MA206

Probability and Statistics
Project Tips for Success

AY26-1

In this document, we will cover the following topics as it pertains to the MA206 Project:

- Provide a brief discussion of *linear regression*
- Offer advice on good datasets and where to find them
- Explain some common pitfalls

So What is Linear Regression?

Think of regression as building a simple model of the world. We try to explain or predict one outcome (the *response*) using one or more pieces of information (the *predictors*). At its heart, regression is a form of **linear modeling**:

$$\hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \epsilon_i$$

Here the β coefficients are the *decision variables* that minimize some total error metric, and ϵ_i represents the leftover difference between the prediction and reality.

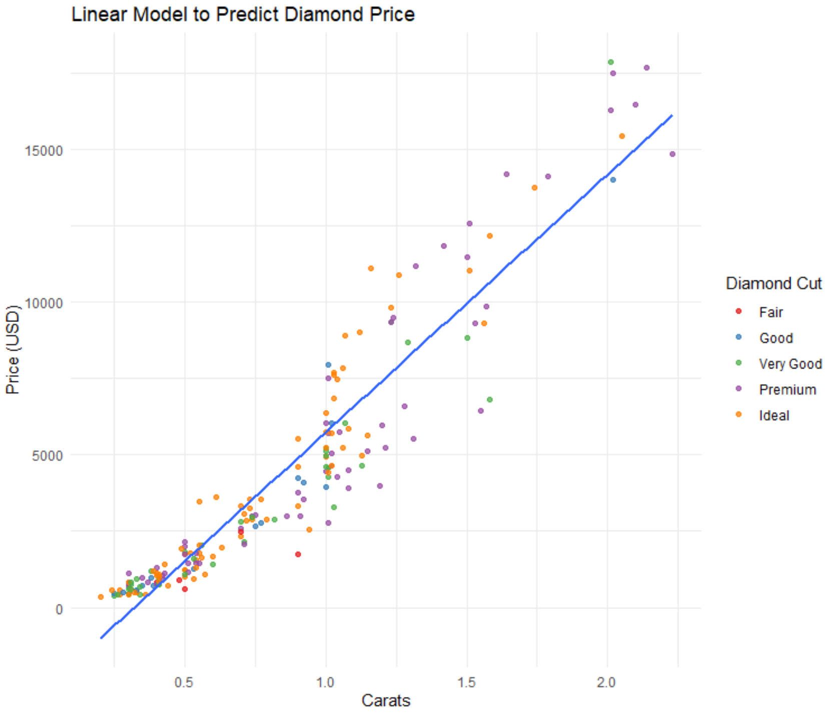

Figure 1: A “line of best fit” drawn to minimize total error.

Quick thought exercise: If you wanted to predict an athlete’s bench press, what information would you like to know? Age? Weight? Training hours? Sport? Each new piece of information can help reduce error and improve predictions.

How Do I Choose a Dataset for Regression?

Not every dataset is a good fit for regression. Keep these guidelines in mind:

- You need a **quantitative response variable** (e.g., exam score, salary, running time).
- Predictors can be **quantitative** (height, weight, study hours) or **categorical** (sex, yes/no responses), but they should plausibly relate to the outcome.
- More predictors are not always better. Start simple, then expand.
- A general rule of thumb is at least 30 observations (rows of data), but
- Merging datasets is allowed! For instance, you could take one dataset on country population and another on GDP, and combine them. This is much easier in `R` than by hand in Excel. If you want to try merging data, please talk with your instructor first.

Where Can I Find Data?

Here are some reliable starting points. Be creative—the most successful projects usually involve topics you are genuinely curious about.

- **Sports**
  - https://www.basketball-reference.com/
  - https://www.baseball-reference.com/
  - https://www.pro-football-reference.com/
  - https://www.pgatour.com/stats
  - https://www.spotrac.com/
  - https://scorenetwork.org/
- **Economics**
  - https://www.federalreserve.gov/data.htm
  - https://www.nber.org/research/data?page=1&perPage=50
  - https://library.bu.edu/economics/datasets
  - https://www.census.gov/topics/business-economy/data/datasets.html
- **Peace and Terrorism**
  - https://www.visionofhumanity.org/maps/#/
- **Music and Games**
  - https://research.atspotify.com/datasets/
  - https://vginsights.com/games-database

- **COVID & Public Health**
  - https://health.google.com/covid-19/open-data/
  - https://data.cdc.gov/
  - https://www.who.int/data
- **Miscellaneous**
  - https://www.gapminder.org/data/
  - https://data.world/
  - https://data.gov/
  - https://www.kaggle.com/
  - https://ourworldindata.org/

Final Advice: Start with a clear question, choose predictors thoughtfully, and let your model tell a story. A simple, well-explained model is often stronger than an overly complicated one. If you need assistance, run it by your instructor sooner rather than later!

What makes a (potentially) “not so great” project?

- Lack of effort
- Not reading the project requirements or prompts
- Not understanding model assumptions
- Fake or synthetic data
- Project partners that work in silos and create a disjointed product
- Small datasets (<100 observations)
- Poorly motivated research question (unanswerable using regression)
- Time variables (Sometimes! They get tricky!)