STAT GR5205 Project
10% Final Grade + EC
Due TBA

The STAT GR5205 Project is a "guided-open-ended" case study
intended to be a capstone on the linear regression models class.

1 Data Description

The data comprise of 166 countries and a number of variables. Most of the data comes from the CIA's "The World Factbook." For a detailed data description, please study The World Factbook website. A few additional variables are also included on `democracy index,` which is taken from Wikipedia: Democracy Index. As reference, the full **data dictionary** or **variable description** is displayed in Section 8.

This project has two parts:

I. Test the two research questions stated below in Section 2. To complete this part, you must fit a multiple linear regression model that includes appropriate variables, functional forms and interactions. You must also include appropriate diagnostics and remedial measures. Once deciding on a model, you are only testing variables related to the research question(s). You must also include appropriate diagnostics and remedial measures. Section 2 describes Part I in more detail and Section 3 describes the required writeup.

II. Build a *"predictive linear regression model"* intended to predict total length of roadways **or** life expectancy. Please choose only one of the two variables for prediction. Section 4 describes Part II in more detail and Section 5 describes the required writeup.

2 Part I: Research Questions

The goal of Part I is to run classic hypothesis testing procedures based on the multiple linear regression model. As a researcher your are interested how infant mortality rates and life expectancy impact the democracy index. The two research questions follow below:

1. Is there a statistically significant relationship between democracy index and infant mortality? (without controlling for life expectancy)
2. Is there a statistically significant relationship between the democracy index and life expectancy? (without controlling for infant mortality)

3 Part I: Writeup

Students are required to type up a final report. The final report should be broken up into the following four sections. Section (IV) has several components.

I. **Introduction:** Include a brief description of the goals of Part I coupled with some exploratory data analysis. Your exploratory analysis should include **a few important plots** and basic summary statistics that help support the research question. Be creative on the exploratory analysis and only include items that you feel are informative.

II. **Statistical Model:** In this section, clearly state your statistical model along with the `R` summary output. Be sure to describe all interactions, functional forms and transformations included in your model. Also include R^2 and R_a^2.

III. **Research Question:** Perform the relevant testing procedure(s) to answer the two research questions stated in Section 2. Also include a brief written summary of your results.

IV. **Appendix**

a. **Model:**

i. Here you will explain in detail what interactions, functional forms and variables you decided to include in the model. Describe if and why a transformation is applied to the response variable. Without overwhelming the TAs, include relevant `R` output and plots that helped you arrive at your statistical model.

b. **Diagnostics:**

i. Include all relevant diagnostic plots.

ii. Include a section on influential observations. For this application, we only care about testing the slopes related to the research question, thus you don't need to include plots for all $(DFBETAS)_j$.

iii. **Anything you Feel Necessary:**

4 Part II: Predictive Model

Build a statistical model intended to **predict** one (and only one) of the following two variables:

- **Total length of roadways**
- **Life expectancy**

Project Part II assesses students on building a **predictive model** as opposed to an **inferential model**, as in Part I. Students can choose from any technique introduced in this class, including:

- Multiple linear regression
- Robust regression (using a robust loss function)
- Ridge regression
- The LASSO

When using multiple linear regression (or MLR), you have total flexibility to include any functional forms and transformations of your variables. Also when using the using MLR, you are encouraged to run a model selection algorithm, e.g., *best subsets, stepwise, LASSO, etc...* I recommend trying a handful of **model candidates** and choose your final model based on its prediction error $MSPE$.

5 Part II: Writeup

Students are required to type up a final report. The final report should be broken up into the following three sections. Section (III) has several components.

I. **Introduction:** Include a brief description of the goals of this analysis coupled with some exploratory data analysis. This is similar to the Part I writeup, except your exploratory analysis should relate to prediction as opposed to inference. Include something that you found interesting based on your final predictive model. This section should be brief.

II. **Statistical Model:** In this section, clearly state your final **predictive model** along with the `R` summary output, or similar. Also include AIC, R^2, R_a^2, $MSPE$ and any other metrics you feel appropriate.

III. **Appendix**

   a. **Model:**

      i. Here you will explain in detail how you chose your final model.

   b. **Diagnostics and Model Validation:**

      i. include the computed $MSPR$ and compare this number to the computed MSE of your final model. In this section display and compare the performance of several models.

      ii. Include any graphics related to diagnostics or model selection. Again... don't overwhelm the TA.

      iii. Include a section on influential observations. For this application, we only care about prediction. Therefore you should only look at $DFFITS$ and *Cook's Distance*. Also note that $DFFITS$ and *Cook's Distance* only apply to the traditional multiple linear regression model.

      iv. **Anything you Feel Necessary:**

6 R Code

- Students should prepare an organized `Rscript` (or `Rmd`) file that complements the written report. This should have a `.R` or `.rmd` extension. Please include comments that help describe your model building process. Also describe your exploratory analysis, relevant diagnostics and how you tested the research questions.
- **Do not** copy and paste the `R` code into your appendix. Only include important `R` code in your final report. Please upload the `Rscript` (or `.Rmd`) file on Canvas by the due date.
- You are allowed to use another programming languages for this project but I personally recommend for students to use `R`. Note that other statistical languages sometimes have different parameterizations of common models. You are **not** allowed to use `Excel`, `Minitab` and `SPSS`.

7 Grading and Length

- This project will be graded on:
  1. Completeness (don't forget to turn in your `R` file also)
  2. Correctness
  3. Organization/neatness
  4. Creativity

  I want to see a nice organized final report for each project part. The report must typed with graphs labeled. **Please do not make the report too long!** (15 pages or less for both Part I and Part II)

- I am granting extra credit on this report. If the class TA feels that you did an excellent job, your hard work will be reflected when calculating final grades.
- This is **NOT** a group project.

8 Data Dictionary

The full data dictionary follows below:

Variable Name	**Variable Description**
name	country name
population	population
birth_rate	number of births per 1,000 population
death_rate	number of deaths per 1,000 population
infant_mortality_rate	infant deaths per 1,000 live births
life_exp_at_birth	overall life expectancy in years
life_exp_at_birth_m	male life expectancy in years
life_exp_at_birth_f	female life expectancy in years
gdpPPP	gross domestic product purchasing power parity in USD
gdpPPP_percap	gross domestic product purchasing power parity per capita in USD
labor_force	domestic labor force
land_area	in sq km
coastline	in km
land_use_agricultural	% of total land used for agriculture
urbanization	% of total population living in urban area
refined_petrol_consumption	in bbl/day
co2_emisssions_energy_consumption	CO2 emissions from consumption of energy in Mt
airports	number of airports in country
region	region
roadways	total roadways in km
democracy_index	**see below**
function_of_government	measure of democracy
political_participation	measure of democracy
political_culture	measure of democracy
civil_liberties	measure of democracy
regime_type	see the link Democracy
continent	continent
health_spend_pct_gdp	Health expenditure as a % of GDP

A few additional variables are also included on **democracy index**. This data is taken from Wikipedia: Democracy Index. The overall democracy index is the average of the other 5 variables:

1. electoral_process_and_pluralism
2. function_of_government
3. political_participation, political_culture
4. civil_liberties