COGS 109: Homework 2

Due Friday, Oct 10 @ 11:59pm PT

Total estimated time to complete: 4-5 hours

Instructions (PLEASE READ)

1. Please copy and paste this entire assignment in your own document and write your answers below each question. **Leave unanswered questions blank** (don't delete them).
   a. 🚨Gradescope page matching: **please tag all pages that contain answers to the correct questions (don't skip any).** This will help our reader a lot when he is grading your homework!
2. For **full competition credit (3pts)**, please choose one of these two options:
   a. Complete Part A (1pt) + Part B (1.5pt) + Part D (0.5pt) = 3 points
   b. Complete 4 questions from Part A (0.5pt) + Part B (1.5pt) + Part C (0.5pt) + Part D (0.5pt) = 3 points
3. Overachiever option! If you choose to complete all parts in their entirety, you can earn the full 3.5 points.

Part A: Concepts (1 pt)

Estimated time to complete: 1.5 hours

These homework questions follow the order of topics in ISLP Ch. 3.1-3.3. Having the textbook open while you do the homework will be very helpful 🙂 .

Lecture 4 - Simple Linear Regression

1. Model setup.
   a. State the **simple linear regression** (SLR) setup (ISLP Eq. 3.1) and define each term in the equation.
   b. List and briefly explain the basic assumptions of the SLR model.
2. **Model fitting.** In your own words, explain how the coefficients of the SLR model are estimated (i.e., how we "fit" the model).
   a. Reference the **Residual Sum of Squares (RSS)** and describe how the minimizing coefficients are obtained. (We unfortunately don't have time to go through the full derivation of the least squares estimators in class. If you're interested, you can review this video in your own time.)
3. Interpretation. In your own words…
   a. Explain the difference between the population regression line and the least squares line.
   b. Define the standard error and explain what it tells us about the estimated parameters.
   c. Define confidence intervals and explain what it tells us about the estimated parameters.
4. **Applied example.** Consider the SLR model: $examScore \approx \beta_0 + \beta_1 \times hoursStudied + \epsilon$
   a. State the null and alternative hypotheses being tested in this model.

   You fit the model and obtain the following results table:

Term	Coefficients (β)	Std. Error	t	p-value
Intercept	3.4	0.80	4.25	<0.001
HoursStudied	1.5	0.20	7.50	<0.001

b. Interpret each *coefficient*: write one sentence for each coefficient explaining its meaning in context of the problem.
c. Are the coefficients statistically significant? Explain how you came to that conclusion.
d. Suppose the 95% confidence intervals (CIs) for $\beta_0 = [1.8, 5.0]$ and for $\beta_1 = [1.1, 1.9]$. Write one sentence of interpretation for each CI in context.
e. For your model, you compute RSE=5.2 and R^2=0.62. Write one sentence for **each** metric that interprets it in the context of the problem.

Lecture 5 - Multiple Linear Regression

5. Model setup.
   a. State the **multiple linear regression** (MLR) setup (ISLP Eq. 3.19) and define each term in the equation.
   b. List and briefly explain the basic assumptions of the MLR model.
6. Interpretation.
   a. Explain why simple and multiple linear regression models might disagree about the relationship between a single predictor and the output.
   b. Describe how to interpret each coefficient β_j in the presence of other predictors.
   c. Explain what overall *F-statistic* tests for. Why might we look at the F-statistic in addition to individual p-values for individual predictors?

Lecture 6 - Other Considerations in Regression

7. Model assumptions and extensions.
   a. Provide one example each of how the additivity and linearity assumptions of the linear regression model can be relaxed.
8. Categorical predictors.
   a. Imagine that someone would like to predict $house\ price$ from $square\ footage,\ location$ and $number\ of\ bedrooms.$
      i. Write down an appropriate regression model.
      ii. Explain one way you could code the qualitative (categorical) predictors in the regression model.

Part B: Coding Exercise (1.5 pt)

Estimated time to complete: 2.5 hours

Based on feedback from Homework 1, we will now provide Jupyter Notebooks (.ipynb files) as code templates for this portion of the homework.

Please download and complete hw2.ipynb. You may work on your own computer or use Google Colab. All cells marked with # YOUR CODE HERE must be completed for the notebook to run.

Overview

In this assignment, you will explore the process of **building and evaluating linear regression models** using the **ISLP College** dataset.

You will:

- Conduct **exploratory data analysis (EDA)** to understand relationships between predictors and the response variable.
- Fit and interpret a **simple linear regression (SLR)** model.
- Extend to a **multiple linear regression (MLR)** model and apply **backward selection** to refine your predictor set.
- Estimate and interpret **model coefficients**, **p-values**, and the **F-statistic**.
- Use the fitted model to **generate predictions** and compute the **root mean squared error (RMSE)** as a measure of test accuracy.
- Create and interpret **diagnostic plots** to evaluate model assumptions and summarize findings.

By the end, you will have a complete workflow for fitting, evaluating, and communicating results from linear regression models.

Submission Instructions

- Please write your answers and include all requested plots **directly below each corresponding question** in THIS DOCUMENT.
- **Export the completed Jupyter notebook as a PDF** and attach it to the end of your homework submission.

Part 1 — Exploratory Data Analysis (EDA)

1. Examine the code and complete cells with the comment # YOUR CODE HERE. Run the notebook.
2. Which predictors did you choose for your EDA (your **top_nums** list)? Paste the scatterplots of each chosen predictor versus Outstate below.
3. Which predictors appear roughly linear with Outstate? Do you observe any curvature, clusters, or outliers?

Part 2 — Simple Linear Regression (SLR)

4. Select one numeric predictor (x1) from your EDA to fit the model Outstate ~ x1. Which variable did you choose?
5. Plots (include below):
   a. Scatterplot of Outstate vs. x1 with the fitted regression line.
   b. Residuals vs. fitted values plot (from your model).
   c. Do you see any violations of the linearity assumptions? Please comment.

Part 3 — Multiple Linear Regression (MLR) with Backward Selection

6. Review ISLP Chapter 3.2.2, especially "Two: Deciding on Important Variables." Briefly describe the three methods for selecting variables in a multiple regression model.
7. Which variables did the backward selection process keep? Write out your final MLR formula after backward selection (e.g., Outstate = β_0 + β_1Top10perc...).
8. Interpret the final model output:

   a. State the null and alternative hypotheses tested by the F-statistic in your final MLR model.
   b. Report the F-statistic and its p-value and explain what they imply about the model overall.
   c. For each kept predictor, write one sentence interpreting the coefficient in context (e.g., "Holding other predictors fixed, a 1-unit increase in Expend is associated with an average increase of ___ in Outstate.").

Part 4 — Prediction, RMSE, and Visual Evaluation

9. What are your training and test RMSE values? Is there evidence of overfitting or underfitting? Explain briefly.
10. Visual assessment (include plots below):
    a. Predicted vs. Actual (test set) scatterplot with a 45° reference line.
    b. Residuals vs. Fitted (test set) plot.
11. Summarize your findings:
    a. Comment on the MLR's model's accuracy and how well it predicts Outstate.
    b. Mention one limitation of your model and a next step you would take to improve it.
12. Finally, please check one more time that you **included all plots and answers below each corresponding question**. Export your Jupyter Notebook as a PDF and attach it to the end of your homework document before submitting. Please tag/match all answers to the correct questions in Gradescope (don't skip any).

Part C: Exam Question (0.5 pt)

Estimated time to complete: 30-45 mins

This section is optional and only needs to be completed if you selected **option (b)**:
4 questions from Part A (0.5pt) + Part B (1.5pt) + Part C (0.5pt) + Part D (0.5pt) = 3 points

Write one (1) exam question (either multiple choice *or* free response) that tests a key concept from this week (listed below).

Writing your own exam questions helps you think about what's most important to understand about a concept and what kinds of answers best demonstrate mastery. Strong student-written questions may be selected for the actual exam (with credit given). All submitted questions will be compiled each week into a study bank for the class.

Guidelines (please read carefully!):

- Please choose between writing a multiple choice (MC) *or* free-response (FR) question:
  - **MC:** include 4 answer choices.
  - **FR:** write a multi-step problem that would take about 10 minutes to complete.
- Clearly **name the concept** your question is designed to test.
- Provide the **correct answer** and a **short explanation**.
- You may (and are encouraged to!) use AI tools (e.g., ChatGPT, Claude, Gemini, etc.) to help brainstorm and generate potential questions. If you do use AI, **please include:**
  - Which tool you used
  - The exact prompt(s) you entered
  - 💡**Tip:** Try experimenting with different AI tools or refining your prompts for better results. It may be helpful to ask the AI to generate many questions so that you may choose the best one.

  - ⚠️WARNING: If you use AI, make sure the questions and answers are **correct, make sense, and actually test material we've covered in class.** You will almost always need to refine the questions that AI generates. The point of this activity is not to copy-paste the output of a chatbot, but to think carefully about what makes a strong exam question and what it should test!

Example AI prompt that you might refine:
"Write a multi-step free-response question (that takes about 10 minutes to complete) testing the concept of ________, framed in a cognitive science context. Generate 5 variations so I can choose the most suitable one, and include the correct answer and a short explanation."

Key concepts from this week:

- Lecture 4
  - Simple linear regression
  - Estimating coefficients and assessing accuracy
  - Residual standard error
- Lecture 5
  - Multiple linear regression
  - Interpreting coefficients and measures of statistical significance
  - Identifying important predictors and variables
- Lecture 6
  - Qualitative predictors (and w/ multiple levels)
  - Extensions of the linear model
  - Potential problems in linear regression (non-linearity, correlation, non-constant variance outliers. high leverage points, colinearity)

Your exam question

Concept tested:
Question:
Answer & explanation:
If you used AI Which tool?: The exact prompt(s) you entered:

Part D: Metacognitive Reflection (0.5 pt)

Estimated time to complete: 15-30mins

Welcome to your first weekly reflection! This portion of your homework is designed to help you reflect on your own learning in order to learn more effectively. Research shows that regularly reflecting on your own learning (a process called *metacognition*) can significantly improve understanding of course material.

Each week you'll complete a reflection using this same Google Form.

For full credit on this portion, please answer at least 5 of the reflection questions. You're also welcome to answer more—the more you reflect, the better you can refine your learning strategies!