Logistic Regression with Measurement Error: Robust Optimization vs Regularization, Applied to Predicting Jerky Treat Toxicity

Nicholas Renegar
Operations Research Center, Massachusetts Institute of Technology, 1 Amherst St, Cambridge, MA 02142
renegar@mit.edu

May 15, 2019

Abstract

Over the last 12 years, there has been an epidemic of acquired fanconi syndrome in dogs, from the consumption of toxic jerky treats, leading to 6,200 illnesses and over 1,140 deaths. Despite the U.S. Food and Drug Administration's best efforts, they have been unable to identify a root cause. We investigate the problem of labeling jerky treats as toxic or nontoxic, based on measures of how dog kidney cells change in-vitro, when an extract of the treat is introduced. Because the experimental setup has covariate measurement error, we consider logistic regression with both regularization and robust optimization to see if we can improve over the nominal model. While robust optimization offers the best results, the performance improvement, as measured by area-under-the-curve, is marginal. We explore how regularization affects the fitted model coefficients in logistic regression, and offer a hypothesis for this.

1 Introduction

In 2007, the U.S. Food and Drug Administration (F.D.A.) began seeing reports of acquired Fanconi syndrome in dogs, resulting from the consumption of pet jerky treats[U.S18]. Fanconi syndrome is a rare, potentially-fatal kidney disease, resulting in the kidneys filtering out beneficial substances normally absorbed into the bloodstream. While Fanconi syndrome is typically a genetic disease, by 2015 the F.D.A. had received complaints related to jerky treat induced illnesses in 6,200 dogs, including over 1,140 deaths.

The complaints have related to a number of types of jerky treats (chicken, duck, and sweet potato), manufactured largely in China, but present in U.S. treats as well. So far, the F.D.A. has failed to identify a cause for the acquired Fanconi syndrome. Unlike pharmaceuticals, where it is easy to see from a mass spectrometry scan where the recipe differs from what is expected, testing jerky-treats is complicated by the food matrix. For chicken jerky treats, in particular, over 2,000 chemicals are identified, making it very hard to tell what the culprit molecule or molecular interactions might be [1].

To combat these illnesses and deaths, the FDA is considering alternative (non-traditional) methods of approaching the problem. For this purpose, they have entered into a collaboration with our MIT food safety team (representing the Operations Research Center, Sloan, the Center for Biomedical Innovation, and the Sinskey Lab in the department of biology), and provided us with labeled toxic and non-toxic jerky treats, as determined by the history of complaints and autopsies of deceased pets. Our goal is to take a jerky treat, and accurately predict whether it is toxic or non-toxic, to serve as a data point for product inspections and recalls, as well as to help further work into determining the root cause of toxicity. To this end, we have developed an experimental procedure roughly as follows: we take an extract of the jerky treat, introduce it into in-vitro dog kidney cells, and using a machine we measure a number of features describing the changes in the kidney cells.

[1] Based on conversations with the FDA and work by the Sinskey Lab at the MIT department of biology

From these cellular features, we would like to predict toxicity. This is complicated by the fact that the features have a significant amount of measurement error. That is, given the same treat extract, when we measure the cellular changes to two samples of in-vitro dog kidney cells, we observe moderately different cellular changes. We are currently having success with nominal logistic regression, ignoring measurement error. We investigate the extent to which we can improve our predictive results, by applying logistic regression with robust optimization (RO) and regularization.

1.1 Literature Review

In statistics, regression with measurement error (often referred to as the study of 'measurement error models' or 'errors-in-variables models') has been an important problem historically. For example in Casella and Berger's *Statistical Inference*, which is the standard textbook for first-year PhD students in statistics, the entire 12th and final chapter is devoted to the study of regression with measurement error [CB01]. In many of the sciences and real-life domains, such as healthcare, biology, chemistry, and physics, measurement error is an inherent and unavoidable aspect of the work.

1.1.1 Linear Regression

For linear regression the measurement error model, is typically stated as:

$$y = \beta^T v + \Delta_y$$

$$v = x + \Delta_x,$$

where the outcome y and measurements x are observed, the true independent features v are unobserved, and Δ_y, Δ_x are unobserved noise terms. Under this model, there are a number of simple results showing the parallels between robust optimization and regularization. For example in class, we saw for the uncertainty set defined by the induced matrix norm, where $\mathcal{U}_1 = \{\Delta \in \mathbb{R}^{n\times m} | \|\Delta\|_{q,p} \leq \rho\}$, we have that:

$$\min_\beta \max_{\Delta\in\mathcal{U}_1} \|y - (X+\Delta)\beta\|_p = \min_\beta \|y - X\beta\|_p + \rho\|\beta\|_q.$$

Similarly for $\mathcal{U}_2 = \{\Delta \in \mathbb{R}^{n\times m} | \|\Delta\|_{p-F} \leq \rho\}$, we have when $\frac{1}{p} + \frac{1}{p^*} = 1$ that:

$$\min_\beta \max_{\Delta\in\mathcal{U}_2} \|y - (X+\Delta)\beta\|_p = \min_\beta \|y - X\beta\|_p + \rho\|\beta\|_{p^*}.$$

Similar parallels between regularization and stochastic optimization also exist, and are sometimes mentioned as an aside in statistics and machine learning courses. For example in Bishop's *Pattern Recognition and Machine Learning*, Question 3.4 is used to show that for linear regression with measurement error, where the covariate error is i.i.d. gaussian with $\Delta_{x_i} \sim N(0, \beta), \forall i$, that minimizing the expected mean square error with respect to the measurement error, is equivalent to a linear ridge regression with penalty parameter $\frac{\beta}{2}$.

1.1.2 Logistic Regression

For logistic regression with measurement error, we have the generalized linear model:

$$\begin{aligned} \mathbb{P}(Y = y|V = v) &= \frac{1}{1 + \mathrm{e}^{-y(\beta^T v + \beta_0)}} \\ V &= X + \Delta_X, \end{aligned} \tag{1}$$

where the outcome $Y \in \{-1, 1\}$ and measurements X are observed, the true independent features V are unobserved, and Δ_X are unobserved noise terms.

In contrast to linear regression, the problem of selecting a logistic regression model with measurement error is very difficult from a traditional, statistics-driven approach. It has been shown that no closed-form solution to maximum likelihood exists [RHPS03]. In addition, it has been shown that the likelihood function is non-convex, making it impossible to solve with gradient descent or other traditional convex optimization methods [HJZW+]. Ignoring the measurement error, and solving maximum likelihood estimation in the traditional way for logistic regression, has been shown to produce biased model coefficient estimates [Bar10].

As a result of these difficulties, a lot of the statistics literature is focused on correcting for this bias. This bias correction is a difficult task, and there are highly-cited papers that use corrections that have guarantees for creating 'less-biased' estimates, as opposed to the ideal unbiased estimates [SC85]. This correction approach has been criticized as well. For example, it has been shown that this correction is imperfect when the methods to accurately assess measurement error are imperfect, and in such cases the corrections can even produce worse results [SSM97].

There have also been methods for non-parametric maximum likelihood estimation that have achieved decent results in practice, but they require full information about the distribution of covariate noise [RHPS03]. Finally, there have been statistics-based approaches to robust logistic regression and classification, to remove the effect of outliers and implicitly adjust for measurement error, as inspired by traditional robust regression (Note: this phrasing does not refer to robust optimization as studied in class) [FMXY14]. This work involves an algorithm that sequentially removes outliers and solves a linear program, to achieve less-biased model coefficients. While regularization seems to be the most common approach in practice, and is mentioned in several of these papers, almost nothing concrete seems to have been said about it's guarantees in this setting [Tib96].

Robust optimization approaches, to logistic regression with measurement error, have only started appearing in the literature recently. The most relevant work to this paper was done in *Robust Classification* by Bertsimas, Dunn, Pawlowski, and Zhuo, and it shows that a robust approach to logistic regression is computationally practical, and often improves upon the nominal model both for synthetic and empirical data [BDPZ19]. The paper also shows that the robust approach can outperform regularization for many empirical datasets, although the results are mixed, with regularization sometimes doing better. Relatively little has been said about the structure of the data, and how it relates to the results, so it is not clear what will be the case for the jerky treat data. Robust optimization for logistic regression has also been studied in the case of bounded measurement error and bounded uncertainty sets, and shown to improve upon lasso [HJZW$^+$]. Finally there is preprint work being done on distributionally robust logistic regression under measurement error, although this is beyond the scope of this paper and will not be studied in the empirical sections [HJZW$^+$].

1.2 Contributions

This paper is application-focused, and seeks to create a better predictive method for the F.D.A. to identify toxic vs. non-toxic jerky treats. We will compare the nominal logistic regression model, assuming no measurement error, to regularized approaches and robust optimization approaches.

We will first do this comparison on synthetic data, created to be structurally similar to the jerky treat data. We will use a training, validation, and testing split, to train this data across multiple hyperparemeter values (related to the penalty term for regularization and uncertainty set for robust optimization), choose the best hyperparameters on the validation set, and then evaluate out-of-sample performance on the test set. We will also evaluate the trained model coefficients compared to the nominal model, to see how different choices of robust optimization and regularization affect the fitted model coefficients.

We will then follow a similar process for the jerky treat data, to evaluate out-of-sample testing performance compared to the true toxicity labels. We investigate whether robust optimization improves upon regularized regression for this real data set, and as a result whether we can help guide the F.D.A. as they investigate jerky treats, and continue to search for the root cause of the problem.

2 Model Formulations

We test several models on the synthetic and jerky treat data sets. These models include the nominal logistic regression MLE formulation, where we assume there is no measurement error, logistic regression with two types of regularization, and robust optimization for logistic regression with three types of uncertainty sets. We fit all six models using the optimization formulations stated below. For all models, we let $Y \in \{-1, 1\}$ be the labels for the data, and we let X be the observed covariates (including measurement error).

2.1 Nominal Model

The nominal model assumes no noise in the observed covariates. From Equation 1, we see that this simplifies to the following:

$$\mathbb{P}(Y = y|X = x) = \frac{1}{1 + \mathrm{e}^{-y(\beta^T x+\beta_0)}}.$$

We fit the nominal model by MLE, and equivalently we solve for the maximum log-likelihood, which gives us the following concave maximization formulation:

$$\max_{\beta,\beta_0} \sum_{i=1}^{n} -\log(1 + \mathrm{e}^{-y_i(\beta^T x_i+\beta_0)}). \tag{2}$$

2.2 Logistic Regression with Regularization

Under logistic regression with regularization, we add an l_q norm penalty term to the maximum likelihood formulation for the nominal model. This penalty term will decrease the fitted model coefficients, compared to the nominal MLE formulation, which will help correct for the bias in our fitted model coefficients under the nominal model. We consider two forms of regularization: lasso (the l_1 norm), and ridge (the l_2 norm). The model formulations are stated below:

2.2.1 Lasso

$$\max_{\beta,\beta_0} \sum_{i=1}^{n} -\log(1 + \mathrm{e}^{-y_i(\beta^T x_i+\beta_0)}) - \lambda\|\beta\|_1. \tag{3}$$

2.2.2 Ridge

$$\max_{\beta,\beta_0} \sum_{i=1}^{n} -\log(1 + \mathrm{e}^{-y_i(\beta^T x_i+\beta_0)}) - \lambda\|\beta\|_2. \tag{4}$$

2.3 Robust Optimization for Logistic Regression

Under robust optimization for logistic regression, we consider an uncertainty set $\mathcal{U}$ of measurement errors, and select β, β_0 to maximize the worst-case log likelihood against all possible values of $\Delta_x \in \mathcal{U}$. That is, we solve the problem:

$$\max_{\beta,\beta_0} \min_{\Delta_x \in \mathcal{U}} \sum_{i=1}^{n} -\log(1 + \mathrm{e}^{-y_i(\beta^T (x_i+\Delta_{x_i})\beta_0)})$$

.

It was shown by Bertsimas, Dunn, Pawlowski, and Zhuo [BDPZ19], that when the uncertainty set is in the form $\mathcal{U} = \{\Delta_x \in \mathbb{R}^{n\times p} | \|\Delta_{x_i}\|_q \leq \rho, i = 1, .., n\}$, that the max-min formulation above is equivalent to:

$$\max_{\beta,\beta_0} \sum_{i=1}^{n} -\log(1 + \mathrm{e}^{-y_i(\beta^T x_i+\beta_0)+\rho\|\beta\|_{q*}})$$

.

For the l_1, l_2, and l_∞ norms, we have the dual norms:

$$\|\beta\|_{1*} = \|\beta\|_\infty$$

$$\|\beta\|_{2*} = \|\beta\|_2$$

$$\|\beta\|_{\infty*} = \|\beta\|_1,$$

all of which give computationally straightforward robust situations. Therefore we consider these three uncertainty sets for our analysis:

$$\mathcal{U}_1 = \{\Delta_x \in \mathbb{R}^{n\times p} | \|\Delta_{x_i}\|_1 \leq \rho, i = 1, .., n\}$$

$$\mathcal{U}_2 = \{\Delta_x \in \mathbb{R}^{n\times p} | \|\Delta_{x_i}\|_2 \leq \rho, i = 1, .., n\}$$
$$\mathcal{U}_\infty = \{\Delta_x \in \mathbb{R}^{n\times p} | \|\Delta_{x_i}\|_\infty \leq \rho, i = 1, .., n\}$$

The three robust formulations are stated below.

2.3.1 Linear Uncertainty Set ($\mathcal{U}_1$)

$$\max_{\beta,\beta_0} \sum_{i=1}^{n} -\log(1 + \mathrm{e}^{-y_i(\beta^T x_i+\beta_0)+\rho\|\beta\|_\infty}). \tag{5}$$

2.3.2 Ellipsoidal Uncertainty Set ($\mathcal{U}_2$)

$$\max_{\beta,\beta_0} \sum_{i=1}^{n} -\log(1 + \mathrm{e}^{-y_i(\beta^T x_i+\beta_0)+\rho\|\beta\|_2}). \tag{6}$$

2.3.3 Box Uncertainty Set ($\mathcal{U}_\infty$)

$$\max_{\beta,\beta_0} \sum_{i=1}^{n} -\log(1 + \mathrm{e}^{-y_i(\beta^T x_i+\beta_0)+\rho\|\beta\|_1}). \tag{7}$$

3 Synthetic Data & Analysis

We first evaluate the performance of the six models from Section 2 on synthetic data. Because the performance of regularization compared to robust optimization for logistic regression has been shown to vary considerably by dataset, we want to construct the synthetic data to be structurally similar to the jerky treat data of interest.

After cleaning the jerky treat data, we perform data exploration, and learn how we should construct our synthetic data. During data cleaning, we normalize so that the four covariates, averaged across the eight experimental replicates of each treat, and then compared across treats, have mean 0 and standard deviation 1. We also see that the correlation across the four covariates is low for each treat. This is roughly equivalent to the true covariates $v^{(1)}, (v^2), (v^3), (v^4)$ from Equation 1, being drawn from i.i.d. $N(0, 1)$ distributions. We next look at the eight replicates of each jerky treat, to determine the measurement error. We find that $\Delta_{x^{(i)}} \sim N(0, 1), \forall i$, and that the $\Delta_{x^{(i)}}$ are very lowly correlated. Fitting the nominal logistic regression model, we find that the coefficients β vary a moderate amount, and we see that our dataset is unbalanced, with roughly 75% of the samples being toxic.

As a result, we create our synthetic data as follows. We have four covariates, and the true values $v^{(i)} \sim N(0, 1), \forall i = 1, .., 4$. The measurement errors are also assumed to be i.i.d. standard normal, to be consistent with the real data, so $\Delta_{x^{(i)}} \sim N(0, 1), \forall i = 1, .., 4$. We then have the observed $x^{(i)} = v^{(i)} + \Delta_{x^{(i)}}$. We set the true model coefficients to be $\beta_1 = 3, \beta_2 = 1, \beta_3 = -1, \beta_4 = -3$, to make some coefficients more predictive than others. We also set $\beta_0 = 3$ to make our dataset unbalanced, with more toxic than nontoxic samples. We then randomly draw y such that $\mathbb{P}(y = 1) = \frac{1}{1+\mathrm{e}^{-(\beta^T v+\beta_0)}}$.

3.1 Model Setup

We create 400 samples of the synthetic data, as described above, and use a 50/25/25 training, validation, and test split. That is, we train our models for various values (of the form 2^n) of the hyperparameters on the 200 training set records (λ for penalized models, and ρ for robust optimization models), then select the best hyperparameter for each model, based on performance for the 100 validation set records, and finally evaluate the model out-of-sample on the 100 testing records. We do this for the six models described in Section 2. We fit the following hyperparameters based on maximizing AUC for the validation data set, as shown in Table 1.

Table 1: Synthetic Data - Hyperparameters

Model	Hyperparameter
Regularization - Lasso	$\lambda = 2^{-1}$
Regularization - Ridge	$\lambda = 2^{1}$
Robust Optimization - Linear Uncertainty	$\rho = 2^{-2}$
Robust Optimization - Ellipsoidal Uncertainty	$\rho = 2^{-5}$
Robust Optimization - Box Uncertainty	$\rho = 2^{-6}$

Table 2: Synthetic Data - Model Coefficients

Model	β_0	β_1	β_2	β_3	β_4
Nominal	1.634	0.889	0.329	-0.188	-0.961
Regularization - Lasso	1.575	0.858	0.309	-0.165	-0.927
Regularization - Ridge	1.568	0.859	0.315	-0.174	-0.926
Robust Optimization - Linear Uncertainty	0.788	0.799	0.131	-0.385	-0.799
Robust Optimization - Ellipsoidal Uncertainty	1.570	0.863	0.315	-0.172	-0.931
Robust Optimization - Box Uncertainty	1.581	0.861	0.305	-0.162	-0.930

3.2 Results - Model Coefficient Comparisons

Comparing the fitted model coefficients β, β_0 across the six models, we find the results as shown in Table 2.

We see that across all regularized and robust approaches, that the best fit models have a dampening of the coefficients as we would expect. As in linear regression, ridge and robust optimization with ellipsoidal uncertainty under this setting both perform a relatively proportional dampening of all of the variables to 0. In contrast, lasso and robust optimization with box uncertainty provide a slightly asymmetric dampening across the variables, although all are still relatively proportional. As we would expect from the model formulations in Section 2, RO with linear uncertainty provides the most uneven dampening.

3.3 Results - Predictive Power

The predictive power for the models was first assessed using the area-under-the-curve (AUC), for the receiver operating characteristic. This measures the ability of a classification model to differentiate between labels, as measured by the false positive and true positive rates, for various choices of the predictive threshold. An AUC of 0.5 corresponds to random guessing, while an AUC of 1.0 corresponds to perfect predictions. The ROC curves are shown in Figure 1.

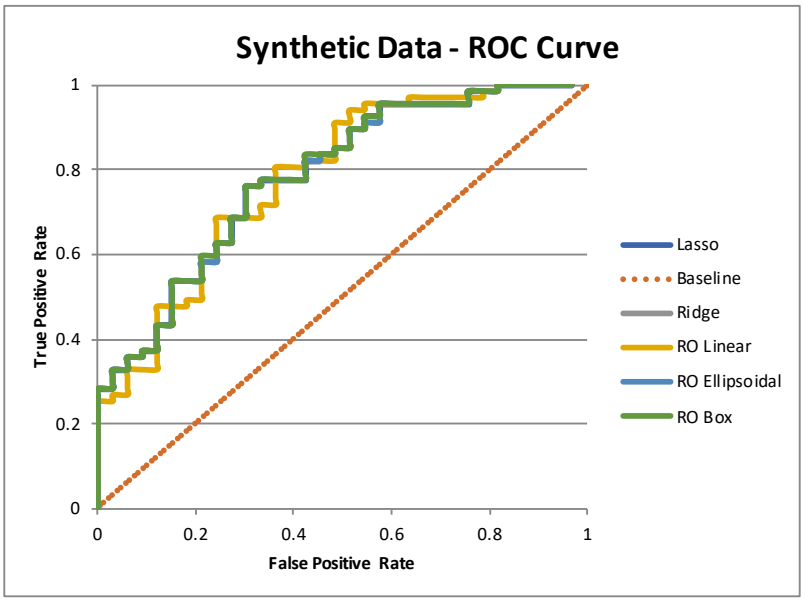
The results are very interesting - the ROC curves are very similar across all of the models, with no clear situations where some models significantly outperform others. After some investigation, it seems like there might be a reasonable explanation for this. Consider the following scenario: assume we fit a model with parameters $\beta = \beta', \beta_0 = \beta_0'$. After applying regularization or robust optimization, we have a model where all parameters have been dampened exactly proportionally, and $\beta = c * \beta', c * \beta_0 = \beta_0'$, for $c \in \mathbb{R}^+$. Now consider the predictions defined as $\hat{y} := \mathbb{P}(Y = 1|X = x) = \frac{1}{1+e^{-c(\beta^T x+\beta_0)}}$. Notice that $\hat{y}$ is monotonic as a function of c. As a result, for any $c \in \mathbb{R}^+$, the order of our predictions $\hat{y}$ will remain unchanged, which means the AUC will remain unchanged. From Table 2, we saw that the coefficient dampening is usually roughly proportional, so as a result it is not surprising that the effect on AUC is negligible.

As an implication of the above, we consider a secondary measure of predictive accuracy: the sum squared error (SSE) of $\hat{y}$, compared to the binary indicator $\mathbb{1}_{y=1}$. That is we define

$$\text{Sq. Err.} := \sum_{i=1}^{n} (\hat{y}_i - \mathbb{1}_{y_i=1})^2.$$

This new metric will capture whether our predictions are more accurate in the regularized or robust approaches, even if the ordering of predictions is the same as in the nominal model. In Table 3,

Figure 1: Synthetic Data - ROC curves



we display the AUC and squared error on the validation and test data sets [2].

Table 3: Synthetic Data - AUC and Squared Error

Model	AUC - Val.	AUC - Test	Sq. Err. - Val.	Sq. Err. - Test
Nominal	0.815	0.783	14.272	17.776
Regularization - Lasso	0.816	0.784	14.242	17.724
Regularization - Ridge	0.818	0.783	14.256	17.607
RO - Linear Uncertainty	0.815	0.782	16.8	**17.112**
RO - Ellipsoidal Uncertainty	**0.818**	0.783	14.287	17.695
RO - Box Uncertainty	0.816	**0.785**	**14.240**	17.743

Under the squared error metric, we see that both regularized and robust optimization approaches outperform the nominal model, with robust optimization with linear uncertainty leading with a decrease of 3.7% for SSE.

4 Jerky Treat Data & Analysis

Next, we repeat the process of evaluating out-of-sample predictive power for the jerky treat data. The jerky treat data contains 192 data points, including 24 jerky treats and 8 experimental replicates for each treat. We have four independent features for each data point, giving scores related to changes in the nucleus, mitochondria, lyosomes, and endoplasmic reticulum. Although we have created 8 experimental replicates for each treat, this is a time consuming process, and we want to create a predictive model that can predict toxicity from a single experimental replicate (as opposed to averaging our independent features across experimental replicates and adjusting for noise in that manner). The toxicity labels are from the F.D.A., with significant vetting, so we assume that these labels are correct.

4.1 Model Setup

We have to follow a slightly more complicated experimental setup, because we only have data from 24 unique jerky treats. The experimental setup is as follows:

[2]This table was updated after completing the poster, using a stronger convergence criteria. Changes in results are very marginal, so the other tables were left the same.

1. We take the 8 replicates of jerky treat number 1 as our testing data set.
2. We take the remaining 184 data points from the 23 other jerky treats as our joint training/validation data set.
3. We perform 23-fold cross-validation on our training/validation data set, to set the hyperparameter for our model.
4. After determining the model hyperparameter, we train our predictive model for treat number 1, on all 184 data points in the training/validation data set.
5. We repeat steps 1-4 to create predictive models for jerky treats number 2-24.
6. We combine all predictive results from steps 4 and 5, and evaluate the combined model performance out-of-sample on all 192 data points.

Because of the model setup, we are unable to reasonably evaluate the effect of model formulation on coefficients or hyperparameters in a simple way, as we did for the synthetic data in Section 3. Instead we focus on our predictive power for evaluation.

4.2 Results - Predictive Power

As in Section 3, the predictive power for the models was assessed using both AUC for the ROC curve, and sum squared error on the binary labels (Sq. Err. $:= \sum_{i=1}^{n}(\hat{y}_i - \mathbb{1}_{y_i=1})^2$). The ROC curves are shown in Figure 2, and the AUC and squared error are reported in Table 4

Figure 2: Jerky Treat Data - ROC curves

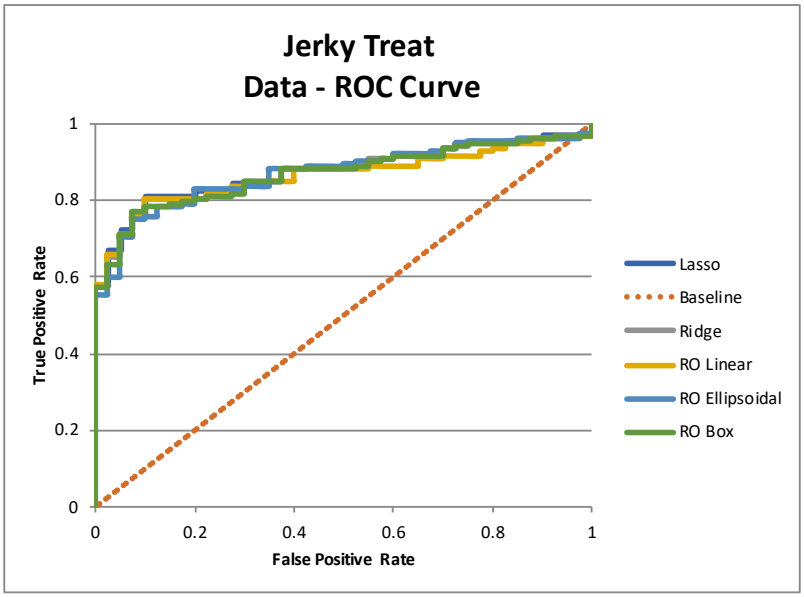


Table 4: Jerky Treat Data - AUC and Squared Error

Model	AUC - Test	Sq. Err. - Test
Nominal	0.873	24.193
Regularization - Lasso	0.874	24.130
Regularization - Ridge	0.871	24.197
RO - Linear Uncertainty	0.870	24.288
RO - Ellipsoidal Uncertainty	**0.876**	**24.092**
RO - Box Uncertainty	0.869	24.161

The results are similar to what we saw for the synthetic data in Section 2 - the ROC curves are very similar across all of the models, with no clear situations where some models significantly outperform others.

We do once again find that robust optimization formulations outperform regularization approaches, but the performance improvement is small. In this case, the performance improvement by using a robust optimization approach, measured by the relative decrease in squared error on the binary indicators, is less than 1%.

5 Conclusion & Discussion

We investigated the problem of predicting jerky treat toxicity, using measurements of how dog kidney cells change in-vitro, when an extract of that treat is introduced. Motivated by the covariate measurement error in the experimental setup, we considered using robust optimization for logistic regression to predict toxicity. We compared this approach to the nominal logistic regression model, as well as logistic regression with regularization.

While the covariate measurement error was substantial, we were not able to significantly improve the AUC compared to the nominal model, through either regularization or robust optimization. We did have a measurable improvement on the sum squared error of the binary toxicity labels, though, with a robust optimization approach. Through robust optimization, we were able to decrease the sum squared error 3.7% for the synthetic data, and 0.5% for the jerky treat data.

In Section 3, we were able to get an intuition for why we might have observed this phenomenon. When we fit a model with parameters $\beta = \beta', \beta_0 = \beta_0'$, If we then dampen model coefficients exactly proportionally, and let $\beta = c * \beta', c * \beta_0 = \beta_0'$, for $c \in \mathbb{R}^+$, then the AUC will remain unchanged, as the function mapping our old predicted toxicity to new predicted toxicity is monotonic. Because this is roughly what happened to the coefficients in Table 2, it is not surprising that AUC changed only marginally across the various models.

We offer the following concluding thoughts, on when robust optimization would be most useful for logistic regression. For robust optimization to strongly improve upon AUC compared to the nominal logistic regression model, we hypothesize that the following two conditions should hold:

1. The MLE estimates for the nominal model should be disproportionate. If MLE for nominal logistic regression produces model coefficients that are all biased by some multiplicative factor $c \in \mathbb{R}^+$ compared to the true best-fit model, then our AUC will already be equivalent to the best-fit model

2. We need to apply uncertainty sets for robust optimization, that will impact the fitted coefficients in a disproportionate way. Otherwise, as we argue above, the AUC will not improve.

In general, robust classification is a very new area, and it will be interesting to see how the heuristics develop as to where robust optimization for classification is the most useful, compared to regularization and the nominal model.

References

[Bar10] Jonathan William Bartlett. Correction for classical covariate measurement error and extensions to life-course studies. A thesis submitted to the University of London for the degree of Doctor of Philosophy, 2010.

[BDPZ19] Dimitris Bertsimas, Jack Dunn, Colin Pawlowski, and Ying Daisy Zhuo. Robust classification. *INFORMS Journal on Optimization*, 1(1):2–34, 2019.

[Ber19] Bertsimas, Dimitris. Lecture 22: Robust Linear Regression. MIT Course 15.094: Robust Modeling, Optimization and Computation, 2019.

[Bis06] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[CB01] George Casella and Roger Berger. *Statistical Inference*. Duxbury Resource Center, 2001.

[FMXY14] Jiashi Feng, Shie Mannor, Huan Xu, and Shuicheng Yan. Robust logistic regression and classification. NIPS, 2014.

[HJZW+] Patrick L. Harrington Jr., Aimee Zaas, Christopher W. Woods, Geoffrey S. Ginsburg, Lawrence Carin, and Alfred O. Hero III. Robust logistic regression with bounded data uncertainties. *Preprint*. http://web.eecs.umich.edu/~hero/Preprints/Harrington_TSP10_dbl.pdf.

[RHPS03] Sophia Rabe-Hesketh, Andrew Pickles, and Anders Skrondal. Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Statistical Modelling*, 3:215–232, 2003.

[SAEK] Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. *arXiv preprint*, (1509.09259v3).

[SC85] Leonard A. Stefanski and Raymond J. Carroll. Covariate measurement error in logistic regression. *The Annals of Statistics*, 13(4):1335–1351, 1985.

[SSM97] Donna Spiegelman, Sebastian Schneeweiss, and Aidan McDermott. Measurement error correction for logistic regression models with an "alloyed gold standard". *American Journal of Epidemiology*, 145(2):184–196, 1997.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[U.S18] U.S. F.D.A. FDA Investigates Animal Illnesses Linked to Jerky Pet Treats. https://www.fda.gov/animal-veterinary/news-events/fda-investigates-animal-illnesses-linked-jerky-pet-treats, 2018.