Ordinary Least Squares, continued

EDS 222

Tamma Carleton
Fall 2021

Announcements/check-in

New office hours location!! We will now meet in the Pine Room (Bren Hall 3526) for office hours so we have more space.

- Assignment #2: Grades and answers by the end of the week
- Assignment #3: Posted, due Friday 5pm
- Labs and repos: No more `git` for labs, please download directly
- Midterm heads up: week after next
- Thank you for the feedback on Slack

Today

Notes on OLS

- Unit conversions, missing data, outliers

Measures of model fit

- Coefficient of variation R^2

Categorical variables

- In `R`, interpretation

Multiple linear regression

- Adding independent variables, interpretation of results
- Nonlinearities
- Adjusted R^2
- Interaction effects [probably next time]

Notes on OLS

Units of X and Y matter

Original regression (Temperature in degrees F)

```r
mod <- lm(Ozone ~ Temp, data=airquality)
summary(mod)
```

```
#> 
#> Call:
#> lm(formula = Ozone ~ Temp, data = airquality)
#> 
#> Residuals:
#>     Min      1Q  Median      3Q     Max 
#> -40.729 -17.409  -0.587  11.306 118.271 
#> 
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)    
#> (Intercept) -146.9955    18.2872  -8.038 9.37e-13 ***
#> Temp           2.4287     0.2331  10.418  < 2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> 
#> Residual standard error: 23.71 on 114 degrees of freedom
#>   (37 observations deleted due to missingness)
```

Units of X and Y matter

New regression (Temperature in degrees C)

```
airquality$TempC <- (airquality$Temp - 32)*5/9
summary(lm(Ozone~TempC, data=airquality))
```

```
#> 
#> Call:
#> lm(formula = Ozone ~ TempC, data = airquality)
#> 
#> Residuals:
#>     Min      1Q  Median      3Q     Max 
#> -40.729 -17.409  -0.587  11.306 118.271 
#> 
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)    
#> (Intercept) -69.2770    10.9182  -6.345 4.65e-09 ***
#> TempC         4.3717     0.4196  10.418  < 2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> 
#> Residual standard error: 23.71 on 114 degrees of freedom
#>   (37 observations deleted due to missingness)
```

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 (*winsorize*)
- Take the log of the variable (This lowers the leverage of large values -- why?)
- 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.

Missing data

Similarly, missing data can affect your results.

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

```r
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.

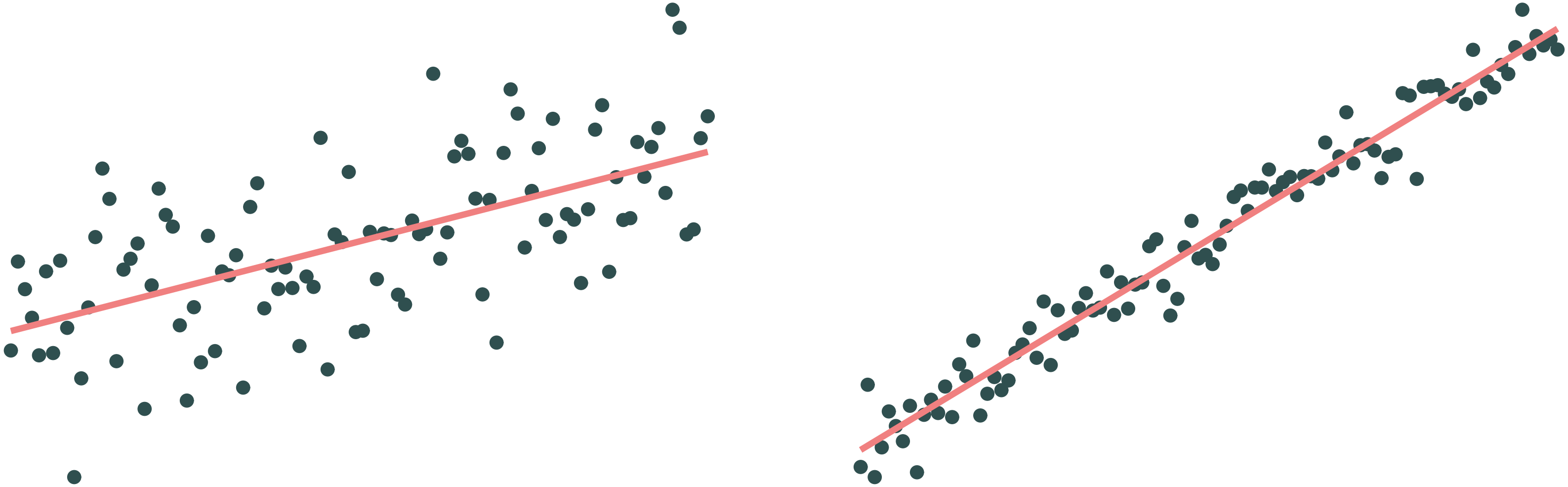
[†]: Or perform almost any operation/function

Measures of model fit

Measures of model fit

Goal: quantify how "well" your regression model fits the data

General idea: Larger variance in residuals suggests our model isn't very predictive



Coefficient of determination

- We already learned one measure of the strength of a linear relationship: correlation, r

- In OLS, we often rely on R^2, the **coefficient of determination**. In simple linear regression, this is simply the square of the correlation.

- Interpretation of R^2: **share of the variance in y that is explained by x**

$$SSR = \text{sum of squared residuals} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2$$

$$SST = \text{total sum of squares} = \sum_i (y_i - \bar{y})^2$$

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \bar{y})^2}$$

Coefficient of determination

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \bar{y})^2}$$

- R^2 varies between 0 and 1: Perfect model with $e_i = 0$ for all i has $R^2 = 1$. $R^2 = 0$ if we just guess the mean $\bar{y}$.

- In more complex models, R^2 is not the same as the square of the correlation coefficient. You should think of them as related but distinct concepts.

Coefficient of determination

About 49% of the variation in ozone can be explained with temperature alone!

```
#> 
#> Call:
#> lm(formula = Ozone ~ Temp, data = airquality)
#> 
#> Residuals:
#>     Min      1Q  Median      3Q     Max 
#> -40.729 -17.409  -0.587  11.306 118.271 
#> 
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)    
#> (Intercept) -146.9955    18.2872  -8.038 9.37e-13 ***
#> Temp           2.4287     0.2331  10.418  < 2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> 
#> Residual standard error: 23.71 on 114 degrees of freedom
#>   (37 observations deleted due to missingness)
#> Multiple R-squared:  0.4877,    Adjusted R-squared:  0.4832 
#> F-statistic: 108.5 on 1 and 114 DF,  p-value: < 2.2e-16
```

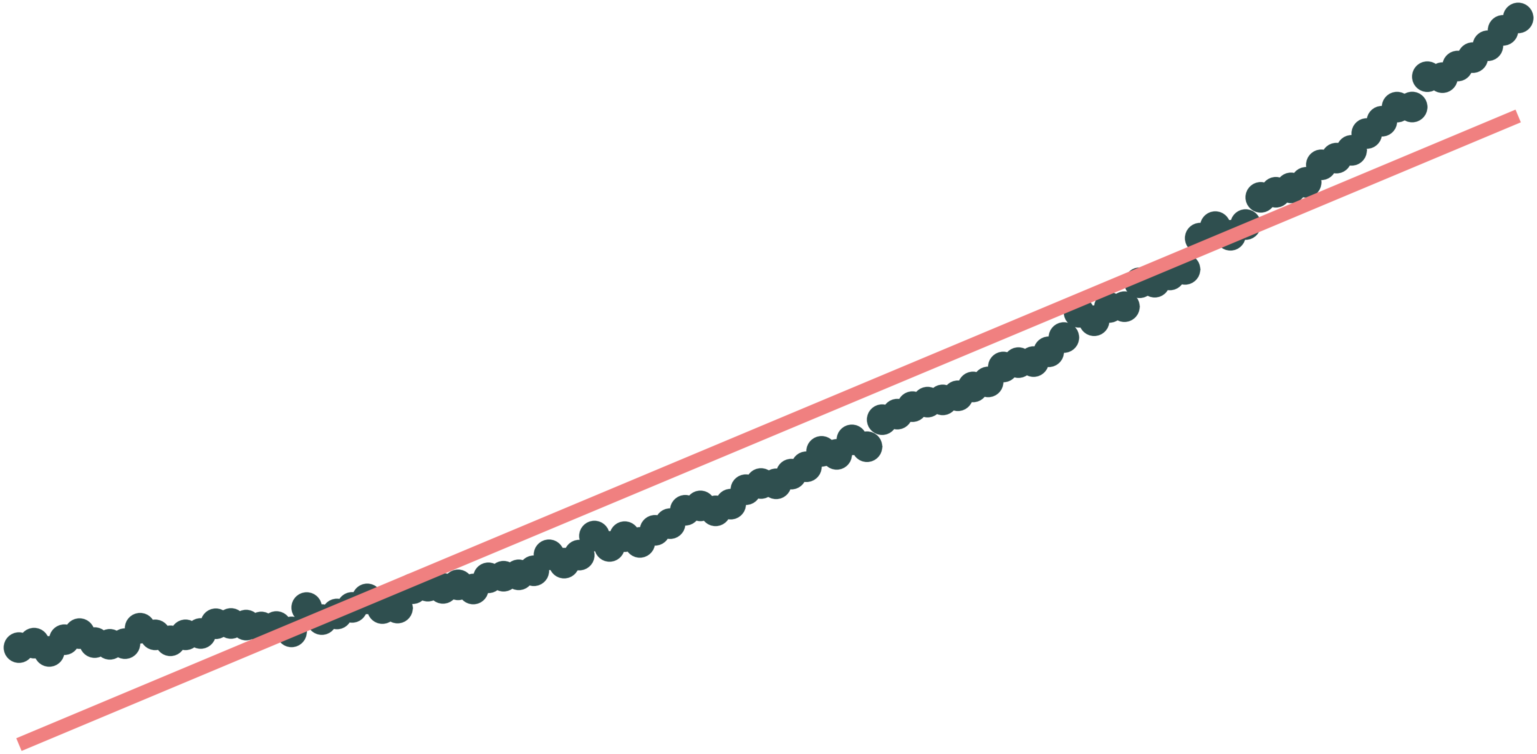
Coefficient of determination

Definition: % of variance in y that is explained by x (and any other independent variables)

- Describes a *linear* relationship between y and $\hat{y}$
- Higher R^2 does not mean a model is "better" or more appropriate
    - Predictive power is not often the goal of regression analysis (e.g., you may just care about getting β_1 right)
    - If you are focused on predictive power, many other measures of fit are appropriate (to discuss in machine learning)
    - Always look at your data and residuals!
- Like OLS in general, R^2 is very sensitive to outliers. Again...always look at your data!

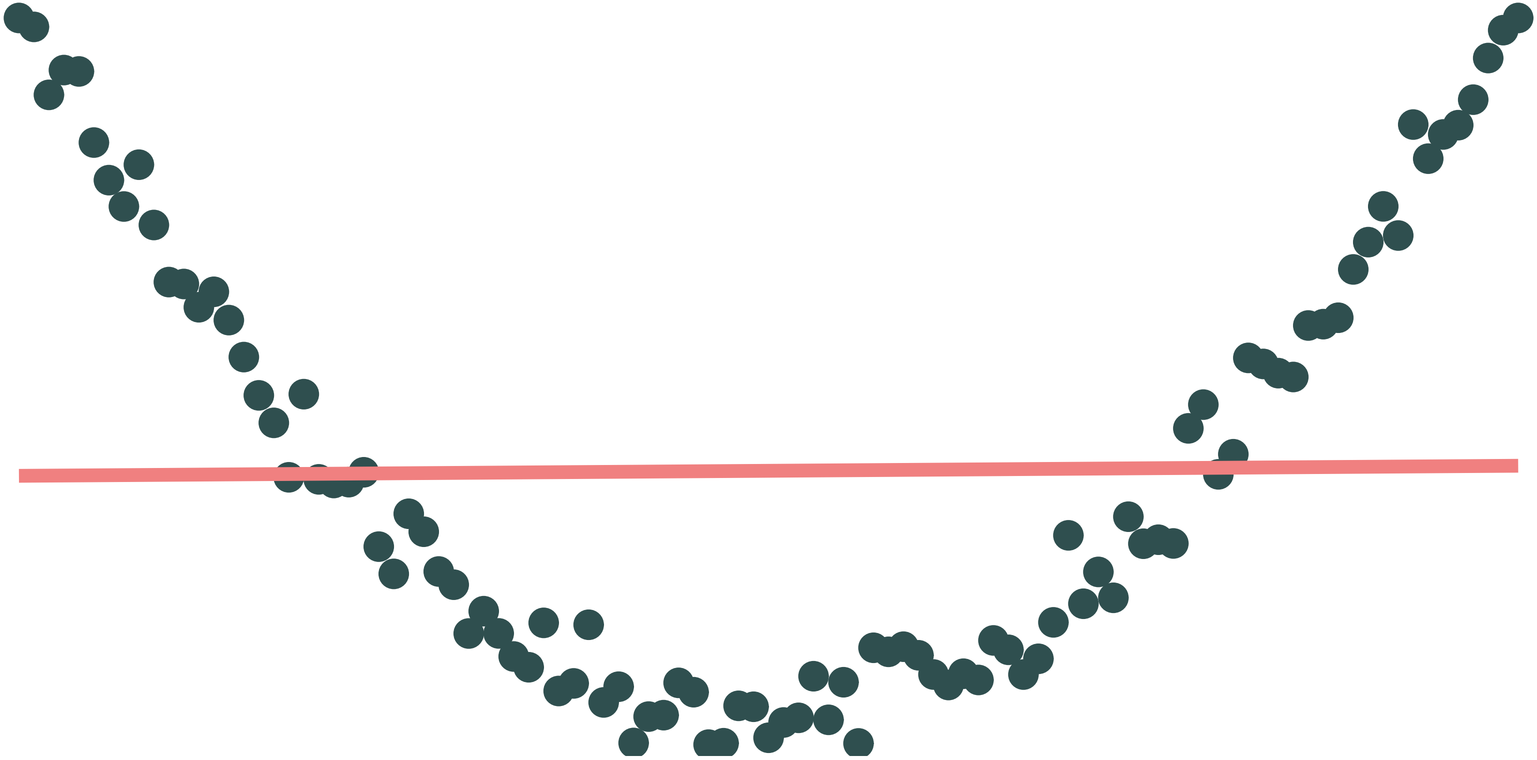
Coefficient of determination

Here, $R^2 = \mathbf{0.94}$. Does that mean a linear model is appropriate?



Coefficient of determination

Here, $R^2 = 0$. Does that mean there is no relationship between these variables?



Indicator/categorical variables

Categorical variables

We have been talking a lot about **numerical** variables in linear regression...

- Ozone levels
- Possom tail lengths
- Temperature and precipitation amounts
- etc.

...but a lot of variables of interest are **categorical**:

- Male/female
- Presence/absence of a species
- In/out of compliance with a pollution standard
- etc.

How do we execute and interpret linear regression with categorical data?

Categorical variables

We use **dummy** or **indicator** variables in linear regression to capture the influence of a categorical independent variable (*x*) on a continuous dependent variable (*y*).

For example, let *x* be a categorical variable indicating the gender of an individual. Suppose we are interested in the "gender wage gap", so *y* is wages. We estimate:

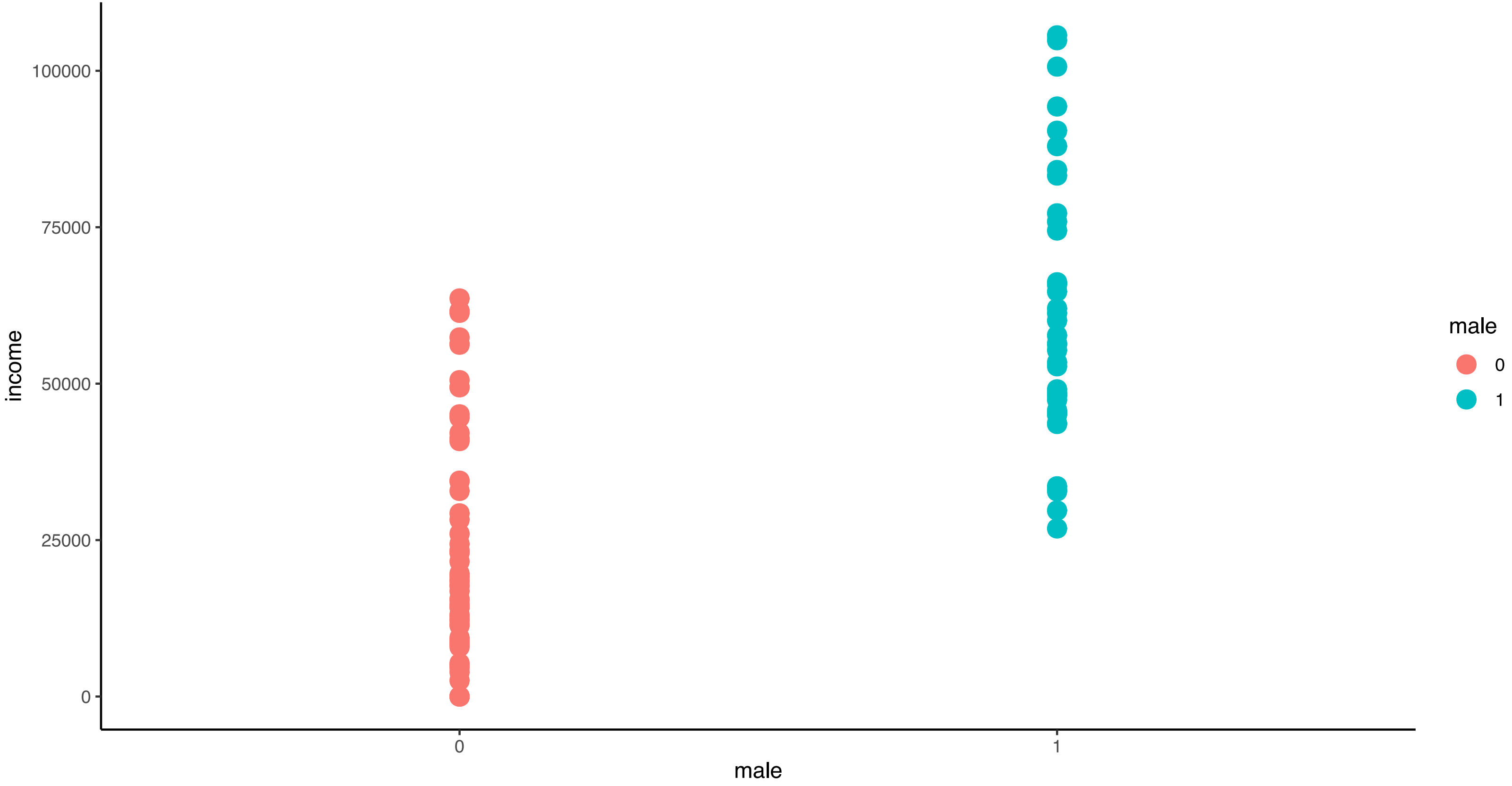
$$y_i = \beta_0 + \beta_1 MALE_i + \varepsilon_i$$

Interpretation [draw it]:

- $MALE_i$ is an **indicator** variable that = 1 when i is male (0 otherwise)
- $\beta_0 =$ average wages if i is **not** male
- $\beta_0 + \beta_1 =$ average wages if i is male
- $\beta_1 =$ average *difference* in wages between males and females

Categorical variables

For a categorical variable with two "levels", OLS reports the difference in means across the two groups



Categorical variables

What if I have many categories?

- E.g., species, education level, age group, …

For example, let *x* be a categorical variable indicating the species of penguin, and *y* is body mass. We estimate:

$$y_i = \beta_0 + \beta_1 SPECIES_i + \varepsilon_i$$

Where **species** can be one of:

- Adelie
- Chinstrap
- Gentoo

Categorical variables

```r
library(palmerpenguins)
head(penguins)
```

```
#> # A tibble: 6 x 8
#>   species island bill_length_mm bill_depth_mm flipper_length_… body_mass_g sex
#>   <fct>   <fct>           <dbl>         <dbl>            <int>       <int> <fct>
#> 1 Adelie  Torge…           39.1          18.7              181        3750 male
#> 2 Adelie  Torge…           39.5          17.4              186        3800 fema…
#> 3 Adelie  Torge…           40.3          18                195        3250 fema…
#> 4 Adelie  Torge…           NA            NA                 NA          NA <NA>
#> 5 Adelie  Torge…           36.7          19.3              193        3450 fema…
#> 6 Adelie  Torge…           39.3          20.6              190        3650 male
#> # … with 1 more variable: year <int>
```

```r
class(penguins$species)
```

```
#> [1] "factor"
```

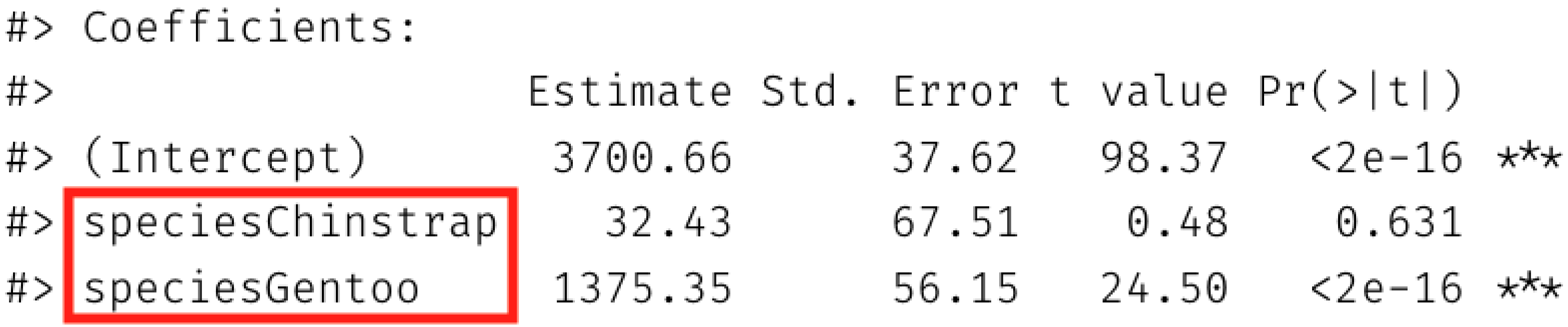
Categorical variables

```
summary(lm(body_mass_g ~ species, data = penguins))
```

```
#> 
#> Call:
#> lm(formula = body_mass_g ~ species, data = penguins)
#> 
#> Residuals:
#>      Min       1Q   Median       3Q      Max 
#> -1126.02  -333.09   -33.09   316.91  1223.98 
#> 
#> Coefficients:
#>                  Estimate Std. Error t value Pr(>|t|)    
#> (Intercept)       3700.66      37.62   98.37   <2e-16 ***
#> speciesChinstrap    32.43      67.51    0.48    0.631    
#> speciesGentoo     1375.35      56.15   24.50   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> 
#> Residual standard error: 462.3 on 339 degrees of freedom
#>   (2 observations deleted due to missingness)
#> Multiple R-squared:  0.6697,	Adjusted R-squared:  0.6677 
#> F-statistic: 343.6 on 2 and 339 DF,  p-value: < 2.2e-16
```

Categorical variables

When your categorical variable takes on k values, R will create dummy variables for $k-1$ values, leaving one as the **reference** group:



```
#> Coefficients:
#>                   Estimate Std. Error t value Pr(>|t|)
#> (Intercept)        3700.66      37.62   98.37   <2e-16 ***
#> speciesChinstrap     32.43      67.51    0.48    0.631
#> speciesGentoo      1375.35      56.15   24.50   <2e-16 ***
```

To evaluate the outcome for the reference group, **set the dummy variables equal to zero for all other groups**.

> Q: What is the average body mass of an Adelie species?
>
> Q: What is the difference in body mass between Chinstrap and Adelie?

Multiple linear regression

More explanatory variables

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

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

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...

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

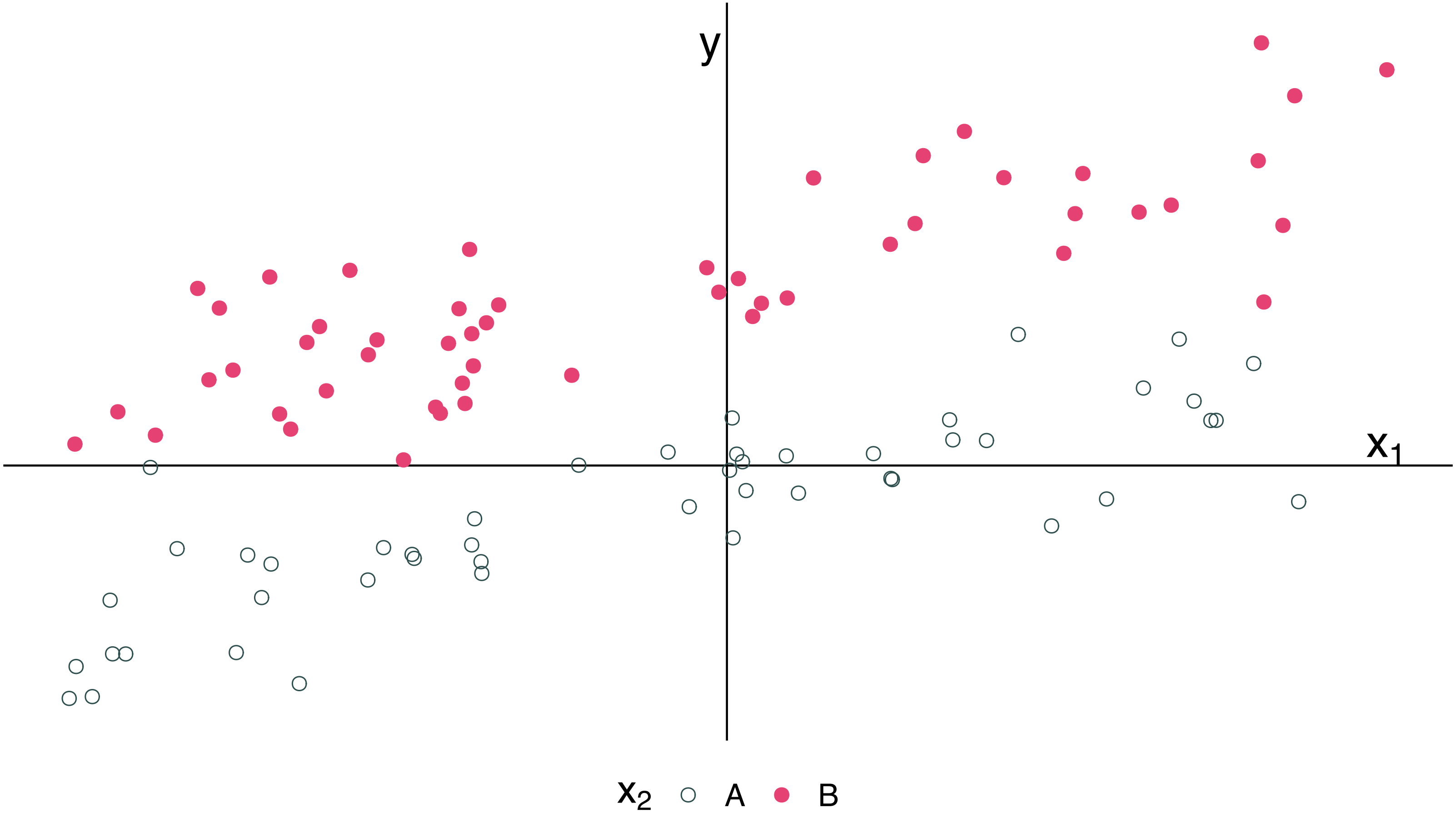
... raises many questions:

- 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

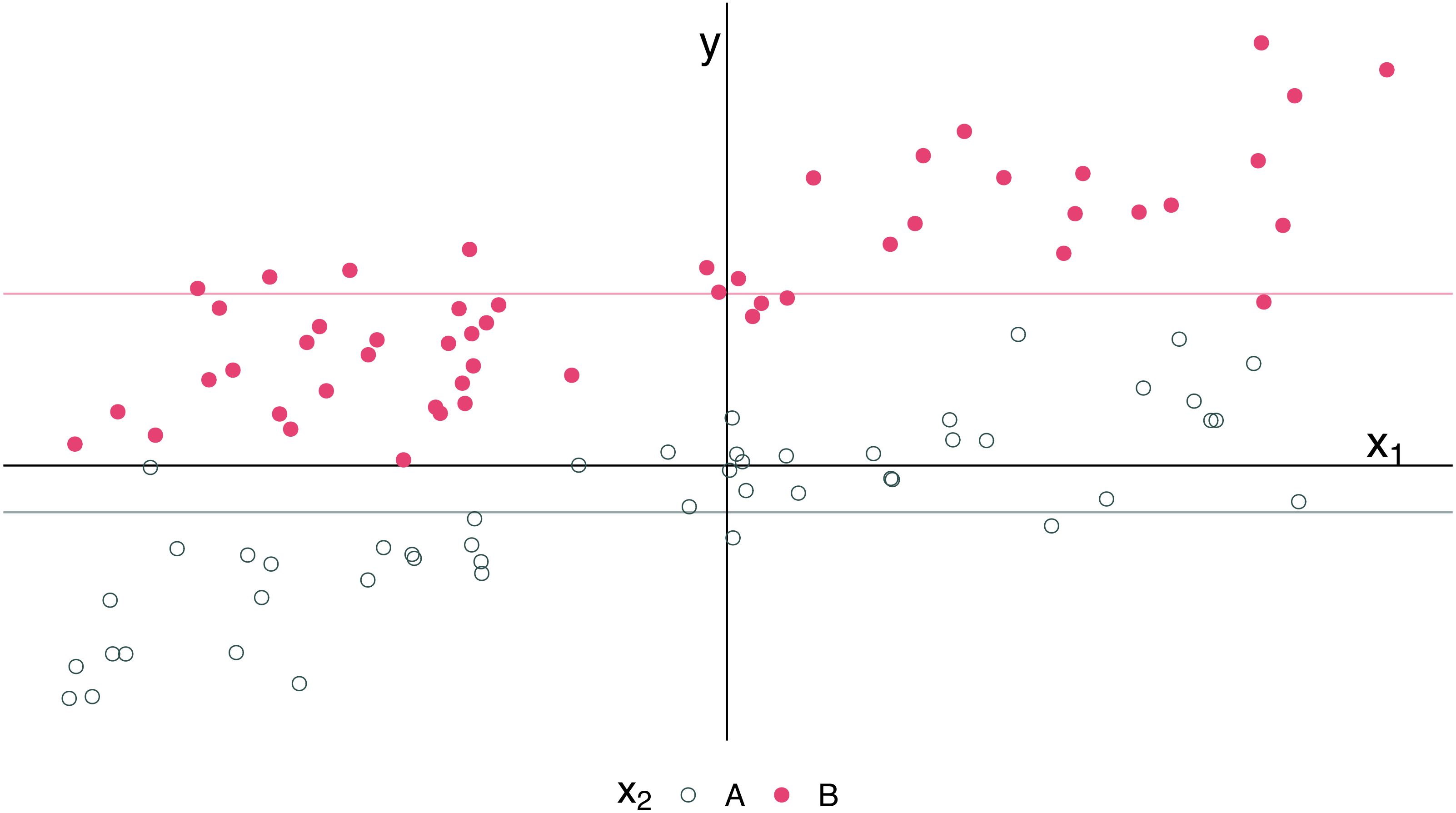
Multiple regression

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



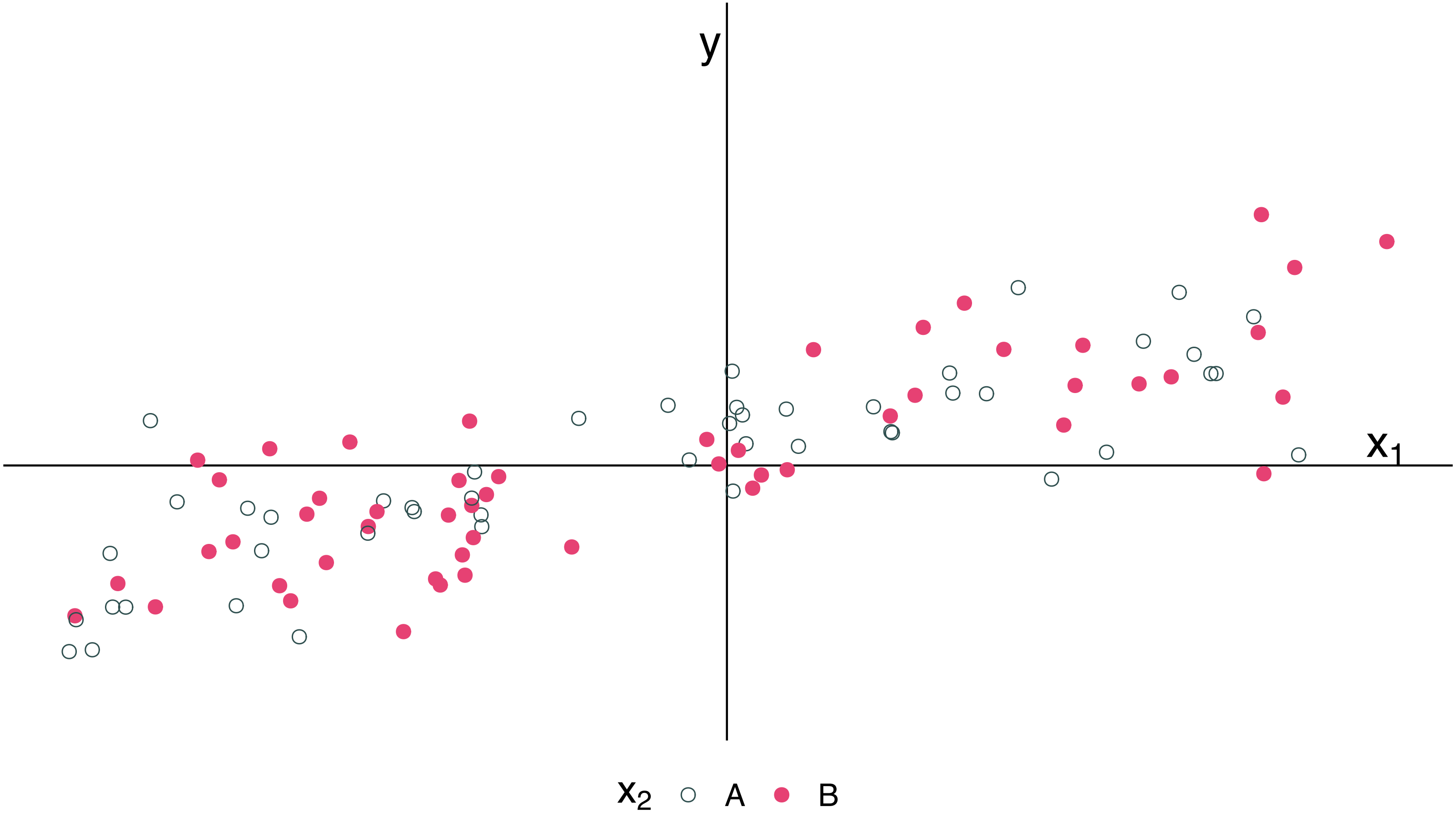
Multiple regression

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



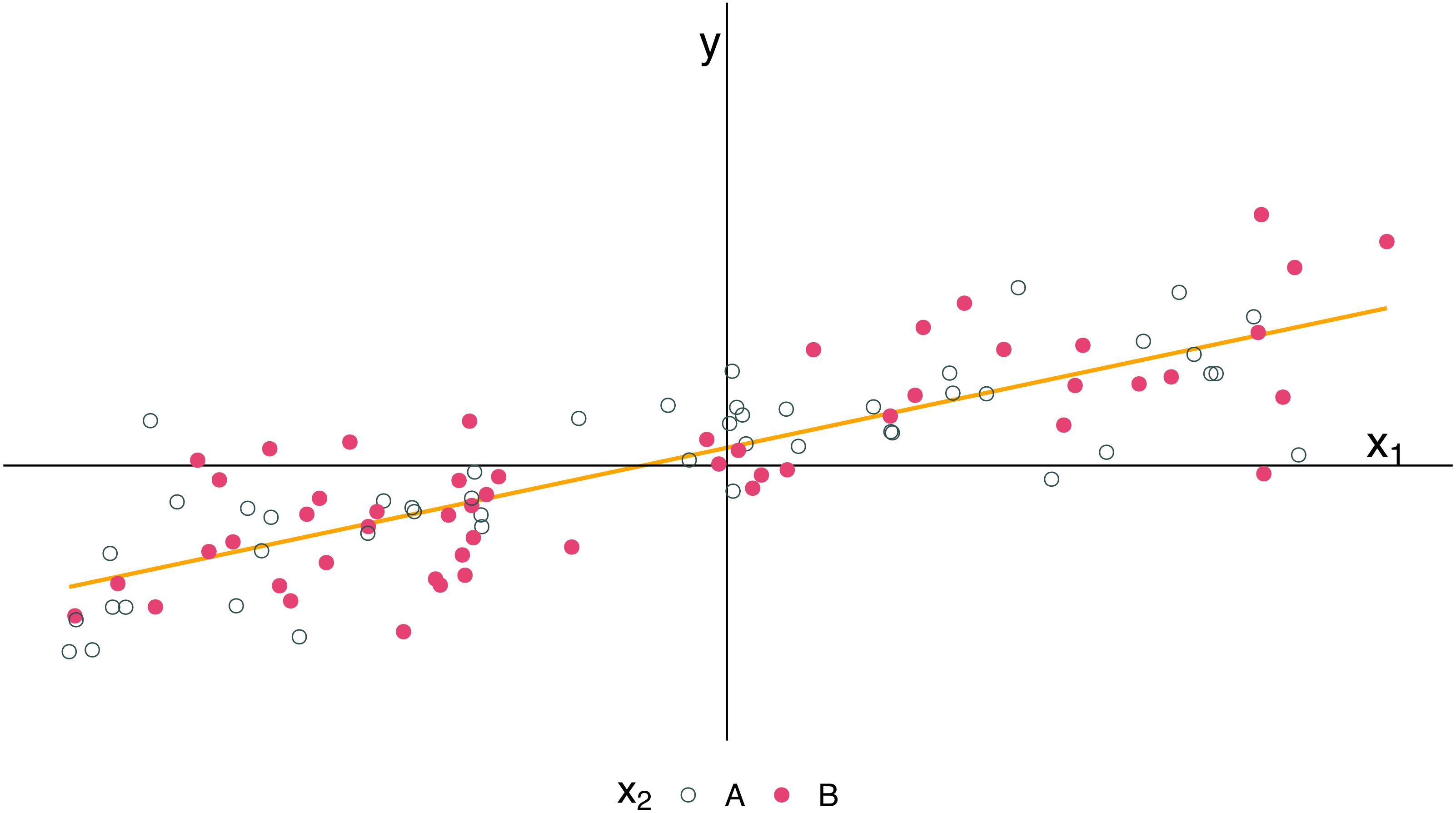
Multiple regression

With groups' means removed:



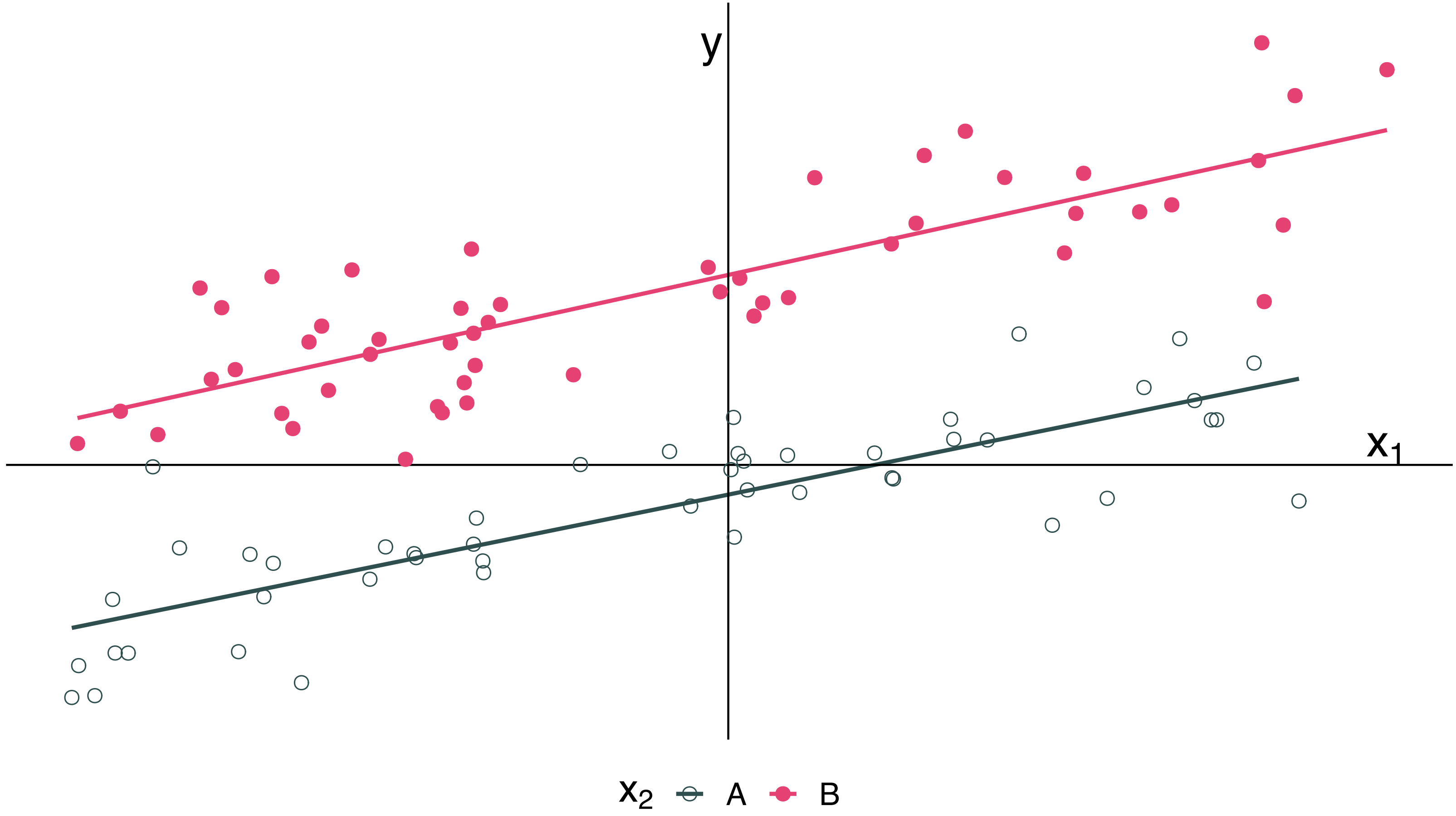
Multiple regression

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



Multiple regression

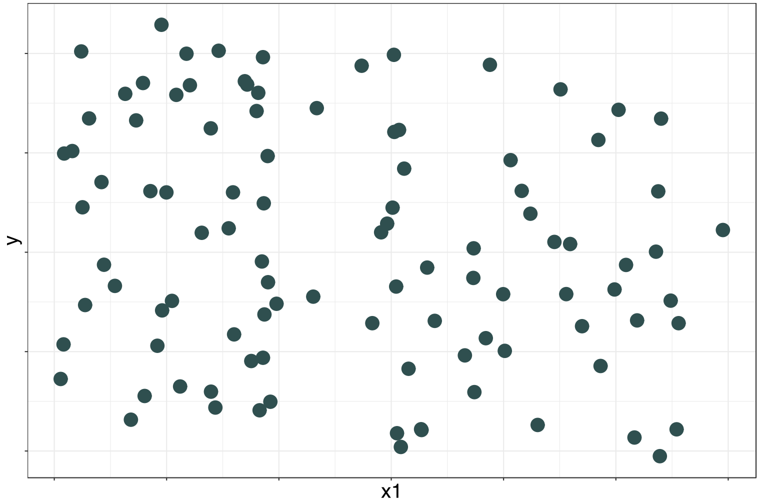
Another way to think about it:

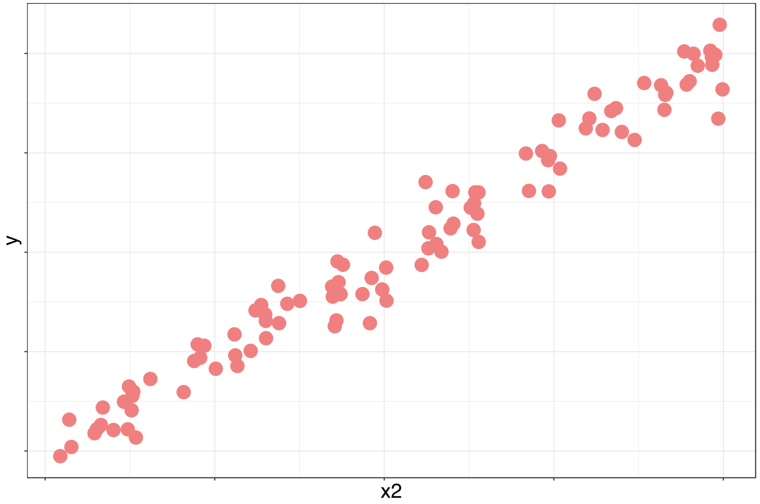


Multiple regression

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

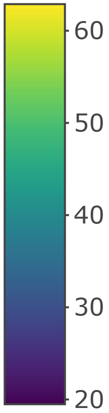
Suppose $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$





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

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

- 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

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 from last week.

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.

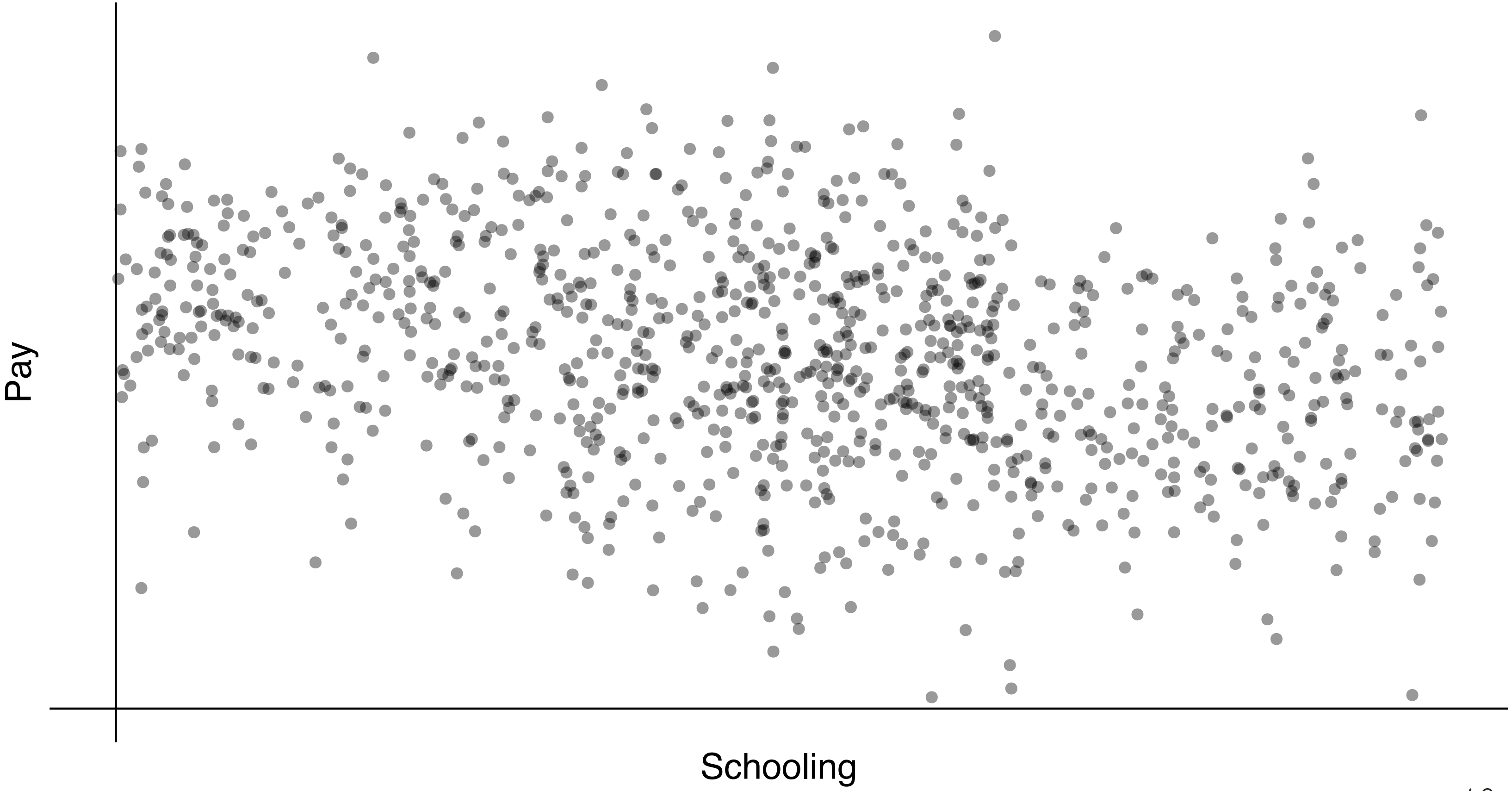
thus

- 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

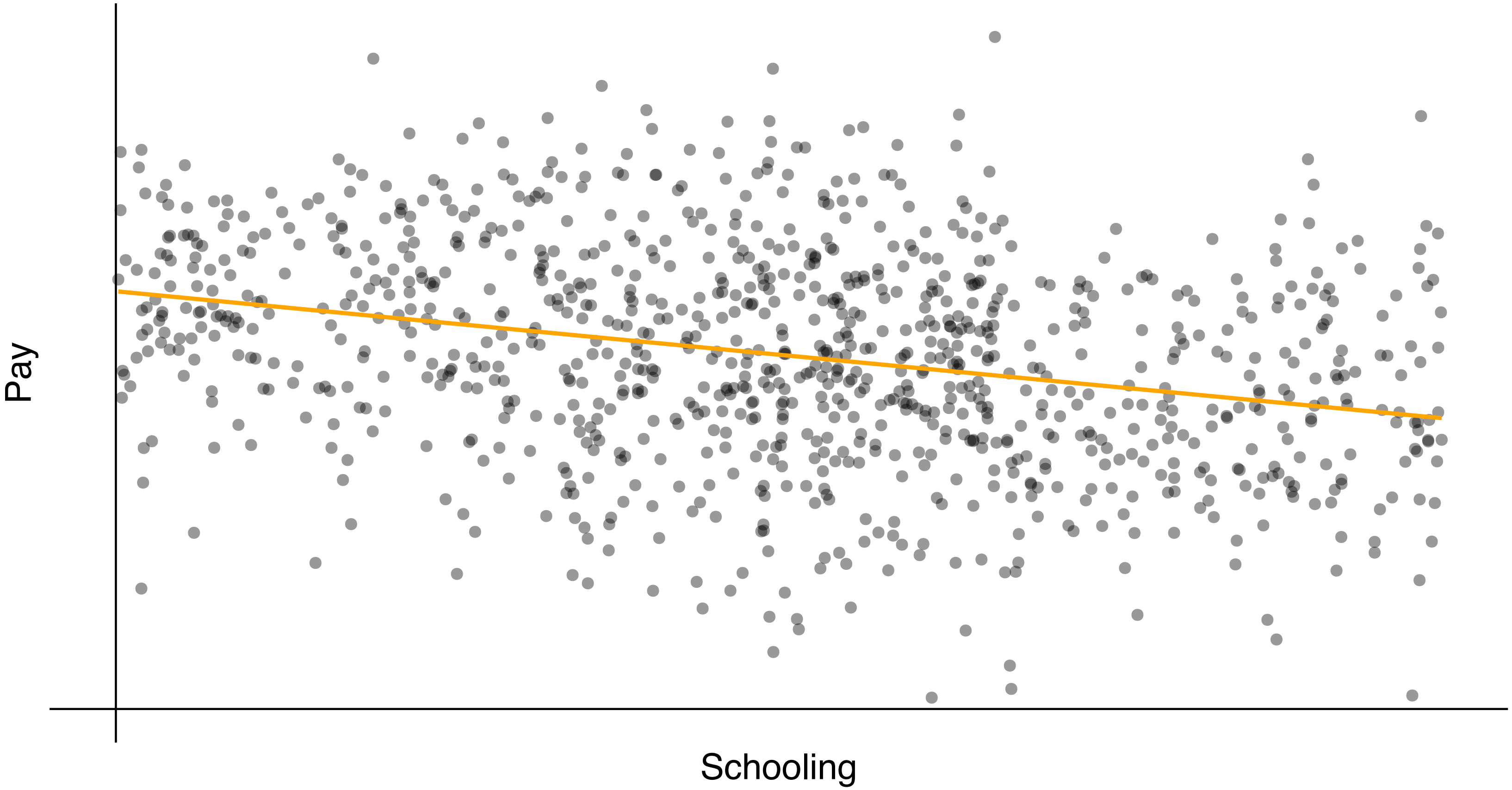
Example, continued: $\text{Pay}_i = 20 + 0.5 \times \text{School}_i + 10 \times \text{Male}_i + u_i$

The relationship between pay and schooling.



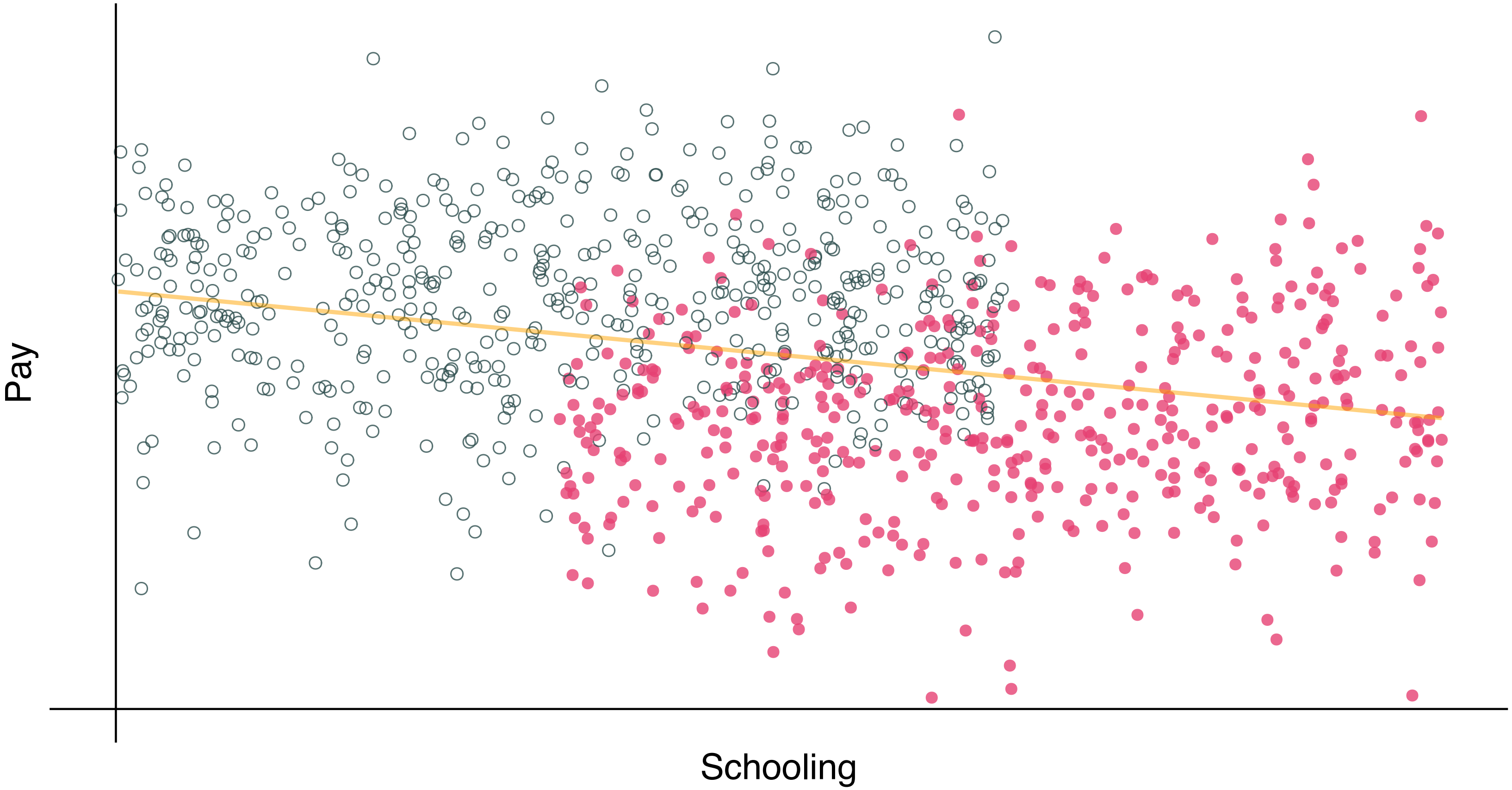
Omitted-variable bias

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



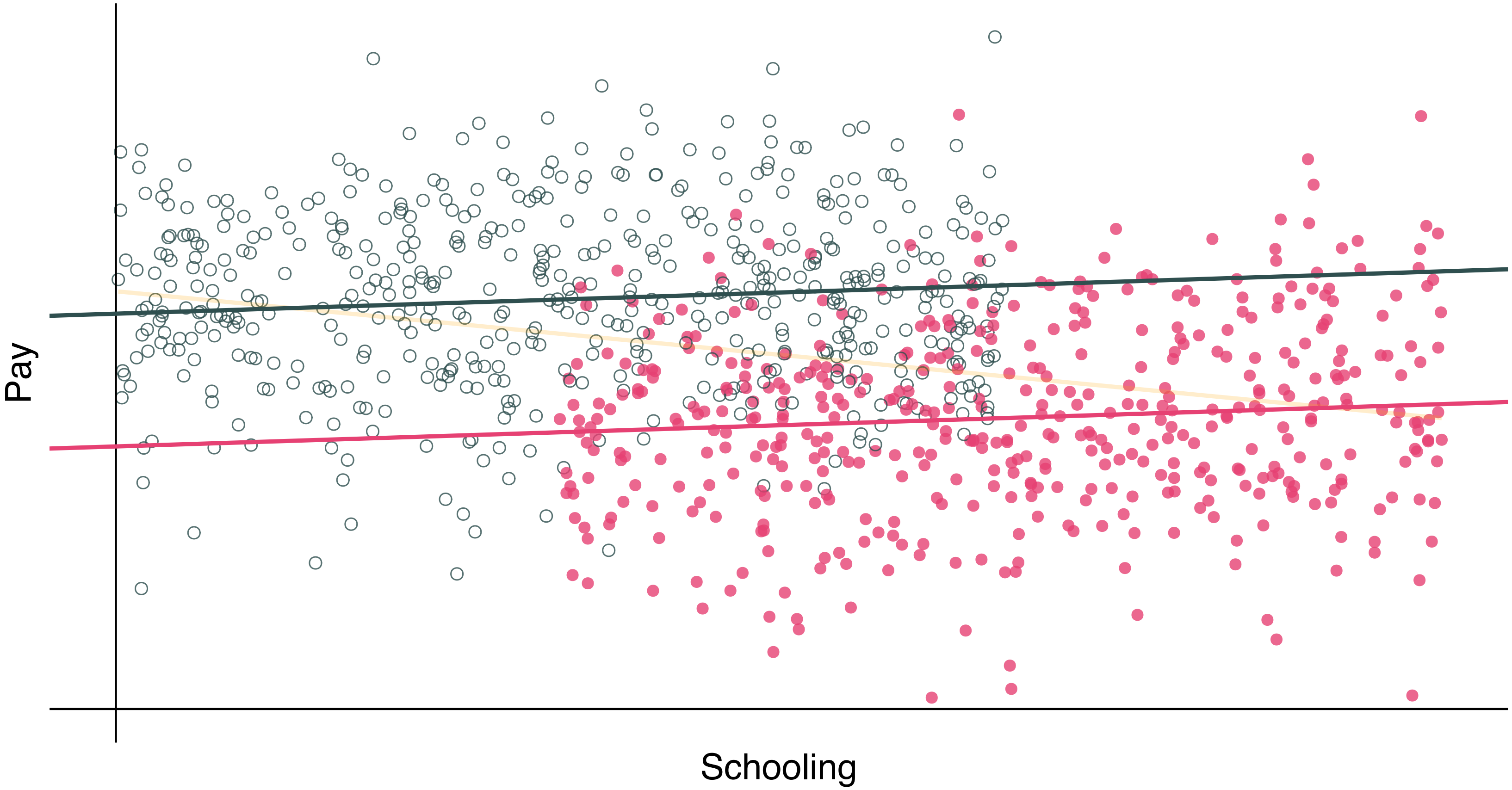
Omitted-variable bias

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



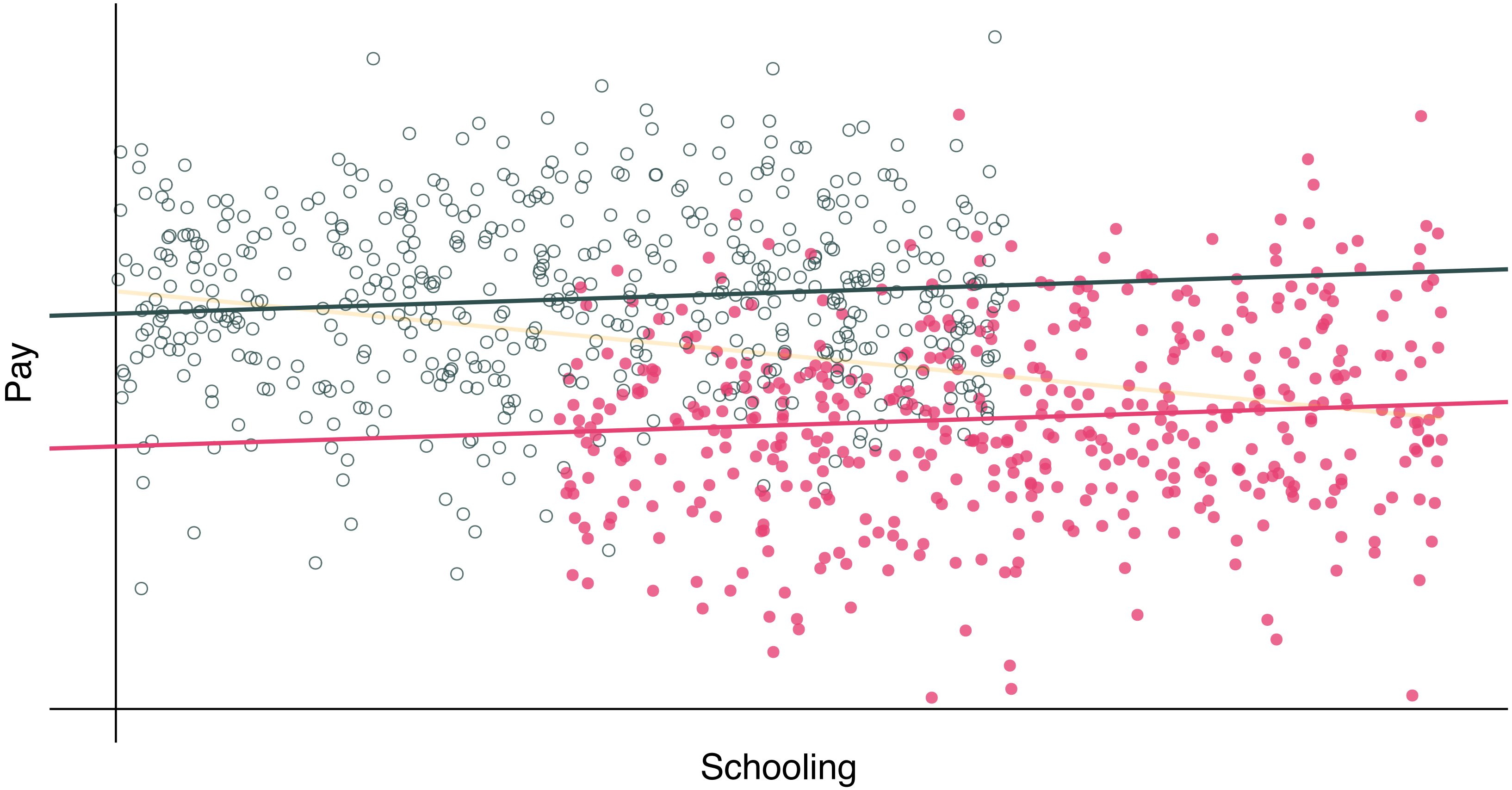
Omitted-variable bias

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



Omitted-variable bias

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



Model fit in multiple regression

Nonlinear transformations

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

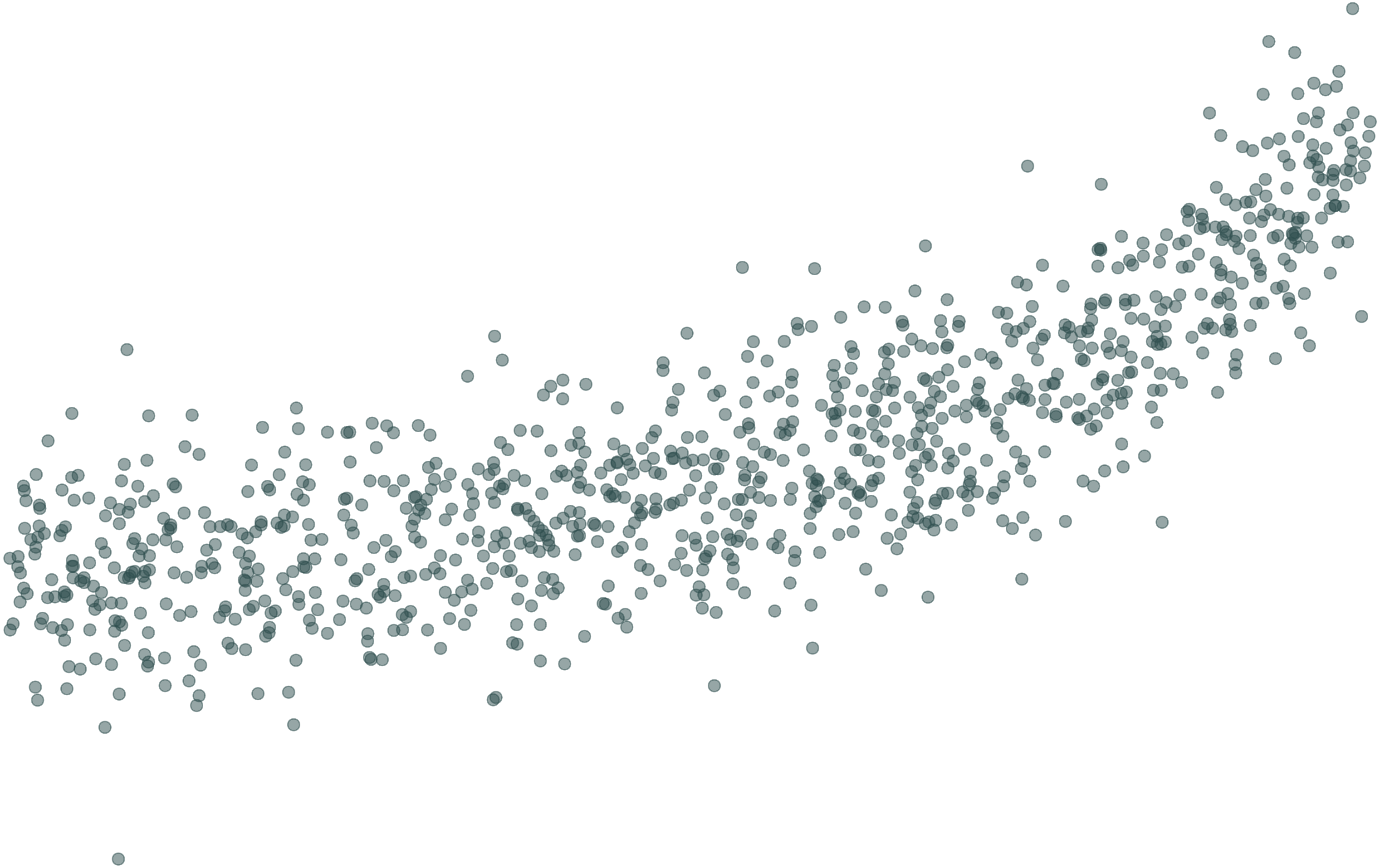
We allow nonlinear relationships between y and the explanatory variables x.

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?



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.

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.

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**

- 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:

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

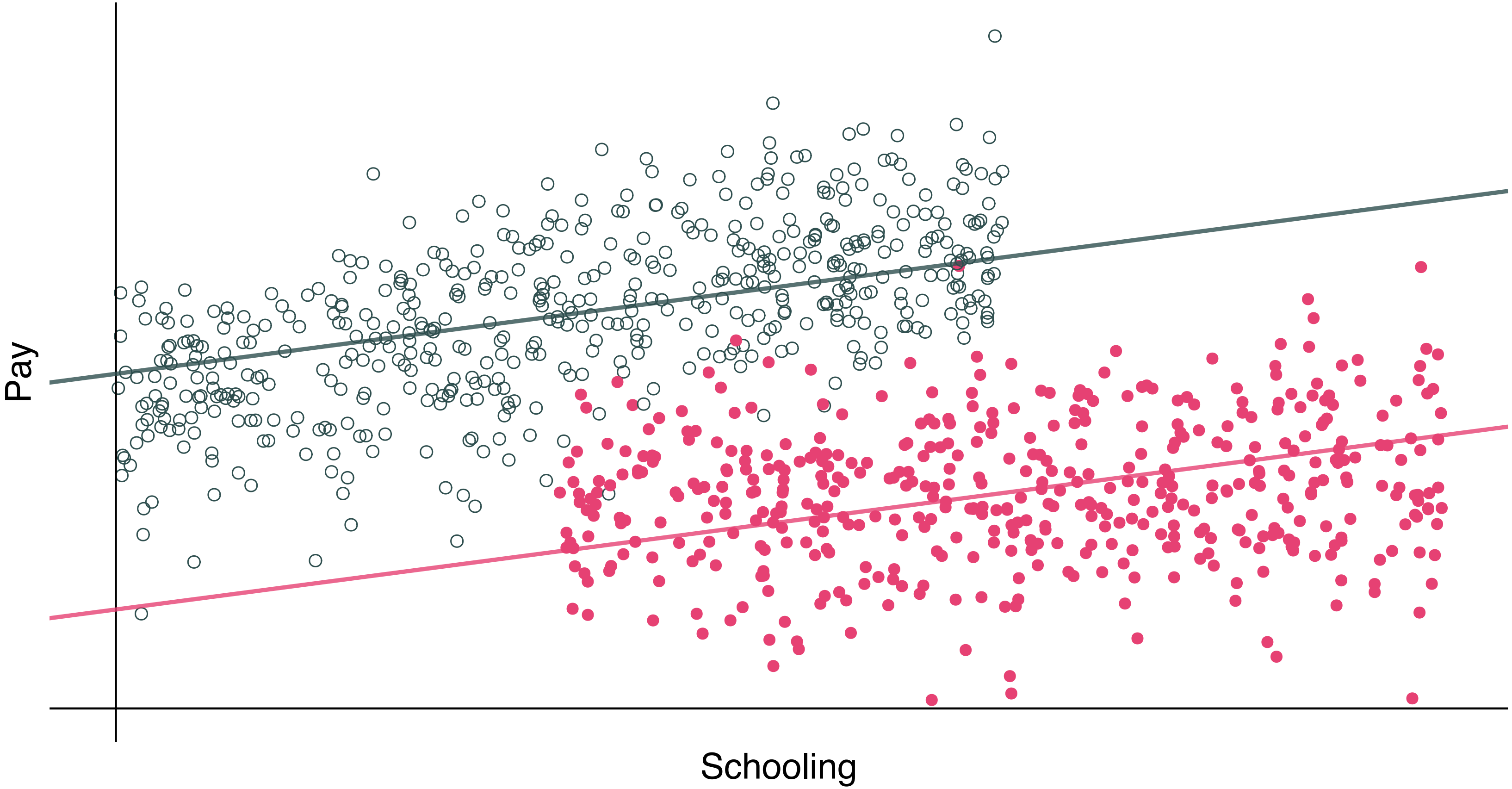
but we can also allow the effect of school to vary by gender:

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

Interactions

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

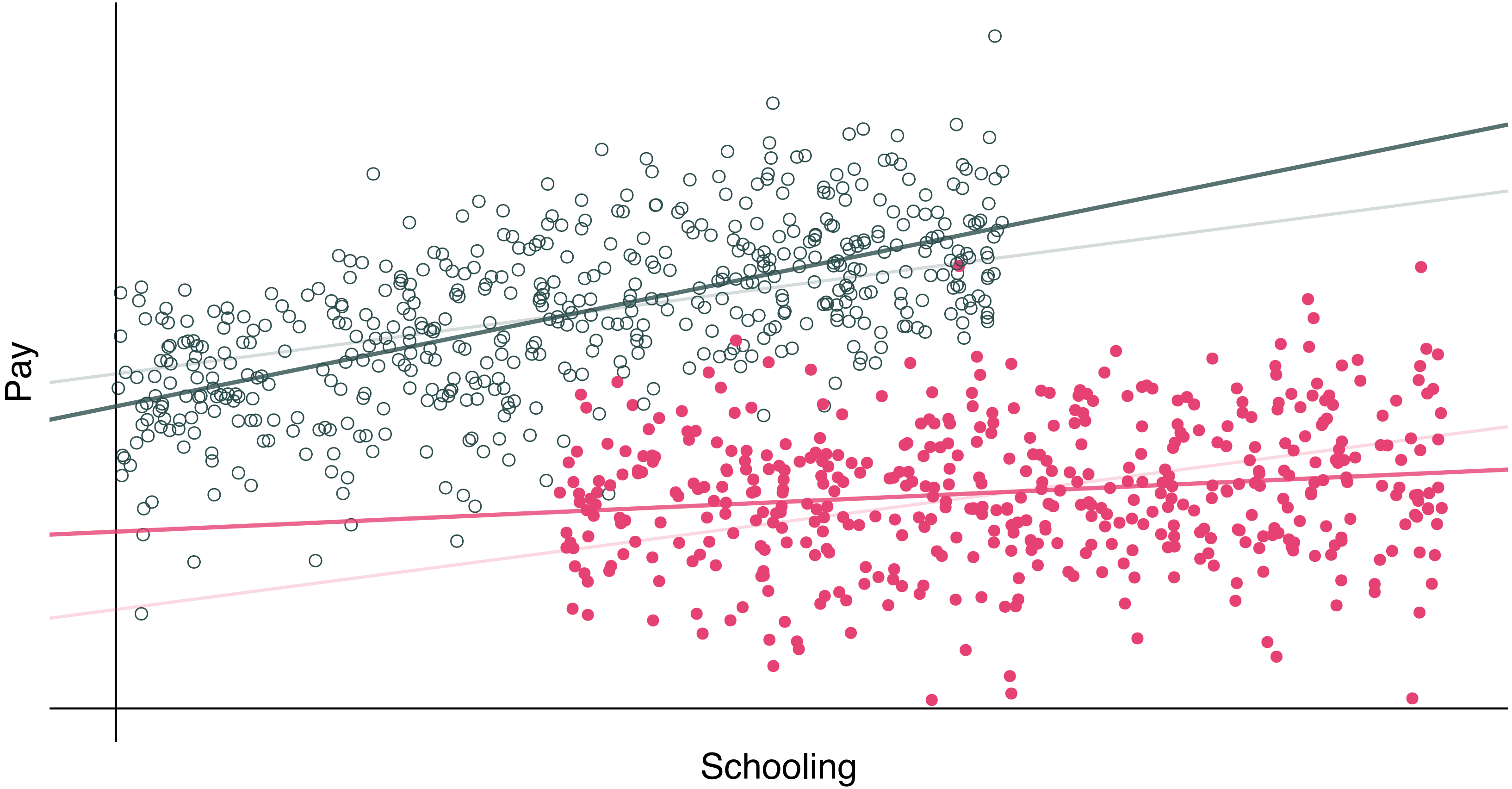
$$\text{Pay}_i = \beta_0 + \beta_1\,\text{School}_i + \beta_2\,\text{Male}_i + u_i$$



Interactions

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

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



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

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