Linear Regression

INFO 370

Learning Objectives

Check-in on course progress

Discuss insights, challenges, and new metrics from **assignment 2**

Contextualize linear regression within the area of **statistical learning**

Introduce simple (univariate) linear regression

Leverage linear regression to **describe how salary is associated with** gender, rank, and experience on a University faculty (notebook-set-4)

Course Check-in

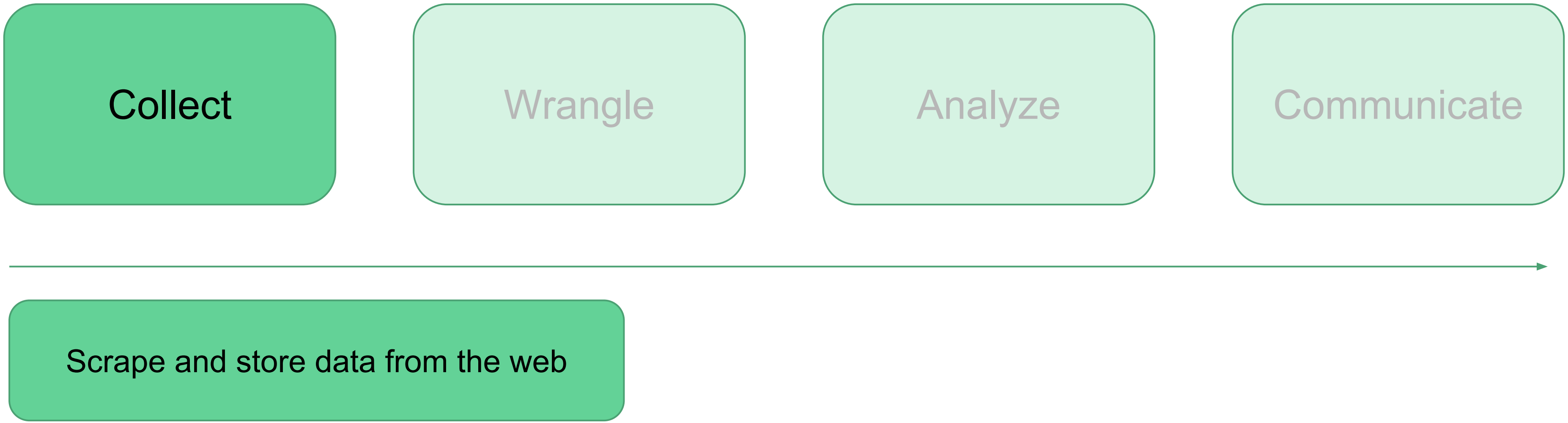
So far...

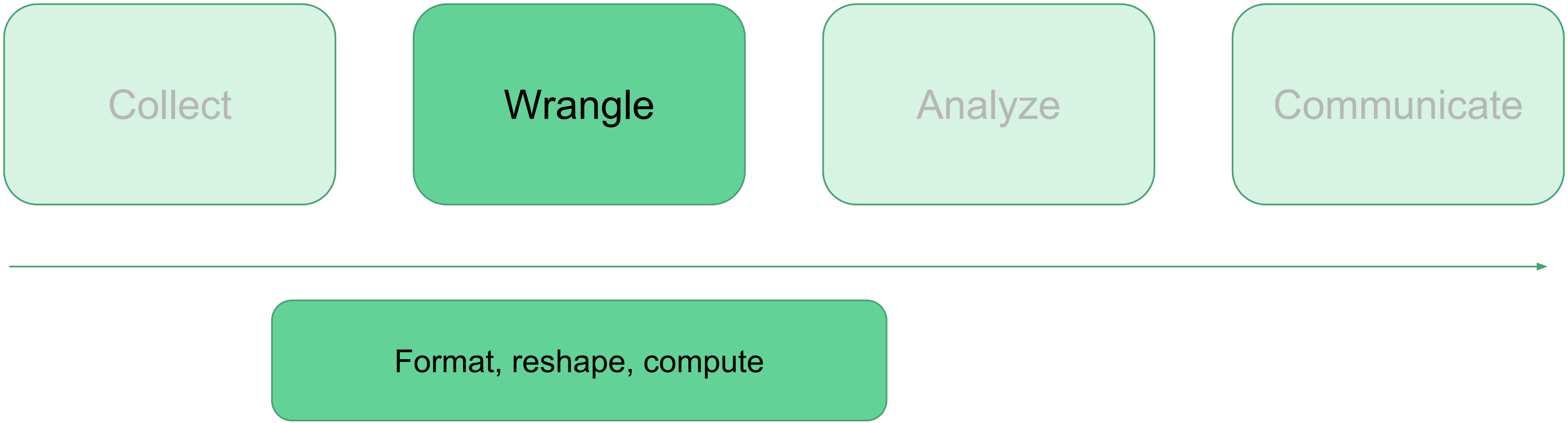
Programming for Data Science (R, Python)

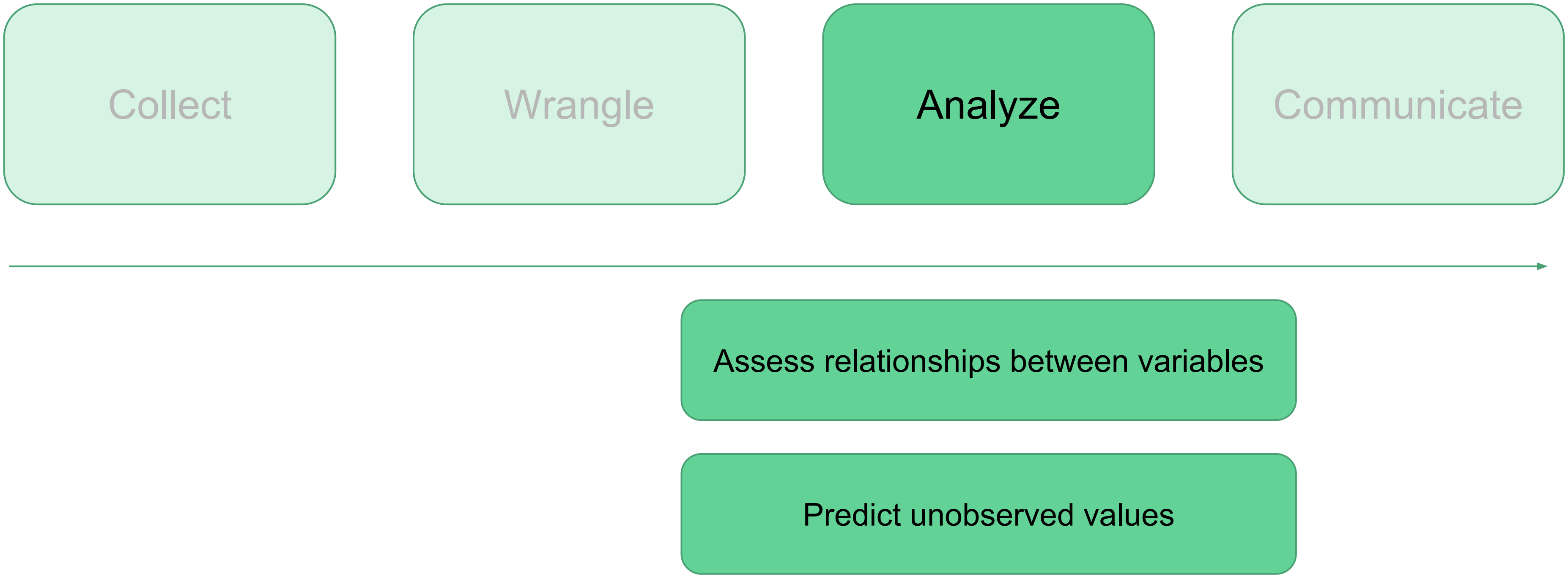
Metric computation

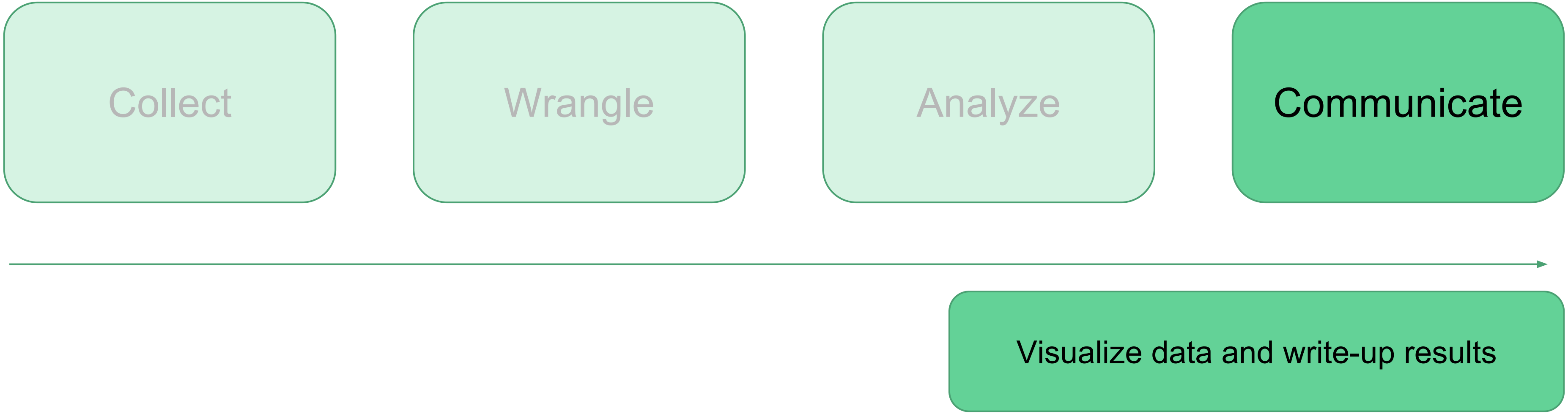
Probability + Statistics fundamentals

Statistical tests as a form of hypothesis testing


Collect
Wrangle
Analyze
Communicate
Scrape and store data from the web


Collect
Wrangle
Analyze
Communicate
Format, reshape, compute


Collect
Wrangle
Analyze
Communicate
Assess relationships between variables
Predict unobserved values


Collect
Wrangle
Analyze
Communicate
Visualize data and write-up results

How's it been going?

INFO 370 Check-in Survey

When survey is active, respond at **PollEv.com/mikefree**

0 surveys done

0 surveys underway

Start the presentation to see live content. Still no live content? Install the app or get help at **PollEv.com/app**

Has this course demanded the appropriate time commitment?

Do students who have taken INFO 201 spend more or less time on the course?

What's next?

Topics

- Regression methods for hypothesis testing
- Machine learning as a tool for prediction

Assignments

- Measuring associations with statistical methods (R3, A3)
- Making predictions with machine learning (A4)
- Data ethics (R4)
- Final projects (groups formed in lab **next week**)

Assignment 2

Segregation Metrics

Share your assignment with those around you and discuss:

- What did you learn about segregation metrics?
- What did you learn about specific cities?
- What did you learn (about R/Rmd)?
- What were the challenges of implementing this assignment?

Linear Regression Context

Statistical learning methods provide a set of tools used to **ask questions about data**. They leverage *mathematical concepts* and *computational abilities* to **make inferences** about relationships, or **make predictions** about unobserved contexts.

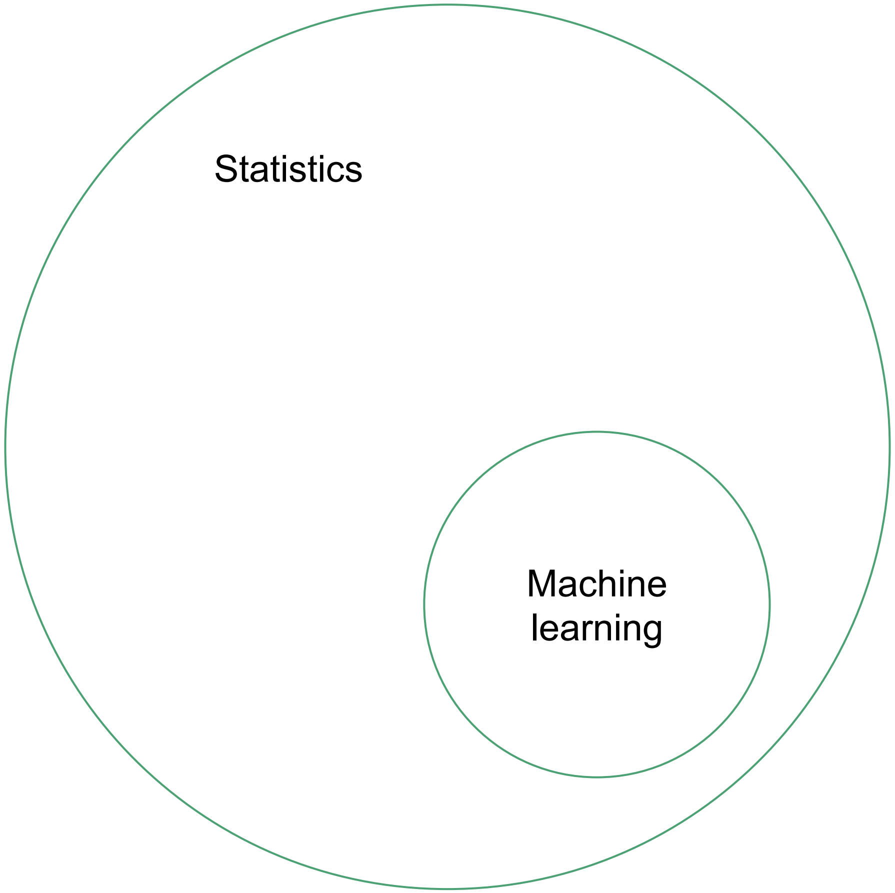
Inference
(interpretability)

Prediction
(accuracy)

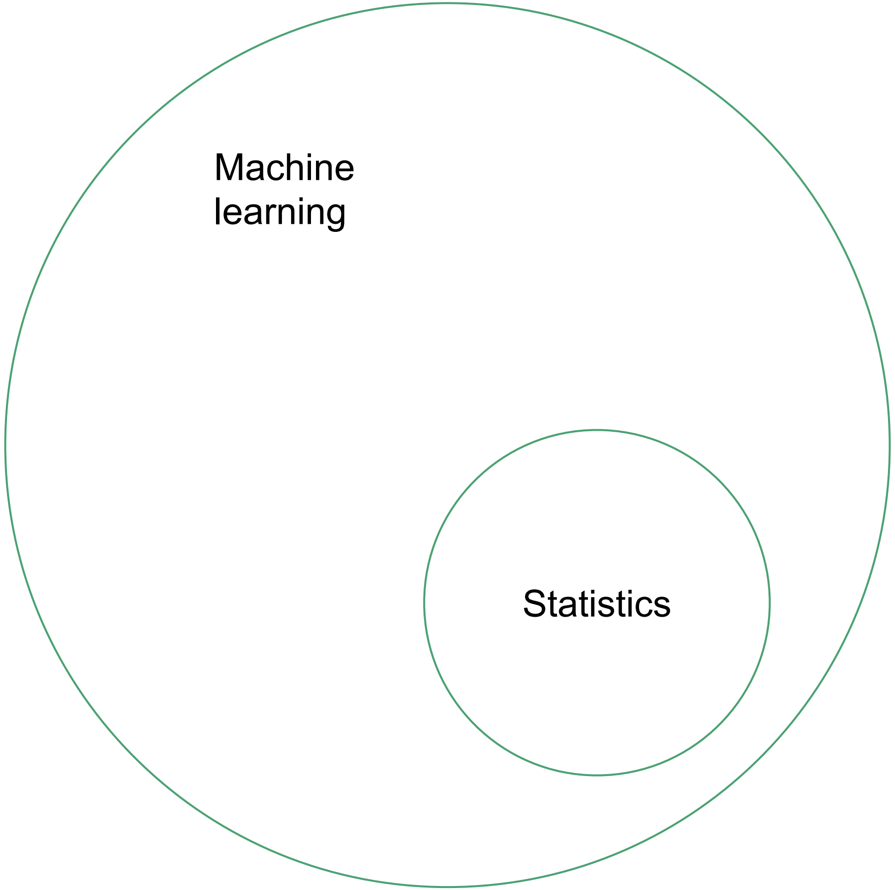
(more stats)

(more ml)

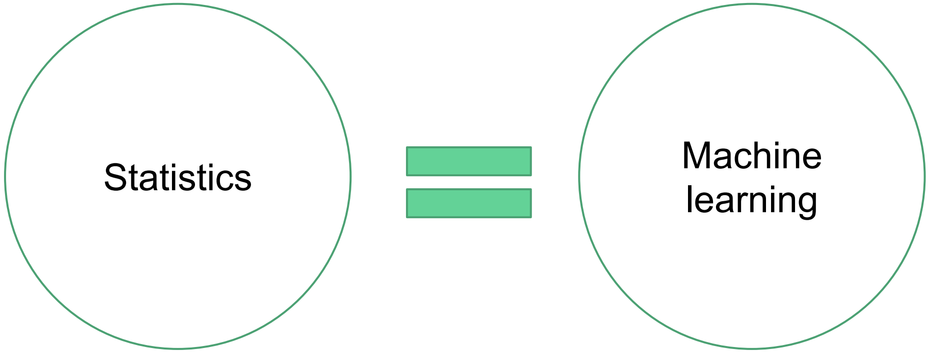
Tradeoffs



Many people argue this....



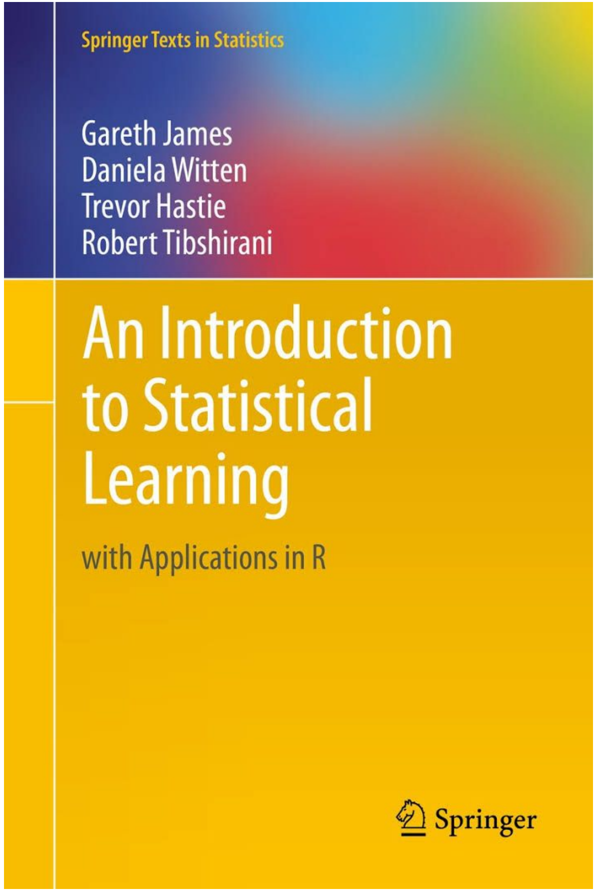
While many others argue this…



Glossary

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = $1,000,000	large grant= $50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

And others think this...



A wonderfully written book, chapter 3 used throughout this lecture (free download)

Advertising budgets (by type) and sales in each city

	X	TV	Radio	Newspaper	Sales
1	1	230.1	37.8	69.2	22.1
2	2	44.5	39.3	45.1	10.4
3	3	17.2	45.9	69.3	9.3
4	4	151.5	41.3	58.5	18.5
5	5	180.8	10.8	58.4	12.9
6	6	8.7	48.9	75.0	7.2
7	7	57.5	32.8	23.5	11.8
8	8	120.2	19.6	11.6	13.2

What questions can we ask of this dataset (download)?

Linear regression questions

1. Is there a relationship between advertising budget and sales?

2. How strong is the relationship between advertising budget and sales?

3. Which media contribute to sales?

4. How accurately can we estimate the effect of each medium on sales?

5. How accurately can we predict future sales?

6. Is the relationship linear?

7. Is there synergy among the advertising media?

	X	TV	Radio	Newspaper	Sales
1	1	230.1	37.8	69.2	22.1
2	2	44.5	39.3	45.1	10.4
3	3	17.2	45.9	69.3	9.3
4	4	151.5	41.3	58.5	18.5
5	5	180.8	10.8	58.4	12.9
6	6	8.7	48.9	75.0	7.2
7	7	57.5	32.8	23.5	11.8
8	8	120.2	19.6	11.6	13.2

Linear regression questions (more generally)

What are the **strength**, **magnitude**, and **uncertainty** associated with the relationships between **independent** and **dependent** variables?

	X	TV	Radio	Newspaper	Sales
1	1	230.1	37.8	69.2	22.1
2	2	44.5	39.3	45.1	10.4
3	3	17.2	45.9	69.3	9.3
4	4	151.5	41.3	58.5	18.5
5	5	180.8	10.8	58.4	12.9
6	6	8.7	48.9	75.0	7.2
7	7	57.5	32.8	23.5	11.8
8	8	120.2	19.6	11.6	13.2

Simple Linear Regression

There are entire graduate courses on Linear Regression.

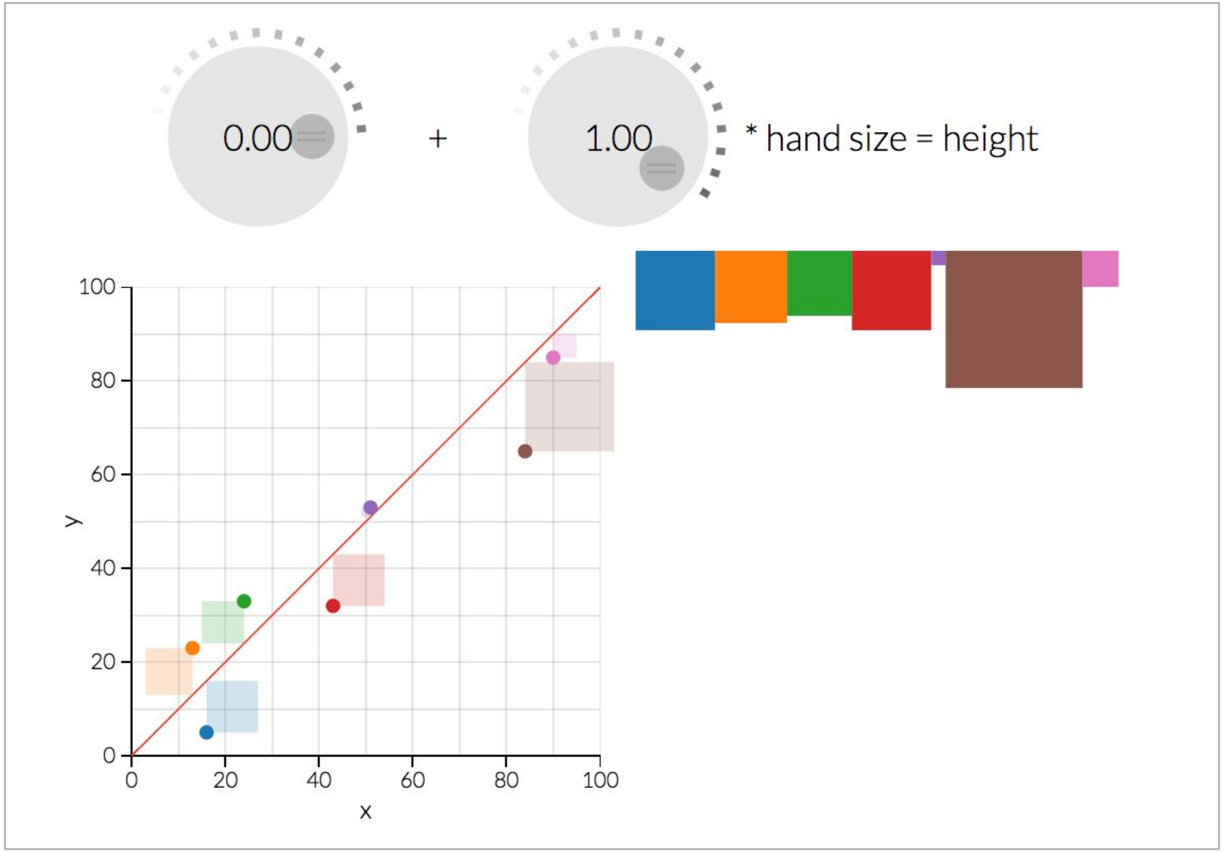
Simple Linear Regression

Estimate a (linear) functional form for predicting a **quantitative** response Y on a **single** predictor variable X ($\approxeq$ "approximately modeled as")

$$Y \approxeq \beta_0 + \beta_1 X$$

We estimates specific betas (*beta hat*) and compute estimates of y (*y hat*) using values of X (where X = x)

$$\hat{y} = \hat{B}_0 + \hat{B}_1 x$$



Estimating the coefficients (betas) -- [link]()

Interpreting Coefficients (Betas)

What is the interpretation of this formula?

$$income \approx 15000 + 2000 * years\text{-}education$$

Intercept of 15,000 (where a line intercepts the y axis -- income with 0 years ed.)

Slope of 2,000 (each additional year of education **is associated with** an increase in income of $2,000)

Assessing Accuracy

Assessing Coefficient Estimate Accuracy

Betas are chosen to minimize the **residual sum of squares**

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

The RSS can be used to **estimate** the **standard error of the residuals (RSE)**

$$RSE = \sigma = \sqrt{RSS/(n-2)}$$

RSE can then be used to estimate the **standard errors of the betas**

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

We can then (finally) compute confidence intervals around our betas

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$$

Interpreting Coefficient Confidence Intervals

Given the following formula:

$$income \approx 15000 + 2000 * years\text{-}education$$

Interpret a confidence interval [1500, 2500] for the *years-education* coefficient

There is a 95% chance that the **true value of B1** (association between years-education and income) falls between 1500 and 2500.

There is only a 5% chance that this data was observed **by chance.**

Hypothesis testing with coefficients

Given a dataset of *income* and *years-education*, what would the **null** and **alternative** hypothesis be?

Use a **t-test** to determine if a beta is **significantly different** from 0

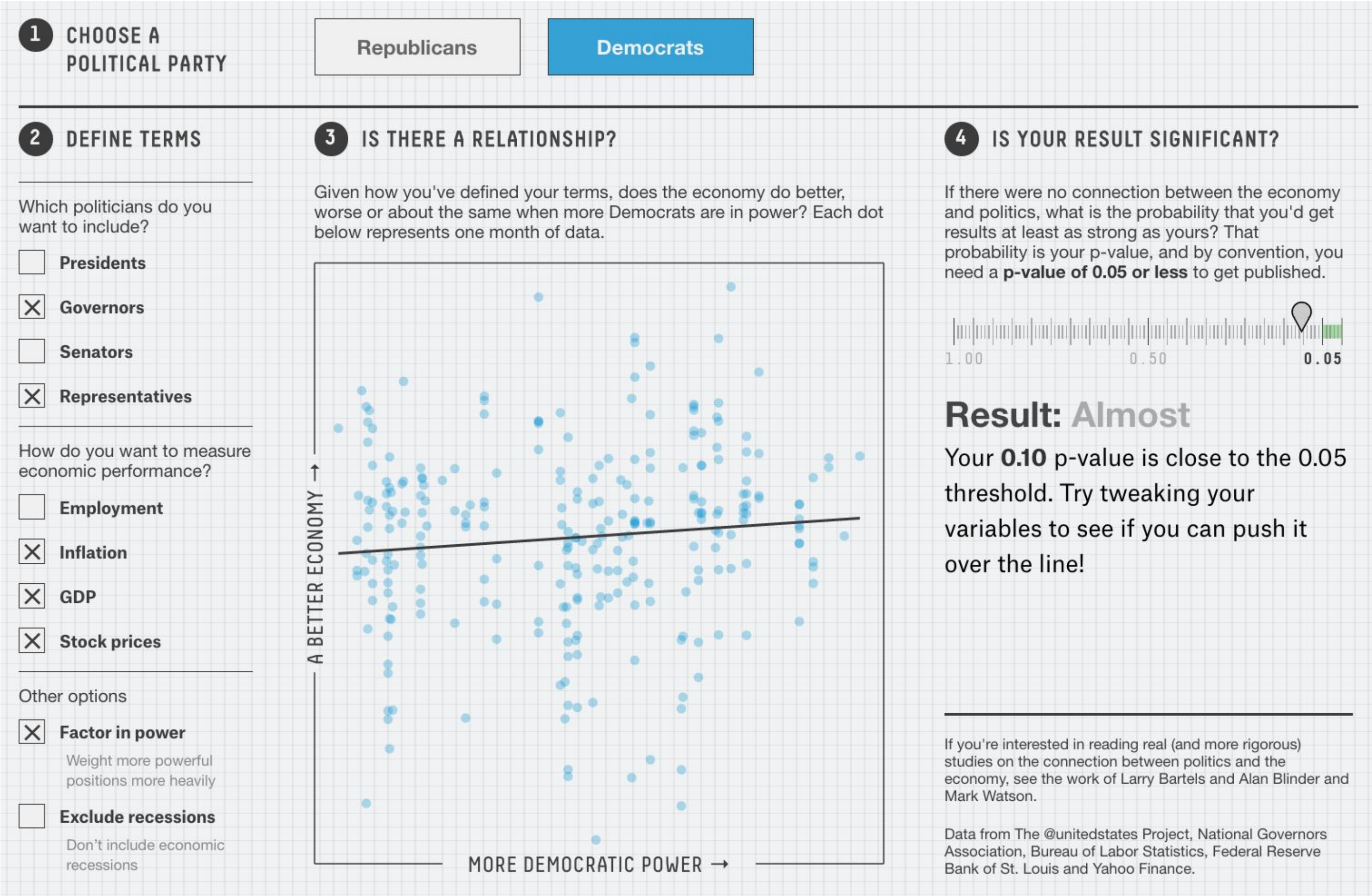
This depends on the standard error of the coefficient

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

We can then calculate the **p-value** as the probability of observing a t statistic (greater than or equal to **t**) if the actual value of B_1 *is 0*

The *p*-value is the probability of getting the observed value of the test statistic, or a value with even greater evidence against H_0, *if the null hypothesis is actually true*.

- (source)



P-hacking (link)

Assessing Model Accuracy: RSE

We can use the root standard error (RSE) as an estimate of the average amount of distance between the data and our fit

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

$$RSE = \sigma = \sqrt{RSS/(n-2)}$$

Note, this is in the **units of Y** (i.e., if RSE = $100, the average difference between our estimate and the data is 100)

This is a measure of **lack of fit**

Assessing Model Accuracy: R Squared

The r-squared statistic provides a measure of **goodness of fit**

It describes the variance in Y that is **explained by** variance in X

Measured as a **proportion**, between 0 and 1

$$Total\ Sum\ of\ Squares = TSS = (y_i - \overline{y})^2$$

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

$$R^2 = \frac{TSS - RSS}{TSS}$$

[class/linear-regression]

Code along!

```
# Set the upstream remote for the 'class' repo, if you haven't already
git remote add upstream https://github.com/info370a-w18/class.git

# Pull in the remote changes for the *master* branch
git checkout master # make sure you're on master
git pull upstream master

# Pull in the remote changes for the *complete* branch
git checkout complete # make sure you're on complete
git pull upstream complete

# Checkout the master branch to start doing the code-along
git checkout master
```

Notebook Set 4

nb-set-4

Upcoming...

r3-modeling due ***Tuesday before class***

Notebook4 due **Friday night**

This week

- Linear and Poisson Regression