Week 6 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:

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## 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 become errors
```

```r
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.
```

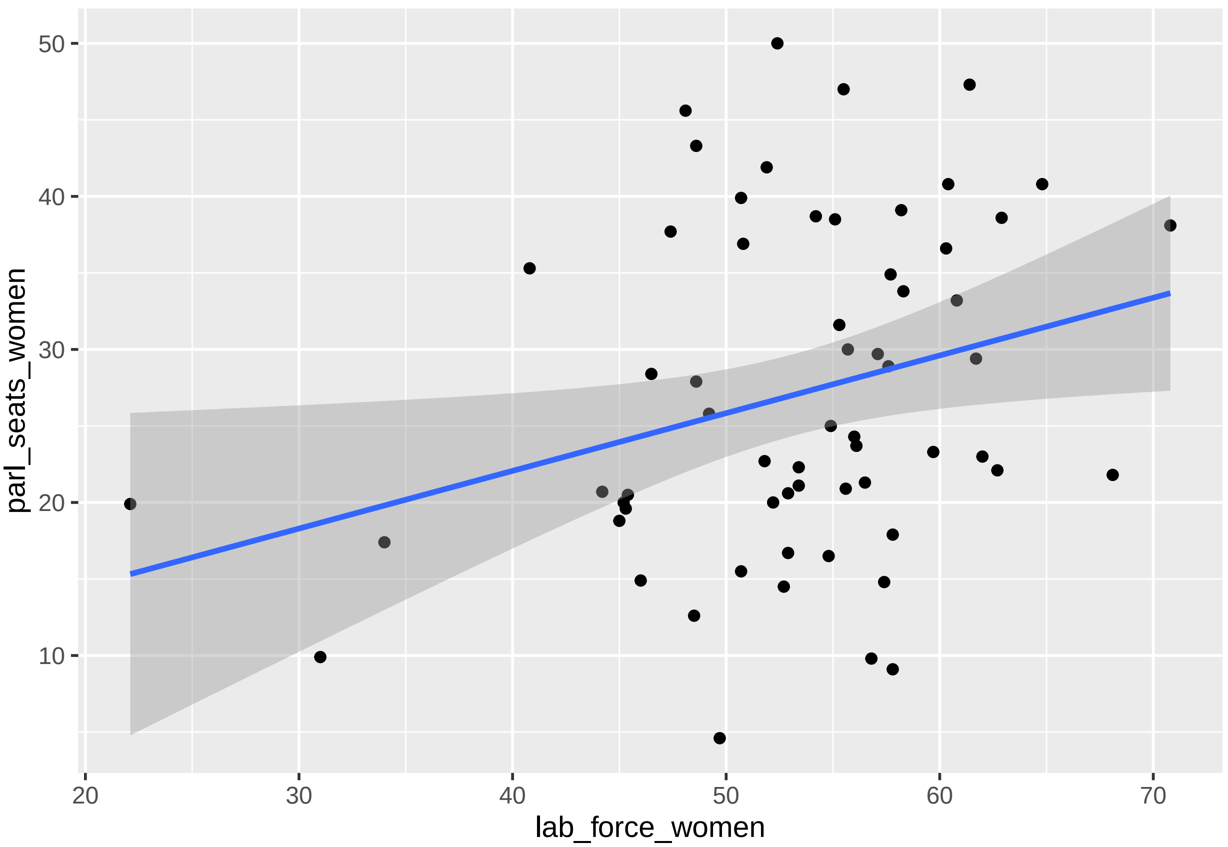
```r
un_gender_data
```

```
## # A tibble: 62 x 12
##    country     parl_seats_women sec_ed_women lab_force_women in_eu gsni_1_bias
##    <chr>                  <dbl>        <dbl>           <dbl> <dbl>       <dbl>
##  1 Norway                  40.8         95.4            60.4     0        41.3
##  2 Ireland                 24.3         81.9            56       1        NA
##  3 Switzerland             38.6         95.6            62.9     0        56.0
##  4 Iceland                 38.1        100              70.8     0        NA
##  5 Germany                 31.6         95.9            55.3     1        62.6
##  6 Sweden                  47.3         89.3            61.4     1        30.0
##  7 Australia               36.6         91              60.3     0        46.2
##  8 Netherlands             33.8         87.6            58.3     1        39.8
##  9 Denmark                 39.1         91.2            58.2     1        NA
## 10 Finland                 47          100              55.5     1        51.2
## # i 52 more rows
## # i 6 more variables: gsni_2_bias <dbl>, gsni_no_bias <dbl>,
## #   political_bias <dbl>, economic_bias <dbl>, educational_bias <dbl>,
## #   physical_bias <dbl>
```

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

```r
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'
```



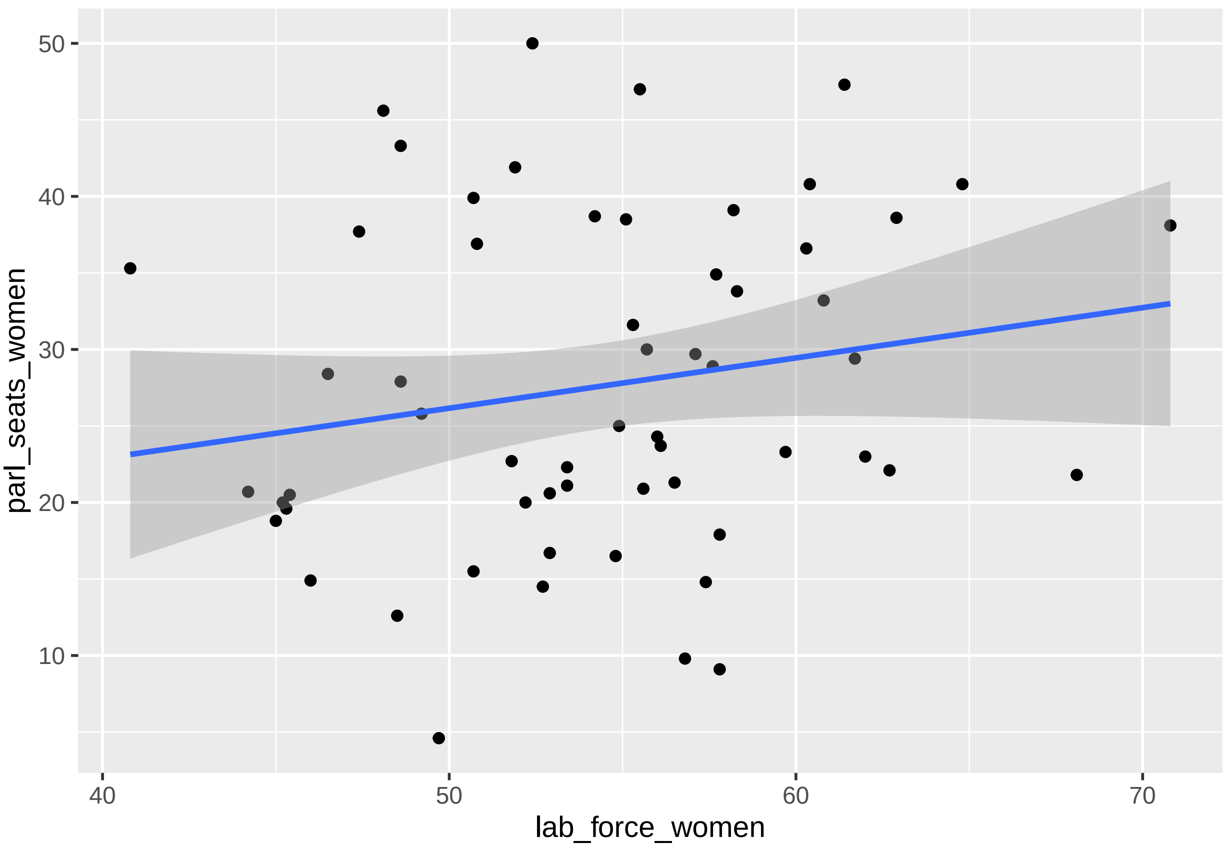
- Describe the relationship between the two variables that you see here.
- Which assumptions might be violated in linear regression if we used this data as is? (There can be more than one)
- What might you suggest the researchers do to improve how well the data meets the assumptions of linear regression?

The researchers decide that the three data points with less 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?
- Are there any other features in this data that you think might violate the assumptions of linear regression?

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?
- Was the relationship between labour force participation and the percentage of parliamentary seats held by women statistically significant?
- Write the regression equation for this model, in the form Y = B_0 + B_1*X
- Finish the following interpretation of the regression slope

For every 1 percentage point increase in women's labour force participation…

- How might the researchers conclude this part of their study?
- Now, write the code below to rerun this analysis using the original data (with the outliers).

```
# Write your own code here
```

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

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.

```
# Write your own code here
```

- Does it look like there might be any association between the two variables from this visualisation?
- Does it look like any of the assumptions of linear regression are broken by this data?
- 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.

```
# Write your own code here
```

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

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:

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.

```
# Write your own code here
```

- What does the relationship between the variables look like (strong? weak? positive? negative?)
- Do the patterns in the data around the regression line look like they violates any of the assumptions of linear regression?
- Now estimate a linear regression model for your chosen variables:

```
# Write your own code here
```

- What was the R-squared value for this model and what does this indicate?
- Was the relationship between the variables statistically significant at a critical value of 0.05 or not?
- Was the linear regression coefficient for the variable you chose (the slope) in a positive or negative direction? What does this mean?
- 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…)
- Write the model intercept, slope, and variables out in the form Y = B_0 + B_1*X
- 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 6 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.