PREDICTING SURGERY DURATION WITH NEURAL HETEROSCEDASTIC REGRESSION

Nathan Ng[1], **Rodney A Gabriel**[2,3], **Julian McAuley**[1], **Charles Elkan**[1], **Zachary C Lipton**[1,2] *
Department of Computer Science[1]
Division of Biomedical Informatics[2]
Department of Anesthesiology[3]
University of California, San Diego
9500 Gilman Drive La Jolla, CA 92093, USA
{nhng, ragabriel, jmcauley, elkan, zlipton}@ucsd.edu

ABSTRACT

Scheduling surgeries is a challenging task due to the fundamental uncertainty of the clinical environment, as well as the risks and costs associated with under- and over-booking. We investigate neural regression algorithms to estimate the parameters of surgery case durations, focusing on the issue of *heteroscedasticity*. We seek to simultaneously estimate the duration of each surgery, as well as a surgery-specific notion of our *uncertainty* about its duration. Estimating this uncertainty can lead to more nuanced and effective scheduling strategies, as we are able to schedule surgeries more efficiently while allowing an informed and case-specific margin of error. Using surgery records from the UC San Diego Health System, we demonstrate potential improvements on the order of 18% (in terms of minutes overbooked) compared to current scheduling techniques, as well as strong baselines that do not account for heteroscedasticity.

1 INTRODUCTION

Healthcare in the United States is expensive and hospital resources are scarce. Healthcare expenditure now exceeds 17% of US GDP (World Bank, 2014), even as surgery wait times steadily increase (Bilimoria et al., 2011). One source of inefficiency (among many) is the inability to fully utilize hospital resources. Since the duration of surgeries cannot be accurately predicted, operating rooms can become congested (when surgeries run long) or lie vacant (when they run short). Over-booking can lead to long wait times and higher costs of labor (due to over-time pay), while under-booking decreases throughput, increasing the marginal cost per surgery.

We seek to address this issue by developing better and more nuanced strategies for surgery case prediction. Our work focuses on a collection of surgery logs recorded in Electronic Health Records (EHRs) at the University of California, San Diego Health System. For each patient, we consider a collection of pre-operative features, including patient attributes (age, weight, height, sex, co-morbidities, etc.), as well as attributes of the clinical environment, such as the surgeon, surgery location, and time. For each procedure, we also know how much time was originally scheduled, in addition to the *actual* (‘ground-truth’) surgery duration, recorded after each procedure is performed.

Our raw dataset consists of 107,755 surgeries. After discarding surgeons and procedures associated with fewer than 100 cases, we are left with 86,945 examples, which we split 80%/8%/12% for training/validation/testing. To handle missing values, we incorporate missing value indicators, following previous work on clinical datasets (Lipton et al., 2016).

We are particularly interested in developing methods that allow us to better estimate the *uncertainty* associated with the duration of each surgery. Traditional regression objectives assume *homoscedasticity*, i.e., constant levels of target variability for all instances. While mathematically convenient, this assumption is clearly violated in data such as ours: As a simple example, operations that are

*Corresponding author, website: http://zacklipton.com expected to take one hour might exhibit greater variance than those projected to take ten minutes; or variability could be higher for complex as opposed to routine operations.

Our approach revisits the idea of heteroscedastic neural network regression. We jointly learn all parameters of a predictive distribution. In particular, we consider Gaussian and Laplace distributions, each of which is parameterized by a mean and standard deviation. Our models allow for increased scheduling efficiency compared to both current practice and neural network baselines that fail to account for heteroscedasticity. Furthermore, our models produce reliable estimates of the variance, which can be used to schedule intelligently, especially when over-booking and under-booking confer disparate costs.

2 Methods

A typical solution to a regression task might consist of finding maximum likelihood parameters of a Gaussian predictive distribution:

$$\boldsymbol{\theta}^{\text{MLE}} = \max_{\theta} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(\frac{-(y_i - \hat{y}_i)^2}{2\hat{\sigma}^2}\right) = \min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(\log(\hat{\sigma}_i) + \frac{(y_i - \hat{y}_i)^2}{2\hat{\sigma}^2}\right) \tag{1}$$

Assuming constant $\hat{\sigma}$, this yields a familiar least-squares objective. In this work, we seek to relax this assumption, predicting both $\hat{y}(\theta, \boldsymbol{x})$ and $\hat{\sigma}(\theta, \boldsymbol{x})$ simultaneously.

We consider prediction using Multilayer Perceptrons (MLPs), though the idea is easily applied to networks of arbitrary architecture. To predict the standard deviation $\hat{\sigma}$ of the predictive distribution, our MLP has two outputs: The first output has linear activation and is used as $\hat{y}$ to calculate the objective; the second output models $\hat{\sigma}$. To enforce positivity of $\hat{\sigma}$, this output undergoes the activation function $\text{softplus}(z) = \log(1 + \exp(z))$.

We extend the same idea to Laplace distributions, which turn out to better describe the target variability for surgery duration, and are also maximum likelihood estimators when optimizing the Mean Absolute Error (MAE). Mean Absolute Error corresponds to the average number of minutes over or underbooked, and is typically the quantity of interest for this type of scheduling task. The Laplace distribution is parameterized by $b = \sqrt{2}\sigma$:

$$\boldsymbol{\theta}^{\text{MLE}} = \max \prod_{i=1}^{n} \frac{1}{2b} \exp\left(\frac{-|y_i - \hat{y}_i|}{b}\right) = \min_{\theta} \sum_{i=1}^{n} \log b + \frac{|y_i - \hat{y}_i|}{b}. \tag{2}$$

3 Experiments

For all experiments, we use MLPs with ReLU activations. For evaluation, we report the root mean squared error (RMSE), mean absolute error (MAE), and negative log-likelihood (NLL). For homoscedastic models, we choose a constant σ that minimizes NLL on the validation set. For heteroscedastic models, we evaluate NLL using the predicted deviations σ_i. We run all experiments using MLPs with 2 hidden layers and a number of hidden units chosen by gridsearch on validation data. All models use dropout regularization and weight decay.

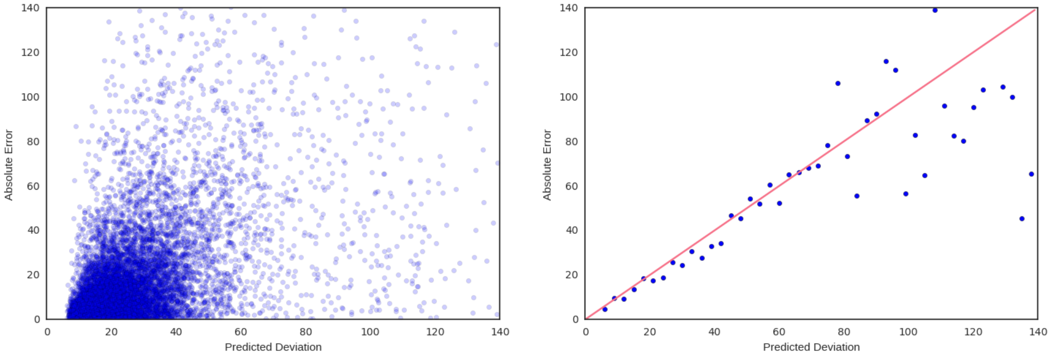
Results are shown in Table 1. Plots in Figure 1 demonstrate that the predicted deviation reliably estimates the observed error and QQ plots (Figure 2) demonstrate that the Laplace distribution appears to fit our targets better than a Gaussian predictive distribution.

Models	**RMSE**	**MAE**	**NLL**
Current Method	49.80	28.87	0.5985
Procedure Means	49.06	27.70	0.5899
Linear Regression	45.23	25.07	0.5084
MLP Gaussian	**44.28**	24.78	0.4888
MLP Gaussian HS	44.48	23.82	0.0852
MLP Laplace	44.56	23.54	0.4619
MLP Laplace HS	44.83	**23.53**	−0.1251
MLP Gamma HS	44.45	23.67	**−0.1540**

Table 1: Performance on test-set data

4 Related Work

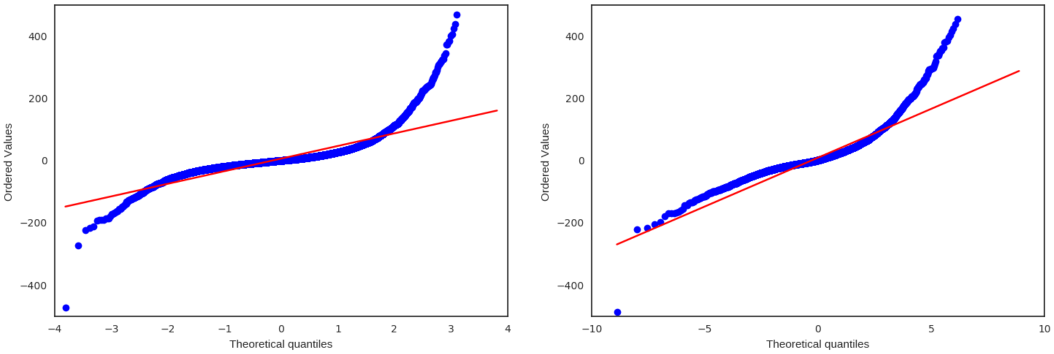
Previous work in in the medical literature addresses the prediction of surgery duration (Eijkemans et al., 2010; Kayış et al., 2015; Devi et al., 2012), accounting for both patient and surgical team characteristics. To our knowledge ours is the first paper to address the problem



(a) Scatter

(b) Binning by predicted $\hat{\sigma}$ with width 3

Figure 1: Plots of predicted $\hat{\sigma}$ against absolute error with Laplace noise model. Averaging over bins of width 3 (b), shows that $\hat{\sigma}_i$ is a reliable estimator of the observed error.



(a) Gaussian QQ Plot

(b) Laplace QQ Plot

Figure 2: QQ plots of observed error for Gaussian and Laplace noise models. The Laplace distribution better describes observed error, with shorter tails at both ends.

with modern deep learning techniques and to the first to model its heteroscedasticity. The idea of neural heteroscedastic regression was first proposed by Nix & Weigend (1994), though they do not share hidden layers between the two outputs, and are only concerned with Gaussian predictive distributions. Williams (1996) use a shared hidden layer and consider the case of multivariate Gaussian distributions, for which they predict the full covariance matrix via its Cholesky factorization. Heteroscedastic regression has a long history outside of neural networks. Le et al. (2005) address a formulation for Gaussian processes. Most related is Lakshminarayanan et al. (2016) which also revisits heteroscedastic neural regression, also using a softplus activation to enforce non-negativity. We show some successful modifications to the above work, such as the use of the Laplace distribution, but our more significant contribution is the application of the idea to clinical medical data.

5 Discussion

We demonstrate the efficacy of heteroscedastic neural regression for predicting surgery duration. Our best models improve upon current practice, reducing MAE by 18.4%. This work could potentially lead to greater throughput and lower costs. We provide further analysis in Appendix A, demonstrating that each model could be used to trade off between over-reserved and under-reserved minutes. Here, the uncertainty information from heteroscedastic models aids decision theory. Our analysis shows that the heteroscedastic Laplace MLP strictly dominates the other models.

REFERENCES

Karl Y Bilimoria, Clifford Y Ko, James S Tomlinson, Andrew K Stewart, Mark S Talamonti, Denise L Hynes, David P Winchester, and David J Bentrem. Wait times for cancer surgery in the united states: trends and predictors of delays. *Annals of surgery*, 253(4):779–785, 2011.

S Prasanna Devi, K Suryaprakasa Rao, and S Sai Sangeetha. Prediction of surgery times and scheduling of operation theaters in optholmology department. *Journal of medical systems*, 2012.

Marinus JC Eijkemans, Mark van Houdenhoven, Tien Nguyen, Eric Boersma, Ewout W Steyerberg, and Geert Kazemier. Predicting the unpredictable: A new prediction model for operating room times using individual characteristics and the surgeon's estimate. *The Journal of the American Society of Anesthesiologists*, 2010.

Enis Kayış, Taghi T Khaniyev, Jaap Suermondt, and Karl Sylvester. A robust estimation model for surgery durations with temporal, operational, and surgery team effects. *Health care management science*, 2015.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016.

Quoc V Le, Alex J Smola, and Stéphane Canu. Heteroscedastic gaussian process regression. In *ICML*, 2005.

Zachary C Lipton, David C Kale, and Randall Wetzel. Modeling missing data in clinical time series with rnns. In *Machine Learning for Healthcare*, 2016.

David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *International Conference on Neural Networks*. IEEE, 1994.

Peter M Williams. Using neural networks to model conditional multivariate densities. *Neural Computation*, 1996.

World Bank. Health expenditure, total (% of gdp), 2014. URL `http://data.worldbank.org/indicator/SH.XPD.TOTL.ZS`.

Appendices

A Economic Analysis

Our aim in predicting the variance of the error is to provide uncertainty information that could be used to make better scheduling decisions. To compare the various approaches economically, we consider the simple case where the cost to over-reserve the room by one minute (procedure finishes early) differs from the cost to under-reserve the room (procedure runs over). We demonstrate how the two quantities can be traded off in Figure 3.

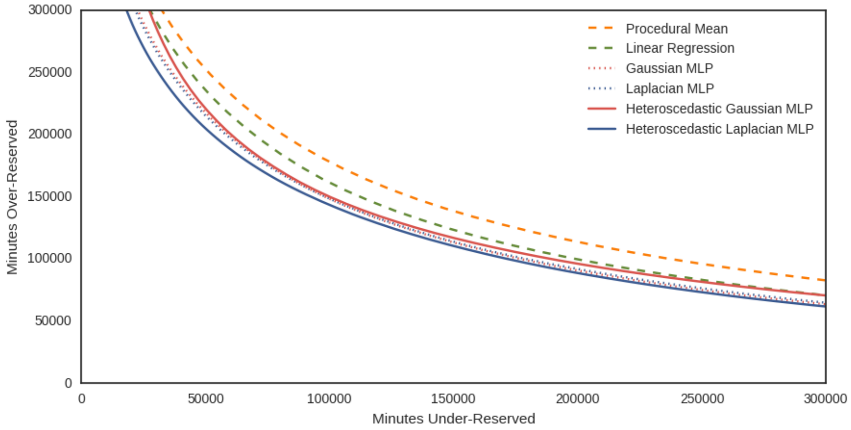


Figure 3: Trade-offs between over and under-reserving operating rooms

For models that don't output variance, we consider scheduled durations of the form $\hat{y} + k$ and $\hat{y} \cdot k$ where k is a data-independent constant. In either case, by modulating k, one books more or less aggressively. The multiplicative approach performed better, likely because long procedures have higher variance than short ones. For heteroscedastic models, we make the trade-off by modulating k for $\hat{y} + \hat{\sigma} \cdot k$. As shown in Figure 3, the heteroscedastic model strictly dominates other methods.