Metrics Review, Part 2

EC 421, Set 3

Connor Lennon
Spring 2021

Prologue

R showcase

`ggplot2`

- Incredibly powerful graphing and mapping package for R.
- Written in a way that helps you build your figures layer by layer.
- Exportable to many applications.
- Party of the `tidyverse`.

`shiny`

- Export your figures and code to interactive web apps.
- Enormous range of applications
  - Distribution calculator
  - Tabsets
  - Traveling salesman

Schedule

Last Time

We reviewed the fundamentals of statistics and econometrics.

Today

We review more of the main/basic results in metrics.

This week

We will post the **first assignment** (focused on *review*) soon.

Multiple regression

Multiple regression

More explanatory variables

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

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

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

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

Multiple regression

More explanatory variables

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

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

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?

Multiple regression

More explanatory variables

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

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

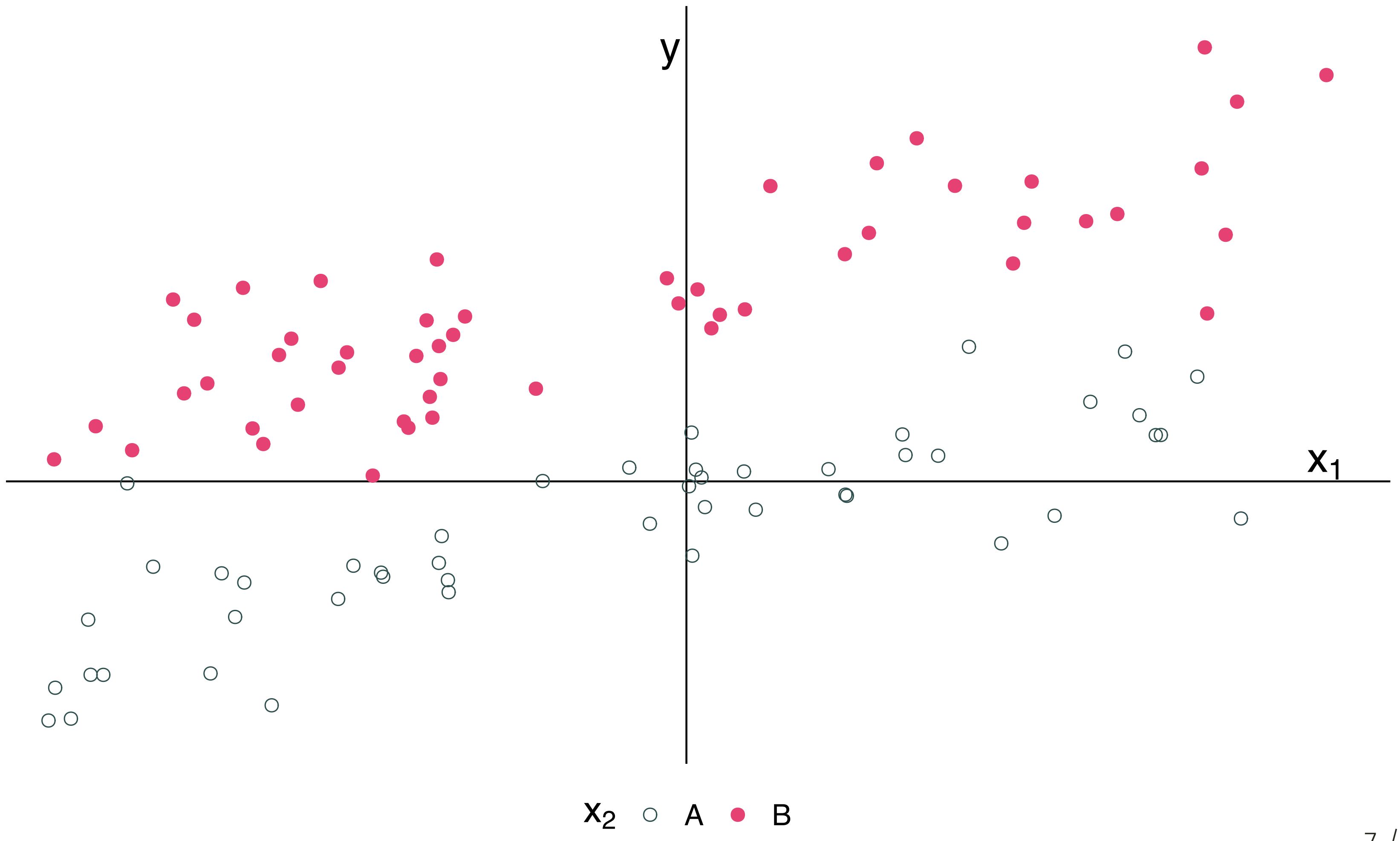
to the land of **multiple linear regression** (one outcome variable and multiple explanatory variables)

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

Why? We can better explain the variation in y, improve predictions, avoid omitted-variable bias, ...

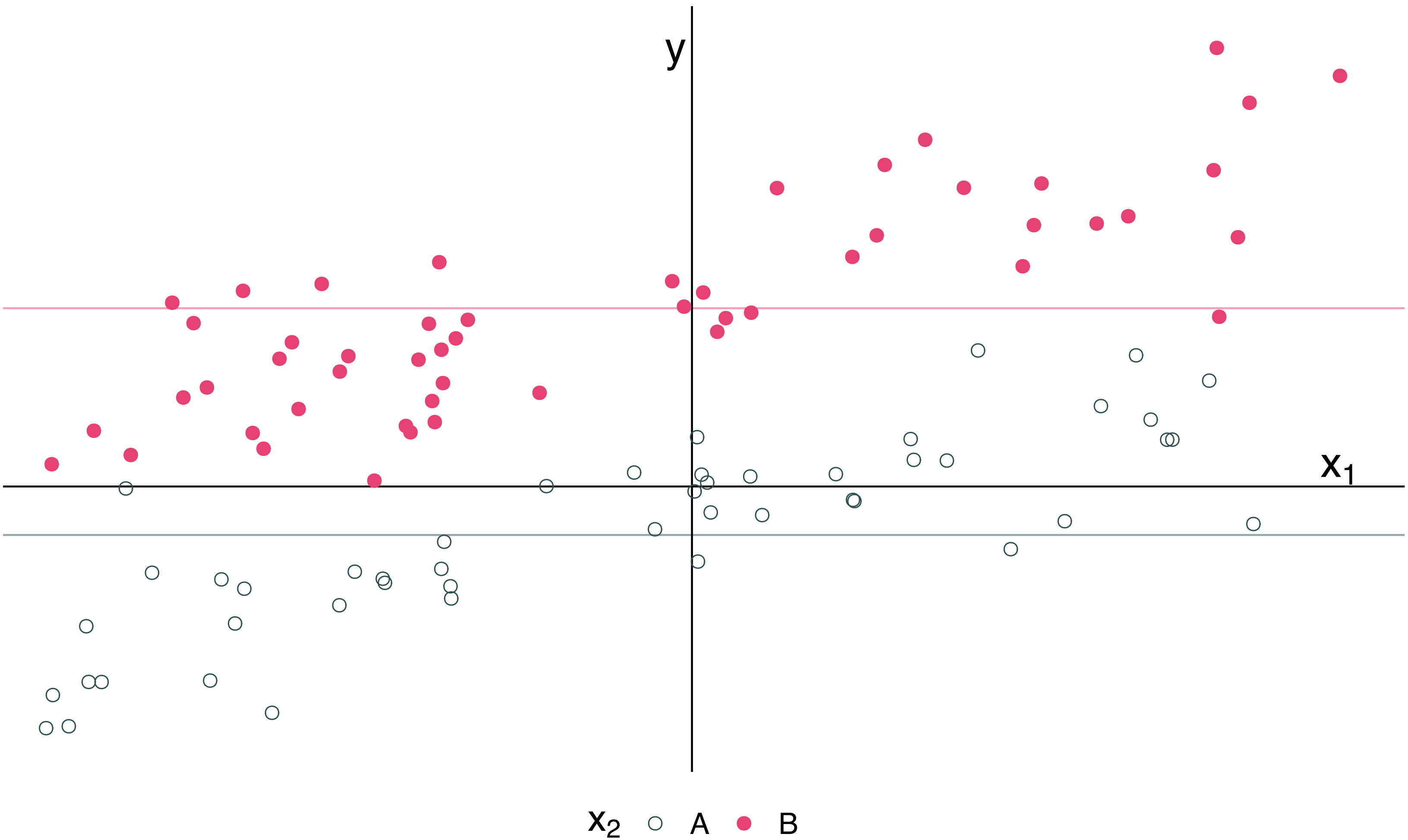
Multiple regression

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



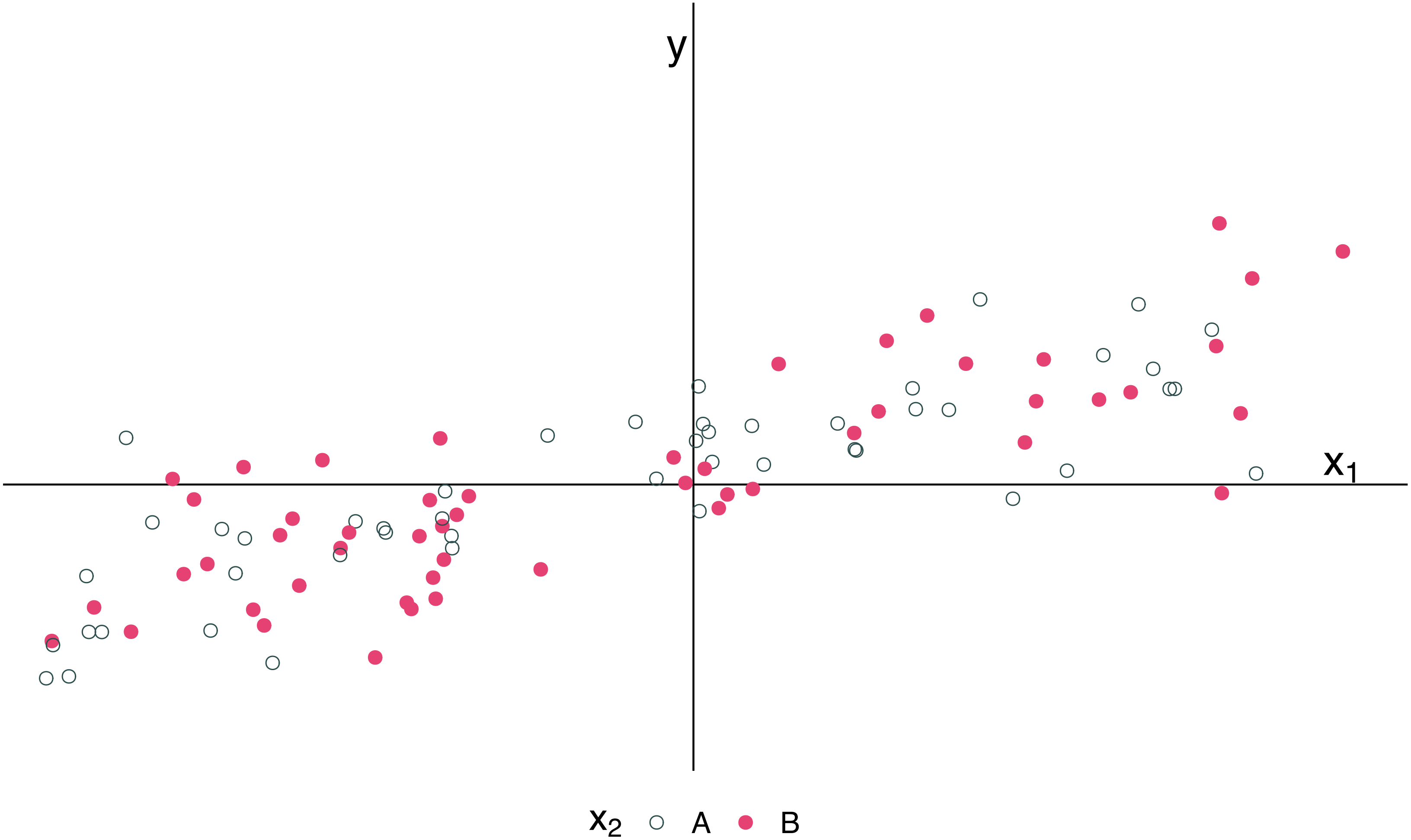
Multiple regression

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



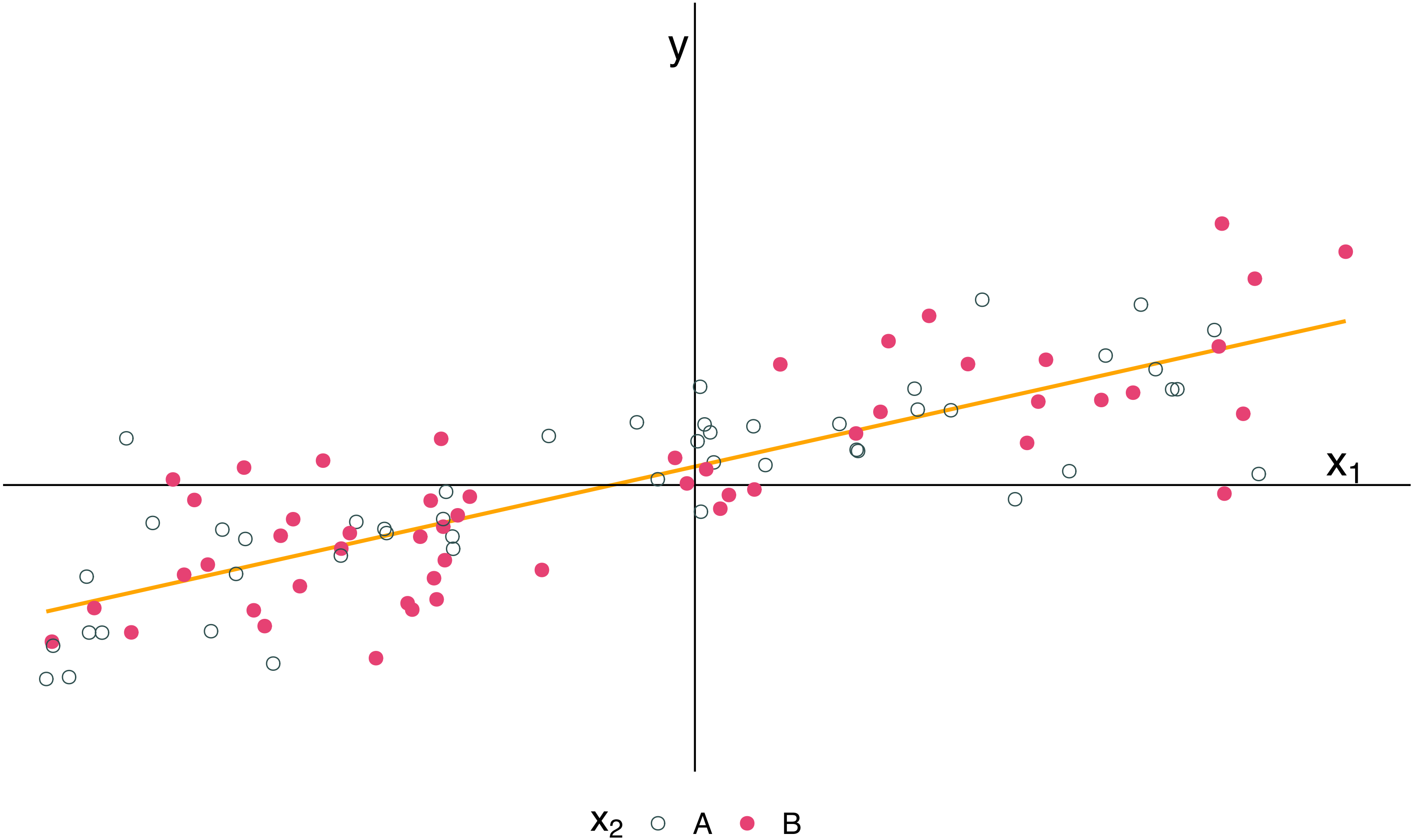
Multiple regression

With groups' means removed:



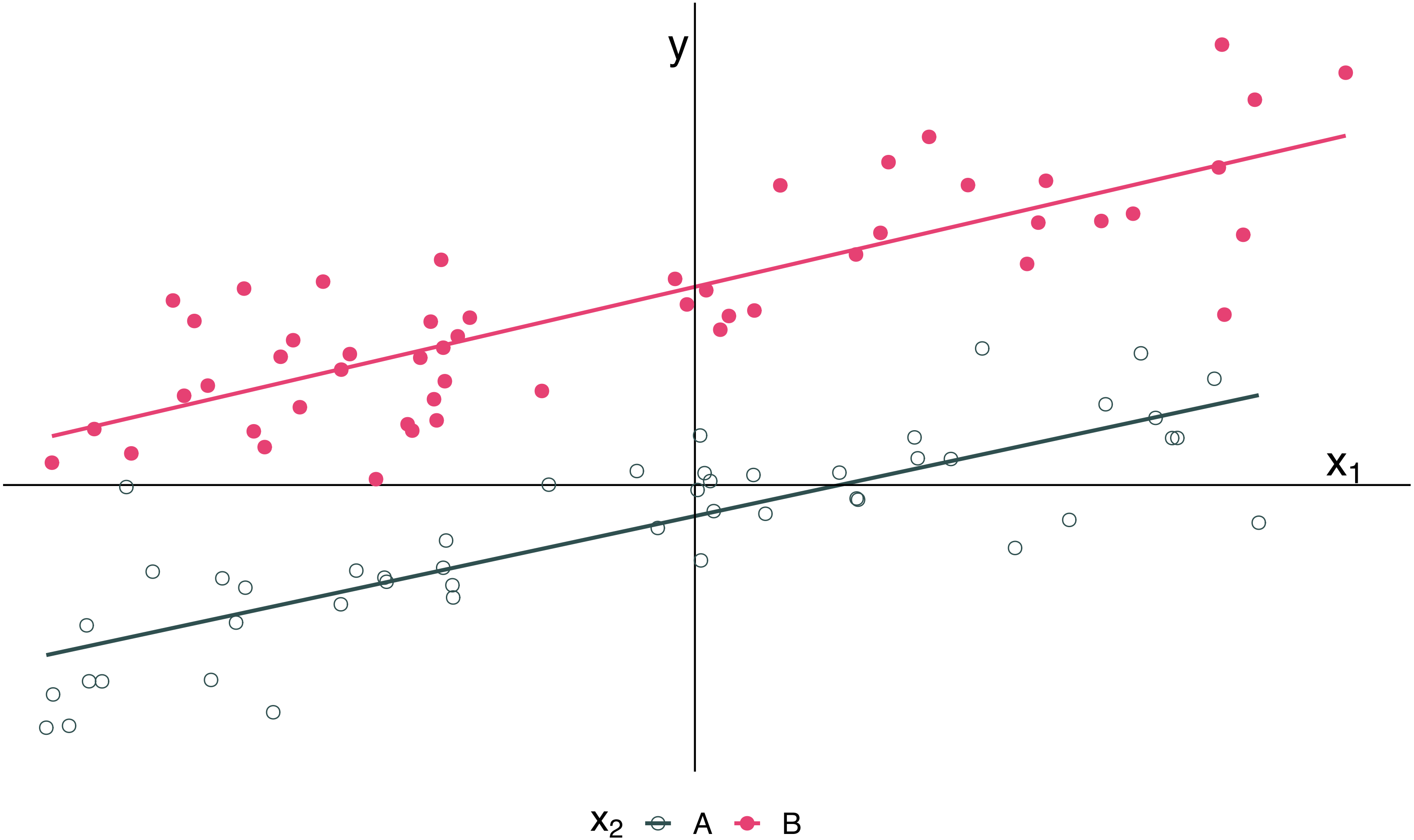
Multiple regression

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



Multiple regression

Another way to think about it:



Multiple regression

Looking at our estimator can also help.

For the simple linear regression $y_i = \beta_0 + \beta_1 x_i + u_i$

$$
\begin{aligned}
\hat{\beta}_1 &= \\
&= \frac{\sum_i \left(x_i - \overline{x} \right) \left(y_i - \overline{y} \right)}{\sum_i \left(x_i - \overline{x} \right)} \\
&= \frac{\sum_i \left(x_i - \overline{x} \right) \left(y_i - \overline{y} \right) / (n-1)}{\sum_i \left(x_i - \overline{x} \right) / (n-1)} \\
&= \frac{\hat{\mathrm{Cov}}(x,\, y)}{\hat{\mathrm{Var}}(x)}
\end{aligned}
$$

Multiple regression

Simple linear regression estimator:

$$\hat{\beta}_1 = \frac{\hat{\mathrm{Cov}}(x, y)}{\hat{\mathrm{Var}}(x)}$$

moving to multiple linear regression, the estimator changes slightly:

$$\hat{\beta}_1 = \frac{\hat{\mathrm{Cov}}(\tilde{x}_1, y)}{\hat{\mathrm{Var}}(\tilde{x}_1)}$$

where $\tilde{x}_1$ is the *residualized* x_1 variable—the variation remaining in x after controlling for the other explanatory variables.

Multiple regression

More formally, consider the multiple-regression model

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u_i$$

Our residualized x_1 (which we named $\tilde{x}_1$) comes from regressing x_1 on an intercept and all of the other explanatory variables and collecting the residuals, *i.e.*,

$$\begin{aligned} \hat{x}_{1i} &= \hat{\gamma}_0 + \hat{\gamma}_2\, x_{2i} + \hat{\gamma}_3\, x_{3i} \\ \tilde{x}_{1i} &= x_{1i} - \hat{x}_{1i} \end{aligned}$$

Multiple regression

More formally, consider the multiple-regression model

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u_i$$

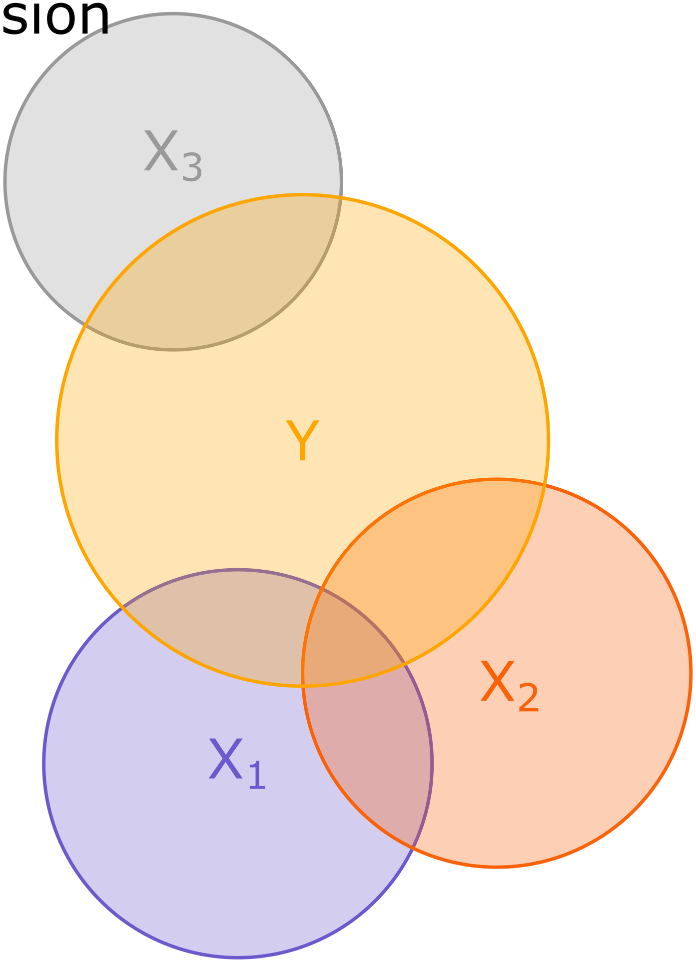
Our residualized x_1 (which we named $\tilde{x}_1$) comes from regressing x_1 on an intercept and all of the other explanatory variables and collecting the residuals, *i.e.*,

$$\begin{aligned}\hat{x}_{1i} &= \hat{\gamma}_0 + \hat{\gamma}_2\, x_{2i} + \hat{\gamma}_3\, x_{3i} \\ \tilde{x}_{1i} &= x_{1i} - \hat{x}_{1i}\end{aligned}$$

allowing us to better understand our OLS multiple-regression estimator

$$\hat{\beta}_1 = \frac{\widehat{\mathrm{Cov}}(\tilde{x}_1,\, y)}{\widehat{\mathrm{Var}}(\tilde{x}_1)}$$

Multiple regression



Multiple regression

Model fit

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)

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

Notice our old friend SSE: $\sum_i \left(y_i - \hat{y}_i\right)^2 = \sum_i e_i^2$.

Multiple regression

Model fit

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)

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

Notice our old friend SSE: $\sum_i \left(y_i - \hat{y}_i\right)^2 = \sum_i e_i^2$.

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

Multiple regression

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

Multiple regression

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

To see this problem, we can simulate a dataset of 10,000 observations on y and 1,000 random x_k variables. **No relations between y and the x_k!**

Pseudo-code outline of the simulation:

Multiple regression

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

To see this problem, we can simulate a dataset of 10,000 observations on y and 1,000 random x_k variables. **No relations between y and the x_k!**

Pseudo-code outline of the simulation:

- Generate 10,000 observations on y
- Generate 10,000 observations on variables x_1 through x_{1000}
- Regressions
  - LM$_1$: Regress y on x_1; record R^2
  - LM$_2$: Regress y on x_1 and x_2; record R^2
  - LM$_3$: Regress y on x_1, x_2, and x_3; record R^2
  - ...
  - LM$_{1000}$: Regress y on x_1, x_2, ..., x_{1000}; record R^2

Multiple regression

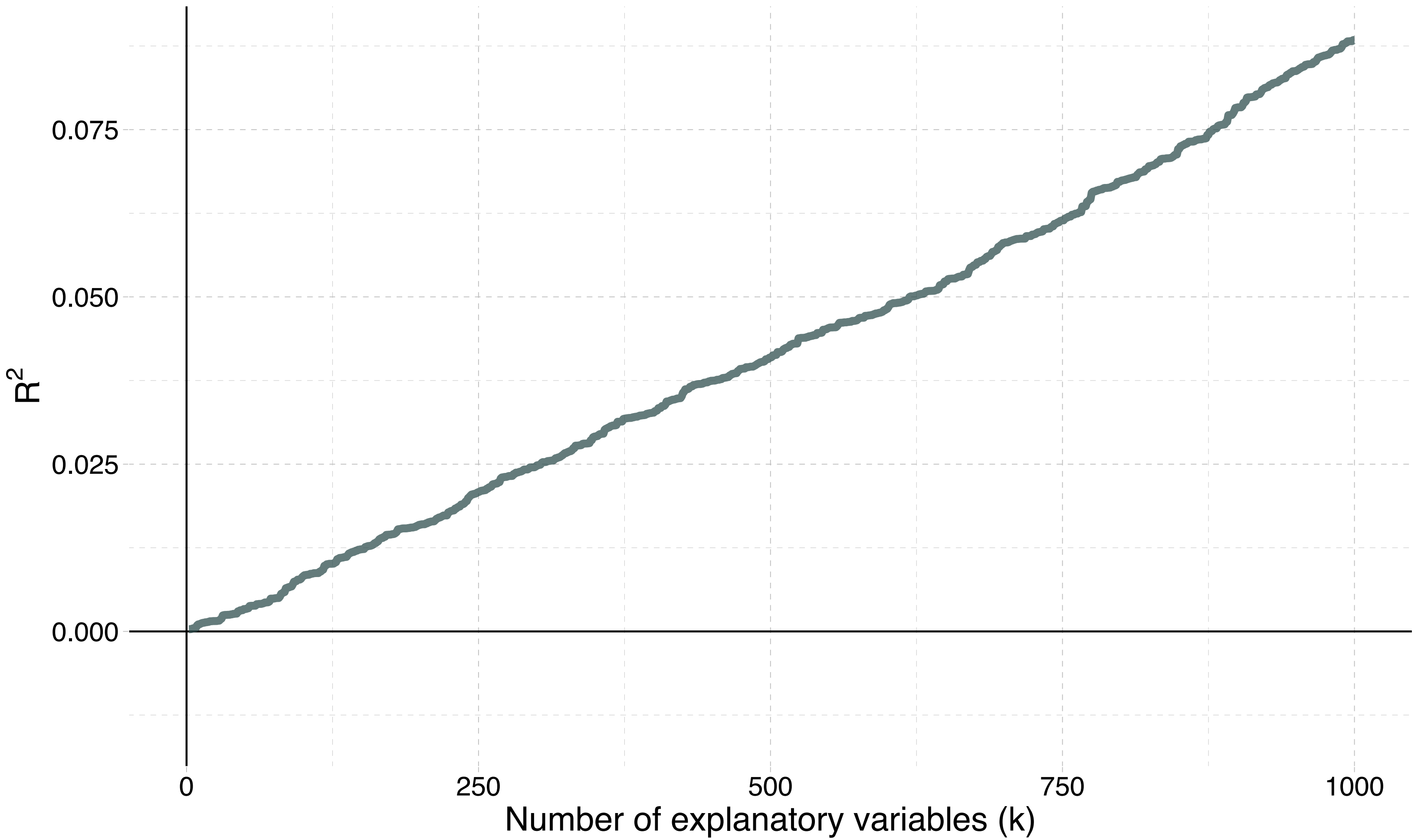
The problem: As we add variables to our model, R^2 *mechanically* increases.

R code for the simulation:

```
set.seed(1234)
y <- rnorm(1e4)
x <- matrix(data = rnorm(1e7), nrow = 1e4)
x %<>% cbind(matrix(data = 1, nrow = 1e4, ncol = 1), x)
r_df <- mclapply(X = 1:(1e3-1), mc.cores = detectCores() - 1, FUN = function(i)
  tmp_reg <- lm(y ~ x[,1:(i+1)]) %>% summary()
  data.frame(
    k = i + 1,
    r2 = tmp_reg %$% r.squared,
    r2_adj = tmp_reg %$% adj.r.squared
  )
}) %>% bind_rows()
```

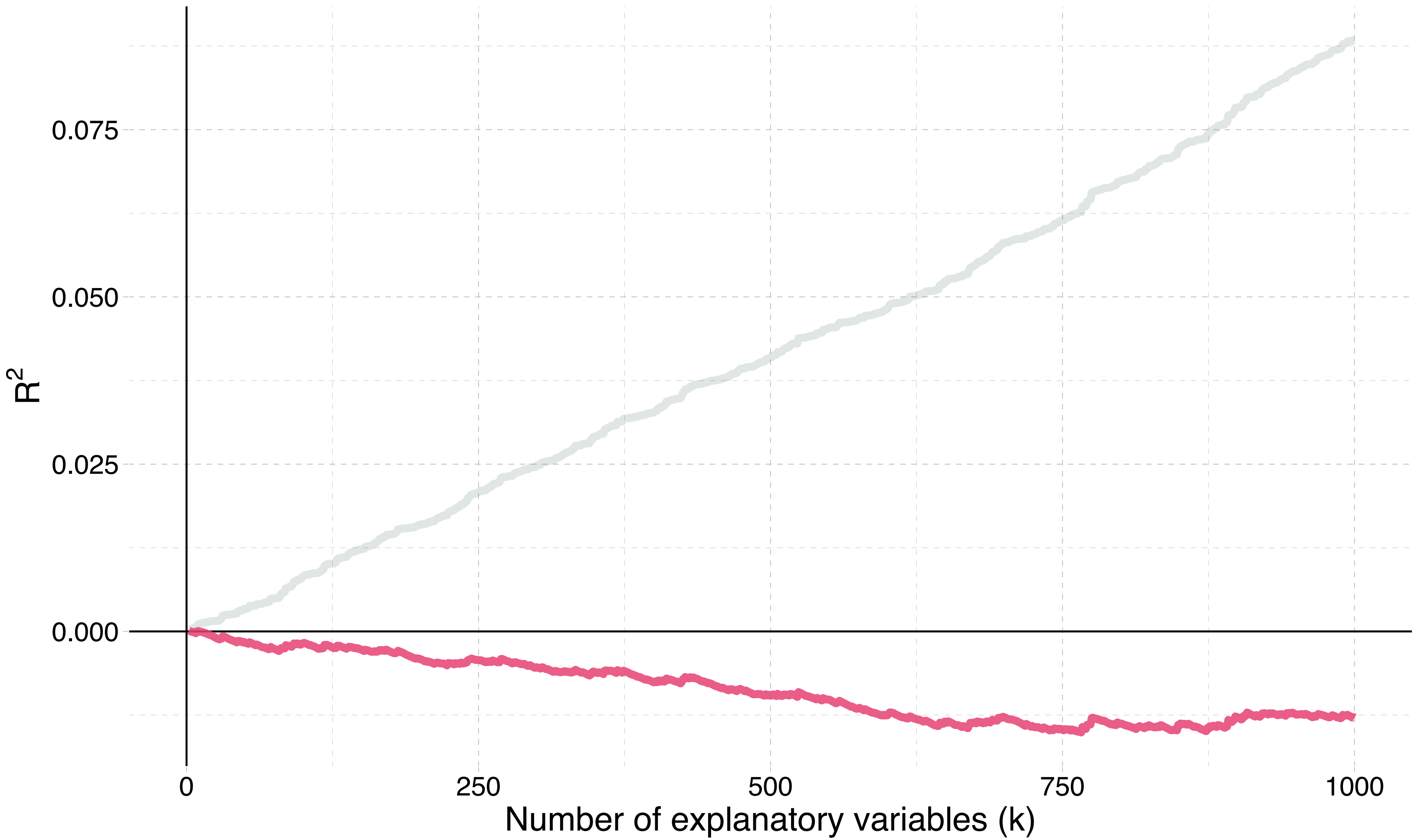
Multiple regression

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



Multiple regression

One solution: Adjusted R^2



Multiple regression

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

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.

Tradeoffs

There are tradeoffs to remember 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 variabless—still omitted-variable bias

Omitted-variable bias

Omitted-variable bias

We'll go deeper into this issue in a few weeks, but as a refresher:

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.

Omitted-variable bias

We'll go deeper into this issue in a few weeks, but as a refresher:

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.

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

From our population model

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

If a study focuses on the relationship between pay and schooling, *i.e.*,

$$\text{Pay}_i = \beta_0 + \beta_1 \text{School}_i + \left(\beta_2 \text{Male}_i + u_i\right)$$

$$\text{Pay}_i = \beta_0 + \beta_1 \text{School}_i + \varepsilon_i$$

where $\varepsilon_i = \beta_2 \text{Male}_i + u_i$.

We used our exogeneity assumption to derive OLS' unbiasedness. But even if $\boldsymbol{E}[u|X] = 0$, it is not true that $\boldsymbol{E}[\varepsilon|X] = \mathbf{0}$ so long as $\beta_2 \neq 0$.

Omitted-variable bias

Example, continued

From our population model

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

If a study focuses on the relationship between pay and schooling, *i.e.*,

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

$$\text{Pay}_i = \beta_0 + \beta_1 \text{School}_i + \varepsilon_i$$

where $\varepsilon_i = \beta_2 \text{Male}_i + u_i$.

Specifically, exogeneity requires that School and Male are unrelated.
Otherwise OLS is biased.

Omitted-variable bias

Example, continued

Let's try to see this result graphically.

The population model:

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

Our regression model that suffers from omitted-variable bias:

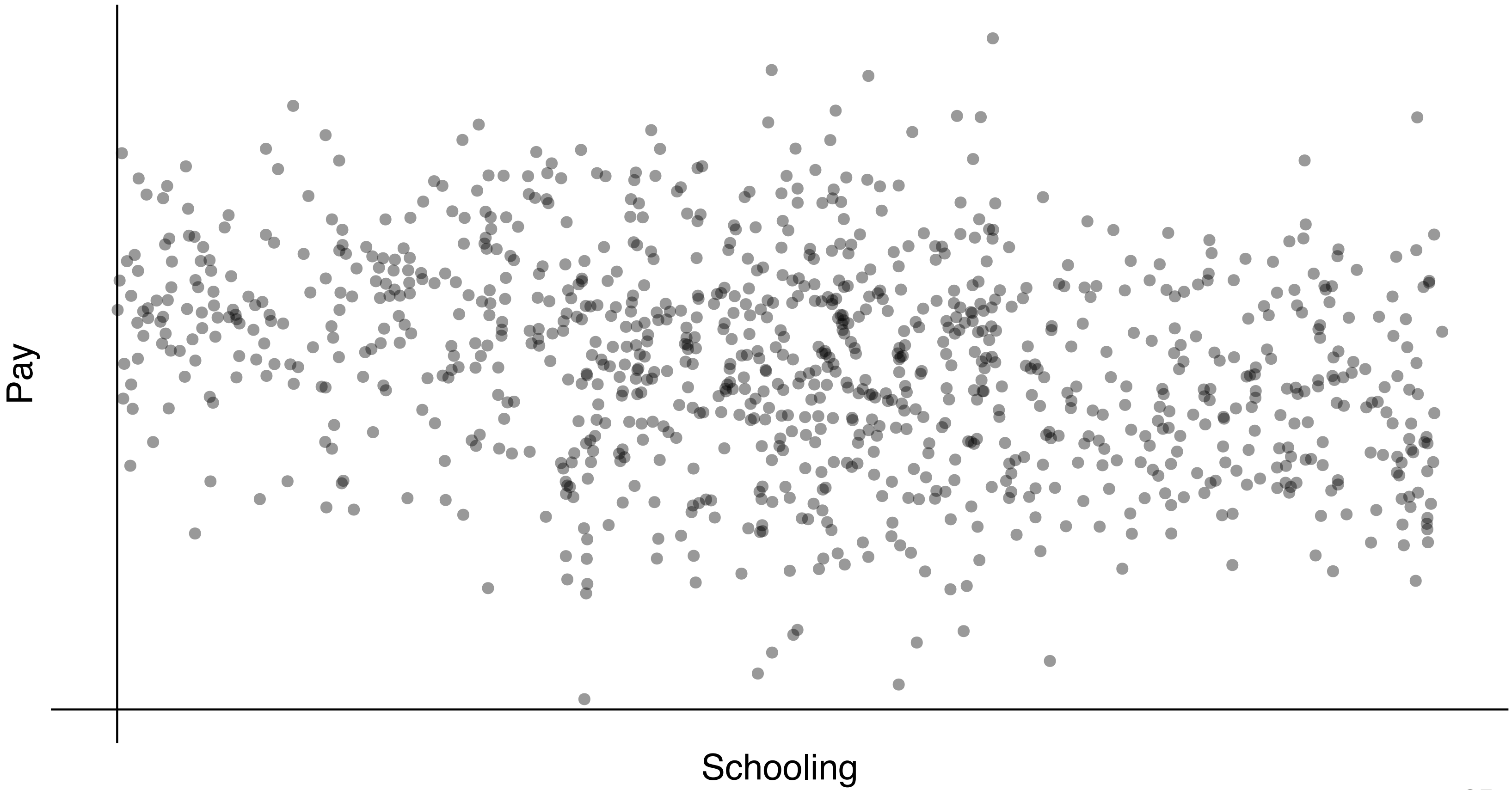
$$\text{Pay}_i = \hat{\beta}_0 + \hat{\beta}_1 \times \text{School}_i + e_i$$

Finally, imagine that women, on average, receive more schooling than men.

Omitted-variable bias

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

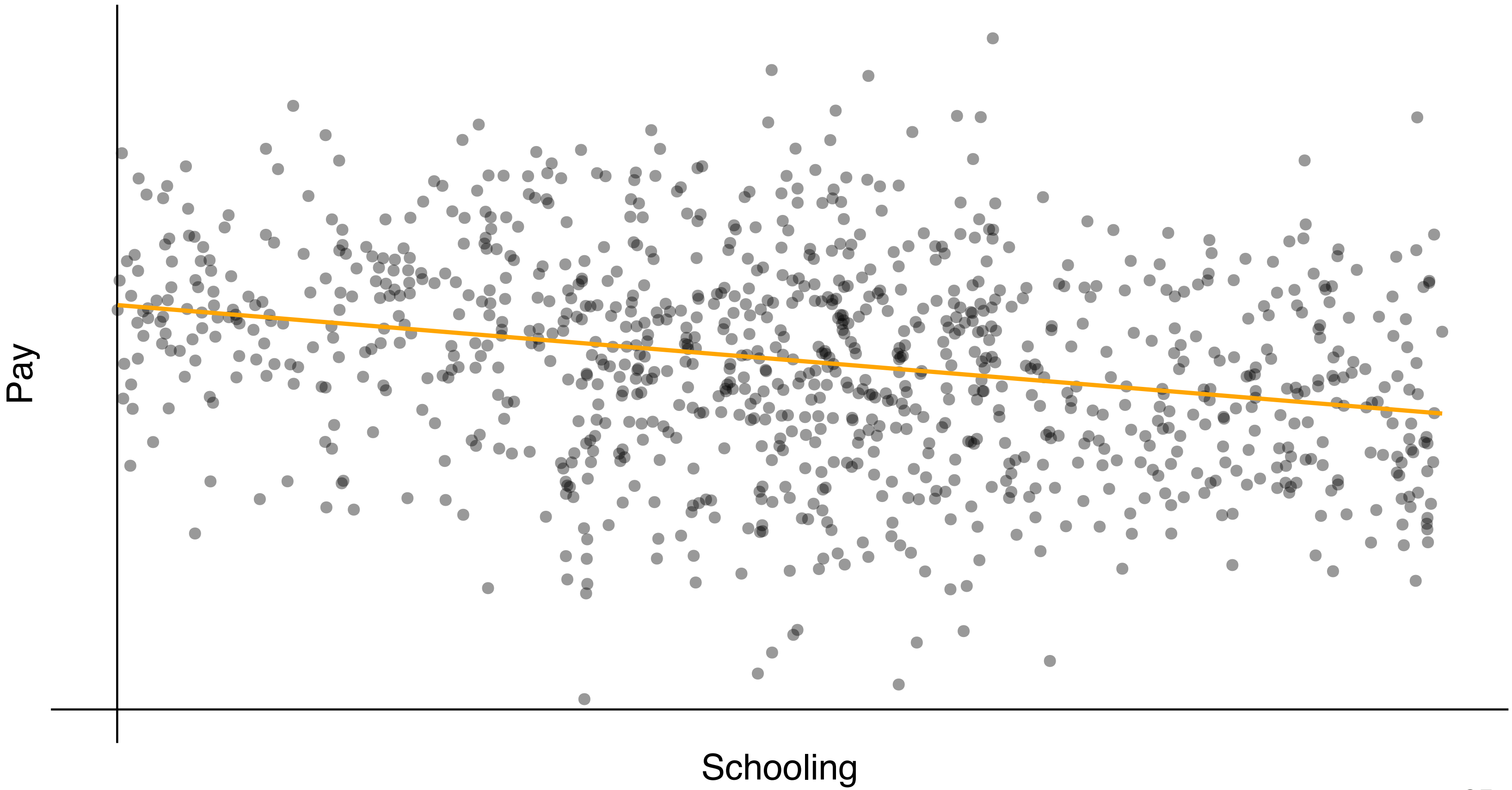
The relationship between pay and schooling.



Omitted-variable bias

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

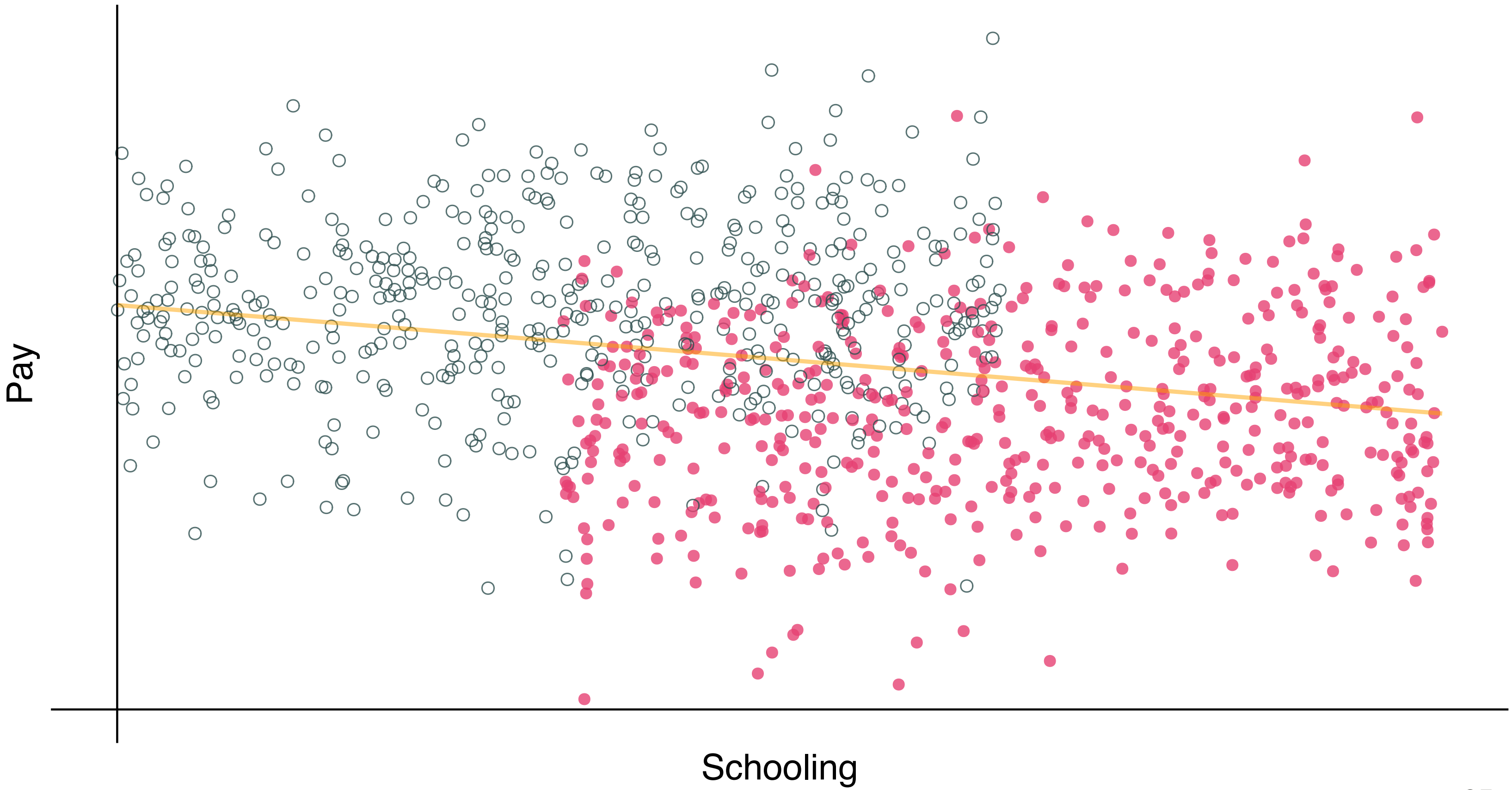
Biased regression estimate: $\widehat{\textbf{Pay}}_i = \mathbf{31.3} + \mathbf{-0.9} \times \textbf{School}_i$



Omitted-variable bias

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

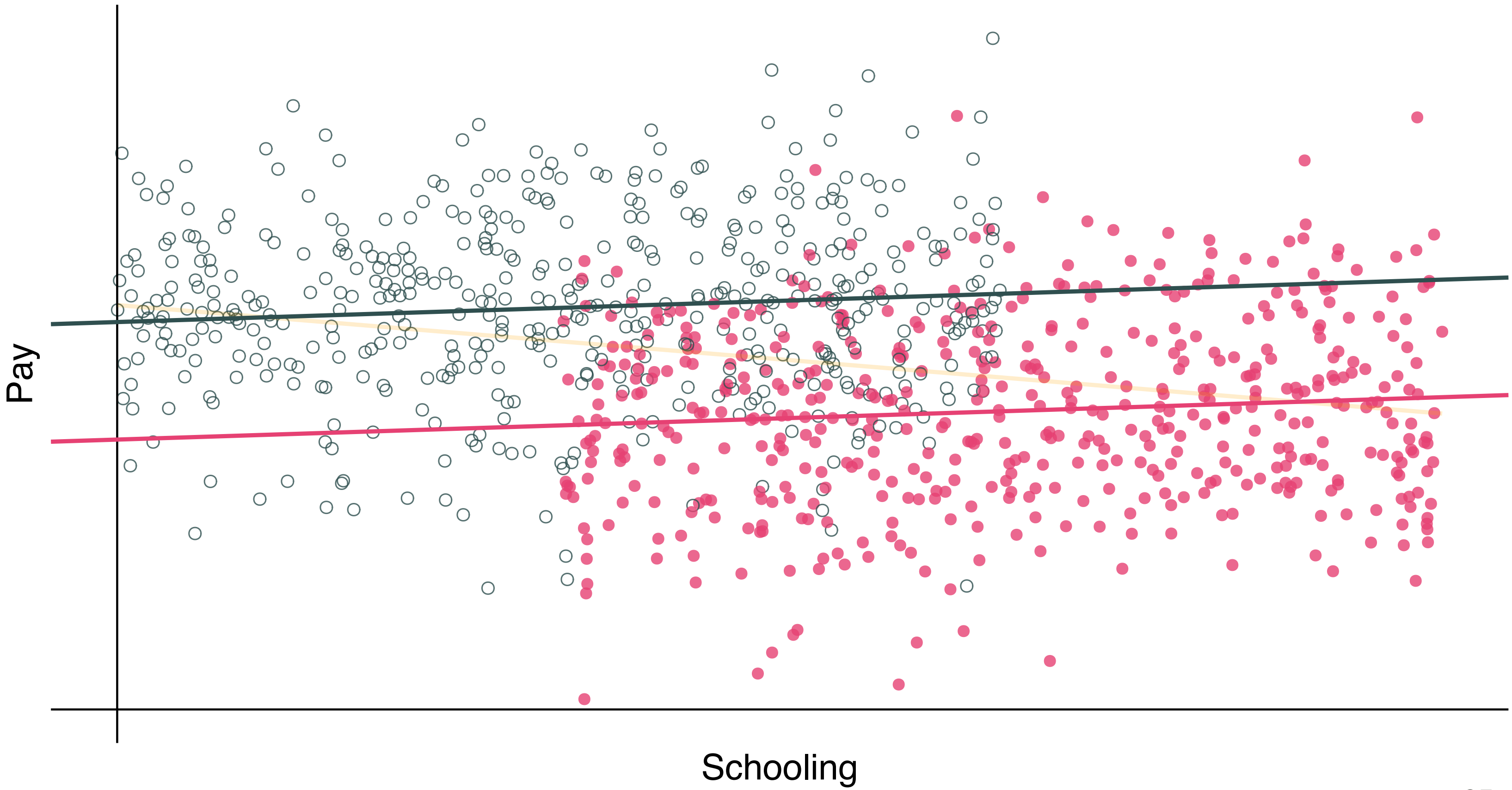
Recalling the omitted variable: Gender (**female** and **male**)



Omitted-variable bias

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

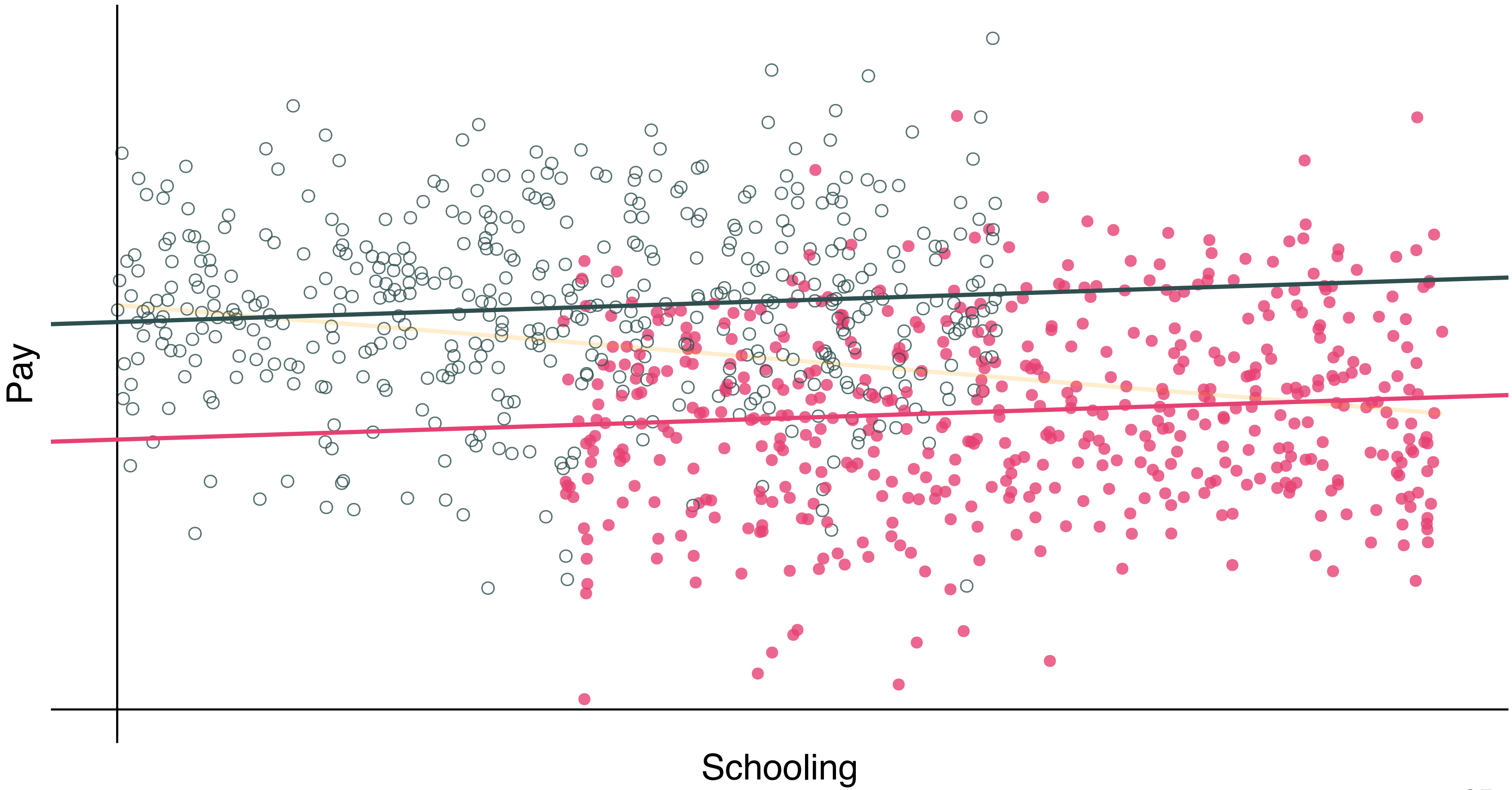
Recalling the omitted variable: Gender (**female** and **male**)



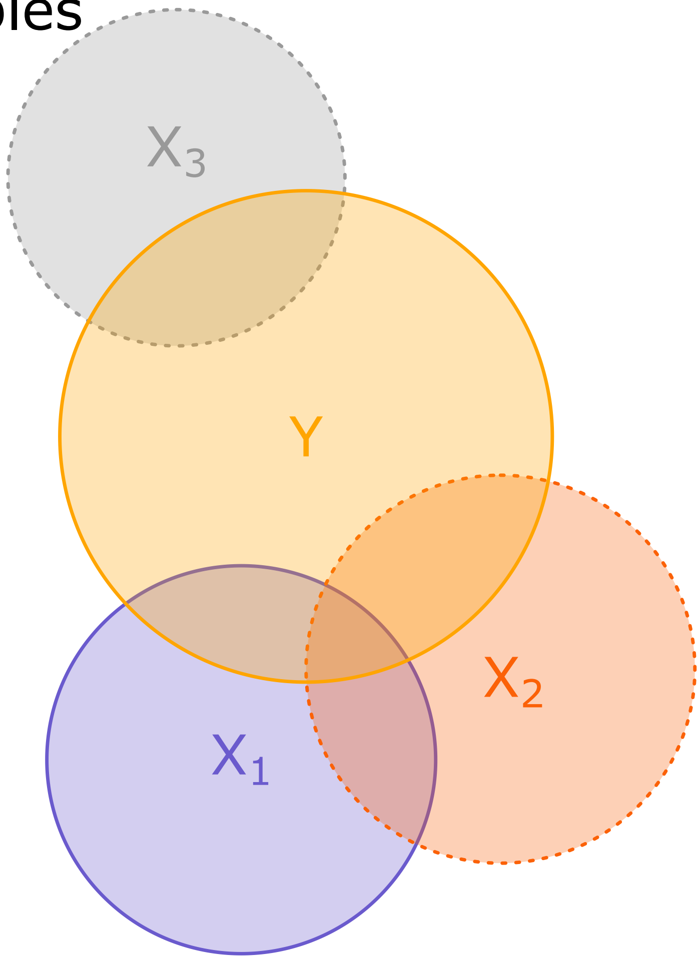
Omitted-variable bias

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

Unbiased regression estimate: $\widehat{\text{Pay}}_i = 20.9 + 0.4 \times \text{School}_i + 9.1 \times \text{Male}_i$



Omitted variables



Omitted-variable bias

Solutions

1. Don't omit variables

2. Instrumental variables and two-stage least squares[†]

Warning: There are situations in which neither solution is possible.

Omitted-variable bias

Solutions

1. Don't omit variables

2. Instrumental variables and two-stage least squares[†]

Warning: There are situations in which neither solution is possible.

1. Proceed with caution (sometimes you can sign the bias).

2. Maybe just stop.

[†]: Coming soon to our lectures.

Interpreting coefficients

Interpreting coefficients

Continuous variables

Consider the relationship

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

where

- Pay_i is a continuous variable measuring an individual's pay
- School_i is a continuous variable that measures years of education

Interpreting coefficients

Continuous variables

Consider the relationship

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

where

- Pay_i is a continuous variable measuring an individual's pay
- School_i is a continuous variable that measures years of education

Interpretations

- β_0: the y-intercept, *i.e.*, Pay when $\text{School} = 0$
- β_1: the expected increase in Pay for a one-unit increase in School

Interpreting coefficients

Continuous variables

Deriving the slope's interpretation:

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

Interpreting coefficients

Continuous variables

Deriving the slope's interpretation:

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

I.e., the slope gives the expected increase in our outcome variable for a one-unit increase in the explanatory variable.

Interpreting coefficients

Continuous variables

If we have multiple explanatory variables, *e.g.*,

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

then the interpretation changes slightly.

Interpreting coefficients

Continuous variables

If we have multiple explanatory variables, *e.g.*,

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

then the interpretation changes slightly.

$$\begin{aligned} \boldsymbol{E}[\text{Pay}|\text{School} = \ell + 1 \land \text{Ability} = \alpha] - \\ \boldsymbol{E}[\text{Pay}|\text{School} = \ell \land \text{Ability} = \alpha] = \\ \boldsymbol{E}[\beta_0 + \beta_1(\ell + 1) + \beta_2\alpha + u] - \boldsymbol{E}[\beta_0 + \beta_1\ell + \beta_2\alpha + u] = \\ [\beta_0 + \beta_1(\ell + 1) + \beta_2\alpha] - [\beta_0 + \beta_1\ell + \beta_2\alpha] = \\ \beta_0 - \beta_0 + \beta_1\ell - \beta_1\ell + \beta_1 + \beta_2\alpha - \beta_2\alpha = \beta_1 \end{aligned}$$

Interpreting coefficients

Continuous variables

If we have multiple explanatory variables, *e.g.*,

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

then the interpretation changes slightly.

$$\begin{aligned}
\boldsymbol{E}[\text{Pay}|\text{School} = \ell + 1 \land \text{Ability} = \alpha] - \\
\boldsymbol{E}[\text{Pay}|\text{School} = \ell \land \text{Ability} = \alpha] = \\
\boldsymbol{E}[\beta_0 + \beta_1(\ell + 1) + \beta_2\alpha + u] - \boldsymbol{E}[\beta_0 + \beta_1\ell + \beta_2\alpha + u] = \\
[\beta_0 + \beta_1(\ell + 1) + \beta_2\alpha] - [\beta_0 + \beta_1\ell + \beta_2\alpha] = \\
\beta_0 - \beta_0 + \beta_1\ell - \beta_1\ell + \beta_1 + \beta_2\alpha - \beta_2\alpha = \beta_1
\end{aligned}$$

I.e., the slope gives the expected increase in our outcome variable for a one-unit increase in the explanatory variable, **holding all other variables constant** (*ceteris paribus*).

Interpreting coefficients

Continuous variables

Alternative derivation

Consider the model

$$y = \beta_0 + \beta_1 \, x + u$$

Differentiate the model:

$$\frac{dy}{dx} = \beta_1$$

Interpreting coefficients

Categorical variables

Consider the relationship

$$\text{Pay}_i = \beta_0 + \beta_1 \, \text{Female}_i + u_i$$

where

- Pay_i is a continuous variable measuring an individual's pay
- Female_i is a binary/indicator variable taking 1 when i is female

Interpreting coefficients

Categorical variables

Consider the relationship

$$\text{Pay}_i = \beta_0 + \beta_1 \, \text{Female}_i + u_i$$

where

- Pay_i is a continuous variable measuring an individual's pay
- Female_i is a binary/indicator variable taking 1 when i is female

Interpretations

- β_0: the expected Pay for males (*i.e.*, when $\text{Female} = 0$)
- β_1: the expected difference in Pay between females and males
- $\beta_0 + \beta_1$: the expected Pay for females

Interpreting coefficients

Categorical variables

Derivations

$$\begin{aligned}\boldsymbol{E}[\text{Pay}|\text{Male}] &= \boldsymbol{E}[\beta_0 + \beta_1 \times 0 + u_i] \\ &= \boldsymbol{E}[\beta_0 + 0 + u_i] \\ &= \beta_0\end{aligned}$$

Interpreting coefficients

Categorical variables

Derivations

$$\begin{aligned}
\boldsymbol{E}[\text{Pay}|\text{Male}] &= \boldsymbol{E}[\beta_0 + \beta_1 \times 0 + u_i] \\
&= \boldsymbol{E}[\beta_0 + 0 + u_i] \\
&= \beta_0
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{E}[\text{Pay}|\text{Female}] &= \boldsymbol{E}[\beta_0 + \beta_1 \times 1 + u_i] \\
&= \boldsymbol{E}[\beta_0 + \beta_1 + u_i] \\
&= \beta_0 + \beta_1
\end{aligned}$$

Interpreting coefficients

Categorical variables

Derivations

$$
\begin{aligned}
\boldsymbol{E}[\text{Pay}|\text{Male}] &= \boldsymbol{E}[\beta_0 + \beta_1 \times 0 + u_i] \\
&= \boldsymbol{E}[\beta_0 + 0 + u_i] \\
&= \beta_0
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{E}[\text{Pay}|\text{Female}] &= \boldsymbol{E}[\beta_0 + \beta_1 \times 1 + u_i] \\
&= \boldsymbol{E}[\beta_0 + \beta_1 + u_i] \\
&= \beta_0 + \beta_1
\end{aligned}
$$

Note: If there are no other variables to condition on, then $\hat{\beta}_1$ equals the difference in group means, *e.g.*, $\overline{x}_{\text{Female}} - \overline{x}_{\text{Male}}$.

Interpreting coefficients

Categorical variables

Derivations

$$
\begin{aligned}
\boldsymbol{E}[\text{Pay}|\text{Male}] &= \boldsymbol{E}[\beta_0 + \beta_1 \times 0 + u_i] \\
&= \boldsymbol{E}[\beta_0 + 0 + u_i] \\
&= \beta_0
\end{aligned}
$$

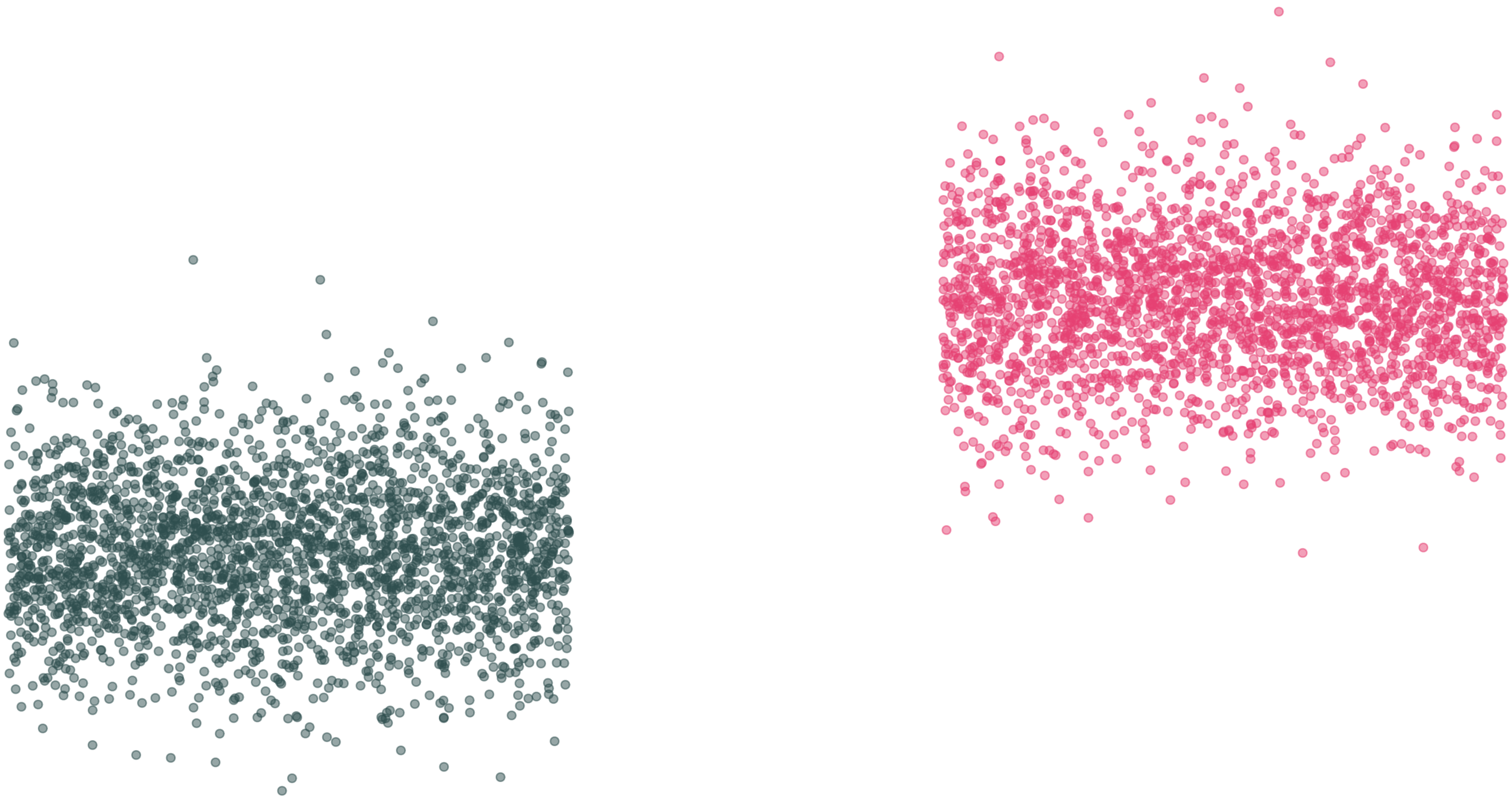
$$
\begin{aligned}
\boldsymbol{E}[\text{Pay}|\text{Female}] &= \boldsymbol{E}[\beta_0 + \beta_1 \times 1 + u_i] \\
&= \boldsymbol{E}[\beta_0 + \beta_1 + u_i] \\
&= \beta_0 + \beta_1
\end{aligned}
$$

Note: If there are no other variables to condition on, then $\hat{\beta}_1$ equals the difference in group means, *e.g.*, $\overline{x}_{\text{Female}} - \overline{x}_{\text{Male}}$.

Note$_2$: The *holding all other variables constant* interpretation also applies for categorical variables in multiple regression settings.

Categorical variables

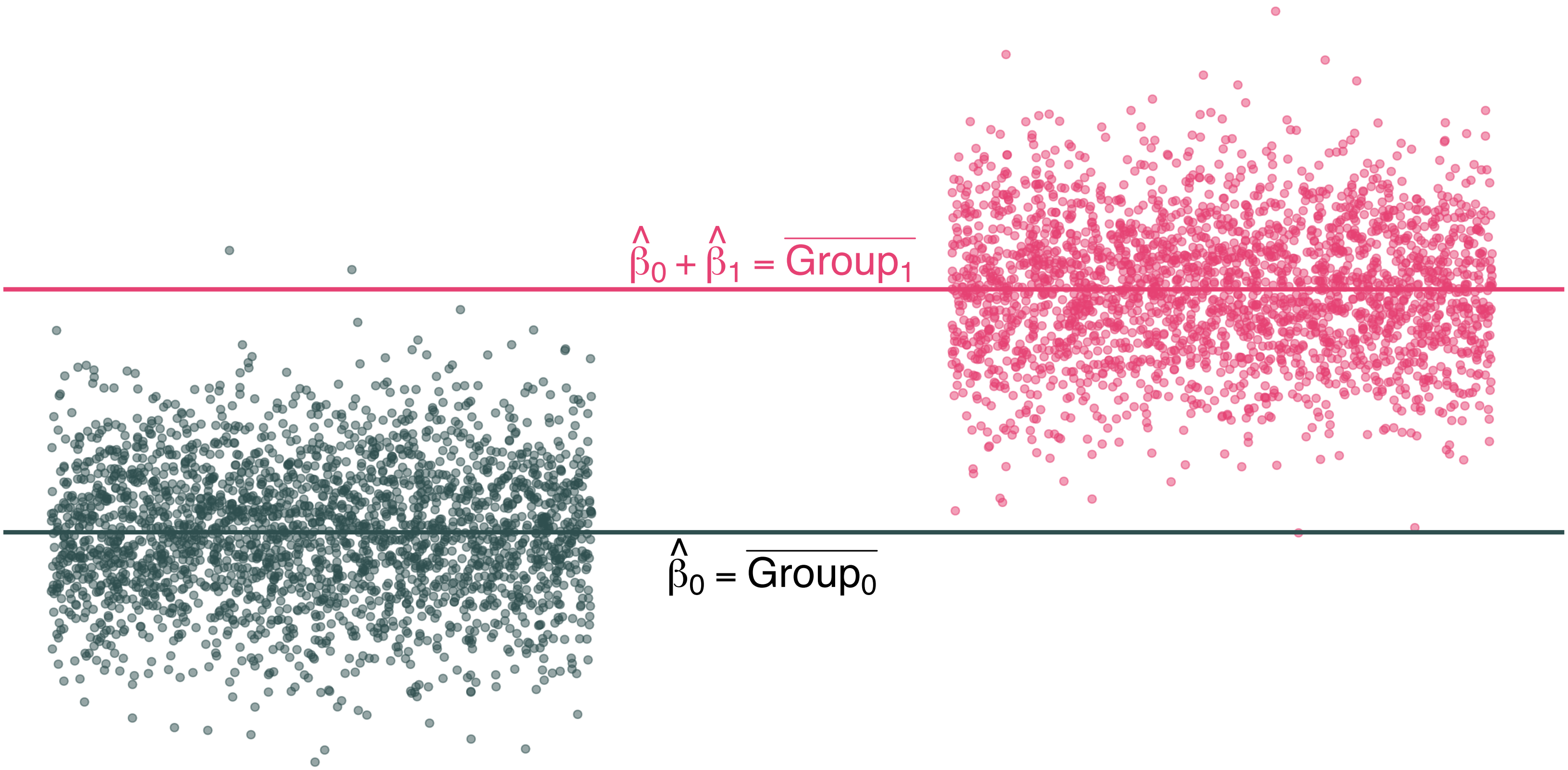
$y_i = \beta_0 + \beta_1 x_i + u_i$ for binary variable $x_i = \{0, 1\}$



Interpreting coefficients

Categorical variables

$y_i = \beta_0 + \beta_1 x_i + u_i$ for binary variable $x_i = \{0, 1\}$



Interpreting coefficients

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?

Interpreting coefficients

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:

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

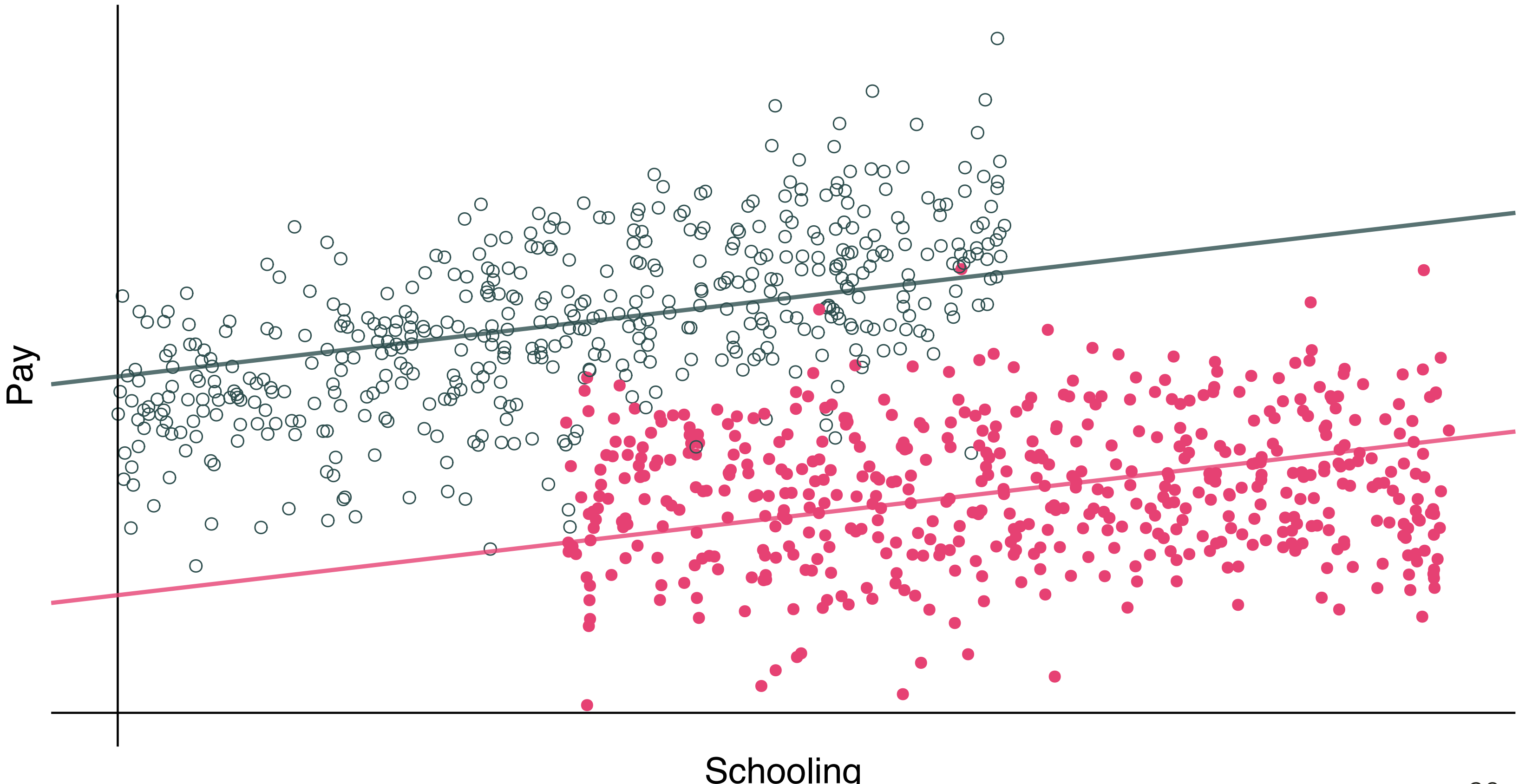
but we can also allow the effect of school to vary by gender:

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

Interpreting coefficients

Interactions

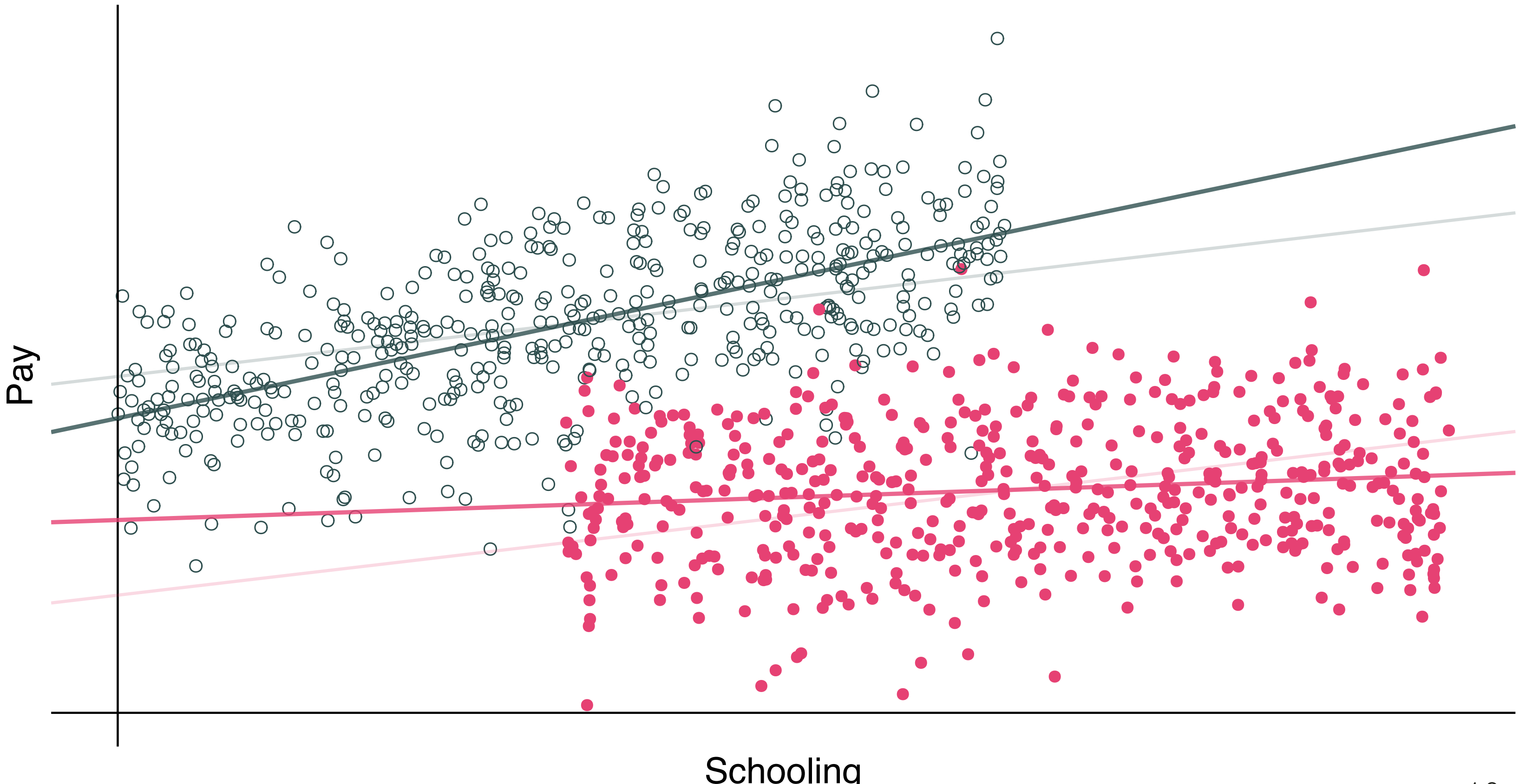
The model where schooling has the same effect for everyone (**F** and **M**):



Interpreting coefficients

Interactions

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



Interpreting coefficients

Interactions

Interpreting coefficients can be a little tricky with interactions, but the key† is to carefully work through the math.

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

Expected returns for an additional year of schooling for women:

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

† As is often the case with econometrics.

Interpreting coefficients

Interactions

Interpreting coefficients can be a little tricky with interactions, but the key[†] is to carefully work through the math.

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

Expected returns for an additional year of schooling for women:

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

Similarly, β_1 gives the expected return to an additional year of schooling for men. Thus, β_3 gives the **difference in the returns to schooling** for women and men.

† As is often the case with econometrics.

Interpreting coefficients

Log-linear specification

In economics, you will frequently see logged outcome variables with linear (non-logged) explanatory variables, *e.g.*,

$$\log(\text{Pay}_i) = \beta_0 + \beta_1 \, \text{School}_i + u_i$$

This specification changes our interpretation of the slope coefficients.

Interpretation

- A one-unit increase in our explanatory variable increases the outcome variable by approximately $\beta_1 \times 100$ percent.

- *Example:* An additional year of schooling increases pay by approximately 3 percent (for $\beta_1 = 0.03$).

Interpreting coefficients

Log-linear specification

Derivation

Consider the log-linear model

$$\log(y) = \beta_0 + \beta_1\, x + u$$

and differentiate

$$\frac{dy}{y} = \beta_1 dx$$

So a marginal change in x (*i.e.*, dx) leads to a $\beta_1 dx$ **percentage change** in y.

Log-linear specification

Because the log-linear specification comes with a different interpretation, you need to make sure it fits your data-generating process/model.

Does x change y in levels (*e.g.*, a 3-unit increase) or percentages (*e.g.*, a 10-percent increase)?

Log-linear specification

Because the log-linear specification comes with a different interpretation, you need to make sure it fits your data-generating process/model.

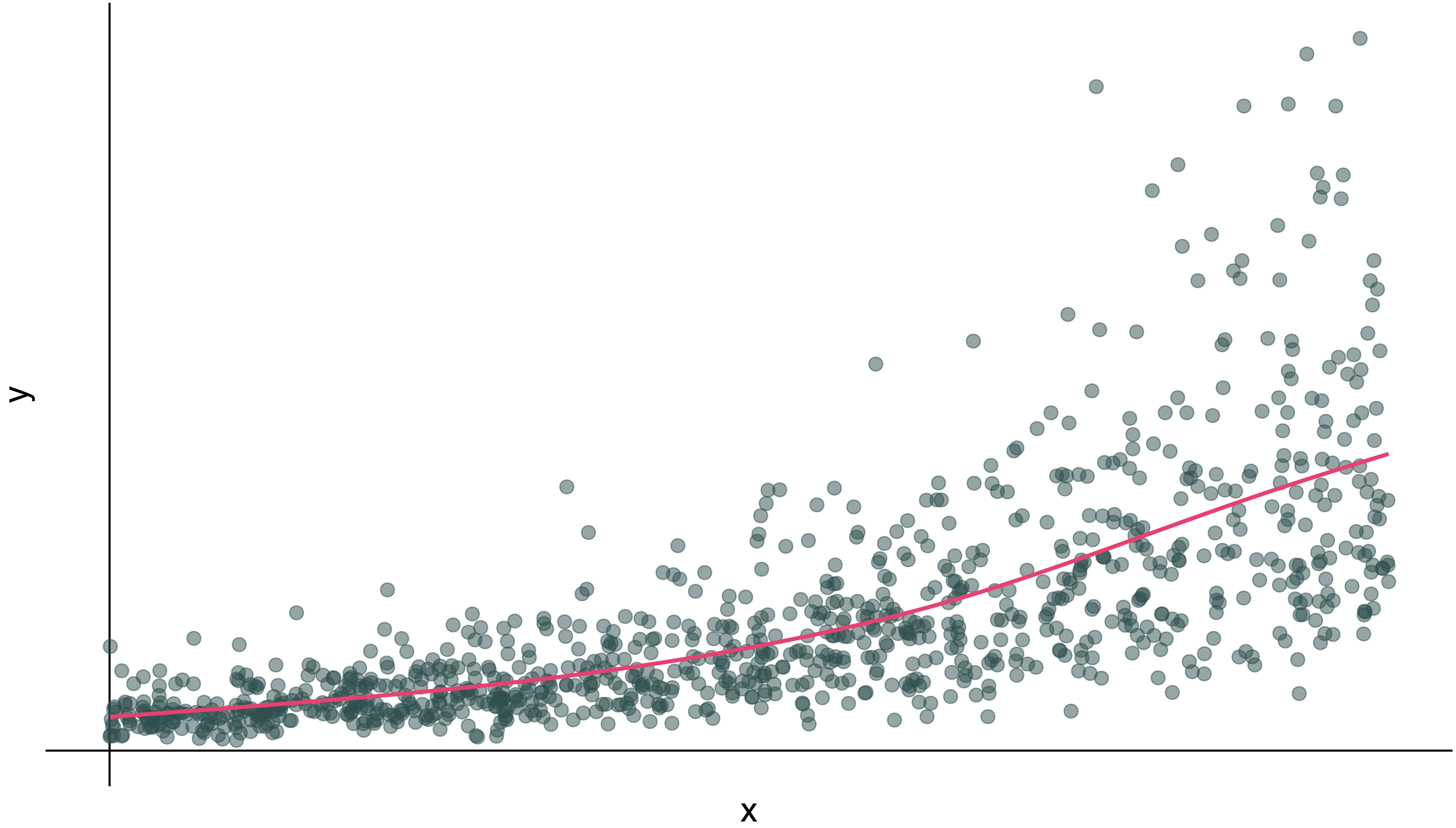
Does x change y in levels (*e.g.*, a 3-unit increase) or percentages (*e.g.*, a 10-percent increase)?

I.e., you need to be sure an exponential relationship makes sense:

$$\log(y_i) = \beta_0 + \beta_1\, x_i + u_i \iff y_i = e^{\beta_0 + \beta_1 x_i + u_i}$$

Interpreting coefficients

Log-linear specification



Interpreting coefficients

Log-log specification

Similarly, econometricians frequently employ log-log models, in which the outcome variable is logged *and* at least one explanatory variable is logged

$$\log(\text{Pay}_i) = \beta_0 + \beta_1 \log(\text{School}_i) + u_i$$

Interpretation:

- A one-percent increase in x will lead to a β_1 percent change in y.
- Often interpreted as an elasticity.

Interpreting coefficients

Log-log specification

Derivation

Consider the log-log model

$$\log(y) = \beta_0 + \beta_1 \log(x) + u$$

and differentiate

$$\frac{dy}{y} = \beta_1 \frac{dx}{x}$$

which says that for a one-percent increase in x, we will see a β_1 percent increase in y. As an elasticity:

$$\frac{dy}{dx}\frac{x}{y} = \beta_1$$

Interpreting coefficients

Log-linear with a binary variable

Note: If you have a log-linear model with a binary indicator variable, the interpretation for the coefficient on that variable changes.

Consider

$$\log(y_i) = \beta_0 + \beta_1 x_1 + u_i$$

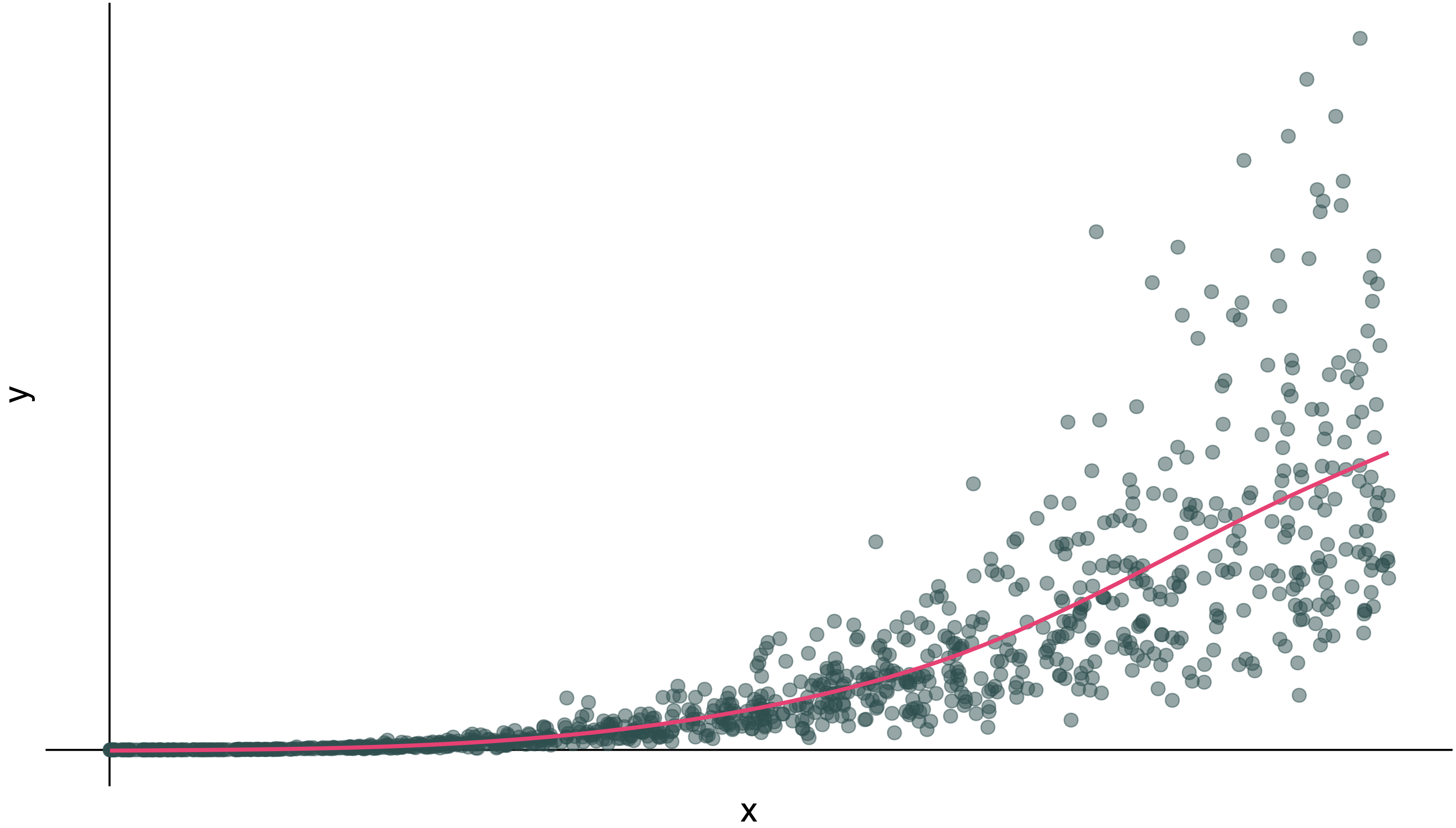
for binary variable x_1.

The interpretation of β_1 is now

- When x_1 changes from 0 to 1, y will change by $100 \times \left(e^{\beta_1} - 1\right)$ percent.
- When x_1 changes from 1 to 0, y will change by $100 \times \left(e^{-\beta_1} - 1\right)$ percent.

Interpreting coefficients

Log-log specification



Additional topics

Additional topics

Inference *vs.* prediction

So far, we've focused mainly **statistical inference**—using estimators and their distributions properties to try to learn about underlying, unknown population parameters. In OLS regression, that looks like this -

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki} + e_i$$

Additional topics

Inference *vs.* prediction

So far, we've focused mainly **statistical inference**—using estimators and their distributions properties to try to learn about underlying, unknown population parameters. In OLS regression, that looks like this -

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki} + e_i$$

Prediction includes a fairly different set of topics/tools within econometrics (and data science/machine learning)—creating models that accurately estimate individual observations.

$$\hat{y}_i = \hat{f}\,(x_1,\, x_2,\, \ldots x_k)$$

Inference *vs.* prediction

Succinctly, in stock-standard econometrics...

Inference *vs.* prediction

Succinctly, in stock-standard econometrics...

- **Inference:** causality, $\hat{\beta}_k$ (consistent and efficient), standard errors/hypothesis tests for $\hat{\beta}_k$, generally OLS
- **Prediction:**[†] correlation, $\hat{y}_i$ (low error), model selection, nonlinear models are much more common

† Includes forecasting.

Additional topics

Inference *vs.* prediction

But I want to broaden this definition - linearity does not enforce causality, the assumptions behind the models do. We can re-frame causality also as a prediction problem all of its own.

Inference *vs.* prediction

But I want to broaden this definition - linearity does not enforce causality, the assumptions behind the models do. We can re-frame causality also as a prediction problem all of its own.

Prediction is primarily interested in

$$\hat{y}_i = \hat{f}\,(x_1,\, x_2,\, \ldots x_k)$$

Additional topics

Inference *vs.* prediction

But I want to broaden this definition - linearity does not enforce causality, the assumptions behind the models do. We can re-frame causality also as a prediction problem all of its own.

Prediction is primarily interested in

$$\hat{y}_i = \hat{f}\left(x_1, x_2, \ldots x_k\right)$$

Whereas - causality is prediction *with missing data*, specifically

$$\hat{y}\_i = \hat{f}\left(x_1, x_2, \ldots x_k \middle| x_{trt} = x_{trt} + c\right)$$

There are lots of useful methods that get ignored because they are non-linear.

Treatment effects and causality

Much of modern (micro)econometrics focuses on causally estimating (*identifying*) the effect of programs/policies, *e.g.*,

- Government shutdowns
- The minimum wage
- Recreational-cannabis legalization
- Salary-history bans
- Preschool
- The Clean Water Act

Additional topics

Treatment effects and causality

Much of modern (micro)econometrics focuses on causally estimating (*identifying*) the effect of programs/policies, *e.g.*,

- Government shutdowns
- The minimum wage
- Recreational-cannabis legalization
- Salary-history bans
- Preschool
- The Clean Water Act

In this literature, the program is often a binary variable, and we place high importance on finding an unbiased estimate for the program's effect, $\hat{\tau}$.

$$\text{Outcome}_i = \beta_0 + \tau\,\text{Program}_i + u_i$$

Transformations

Our linearity assumption requires

1. **parameters enter linearly** (*i.e.*, the β_k multiplied by variables)
2. the u_i **disturbances enter additively**

We allow nonlinear relationships between y and the explanatory variables.

Transformations

Our linearity assumption requires

1. **parameters enter linearly** (*i.e.*, the β_k multiplied by variables)
2. the u_i **disturbances enter additively**

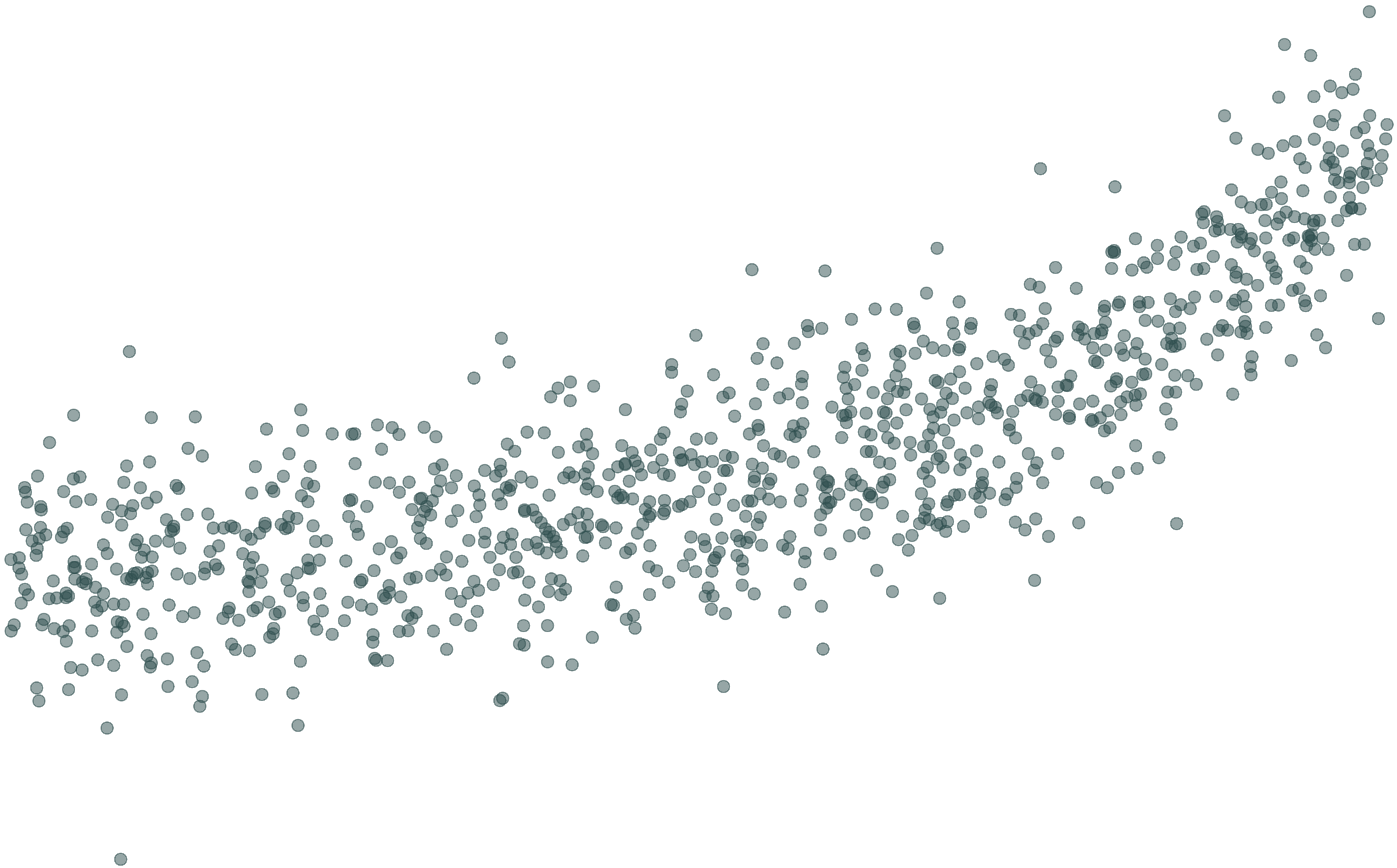
We allow nonlinear relationships between y and the explanatory variables.

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:** $y_i = \beta_0 + \beta_1 x_1 + \beta_2 \, \mathbb{I}(x_1 \geq 100) + u_i$

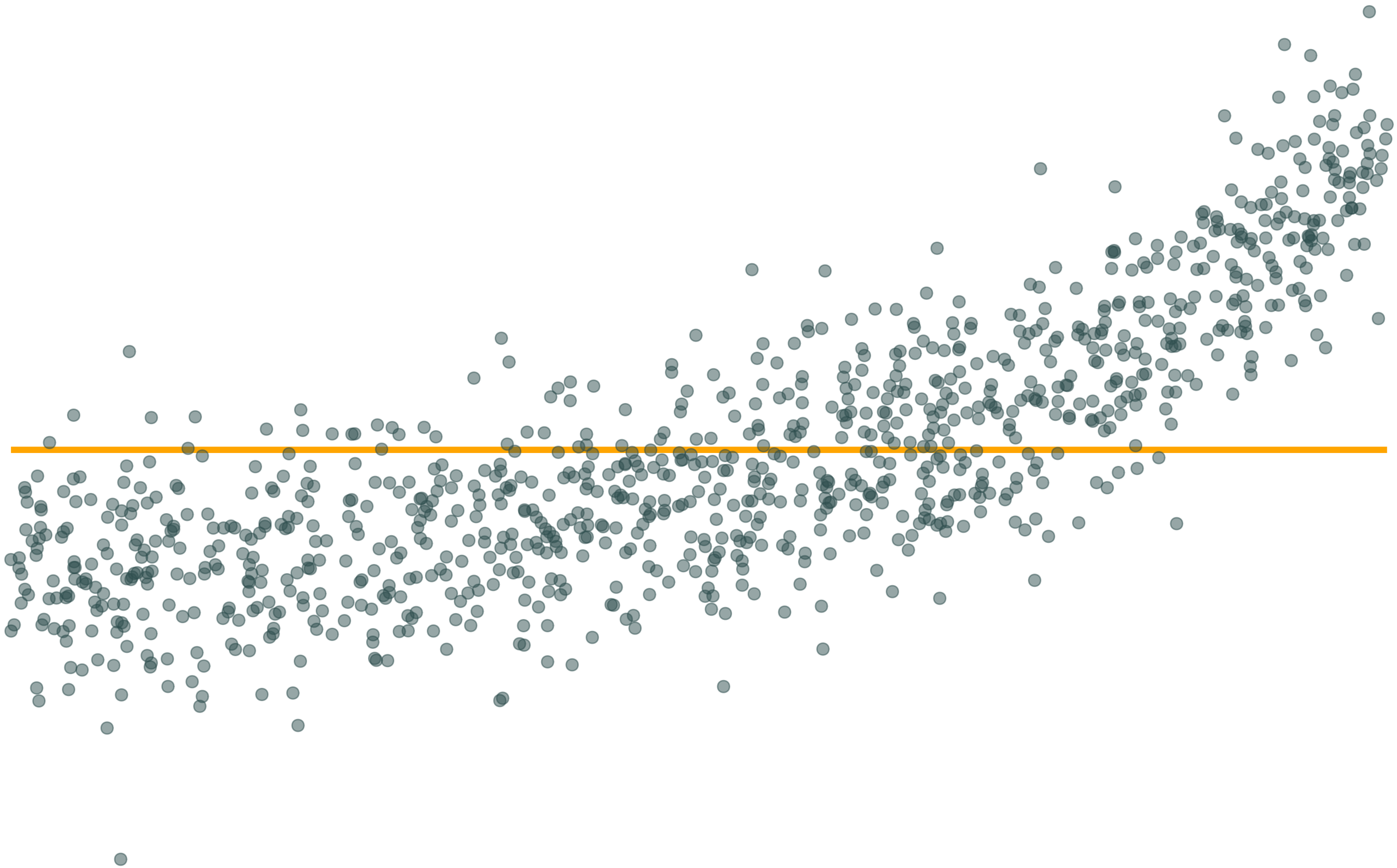
Additional topics

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



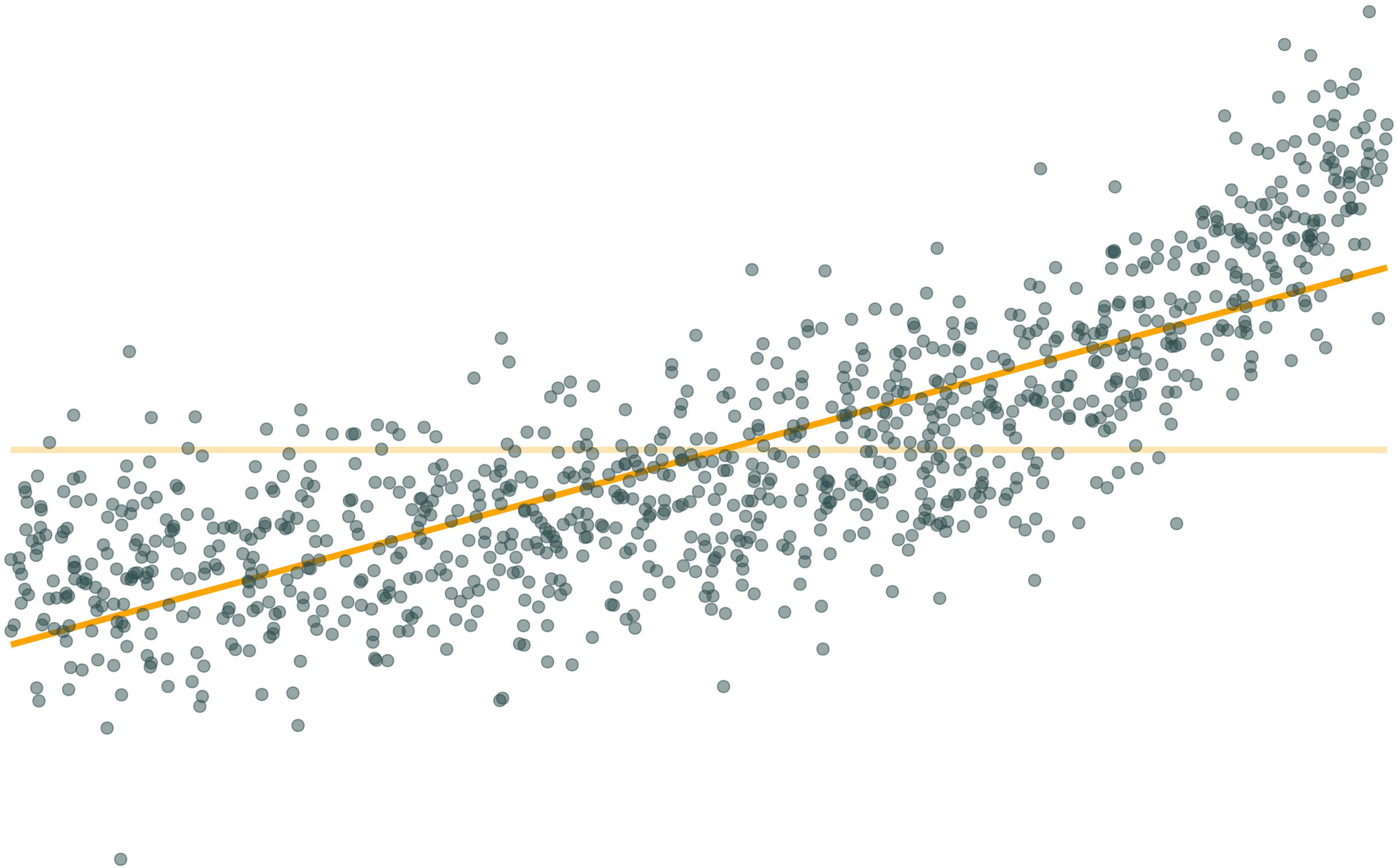
Additional topics

$y_i = \beta_0 + u_i$



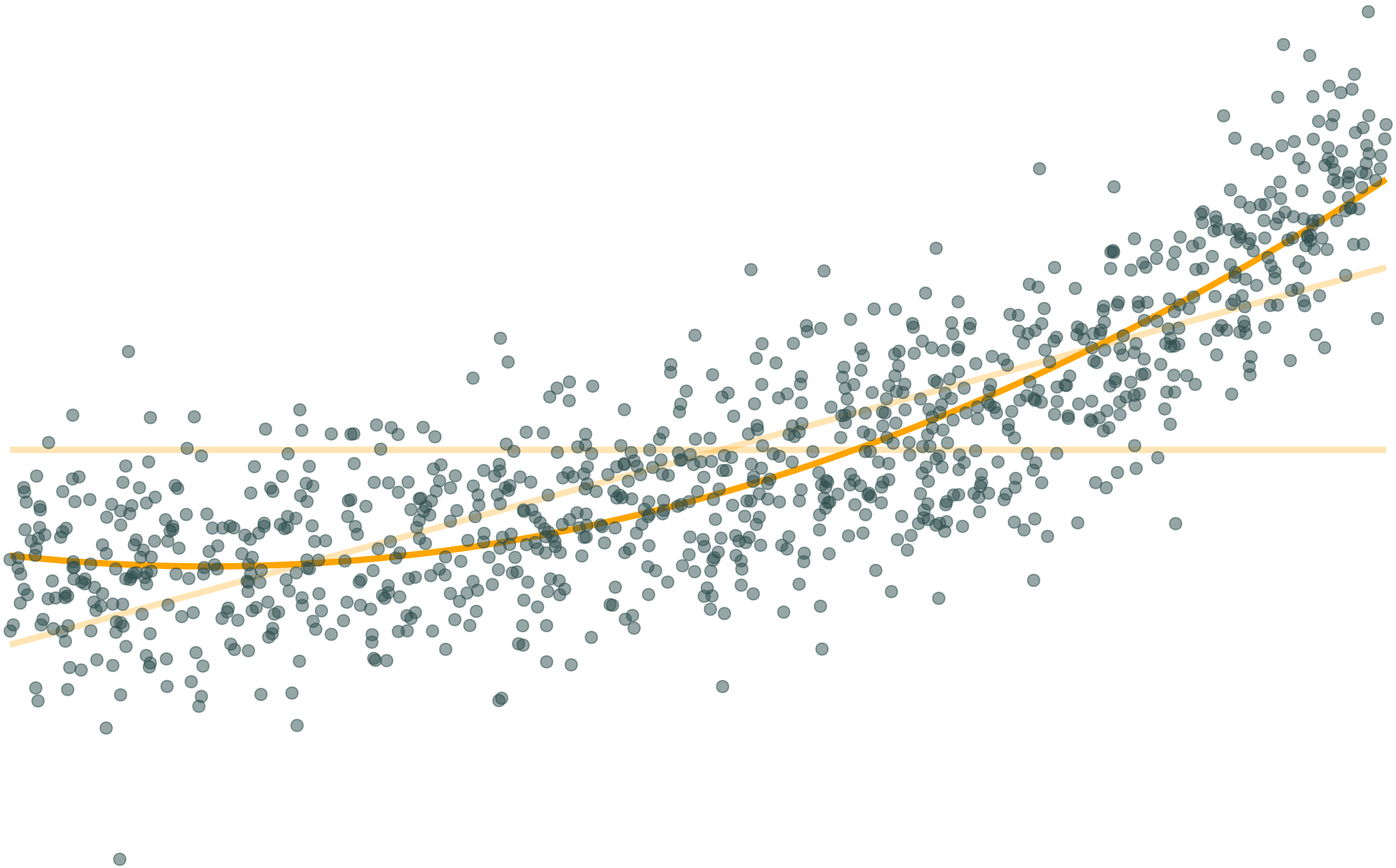
Additional topics

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



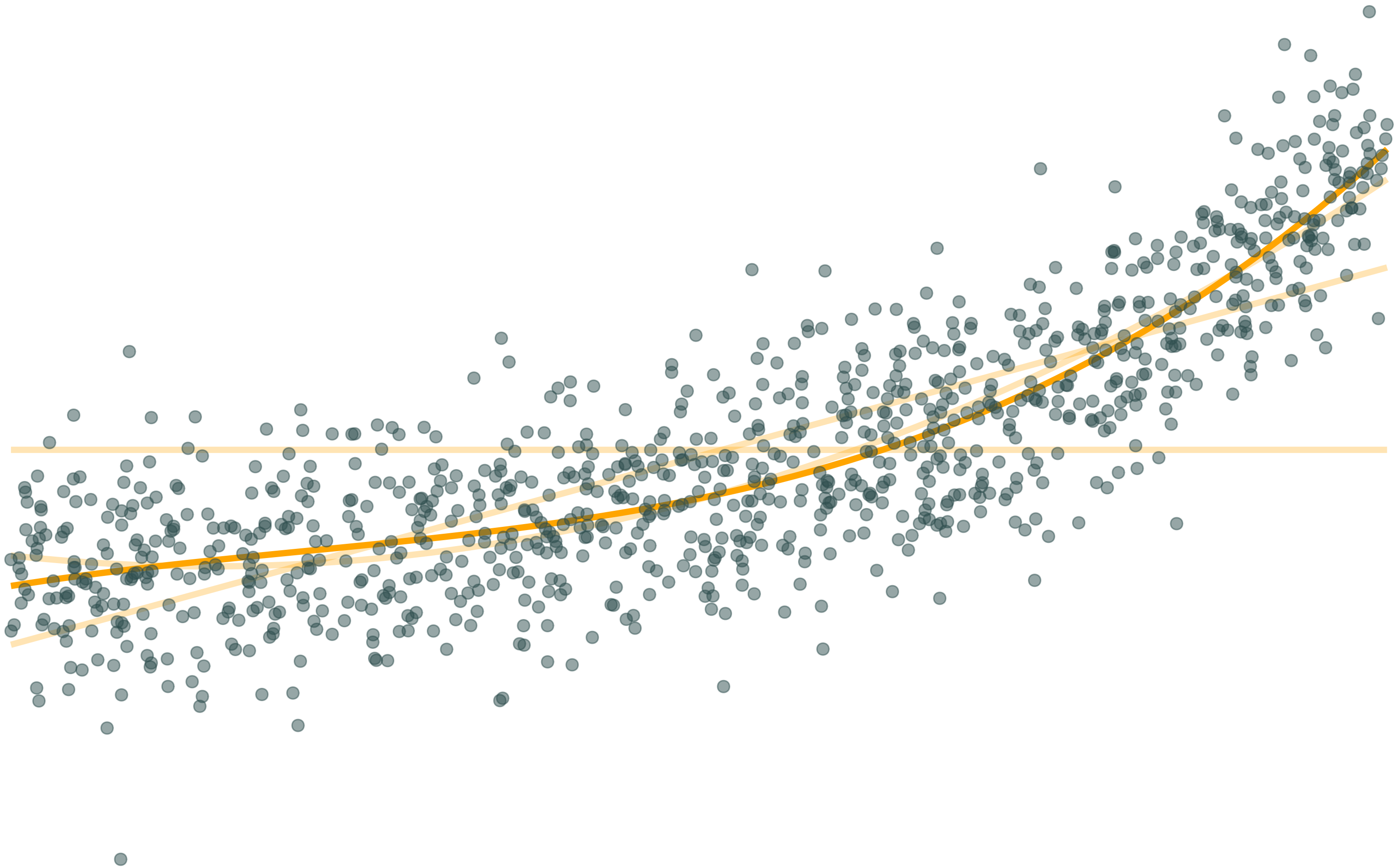
Additional topics

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



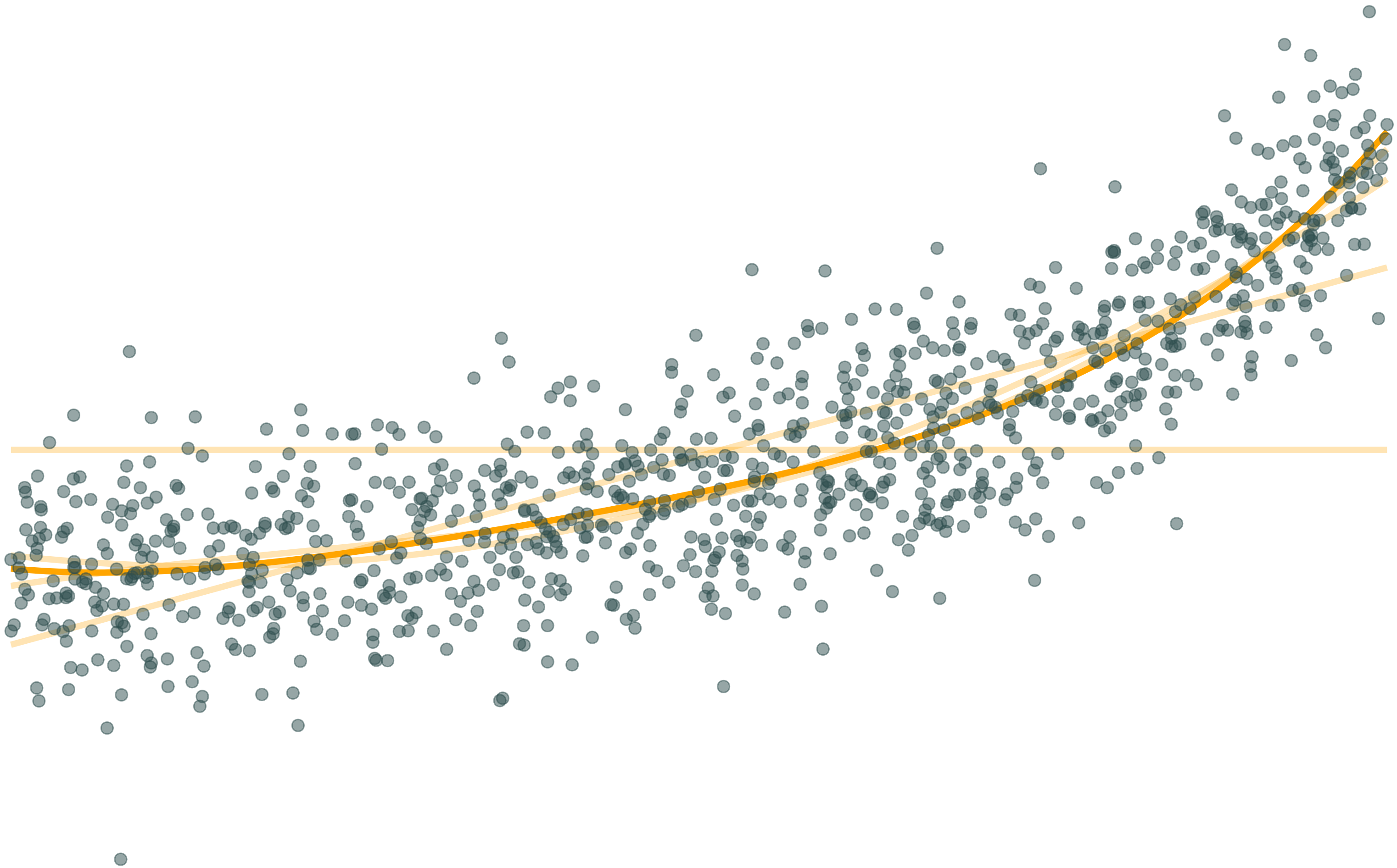
Additional topics

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



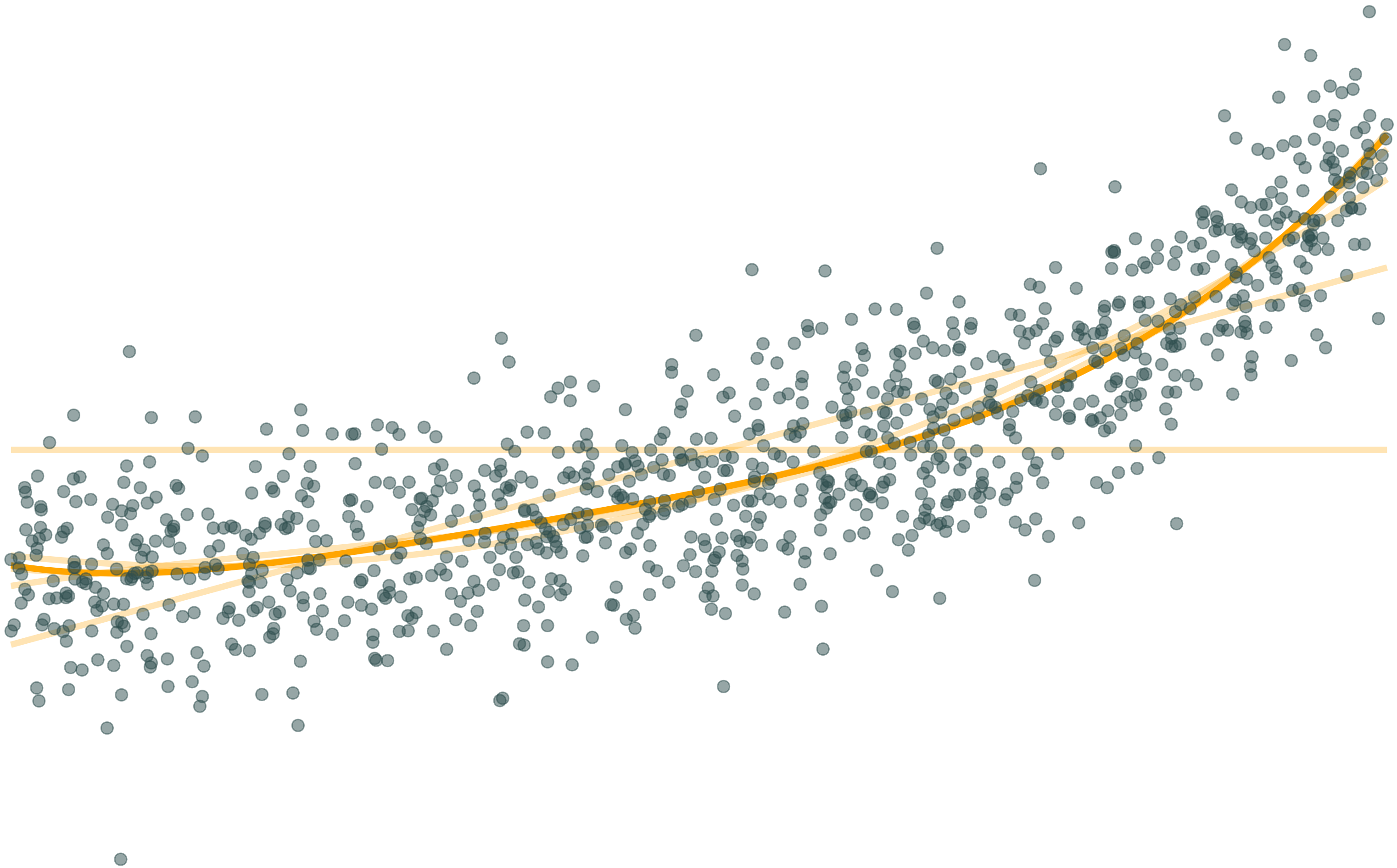
Additional topics

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



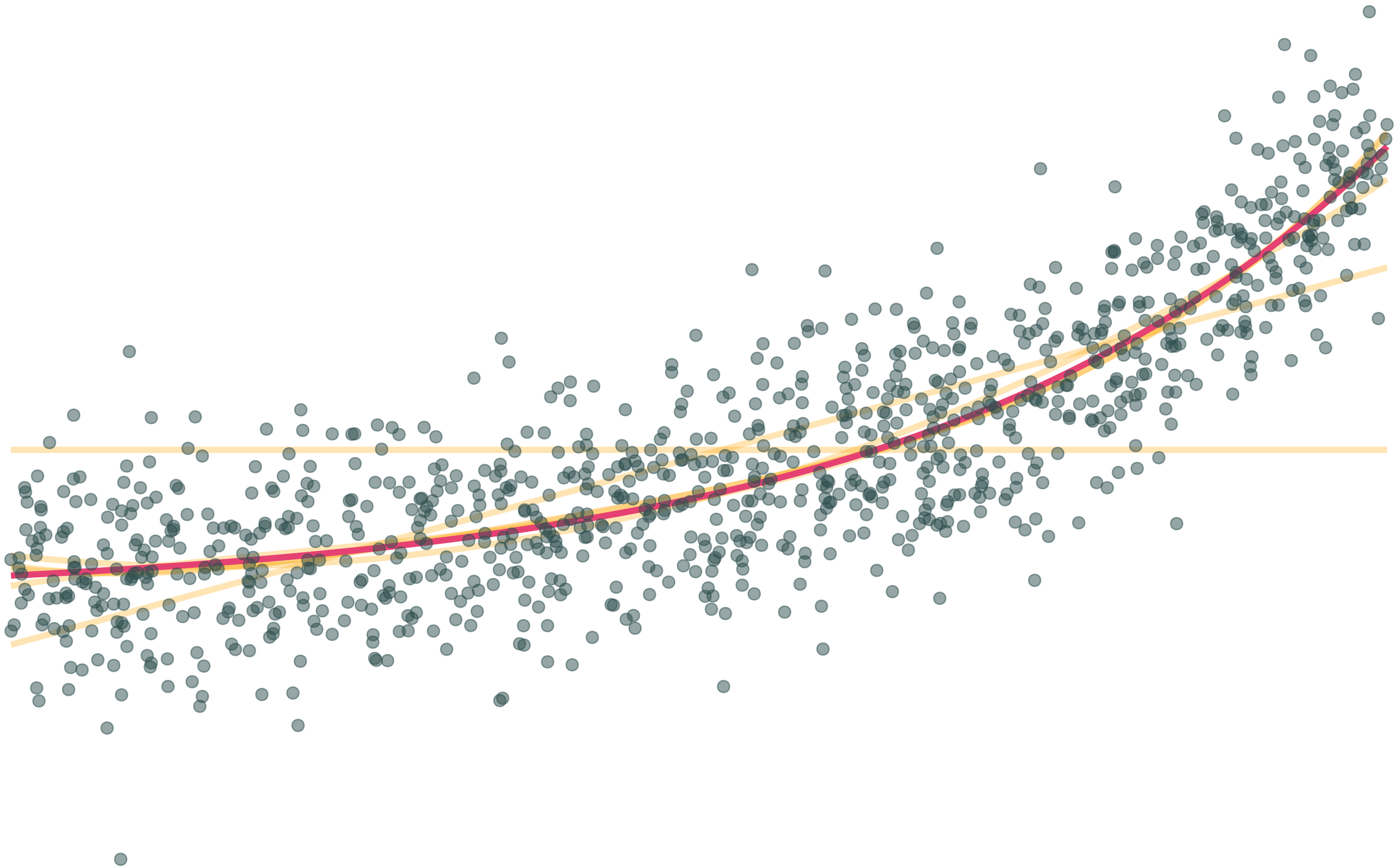
Additional topics

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



Additional topics

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



Outliers

Because OLS minimizes the sum of the **squared** errors, outliers can play a large role in our estimates.

Common responses

- Remove the outliers from the dataset
- Replace outliers with the 99th percentile of their variable (*Windsorize*)
- Take the log of the variable to "take care of" outliers
- Do nothing. Outliers are not always bad. Some people are "far" from the average. It may not make sense to try to change this variation.

Additional topics

Missing data

Similarly, missing data can affect your results.

`R` doesn't know how to deal with a missing observation.

```
1 + 2 + 3 + NA + 5
```

```
#> [1] NA
```

If you run a regression[†] with missing values, `R` drops the observations missing those values.

If the observations are missing in a nonrandom way, a random sample may end up nonrandom.

[†]: Or perform almost any operation/function