Data Science (Cpts 575)
Fall 2019
Patrick Keppler
Matthew Mietchen

Class Project Final Report

Parameter Emulation in a Disease Transmission Model

Abstract

The study of pathogen transmission within an Intensive Care Unit (ICU) is an important tool to reduce the burden of patient infections. Using a stochastic compartmental model of an ICU, our outcome of interest or the number of methicillin-resistant Staphylococcus aureus (MRSA) acquisitions of patients, were simulated from randomly chosen model parameter values. Data cleaning and exploratory data analysis (EDA) were performed prior to building several regression models in an attempt to find a machine learning algorithm that could predict a parameter value of the model that is difficult to obtain. Our results found that a linear model using pair-wise interactions of all parameters would be the best fit model to predict acquisitions, which was an important first step in determining whether our parameter of interest may have some relationship with the outcome. In conclusion, our methods found that there are linear relationships within the simulated data and that acquisitions could be predicted, however, building a model with our parameter (psi) as the outcome, proves to be more difficult. Further investigation is required to evaluate whether an actual machine learning algorithm could be built in order to reduce computational time of fitting the parameters to a model for simulation.

Introduction

Healthcare associated pathogens, such as MRSA, continue to be problematic in healthcare settings, especially in an ICU. While the infections have been declining over recent years, the rate of decline is slowing. A recent CDC report found that in 2017, nearly 120,000 *Staphylococcus aureus* bloodstream infections were reported and close to 20,000 associated deaths occurred [1]. Treatment of antimicrobial infections are problematic and difficult. One type of intervention is to reduce the spread of pathogens, such as MRSA, rather than solely focusing on the treatment. Building mathematical models to study these interactions based on population structure of those who are at risk of becoming colonized is less expensive and easier to conduct

than traditional clinical trials or cohort studies. ICUs are important and very convenient to study MRSA due to the typical structure they naturally take when patients are assigned to healthcare staff in groups. It is also important to build a model with strict interactions, however, that is unrealistic when random events may dictate a staff member interacting with unassigned patients. As with most infectious disease stochastic models, a very important question is whether a more simplified model can give an acceptable approximation or do we need to move toward complex and large models to further understand the transmission dynamics. The model used for this type of study includes several important features or parameters based on real-world interactions and processes.

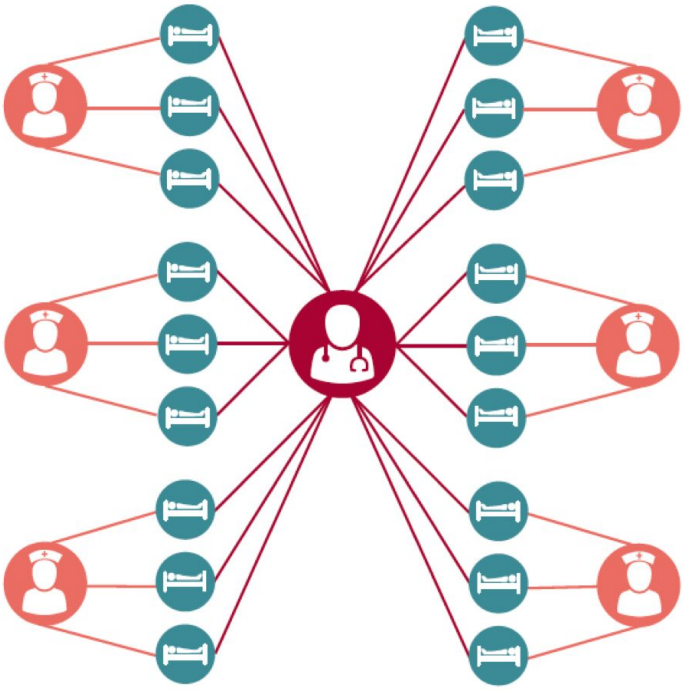
A stochastic compartmental model of a hospital ICU was used to simulate data for the project [2]. The number of patients colonized with MRSA was the model output. The structure of this ICU takes on a metapopulation characteristic as the nurses interact randomly with a specific set of patients that are connected by a single individual, in this case, a physician (Figure 1). This stochastic compartmental model had several assumptions; if a patient is discharged, a new admission occurs immediately, the ICU consists of single-occupant patient rooms only, patients do not enter other rooms, nurses and the MD do not interact with each other or any other fomites such as medical devices or surfaces that may cause contamination, no treatment or interventions are performed for colonized patients, and standard hospital contact precaution guidelines are followed to detect MRSA colonization with perfect accuracy.

MRSA acquisitions are dependent upon the model parameters (Appendix Table 1). These parameters are estimated and determined by published clinical trials and literature. In a typical ICU, most of these parameters are known or tracked by the hospital infection control program, or the program at least has a good idea or estimate. However, there is one parameter, psi, which is the probability of successful colonization of a patient given contact with a contaminated nurse, that is not typically known or available. Psi therefore has to be fitted to the model using a method called Approximate Bayesian Computation (ABC) in order to get a posterior estimate [3]. This parameter is important as it describes the probability of MRSA transmission. However, estimating this parameter comes at a high computational cost. Therefore, our main objective and motivation of this study was to explore and determine if Machine Learning could estimate psi when the other model parameters and the MRSA acquisitions are known. This would be beneficial by saving computational time, and most importantly, be used to potentially reduce MRSA transmission.

To achieve the project objective, several regression models were built and compared for best fit to the data with the acquisitions as the outcome. It was important to begin with the response as acquisitions first to determine if any relationships were present in the data. Linear relationships were found, and a model was selected from various methods. These findings were very important to begin exploring how the parameter of interest, psi, could be predicted using the other known

parameters. Unfortunately, in an attempt to build a linear model to predict psi, the models performed poorly. It is important to reiterate that predicting psi may not be possible due to the noise of the system and was already known to be a possible result. Reevaluation of the project will need to occur next, as well as looking into other possible methodologies to apply to the analysis.

Figure 1. Stochastic Compartmental Model with a Metapopulation Type Structure of an ICU



Problem Definition

The main objective of this research was to develop a machine learning algorithm to predict a model parameter called psi. In order to determine the feasibility of achieving this objective, developing a predictive linear regression model for the outcome variable, or the number of MRSA acquisitions, would indicate if regression modelling for this data is a viable approach.

Motivations for the research are focused on being able to use machine learning to expedite the prediction process of an important, but difficult to obtain, parameter. In order to simulate the ICU model for a specific ICU or under certain circumstances, such as an outbreak, the parameter psi would take several days to estimate using current methodologies, such as the ABC method. An ideal alternative method should be able to increase the speed of prediction with minimal loss to precision.

Models/Algorithms/Measures

Several techniques and methods were used to investigate the research problem, as described in the previous section. Exploratory Data Analysis (EDA) including scatter plots, box plots, histograms were important to understand the data structure and visualize relationships. Principal Components Analysis (PCA) was another measure used to further understand the dimensionality of the data. Linear regression models were built and model comparison/selection was achieved by backward-selection techniques where removing specific features from the model that did not seem to have a relationship or explained the data. Another important algorithm used for evaluation of model performance and also assisted in model selection was the Least Absolute Shrinkage and Selection Operator or LASSO model. These models, algorithms, and measures were important for the analysis and investigation into addressing the research problem.

Implementation/Analysis

Data Cleaning and EDA

The source of the data used for training the model was a series of executions of a stochastic simulation. Each observation in the data frame represented a different combination of parameters selected at random, which were used in the simulation to produce an output of acquisitions. Because the source of the data was generated, the data was already known not to contain missing or improperly-recorded values. Thus, data imputation and tidying was not a necessary step in processing the data.

One challenge more unique to simulated data was the inclusion of observations that could not be justified in a real-world setting. This problem is similar to the problem of outliers in real-world data; however, the cause of these two problems are different. While outliers are often caused either by errors in recording or small numbers of observations of highly unusual behavior, the observations in this data that have the appearance of outliers are properly generated by the simulation. Even though these observations are possible to generate through the simulation, the inclusion of these observations will skew any models built off of the simulated data.

By examining summary statistics and boxplots of select features, it was determined that the most unrealistic and problematic data was contained in the top five percent of acquisition cases. Observations with less acquisitions than the top five percent were considered reasonable to keep in the dataset. While the possibility of removing the bottom five percent of data was also considered, the inclusion of observations with close to zero acquisitions was ultimately determined both realistic and desirable.

PCA

The simulated data predicts acquisitions using eleven different parameters. Consequently, the goal of Principal Components Analysis (PCA) was to investigate the possibility that a lower-dimensional model could be constructed. In a typical application of PCA, the number of chosen principal components selected for use in a model corresponds to the proportion of variance explained (PVE). Each principal component explains a fraction of the overall PVE, with principal components being selected such that the PVE of each additional principal component is less than the PVE of the previous principal component.

In the ideal case, a very small number of principal components will account for a large amount of the PVE, followed by rapid diminishing returns for additional principal components. For this project we wanted to investigate the possibility that the current set of features could be reduced with minimal loss of predictive power.

Regression Models

Since the goal of the project was to identify a model for predicting continuous outputs, such as acquisitions, or continuous parameters, the approach chosen was selecting linear regressions with different feature sets for comparison. The first model produced was called the "Full" model, containing only the eleven parameters from the data set.

The next model produced was the "Interaction" model, which included the eleven features as well as the pairwise interactions between the features. The motivation for this model was that despite the parameters for each observation being chosen at random and therefore uncorrelated, the impact that these parameters have on the outcome may not be independent.

Finally, the last model produced directly during regression was the "Reduced" model. This model examines the summary statistics of the interaction model, and removes the interactions and features not considered statistically significant. The significance of a feature related to the feature's p-value; features with a p-value of above 0.05 were excluded from the model, unless interactions involving this feature were present with p-values below 0.05.

LASSO

In addition to three regression models described above, the LASSO method was used as an alternative regression technique. This method applies penalties to large coefficients in order to reduce variability in the model. In some cases, the application of this penalty can result in feature coefficients being computed as exactly zero, meaning that their impact is completely removed from the model. Because of this property, LASSO could also be used to validate the inclusion or removal of features in the reduced model. Due to the indication in the construction of earlier models that interactions were significant, the LASSO model was also built using interactions.

Results and Discussion

Data Cleaning and EDA

As described in the previous section, while the source of data ensured that no problems were encountered concerning missing values, some simulation outputs resulted in highly unrealistic scenarios that should not be included for training purposes. For instance, the maximum recorded value for acquisitions in the set was 12,100, over 100 times the value of the median of 116, and still substantially larger than the third quartile value of 177. As seen in Figure 2, due to the large unrealistic observations, the plot is skewed too far to be interpreted. This boxplot contains one categorical feature called bed size against the acquisitions over the entire dataset and shows how much these large acquisitions affected the visualization of the data. However, once these large observations were removed, a better visualization of the relationship is shown in Figure 3. These results are helpful to understand the observed relationship and justified the removal of those extreme values.

Figure 2. A boxplot which contains the full range of acquisition values in the dataset

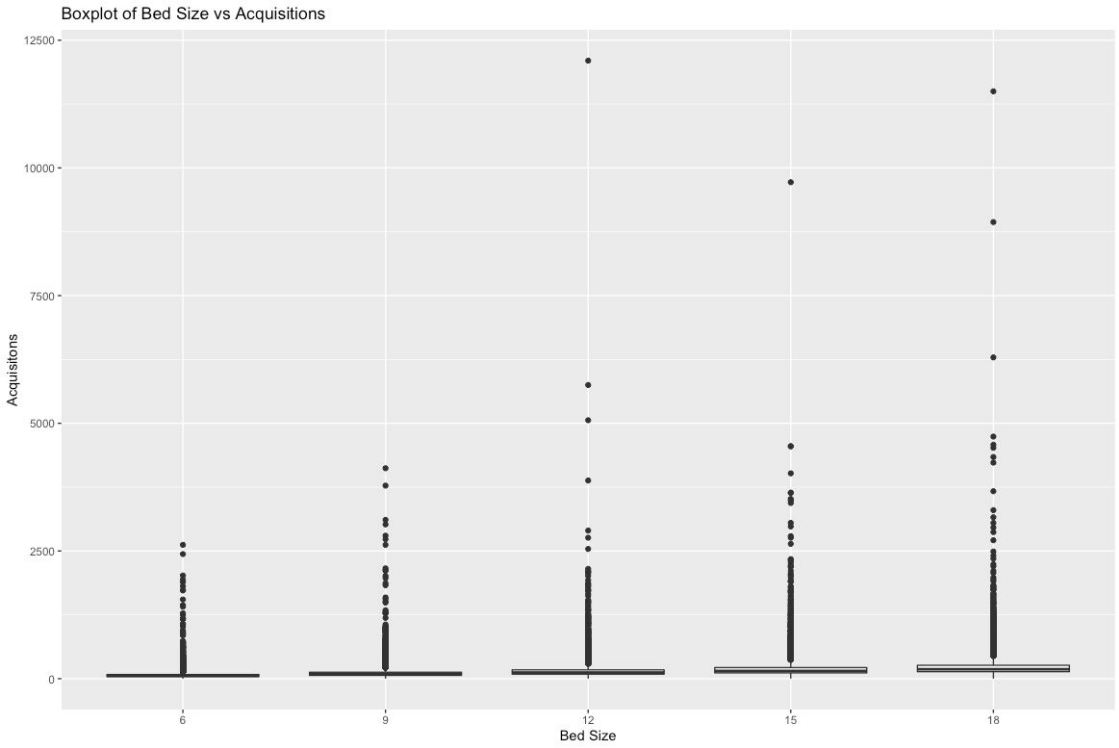
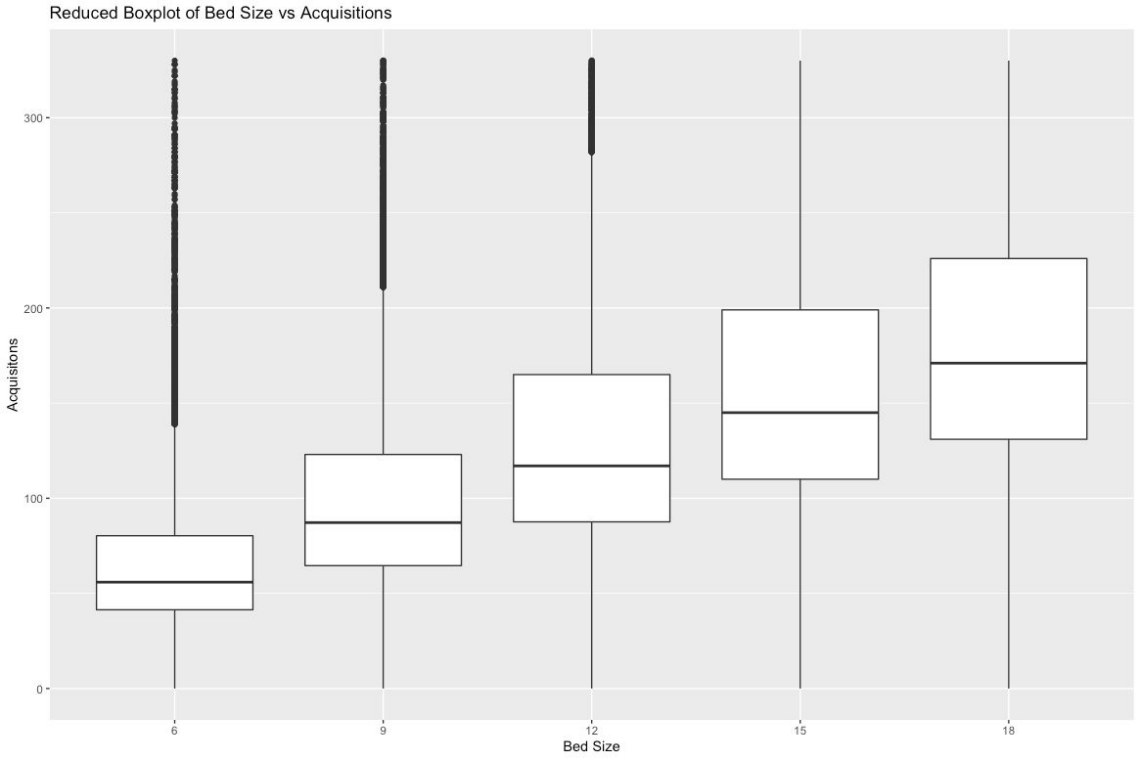


Figure 3. A boxplot without the top 5 percent of acquisition values in the dataset



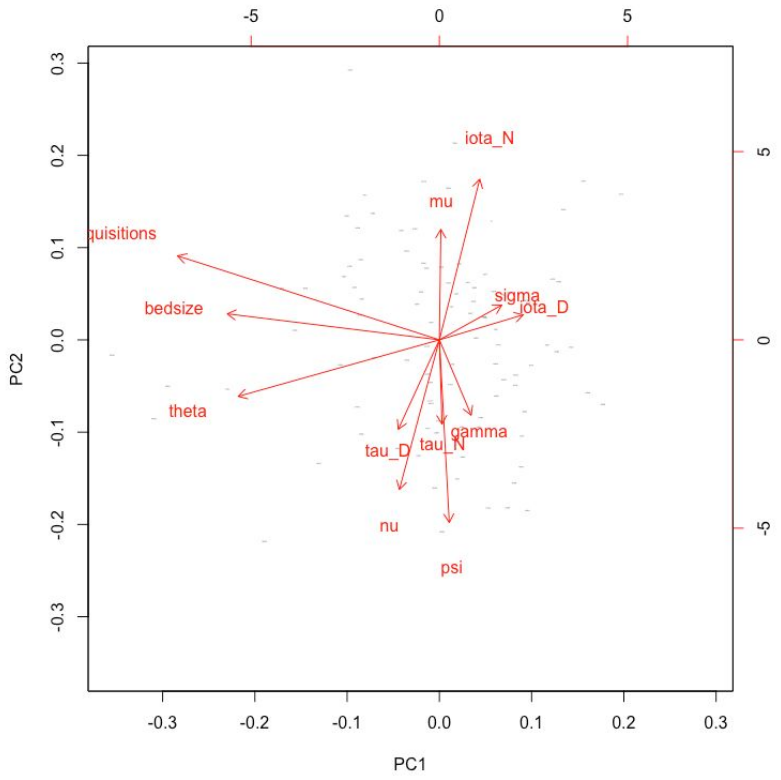
There was an expected positive relationship between bedsize and acquisitions, which supports the reasoning that when more patients are added to the system, more acquisitions should be observed. However, with high values for the observations being observed, the extreme values are unlikely to occur in any healthcare setting. There would never be over 12,000 acquisitions realistically in the natural world. To reiterate, even in a situation where we looked at the largest ICUs in the dataset, containing 18 beds, this means that acquisitions would need to occur on average 33 times per day for the entire year. Therefore, limiting the dataset to a more realistic maximum limit is useful in order to visualize the data.

While the data visualization was useful for identifying the impact of categorical features on acquisitions, there were challenges involved in doing the same for the continuous features. Attempting to use scatter plots to show the relationships between continuous values was impeded by the quantity of data points. Attempts at including plots with a small sample of points were also unsuccessful, as were alternative plots such as density graphs.

PCA

The principal components analysis for this project was to investigate whether the data could be modeled effectively by a smaller set of features. Figure 4 shows how the features map relative to the first two principal components.

Figure 4. The biplot for the first two principal components with loading vectors.



For the data in this project, the initial principal component explained a relatively small amount of the variance (under 15 percent), and the next ten principal components explained an approximately even amount of the remaining variance. This result indicated that principal components could not be used to effectively reduce the dimensionality of the model. Figures 5 and 6 show the PVE and the cumulative PVE from the PCA output, respectively.

Figure 5. Principal component scree plot

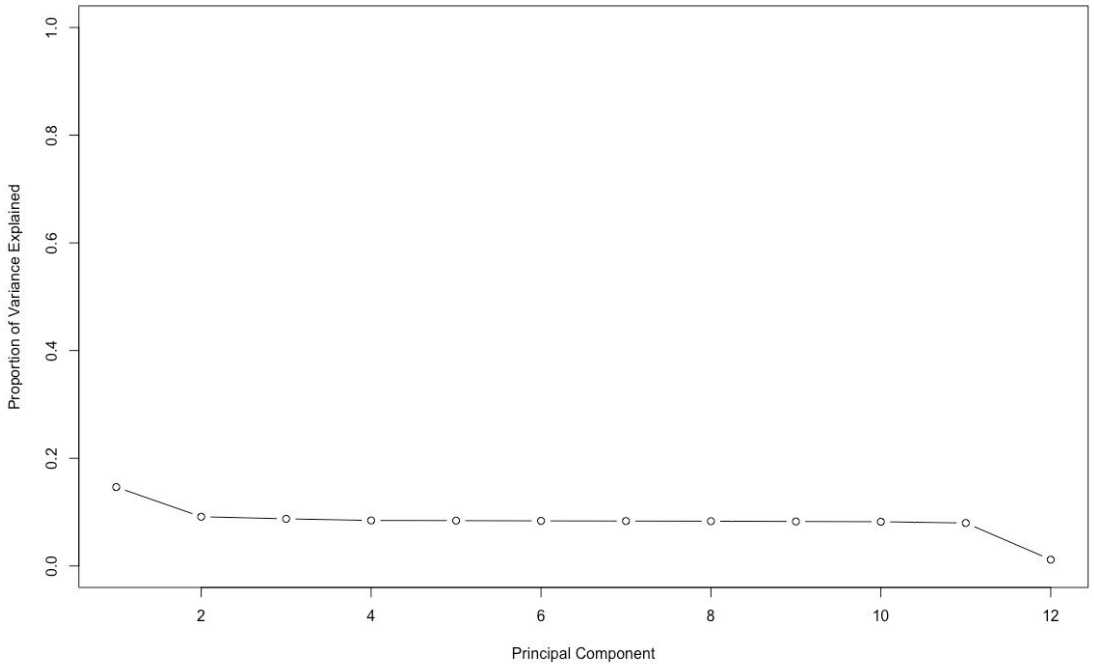
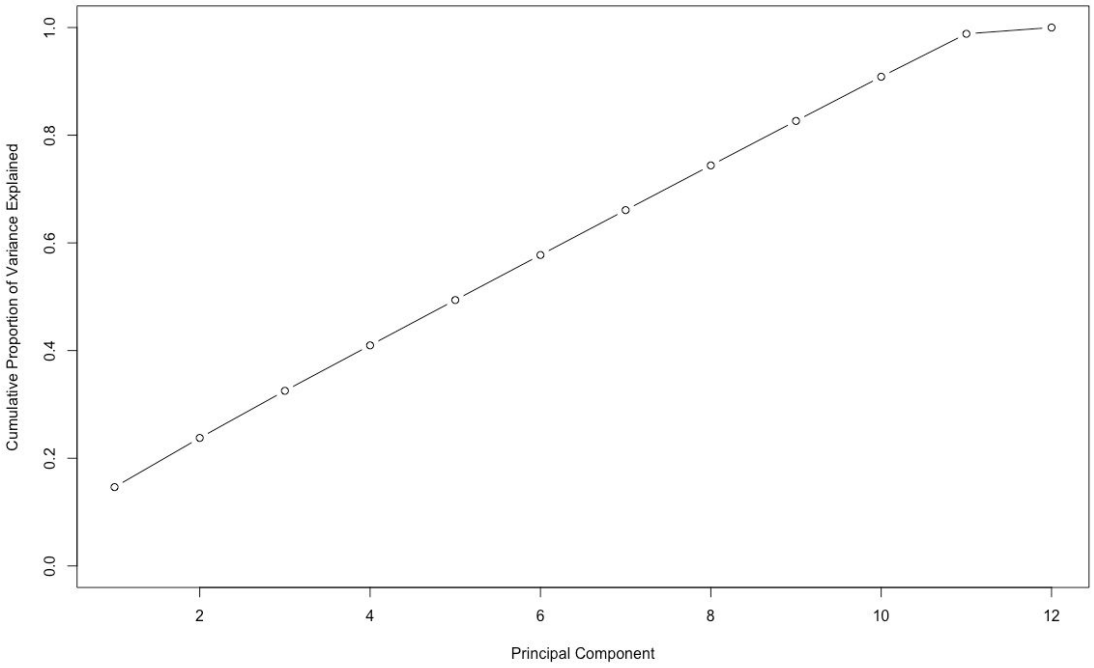


Figure 6. Cumulative principal component scree plot



Model Evaluation

To measure the performance of the models created for this project, a number of different metrics were considered. For the first three regressions, the adjusted R-squared values were extracted from the summary statistics of each model. The Akaike information criterion (AIC) was also computed for each of the models, using generalized linear models with the corresponding set of features to produce the AIC. The model scores using both of these evaluation metrics can be seen in Table 1.

The last evaluation metric chosen was root mean squared error (RMSE). This metric was chosen specifically because of the error term being in the same units and scale as the output of the model. For example, a RMSE of 36 for acquisitions implies that the expected error of prediction using this model is 36 acquisitions.

RMSE was computed through the use of a test set. Each model generated a set of predictions for the test data (called $\hat{y}$ collectively), which were compared to the actual values from the test set (called y collectively) using the following formula:

$$\text{RMSE} = \sqrt{\frac{\sum_i (\hat{y}_i - y_i)^2}{N}}$$

RMSE was applied to the first three regressions as well as the LASSO regression in Table 2.

The adjusted R-squared values the models indicated strong fits, especially for models that included interactions. By this metric, there was no clear advantage to including all interactions over using the reduced model, but there is a substantial advantage in using these two models over the full model.

The AIC is more difficult to interpret in an objective manner, but is still useful for comparing models trained from the same data. A lower AIC indicates better performance for the model. In this case, the AIC for the reduced model is slightly lower than the interaction model, indicating slightly better performance. However, these two scores are close enough to indicate that the difference in performance is again negligible. Once again, the two models containing interactive effects vastly outperformed the model without.

Table 1. Linear Regression Model Comparison

Linear Model	**AIC**	**Adjusted R^2**
Full	951,205	0.7241
Interaction	887,554	0.8589
Reduced	887,536	0.8589

The RMSE scores for the models were the primary area where the models did not perform as well as hoped. The same pattern of models with interactions outperforming the model without holds true here as well, as a lower RMSE is preferred. The LASSO regression performed very comparably to the other interaction models. However, the Interaction model should be considered the best model from these results due to its performance in RMSE. Because of the role of RMSE in measuring the prediction capability of a model, RMSE is the most important metric for model evaluation in this problem. Despite the similar performance of the best models, the RMSE is relatively high in context. If a hospital wants to approximate acquisitions using one of these models, an expected error of nearly 26 acquisitions is likely too high to be useful.

Table 2. Linear Model Performance Using RMSE

Linear Model	**RMSE**
Full	36.276
Interaction	25.885
Reduced	25.890
LASSO	25.893

<u>Psi Prediction</u>

In an attempt to build a prediction model for our parameter of interest, psi, the Interaction and LASSO models were chosen due to their performance in the acquisition prediction. These two models were also chosen in order to compare different approaches to regression for prediction of psi. The RMSE result for the LASSO and Interaction models were 0.289 and 0.287, respectively. Given the value range of psi being from 0 to 1, these results indicate low confidence in the models' abilities to predict psi.

Related Work

Healthcare associated infections spreading within and throughout a hospital setting, such as an ICU, is being addressed at many fronts. Most studies and research over the last decade has contributed to the understanding of specific issues such as hand hygiene, surface contamination, length of stay in the ICU, and isolation precautions. Other research continues to focus on the medical devices and trying to identify the source of contamination. However, limited research has been done evaluating the structure of the ICU and the population interactions between healthcare workers and patients. Partly, this is due to logistical and ethical limitations. Related works, such as the stochastic compartmental model described throughout this current research has attempted to answer this question. In addition, methods such as the approximate Bayesian computation (ABC) algorithm has advanced the ability to estimate unknown posteriors for certain parameters such as the one of interest in this research, psi.

Conclusion

This project attempted to first determine whether regression modelling could be used to effectively and efficiently predict the output of a stochastic simulation, given a set of parameters. By most measures such as R-squared and AIC, the regression models with interaction terms were largely successful at fitting to the data. However, the prediction error computed through RMSE needed to be lower for any of the regression models to be considered as a replacement for previous prediction methods. Overall, given the output measures, the Interaction model would be the preferred model due to the additional performance not coming at a large cost.

The models created for predicting acquisitions served as evidence that such a method could also be used as a method for predicting a parameter when the acquisitions and other parameters are known. In this respect, both the linear and LASSO regressions failed to produce acceptable results when attempting to predict psi. Both by R-squared and RMSE metrics, models used in prediction of psi were far less effective, to the point that the certainty in the prediction is too low to be useful at all.

While the results indicate some potential for regression to predict simulation outputs, the results strongly suggested regression would not be an effective method for predicting a parameter given the output. Future work will involve investigating other possible efficient methods of parameter prediction.

Bibliography

1) Kourtis A, Hatfield K, Baggs J. Vital Signs: Epidemiology and Recent Trends in Methicillin-Resistant and in Methicillin-Susceptible Staphylococcus aureus Bloodstream Infections — United States. *MMWR Morb Mortal Wkly Rep*. 2019;(68):214-219. doi:http://dx.doi.org/10.15585/mmwr.mm6809e1

2) Mietchen M, Short C, Samore M, Lofgren E. Population Structure Drives Differential Methicillin resistant *Staphylococcus aureus* Colonization Dynamics. ***2019***. DOI: https://doi.org/10.1101/19002402

3) T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf, "Approximate Bayesian Computation scheme for parameter inference and model selection in dynamical systems," J. R. Soc. Interface, vol. 6, no. 31, pp. 187–202, Feb. 2009.

Appendix (Supplementary Material)

GitHub link
https://github.com/Mmietchen/Model-Emulation-Project-CptS-575-

Table 1.

Parameters for modeling the acquisition of methicillin-resistant Staphylococcus aureus in an Intensive Care Unit			
Parameter	Parameter Description	Parameter Value	Source(s)
ρ	Contact rate between patients and HCWs	4.154 (# of direct care tasks/hour)	[22], [23]
ρ_N	Contact rate between patients and nurses	3.973 (# of nurse direct care tasks/hour)	[22], [23]
ρ_D	Contact rate between patients and physician	0.181 (# of physician direct care tasks/hour)	[22], [23]
σ	Probability that a HCW's hands are contaminated from a single contact with a colonized patient	0.054	[13]
$\Psi_{\text{Metapopulation}}$	Probability of successful colonization of an uncolonized patient due to contact with a contaminated HCW in metapopulation structure	0.0464	Fitted to [17]
θ	Probability of discharge	4.39 $days^{-1}$	[17]
ν_u	Proportion of admissions uncolonized with MRSA	0.9221	[17]
ν_c	Proportion of admissions colonized with MRSA	0.0779	[17]
ι	Effective hand-decontaminations/hour (direct care tasks × hand hygiene compliance × efficacy)	5.740 (10.682 direct care tasks/hour × 56.55% compliance × ~ 95% efficacy)	[17], [22]–[24]
ι_N	Effective nurse hand-decontaminations/hour	6.404 (11.92 direct care tasks/hour × 56.55% compliance × ~ 95% efficacy)	[17], [22]–[24]
ι_D	Effective physician hand-decontaminations/hour	1.748 (3.253 direct care tasks/hour × 56.55% compliance × ~ 95% efficacy)	[17], [22]–[24]
τ	Effective gown or glove changes/hour (2 × # of visits × compliance)	2.445 (2.957 changes/hour × 82.66% compliance)	[13], [17], [20]
τ_N	Effective nurse gown or glove changes/hour	2.728 (3.30 changes/hour × 82.66% compliance)	[13], [17], [20]
τ_D	Effective physician gown or glove changes/hour	0.744 (0.90 changes/hour × 82.66% compliance)	[13], [17], [20]
μ	Natural decolonization rate	20.0 $days^{-1}$	[21]