The Use of Survival Analysis Techniques Among Highly Censored Data Sets

Shelby Marie Cummings

April 18, 2013

**Abstract**

 The purpose of this research project was to look at the various survival analysis techniques and determine if there was either a way of fixing these methods or a better method to use in the case of data sets with a large percentage of censored data points. Censored data points are data points where an exact survival time is not known, simple a lower bound on the survival time is known. This research used data simulated with the Weibull distribution to look at how well current survival analysis techniques work as censoring rate changes and found an improved model that can better predict the real world survival model in the case of high amounts of minimum censoring.

**Introduction**

Every day, scientists and medical researchers are working around the clock to develop new medicines, techniques, and surgical procedures to enhance the quality and length of the human life. Whether it is a preventative measure that can influence the entire population, or a new medication that can better treat a particular medical condition, these discoveries are vital to the medical community and the population as a whole. Once someone has come up with a new medical technique, how can one tell that this new technique actually works better than the previous methods. A researcher starts by conducting studies and collecting data on the life spans of patients receiving both the old and the new treatment methods. They then look at these life spans and see if there are any differences. Then the question becomes, how can we tell whether these differences are simply due to chance, or are statistically significant?

At the very core of statistics is determining whether or not the differences we see in the world are actually systematic differences or if they are just occurring by chance. Determining statistical significance would appear to be a good starting point to determine whether or not these new treatments are actually working better than the previous methods. The question then arises, what makes a treatment “better?” Is a better treatment one where the average life span of recipients is longer? Or does it go beyond that? Is the best treatment one where everyone lives at least 1 year? This is where survival analysis comes into play. Survival analysis is a prescribed way of measuring the survivability of patients receiving different types of treatment.

**Background**

The general idea of survival analysis is to create what are known as survival models or survival curves. A survival model or curve is a way of representing how long a patient with a given disease or condition will survive (Balakrishnan 2004). While there are quite a few different methods of doing this, the overall idea is to create a model showing the probability of surviving a certain length of time for each type of treatment. These models are based off of empirical data collected by medical researchers of people who are actually being treated using either the new or the old method(Guo 1991). The most simple versions work based solely on how long each individual survived, but more complicated versions also exist, which can take into account a number of other variables (medical history, pre-existing diseases, lifestyle, ethnicity, age, gender, etc.), which are considered to be important in the medical world (Selvin 2008).

For example, Table 1 is a sample, simulated data set depicting the kind of data a researcher would have at the end of his study.



**Table 1:** This is a table of sample survival analysis data that would be available to a researcher at the conclusion of a study.

From this data, the researcher would create a survival model (Figure 2) that showed the percentage of people surviving as a function of time.



**Figure Two:** This shows a stepwise survival model from empirical data in Table 1.

This step function survival model takes the raw data and looks at what percentage of the participants are still living at each point in time. This survival model is then made into a “smooth curve” survival model by fitting a distribution, normally a Weibull distribution, to the original step function. Figure 3 shows what a typical, smooth Weibull approximation looks like. It is this model that is used as the end survival model.



**Figure 3:** This is the Weibull Approximation to the Survival Model in Figure 2.

The idea of a survival model is useful for many reasons, to doctors and patients alike. For doctors and researchers, it is considered to be more reliable and paint a better picture of the scenario as a whole, because instead of focusing on a single statistic, like the mean or the median lifetime, it looks at survival over a period of time. For patients, it really gives a better idea of what to expect when it comes to their condition. For example, a person diagnosed with a particular type of cancer may be told that the average survival time is three to five years. But for someone who has just been diagnosed with a serious illness such a cancer, more information is generally desired. For example, a recently diagnosed patient may want to know how many years 90% of patients live, or the percentage of people that live beyond 1, 5, 15, etc. years. When describing data, the average, or some sort of measure of center is generally a good place to start, but it really doesn’t fully tell what is going on with the data. Even though the average lifespan for a particular disease may be three to five years, many people will die from this condition before three or after five years. This average doesn’t let researchers and patients know if the data is skewed (meaning the distribution of survival times is not symmetrical) and it doesn’t take into account any other variables, such as history or other health conditions. By creating survival curves, specifically ones that take other variables into account, a patient is provided with a much better idea of what their prognosis actually is (Selvin 2008).

It should be noted that survival analysis is also useful in a variety of different fields, not just in medicine (Fleming and Harrington 1991). For example, survival analysis can be very useful to determine the “lifespan” of a product or piece of equipment. While this is more commonly referred to as reliability testing, it is in essence the same as survival analysis. Survival analysis techniques can also be used for subjects that are not human. For example, it can be used in a variety of biological studies to determine the lifespan of different animals, or how long plants survive when living in different conditions. Because the main use of survival analysis currently is in the medical field, this paper will use that as its main focus.

Survival analysis as a relatively new area of study still has its flaws. One of the biggest issues with survival analysis is the problem of censoring. Censoring is the issue of incomplete data. Due to the nature of this type of data, it is very common that researchers do not have an exact “survival time” for a particular patient. When a piece of data is “incomplete,” we call this data censored. There are two main types of censoring: right censoring and left censoring. Right censoring occurs when the data is unknown on the “right hand side,” or at the ending time. We know that a patient survived at least so long, but are not sure how much, if any, longer they have lived. Right hand censoring is the more common type of censoring. There are three main reasons for this type of censoring to occur:

1. The patient is still alive at the end of the study

2. The patient withdraws from the study

3. The researcher loses track of the patient.

The most common of these reasons is the first one, that a patient is still alive at the end of the study. Since these medicines and treatment options are new, researchers don’t have many years worth of data on them. Many times, a patient will be completely cured from a disease and go on to live for a normal lifespan, but this type of information isn’t gathered in a five or ten year study. So, at the end of the study, the researcher can only report that a patient lived at minimum, the length of time the patient was in the study, and after that, we don’t know what happened. Because right censoring is so common, many survival analysis techniques are able to account for this in their predictions(Houwelingen and Hein 2012).

A less common type of censoring is known as left censoring, which occurs when the data is cut off at the “left hand side,” or the starting time. This doesn’t occur with testing different types of treatments as much as when researchers are just looking at life expectancies of people with certain disorders in general. For example, if a researcher wanted to create a survival curve showing the survivability of people with AIDS, they would begin by finding a group of people who have been diagnosed with AIDS, and follow them to see how long they live, once they have been diagnosed with AIDS. The problem with this is that the date of diagnosis is not the same as the date they actually contracted the disease. The date of diagnosis simply tells a researcher that by that date, this individual had contracted the disease. When that person dies, we know that they lived with AIDS from the date of diagnosis until the date of death, plus whatever time they lived with the disease before being diagnosed (Klein and Moescherger 2003).

An additional way that I have come up with and used to categorize censoring is minimum vs. maximum censoring. Minimum censoring occurs when the data values most likely to be censored are the smallest values in a data set. This occurs in certain types of research studies where researchers do not begin the study with their full set of participants. With rare diseases, such as particular kinds of cancers, it is hard to find a lot of participants, so researchers will add more participants to an ongoing study as they find them. In this case, the people with lower survival times are likely to be those that were diagnosed and came into the study after it initially began. Since they were diagnosed at a later point in time, they are more likely to still be alive at the end of the study and have a censored data point.

Maximum censoring occurs when the data values most likely to be censored are the largest values in a data set. This occurs in research studies where the researchers begin the study with a set group of people they will follow through the duration of the study. In this case, those that are most likely to be censored are those participants that survive the longest, for two reasons. The first reason is that the longer a participant is in a study, the greater the chance of them getting lost by the researcher. The second reason is that those who live beyond the time period of the study will all be censored, and since everyone began the study at the same time, the participants that live beyond the time period of the study will also be the people who have the longest survival times.

In all cases, left and right censoring as well as maximum and minimum censoring, researchers are working with the min(Ci, Ti), which is the minimum of the censored survival time and the actual survival time. We know that the patient lived at least as long as the censored time, but we don’t know how much longer they lived.

In Table 4, we see an extension of the early example used to depict a survival model. In the original example, the only data we used was the subject number and the survival time. Now that we know about censored data, we can add that to the example. Each subject has three different survival times: the actual survival time, the censored time, and the experimental survival time. 

**Table 4:** This table extends the survival data example in Table 1.

The actual survival time is the actual amount of time that the patient survived. The censored time is the length of time of the study. The experimental survival time is the minimum of the actual and the censored survival times. The experimental survival time is the only piece of data that the researcher has to use for their analysis, in addition to an indicator variable denoting whether or not a particular subject’s experimental survival time is censored or not. For the purposes of this research and due to the coding of the software used, the indicator variable 0 means censored, while 1 means uncensored. Table 5 below shows an example of what data from the chart above a researcher would actually have available to analyze.



**Table 5:** This table is the survival analysis data that is available to the researcher.

We can also classify censored data as being informative, or non-informative. In the case of informative censored data, we have some sort of other background information that can give us clues about the censored patient. For example, if a person who is responding well to treatment drops out of a study, it is generally an indication that the person is doing well and has a survival time higher than the censored time; otherwise they would have continued treatment. On the other hand, if a person is not responding well to treatment and decides to drop out of the study, it could be a clue that this person’s actual survival time is not much greater than their censored survival time. Informative censoring really doesn’t affect the survival curves created, but the information can be taken into account when analyzing the survival curves that are created.

Unfortunately, most of the time researchers do not have the luxury of informative censoring. Most of the time, the data is censored because a patient has not yet died at the point in time a study has ended. In this case, researchers have no idea how much longer the patient will live. Some would say the best way to deal with this is to simply throw out the censored data. But there are two problems with this approach. First off, this is removing very valuable data. Especially in cases where researchers are looking at survivability of rare diseases, it is hard enough to find enough participants to start with. Once a researcher starts throwing out all the patients who are censored, they are stuck with an even smaller sample size, and they run the risk of having data that doesn’t really reflect what is going on in the population as a whole (Machin 2006).

The second problem with doing this is that researchers would not only be eliminating valuable data, but they would also be systematically eliminating data. In general, for right hand censoring, the patients who have censored survival times are generally the patients who are living longer, because, for the most part, the censored data points are the ones who are still alive at the end of the study. By eliminating these data points, you are systematically removing only the higher survival times, and are therefore biasing the data. This would give inaccurate survival curves, so we must come up with another solution. But waiting for each person in the study to die isn’t really a feasible solution either, because that would take so long that an even better treatment would already be out by the time the study was over, and the treatment that so much time and energy was spent researching would already be obsolete. So researchers and statisticians must learn to work with the data they have, censored times and all.

**Problem Motivation**

For the most part, survival analysis models used to create survival curves are fairly sturdy and robust when the censoring rate is relatively low. Not all models are built equally, and some will always outperform others, but in most cases, as long as the rate of censoring is below 10%, the survival models produced will be fairly accurate. This seems like a reasonable number, but in many cases, it just isn’t feasible to get a censoring rate that low. This problem was first brought to my attention during a summer research internship.

Under the assistance of our faculty advisor, my research team worked to create survival models for pediatric patients undergoing various types of liver transplants. In this case, the hope is that a child will receive a liver transplant at a young age, and continue to live a healthy life with a normal lifespan. For a study of this nature, it is quite reasonable to expect that, if the transplant goes as planned, there is a good chance of the child living five or even ten years after they received the transplant, if not more. For a study of this nature, researchers hope and expect that the censoring rate of a ten-year study would be more than the prescribed 10%. In the study my research group was using, there ended up being an overall censoring rate of 92.1%, with group censoring rates ranging from 87.9% to 93.5%. We worked hard to find literature on what to do with censoring levels this extreme, but we found nothing. This level was so extreme; it was beyond the normal bound of “high levels of censoring,” which is normally defined as any censoring level between 30% and 50%. We tried to find previous studies dealing with data this censored, but everything we could find used methods specific to the situation/data, and therefore were not applicable to our data set. We continued on with the analyses, and created the survival models using the information we had, but felt that any model we came up with was inconclusive due to this extreme level of censoring.

Ever since, I have wondered, was there something else we could have done. Was there another method we could have used or was there a way we could alter the current method so that we could create a more reliable and accurate survival model, even with such a censored data set? It seemed like there had to be something that could be done. And I set out to determine what, if anything, could possibly fix that problem.

**Research Process**

I started out this project by looking first at how survival analysis techniques work under ideal conditions. In order to determine how these techniques are affected by different levels of censoring, it is important to understand the underlying principles of how they work normally. I began by looking at two of the most common survival analysis methods: the Kaplan-Meier Method and the Cox Proportional Hazards Model. Both of these models try to create survivability curves by determining the probability of living past time t, but do so in different ways.

I then began experimenting with some of my own fabricated data to get a feel for how these two methods behaved when used with censored data. I created a random data set, using the Weibull distribution with a shape parameter of 10 and a scale parameter of 2. I considered this to be the equivalent of a data set that was complete with no censored values. The reason I chose to use the Weibull distribution is because it is one of the most widely used distributions to model survival curves. Between the shape parameter and the scale parameter, it can be manipulated to create a variety of survival models that tend to fit almost all survival data. The formula for the density of a Weibull distribution is , t > 0, α > 0, β>0, where α is the shape parameter and β is the scale parameter (Randles 1991).



**Figure 6:** Weibull Distributions with varied Shape Parameters to show the variety of ways a Weibull Distribution can look.

The figures, Figure 6 and Figure 7, show what happens when you vary the shape and the scale parameters. The X in the cases of these two graphs can show a variety of things, in the case of survival analysis, the X value is the survival time.



**Figure 7:** Weibull Distributions with varied Scale Parameters to show the variety of ways a Weibull Distribution can look.

Once a researcher has a Weibull density model, they can then very easily take the CDF, or cumulative distribution function, of the model and subtract it from one to get a survival model. The original probability density function shows the probability of a patient surviving in a specific time range. The CDF shows the probability of a patient living no more than a certain amount of time. The survival model is the opposite of the cumulative probability. It shows the probability of a person living, at minimum, a certain amount of time. Figure 8 shows the progression from Weibull density function, to CDF, to survival model.



**Figure 8:** Weibull Density Model, Cumulative Density Function, and Survival Model.

I then manually went in and censored this data set using different degrees of censoring. For my first data set, I used censoring values of 1%, 5%, 10%, 15%, 20%, and 25%. Since, for the most part, censored survival times are normally the longest survival times, I changed the largest 1%, 5%, 10%, 15%, 20%, and 25% of values to be censored. This did not change the actual value recorded; it only added an indicator to the value to say that it was an incomplete piece of data, and that all that was known was that the true survival time was no less than the given value. For each level of censoring I ran tests to create both a Kaplan-Meier Curve, as well as a Cox Proportional Hazards model, which I compared to the Kaplan-Meier Curve and the Cox Proportional Hazards model curve for the original data set with no censored values. The Kaplan-Meier Curve and the Cox Proportional Hazards Model are currently the two techniques most widely used in survival analysis to create estimates for the actual survival curve parameters. For the most part, I saw that the Kaplan-Meier Curve as well as the Cox Proportional Hazards model both retained their original shapes as the level of censoring increased, but the Kaplan-Meier Curve stayed truer to the scale of the original data set.

After viewing the survival curves created for each censoring level and visually inspecting them for any differences, I decided to find a more quantitative way to determine what was happening at each censoring level. I chose to work exclusively with the Kaplan-Meier Curve because it seemed to be doing a better job of representing the true nature of the data, even with the amounts of censoring that I had tested. Also, the Kaplan-Meier Curve is much more widely used in the field of survival analysis. I again created a random survival data set from the Weibull distribution with a shape parameter of 10 and a scale parameter of 2. At each level of censoring that had been used previously, I created a Kaplan-Meier curve, which gave me an estimate to the shape and scale parameter of the distribution the data had come from. I was looking to see if there were any patterns in how these parameters changed as the censoring level increased. With this single data set, it seemed like the shape parameter increased linearly as the censoring value increased. Due to this indication, I decided to do a larger scale simulation to see how censoring affected the shape parameter.

Using Minitab, I began to run a series of simulations that would create many different data sets. For each simulation, I specified the sample size, the number of datasets to use, the different shape and scale parameters to use, and the number of censorings to do. Figure 9 shows the simulation frame used.



**Figure 9:** Simulation Frame used to Collect Data.

I first began by letting the simulator randomly censor the data. For each censoring rate, it would randomly pick that percentage of the points and give them the “censored indicator.” Again, it did not change these data values; it simply marked those points as being censored. The simulator then used the Kaplan-Meier Estimate to estimate the shape and scale parameter of the distribution the data came from. I decided that the main focus of my studies would be to determine how well the Kaplan-Meier estimates the shape parameter and try and create my own method or correction factor that could give a better estimate of the shape parameter. When the simulation was finished running, I was supplied with a lot of data about the data set. My first task was to determine what of this data was important and could tell me something valuable about my data set. I obviously wanted to look at the estimated shape parameter and compare it to what the true shape parameter was in order to determine how far off the estimates were once the data became censored. But, the big question was, what should I use for the “true” shape parameter, because there were two things that seemed appropriate to use. My first choice was to use the shape parameter I had supplied the simulator with. When I supplied the generator with the shape parameter, it created a random Weibull distribution from that parameter, so it made some intuitive sense to use that as my true value. But the simulator also calculated the shape parameter given the entire generated data set, before any of the values were censored. Table 10 below shows the typical differences between the supplied shape and scale parameters that I put into the simulator and the calculated shape and scale parameters for the data set.

 To me, it made more sense to use the calculated parameters as my “true values.” It was indicative of what the shape and scale parameter would be if none of the data were censored, which is what I was ultimately trying to predict.

 **Table 10:** Typical Differences between Supplied and Calculated Parameters.

Once I decided on what my true value was, I created a “Bias” column in my data set, which took the estimated shape parameter and subtracted the true value. I then used this to create a “Percent Bias” column, by dividing the “Bias” by the true value and multiplying by 100. The resulting data set is shown in Figure 11.



**Figure 11:** Data Created from Simulation.

Using a scatterplot, shown in Figure 12, I looked at how the Percent Bias changed as the censoring rates increased. Each dot in the scatter plot represents a single run through the simulation; for each run in the simulation, the percent bias was calculated and this scatterplot gives a visual representation of how the percent bias changes as the censoring rate increases.



**Figure 12:** Scatterplot showing how the percent bias changes as the censoring rate increases for random censoring

As seen in the scatterplot, it appears as though the bias is centered around zero for all levels of censoring when the censoring is done randomly. I ran some basic statistics, with the means being reported in Table 13, on the data to see if this observation held true.

Censor

Variable Rate Mean

PercentBias 25 3.62

 35 1.24

 45 1.27

 55 -0.871

 65 1.145

 75 0.162

 85 0.715

 95 -0.258

**Table 13:** Basic Statistics for Percent Bias by Censoring Rate for Random Censoring

 Based on these statistics, which backed up what I saw in the scatterplot, it appeared as though when censoring was done randomly, the amount of bias in the estimate of the shape parameter was, on average, zero. That is not to say that each estimate was perfect and did a good job of estimate the true value of the shape parameter. By looking at the scatterplot, I could tell that some of the values had a Percent Bias as high as 100%. What this means is that there didn’t appear to be a systematic change in the bias as the censoring rate increased. The estimators did not begin to overestimate or underestimate as the censoring rate increased. Because of this, there wasn’t really anything that I would be able to fix in the estimator.

 I then began looking at other ways of censoring, other than randomly picking points to be censored. On one hand, random censoring seemed to make sense with what happened in the real world. The researcher had no way of knowing which people would be censored. But just because the researcher doesn’t always know the reason behind why subjects are censored doesn’t mean that subjects are censored randomly. There has to be something behind why certain subjects are censored. My first thought was that subjects with longer survival times are more likely to be censored. Because most of these studies are typically 5 or 10 year long studies, if someone is still living at the end of the study, they will become a censored value. The people that are still living at the end of the study are the ones who have lived the longest, and therefore, the people living the longest may be more likely to be censored.

 I decided to rewrite part of the simulator in order to change the random censoring to a “maximum censoring.” This time, instead of taking the censoring rate and randomly picking that percentage of values to be censored, the simulator ordered the values from least to greatest and let the top percentage of them be censored. I again went through once the simulator had finished running and added in the bias and the percent bias.

 In order to see if the percent bias increased as a function of censoring rate, I created a scatterplot, Figure 14, of censoring rate vs. percent bias.



**Figure 14:** Scatterplot showing how the percent bias changes as the censoring rate increases for Maximum Censoring.

 As with the random censoring, it appeared as though the percent bias of this estimator was centered around zero, which would mean the estimator is unbiased. This time, it appears as though the variability increased as the rate of censoring increased. In order to check these observations, I ran some basic statistics, Table 15, on my simulated data. This data confirmed my suspicion that the variability was increasing as the censoring rate was increasing, but the mean was staying centered around zero.

Censor

Variable Rate Mean StDev

PercentBias 5 0.019 0.931

 15 0.941 3.456

 25 -0.197 4.100

 35 -2.007 5.750

 45 2.664 8.250

 55 2.25 12.56

 65 -2.33 14.64

 75 3.73 23.71

**Table 15:** Basic Statistics for Percent Bias by Censoring Rate for Maximum Censoring.

As seen in Figure 16, I graphed the mean of the percent bias over the different censoring rates and ran a linear regression model to see if there was any change in the bias as the censoring rate increased

**Figure 16:** Mean Percent Bias as Censoring Rates Increase for Maximum Censoring.

While the linear regression trend line shows what could be a slight increase in the mean bias as the censoring rate increases, the R2 value of this regression analysis is only 8.4%. With a value so low, it is reasonable to believe that the mean doesn’t change due to the censoring rate and that the estimators created using this are unbiased.

 I then graphed the standard deviation of the percent bias against the censoring rate and ran a linear regression model to see if the standard deviation was truly increasing as the censoring rate increased (Figure 17).



**Figure 17:** Standard Deviation of Percent Bias as Censoring Rate increases for Maximum Censoring.

The scatterplot and resulting linear regression model show a distinct upward trend in the standard deviations of the percent bias as the censoring rate increases. With an R2 value of 89.9%, we can be fairly certain that there is a real association between the censoring rate and the standard deviation.

Unfortunately, there really isn’t that much that can be done to fix the standard deviation from increasing this way. It can be decreased by increasing the overall sample size, but there is really nothing in the way of the model that can be done to fix the problem of the standard deviation increasing as the censoring rate increases.

Since this estimator seems to be unbiased and there is no good way to fix the problem with the standard deviation, there wasn’t much I could do to try to fix this estimator. It was encouraging to find that when data are censored in this way, the model works fairly well for estimating the parameters. But what happens when data is not censored in this manner? One of my assumptions was that the data points most likely to be censored were those that were the largest values. My thought process was that researchers would begin the study with all of the participants at the same time and those who were still surviving at the end of the study would be the ones who were censored. By this logic, it would make sense that the largest data values would be the ones that were censored. This certainly makes sense some of the time, because many research studies are carried out this way. But this is not always true. Sometimes a research will begin their study with whatever participants then can find and will add more participants as they find them. This is especially prevalent in research that is done on rare diseases, because sometimes it takes quite a while to find participants. Because of this, people joining later in the study are more likely to be censored, because they haven’t had the disease as long and are more likely to continue living. This would mean that the people most likely to be censored would be those with the lowest data values.

In order to look at how this type of censoring affected the estimated parameter values, I changed the simulator to create data that looked at “minimum censoring.” Instead of the maximum values being marked as censored, now the minimum values were being marked as censored. I again ran the simulator, using a variety of shape and scale parameters and censoring rates from 5% to 85%. I calculated the bias and the percent bias for the shape estimate given by the Kaplan-Meier estimate as I had done with previous data sets. This time though, it appeared as thought the percent bias increased linearly as the censoring rate increased. I used the scatterplot in Figure 18 of the percent bias against the censoring rate to pictorially look at this trend.



**Figure 18:** Scatterplot showing how the percent bias changes as the censoring rate increases for Minimum Censoring.

This scatterplot certainly confirmed my suspicions that as the censoring rate increased, so did the percent bias. I looked at some basic statistics (Table 19), such as the mean and standard deviation, and graphed those as a function of censoring rate as well (Figure 20 and Figure 21) to get a better picture of my data.

Censor

Variable Rate Mean StDev

PercentBias 5 28.86 7.45

 15 58.79 14.01

 25 86.39 19.68

 35 119.09 31.73

 45 141.78 29.67

 55 175.84 37.50

 65 221.7 53.3

 75 247.3 69.0

 85 338.4 95.6

**Table 19:** Basic Statistics for Percent Bias by Censoring Rate for Minimum Censoring.



**Figure 20:** Mean Percent Bias as Censoring Rates Increase for Minimum Censoring.



**Figure 21:** Standard Deviation of Percent Bias as Censoring Rate increases for Minimum Censoring

**Results**

These statistics and scatterplots of how they changed with the censoring rate confirmed my suspicion that the amount of bias increased as the censoring rate increased. This made me feel very confident that I could find a correction factor of some sort to make the Kaplan-Meier a better estimate. I ran a linear regression model to see if I could find a way to better estimate the actual parameter value based on the Kaplan-Meier estimate from Minitab. Since the bias was increasing as the censoring rate increased, it meant that as the censoring rate increased, the estimators were over-predicting by larger amounts. Linear regression would look to find the trends as to how this over-prediction is occurring and would make a better estimate for the shape estimate. The variables I wanted to use to make this better prediction were the original calculated shape parameter (which was over predicting) as well as the censoring rate of the observation. Figure 22 below shows the outcome of running linear regression on this data.

The regression equation is

Shape Est Full = 1.55 + 0.302 Shape Est Cens - 0.0241 Censor Rate

Predictor Coef SE Coef T P

Constant 1.5527 0.1182 13.13 0.000

Shape Est Cens 0.30173 0.01125 26.82 0.000

Censor Rate -0.024135 0.002427 -9.95 0.000

S = 0.851403 R-Sq = 77.1% R-Sq(adj) = 76.9%

**Figure 22:** Minitab Output when Running Linear Regression on the Data Set.

The resulting equation I got was Shape Estimate Full = 1.55 + 0.302\*Shape Estimate Censored – 0.0241\*Censor Rate, where the shape estimate full is the actual parameter value that we are trying to estimate, the shape estimate censored is the Kaplan-Meier Estimate from Minitab, and the censor rate is the percentage of the data that is censored as a percentage from 0 to 100. With an adjusted R2 value of 76.9%, we know this model is doing a fairly good job of predicting the actual parameter value based solely on the Kaplan-Meier Estimate and the Censoring Rate.

I then estimated all of my simulated values using my own estimate, and created a column with their bias and percent bias, as I had previously done with the Kaplan-Meier Estimate. In Figure 23, I created a scatterplot of the percent bias as a function of censoring rate for my own estimate.



**Figure 23:** Scatterplot of Percent Bias as Censoring Rate increases for my Estimate when Minimum Censoring is used.

Based on this scatterplot, there still appeared to be some sort of trend in the percent bias as a function of the censoring rate, but to see what was really going on, I decided to look at some descriptive statistics (Table 24).

Censor

Variable Rate Mean StDev

My PercentBias 5 72.0 103.8

 15 61.6 85.9

 25 47.3 69.4

 35 32.6 54.6

 45 14.90 34.51

 55 3.78 19.52

 65 -4.38 16.06

 75 -18.90 28.52

 85 -12.7 53.3

**Table 24:** Basic Statistics for Percent Bias by Censoring Rate for Minimum Censoring with my Estimator.

While it appeared as though the estimate was still biased, the new estimate appears to have less bias than the general Kaplan-Meier Estimate. The means of the percent bias are closer to zero, while the standard deviations seem comparable. To look further into how well the estimate was working, I picked a single shape and scale parameter, and compared how my estimate worked to how the original Kaplan-Meier Estimate worked. I used all the values in my simulation with a shape parameter of 5 and a scale parameter of 2 and estimated them using both the Kaplan-Meier Estimate and my own estimate. In Figure 25 and Figure 26, I created scatterplots of both of these estimates as a function of censoring rate. Take note of the drastically different scales used in each of these scatterplots; although their shapes are very similar, the true difference is in the scale.



**Figure 25:** Kaplan-Meier Estimator as Censoring Rate Increases (True Value = 5).



**Figure 26:** My Estimator as Censoring Rate Increases (True Value = 5).

Although both of the estimates start by underestimating the parameter value at low levels of censoring and end up overestimating the parameter value at high levels of censoring, the degree to which they do so is very different. In this case, the actual parameter value should have been right around 5. The Kaplan-Meier does very well estimating the actual parameter values when the censoring rate is less than 25%, but when censoring levels get much above that, it begins to grossly overestimate, even estimating parameter values as high as the low 30’s. My estimate, while it starts out underestimating more so than the Kaplan-Meier Estimate, never gets that far off. The highest value it estimates is around nine. In this case, my estimator works best for censoring levels between 35% and 75%.

This is consistent with the results I have found for the whole data set. It appears as though this method of estimation works best for mid-range levels of censoring. It still works for censoring levels both above and below this range, but the estimates aren’t as accurate and tend to be more biased. This method works with sample sizes of approximately 50 participants or greater. Once the number of participants drops below 50, the estimates are not as accurate, especially under high levels of censoring.

**Conclusions and Future Work**

Although this method does not eliminate the bias in the parameter estimate, it does lower the amount of bias and creates estimates that are much closer to the original value than the ones produced by the Kaplan-Meier estimate. Since it appears as though there is somewhat of a pattern in the bias as the censoring rate changes (it underestimates at low levels of censoring and overestimates at high levels of censoring), future work would include trying to rework the model so that the bias is closer to zero. Other future work would include looking at different ways of simulating censoring as to create a method that most closely mirrors the censoring that occurs in real life. Overall though, this method, as it is, does a very good job of improving upon the previous Kaplan-Meier Estimate, especially in the middle range censoring values.

Works Cited

Balakrishnan. *Handbook of Statistics: Advances in Survival Analysis*. Vol. 23. North Holland, 2004. Print.

Fleming, Thomas R., and David P. Harrington. *Counting Processes and Survival Analysis*. New York: Wiley, 1991. Print.

Guo, Shenyang. *Survival Analysis*. Oxford UP, 2010. Print.

Hall, Gaineford J. *Outliers Matter in Survival Analysis*. Rand Corporation, 1982. Print.

Houwelingen, J. C. Van., and Hein Putter. *Dynamic Prediction in Clinical Survival Analysis*. Boca Raton: CRC, 2012. Print.

Ibrahim, Joseph George., Ming-Hui Chen, and Debajyoti Sinha. *Bayesian Survival Analysis*. New York: Springer, 2001. Print.

Klein, John P., and Melvin L. Moeschberger. *Survival Analysis*. 2nd ed. Springer, 2003. Print.

Machin, David, Yin Bun Cheung, and Mahesh KB Parmar. *Survival Analysis: A Practical Approach*. 2nd ed. Wiley, 2006. Print.

Nikulin, M. S. *Parametric and Semiparametric Models with Applications to Reliability, Survival Analysis, and Quality of Life*. Boston: Birkhäuser, 2004. Print.

Randles, Ronald H., and Douglas A. Wolfe. *Introduction to the Theory of Nonparametric Statistics*. Krieger, 1991. Print.

Selvin, S. *Survival Analysis for Epidemiologic and Medical Research: A Practical Guide*. Cambridge: Cambridge UP, 2008. Print.

Wienke, Andreas. *Frailty Models in Survival Analysis*. Boca Raton, FL: CRC, 2011. Print.