PREDICTIVE MODELLING OF COVID-19 STIMULUS FUNDS PAID FOR NURSING HOME QUALITY INCENTIVE PROGRAM

Omar Al-Azzam and Paul Court

Department of Computer Science and Information Technology,
St. Cloud State University, St. Cloud, MN 56301, USA

ABSTRACT

Painstaking measures should be taken to determine how federal dollars are spent. Proper justification for allocation of funds rooted in logic and fairness leads to trust and transparency. The COVID-19 pandemic has warranted rapid response by government agencies to provide vital aide to those in need. Decisions made should be evaluated in hindsight to see if they indeed achieve their objectives. In this paper, the data collected in the final four months of 2020 to determine funding for nursing home facilities via the Quality Incentive Program will be analysed using data mining techniques. The objective is to determine the relationships among numeric variables and formulae given. The dataset was assembled by the Health Resources and Services Administration. Results are given for the reader's insight and interpretation. With the data collection and analytical process, new questions come to light. These questions should be pondered for further analysis.

KEYWORDS

Predictive modelling, Cross validation, Linear Regression.

1. INTRODUCTION

Over the span of approximately a year, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a virus that has come to be known as COVID-19, emerged as a major health concern for people over the entire globe [1]. Oddly, older people are disproportionately affected by its adverse consequences. These devastating effects have placed an inordinate financial burden on congregate care facilities for the elderly. Roughly eighty percent of COVID- 19 deaths in the United States have been people ages 65 and older. A person 85 years of age or older has a rate of death 8700 times higher than a person 5 – 17 years of age [2]. For this reason, government funding of nursing home care facilities is of vital importance. The Department of Health and Human Services provided $2 billion as incentive payments to nursing home facilities that found techniques to lower COVID-19 infection rates and facility mortality. The Coronavirus Aid, Relief, & Economic Security (CARES) Act, a bipartisan group, collaborated with the Paycheck Protection Program and Health Care Enhancement Act (PPPHCEA) and the Coronavirus Response and Relief Supplemental Appropriations (CRRSA) Act to form an alliance tasked with determining a fair method of allocating relief funds to hospitals, nursing homes, and other front line health care facilities. Their express purpose was to assist with coronavirus expenses incurred by these facilities. These procedures would be deemed a success if the facility were able to demonstrate reduction relative to their facility's county infection rate and mortality versus national metrics [3][4].

David C. Wyld et al. (Eds): CSTY, SIGI, MaVaS - 2021
pp. 13-25, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.112202

1.1. The Research

This research will discuss an analysis of the Nursing Home Quality Incentive Program (QIP). The program was designed for the purpose of defending nursing home patients across the country against severe outcomes due to the ongoing pandemic. The program designed a method of allocating funds to facilities based on their performance [3][5]. Mortality rates that significantly exceeded the national average in each month could not receive payment for that month. To ensure fairness in the process of allocating funds via QIP, the facility needed to meet the eligibility requirements for the performance period. A metric was created that involves infection rate and mortality rate to those facilities that followed the guidelines for QIP data reporting [5]. The specific data gathered to determine QIP funding will be discussed in detail and relationships among data will be presented to detect any anomalies in funding. The evidence presented will help to make future decisions about funding patterns and processes more astutely.

1.2. The analysis processes

When data scientists are faced with predicting results for a given dependent variable, the usual process is to examine each data variable individually for shape, centre, and variability, noting any curious results that may expose themselves. Next, pairing data variables is done to evaluate strength of fit, association, and the nature of the relationship. Strength and association can be measured using correlation (positive and negative association), but the nature of the relationship (linear or non-linear) can be difficult to determine superficially [6]. Multivariate evaluations combine the effects of independent variables on a dependent variable, usually for the purpose of predicting future outcomes. Multiple independent variable predictions are often made either by ordinary least squares or cross validation techniques [7]. Figure 1 depicts this type of process. Once an iteration of the process has been completed, further investigation is usually warranted to study associations or anomalies detected in the previous cycle.

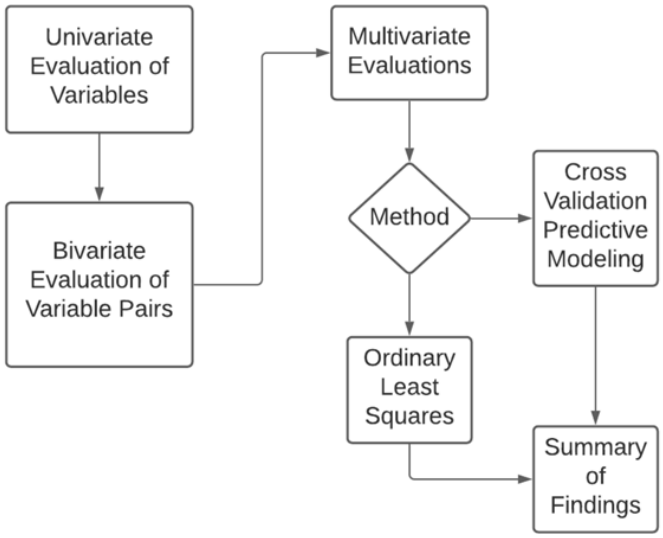

Figure 1. A standard evaluation process for critiquing datasets for the purpose of predicting values for a dependent variable. This investigative process can (and is often) repeated as anomalies are discovered.

1.3. Dataset information

The data studied in this research has 33,305 entries from across the country. Among the fifteen attributes noted are CCN number, Facility Name, City, State, Zip Code, Total Resident Weeks (TRW), Total COVID Infections (TCI), Facility Infection Rate Per 1000 Resident Weeks (FIR), County Infection Rate Per 1000 Resident Weeks (CIR), Infection Performance Score (IPS), Infection Performance Score Capped (IPSC), Mortality Adjustment (MA), Performance Month (PM), and Final Payment (FP) [3].

The CCN number is known as the Centers for Medicare and Medicaid Services certification number, sometimes called the Medicare Provider Number. It is a facility identification number used by government agencies. After the identification of the facility and its location, the numeric data in the report are a described in further detail. The data value Total Resident Weeks (TRW) is the number of residents reported in a performance period using the sum of total beds occupied reduced by the number of COVID admissions in each week of the performance period. A facility that meets the requirements will be assigned a performance score for that performance period (one month).

1.4. Calculations for Infection Performance Score and Mortality Adjustment

The steps that follow briefly summarize the performance score calculation:

Step 1: The in-facility infections in a performance period are determined by summing each week's infections as reported.
Step 2: The number of resident weeks is the sum of the total beds occupied in a facility reduced by the number of reported COVID admissions in each week of the performance period.
Step 3: The facility infection rate for a performance period is the ratio of the infections reported in a facility to the total resident weeks as calculated in steps 1 and 2.
Step 4: The county infection rate is a sum of the ratio of infections reported each week, to the total county resident weeks.
Step 5: Assuming the facility infection rate mirrors the county infection rate, an expected number of facility infections is calculated.
Step 6: Finally, the difference between the estimated and the actual infections is calculated. The results are the infection performance score for the given facility in the given county [5]. Formulae for these calculations can be found in the referenced materials.

The infection performance score provides a metric to determining how well a facility is performing versus their county and other facilities throughout the nation.

In much the same way, the mortality adjustment calculates how well a facility performs. Using county mortality estimates, an expected facility mortality is calculated. A score is then assigned to represent the difference between expected and actual deaths.

2. LITERATURE REVIEW

As mentioned, COVID-19 has disproportionally affected the lives of older Americans. Less than one percent of the Americans live in what could be categorized as a long-term care facility. However, residents and facility employees make up about forty percent of COVID-19 deaths [8]. Research about this topic indicates mixed opinions on the effectiveness of formulae applied to the data collected for the purpose of assigning funding to long-term care facilities. Usually, flat fees are charged, but recently, extra fees for specialty services have changed revenue patterns. "These

performance payments are an important stimulus for nursing homes fighting to improve their performance in a dire situation," Terry Fulmer, president of the John A. Hartford Foundation and a member of a commission on coronavirus safety in nursing homes. "As we approach the rollout of safe and effective vaccines for our most vulnerable, we continue the innovative program we created this year to incentivize and assist nursing homes in battling COVID-19 and applying the right infection control practices," said HHS Secretary Alex Azar. "This half a billion dollars in incentive payments will reward nursing homes that have shown results in their tireless work to keep their residents safe from the virus." [9]

States that have controlled the virus well in their communities, however, would typically be placed at a funding allocation disadvantage. As a congressional delegation noted, "The fact that there is a lower level of COVID-19 spread in the community in New Hampshire does not mean that Granite State nursing facilities do not need support. That is why incorporating measures of overall community spread of COVID-19 (outside of nursing facilities) into the formula is so damaging for states like New Hampshire." [10] In another example, Wisconsin's nursing home population comprises less than two percent of the nation's nursing home population but are receiving more than four percent of the emergency funds. October saw the state's positivity rate soar, making it easier for facilities to perform well with the QIP metrics [9].

Mathematicians have attempted to model natural occurring events, such as the spread of disease. This pandemic is no exception. Models can help predict the impact of these events not only by estimating cases and mortality, but necessities such as peak need for hospital beds. For a prediction to be effective, it must consider infection rate of detected and undetected cases, number of susceptible people in a population, along with those who are immune to infection. These models need to consider parameters that cannot be practically measured [11]. Predictive models should consider how public assistance can be most effectively allocated. Prognostications should tie economics to the instance of disease and the extent to which the virus has impacted individual's financial health. Factors should be considered to manage such predictions on economic impact include lockdown measures, job loss, health related expenses, socio-economic status, ethnicity, loan eligibility, even social distancing [12].

3. METHODOLOGY

To better understand how Nursing Home Quality Incentive Program funds have been distributed, its dataset will be analysed diligently. A summary of the numeric data collected for purpose of distributing funds will be organized and evaluated. This will be followed by paired comparisons between variables to see if associations exist. The relationship between pairs of variables will be evaluated using correlation as the metric. These relationships will help to narrow the focus of an ordinary least squares regression analysis with multiple independent variables. The regression equation will help interpret whether funding allocations are justified. Finally, a ten-fold cross validation procedure will be done to test the model created for the final payment of funds.

3.1. Conjectures

A conjecture would be that facility infection rate per 1000 resident weeks would be positively related to county infection rate per 1000 resident weeks. Also expected is a positive relationship between facility infection rate per 1000 resident weeks and the infection performance score. Lastly, a positive association with facility and county infection rate per 1000 resident weeks and final payment or infection performance score and final payment should be found. This would indicate a justified method for assigning funding to the facilities. High positive correlations among any of these variables of interest would not be surprising. One of the most common

methods for calculating correlation is the Pearson's correlation coefficient. It is found by summing the distance each data item is from the mean then dividing by the standard deviation [6]. For instance, there would be a high degree of positive correlation between overtime worked and money earned by an employee. If hours of overtime worked increases, there should be a somewhat predictable increase in money earned.

Equation 1. Pearson's Correlatioin Coefficient.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

The variables, however have to have a justifiable linear association for correlation to be relevant. Correlation, by nature of its calculation will always be between negative one and one, with negative one or one being perfect correlation. A high positive correlation for this scenario would be $r = 0.85$ or higher, indicating the variables are strongly associated. The associations will be analysed further by taking the data in monthly segments. If monthly correlations increase or decrease, the rationale for these changes will be considered.

3.2. Outline of the process

With these preliminary activities complete, ordinary least squares regression will consider the combined effects of the six major independent variables in the dataset: Total Resident Weeks, Total Covid Infections, Facility Infection Rate, County Infection Rate, Infection Performance Score, and Mortality Adjustment. These multiple independent variables will be used to create a linear regression model to predict Final Payment, the dependent variable. The goodness of fit for this model will be evaluated using correlation coefficient and probability models.

Next, predictive modelling using cross validation methods will be examined. *K*-fold cross validation consists of partitioning data into disjoint groups. One data segment is held out as the testing partition, developing the prediction model using the remaining partitions as a training set. When the model is trained, it can be compared to the actual data in the unused partition. Repeated *K* times and combined, an overall model will be produced and evaluated.

Other tools used to evaluate the relationship among the variables will include graphical analysis. We must look at the data to help understand them. This will assist in detecting underlying dimensions in the data.

4. RESULTS

A summary of the numeric data collected in the last four months of 2020 by the Centers for Disease Control and Prevention will be followed by pairing variables with Final Payment. Linear regression analysis with correlation as the metric will be used to evaluate the relationships.

4.1. Univariate evaluation

Figure 2 summarizes the state-by-state frequencies of nursing home facilities involved in the Provider Relief Fund COVID-19 Nursing Home Quality Incentive Program. All fifty states are represented as well as the District of Columbia and Puerto Rico.

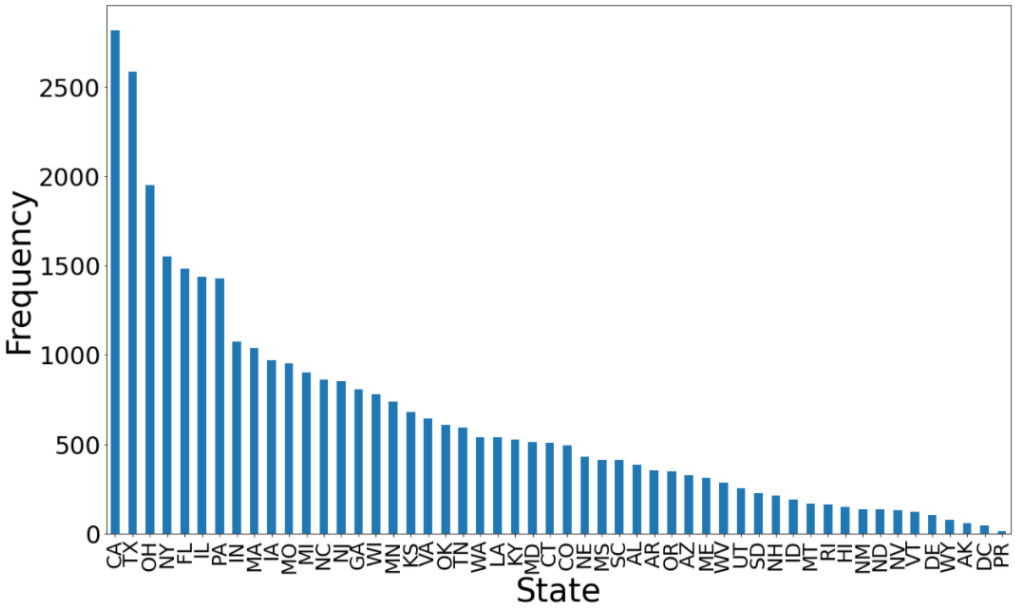

Figure 2. State by state, including District of Columbia and Puerto Rico, enrolment of nursing home facilities in the QIP. Each state with DC and PR listed on the horizontal axis and the number of facilities in each state on the vertical axis.

A histogram of the primary numeric variables represented in the dataset can be found in Figure 3. Each distribution shows a distinct skew. The top left, Total Resident Weeks is the least skewed, whereas Facility Infection Rate per 1000 Resident Weeks is the most skewed. The similar shapes would lead to the belief that these data are highly correlated.

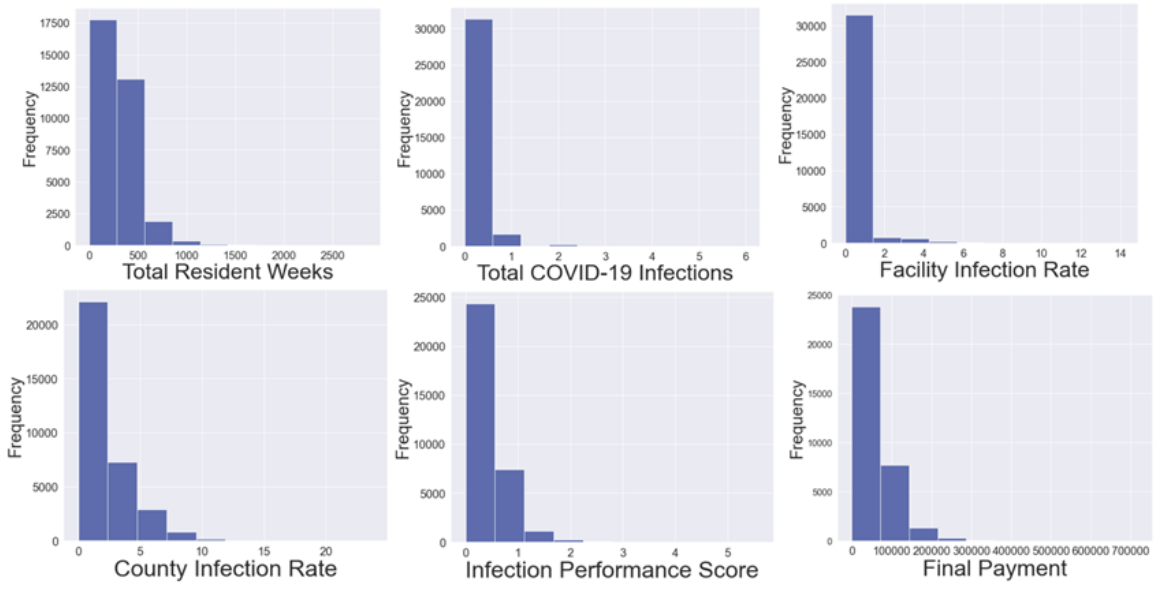

Figure 3: Top left; histogram of Total Resident Weeks (*x*-axis show TRW raw numbers vs. frequency of occurence in the data set on the *y*-axis), Top middle; histogram of Total Covid Infections (*x*-axis show TCI raw numbers vs. frequency of occurence in the data set on the *y*-axis), Top right; histogram of Facility Infection Rate per 1000 Resident Weeks (*x*-axis show FIR in rate per 1000 resident weeks vs. frequency of occurence in the data set on the y-axis), Bottom left; County Infection Rate per 1000 Weeks (*x*-axis show CIR rate per 1000 resident weeks vs. frequency of occurence in the data set on the *y*-axis), Bottom middle;

Infection Performance Score (*x*-axis show IPS vs. frequency of occurence in the data set on the *y*-axis), Bottom right; Final Payment (*x*-axis in Dollars vs. frequency of occurrence in the data set on the *y*-axis).

A numeric snapshot of each of the variables is provided in Table 1. It depicts in the dataset for the Provider Relief Fund COVID-19 Nursing Home Quality Incentive Program. The skewing of the data is again evident. The mean of each data set is larger than the median, except Infection Performance Score Capped. The minimum and maximum values for each variable also add depth to the variability and amount of skew present.

Table 1. A snapshot of statistics for the variables in the dataset. Note the mean and median of each. All are skewed to the right except Infection Performance Score Capped.

	TRW	**TCI**	**FIR**	**CIR**	**IPS**	**IPSC**	**MA**	**FP**
Mean	307.73	0.07	0.20	2.20	0.43	0.73	0.06	57,812.48
Std	195.50	0.31	0.87	2.02	0.37	0.44	0.09	49,371.35
Min	1.00	0.00	0.00	0.01	0.00	0.00	-0.2	100.67
Med	273.00	0.00	0.00	1.48	0.34	1.00	0.00	45,629.29
Max	2849.00	6.00	14.18	23.78	5.60	1.00	0.20	718,593.32

4.2. Bivariate evaluation

A heatmap of correlations among the numeric variables in the QIP data is shown in Figure 4. It compares the variables two at a time and displays the Pearson's correlaion coeffeiceient for each pair. The heatmap is symmetric with the diagonal comparing the data variable to itself and thus creating a correlation of perfect fit, $r = 1.0$. It can be disregarded. The calculations shown in Figure 4 were done using Jupyter/scipy stats package to calculate the Pearson's correlation coefficient in the standard way. The formula assumes a linear realtionship between the bivariate data and represents the ratio of the covariance of the variables to the product of their standard deviations. The heatmap shows surprising results. Among them, Final Payment funding is almost perfectly correlated with the Infection Performance. Total Covid Infection and Facility Infection Rate per 1000 Resident Weeks ranked a distant second among the relationships analysed. Final Payment vs. Total Resident Weeks, Total Resident Weeks vs. Infection Performance Score and County Infection Rate per 1000 Resident Weeks vs. Infection Performance Score round out the top five highly correlated bivariate relationships, the last having a correlation coefficient of 0.42. Referring to the original assumptions, facility infection rate per 1000 resident weeks would be positively related to county infection rate per 1000 resident weeks (true, however $r = 0.39$). Also expected is a positive relationship between facility infection rate per 1000 resident weeks and the infection performance score (true, but $r = 0.038$). Lastly, a positive association with facility infection rate and county infection rate per 1000 resident weeks and final payment ($r = 0.058$ and $r = 0.42$, respectively) should be found. This would indicate a justified method for assigning funding to the facilities.

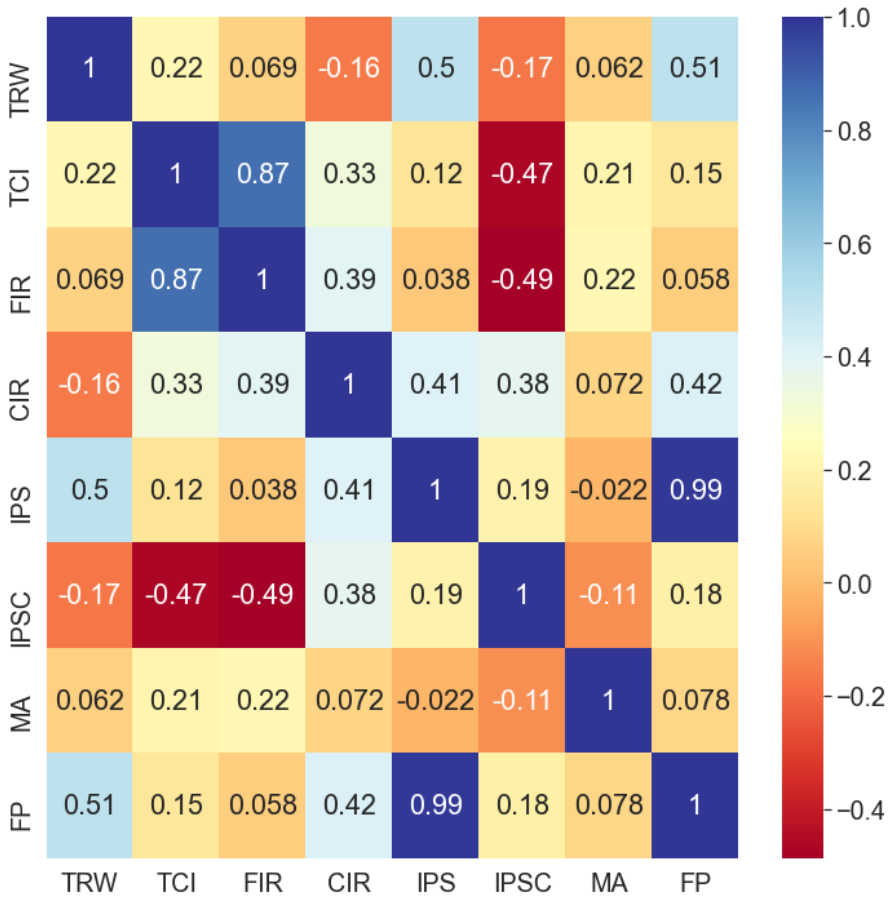

Figure 4: Heat map of numeric data bivariate correlations. The diagonal shows the variable's relationship with itself (r = 1.0). Most correlations are not very strong ($|r| < 0.50$). However, Final Payment with Infection Performance Score (r = 0.99) and Facility Infection Rate with Total Covid Infection (r = 0.87) show that these variables could have linear association.

When each of the top five relationships are investigated more closely, the month-by-month relationships showed interesting results as captured in Table 2 and depicted graphically in Figure 5. If compared to the overall correlation, the monthly correlation for final payment and infection performance score along with total covid infection and facility infection rate per 1000 resident weeks remained consistent. However, there were noticeable increases in correlation between final payment and total resident weeks, indicating that facilities with larger populations received more funding. Total resident weeks and infection performance score also showed increased correlation over the four-month period.

Table 2. An analysis of related variables separated by month

Relationship	Overall	Sept	Oct	Nov	Dec
FP vs IPS	0.99	0.99	0.99	0.99	0.99
TCI vs FIR	0.87	0.94	0.88	0.85	0.87
FP vs TRW	0.51	0.36	0.39	0.75	0.83
TRW vs IPS	0.50	0.35	0.38	0.75	0.83
CIR vs IPS	0.41	0.74	0.65	0.07	0.02

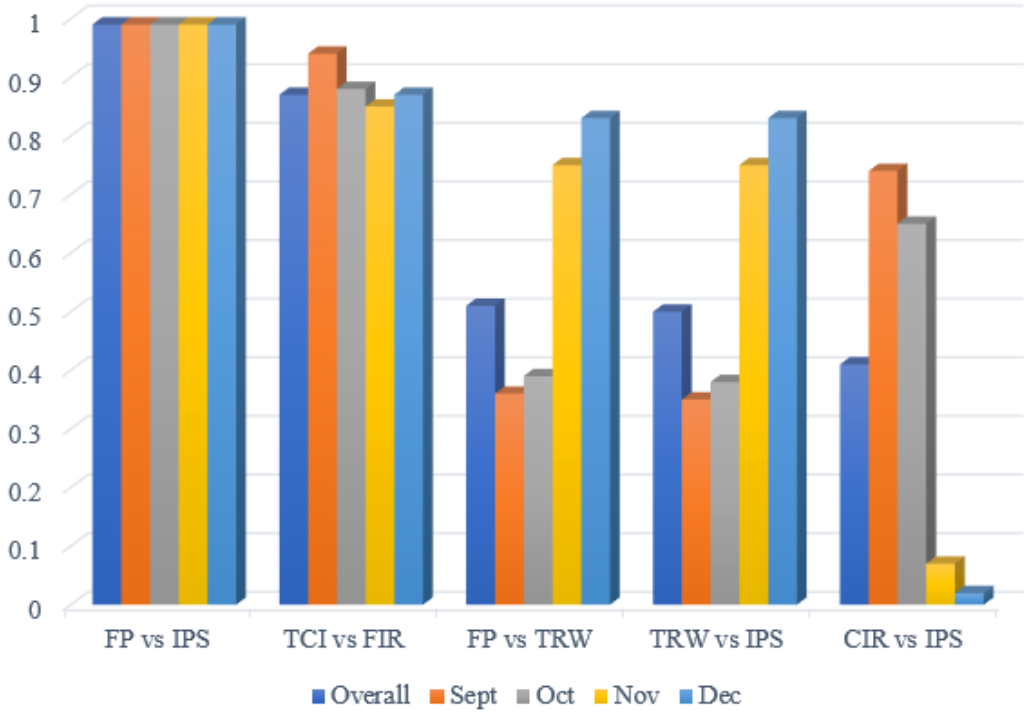

Figure 5. Graphical representation of the analysis of related variables separated by month. Overall correlation is followed by monthly results.

County infection rate per 1000 resident weeks and infection performance score correlation weakened significantly, which was a goal of the program. Recall that performance scores for a successful facility is based on having facility infection rates that do not reflect the county facility rate. However, note a decline in facility participation in the QIP over this period due to failure to meet program requirements tied to under performance.

4.3. Check for normality

Since each of the data was skewed when analysed graphically and statistically, a logarithmic transformation was applied to examine whether the data could be normalized. The results are summarized in Figure 6. If the dataset could have zero as a value, it was adjusted so the transformation could be performed. The transformation normalizes Total Resident weeks and Final Payment quite well and County Infection rate and Infection Performance score reasonably well. Total Covid Infections and Facility Infection Rate were not normalized by the logarithmic transformation. If the assumption made previously is correct, there should be a high degree of association between TRW and FP, some association between CIR and IPS, and the relationship between TCI and FIR should be weak. A formalized hypothesis regarding the relationships found in this data set will reflect the assumptions made earlier.

4.4. Formal hypotheses

Each of these hypotheses will be tested against the following two-tailed null hypothesis:

H0: The relationship expected is not present in the data.
H1: County infection rate will have a positive impact on facility infection rate.
H2: Facility infection rate will have a positive influence on infection performance score.

H3: Facility infection rate and county infection rate will dictate a facility's infection performance score and therefore their final payment.

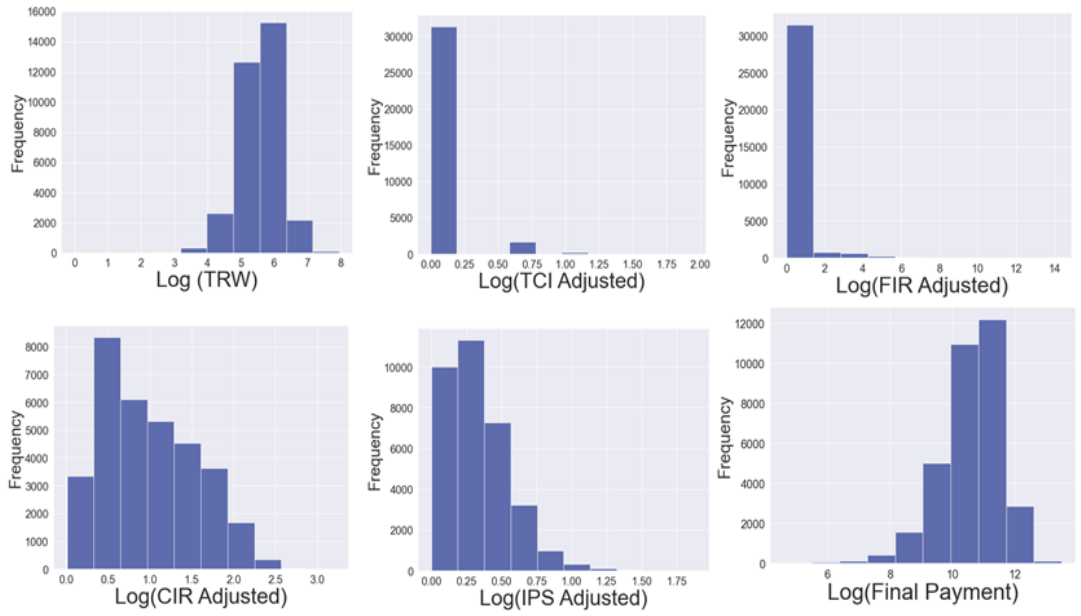

Figure 6: Top left; frequencies vs. log(Total Resident Weeks), Top middle; frequencies vs. log(Total Covid Infections), Top right; frequencies vs. log(Facility Infection Rate per 1000 Resident Weeks), Bottom left; frequencies vs. log(County Infection Rate per 1000 Weeks), Bottom middle; frequencies vs. log(Infection Performance Score), Bottom right; frequencies vs. log(Final Payment). All figures show a histogram with the frequencies on the *y*-axis and logarithmic transformations to numeric data variables found in Figure 2 on the *x*-axis. Data sets with a zero element were adjusted to allow a logarithmic transformation.

4.5. Multivariate evaluation

Using ordinary least squares regression with multiple independent variables Total Resident Weeks, Total Covid Infections, Facility Infection Rate, County Infection Rate, Infection Performance Score, and Mortality Adjustment a model to predict the Final Payment was produced. The resulting coefficients and measures of evaluation are found in Table 3. The *t*-values help to show the statistical significance of each coefficient. Probabilities under the heading $P > |t|$ give the likelihood of a result of this nature occurring by chance, assuming the two-tailed null hypothesis to be true. If the *P*-value is less than the confidence level (usually $\alpha = 0.05$), it indicates a statistically significant result [13]. Facility Infection Rate has a negative impact on the Final Payment. The *t*-value of -18 suggests that the assumption that FIR has no influence on FP is extremely unlikely. The coefficient for FIR is -990.130, indicating that its values detract from the Final Payment prediction. County infection rate also has a negative impact (t = -3.153 with coefficient -49.026), but in a much smaller magnitude. Infection Performance Score had an extreme positive impact on the prediction of the Final Payment (t = 1495.260 with coefficient 134,500). Total Covid Infection also had a positive effect on Final Payment (t = 19.867 with coefficient 2945.970), but in a much smaller way, in comparison. The coefficient of determination (r^2 = 0.993) for the model shows the proportion of the variance in Final Payment that can be attributed to the multiple independent variables assessed. Standard errors for each variable were calculated using the basic standard error of the estimate of the coefficient as provided by the ordinary least squares regression feature in python.

A ten-fold cross validation procedure was applied to the data, partitioning the data into ten disjoint segments. The dependent variable choice for the models was again Final Payment. The independent variables used were the same as in the ordinary least squares calculation. Models were trained on nine of the folds, using the tenth to evaluate the model's skill.

Table 3. Coefficients of the ordinary least squares regression model with probability values for each independent variable

	Coefficient	**Standard Error**	***t*-value**	***P* > |*t*|**
Intercept	-2465.098	55.510	-44.408	0.000
TRW	-0.981	0.160	-6.152	0.000
TCI	2945.970	148.282	19.867	0.000
FIR	-990.130	54.284	-18.240	0.000
CIR	-49.026	15.550	-3.153	0.002
IPS	1.345e+05	89.920	1495.260	0.000
MA	5.403e+04	246.312	219.349	0.000

This process was repeated using each of the ten disjoint folds as the test partition. The combined validation score for the ten folds was 0.9891. This number estimates the skill of the model in its overall performance on the test folds. It is expected that the fitted model will perform better on some folds than others. The overall results in this case are very consistent. Table 4 displays the cross-validation score for each of the ten train-test splits.

Whenever regression equations are used, a necessary procedure is to check for interesting residual results. In this case, a plot of the predicted final payment versus the residual values showed a cone shape, indicating that as payments increased, variability increased. However, when examined on a per dollar basis, the smaller final payments showed more variability.

Overall, the ordinary least squares regression and cross validation evaluations show that the infection performance score justifies the final payment allocation statistically. The methodology for determining the infection performance score may be a topic for debate for future decision-makers, but the final payments received by facilities statistically adheres to the infection performance scores calculated.

Table 4. Cross validation scores for the ten-fold cross validation procedure performed. The *k*-value indicates which of the ten disjoint folds was used to test the data, using the remaining nine as the training data

k	**Validation Score**
1	0.9880
2	0.9857
3	0.9898
4	0.9934
5	0.9938
6	0.9924
7	0.9905
8	0.9934
9	0.9800
10	0.9839

When Final Payment is predicted using the variables not contrived via the formulas provided (IPS and MA), there is a moderate positive correlation (r = 0.546). The positive contributors to Final Payment are Total Covid Infections and Total Resident Weeks with Facility Infection Rate

as the largest detractor. County infection rate was positive, but its coefficient was essentially zero. These results tend to support the design of the infection performance score and the mortality adjustment.

5. CONCLUSIONS

This analysis of the Quality Incentive Program's funding allocation to nursing home facilities across the nation examined the numeric variables and formulae involved. As was demanded by the severity of the pandemic, a rapid response was necessary. Proper evaluation of the response measures is always a good practice to help determine whether a program's goals were met and to help with future decisions about funding of this nature. The program goals were to derive a system to distribute funding tied to facility's performance versus infections and mortality in their surrounding communities. As seen, the infection performance score and final payment are very strongly correlated, providing a transparent method of allocating funds. The data studied for the four-month period also showed successful facilities did manage to out-perform their communities in disease infection and mortality. To improve the quality of the discussion started in this research, data over a longer period would help to determine if trends in funding distributions could be noted. Also of interest, would be performance of facilities that were disqualified from participation in the program as compared to their counterparts who were able to remain in the program. This comparison would shed light on whether the program indeed met the goal of assisting facilities through the pandemic. An extensive inspection from numerous viewpoints is necessary when dealing with such a complex scenario. The results are given for the reader's insight and interpretation. As usual in data analytics, more questions are unearthed that may pique interest and merit further study.

ACKNOWLEDGEMENTS

The authors would like to thank SOLAP Interactive Visualization Platform (grant 211246) for making this research possible.

REFERENCES

[1] World Health Organization, "Naming the coronavirus disease (COVID-19) and the virus that causes it", *WHO*. Accessed on: March 14, 2021. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it

[2] Center for Disease Control and Prevention, (Apr. 16, 2021) "Older adults at greater risk of requiring hospitalization or dying if diagnosed with COVID-19", *CDC*. Accessed on: Apr. 26, 2021. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/older-adults.html

[3] Health Resources and Services Administration, (2020) "Provider Relief Fund COVID-19 Nursing Home Quality Incentive Program." *DATA.gov*. Accessed on: Mar. 14, 2021. [Online]. Available: https://catalog.data.gov/dataset?publisher=HRSA

[4] Department of Health and Human Services, (Oct. 28, 2020) "Trump Administration distributes incentive payments to nursing homes curbing COVID-19 deaths and infections", *HHS*. Accessed on: February 10, 2021. [Press Release]. [Online]. Available: https://www.hhs.gov/about/news/2020/10/28/trump-administration-distributes-incentive-payments-to-nursing-homes-curbing-covid-19-deaths-and-infections.html

[5] Department of Health and Human Services, (Dec. 7, 2020) "Nursing home quality incentive program methodology", *HHS*. Accessed on: Mar. 18, 2021. [Online]. Available: https://www.hhs.gov/sites/default/files/nursing-home-qip-methodology.pdf

[6] Stojiljkovic, Mirko, "NumPy, SciPy, and Pandas: Correlation with Python", *Real Python*. Accessed on: Mar. 27, 2021. [Online]. Available: https://realpython.com/numpy-scipy-pandas-correlation-python/#example-numpy-correlation-calculation

[7] Zakeri, Z., Mansfield, N., Sunderland, C., and Omurtag, A., (2020) "Cross-validation models of continuous data from simulation and experiment by using liner regression and artificial neural networks." *Elsevier*. Volume 21.

[8] Udow-Phillips, M. and Rontal, R. (2020) "Reforms needed after systemic flaws in nursing homes worsen outcomes from COVID-19." *American Bar Association.* Volume 41, Issue number 6. Accessed on: Feb. 4, 2021. [Online]. Available: https://www.americanbar.org/groups/law_aging/publications/bifocal/vol-41/vol-41--issue-no-6--july-august-2020-/systemic-flaws--tragic-outcomes/

[9] Englund, W., (Dec. 9, 2020) "For the first time, the U.S. will reward nursing homes for controlling the spread of infectious disease", *Washington Post.* Accessed on: Mar. 8, 2021. [Online]. Available: https://www.washingtonpost.com/business/2020/12/09/nursing-home-infection-control/

[10] Williams, B., (Dec. 8, 2020) "Nursing home quality incentive program? Hardly", *McKnight's Long-Term Care News*. Accessed on: February 4, 2021. [Online]. Available: https://www.mcknights.com/blogs/guest-columns/nursing-home-quality-incentive-program-hardly/

[11] Ivorra, B., Ferrandez, M.R., Vela-Perez, M, and Ramos, A.M., (Apr. 30, 2020) "Mathematical modeling of the spread of the coronavirus disease 2019 (COVID-19) taking into account the undetected infections. The case of China", *Elsevier Public Health Emergency Collection.* Accessed on: March 30, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7190554/

[12] COVID Economics, "About Covid Economics", Centre for Economic Policy Research. Accessed on: Feb. 19, 2021. [Online]. Available: https://cepr.org/content/covid-economics-vetted-and-real-time-papers-0

[13] Prettenhofer, P., (Feb. 8, 2014) "Ordinary least squares in Python", *DataRoot*. Accessed on: Feb. 18, 2021. [Online]. Available: https://www.datarobot.com/blog/ordinary-least-squares-in-python/

AUTHORS

Dr. Omar Al-Azzam is an Associate Professor of Software Engineering in the Department of Computer Science and Information Technology (CSIT) at Saint Cloud State University (SCSU). Dr. Al-Azzam earned his BSc and MSc from Yarmouk University, Jordan and PhD from North Dakota State University (NDSU). Dr. Al-Azzam main research interests are big data analytics, bioinformatics and data mining.

Paul Court is a graduate student in the Professional Science Master of Software Engineering (PSMSE) program at Saint Cloud State University (SCSU) in the Department of Computer Science and Information Technology (CSIT). Mr. Court earned a MEd in Mathematics from the University of Minnesota and a BA in Mathematics from the University of Minnesota, Morris.