Jeffrey_hw7_exercise1

Jack Jeffrey

HW 7 Bivariate Regression

Question 1

```
# load necessary packages
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 beco
```

```
# specify path to world data
getwd()
```

```
[1] "/Users/jackjeffrey/Documents/Poli502_Jeffrey/hw7"
```

```
setwd("/Users/jackjeffrey/Documents/Poli502_Jeffrey/Data")
# assign world data an object
world <- read.csv("world.csv")
# view summary statistics for women and gdp variables
summary(world$women09)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00    9.70   15.55   17.18   22.95   56.30      11
```

```r
length(world$women09)
```

```
[1] 191
```

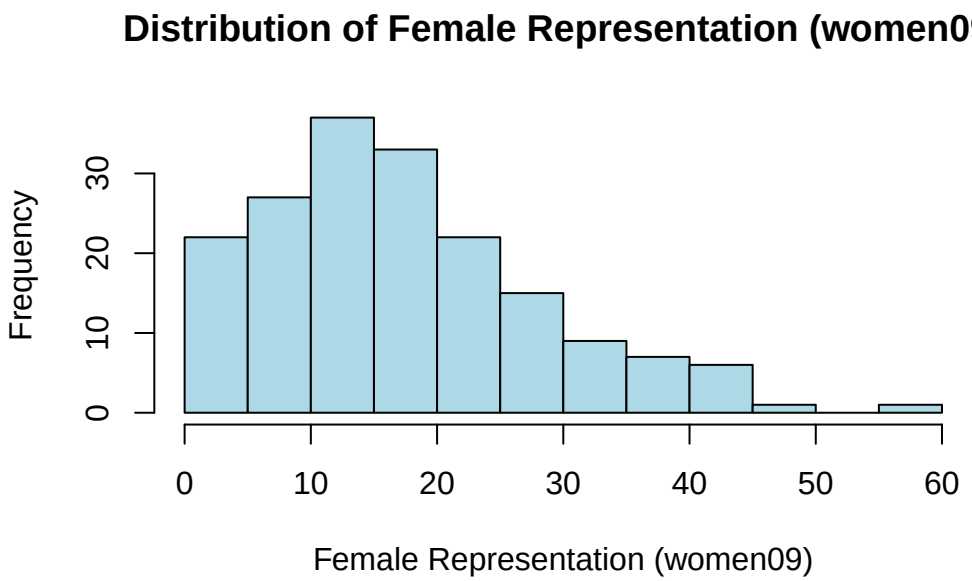
```r
# mean of 17.18 with 11 NA's
# 191 countries
summary(world$gdp10_thou)
```

```
Length  Class   Mode
     0   NULL   NULL
```

```r
# length 0 class and mode both null
```

Graphical univariate analysis

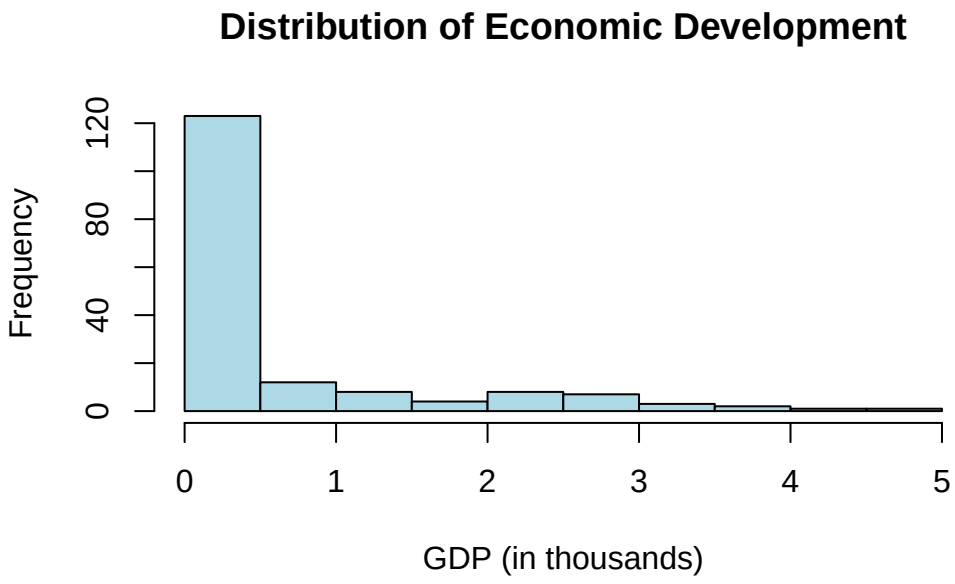
```r
# Histogram for (women09) variable
hist(world$women09, main="Distribution of Female Representation (women09)",
     xlab="Female Representation (women09)", col="lightblue", border="black",
     breaks=15)
```



```r
# female representation is the highest from 10-20 and starts to get progressively lower.
# Remove rows with missing data in either 'gdp_10_thou' or 'women09'
world_clean <- world[!is.na(world$gdp_10_thou) & !is.na(world$women09), ]
# view summary of gdp_10_thou cleaned
summary(world_clean$gdp_10_thou)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0090  0.0537  0.1941  0.6258  0.6495  4.7354
```

```r
# mean of 0.6258 with min of 0.0090 and max of 4.7354
# Histogram of economic developement variable
hist(world_clean$gdp_10_thou, main = "Distribution of Economic Development",
     xlab = "GDP (in thousands)", col = "lightblue", border = "black")
```




```r
# gdp recorded in thousands, highest peak in the 0-1 range with a frequency/countries over 1
```

Question 2.

```r
length(world_clean$women09)
```

```
[1] 169
```

```r
# With NA's removed its down to 169 countries
length(world_clean$gdp_10_thou)
```

```
[1] 169
```

```r
# With NA's removed length is no longer null and there are 169 countries
```

Question 3.

```r
# Calculate correlation coeficcient
cor_coefficient <- cor(world_clean$women09, world_clean$gdp_10_thou)
cor_coefficient
```

```
[1] 0.3050866
```

```r
# The correlation coeficcient is 0.305, positive moderate linear
# relationship, as economic developement increases female representation
# slighly increases
# Test for statistical significance
cor_test <- cor.test(world_clean$women09, world_clean$gdp_10_thou)
cor_test
```

```
    Pearson's product-moment correlation

data:  world_clean$women09 and world_clean$gdp_10_thou
t = 4.14, df = 167, p-value = 5.501e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1615677 0.4359677
sample estimates:
      cor
0.3050866
```

```r
# p-value of 5.501e-05 is much smaller 0.05 so we can reject the null, the # relationship is
# confidence interval
# 95 confidence interval of 0.161, 0.436
# T score of 4.14 indicating a significant relationship
# degrees of freedom equal to 167
```

Question 4

```
# Null Hypothesis (H0): (gdp_10_thou) has no effect on (women09).The slope # of the regressi

# Alternative Hypothesis (H1):(gdp_10_thou) has a significant effect on

# (women09).The slope of the regression line (beta_1) is not zero.
```

Question 5

```
# Run the linear regression model
model <- lm(women09 ~ gdp_10_thou, data = world_clean)
# View regression output
summary(model)
```

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

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
Multiple R-squared:  0.09308,   Adjusted R-squared:  0.08765
F-statistic: 17.14 on 1 and 167 DF,  p-value: 5.501e-05
```

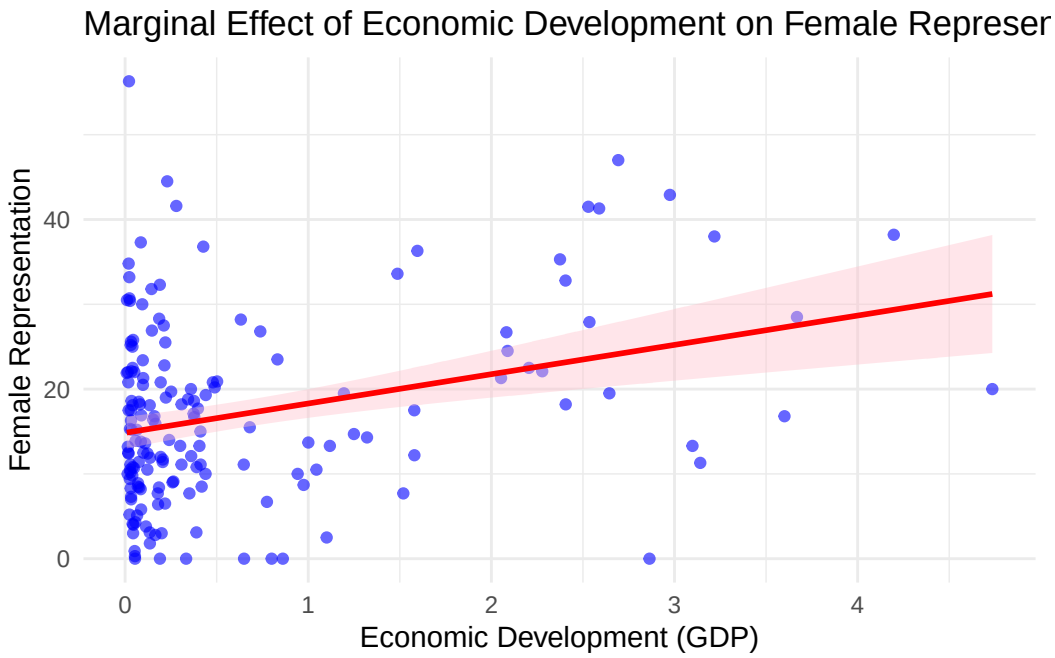
```
# Estimated regression equation
# (a) women09 = 14.8430 + 3.4574*gdp_10_thou
# Sign for the coefficient of X
# (b) X coefficient equals 3.4574 and is positive
```

```r
# Size of coefficient X
# (c) Size is 3.4574, for every unit increase of X, Y increases by 3.4574
# Significance and condidence level | if increased by 10,000 dollars
# women09 would increase by 3.4574% and if increased by 1,000 dollars
# women09 would increase by 0.34574%.
# (d) P-value of 5.501e-05 is much smaller than 0.05 and is statistically # significant at t
# hypothesis.
```

Question 6

```r
# graph regression results
ggplot(data = world_clean, aes(x = gdp_10_thou, y = women09)) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_smooth(method = "lm", color = "red", fill = "pink", se = TRUE) +
  labs(
    title = "Marginal Effect of Economic Development on Female Representation",
    x = "Economic Development (GDP)",
    y = "Female Representation"
  ) +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Graph interpretation

```
# The scatter plot points show the actual data, while the red regression line indicates the p
```

Question 7

```
# vew regression statistics
summary(model)
```

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

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
Multiple R-squared:  0.09308,   Adjusted R-squared:  0.08765
F-statistic: 17.14 on 1 and 167 DF,  p-value: 5.501e-05
```

```
# Based on the R sqaured resukt of 0.09308, only a 9.31% of women09 is
# explained by gdp_10_thou, this is too small of a value to confident that # women09 is cause
# There is a residual standard error of 10.38% which is quite large
# considering women09 is in percentages and is probably too large to feel
# good about the model.
```

Question 8

```
# obtain predicted values for women09
world_clean$predicted_women09 <- predict(model)
# find the predicted and actual values for women09 in Rwanda
rwanda_predicted <- world_clean$predicted_women09[world_clean$country == "Rwanda"]
rwanda_actual <- world_clean$women09[world_clean$country == "Rwanda"]
# Calculate difference between actual and predicted values
residual_rwanda <- rwanda_actual - rwanda_predicted
# evaluate values
cat("Predicted value for Rwanda:", rwanda_predicted, "\n")
```

```
Predicted value for Rwanda: 14.91632
```

```
cat("Actual value for Rwanda:", rwanda_actual, "\n")
```

```
Actual value for Rwanda: 56.3
```

```
cat("Difference (residual) for Rwanda:", residual_rwanda, "\n")
```

```
Difference (residual) for Rwanda: 41.38368
```

```
# interpretation - the models predcited value verses the actual value is
# noticeably different. The residual is 41.38 which indicates the model is
# not well suited to fit Rwanda.
```

Question 9

```
# Create a smaller subset for PR and non PR countries
pr_countries <- world_clean[world_clean$pr_sys == "Yes", ]
non_pr_countries <- world_clean[world_clean$pr_sys == "No", ]
# View observations for each object
cat("Number of PR countries:", nrow(pr_countries), "\n")
```

```
Number of PR countries: 63
```

```
cat("Number of Non-PR countries:", nrow(non_pr_countries), "\n")
```

```
Number of Non-PR countries: 106
```

```
# 63 PR countries
# 106 Non-PR countries

# Regression for PR countries
pr_model <- lm(women09 ~ gdp_10_thou, data = pr_countries)
cat("Regression results for PR countries:\n")
```

```
Regression results for PR countries:
```

```
summary(pr_model)
```

```

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

Residuals:
    Min      1Q  Median      3Q     Max
-17.357  -8.724  -1.473   7.316  36.847
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   19.375      1.748  11.084 3.01e-16 ***
gdp_10_thou    3.641      1.218   2.989  0.00403 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.25 on 61 degrees of freedom
Multiple R-squared:  0.1278,    Adjusted R-squared:  0.1135
F-statistic: 8.935 on 1 and 61 DF,  p-value: 0.004028
```

```r
# Regression for Non-PR Countries
non_pr_model <- lm(women09 ~ gdp_10_thou, data = non_pr_countries)
cat("Regression results for Non-PR countries:\n")
```

```
Regression results for Non-PR countries:
```

```r
summary(non_pr_model)
```

```

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

Residuals:
     Min       1Q   Median       3Q      Max
-17.6962  -6.6159  -0.6734   5.3486  20.2055

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   12.956      1.001  12.940   <2e-16 ***
gdp_10_thou    1.655      1.081   1.531    0.129
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.685 on 104 degrees of freedom
Multiple R-squared:  0.02204,   Adjusted R-squared:  0.01263
F-statistic: 2.343 on 1 and 104 DF,  p-value: 0.1288
```

Question 10

```
# Extract regression coeficcients from PR model
intercept <- coef(pr_model)[1]
slope <- coef(pr_model)[2]
# Print regression equation
cat("Estimated regression equation for PR countries: women09 =", round(intercept, 2), "+", ro
```

```
Estimated regression equation for PR countries: women09 = 19.38 + 3.64 * gdp_10_thou
```

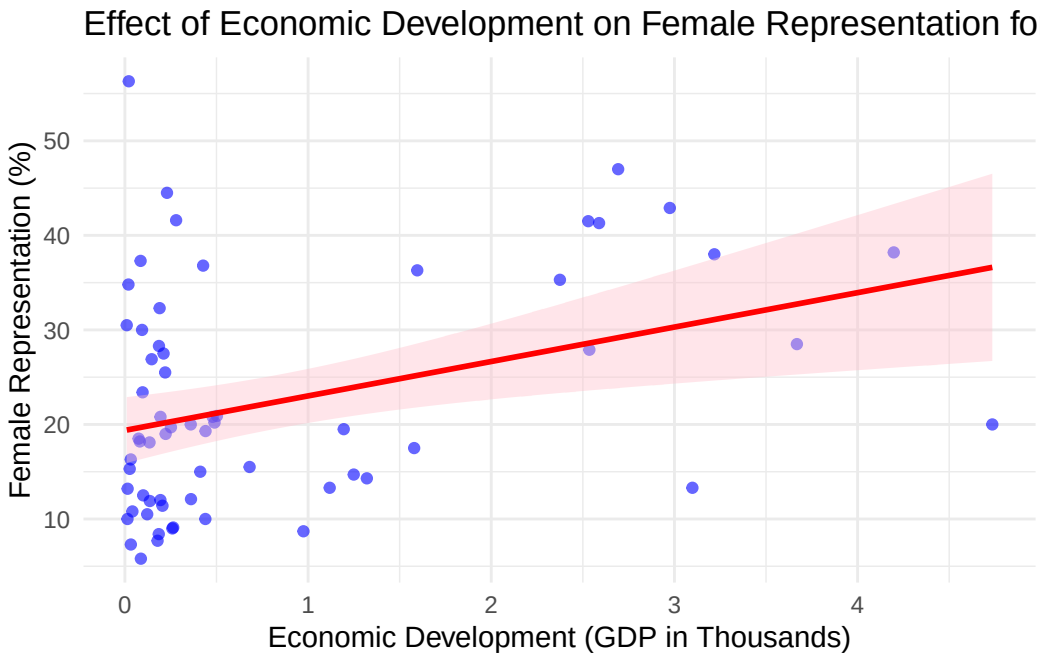
```
# Estimated regression equation for PR countries: women09 = 19.38 + 3.64 *
# gdp_10_thou
# Interpretation
cat("For PR countries, every additional unit increase in GDP (in thousands) is associated wit
    round(slope, 2), "% in female representation.\n")
```

```
For PR countries, every additional unit increase in GDP (in thousands) is associated with an
```

```
# For PR countries, every additional unit increase in GDP (in thousands) is associated with a

# Graph scatterplot for PR countries
ggplot(pr_countries, aes(x = gdp_10_thou, y = women09)) +
  geom_point(color = "blue", alpha = 0.6) +  # Add scatterplot points
  geom_smooth(method = "lm", color = "red", fill = "pink", se = TRUE) +  # Add regression lin
  labs(
    title = "Effect of Economic Development on Female Representation for PR Countries",
    x = "Economic Development (GDP in Thousands)",
    y = "Female Representation (%)"
  ) +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Graph Interpretation

```
# This is a weak positive relationship. The wide confidence intervals
# suggest more variability in the data. There relationship between the two
# variables is weaker in Non-PR compared to PR countries.
```

Question 11

```
# Extract regression coefficients for Non-Pr countries
intercept_nonpr <- coef(non_pr_model)[1]
slope_nonpr <- coef(non_pr_model)[2]
# Print regression equation
cat("Estimated regression equation for non-PR countries: women09 =", round(intercept_nonpr, 2
```

```
Estimated regression equation for non-PR countries: women09 = 12.96 + 1.66 * gdp_10_thou
```

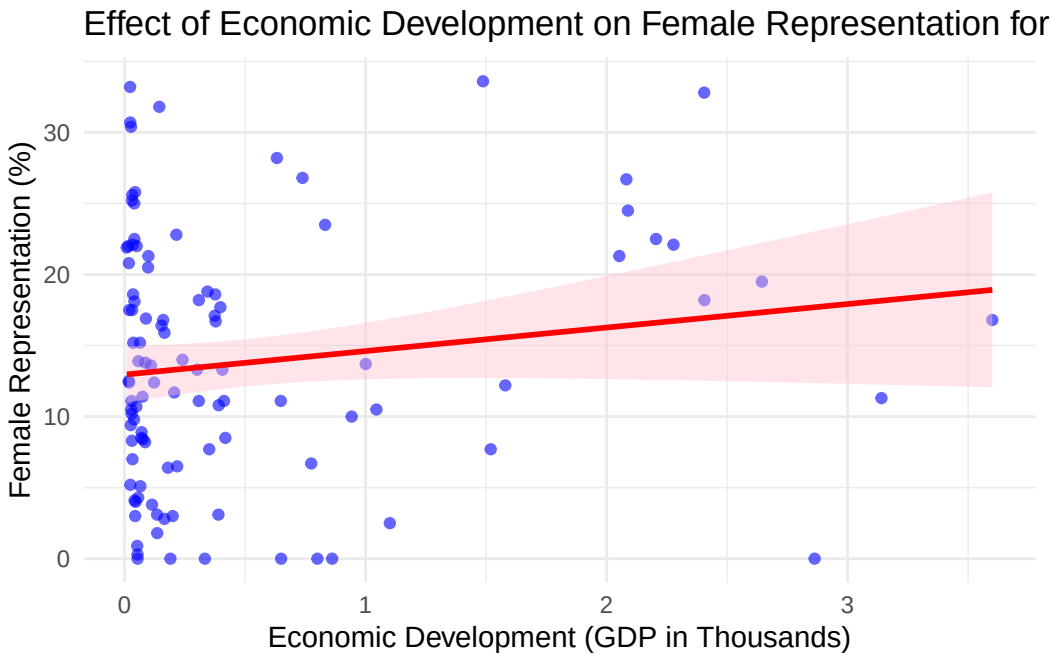
```r
# Estimated regression equation for non-PR countries: women09 = 12.96 +
# 1.66 * gdp_10_thou
# Interpretation
cat("For non-PR countries, every additional unit increase in GDP (in thousands) is associated
    round(slope_nonpr, 2), "% in female representation.\n")
```

```
For non-PR countries, every additional unit increase in GDP (in thousands) is associated with
```

```r
# For non-PR countries, every additional unit increase in GDP (in
# thousands) is associated with an increase of 1.66 % in female
# representation.

# Graph scatterplot for Non-Pr countries
ggplot(non_pr_countries, aes(x = gdp_10_thou, y = women09)) +
  geom_point(color = "blue", alpha = 0.6) +  # Scatterplot points
  geom_smooth(method = "lm", color = "red", fill = "pink", se = TRUE) +  # Regression line wi
  labs(
    title = "Effect of Economic Development on Female Representation for Non-PR Countries",
    x = "Economic Development (GDP in Thousands)",
    y = "Female Representation (%)"
  ) +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Question 12

```
# Find predicted values for PR countries
pr_countries$predicted_women09 <- predict(pr_model, newdata = pr_countries)
# Find predicted and actual values for Rwanda
rwanda_predicted <- pr_countries$predicted_women09[pr_countries$country == "Rwanda"]
rwanda_actual <- pr_countries$women09[pr_countries$country == "Rwanda"]
# Print predicted and actual values
cat("Predicted Female Representation for Rwanda:", round(rwanda_predicted, 2), "%\n")
```

```
Predicted Female Representation for Rwanda: 19.45 %
```

```
cat("Actual Female Representation for Rwanda:", round(rwanda_actual, 2), "%\n")
```

```
Actual Female Representation for Rwanda: 56.3 %
```

```
# Predicted Female Representation for Rwanda: 19.45 %
# Actual Female Representation for Rwanda: 56.3 %
```

```r
# Calculate and print residual
residual_rwanda <- rwanda_actual - rwanda_predicted
cat("Difference (Residual) between predicted and actual value for Rwanda:", round(residual_rw
```

```
Difference (Residual) between predicted and actual value for Rwanda: 36.85 %
```

```r
# Difference (Residual) between predicted and actual value for Rwanda:
# 36.85 %
# The model underestimated the actual value by 36.85%. Based on the model # it cannot be conl
```