Survival Analysis on Dogs and Cats in Animal Shelters

Jessica Ko
UC Berkeley
jessicak@berkeley.edu

1 Introduction

Animal shelters are often overcrowded with animals, and they aren't always adopted. In this paper, I want to explore whether there are correlations between animal characteristics and survival once they enter an animal shelter over time. Understanding these correlations can help improve animal shelters and the well-being of these animals. Survival analysis using the Cox Proportional Hazards Model, which is related to a generalized linear model, is applied to the Austin Animal Shelter dataset to explore survival of animals in shelters [1].

Traditionally, survival analysis is used in bio-statistics to determine the chances of a patient surviving after undergoing some treatment. This can be used for survival analysis of cancer patients after receiving chemotherapy. Data is recorded from the patient over time and the outcome after the study is noted. Some possible outcomes are dying, being cured, or exiting the study. For this dataset, it is assumed that all the animals are "treated" by entering the animal shelter. Survival analysis can be used for areas outside of medicine such as the Austin Animal Shelter dataset.

1.1 Problem Statement

The goal of the paper is to find probabilities of dogs and cats in animal shelters surviving, not being euthanized or dying, over time and compare it with different animal characteristics like color, etc. Survival is defined as an animal being adopted or transferred to another center. Failure, or death, is defined as when an animal dies or is euthanized.

The goal is to find the survival function, $S(t) = Pr(T_{death} > t)$, which is the probability of surviving after time t. In survival analysis, it is defined as

$$S(t) = \exp\left(-\Lambda(t)\right)$$

$\Lambda(t)$ is the cumulative hazard, or cumulative risk, is defined as [1]

$$\Lambda(t) = \int_0^t \lambda(x)dx$$

where $\lambda(t)$ is the hazard function, the instantaneous probability of death at time t, given survival until t. [2] It can also be rewritten as

$$\lambda(t) = -\frac{d}{dt}\log S(t)$$

Every animal i must keep track of a time t_i where a death or a censor event occurs. Censoring is when the event of interest has not yet occurred. In this dataset, animals who are adopted or transferred to another animal shelter fall under this category because it is possible for an adopted animal to re-enter the shelter and the transferred animal is still in another shelter. An event of failure is still possible for these animals. In addition, a failure indicator D_i is used to determine what type of event occurred. It is marked as 1 if a failure occurred and 0 for censoring.

1.2 Dataset

The dataset contains 49,970 animals recorded from 2013 to present such as outcome time, outcome type (adoption, transfer, euthanasia, or death), etc. Even though the animals have different dates for when they entered the animal shelter, the start times can be interpreted as starting at the same with no loss of generality [6]. The time is measured as the difference between intake time and outcome time. The dataset leads itself well to using binary features. Some of the data is intake type (Owner,Stray, or PublicAssist), injured/sick or not, pregnant or not, dog or cat, spayed/neutered or not, gender, purebred or mix, and black colored or not.

2 Cox Proportional Hazard Model

Many different factors can affect the survival of animals in pet shelters as mentioned in Section 1.2. In order to understand the correlations between these features like color, breed, and etc. on survival, an appropriate model must be used. The Austin Animal Shelter dataset does not have any time-dependent features; however, the Cox Model allows for this kind of features if relevant. The Cox Proportional Hazards Model accurately depicts these interactions in the hazard function [4]. Given a vector $x = (x_1, \ldots, x_d)$ of d features and a parameter $\beta \in \mathbb{R}^d$.

$$\lambda(t|x) = \lambda_0(t)e^{\beta^\top x}$$

The baseline hazard function $\lambda_0(t)$ does not need to be specified for the Cox model, making it semi-parametric.

The Cox Model is robust because the baseline hazard function does not to be specified, which allows freedom. The model only needs to satisfy the proportional hazard assumption, which is that the hazard of one sample is proportional to the hazard of another sample [5]. The dataset is shown to satisfy the proportional hazard model in Section 5.1. To exhibit this assumption, consider two samples x_1 and x_2. The ratio is not dependent on time.

$$\begin{aligned}\frac{\lambda(t|x_1)}{\lambda(t|x_1)} &= \frac{\lambda_0(t)e^{\beta^\top x_1}}{\lambda_0(t)e^{\beta^\top x_2}} \\ &= \beta^\top(x_1 - x_2)\end{aligned}$$

3 β Estimation by Partial Likelihood

The parameter β can be found by maximizing the partial likelihood because the hazard function is not specified.

3.1 Equivalence to Poisson Regression

The partial likelihood of the Cox Model can be fitted by the likelihood of Poisson regression, a generalized linear model, because the likelihoods are proportional to each other [6]. As a result, the estimates of β will be the same. This is very advantageous because the Cox Model can be fitted using software for generalized linear models like in R. Alternatively, the estimate from the Cox Model can be used for Poisson regression. In Section 4, a coordinate descent method is proposed for solving the maximum partial likelihood of the Cox Model.

The Cox Model can be interpreted in terms of a Poisson regression. Given the cumulative hazard $\Lambda(t)$, the estimates of β can be obtained by treating the failure indicator D_i as Poisson distributed with mean $\mu_i = \Lambda(t_i)e^{\eta_i}$ where $\eta = \beta^\top x$.The link function is modified to be $\beta^\top x = \log(\mu_i) - \log(\Lambda(t_i))$ [7].

3.2 Partial Likelihood

The partial likelihood for the Cox Model can be written as

$$\begin{aligned}\mathcal{L}(\beta) &= \prod_{\{i:D_i=1\}} \frac{\lambda_0(t_i)e^{\beta^\top x_i}}{\sum_{j\in R_i}\lambda_0(t_i)e^{\beta^\top x_j}} \\ &= \prod_{\{i:D_i=1\}} \frac{e^{\beta^\top x_i}}{\sum_{j\in R_i} e^{\beta^\top x_j}}\end{aligned}$$

where the risk set R_i is the set of indexes of samples with death or censor time occurs after t_i. This represents the probability of failure occurring to a sample at time t_i among those at risk at time t_i. The semi-parametric property can be exhibited here because the baseline hazard λ_0 gets canceled out.

However, the partial likelihood function does not take tied events into account. Tied events occur if the number of deaths d_i at time t_i is greater than 1. The partial likelihood above does not incorporate ties, so the probabilities are not as accurate. Breslow introduces a different partial likelihood function to deal with the ties [8]

$$\mathcal{L}(\beta) = \prod_{\{i:D_i=1\}} \frac{e^{\beta^\top x_i}}{\left[\sum_{j\in R_i} e^{\beta^\top x_j}\right]^{d_i}}$$

The paper will refer to the Breslow ties as the partial likelihood because ties occur in the dataset.

3.3 Minimize Negative Log Likelihood

The parameter β can be found by minimizing by the negative log likelihood $\ell(\beta)$. The log likelihood $\ell(\beta)$ is defined as

$$\begin{aligned}\ell(\beta) &= \log(\mathcal{L}(\beta)) \\ &= \sum_{\{i|D_i=1\}} \log \frac{e^{\beta^\top x_i}}{\left[\sum_{j\in R_(t_i)} e^{\beta^\top x_j}\right]^{d_i}}\end{aligned}$$

The minimization of the negative log likelihood with ℓ_1 regularization is formed below

$$\min_\beta - \sum_{\{i|D_i=1\}} \log \frac{e^{\beta^\top x_i}}{\left[\sum_{j\in R_(t_i)} e^{\beta^\top x_j}\right]^{d_i}} + s\|\beta\|_1$$

Regularization is included because there are many benefits such as being more accurate than stepwise selection and yielding interpretable models [9]. In addition, the regularization prevents degenerate behavior when there are many more predictors than observations.

4 Coordinate Descent on Minimum Partial Likelihood

Cyclic coordinate descent is used to solve the minimum partial likelihood. The algorithm cycles between fixing each index and solving the minimization problem. Because the log likelihood $\ell(\beta)$ is convex and smooth, coordinate descent will converge to the correct solution. When β has a large dimension, coordinate descent is advantageous over other methods like gradient descent. The dataset originally had more predictors; however, a few made the model not satisfy the proportional assumption, so they had to be removed from the final model.

Each individual minimization problem for a fixed index is solved by the bisection method. The benefit of this method is calculating the derivative with respect to one variable, which is not as costly as calculating the gradient [10]. The method starts with an interval where the optimal parameter β^*

falls under. The interval for each bisection method is $\left[\frac{\ell(0)}{s}, \frac{-\ell(0)}{s}\right]$ and can be shown below:

$$\begin{aligned} -\ell(\beta^*) + s\|\beta^*\|_1 &\leq -\ell(0) + s\|0\|_1 \\ s\|\beta^*\|_1 &\leq -\ell(0) + \ell(\beta^*) \\ \|\beta^*\|_1 &\leq \frac{1}{s}(-\ell(0) + \ell(\beta^*)) \\ \|\beta^*\|_1 &\leq \frac{-\ell(0)}{s} \end{aligned}$$

The likelihood $\mathcal{L}(\beta)$ is a product of probabilities, so it lies between $[0, 1]$. It is important to note that $\forall\beta\ \ell(\beta) \leq 0$ because $\ell(\beta) = \log(\mathcal{L}(\beta))$. The term can be dropped out from the expression above.

The algorithm with coordinate descent and the bisection method is shown in Algorithm 1. The derivative for the log likelihood, $f'_k(\beta_k)$, with respect to index k where $x_{s,t}$ indicates the t feature of sample s is below. For coordinate descent, all indexes in β are fixed except β_k. Because the ℓ_1 is not differentiable at 0, a subgradient is introduced to handle this case [11]. Thus, the derivative of $\|x\|_1 = g(x)$.

$$g(x) = \begin{cases} +1, & \text{if } x > 0 \\ -1, & \text{otherwise} \end{cases}$$

Using this as the derivative for ℓ_1 norm,

$$\begin{aligned} f'_k(\beta_k) &= \frac{d}{d\beta_k}\left[-\ell(\beta) + s\|\beta\|_1\right] \\ &= \sum_{\{i|D_i=1\}} \left[x_{i,k} - \left(\frac{d_i}{\sum_{j\in R_i} e^{\beta^\top x_j}}\right)\left(\sum_{j\in R_i} x_{j,k} e^{\beta^\top x_j}\right)\right] + sg(\beta_k) \end{aligned}$$

initialize β ;
while β *Not Converged* **do**
  for *index* i *in* β **do**
    Fix all indexes except i in β;
    Consider interval $[l = \frac{\ell(0)}{s}, u = \frac{-\ell(0)}{s}] \in \beta_i^*$;
    while β_i *Not Converged* **do**
      $x \leftarrow (l + u)/2$;
      if $f'_i(x) < 0$ **then**
        $l \leftarrow x$;
      else
        $u \leftarrow x$;
      end
    end
    $\beta_i \leftarrow x$;
  end
end

Algorithm 1: Coordinate Descent

5 Results

In this section, the fit of the Cox Model on the dataset is assessed.The data must satisfy some conditions in order to be properly used with the model. The survival curves are also shown in this section to demonstrate the correlation of certain features and survival. In R, the survival and glmnet library were used for evaluating the Cox Model fit and selecting a λ. I wrote my own python code to find the minimum β and graph the survival curves.

5.1 Assessing Cox Model Fit

In order to use the Cox Model, the proportional hazard assumption must hold for the data. The assumption can be verified with a p-value > 0.05. In addition, the Schoenfeld residuals plots should have a slope of 0 [12].

At first, I tried a few different combinations of the features and chose the features to investigate based on the p-value given from R. The values are shown in the figures below. Figure 1 and 2 do not satisfy the proportional hazard assumption because the p-values are respectively 0 and 0.00239, which are smaller than 0.05. Figure 3 has a p-value of 0.252, which is above 0.05. This means that the proportional hazard assumption is satisfied. In the next section, I will graph the survival curves for the SpayNeuter, Gender, and Color features.

Figure 1: P-value Choosing Most Features

	rho	chisq	p
Owner	-0.18801	95.9222	0.00e+00
Stray	-0.21705	118.7589	0.00e+00
Normal	0.02639	1.8576	1.73e-01
InjuredSick	-0.04477	5.3810	2.04e-02
AnimalType	0.08816	20.5145	5.92e-06
SpayNeuter	-0.04053	4.4404	3.51e-02
Gender	0.02234	1.4298	2.32e-01
AgeIntake	0.01127	0.3272	5.67e-01
Breed	-0.00846	0.1906	6.62e-01
Color	0.00582	0.0912	7.63e-01
GLOBAL	NA	370.7772	0.00e+00

Figure 2: P-value Choosing Some Features

	rho	chisq	p
SpayNeuter	-0.0316	2.67	0.10204
Gender	-0.0284	2.40	0.12134
AgeIntake	-0.0316	2.65	0.10342
Breed	-0.0632	10.66	0.00110
Color	-0.0212	1.21	0.27115
GLOBAL	NA	18.49	0.00239

Figure 3: P-value Choosing a Few Features

	rho	chisq	p
SpayNeuter	-0.03026	2.446	0.118
Gender	-0.00884	0.209	0.647
Color	-0.02305	1.423	0.233
GLOBAL	NA	4.088	0.252

The Schoenfeld residuals also exhibit the assumption. The slope of all these graphs is 0. The Figures 4, 5, and 6 show solid and flat black lines.

Figure 4: Schoenfeld Residuals for Spay or Neuter

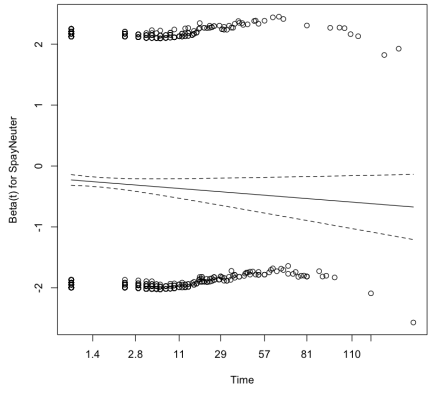


Figure 5: Schoenfeld Residuals for Color

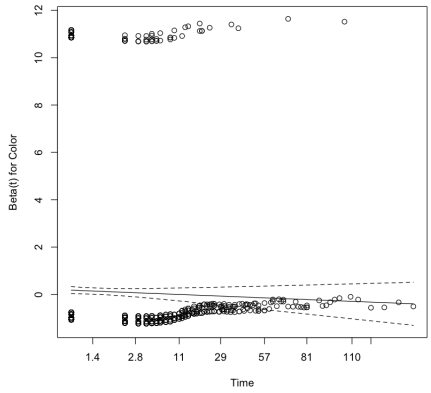
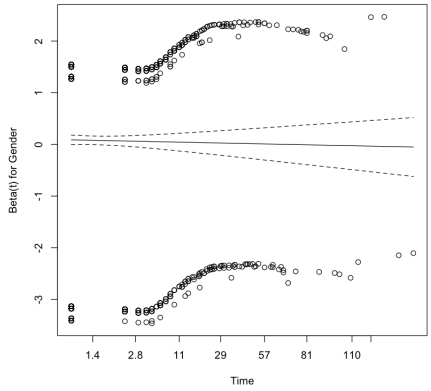


Figure 6: Schoenfeld Residuals for Gender



5.2 Survival Curves

Survival curves allow the correlation between the features and survival to become evident. Because the proportional hazard assumption holds for the features picked, the curves will not cross each other, which indicates how one feature correlates more with survival. The survival function for the Cox Model at time t for a given sample x is defined as

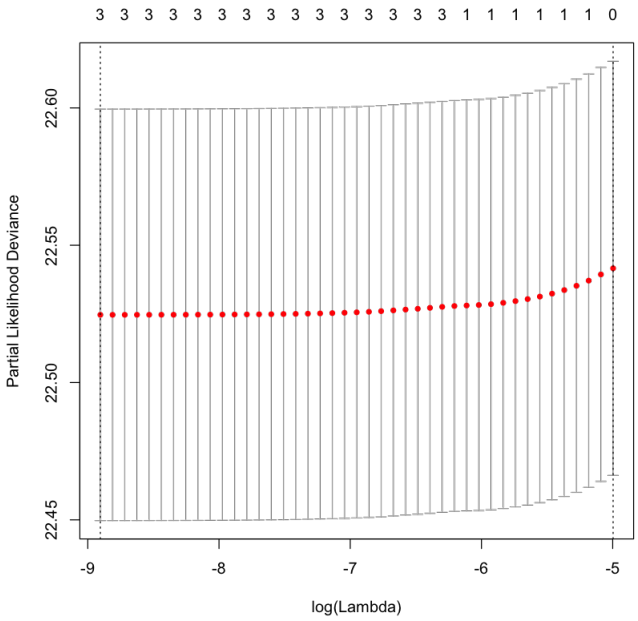
$$\begin{aligned} S(t|x) &= \exp(-\Lambda(t)) = \exp\left(\int_0^t \lambda(s|x)ds\right) \\ &= \exp\left(-\int_0^t \lambda_0(s)e^{\beta^\top x}ds\right) \\ &= \exp\left(-e^{\beta^\top x}\int_0^t \lambda_0(s)ds\right) \\ &= S_0(t)^{e^{\beta^\top x}} \end{aligned}$$

where the cumulative baseline hazard and baseline survival are $\Lambda_0(t) = \int_0^t \lambda_0(s)ds.$ and $S_0(t) = e^{-\Lambda_0(t)}$ [13]. The cumulative baseline hazard is estimated as $\Lambda_0(t) = \sum_{i:t_i<t} \frac{d_i}{\sum_{j\in R_i} e^{\beta^\top x_j}}$.

In order to graph the survival curves, the best λ needs to be chosen. The λ with the lowest cross-validated deviance is chosen for the model [3]. The deviance is calculated by doing cross validation.

For the dataset, I used 10-fold cross validation. Figure 7 shows the deviances and it seems to be slowly increasing as $\log(\lambda)$ increases, so I picked $\log(\lambda) = -8.9$.

Figure 7: Picking the Best λ



From finding β above, the survival curves can now be graphed. The mean of the data has to be prepared in order to graph the curves because the survival function is dependent on time. For the feature of interest, separate the data by the classes. For example, the data is separated by male and female for gender. After, the mean of each feature is calculated. However, the index of the feature of interest is marked the respective value like 1 for female and 0 for male. The two mean vectors, one for each binary class, is then graphed using the survival function $S(t|x)$ for all failure times. The following figures are results of the survival curves. Each downward step in the graphs indicates an event of failure.

Figure 8:

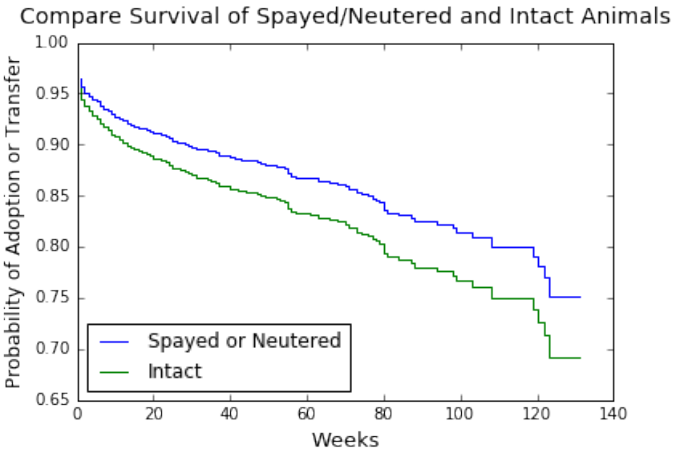


Figure 9:

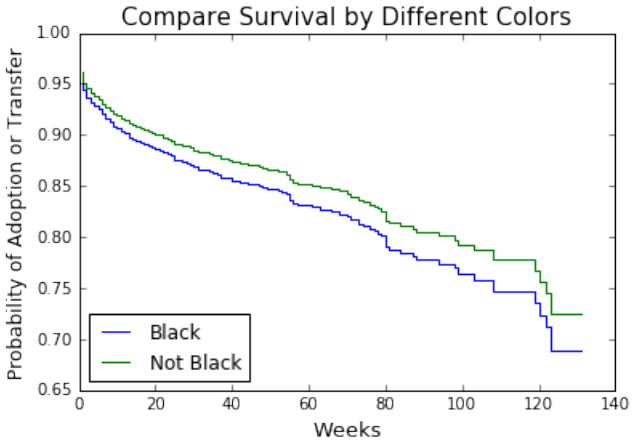


Figure 10:

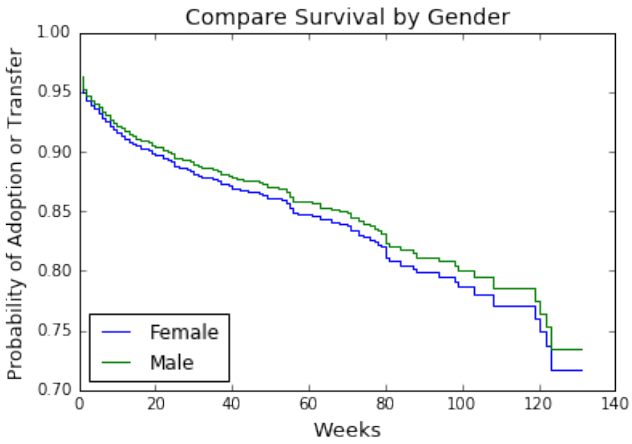


Figure 8 shows that spayed or neutered animals correlate more with survival. This makes sense because adopting a pet who has already undergone the surgery will save costs and there are benefits for spaying or neutering. Figure 9 shows that black colored animals correlated with lower survival. Black cats are sometimes symbolized as bad luck, and darker colored dogs are often portrayed as being aggressive. Compared to the other factors, gender does not seem to correlate with survival as greatly as shown in Figure 10.

6 Conclusion

In this paper, the correlation of different characteristics and probability of survival in animals in the Austin Animal Center are explored. The Cox Proportional Hazards Model is used to model the survival function. It has many benefits such as not specifying the baseline hazard. The partial likelihood is equivalent to the likelihood of Poisson regression and can also be understood in terms of the Poisson regression. The partial likelihood of the Cox Model is solved by coordinate descent and the bisection method. For this dataset, the proportional hazards assumption holds for the SpayNeuter, Color, and Gender feature, so the Cox Model can be applied on them. The curves show that spayed or neutered animals have an advantage for surviving.

Because other features did not satisfy the proportional hazards assumption, they could not be analyzed with the model. This really limited what kind of features could be explored, even though there are several other features available in the dataset like breed and intake type. The modified version,

the stratified general Cox, could adjust those features that do not satisfy the proportional hazards assumption [14]. More work can be done in the future to use more features for the model.

References

[1] Austin Animal Center (2016) http://www.austintexas.gov/department/animal-services

[2] Rodriguez, G. (2007). Lecture Notes on Generalized Linear Models. URL: http://data.princeton.edu/wws509/notes/

[3] Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical learning with sparsity: The lasso and generalizations. Boca Raton: CRC Press, Taylor Francis Group.

[4] D. R. Cox and D. Oakes. Analysis of survival data, volume 21. CRC Press, 1984.

[5] Cayé, T. Evaluating Proportional Hazards Assumption.

[6] Whitehead, J. (1980). Fitting Cox's regression model to survival data using GLIM. Applied Statistics, 268-275.

[7] McCullagh, P., Nelder, J. A. (1989). Generalized linear models. London: Chapman and Hall.

[8] Breslow, N. (1974). Covariance Analysis of Censored Survival Data. Biometrics, 30(1), 89-99. doi:10.2307/2529620

[9] Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. Statistics in medicine, 16(4), 385-395.

[10] D'Aspremont, A. Convex Optimization.

[11] Boyd, S. Subgradients. Notes for EE364b, Stanford University, Winter 2006-07

[12] Applied Survival Analysis Chapter 6: Assessment of Model Adequac. IDRE UCLA

[13] Rodriguez, G. Non-Parametric Estimation in Survival Models.

[14] Borsi, L. (2011, March). The stratified Cox Procedure.