Applied Quantitative Methods for the Social Sciences II
MA in Social Sciences, UC3M, Spring 2026

Assignment 4: Model Interpretation and Diagnostics

Instructions:

- Deadline: **March 5, 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
- I'll upload a solution file to the website after next class

Contents

1 Part 1: In-Class (Corruption and Wealth)

In this lab, we analyze the relationship between corruption and economic development using cross-country data. You will practice computing predicted values, marginal effects, and creating publication-quality tables and plots. Recall from the lecture that raw coefficients are rarely enough — we need to compute **quantities of interest** to communicate results effectively.

1.1 Setup and data exploration

Download the data here:

- github.com/franvillamil/AQM2/tree/master/datasets/other

a) Load the `corruption.dta` dataset using `readstata13::read.dta13()`. Key variables:
   - `cname` — country name
   - `ti_cpi` — Corruption Perceptions Index (0–10 scale, higher = less corrupt)
   - `undp_gdp` — GDP per capita (PPP, in dollars)

b) Drop observations with missing values on `ti_cpi` or `undp_gdp`. How many countries remain?

c) Compute summary statistics for `ti_cpi` and `undp_gdp`. In a comment, note the range and standard deviation of each variable. Is GDP per capita right-skewed?

1.2 Exploratory visualization

a) Create a scatter plot of `ti_cpi` (y-axis) against `undp_gdp` (x-axis) using `geom_point()`. Add a smooth line with `geom_smooth(method = "lm")`.

b) In a comment, describe the pattern. Does the relationship look linear?

c) Now create a second scatter plot with `log(undp_gdp)` on the x-axis. Does the log transformation improve the linearity of the relationship?

1.3 Bivariate regression

a) Estimate a bivariate regression of corruption on GDP per capita:
   `m1 = lm(ti_cpi ~ undp_gdp, data = df)`.

b) Print the results using `summary()` or `broom::tidy()`. In a comment, interpret the coefficient on `undp_gdp`. What is the predicted change in the corruption index for a $10,000 increase in GDP per capita?

c) Compute the 25th and 75th percentiles of GDP per capita using `quantile()`. Then use `predictions()` to get predicted corruption scores at these values:

```
library(marginaleffects)
predictions(m1, newdata = datagrid(undp_gdp = c(q25, q75)))
```

In a comment, report the predicted values and their 95% confidence intervals. What is the difference in predicted corruption between a country at the 25th percentile vs. the 75th percentile of GDP?

1.4 Non-linear specifications

a) Estimate a model using the log of GDP per capita:
`m2 = lm(ti_cpi ~ log(undp_gdp), data = df)`.

b) Interpret the coefficient on `log(undp_gdp)`. Recall from the lecture that this is a level-log model: a 1% increase in GDP is associated with a change of $\beta_1/100$ in the corruption index. In a comment, compute the predicted change in corruption for a **doubling** of GDP per capita (hint: $\log(2) \approx 0.693$, so the change is approximately $\beta_1 \times 0.693$).

c) Estimate a model with a quadratic GDP term:
`m3 = lm(ti_cpi ~ undp_gdp + I(undp_gdp^2), data = df)`.

d) Compare the R^2 of all three models. Which specification fits the data best? In a comment, explain why a non-linear specification might be appropriate for this relationship.

1.5 Marginal effects

a) For the log model (`m2`), compute the average marginal effect of GDP using:
`avg_slopes(m2, variables = "undp_gdp")`.

b) In a comment, explain why the AME differs from the raw coefficient on `log(undp_gdp)`. What does the AME tell you in substantive terms?

c) For the quadratic model (`m3`), compute marginal effects at specific GDP values using:

```
slopes(m3, variables = "undp_gdp",
       newdata = datagrid(undp_gdp = c(2000, 10000, 30000)))
```

In a comment, describe how the marginal effect of GDP changes as countries become richer. Does the effect diminish?

1.6 Prediction plots

a) Create a prediction plot for the log model:
`plot_predictions(m2, condition = "undp_gdp")`. Save the plot.

b) Create a prediction plot for the quadratic model (`m3`) on the same variable. Save this plot too.

c) In a comment, compare the two plots. Do the models tell a similar story about the corruption–wealth relationship? Where do they diverge?

1.7 Residual diagnostics

a) Use `broom::augment(m1)` to get residuals and fitted values from the level-level model. Create a scatter plot of residuals (`.resid`) vs. fitted values (`.fitted`). Does the plot suggest non-linearity or heteroskedasticity?

b) Now do the same for the log model (`m2`). Does the log transformation improve the residual pattern?

c) Identify influential observations using Cook's distance. Use `plot(m2, which = 4)` or compute Cook's distance manually with `cooks.distance(m2)`. Which countries (if any) have Cook's distance above $4/n$? Look up their names.

d) In a comment, discuss: should these influential observations be removed? What would you recommend as a robustness check?

1.8 Publication-quality table

a) Create a regression table comparing all three models using:

```
library(modelsummary)
modelsummary(
  list("Level-Level" = m1, "Level-Log" = m2, "Quadratic" = m3),
  vcov = "robust",
  stars = TRUE,
  gof_map = c("r.squared", "nobs"))
```

b) In a comment, summarize: which model would you choose for a final presentation, and why?

2 Part 2: Take-Home (Wealth and Infant Mortality)

We now turn to another cross-country question: the relationship between national income and infant mortality. This exercise asks you to build and compare multiple specifications, compute predicted values for specific scenarios, and create a publication-quality visualization.

Download `infantmortality.dta` from:

- github.com/franvillamil/AQM2/tree/master/datasets/other

Key variables:

- `country` — country name
- `region` — world region (Africa, Americas, Asia, Europe)
- `income` — per-capita income (dollars)
- `infant` — infant mortality rate (per 1,000 live births)
- `oil` — oil-exporting country (yes/no)

2.1 Data exploration

a) Load the dataset and print summary statistics for all variables. How many countries are in the data?
b) Create a histogram of `infant` and a histogram of `income`. Are either of them right-skewed?
c) Create a scatter plot of `infant` (y-axis) against `income` (x-axis), coloring points by `region`. Describe the relationship in a comment.
d) Create the same scatter plot but using `log(income)` on the x-axis and `log(infant)` on the y-axis. Does the log-log relationship look more linear?

2.2 Comparing specifications

a) Estimate a level-level model:
`m1 = lm(infant ~ income, data = df).`
b) Estimate a log-log model:
`m2 = lm(log(infant) ~ log(income), data = df).`
c) Interpret the coefficient on income in each model:
   - In `m1`: what is the predicted change in infant mortality for a $1,000 increase in income?
   - In `m2`: recall that the log-log coefficient is an **elasticity**. What does it mean here? (e.g., "A 10% increase in income is associated with a _% change in infant mortality.")
d) Create a residuals vs. fitted values plot for both models. Which specification has a better residual pattern? Discuss in a comment.

2.3 Multiple regression with controls

a) Estimate a log-log model with controls for region and oil-exporting status: `m3 = lm(log(infant) ~ log(income) + region + oil, data = df)`.
b) Print the results. In a comment, interpret the coefficient on `log(income)`: does controlling for region and oil status change the income effect?
c) Interpret the coefficient on the Africa region indicator (relative to the reference category). What does it tell you about infant mortality in Africa, controlling for income?
d) Compute average marginal effects using `avg_slopes(m3)`. Focus on the AME of `income` and report it in a comment.

2.4 Interaction: oil status and income

a) Estimate a model with an interaction between oil status and log income: `m4 = lm(log(infant) ~ log(income) * oil + region, data = df)`.
b) Use `avg_slopes(m4, variables = "income", by = "oil")` to compute the marginal effect of income separately for oil-exporting and non-oil countries.
c) In a comment, discuss: does the relationship between income and infant mortality differ for oil-exporting countries? What might explain this?
d) Plot how the marginal effect of income varies by oil status: `plot_slopes(m4, variables = "income", condition = "oil")`. Save the plot.

2.5 Predicted values for specific scenarios

a) Using model `m3` (without interaction), compute predicted infant mortality rates for:
   - A non-oil African country with income = $1,000
   - A non-oil European country with income = $20,000
   - An oil-exporting country in the Americas with income = $10,000

   Use:

```
predictions(m3,
  newdata = datagrid(
    income = c(1000, 20000, 10000),
    region = c("Africa", "Europe", "Americas"),
    oil = c("no", "no", "yes")))
```

   Note: since the outcome is `log(infant)`, you need to exponentiate the predictions to get infant mortality in the original scale. Use `exp()` on the `estimate` column.
b) In a comment, discuss the predicted values. Are they plausible? How large is the gap between the African and European scenarios?

2.6 Publication-quality visualization

a) Create a prediction plot showing predicted infant mortality across income levels, separately by region:

```
plot_predictions(m3, condition = c("income", "region"))
```

Customize the plot to make it suitable for a general audience: add informative axis labels, a title, and use `theme_minimal()` or similar. Save the plot.

b) In a comment (5–10 sentences), discuss: what does this plot tell a general audience about the relationship between wealth and infant mortality? What role does geography play? What are the main limitations of this analysis (e.g., omitted variables, reverse causality, ecological fallacy)?

2.7 Diagnostics and robust inference

a) Create a residuals vs. fitted values plot for `m3`. Does the plot suggest heteroskedasticity?

b) Create a regression table comparing all four models with robust standard errors:

```
modelsummary(
  list("Level" = m1, "Log-Log" = m2,
       "Controls" = m3, "Interaction" = m4),
  vcov = "robust",
  stars = TRUE,
  gof_map = c("r.squared", "nobs"))
```

c) Compare the robust and default standard errors for `m3`. Run `modelsummary()` with and without `vcov = "robust"`. Do the conclusions change? Why use robust SEs?

3 Data Sources

Both datasets are available at the course GitHub repository:

- Corruption data: github.com/franvillamil/AQM2/tree/master/datasets/other (`corruption.dta`)
- Infant mortality data: same folder (`infantmortality.dta`)

4 Submission

Commit your file to your GitHub repository before the deadline. Put it in a different folder, e.g. `assignment4`. 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