Assignment 8: Final – BJA Report – Box-Jenkins Model: BJA

Predict 411

Section 56

Winter Quarter

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

School of Continuing Studies

Northwestern University

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Daniel Prusinski

In compliance for Master of Science Predictive Analytics

Bachelor of Science Business Marketing

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Program Analyst

Wooddale Church

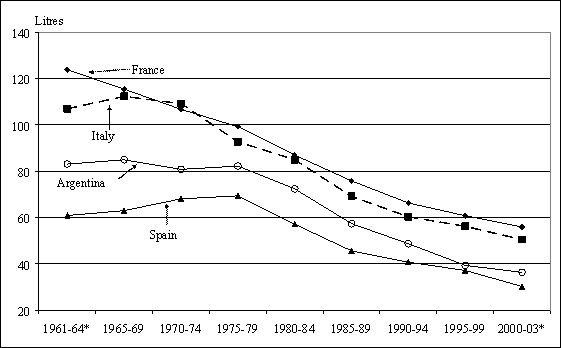
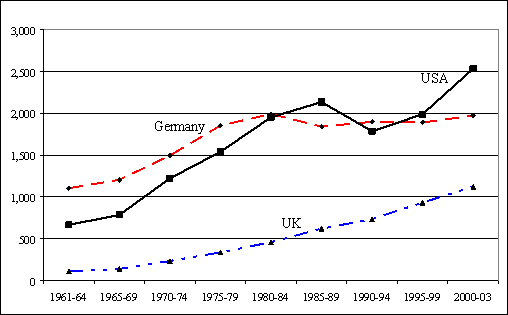
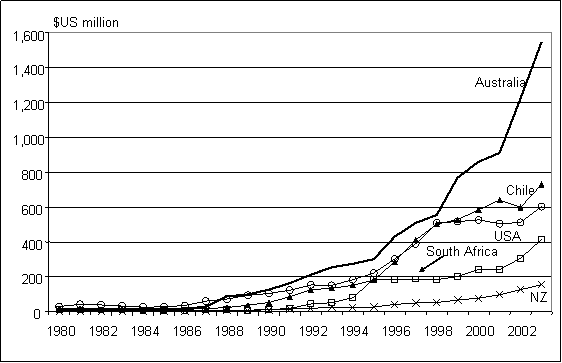
6630 Shady Oak Road

Eden Prairie, MN 55344

**Executive Summary**

From 1935-1954, the United States faced some of its hardest times as well as best economic growth periods. The Great Depression, World War II, Reconstruction, and the Korean War all took place in this twenty-year span. US companies that were titans of industry prior to The Great Depression were susceptible to the macro world trends. As a result of these major world trends, analyzing financial data from this time period requires different modeling techniques. The familiar Ordinary Least Squares (OLS) model proves insufficient given that the stringent assumptions cannot be validated based on heteroscedasticity amongst residuals between the manufacturing firms being studied. In response to the heteroscedasticity, Seemingly Unrelated Regression (SUR), and Feasible Generalized Least Squares (FGLS) yield models that are more robust and satisfy the general assumptions for regression analysis. The data from this exploratory data analysis (EDA) reflects major world events and demonstrates the financial winners for this time period. General Motors, General Electric, and Chrysler emerged from this time period as winner in respect to growth and increased financial valuation. Through this EDA, economists can better understand contagions and develop strategies to offset such events.

**Introduction**

Global consumption for wine has steadily increases over the past ten years. The world’s largest wine producing countries have experienced an overall decline in wine consumption. These two factors have driven down the global price for grapes, as well as increased the exportation of wine from traditional wine consuming countries to new markets (Wittwer & Rothfield). The graph to the right shows the consumption rates of wine (liters) for the four largest countries. It can be seen that there has been a negative trend of wine consumption for the last 50 years in the traditional wine consumer countries. From the 1960’s on, the world shifted to a more inclusive trading platform, which resulted in additional spirits in these countries. On the global stage, the USA, Germany, and the UK all increased their consumption of wine and importation of wine during the same time period. Essentially the two graphs show that while wine consumption decreased in France, Italy, Spain, and Argentina, it increases in the US, UK and Germany. While new markets were expanded into over the past 40 years, it is vitally important to analyze which countries exported the most wine during this time period.  The graph to the left shows which countries exportation of wine grew the largest over the past 20years. It can clearly be seen that Australia, Chile, and the US have been winning the export war of wine. Analyzing Australia’s production and exportation of wine is a helpful exercise in understanding successful export strategies.

Given the preliminary information above, I expect that as time increases so too does wine consumption as well as exportation for the Australian wine industry. The dataset in this EDA records the monthly sales of Australian wine makers. Specifically, the objective for this EDA is to utilize the Box-Jenkins methodology to analyze the wine data set, fit an appropriate model, and forecast 10 periods ahead.

**Analysis**

In order to meet the objective of exploring the relationship between the recidivism rate and the 12 variables, an exploratory data analysis (EDA) must be conducted. This EDA will start with a basic Proc Lifereg in SAS to begin the modeling process. From that model, analysis will be made based on the distribution that best fits the data.. As a data scientist in training, I am inculcating a paradigm of which to study data. While this paradigm is redundant report to report, it is training me to have the correct mindset. I will be utilizing the EDA paradigm and structure put forth by Bruce Ratner found in his book *Statistical and Machine-Learning Data Mining*:

Problem/Objective: Explore the relationship between the recidivism rate and 12 variables. Their interpretations were explained above.

Data: The data has been aggregated and has been supplied from management. There are no missing values.

Analysis: I will explore the data via simple statistics. After the initial analysis, I will start with fitting the data with a Proc Lifereg in SAS to pick a distribution.

Model: Proc Phreg in SAS will be used to fit a proportional hazard model to the dataset. In addition, I will analyze additional variables and weigh the efficacy of adding them to the model.

Results/Interpretation: Once the model has been created, an assessment of the model adequacy will be conducted to discern how well the model fits that data and the statistical backing of the model.

A properly executed EDA for management must reflect that the data was the driving force behind constructing the model. The steps outlined above are ordered such that the data drives building the best model, and the analyst’s personal bias is mitigated.

**Data**

Of interest is the time until they return to prison. The information was collected retrospectively by looking at records in April 1984, so the maximum possible length of observation is 81 months. Although there are several variables in the data set, the ones of interest are:

There are a total of 1445 observations with 0 completely missing values per row. Management has requested the focus to be on the following variables: durat - the duration in months until return to prison, cens - the censoring indicator variable, workprg - an indicator of participation in a work program, priors - the number of previous convictions, tserved - the time served rounded to months, felon - an indicator of felony sentences, alcohol - an indicator of alcohol problems, drugs - an indicator of drug use history, black - an indicator for African Americans, married - an indicator if married when incarcerated, educ - the number of years of schooling, and age - in months. The data is comprised of 18 total variables over 81 months. Thus, each observation, which is an individual, has 18 data values and there are 1,445 observations in this dataset which equals a total of 26,010 data values being studied in this EDA. Below you will find general descriptive statistics of the variables and their correlation with the response variable.

Durat - the duration in months until return to prison.

Cens - the censoring indicator variable.

**Results**

**Future Work**

Further recommendations on how this study can be improved upon are the following:

* Time permitting, it would be great to utilize the heteroscedasticity diagnostic tests to further validate this data set.
* It would be an absolute party to analyze other companies from different countries and compare their data output during this time period. Countries of interest would include Japan, Germany, Australia, New Zealand, and Switzerland. From this study, one could ascertain from an economic standpoint the winning and losing economies.
* Economists often compare the Great Recession to The Great Depression, and one could analyze the same companies data for these different time periods and look for similar trends.

Through this initial EDA, coupled with the future work recommendations, economists would gather pertinent information in regard to economic recovery macroeconomic events.

**References**

"recidivist". Oxford Dictionaries. April 2010. Oxford Dictionaries. April 2010. Oxford University Press. 18 February 2013 <http://oxforddictionaries.com/definition/english/recidivist>

Ajmani, V. (2009). *Applied Econometrics Using the SASSystem*. Hoboken: John Wiley & Sons.