Homework Week 10

Conor Craig

2023-11-06

Question 1

examining the data

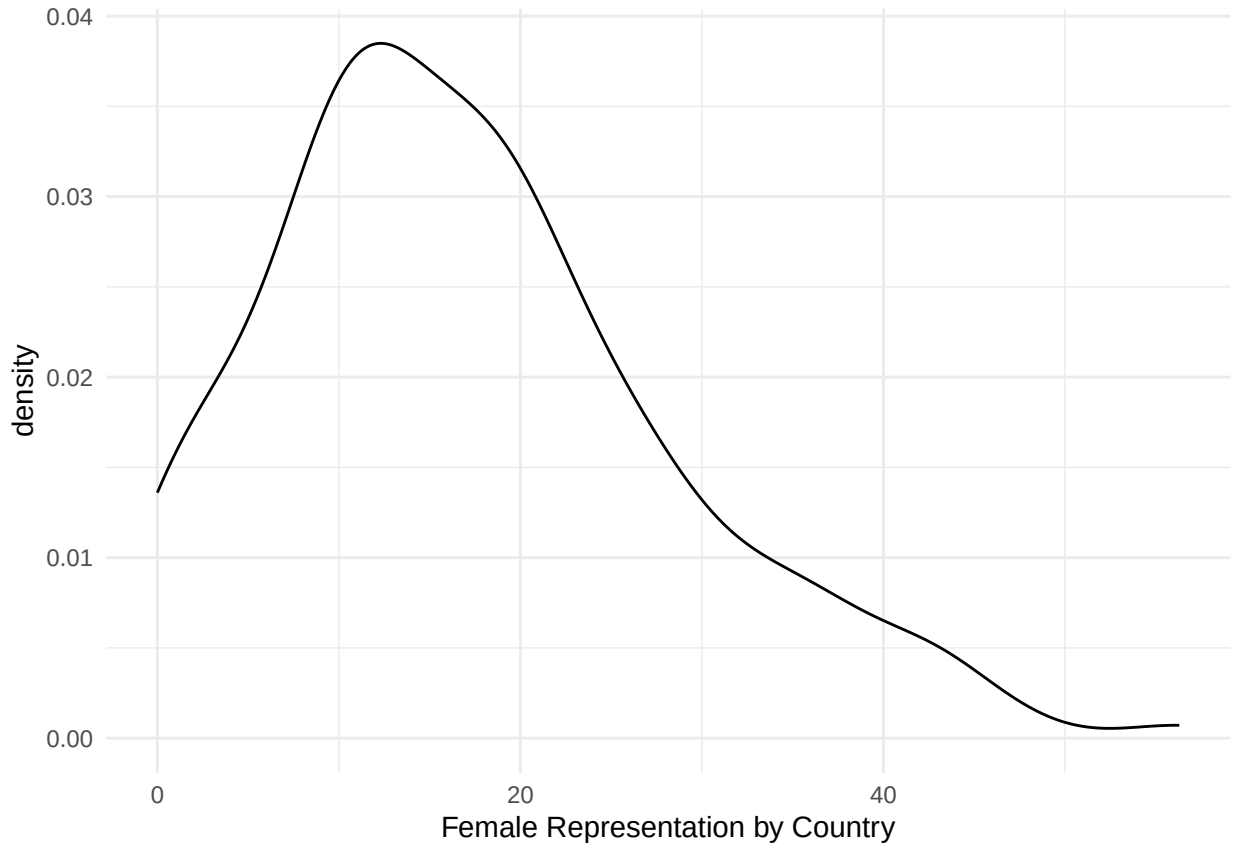
```
# I purposely didn't remove NA in this question since it's request in the next
# question
fem_rep <- world$women09
econ_devl <- world$gdp_10_thou

summary(fem_rep)
```

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

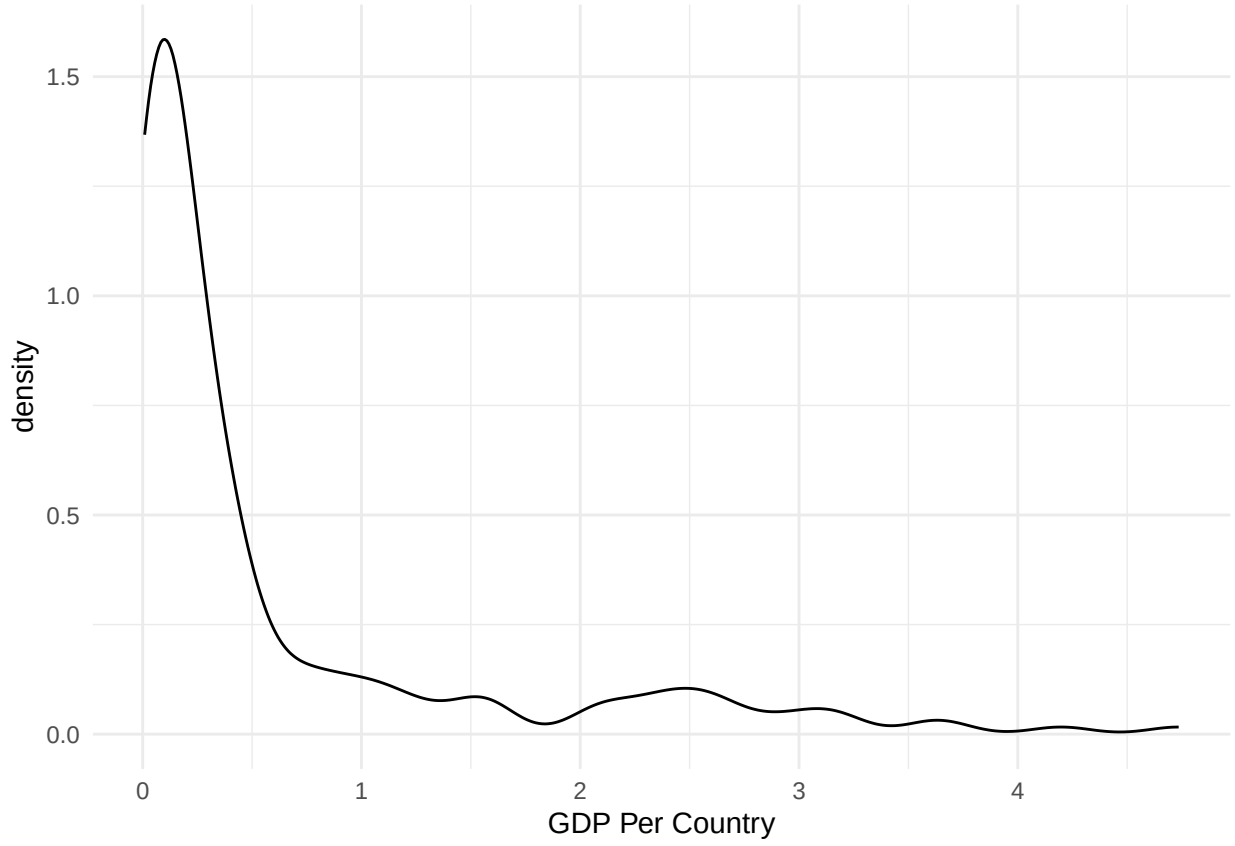
```
summary(econ_devl)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0090  0.0503  0.1897  0.6018  0.6320  4.7354      14
```

```
# Distribution Plots
fem_rep_plot <- ggplot (world, aes (x = world$women09)) + geom_density() +
  xlab ("Female Representation by Country") + theme_minimal()
fem_rep_plot
```




```
econ_devl_plot <- ggplot (world, aes (x = world$gdp_10_thou)) + geom_density() +
  xlab ("GDP Per Country") + theme_minimal()
econ_devl_plot
```




```
# Histograms. Not shown for brefity
fem_rep_plot_hist <- ggplot (world, aes (x = world$women09)) + geom_histogram() +
  xlab ("Female Representation by Country") + ylab ("Number of Countries") + theme_mini
fem_rep_plot_hist
econ_devl_plot_hist <- ggplot (world, aes (x = world$gdp_10_thou)) + geom_histogram() +
econ_devl_plot_hist
```

Question 2

Removing NA's

```
data <- data.frame (fem_rep = fem_rep, econ_devl = econ_devl)
data_no_NA <- na.omit (data)
length(data_no_NA$fem_rep)
```

```
## [1] 169
```

```r
# There are 169 observations left
```

Question 3

Correlation coefficient.

```r
cor_matrix <- as.matrix(data_no_NA)
rcorr(cor_matrix, type = "pearson")
```

```
##           fem_rep econ_devl
## fem_rep      1.00      0.31
## econ_devl    0.31      1.00
##
## n= 169
##
##
## P
##           fem_rep econ_devl
## fem_rep             0
## econ_devl  0
```

```r
# the relationship is positive and significant.
```

Question 4

Determining the null and alternative hypotheses

```r
# the null hypothesis is that economic growth has no effect on female
# representation
# The alternative hypothesis is that economic growth does have an
# effect on female representation. This can be either positive or negative.
```

Question 5

(a) Estimated Regression Equation of economic development on female representation

(b) What does the estimated regression equation look like?

```r
x <- lm(fem_rep ~ econ_devl, data = data_no_NA)
```

(b) What is the sign of the coefficient for X?

```r
summary(x)
```

```
##
## Call:
## lm(formula = fem_rep ~ econ_devl, data = data_no_NA)
##
## 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 ***
## econ_devl     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
```

```r
# The coefficient is positive
```

(c) Size of the coefficient.

```r
# the effect size is 3.457. Meaning that if GDP increases by 10,000, female
# representation is predicted to increase by 3.457%. If GDP increased by $1,000
# female representation is predicted to increase by .3457%
```

(d) Is the estimated coefficient statistically significant? At what confidence level?

```r
summary(x)
```

```
##
## Call:
## lm(formula = fem_rep ~ econ_devl, data = data_no_NA)
##
```

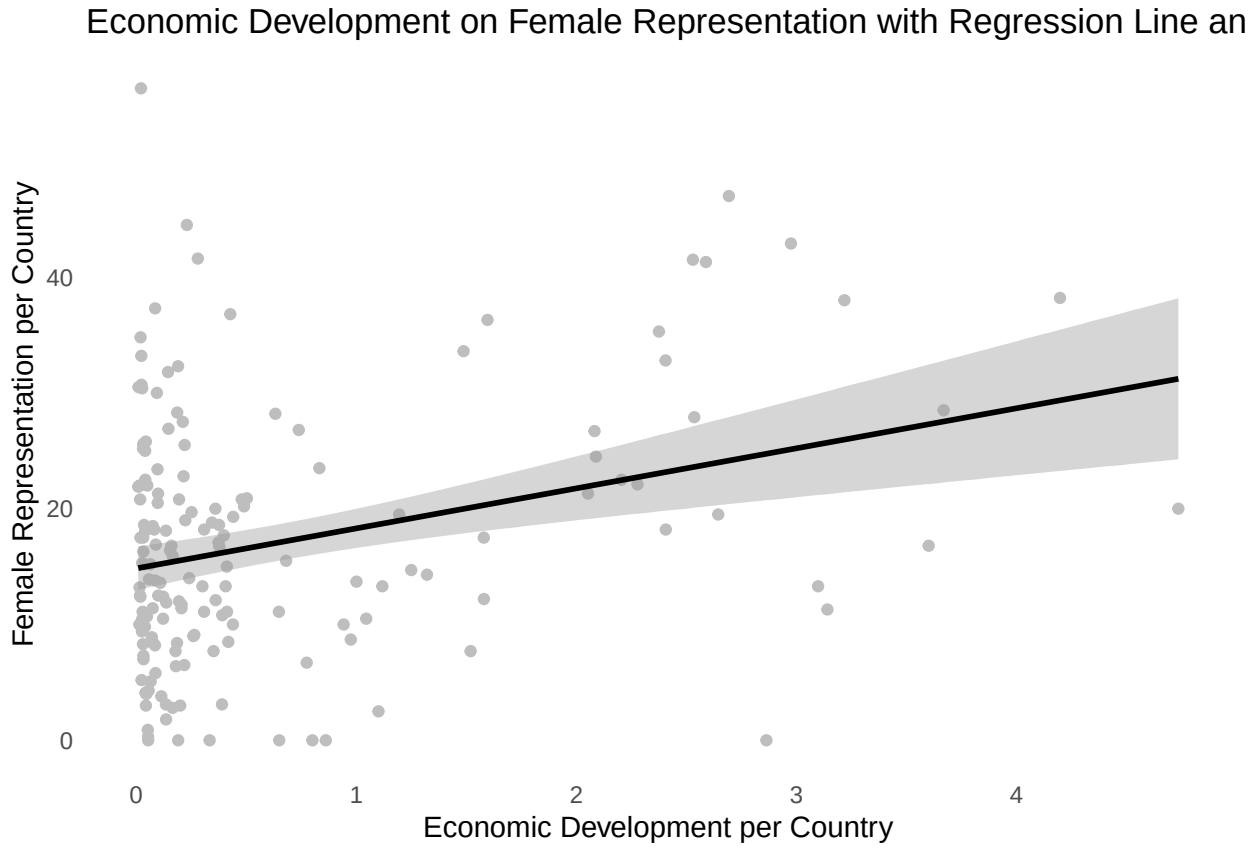
```
## Residuals:
##    Min     1Q Median     3Q    Max
## -24.74  -6.74  -1.62   5.78  41.38
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.8430     0.9542   15.56  < 2e-16 ***
## econ_devl     3.4574     0.8351    4.14  5.5e-05 ***
## ---
## Signif. codes:  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
```

```
# yes, it is significant at the 99.9% confidence interval
```

Question 6

Graphing X on Y with regression line and confidence interval

```
ggplot(data_no_NA, aes(x = econ_devl, y = fem_rep)) +
  geom_point(color = "grey", alpha = 1) +
  geom_smooth(method = "lm", se = TRUE, color = "black") +
  xlab ("Economic Development per Country") +
  ylab ("Female Representation per Country") +
  labs(title = "Economic Development on Female Representation with Regression Line and 9
  theme_minimal() +
  theme(panel.grid.major = element_blank(),
            panel.grid.minor = element_blank())
```




```
# alternatively, this method also works. Also not shown for brefity
pr <- ggpredict(x, c("econ_devl"))
plot(pr) + theme(legend.position = "bottom") + xlab("Civic Community") + ylab("Institut
```

Question 7

Goodness of fit

(a) How much of the variation in Female Representation is explained by Economic Development? Would you say this is big enough?

```
# about 9% of the variance in female representation is accounted for by economic
# development. I would say this is not big enough to explain female
# representation (found from the R^2)
```

(b) How far off are our predictions, on average? Would you say this is too big?

```
# The residual standard error is 10.38. Meaning our predictions are quite far
# from the actual observed data on female representation. Yes, I would say this
# is too far.
```

Question 8

Predicting values.

```
country <- world$country
pr_sys <- world$pr_sys
new_data <- data.frame (country = country,
                        fem_rep = fem_rep, econ_devl = econ_devl,
                        pr_sys = pr_sys)
new_data_no_na <- na.omit (new_data)

new_data_no_na $ y_hat <- predict (x)
subset(new_data_no_na, country == "Rwanda")$y_hat
```

```
## [1] 14.91632
```

```
subset(new_data_no_na, country == "Rwanda")$fem_rep
```

```
## [1] 56.3
```

```
# The prediction for Rawanda is very far off. The model predicts that female
# representation will be 14.9%, but the observed data is much higher with 56.3%
# female representation
```

Question 9

Separating countries by PR system.

```
# Regression for PR systems
pr_sys_data <- subset(new_data_no_na, pr_sys == "Yes")
pr_regress <- lm(fem_rep ~ econ_devl, data = pr_sys_data)

# Regression for non-PR systems
non_pr_sys_data <- subset(new_data_no_na, pr_sys == "No")
non_pr_regress <- lm(fem_rep ~ econ_devl, data = non_pr_sys_data)
```

Question 10

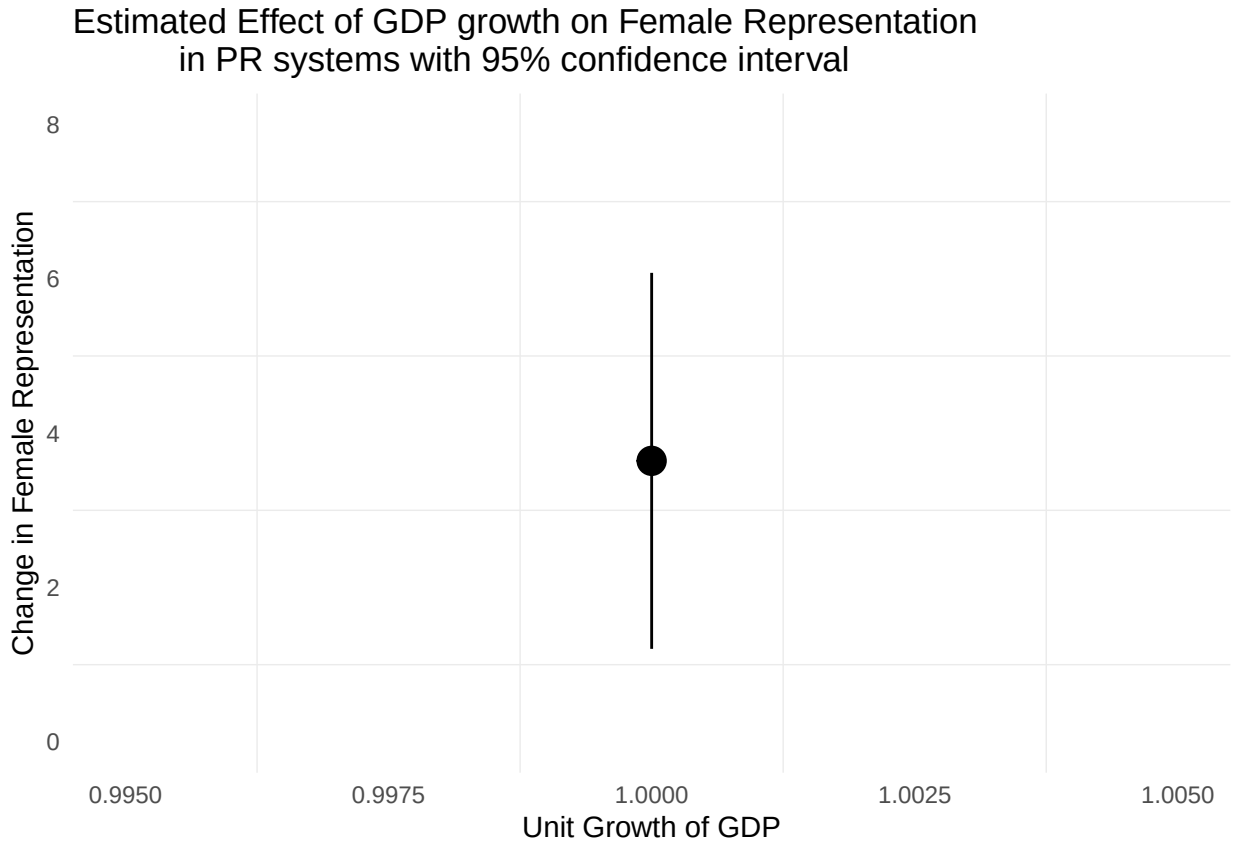
Estimated effect of economic development on female representation

```
#Yields the actual point estimate of beta
summary(pr_regress)
```

```
##
## Call:
## lm(formula = fem_rep ~ econ_devl, data = pr_sys_data)
##
## 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 ***
## econ_devl      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
```

```
# building components of pr graphs
beta_estimate <- 3.641
e <- 1.218 *2
lower_ci <- beta_estimate + e
upper_ci <- beta_estimate - e

# Create a data frame for ggplot
pr_data_point_graph <- data.frame(
  Estimate = beta_estimate,
  ymin = lower_ci,
  ymax = upper_ci)

# graph
point_est_plot <- ggplot(pr_data_point_graph, aes(x = 1, y = Estimate)) +
  geom_pointrange(aes(ymin = ymin, ymax = ymax), color = "black", size = 1) +
  theme_minimal() +theme(panel.grid.major = element_blank()) +
  ylim (0, 8) + xlim (.995, 1.005) +
  xlab ("Unit Growth of GDP") +
```

```
  ylab ("Change in Female Representation") +
  ggtitle ("Estimated Effect of GDP growth on Female Representation
           in PR systems with 95% confidence interval")
point_est_plot
```




```
# the estimated effect of economic development on female representation in
# PR systems is 3.641
```

Question 11

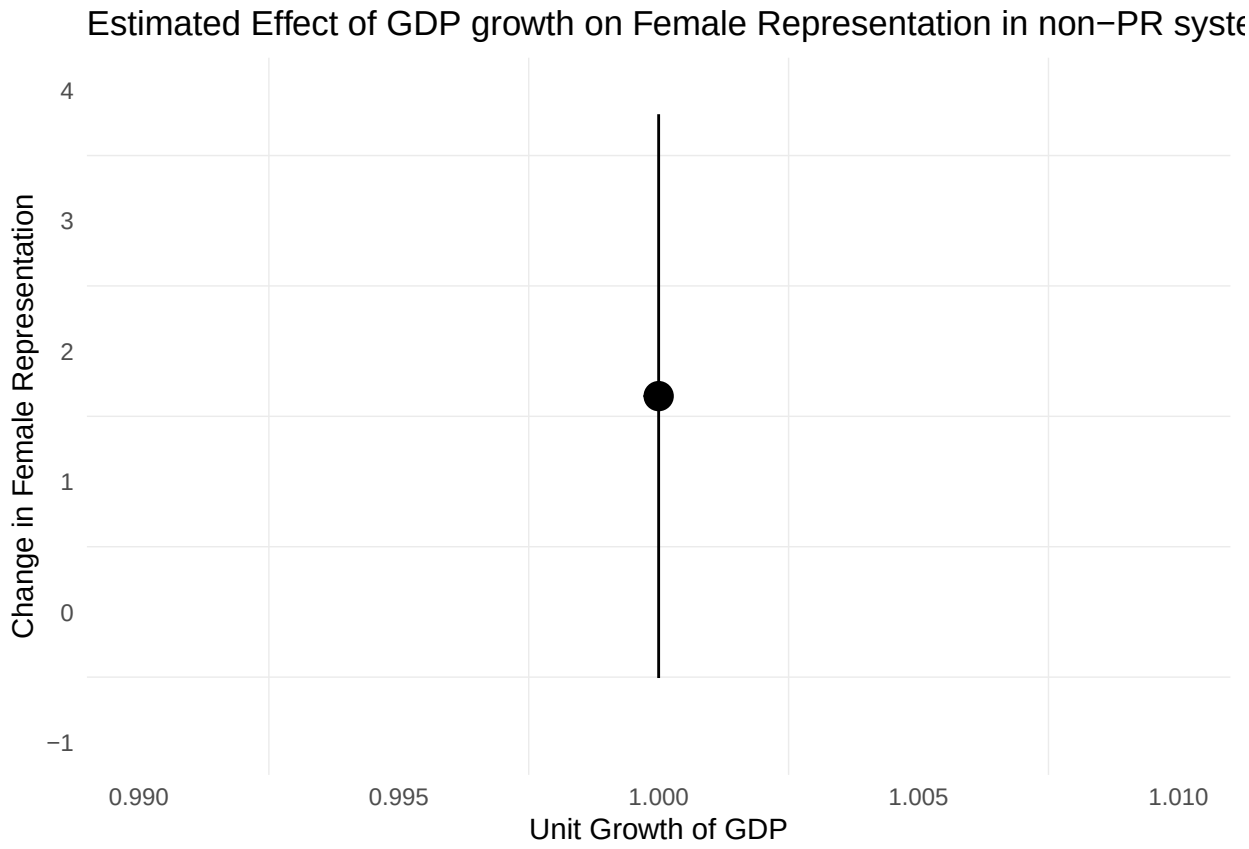
Estimated effect of economic development in non-PR countries

```
#Yields the actual point estimate of beta
summary(non_pr_regress)
```

```
##
## Call:
## lm(formula = fem_rep ~ econ_devl, data = non_pr_sys_data)
```

```
##
## 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 ***
## econ_devl      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
```

```r
# building components of non_pr graphs
beta_estimate_non_pr <- 1.655
e_non_pr <- 1.081 *2
lower_ci_non_pr <- beta_estimate_non_pr + e_non_pr
upper_ci_non_pr <- beta_estimate_non_pr - e_non_pr

# Create a data frame for ggplot
non_pr_data_point_graph <- data.frame(
  Estimate = beta_estimate_non_pr,
  ymin = lower_ci_non_pr,
  ymax = upper_ci_non_pr)

# graph
point_est_plot_non_pr <- ggplot(non_pr_data_point_graph,
                                aes(x = 1, y = Estimate)) +
  geom_pointrange(aes(ymin = ymin, ymax = ymax), color = "black", size = 1) +
  theme_minimal() +theme(panel.grid.major = element_blank()) +
  ylim (-1, 4) + xlim (.99, 1.01) +
  xlab ("Unit Growth of GDP") +
  ylab ("Change in Female Representation") +
  ggtitle ("Estimated Effect of GDP growth on Female Representation in non-PR systems wi
point_est_plot_non_pr
```




```
# the estimated effect of economic growth on female representation in non pr
# systems is 1.655
```

Question 12

Comparing predicted and actual value of female representation for Rwanda

```
pr_sys_data $ y_hat <- predict (pr_regress)
subset (pr_sys_data, country == "Rwanda")$y_hat
```

```
## [1] 19.45267
```

```
subset (pr_sys_data, country == "Rwanda")$fem_rep
```

```
## [1] 56.3
```

```
# The prediction is a little closer but still quite off. The predicted value is
# 19.45 while the actual value is still 56.3
```

End of file