Assignment 2 – Solutions: Part 1 (QoG Dataset)

Applied Quantitative Methods II, UC3M

1. Setup and data preparation

a) Load and rename variables:

```
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) Using `na.omit()` drops too many observations because `green_seats` is only available for a small subset of (mostly European) countries. Since the exercises focus on the other variables, it is better to keep the full sample and let R handle missing values automatically in each regression.

```
# na.omit() would leave very few countries -- not recommended here
nrow(df)
```

```
## [1] 194
```

c) Summary statistics:

```
summary(df)
```

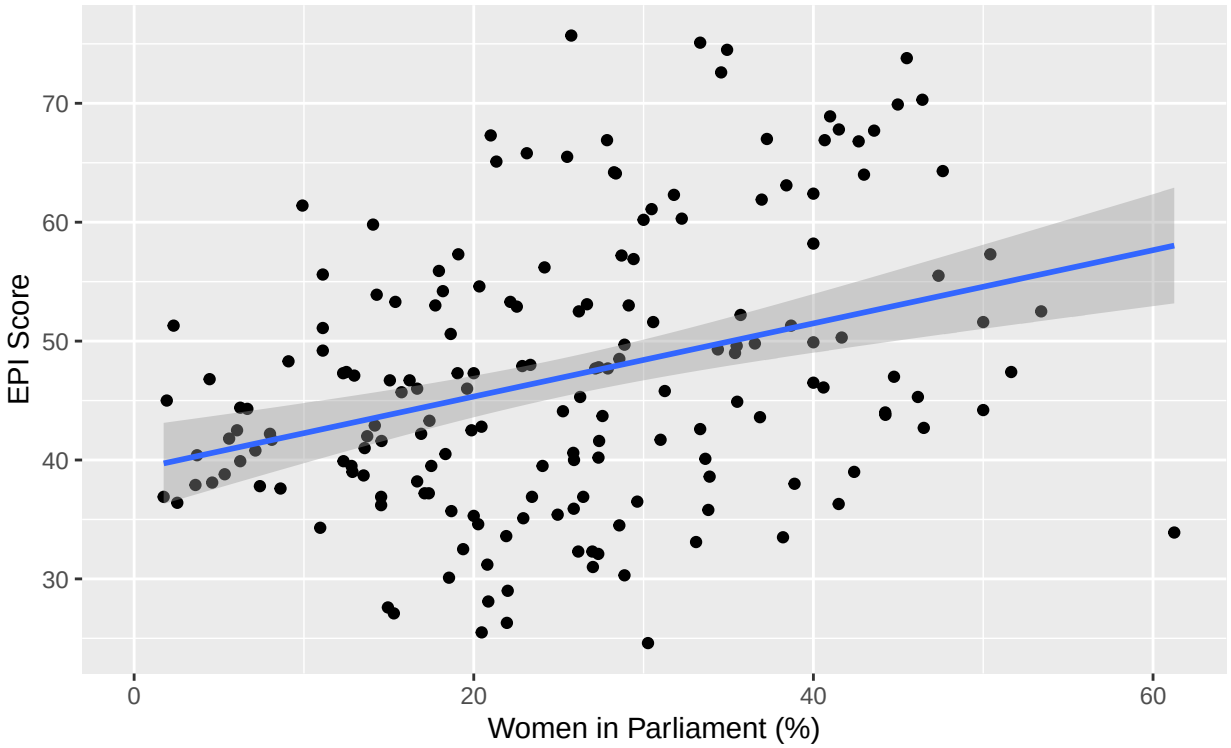
```
##    country               epi          women_parl       gov_eff
##  Length:194         Min.   :24.60   Min.   : 0.00   Min.   :-2.21688
##  Class :character   1st Qu.:38.65   1st Qu.:15.36   1st Qu.:-0.68971
##  Mode  :character   Median :45.70   Median :25.10   Median :-0.07002
##                     Mean   :46.99   Mean   :24.99   Mean   : 0.02314
##                     3rd Qu.:53.20   3rd Qu.:33.68   3rd Qu.: 0.70781
##                     Max.   :75.70   Max.   :61.25   Max.   : 2.20130
##                     NA's   :15      NA's   :2       NA's   :2
##   green_seats
##  Min.   : 0.000
##  1st Qu.: 0.000
##  Median : 0.300
##  Mean   : 5.500
##  3rd Qu.: 8.725
```

```
##  Max.   :45.500
##  NA's   :158
```

2. Exploratory visualization

a–b) Scatter plot with linear fit:

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



c) Positive relationship: countries with more women in parliament tend to have higher EPI scores. This likely reflects that both variables are associated with broader development and governance quality.

3. Bivariate regression

a) Run the bivariate model:

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

b) Extract results:

```
tidy(m1)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   39.2      1.83       21.4  5.91e-51
```

```
## 2 women_parl     0.308    0.0646      4.76 3.99e- 6
```

c) The coefficient on `women_parl` indicates the predicted change in EPI score for each additional percentage point of women in parliament. To get the predicted difference between the 25th and 75th percentile, we can either multiply the coefficient by the IQR or use `predict()`. Both give the same result:

```
p25 = quantile(df$women_parl, 0.25, na.rm = TRUE)
p75 = quantile(df$women_parl, 0.75, na.rm = TRUE)

# Option 1: multiply coefficient by IQR
coef(m1)["women_parl"] * (p75 - p25)
```

```
## women_parl
##   5.638584
```

```
# Option 2: use predict()
pred = predict(m1, newdata = data.frame(women_parl = c(p25, p75)))
pred[2] - pred[1]
```

```
##      75%
## 5.638584
```

4. Multiple regression

a) Add government effectiveness:

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

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  44.2       1.34       33.0  3.94e-77
## 2 women_parl    0.0979    0.0480      2.04 4.32e- 2
## 3 gov_eff       8.71      0.647      13.5  8.53e-29
```

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 (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:

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

c) Verify the OVB formula:

```r
# Right-hand side: beta1_mult + beta2_mult * delta
round(beta1_mult + beta2_mult * delta, 4)
```

```
## [1] 0.3307
```

```r
# Left-hand side: beta1_biva
round(beta1_biva, 4)
```

```
## [1] 0.3078
```

Both values match, confirming the OVB formula.

d) The bias is positive because `gov_eff` is positively correlated with both `women_parl` ($\tilde{\delta} > 0$) and with `epi` ($\hat{\beta}_2 > 0$). This inflated the bivariate estimate.

6. Robust standard errors

a) Classical SEs:

```r
modelsummary(m2, output = "markdown")
```

	Model 1
(Intercept)	44.214
	(1.338)
women_parl	0.098
	(0.048)
gov_eff	8.710
	(0.647)
Num.Obs.	178
R2	0.565
R2 Adj.	0.560
AIC	1233.6
BIC	1246.3
Log.Lik.	-612.794
RMSE	7.57

b) Robust SEs:

```r
modelsummary(m2, vcov = "robust", output = "markdown")
```

	Model 1
(Intercept)	44.214
	(1.282)
women_parl	0.098
	(0.047)

	Model 1
gov_eff	8.710
	(0.670)
Num.Obs.	178
R2	0.565
R2 Adj.	0.560
AIC	1233.6
BIC	1246.3
Log.Lik.	-612.794
RMSE	7.57
Std.Errors	Robust

c) SEs may differ somewhat but conclusions typically don't change with this sample.

7. Presenting results

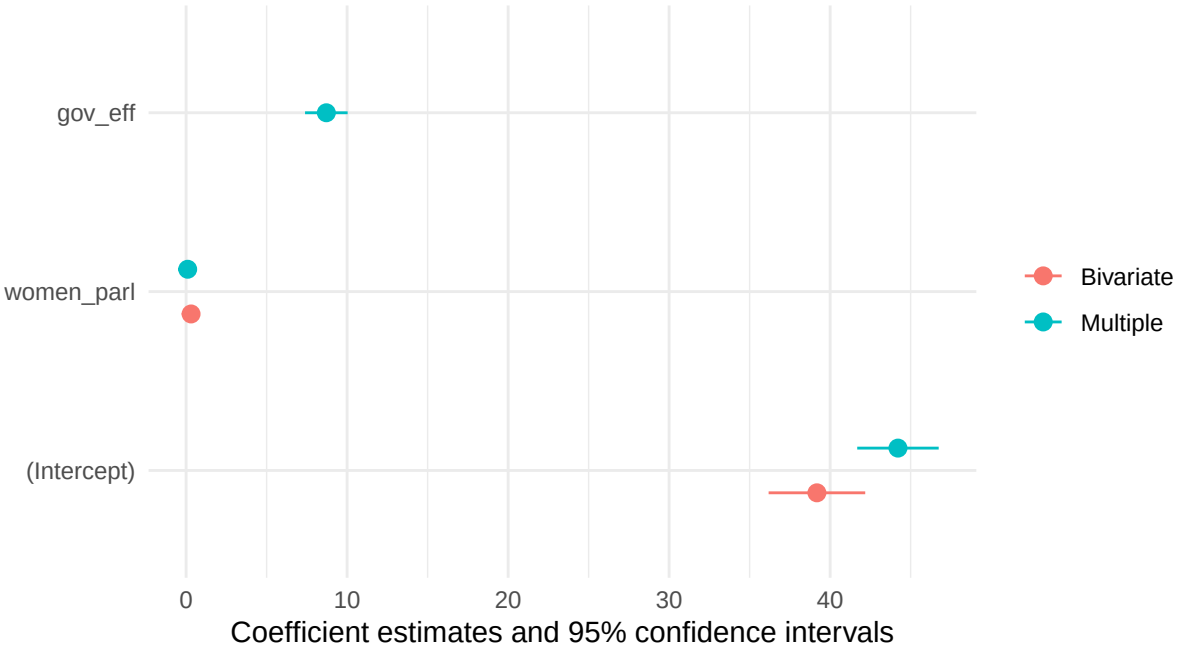
a) Side-by-side table:

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

	Bivariate	Multiple
(Intercept)	39.180	44.214
	(1.520)	(1.282)
women_parl	0.308	0.098
	(0.063)	(0.047)
gov_eff		8.710
		(0.670)
Num.Obs.	178	178
R2	0.114	0.565
R2 Adj.	0.109	0.560
AIC	1358.0	1233.6
BIC	1367.6	1246.3
Log.Lik.	-676.014	-612.794
RMSE	10.79	7.57
Std.Errors	Robust	Robust

b–c) Coefficient plot:

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



8. Extra: effect size

Several strategies to assess effect size:

- Compare the predicted change for a meaningful shift in the predictor (e.g., one SD or the IQR) relative to the SD of EPI.
- Standardize both variables and re-run the regression to get a standardized ("beta") coefficient.
- Use substantive knowledge about whether the predicted change matters in practice.