Modeling the Association between hospital capacity and community vulnerabilities

Phase II Report

William O. Agyapong Prince Appiah Eti Nyamekeh Baffoe

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

Abstract

We analyzed the relationship between hospital capacity and community vulnerabilities using a multiple regression procedures. Overall, it was discovered that the predictor variable Low Income Area LIA County SAIPE Poverty Percentage, is_metro_micro, tribal community, hardest hit area score, rural score and the response variable inpatient beds 7 day average explained to a greater 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 the hospital subtype. Variations in Low Income Area LIA County SAIPE Poverty Percentage, is_metro_micro, and tribal community have positive influence on hospital bed capacity, while hardest hit area score, rural score and hospital subtype affect hospital bed capacity negatively.

1 Introduction

In this project, we take a multiple regression approach to try to understand the association between hospital beds capacity and community vulnerabilities. At this stage, we seek to identify the best set of variables including at least one community vulnerability measure and other relevant factors that are able to significantly explain our measure of hospital capacity, **inpatient_beds_7_day_avg**.

1.1 Background and Motivation

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 USA 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 a 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, the need to investigate the nature of the relationship that exists between hospital beds capacity and community vulnerabilities cannot be overemphasized. This has the potential of providing great insights and guidelines to policy and decision makers to help them take necessary actions to prevent strained ICU capacity from compounding COVID-19 inequities. In this report, we conduct a multiple linear regression analysis to investigate whether hospital beds capacity measured by the **inpatient_beds_7_day_avg** variable is associated with community vulnerabilities using data obtained from the U.S. Department of Health and Human Services Protect database.

1.2 Objectives

Among other things, this report aims to address the following research question.

How is hospital capacity associated with community vulnerabilities?

Using **inpatient_beds_7_day_avg** as a quantifier for hospital capacity, our goal is to identify an appropriate regression model that utilizes the appropriate community vulnerability measure(s) and other relevant factors to explain much of the variability in hospital capacity, while serving as a reliable predictive model.

Our initial hypothesis is that the vulnerability measure, ***low income area county poverty percentage***, and a potential confounding variable ***hospital subtypes*** are likely to have the most significant effect on the dependent variable 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 nine variables obtained from a facility-level hospitalization data as well as a community-level vulnerability data. The following table provides information about these variables used in the study. As indicated by the role column of Table 1, our target or response variable is **inpatient_beds_7_day_avg** which serves as a measure of hospital capacity and constitute the seven-day average of the reports provided for a given facility for that element during that collection week, five independent (predictor) variables measuring community vulnerability, and two potential confounding variables, hospital subtype and city.

Table 1: Definition of Variables of interest

Variable Name	Description	Type of Measure	Data Type	Role
inpatient beds 7 day average	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 average	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 average	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/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
LIA_CT_PP	Low Income Area (LIA) Census Track Poverty Percentage	Community vulnerability	Numeric/continuous	Predictor
Rural/Rural_Score	An indicator of whether the facility is at a rural location or not	Community vulnerability	Integer/Categorical	Predictor
Tribal Community	An indicator of whether the facility is found in a tribal community. Possible values are fully tribal, non−tribal, and partial tribal.	Community vulnerability	Character/Categorical	Predictor
Hospital subtype	The sub-type of the facility reporting. Valid values are: Children's Hospitals, Critical Access Hospitals, Long Term, Psychiatric, Rehabilitation & Short Term.	-	Factor/Categorical	Potential confounder
City	The city of the facility reporting.	-	Factor/Categorical	Potential confounder

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 databases 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 volume of the data, only 5009 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) to be used as the sole target variable. These observations were then merged with the community-level vulnerability data through the use of an inner join on the Federal Information Processing Standard (FIPS) code of the location of the hospital existing in both data sets.

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 for phase I of our study because they provide an evaluation of the most vulnerable communities in our population. However, in this second phase of the study we first examined the differences between using the raw versus the scored variables and realized that the choice of one did not impact our results in any way, so we decided to use the raw variables to facilitate easy interpretation of our results.

1.7 Bias

- We believe that the study data is not representative of the study population since we simply took the first **5009** observations from the hospital capacity dataset as a way of obtaining a manageable sample size. 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.
- There is a high potential for our final model to overfit the data set used to build the model. To guard against this phenomenon, we employed a cross-validation procedure where the full data set was partitioned into a training set for training the model and a validation set for assessing the reasonableness and predictive ability of the final selected model.
- Another potential source of bias is the high level of class imbalance seen in the distribution of the categorical predictors. Some levels of all these four qualitative measures, `HHA_Score`, `Rural Score`, `Tribal Community`, `is_metro_micro`, and `hospital subtype`, are disproportionately represented. This can lead to a serious problem where our findings would be unfairly biased towards a particular subgroup. After consultation with our project advisor in the person of Dr. Amy Wagler, it was

decided that the treatment of class imbalance is beyond the scope of the project, so we did not address this potential source of bias in this second phase 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 in Section 2 for what we considered to be non-reported values. In the end, the data used for the analysis had 45797 observations making up our study size.

1.9 Statistical methods

1.9.1 Regression model and Assumptions

The main statistical method used in this project is the multiple regression model

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 Xi2 + \beta_3 X_{i3} + ... + \beta_{p-1} X_{ip-1} + \epsilon_i$ where

- The parameters are β_j where $j = 0, 1, 2, ..., p-1$
- ϵ_i are independent $N(0, \sigma^2)$
- X_{ij} denotes the jth independent variable for $i = 1, ..., n$.
- β_j measures the effect X_j has on the independent variable Y Note that $X_j = X_{ij}$.
- Generally the Multiple linear Regression can be represented in matrix form as:

$$Y_{nxn} = X'_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$$

where

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \begin{bmatrix} 1 & X_{11} & X_{12} ... X_{1n} \\ 1 & X_{11} & X_{12} ... X_{1n} \\ \vdots & \vdots & \vdots ... \vdots \\ 1 & X_{n1} & X_{n2} ... X_{np} \end{bmatrix}, \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}, \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

We check for the following assumptions.

- There must be a linear relationship between the outcomes(dependent) variables and the independent variables: We use the scatter plot to check for this linear relationship.
- Multivariate Normality: Multiple regression assumes that the residuals are normally distributed.
- No Multicollinearity: Multiple regression assumes that the independent variables are not highly correlated with each other.We test this assumption using the Variance Inflation factor(VIF) values.
- Homoscendasticity: This assumption states that the variance of error terms are similar across the values of the independent variables: Plotting the standardized residual versus the predicted shows us whether the points are equally distributed across all values of the independent variables.

1.9.2 Effect Of Multicollinearity

Multicollinearity occurs when there is correlation between independent variables in a regression model, this is a problem because as seen in the assumptions above the independent variables should be independent,but if the degree of the correlation between variables is high enough, it can cause problems when fitting the moideland interpreting the results. This is because our key goal for this analysis is to isolates the relationship between each independent variable and dependent variable.

Therefore the interpretation of the regression coefficients is it that it represents the mean change in the dependent variable for every one unit change in an independent variable when we hold all of the other independent variables constant. Therefore it becomes very difficult for the model to estimate the relationship between each independent variable and the dependent variable independently because the independent variables tend to change in unison.

1.9.3 Model Selection

The data source presented to us with numerous variables that could serve as potential predictors. Obviously, using all available variables will amount to a needlessly complex model. Thus, after choosing the important measure of hospital capacity to model, we critically assessed each one of the community vulnerability measures based on information we gathered about them in order to ascertain whether they could form part of our initial predictor set. Six variables, `LIA_CS_PP`, `LIA_CT_PP`, `HHA`, `is_micro_metro`, `tribal community`, and `rural` stood out tall. In addition, we searched our data for variables that were

not direct determinants of community vulnerabilities but could potentially influence our model and settled on the `hospital subtype` and the `city` variables. This brought our predictor set to eight variables, to start with. Subsequently, a backward stepwise regression procedure was employed to help us narrow down our predictor set to the "best" subset of variables. Finally, using the final model reported by the automatic search procedure as a baseline model, we tried several other combinations of variables already in the model and those that were dropped including their interactions to arrive at our final model. It is noteworthy that our personal judgment guided by the goal of obtaining a more explanatory model played a key role in this pursuit. By and large, we were inspired by Kutner et al. (2004)'s remarks that "Judgment needs to play an important role in model building for exploratory studies. Some explanatory variables may be known to be more fundamental than others and therefore should be retained in the regression model if the primary purpose is to develop a good explanatory model".

1.9.4 Model Assessement

We diagnosed the appropriateness of our models using diagnostic plots such as the 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. We used the plot of the Residual versus Leverage to check the existence of any outliers and any possible influential observations. We also, checked the linear relationship assumption with the residual versus fitted plot as well as the **Added-Variable Plots**, both the response variable \textbf{inpatient_beds_7_day_avg} and the predictor variables were all regressed against each other predictor variables which were already in our model. We obtained the residuals for each. The plot of the residuals against the other set showed that there was a marginal contribution of the candidate. This was done largely to ensure that the various assumptions required for linear regression model listed above were reasonably satisfied. We therefore relied on the these diagnostics to select our final model.

To check the multicollinearity between our predictor variables we used the variance inflation factor (VIF). We calculated the VIF value for each predictor variables start and realized they were all below 10, suggeting the absence of multicollinearity.

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.

1.9.5 Model Validation

To be able to validate our candidate selected model, both internally and externally, we used a holdout sample to check the model and its predictive ability or the tendency of the model to generalize well to new, unseen data. As a result, we split the data set into 70% training set and 30% testing set corresponding to **32061** and **13736** training and validation samples, respectively. Given our large sample size, this partition ensured a sufficiently large model-building data set to ensure the development of a reliable model.

We compute mean squared prediction error as follows:

$$\text{MSPR} = \frac{\sum_{i=1}^{n_v}(Y_i - \bar{Y})^2}{n_v}$$

where:

Y_i is the observation of the response variable in the i^{th} validation case

$\bar{Y}$ is the predicted value for the i^{th} validation case based on the model-building data set.

n_v is the number of observations in the validation or test set.

All analyses were performed in the R Statistical Software version 4.0.2 (2020-06-22).

2 Analysis and Results

In this section, we present key results from our exploratary data analysis as well as the regression modeling procedures.

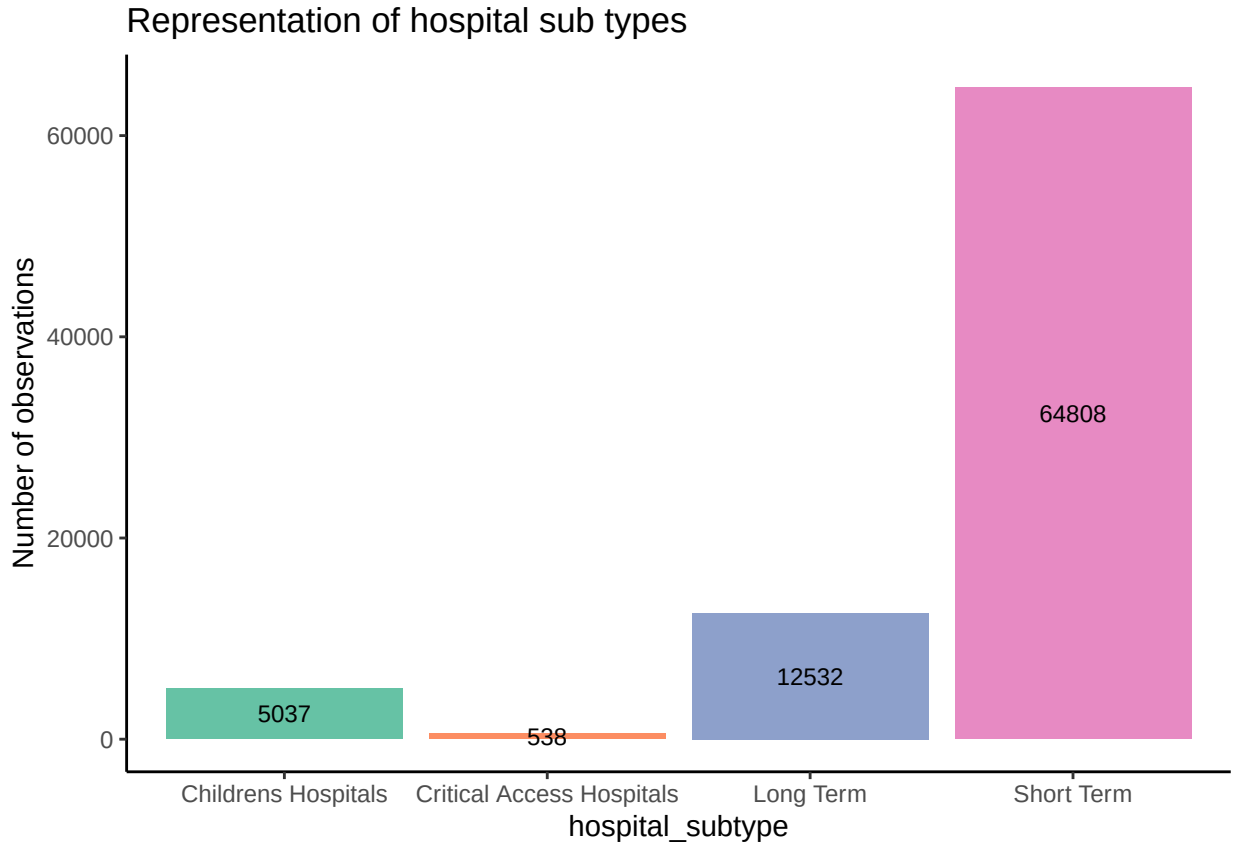
2.1 Descriptive data

We begin our analysis by providing both numerical summaries and graphs to enhance our understanding of the underlying data for the study.

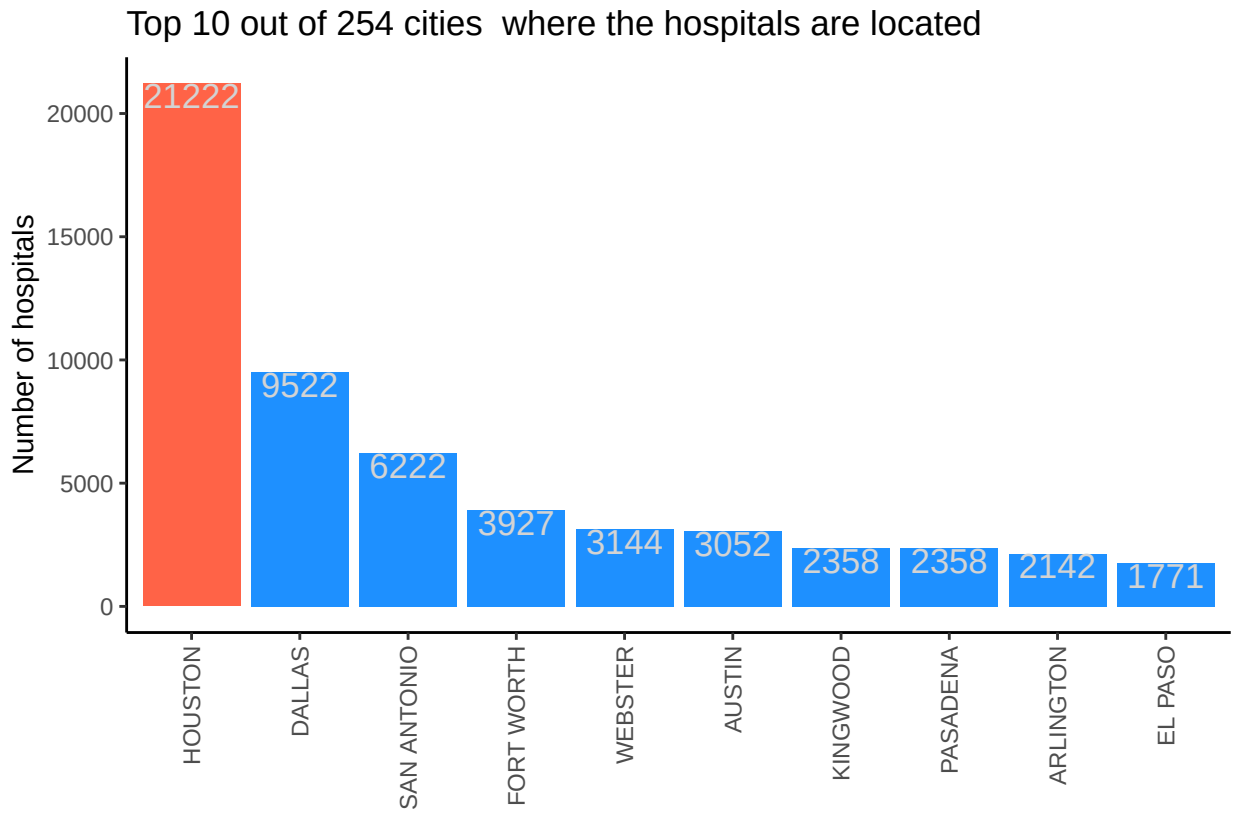
2.1.1 Characteristics of study participants

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
LIA_CT_PP	341	0.4113
inpatient_beds_7_day_avg	222	0.2677
LIA_CS_PP	0	0.0000
is_metro_micro	0	0.0000
HHA_Score	0	0.0000
Tribal_Community	0	0.0000
Rural_Score	0	0.0000
hospital_subtype	0	0.0000
city	0	0.0000

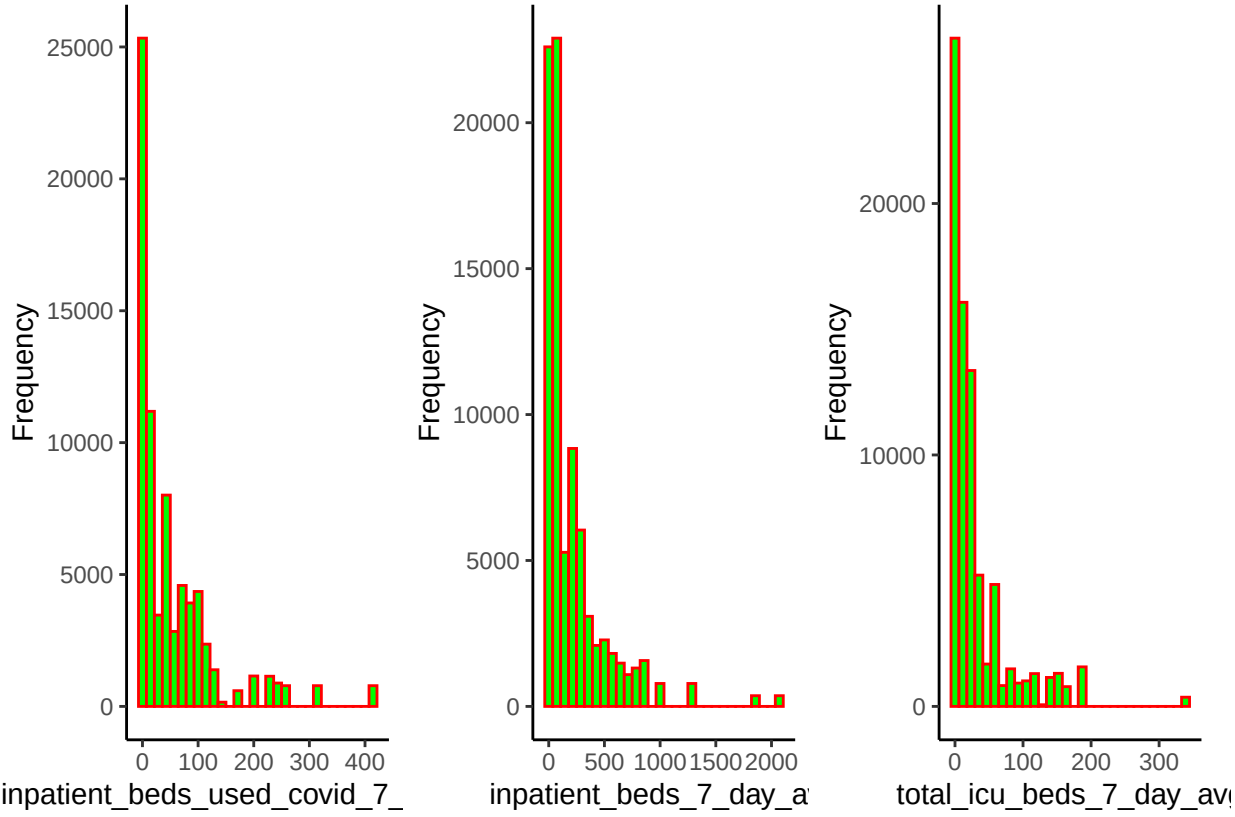
Table 3: Missing values in dependent variables from the covid data before merge

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.

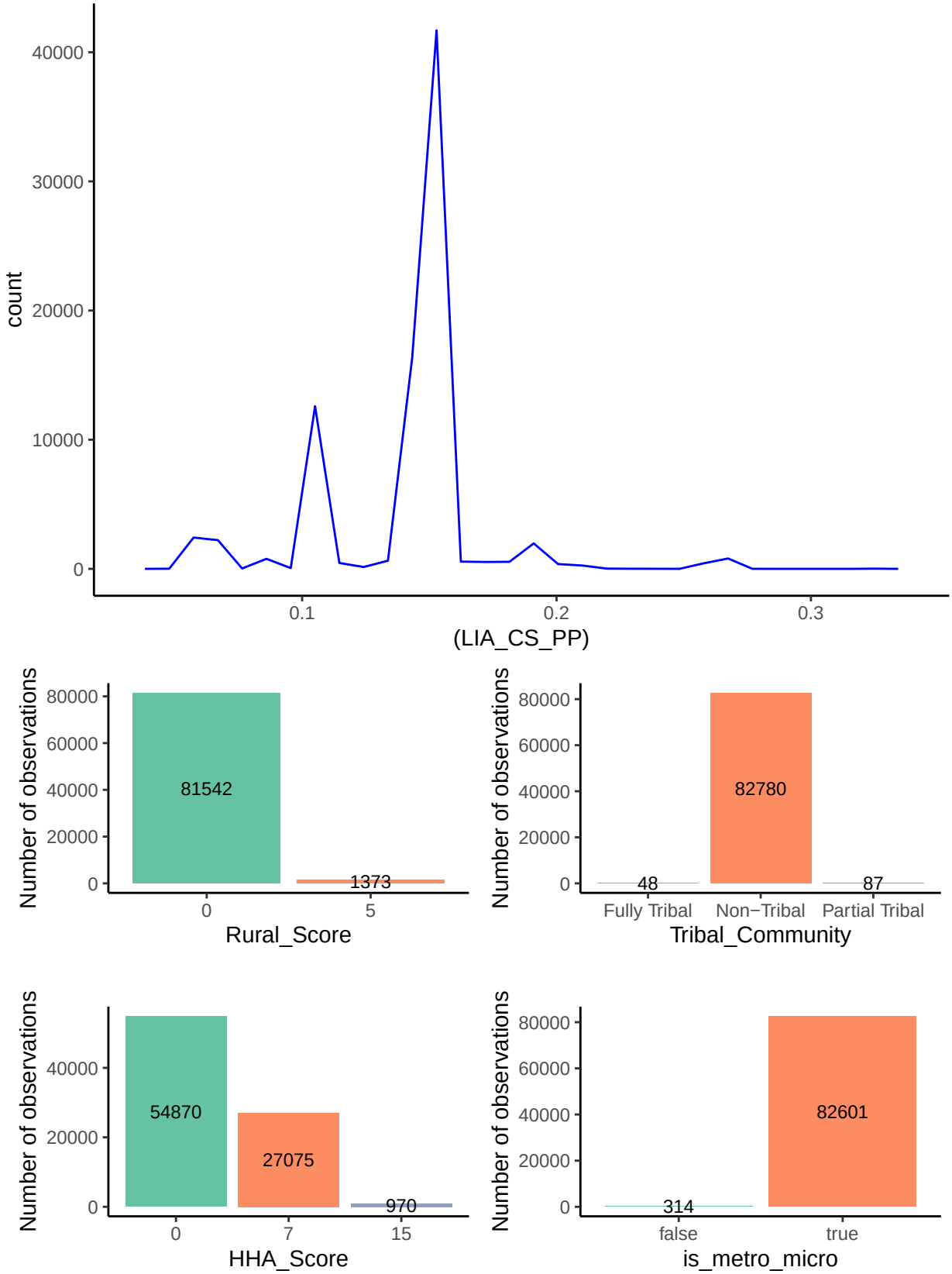
2.1.2.1 Non-reported values in dependent variables We also observed that the hospital capacity measures chosen for our dependent variables had zeros (0) in them. This was quite surprising at first because we did not expect to see zeros in these variables when the maintainers of the data ***suppressed 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

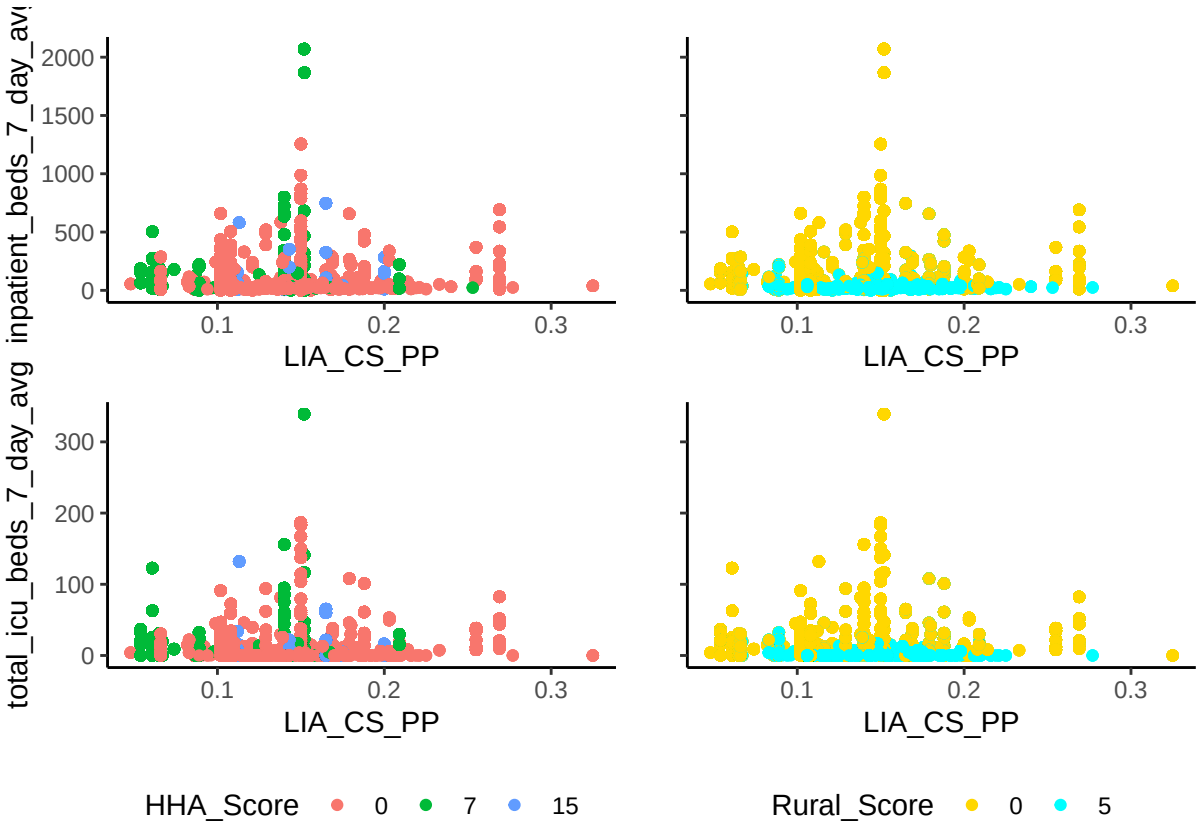


The distribution of the continuous independent/predictor variable, `Low Income Area County SAIPE`

`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 other 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 to present them here for the sake of brevity.

We also explored how each dependent variable is distributed across the levels of each one of the categor-

ical 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

Selecting **inpatient_beds_7_day_avg** as the response of interest, our modeling process began with an automatic search for important variables by performing a backward elimination procedure using the base R `step()` function. The full model utilized at the start of the algorithm contained eight variables described in Section **1.9.2**. Surprisingly, a thorough investigation revealed that whenever the `city` variable was included in the full model, all community vulnerability measures became insignificant such that our automatic selection procedure always dropped all of them from the model, leaving only `hospital subtype` and `city`. Meanwhile, removing `city` resulted in a substantial drop in performance in terms of R^2 and the residual standard error and substantial improvement in model assumption violations.

In any case, the many **116** levels of the city variable were probably causing overfitting, which is what you get when you have a qualitative variable with so many levels. Again, the resulting model deviated greatly from the aims of the project, in that none of the community vulnerability measures were retained. So we decided to remove the city variable, and conducted another automatic search where, to our surprise, all the community vulnerability measures together with the `hospital subtype` were retained in the final model, but with a huge drop in model performance.

In the end, we had to sacrifice a model with high performance but possible overfitting for a simple model with high explanatory value consistent with our objectives. The Table 2 shows our selected model at this stage, while Table 3 presents its performance alongside the other model when city was considered.

Table 4: Parameter estimates of selected model suggested by backward stepwise algorithm

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-78.90	135.977	-0.5802	0.5618	-345.42	187.62
LIA_CS_PP	1692.74	56.223	30.1073	0.0000	1582.54	1802.94
is_metro_microtrue	155.10	91.593	1.6933	0.0904	-24.43	334.62
HHA_Score7	33.32	4.000	8.3279	0.0000	25.47	41.16
HHA_Score15	-53.08	16.960	-3.1300	0.0017	-86.32	-19.84
Tribal_CommunityNon-Tribal	74.23	98.129	0.7565	0.4494	-118.10	266.57
Tribal_CommunityPartial Tribal	-120.60	110.944	-1.0871	0.2770	-338.06	96.85
Rural_Score5	-157.67	16.948	-9.3033	0.0000	-190.89	-124.45
hospital_subtypeCritical Access Hospitals	-290.33	162.814	-1.7832	0.0746	-609.45	28.79
hospital_subtypeLong Term	-352.99	11.476	-30.7574	0.0000	-375.48	-330.49
hospital_subtypeShort Term	-27.17	7.781	-3.4912	0.0005	-42.42	-11.91

Apart from the tribal community, critical access level of hospital subtype, and is_metro_micro predictors, all other terms have strong evidence of significance, as suggested by the small p-values (reasonably less than 5%) and the confidence intervals (not including 0).

The next subsections present detailed analyses conducted on the initial best model for reported in Table 4.

2.2.1 Further analysis on the selected model

At this stage, we present results obtained from the residual diagnostics performed to evaluate model assumptions.

From Figure 6 the residual versus fitted plot suggests serious violation of the equal error variance assumption. There also appears to be a non-linear relationship with the presence of high outlying observations. Moreover, the normal probability plot departs substantially from a linear trend, showing that the distribution of the error terms is not normal. As a remedial measure to help correct most of the violations identified, we fitted two models, one with a log transformation of the response variable, inpatient_beds_7_day_avg, denote this by Model A, and another denoted Model B involving a simultaneous log transformation of the response and LIA_CS_PP, the only numeric predictor variable. Residual plots are presented in Figures 7 and 8 for the respective models with not much observable differences. We can observe a dramatic improvement in all the plots with respect to violations of model assumptions. There is a fairly linear relationship and the residuals are also fairly normal. However, there appears to be unequal variance of the residuals as the variances decrease to the left and some outliers still remain. Table 5 presents numerical test for confirmation or otherwise of the nonconstant error variance.

Figure 6: Diagnostic plots for model with initial selected model

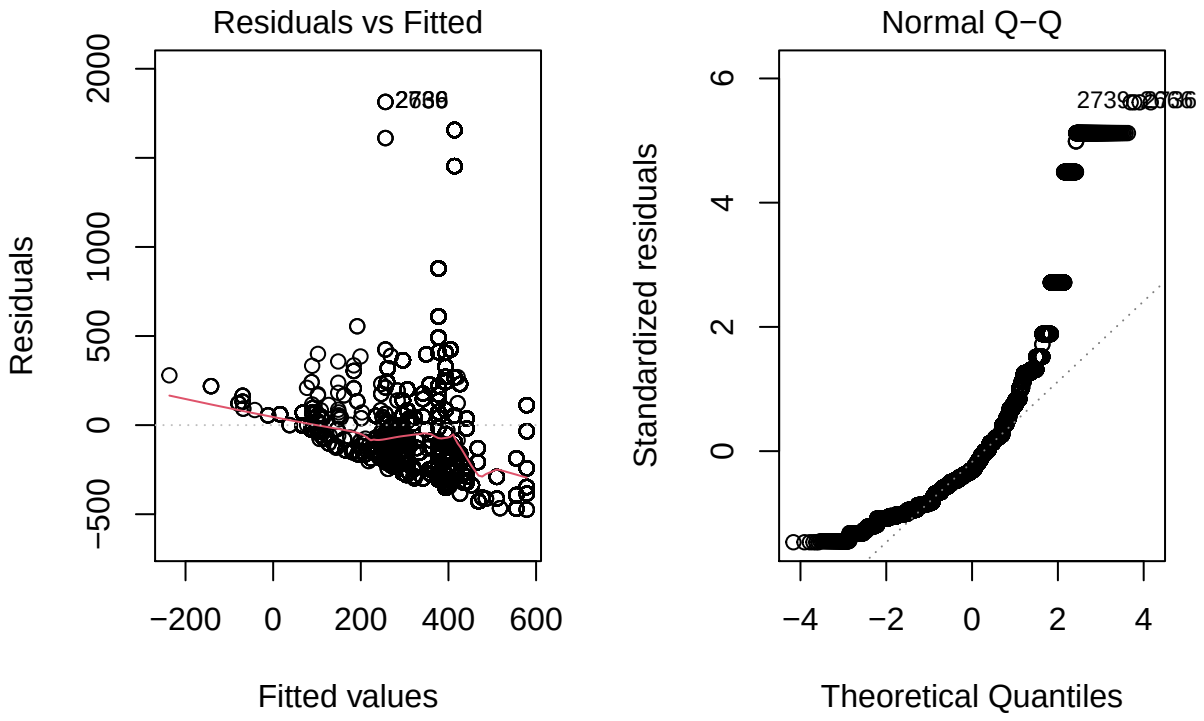


Figure 7: Diagnostic plots for model with log transformation of the response variable

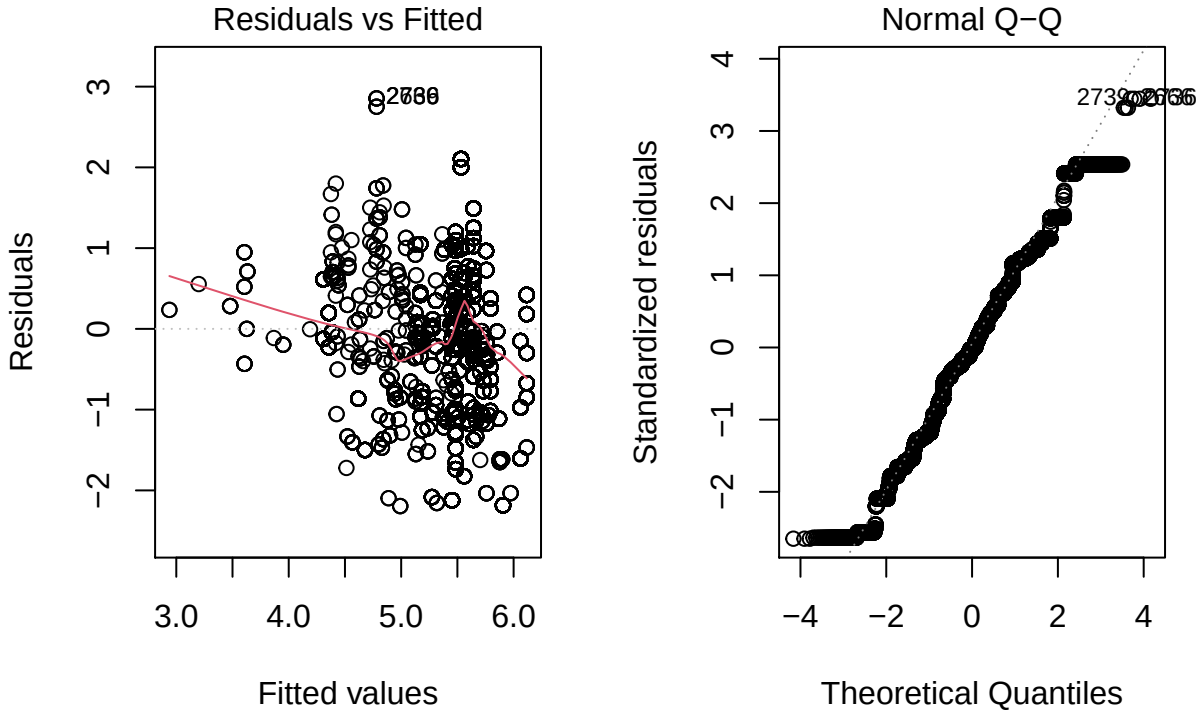


Figure 8: Diagnostic plots for model with simultaneous log transformation of the response variable and LIA_CS_PP

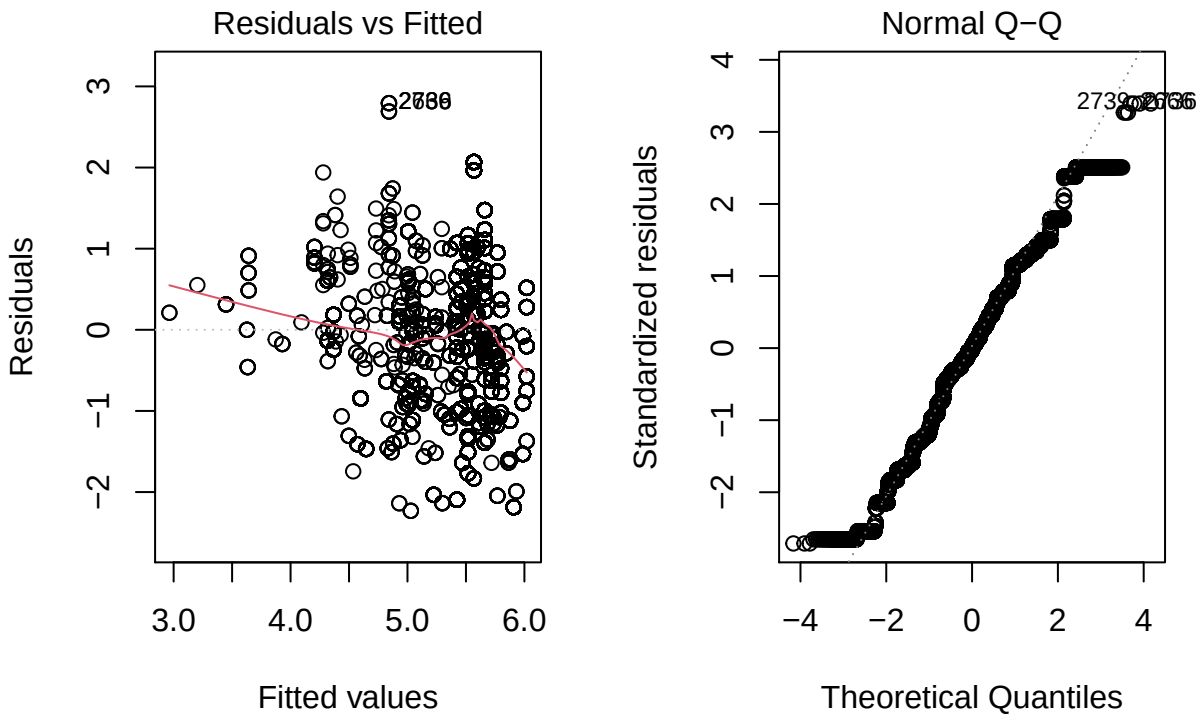
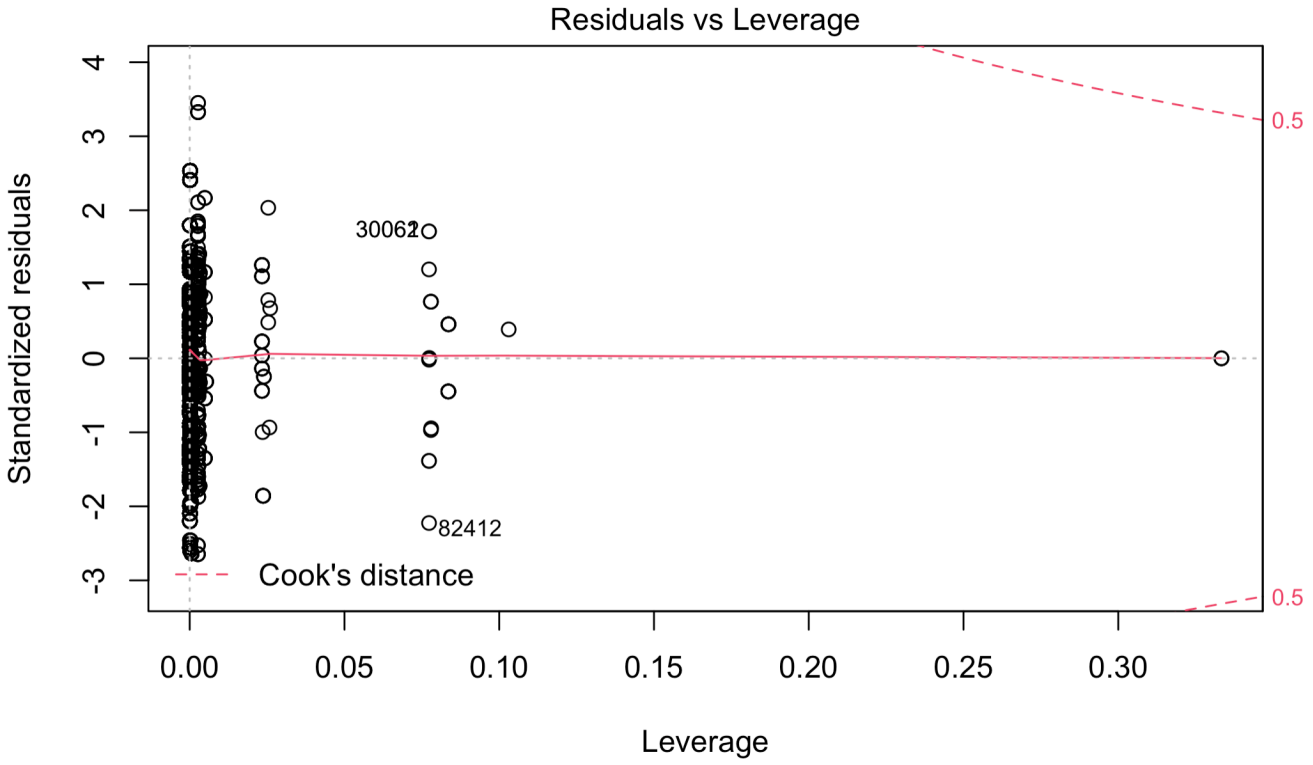


Table 5: Breusch-Pagan test for noncanstant error variance

	Test statistic	Degrees of freedom	P-value
Model A	1590	10	0
Model B	1782	10	0

The null hypothesis for the Breusch-Pagan test states that the error variance is constant, while the alternative says otherwise. Hence, at 5% significance level, the results in table 5 confirm that the constant error variance assumption is not satisfied. Note however that the test statistic increased for the simultaneous log transformed model, suggesting that the single log transformation of the response is probably better.

With regard to outlying observations, we do see that some observations are beyond 3 standard deviations of the residuals. We therefore computed the DFFITS, Coook's Distance and DFBETAS to determine whether such observations are influential or not. All measures suggested that those outlying observations are not influential, that is their exclusion do not cause substantial changes in the fitted model. Figure 9 also provides evidence for this observation as none of the points fall above or below the Cook's distance (dotted red lines).



Based on our analyses up to this point, we settled on the model with a transformed response variable as a function of LIA_CS_PP, is_metro_micro, HHA_Score, Tribal_Community, Rural_Score, and hospital_subtype. Parameter estimates of the chosen model and model performance metrics are presented in Tables 6 and 7, respectively.

Table 6: Parameter estimates of the final model

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	3.5485	0.3487	10.177	0.0000	2.8651	4.2319
LIA_CS_PP	3.9791	0.1442	27.601	0.0000	3.6966	4.2617
is_metro_microtrue	1.2340	0.2349	5.254	0.0000	0.7737	1.6944
HHA_Score7	-0.1189	0.0103	-11.589	0.0000	-0.1390	-0.0988
HHA_Score15	-0.1124	0.0435	-2.583	0.0098	-0.1976	-0.0271
Tribal_CommunityNon-Tribal	0.3749	0.2516	1.490	0.1362	-0.1182	0.8681
Tribal_CommunityPartial Tribal	-0.2932	0.2845	-1.031	0.3027	-0.8508	0.2644
Rural_Score5	-0.7511	0.0435	-17.283	0.0000	-0.8363	-0.6659
hospital_subtypeCritical Access Hospitals	-1.5673	0.4175	-3.754	0.0002	-2.3856	-0.7490
hospital_subtypeLong Term	-1.2871	0.0294	-43.736	0.0000	-1.3448	-1.2294
hospital_subtypeShort Term	-0.1113	0.0200	-5.576	0.0000	-0.1504	-0.0721

Table 7: Parameter estimates of the final model

r.squared	adj.r.squared	MSE	F.statistic	p.value
0.1213	0.121	0.6879	442.4	0

2.3 Validation of the Final Model

To valid the final trained regression model against new data, we first reestimated the final selected model on the testing set described previously. Results comparing how the reestimated model performed against the chosen model fitted to the training data is presented in Table 8.

Table 8: Comparing results of the final model fitted to the training and testing data sets

Model	r.squared	adj.r.squared	sigma	statistic	p.value
Final model	0.1213	0.121	0.8294	442.4	0
Validated model	0.1226	0.122	0.8258	191.8	0

We observe from the above results that the models for the train and test data have approximately the same r.squared, adj.r.squared, sigma. Hence we have consistency between the two models results. Thus, the results provide strong support that the chosen regression model is applicable under broader circumstances than those related to the original data. Also, we found out that the parameter estimates for the two models compared very well.

Additionally, we computed the mean squared prediction error (MSPR) using the formula in section 1.9.5 and obtained a value of **0.682**. Since the value of MSPR is approximately the same as the value for MSE of **0.6879** from the trained model, it implies that MSE for the selected regression model is not seriously biased and gives an appropriate indication of the predictive ability or power of the model.

3 Discussion

3.1 Key results

The study suggested that a multiple linear regression model involving a log transform of the response variable **inpatient_beds_7_day_avg** with five community vulnerability measures, **Low Income Area LIA County SAIPE Poverty Percentage**, **LIA_CS_PP**, **is_metro_micro**, **HHA_Score**, **Tribal_Community**, **Rural_Score**, and **hospital_subtype**, provides the best model for describing how hospital capacity is associated with community vulnerabilities. All model assumptions required were found to be fairly satisfied. The residual plot revealed some potential outliers whose influence on our models were found not to be too impactful on the performance of the final model.

We validated our final model on a holdout data set and found that the results provided strong support

that the chosen regression model is applicable under broader circumstances than those related to the original data.

3.2 Limitations

We want to formally put on record that the findings in this report 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.
- We could not control for possible confounding likely to result from the different cities where the various hospitals are located since there were too many cities leading to overfitting. Future studies can consider ways of collapsing the cities into meaningful subgroups of smaller size.
- Lack of adequate previous research studies on the topic.
- The study only focused on the averages of the hospital capacity measures when the raw values could have been used.

3.3 Interpretation

We observed that at 5% significance level there is sufficient evidence that the slope coefficients, with the exception of the coefficients associated with the **tribal community** variable, and the intercept for our final model are statistically significant. Changes in `Low Income Area LIA County SAIPE Poverty Percentage`, `is_metro_micro`, and `tribal community` have positive influence on hospital bed capacity, while `hardest hit area score`, `rural score` and `hospital subtype` affect hospital bed capacity negatively.

Also, evaluating the model on a testing data yielded $R^2 = 0.1226$ and adjusted $R^2 = 0.122$. Thus, we see that about 12% of the variation in `inpatient_beds_7_day_avg` is accounted for by the regression model containing the five community vulnerability measures.

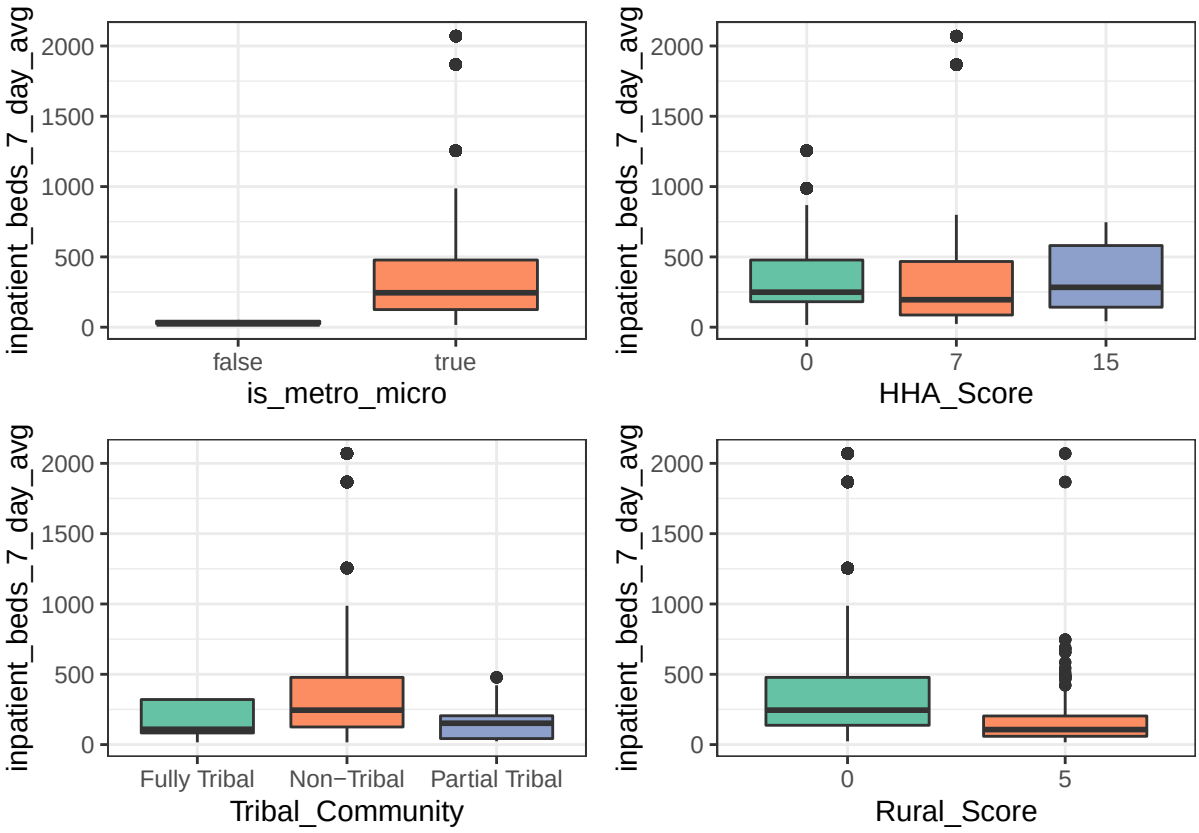
In conclusion, our findings show that hospital capacity is fairly associated with community vulnerability measures.

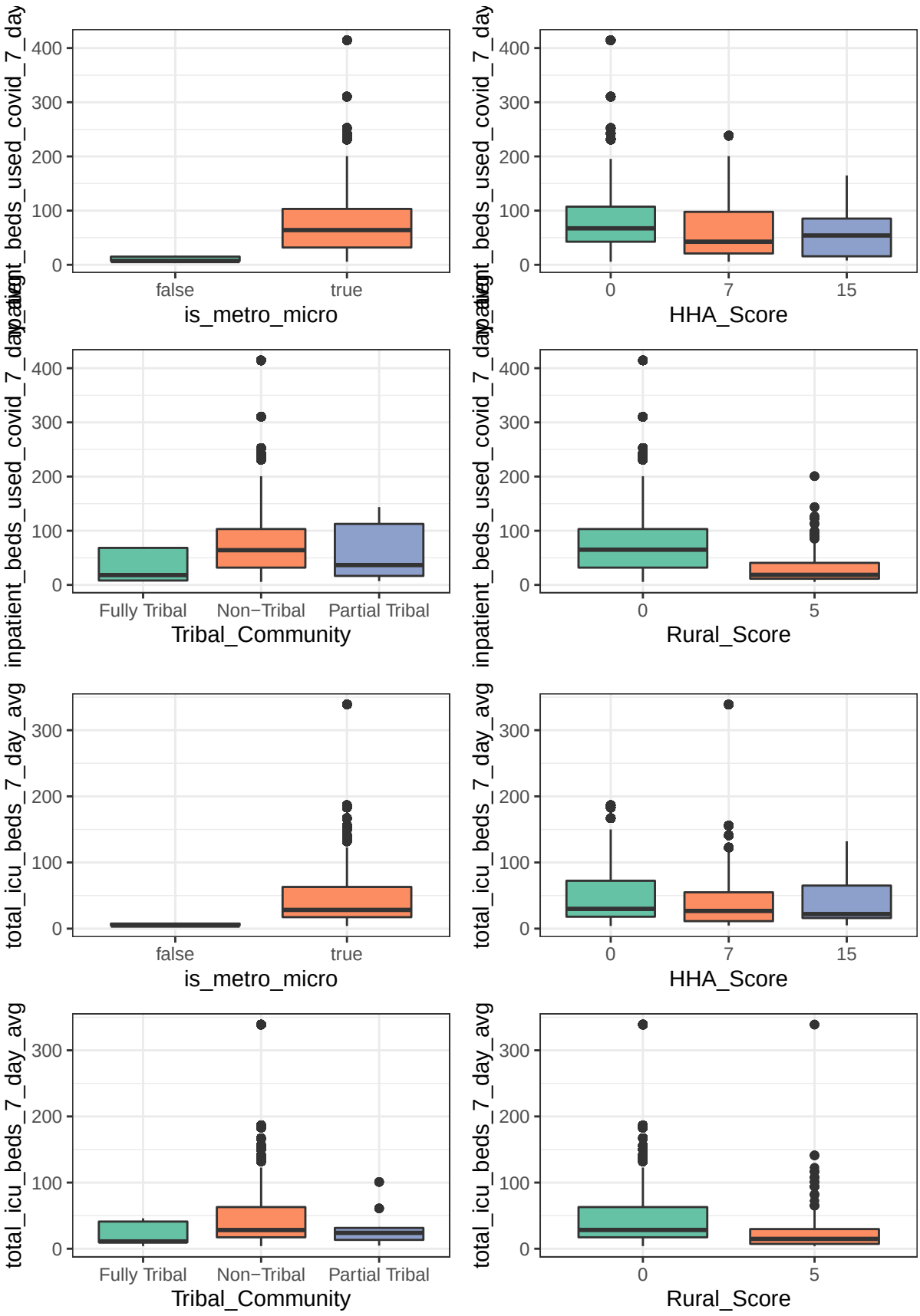
3.4 Generalisability

We found out that there was consistency of the results for the final and validated model. Indeed, our final model exhibited both internal and external validity. Thus, to a great extent our final model has great predictive ability and generalizes well to the population or data under consideration. However, the fact that the study was limited to only the State of Texas also means that, though we had a great final model, our results may not fit well for a data for the entire US for decision making that will affect the whole nation.

4 Appendix

4.1 A: Effect of categorical independent variables on the dependent variables

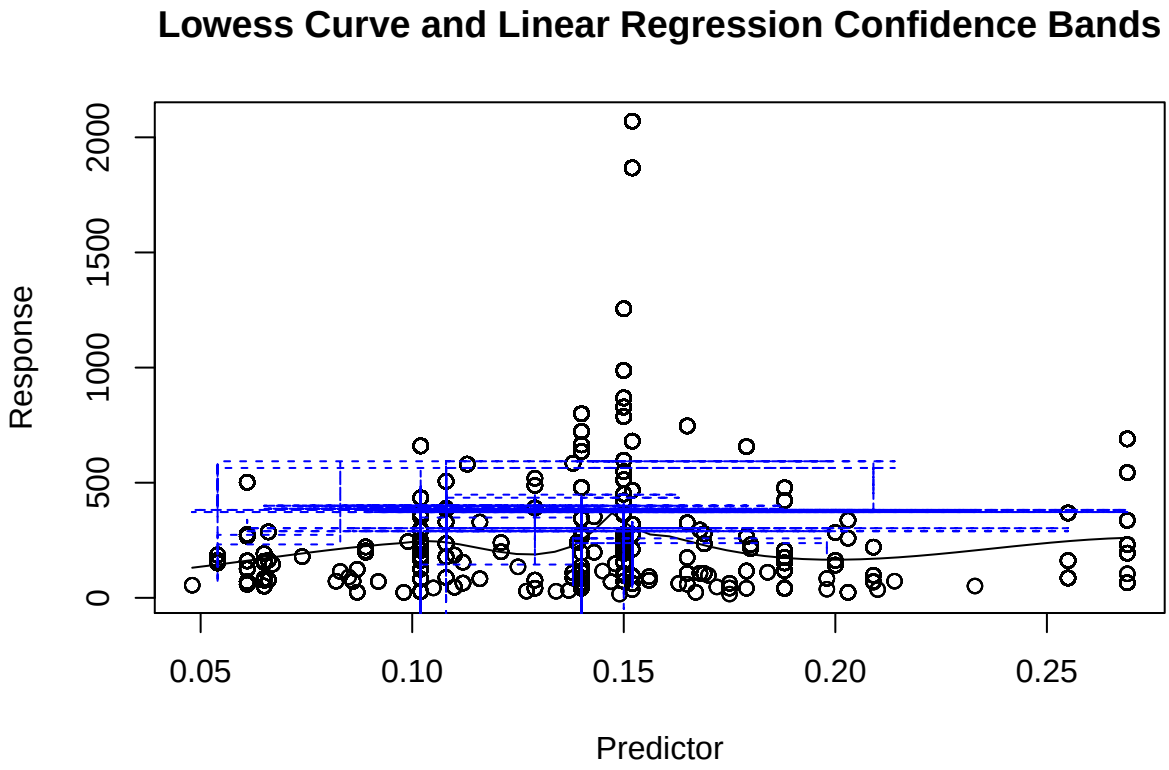


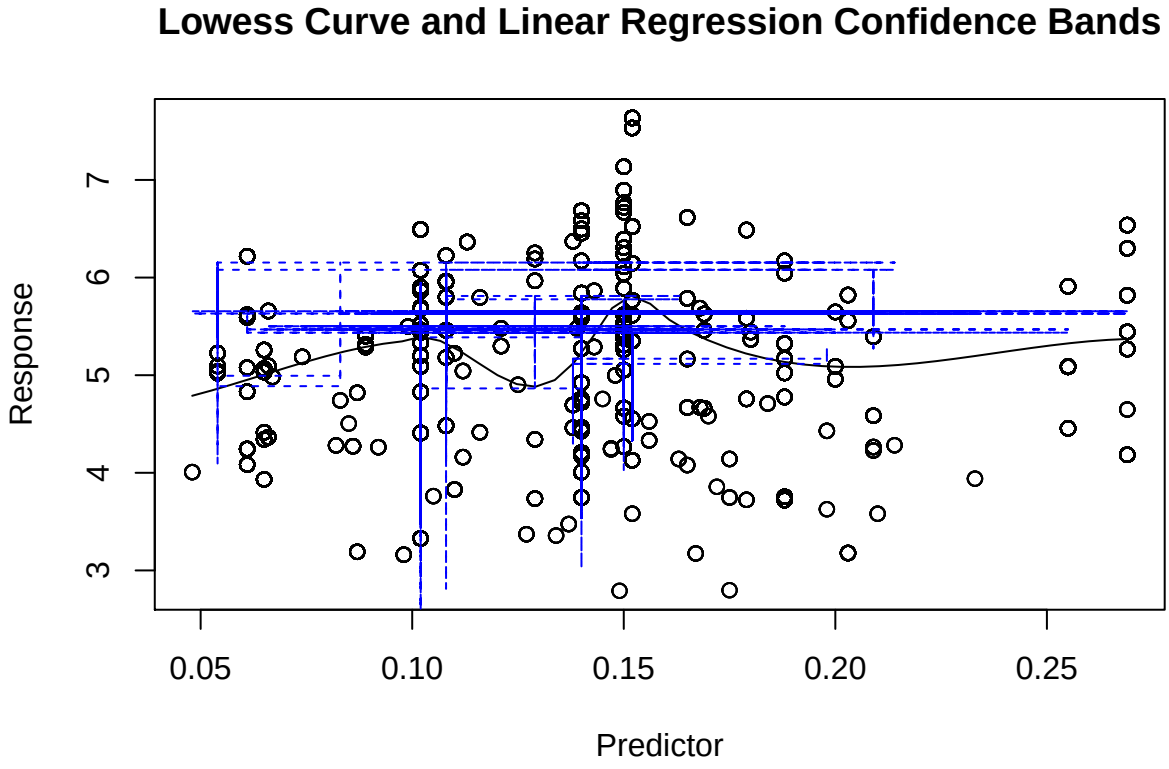

inpatient_beds_used_covid_7_day_avg
total_icu_beds_7_day_avg
0
100
200
300
400
false
true
is_metro_micro
0
7
15
HHA_Score
Fully Tribal
Non−Tribal
Partial Tribal
Tribal_Community
0
5
Rural_Score

4.2 B

```
##
##  Pearson's product-moment correlation
##
## data:  corr[, 3] and corr[, 1]
## t = 361, df = 32059, p-value <2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8938 0.8981
## sample estimates:
##    cor
## 0.8959
```

```
##
##  Pearson's product-moment correlation
##
## data:  corr[, 3] and corr[, 1]
## t = 2010, df = 32059, p-value <2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9960 0.9961
## sample estimates:
##    cor
## 0.9961
```





5 References

- Kutner, M. H., Nachtsheim, C. J., Neter, J., & Wasserman, W. (2004). Applied linear regression models (Vol. 4). New York: McGraw-Hill/Irwin.

- Hospital capacity data source: https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/anag-cw7u
- Community Vulnerability measures data source: https://healthdata.gov/Health/COVID-19-Community-Vulnerability-Crosswalk-Crosswa/x2y5-9muu
- 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