# Week 7 Exercise

Calum Webb

27/10/2022

## United Nations Human Development Program: Gender Inequality and Societal Gender Bias

This week's exercise uses data from the UNHDP's Gender Inequality Index and Gender Social Norms Index for 'highly developed' countries. This includes the following variables:

### Gender Inequality Index Variables (GII)

- `parl_seats_women`: Percentage of parliamentary seats held by women
- `sec_ed_women`: Percentage of women with at least some secondary school education
- `lab_force_women`: Percentage of women participating in the labour force

### Gender Socian Norms Index

- `gsni_1_bias`: Percentage of the population holding at least one social bias against women.
- `gsni_2_bias`: Percentage of the population holding at least two social biases against women.
- `gsni_no_bias`: Percentage of the population holding no social biases against women.
- `political_bias`: Percentage of the population who think that men make better political leaders than women, or who disagree that women have the same rights as men.
- `economic_bias`: Percentage of the population who agree that university is more important for a man than for a woman.
- `educational_bias`: Percentage of the population who believe that men should have more right to a job than women or that men make better business executives than women.
- `physical_bias`: Percentage of the population who believe that intimate partner physical violence against women or the restriction of women's reproductive rights is justified.

There is also a variable for whether the country is part of the European Union (`in_eu`).

## Part I: Interpreting regression output and regression lines

The research team started by exploring the relationship between women's labour force participation and representation in parliamentary democracies.

They began by reading the data into `R` and looking at its structure:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
un_gender_data <- read_csv("un_gii_gsni.csv")
```

```
## Rows: 62 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): country
## dbl (11): parl_seats_women, sec_ed_women, lab_force_women, in_eu, gsni_1_bia...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
un_gender_data
```

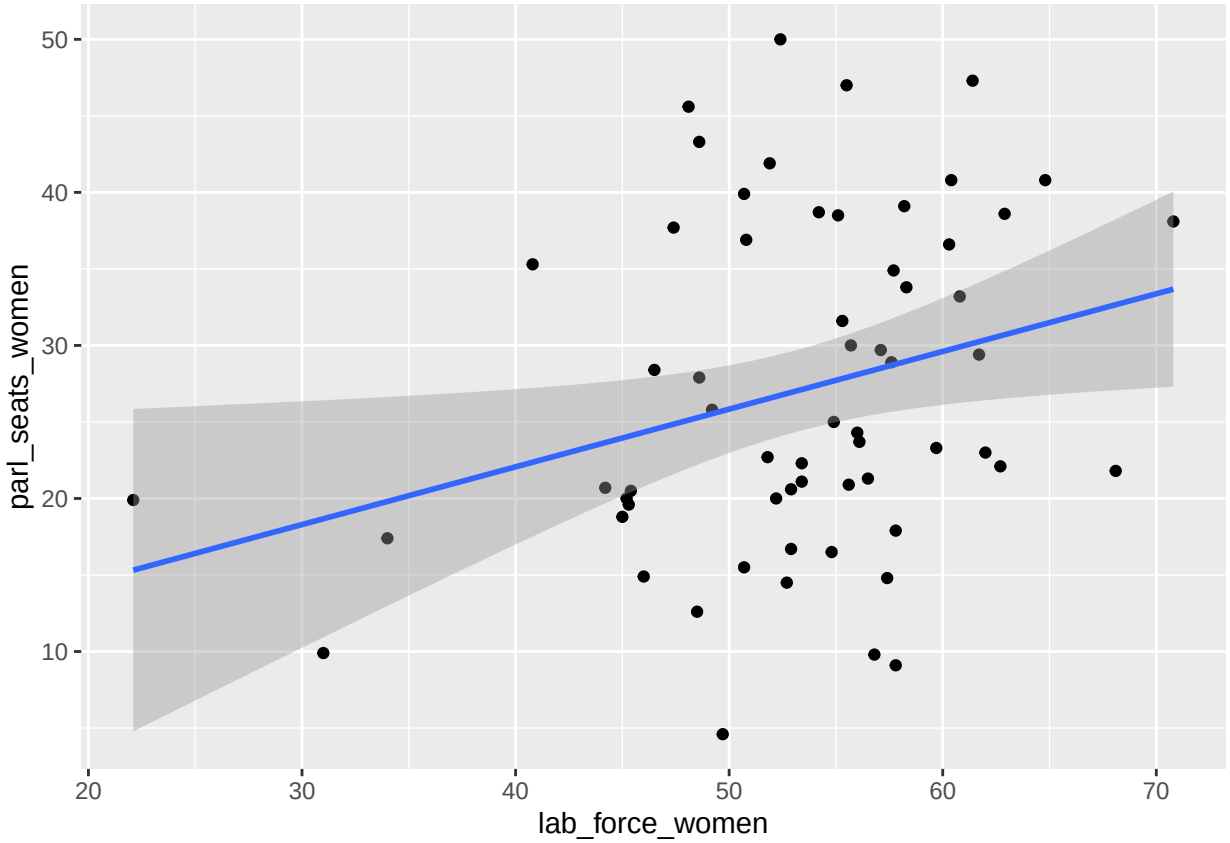
```
## # A tibble: 62 x 12
##    country parl_~1 sec_e~2 lab_f~3 in_eu gsni_~4 gsni_~5 gsni_~6 polit~7 econo~8
##    <chr>     <dbl>   <dbl>   <dbl> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 Norway     40.8    95.4    60.4     0    41.3    16      58.7    19.5   21.8
##  2 Ireland    24.3    81.9    56       1    NA      NA      NA      NA     NA
##  3 Switze~    38.6    95.6    62.9     0    56.0    26.9    44.0    20.6   29.8
##  4 Iceland    38.1   100      70.8     0    NA      NA      NA      NA     NA
##  5 Germany    31.6    95.9    55.3     1    62.6    33.1    37.4    26.6   30.9
##  6 Sweden     47.3    89.3    61.4     1    30.0    10.8    70.0    16.0    9.16
##  7 Austra~    36.6    91      60.3     0    46.2    23      53.8    32.5   18.1
##  8 Nether~    33.8    87.6    58.3     1    39.8    15.9    60.2    21.3   13.6
##  9 Denmark    39.1    91.2    58.2     1    NA      NA      NA      NA     NA
## 10 Finland    47     100      55.5     1    51.2    22.7    48.8    24.6   23.1
## # ... with 52 more rows, 2 more variables: educational_bias <dbl>,
## #   physical_bias <dbl>, and abbreviated variable names 1: parl_seats_women,
## #   2: sec_ed_women, 3: lab_force_women, 4: gsni_1_bias, 5: gsni_2_bias,
## #   6: gsni_no_bias, 7: political_bias, 8: economic_bias
```

---

They then created a scatterplot between the indepedent variable (women's labour force participation %) and the dependent variable (women's representation in parliamentary democracies %).

```
un_gender_data %>%
  ggplot() +
  geom_point(aes(x = lab_force_women, y = parl_seats_women)) +
  geom_smooth(aes(x = lab_force_women, y = parl_seats_women), method = "lm")
```

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




- Describe the relationship between the two variables that you see here.

There appears to be a positive linear relationship between the two variables, but it could be quite weak given the difference between the points and the line.

- Which assumptions might be violated in linear regression if we used this data as is? (There can be more than one)

There are three points in the bottom left quadrant of the plot that could be considered outliers or leverage points; the line is likely to be much flatter if these were removed. If these were kept, there could be some issues with heteroscedasticity, as the variance of the points around the line seems to increase at higher values of X.

- What might you suggest the researchers do to improve how well the data meets the assumptions of linear regression?

The researchers may wish to consider removing the three points that could be considered outliers or leverage points due to their strong influence. This would also correct the issue of heteroscedasticity. However, they should also consider whether there are reasons (such as some kind of grouping) that might explain why these points are so far away from many of the other points.

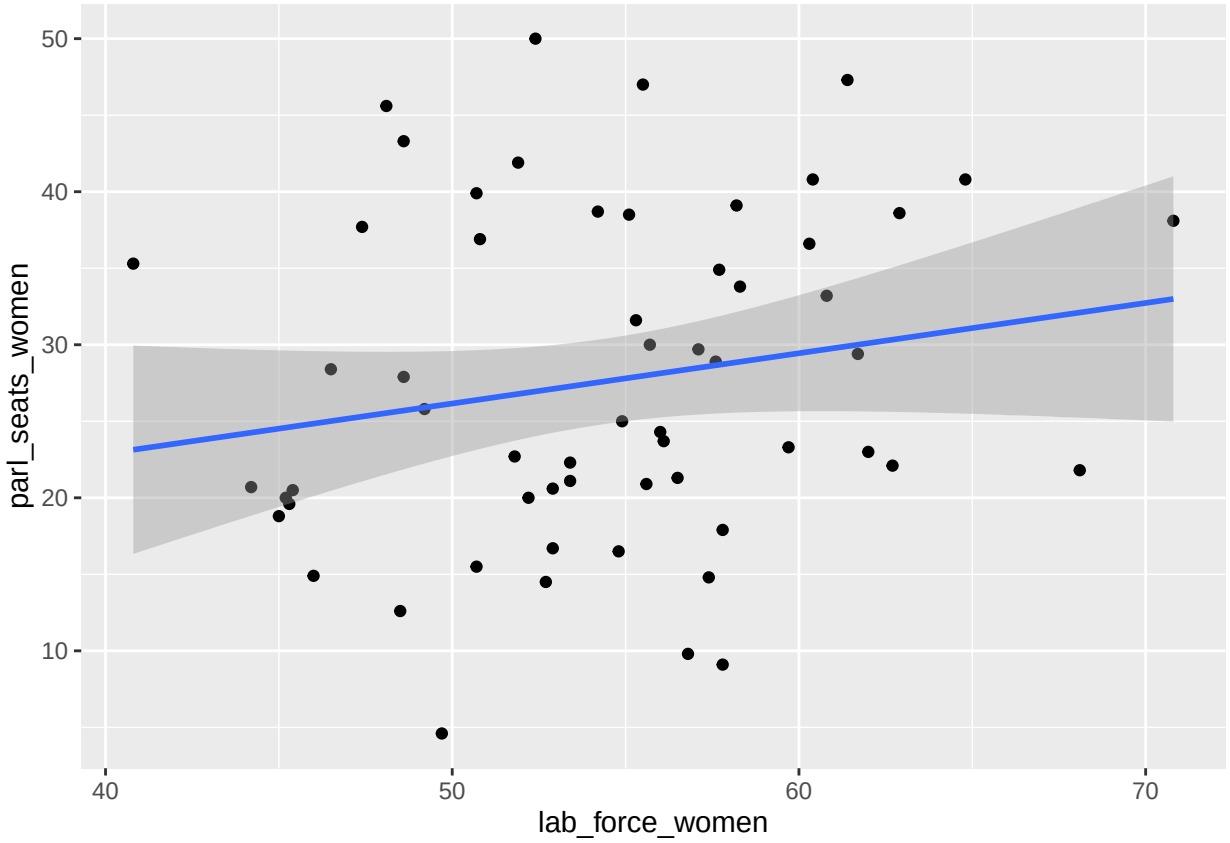
---

The researchers decide that the three data points with fewer than 35% of women in the labour market could be considered outliers, and choose to remove them before estimating their linear regression model. They save the data without the outliers as `un_gender_data_2`. They then re-run their scatterplot to see how the linear association has changed.

```
un_gender_data_2 <- un_gender_data %>%
                      filter(lab_force_women > 35)
```

```
un_gender_data_2 %>%
  ggplot() +
  geom_point(aes(x = lab_force_women, y = parl_seats_women)) +
  geom_smooth(aes(x = lab_force_women, y = parl_seats_women), method = "lm")
```

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




- How has the regression line changed?

The regression line has become flatter (the slope has decreased and intercept has decreased).

- Are there any other features in this data that you think might violate the assumptions of linear regression?

Linearity: It looks okay, a curved line doesn't look like it would fit the data much better than the straight line. Homoscedasticity: This assumption is probably met, though there might be some increased variance in the residuals closer to the middle. Outliers & leverage points: None of these remaining data points now look like they would be considered outliers or leverage points. Normality of residuals: The residuals appear to be normally distributed (the average distance and spread above the line looks similar to the average distance and spread below the line)

---

The researchers now decide that they can estimate their linear regression model using this processed data. They write the code to do so below:

```
model_1 <- lm(data = un_gender_data_2, formula = parl_seats_women ~ lab_force_women)

summary(model_1)
```

```
##
## Call:
## lm(formula = parl_seats_women ~ lab_force_women, data = un_gender_data_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.462  -7.048  -2.770   8.991  23.051
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9.7371    12.5005   0.779    0.439
## lab_force_women   0.3285     0.2285   1.438    0.156
##
## Residual standard error: 10.69 on 57 degrees of freedom
## Multiple R-squared:  0.035,  Adjusted R-squared:  0.01807
## F-statistic: 2.067 on 1 and 57 DF,  p-value: 0.156
```

- How large was the R-squared value and what does this mean?

The R-squared value was 0.035. This means that only 3.5% of the variance in parliamentary seats held by women could be explained by the percentage of women in the labour force in each country. This is not likely to be a strong predictor.

- Was the relationship between labour force participation and the percentage of parliamentary seats held by women statistically significant?

The p-value for the slope of the percentage of women in the labour force was 0.156; this is higher than our typical critical value of 0.05. This means that, if the data is a random sample from a wider population, the estimate from our observation is within the range of what we would expect if the real relationship between percentage of women in the labour force and the percentage of parliamentary seats held by women was 0 in the entire population. In other words, the result was not statistically significant.

**However**, it would be sensible to think about whether the assumptions for hypothesis testing are met here. Is this a random sample of countries or is it non-random in some way? Or is the sample a complete reflection of the population of interest? In either of the latter two cases it would not be appropriate to use statistical significance in this way. Because the countries do not appear to have been sampled randomly they may refer to the entire population of interest (OECD countries), in which case we should not use a hypothesis test and can interpret the effect size to be the effect in the population, or they may be an opportunity sample of countries (e.g. just those countries that had data available), in which case we should be very cautious about generalising outside of this specific sample.

- Write the regression equation for this model, in the form Y = B_0 + B_1*X

parl_seats_women = 9.737 + 0.329 * lab_force_women

- Finish the following interpretation of the regression slope

For every 1 percentage point increase in women's labour force participation there was a mean change in the percentage of parliamentary seats held by women of approximately 0.33 percentage points.

- How might the researchers conclude this part of their study?

The researchers might conclude that while there was only a slight association between the two variables: women's labour market participation was not a strong predictor of women's representation in parliament, at a country-wide level. This might be because this measure does not capture the type of labour market participation in each country, or account for the fact that more wealthy countries may have lower labour market participation in general. Additionally, countries with limited employment opportunities may have lower participation in general. They might also mention that because the sample is unlikely to be a random sample of a larger population the significance of the relationship cannot be relied upon.

- Now, write the code below to rerun this analysis using the original data (with the outliers).

```
model_2 <- lm(data = un_gender_data, formula = parl_seats_women ~ lab_force_women)

summary(model_2)
```

```
##
## Call:
## lm(formula = parl_seats_women ~ lab_force_women, data = un_gender_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.123  -7.032  -2.544   7.646  23.259
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       6.9863     8.8215   0.792   0.4315
## lab_force_women   0.3770     0.1641   2.298   0.0251 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.5 on 60 degrees of freedom
## Multiple R-squared:  0.08089,    Adjusted R-squared:  0.06557
## F-statistic:  5.28 on 1 and 60 DF,  p-value: 0.02507
```

- How does the inclusion of the outliers/leverage points change the regression estimates and the research findings?

The estimated slope for the percentage of women in the labour force increases from 0.33 to 0.38 with the inclusion of the outliers. The association between the two variables also becomes statistically significant ($p<0.05$).

While the estimates are somewhat similar (a 1 percentage point increase in the proportion of women in the labour force being associated with a 0.33 or 0.38 percentage point increase in the percentage of parliamentary seats held by women), the fact that the results are now "statistically significant" might substantially change the conclusions – this is quite concerning when the use of statistical significance might not actually be appropriate in this case!

**NOTE:** Notice that throughout this section I have been using **percentage point** rather than percentage. This is because the scale both variables are on is a percentage scale (0-100), and because if I were to use **percentage** not **percentage point** there might be some ambiguity in whether I am referring to an absolute or a relative increase. For example:

A 10 **percent** increase from a value of 25 percentage points would be 27.5 (25%p + 10% of 25 = 27.5).

A 10 **percentage point** increase from a value of 25 percentage points would be 35 percentage points (25%p + 10%p = 35%).

It's not always possible to avoid making this mistake, but try to do so. Also remember that **the Y and X variables are always interpreted on the scale they are on unless they have been transformed in some way**. If, instead, our Y variable was parliamentary seats out of every 100,000 women in the country, we would instead be talking about increases of B1 more parliamentary seats per 100,000 women in the country.

---

## Part II: Exploring the association between societal gender biases and women's political representation

Your task is to explore the other part of the researcher's questions:

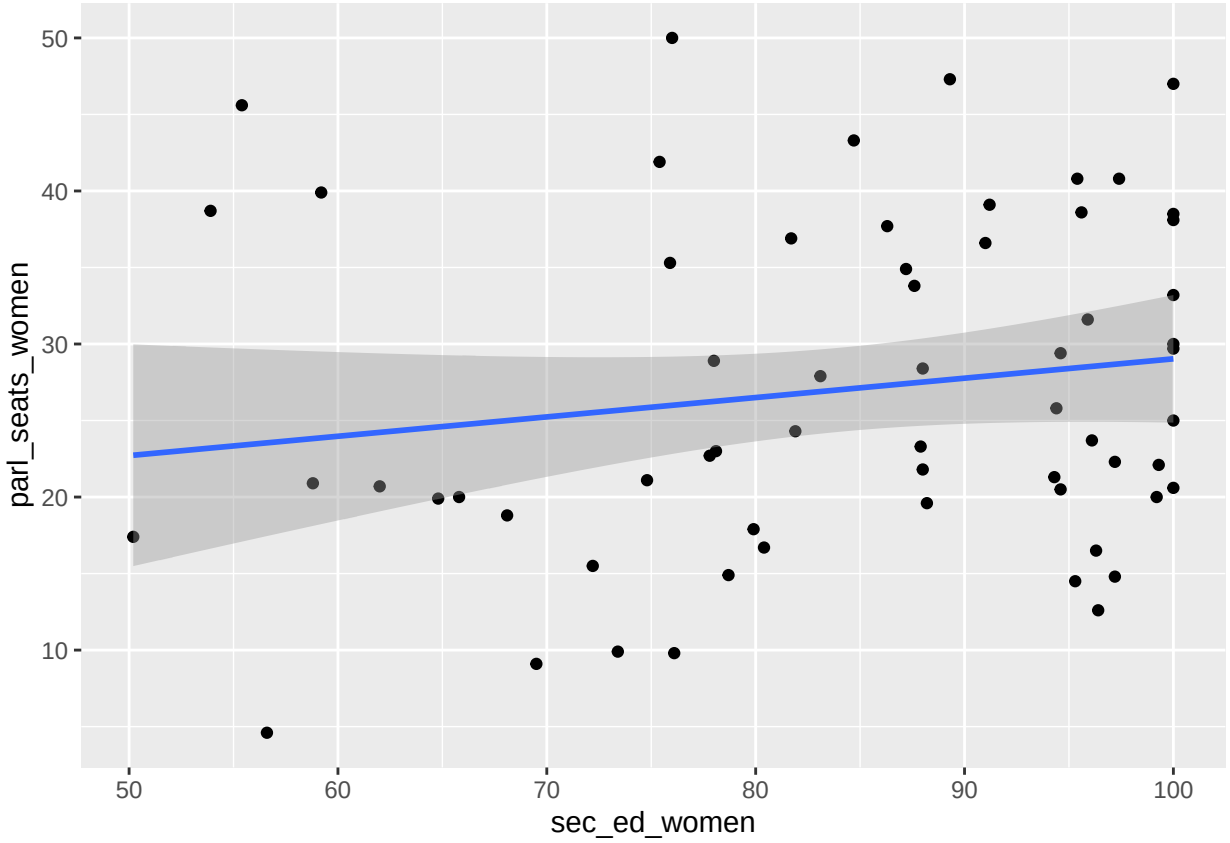
**The association between secondary school education rates for women and women's parliamentary representation**

---

- Create a scatterplot with a regression line to explore the above association.

```
un_gender_data %>%
  ggplot() +
  geom_point(aes(x = sec_ed_women, y = parl_seats_women)) +
  geom_smooth(aes(x = sec_ed_women, y = parl_seats_women), method = "lm")
```

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




- Does it look like there might be any association between the two variables from this visualisation?

No, the regression line is only slightly positive.

- Does it look like any of the assumptions of linear regression are broken by this data?

Generally, it looks like none of the assumptions of linear regression are violated here. There may be some non-normality of residuals or heteroscedasticity at higher values of the independent variable.

- Write the code to make any changes to the data (if you think they are necessary), and then estimate a linear regression model and check the results.

```r
model_3 <- lm(data = un_gender_data, formula = parl_seats_women ~ sec_ed_women)

summary(model_3)
```

```
##
## Call:
## lm(formula = parl_seats_women ~ sec_ed_women, data = un_gender_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.942  -7.638  -3.091   9.255  24.005
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.38336    8.42053   1.946   0.0564 .
## sec_ed_women  0.12647    0.09875   1.281   0.2052
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.81 on 60 degrees of freedom
## Multiple R-squared:  0.02661,    Adjusted R-squared:  0.01039
## F-statistic:  1.64 on 1 and 60 DF,  p-value: 0.2052
```

- Interpret the output from the regression model — start with the model general statements and then move down to more specific ones.

The model has an R-squared value of 0.0266, meaning only around 2.7% of the variance in the percentage of parliamentary seats held by women could be explained by the percentage of women with at least some secondary school education. The relationship between the two variables was not statistically significant (p = 0.205). The regression coefficient was positive, and every 1 unit increase in the percentage of women with secondary education was associated with an average 0.126 unit increase in the percentage of parliamentary seats that were held by women.

---

**The association between societal gender bias and women's parliamentary representation.** Your next task is to explore whether there is a linear association between women's political representation and one form of societal gender bias in the countries in the sample.

- Pick which form of societal gender bias (use the descriptions at the start of the document) and justify (write a rationale for) why you have chosen to look at this form of gender bias below:

Political gender bias, the proportion of the population who believe that men make better leaders than women or that women are not entitled to the same rights as men, may be a more important contributor to the proportion of parliamentary seats held by women than either the female population's educational opportunities or their representation in the labour market. Strong bias against women is likely to impede their chances of election in democratic societies. Further, it is important to understand how biased gender norms in society create barriers to women's political representation rather than placing the emphasis solely on women's capacity or personal qualities (e.g. actively in work or in education).

If you are in class, or working with someone else, ask them to pick a different form of societal gender bias to explore and compare the model estimates.

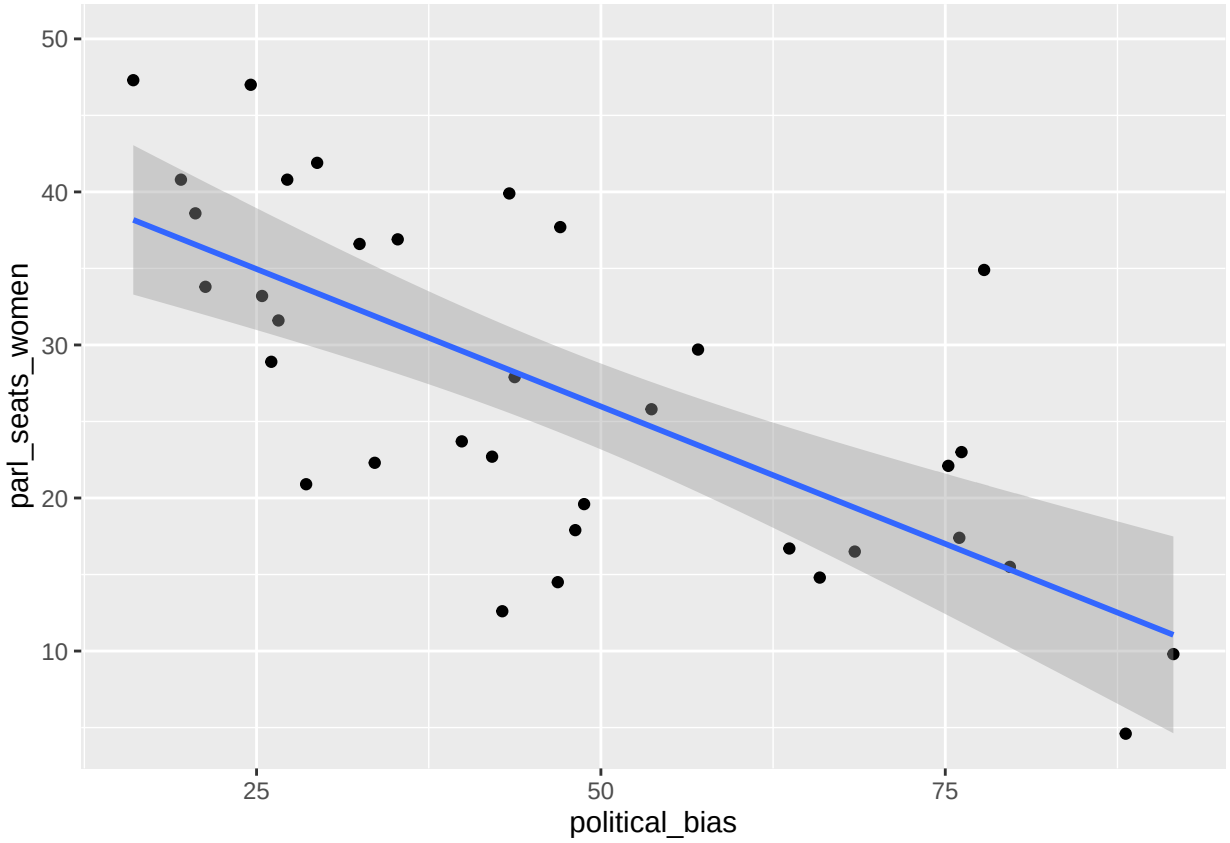
- Start by creating a scatterplot between the two variables. Include a linear regression line using the `geom_smooth()` function, with the `method = "lm"` argument, like above.

```
un_gender_data %>%
  ggplot() +
  geom_point(aes(x = political_bias, y = parl_seats_women)) +
  geom_smooth(aes(x = political_bias, y = parl_seats_women), method = "lm")
```

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

```
## Warning: Removed 27 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 27 rows containing missing values (geom_point).
```




- What does the relationship between the variables look like (strong? weak? positive? negative?)

There appears to be a fairly strong negative association between political bias against women in society and the percentage of parliamentary seats held by women - the higher the political bias against women in the country, the lower the percentage of parliamentary seats held by women, on average.

- Do the patterns in the data around the regression line look like they violates any of the assumptions of linear regression?

It is difficult to tell due to the small number of data points - there are no serious outliers, though the data point at X = ~76 and Y = ~35 deviates somewhat from the general trend. There may be some heteroscedasticity, residuals seem to be slightly more spread out in the middle of the regression line compared to the ends. Similarly, at early values of X the residuals seem to skew slightly towards the positive (above the line), rather tham below the line. This might indicate, in relation to linearity, that a slightly curved line would fit the data marginally better than a simple linear regression. Overall, linear regression appears to be appropriate for the data.

- Now estimate a linear regression model for your chosen variables:

```r
# This time, I am going to standardise the predictor using the scale() function
# This changes the scale of the predictor to z-scores, where now instead of a 1
# unit increase in X meaning a one percentage point increase, it will instead mean
# a one *standard deviation* increase.
# Scaling the predictor variables in this way can be useful for comparing the
# relative strength of different (continuous, approximately normally distributed)
# predictors (independent variables). It also makes it so that the mean of the predictors
# are all 0 - this changes the meaning of the intercept to be "the predicted value of Y
# at the mean value of X)
# Note: if you also standardised the outcome variable (dependent), then the slope coefficient (Estimate
# would be interpreted in the same way as a Pearson's R correlation.

model_4 <- lm(data = un_gender_data, formula = parl_seats_women ~ scale(political_bias))
summary(model_4)
```

```
##
## Call:
## lm(formula = parl_seats_women ~ scale(political_bias), data = un_gender_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.9528  -5.7917  -0.3296   5.9358  18.9084
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             27.083      1.366  19.830  < 2e-16 ***
## scale(political_bias)   -7.792      1.386  -5.623 2.93e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.08 on 33 degrees of freedom
##   (27 observations deleted due to missingness)
## Multiple R-squared:  0.4893, Adjusted R-squared:  0.4738
## F-statistic: 31.62 on 1 and 33 DF,  p-value: 2.931e-06
```

- What was the R-squared value for this model and what does this indicate?

The R-squared value for this model was 0.489, which indicates that the variance in values of political bias against women were able to explain 48.9% of the variance in parliamentary seats held by women.

- Was the relationship between the variables statistically significant at a critical value of 0.05 or not?

The association between political bias against women and parliamentary representation of women was statistically significant ($p < 0.05$), however, given that these data are not a random sample of countries it may be inappropriate to use a hypothesis test.

- Was the linear regression coefficient for the variable you chose (the slope) in a positive or negative direction? What does this mean?

As societal political bias against women increases, parliamentary representation of women tends to decrease.

- Describe the relationship between your independent variable and the dependent variable (parliamentary seats) using the estimate (e.g. for a 1 per cent increase in...)

A 1 standard deviation increase in the percentage of people holding gender biased political views was associated with a 7.79 percentage point decrease in the percentage of parliamentary seats held by women, on average (95% Confidence Interval: -10.509, -5.075).

Notice here that I included a 95% confidence interval. To calculate this, I took the estimate (-7.792) and then subtracted 1.96 times the standard error, for the lower bound, and added 1.96 times the standard error for the upper bound:

```
-7.792 - 1.96*1.386
```

```
## [1] -10.50856
```

```
-7.792 + 1.96*1.386
```

```
## [1] -5.07544
```

This works because we assume that our estimate has some uncertainty that is proportional to our sample size, and that this error is normally distributed. Therefore, we can get a 95% confidence interval by taking our estimate as a mean and using the standard error (Std. Error) as our standard deviation, and then applying the Empirical Rule (95% of observed values within ±1.96 standard deviations).

- Write the model intercept, slope, and variables out in the form Y = B_0 + B_1*X

parl_seats_women = 27.083 + -7.792 * (political_bias - mean(political bias)) / sd(political bias)

Note that I've included the long formula for scaling a variable here, you can compare this to see that it gives the same results:

```
tibble(
  scale(un_gender_data$political_bias), # results using scale
  (un_gender_data$political_bias - mean(un_gender_data$political_bias, na.rm = TRUE)) / sd(un_gender_da
)
```

```
## # A tibble: 62 x 2
##    `scale(un_gender_data$political_bias)`[,1] ... / sd(un_gender_data$politica~1
##                                         <dbl>                               <dbl>
##  1                                     -1.26                               -1.26
##  2                                     NA                                  NA
##  3                                     -1.22                               -1.22
##  4                                     NA                                  NA
##  5                                     -0.938                              -0.938
##  6                                     -1.42                               -1.42
##  7                                     -0.666                              -0.666
##  8                                     -1.18                               -1.18
##  9                                     NA                                  NA
## 10                                     -1.03                               -1.03
## # ... with 52 more rows, and abbreviated variable name
## #   1: `... / sd(un_gender_data$political_bias, na.rm = TRUE)`
```

- Compare your results with someone else who looked at a different form of societal gender bias and its association with women's parliamentary representation. Is one a stronger predictor than the other? Describe the differences.

---

## Week 7 Challenge

Well done — hopefully the above exercises will have given you some good practice interpreting linear regression models and we will build on this further in Week 8. Interpretation of these models gets easier with practice.

You can choose to instead focus on what you've learned here on your assessment 1 work, if you wish. But if you would like some additional challenges, here are some suggestions:

- Using the inclusion of binary categorical variables in regression (something we will cover in more detail in Week 8), design a model to test whether societal gender bias is higher or lower in countries that are part of the European Union.
- Return to some of the data that we used for running ANOVA tests or for Pearson's correlations (R) from Week 4 and 5; try creating a linear regression model equivalent of these tests and make notes on how (1) the interpretation differs — are slope coefficients more informative? — (2) how the p-values change (or don't change) and (3) whether any of the test statistics that are created are the same. This will give you some good practice moving the data into the directory you are working in, reading in the data, and making any transformations if you need to (e.g. creating new binary categorical variables from regular categorical variables using the `ifelse()`, `case_when()`, or the `to_dummy()` function from the `sjmisc` package.