3-23-2018

# Parametric Survival Analysis of US Air Force Rated Officer Retention

Jacob R. Lindell


# PARAMETRIC SURVIVAL ANALYSIS OF US AIR FORCE RATED OFFICER RETENTION

THESIS

Jacob R Lindell, 2nd Lt

AFIT-ENS-MS-18-M-136

**DEPARTMENT OF THE AIR FORCE**
**AIR UNIVERSITY**

***AIR FORCE INSTITUTE OF TECHNOLOGY***

**Wright-Patterson Air Force Base, Ohio**

AFIT-ENS-MS-18-M-136

# PARAMETRIC SURVIVAL ANALYSIS OF US AIR FORCE RATED OFFICER RETENTION

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Research

Jacob R Lindell, B.S.

2nd Lt, USAF

22 March 2018

AFIT-ENS-MS-18-M-136

# PARAMETRIC SURVIVAL ANALYSIS OF US AIR FORCE RATED OFFICER RETENTION

THESIS

Jacob R Lindell, B.S.
2nd Lt, USAF

Committee Membership:

PhD Seong-Jong Joo,
Chair

PhD Raymond R. Hill
Member

AFIT-ENS-MS-18-M-136

# Abstract

Personnel retention is a very important topic in both the private and public sectors. Not only do companies need to make sure they have the right people, they need to have the right amount of people. Within the public sector, specifically the US Air Force, maintaining appropriate manning comes in two phases; bringing in the right amount of people each year, and retaining enough people from year to year. These two aspects go hand in hand; if the Air Force knows how many people they will lose in a given year, they can bring in the exact number of people they need to make sure they maintain their end strength requirements. As part of the effort to ensure proper military accessions, the Air Force uses retention models to assist in predicting the future retention patterns. Not only does the Air Force want to make sure they meet their end strength requirements, they want to make sure they bring in the correct amount of people to each career field. The career fields, Air Force Specialty Codes (AFSCs), have a personnel requirement each year in order to accomplish that AFSCs mission. In this study, semiparametric survival analysis was used to determine the significant factors in predicting the rated officer career retention rates. The variables considered were sex, marital status, whether or not an officer had dependents, whether or not an officer was prior enlisted, whether or not an officer graduated as a distinguished graduate, and the institution from where the officer was commissioned. All of these factors were significant for the rated officer career field, which was validated using survival analysis. All of these factors are included in the survival analysis, which took the variables and created a survival curve fit to a specific distribution; the log-logistic. This survival curve was compared to previously

done survival analysis to determine the best decision for use in manpower retention prediction for the Unted States Air Force.

*To the fam. Thanks. #squadup*

# Acknowledgements

I would like to express my gratitude to my AFIT reearch advisors, Dr. Seong-Jong Joo and Dr Raymond Hill. They helped me get to where I needed to be. Thank you to AFIT professor Major Freels, PhD, who was a huge help in R package and allowed me to ask for his help very often.

# Table of Contents


| | | Page |
|---|---|---|
| | **Abstract** | iv |
| | **List of Tables** | x |
| | **List of Figures** | xi |
| **I** | **Introduction** | 1 |
| 1.1 | Problem Background | 1 |
| 1.2 | Research Scope | 3 |
| 1.3 | Issues, Needs, and Limitations | 4 |
| 1.4 | Thesis Outline | 4 |
| **II** | **Literature Review** | 5 |
| 2.1 | Introduction | 5 |
| 2.2 | Modeling Techniques | 5 |
| 2.3 | Methodologies | 9 |
| | Model Comparison Techniques | 12 |
| 2.4 | Antecedent Studies | 14 |
| **III** | **Methodology** | 17 |
| 3.1 | Introduction | 17 |
| 3.2 | Data Preparation | 17 |
| 3.3 | Career Field Analysis | 19 |
| | ABM CYOS | 20 |
| | CSO CYOS | 21 |
| | Pilot CYOS | 21 |
| 3.4 | Survival Analysis | 22 |
| | Introduction | 22 |
| | Nonparametric Survival Analysis | 23 |
| | Semiparametric Survival Analysis | 29 |
| | Parametric Survival Analysis | 31 |
| **IV** | **Results and Analysis** | 40 |
| 4.1 | Introduction | 40 |

4.2 Results . . . 40
Distribution Selection . . . 40
Parametric vs Semiparametric . . . 44
YOS as independent variable . . . 45
4.3 Analysis . . . 46
Residuals . . . 46
**V Conclusion** . . . 51
5.1 Limitations of Work . . . 51
5.2 Follow-On Research . . . 51
Appendix A . . . 53
5.3 Data Example . . . 53
5.4 Marital Status . . . 54
5.5 Prior Service . . . 55
Appendix B . . . 57
**Bibliography** . . . 71

# List of Tables


| Table | | Page |
|---|---|---|
| 1 | Wald's p-values from each CYOS and Career Field | 19 |
| 2 | ABM p-values | 20 |
| 3 | CSO p-values | 21 |
| 4 | Pilot p-values | 22 |
| 5 | Parametric Model Parameters | 34 |
| 6 | PH Model Parameters | 35 |
| 7 | AIC for ABM | 41 |
| 8 | AIC for CSO | 41 |
| 9 | AIC for Pilot | 42 |
| 10 | AIC for ABM: Parametric vs Semiparametric | 44 |
| 11 | AIC for CSO: Parametric vs Semiparametric | 44 |
| 12 | AIC for Pilot: Parametric vs Semiparametric | 44 |
| 13 | AIC for ABM, CSO, and Pilot with YOS | 45 |
| 14 | AIC with YOS comparisons | 46 |
| 15 | Sum of Squared Error for Parametric model | 47 |
| 16 | Sum of Squared Error for Parametric model | 49 |
| 17 | Sum of Squared Error for Parametric model | 50 |

# List of Figures


| Figure | | Page |
|---|---|---|
| 1 | Survival Analysis Sustainment Curve | 14 |
| 2 | Kaplain-Meier plots (Sex) | 24 |
| 3 | Kaplain-Meier plots (Marital Status) | 25 |
| 4 | Kaplain-Meier plots (Dependents) | 26 |
| 5 | Kaplain-Meier plots (Commissioning Source) | 27 |
| 6 | Kaplain-Meier plots (Distinguished Graduate) | 28 |
| 7 | Kaplain-Meier plots (Prior Service) | 29 |
| 8 | Cox PH Plots | 30 |
| 9 | Parametric plot for ABM | 32 |
| 10 | Parametric plot for CSO | 32 |
| 11 | Parametric plot for Pilot | 33 |
| 12 | Parametric plot for ABM with YOS | 37 |
| 13 | Parametric plot for CSO with YOS | 38 |
| 14 | Parametric plot for Pilot with YOS | 38 |
| 15 | Parametric Distributions for ABM | 43 |
| 16 | Residual plot for Parametric model (ABM) | 47 |
| 17 | Residual plot for Parametric model (CSO) | 48 |
| 18 | Residual plot for Parametric model (Pilot) | 49 |
| 19 | Kaplain-Meier plots (Marital Status) | 54 |
| 20 | Kaplain-Meier plots (Prior Service) | 55 |
| 21 | Kaplain-Meier plots (Prior Service) | 56 |

# PARAMETRIC SURVIVAL ANALYSIS OF US AIR FORCE RATED OFFICER RETENTION

# I. Introduction

## 1.1 Problem Background

1973 was the first year the United States no longer used conscription to constitute the military. The All Volunteer Force (AVF) was established, and since then the United States military has remained an all volunteer force [1]. The All Volunteer Force brought along the challenge of bringing in an adequate number of people into the military, and keeping those individuals in for as long as desired.

The United States Air Force receives an end strength cap on the size of their force each fiscal year by Congress. The end strength manpower is calculated by how many members are in the Air Force at the beginning of the year, adding the number of personnel acquired during that fiscal year, and subtracting the number of personnel lost during the fiscal year. These numbers drive the recruiting of new members brought in during each fiscal year. If the Air Force has less individuals than the end strength requirement, they can bring in more members to fill their needs. However, if the Air Force has more individuals than the end strength requirement, those extra members may need to be released to meet the end strength requirement. This inconsistency of maintaining the desired manpower end strength affects all career fields within the Air Force.

The Air Force is split into two active duty components: enlisted and commissioned

officers. The difference between these two components is that commissioned officers must attend college before entering the Air Force, and then attend a commissioning source such as the USAFA, Officer Training School (OTS) or civilian university Reserve Officers' Training Corps (ROTC). Within these active duty components are career fields, split up by the type of mission each supports, known as Air Force Specialty Codes (AFSCs). The officer career fields are split into rated, non-rated line, and non-rated non-line AFSCs.

The rated AFSCs main mission is flying, while the non-rated officers' functional capacity is to support the flying mission. The rated officers are split into four career fields; Pilot, Combat Systems Officer (CSO), Air Battle Manager (ABM), and Remotely Piloted Aircrafts (RPA). These rated career fields require extensive initial training to equip personnel to perform the required task of their career field. Subsequently there are continuous training upgrades in order to remain qualified as a rated officer. If a rated officer does not remain qualified, they lose their aeronautical rating, are no longer rated, and no longer equipped to do their job.

The training required to become a rated officer is expensive and time consuming, but varies by career field. Pilots have a year of initial pilot training, after which their airframe is decided. Pilots then train for their specific airframe, which requires 4 to 8 months. Training a pilot is the most expensive training for any career field. According to the United States General Accounting Office (GAO), in 1999 the cost to send one pilot through basic pilot training was $1 million, and the cost to fully train a pilot was $9 million. As with any officer's training, pilots incur a payback, otherwise known as an Active Duty Service Commitment (ADSC), upon completion of any training. Once a pilot's (ADSC) is served, they can voluntarily separate from the Air Force. Not only does the Air Force lose the millions of dollars invested in training that rated officer, they now have to train another officer to fill the void created [2].

An ADSC is a contractual obligation for an individual to serve the Air Force for a specified length of time. The first ADSC acquired, the accessions ADSC, is four years, except for

officers coming from the United States Air Force Academy, which is five years. The ADSC changes for rated officers once they have completed their intial rated officer training; ten years for pilots, six years for CSOs and RPAs, and three years for ABMs.

Once the initial ADSC has expired, an officer can voluntarily separate from the Air Force. In a study by Buyer and Abercrombie [2], the Air Force projected having a pilot shortage of between 1,900 and 2,155 pilots. Some of the reasons for this shortage were reduced accessions, training delays, or increased desire for rated officers outside of the Air Force. The reduction of force in the mid 1990s caused the Air Force to reduce their pilot accessions, leading to less pilots being trained and ready to fly by the end of the 1990s. This constant flux is part of the reason the Air Force has trouble meeting their pilot manpower requirements, but the main reason is that pilots separate from the Air Force before their replacement has been trained [2].

Ideally the Air Force would bring in and train the exact number of rated officers they need and none of those officers would separate. However, this is not feasible because of the outside factors that continuously come into play causing pilots, and non pilots, to leave the Air Force. The voluntary separation of rated officers becomes very time and cost sensitive, leading the Air Force to offer multiple bonuses throughout a rated officer's career in hopes of retaining more officers. To combat the voluntary separations and rated officer shortages, this research will identify those important factors that lead to rated officers leaving the Air Force, and create a model to help predict the future retention of rated officers.

## 1.2 Research Scope

The officer corps, although it only makes up 20% of the total force, accounts for 100% of the pilots, CSOs, and ABMs in the Air Force. Due to all of the training needed for rated officers, there is a lot of time and money invested into the rated officer corps. Retaining these officers is a priority for the Air Force because of the time and money it takes to train

them. This research focuses on modeling ABMs', CSOs', and Pilots' behavior over 40 years. Specifically, this research examines a parametric approach to survival analysis in hopes of improving the Air Force's rated officer retention predictions.

## 1.3 Issues, Needs, and Limitations

This research utilized rated officer data from 2006-2015. The data comes from the Headquarters Air Force Directorate of Personnel officer (HAF/A1PF), and was extracted from the Military Personnel Delivery System (MilPDS). Over this time period, career fields have combined while others split; therefore, HAF/A1 had to change old data and fill in missing data points.

MilPDS is prone to contain errors within the data, which can be due to glitches in the system, input errors, or just missing data. The assumptions made by HAF/A1 when creating and adapting the data have been inherited into this research.

## 1.4 Thesis Outline

Chapter 2 presents previous, relevant data to the topic of survival analysis and manpower retention. Chapter 3 is the methodology section that describes survival analysis and finding the best model for retention. Chapter 4 is the results and analysis section that compares the parametric and semiparametric models and examines if the parametric model is a good fit to the data. Chapter 5 is the conclusion and provides summary remarks regarding the research.

# II. Literature Review

## 2.1 Introduction

This chapter reviews previous relevant work that uses survival analysis based on parametric models for manpower retention approximations.

The challenge with the military and their manpower problem is that it cannot be solved from the top down; if the Air Force needs a Lieutenant Colonel, they cannot go outside of the organization and find an individual that could fill this void. Every person in the Air Force started from the bottom rank level of the Air Force. Once an individual no longer wants to be in the Air Force, as long as their Active Duty Service Commitment has expired, they are allowed to leave the Air Force and become a civilian.

Manpower is essential for the mission of protecting the country from all enemies foreign and domestic, so it is necessary to have the correct amount of people each year to fill the manpower requirements so the Air Force may operate as efficiently and effectively as possible. Predicting the amount of people that may leave the Air Force in a given year allows the Air Force to prepare for these separations years in advance; allowing them to have the amount of officers at each rank required each year.

## 2.2 Modeling Techniques

Within manpower prediction, methodologies range from regression to survival analysis. Regression analysis is a commonly used technique in military personnel management because it is simple, yet effective. "Regression analysis is a statistical technique for investigating and modeling the relationship between variables" [3]. The most commonly used type of

regression, linear regression, models the linear relationship between the response variable ($y$) and the regressor variable ($x$), or variables. The linear relationship is determined by fitting a "line of best fit". This line of best fit is fit with the purpose of being able to use the different regressor variable inputs, usually values that have not yet been evaluated but are within the data, and best predict the response variable outcome. The commonly used least squares method's goal in the line of best fit is to minimize the Sum of Squared Error for the residuals; the difference between the actual value and the predicted value for all regressor values. A measure of accuracy of this line is portrayed in the $R^2$ value, which ranges from 0 to 1, and is used to show how much of the variance in the response variable can be explained by the regressor variable(s). Once this regression equation is established, new regressor variables can be used to predict the associated response variable.

In some circumstances, such as with binary response data, linear regression analysis may not be useful. This is where logistic regression comes in. "Logistic regression is a mathematical modeling approach that can be used to describe the relationship of several $x$'s to a dichotomous dependent variable" [4]. The logistic model and linear model are similar in theory, but the difference between the two is linear regression has a continuous dependent variable ($y$) while logistic regression is used more commonly with categorical variables; values that are integer in nature, and has a non-continuous dependent variable.

The strength of both linear and logistic regression analysis is interpolation; determining a response based on regressor variables that are within the data range modeled. The weakness of linear and logistic regression analysis is the counterpart, extrapolation; determining a response variable based on regressor variables that have not already occurred. Extrapolation is estimating a value based on data outside the range used to create the model. When extrapolation is the goal of the research, using a method other than regression analysis may be more useful to give a viable result [3].

Simulation is another modeling approach, but is more commonly used with systems or processes. "A simulation is the imitation of the operation of a real-world process or system

over time" [5]. Simulation is mainly used for complicated systems because those systems are commonly hard to predict using a clean, analytical method. Analysts use varying software products like *Microsoft Excel* [6], or more simulation specific like *Arena* [7] or *SIMIO* [8]. These analytic software programs are used to simulate real-world scenarios and calculate different statistics that give information needed regarding the real-world scenario. "Once developed and validated, a model can be used to investigate a wide variety of 'what if' questions about the real-world system" [9]. Although a simulation does not always yield a 100% solution, it is one of the better ways to gain insight into a real-world problem.

Within the broad area of simulation is the agent-based simulation, which is where "a system is modeled as a collection of autonomous decision-making entities called agents". The actions of and the interactions between the agents are considered to determine the effects they have on the system as a whole; "the whole is more than the sum of its parts because of the interactions between the parts" [10]. The agents are given a set of rules and the model is used to see how those rules shape the entities individually, and then the entire system as a whole. Like simulation, agent-based modeling can model real-world scenarios, but unlike simulation, agent-based simulation is better for modeling actions of and interactions between autonomous agents. Agent-based simulation is better at modeling unexpected behavior of the autonomous agents, while simulation is best with predictable interactions of agents.

Survival analysis techniques have been used recently to better analytically predict manpower attrition in both the public and private sectors. Survival analysis is "a loosely defined statistical term that encompasses a variety of statistical techniques for analyzing positive-valued random variables. Typically, the value of the random variable is the time to failure of a physical component... or the time to the death of a biological unit" [@]. In survival analysis "there are three basic elements which must be well defined: a time origin, a scale for measuring time, and an event" [11]. The time origin is most commonly the beginning of a study (time = 0). The scale for measuring time varies with the event being measured. Examples of events are birth, death, attracting a disease or anything that has a binary out-

come; 0 meaning the event has not yet occurred, 1 meaning the event has occurred. The survival function gives the probability of whatever event is in question, in our case manpower, surviving or retaining up to that specified time [12].

One of the advantages survival analysis has over linear and logistic regression is how it deals with censored data. According to Miller (2011), there are three types of data censorship: type 1 censoring, type 2 censoring, and random censoring. Type 1 censoring occurs when the trial ends at a predetermined time, so only the occurrences before the time expires are being recorded. Type 2 censoring is when the trial ends when a specified fraction or when a number of events occur. Random censoring is best explained in an example. Miller uses a medical study: "In a clinical trial, patients may enter the study at different times; then each is treated with one of several possible therapies. We want to observe their lifetimes, but censoring occurs in one of the following forms": loss to follow up; we never see the patient again, drop out; patient refuses to continue treatment but we are still in contact, and termination of study; study ends [13].

Survival analysis is useful in aspects where regression fails. For example, when regression techniques see some data as an outlier, and remove that data, survival analysis includes that data.

Survival analysis can be broken into three models; nonparametric, semiparametric, and parametric models. The three types of models differ in flexibility; nonparametric being the most flexible, and parametric being the least flexible. Understanding the differences between the models allows for accurate model selection to best represent the survival data.

In nonparametric models "there is no assumptions about the shape of the hazard function or about how covariates may affect that shape" [13]. Nonparametric models are generally the starting point for survival analysis because they allow conducting survival analysis without meeting any assumptions. These nonparametric models cannot handle more than one covariate, so models with multiple covariates are usually handled using semiparametric or parametric models.

Semiparametric models make no assumptions about the shape of the hazard function, but they do "make a strong assumption about how the covariates affect the shape of the hazard function between groups over time" [13]. The hazard function is the ratio of the probability density function and the survival function. The lack of a distributional assumption gives the semiparametric model more flexibility and robustness, so they generally fit the data well.

The parametric function differs in that the outcomes follow a certain assumed distribution and the relationship between the covariates and survival also have assumptions. These assumptions determine whether or not a proportional hazards or an accelerated failure time model is chosen. In the proportional hazards models, the covariates have a multiplicative effect on the hazard function. In accelerated failure time models, the survival time is interpreted to either accelerated, time scale factor greater than 1, or decelerate, time scale factor less than 1 [14]. Parametric functions are useful because once a distribution is chosen, the models can be completed without any sample data.

## 2.3 Methodologies

A major concern for colleges and universities is the rate at which students drop out. Educators want to keep students in their programs until completion [15]. Min et al. set out to determine the significant factors that influence as to whether or not a student is likely to drop from the engineering majors of their respective undergraduate universities. The research found that the significant factors differ between semesters; white and female students tend to leave engineering schools earlier than average, students with lower SAT math scores tend to leave during the second semester, and a student's SAT math score better predicts the likelihood of them dropping out than their SAT verbal score. The universities and colleges can now use this information to better understand their student population and see if they want to change the criteria of acceptance to an engineering school to decrease the student drop-out rate.

There have been a lot of studies about employee turnover rate in the civilian sector. Cotton and Tuttle [16] used Ordinary Least Squares and piecewise regression models to determine that almost all of the variables they studied; external, work-related, and personal factors, were significant and related to turnover. Some of these variables were: unemployment rate, pay, job performance, age, gender, and education. Most of these factors apply to the military, but unlike in the civilian world, the military cannot change some factors such as: pay, satisfaction with pay, tenure, etc.

The military promotes from within and cannot change individual wages based on performance or lack thereof. Thus, some of the factors exhibited in [16] do not pertain to military personnel turnover. Studying the factors in the civilian sector can help the military get a better understanding of their own retention rates, especially when the military's reenlistment rates change based on some of the factors within the civilian sector.

Gass [17] discusses the various types of military manpower models used in the 1980s. He stated that the problem manpower modeling addresses is to "determine the number of personnel and their skills that best meets the future operational requirements of an enterprise". Gass starts with transition rate models, also known as Markov chains, to answer the question: "what is the composition of the force at the end of a planning period?" Markov chains can be used to estimate new hires, separations, retirements, etc. This is done by using different factors deemed relevant to manpower progression. These factors are: time hired, skills, function, and job title. Gass notes that with large personnel systems, such as the military, Markov models may be difficult to use because the steady-state forecasts they generate may not be accurate enough due to the transition rates changing over time [17].

Even with this difficulty, Gass [17] notes how the US Army used Markov models to project the flow of the US Army enlisted force. Since the Army was seeing a decrease in recruits, they found it necessary to increase reenlistments to keep the necessary manpower to meet requirements. This increase in reenlistments eventually led to an "older" force, which in turn prompted a decrease in the promotion rate to counteract the earlier decision to increase

recruitment.

Gass [17] also discusses how network-flow problems were adapted from their usual use in scheduling to personnel flows. In a personnel flow approach, the source nodes represent initial personnel inventories, the intermediate nodes are used to meet grade and skill goals, and the sink nodes were final personnel inventories.

Although there are difficulties modeling personnel using the network-flow models, Gass [17] discusses a model he used for officer personnel flow. When using a network-flow approach, the problem begins to grow rapidly when the number of grades, commands, and skills increase, but these problems are still solvable.

In the 1980s, the military used the Accession Supply Costing and Requirements Model (ASCAR) to reach a given personnel requirement and to optimize the mix of new recruits. ASCAR determines the annual accessions needed to meet the service's end strength, man year, quality requirements, and the cost of the force over a 15-year time horizon, ASCAR uses five steps; analyze historical data, simulate one-year losses, evaluate the new recruits based on demographic and qualitative statistics, apply goal programming to meet end strength requirements, and assess cost factors for alternative manpower policies. ASCAR uses goal programming to determine these factors, but the problem with goal programming is that it is not pareto efficient, therefore the solution cannot be completely trusted [18].

The US Navy used a statistical approach to their manpower projection models to find the right number and types of people for their workforce[19]. Historical performance data were analyzed to identify correlations between the model parameters and the retention outcomes which were used to predict future outcomes. Cashbaugh [19] used agent-based simulation to improve Navy personnel forecasting to model social, economic, quality of life, and incentives factors.

Capon and Chernyshenko [20] studied applying civilian retention theory to the New Zealand military. They note that up until their study, a majority of the military retention studies rely too heavily on demographic and organizational characteristics; examples of demo-

graphic are gender or race, while examples of organizational are male/female ratio and length of overseas assignments. They degrade this ideology because of three reasons: demographic characteristics cannot be changed, recruiting based on these demographics would further diminish the force, and focusing on demographics neglects the actual cause of turnover. They touch on the fact that retention and turnover commonly go hand in hand, but that does not account for the voluntary or involuntary turnover. Capon and Chernyshenko [20] concluded that civilian retention models can be useful when looking at military retention.

Orrick [21] addressed a Marine Corps problem; they needed to increase their end strength manpower by 20,000 over a three-year time period. Orrick focused on the attrition aspect of manpower forecasting; he needed to determine how many marines were going to leave, and when, in order for the Marine Corps to bring in enough individuals to meet their end strength requirements.

Orrick [21] used logistic regression models to forecast manpower losses for three fiscal years; 2005, 2006, and 2007. The primary aspect of his work was comparing End of Active Service losses to non-End of Active Service losses. He used Receiver Operating Characteristic (ROC) curves to assess the performance of his models, which showed than his logistic models performed well. Orrick did not look at filling the job requirements each year with certain ranks or grades, but rather having enough people at the end of year 3 to fulfill the 20,000 troop increase, which led to his future research recommendation.

Orrick [21] suggested survival analysis since it "has proven to be a very useful tool in its predictions based on attributes of a representative sample of the entire population" but the data would not support this kind of analysis [21].

### *2.3.1 Model Comparison Techniques*

With all of the different models available for use, choosing a best model is important. One of the most common ways of doing this is using an $R^2$ value. The $R^2$ value, which ranges from 0 to 1, is used to show how much of the variance in the response variable can

be explained by the regressor variable(s). An $R^2$ value of 0 means 0% of the variance in the response variable can be explained by the regressor variable(s), while an $R^2$ value of 1 signifies all of the variance in the response variable is explained by the regressor variables. The closer to 1 the $R^2$ gets, the better the model fit. This method is commonly used in linear and logistic regression because the user is attempting to fit a line through data points in order to interpolate data points that have not yet occurred.

The goodness-of-fit test is similar to that of the $R^2$ value since it portrays how well the data are fit by a given model. The goodness-of-fit test uses hypothesis testing. Hypothesis testing is a method of statistical inference used to make a decision in the face of uncertainty. Hypothesis testing involves a null hypothesis and an alternative hypothesis, with the purpose of proving the null hypothesis to be false; to "reject the null". Under the null hypothesis, a statistical test is formed, in many goodness-of-fit cases, a chi-squared distribution-based test. The hypothesis is tested against the chi-squared distribution to see if data observed is as hypothesized or unlikely to be as hypothesized.

The goodness-of-fit can then be extended with the likelihood ratio test; the goodness-of-fit of two models, one of which is the special case of the other. The likelihood ratio test is based on the likelihood ratio, which shows how much more likely data is to be under one model rather than the other model. The likelihood ratio is then used to either compute a p-value or to be compared against a test statistic to see if the null hypothesis can be rejected or not.

A model comparison tool used commonly in survival analysis is the Akaike Information Criterion (AIC). "The AIC is used to test whether you have the appropriate model fit between the competing non-nested models" [14]. A low number for an AIC is better, and the equation to calculate the AIC statistic is:

$$AIC = -2 \log L + 2(c + p + 1) \tag{1}$$

where $L$ is the log-likelihood, $c$ is the number of covariates in the model and $p$ is the

number of parameters in the model [14]. The AIC is used to compare survival models because it takes into account the log-likelihood and number of model parameters, and it penalizes models for lack of parsimony, or too many model parameters. The AIC can be used to compare nonparametric, semiparametric, or parametric survival models, or any combination of the three against each other.

## 2.4 Antecedent Studies

The Air Force produces sustainment lines based on the 5-year historical retention rate. Based on the year, the Air Force could either bring in too many people or not enough people. If the Air Force brings in too few people one year, they want to bring in more than the expected number of people the following year in order to balance and approach the average number of people needed. Recently, manpower predictions have been examined by using survival analysis techniques. Figure **??** shows a sustainment line from Zimmerman's [22] work on enlisted force retention.

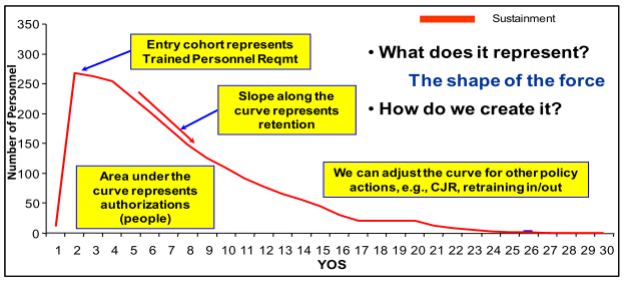


**Figure 1.** Survival Analysis Sustainment Curve

This work is a follow-on of various previous explorations into Air Force manpower predictions using Survival Analysis. Schofield [23] and Zens [24] explored attrition behavior for the non-rated officer career fields, Franzen [25] explored the rated officer career fields, and Zimmerman [22] explored the enlisted career fields.

Schofield [23], used both logistic regression and survival analysis to produce an updated model for use as the sustainment line process for Air Force manpower predictions. Schofield

filtered the data set to only include non-rated officers within four generic career fields: Acquisitions, Logistics, Support, and Non-Rated Officers. Schofield used logistic regression to determine the significant factors in predicting the non-rated Air Force line officer retention. Schofield found that commissioning year, gender, commissioning source, number of years served as a prior enlisted member, career field, and distinguished graduate status were significant, with a few exceptions within different year brackets.

Schofield used these six significant factors in survival analysis in order to create the sustainment lines for each of the selected career fields. Schofield found 99 survival functions were necessary in explaining the selected non-rated career fields, with each survival function being based on different combinations of variable settings. Schofield found that the methods used were as effective as the current model utilized by HAF/A1 [23].

Zens [24] used Schofield's work to forecast future personnel levels within the same four career fields in order to predict "the number of personnel who will remain in each of these career fields over the next 30 years". Zens then measured the stability of those career fields based on the mean and standard deviations for the coefficients of variation with the hopes of decreasing personnel costs and enhancing understanding of officer behavioral patterns. The survival curves enabled predicting retention rates for the 30 year time period.

Franzen [25] also used logistic regression to identify the significant factors contributing to the attrition of rated officers and then applying survival analysis to build a model capable of predicting attrition rates for rated officers. Franzen used demographic, economic, and political data within her significant factors analysis because of how these three factor groups can all influence the rated officer attrition.

Franzen used the same six explanatory variables as Zens and Schofield and analyzed these variables within each career field. Franzen started with nonparametric survival analysis; using Kaplan-Meier plots to show how manpower survival changes over time, but then applied semiparametric analysis to this same data. Using Cox Proportional Hazards regression, Franzen used economic and political factors in a Cox Proportional Hazard model to determine

that there are six demographic factors; gender, number of years served as an enlisted member, career field selection, commissioning year, commissioning source, and distinguished graduate status, and one economic factor; male-to-female ratio, that are statistically significant in modeling the retention behavior or rated officers. Franzen found that this methodology yielded similar results to that of HAF/A1 [25].

Zimmerman [22] used survival analysis to examine enlisted force retention. Zimmerman also used logistic regression to find the significant factors in predicting retention rates, but found some significant factors that differed from the officer data analyzed by Zens, Schofield, and Franzen. Zimmerman found that grade, gender, race, marital status, number of dependents, and years of service were the significant factors in predicting retention. Grade and years of service for enlisted personnel are similar to the commissioning year and number of years served as a prior enlisted member for officers, but the main difference is that while officers have almost automatic promotion rates up to Captain, the enlisted force only automate promotes to E4. The enlisted data used was examined at an aggregate level, separated by year, AFSC, and Selective Reenlistment Bonus for each AFSC. The four AFSCs Zimmerman examined were: Airfield Management (1C7XI), Operations Intelligence (1N0X1), Survival, Evasion, Resistance and Escape (1T0X1), and Mental Health Services (4C0X1). Zimmerman used Kaplan-Meier estimation to compare the survival analysis techniques used in her work to the current sustainment lines used by HAF/A1. These Kaplan-Meier plots were then broke out into gender and marital status to analyze the survival of manpower based on different significant factors.

Schofield, Zens, Franzen, and Zimmerman all came to the conclusion that the Air Force's current personnel retention models are as accurate, asthe models they created. Their recommendations were to continue to use the current retention models.

# III. Methodology

## 3.1 Introduction

Nonparametric survival analysis is used to examine trends and differences between the chosen variables. Semiparametric survival analysis is used to identify significant factors within the different career fields and for comparison to parametric survival models. Parametric survival analysis is then used to create models for predicting retention given these significant variables.

## 3.2 Data Preparation

The data used for the survival analysis is a compilation of all monthly extracts provided by HAF/A1 from January 2006 until December 2015. This data was previously modified by Franzen [25] in her thesis work, and providesa breakdown of each career field chosen and into different Current Years of Service (CYOS) groups. The data was given as *SAS* files and was read into *R* for the analysis.The CYOS groups are: 0-6, 4-8, 8-14, 12-19, 20-22, with each career field containing these CYOS groups.

Unfortunately, the personnel records given are prone to censoring because some response values are not observed within the time frame of the records provided, or some values of the variables may not fall in the specified range. This leads to some incomplete records, which led to the use of overlapping CYOS'. The officer's records are only included in a given CYOS group if their career spans the entirety of the group; if an officer leaves the Air Force at 10 years, their records are in the 0-6 CYOS, the 4-8 CYOS, but not the 8-14 CYOS. This overlapping prevents data truncation. The data was given with an individual for each line

(data point), with all variables describing that individual while within the CYOS.

Survival analysis needs a *status* variable which indicates whether or not an officer has separated. Survival analysis within $R$ requires both a time component and a *status* variable, with 0 indicating an individual remaining in the Air Force and 1 indicating the individual leaving the Air Force. To add a *status* variable, each member was expanded out for every year until their YOS (Years of Service). For example, if an individual stayed in the Air Force for 12 years (YOS = 12), twelve data-points are created, with the only variable changing being the *status*; which equals 0 for the first 11 years, and equals 1 once the individual's data has been replicated 12 times.

Previous work done in United States Air Force personnel retention analyzed the data and were able to extract six variables necessary to best predict the model. This work was done by Schofield [23] by a logistic regression model in *SAS* and continued in this work. The covariates included in the model are Sex (M = Male, F = Female), Marital Status (0 = Single, 1 = Married, 2 = Previously married but no longer), Dependents (0 = No dependents, 1 = Dependents), Commissioning Source (1 = Other, 2 = United States Air Force Academy (USAFA), 3 = Reserve Officer Training Corp (ROTC), 4 = Air Force Officer Training School (OTS)), Distinguished Graduate (0 = Not a Distinguished Graduate, 1 = Distinguished Graduate), and Prior Service (0 = No prior service, 1 = Prior service).

These variables were used by Schofield [23] and Franzen [25]. This current effort reviewed the data to determine if there were any variables that may give a better answer for the survival models. Upon review of the data, four variables were chosen and further analyzed to determine if any of them would help the survival analysis: *race*, *citizenship*, *religious denomination*, and *place of birth (country/state)*. Race was previously used by Zimmerman [22] for the enlisted force. These variables were chosen because they were categorical in nature and seemed to be potentially be useful. None of the variables tested were significant, so the variables used for the analysis were those given by Schofield [23] and Franzen [25].

## 3.3 Career Field Analysis

Wald's p-values were used to determine the variables' significance. Previously the importance of the variables was established using linear regression by Schofield [23]. A proportional hazards (PH) model allows for a distributional input to determine the significance of variables. This is done to reduce the difference between linear regression and survival analysis, which in turn should allow for the best representation of the significant variables. The variables *dependents (DEPENDS)*, *distinguished graduate (DG)*, and *prior service (PRIORSVC)* are binary variables (0 or 1) and *sex*, *marital status (MARITALSTAT)*, along with *Commissioning Source (SOC)* are categorical variables.

Before using the specific CYOS's within the career fields, the broad career fields are evaluated in order to determine if the variables are significant. Each career field was analyzed, and Wald's p-values were extracted as shown in Table 1.

**Table 1.** Wald's p-values from each CYOS and Career Field

| Variables | ABM | CSO | Pilot |
|---|---|---|---|
| Sex | 0.003 | <0.001 | <0.001 |
| Marital Status | <0.001 | <0.001 | <0.001 |
| Dependents | <0.001 | <0.001 | <0.001 |
| Source of Commissioning | <0.001 | <0.001 | 0.5 |
| Distinguished Graduate | <0.001 | <0.001 | <0.001 |
| Prior Service | <0.001 | <0.001 | <0.001 |
| Individuals | 2543 | 7300 | 23863 |
| Observations | 32968 | 110043 | 327130 |

The initial analysis shows that commissioning source is insignificant for pilots. This is a red flag because we do not want any of our variables chosen to be insignificant at any point in time; we want the variables that best explain the data for the entirety of the data. Because a p-value less than 0.05 shows significance, a p-value of 0.5 means that the variable is insignificant for the survival analysis focused on pilots. This can be explained, however.

Within the Air Force it is hard to get a pilot slot, and it is even harder for individuals to receive a pilot slot that are not from either USAFA or ROTC. Within the data, 88% of the pilots are either from USAFA or ROTC, so there may not be enough data for OTS and other commissioning sources to balance out the effects of commissioning source on survival. Further, when breaking down the data into the CYOS', *commissioning source* is only found insignificant once, as shown in Table 4.

### *3.3.1 ABM CYOS*

For the analysis of the significant variables within the different CYOS subsets, Wald chi-squared p-values are used. The Wald's p-value is used to test the significance of variables within a data set and helps examine how those career fields differ over the CYOS subsets. This helps see if some variables are more important than others at certain times in an officer's career, or not at all. Table 2 shows the p-values for the different CYOS' within the ABM career field.

**Table 2.** ABM p-values

| CYOS | Sex | Marital Status | Dependents | Commissioning Source | Distinguished Graduate | Prior Service | Observations |
|---|---|---|---|---|---|---|---|
| 0 to 6 | 0.001 | <0.001 | <0.001 | 0.061 | <0.001 | <0.001 | 2084 |
| 4 to 8 | <0.001 | <0.001 | <0.001 | 0.505 | <0.001 | <0.001 | 1837 |
| 8 to 14 | 0.012 | 0.001 | <0.001 | 0.001 | <0.001 | <0.001 | 1159 |
| 12 to 19 | 0.808 | 0.626 | <0.001 | 0.005 | 0.367 | <0.001 | 683 |
| 20 to 22 | 0.633 | 0.798 | 0.163 | 0.052 | 0.545 | <0.001 | 498 |

*Prior Service* is significant for every CYOS subset, while *Dependents* is significant for four of the five CYOS subsets. As the CYOS' increase in years, the less variables are significant. This could be due to ABM being a small data set, only 2,551 officers observed in the given data, in comparison to the other career fields being analyzed. It can be difficult to distinguish a correlation between the covariates and the response variable with a smaller

data set. Further analysis can determine the accurate significance of these variables.

### *3.3.2 CSO CYOS*

CSO is a larger career field, so it is easier to accurately determine the significance of the covariates. The p-values for CSO are shown in Table 3.

**Table 3.** CSO p-values

| CYOS | Sex | Marital Status | Dependents | Commissioning Source | Distinguished Graduate | Prior Service | Observations |
|---|---|---|---|---|---|---|---|
| 0 to 6 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | 6462 |
| 4 to 8 | <0.001 | <0.001 | <0.001 | 0.001 | <0.001 | <0.001 | 5945 |
| 8 to 14 | <0.001 | 0.008 | <0.001 | 0.319 | <0.001 | <0.001 | 4293 |
| 12 to 19 | <0.001 | 0.199 | 0.08 | <0.001 | <0.001 | <0.001 | 3232 |
| 20 to 22 | 0.29 | 0.8 | 0.963 | 0.001 | 0.002 | <0.001 | 2587 |

Once again, *prior service* is significant for all CYOS subsets, along with *distinguished graduate*. *Marital status* and *dependents* start out as significant, for CYOS subsets 0-6, 4-8, and 8-14, but become insignificant near the end of an officer's careers. This could be due to the needed stability of a career at the beginning of a marriage or family, but as the officer ages, that stability may be less important. *Sex* becomes insignificant in the last CYOS subset, which could be due to the fact that good officers are good officers, and their gender does not matter. *Commissioning Source* is only insignificant in the middle of an officer's career.

### *3.3.3 Pilot CYOS*

The pilot subset is the largest subset, which allows for the most accurate p-values. The Wald's p-values are shown in Table 4.

**Table 4.** Pilot p-values

| CYOS | Sex | Marital Status | Dependents | Commissioning Source | Distinguished Graduate | Prior Service | Observations |
|---|---|---|---|---|---|---|---|
| 0 to 6 | <0.001 | <0.001 | <0.001 | 0.024 | <0.001 | <0.001 | 21548 |
| 4 to 8 | <0.001 | <0.001 | <0.001 | 0.001 | <0.001 | <0.001 | 5945 |
| 8 to 14 | <0.001 | 0.008 | <0.001 | 0.319 | <0.001 | <0.001 | 4293 |
| 12 to 19 | <0.001 | 0.199 | 0.08 | <0.001 | <0.001 | <0.001 | 3232 |
| 20 to 22 | 0.29 | 0.8 | 0.963 | 0.001 | 0.002 | <0.001 | 2587 |

Once again, *prior service* and *distinguished graduate* are significant for all CYOS', *sex* is only insignificant at the end of a career, *marital status* and *dependents* become insignificant near the end of a career, and *Commissioning Source* is insignificant in the middle of a career.

Pilot and CSO have the same significant variables, so it seems that more observations within a data set lead to the "true" solution. ABM may not have enough data points to give us the exact variable significance, but we can assume that as the number of data points increases, the p-value table would converge to look like that of CSO and Pilot. In comparison to CSO and Pilot, ABM seems to be under fitting the data; variables that should be significant are shown as insignificant. This is probably because of the relatively small size of the ABM data set in comparison to CSO and Pilot, so all variables are going to continue to be used for the ABM data set with hopes of correctly fitting the data if it were larger.

## 3.4 Survival Analysis

### *3.4.1 Introduction*

The data given has 628 variables for each individual within the 2006-2015 data. Six key variables are chosen, and the *status* variable was created. Survival analysis is used to estimate survival function curves, which in turn yields insight into the significant factors for officer retention.

There are three types of survival analysis. Nonparametric survival analysis is the baseline, and does not make any predictions about how the covariates affect the shape, nor any prediction of a distributional fit within the data. Nonparametric survival analysis is most commonly done using life tables and Kaplan-Meier plots. Semiparametric analysis makes an assumption about how the covariates affect the survival function, but does not make any assumptions on a distributional fit within the data. Semiparametric analysis most commonly uses Cox Proportional hazards regression for the analysis. These forms of analysis have both been done previously for Air Force officer retention, most recently by Franzen [25]. Parametric survival analysis, however, makes an assumption for how the covariates affect the shape and makes a distributional prediction for the shape of the model, i.e. exponential or log-logistic [14]. *R* code was used to conduct all three types of survival analysis. The code can be found in Appendix B.

### *3.4.2 Nonparametric Survival Analysis*

Nonparametric survival analysis is done using the *survfit* command in *R*. This command computes an estimate of a survival curve for censored data. The Kaplan-Meier plots are useful because they can show survival and hazard functions without any underlying assumptions of how the binary or categorical variables affect the shape of the curves and allow for us to see how the variables differ between the career fields.

The depiction of the survival curves for ABM, CSO, and Pilot with respect to *sex* are seen in Figure 2. For all of the career fields, *sex* is significant until the last CYOS subset, which is depicted in the Kaplan-Meier plots.

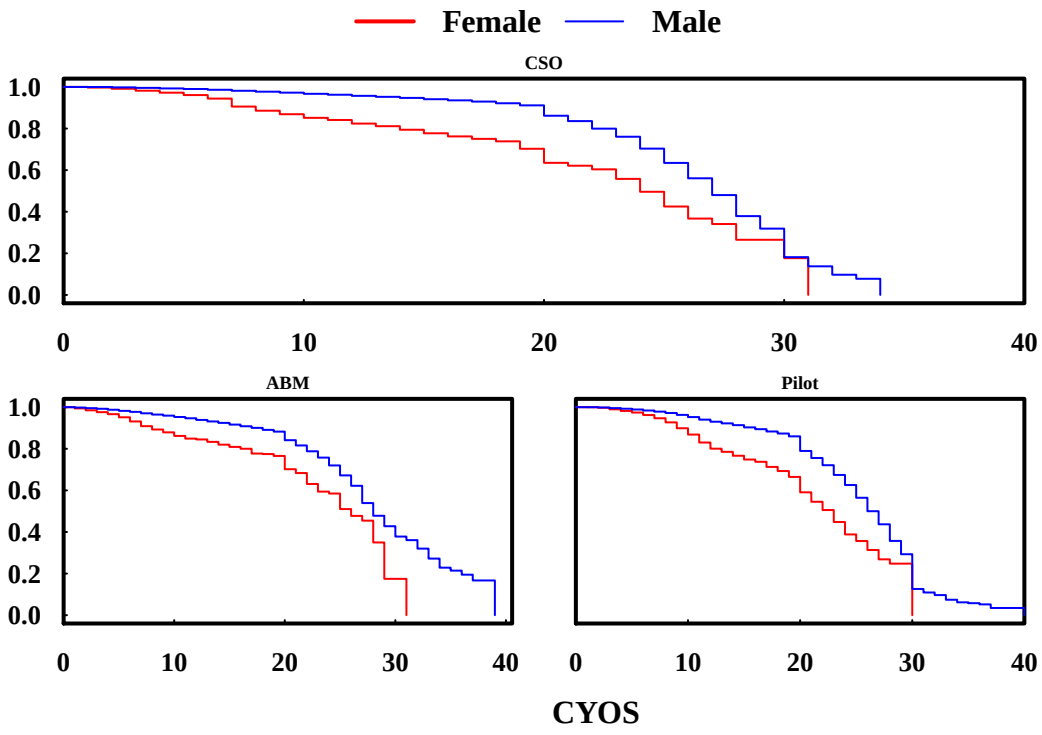


**Figure 2.** Kaplain-Meier plots (Sex)

The plots all look similar, but one thing of note is that in all three plots males retain in the Air Force longer tham females. All of the female survival curves stop before the male survival curves, indicating a lack of data past that CYOS.

Figure 3 shows the Kaplan-Meier plots of ABM and CSO split up by *marital status*. ABM and CSO are the only plots shown because Pilot, as shown in Appendix A, has a very similar Kaplan-Meier plot to CSO for *marital status*.

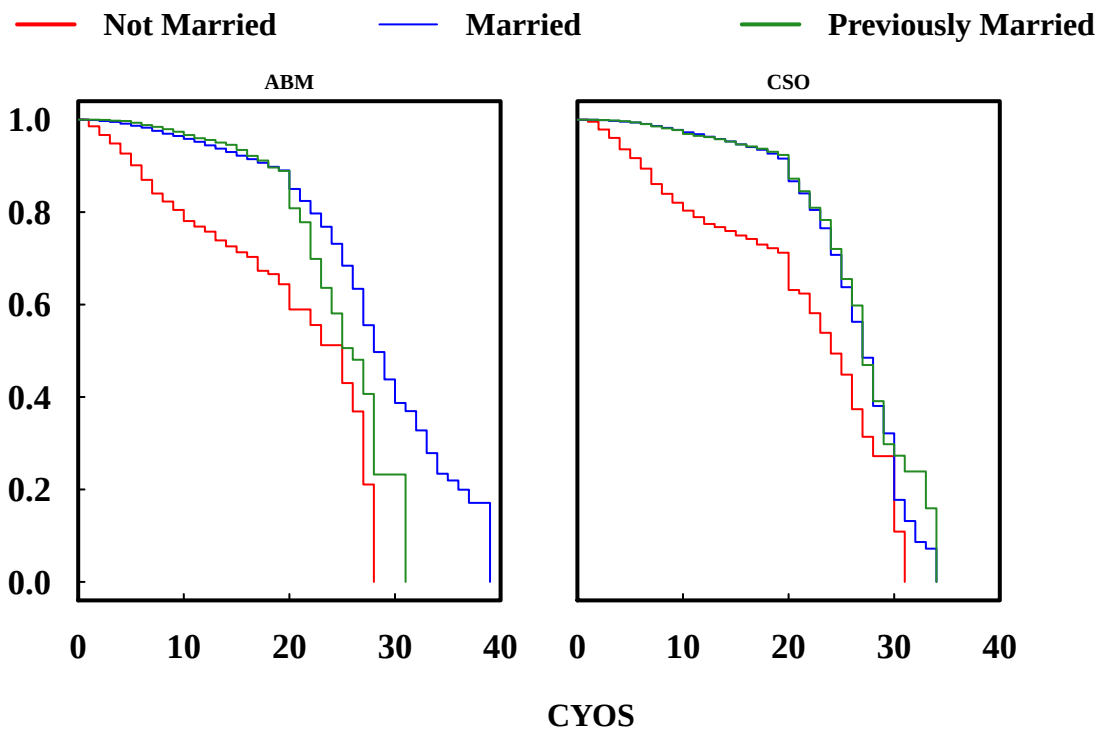


**Figure 3.** Kaplain-Meier plots (Marital Status)

In Figure 3 it seems early on, for all three career fields, married and previously married are similar and stay above not married until at least year 20. For ABM, married takes over predominantly as having the highest retention rate, but for all of the career fields not married has the lowest retention probability. This could be because without a spouse an officer might not need as much stability and may be more willing to separate and try their hand at the public sector.

Figure 4 shows all three Kaplan-Meier plots in respect to *dependents*, with 0 being no dependents and 1 being one or more dependents.

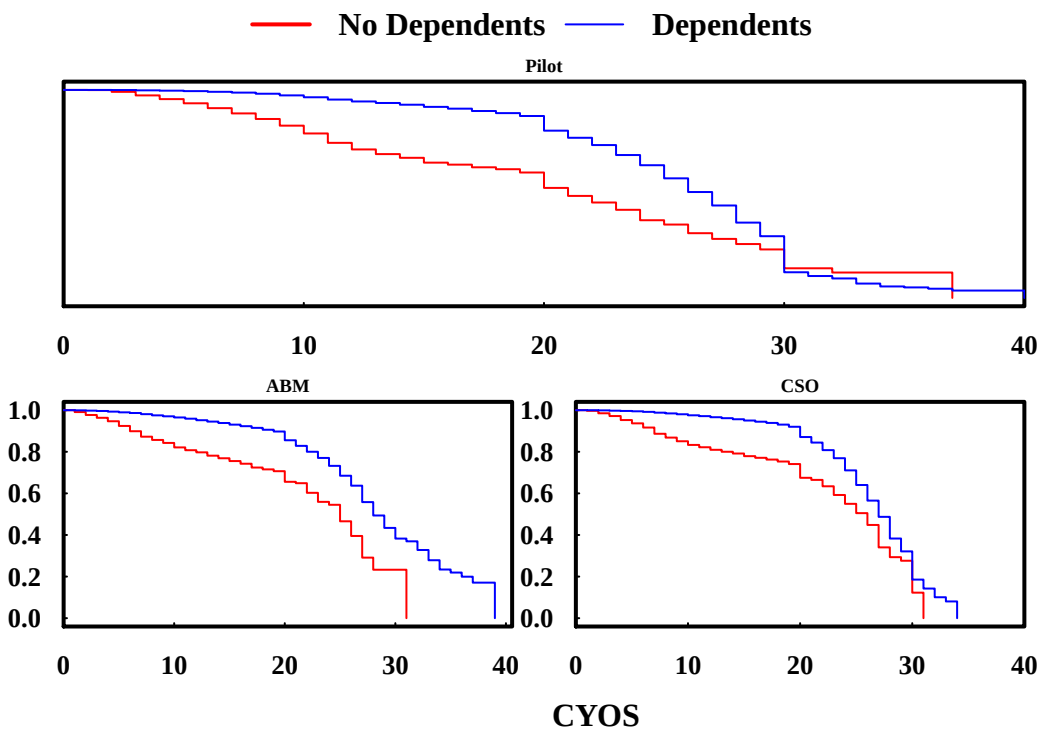


**Figure 4.** Kaplain-Meier plots (Dependents)

There is some variation between the three Kaplan-Meier plots, especially with the pilot plot showing the crossover from having dependents to not having dependents being the dominant retention curve at about year 30. CSO stays level until about 20 years for the officers having dependents, then begins to drop off, but, like ABM, always stays above not having dependents.

Figure 5 shows the *commissioning source* Kaplan-Meier plots. *SOC* 1 represents other commissioning sources, *SOC* 2 is USAFA, 3 is ROTC, and 4 is OTS.

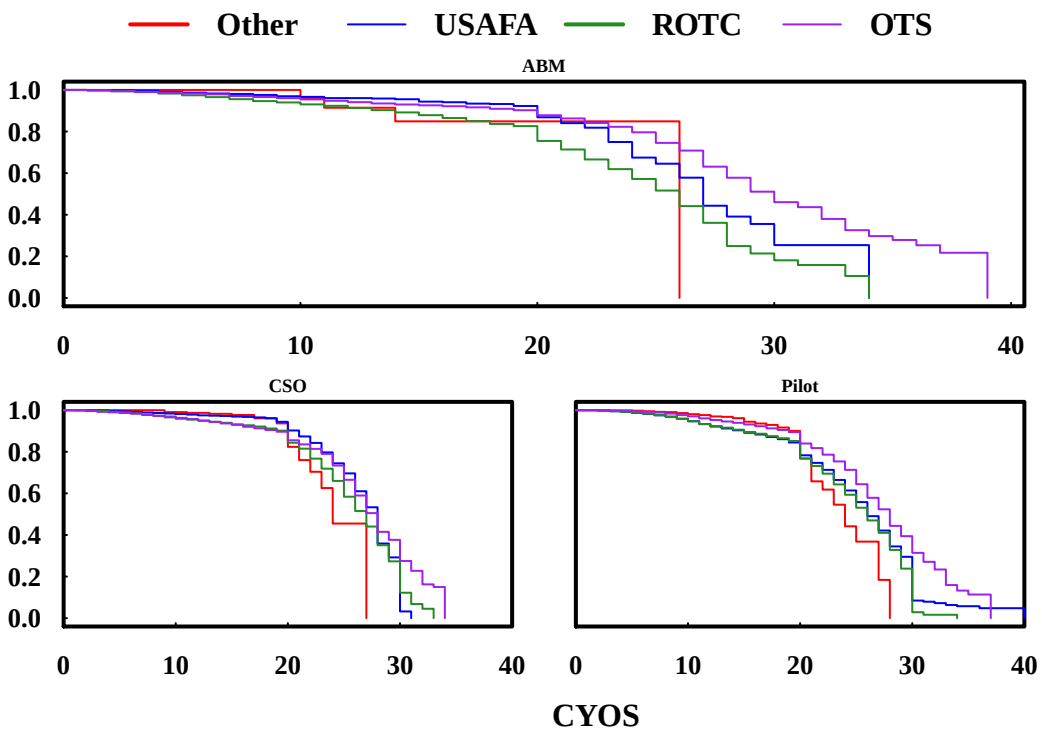


**Figure 5.** Kaplain-Meier plots (Commissioning Source)

The *commissioning source* plots show the most crossover between variables. ABM has *other commissioning source* being the dominant retention curve until 10 years, then *USAFA* until 20, *other commissioning source* for a few more years after that, and then *OTS* from 27 years on. This change toward the end could account for the fact that only about 20% of the officers within the data stay until the 20 year mark so there is not a large sample.

All of the Kaplan-Meier plots for *commissioning source* have a lot of crossover, so it may be hard to pinpoint a *commissioning source* that is more likely to stay in. CSO once again is very level to year 20, 50% stay past year 15, 40% stay past year 20. This attrition rate is much lower than the rest of the Air Force, so CSO seems to be stable; once an officer enters the career field there is a 40% chance they retire at 20 years.

Figure 6 illustrates retention for *distinguished graduates* for all three career fields.

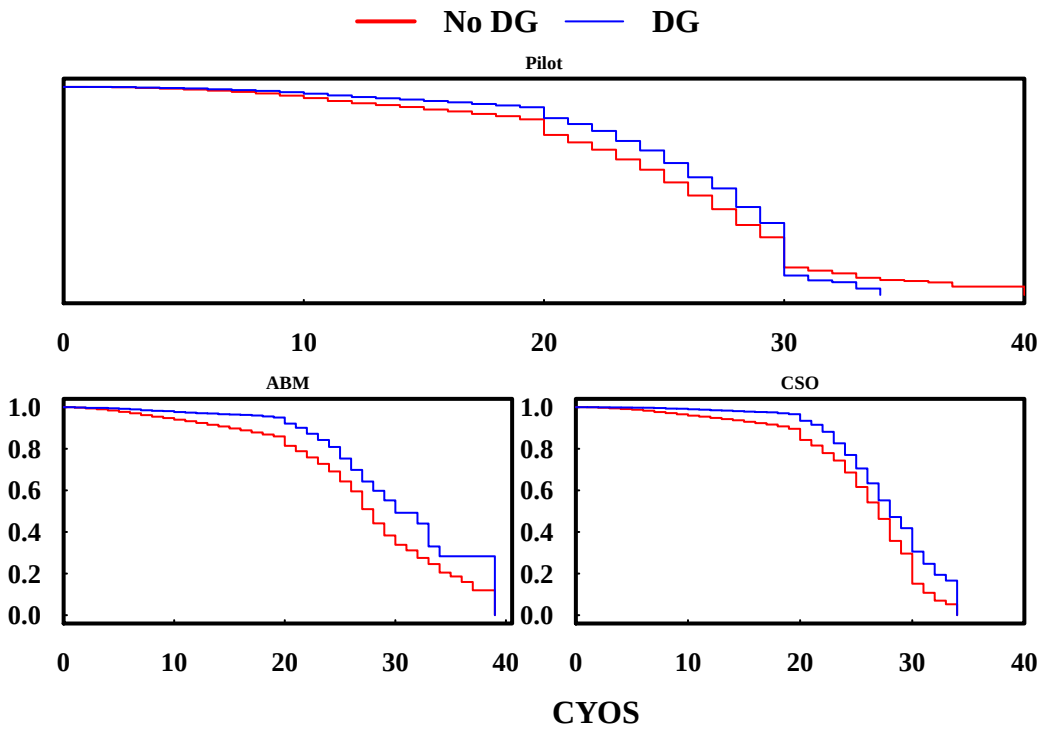


**Figure 6.** Kaplain-Meier plots (Distinguished Graduate)

All three of the plots seem to have a steady trend between the two *distinguished graduate* strata, with the only crossover being with pilot at the 30-year mark. There does not seem to be too much of a difference between the two *distinguished graduate* categories, but *distinguished graduates* do have a slightly higher survival than *non-distinguished graduates* except at the year 30 mark for pilots.

Figure 7 shows the retention curves split by *prior service* for only CSO, because the other two career fields' plots look very similar.

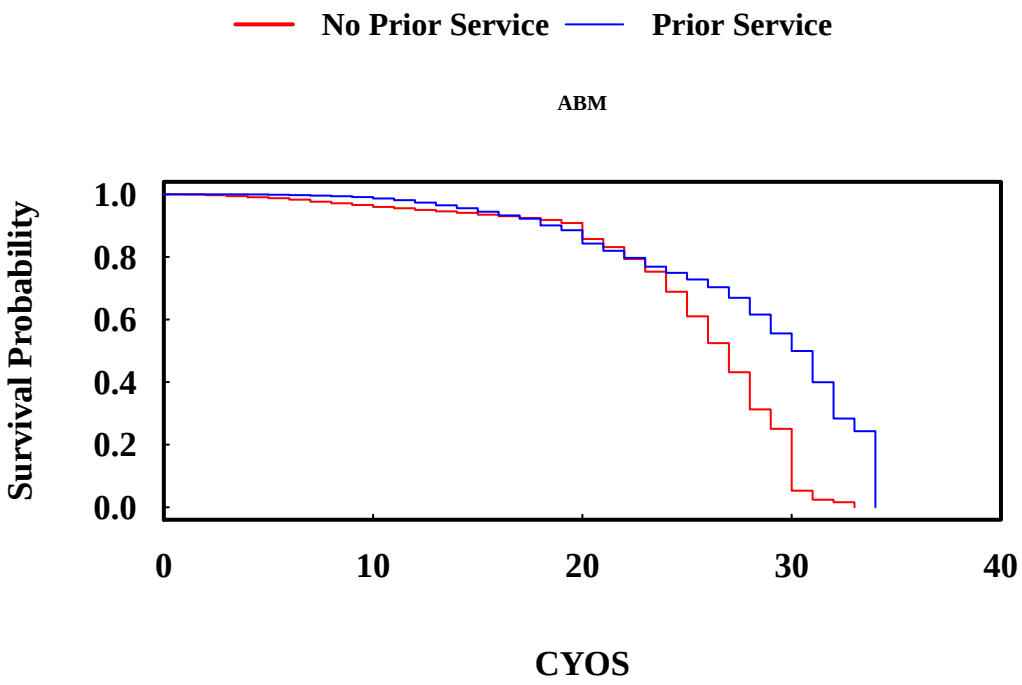


**Figure 7.** Kaplain-Meier plots (Prior Service)

From about year 17 to year 22, not being *prior service* is above being *prior service*, the only time that happens. This might be because officers with *prior service* are closer to retirement because of the previous time they served, so at year 17 officers without *prior service* are almost at retirement and are more likely to stay in, but the prior service officers may be past their 20-year mark at this 17-year commissioned point.

### *3.4.3 Semiparametric Survival Analysis*

Semiparametric survival analysis makes assumptions about how the covariates affect the shape of the hazard function, and commonly involves Cox Proportional Hazards regression models. However, semiparametric models do not assume specific probability distributions for model estimation. Instead, they assume an arbitrary probability distribution. As a result, the interpolation of their estimates is relative. *Coxph* and *phreg* are used within *R* to create the Cox survival functions and to find the Wald's p-value of the variables, shown in

Table 1, where all variables are significant except *commissioning source* for Pilot, which was explained earlier. Semiparametric survival analysis was used previously for the Air Force's ABM, CSO, and Pilot career fields by Franzen [25]. The semiparametric survival analysis is illustrated in order for comparison to parametric survival analysis, done in the next section.

Figure 8 illustrates the semiparametric survival curve for the officer subsets. Unlike other survival analysis software like *SAS*, *R* does not provide a graph for all different combinations of the survival possibilities (different combinations of the categorical variables). This inhibits the ability to compare the groupings of the survival curves, but the ultimate survival curve is still produced in *R* just like as in *SAS*.

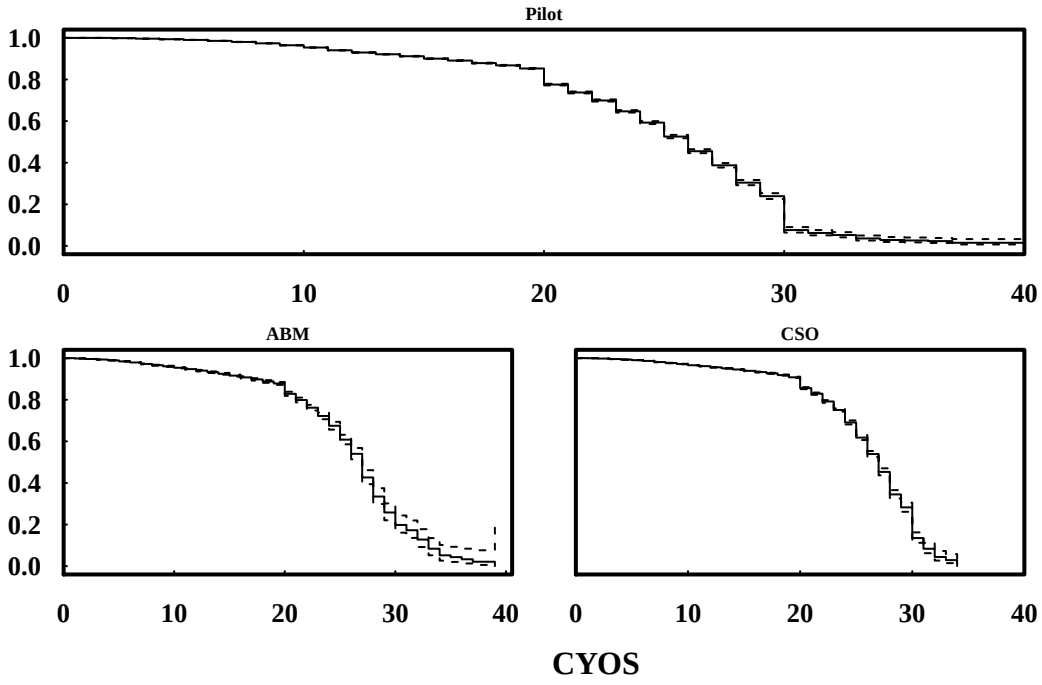


**Figure 8.** Cox PH Plots

In Figure 8, ABM and Pilot look similar, but Pilot has a less gradual drop off starting at 20 years. Once again CSO only graphs out to year 30 because of the lack of survival past year 30 in the CSO career field. Statistical metrics are taken from the semiparametric models and used to compare against the parametric model shown in the Results and Analysis section.

### *3.4.4 Parametric Survival Analysis*

Parametric survival analysis has the most assumptions within the different survival analysis techniques; it makes an assumption about how the covariates affect the shape and it makes an assumption that the data follows a probabilistic distribution. Within *R*, parametric survival analysis is done using the *flexsurvreg* command, which allows for the input of distributions to fit the survival data to.

The *flexsurvreg* function has ten possible distributions; generalized gamma (stable), generalized gamma (original), generalized F (stable), generalized F (original), weibull, gamma, exponential, log-logistic, log-normal, and gompertz. In our calculations to find the best distributional fit, weibull was not used because of its incompatibility with the data. The distribution used is the log-logistic distribution, and the reasoning is shown in the Results and Analysis section.

The parametric survival model for the three career fields are shown in Figures 9, 10, and 11.

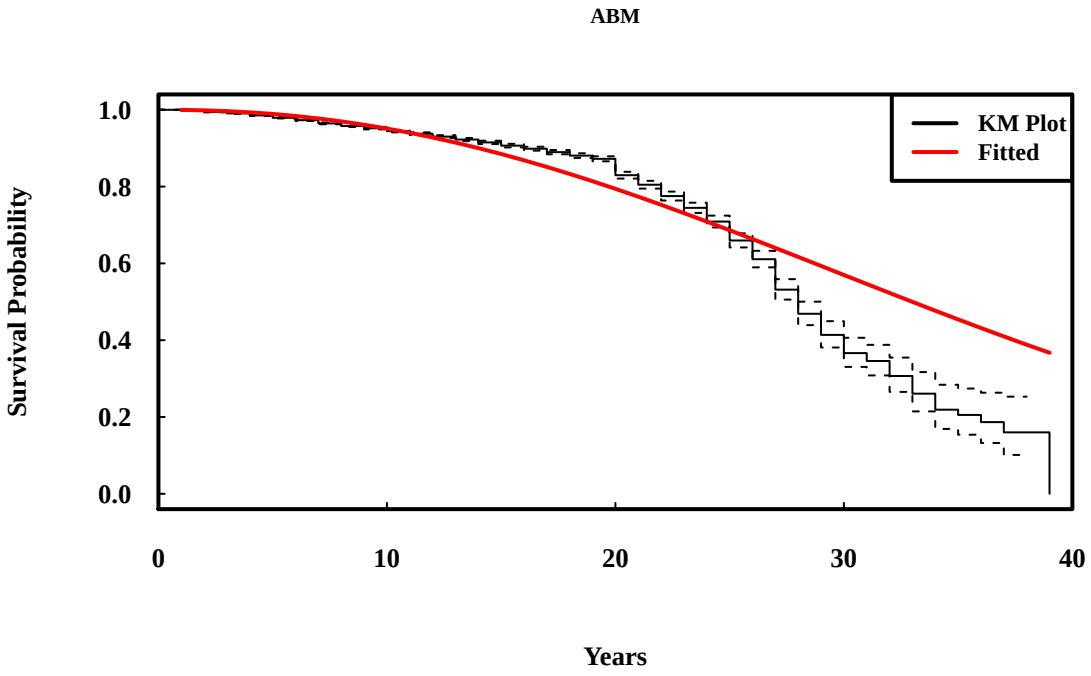


**Figure 9.** Parametric plot for ABM

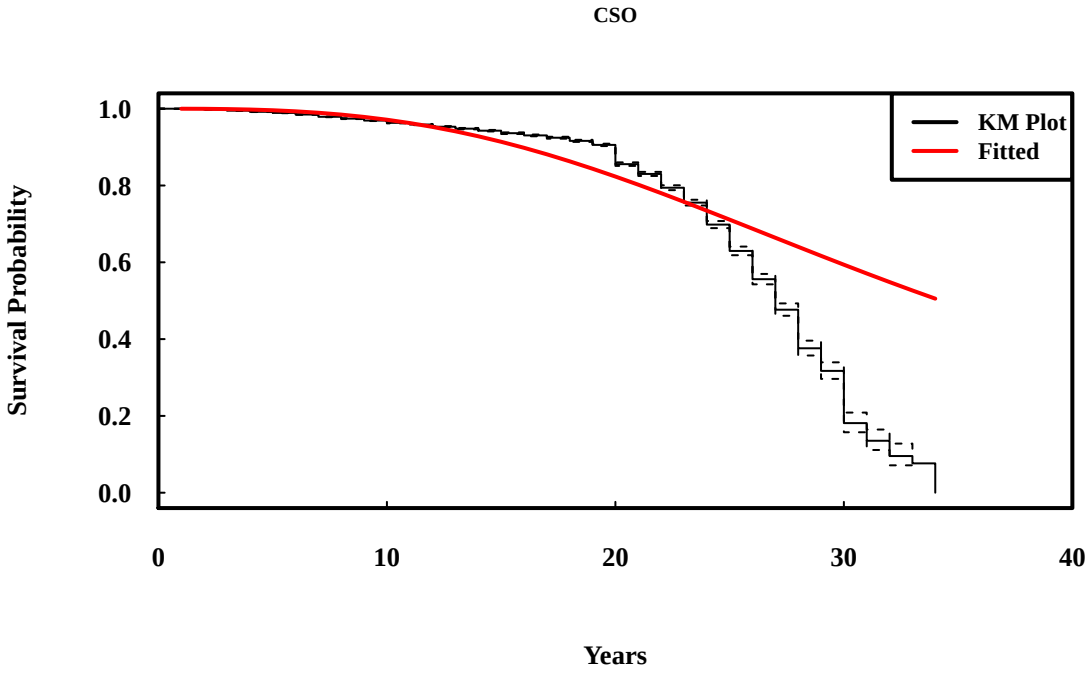


**Figure 10.** Parametric plot for CSO

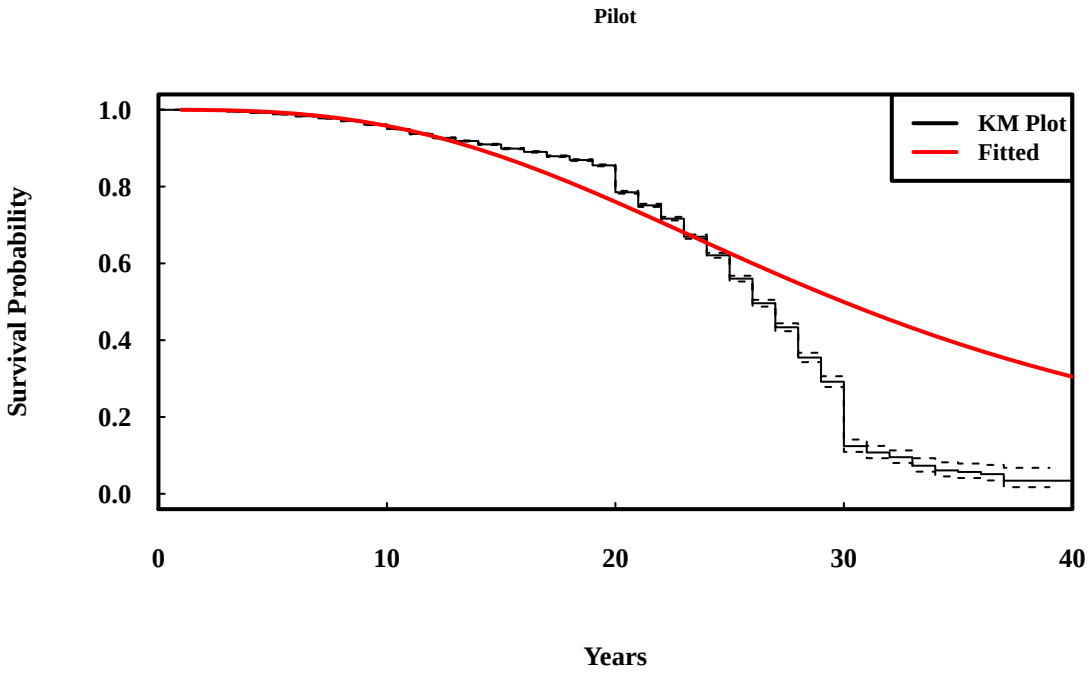


**Figure 11.** Parametric plot for Pilot

For Figures 9, 10, and 11, the red line is the survival estimate using the parametric model, while the black line is the Kaplan-Meier model for the same data. The parametric estimate for ABM is very close to that of the Kaplan-Meier plot, while the other two career fields show a substantial difference between the two methods of survival analysis. All three parametric survival curves are done using the log-logistic distribution and are examined further in the Akaike Information Criterion subsection of the Results and Analysis section. The results of the variable estimates for all three career fields are shown in Table 5.

**Table 5.** Parametric Model Parameters

| Variables | ABM | CSO | Pilot |
|---|---|---|---|
| Sex (Male) | 0.88956 | 0.95761 | 0.96387 |
| Marital Status (Married) | 0.82956 | 0.8783 | 0.88496 |
| Marital Status (No Longer) | 0.06267 | 0.04772 | 0.03994 |
| Dependents (Yes) | 0.82028 | 0.88797 | 0.88049 |
| SOC (USAFA) | 0.07798 | 0.11572 | 0.43161 |
| SOC(ROTC) | 0.48699 | 0.53732 | 0.43686 |
| SOC(OTS) | 0.43318 | 0.3419 | 0.1268 |
| DG (Yes) | 0.11032 | 0.11424 | 0.16474 |
| Shape | 2.3003 | 2.86046 | 3.015904 |
| Scale | 17.1203 | 17.15319 | 13.525771 |

Table 5 shows the parameter estimates found in the above plots. The parameters allow for other log-logistic regressions to be done without obtaining any survival data and fitting the model to the model parameters. These plots and model parameters are found using an Accelerated Failure Time model, which is discussed further below.

### Proportional Hazards Model

Proportional Hazards regression (PH) is a form of parametric survival analysis which the assumes that the effect of the covariates on the hazard function are to increase or decrease proportionally throughout all time periods [26]. A log-logistic distribution is not available for a PH regression, so an exponential distribution is used on the ABM data set as an example in Table 6.

**Table 6.** PH Model Parameters

| Variables | Coefficient |
|---|---|
| Sex (Male) | -0.003 |
| Marital Status (Married) | -0.612 |
| Marital Status (No Longer) | -0.441 |
| Dependents (Yes) | -0.15 |
| SOC (USAFA) | 0.062 |
| SOC(ROTC) | 0.713 |
| SOC(OTS) | 0.049 |
| DG (Yes) | 0.073 |
| Prior Service (Yes) | -2.223 |

PH regression's covariates behave similar to those in a linear or logistic regression; the coefficients on the variables are a multiplier for all time periods. If the estimated coefficient is negative, this indicates a lower risk of an event (retirement) with the increase of that variable. Since all of the variables are categorical and non-continuous, a negative value means retirement is less likely to happen. The PH regression model shown in Table 6 yields an AIC value of *26576.17* for the exponential distribution, while the log-logistic Accelerated Failure Time model yields an AIC of *21331.11* and the exponential Accelerated Failure Time model yields an AIC of *24608.87*. For the specific data set, Accelerated Failure Time models yield the best result.

When using the *flexsurvreg* command, the distribution that is best fit determines whether or not the PH or AFT model is best; the PH model is only used for the exponential and gompertz distributions. Outside of the *flexsurvreg* command, the proportional hazards model can only be used for three distributions; gompertz, exponential, and weibull [27]. If another distribution is chosen that is within one of these models the proportional hazards model can be chosen, otherwise the accelerated failure time model is appropriate. Since all of the

distributions chosen were log-logistic, the accelerated failure time models were used.

### Accelerated Failure Time Model

Accelerated failure time models (AFT) differ from proportional hazards model. While proportional hazards models assume that the effect of the covariate is to multiply the hazard by some constant, accelerated failure time models assume that the effect of the covariate is to speed up or slow down time until an event. In other words, in the PH model the parameter estimates act multiplicatively on the variables while in an AFT model the parameter estimates act multiplicatively on time [28]. The results of the AFT regression for the ABM, CSO, and Pilot data are shown in Table 5.

If the estimated coefficient is negative, this implies decreasing survival times and shorter expected durations. For example, in ABM when *commissioning source (SOC)* is 3 or 4 (ROTC or OTS), the expected time of survival is less than the baseline (SOC=1, other). For those officers from USAFA (SOC = 2), the expected time of survival is longer than that of the baseline. This could be because officers commissioned through OTS are generally, but not always, older than their counterparts from the USAFA, ROTC, and other commissioning sources. For *sex*, when an officer is male, the parameter estimate for ABM is *0.88956*, and taking the exponent of that value yields *2.43*, which means that in comparison to females, males reach survival half as slow (*1/2.43*). In other words, females are twice as fast to get out as compared to males.

### YOS as independent variable

Franzen [25] used only the variables *sex*, *marital status*, *dependents*, *commissioning source*, *distinguished graduate*, and *prior service* in her semiparametric survival analysis models for rated officers. In order to compare our model to hers, we saw it fit to use the same variables in order to determine an exact comparison of survival analysis techniques. Doing this, one thing we missed out on was using the years of service of an Air Force officer

in hopes of better predicting the retention rates of officers. As expected, the amount of years an individual stays in the Air Force will affect the ability of us to model whether or not an officer gets out of the Air Force. In order to add on to the current methodology, we included *years of service* as an independent variable to better predict retention. Years of service of an officer and the CYOS subgroups are highly correlated; so the model was only done on the entire data set for each career field. The plots for ABM, CSO, and Pilot are shown in Figures 12, 13, 14. The log-logistic distribution was selected and used in order to keep consistency between the two methodologies.

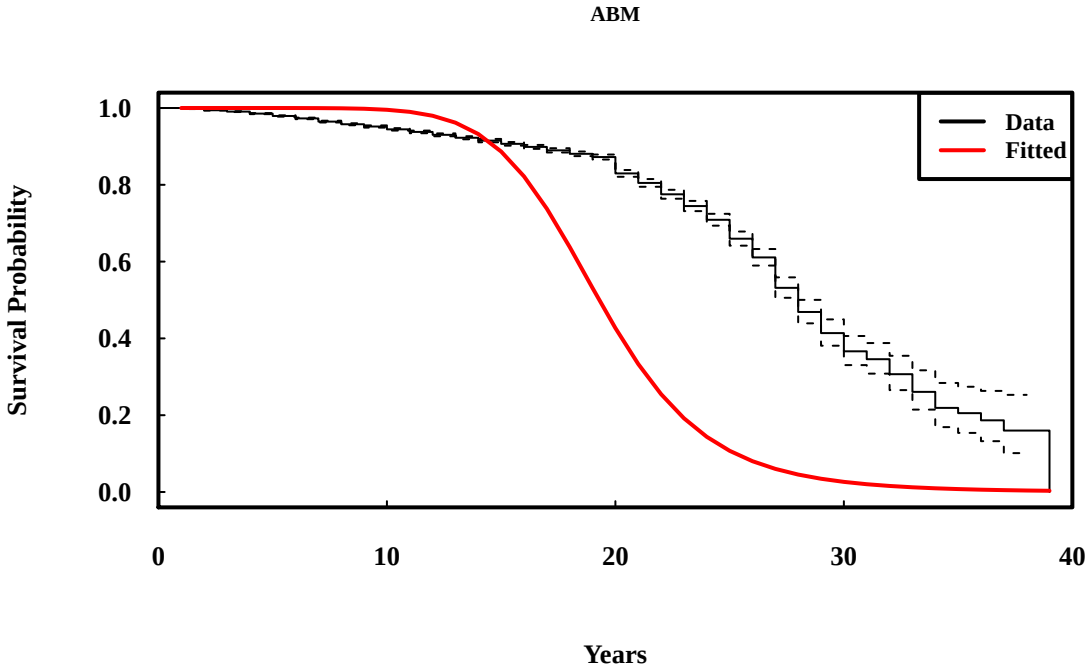


**Figure 12.** Parametric plot for ABM with YOS

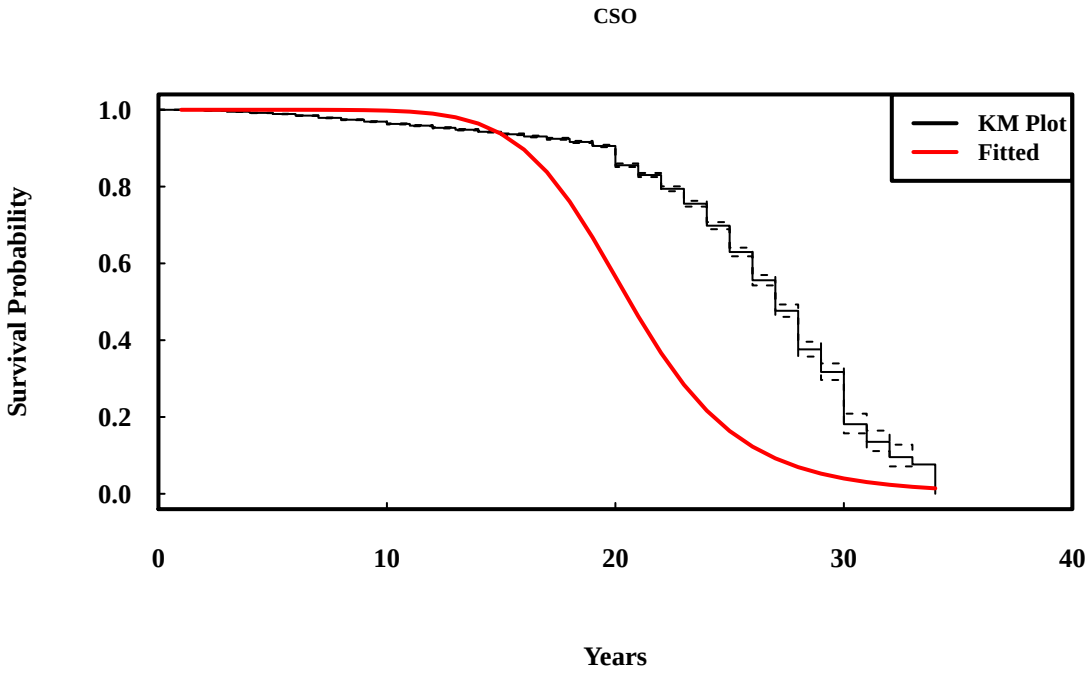


**Figure 13.** Parametric plot for CSO with YOS

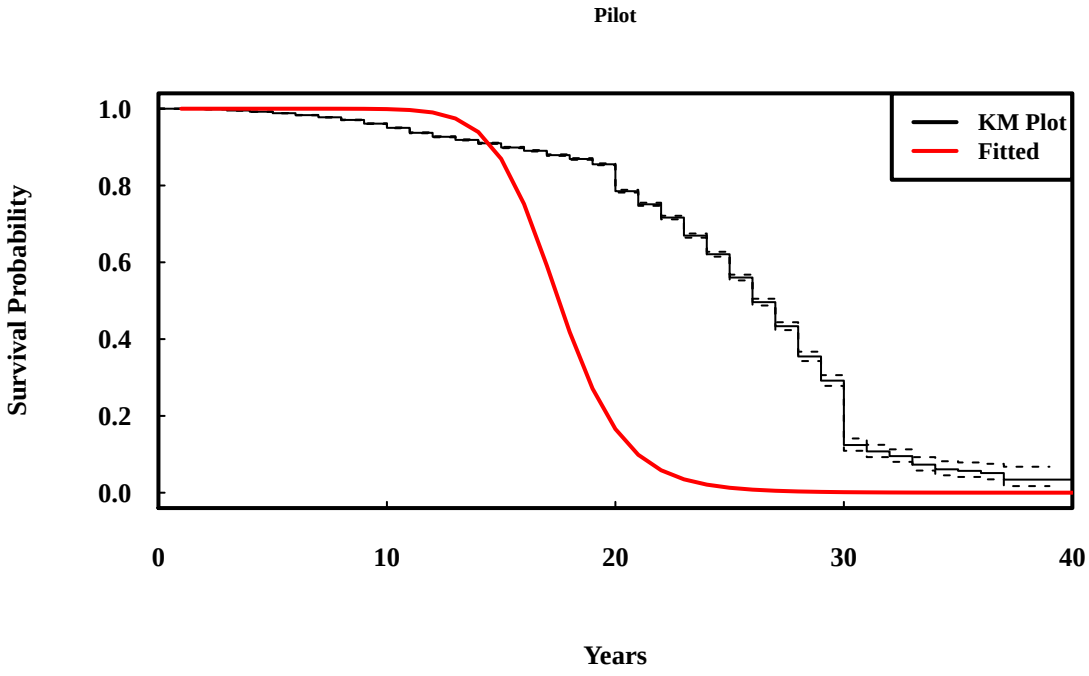


**Figure 14.** Parametric plot for Pilot with YOS

Figures 12, 13, 14 show the new parametric plots using the log-logistic distribution while including Years of Service as an independent variable. The plots look like they predict the data worse; the remain at about 100% until year 13, then drop from 100% to 0% rapidly. This looks to show that adding Years of Service would not benefit better modeling of retention rates, but that will be examined more in the Results and Analysis section.

# IV. Results and Analysis

## 4.1 Introduction

Survival Analysis has produced survival lines for the officers within the ABM, CSO, and Pilot career fields. The semiparametric and parametric models are compared to determine the best model to fit the data.

## 4.2 Results

### *4.2.1 Distribution Selection*

To determine the best distributional fit for the parametric models, the comparison metric used is the Akaike Information Criterion (AIC). The AIC was calculated for all CYOS' within each career field, and the career field as a whole using both semiparametric and fully parametric survival analysis. The AIC is a common metric used to compare semiparametric and parametric models, illustrating which model performs better [29]. The AIC table comparing the distributions for ABM is in Table 7.

**Table 7.** AIC for ABM

| Distribution | AIC0to6 | AIC4to8 | AIC8to14 | AIC12to19 | AIC20to22 | AIC |
|---|---|---|---|---|---|---|
| Exponential | 24608.87 | 22151.45 | 14809.40 | 9060.22 | 6695.26 | 24608.87 |
| Generalized F (original) | 21511.70 | 18624.33 | 10885.54 | 5604.58 | 4019.07 | 21511.70 |
| Generalized Gamma (stable) | 21509.19 | 18972.63 | 10953.51 | 5711.78 | 3740.49 | 21509.19 |
| Generalized Gamma (original) | 21326.88 | 18578.07 | 11065.23 | 6268.34 | 3850.47 | 21326.88 |
| Gompertz | 21674.78 | 19288.46 | 11370.74 | 6115.91 | 4464.00 | 21674.78 |
| Log-logistic | 21331.11 | 18543.35 | 10808.54 | 5580.95 | 3705.29 | 21331.11 |
| Log Normal | 21368.32 | 18584.98 | 10951.07 | 5602.81 | 3638.30 | 21368.32 |
| Weibull | 21282.74 | NA | NA | NA | NA | 26580.17 |
| Gamma | 21429.02 | 18550.13 | 11193.67 | 6158.81 | 4276.95 | 27210.28 |
| Generalized F (stable) | 21395.24 | 18622.49 | 10942.96 | 5934.13 | 3964.84 | 27049.35 |

The heading labels show the CYOS groups: 0-6, 4-8, 8-14, 12-19, 20-22. The log-logistic distribution is the best for three of the five subgroups, with weibull and log normal being the other two. The weibull distribution has a lot of NAs which inhibits the ability to choose that distribution; it cannot work with the data sometimes. The generalized gamma (original) distribution has the lowest overall AIC, with the log-logistic less than 0.1% behind. This does not cause any concern for picking the log-logistic because of how close the AICs are to each other.

The AIC table for the parametric distributions for CSO can be found in Table 8.

**Table 8.** AIC for CSO

| Distribution | AIC0to6 | AIC4to8 | AIC8to14 | AIC12to19 | AIC20to22 | AIC |
|---|---|---|---|---|---|---|
| Exponential | 78772.74 | 73478.43 | 55324.57 | 42753.85 | 34633.09 | 86630.08 |
| Generalized F (original) | 66693.08 | 60203.00 | 40263.59 | 25750.54 | 19766.33 | 77559.79 |
| Generalized Gamma (stable) | 68710.11 | 61649.02 | 40456.75 | 25530.37 | 19884.51 | 80714.11 |
| Generalized Gamma (original) | 68453.58 | 62917.52 | 41017.86 | 25584.70 | 19806.49 | 78762.38 |
| Gompertz | 69400.23 | 63866.29 | 40318.25 | 32916.91 | 24289.22 | 79000.94 |
| Log-logistic | 66508.04 | 59836.59 | 39337.68 | 25264.08 | 18607.15 | 77489.62 |
| Log Normal | 67320.81 | 60653.88 | 40025.79 | 25353.21 | 18358.40 | 78570.17 |
| Weibull | 65910.37 | NA | NA | NA | NA | 77258.49 |
| Gamma | 66612.68 | 60128.28 | 40770.22 | 29600.52 | 23462.59 | 78829.83 |
| Generalized F (stable) | 66797.81 | 59891.82 | 40035.31 | 29823.30 | 19981.76 | 77971.74 |

Once again, the weibull distribution cannot evaluate the data for a lot of the CYOS subsets, even though it has the best overall AIC. The log-logistic falls behind again by 0.1% and does not have any missing values. The generalized gamma (original) is now 6th, showing inconsistency. Log-logistic has been the most consistent throughout, and like for ABM, was the best at predicting the middle years of data for the CSO career field.

The AIC table for the parametric distributions for pilot can be found in Table 9.

**Table 9.** AIC for Pilot

| Distribution | AIC0to6 | AIC4to8 | AIC8to14 | AIC12to19 | AIC20to22 | AIC |
|---|---|---|---|---|---|---|
| Exponential | 253516.21 | 232664.79 | 175097.44 | 111833.10 | 82825.73 | 274236.13 |
| Generalized F (original) | 213849.39 | 191851.94 | 131747.12 | 70588.29 | 50179.96 | 241548.05 |
| Generalized Gamma (stable) | 216052.15 | 190646.71 | 133484.81 | 71475.57 | 49137.20 | 247593.62 |
| Generalized Gamma (original) | 220387.94 | 190543.75 | 132978.19 | 79249.15 | 48530.41 | 242366.90 |
| Gompertz | 244227.51 | 215610.92 | 158996.00 | 82343.91 | 56003.24 | 248719.76 |
| Log-logistic | 213804.85 | 189678.57 | 131545.15 | 70191.14 | 46155.58 | 241442.74 |
| Log Normal | 214770.98 | 190222.56 | 132651.52 | 70629.18 | 45232.73 | 243692.99 |
| Weibull | 213253.80 | 189142.90 | NA | NA | NA | 241083.00 |
| Gamma | 213889.80 | 189806.30 | 133299.40 | 72464.74 | 52266.63 | 242115.80 |
| Generalized F (stable) | 215867.90 | 191052.40 | 131872.00 | 71108.15 | 48429.36 | 242821.20 |

The weibull still has NA values, but is ranked number one on datasets it can evaluate. The log-logistic is second, behind by less than 0.1%, showing that it is the most consistently good distribution of the available options.

When looking at the tables for all of the AICs given their proper distribution and CYOS group, there is a noticeable difference between the AICs within the different CYOS'. Starting with the first CYOS (0-6), the AIC is generally higher than the other CYOS groups, and as the CYOS' change with time, the AICs lower. The early CYOS' have more officers, leading to more data points, which ultimately increases the AIC because it takes into account the amount of data being analyzed. The AICs cannot be compared across career fields and CYOS' because of the difference in datapoints, but within the same career field and CYOS the AIC can be compared.

When the weibull does not have an N/A, it is usually the best. This inconsistency, however, leads us to stay away from the weibull distribution. On average the log-logistic distribution is the best, and it is the most consistent; always coming in either first or second. The log-logistic distribution is the distribution chosen for the parametric survival analysis.

Figure 15 shows the parametric distributions; exponential, log-normal, gompertz, general gamma (original), generalized gamma (stable), log-logistic, and weibull, in comparison to the Kaplan-Meier plot for ABM.

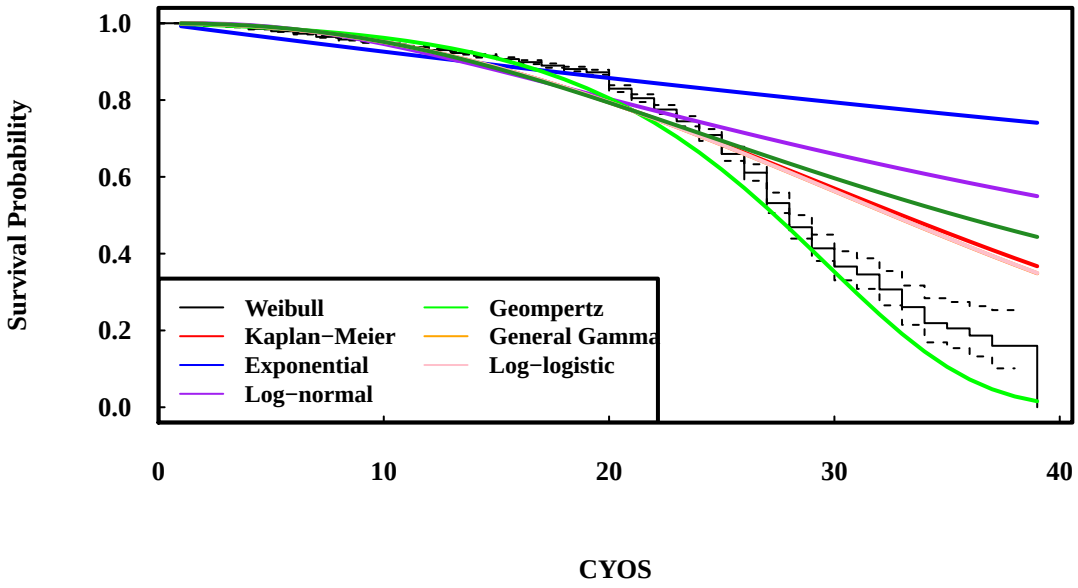


**Figure 15.** Parametric Distributions for ABM

Figure 15 shows the different distribution options for a parametric model. The chosen distribution, log-logistic, is in dark green. The weibull, log-normal, generalized gamma (original), generalized gamma (stable), and the log-logistic are similar in the beginning of the data; from approximately CYOS' 1 to 10. The gompertz distribution seems to fit the data well for the last 20 years, which we know is a problem for the log-logistic distribution. The reason this is occuring, but is still not the best distributional option, is because the

log-logistic fits the data very well until approximately year 27, which makes up for the latter years.

### *4.2.2 Parametric vs Semiparametric*

Using log-logistic means next using the AIC to make comparisons between the parametric and semiparametric models. The AIC is used to compare the semiparametric and parametric models. The AIC table for ABM can be found in Table 10.

**Table 10.** AIC for ABM: Parametric vs Semiparametric

| Distribution | AIC0to6 | AIC4to8 | AIC8to14 | AIC12to19 | AIC20to22 | AIC |
|---|---|---|---|---|---|---|
| Log-logistic | 21331.11 | 18543.35 | 10808.54 | 5580.95 | 3705.29 | 21331.11 |
| Semiparametric | 34922.99 | 30208.75 | 17604.41 | 9397.01 | 6360.3 | 34922.99 |
| Percent Difference | 38.9 | 38.6 | 38.6 | 40.6 | 41.7 | 38.9 |

The AIC for log-logistic is better than for the semiparametric model, with a percent change ranging from 38.6% to 41.7%. The AIC comparison for CSO is in Table 11.

**Table 11.** AIC for CSO: Parametric vs Semiparametric

| Distribution | AIC0to6 | AIC4to8 | AIC8to14 | AIC12to19 | AIC20to22 | AIC |
|---|---|---|---|---|---|---|
| Log-logistic | 66508.04 | 59836.59 | 39337.68 | 25264.08 | 18607.15 | 77489.62 |
| Semiparametric | 122895.92 | 111636.07 | 76313.61 | 54300.88 | 41685.5 | 141056.725 |
| Percent Difference | 45.9 | 46.4 | 48.5 | 53.5 | 55.4 | 45.1 |

The CSO career field is the second biggest dataset, so the difference in AIC between the two models starts to get bigger. The AICs in general are also getting bigger. The AIC comparison for pilot can be found in Table 12.

**Table 12.** AIC for Pilot: Parametric vs Semiparametric

| Distribution | AIC0to6 | AIC4to8 | AIC8to14 | AIC12to19 | AIC20to22 | AIC |
|---|---|---|---|---|---|---|
| Log-logistic | 213804.85 | 189678.57 | 131545.15 | 70191.14 | 46155.58 | 241442.74 |
| Semiparametric | 462150.7 | 411872.81 | 286261.89 | 162418.31 | 112436.92 | 516648.61 |
| Percent Difference | 53.7 | 53.9 | 54 | 56.8 | 58.9 | 53.3 |

The pilot dataset is the largest, so the AICs have also gotten much bigger. the percent difference did not change much from the CSO to the pilot career fields, but the percent difference between the log-logistic model and the semiparametric model is around 50% for these two datasets.

Tables 10, 11, and 12 show the AIC for all three career fields at each CYOS, along with the total AIC for the data across the entire career field. The percent difference between the log-logistic AIC and the semiparametric AIC ranges from 38.6% to 58.9%. Since the number of parameters and covariates, along with the datasets, are the same size, that means the log-likelihood is drastically different between the two. Tables 10, 11, and 12 show that the parametric model models the data better than the semiparametric model.

### *4.2.3 YOS as independent variable*

Table 13 shows the new AIC comparison tables with including Years of Service as an independent variable for all career fields.

**Table 13.** AIC for ABM, CSO, and Pilot with YOS

| Distribution | ABM | CSO | Pilot |
|---|---|---|---|
| Exponential | 26123.32 | 80236.50 | 255734.30 |
| Generalized F (original) | 18136.23 | 50379.74 | 159517.20 |
| Generalized Gamma (Stable) | 18247.28 | 59034.05 | 179155.80 |
| Generalized Gamma (original) | 18347.19 | 57593.09 | 187481.90 |
| Gompertz | 5927.77 | 24476.59 | 7457.80 |
| Log-logistic | 16188.06 | 45743.46 | 129940.00 |
| Log Normal | 17395.87 | 47687.04 | 140091.30 |
| Weibull | 15424.34 | 41760.59 | 120281.60 |
| Gamma | 19710.90 | 61084.95 | 112088.80 |
| Generalized F (stable) | 19243.89 | 54145.47 | 171661.60 |

For ease of observation, table 14 takes the log-logistic distribution from the original model (without Years of Service) and the new model (with Years of Service) to better show the differences between the two.

YOS? ABM CSO Pilot 1 No 21331.11 77489.62 241442.7 2 Yes 16188.06 45743.46 129940.0

**Table 14.** AIC with YOS comparisons

| YOS? | ABM | CSO | Pilot |
|---|---|---|---|
| No | 21331.11 | 77489.62 | 241442.74 |
| Yes | 16188.06 | 45743.46 | 129940.00 |

Table 14 shows the two models side-by-side; the top row of the table showing the original model without Years of Service as an independent variable, and the bottom row of the table showing Years of Service as an independent variable. The row with Years of Service as an independent variable has a better AIC than the model without Years of Service as an independent variable. This is alarming after looking at the plots shown in the Methodology section because the plots did not look like they fit the data well, but their AICs are better than the original model. Even though the AICs are better, the plots are cause for concern, so the model with Years of Service is not analyzed any further.

## 4.3 Analysis

### *4.3.1 Residuals*

The data and the parametric models are compared to test how well the parametric model models the survival data. The survival probability at each Year (CYOS = 1-39) is extracted from the data and the parametric model, then the difference between the two points for a given CYOS are squared. The residual plots are shown in Figures 16, 17, and 18. The sum of the squared error (SSE) is shown in Tables 15, 16, and 17, for ABM, CSO, and pilot, respectively.

#### ABM

The SSE and residual plot for the ABM career field is shown in Table 15 and Figure 16.

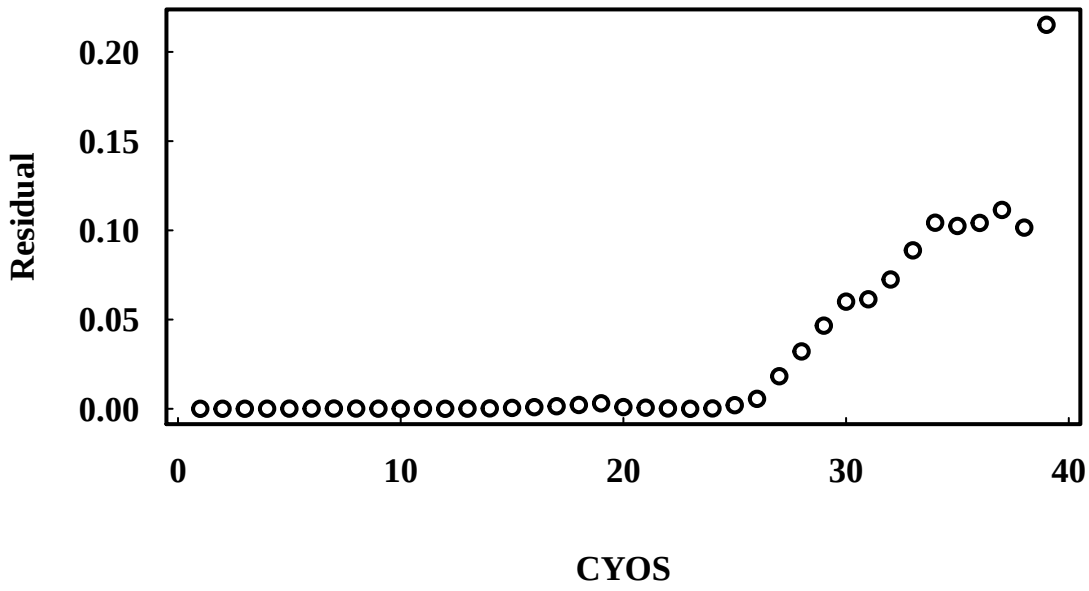


**Figure 16.** Residual plot for Parametric model (ABM)

Figure 16 shows that the residuals are pretty steady until year 20, where officers can retire from the Air Force and get their full pension. It is expected that more officers get out at this point. The residuals then return to a very steady line until year 27, where the residuals begin to increase rapidly. This could be because there are such few people staying in this late in their career that the model sacrifices the small fraction of data points in this area in order to fit the data points in the beginning of the model better. The model decides it is better to fit the earlier years more accurately because there are more data points, but in turn sacrifices the accuracy of the later years.

The SSE for both the total model and until year 27 are shown in Table 15.

**Table 15.** Sum of Squared Error for Parametric model

| SSE | <= 27 years |
|---|---|
| 1.14 | 0.04 |

The SSE for the total model is high, but looking at the residual plot in Figure 16, when the time gets greater than 27 years the residual plot shows larger residuals. Having an SSE closer to 0 is better because it means that the model being tested models the data perfectly; there is no difference between the data and the model. Although this is not the case, an SSE of 0.037 shows a very strong model for modeling years 1-27.

### CSO

The SSE and residual plot for the CSO career field is shown in Table 16 and Figure 17.

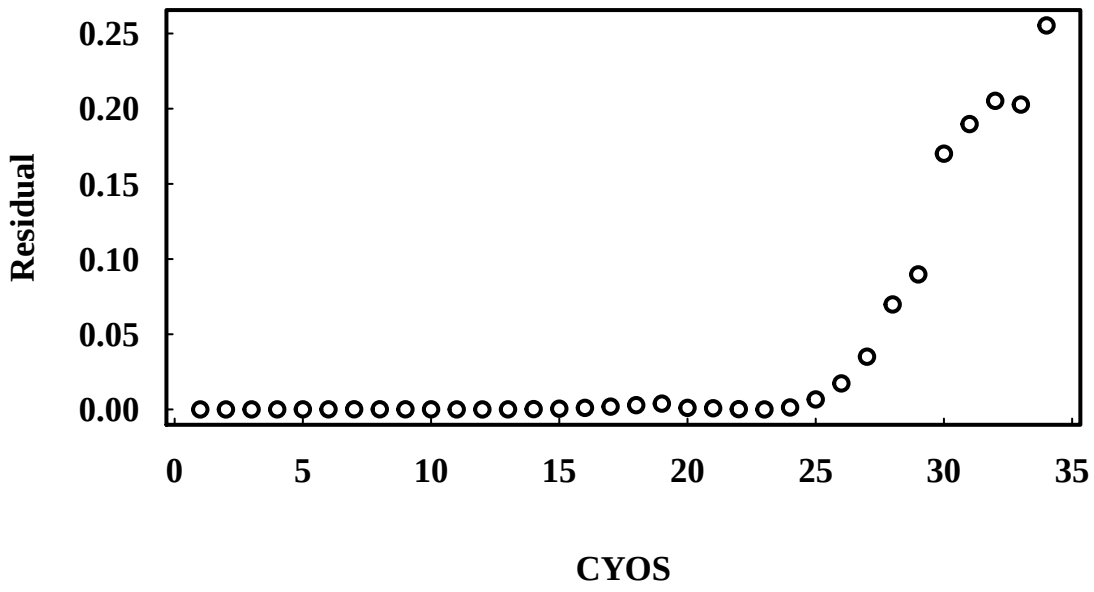


**Figure 17.** Residual plot for Parametric model (CSO)

The natural cut-off for the CSO career field is at about 25 years looking at Figure 17. There seems to be a jump in residuals at the 20 year mark which makes sense because there is more than an average amount of people getting out once they hit the 20 year mark. The sharp increase in residuals at the 25 year mark could be because the number of officers getting out each year is different than the average. The model is sacrificing the accuracy in

the later years for the almost perfect accuracy up to year 25, so if the retention rates change at all, like they do after year 25, then the model will do a poor job predicting.

The SSE for the CSO data set is shown in Table 16.

**Table 16.** Sum of Squared Error for Parametric model

| SSE | <= 25 years |
|---|---|
| 1.26 | 0.02 |

Table 16 shows what we thought; the residuals are good for for years 1-25, but increase rapidly after year 25. The SSE of the total model is *1.256*, but the residuals for years 1-25 are 0.021.

### Pilot

The SSE table and residual plot can be found in Table 17 and Figure 18

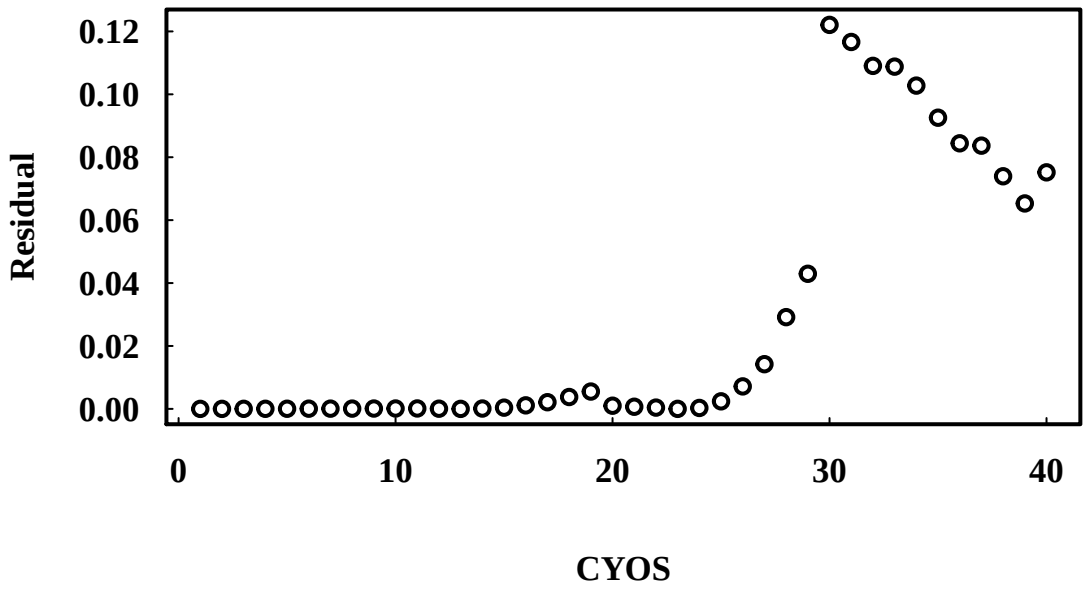


**Figure 18.** Residual plot for Parametric model (Pilot)

From Figure 18 the residuals once again peak at year 20, indicating officers leave the Air Force at year 20 in comparison to what is expected. From year 25 until year 40, the residuals act in a strange way. The residuals continually increase until year 30, and then begin to decrease until the data set ends at year 40. This shows that the model does well at representing the data until year 25, and then the model begins to do a poor job predicting the data until year 30, where the residuals peak at *0.12*. The residuals then decrease from year 30 to 40. At year 30 the model did the worst job at predicting the actual data.

**Table 17.** Sum of Squared Error for Parametric model

| SSE | <= 25 years |
|---|---|
| 1.15 | 0.02 |

For Table 17, once again the model does a good job at predicting the earlier data, but a bad job at predicting the later data. Since the model is following a distribution, it does not account well for data that does not follow the log-logistic regression, or at least resemble it. Although this occurs, the model still does a good job of predicting the data up to 25 years, with an SSE of 0.019.

The residual plots (Figures 16, 17, and 18) allow for the visualization of how well the model is predicting the data. For all three career fields the model does good at predicting the data for a set amount of time (27 for ABM, 25 for both CSO and pilot). The model seems to show favoritism for the earlier data, which makes sense because the earlier data has more data points. Of the 327130 data points used for pilot (including each pilot each year until their CYOS), 98.8% of the data is before year 25. The model fits the data before year 25 well because there is so few data points after year 25 that the earlier group requires all of the attention. This is why when looking in Figure 15, the gompertz distribution looks like it fits the later data much better than that of the log-logistic, but because of how well the log-logistic fits 98.8% of the data, it is the best distribution for fitting the data.

# V. Conclusion

Parametric models allow for future survival models to be established without any data. Once the distributional parameters are established for a career field, those parameters can create new survival curves without anymore data. Shown by their AICs, parametric survival analysis is better at modeling the data provided by HAF/A1 than semiparametric survival analysis. We recommend using parametric survival analysis in place of semiparametric survival analysis when trying to predict attrition rates of rated officers in the Air Force.

## 5.1 Limitations of Work

The goal of this work is to provide insight to personnel management officers for a specific subpopulation of the United States Air Force; officers in the career fields of ABM, CSO, and pilot. Each survival function applies directly to the respective career field, and although the parametric distribution chosen is the same, the parameters differ for each career field. These parametric distributions cannot be applied to other career fields because, as we even saw in this data, every career field has different trends of attrition rates.

## 5.2 Follow-On Research

Recommended additional research could include conducting regression and survival analysis on other Air Force subpopulation to determine if other subpopulations' survival follows a distributional fit, and if it does, if it is the log-logistic. Within this data, the career fields chosen could be broken down even more. There are numerous types of aircraft, and the pilots all retain differently.

Another follow-on study could be conducted with the data after the new Blended Retirement System goes into place in 2018. This new retirement system gives benefits to people that get out of the Air Force before the traditional 20-year mark, which may cause less people to stay in for 20 years.

Additionally, future work could include the collection of different, time series data points by HAF/A1 to allow for the better prediction of retention. In our data set, an officer was either single, married, or divorced/widowed, but these things can change throughout an officer's career. An officer can be single for 4 years, married for 8 years, and then divorced for 2 years before they decide to retire. This information would allow for a more accurate representation of the categorical variables used, and probably a better predictive accuracy to the attrition.

Finally, the work done with using Years of Service as an independent variable showed promise, but not enough time permitted to complete this delve into a different modeling technique. There is a reason the plots look like they do not model the data well but have a low AIC, but this was not analyzed in depth and could be in the future.

# Appendix A

## 5.3 Data Example

| | IDNumber | Year | SEX | MARITALSTAT | DEPENDS | SOC | DG | PRIORSVC | YOS | Status |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 12 | 1 | M | 1 | 1 | 3 | 0 | 0 | 4 | 0 |
| 2 | 12 | 2 | M | 1 | 1 | 3 | 0 | 0 | 4 | 0 |
| 3 | 12 | 3 | M | 1 | 1 | 3 | 0 | 0 | 4 | 0 |
| 4 | 12 | 4 | M | 1 | 1 | 3 | 0 | 0 | 4 | 1 |
| 5 | 27 | 1 | M | 0 | 0 | 4 | 0 | 0 | 10 | 0 |
| 6 | 27 | 2 | M | 0 | 0 | 4 | 0 | 0 | 10 | 0 |
| 7 | 27 | 3 | M | 0 | 0 | 4 | 0 | 0 | 10 | 0 |
| 8 | 27 | 4 | M | 0 | 0 | 4 | 0 | 0 | 10 | 0 |
| 9 | 27 | 5 | M | 0 | 0 | 4 | 0 | 0 | 10 | 0 |
| 10 | 27 | 6 | M | 0 | 0 | 4 | 0 | 0 | 10 | 0 |
| 11 | 27 | 7 | M | 0 | 0 | 4 | 0 | 0 | 10 | 0 |
| 12 | 27 | 8 | M | 0 | 0 | 4 | 0 | 0 | 10 | 0 |
| 13 | 27 | 9 | M | 0 | 0 | 4 | 0 | 0 | 10 | 0 |
| 14 | 27 | 10 | M | 0 | 0 | 4 | 0 | 0 | 10 | 1 |

## 5.4 Marital Status

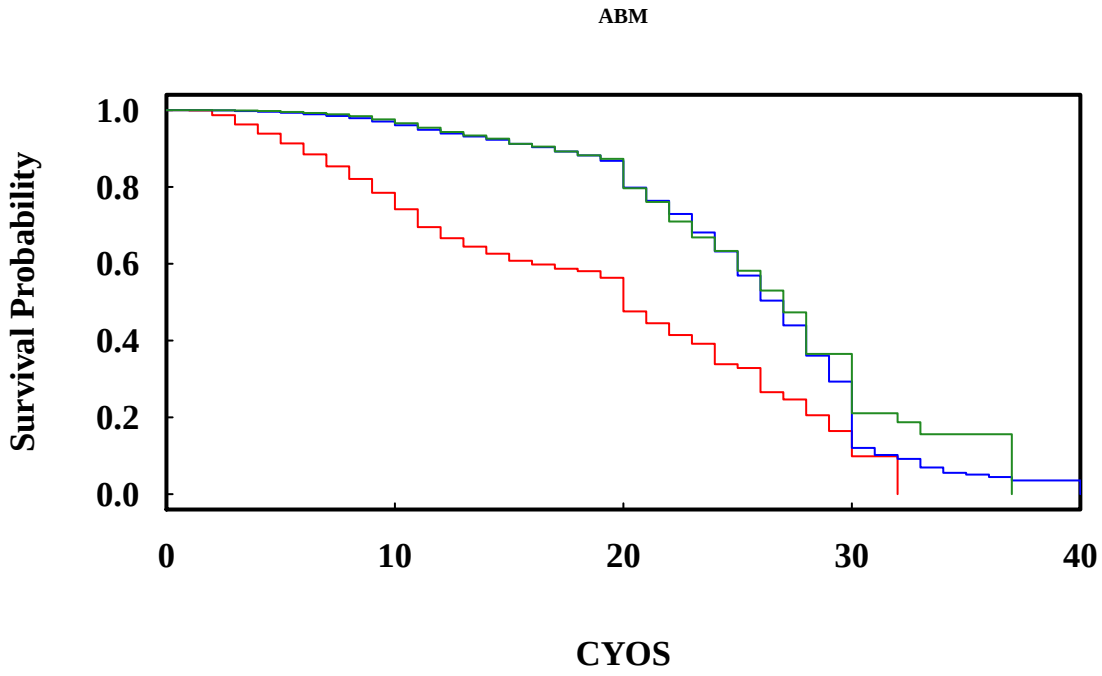


**Figure 19.** Kaplain-Meier plots (Marital Status)

## 5.5 Prior Service

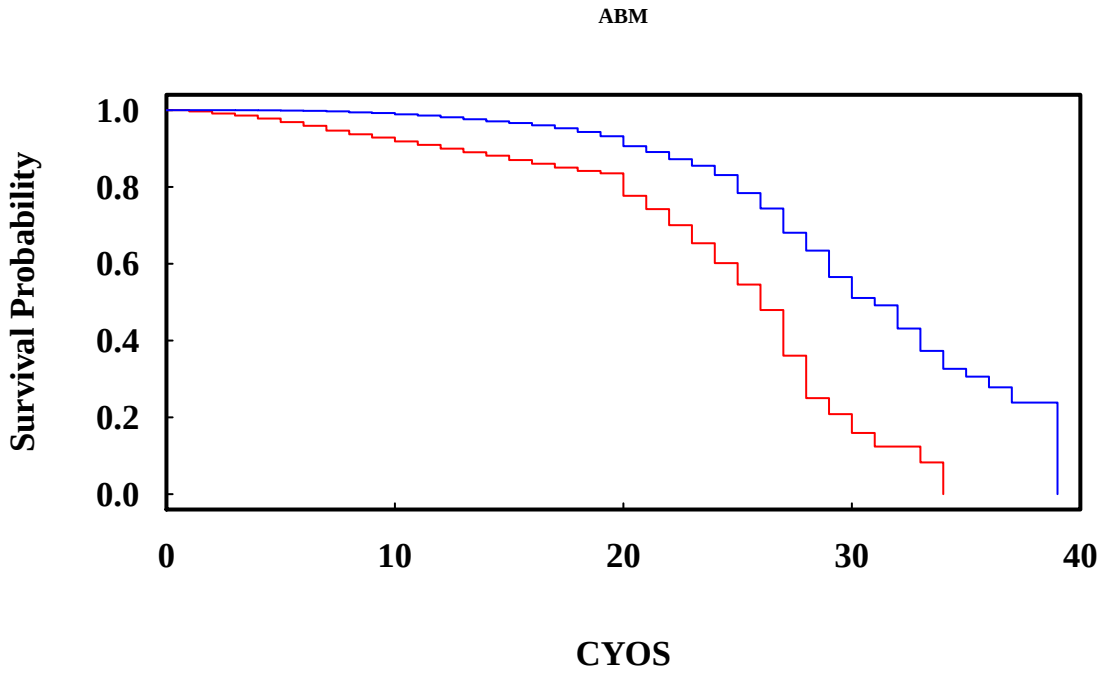


**Figure 20.** Kaplain-Meier plots (Prior Service)

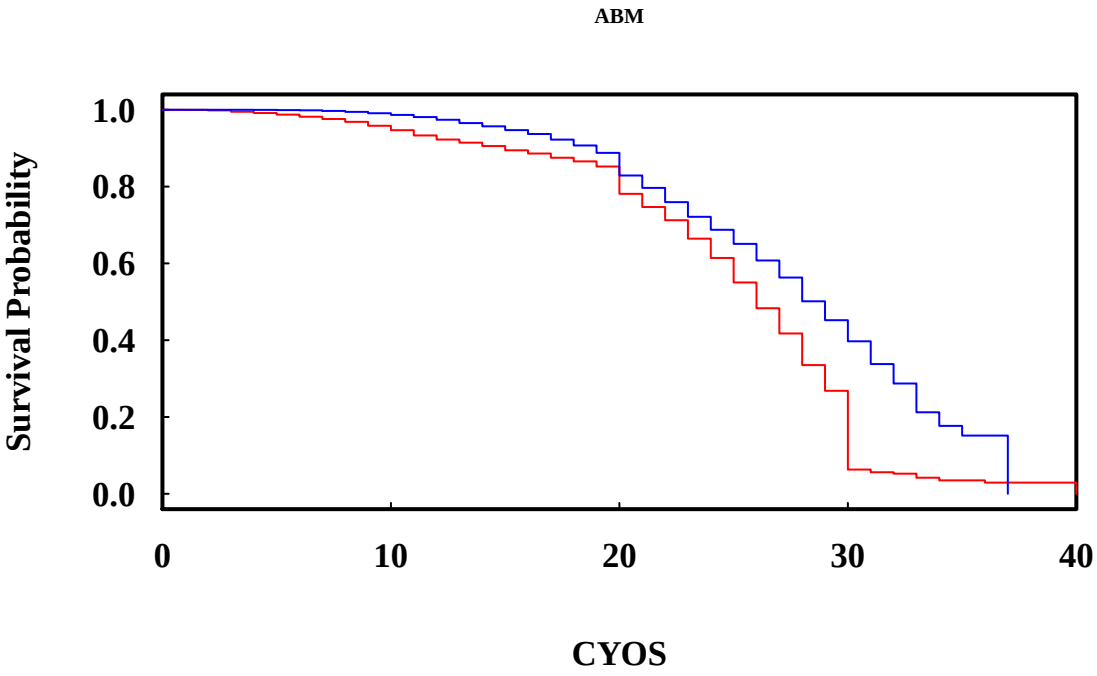


**Figure 21.** Kaplain-Meier plots (Prior Service)

# Appendix B

```
# Use ABM as example dataset.
# Code is the same for CSO and Pilot

## Packages

library(sas7bdat)
library(dplyr)
library(survival)
library(flexsurv)
library(knitr)
library(eha)
library(data.table)
library(kableExtra)
library(formattable)
library(survminer)



## Load the Data


abmall20<-read.sas7bdat("data/ABM/abmall20.sas7bdat",
```

```
                          debug=FALSE)
abm0to6 <-read.sas7bdat("data/ABM/abm0to6.sas7bdat",
                          debug=FALSE)
abm4to8 <-read.sas7bdat("data/ABM/abm4to8.sas7bdat",
                          debug=FALSE)
abm8to14 <-read.sas7bdat("data/ABM/abm8to14.sas7bdat",
                           debug=FALSE)
abm12to19 <-read.sas7bdat("data/ABM/abm12to19.sas7bdat",
                            debug=FALSE)
abm20to22 <-read.sas7bdat("data/ABM/abm20to22.sas7bdat",
                            debug=FALSE)
abm <-read.sas7bdat("data/ABM/abm.sas7bdat",
                      debug=FALSE)



# Data Cleaning




# If using other file than abm, replace abm with data file name
load("data/ABM/abm.RData")
attach(abm)

# abmshort takes only the necessary variables
abmshort <- subset(abm, select=c("SSAN", "YOS_EOP", "SEX",
                                 "MARITALSTAT", "DEPENDS",
                                 "SOC", "DG", "PRIORSVC"))
```

```r
year <- c("Years of Service")
abmshort[,year] <- NA
abmshort$`Years of Service` <- abmshort$YOS_EOP
colnames(abmshort)[2] <- "Year"
colnames(abmshort)[9] <- "YOS" # configure data frame for liking

# delete duplicates
abmshort <- abmshort[!duplicated(abmshort), ]

# SOC has non-numeric variables,
# so turns non-numeric to NA
abmshort$SOC <- as.numeric(as.character(abmshort$SOC))

# some variables have DG = "NaN",
# throw out individuals without a DG

colSums(is.na(abmshort))
nandg <- with(abmshort, which(DG=="NaN", arr.ind = TRUE))
abmshort <- abmshort[-nandg, ]

# abmshortexp makes a new datapoint per year
abmshortexp <- abmshort[rep(row.names(abmshort),
                            abmshort$Year),] %>%
    group_by(SSAN) %>%
    mutate(Year = 1:n()) # loops the year for each ssan

status <- c("Status")
```

```
abmshortexp[,status] <- NA # creates new blank variable



# make variables factors
abmshortexp$DG <- as.factor(abmshortexp$DG)
abmshortexp$PRIORSVC <- as.factor(abmshortexp$PRIORSVC)
abmshortexp$Status <- as.factor(abmshortexp$Status)
abmshortexp$MARITALSTAT <- as.factor(abmshortexp$MARITALSTAT)
abmshortexp$DEPENDS <- as.factor(abmshortexp$DEPENDS)
abmshortexp$SOC <- as.factor(abmshortexp$SOC)

# make status = 1 when person retires
abmshortexp$Status <- ifelse(abmshortexp$Year ==
                               abmshortexp$YOS,1,0)

AICcompare <- matrix(ncol = 2, nrow = 2) # put the AICs into here for compa
colnames(AICcompare) <- c("Type of SA", "AIC")



## Nonparametric Survival Analysis



attach(abmshortexp)
S <- Surv(Year,  Status)
abmshortexp2 <- abmshortexp
abmshortexp2$SurvObj <- with(abmshortexp,
                             Surv(Year, Status == 1))
```

```
res.km <- survfit(Surv(Year, Status) ~ SEX,
                  data = abmshortexp2)

plot(res.km, xlab = "Commissioned Whole Years of Service",
     ylab = "Survival Probability")



## Semiparametric



attach(abmshortexp)
S <- Surv(Year, Status)
res.cox <- coxph(data = abmshortexp, S ~
                   SEX + MARITALSTAT + DEPENDS +
                   SOC + DG + PRIORSVC)
#summary(res.cox)
AICcompare[1,1] <- "Semiparametric"
AICcompare[1,2] <- round(AIC(res.cox), 3)
res.coxsurv <- survfit(res.cox)
plot(res.coxsurv)



## Parametric

Dist <- c("exp", "lnorm", "gompertz",
          "gengamma", "gengamma.orig", "genf.orig",
```

```
          "llogis", "weibull")
AIC <- matrix(ncol = 2, nrow = 8)

attach(abmshortexp)
S <- Surv(Year, Status)

for(i in 1:nrow(AIC)){
  AIC[i,1] <- Dist[i]
  model <- flexsurvreg(S ~ SEX + MARITALSTAT + DEPENDS +
                         SOC + DG + PRIORSVC, dist=Dist[i],
                       method = "Nelder-Mead")
                        # gamma, genf, weibull?
                          #weibull did not work for
  AIC[i,2] <- AIC(model)

}
colnames(AIC) <- c("Distribution", "AIC")
AIC <- transform(AIC, Distribution =
                   as.character(Distribution),
                 AIC = as.factor(AIC))
#### clean up AIC
AIC$AIC <- as.numeric(levels(AIC$AIC)[AIC$AIC])
AIC$AIC <- round(AIC$AIC,2)
AIC <- AIC[order(AIC$AIC),]
rownames(AIC) <- c()

if (AIC[1,2] < 0) {
```

```
   Disti <- AIC[2, 1]
} else {
   Disti <- AIC[1, 1]
}

S <- Surv(Year, Status)
Disti <- "llogis"
model5 <- flexsurvreg(S ~ SEX + MARITALSTAT + DEPENDS +
                          SOC + DG + PRIORSVC,
                      dist=Disti, data = abmshortexp,
                      method = "Nelder-Mead")

AICcompare[2,1] <- Disti
AICcompare[2,2] <- round(AIC(model5),3) # AIC compares the two
plot(model5, ylab="Survival Probability",
     xlab="Years", main = NULL,
     cex.lab = 0.8, cex.axis = 0.8, cex.main=0.75)
legend("topright",legend=c("KM Plot","Fitted"),
       lty=c(1,1),
       col=c("black","red"), cex=0.75)

colnames(AICcompare) <- c("Distribution", "AICcompare")
AICcompare <- transform(AICcompare, Distribution =
                            as.character(Distribution),
                  AICcompare = as.factor(AICcompare))
#### clean up AICcompare
AICcompare$AICcompare <- as.numeric(levels(AICcompare$AICcompare)
```

```
                                        [AICcompare$AICcompare])
AICcompare$AICcompare <- round(AICcompare$AICcompare,2)
AICcompare <- AICcompare[order(AICcompare$AICcompare),]
rownames(AICcompare) <- c()



# PH paramerization

Expph <- phreg(formula = Surv(Year, Status) ~
                 SEX + MARITALSTAT + DEPENDS +
                 SOC + DG + PRIORSVC,
               data = abmshortexp)
summary(Expph)

# AIC compare

abmAIC <- readRDS(file =
                    "data/ABM/abmAIC.rds")
abm0to6AIC <- readRDS(file =
                      "data/ABM/abm0to6AIC.rds")
abm4to8AIC <- readRDS(file =
                        "data/ABM/abm4to8AIC.rds")
abm8to14AIC <- readRDS(file =
                         "data/ABM/abm8to14AIC.rds")
abm12to19AIC <- readRDS(file =
                          "data/ABM/abm12to19AIC.rds")
abm20to22AIC <- readRDS(file =
```

```
                              "data/ABM/abm20to22AIC.rds")
abmall20AIC <- readRDS(file =
                             "data/ABM/abmall20AIC.rds")



abmAICcompare <- readRDS(file =
                               "data/ABM/abmAICcompare.rds")
abm0to6AICcompare <- readRDS(file =
                                  "data/ABM/abm0to6AICcompare.rds")
abm4to8AICcompare <- readRDS(file =
                                  "data/ABM/abm4to8AICcompare.rds")
abm8to14AICcompare <- readRDS(file =
                                   "data/ABM/abm8to14AICcompare.rds")
abm12to19AICcompare <- readRDS(file =
                                    "data/ABM/abm12to19AICcompare.rds")
abm20to22AICcompare <- readRDS(file =
                                    "data/ABM/abm20to22AICcompare.rds")
abmall20AICcompare <- readRDS(file =
                                   "data/ABM/abmall20AICcompare.rds")



#### clean up AICcompare
colnames(abmAICcompare) <- c("Distribution", "AIC")
abmAICcompare <- transform(abmAICcompare, Distribution =
                               as.character(Distribution),
                  AIC = as.factor(AIC))
```

```
abmAICcompare$AIC <- as.numeric(levels(abmAICcompare$AIC)
                                  [abmAICcompare$AIC])
abmAICcompare$AIC <- round(abmAICcompare$AIC,2)
rownames(abmAICcompare) <- c()



abmAICall <- rbind(abmAIC, abmAICcompare)
# delete duplicates
abmAICall <- abmAICall[!duplicated(abmAICall), ]
abmAICall <- abmAICall[order(abmAICall$AIC),]
abmAICall
##saveRDS(abmAICall, "data/abm/abmAICall.rds")



abmAICall <- readRDS(file =
                       "data/ABM/abmAICall.rds")
abm0to6AICall <- readRDS(file =
                           "data/ABM/abm0to6AICall.rds")
abm4to8AICall <- readRDS(file =
                           "data/ABM/abm4to8AICall.rds")
abm8to14AICall <- readRDS(file =
                            "data/ABM/abm8to14AICall.rds")
abm12to19AICall <- readRDS(file =
                             "data/ABM/abm12to19AICall.rds")
abm20to22AICall <- readRDS(file =
                             "data/ABM/abm20to22AICall.rds")
# abmall20AICall <- readRDS(file =
```

```
"data/ABM/abmall20AICall.rds")

Dist <- c("gompertz", "llogis", "gengamma.orig", "gengamma",
          "lnorm", "genf.orig", "exp", "weibull",
          "gamma", "genf", "Semiparametric")

colnames(abm0to6AICall) <- c("Distribution", "AIC0to6")
colnames(abm4to8AICall) <- c("Distribution", "AIC4to8")
colnames(abm8to14AICall) <- c("Distribution", "AIC8to14")
colnames(abm12to19AICall) <- c("Distribution", "AIC12to19")
colnames(abm20to22AICall) <- c("Distribution", "AIC20to22")
colnames(abmAICall) <- c("Distribution", "AIC")



abmAICforall <- Reduce(function(...) merge(..., all=TRUE),
       list(abm0to6AICall, abm4to8AICall, abm8to14AICall,
            abm12to19AICall, abm20to22AICall, abmAICall))
abmAICforall <- rbind(abmAICforall, AICextra)
abmAICforall[1,1] <- "Exponential"
abmAICforall[2,1] <- "Generalized F (original)"
abmAICforall[3,1] <- "Generalized Gamma (stable)"
abmAICforall[4,1] <- "Generalized Gamma (original)"
abmAICforall[5,1] <- "Gompertz"
abmAICforall[6,1] <- "Log-logistic"
abmAICforall[7,1] <- "Log Normal"
abmAICforall[8,1] <- "Semiparametric"
```

```
abmAICforall[9,1] <- "Weibull"
abmAICforall[10,1] <- "Gamma"
abmAICforall[11,1] <- "Generalized F (stable)"

abmAICforall <- abmAICforall[c(1:7,9:11, 8), ]
#abmAICforall <- abmAICforall[, -c(7)]

#abmAICforall <- abmAICforall[order(abmAICforall$Distribution),]


saveRDS(abmAICforall, "data/ABM/abmAICforall.rds")

# abmAICforall <- readRDS(file = "data/ABM/abmAICforall.rds")
# readRDS(file = "data/ABM/abmAICforall.rds")


#AICextra


colnames(AICextra) <- c("Distribution", "AIC0to6", "AIC4to8",
                        "AIC8to14", "AIC12to19", "AIC20to22", "AIC")
AICextra <- transform(AICextra, Distribution = as.character(Distribution),
                  AIC0to6 = as.factor(AIC0to6), AIC4to8 =
                    as.factor(AIC4to8), AIC8to14 = as.factor(AIC8to14),
                  AIC12to19 = as.factor(AIC12to19), AIC20to22 =
                    as.factor(AIC20to22), AIC = as.factor(AIC))
#### clean up AIC
```

```
AICextra$AIC0to6 <- as.numeric(levels(AICextra$AIC0to6)
                                [AICextra$AIC0to6])
AICextra$AIC4to8 <- as.numeric(levels(AICextra$AIC4to8)
                                [AICextra$AIC4to8])
AICextra$AIC8to14 <- as.numeric(levels(AICextra$AIC8to14)
                                 [AICextra$AIC8to14])
AICextra$AIC12to19 <- as.numeric(levels(AICextra$AIC12to19)
                                  [AICextra$AIC12to19])
AICextra$AIC20to22 <- as.numeric(levels(AICextra$AIC20to22)
                                  [AICextra$AIC20to22])
AICextra$AIC <- as.numeric(levels(AICextra$AIC)
                           [AICextra$AIC])
AICextra$AIC0to6 <- round(AICextra$AIC0to6,2)
AICextra$AIC4to8 <- round(AICextra$AIC4to8,2)
AICextra$AIC8to14 <- round(AICextra$AIC8to14,2)
AICextra$AIC12to19 <- round(AICextra$AIC12to19,2)
AICextra$AIC20to22 <- round(AICextra$AIC20to22,2)
AICextra$AIC <- round(AICextra$AIC,2)
rownames(AIC) <- c()


# Compare models


abm0to6semimodel <- readRDS(file =
                              "data/ABM/abm0to6semimodel.rds")
abm0to6paramodel <- readRDS(file =
```

```
                                "data/ABM/abm0to6paramodel.rds")
abm4to8semimodel <- readRDS(file =
                                "data/ABM/abm4to8semimodel.rds")
abm4to8paramodel <- readRDS(file =
                                "data/ABM/abm4to8paramodel.rds")
abm8to14semimodel <- readRDS(file =
                                 "data/ABM/abm8to14semimodel.rds")
abm8to14paramodel <- readRDS(file =
                                 "data/ABM/abm8to14paramodel.rds")
abm12to19semimodel <- readRDS(file =
                                  "data/ABM/abm12to19semimodel.rds")
abm12to19paramodel <- readRDS(file =
                                  "data/ABM/abm12to19paramodel.rds")
abm20to22semimodel <- readRDS(file =
                                  "data/ABM/abm20to22semimodel.rds")
abm20to22paramodel <- readRDS(file =
                                  "data/ABM/abm20to22paramodel.rds")
abmsemimodel <- readRDS(file =
                            "data/ABM/abmsemimodel.rds")
abmparamodel <- readRDS(file =
                            "data/ABM/abmparamodel.rds")


plot(abmparamodel, ylab="Survival Probability",
     xlab="Years", main = NULL,
    cex.lab = 0.8, cex.axis = 0.8, cex.main=0.75)
legend("topright",legend=c("KM Plot","Fitted"),
       lty=c(1,1),col=c("black","red"), cex=0.75)
```

# Bibliography


[1] Bernard Rostker. I want you! the evolution of the all-volunteer force. Technical report, Santa Monica, CA: RAND Corporation, 2006.

[2] Steve Buyer and Neil Abercrombie. Actions needed tobetter define pilotrequirements andpromote retention. Technical report, United States General Accounting Office, 1999.

[3] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis, 4th edition*. John Wiley & Sons, 2015.

[4] David G Kleinbaum, Mitchel Klein, and ER Pryor. *Logistic regression: a self-learning text*. Springer, New York, USA, 2002.

[5] W David Kelton, Jeffrey S Smith, and David T Sturrock. *Simio & simulation: Modeling, analysis, applications*. Learning Solutions, 2011.

[6] Microsoft excel, 2016. URL https://www.microsoft.com/en-us/store/d/office-365-personal/.

[7] Arena, 2016. URL https://www.arenasimulation.com/.

[8] Simio: Simulation, production planning and scheduling software, 2017. URL https://www.simio.com/.

[9] Jerry Banks, II Carson, Barry L Nelson, and David M Nicol. *Discrete-event system simulation, 5th edition*. Pearson, 2005.

[10] Eric Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, 99(suppl 3):7280–7287, 2002.

[11] Goran Brostrom. *Event History Analysis with R*. CRC Press, 2012.

[12] Odd Aalen, Ornulf Borgan, and Hakon Gjessing. *Survival and event history analysis: a process point of view*. Springer Science & Business Media, 2008.

[13] Rupert G Miller Jr. *Survival analysis*, volume 66. John Wiley & Sons, 2011.

[14] Melinda Mills. *Introducing survival and event history analysis*. Sage Publications, 2011.

[15] Youngkyoung Min, Guili Zhang, Russell A. Long, Timothy J. Anderson, and Matthew W. Ohland. Nonparametric survival analysis of the loss rate of undergraduate engineering students. *Journal of Engineering Education*, 100(2):349–373, 2011. ISSN 2168-9830. doi: 10.1002/j.2168-9830.2011.tb00017.x. URL http://dx.doi.org/10.1002/j.2168-9830.2011.tb00017.x.

[16] John T Warner. *The economics of military manpower*. Elsevier Science, 1995.

[17] Saul I Gass. Military manpower planning models. *Computers & Operations Research*, 18(1):65–73, 1991.

[18] Roger W Collins, Saul I Gass, and Edward E Rosendahl. The ascar model for evaluating military manpower policy. *Interfaces*, 13(3):44–53, 1983.

[19] Dave Cashbaugh. Managing retention use of simulation and optimization. Technical report, Navy Personnel Research Studies and Technology Millington TN, 2010.

[20] John Capon, Oleksandr S Chernyshenko, and Stephen Stark. Applicability of civilian retention theory in the new zealand military. *New Zealand Journal of Psychology*, 36 (1):50, 2007.

[21] Sanford C Orrick. *Forecasting Marine Corps enlisted losses*. PhD thesis, Monterey, California. Naval Postgraduate School, 2008.

[22] Jamie T Zimmerman. *Application of Enlisted Force Retention Levels and Career Field Stability*. Master's thesis, Air Force Institute of Technology Wright-Patterson AFB OH Graduate School of Engineering and Management, 2017.

[23] Jill A Schofield. *Non-Rated Air Force Line Officer Attrition Rates Using Survival Analysis*. Master's thesis, Air Force Institute of Technology Wright-Patterson AFB OH Graduate School of Engineering and Management, 2015.

[24] Christine L Zens. *Application of Non-Rated Line Officer Attrition Levels and Career Field Stability*. Master's thesis, Air Force Institute of Technology Wright-Patterson AFB OH Graduate School of Engineering and Management, 2016.

[25] Courtney N Franzen. *Survival Analysis of US Air Force Rated Officer Retention*. Master's thesis, Air Force Institute of Technology Wright-Patterson AFB OH Graduate School of Engineering and Management, 2017.

[26] German Rodriguez. Parametric survival models. *Princeton University*, 2010. URL http://data.princeton.edu/pop509/ParametricSurvival.pdf.

[27] Jiezhi Qi. Comparison of proportional hazards and accelerated failure time models. Technical report, University of Saskatchewan, 2009.

[28] Patrick Breheny. Accelerated failure time models. University Lecture, 2015.

[29] Mohamad Amin Pourhoseingholi, Azadeh Safaee, and Alireza Abadi. Comparing cox regression and parametric models for survival of patients with gastric carcinoma. *Asian Pacific Journal of Cance Prevention*, pages 412 – 416, 2007.

| REPORT DOCUMENTATION PAGE | Form Approved OMB No. 0704-0188 |
|---|---|



| 1. REPORT DATE (DD-MM-YYYY) | 2. REPORT TYPE | 3. DATES COVERED (From - To) |
|---|---|---|
| 22-03-2018 | Thesis | October 2016-March 2018 |

**4. TITLE AND SUBTITLE**
PARAMETRIC SURVIVAL ANALYSIS OF US AIR FORCE RATED OFFICER RETENTION

**5a. CONTRACT NUMBER**

**5b. GRANT NUMBER**

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**
Lindell, Jacob, R, 2nd Lt, USAF

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**
AFIT-ENS-MS-18-M-136

**7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(ES)**
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/Department of Operational Sciences)
2950 Hobson Way, Building 640
WPAFB OH 45433-7201

**8. PERFORMING ORGANIZATION REPORT NUMBER**
AFIT-ENS-MS-18-M-136

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
AF/A1PF
Dr. Gerald Diaz
1500 W Perimeter Rd JB Andrews, MD 20762
gerald.diaz.civ@mail.mil

**10. SPONSOR/MONITOR ACRONYM(S)**
AF/A1PF

**11. SPONSOR/MONITOR REPORT NUMBER(S)**

**12. DISTRIBUTION/AVAILABILITY STATEMENT**
Distribution Statement A. Approved for Public Release; Distribution Unlimited

**13. SUPPLEMENTARY NOTES**
This material is declared a work of the
U.S. Government and is not subject to copyright protection in the United States

**14. ABSTRACT**
Personnel retention is a very important topic in both the private and public sectors. Not only do companies need to make sure they have the right people, they need to have the right amount of people. Within the public sector, specifically the US Air Force, maintaining appropriate manning comes in two phases; bringing in the right amount of people each year, and retaining enough people from year to year. These two aspects go hand in hand; if the Air Force knows how many people they will lose in a given year, they can bring in the exact number of people they need to make sure they maintain their end strength requirements. As part of the effort to ensure proper military accessions, the Air Force uses retention models to assist in predicting the future retention patterns. Not only does the Air Force want to make sure they meet their end strength requirements, they want to make sure they bring in the correct amount of people to each career field. The career fields, Air Force Specialty Codes (AFSCs), have a personnel requirement each year in order to accomplish that AFSCs mission. In this study, semiparametric survival analysis was used to determine the

**15. SUBJECT TERMS**

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | Dr. Seong-Jong Joo |
| U | U | U | UU | 85 | 19b. PHONE NUMBER (Include area code) (937)255-3636 x7469 seong-jong.joo@afit.edu |