# Case Study 1 Report

Practicing Physicians by County

**Group Member**

Bingheng Li (bl34)

Chengbo Li (chengbo4)

Jiyang Xu (jiyangx3)

## Introduction

The relationship between public health and wealth and poverty is always a prevailing controversial topic even before the prevalence of Covid-19. Some argue that wealth inequality is a vital public health issue. This is a long-standing social dilemma waiting to be solved. Meanwhile, the gap between wealth and poverty can lead to inequalities in other areas like education and security levels. In this case study, we proposed a regression model to predict the number of practicing physicians based on the data set including the selected county demographic information in the most populous counties from 1990 to 1992. Through the rich diversity of variables provided to us, we are able to analyze how some of the predictors contribute to distinct medical resources of each county, namely active physicians.

## Methodology

Our group started the model selection process by cleaning up the CDI dataset, then we did the model selection to drop the insignificant predictors and the highly correlated predictors. In the end, we performed diagnostics and Box-Cox transformation on our model.

## 1 Data Preprocessing

After observing the dataset, we found three discrete predictors, which are ID, county, and state. While multiple linear regression only naturally works with continuous variables and transforming the discrete variables into continuous variables is time-consuming, we decided to drop these predictors.

## 2 Model Selection

The correlation matrix is inspected to seek any high-correlated predictors. We identify correlations that are higher than 0.9 as high correlations. Observing the correlation matrix (Appendix 2.1), we found three predictors that are highly correlated to each other. Specifically, ‘hospital.beds’ and ‘total.population’ have a correlation of 0.923738360; ‘total.personal.income’ and ‘total.population’ have a correlation of 0.986747626; ‘hospital.beds’ and ‘total.personal.income’ have a correlation of 0.902061545. Since collinearity among predictors affects our model, we plan to combine the number of hospital beds and the total population together. More specifically, we will create a new column called “hospital.beds.avg”, which will be derived by dividing the number of hospital beds by the total population for each row. We will then remove the ‘hospital.beds’ column and the ‘total.population’ column. Running the correlation matrix (Appendix 2.2) again, we can see that all predictors have a correlation that is less than 0.9 with each other.

After cleaning up the dataset, we fitted a multiple linear regression model with all the predictors left (Appendix 2.3). We found that the multiple determination R^2 is approximately 93%,

representing this model can explain around 93% variation in the response variable. However, a high R^2 does not always mean the regression model is a useful one.

Starting the model selection process, we decided to perform a hypothesis test for all predictors. The null hypothesis is that the model with all predictors is a more adequate one while the alternative hypothesis is that the model with only the intercept is better. Based on the summary table (Appendix 2.3), the F statistic is equal to 445.3 and it follows an F distribution with p=12 (excluding the intercept) and degrees of freedom of 427. The corresponding p-value is equal to 2.2e-16, which is smaller than $\alpha = 0.05$. This implies that we reject the null and conclude that the model with all the predictors is more adequate compared to the intercept-only model. Next, we plan to conduct single predictor tests on each of the predictors that have a high p-value that is greater than 0.05. According to our summary table for our regression model(Appendix 2.3), we found the following predictors have a p-value that is smaller than 0.05: 'bachelor%', 'total.personal.income', 'hospital.beds.avg', which means we would test for rest of the predictors.

We planned to use the "backward" approach, where we started with the full model and removed predictors one at a time. We started with the predictor with the largest p-value. For all of our single predictor tests, the null hypothesis will be that the predictor is not significant and can be removed from the regression model ($\beta_j = 0$), and the alternative hypothesis will be that the predictor can not be removed from the model. Using the ANOVA test, all predictors except for 'below.poverty.level' have a p-value that is larger than 0.05, which means that we failed to reject their null hypotheses and had to drop them from the model. For the predictor 'below.poverty.level', based on the ANOVA table, the p-value is 8.903e-06, which is smaller than the 0.05 significance level. Therefore, we reject the null hypothesis and we conclude that we shall not drop the predictor 'below.poverty.level%' from our full model.

Now, we have successfully configured our MLR model with the following predictors: 'bachelor%', 'below.poverty.level%', 'total.personal.income', and 'hospital.beds.avg'.

# 3 Diagnostic

In this part, we will conduct diagnostics on our reduced model. First, we detected the unusual observations: (good/bad) high leverage points, outliers, and high influential points. Second, we checked the model assumptions: the constancy of variance, normality, non-linearity, and collinearity. In the end, we also used box-cox transformation to remedy departures from the normality assumption and reduce non-linearity.

## 3.1 Detect Unusual Observation

### 3.1.1 High-Leverage Point

In order to find high-leverage points, we first checked the half-normal plot and concluded that we can clearly see some high-leverage points (Appendix 3.1.1.1). In addition, we used R code to find

the observations with leverages more than 2p/n (twice the mean leverage), and we observed that there were 30 high leverage points, representing about 6.8% of the observations. Next, we determined the good leverage points and bad leverage points by finding the lower quartile (Q1) and the upper quartile (Q3). We filter the data frame of high-leverage points and extract the ones outside the interquartile range, and there were 13 “bad” high-leverage points (Appendix 3.1.1.2).

### 3.1.2 Outlier

We used the Bonferroni test to test the outliers and we studentized residuals, which follow a T distribution with a degree of freedom of (n - p - 1) where n is the sample size and p is the number of predictors (n = 440, p = 5). After we got the Bonferroni critical t value (-3.895092), we sorted the residuals in descending order and found the studentized residuals exceed the Bonferroni critical value. We concluded that there are four outliers in the data set: 50, 11, 67, and 48 (Appendix 3.1.2.1).

### 3.1.3 Highly Influential Point

We checked the high influential points by using Cook’s distance and we found that there were no influential points in the dataset since there was no point’s Cook’s distance larger than 1 (Appendix 3.1.3.1). Still, we plotted Cook’s distance calculated for every observation (Appendix 3.1.3.2) and got the same result.

## 3.2 Checking Model Assumptions

### 3.2.1 Checking for Constant Variance

We started checking the constant variance by using a residuals plot and specifically the residuals against fitted values (Appendix 3.2.1.1), and we observed that the points on the plot are not randomly scattered around the zero line, so we conclude that the constant variance assumption is not satisfied. Also, we conducted the Breusch-Pagan test with the null hypothesis: the variance is constant, and the alternative hypothesis: the variance is not constant. The p-value is 3.413e-09 for the studentized Breusch-Pagan test. Therefore, we reject the null hypothesis and conclude that the constant variance assumption is not satisfied.

### 3.2.2 Checking for Normality

We checked the normality assumption by using two plots: the QQ plot and the histogram of the residuals. It seems that the points in the QQ-plot do not fall on a straight line (Appendix 3.2.2.1). Also, we plotted the histogram of the residuals, and it departures the normality assumption. Since our data set has 440 observations, we use the Kolmogorov-Smirnov Test with the null hypothesis: the distribution is normal, and the alternative hypothesis: the distribution is not normal. The p-value is 2.2e-16. Therefore, we reject the null hypothesis and conclude that the normality assumption is not satisfied.

### 3.2.3 Checking for Linearity

We checked the structure of the relationship between the predictors and response by checking them one by one after dropping the response (‘active.physicians’). Specifically, we check the linearity with respect to 'bachelor%', 'below.poverty.level%', 'total.personal.income', and 'hospital.beds.avg'.

Observing the four plots, the points appeared to be randomly scattered around the fitted regression line and the blue line is not horizontal, (Appendix 3.2.3.1). Therefore, the linearity assumption is satisfied here.

#### 3.2.4 Checking for Collinearity

We checked collinearity in our data by investigating the new correlation matrix of the reduced model and we concluded that there were no high-correlated predictors (>= 0.9). (Appendix 3.2.4.1)

### 3.3 Box-Cox Transformation

Since we failed the normality assumption for our model, we decided to perform the Box-Cox Transformation to remedy the departures from the normality assumption and reduce non-linearity. The basic idea behind Box-Cox is to find λ for which Y^λ follows a Normal distribution. In order to find λ, we plotted the log-likelihood function vs. λ graph (Appendix 3.3.1). Based on the output, we say that a value of lambda near 0.1553885 would probably fix the departure from the normality assumption. Here, we decided to round up λ to 0, which means that we will take the logarithm of the response variable Y. After performing operations on R, we have successfully transformed our model. We decided to run the Kolmogorov-Smirnov Normality Test on our model again. The null hypothesis indicates a normal distribution while the alternative hypothesis indicates a non-normal distribution. Running the test again gives us a p-value of 3.819e-08, which is still smaller than the 0.05 significance level; therefore, we reject the null hypothesis of normality and conclude that the normality assumption is not satisfied, unfortunately.

## 4 Result

### 4.1 Diagnostic for Transformed Model

Since we performed the Box-Cox transformation and altered our model, we needed to perform the diagnostics that we performed on our reduced model again on the transformed model. As we utilized the same methods that we used before to detect unusual observations and check model assumptions, we have only included the results for the diagnostics. We concluded that there were 30 high-leverage points, representing 6.8% of the observations, and 13 “bad” high-leverage points. Moreover, there was only one outlier in the transformed model in the data set: data point 1. In addition, we observed no high influential points. Testing for the constant variance assumption. we delivered a p-value of 2.2e-16, which is smaller than the significance level of 0.05, so we reject the null and conclude that the constant variance assumption is not satisfied in the transformed model. Lastly, we found that the predictor ‘total.personal.income’ appeared to be non-linear, while the other three predictors still have a linear relationship with the response.

## Conclusion

The relationship between public health and the economic level is controversial yet critical to address whether in the past, present, or future. Medical resources are vital to everyone, and every person, no matter wealthy or poor, should be provided with equal rights to access any medical resources and enjoy medical care. Based on our model created for predicting active physicians in this case study, we recognized the percentage of the adult population with bachelor's degrees, the percentage of 1990 CDI population with income below the poverty level, the average number of beds, cribs, and bassinets in the hospital during 1990 (we use the total number of hospital beds divide by total population in each county to get this data), and total personal income as meaningful predictors for the response. This model indicates that wealth has an underlying relationship with the number of practicing physicians in each county, as the predictors we selected for the model are all somehow involved with capital. Moreover, according to our model, we can clearly see that the effect of the average number of hospital beds on the prediction of active physicians in counties is not as significant as the other three variables, namely the percentage of the adult population with bachelor's degrees, the percentage of the population with income below the poverty level and the total personal income. Therefore, a clear correlation between capital and medical resources is suggested.

We cannot determine whether this phenomenon is a beneficial motivator or a deterioration factor for society, and we simply hope that no matter young or old, rich or poor, we human beings can be provided with approximately equal medical resources.

# Appendix:

## 2.1

| | land.area | total.population | population%(18-24) | population%(65+) | hospital.beds | crimes | high.school% |
|---|---|---|---|---|---|---|---|
| land.area | 1.000000000 | 0.173083353 | -0.05487812 | 0.005770871 | 0.073047270 | 0.12947537 | -0.09859811 |
| total.population | 0.173083353 | 1.000000000 | 0.07837212 | -0.029037393 | 0.923738360 | 0.88633185 | -0.01742690 |
| population%(18-24) | -0.054878125 | 0.078372117 | 1.00000000 | -0.616309639 | 0.074531907 | 0.08994063 | 0.25058429 |
| population%(65+) | 0.005770871 | -0.029037393 | -0.61630964 | 1.000000000 | 0.053278417 | -0.03529032 | -0.26825176 |
| hospital.beds | 0.073047270 | 0.923738360 | 0.07453191 | 0.053278417 | 1.000000000 | 0.85684988 | -0.11191638 |
| crimes | 0.129475371 | 0.886331846 | 0.08994063 | -0.035290324 | 0.856849883 | 1.00000000 | -0.10632840 |
| high.school% | -0.098598111 | -0.017426900 | 0.25058429 | -0.268251758 | -0.111916382 | -0.10632840 | 1.00000000 |
| bachelor% | -0.137237736 | 0.146813850 | 0.45609703 | -0.339228765 | 0.100426534 | 0.07707652 | 0.70778672 |
| below.poverty.level% | 0.171343348 | 0.038019509 | 0.03397551 | 0.006578474 | 0.172793840 | 0.16440566 | -0.69175048 |
| unemployment% | 0.199209277 | 0.005351703 | -0.27852706 | 0.236309411 | 0.007523992 | 0.04355675 | -0.59359579 |
| percent.capita.Income | -0.187715132 | 0.235610188 | -0.03164843 | 0.018590706 | 0.194808180 | 0.11753914 | 0.52299613 |
| total.personal.income | 0.127074261 | 0.986747626 | 0.07116151 | -0.022733151 | 0.902061545 | 0.84309805 | 0.04335573 |
| geographic.region | 0.362868243 | 0.069437072 | 0.05241407 | -0.173291567 | -0.003106920 | 0.09130732 | -0.01005506 |

| | bachelor% | below.poverty.level% | unemployment% | percent.capita.Income | total.personal.income | geographic.region |
|---|---|---|---|---|---|---|
| land.area | -0.13723774 | 0.171343348 | 0.199209277 | -0.18771513 | 0.12707426 | 0.36286824 |
| total.population | 0.14681385 | 0.038019509 | 0.005351703 | 0.23561019 | 0.98674763 | 0.06943707 |
| population%(18-24) | 0.45609703 | 0.033975512 | -0.278527058 | -0.03164843 | 0.07116151 | 0.05241407 |
| population%(65+) | -0.33922877 | 0.006578474 | 0.236309411 | 0.01859071 | -0.02273315 | -0.17329157 |
| hospital.beds | 0.10042653 | 0.172793840 | 0.007523992 | 0.19480818 | 0.90206155 | -0.00310692 |
| crimes | 0.07707652 | 0.164405659 | 0.043556752 | 0.11753914 | 0.84309805 | 0.09130732 |
| high.school% | 0.70778672 | -0.691750483 | -0.593595788 | 0.52299613 | 0.04335573 | -0.01005506 |
| bachelor% | 1.00000000 | -0.408423848 | -0.540906913 | 0.69536186 | 0.22223013 | 0.02029897 |
| below.poverty.level% | -0.40842385 | 1.000000000 | 0.436947236 | -0.60172504 | -0.03873934 | 0.27098485 |
| unemployment% | -0.54090691 | 0.436947236 | 1.000000000 | -0.32214439 | -0.03387633 | -0.05437857 |
| percent.capita.Income | 0.69536186 | -0.601725039 | -0.322144395 | 1.00000000 | 0.34768161 | -0.22249375 |
| total.personal.income | 0.22223013 | -0.038739339 | -0.033876330 | 0.34768161 | 1.00000000 | 0.03768546 |
| geographic.region | 0.02029897 | 0.270984846 | -0.054378572 | -0.22249375 | 0.03768546 | 1.00000000 |

## 2.2

| | land.area | population%(18-24) | population%(65+) | crimes | high.school% | bachelor% | below.poverty.level% |
|---|---|---|---|---|---|---|---|
| land.area | 1.000000000 | -0.05487812 | 0.005770871 | 0.12947537 | -0.09859811 | -0.13723774 | 0.171343348 |
| population%(18-24) | -0.054878125 | 1.00000000 | -0.616309639 | 0.08994063 | 0.25058429 | 0.45609703 | 0.033975512 |
| population%(65+) | 0.005770871 | -0.61630964 | 1.000000000 | -0.03529032 | -0.26825176 | -0.33922877 | 0.006578474 |
| crimes | 0.129475371 | 0.08994063 | -0.035290324 | 1.00000000 | -0.10632840 | 0.07707652 | 0.164405659 |
| high.school% | -0.098598111 | 0.25058429 | -0.268251758 | -0.10632840 | 1.00000000 | 0.70778672 | -0.691750483 |
| bachelor% | -0.137237736 | 0.45609703 | -0.339228765 | 0.07707652 | 0.70778672 | 1.00000000 | -0.408423848 |
| below.poverty.level% | 0.171343348 | 0.03397551 | 0.006578474 | 0.16440566 | -0.69175048 | -0.40842385 | 1.000000000 |
| unemployment% | 0.199209277 | -0.27852706 | 0.236309411 | 0.04355675 | -0.59359579 | -0.54090691 | 0.436947236 |
| percent.capita.Income | -0.187715132 | -0.03164843 | 0.018590706 | 0.11753914 | 0.52299613 | 0.69536186 | -0.601725039 |
| total.personal.income | 0.127074261 | 0.07116151 | -0.022733151 | 0.84309805 | 0.04335573 | 0.22223013 | -0.038739339 |
| geographic.region | 0.362868243 | 0.05241407 | -0.173291567 | 0.09130732 | -0.01005506 | 0.02029897 | 0.270984846 |
| hospital.beds.avg | -0.141233520 | 0.02952439 | 0.247147869 | 0.07789072 | -0.21116247 | -0.04541826 | 0.371398926 |

| | unemployment% | percent.capita.Income | total.personal.income | geographic.region | hospital.beds.avg |
|---|---|---|---|---|---|
| land.area | 0.19920928 | -0.18771513 | 0.127074261 | 0.36286824 | -0.141233520 |
| population%(18-24) | -0.27852706 | -0.03164843 | 0.071161515 | 0.05241407 | 0.029524392 |
| population%(65+) | 0.23630941 | 0.01859071 | -0.022733151 | -0.17329157 | 0.247147869 |
| crimes | 0.04355675 | 0.11753914 | 0.843098049 | 0.09130732 | 0.077890722 |
| high.school% | -0.59359579 | 0.52299613 | 0.043355729 | -0.01005506 | -0.211162472 |
| bachelor% | -0.54090691 | 0.69536186 | 0.222230125 | 0.02029897 | -0.045418264 |
| below.poverty.level% | 0.43694724 | -0.60172504 | -0.038739339 | 0.27098485 | 0.371398926 |
| unemployment% | 1.00000000 | -0.32214439 | -0.033876330 | -0.05437857 | -0.062487824 |
| percent.capita.Income | -0.32214439 | 1.00000000 | 0.347681610 | -0.22249375 | -0.053550037 |
| total.personal.income | -0.03387633 | 0.34768161 | 1.000000000 | 0.03768546 | 0.006323904 |
| geographic.region | -0.05437857 | -0.22249375 | 0.037685456 | 1.00000000 | -0.113622302 |
| hospital.beds.avg | -0.06248782 | -0.05355004 | 0.006323904 | -0.11362230 | 1.000000000 |

## 2.3

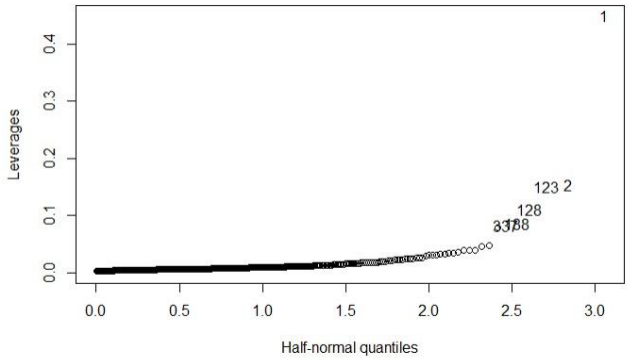
```
Call:
lm(formula = active.physicians ~ ., data = CDI)

Residuals:
    Min      1Q  Median      3Q     Max
-1507.8  -228.9   -10.5   171.9  2889.9

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -6.282e+02  6.597e+02  -0.952   0.3415
land.area            -2.172e-02  1.555e-02  -1.396   0.1634
`population%(18-24)`  1.078e+01  8.133e+00   1.326   0.1856
`population%(65+)`    4.939e+00  7.490e+00   0.659   0.5100
crimes                2.433e-04  7.416e-04   0.328   0.7430
`high.school%`       -8.412e+00  6.125e+00  -1.373   0.1703
`bachelor%`           1.734e+01  6.960e+00   2.491   0.0131 *
`below.poverty.level%` 1.658e+01  9.604e+00   1.726   0.0850 .
`unemployment%`      -2.902e+00  1.288e+01  -0.225   0.8218
percent.capita.Income -5.906e-03  1.179e-02  -0.501   0.6167
total.personal.income  1.295e-01  3.532e-03  36.670   <2e-16 ***
geographic.region     -3.862e+00  2.501e+01  -0.154   0.8773
hospital.beds.avg      1.440e+05  1.390e+04  10.360   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 445.3 on 427 degrees of freedom
Multiple R-squared:  0.9398,    Adjusted R-squared:  0.9381
F-statistic: 555.4 on 12 and 427 DF,  p-value: < 2.2e-16
```

### 3.1.1.1




### 3.1.1.2

| | active.physicians <int> | bachelor% <dbl> | below.poverty.level% <dbl> | total.personal.income <int> | hospital.beds.avg <dbl> |
|---|---|---|---|---|---|
| 1 | 23677 | 22.3 | 11.6 | 184230 | 0.003125295 |
| 2 | 15153 | 22.8 | 11.1 | 110928 | 0.004221296 |
| 3 | 7553 | 25.4 | 12.5 | 55003 | 0.004417360 |
| 4 | 5905 | 25.3 | 8.1 | 48931 | 0.002473563 |
| 5 | 6062 | 27.8 | 5.2 | 58818 | 0.002642129 |
| 6 | 4861 | 16.6 | 19.5 | 38658 | 0.003886704 |
| 8 | 3823 | 13.7 | 16.9 | 36872 | 0.004494037 |
| 48 | 4635 | 49.9 | 2.7 | 22772 | 0.001990682 |
| 67 | 5674 | 27.7 | 14.4 | 15369 | 0.009269385 |
| 70 | 3368 | 31.6 | 15.4 | 14808 | 0.008871240 |

1–10 of 13 rows

Previous 1 2 Next

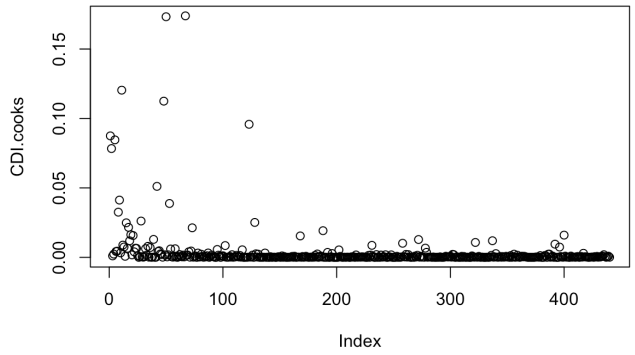
### 3.1.2.1

```
      50       11       67       48
7.099874 6.354710 6.233780 3.915906
```

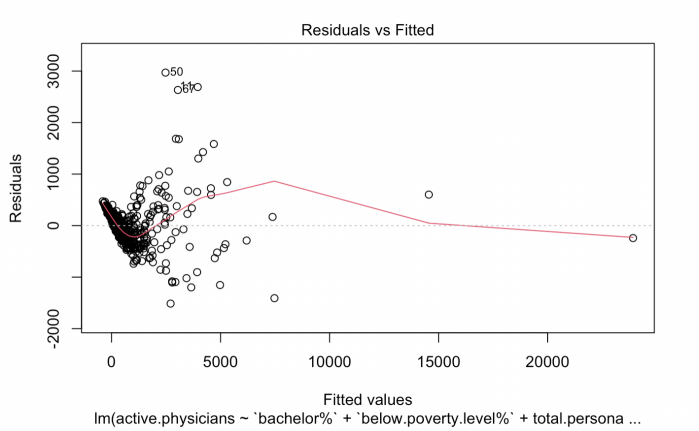
### 3.1.3.1

```
       67         50         11         48        123          1          5          2         42          9
53          8         28        128         15
0.17387954 0.17324210 0.12037705 0.11246926 0.09581769 0.08747078 0.08459801 0.07832728 0.05112855 0.04123943
0.03878746 0.03254138 0.02615969 0.02513721 0.02486516
```

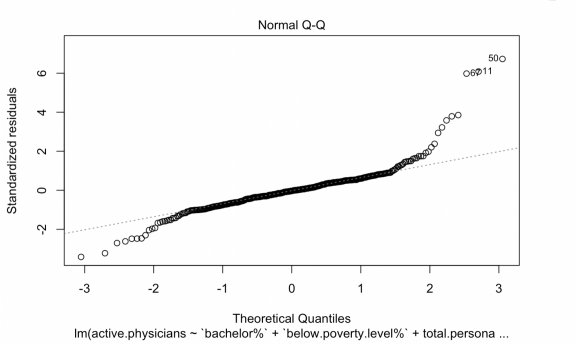
### 3.1.3.2



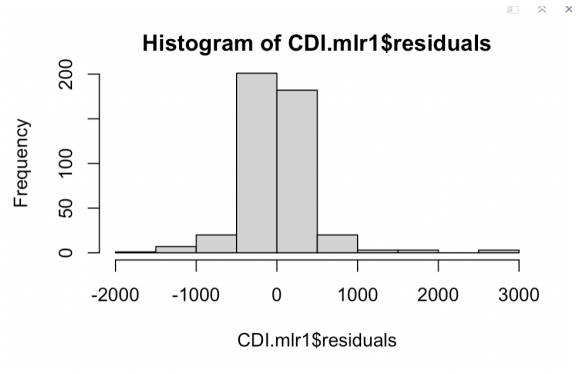

### 3.2.1.1



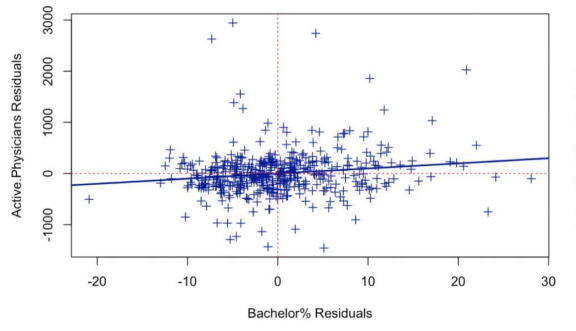

### 3.2.2.1

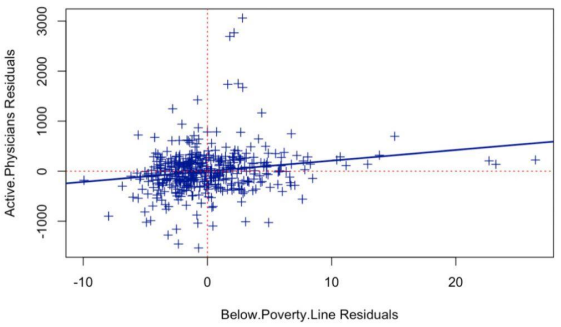


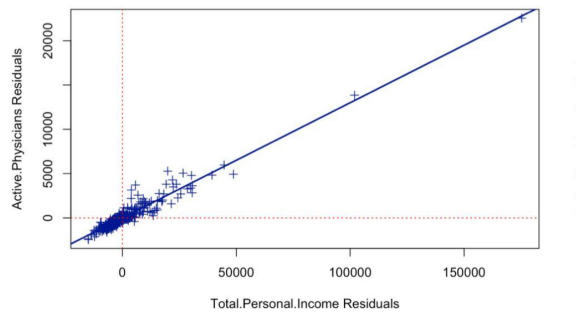

### 3.2.2.2

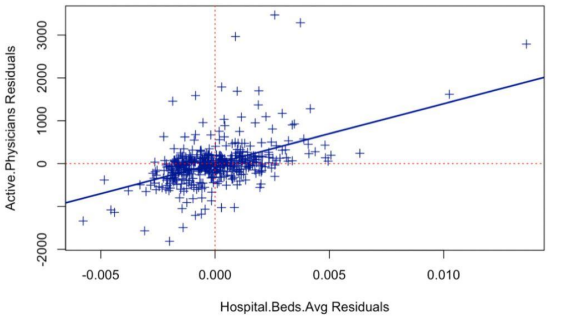



### 3.2.3.1



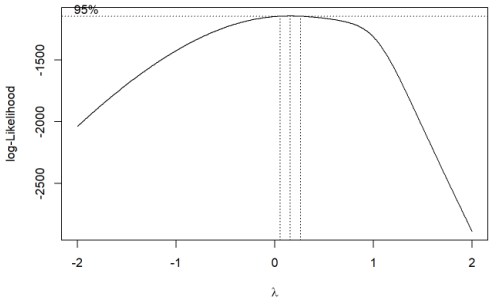




---

### 3.2.4.1

| | bachelor% | below.poverty.level% | total.personal.income | hospital.beds.avg |
|---|---|---|---|---|
| bachelor% | 1.00 | -0.41 | 0.22 | -0.05 |
| below.poverty.level% | -0.41 | 1.00 | -0.04 | 0.37 |
| total.personal.income | 0.22 | -0.04 | 1.00 | 0.01 |
| hospital.beds.avg | -0.05 | 0.37 | 0.01 | 1.00 |

### 3.3.1




## 4.2

```
Call:
lm(formula = active.physicians ~ `bachelor%` + `below.poverty.level%` +
    hospital.beds.avg, data = CDI_red)

Residuals:
    Min      1Q  Median      3Q     Max
-2447.5  -646.7  -307.0    59.5 22542.7

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            -1411.07     360.97  -3.909 0.000107 ***
`bachelor%`               68.98      11.67   5.910 6.89e-09 ***
`below.poverty.level%`    47.90      20.64   2.321 0.020766 *
hospital.beds.avg      144488.26   43886.62   3.292 0.001075 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1695 on 436 degrees of freedom
Multiple R-squared:  0.1093,    Adjusted R-squared:  0.1031
F-statistic: 17.83 on 3 and 436 DF,  p-value: 6.256e-11
```