# Jeffrey_hw8

Jack Jeffrey

## HW8

## setup

```r
# Set consistent root directory for all chunks
knitr::opts_knit$set(root.dir = "/Users/jackjeffrey/Documents/Poli502_Jeffrey/Data")
# Set CRAN mirror to avoid errors during package installation
options(repos = c(CRAN = "https://cran.rstudio.com/"))

install.packages("effects")  # Only if not already installed
```

```
The downloaded binary packages are in
    /var/folders/5g/td22kj7s5q9frby6hwm2szv80000gn/T//RtmpEdoKX8/downloaded_packages
```

```r
library(effects)
```

```
Loading required package: carData
```

```
lattice theme set by effectsTheme()
See ?effectsTheme for details.
```

```r
getwd()
```

```
[1] "/Users/jackjeffrey/Documents/Poli502_Jeffrey/hw8"
```

```r
setwd("/Users/jackjeffrey/Documents/Poli502_Jeffrey/Data")


world <- read.csv("world.csv")


world <- read.csv("/Users/jackjeffrey/Documents/Poli502_Jeffrey/Data/world.csv")
```

## 1. Estimate a constant-only model

```r
# Fit a constant only model
model_constant <- lm(women09 ~ 1, data = world)
# View summary of model
summary(model_constant)
```

```
Call:
lm(formula = women09 ~ 1, data = world)

Residuals:
    Min      1Q  Median      3Q     Max
-17.177  -7.477  -1.627   5.773  39.123

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.1772     0.8238   20.85   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.05 on 179 degrees of freedom
  (11 observations deleted due to missingness)
```

```r
# Estimate of intercept/mean equals 17.17772 which is 17% female representation
# across all countries on average, standard error of 0.8238, P-value below zero
# indicates the result is statistically significant meaning there is a clear
# average.
# 110 Observations not included due to NA's.
```

## 2. Estimate a model that uses per capita GDP (gdp_10_thou) as the main independent variable.

```
# Fitting a linear model with women09 variable as dependent and gdp_10_thou
# as independent variable
model_gdp <- lm(women09 ~ gdp_10_thou, data = world)
# view summary of model
summary(model_gdp)
```

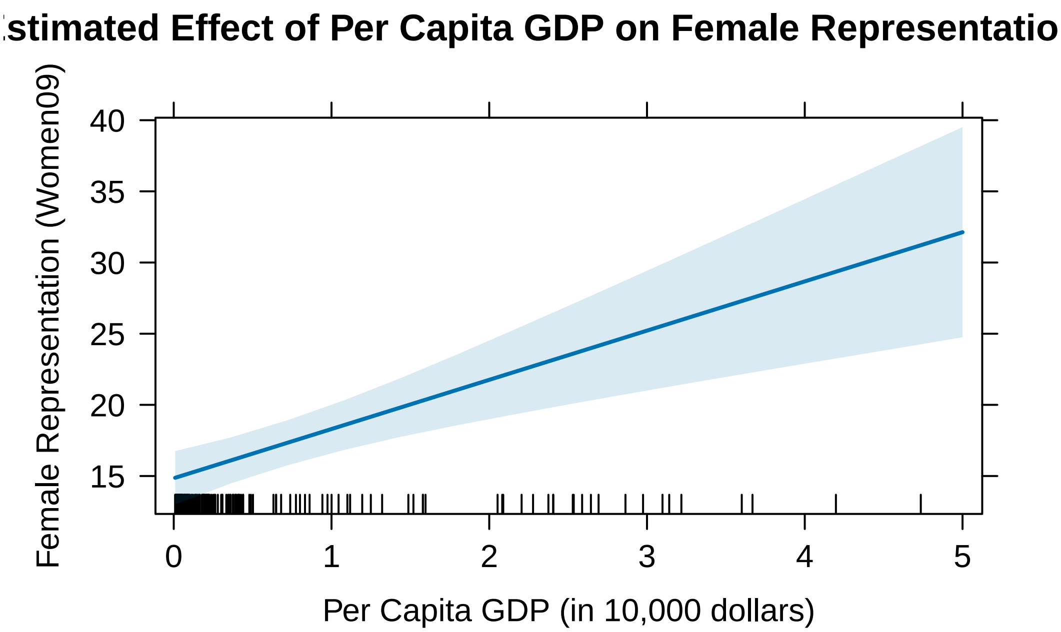
```
Call:
lm(formula = women09 ~ gdp_10_thou, data = world)

Residuals:
   Min     1Q Median     3Q    Max 
-24.74  -6.74  -1.62   5.78  41.38 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  14.8430     0.9542   15.56  < 2e-16 ***
gdp_10_thou   3.4574     0.8351    4.14  5.5e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.38 on 167 degrees of freedom
  (22 observations deleted due to missingness)
Multiple R-squared:  0.09308,	Adjusted R-squared:  0.08765 
F-statistic: 17.14 on 1 and 167 DF,  p-value: 5.501e-05
```

```
# The intercept is 14.8430. When GDP (gdp_10_thou) is 0, the baseline female
# representation would be approximately 14.84%. The coefficient for GDP is 3.4574, meaning
# for each increase of 1,000 dollars in per capita GDP, female representation is
# expected to increase by 3.46%. The p-value for gdp_10_thou is 5.5e-05, which is
# much smaller than 0.05. This means that the relationship between GDP and female
# representation is statistically significant.
# The R-squared value is low (9.3%), which suggests that while GDP
# is statistically significant, it explains only a small portion
# of the variation in female representation.
```

### 3. Create a graph that shows the estimated effect of per capita GDP on female representation

### using the effect function.

```
# Create effect graph for GDP per capita's effect on female representation
effect_gdp <- effect("gdp_10_thou", model_gdp)
plot(effect_gdp, main = "Estimated Effect of Per Capita GDP on Female Representation",
     xlab = "Per Capita GDP (in 10,000 dollars)", ylab = "Female Representation (Women09)")
```




### 4. Estimate a model that uses a dummy variable that measures electoral system (pr_sys) as the

### main independent variable.

```
# Investigate PR system variable to see if it is binary
unique(world$pr_sys)
```

```
[1] "No"  "Yes"
```

```r
# Estimate a model with PR system as the main independent variable
model_pr_sys <- lm(women09 ~ pr_sys, data = world)
# summarize result
summary(model_pr_sys)
```

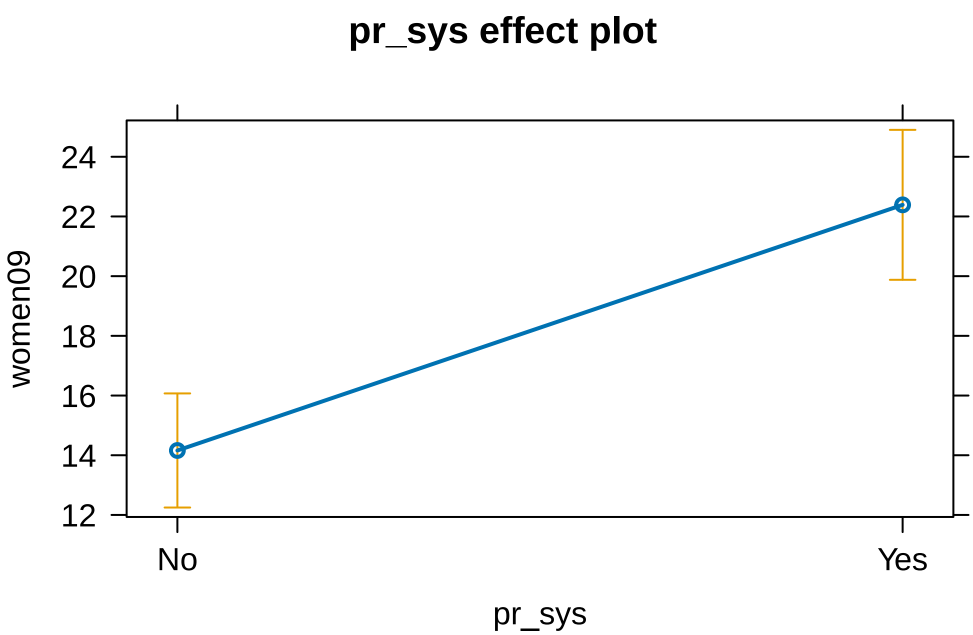
```

Call:
lm(formula = women09 ~ pr_sys, data = world)

Residuals:
    Min      1Q  Median      3Q     Max 
-16.589  -7.835  -1.860   7.140  33.911 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  14.1596     0.9686  14.619  < 2e-16 ***
pr_sysYes     8.2297     1.5996   5.145 7.02e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.34 on 178 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.1295,    Adjusted R-squared:  0.1246 
F-statistic: 26.47 on 1 and 178 DF,  p-value: 7.019e-07
```

```r
# Proportional representation has a positive effect on female representation,
# with a coefficient of 8.23. P-value is 7.02e-07 and highly significant,
# meaning the electoral system has a statistically significant impact
# on female representation. The R-squared value is quite low, suggesting
# that the electoral system alone explains only about 12.95% of the
# variation in female representation. For countries with a non-proportional
# representation electoral system, the average
# female representation is about 14.16.
```

### 5. Create a graph that shows the estimated effect of electoral system on female representation

### using the effect function.

```
# Create an effect graph for the effect of electoral systems
# on female representation
effect_pr_sys <- effect("pr_sys", model_pr_sys)
# Plot the effect
plot(effect_pr_sys)
```




```
# Based on the plot, proportional representation systems result in
# much higher female representation
```

### 6. Estimate a model that includes per capita GDP AND electoral system dummy at the same time.

```r
# Create a multiple regression model
model_gdp_prsys <- lm(women09 ~ gdp_10_thou + pr_sys, data = world)
# View results
summary(model_gdp_prsys)
```

```
Call:
lm(formula = women09 ~ gdp_10_thou + pr_sys, data = world)

Residuals:
    Min      1Q  Median      3Q     Max
-20.371  -7.872  -1.266   6.399  36.148

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.3922     1.0248  12.092  < 2e-16 ***
gdp_10_thou   2.7864     0.7948   3.506 0.000585 ***
pr_sysYes     7.7007     1.5713   4.901 2.25e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.73 on 166 degrees of freedom
  (22 observations deleted due to missingness)
Multiple R-squared:  0.2077,    Adjusted R-squared:  0.1982
F-statistic: 21.76 on 2 and 166 DF,  p-value: 4.049e-09
```

```r
# Interpretation - The residual standard error of 9.73 is the average difference
# between the actual values of female representation and the values
# predicted by the model. 20.77%  of the variation in female representation
# is explained by the per capita GDP and electoral system combined based on the R-squared val
# With a p-value of 4.049e-09, this model is highly significant.
```

## 7. Comparing the four models you have estimated so far, which one fits the data best?

```r
# Model 4 fits the data best of the four models as it has
# the lowest Residual Standard Error and highest Adjusted R-squared
# meaning it accounts for more of the variation in female representation after penalizing fo
```

## 8. Create a graph that shows the effect of per capita GDP on Y for countries that adopt a

## proportional representation system.

```
# filter out missing data in pr_sys (ensure pr_sys is a factor with 1/0 values or "Yes"/"No")
world$pr_sys <- as.factor(world$pr_sys)  # make sure it's a factor
levels(world$pr_sys)  # check factor levels to ensure correct filtering
```

```
[1] "No"  "Yes"
```

```
# Filter data for countries with proportional representation system
pr_data <- subset(world, pr_sys == "Yes")  # Adjust to match your factor level

# Define the model
pr_model <- lm(women09 ~ gdp_10_thou, data = pr_data)

# Generate predictions
gdp_seq <- seq(min(pr_data$gdp_10_thou, na.rm = TRUE),
               max(pr_data$gdp_10_thou, na.rm = TRUE),
               length.out = 100)

# Create data frame from predictions
prediction_data <- data.frame(gdp_10_thou = gdp_seq)

# Predict Y for GDP values
prediction_data$women09_pred <- predict(pr_model, newdata = prediction_data)

# Load tidyverse
library(tidyverse)
```

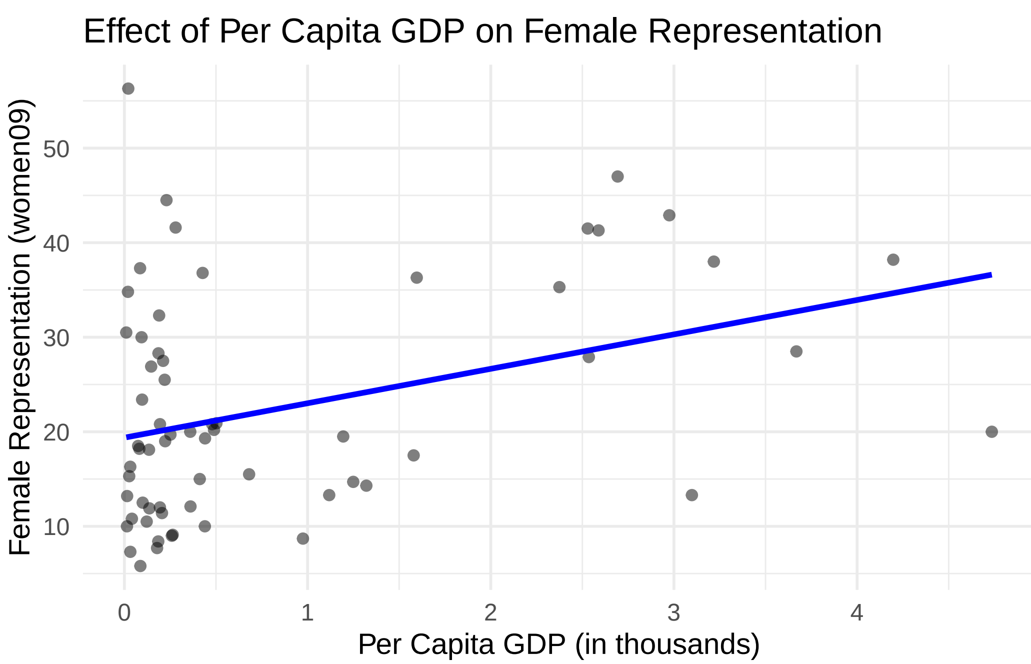
```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
```

```
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
# Graph plot
ggplot(pr_data, aes(x = gdp_10_thou, y = women09)) +
  geom_point(alpha = 0.5) +  # Scatterplot of actual data
  geom_line(data = prediction_data, aes(x = gdp_10_thou, y = women09_pred),
            color = "blue", size = 1) +  # Predicted effect line
  labs(
    title = "Effect of Per Capita GDP on Female Representation",
    x = "Per Capita GDP (in thousands)",
    y = "Female Representation (women09)"
  ) +
  theme_minimal()
```

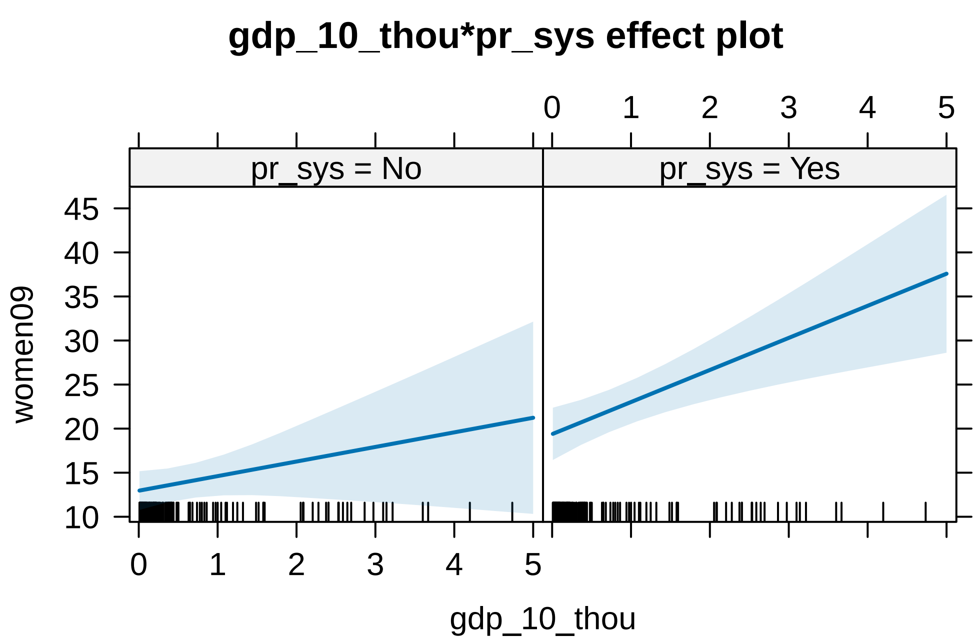
```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

```
Warning: Removed 4 rows containing missing values or values outside the scale range
(`geom_point()`).
```



### 9. Try creating the same graph by providing "gdp_10_thou:pr_sys" as the term.

```
# Model with interaction term
model_interaction <- lm(women09 ~ gdp_10_thou * pr_sys, data = world)

# Plot effect for interaction term
effect_gdp_pr_interaction <- Effect(
  focal.predictors = c("gdp_10_thou", "pr_sys"),
  mod = model_interaction,
  given.values = list(pr_sysYes = 1)
)

# Plot graph
plot(effect_gdp_pr_interaction)
```



## 10. Estimate a regression model of female representation that uses region as the main

## independent variable.

```r
# Ensure region is a factor variable
world$region <- as.factor(world$region)
# Estimate regression model
model_region <- lm(women09 ~ region, data = world)
# View model summary
summary(model_region)
```

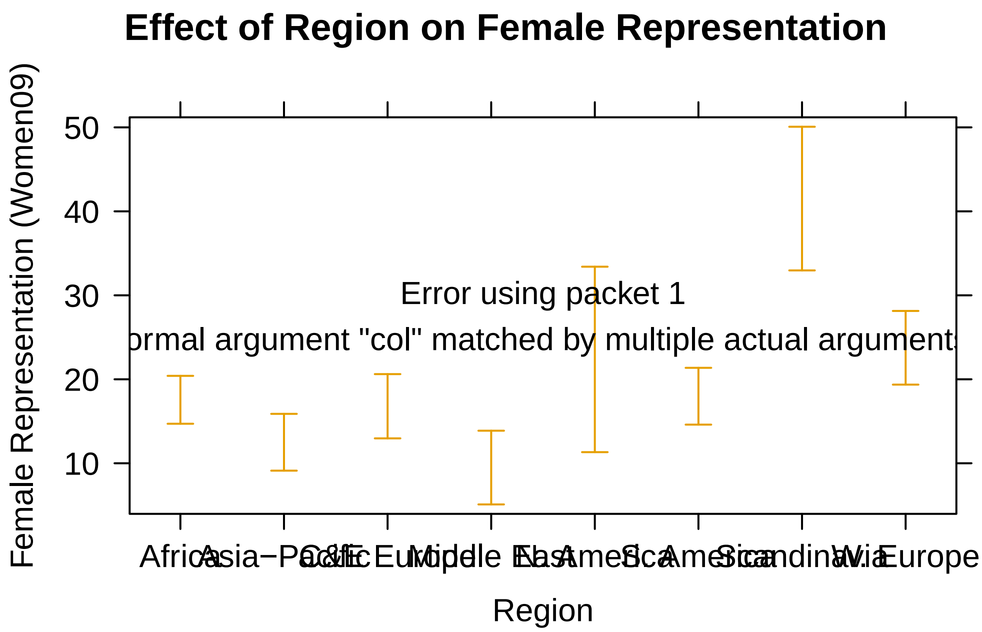
```
Call:
lm(formula = women09 ~ region, data = world)

Residuals:
    Min      1Q  Median      3Q     Max 
-17.984  -7.135  -1.746   4.914  38.740 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)         17.5600     1.4445  12.157  < 2e-16 ***
regionAsia-Pacific  -5.0569     2.2407  -2.257   0.0253 *  
regionC&E Europe    -0.7680     2.4171  -0.318   0.7511    
regionMiddle East   -8.0705     2.6511  -3.044   0.0027 ** 
regionN. America     4.8067     5.7780   0.832   0.4066    
regionS. America     0.4244     2.2407   0.189   0.8500    
regionScandinavia   23.9600     4.5679   5.245 4.54e-07 ***
regionW. Europe      6.1926     2.6511   2.336   0.0207 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.69 on 172 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.2615,    Adjusted R-squared:  0.2314 
F-statistic:   8.7 on 7 and 172 DF,  p-value: 3.929e-09
```

```
# The regions with statistically significant results are - Asia-Pacific, Middle
# East, Scandinavia, and Western Europe.
# The Scandinavia region has the highest positive impact
# on female representation, while the Middle East has the
# largest negative impact. The Multiple R-squared of 0.2615
# indicates that approximately 26.15% of the variability
# in female representation is explained by the region variable.
```

## 11. Create an effect plot that shows the relationship between region and female

## representation.

```
# Generate effect of region on female representation
effect_region <- Effect(focal.predictors = "region", mod = model_region)
# Plot the effect
plot(effect_region, main = "Effect of Region on Female Representation",
     xlab = "Region", ylab = "Female Representation (Women09)",
     col = "blue", cex.lab = 1.2, cex.axis = 1.2)
```



## 12. Estimate a regression model of female representation on per capita GDP that controls for

## region.

```r
# Fit the model with per capita GDP and region as independent variables
model_gdp_region <- lm(women09 ~ gdp_10_thou + region, data = world)
# View model summary
summary(model_gdp_region)
```

```
Call:
lm(formula = women09 ~ gdp_10_thou + region, data = world)

Residuals:
    Min      1Q  Median      3Q     Max
-17.123  -6.959  -1.845   4.348  38.563

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         17.7141     1.4194  12.480  < 2e-16 ***
gdp_10_thou          1.0914     1.2707   0.859 0.391693
regionAsia-Pacific  -4.9372     2.2856  -2.160 0.032254 *
regionC&E Europe    -1.2315     2.3610  -0.522 0.602678
regionMiddle East  -11.0675     2.8079  -3.942 0.000121 ***
regionN. America     2.2842     6.1856   0.369 0.712410
regionS. America    -0.9546     2.2279  -0.428 0.668890
regionScandinavia   20.3981     5.8626   3.479 0.000648 ***
regionW. Europe      3.1906     3.9828   0.801 0.424264
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.38 on 160 degrees of freedom
  (22 observations deleted due to missingness)
Multiple R-squared:  0.2903,    Adjusted R-squared:  0.2549
F-statistic: 8.183 on 8 and 160 DF,  p-value: 2.991e-09
```

```r
# The model shows the effect of GDP after controlling for region.
# The Adjusted R-squared of 0.2643 is slightly higher than previous models.
```

# # 13. Estimate a regression model of female representation on frac_eth3, a

# three-category ordinal variable that measures levels of ethnic fractionalization.

```r
# Create a model using ethnic fractionalization as a categorical nominal variable
model_frac_eth3 <- lm(women09 ~ factor(frac_eth3), data = world)
# view summary of model
summary(model_frac_eth3)
```

```
Call:
lm(formula = women09 ~ factor(frac_eth3), data = world)

Residuals:
    Min      1Q  Median      3Q     Max
-17.457  -7.502  -1.610   5.493  39.416

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)               11.967      6.422   1.864   0.0641 .
factor(frac_eth3)High      5.469      6.583   0.831   0.4072
factor(frac_eth3)Low       5.491      6.578   0.835   0.4050
factor(frac_eth3)Medium    4.918      6.588   0.746   0.4564
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.12 on 176 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.004347,  Adjusted R-squared:  -0.01262
F-statistic: 0.2562 on 3 and 176 DF,  p-value: 0.8569
```

## 14. Based on the results, do you think ethnic fractionalization levels have a positive/negative impact on female representation?

```
# Based on the regression results,
# we cannot conclude that ethnic fractionalization
# levels have a statistically significant positive
# or negative impact on female
# representation in this dataset.
# The p-values for all categories of ethnic fractionalization
# are high, and the model as a whole is not significant.
```