Ordinary Least Squares

EDS 222

Tamma Carleton
Fall 2022

Today

Relationships between variables

- Covariance, correlation

Ordinary Least Squares

- Finding the "best fit" line, properties of OLS, assumptions of OLS

Interpreting OLS output

- Slopes, intercepts, unit conversions

Measures of model fit

- Coefficient of variation

Notes on OLS

- Missing data, outliers

Announcements/check-in

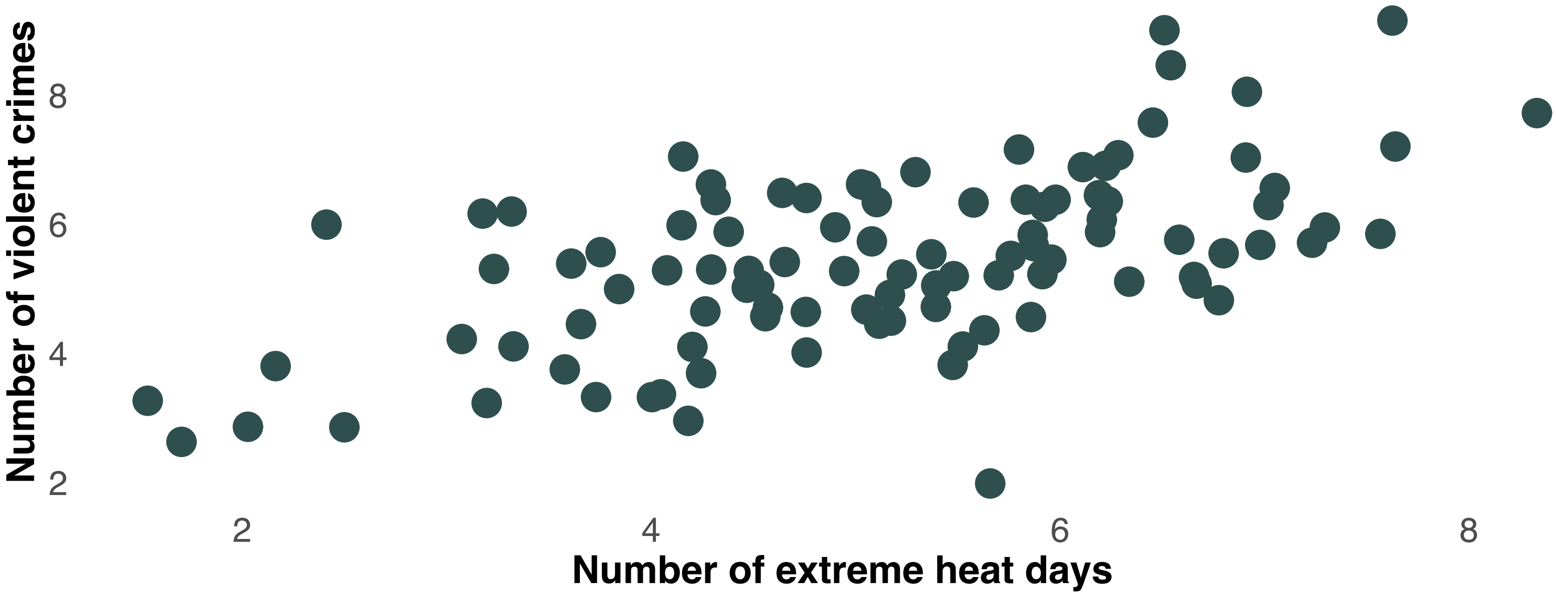
- Assignment #1: Grading and answers early next week
- Assignment #2: To be posted this week, due 10/13, 9am
- Flag on IMS and linear regression

Relationships between variables

Two random variables

Often we are interested in the *relationship* between two (or more) random variables.

E.g., heat waves and heart attacks, nitrogen fertilizer and water pollution



Note: these are simulated data. But the violence-temperature link is real!
See here for a summary of research.

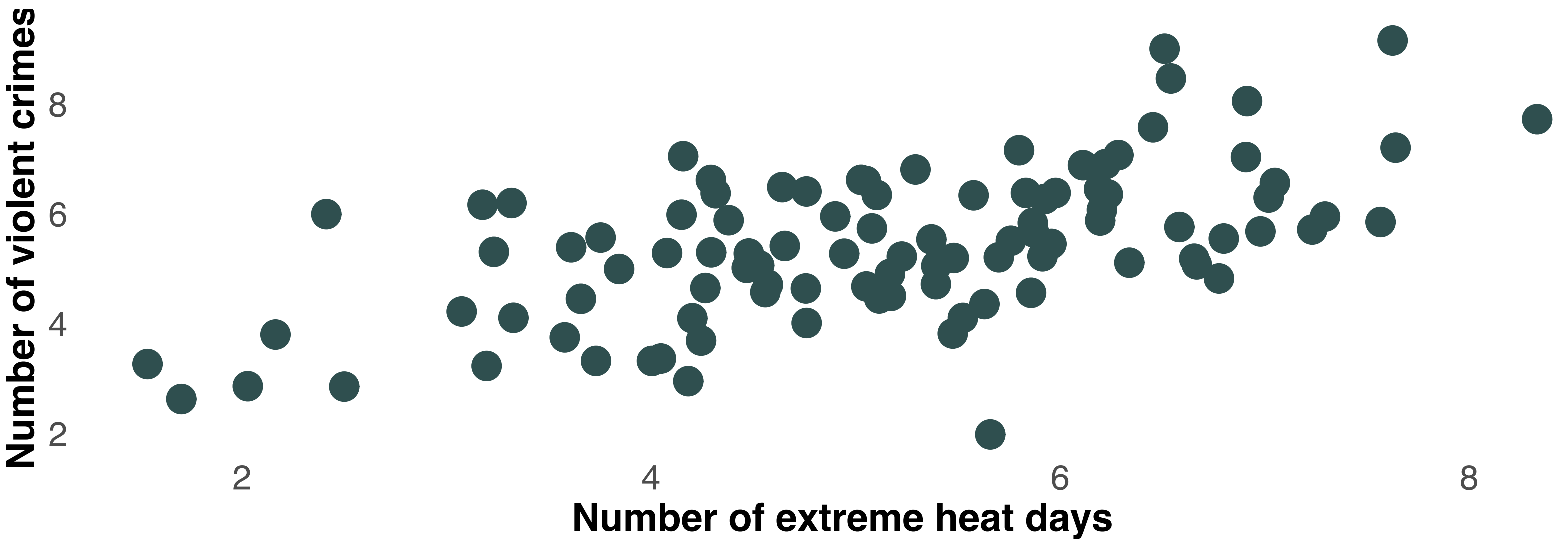
Two random variables

What metrics can we use to characterize the *relationship* between two variables?

There are lots. But let's start with...

1. Covariance

2. Correlation



Covariance

Variance indicates how dispersed a distribution is (average squared deviation from the mean)

Covariance is a measure of the *joint* distribution of two variables

- Higher values of X correspond to higher values of $Y \to$ **positive** covariance
- Higher values of X correspond to lower values of $Y \to$ **negative** covariance

In the population:

$$Cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)] = E[XY] - \mu_x\mu_y$$

In the sample:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Covariance

Variance indicates how dispersed a distribution is (average squared deviation from the mean)

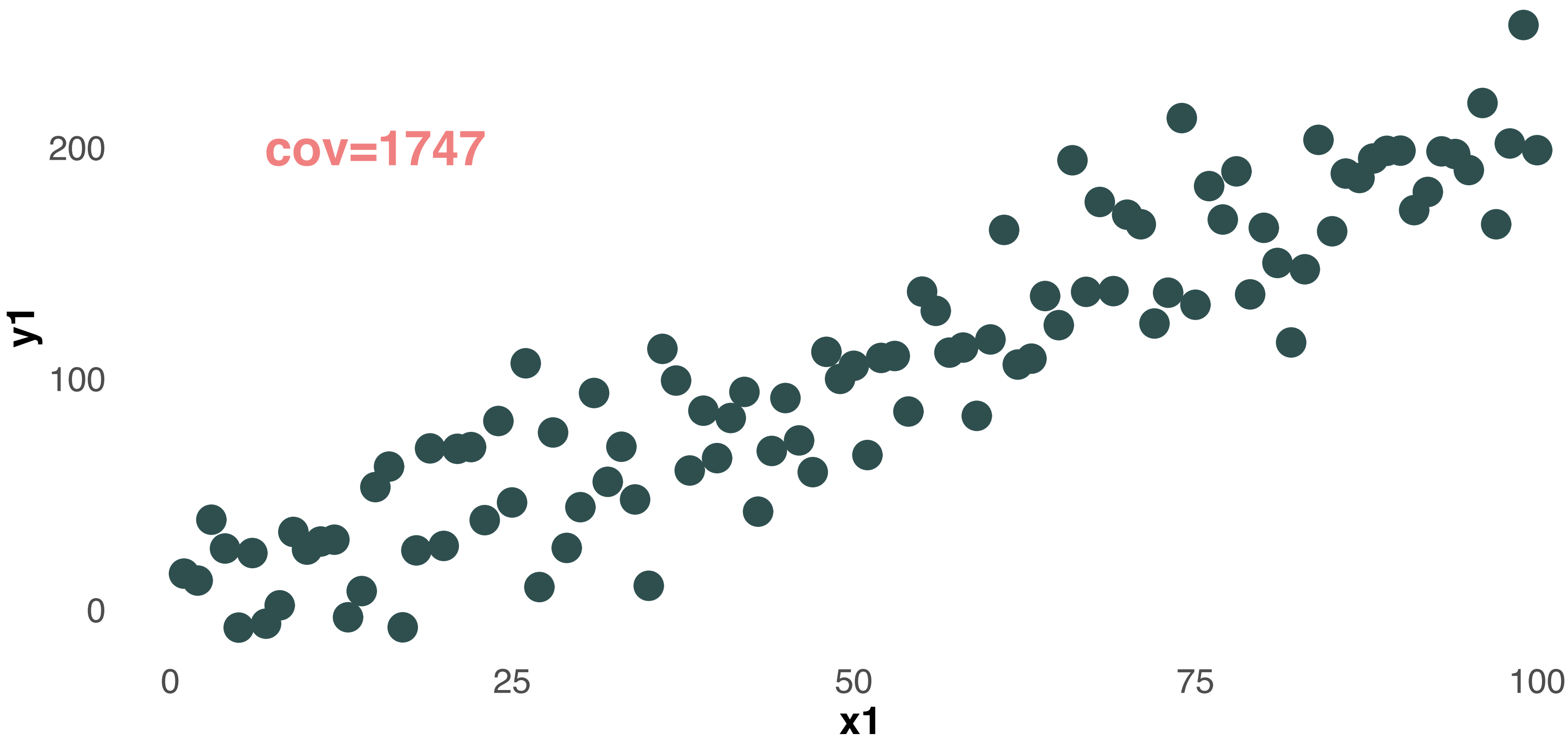
Covariance is a measure of the *joint* distribution of two variables

- Higher values of X correspond to higher values of Y $\rightarrow$ **positive** covariance
- Higher values of X correspond to lower values of Y $\rightarrow$ **negative** covariance

The **sign** of s_{xy} tells us the sign of the linear relationship between X and Y, but the **magnitude** depends on the units of the variables and is therefore difficult to interpret

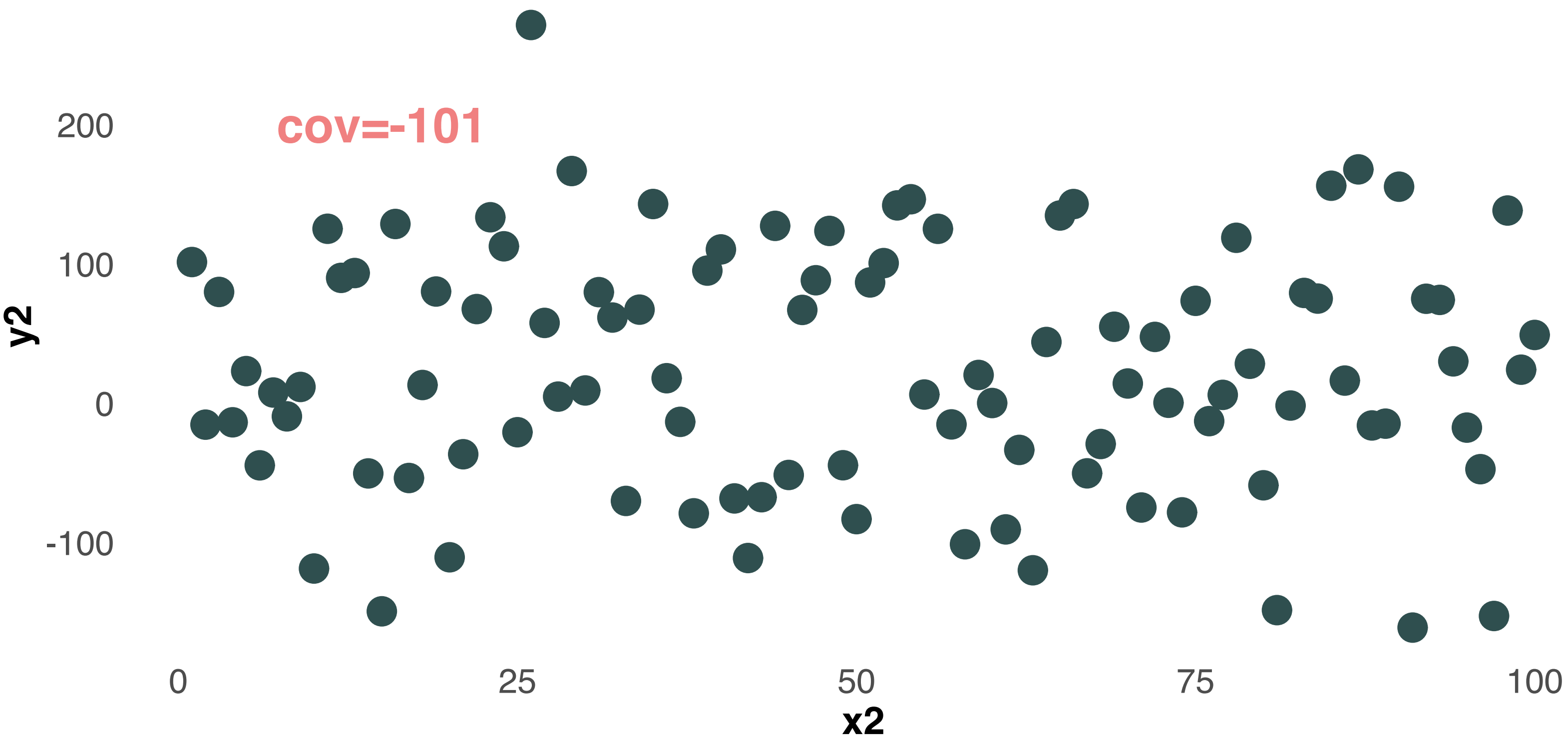
Covariance

Example: positive covariance



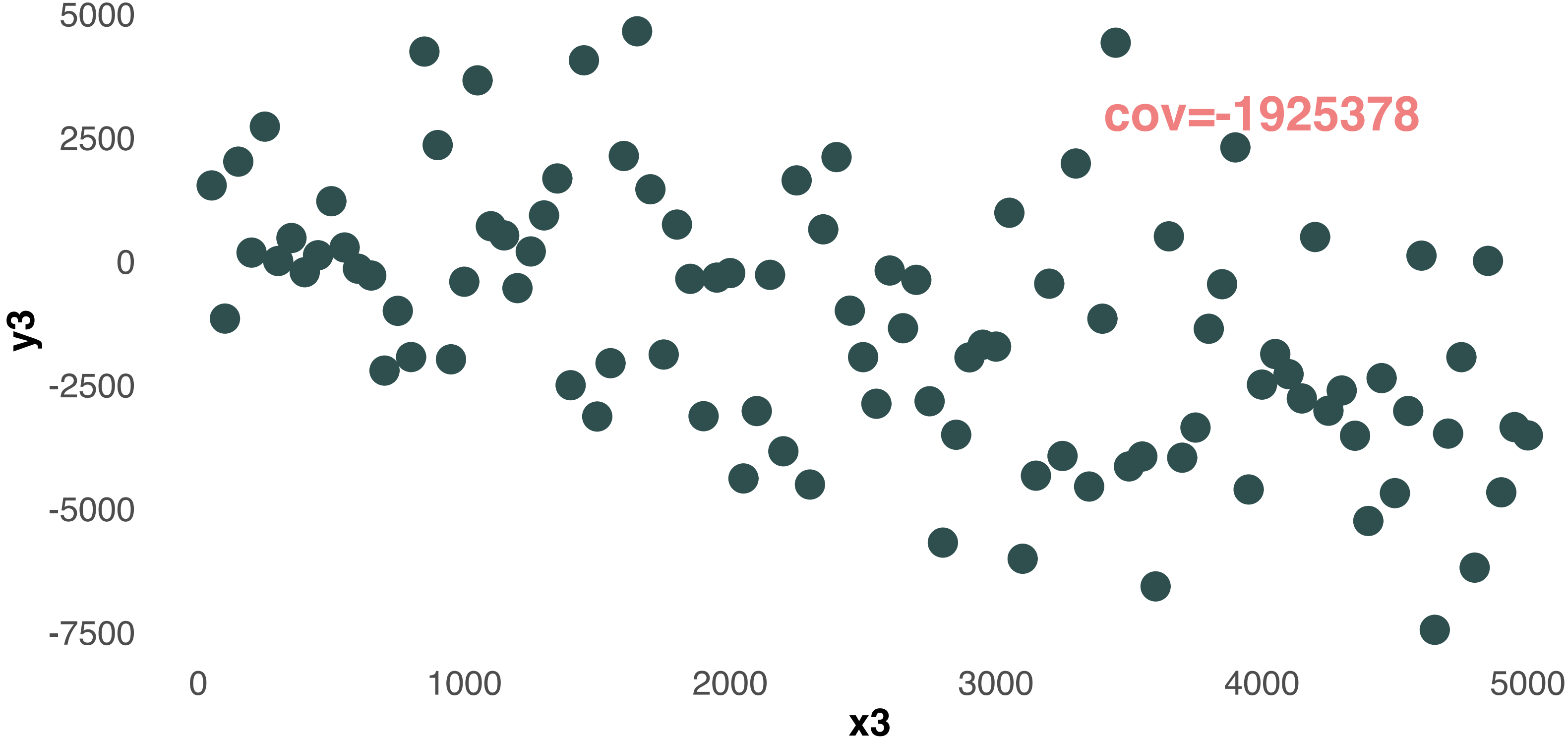
Covariance

Example: zero covariance



Covariance

Example: Negative covariance



How do I interpret these units?! Hard to compare across these three graphs...

Correlation

Correlation allows us to normalize covariance into interpretable units

The sign still tells us about the nature of the (linear) relationship between two variables:

- **positive** covariance $\rightarrow$ **positive** correlation (and vice versa)

But now, the magnitude is interpretable:

- Ranges from -1 to 1, with magnitude indicating *strength* of the relationship

Correlation

Correlation allows us to normalize covariance into interpretable units

In the population:

$$\rho_{X,Y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$

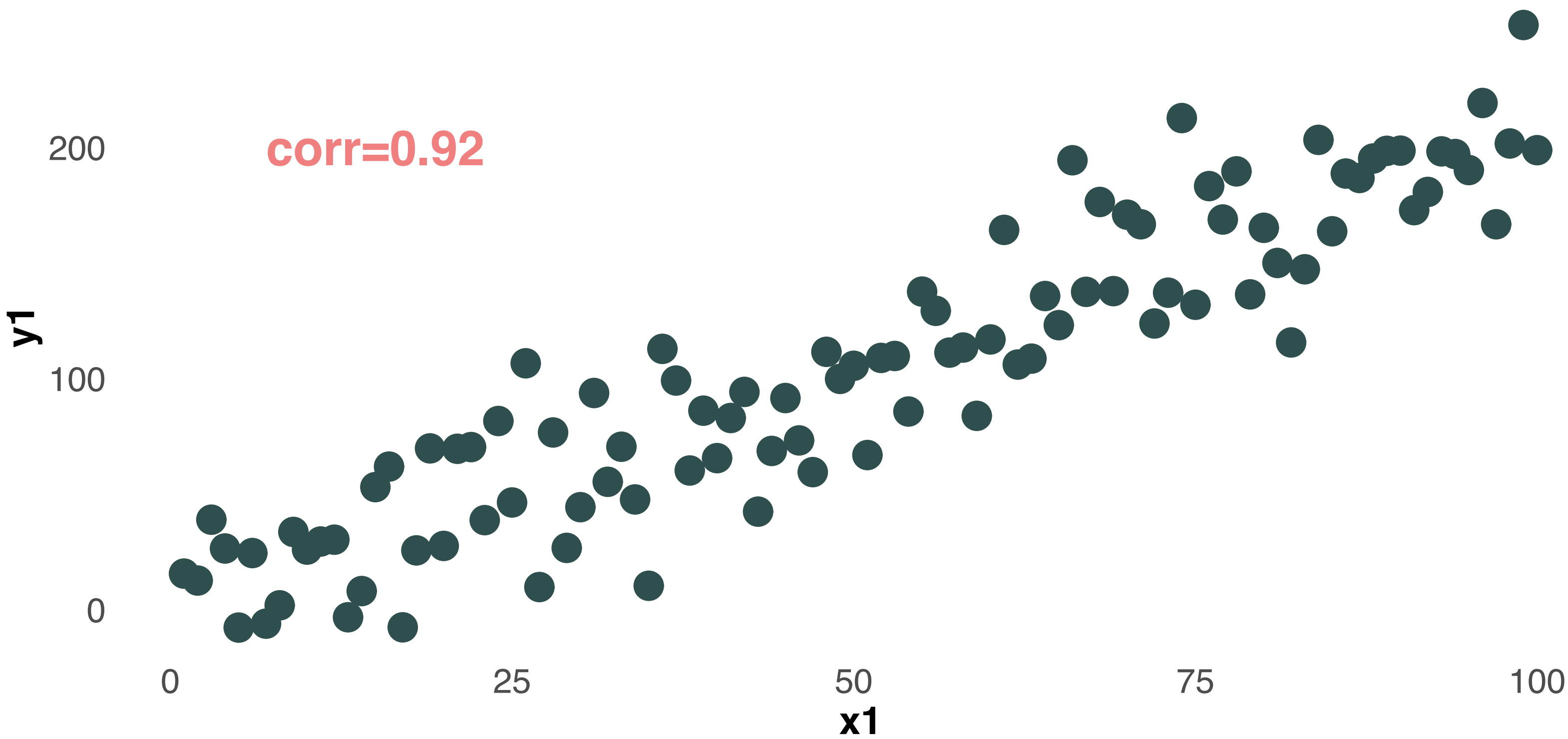
In the sample:

$$r_{x,y} = \frac{s_{x,y}}{s_x s_y} = \frac{1}{(n-1)s_x s_y} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Want to proove that $-1 \leq r_{x,y} \leq 1$? Key result: Cauchy-Schwarz Inequality tells us that $|cov(X,Y)|^2 \leq var(X)var(Y)$.

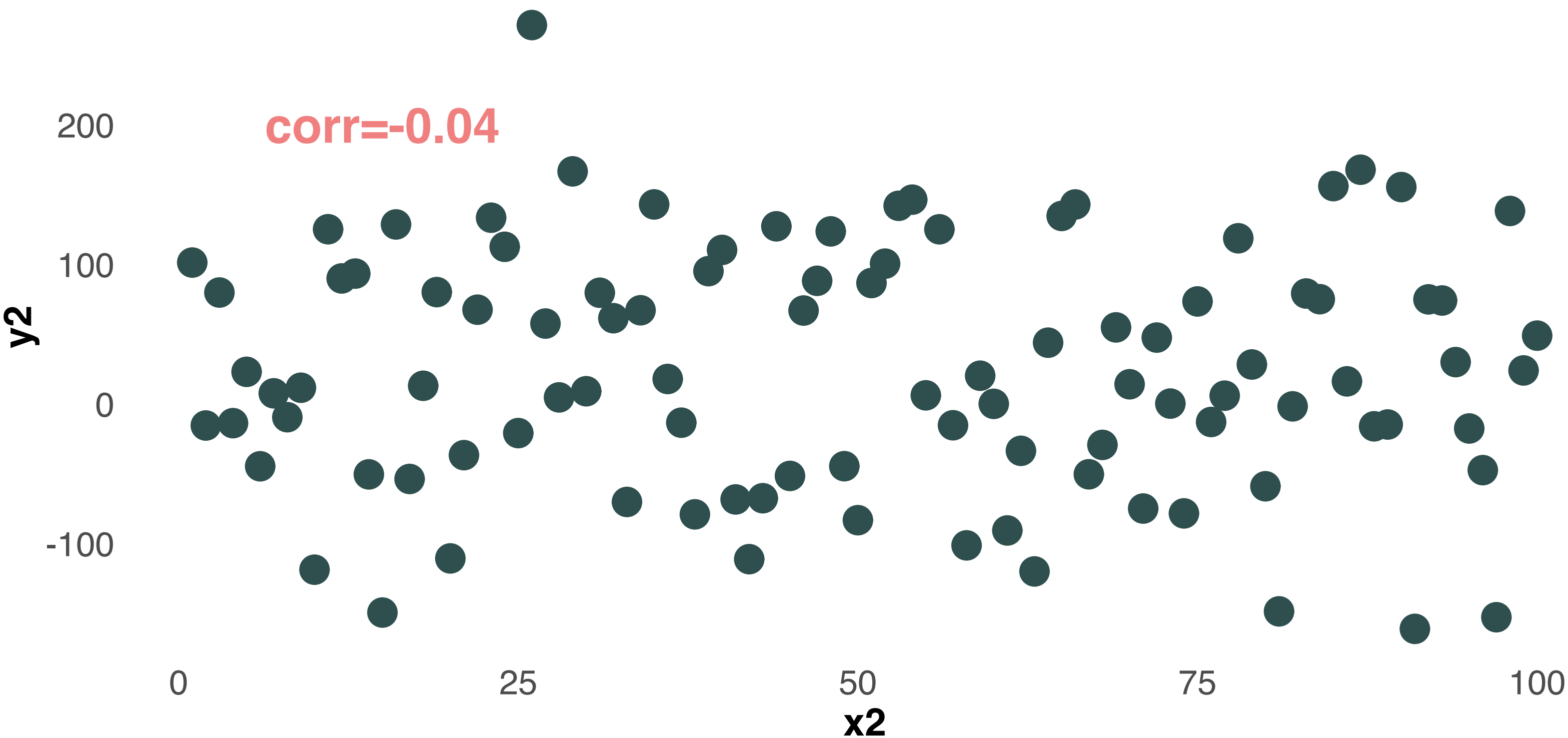
Correlation

Example: Strong positive correlation



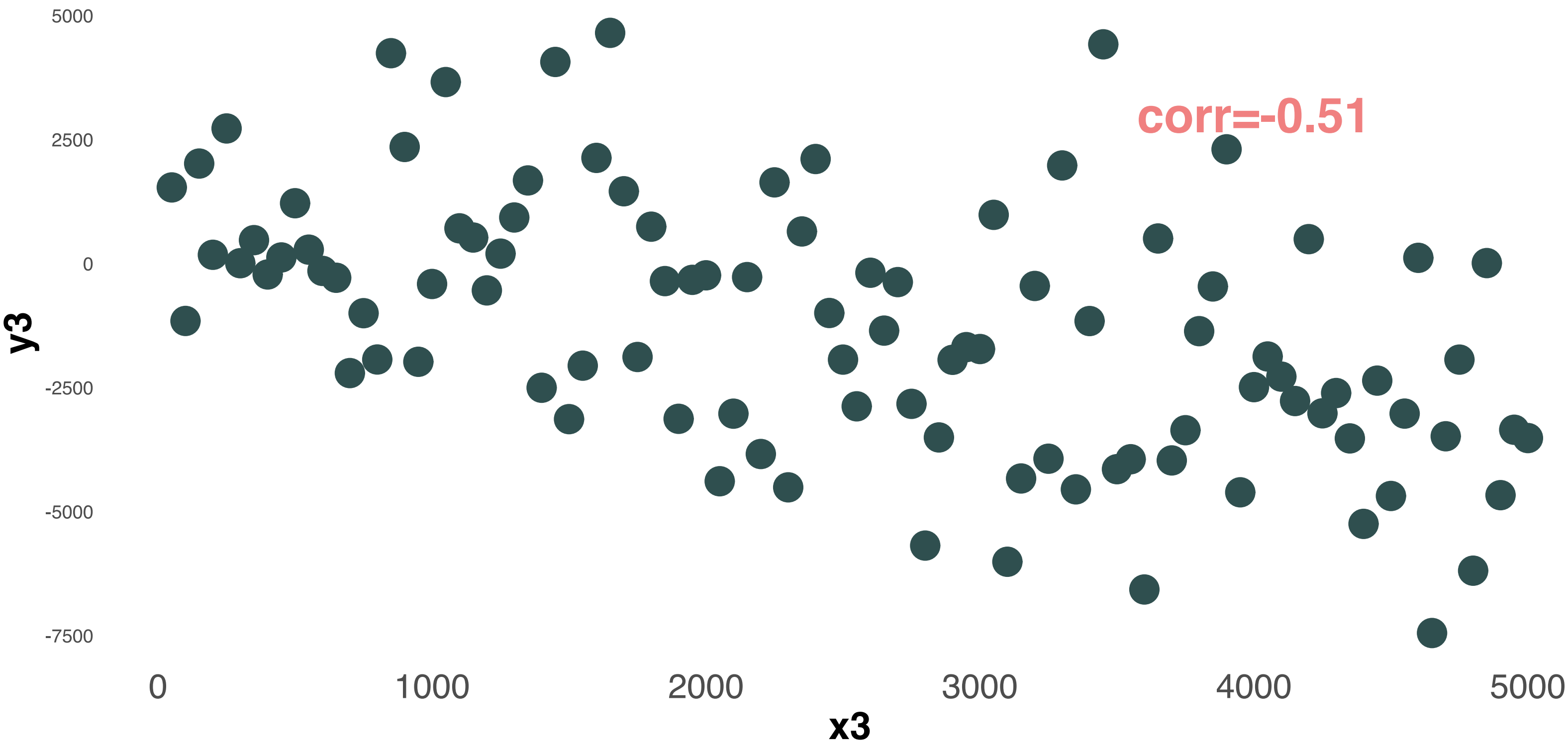
Correlation

Example: zero correlation



Correlation

Example: Moderate negative correlation



Ordinary Least Squares

Linear regression

Covariance and correlation give us a single summary of the **strength** of the relationship between two random variables Y and X...

...but we want to know more!

In particular, we are often interested in the **linear** relationship between X and Y:

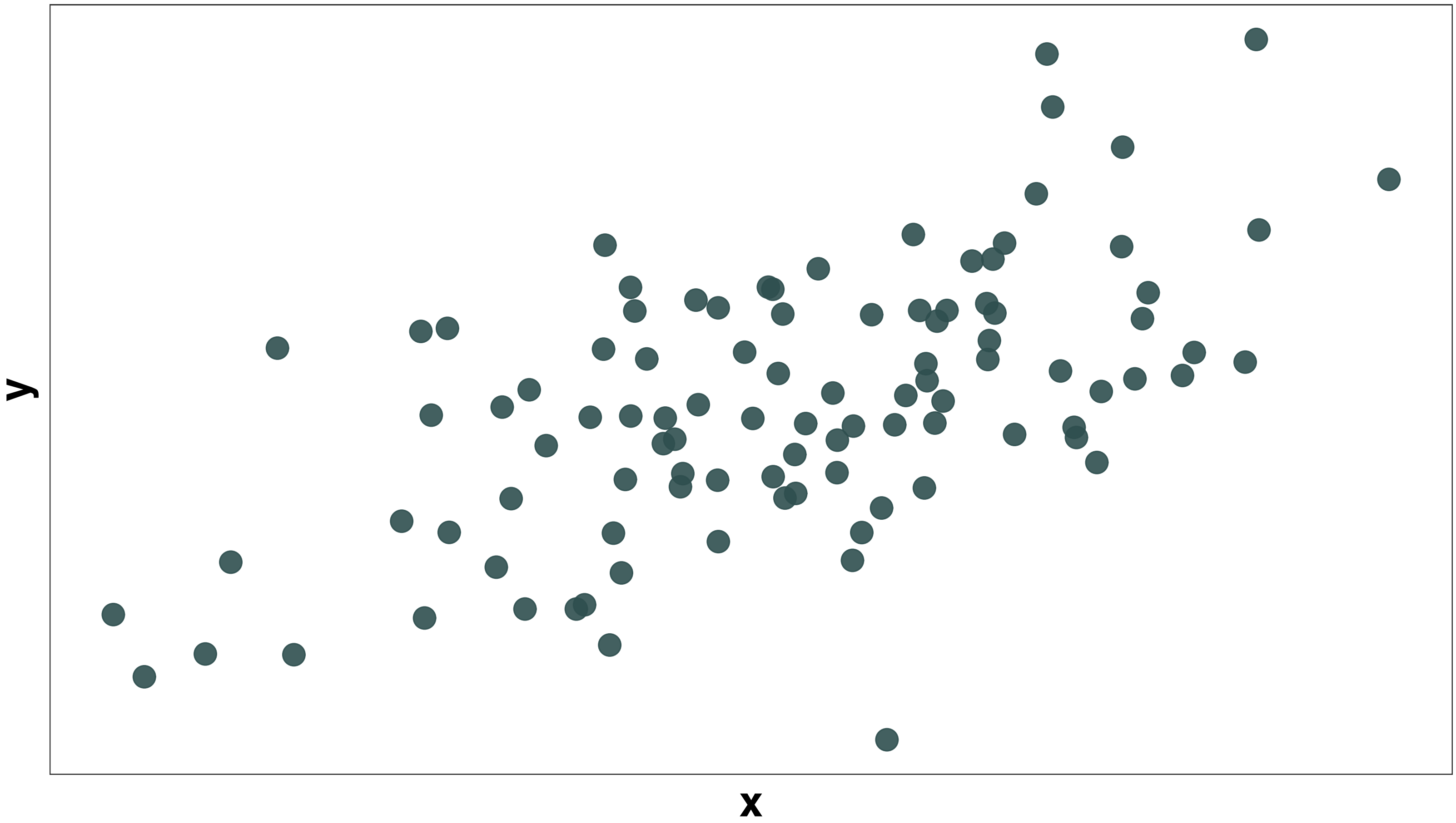
In the **population**:

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

Can we use our sample to estimate β_0 (the intercept) and β_1 (the slope)?

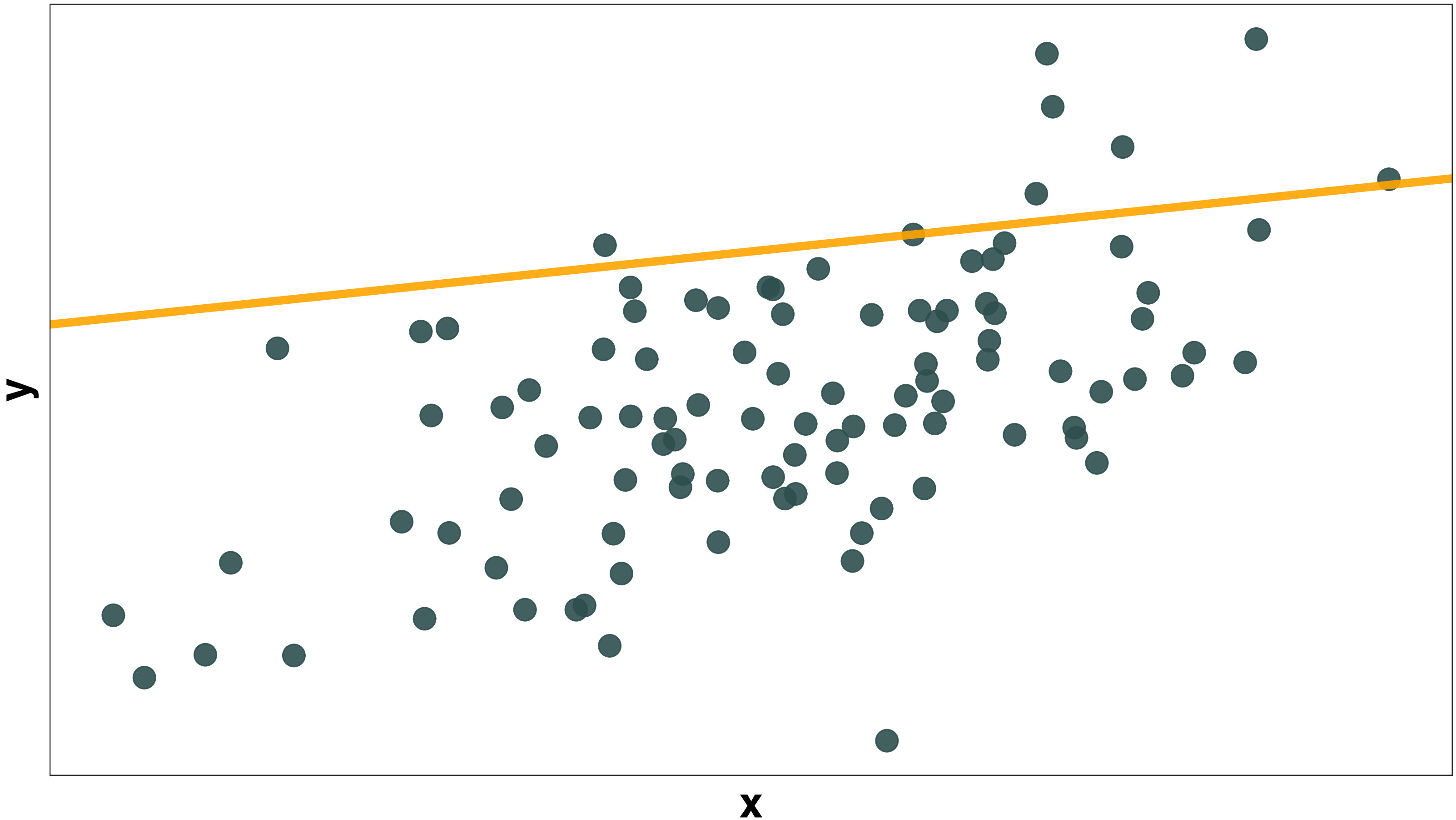
Finding a "best fit" line

Consider some sample data.



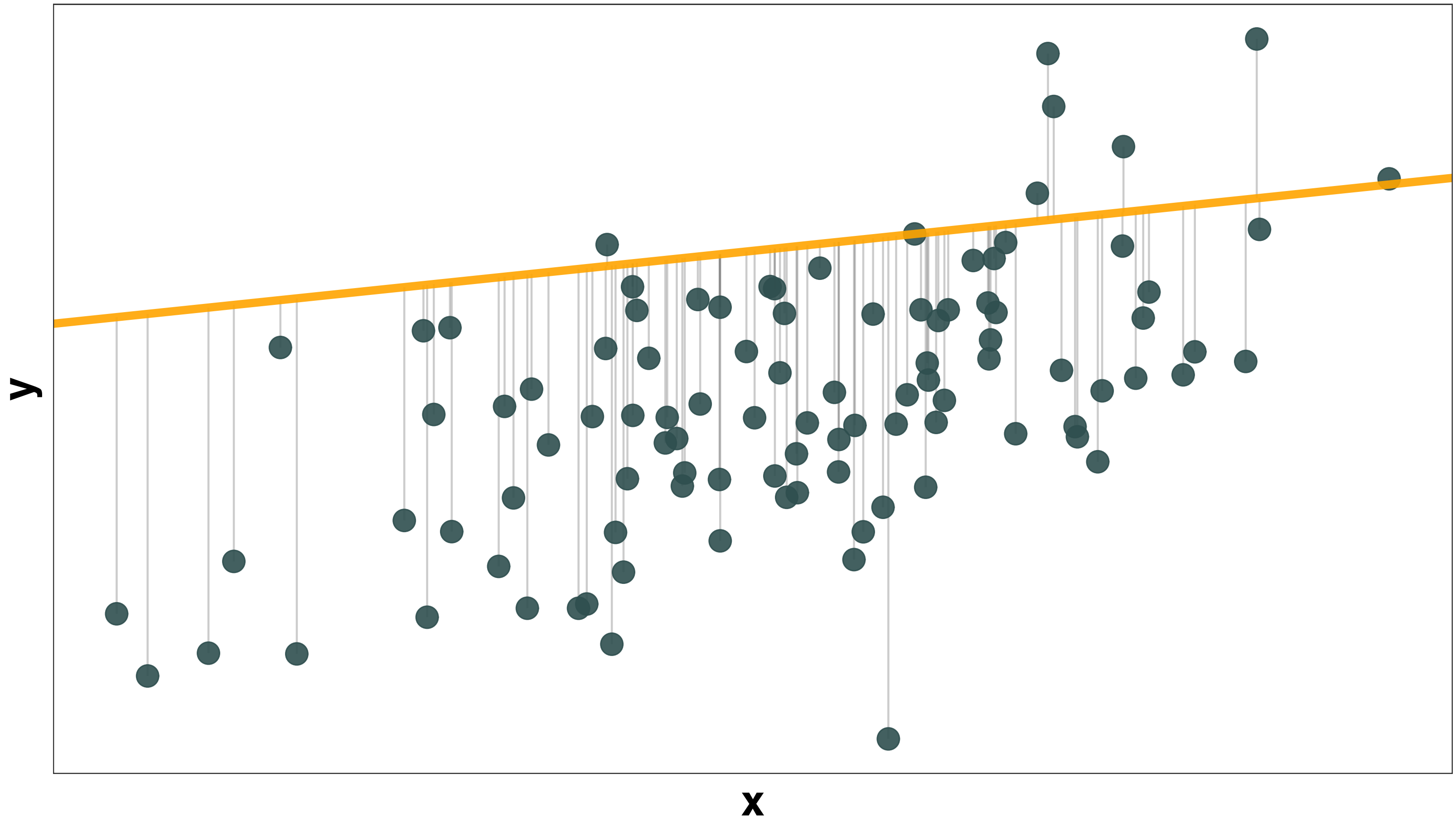
Finding a "best fit" line

For any line $\left(\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x\right)$



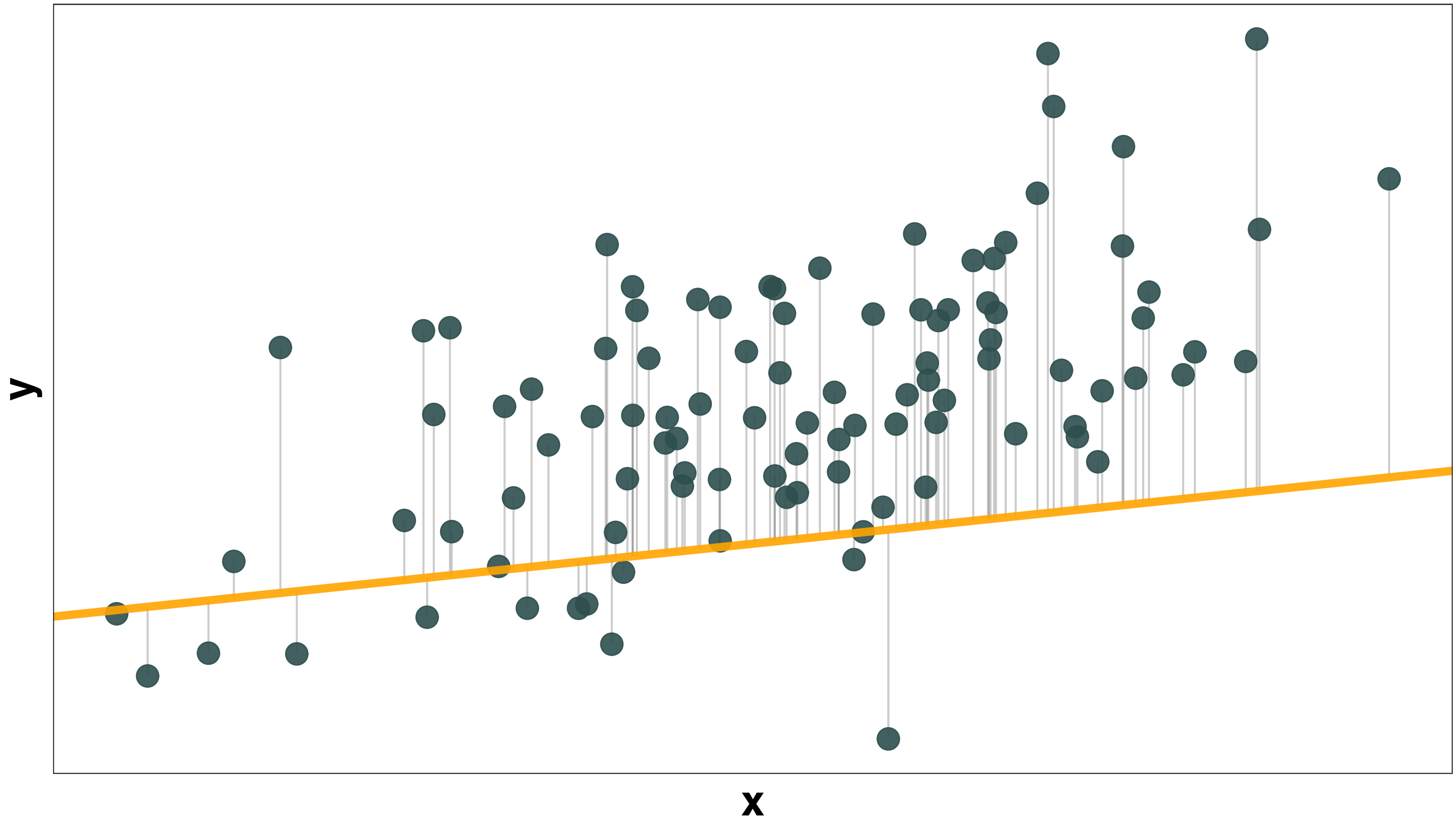
Finding a "best fit" line

For any line $\left(\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x\right)$, we can calculate errors: $e_i = y_i - \hat{y}_i$



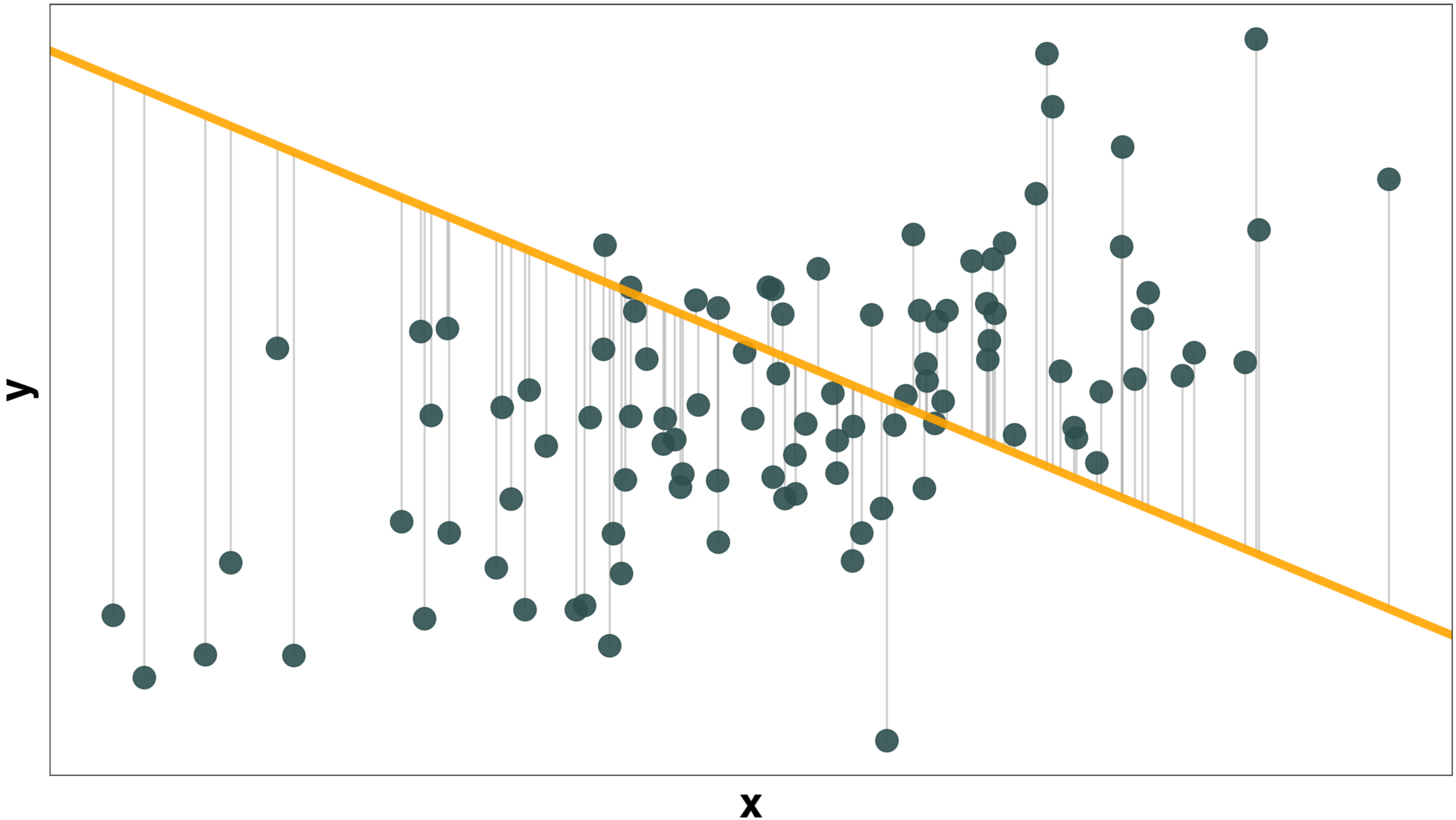
Finding a "best fit" line

For any line $\left(\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x\right)$, we can calculate errors: $e_i = y_i - \hat{y}_i$



Finding a "best fit" line

For any line $\left(\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x\right)$, we can calculate errors: $e_i = y_i - \hat{y}_i$



Ordinary Least Squares

OLS chooses a line that minimizes the **sum of squared errors**:

$$SSE = \sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x)^2$$

In other words, OLS gives us a combination of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes the SSE.

Now you see where "least squares" comes from!

In R:

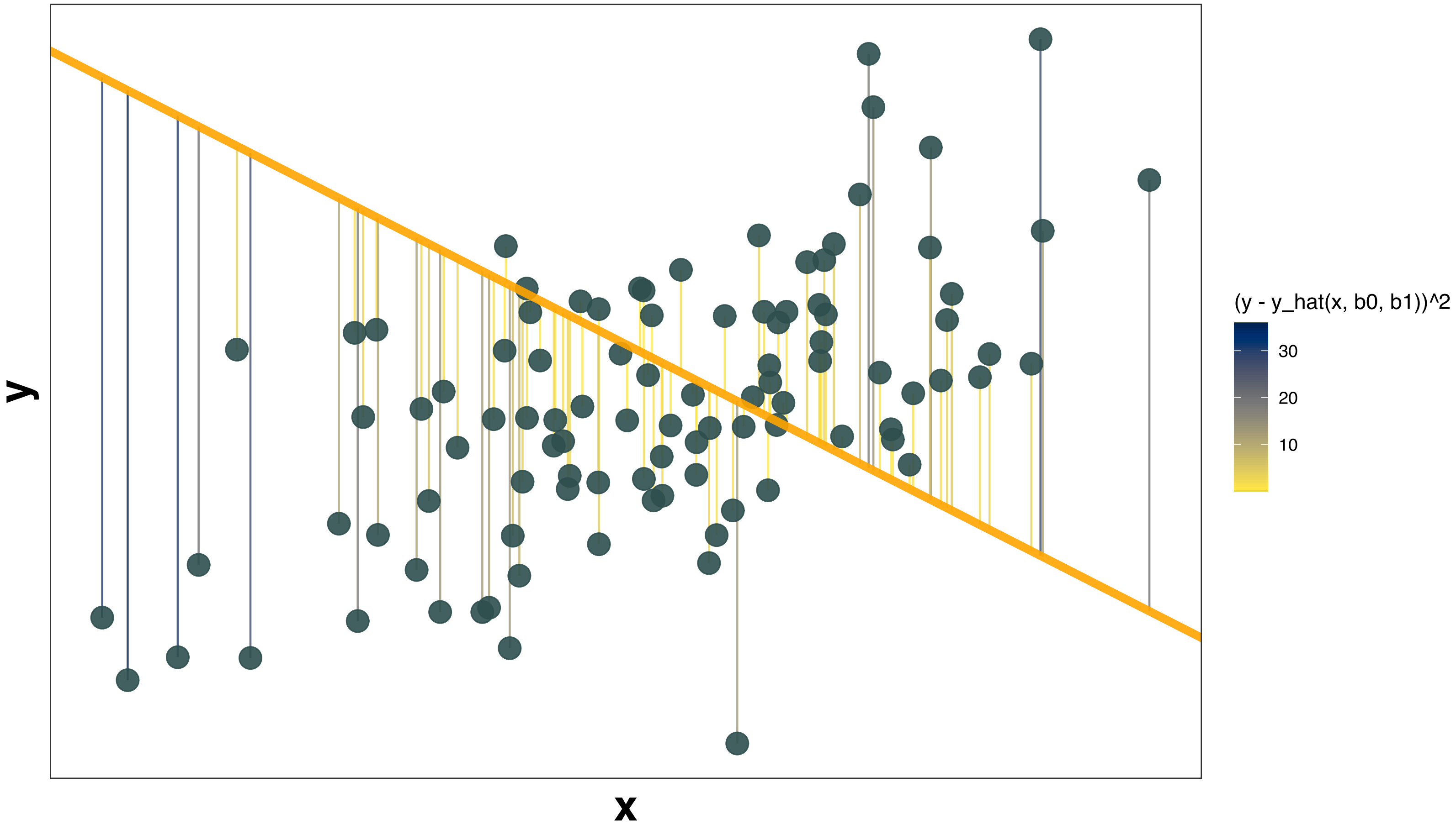
```
library(stats)
```

```
lm(y ~ x, my_data)
```

Note: SSE is also called "sum of squared residuals" or SSR

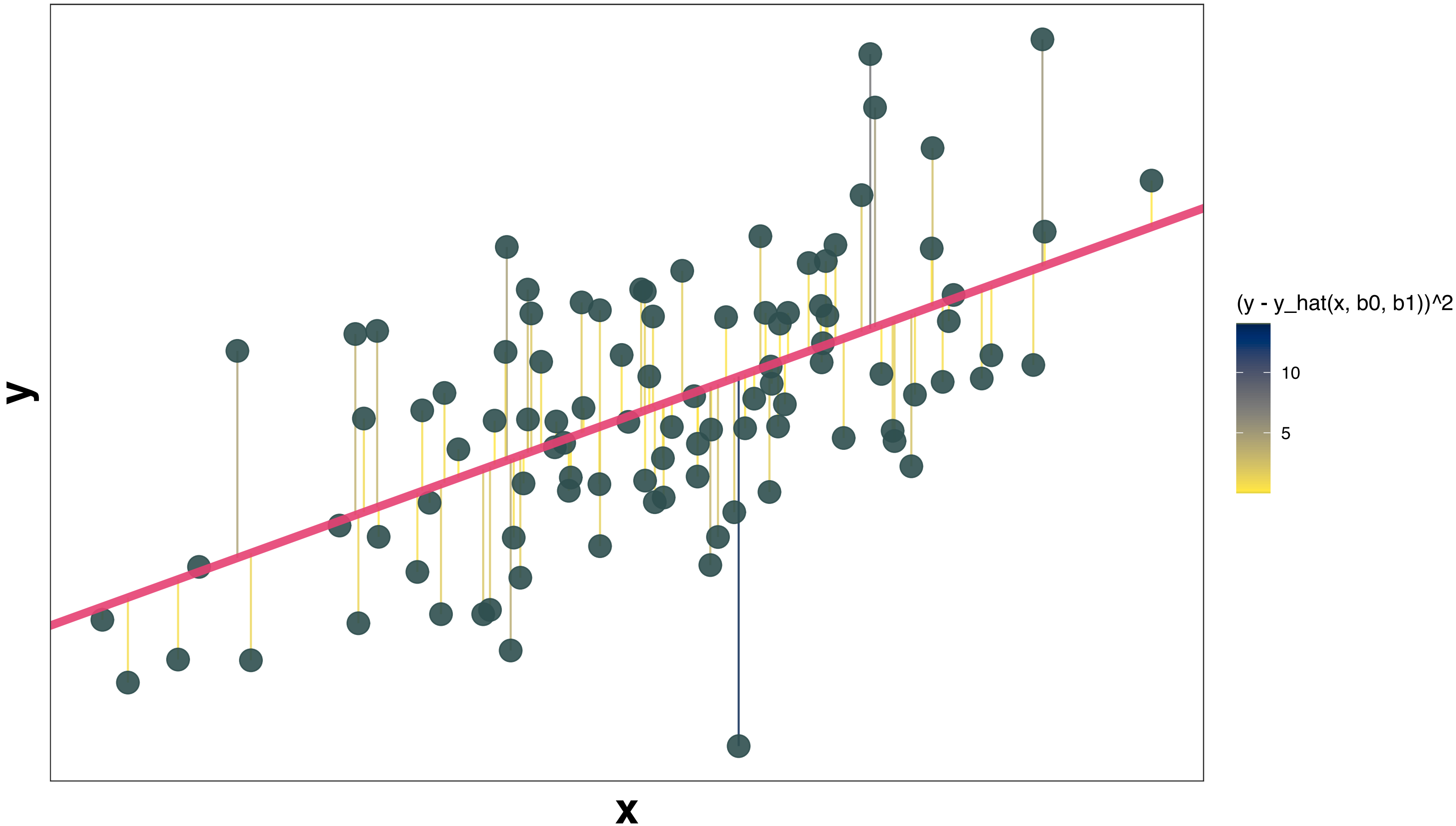
Ordinary Least Squares

SSE squares the errors $\left(\sum e_i^2\right)$: bigger errors get bigger penalties.



Ordinary Least Squares

The OLS estimate is the combination of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes SSE.



OLS, formally

In simple linear regression, the OLS estimator comes from choosing the $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squared errors (SSE), *i.e.*,

$$\min_{\hat{\beta}_0,\, \hat{\beta}_1} \mathrm{SSE}$$

but we already know $\mathrm{SSE} = \sum_i e_i^2$. Now use the definitions of e_i and $\hat{y}$.

$$e_i^2 = \left(y_i - \hat{y}_i\right)^2 = \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

this expands to:

$$e_i^2 = y_i^2 - 2 y_i \hat{\beta}_0 - 2 y_i \hat{\beta}_1 x_i + \hat{\beta}_0^2 + 2 \hat{\beta}_0 \hat{\beta}_1 x_i + \hat{\beta}_1^2 x_i^2$$

OLS, formally

Choose the $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squared errors (SSE), *i.e.*,

$$\min_{\hat{\beta}_0,\, \hat{\beta}_1} \sum_i e_i^2$$

Derivation: Minimizing a multivariate function requires (**1**) first derivatives equal zero (the *1st-order conditions*) and (**2**) second-order conditions (concavity).

See extra slides if you want the full derivation. Basically, we take the first derivatives of the SSE above with respect to β_0 and β_1, set them equal to zero, and solve for β_0 and β_1.

OLS, formally

The OLS estimator for the slope is:

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2} = \frac{cov(x, y)}{var(x)}$$

and the intercept:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Note that the expression for $\hat{\beta}_0$ can be rearranged to show us that our regression line always passes through the sample mean of x and y.

Let's collect some definitions

True **population** relationship:

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

Estimated **sample** relationship:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- **Dependent variable** = regressand = y
- **Independent variable** = explanatory variable = regressor = x
- **Residual** = sample error = $y_i - \hat{y}_i$
- Estimated **intercept** coefficient = $\hat{\beta}_0$
- Estimated **slope** coefficient = $\hat{\beta}_1$

Why choose the OLS line?

There are many possible ways to define a "best fit" linear relationship. For example:

- Least absolute deviations: minimize $\sum_i |y_i - \hat{y}_i|$
- Ridge regression: minimize $\sum_i \left[(y_i - \hat{y}_i)^2 + \lambda \sum_k \hat{\beta}_k^2\right]$
- ...

Why choose the OLS line?

There are many possible ways to define a "best fit" linear relationship.

So why do we often rely on OLS?

- Under a key set of assumptions, OLS satisfies some very desirable properties that most statisticians, economists, political scientists put emphasis on
- However, you will see many other linear (and nonlinear) estimators in machine learning
- What estimator you use depends on what the goal of your analysis is, but OLS is the best option a LOT of the time

Why choose the OLS line?

Under key assumptions, OLS satisfies two desirable properties:

- OLS is **unbiased**.
- OLS has the **minimum variance** of all unbiased linear estimators.

Let's dig into each of these for a moment so you can appreciate how amazing OLS is.

OLS property #1: Unbiasedness

Under a key set of assumptions (we'll get into these in a few slides), OLS is **unbiased**

Unbiasedness:

On average (after *many* samples), does the estimator tend toward the true population value?

More formally: The mean of estimator's distribution equals the population parameter it estimates:

$$\underset{\beta}{\mathrm{Bias}}\left(\hat{\beta}\right) = \boldsymbol{E}\left[\hat{\beta}\right] - \beta$$

OLS property #1: Unbiasedness

Under a key set of assumptions (we'll get into these in a few slides), OLS is **unbiased**

Unbiasedness:

On average (after *many* samples), does the estimator tend toward the true population value?

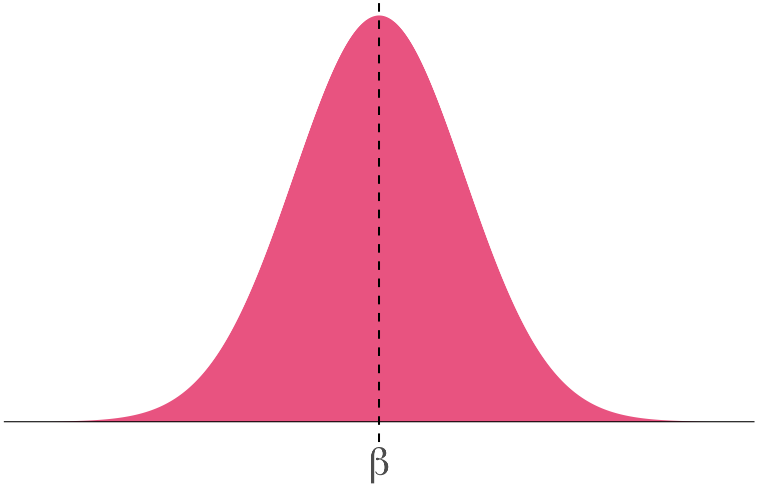
$\rightarrow$ You should think about the distribution of $\hat{\beta}$ values as the distribution of regression results you would get if you could draw many random samples from the population and generate a new $\hat{\beta}$ every time.

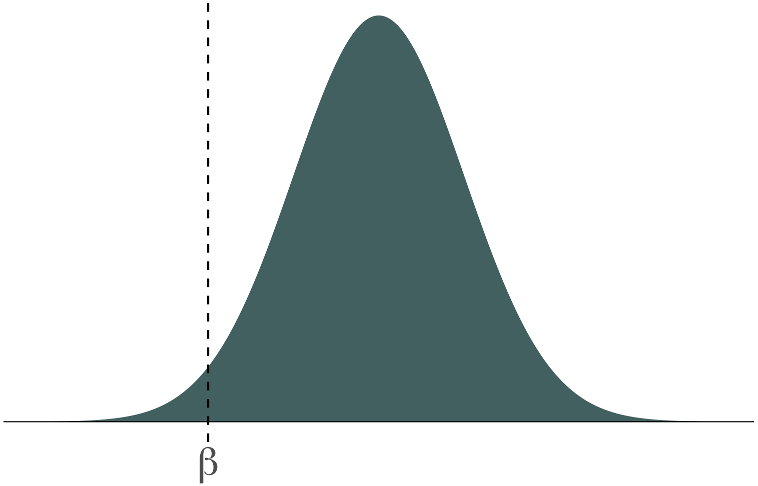
$\rightarrow$ In two weeks we'll talk a lot more about uncertainty in and distributions of estimators like $\hat{\beta}$.

OLS property #1: Unbiasedness

Unbiased estimator: $\boldsymbol{E}\left[\hat{\beta}\right] = \beta$

Biased estimator: $\boldsymbol{E}\left[\hat{\beta}\right] \neq \beta$





Distributions show probability density function of $\hat{\beta}$ estimates recovered from many different randomly drawn samples.

OLS property #2: Lowest variance

Under a key set of assumptions (again, let's wait a couple slides), OLS is the estimator with the **lowest variance**

Lowest variance:

Just as we discussed when defining summary statistics, the central tendencies (means) of distributions are not the only things that matter. We also care about the **variance** of an estimator.

$$\mathrm{Var}\left(\hat{\beta}\right) = \boldsymbol{E}\left[\left(\hat{\beta} - \boldsymbol{E}\left[\hat{\beta}\right]\right)^2\right]$$

Lower variance estimators mean we get estimates closer to the mean in each sample.

OLS property #2: Lowest variance

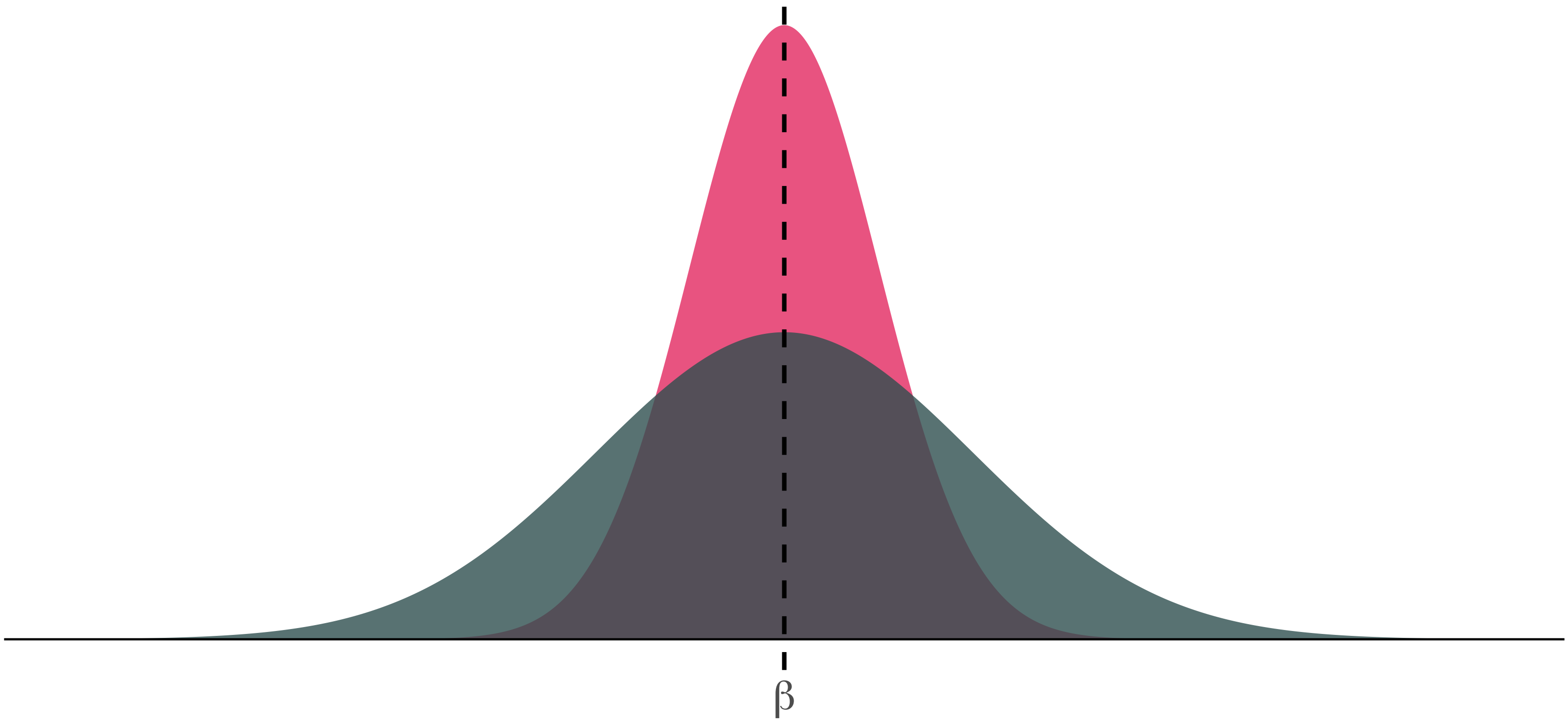
Under a key set of assumptions (again, let's wait a couple slides), OLS is the estimator with the **lowest variance**

Lowest variance:

Just as we discussed when defining summary statistics, the central tendencies (means) of distributions are not the only things that matter. We also care about the **variance** of an estimator.

$\rightarrow$ Again, think about the distribution of $\hat{\beta}$ values as the distribution of regression results you would get if you could draw many random samples from the population and generate a new $\hat{\beta}$ every time.

OLS property #2: Lowest variance



Properties of OLS

Property 1: Bias.

Property 2: Variance.

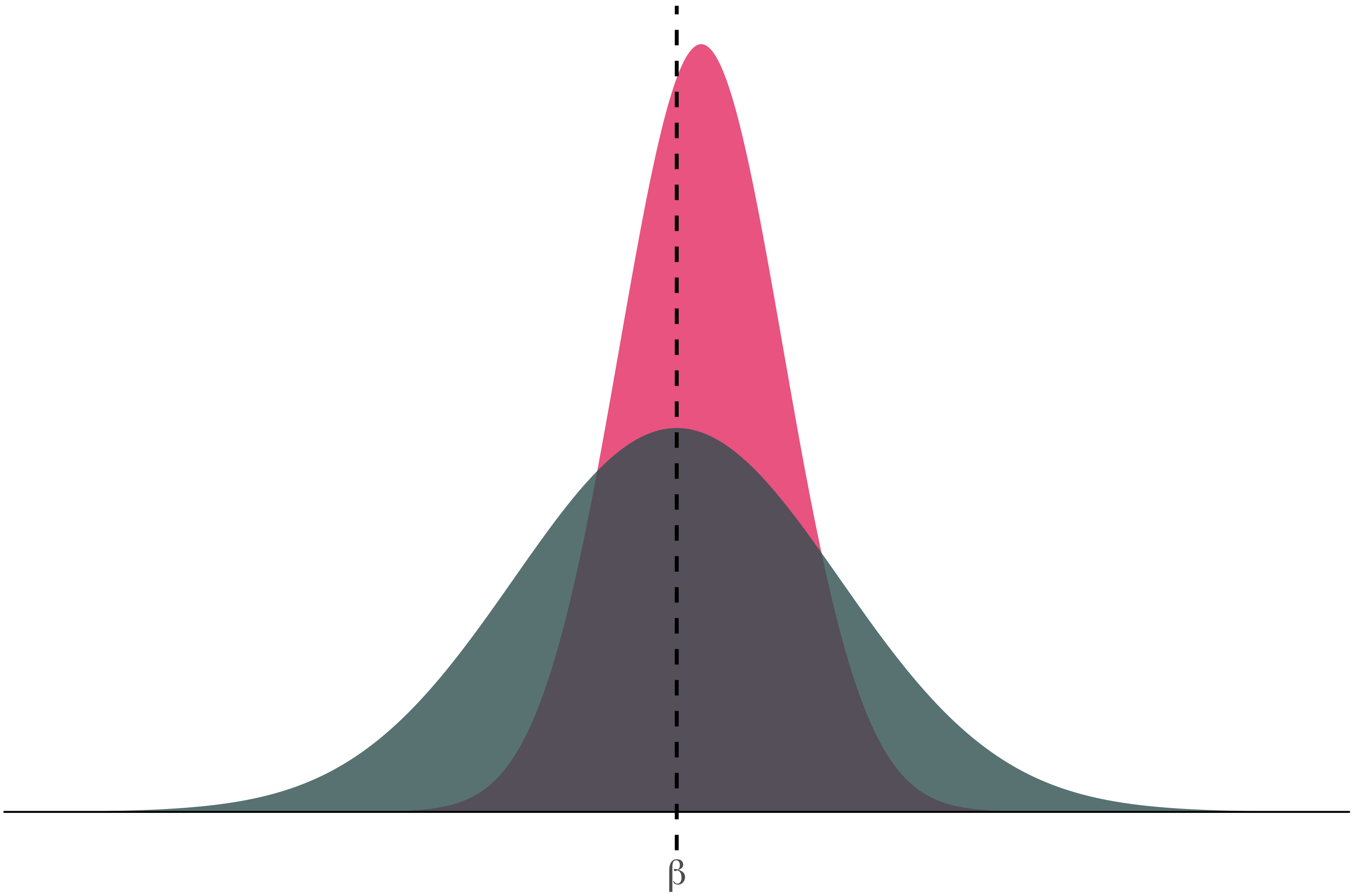
Subtlety: The bias-variance tradeoff.

Should we be willing to take a bit of bias to reduce the variance?

In much of statistics, we choose unbiased estimators. But other disciplines (especially computer science) will choose estimators that sacrifice some bias in exchange for lower variance.

You'll learn more about these estimators (e.g., ridge regression) in EDS 232 👀

The bias-variance tradeoff.



OLS: Assumptions

These very nice properties depend on a key set of assumptions:

1. The population relationship is linear in parameters with an additive disturbance.

2. The $\boldsymbol{X}$ variable is **exogenous**, *i.e.*, $\boldsymbol{E}[u \mid X] = 0$.

   - I.e., is there no other variable correlated with $\boldsymbol{X}$ that also affects $\boldsymbol{Y}$
   - You will talk a lot more about this in EDS 241 👀

3. The $\boldsymbol{X}$ variable has variation (and if there are multiple explanatory variables, they are not perfectly collinear)

   - Recall, $\boldsymbol{var}(\boldsymbol{x})$ is in the denominator of the OLS slope coefficient estimator!

OLS: Assumptions

These very nice properties depend on a key set of assumptions:

1. The population relationship is linear in parameters with an additive disturbance.

2. Our X variable is **exogenous**, *i.e.*, $\boldsymbol{E}[u \mid X] = 0$.

3. The X variable has variation.

4. The population disturbances u_i are independently and identically distributed as **normal** random variables with mean zero ($\boldsymbol{E}[u] = 0$) and variance σ^2 (*i.e.*, $\boldsymbol{E}\left[u^2\right] = \sigma^2$)

    - Independently distributed and mean zero jointly imply $\boldsymbol{E}\left[u_i u_j\right] = 0$ for any $i \neq j$
    - Constant variance means errors cannot vary with X (this is called "homoskedasticity")

OLS: Assumptions

Different assumptions guarantee different properties:

- Assumptions (1), (2), and (3) make OLS **unbiased**
- Assumption (4) gives us an unbiased estimator for the **variance** of our OLS estimator (we will talk more about this when covering *inference* in a couple weeks)

We will discuss the many ways real life may **violate these assumptions**. For instance:

- Non-linear relationships in our parameters/disturbances (or misspecification) $\rightarrow$ e.g., logistic regression
- Disturbances that are not identically distributed and/or not independent $\rightarrow$ lectures on *inference*
- Violations of exogeneity (especially omitted-variable bias) $\rightarrow$ mostly covered in EDS 241

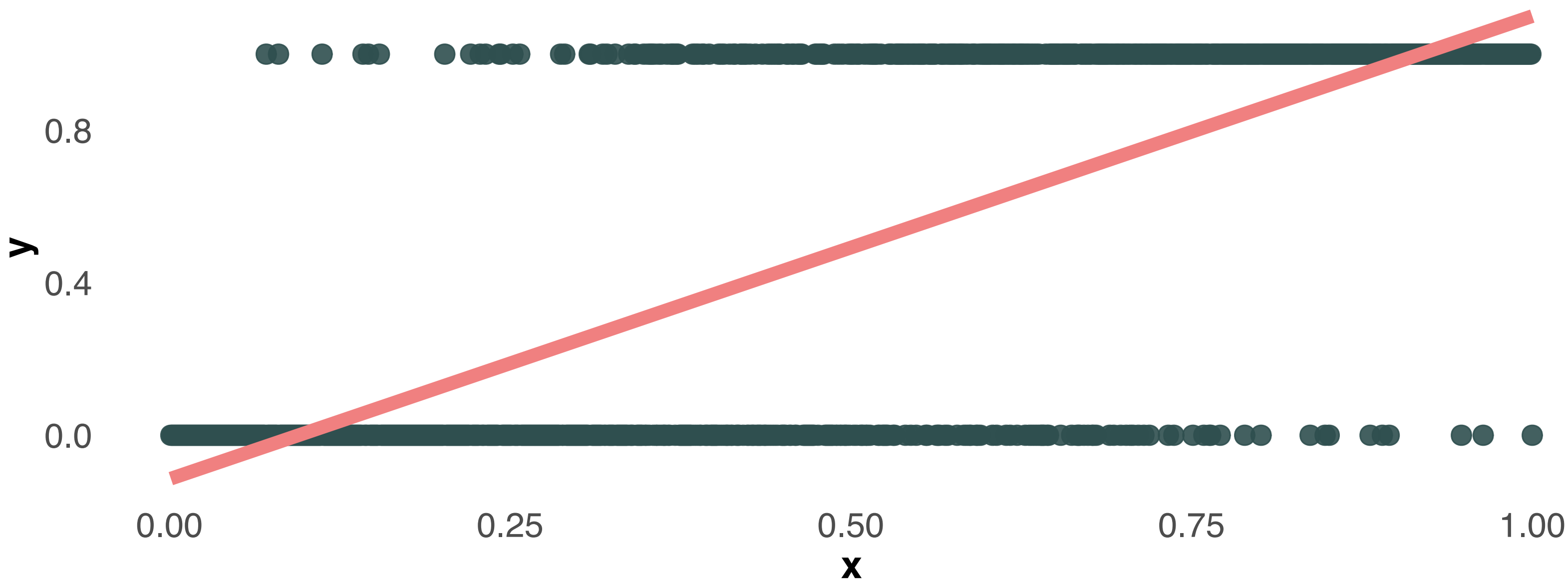
OLS: Assumptions

Q: Can we test these assumptions?

> A: Some of them.

Assumption 1: Linear in parameters.

You can look at your data to see if this might be reasonable.



OLS: Assumptions

Q: Can we test these assumptions?

> A: Some of them.

Assumption 1: Linear in parameters.

You can look at your data to see if this might be reasonable.

- Note: this assumption does not require your model to be linear in $\boldsymbol{X}$! As we discuss later, nonlinear relationships in $\boldsymbol{X}$ *can* be easily accommodated with OLS:

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

This equation was estimated using OLS to give the nonlinear relationship on the next slide.

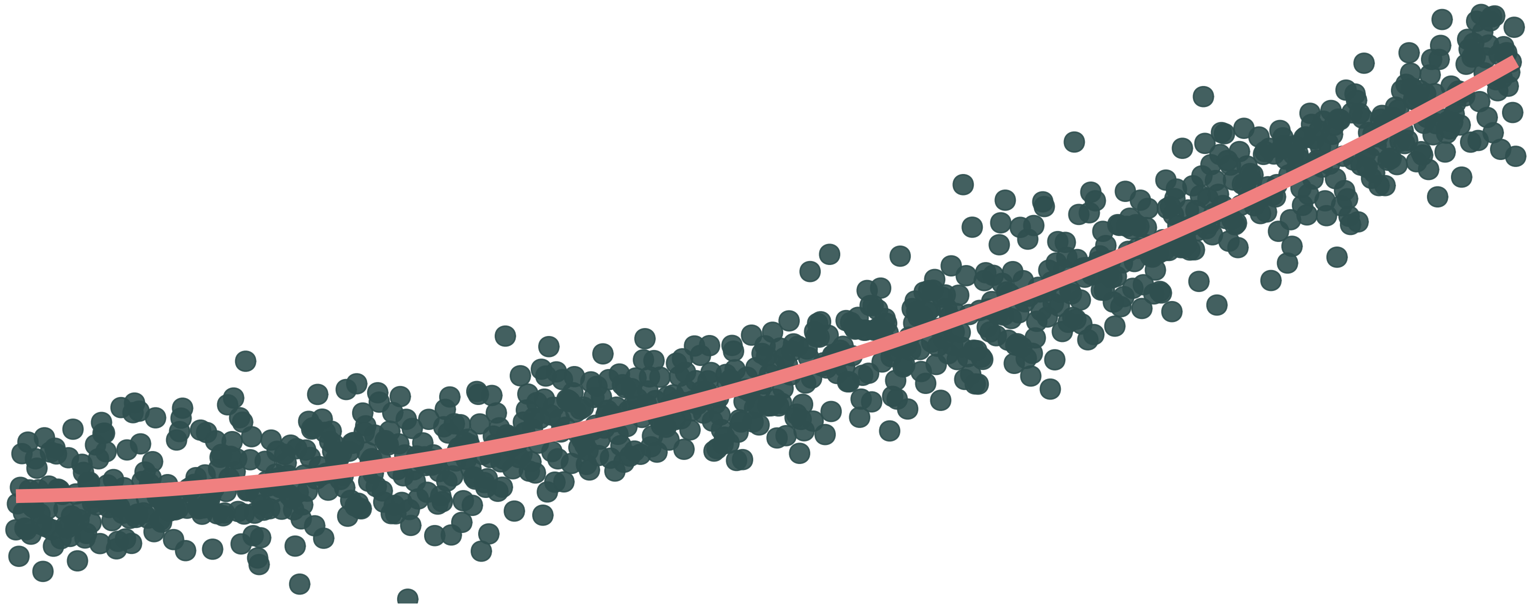
OLS: Assumptions

Q: Can we test these assumptions?

> A: Some of them.

Assumption 1: Linear in parameters.

You can look at your data to see if this might be reasonable.



OLS: Assumptions

Q: Can we test these assumptions?

> A: Some of them.

Assumption 1: Linear in parameters.

Example of a population relationship that is *not* linear in parameters:

$$Y = e^{\beta_0 + \beta_1 X + u}$$

OLS: Assumptions

Q: Can we test these assumptions?

> A: Some of them.

Assumption 2: Exogeneity

$$\boldsymbol{E}[u \mid X] = 0$$

This is not a testable assumption!

There are a lot of methods designed to probe this assumption, but it's fundamentally untestable since there are infinite possible correlates of X and Y that are unobservable to the researcher.

In general, you should always think about what is in u that may be correlated with X.

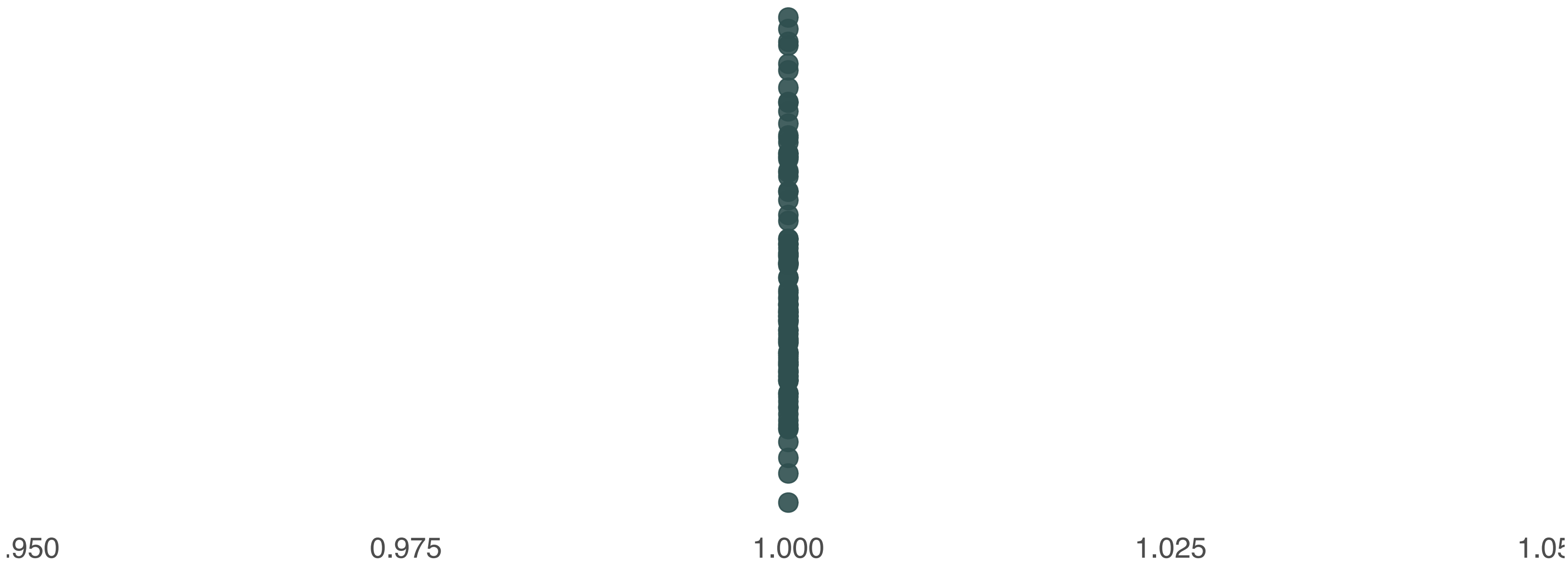
OLS: Assumptions

Q: Can we test these assumptions?

> A: Some of them.

Assumption 3: X has variation.

This is very easy to test:



OLS: Assumptions

Q: Can we test these assumptions?

> A: Some of them.

Assumption 4: The population disturbances u_i are independently and identically distributed as **normal** random variables with mean zero and variance σ^2

Use the residuals from your regression to investigate this assumption

Step 1: Run linear regression

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

Step 2: Generate residuals

$$e_i = y_i - \hat{y}_i$$

OLS: Assumptions

Q: Can we test these assumptions?

> A: Some of them.

Assumption 4: The population disturbances u_i are independently and identically distributed as **normal** random variables with mean zero and variance σ^2

Use the residuals from your regression to investigate this assumption

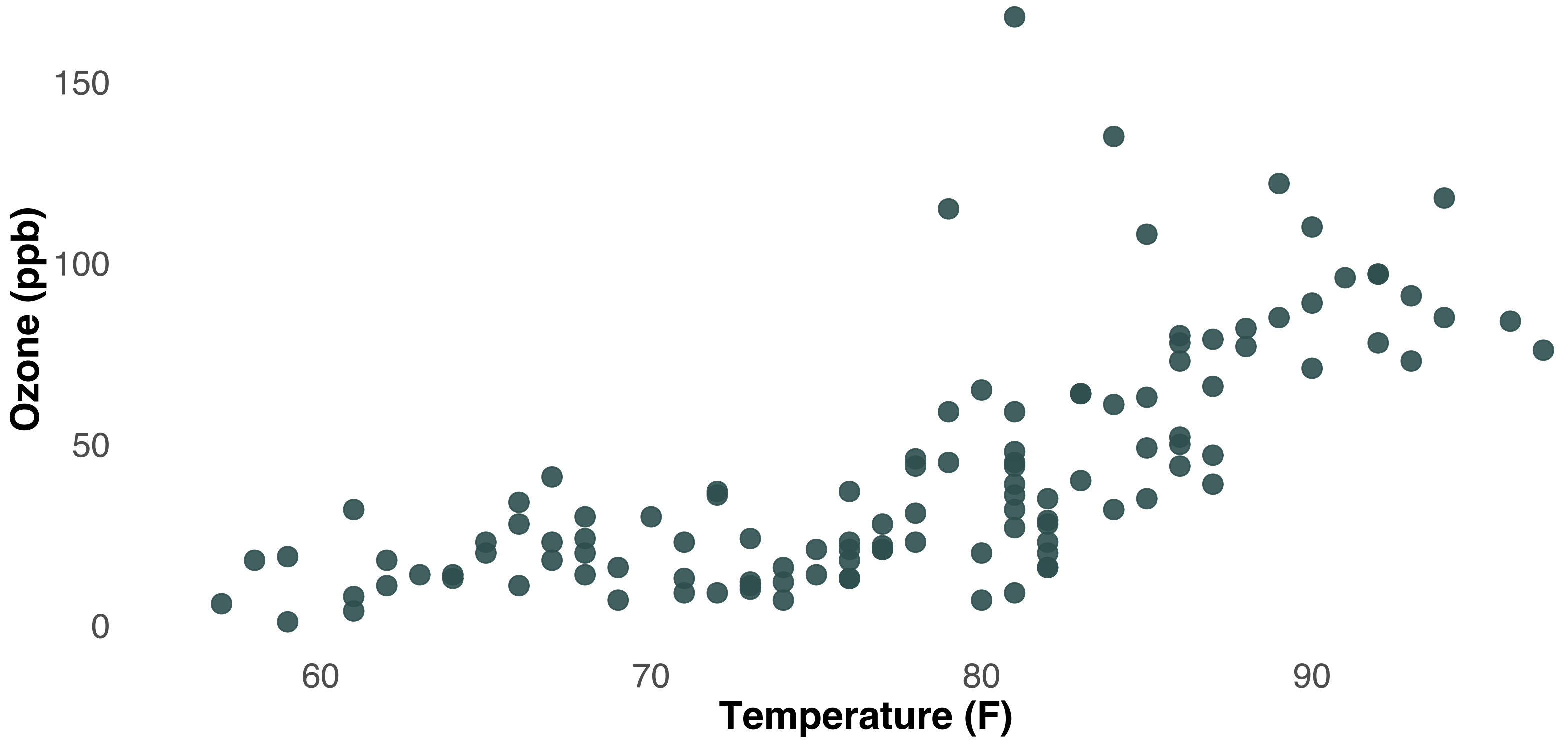
Step 3: Plot and investigate residuals [draw these examples]

- histogram (are they normal?)
- plot of e_i against X (are they uncorrelated? does the variance depend on X?)

Interpreting regression results

Interpreting OLS results

Example: Ozone increases due to temperature (NYC)



Interpreting OLS results

Example: Ozone increases due to temperature (NYC)

We can use `lm(y~x, my_data)` in R to run a linear regression of y on x, including a constant term.

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

Interpreting OLS results

Example: Ozone increases due to temperature (NYC)

`summary()` then lets us see the regression results.

How do we interpret these??

Interpreting OLS results

```
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)
#> Multiple R-squared:  0.4877,    Adjusted R-squared:  0.4832
#> F-statistic: 108.5 on 1 and 114 DF,  p-value: < 2.2e-16
```

Interpreting OLS results

$$Ozone_i = \beta_0 + \beta_1 Temp_i + \varepsilon_i$$

```
#> 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 ***
```

- **Slope**: Change in y for a one unit change in x.
  - Here: On average, we expect to see ozone increase by 2.4 ppb for each 1 degree F increase in temperature.
- **Intercept**: Level of y when $x = 0$.
  - Here: On average, we expect Ozone to be -147 ppb when temperature is 0 degrees F.
  - **CAREFUL** with extrapolation! This doesn't even make sense!

Interpreting OLS results

$$Ozone_i = \beta_0 + \beta_1 Temp_i + \varepsilon_i$$

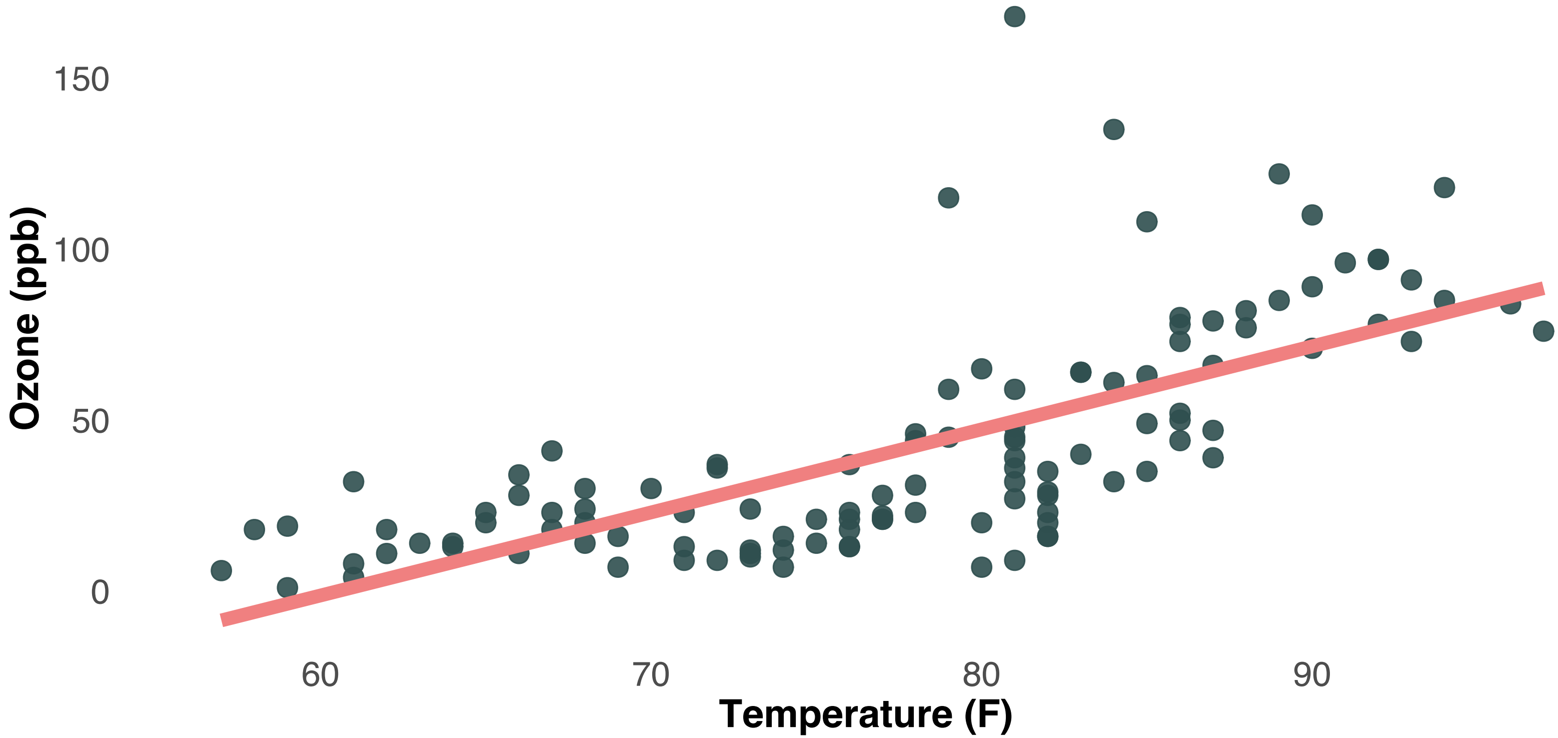
```
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)
#> (Intercept) -146.9955    18.2872  -8.038 9.37e-13 ***
#> Temp           2.4287     0.2331  10.418  < 2e-16 ***
```

- Standard error, t-value, and Pr(>t): These all concern **uncertainty** around our parameter estimates. We will tackle these fully in a week or so.

Interpreting OLS results

Visualizing our predicted model using `geom_smooth()`

> Where is β_0? Where is β_1?



Interpreting OLS results

Units matter!

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

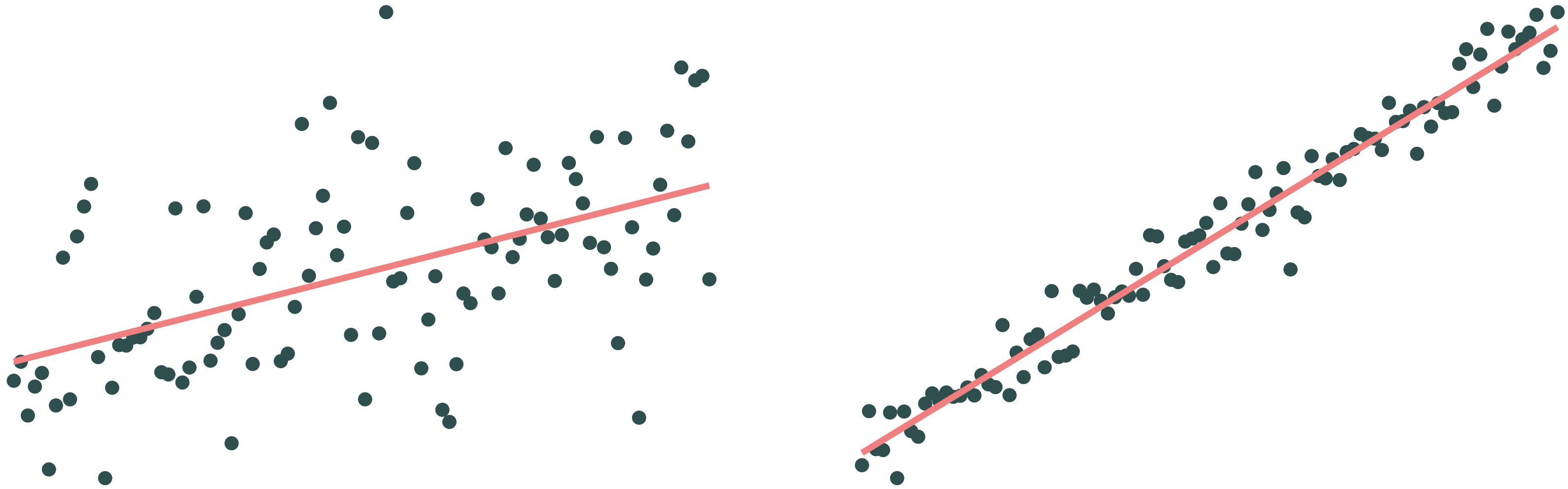
```
#>
#> Call:
#> lm(formula = Ozone ~ TempC, data = airquality)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -40.729 -17.409  -0.587  11.306 118.271
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) -69.2770    10.9182  -6.345 4.65e-09 ***
#> TempC         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
```

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
```

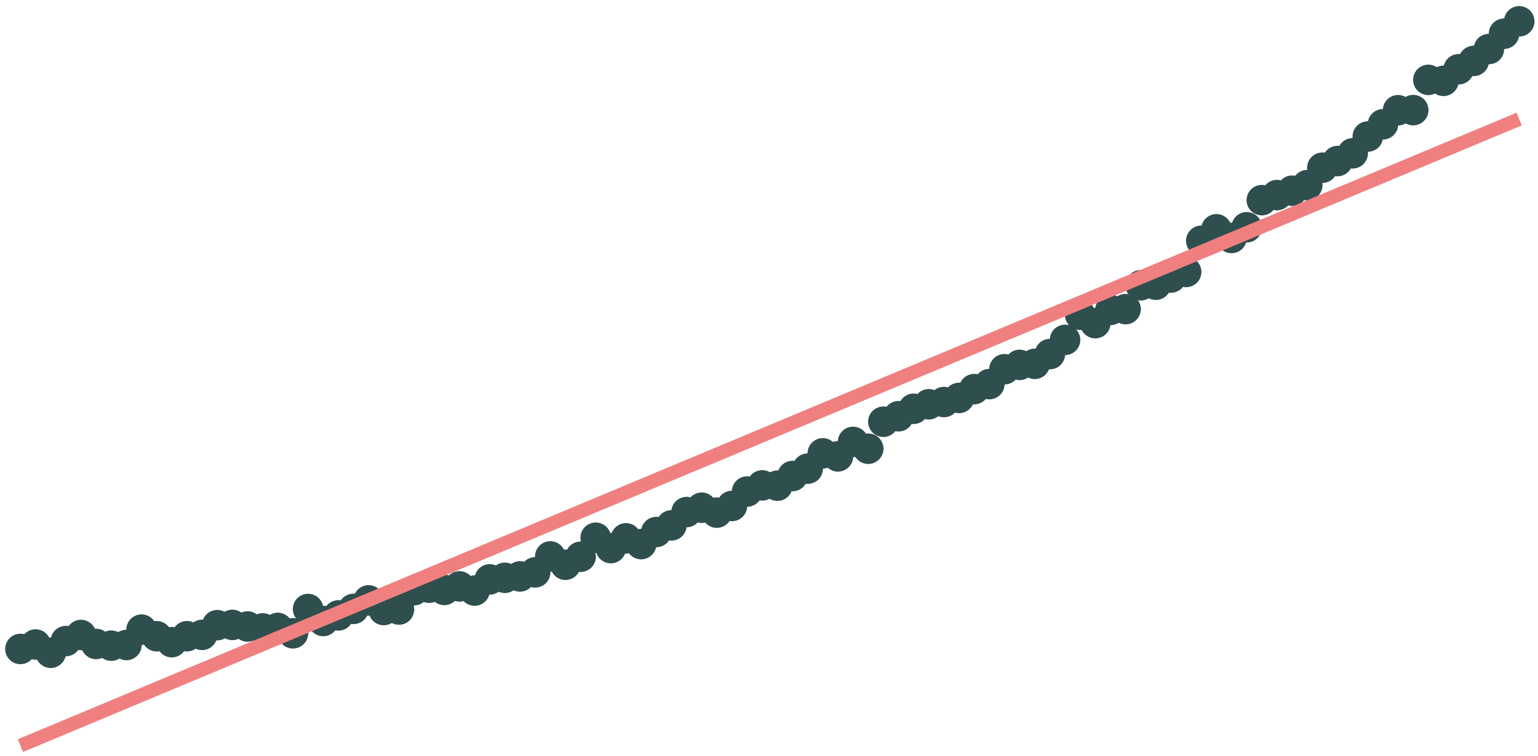
Coefficient of determination

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!

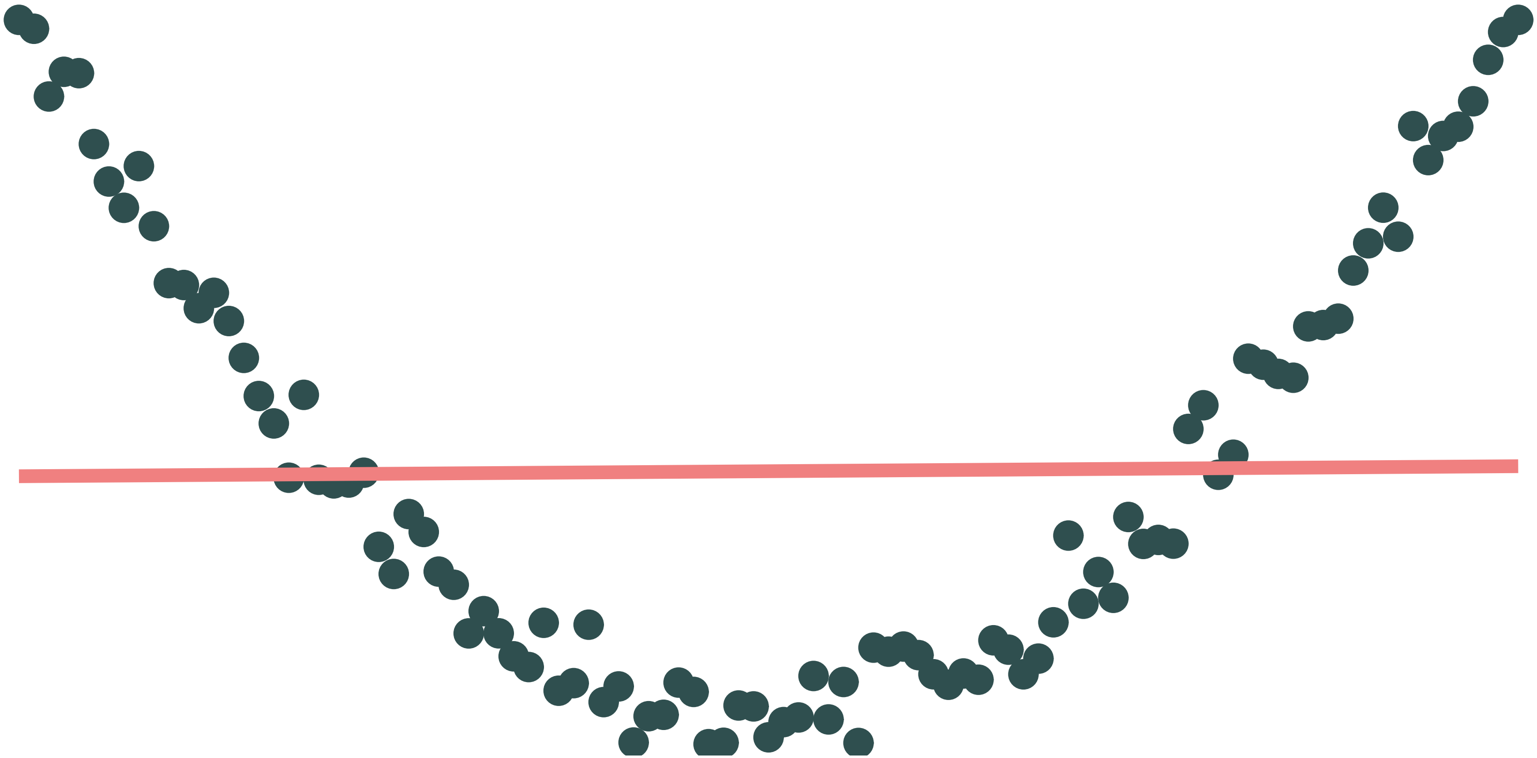
Coefficient of determination

Here, $R^2 = 0.94$. 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?



Important 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.

```
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.

- This is *systematic non-response* from Lecture 01

[*]: Or perform almost any operation/function

Multiple linear regression

Multiple linear regression (preview)

The true population model probably involves **other regressors**:

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$ This 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?

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

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

Extra slides

OLS, formally

In simple linear regression, the OLS estimator comes from choosing the $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squared errors (SSE), *i.e.*,

$$\min_{\hat{\beta}_0,\, \hat{\beta}_1} \mathrm{SSE}$$

but we already know $\mathbf{SSE} = \sum_i e_i^2$. Now use the definitions of e_i and $\hat{y}$.

$$e_i^2 = \left(y_i - \hat{y}_i\right)^2 = \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

this expands to:

$$e_i^2 = y_i^2 - 2y_i\hat{\beta}_0 - 2y_i\hat{\beta}_1 x_i + \hat{\beta}_0^2 + 2\hat{\beta}_0\hat{\beta}_1 x_i + \hat{\beta}_1^2 x_i^2$$

Recall: Minimizing a multivariate function requires (**1**) first derivatives equal zero (the *1st-order conditions*) and (**2**) second-order conditions (concavity).

OLS, formally

We're getting close. We need to **minimize SSE**. We've showed how SSE relates to our sample (our data: x and y) and our estimates (*i.e.*, $\hat{\beta}_0$ and $\hat{\beta}_1$).

$$\text{SSE} = \sum_i e_i^2 = \sum_i \left(y_i^2 - 2y_i\hat{\beta}_0 - 2y_i\hat{\beta}_1 x_i + \hat{\beta}_0^2 + 2\hat{\beta}_0\hat{\beta}_1 x_i + \hat{\beta}_1^2 x_i^2 \right)$$

For the first-order conditions of minimization, we now take the first derivates of SSE with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$.

$$\frac{\partial \text{SSE}}{\partial \hat{\beta}_0} = \sum_i \left(2\hat{\beta}_0 + 2\hat{\beta}_1 x_i - 2y_i \right) = 2n\hat{\beta}_0 + 2\hat{\beta}_1 \sum_i x_i - 2\sum_i y_i \qquad = 2n\hat{\beta}_0$$

where $\overline{x} = \frac{\sum x_i}{n}$ and $\overline{y} = \frac{\sum y_i}{n}$ are sample means of x and y (size n).

OLS, formally

The first-order conditions state that the derivatives are equal to zero, so:

$$\frac{\partial \mathrm{SSE}}{\partial \hat{\beta}_0} = 2n\hat{\beta}_0 + 2n\hat{\beta}_1\overline{x} - 2n\overline{y} = 0$$

which implies

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}$$

Now for $\hat{\beta}_1$.

OLS, formally

Take the derivative of SSE with respect to $\hat{\beta}_1$

\dfrac{\partial \text{SSE}}{\partial \hat{\beta}_1} &= \sum_i \left(2 \hat{\beta}_0

$$= 2n\hat{\beta}_0\overline{x} + 2\hat{\beta}_1 \sum_i x_i^2 - 2\sum_i y_i x_i$$

set it equal to zero (first-order conditions, again)

$$\dfrac{\partial \text{SSE}}{\partial \hat{\beta}_1} = 2n\hat{\beta}_0\overline{x} + 2\hat{\beta}_1 \sum_i x_i^2 - 2\sum_i y_i x_i = 0$$

and substitute in our relationship for $\hat{\beta}_0$, *i.e.*, $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}$. Thus,

$$2n\left(\overline{y} - \hat{\beta}_1\overline{x}\right)\overline{x} + 2\hat{\beta}_1 \sum_i x_i^2 - 2\sum_i y_i x_i = 0$$

OLS, formally

Continuing from the last slide

$$2n\left(\overline{y} - \hat{\beta}_1 \overline{x}\right)\overline{x} + 2\hat{\beta}_1 \sum_i x_i^2 - 2\sum_i y_i x_i = 0$$

we multiply out

$$2n\overline{y}\,\overline{x} - 2n\hat{\beta}_1 \overline{x}^2 + 2\hat{\beta}_1 \sum_i x_i^2 - 2\sum_i y_i x_i = 0$$

$$\implies 2\hat{\beta}_1 \left(\sum_i x_i^2 - n\overline{x}^2\right) = 2\sum_i y_i x_i - 2n\overline{y}\,\overline{x}$$

$$\implies \hat{\beta}_1 = \frac{\sum_i y_i x_i - 2n\overline{y}\,\overline{x}}{\sum_i x_i^2 - n\overline{x}^2} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}$$

OLS, formally

Done!

We now have (lovely) OLS estimators for the slope

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}$$

and the intercept

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

And now you know where the *least squares* part of ordinary least squares comes from. 🎊