Assignment 4 – Solutions: Part 2 (Wealth and Infant Mortality)

Applied Quantitative Methods II, UC3M

1. Data exploration

a) Load the dataset and summary statistics:

```
library(dplyr)
library(broom)
library(ggplot2)
library(modelsummary)
library(marginaleffects)
library(readstata13)

df = read.dta13("https://raw.githubusercontent.com/franvillamil/AQM2/refs/heads/master/datasets/other/infa
summary(df)
```

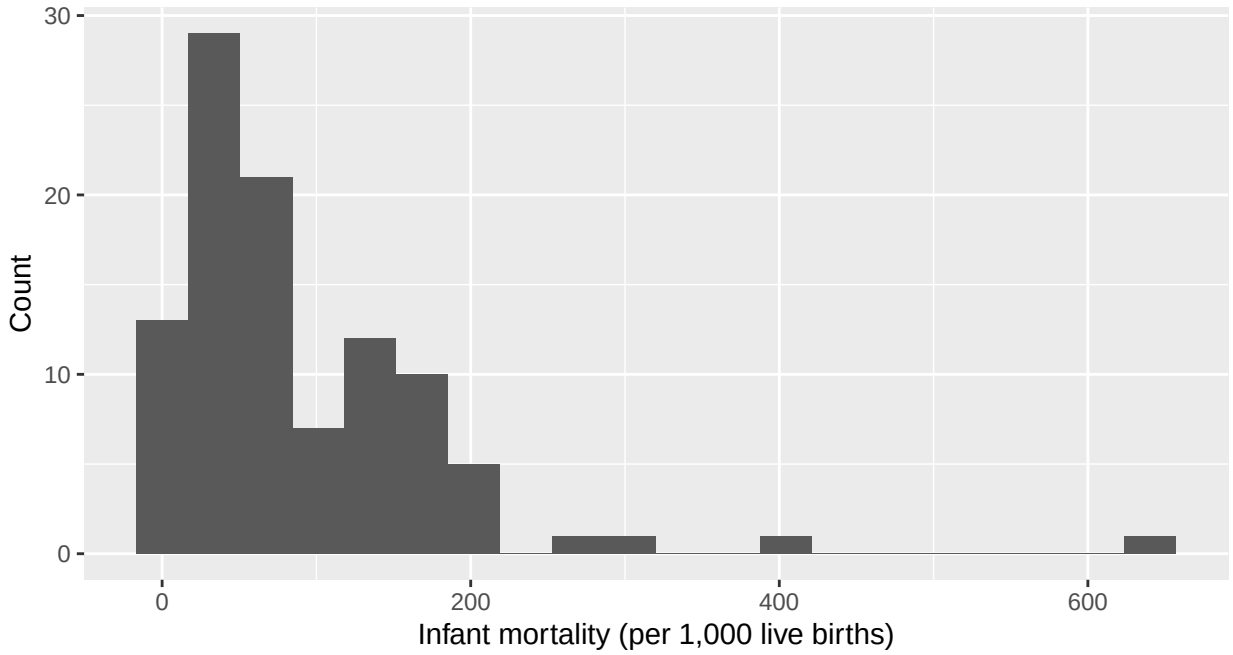
```
##    country             region              income         infant
##  Length:101         Length:101         Min.   :  50   Min.   :  9.60
##  Class :character   Class :character   1st Qu.: 130   1st Qu.: 26.20
##  Mode  :character   Mode  :character   Median : 334   Median : 60.60
##                                        Mean   :1022   Mean   : 89.05
##                                        3rd Qu.:1191   3rd Qu.:129.40
##                                        Max.   :5596   Max.   :650.00
##      oil
##  Length:101
##  Class :character
##  Mode  :character
##
##
##
```

```
nrow(df)
```

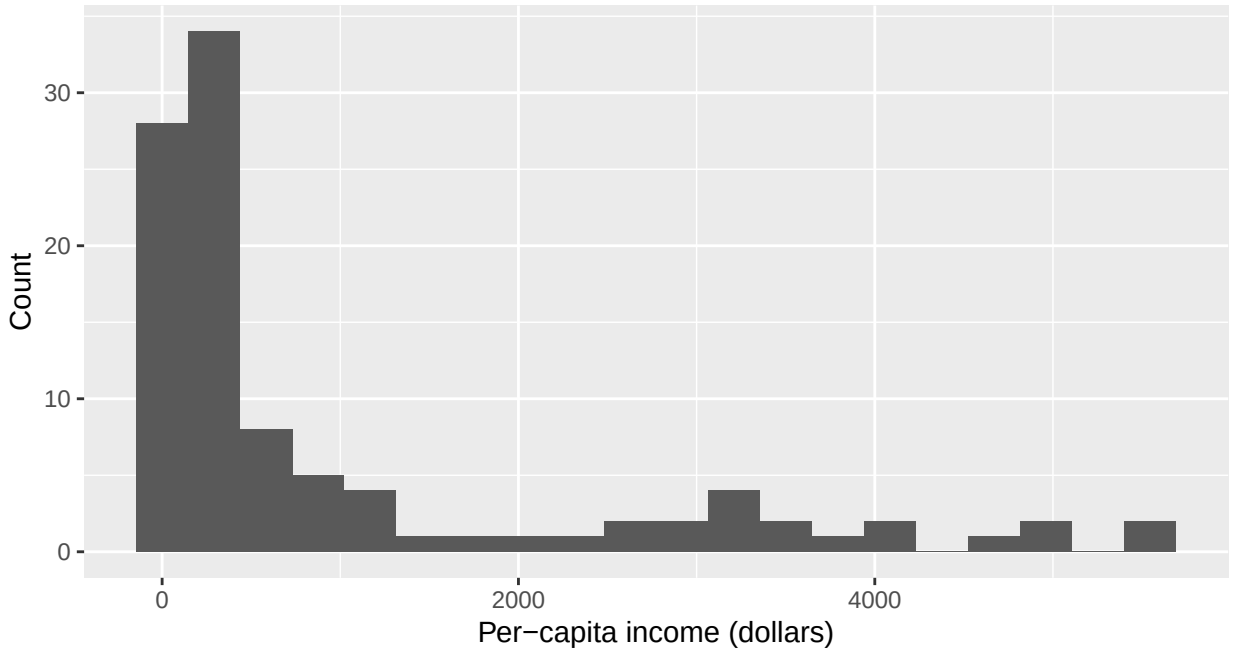
```
## [1] 101
```

b) Histograms of infant mortality and income:

```
ggplot(df, aes(x = infant)) +
  geom_histogram(bins = 20) +
  labs(x = "Infant mortality (per 1,000 live births)", y = "Count")
```



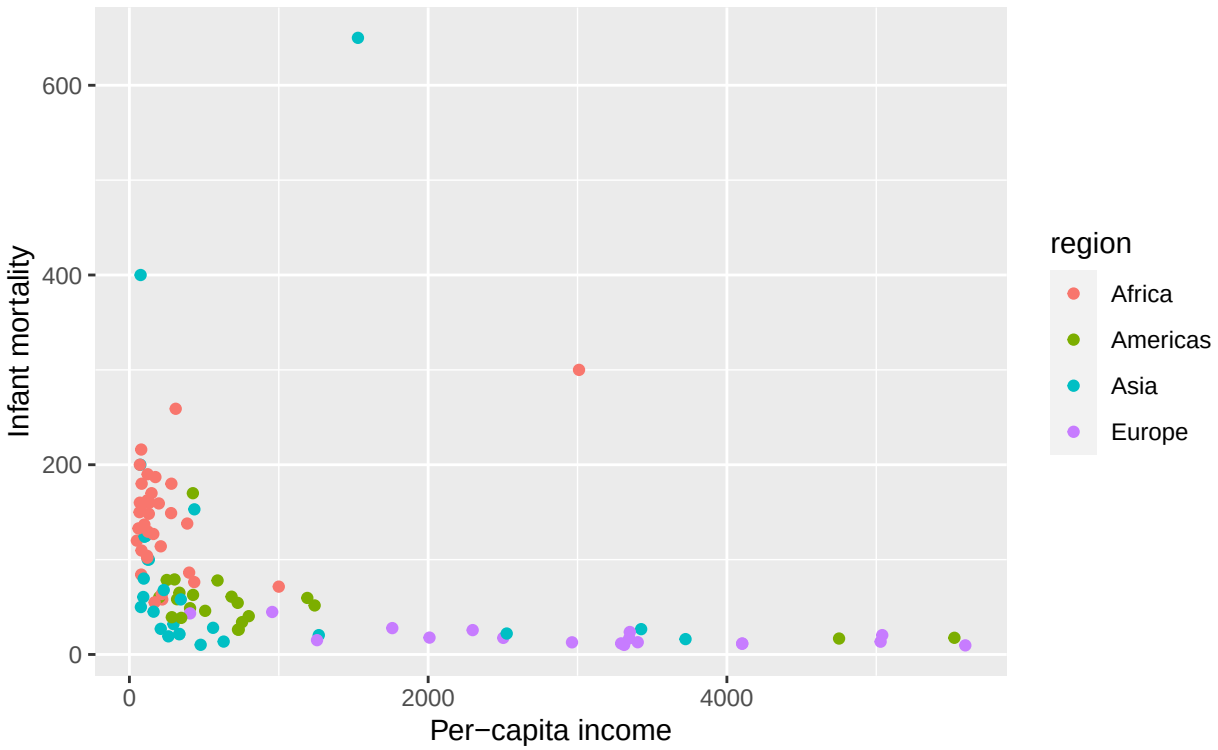

```
ggplot(df, aes(x = income)) +
  geom_histogram(bins = 20) +
  labs(x = "Per-capita income (dollars)", y = "Count")
```



Both variables are right-skewed: most countries have low income and low-to-moderate infant mortality, with long right tails.

c) Scatter plot colored by region:

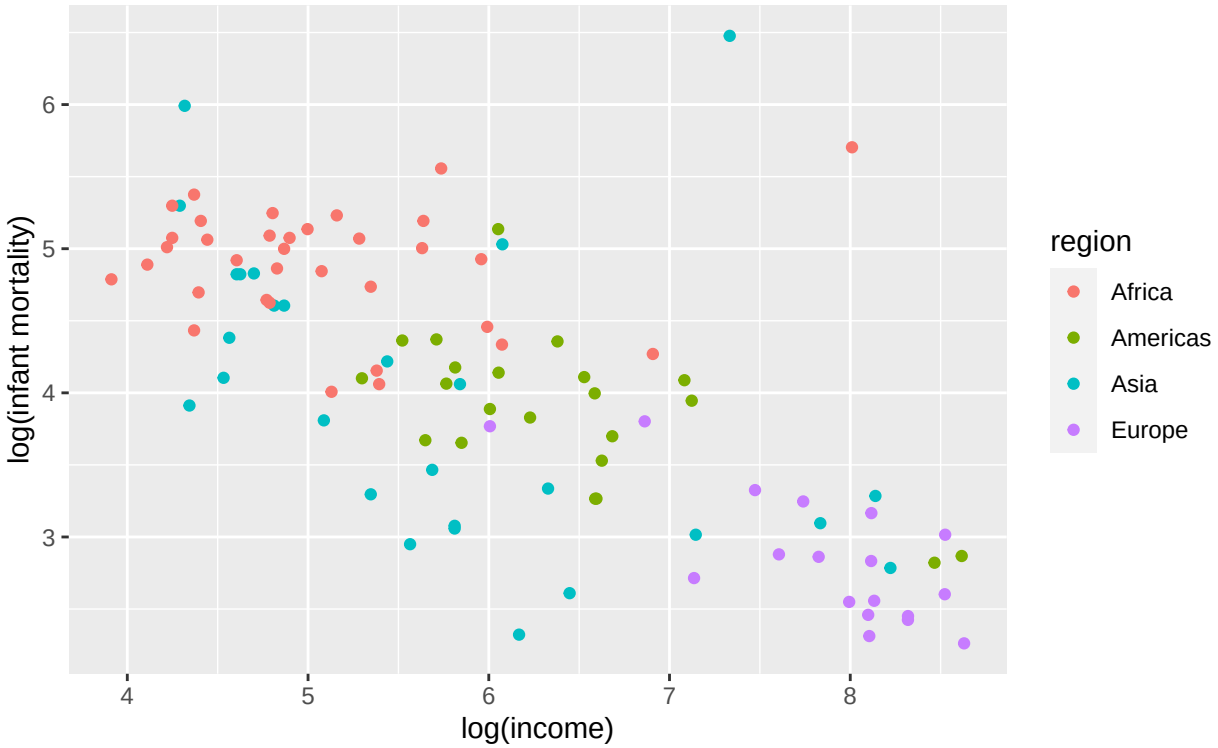
```
ggplot(df, aes(x = income, y = infant, color = region)) +
  geom_point() +
  labs(x = "Per-capita income", y = "Infant mortality")
```



There is a clear negative relationship: richer countries have lower infant mortality. The relationship is non-linear, with a steep decline at low income levels that flattens out. African countries tend to have the highest mortality and lowest income.

d) Log-log scatter plot:

```
ggplot(df, aes(x = log(income), y = log(infant), color = region)) +
  geom_point() +
  labs(x = "log(income)", y = "log(infant mortality)")
```



The log-log relationship looks much more linear, suggesting that a log-log specification is appropriate.

2. Comparing specifications

a) Level-level model:

```
m1 = lm(infant ~ income, data = df)
tidy(m1)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) 110.      10.5         10.5  9.96e-18
## 2 income       -0.0209   0.00600     -3.48 7.35e- 4
```

b) Log-log model:

```
m2 = lm(log(infant) ~ log(income), data = df)
tidy(m2)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    7.15      0.317      22.6  7.93e-41
## 2 log(income)   -0.512     0.0512     -9.99 1.14e-16
```

c) In the level-level model, the coefficient on income gives the predicted change in infant mortality for a one-dollar increase in income. For a $1,000 increase:

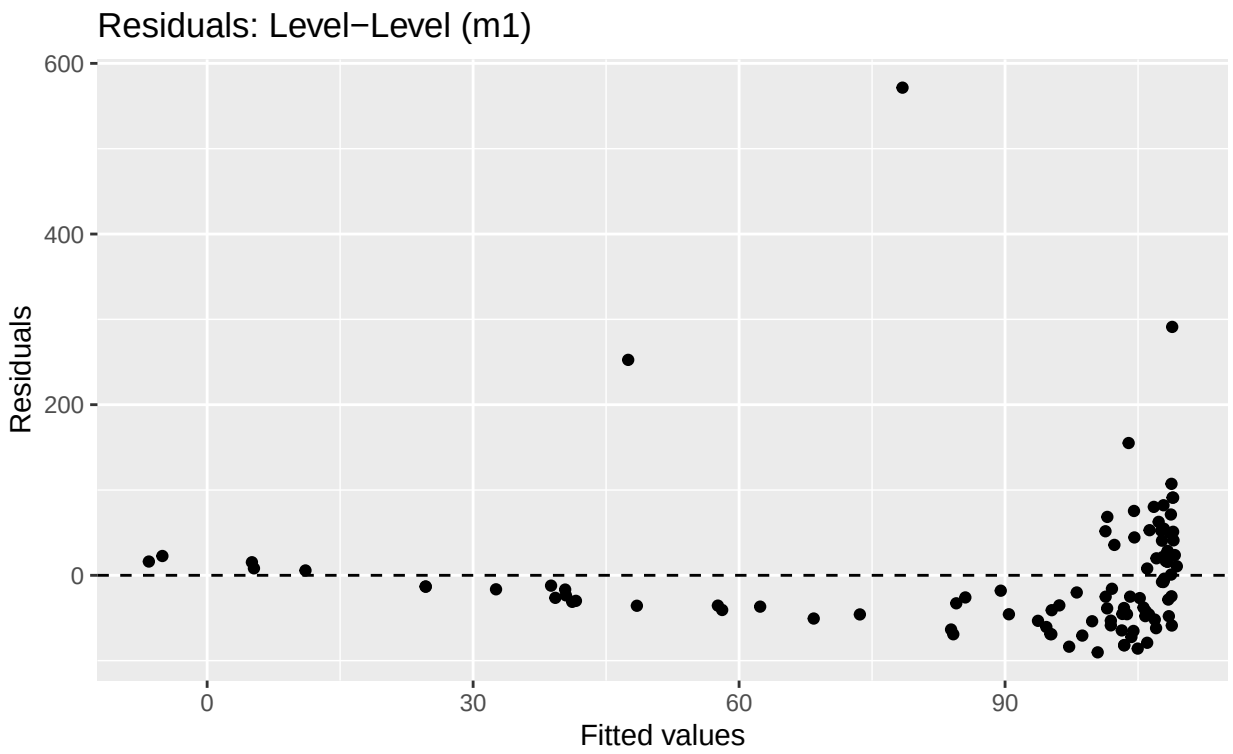
```
coef(m1)["income"] * 1000
```

```
##    income
## -20.90658
```

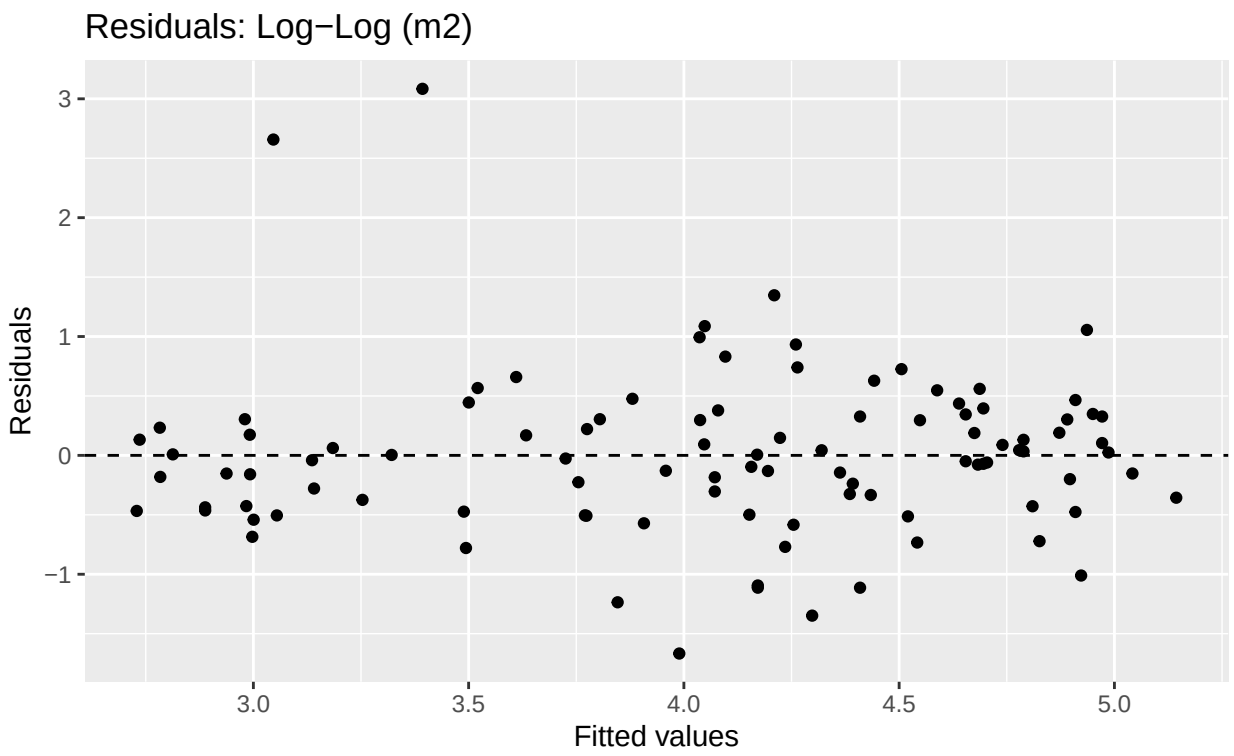
In the log-log model, the coefficient is an elasticity: a 1% increase in income is associated with a change of approximately that percentage in infant mortality. For a 10% increase in income, the predicted change in infant mortality is roughly 10 times the coefficient in percentage terms.

d) Residual plots for both models:

```
m1_aug = augment(m1)
ggplot(m1_aug, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "Fitted values", y = "Residuals", title = "Residuals: Level-Level (m1)")
```




```
m2_aug = augment(m2)
ggplot(m2_aug, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "Fitted values", y = "Residuals", title = "Residuals: Log-Log (m2)")
```



The level-level model shows a strong curved pattern in the residuals and clear heteroskedasticity. The log-log model substantially improves both issues, producing a much more random scatter of residuals.

3. Multiple regression with controls

a) Log-log model with controls:

```
m3 = lm(log(infant) ~ log(income) + region + oil, data = df)
tidy(m3)
```

```
## # A tibble: 6 x 5
##   term           estimate std.error statistic  p.value
##   <chr>             <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)       6.55     0.350      18.7  1.14e-33
## 2 log(income)      -0.340    0.0666     -5.10 1.70e- 6
## 3 regionAmericas   -0.550    0.184      -2.98 3.66e- 3
## 4 regionAsia       -0.713    0.158      -4.52 1.75e- 5
## 5 regionEurope     -1.03     0.257      -4.03 1.14e- 4
## 6 oilyes            0.640    0.225       2.84 5.44e- 3
```

b) The coefficient on `log(income)` represents the elasticity of infant mortality with respect to income, controlling for region and oil-exporting status. Comparing with the bivariate log-log model, controlling for region and oil status changes the income coefficient, indicating that some of the bivariate association was driven by regional differences and oil wealth.

c) The coefficient on the Africa region indicator (relative to the reference category) shows the log-difference in infant mortality between African countries and the reference group, holding income and oil status constant. The positive coefficient indicates that African countries have higher infant mortality than the reference group even after accounting for income differences.

d) Average marginal effects:

```
avg_slopes(m3)
```

```
##
##    Term          Contrast Estimate Std. Error     z Pr(>|z|)    S   2.5 %
##  income dY/dX             -0.00159   0.000311 -5.10  < 0.001 21.5 -0.0022
##  oil    yes - no           0.64021   0.225052  2.84  0.00444  7.8  0.1991
##  region Americas - Africa -0.54984   0.184492 -2.98  0.00288  8.4 -0.9114
##  region Asia - Africa     -0.71292   0.157572 -4.52  < 0.001 17.3 -1.0218
##  region Europe - Africa   -1.03383   0.256717 -4.03  < 0.001 14.1 -1.5370
##     97.5 %
##  -0.000978
##   1.081304
##  -0.188239
##  -0.404087
##  -0.530671
##
## Columns: term, contrast, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high
## Type:  response
```

The AME of income tells us the average predicted change in log(infant mortality) for a one-dollar increase in income, computed across all observed values. Because the model is in logs, the marginal effect on the original scale depends on the income level.

4. Interaction: oil status and income

a) Interaction model:

```
m4 = lm(log(infant) ~ log(income) * oil + region, data = df)
tidy(m4)
```

```
## # A tibble: 7 x 5
##   term               estimate std.error statistic  p.value
##   <chr>                 <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)           7.00     0.342      20.5  2.06e-36
## 2 log(income)          -0.432    0.0657     -6.58 2.60e- 9
## 3 oilyes               -4.38     1.24       -3.52 6.59e- 4
## 4 regionAmericas       -0.416    0.174      -2.39 1.87e- 2
## 5 regionAsia           -0.629    0.147      -4.27 4.66e- 5
## 6 regionEurope         -0.759    0.247      -3.07 2.77e- 3
## 7 log(income):oilyes    0.808    0.197       4.10 8.89e- 5
```

b) Marginal effect of income by oil status:

```
avg_slopes(m4, variables = "income", by = "oil")
```

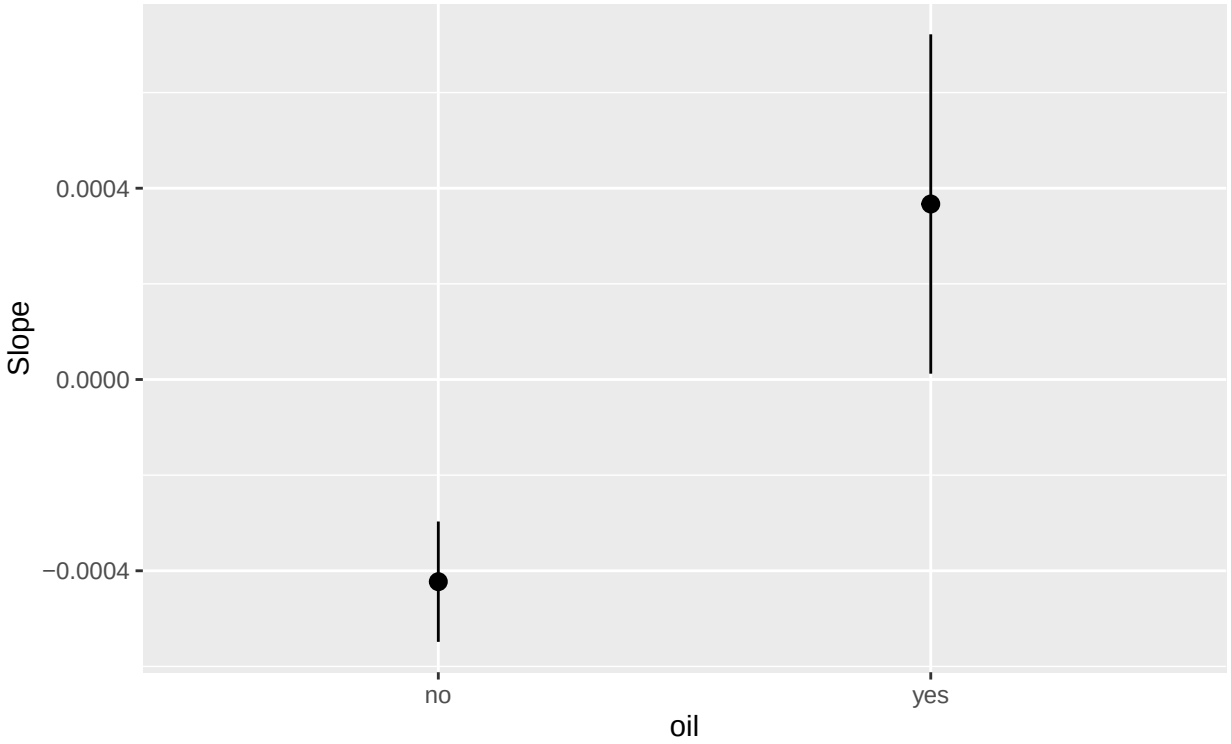
```
##
##    Term    Contrast oil Estimate Std. Error     z Pr(>|z|)    S      2.5 %
##  income mean(dY/dX) no  -0.00208   0.000316 -6.58   <0.001 34.3 -0.0027032
##  income mean(dY/dX) yes  0.00111   0.000548  2.03   0.0427  4.6  0.0000367
```

```
##    97.5 %
##  -0.00146
##   0.00218
##
## Columns: term, contrast, oil, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high, pr
## Type:  response
```

c) The interaction allows the income-mortality relationship to differ for oil-exporting and non-oil countries. Oil-exporting countries may have high income from resource extraction without the broader economic development (health infrastructure, education, governance) that normally accompanies income growth. This "resource curse" pattern can weaken the income-mortality link.

d) Plot marginal effects by oil status:

```
p1 = plot_slopes(m4, variables = "income", condition = "oil")
p1
```




```
ggsave("slopes_oil.png", p1, width = 6, height = 4)
```

5. Predicted values for specific scenarios

a) Predicted values for three scenarios:

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

```
##
##  income   region oil Estimate Std. Error     z Pr(>|z|)     S 2.5 % 97.5 %
##    1000 Africa   no      4.20      0.163 25.84   <0.001 486.7  3.89   4.52
##    1000 Africa   yes     4.84      0.238 20.38   <0.001 304.4  4.38   5.31
##    1000 Americas no      3.65      0.133 27.53   <0.001 551.8  3.39   3.91
##    1000 Americas yes     4.29      0.237 18.16   <0.001 242.4  3.83   4.76
##    1000 Europe   no      3.17      0.153 20.71   <0.001 314.2  2.87   3.47
##    1000 Europe   yes     3.81      0.283 13.46   <0.001 134.7  3.26   4.37
##   10000 Africa   no      3.42      0.298 11.50   <0.001  99.2  2.84   4.01
##   10000 Africa   yes     4.06      0.322 12.61   <0.001 118.6  3.43   4.69
##   10000 Americas no      2.87      0.229 12.54   <0.001 117.5  2.42   3.32
##   10000 Americas yes     3.51      0.276 12.74   <0.001 121.1  2.97   4.05
##   10000 Europe   no      2.39      0.166 14.42   <0.001 154.1  2.06   2.71
##   10000 Europe   yes     3.03      0.264 11.49   <0.001  99.1  2.51   3.54
##   20000 Africa   no      3.19      0.341  9.34   <0.001  66.5  2.52   3.86
##   20000 Africa   yes     3.83      0.357 10.73   <0.001  86.7  3.13   4.53
##   20000 Americas no      2.64      0.269  9.82   <0.001  73.2  2.11   3.16
##   20000 Americas yes     3.28      0.302 10.85   <0.001  88.6  2.68   3.87
##   20000 Europe   no      2.15      0.195 11.06   <0.001  92.1  1.77   2.53
##   20000 Europe   yes     2.79      0.275 10.17   <0.001  78.3  2.25   3.33
##
## Columns: rowid, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high, infant, income,
## Type:  response
```

Since the outcome is `log(infant)`, we exponentiate to get infant mortality in the original scale:

```
preds$estimate_original = exp(preds$estimate)
preds %>% select(income, region, oil, estimate, estimate_original)
```

```
##
##  Estimate
##      4.20
##      4.84
##      3.65
##      4.29
##      3.17
##      3.81
##      3.42
##      4.06
##      2.87
##      3.51
##      2.39
##      3.03
##      3.19
##      3.83
##      2.64
##      3.28
##      2.15
##      2.79
```

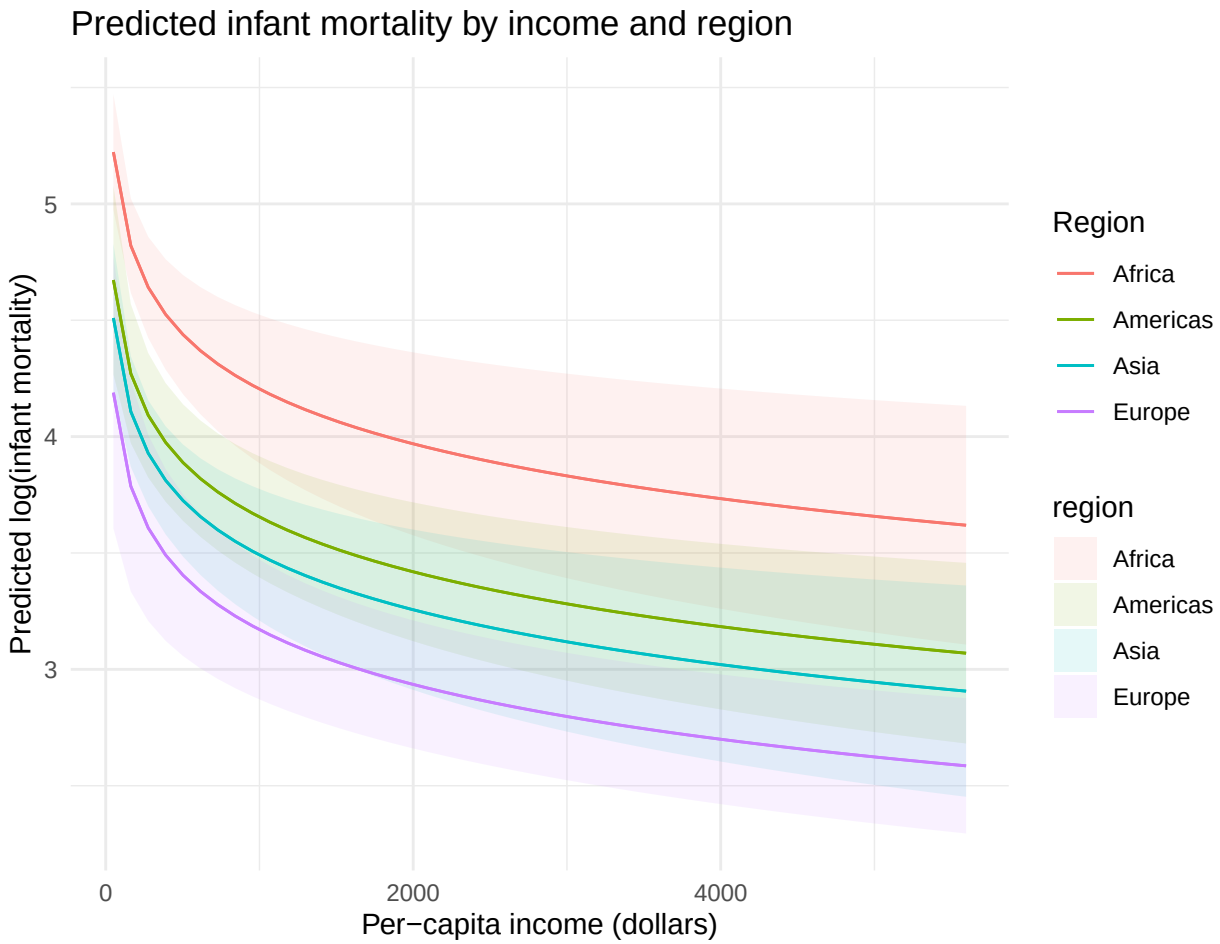
```
##
## Columns: income, region, oil, estimate, estimate_original
```

b) The predicted infant mortality rates are plausible given what we know about global health patterns. The gap between the African low-income scenario and the European high-income scenario is substantial, reflecting the combined effect of income differences and regional factors. The Americas oil-exporting scenario falls between the two.

6. Publication-quality visualization

a) Prediction plot by region:

```
p2 = plot_predictions(m3, condition = c("income", "region")) +
  labs(
    x = "Per-capita income (dollars)",
    y = "Predicted log(infant mortality)",
    title = "Predicted infant mortality by income and region",
    color = "Region") +
  theme_minimal()
p2
```




```
ggsave("pred_plot_region.png", p2, width = 7, height = 5)
```

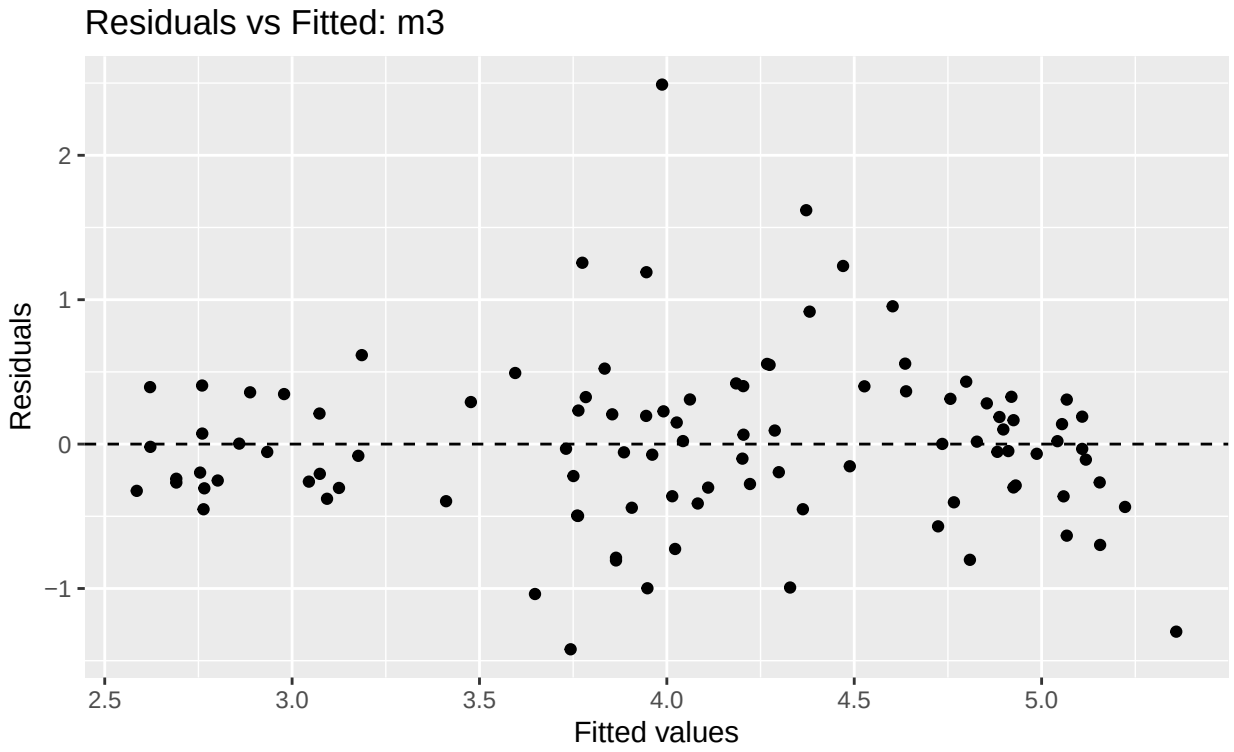
b) This plot shows that infant mortality decreases with income across all regions, but at different levels. African

countries have consistently higher predicted infant mortality than other regions at the same income level, reflecting structural factors beyond income—such as disease burden, health system capacity, and historical underinvestment. European countries have the lowest predicted mortality at any given income level. The income-mortality gradient is steepest at low income levels, meaning that economic growth has the largest potential health impact in the poorest countries. However, this analysis has important limitations. First, it is cross-sectional and cannot establish causation: countries that invest in health may also grow economically, creating reverse causality. Second, many relevant variables are omitted (education, governance, health spending). Third, the ecological fallacy means that country-level relationships may not apply to individuals within countries. Finally, the log-log specification, while fitting the data well, imposes a specific functional form assumption.

7. Diagnostics and robust inference

a) Residuals vs. fitted for model 3:

```
m3_aug = augment(m3)
ggplot(m3_aug, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "Fitted values", y = "Residuals", title = "Residuals vs Fitted: m3")
```



The residual pattern looks reasonably random, though some heteroskedasticity may be present. This motivates the use of robust standard errors.

b) Regression table comparing all four models:

```
modelsummary(
  list("Level" = m1, "Log-Log" = m2,
       "Controls" = m3, "Interaction" = m4),
  vcov = "robust",
```

```
  stars = TRUE,
  gof_map = c("r.squared", "nobs"),
  output = "markdown")
```

	Level	Log-Log	Controls	Interaction
(Intercept)	110.421***	7.146***	6.552***	7.005***
	(9.037)	(0.276)	(0.398)	(0.282)
income	-0.021***			
	(0.004)			
log(income)		-0.512***	-0.340***	-0.432***
		(0.049)	(0.078)	(0.057)
regionAmericas			-0.550**	-0.416*
			(0.184)	(0.162)
regionAsia			-0.713***	-0.629**
			(0.207)	(0.203)
regionEurope			-1.034***	-0.759***
			(0.256)	(0.211)
oilyes			0.640	-4.383
			(0.472)	(3.019)
log(income) × oilyes				0.808+
				(0.481)
R2	0.109	0.502	0.646	0.700
Num.Obs.	101	101	101	101

Note: + p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

c) Compare robust and default standard errors for m3:

```
modelsummary(
  list("Default SEs" = m3, "Robust SEs" = m3),
  vcov = list("classical", "robust"),
  stars = TRUE,
  gof_map = c("r.squared", "nobs"),
  output = "markdown")
```

	Default SEs	Robust SEs
(Intercept)	6.552***	6.552***
	(0.350)	(0.398)
log(income)	-0.340***	-0.340***
	(0.067)	(0.078)
regionAmericas	-0.550**	-0.550**
	(0.184)	(0.184)
regionAsia	-0.713***	-0.713***
	(0.158)	(0.207)
regionEurope	-1.034***	-1.034***
	(0.257)	(0.256)

	Default SEs	Robust SEs
oilyes	0.640**	0.640
	(0.225)	(0.472)
R2	0.646	0.646
Num.Obs.	101	101

Note: + p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

The robust standard errors may differ from the default ones, particularly if heteroskedasticity is present. When differences are small, the conclusions are robust. We use robust standard errors as a safeguard against heteroskedasticity, which is common in cross-country data where variance in outcomes often depends on the level of development.