PSY 503: Foundations of Psychological Methods

Lecture 19: Heterogeneity and Experimental Regression Adjustments

Robin Gomila

Princeton

November 11, 2020

Heterogeneity

Heterogeneity

- So far, we have considered cases in which β_1 was constant across subgroups

Heterogeneity

- So far, we have considered cases in which β_1 was constant across subgroups
- What happens if it's not?

Heterogeneity

- So far, we have considered cases in which β_1 was constant across subgroups
- What happens if it's not?
- Let's open R Studio!

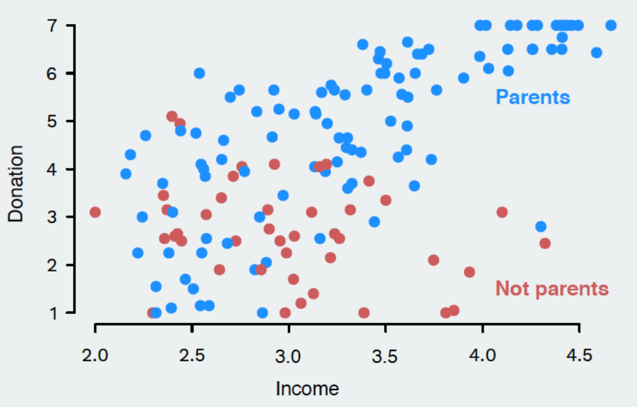
Why interaction terms?

- Interaction terms will allow you to let the **slope** on one variable vary as a function of another variable

Why interaction terms?

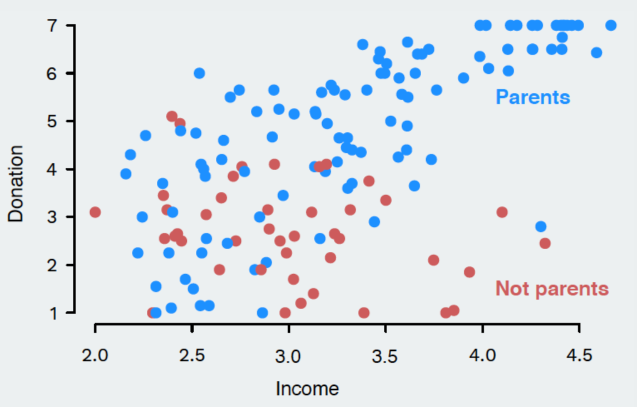
- Interaction terms will allow you to let the **slope** on one variable vary as a function of another variable
- Let's explore a different hypothetical dataset describing the relationship between income and donations conditional on having children

Let's see the data



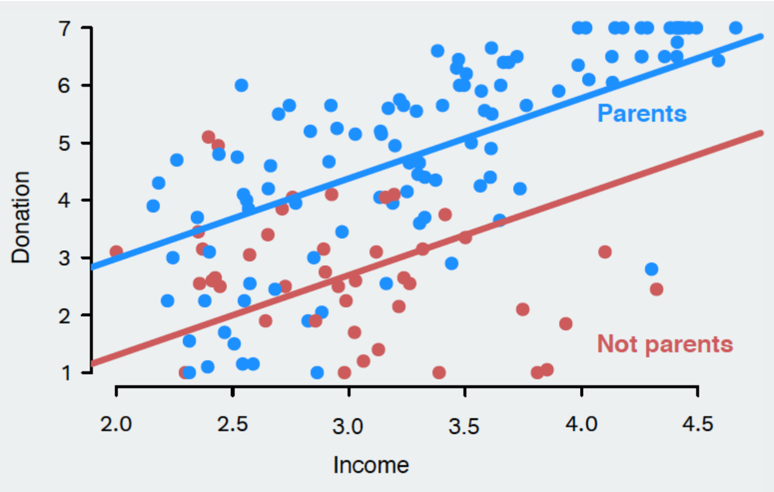
- Does it look like there's heterogeneity?

Let's see the data



- Does it look like there's heterogeneity? What happens if we simply control for "parents" additively?

Controlling for "parent" additively



- The regression is a poor fit for non-parents. Can we allow for different slopes for each group?

Interactions with a binary variable

- Let X_2 be binary

Interactions with a binary variable

- Let X_2 be binary
- In this case, $X_2 = 1$ for parents

Interactions with a binary variable

- Let X_2 be binary
- In this case, $X_2 = 1$ for parents
- We can add another covariate to the baseline model that allows the effect of income to vary by parental status

Interactions with a binary variable

- Let X_2 be binary
- In this case, $X_2 = 1$ for parents
- We can add another covariate to the baseline model that allows the effect of income to vary by parental status
- This covariate is called an interaction term and it is the product of the two marginal variables of interest: $income \times parent$

Interactions with a binary variable

- Let X_2 be binary
- In this case, $X_2 = 1$ for parents
- We can add another covariate to the baseline model that allows the effect of income to vary by parental status
- This covariate is called an interaction term and it is the product of the two marginal variables of interest: $income \times parent$
- Here is the model with the interaction term:

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2 + \widehat{\beta}_3 X_1 X_2$$

Two lines in one regression

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1_i} + \widehat{\beta}_2 X_{2_i} + \widehat{\beta}_3 X_{1_i} X_{2_i}$$

- How can we interpret this model?

Two lines in one regression

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1_i} + \widehat{\beta}_2 X_{2_i} + \widehat{\beta}_3 X_{1_i} X_{2_i}$$

- How can we interpret this model? We can plug in the two possible values of X_{2i}

Two lines in one regression

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1_i} + \widehat{\beta}_2 X_{2_i} + \widehat{\beta}_3 X_{1_i} X_{2_i}$$

- How can we interpret this model? We can plug in the two possible values of X_{2i}
- When $X_{2_i} = 0$:

$$\begin{aligned}
\widehat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_{1_i} + \widehat{\beta}_2 X_{2_i} + \widehat{\beta}_3 X_{1_i} X_{2_i} \\
&= \widehat{\beta}_0 + \widehat{\beta}_1 X_{1_i} + \widehat{\beta}_2 \times 0 + \widehat{\beta}_3 X_{1_i} \times 0 \\
&= \widehat{\beta}_0 + \widehat{\beta}_1 X_{1_i}
\end{aligned}$$

Two lines in one regression

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1_i} + \widehat{\beta}_2 X_{2_i} + \widehat{\beta}_3 X_{1_i} X_{2_i}$$

- How can we interpret this model? We can plug in the two possible values of X_{2i}
- When $X_{2_i} = 0$:

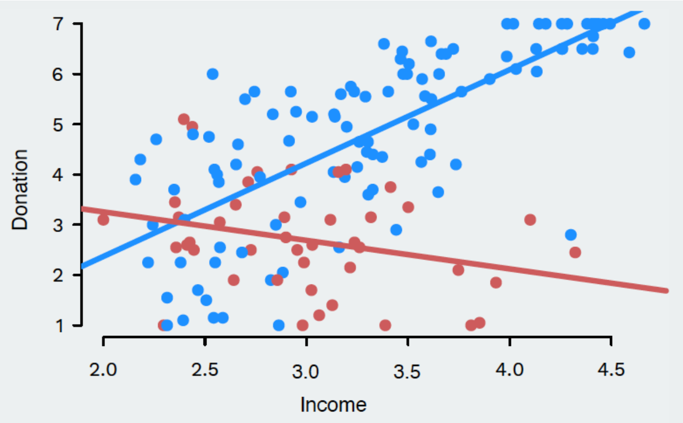
$$\begin{aligned}\widehat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_{1_i} + \widehat{\beta}_2 X_{2_i} + \widehat{\beta}_3 X_{1_i} X_{2_i} \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_{1_i} + \widehat{\beta}_2 \times 0 + \widehat{\beta}_3 X_{1_i} \times 0 \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_{1_i}\end{aligned}$$

- When $X_{2_i} = 1$:

$$\begin{aligned}\widehat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_{1_i} + \widehat{\beta}_2 X_{2_i} + \widehat{\beta}_3 X_{1_i} X_{2_i} \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_{1_i} + \widehat{\beta}_2 \times 1 + \widehat{\beta}_3 X_{1_i} \times 1 \\ &= (\widehat{\beta}_0 + \widehat{\beta}_2) + (\widehat{\beta}_1 + \widehat{\beta}_3) X_{1_i}\end{aligned}$$

Example interpretation of the coefficients

	Intercept for X_{2i}	Slope for X_{2i}
$X_{2i} = 0$	$\widehat{\beta}_0$	$\widehat{\beta}_1$
$X_{2i} = 1$	$\widehat{\beta}_0 + \widehat{\beta}_2$	$\widehat{\beta}_1 + \widehat{\beta}_3$



General interpretation of the coefficients

- $\widehat{\beta}_0$: average value of Y_i when both X_{1_i} and X_{2_i} are equal to 0

General interpretation of the coefficients

- $\widehat{\beta}_0$: average value of Y_i when both X_{1_i} and X_{2_i} are equal to 0
- $\widehat{\beta}_1$: a one-unit change in X_{1_i} is associated with a $\widehat{\beta}_1$-unit change in Y_i when $X_{2_i} = 0$

General interpretation of the coefficients

- $\widehat{\beta}_0$: average value of Y_i when both X_{1_i} and X_{2_i} are equal to 0
- $\widehat{\beta}_1$: a one-unit change in X_{1_i} is associated with a $\widehat{\beta}_1$-unit change in Y_i when $X_{2_i} = 0$
- $\widehat{\beta}_2$: average difference in Y_i between $X_{2_i} = 1$ group and $X_{2_i} = 0$ group when $X_{1_i} = 0$

General interpretation of the coefficients

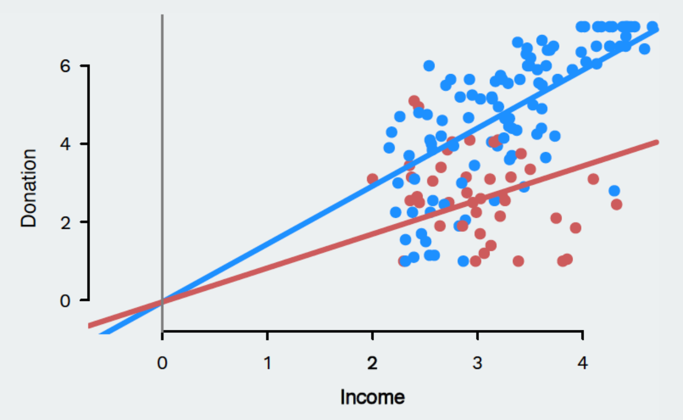
- $\widehat{\beta}_0$: average value of Y_i when both X_{1_i} and X_{2_i} are equal to 0
- $\widehat{\beta}_1$: a one-unit change in X_{1_i} is associated with a $\widehat{\beta}_1$-unit change in Y_i when $X_{2_i} = 0$
- $\widehat{\beta}_2$: average difference in Y_i between $X_{2_i} = 1$ group and $X_{2_i} = 0$ group when $X_{1_i} = 0$
- $\widehat{\beta}_3$: change in the effect of X_{1_i} on Y_i between $X_{2_i} = 1$ group and $X_{2_i} = 0$ group

Lower order terms

- **Principle of Marginality**: Always include the lower order terms!
- Imagine that we ommitted the lower order term:

Lower order terms

- **Principle of Marginality**: Always include the lower order terms!
- Imagine that we ommitted the lower order term:



Omitting lower order terms

$$\widehat{Y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1_i} + 0 \times X_{2_i} + \widehat{\beta}_3 X_{1_i} X_{2_i}$$

Omitting lower order terms

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1_i} + 0 \times X_{2_i} + \widehat{\beta}_3 X_{1_i} X_{2_i}$$

- This model does not allow for a difference between parents and non-parents when income is 0
  - This distorts slope estimates
  - Very rarely justified
  - Yet, for some reasons, people do it...

A note on interactions with non-binary discrete and continuous variables

- Same principle!
- Plug in values in the equation to get the marginal effect of X_{1_i} conditional on specific values of X_{i_2}

Summary for interactions

- Do not omit lower order terms because it usually imposes unrealistic restrictions

Summary for interactions

- Do not omit lower order terms because it usually imposes unrealistic restrictions
- Do not interpret the coefficients on the lower order terms as marginal effects

Summary for interactions

- Do not omit lower order terms because it usually imposes unrealistic restrictions
- Do not interpret the coefficients on the lower order terms as marginal effects
  - They give the marginal effect only for the case where the other variable is equal to zero

Summary for interactions

- Do not omit lower order terms because it usually imposes unrealistic restrictions
- Do not interpret the coefficients on the lower order terms as marginal effects
  - They give the marginal effect only for the case where the other variable is equal to zero
- The p-value on the interaction term can be used as a test against the null of no interaction, but significant tests for the lower order term rarely make sense

Polynomials

Polynomial terms

- Polynomial terms are a special case of the continuous variable interactions
  - where $X_1 = X_2$

Polynomial terms

- Polynomial terms are a special case of the continuous variable interactions
  - where $X_1 = X_2$

$$\begin{aligned}
\widehat{Y_i} &= \widehat{\beta}_0 + \widehat{\beta}_1 X_{1_i} + \widehat{\beta}_2 X_{1_i} + \widehat{\beta}_3 X_{1_i} X_{1_i} \\
&= \widehat{\beta}_0 + (\widehat{\beta}_1 + \widehat{\beta}_2) X_{1_i} + \widehat{\beta}_3 X_{1_i} X_{1_i} \\
&= \widehat{\beta}_0 + (\widehat{\beta}_1 + \widehat{\beta}_2) X_{1_i} + \widehat{\beta}_3 X_{1_i}^2
\end{aligned}$$

Polynomial terms

- Polynomial terms are a special case of the continuous variable interactions
  - where $X_1 = X_2$

$$\begin{aligned}
\widehat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_{1_i} + \widehat{\beta}_2 X_{1_i} + \widehat{\beta}_3 X_{1_i} X_{1_i} \\
&= \widehat{\beta}_0 + (\widehat{\beta}_1 + \widehat{\beta}_2) X_{1_i} + \widehat{\beta}_3 X_{1_i} X_{1_i} \\
&= \widehat{\beta}_0 + (\widehat{\beta}_1 + \widehat{\beta}_2) X_{1_i} + \widehat{\beta}_3 X_{1_i}^2
\end{aligned}$$

- This is called a **second order polynomial** in X_1, which we generally write:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$$

Polynomial terms

- Polynomial terms are a special case of the continuous variable interactions
  - where $X_1 = X_2$

$$
\begin{aligned}
\widehat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_{1_i} + \widehat{\beta}_2 X_{1_i} + \widehat{\beta}_3 X_{1_i} X_{1_i} \\
&= \widehat{\beta}_0 + (\widehat{\beta}_1 + \widehat{\beta}_2) X_{1_i} + \widehat{\beta}_3 X_{1_i} X_{1_i} \\
&= \widehat{\beta}_0 + (\widehat{\beta}_1 + \widehat{\beta}_2) X_{1_i} + \widehat{\beta}_3 X_{1_i}^2
\end{aligned}
$$

- This is called a **second order polynomial** in X_1, which we generally write:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$$

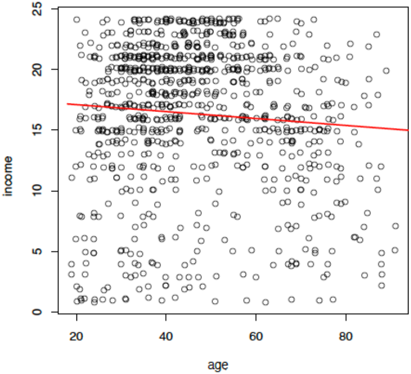
- A third order polynomial is given by:
  $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_i$

Polynomial example: Income and Age

- Classic example of a non-linear relationship between two variables: income and age

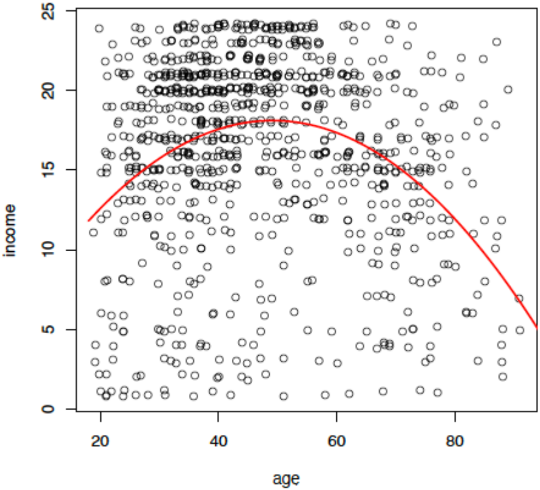
Polynomial example: Income and Age

- Classic example of a non-linear relationship between two variables: income and age
- Here we can see that simple linear regression does not seem to fit the data

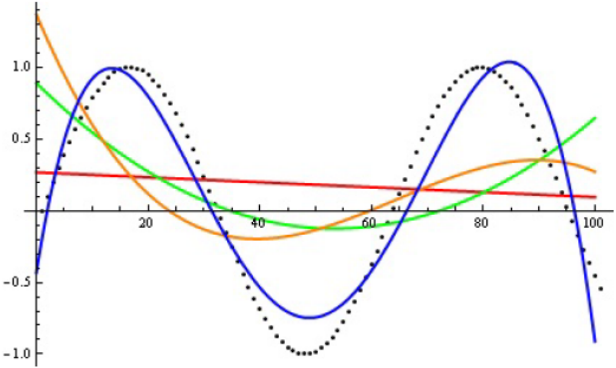


Polynomial example: Income and Age

- A second order polynomial in age fits the data a lot better:



Higher order polynomials



Covariate Adjustments in Experimental Designs

Why include PRETREATMENT covariates in experimental data analysis?

Why include PRETREATMENT covariates in experimental data analysis?

- In experiments: Regression analysis with and without PRETREATMENT covariate adjustments is unbiased!
  - Random assignment to conditions creates two groups that are, in expectation, identical prior to treatment. This implies pretreatment covariates are not correlated with the treatment

Why include PRETREATMENT covariates in experimental data analysis?

- In experiments: Regression analysis with and without PRETREATMENT covariate adjustments is unbiased!
  - Random assignment to conditions creates two groups that are, in expectation, identical prior to treatment. This implies pretreatment covariates are not correlated with the treatment
- So why adjust for covariates?

Why include PRETREATMENT covariates in experimental data analysis?

- In experiments: Regression analysis with and without PRETREATMENT covariate adjustments is unbiased!
  - Random assignment to conditions creates two groups that are, in expectation, identical prior to treatment. This implies pretreatment covariates are not correlated with the treatment
- So why adjust for covariates? Controlling for pretreatment covariates:
  - Increases precision
  - Increases statistical power

Why include PRETREATMENT covariates in experimental data analysis?

- In experiments: Regression analysis with and without PRETREATMENT covariate adjustments is unbiased!
  - Random assignment to conditions creates two groups that are, in expectation, identical prior to treatment. This implies pretreatment covariates are not correlated with the treatment
- So why adjust for covariates? Controlling for pretreatment covariates:
  - Increases precision
  - Increases statistical power
- Make sure to pre-register your model!

Including covariates in regression analysis

Let's open R studio!

Any drawbacks to adjusting for pretreatment covariates?

Any drawbacks to adjusting for pretreatment covariates?

- In small sample ($n < 20$): Including pretreatment covariates introduces bias in the estimate of the ATE

Any drawbacks to adjusting for pretreatment covariates?

- In small sample ($n < 20$): Including pretreatment covariates introduces bias in the estimate of the ATE
- In larger samples ($n > 20$): Bias becomes negligible

Any drawbacks to adjusting for pretreatment covariates?

- In small sample ($n < 20$): Including pretreatment covariates introduces bias in the estimate of the ATE
- In larger samples ($n > 20$): Bias becomes negligible
- If you are going to include more than one pretreatment covariates, make sure sample is large enough!

Any drawbacks to adjusting for pretreatment covariates?

- In small sample ($n < 20$): Including pretreatment covariates introduces bias in the estimate of the ATE
- In larger samples ($n > 20$): Bias becomes negligible
- If you are going to include more than one pretreatment covariates, make sure sample is large enough!
  - How? Use Monte-Carlo simulation...

Including covariates in regression analysis

- **Key assumption:**

Including covariates in regression analysis

- **Key assumption:** Covariates are unaffected by treatment assignment
  - i.e., the schedule of covariates is FIXED regardless of whether an individual was assigned to treatment vs. control condition

Including covariates in regression analysis

- **Key assumption:** Covariates are unaffected by treatment assignment
  - i.e., the schedule of covariates is FIXED regardless of whether an individual was assigned to treatment vs. control condition
- **IMPLICATION: DO NOT adjust for posttreatment covariates!**
  - It will bias ATE
  - It will turn your experimental study into a correlational study
  - It will not allow you to make **any** causal claims

What is a posttreatment covariate?

- Any variables that: were measured after the treatment / may be affected by the treatment

Why you SHOULD NOT adjust for posttreatment covariate?

- Adjusting for posttreatment covariates **introduces bias** in your estimate of the average treatment effect, *even if treatment was randomly assigned to participants*
  - That is because variables that are affected by treatment cannot possibly be independent of treatment assignment
- Pervasive issue in social psychology —so called mediation analysis!

Why you SHOULD NOT adjust for posttreatment covariate?

Back to R Studio!