AIR FORCE OFFICER ATTRITION: AN ECONOMETRIC ANALYSIS

THESIS

Jacob T Elliott, 1st Lt

AFIT-ENS-MS-18-M-118

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENS-MS-18-M-118

AIR FORCE OFFICER ATTRITION: AN ECONOMETRIC ANALYSIS

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Research

Jacob T Elliott, BS

1st Lt, USAF

22 March 2018

AFIT-ENS-MS-18-M-118

AIR FORCE OFFICER ATTRITION: AN ECONOMETRIC ANALYSIS

THESIS

Jacob T Elliott, BS
1st Lt, USAF

Committee Membership:

Raymond R. Hill, PhD
Chair

Major Thomas P. Talafuse, PhD
Member

AFIT-ENS-MS-18-M-118

Abstract

Many organizations are concerned, and struggle, with personnel management. Training personnel is expensive, so there is a high emphasis on understanding why and anticipating when individuals leave an organization. The military is no exception. Moreover, the military is strictly hierarchical and must grow all its leaders, making retention all the more vital. Intuition holds that there is a relationship between the economic environment and personnel attrition rates in the military (e.g. when the economy is bad, attrition is low). This study investigates that relationship in a more formal manner. Specifically, this study conducts an econometric analysis of U.S. Air Force officer attrition rates from 2004-2016, utilizing several economic indicators such as the unemployment rate, labor market momentum, and labor force participation. Dynamic regression models are used to explore these relationships, and to generate a reliable attrition forecasting capability. This study finds that the unemployment rate significantly affects U.S. Air Force officer attrition, reinforcing the results of previous works. Furthermore, this study identifies a time lag for that relationship; unemployment rates were found to affect attrition two years later. Further insights are discussed, and paths for expansion of this work are laid out.

Acknowledgements

I am incredibly grateful to my advisor, Dr. Hill, for getting on my case when I needed it and above all else, for being patient. I also want to thank my sponsor, the Strategic Analysis branch of the Force Management Division of Headquarters Air Force (HAF/A1XDX) for providing the personnel data and research guidance.

Table of Contents

List of Tables

List of Figures

AIR FORCE OFFICER ATTRITION: AN ECONOMETRIC ANALYSIS

I. Introduction

1.1 Background

As with any large organization, the personnel management functions of the components of the Department of Defense (DoD) are concerned with personnel retention. However, since the DoD must grow all its leaders from an entry level, retention is far more important and challenging.

The DoD has long offered an all-or-nothing 20-year retirement: stay to 20 years and you are eligible for retirement benefits, leave before 20 years and you have nothing. This 20-year goal has certainly been a positive retention motivator.

The new blended retirement system will change the all-or-nothing aspect of military retirement. Personnel can now leave before 20 years with some level of retirement benefit. These new options will surely change military retention patterns. How the patterns will change is unknown.

Part of the military strategy to keep retention at desired levels is to increase pay levels of targeted personnel groups with retention bonuses. Clearly, military members offered such a bonus must consider the bonus and retaining versus civilian pay potential if the member separates.

This research is a study of military retention as affected by economic measures used

as indicators of civilian employment potential. An important caveat is that the study is based on pre-blended retirement systems. The blended system is simply too new to provide meaningful trend data.

1.2 Scope

For both releasability and compatibility reasons, the Air Force personnel data used in this work has been aggregated to the national level, limiting the detail to which relationships can be explored. This was done to match the national economic data available, and to protect personal information of the individuals included in the analysis.

The military personnel data concerns those serving during the 2004-2017 timeframe, and the economic data matches. Some extraordinary events occurred during that period, notably the Great Recession beginning in 2008, which may have altered normal military retention behavior. The U.S. military is also transitioning to a new retirement system. It is possible that any relationships revealed in this thesis will be affected differently by the new retirement system.

1.3 Assumptions and Limitations

As with any analytic endeavor, several assumptions are made in order to faciliate the modeling of real world phenomena. Perhaps most central to this thesis is the assumption that there exists at least one economic indicator (but ideally many) that helps inform an individual military member's decision to stay or leave active duty service. It is also assumed that if these variables do not directly inform individual retention decisions, they serve as adequate proxies for unobservable or abstract factors that do influence the individual's decision. For instance, members may not follow the movements of the Consumer Price Index (CPI), but that movement should provide information on the cost of living which may affect

the decision to stay in the military. Naturally, it is assumed the collective individual behaviors adequately aggregate so that the data employed is reflective of the collective individual behaviors. We also assume that the skills held by the Air Force officer corps are largely transferrable to civilian labor markets. Standard assumptions asociated with regression modeling and forecasting are made (independent, normal, and homoscedastic errors) and are tested, as well.

1.4 Outline

This chapter introduced the retention problem investigated and discussed the foundational motivations and thoughts underpinning the thesis. The next chapter reviews the related literature - the efforts used to better frame the problem and previous attempts to model it. The third chapter focuses on the methodology, documenting how and why the data were attained (i.e. sources and selection criteria), as well as any transformations necessary to conduct the analysis. Chapter III continues by discussing the modeling procedure in detail, including general steps and specific mathmatical formulations. Lastly, the results are examined and insights or conclusions are highlighted in Chapter IV.

II. Literature Review

2.1 Chapter Overview

Managing personnel and modeling retention behaviors have, appropriately, long been a concern of the Department of Defense as well as almost any non-military organization. This chapter summarizes the retention problem, examines previous research endeavors, and finally discusses the impetus for the econometric approach used in this research.

2.2 The Military Retention Problem

All organizations have some problem associated with retaining their people. This is especially true of the military, wherein members are routinely confronted with deployments, long duty hours, and frequent relocations - factors generally not found in non-military organizations. These factors produce high stress on the military members and their families, who play a significant role in a member's retention decision [1]. Evidence suggests that individuals serving in the military are generally more tolerant of these conflicts [2], but the causes of attrition involve more than just familial concerns. Kane [3] argues the military suffers from a chronic personnel mismanagement problem: members' merit is not always rewarded nearly as well as it is in the private sector, in terms of personal recognition and upward movement, partly due to heavy bureaucratic restrictions. This disparity can lead to frustration and job dissatisfaction, damaging the member's commitment to the organization and incentivizing their attrition behavior [2].

Compounding the internal frustrations, civilian labor markets can offer intense incentives for leaving. Barrows [4] details the mechanisms underpinning U.S. Air Force pilot attrition

to civilian airlines, framing the problem with human capital theory. The military offers a unique opportunity for developing highly desired skill sets, placing members in positions of high stress, and providing them responsibility at early stages of professional development [3]. Furthermore, evidence suggests that the military as an insitution is quite adept at attracting intelligent and capable individuals [5]. Providing innately talented individuals with a high degree of general and specific training fosters the development of high-performers with desirable and broadly applicable skill sets. Therein lies the problem. Civilian firms are typically more flexible in their ability to compensate such individuals through organizational advancement and wage, often outcompeting the military [3]. These phenomena are in direct contradiction to the principles for successful retention laid out by Asch [6]. Asch explains that in order for military compensation to be attractive, it needs to be at least as great as the members' expected wages and benefits as would be offered by civilian labor markets. Compensation should also be contigent upon performance, reflecting the individual's value to the organization, to maintain motivation and disincentivize attrition [6]. In order to help best determine compensation, then, it behooves the military to develop methods for anticipating the effects of labor market conditions on military members' retention decisions.

2.3 Previous Research

There have been many forays into personnel retention modeling and forecasting. Saving et al. [7] find a significant interaction between labor markets and military retention by analyzing individual career fields within the U.S. Air Force. Their results indicate that demographic factors such as race and education level are influential to retention at early stages, but exhibit diminished effects as careers progress. Additionally, their work supports the conjecture that civilian wages, unemployment rates, and other economic variables affect military retention.

In 1987, Grimes [8] investigated the retention problem by applying a variety of regression

methods (ordinary multiple linear regression, with logarithmic tranformations on response and/or explanatory variables) to try and predict officer loss estimates 6-12 months in the future. He was unable to provide adequate effects estimates or reliable predictions, concluding that the chronological nature of the data led to serial correlation errors.

Fugita and Lakhani [1] use survey and demographic data compiled by the Defense Manpower Data Center to estimate hierarchical regression equations to describe retention behaviors in Reservists and Guard members. Hierarchical regression models are useful when there exists some causal ordering among predictors, as is often the case with demographic and economic data. This causal relationship can lead to high multicollinearity, increasing the estimated standard error of coefficient estimates and resulting in non-significant predictors. They find that, for both officers and enlisted, retention probabilities tend to rise with increased earnings, years of service, and spousal attitude towards retention. Their work reinforces the importance of including demographic variables in retention modeling, and that wages are in the forefront of a member's mind when deciding to stay.

Gass [9] takes a more general view by modeling the manpower problem in three different ways: as a Markov chain with fixed transition rates between nodes, as a minimum-cost network flow problem, and as a goal-programming problem. While potentially easier to interpret, these models can present a too-sanitized picture of an enormously complex system, particularly the current military personnel system.

Barrows [4] analyzes retention, specifically for Air Force pilots, through the lens of human capital and internal labor market theories. He argues two points important to this thesis: the degree of specific training is inversely correlated with attrition, and that the Air Force personnel system suffers from the inefficiences typical of an internal labor market.

To Barrows' first point, the military offers a high degree of general and specific training. General training is conducive to attrition, as it allows the individual to more easily transfer between military and non-military jobs. Specific training decreases worker transferability and helps improve military retention. This effect is seen in differing retention rates between

general pilots (e.g. cargo, heavies) and those with more specific skill sets (e.g. helicopters, fighters). One can imagine this would also reveal itself in the non-rated officer population; that is, career fields with transferable skill sets suffer more from attrition than those with specific skill sets. For instance, logistics or inventory specialists are more general than aircraft or missile maintenance, which tends to be more military specific.

Regarding Barrows' second point, workers are somewhat insulated from the competition posed by outside labor markets (e.g. Field-grade officers do not have to worry about civilians being hired specifically to replace them), and are paid according to position as opposed to productivity. Shielding employees from outside competition can possibly remove incentive for performance; individuals who feel more secure in their jobs may not try as hard. Not paying according to performance can also be damaging in two ways: high-performers can feel undervalued and motivated to leave, and under-performers could be receiving more than they produce.

Looking to the Navy, specifically Junior Surface Warfare Officers (SWOs), Gjurich [10] found that one of the most important factors affecting retention was marital status. Single officers are more likely to leave than those with families. This actually may be a proxy for risk aversion. Those officers with dependents may be less likely to risk unemployment by leaving the military, choosing instead to retain and keep a relatively secure job. Again, the importance of demographic factors was reinforced, but little is said of the economic considerations.

In 2002, Demirel [11] used logit regression to analyze retention behaviors for officers at the end of their initial service obligation and at ten years of service. While the focus of this endeavor was to identify any changes in retention related to commissioning source, several other demographic factors - such as marital status, education level, and gender - were found to be statistically significant. This reinforces conclusions about demographic factors drawn by previous research efforts, and shows evidence that these trends generally apply to the military population, instead of particular service branches.

Ramlall [12] takes a less technical approach and surveys the existing employee motivation theories to offer an explanation of how employee motivations affect retention, and how the disregard for the principles contained therein motivate attrition. Many causes are discussed, and a few are consistent (or at least common) amongst the spectrum of motivation theories. When wages and promotions are not viewed as tied to performance, individuals are disincentivized and do not feel as loyal to the institution. Also, a lack of flexibility within job scheduling and structure is seen as disloyal or disrepectful to the individual. Lastly, when managers fail to act as coaches or are not seen as facilitators to employees' careers, turnover rates tend to be greater. Given that civilian labor markets are generally more flexible in both pay structure and work scheduling, Ramlall's research underpins the importance of incorporating civilian labor market conditions.

More recently, Schofield [13] employs a logisitic regression model to identify key demographic factors influencing the retention decisions of non-rated Air Force Officers. She finds that career field grouping, distinguished graduate status at commissioning source, years of prior enlistment, and several other structural variables were significant. She then utilizes these factors to generate a series of survival functions describing retention patterns and behavior. Again, the importance of demographic factors is reinforced. However, any possible effects of economic factors were unexplored.

Looking at the rated officer corps, Franzen [14] takes a similar approach to Schofield [13] using logisitic regression to identify significant factors and generating survival functions. However, Franzen's work differs from Schofield by choosing to also assess the influence of economic, demographic, and other variables exogenous to the military. She finds that marital status, number of dependents, gender, source of commissioning, prior enlisted service, and the New Orders value from the Advance Durable Goods Report were all significant. The first couple of factors support the notion that familial strain caused by military service affects retention, the next few factors (gender, source of commissioning, and prior service) reaffirm the work conducted by Schofield. The last variable, New Orders, suggests that indicators of

economic health play some role in retention decisions. This last observation is a motivation for this thesis research.

In that vein is the work conducted by Jantscher [15] where she conducts correlation analysis to determine the relationship between a host of economic indicators and retention rates for each Air Force Specialty Code (AFSC). The results of the preliminary correlation analysis provide a subset of economic indicators shown to be correlated with retention, such as unemployment rates, gross national savings, real GDP growth, etc. She then attempts to form a regression model to forecast retention, but was unable due to achieve an adequate model due to high multicollinearity between many of the indicators. Nonetheless, her correlation analysis provides a starting point from which additional modeling techniques may be applied.

2.4 Insights

Several key themes arise based on this review of the literature:

- Demographic and economic factors can play a significant role in a member's attitude towards retention;
- Military members are aware of and incorporate opportunities in the civilian labor market when deciding to remain in or leave military service;
- Logistic regression on demographic data yields promising results when predicting whether an individual will remain in service, but may be innappropriate for modeling aggregate trends; and
- Effects estimation of economic factors through regression can be difficult, as many indicators are highly correlated.

What is also apparent is that there are several topics yet unexplored:

- Modeling the military population with performanced-based pay structures and advancement schemes to estimate effects on retention;

- Determining how comparable the military population is to the civilian, and how easily the professional skills sets exhibited by the former transfer to the latter; and
- Applying other forecasting techniques (ARIMA, Exponential Smoothing, Dynamic Regression) to retention data to help achieve models that provide insight into the military retention problem.

This thesis research focuses on the last point. The research goal is to forecast Air Force Non-rated officer retention with a dynamic regression model in order to estimate the effects of different economic indicators. This is approach covered in the next chapter.

III. Analysis and Results

3.1 Data Composition

3.1.1 Introduction

Predictive and descriptive analyses begin with attaining an understanding of the data. Every data set has its idiosyncracies, its own unique challenges. Understanding these characteristics and the meaning of the data - what the variables represent and how they might interact with each other - is key to any successful analytic endeavor. Below, the data used in this research are described in detail to include its sources, meaning, and peculiarities.

3.1.2 HAF/A1XDX

The Strategic Analysis branch of the Force Management Division of Headquarters Air Force (AF/A1XDX) provided the data on Air Force personnel used in this research. The data are extracted from the Military Personnel Data System (MilPDS), a database containing Air Force personnel data for every airman over his or her career. The data are input by trained personnelists or are automatically updated within the system (e.g., age will automatically increase). The data were originally split into two separate `.sas7bdat` files, one containing monthly attrition numbers for each Air Force Specialty Code (AFSC) and the other detailing monthly assigned levels for each AFSC. Each file contains information starting in October of 2004 through September of 2017, for a total of 156 observations across 67 AFSCs.

3.1.3 Federal Reserve Bank of St. Louis

The Federal Reserve Bank of St. Louis is one of 13 banking entities which comprise the United States' central bank (the others being 11 regional reserve banks and the Board of Governors). As a whole, the central bank is responsible for determining and enacting monetary policy for the U.S. Many of these entities maintain expansive databases containing information about the U.S. economic environment - financial data, national employment statistics, private sector business data, etc. Fortunately, the Federal Reserve Bank of St. Louis offers public access to the Federal Reserve Economic Data (FRED) database via an online interface. From this interface, historical data on several economic indicators were retrieved for this research: the nation unemployment rate (both seasonally adjusted and non-adjusted), the labor force participation rate (LFPR), job openings (adjusted and not), total nonfarm job quits, the labor market momentum index, real GDP per capita, and the consumer price index (CPI). Each indicator consists of monthly recordings across varying time spans (e.g. 1990-2016 or 2001-2017).

The LFPR is the percentage of the population actively employed or looking for employment. Changes to the participation rate can give insight into the strength of the economy - e.g. rising participation is usually associated with economic growth. When paired with unemployment rates, the LFPR can also reveal people's attitude about the economy. For example, the steady decline of participation from 2010 onward (seen in Figure 1) might indicate that the decrease in unemployment over the same period is somewhat exaggerated; people seeking, but unable to find work may become discouraged and exit the labor force, artificially decreasing the unemployment rate. It is possible that this perception of economic health affects military retention decisions. In this research, LFPR is restricted to members of the civilian labor force with at least a baccalaureate degree and no younger than 25 years of age. This subset of the civilian labor force most closely matches the characteristics of military officers.

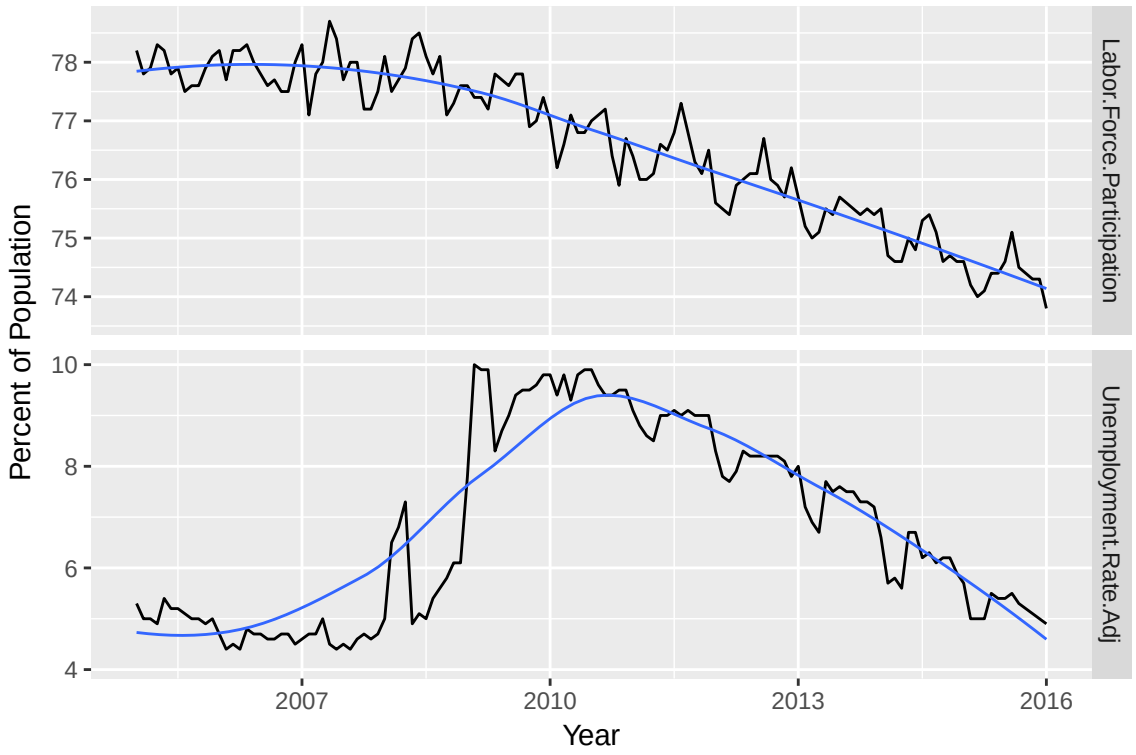

Figure 1. Participation and Unemployment

It is assumed that the skillsets of the target population (Air Force officers) are most transferrable to those jobs covered by nonfarm payrolls. Nonfarm is a category of the labor force that excludes proprietors, private household employees, unincorporated self-employment, unpaid volunteers, and farm employees [16]. Job quits are generally voluntary separations and may reflect workers' willingness to leave the job; it may be that the a higher propensity to volutarily leave a job translates to a positive outlook on obtaining another and the economy as a whole.

The labor market momentum index compares current labor market conditions to historical averages. A negative value indicates conditions below the long-term average, and a positive value indicates favorable conditions. The CPI examines the weighted average price of a basket of consumer goods and services; it is used to estimate the cost of living. There is some uncertainty involving employment in separation from the military, so cost of living information may be especially important to the retention decision as the military is excluded from CPI statistics.

By including these variables in a regression model and estimating their effects on military attrition trends, this work seeks to capture military members' perceptions of economic health and job prospects, and use that information as a means to forecast Air Force officer attrition.

3.1.4 Cleaning and Preparation

Perfect data are rarely found or received outside of the classroom, and such is the case here. Before exploration and modeling, several steps helped produce a useable data set.

The personnel data is first converted from long to wide format. Originally, the personnel data comes with three variables: Air Force Specialty Code (AFSC), Date, and Separations. This form is not conducive to modeling. A new variable is thus created for each category in AFSC containing the associated separation counts. This procedure generates missing values, which then must be dealt with appropriately. Missing values can result from several underlying issues: data storage corruption, entry errors, miscommunication between software, none of which apply here. Since the attrition data is a monthly count of people exiting USAF service, the intuition is that these missing values simply represent a lack of an observation (i.e. zero separations). This is confirmed by the data's provider. Therefore all missing values in the personnel data are replaced with zero. Initially, observation dates are stored as the number of days since 1 Jan 1960 (the standard for SAS). This is transformed into YYYY-MM-DD to facilitate its merging with the economic data. An additional column is tabulated, the total separations across all AFSCs. This column total is the response used

Table 1. Selected Economic Indicators

Variable	**Description**
Labor Market Momentum Index	Compares current market conditions to long-run average
CPI	Weighted average price of a basket of goods and services
Nonfarm Jobs Openings	Unfilled positions at the end of the month in the nonfarm sector
Real GDP per Capita	Measure of economic output per person, adjusted for inflation
Nonfarm Job Quits	Voluntary separations from jobs in the nonfarm sector
Unemployment Rate	Percentage of unemployed individuals in the labor force
Labor Force Participation Rate	Percentage of the population either employed or actively seeking work

for the modeling efforts

The economic data do not require much treatment as they come from a professionally managed database. One of the indicators, real GDP per capita, occurs in quarterly intervals while the rest are monthly. To make data comparable, the quarterly values are applied across each month in the quarter (e.g. the observation for Q1 2006 is applied over January, Febuary, and March 2006). Then, variables are also renamed for clarity. Finally, economic data are merged with the personnel data through an inner join, preserving only those observations with dates common to both data sets.

3.2 Model Selection

3.2.1 Introduction

General modeling practices involve horizontally splitting the original data set into at least two, sometimes three, subsets. This ensures model fitting and assessment are independent processes. There are many ways to generate these subsets, each particular to the structure of the data. With time-series data, as in this research, the typical approach is to retain roughly the first 80 percent of the data for model fitting, leaving the rest for model assessment. These two sections are respectively known as the training and validation sets. The training set is used to estimate model parameters, which are then used for predictions on subsequent observations. These predictions are compared against the validation set - actual, observed data - as a means of assessing model performance. Model performance is assessed using three criteria: the corrected Akaike Information Criteria (AICc), training root mean square error (training RMSE), and validation root mean square error (validation RMSE). Generally, better model performance is associated with lower scores for each criteria, so 'good' models are identified by having lower scores relative to other models. The training/validation approach is applied to each modeling technique employed.

This endeavor utilizes two modeling techniques for forecasting: na\"ive models and dy-

namic regression models (also known as transfer functions). The former is a far simpler technique and is used as a baseline. The latter is a bit more complex. Dynamic regression has two major components, regression and time-series, each with their own assumptions and requirements.

Regression models with multiple predictor variables assume independence of those predictors (also called regressors or exogeneous variables). All regression models assume errors are normally and independently distributed around zero with constant variance. The regression portion is primarily concerned with coefficients of the predictor variables. These coefficients provide insight as to which predictors have a statisically significant effect in explaining the variability in the data.

ARIMA models are used to address the pecularities of time series data, and a brief review of those characteristics is necessary to understand the analysis presented later in this chapter. Foremost is the concept of autocorrelation, which is when a variable (e.g. the temperature) depends on previous observations of itself. Another concept central to subsequent modeling efforts is that of stationarity. A stationary variable is one that does not exhibit mean changes, such as caused by trend or seasonality effects - when plotted over time. Stationarity is requisite for generating reliable forecasts with time-series models. Last is a matter of notation. In this work, backshift notation is used to indicate backwards time steps, denoted with B and is defined below:

For a single step back,

$$By_t = y_{t-1},$$

for two steps back,

$$B^2 y_t = y_{t-2},$$

and in general,

$$B^k y_t = y_{t-k}.$$

3.2.2 Initial Exploration

First, the data are examined visually. Plotting the response, total separations over all career fields, in Figure 2 shows significant spikes during 2005, '06, '07, and '14. It is known that during these periods, special spearation incentive programs were introduced by th Air Fofce to artificially downsize the force. The effects of these periods merit investigation later on, as they could negatively affect model prediction performance.

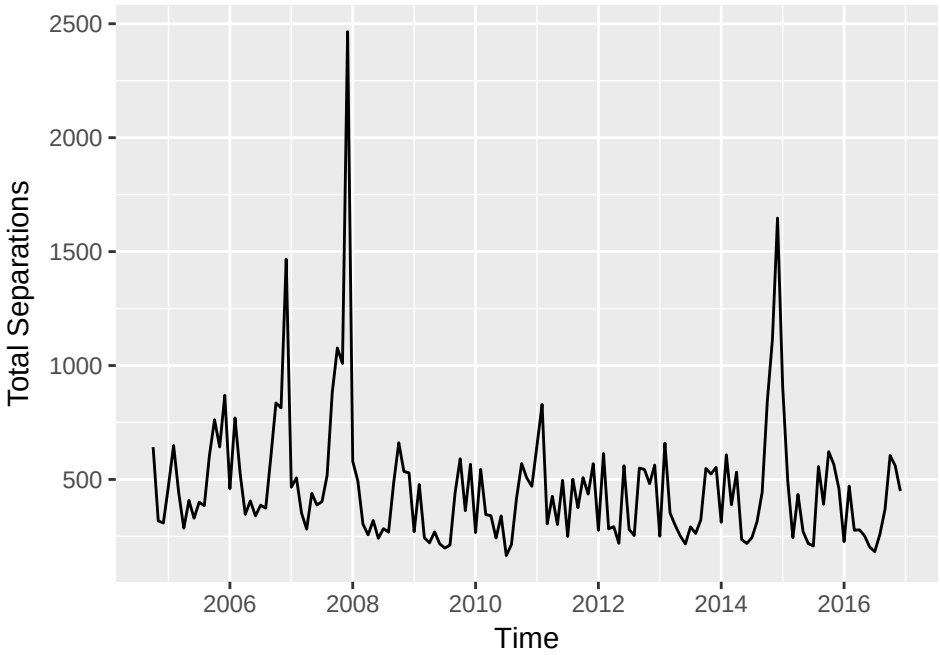

Figure 2. Monthly Officer Separations

No seasonality is immediately obvious in Figure 2. However, if each year is plotted separately, a clearer picture emerges. First, Figure 3 shows that the extreme points noticed noticed above seem to be relegated to the November-December time frame. Second, it is easier to witness the seasonality: bowing across the year, with higher counts at the beginning and end.

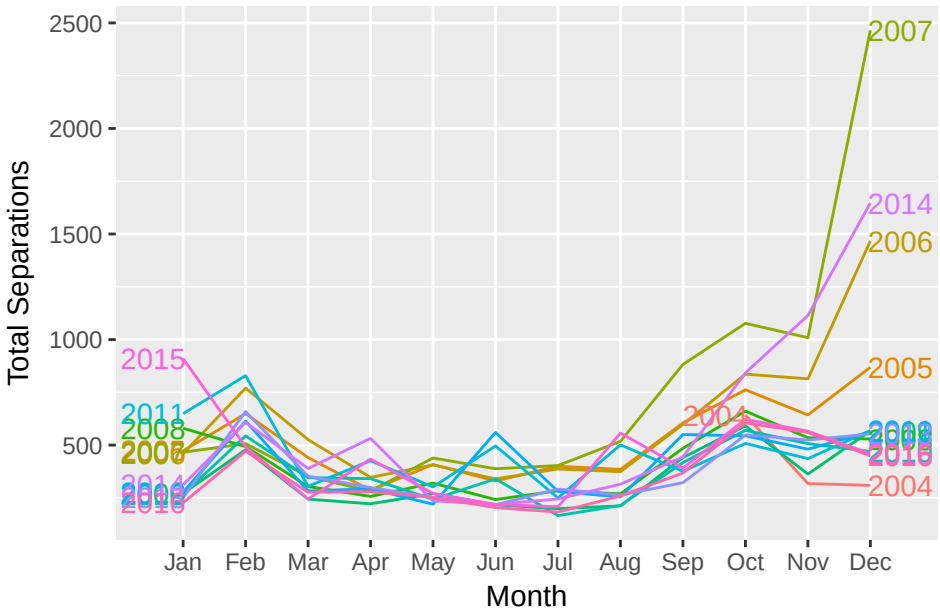

Figure 3. Seasonal Plot: Total Separations

Considering these plots, it is expected that a seasonal model performs best and some alteration will have to be made to accomodate the outliers. To confirm, na\"ive models are fit to the data and the results are examined. Beyond revealing seasonality and outlier effects, fitting na\"ive models establishes a baseline to compare against later models. Na\"ive models are very simplistic, so if later models perform worse or only marginally better, it implies they are not capturing much information.

Figure 4 gives evidence to the negative effects of outliers. Notice the large confidence intervals surrounding the na\"ive forecast and the 2014 spike carried through in the seasonal forecast.

```
## <ScaleContinuousPosition>
##  Range:
##  Limits:     0 --     1

## <ScaleContinuousPosition>
##  Range:
##  Limits:     0 --     1
```

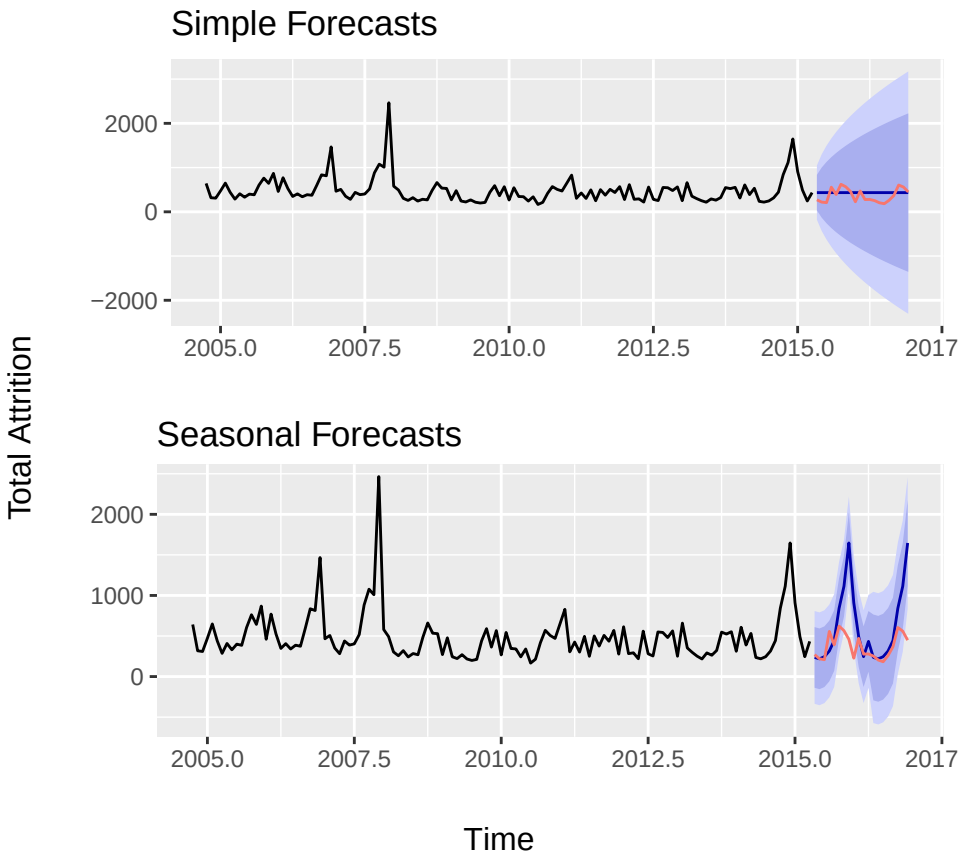

Figure 4. Simple and Seasonal Naïve Forecasts

Tables 2 and 3 show different error metrics for each of the two models. Judging by root mean square error (RMSE), the seasonal model generally fits the training data better, possibily indicating presence of seasonality effects. However, there is a large disparity between validation RMSEs, possibly caused by the major spike in 2014 - reaffirming the earlier intuition about outlier effects.

Table 2. Naïve Results

	ME	**RMSE**	**MAE**	**MPE**	**MAPE**	**MASE**
Training set	-1.651	312.523	195.984	-12.530	42.093	1.261
Test set	-62.600	160.642	144.300	-37.265	51.569	0.928

It is known that during years 2005, '06, '07, and '14 special separation programs were implemented. Given the effect those years appear to have on modeling, they must be accomodated before continuing. Before deciding how, the explicit points in question need to

Table 3. Seasonal Naïve Results

	ME	**RMSE**	**MAE**	**MPE**	**MAPE**	**MASE**
Training set	16.496	291.791	155.452	-7.374	30.452	1.000
Test set	-238.850	454.288	271.450	-59.852	67.315	1.746

be identified. To help, refer to Figure 3. As noted above, the spikes generally occur in November and December. However, the observations from 2005 are close enough to those from other years that they may have resulted naturally. Minimal removal of information from the data set is desired, removing only that which is misleading. So, November and December observations from 2006, '07, and '14 are selected for replacement.

Given the seasonality in the data set, the replaced values should stem from matching observations in previous years, as opposed to previous observations within the same year. The outliers are replaced (or imputed) with the arithmetic mean of all years not being replaced (e.g. November 2006, '07, '14 are replaced with the mean separations in November for all other years).

Replotting the response in Figure 5 shows a much better behaved data set. The data look fairly stationary, setting the stage for developing more complex forecasting models.

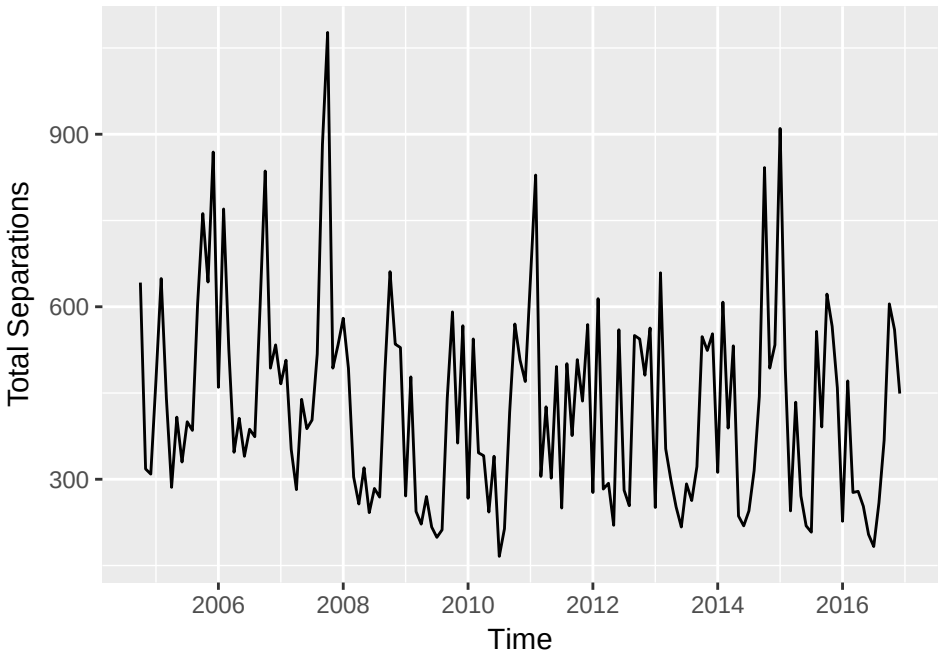

Figure 5. Separations - Outliers Removed

With the outliers replaced, seasonal effects are much more apparent (see Figure 6), further enforcing the need for a seasonal model.

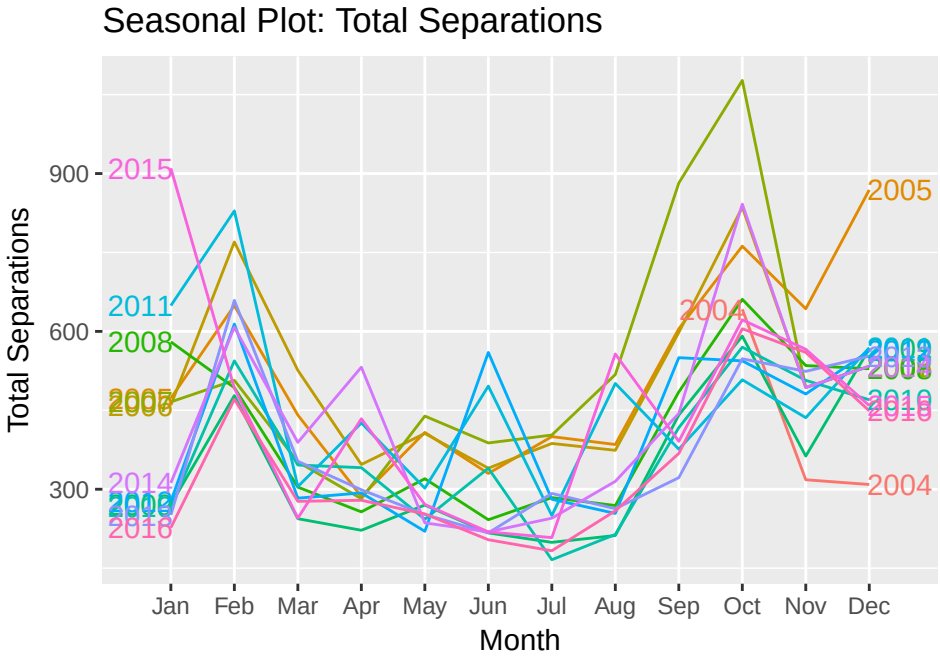

Figure 6. Seasonal Plot: Outliers Removed

Table 4 compares the seasonal na\"ive RMSEs before and after imputing the identified

outliers. The results indicate that removing and replacing the extreme values for November and December improved the model. This is further reflected by the forecast (shown in blue) in Figure 7, which follows the validation data (shown in orange) more closely than those in Figure 4. Overall, these results imply that imputation of the selected observations was useful.

Table 4. Seasonal Naïve RMSE Comparison

	Raw Data	**Imputed Data**
Training	291.791	161.262
Validation	454.288	186.584

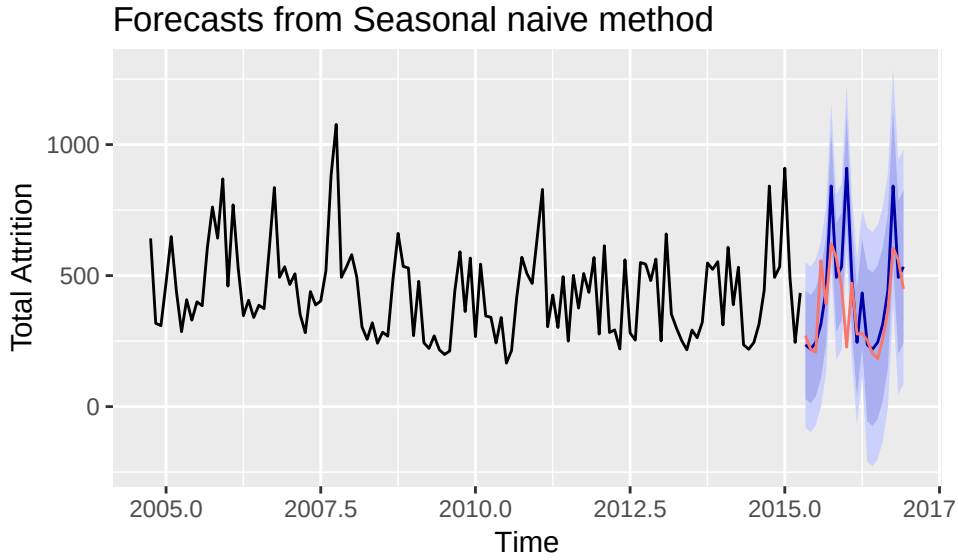

Figure 7. Seasonal Naïve Forecast After Imputation

3.2.3 Dynamic Regression

Na\"ive models are simple, regression models adequately involve exogenous predictor variables, and time series model adequately handle autoregressive components of data. Individually these models are useful, but cannot handle both exogenous and autoregressive components.

Dynamic regression is a regression model with an ARIMA model fit to the errors. The regression piece allows use of independent variables in predicting a response, and the ARIMA

portion helps model the autoregressive information which can exist in time-series data. The general formulation of a dynamic regression model with ARIMA(1,1,1) errors:

$$y_t = \beta_0 + \beta_1 x_{1,t} + ... + \beta_k x_{k,t} + n_t$$

where,

$$(1 - \phi_1 B)(1 - B)n_t = (1 + \theta_1 B)e_t$$

and e_t is white noise. The ϕ_1 is the non-seasonal autoregressive coefficient while the θ_1 is the non-seasonal moving average coefficient. The $(1 - B)$ indicates the errors are also subjected to a single order of differencing to achieve a stationary time-series in the error term.

Fitting a dynamic regression model requires taking several steps to ensure key assumptions are not violated. First is to address the issue of collinearity. Collinearity between predictor variables implies a dependent relationship and can lead to innaccurate coefficient estimates, a result contrary to the goal of any modeling effort. To avoid this pitfall, a correlation matrix of possible regressors is compiled and examined. A correlation matrix shows how collinear each pair of indicators is. A high correlation coefficient between indicators implies collinearity, meaning the variables should not be used together in the model. Figure 8 shows the correlation between all pair-wise combinations of variables in the economic data set. There are many instances of collinearity, which is expected since many economic indicators are constructed from similar information. There are some independent subsets, though, and these are the candidates for the dynamic regression model.

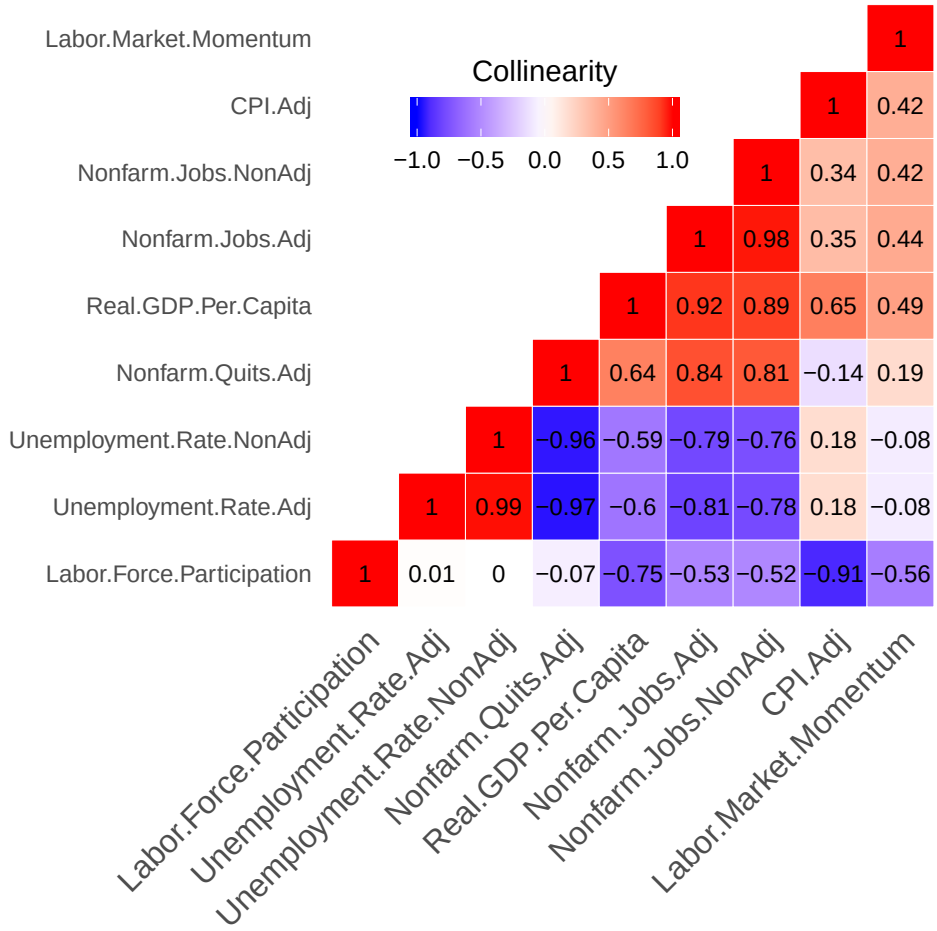

Figure 8. Correlation Matrix - Economic Indicators

Given their low correlation, Unemployment Rate (Adj.), Labor Force Participation Rate, and Labor Market Momentum Index are selected as independent variables. To ensure the assumptions made by the ARIMA piece of the model, the stationarity of the regressors is checked. Though trend and seasonality components can be incorporated through regression techniques, ARIMA models require stationarity. Non-stationary variables can produce inconsistent coefficient estimates, even if they are independent. To assess, a plot of the three indicators in Figure 9 is generated.

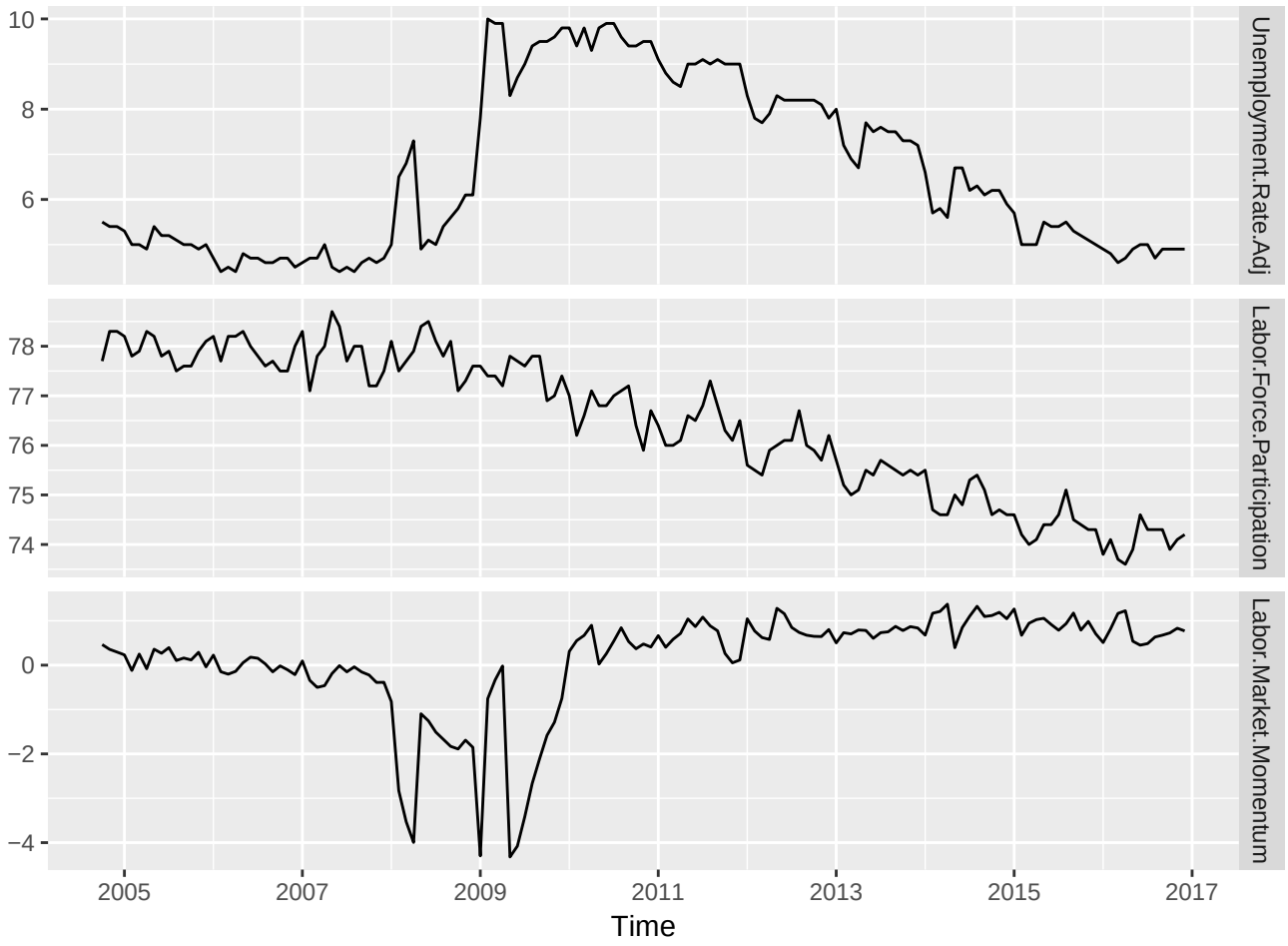

Figure 9. Economic Indicators - Raw

Each of the three indicators show evidence of a trend or changing mean(i.e. are non-stationary). To handle this, the data are differenced. The resulting data are shown below in Figure 10.

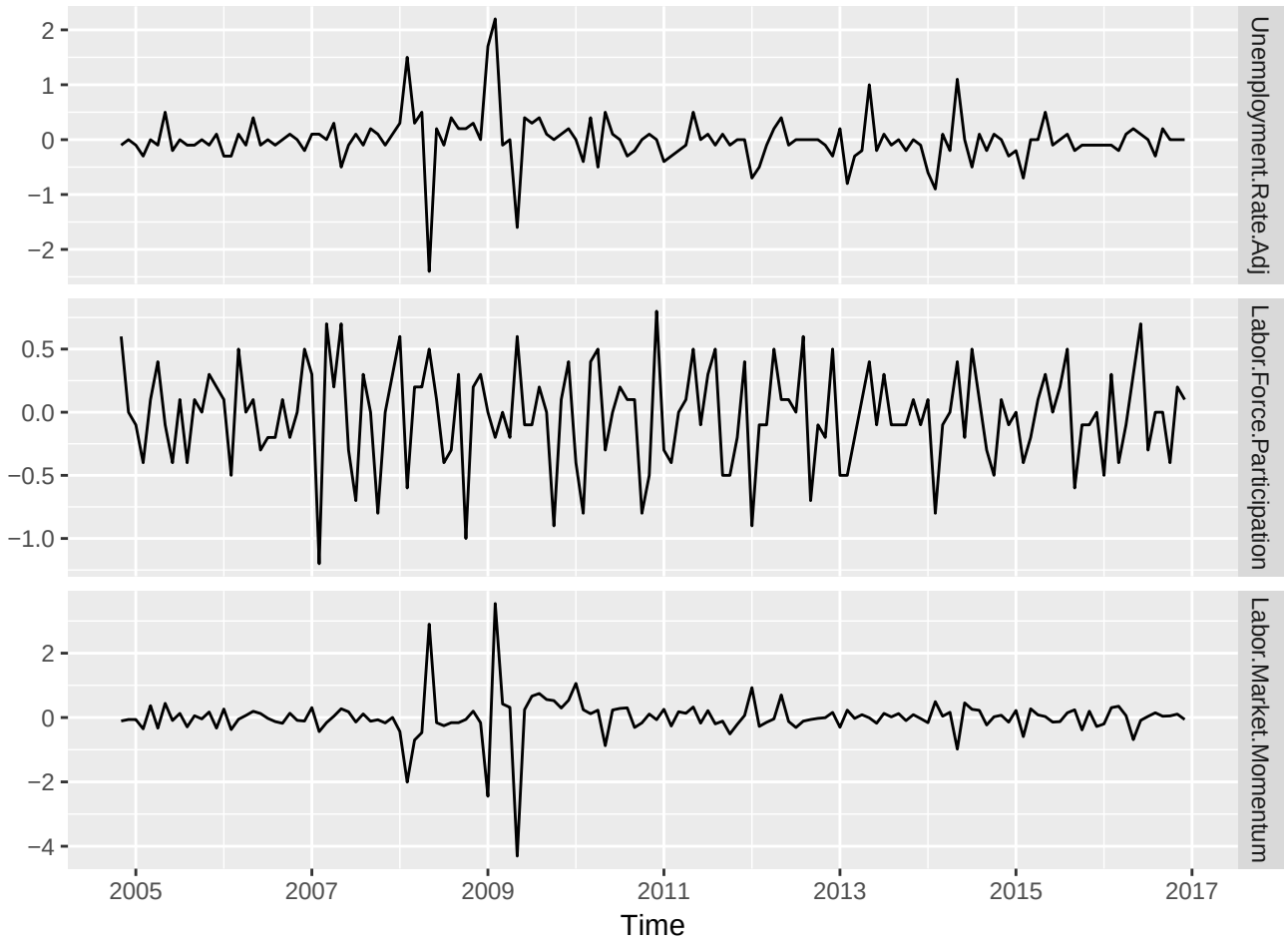

Figure 10. Economic Indicators - Differenced

Simple differencing produces the desired effect, the data are stationary. It is important to note the regressors now show the month-to-month change, which will be pertinent when interpreting results.

With stationary and independent economic indicators, model formulation can transition from regression to the ARIMA portion. Up to six parameters can be specified and estimated: the order of autoregression, degree of differencing, and order of the moving average ($p, d,$ and q, respectively) and their seasonal counterparts ($P, D,$ and Q). Many combinations are considered. A range is specified for each parameter, and a model is fit for every combination within the specified ranges; the model with the lowest corrected Akaike Information Criteria (AICc) is selected. For the first, and all subsequent, dynamic regression models in this work, the following ranges/values were used: $p, q \in [0, 5]$, $d, D = 0$, and $P, Q \in [0, 2]$.

For this first pass, the model selected was a regression model with a fourth-order moving average and first-order seasonal autoregression on the errors. More explicitly:

$$y = \beta_0 + \beta_1 x'_{1,t} + \beta_2 x'_{2,t} + \beta_3 x'_{3,t} + n_t$$

where,

$$(1 - \Phi_1 B^{12}) n_t = (1 + \theta_1 B + \theta_2 B^2 + \theta_3 B^3 + \theta_4 B^4) e_t$$

and,

$$x'_{i,t} = x_{i,t} - x_{i,t-1}$$

The coefficient estimates are shown in Table 5:

Table 5. Estimated Coefficients - Initial Model

	θ_1	θ_2	θ_3	θ_4	Φ_1	β_0	β_1	β_2	β_3
Coeff	0.218	0.145	0.336	0.260	0.576	429.875	-13.390	-15.949	-2.819
StdErr	0.087	0.088	0.092	0.092	0.082	47.602	22.286	35.016	11.494

Model assessment involves analysis of the residuals. Residuals are examined for evidence of remaining autocorrelation, satisfaction of normality assumptions ($e_t \sim N(0, \sigma^2)$), and outlier effects. Figure 11 provides the plots used to answer those questions. The top subfigure plots the raw model residuals, and is used to identify possible trends, seasonality, or heteroscedasticity. Fortunately, none of those features are apparent. The bottom-left is used to examine significant autocorrelation in the residuals; significant correlations would indicate a possible violation of the independence of the residuals. The current model's results only show one lag-period with significant autocorrelation, which may mean that there is information unnaccounted for by the current model. Overall autocorrelation, however, appears insignificant, as further evidenced by the results of a Ljung-Box test for autocorrelation (Table 6). The bottom-right plot shows a histogram of the residuals, comparing the raw distribution against the ideal normal. The plot shows slight skewness, but overall the data appear normal. Thus, there is little, if any, misbehavior in the model's residuals.

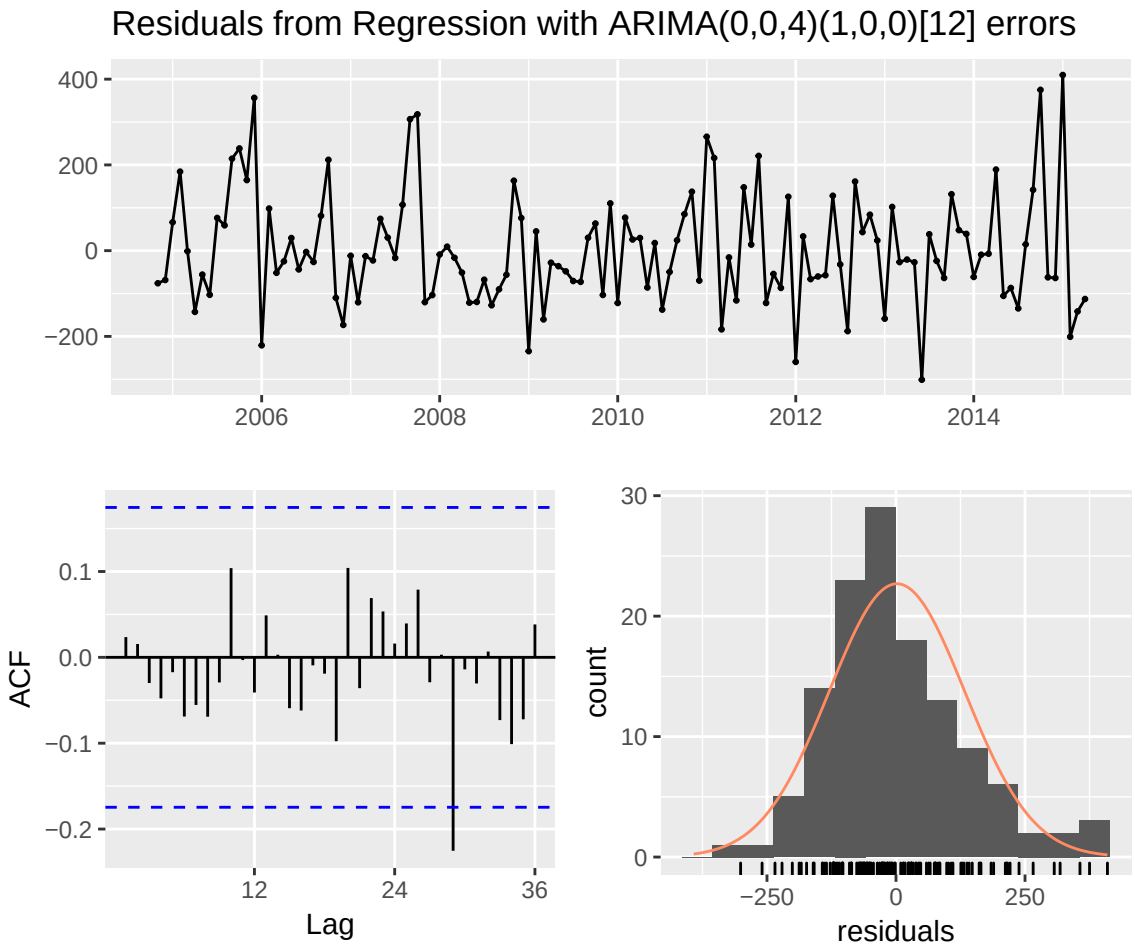

Figure 11. Initial Attrition Model - Residual Analysis

Table 6. Initial Model - Autocorrelation Test

Test type	Test statistic	p-value
Box-Ljung test	10.121	0.812

Forecasts are generated from the training data and compared against the validation data. Figure 12 plots the training and validation data against the model predictions. Large movements are generally captured, even if not perfectly forecast. To compare modeling performance, the RMSEs from the models are compared in Table 7.

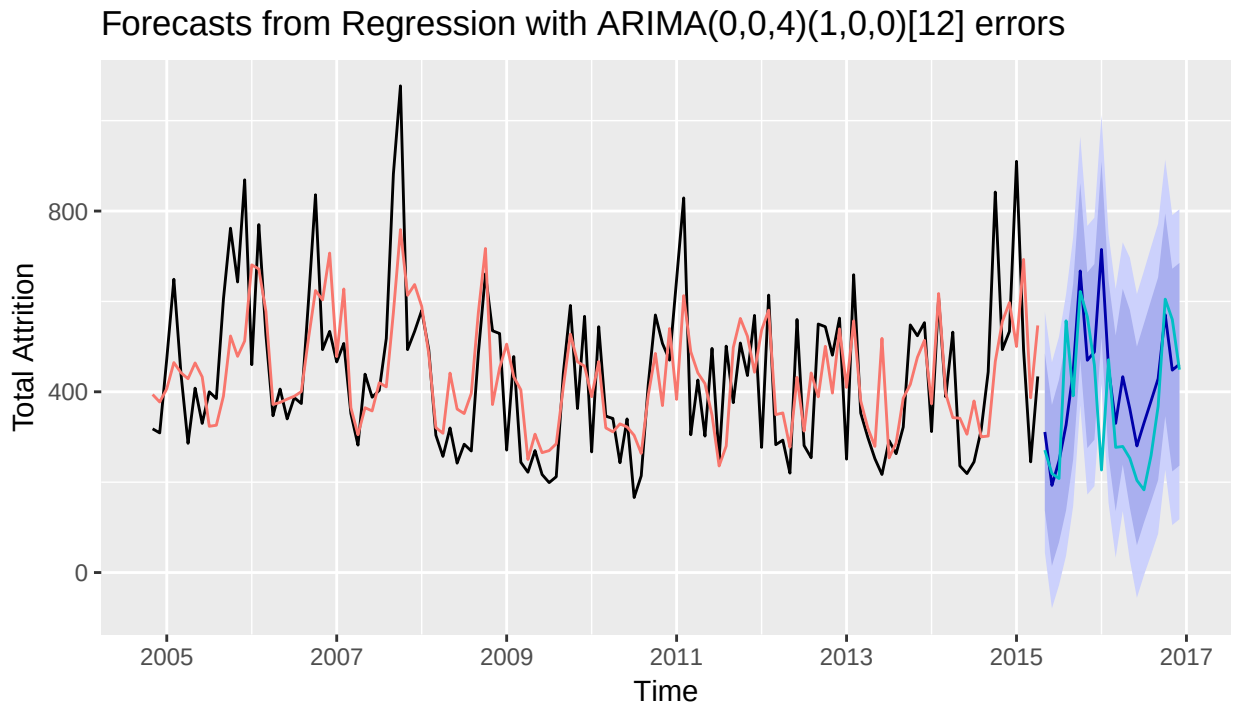

Figure 12. Initial Model - Forecasts Against Validation Data

Table 7. Model RMSE Comparison

	Simple Na\"ive	**Seasonal Na\"ive**	**Dynamic Regression**
Training	199.832	161.262	130.746
Validation	160.642	186.584	142.988

Dynamic regression demonstrates greater ability to forecast the attrition data than the na\"ive models. However, the high standard errors of the regression coefficients (β_1, β_2, and β_3 in Table 5) indicate that none of the economic indicators are statistically significant. This means the ARIMA model handles all the forecasting and the regression provides little insight. Essentially, the economic predictor variables do not explain much of the data variability. This could be for several reasons:

- With differencing, the indicators represent month-to-month changes. For most observations in the data set those changes are marginal, resulting in an insignificant effect on attrition, at least numerically.

- The economic and personnel data are both aggregated to the national level. It is possible that such a degree of aggregation includes enough noise to mask any economic effects.

- As they are, the indicators show only the previous month's change. The regression coefficients represent the effects last month's changes have on this month's attrition. Intuitively, this does not seem correct. Voluntary separation from the military is a long, bureaucratic process; as such, it is more probable that members decide to leave the military more than a month ahead of time.

Unfortunately, there is not much that can be done about the first point. As mentioned earlier, the indicators must be stationary in order to ensure the reliability of any potential effects, and the data must be differenced to be stationary.

3.2.4 Lagged Economic Indicators

Occasionally with time-series data, the effect of one variable on another is not be immediately observed. Consider a production firm redirecting profit towards self-investment. Ideally, this investment will lead to enhanced production capacity and higher revenue, though likely at a much later date. In the same sense, the current economic conditions could have a greater effect on attrition 12 months from now than they do today. In this section, the relationships between attrition and the lagged economic inidicators are explored.

The economic data is observed monthly over 12 years, and so there are many possible lag-periods to consider. It is also possible that the best lag-period is not identical for all predictors, so several combinations of different predictors lagged to different periods should be tested. This results in a very large test space. To decrease computational requirements, lag-periods are restricted to 0, 6, 12, 18, and 24 months. A separate dynamic regression model is generated for every combination of predictor and lag-period. This amounts to 125 dynamic regression models. The models are evaluated and compared on three metrics:

AICc, training RMSE, and validation RMSE. Any models that peform well by comparison are inspected further.

Table 8 below summarizes the values for each performance metric. Note that the minimum value for each category are below those seen in the previous model. This suggests that lagging the model's predictors can yield better results than using current values. The lagged models are thus invesitgated in greater detail.

Table 8. Summary Statistics - Lag Results

AICc	Training.RMSE	Validation.RMSE
Min. :1291	Min. :127.0	Min. :122.3
1st Qu.:1299	1st Qu.:133.0	1st Qu.:156.3
Median :1372	Median :134.8	Median :163.9
Mean :1361	Mean :134.6	Mean :164.8
3rd Qu.:1378	3rd Qu.:137.7	3rd Qu.:175.1
Max. :1613	Max. :140.4	Max. :187.6

The 1st quartiles of each performance criteria are used to filter the set of models, seeking models which perform well in all three categories. Only one model does, when the unemployment rate is lagged by 24 months, labor force participation rate by 18 months, and labor market momentum by 24 months.

Table 9. High Performance Across All Criteria

UR.lag	LFPR.lag	LMM.lag	AICc	Training.RMSE	Validation.RMSE
lag24	lag18	lag24	1292.005	126.9902	146.6523

Inspecting the model further reveals that one of the economic indicator is a significant predictor, unemployment rate lagged at 24 months. Unfortunately, none of the other predictors are significant in this model (shown in Table 10).

Top performers are also identified by comparing the best five models from each criteria individually and looking for commonalities. The results in Table 11 show that only AICc and Training RMSE have commonalities. The models in common are where the unemployment rate, labor force participation rate, and labor market momentum are respectively lagged

Table 10. Best Across All Criteria

term	estimate	std.error
ma1	-0.6686164	0.0993732
ma2	-0.2256966	0.1119664
sar1	0.8797210	0.1779036
sma1	-0.4942166	0.4519305
UR.lag.train[, "lag24"]	83.5668948	24.7823250
LFPR.lag.train[, "lag18"]	66.5150439	35.9139704
LMM.lag.train[, "lag24"]	-2.3988843	13.6970392

at (24, 18, 6), (24, 18, 18), and (24, 18, 24). The identified models are next inspected individually for coefficient significance.

Table 11. Common High Performers

UR.lag	LFPR.lag	LMM.lag	AICc	Training.RMSE	Validation.RMSE
Best by AICc					
lag24	lag18	lag6	1290.809	128.5034	164.7404
lag24	lag18	lag18	1290.939	128.5950	163.3188
lag24	lag18	lag12	1291.484	128.9926	165.1801
lag24	lag18	lag0	1291.670	129.1358	163.8772
lag24	lag18	lag24	1292.005	126.9902	146.6523
Best by Training RMSE					
lag24	lag18	lag24	1292.005	126.9902	146.6523
lag24	lag24	lag12	1292.050	128.2139	175.2866
lag24	lag24	lag6	1292.106	128.3596	171.1853
lag24	lag18	lag6	1290.809	128.5034	164.7404
lag24	lag18	lag18	1290.939	128.5950	163.3188
Best by Validation RMSE					
lag0	lag0	lag24	1306.135	135.0200	122.3425
lag0	lag6	lag0	1528.092	133.3393	134.2465
lag0	lag0	lag6	1527.850	133.2320	136.0557
lag6	lag6	lag6	1528.026	133.2758	138.7995
lag6	lag0	lag6	1528.028	133.3174	138.8589

The last common model (24, 18, 24) has already been inspected; it is the same one in Table 10. That leaves two models for comparison (24, 18, 6) and (24, 18, 18). Their coefficients are summarized in Tables 12 and 13 below, respectively. Both tables show similar results as the

previous model (24, 18, 24): The unemployment rate is a significant predictor, labor force particpation has a large effect but with high standard error, and labor market momentum has a small effect with a large standard error.

Table 12. Common High Performer 1

term	estimate	std.error
ma1	-0.6988796	0.0973710
ma2	-0.2255698	0.1092249
sar1	0.6372280	0.0883485
UR.lag.train[, "lag24"]	92.7925361	26.0922804
LFPR.lag.train[, "lag18"]	60.2058317	35.8537777
LMM.lag.train[, "lag6"]	10.9102055	11.7262875

Table 13. Common High Performer 2

term	estimate	std.error
ma1	-0.6768423	0.0997710
ma2	-0.2515895	0.1139329
sar1	0.6368839	0.0873641
UR.lag.train[, "lag24"]	93.4916257	25.9975212
LFPR.lag.train[, "lag18"]	61.6302177	35.3374730
LMM.lag.train[, "lag18"]	-10.2948441	11.9100709

Recall in Figure 8 that labor market momentum and labor force particpation rate are moderately correlated. Investigation into labor market momentum reveals that it is a combination of many economic indicators, including labor force participation. In short, labor market momentum repeats the information represented by the other two predictors and could be introducing multicollinearity issues. Given that information and the collection of estimates discussed above, labor market momentum is removed.

Lagged variable analyis is repeated with labor market momentum excluded from the list of possible predictors. 25 dynamic regression models are generated (with the identical lag periods) and compared. This time, no model falls under the 1st quartile for all performance criteria. Comparing the top five performers for each criteria does yield results. A dynamic

regression model with the unemployment rate lagged by 24 months and labor force participation rate lagged by 18 months falls into the top five performers under AICc and Training RSME. Notice that the repsective lag periods are identical to those in earlier models. Table 14 summarizes the coefficient estimates.

Table 14. Top Model - Reduced Model

term	estimate	std.error
ma1	-0.6951414	0.0974040
ma2	-0.2326008	0.1099844
sar1	0.6315523	0.0887194
UR.lag.train[, "lag24"]	93.0157912	26.2632321
LFPR.lag.train[, "lag18"]	62.4380155	35.8104021

Unfortunately, earlier trends in coefficient estimates hold. The unemployment rate has a noticeable and significant effect, and the labor force participation rate does not. It is worth mentioning, however, that the coefficient estimates are very close to those in earlier models, implying that labor market momentum is a redundant predictor. While these models provide some evidence of significant effects, it is possible that other combinations of economic indicators are better fit. In the next section, this idea is explored.

3.2.5 Alternative Economic Subsets

Section 3.2.3 notes that there exist other subsets of economic indicators with low correlation, besides unemployment, labor force participation, and labor market momentum. In particular, nonfarm job quits and labor force particpation have a correlation coefficient of -0.07, the next lowest after unemployment and labor force participation. This motivates examination of dynamic regression models which include the former pair. The models are generated and analyzed. Each predictor is lagged at 0, 6, 12, 18, and 24 months, and a dynamic regression model is produced for every combination. These models are compared by AICc, training RMSE, and validation RMSE, seeking to identify top performers.

From this analysis, only one model produces results comparable to previous models.

Table 15 displays the specified model and its perfomance scores. Though the scores are marginally worse than the best models from the previous iteration, the coefficient estimates are more important in providing insight about attrition.

Table 15. Alternative Predictors - Best Model

Quits.lag	LFPR.lag	AICc	Training.RMSE	Validation.RMSE
lag24	lag24	1301.586	137.252	192.0631

Unfortuntely, Table 16 shows no evidence that either predictor is statistically significant. The combination of labor force participation and nonfarm quits does not affect attrition. As with the intitial models, predictive capacity is likely managed by the ARIMA portion alone.

Table 16. Alternative Predictors - Coefficient Estimates

term	estimate	std.error
ma1	-0.8384799	0.0956497
sar1	0.7041322	0.0841965
Quits.lag.train[, "lag24"]	-0.1078249	0.0741378
LFPR.lag.train[, "lag24"]	63.5418324	43.4367079

3.2.6 Summary of Analysis

This section analysed nine economic indicators and hundreds of models, in several iterations. The best models were identified and explored, revealing some trends and signifcant predictors. The insights gleaned from this process are summarized and discussed in the next chapter.

IV. Conclusions and Insights

The research sought to identify economic indicators with statistically significant effects on attrition and to specify a mathematical model with which to build reliable forecasts of attrition. Regarding the former, nine separate economic indicators were initially considered (summarized in Table 1). Correlation analysis was used to identify subsets of variables exhibiting the least interdependence to avoid the effects of multicollinearity. Initial modeling found no statistically significant effects. However, subsequent attempts found that lagged economic indicators were significant. Specifically, the unemployment rate lagged by 24 months was found to be statistically signifcant in all of the top performing models. No other variables analyzed (labor market momentum, labor force participation, and nonfarm job quits) showed evidence of significant effects. Regarding forecasting capacity, all of the top performing dynamic regression models explored in Chapter III exhibited lower training and validation RMSE than the na\"ive models. That is, the dynamic regression technique, regardless of predictor significance, is better adapted for forecasting attrition than simply applying previous observations forward. This work confirms results found in previous endeavors (such as Jantscher [15]), reinforcing the relevance of the unemployment rate to attrition. This work builds on that knowledge by also providing a timeline, finding evidence that the current unemployment rate has a significant effect on attrition two years later.

Many possibilites were addressed in this work; four na\"ive models and 33,792 unique dynamic regression models (176 regression specifications, each with 192 ARIMA variations). By no means does this encapsulate the total set of possible models. There are many avenues left unexplored by this research. Future work in this area could, first, explore different subsets of the economic indicators identified. The initial set of economic indicators can also easily be expanded; the relevant data is freely accessible to the public. Only five unique lag-periods

were investigated, and the procedure discussed could easily incorporate more vaired time lags (at the cost of more time and computational complexity). Lastly, all data considered were aggregated to the national level and only total attrition across U.S. Air Force officers was evaluated. Future work in this area could, and should, investigate possible differences across AFSC groupings and, if possible, at a higher level of fidelity (e.g. state or county vs. national aggregates).

Appendix A. R Code

```
# check for req'd packages, install if not present
list.of.packages <- c("tidyverse", "lubridate", "sas7bdat", "fpp2", "reshape2",
                      "stargazer", "knitcitations", "RefManageR", "xtable",
                      "kableExtra", "zoo", "tictoc")
new.packages <- list.of.packages[!(list.of.packages %in% installed.packages()[,"Package"])]
if(length(new.packages)) install.packages(new.packages, repos="http://cran.us.r-project.org")

#library loadout
library(sas7bdat)
library(zoo)
library(lubridate)
library(reshape2)
library(kableExtra)
library(knitr)
library(gridExtra)
library(tictoc)
library(tidyverse)
library(fpp2)


# set directory for lazy data referencing - allow switch between macOS and Windows
# Basically just set working directory to wherever local repo is held
#setwd("~/Documents/Grad School/Thesis/github/afit.thesis/")
#setwd("C:/Users/Jake Elliott/Desktop/afit.thesis/")

#auto redirect working directory to file's location
setwd(dirname(sys.frame(1)$ofile))

# source file containing functions created for this analysis
#source("~/Documents/Grad School/Thesis/github/afit.thesis/custom-functions.R")
#source("C:/Users/Jake Elliott/Desktop/afit.thesis/custom-functions.R")

#auto redirect working directory to function file's location
source(paste0(dirname(sys.frame(1)$ofile), "/custom-functions.R"))


####################
#   Import Data    #
####################

# Personnel data
# simple list of AFSCs and description
afsc_list <- read.sas7bdat("Data/lu_ao_afs.sas7bdat")
# monthly records of assigned levels, broken out by AFSC - currently in longform
assigned <- read.sas7bdat("Data/assigned_levels.sas7bdat") %>%
  spread(AFS, Assigned)
# monthly separation counts, broken out by AFSC - currently in longform
```

```
attrition <- read.sas7bdat("Data/separations_count.sas7bdat") %>%
  spread(AFS, Separation_Count)

# Econ data
setwd("Data")
econ_data <- list.files(pattern = "*_natl.csv") %>%
  # Import data sets
  lapply(read_csv) %>%
  # Merge all sets
  Reduce(function(x,y) merge(x, y, by = "DATE"), .) %>%
  # Store data as tibble
  as.tibble()

# Now, because gdp per cap is observed only on a quarterly basis, we add it separately
# in order to handle the missing values resulting from a merge. This merge is
# from the previous in that it keeps all values from the existing data set, creating NAs
# where gdp_per_cap does not match with other observations. Merging the quarterly
# gdp_per_cap data with the monthly indicators results in NAs in gdp_per_cap where
# the dates do not match. We then handle NAs by extending the quarterly records throughout
# their respective quarters (e.g. the GDP per cap for Q1 2014 is applied to Jan-Mar 2014).
# Simultaneously, we will rename the variables for interpretability.

# Read in GDP per capita
gdp_per_cap <- read_csv("real09_gdp_percap.csv")
# Combine gdp per cap with econ_data, using a left-join (all.x = TRUE) to preserve the main data set
econ_data <- merge(econ_data, gdp_per_cap, by = "DATE", all.x = TRUE) %>%
  as.tibble() %>%
  # Rename column headers to something more meaningful
  select(Unemployment.Rate.Adj = UNRATE, Unemployment.Rate.NonAdj = UNRATENSA,
         CPI.Adj = CPIAUCSL, Nonfarm.Jobs.Adj = JTSJOL,
         Nonfarm.Jobs.NonAdj = JTUJOL, Labor.Force.Participation = LNS11327662,
         Labor.Market.Momentum = FRBKCLMCIM, Real.GDP.Per.Capita = A939RX0Q048SBEA,
         Nonfarm.Quits.Adj = JTSQUL, Date = DATE)

# The na.locf() command below carries a value forward through NAs until the next non-empty value is met;
# this is how we choose to represent the gdp per capita through an entire quarter
econ_data$Real.GDP.Per.Capita <- as.numeric(econ_data$Real.GDP.Per.Capita) %>%
  na.locf()


#######################################
#   Data Cleaning and Preparation     #
#######################################

# The dates in the personnel data are in total days since 1 Jan 1960 (SAS default)
# We'll need to reformat the date into something readable and mergeable with our
# econ set. Also, we create a new column - the total number of separations across
# all AFSCs. We'll also create totals for different categories of officers: rated,
# non-rated, line, etc.
attrition <- mutate(attrition, Total = rowSums(attrition[-1], na.rm = TRUE)) %>%
  mutate(temp = EOP_Date * 86400) %>%
  mutate(Date = as.POSIXct(temp, origin = "1960-01-01")) %>%
  within(rm(temp, EOP_Date))

# repeat prodcedure for assigned data
```

```r
assigned <- mutate(assigned, Total = rowSums(assigned[-1], na.rm = TRUE)) %>%
  mutate(temp = EOP_Date * 86400) %>%
  mutate(Date = as.POSIXct(temp, origin = "1960-01-01")) %>%
  within(rm(temp, EOP_Date))

# However, the date variables differ slightly between the econ and personnel sets:
# Though both represent monthly observations, the econ set default to the first
# of each month, and the personnel data defaulted to the last
# (e.g. October 2004 is represented as 01-10-2004 in the former, and 31-10-2004 in the latter)
# To handle this, we'll create new date variables in each set that have the days
# trimmed off, then merge. Merging isn't strictly necessary, but it is a convenient
# way to only keep those observations common to both data sets.
econ_data <- mutate(econ_data, "Date1" = paste(year(Date), month(Date)))
attrition <- mutate(attrition, "Date1" = paste(year(Date), month(Date)))
# Merge data sets
df <- merge(econ_data, attrition, by = "Date1")

# Next, we see many NAs within the attrition data set. Given the data's nature,
# our intuition was that these missing values aren't a result of encoding error
# or similar, but rather an indication that no separations occurred during
# that period (i.e. NAs indicate no separations were observed, instead of
# indicating some sort of error). This intuition was confirmed by the data's
# provider, HAF/A1FPX.
df[is.na(df)] <- 0

# Next we'll go ahead and drop all of our date variables. When we use df
# to create a time series object, the date variables become redundant.
df <- df[, !(names(df) %in% c("Date1", "Date.x", "Date.y"))]

# Now we'll initialize the time series object - start = Oct 2004, freq = 12 -
# and create the validation and training sets. Since we're only really interested
# in the Total column for modeling purposes
df.ts.1 <- ts(df, start = c(2004, 10), frequency = 12)
train.ts.1 <- subset(df.ts.1, end = 127)
val.ts.1 <- subset(df.ts.1, start = 128)

########################################
#   Initial Exploration and Modeling    #
########################################

# Let's take an initial, unmodified look at our response - total separations across
# all officer AFSCs. We see some pretty substantial spikes; fortunately, we know
# from the sponsor that they are artificially high. Special incentive programs
# for separation were implemented in the same years containing the spikes. So,
# we can do something about those observations - remove them, impute and replace, etc.
autoplot(df.ts.1[,'Total'], ylab = "Total Separations")

# We also will want to look for evidence of seasonality or for any one year that
# stands out. Grouping the separations by year, we can see that the tail ends of
# 2005, 2006, 2007 and 2014 were higher than other years (we saw this in the
# previous plot as well). Aside from those periods, however, no individual year
# stands out. We do notice, though, that separations appear to have a bowed shape
# as the years progress. That is, slightly higher levels of separation at the beginning
# and ends of the calenday year, with lower rates of separation during summer months.
# We may have to account for this seasonality in our modeling by transforming the data.
```

```
p <- ggseasonplot(df.ts.1[,'Total'], year.labels = TRUE, year.labels.left = TRUE) +
  ylab("Total Separations") +
  ggtitle("Seasonal Plot: Total Separations") +
  theme(legend.position = "none")
p

# However, we might want to try modeling before making any alterations to the data.
# It very well could be that we don't need to replace or impute values, that we're
# able to forecast fairly accurately without adjustments. We'll start with
# some naive forecasts against which we compare future forecasts.

n.1 <- naive(train.ts.1[,"Total"], h = dim(val.ts.1)[1])
sn.1 <- snaive(train.ts.1[,"Total"], h = dim(val.ts.1)[1])

n.1.error <- accuracy(n.1, val.ts.1[,"Total"])
sn.1.error <- accuracy(sn.1, val.ts.1[,"Total"])

# There are two takeaways from results below:

# First, we see that the seasonal model performs worse on the validation set,
# indicating that it is possibly overfit or overly affected by some outliers
kable(n.1.error, caption = "Na\\\"ive Performance", digits = 3, align = 'c')
kable(sn.1.error, caption = "Seasonal Na\\\"ive Performance", digits = 3, align = 'c')



# Plotting the forecasts against the validation data, we can see that outliers
# might be the source of the problem. The last spike, around 2014-2015, is carried
# through in the forecasts, resulting in high errors. We know from an in-depth
# discussion of the actual data that spike is an aberration. From this, we infer
# that outliers are going to be a problem, and probably ought to be handled.
autoplot(n.1) +
  autolayer(val.ts.1[,"Total"]) +
  theme(legend.position = "none") +
  ylab("Total Attrition")

autoplot(sn.1) +
  autolayer(val.ts.1[,"Total"]) +
  theme(legend.position = "none") +
  ylab("Total Attrition")

# First, though, we need to identify which exact data points are outliers. We can
# refer back to our season plot to help. On visual inspection, it appears that
# roughly Oct-Dec of '05-'07 and '14 stand out (possibly Sep '07, as well).
# These observations are backed by insight provided by the data's sponsor - HAF/A1.
# From them, we've found out that in 2006, '07, and '14 special separation
# programs were instituted in order to incentivize attrition. So, these outlying
# points probably reflect the effects of those special programs.
p

# To handle these, we'll calculate the average values of all other years during
# months and replace the current values. First, let's create slices of our
# response containing the months we're concerned with - December and November.
# We also want to grab the correpsonding indices for updating our series later.

dec <- subset(df.ts.1[,'Total'], month = 12)
```

```r
dec.ind <- which(cycle(df.ts.1[,'Total']) == 12)
nov <- subset(df.ts.1[,'Total'], month = 11)
nov.ind <- which(cycle(df.ts.1[,'Total']) == 11)

# oct <- subset(df.ts.1[,'Total'], month = 10)
# sep <- subset(df.ts.1[,'Total'], month = 9)

# Referring back to p, and combining the graphical insights with information
# from the data's sponsor, we assume that 2006, '07, and '14 are the years which
# saw the largest effects from the separation incentive programs - i.e. artificial
# attrition. Those correspond the the 3rd, 4th, and 11th indices. So now, we
# replace those observations with the average of the non-aberrant years.

dec[c(3,4,11)] <- mean(dec[-c(3,4,11)])
nov[c(3,4,11)] <- mean(nov[-c(3,4,11)])

# And finally, we place these values back into the original series.
df.ts.1[dec.ind, 'Total'] <- dec
df.ts.1[nov.ind, 'Total'] <- nov

# Revisiting the response and seasonality plots, we can more easily see the
# effects of seasonality, and a much more stationary data set without any
# egregious outliers.

autoplot(df.ts.1[,'Total'], ylab = "Total Separations")

ggseasonplot(df.ts.1[,'Total'], year.labels = TRUE, year.labels.left = TRUE) +
  ylab("Total Separations") +
  ggtitle("Seasonal Plot: Total Separations") +
  theme(legend.position = "none")

# Lastly, we might want to retrain and assess our naive models so see if removing
# those outliers effected much

# store split index
set.split <- 127
# New train and val sets
train.ts.2 <- subset(df.ts.1[,'Total'], end = set.split)
val.ts.2 <- subset(df.ts.1[,'Total'], start = set.split+1)

# Train models and generate errors
n.2 <- naive(train.ts.2, h = length(val.ts.2))
sn.2 <- snaive(train.ts.2, h = length(val.ts.2))

n.2.error <- accuracy(n.2, val.ts.2)
sn.2.error <- accuracy(sn.2, val.ts.2)

# Compare the errors
kable(n.1.error, caption = "Na\\\"ive Performance")
kable(n.2.error, caption = "Na\\\"ive Performance")

kable(sn.1.error, caption = "Seasonal Na\\\"ive Performance")
kable(sn.2.error, caption = "Seasonal Na\\\"ive Performance")

# Aaaaaaand compare plots, forecasts
```

```r
# Nothing too special about these
autoplot(n.2) +
  autolayer(val.ts.2) +
  theme(legend.position = "none") +
  ylab("Total Attrition")

# Here we see that the variation in the validation set is more closely followed
# by the forecasts. We infer that removing the outliers was beneficial.
autoplot(sn.2) +
  autolayer(val.ts.2) +
  theme(legend.position = "none") +
  ylab("Total Attrition")

# Though our naive models aren't particularly useful for providing forecasts
# (or identifying key economic indicators), they're useful for providing a
# baseline comparison for other models. The idea being naive models are our
# simplest methods, and we'll compare the performance of more sophisticated
# against them - models such as...

# A multivariate regression model with ARIMA errors. Why this? Because a
# multivariate regression model allows us to include outside variables (economic
# indicators) to help predict a response (attrition). The problem with just
# regression, though, is that regression assumes independent errors, and we
# often find autocorrelation with time-series data. Enter the
# ARIMA: fitting an ARIMA model on our regression error, then, allows us to
# handle the autocorrelative nature of the data, but does not allow room for
# any exogeneous information (i.e. info other than the response).

# Separately, each of those methods provides roughly half of what we're looking
# to model. So, by our powers combined: We'll relax the assumption of independent
# errors in the regression model, and instead assume that they ARE autocorrelated.
# And since we have a model for predicting autocorrelated data, we now treat the
# 'error' term in the regression model as its own ARIMA model (technical
# formulation is in the thesis; you can also just search Google for
# "Regression with ARIMA errors"). We're left with a model that, when correctly
# specified, should provide both forecasts for our response and insight as to
# which variables contribute to those forecasts (response variance, etc).

# Now, before go fitting our data, we need to take some steps to ensure we're
# fitting it properly - there are other assumptions involved here. First, we need
# independent regressors. Collinearity between our regressor variables will
# inflate regression coefficients' variances; we won't have a good idea of how
# influential our economic indicators are. To avoid these issues, we'll build a
# heat map showing the correlation for every pairwise combination in our set of
# economic indicators.

# generate actual correlation heatmap
# Note: reorder.cormat(), get.upper.tri(), and %ni% are custom functions whose code
# can be found in the script custom-functions.R
df[which(names(econ_data) %ni% c("Date", "Date1"))] %>%
  cor() %>%
  round(2) %>%
  reorder.cormat() %>%
  get.upper.tri() %>%
  melt(na.rm = TRUE) %>%
```

```
  ggplot(aes(Var2, Var1, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Collinearity") +
  theme_minimal() + # minimal theme
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                   size = 12, hjust = 1)) +
  coord_fixed() +
  geom_text(aes(Var2, Var1, label = value), color = "black", size = 3) +
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank(),
    axis.ticks = element_blank(),
    legend.justification = c(1, 0),
    legend.position = c(0.6, 0.7),
    legend.direction = "horizontal") +
  guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                               title.position = "top", title.hjust = 0.5))

# The heatmap shows many isntances of collinearity, which is expected - many
# economic indicators are variations or flavors of the same information. However,
# some non-correlated groups are shown. For our initial model, we select the
# Labor Force Participation Rate, Market Momentum Index, and the Unemployment
# Rate. Though the first two have a noticeable correlation, we suspect they will
# still provide information for our model - the correlation isn't too strong
# anyway. The model can be adjusted and specified later, this is just a first
# stab.

# grab index of econ vars: LFPR, Unem.Adj, LMM
econ.vars <- which(names(df) %in% c("Labor.Force.Participation",
                                    "Unemployment.Rate.Adj",
                                    "Labor.Market.Momentum"))

# Now that we've selected our regressors, we need to check for stationarity. We're
# looking for evidence of non-zero trend, seasonality, etc.

autoplot(df.ts.1[,econ.vars], facets = TRUE) +
  ylab("")

# Yikes, okay not good. Definitely non-stationary. Let's try differencing, and
# see if that improves the situation.

econ.vars.d <- diff(df.ts.1[,econ.vars])

autoplot(econ.vars.d, facets = TRUE) +
  ylab("")

# Oh, yea - way better. Differenced, these variables look useable. Now while
# we're looking at differencing, we should look to see if our response also
# needs to be differenced. Remember, we have this:
```

```r
autoplot(df.ts.1[,"Total"]) +
  ylab("Total Attrition")

# Hmm, nothing too crazy, actually - is there any statistical evidence
# for differencing?

# regular differencing
ndiffs(df.ts.1[,"Total"])
# seasonal differencing
nsdiffs(df.ts.1[,"Total"])

# Neat! However, before we can build the model, we have to account for the
# differencing performed on our regressors. With simple differencing, we lose
# the first observation, and so we remove the first observation from our
# response.

head(df.ts.1[,"Total"])
head(econ.vars.d[,1])

response.ts <- ts(df.ts.1[-1,"Total"], start = c(2004, 11), frequency = 12)

head(response.ts)

# We also need to split up the econ vars into training and test sets

# store split index
set.split <- 126

# subset our response
train.ts.3 <- subset(response.ts, end = set.split)
val.ts.3 <- subset(response.ts, start = set.split+1)

# subset econ variables
econ.vars.d.train <- subset(econ.vars.d, end = set.split)
econ.vars.d.val <- subset(econ.vars.d, start = set.split+1)

# We'll utilize the auto arima function from fpp2
tic("dynamic regression")
dyn.reg.1 <- auto.arima(train.ts.3, xreg = econ.vars.d.train, trace = TRUE,
                        stepwise = FALSE, approximation = FALSE)
toc()
# The first thing we'll want to check, after the model summary, is how our
# residuals behave. Do they appear to satisfy normality assumptions? Are there
# any outliers? Evidence of leftover autocorrelation? Essentially, we're
# checking to see if there's anything other than white noise in the error term.

# No - to all of the above. ACF plots look clean, raw residuals seem
# noisy, and also have a roughly normal distribution. Furthermore, a Ljung-Box
# test shows no evidence that the data aren't normal (p: 0.8121).
checkresiduals(dyn.reg.1)

# Let's generate some forecasts then
dyn.reg.1.f <- forecast(dyn.reg.1, xreg = econ.vars.d.val, h = 20)

# We'll also take a look at some accuracy measures. According to RMSE and MASE,
```

```r
# the dynamic regression model performs better than a seasonal naive estimation;
# that's good news - we're getting closer towards accurate forecasting.
accuracy(dyn.reg.1.f, val.ts.3)

# And then plot those forecasts over the actual data
autoplot(dyn.reg.1.f) +
  autolayer(dyn.reg.1$fitted) +
  autolayer(val.ts.3) +
  theme(legend.position = "none") +
  ylab("Total Attrition")

# In short, our residuals are clean, error is improved from that of naive methods,
# and forecasts track the validation set fairly well. There's one glaring problem,
# however. Looking back at our model summary, it's clear that none of the estimated
# coefficients for the economic indicators are statistically significant. Boo.
# This could be for several reasons:
#
# 1) indicators are month-to-month changes, don't have large enough fluctuations
# to cause significant changes
#
# 2) no lagged information - i.e. current economic info probably doesn't affect
# current attrition rate
#
# 3) data might be too aggregated, contains too much noise to establish significant
# relationships

# In summary, dynreg gives better forecasts than naive models, but current
# specification doesn't reveal much in the way of economic insight

####################################
#   Specification - Lagged Effects  #
####################################

# Initially, we'll test 6, 12, 18, and 24 months for each indicator. We'll
# build a framework for any desired time-lag later.


# lazy, inefficient lagged variable build - if there's time later we'll build a
# function and generalize this process

Unemployment.Rate.lag <- cbind(
  lag0 = econ.vars.d[,"Unemployment.Rate.Adj"],
  lag6 = stats::lag(econ.vars.d[,"Unemployment.Rate.Adj"], -6),
  lag12 = stats::lag(econ.vars.d[,"Unemployment.Rate.Adj"], -12),
  lag18 = stats::lag(econ.vars.d[,"Unemployment.Rate.Adj"], -18),
  lag24 = stats::lag(econ.vars.d[,"Unemployment.Rate.Adj"], -24)
)

Labor.Force.Participation.lag <- cbind(
  lag0 = econ.vars.d[,"Labor.Force.Participation"],
  lag6 = stats::lag(econ.vars.d[,"Labor.Force.Participation"],-6),
  lag12 = stats::lag(econ.vars.d[,"Labor.Force.Participation"], -12),
  lag18 = stats::lag(econ.vars.d[,"Labor.Force.Participation"], -18),
  lag24 = stats::lag(econ.vars.d[,"Labor.Force.Participation"], -24)
)
```

```
Labor.Market.Momentum.lag <- cbind(
  lag0 = econ.vars.d[,"Labor.Market.Momentum"],
  lag6 = stats::lag(econ.vars.d[,"Labor.Market.Momentum"], -6),
  lag12 = stats::lag(econ.vars.d[,"Labor.Market.Momentum"], -12),
  lag18 = stats::lag(econ.vars.d[,"Labor.Market.Momentum"], -18),
  lag24 = stats::lag(econ.vars.d[,"Labor.Market.Momentum"], -24)
)

# create train and val splits
UR.lag.train <- subset(Unemployment.Rate.lag, end = set.split)
UR.lag.val <- subset(Unemployment.Rate.lag, start = set.split+1,
                     end = dim(econ.vars.d)[1])

LFPR.lag.train <- subset(Labor.Force.Participation.lag, end = set.split)
LFPR.lag.val <- subset(Labor.Force.Participation.lag, start = set.split+1,
                     end = dim(econ.vars.d)[1])

LMM.lag.train <- subset(Labor.Market.Momentum.lag, end = set.split)
LMM.lag.val <- subset(Labor.Market.Momentum.lag, start = set.split+1,
                     end = dim(econ.vars.d)[1])

# Initialize table to store results of loop. We are going to capture the
# variable combination, AICc, training RMSE, and validation RMSE. Tibble
# generated, saved to local .rds so we won't have to re-run this awful loop;
# runtime is approximately 1 hr on 4-core machine, process run in parallel


# lag.results <- tibble("UR.lag" = rep(NA, 125),
#                       "LFPR.lag" = rep(NA, 125),
#                       "LMM.lag" = rep(NA, 125),
#                       "AICc" = rep(NA, 125),
#                       "Training.RMSE" = rep(NA, 125),
#                       "Validation.RMSE" = rep(NA, 125))
#
# m <- 1
# for(i in c(1:5)){
#   for(j in c(1:5)){
#     for(k in c(1:5)){
#
#       xreg.train <- cbind(UR.lag.train[,i],
#                           LFPR.lag.train[,j],
#                           LMM.lag.train[,k])
#
#       xreg.val <- cbind(UR.lag.val[,i],
#                         LFPR.lag.val[,j],
#                         LMM.lag.val[,k])
#
#       dyn.model <- auto.arima(train.ts.3,
#                               xreg = xreg.train,
#                               stepwise = FALSE,
#                               approximation = FALSE,
#                               parallel = TRUE)
#
#       dyn.model.f <- forecast(dyn.model, xreg = xreg.val, h = 20)
#
```

```
#       dyn.model.err <- accuracy(dyn.model.f, val.ts.3)
#
#       lag.results[m, "UR.lag"] <- colnames(UR.lag.train)[i]
#       lag.results[m, "LFPR.lag"] <- colnames(LFPR.lag.train)[j]
#       lag.results[m, "LMM.lag"] <- colnames(LMM.lag.train)[k]
#       lag.results[m, "AICc"] <- dyn.model$aicc
#       lag.results[m, "Training.RMSE"] <- dyn.model.err[1,2]
#       lag.results[m, "Validation.RMSE"] <- dyn.model.err[2,2]
#
#       m <- m + 1
#     }
#   }
# }
#
# saveRDS(lag.results, "lagResults.rds")

# read in compiled lag.results
lag.results <- readRDS("lagResults.rds")

# summarize the selection criteria
summary(lag.results[,4:6])

# filter tibble to return rows where each metric is below respective first quartile
lag.results %>%
  filter(lag.results[,"Validation.RMSE"] <= 156.4 &
           lag.results[,"Training.RMSE"] <= 133.1 &
           lag.results[,"AICc"] <= 1299)

# only one result: 24, 18, 24, might want to go back and loosen filter criteria - for now
# let's investigate that one model

# best across three
xreg.train <- cbind(UR.lag.train[,"lag24"],
                    LFPR.lag.train[,"lag18"],
                    LMM.lag.train[,"lag24"])

xreg.val <- cbind(UR.lag.val[,"lag24"],
                    LFPR.lag.val[,"lag18"],
                    LMM.lag.val[,"lag24"])

dyn.reg.2 <- auto.arima(train.ts.3,
                        xreg = xreg.train,
                        stepwise = FALSE,
                        approximation = FALSE)

dyn.reg.2.f <- forecast(dyn.reg.2, xreg = xreg.val, h = 20)

autoplot(dyn.reg.2.f) +
  autolayer(dyn.reg.2$fitted) +
  autolayer(val.ts.3) +
  theme(legend.position = "none") +
  ylab("Total Attrition")

# we can also identify the 'top' model by the minimum of each criteria
top.models.1 <- rbind(lag.results %>%
```

```
                      filter(AICc == min(AICc)),
                    lag.results %>%
                      filter(Training.RMSE == min(Training.RMSE)),
                    lag.results %>%
                      filter(Validation.RMSE == min(Validation.RMSE)))

# take best five models according to each criteria and see if there are any
# commonalities

best.by.AICc <- lag.results %>%
                  arrange(AICc) %>%
                  head(5)

best.by.trainingRMSE <- lag.results %>%
                        arrange(Training.RMSE) %>%
                        head(5)

best.by.validationRMSE <- lag.results %>%
                          arrange(Validation.RMSE) %>%
                          head(5)

inner_join(best.by.AICc, best.by.trainingRMSE)
inner_join(best.by.AICc, best.by.validationRMSE)
inner_join(best.by.trainingRMSE, best.by.validationRMSE)

# Only best.by.AICc and best.by.training.MSE have a model in common

# we'll look more closely at the best model from each category and the model
# common to best.by.AICc and best.by.training.MSE - that's four more models

# Best from each category:

# AICc: 24, 18, 6
xreg.train <- cbind(UR.lag.train[,"lag24"],
                    LFPR.lag.train[,"lag18"],
                    LMM.lag.train[,"lag6"])

xreg.val <- cbind(UR.lag.val[,"lag24"],
                  LFPR.lag.val[,"lag18"],
                  LMM.lag.val[,"lag6"])

dyn.reg.3 <- auto.arima(train.ts.3,
                        xreg = xreg.train,
                        stepwise = FALSE,
                        approximation = FALSE)
checkresiduals(dyn.reg.3)

# AICc: 24, 18, 18
xreg.train <- cbind(UR.lag.train[,"lag24"],
                    LFPR.lag.train[,"lag18"],
                    LMM.lag.train[,"lag18"])

xreg.val <- cbind(UR.lag.val[,"lag24"],
                  LFPR.lag.val[,"lag18"],
                  LMM.lag.val[,"lag18"])
```

```
dyn.reg.6 <- auto.arima(train.ts.3,
                        xreg = xreg.train,
                        stepwise = FALSE,
                        approximation = FALSE)
saveRDS(dyn.reg.6, "dynReg6.rds")

# trainingMSE: 24, 18, 24 - already done, best under 1st quartiles

# validationMSE: 0, 0, 24 - problem: not interested in predictions based on
# current data, doesn't allow for forecasts into future - purely reactive, not
# proactive information

# most common model: 24, 18, 6 - already looked at with dyn.reg.3

# Q: So...what does dyn.reg.3 show?
# A: Similar pattern with previous models - UR is significant, LFPR is almost
# significant, and LMM is not. Investigation into LMM reveals that it is a
# result of Principle component analysis of several indicators, including
# UR. Explains the .56 corr with LFPR from the heatmap, and indicates that
# LMM and LFPR capture similar information. Corr might be causing
# inefficiencies in coeff of other two variables - UR and LFPR. Try dropping
# LMM and re-running lag analysis.

# lag.results.2 <- tibble("UR.lag" = rep(NA, 25),
#                         "LFPR.lag" = rep(NA, 25),
#                         "AICc" = rep(NA, 25),
#                         "Training.RMSE" = rep(NA, 25),
#                         "Validation.RMSE" = rep(NA, 25))
#
# m <- 1
# for(i in c(1:5)){
#   for(j in c(1:5)){
#
#       xreg.train <- cbind(UR.lag.train[,i],
#                           LFPR.lag.train[,j])
#
#       xreg.val <- cbind(UR.lag.val[,i],
#                         LFPR.lag.val[,j])
#
#       dyn.model <- auto.arima(train.ts.3,
#                               xreg = xreg.train,
#                               stepwise = FALSE,
#                               approximation = FALSE,
#                               parallel = TRUE)
#
#       dyn.model.f <- forecast(dyn.model, xreg = xreg.val, h = 20)
#
#       dyn.model.err <- accuracy(dyn.model.f, val.ts.3)
#
#       lag.results.2[m, "UR.lag"] <- colnames(UR.lag.train)[i]
#       lag.results.2[m, "LFPR.lag"] <- colnames(LFPR.lag.train)[j]
#       lag.results.2[m, "AICc"] <- dyn.model$aicc
#       lag.results.2[m, "Training.RMSE"] <- dyn.model.err[1,2]
#       lag.results.2[m, "Validation.RMSE"] <- dyn.model.err[2,2]
#
```

```
#       m <- m + 1
#   }
# }
#
# saveRDS(lag.results.2, "lagResults2.rds")

lag.results.2 <- readRDS("lagResults2.rds")

# Now that we have a data set with only 2 variables - UR and LFPR - let's
# the best models in the same manner as before

summary(lag.results.2[,3:5])

lag.results.2 %>%
  filter(lag.results.2[,"Validation.RMSE"] <= 151.8 &
           lag.results.2[,"Training.RMSE"] <= 133.9 &
           lag.results.2[,"AICc"] <= 1303)

# No single model falls below the 1st quartile for all three. Look at best by
# each criteria

top.models.2 <- rbind(lag.results.2 %>%
                        filter(AICc == min(AICc)),
                      lag.results.2 %>%
                        filter(Training.RMSE == min(Training.RMSE)),
                      lag.results.2 %>%
                        filter(Validation.RMSE == min(Validation.RMSE)))

# take best five models according to each criteria and see if there are any
# commonalities

best.by.AICc.2 <- lag.results.2 %>%
  arrange(AICc) %>%
  head(5)

best.by.trainingRMSE.2 <- lag.results.2 %>%
  arrange(Training.RMSE) %>%
  head(5)

best.by.validationRMSE.2 <- lag.results.2 %>%
  arrange(Validation.RMSE) %>%
  head(5)

inner_join(best.by.AICc.2, best.by.trainingRMSE.2)
inner_join(best.by.AICc.2, best.by.validationRMSE.2)
inner_join(best.by.trainingRMSE.2, best.by.validationRMSE.2)

# Best model by AICc from 2-variable (24, 18) has slightly better AICc
# and validation RMSE than that of 3-variable (24, 18, 6). Also, results from
# 2 model are comparable to our 'best' model from 3-variable (24, 18, 24).
# Let's look at the coefficients:

# lag2 best by AICc: 24, 18
xreg.train <- cbind(UR.lag.train[,"lag24"],
                    LFPR.lag.train[,"lag18"])
```

```
xreg.val <- cbind(UR.lag.val[,"lag24"],
                  LFPR.lag.val[,"lag18"])

dyn.reg.4 <- auto.arima(train.ts.3,
                        xreg = xreg.train,
                        stepwise = FALSE,
                        approximation = FALSE)
checkresiduals(dyn.reg.4)



# From this 'round' we can say our 'best' model is the 24,18: minimizes
# information loss, and provides similar results to the 'best' model from
# previous round. When faced with similar results, pick simplest - Occam's razor
# Now, we are getting mild results with this variable selection. Let's try
# including a different subset of variables. Referring back to the heatmap, we
# can see other subsets with low collinearity: LFPR and nonfarm quits, nonfarm
# quits and cpi. However, nonfarm quits and cpi are highly negatively correlated
# so we can either choose one to place with nonfarm quits or do both groups
# separately. FOr now, let's start with LFPR and nonfarm quits - if those
# results don't look great, we'll try the other subset.

#LFPR and NonfarmQuits

# need to difference nonfarmquits
econ.vars.2 <- which(names(df) %in% c("Labor.Force.Participation",
                                      "Unemployment.Rate.Adj",
                                      "Labor.Market.Momentum",
                                      "Nonfarm.Quits.Adj",
                                      "CPI.Adj"))

econ.vars.2.d <- diff(df.ts.1[,econ.vars.2])

autoplot(econ.vars.2.d[,c("CPI.Adj", "Nonfarm.Quits.Adj")], facets = TRUE)

# create lag set for Nonfarm quits, LFPR already exists
Quits.lag <- cbind(
  lag0 = econ.vars.2.d[,"Nonfarm.Quits.Adj"],
  lag6 = stats::lag(econ.vars.2.d[,"Nonfarm.Quits.Adj"], -6),
  lag12 = stats::lag(econ.vars.2.d[,"Nonfarm.Quits.Adj"], -12),
  lag18 = stats::lag(econ.vars.2.d[,"Nonfarm.Quits.Adj"], -18),
  lag24 = stats::lag(econ.vars.2.d[,"Nonfarm.Quits.Adj"], -24)
)

# create train and val splits for nonfarm quits
Quits.lag.train <- subset(Quits.lag, end = set.split)
Quits.lag.val <- subset(Quits.lag, start = set.split+1,
                    end = dim(econ.vars.2.d)[1])

lag.results.3 <- tibble("Quits.lag" = rep(NA, 25),
                        "LFPR.lag" = rep(NA, 25),
                        "AICc" = rep(NA, 25),
                        "Training.RMSE" = rep(NA, 25),
                        "Validation.RMSE" = rep(NA, 25))

# m <- 1
```

```
# for(i in c(1:5)){
#   for(j in c(1:5)){
#
#       xreg.train <- cbind(Quits.lag.train[,i],
#                           LFPR.lag.train[,j])
#
#       xreg.val <- cbind(Quits.lag.val[,i],
#                         LFPR.lag.val[,j])
#
#       dyn.model <- auto.arima(train.ts.3,
#                               xreg = xreg.train,
#                               stepwise = FALSE,
#                               approximation = FALSE,
#                               parallel = TRUE)
#
#       dyn.model.f <- forecast(dyn.model, xreg = xreg.val, h = 20)
#
#       dyn.model.err <- accuracy(dyn.model.f, val.ts.3)
#
#       lag.results.3[m, "Quits.lag"] <- colnames(Quits.lag.train)[i]
#       lag.results.3[m, "LFPR.lag"] <- colnames(LFPR.lag.train)[j]
#       lag.results.3[m, "AICc"] <- dyn.model$aicc
#       lag.results.3[m, "Training.RMSE"] <- dyn.model.err[1,2]
#       lag.results.3[m, "Validation.RMSE"] <- dyn.model.err[2,2]
#
#       m <- m + 1
#   }
# }
#
# saveRDS(lag.results.3, "lagResults3.rds")

lag.results.3 <- readRDS("lagResults3.rds")

# top model for each
top.models.3 <- rbind(lag.results.3 %>%
                        filter(AICc == min(AICc)),
                      lag.results.3 %>%
                        filter(Training.RMSE == min(Training.RMSE)),
                      lag.results.3 %>%
                        filter(Validation.RMSE == min(Validation.RMSE)))

summary(lag.results.3[,3:5])

lag.results.3 %>%
  filter(lag.results.3[,"Validation.RMSE"] <= 152.9 &
           lag.results.3[,"Training.RMSE"] <= 135.6 &
           lag.results.3[,"AICc"] <= 1303)

# none fall under 1st quartile for all three

# look at top 5 from each
best.by.AICc.3 <- lag.results.3 %>%
  arrange(AICc) %>%
  head(5)
```

```r
best.by.trainingRMSE.3 <- lag.results.3 %>%
  arrange(Training.RMSE) %>%
  head(5)

best.by.validationRMSE.3 <- lag.results.3 %>%
  arrange(Validation.RMSE) %>%
  head(5)

inner_join(best.by.AICc.3, best.by.trainingRMSE.3)
inner_join(best.by.AICc.3, best.by.validationRMSE.3)
inner_join(best.by.trainingRMSE.3, best.by.validationRMSE.3)

# Only trainingRMSE and ValidationRMSE have one in common, and model isn't
# usful as it uses lag0 variables (i.e. current data)

# let's compare the top models for each 'round' so far (order is AIC, train, val)
top.models.1
top.models.2
top.models.3

# only min AICc from 3rd round (LFPR and nonfarmquits) look comparable to other
# models, let's inspect the model more closely (24,24)

xreg.train <- cbind(Quits.lag.train[,"lag24"],
                    LFPR.lag.train[,"lag24"])

xreg.val <- cbind(Quits.lag.val[,"lag24"],
                  LFPR.lag.val[,"lag24"])

dyn.reg.5 <- auto.arima(train.ts.3,
                        xreg = xreg.train,
                        stepwise = FALSE,
                        approximation = FALSE)
checkresiduals(dyn.reg.5)

#results: residuals look 'okay', but not a clean as previous models, and
# none of the coefficients look to be significant

# final choice:
# dyn.reg.4: URlag24, LFPRlag18

#save all models used so they do not have to be regnerated
saveRDS(dyn.reg.1, "dynReg1.rds")
saveRDS(dyn.reg.2, "dynReg2.rds")
saveRDS(dyn.reg.3, "dynReg3.rds")
saveRDS(dyn.reg.4, "dynReg4.rds")
saveRDS(dyn.reg.5, "dynReg5.rds")
```

Bibliography

[1] Stephen S Fugita and Hyder A Lakhani. The Economic and Noneconomic Determinants of Retention in the Reserve/Guard Units. Technical report, DTIC Document, 1991.

[2] John Capon, Oleksandr S Chernyshenko, and Stephen Stark. Applicability of Civilian Retention Theory in the New Zealand Military. *New Zealand Journal of Psychology*, 36 (1):50, 2007.

[3] Tim Kane. *Bleeding Talent: How the US Military Mismanages Great Leaders and Why It's Time for a Revolution*. Palgrave Macmillan, 2012.

[4] Stephen P Barrows. Air Force Pilot Retention: An Economic Analysis. Technical report, DTIC Document, 1993.

[5] Beth Asch and James R Hosek. Looking to the Future: What Does Transformation Mean for Military Manpower and Personnel Policy. Technical report, DTIC Document, 2004.

[6] Beth J Asch. Designing Military Pay: Contributions and Implications of the Economics Literature. Technical report, RAND Corporation, 1993.

[7] Thomas R Saving, Brice M Stone, Larry T Looper, and John N Taylor. Retention of Air Force Enlisted Personnel: An Empirical Examination. Technical report, DTIC Document, 1985.

[8] Gary R Grimes. The Effects of Economic Conditions on Overall Air Force Officer Attrition. Technical report, DTIC Document, 1987.

[9] Saul I. Gass. Military Manpower Planning Models. *Computers & Operations Research*, 18(1):65 – 73, 1991. ISSN 0305-0548. doi: https://doi.org/10.1016/0305-0548(91)90043-Q. URL http://www.sciencedirect.com/science/article/pii/030505489190043Q.

[10] Gregory D Gjurich. *A Predictive Model of Surface Warfare Officer Retention: Factors Affecting Turnover*. PhD thesis, Monterey, California. Naval Postgraduate School, 1999.

[11] Turgay Demirel. A Statistical Analysis of Officer Retention in the U. S. Military. Master's thesis, Naval Postgraduate School, 2002. URL https://pdfs.semanticscholar.org/d117/d7abcff4ea6ac90ba8a206e80cb9515ce235.pdf.

[12] Sunil Ramlall. A Review of Employee Motivation Theories and their Implications for Employee Retention within Organizations. *Journal of American Academy of Business*, 5(1/2):52–63, 2004.

[13] Jill A Schofield. Non-Rated Air Force Line Officer Attrition Rates Using Survival Analysis. Master's thesis, Air Force Institute of Technology, 2015.

[14] Courtney N Franzen. Survival Analysis of US Air Force Rated Officer Retention. Master's thesis, Air Force Institute of Technology, 2017.

[15] Helen L Jantscher. An Examination of Economic Metrics as Indicators of Air Force Retention. Master's thesis, Air Force Institute of Technology, 2016.

[16] All Employees: Total Nonfarm Payrolls, Dec 2017. URL https://fred.stlouisfed.org/series/PAYEMS.

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
01-03-2018	Thesis	

4. TITLE AND SUBTITLE
Air Force Officer Attrition: An Econometric Analysis

5a. CONTRACT NUMBER
DACA99-99-C-9999

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Elliott, Jacob, T, 1st Lt, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(ES)
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/Department of Operational Sciences)
2950 Hobson Way, Building 640
WPAFB OH 45433-7301

8. PERFORMING ORGANIZATION REPORT NUMBER

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

10. SPONSOR/MONITOR ACRONYM(S)

11. SPONSOR/MONITOR REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Many organizations are concerned, and struggle, with personnel management. Training personnel is expensive, so there is a high emphasis on understanding why and anticipating when individuals leave an organization. The military is no exception. Moreover, the military is strictly hierarchical and must grow all its leaders, making retention all the more vital. Intuition holds that there is a relationship between the economic environment and personnel attrition rates in the military (e.g. when the economy is bad, attrition is low). This study investigates that relationship in a more formal manner. Specifically, this study conducts an econometric analysis of U.S. Air Force officer attrition rates from 2004-2016, utilizing several economic indicators such as the unemployment rate, labor market momentum, and labor force participation. Dynamic regression models are used to explore these relationships, and to generate a reliable attrition forecasting capability. This study finds that the unemployment rate significantly affects U.S. Air Force officer attrition, reinforcing the results of previous works. Furthermore, this study identifies a time lag for that relationship: unemployment rates were

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE		30	19b. PHONE NUMBER (Include area code) 937-255-3636 x7469advisor@afit.edu