PSYCH 201B

Statistical Intuitions for Social Scientists

Modeling data III

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

02/05/2025

PSYCH 201B

Statistical Intuitions for Social Scientists

II $^{1}/_{2}$

Modeling data ~~III~~

Franny “ate” my prep time

=

short class today

02/05/2025

Today's Plan

1. Recap
    - General Linear Model
    - Ordinary Least Squares
2. Exploring model assumptions

Recap

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

The **G**eneral **L**inear 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 → $\epsilon^2 = \sum (y - X\hat{b})^2$

1 model for solving *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

1 model for solving *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

1 model for solving *many* equations

2 unknown parameters we're *estimating*

Intercept
linear algebra for "mean"

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

Dependent Variables

Independent Variable(s)

1 model for solving *many* equations

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

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

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	$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$	coefficients betas slopes effects parameter- estimates

Regression(s) are just "flavors" of GLM

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

Simple (univariate) regression

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

one predictor

one column wide

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

Regression(s) are just "flavors" of GLM

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

Multiple regression

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

many predictors

many columns wide

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

Regression(s) are just “flavors” of GLM

outcome/target
measurements
output variable

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

predictors/features
design matrix
input variables

$$\begin{bmatrix} \vdots \\ \vec{x}_n \\ \vdots \end{bmatrix}$$

effects
parameter-
estimates

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

Analytic approach to estimating **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	$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$	coefficients betas slopes effects parameter-estimates

The General Linear Model (GLM)

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

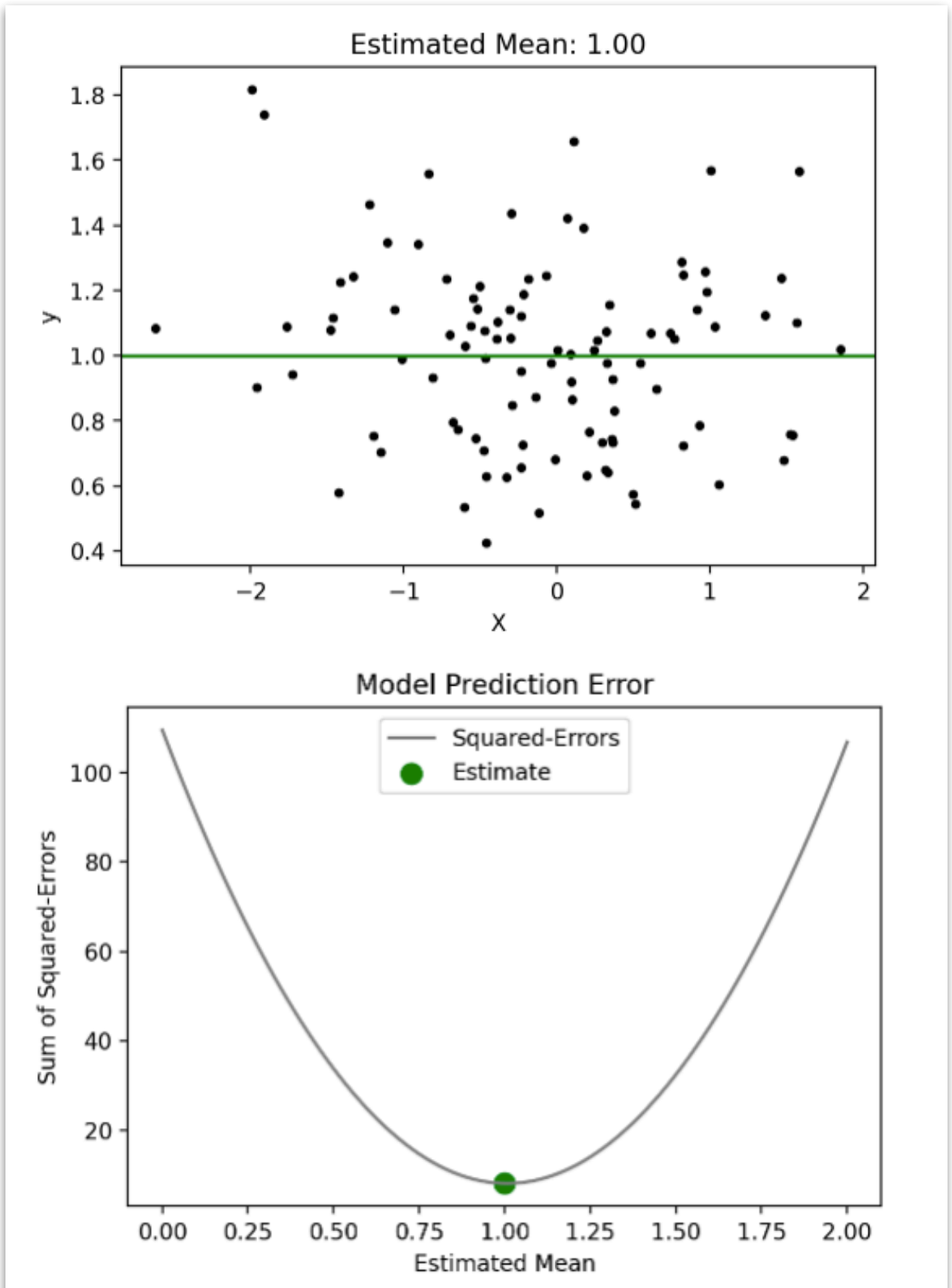
Ordinary Least Squares (**OLS**):
an *analytic* (closed-form) equation for estimating betas that minimizes SSE!

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

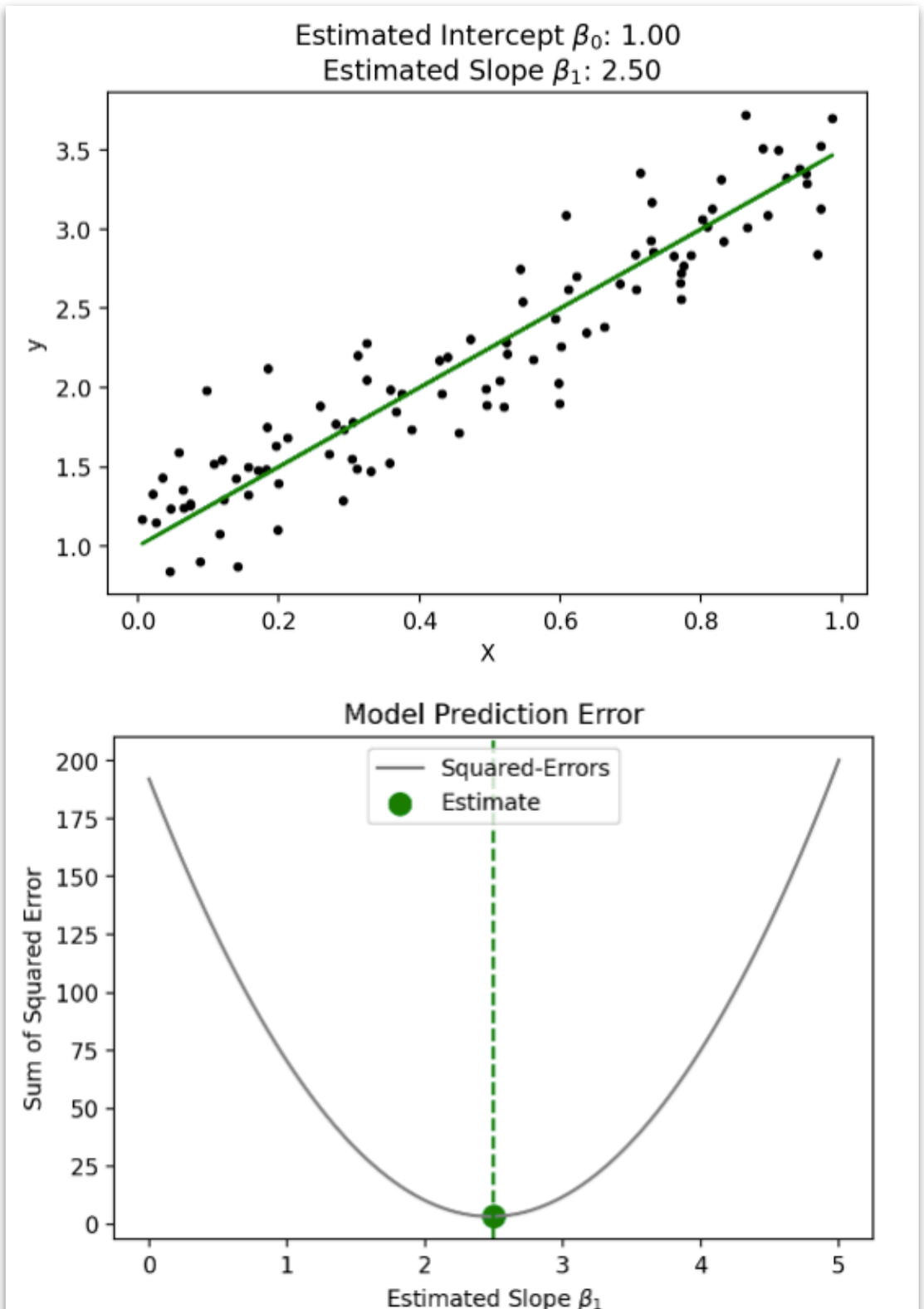
Just like **mean** = best *single* estimate that **minimizes SSE**

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

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



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



What is OLS doing?

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

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?

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

If this gets the slope of *one* IV…

$$\hat{b}_1 = \frac{covariance(x_1, y)}{\sigma_{x_1} \sigma_{x_1}}$$

Even if you don't know linear algebra, try to notice how much these formulas look alike

…the matrix inverse is like getting the slope of *all* IVs together

$$\hat{\beta} = \frac{X^T y}{X^T X}$$

Regression vs correlation

Regression

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

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

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

Regression estimate -> Correlation

Regression

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

Correlation

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

Scale beta by
ratio of **IV-to-DV** standard deviations

$$\hat{r}_{x_1 y} = \hat{\beta}_1 * \frac{\sigma_x}{\sigma_y}$$

Regression estimate -> Correlation

Regression

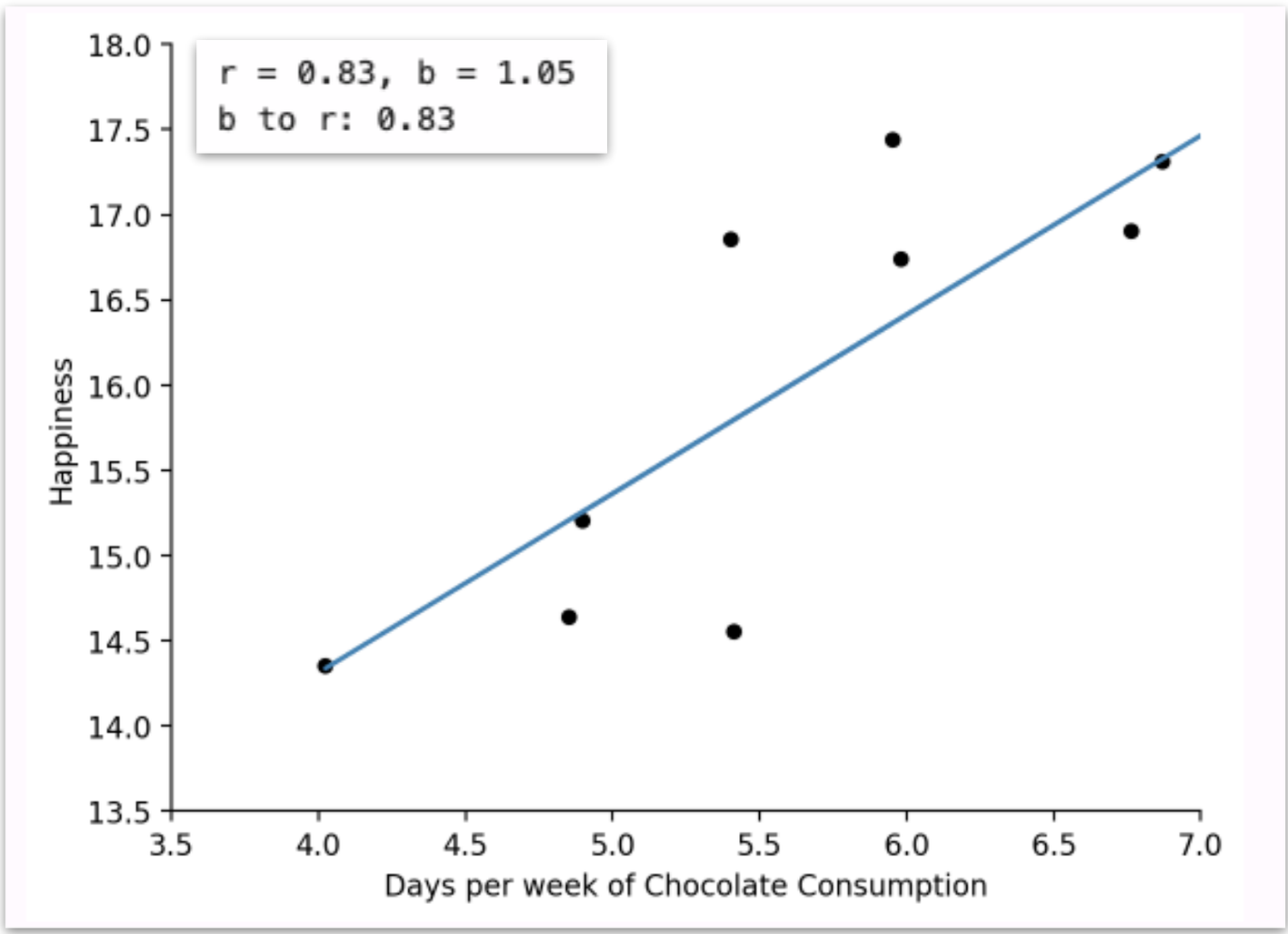
$$\hat{\beta}_1 = \frac{covariance(x_1, y)}{\sigma_{x_1}\sigma_{x_2}}$$

Correlation

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

Scale beta by
ratio of **IV-to-DV** standard deviations

$$\hat{r}_{x_1 y} = \hat{\beta}_1 * \frac{\sigma_x}{\sigma_y}$$



Correlation -> Regression estimate

Regression

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

Correlation

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

Scale correlation by
ratio of **DV-to-IV** standard deviations

$$\hat{\beta}_1 = \hat{r}_{x_1 y} * \frac{\sigma_y}{\sigma_x}$$

Correlation -> Regression estimate

Regression

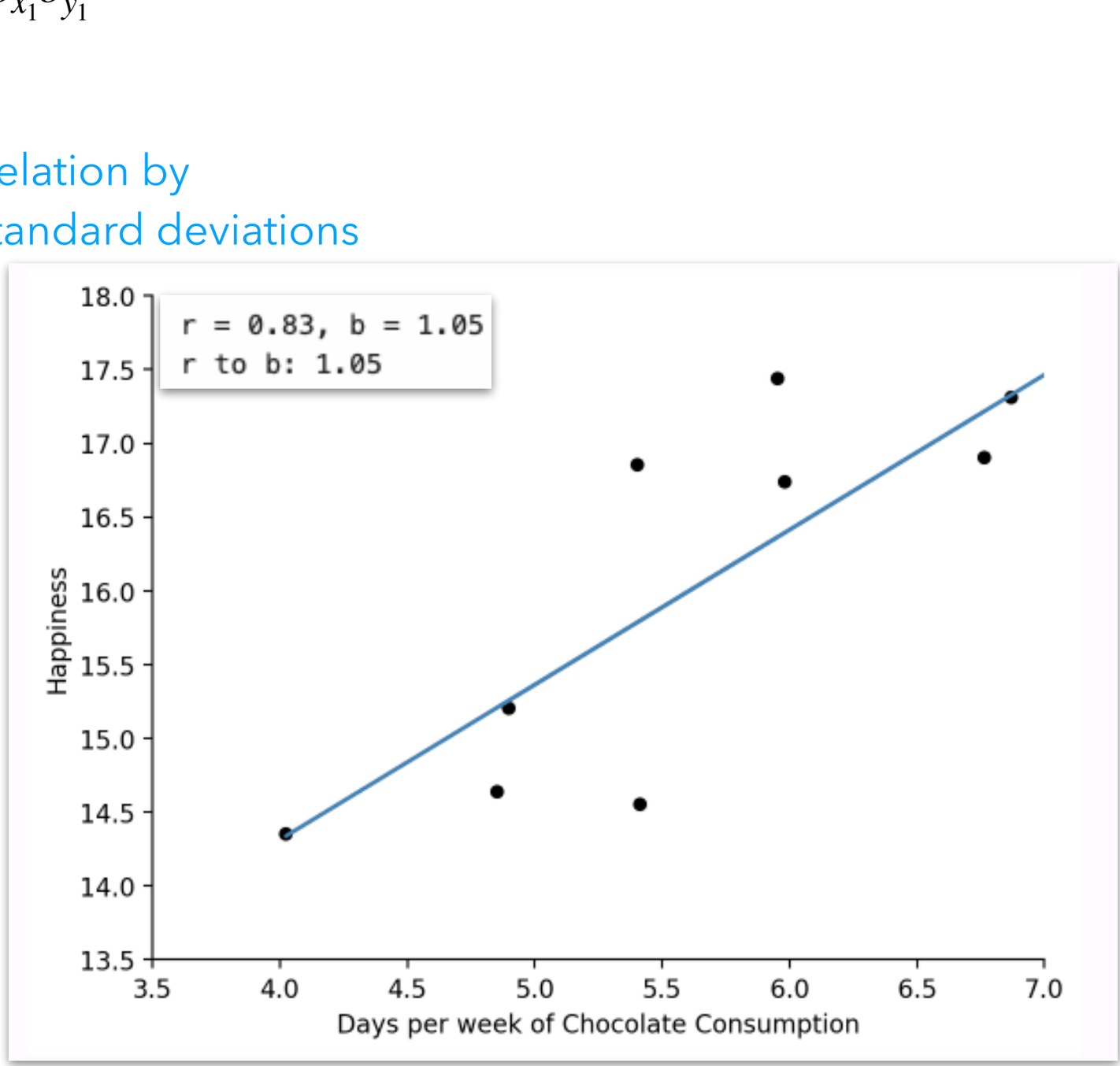
$$\hat{\beta}_1 = \frac{covariance(x_1, y)}{\sigma_{x_1}\sigma_{x_2}}$$

Correlation

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

Scale correlation by
ratio of **DV-to-IV** standard deviations

$$\hat{\beta}_1 = \hat{r}_{x_1 y} * \frac{\sigma_y}{\sigma_x}$$



OLS Summary

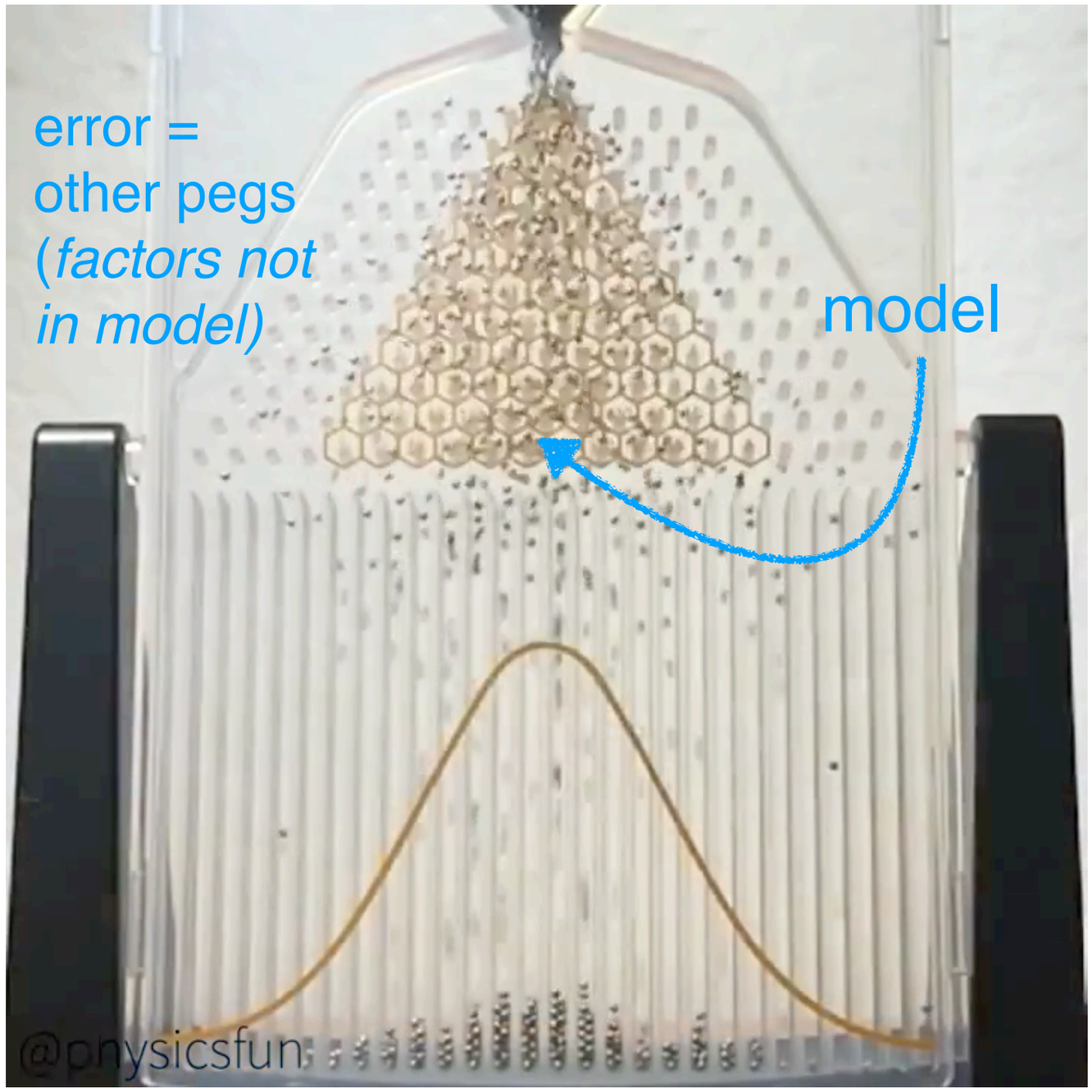
- Analytic approach for finding **best linear relationship** between 1 (or more) IVs and 1 DV
- Best = **minimized** sum-of-squared errors
- Lets us collect our IVs into a **design matrix**
- **Matrix inversion** to remove similarity between IVs (columns of design matrix)
- **Dot product** to calculate similarity DV and IVs after removing similarity between IVs
- Correlation = Regression ***without* accounting for similarity** between IVs

Questions?

Model *assumptions*

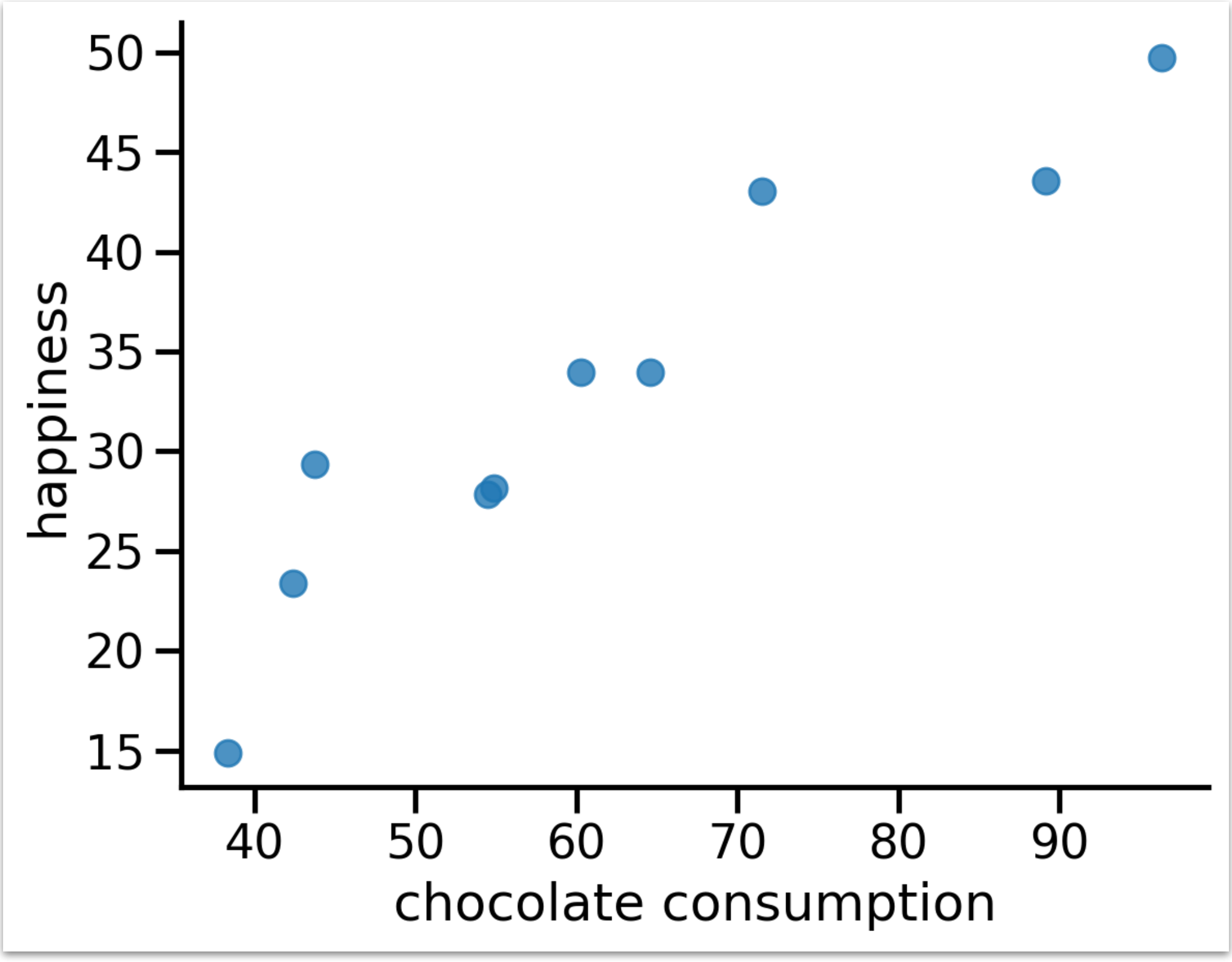
Error = Data - Model

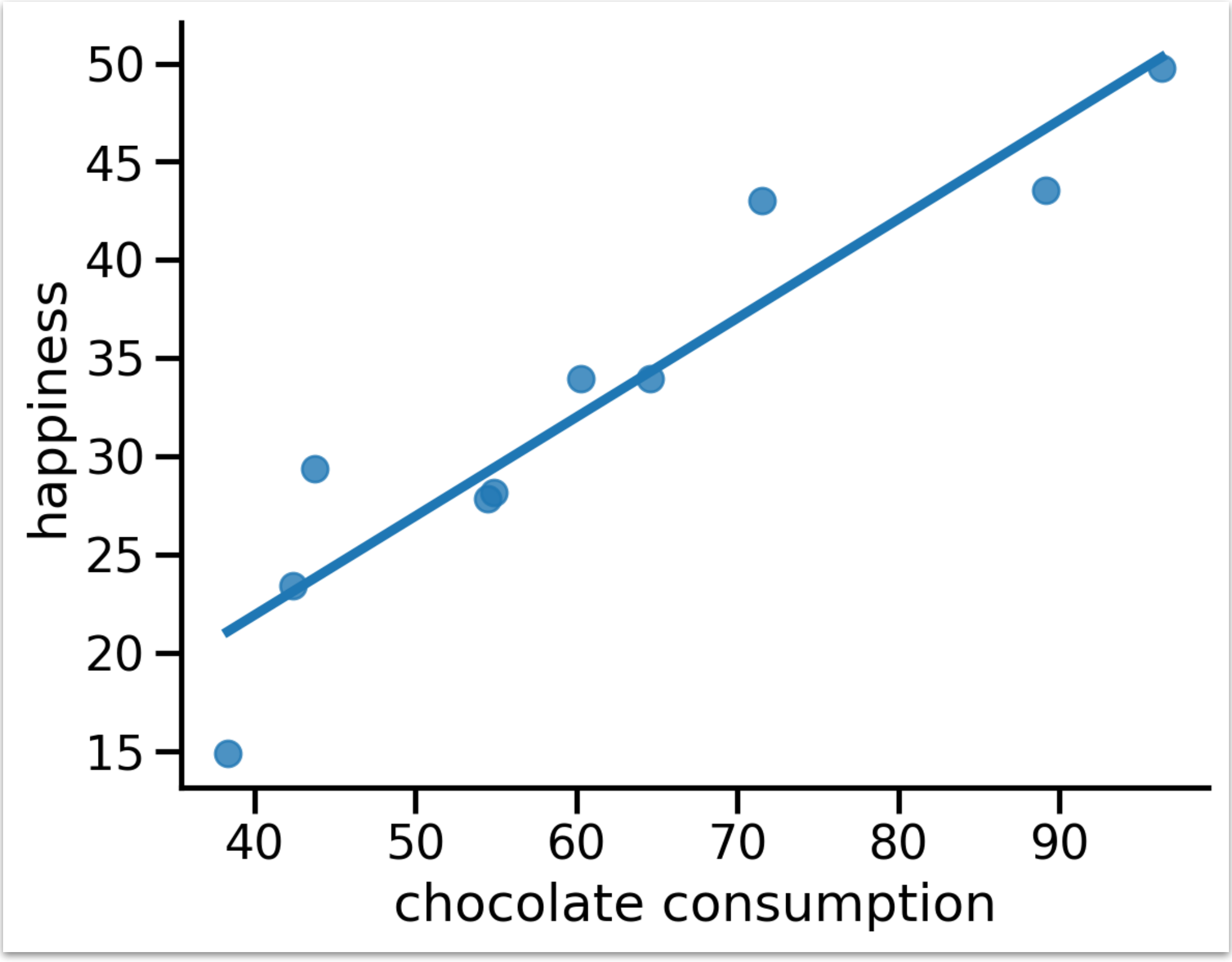
1. We **assume** that the **errors** are due to (*a potentially large number of*) factors that we didn't take into account.

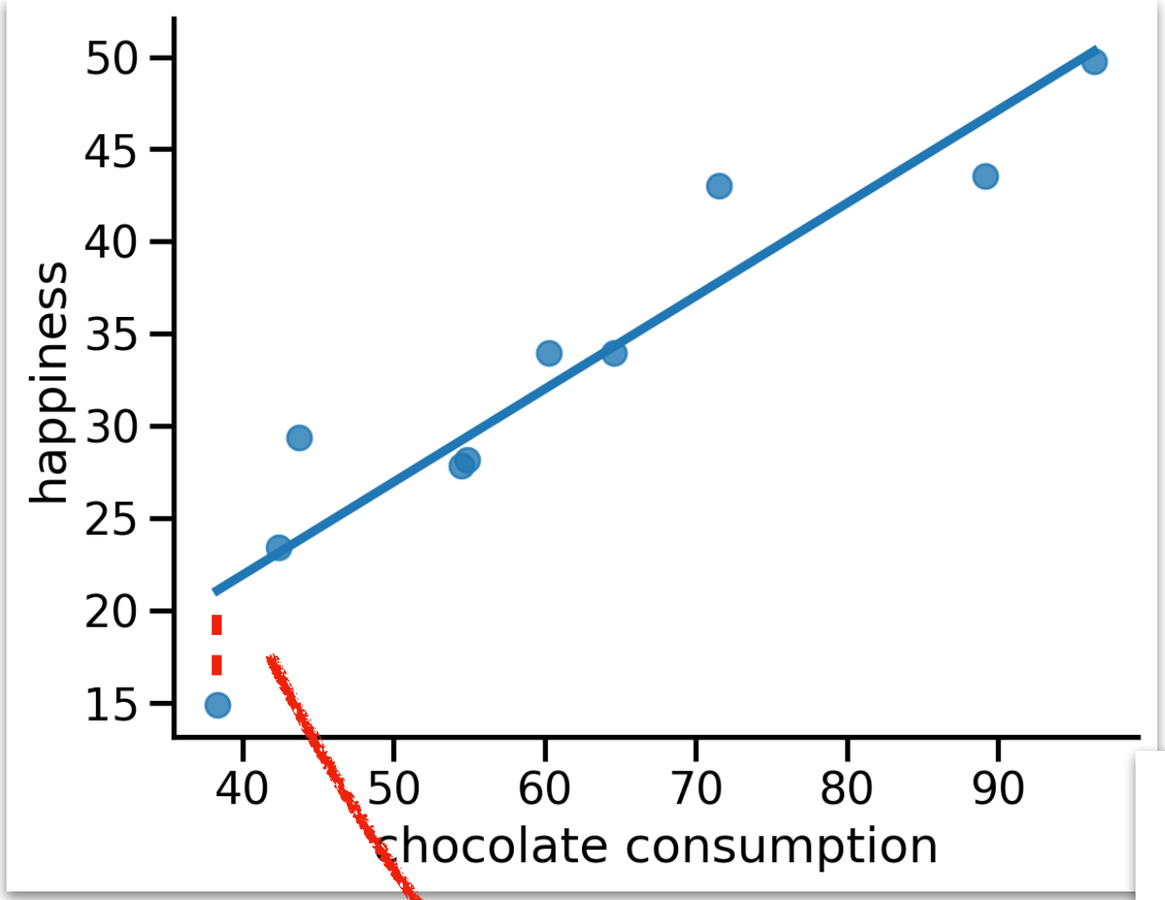


2. We assume that each of these factors influences the data in **an additive way** (some pulling in one, others pulling in another direction).

Result: **Normally Distributed Errors**

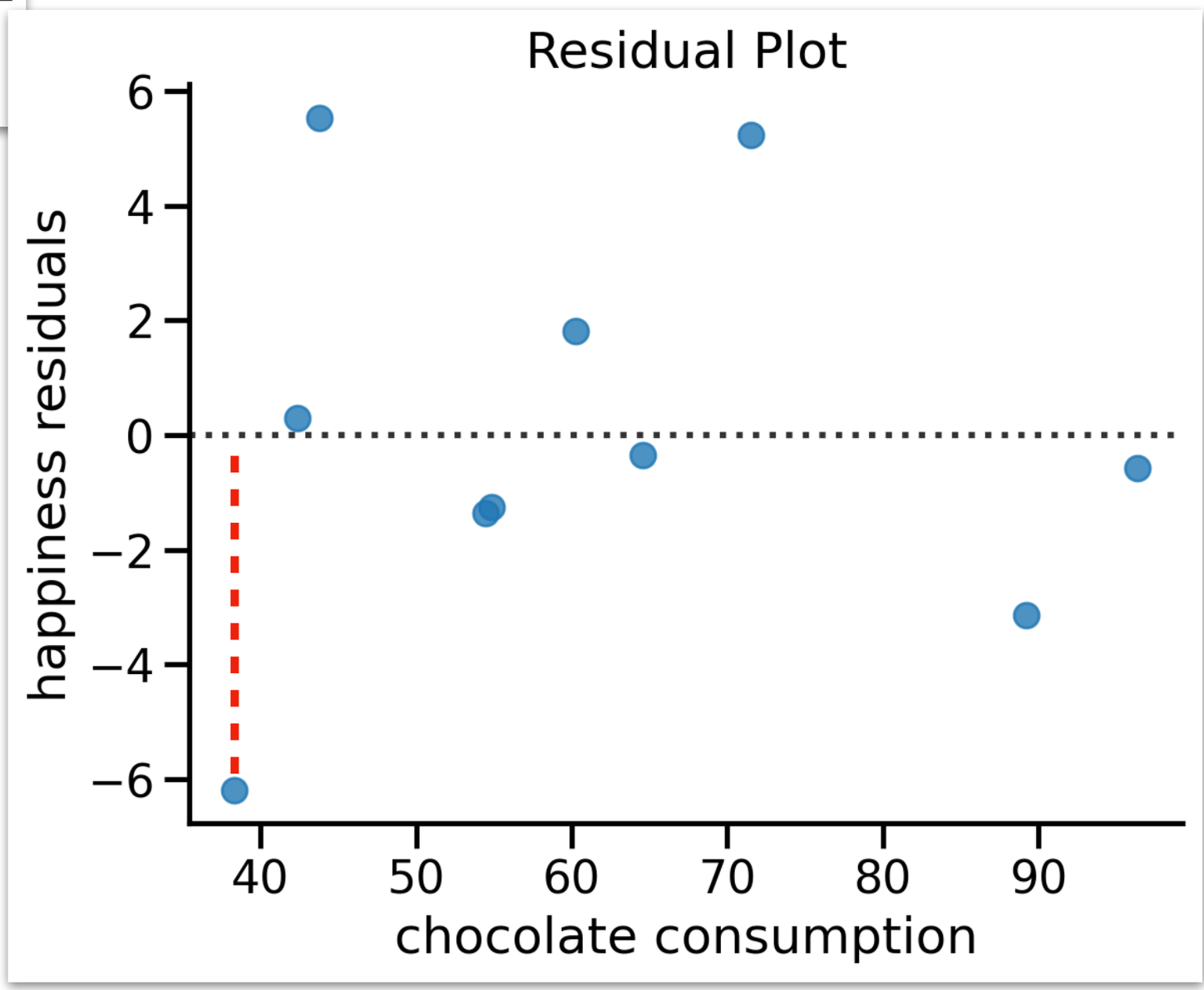

50
45
40
35
30
25
20
15
happiness
40
50
60
70
80
90
chocolate consumption

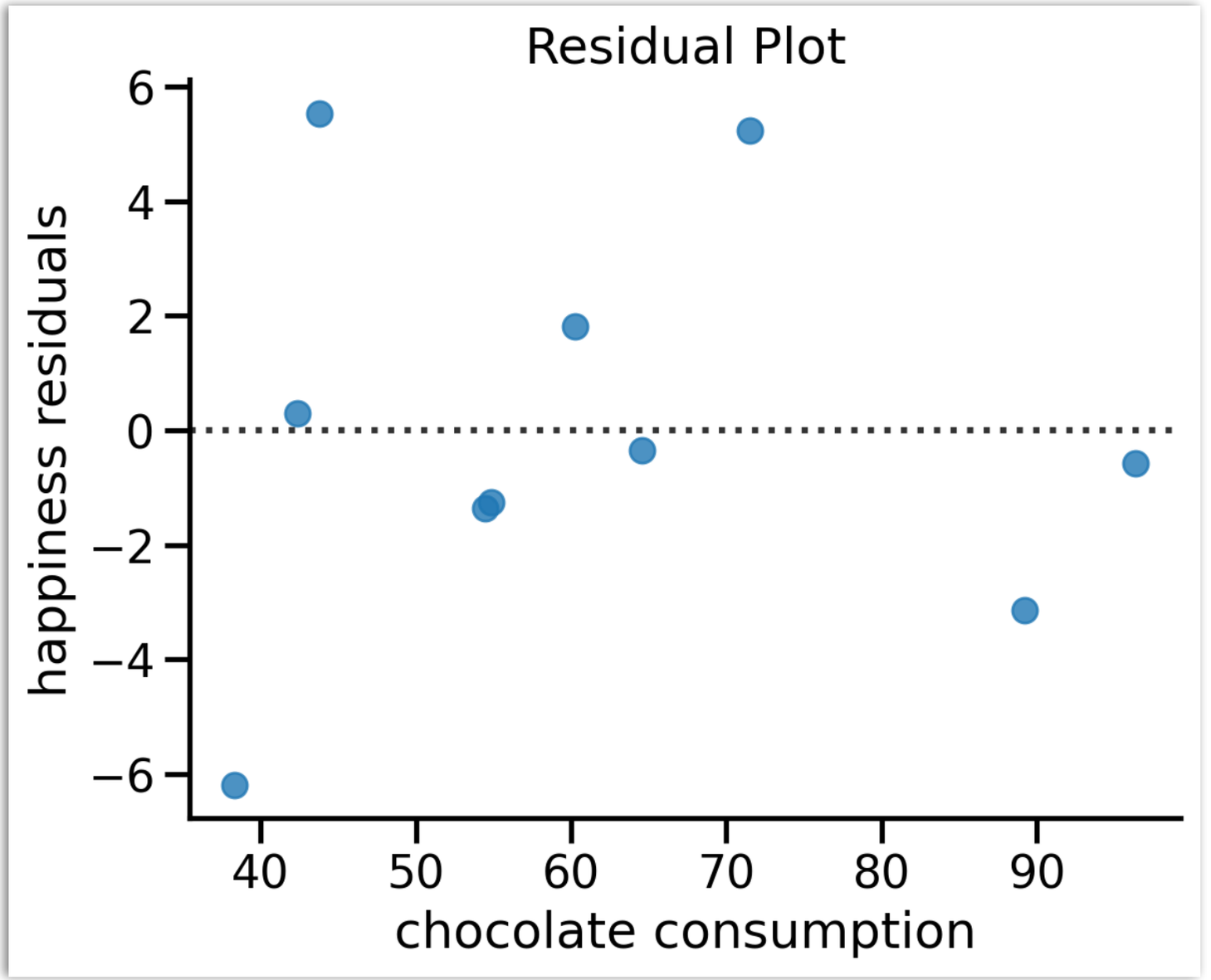

50
45
40
35
30
25
20
15
happiness
40
50
60
70
80
90
chocolate consumption



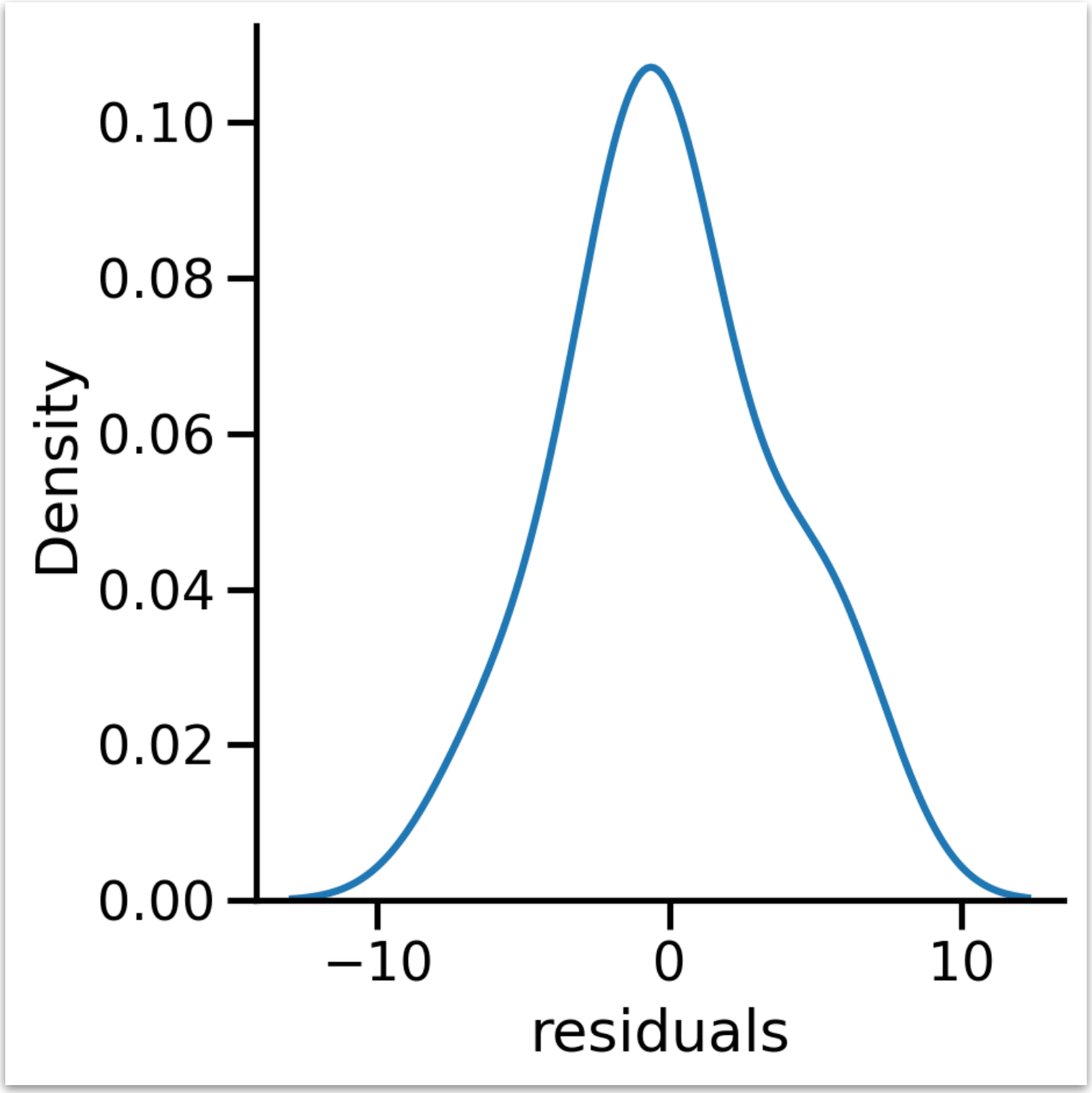
Model errors

Residual plots let us **visually inspect** our assumptions




Residual Plot
6
4
2
0
−2
−4
−6
happiness residuals
40
50
60
70
80
90
chocolate consumption

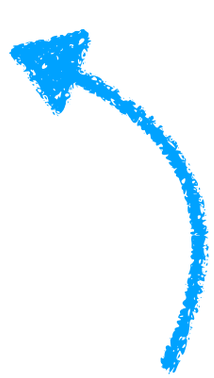
Normally Distributed Errors



Look for **structure** in your residuals

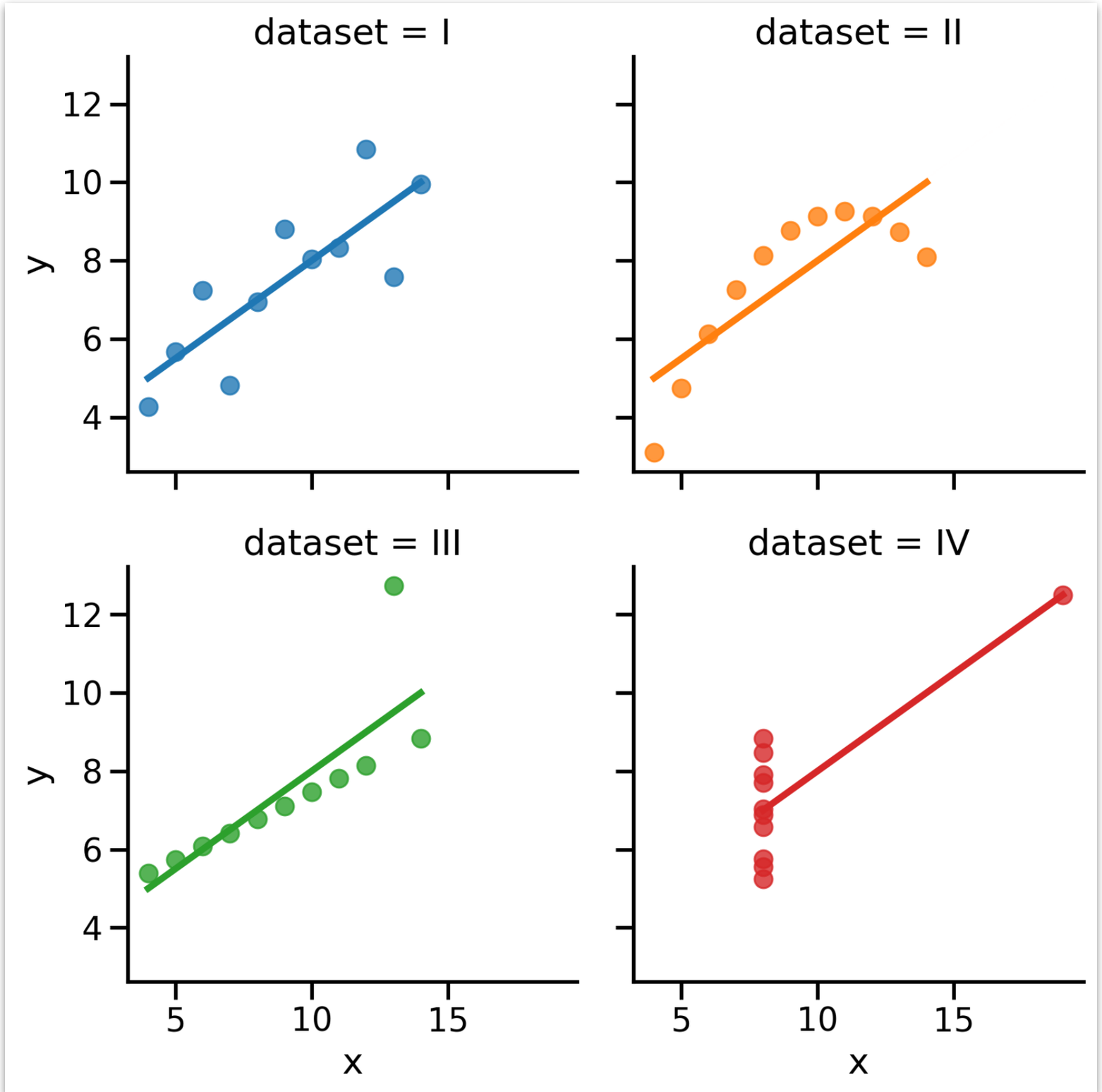
Should be **i**ndependent & **i**dentically **distributed (i.i.d)**

- Independent
- Constant Variance
- Uncorrelated
- Normally distributed

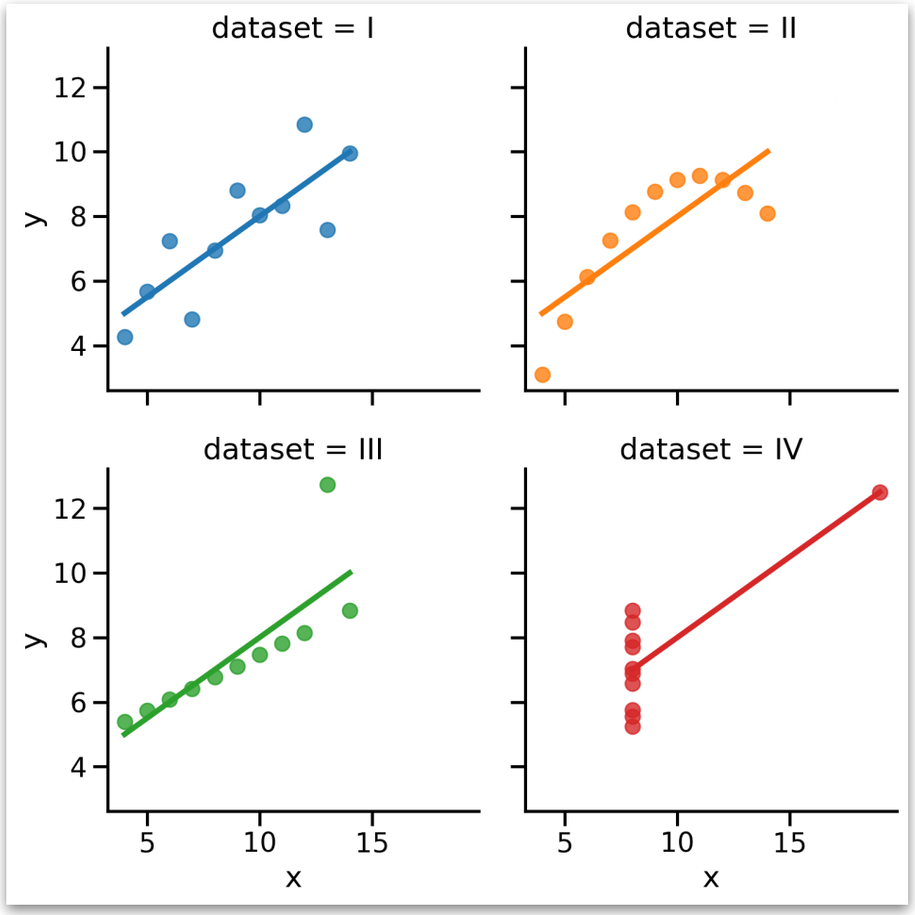


Remember: the dependent variable doesn't need to be normally distributed!

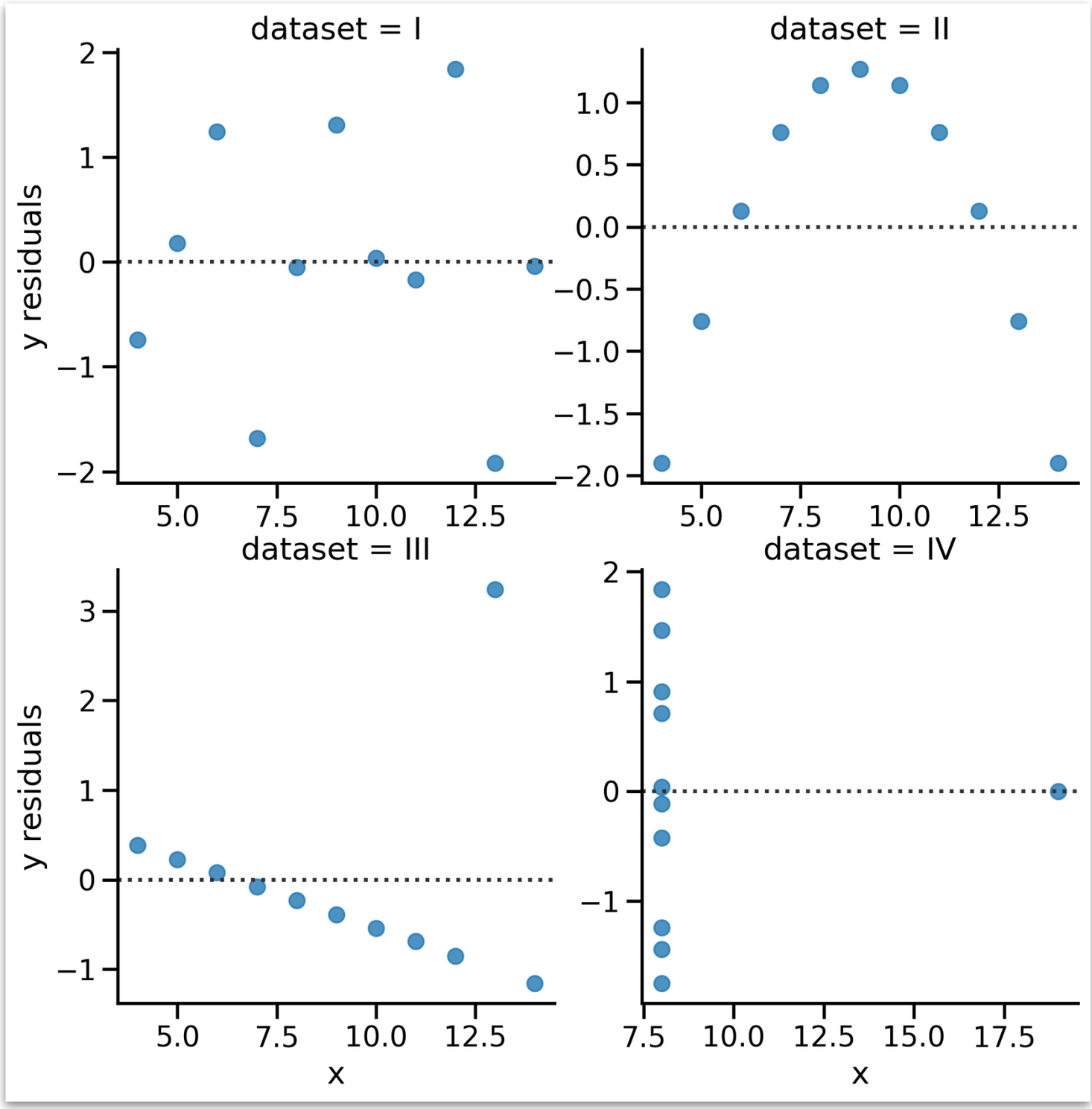
Remember Anscombe's Quartet?



Remember Anscombe's Quartet?



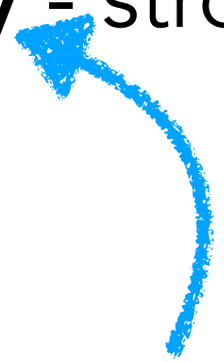
Are these normal?



Look for **structure** in your residuals

Should be **i**ndependent & **i**dentically **distributed (i.i.d)**

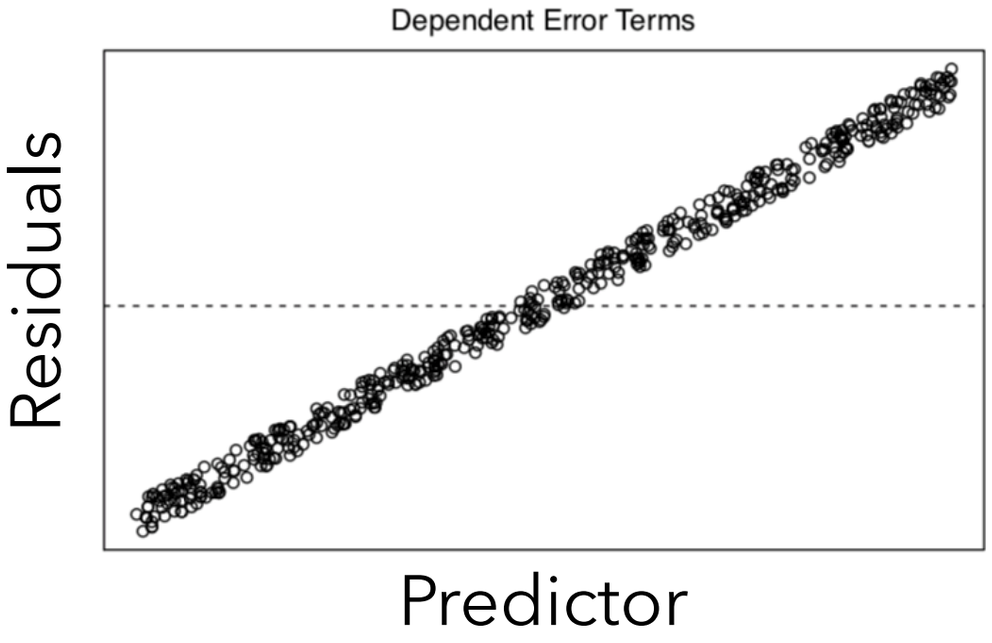
- Independent
- Constant Variance
- Uncorrelated
- Normally distributed
- **No multi-collinearity** - strong correlations between IVs



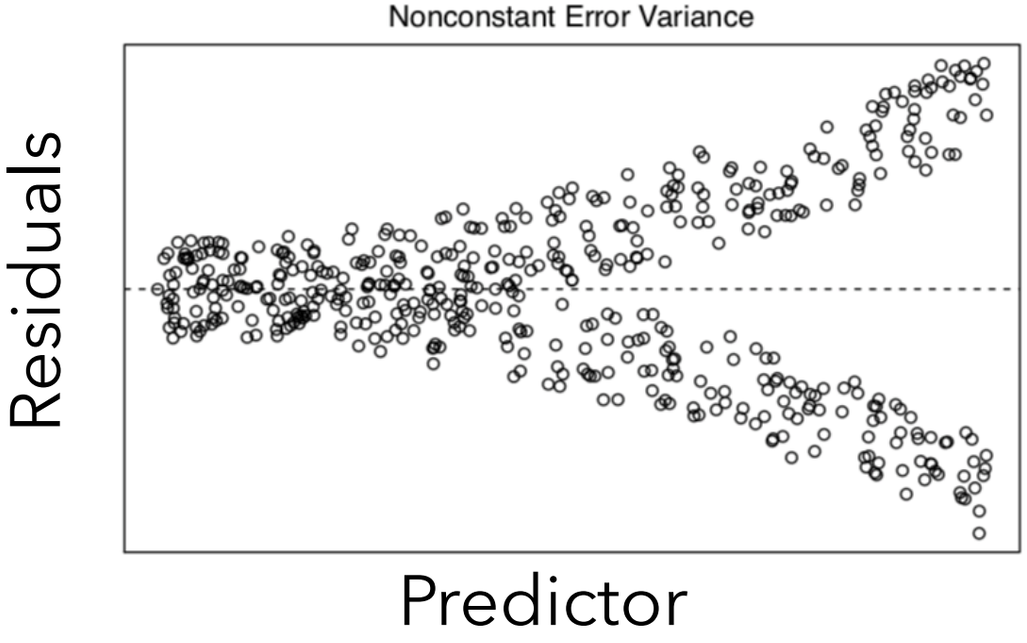
We'll explore this more next week

Wrap-up - Brief Discussion

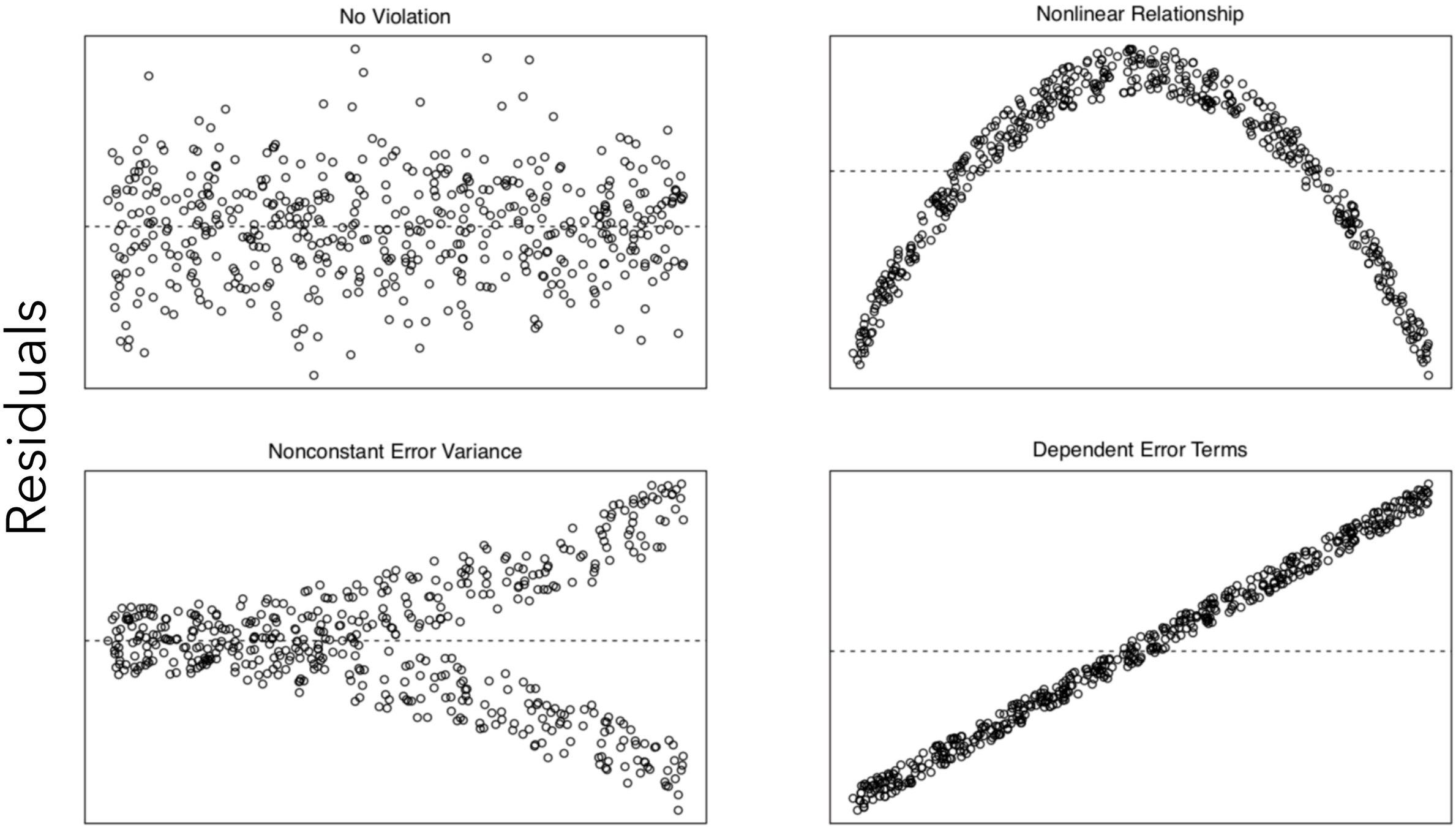
What do you think happens to our estimates or model comparisons if our errors are **not independent?**



What do you think happens to our estimates or model comparisons if our errors have **non-constant variance?**



Violating assumptions: examples



Next time

- **Out-of-sample** model evaluation
- Interpreting model **parameters**
- Model **parameter inference**
- Linear Regression in Python…

HW 2 Due Monday @ midnight

Thanks for being accommodating!