PSYCH 201B

Statistical Intuitions for Social Scientists

Modeling data II

You can download these slides:
course website > Week 5 > Overview

02/04/2025

Today's Plan

- First Half-ish
  - Quick recap
  - Estimating models
  - Ordinary-Least-Squares (OLS)
- BREAK
- Second Half-ish
  - "Fun" interactive walk-through on your own
  - Any qs about HW

Quick recap

What is a **statistical** model?

A **theory** of how **observed data** were **generated**

$$\texttt{Data} = \texttt{Model} + \texttt{Error}$$

what's a good model?

a good model balances

fit and **complexity**

What is a model?

A **theory** of how **observed data** were **generated**

$$\texttt{Data} = \texttt{Model} + \texttt{Error}$$

how shall we define this?

Residual: the part that's left over after we have used the model to predict/explain the data

$$\text{Error} = \text{Data} - \text{Model}$$

Residual: the part that's left over after we have used the model to predict/explain the data

$$\text{Error} = \text{Data} - \hat{Data}$$

Model predictions

"hats" on variables are *estimated* not measured quantities

$$\text{Error} = \text{Data} - \text{Model}$$

assumed to be normally distributed

don't need to be normally distributed!!

very common misconception!!!

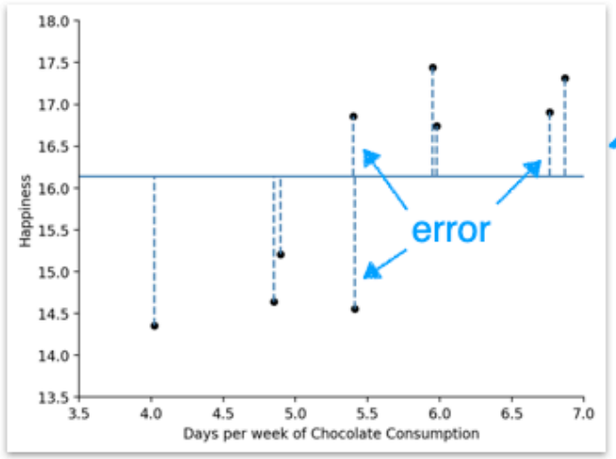
Modeling Data: Summary

$$\text{Data} = \text{Model} + \text{Error}$$

what's a good model?

a good model balances **fit** and **complexity**

concretely: fit models to minimize the **sum-of-squared-errors**



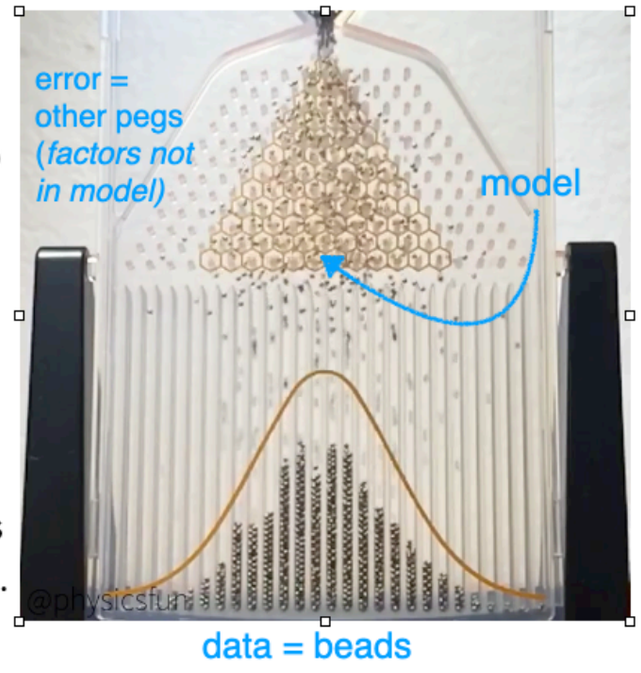
why squared error?

- positive and negative prediction errors don't cancel out
- larger errors are weighted more

$$\text{Error} = \text{Data} - \text{Model}$$

1. We **assume** that the **errors** are due to (*a potentially large number of*) factors that we didn't take into account.

2. We assume that each of these factors influences the data in **an additive way** (some pulling in one, others pulling in another direction).



Result: **Normally Distributed Errors**

Residual: the part that's left over after we have used the model to predict/explain the data

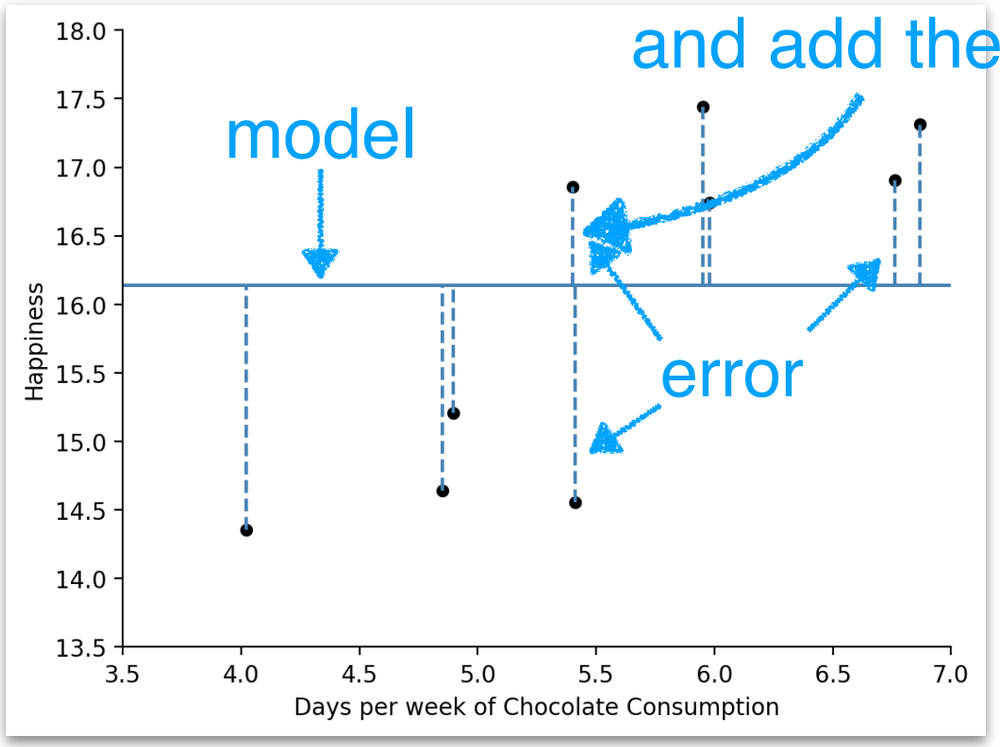
$$\text{Error} = \text{Data} - \text{Model}$$

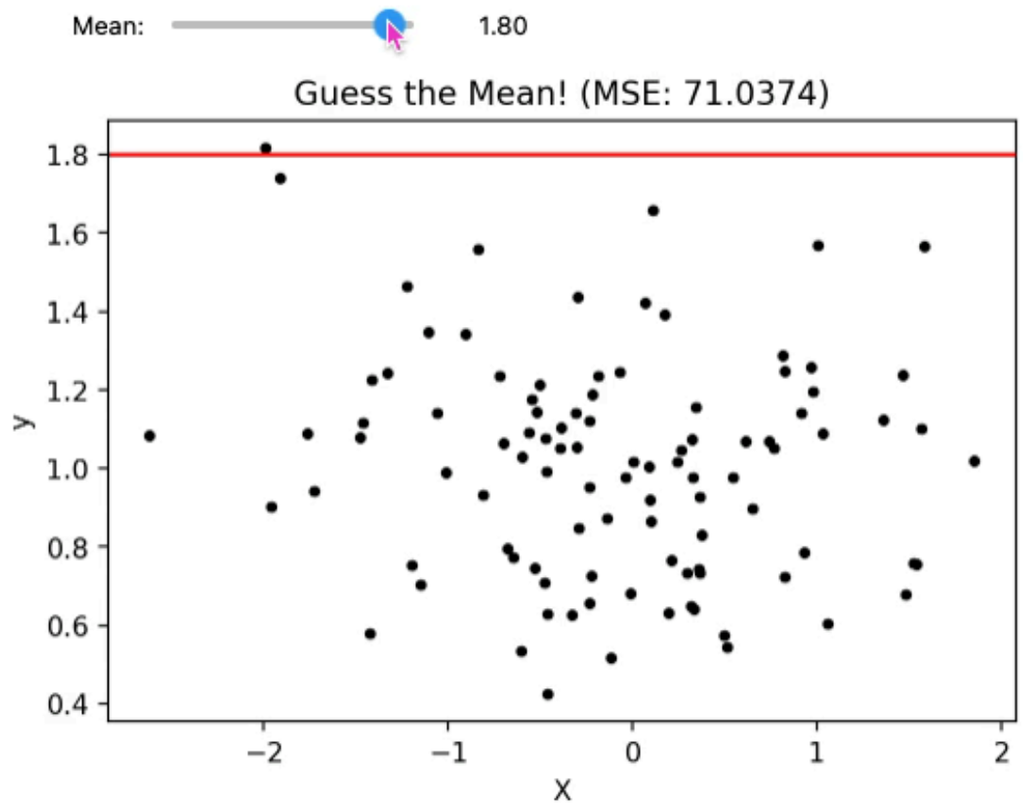
assumed to be normally distributed

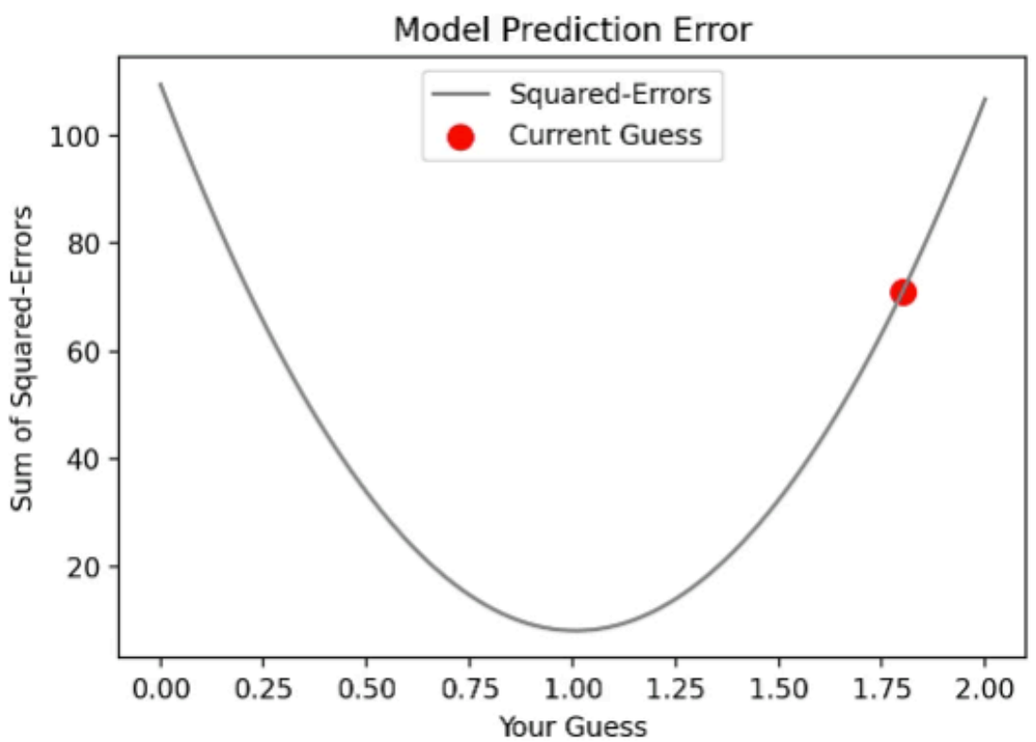
don't need to be normally distributed!!

very common misconception!!!

The **mean** is the **best 1 parameter model** when best = minimize **sum-of-squared** error





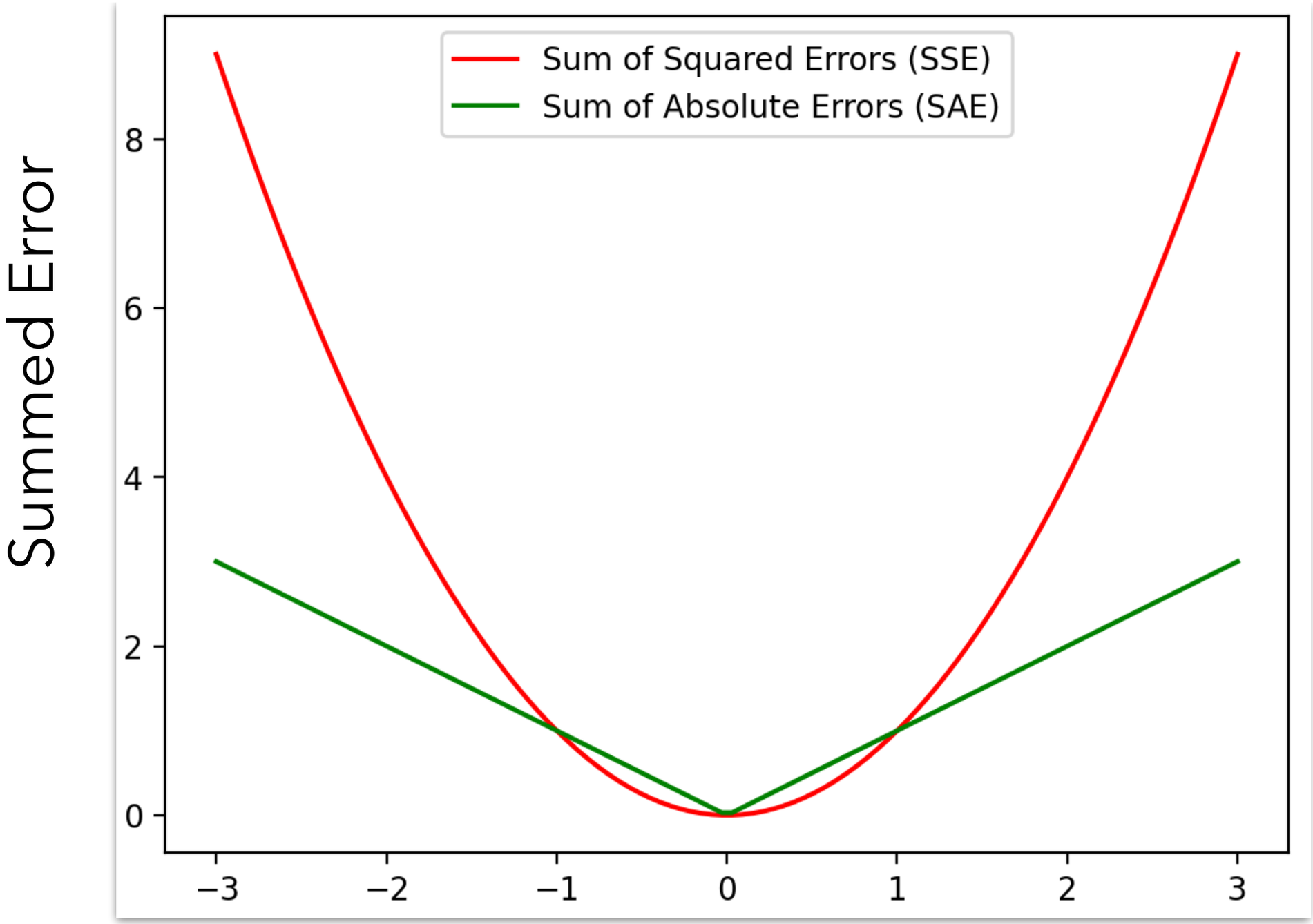


This is **why** the median is a *robust estimator* (model) compared to the mean

New perspective: ***Outliers*** *as* ***Predictions*** *with very high* ***Error***

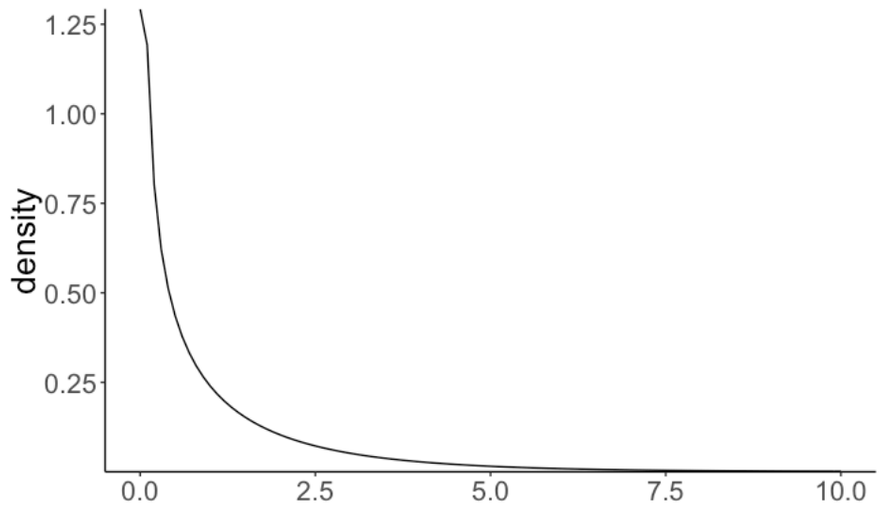
Sum-of-Squared-Error (SSE) - grows *quadratically* with error - **sensitive to outliers**

Sum-of-Absolute-Error (SAE) - grows *linearly* with error - **more robust to outliers**



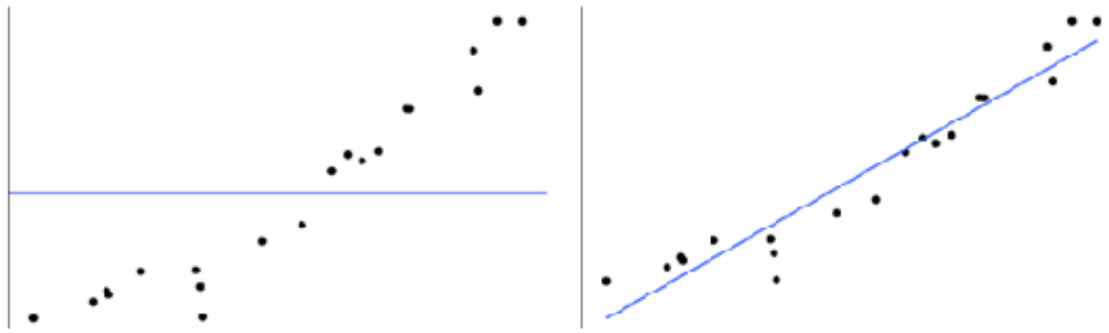
Hypothesis testing as model comparison

1. Start with research question
2. Formulate **hypothesis as a comparison** between *compact* and *augmented* model
3. Fit parameters in each model
4. Calculate the **proportional reduction of error** (PRE) to **compare models'** ability to *predict* data (compact model vs augmented model)
5. Decide whether PRE is worth it using F distribution



Summary: Hypothesis testing as model comparison

Why model fits data best?



Compact model

model_C: $Y_i = \beta_0 + \text{ERROR}$

Augmented model

model_A: $Y_i = \beta_0 + \beta_1 X_i + \text{ERROR}$

$$\text{ERROR}(C) \geq \text{ERROR}(A)$$

Proportional reduction in error (PRE)

$$\text{PRE} = \frac{\text{ERROR}(C) - \text{ERROR}(A)}{\text{ERROR}(C)}$$

The **worth it?** question

fit to the data

1 parameter(s) → model_2: $Y_i = \beta_0 + \text{ERROR}$

worth it?

additional predictor

2 parameter(s) → model_3: $Y_i = \beta_0 + \beta_1 X_i + \text{ERROR}$

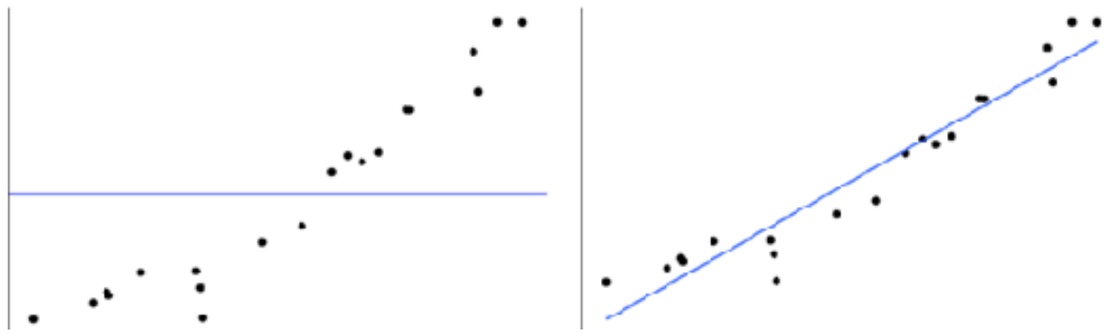
fit to the data

- to answer the worth it? question we need inferential statistics (define ERROR, sampling distributions, etc.)
- more likely to be worth it if:
  1. **PRE** is high
  2. the number of additional parameters in A compared to C is low (**PRE per additional parameter**)
  3. the number of parameters we *could have, but didn't use* to create model_A is high; it costs fewer *degrees-of-freedom*

more impressive if the number of parameters (p) far fewer than number of observations (n)

Summary: Hypothesis testing as model comparison

Why model fits data best?



Compact model

model_C: $Y_i = \beta_0 + \text{ERROR}$

Augmented model

model_A: $Y_i = \beta_0 + \beta_1 X_i + \text{ERROR}$

$$\text{ERROR}(C) \geq \text{ERROR}(A)$$

Proportional reduction in error (PRE)

$$\text{PRE} = \frac{\text{ERROR}(C) - \text{ERROR}(A)}{\text{ERROR}(C)}$$

The **worth it?** question

fit to the data

1 parameter(s) → model_2: $Y_i = \beta_0 + \text{ERROR}$

worth it?

additional predictor

2 parameter(s) → model_3: $Y_i = \beta_0 + \beta_1 X_i + \text{ERROR}$

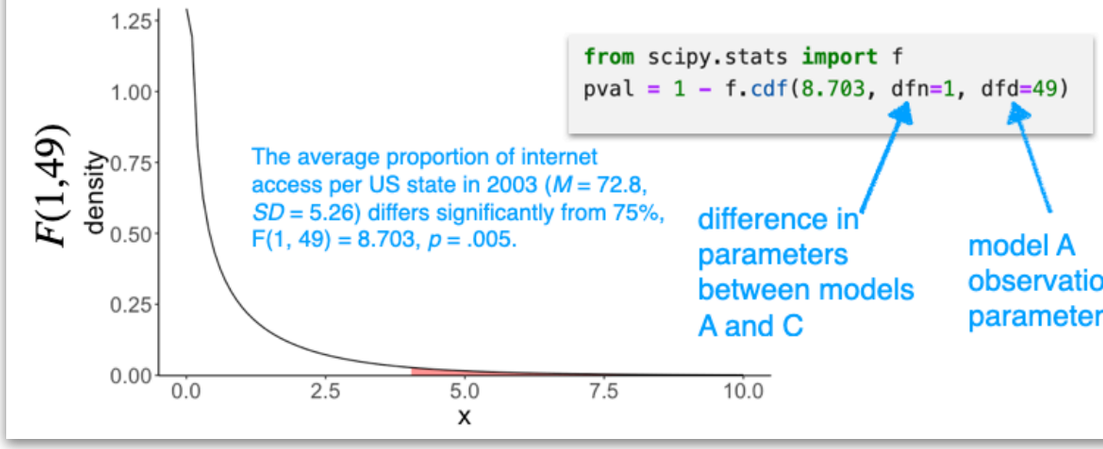
fit to the data

Decide whether it's worth it

$$F = \frac{\text{PRE}/(\text{PA} - \text{PC})}{(1 - \text{PRE})/(n - \text{PA})}$$

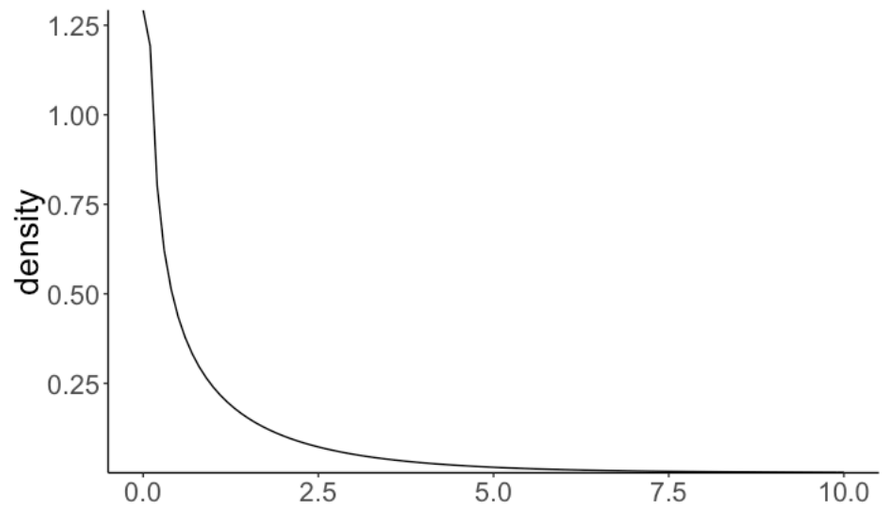
$$= \frac{.15/(1-0)}{(1-.15)/(50-1)} = 8.703$$

$p = .00486$



Hypothesis testing as model comparison

1. Start with research question
2. Formulate **hypothesis as a comparison** between *compact* and *augmented* model
3. **Fit parameters in each model** ← **?**
4. Calculate the **proportional reduction of error** (PRE) to **compare models'** ability to *predict* data (compact model vs augmented model)
5. Decide whether PRE is worth it using F distribution

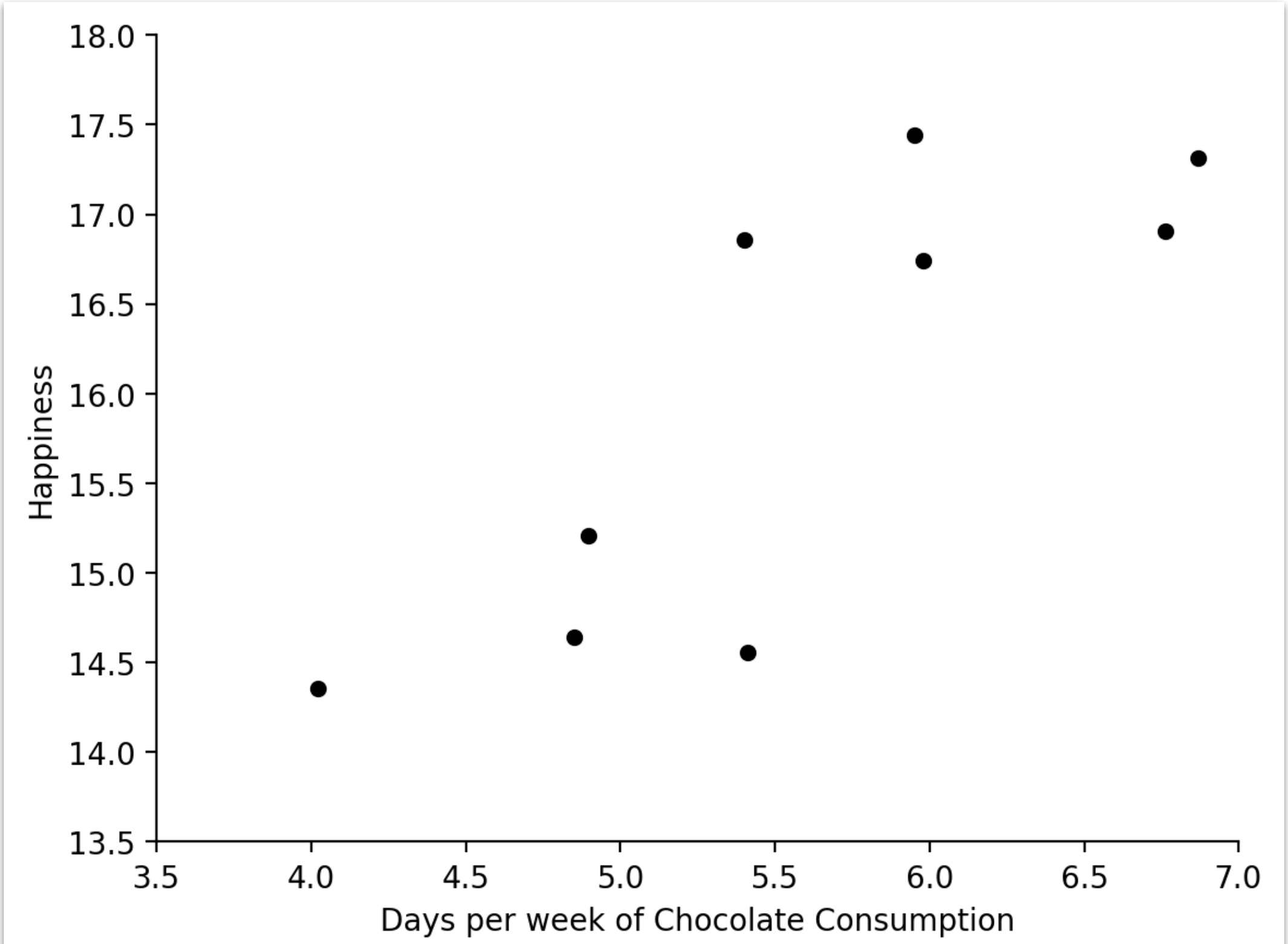


Fundamental concepts of statistics

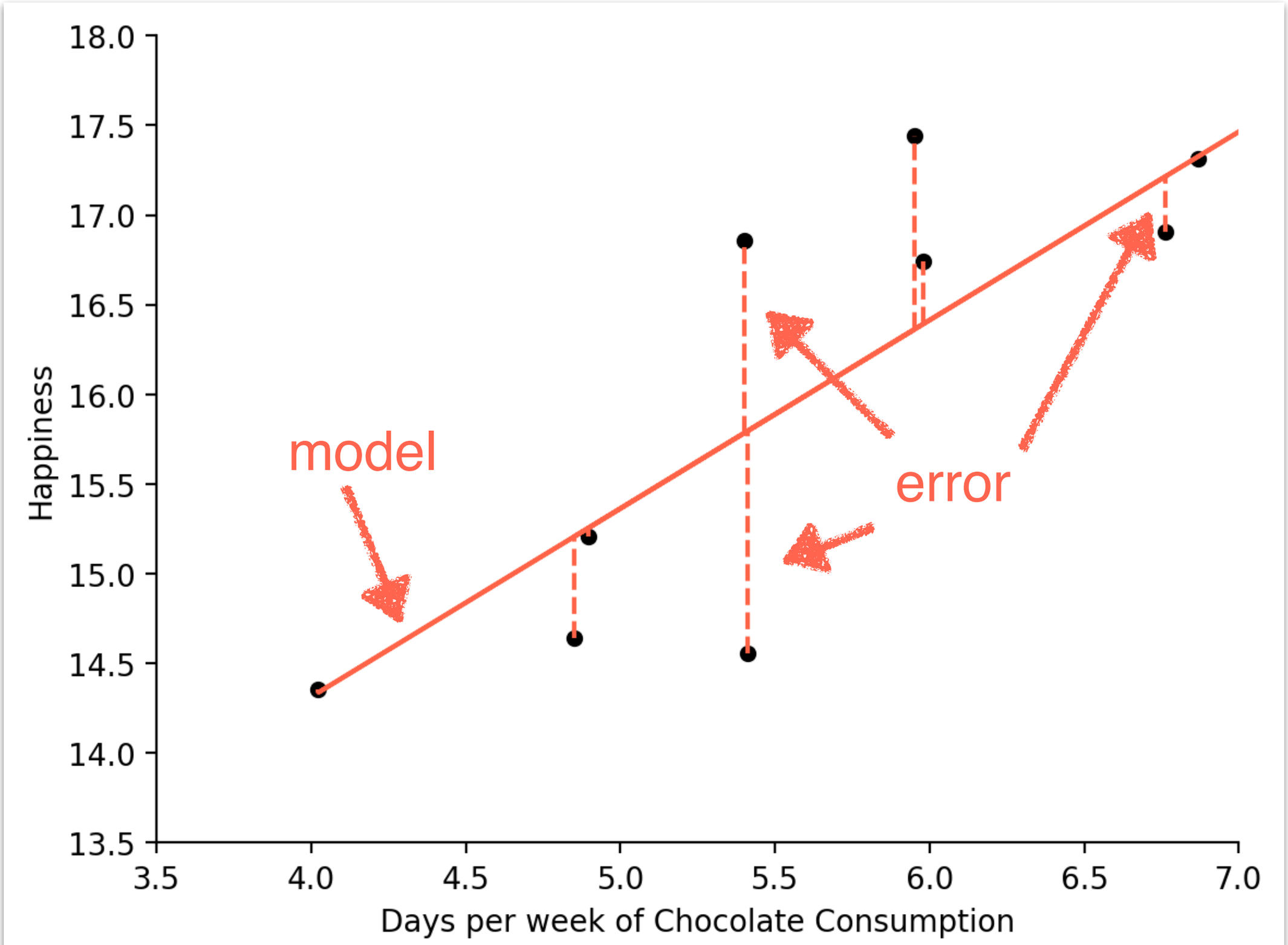
- **Aggregation**
  - describe and *compress* the data into a summary
- ***Re*-Sampling**
  - facilitate *generalizing* to unseen data
- **Uncertainty**
  - quantifying *trust* in our estimates based on different sources of error
- **Learning**
  - using data to *update* our estimates

Model *estimation*

Data = Model + Error



Data = Model + Error



Linear model

$$\texttt{happiness}_{\texttt{prediction}} = \texttt{mean}_{\texttt{happiness}} + \texttt{slope}_{\texttt{chocolate}} + \texttt{error}$$

$$\hat{Y}_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

the model is a **linear combination** of predictors

Linear model

$$\texttt{happiness}_{\texttt{prediction}} = \texttt{mean}_{\texttt{happiness}} + \texttt{slope}_{\texttt{chocolate}} + \texttt{error}$$

$$\hat{Y}_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Dependent Variables

Independent Variable(s)

Linear model

$$\texttt{happiness}_{\texttt{prediction}} = \texttt{mean}_{\texttt{happiness}} + \texttt{slope}_{\texttt{chocolate}} + \texttt{error}$$

$$\hat{Y}_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

model **intercept**
(mean of happiness)

chocolate **slope**
(change in happiness for 1-unit change in chocolate consumption)

How do we figure this out?

Two fundamental approaches for estimation

1. Iterative approach (*guess-timation)*
    - Guess -> predict -> error -> update guess

Iterative Approach

$$\texttt{happiness}_{\texttt{prediction}} = \texttt{mean}_{\texttt{happiness}} + \texttt{slope}_{\texttt{chocolate}} + \texttt{error}$$

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

$$= 0.5 + 11 * X_i$$

random initial guesses

repeat

$\hat{Y}_i =$ ← make prediction

$Y_i - \hat{Y}_i$ ← calculate error

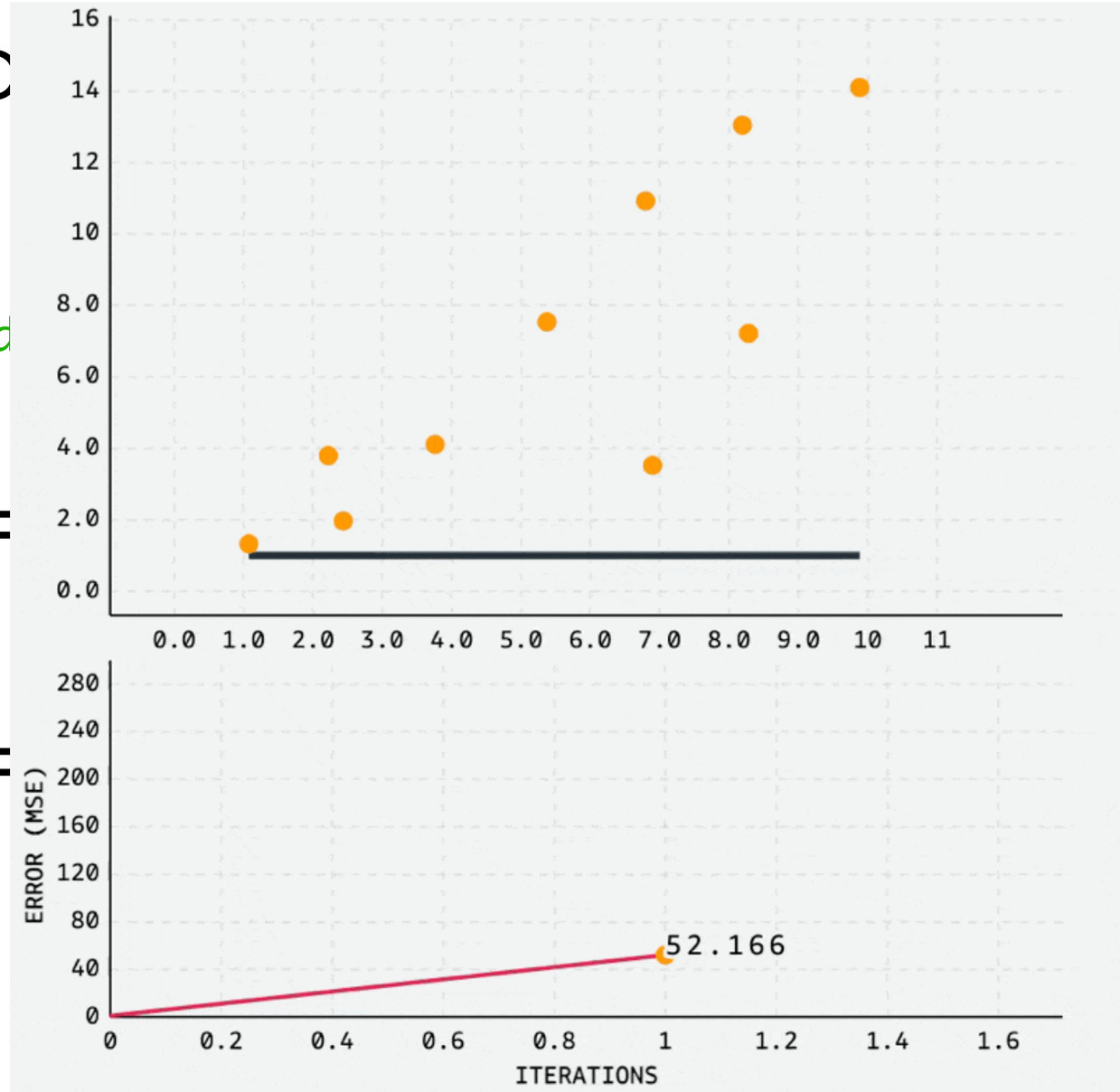
update guess

Iterative App

$happiness_{pred}$ late + error

$Y_i =$

$Y_i =$



esses

repeat

$Y_i = \hat{Y}_i$ ← model prediction

$Y_i - \hat{Y}_i$ ← prediction error

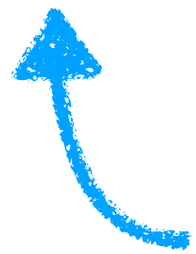
update guess

Two fundamental approaches for estimation

1. Iterative approach (*guess-timation)*
    - Guess -> predict -> error -> update guess
    - Works for almost any model if you can guess “intelligently”

Two fundamental approaches for estimation

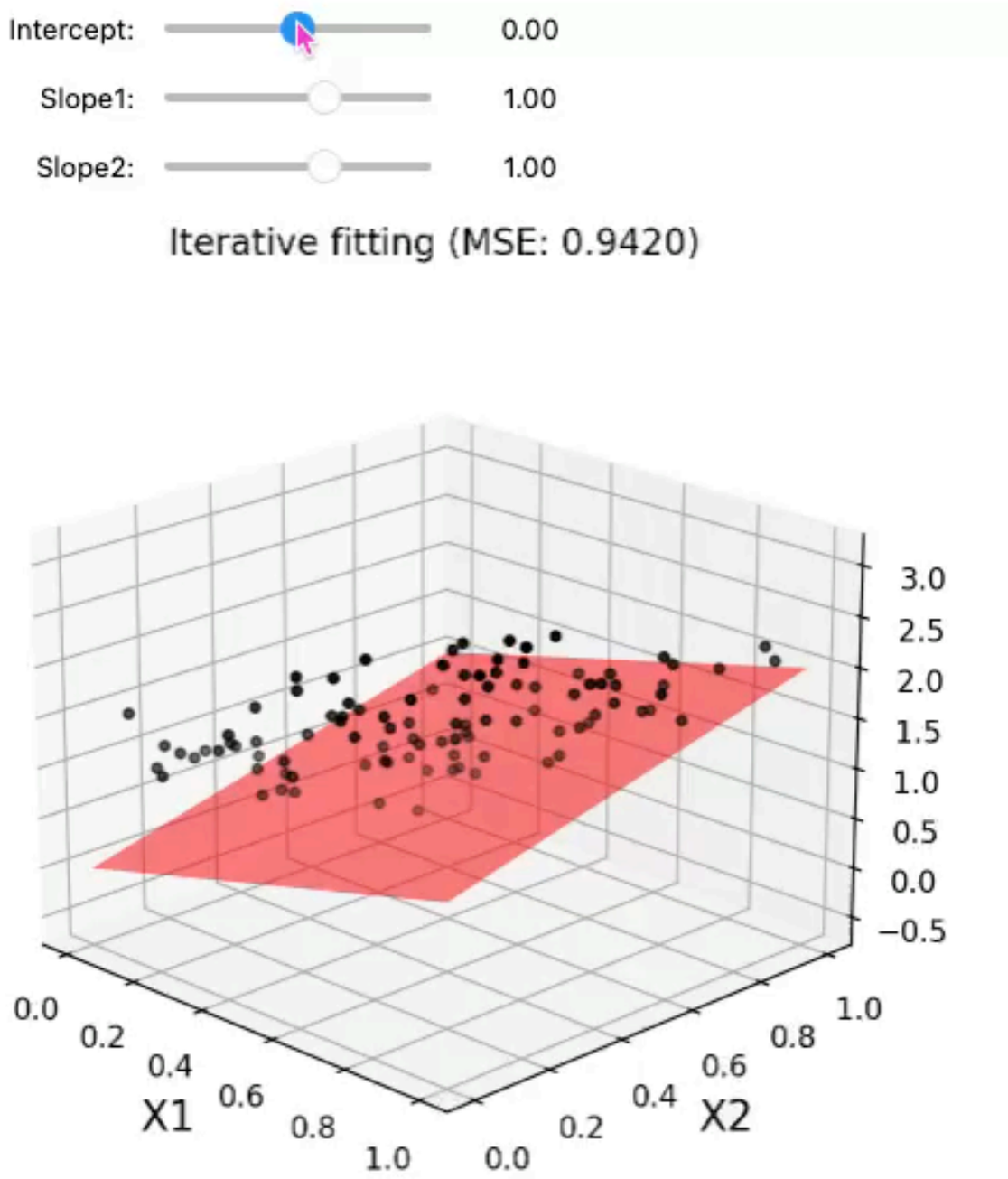
1. Iterative approach (*guess-timation)*
    - Guess -> predict -> error -> update guess
    - Works for almost any model if you can guess “intelligently”
        - Stochastic Gradient Descent (SGD)



we won’t cover this (but you should be aware)
- work-horse for modern machine-learning

Two fundamental approaches for estimation

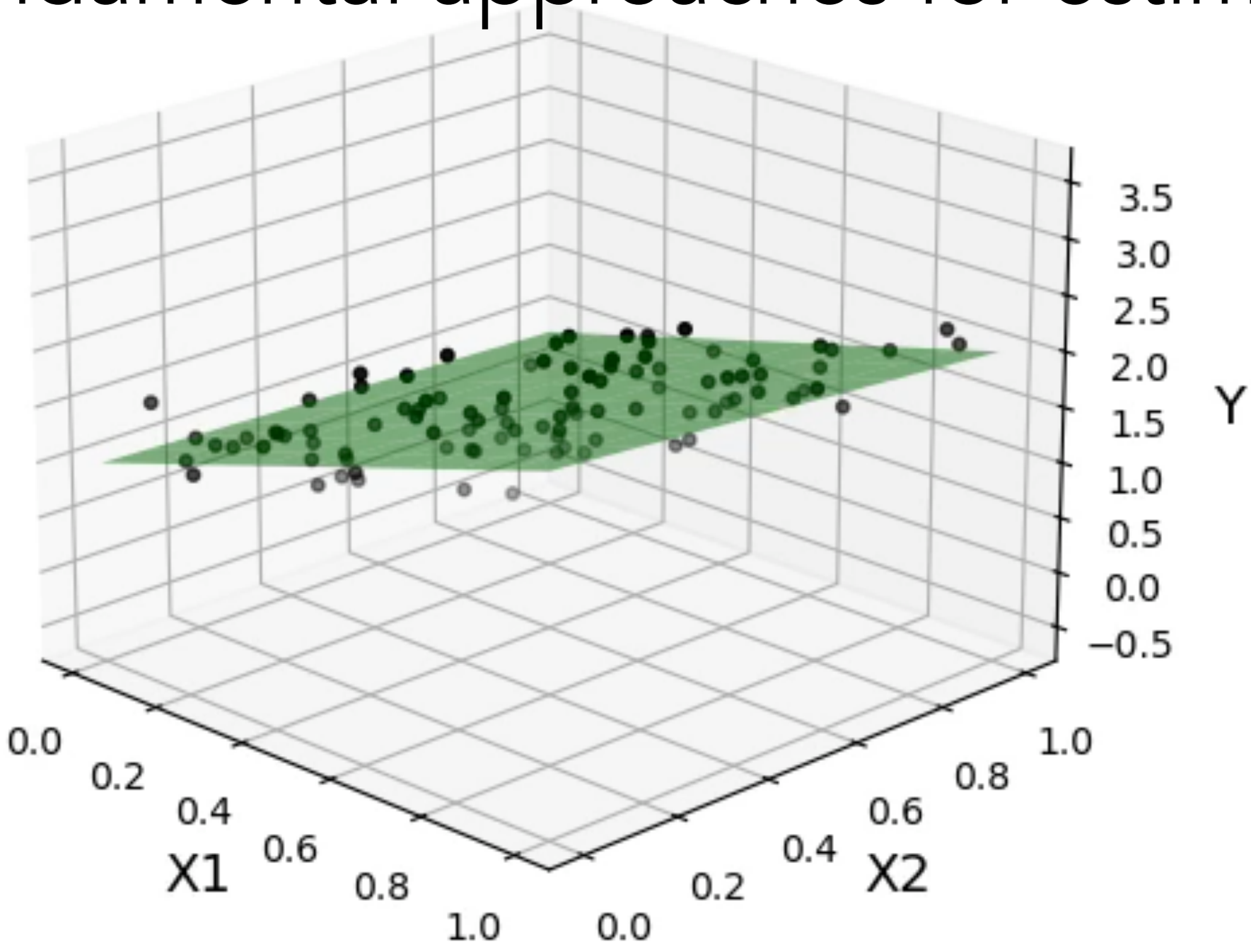
1. **Iterati**



- Guess guess
- Works an guess "intel
- Sensi
  - You
  - How
- *These ne!*
- Especially if we add more independent variables…

Two fundamental approaches for estimation

1.

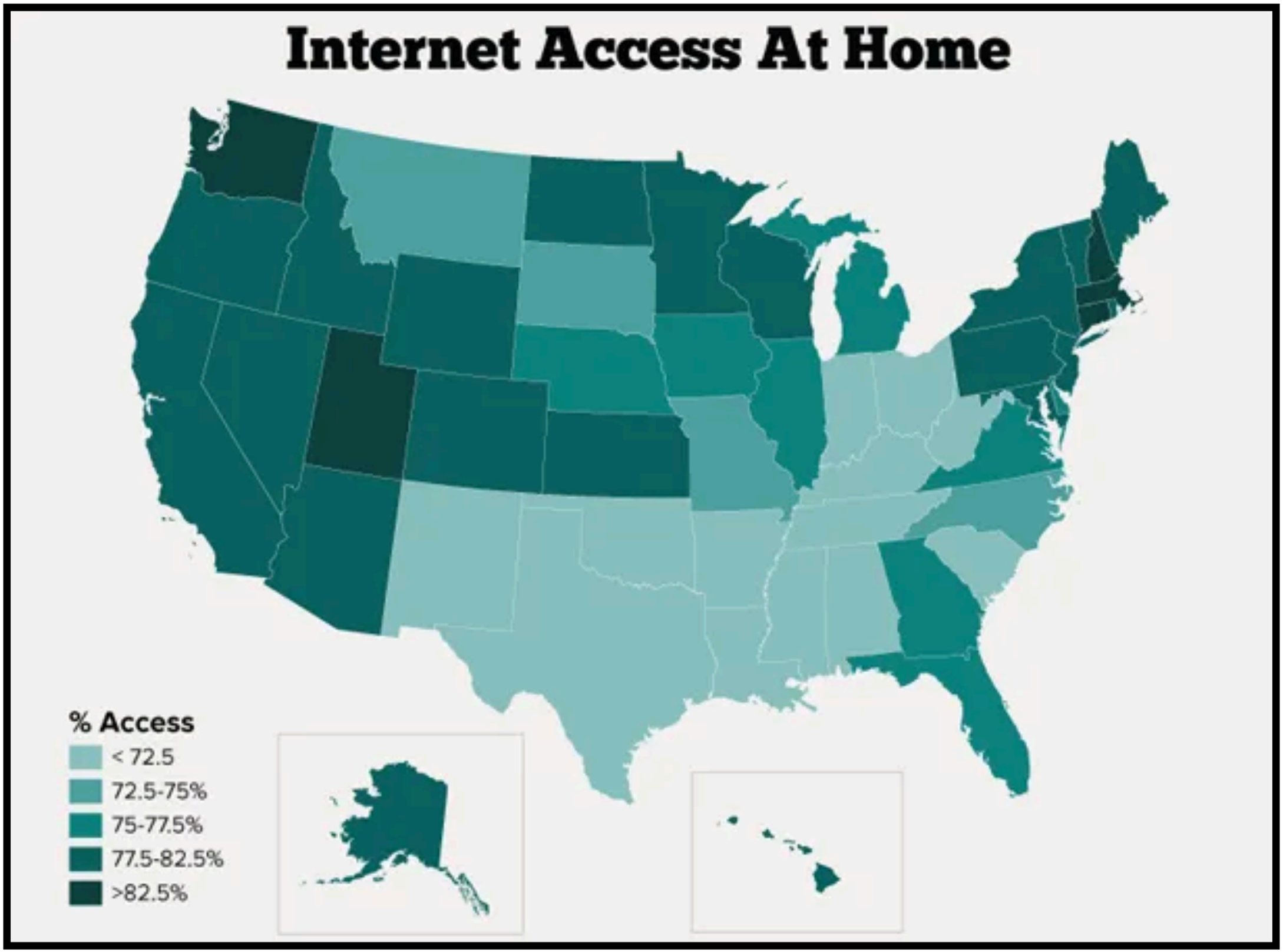


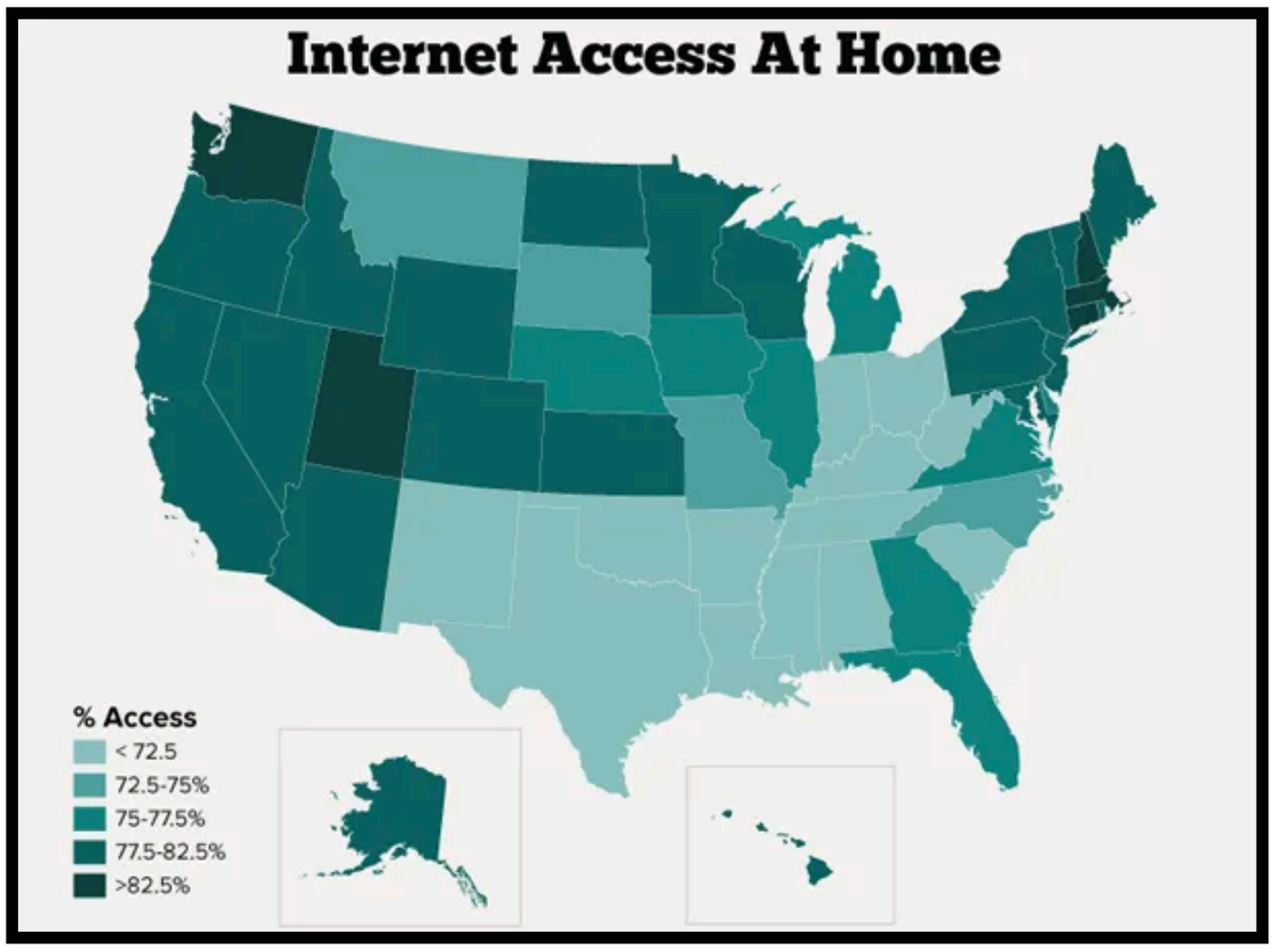
- Especially if we add more independent variables…

Two fundamental approaches for estimation

1. Iterative approach
2. **Analytic approach**
    - Linear algebra to the rescue!

Internet Access At Home





shape: (50, 6)

State	Internet	College	Auto	Density	SES
str	f64	f64	f64	f64	f64
"AK"	79.0	28.0	1.17	1.2	41.264052
"AL"	63.5	23.5	1.34	94.4	32.150157
"AR"	60.9	20.6	1.7	56.0	31.428738
"AZ"	73.9	27.4	1.27	56.3	39.190893
"CA"	77.9	31.0	0.84	239.1	40.817558
…	…	…	…	…	…
"VT"	75.3	35.7	0.98	67.9	37.211926
"WA"	78.9	32.7	0.8	101.2	38.197205
"WI"	73.0	27.7	0.96	105.0	37.27749
"WV"	64.9	18.9	1.64	77.1	30.836102
"WY"	75.5	26.6	1.66	5.8	37.53726

From 1 equation to *many* equations

$\text{internet}_{\text{prediction}} = \text{mean}_{\text{internet}} + \text{slope}_{\text{SES}}$

$$\hat{Y}_i = \beta_0 + \beta_1 X_i$$

$$79 = \beta_0 + \beta_1 * 41.26$$
$$63.5 = \beta_0 + \beta_1 * 32.15$$
$$60.9 = \beta_0 + \beta_1 * 31.43$$
$$\vdots$$
$$73.9 = \beta_0 + \beta_1 * 37.53$$

shape: (50, 6)

State	Internet	College	Auto	Density	SES
str	f64	f64	f64	f64	f64
"AK"	79.0	28.0	1.17	1.2	41.264052
"AL"	63.5	23.5	1.34	94.4	32.150157
"AR"	60.9	20.6	1.7	56.0	31.428738
"AZ"	73.9	27.4	1.27	56.3	39.190893
"CA"	77.9	31.0	0.84	239.1	40.817558
…	…	…	…	…	…
"VT"	75.3	35.7	0.98	67.9	37.211926
"WA"	78.9	32.7	0.8	101.2	38.197205
"WI"	73.0	27.7	0.96	105.0	37.27749
"WV"	64.9	18.9	1.64	77.1	30.836102
"WY"	75.5	26.6	1.66	5.8	37.53726

From 1 equation to many equations

$$
\begin{aligned}
79 &= \beta_0 + \beta_1 * 41.26 \\
63.5 &= \beta_0 + \beta_1 * 32.15 \\
60.9 &= \beta_0 + \beta_1 * 31.43
\end{aligned}
\qquad
\begin{bmatrix} 79 \\ 63.5 \\ 60.9 \end{bmatrix} = \begin{bmatrix} 1 & 41.26 \\ 1 & 32.15 \\ 1 & 31.43 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}
$$

Vector Matrix Vector

From 1 equation to many equations

Intercept
linear algebra for “mean”

2 unknown parameters
we’re *estimating*

$$\begin{bmatrix} 79 \\ 63.5 \\ 60.9 \end{bmatrix} = \begin{bmatrix} 1 & 41.26 \\ 1 & 32.15 \\ 1 & 31.43 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Dependent Variables

Independent Variable(s)

From 1 equation to many equations

$$\begin{bmatrix} 79 \\ 63.5 \\ 60.9 \end{bmatrix} = \begin{bmatrix} 1 & 41.26 \\ 1 & 32.15 \\ 1 & 31.43 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_0 \\ 1 & x_1 \\ \vdots & \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$\hat{y} = X\beta$$

The General Linear Model (GLM)

The General Linear Model (GLM)

$$y = X\beta + \epsilon$$

Simple (univariate) regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$$

one predictor

Multiple regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \epsilon_i$$

many predictors

y ← X β

dependent variable
response variable
regressand
outcome/target
measurements
output variable

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

independent variables
explanatory variables
regressors
predictors/features
design matrix
input variables

$$\begin{bmatrix} \text{---} \ \vec{x}_1 \ \text{---} \\ \text{---} \ \vec{x}_2 \ \text{---} \\ \vdots \\ \text{---} \ \vec{x}_n \ \text{---} \end{bmatrix}$$

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$$

coefficients
betas
slopes
effects
parameter-estimates

The General Linear Model (GLM)

$$y = X\beta + \epsilon$$

Models a scalar variable ***y*** as a linear function of predictor variables ***X*** by minimizing:

sum of squared errors → $\epsilon^2 = \sum (y - X\hat{b})^2$

y ← X β

dependent variable
response variable
regressand
outcome/target
measurements
output variable

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

independent variables
explanatory variables
regressors
predictors/features
design matrix
input variables

$$\begin{bmatrix} \text{---} \ \vec{x}_1 \ \text{---} \\ \text{---} \ \vec{x}_2 \ \text{---} \\ \vdots \\ \text{---} \ \vec{x}_n \ \text{---} \end{bmatrix}$$

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$$

coefficients
betas
slopes
effects
parameter-estimates

The General Linear Model (GLM)

$$y = X\beta + \epsilon$$

how do we find the best β that minimizes SSE?

Models a scalar variable ***y*** as a linear function of predictor variables ***X*** by minimizing:

sum of squared errors → $\epsilon^2 = \sum (y - X\hat{b})^2$

y ← X β

dependent variable response variable regressand outcome/target measurements output variable	$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$	independent variables explanatory variables regressors predictors/features design matrix input variables	$\begin{bmatrix} \vec{x}_1 \\ \vec{x}_2 \\ \vdots \\ \vec{x}_n \end{bmatrix}$	$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$ coefficients betas slopes effects parameter- estimates

The General Linear Model (GLM)

$$y = X\beta + \epsilon$$

how do we find the best β that minimizes SSE?

Ordinary Least Squares (**OLS**):
an *analytic* (closed-form) equation for estimating betas that minimizes SSE!

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

y		X	β	
dependent variable response variable regressand outcome/target measurements output variable	$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$	independent variables explanatory variables regressors predictors/features design matrix input variables	$\begin{bmatrix} \text{—} \vec{x}_1 \text{—} \\ \text{—} \vec{x}_2 \text{—} \\ \vdots \\ \text{—} \vec{x}_n \text{—} \end{bmatrix}$	$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$	coefficients betas slopes effects parameter-estimates

The General Linear Model (GLM)

$$y = X\beta + \epsilon$$

Ordinary Least Squares (**OLS**):
an *analytic* (closed-form) equation for estimating betas that minimizes SSE!

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

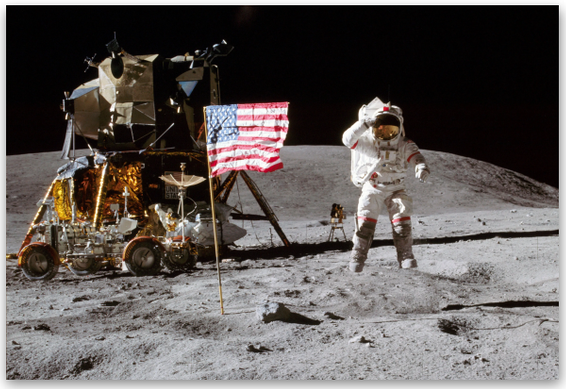
Just like **mean** = best *single* estimate that **minimizes SSE**

OLS = best **slope** estimate(s) that **minimizes SSE**

Ordinary Least Squares (**OLS):**
an *analytic* (closed-form) equation for estimating betas that minimizes SSE!

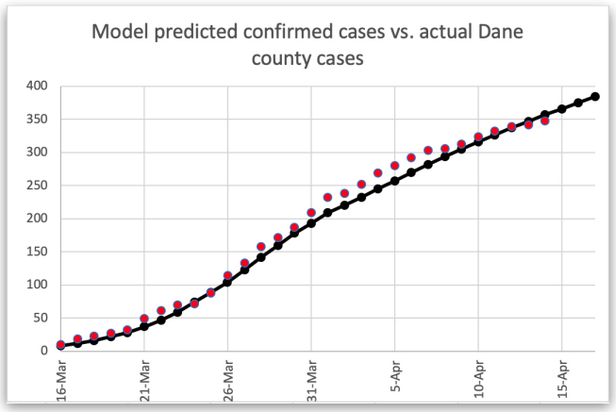
$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Foundational equation that underlies *most* of science & engineering



Getting to the Moon
Calibrating shuttle gyroscopes, accelerometers, flight paths, navigation

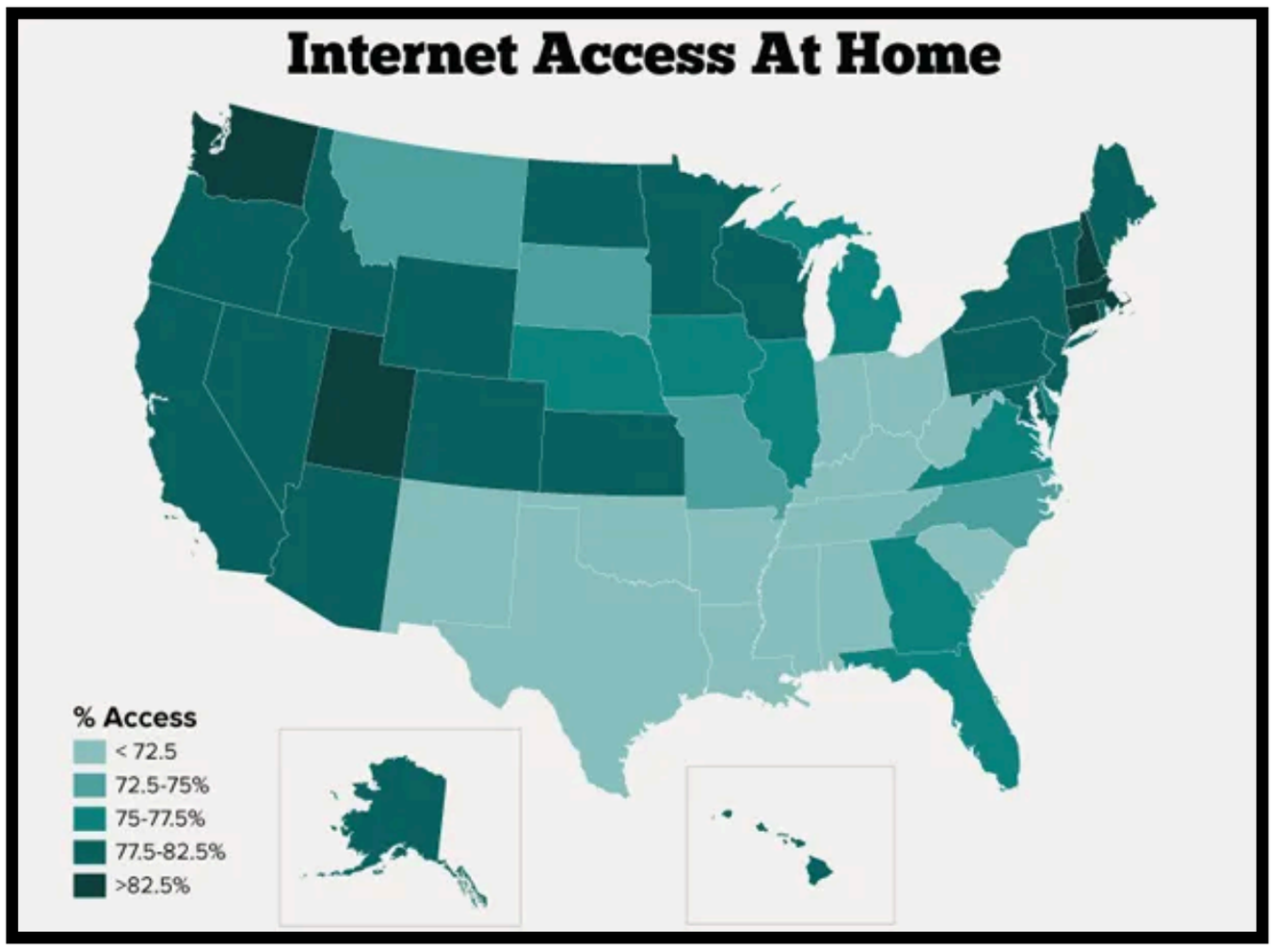
Building bridges
Structural stress-testing of metal, concrete, failure rates



Predicting epidemic spread
Infection rates, mortality, intervention efficacy

What is it doing?

$$\hat{\beta} = (X^T X)^{-1} X^T y$$



shape: (50, 6)

State	Internet	College	Auto	Density	+	SES
str	f64	f64	f64	f64		f64
"AK"	79.0	28.0	1.17	1.2		41.264052
"AL"	63.5	23.5	1.34	94.4		32.150157
"AR"	60.9	20.6	1.7	56.0		31.428738
"AZ"	73.9	27.4	1.27	56.3		39.190893
"CA"	77.9	31.0	0.84	239.1		40.817558
…	…	…	…	…		…
"VT"	75.3	35.7	0.98	67.9		37.211926
"WA"	78.9	32.7	0.8	101.2		38.197205
"WI"	73.0	27.7	0.96	105.0		37.27749
"WV"	64.9	18.9	1.64	77.1		30.836102
"WY"	75.5	26.6	1.66	5.8		37.53726

What is it doing?

$$\hat{y} = X\beta$$

$\texttt{internet}_{\texttt{prediction}} = \texttt{slope}_{\texttt{SES}} + \texttt{slope}_{\texttt{Density}}$

$y =$

Internet
f64
79.0
63.5
60.9
73.9
77.9
…
75.3
78.9
73.0
64.9
75.5

$X =$

Density	SES
f64	f64
1.2	41.264052
94.4	32.150157
56.0	31.428738
56.3	39.190893
239.1	40.817558
…	…
67.9	37.211926
101.2	38.197205
105.0	37.27749
77.1	30.836102
5.8	37.53726

Similarity between predictors
uncentered co-variance
(dot product!)

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Similarity between predictors and outcome

What is it doing?

$$\hat{y} = X\beta$$

$\texttt{internet}_{\texttt{prediction}} = \texttt{slope}_{\texttt{SES}} + \texttt{slope}_{\texttt{Density}}$

Matrix inverse
think division

$$x * \frac{1}{5} = \frac{x}{5}$$

Similarity between predictors
uncentered co-variance
(dot product!)

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Similarity between
predictors and outcome

What is it doing?

Calculate the *similarity* between each X and y after removing the covariance between Xs

*What is **unique contribution** of each x in predicting y?*

Matrix inverse
think division

$$x * \frac{1}{5} = \frac{x}{5}$$

Similarity between predictors
uncentered co-variance
(dot product!)

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Similarity between predictors and outcome

What is it doing?

*What is **unique contribution** of Density and SES in predicting Internet*

$y =$

Internet
f64
79.0
63.5
60.9
73.9
77.9
…
75.3
78.9
73.0
64.9
75.5

$X =$

Density	SES
f64	f64
1.2	41.264052
94.4	32.150157
56.0	31.428738
56.3	39.190893
239.1	40.817558
…	…
67.9	37.211926
101.2	38.197205
105.0	37.27749
77.1	30.836102
5.8	37.53726

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$\frac{\hat{\beta}_1 = covariance(x_1, y)}{\sigma_{x_1}\sigma_{x_1}} \qquad \frac{\hat{\beta}_2 = covariance(x_2, y)}{\sigma_{x_2}\sigma_{x_2}}$$

Regression vs correlation

Regression

*What is **unique contribution** of Density and SES in predicting Internet*

$$\hat{\beta}_1 = \frac{covariance(x_1, y)}{\sigma_{x_1}\sigma_{x_1}} \qquad \hat{\beta}_2 = \frac{covariance(x_2, y)}{\sigma_{x_2}\sigma_{x_2}}$$

Only denominator is different!

Correlation

*What is the **non-unique contribution** of Density and SES in predicting Internet*

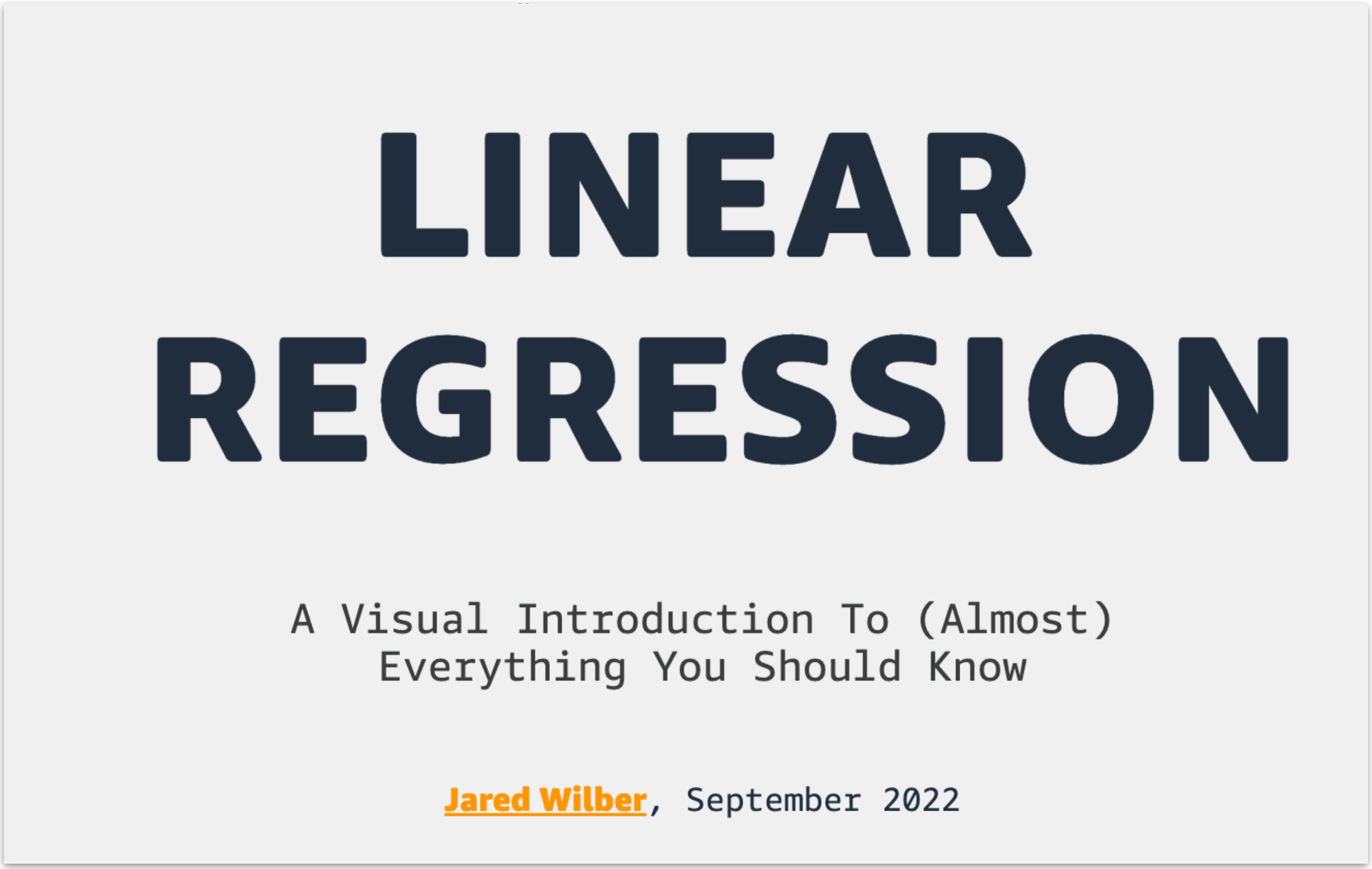
$$\hat{r}_{x_1 y} = \frac{covariance(x_1, y)}{\sigma_{x_1}\sigma_{y_1}} \qquad \hat{r}_{x_2 y} = \frac{covariance(x_2, y)}{\sigma_{x_2}\sigma_{y_1}}$$

Correlation doesn't account for shared variance between predictors

Two fundamental approaches for estimation

1. Iterative approach
2. **Analytic approach (OLS)**
    - Linear algebra to the rescue!
    - Lets us collect our IVs into a **design matrix**
    - **Matrix inversion** to remove similarity between IVs (columns of design matrix)
    - Calculate similarity between IVs and DV after removing similarity between IVs
    - **best** slope between 1 or more Xs and y ⟶ **minimize** sum-of-squared-error

Your turn: interactive walkthrough



https://mlu-explain.github.io/linear-regression/