STAT 410: Linear Regression
Final Project

Prof. Daniel R. Kowal

Spring 2022

Description

The goal of the project is to conduct a full regression analysis on a real dataset of interest to you. While the choice of the research question and the statistical tools are up to you, the basic requirements are (i) you must use a real dataset and (ii) you must include a regression analysis. The projects will be completed individually and the work must adhere to the Rice Honor Code.

Project Requirements

The project will feature three components:

1. **Project proposal:** each of you will submit a **one page** proposal for your project. The proposal should include the following:

   (i) a description of the **research question** you will answer and why it interests you;

   (ii) an overview of the **dataset** (or datasets) you will use; and

   (iii) a summary of which **statistical methods** you will use and why.

   The project proposal is due on **March 31 at 1:00pm** and is worth **5%** of the project grade.

2. **Peer review:** each of you will review one project proposal. You will also receive the review of your proposal by another classmate. The review should be at least **one paragraph** and must include the following:

   (i) at least one **supportive comment** about the strengths of the proposal;

   (ii) at least one **critical comment** about the limitations of the proposal;

   (iii) at least one **suggestion** for improving the proposal, such as a variation of the research question, an alternative dataset, or a statistical method that might help.

   The peer review is due on **April 4 at 1:00pm** and is worth **5%** of the project grade.

3. **Final report:** the final report will document your analysis and conclusions. It should be **no more than 10 pages** (excluding references and code) and include the following:

   (i) Project statement: what is the purpose of your analysis? What effect are you investigating? Why?

   (ii) Data description: how were the data collected? How do these data help you answer the relevant research questions?

   (iii) Exploratory data analysis: what are the basic features of the data? What are the variables? How are the variables related? Are there any unusual patterns?

   (iv) Data analysis: what is the statistical model (or models) that you are using? Why is this an appropriate model to use? Do the model diagnostics contradict the model assumptions? How do you interpret the results from the statistical analysis in the context of your research question?

   (v) Summary and discussion: what are the main conclusions of your analysis? What are the main limitations? What might be investigated in future research?

   (vi) References: what are the sources for the data and any additional statistical resources that you used to support your analysis?

   (vii) Code and data: include the code and data to reproduce your analysis. The code should include clear comments and should run correctly without errors.

   The final report is due on **April 29 at 11:59pm** and is worth **90%** of the project grade.

All documents must be submitted online through Canvas and use **12pt font, 1 inch margins, and double-spacing**.

Resources for selecting datasets

Many interesting datasets are publicly available. Here are some examples of repositories:

- Urban data platform (Houston): `kinderudp.org`
- FRED economic data: `fred.stlouisfed.org`
- Google dataset search: `datasetsearch.research.google.com`
- Nature scientific data: `nature.com/sdata`
- Kaggle: `kaggle.com`
- FiveThirtyEight: `data.fivethirtyeight.com`
- Our World in Data: `ourworldindata.org`
- Police data initiative: `policedatainitiative.org/datasets`
- CDC WONDER: `wonder.cdc.gov`

You are also welcome to collect your own data, such as surveys of your college or classmates, biophysical measurements on yourself over time, etc.

Suggestions for statistical analysis

Here are some suggestions for components of a statistical analysis. These are *suggestions*, not *requirements*, and other approaches or variations may be used. However, the statistical analysis must include a regression component and should be chosen to answer your research question of interest.

- Multiple linear regression: model the association between one response variable and p predictor variables. Are there hypothesis tests or confidence intervals that can address your research question? Are transformations (of the predictors and/or the response) needed? Should interactions or dummy variables be included?
- Bootstrap: how does inference from the bootstrap (e.g., confidence intervals) compare to that under the Gaussian assumption? Do the conclusions change? If so, why?
- Nonlinear regression: use polynomial regression, splines, or additive models to capture nonlinear effects in one or more variables. Are the nonlinear terms necessary? If so, are the nonlinear effects interpretable and explainable?
- Prediction: is there a statistical model and/or a set of variables that can predict a response variable of interest? For example, if you fit model to a subset of "training" data, can you predict the remaining "testing" data values? And do the prediction intervals "cover" the right proportion (e.g., 95%) of values in the testing data?
- Variable selection: which variables are needed in the model? Which variables are important? Are these results interpretable?
- **Recommendation:** avoid binary response variables!