Final project: Survival analysis of cetaceans in captivity

Kerinna Good and Josh Wagner

The `all_cetaceans2` data set used in this analysis contains information on 1,416 dolphins and whales living in captivity in the US between 1943 and 2017. Our analysis explores the relationship between cetacean survival rates in captivity and type of species, sex, whether animals were born in the wild or captivity, current location (as of the recorded status date), the percentage of animals' life lived in captivity, whether animals were transferred between facilities during their lifetime, and whether the animals were transferred from foreign countries. Our research question asks, do animals that spend more of their lives in captivity live longer or shorter lives than animals that spend less of their lives in captivity? We hypothesize that animals that spend more of their lives in captivity will live shorter lives than animals that spend less of their lives in captivity. The results of our analysis could help inform best practices for holding dolphins and whales in captivity, and further could inform policies geared toward reducing cetacean fatalities in captivity.

Our data comes from an article in The Pudding published by Amber Thomas in 2017 titled "Free Willy and Flipper by the Numbers" (Thomas (2017)). Specifically, we use a data set titled `all_cetaceans2` which is a modified version of the `allCetaceanData` data set published on Data.world (Data.world (2017)). The data were originally collected from the National Marine Mammal Inventory (curated by the National Oceanic and Atmospheric Administration) and the crowd-sourced website Ceta-Base (Thomas (2017)). Starting with the `allCetaceanData` data set, we renamed several variables for clarity, mutated a new variable for status date that lists the status date for alive cetaceans as May 7, 2017 (the date at which each animal's status was evaluated) opposed to NA, mutated a new variable for each cetacean's age in years, mutated a new variable to describe the percentage of an animal's life spent in captivity, simplified the status levels to dead or alive (e.g. stillborn animals were considered dead, released animals were considered alive), simplified the acquisition levels to born into captivity, wild, or unknown (e.g. stillborn animals were considered born into captivity, rescued animals were considered wild), simplified the species variable to bottlenose or not bottlenose (to reduce the original thirty-seven species levels to two), simplified the current location variable to the ten most common locations (to reduce the original eighty locations), and removed observations with impossible birth years relative to origin year and status year.

The code book describing variables used in our analysis is below.

Variable	Class	Description
species	character	Species of animal (whale or dolphin)
sex	character	Sex of animal
acquisition2	character	Method through which an animal was brought into captivity
currentlocation2	character	Location of animal at date of recorded status
transfer	binary	Whether current location matches origin location
foreigntransfer	character	Whether an animal was transferred from outside the US
status2	binary	Binary indicator for status (dead or alive)
age	integer	Age of animal at date of recorded status
captivity	integer	Percentage of animal's life lived in captivity

Table 1 (below) shows summary statistics for the main variables of interest in our analysis. The mean percentage of life spent in captivity for cetaceans in the `all_cetaceans2` data set is 0.91 percent. The proportion of cetaceans born in the wild and born into captivity is roughly proportional at 51% and 48%, respectively. The status of 67% of cetaceans in our data set was recorded at a different location than where the cetacean entered captivity, telling us that they were transferred (either from the wild or another facility) at least once in their lifetime.

Table 1

Characteristic	**N = 1,416**
age	12 (5, 22)
status2	930 (66%)
species2	NA
aOther	252 (18%)
Bottlenose	1,164 (82%)
sex	NA
F	748 (53%)
M	635 (45%)
U	33 (2.3%)
acquisition2	NA
Born	678 (48%)
Unknown	13 (0.9%)
Wild	725 (51%)
captivity	0.91 (0.70, 1.00)
Unknown	140
transfer	944 (67%)
foreigntransfer	NA
Foreign	19 (1.3%)
US	1,397 (99%)

We conducted exploratory data analysis on the relationship between our categorical variables of interest (species, sex, acquisition method, transfer status, foreign transfer, and current location) and cetacean survival rates. A plot of the Kaplan-Meier survival time curve for acquisition method shown below (Figure 1) suggests that cetaceans born into captivity tend to have lower survival rates at the beginning of their lives and greater survival rates at the end of their lives when compared to cetaceans captured from the wild. It appears that after 17 years, cetaceans born in captivity tend to outlive cetaceans captured from the wild. Figure 1 shows that the proportional hazards assumption may not be met since the hazard curves for cetaceans born in captivity and cetaceans captured from the wild are not proportional (the two lines cross at approximately 17 years). Also apparent is a steep fatality at age zero among cetaceans born into captivity. The steep fatality represents cetaceans that were born into captivity and died soon after birth (or were miscarried). The difference between survival of cetaceans born in captivity and those born in the wild at very young ages is likely exaggerated since only wild cetaceans that survive to a certain age are represented in the data set. The lack of representation of young deaths and stillbirths among cetaceans in the wild may account, in part, for the violation of the proportional hazards assumption.

Figure 2 shows that the proportional hazards assumption is more reasonable for a variable that is not directly related to whether a cetacean was born in captivity or in the wild. The Kaplan-Meier survival curve in Figure 2 plots the survival rate by sex and shows that the survival times for male and female cetaceans appear to be roughly proportional.

Figure 1

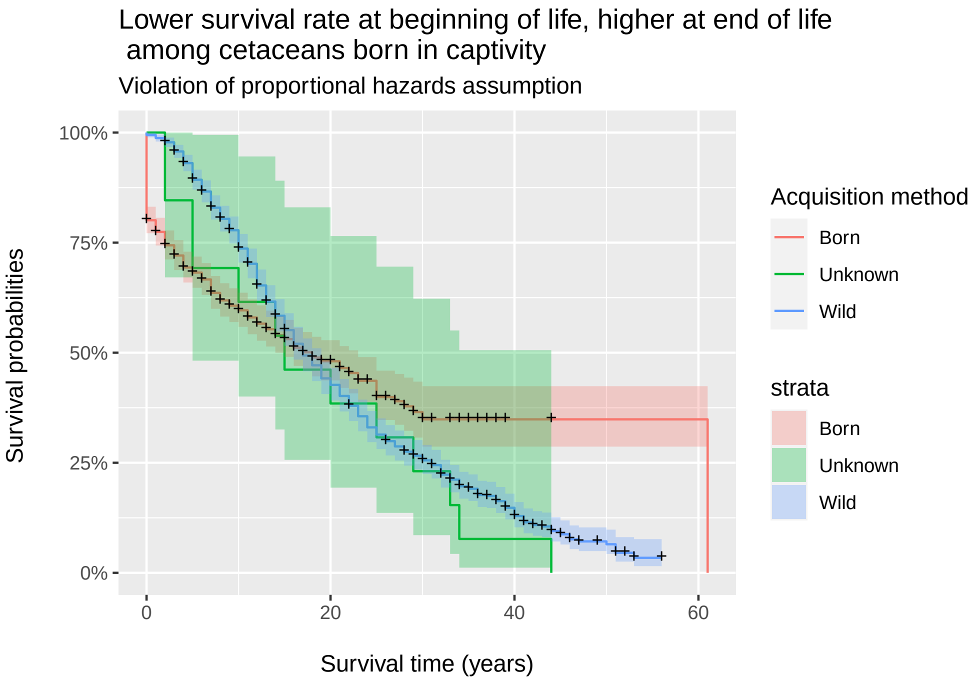
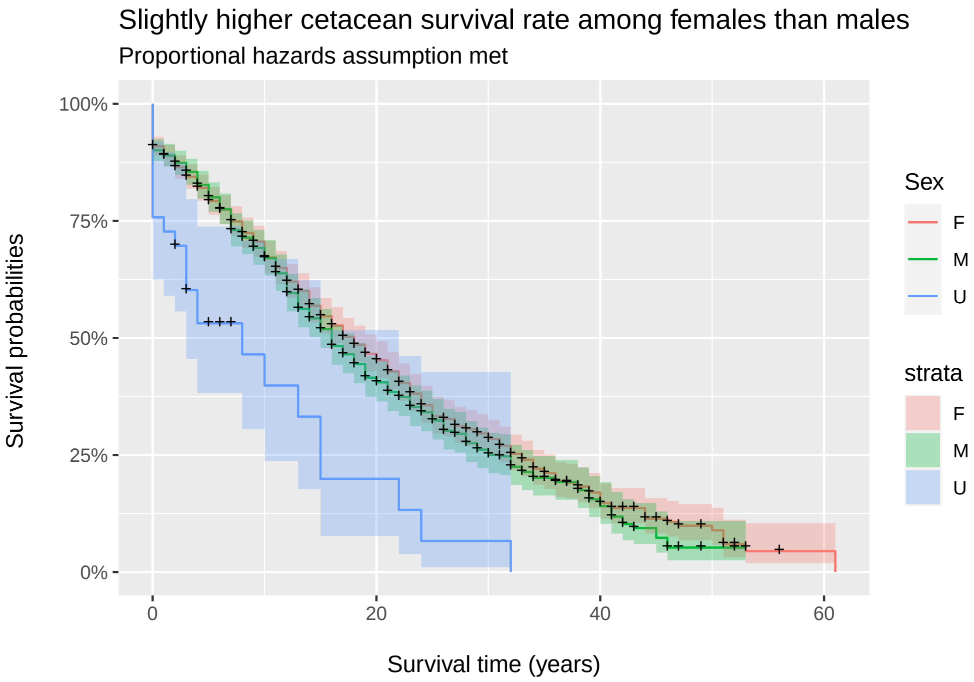


Figure 2



Code for survival visualizations from dkmathstats Website (David (n.d.)).

Methodology

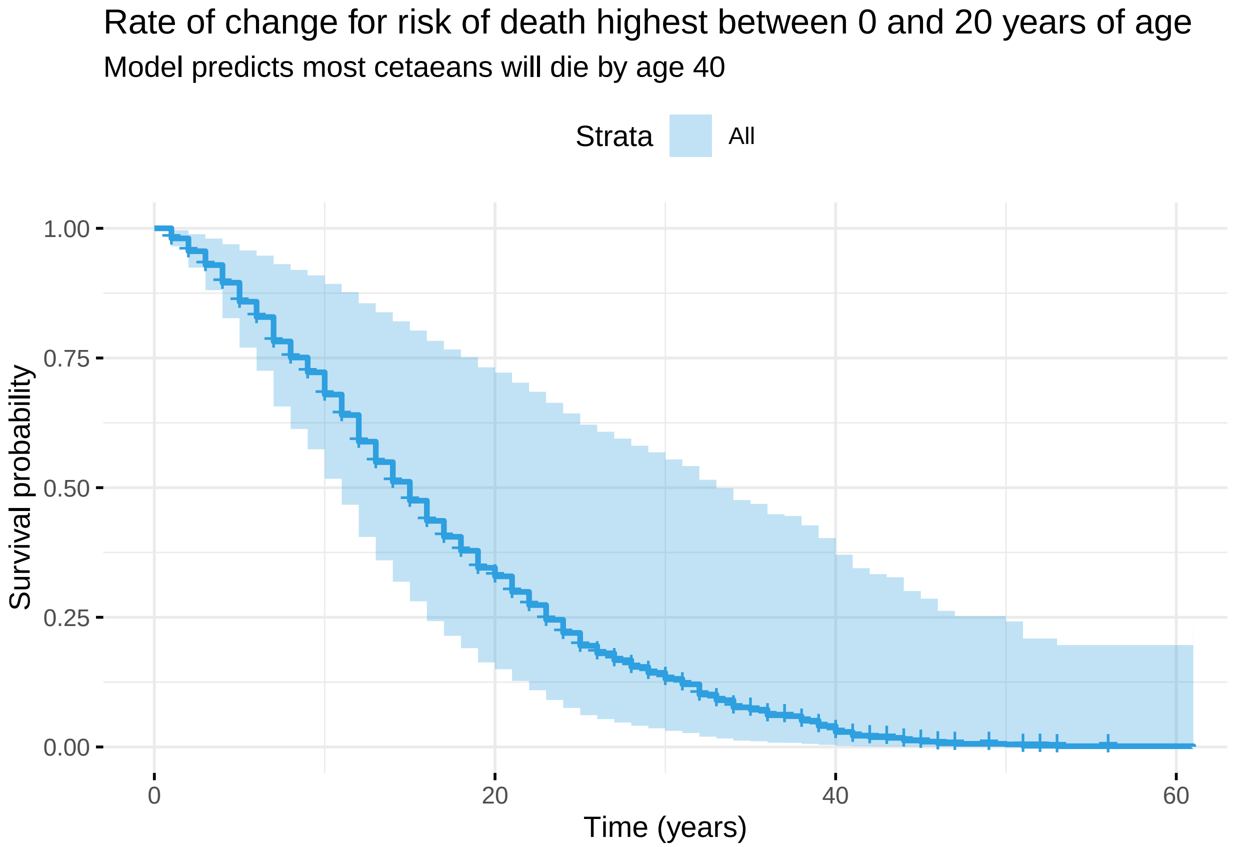
We chose to fit a multivariable Cox proportional hazards model because we want to predict survival rates of cetaceans in captivity based on a mix of categorical, binary, and numeric variables. Thus, the Cox proportional hazards model proved most appropriate in how it allows one to "evaluate simultaneously the effect of several factors on survival" (Kassambara (n.d.)). We then evaluated whether or not the assumptions of the Cox Proportional-Hazards Model were violated for our data set. First, the independence of observations assumption is reasonably satisfied because knowing something about one Dolphin does not tell you anything about another, seeing that we controlled for location in the model. Secondly, the assumption for the Cox proportional hazards model that censoring and death are independant is shown to be satisfied because there is no relationship between an animal being censored and its survival. Namely, an animal being released or living past the end of the observation period and thus being censored from the study does not affect the likelihood of its survival. Finally, the proportional hazards assumption of the Cox proportional hazards model is shown to be violated in that the hazard curves of different groups cross. The proportional hazards assumption states that "the hazard of the event in any group is a constant multiple of the hazard in any other" and thus the hazard curves of different groups crossing violates this assumption seeing that the group's hazard is no longer a constant multiple of another (Kassambara (n.d.)). However, we will proceed on in spite of this but make sure to take note of it when interpreting our results.

The outcome variable of interest in our model is the hazard rate, or the risk of death for cetaceans living in captivity. The time variable is `age` (in years) and the event of interest is `status2`, or the status of an animal as alive (0) or dead (1). In terms of predictors variables, then, we considered all unique, informative, and usable categorical and numeric variables in the `all_cetaceans2` data set. We identified species, sex, acquisition method, transfer status, foreign transfer status, and captivity as meeting the criteria for predictors. Other variables found in the `allCetaceanData` data set such as mother, father, origin location, list of transfers, and cause of death had too many variable levels to use in the model. We also included a simplified version of the current location variable in our model in order to control for differences in survival rate based on the facility where cetaceans are held in captivity and prevent a violation of independence, as mentioned above.

Results

The final model is printed below.

Predictor	Coefficient	Std. Error	Statistic	P-value
species2Bottlenose	-0.4062	0.0966	-4.2035	0.0000
sexM	0.1738	0.0751	2.3128	0.0207
sexU	0.7554	0.2674	2.8250	0.0047
acquisition2Unknown	-0.5234	0.3130	-1.6720	0.0945
acquisition2Wild	-0.3089	0.1370	-2.2545	0.0242
captivity	-2.3513	0.1502	-15.6559	0.0000
transfer	-0.4102	0.1550	-2.6467	0.0081
foreigntransferUS	0.4172	0.3285	1.2700	0.2041
currentlocation2Discovery Cove	-0.9416	0.3119	-3.0190	0.0025
currentlocation2Dolphin Research Center	-0.2323	0.1737	-1.3373	0.1811
currentlocation2Marine Life Oceanarium	0.8279	0.1617	5.1212	0.0000
currentlocation2Marineland Florida	0.0328	0.1840	0.1783	0.8585
currentlocation2Miami Seaquarium	0.5366	0.1816	2.9540	0.0031
currentlocation2SeaWorld Orlando	-0.0575	0.1568	-0.3670	0.7136
currentlocation2SeaWorld San Antonio	0.0888	0.1604	0.5535	0.5799
currentlocation2SeaWorld San Diego	-0.4571	0.1502	-3.0433	0.0023
currentlocation2U.S. Navy	-0.4717	0.1160	-4.0670	0.0000
currentlocation2Unknown	0.0937	0.2175	0.4307	0.6667



Code for visualization from Statistical tools for high-throughput data analysis (Kassambara (n.d.)).

The survival plot above shows that the rate at which survival probability decreases is roughly constant between the ages of zero and twenty, then slows down between the ages of twenty and thirty, then tapers off between thirty and sixty years of age. The model output shows that the predictors for species (Bottlenose), sex (male), sex (unknown), acquisition method (wild), percentage of life spent in captivity, transfer between US facilities, current location (Discovery Cove), current location (Marine Life Oceanarium), current location (Miami Seaquarium), current location (SeaWorld San Diego), and current location (U.S. Navy) all have significant p-values at the $\alpha = 0.05$ significance level. The p-value for the `captivity` variable was notably small at less than $2e^{-16}$.

To then analyze the results that directly pertain to our research question, we will interpret the captivity, acquisition method, and transfer status parameters. First, the parameter corresponding to `captivity` can be interpreted as meaning that a Dolphin that has lived in captivity 1% of its life more than another is predicted to have $e^{-2.3513} \approx 0.095$ times the hazard of death than the other, while controlling for all other variables in the model. This result addresses our research question because it suggests that Dolphins which live greater percentages of their lives in captivity actually tend to live longer. Next, in looking at the categories of the `acquisition` parameters, the parameter corresponding to acquisition from the wild can be interpreted as meaning that a Dolphin that was taken from the wild into captivity is predicted to have approximately $e^{-0.3089} \approx 0.734$ times the hazard of death compared to a Dolphin born into captivity, while controlling for all other variables in the model. Similarly, a Dolphin whose acquisition method is unknown is predicted to have approximately $e^{-0.5234} \approx 0.593$ times the hazard of death compared to a Dolphin born into captivity, while controlling for all other variables in the model. Both of these results help us answer our research question because they have to do with how the animals entered captivity and thus shed light on the effects of captivity on the animals. Specifically, we learned that animals that were taken from the wild into captivity, or came into captivity through unknown means tended to live longer lives compared to those born into captivity.

Finally, the `transfer` parameter can be interpreted as meaning that a Dolphin whose current location does not match their origin location (whether the Dolphin was transferred at some point in its life) is predicted to have approximately $e^{-0.4102} \approx 0.664$ times the hazard of death compared to a Dolphin whose current location does match their origin location. This result contributes to answering our research question because it sheds light on the effects of a specific attribute of captivity: being transferred. Namely, we learned that Dolphins that have been transferred at least once in their lives tend to live longer than those that are not.

Discussion

Our model predicts that while controlling for all other variables in the model, cetaceans born in the wild tend to live longer in captivity than cetaceans born into captivity. Further, cetaceans living at a different location than their origin location (this includes cetaceans born in the wild and brought into captivity) tend to live longer than cetaceans that are living at the same location as their origin location (while holding all other variables constant). At the same time, the model predicts that cetaceans that live higher percentages of their lives in captivity tend to live longer than cetaceans that live lower percentages of their lives in captivity (while holding all other variables constant). These results do not support our hypothesis that cetaceans that spend more of their lives in captivity will live shorter lives than cetaceans that spend less of their lives in captivity. That said, there is evidence to suggest that being born in the wild but transferred into captivity at a young age may have some survival advantages.

Our results may seem contradictory. As mentioned in the introduction, our interpretations and conclusions are limited by the fact that the data set we used doesn't contain any observations of stillbirths, miscarriages, or infant mortality that occur in the wild. Therefore, the difference between cetacean survival rates among cetaceans born in the wild and born into captivity at zero years of age is likely more extreme in our data set and in our model than in the population. If cetacean stillbirths, miscarriages, and infant mortality among wild cetaceans weren't underrepresented in our data set, we might see that cetaceans born in the wild tend to live shorter lives than cetaceans in captivity at all ages.

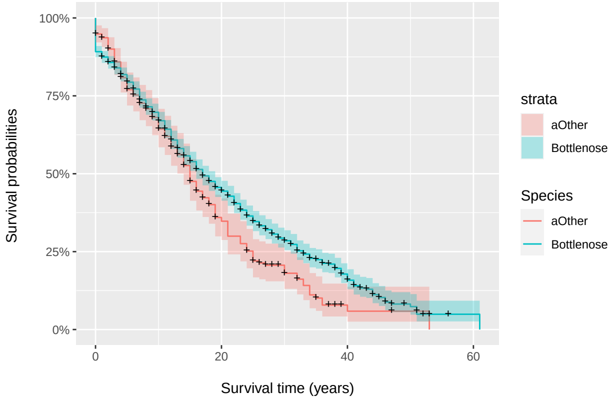
Another limitation of our analysis is that we have reason to believe that some of the data in the `all_cetaceans2` data set is unreliable. We removed 84 observations with impossible birth dates relative to origin date or status date (e.g., the status date was listed as 2003 and the birth date was listed as 2009). These impossible observations suggest that there was human error involved in the collection of the data used in our analysis. The data set description in the article from The Pudding states that some of the data was collected from the crowd-sourced website Ceta-Base. Further analysis may want to only consider data collected from verified official sources to reduce human error and address the reliability issues of the data set used in our analysis.

When initially considering how to select our predictor variables, we considered using a variable selection technique such as LASSO regression or all subset selection. However, in the case of LASSO regression we found that using LASSO as a means of selecting variables for a Cox model was relatively new and involved statistical principles which were beyond the scope of our current understanding (Tibshirani (1997)). Similarly, in researching the feasibility of using all subset selection, we found that the answer to which variable selection technique is best for the Cox proportional hazards model is still up for debate (Ekman (2017), Fan and Li (2002), Petersson and Sehlstedt (2018)). Thus, to avoid misusing the appropriate variable selection technique for our project, we decided to instead use all of the unique, informative, and usable predictor variables from our data set in our model. As a result, it is possible that a more effective model could be derived if future researchers were to beforehand use one of the aforementioned variable selection techniques.

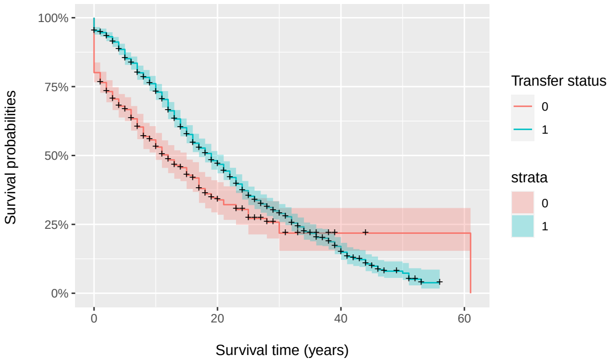
Additionally, as mentioned in the methodology section, not all of the assumptions for the Cox Proportional Hazards Model were shown to be satisfied. Subsequently, the interpretations and conclusions we are able to draw from our model are inherently limited. Future work could be done to address the assumption violations and perform a more rigorous analysis.

Appendix

The following figures explore the relationship between survival rates and predictor variable included in our model but not necessary for our argument in this project.

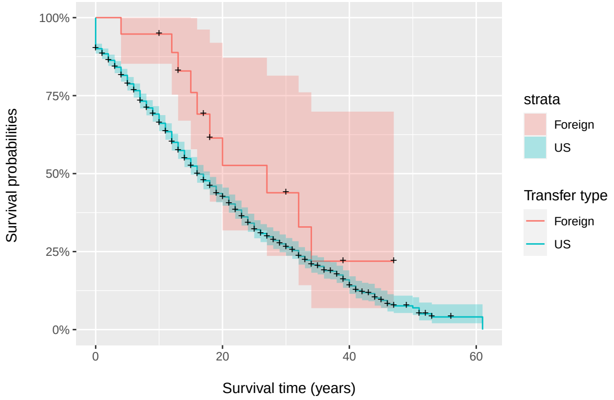


Survival time of cetaceans in captivity by species

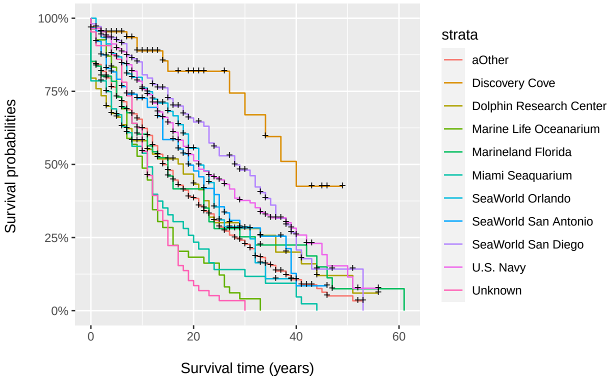


Lower survival rate at beginning of life, higher at end of life among cetaceans transferred between facilities

Violation of proportional hazards assumption



Survival time of cetaceans in captivity by transfer type



Cetaceans currently at Discovery Cove have highest survival rate

Violation of proportional hazards assumption

References

Data.world. 2017. "Captive Whales and Dolphins in the US (1938 - 2017)." Data.world. https://data.world/the-pudding/cetaceans/activity.

David. n.d. "Plotting Kaplan-Meier Survival Times Curves in r with Ggplot2." dkmathstats Website. https://dk81.github.io/dkmathstats_site/index.html.

Ekman, Anna. 2017. "Variable Selection for the Cox Proportional Hazards Model: A Simulation Study Comparing the Stepwise, Lasso and Bootstrap Approach." PhD thesis, Umeå University. http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-130521.

Fan, Jianqing, and Runze Li. 2002. "Variable Selection for Cox's Proportional Hazards Model and Frailty Model." *The Annals of Statistics* 30 (1): 74–99. https://doi.org/10.1214/aos/1015362185.

Kassambara, Alboukadel. n.d. "Cox Proportional-Hazards Model." Statistical tools for high-throughput data analysis. http://www.sthda.com/english/wiki/cox-proportional-hazards-model.

Petersson, Simon, and Klas Sehlstedt. 2018. "Variable Selection Techniques for the Cox Proportional Hazards Model: A Comparative Study." PhD thesis, University of Gothenburg School of Business. https://core.ac.uk/download/pdf/152600668.pdf.

Thomas, Amber. 2017. "Free Willy and Flipper by the Numbers." https://pudding.cool/2017/07/cetaceans/.

Tibshirani, R. 1997. "The Lasso Method for Variable Selection in the Cox Model." *Statistics in Medicine* 16 (4): 385–95.