Best Friends Animal Society: Report on Estimation Validity of Regression Modeling

Courtney Paulson, Ph.D.[1]

Aim of this Report

This report aims to evaluate the benefits and limitations of the Best Friends Animal Society's approach to shelter data modeling. In particular, this report focuses on Best Friends' estimation of missing shelter data and provides an overall assessment of the validity of the methodology.

Introduction and Overview of Validation

For this validation, models were evaluated with two data sets: a pre-pandemic data set from prior to the COVID-19 pandemic in March 2020 and a post-pandemic data set from after March 2020. In the provided methodology, a randomized 80% of the data in each set was used as training data, while the remaining 20% was separated as testing data.

Three models were constructed for each of the training data sets, using the most recent rolling 12 months of data for a particular shelter:

(1) A model to predict gross intake of a shelter,
(2) A model to predict net intake of a shelter, and
(3) A model to predict non-live outcomes, where *Non-Live Outcomes* is defined as animals lost through euthanasia, died in care, or lost in care.

All models are log-linear models, where the predicted (response) variable is the log of gross intake, the log of net intake, and the log of non-live outcomes respectively. As part of the verification process, the appropriateness of using a logarithmic transformation was also evaluated by checking the skew of the raw response variables; see Appendix for further discussion, including data visualizations for the distribution of each response variable.

In all cases, the model used is a robust regression model[2]. All models were run on training data and predicted on testing data. Logarithmic predictions were generated from the model and then exponentiated to obtain direct estimates. To predict non-live outcomes, a threshold of 10% was applied to the direct estimates. This adjusts for the inability of the log-linear model to predict exactly 0, and it also accounts for the 90% no-kill benchmark (as "killed" animals is defined as the number of animals that die at a particular shelter above the 90% no-kill benchmark).

---

[1] Assistant Professor of Statistics, University of New Hampshire; Courtney.Paulson@unh.edu

[2] These models were operationalized through the `rlm` function in the `MASS` package of the `R` software. The models evaluated specifically use the "MM" method of the rlm function.

Benefits of Estimation

The estimation here particularly benefits from the use of a log-linear robust regression, which is uniquely appropriate for this problem given the skewed intake variables and the higher probability of outliers in shelter data (both evident in the histograms of the Appendix, for example). Converting the intake variables to logarithmic variables improves the prediction accuracy of the model by increasing the variability of the responses, and the use of a robust regression model here reduces the outlier effect to provide a stronger basis for predictions. Model accuracy was significantly reduced when testing non-robust regression models, such as traditional least-squares model estimation, and removing the logarithmic conversion.

The most obvious benefit of the model estimation, however, is in the testing data accuracy itself. For verification, this report considers two prediction metrics:

(1) Accuracy: The percentage of shelters in the testing data accurately predicted as kill or no-kill. An accuracy of 100% would be ideal, with no false positives (shelters predicted to be no-kill but are not) and no false positives (shelters not predicted to be no-kill but are).

(2) Absolute Percent Difference: The absolute percentage difference between the actual testing data values and the predicted testing data values. Absolute Percent Difference is calculated for Gross Intake and Net Intake[3].

The baseline threshold for accuracy in this data is roughly 55% across all testing data sets[4]. Although 100% would indicate a perfect model, an accuracy above 55% indicates the model is identifying no-kill status more often than expected. Further, in the absence of a perfect 100% accuracy, false positives and false negatives should be balanced, indicating a model is not biased toward predicting a particular category.

For fair model verification, accuracy and percent difference are calculated using an 80% training/20% testing data split over 100 randomized runs of models and predictions for both pre- and post-pandemic data[5]. The average (mean) results across all 100 runs are presented in the following table:

Metric	Pre-Pandemic	Std. Error	Post-Pandemic	Std. Error
Accuracy	80.55%	0.25%	86.77%	0.22%
False Negative Rate	11.58%	0.24%	7.65%	0.17%
False Positive Rate	7.87%	0.16%	5.58%	0.14%
Abs. Percent Difference (Gross)	48.50%	0.69%	12.22%	0.57%
Abs. Percent Difference (Net)	34.17%	1.51%	18.77%	0.76%

[3] Note that percent difference is not calculated for animals killed, since many shelters have zero animals killed by virtue of being a no-kill shelter and thus invalidate percent difference calculations.
[4] Approximately 55% of the shelters in the randomized testing data sets are collected “no-kill” shelters, indicating an accuracy of 55% is possible by simply guessing all shelters as “no-kill.”
[5] That is, 100 separate runs of choosing a random 80% of data as training and 20% of data as testing, with predictions calculated on the 20% testing data for each run. Results were also run with 70%/30% and 90%/10% data splits and found to be consistent.

Overall, as expected the more recent post-pandemic data shows a higher accuracy than the less recent pre-pandemic data. The standard errors are all remarkably low, indicating the model performance is highly consistent. The false positive rates and false negative rates are both low and roughly equal, indicating the model not only has high accuracy but also low bias. The model appears to be uniquely suited to predict the status of a shelter ("no-kill" or not) when missing status but not the remaining variables.

Of particular note, the absolute percent differences for the post-pandemic data are extremely low for a robust regression model with gross and net intake values exceeding 40,000 in the data. Although the goal of predicting no-kill shelters is highly successful (demonstrated at nearly 87% accuracy on average), the total animals predicted for intake is off by only 12% to 19% (again, on average[6]). For a shelter with a gross intake of 2000 animals, the model would predict a gross intake of 1760 to 2240 animals. This percent difference means categorizing shelters by size is highly accurate as well, as the actual difference is unlikely to separate large shelters from small shelters.

Limitations of Estimation

Although the model prediction results above demonstrate the model is currently highly effective in predicting both the status of a shelter as no-kill as well as the total intake to the shelter, the model does appear to have some minor limitations in its current form:

(1) Although robust estimation offers a strong basis for modeling this problem, the nature of the data might warrant consideration of a time series or an econometric "pre- and post-" model given the separation of the data into before and after the pandemic. Initial testing of both time series and econometric modeling did not yield any immediately promising results, however, which indicates that pre- and post-pandemic data is perhaps best modeled separately.
(2) The data appears to be reasonably large and complete. For shelters with less complete data, these regression models may not be as appropriate (though any model will fail with sufficiently inaccurate or incomplete data, so this is unlikely to present significant issues given Best Friends' current data collection approach). As noted above, the accuracy of the model indicates a high success rate in predicting missing "no-kill" status if all other variables are present.
(3) As expected with log-linear regression, the model algorithm does fail to converge occasionally with particular randomized subsets of the data. This occurred in less than 1% of cases and is unlikely to present issues for general modeling but should be noted as a possible issue.
(4) Also due to the use of the log-linear regression, a value of 1 is added to the shelter totals to create a non-zero value to be used in mathematical modeling (as the log of 0 is a negative infinite value). This does not impact modeling results in any way, but it again should be noted in reporting as the adjustment may cause confusion, particularly since it adjusts the total of some non-live outcomes from 0 to 1.

[6] Also of note, the (average) median percent difference values are 12.51% for the post-pandemic gross intake and 19.43% for the post-pandemic net intake. This indicates the averages here are not skewed by accurately predicting small shelters but inaccurately predicting large shelters or vice versa.

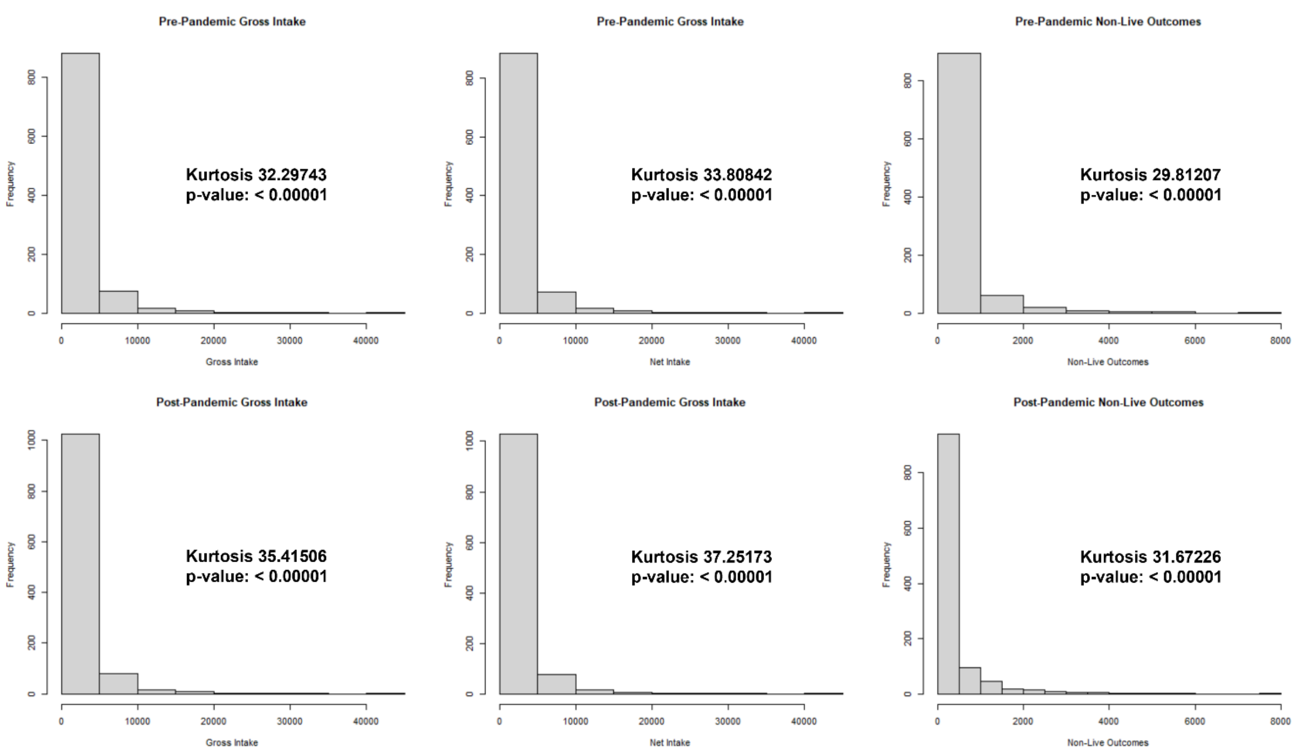
Overall Assessment of Estimation

The regression models used here demonstrate significant improvement over baseline accuracy, including highly consistent predictions for both no-kill status and total intake. The choice of a log-linear robust regression adjusts for the skewed nature of the response variables, and the model demonstrated improved accuracy over non-robust regression modeling and raw data without the logarithmic transformation. More importantly, the model results were robust to changes in data split and showed statistically small variation in accuracy across multiple runs.

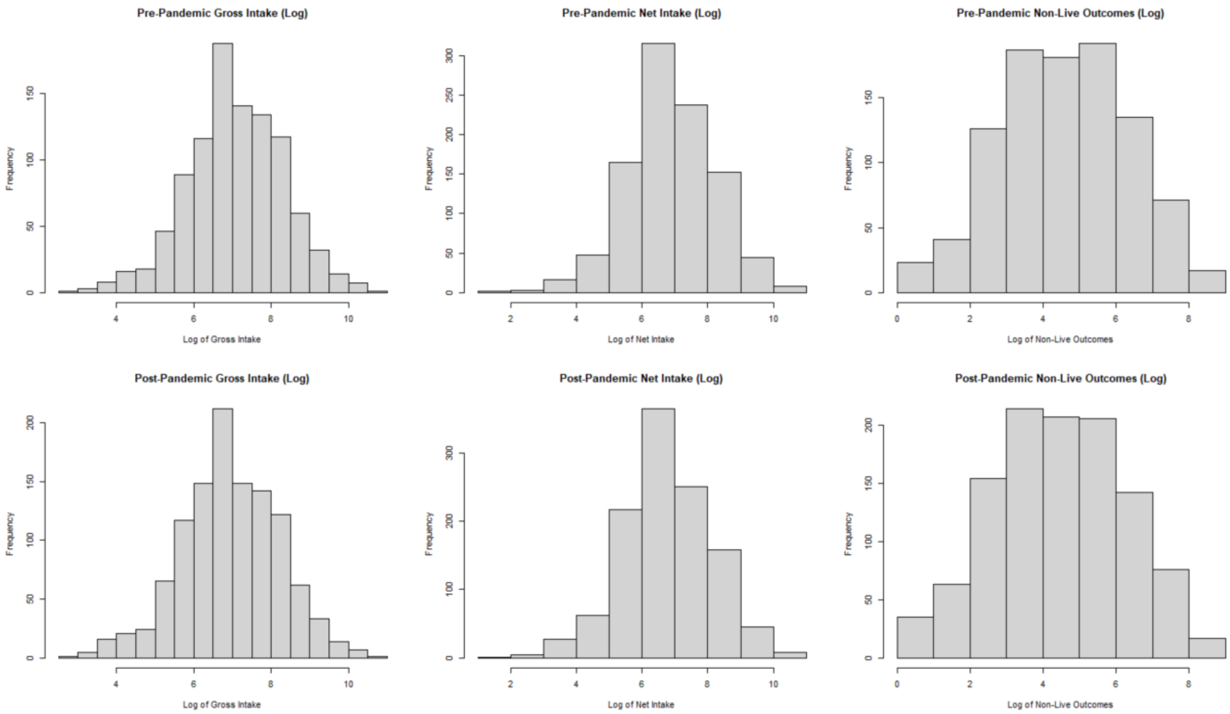
Best Friends Animal Society has developed a highly accurate model with strong confidence in model results. The model appears to be well-validated in addition to demonstrating strong performance in practice. Model verification identified no significant concerns for implementing the model across new or expanded data for analysis, particularly in predicting missing shelter designation ("no-kill" or not).

Appendix: Appropriateness of Log-Linear Modeling

To evaluate the appropriateness of using a log-adjusted predictor variable for both gross intake and net intake, histograms for each of the four variables were generated to check the skew of each variable. The figure below shows all four variables with severe right skew, indicating a need for logarithmic adjustment; the calculated kurtosis for all four is well above the standard threshold of 3.00, and the Jarque-Bera test for normality further verifies all four variables with a p-value of almost zero (rejecting the null hypothesis that the data is normal and unskewed).



For comparison, the same figure was also generated with the log transformation of each variable:



Log-linear modeling is a more appropriate choice and improves the quality of predictions.