HW 2

2023-01-12

Let's begin by loading in the data.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      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()
```

```
# Set our ggplot theme from the outset
theme_set(theme_light())
# Read in the data
gender_employment <- read_csv("../Data/gender_employment.csv")
```

```
## Rows: 2088 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): occupation, major_category, minor_category
## dbl (9): year, total_workers, workers_male, workers_female, percent_female, ...
##
## 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.
```

```
# Glimpse at the data
glimpse(gender_employment)
```

```
## Rows: 2,088
## Columns: 12
## $ year                   <dbl> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, ~
## $ occupation             <chr> "Chief executives", "General and operations mana~
## $ major_category         <chr> "Management, Business, and Financial", "Manageme~
## $ minor_category         <chr> "Management", "Management", "Management", "Manag~
## $ total_workers          <dbl> 1024259, 977284, 14815, 43015, 754514, 44198, 10~
## $ workers_male           <dbl> 782400, 681627, 8375, 17775, 440078, 16141, 7287~
## $ workers_female         <dbl> 241859, 295657, 6440, 25240, 314436, 28057, 3683~
## $ percent_female         <dbl> 23.6, 30.3, 43.5, 58.7, 41.7, 63.5, 33.6, 27.5, ~
## $ total_earnings         <dbl> 120254, 73557, 67155, 61371, 78455, 74114, 62187~
## $ total_earnings_male    <dbl> 126142, 81041, 71530, 75190, 91998, 90071, 66579~
## $ total_earnings_female  <dbl> 95921, 60759, 65325, 55860, 65040, 66052, 55079,~
## $ wage_percent_of_male   <dbl> 76.04208, 74.97316, 91.32532, 74.29179, 70.69719~
```

Now lets visualize the data in respect to the trend we will be looking at.

```
gender_employment %>%
  ggplot(aes(x = year, y = wage_percent_of_male)) +
  geom_jitter(alpha = 0.1) +
  geom_smooth(method = "lm") +
  labs(title = "Women's earnings with respect to men's",
       y = "% of Men's Income",
       x = "Year")
```

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

```
## Warning: Removed 846 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 846 rows containing missing values (`geom_point()`).
```

Question 1

```
# first let's relevel our categorical variable
gender_employment <- gender_employment %>%
```

```
  mutate(major_category = as.factor(major_category),
         major_category = relevel(major_category, ref = "Management, Business, and Financial"))

# next, let's create our model
parallel_model <- lm(wage_percent_of_male ~ major_category + year, data = gender_employment)

# lastly, let's summarize the data
library(broom)
tidy(parallel_model)
```

```
## # A tibble: 9 x 5
##   term                                      estimate std.e~1 stati~2  p.value
##   <chr>                                        <dbl>   <dbl>   <dbl>    <dbl>
## 1 (Intercept)                                -307.    459.     -0.669 5.04e- 1
## 2 major_categoryComputer, Engineering, and Sc~    6.32    0.946   6.68  3.56e-11
## 3 major_categoryEducation, Legal, Community S~    5.76    0.985   5.84  6.53e- 9
## 4 major_categoryHealthcare Practitioners and ~    5.52    1.10    5.00  6.41e- 7
## 5 major_categoryNatural Resources, Constructi~    4.91    1.24    3.95  8.15e- 5
## 6 major_categoryProduction, Transportation, a~   -1.31    0.960  -1.37  1.72e- 1
## 7 major_categorySales and Office                  3.33    0.858   3.88  1.11e- 4
## 8 major_categoryService                           6.08    0.885   6.87  1.03e-11
## 9 year                                            0.192   0.228   0.844 3.99e- 1
## # ... with abbreviated variable names 1: std.error, 2: statistic
```

Can we say anything about overall trends by year? Each year, the male's wage goes up by 19.2%.

1. *Calculate the wage percentage of male income for Sales and Office occupations in 2015.*

$$-306.72 + 6.32 * 0 + 5.76 * 0 + 5.52 * 0 + 4.91 * 0 - 1.31 * 0 + 3.33 * 1 + 6.08 * 0 + 0.19 * 2015 = 79.46$$

(I am confused here because I don't know how to convert this number I got into a percentage (like I did above) or if I should have converted something else before?) ALSO, I decided to use the actual year in the year category because in question 3 they replace year with the actual year.

MY INTERPRETATION: Sales and Office Occupations male's income will increase by 2015 79.461 more than Management, Business, and Financial. (Not sure if the reference category needs to be mentioned with the interpretation)

2. *Calculate the wage percentage of male income for Service occupations in 2016.*

$$-306.72 + 6.32 * 0 + 5.76 * 0 + 5.52 * 0 + 4.91 * 0 - 1.31 * 0 + 3.33 * 0 + 6.08 * 1 + 0.19 * 2016 = 82.4$$

MY INTERPRETATION: Service male's income will increase by 2016 82.4 more than Management, Business, and Financial.

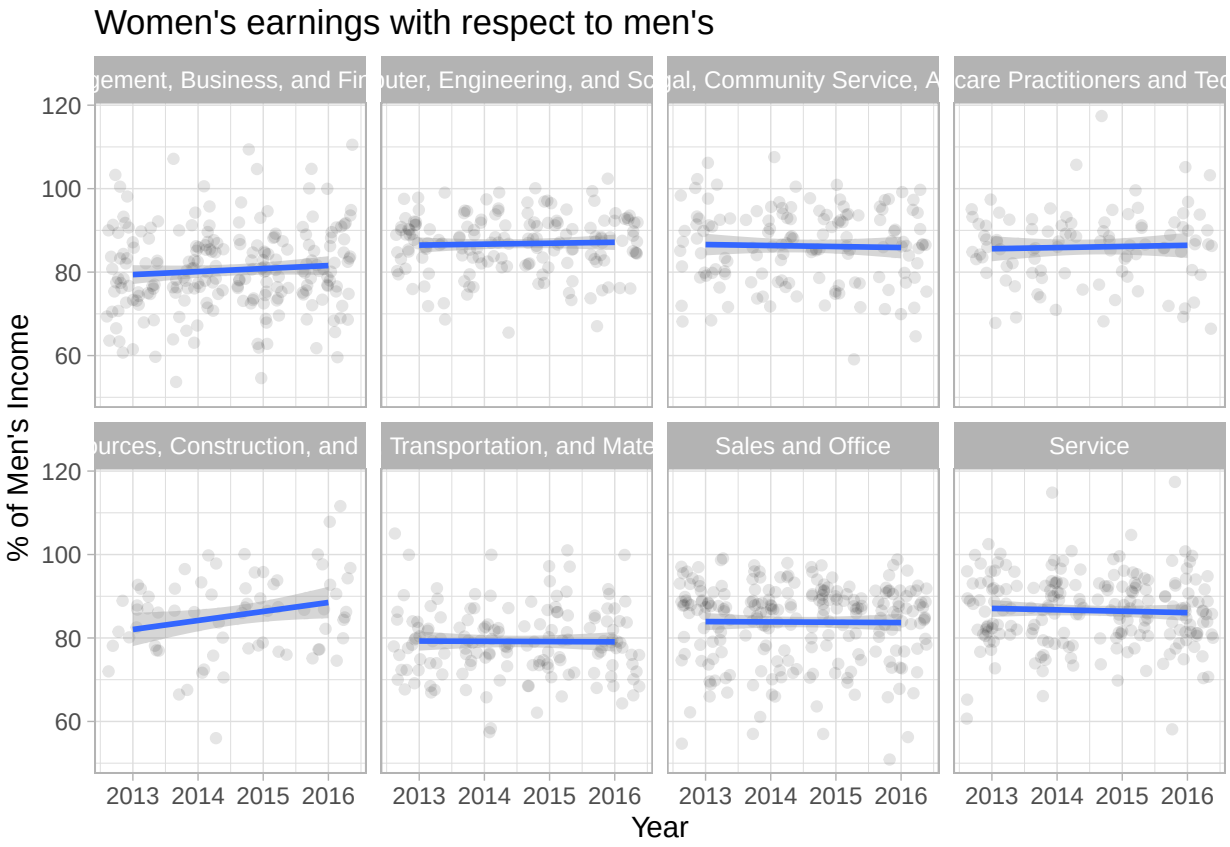
Question 2

```
gender_employment%>%
  ggplot(aes(x = year, y = wage_percent_of_male)) +
  geom_jitter(alpha = 0.1) +
  geom_smooth(method = "lm") +
  labs(title = "Women's earnings with respect to men's",
       y = "% of Men's Income",
       x = "Year") +
  facet_wrap(~major_category, nrow = 2)
```

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

```
## Warning: Removed 846 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 846 rows containing missing values (`geom_point()`).
```



Looking at this, the parallel rends assumption is not warranted because the slopes are not the same across categories. For example, I notice that the Natural Resources, Construction, and Maintenance group has a much steeper slope than most of the other categories. Additionally, the slope of Service is almost 0, but seems to be slightly negative.

Question 3

Based on this observation, now let's try fitting the model as an interaction.

```
interaction_model <- lm(wage_percent_of_male ~ major_category * year, data = gender_employment)

# lastly, let's summarize the data
library(broom)
tidy(interaction_model)
```

```
## # A tibble: 16 x 5
##    term                                      estimate std.e~1 stati~2 p.value
##    <chr>                                        <dbl>   <dbl>   <dbl>   <dbl>
##  1 (Intercept)                              -1.37e+3 1.11e+3  -1.24    0.216
##  2 major_categoryComputer, Engineering, and Sc~  1.00e+3 1.70e+3   0.589   0.556
##  3 major_categoryEducation, Legal, Community S~  1.94e+3 1.77e+3   1.09    0.275
##  4 major_categoryHealthcare Practitioners and ~  9.06e+2 1.99e+3   0.456   0.649
##  5 major_categoryNatural Resources, Constructi~ -2.89e+3 2.23e+3  -1.29    0.196
##  6 major_categoryProduction, Transportation, a~  1.58e+3 1.73e+3   0.909   0.363
##  7 major_categorySales and Office                1.61e+3 1.54e+3   1.05    0.296
##  8 major_categoryService                         2.14e+3 1.59e+3   1.34    0.180
##  9 year                                          7.20e-1 5.49e-1   1.31    0.190
## 10 major_categoryComputer, Engineering, and Sc~ -4.95e-1 8.45e-1  -0.585   0.559
## 11 major_categoryEducation, Legal, Community S~ -9.59e-1 8.80e-1  -1.09    0.276
## 12 major_categoryHealthcare Practitioners and ~ -4.47e-1 9.86e-1  -0.453   0.651
## 13 major_categoryNatural Resources, Constructi~  1.44e+0 1.11e+0   1.30    0.195
## 14 major_categoryProduction, Transportation, a~ -7.84e-1 8.61e-1  -0.910   0.363
## 15 major_categorySales and Office:year          -7.98e-1 7.65e-1  -1.04    0.297
## 16 major_categoryService:year                   -1.06e+0 7.92e-1  -1.34    0.182
## # ... with abbreviated variable names 1: std.error, 2: statistic
```

What would the estimate be for "Computer, Engineering, and Science" for 2016.
(this time, I will only include things in my equation that are "turned on" like in the example, instead of writing out the things that will receive 0s.)

$$-1370.47 + 0.72 * 2016 + 10002.85 * 1 - 0.49 * 2016 * 1 = 9096.06$$

What about for Service?

$$-1370.47 + 0.72 * 2016 + 2137.65 * 1 - 1.06 * 2016 * 1 = 81.72$$

I notice that the estimates for each are not as close as they were when we used the parallel model. This makes sense as interaction models allow the slopes to be different while the parallel model seems to constrain the slopes so that they can be parallel.

Question 4

Why would we choose to build a model that assumes parallel trends?
The parallel model is easier, and interaction models are more complex. When the complexity is not warranted, there is no reason to use such a complicated model like an interaction. For example, when the slopes seem not to differ as much, then adding the complexity of an interaction model makes no sense.

Question 5

We will start by building a simple model where `wage_percent_of_male` is the outcome variable and `year` is the predictor.

```
# build model
simple_fit <- lm(wage_percent_of_male ~ year, data = gender_employment)

# summarize
tidy(simple_fit)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept) -322.      479.       -0.671   0.502
## 2 year           0.201     0.238     0.847   0.397
```

With every one year increase, there is a 20.14% increase in men's income.

```
gender_employment %>%
  select(year, wage_percent_of_male, percent_female) %>%
  cor(use = "complete.obs")
```

```
##                            year wage_percent_of_male percent_female
## year                1.000000000           0.02403895    0.004998286
## wage_percent_of_male 0.024038950           1.00000000    0.111464461
## percent_female       0.004998286           0.11146446    1.000000000
```

Now we have the correlation coefficients between year, percent of men's income, and percent of females.

We see pretty weak correlations between the variables (0.02, 0.005, 0.11).

Now that we see there is weak correlation however, we really want to see if the relationship between year andd the paygap are conditional on the proportion of women who work in an occupation. To do this we will add proportion of women who work in an occupation to our linear model.

```
# create the model
multiple_fit <- lm(wage_percent_of_male ~ year + percent_female, data = gender_employment)

# summarize
tidy(multiple_fit)
```

```
## # A tibble: 3 x 5
##   term            estimate std.error statistic   p.value
##   <chr>              <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    -314.      477.        -0.660 0.510
## 2 year              0.197     0.237      0.832 0.406
## 3 percent_female    0.0425    0.0108     3.94  0.0000843
```

It seems that a change in year has more of an impact on men's income than the proportion of women who work in the occupation (19.69% vs. 4.25%). For me this makes sense; however, I did think that there would be a higher impact from the proportion of women within the occupation... even if a negative impact.

Question 6

Briefly tell me, in your own words, what R-squared is. R-squared has to do with variation in the dependent variable around the mean that is explained by the model. In other words, the R-squared value shows you the variation around the regression line.

```
# let's look at the r-squared value for the model without the condition
simple_glanced <- glance(simple_fit)
simple_glanced$r.squared
```

```
## [1] 0.0005778711
```

```
# let's look at the r-squared value for the model with the condition
multiple_glanced <- glance(multiple_fit)
multiple_glanced$r.squared
```

```
## [1] 0.01297574
```

Since the multiple fit model has a high R-squared value, it seems like this model explains more variation from the dependent variable than the simple model. This means that the multiple variable model is likely a better fit.

The simple model has an r-squared value of 0.05% and the r-squared value for the multiple variable models is 1.30%.