Multiple Linear Regression and Interactions

EDS 222

Tamma Carleton
Fall 2021

Announcements/check-in

Slowing down the pace for the next 1.5 weeks In response to very helpful feedback, I've decided to push hypothesis testing and inference to *after the midterm exam*.

- No more assignments before the midterm
- Tuesday 10/26: Review. Please come with questions.

Announcements/check-in

Slowing down the pace for the next 1.5 weeks In response to very helpful feedback, I've decided to push hypothesis testing and inference to *after the midterm exam*.

- No more assignments before the midterm
- Tuesday 10/26: Review. Please come with questions.
- Assignment #2: Grades posted. Focus on interpretation!

Announcements/check-in

Slowing down the pace for the next 1.5 weeks In response to very helpful feedback, I've decided to push hypothesis testing and inference to *after the midterm exam*.

- No more assignments before the midterm
- Tuesday 10/26: Review. Please come with questions.
- Assignment #2: Grades posted. Focus on interpretation!
- Assignment #3: Grading TBD, answer key mid-week

Announcements/check-in

Slowing down the pace for the next 1.5 weeks In response to very helpful feedback, I've decided to push hypothesis testing and inference to *after the midterm exam*.

- No more assignments before the midterm
- Tuesday 10/26: Review. Please come with questions.
- Assignment #2: Grades posted. Focus on interpretation!
- Assignment #3: Grading TBD, answer key mid-week
- Reminder: Office hours in the Pine Room (Bren Hall 3526)

Midterm Exam

Two parts:

Midterm Exam

Two parts:

Part 1: Short answer questions (~4)

- Focus on definitions of key concepts
- You should know key definitions (e.g., expectation/mean, median, variance, R^2, OLS slope and intercept formulas for simple linear regression)
- You do not need to memorize math rules (e.g., $var(ax + b) = a^2 var(x)$)
- Be able to interpret probability distributions, scatter plots, QQ-plots, boxplots

Midterm Exam

Two parts:

Part 2: Long answer questions (~2)

- Each question poses a data science problem and walks you through a set of analysis steps
- Very similar to assignments but focused on interpretation of existing code and output
- May include some minimal pseudo-coding

Today

Interpreting multiple linear regression

"All else equal", parallel slopes model

Today

Interpreting multiple linear regression

"All else equal", parallel slopes model

Omitted-variable bias

The challenge and some solutions

Today

Interpreting multiple linear regression

"All else equal", parallel slopes model

Omitted-variable bias

The challenge and some solutions

Adjusted R^2

Correction to coefficient of determination

Today

Interpreting multiple linear regression

"All else equal", parallel slopes model

Omitted-variable bias

The challenge and some solutions

Adjusted R^2

Correction to coefficient of determination

Interaction effects

Implementation and interpretation

Today

Interpreting multiple linear regression

"All else equal", parallel slopes model

Omitted-variable bias

The challenge and some solutions

Adjusted R^2

Correction to coefficient of determination

Interaction effects

Implementation and interpretation

Multicollinearity

Problems and (some) solutions

Multiple linear regression

More explanatory variables

We're moving from **simple linear regression** (one outcome variable and one explanatory variable)

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

More explanatory variables

We're moving from **simple linear regression** (one outcome variable and one explanatory variable)

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

to the land of **multiple linear regression** (one outcome variable and multiple explanatory variables)

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$$

More explanatory variables

We're moving from **simple linear regression** (one outcome variable and one explanatory variable)

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

to the land of **multiple linear regression** (one outcome variable and multiple explanatory variables)

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$$

Why?

More explanatory variables

We're moving from **simple linear regression** (one outcome variable and one explanatory variable)

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

to the land of **multiple linear regression** (one outcome variable and multiple explanatory variables)

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$$

Why? We can better explain the variation in y, improve predictions, avoid omitted-variable bias (i.e., second assumption needed for unbiased OLS estimates), ...

More explanatory variables

Multiple linear regression...

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$$

More explanatory variables

Multiple linear regression…

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$$

… raises many questions:

More explanatory variables

Multiple linear regression...

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$$

... raises many questions:

- Which x's should I include? This is the problem of "model selection".

More explanatory variables

Multiple linear regression...

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$$

... raises many questions:

- Which x's should I include? This is the problem of "model selection".
- How does my interpretation of β_1 change?

More explanatory variables

Multiple linear regression...

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$$

... raises many questions:

- Which x's should I include? This is the problem of "model selection".
- How does my interpretation of β_1 change?
- What if my x's interact with each other? E.g., race and gender, temperature and rainfall.

More explanatory variables

Multiple linear regression...

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$$

... raises many questions:

- Which x's should I include? This is the problem of "model selection".
- How does my interpretation of β_1 change?
- What if my x's interact with each other? E.g., race and gender, temperature and rainfall.
- How do I measure model fit now?

More explanatory variables

Multiple linear regression...

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$$

... raises many questions:

- Which x's should I include? This is the problem of "model selection".
- How does my interpretation of β_1 change?
- What if my x's interact with each other? E.g., race and gender, temperature and rainfall.
- How do I measure model fit now?

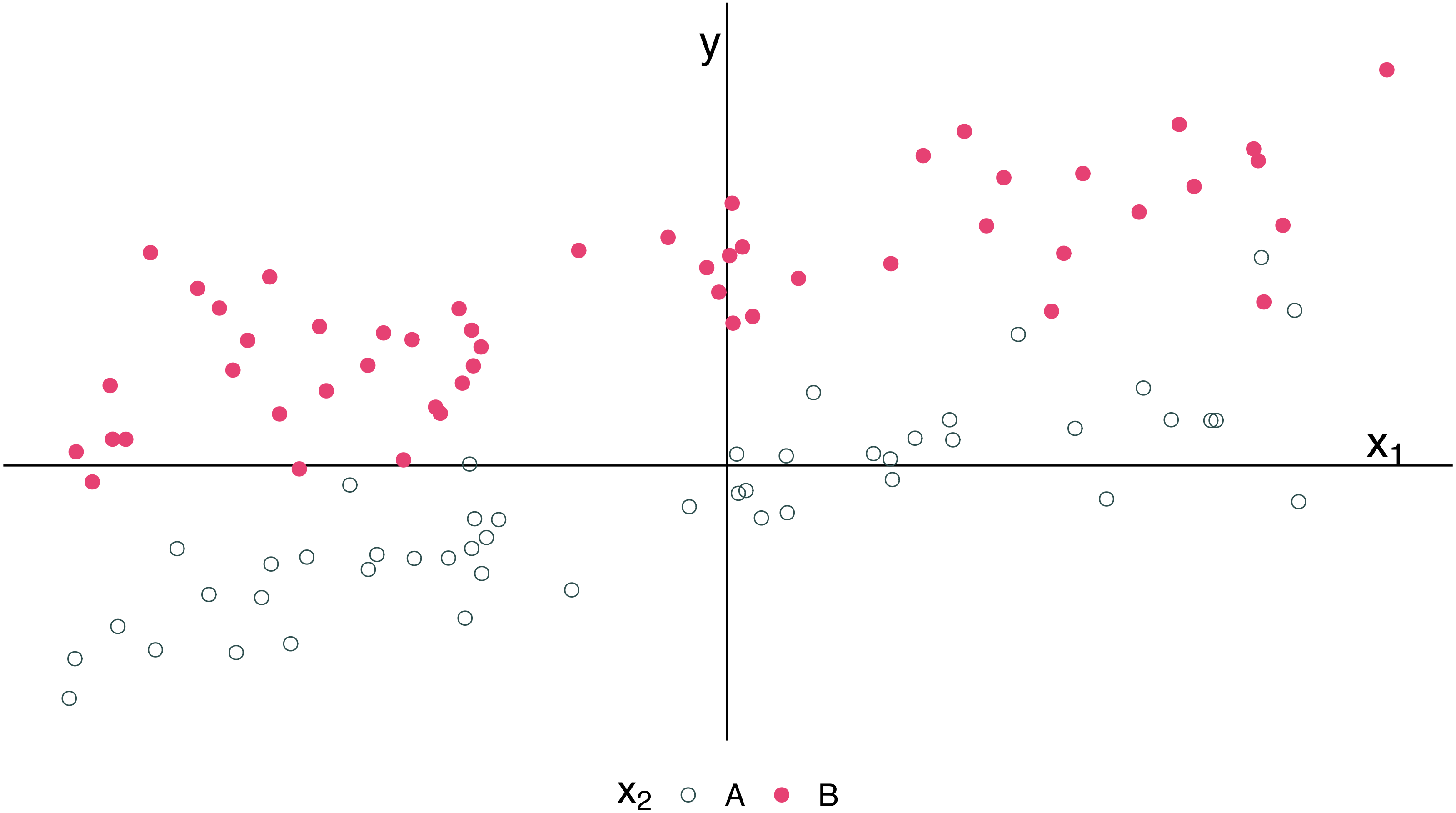
We will dig into each of these here, and you will see these questions in other MEDS courses

Multiple regression

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$ x_1 is continuous x_2 is categorical

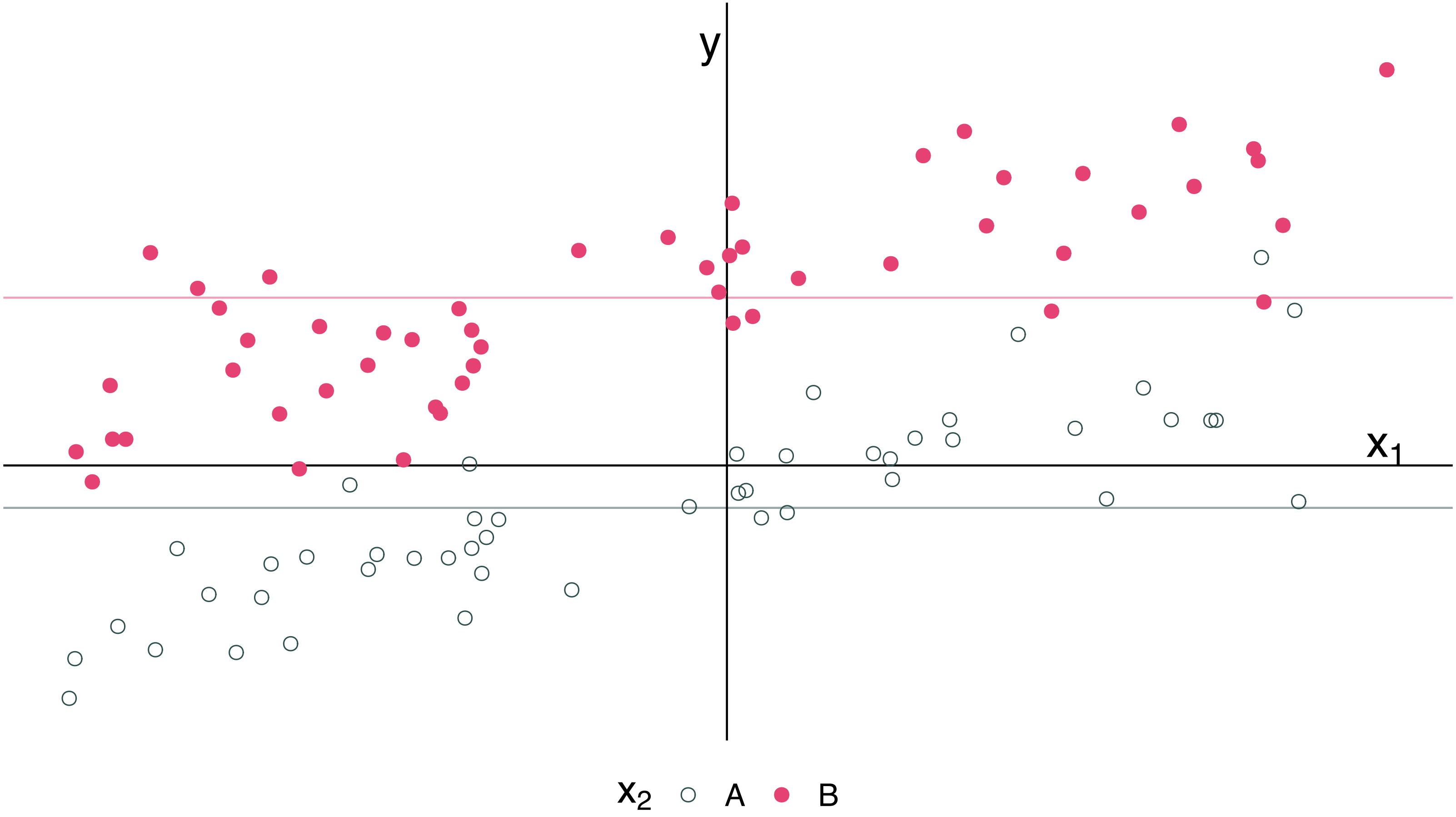
Multiple regression

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$ x_1 is continuous x_2 is categorical



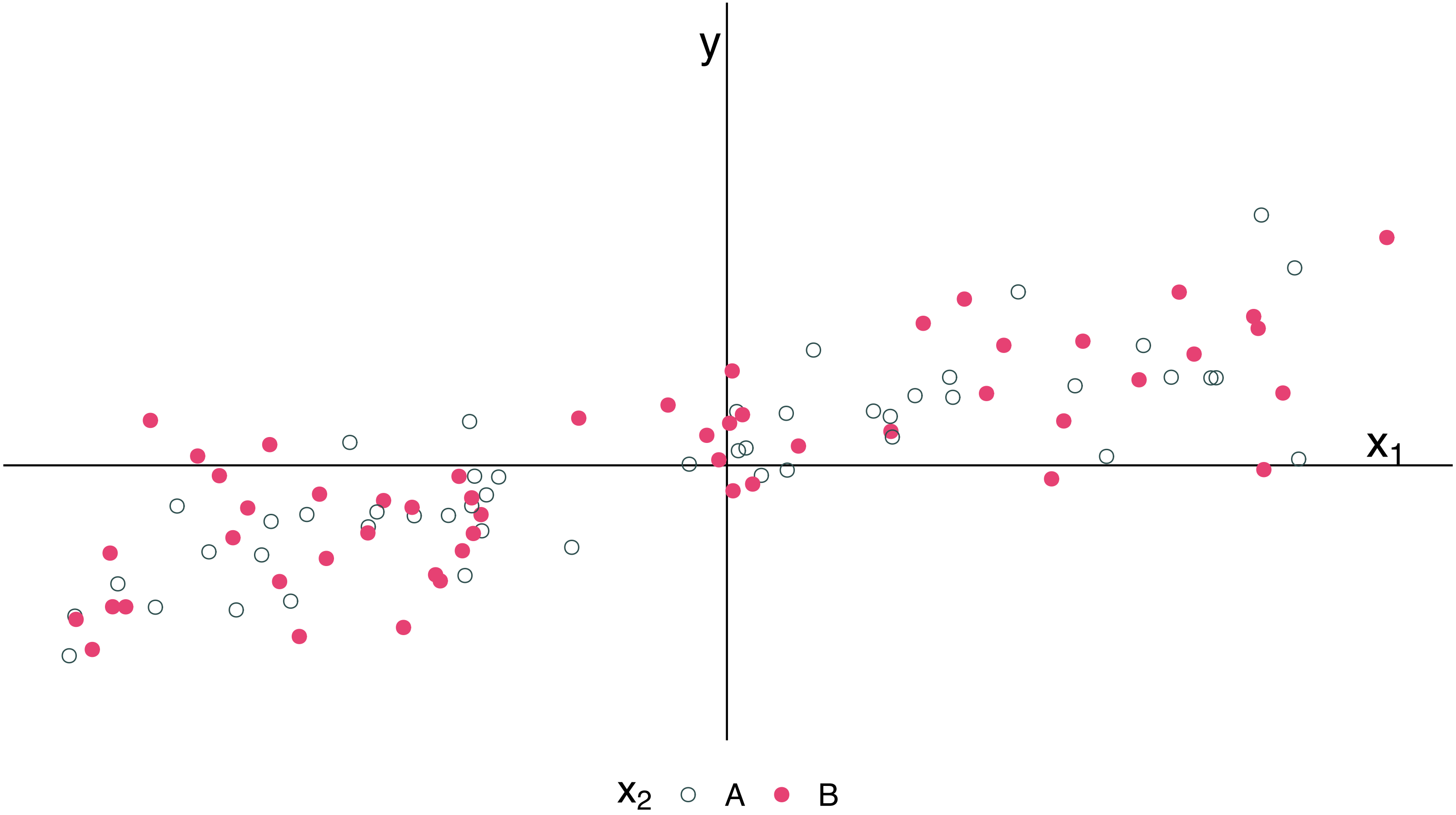
Multiple regression

The intercept and categorical variable x_2 control for the groups' means.



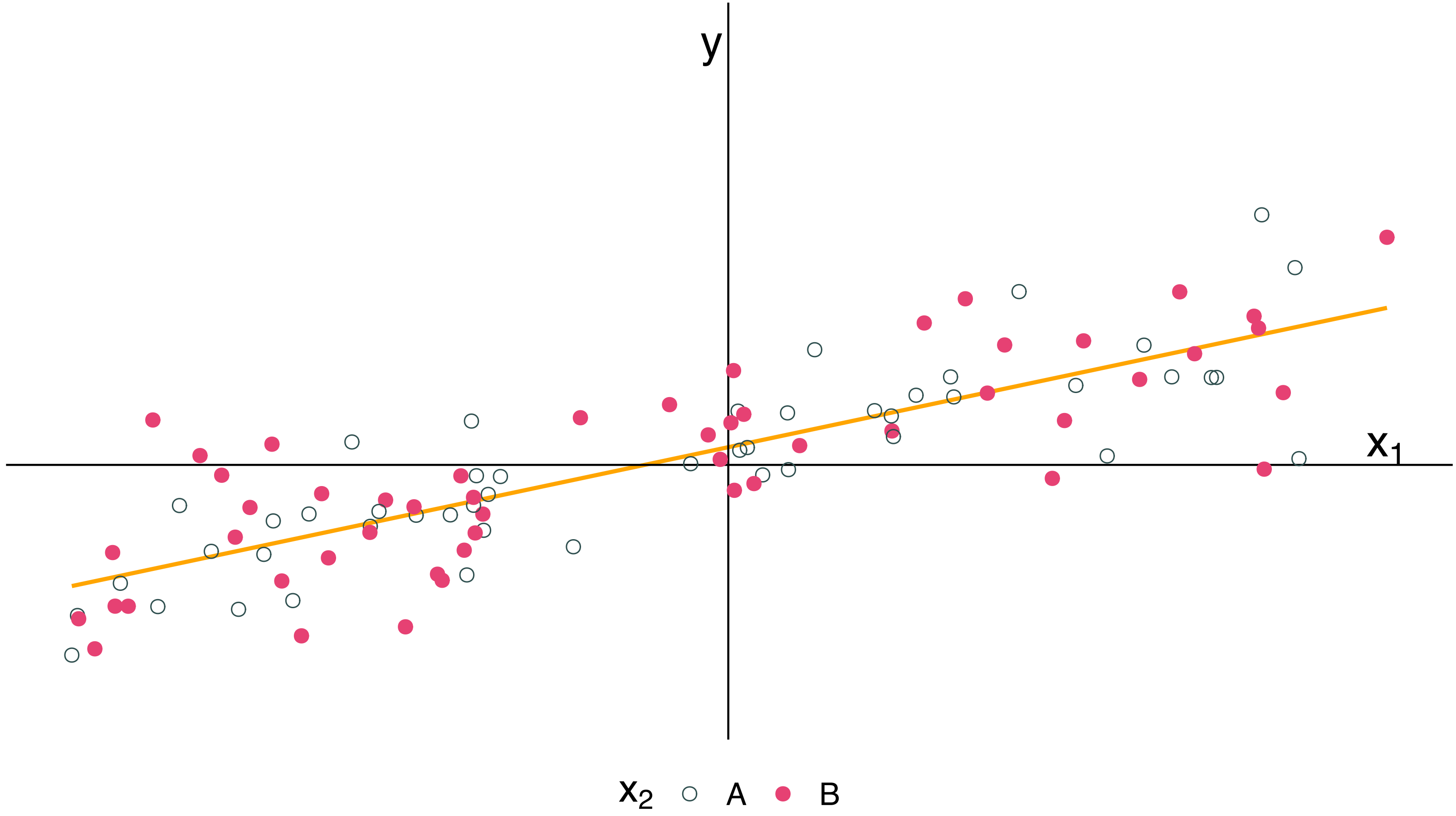
Multiple regression

With groups' means removed:



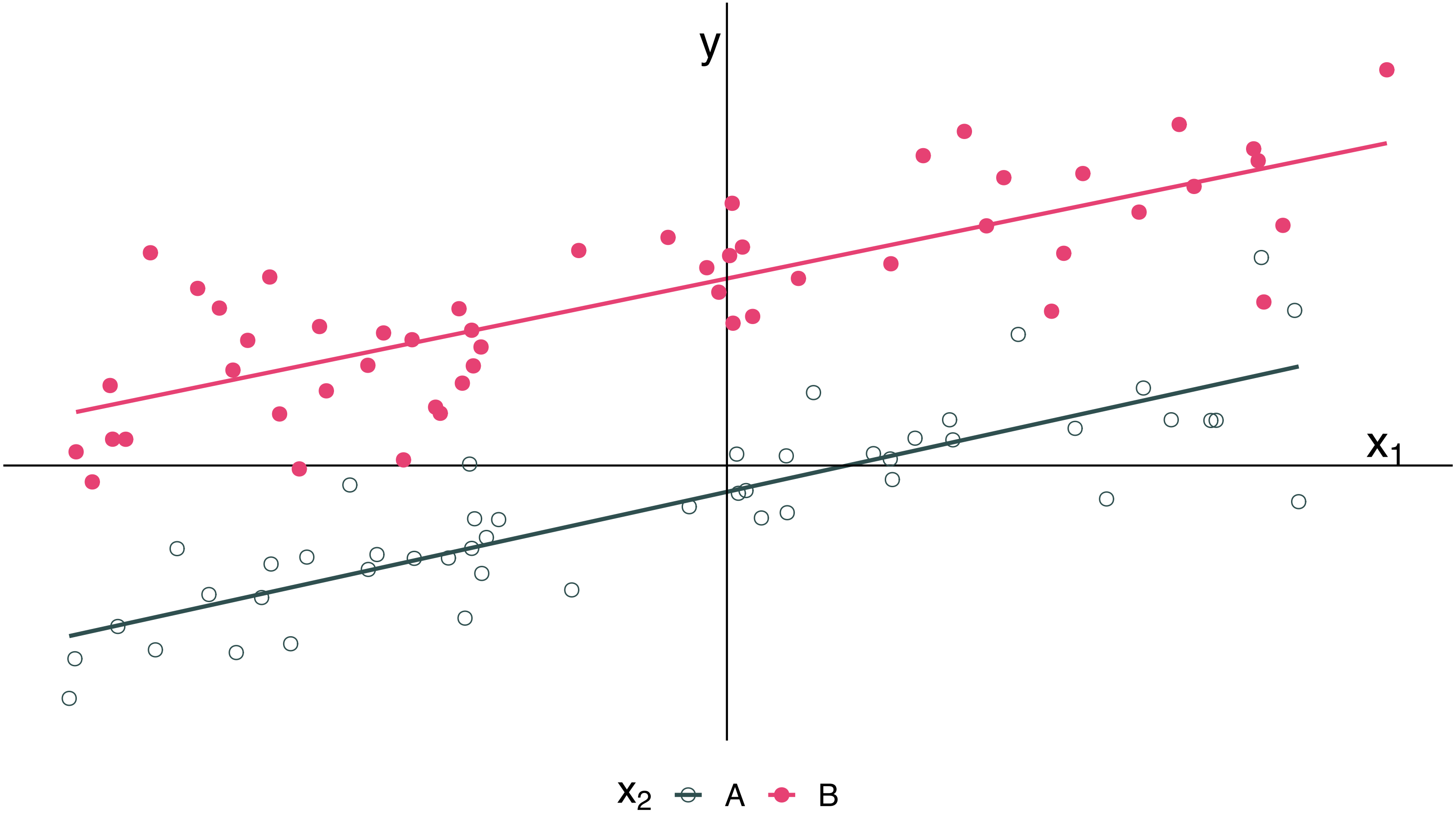
Multiple regression

$\hat{\beta}_1$ estimates the relationship between y and x_1 after controlling for x_2.



Multiple regression

This regression gives the "parallel slopes" model:



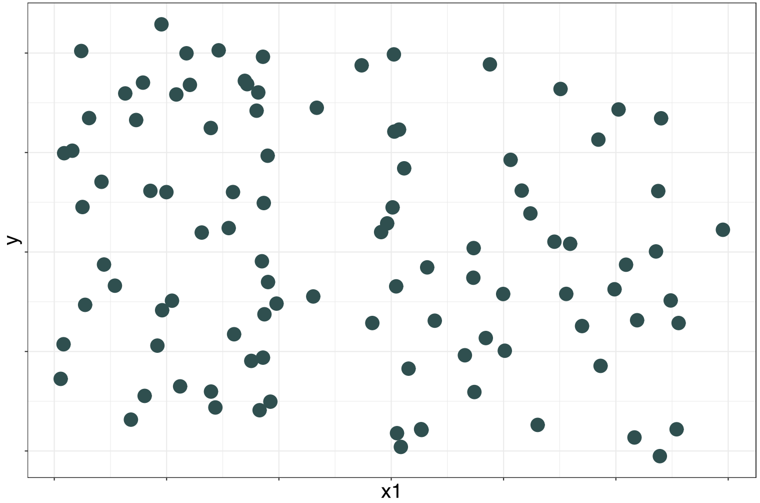
Multiple regression

More generally, how do we think about multiple explanatory variables?

Multiple regression

More generally, how do we think about multiple explanatory variables?

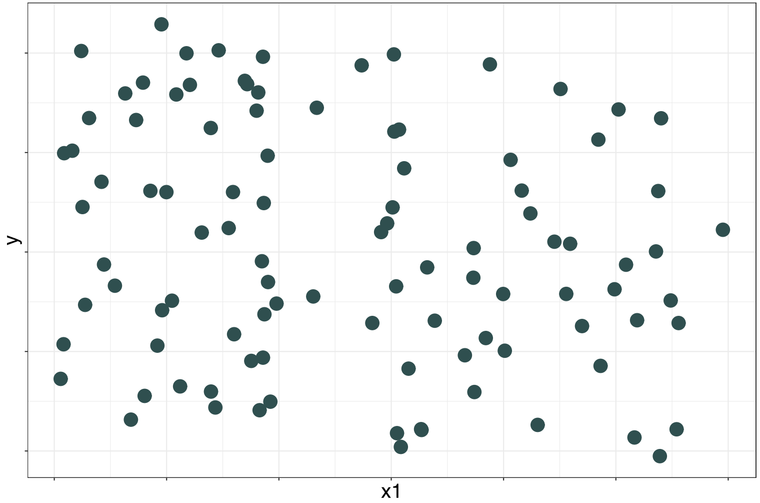
Suppose $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$ where x_1 and x_2 are both continuous numerical variables

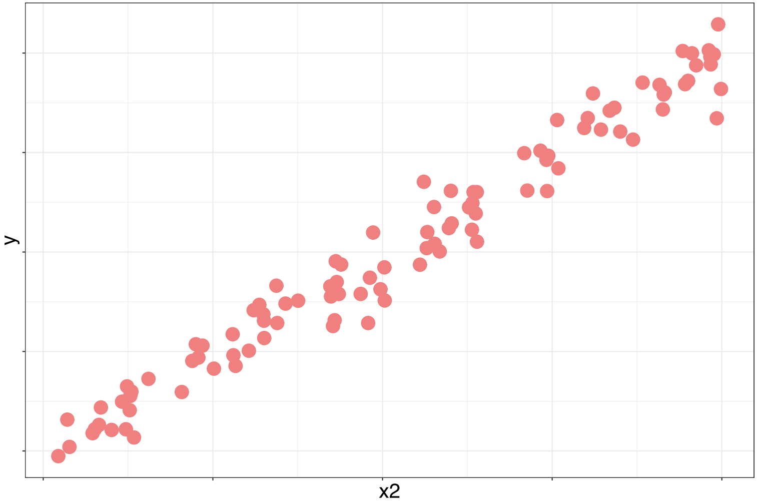


Multiple regression

More generally, how do we think about multiple explanatory variables?

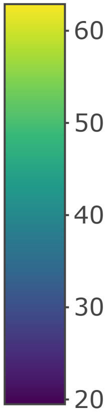
Suppose $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$ where x_1 and x_2 are both continuous numerical variables





Multiple regression

More generally, how do we think about multiple explanatory variables?



Multiple regression

With **many** explanatory variables, we visualizing relationships means thinking about **hyperplanes** 🤯

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + u_i$$

Math notation looks very similar to simple linear regression, but *conceptually* and *visually* multiple regression is **very different**

Multiple regression

Interpretation of coefficients

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + u_i$$

Multiple regression

Interpretation of coefficients

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + u_i$$

- β_k tells us the change in y due to a one unit change in x_k when **all other variables are held constant**

Multiple regression

Interpretation of coefficients

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + u_i$$

- β_k tells us the change in y due to a one unit change in x_k when **all other variables are held constant**

- This is an "all else equal" interpretation

Multiple regression

Interpretation of coefficients

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + u_i$$

- β_k tells us the change in y due to a one unit change in x_k when **all other variables are held constant**

- This is an "all else equal" interpretation

- E.g., how much do wages increase with one more year of education, *holding gender fixed*?

Multiple regression

Interpretation of coefficients

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + u_i$$

- β_k tells us the change in y due to a one unit change in x_k when **all other variables are held constant**
- This is an "all else equal" interpretation
- E.g., how much do wages increase with one more year of education, *holding gender fixed*?
- E.g., how much does ozone increase when temperature rises, *holding NOx emissions fixed*?

Tradeoffs

There are tradeoffs to consider as we add/remove variables:

Fewer variables

- Generally explain less variation in y
- Provide simple interpretations and visualizations (*parsimonious*)
- May need to worry about omitted-variable bias

More variables

- More likely to find *spurious* relationships (statistically significant due to chance—does not reflect a true, population-level relationship)
- More difficult to interpret the model
- You may still miss important variables—still omitted-variable bias

Omitted-variable bias

Omitted-variable bias

You will study this in much more depth in EDS 241, but here's a primer.

Omitted-variable bias (OVB) arises when we omit a variable that

1. affects our outcome variable y

2. correlates with an explanatory variable x_j

As it's name suggests, this situation leads to bias in our estimate of β_j. In particular, it violates Assumption 2 of OLS (see week 03 slides).

Omitted-variable bias

You will study this in much more depth in EDS 241, but here's a primer.

Omitted-variable bias (OVB) arises when we omit a variable that

1. affects our outcome variable y

2. correlates with an explanatory variable x_j

As it's name suggests, this situation leads to bias in our estimate of β_j. In particular, it violates Assumption 2 of OLS (see week 03 slides).

Note: OVB Is not exclusive to multiple linear regression, but it does require multiple variables affect y.

Omitted-variable bias

Example

Let's imagine a simple model for the amount individual i gets paid

$$\text{Pay}_i = \beta_0 + \beta_1 \text{School}_i + \beta_2 \text{Male}_i + u_i$$

where

- School_i gives i's years of schooling
- Male_i denotes an indicator variable for whether individual i is male.

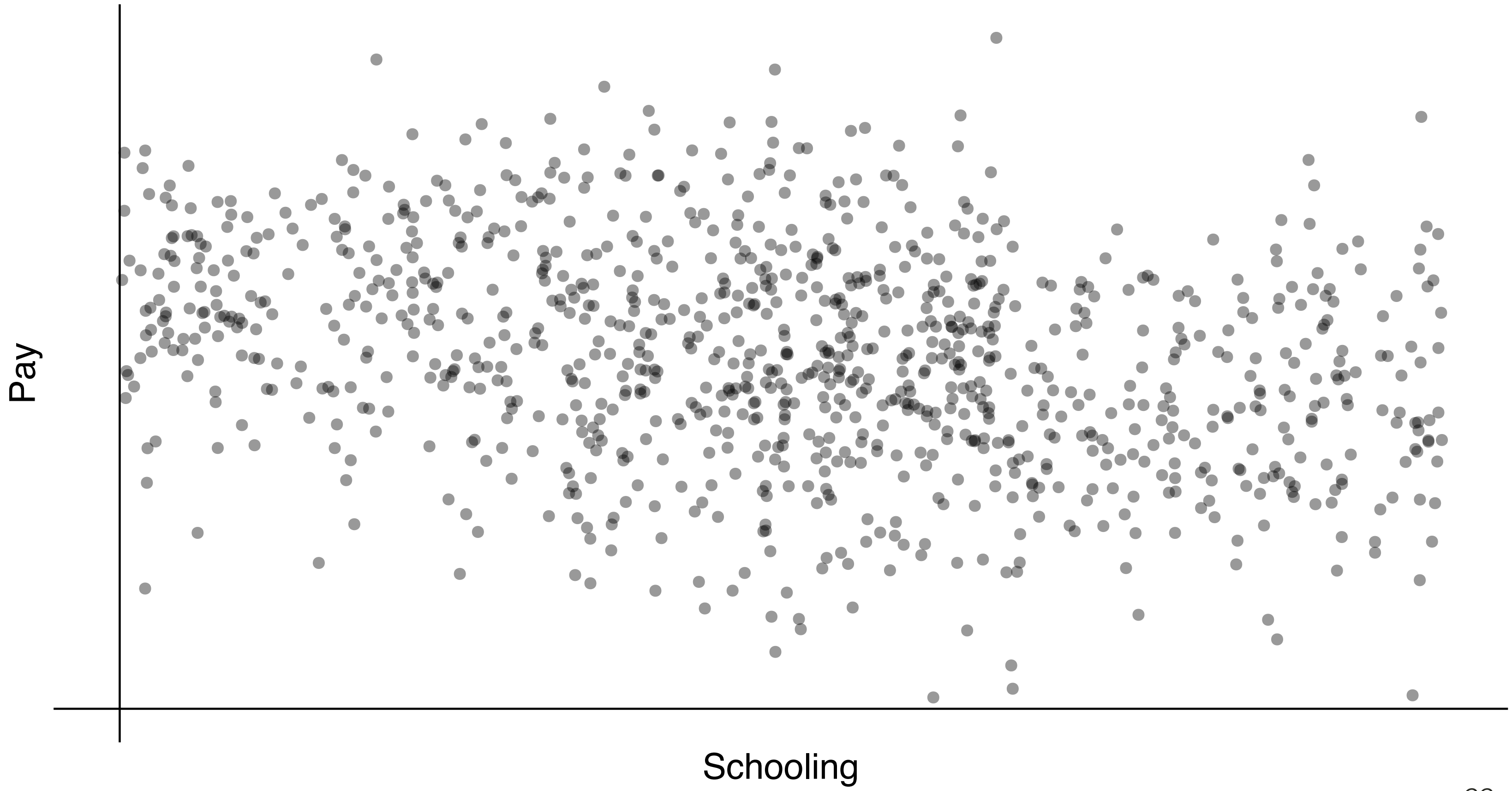
thus

- β_1: the returns to an additional year of schooling (*ceteris paribus*)
- β_2: the premium for being male (*ceteris paribus*)
  If $\beta_2 > 0$, then there is discrimination against women—receiving less pay based upon gender.

Omitted-variable bias

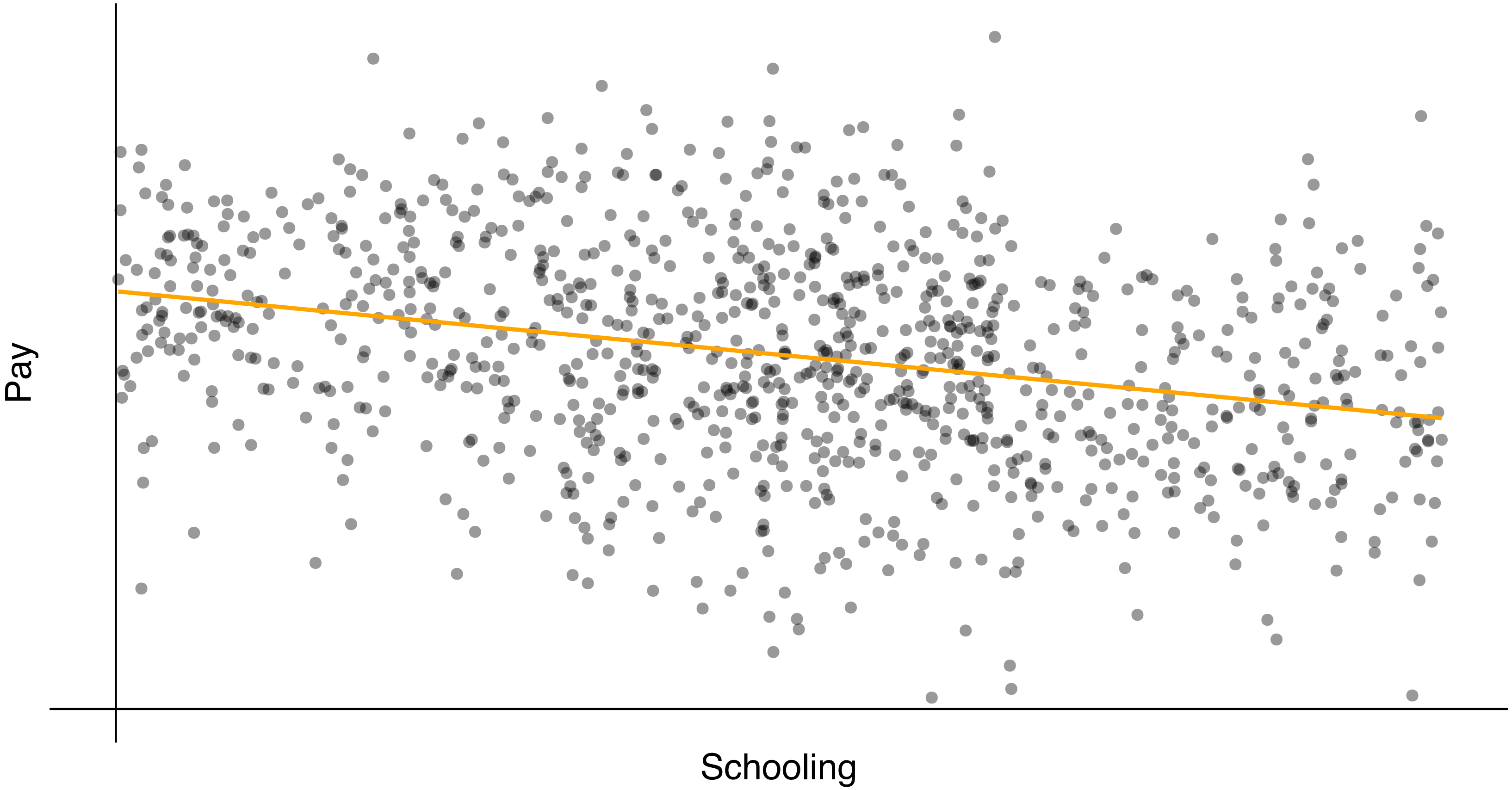
Example, continued: $\text{Pay}_i = 20 + 0.5 \times \text{School}_i + 10 \times \text{Male}_i + u_i$

The relationship between pay and schooling.



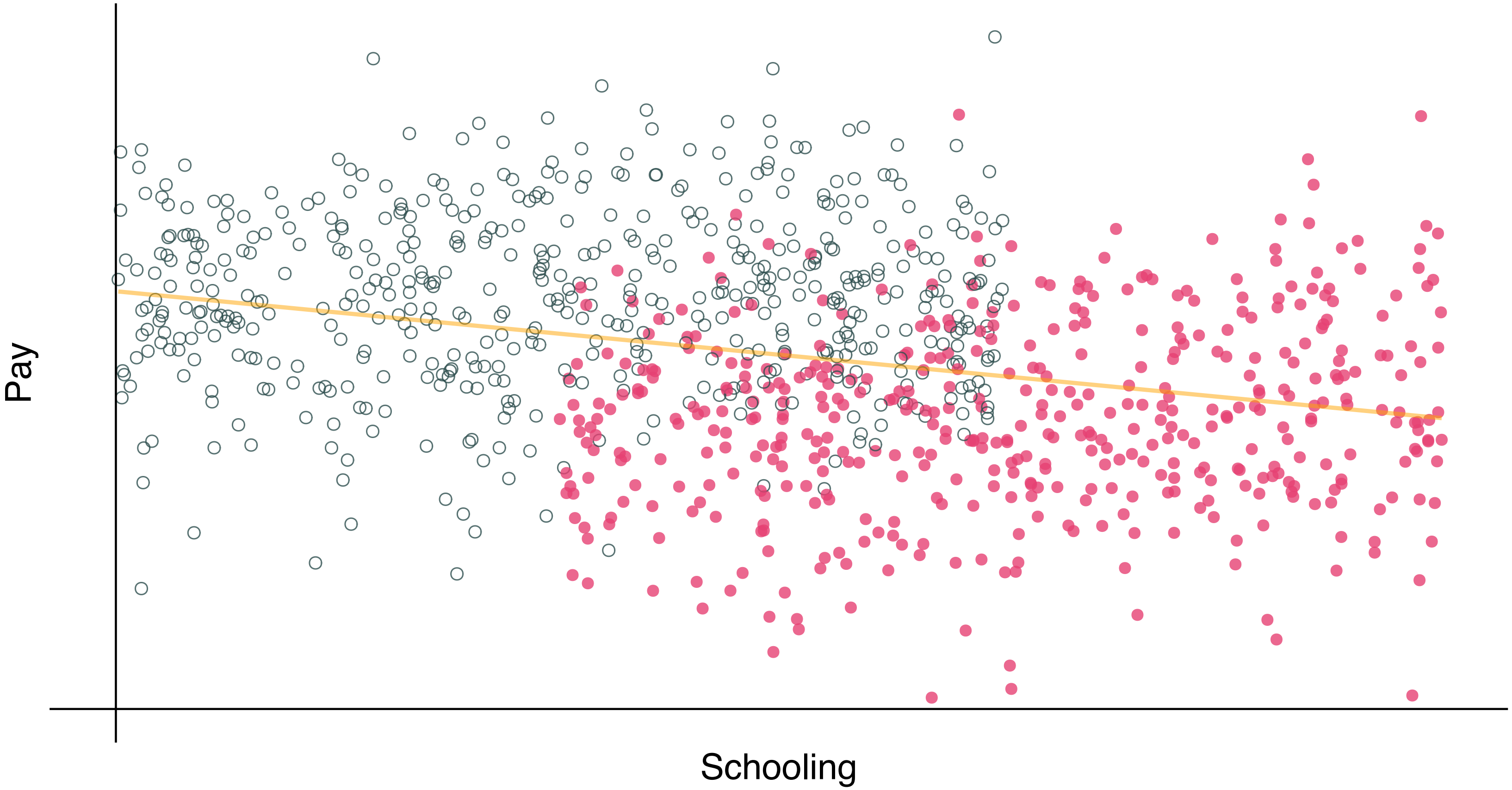
Omitted-variable bias

Biased regression estimate: $\widehat{\mathbf{Pay}}_i = \mathbf{32.2} + \mathbf{-1.1} \times \mathbf{School}_i$



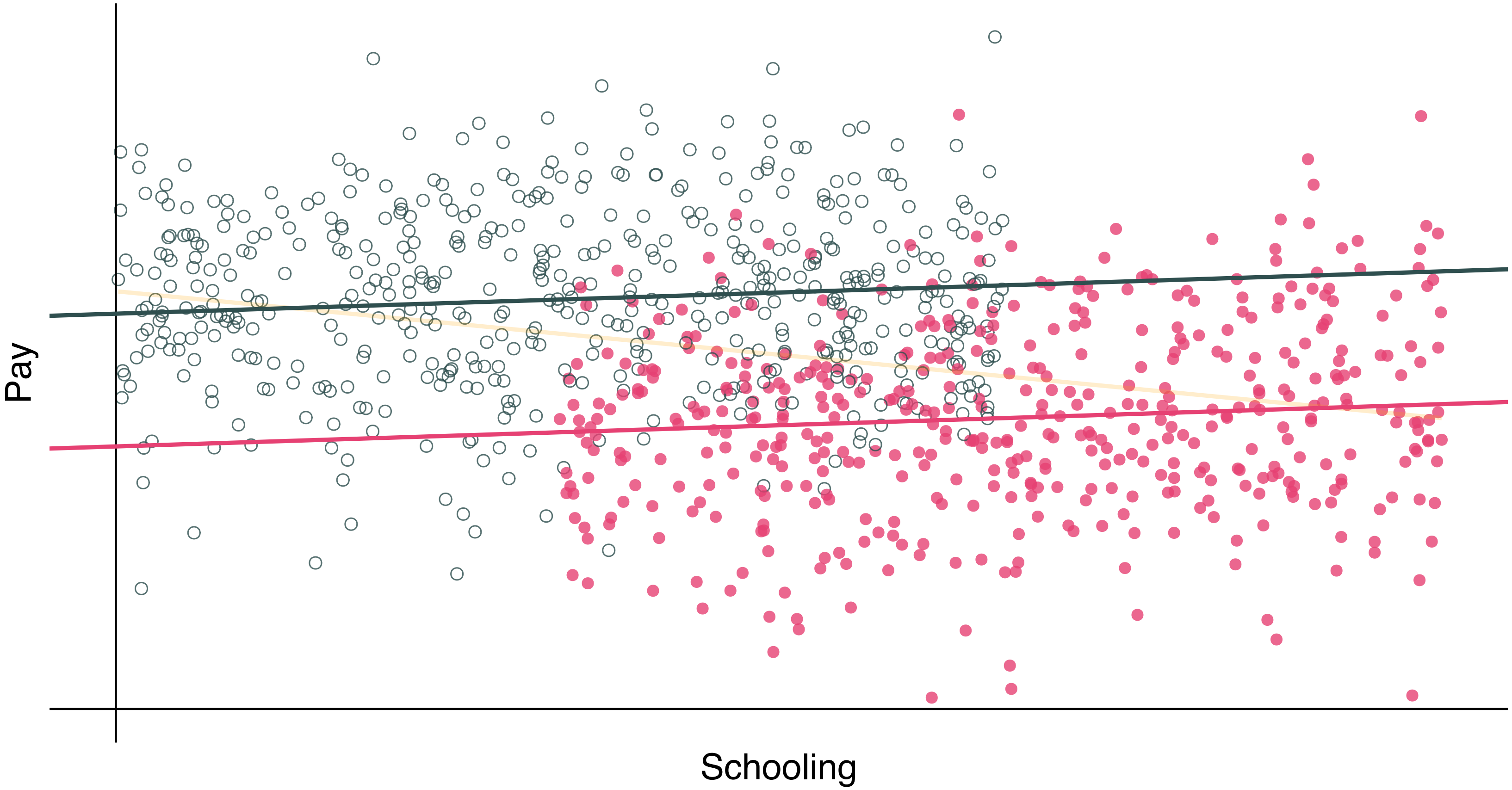
Omitted-variable bias

Recalling the omitted variable: Gender (**female** and **male**)



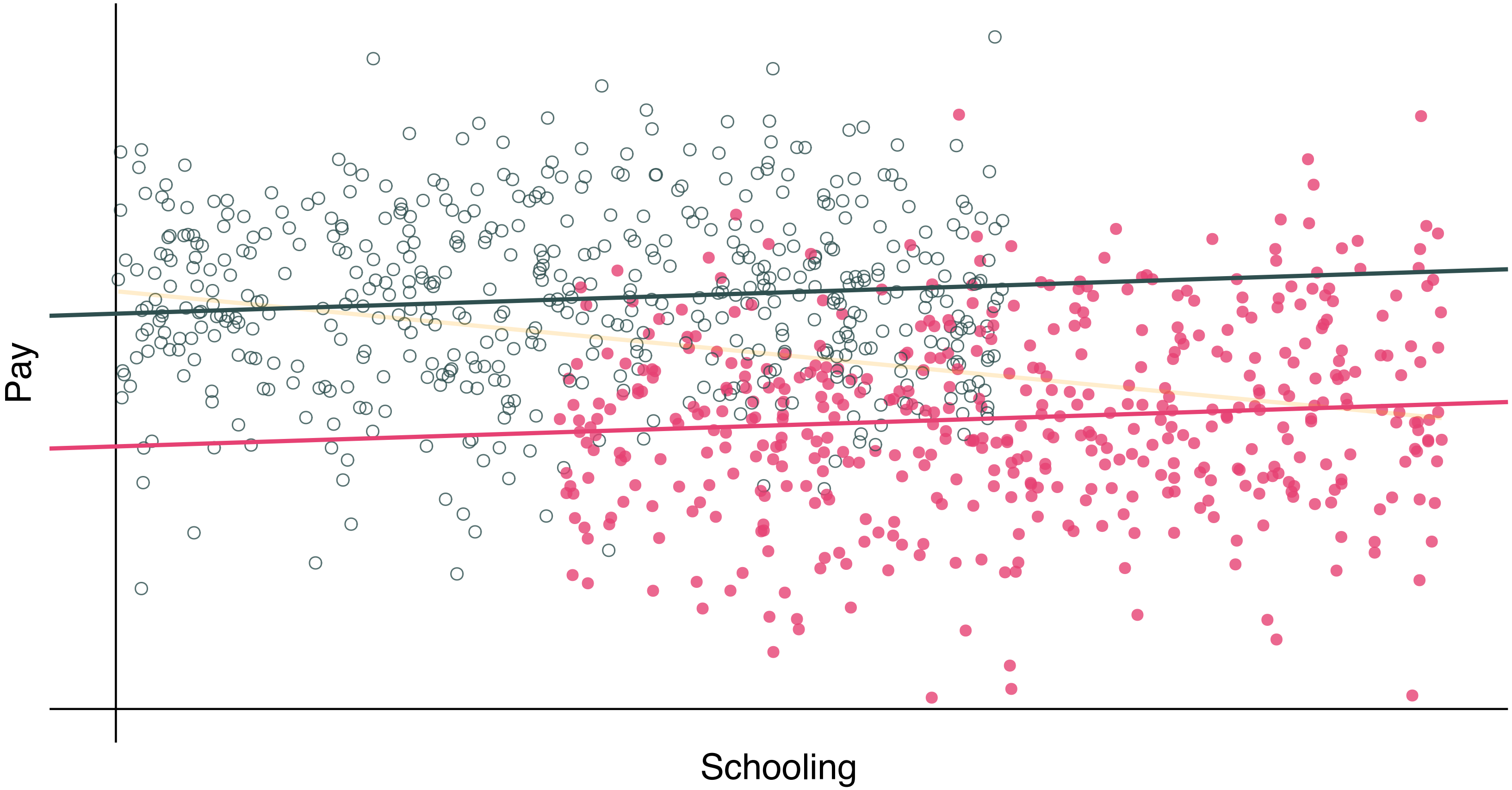
Omitted-variable bias

Recalling the omitted variable: Gender (**female** and **male**)



Omitted-variable bias

Unbiased regression estimate: $\widehat{\text{Pay}}_i = 20.3 + 0.4 \times \text{School}_i + 10.2 \times \text{Male}_i$



Adjusted R^2

Nonlinear transformations

Our linearity assumption requires that **parameters enter linearly** (*i.e.*, the β_k multiplied by variables)

We allow nonlinear relationships between y and the explanatory variables x.

Nonlinear transformations

Our linearity assumption requires that **parameters enter linearly** (*i.e.*, the β_k multiplied by variables)

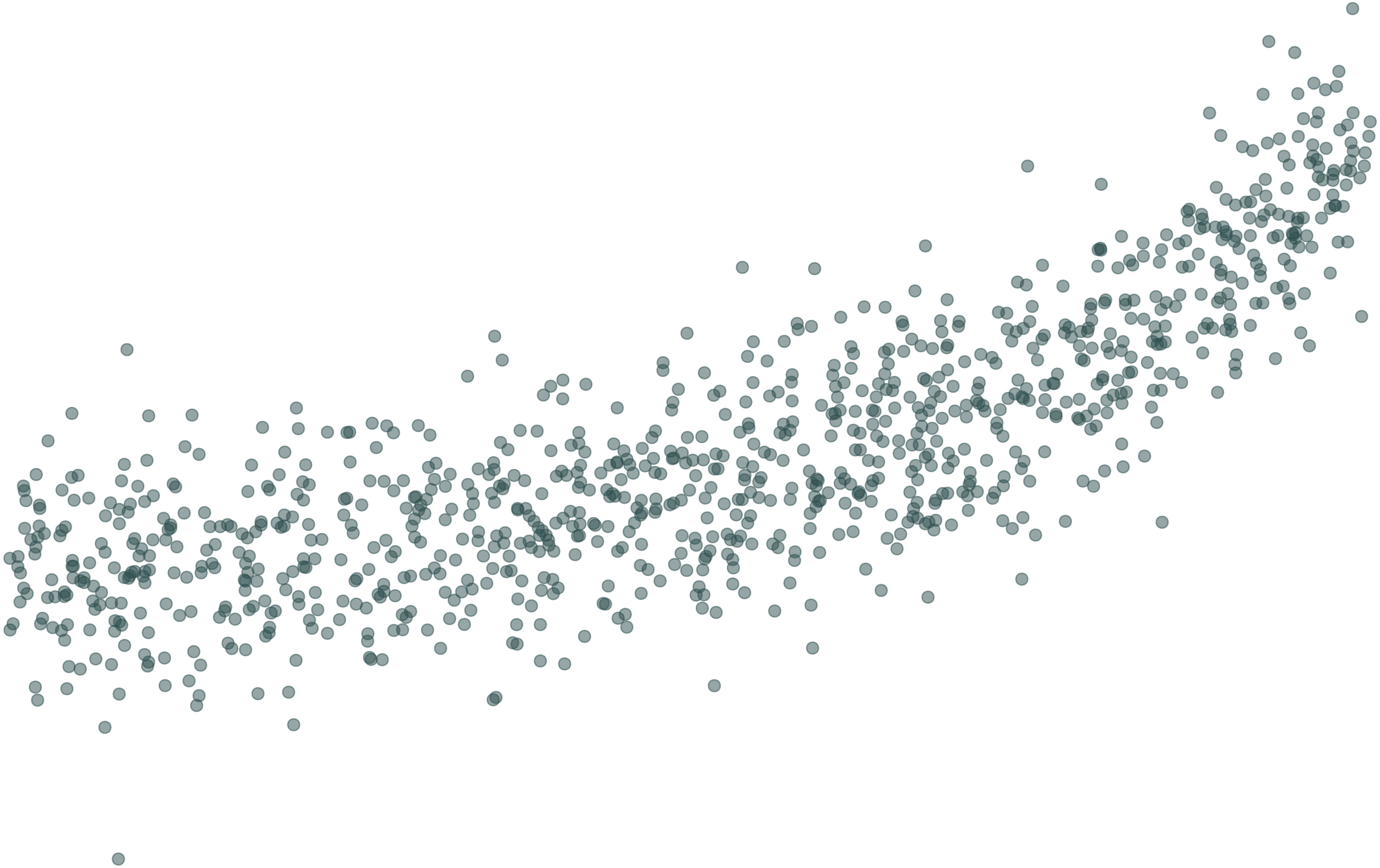
We allow nonlinear relationships between y and the explanatory variables x.

Examples

- **Polynomials** and **interactions:** $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2^2 + \beta_5 \left(x_1 x_2\right) + u_i$
- **Exponentials** and **logs:** $\log(y_i) = \beta_0 + \beta_1 x_1 + \beta_2 e^{x_2} + u_i$
- **Indicators** and **thresholds:** $y_i = \beta_0 + \beta_1 x_1 + \beta_2 \, \mathbb{I}(x_1 \geq 100) + u_i$

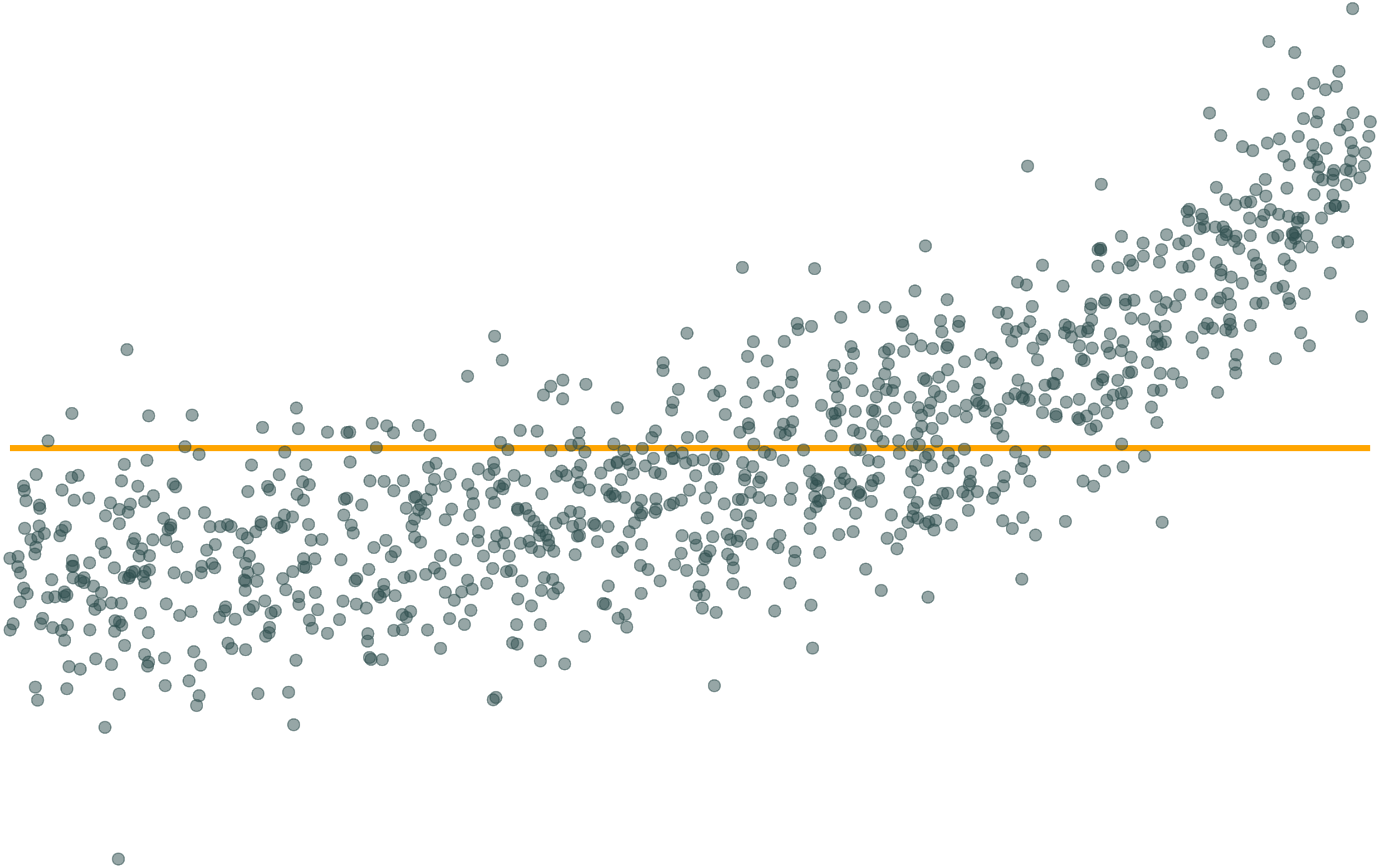
Nonlinear transformations

Transformation challenge: (literally) infinite possibilities. What do we pick?



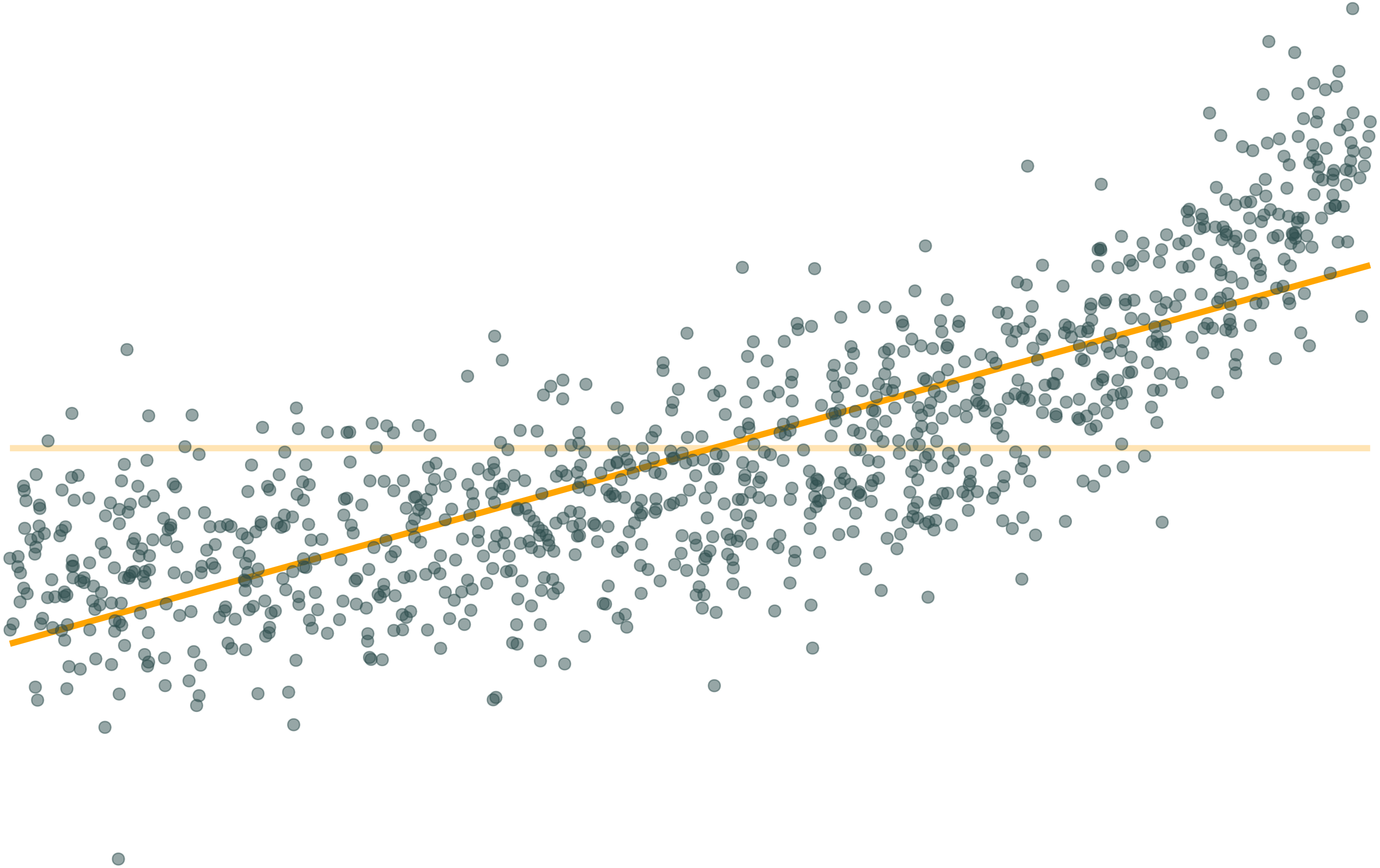
Nonlinear transformations

$y_i = \beta_0 + u_i$



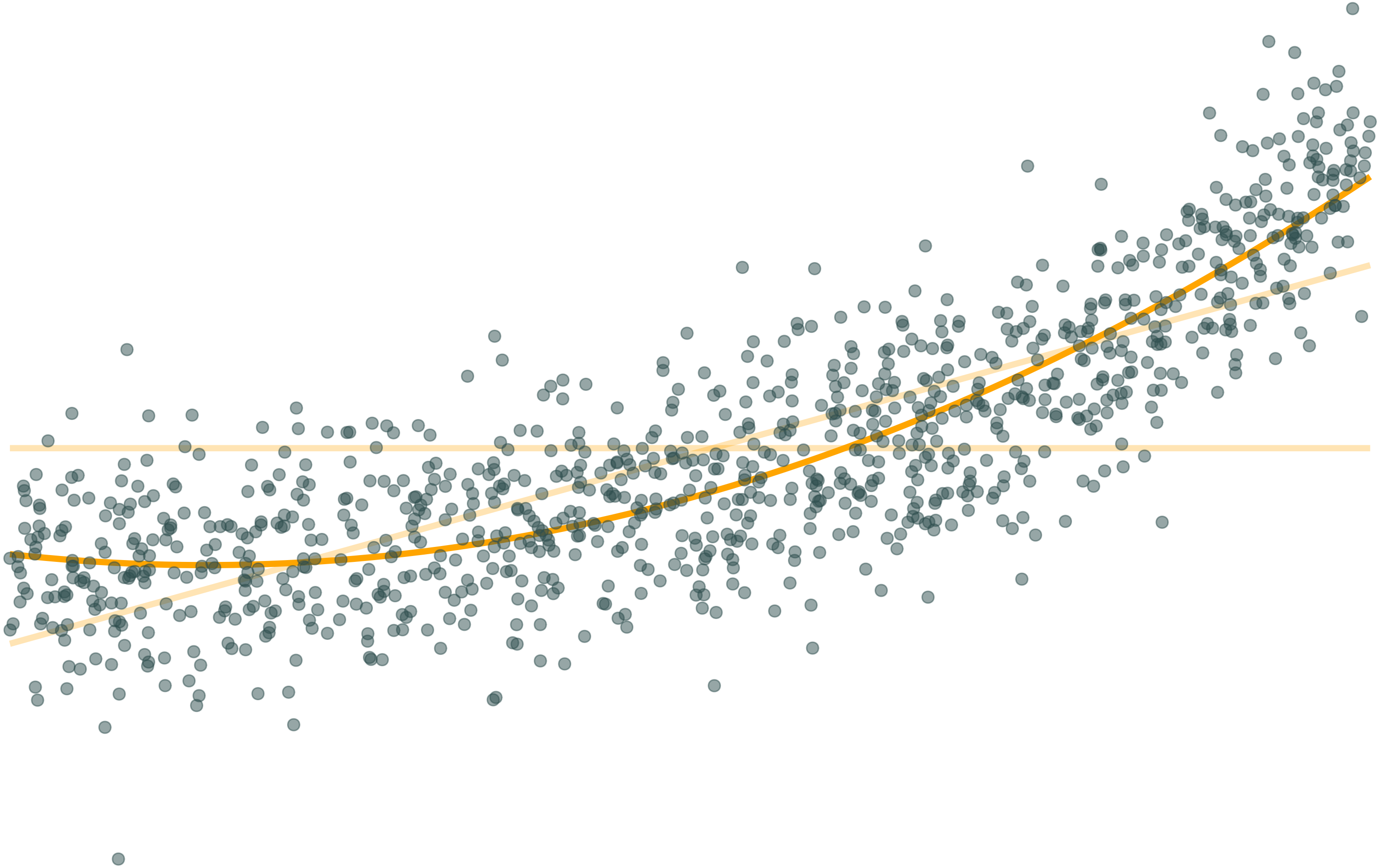
Nonlinear transformations

$y_i = \beta_0 + \beta_1 x + u_i$



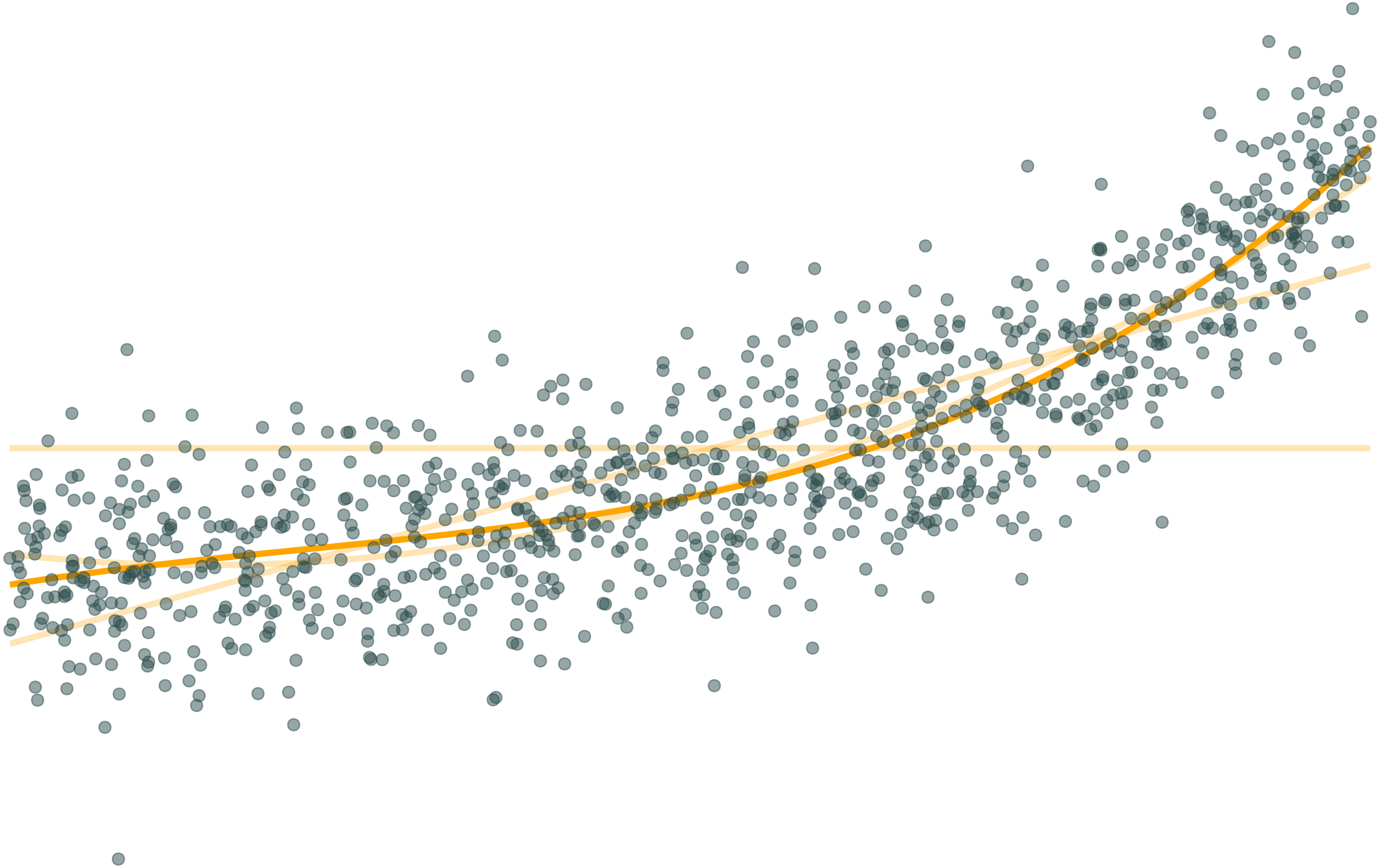
Nonlinear transformations

$y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + u_i$



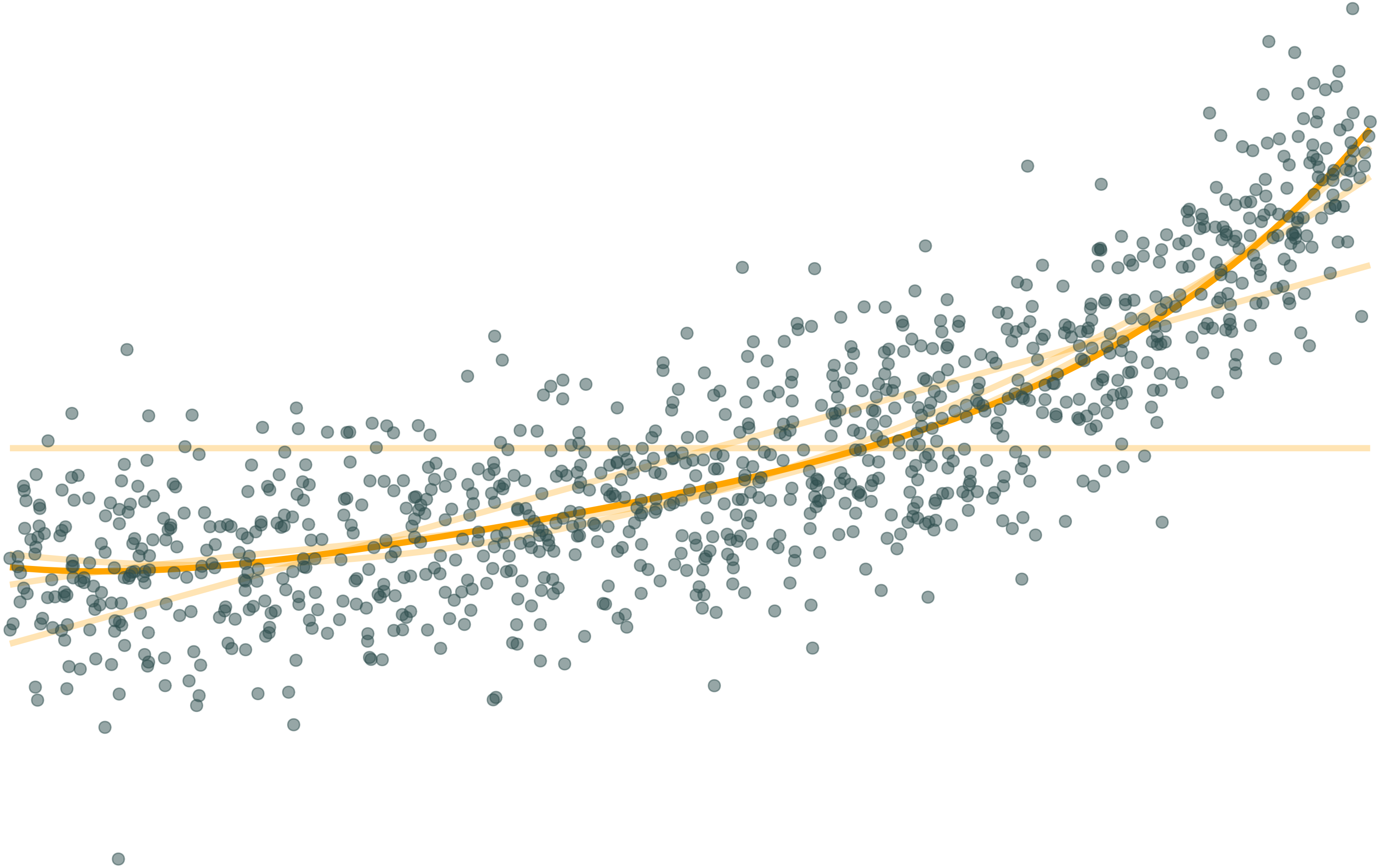
Nonlinear transformations

$y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + u_i$



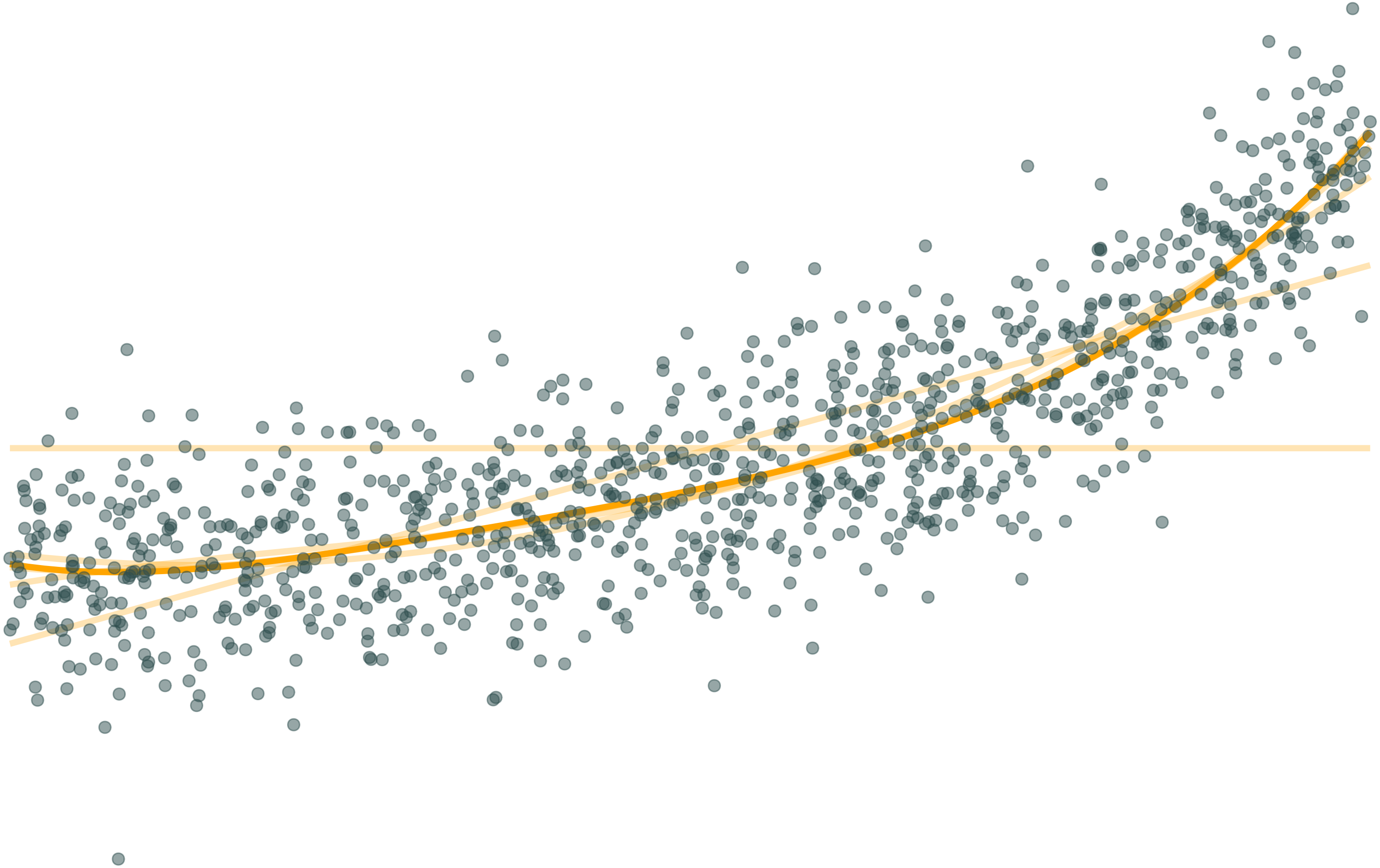
Nonlinear transformations

$$y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + u_i$$



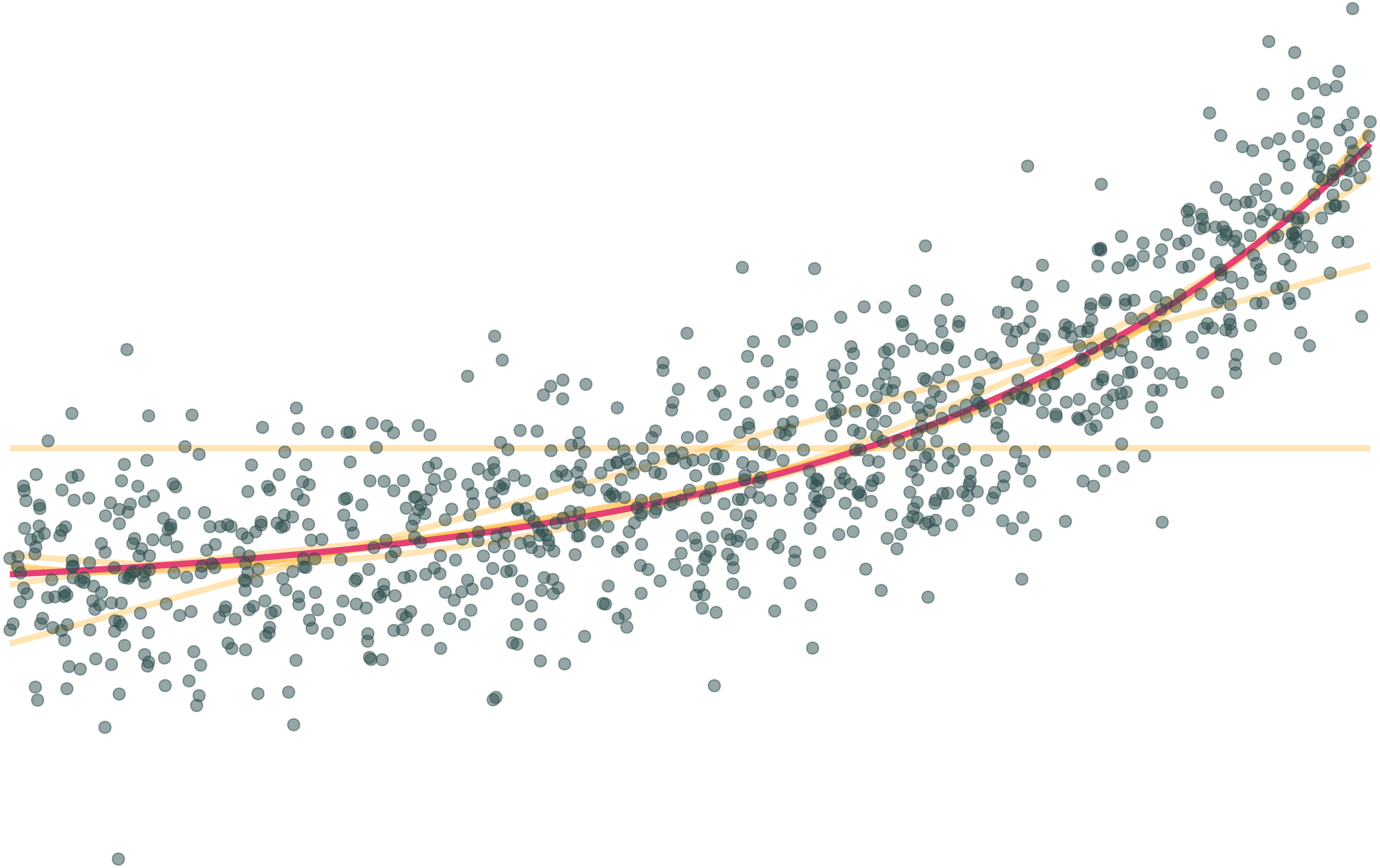
Nonlinear transformations

$$y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5 + u_i$$



Nonlinear transformations

Truth: $y_i = 2e^x + u_i$



Model fit with multiple regressors

Measures of *goodness of fit* try to analyze how well our model describes (*fits*) the data.

Model fit with multiple regressors

Measures of *goodness of fit* try to analyze how well our model describes (*fits*) the data.

Common measure: R^2 [R-squared] (*a.k.a.* coefficient of determination)

$$R^2 = 1 - \frac{\sum_i \left(y_i - \hat{y}_i\right)^2}{\sum_i \left(y_i - \overline{y}\right)^2} = 1 - \frac{\sum_i e_i^2}{\sum_i \left(y_i - \overline{y}\right)^2}$$

Recall $\sum_i \left(y_i - \hat{y}_i\right)^2 = \sum_i e_i^2$ is the "sum of squared errors".

Model fit with multiple regressors

Measures of *goodness of fit* try to analyze how well our model describes (*fits*) the data.

Common measure: R^2 [R-squared] (*a.k.a.* coefficient of determination)

$$R^2 = 1 - \frac{\sum_i \left(y_i - \hat{y}_i\right)^2}{\sum_i \left(y_i - \overline{y}\right)^2} = 1 - \frac{\sum_i e_i^2}{\sum_i \left(y_i - \overline{y}\right)^2}$$

Recall $\sum_i \left(y_i - \hat{y}_i\right)^2 = \sum_i e_i^2$ is the "sum of squared errors".

R^2 literally tells us the share of the variance in y our current models accounts for. Thus $0 \leq R^2 \leq 1$.

Model fit with multiple regressors

The problem: As we add variables to our model, R^2 *mechanically* increases.

Model fit with multiple regressors

The problem: As we add variables to our model, R^2 *mechanically* increases.

Intuition: Even if our added variable has *no true relation to* y, it can help lower e_i by fitting to the sampling noise

Model fit with multiple regressors

The problem: As we add variables to our model, R^2 *mechanically* increases.

Intuition: Even if our added variable has *no true relation to* y, it can help lower e_i by fitting to the sampling noise

One solution: Penalize for the number of variables, *e.g.*, adjusted R^2:

$$\overline{R}^2 = 1 - \frac{\sum_i \left(y_i - \hat{y}_i\right)^2/(n-k-1)}{\sum_i \left(y_i - \overline{y}\right)^2/(n-1)}$$

Note: Adjusted R^2 need not be between 0 and 1.

Model fit with multiple regressors

We often use measures of model fit (or model "performance") to help choose a regression model from among multiple possibilities

- Adjusted R^2 is just one of **many possible performance metrics**

Model fit with multiple regressors

We often use measures of model fit (or model "performance") to help choose a regression model from among multiple possibilities

- Adjusted R^2 is just one of **many possible performance metrics**
- For example, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Mean Squared Error (MSE), ...

Model fit with multiple regressors

We often use measures of model fit (or model "performance") to help choose a regression model from among multiple possibilities

- Adjusted R^2 is just one of **many possible performance metrics**
- For example, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Mean Squared Error (MSE), …
- Lots more on the topic of model selection in EDS 232 👀

Model fit with multiple regressors

We often use measures of model fit (or model "performance") to help choose a regression model from among multiple possibilities

- Adjusted R^2 is just one of **many possible performance metrics**
- For example, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Mean Squared Error (MSE), ...
- Lots more on the topic of model selection in EDS 232 👀
- Don't forget the *theory* behind your data science!

Interactions

Interactions

Interactions allow the effect of one variable to change based upon the level of another variable.

Examples

1. Does the effect of schooling on pay change by gender?
2. Does the effect of gender on pay change by race?
3. Does the effect of schooling on pay change by experience?

Interactions

Previously, we considered a model that allowed women and men to have different wages, but the model assumed the effect of school on pay was the same for everyone:

$$\mathbf{Pay}_i = \beta_0 + \beta_1\,\mathbf{School}_i + \beta_2\,\mathbf{Male}_i + u_i$$

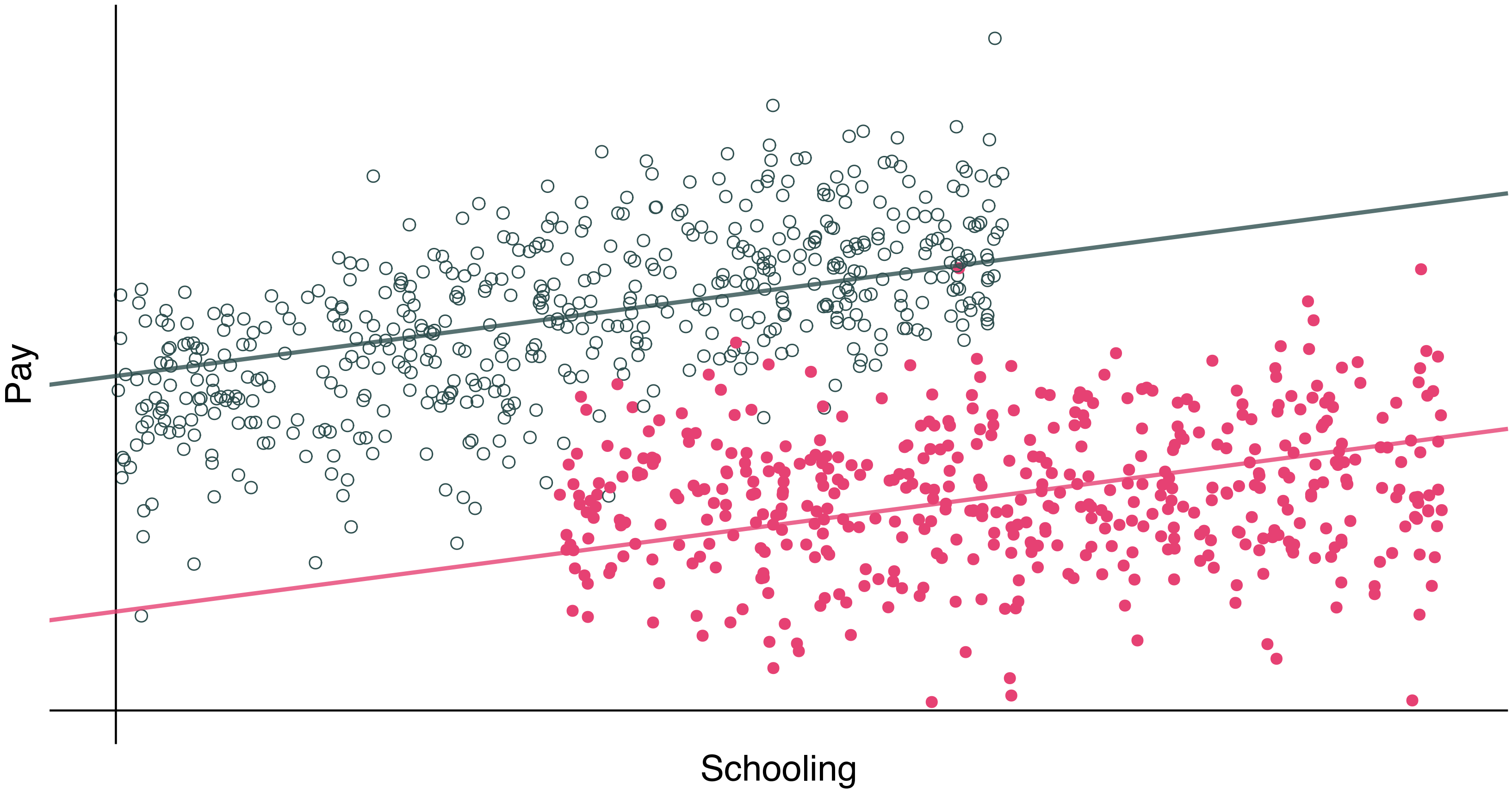
but we can also allow the effect of school to vary by gender:

$$\mathbf{Pay}_i = \beta_0 + \beta_1\,\mathbf{School}_i + \beta_2\,\mathbf{Male}_i + \beta_3\,\mathbf{School}_i \times \mathbf{Male}_i + u_i$$

Interactions

The model where schooling has the same effect for everyone (**F** and **M**):

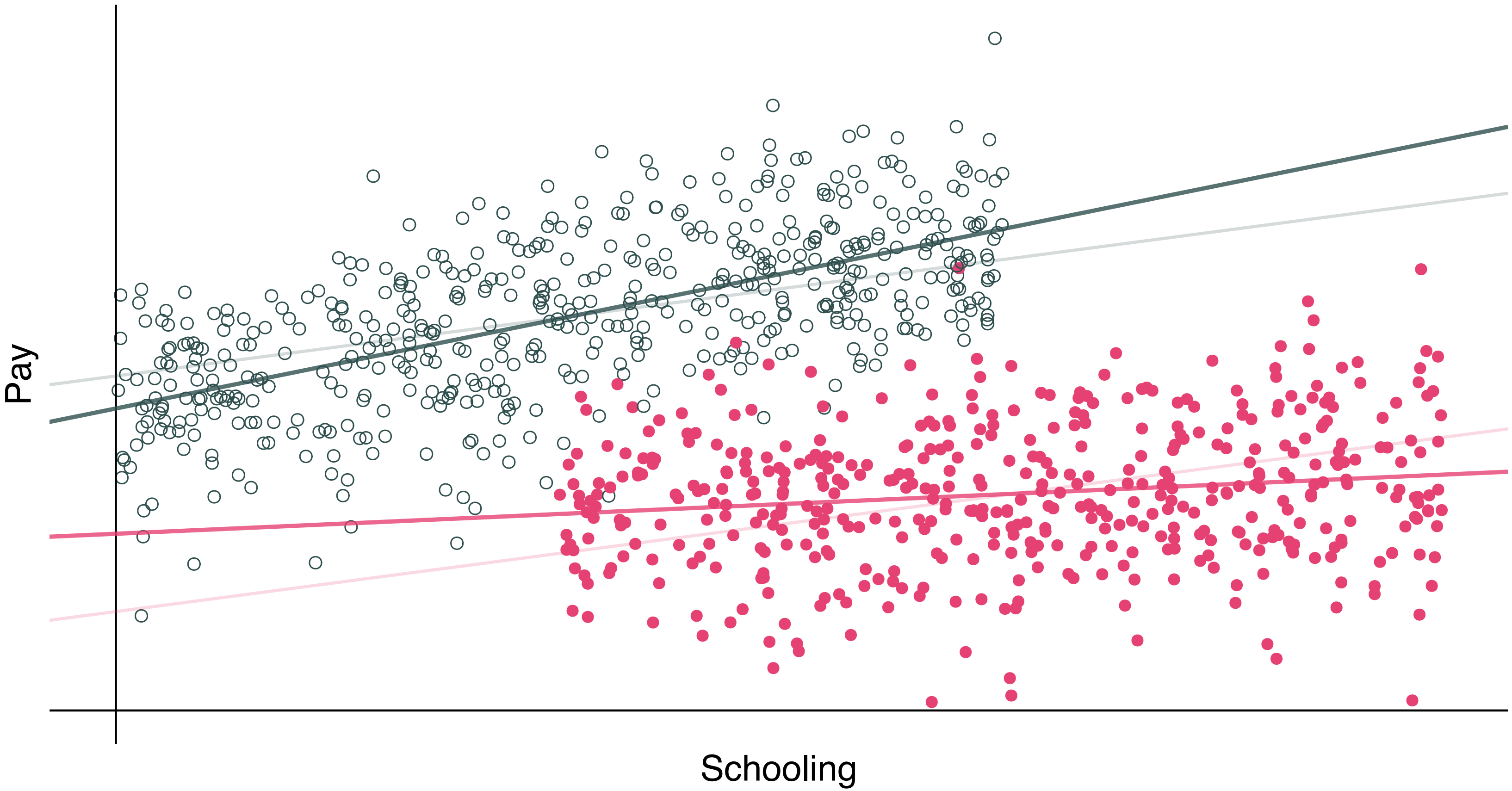
$$\text{Pay}_i = \beta_0 + \beta_1 \, \text{School}_i + \beta_2 \, \text{Male}_i + u_i$$



Interactions

The model where schooling's effect can differ by gender (**F** and **M**):

$$\text{Pay}_i = \beta_0 + \beta_1\,\text{School}_i + \beta_2\,\text{Male}_i + \beta_3\,\text{School}_i \times \text{Male}_i + u_i$$



Interactions

Interpreting coefficients can be a little tricky -- carefully working through the math helps.

$$\mathbf{Pay}_i = \beta_0 + \beta_1 \, \mathbf{School}_i + \beta_2 \, \mathbf{Female}_i + \beta_3 \, \mathbf{School}_i \times \mathbf{Female}_i + u_i$$

Expected returns for an additional year of schooling for **women**:

$$\begin{aligned} \boldsymbol{E}[\text{Pay}_i | \text{Female} \wedge \text{School} = \ell + 1] - \boldsymbol{E}[\text{Pay}_i | \text{Female} \wedge \text{School} = \ell] = \\ \boldsymbol{E}[\beta_0 + \beta_1(\ell + 1) + \beta_2 + \beta_3(\ell + 1) + u_i] - \boldsymbol{E}[\beta_0 + \beta_1 \ell + \beta_2 + \beta_3 \ell + u_i] = \\ \beta_1 + \beta_3 \end{aligned}$$

Interactions

Interpreting coefficients can be a little tricky -- carefully working through the math helps.

$$\mathbf{Pay}_i = \beta_0 + \beta_1\,\mathbf{School}_i + \beta_2\,\mathbf{Female}_i + \beta_3\,\mathbf{School}_i \times \mathbf{Female}_i + u_i$$

Expected returns for an additional year of schooling for **men**:

$$\begin{aligned} \boldsymbol{E}[\mathrm{Pay}_i | \mathrm{Male} \wedge \mathrm{School} = \ell + 1] - \boldsymbol{E}[\mathrm{Pay}_i | \mathrm{Male} \wedge \mathrm{School} = \ell] &= \\ \boldsymbol{E}[\beta_0 + \beta_1(\ell + 1) + u_i] - \boldsymbol{E}[\beta_0 + \beta_1 \ell + u_i] &= \\ \beta_1 & \end{aligned}$$

Interactions

Interpreting coefficients can be a little tricky -- carefully working through the math helps.

$$\textbf{Pay}_i = \beta_0 + \beta_1 \, \textbf{School}_i + \beta_2 \, \textbf{Female}_i + \beta_3 \, \textbf{School}_i \times \textbf{Female}_i + \boldsymbol{u}_i$$

Expected returns for an additional year of schooling for **men**:

$$\begin{aligned} \boldsymbol{E}[\text{Pay}_i | \text{Male} \wedge \text{School} = \ell + 1] - \boldsymbol{E}[\text{Pay}_i | \text{Male} \wedge \text{School} = \ell] &= \\ \boldsymbol{E}[\beta_0 + \beta_1(\ell + 1) + \boldsymbol{u}_i] - \boldsymbol{E}[\beta_0 + \beta_1 \ell + \boldsymbol{u}_i] &= \\ \beta_1 & \end{aligned}$$

Thus, β_3 gives the **difference in the returns to schooling** for women and men.

Interactions

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} \times x_{2i} + u_i$$

In general, interaction models should be used when **the level of one variable influences the relationship between the outcome and another variables**

Interactions

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} \times x_{2i} + u_i$$

In general, interaction models should be used when **the level of one variable influences the relationship between the outcome and another variables**

For example:

- Income changes the relationship between extreme heat and mortality (Carleton et al., 2021)

Interactions

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} \times x_{2i} + u_i$$

In general, interaction models should be used when **the level of one variable influences the relationship between the outcome and another variables**

For example:

- Income changes the relationship between extreme heat and mortality (Carleton et al., 2021)
- Gender changes the relationship between air pollution and labor productivity (Graff-Zivin and Neidell, 2021)

Interactions

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} \times x_{2i} + u_i$$

In general, interaction models should be used when **the level of one variable influences the relationship between the outcome and another variables**

For example:

- Income changes the relationship between extreme heat and mortality (Carleton et al., 2021)
- Gender changes the relationship between air pollution and labor productivity (Graff-Zivin and Neidell, 2021)
- Other examples?

Interactions

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} \times x_{2i} + u_i$$

Interpreting interaction models means you have to consider the interaction term when computing slopes.

For example: What is the "slope" of the relationship between y and x_1?

Interactions

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} \times x_{2i} + u_i$$

Interpreting interaction models means you have to consider the interaction term when computing slopes.

For example: What is the "slope" of the relationship between y and x_1?

$$\begin{aligned} \boldsymbol{E}[y_i|x_{i2}, x_{i1} = \ell + 1] - \boldsymbol{E}[y_i|x_{i2}, x_{i1} = \ell] &= \\ \boldsymbol{E}[\beta_0 + \beta_1(\ell + 1) + \beta_3(\ell + 1) \times x_{i2} + u_i] - \boldsymbol{E}[\beta_0 + \beta_1 \ell + \beta_3(\ell) \times x_{i2} + u_i] &= \\ \beta_1 + \beta_3 x_{i2} \end{aligned}$$

Note: higher x_{i2} increases the slope of the relationship between y and x_1!

The inverse is also true.

For two continous random variables, we now have infinitely many slopes for each variable, depending on the level of the other independent variable.

Multicollinearity

Multicollinearity

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$$

What is it?

- When 2 (*collinearity*) or more (*multicollinearity*) of your independent variables are highly correlated with one another

Multicollinearity

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$$

What is it?

- When 2 (*collinearity*) or more (*multicollinearity*) of your independent variables are highly correlated with one another

What is the problem?

- Coefficients change *substantially* with small changes in independent variables
- Illogical/unexpected coefficients

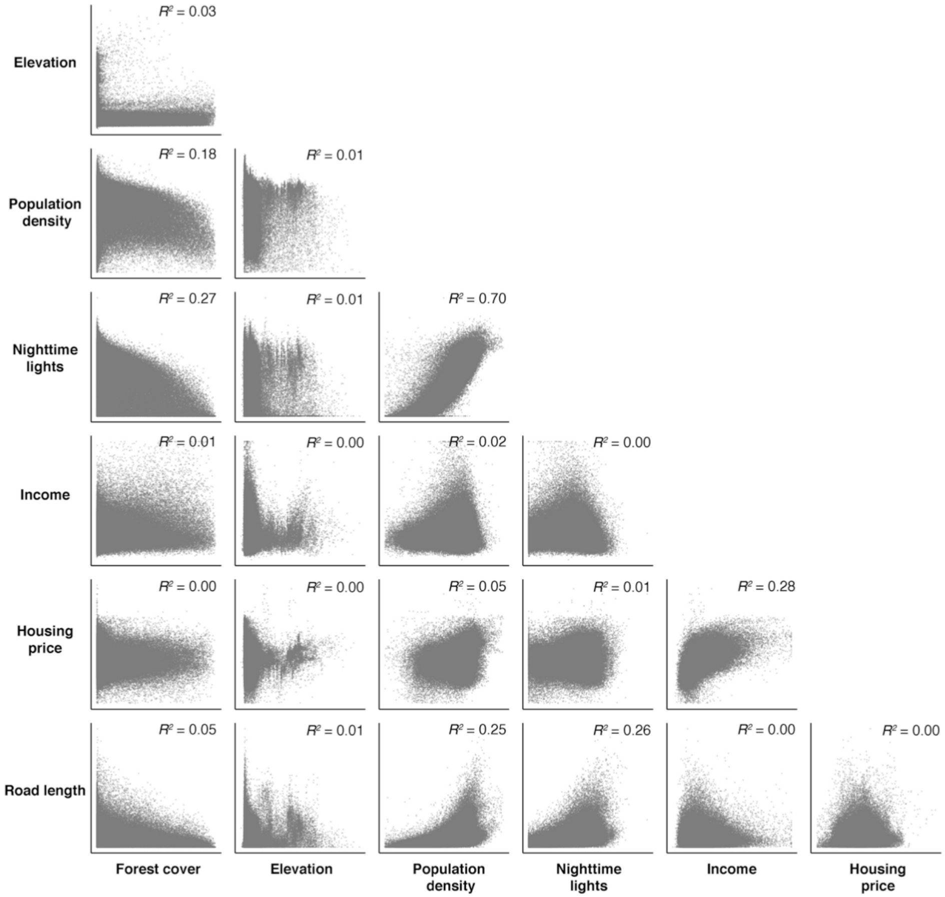
Multicollinearity

Why might it happen?

- Too many independent variables ("overspecified" model)
- Including dummy variable for your reference group
- True population correlation between variables is high

Multicollinearity

Easy check: `ggpairs()`, `pairs()`, etc.



Multicollinearity

What to do about it?

- More data helps, if possible
- Check if some variables should be omitted based on theory/conceptual model (e.g., reference group dummy)?
- Eliminate highly correlated variables (ensure your interpretation changes accordingly)
    - E.g., temperature and humidity

Slides created via the R package **xaringan**.

Some slides and slide components were borrowed from Ed Rubin's awesome course materials.