Lab 1 - Linear Regression: Models, Variability, and Misleading Certainty

Overview

In this lab, you will investigate how linear regression models behave when confronted with uncertainty, limited data, and unrealistic assumptions. While regression lines are powerful tools for describing relationships between variables, they are ultimately approximations of an underlying reality rather than exact truths.

Through a sequence of guided challenges, you will examine how regression lines can vary from sample to sample, how linear models can fail when extrapolated beyond the observed data, and how measures such as the coefficient of determination (R^2) should be interpreted with care.

AI Use Policy

In this lab, you will work in collaboration with an **AI agent (ChatGPT, Gemini, Claude, etc.)**, but you remain the chief statistician throughout the assignment. The AI should be treated as a support tool rather than a substitute for your own reasoning. You may use it as a brainstorming partner to help imagine scenarios and ask better questions, as a data generator capable of producing datasets efficiently, as a prediction machine whose outputs you are encouraged to test and challenge, and as a statistical sparring buddy that you are expected to question, critique, and occasionally prove wrong using sound statistical reasoning.

You are responsible for directing the investigation, performing all analysis and calculations, and communicating the final conclusions in your own words. While imaginative or playful contexts are encouraged, all scenarios must remain appropriate and non-offensive, and all written responses must be clear, concise, and academically appropriate.

Deadline: February 23; 4:00 pm

Part 1 – The Multiverse of Samples

How stable are regression results when they are based on small samples?

Task. In this part, you will use an AI agent to explore how regression lines can vary when multiple small samples are drawn from the same population. Your goal is to visualize and compare these differences and reflect on what they imply about the reliability of conclusions drawn from limited data.

Instructions.

1. Ask the AI to describe a large fictional population in which two quantitative variables have a clear linear relationship. Be creative in your choice of context.
2. From this same population, ask the AI to generate three distinct random samples, each of size $n = 8$.
3. For each sample, compute the least-squares regression line.
4. Create a single scatter plot using the data from one of the samples, and overlay the regression lines from all three samples on the same graph.

What You Need to Submit.

- A brief description of the population.
- The three datasets (Sample A, Sample B, and Sample C).
- One scatter plot showing the data from one sample with all three regression lines displayed.
- A short written analysis addressing the following:
  - Compare the slopes (b_1) and intercepts (b_0) of the three regression lines. How similar or different are they?
  - How might conclusions differ if only Sample A were collected instead of Sample B or Sample C?
  - What does this suggest about the reliability of regression results based on small samples?

Part 2 – To Infinity and Beyond

What can go wrong when a linear model is extended beyond the data on which it was built?

Task. In this part, you will use an AI agent to construct a scenario in which a linear relationship is reasonable over a limited interval but breaks down when extrapolated far beyond the observed data. You will compute a regression model, use it to make an extreme prediction, and then critically evaluate the result.

Instructions.

1. Choose a real-world or fictional scenario in which two quantitative variables exhibit an approximately linear relationship over a limited range. Examples include growth over a short time period, heating or cooling processes, or the early adoption of a new technology.
2. Ask the AI to generate a dataset containing at least 10 observations representing this *safe zone* where a linear model is reasonable.
3. Compute the least-squares regression line for your dataset.
4. Use your regression equation to predict a value for y corresponding to an x-value far outside the domain of your data.

What You Need to Submit.

- A brief description of your scenario, clearly explaining why a linear model is reasonable over the observed range.
- The dataset used to build the regression model.
- The regression equation.
- The extrapolated prediction, including the value of x used.
- A short written analysis addressing the following:
  - What value does the model predict?
  - Why is this prediction unrealistic, impossible, or misleading in the real world?
  - What type of mathematical model might be more appropriate for describing this scenario over a longer time span?

Part 3 – The Tale of Two Fits

What does the coefficient of determination actually tell us about a regression model and what does it leave out?

Task. In this part, you will use an AI agent to generate two datasets that share approximately the same regression line but have very different R^2 values. By comparing these datasets, you will investigate what R^2 does—and does not—tell us about the quality and usefulness of a linear model.

Instructions.

1. Ask the AI to generate two datasets involving the same explanatory and response variables.
2. The first dataset should exhibit a strong linear relationship (e.g. $R^2 > 0.90$).
3. The second dataset should exhibit a much weaker linear relationship (e.g. $R^2 < 0.40$).
4. The regression equations for both datasets should be approximately the same (for example, $y = 2x + 10$).
5. Create a scatter plot for each dataset and display the corresponding regression line.

What You Need to Submit.

- The two datasets.
- Two scatter plots (one for the high R^2 dataset and one for the low R^2 dataset), each with its regression line shown.
- The regression equation and R^2 value for each dataset.
- A short written analysis addressing the following:
  - How do the two plots differ visually, despite having similar regression equations?
  - Describe a real-world context in which a relatively low R^2 value could still represent a meaningful or important finding.
  - Describe a context in which an R^2 value close to 1 might still be considered insufficient.
  - In your own words, explain why R^2 should be interpreted as a measure of scatter rather than as a guarantee of model validity.