Causal Analysis: Effects of Nursing Home Facilities on Health Inspection Rating

Jonah Winninghoff *Springboard Data Science*

The U.S. Centers for Medicare and Medicaid Services (CMS) use the star rating system to assist consumers with choosing the nursing home since 2009. However, during the COVID-19 pandemic, one of these studies demonstrates that the star rating system is a failure largely due to a lack of data audit and self-report bias. Two approaches, data science and econometrics, are applied in this analysis. The objective is the same by attempting to establish a causal relationship using the health inspection rating as a metric, rather than by relying on self-report rating results. There are two findings: the number of certified beds and the amount of hours with residents assigned by license practical nurses are negative contributors to the star rating system. Several solutions are discussed (for example, AI solution and LPN-to-RN career pathway pilot program).

Keywords: Bayes Optimal Feature Selection, Lasso regularization, Big Data, Gauss Markov Assumptions, Probit

---

October 12, 2021

INTRODUCTION

Established by U.S. Centers for Medicare and Medicaid Services (CMS) in 2009, the star rating is a system that helps people to make decisions in which nursing home the senior residents would reside. The system consists of three elements that it has been subjected to, which is in-person inspection evaluation, the amount of time the nurses spend with residents, and the quality of care that the residents receive. The in-person inspection evaluation accounts for most parts of the system. The system became recently popular. However, the problem in this system emerged when the COVID-19 pandemic occurred. On March 13, 2021, the virus killed more than 130,000 residents, which accounts for, roughly, 25% of total deaths in the United States. The number of deaths remains underestimated. It is important to note that this event is historically tragic. This pandemic offers key insights that the rating itself is a failure due to *moral hazard*. For example, the study shows that the number of COVID-19 deaths at five-star facilities are not significantly different from one-star facilities (Silver-Greenberg and Gebeloff, 2021).

The recent study demonstrates the data collection issues with CMS databases. A failure of this system is largely due to a lack of data audit and self-report bias. Tethered to the moral hazard, this self-reported behavior is prone to deviating from cooperative behavior that the report itself expects to be truthful. The problem with this system is that the facilities have incentives to inflate their report by vying to offset revenue against cost. The facilities perceive this effort, rather than as a consequence, as an opportunity cost that they are willing to compensate for positive profit.

Consequently, the consumers face *adverse selection*. For example, the consumers know less about internal situations than the staff and employers. The consumers' inability to distinguish different levels of quality care services in the facilities lead to a decline in public trust that could endanger both nursing home industry, and citizen-government relations. For that specific reason, the causal analysis is a linchpin in addressing this moral hazard problem. There are two different frameworks tested in this analysis, which are data science and econometrics. The data science approach searches the patterns in data and developing models tested against data. However, the econometric approach begins by writing a causal model of economic behavior and its underlying assumptions, followed by determining whether the available data fits in the causal model. This analytic process is to address what the next step this action should be taken.

EXPLORATORY DATA ANALYSIS

Two datasets obtained from CMS databases are Minimum Data Set (MDS) Quality Measures and Provider Information datasets. The MDS Quality Measures dataset contains over 15,000 different providers from 50 states plus District of Columbia. None of the variables holds predictive power for measure quality. However, this dataset does have one predictor (or independent variable) called zip code that can predict measure quality. The R-squared score for simple regression using this predictor accounts up to 20.9%. This regressor explains at least 20% of the variance for target variable (or dependent variable).

The problem with this approach is that this predictor is *proxy* for race. Any models using the features correlating to the protected classes is a form of indirect discrimination, which is known to be proxy discrimination. If implementing the algorithm, the disparate impact may be occurring (Datta et al., 2017). For example, the nursing home located in a certain zip code may be penalized by this algorithm even if this institution provides high quality of services, rather than an attempt to establish a causal relationship. In the nutshell, this dataset has certain features that are useful for statistical inferences.

For example, the measure quality in this dataset is in use to describe the overall rating of what each facility does well associated with every mea-

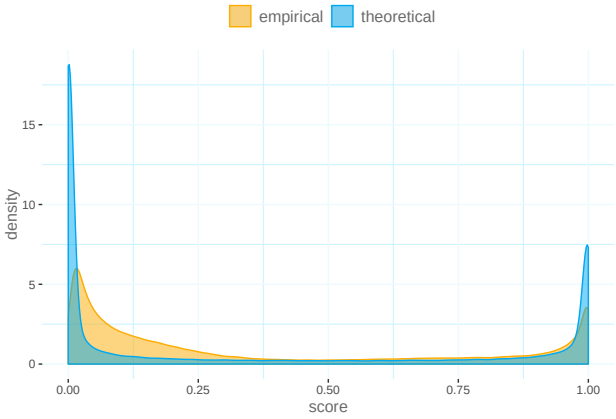
sure code. The measure code associates with measure description that explains how this score is calculated. This information can provide more insights of how each facility performs particular tasks. But, as indicated by the complete data quality report, this score is not normally distributed. The Empirical Cumulative Distribution Function (ECDF) tool is undertaken to compare empirical distribution with theoretical distribution and determine if the empirical distribution is parametric.

This approach is more preferable over the histogram approach since it does not have binning bias. The binning bias refers to the change in the number of bins that changes meanings in descriptive statistics. For example, the histogram shows, when the size of bins is small, the distribution is Gaussian. When the size of bins is increased, the distribution turns bimodal. The point in mentioning this is that the ECDF approach should be in use to better infer what is presented.

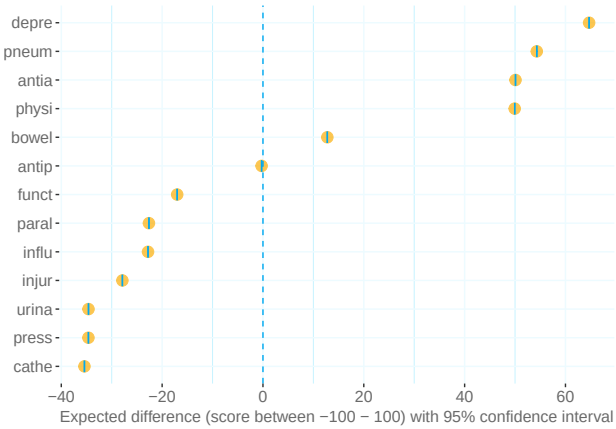
This distribution should be identical to or, at least, consistent with the theoretical one. Given that this score is $\epsilon(0,100)$ and cross-sectional, the continuous version of binomial distribution called Beta distribution is in use to compare this distribution. Both alpha and beta parameters are unknown, so the identifications should be established. These identifications help to construct the theoretical Beta distribution using random generator (Sinharay, 2010). Both identifications for alpha and beta parameters are:

$$\hat{\alpha} = \bar{x}\left[\frac{\bar{x}(1-\bar{x})}{s^2-1}\right]$$
$$\hat{\beta} = (1-\bar{x})\left[\frac{\bar{x}(1-\bar{x})}{s^2-1}\right]$$

As the result shown, the ECDF descriptive statistics concludes that the empirical distribution is not consistent with the theoretical Beta distribution. Given that the sampling size is larger than 200,000, this distribution is less likely to alter when the size increases due to Law of Large Number. This score is nonparametrically distributed. The permutation test can be instead used for testing the null hypothesis. The adjusted density plot is more explainable for the broader audience. As indicated by this plot, the kurtotic value in theoretical distribution is greater than this value in empirical distribution.



The null hypothesis for the permutation test is that two groups are within the same distribution. More specifically, the permutation test is to evaluate 13 different measure codes. Making this analysis more explainable is by substituting each measure code with every key term that describes what this score calculates. The alpha level for null hypothesis is originally 5%. When the Bonferroni correction is applied, this level lowers to 0.38%. This correction is to ensure that the chance of Type I Error (or Type Alpha Error) is minimal. This approach is useful for multiple comparisons of hypotheses, albeit being conservative (Freund et al., 2010).



Two key parameters are included in this analysis.

The first parameter is observe difference by averaging observed scores with particular measure code and subtracting by average observed scores without this code. The second parameter is confidence interval obtained from the permutation test.

As indicated by the plot, the confidence interval is difficult to be seen due to small standard errors. The minimal and maximal margin of errors are 0.001, and 0.002, respectively. More likely, the effect of observed differences is precise. Each interval is unlikely to alter when the sampling size changes since the sampling size for each group is 10,000 or more. But this plot is not established with causal inference since this score is based on self-reported data (CMS, 2021).

On the other hand, this information is insightful. For example, the observed difference for depressive-related measure code is higher than 60 while the observed difference for catheter-related measure code is lower than -40. Intuitively, the facilities perform excellently with treating residents who have depressive symptoms while they do not do well with residents who have catheter inserted and left in their bladders. The observed difference for treating residents who receive antipsychotic medication is near zero.

The second dataset called Provider Information has over 80 features and at least 15,283 entities. At least 70 features are usable for prediction, albeit the variety of predictive power. The challenge with this dataset is that several features are redundant (i.e., some are perfectly correlated) and feature selection requires to rely on several automation techniques. Secondly, there is potential leakage in this dataset. The leakage refers to which the features in the training process do not exist when integrating the production, in turn, causes the predictive scores to overestimate. This is a common mistake in data science.

For example, the feature called total weighted health survey score is in use to predict the health inspection rating but this feature will not exist when the predictive model is being deployed. The objective is to create an insightful analysis, rather than attempt to maximize prediction accuracy. For that reason, the automation is to identify the large number of features that considers to be leakages. Not only that, the investigation is undertaken to ensure that several leaked features are overlooked. As a result, there are 30 leaked features found in this dataset. As a result, the number of features is reduced to 37 features.

Predictive Modeling

This dataset has undergone complete data quality analysis, wrangling, and exploratory data analysis in order to ensure that it is cleaned and none of features are leaked. The next step is to use two different automations to identify predictors that help to yield higher model performance. The threshold for modeling prediction should explain at least 80% of the variance for the target variable. More importantly, this condition should be met when a model generalizes to unseen data. Not only that, the model should be well-calibrated. The calibration refers to the difference between predicted and actual values. For that reason, this analysis separates this dataset into training set, validation set, and testing set by 60-20-20.

Two automations are the least absolute shrinkage and selection operator (lasso) regularization, and Bayes optimal feature selection. The lasso regularization is, especially, a popular method that proves to be successful in data science. Its function is the residual sum of squares plus the penalty almost similar to ridge function. But the penalty is lambda multiplied by sum of absolute value of coefficients. This function is robust to the outliers, but it is not differentiable due to piecewise step function in derivative (Boehmke, 2021). This algorithm has the ability to adjust coefficients of unimportant predictors to zero.

The second automation performs a similar task but it is an iterative process by balancing its

needs of exploration and exploitation depending on the probabilistic model. This approach is capable of efficiently addressing complex derivative-free problems. Three functions are in use to search hyperparameters. There is an objective function that possesses one true shape hidden in the model. This shape is not completely observable. Acquiring its true shape needs extensive time of iterative process, which is expensive to compute. This function can only reveal several data points. The surrogate function is a probabilistic model that is built to exploit what is known while the acquisition function is to calculate a vector of hyperparameters that likely yield higher local maximum of objective function. One of these studies shows that this approach has the ability to reduce the number of features and yield higher model performance, competing with many state-of-the-art methods (Ahmed et al., 2015).

Using both approaches make more sense due to the large number of features. There are 36 features in total. In other words, the total possible combinations are 68,719,476,736. Based on the analysis, the best lambda value for lasso is 0.005 because the R-squared score is 32.19% for the validation set. The Bayes optimal feature selection is to search the highest local maximum of R-squared score using binary vector. As a result, this score is 32.87%. However, the number of features for that approach in total is 21 while the number of features in total for lasso is 16. The simpler model is better and less likely to overfit.

Capturing more variance for the target variable, the light gradient boosting model with regularization is in use. Prior to explaining more results, the concept of this algorithm needs to be elaborated. The gradient boosting model is, by principle, to make a prediction using the training set and its residual error, rather than by adjusting the weight of data points. This process is iterative, and generally, unparallelizable. Two hyperparameters for this algorithm are the maximal depth of decision tree and the number of sequential trees. The learning rate is the regularization parameter that serves as a buffer. This parameter is to slow down the speed of learning speed that could otherwise lead this model to overfitting (Park and Ho, 2018). In the final stage, this model makes a prediction by summing up all sequential trees.

There is a distinction between the light gradient boosting and XGBoost. The first algorithm is leaf-wise while the second algorithm is level-wise tree growth. The leaf-wise tree growth has uneven splits of nodes that, unfortunately, may more be prone to overfitting. At least, the maximal depth is in use to control this risk. On the other hand, this algorithm is more memory-efficient and it makes parallelizability possible (Khandelwal, 2017).

But unfortunately, when the Bayesian search theory is in use to optimize this algorithm, the R-squared score is 43.47% for validation set at maximum. This score is 41.92% for the testing set while the adjusted R-squared score is 41.62%. The mean absolute error (MAE) score is 0.77. This prediction model can explain less than a half variance for the target variable, which is well below the threshold. This model is not well calibrated, either. For example, if the model predicts that the health inspection rating is 3.5, the actual score may fall somewhere between 2.73 and 4.27. This error score indicates how large the error is by averaging the difference between predicted and actual values.

Econometric Method

To the extent that the causal relationship fails to establish, the previous discussion indicates the limits of this approach. This approach makes an effort to explain the nature of relationship by testing the model against data. In this section, the analysis turns to the model side instead of searching the data pattern. In other words, addressing this problem is by adopting the econometric approach.

In social science, the R-squared score falling between 0.2 and 0.3 is considered to be accept-

able. Some suggest that the minimal requirement for this approach is that the adjusted R-squared score must be positive. But in technical sense, the stipulation with model misspecification is much relevant, following by Gauss Markov assumptions. Even though the data science approach selects features that yield higher model performance, the econometric approach is human-centric. This approach begins by writing the causal model and its underlying assumptions. In the nutshell, this approach is to test data against model.

As mentioned earlier, Gauss Markov assumptions are fundamental of the econometric approach, though the causal model may depart from some assumptions, as the following: linearity in parameters (A1), no perfect collinearity (A2), zero conditional mean error (A3), homoskedasticity and no serial correlation (A4), and normality of the error (A5). This causal model does not follow A1 and A4 assumptions. The endogenous variable (or target variable) is limited between 1 and 5 so that it is Limited Dependent Variable (LDV). The multiple ordinary least squares (OLS) are not suitable for LDV. The popular solution is logarithmic transformation is, unfortunately, misspecified. This particular model misspecification is called Duan's Smear.

$$
\begin{aligned}
&log(y_i) = x_i\beta + u_i \\
&e^{log(y_i)} = e^{x_i\beta + u_i} \\
&E[y_i|x_i] = e^{x_i\beta} \int e^{u_i} \partial u_i
\end{aligned}
$$

This analysis demonstrates the mathematical issue with assumption that error term is independent of exogenous variable (or predictor). The logarithmic transformation may be in violation of A3 due to endogeneity. This term refers to which the exogenous variable is correlated with error term.

$$e^{x_i\beta} + u_i \not\approx e^{x_i\hat{\beta}} \frac{1}{n} \Sigma_{i=1}^{n} e^{\hat{u}_i}$$

As indicated by this equation, the logarithmic transformation is not approximately equal to true population logit regression. The nonlinear least square may solve this problem. For this analysis, the nonlinear least square is called Probit with Quasi-Maximum Likelihood Estimate condition (QMLE). This model is more efficient to heteroskedasticity. The heteroskedasticity is the variance of residual term that is not constant through the regression. For that reason, based on the standard asymptotic theorems, the causal model is consistent and asymptotically normal where $V \not\propto A$. The condition for consistency is:

$$\sqrt{n}(\hat{\theta}_{QMLE} - \theta_0) \sim^a N(0, A^{-1}VA^{-1})$$

where

$$
\begin{aligned}
&A = -E[H(w_i, \theta_0)] \\
&V = E[s(w_i, \theta_o)s(w_i, \theta_0)']
\end{aligned}
$$

The Hessian matrix $H(w_i, \theta)$ is the definiteness of the matrix of second-order partial derivatives that identify the type of extremum while the Score vector $s(w_i, \theta)$ is to maximize or minimize using the fist partial derivatives. The consistent condition is established due to the inconstant variance of the residual term. The Breusch-Pagan test confirms that the null hypothesis that the error variances are all equal is rejected at the significance level. In other words, at least one coefficient is heteroskedastic (Goldstein, 2021). The causal model is contemporaneously exogenous, rather than strictly exogenous. This model is using this Probit function equal to $\phi(.)$ is established, as follows:

$$y_i = \phi(\beta_0 + \beta_1 bed_i + \beta_2 hr_i + cond_i\beta) + u_i$$

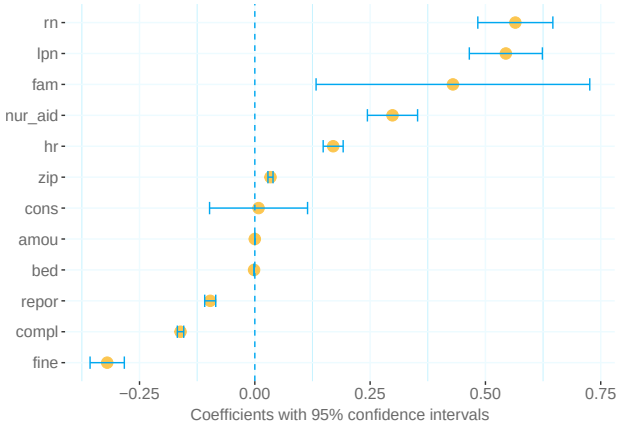
The bed_i represents the number of certified beds while the hr_i represents the amount of hours per residents each day. The $cond_i\beta$ is a vector of conditions that include the level of quality care and competency the staff provides, type of location area and kind of environment residents live in, and kind of vulnerability residents have. The proxies for competency using dummy variables are job titles like, for example, nurse aids, license practical nurses, and registered nurses.

The proxies for quality are total number and charged amount of fines, number of substantiated complaints, and number of facility reported incidents. The proxy for environment is installed automatic sprinkler system while the exogenous variable for location area is National Area regional code. The proxies for vulnerability using dummy variables are family and abuse icon. However, as indicated by Wald test, the null hypothesis that the coefficients of installed automatic sprinkler system and abuse icon both equal zero are failed to reject. Both variables are removed.

Actionable Insights

When the causal relationship is established, each coefficient of exogenous variables is statistically different from zero at 5% alpha level using Huber-White (HCO covariance type) robust standard errors. The robust standard errors are in use when the homoskedasticity does not exist. The pseudo-R-squared score is 32.54%, which is considerably high for social science. The p-value for LLR is below 5% alpha level. Intuitively, the null hypothesis for LLR is that the fit of the intercept-only model and causal model are equal is rejected.

At least one coefficient of exogenous variable, intuitively, has important contributions to this model. Except for the p-value for intercept, each p-value for every coefficient is below 0.01. The p-value for intercept equals 0.881, so this coefficient is not different from zero.



Being compared to all parameters, the registered and license practical nurses have significant effects on health inspection rating as expected. For example, when the registered and license practical nurses are present in facilities, these facilities have a higher chance of receiving at least 4 point in health inspection rating. Interestingly, with the nurse aides being present in facilities, their positive effect on health inspection rating is relatively less.

When the exogenous variable hr_i interacts with license practical nurses, registered nurses, and nurse aides, the average partial effects (APEs) are altered. Prior to discussing this analysis more, the concept of APE needs to be introduced. One most common approach in the econometric analysis is that partial effect is in use. But this effect is not usable due to the nonlinear causal relationship. The APE is a change in x_i on y_i. This value is extracted by averaging partial effects.

When the interaction term is applied, the APE in respect to hr_i is 0.24. However, more surprisingly, the coefficient of hr_i interacting with nurse aides is -0.04 with margin of error equal to -0.05. The increase in number of hours that the nurse aides spend with residents each day has little to no impact on health inspection rating. The coefficient of hr_i interacting with license practical nurses is, however, -0.0881 with margin of error equal to 0.07. The incremental increase in the number of hours that license practical nurses spend has a negative impact on health inspection rating at the significance level. The registered nurses still have a positive impact on this rating at the significance level.

Another important finding is that having family members on the council has impactful implications on this rating. The coefficient of this variable is 0.43 with margin of error equal 0.30. But the coefficient of bed_i interacting with this variable is -0.011 with margin of error equal to 0.008. The increase in the number of certified beds, *cerberus paribus*, lead to the decrease in this rating. This decline is exacerbated when family members

get involved.

Succinctly, if license practical nurses spend more time with residents, the health inspection rating point is slightly lower but it remains positive. The increase in the number of certified beds is, mainly, responsible for negative effects on health inspection rating. The increase in the number of certified beds with registered nurses does not improve the chance of receiving good health inspection rating at the significance level. The license practical nurses still have more negative effects on this rating. Tethered to this causal analysis, the measure quality for depressive-related code is higher while this score for catheter-related code is lower. There is potential linkage between roles of nurses. For example, the role of registered nurses is to administer medication and treatments while the role of license practical nurses comforts the residents and provides the basic cares including inserting catheters.

Future Work

As mentioned earlier, there is potential linkage between the roles of nurses and measure codes. However, the linkage is indetermined since the MDS and Provider Information datasets are incompatible based on the record linkage. The future study hopefully has access to CMS databases that allow to establish the linkage to estimate what effects would be. Such information can determine how to properly address the moral hazard problem in an implicate manner. The increase in monitoring is one of many ways to reduce this behavior. This study shows that when all other variables are equal, registered nurses have positive impacts on health inspection rating while the licensed practical nurses have negative impacts. In other words, this causal analysis may have another solution.

For example, the CMS pilot program could be established by either helping provide financial aid for LPN-to-RN career pathway or promoting awareness of existing programs. The randomized controlled trials (RCT) should be in use to identify if either approaches have real impacts on this rating, and hopefully, alleviate moral hazard. An alternative study is the existing dataset that can establish a causal relationship by comparing particular facilities that have undergone LPN-to-RN training program with these that do not.

Besides this, the direct solution to inflation in self-report data is much possible by adopting AI solutions. For example, if the data audit is undertaken to identify inflation in self-report data and adjust them, this information can be in use to create either machine learning or deep learning in order to make data audit cost-effective and reduce moral hazard. This future study would make a significant contribution.

Conclusion

The causal relationship shows that the number of certified beds decreases the chance of receiving good health nurse rating. Having license practical nurses present affects this rating in several different ways. However, this role has negative impacts when interacting with both amount of hours spending with residents each day and number of certified beds while the registered nurses have positive impacts on this rating. The financial or awareness promotional pilot program needs to be studied using RCT to determine if this program helps to improve this rating, and hopefully, alleviate moral hazard. However, the most direct solution is AI solution that could automate to identify whether the inflation in self-report data exists. This approach can also be tested to see if it helps improve public trust from consumers in facilities.

References

"Technical Details." *Nursing homes including rehab services*, the Centers for Medicare and Medicaid Services, Sep. 2021. https://data.cms.gov/provider-data/topics/nursing-homes/technical-details#health-inspections

Adams, Ryan P. "Practical Bayesian Optimization of Machine Learning Algorithms." School

of Engineering and Applied Sciences Harvard University, 2012.

Ahmed, S., Narasimhan, H., and Agarwal, S. "Bayes Optimal Feature for Supervised Learning with General Performance Measures." arXiv, 2015. http://auai.org/uai2015/proceedings/papers/72.pdf

Boehmke, B. "Regularized Regression." *UC Business Analytics R Programming Guide*, University of Cincinnati, 2021. http://uc-r.github.io/regularized_regression#lasso

Datta, A., Fredrikson, M., Ko, G., Mardziel, P., and Sen, S. "Proxy Discrimination in Data-Driven Systems." *Theory and Experiments with Learnt Programs*, arXiv, 25 Jul. 2017. https://arxiv.org/abs/1707.08120

Freud, RJ., Wilson, WJ., and Mohr, DL. "Inferences for Two or More Means." *Statistical Methods, Third Edition*, Academic Press, 2010. https://doi.org/10.1016/B978-0-12-374970-3.00006-8

Goldstein, Nathan. "Lecture 1. Foundations of Microeconometrics." *Microeconometrics*, Zanvyl Krieger School of Arts and Sciences Johns Hopkins University, 2021.

Kaynig-Fattkau, V., Blitzstein, J., and Pfister, H. "CS109 - Data Science." *Decision Trees*, Harvard University, 2021. https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=c22cbde8-94dd-42ad-86ef-09 1448ad02e4

Khandelwal, P. "Which algorithm takes the crown: Light GBM vs XGBOOST?" Analytics Vidhay, 12 Jun 2017. analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs-xgboost/

Park, Y. and Ho JC. "PaloBoost." *An Overfitting-robust TreeBoost with Out-of-Bag Sample Regularization Techniques*, Emory University, 22 Jul 2018. http://arxiv-export-lb.library.cornell.edu/pdf/1807.08383

Silver-Greenberg, Jessica, and Robert Gebeloff. "Maggots, Rape and Yet Five Stars: How U.S. Ratings of Nursing Homes Mislead the Public." *How U.S. Ratings of Nursing Homes Mislead the Public*, New York Times, 13 Mar. 2021.

Sinharay, S. "Coninuous Probability Distributions." *The International Encyclopedia of Education*, Elsevier Science, 2010. https://doi.org/10.1016/B978-0-08-044894-7.01720-6

Streiner, David L. "Unicorns Do Exist: A Tutorial on"Proving" the Null Hypothesis." *Research Methods in Psychiatry*, The Canadian Journal of Psychiatry, Dec. 2003.