LECTURE 32

# Inference in Multiple Linear Regression

Regression coefficients

**Data CSCI 3022**

Content credit: Acknowledgments

# Announcements

- Yesterday's TA notebook session and video posted on Canvas:  Nb 12 (Modeling and Analyzing Covid Cases)

- Project Part 1 due tomorrow Thursday at 11:59pm MT  (plan accordingly - no late submissions accepted)

- Last quiz (quiz 10) Friday (Scope:  HW 11, Lessons 26-28)

- Project Part 2 released later tonight -- due next Thursday:
  - Since we've had a delay in releasing Project 2 I've switched one of the questions to be extra credit!

# Three Reasons for Building Models

## Reason 1:

Previous lecture, and Project Part 2

To make **accurate predictions** about unseen data.

- Can we predict if an email is spam or not?
- Can we generate a one-sentence summary of this 10-page long article?

Other times, we care more about making extremely accurate predictions, at the cost of having an uninterpretable model. These are sometimes called **black-box models**, and are common in fields like deep learning.

## Reason 2:

Today!

To explain **complex phenomena** occurring in the world we live in.

- How are the parents' average heights related to the children's average heights?
- How do an object’s velocity and acceleration impact how far it travels?

Often times, we care about creating models that are **simple and interpretable**, allowing us to understand what the relationships between our variables are.

## Reason 3:

To make **causal inferences** about if one thing causes another thing.

- Can we conclude that smoking *causes* lung cancer?
- Does a job training program cause increases in employment and wage?

Much harder question because most statistical tools are designed to infer association not causation

This won’t be the focus of this class, but will be if you go on to take more advanced classes

Most of the time, we want to strike a balance between **interpretability** and **accuracy**.

# Model Notation Review

**Model Notation Review**

Interpreting Regression Coefficients

Bootstrapping Test for a Regression Coefficient

Collinearity

Correlation vs. Causation

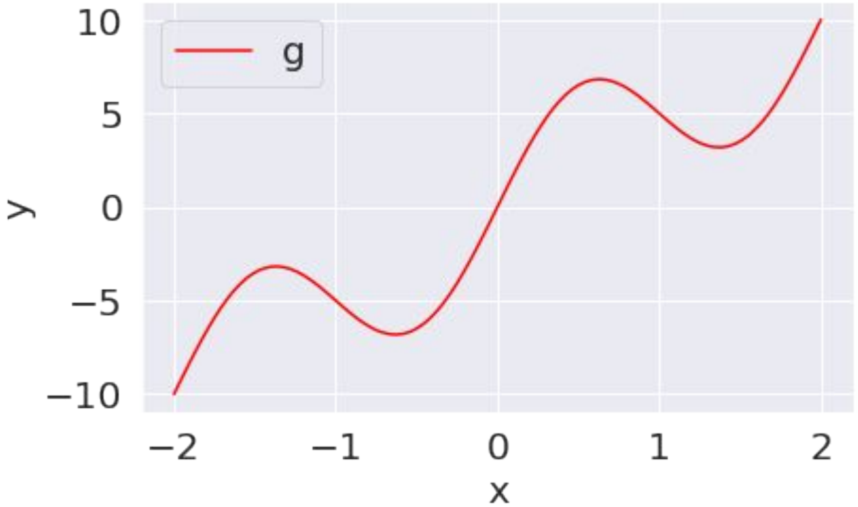
[Extra] Review of the Bootstrap

[Extra] Derivation of Bias-Variance Decomposition

# Modeling: Estimating a Relationship

What if we wanted to estimate the relationship between input $x$ and random response $Y$ ?

$$Y = g(x) + \epsilon$$



We would like to find the true relationship $g$.

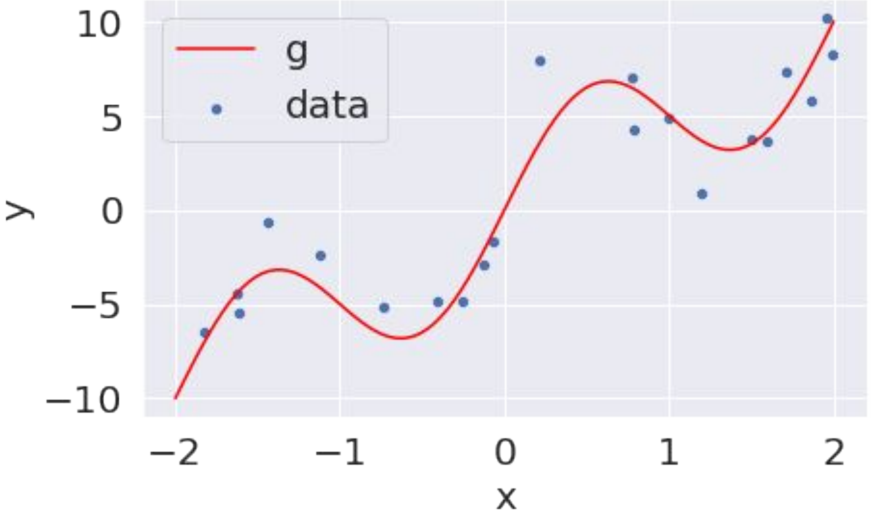
Each individual in the population has:

- **Fixed features** $x$ , and hence fixed $g(x)$.

# Modeling: Estimating a Relationship

What if we wanted to estimate the relationship between input $x$ and random response $Y$ ?

$$Y = g(x) + \epsilon$$



We would like to find the true relationship $g$.

Each individual in the population has:

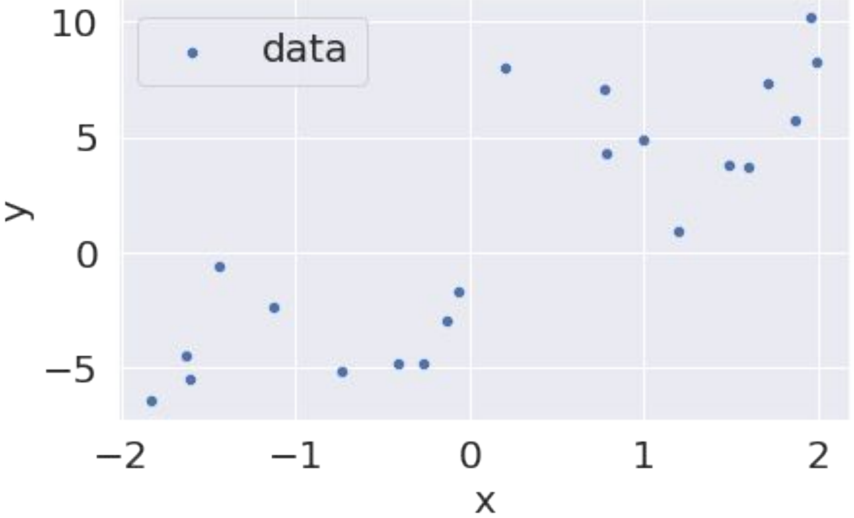
- **Fixed features** $x$ , and hence fixed $g(x)$.
- Random **error/noise** $\epsilon$
- Random **observation/response** $Y = g(x) + \epsilon$

Errors $\epsilon$ are assumed expectation 0 (“zero mean”) and i.i.d. across individuals.

# Modeling: Estimating a Relationship

What if we wanted to estimate the relationship between input $x$ and random response $Y$ ?

$$Y = g(x) + \epsilon$$



We would like to find the true relationship $g$.

Each individual in the population has:

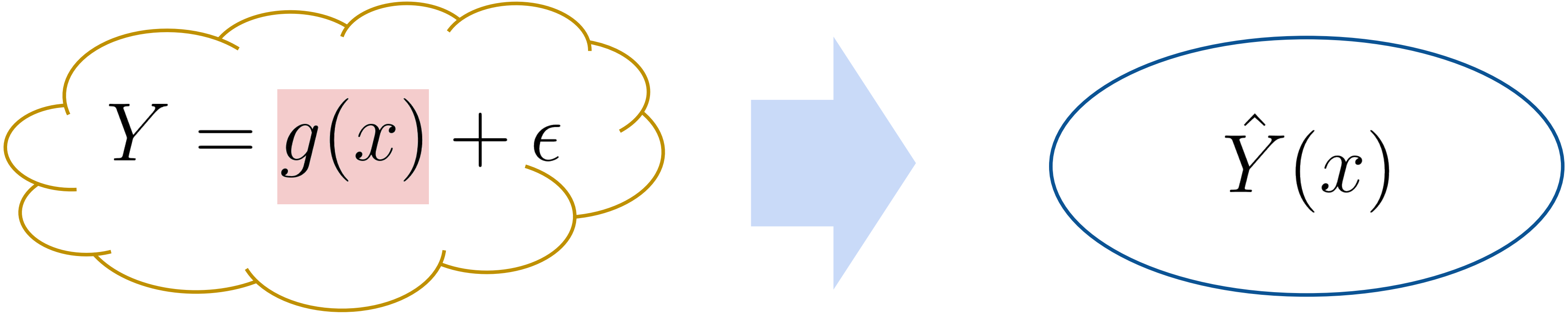
- **Fixed features** $x$ , and hence fixed $g(x)$.
- Random **error/noise** $\epsilon$
- Random **observation/response** $Y = g(x) + \epsilon$

Errors $\epsilon$ are assumed expectation 0 (“zero mean”) and i.i.d. across individuals

We only can only observe our random sample. From this we’d like to estimate the true relationship $g$.

# Modeling: Estimating a Relationship

What if we wanted to estimate the relationship between input $x$ and random response $Y$ ?



$$Y = g(x) + \epsilon$$

$$\hat{Y}(x)$$

We would like to find the true relationship $g$.

Each individual in the population has:

- **Fixed features** $x$ , and hence fixed $g(x)$.
- Random **error/noise** $\epsilon$
- Random **observation/response** $Y = g(x) + \epsilon$

Errors $\epsilon$ are assumed expectation 0 (“zero mean”) and i.i.d. across individuals

We build a **model** for predictions based based on our observed sample of $(x, y)$ pairs. Our model **estimates** the true relationship $g$.

At every $x$, our **prediction** for Y is $\hat{Y}(x)$.

# Interpreting Regression Coefficients

Review

**Interpreting Regression Coefficients**

Bootstrapping Test for a Regression Coefficient

Collinearity

Correlation vs. Causation

[Extra] Review of the Bootstrap

[Extra] Derivation of Bias-Variance Decomposition

# Inference for Linear Regression

Assume the true relationship is linear:

$$f_{\theta}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \cdots + \theta_p x_p + \epsilon$$

Unknown true parameters $\theta$

Our estimation from our sample (design matrix $\mathbb{X}$, response vector $\mathbb{Y}$):

$$f_{\hat{\theta}}(x) = \hat{\theta}_0 + \hat{\theta}_1 x_1 + \hat{\theta}_2 x_2 \cdots + \hat{\theta}_p x_p$$

Estimated parameters $\hat{\theta}$

The meaning of "slope":

1. What if the true parameter $\theta_1$ is 0?
2. Can we figure out whether it is positive or negative?
3. What does the parameter $\theta_1$ even mean?

What can we **infer** about our true parameter given our estimate $\hat{\theta}_1$?

# Inferring Zero slope

Our **estimation** from our sample (design matrix $\mathbb{X}$ , response vector $\mathbb{Y}$ ):

$$f_{\hat{\theta}}(x) = \hat{\theta}_0 + \hat{\theta}_1 x_1 + \hat{\theta}_2 x_2 \cdots + \hat{\theta}_p x_p$$

This estimate $\hat{\theta}_1$ for the true parameter $\theta_1$ depends our sample.

What if the true parameter $\theta_1$ was 0?

Then the feature $x_1$ has **no effect** on the response!

# Approximate confidence interval for true slope

How do we test if the true parameter $\theta_1$ was 0?

- We get one estimate $\hat{\theta}_1$ from our sample of size $n$.
- But we must imagine all the other random samples that could have happened, and draw our conclusion based on this distribution of estimates.

Enter **hypothesis testing**!

**Null hypothesis**: The true parameter $\theta_1$ is 0.

Alternative hypothesis: The true parameter $\theta_1$ is not 0.

If your p-value is small, reject the null hypothesis at the cutoff level (say, 5%)

Ruling out 0 almost always means determining the sign of $\theta_1$

Equivalently (duality argument):

- Compute an approximate 95% confidence interval with **bootstrapping**.
- If the interval does not contain 0, reject the null hypothesis at the 5% level.
- Otherwise, data are consistent with null hypothesis (the true parameter *could* be 0).

# Bootstrapping Test for a Regression Coefficient

Review

The Bias-Variance Tradeoff

Interpreting Regression Coefficients

**Bootstrapping Test for a Regression Coefficient**

Collinearity

Correlation vs. Causation

[Extra] Review of the Bootstrap

[Extra] Derivation of Bias-Variance Decomposition

## Demo

# Recall from HW 11: The Snowy Plover

Data on the tiny Snowy Plover bird was collected at the Point Reyes National Seashore.

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

Assumed true relationship for newborn weight $Y = f_\theta(x)$:

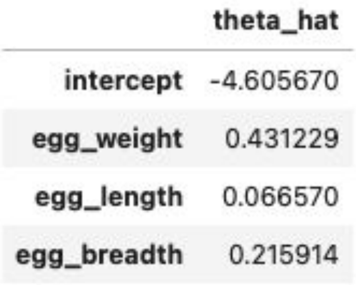
$$f_\theta(x) = \theta_0 + \theta_1 \text{egg\_weight} + \theta_2 \text{egg\_length} + \theta_3 \text{egg\_breadth} + \epsilon$$

# Estimating the Snowy Plover

Assumed true relationship for newborn weight $Y = f_\theta(x)$:

$$f_\theta(x) = \theta_0 + \theta_1 \text{egg\_weight} + \theta_2 \text{egg\_length} + \theta_3 \text{egg\_breadth} + \epsilon$$

Estimated model for newborn weight $\hat{Y} = f_{\hat{\theta}}(x)$:



| | | theta_hat |
|---|---|---|
| $\hat{\theta}_0$ | **intercept** | -4.605670 |
| $\hat{\theta}_1$ | **egg_weight** | 0.431229 |
| $\hat{\theta}_2$ | **egg_length** | 0.066570 |
| $\hat{\theta}_3$ | **egg_breadth** | 0.215914 |

## Demo

Is this the right linear model for newborn weight?

Let's test the **null hypothesis**: The true parameter $\theta_1$ is 0.

## Demo

# Bootstrapped Confidence Interval for $\theta_1$

We can estimate the distribution of $\hat{\theta}_1$ by bootstrapping.

Bootstrap the sample to build an **approximate 95% confidence interval** for the parameter $\theta_1$:

```
sample_df = … # call this the bootstrap population
n = len(sample_df)
estimates = []
repeat 10000 times:
    # resample ... ? times with replacement
    resample = ...
    ...
    estimate = ...
    estimates.append(estimate)
lower = np.percentile(estimates, ...)
upper = np.percentile(estimates, ...)
conf_interval = (lower, upper)
```

**1.** (Bootstrap review) Why must we resample **with replacement**?
**2.** What goes in the blanks?

Demo

## Bootstrapped Confidence Interval for $\theta_1$

We can estimate the distribution of $\hat{\theta}_1$ by bootstrapping.

Bootstrap the sample to build an **approximate 95% confidence interval** for the parameter $\theta_1$:

```
sample_df = … # call this the bootstrap population
n = len(sample_df)
estimates = []
repeat 10000 times:
    # resample n times with replacement
    resample = sample_df.sample(n, replace=True)
    ... # fit the new model to the new resampled X, y
    estimate = get_theta_hat1(model)
    estimates.append(estimate)
lower = np.percentile(estimates, 2.5)
upper = np.percentile(estimates, 97.5)
conf_interval = (lower, upper)
```

## Bootstrapped Confidence Interval for $\theta_1$

We can estimate the distribution of $\hat{\theta}_1$ by bootstrapping.

Bootstrap the sample to build an **approximate 95% confidence interval** for the parameter $\theta_1$:

Our bootstrapped 95% confidence interval for the true $\theta_1$:

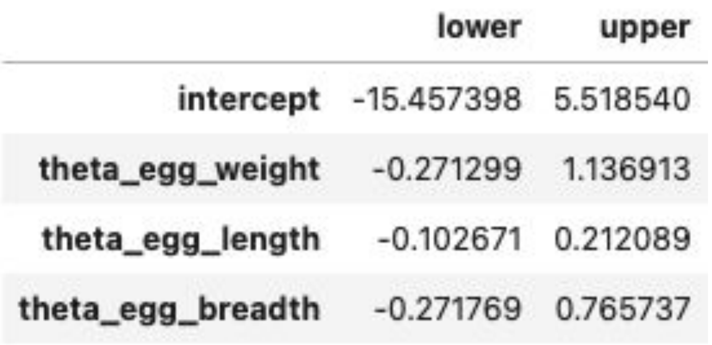
```
(-0.262, 1.115)
```

We cannot reject the null hypothesis at cutoff 5% (our true parameter $\theta_1$ could be 0).

**Demo**

# Are all of our true parameters 0?

Let's bootstrap 95% confidence intervals for all our parameters:



| True param | | lower | upper |
|---|---|---|---|
| $\theta_0$ | **intercept** | -15.457398 | 5.518540 |
| $\theta_1$ | **theta_egg_weight** | -0.271299 | 1.136913 |
| $\theta_2$ | **theta_egg_length** | -0.102671 | 0.212089 |
| $\theta_3$ | **theta_egg_breadth** | -0.271769 | 0.765737 |

**Wait….something's off here!**

## Demo

# Collinearity

Review

Interpreting Regression Coefficients

Bootstrapping Test for a Regression Coefficient

**Collinearity**

Correlation vs. Causation

# Collinearity

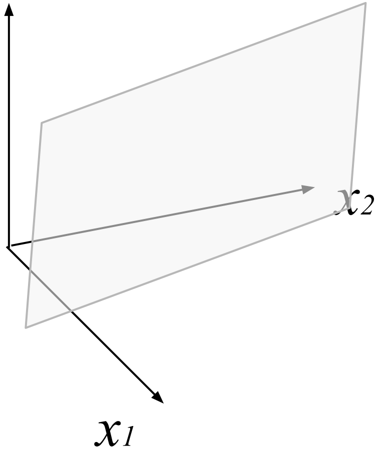
Our estimation from our sample (design matrix $\mathbb{X}$, response vector $\mathbb{Y}$):

$$f_{\hat{\theta}}(x) = \hat{\theta}_0 + \hat{\theta}_1 x_1 + \hat{\theta}_2 x_2 \cdots + \hat{\theta}_p x_p$$

The **slope** $\hat{\theta}_p$ measures the change in $y$ per unit change in $x_p$, **provided all the other variables are held constant**.

$$\text{predicted weight} = a_0 + a_1 \cdot \text{length} + a_2 \cdot \text{sleep}$$

If two cats have a 1 inch height difference **and the same hours of sleep**, their estimated weight difference is $a_1$.



If variables are **related** to each other, then **interpretation fails!**
E.g., if a change in length always came with a change in sleep

# Collinearity

If features are related to each other, it might not be possible to have a change in one of them **while holding the others constant**.

- **Example:** we can't change only one column of a one-hot encoding
- Then the individual slopes are more difficult to interpret

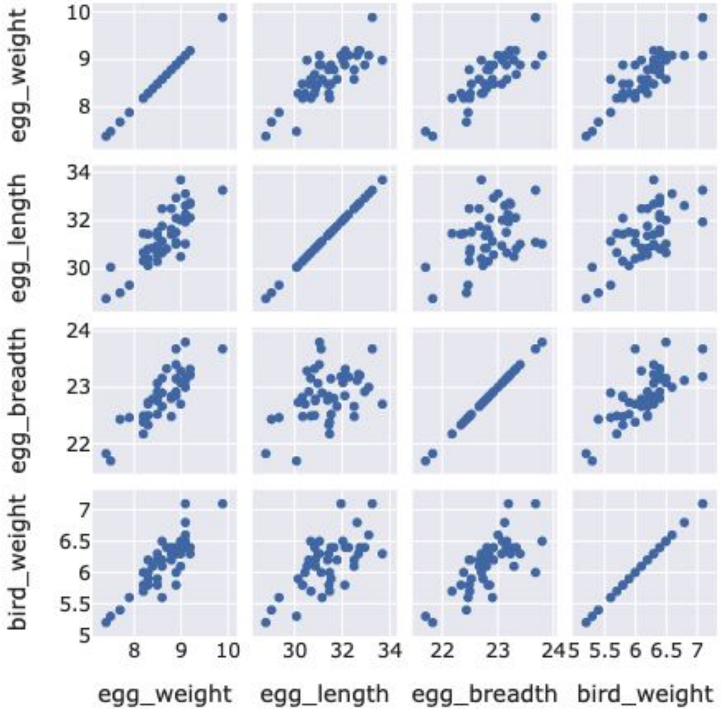
**Collinearity**: When a feature can be predicted pretty accurately by a **linear** function of the others, i.e., the feature is highly correlated with the others.

- Slopes are hard to interpret
- $\mathbb{X}^T\mathbb{X}$ might not be invertible, i.e., solution might not be uniquely determined
- Small changes in the data sample can lead to big changes in the estimated slopes
- Also known as **multicollinearity**

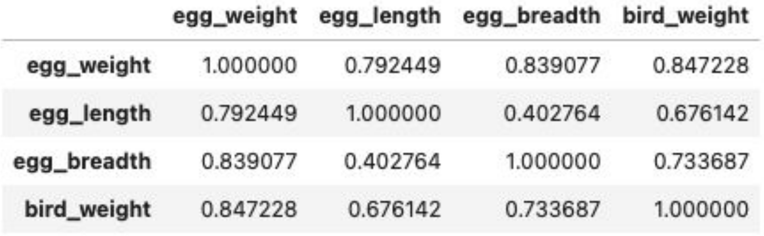
**Why?** Suppose p = 3, and $X_1 \approx X_2 + X_3$. What if we **increase** $\theta_1$ by 10 and also **decrease** $\theta_2, \theta_3$ by 10?

- Predictions hardly change at all (but coefficients changed a lot!)
- Means there are very dissimilar models that are nearly indistinguishable from the data

# Cross-wise comparison of egg features



```
px.scatter_
    matrix(eggs)
```



| | egg_weight | egg_length | egg_breadth | bird_weight |
|---|---|---|---|---|
| **egg_weight** | 1.000000 | 0.792449 | 0.839077 | 0.847228 |
| **egg_length** | 0.792449 | 1.000000 | 0.402764 | 0.676142 |
| **egg_breadth** | 0.839077 | 0.402764 | 1.000000 | 0.733687 |
| **bird_weight** | 0.847228 | 0.676142 | 0.733687 | 1.000000 |

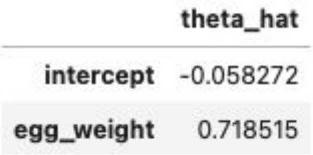
```
eggs.corr()
```

## Demo

## A more interpretable model

If we instead assume a true relationship using only egg weight:

$$f_\theta(x) = \theta_0 + \theta_1 \text{egg\_weight} + \epsilon$$



|  |  | theta_hat |
|---|---|---|
| $\hat{\theta}_0$ | **intercept** | -0.058272 |
| $\hat{\theta}_1$ | **egg_weight** | 0.718515 |

This model performs almost as well as our other model (RMSE 0.0464, old RMSE 0.0454), and the confidence interval for the true parameter $\theta_1$ doesn't contain zero

You found CI this on HW 11!

```
(0.604, 0.819)
```

In retrospect, it's no surprise that the weight of an egg best predicts the weight of a newly-hatched chick.

# Demo

A model with **highly correlated variables** prevents us from interpreting how the variables are related to the prediction.

## Reminder: Assumptions matter

Keep the following in mind:

- All inference assumes that the regression model holds.
- If the model doesn't hold, the inference might not be valid.
- If the assumptions of the bootstrap don't hold, i.e.
  - Sample size n is large
  - Sample is representative of population distribution (drawn IID, unbiased)

  …then the results of the bootstrap might not be valid.

# How accurate are air quality measurements?

SF Bay Bridge, 09/2020

## Demo

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)

# Calibration Model

**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.

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

## Demo

Textbook (Ch12, Ch 17)

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**.

## Demo

# Calibration Model

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 PurpleAir measurements from the AQS sensor measurements by incorporating Relative Humidity:

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

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.

# Recap: Correlation vs. Causation

Review

Interpreting Regression Coefficients

Bootstrapping Test for a Regression Coefficient

Collinearity

**Correlation vs. Causation**

# Correlation and Causation

What does $\theta_j$ mean in our regression?

- Holding other variables fixed, how much should our **prediction** change with $X_j$?
- For simple linear regression, this boils down to the **correlation coefficient**:
  - Does having more x predict more y (and by how much)?

**Which** of these questions can be answered using the data alone?

- Is college GPA higher for students who win a certain scholarship?
- Does getting the scholarship **improve** students' GPAs?

# Prediction vs Causation

Questions about **correlation / prediction**:

- Are homes with granite countertops worth more money?
- Is college GPA higher for students who win a certain scholarship?
- Are breastfed babies less likely to develop asthma?
- Do cancer patients given some aggressive treatment have a higher 5-year survival rate?
- Are people who smoke more likely to get cancer?

These sound like **causal questions**, but **they are not**!

# Prediction vs Causation

Questions about **correlation / prediction**:

- Are homes with granite countertops worth more money?
- Is college GPA higher for students who win a certain scholarship?
- Are breastfed babies less likely to develop asthma?
- Do cancer patients given some aggressive treatment have a higher 5-year survival rate?
- Are people who smoke more likely to get cancer?

Questions about **causality**:

- How much do granite countertops **raise** the value of a house?
- Does getting the scholarship **improve** students' GPAs?
- Does breastfeeding **protect** babies against asthma?
- Does the treatment **improve** cancer survival?

- Does smoking **cause** cancer?

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

Note: Regression coefficients θ are sometimes called "effects." This can be deceptive!

# Prediction vs Causation

**Only one** of these questions can be answered using the data alone:

✅ **Predictive question:** Are breastfed babies healthier?
❌ **Causal question:** Does breastfeeding improve babies' health?

Possible explanations for **why** breastfed babies would be healthier, on average:

1. **Causal effect:** breastfeeding makes babies healthier
2. **Reverse causality**: healthier babies more likely to successfully breastfeed
3. **Common cause**: healthier / richer parents have healthier babies **and** are more likely to breastfeed

We cannot tell which explanations are true (or to what extent) just by observing (x,y) pairs!

Causal questions implicitly involve **counterfactuals** (an event that did not happen):

- **Would** the **same** breastfed babies have been less healthy **if** they **hadn't** been breastfed?
- Explanation 1 implies they would be, explanations 2 and 3 do not

# Confounders

Let T: Treatment, e.g., alcohol use.

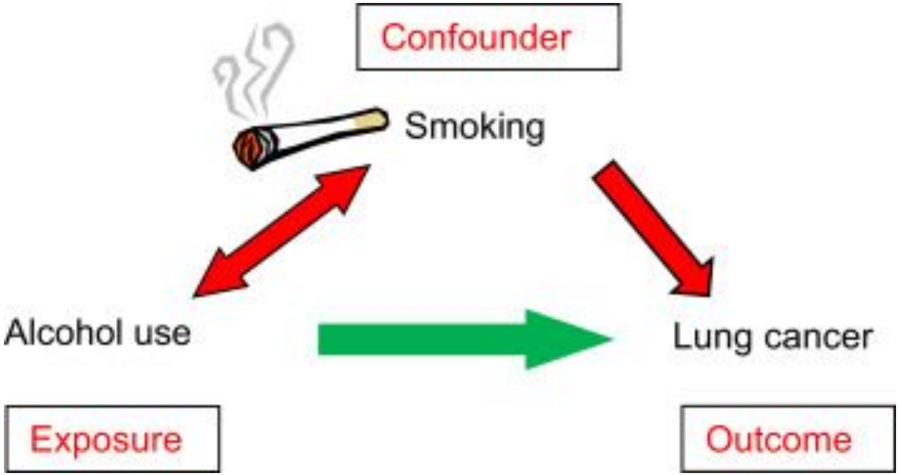
Y: Outcome, e.g., lung cancer.

Suppose we observe that people who drink are more likely to have lung cancer.

A **confounder** is a variable that affects both T and Y, distorting the correlation between them.

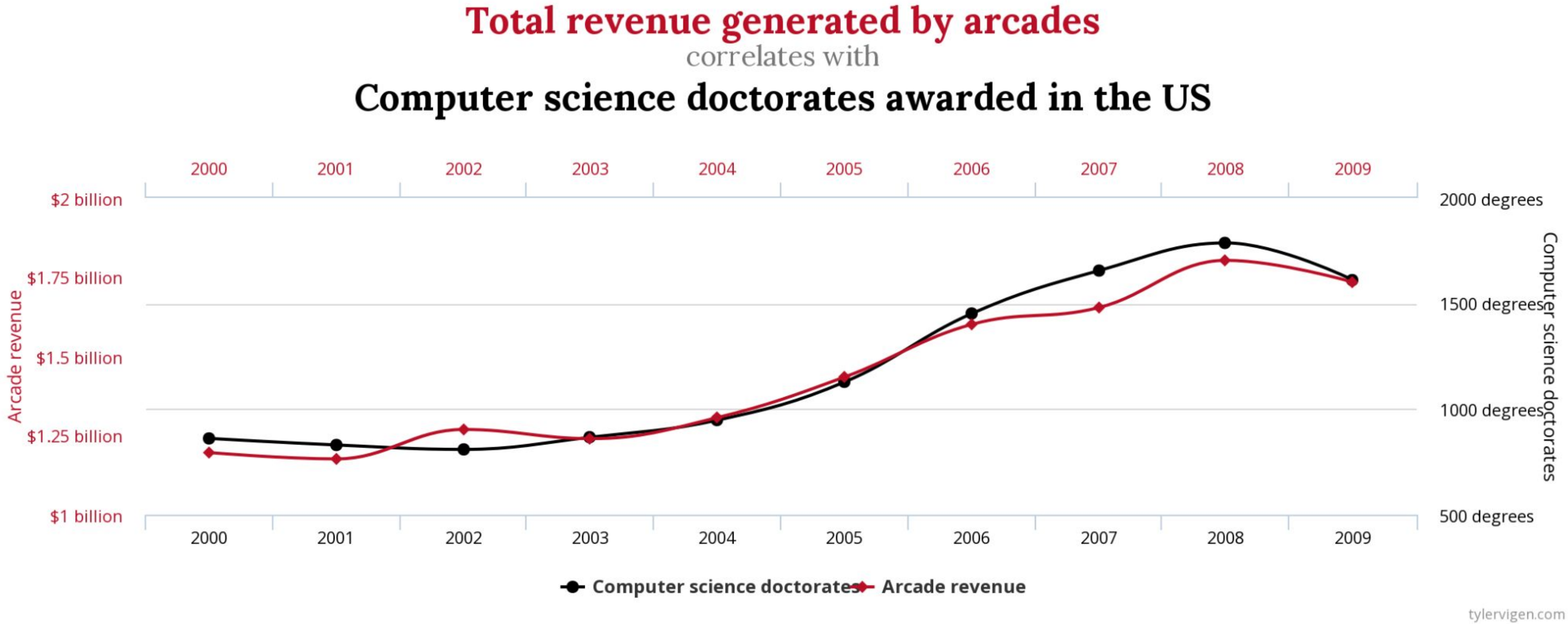
- (e.g. rich parents → breastfeeding, baby's health)
- Can be a measured covariate (feature), or unmeasured variable we don't know about!

Confounders generally cause problems.



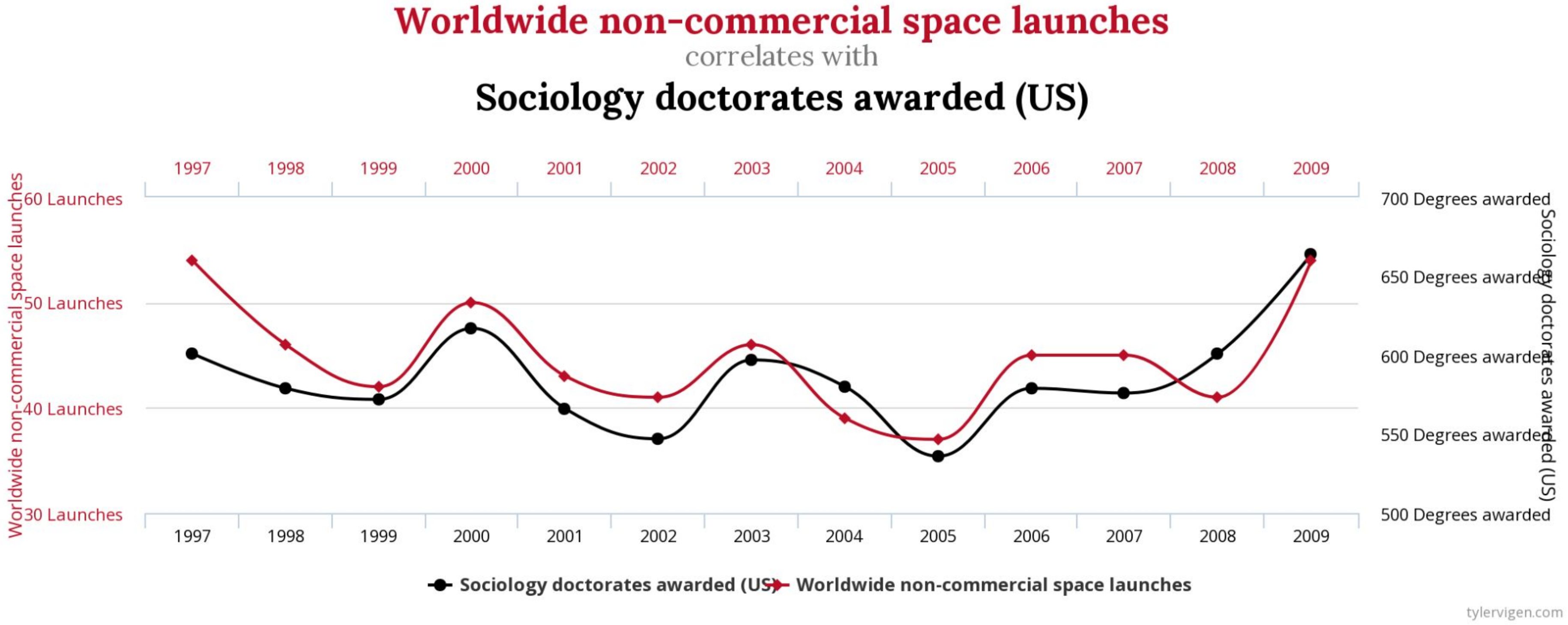
**Common assumption:** all confounders are observed (**ignorability**).

# Spurious Correlations



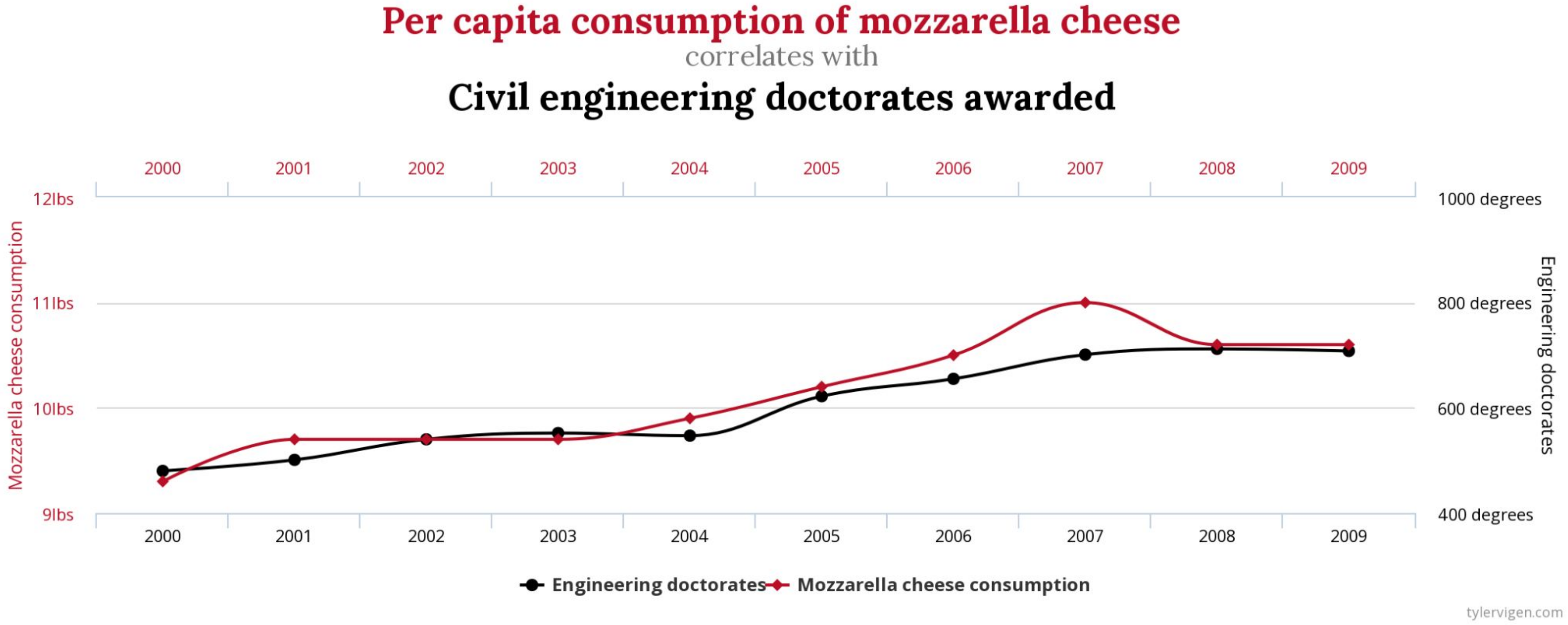
link

# Spurious Correlations



link

# Spurious Correlations



link

# How to perform causal inference?

**Ignorability**: All confounders are observed, i.e., covariates (data features) contain confounders.

In a **randomized experiment**:

- Randomly assign participants into two groups (the **treatment** and the **control** group) and then apply the treatment to the treatment group only.
- We assume ignorability and gather as many measurements as possible.
- Often not practical: randomly assigning treatments to participants is impractical or unethical!

In an **observational study**:

- Obtain two participant groups separated based on some identified treatment variable.
- Cannot assume ignorability: the participants could have separated into the two groups based on other covariates!
- There could also be unmeasured confounders!

There is an entire field of statistics called **causal inference** which studies causal models (i.e., treatment/covariate/response variables) in the context of observational studies.