error measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values. This means that the larger the MSE value, the larger the model error.

The formula to calculate MSE is as follows:

$$MSE = \frac{SSE}{n-2} = \frac{\sum e_i^2}{n-2}$$

Where n is the number of observations, and e_i^2 is the square of each individual error.

Each of the three predictor variables have large MSE values.

Model 1 (Total Population): 372203.5
Model 2 (Total Number of Beds): 310191.9
Model 3 (Total Income): 324539.4
From these results, we can conclude that the predictor variable of Total Income leads to the smallest variability around the fitted regression line. This is due to it having the smallest MSE value of the 3 predictor variables. The fact that the MSE's are so large indicates the variance or the bias in the estimators are also large.

Part 2: *Measuring Linear Associations*

R^2 is the Coefficient of Determination. It measures how well a statistical model predicts an outcome. It is the proportion of the variance for a dependent variable that's explained by an independent variable in a regression model. To calculate the R^2 criterion, the summary function in RStudio was used.

Our results by calculating the Coefficient of Determination for each of the three predictor variables were as follows:

Total Population: 0.8841
Total Number of Beds: 0.9034
Total Income: 0.8989

The higher the value of R^2, the better the model. In this case, the results show the predictor variable of Total Number of Beds accounts for the largest reduction in the variability in the number of active physicians.

Part 3: *Inference about regression parameters*

In this part, we are trying to obtain an interval estimate for the β_1 value of each of the 4 regions using a 90% confidence coefficient. The confidence coefficient is simply the proportion of samples of a given size that may be expected to contain the true mean.

The β_1 values were previously obtained when the regressing Y value of per capita income against the X value of percentage of individuals in a county having at least a bachelor's degree. The 4 β_1 values in region order are 522.2, 238.7, 330.6 and 440.3.

Assuming Normality, the following formula is used to calculate the confidence interval.

$$\hat{\beta}_1 \pm t_{n-2}\left(1-\frac{\alpha}{2}\right) \text{s.e.}(\hat{\beta}_1),$$

where

$$\text{s.e.}(\hat{\beta}_1) = \sqrt{\frac{\text{MSE}}{\sum_i (x_i - \bar{x})^2}}.$$

In our case, we used an alpha value of 0.1.

The results for each region were as follows:
Region 1: [460.518, 583.8]
Region 2: [193.486, 283.853]
Region 3: [285.708, 375.516]
Region 4: [364.759, 515.873]

To understand these regression models, we can perform statistical analysis to further test the degree of differences between two or more groups of an experiment. The Analysis of Variance can be carried out for each of the regression models to review how the sum of squares are distributed according to source of variation. This in turn will let us figure out whether the regression model employed provides a better fit than a model that contains no independent variables (F-Test).

ANOVA TABLES:

Region 1:

Variation	Sum Squared	Mean Square	Degree of Freedom	F value
Regression	1450517671	1450517671	1	197.75
Error	740835765	7335008	101	
Total	2191353436	1457852679	102	

Region 2:

Variation	Sum Squared	Mean Square	Degree of Freedom	F value
Regression	338907694	338907694	1	76.826
Error	467602149	4411341	106	
Total	806509843	343319035	107	

Region 3:

Variation	Sum Squared	Mean Square	Degree of Freedom	F value
Regression	1109873245	1109873245	1	148.49
Error	1121152411	7474349	150	
Total	2231025656	1117347594	151	

Region 4:

Variation	Sum Squared	Mean Square	Degree of Freedom	F value
Regression	773745787	773745787	1	94.195
Error	616073841	8214318	75	
Total	1389819628	781960105	76	

With the results, we can now do F-tests for each region and determine whether the percentage of individuals in a county having at least a bachelor's region and per capita income are jointly significant.

$H_0 : \rho = 0$, $H_1 : \rho \neq 0$. We reject the null hypothesis, if the p-value of the F test is smaller than the alpha value used or if the F value calculated from the data is larger than the F critical value. The P value is the probability of getting a result at least as extreme as the one that was actually observed

Through the ANOVA function in R, we were also able to receive the p-values for each F value. The alpha value used for the F tests was 0.05, meaning if the p-values is less than this, then we can determine that our results are significant and can reject the null hypothesis.

For Region 1: The p-value was $<2.2\text{x}10^{-16}$ which is smaller than 0.05. Therefore we can reject H_0 and conclude that the two variables are not statistically independent from each other.

For Region 2: The p-value was $3.344\text{x}10^{-14}$ which is smaller than 0.05. Therefore we can reject H_0 and conclude that the two variables are not statistically independent from each other.

For Region 3: The p-value was $<2.2\text{x}10^{-16}$ which is smaller than 0.05. Therefore we can reject H_0 and conclude that the two variables are not statistically independent from each other.

For Region 4: The p-value was 6.856x10-15which is smaller than 0.05. Therefore we can reject H_0 and conclude that the two variables are not statistically independent from each other.

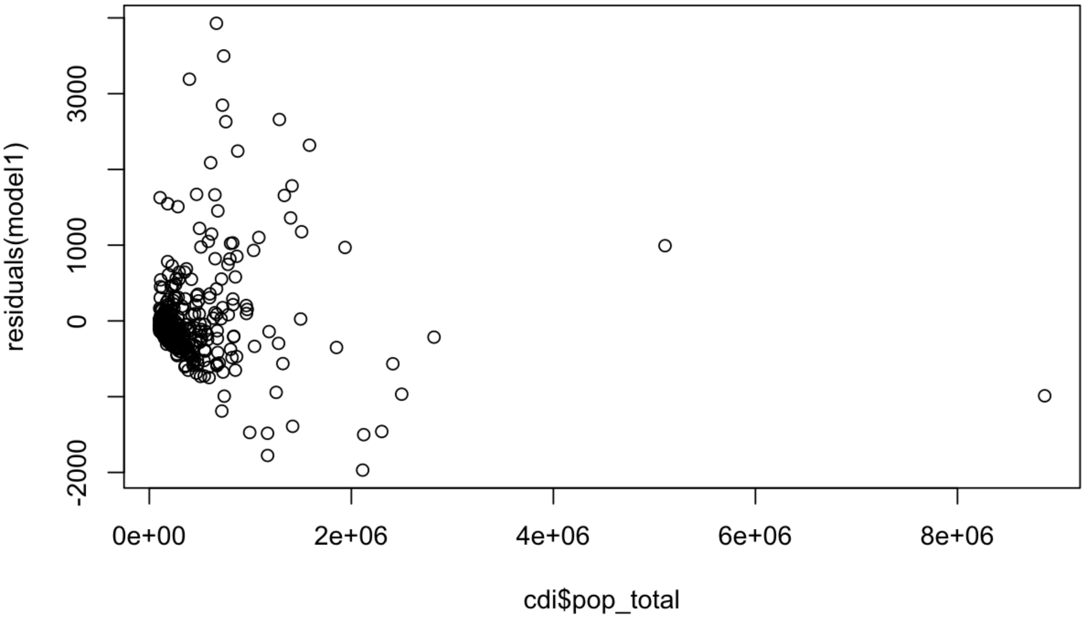
From the F-tests, we can gather that the percentage of individuals in a county having at least a bachelor's region and per capita income are jointly significant for all 4 regions.

<u>Part 4:</u> Regression Diagnostics

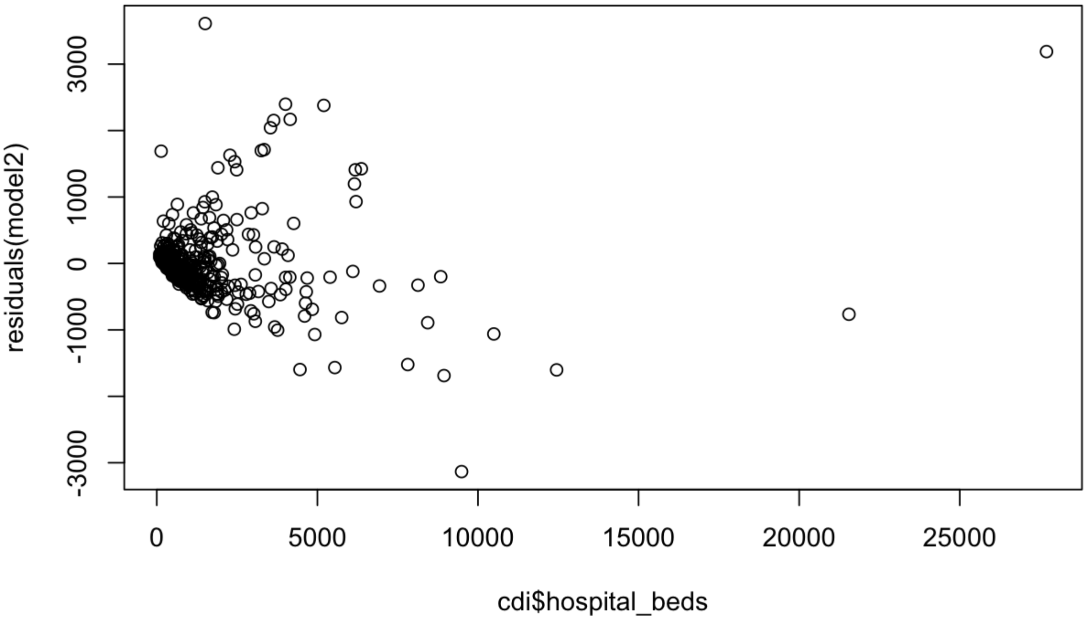
For each of the three fitted regression models, we can plot the residuals against X in a scatter plot and then a normal probability plot to determine if the assumptions surrounding the regression models are true or not. This will allow us to determine whether the linear regression model (2.1) is a more appropriate model in each case.

These are the 3 scatterplots with each model's residuals on the Y axis and the respective variable on the X axis.

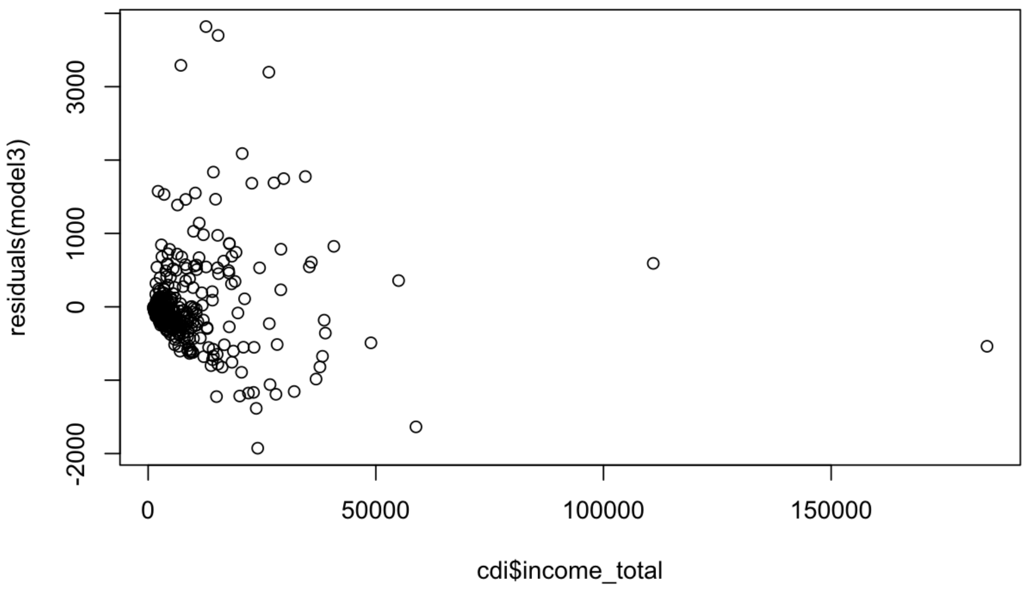
Model 1 Residuals vs Population Total:



Model 2 Residuals vs Total Number of Hospital Beds:



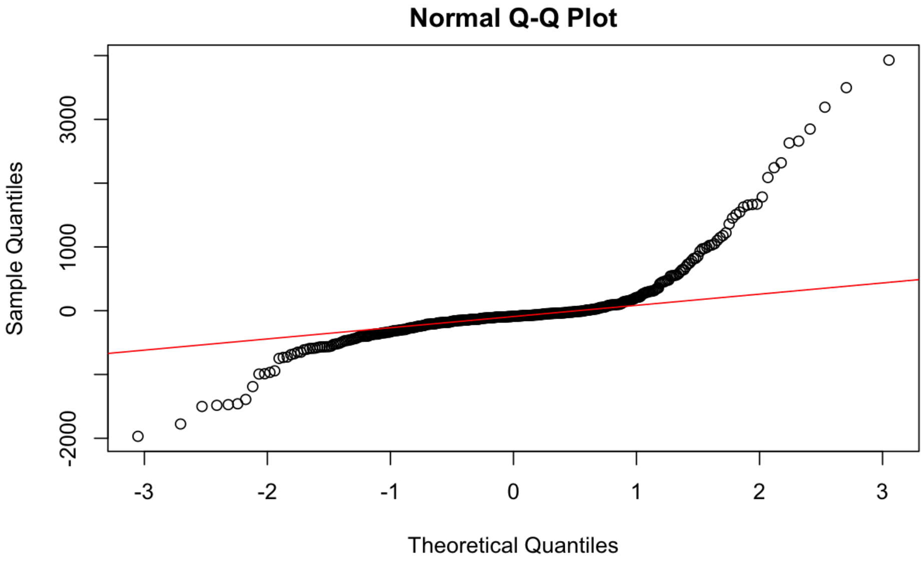
Model 3 Residuals vs Total Income:



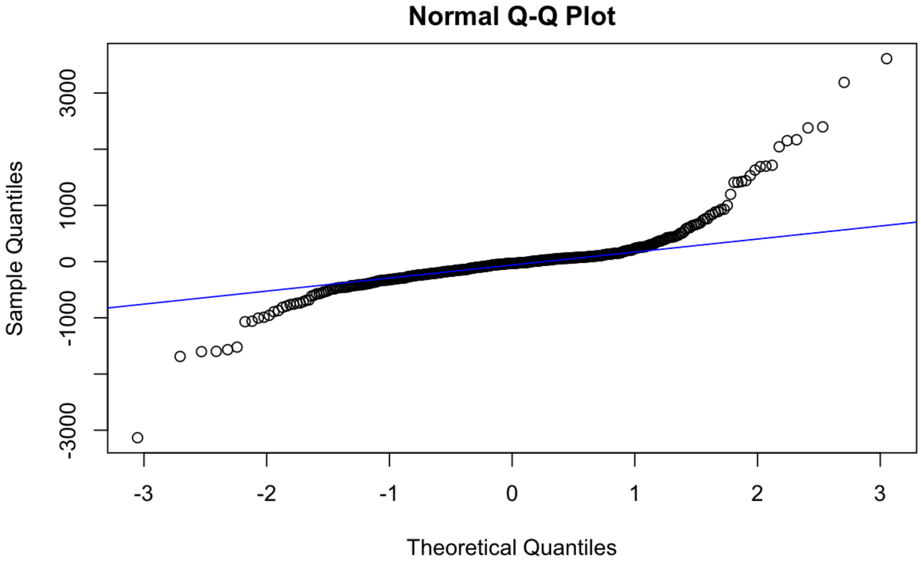
Upon reviewing these plots, we can detect from all 3 that there is the clear aspect of randomness. This confirms the assumption that the residuals are all independent and do not effect each other. With this information, we will now plot Normal Probability Plots to check for the normality assumption of regression errors. We will be plotting each residual against its Expected value and this should form a line with a constant gradient.

Normal Probability Plots:

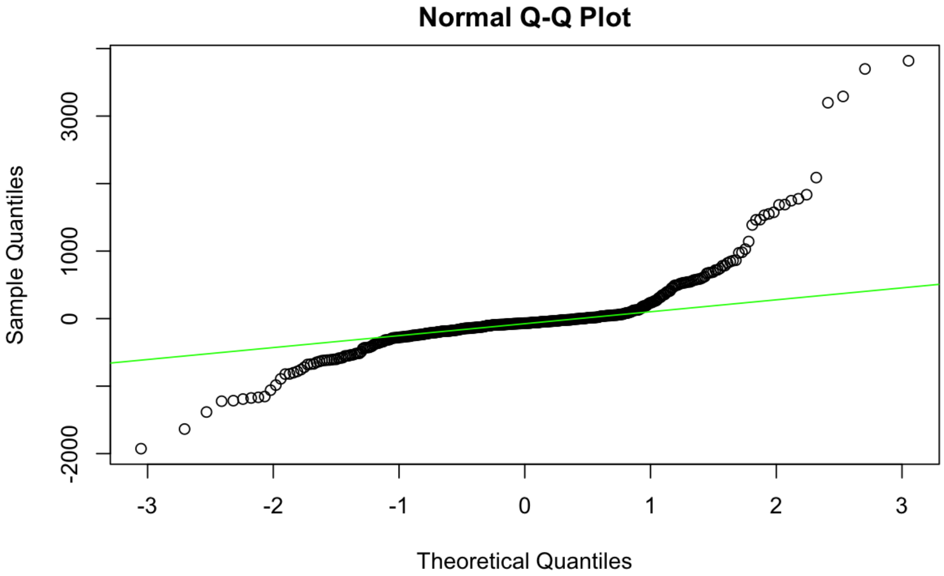
Residuals 1 vs E$((\varepsilon(hat))$



Residuals 2 vs E($(\varepsilon(hat))$)



Residuals 3 vs E($(\varepsilon(hat))$)



In these plots, we see that each residual normal probability plot does not actually follow the straight line. The main issue that can be deducted from all 3 plots is the face that there are far outliers that are completely off the expected line. This creates problems as the residuals may therefore not follow a normal distribution meaning that the linear regression model (2.1) could be a better model to implement.

Part 5: Discussion

In review of the data, the linear regression model (1.1) on face value looked to be good fits for the relationships between the number of active physicians and the 3 predictor variables. However, upon further inspection through regression diagnostics on residuals, it can be seen that linear regression model (2.1) would be a better fit for 3 relationships. We came to this conclusion because despite the residuals showing randomness, did not seem to follow a Normal Distribution in the Normal Q-Q plot.

Each predictor variable used for the 3 relationships showed that they were not independent to the Y variable. As well as in each region, per capita income was significantly joint and dependent on the percentage of those with at least a bachelor’s degree. It is possible that this observational data demonstrates the correlation between these in which certain governmental decisions to push for more people to get bachelor’s degrees or invest more money into hospital beds in a push to get more active physicians.

To improve the linear regression models, a possible way to improve them would be to increase the degrees of freedom. In general, this would be through increasing the sample size and in turn mean there is more power to reject a false null hypothesis and find a significant result.

Appendix:

Here is the code that was used to run the models and analysis.

Used to fit linear regression models in part 1:

```
model1<- lm(active_physicians ~ pop_total, data = cdi)
model2<- lm(active_physicians ~ hospital_beds, data = cdi)
model3<- lm(active_physicians ~ income_total, data = cdi)
```

Used to plot linear regression models:

```
plot(cdi$pop_total, cdi$active_physicians)
abline(model1) # this adds a regression line
plot(cdi$hospital_beds, cdi$active_physicians)
abline(model2)
plot(cdi$income_total, cdi$active_physicians)
abline(model3)
```

To obtain theresiduals as a collective of individual data we used the residuals() function

```
sum(residuals(model1)^2) / (n-2)
sum(residuals(model2)^2) / (n-2)
sum(residuals(model3)^2) / (n-2)
#This allowed us to find the MSE value for each model
```

We used the summary() function to obtain the R^2 values, information on residuals, the F-statistics for each of the 3 models.

```
summary(model1)
summary(model2)
summary(model3)
```

Code used to model the regression models split into the specific 4 regions:

```
model_region1 <- lm(income_percap ~ pct_bachelors,
  data = cdi[cdi$region == 1,]
model_region2 <- lm(income_percap ~ pct_bachelors,
  data = cdi[cdi$region == 2,])
model_region3 <- lm(income_percap ~ pct_bachelors,
  data = cdi[cdi$region == 3,])
model_region4 <- lm(income_percap ~ pct_bachelors,
  data = cdi[cdi$region == 4,])
```

Code used to create 4 confidence intervals at a 90% level:

```
confint(model_region1, "pct_bachelors", level = 0.9)
confint(model_region2, "pct_bachelors", level = 0.9)
confint(model_region3, "pct_bachelors", level = 0.9)
confint(model_region3, "pct_bachelors", level = 0.9)
```

The anova() function was used to create an ANOVA table that includes F-test results:

```
anova(model_region1)
anova(model_region2)
anova(model_region3)
anova(model_region4)
```

Code used to plot the residual scatter graph:

```
plot(cdi$pop_total, residuals(model1))
plot(cdi$hospital_beds, residuals(model2))
plot(cdi$income_total, residuals(model3))
```

Code used to plot the Normal Probability Plots:

```
qqnorm(residuals(model1))
qqnorm(residuals(model2))
qqnorm(residuals(model3))
```