Problem Set 3: Applied Regression (II)

Applied Quantitative Methods for the Social Sciences II

Carlos III–Juan March Institute, Spring 2026

Instructions:

- **Deadline**: **February 26, before class**
- Submit your work as a `.R` file called `ps3.R` in your GitHub repository
- Use comments in your R code to answer conceptual questions and explain your analysis
- For visualizations, use the `marginaleffects` package
- You are encouraged to work together, but each person must submit their own code

1 Conceptual Questions

Answer these questions using comments in your R script.

1.1 Question 1: Interaction Effects

Consider the model: $Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 (X \times Z) + \varepsilon$

a) What is the marginal effect of X on Y? Show how it depends on Z.

b) Suppose $\beta_1 = 0.5$, $\beta_3 = -0.1$, and Z ranges from 0 to 10. At what value of Z does the effect of X become zero?

c) Why is it incorrect to interpret β_1 as "the effect of X" in this model?

d) A researcher estimates this model and finds that β_1 is not statistically significant. They conclude that "X has no effect." Explain why this conclusion is problematic.

1.2 Question 2: Non-linear Relationships

a) Consider the model $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$. What is the marginal effect of X on Y? At what value of X is the marginal effect zero?

b) Suppose $\beta_1 = 2$ and $\beta_2 = -0.1$. Sketch the relationship between X and Y. Is it U-shaped or inverted U-shaped?

c) In the log-log model $\log(Y) = \beta_0 + \beta_1 \log(X) + \varepsilon$, what is the interpretation of β_1?

1.3 Question 3: Standard Errors and Inference

a) Explain what heteroskedasticity is and why it affects standard errors but not point estimates.

b) A researcher reports robust standard errors in their analysis. What does this mean? When should robust standard errors be used?

c) Explain the difference between statistical significance and practical significance. Give an example where a result is statistically significant but not practically significant.

2 Applied Analysis: Gapminder Data

For this problem set, you will use the Gapminder dataset, which contains country-level data on life expectancy, GDP per capita, and population over time.

```
install.packages("gapminder")
library(gapminder)
data(gapminder)
```

You will also need the `marginaleffects` package for visualization:

```
install.packages("marginaleffects")
library(marginaleffects)
```

2.1 Question 4: Non-linear Relationships

Focus on the year 2007 for this analysis.

a) Create a scatter plot of GDP per capita versus life expectancy. Based on the plot, comment on whether you think a linear model is appropriate.

b) Estimate three models:

- Model 1: $\text{lifeExp} = \beta_0 + \beta_1 \cdot \text{gdpPercap} + \varepsilon$
- Model 2: $\text{lifeExp} = \beta_0 + \beta_1 \cdot \log(\text{gdpPercap}) + \varepsilon$
- Model 3: $\text{lifeExp} = \beta_0 + \beta_1 \cdot \text{gdpPercap} + \beta_2 \cdot \text{gdpPercap}^2 + \varepsilon$

Print summaries of the results.

c) Compare the R^2 values of the three models. Which model fits the data best?

d) Using Model 2, interpret the coefficient on $\log(\text{gdpPercap})$ in a comment. What happens to life expectancy when GDP per capita doubles?

2.2 Question 5: Interaction Effects

Now we will examine whether the relationship between GDP and life expectancy varies by continent.

a) Estimate a model with $\log(\text{gdpPercap})$, continent, and their interaction:

$$\text{lifeExp} = \beta_0 + \beta_1 \cdot \log(\text{gdpPercap}) + \beta_2 \cdot \text{continent} + \beta_3 \cdot \log(\text{gdpPercap}) \times \text{continent} + \varepsilon$$

Print the results.

b) Using the `marginaleffects` package, calculate the marginal effect of $\log(\text{gdpPercap})$ for each continent. Print the estimates and 95% confidence intervals.

c) Create a plot showing the predicted life expectancy across the range of GDP per capita for each continent. Include confidence bands. Save the plot.

d) Interpret your findings substantively in a comment. Does the relationship between wealth and health differ by continent? What might explain these differences?

2.3 Question 6: Presenting Results

a) Using your interaction model from Question 5, calculate the predicted life expectancy for:

- A poor African country (GDP per capita = $1,000)
- A middle-income Asian country (GDP per capita = $10,000)
- A wealthy European country (GDP per capita = $40,000)

Print both point estimates and 95% confidence intervals.

b) Calculate the "first difference": How much higher is life expectancy in a wealthy European country compared to a poor African country? Print the 95% confidence interval for this difference.

c) Create a visualization that effectively communicates the key findings from your analysis. Save it and explain in a comment why you chose this particular visualization.

2.4 Question 7: Diagnostics

a) Create a residuals vs. fitted values plot for your interaction model. Comment on whether there are any patterns that suggest model misspecification.

b) Test for heteroskedasticity using the Breusch-Pagan test (use the `lmtest` package). Print the test statistic and p-value.

c) Re-estimate your model using robust standard errors (use the `sandwich` or `estimatr` package). How do the confidence intervals change?

3 Synthesis Question

3.1 Question 8

In approximately 300 words (as comments in your R script), discuss the importance of moving beyond simple regression tables when communicating research findings. Drawing on your analysis of the Gapminder data, explain:

- Why predicted values and marginal effects are more informative than regression coefficients alone
- The value of visualizations in understanding complex relationships (like interactions)
- How to effectively communicate uncertainty in your estimates

4 Submission

Commit your `ps3.R` file to your GitHub repository before the deadline. Make sure your repository is public so I can access it.

Your R script should:

- Be well-organized with clear section headers (using comments)
- Include all code needed to reproduce your analysis
- Include your answers to conceptual questions as comments
- Save any plots to files (e.g., using `ggsave()`)
- Run without errors from top to bottom