3-2021

An Examination of Civilian Retention in the United States Air Force

William F. Wilson

AN EXAMINATION OF CIVILIAN RETENTION IN THE UNITED STATES AIR FORCE

THESIS

William F. Wilson, Captain, USAF

AFIT-ENS-MS-21-M-195

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENS-MS-21-M-195

AN EXAMINATION OF CIVILIAN RETENTION IN THE UNITED STATES AIR FORCE

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

William F. Wilson, B.S.

Captain, USAF

March 2021

AFIT-ENS-MS-21-M-195

AN EXAMINATION OF CIVILIAN RETENTION IN THE UNITED STATES AIR FORCE

THESIS

William F. Wilson, B.S.
Captain, USAF

Committee Membership:

Dr. Raymond R. Hill
Chair

Dr. Bruce A. Cox
Member

Abstract

The backbone of the United States Air Force is undoubtedly the large civilian workforce that supplements the great work that is accomplished. Many research studies have been conducted on officer and enlisted personnel to ensure that the career fields are properly developed and managed to meet the ever growing demands of the military's varied missions, but no recent studies have focused on the civilian workforce. Striking a balance between new and experienced employees is paramount to success given the ever-changing economic and political landscapes where we find ourselves.

The first part of the research uses logistic regression to determine the factors that are important for retention in the civilian workforce over the last ten years (2010-2019). The six variables analyzed were age, gender, race, education level, prior military status, and years of service; all six were significant. Further breakdowns showed differences between the occupational series and between white-collar and laborer positions. Odds ratios indicate the disparity between having a certain qualification or not.

The second part of the study uses survival analysis in the form of Kaplan-Meier survival curves and a Cox proportional hazards model to create unique survival curves that display the probability of remaining in employment given the number of years of service for a particular group. Future personnel management decisions can be enhanced using these curves as a basis for understanding the recent retention trends of the civilian workforce.

This research is dedicated to my loving wife, for all of the motivation and encouragement to see things through.

Table of Contents

Page

List of Figures

List of Tables

AN EXAMINATION OF CIVILIAN RETENTION IN THE UNITED STATES AIR FORCE

I. Introduction

1.1 Problem Background

The United States Air Force employs a workforce of over 170,000 civilian employees across many disciplines and locations [1] [2]. This workforce is responsible for a wide range of activities, such as maintaining military air superiority, assisting military efforts to combat terrorism, and even make sure paychecks are released on time. As such, the civilian workforce may be viewed as a linchpin that solidifies the Air Force as the powerhouse that it is. It is vital to the nation that this workforce be maintained and monitored to ensure that there is a sufficient number of employees to produce outstanding work quality. Military officer personnel levels are maintained by the Defense Officer Personnel Management Act (DOPMA) [3]. Unlike military positions, civilian employee positions do not have a congressional constraint and are only limited by the number of billets available and the financial situation of an organization to support the position. Civilian employees are also not constrained to Active Duty Service Commitments (ADSC) and as such have the ability to leave or transfer to another government agency when they please making workforce planning far into the future more difficult.

1.2 Research Scope

The purpose of the research is to demonstrate a method to hel in the tracking and planning for the civilian workforce. This is achieved through the use of logistic regression and survival analysis. The Air Force does not currently model civilian retention this way. The scope of this research is to find warning indicators that can alert senior leaders to individuals or groups of people that are likely to have the disposition to leave employment with the Air Force. Once these people or groups have been identified, proper incentives can be given, or planning undertaken, to mitigate the turnover time to fill the empty position.

1.3 Issues, Needs, and Limitations

Headquarters Air Force Directorate of Personnel (HAF/A1) provided the extracts from the Air Force's personnel database, the Military Personnel Data System (MilPDS). This database contains all of the personal information of the civilian employees. The data used in this analysis spans the period from January 2010 to December 2019. Despite the efforts of many people maintaining the database by constantly updating and correcting the information, the data are bound to have a few mistakes but it is assumed that these errors do not make up a sizeable portion of the data and as such would not affect the results.

The purpose of this research is to provide a reproducible product that monitors the civilian workforce for warning signs utilizing software that models civilian workforce retention as a function of various covariates. To make the end-product sustainable and reproducible, a self-imposed limitation was to only utilize SAS and Excel because most analysts in HAF/A1 have access to these two programs. SAS is mostly used by the personnel analysts for data mining and collecting summary statistics that are exported to Excel for formatting into charts and graphs. However, SAS is a very

powerful analytical tool that can be used to automate many complex algorithms and perform detailed analysis on very large data sets.

1.4 Thesis Outline

Chapter 2 reviews methodologies and research that has been conducted on past retention studies both in the civilian sector and in multiple branches of the military. Chapter 3 describes the data source, MilPDS, and the extracts provided for the analysis. Chapter 4 discusses the logistic regression model developed and the results obtained. Chapter 5 shows the application of the Kaplan-Meier survival analysis and Cox proportional hazards model. Chapter 6 details the limitations of the research and recommendations for follow-on research.

II. Background and Literature Review

Many studies have been conducted on the retention trends of military personnel, but few have concentrated on government civilian personnel. However, that prior research is invaluable for examining the methodologies used. The review presented here examines these past works.

2.1 Modeling Techniques

A variety of techniques have been used to study both military and public civilian personnel management. Some of the more common techniques have been regression analysis, logistic regression, survival analysis, and simulation.

"Regression analysis is a statistical technique for investigating and modeling the relationship between variables" [4]. The response variable y is plotted in n-dimensional space and a linear equation is fit to the data to minimize the deviation from the plotted points to the estimated hyperplane. Many high and low order equations can be fit to the data in the hopes of estimating the true relationship of the response to the regressors. For manpower analysis, this might entail attempting to find a relationship between the total number of years employed and the attributes of the employees. Typical attributes include age, sex, race, education level, previous work experience, etc.

While multiple linear regression is primarily concerned with a quantitative response, logistic regression can handle qualitative responses. This technique is primarily used for binary responses but can be altered for responses with more than two levels [5]. An example of logistic regression might seek to classify a person as risky or not risky for a bank loan based on their past loan payments, credit history, and other related attributes. Logistic regression is a generalized version of linear regression that

takes advantage of the logit function to transform the range of the original function to an infinite range [4]. The logit function maps the inputs from the regressors to the probability of being a certain response. Odds ratios obtained from the function allow the analyst to measure the estimated increase of the probability of success given a one-unit increase in the regressor [4].

Survival analysis aims to predict the time until an event occurs. The time between the start of the trial and the occurrence of the event of interest is labeled as the survival time. This technique is used extensively in health studies such as predicting the occurrence of a tumor reappearing or the death of a patient with a debilitating disease [6]. In the case of personnel management, the survival time would equate to the time between when a person enters and leaves employment. A curve is then constructed to represent the total number of survivors from time zero to the time of the last survivor "dying." Usually, a survival curve is constructed for all feasible combinations of regressors used in the analysis. "For example, in medical follow-up studies to determine the distribution of survival times after an operation, contact with some individuals will be lost before their death, and others will die from causes it is desired to exclude from consideration. Similarly, observation of the life of a vacuum tube may be ended by breakage of the tube, or a need to use the test facilities for other purposes. In both examples, incomplete observations may also result from a need to get out a report within a reasonable time" [7]. This causes a censoring problem that is overcome by using the Kaplan-Meier method to estimate the survival outcomes of the observations [7].

"A simulation is the imitation of the operation of a real-world process or system over time. Whether done by hand or on a computer, simulation involves the generation of an artificial history of a system and the observation of that artificial history to draw inferences concerning the operating characteristics of a system" [8]. Many simu-

lation software packages exist, each with their strengths and weaknesses. Depending on the complexity of the problem at hand, different software is required. Microsoft Excel, Java, and C++ can be leveraged to create simple models while more complex simulations may require specialized software products to successfully implement the simulation. Two commonly used software packages for simulating complex systems are Simio [9] and Arena [10]. Statistics such as interarrival times, the time between failures, the average number waiting to be served, and the number of entities that renege at service can all be easily captured using a dedicated simulation software package. Simulations can be deterministic or stochastic depending on the inclusion or exclusion of uncertainty via random variables. They can also be static or dynamic based on whether time is considered. Simulations can also be continuous or discrete depending on whether the model's state changes at discrete points or continuously through the running of the simulation [8].

Discrete-event simulations are arguably the most common type of simulation. "Discrete-event simulation is the modeling of systems in which the state variable only changes at a discrete set of points in time" [8]. A familiar example of this would be a checkout stand at a grocery store. The state variable in question might be the total number of people waiting to be serviced and the customers are the entities. Customers arrive at the checkout stand, wait to be serviced, and finally leave the grocery store. A simulation to monitor civilian personnel would attempt to simulate workers entering and leaving the workforce at set points in time. System performance metrics such as the number of workers, number of empty slots, and turnover time of employees would be worth collecting.

A more complicated but useful simulation tool is the application of agent-based modeling. "Agent-based modeling is a method for simulating the actions and interactions of autonomous individuals (the agents) in a network, with a view of assessing

their effects on the system as a whole. Agents may be people or animals or other entities that have agency, meaning that they are not passive, they actively make decisions, retain memory of past situations and decisions, and exhibit learning" [8]. Weimer, Miller, and Hill give a more concise definition: "An ABM (agent-based model) is a simulation framework, using primarily the discrete-event scheduling paradigm, where the entities within the simulation have a greater degree of autonomy in movement and decision making than generally found in simulation models" [11]. This type of advanced modeling is computer resource intensive and requires fine tuning of the agents' behavior and tolerances for certain activities.

System dynamics is a type of simulation that uses feedback loops to model the system instead of relying on cause-effect relationships. Stocks and flows models are used in system dynamics as a way of measuring quantities in a system over time. "Stocks and flows – the accumulation and dispersal of resources – are central to the dynamics of complex systems" [12]. A stocks and flows model for retention would use people as the "stock" and simulate the "flow" of people into and out of the system. It would incorporate feedback loops such as incentives to stay and policies and decisions that cause people to leave.

Chi-squared automatic interaction detection (CHAID) is an algorithm that predicts response behavior by dividing "a data set in exclusive and exhaustive segments that differ with respect to the response variable. The segments are defined by a tree structure of a number of independent variables, the predictors. To each segment of individuals, CHAID assigns a probability of response" [13]. Another commonly used algorithm related to CHAID that also uses a tree-like structure is classification and regression trees (CART). CHAID is used for problems with many categorical variables while CART is primarily used for problems with many continuous variables [13]. For a retention study, CHAID is preferred because most of the variables are categorical.

The leaf with the highest probability would denote the type of person least likely to end employment while the lowest probability leaf would denote the type of person with the highest probability to end employment. Certain targeted people or groups could then be incentivized through various means to not leave employment.

2.2 Previous Work Related to US Military Manpower

The examples defined here are by no means the definitive sources of knowledge on the subject of personnel modeling, but they give insight into the techniques that can be used to manage personnel.

Hall [14] employed survival analysis to model enlisted marine core retention. He used parametric models to form exhaustive subsets of the population. Some breakouts of the population were gender, race, and occupational field. Five parametric models were used to analyze the enlisted personnel: Exponential, Weibull, Gompertz, Log-Logistic, and Log-Normal. Residual plot and Akaike Information Criteria (AIC) analysis showed the Gompertz model to be superior. Models for various occupational fields, the Air Force equivalent of an AFSC, were conducted. The results showed that each breakout's behavior was unique and thus Hall concluded that they should be modeled separately instead of combined.

Schofield [15] used logistic regression and survival analysis to observe the attrition behavior of non-rated line officers from various career fields. Acquisitions, Logistics, Non-Rated Operations, and Support classifications were looked at with each of these being further broken down into the contained subpopulations. Six demographic variables were utilized in the analysis: 1) commissioning yeargroup, 2) gender, 3) source of commission, 4) number of years enlisted, 5) career field grouping, and 6) distinguished graduate at commissioning source. Yeargroup is determined by the year the officer was commissioned and career field grouping is based on the first digit of the

officer's core AFSC. Odds ratios of retention were found using each of these variables in the logistic regression analysis.

A parametric form for the survival analysis was needed because all of the variables used were categorical. Cox proportional hazards were used to handle the data censoring along with stepwise regression to ensure that only the necessary variables were utilized. In total, 99 survival functions were built to handle all possible characterizations of the population. After validation, it was concluded that this was a good method to track non-rated officer attrition because it gave very similar results to the currently employed method.

Franzen [16] extended Schofield's [15] work by analyzing the Air Force's rated community. Rated officers come from the pilot, air battle manager (ABM), and combat systems operator (CSO) career fields. Again, logistic regression was leveraged to find odds ratios for a similar set of variables. The demographic variables in Franzen's analysis were 1) marital status, 2) gender, 3) source of commission, 4) distinguished graduate at commissioning source, 5) prior enlisted service, 6) binary for any dependents. The results also showed promise as a method to track rated officer attrition.

According to Hill, Miller, and McIntyre, "some of the critical issues facing the military in the aggregate include: how to structure the military given the uncertainty of the future; how to maintain a viable military-industrial complex given the uncertain future; and how to allocate limited defense dollars among the services" [17]. Davis [18] states that the DoD has three views of models. "Live simulations involve real people and real systems; virtual simulations involve real people using simulators (e.g., flight simulators); and constructive simulations are what we usually think of as models, war games, and simulations" [18].

Hill, Miller, and McIntyre state that the Air Force has used simulation to study a variety of topics to include modeling an Autonomic Logistics System (ALS), support

equipment reduction, and Army recruiting [17]. The ALS study intends to equip fighter aircraft with the ability to self-diagnose faults in the system, which would enable the logistics systems to be more agile and less reactive. The ultimate goal of the study was to provide insights to the designers of the maintenance component into where research emphasis should be placed.

The support equipment reduction study discussed by Hill, Miller, and McIntyre [17] looked into reducing the amount of maintenance equipment and replacement parts sent on a deployment for repairing aircraft. This would theoretically enable the deployed forces to move quicker, but comes at the cost of reduced mission effectiveness. Finding the right balance for this trade-off is key to fulfilling the mission requirements. The third simulation project looks into gain insights into Army recruiting at local stations. This topic was looked at for three research projects. The first project was very basic and only modeled three recruiters and only one type of prospect. The final research project was the most true-to-life representation of the system by modeling three recruiters with varying abilities, seasonal fluctuations, and nine prospect types. The insights gained from these studies enabled analysts, recruiters, and decision-makers to make better decisions about how to successfully recruit more individuals.

Castro and Huffman [19] analyzed the retention intentions of 289 United States Army enlisted and junior officers that were stationed in Germany and Italy. Soldier's opinions were elicited via survey on their intentions to stay in the military or leave the service at some point in the future. This survey data, along with the soldier's demographic data, were used to generate multiple CHAID and logistic regression models. CHAID was used to analyze the survey data. These results were used as inputs to the logistic regression models that predicted the chance of a person leaving the force. Their analyses showed that both the survey data and the demographic data were required to obtain an accurate model [19]. An obvious problem with this

approach is that survey data can be widely inaccurate and people willing to provide truthful and accurate data are hard to find.

2.3 Previous Work Related to Non-US Military Manpower

Parker and Marriott [20] used a stocks and flows model to track employees of different pay bands. Using cost as the main factor, the model balanced the pay-grade constraints that were forced on the organization. This method of study allowed management to alter the numbers of certain attributes within the system and allowed them to gain insight into their workforce. This type of modeling could be used to model Air Force employees but comes at the cost of aggregation when considering the entirety of the civilian population.

Cho et al. [21] used survival analysis to estimate survival curves of Korean nurses whose first job after graduating was as a full-time registered nurse in a hospital. The 351 participants were asked a series of questions over the course of three years from 2006 to 2008 about factors related to education, the hospital where they worked, individual and family qualities, as well as job dissatisfaction. The hospitals were measured by whether the nurses were unionized, small or large, and whether it was in a large metropolitan area. The purpose of the study was to identify significant factors that led to higher turnover rates among nurses. Doing so would allow the Korean government and healthcare system to intervene and make the work environments more hospitable to the nurses. The researchers found that only 54% of nurses were still employed after three years at their first job. They postulate that job satisfaction as well as the hospital's characteristics were major contributors to the nurses' turnover.

Two papers examine rural doctor retention in Australia and focus on identifying the factors that lead to increased retention. Bailey et al. [22] analyzed rural Western Australian doctor retention using survival analysis and Cox proportional hazards

regression. Typically population health is correlated with a higher concentration of doctors. This results in a health inequity where larger cities are healthier than rural towns. The purpose of the study was to determine which factors were associated with retention and to analyze if the recent intervention by the Australian government to improve retention of rural general practitioners was successful. Western Australia has typically relied on recruiting foreign doctors to supplement the declining workforce. 1154 doctors were analyzed over 10 years. A 7% increase in retention after five years was found after the government intervention. This improvement is attributable to doctors that began as general practice registrars as opposed to non-registrars. This methodology is similar to the analysis performed in this thesis in that the purpose is to identify the important retention factors.

Russell et al. [23] also studied important factors leading to increased retention of Australian doctors, but concentrated on the most populated state, New South Wales. 3,354 physicians practicing between 2003 and 2012 were studied using a survival analysis method. They determined that geographic location, population size, country of primary medical degree, procedural activity, and VMO (the right to provide medical services in a public hospital) were the most important factors related to retention. Australian-trained, non-procedural physicians that had VMO rights were the most likely to stay in their location whereas physicians working in a small town with less than 5,000 people were likely to leave. Coastal area doctors were likely to stay due to most large cities being located on the coast. They also stated that non-Australian trained doctors who did not prefer to serve in a remote location were the most likely candidates to leave employment.

Zini et al. [24] studied the burnout levels of Israeli dentists using multiple logistic regression. After a few studies in Northern Ireland, Netherlands, and the United Kingdom indicated that dentists were suffering from extreme levels of burnout, it was

proposed that a similar study be conducted on Israeli dentists to gauge the level of dissatisfaction with their current employment. A questionnaire focused on burnout was answered by 320 dentists. These answers were compared to the dentists' sex, age, specialization, and experience. Multiple logistic regression was employed against the ten burnout questions. It was found that females were more likely to feel depressed and physically weak and that dentists with only 10-20 years of experience were likely to feel helpless or worthless. Finally, general dentists were likely to feel tired, helpless, physically weak, and worthless. This information was crucial to understanding the behavior and thoughts of the Israeli dentists and is hopefully being used to curb the feelings of the dentists.

Capon, Chernyshenko, and Stark [25] state that most military retention studies have focused on demographic data mining, but note that this has its limitations. "1) many demographic characteristics, such as gender, are inherent and cannot readily be changed; 2) recruiting policy based on demographics would further decrease the already diminishing source of potential recruits; and 3) although data mining can result in relatively high predictive validities, such approaches are ill-suited for building a theory of military retention/turnover" [25]. Because of these limitations, they apply civilian retention methodology which models motivated personal choice to the enlisted personnel of the New Zealand Army.

Capon, Chernyshenko, and Stark [25] utilize surveys of the target group that asked questions related to 1) job involvement, 2) organizational commitment, 3) perceived organizational support, 4) work satisfaction, 5) work-family conflict, 6) community involvement, 7) dispositions, 8) met expectations, and 9) intentions to remain. Unfortunately, the Air Force does not collect survey information related to similar categories for civilian personnel. The research conducted in this paper is in direct contrast to the methodology described by these researchers and attempts to apply military retention

methodology to the civilian workforce.

Most civilian workforces use techniques similar those in to Capon, Chernyshenko, and Stark [25] to study retention among their employees. Surveys are conducted to screen people for possible unhappiness or willingness to find employment elsewhere. Across all branches of the military, most retention studies have concentrated on modeling using personnel factors. The assumption is that people with similar qualities such as time in service, number of dependants, and gender all affect a service members likelihood of leaving the force. The research conducted here uses the military style of retention modeling to find groups of government civilians that are likely to leave as predicted by their personal attributes.

III. Data Review

3.1 Introduction

The first step in any analysis is the thorough dissection of the data to uncover insights and rectify any possible mistakes. Unfortunately, the amount of data provided for this study is too unwieldy to check every file for minuscule mistakes. An assumption made about the data is that it is a true-to-life representation of the employees at the time of recording with only small mistakes that in aggregate only make up a small percentage of the data. This chapter provides more information on MilPDS and the data.

3.2 MilPDS

The Air Force stores all of its personnel data in MilPDS. This includes military officers, enlisted personnel, and civilian employees. Each person has hundreds of data fields that are populated with information such as name, age, sex, duty title (current and history), rank, awards, educational history, etc. Thousands of technicians create and change these entries many times every day. This makes the data very unstable and prone to errors. These errors are corrected when noticed, but some errors still arise. Some changes to the data require the member to initiate the change and if this does not happen, the data is incomplete or wrong. There are many reasons why the data could be incorrect and a comprehensive list of these reasons would be too long to list here. The database is maintained regularly, but sometimes the data can become corrupted. This is overcome by keeping daily and monthly backups of the data for historical purposes.

3.3 Extracts

Personnel data analysis at HAF/A1 is usually performed using extracts taken from MilPDS at the end of the month. These extracts are snapshots in time taken at close to the same time every month. If the extract is created before a change to fix an error is entered into the system, then the record will still contain the error. Luckily, some mistakes are noticeable and are altered on the back end. The analysts at HAF/A1 and other supporting agencies such as the Air Force Personnel Center (AFPC) have developed programs to "clean" the data. This typically involves scrubbing a persons record and looking for any possible and well known mistakes such as projected rank being a Captain when the person is a Lt. Colonel. These types of mistakes are not common, but do occur from time to time.

A cursory glance at the data revealed that some individuals did have bad information. This includes people with negative ages or an age too young to work in federal service. Others include being older than the time they spent working for the government. Lastly, people with impossible occupational series codes were removed. Out of the 404,358 total number of records, 1,172 (0.3%) were deleted for these reasons.

To perform the analysis, a monthly collection of data sets that contained all individuals that left the service was created. This was achieved by comparing two successive monthly files to see if a person was in the following monthly file. The code for this is shown in Appendix A.

3.4 Dates and Variables

Over seventeen years of monthly extracts were provided for this analysis. Undoubtedly, programs and incentives intended to increase retention in the civilian workforce have and will continue to change. Selecting too little data would only model more recent trends and selecting too much data could lead to results that do

not adequately reflect the intentions of the present employees. With this in mind, the most current ten years of data were selected to represent the current retention trends. This time frame covers the dates between January 2010 and December 2019.

Most literature has concentrated on a few categorical variables such as age, sex, and race. Four variables other than these classic demographics were chosen for this analysis. From my past experience working as a DoD employee and as a personnel analyst at AFPC, I chose seven variables as necessary to review for their importance to retention: Age, Gender, Race, Education Level, Years of Service (YOS), Prior Military Service, and Occupational Series. Prior military service and occupational series are variables that are unique to the civilian data sets when compared to the officer and enlisted data sets.

Age, gender, race, occupational series, and YOS were not altered from the original files. Education level and prior military service were constructed as binary variables. Education level was reduced down to whether or not the person had a Master's degree or PhD as of their last known record. Prior military service was changed to only reflect if a person served in the military and does not include which service for which they volunteered. For both constructed variables, a "1" indicates the positive occurrence for that attribute. Occupational series was not used as a covariate in the logistic or survival analysis, but it was used to divide the people into smaller groupings. The rule used to separate people by occupational series was obtained from the Office of Personnel Management's (OPM) website [26] where people are grouped according to the first two digits of their occupational series code.

IV. Analysis

4.1 Logistic Regression

Logistic regression was used to determine which variables were significant to predicting if a person would retain in the civil service. The response variable (Retain) for logistic regression is binary and the predictor variables can be categorical or continuous. The variables used were age (AGE), gender (GENDER), race (RACE_GRP), prior military service (PRIOR_MIL), and years of service (YOS).

To perform the logistic regression analysis, all of the inventory and loss files spanning the time frame were combined separately and only the last known record of a person was kept. These two files were then combined and again the last known record was kept. If a person had separated during the time frame, that record would come from the loss files so they were marked with "Retain=0." If a person had not separated during the time frame, that record would come from inventory files so they were marked with "Retain=1." A few variables were added to records to simplify them and records with bad data were deleted from the cohort. This process is shown in Appendix B. Lastly, the PROC LOGISTIC command was used to perform the analysis. The code for this is shown in Appendix C. Table 1, obtained from OPM's occupational handbook [26], shows the titles for each four-digit grouping of occupational series. The 4000 and 9000 groupings did not have any Air Force employees over the last ten years so they are not studied.

Table 1: List of all OPM Occupational Series Groups

Occ Ser	Title
0000	Miscellaneous Occupations Group
0100	Social Science, Psychology, And Welfare Group

0200	Human Resources Management Group
0300	General Administrative, Clerical, And Office Services Group
0400	Natural Resources Management And Biological Sciences Group
0500	Accounting And Budget Group
0600	Medical, Hospital, Dental, And Public Health Group
0700	Veterinary Medical Science Group
0800	Engineering And Architecture Group
0900	Legal And Kindred Group
1000	Information And Arts Group
1100	Business And Industry Group
1200	Copyright, Patent, And Trademark Group
1300	Physical Sciences Group
1400	Library And Archives Group
1500	Mathematical Sciences Group
1600	Equipment, Facilities, And Services Group
1700	Education Group
1800	Inspection, Investigation, Enforcement, And Compliance Group
1900	Quality Assurance, Inspection, And Grading Group
2000	Supply Group
2100	Transportation Group
2200	Information Technology Group
2500	Wire Communications Equipment Installation And Maintenance Family
2600	Electronic Equipment Installation And Maintenance Family
2800	Electrical Installation And Maintenance Family
3100	Fabric And Leather Work Family

3300	Instrument Work Family
3400	Machine Tool Work Family
3500	General Services And Support Work Family
3600	Structural And Finishing Work Family
3700	Metal Processing Family
3800	Metal Work Family
3900	Motion Picture, Radio, Television, And Sound Equipment Operation Family
4000	Lens And Crystal Work Family
4100	Painting And Paperhanging Family
4200	Plumbing And Pipefitting Family
4300	Pliable Materials Work Family
4400	Printing Family
4600	Wood Work Family
4700	General Maintenance And Operations Work Family
4800	General Equipment Maintenance Family
5000	Plant And Animal Work Family
5200	Miscellaneous Occupations Family
5300	Industrial Equipment Maintenance Family
5400	Industrial Equipment Operation Family
5700	Transportation/Mobile Equipment Operation Family
5800	Transportation/Mobile Equipment Maintenance Family
6500	Ammunition, Explosives, And Toxic Materials Work Family
6600	Armament Work Family
6900	Warehousing And Stock Handling Family
7000	Packing And Processing Family

7300	Laundry, Dry Cleaning, And Pressing Family
7400	Food Preparation And Serving Family
7600	Personal Services Family
8200	Fluid Systems Maintenance Family
8600	Engine Overhaul Family
8800	Aircraft Overhaul Family
9000	Film Processing Family

The assumptions about the model and variables, as well as the issue of multi-collinearity between variables are deserve discussion. Figures 1 and 2 display the distributions and normality plots for the two continuous variables, Age and YOS, respectively. The best outcome when looking at a normality plot is to see the points fall perfectly along the diagonal line. In the current situation, the bottom tail on both graphs shoots out to the left, but the majority of points on the graph show that the normality assumptions on Age and YOS are satisfied. The large number of employees that have 0-1 YOS contribute to the tail the most.

The issue of potential multicollinearity among the predictor variables is examined by finding the correlation measures among the predictor variables. Correlation is a measure of the strength of the linear relationship between variables. Values fall between -1 and 1 with the extreme values representing a highly negative and highly positive relationship, respectively. Examining predictor variable correlation helps ensure that two variables that provide the same predictive capability do not both enter the model. The cutoff point used to indicate high correlation was 0.80. A variable with a correlation value larger than this is not used in the model. The largest correlation value between the continuous predictor variables Age and YOS is 0.66935. This indicates some positive correlation between the two, but it does not

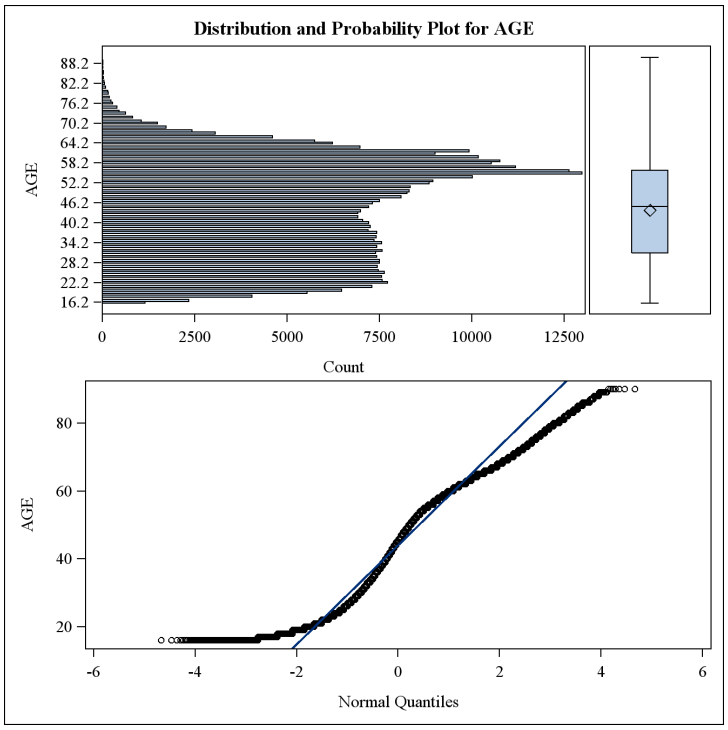

Figure 1: Normal Probability Plot for Age

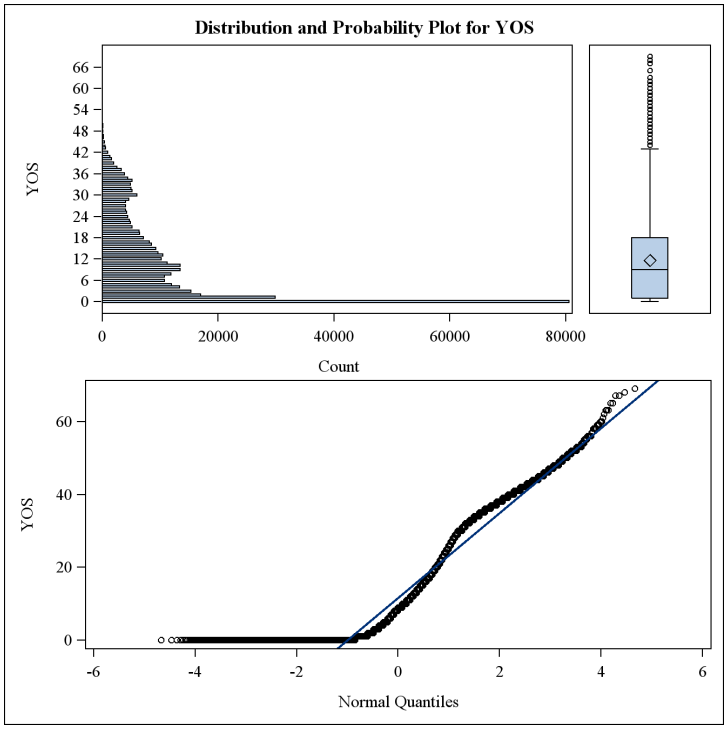

Figure 2: Normal Probability Plot for YOS

meet the threshold established, thus both variables were screened for inclusion in the model.

Table 2 shows the correlation values produced by SAS between the categorical variables. For the correlations with race, the "contingency coefficient" is displayed which takes into account the multiple categories of race. This value is only appropriate for variables with more than two categories. Correlation values between continuous and categorical variables were found by calculating the point-biserial correlation coefficient which is mathematically equivalent to the Pearson correlation coefficient. The values are shown in Table 3. No values from either table were greater than 0.80 which, based on our cut off value, indicates there is no issue with multicollinearity.

Table 2: Correlations Between Categorical Variables

	Retain	**Gender**	**Race_Grp**	**HighEd**	**Prior_Mil**
Retain	1	0.1525	0.0308	0.1458	0.1328
Gender		1	0.1207	0.0617	0.3823
Race_Grp			1	0.1035	0.1879
HighEd				1	0.1228
Prior_Mil					1

Table 3: Correlations Between Categorical and Continuous Variables

	Age	**YOS**
Retain	0.15776	0.10939
Gender	0.19328	0.14335
Race_Grp	0.01336	0.03515
HighEd	0.16151	0.10661
Prior_Mil	0.32775	0.21999

The results of the logistic regression performed on the entire collection of people in the data set are shown in Table 4. A yellow box indicates that the covariate was not significant in predicting that group's likelihood of retention. A red box indicates that no estimate for the covariate could be calculated due to the entire population being of one type or that the occupational series grouping did not converge. The occupations that did not converge had too few data points. These groupings are not examined further because no accurate information could be obtained from them.

Table 4: Analysis of Effects for All Occupational Series

Occ Ser	Obs	Age	Gender	Race	Higher Ed	Prior Mil	YOS
All	403,060	<.0001	<.0001	<.0001	<.0001	<.0001	0.0514

The standard acceptability threshold for hypothesis tests of $\alpha = 0.05$ was used as a rough cutoff point. For the entire population, the only covariate to not meet this criteria was YOS, but only by a small margin so it was deemed important for predicting retention. The convergence criteria used by SAS was satisfied and all three of the overall model hypothesis tests showed the overall model was significant. The same procedure was performed for each two-digit grouping of the occupational series and is shown in Table 5.

Table 5: Analysis of Effects for Each Occupational Series

Occ Ser	Obs	Age	Gender	Race	Higher Ed	Prior Mil	YOS
0000	18,764	<.0001	0.005	<.0001	<.0001	<.0001	<.0001
0100	26,509	<.0001	0.0401	<.0001	<.0001	<.0001	<.0001
0200	8,773	0.2058	0.0061	<.0001	<.0001	<.0001	0.1108
0300	68,847	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
0400	727	0.5584	0.3716	0.4469	0.0008	0.0845	0.1638

0500	16,306	0.6072	<.0001	<.0001	<.0001	0.0034	0.0793
0600	9,871	<.0001	0.0005	0.0002	<.0001	0.0087	0.3997
0700	276	0.6974	0.3771	0.9996	0.4241	0.3436	0.9559
0800	29,557	0.0154	0.0006	<.0001	<.0001	<.0001	<.0001
0900	2,347	<.0001	0.0383	0.1306	<.0001	0.0014	0.7737
1000	3,240	0.7956	0.0081	<.0001	0.0002	0.0029	0.3485
1100	28,386	<.0001	<.0001	<.0001	<.0001	0.0749	<.0001
1200	26	0.1957	0.5772	0.9971	0.9425	0.7511	0.2844
1300	2,528	0.0012	0.354	0.6306	0.0352	0.9876	0.0024
1400	1,847	0.0485	0.0011	0.0384	<.0001	0.8979	0.0003
1500	4,538	0.0003	0.0678	0.0122	0.905	<.0001	<.0001
1600	5,347	<.0001	0.0082	0.3263	<.0001	0.1436	<.0001
1700	37,844	<.0001	0.002	<.0001	<.0001	<.0001	<.0001
1800	1,483	0.4491	0.2443	0.5928	0.0077	0.0005	0.0244
1900	1,665	<.0001	0.0116	0.2801	0.2176	0.0317	<.0001
2000	9,400	<.0001	<.0001	<.0001	<.0001	<.0001	0.4513
2100	9,569	<.0001	0.0501	<.0001	0.0009	0.0035	<.0001
2200	12,047	<.0001	<.0001	0.3253	0.0056	0.0363	<.0001
2500	394	0.3858	0.3563	0.1628	0.9231	0.854	0.0339
2600	5,316	0.3637	0.0813	<.0001	0.3598	<.0001	<.0001
2800	3,567	0.0408	0.0344	<.0001	0.9493	<.0001	<.0001
3100	201	0.7578	0.26	0.0394	0.9541	0.0338	0.7303
3300	166	0.4823	0.0037	0.9256	0.9091	0.4925	0.0754
3400	1,386	0.0225	0.1211	0.0031	0.3293	<.0001	<.0001
3500	17,709	<.0001	0.0243	<.0001	0.9676	<.0001	<.0001

3600	344	0.0486	0.9671	0.2827		0.8324	0.3047
3700	1,545	0.066	0.0075	0.8287	0.5878	<.0001	<.0001
3800	6,874	0.0535	0.0004	<.0001	0.5199	<.0001	<.0001
3900	9	0.92		0.7174		0.9604	0.8193
4100	1,928	0.1365	0.8873	0.0005	0.5078	<.0001	0.1322
4200	1,227	0.03	0.168	0.8585	0.4571	0.1274	<.0001
4300	766	0.9035	0.3308	0.9126	0.4371	<.0001	0.0001
4400	7	1	0.9737	1		0.9329	1
4600	980	0.1442	<.0001	<.0001	0.9626	0.2187	0.1181
4700	4,363	0.0848	0.641	<.0001	0.0204	0.0027	0.6166
4800	1,250	0.3709	0.4734	0.0014	0.2293	0.0472	0.0321
5000	1,258	0.7859	<.0001	<.0001	0.7629	0.1404	0.0087
5200	73	0.0342	0.9562	0.9798	0.9688	0.0405	0.1341
5300	3,342	0.0002	0.3054	0.1557	0.5159	<.0001	<.0001
5400	1,476	0.002	0.4014	<.0001	0.7426	0.016	0.0076
5700	5,536	<.0001	0.002	<.0001	0.4221	0.4746	<.0001
5800	2,386	0.6067	0.1019	<.0001	0.6447	<.0001	0.1826
6500	250	0.1546	0.6893	0.5908	0.4167	0.3337	0.4417
6600	1,279	0.0966	0.1432	0.0227	0.8865	0.1575	<.0001
6900	4,554	0.0005	0.3507	<.0001	0.5949	<.0001	<.0001
7000	495	0.9652	0.6825	0.256	0.7831	0.0114	0.6102
7300	330	0.4917	0.0349	<.0001	0.9857	0.2658	0.0574
7400	19,209	<.0001	<.0001	<.0001	0.7562	0.002	<.0001
7600	44	0.1495	0.7479	0.995		0.9432	0.0589
8200	2,164	0.3676	0.3374	0.991	0.8105	<.0001	<.0001

8600	2,590	0.1408	0.1269	0.0296	0.9754	<.0001	<.0001
8800	10,145	0.8705	0.0261	0.0431	0.3197	<.0001	<.0001

The factors become less significant as the data are split into smaller groups. Since the purpose of this study is to look at retention issues on a macro scale, no deeper divisions were pursued for fear of having too shallow of a population with which to work. Groupings that had a very small number of people over the last ten years did not satisfy the convergence criteria and thus some estimates are missing. For these smaller occupations, the remaining estimates are likely biased and should not be trusted so they were not examined further.

OPM's occupational handbook [26] makes a clear divide between white-collar and trade, craft, and labor jobs. Any two-digit occupational series code between "0000" and "2200" is considered white-collar and any code between "2500" and "9000" is considered a tradecraft. It is expected, or in some cases mandatory, that white-collar workers have at least a master's degree while this is not so expected for trade laborers. This is apparent when looking at the higher education covariate. Ignoring the occupational series that did not converge (red), eighteen of twenty white-collar occupations showed higher education as important for retention and only one of twenty-one trade labor occupations showed it as important. This distinction indicates that the two groups have different reasons for retaining or not retaining.

Figures 3 and 4 display the standard influence charts produced by SAS. As a reminder, "Retain=0" which is shown by a blue circle indicates that the person has left employment during the time frame. Similarly, "Retain=1" which is shown by a red cross indicates that the person was still employed at the end of the time frame. The Pearson and Deviance residual charts show no major deviations between the actual and predicted values. As well, the leverage graphs do not show any noticeable

outliers. Because the model assumptions were met, the correlations do not show any issues with multicollinearity, and the influence diagnostics do not show any worrying signs, the logistic regression model remains unchanged.

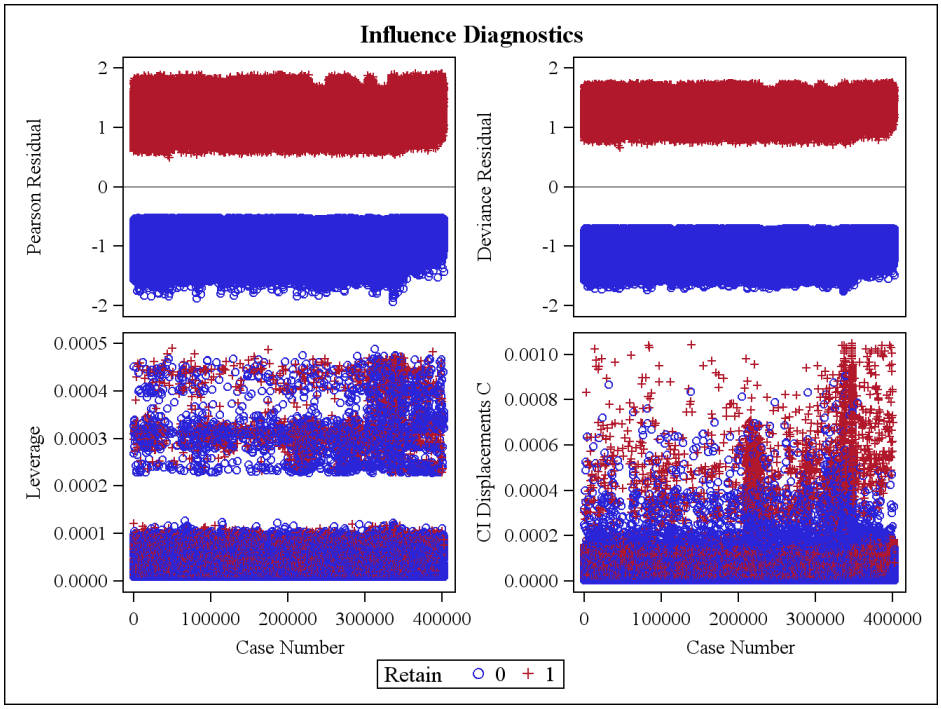

Figure 3: Logistic Regression Influence Diagnostics (Part 1)

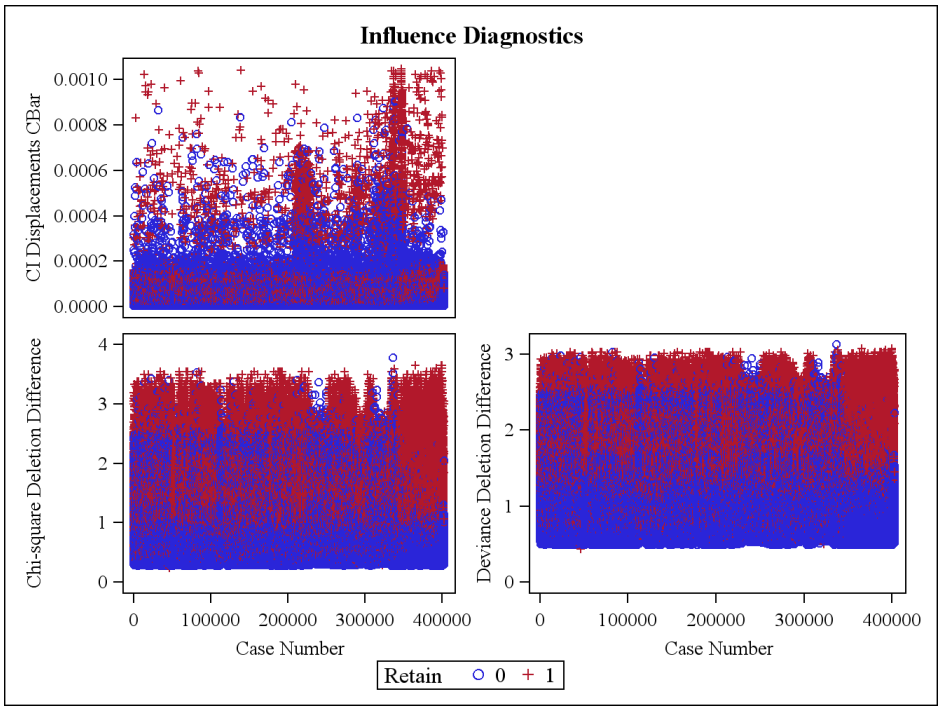

Figure 4: Logistic Regression Influence Diagnostics (Part 2)

4.2 Odds Ratios

Odds ratios can be calculated when performing logistic regression. Odds ratios are interpreted as the likelihood over the baseline for retention. The baseline odds ratio is always one. When compared to the baseline, a number above one indicates that it is n times more likely to retain while a number below one is n times less likely to retain. For example, Figure 5 shows that for the entire force, females are 0.627 times less likely to retain compared to males. A potential answer for why this is happening might be that women are more likely to fall into a traditional role in the family where they take care of the children, while the man is likely to stay employed in the workforce.

Figure 6 shows Black employees are the only race less likely to retain than White employees. All other race categories are more likely to retain than White employees in aggregate. Black employees were 0.939 times less likely to retain over White em-

ployees while Multi-race employees were 1.380 times more likely to retain over White employees in aggregate. There is not a large difference in odds ratios between races unlike between the genders.

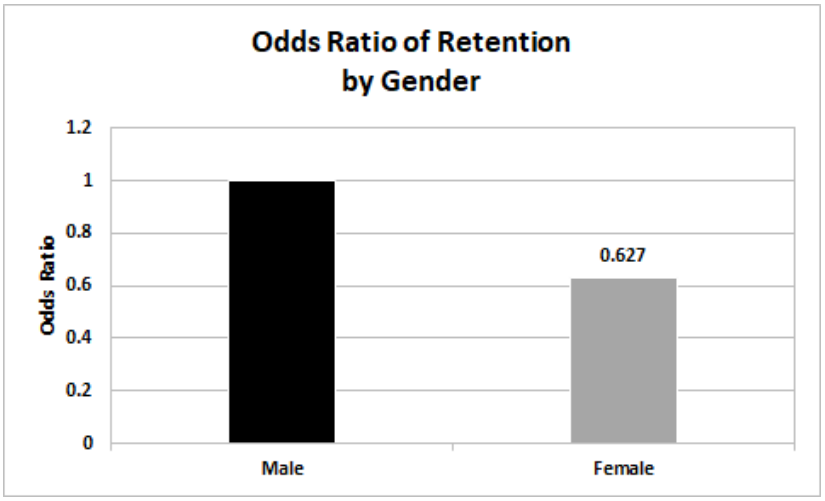

Figure 5: Odds Ratio of Retention by Gender (All)

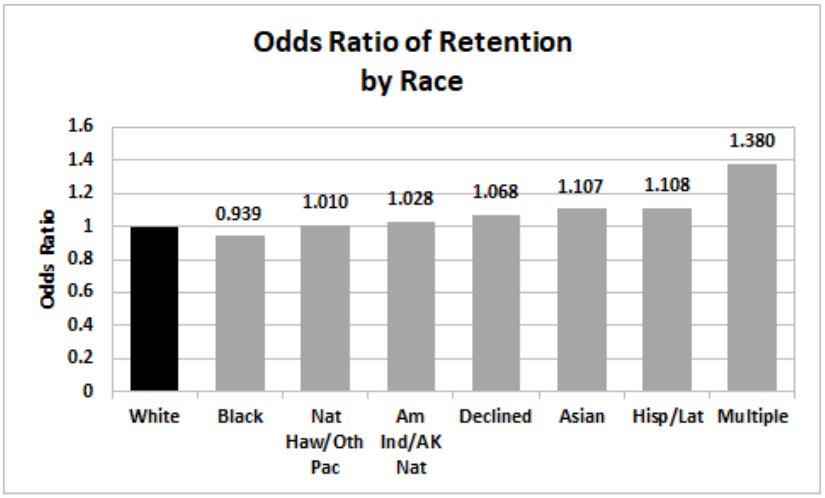

Figure 6: Odds Ratio of Retention by Race (All)

Figure 7 shows individuals that do not have a Master's degree or PhD are 0.520 times less likely to retain when compared to people that do have those degrees. Because OPM splits the occupational series into white-collar and laborers, this topic will be revisited by looking at a few occupational series' odds ratios. Figure 8 indicates

that employees with any prior military service are more likely to retain than people without prior military service. The people with no prior service are 0.821 times less likely to retain. A potential reason for this occurring might be that prior military personnel are more acclimated to the workload and structure that comes with working in a military environment and potentially have a greater sense of patriotism.

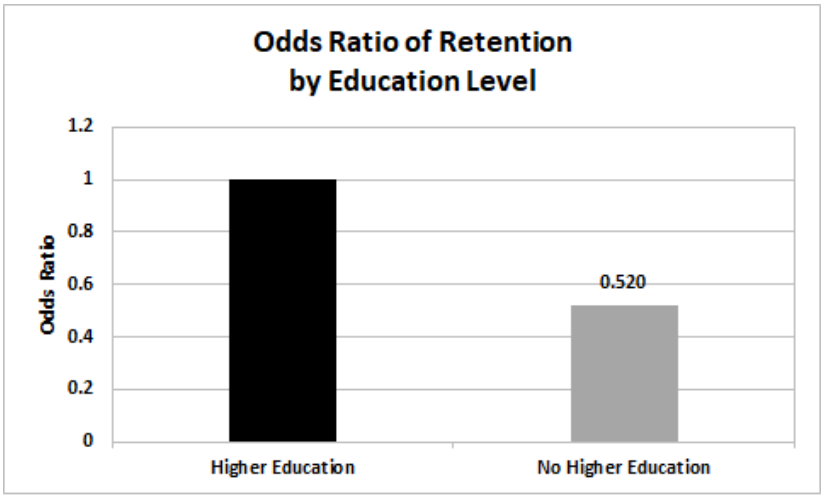

Figure 7: Odds Ratio of Retention by Education Level (All)

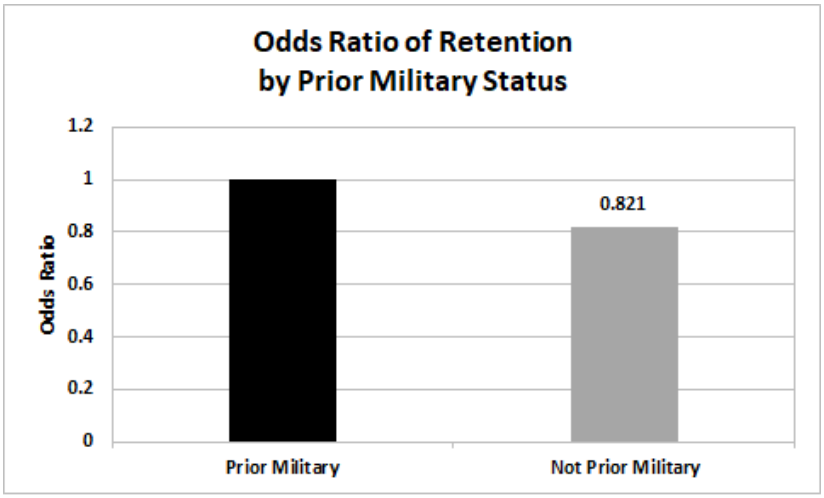

Figure 8: Odds Ratio of Retention by Prior Military Status (All)

For the continuous variables (Age and YOS), the interpretation of the odds ratio is slightly different. For a one unit increase in age, the odds of retaining increase

by a multiplicative factor of 1.015. For example, say the likelihood of retaining of a 30-year-old is 10. Then, after a year, their new likelihood, holding all else equal, is 10.15. Similarly, for every one unit increase in YOS the odds of retaining increase by a factor of 1.001.

Two white-collar and two laborer occupational series groups were chosen to compare. The 0800 (Engineering and Architecture), 1700 (Education), 3500 (General Services and Support Work), and 7400 (Food Preparation and Serving) occupational series were selected because they constitute a greater portion of civilian employees compared to other occupations. Figure 9 shows that only for the 3500 series do females have a better retention than males. Females working in the 3500 occupational series are 1.110 times more likely to retain when compared to males. Conversely, females are 0.885, 0.888, and 0.825 times less likely to retain in the 0800, 1700, and 7400 occupational series, respectively, when compared to men.

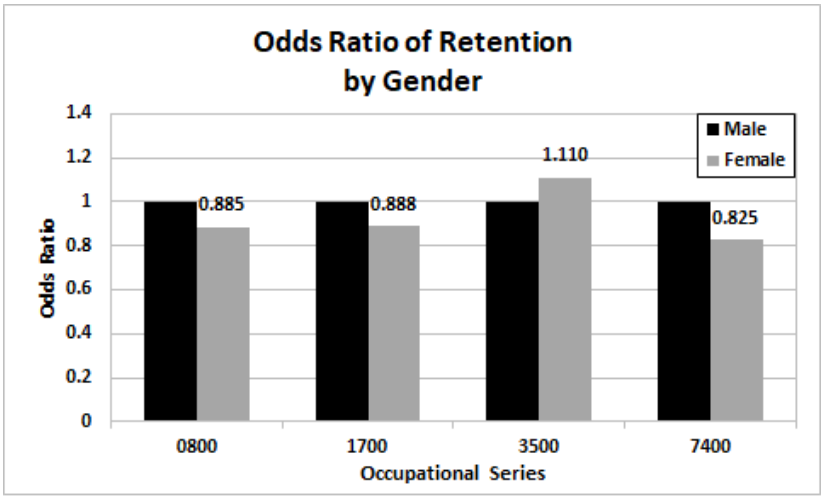

Figure 9: Odds Ratio of Retention by Gender (Separate)

Figure 10 shows the odds ratios for each race. People that did not indicate a race were 0.534 times less likely to retain in the 0800 occupational series, but were 2.568 times more likely to retain in the 7400 occupational series when compared to white employees. All races were more likely to retain in the 1700 occupational series than

white employees.

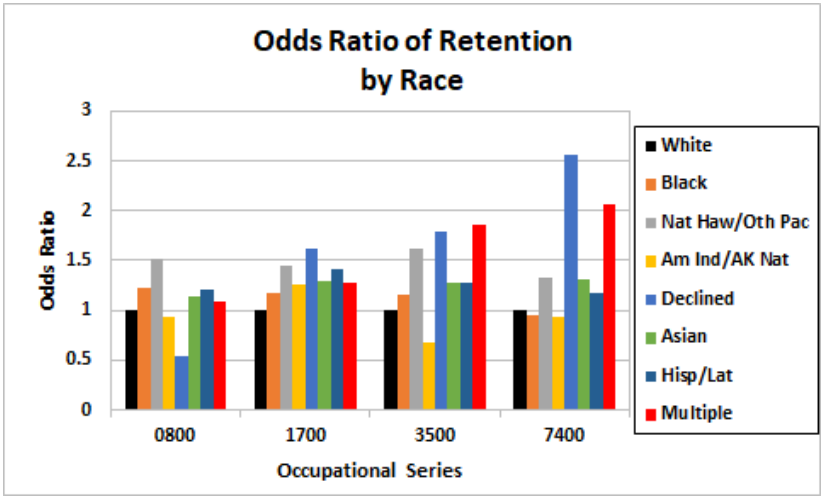

Figure 10: Odds Ratio of Retention by Race (Separate)

Figure 11 indicates that there is a noticeable difference between the white-collar and laborer jobs when comparing people with and without a higher education. For the 0800 and 1700 series, employees without a Master's degree or PhD are 0.628 and 0.640 times less likely to retain, respectively, than people with those degrees. For the 3500 and 7400 series, employees without a Master's degree or PhD are 0.986 times less likely and 1.076 times more likely to retain, respectively, than people with those degrees. A possible explanation for this occurrence could be that the two white-collar series are more likely to expect a higher level of education. Leaving people that can not meet this expectation with higher stress levels leading to their resignation or they are pushed out because of poor performance.

Figure 12 shows the prior military status odds ratios for the four occupational series. Figure 8 showed that in the aggregate, prior military were more likely to retain. However, in three of the four occupations shown, non-prior military individuals are more likely to retain. The difference between the overall and the subdivided results can most likely be attributed to the differences between the occupations. Some occupations might suit the mindset and expectations that a prior military employee

might have. In the 0800, 3500, and 7400 occupational series, people with no prior service were 1.350, 1.720, and 1.430 times more likely to retain, respectively, than people with prior military service. In the 1700 occupational series, non-prior military were 0.616 times less likely to retain.

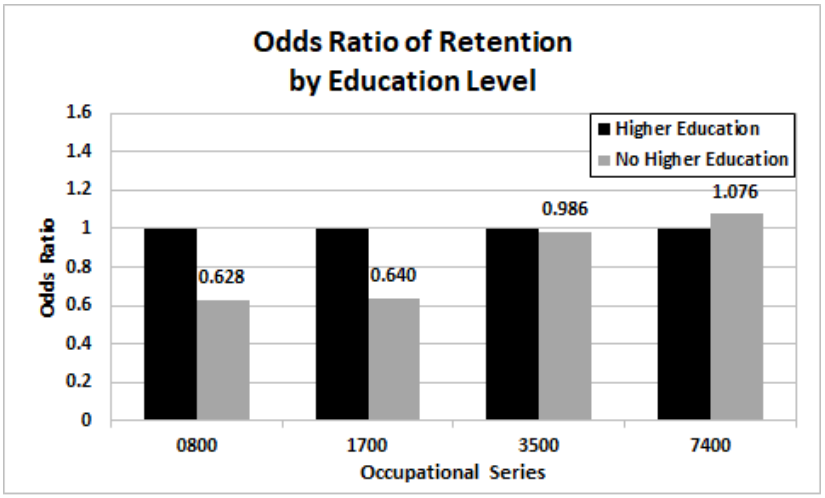

Figure 11: Odds Ratio of Retention by Education Level (Separate)

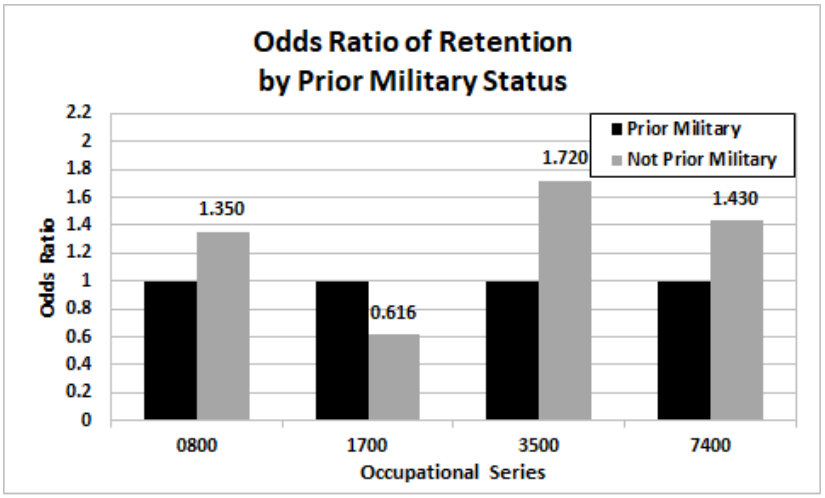

Figure 12: Odds Ratio of Retention by Prior Military Status (Separate)

Using logistic regression as a means of predicting retention has its drawbacks when the data is censored. A censored dataset has observations where the whole time frame is not captured. In this case, the data used contains individuals that have

not yet left employment; otherwise known as right-censored data. Logistic regression does not handle this type of data very well, but it can still be used as a baseline for determining important covariates. Survival analysis is used to handle censored data and is explored next.

V. Application

5.1 Survival Analysis

Survival analysis is used to model the factors that affect the time of an event. The graphs created using this technique will show a non-increasing line indicating the survival probability as a function of time. Given a length of employment, we can estimate the survival probability for that individual. This is accomplished in SAS using the functions PROC LIFETEST and PROC PHREG. The former function is used as a means of nonparametric estimation using the Kaplan-Meier method and the latter creates a semi-parametric Cox proportional hazards model that can be used to test for the strength and significance of the effects.

Understanding survival analysis curves requires understanding probability distribution curves. The probability density function, PDF, is the function that describes the relative likelihood of observing a particular value. In the case of retention this is showing the likelihood of surviving a a certain amount of time. The PDF is denoted by $f(t)$. Another useful function is the cumulative density function, CDF, denoted by $F(t)$. The CDF describes the likelihood of observing a value less than or equal to some time t and is calculated by integrating the PDF. This is shown in equation 1. Consequently, the PDF can be obtained by taking the derivative of the CDF as shown in equation 2.

$$F(t) = \int_0^t f(u)du \tag{1}$$

$$f(t) = \frac{dF(t)}{dt} \tag{2}$$

The reliability function, commonly referred to as the survival function, R(t), de-

scribes the likelihood of surviving beyond time t. See Equation 3 for the derivation. Two other useful functions are the hazard rate function and the cumulative hazard rate function which are shown in equations 4 and 5, respectively. The hazard rate function, describes the relative likelihood of a failure at some time t, conditional on the survival up to that time. The cumulative hazard rate function accumulates the instantaneous hazards over time.

$$R(t) = 1 - F(t) \tag{3}$$

$$\lambda(t) = \frac{f(t)}{R(t)} \tag{4}$$

$$\Lambda(t) = \int_0^t \lambda(u)du \tag{5}$$

5.2 Kaplan-Meier Survival Functions

The population of people that left employment was initially analyzed. This data excludes people that were still working for the Air Force at the end of December 2019, AKA censored individuals. This will demonstrate the five functions while a more in depth look at the data with censored individuals included follows. The SAS code to create all of the graphs can be found in Appendix D. Figure 13 shows summary statistics for each group of people. "$Retain = 0$" means that the person has left employment while "$Retain = 1$" means that they are still employed at the end of December 2019.

Figure 14 displays a histogram of the variable YOS along with a smoothed approximation of the PDF. From this we can see that around 50% of people will leave employment in the first five years, and the remaining will leave over the following few

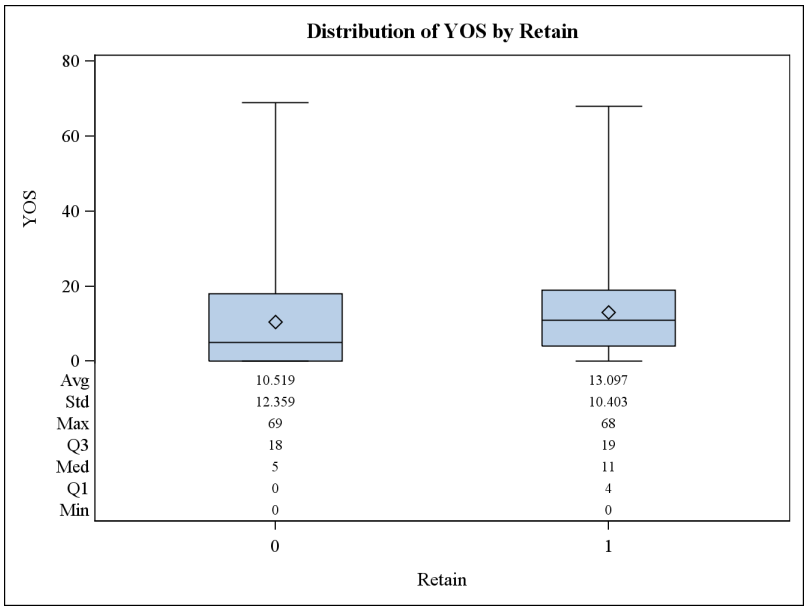

Figure 13: Summary Statistics

decades. Figure 15 displays the CDF of the data. This confirms that roughly 50% of individuals left employment with the Air Force in the first five years. Figure 16 displays the survival curve. The number "At Risk" is displayed along the bottom of the chart and 95% confidence intervals are shown by the shaded region around the line.

Figure 17 displays the hazard rate function. As expected, given the previous graphs, the rate that people leave employment is large at the beginning of employment, flattens out, then slowly increases after twenty years. There is a decline in the rate of people leaving around the 45 year mark possibly due to the individuals not wanting to seek other employment and just wait until they feel comfortable with the amount of retirement savings they have accrued. The decrease happens gradually until around sixty years of employment where the rate spikes and almost becomes asymptotic. Figure 18 displays the cumulative hazard rate over time.

The purpose of displaying these graphs is to show graphically the classical way of viewing employment by considering only the people that have left employment. In

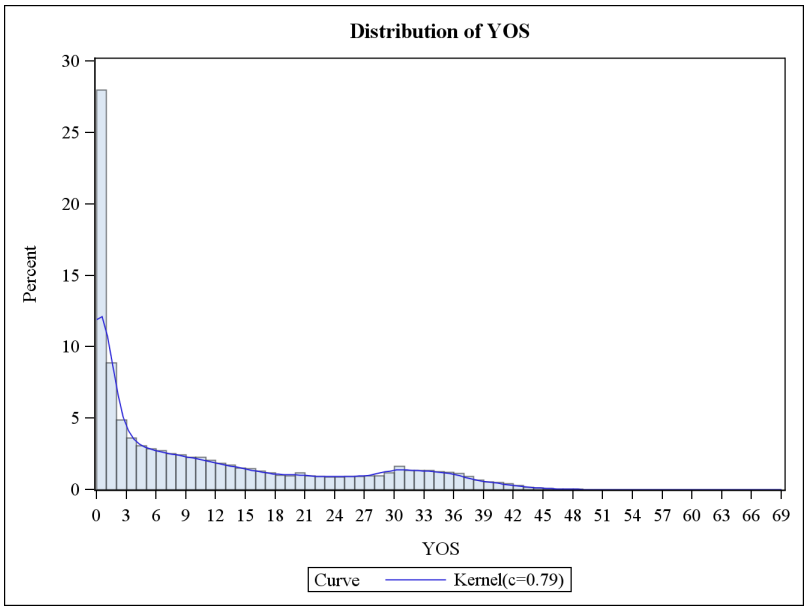

Figure 14: PDF of YOS (excluding censors)

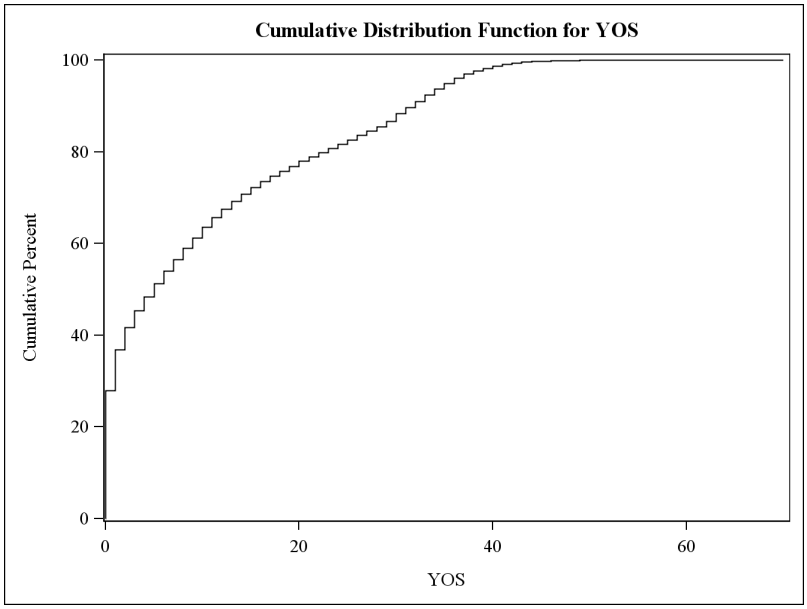

Figure 15: CDF of YOS (excluding censors)

truth, the people that are still employed can also be used in the reliability and survival analysis. The main purpose for using survival analysis is to include people that are still employed, AKA the censored individuals. Minor modifications were made to the code to include these people in the analysis and the code is in Appendix D. Only the

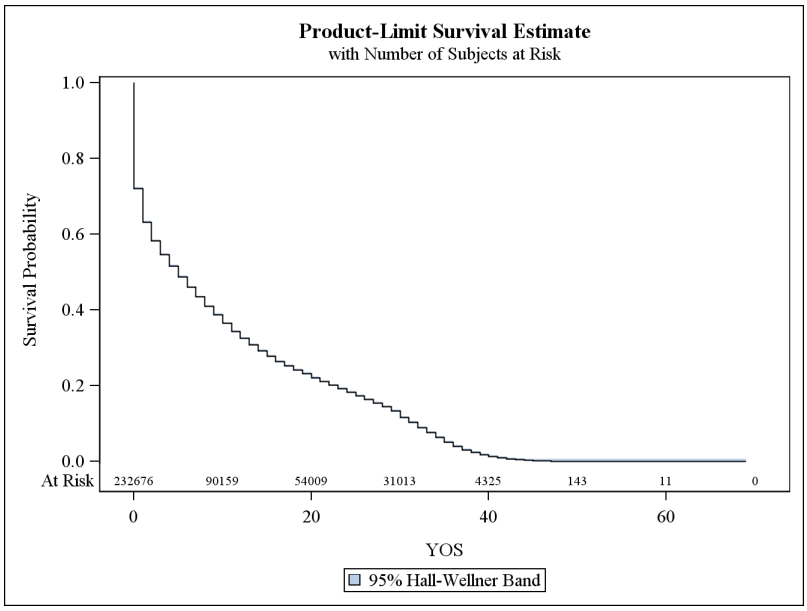

Figure 16: Survival Curve (excluding censors)

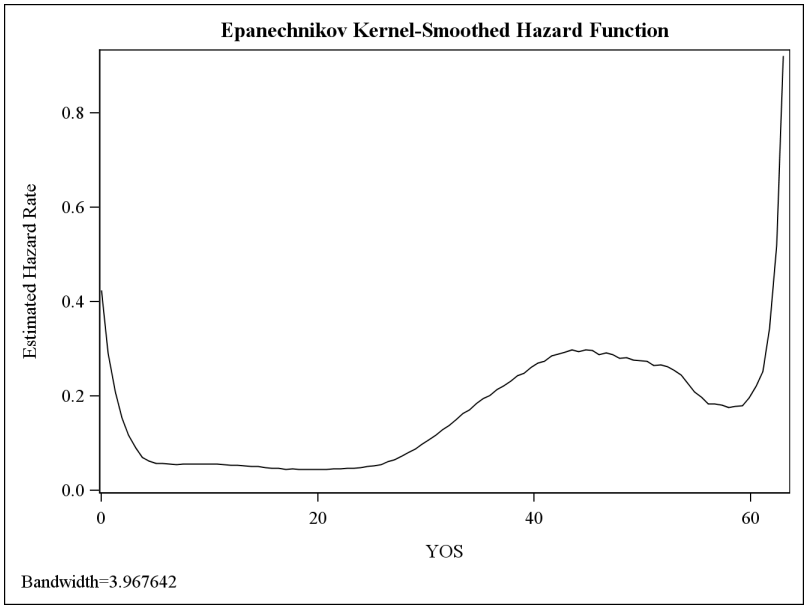

Figure 17: Hazard Rate Function (excluding censors)

Kaplan-Meier survival curves are discussed.

Figures 19 through 25 display survival curves with the censored individuals included. The censored records are indicated by "+" signs on the curves. Figures 16 and 19 show the difference in the curves with and without the censored individuals

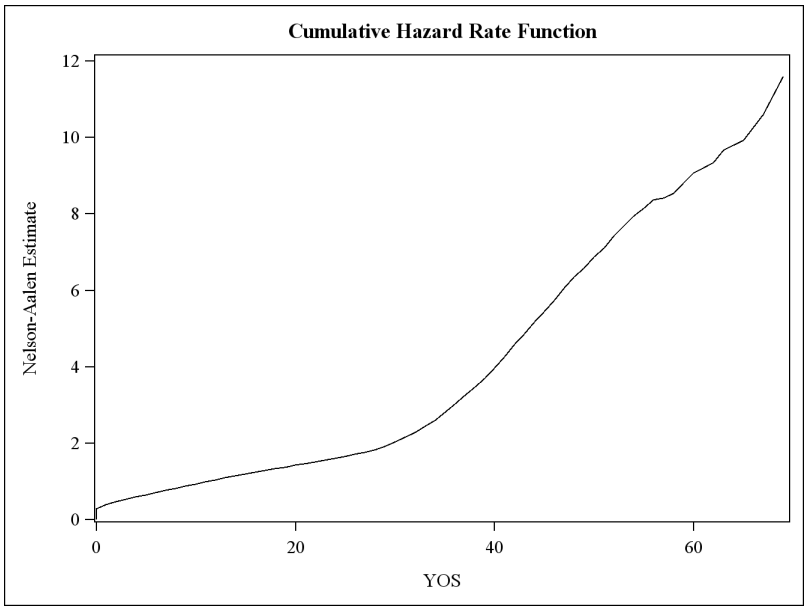

Figure 18: Cumulative Hazard Rate Function (excluding censors)

included. The 95% confidence bands are indicated graphically around the survival curve by the shaded region. Graphs that have multiple breakouts also test for homogeneity. This uses a logrank test where low p-values, shown in the upper right corner of the graph, indicate that the breakouts are distinct.

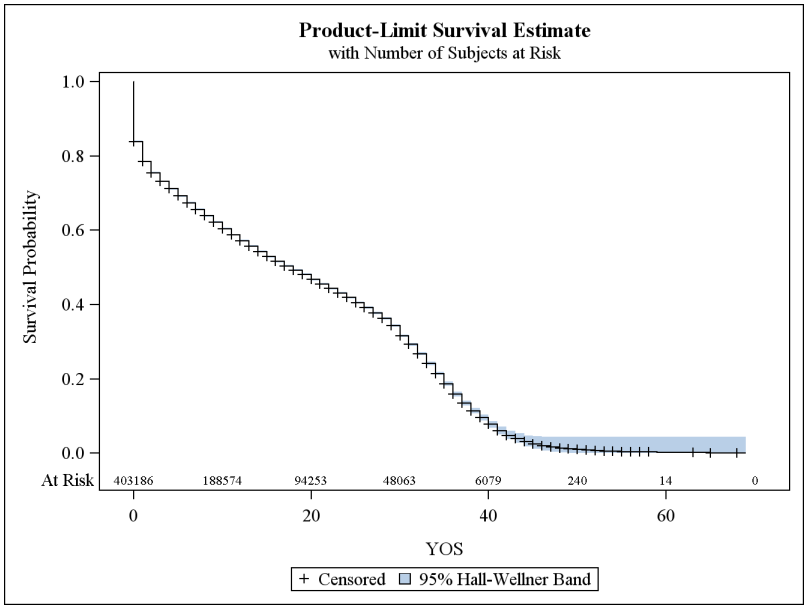

Figure 19: Survival Curve for All Employees

Figure 20 compares the survival curves between males and females. Males are shown by the red line and females by the blue line. The p-value is <0.0001 which indicates that the two groups of people are statistically different from one another. Male employees are more likely across all time to "survive" than females which is the same conclusion that was drawn from the odds ratio data from Figure 5. In the first year of employment, the reliability of females drops to roughly 75% while males only drops to around 90%.

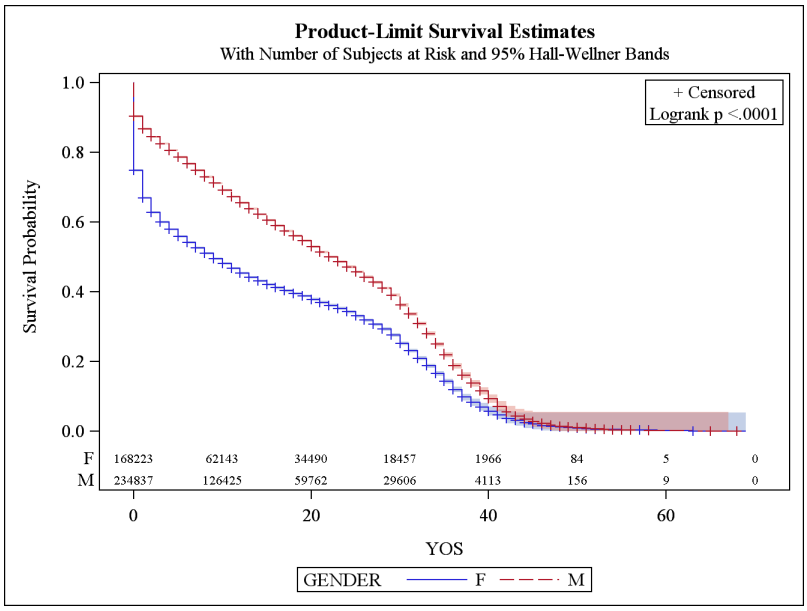

Figure 20: Survival Curve by Gender

Figures 21 and 22 show the survival curves for each race. Interestingly, while Figure 6 shows Multi-race employees as having the best odds ratio, the survival curve tells a different story. This group of people has a very steep drop off at the beginning of their employment. Every race group besides "Declined to Respond" shows a steep drop in the first few years followed by a slow decline in survival probability. No explanation can be found for the anomalous behavior exhibited by the declined group.

Education level is examined using Figure 23. The logrank test returns a p-value of <0.0001 — affirming that the two groups of people are distinct from one another.

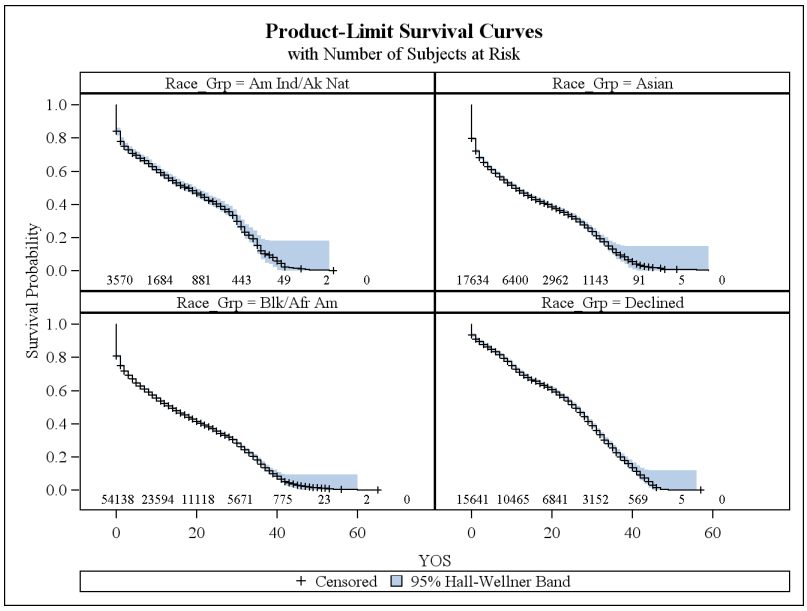

Figure 21: Survival Curve by Race (Part 1)

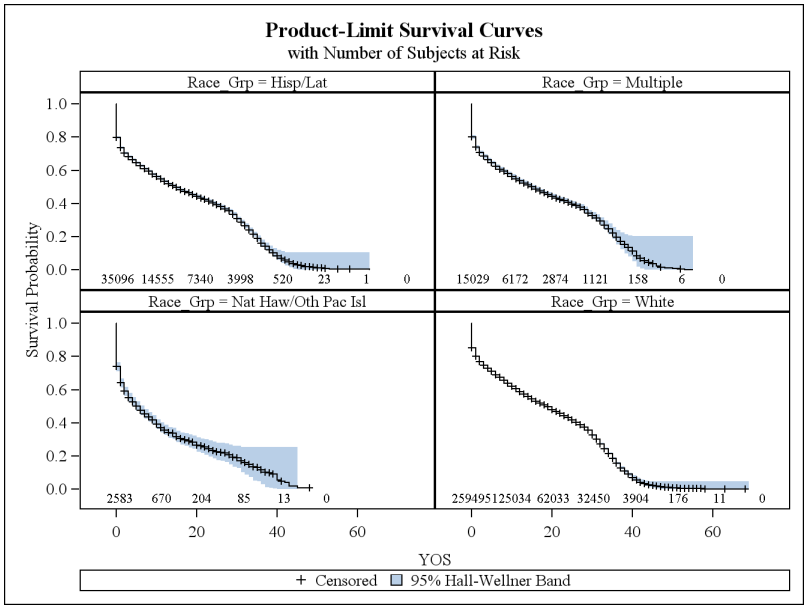

Figure 22: Survival Curve by Race (Part 2)

Individuals with a master's degree or higher have a better survival curve than those without those degrees. This is a similar result when looking at the odds ratios. Practically no employees with an advanced degree leave employment during the first year while nearly 20% of people without an advanced degree leave during the first

year. In both cases, the rate of leaving employment increases around the 30 year mark. There is a correlation between higher education and better paying jobs. The behavior of the more educated employees could be due to the positions they hold being better paying. There is typically no reason for a person to leave employment for a worse paying job. Conversely, the less educated might be more enticed leave employment with the Air Force if they find a better paying job in the private sector.

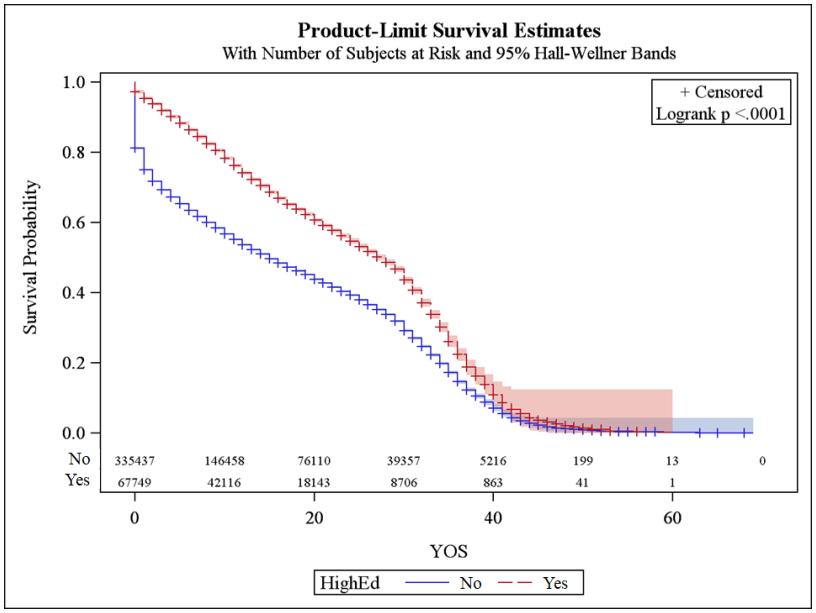

Figure 23: Survival Curve by Education Level

The survival curve for prior military service is shown in Figure 24. The logrank test returns a p-value of <0.0001. The results are similar to the education level chart in that people with prior service are not likely to leave employment during the first year while roughly 25% of people without prior service will leave. The rate of leaving employment spikes around 30 years for both groups. This result is confirmed when considering the odds ratios. Employees with prior military service have a higher odds ratio than those without prior military service. This behavior could be due to prior military employees being more acclimated to the workload and structure that comes with working in a military environment and potentially have a greater sense

of patriotism.

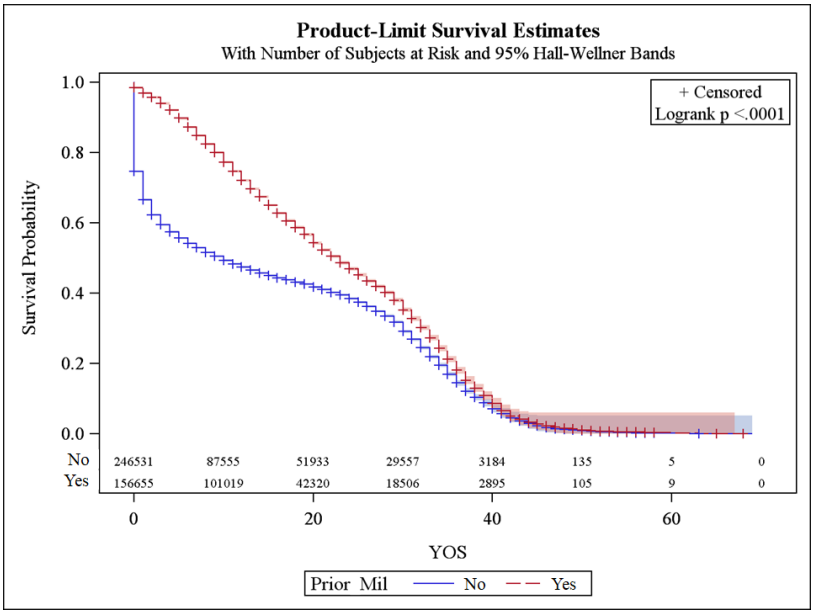

Figure 24: Survival Curve by Prior Military Status

Lastly, Figure 25 displays the survival curve for the four occupational series of interest. The 0800 occupational series group preformed the best with very few leaving during the first year of employment and the survival probability staying constant until around 30 years where it is expected that the rate of leaving employment spikes. The 1700, 3500, and 7400 occupational series groups performed similarly as each had about 50% of their employees leave in the first year.

A means of identifying problematic subdivisions of the population can be found by utilizing the survival curves. Each career field or demographic of the population is different and has varying reasons as to why they might have a retention issue. Attempting to clarify the reasons for this behavior is beyond the scope of this thesis. Only a means of identifying the trends and disparities is presented.

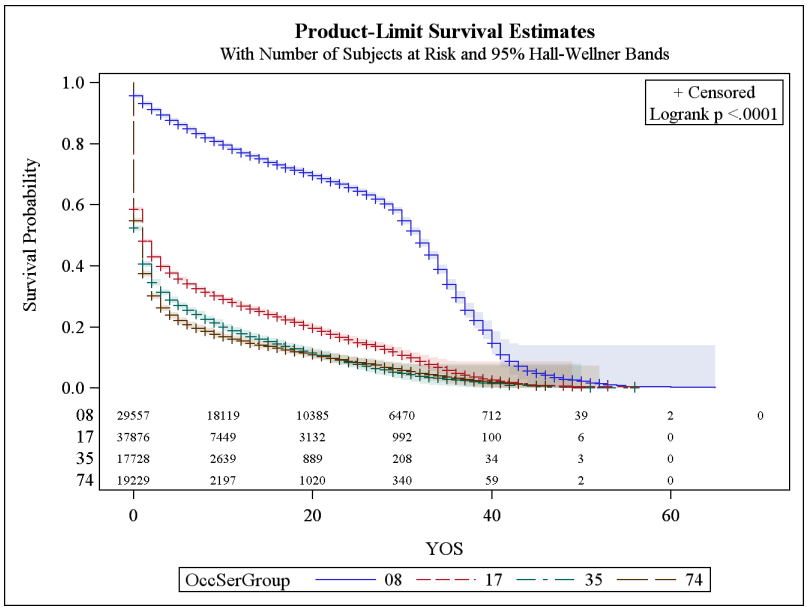

Figure 25: Survival Curve by Occupational Series

5.3 Cox Proportional Hazard Regression

The Cox proportional hazards model is used to estimate the strength and significance of the effects. This is accomplished in SAS via the PROC PHREG function. The code for this is in Appendix E. An initial model was created using a stepwise procedure. The entering criteria was $\alpha = 0.20$ and the exiting criteria was $\alpha = 0.05$. This parameter setting forces only the effects and interactions with a p-value less than 0.20 to enter the regression model while kicking out those where the p-value rises above 0.05 given other effects are in the model. The Wald Chi-Square and p-values from the resulting model are shown in Table 6. Table 7 displays the estimate of the effect, the standard error, and the corresponding p-value for the breakdown of the parameter estimates. The overall model has a p-value of <.0001.

Table 6: Cox Stepwise Model Results

Effect	Wald Chi-Square	p-value
AGE	1627.4243	<.0001

GENDER	355.2342	<.0001
Race_Grp	131.9684	<.0001
HighEd	275.4784	<.0001
Prior_Mil	224.5986	<.0001
AGE*GENDER	316.7226	<.0001
AGE*Race_Grp	122.3004	<.0001
AGE*HighEd	292.4252	<.0001
AGE*Prior_Mil	521.5993	<.0001
GENDER*Race_Grp	50.6132	<.0001
GENDER*Prior_Mil	18.5615	<.0001
Race_Grp*HighEd	76.8291	<.0001
Race_Grp*Prior_Mil	32.9235	<.0001
HighEd*Prior_Mil	281.477	<.0001
AGE*GENDER*Race_Grp	52.8791	<.0001
AGE*GENDER*Prior_Mil	136.7701	<.0001
AGE*Race_Grp*HighEd	86.0901	<.0001
AGE*Race_Grp*Prior_Mil	33.5442	<.0001
AGE*HighEd*Prior_Mil	99.333	<.0001
GENDER*Race_Grp*Prior_Mil	112.4531	<.0001
Race_Grp*HighEd*Prior_Mil	36.0256	<.0001

Table 7: Cox Stepwise Model Results Breakdown

Label	Estimate	Std Error	p-value
Age	-0.03602	0.000893	<.0001
Gender (F)	0.7731	0.04102	<.0001

Race (Am Ind/Ak Nat)	0.56347	0.37935	0.1374
Race (Asian)	0.6145	0.16813	0.0003
Race (Blk/Afr Am)	0.94101	0.09286	<.0001
Race (Declined)	-1.00706	0.54111	0.0627
Race (Hisp/Lat)	0.38865	0.1303	0.0029
Race (Multiple)	0.77954	0.19905	<.0001
Race (Nat Haw/Oth Pac Isl)	1.34264	0.5697	0.0184
HighEd (0)	0.83531	0.05033	<.0001
Prior_Mil (0)	0.88946	0.05935	<.0001
Gender (F)*Age	-0.01508	0.0008473	<.0001
Race (Am Ind/Ak Nat)*Age	-0.00939	0.00683	0.1693
Race (Asian)*Age	-0.00959	0.00321	0.0028
Race (Blk/Afr Am)*Age	-0.01681	0.00177	<.0001
Race (Declined)*Age	0.01007	0.00887	0.2564
Race (Hisp/Lat)*Age	-0.00945	0.0025	0.0002
Race (Multiple)*Age	-0.01683	0.00393	<.0001
Race (Nat Haw/Oth Pac Isl)*Age	-0.02715	0.01147	0.018
HighEd (0)*Age	-0.01586	0.0009275	<.0001
Prior_Mil (0)*Age	-0.02584	0.00113	<.0001
Gender (F)*Race (Am Ind/Ak Nat)	0.14128	0.19581	0.4706
Gender (F)*Race (Asian)	-0.38401	0.09458	<.0001
Gender (F)*Race (Blk/Afr Am)	-0.08108	0.05149	0.1154
Gender (F)*Race (Declined)	0.08803	0.20086	0.6612
Gender (F)*Race (Hisp/Lat)	-0.29736	0.06422	<.0001
Gender (F)*Race (Multiple)	-0.36416	0.09474	0.0001

Gender (F)*Race (Nat Haw/Oth Pac Isl)	-0.44328	0.22923	0.0531
Gender (F)*Prior_Mil (0)	-0.18366	0.04263	<.0001
Race (Am Ind/Ak Nat)*HighEd (0)	-0.75526	0.36268	0.0373
Race (Asian)*HighEd (0)	-0.54813	0.14854	0.0002
Race (Blk/Afr Am)*HighEd (0)	-0.59889	0.08916	<.0001
Race (Declined)*HighEd (0)	-1.05842	0.4394	0.016
Race (Hisp/Lat)*HighEd (0)	-0.36383	0.12719	0.0042
Race (Multiple)*HighEd (0)	-0.64678	0.19178	0.0007
Race (Nat Haw/Oth Pac Isl)*HighEd (0)	-1.36654	0.55751	0.0142
Race (Am Ind/Ak Nat)*Prior_Mil (0)	-0.19865	0.26372	0.4513
Race (Asian)*Prior_Mil (0)	-0.48906	0.14052	0.0005
Race (Blk/Afr Am)*Prior_Mil (0)	-0.06051	0.06685	0.3654
Race (Declined)*Prior_Mil (0)	1.83589	0.42308	<.0001
Race (Hisp/Lat)*Prior_Mil (0)	0.02336	0.08894	0.7928
Race (Multiple)*Prior_Mil (0)	0.06228	0.14412	0.6656
Race (Nat Haw/Oth Pac Isl)*Prior_Mil (0)	0.01418	0.34598	0.9673
HighEd (0)*Prior_Mil (0)	1.0065	0.05999	<.0001
Gender (F)*Race (Am Ind/Ak Nat)*Age	-0.00518	0.0036	0.1498
Gender (F)*Race (Asian)*Age	0.0062	0.00162	0.0001
Gender (F)*Race (Blk/Afr Am)*Age	-0.0007001	0.0009818	0.4758
Gender (F)*Race (Declined)*Age	-0.00208	0.00196	0.2887
Gender (F)*Race (Hisp/Lat)*Age	0.0054	0.0012	<.0001
Gender (F)*Race (Multiple)*Age	0.00596	0.00176	0.0007
Gender (F)*Race (Nat Haw/Oth Pac Isl)*Age	0.00928	0.00443	0.0361
Gender (F)*Prior_Mil (0)*Age	0.01048	0.000896	<.0001

Race (Am Ind/Ak Nat)*HighEd (0)*Age	0.01493	0.00657	0.0231
Race (Asian)*HighEd (0)*Age	0.01004	0.00279	0.0003
Race (Blk/Afr Am)*HighEd (0)*Age	0.01251	0.00172	<.0001
Race (Declined)*HighEd (0)*Age	0.018	0.00706	0.0108
Race (Hisp/Lat)*HighEd (0)*Age	0.00539	0.00246	0.0283
Race (Multiple)*HighEd (0)*Age	0.01331	0.00384	0.0005
Race (Nat Haw/Oth Pac Isl)*HighEd (0)*Age	0.03102	0.01133	0.0062
Race (Am Ind/Ak Nat)*Prior_Mil (0)*Age	0.00194	0.00431	0.6524
Race (Asian)*Prior_Mil (0)*Age	0.00965	0.00251	0.0001
Race (Blk/Afr Am)*Prior_Mil (0)*Age	0.00202	0.00115	0.0799
Race (Declined)*Prior_Mil (0)*Age	-0.02189	0.00665	0.001
Race (Hisp/Lat)*Prior_Mil (0)*Age	-0.00181	0.00148	0.2201
Race (Multiple)*Prior_Mil (0)*Age	-0.00359	0.0025	0.151
Race (Nat Haw/Oth Pac Isl)*Prior_Mil (0)*Age	0.00042	0.00557	0.9399
HighEd (0)*Prior_Mil (0)*Age	-0.01146	0.00115	<.0001
Gender (F)*Race (Am Ind/Ak Nat)*Prior_Mil (0)	0.28634	0.11661	0.0141
Gender (F)*Race (Asian)*Prior_Mil (0)	0.36378	0.06675	<.0001
Gender (F)*Race (Blk/Afr Am)*Prior_Mil (0)	-0.23597	0.03143	<.0001
Gender (F)*Race (Declined)*Prior_Mil (0)	-0.21865	0.17113	0.2014
Gender (F)*Race (Hisp/Lat)*Prior_Mil (0)	0.06933	0.04291	0.1062
Gender (F)*Race (Multiple)*Prior_Mil (0)	0.10619	0.06523	0.1035
Gender (F)*Race (Nat Haw/Oth Pac Isl)*Prior_Mil (0)	-0.04403	0.15061	0.77
Race (Am Ind/Ak Nat)*HighEd (0)*Prior_Mil (0)	-0.06155	0.15697	0.695
Race (Asian)*HighEd (0)*Prior_Mil (0)	0.18967	0.07485	0.0113
Race (Blk/Afr Am)*HighEd (0)*Prior_Mil (0)	0.16464	0.03986	<.0001

Race (Declined)*HighEd (0)*Prior_Mil (0)	-0.182	0.19181	0.3427
Race (Hisp/Lat)*HighEd (0)*Prior_Mil (0)	0.22037	0.05699	0.0001
Race (Multiple)*HighEd (0)*Prior_Mil (0)	-0.03848	0.08908	0.6658
Race (Nat Haw/Oth Pac Isl)*HighEd (0)*Prior_Mil (0)	0.24929	0.25088	0.3204

With this model all possible survival curves can be established. To demonstrate the breadth of possibilities, baseline covariates were created and plugged into the model for varying measures of the effects. Figure 26 displays all of these possibilities. Readers should bear in mind that just because a curve exists does not mean that it is very likely. The population of people that a curve would represent could apply to a very small subset of the population.

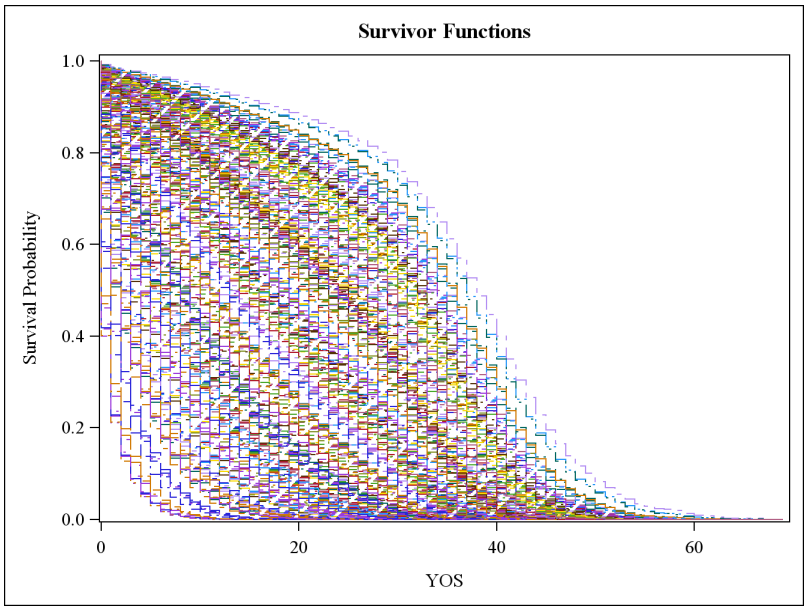

Figure 26: Breadth of Possible Survival Curves

Figure 27 displays the best and worst possible scenarios according to the odds ratios calculated by logistic regression. Covariate set 1 represents a 60-year-old, multi-racial, male with a higher education degree and prior military service. Covariate set 2 represents a 20-year-old, Black, female with no higher education degree or prior

military service. The two survival curves show the stark difference in the employment times of individuals. The first person is very likely to stay in the service for a long time while the other is very likely to leave in the first year.

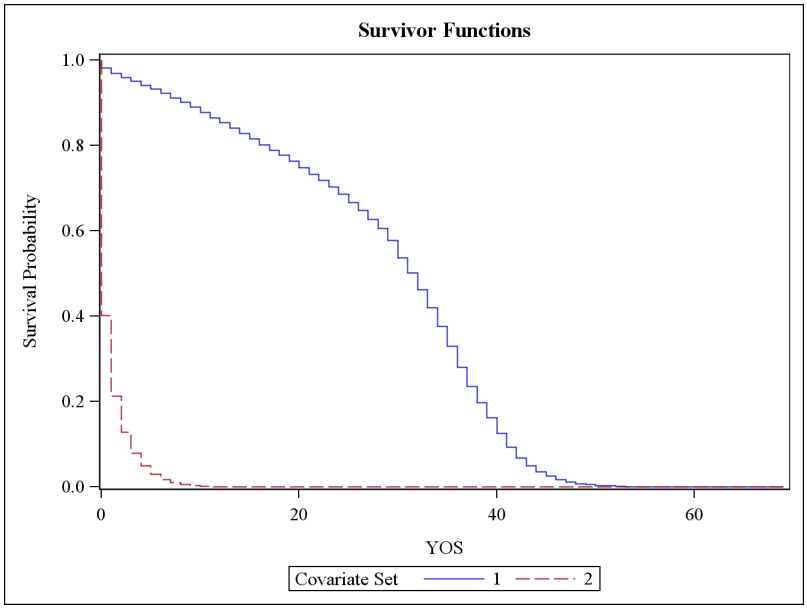

Figure 27: Best and Worst Survival Curves (Using Odds Ratios)

Plots of the Martingale and Deviance residuals are shown in Figures 28 and 29, respectively. The Martingale residual plot suggests that a few data points are potential outliers. 131 observations are shown to have a residual value less than -5. 85% of the outliers had at most one year of service with no prior military service and no higher education degree. No substantial reasons were found that warranted removing the observations from the analysis. The Deviance residual plot shows some issues with non-constant variance. The spread of the residuals shrinks as the predictions become larger. This skewness and pattern of the residuals is not unprecedented given the categorical nature of the model terms, so no remedial action is taken.

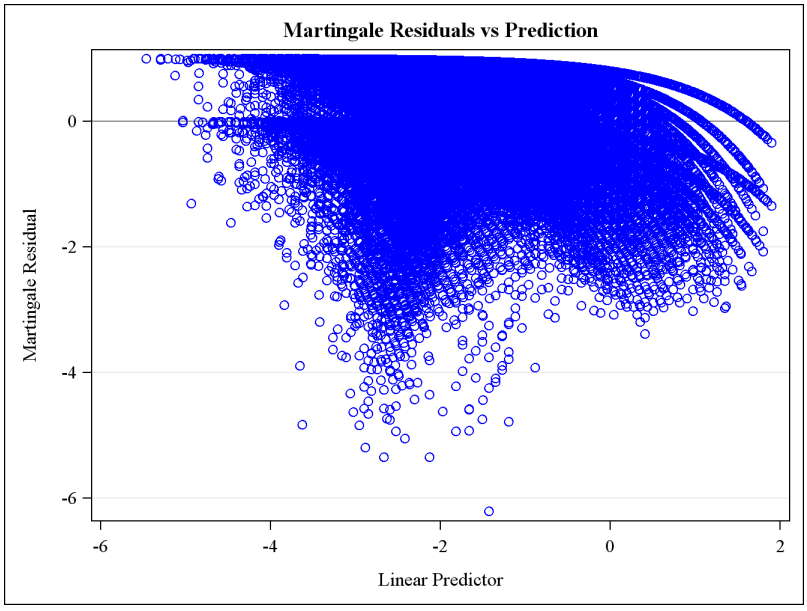

Figure 28: Martingale Residual Plot

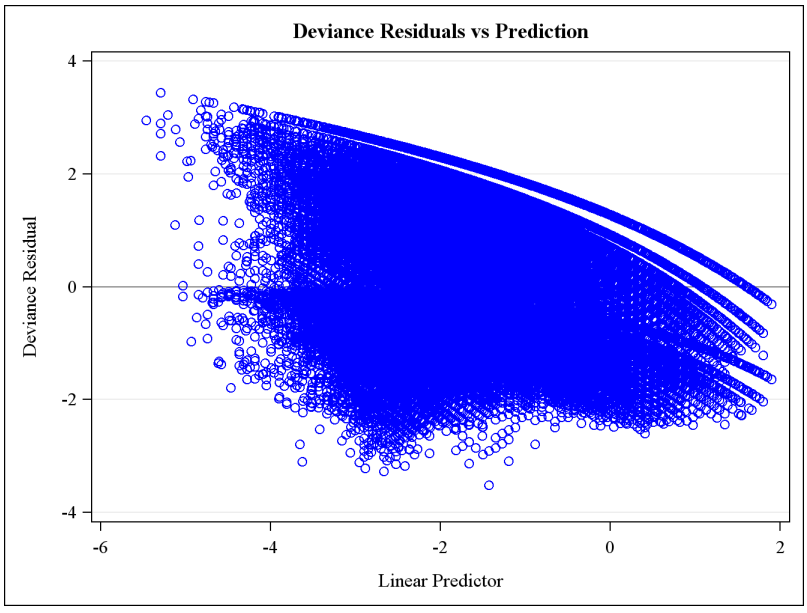

Figure 29: Deviance Residual Plot

VI. Conclusion

6.1 Limitations

The data used in this analysis concentrates on a few select demographics. With access to more data, a better understanding of the workforce could be obtained. Many prior theses and research have included variables concerned with the family such as marital status and number of dependents. Other variables that could provide useful information are economic status and political affiliation. While this information would likely be hard to obtain, the reasoning behind the collection of the data remains the same — to enhance the prediction of employees retainability.

The data itself are understandably messy given the number of people employed by the government and the number of technicians allowed to alter the database. This forces pre-cleaning of the data to remove impossibilities. While every effort to correct mistakes was taken, some lesser known mistakes could have slipped through the cracks. Efforts by the data provider and the recipient to perform a more rigorous and complete cleaning should be attempted before another similar analysis is conducted.

Civilian retention has not been studied much by the Air Force, leading to potentially poor management of the civilian workforce. This, coupled with the inability to hire workers efficiently, has led to some shortcomings in the workforce. Policies and objectives that target certain groups of people, such as only females or employees with no prior military service, can never be used because of the perception of being biased towards the other groups of people. Instead, improved or modified department-level programs and incentives should be offered to those where the need is the greatest and the retention is lowest.

6.2 Key Takeaways

Logistic Regression results indicate age, gender, race, education level, prior military service, and years of service are all significant in determining the likelihood of retention. The analysis shows that there is a noticeable difference between the white-collar and laborer positions in relation to the significance of higher education on the likelihood of retention. The odds ratios indicate males retain better than females, Multi-racial employees are the most likely race group to retain, Black employees are the least likely to retain, individuals with a Master's degree or PhD are more likely to retain than people without those degrees, and employees with prior military service are more likely to retain over employees without. Several occupational series groups were analyzed. Some groups showed conflicts with the overall population's results leading to the conclusion that each occupational series needs to be analyzed separately in order to accurately account for the nuances between them.

Survival analysis was used to handle the censored nature of the data. The overall Kaplan-Meier survival curve, shown in Figure 19, indicates that roughly 16% of new-hires leave employment with the Air Force in the first year. A more in-depth study of this phenomenon should be conducted to determine the root causes of such an exodus. Roughly 47% of all employees will stay 20 years and around 8% will make it to 40 years. Similar results to logistic regression were found when comparing the various breakouts of employees. Males were more likely to retain than females, highly educated and prior military employees were more likely to retain over their counterparts, and some occupational series retained better than others.

The Cox proportional hazards model was created using a stepwise procedure and ended with 21 statistically significant main effects and interactions. The martingale residual plot shows some outliers, but no concrete reasoning warranted their removal. The model was used to show the breadth of possible survival curves in Figure 26.

Figure 27 shows the best and worse performing survival curves according to the results of the odds ratios. The Cox model shows that there is a wide range of survival curves that exist. Future retention-related decisions can use this information as a basis for trying to target and improve the weakest performing groups.

6.3 Future Research

A possible avenue to extend this research investigates whether the economy has influence in the retention of employees. Typically, when the economy is booming, people are more inclined to leave the service and find employment in the private sector. This is because of the system the government uses to pay its employees. People are given a grade or rank and are paid accordingly. In the private sector, pay is roughly based on experience or work ethic. If people believe they can make more money by leaving the government then it might be worthwhile to study the effects of the economy on retention in the civilian workforce.

In a similar vein, the current political environment could also be analyzed for any influence on retention. Trends in the political landscape could also push people to seek employment with the private sector. Admittedly, this might be too cumbersome of a project to quantify, especially if general trends can not be established. As well, retention could be analyzed as a function of time. Has retention increased or decreased over the years and if it has, can we link these changes to specific events or policies?

6.4 Conclusion

Logistic regression is used as a means of identifying high risk groups. Odds ratios can be established that indicate relative to the other breakouts the likelihood of retaining. The results showed that men with advanced academic degrees and prior military service fared better than people without those qualities. The results of the

logistic regression and survival analysis had similar outcomes, but survival analysis should be used in the case where censored data is concerned. Logistic regression is not explicitly designed to analyze this type of data.

By using survival analysis, the behavior of certain groups of people can be found through time, even with censored data. Groups of people that perform very poorly in the first few years could be examined closer and surveyed for the reason of their departure. If too many people leave from one year group, this could lead to bathtub effects where there are not enough experienced people to fill the more senior roles when those positions become vacant. The civilian situation is better than the officer and enlisted ranks where people are placed into higher ranking positions within the rank structure. Civilian positions can be filled by any willing and qualified individual within and outside of the government so the threat of a bathtub is more manageable. Having a healthy workforce includes having a good ratio of new to experienced personnel which is different for every occupational series and grouping, but is vital to ensuring the mission is accomplished.

The Cox proportional hazards model is a stepping stone to a more accurate model. As discussed previously, more variables should be added if they provide a better estimate of the number of years of service and the data should be scrutinized to a higher degree to weed out any impossibilities. This method is semi-parametric, so it requires the use of explanatory variables which is a downside when the Kaplan-Meier estimate produces similar results without the need to use explanatory variables.

Programs and incentives that would increase retention for high-priority or at-risk groups should be considered where there is large turnover in the first few years of employment. The inverse could also happen, where employees are staying around longer than is necessary for a healthy workforce. It might be more economical to incentivize a person with 25 years of service to leave employment, replace that newly

vacant position with a new person from within the organization, and fill that newly opened slot with an intern or a new-hire. Regardless, the health of the civilian workforce is vital for a well run Air Force.

Appendix A. SAS Code for Loss Files

```
/*Insert all file date extensions needed (in order).*/
%Let file1=200912;
%Let file2=201001;
%Let file3=201002;

⋮

%Let file121=201912;

Libname CIVINV "<insert filepath to inventory folder>";
Libname CIVLOSS "<insert filepath to loss folder>";

%MACRO CREATE_LOSS_FILES;
    %Do i = 1 %To 120;
        %Let j = %sysevalf(&i. + 1);
        Data Civloss.Civloss&&file&j..;
            Merge Civinv.Civinv&&file&i.. (IN=A)
                  Civinv.civinv&&file&j.. (IN=B);
            By SSAN;
            If A and ~B Then Output;
        Run;
    %End;
%MEND;
%CREATE_LOSS_FILES;
```

Appendix B. SAS Code for Combined File

```
/*Setting the library for the civilian data files.*/
Libname CIVINV "<insert filepath to inventory folder>";
Libname CIVLOSS "<insert filepath to loss folder>";

%Let start_yr = 2010;
%Let end_yr = 2019;

/*This macro will create a combined file*/
/*for the entire date range supplied.*/
%MACRO LOOP_ALL_YEARS_COMB;
    /*Stacking all of the inventory files then only*/
    /* keeping the last record for each person.*/
    Data Civinv_&start_yr._&end_yr.;
        Set
        %Do i = &start_yr. %To &end_yr.;
            Civinv.civinv&i.:
        %End;
        ;
        By SSAN;
        If last.SSAN;
        Retain = 1;
    Run;

    /*Stacking all of the loss files then only*/
    /* keeping the last record for each person.*/
    Data Civloss_&start_yr._&end_yr.;
        Set
        %Do i = &start_yr. %To &end_yr.;
            Civloss.civloss&i.:
        %End;
        ;
        By SSAN;
        If last.SSAN;
        Retain = 0;
    Run;

    /*Combining the inventory and loss files and*/
    /* keeping the last record per person.*/
    Data Civcomb_&start_yr._&end_yr.;
        Set Civinv_&start_yr._&end_yr.
```

```
            Civloss_&start_yr._&end_yr.;
        By SSAN;
        If last.SSAN;
    Run;

    /*Altering a couple of variables.*/
    Data Civcomb_&start_yr._&end_yr.;
        Set Civcomb_&start_yr._&end_yr.;

        /*Recoding high education level.*/
        /*Master's or higher = 1, else = 0*/
        If EdLevel IN ("MA", "PHD/PROF DEG") Then HighEd = 1;
            Else HighEd = 0;

        /*Recoding the prior military service.*/
        If AFPC_Prior_Mil = "NO MILITARY CREDIT" Then Prior_Mil = 0;
            Else Prior_Mil = 1;

        /*Recording the occupational series identifier by grouping.*/
        OccSerGroup = substr(OCCSER, 1, 2);

        /*Records with inconsistent ages, YOS, very old individuals,*/
        /* and bad OCCSER codes removed.*/
        If 16 <= AGE <= 90;
        If YOS >= 0;
        If (AGE - 15) >= YOS;
        If OccSerGroup NE "0W";
    Run;
%MEND;
%LOOP_ALL_YEARS_COMB;
```

Appendix C. SAS Code for Logistic Regression

```
/*List all of the occupational series.*/
%Let OccSerGroup1 = 00;
%Let OccSerGroup2 = 01;
%Let OccSerGroup3 = 02;
%Let OccSerGroup4 = 03;
%Let OccSerGroup5 = 04;
%Let OccSerGroup6 = 05;
%Let OccSerGroup7 = 06;
%Let OccSerGroup8 = 07;
%Let OccSerGroup9 = 08;
%Let OccSerGroup10 = 09;
%Let OccSerGroup11 = 10;
%Let OccSerGroup12 = 11;
%Let OccSerGroup13 = 12;
%Let OccSerGroup14 = 13;
%Let OccSerGroup15 = 14;
%Let OccSerGroup16 = 15;
%Let OccSerGroup17 = 16;
%Let OccSerGroup18 = 17;
%Let OccSerGroup19 = 18;
%Let OccSerGroup20 = 19;
%Let OccSerGroup21 = 20;
%Let OccSerGroup22 = 21;
%Let OccSerGroup23 = 22;
%Let OccSerGroup24 = 25;
%Let OccSerGroup25 = 26;
%Let OccSerGroup26 = 28;
%Let OccSerGroup27 = 31;
%Let OccSerGroup28 = 33;
%Let OccSerGroup29 = 34;
%Let OccSerGroup30 = 35;
%Let OccSerGroup31 = 36;
%Let OccSerGroup32 = 37;
%Let OccSerGroup33 = 38;
%Let OccSerGroup34 = 39;
%Let OccSerGroup35 = 41;
%Let OccSerGroup36 = 42;
%Let OccSerGroup37 = 43;
%Let OccSerGroup38 = 44;
%Let OccSerGroup39 = 46;
```

```
%Let OccSerGroup40 = 47;
%Let OccSerGroup41 = 48;
%Let OccSerGroup42 = 50;
%Let OccSerGroup43 = 52;
%Let OccSerGroup44 = 53;
%Let OccSerGroup45 = 54;
%Let OccSerGroup46 = 57;
%Let OccSerGroup47 = 58;
%Let OccSerGroup48 = 65;
%Let OccSerGroup49 = 66;
%Let OccSerGroup50 = 69;
%Let OccSerGroup51 = 70;
%Let OccSerGroup52 = 73;
%Let OccSerGroup53 = 74;
%Let OccSerGroup54 = 76;
%Let OccSerGroup55 = 82;
%Let OccSerGroup56 = 86;
%Let OccSerGroup57 = 88;

%MACRO PERFORM_LOG_REG;
    /*Performing logistic regression on the entire AF.*/
    ODS GRAPHICS ON;
    ODS RTF FILE="C:\Users\wwilson1\Documents\Wilson Thesis\
                  Logistic Regression\LogReg_&start_yr._&end_yr..doc";
        Proc Logistic Data = Civcomb_&start_yr._&end_yr.;
        Class Gender Race_Grp HighEd Prior_Mil;
        Model Retain(Event = "1") =
            Age Gender Race_Grp HighEd Prior_Mil YOS;
        Run;
    ODS RTF CLOSE;
    ODS GRAPHICS OFF;

    /*Performing logistic regression on each occupational series.*/
    %Do i = 1 %To 57;
        ODS GRAPHICS ON;
        ODS RTF FILE="C:\Users\wwilson1\Documents\Wilson Thesis\
                      Logistic Regression\
                      LogReg_&start_yr._&end_yr._&&OccSerGroup&i...doc";
            Proc Logistic Data = Civcomb_&start_yr._&end_yr.
                          (Where=(OccSerGroup = "&&OccSerGroup&i.."));
            Class Gender Race_Grp HighEd Prior_Mil;
            Model Retain(Event = "1") =
                Age Gender Race_Grp HighEd Prior_Mil YOS;
```

```
            Run;
        ODS RTF CLOSE;
        ODS GRAPHICS OFF;
    %End;
%MEND;
%PERFORM_LOG_REG;

/*Checking to see if the continuous vars are normally dist.*/
ODS GRAPHICS ON;
ODS RTF FILE="C:\Users\wwilson1\Documents\Wilson Thesis\
             Logistic Regression\Norm_&start_yr._&end_yr._Age.doc";
    Proc Univariate
        Data = Civcomb_&start_yr._&end_yr. plots;
        Var Age;
    Run;
ODS RTF CLOSE;
ODS GRAPHICS OFF;

ODS GRAPHICS ON;
ODS RTF FILE="C:\Users\wwilson1\Documents\Wilson Thesis\
             Logistic Regression\Norm_&start_yr._&end_yr._YOS.doc";
    Proc Univariate
        Data = Civcomb_&start_yr._&end_yr. plots;
        Var YOS;
    Run;
ODS RTF CLOSE;
ODS GRAPHICS OFF;


/*Checking the correlation values between the continuous variables.*/
/*Pearson Correlation Coefficient*/
ODS GRAPHICS ON;
ODS RTF FILE="C:\Users\wwilson1\Documents\Wilson Thesis\
             Logistic Regression\Corr_Cont_&start_yr._&end_yr..doc";
    Proc Corr
        Data = Civcomb_&start_yr._&end_yr.;
        Var Age YOS;
    Run;
ODS RTF CLOSE;
ODS GRAPHICS OFF;
```

```
/*Checking the correlation values between the categorical variables.*/
/*Phi-Coefficient for 2x2 contingency tables.*/
/*Contingency Coefficient for tables larger than 2x2.*/
/* CC is adjustment to phi for larger tables.*/
%Let cat1 = Retain;
%Let cat2 = Gender;
%Let cat3 = Race_Grp;
%Let cat4 = HighEd;
%Let cat5 = Prior_Mil;

%MACRO CREATE_P_FOR_CAT;
    %Do i = 1 %To 5;
        %Do j = 1 %To 5;
            %If %sysevalf(&j. > &i.) %Then %Do;
                ODS GRAPHICS ON;
                ODS RTF FILE="C:\Users\wwilson1\Documents\Wilson Thesis\
                             Logistic Regression\
                             Corr_Cat_&i._&j._&start_yr._&end_yr..doc";
                    Proc Freq
                        Data = Civcomb_&start_yr._&end_yr.;
                        Table &&cat&i..*&&cat&j..
                        / nopercent norow nocol chisq;
                    Run;
                ODS RTF CLOSE;
                ODS GRAPHICS OFF;
            %End;
        %End;
    %End;
%MEND;
%CREATE_P_FOR_CAT;

/*Checking the correlation values between*/
/*the categorical and continuous variables.*/
/*Point-Biserial Correlation Coefficient*/
/* Special case of Pearson CC*/
/* Assumes continuous vars are norm dist and homoscedastic*/
Data CivComb_w_binary;
    Set Civcomb_&start_yr._&end_yr.;
    Gender_Trans = (Gender="M");
    If Race_Grp = "Am Ind/Ak Nat"       Then Race_Trans = 0;
    If Race_Grp = "Asian"               Then Race_Trans = 1;
    If Race_Grp = "Blk/Afr Am"          Then Race_Trans = 2;
    If Race_Grp = "Declined"            Then Race_Trans = 3;
```

```
    If Race_Grp = "Hisp/Lat"            Then Race_Trans = 4;
    If Race_Grp = "Multiple"            Then Race_Trans = 5;
    If Race_Grp = "Nat Haw/Oth Pac Isl" Then Race_Trans = 6;
    If Race_Grp = "White"               Then Race_Trans = 7;
Run;

ODS GRAPHICS ON;
ODS RTF FILE="C:\Users\wwilson1\Documents\Wilson Thesis\
            Logistic Regression\Corr_Cat_Cont_&start_yr._&end_yr..doc";
    Proc Corr
        Data = CivComb_w_binary;
        Var Age YOS;
        With Retain Gender_Trans Race_Trans HighEd Prior_Mil;
    Run;
ODS RTF CLOSE;
ODS GRAPHICS OFF;
```

Appendix D. SAS Code for Kaplan-Meier Survival Analysis

```
%MACRO CREATE_DEMO_GRAPHS;
    /*Creating the histogram/smoothed pdf curve with people*/
    /*that have left employment.*/
    ODS GRAPHICS ON;
    ODS RTF FILE="C:\Users\wwilson1\Documents\Wilson Thesis\
                  Survival Analysis\Demo\PDF_&start_yr._&end_yr..doc";
    Proc Univariate
        Data = Civcomb_&start_yr._&end_yr.(where=(Retain=0));
        Var YOS;
        Histogram YOS/kernel endpoints=(0 to 60 by 1);
    Run;
    ODS RTF CLOSE;
    ODS GRAPHICS OFF;

    /*Creating the CDF plot with people that have left employment.*/
    ODS GRAPHICS ON;
    ODS RTF FILE="C:\Users\wwilson1\Documents\Wilson Thesis\
                  Survival Analysis\Demo\CDF_&start_yr._&end_yr..doc";
    Proc Univariate
        Data = Civcomb_&start_yr._&end_yr.(where=(Retain=0));
        Var YOS;
        CDFPLOT YOS;
    Run;
    ODS RTF CLOSE;
    ODS GRAPHICS OFF;

    /*Creating the KM curve with people that have left employment.*/
    ODS GRAPHICS ON;
    ODS RTF FILE="C:\Users\wwilson1\Documents\Wilson Thesis\
                  Survival Analysis\Demo\KM_&start_yr._&end_yr..doc";
    ODS EXCLUDE ProductLimitEstimates;
    Proc Lifetest
        Data = Civcomb_&start_yr._&end_yr. (where=(Retain=0))
        Plots = Survival(atrisk cb test);
        Time YOS*Retain(1);
    Run;
    ODS RTF CLOSE;
    ODS GRAPHICS OFF;
```

```
/*Creating the hazard rate function with people*/
/*that have left employment.*/
ODS GRAPHICS ON;
ODS RTF FILE="C:\Users\wwilson1\Documents\Wilson Thesis\
              Survival Analysis\Demo\
              Hazard_&start_yr._&end_yr..doc";
ODS EXCLUDE ProductLimitEstimates;
Proc Lifetest
    Data = Civcomb_&start_yr._&end_yr. (where=(Retain=0))
    Plots = Hazard;
    Time YOS*Retain(1);
Run;
ODS RTF CLOSE;
ODS GRAPHICS OFF;

/*Creating the cumulative hazard rate function with people*/
/*that have left employment.*/
ODS GRAPHICS ON;
ODS RTF FILE="C:\Users\wwilson1\Documents\Wilson Thesis\
              Survival Analysis\Demo\
              CumHazard_&start_yr._&end_yr..doc";
ODS EXCLUDE ProductLimitEstimates;
ODS OUTPUT ProductLimitEstimates = ple;
Proc Lifetest
    Data = Civcomb_&start_yr._&end_yr. (where=(Retain=0))
    Nelson;
    Time YOS*Retain(1);
Run;
Proc Sgplot
    Data = ple;
    Series x=YOS y=CumHaz;
    Title "Cumulative Hazard Rate Function";
Run;
ODS RTF CLOSE;
ODS GRAPHICS OFF;

/*Creating the box plot of the data. Need to sort first.*/
Proc Sort
    Data = Civcomb_&start_yr._&end_yr.;
    By Retain;
Run;
```

```
    ODS GRAPHICS ON;
    ODS RTF FILE="C:\Users\wwilson1\Documents\Wilson Thesis\
                  Survival Analysis\Demo\
                  BoxPlot_&start_yr._&end_yr..doc";
    Proc BoxPlot
        Data = Civcomb_&start_yr._&end_yr.;
        Plot YOS*Retain;
        insetgroup min Q1 Q2 Q3 max stddev mean/
            header = "Overall Statistics";
    Run;
    ODS RTF CLOSE;
    ODS GRAPHICS OFF;
%MEND;
%CREATE_DEMO_GRAPHS;


/*These are all of the stratifications looked at.*/
%Let Strata1 = Gender;
%Let Strata2 = Race_Grp;
%Let Strata3 = HighEd;
%Let Strata4 = Prior_Mil;
/*These are all of the occupational series.*/
%Let OccSerGroup1 = 00;
%Let OccSerGroup2 = 01;
%Let OccSerGroup3 = 02;
%Let OccSerGroup4 = 03;
%Let OccSerGroup5 = 04;
%Let OccSerGroup6 = 05;
%Let OccSerGroup7 = 06;
%Let OccSerGroup8 = 07;
%Let OccSerGroup9 = 08;
%Let OccSerGroup10 = 09;
%Let OccSerGroup11 = 10;
%Let OccSerGroup12 = 11;
%Let OccSerGroup13 = 12;
%Let OccSerGroup14 = 13;
%Let OccSerGroup15 = 14;
%Let OccSerGroup16 = 15;
%Let OccSerGroup17 = 16;
%Let OccSerGroup18 = 17;
%Let OccSerGroup19 = 18;
```

```
%Let OccSerGroup20 = 19;
%Let OccSerGroup21 = 20;
%Let OccSerGroup22 = 21;
%Let OccSerGroup23 = 22;
%Let OccSerGroup24 = 25;
%Let OccSerGroup25 = 26;
%Let OccSerGroup26 = 28;
%Let OccSerGroup27 = 31;
%Let OccSerGroup28 = 33;
%Let OccSerGroup29 = 34;
%Let OccSerGroup30 = 35;
%Let OccSerGroup31 = 36;
%Let OccSerGroup32 = 37;
%Let OccSerGroup33 = 38;
%Let OccSerGroup34 = 39;
%Let OccSerGroup35 = 41;
%Let OccSerGroup36 = 42;
%Let OccSerGroup37 = 43;
%Let OccSerGroup38 = 44;
%Let OccSerGroup39 = 46;
%Let OccSerGroup40 = 47;
%Let OccSerGroup41 = 48;
%Let OccSerGroup42 = 50;
%Let OccSerGroup43 = 52;
%Let OccSerGroup44 = 53;
%Let OccSerGroup45 = 54;
%Let OccSerGroup46 = 57;
%Let OccSerGroup47 = 58;
%Let OccSerGroup48 = 65;
%Let OccSerGroup49 = 66;
%Let OccSerGroup50 = 69;
%Let OccSerGroup51 = 70;
%Let OccSerGroup52 = 73;
%Let OccSerGroup53 = 74;
%Let OccSerGroup54 = 76;
%Let OccSerGroup55 = 82;
%Let OccSerGroup56 = 86;
%Let OccSerGroup57 = 88;

%MACRO KAPLAN_MEIER;
    /*Censored entries are ones that have not left employment.*/
    /*(AKA people currently working, retain=1)*/
    /*Creating the Kaplan-Meier estimate of the*/
```

```
/*survival curve for all of the data.*/
/*The "exclude" line suppresses the table from being created.*/

ODS GRAPHICS ON;
ODS RTF FILE="C:\Users\wwilson1\Documents\Wilson Thesis\
             Survival Analysis\Survival Curves\
             KM_All_&start_yr._&end_yr..doc";
ODS EXCLUDE ProductLimitEstimates;
Proc Lifetest
    Data = Civcomb_&start_yr._&end_yr.
    Outs = All
    Plots = Survival(atrisk cb test);
    Time YOS*Retain(1);
Run;
ODS RTF CLOSE;
ODS GRAPHICS OFF;

/*Creating the Kaplan-Meier estimate of the survival curve*/
/*for certain demographic variables.*/
%Do i = 1 %To 4;
    ODS GRAPHICS ON;
    ODS RTF FILE="C:\Users\wwilson1\Documents\Wilson Thesis\
                  Survival Analysis\Survival Curves\
                  KM_&&Strata&i.._&start_yr._&end_yr..doc";
    ODS EXCLUDE ProductLimitEstimates;
    Proc Lifetest
        Data = Civcomb_&start_yr._&end_yr.
        Outs = All_&&Strata&i..
        Plots = Survival(atrisk cb test);
        Strata &&Strata&i..;
        Time YOS*Retain(1);
    Run;
    ODS RTF CLOSE;
    ODS GRAPHICS OFF;
%End;

/*Creating the Kaplan-Meier estimate of the survival curve*/
/*for all of the occupational series.*/
%Do i = 1 %To 57;
    ODS GRAPHICS ON;
    ODS RTF FILE="C:\Users\wwilson1\Documents\Wilson Thesis\
                  Survival Analysis\Survival Curves\
                  KM_&&OccSerGroup&i.._&start_yr._&end_yr..doc";
```

```
    ODS EXCLUDE ProductLimitEstimates;
    Proc Lifetest
        Data = Civcomb_&start_yr._&end_yr.
            (Where=(OccSerGroup="&&OccSerGroup&i.."))
        Outs = All_&&OccSerGroup&i..
        Plots = Survival(atrisk cb test);
        Time YOS*Retain(1);
    Run;
    ODS RTF CLOSE;
    ODS GRAPHICS OFF;
%End;

/*This will make graphs for race in panels.*/
ODS GRAPHICS ON;
ODS RTF FILE="C:\Users\wwilson1\Documents\Wilson Thesis\
             Survival Analysis\Survival Curves\
             KM_Race_Grp_v2_&start_yr._&end_yr..doc";
ODS EXCLUDE ProductLimitEstimates;
Proc Lifetest
    Data = Civcomb_&start_yr._&end_yr.
    Outs = All_Race_Grp_v2
    Plots = Survival(atrisk cb test strata=panel);
    Strata Race_Grp;
    Time YOS*Retain(1);
Run;
ODS RTF CLOSE;
ODS GRAPHICS OFF;

/*This will make graphs for the four OccSer's of interest.*/
ODS GRAPHICS ON;
ODS RTF FILE="C:\Users\wwilson1\Documents\Wilson Thesis\
             Survival Analysis\Survival Curves\
             KM_OccSerGroupings_&start_yr._&end_yr..doc";
ODS EXCLUDE ProductLimitEstimates;
Proc Lifetest
    Data = Civcomb_&start_yr._&end_yr.
        (Where=(OccSerGroup IN ("08", "17", "35", "74")))
    Outs = All_OccSerGroupings
    Plots = Survival(atrisk cb test);
    Strata OccSerGroup;
    Time YOS*Retain(1);
Run;
ODS RTF CLOSE;
```

```
    ODS GRAPHICS OFF;
%MEND;
%KAPLAN_MEIER;
```

Appendix E. SAS Code for Cox Proportional Hazards Model

```
/*Cox Regression model*/
/*Determining which covariates are important via stepwise.*/
ODS GRAPHICS ON;
ODS RTF FILE="C:\Users\wwilson1\Documents\Wilson Thesis\
              Survival Analysis\Survival Curves\Cox_StepwiseReg.doc";
Proc Phreg
    Data = Civcomb_&start_yr._&end_yr.
    Plots = survival;
    Class Gender Race_Grp HighEd Prior_Mil;
    Model YOS*Retain(1) = Age|Gender|Race_Grp|HighEd|Prior_Mil
        /Selection = Stepwise SLENTRY = 0.2 SLSTAY = 0.05 Details;
    Output OUT = Outp XBETA = Xbeta RESMART = Mart RESDEV = Dev;
Run;
ODS RTF CLOSE;
ODS GRAPHICS OFF;

/*Checking Martingale Residual plot*/
ODS GRAPHICS ON;
ODS RTF FILE="C:\Users\wwilson1\Documents\Wilson Thesis\
              Survival Analysis\Survival Curves\Cox_Martingale.doc";
TITLE "Martingale Residuals vs Prediction";
Proc Sgplot
    Data = Outp;
    YAXIS GRID;
    REFLINE 0 / AXIS = y;
    SCATTER Y = Mart X = Xbeta /
        markerattrs=(color=blue symbol=circle);
Run;
ODS RTF CLOSE;
ODS GRAPHICS OFF;

/*Checking Deviance Residual plot*/
ODS GRAPHICS ON;
ODS RTF FILE="C:\Users\wwilson1\Documents\Wilson Thesis\
              Survival Analysis\Survival Curves\Cox_Deviance.doc";
TITLE "Deviance Residuals vs Prediction";
Proc Sgplot
    Data = Outp;
    YAXIS GRID;
```

```
    REFLINE 0 / AXIS = y;
    SCATTER Y = Dev X = Xbeta /
        markerattrs=(color=blue symbol=circle);
Run;
ODS RTF CLOSE;
ODS GRAPHICS OFF;

/*Looking at the large magnitude martingale residuals.*/
Data Large_Mart;
Set Outp (Where=(Mart <= -5));
Run;

Proc Freq
    Data = Civcomb_&start_yr._&end_yr.;
    Tables Age / norow nocol nocum;
Run;
Proc Freq
    Data = Large_Mart;
    Tables Age / norow nocol nocum;
Run;

Proc Freq
    Data = Civcomb_&start_yr._&end_yr.;
    Tables Gender / norow nocol nocum;
Run;
Proc Freq
    Data = Large_Mart;
    Tables Gender / norow nocol nocum;
Run;

Proc Freq
    Data = Civcomb_&start_yr._&end_yr.;
    Tables Race_Grp / norow nocol nocum;
Run;
Proc Freq
    Data = Large_Mart;
    Tables Race_Grp / norow nocol nocum;
Run;

Proc Freq
    Data = Civcomb_&start_yr._&end_yr.;
    Tables HighEd / norow nocol nocum;
Run;
```

```
Proc Freq
    Data = Large_Mart;
    Tables HighEd / norow nocol nocum;
Run;

Proc Freq
    Data = Civcomb_&start_yr._&end_yr.;
    Tables Prior_Mil / norow nocol nocum;
Run;
Proc Freq
    Data = Large_Mart;
    Tables Prior_Mil / norow nocol nocum;
Run;

Proc Freq
    Data = Civcomb_&start_yr._&end_yr.;
    Tables YOS / norow nocol nocum;
Run;
Proc Freq
    Data = Large_Mart;
    Tables YOS / norow nocol nocum;
Run;

/*Creating a comprehensive list of possible covariates.*/
Data Covariates1;
    Length Race_Grp $19;
    Infile datalines dsd missover;
    Input Age Gender $ Race_Grp $ HighEd Prior_Mil;
    Datalines;
    20, F, Am Ind/Ak Nat, 0, 0
    20, F, Am Ind/Ak Nat, 0, 1
    20, F, Am Ind/Ak Nat, 1, 0
    20, F, Am Ind/Ak Nat, 1, 1
    20, F, Asian, 0, 0
    20, F, Asian, 0, 1
    20, F, Asian, 1, 0
    20, F, Asian, 1, 1
    20, F, Blk/Afr Am, 0, 0
    20, F, Blk/Afr Am, 0, 1
    20, F, Blk/Afr Am, 1, 0
    20, F, Blk/Afr Am, 1, 1
    20, F, Declined, 0, 0
    20, F, Declined, 0, 1
```

```
20, F, Declined, 1, 0
20, F, Declined, 1, 1
20, F, Hisp/Lat, 0, 0
20, F, Hisp/Lat, 0, 1
20, F, Hisp/Lat, 1, 0
20, F, Hisp/Lat, 1, 1
20, F, Multiple, 0, 0
20, F, Multiple, 0, 1
20, F, Multiple, 1, 0
20, F, Multiple, 1, 1
20, F, Nat Haw/Oth Pac Isl, 0, 0
20, F, Nat Haw/Oth Pac Isl, 0, 1
20, F, Nat Haw/Oth Pac Isl, 1, 0
20, F, Nat Haw/Oth Pac Isl, 1, 1
20, F, White, 0, 0
20, F, White, 0, 1
20, F, White, 1, 0
20, F, White, 1, 1
20, M, Am Ind/Ak Nat, 0, 0
20, M, Am Ind/Ak Nat, 0, 1
20, M, Am Ind/Ak Nat, 1, 0
20, M, Am Ind/Ak Nat, 1, 1
20, M, Asian, 0, 0
20, M, Asian, 0, 1
20, M, Asian, 1, 0
20, M, Asian, 1, 1
20, M, Blk/Afr Am, 0, 0
20, M, Blk/Afr Am, 0, 1
20, M, Blk/Afr Am, 1, 0
20, M, Blk/Afr Am, 1, 1
20, M, Declined, 0, 0
20, M, Declined, 0, 1
20, M, Declined, 1, 0
20, M, Declined, 1, 1
20, M, Hisp/Lat, 0, 0
20, M, Hisp/Lat, 0, 1
20, M, Hisp/Lat, 1, 0
20, M, Hisp/Lat, 1, 1
20, M, Multiple, 0, 0
20, M, Multiple, 0, 1
20, M, Multiple, 1, 0
20, M, Multiple, 1, 1
20, M, Nat Haw/Oth Pac Isl, 0, 0
```

```
20, M, Nat Haw/Oth Pac Isl, 0, 1
20, M, Nat Haw/Oth Pac Isl, 1, 0
20, M, Nat Haw/Oth Pac Isl, 1, 1
20, M, White, 0, 0
20, M, White, 0, 1
20, M, White, 1, 0
20, M, White, 1, 1
30, F, Am Ind/Ak Nat, 0, 0
30, F, Am Ind/Ak Nat, 0, 1
30, F, Am Ind/Ak Nat, 1, 0
30, F, Am Ind/Ak Nat, 1, 1
30, F, Asian, 0, 0
30, F, Asian, 0, 1
30, F, Asian, 1, 0
30, F, Asian, 1, 1
30, F, Blk/Afr Am, 0, 0
30, F, Blk/Afr Am, 0, 1
30, F, Blk/Afr Am, 1, 0
30, F, Blk/Afr Am, 1, 1
30, F, Declined, 0, 0
30, F, Declined, 0, 1
30, F, Declined, 1, 0
30, F, Declined, 1, 1
30, F, Hisp/Lat, 0, 0
30, F, Hisp/Lat, 0, 1
30, F, Hisp/Lat, 1, 0
30, F, Hisp/Lat, 1, 1
30, F, Multiple, 0, 0
30, F, Multiple, 0, 1
30, F, Multiple, 1, 0
30, F, Multiple, 1, 1
30, F, Nat Haw/Oth Pac Isl, 0, 0
30, F, Nat Haw/Oth Pac Isl, 0, 1
30, F, Nat Haw/Oth Pac Isl, 1, 0
30, F, Nat Haw/Oth Pac Isl, 1, 1
30, F, White, 0, 0
30, F, White, 0, 1
30, F, White, 1, 0
30, F, White, 1, 1
30, M, Am Ind/Ak Nat, 0, 0
30, M, Am Ind/Ak Nat, 0, 1
30, M, Am Ind/Ak Nat, 1, 0
30, M, Am Ind/Ak Nat, 1, 1
```

```
30, M, Asian, 0, 0
30, M, Asian, 0, 1
30, M, Asian, 1, 0
30, M, Asian, 1, 1
30, M, Blk/Afr Am, 0, 0
30, M, Blk/Afr Am, 0, 1
30, M, Blk/Afr Am, 1, 0
30, M, Blk/Afr Am, 1, 1
30, M, Declined, 0, 0
30, M, Declined, 0, 1
30, M, Declined, 1, 0
30, M, Declined, 1, 1
30, M, Hisp/Lat, 0, 0
30, M, Hisp/Lat, 0, 1
30, M, Hisp/Lat, 1, 0
30, M, Hisp/Lat, 1, 1
30, M, Multiple, 0, 0
30, M, Multiple, 0, 1
30, M, Multiple, 1, 0
30, M, Multiple, 1, 1
30, M, Nat Haw/Oth Pac Isl, 0, 0
30, M, Nat Haw/Oth Pac Isl, 0, 1
30, M, Nat Haw/Oth Pac Isl, 1, 0
30, M, Nat Haw/Oth Pac Isl, 1, 1
30, M, White, 0, 0
30, M, White, 0, 1
30, M, White, 1, 0
30, M, White, 1, 1
40, F, Am Ind/Ak Nat, 0, 0
40, F, Am Ind/Ak Nat, 0, 1
40, F, Am Ind/Ak Nat, 1, 0
40, F, Am Ind/Ak Nat, 1, 1
40, F, Asian, 0, 0
40, F, Asian, 0, 1
40, F, Asian, 1, 0
40, F, Asian, 1, 1
40, F, Blk/Afr Am, 0, 0
40, F, Blk/Afr Am, 0, 1
40, F, Blk/Afr Am, 1, 0
40, F, Blk/Afr Am, 1, 1
40, F, Declined, 0, 0
40, F, Declined, 0, 1
40, F, Declined, 1, 0
```

```
40, F, Declined, 1, 1
40, F, Hisp/Lat, 0, 0
40, F, Hisp/Lat, 0, 1
40, F, Hisp/Lat, 1, 0
40, F, Hisp/Lat, 1, 1
40, F, Multiple, 0, 0
40, F, Multiple, 0, 1
40, F, Multiple, 1, 0
40, F, Multiple, 1, 1
40, F, Nat Haw/Oth Pac Isl, 0, 0
40, F, Nat Haw/Oth Pac Isl, 0, 1
40, F, Nat Haw/Oth Pac Isl, 1, 0
40, F, Nat Haw/Oth Pac Isl, 1, 1
40, F, White, 0, 0
40, F, White, 0, 1
40, F, White, 1, 0
40, F, White, 1, 1
40, M, Am Ind/Ak Nat, 0, 0
40, M, Am Ind/Ak Nat, 0, 1
40, M, Am Ind/Ak Nat, 1, 0
40, M, Am Ind/Ak Nat, 1, 1
40, M, Asian, 0, 0
40, M, Asian, 0, 1
40, M, Asian, 1, 0
40, M, Asian, 1, 1
40, M, Blk/Afr Am, 0, 0
40, M, Blk/Afr Am, 0, 1
40, M, Blk/Afr Am, 1, 0
40, M, Blk/Afr Am, 1, 1
40, M, Declined, 0, 0
40, M, Declined, 0, 1
40, M, Declined, 1, 0
40, M, Declined, 1, 1
40, M, Hisp/Lat, 0, 0
40, M, Hisp/Lat, 0, 1
40, M, Hisp/Lat, 1, 0
40, M, Hisp/Lat, 1, 1
40, M, Multiple, 0, 0
40, M, Multiple, 0, 1
40, M, Multiple, 1, 0
40, M, Multiple, 1, 1
40, M, Nat Haw/Oth Pac Isl, 0, 0
40, M, Nat Haw/Oth Pac Isl, 0, 1
```

```
40, M, Nat Haw/Oth Pac Isl, 1, 0
40, M, Nat Haw/Oth Pac Isl, 1, 1
40, M, White, 0, 0
40, M, White, 0, 1
40, M, White, 1, 0
40, M, White, 1, 1
50, F, Am Ind/Ak Nat, 0, 0
50, F, Am Ind/Ak Nat, 0, 1
50, F, Am Ind/Ak Nat, 1, 0
50, F, Am Ind/Ak Nat, 1, 1
50, F, Asian, 0, 0
50, F, Asian, 0, 1
50, F, Asian, 1, 0
50, F, Asian, 1, 1
50, F, Blk/Afr Am, 0, 0
50, F, Blk/Afr Am, 0, 1
50, F, Blk/Afr Am, 1, 0
50, F, Blk/Afr Am, 1, 1
50, F, Declined, 0, 0
50, F, Declined, 0, 1
50, F, Declined, 1, 0
50, F, Declined, 1, 1
50, F, Hisp/Lat, 0, 0
50, F, Hisp/Lat, 0, 1
50, F, Hisp/Lat, 1, 0
50, F, Hisp/Lat, 1, 1
50, F, Multiple, 0, 0
50, F, Multiple, 0, 1
50, F, Multiple, 1, 0
50, F, Multiple, 1, 1
50, F, Nat Haw/Oth Pac Isl, 0, 0
50, F, Nat Haw/Oth Pac Isl, 0, 1
50, F, Nat Haw/Oth Pac Isl, 1, 0
50, F, Nat Haw/Oth Pac Isl, 1, 1
50, F, White, 0, 0
50, F, White, 0, 1
50, F, White, 1, 0
50, F, White, 1, 1
50, M, Am Ind/Ak Nat, 0, 0
50, M, Am Ind/Ak Nat, 0, 1
50, M, Am Ind/Ak Nat, 1, 0
50, M, Am Ind/Ak Nat, 1, 1
50, M, Asian, 0, 0
```

```
50, M, Asian, 0, 1
50, M, Asian, 1, 0
50, M, Asian, 1, 1
50, M, Blk/Afr Am, 0, 0
50, M, Blk/Afr Am, 0, 1
50, M, Blk/Afr Am, 1, 0
50, M, Blk/Afr Am, 1, 1
50, M, Declined, 0, 0
50, M, Declined, 0, 1
50, M, Declined, 1, 0
50, M, Declined, 1, 1
50, M, Hisp/Lat, 0, 0
50, M, Hisp/Lat, 0, 1
50, M, Hisp/Lat, 1, 0
50, M, Hisp/Lat, 1, 1
50, M, Multiple, 0, 0
50, M, Multiple, 0, 1
50, M, Multiple, 1, 0
50, M, Multiple, 1, 1
50, M, Nat Haw/Oth Pac Isl, 0, 0
50, M, Nat Haw/Oth Pac Isl, 0, 1
50, M, Nat Haw/Oth Pac Isl, 1, 0
50, M, Nat Haw/Oth Pac Isl, 1, 1
50, M, White, 0, 0
50, M, White, 0, 1
50, M, White, 1, 0
50, M, White, 1, 1
60, F, Am Ind/Ak Nat, 0, 0
60, F, Am Ind/Ak Nat, 0, 1
60, F, Am Ind/Ak Nat, 1, 0
60, F, Am Ind/Ak Nat, 1, 1
60, F, Asian, 0, 0
60, F, Asian, 0, 1
60, F, Asian, 1, 0
60, F, Asian, 1, 1
60, F, Blk/Afr Am, 0, 0
60, F, Blk/Afr Am, 0, 1
60, F, Blk/Afr Am, 1, 0
60, F, Blk/Afr Am, 1, 1
60, F, Declined, 0, 0
60, F, Declined, 0, 1
60, F, Declined, 1, 0
60, F, Declined, 1, 1
```

```
60, F, Hisp/Lat, 0, 0
60, F, Hisp/Lat, 0, 1
60, F, Hisp/Lat, 1, 0
60, F, Hisp/Lat, 1, 1
60, F, Multiple, 0, 0
60, F, Multiple, 0, 1
60, F, Multiple, 1, 0
60, F, Multiple, 1, 1
60, F, Nat Haw/Oth Pac Isl, 0, 0
60, F, Nat Haw/Oth Pac Isl, 0, 1
60, F, Nat Haw/Oth Pac Isl, 1, 0
60, F, Nat Haw/Oth Pac Isl, 1, 1
60, F, White, 0, 0
60, F, White, 0, 1
60, F, White, 1, 0
60, F, White, 1, 1
60, M, Am Ind/Ak Nat, 0, 0
60, M, Am Ind/Ak Nat, 0, 1
60, M, Am Ind/Ak Nat, 1, 0
60, M, Am Ind/Ak Nat, 1, 1
60, M, Asian, 0, 0
60, M, Asian, 0, 1
60, M, Asian, 1, 0
60, M, Asian, 1, 1
60, M, Blk/Afr Am, 0, 0
60, M, Blk/Afr Am, 0, 1
60, M, Blk/Afr Am, 1, 0
60, M, Blk/Afr Am, 1, 1
60, M, Declined, 0, 0
60, M, Declined, 0, 1
60, M, Declined, 1, 0
60, M, Declined, 1, 1
60, M, Hisp/Lat, 0, 0
60, M, Hisp/Lat, 0, 1
60, M, Hisp/Lat, 1, 0
60, M, Hisp/Lat, 1, 1
60, M, Multiple, 0, 0
60, M, Multiple, 0, 1
60, M, Multiple, 1, 0
60, M, Multiple, 1, 1
60, M, Nat Haw/Oth Pac Isl, 0, 0
60, M, Nat Haw/Oth Pac Isl, 0, 1
60, M, Nat Haw/Oth Pac Isl, 1, 0
```

```
    60, M, Nat Haw/Oth Pac Isl, 1, 1
    60, M, White, 0, 0
    60, M, White, 0, 1
    60, M, White, 1, 0
    60, M, White, 1, 1
    ;
Run;

/*Showing the breadth of possible survival curves.*/
TITLE;
ODS GRAPHICS ON;
ODS RTF FILE="C:\Users\wwilson1\Documents\Wilson Thesis\
             Survival Analysis\Survival Curves\Cox_Covariates1.doc";
Proc Phreg
    Data = Civcomb_&start_yr._&end_yr.
    Plots(Overlay) = survival;
    Class Gender Race_Grp HighEd Prior_Mil;
    Model YOS*Retain(1) = AGE
                          GENDER
                          Race_Grp
                          HighEd
                          Prior_Mil
                          AGE*GENDER
                          AGE*Race_Grp
                          AGE*HighEd
                          AGE*Prior_Mil
                          GENDER*Race_Grp
                          GENDER*Prior_Mil
                          Race_Grp*HighEd
                          Race_Grp*Prior_Mil
                          HighEd*Prior_Mil
                          AGE*GENDER*Race_Grp
                          AGE*GENDER*Prior_Mil
                          AGE*Race_Grp*HighEd
                          AGE*Race_Grp*Prior_Mil
                          AGE*HighEd*Prior_Mil
                          GENDER*Race_Grp*Prior_Mil
                          Race_Grp*HighEd*Prior_Mil;
    Baseline Covariates = Work.Covariates1
    Out = CovOut1 SURVIVAL = _all_;
Run;
ODS RTF CLOSE;
ODS GRAPHICS OFF;
```

```
/*Creating a list of best/worst covariates based on odds ratios.*/
Data Covariates2;
    Length Race_Grp $19;
    Infile datalines dsd missover;
    Input Age Gender $ Race_Grp $ HighEd Prior_Mil;
    Datalines;
    60, M, Multiple, 1, 1
    20, F, Blk/Afr Am, 0, 0
    ;
Run;

/*Showing the best and worst survival curves based on odds ratios.*/
ODS GRAPHICS ON;
ODS RTF FILE="C:\Users\wwilson1\Documents\Wilson Thesis\
             Survival Analysis\Survival Curves\Cox_Covariates2.doc";
Proc Phreg
    Data = Civcomb_&start_yr._&end_yr.
    Plots(Overlay) = survival;
    Class Gender Race_Grp HighEd Prior_Mil;
    Model YOS*Retain(1) = AGE
                          GENDER
                          Race_Grp
                          HighEd
                          Prior_Mil
                          AGE*GENDER
                          AGE*Race_Grp
                          AGE*HighEd
                          AGE*Prior_Mil
                          GENDER*Race_Grp
                          GENDER*Prior_Mil
                          Race_Grp*HighEd
                          Race_Grp*Prior_Mil
                          HighEd*Prior_Mil
                          AGE*GENDER*Race_Grp
                          AGE*GENDER*Prior_Mil
                          AGE*Race_Grp*HighEd
                          AGE*Race_Grp*Prior_Mil
                          AGE*HighEd*Prior_Mil
                          GENDER*Race_Grp*Prior_Mil
                          Race_Grp*HighEd*Prior_Mil;
    Baseline Covariates = Work.Covariates2
    Out = CovOut2 SURVIVAL = _all_;
```

```
Run;
ODS RTF CLOSE;
ODS GRAPHICS OFF;
```

Bibliography

1. Air force civilian service homepage. `https://afciviliancareers.com/`. (accessed June 12, 2020).

2. Office of personnel management federal employment reports. `https://www.opm.gov/policy-data-oversight/data-analysis-documentation/federal-employment-reports/reports-publications/federal-civilian-employment/`. (accessed June 12, 2020).

3. Defense officer personnel management act. `https://www.govtrack.us/congress/bills/96/s1918/text`. (accessed June 12, 2020).

4. D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to Linear Regression Analysis, 5th Edition*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2012.

5. G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer-Verlag, New York, 2013.

6. L. Cui, H. Li, W. Hui, S. Chen, L. Yang, Y. Kang, Q. Bo, and J. Feng. A deep learning-based framework for lung cancer survival analysis with biomarker interpretation. *BMC Bioinformatics*, 21(1):112, 2020.

7. E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

8. J. Banks, J. S. Carson, B. L. Nelson, and D. M. Nicol. *Discrete-event system simulation*. Prentice Hall, Upper Saddle River, 2010.

9. Simio simulation software. `https://www.simio.com/index.php`. (accessed June 21, 2020).

10. Arena simulation software. `https://www.arenasimulation.com/`. (accessed June 21, 2020).

11. C. Weimer, J. O. Miller, and R. R. Hill. Agent-based modeling: An introduction and primer. In *Proceedings of the 2016 Winter Simulation Conference*. IEEE, 2016.

12. J. D. Sterman. System dynamics modeling: Tools for learning in a complex world. *California Management Review*, 43(4):8–25, 2001.

13. M. van Diepen and P. H. Franses. Evaluating chi-squared automatic interaction detection. *Information Systems*, 31(8):814–831, 2006.

14. J. T. Hall. Forecasting marine corps enlisted attrition through parametric modeling. Master's thesis, Naval Postgraduate School, Monterrey, CA, 2009.

15. J. 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, 2015.

16. C. N. Franzen. Survival analysis of us air force rated officer retention. Master's thesis, Air Force Institute of Technology, Wright-Patterson AFB, OH, 2017.

17. R.R. Hill, J. O. Miller, and G. A. McIntyre. Applications of discrete event simulation modeling to military problems. In *Proceeding of the 2001 Winter Simulation Conference*, pages 780–788. IEEE, 2001.

18. P. K. Davis. Distributed interactive simulation in the evolution of dod warfare modeling and simulation. *Proceedings of the IEEE*, 83(8):1138–1155, 1995.

19. C. A. Castro and A. H. Huffman. Predicting retention rates of us soldiers stationed in europe. Technical report, United States Army Medical Research Unit-Europe, 2002.

20. S. R. Parker and J. A. Marriott. Personnel forcecasting strategic workforce planning: a proposed simulation cost modeling methodology. In *Proceedings of the 1999 Winter Simulation Conference*, pages 1410–1414. IEEE, 1999.

21. S. Cho, J. Y. Lee, B. A. Mark, and S. Yun. Turnover of new graduate nurses in their first job using survival analysis. *Journal of Nursing Scholarship*, 44(1):63–70, 2012.

22. B. E. S. Bailey, R. G. Wharton, and C. D. J. Holman. Glass half full: Survival analysis of new rural doctor retention in western australia. *The Australian Journal of Rural Health*, 24(4):258–264, 2016.

23. D. J. Russell, J. S. Humphreys, M. R. McGrail, W. I. Cameron, and P. J. Williams. The value of survival analyses for evidence-based rural medical workforce planning. *Human Resources for Health*, 11(65):1–18, 2013.

24. A. Zini, Y. Zaken, H. Ovadia-Gonen, J. Mann, and Y. Vered. Burnout level among general and specialist dentists: A global manpower concern. *Occupational Medicine and Health Affairs*, 1(128):1–4, 2013.

25. J. Capon, O. Chernyshenko, and S. Stark. Applicability of civilian retention theory in the new zealand military. *New Zealand Journal of Psychology*, 36(1):50–56, 2007.

26. Classification & qualifications. `https://www.opm.gov/policy-data-oversight/classification-qualifications/classifying-general-schedule-positions/`. (accessed September 27, 2020).

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
25–03–2021	Master's Thesis	Sept 2019 — Mar 2021

4. TITLE AND SUBTITLE

An Examination of Civilian Retention in the United States Air Force

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Wilson, William F., Capt, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENS-MS-21-M-195

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AF/A1XD
Douglas A. Boerman
1550 W. Perimeter Rd,, Rm 4710
Joint Base Andrews NAF Washington, MD 20762-5000
Email: douglas.a.boerman.civ@mail.mil

10. SPONSOR/MONITOR'S ACRONYM(S)

AF/A1XD

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

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A:
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The backbone of the United States Air Force is undoubtedly the large civilian workforce that supplements the great work that is accomplished. Many research studies have been conducted on officer and enlisted personnel to ensure that the career fields are properly developed and managed to meet the ever growing demands of the military's varied missions, but no recent studies have focused on the civilian workforce. Striking a balance between new and experienced employees is paramount to success given the ever-changing economic and political landscapes where we find ourselves. The first part of the research uses logistic regression to determine the factors that are important for retention in the civilian workforce over the last ten years (2010-2019). The six variables analyzed were age, gender, race, education level, prior military status, and years of service; all six were significant. Further breakdowns showed differences between the occupational series and between white-collar and laborer positions. Odds ratios indicate the disparity between having a certain qualification or not. The second part of the study uses survival analysis in the form of Kaplan-Meier survival curves and a Cox proportional hazards model to create unique survival curves that display the probability of remaining in employment given the number of years of service for a particular group. Future personnel management decisions can be enhanced using these curves as a basis for understanding the recent retention trends of the civilian workforce.

15. SUBJECT TERMS

personnel analysis, civilian workforce, retention, logistic regression, survival analysis

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. Raymond R. Hill, AFIT/ENS
U	U	U	UU	101	19b. TELEPHONE NUMBER *(include area code)* (937) 255-3636, x7469; Raymond.Hill@afit.edu