Ordinary Least Squares, continued

EDS 222

Tamma Carleton
Fall 2023

Announcements/check-in

- Assignment #1: Grades posted
  - Please ensure your `.html` file is compiled and pushed to GitHub
  - Please do not push data to GitHub (generally a good rule to follow)
  - Sandy to go over some areas of confusion

Announcements/check-in

- Assignment #1: Grades posted
  - Please ensure your `.html` file is compiled and pushed to GitHub
  - Please do not push data to GitHub (generally a good rule to follow)
  - Sandy to go over some areas of confusion
- Assignment #2: Due 10/20, 5pm

Announcements/check-in

- Assignment #1: Grades posted
  - Please ensure your `.html` file is compiled and pushed to GitHub
  - Please do not push data to GitHub (generally a good rule to follow)
  - Sandy to go over some areas of confusion
- Assignment #2: Due 10/20, 5pm
- Reiteration of COVID/illness policy

Today

Notes on OLS

- Outliers, missing data

Today

Notes on OLS

- Outliers, missing data

Measures of model fit

- Coefficient of variation R^2

Today

Notes on OLS

- Outliers, missing data

Measures of model fit

- Coefficient of variation R^2

Categorical variables

- In R, interpretation

Today

Notes on OLS

- Outliers, missing data

Measures of model fit

- Coefficient of variation R^2

Categorical variables

- In R, interpretation

Multiple linear regression

- Adding independent variables, interpretation of results

Notes on OLS

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 99^{th} 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.

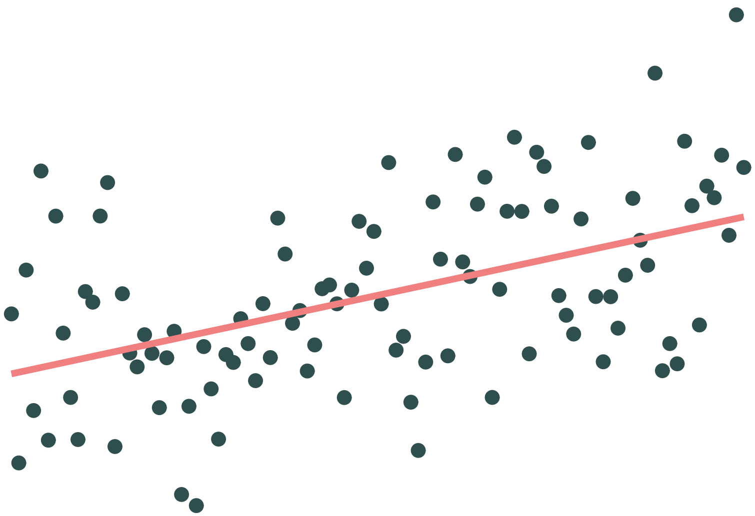
[†]: Or perform almost any operation/function

Measures of model fit

Measures of model fit

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

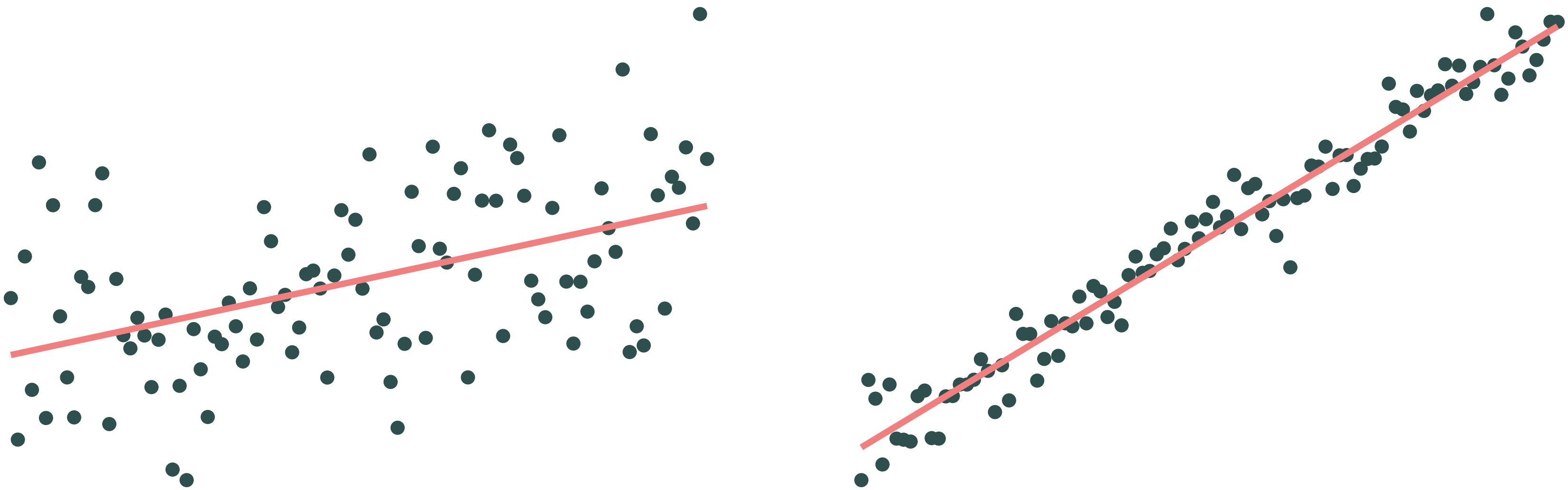
General idea: Larger variance in residuals suggests our model isn't very predictive



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

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 your regression model**

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 your regression model**

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

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

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

Coefficient of determination

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

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 $\bar{y}$.

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)

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 $\hat{y}$

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 can be appropriate (to discuss in machine learning)
  - Always look at your data and residuals!

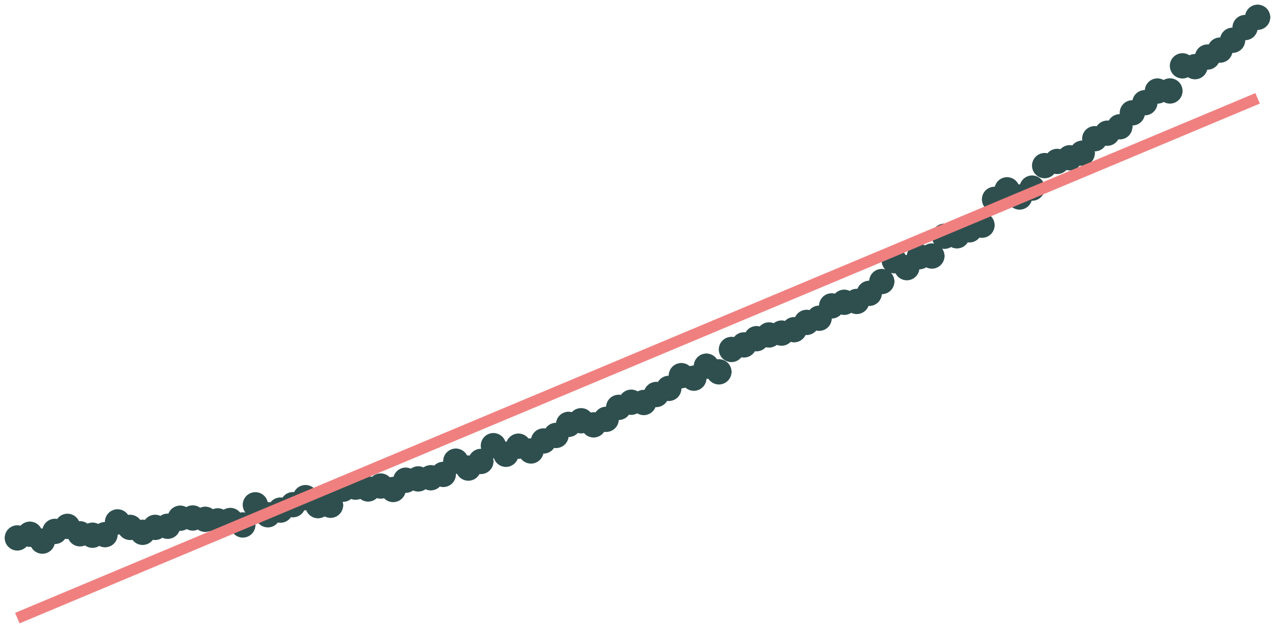
Coefficient of determination

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

- Describes a *linear* relationship between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$
- Higher $\boldsymbol{R}^2$ 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 can be appropriate (to discuss in machine learning)
  - Always look at your data and residuals!
- Like OLS in general, $\boldsymbol{R}^2$ is very sensitive to outliers. Again…always look at your data!

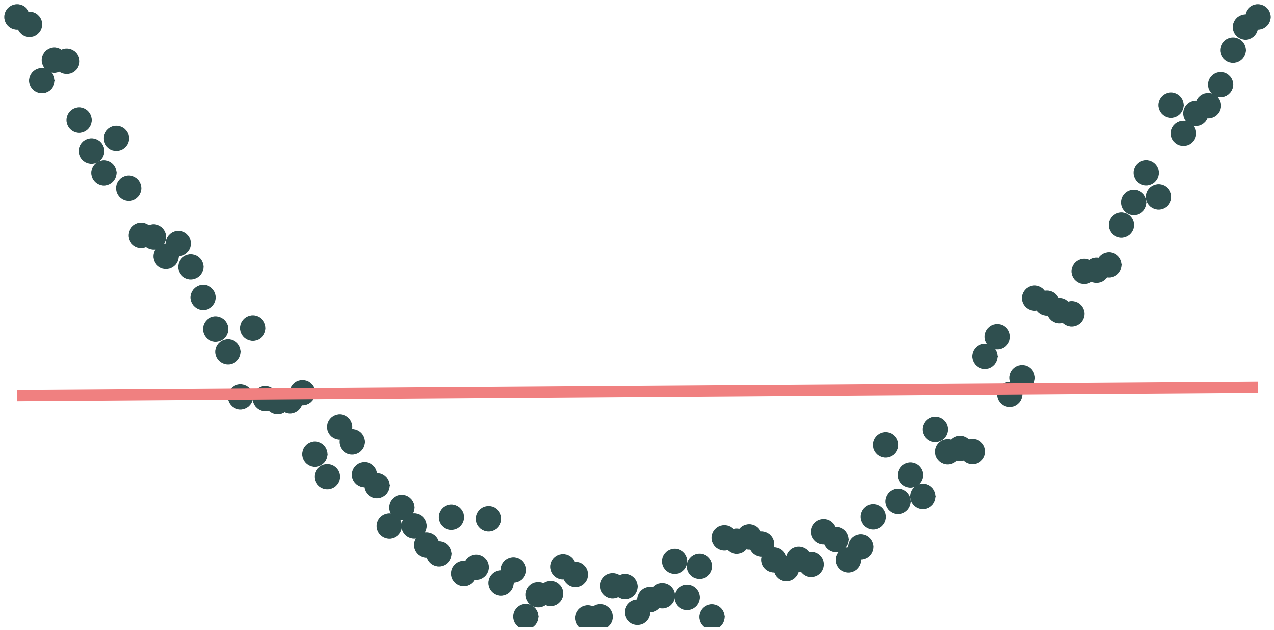
Coefficient of determination

Here, $R^2 = 0.94$ for a model of $y = \beta_0 + \beta_1 x + \epsilon$. Does that mean a linear relationship with x is appropriate?



Coefficient of determination

Here, $R^2 = 0$ for a model of $y = \beta_0 + \beta_1 x + \epsilon$. 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
- Crab size
- Temperature and precipitation amounts
- etc.

Categorical variables

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

- Ozone levels
- Crab size
- 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.

Categorical variables

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

- Ozone levels
- Crab size
- 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).

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 income We estimate:

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

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 income We estimate:

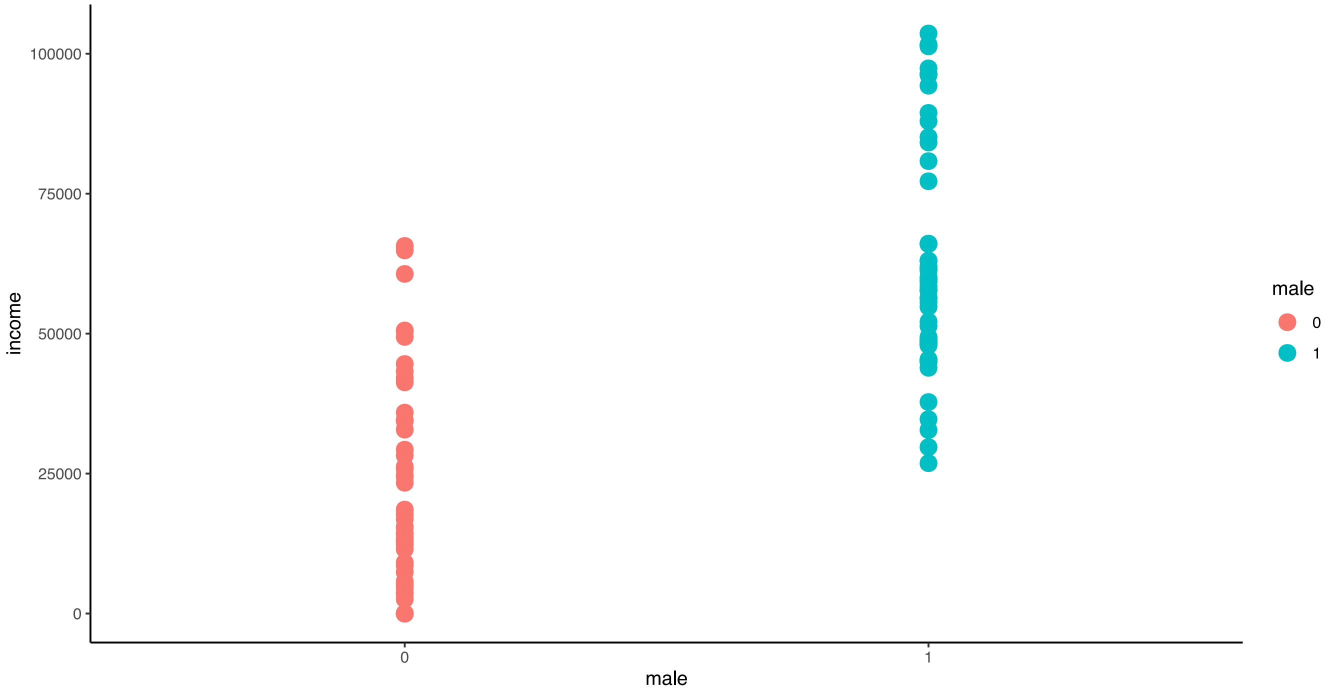
$$y_i = \beta_0 + \beta_1 MALE_i + \varepsilon_i$$

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", the OLS slope coefficient is 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 × 8
#>   species island    bill_length_mm bill_depth_mm flipper_l…¹ body_…² sex    year
#>   <fct>   <fct>              <dbl>         <dbl>       <int>   <int> <fct> <int>
#> 1 Adelie  Torgersen           39.1          18.7         181    3750 male   2007
#> 2 Adelie  Torgersen           39.5          17.4         186    3800 fema…  2007
#> 3 Adelie  Torgersen           40.3          18           195    3250 fema…  2007
#> 4 Adelie  Torgersen           NA            NA            NA      NA <NA>   2007
#> 5 Adelie  Torgersen           36.7          19.3         193    3450 fema…  2007
#> 6 Adelie  Torgersen           39.3          20.6         190    3650 male   2007
#> # … with abbreviated variable names ¹flipper_length_mm, ²body_mass_g
```

```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

What is going on here?? One x variable turned into multiple slope coefficients? 🤔

Categorical variables

What is going on here?? One x variable turned into multiple slope coefficients? 🤔

R is turning our regression

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

where *SPECIES* is a categorical variable indicating one of three species, into:

$$y_i = \beta_0 + \beta_1 CHINSTRAP_i + \beta_2 GENTOO_i + \varepsilon_i$$

where *CHINSTRAP* and *GENTOO* are dummy variables for the Chinstrap and Gentoo species, respectively.

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 ***
```

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)

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

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?

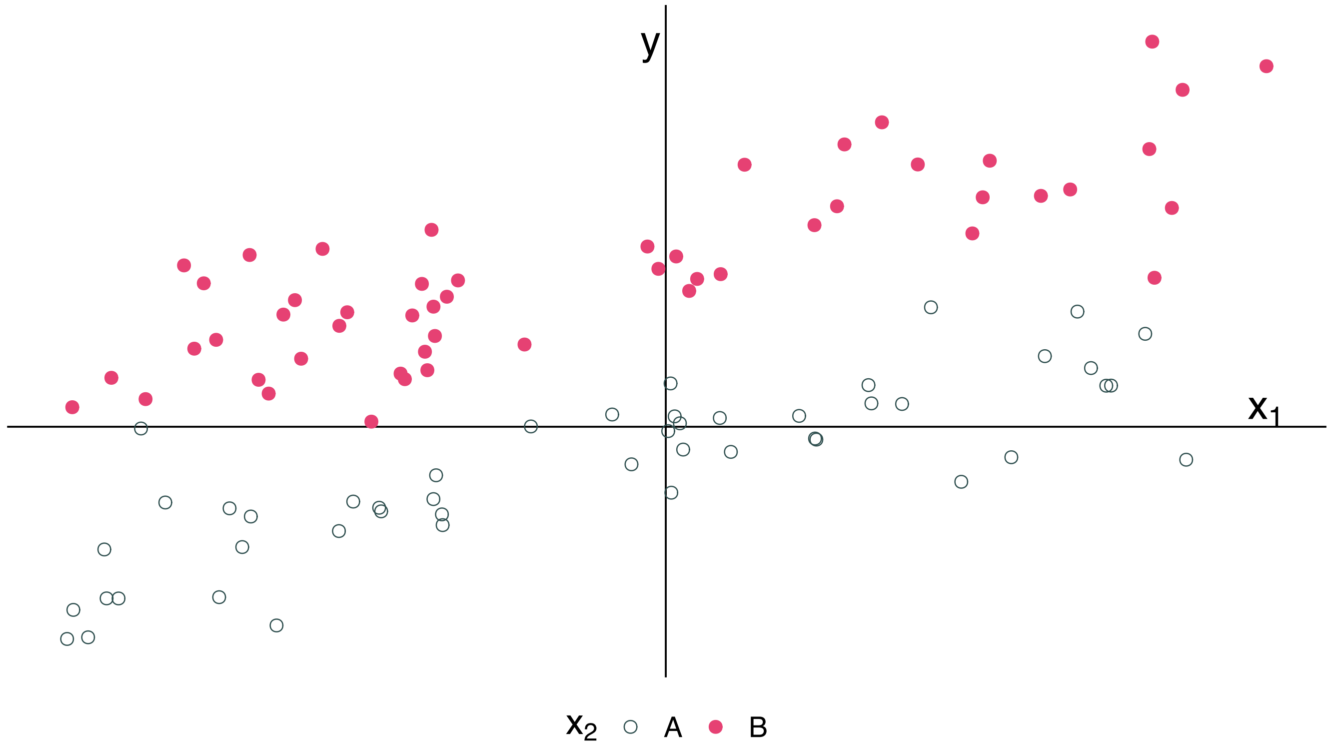
We will dig into each of these here, and you will see these questions in other MEDS courses

Multiple regression

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

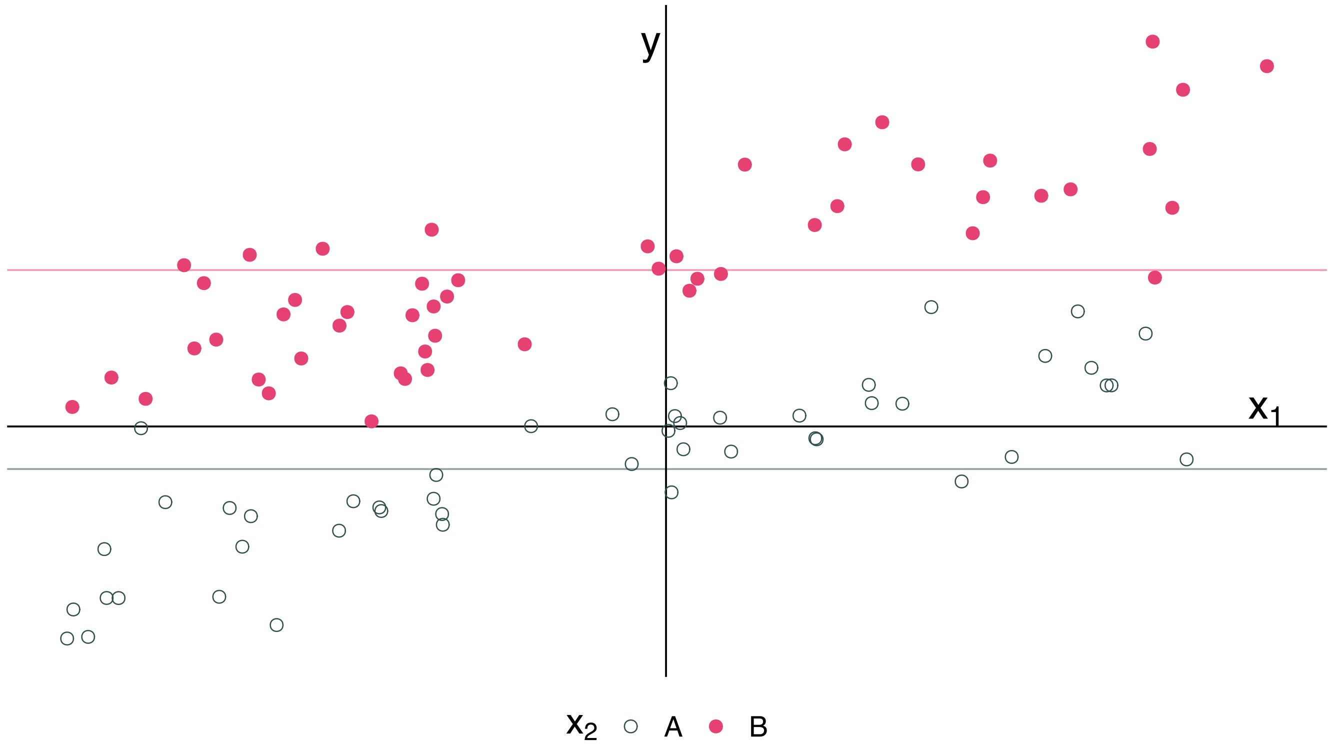
Multiple regression

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



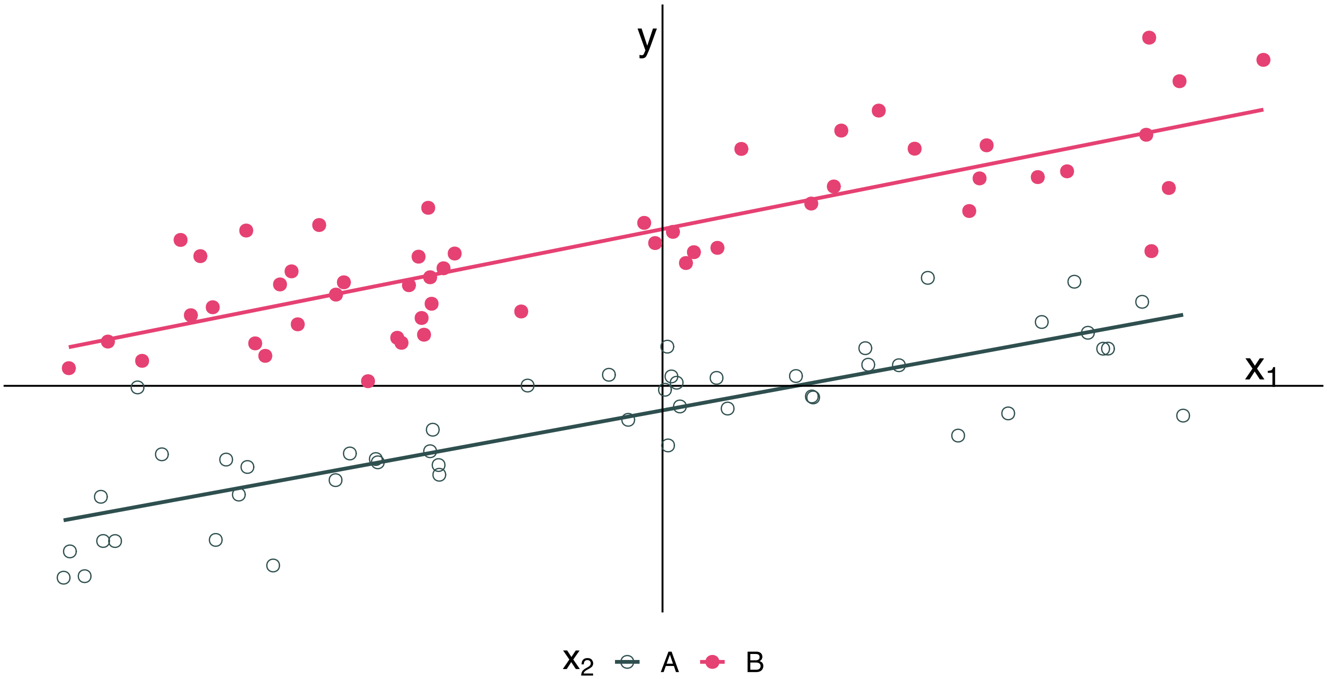
Multiple regression

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



Multiple regression

$\hat{\beta}_1$ estimates the relationship between y and x_1 after controlling for x_2. This is often called the "parallel slopes" model (one slope β_1 for each of the groups in x_2)



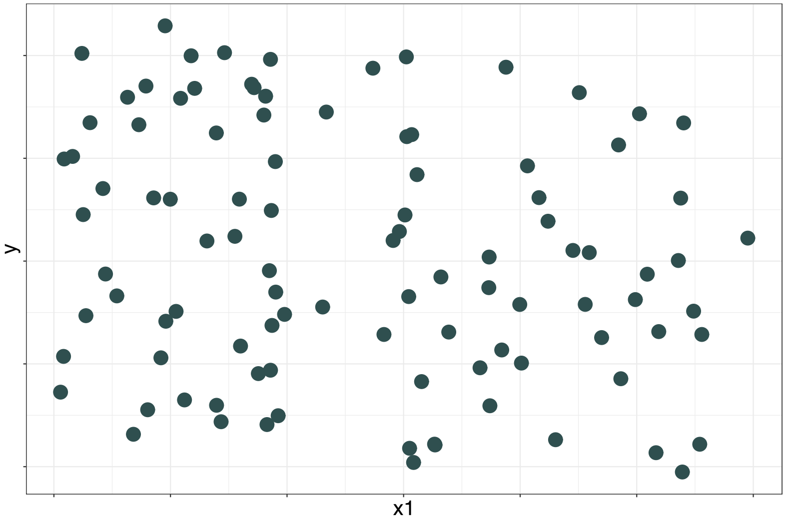
Multiple regression

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

Multiple regression

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

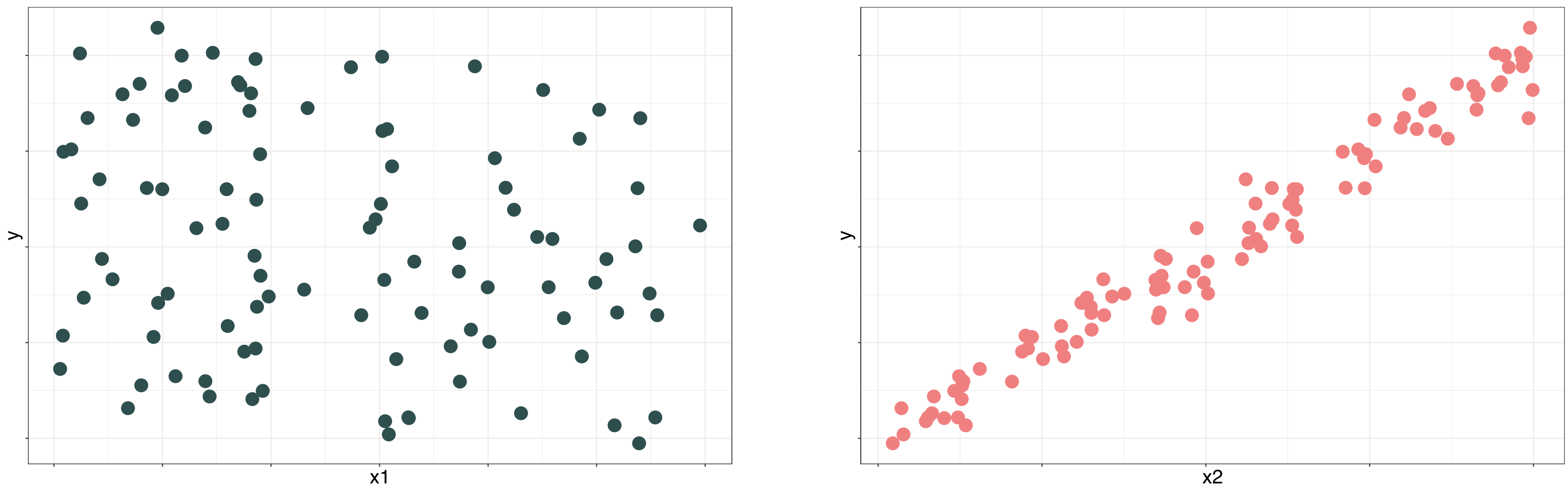
Suppose $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$



Multiple regression

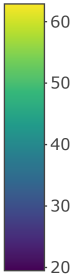
More generally, how do we think about multiple explanatory variables?

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

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

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

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.

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 cancer rates in census tract i:

$$\text{Cancer rate}_i = \beta_0 + \beta_1 \text{UV radiation}_i + \beta_2 \text{TRI}_i + u_i$$

where

- UV radiation_i gives the average UV radiation in tract i (mW/cm^2)
- TRI_i denotes an indicator variable for whether tract i has a Toxics Release Inventory facility

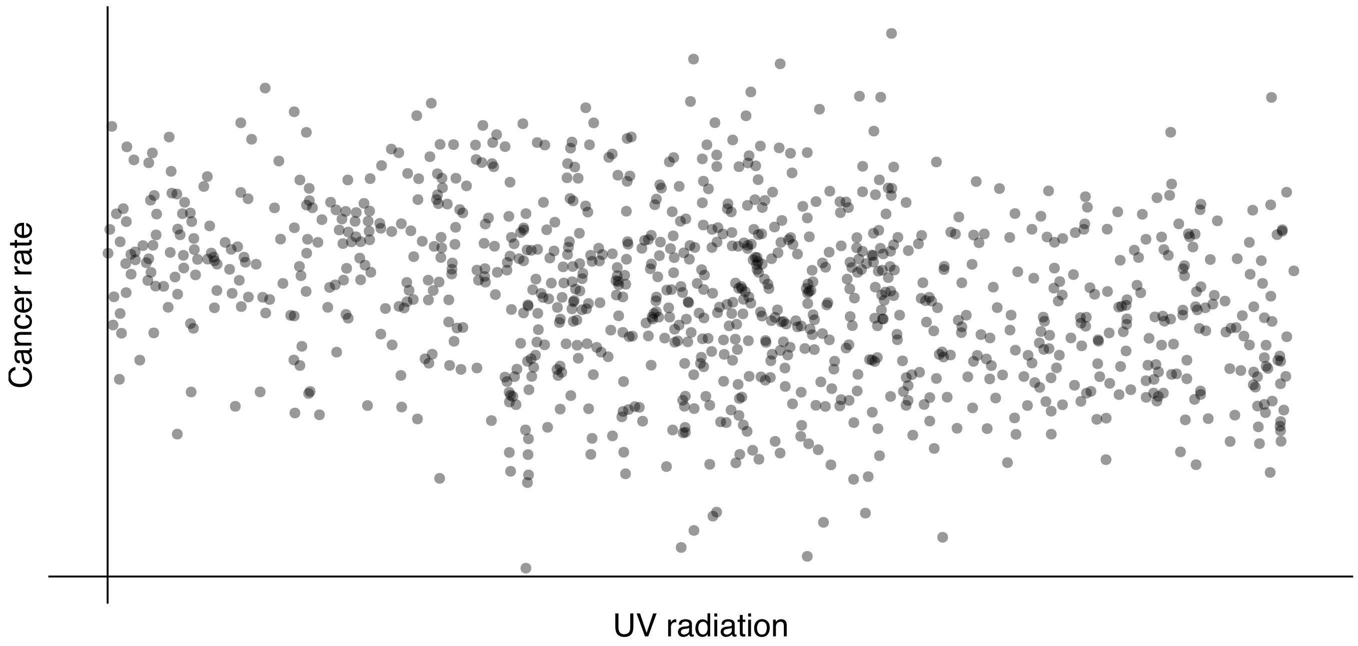
thus

- β_1: the change in cancer rate associated with a 1 mW/cm^2 increase in UV radiation (*ceteris paribus*)
- β_2: the difference in avg. cancer rates between TRI and non-TRI census tracts (*ceteris paribus*)
  If $\beta_2 > 0$, then TRI tracts have higher cancer rates

Omitted-variable bias

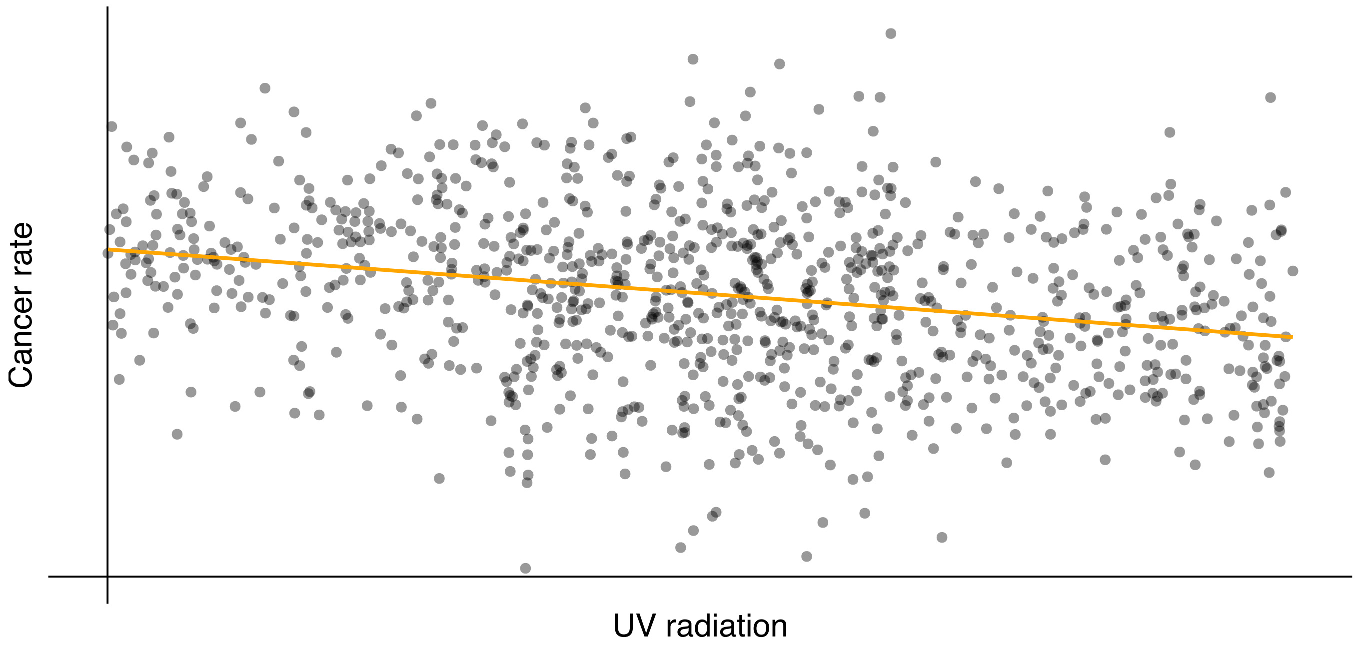
"True" relationship: $\text{Cancer rate}_i = 20 + 0.5 \times \text{UV radiation}_i + 10 \times \text{TRI}_i + u_i$

The relationship between cancer rates and UV radiations:



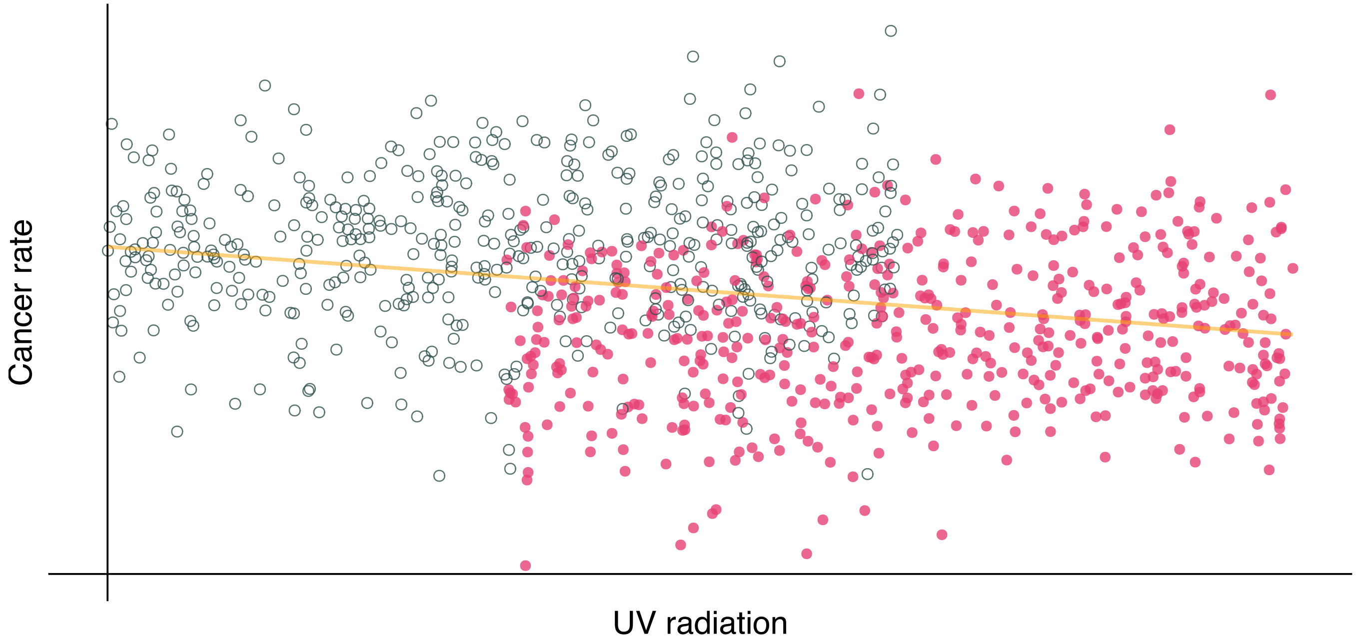
Omitted-variable bias

Biased regression estimate: $\widehat{\textbf{Cancer rate}}_i = \mathbf{31.3} + \mathbf{-0.9} \times \textbf{UV radiation}_i$



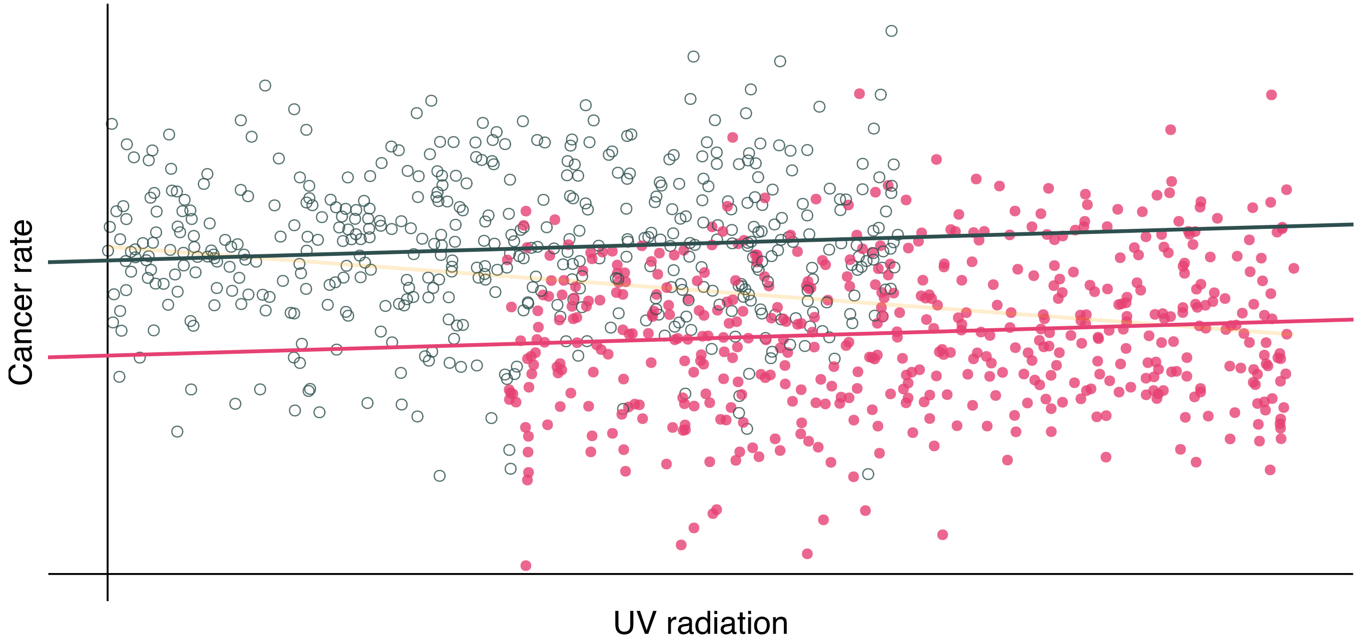
Omitted-variable bias

Recalling the omitted variable: TRI (**non-TRI** and **TRI**)



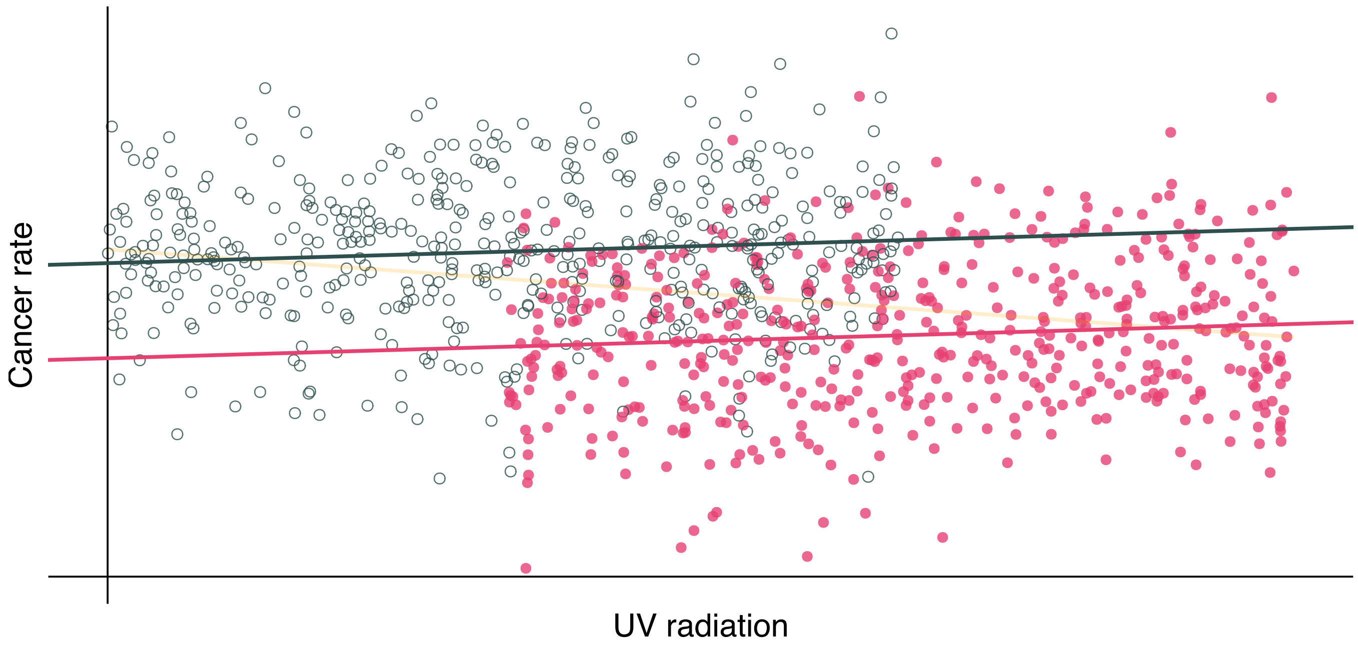
Omitted-variable bias

Recalling the omitted variable: TRI (**non-TRI** and **TRI**)



Omitted-variable bias

Unbiased regression estimate: $\widehat{\text{Cancer rate}}_i = 20.9 + 0.4 \times \text{UV radiation}_i + 9.1 \times \text{TRI}_i$



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

Some slide components come from Ed Rubin's awesome course materials.