**Econ 465 Final Project Milestone #1**

**Group and Topic Selection**

**25 Points – Due Lesson 8**

For this milestone, you will submit the names of your proposed group members, proposed project, and evidence that you have the information needed to complete the project. Please do the following steps, turn in a typed document with the “Milestone 1 deliverable” information at the beginning of class on **lesson 8**.

**Select a Group**

Groups may contain 2 to 3 cadets. Groups must be in the same section (so whole group can be available for the final presentation).

**Review and Select a Topic**

You will select a problem to solve or dataset to analyze. There are numerous data analysis problems on the kaggle website (<https://www.kaggle.com/competitions>) which you may work on for this project, or you can find an econometric dataset to analyze. Some possible sources for time-series econometric datasets are sites that track commodities prices, such as:

<http://www.indexmundi.com/commodities/>

<http://www.worldbank.org>

<http://research.stlouisfed.org/fred2/tags/series>

<https://www.quandl.com/c/markets/commodities>

If choosing a kaggle.com project, be sure you understand the broader context of the competition, the specific goal it is trying to achieve, and the data that are provided. Please be aware that not all problems are created equal. You may want to invest some time browsing the competitions and even inspecting the data and thinking about how you would address the problem. You might find that some are more or less time-consuming that you initially realized. For the econometric time-series datasets, be sure that enough data is available at a fine-enough fidelity to conduct thorough and adequate time-series analysis.

Although there are many interesting competitions, some are not appropriate for this course. If you find another competition you would rather do, you are required to review the following checklist and obtain approval *before* starting work. If the project meets all criteria, your Milestone 1 deliverable will include evidence that the criteria are met.

1. The competition must be focused on prediction, forecasting, or other use of econometric methods discussed in Econ 365/465. “Visualization” competitions are generally not appropriate, unless you can propose a specific question focused on econometric modeling.

2. The competition must be “open to all,” and you must be able to download the data files. Some competitions are “invitation-only,” while others block access to new entrants after a specified time. Make sure you can click on the data file links and complete the downloads.

3. The data files must be sufficiently small in size that Stata and your laptop are capable of handling the data. You should be aware of three limitations: (a) Stata/IC allows up to 2,047 variables in your dataset; (b) Stata/IC allows up to 798 independent variables in any model; (c) the entire dataset must be loaded into your computer’s RAM. The first two are unlikely to be a constraint unless you have a very large number of dummy variables. Given that you are likely to create new variables (which will further increase the memory requirements), a good rule of thumb is that the uncompressed data files should total no more than 1 GB (although smaller files will be easier to handle).

4. The dataset must be sufficiently large to allow for meaningful analysis. This will primarily be a problem with time-series data, especially if you can only get annual data back a few decades. At best you should strive to find quarterly or monthly data back 50 years or so.

**Scope the Problem and Background Research**

Once you have a topic, you should come up with a specific and well-scoped problem statement. This is what you will endeavor to use econometrics to solve. It should be specific to the topic at hand and answering it should address the customer's main question. For Kaggle projects, this will probably be given to you. If you just select a time-series data set from some web site, you will have to come up with the problem statement yourself.

At this time, some background information may also be helpful to aid your audience. There's a reason you picked this topic and that should come out in the background research you do. What motivated you to pursue this topic? Answer that with some outside referenced sources (at least two) and you'll be able to explain to the reader why you care about this topic. Think of this as a miniature “literature review” that aims to put the problem into a broader context. This literature could include (among other things) existing academic research on the topic or reports or newspaper articles that emphasize the importance of the problem. In addition to being included in your report, this may help to inform your analysis.

**Acquire Data**

If you join a competition on kaggle.com, you will be asked to create a free account on kaggle and agree to the terms of use before accessing the data. You do not need to do any more with the account than download the data. Be sure to download all relevant files and documentation (including a description of the project) as soon as possible; some competitions block access to the data after a certain point. If you are unable to access the data, the instructor cannot either.

If you decide to analyze a time-series data set, be sure to keep a copy of the original raw dataset as it was downloaded from the original source page. You should then "clean-up" this raw data to create a "tidy" dataset that is ready for use in econometric model building. Along with your tidy dataset, you should have a "code book" describing each variable and its values in the tidy dataset. Finally, there should be some fairly specific documentation about what you did to the raw data to create the tidy data set.

When conducting econometric modeling and forecasting, it is common to develop a model using a “training” dataset (often a random subsample of a larger dataset), then test the model using a “validation” dataset. The purpose of this procedure is to prevent “over-fitting” the model, which would overstate its ability to make out-of-sample predictions. Given that the validation sample generally does not contain the outcome, you should focus on the training dataset and concentrate your modeling efforts on it. If you have enough data and want to try your model as a forecasting tool, that will only help your grade on the assignment.

Entering a competition is not a required part of this assignment. The submission requirements, evaluation criteria, and deadlines of the competitions differ from those in this assignment. However, you are welcome to submit a response to the competition as long as you satisfy the rules specified in that competition. Your choice to do so will not affect your grade.

In this section, you should make the reader very confident in the fact that you have acquired the necessary data to conduct the research, you understand something about the data (what type of data it is, how many observations and variables it contains, and how much of the data is missing). You should also have a good plan for how you are going to process the data into a dataset set ready for analysis. Lastly, you should ensure that your software (Stata and **R** for some of you) will be able to handle the volume of the data.

**Analysis Plan**

Once you've described how you plan to tidy up your dataset, you should describe the basic analysis you plan to conduct. At a minimum, this should include an econometric model, one or more statistical tests, and a description of how estimating that model and performing those tests will answer your problem. This is just a plan, so you should also consider contingencies like some types of issues you believe you might encounter and any potential hindrances to analysis you think you'll have to deal with.

**Milestone 1 Deliverables**

After completing the above, your group should write up a small report that addresses the following:

**1.** **Administrative Details:** List the names of all members of your group. Also provide a link to the website for your project.

**2. Project Goal:** What motivated you to select this project? Provide some background information about the topic using at least two outside sources (not including the actual dataset). Finally, craft a clear and concise problem statement that will be the focus of your analysis and ultimately be what you address in your conclusion.

**3. Data:** Provide the teacher a statement that you have successfully acquired the data. Give some basic descriptive information about the data (size, type, degree of missing variables, etc.). Outline some of the data processing activities you believe you will have to complete. Lastly, assure the reader that your computer and Stata version can handle the dataset you've selected.

**4. Analysis Plan:** Depict your plan for how you will use econometrics to answer the problem statement. What kind of model(s) do you expect to create? What issues to you expect to encounter? Are there any potential hindrances you foresee? How will address and mitigate them?

Consider milestone 1 your proposal. Expect to write between 2 to 5 pages to adequately communicate this information to your teacher. The goal is to provide information to your instructor about what you want to study with some background, context, and assurances that the data is available for your use. You should not do much analytical work on your project before receiving approval on the project subject matter from your instructor. The instructor will ensure that all cadets are part of a group, and that a wide variety of projects are selected. If you wish, you may turn in Milestone 1 ahead of the deadline to receive early approval to begin work.

Group Members: Jacob Lindell and Patrick Mackintosh

Lt Col Gaupp

M3

Milestone 1

The project we would like to do was not found on Kaggle, but is more based off of a sports analytics project we did last semester in our OR 495 class. The project is to take NFL play-by-play data from 2013 and 2014 and use it to predict the next play that should be called based on different aspects; down, yardage, the offensive team’s strengths, and the defensive team’s weaknesses. Cadet Mackintosh and I are very interested in sports, so this is right up our alley.

We think that football analytics is coming along and going to be much more prevalent in the near future, and this will be one of the ways it will be used. We have heard that Gary Kubiak, the head coach of the Denver Broncos’, is trying to replace one of his coordinators on the headset with an analyst that will be running regressions and using statistics to determine the next play that should be run (Wagner-McGough). One of the very interesting analyses online is from the Green Bay vs. Seattle NFC Championship Game, in which Brian Burke breaks down multiple decisions throughout the game and statistically shows how each decision changed the outcome of the game (Burke).

The data that we will be using was given by Major Pietz during the OR 495 class last semester for the purpose of determining if team’s that run more than 20 times per game are more likely to win. The data includes all of the information collected during a given play; the offensive team, down, yardage, play choice (run or pass), play selected (run right, deep pass), player used, etc. We will need to organize all of the plays by team, and figure out a team’s strengths and weaknesses on both sides of the field in order to best select the next play that should be called. The data is given in Excel and is compatible with Stata.

The type of model used could be tricky because our prediction is categorical. Football plays are multifaceted, they could go left or right, throw or run, and each team has their own unique playbook. The first approach to this would be to have a multi-layered model with binary outputs. Probably multiple logit or probit models that would first predict the type of play and then where that play is targeted.

The inputs in the model are offensive team, down, yardage, play choice, play specification, gained yards, player used, etc. However, when it comes to play calling there are a lot of factors that go into it. Many lurking variables as well as variables that could have some multi-collinearity could be present. For example, certain teams could correlate with pass plays because of their roster, while maybe other teams might not be passing much because their star quarterback is out. Exploring the data beforehand to determine all these factors is a major step in keeping the integrity of our model.

Throughout the season, we can use information from previous games; the preseason, or the early weeks of the season, in order to evaluate the last portion of the season. The ultimate test would be to test our predictions against the Denver Broncos’ play-call, since they are one of the known teams that actively use analytics.

Works Cited

Burke, Brian. "GB-SEA Analysis." *Advanced Football Analytics*. N.p., 19 Jan. 2015. Web.

Wagner-McGough, Sean. "Broncos' Director of Analytics Will Help Gary Kubiak with Game Decisions." *CBSSports.com*. N.p., n.d. Web. 27 Aug. 2015.