4055801

Join at slido.com #4055801

ⓘ Click **Present with Slido** or install our [Chrome extension] to display joining instructions for participants while presenting.

4055801

LECTURE 19

Parameter Inference and the Bootstrap

Bias-Variance Tradeoff, regression coefficients

Data 100/Data 200, Spring 2025 @ UC Berkeley

Narges Norouzi and Josh Grossman

Content credit: Acknowledgments

Announcements

4055801

Thanks for your flexibility today ❤️ Not the ideal week to be sick!

Josh's office hours moved to **Mondays 5-7pm in Evans 421**. Come hang!

- Creating parallel universes
- Prediction versus Inference
- Regression inference
- Collinearity

Today's Roadmap

Lecture 19, Data 100 Spring 2025

Creating parallel universes

Lecture 19, Data 100 Spring 2025

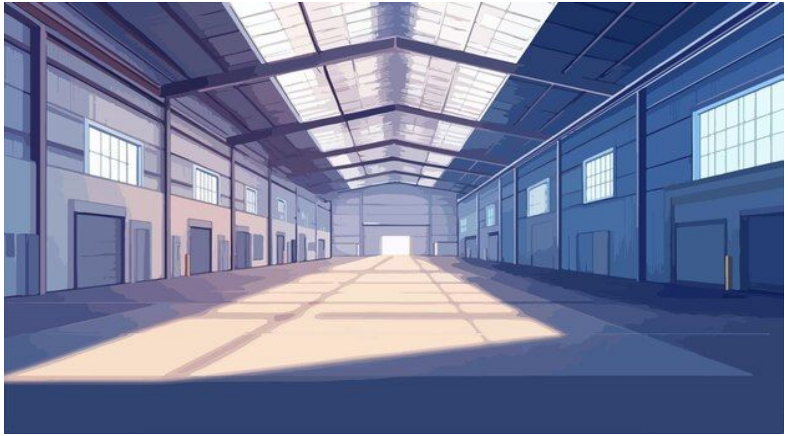
- **Creating parallel universes**
- Prediction versus Inference
- Regression inference
- Collinearity

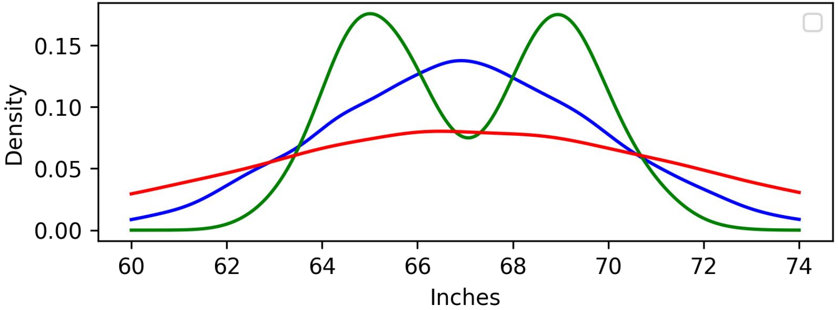
4055801

A familiar approach: Estimate the true mean with a sample mean

4055801

Warehouse with all **32,000 heights** of Berkeley undergrads on slips of paper (a **population**):





Possible distributions of the raw data

i.i.d. random sample of 100 heights:
$X_1, X_2, \ldots X_{100}$

Sample mean is 68.1 inches.

$\bar{X}_{100}$ = 68.1 inches

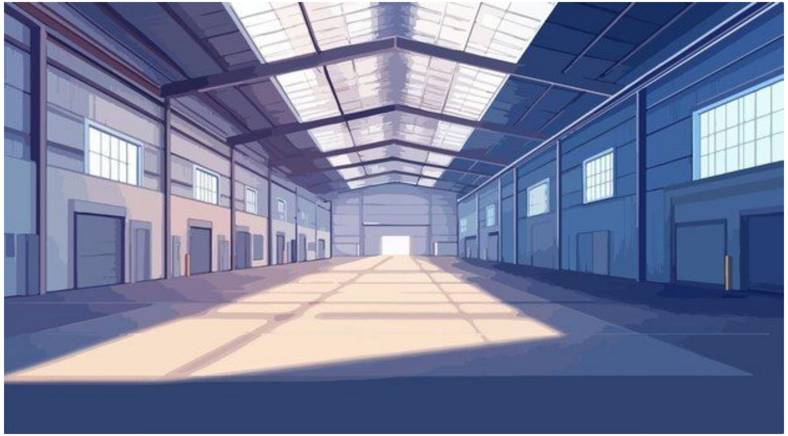
Our "best guess" for the population mean is 68.1 inches.

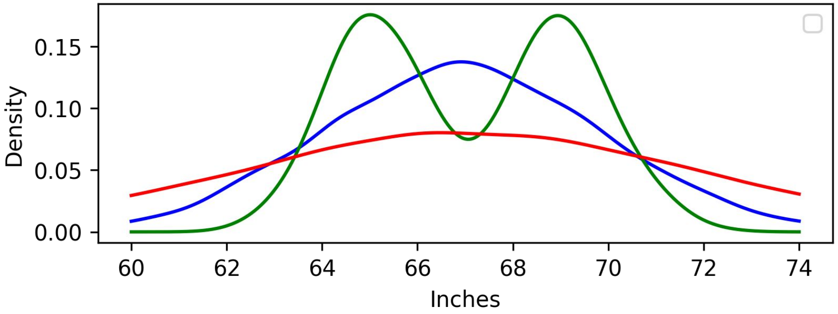
Harder Q: **How do we know if 68.1 inches is a "good" estimate?**

Thinking about a sample we could have observed

4055801

Warehouse with all **32,000 heights** of Berkeley undergrads on slips of paper (a **population**):





Possible distributions of the raw data

Our universe (Observed sample):

i.i.d. random sample of 100 heights:
$X_1, X_2, \ldots X_{100}$

Sample mean is 68.1 inches.

$\bar{X}_{100}$ = 68.1 inches

A parallel universe (An unobserved sample):

i.i.d. random sample of 100 heights:
$X_1, X_2, \ldots X_{100}$

Sample mean is **69.2** inches.

$\bar{X}_{100}$ = **69.2** inches

There are many possible samples we could have observed!

4055801

Warehouse with all **32,000 heights** of Berkeley undergrads on slips of paper (a **population**):

There are (effectively) infinite possible samples of size 100 we could have drawn! But, we observe **just one sample**.

$\bar{X}_{100,\text{Sample 1}}$ = 68.1 inches

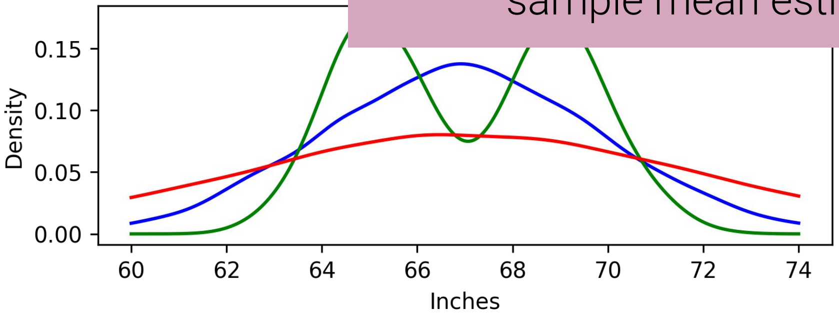
$\bar{X}_{100,\text{Sample 2}}$

$\bar{X}_{100,\text{Sample 3}}$

. . .

$\bar{X}_{100,\text{Sample }\infty}$

🐘 **Elephant in the room**: If we can't observe parallel universes, how can we say anything about the variability of our sample mean estimator?



Possible distributions of the raw data

Bootstrapping (Data 8) from a bag of M&Ms

4055801

Observed bag of M&Ms: Proportion primary-colored is 9/18

Randomly sample
with replacement

10,000 synthetic parallel universes!

Synthetic bag 1: Proportion primary-colored is 7/18

Synthetic bag 2: Proportion primary-colored is 11/18

• • •

Synthetic bag 10,000: Proportion primary-colored is 11/18

The Bootstrap (Data 8): Constructing synthetic parallel universes

4055801

Big assumption: The distribution of our sample data resembles the unknown population distribution.

Our i.i.d. random sample of 100 heights: $X_1, X_2, \ldots X_{100}$

We can't sample again from the population!

Randomly resample **with replacement** (i.e., allow duplicates)

Real "best guess"

$\bar{X}_{100,\text{Sample } 1}$ = 68.1 inches

$\bar{X}_{100,\text{Sample } 2}$ Synthetic "best guess"

$\bar{X}_{100,\text{Sample } 3}$ Synthetic "best guess"

. . .

$\bar{X}_{100,\text{Sample } \infty}$ Synthetic "best guess"

The Bootstrap (Data 8): We're not limited to the sample mean!

4055801

Big assumption: The distribution of our sample data resembles the unknown population distribution.

An i.i.d. random sample of size n: $X_1, X_2, \ldots X_n$

Randomly resample **with replacement** (i.e., allow duplicates)

Real "best guess"

$\hat{\theta}_{n,\text{Sample 1}}$

$\hat{\theta}_{n,\text{Sample 2}}$

Synthetic "best guess"

$\hat{\theta}_{n,\text{Sample 3}}$

Synthetic "best guess"

. . .

$\hat{\theta}_{n,\text{Sample } \infty}$

Synthetic "best guess"

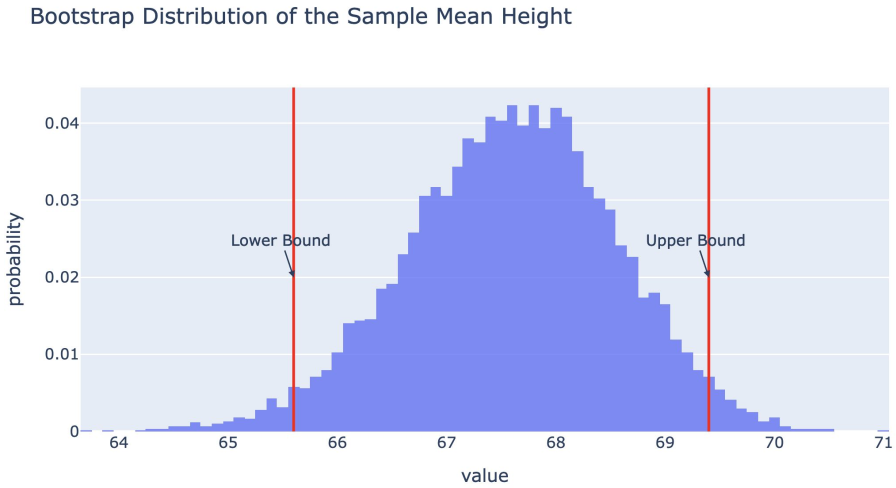
$\boldsymbol{\theta}$ is a property of the population distribution. For example, the true mean μ, the median, 75th percentile, …

$\hat{\theta}_n$ is an **estimator** of $\boldsymbol{\theta}$ calculated with a sample of size **n**.

For example, $\bar{X}_n$ is an estimator of μ.

Data 8 Review: Bootstrapping

How do we communicate the uncertainty of our best guess of the average height of UC Berkeley undergraduates?



Note: Lots of today's content is presented and explained thoroughly in the demos, so be sure to **work through the demos** in addition to the slides!

Demo

lec19.ipynb

In what situations might the bootstrap do a bad job of simulating the uncertainty around our best guess?

Bootstrap Limitations

4055801

The quality of our bootstrapped distribution depends on the quality of our original sample.

1. Works better for **large random samples**
   - A larger sample is more likely to resemble the population. At least 30, 50, 100… it depends.

2. Works poorly when the **population distribution is heavily skewed**
   - Imagine Bill Gates lived in Berkeley (i.e., heavily skewed income distribution)
   - The true average income in Berkeley is really high because of Bill.
   - But, if we take a random sample, we're unlikely to select Bill, so our bootstrapping procedure won't produce an accurate confidence interval for the mean.

3. Works poorly when **estimator is extreme** (i.e., the maximum or minimum)
   - For example, the true population max is guaranteed to be the same as your sample maximum or higher. No easy way to know about larger possible values.

Prediction versus Inference

Lecture 19, Data 100 Spring 2025

- Creating parallel universes
- **Prediction versus Inference**
- Regression inference
- Collinearity

4055801

Prediction has been our focus in Data 100

4055801

So far in Data 100, we have mostly thought about **prediction** problems.

In other words, we have focused on getting our predictions ($\hat{Y}_i$'s) **as close as possible** to the true outcomes (Y_i's).

We have spent less time on the **interpretation** of the relationships between X and Y, and **why** we decide to include or exclude certain features.

This difference illustrates the difference between **prediction** and **inference**.

Prediction and inference

4055801

Prediction Problems 🎯
Goal: Get Ŷ close to Y

How much will the stock market go up tomorrow?

Is this credit card charge fraudulent?

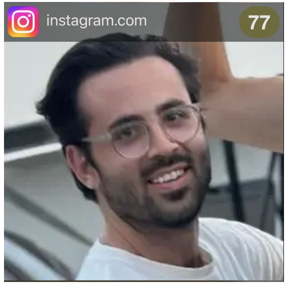
Can my phone accurately detect my face?

Inference Problems 🔬
Goal: How+why does X relate to Y?

What is the <u>effect</u> of getting a college degree on life outcomes?

What is the <u>effect</u> of a drug?

How does raising the minimum wage <u>affect</u> the unemployment rate?



An unsuccessful prediction attempt!

Overlap of prediction and inference

4055801

There can be overlap between prediction and inference problems.

For example, a credit card agency builds a model to accurately **predict** credit scores. A customer might want to know **why** their credit score is lower than expected.

Key Factor(s) Affecting Your FICO® Score:

1. **Lack of recent installment loan information**
   FICO® Scores consider recent non-mortgage installment loans (such as auto or student loans) information on a person's credit report. Your score was impacted because your credit report shows no recent non-mortgage installment loans or insufficient recent information about your loans.

2. **Too many accounts with balances**
   FICO® Scores consider the total number of accounts a consumer holds with balances, including credit card balance amounts that appear from the most recent account statements—even if that balance was paid off. Your score was impacted by having too many accounts with balances.

Is identifying the most likely next word in a sequence, akin to ChatGPT, primarily a prediction problem or an inference problem?

4055801

There is indeed research exploring **why** large-language models (LLMs) decide to choose one word or answer over another.

But, companies that produce LLMs are **primarily** interested in predicting the word that will satisfy you the most. Understanding why that word is selected is a **secondary** concern.

Keep in mind that understanding why a particular word is selected could allow you to make better predictions in the future. Prediction and inference complement each other!

Inference is closely associated with both correlation and causality

4055801

Correlational inference

Are homes with granite countertops worth more money?

Do people with college degrees have higher lifetime earning?

Are people who smoke more likely to get cancer?

Causal inference (harder!)

How much do granite countertops **raise** the value of a house?

Does getting a college degree **increase** lifetime earnings?

Does smoking **cause** cancer?

Causal questions are about the **effects** of **interventions**, not just passive observation.

Causal inference can be difficult or impossible!

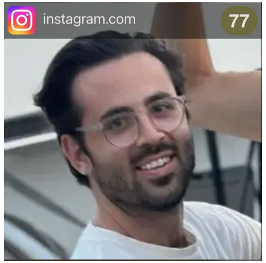
4055801

Correlation: Do people with college degrees have higher lifetime earnings?
Causation: Does receiving a college degree **increase** lifetime earnings?

In Data 8, you learned that **randomization** is required in order to infer causality.

It would be **unethical** to randomly assign students to college degrees, and hard to keep track of them over a lifetime. So, we're limited to the tools of **correlational analysis**.

Take Stat 156 to learn more about how we can infer causality **without (!)** randomized experiments.



Degree → Earnings w/ degree

No degree → Earnings w/o degree

Regression Inference

Lecture 19, Data 100 Spring 2025

- Creating parallel universes
- Prediction versus Inference
- **Regression inference**
- Collinearity

4055801

Correlational inference as regression

Suppose we have a dataset of **lifetime earnings** and **college degree** status.

The correlational relationship between earnings and degrees can be written as a regression:

$$\widehat{\text{Earnings}} = \hat{\theta}_0 + \hat{\theta}_1(\text{Has college degree})$$

$\hat{\theta}_0$: What are the predicted lifetime earnings for someone **without** a college degree?

$\hat{\theta}_1$: How much higher are predicted earnings for someone **with** a degree, **relative** to someone **without** a degree?

Is this strong evidence that getting a college degree increases lifetime earnings?

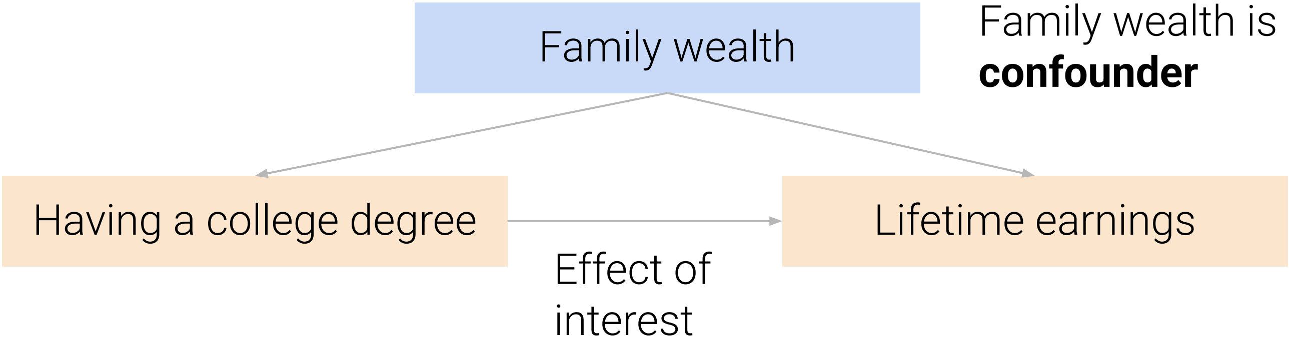
A regression does not guarantee your analysis is causal

$$\widehat{\text{Earnings}} = \hat{\theta}_0 + \hat{\theta}_1(\text{Has college degree})$$

Suppose we fit this OLS model to a randomly sampled dataset of degree holders and non-degree holders. As it turns out $\hat{\theta}_1$ >> 0.

Is this strong evidence that getting a college degree **increases** lifetime earnings?

No. People with college degrees may just be more likely to have other traits that increase lifetime earnings.



Family wealth is a **confounder**

It looks like wealth is a potential **confounder**. So, let's **adjust** for it in our OLS model:

$$\widehat{\text{Earnings}} = \hat{\theta}_0 + \hat{\theta}_1(\text{Has college degree}) + \hat{\theta}_2(\text{Family wealth in dollars})$$

$\hat{\theta}_0$: What are the predicted lifetime earnings for someone without a college degree **and with zero family wealth**?

$\hat{\theta}_1$: How much higher are predicted earnings for a degree-holder, relative to someone without a degree, **holding family wealth constant**?

$\hat{\theta}_2$: Slido!

What is the interpretation of θ2_hat here?

Interpreting a new coefficient

$$\widehat{\text{Earnings}} = \hat{\theta}_0 + \hat{\theta}_1(\text{Has college degree}) + \hat{\theta}_2(\text{Family wealth in dollars})$$

$\hat{\theta}_2$: How much higher are predicted lifetime earnings for each additional dollar of family wealth, assuming we're **comparing two people with the same degree status**?

$$\widehat{\text{Earnings}}_{\text{Degree, \$1000}} =$$

$$\widehat{\text{Earnings}}_{\text{Degree, \$1001}} =$$

$$\widehat{\text{Earnings}}_{\text{No Degree, \$1000}} =$$

$$\widehat{\text{Earnings}}_{\text{No Degree, \$1001}} =$$

Interpreting a new coefficient

4055801

$$\widehat{\text{Earnings}} = \hat{\theta}_0 + \hat{\theta}_1(\text{Has college degree}) + \hat{\theta}_2(\text{Family wealth in dollars})$$

$\hat{\theta}_2$: How much higher are predicted lifetime earnings for each additional dollar of family wealth, assuming we're **comparing two people with the same degree status**?

$$\widehat{\text{Earnings}}_{\text{ Degree, \$1000}} = \hat{\theta}_0 + \hat{\theta}_1 + 1000 * \hat{\theta}_2$$

Difference is $\hat{\theta}_2$

$$\widehat{\text{Earnings}}_{\text{ Degree, \$1001}} = \hat{\theta}_0 + \hat{\theta}_1 + 1001 * \hat{\theta}_2$$

$$\widehat{\text{Earnings}}_{\text{ No Degree, \$1000}} = \hat{\theta}_0 + 1000 * \hat{\theta}_2$$

Difference is also $\hat{\theta}_2$

$$\widehat{\text{Earnings}}_{\text{ No Degree, \$1001}} = \hat{\theta}_0 + 1001 * \hat{\theta}_2$$

Do we now have strong evidence that getting a college degree increases lifetime earnings?

Confounders are lurking everywhere

4055801

$$\widehat{\text{Earnings}} = \hat{\theta}_0 + \hat{\theta}_1(\text{Has college degree}) + \hat{\theta}_2(\text{Family wealth in dollars})$$

We fit this OLS model. As it turns out $\hat{\theta}_1$ >> 0. Do we now have strong evidence that getting a college degree increases lifetime earnings? **No.**

There could be other **observed** confounders, like health, demographics, and geography.

There could also be **unobserved** confounders, like intrinsic motivation and values.

We cannot be certain we've isolated the causal effect!

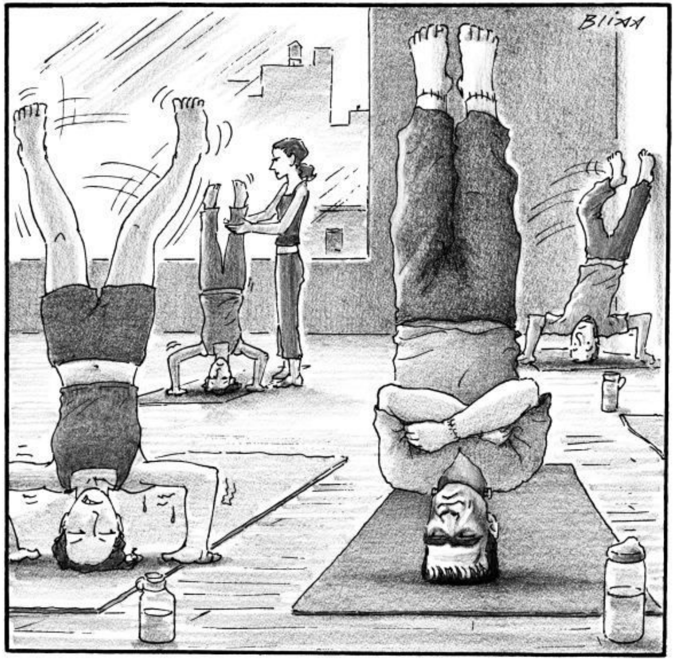
Common assumption: All confounders are observed and adjusted for (**ignorability**).

This assumption implies a causal relationship, but the assumption cannot be verified!

Important to be **transparent** about how to interpret your model.

2-minute stretch break!

Lecture 19, Data 100 Spring 2025



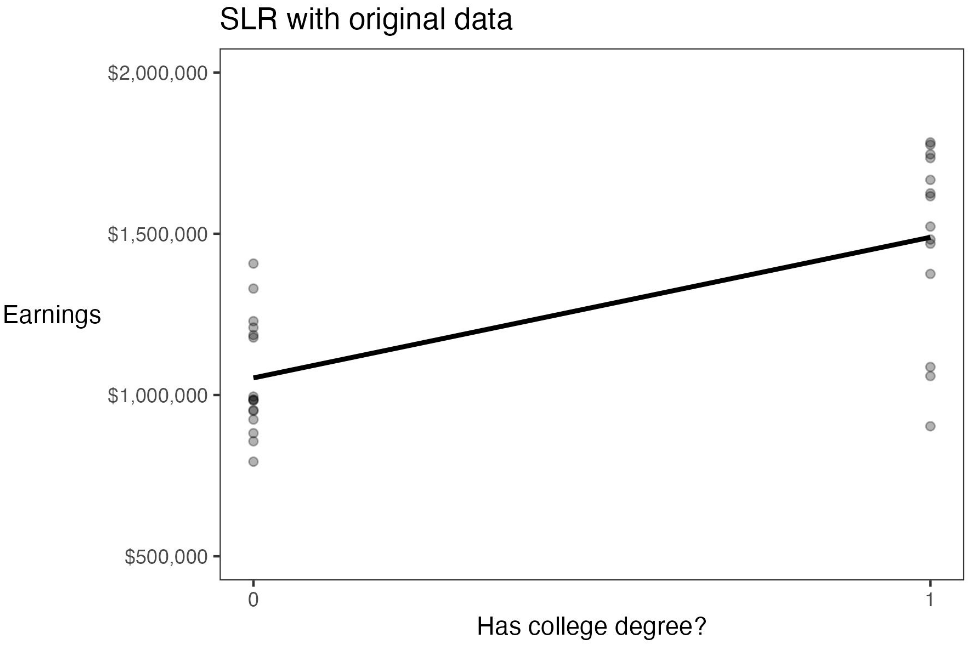
4055801

Correlational inference as regression

Let's return to our original regression:

$$\widehat{\text{Earnings}} = \hat{\theta}_0 + \hat{\theta}_1(\text{Has college degree})$$

$\hat{\theta}_1$ is our "best guess" of the association between earnings and degrees, without adjusting for any other variable.



How could we measure uncertainty in $\hat{\theta}_1$?

Once again, the **bootstrap**!

Correlational inference as regression

4055801

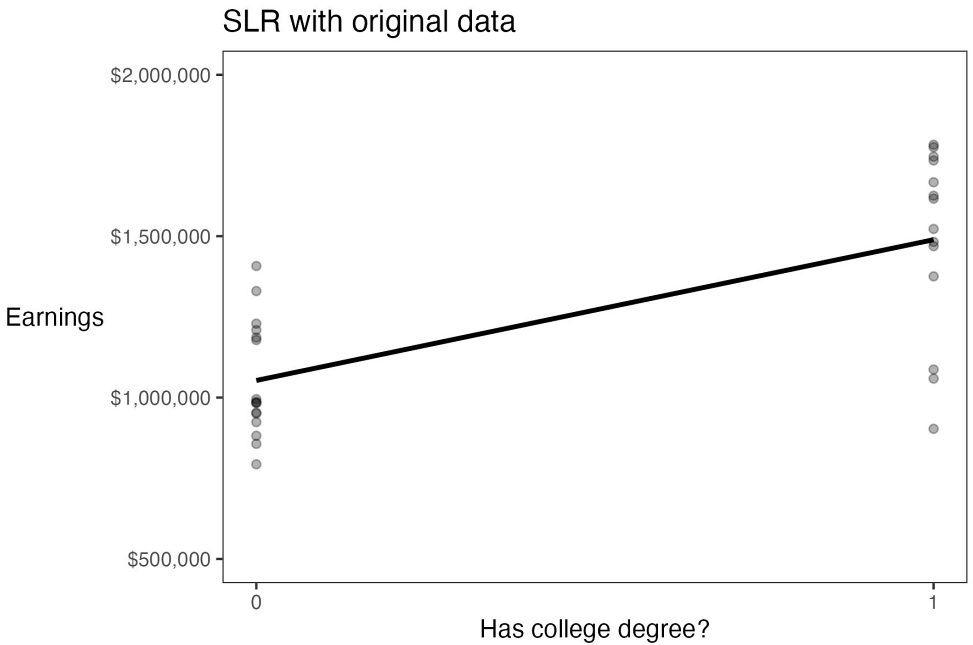
Let's return to our original regression:

$$\widehat{\text{Earnings}} = \hat{\theta}_0 + \hat{\theta}_1(\text{Has college degree})$$

$\hat{\theta}_1$ is our "best guess" of the association between earnings and degrees, without adjusting for any other variable.

How could we measure uncertainty in $\hat{\theta}_1$?

Once again, the **bootstrap**!

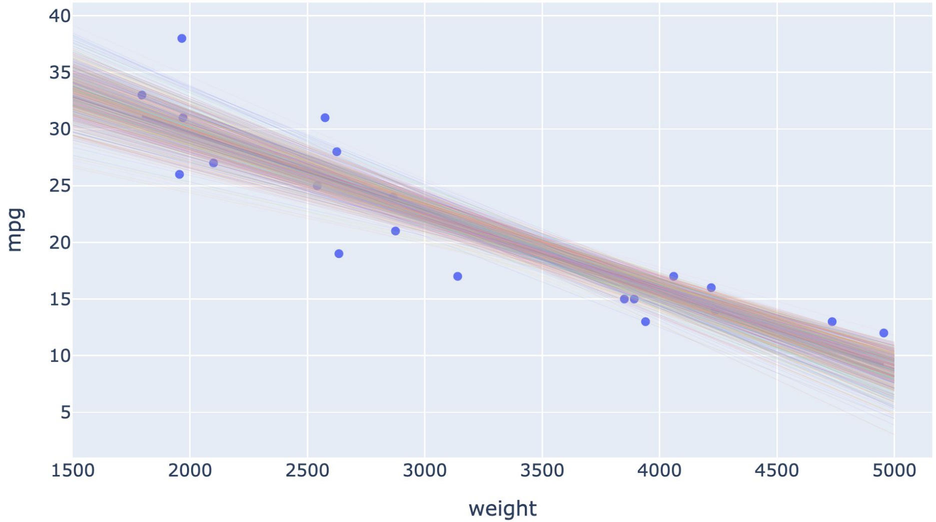


The image above is a GIF. Be sure to view in slideshow mode!

Bootstrapping Regression Coefficients

4055801

How do we create confidence intervals (CIs) for regression coefficients? How do we conduct hypothesis tests with regression coefficients?



<u>Note</u>: Lots of today's content is presented and explained thoroughly in the demos, so be sure to **work through the demos** in addition to the slides!

Demo

lec19.ipynb

Collinearity

Lecture 19, Data 100 Spring 2025

- Creating parallel universes
- Prediction versus Inference
- Regression inference
- **Collinearity**

4055801

OLS inference requires careful thought about feature selection

Adding new terms to a regularized model often improves **predictive** performance.

So, as long as we're at it, let's add another variable to our lifetime earnings regression:

$$\widehat{\text{Earnings}} = \hat{\theta}_0 + \hat{\theta}_1(\text{Has college degree}) + \hat{\theta}_2(\text{Family wealth in dollars}) + \underbrace{\hat{\theta}_3(\text{Has lived in a college town})}$$

Is adding this term appropriate if we want to make **inferences** about $\hat{\theta}_1$?

Probably not.

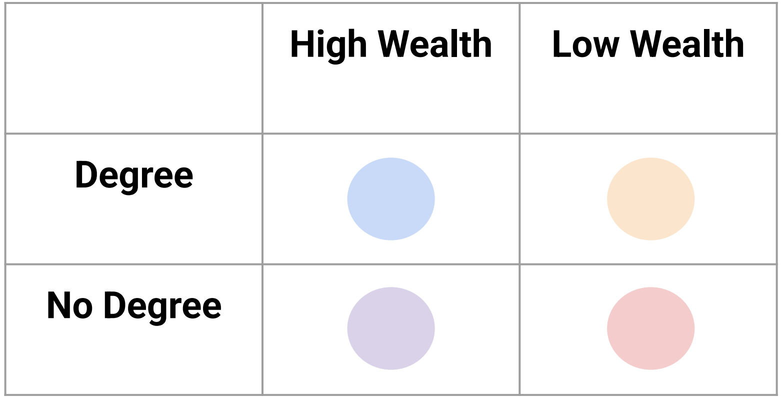
Beware of strong multicollinearity when conducting inference

4055801

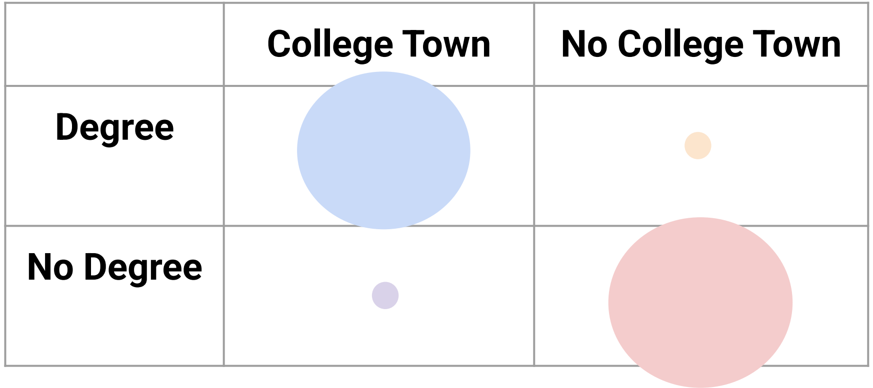
$$\widehat{\text{Earnings}} = \hat{\theta}_0 + \hat{\theta}_1(\text{Has college degree}) + \hat{\theta}_2(\text{Family wealth in dollars}) + \hat{\theta}_3(\text{Has lived in a college town})$$

Having a college degree and living in a college town are **very highly correlated** features. To the regression model, these features look very similar!

The regression model does not know that a college degree is more likely to change earnings than living in a college town. This is **causal domain knowledge**.



Plenty of individuals in each group for comparison.



Comparison is very sensitive to the training data.

Beware of strong multicollinearity when conducting inference

4055801

$$\widehat{\text{Earnings}} = \hat{\theta}_0 + \hat{\theta}_1(\text{Has college degree}) + \hat{\theta}_2(\text{Family wealth in dollars}) + \hat{\theta}_3(\text{Has lived in a college town})$$

The fitting process of OLS minimizes RMSE (i.e., it maximizes predictive performance).

Because degree-status and living in a college town tend to have the **same** value, we could **increase** $\hat{\theta}_1$ by \$X and **decrease** $\hat{\theta}_3$ by \$X without changing the RMSE all that much.

So, the $\hat{\theta}_1$ and $\hat{\theta}_3$ coefficients will be sensitive to the training data (i.e., **high variance**).

This high variance does not harm **predictive** performance, but it does harm the **validity** of $\hat{\theta}_1$ as a measure of the association between college degrees and lifetime earnings.

<u>Lesson</u>: If you want to make inferences about a parameter, don't include features that are highly correlated with that parameter's feature.

4055801

We made it!

Lecture 19, Data 100 Spring 2025

- Creating parallel universes
- Prediction versus Inference
- Regression inference
- Collinearity

4055801

LECTURE 19

Parameter Inference and the Bootstrap

Data 100/Data 200, Spring 2025 @ UC Berkeley

Narges Norouzi and Josh Grossman

Content credit: Acknowledgments

Bonus demos

How accurate are air quality measurements?

4055801

SF Bay Bridge, 09/2020

Demo

Data 100 textbook (Ch12, Ch 17)

Two common sources of air quality information:

Air Quality System (AQS):

- (+) High-quality, well-calibrated, publicly available, government-run. Gold standard for accuracy
- (–) Expensive (~$15k-40k) and far apart.
- (–) Hourly/delayed reports because of extensive calibration

PurpleAir sensors (link)

- (+) Cheap (~$250), can be installed at home for personal use
- (+) Measurements every two minutes, denser coverage
- (–) Less accurate than AQS (see Josh Hug's post)

How do we use nearby AQS sensor measurements to improve PurpleAir measurements?

Focus on PM2.5 particles (particles < 2.5μm)

Demo

Data 100 textbook (Ch12, Ch 17)

Calibration Model

4055801

Goal: Create a model that predicts PM2.5 readings as accurately as possible.

- Build a model that adjusts PurpleAir (PA) measurements based on nearby **AQS measurements** (AQS, true air quality).

$$PA \approx \theta_0 + \theta_1 AQS$$

- Then, invert model to predict **true air quality** from PA measurements.

$$\text{True Air Quality} \approx -\frac{\theta_0}{\theta_1} + \frac{1}{\theta_1} PA$$

Side note: Why perform this “inverse regression”?

- Intuitively, AQS measurements are “true” and have no error.
- **A linear model takes a “true” x value input and minimizes the error in the y direction.**
- Algebraically identical, but **statistically different**.

Calibration Model

4055801

Focus on original linear model (instead of algebraic step 2):

1. Build a model that adjusts PurpleAir (PA) measurements based on nearby **AQS measurements** (AQS, true air quality).

$$PA \approx \theta_0 + \theta_1 AQS$$

2. Karoline Barkjohn, Brett Gannt, and Andrea Clements from the US Environmental Protection Agency developed a model to improve the PuprleAir measurements from the AQS sensor measurements by incorporating Relative Humidity:

$$\mathrm{PA} \approx \theta_0 + \theta_1 \mathrm{AQS} + \theta_2 \mathrm{RH}$$

Barkjohn and group's work is now used in the official US government maps, like the AirNow Fire and Smoke map, includes both AQS and PurpleAir sensors, and applies Barkjohn's correction to the PurpleAir data.

Demo

The Snowy Plover

4055801

Data on the tiny Snowy Plover bird was collected by a former Berkeley student at the Point Reyes National Seashore.

The bigger a newly hatched chick, the more likely it is to survive.

Demo

Highly collinear!

$$\widehat{\text{Newborn weight}} = \hat{\theta}_0 + \hat{\theta}_1 \text{egg\_weight} + \hat{\theta}_2 \text{egg\_length} + \hat{\theta}_2 \text{egg\_breadth}$$