SOC6707 Intermediate Data Analysis

Monica Alexander

Week 7: Interactions, and thinking about problems

Announcements

- Mode of delivery
- Content
- Presentations?
- A2
- Recording plan for EDA related stuff

Today: seminar at 12pm

Harvey J. Nicholson: 'Considering within-group heterogeneity to explore Black Americans' feelings toward Asian Americans: the case of African Americans and Black Caribbeans'

Interaction terms

Effect moderation

- Effect moderation refers to the situation where the partial effect of one explanatory variable differs or changes across levels of another explanatory variable
  - e.g. the association between income and age may vary by education level
- All of the models we have considered thus far constrain the partial effects of the explanatory variables to be invariant, but this may not be appropriate
- If a model constrains partial effects to be invariant when in fact they are not, our estimates will be wrong

We can accommodate effect moderation through the use of **interaction terms**

Interaction terms

Example of an MLR model with an interaction term:

$$\begin{aligned} Y_i &= E\left(Y_i \mid X_{i1}, X_{i2}\right) + \varepsilon_i \\ &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} \end{aligned}$$

- How should we interpret the parameters in an MLR model with interaction terms?
- First, let's look at an example

Example

- What is the association between TFR, life expectancy and region?
- Does the association between TFR and life expectancy differ based on whether country is in Developed Regions or not?

Example in R

```
country_ind_2017 <- country_ind %>%
  filter(year==2017) %>%
  mutate(dev_region = ifelse(region=="Developed regions", "yes", "no"))

summary(lm(tfr ~ life_expectancy + dev_region + life_expectancy*dev_region, data = country_ind_2017))
```

```
##
## Call:
## lm(formula = tfr ~ life_expectancy + dev_region + life_expectancy *
##     dev_region, data = country_ind_2017)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.23326 -0.29618 -0.02426  0.28744  2.54832
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     13.52646    0.52158  25.933  < 2e-16 ***
## life_expectancy                 -0.14454    0.00722 -20.019  < 2e-16 ***
## dev_regionyes                  -12.95159    2.91594  -4.442 1.59e-05 ***
## life_expectancy:dev_regionyes    0.15711    0.03557   4.417 1.76e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6164 on 172 degrees of freedom
## Multiple R-squared:  0.7784, Adjusted R-squared:  0.7745
## F-statistic: 201.4 on 3 and 172 DF,  p-value: < 2.2e-16
```

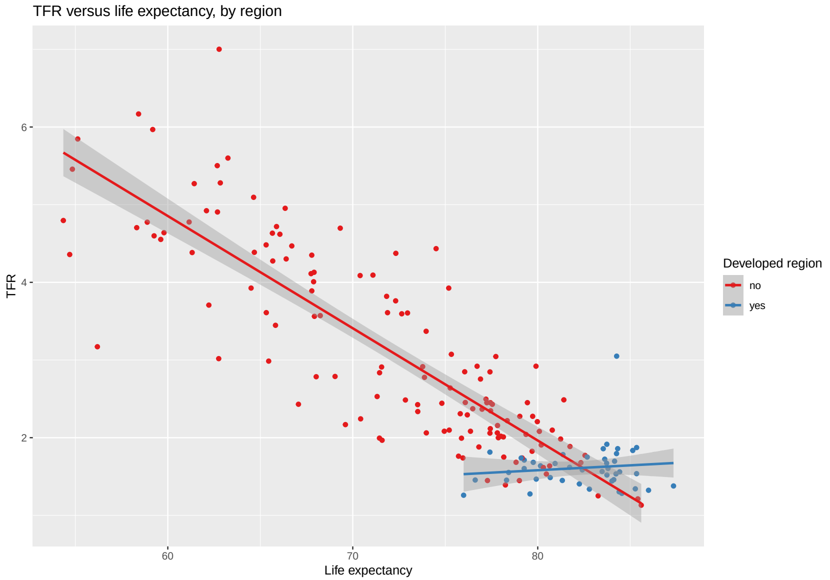
Example

$$Y_i = 13.5 - 0.14X_1 - 13.0X_2 + 0.16X_1X_2$$

Some interpretations

- for non-developed regions, 1 year increase in life expectancy associated with 0.14 decrease in TFR
- for developed regions, a 1 year increase in life expectancy associated with a 0.02 increase in TFR

Visualizing interactions



Interaction terms

Now, let's take a look at how $E(Y_i \mid X_{i1}, X_{i2})$ changes with a unit increase in X_{i1} in the general case

Interaction terms

$$E(Y_i \mid X_{i1}, X_{i2}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2}$$

In this model, the change in the expected value of Y_i associated with a unit increase in X_{i1} is given by

$$E(Y_i \mid X_{i1} = x_1 + 1, X_{i2} = x_2) - E(Y_i \mid X_{i1} = x_1, X_{i2} = x_2) = \beta_1 + \beta_3 x_2$$

- The partial effect of X_{i1} now depends on the value to which we set the other explanatory variable, X_{i2}
- Note that when $X_{i2} = 0$, this expression simplifies to β_1, or in other words, β_1 is the change in the expected value of Y_i associated with a unit increase in X_{i1} specifically when $X_{i2} = 0$

Interaction terms

Next, let's take a look at how the partial effect of $X_{i1}, \beta_1 + \beta_3 x_2$, changes with a unit increase in X_{i2}

The change in the partial effect of X_{i1} associated with a unit increase in X_{i2} is given by

$$\begin{aligned}&[E(Y_i \mid X_{i1} = x_1 + 1, X_{i2} = x_2 + 1) - E(Y_i \mid X_{i1} = x_1, X_{i2} = x_2 + 1)] \\ &- [E(Y_i \mid X_{i1} = x_1 + 1, X_{i2} = x_2) - E(Y_i \mid X_{i1} = x_1, X_{i2} = x_2)] = \beta_3\end{aligned}$$

In words, β_3 represents the amount by which the partial effect of X_{i1} differs across levels of the other explanatory variable, X_{i2}

Interaction terms

- The previous slides may take a little getting used to
- In reality, one of our explanatory variables (say X_{i2}) is a binary variable (so either 0 or 1)
- This simplifies the interpretation of the interaction term

Another example

```r
gss <- gss %>% mutate(age_c = age - mean(age))
mod <- lm(age_at_first_marriage~ age_c + has_bachelor_or_higher, data = gss)
summary(mod)
```

```
##
## Call:
## lm(formula = age_at_first_marriage ~ age_c + has_bachelor_or_higher,
##     data = gss)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.046  -3.379  -1.254   2.026  27.226
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              24.489886   0.116125 210.892   <2e-16 ***
## age_c                    -0.061697   0.006172  -9.996   <2e-16 ***
## has_bachelor_or_higherYes  1.982372   0.184405  10.750   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.295 on 5265 degrees of freedom
##   (15334 observations deleted due to missingness)
## Multiple R-squared:  0.04454,    Adjusted R-squared:  0.04417
## F-statistic: 122.7 on 2 and 5265 DF,  p-value: < 2.2e-16
```

With interaction

```r
mod2 <- lm(age_at_first_marriage~ age_c*has_bachelor_or_higher, data = gss)
summary(mod2)
```

```
##
## Call:
## lm(formula = age_at_first_marriage ~ age_c * has_bachelor_or_higher,
##     data = gss)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.970  -3.372  -1.211   2.018  27.328
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    24.58532    0.12365 198.837  < 2e-16 ***
## age_c                          -0.06882    0.00694  -9.916  < 2e-16 ***
## has_bachelor_or_higherYes       1.62500    0.24374   6.667 2.88e-11 ***
## age_c:has_bachelor_or_higherYes  0.03397    0.01516   2.241   0.0251 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.293 on 5264 degrees of freedom
##   (15334 observations deleted due to missingness)
## Multiple R-squared:  0.04545,    Adjusted R-squared:  0.0449
## F-statistic: 83.54 on 3 and 5264 DF,  p-value: < 2.2e-16
```

Visualizing

Interpretation

```
##
## Call:
## lm(formula = age_at_first_marriage ~ age_c * has_bachelor_or_higher,
##     data = gss)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.970  -3.372  -1.211   2.018  27.328
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    24.58532    0.12365 198.837  < 2e-16 ***
## age_c                          -0.06882    0.00694  -9.916  < 2e-16 ***
## has_bachelor_or_higherYes       1.62500    0.24374   6.667 2.88e-11 ***
## age_c:has_bachelor_or_higherYes  0.03397    0.01516   2.241   0.0251 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.293 on 5264 degrees of freedom
##   (15334 observations deleted due to missingness)
## Multiple R-squared:  0.04545,    Adjusted R-squared:  0.0449
## F-statistic: 83.54 on 3 and 5264 DF,  p-value: < 2.2e-16
```

Problems

Example

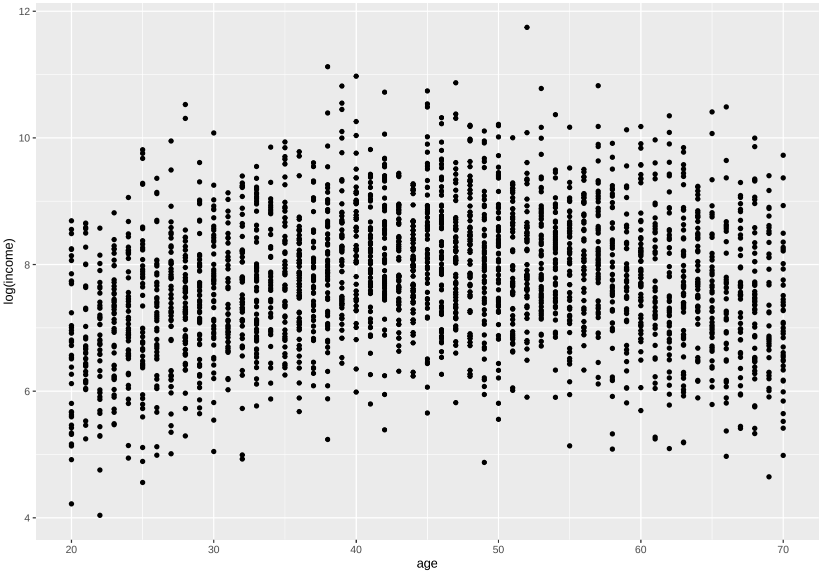
Pretend we are interested in studying the association between education and income. We have a dataset with income (weekly $), age (years) and years of schooling. Note that some incomes are missing.

```
## # A tibble: 5,100 x 3
##      age   yrs income
##    <int> <int>  <dbl>
##  1    20    14  1392.
##  2    20    10    NA
##  3    20     8  5949.
##  4    20    12   269.
##  5    20     9    NA
##  6    20     9    NA
##  7    20    13   204.
##  8    20     9    NA
##  9    20    10    NA
## 10    20    15   277.
## # ... with 5,090 more rows
```

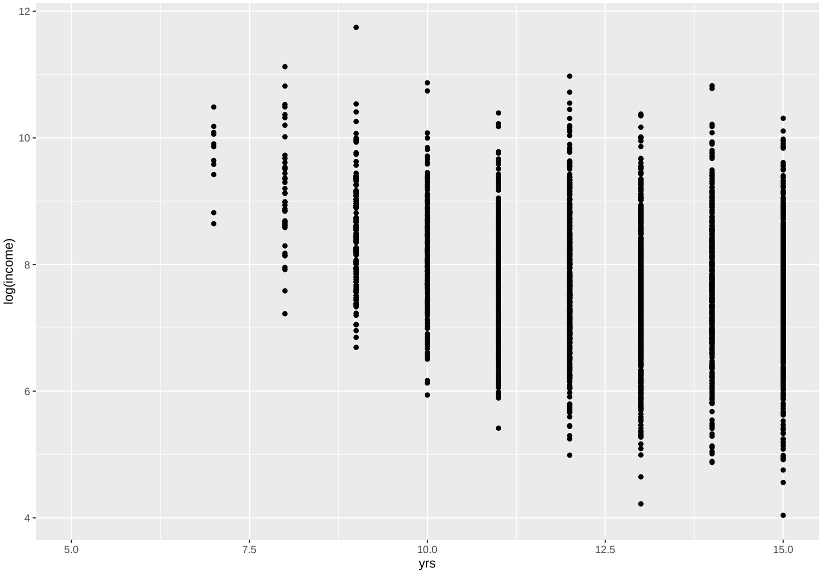
Summaries

```
## # A tibble: 11 x 4
##      yrs mean_log_income     n n_income_missing
##    <int>           <dbl> <int>            <int>
##  1     5           NaN     413              413
##  2     6           NaN     463              463
##  3     7             9.70  458              447
##  4     8             9.15  486              445
##  5     9             8.53  490              359
##  6    10             8.17  452              216
##  7    11             7.80  438              100
##  8    12             7.72  465               45
##  9    13             7.55  487               11
## 10    14             7.66  476                2
## 11    15             7.49  472                0
```

Age versus log income



Education versus log(income)



Problem: mis-specification

Regression

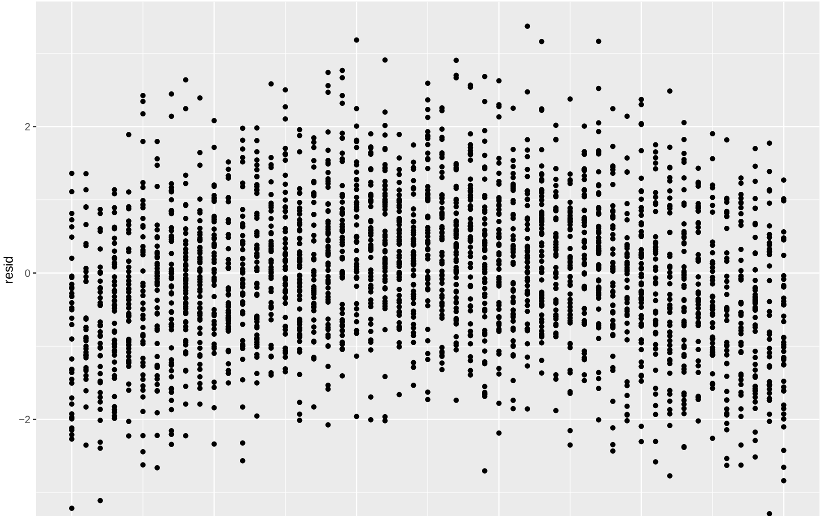
Note missing values get dropped from regression

```r
d <- d %>% mutate(log_income = log(income))
mod <- lm(data = d, log_income ~ age+yrs)
summary(mod)
```

```
##
## Call:
## lm(formula = log_income ~ age + yrs, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2881 -0.6877  0.0129  0.6982  3.3710
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.221695   0.148425   62.13  < 2e-16 ***
## age          0.010235   0.001406    7.28 4.42e-13 ***
## yrs         -0.153262   0.010541  -14.54  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.013 on 2596 degrees of freedom
##   (2501 observations deleted due to missingness)
## Multiple R-squared:  0.09298,    Adjusted R-squared:  0.09228
## F-statistic: 133.1 on 2 and 2596 DF,  p-value: < 2.2e-16
```

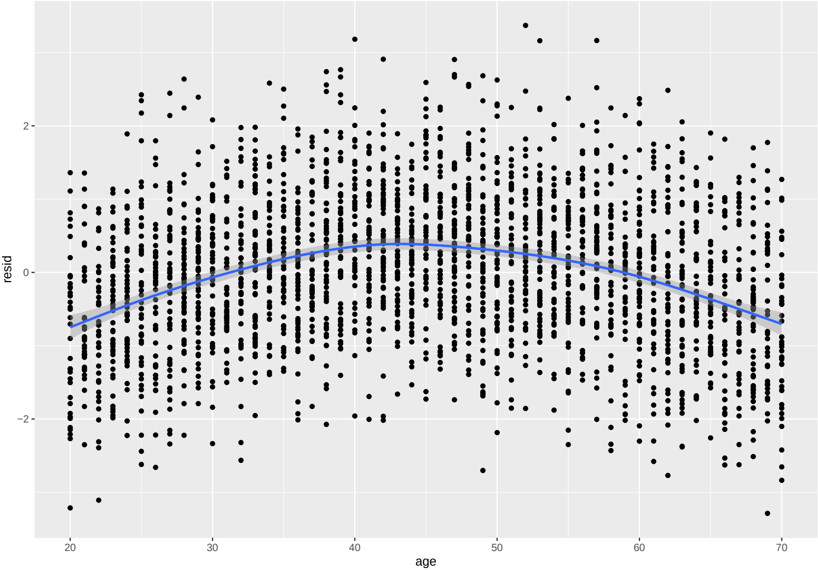
Residuals

```
df_resid <- tibble(resid = residuals(mod),
                   age = d %>%
                     drop_na() %>%
                     select(age) %>%
                     pull())
ggplot(data = df_resid, aes(age, resid)) + geom_point()
```



Residuals

```
ggplot(data = df_resid, aes(age, resid)) + geom_point()+ geom_smooth()
```



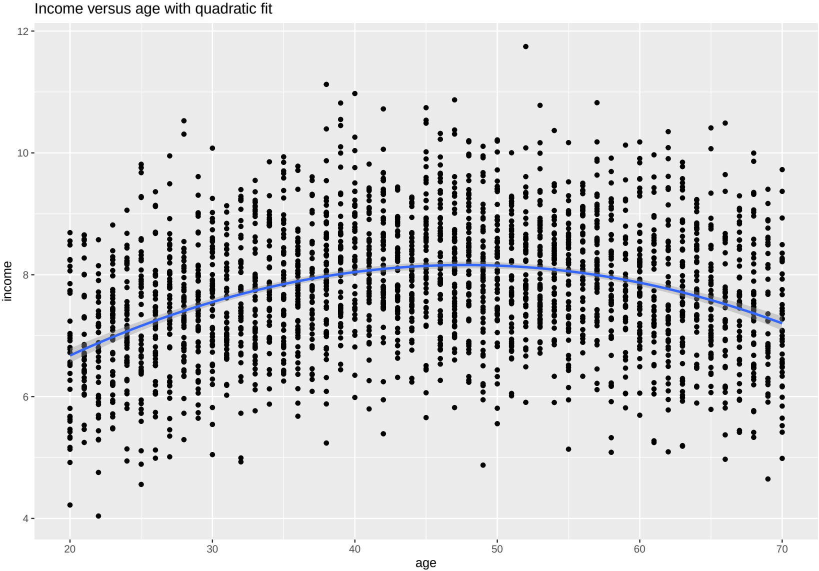
Rerun model

```
d <- d %>% mutate(age_sq = age^2)
mod2 <- lm(data = d, log_income ~ age+age_sq + yrs)
summary(mod2)
```

```
##
## Call:
## lm(formula = log_income ~ age + age_sq + yrs, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.05843 -0.64437  0.00297  0.62313  3.13452
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.7103425  0.2374233   24.05   <2e-16 ***
## age          0.1758546  0.0091503   19.22   <2e-16 ***
## age_sq      -0.0018383  0.0001005  -18.29   <2e-16 ***
## yrs         -0.1414379  0.0099433  -14.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9537 on 2595 degrees of freedom
##   (2501 observations deleted due to missingness)
## Multiple R-squared:  0.1966, Adjusted R-squared:  0.1956
## F-statistic: 211.6 on 3 and 2595 DF,  p-value: < 2.2e-16
```

Bonus! you just learnt polynomial regression

How to interpret? Easiest to plot the relationship, and pick a few ages to calculate the effect.



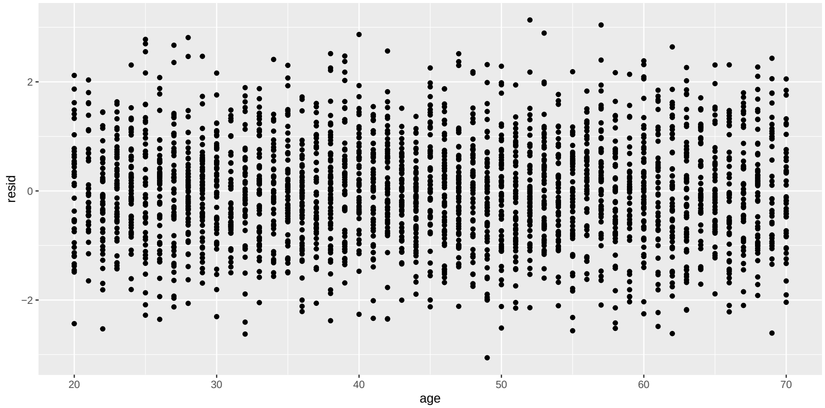
Interpretation

Effect of age when age equals:

- 25:
- 45:
- 65:

Residuals

```
df_resid <- tibble(resid = residuals(mod2),
                   age = d %>%
                     drop_na() %>%
                     select(age) %>%
                     pull())
ggplot(data = df_resid, aes(age, resid)) + geom_point()
```



Problem: missing data and collider bias

Missing data

- We have quite a lot of missing observations of income in this dataset
- We can still run regressions in R, `lm` doesn't mind at all
- Just drops the missing rows

```
##
## Call:
## lm(formula = log_income ~ age + yrs, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2881 -0.6877  0.0129  0.6982  3.3710
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.221695   0.148425   62.13  < 2e-16 ***
## age          0.010235   0.001406    7.28 4.42e-13 ***
## yrs         -0.153262   0.010541  -14.54  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.013 on 2596 degrees of freedom
##   (2501 observations deleted due to missingness)
## Multiple R-squared:  0.09298,    Adjusted R-squared:  0.09228
## F-statistic: 133.1 on 2 and 2596 DF,  p-value: < 2.2e-16
```

Missing data

- If we use our model to make inferences about the relationship between education and income for the whole population, what are we assuming?

Missing at random

- Assuming the people we have are representative of the broader population
- The relationship between education and income we see is true for those missing also
- There's no systematic reason for people being missing

But we know this isn't true

- The people with observed values of income are more likely to have more years of schooling
- It's very conceivable that the relationship between education and income may be different for those with missing observations

Collider bias

- Colliders (e.g. non-response bias)
- Schooling and income both influence survey response
- Conditioning on survey response creates a non-causal association between schooling and income
- In our example, higher income and education both increase the chance of response, then conditioning on responding to the survey (i.e. only looking at non-missing values) if someone has a relatively high education then it's more likely they have a lower income. This creates a non-causal negative association between education and income

What can you do about missing data

Broadly, there are two strategies:

1. Remove
2. Impute

Removal of missing data

- This is essentially what we've been doing! All rows with any missing values for variables that go into the regression are removed
- This is okay, as long as you know what's being removed
- May be useful to remove variables that have a lot of unexplained missingness from your analysis

Imputation

We won't cover in this class, but broadly

- Make a decision to impute some reasonable values for some/all missing data
- e.g. mean, median, mode, group means, modeled based predictions
- more complex strategies: multiple imputation

Monica is not a fan because it's hard to propagate and quantify uncertainty (and we all about quantifying uncertainty, why else are we running regressions)

Embrace the missingness

- Try to understand what's missing and how it might affect your conclusions
- If appropriate, you may be able to redefine your research question
- Reconsider using variables that have a lot of missingness