**Hospital Bed Usage during the Covid-19 Pandemic in the United States**

Fahim Karim

Ardita Elmazi

Xinyuan Yang

Dan Prengel

**Business Case**: The Covid-19 pandemic has shown that many hospitals across the Unites States were not prepared for an influx of patients and provide the needed services. Covid-19 case projection data shows that there was more demand for hospital beds than available beds, especially during the peak of pandemic. In the new future, how can we ensure the available beds in hospitals meet the demand?

**Business Questions**: How many hospital beds are needed across the United States should a pandemic hit in the future to minimize shortages and enable hospitals to treat their patients? When is the peak of the season that hospitals need to have more beds available considering that Covid-19 cases went up during spring and fall of 2020?

**Analytics Question**: What is the number of beds that the hospital needs to have available to treat future patients by analyzing the total number of beds available, the number of beds used for covid-19 cases and the number of positive cases from year 2020?

To help achieve that, this predictive analytic case will use the current data about hospitalization and predict the potential number of beds that might be needed in the upcoming weeks.

**Data**

Our data combines two datasets together. One dataset comes from HealthData.gov website that contains over 21,400,000 entries on the hospitals’ beds utilization across different states with a timeline from March 2020 to March 2021. After we cleaned the data and removed missing data points, there are 272,937 data remaining for our analysis. The other dataset is from John Hopkins website, which contains COVID-19 cases for each state across a similar timeline as our first dataset. We combined the two together by each state, so that we know the positive case numbers on that specific date. Since our aim is to predict the number of beds needed in a hospital based on available beds, inpatient beds used and active Covid-19 cases, we identified the number of beds that are in-use by a Covid at each hospital as our dependent variable. Independent variables include the number of total beds in a hospital, the number of beds used and active Covid-19 cases. All of our data is at the state level and we are unable to make distinctions between urban and rural areas in a given state. Covid-19 numbers and bed data is given daily over the course of the pandemic (3/23/2020 through 3/27/2021).

We count inpatient bed usage if a patient has occupied a bed at least one night for an examination or for treatment. Inpatient bed usage by Covid patients is just that, and includes any patient hospitalized for Covid-19 or suspected Covid-19 for one or more nights. New positive cases of Covid-19 are reported daily from each state to the CDC.

Our variables are all quantitative and take numerical values from 1 to thousands.

**Data overview**

From descriptive analysis (Appendix, Descriptive Analytics 1) and the ANOVA (Appendix, Descriptive

Analytics 2) test results, we found that inpatient used for covid has a strong positive correlation with

the number of new cases, inpatient beds number and used inpatient beds. For the state variable, we run

another ANOVA test on this binary variable, the box plot (Appendix, Descriptive Analytics 3) shows that

the difference in inpatient beds usage among states is quite significant.

In order to predict inpatient bed used for covid, we have to know how many positive cases are reported, the

total number of inpatient beds, and how many inpatient beds are used already. We will also include state

and date in our model as a binary variable and time series variable. So we will have to do variable

transformations later.

**Models**

Our project starts with a simple OLS regression:

covid\_bed\_usage = (B) + (B)new\_covid\_cases + (B)beds\_available + error

Analyzing this model and the output, we can then begin to plan for adjustments, including using time/date

information, state, and addressing any shrinkage.

**Initial Results**

In our initial observations, we see that our dependent variable (Y) is not continuous, has severe

multicollinearity and is heteroscedastic. Since we cannot have negative beds, our Y is not

continuous, and we will have to adjust our model appropriately. Our residuals in the QQ plot

below are not normally distributed. The histogram also shows a right skew to our data, further

showing the non-normal distribution of our data. To rectify this issue and violation of OLS

Assumption 2, we will log our variable to achieve a normal distribution.

We also run into multicollinearity, violating Assumption 3. The VIF for inpatent\_beds and

inpatient\_beds\_used is higher than 10 and the VIF for new\_positive\_cases\_reported is lower than 10.

To correct for this we will need to utilize RIDGE or LASSO in our model. Assumption 4 is also violated,

and our data does not move linearly. In our follow-up analysis we will utilize nonlinear models and

evaluation methods.

Our data also runs into problems from Assumption 5, as daily data is contingent on the previous day(s).

To correct, we will lag our data to attempt to control for this influence.

OLS Assumptions:

**Assumption no.1** :Y is continuous- numerical

**Assumption no2.**: Errors are normally distributed.

From the QQ plot we can see that the data depart from the line and therefore the residuals deviates from the

line at the tail more than 70% of the data. Therefore the residuals are not normally distributed. In this case we

have to transform the outcome variable to Log (y) to achieve normal distribution. Also our outcome variable

doesn't appear to be normally distributed as there is no bell shape in the histogram and shows that it is right

skewed

**Assumption no.3** : Predictors are independent.

There is a severe multicollinearity and the predictors are correlated. VIF for inpatient\_beds and inpatient\_

beds\_used is higher than 10 while the VIF for new\_positive\_cases\_reported is lower than 10. To solve this we

need to try other models such as RIDGE or LASSO.

**Assumption no4**: X’s and Y have linear relationship.

From the scatter plot, we can see that there is no linear relationship between X’s and Y. They have a nonlinear

Relationship. The correction for this is to use a nonlinear model.

**Assumption no.5**: Observations are independent and the data does not influence each other.

**Assumption no.6**: Errors are independent.

From the Durbin Watson test we can see that there is correlation between variables and errors.

The p-value of the residual regression is significant, P<0.05 then the errors are correlated

with the predicted values and heteroscedasticity is present.

If the errors are correlated then there is a missing variable on the model. When 2 predictors are correlated

and you omit one, it causes the included variable to be biased.

**Assumption no7:** Errors average is 0.

**Assumption no.8**: Errors variance is constant.

From the residual plot, we can see that the errors are not constant but create a pattern therefore there is

Heteroscedasticity. If the residuals are uneven, this cause the observations with large errors to pull

the regression strongly in both directions, thus increasing variance.

We now opt to do a shrinkage model like Ridge or LASSO to fight multicollinearity:

After performing Ridge Regression and LASSO and finding our lambda for both,

we have the choice to choose the best lambda for our predict model or

we can use a Net Elastic Regression model where we take weighted values from both Ridge or

LASSO and give us output using this hybrid model.

**Conclusions**

Predicting the number of beds a hospital may need for Covid-19 or the next pandemic to hit The

United States is a difficult task. As seen above, our current data has many issues to prevent us from

using a simple OLS model, and as we begin to dive deeper into this project we must complicate

our model in order to craft a useful and accurate prediction of needed hospital beds. In fine-tuning

our model, we may find more success expanding our model to include factors outside of covid cases

and available hospital beds. Since Covid-19 has been very politicized in the US, we may find success

adding a political variable to help determine how the citizens of a given state may be reacting to the

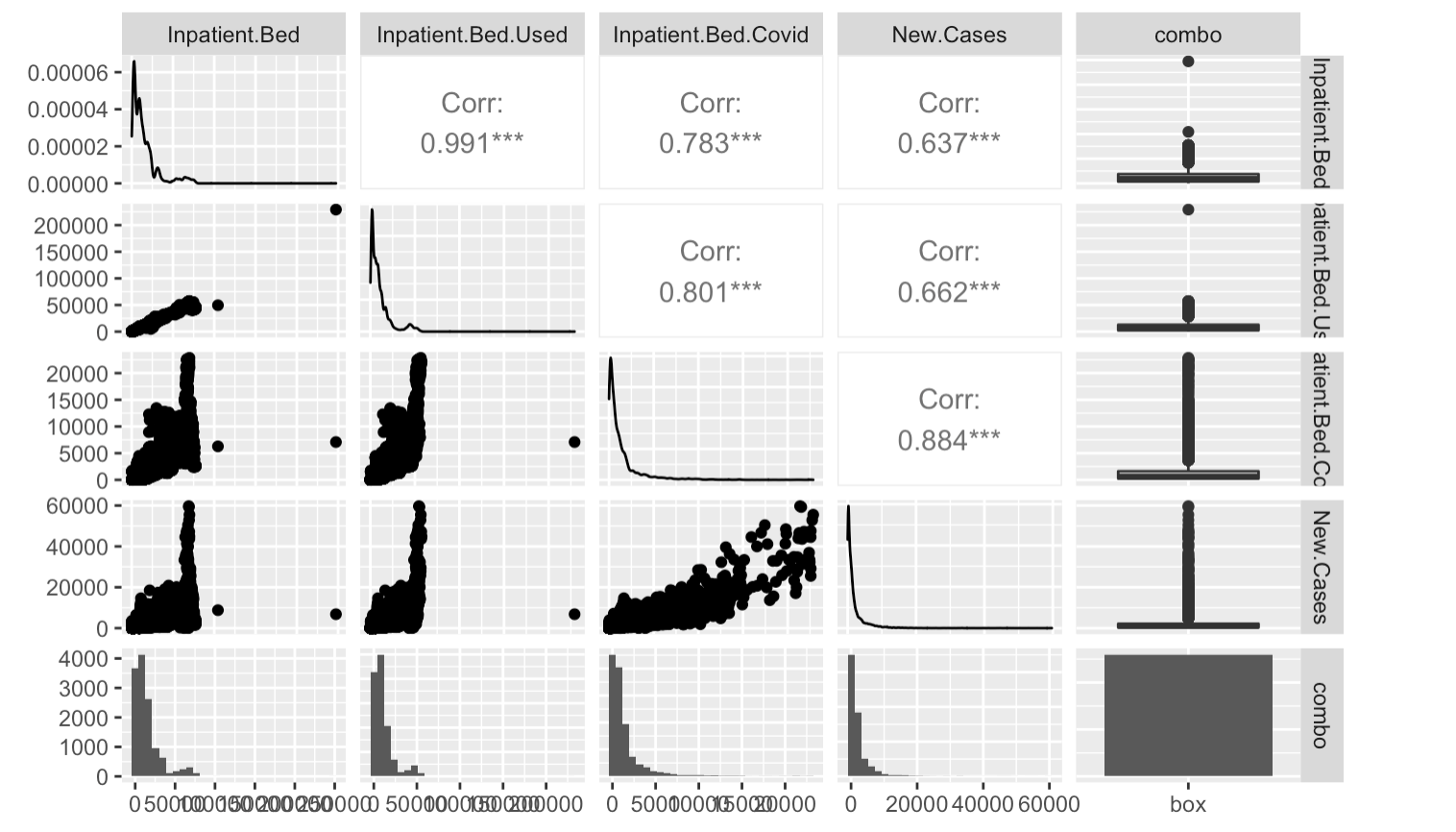
pandemic (are they wearing masks? Is the state “open”?). We may also find more success in including

demographic information, as previous findings show older residents, people of color, and young

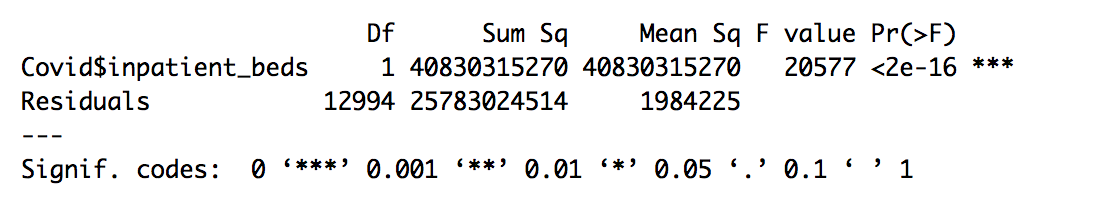
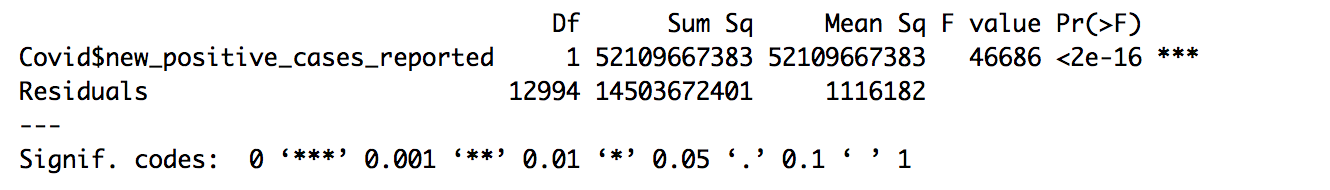
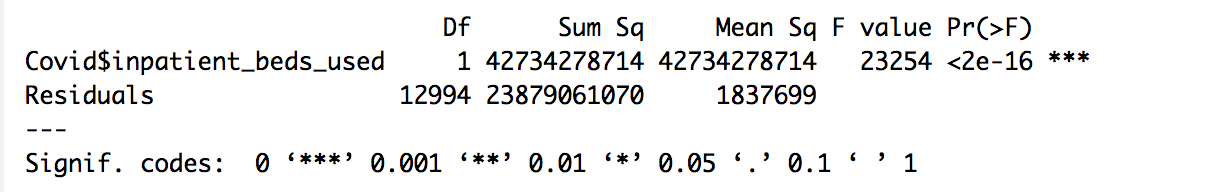
children all have very different responses to Covid-19 and the need to be hospitalized. This analysis

provides a great starting point for future research.

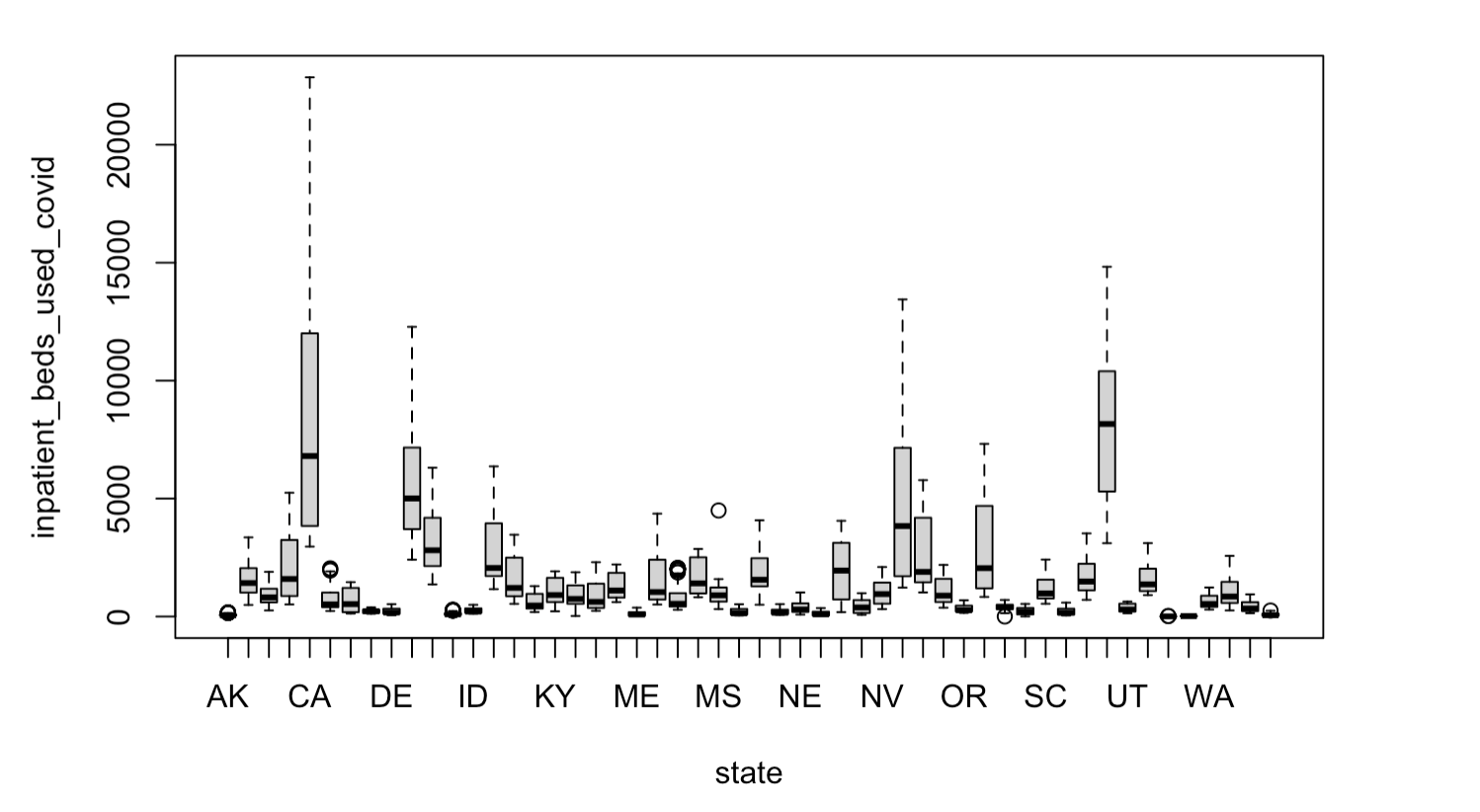
Appendix:

Descriptive Analytics 1

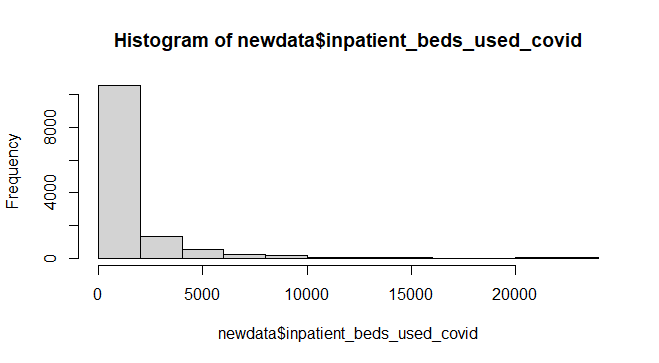
Descriptive Analytics 2



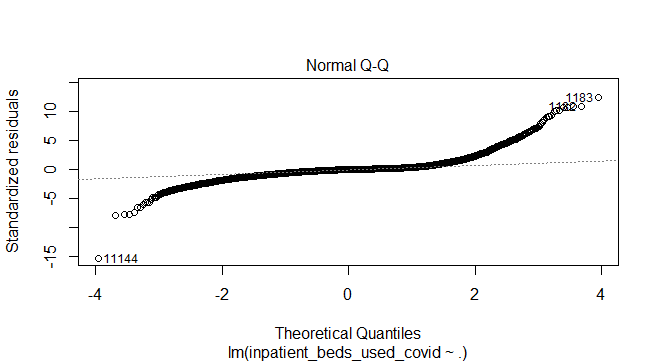
Descriptive Analytics 3



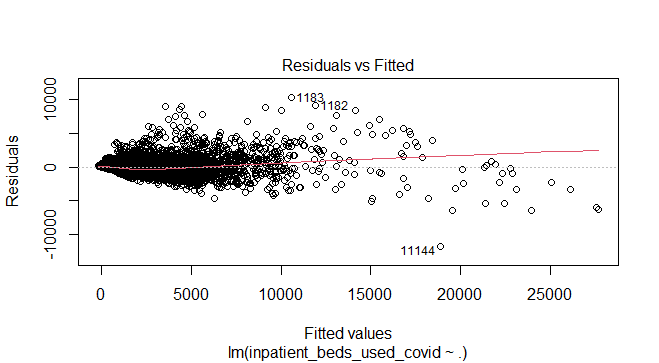
Assumption 1



Assumption 2:



Assumption 3:

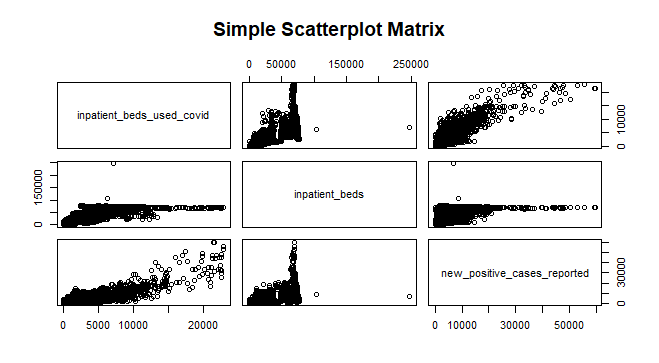


vif(lm.ols)

#inpatient\_beds inpatient\_beds\_used new\_positive\_cases\_reported

57.880474 61.333787 1.853703

Assumption 4:



Assumption 5: DW Test outcome

durbinWatsonTest(lm.ols)

lag Autocorrelation D-W Statistic p-value

1 0.8783205 0.2433569 0

Alternative hypothesis: rho != 0

Ridge Regression:

The best lambda is 200.2416

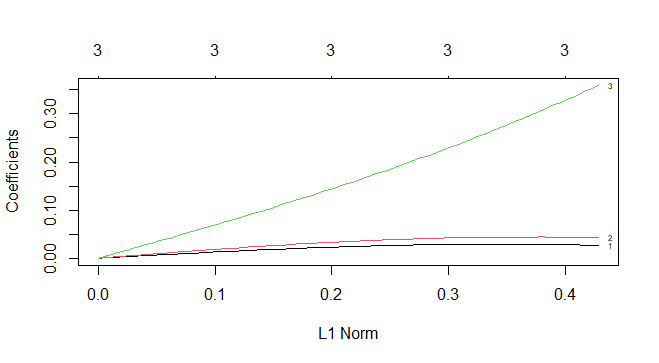
Ridge Regression coefficients are :

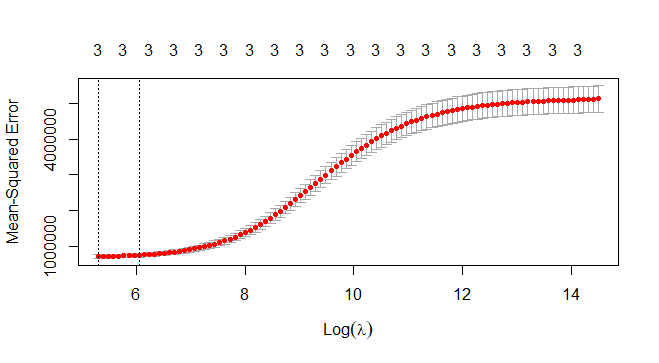
(Intercept) -135.59905658

inpatient\_beds 0.02637977

inpatient\_beds\_used 0.04351414

new\_positive\_cases\_reported 0.35882752





For LASSO Model:

The best Lambda is 12.00417

(Intercept) -151.86901304

inpatient\_beds 0.01031134

inpatient\_beds\_used 0.06053998

new\_positive\_cases\_reported 0.39636543

Best Lambda Log(Lambda) Best 10FCV MSE

12.00417 2.485254 697508.3

