

Assignment 2: Applied Regression

Instructions:

- **Deadline:** February 19, before class
- Submit your work in a separate folder in your GitHub repository
 - You can include only the R file or additional ones (e.g. pdf with results)
- **Always use comments** in your R code – and use them to answer questions
- You are encouraged to work together, but each person must submit their own code
- Plan is to start Part 1 in class and complete Part 2 at home

Contents

1	Part 1: QoG Dataset	2
2	Part 2: Take-Home Exercises (STAR Dataset)	4
3	Data Sources	5
4	Submission	5

1 Part 1: QoG Dataset

In this lab, we explore cross-country data from the Quality of Government (QoG) dataset. You will practice bivariate and multiple regression, demonstrate omitted variable bias, use robust standard errors, and present results with `modelsummary`. Follow along in class.

1.1 Setup and data preparation

Download the QoG Standard cross-section dataset from gu.se/en/quality-government/qog-data/data-downloads/standard-dataset (CSV format).

a) Load the dataset and select the following variables. Rename them for convenience:

- `cname` — country name
- `epi_epi` — Environmental Performance Index (EPI) score (our outcome)
- `wdi_wip` — share of women in parliament (%)
- `wbgi_gee` — government effectiveness (World Bank governance indicator)
- `gpp_grnst` — green party seat share in parliament (%)

Suggested names: `country`, `epi`, `women_parl`, `gov_eff`, `green_seats`.

- b) Drop observations with missing values on any of these variables. How many countries remain?
- c) Print summary statistics for all variables (e.g., using `summary()` or `skimr::skim()`).

1.2 Exploratory visualization

- a) Create a scatter plot of `women_parl` (x-axis) vs. `epi` (y-axis).
- b) Add a linear fit using `geom_smooth(method = "lm")`.
- c) In a comment, describe what you see. Is there a relationship? What is its direction?

1.3 Bivariate regression

- a) Run a bivariate regression: `lm(epi ~ women_parl, data = df)`.
- b) Extract the results using `broom::tidy()`.
- c) Interpret the coefficient on `women_parl` in a comment. What is the predicted difference in EPI between a country at the 25th percentile and one at the 75th percentile of women in parliament?

1.4 Multiple regression

- a) Run a multiple regression adding `gov_eff` as a control:
`lm(epi ~ women_parl + gov_eff, data = df)`.
- b) Compare the coefficient on `women_parl` between the bivariate and multiple regression. Does it change? In what direction? Explain in a comment what this suggests.

1.5 Demonstrating OVB

The omitted variable bias formula says: $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \cdot \tilde{\delta}$, where $\tilde{\beta}_1$ is the bivariate coefficient, $\hat{\beta}_1$ and $\hat{\beta}_2$ are the multiple regression coefficients, and $\tilde{\delta}$ is the coefficient from regressing the omitted variable on the included variable.

- a) From the regressions above, write down $\tilde{\beta}_1$ (bivariate), $\hat{\beta}_1$ (multiple), and $\hat{\beta}_2$ (multiple).
- b) Run the auxiliary regression: `lm(gov_eff ~ women_parl, data = df)`. Extract $\tilde{\delta}$.
- c) Verify the OVB formula: check that $\hat{\beta}_1 + \hat{\beta}_2 \times \tilde{\delta} \approx \tilde{\beta}_1$ (up to rounding).
- d) In a comment, interpret what this means: why did the coefficient on `women_parl` change when we added `gov_eff`?

1.6 Robust standard errors

- a) Using `modelsummary()`, print the multiple regression results with default (classical) standard errors.
- b) Now print the same model with robust standard errors: `modelsummary(model, vcov = "robust")`.
- c) Compare the SEs. Do they differ substantially? Do any conclusions change?

1.7 Presenting results

- a) Create a table comparing the bivariate and multiple regression models side by side, using robust SEs for both:
`modelsummary(list(m1, m2), vcov = "robust")`.
- b) Create a coefficient plot using `modelsummary::modelplot()` comparing both models.
- c) Save the plot using `ggsave()`.

2 Part 2: Take-Home Exercises (STAR Dataset)

The Project STAR (Student/Teacher Achievement Ratio) experiment randomly assigned students in Tennessee to small classes (13–17 students), regular classes (22–25 students), or regular classes with a teacher’s aide. We use data from this experiment to practice applied regression.

Download `star.csv` from the course page. Each observation is a student. Key variables include:

- `classtype`: class type (1 = small, 2 = regular, 3 = regular + aide)
- `race`: student race (1 = White, 2 = Black, 3 = Asian, 4 = Hispanic, 5 = Native American, 6 = Other)
- `yearssmall`: years spent in a small class (0–4)
- `hsgrad`: graduated high school (0/1)
- `g4math`: 4th grade math test score
- `g4reading`: 4th grade reading test score

2.1 Data preparation

- Load `star.csv`.
- Create a factor variable for `classtype` with labels: "Small", "Regular", "Regular+Aide".
- Create a factor variable for `race` with labels: "White", "Black", "Asian", "Hispanic", "Native American", "Other".
- Create a binary variable `small` that equals 1 if `classtype == "Small"` and 0 otherwise.
- Report the number of observations and the number of non-missing observations for `g4reading` and `g4math`.

2.2 Comparing groups

- Calculate the mean 4th grade reading score by class type. Which group scores highest?
- Run a bivariate regression of `g4reading` on `small`. Interpret the coefficient.
- Verify that the regression coefficient equals the difference in means between small and regular+aide classes. (Hint: compare with the grouped means from part a.)
- Repeat the bivariate regression for `g4math`. Is the pattern similar?

2.3 Adding controls

- Run a multiple regression of `g4reading` on `small`, `race`, and `yearssmall`.
- Compare the coefficient on `small` with the bivariate model. Does it change much? What does this tell you about the quality of the randomization?
- Interpret the coefficient on `yearssmall`. What does it capture?

2.4 Interactions

- a) Does the effect of being in a small class differ by race? Fit the following model:
`lm(g4reading ~ small * race + yearssmall, data = df).`
- b) Print the results using `broom::tidy()`.
- c) What is the estimated effect of a small class for White students? For Black students?
(Use the coefficients to calculate.)
- d) In a comment, discuss whether the interaction is substantively meaningful.

2.5 Presenting results

- a) Create a table with `modelsummary()` comparing all your reading score models (bivariate, multiple, interaction), using robust standard errors.
- b) Create a coefficient plot with `modelplot()` for the three models.
- c) Save both outputs.

2.6 Brief discussion

In a comment (5–10 sentences), discuss:

- a) What does the STAR data suggest about the effect of small class sizes on student achievement?
- b) Why is this evidence more credible than a typical observational study of class size?
- c) Are there any limitations or caveats based on what you observed in the data?

3 Data Sources

- **QoG Standard dataset:** <https://www.gu.se/en/quality-government/qog-data/data-download-standard-dataset>. Use the cross-section version (CSV format), or the file provided on the course page.
- **STAR dataset:** `star.csv`, provided on the course page. Original data from the Tennessee STAR experiment (Krueger, 1999).

4 Submission

Commit your `ps2.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 and interpretations as comments

- Save any plots to files (e.g., using `ggsave()`)
- Run without errors from top to bottom