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

Assignment 2 – Solutions: Part 1 (QoG Dataset)

Contents

1 Setup and data preparation

a) Load and rename variables:

```r
library(dplyr)
library(broom)
library(ggplot2)
library(modelsummary)

qog = read.csv("https://www.qogdata.pol.gu.se/data/qog_std_cs_jan26.csv")

df = qog %>%
  select(country = cname, epi = epi_epi, women_parl = wdi_wip,
         gov_eff = wbgi_gee, green_seats = cpds_lg)
```

b) Drop missing values:

```r
df = df %>% na.omit()
nrow(df)
```

Around 30–35 countries remain (exact number depends on dataset version). The small sample is due to green_seats being available only for a subset of (mostly European) countries.

c) Summary statistics:

```r
summary(df)
```

2 Exploratory visualization

a–b) Scatter plot with linear fit:

```r
ggplot(df, aes(x = women_parl, y = epi)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Women in Parliament (%)", y = "EPI Score")
```

c) There is a positive relationship: countries with more women in parliament tend to have higher environmental performance scores. This likely reflects that both variables are associated with broader development and governance quality.

3 Bivariate regression

a) Run the bivariate model:

```r
m1 = lm(epi ~ women_parl, data = df)
```

b) Extract results:

```r
tidy(m1)
```

c) Interpretation: the coefficient on `women_parl` indicates the predicted change in EPI score for each additional percentage point of women in parliament. For the IQR difference (e.g., from about 20% to 40%), multiply the coefficient by 20 to get the predicted difference in EPI. For example, if $\hat{\beta} \approx 0.5$, the predicted difference would be around 10 EPI points.

4 Multiple regression

a) Add government effectiveness:

```
m2 = lm(epi ~ women_parl + gov_eff, data = df)
tidy(m2)
```

b) The coefficient on `women_parl` decreases substantially once `gov_eff` is included. This suggests that part of the bivariate association was driven by government effectiveness being correlated with both women in parliament and environmental performance. In other words, the bivariate model suffered from omitted variable bias.

5 Demonstrating OVB

Recall the OVB formula: $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \cdot \tilde{\delta}$

a) Extract the relevant coefficients:

```
beta1_biva = tidy(m1) %>% filter(term == "women_parl") %>%
  pull(estimate)
beta1_mult = tidy(m2) %>% filter(term == "women_parl") %>%
  pull(estimate)
beta2_mult = tidy(m2) %>% filter(term == "gov_eff") %>%
  pull(estimate)
```

b) Auxiliary regression:

```
aux = lm(gov_eff ~ women_parl, data = df)
delta = tidy(aux) %>% filter(term == "women_parl") %>%
  pull(estimate)
```

c) Verify:

```
round(beta1_mult + beta2_mult * delta, 2)
round(beta1_biva, 2)
```

Both values should match (up to rounding), confirming the OVB formula.

d) Interpretation: the bivariate coefficient on women in parliament was inflated because `gov_eff` is positively correlated with both `women_parl` ($\tilde{\delta} > 0$) and with `epi` ($\hat{\beta}_2 > 0$). The omitted variable bias was therefore positive, making the bivariate estimate larger than the multiple regression estimate.

6 Robust standard errors

a) Classical SEs:

```
modelsummary(m2)
```

b) Robust SEs:

```
modelsummary(m2, vcov = "robust")
```

c) With a small cross-sectional sample like this, robust SEs may differ somewhat from classical SEs but typically not enough to change substantive conclusions. The direction of differences depends on the heteroskedasticity pattern.

7 Presenting results

a) Side-by-side table:

```
modelsummary(list("Bivariate" = m1, "Multiple" = m2),
             vcov = "robust")
```

b) Coefficient plot:

```
modelplot(list("Bivariate" = m1, "Multiple" = m2),
          vcov = "robust")
```

c) Save:

```
p = modelplot(list("Bivariate" = m1, "Multiple" = m2),
              vcov = "robust")
ggsave("coefplot_qog.png", p, width = 7, height = 4)
```

8 Extra: effect size

Several strategies:

- Compare the predicted change for a meaningful shift in the predictor (e.g., one standard deviation or the IQR) relative to the range or standard deviation of EPI.
- Standardize both variables and re-run the regression to get a standardized coefficient ("beta coefficient").
- Use substantive knowledge: is a 5-point EPI difference meaningful in practice?