PSYCH 201B

Statistical Intuitions for Social Scientists

A few more big ideas & what is a model?

01/14/2025

Today's Plan

1. Hear me blab for ~20-30min

2. Rest of the time for lab (will post HW in a few)

Today's Plan

1. Hear me blab for ~20-30min

2. Rest of the time for lab (will post HW in a few)

Last time…

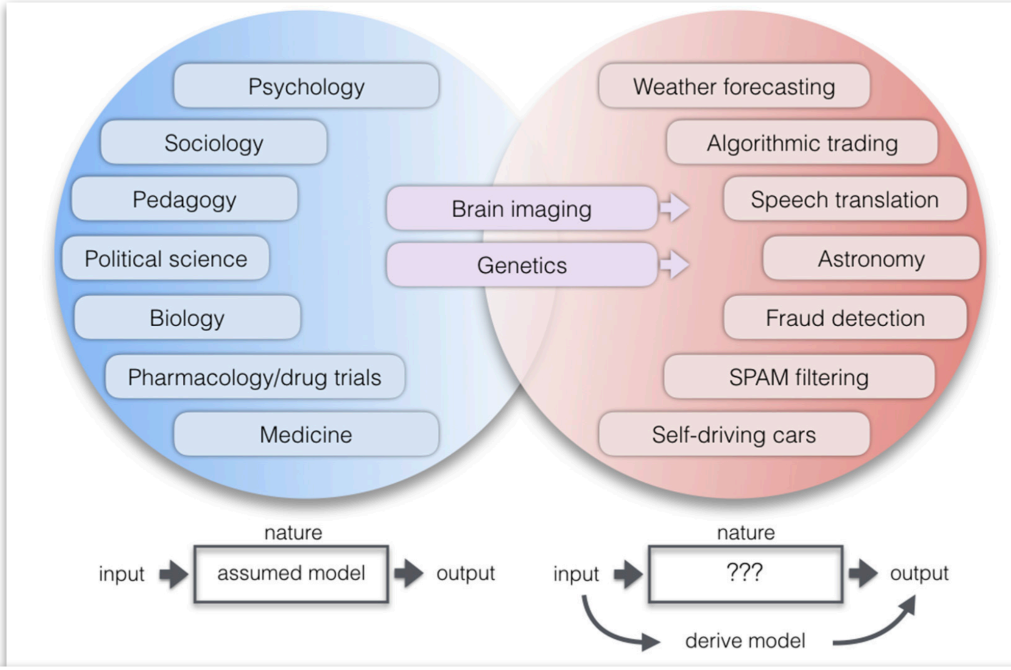
The four fundamental intuitions

- **Aggregation**
  - Gives us a model
- **Sampling**
  - Tells us where it applies
- **Uncertainty**
  - Keeps us honest
- **Learning**
  - Forces iteration

Statistics is not about being right

It's the process of becoming less wrong over time

The two cultures of statistics



„Truth" is in the model	1	„Truth" is in the data
„Long data" (n samples > p variables)	2	„Wide data" (n samples << p variables)
Make a set of judicious assumptions	3	Make least assumptions possible
Test a model for the data deductively	4	Learn a model from the data inductively
Impose mathematical rigor	5	Let the data speak for themselves
Choose a model before visiting the data	6	Choose a model as the data are visited
More confirmatory than exploratory	7	More exploratory than confirmatory
Model assumptions are largely explicit	8	Model assumptions are largely implicit
Data are often experimental	9	Data are often observational
Tractable models with few parameters	10	Expressive models with many parameters

GLM as our model "organism"

Explanation ⟷ **Prediction**

Classical inference

Theory		
null-hypothesis testing & multiple comparisons	bias-variance decomposition	Vapnik-Chervonenkis dimensions & curse of dimensionality
degrees of freedom df		hypothesis space $\mathcal{H}$
asymptotic consistency		finite-sample theorems
Invalidations of inferential process		
double dipping/circular analysis	post-selection inference	data snooping/peeking
Outcome metrics		
(in-sample) p values sensitivity/specificity effect size/power	explained variance metrics AUC/ROC curve confidence intervals	out-of-sample prediction accuracy/precision/recall/F1 scores learning curves certainty estimates via bootstrap
Representative methods		
Student's *t*-test *F*-test ANOVA Binomial test X^2-test linear regression	general(ized) linear model	support vector machines LASSO/ridge regression/elastic net logistic regression nearest neighbors random forests kernel methods („deep") neural networks

Out-of-sample generalization

Bzdok & Ionnadis, 2019

A few more big ideas

1) Law of Large Numbers

2) Central Limit Theorem

3) No Free Lunch Theorem

Important Caveat!

Laws $\neq$ Guarantees

“Laws” and “Theorems” are asymptotic intuitions about what *should* happen in the limit — not what *will* happen in your data

when you have **infinite** data

1) Law of Large Numbers (LLN)

- **Intuition**
  - As we average more independent observations our estimate stabilizes
  - Stabilize = converge to some fixed value
  - LLN reduces noise — does not imply correctness
- **Why?**
  - Random fluctuations cancel out — but systematic bias remains
- **In practice**
  - Averaging many noisy self-reports of mood
  - Averaging many decision-making gambles

2) Central Limit Theorem (CLT)

- **Intuition**
  - Distribution of estimator converges to normal distribution even if data distribution are not normal
  - estimator = aggregated point-value based on data
    - e.g. *mean*, *median*, *variance*
- **Why?**
  - the power of aggregation — we can model aggregates (e.g. mean) with normal assumptions even when underlying data is not normal!
- **In Practice**
  - Reaction times (RT) are skewed — mean RT is not
  - t-tests work even if data is not normally distributed

3) No Free-Lunch Theorem

- **Intuition**
  - Any method that performs well somewhere — must perform poorly elsewhere
  - There is no universally best model, estimator, algorithm etc
- **Why?**
  - Performance always depends on: data-generating process, assumptions, errors/loss-functions
- **In Practice**
  - Models/methods are tools not truths — usefulness is conditional
  - If there were a universal "best test" — we would only need 1 stats textbook!

Ok so what *is* a model?

What is a model?

General: A general mathematical **function** that transforms *inputs* into *outputs*

$$\text{output} = f(\text{input})$$

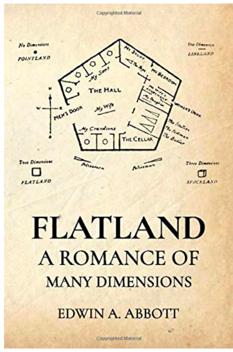
$$\text{happiness} = f(\text{chocolate})$$

$$y = f(x)$$

$$y = f(x1,\ x2, \ldots)$$

We care about this

What is a model?



General: A general mathematical **function** that transforms *inputs* into *outputs*

$$\text{output} = f(\text{input})$$

$$\text{happiness} = f(\text{chocolate})$$

$$y = f(x)$$

$$y = f(x1,\ x2, \ldots)$$

It's our theory in *math*

Jolly & Chang, 2019

What is a model?

"A model is a logical story expressed in *math* (code)"

~ Andrew Gelman

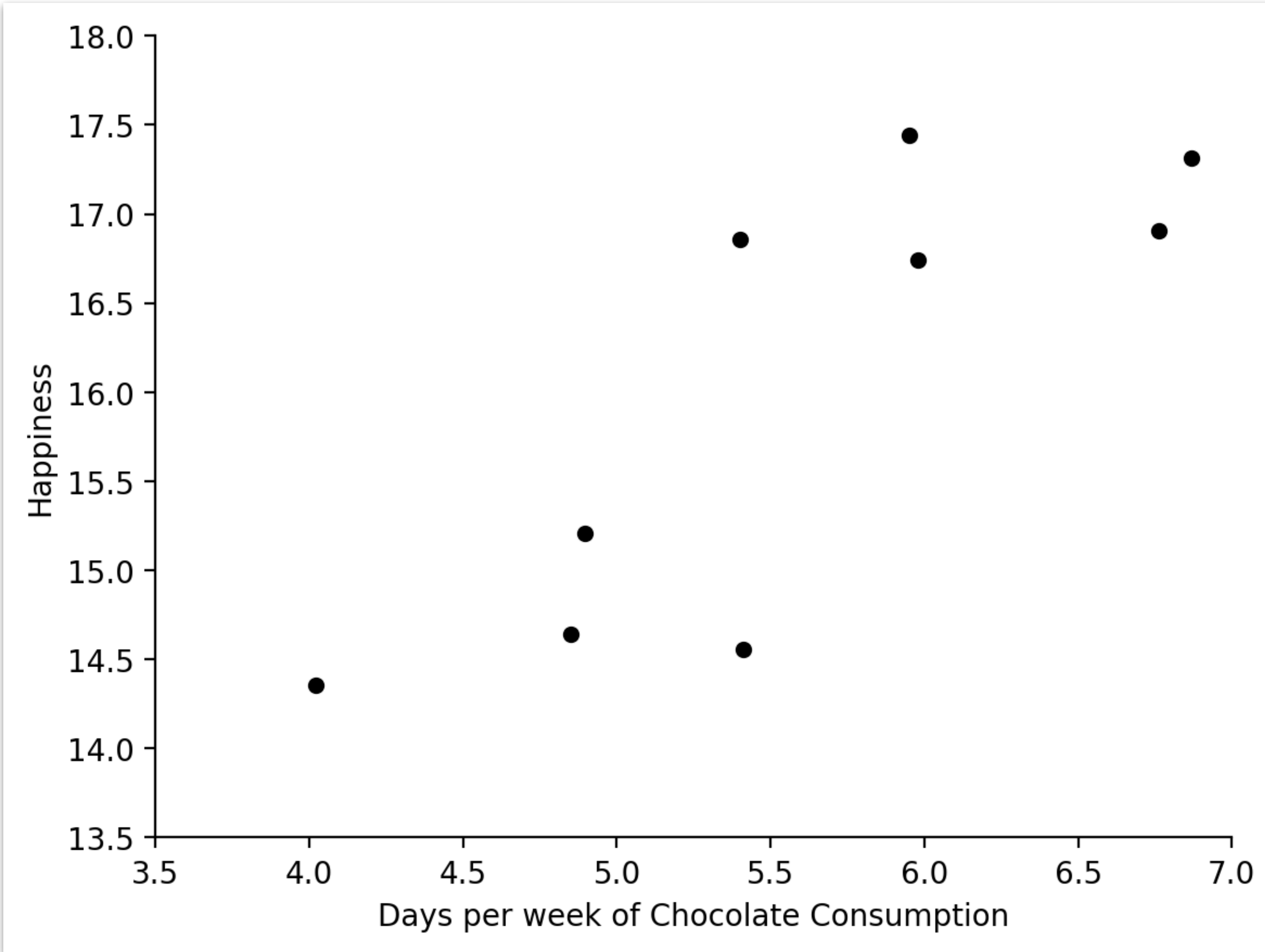
one of the most famous statisticians; inventor of Stan; Prof at Columbia

What is a **statistical** model?

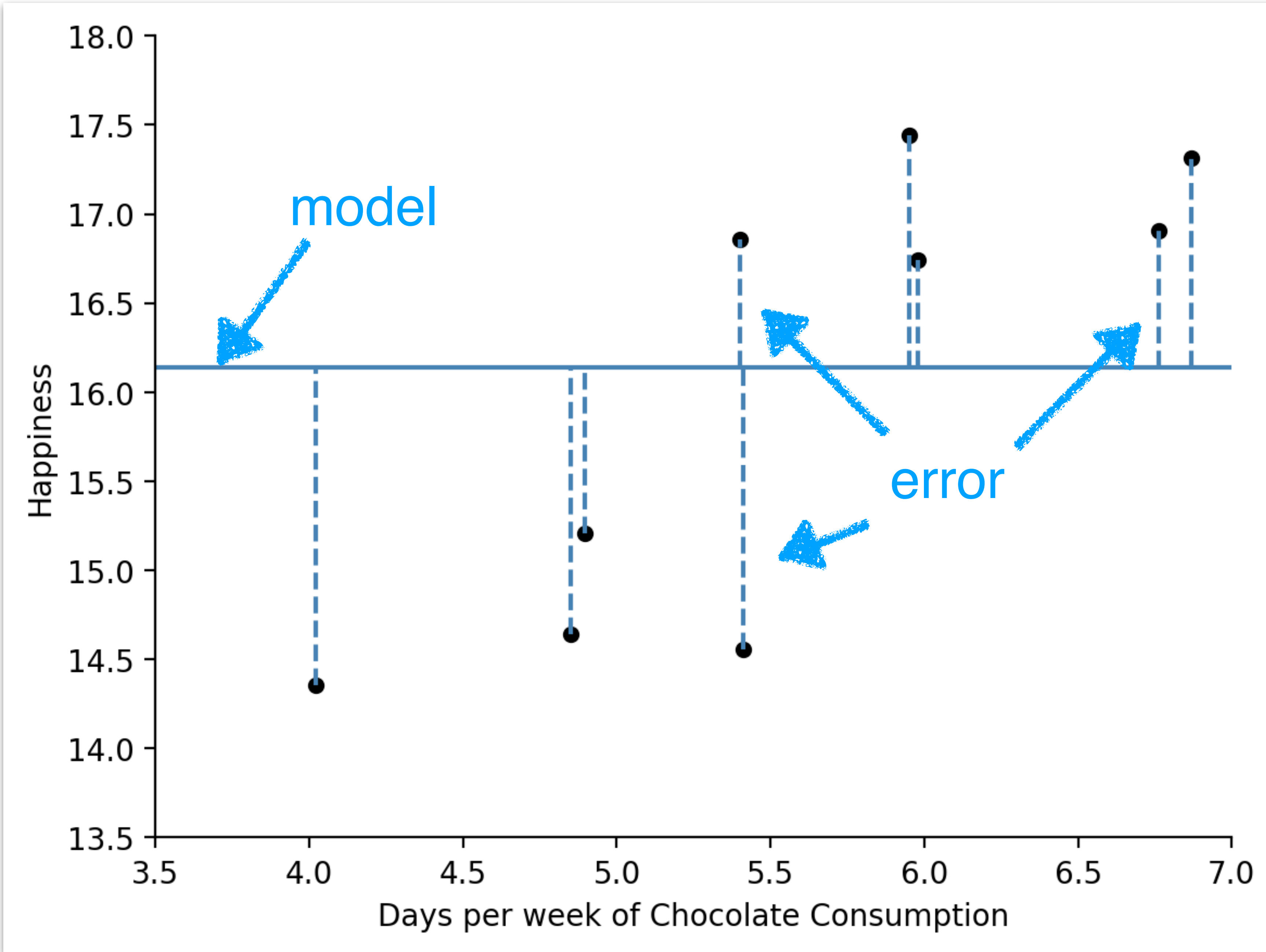
A **theory** of how **observed data** were **generated**

```
Data = Model + Error
```

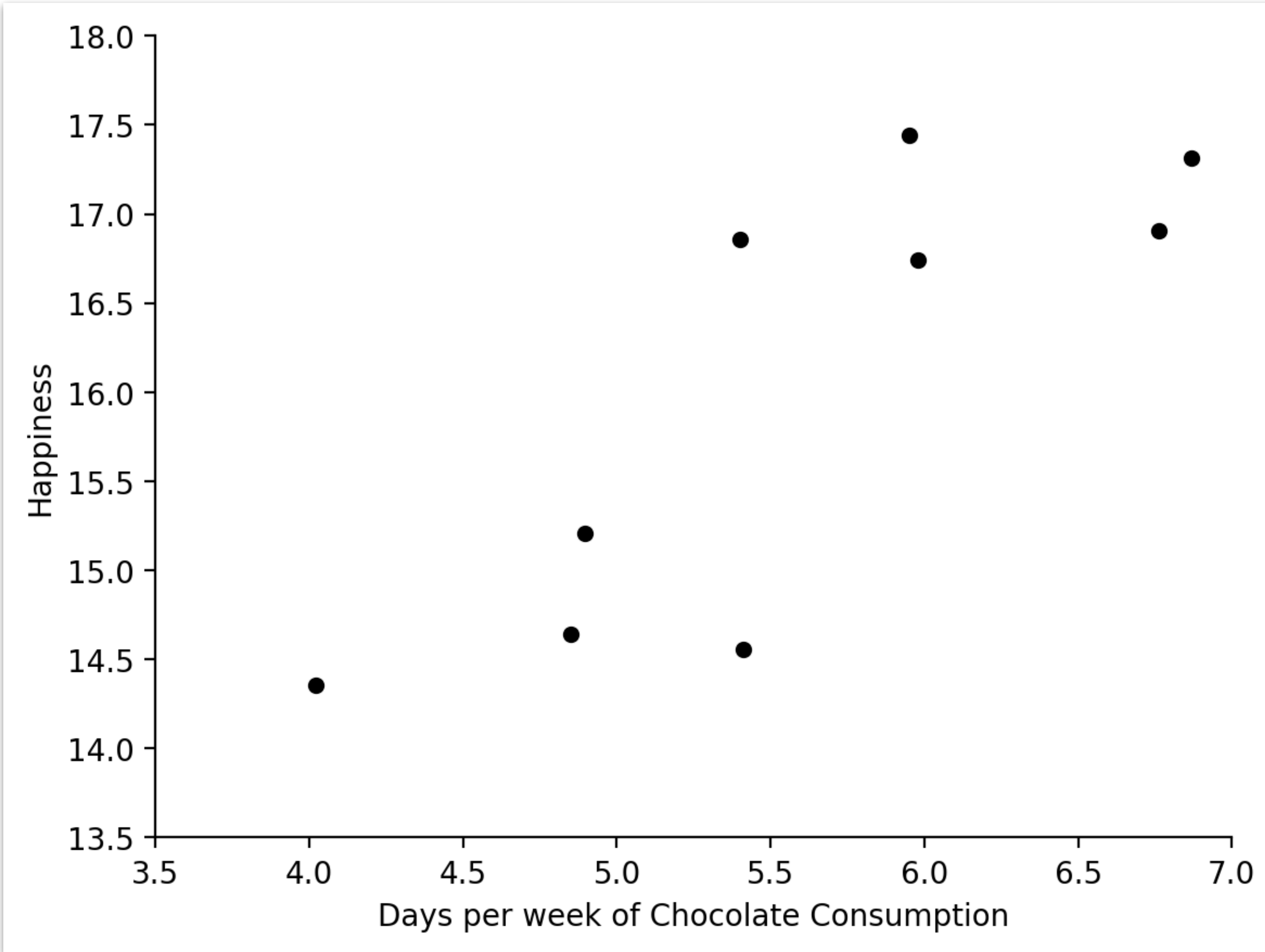
Data = Model + Error



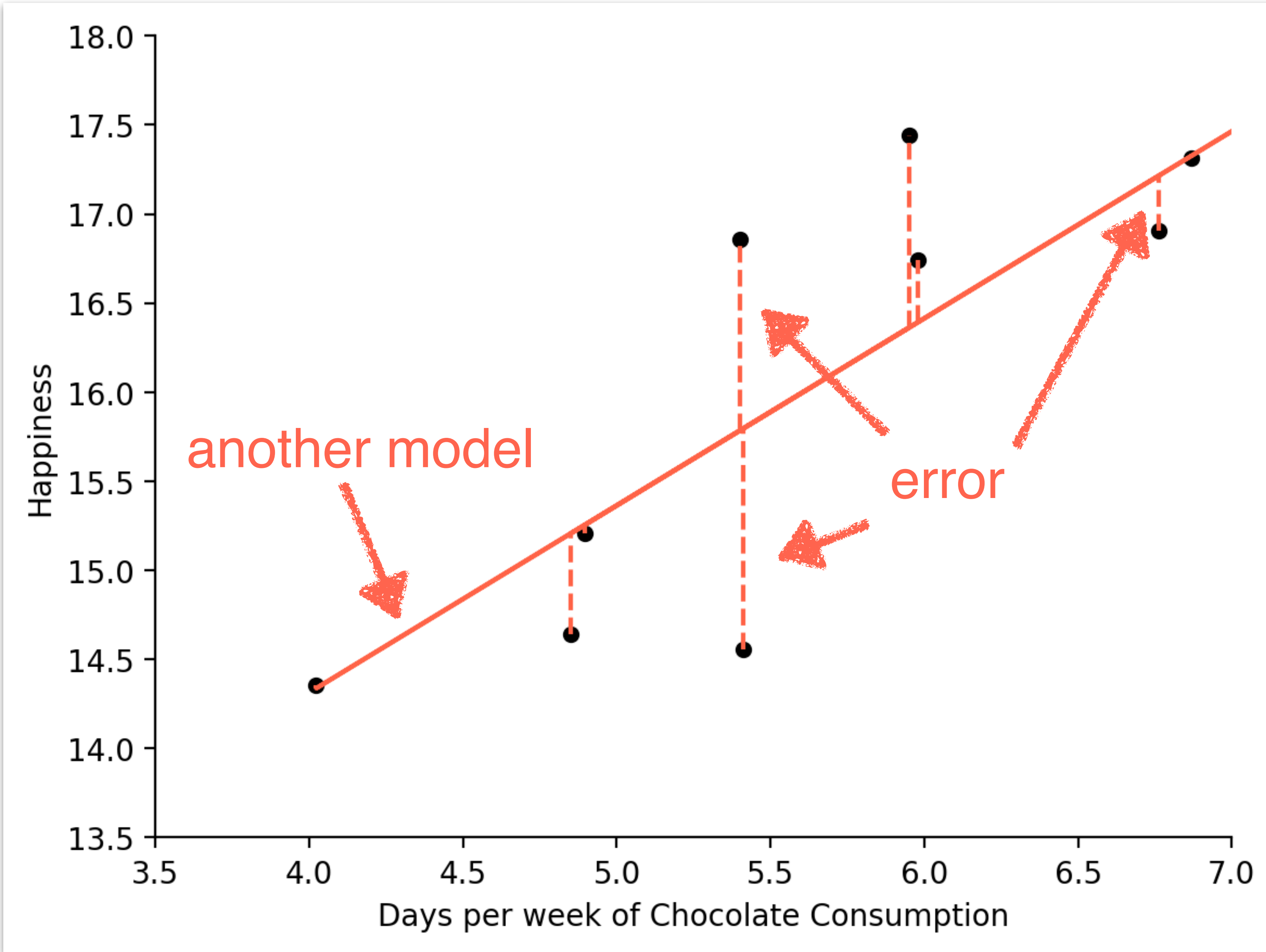
Data = Model + Error



Data = Model + Error



Data = Model + Error



What is a model?

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

```
Data = Model + Error
```

how shall we
define this?

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

```
Error = Data - Model
```

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

```
Error = Data - Model
```

Model **predictions**

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

```
Error = Data - Model
```

To reduce error we can:

improve data quality

e.g. run good experiments

improve the model

e.g. make predictions using additional information

Error = Data - Model

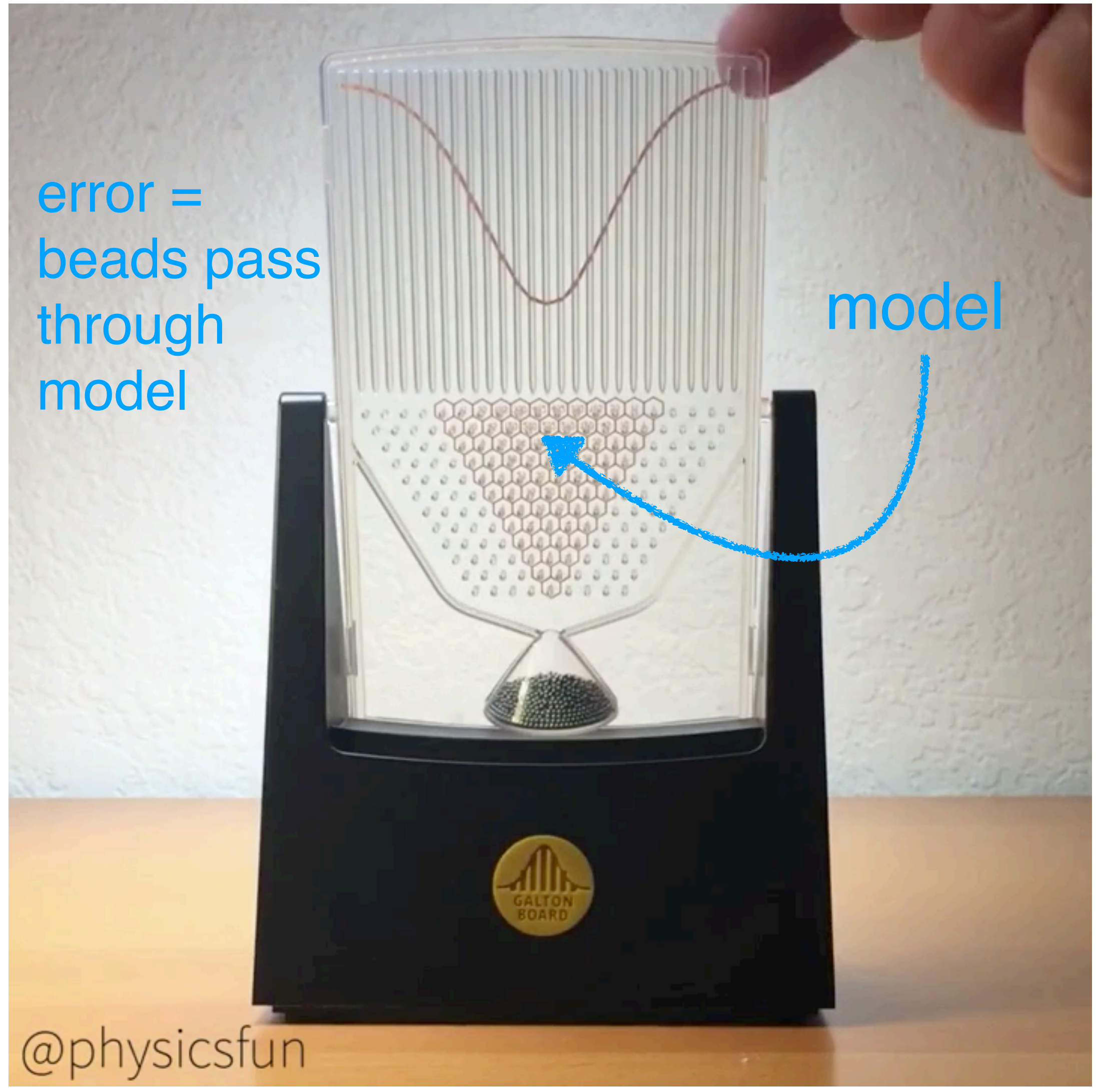
1. We **assume** that the **errors** are due to (*a potentially large number of*) factors that we didn't take into account.

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

Error = Data - 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).



data = beads

Result: **Normally Distributed Errors**

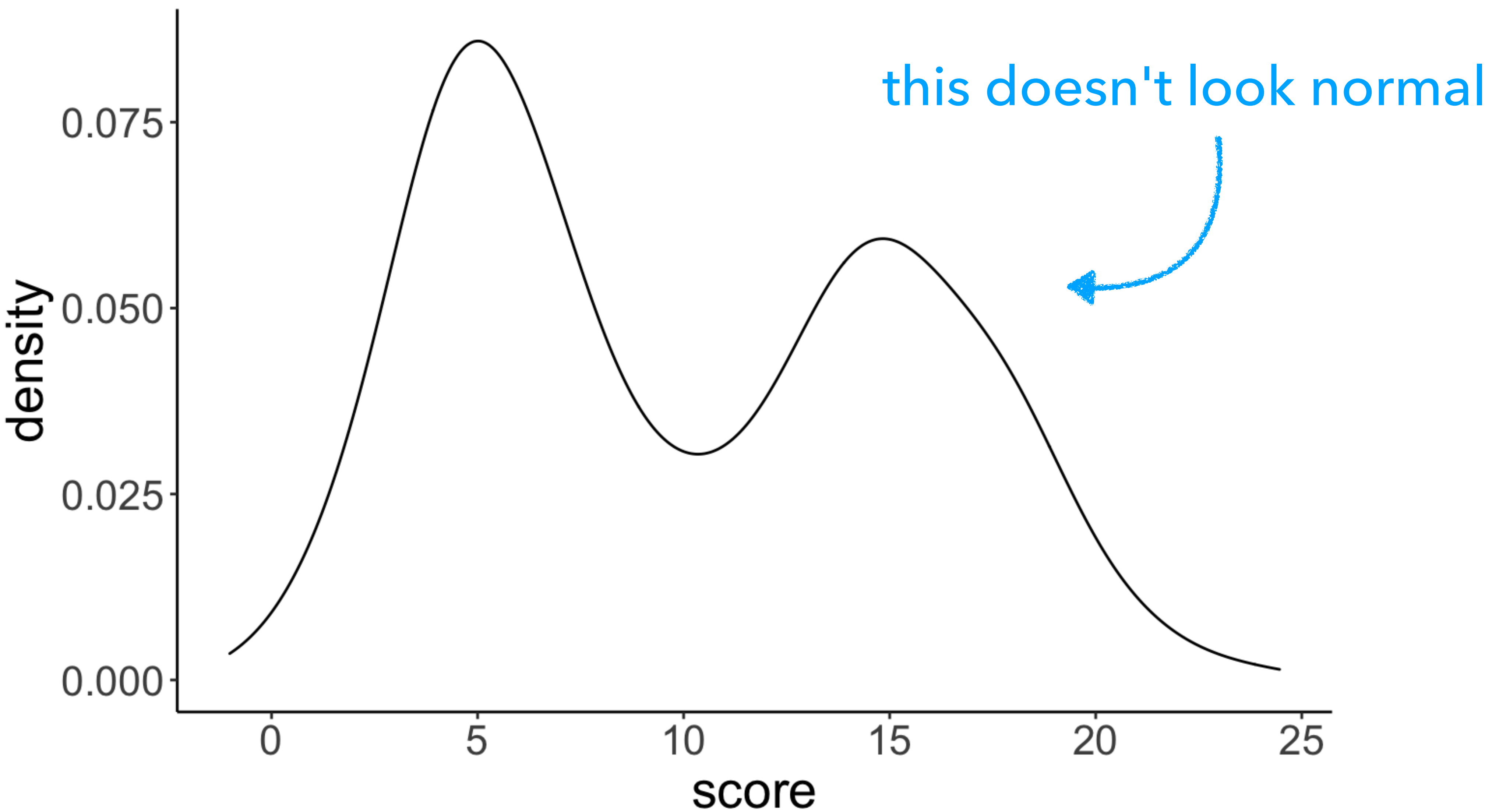
`Error = Data - Model`

assumed to
be normally
distributed

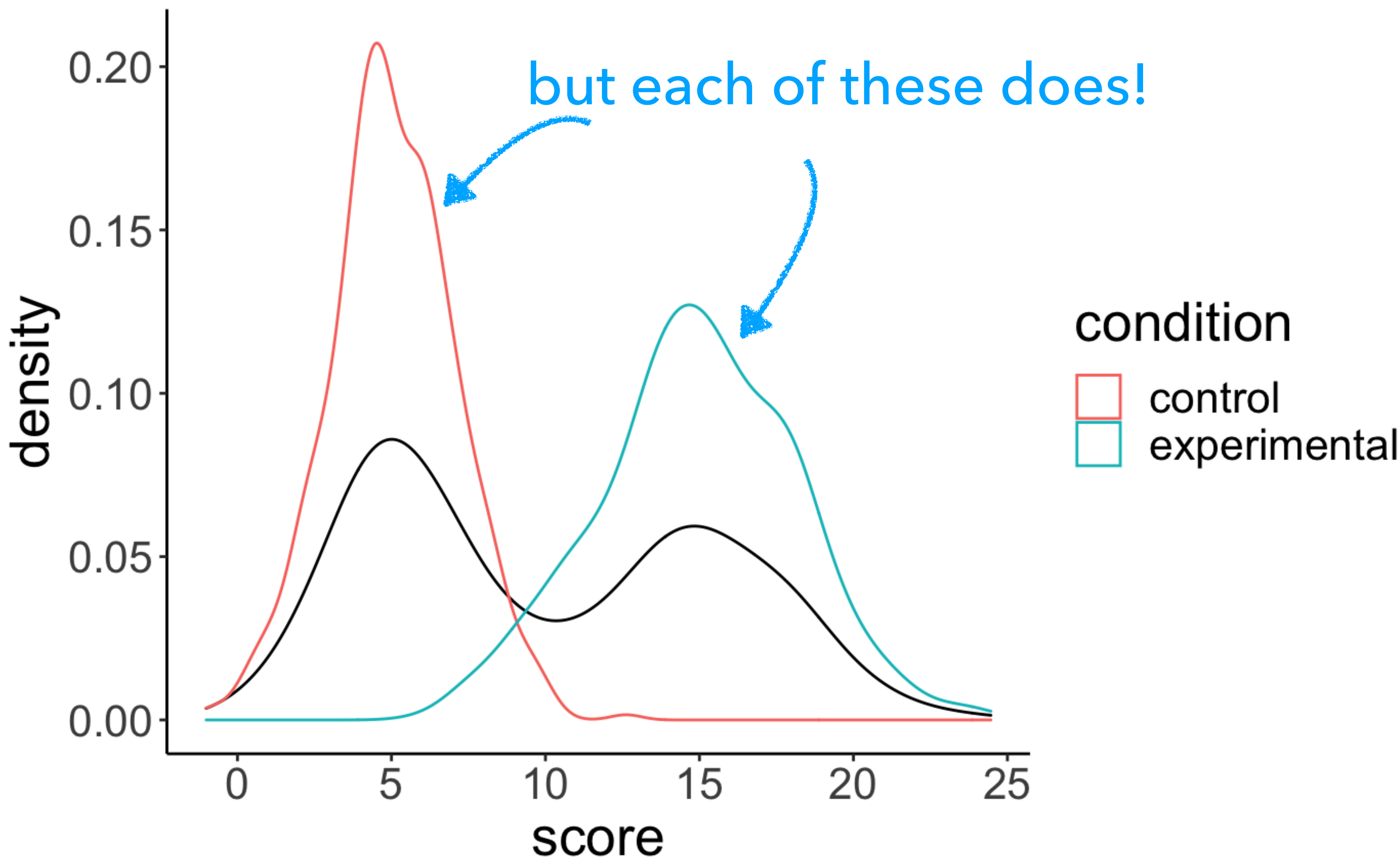
don't need to
be normally
distributed!!

This is the CLT in action!

Distribution of **data**

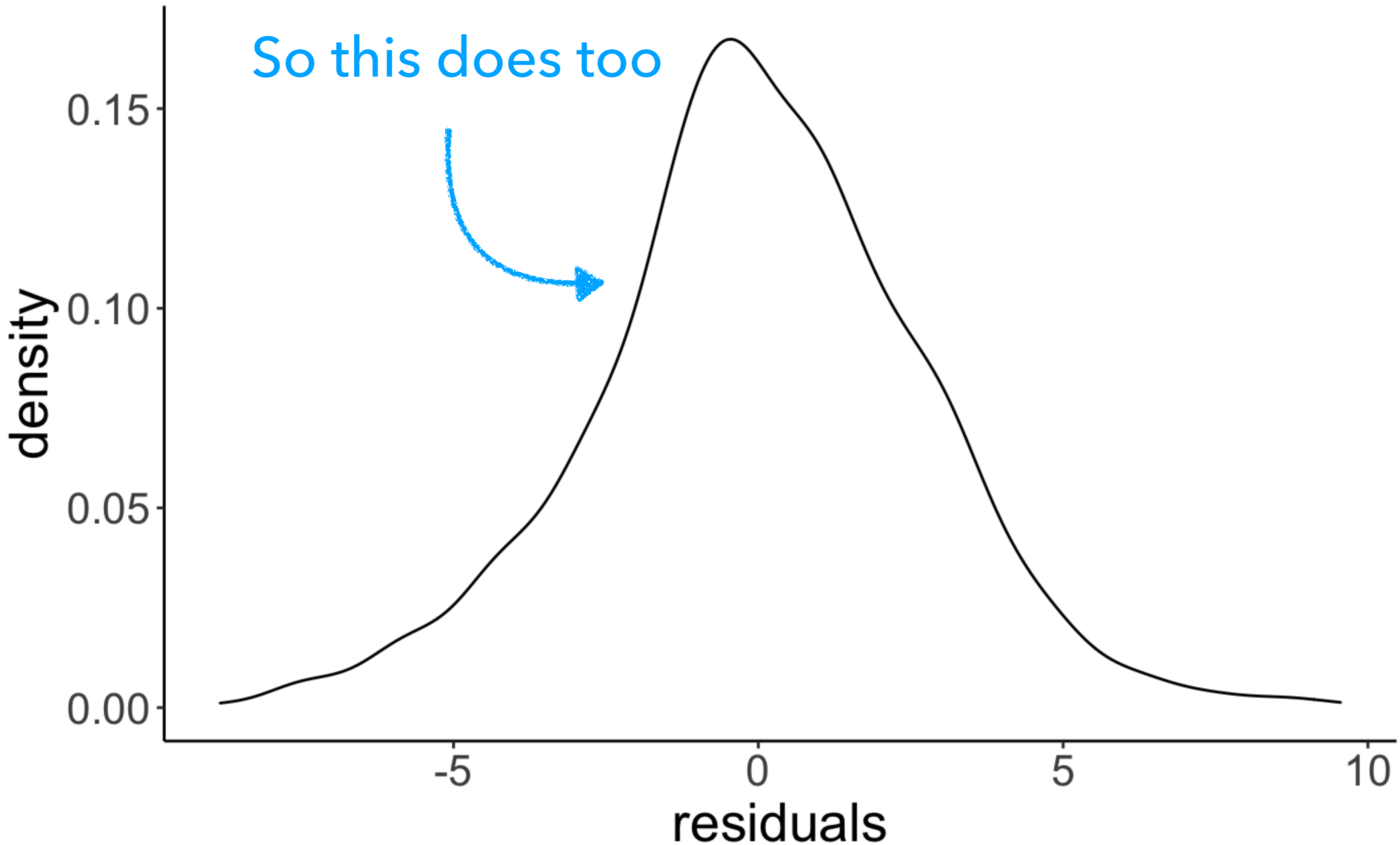


Distribution of **data**



Distribution of **residuals**

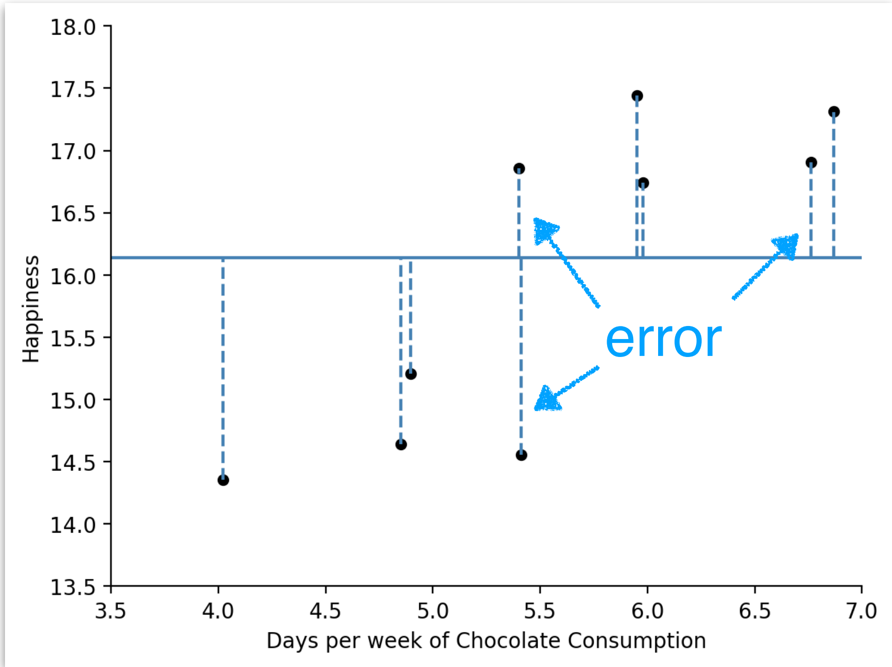
Error = Data - Model



Error = Data - Model

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

square these
and add them up



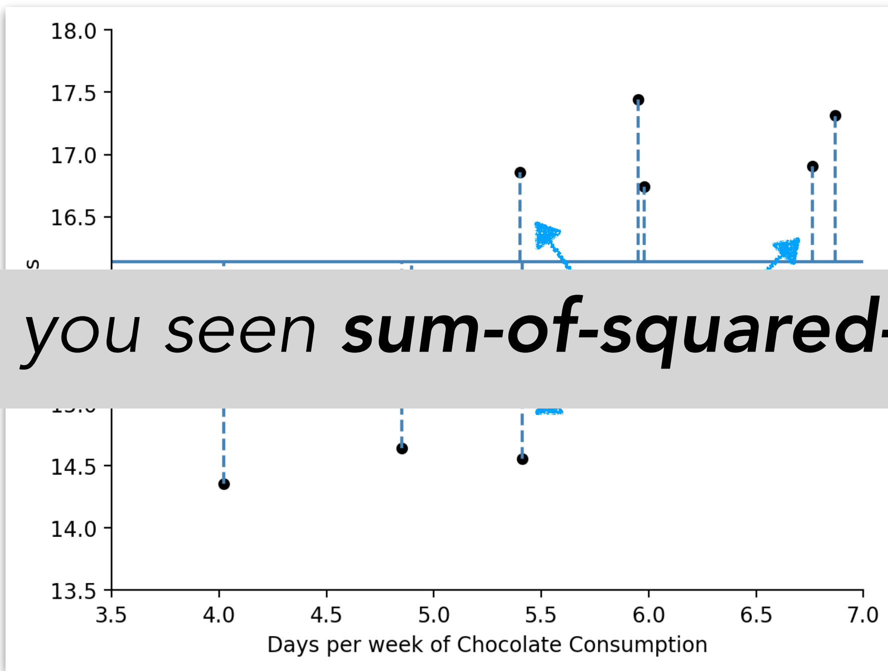
why squared error?

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

Error = Data - Model

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

square these
and add them up



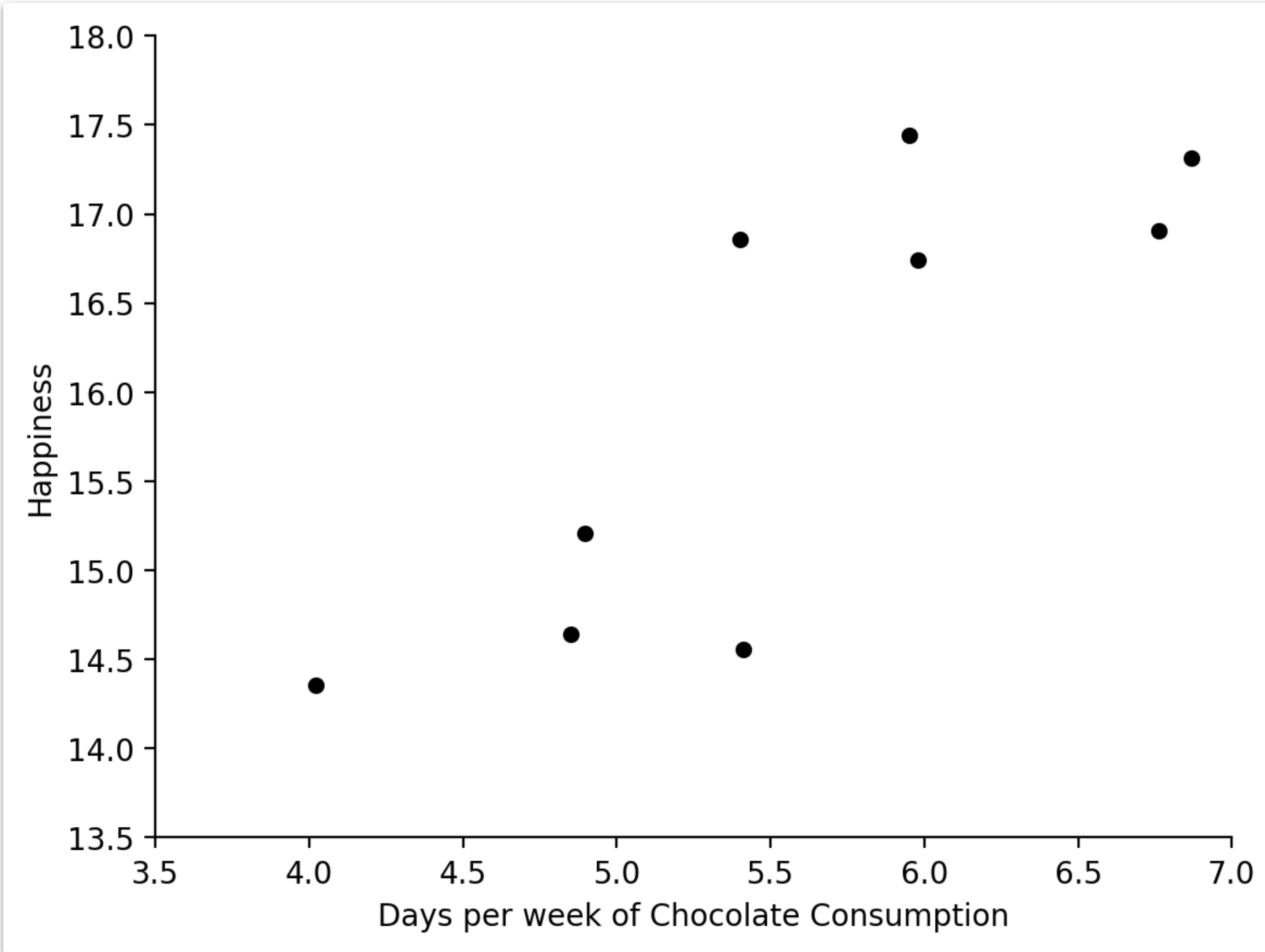
*...where have you seen **sum-of-squared-errors** before....*

why squared error?

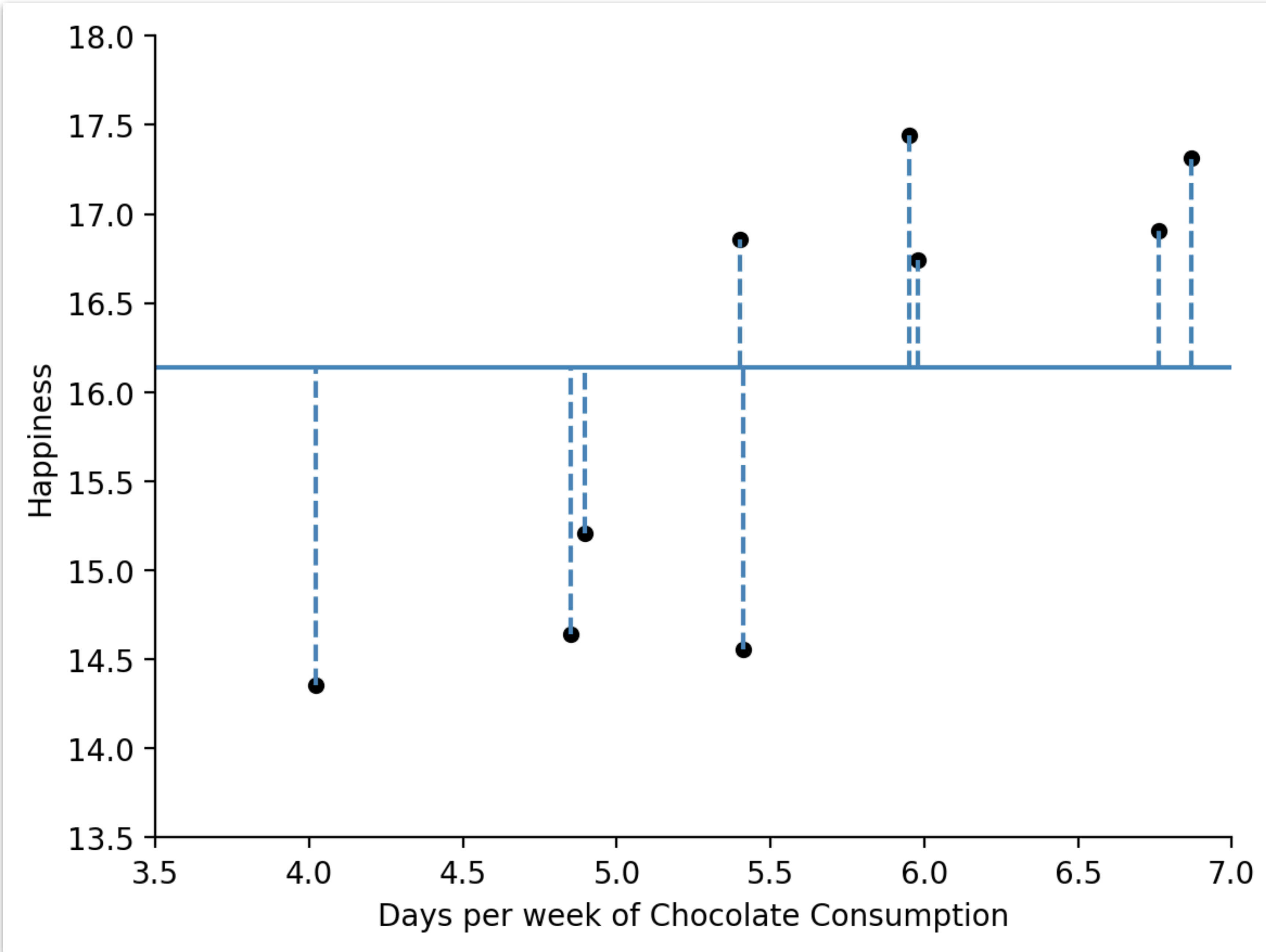
- positive and negative prediction errors don't cancel out
- larger errors are weighted more

The mean as a model

Is there a relationship between chocolate consumption and happiness?

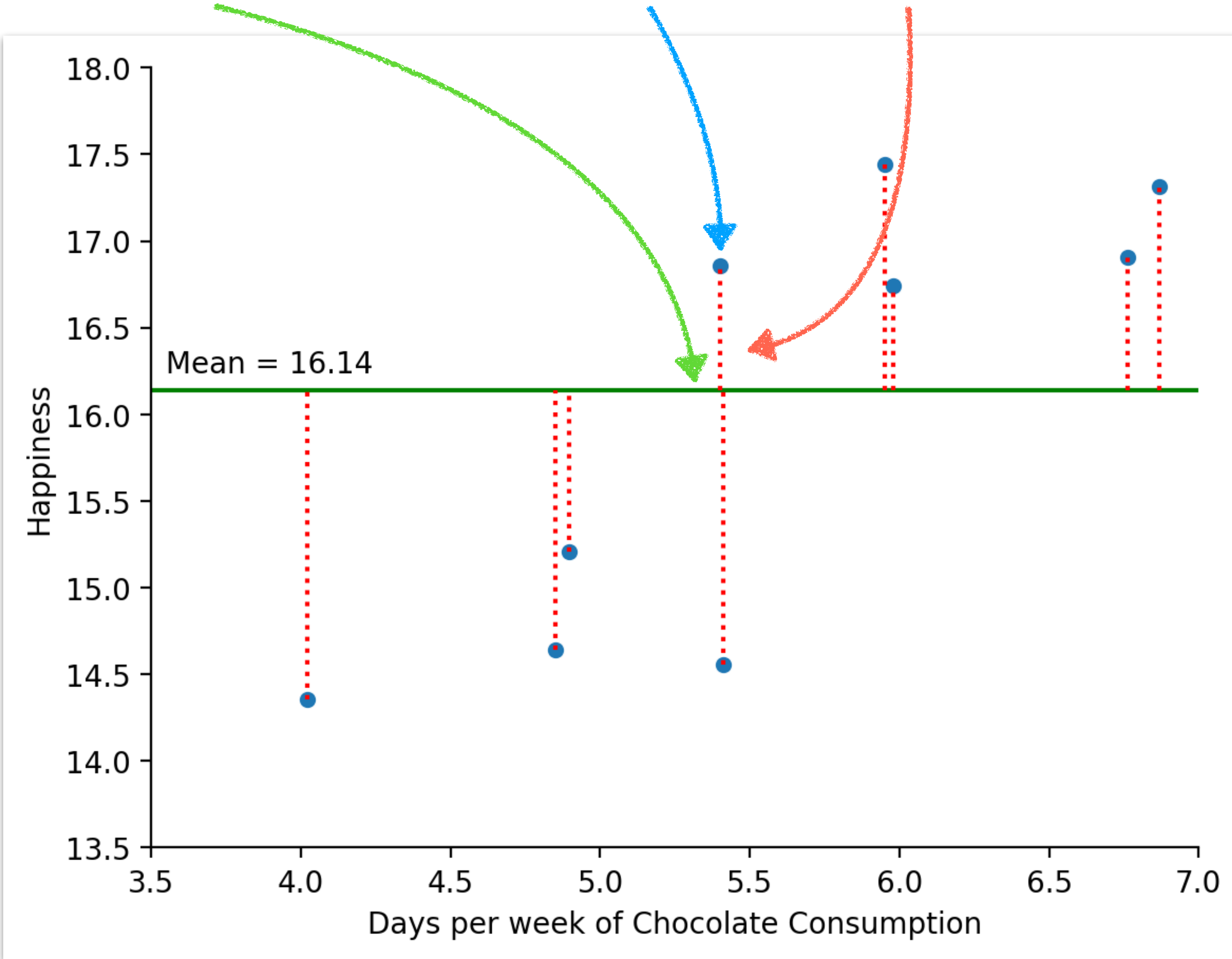


Is there a relationship between chocolate consumption and happiness?



The ***mean*** as a **model** of happiness:

$$\text{happiness}_{\text{prediction}} = \text{mean(happiness)} + \text{error}$$



The ***mean*** as a **model** of happiness:

```
Data = Model + Error
```

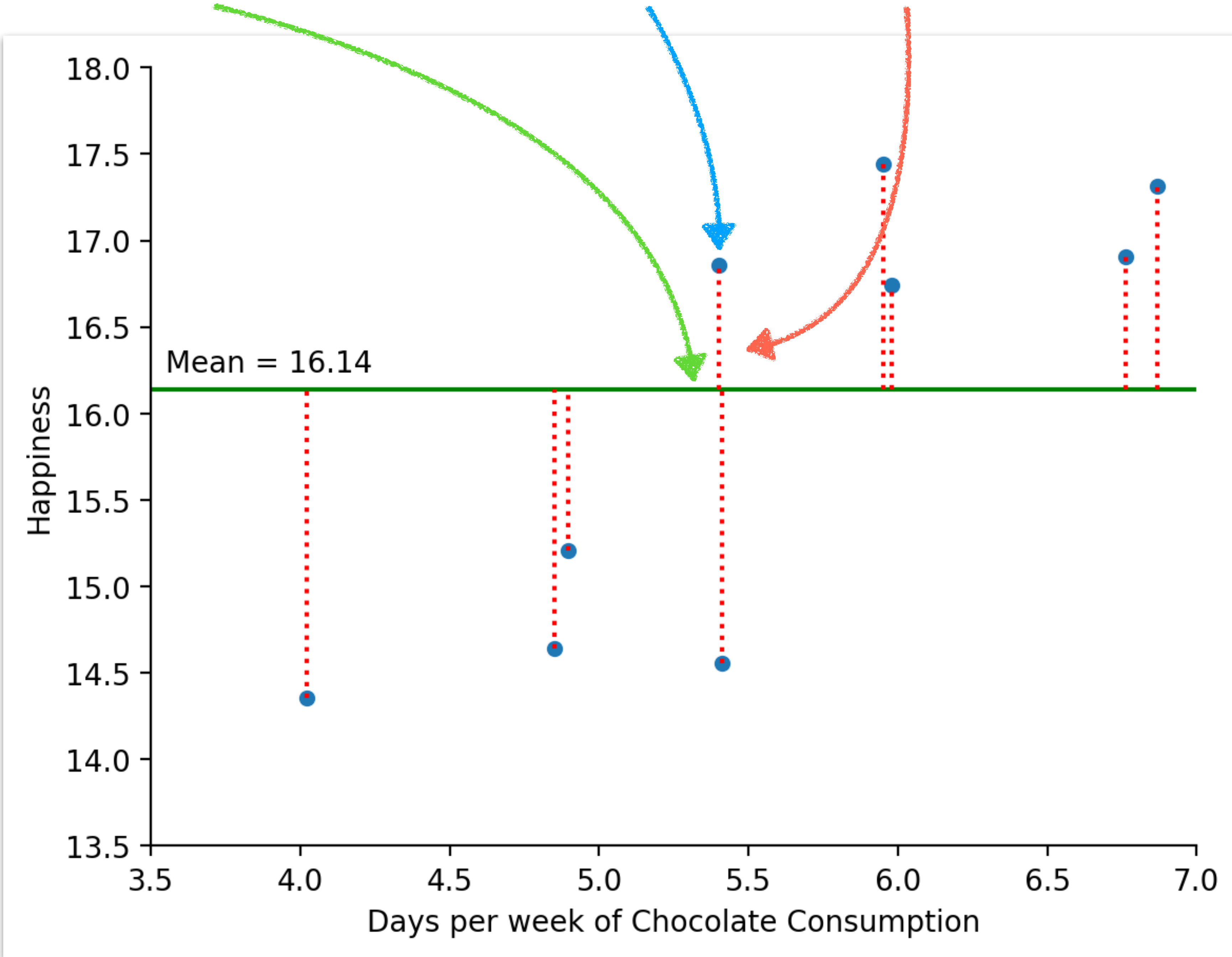
```
happiness_prediction = mean(happiness) + error
```

$$Y_i = \beta_0 + \epsilon_i$$

estimated from data

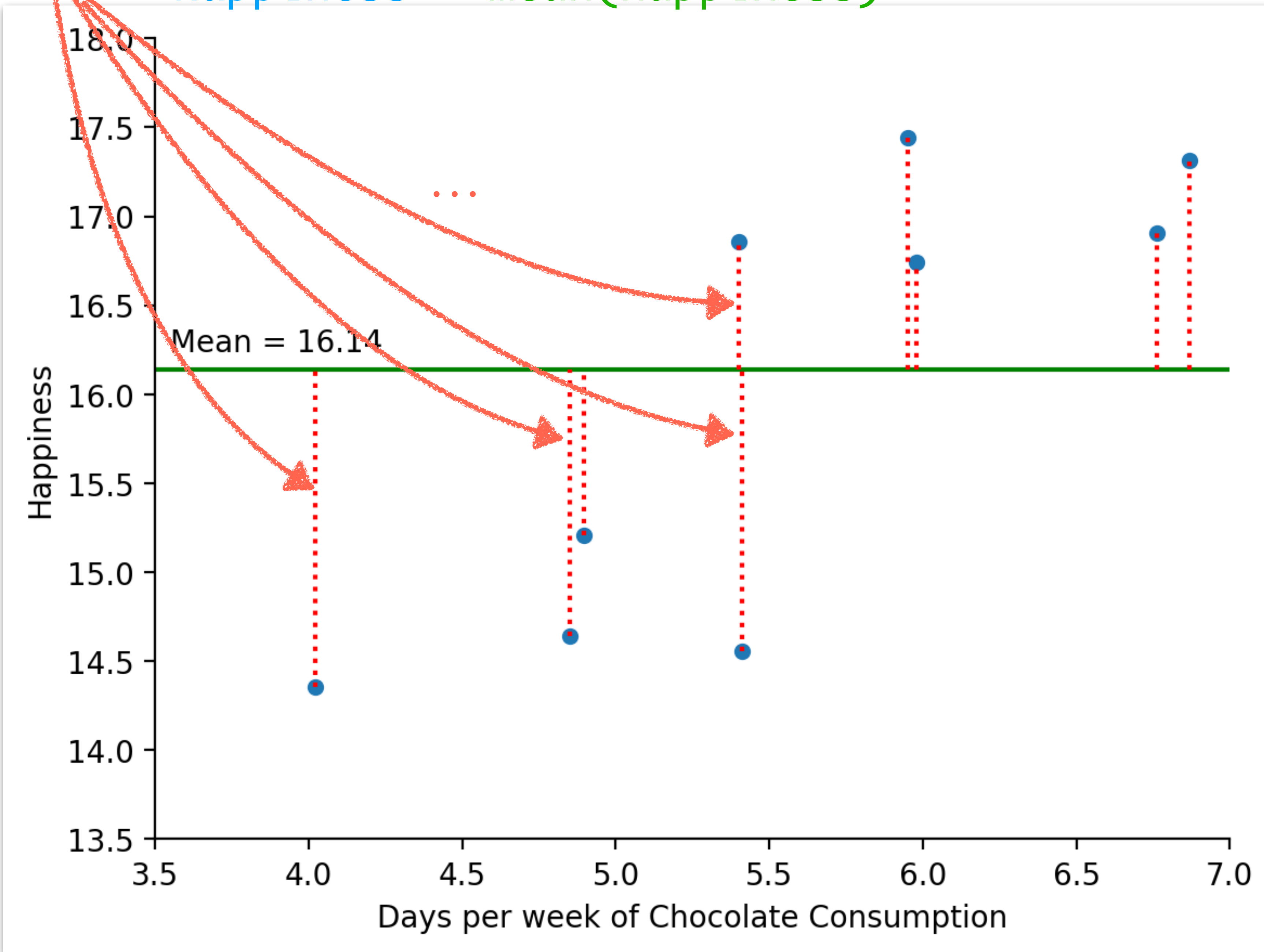
The ***mean*** as a model of happiness:

$\text{happiness}_{\text{prediction}} = \text{mean}(\text{happiness}) + \text{error}$



The ***variance*** as the average ***error*** of the model

```
error = happiness - happiness_prediction
      = happiness - mean(happiness)
```



The ***variance*** as the average ***error*** of the model

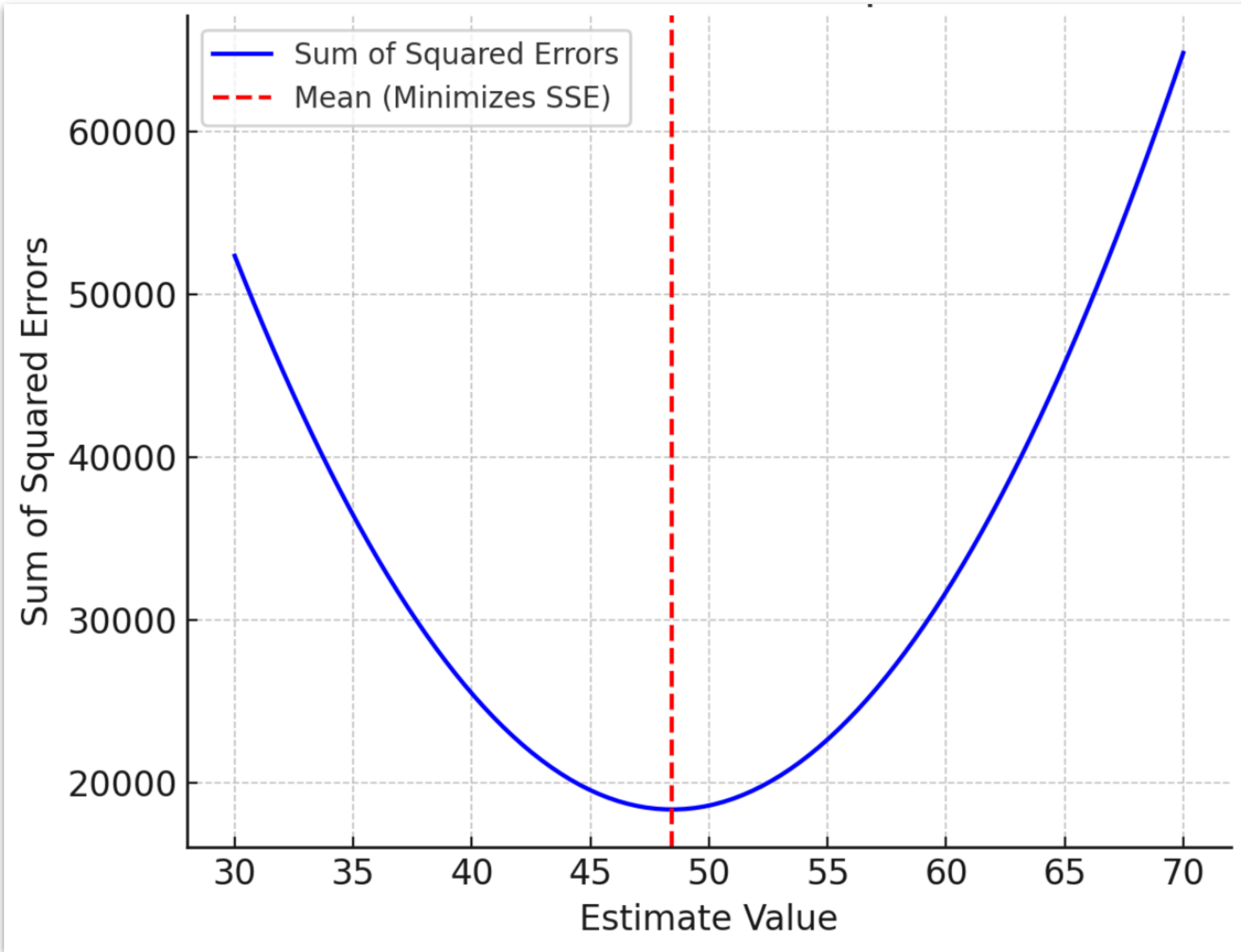
model = sample mean $\bar{x}$

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N} = \frac{\text{SSE}}{N} = MSE$$

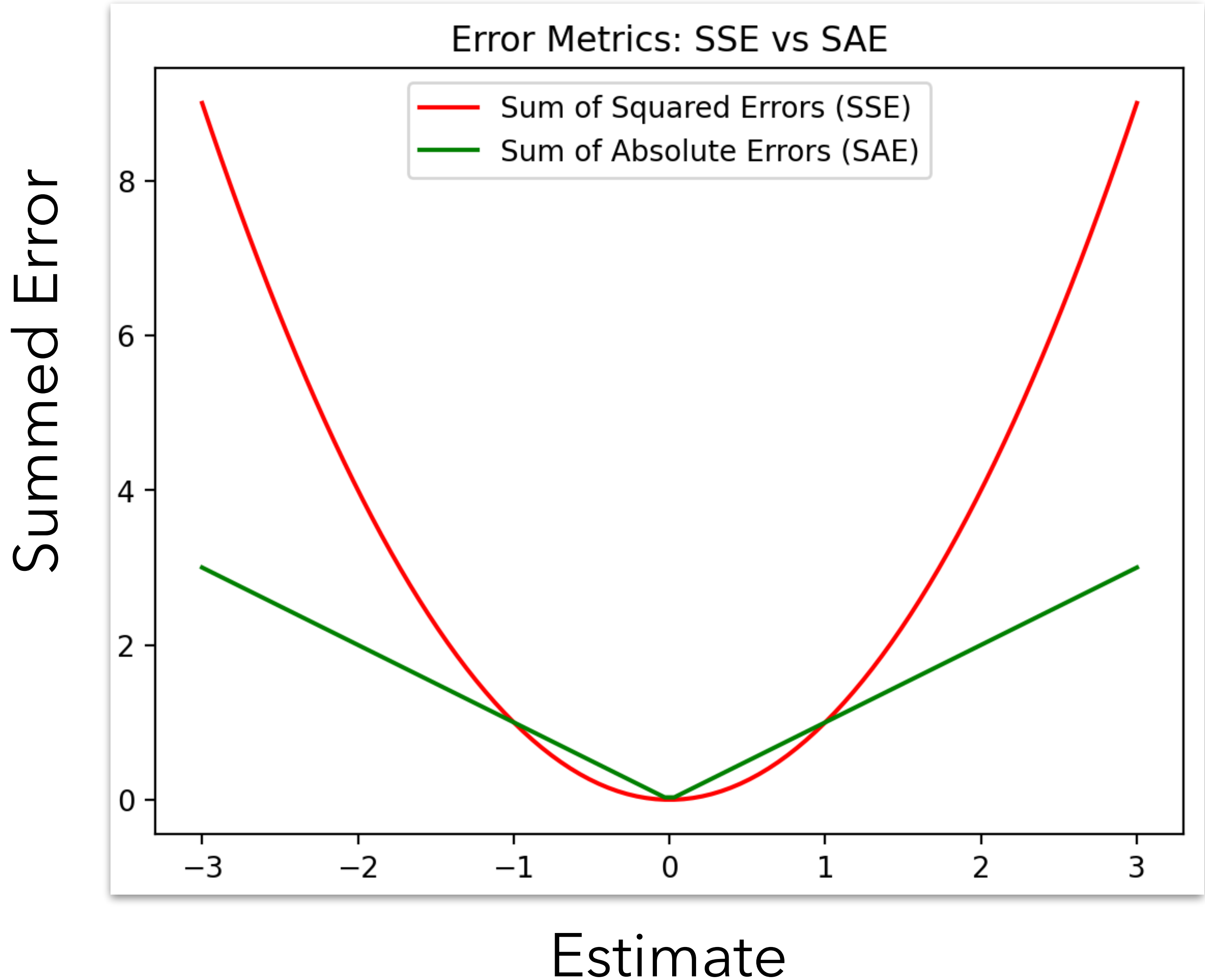
that's variance

error =
average distance² of
each data point from
model

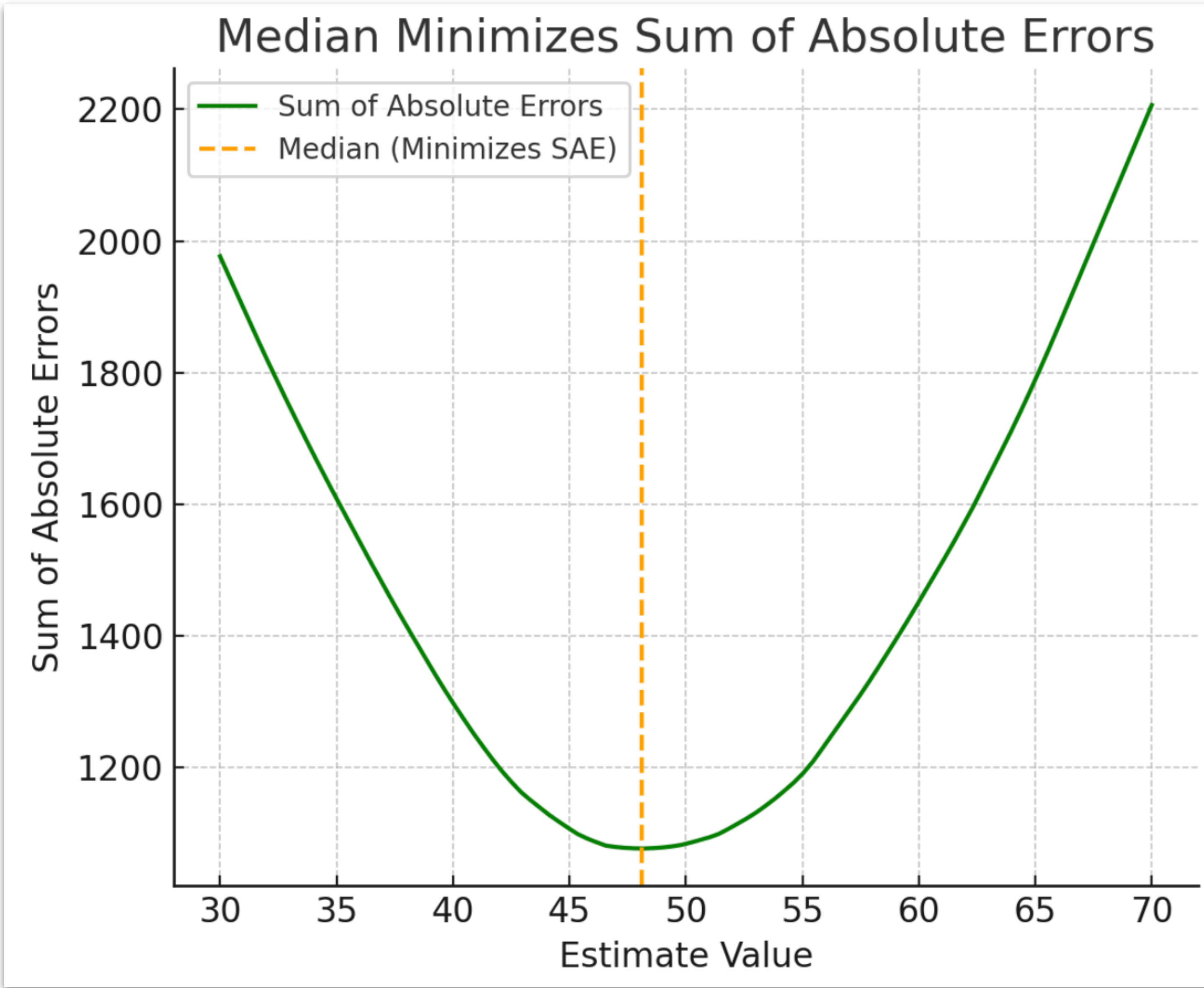
The mean is the **best 1 parameter model** when best = minimize sum-of-squared error



We could calculate error differently though… what about absolute value *instead* of squaring?



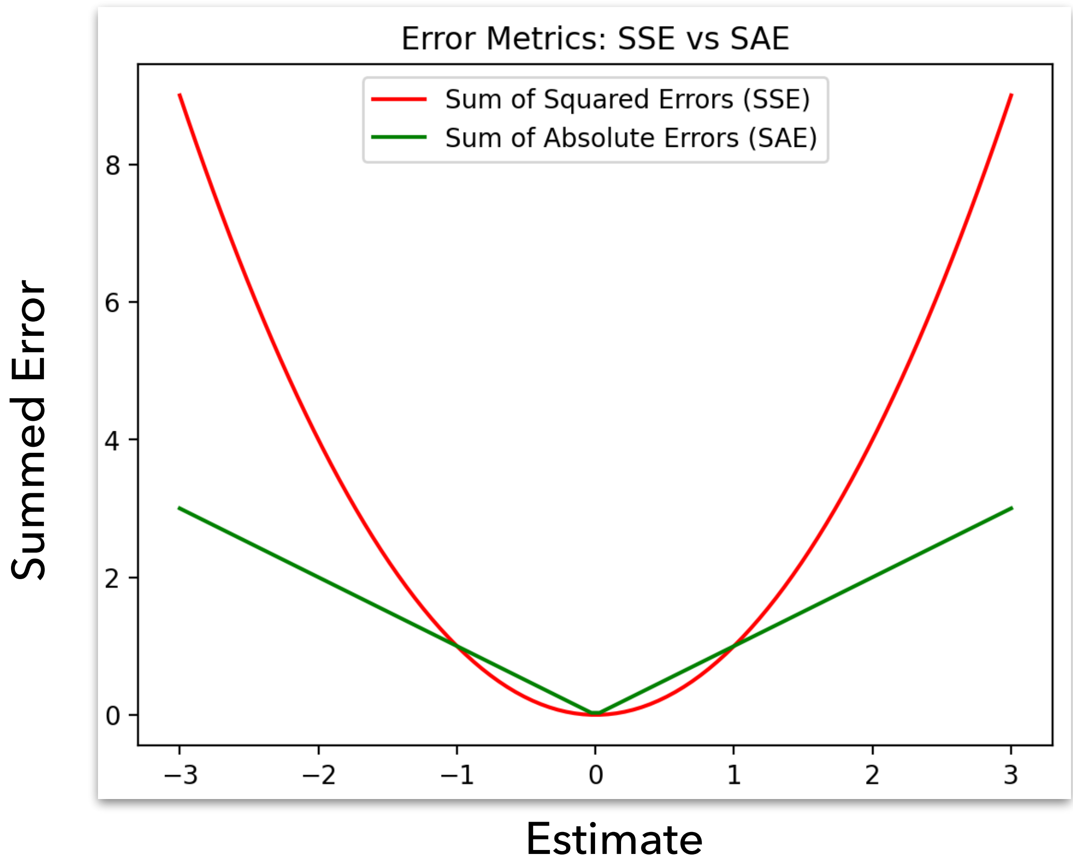
The median is the **best 1 parameter model** when best = minimize sum-of-absolute error



Cultural Integration:

Data outliers == worst model predictions

- **Sum-of-Squared-Errors (SSE)**
  - Grows quadratically with worse prediction — **why** it's **sensitive to outliers**
- **Sum-of-Absolute-Error (SAE)**
  - Grows linearly with worse prediction — **why** it's more "**robust" to outliers**



Today's Plan

1. Hear me blab for ~20-30min

2. Rest of the time for lab (will post HW in a few)