**CP 6025: Advanced Planning Methods**

**Group Project Assignment**

You will work in assigned groups on a statistics project with four deliverables worth 25% of your grade.

The purpose of this project is to increase your familiarity with applying statistical methods to real world problems, research questions and data. With this project you will also practice your teamwork and presentation skills.

Overview of deliverables and due dates:[[1]](#footnote-1)

1. Research question, background theory, variables (5%): due 10/23
2. Cleaned dataset and data dictionary (5%): due 11/11
3. Model building, regression analysis, discussion of results (10%): due 12/1 (before lab)
4. Presentation (5%): due and presented on 11/29 (in lab)

Peer assessments of your contribution to the group project will factor into your individual grade. These will be administered after the presentation.

1. **Deliverable 1: Research question, background theory, and variables**

You will be required to provide the following information as answers to a series of questions, not in the form of a paper, typically 1-3 paragraphs each.

1. You will describe the problem area/topic you are interested to provide readers the background of the problem and to enable readers to understand the context/circumstances surrounding your research question.
2. You will also describe one research question you will investigate with multivariate regression. You might actually have a few questions you’re interested in but choose one main research question. When formulating and stating your research question, be mindful of the following:

* When forming your research question, try to clarify your outcome variable (i.e., dependent variable) and explanatory variable(s) (i.e., independent variable(s) of your interest). This would not be the place to describe all your variables (your controls).
* Clearly delineate your hypothesized relationship between the variables. It could be helpful for you to think of them in a form of null and alternative hypothesis.

1. You will describe the variables you suspect you will need to answer your research question, including control variables. Feel free to use tables to organize this content as a supplement to the written response.
2. You will use the Georgia Tech library online database and/or Google Scholar to perform a literature search for prior research related to your topic. Provide the citation and a brief write-up for 4-6 papers which help you to deduce which variables will be most useful for running a multiple regression. Describe their analysis and how it will inform your own. This literature search should also help you formulate a strong research question and help you think about how to set up a model. Note, in a formal paper and analysis, we would conduct a much more extensive literature review but that is not the focus of CP 6025.
3. **Deliverable 2: Cleaned dataset and data dictionary**

Your topic and research question should guide the data you are looking for. Data can be harder to come by than anticipated and take longer to clean, hence you should try to find data early.

You will upload a cleaned dataset, data dictionary, and the R script you used to clean your data to Canvas. For the cleaned dataset, ensure that it is in .csv format. You can use write.csv() function to save your cleaned data frame from R into a .csv file.

* Some of the important issues that must be clearly delivered to your readers (i.e., in this case, your instructors) include:
  + Whether variables have been transformed in any way;
  + Units or scales they are measured in (e.g., proportion in decimals or percentage / meter or kilometer / standardized or not standardized, etc.);
  + Unit of analysis (i.e., what do the rows represent? are they census tracts? persons? etc.)

For the format of data dictionary, you may refer to the MBB education data dictionary as an example. It's also common to format data dictionaries in excel. We are more concerned with understanding what your variables mean than with how you format the data dictionary.

For the R script, please make thorough comments so that it is easy to understand what each step of code is doing (e.g., subsetting, dropping NAs, joining tables, etc.) even without reading the code itself.

Instructions for data searching at the end of this document.

1. **Deliverable 3: Model Building, Regression Analysis & Discussion of Results**

You will be required to provide the following information as answers to a series of questions, not in the form of a paper, typically a few paragraphs each.

1. Using the combination of the results from stepwise regression, theory, and your understanding on the subject (developed in deliverable #1), develop and test various model specifications. In doing so, consider whether including an interaction term and/or a higher order term would be appropriate. Report various models that you tested in tables using stargazer or sjPlot package. Clearly label the model names so that when you discuss this process in your write up the reader can clearly follow it. Select your preferred model and describe why you prefer it.
2. Diagnose potential problems in your data and the preferred model specification. Check for the common problems (e.g., outliers, heteroskedasticity, non-linearity, multicollinearity) and report whether they are identified in your data/model and if so, how you attempted to address them.
3. Report your findings and conclusion from your model. Components that should be included are:
   * Interpretation of regression coefficients (size, direction, significance, etc.) including confidence intervals.
   * Practical vs. statistical significance
   * Goodness of fit of the model (Multiple and adjusted multiple R2, F-test, etc.).
   * Your overall conclusion regarding your research question. What did you find? What are the implications for practice?
4. What are the limitations of your data and analysis? In your answer, clearly describe what limitations they have and why you think they have them. A few examples of limitations are shown below.
   * Data - Lack of population representativeness, biased sample, data contamination, item non-response from survey, etc.
   * Validity of findings - Weak internal validity, not generalizable findings, etc.
   * Causality - Possibility of reverse causation, incorrectly framing association as causation, etc.
5. **Deliverable 4: Presentation**

Groups will present their findings briefly in lab using PowerPoint, Keynote, etc. The instructor will provide a template for you to follow as you structure your presentation. There are well established norms to giving these kinds of presentations.

You may split the task of presenting among the group members as you wish, but all should contribute to this deliverable. Presentations should be 8-10 minutes and we will allow a few minutes for questions after each from the audience.

You should carefully watch your classmates presentations and be prepared to ask questions of their method, findings, context, etc.

Please conclude the presentation with a short social media blurb that summarizes your findings, for any platform you wish.

**Instructions for Data Search:**

The goal is to find a dataset that’s related to a topic you’re interested in. It might be that you can combine a couple datasets to get all the variables you’re interested in. To do that, you’d need a common variable across both files, like a Census Tract. You’d also need to make sure your datasets are on the same year.

We want to find a dataset with 100+ observations and multiple variables (I’d say at least 10) to be able to use it for future work in the class. You must have one variable that is a location that can be mapped. Think you have a source? Show it to the TA to get the ok!

Here are some starting places to get you going:

1) See what’s available through the Georgia Tech library’s topic guides:

a) All guides: <https://libguides.gatech.edu/?b=s>

b) Two specific city planning guides:

i. <https://libguides.gatech.edu/?b=s>

ii. <https://libguides.gatech.edu/urbanpubpol/articles>

2) Spend 10-15 minutes searching the CP guides as well as any others you find that might have topics related to your interests.

3) Check out the “Data Sources” page I created on Canvas that you’ve had access to all semester. You’re welcome to add to this list!

4) Not finding what you’re looking for? Turn to Google! Run a few Google searches for data. See what researchers have been using by reading recent reports or journal articles.

5) In many cases, you may want to go out and collect data on your own if you don’t find what you’re looking for exactly. In this class we’re going to stick with what’s publicly available.

1. Note, these deliverable names do not match the syllabus. In restructuring the group project slightly to improve upon last year some reordering was necessary. [↑](#footnote-ref-1)