Modeling Association between hospital capacity and community vulnerabilities

Phase I Report

Willliam O. Agyapong Prince Appiah Eti Nyamekeh Baffoe

University of Texas, El Paso (UTEP), Department of Mathematical Sciences

Abstract

In this report, we analysed the relationship between hospital capacity and community vulnerabilities using a simple linear regression model. Overall, it was discovered that the predictor variable Low Income Area LIA County SAIPE Poverty Percentage and the two response variables inpatient_beds_7_day_avg and total_icu_beds_7_day_avg explained to some extent how hospital capacity is associated with community vulnerabilities as compared to the other predictor variables. Hospital capacity was also found to be influenced, to some degree, by whether a hospital is located in a hardest hit area.

1 Introduction

This is the first of a study that seeks to understand the association between community vulnerabilities and hospital beds capacity.

1.1 Background/rationale

Undoubtedly, the COVID-19 pandemic has affected almost every facet of life worldwide. However, areas such as hospital care units and industries have been tremendously affected. The USA was not left out, many hospitals were filled out leading to a serious shortage of hospital beds, especially for intensive care units (ICUs). As the pandemic rises day by day, hospitals have been overburdened or occupied with victims of the pandemic. In such situations, hospitals in vulnerable communities may be more prone to exceeding hospital beds capacity. For instance, Tsai et al (2022) studied the association of community level social vulnerability with US acute care hospital intensive care unit capacity during this period of Covid-19 pandemic and found that 63% of hospitals reached critical ICU capacity for at least two weeks during the study period, while the surge of COVID-19 cases appeared to be crowding out non-COVID-19-related intensive care needs, showing how the association between social vulnerability and critical ICU capacity highlights underlying structural inequities in health care access. Again, according to report by the Office of Inspector General of the U.S. Department of Health and Human Services, hospitals reported that the covid-19 pandemic has significantly strained health care delivery.

Therefore, there is the need to investigate or understand the kind of relationship that exists between hospital beds capacity and community vulnerabilities. This has the potential of providing great insights to decision makers to take action to prevent strained ICU capacity from compounding COVID-19 inequities. In this report, we conduct a simple linear regression analysis to investigate whether hospital beds capacity is associated with community vulnerabilities using data obtained from the U.S. Department of Health and Human Services Protect database.

1.2 Objectives

The main objective of this report is to answer the following research question.

How is hospital capacity associated with community vulnerabilities?

For each of the outcome variables presented in Table 1, our goal is to identify the appropriate community vulnerability measure that is able to explain much of the variability to help us achieve

the main objective of the study. This leads us to the specific objectives described below:

Is there a linear relationship between a particular hospital capacity measure and any of the community vulnerability measures? If so, which of the community vulnerability measure provides the largest reduction in the variability of the given outcome (hospital capacity measure)?

Our initial hypothesis is that the vulnerability measure, low income area county poverty percentage, is likely to have the most significant effect on the dependent variables measuring hospital capacity.

1.3 Setting

Data for this report span the period from July 15, 2020 to `January 7, 2022` and were collected from selected hospitals in the US as described in the participants section. Part of the population were first recruited on June 1, 2020, with the remaining joining on July 15 the same year.

1.4 Participants

Our study participants consists of 444 hospital facilities spread across various counties in the State of Texas selected from a hospital population that includes all hospitals registered with Centers for Medicare & Medicaid Services (CMS) as of June 1, 2020 and non-CMS hospitals that have reported since July 15, 2020. It does not include psychiatric, rehabilitation, Indian Health Service (IHS) facilities, U.S. Department of Veterans Affairs (VA) facilities, Defense Health Agency (DHA) facilities, and religious non-medical facilities. Our study, however, focused on a subsection comprising hospitals in the state of Texas.

1.5 Variables

The study focuses on seven variables obtained from a facility-level hospitalization data as well as a community-level vulnerability data. The following table provides information about these seven variables used in the study. As indicated by the role column, there are three dependent (response) variables which are measures of hospital capacity and constitute the seven-day averages of the reports provided for a given facility for that element during that collection week, while

the remaining four variables, measuring community vulnerability, are the independent (predictor) variables.

Table 1: Variables of interest

Variable Name	Description	Type of Measure	Data Type	Role
inpatient_beds_7_day_avg	Average number of total number of staffed inpatient beds in your hospital including all overflow, observation, and active surge/expansion beds used for inpatients (including all ICU beds) reported in the 7-day period.	Hospital bed capacity	Numeric/continuous	Outcome/ Response
inpatient_beds_used_covid_7_day_avg	Average of reported patients currently hospitalized in an inpatient bed who have suspected or confirmed COVID-19 reported during the 7-day period.	Hospital bed capacity	Numeric/continuous	Outcome/ Response
total_icu_beds_7_day_avg	Average number of total number of staffed inpatient ICU beds reported in the 7-day period.	Hospital bed capacity	Numeric/continuous	Outcome/ Response
is_metro_micro	This is based on whether the facility serves a Metropolitan or Micropolitan area. True if yes, and false if no.	Community vulnerability	Binary/categorical	Predictor
HHA_Score	Hardest Hit Area Score	Community vulnerability	Integer/Categorical	Predictor
LIA_CS_PP	Low Income Area (LIA) County SAIPE - (Poverty Percentage)	Community vulnerability	Numeric/continuous	Predictor
Rural_Score	An indicator of whether the facility is at a rural location or not	Community vulnerability	Integer/Categorical	Predictor

1.6 Data sources/measurement

The variables listed in the previous section come from two sources; a facility-level hospitalization data and community vulnerability data, both of which were accessed from the US Department of

Health and Human Services Protect database for **COVID-19 Reported Patient Impact and Hospital Capacity by Facility** and **COVID-19 Community Vulnerability Crosswalk - Crosswalk by Census Tract**, respectively. Due to the large size of the data, only 5010 observations from the hospital capacity data were used. We also limited ourselves to only the averages derived based on the number of values collected for a given hospital in a collection week (Friday to Thursday). These observations were then merged with the community-level vulnerability data having **72836** data points. We eventually restricted the scope of the study to the State of Texas which left us with a final data set consisting of **82915** observations.

According to the data sources, FCC's scoring procedure was used to weigh the community vulnerability measures including Hardest Hit Area (HHA), Low Income Area, Tribal Community, and Rural Community. We chose the scored variables because they provide an evaluation of the most vulnerable communities in our population.

1.7 Bias

- One source of bias could come from the cases where `Low Income Area County SAIPE Peverty Percentage`, the only continuous predictor variable, emerged as the best independent or predictor variable in our initial modeling for narrowing down to one model per each dependent variable. This is because the overall relationship observed could be different when the sub-populations are considered. To address this potential bias, we considered running other analyses where one of the significant categorical variable was used to split the data into subgroups and refitted the models on the individual subgroups.
- We also believe the study data was not representative of the study population since we simply took the first 5010 observations from the hospital capacity dataset. This source of bias can be addressed by taking a good random sample from the large hospital capacity data from the original source, but this was clearly beyond our control.
- Another potential source of bias is the high level of class imbalance seen in the distribution of the categorical predictors. Some levels of all four community vulnerability measures,

`HHA_Score`, `Rural Score`, `Tribal Community` and `is_metro_micro`, are disproportionately represented. We defer the treatment of this bias to the second stage of the study.

1.8 Study size

There were $761,663$ observations in the original merged data provided for the analysis. Limiting the study to the State of Texas brought the number of observations down to $82,915$. We then removed missing data arising from data suppression that was applied to hospital capacity average measures less than four (4) by the maintainers of the data and non-reporting by some of the facilities. Please see the Missing Data section for what we considered to be non-reported values. In the end, the data used for the analysis had $45,991$ observations making up our study size.

1.9 Statistical methods

1.9.1 Regression model

The statistical method used here is simple linear regression. Under this method, we have the general model $Y_i = \beta_1 X_i + \beta_0 + \epsilon_i$, $i = 1, \cdots, n$, where n is the sample size and ϵ_i is normally distributed with mean zero and variance δ^2. Under this model, we check the following assumptions:

1. Constancy of the error variance. This was checked by using Breusch Pagan test
2. Normality of the error variance. This was checked using the normal probability plot and the Normal QQ plot.
3. Linearity of the model. This was checked using the scatter plot and the residual versus fitted plot
4. Normality of the residuals using coefficient of correlation between the ordered residuals and their expected values under normality.
5. Check the significance of the slope and intercept using confidence intervals of the slope and the intercept.

1.9.2 Model Assessement

We diagnosed the appropriateness of our models using diagnostic plots such residual versus fitted plots, normality plots as well as numerical tests including `Bruesch-Pagan` test for non-constancy of error variance, and the `coefficient of correlation` test for normality. This was done largely to ensure that the various assumptions required for the simple linear regression model list above were reasonably satisfied. We therefore relied on the these diagnostics to to select our best models.

Other performance metrics such the coefficient of determination (R^2) and the mean squared error (MSE) were also utilized.

- The R-squared (coefficient of determination) is used to determine the amount of variability in the dependent/response that is accounted for by the regression model. The strength and weakness of the fitted When R-square is between 0 to 0.5 we say there is a weak fitted linear model, when it is between 0.5 to 0.75, we say that the regression model is moderate and when it is in-between 0.75 to 1, we say the model is strong. Because the r-square is not enough evidence to determine whether a model is well fitted, we did not base our analysis sorely on the R-squaered.

2 Results

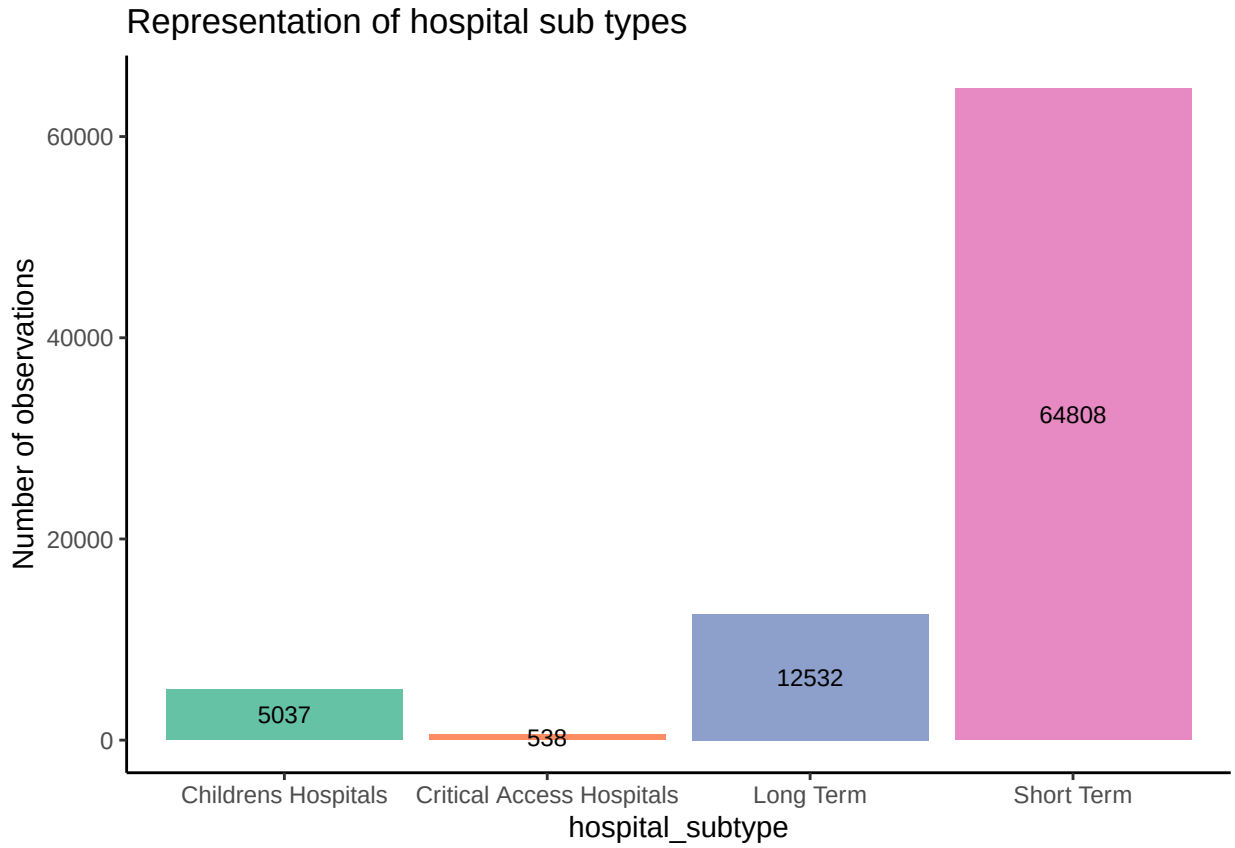
In this section, we present relevant results from our exploratary data analysis and regression modeling procedures.

2.1 Descriptive data

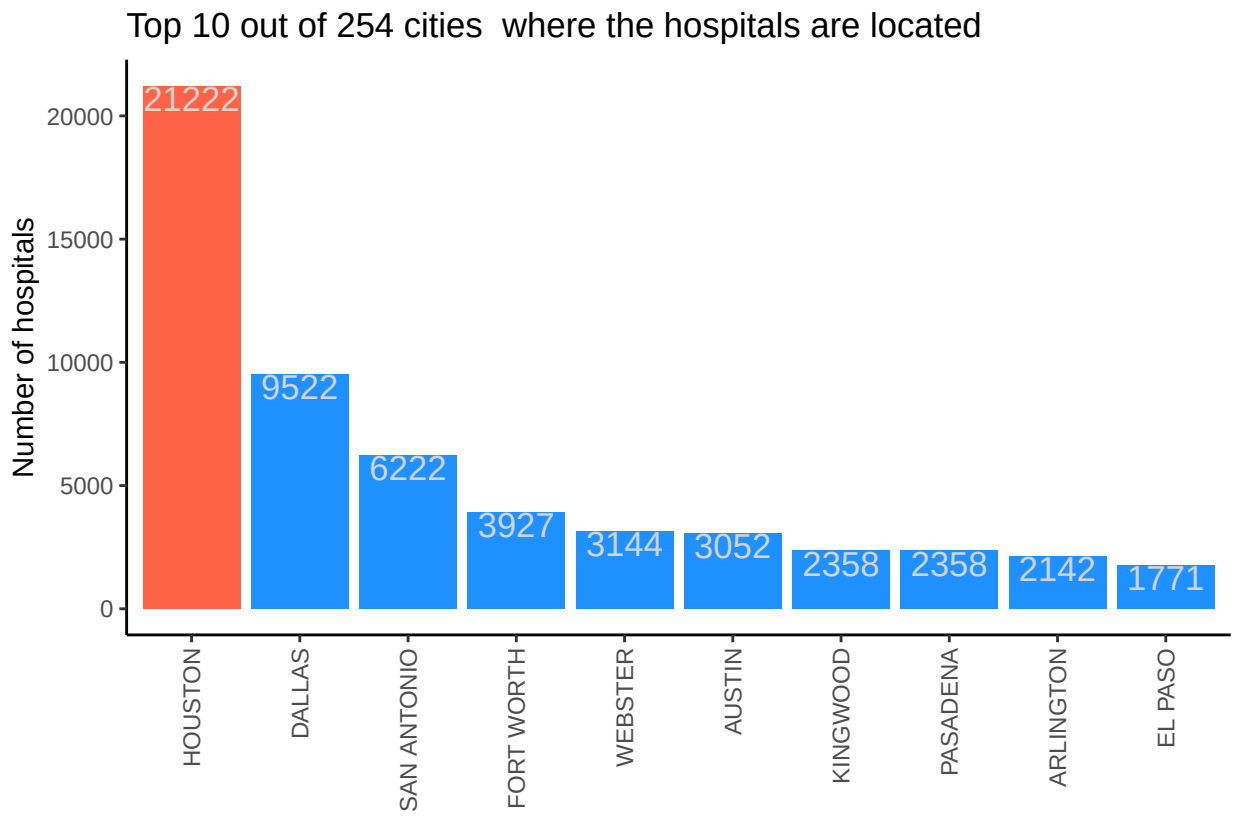
We provide both numerical summaries and graphs to enhance our understanding of the underlying data for the study.

2.1.1 Characteristics of study participants

As already identified in the introduction, our study participants consist of all hospitals in Texas registered with Centers for Medicare & Medicaid Services (CMS) as of June 1, 2020 and non-CMS hospitals that have reported since July 15, 2020. The graphs below provides information about how these participants are distributed in terms of hospital subtypes and cities.



Here, we also see unequal representation with critical access hospitals being low as expected.



Most of the participating hospitals come from Houston, followed by Dallas with the city of El Paso coming last on the list. This may not be a fair representation since the populations at these cities differ greatly from each other. Hence, the population of these cities needs to be taken into account when interpreting the figures.

2.1.2 Investigating missing data

Table 2: Missing values in the merged data set

variable	n_miss	pct_miss
inpatient_beds_used_covid_7_day_avg	9169	11.0583
total_icu_beds_7_day_avg	4332	5.2246
inpatient_beds_7_day_avg	222	0.2677
is_metro_micro	0	0.0000
HHA_Score	0	0.0000
LIA_CS_PP	0	0.0000
Tribal_Community	0	0.0000
Rural_Score	0	0.0000

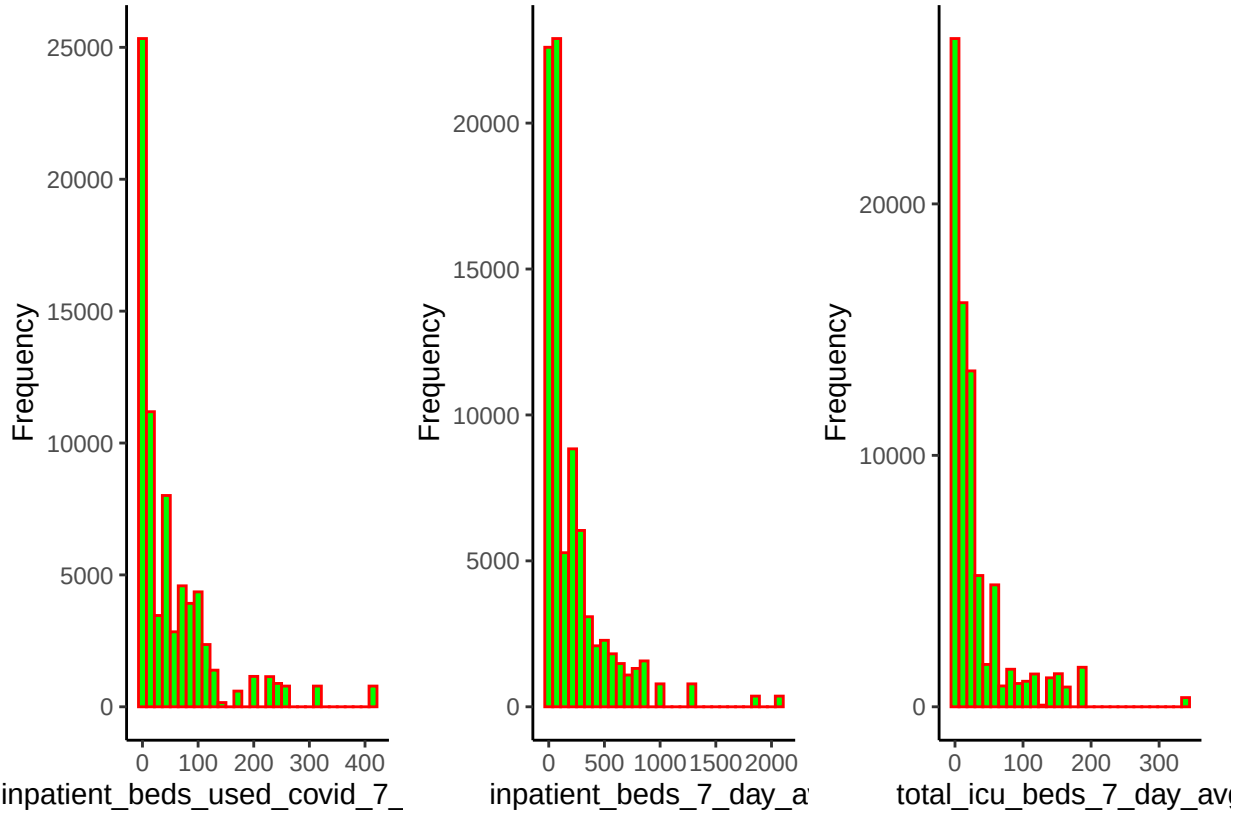
Table 3: Missing values in dependent variables from the covid data before merge

variable	n_miss	pct_miss
inpatient_beds_used_covid_7_day_avg	1249	24.9351
total_icu_beds_7_day_avg	258	5.1507
inpatient_beds_7_day_avg	17	0.3394

We see from **Table 2** that only the three independent variables have missing values. It turns out that most of the missing values were created by the data merge between the hospital capacity data and the community vulnerability data as revealed by **Table 3**. We take a very simplistic approach of **deleting the missing values** as a means of treatment since most of the missing values are artificial. Again, it is important to note that the actual missing values denote averages that were less than 4.

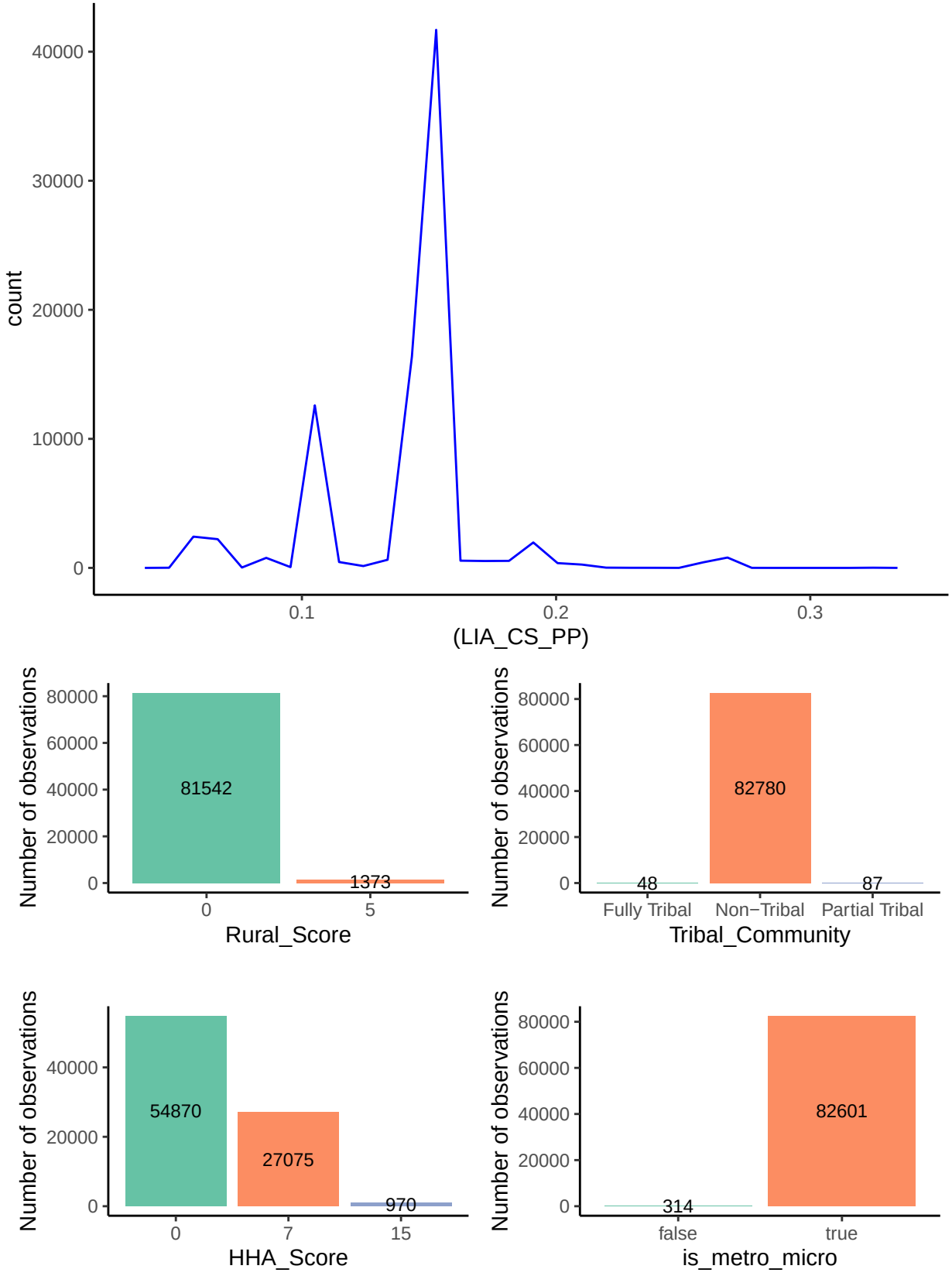
2.1.2.1 Non-reported values in dependent variables We also observed that the hospital capacity measures chosen for our dependent variables had zeros (0) in them. This was quite surprising at first because we did not expect to see zeros in these variables when the maintainers of the data ***suppressed all averages less than four (4) and replaced them with -999,999*** which were then marked as missing values in our version of the data. After digging deeper we realized that these zeros could represent non-reported values by some of the facilities since our data source stated that "No statistical analysis is applied to impute non-response". By this reasoning and the fact that there was no information to determine the reasons leading to non-responses, we decided to represent zeros (0) in our dependent variables as missing values and treated them in the same way as described above.

2.1.3 Distribution of dependent variables



The distributions of all three dependent variables are identical and heavily right skewed. There also appears to be outliers. The skewness suggests that one of two transformations, logarithm, and a power transformation (square root or cube root), might be appropriate. We learn from these distributions that similar results may be obtained from modeling the variables.

2.1.4 Distribution of independent variables

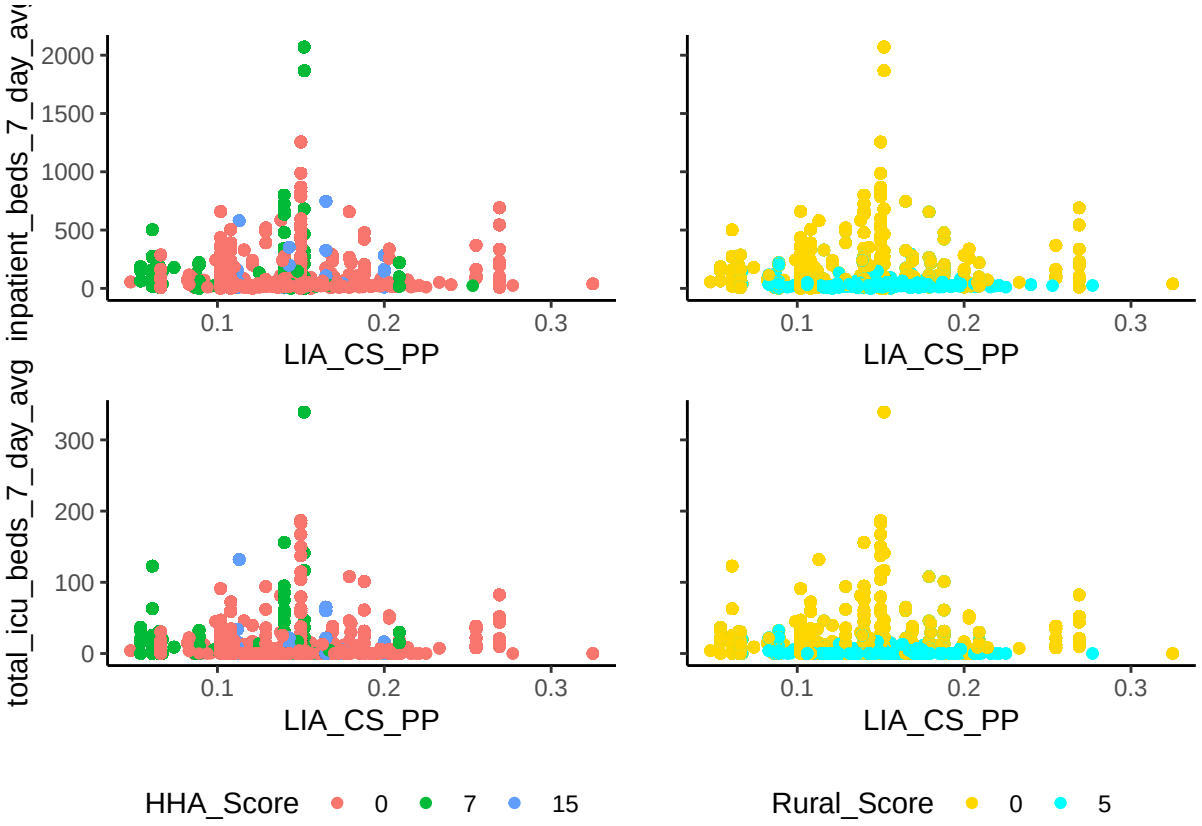


The distribution of the continuous independent/predictor variable, `Low Income Area County`

`SAIPE Poverty Percentage` (LIA_CS_PP), looks multimodal and right skewed with some possible outliers at the extreme ends. This measurement appears to have come from two underlying sub-populations as portrayed by the two high peaks.

Turning our focus to the categorical independent variables, we observe that some of the levels of each variable are extremely disproportionately represented. This is a clear sign of class imbalance which would have to be dealt with appropriately in the modeling phase in order to avoid any potential bias.

2.1.5 Effect of independent variables on the dependent variables



The above graphs depict the relationship between the only continuous independent variable, `Low Income Area County SAIPE Poverty Percentage` (LIA_CS_PP), and two dependent variables, `inpatient beds 7 day average` and `total ICU beds 7 day average`, split into subgroups defined by `Hardest Hit Area Score` and `rural score`. From these plots we see that `LIA_CS_PP` does not appear to have any interesting relationship with the two dependent variables, and the subgroups are also not well separated. Similar observations were made with the other variables so

we decided not present them here.

We also explored how each dependent variable is distributed across the levels of each one of the categorical independent variables which can be found in **Appendix C**. In general, the plots suggest that the categorical variables have some effect on all the dependent variables. However, some of them appear to have relatively stronger effect than others.

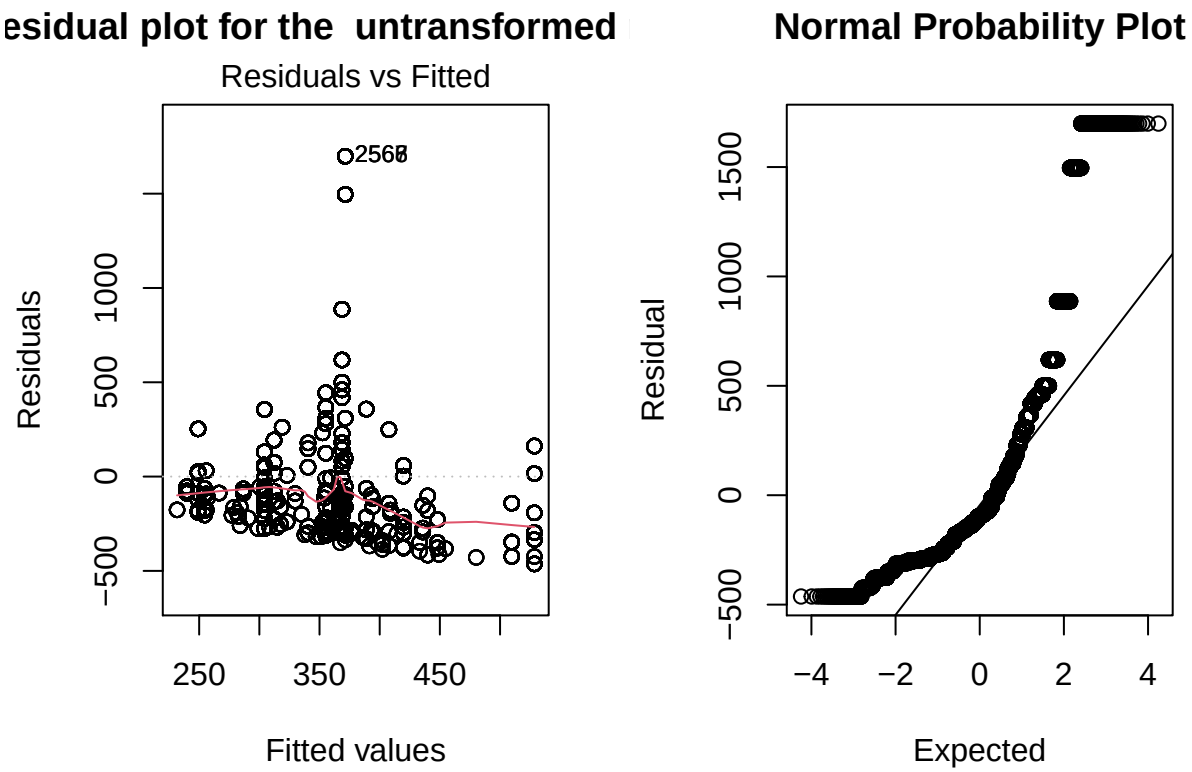
2.2 Modeling

We began our modeling by first running multiple simple linear regression models involving each dependent variable and all the independent variables as a means of obtaining a single best independent variable that could be used for further analysis. As revealed from the EDA, the distribution of all the dependent variables are heavily skewed to the right. This initial observation suggests that a transformation of some kind of the dependent variables would be necessary. Therefore, to give each independent variable a fair chance, we fitted four different models involving no transformation, log transformation, square and cube root transformations of the dependent variables in our initial attempt to selecting the "best" model without any rigorous analyses. At this stage, we simply compared models by looking at overall model performance in terms of the proportion of variability explained (R^2) , the mean squared error (MSE), and any evidence of a linear relationship from the p-values corresponding to the F-statistics. Tables of results showing how each independent variable performed can be found in **Appendix A**. Reported below is a table of each dependent variable with the associated independent variable selected.

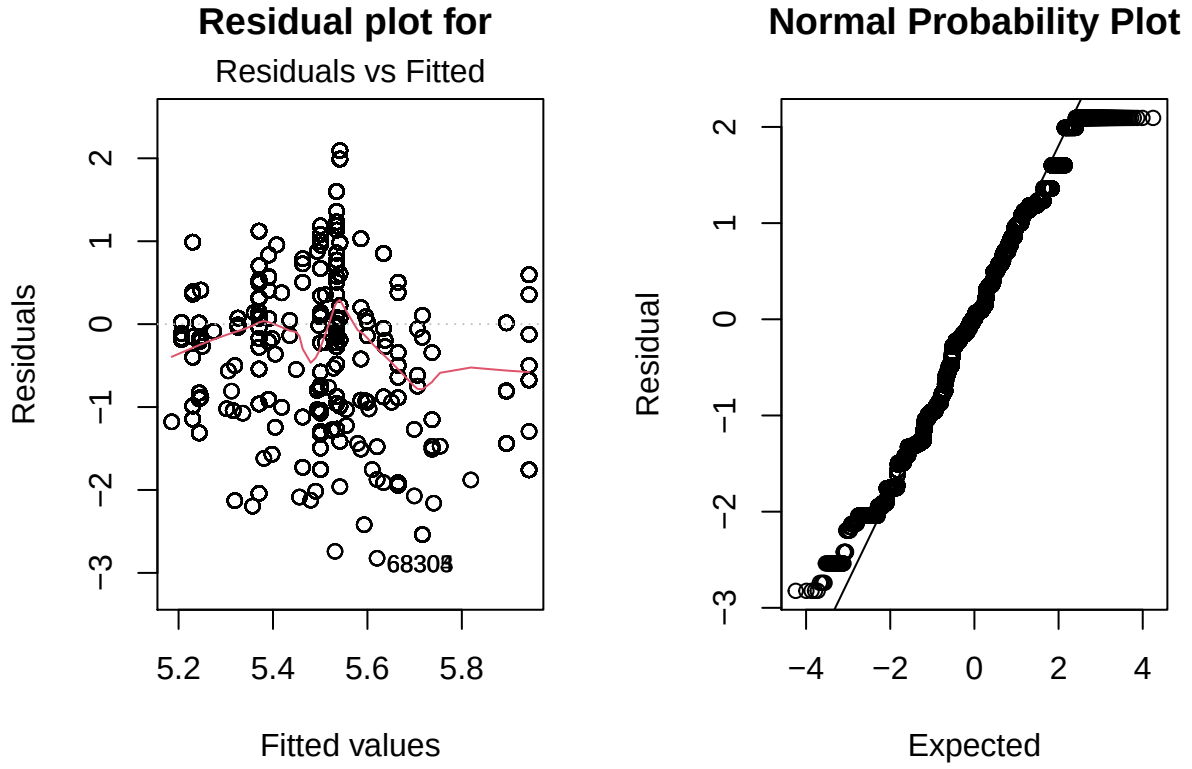
The next subsections present detailed analyses conducted on the initial best model for each of the three dependent variables.

2.2.1 Further analysis on the best model for `inpatient_beds_7_day_avg`

Diagnostic plots for model with untrasfromed variables



Diagnostic plots for model with log transformation of the dependent variable



The following two tables provide results for the Bruesche-Pagan test for nonconstant variance for the untransformed and the log transformed models, respectively.

Table 4: Breusch-Pagan test

	BP.statistic	p.value	chisq
BP	2407	0	3.841

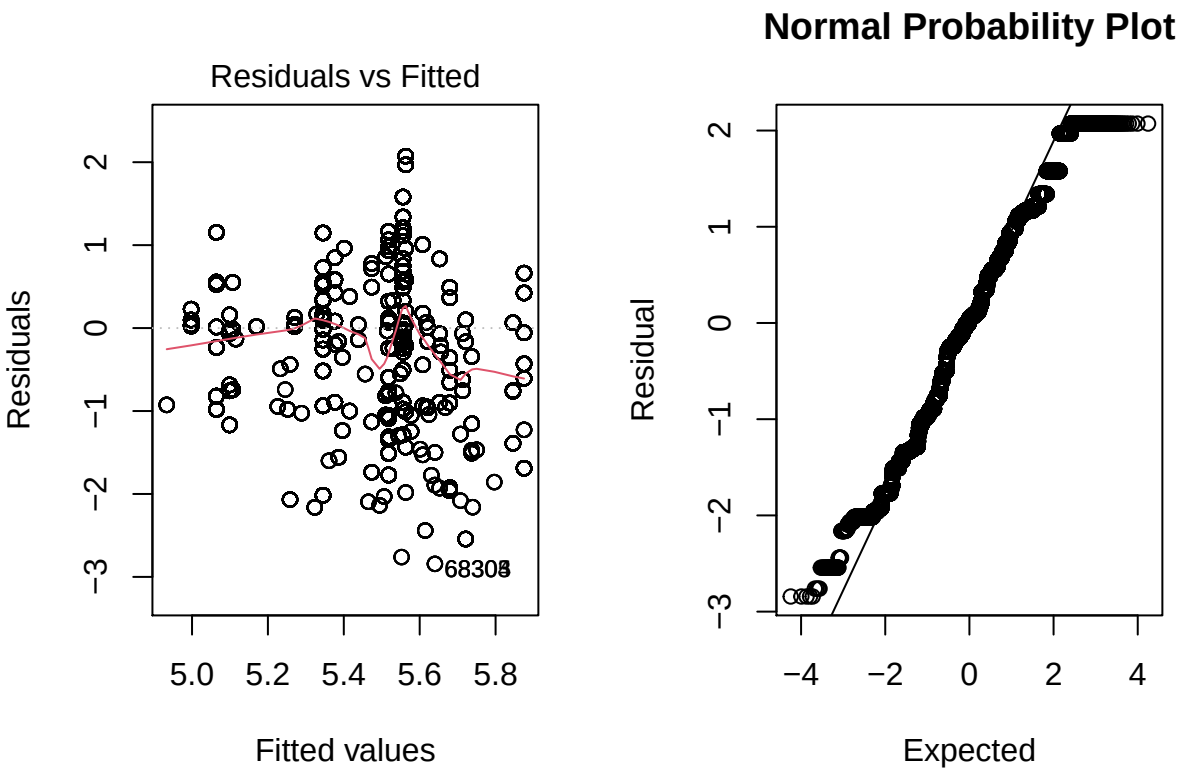
Table 5: Breusch-Pagan test

	BP.statistic	p.value	chisq
BP	468.6	0	3.841

We first checked the model on the untransformed data and realized that most of the underlying simple linear regression were violated. Also, the predictor variable LIA_CS_PP performed well with the untransformed response variable in terms of the residual plot, normal probability plot, scatter plot, lowess curve and regression confidence bands. Looking at the histogram of the response variable it was right skewed so we performed log transformation. We observe that the residual plot for the log transform does not show any pattern of nonlinearity and the normal QQ plot or the normal probability plot is better than the plots for the untransformed data, eventhough its not normal. The residual plot for the untansform has an "n" shape. We also see that the scatter plot for the log transform depicts linearity as compared with the untransform.

We obtained the coefficient of correlation between the ordered residuals and their expected values under normality- and we had the critical value *0.987* less than *0.995* (the correlation coefficient) for the log transform with significance level of 0.05. Thus, we conclude that at *5%* significance level, the residuals are normality distributed. The untransformed model failed the test for the coefficient of correlation between the ordered residuals and their expected values under normality.By Breusch-Pagan test the error variance is not constant. Hence, based on these result, we see that the log transform appeared to be better than the untransformed. In view of that, we proceeded to do a simultaneous log transform for both the predictor variable (LIA_CS_PP) and the response variable (inpatient_beds_7_day_avg)

Simultaneous log transform



After the simultaneous log transform, we observed that the $R^2 = 0.0268039$ has increased as compared to the $R^2 = 0.016626$ of the log transform of only the response variable. There has been an improvement in the residual plot. Even though the error variance is not constant, it is better than the previous one. There is a great improvement in normal probability plot as compared to only the log transform just on the response variable. The residual plot also do not depict non-linearity pattern. We obtained the coefficient of correlation between the ordered residuals and their expected values under normality - and we had the critical value *0.987* less than *0.995* (the correlation coefficient) for the log transform with significance level of 0.05. Thus, we conclude that at *5%* significance level, the residuals are normality distributed.

Reporting the final model:

Table 6: Parameter estimates

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	6.5925	0.0310	212.67	0	6.5317	6.6532
log(LIA_CS_PP)	0.5464	0.0154	35.59	0	0.5163	0.5765

Table 7: Overall Model Performance Results

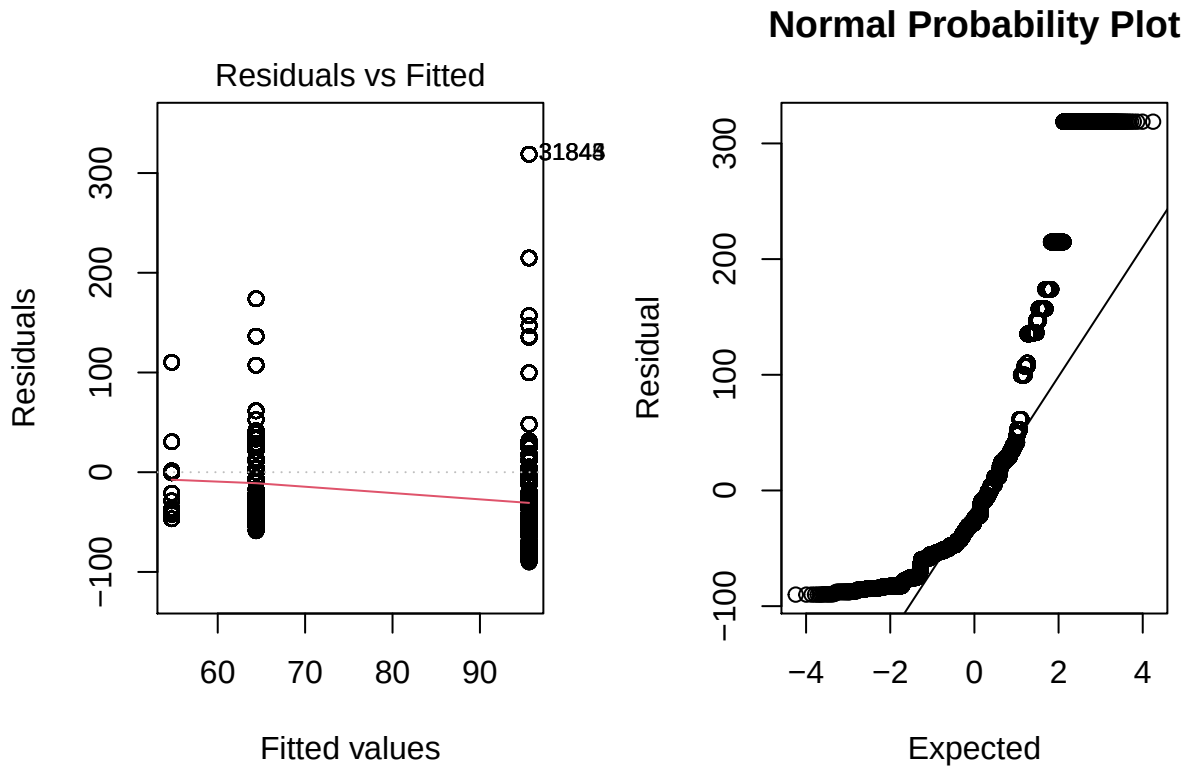
r.squared	adj.r.squared	MSE	F.statistic	p.value
0.0268	0.0268	0.76	1267	0

$$\widehat{log(Y)} = 6.5925 + 0.5464X$$

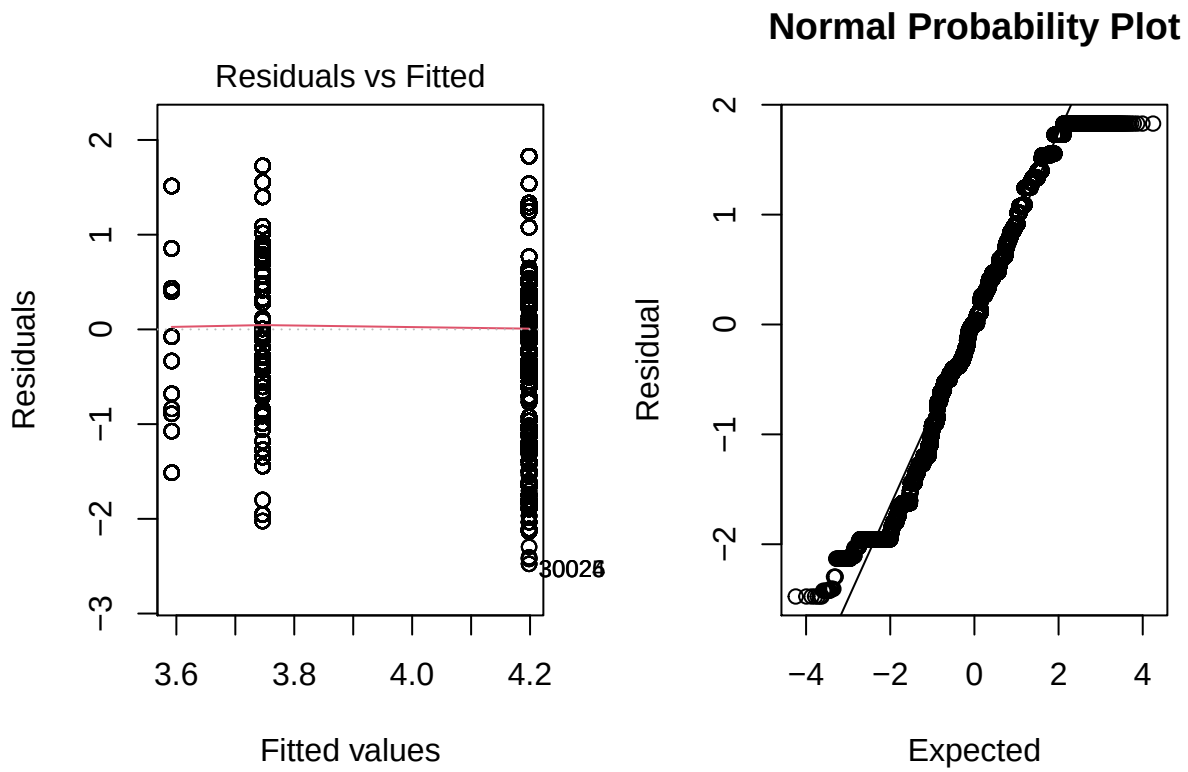
where Y is inpatient_beds_7_day_avg and X is `Low Income Area County SAIPE Poverty Percentage (LIA_CS_PP)` . The model signifies that for every unit change in the the log of `LIA_CS_PP`, the mean of inpatient_beds_7_day_avg increases by **0.5464** in log units.

2.2.2 Further analysis on our best model for the `inpatient_beds_used_covid_7_day_avg`

Diagnostic plots for model with untrasfromed variables



Diagnostic plots for model with log transform of the dependent variable



The following two tables provide results for the Bruesche-Pagan test for nonconstant variance for the untransformed and the log transformed models, respectively.

Table 8: Breusch-Pagan test

	BP.statistic	p.value	chisq
BP	3066	0	3.841

Table 9: Breusch-Pagan test

	BP.statistic	p.value	chisq
BP	317.3	0	3.841

The residual versus fitted plot suggests increasing error variance, which is also supported by the Bruesch-Pagan test as this numerical test indicated a violation of the constant variance assumption. On the other hand, the normality illustrates that the distribution of the error terms does not depart substantially form normality. Notwithstanding, we conclude with this as our best model since all other forms of transformations could not remedy the violation noticed.

Reporting results for final model:

Table 10: Parameter estimates

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	4.1983	0.0053	790.60	0	4.1879	4.2087
HHA_Score7	-0.4519	0.0089	-50.83	0	-0.4694	-0.4345
HHA_Score15	-0.6062	0.0394	-15.39	0	-0.6834	-0.5290

Table 11: Overall Model Performance Results

r.squared	adj.r.squared	MSE	F.statistic	p.value
0.0557	0.0556	0.8245	1356	0

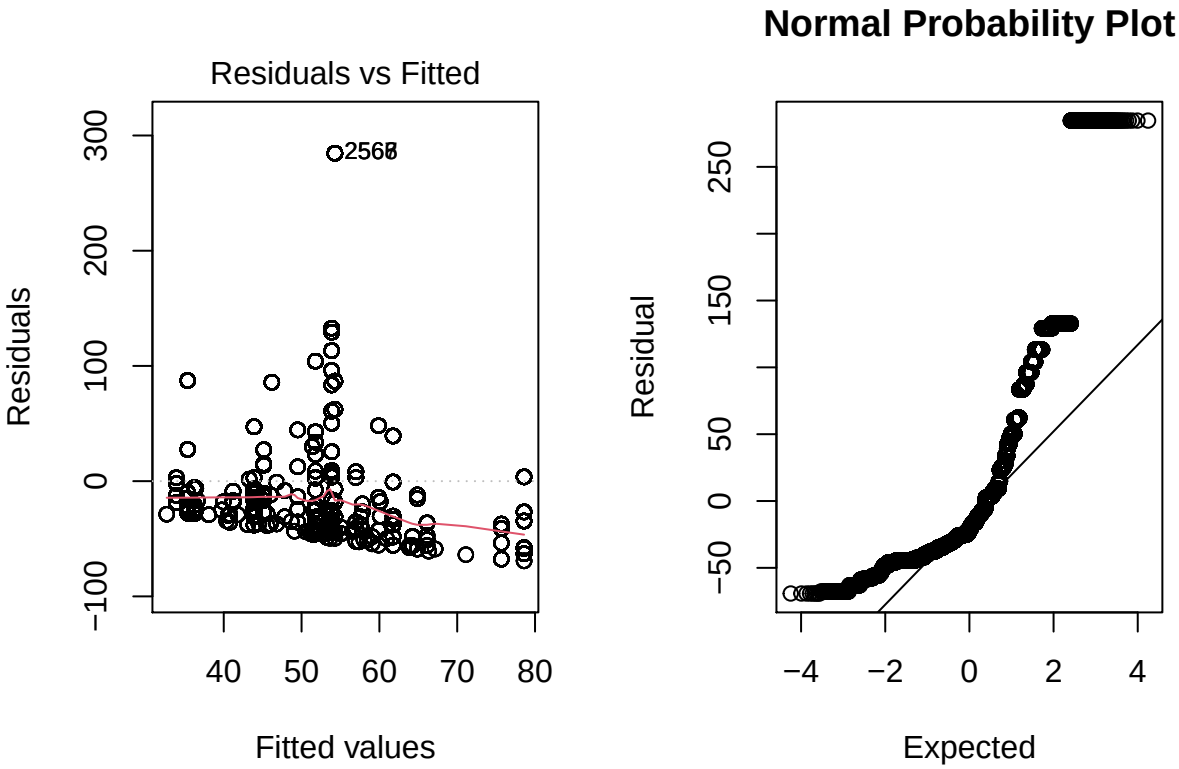
From the table of results, the estimated regression model is given as

$$\hat{\mathrm{Y}} = 4.2242 - 0.5546X_1 - 0.7447X_2$$

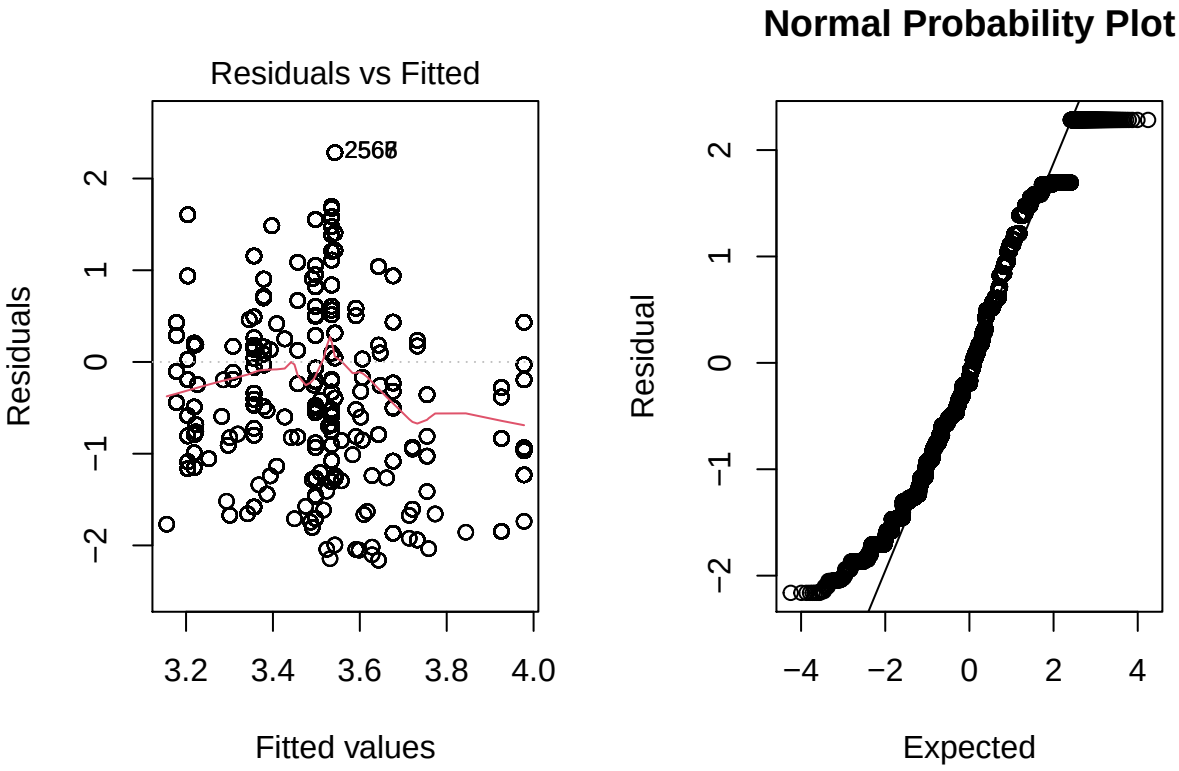
where Y is `inpatient_beds_used_covid_7_day_avg`, X_1 is 1 for level 7 and 0 otherwise, and X_2 is 1 for level 15 and 0 otherwise.

2.2.3 Further analysis on our best model for the `total_icu_beds_7_day_avg`

Diagnostic plots for model with untrasfromed variables



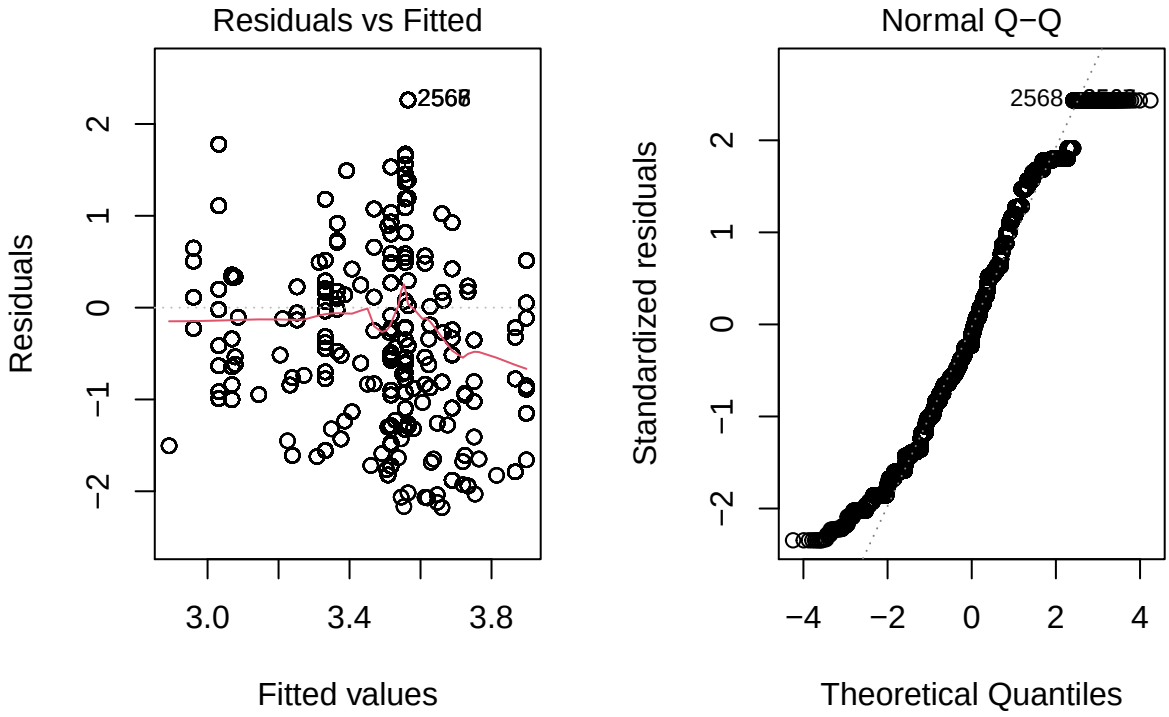
Diagnostic plots for model with log transform of the response variable



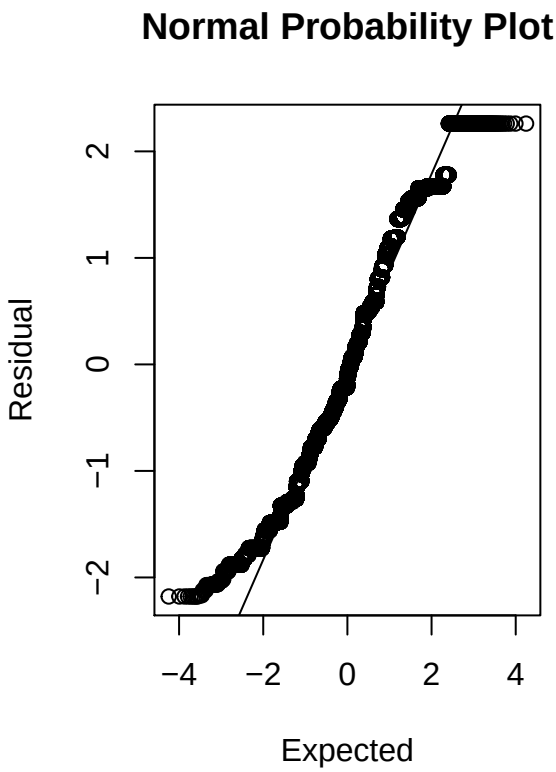
The results here is very similar to the ones obtained for the `inpatient_used_bed_7_day_avg` versus `LIA_CS_PP` models above. The same observations can be made concerning the normality

of the errors as well as the residual versus fitted plot. Also, the coefficient of correlation for normality gave the same evidence. Once again, we still have some violation of the constancy of the error variance assumption with evidence from the Breusch-Pagan test so we proceed to perform a simultaneous log transform.

Diagnostic plots for simultaneous log transformation




```
##
##  Pearson's product-moment correlation
##
## data:  corr[, 3] and corr[, 1]
## t = 1551, df = 45989, p-value <2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9904 0.9907
## sample estimates:
##    cor
## 0.9906
```



We observed after the simultaneous log transform that the $R^2 = 0.0269693$ has increased as compared to the $R^2 = 0.0172882$ of the log transform of only the response variable. There has been an improvement in the residual plot.Even though the error variance is not constant, it is better than the previous one. There is a great improvement in normal probability plots as compared to only the log transform of only the response variable. The scatter plot also do not depict nonlinearity pattern. There seems to be linearity. We obtained the coefficient of correlation between the ordered residuals and their expected values under normality- and we had the critical value *0.987* less than *0.991* (the correlation coefficient) for the log transform with significance level of 0.05. Thus, we conclude that at *5%* significance level, the residuals are normality distributed.

We conclude with the following with the model involving the simultaneous log transformation as our best model for `total_icu_beds_7_day_avg` as reported below.

Final Model

Table 12: Parameter estimates

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	4.6652	0.0330	141.2	0	4.6005	4.7300
log(LIA_CS_PP)	0.5841	0.0164	35.7	0	0.5521	0.6162

Table 13: Overall Model Performance Results

r.squared	adj.r.squared	MSE	F.statistic	p.value
0.027	0.0269	0.8632	1275	0

$$\widehat{log(Y)} = 4.6652 + 0.5841X$$

where Y is `total_icu_beds_7_day_avg` and X is `Low Income Area County SAIPE Poverty Percentage (LIA_CS_PP)` . The model signifies that for every unit change in the the log of `LIA_CS_PP`, the mean of `total_icu_beds_7_day_avg` increases by **0.5841** in log units.

2.3 Other Analysis (Interactions)

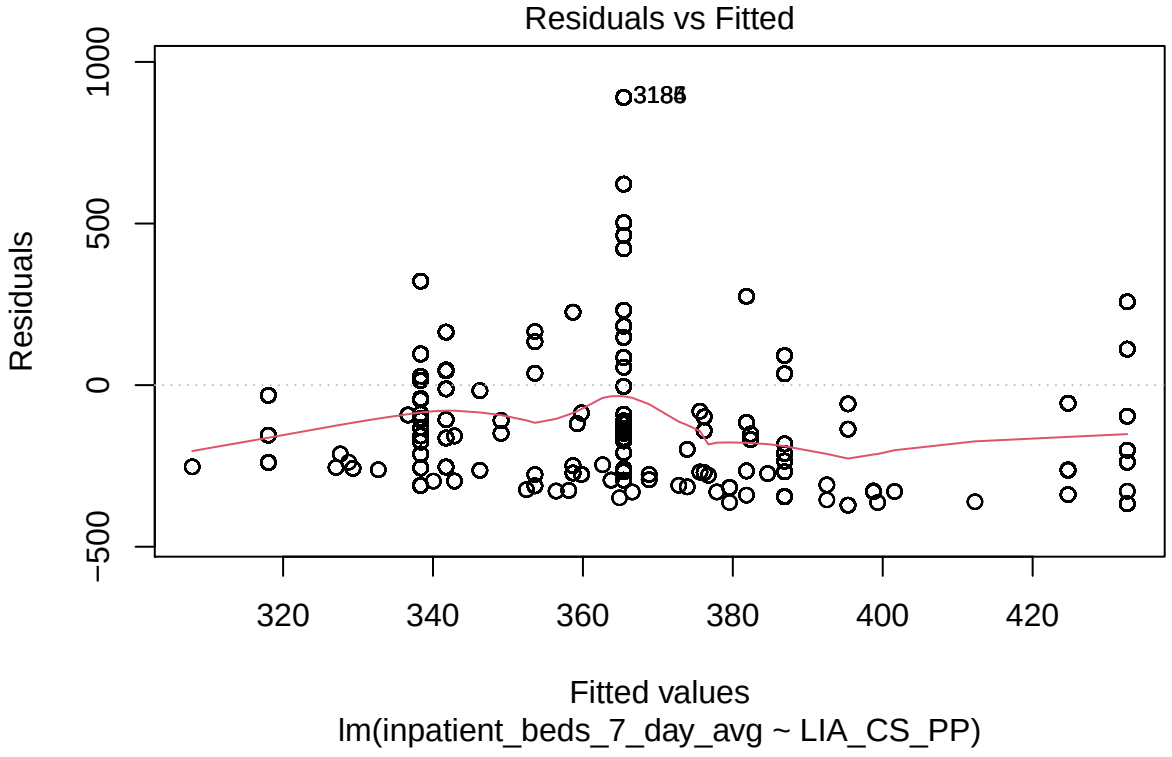
We sought to identify any interactions with the `Hardest Hit Area Score` on the models involving the depend variables and the only continuous independent varialbe `Low Income Area County SAIPE Poverty Percentage`. Here, `Hardest Hit Area Score` was used to create subpopulations. However, it turned out the results were not meaningful as most of the underlying assumptions of simple linear regression were greatly violated.

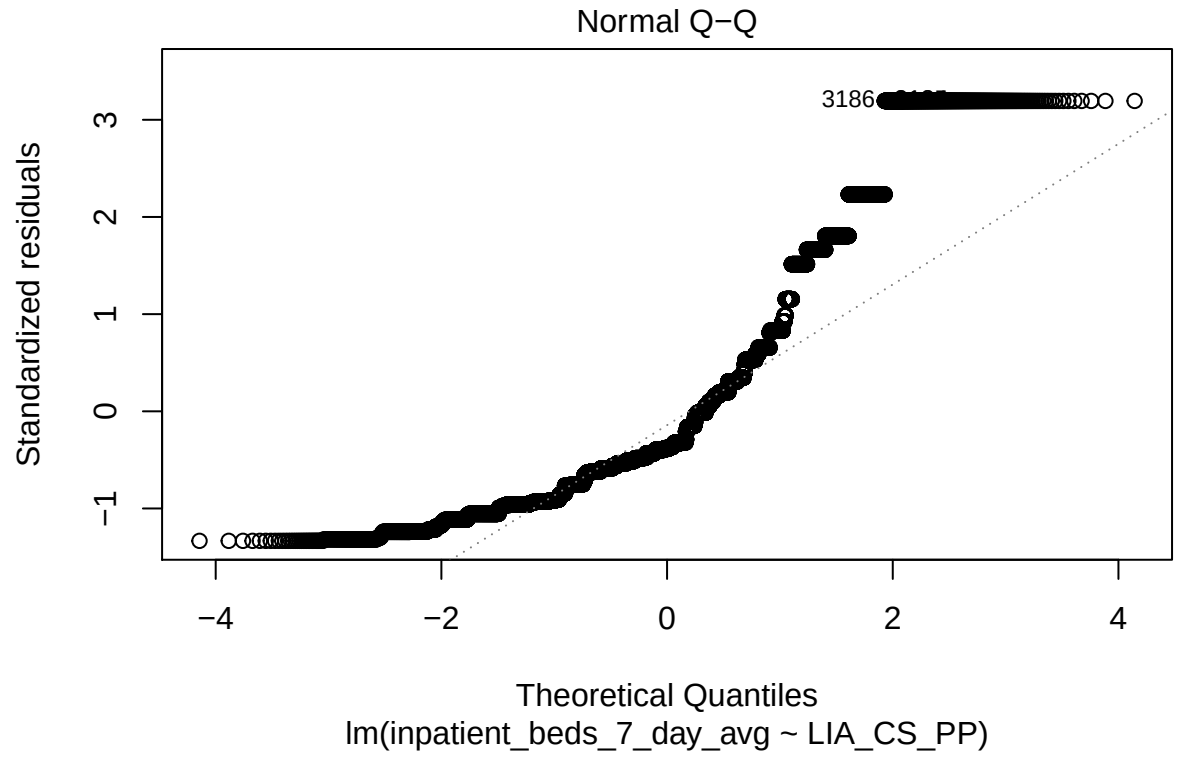
Remarks

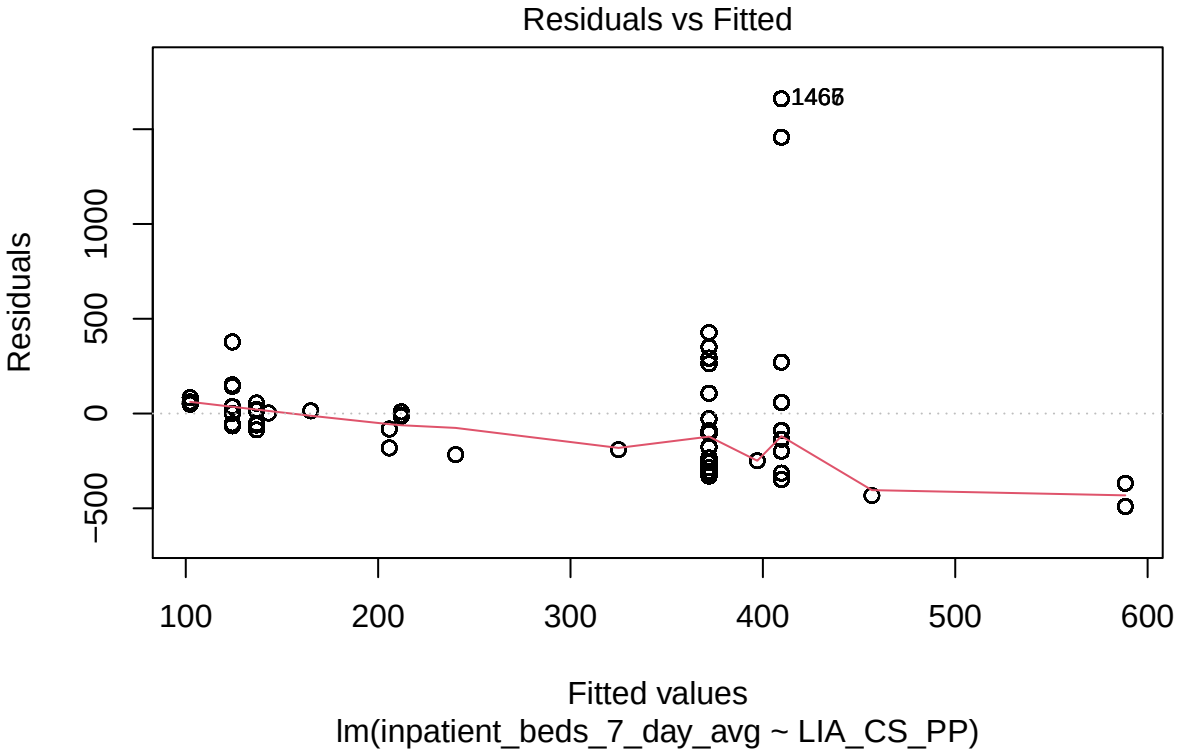
The level 15 for HHA_score had the best R squared among other factors. However, the residual plot reveal large nonconstancy of the error variance and doesnot depict linearity among all the models for the different levels of the HHA_score. Also, the normal QQ or normal probability plot deviates higher from normality. So, we see that these interactions do not give any meaningful results.

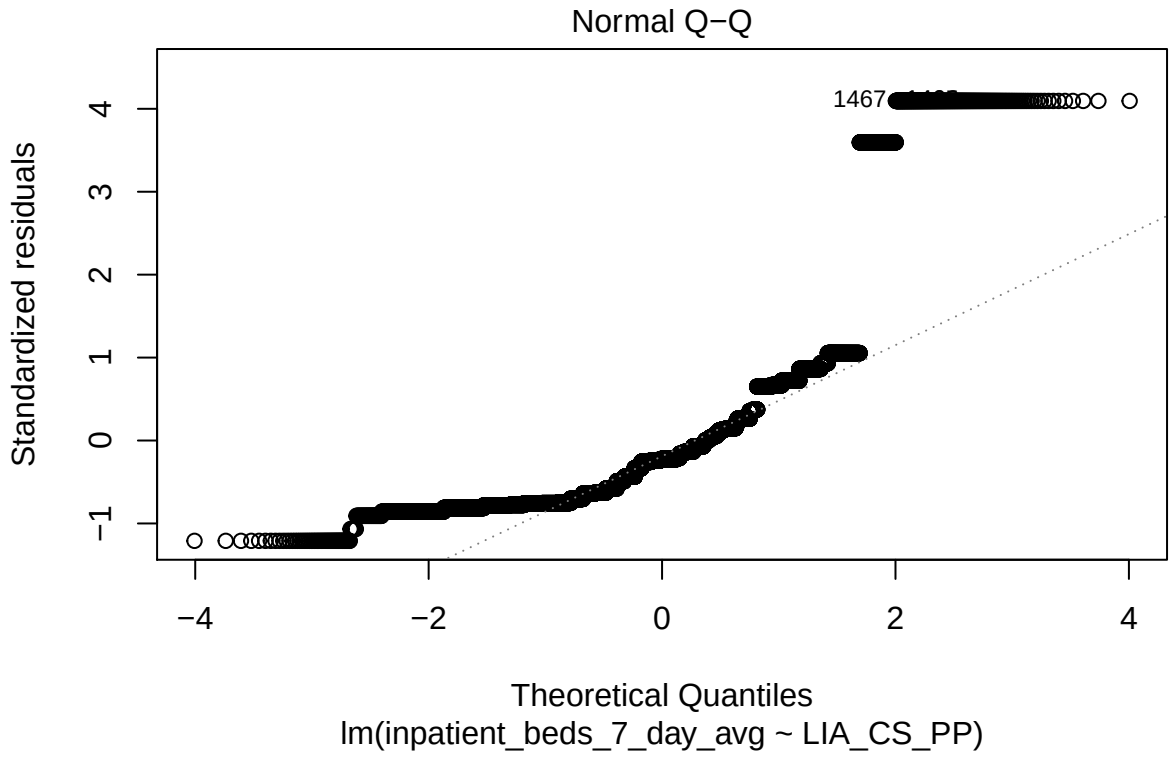
HHA_Score	term	estimate	std.error	statistic	p.value	conf.low	conf.high
0	(Intercept)	280.77	7.192	39.038	0	266.67	294.87
0	LIA_CS_PP	564.52	48.594	11.617	0	469.27	659.76
7	(Intercept)	-67.08	13.874	-4.835	0	-94.27	-39.89
7	LIA_CS_PP	3136.28	103.614	30.269	0	2933.19	3339.38
15	(Intercept)	607.54	52.705	11.527	0	504.01	711.08
15	LIA_CS_PP	-1748.73	323.361	-5.408	0	-2383.93	-1113.52

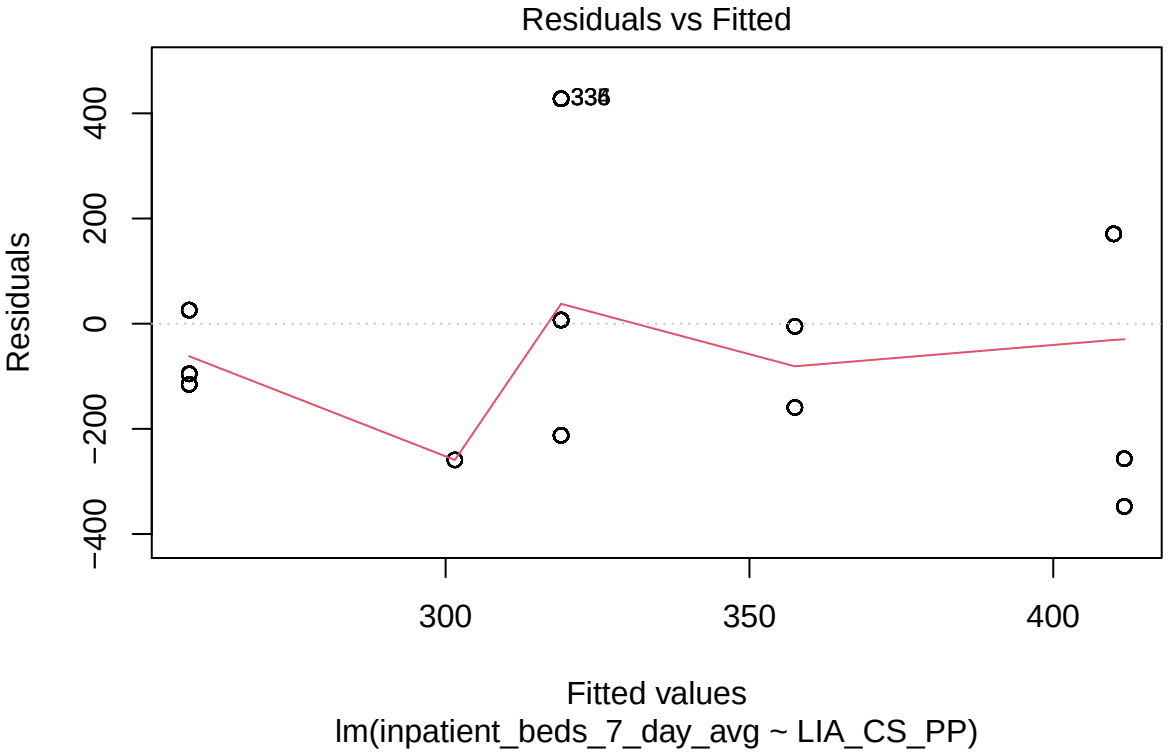
HHA_Score	r.squared	adj.r.squared	F.statistic	p.value	MSE
0	0.0046	0.0046	134.95	0	77630
7	0.0535	0.0534	916.20	0	164416
15	0.0515	0.0497	29.25	0	50628

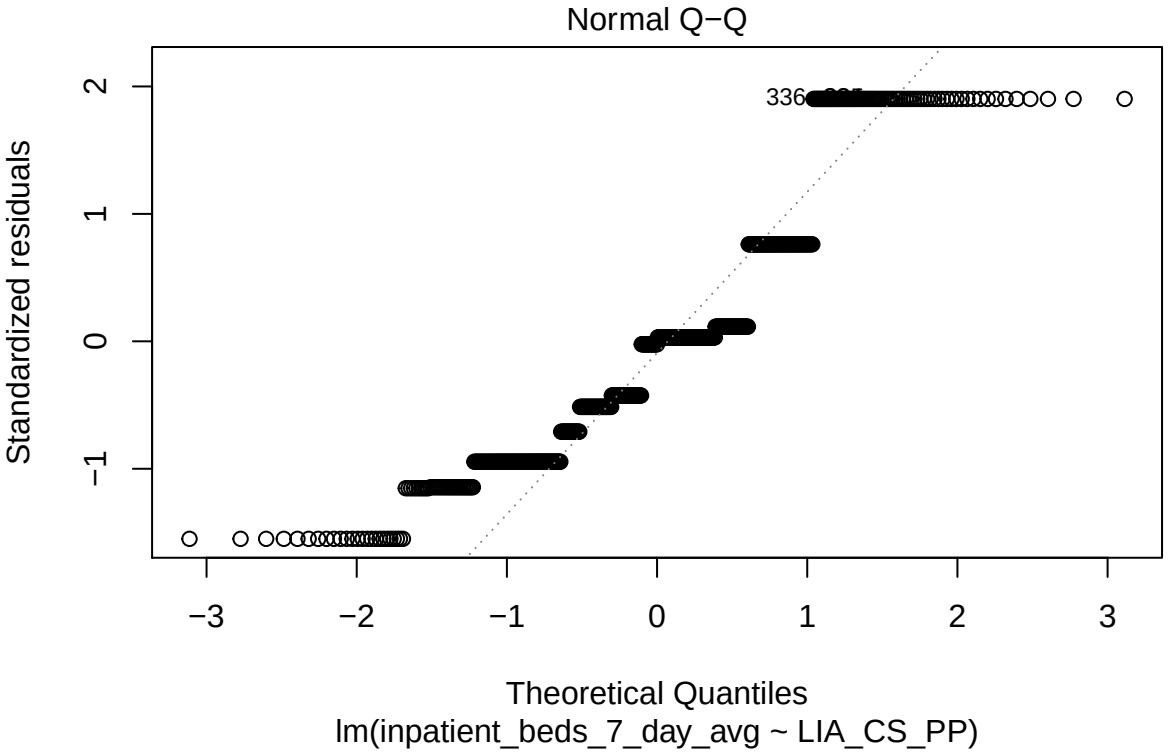



Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(inpatient_beds_7_day_avg ~ LIA_CS_PP)
3186


Residuals vs Fitted
Residuals
Fitted values
lm(inpatient_beds_7_day_avg ~ LIA_CS_PP)


Normal Q−Q
Standardized residuals
1467
−1
0
1
2
3
4
−4
−2
0
2
4
Theoretical Quantiles
lm(inpatient_beds_7_day_avg ~ LIA_CS_PP)


Residuals vs Fitted
Residuals
−400
−200
0
200
400
300
350
400
Fitted values
lm(inpatient_beds_7_day_avg ~ LIA_CS_PP)



Remarks The level 7 for HHA_score had the best R squared among other factors. However, the residual plot reveal large nonconstancy of the error variance and doesnot depict linearity among all the models for the different levels of the HHA_score. Also, the normal QQ or normal probability plot deviates higher from normality. So, we see that these interactions do not give any meaningful results.

3 Discussion

3.1 Key results

The study suggested that simultaneous log transform of the response variables `inpatient_beds_7_day_avg` and `total_icu_beds_7_day_avg` with the predictor variable `Low Income Area LIA County.SAIPE Poverty Percentage` describe how hospital capacity is associated with community vulnerabilities. We also found out that the variables `inpatient_beds_used_covid_7_day_avg` and `HHA_Score` describe how hospital capacity is associated with community vulnerabilities. Overall, the study revealed that the log simultaneous transformation explained well how hospital capacity is

associated with community vulnerabilities. This shows that hospital capacity is associted with some of the community vulnerability measures.

The residual plot revealed some potential outliers whose influence on our models will be treated in our next analysis.

3.2 Limitations

We want to formally put on record that the results and conclusions have to be considered alongside the follow caveats or limitations:

- The data provided to us for the analysis consisted of some observations but not all the data covering the entire population. For instance, due to the large size of the original data sets only a few were subsetted from the Covid hospital capacity data set. Since this selection was not random we think the final data used is not representative of the study population.
- The model selection procedure from the initial stage for selecting the appropriate predictor variable for each response variable is somewhat rudimentary. We know better selection procedures, such as stepwise regression, exist but clearly the scope of what has been studied at this stage of the course does not allow for that.
- We were constrained in this study to use only a simple linear regression model when a multiple regression model or other advanced modeling algorithms would have led to better performance.
- Lack of adequate previous research studies on the topic.
- We think they should have provided the raw data and not just the averages since the averages serves as the center of the observation if there are no outliers of any leverage points.

3.3 Interpretation

We observed that at 5% significance level, the slope and intercept for the three models both fall in their respective confidence intervals. Hence, the slope and intercept for the three models are statistically significant.

We also observed that the model between `total_icu_beds_7_day_avg` and `Low Income Area LIA County.SAIPE Poverty Percentage` had R^2 value of *0.0269693*, *R^2=0.0268039* for the model between inpatient_beds_7_day_avg and `Low Income Area LIA County.SAIPE Poverty Percentage` and *R^2=0.0556804* the model between inpatient_beds_used_covid_7_day_avg and HHA.Score is *R^2=0.0556804*. Thus, we see that about 5% of the variation in inpatient_beds_used_covid_7_day_avg is explained by using `HHA.Score` to predict inpatient_beds_used_covid_7_day_avg.Also, 2% of the variation in total_icu_beds_7_day_avg is explained by using `Low Income Area LIA County.SAIPE Poverty Percentage` to predict total_icu_beds_7_day_avg and 2% of the variation in inpatient_beds_7_day_avg is explained by using `Low Income Area LIA County.SAIPE Poverty Percentage` to predict inpatient_beds_7_day_avg.

3.4 Generalisability

Overall, we think our results do not generalize well to the population under consideration for the followin reasons. Instead of simply subsetting the first 5010 observations to obtain a manageable sample size, a random sample representative of the population could have been obtained to enhance the generalizability of our results to the entire population and probably beyond. Moreover, the fact that the study was limited to only the State of Texas also means that, though the study data was collected across US, sadly our results cannot be generalized to the entire US for decision making that will affect the whole nation.

4 Appendix

4.1 A: Models from which the initial "best" models were selected for Further Analyses

4.1.1 Initial Model results for `inpatient_beds_7_day_avg`

Table 14: Response: inpatient-beds-7-day-avg untransformed

predictor	r.squared	adj.r.squared	MSE	F.statistic	p.value	AIC
Tribal_Community	0.0004	0.0004	111489	9.47	1e-04	665015
Rural_Score	0.0032	0.0032	111177	146.98	0e+00	664885
is_metro_micro	0.0004	0.0003	111492	16.76	0e+00	665015
HHA_Score	0.0009	0.0009	111431	21.47	0e+00	664991
LIA_CS_PP	0.0179	0.0179	109535	838.85	0e+00	664200

Table 15: Response: inpatient-beds-7-day-avg with log transform

predictor	r.squared	adj.r.squared	MSE	F.statistic	p.value	AIC
Tribal_Community	0.0008	0.0008	0.7803	18.40	0	119114
Rural_Score	0.0104	0.0103	0.7728	481.31	0	118670
is_metro_micro	0.0021	0.0021	0.7793	98.43	0	119051
HHA_Score	0.0234	0.0234	0.7626	551.93	0	118060
LIA_CS_PP	0.0166	0.0166	0.7680	777.55	0	118378

Table 16: Response: inpatient-beds-7-day-avg with square root transform

predictor	r.squared	adj.r.squared	MSE	F.statistic	p.value	AIC
Tribal_Community	0.0006	0.0005	58.51	13.47	0	317669
Rural_Score	0.0061	0.0061	58.19	283.48	0	317411
is_metro_micro	0.0009	0.0009	58.49	41.24	0	317653
HHA_Score	0.0091	0.0090	58.01	210.78	0	317276
LIA_CS_PP	0.0190	0.0189	57.43	888.83	0	316813

Table 17: Response: inpatient-beds-7-day-avg with cube root transform

predictor	r.squared	adj.r.squared	MSE	F.statistic	p.value	AIC
Tribal_Community	0.0007	0.0006	3.682	15.00	0	190472
Rural_Score	0.0074	0.0074	3.657	343.14	0	190158
is_metro_micro	0.0012	0.0012	3.680	55.53	0	190444
HHA_Score	0.0136	0.0135	3.635	316.20	0	189873
LIA_CS_PP	0.0185	0.0185	3.616	868.27	0	189639

4.1.2 Initial Model results for `inpatient_beds_used_covid_7_day_avg`

Table 18: Response: inpatient-beds-used-covid-7-day-avg untransformed

predictor	r.squared	adj.r.squared	MSE	F.statistic	p.value	AIC
Tribal_Community	0.0003	0.0002	6332	6.372	0.0017	533098
Rural_Score	0.0053	0.0052	6300	243.523	0.0000	532865
is_metro_micro	0.0003	0.0003	6331	15.396	0.0001	533093
HHA_Score	0.0366	0.0365	6102	872.339	0.0000	531398
LIA_CS_PP	0.0030	0.0029	6315	136.540	0.0000	532972

Table 19: Response: inpatient-beds-used-covid-7-day-avg with log transform

predictor	r.squared	adj.r.squared	MSE	F.statistic	p.value	AIC
Tribal_Community	0.0006	0.0005	0.8727	13.275	0.000	124257
Rural_Score	0.0125	0.0125	0.8622	583.355	0.000	123702
is_metro_micro	0.0014	0.0013	0.8719	63.047	0.000	124219
HHA_Score	0.0557	0.0556	0.8245	1355.806	0.000	121649
LIA_CS_PP	0.0000	0.0000	0.8731	1.741	0.187	124280

Table 20: Response: inpatient-beds-used-covid-7-day-avg with square root transform

predictor	r.squared	adj.r.squared	MSE	F.statistic	p.value	AIC
Tribal_Community	0.0004	0.0004	14.59	9.384	1e-04	253807
Rural_Score	0.0086	0.0086	14.47	398.851	0e+00	253427
is_metro_micro	0.0007	0.0007	14.59	32.702	0e+00	253791
HHA_Score	0.0451	0.0451	13.94	1086.247	0e+00	251703
LIA_CS_PP	0.0014	0.0013	14.58	62.637	0e+00	253761

Table 21: Response: inpatient-beds-used-covid-7-day-avg with cube root transform

predictor	r.squared	adj.r.squared	MSE	F.statistic	p.value	AIC
Tribal_Community	0.0005	0.0004	1.506	10.65	0	149368
Rural_Score	0.0099	0.0099	1.492	460.92	0	148928
is_metro_micro	0.0009	0.0009	1.506	41.42	0	149345
HHA_Score	0.0486	0.0485	1.434	1174.31	0	147098
LIA_CS_PP	0.0008	0.0008	1.506	35.76	0	149351

4.1.3 Initial results for `total_icu_beds_7_day_avg`

Table 22: Response: total-icu-beds-7-day-avg untransformed

predictor	r.squared	adj.r.squared	MSE	F.statistic	p.value	AIC
Tribal_Community	0.0002	0.0002	2915	5.742	0.0032	497426
Rural_Score	0.0028	0.0028	2908	131.435	0.0000	497304
is_metro_micro	0.0003	0.0003	2915	12.895	0.0003	497422
HHA_Score	0.0033	0.0033	2906	77.141	0.0000	497283
LIA_CS_PP	0.0164	0.0164	2868	766.238	0.0000	496675

Table 23: Response: total-icu-beds-7-day-avg with log transform

predictor	r.squared	adj.r.squared	MSE	F.statistic	p.value	AIC
Tribal_Community	0.0004	0.0003	0.8868	8.247	3e-04	124996
Rural_Score	0.0078	0.0078	0.8801	362.912	0e+00	124649
is_metro_micro	0.0013	0.0013	0.8859	61.172	0e+00	124949
HHA_Score	0.0135	0.0134	0.8751	314.370	0e+00	124388
LIA_CS_PP	0.0173	0.0173	0.8717	809.053	0e+00	124208

Table 24: Response: total-icu-beds-7-day-avg with square root transform

predictor	r.squared	adj.r.squared	MSE	F.statistic	p.value	AIC
Tribal_Community	0.0003	0.0003	1.212	7.236	7e-04	139350
Rural_Score	0.0057	0.0057	1.205	262.335	0e+00	139101
is_metro_micro	0.0008	0.0008	1.211	36.546	0e+00	139326
HHA_Score	0.0093	0.0093	1.201	216.047	0e+00	138934
LIA_CS_PP	0.0180	0.0180	1.190	843.737	0e+00	138526

Table 25: Response: total-icu-beds-7-day-avg with cube root transform

predictor	r.squared	adj.r.squared	MSE	F.statistic	p.value	AIC
Tribal_Community	0.0003	0.0003	10.29	6.819	0.0011	237747
Rural_Score	0.0048	0.0048	10.25	221.592	0.0000	237537
is_metro_micro	0.0006	0.0006	10.29	28.048	0.0000	237730
HHA_Score	0.0075	0.0075	10.22	174.062	0.0000	237414
LIA_CS_PP	0.0180	0.0179	10.11	841.599	0.0000	236924

4.2 B: Model results for other analysis (subgroup)

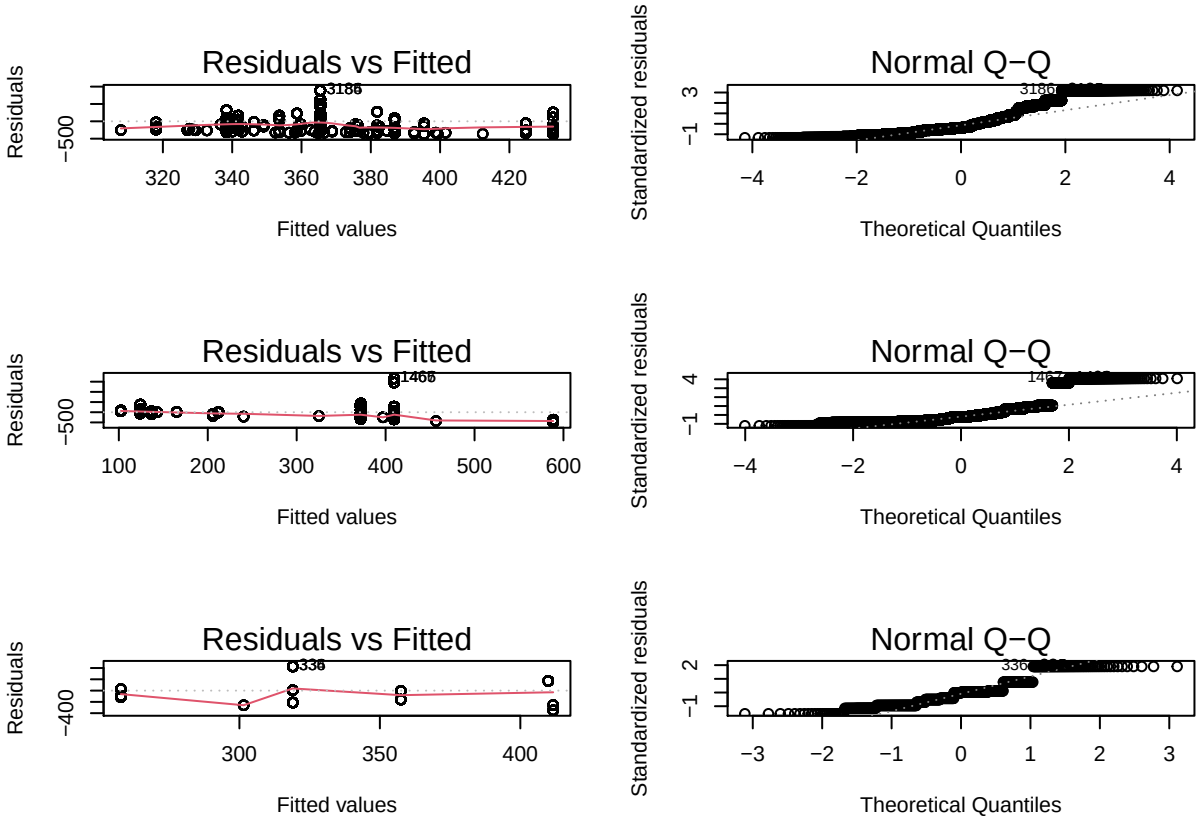
Modeling inpatient_beds_7_day_avg as a function of LIA_CS_P on the subgroups of HHA_Score

Table 26: Parameter estimates

HHA_Score	term	estimate	std.error	statistic	p.value	conf.low	conf.high
0	(Intercept)	280.77	7.192	39.038	0	266.67	294.87
0	LIA_CS_PP	564.52	48.594	11.617	0	469.27	659.76
7	(Intercept)	-67.08	13.874	-4.835	0	-94.27	-39.89
7	LIA_CS_PP	3136.28	103.614	30.269	0	2933.19	3339.38
15	(Intercept)	607.54	52.705	11.527	0	504.01	711.08
15	LIA_CS_PP	-1748.73	323.361	-5.408	0	-2383.93	-1113.52

Table 27: Overall Model Performance Results

HHA_Score	r.squared	adj.r.squared	F.statistic	p.value	MSE
0	0.0046	0.0046	134.95	0	77630
7	0.0535	0.0534	916.20	0	164416
15	0.0515	0.0497	29.25	0	50628



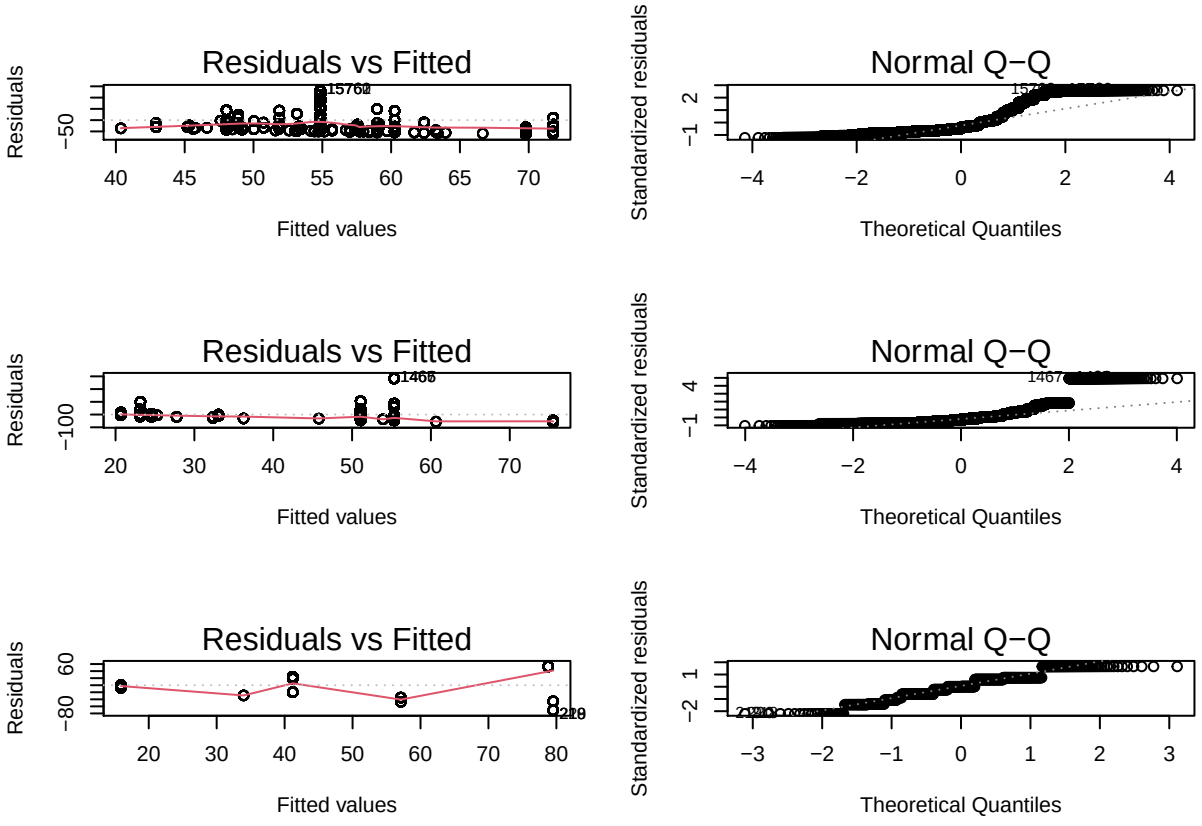
Modeling total_icu_beds_7_day_avg as a function of LIA_CS_P on the subgroups of HHA_Score

Table 28: Parameter estimates

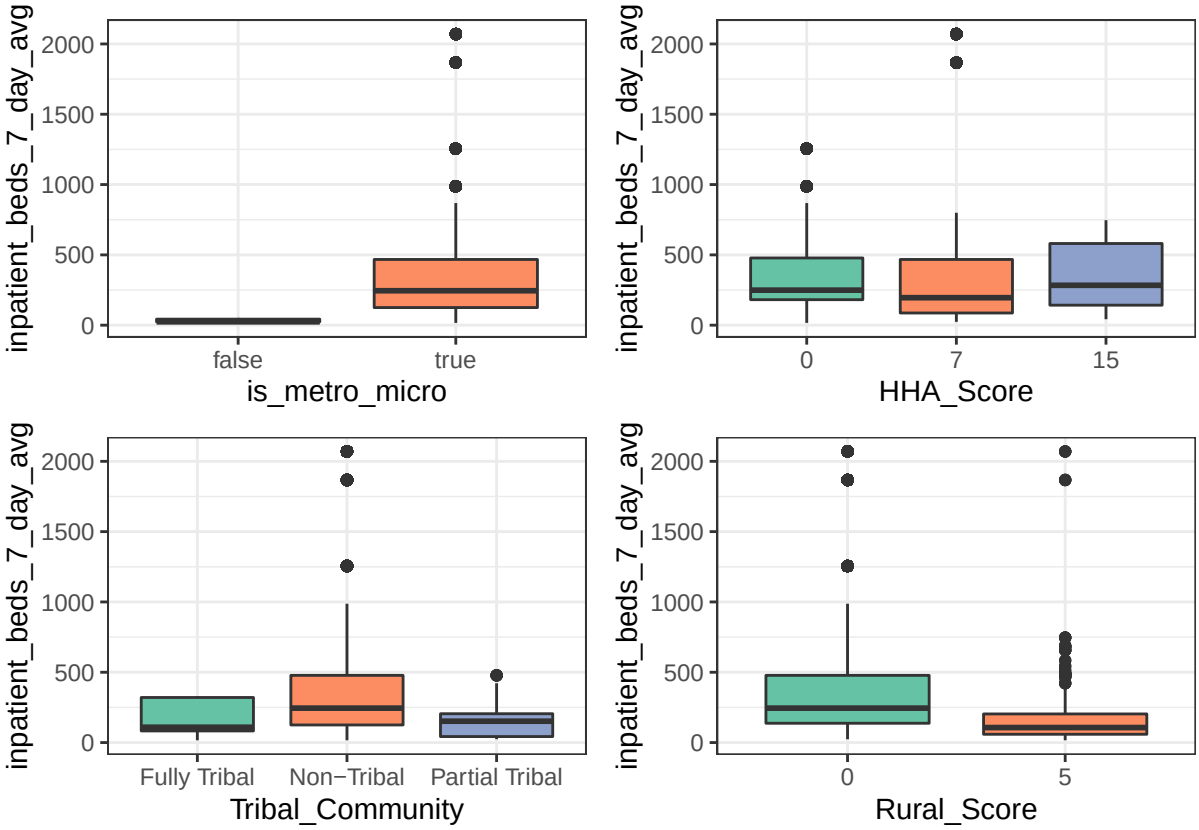
HHA_Score	term	estimate	std.error	statistic	p.value	conf.low	conf.high
0	(Intercept)	33.538	1.320	25.4032	0.0000	30.951	36.126
0	LIA_CS_PP	142.139	8.920	15.9345	0.0000	124.655	159.623
7	(Intercept)	1.574	1.971	0.7985	0.4246	-2.289	5.437
7	LIA_CS_PP	353.723	14.719	24.0313	0.0000	324.871	382.574
15	(Intercept)	160.505	7.556	21.2414	0.0000	145.662	175.348
15	LIA_CS_PP	-723.043	46.359	-15.5965	0.0000	-814.110	-631.976

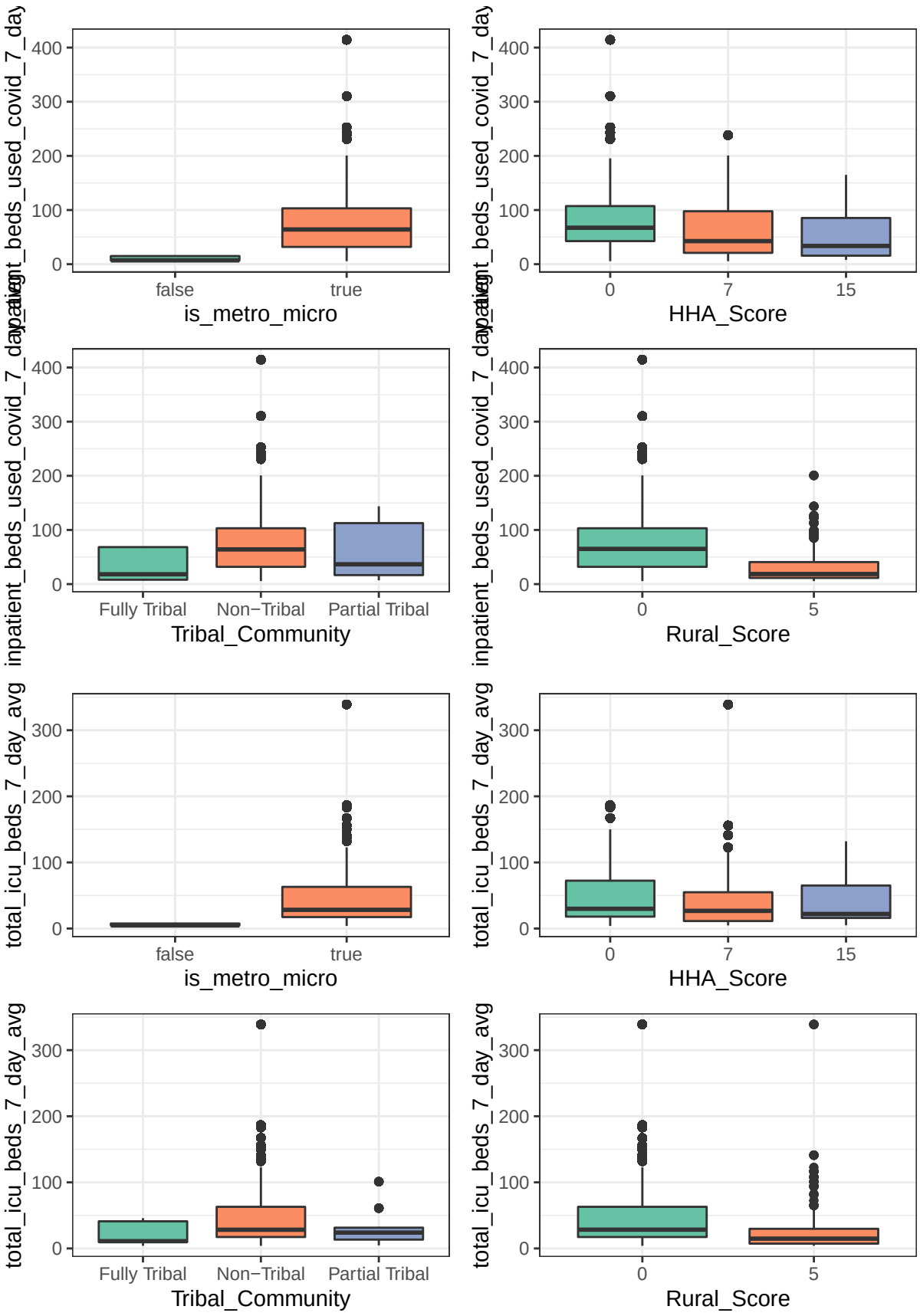
Table 29: Overall Model Performance Results

HHA_Score	r.squared	adj.r.squared	F.statistic	p.value	MSE
0	0.0086	0.0086	253.9	0	2616
7	0.0344	0.0343	577.5	0	3318
15	0.3110	0.3097	243.3	0	1041



4.3 C: Effect of categorical independent variables on the dependent variables

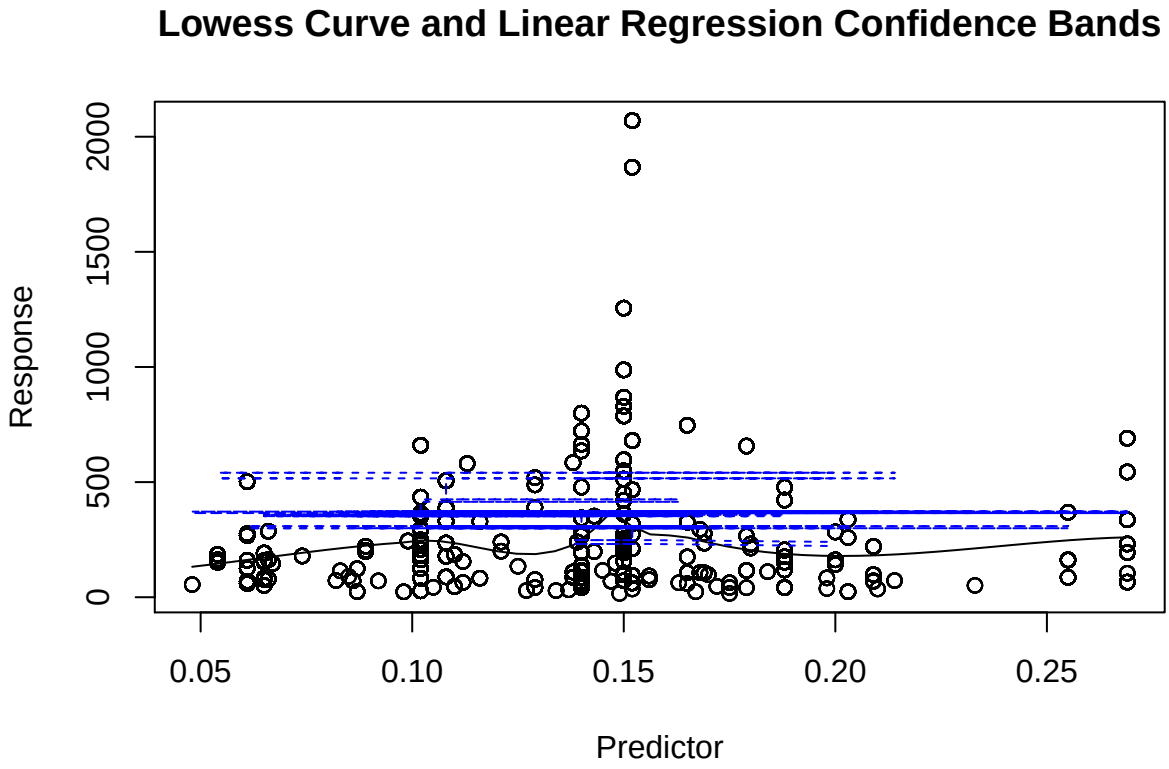


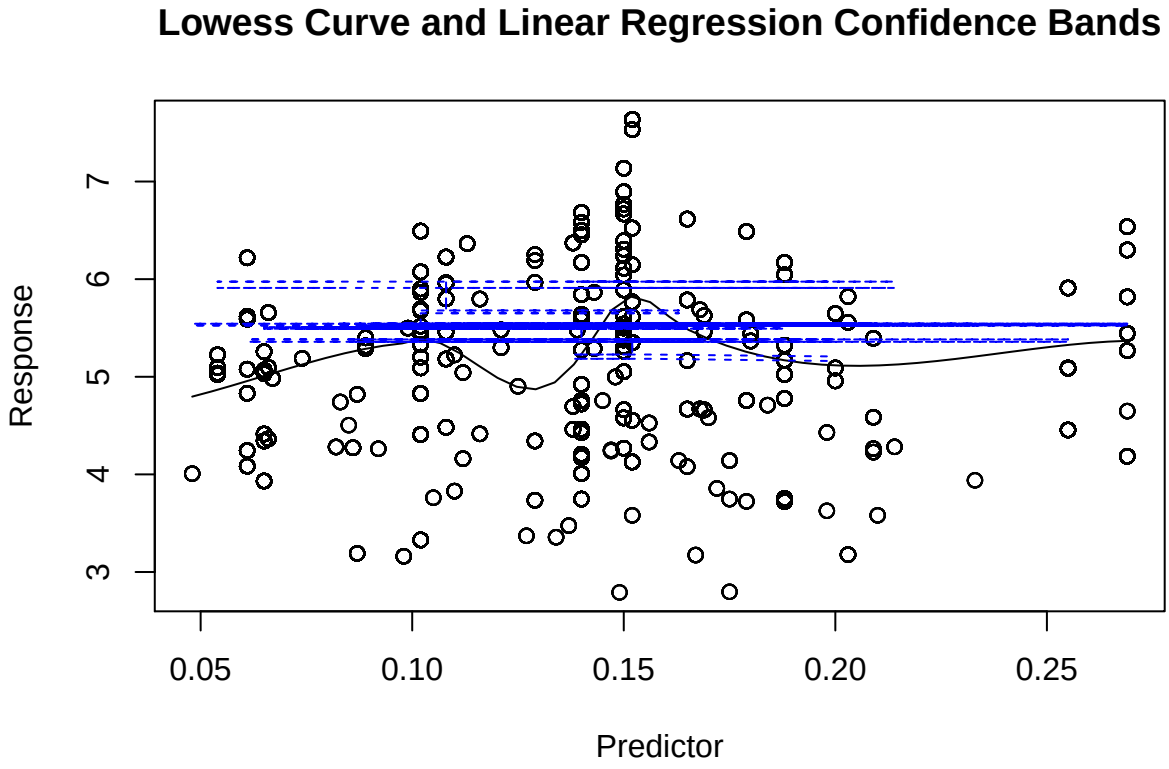

inpatient_beds_used_covid_7_day_avg
0
100
200
300
400
false
true
is_metro_micro
0
7
15
HHA_Score
Fully Tribal
Non−Tribal
Partial Tribal
Tribal_Community
0
5
Rural_Score
total_icu_beds_7_day_avg

4.4 D

```
##
##  Pearson's product-moment correlation
##
## data:  corr[, 3] and corr[, 1]
## t = 413, df = 45989, p-value <2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8853 0.8892
## sample estimates:
##    cor
## 0.8873
```

```
##
##  Pearson's product-moment correlation
##
## data:  corr[, 3] and corr[, 1]
## t = 2239, df = 45989, p-value <2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9954 0.9955
## sample estimates:
##    cor
## 0.9954
```


Lowess Curve and Linear Regression Confidence Bands
Response
0
500
1000
1500
2000
0.05
0.10
0.15
0.20
0.25
Predictor


Lowess Curve and Linear Regression Confidence Bands
Response
3
4
5
6
7
0.05
0.10
0.15
0.20
0.25
Predictor

5 References

- Tsai, Thomas C., et al. "Association of community-level social vulnerability with US acute care hospital intensive care unit capacity during COVID-19." Healthcare. Vol. 10. No. 1. Elsevier, 2022.
- Grimm, Christi A. "Hospitals reported that the COVID-19 pandemic has significantly strained health care delivery." (2021). Accessed from (https://oig.hhs.gov/oei/reports/OEI-09-21-00140.pdf) on 03/08/2022