Capstone: Linear Regression Analysis and Prediction of Hospital Capacity in Idaho and Oregon

Steven Duchene

Colorado State University Global

MIS 480: Business Analytics and Information Systems Capstone

Justin Bateh

February 7, 2021

Capstone: Linear Regression Analysis and Prediction Hospital Capacity in Idaho and Oregon

Over the years, the processes of business intelligence have transformed from the simple analyses of small quantities of data to the analysis of immense quantities of data that were once thought to be too large and complex to be investigated by traditional data-processing software. Industries such as telecommunications use business intelligence processes to provide improved customer service and problem solving while the retail sector uses data analytics to provide personalized customer experiences, improve in store experiences, and even track the amount of time customers spend in specific sections of the store. However, data analytics is not limited to use cases where the goal is to improve customer experiences or increase revenues. Sectors including finance, insurance, fitness, and government are also taking advantage of the over two and a half quintillion bytes of data generated by humans daily (Belfiore, 2019).

One industry not often thought of in the context of big data and data analytics is that of the healthcare sector. Healthcare organizations can extract key information and intelligence through the analysis of healthcare data sources such as physician entries, visit reasons, laboratory data, diagnoses, emergency room visits, electronic health records, and a plethora of other data sources; uncovering new trends, patterns, and other information. This information can be used for several different purposes including determination of premium rates, disease trends, care discrepancies, and most importantly, improving patient outcomes for both chronic disease and emergent conditions to improve both treatment and mortality. Given the pandemic that we are currently experiencing, the healthcare systems ability to be prepared for both COVID related and normal admissions is of critical importance; the consequence of being ill prepared including healthcare systems becoming overwhelmed. Two methods through which healthcare predications can be made include linear regression and ARIMA (Auto Regressive Moving Average).

Fortunately, datasets related to both SARS-CoV-2 and hospital admissions are readily available for analysis. Cloud Service providers such as Amazon Web Services and Google Cloud have compiled and made available a plethora of datasets related to the COVID-19 pandemic. Through a rudimentary review of both Google Cloud COVID-19 Public Dataset Program and the Amazon Web Services public data lake, over 30 different datasets were available for analysis. Additionally, government websites such as healthdata.gov, CDC.gov, and HHS.gov all provide additional datasets related to the pandemic. Analyses conducted as part of this research utilized the COVID-19 Hospital Data Coverage Report obtained from the healthdata.gov website (HealthData.gov, 2021). The dataset is updated daily and the data for this research was obtained on January 18, 2021. The following sections of this report will include a series of analyses, graphs, tables, and predictions conducted using the Hospital Data Coverage Report for the states of Idaho and Oregon. Additionally, it will provide details with respect to the organizations selected, how the data was accessed and stored, and the business intelligence tool utilized. It will also provide details with respect to the respective coding used during this research. The coding utilized throughout this project has also been uploaded into GitHub and can be accessed using the following link: https://github.com/sduchene-208.

Organizational Information

The organizations selected for which the following analyses will apply include the hospitals located in the States of Idaho and Oregon, which includes over 110 hospitals and medical centers. Given that healthcare capacity is a critical factor in an area being able to sufficiently handle both the pandemic conditions as well as standard hospital inflow, the health systems located in these states were selected in order to provide a predictive toolset, for a specifical geographic area, to ensure that the hospital systems remains operational and that effective healthcare is provided to the citizens of Idaho and Oregon.

Dataset, Storage Methods, and Rationale for Selection

The dataset selected for this project was obtained from the healthdata.gov website. Specifically, the dataset contains data related to hospital coverage and associated measurements. A data dictionary related to this dataset can be obtained from https://healthdata.gov/covid-19-reported-patient-impact-and-hospital-capacity-state-data-dictionary (HealthData.gov, 2020). Initially, the dataset contained information related to the entire United States. For the purpose of making this research applicable to a specific geographical area, one of which sees continuous travel between the states, the dataset was modified to include only Idaho and Oregon datapoints. The rationale for selecting this dataset was two-fold. Firstly, the dataset provides critical information related to hospital capacities. As the pandemic shows no signs of slowing at this time, using data that includes these factors may allow for predictive capabilities related to hospital capacities. Secondly, this dataset provides information specific to geographical areas; allowing the researcher to conduct analyses specific to areas of concern and interest.

The data was stored in a PostgreSQL database termed COVID_DATA and contained three tables: Idaho_oregon_combined, idaho_data, and oregon_data. Figure 1 demonstrates the coding utilized for the

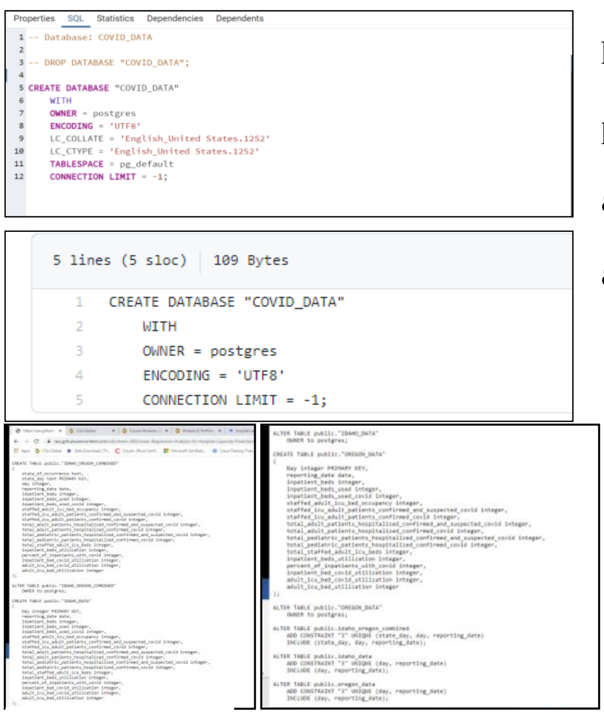

Figure 1: Database

creation of the database and associated tables. Transformations and loading of the database were completed using Pentaho. The data dictionary detailing each variable has been loaded into GitHub.

Problem Statement

The COVID-19 pandemic, in its current form, continues to spread virtually uncontrolled, resulting in hospitals and health systems becoming inundated with critically ill patients. As coronavirus infections, and associated hospitalizations continue an upward trajectory, governments and health systems around the nation are working to address multiple different public health and epidemiological concerns. As the pandemic is, for lack of a better term, unprecedented, there is an absence of historical data available to utilize with respect to the conduction of predictive and prescriptive analytics. This study seeks to improve the predictive capabilities for the 2021-22 through a critical analysis of Hospital Coverage for the years 2020-21, beginning with the discovery dates of SARS-CoV-2 in the states of Idaho and Oregon. The analyses were limited to five critical variables related to hospital utilization.

Business Intelligence Tool Selection

The business intelligence tool selected for this research is SAS Studio. Models including, linear regression and others will be utilized throughout this research project. Additionally, the SAS Viya platform was also utilized in order to create more visually appealing illustrations as well as conduct additional analyses. SAS Viya is a platform for AI, analytics and data processing based on a scalable, cloud-native architecture. Capabilities of both SAS Studio and SAS Viya include AI, data analysis, predictive analytics, and data management.

Analysis of Data and Data Visualizations

When presenting data, data visualizations provide a way for the human brain to process complex information more rapidly than simply looking over spreadsheets and raw data. As the data visualization is a visual representation of the data in question, choosing the right visualizations will allow for the researcher to identify trends patterns within their data; identifying areas that need improvement or factors that have a direct influence on the results. As such, data visualizations are a powerful method to present compelling data (Qin, Luo, Tang, &

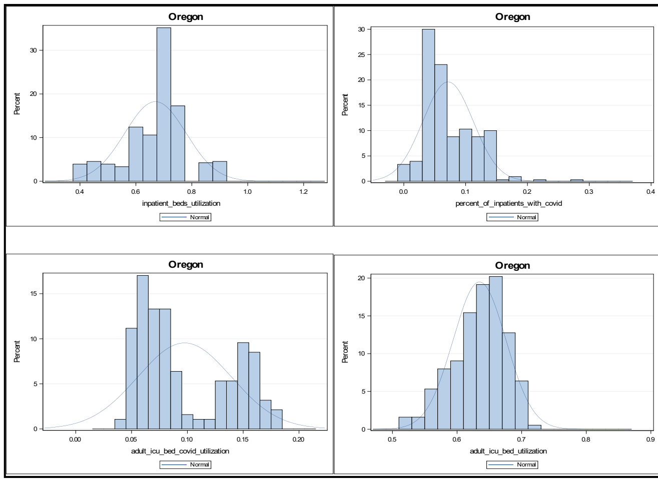

Figure 2: Oregon Histograms

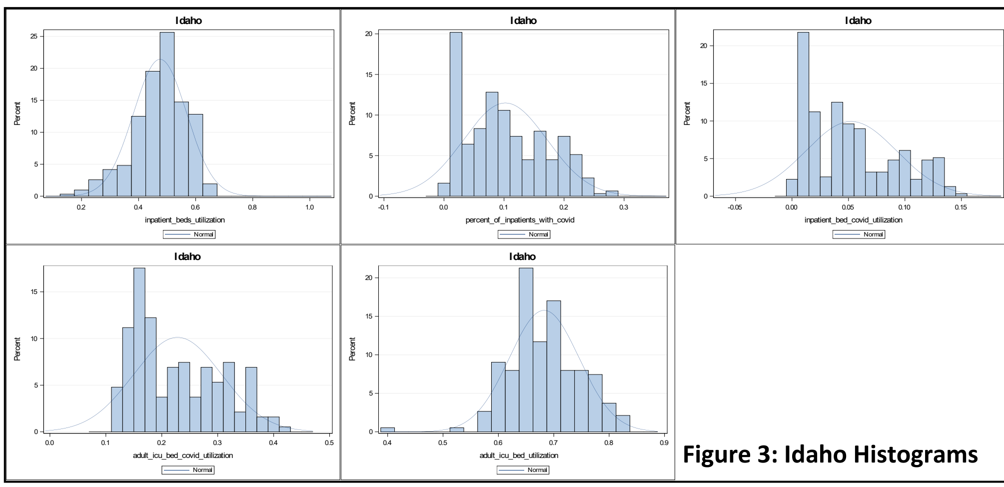

Figure 3: Idaho Histograms

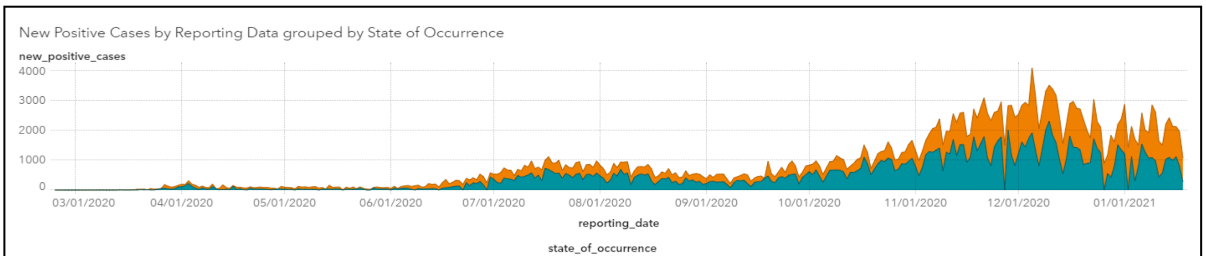

Figure 4: Positive Cases for Idaho and Oregon

Li, 2019). The data visualizations and tables are presented below. The applicable code has been uploaded to GitHub. A brief discussion will follow figures 5 and 6.

Summary Statistics for Oregon

Variable	Label	Mean	Std Dev	Minimum	Maximum	N	N Miss	Coeff of Variation	Skewness	Kurtosis
inpatient_beds_utilization	inpatient_beds_utilization	0.6698512	0.1093377	0.3948032	0.8829787	330	0	16.3226885	-0.5208436	0.4554255
percent_of_inpatients_with_covid	percent_of_inpatients_with_covid	0.0715329	0.0406955	0	0.2807018	330	0	56.8906937	0.9249708	1.4603113
inpatient_bed_covid_utilization	inpatient_bed_covid_utilization	0.0473292	0.0280542	0	0.2051282	330	0	59.2746038	1.1813596	2.4917622
adult_icu_bed_covid_utilization	adult_icu_bed_covid_utilization	0.0974689	0.0418278	0.0392157	0.1820546	188	142	42.9139849	0.5328011	-1.2739932
adult_icu_bed_utilization	adult_icu_bed_utilization	0.6347468	0.0409406	0.5149254	0.7275064	188	142	6.4499116	-0.5247197	-0.0623901

Summary Statistics for Idaho

Variable	Label	Mean	Std Dev	Minimum	Maximum	N	N Miss	Coeff of Variation	Skewness	Kurtosis
inpatient_beds_utilization	inpatient_beds_utilization	0.4752317	0.0931148	0.1600000	0.6425876	312	0	19.5935555	-0.7311983	0.4712312
percent_of_inpatients_with_covid	percent_of_inpatients_with_covid	0.1021539	0.0694710	0	0.2857143	312	0	68.0062379	0.4192678	-0.8755639
inpatient_bed_covid_utilization	inpatient_bed_covid_utilization	0.0524196	0.0401796	0	0.1468676	312	0	76.6499543	0.5839555	-0.8197299
adult_icu_bed_covid_utilization	adult_icu_bed_covid_utilization	0.2281517	0.0788030	0.1122807	0.4144737	188	124	34.5397603	0.5189020	-0.9238554
adult_icu_bed_utilization	adult_icu_bed_utilization	0.6826447	0.0631935	0.4011407	0.8371336	188	124	9.2571520	-0.1471423	1.3131476

Figure 5: Summary Statistics

Pearson Correlation Coefficients
Number of Observations

	inpatient_beds_utilization	percent_of_inpatients_with_covid	inpatient_bed_covid_utilization
day day	0.88946 312	0.69308 312	0.81867 312

Pearson Correlation Coefficients
Number of Observations

	adult_icu_bed_covid_utilization	adult_icu_bed_utilization
day day	0.77322 188	0.39187 188

Pearson Correlation Coefficients
Number of Observations

	inpatient_beds_utilization	percent_of_inpatients_with_covid	inpatient_bed_covid_utilization
day day	0.27306 330	0.34526 330	0.51091 330

Pearson Correlation Coefficients
Number of Observations

	adult_icu_bed_covid_utilization	adult_icu_bed_utilization
day day	0.85288 188	0.60457 188

Figure 6: Correlation Analysis

I. Inpatient Bed Utilization: Oregon data indicates, through analysis of the skewness value, that the data distribution is fairly symmetrical while Idaho datapoints appear to skew slightly to the left. Based on a distribution analysis conducted through SAS, Idaho appears to exhibit a Weibull-like distribution while Oregon slightly exhibits Weibull and hypergeometric characteristics.

II. Percent of IPs with COVID: Skewness analysis indicates that the Oregon Data is skewed relatively strongly to the right while the Idaho data again appears to be fairly symmetrical. Distribution analysis indicates Idaho exhibits both beta and Weibull-like distribution while Oregon seems to exhibit a Poisson-like distribution.

III. Inpatient Bed COVID Utilization: Skewness analysis indicates Oregon exhibits a strong positive skew while Idaho exhibits fairly symmetrical distribution with a slight skew to the right. Oregon appears to exhibit a log-normal distribution while Idaho appears slightly beta-binomial.

IV. Adult ICU COVID Bed Utilization: Interestingly, both Idaho and Oregon exhibited relatively symmetrical distributions. Oregon exhibits a bimodal distribution while Idaho appears to exhibit a Gamma-like distribution.

V. Adult ICU Bed Utilization: Oregon datapoints appear to be relatively symmetrical while Idaho datapoints are exhibiting a slight left skew. The Oregon distribution appears to be Weibull-like while Idaho is exhibiting both log-normal and Gamma characteristics.

Methods and Model Rationale

Prior to determining the appropriate regression method, pre-analysis activities were completed via the creation of a scatter plot matrix using the "proc sgscatter" functionality in SAS. As evidenced in figure 7, for the most part, the datapoints for both Oregon and Idaho appear to be linear in nature. As such, linear regression models were built using both the Idaho and Oregon datasets. Models were created in SAS using the "proc glmselect" and "proc reg" functionalities and the SAS coding was uploaded to GitHub. The five variables were analyzed versus the day variable.

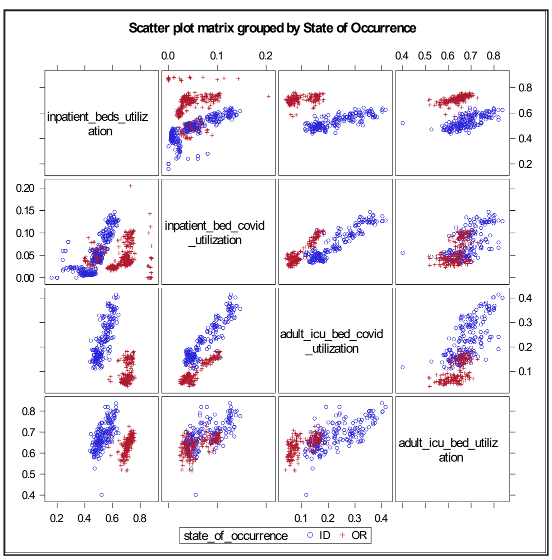

Figure 7: Scatter Plot Matrix

As mentioned above, the ability of the health systems to be prepared for increases in critically ill COVID patients as well as non-COVID patients, is of great importance. Effective management cannot be achieved without a collective effort to forecast hospital demands (Qian, Alaa, & Schaar, 2020). Linear regression was selected as the method of choice for construction of the predictive model for several reasons. Firstly, linear regression can be applied to a broad range of data types, including time series, cross-sectional, pooled, or panel data (Nyce & Cpcu, 2007). Secondly, it is a simple algorithm of which can be readily developed and readily applied. Thirdly, in contrast to some of the other machine learning algorithms, linear regression has a substantially lower time complexity (Singh, 2020). Disadvantages of the linear regression model, however, include its sensitivity to outlier datapoints and that it assumes the datapoints are independent of one another. For variables that were not correlated significantly with “Day,” an ARIMA model was utilized using the “proc arima” SAS functionality. The correlation data can be viewed in figure 8.

state_of_occurrence=ID

1 With Variables:	day
5 Variables:	inpatient_beds_utilization percent_of_inpatients_with_covid inpatient_bed_covid_utilization adult_icu_bed_covid_utilization adult_icu_bed_utilization

Pearson Correlation Coefficients
Number of Observations

	inpatient_beds_utilization	percent_of_inpatients_with_covid	inpatient_bed_covid_utilization	adult_icu_bed_covid_utilization	adult_icu_bed_utilization
day day	0.88946 312	0.69308 312	0.81867 312	0.77322 188	0.39187 188

state_of_occurrence=OR

1 With Variables:	day
5 Variables:	inpatient_beds_utilization percent_of_inpatients_with_covid inpatient_bed_covid_utilization adult_icu_bed_covid_utilization adult_icu_bed_utilization

Pearson Correlation Coefficients
Number of Observations

	inpatient_beds_utilization	percent_of_inpatients_with_covid	inpatient_bed_covid_utilization	adult_icu_bed_covid_utilization	adult_icu_bed_utilization
day day	0.27306 330	0.34526 330	0.51091 330	0.85288 188	0.60457 188

Figure 8: Correlation Analysis

Results

After conducting an analysis with respect to the variables correlated with day (figure 8), the variables highlighted by the red square exhibited strong to relatively strong correlation with the day variable and, as such, were subjected to the linear regression model. The variables highlighted in green were subjected to an ARIMA model. When determining how accurate or appropriate the ARIMA model is, there are several steps that can be taken to ensure an accurately fit ARIMA model. A major rule of thumb with respect to ARIMA models is to determine how many positive autocorrelations are associated with a high number of lags. If the series initially shows strong positive autocorrelation, application of nonseasonal differencing will reduce the amount of autocorrelation (Business, n.d.). This was the case with the ARIMA models fit below. As such, a differencing value of 1 was utilized, which provided the lowest standard deviation values. The results are presented below with respect to each state. 95% confidence limits were utilized in these analyses.

I. Idaho

a. Inpatient Bed Utilization: This model demonstrated a substantially strong capability for prediction with an R^2 value of .7911 and a P-value of less than .0001. The residual plots indicated a good fit. The equation was Inpatient Bed Utilization = .332 + .00092(day).

b. Percent of Inpatients with COVID: This model demonstrated a moderately strong capability for prediction with an R^2 value of .4804 and a P-value of less than .0001. The residual plots indicated a good fit. The equation was Percent of Inpatients with COVID = .01862 + .000534(day).

c. Inpatient Bed COVID Utilization: This model demonstrated a moderately strong capability for prediction with an R^2 value of .6702 and a P-value of less than .0001. The residual plots indicated a good fit. The equation was Inpatient Bed Utilization = -.00465 + .000365(day).

d. ICU Bed COVID Utilization: This model demonstrated a moderately strong capability for prediction with an R^2 value of .5979 and a P-value of less than .0001. The residual plots indicated a good fit. The equation was Inpatient Bed Utilization = -.01652 + .00112(day).

e. ICU Bed Utilization: Due to a weak correlation with day, this variable was subjected to an ARIMA model. It provided a 14-day forecast. The first five days are shown in figure 9.

Forecasts for variable adult_icu_bed_utilization

Obs	Forecast	Std Error	95% Confidence Limits	
313	0.6486	0.0473	0.5559	0.7414
314	0.6479	0.0669	0.5167	0.7791
315	0.6472	0.0820	0.4865	0.8078
316	0.6465	0.0946	0.4610	0.8320
317	0.6458	0.1058	0.4384	0.8532

Figure 9: ARIMA Forecast

II. Oregon

a. Inpatient Bed Utilization: Due to a weak correlation with day, an ARIMA model was fitted for this variable. The results can be seen in figure 10.

Forecasts for variable inpatient_beds_utilization				
Obs	Forecast	Std Error	95% Confidence Limits	
331	0.6985	0.0357	0.6286	0.7685
332	0.6981	0.0505	0.5991	0.7970
333	0.6976	0.0618	0.5764	0.8188
334	0.6971	0.0714	0.5571	0.8370
335	0.6966	0.0798	0.5401	0.8530

Figure 10: ARIMA

b. Percent of Inpatients with COVID: Due to a weak correlation with day, this variable was subjected to an ARIMA model. It provided a 14-day forecast. The first five days are shown in figure 11.

Forecasts for variable percent_of_inpatients_with_covid				
Obs	Forecast	Std Error	95% Confidence Limits	
331	0.0995	0.0177	0.0649	0.1341
332	0.0998	0.0250	0.0508	0.1488
333	0.1001	0.0306	0.0401	0.1601
334	0.1004	0.0353	0.0311	0.1697
335	0.1007	0.0395	0.0233	0.1781

Figure 11: ARIMA Model

c. Inpatient Bed COVID Utilization: Due to a weak correlation with day, this variable was subjected to an ARIMA model. It provided a 14-day forecast. The first five days are shown in figure 12.

Forecasts for variable inpatient_bed_covid_utilization				
Obs	Forecast	Std Error	95% Confidence Limits	
331	0.0697	0.0148	0.0406	0.0987
332	0.0699	0.0210	0.0288	0.1110
333	0.0701	0.0257	0.0198	0.1204
334	0.0703	0.0296	0.0122	0.1284
335	0.0705	0.0331	0.0056	0.1355

Figure 12: ARIMA Model

d. ICU Bed COVID Utilization: This model demonstrated a substantially strong capability for prediction with an R^2 value of .7274 and a P-value of less than .0001. The residual plots indicated a good fit. The equation was Inpatient Bed Utilization = -.05758 + .000.000654(day).

e. ICU Bed Utilization: Due to a weak correlation with day, this variable was subjected to an ARIMA model. It provided a 14-day forecast. The first five days are shown in figure 13.

Forecasts for variable adult_icu_bed_utilization				
Obs	Forecast	Std Error	95% Confidence Limits	
331	0.6335	0.0297	0.5754	0.6917
332	0.6341	0.0420	0.5519	0.7164
333	0.6348	0.0514	0.5340	0.7355
334	0.6354	0.0593	0.5191	0.7517
335	0.6360	0.0663	0.5060	0.7661

Figure 13: ARIMA Model

Discussion

Over the past few months, both Idaho and Oregon have been experiencing some of the highest infection and positivity rates for COVID infections. This has the proclivity to result in critically high hospital capacities; requiring additional resources and supplies as well as increased staffing. Essential care services, such as ICU beds, mechanical ventilation units, and medical staff, are already limited, with critically ill patients afflicted with other diseases already consuming much of the available resources (Ojan & Alaa, 2020). As this has the ability to place unprecedented stress on the healthcare system, an effective model of prediction was necessary to ensure that hospital capacity does not exceed the systems resources. As seen in the prior section, this was achieved through the use of both linear regression and autoregressive models.

Interestingly, with respect to the State of Idaho, the linear regression model exhibited strong predictive capabilities for four out of the five variables subjected to the model. The fit plots for those variables, with 95% confidence levels, are seen in figure 14. While it would appear that the linear regression model could serve as a good candidate for prediction of hospital capacities for the State of Idaho, it should not be relied upon solely as infection behaviors can be highly unpredictable and subject to both seasonal trends and community events. Contrarily, with respect to the State of Oregon, the linear regression models did not appear to be a strong candidate for prediction of hospital capacities. This was evidenced by the correlation coefficients in figure 8. As such, these variables were subjected to an ARIMA model. Interestingly, upon first applying the ARIMA model to the Oregon data, an issue occurred where all predictions were matching the predicted mean. After analysis of the initial ARIMA results, it was determined that a higher order of differencing would be required. Once the models were run with a differencing value of 1, the

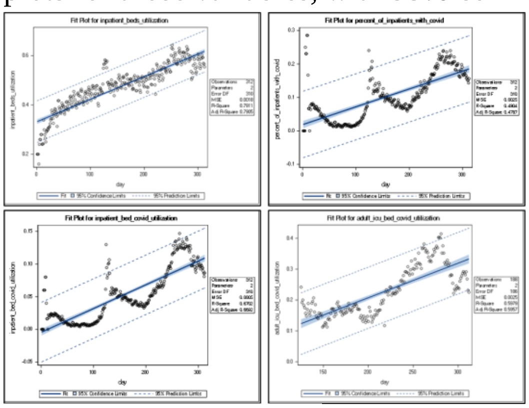

Figure 14: Fit Plots for Idaho

models appeared to be successful in their predictions. However, this does not imply that the models were without limitations.

There are several limitations with respect to this analysis. Firstly, the linear regression and ARIMA models collect input data without taking into consideration the logistic actions, or lack thereof, being taken during the process. However, the model's results were highly indicative of the expected, viral trajectory for these communities. Secondly, the datasets contained only datapoints from the beginning of the pandemic for respective communities on. As such, the datasets, and their subsequent analyses, fail to include historical data prior to the beginning of the pandemic as data was not actively collected prior to the pandemic. Were that data to have been available and used in the analyses, it may have provided for greater predictive capabilities by including data not skewed by the pandemic conditions. Finally, the analyses above are static in nature, in that, as day + N increases, the ability of model to accurately predict decreases. In order to provide more accurate measurements and predictions, it would serve the organizations to develop, for lack of a better term, live prediction models incorporating daily, updated datapoints. This could be achieved through the use of creative dashboards and advanced business analytics.

Benefits of Using Business Intelligence

In the unprecedented conditions of this pandemic, healthcare decision makers are facing an ever-increasing demand for clinical intelligence in order to be adequately prepared for surges in COVID-related illnesses. Business intelligence activities can and should be utilized to make informed decisions with regards to hospital capacities, staffing requirements, and identification of concerning trends in both infections and related hospitalizations. Benefits of using business intelligence in the healthcare setting includes factors such as easier data access, savings in time, better decision-making, improved results and improved financial performance (Sechi et al, 2020).

In order to make accurate and relevant decisions with regards to healthcare resources, it is pertinent to develop models and analyses which can not only provide predictive capabilities but also contribute to the improved operations of the systems as well as offer improved hospitalization outcomes through adequate staffing. Managers and consumers in the health sector need real-time information to better manage data and to produce the information and knowledge they need to improve the quality of healthcare services and minimize risks (Mettler & Vimarlund, 2009). This study has aimed to assist in empowering the respective healthcare systems improving the quality of their healthcare services as well as be adequately prepared for potential surges in COVID-related illnesses and hospitalizations.

Conclusion

Healthcare organizations can extract key information and intelligence through the analysis of healthcare data sources such as physician entries, visit reasons, laboratory data, diagnoses, emergency room visits, and a plethora of other data sources; uncovering new trends, patterns, and other information. Improving healthcare quality and capacity is an issue of both public health and national security. Given the pandemic that we are currently experiencing, the healthcare systems ability to be prepared for both pandemic related admissions and typical admissions as the pandemic presents a real threat of overwhelming healthcare systems. The research conducted in this report attempted to provide such a measure through the use of linear regression and autoregressive analyses. While overall, the models appeared to exhibit moderately strong predictive capabilities, further research is necessary in the realm of predictive analytics; specifically, the creation for forecasting models such as stochastic theory mathematical models and data science/machine learning techniques. Further research into these models is necessary and will only serve to assist healthcare systems in being more prepared as the pandemic continues.

References

Belfiore, M. (2019, May 24). How 10 industries are using big data to win big. Retrieved January 10, 2021, from https://www.ibm.com/blogs/watson/2016/07/10-industries-using-big-data-win-big/.

Business, F. (n.d.). Identifying the order of Differencing in An ARIMA model. Retrieved February 05, 2021, from https://people.duke.edu/~rnau/411arim2.htm

HealthData.gov. (2020, November 18). COVID-19 Reported Patient Impact and Hospital Capacity by State Data Dictionary. Retrieved February 01, 2021, from https://healthdata.gov/covid-19-reported-patient-impact-and-hospital-capacity-state-data-dictionary

HealthData.gov. (2021, January 21). United States COVID-19 cases and deaths by state over time. Retrieved January 21, 2021, from https://healthdata.gov/dataset/united-states-covid-19-cases-and-deaths-state-over-time

Mettler, T., & Vimarlund, V. (2009). Understanding business intelligence in the context of healthcare. *Health Informatics Journal, 15*(3), 254-264. doi:10.1177/1460458209337446

Nyce, C., & Cpcu, A. (2007). Predictive analytics white paper. *American Institute for CPCU. Insurance Institute of America*, 9-10

Qian, Z., Alaa, A. M., & Schaar, M. V. (2020). CPAS: The UK's national machine learning-based hospital capacity planning system for COVID-19. *Mach Learn, 110*, 15-35. doi:https://doi.org/10.1007/s10994-020-05921-4

Qin, X., Luo, Y., Tang, N., & Li, G. (2019). Making data visualization more efficient and effective: A survey. *The VLDB Journal*, 29(1), 93-117. doi:10.1007/s00778-019-00588-3.

Sechi, G. M., Migliori, M., Dassi, G., Pagliosa, A., Bonora, R., Oradini-Alacreu, A., Odone, A., Signorelli, C., Zoli, A., & Response Team, A. C. (2020). Business Intelligence applied to

Emergency Medical Services in the Lombardy region during SARS-CoV-2 epidemic. *Acta bio-medica : Atenei Parmensis*, *91*(2), 39–44. https://doi.org/10.23750/abm.v91i2.9557

Singh, N. (2020, July 27). Advantages and Disadvantages of Linear Regression. Retrieved February 02, 2021, from https://iq.opengenus.org/advantages-and-disadvantages-of-linear-regression/#:~:text=Linear%20regression%20has%20a%20considerably,is%20very%20easy%20to%20master.