



# PSYCH 201B

## *Statistical Intuitions for Social Scientists*

# Modeling data III

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



# 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 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  $\longrightarrow \epsilon^2 = \sum (y - X\hat{b})^2$

# 1 model for solving many equations

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

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

$$79 = \beta_0 + \beta_1 * 41.26$$

$$63.5 = \beta_0 + \beta_1 * 32.15$$

$$60.9 = \beta_0 + \beta_1 * 31.43$$

.

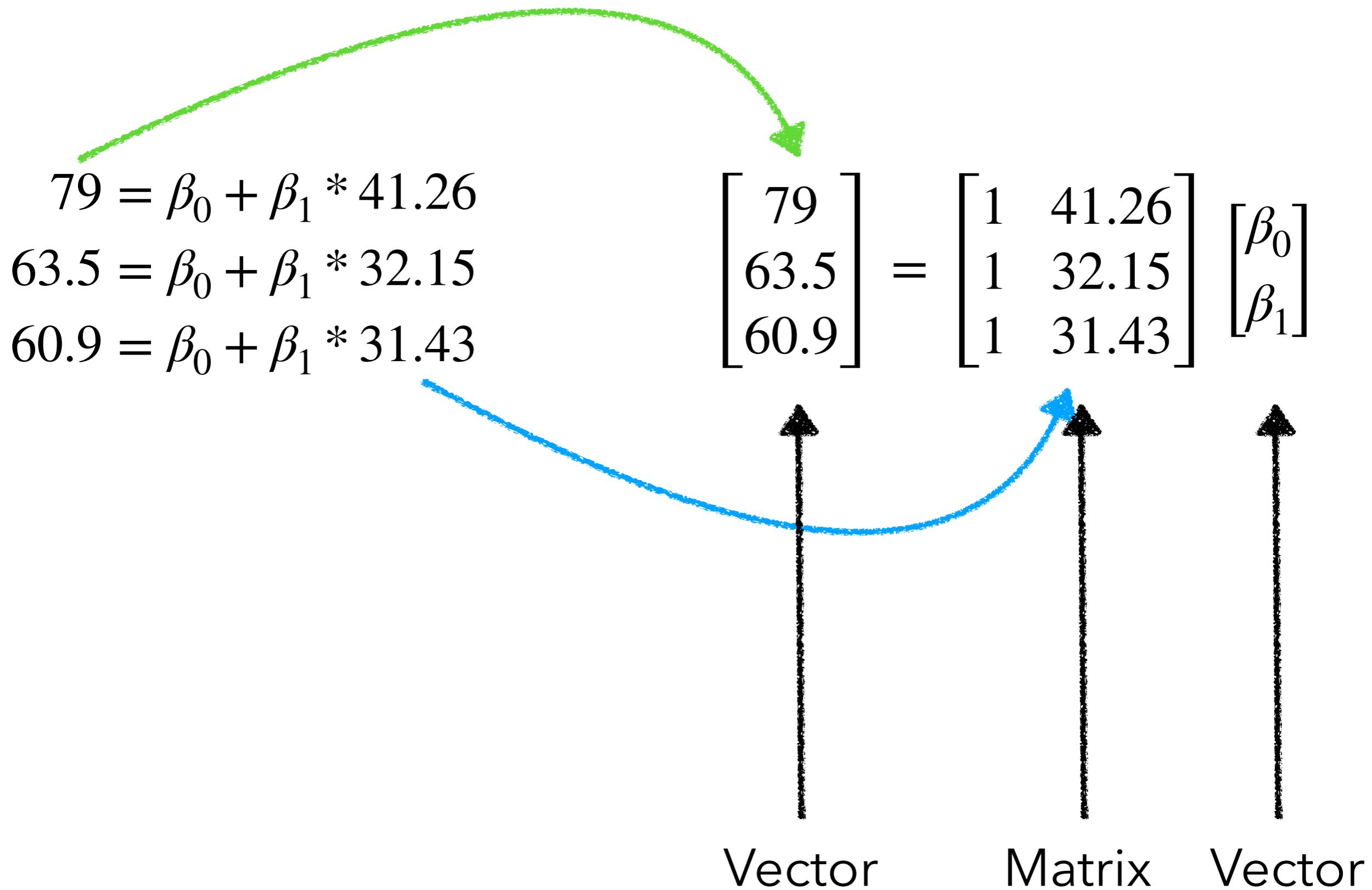
.

.

$$73.9 = \beta_0 + \beta_1 * 37.53$$

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



# 1 model for solving 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

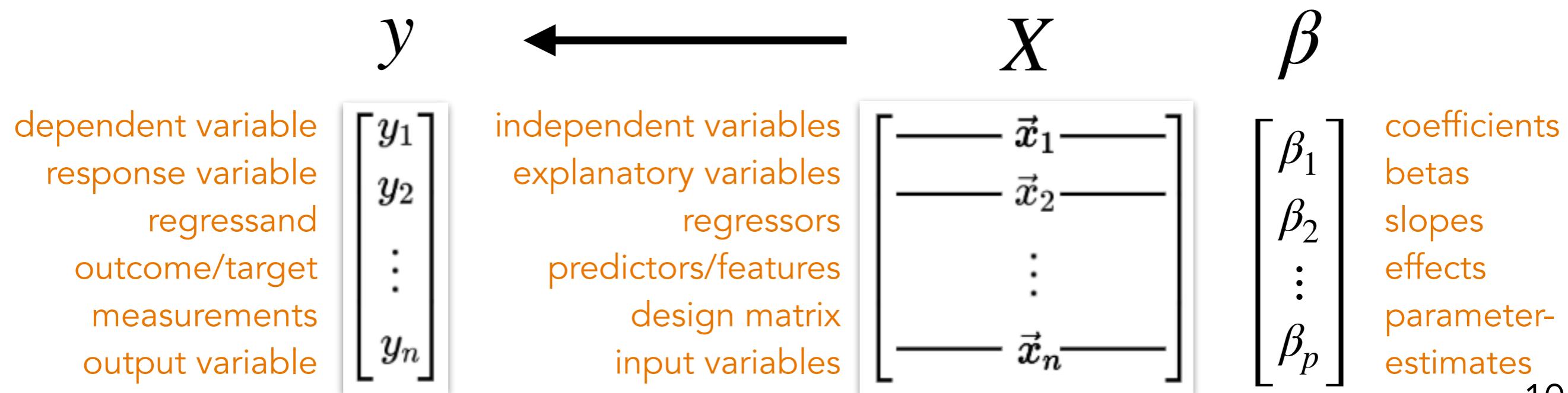
Independent Variable(s)

The diagram illustrates a linear regression model. It shows a matrix equation where three dependent variables (79, 63.5, 60.9) are expressed as a product of a matrix of independent variables (1, 41.26; 1, 32.15; 1, 31.43) and a vector of parameters ( $\beta_0$ ,  $\beta_1$ ). A green arrow points from the dependent variables to the left side of the equation, and a blue arrow points from the independent variables to the right side. Handwritten annotations in blue highlight the intercept concept and the estimation of two parameters.

# 1 model for solving many equations

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

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



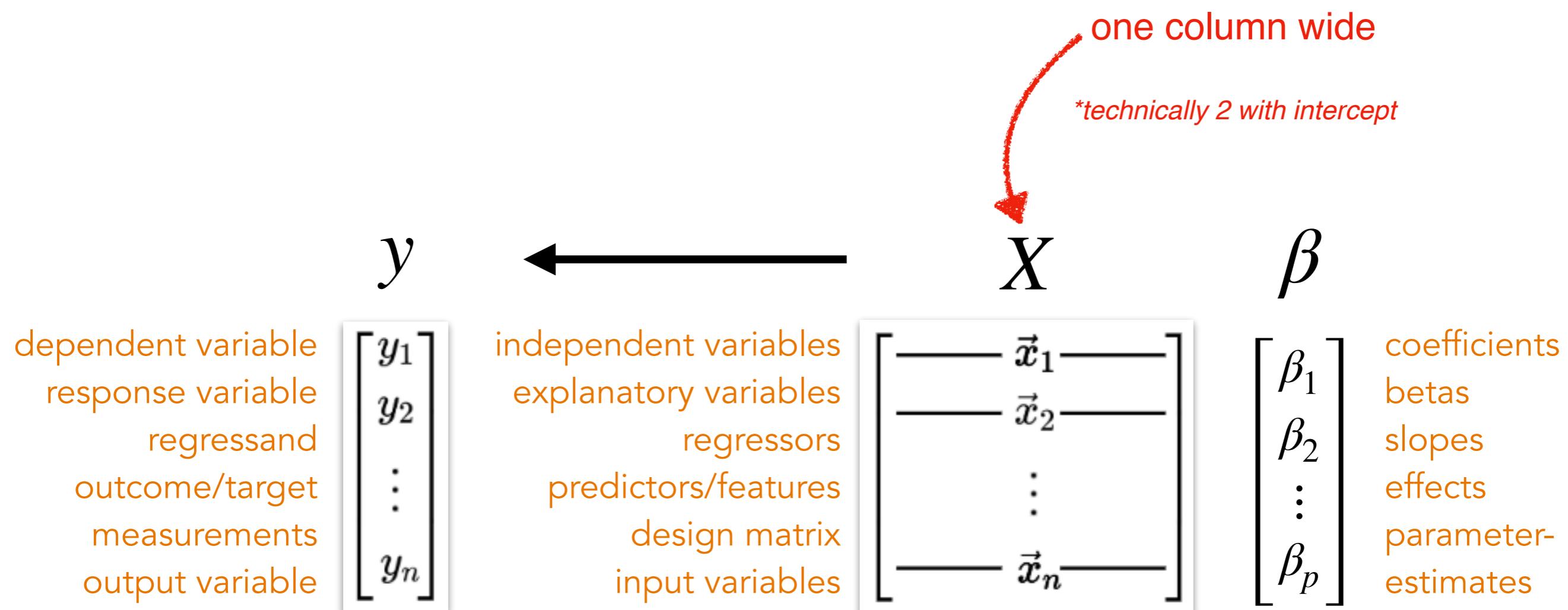
# Regression(s) are just “flavors” of GLM

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

Simple (univariate) regression

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

one predictor



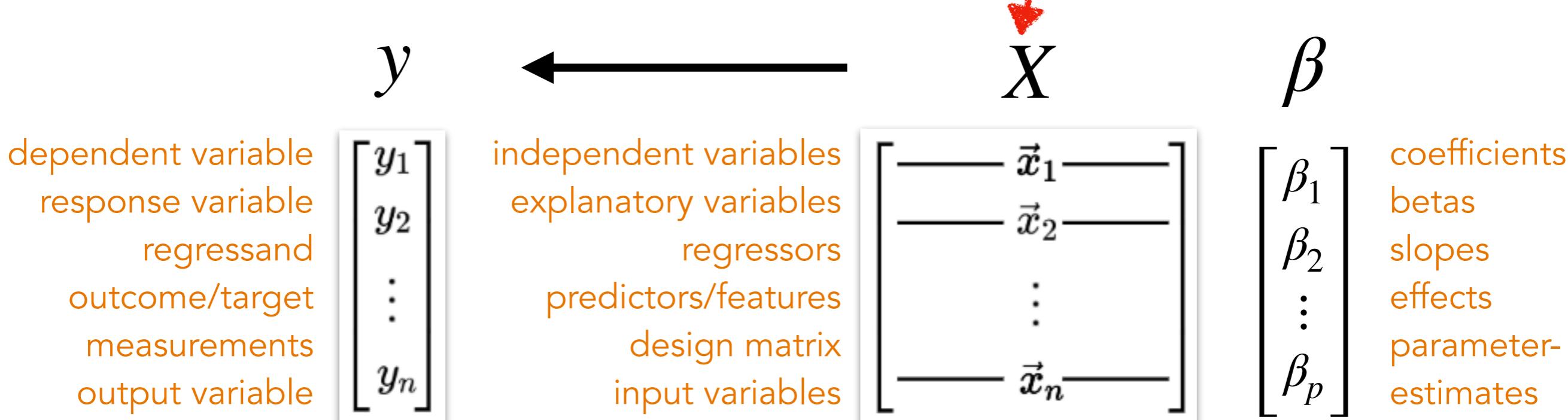
# Regression(s) are just “flavors” of GLM

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

Multiple regression

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

many predictors



# 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

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

effects  
parameter-  
estimates

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

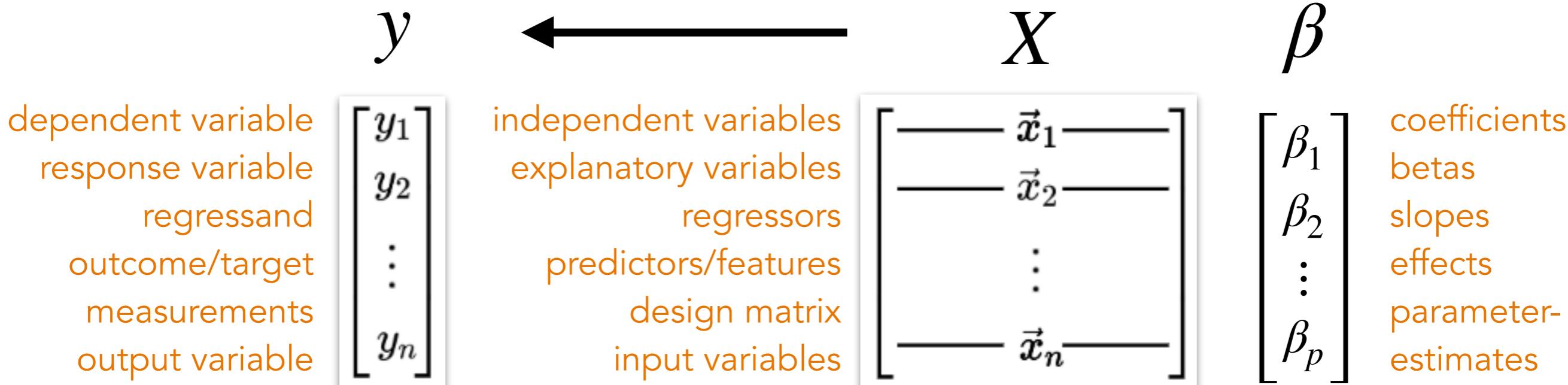
# Analytic approach to estimating GLM

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

how do we find the best  $\beta$  that minimizes SSE?

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

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



# Analytic approach to estimating GLM

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

how do we find the best  $\beta$  that minimizes SSE?

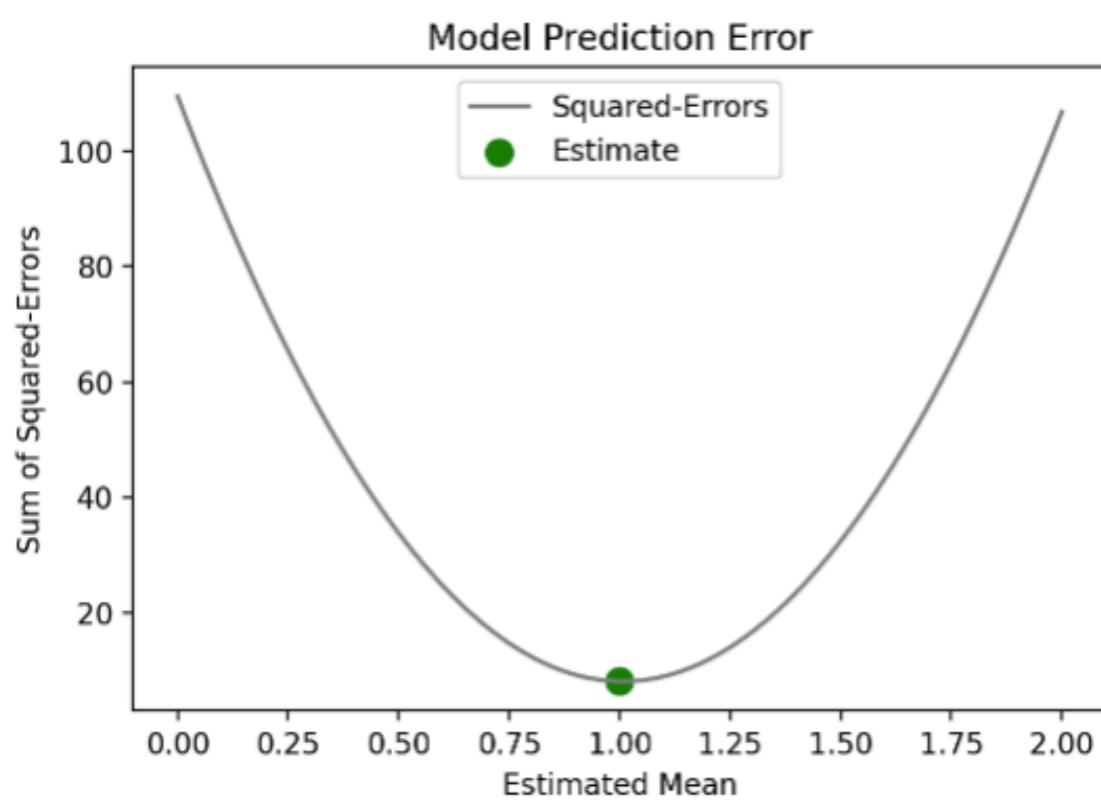
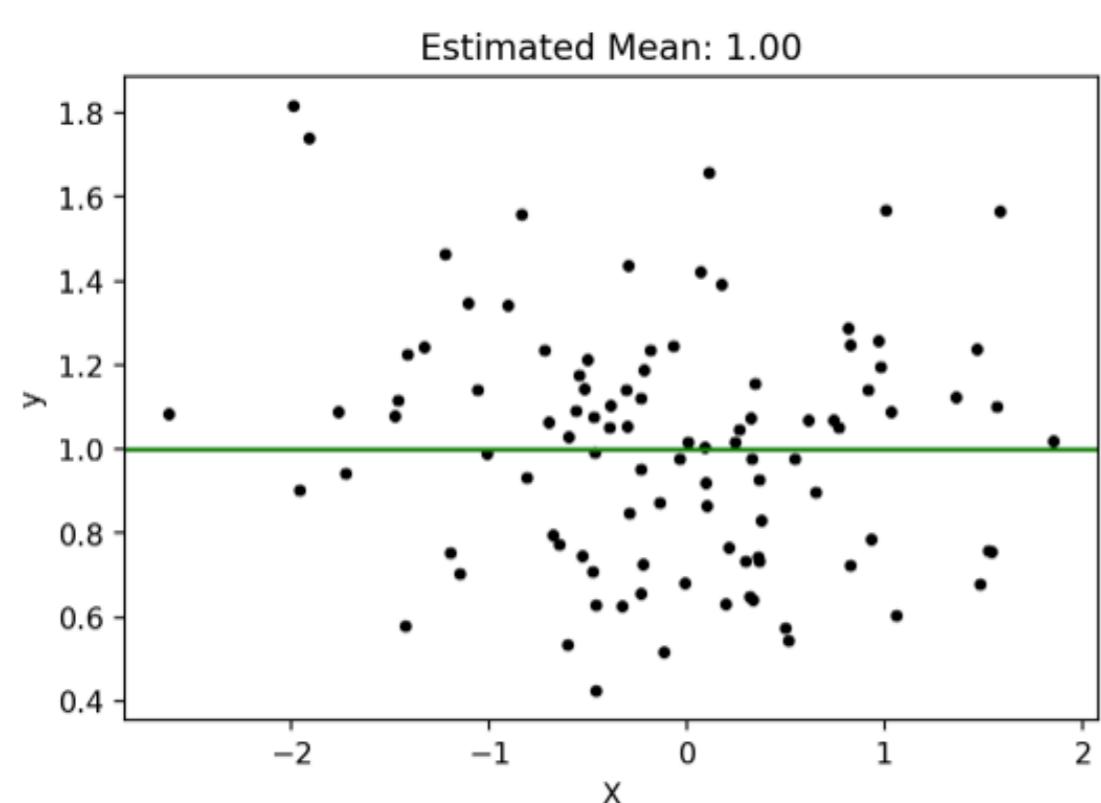
Ordinary Least Squares (**OLS**):  
an *analytic* (closed-form) equation for  
estimating betas that minimizes SSE!

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

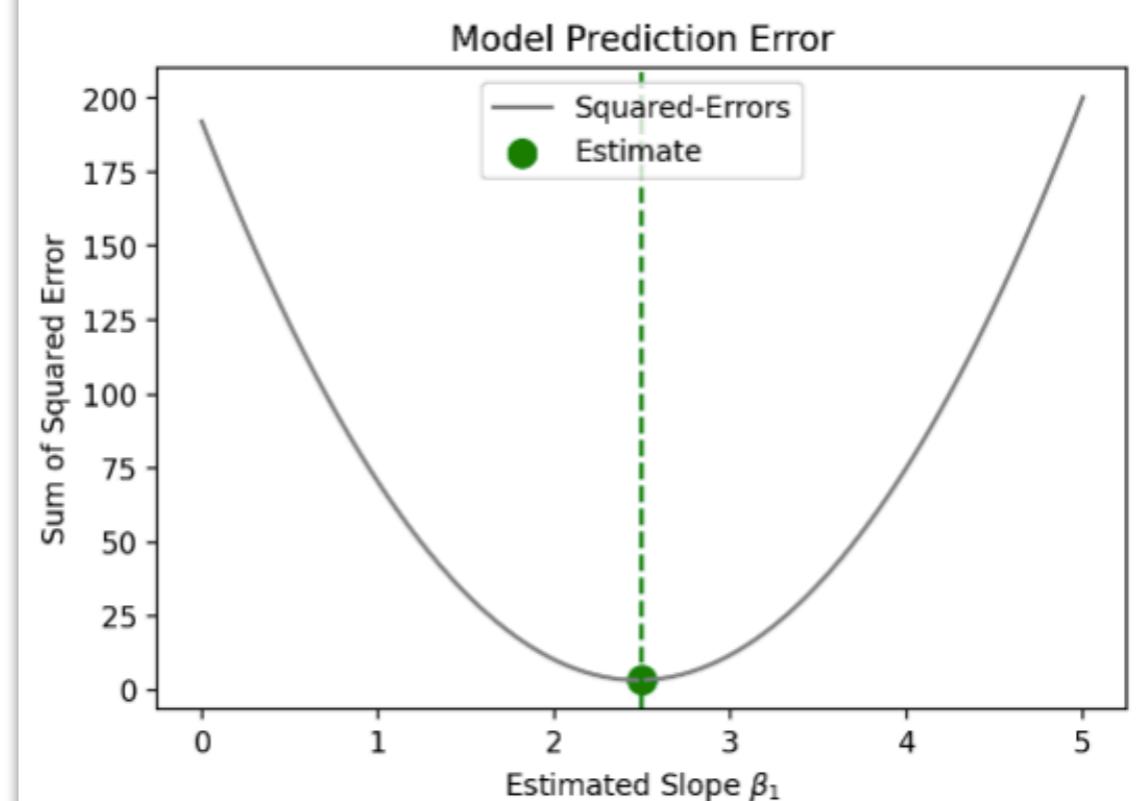
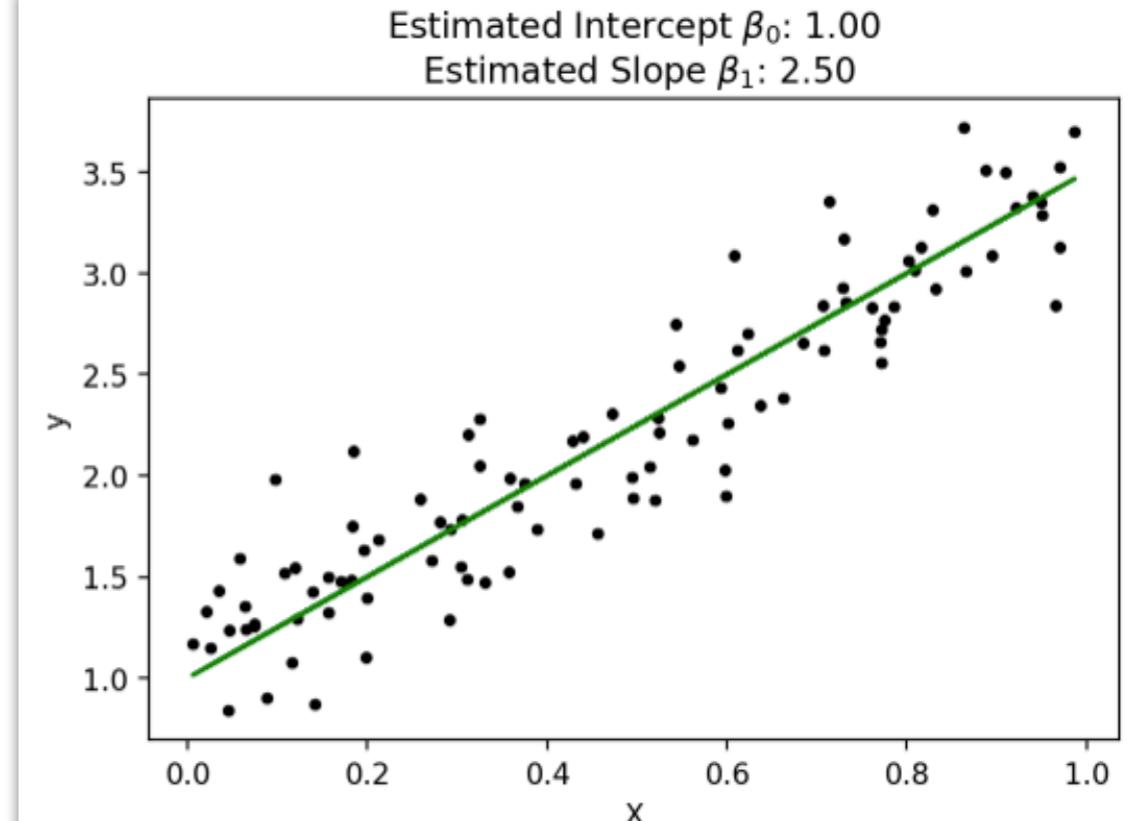
Just like  $\text{mean} = \text{best single estimate}$  that minimizes SSE

OLS =  $\text{best slope estimate(s)}$  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!

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

$y$	$X$	$\beta$
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	$\begin{bmatrix} \vec{x}_1 \\ \vec{x}_2 \\ \vdots \\ \vec{x}_n \end{bmatrix}$ 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?*

$$\hat{\beta} = \underbrace{(X^T X)^{-1}}_{\text{Matrix inverse}} \underbrace{X^T y}_{\text{"undo" or "divide"}}$$

Similarity between predictors

Similarity between predictors and outcome

Matrix inverse  
"undo" or "divide"

# What is it doing?

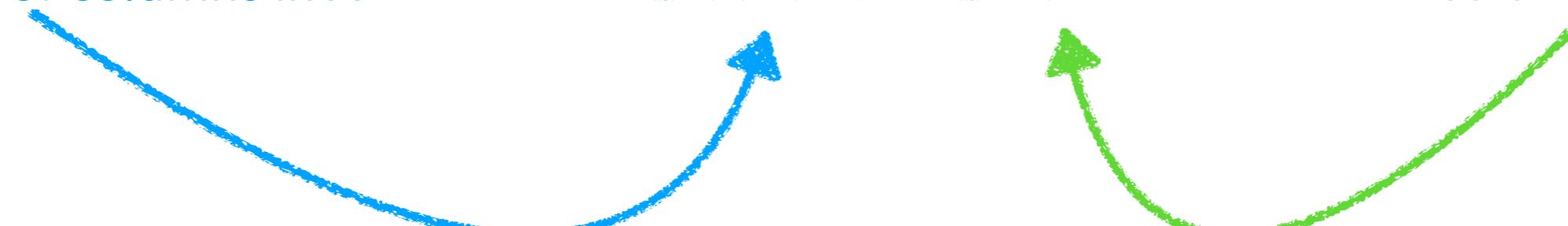
Calculate the *similarity* between each X and y  
after removing the covariance between Xs

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

Matrix-Matrix Dot-Product:  
like *looping over all possible pairs of columns in X*

$$\hat{\beta} = \underbrace{(X^T X)^{-1}}_{\text{Matrix-Matrix Dot-Product}} \underbrace{X^T y}_{\text{Matrix - vector dot-product}}$$

Matrix - vector dot-product:  
like a *looping over each column of X!*



# What is it doing?

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

Matrix inverse  
"undo" or "divide"

If this gets the slope of one IV...

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

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  $X_1$  and  $X_2$  in predicting  $Y$*

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

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

Normalize by ONLY  $Xs$   
OR  
Normalize by  $Xs$  AND  $Y$

## Correlation

*What is the non-unique contribution of  $X_1$  and  $X_2$  in predicting  $Y$*

$$\hat{r}_{x_1y} = \frac{\text{covariance}(x_1, y)}{\sigma_{x_1} \sigma_{y_1}}$$

$$\hat{r}_{x_2y} = \frac{\text{covariance}(x_2, y)}{\sigma_{x_2} \sigma_{y_1}}$$

**Correlation doesn't account for shared variance between predictors**

# Regression estimate -> Correlation

Regression

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

Correlation

$$\hat{r}_{x_1y} = \frac{\text{covariance}(x_1, y)}{\sigma_{x_1} \sigma_{y_1}}$$

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

$$\hat{r}_{x_1y} = \hat{\beta}_1 * \frac{\sigma_x}{\sigma_y}$$

# Regression estimate -> Correlation

Regression

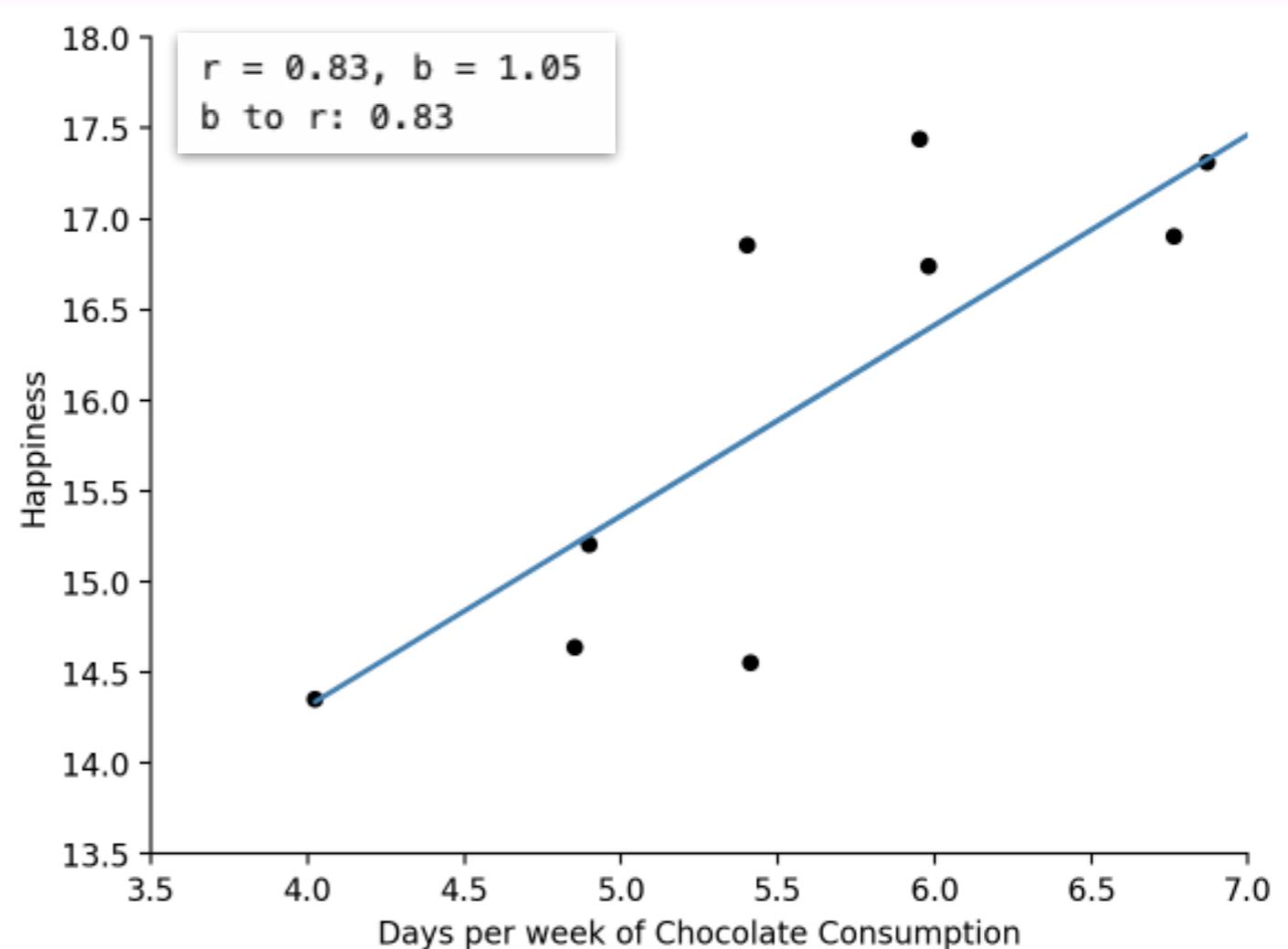
$$\hat{\beta}_1 = \frac{\text{covariance}(x_1, y)}{\sigma_{x_1} \sigma_{x_2}}$$

$$\hat{r}_{x_1y} = \frac{\text{covariance}(x_1, y)}{\sigma_{x_1} \sigma_{y_1}}$$

$$\hat{r}_{x_1y} = \hat{\beta}_1 * \frac{\sigma_x}{\sigma_y}$$

Correlation

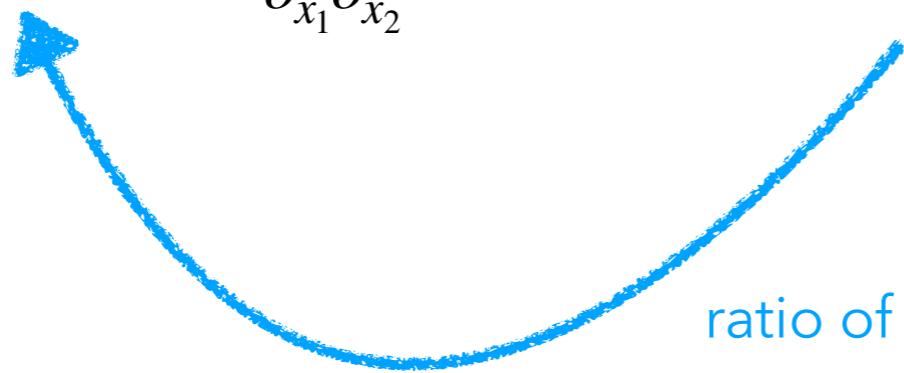
Scale beta by  
ratio of **IV-to-DV** standard deviations



# Correlation -> Regression estimate

Regression

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



Correlation

$$\hat{r}_{x_1y} = \frac{\text{covariance}(x_1, y)}{\sigma_{x_1} \sigma_{y_1}}$$

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

$$\hat{\beta}_1 = \hat{r}_{x_1y} * \frac{\sigma_y}{\sigma_x}$$

# Correlation -> Regression estimate

Regression

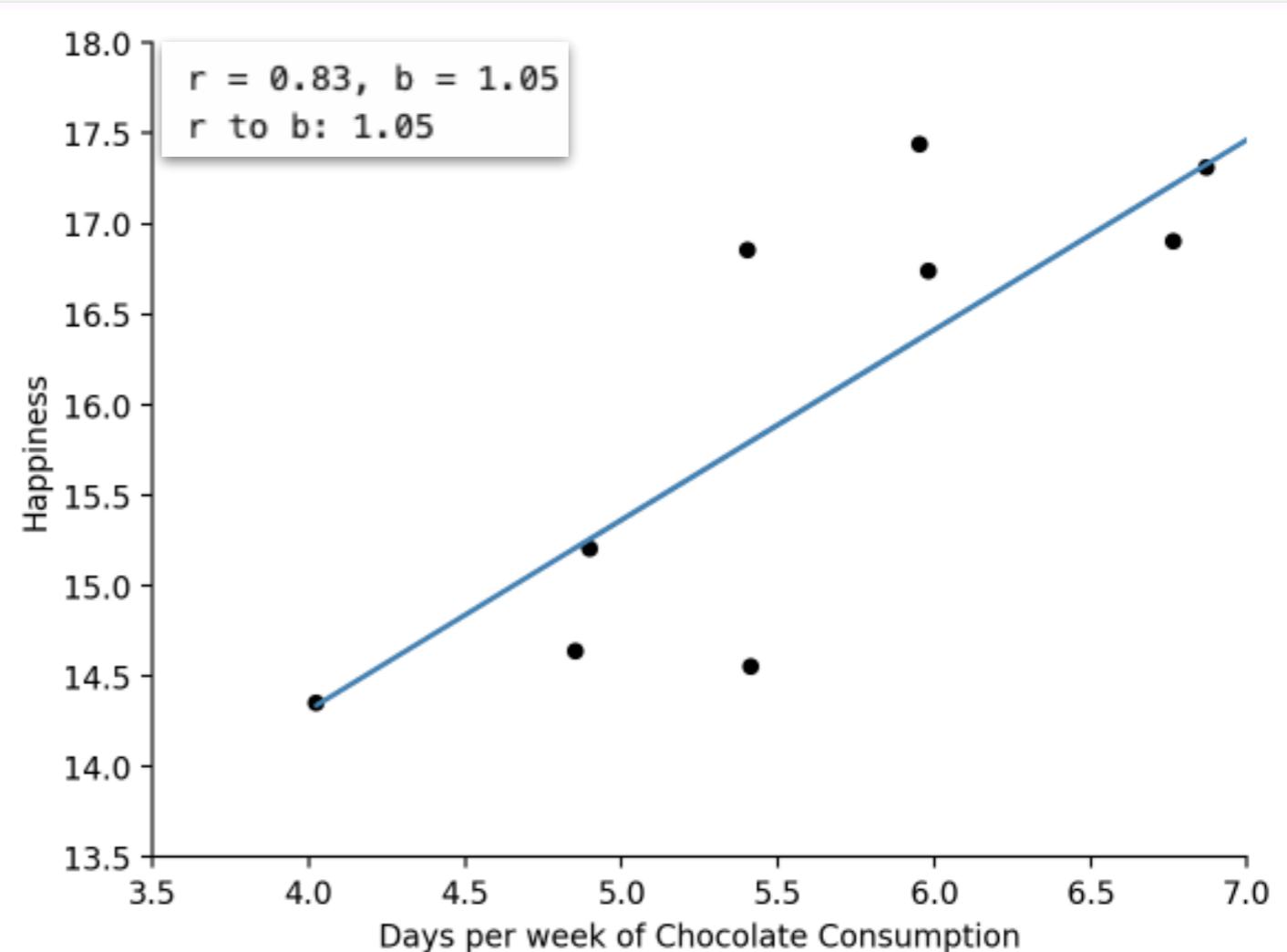
$$\hat{\beta}_1 = \frac{\text{covariance}(x_1, y)}{\sigma_{x_1} \sigma_{x_2}}$$

$$\hat{\beta}_1 = \hat{r}_{x_1y} * \frac{\sigma_y}{\sigma_x}$$

Correlation

$$\hat{r}_{x_1y} = \frac{\text{covariance}(x_1, y)}{\sigma_{x_1} \sigma_{y_1}}$$

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



# Recap Summary

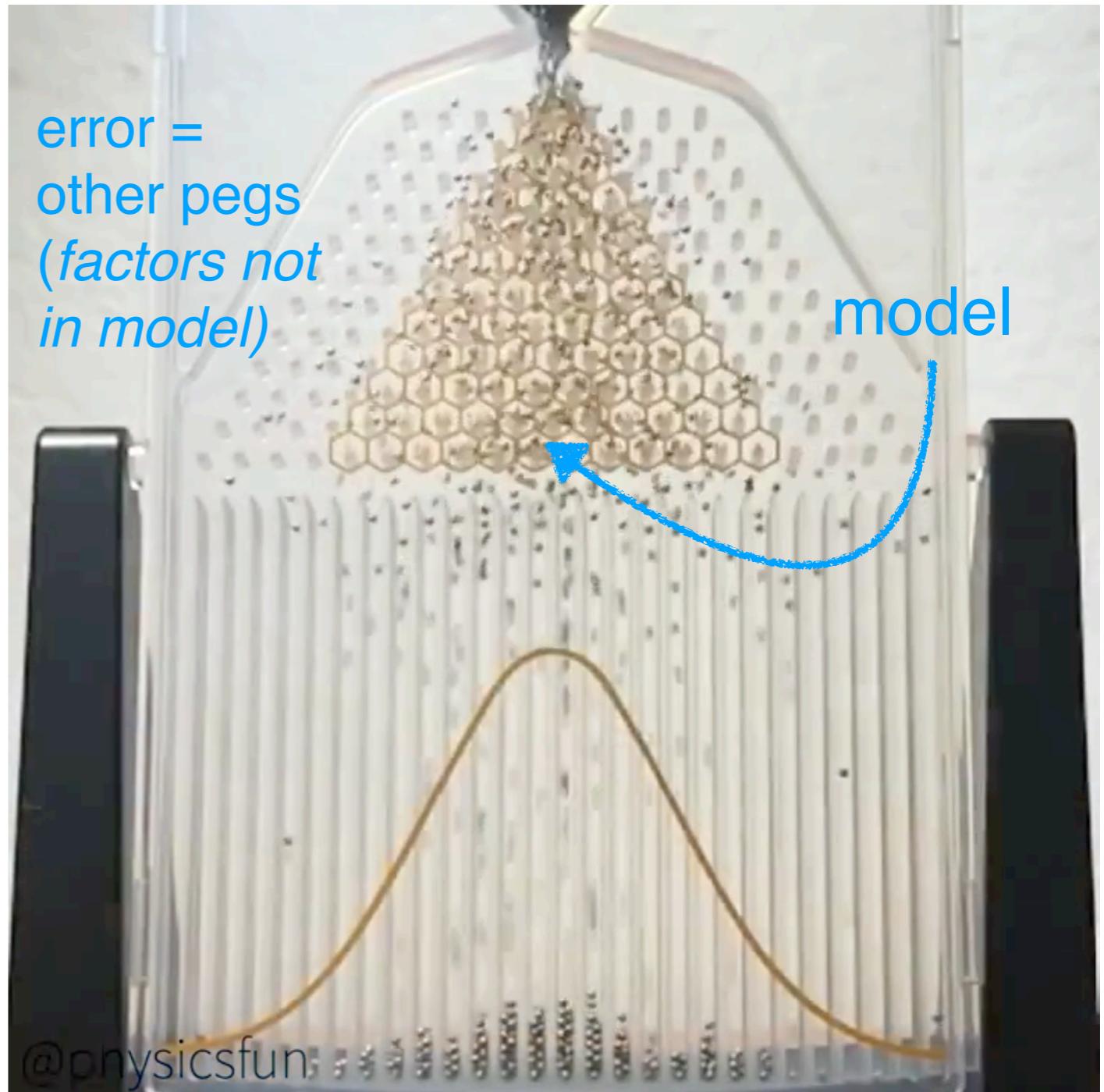
- OLS = **analytic** approach for finding **best linear relationship** between 1 (or more) IVs and 1 DV
- Best = **minimized** sum-of-squared errors
- Lets us collect our IVs into a **design matrix**
- **Matrix inversion** to remove similarity between IVs (columns of design matrix)
- **Dot product**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

## Questions?

# **Model *assumptions***

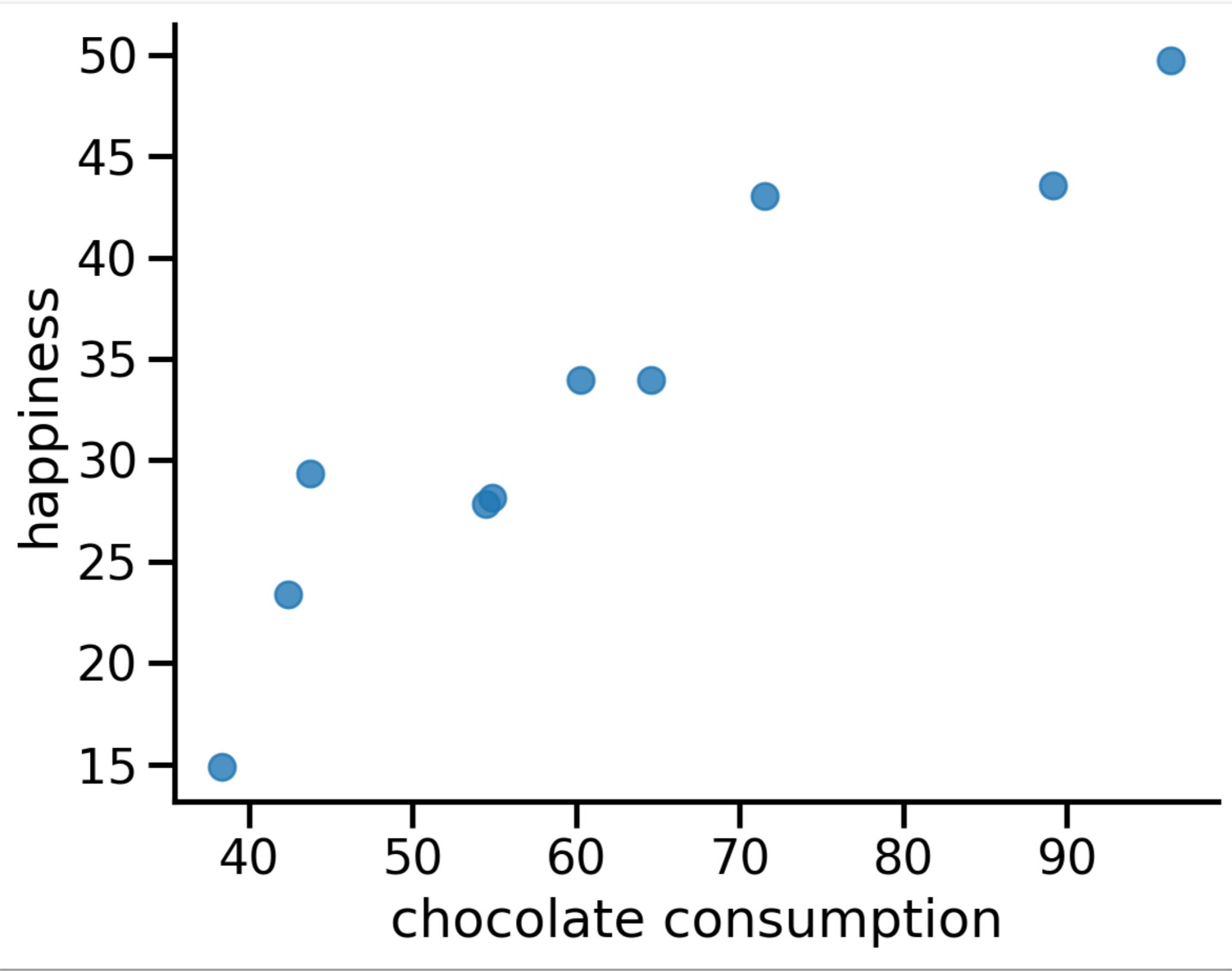
$$\text{Error} = \text{Data} - \text{Model}$$

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).

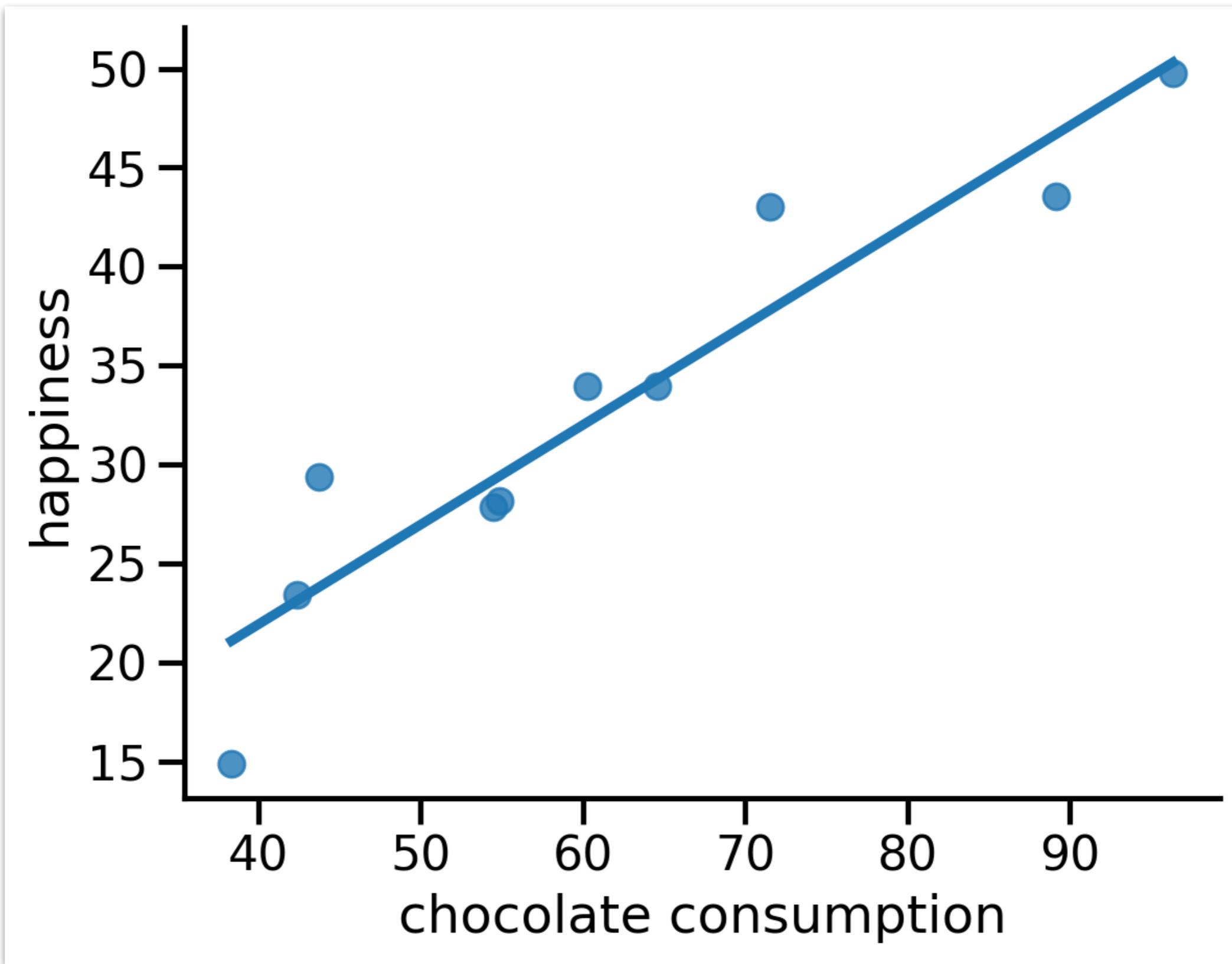


data = beads

Result: Normally Distributed Errors

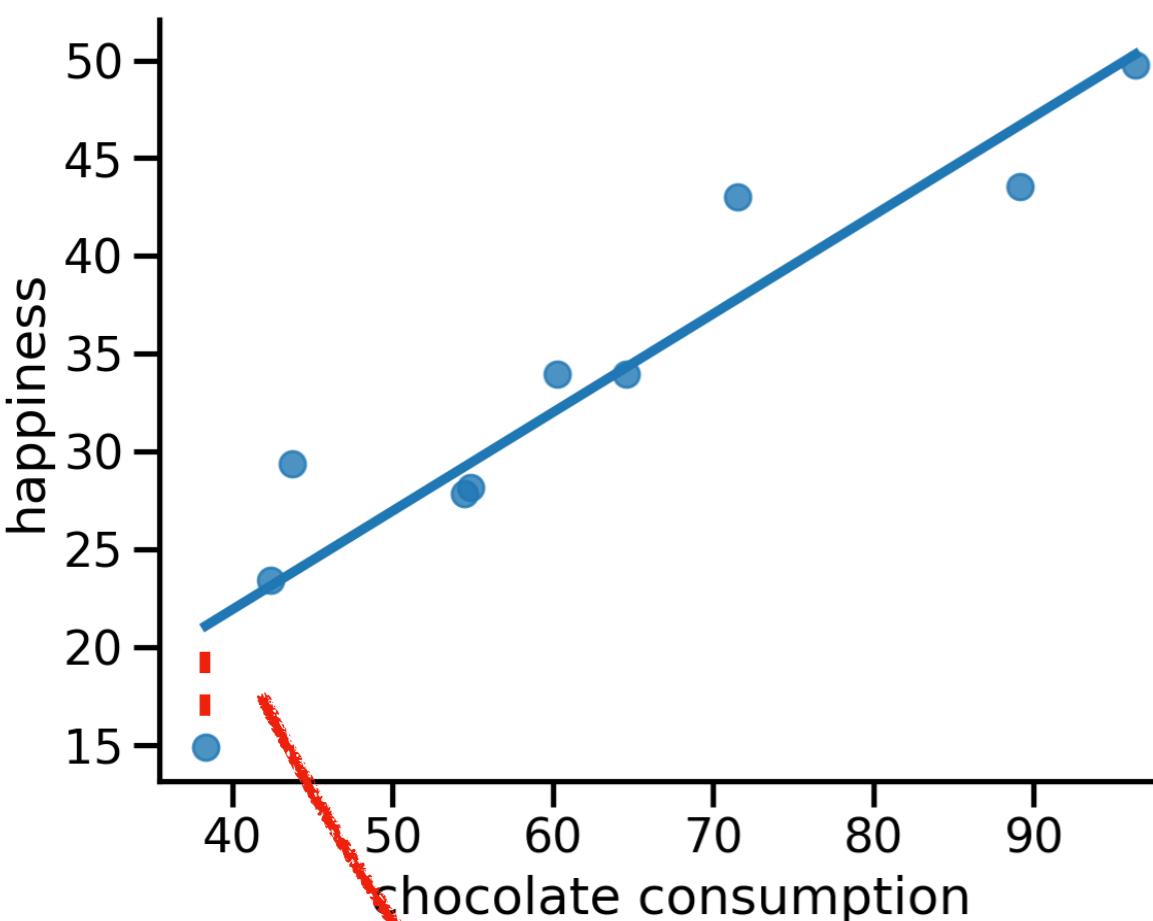


# How do we test our assumptions?



# Visualize our model errors!

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



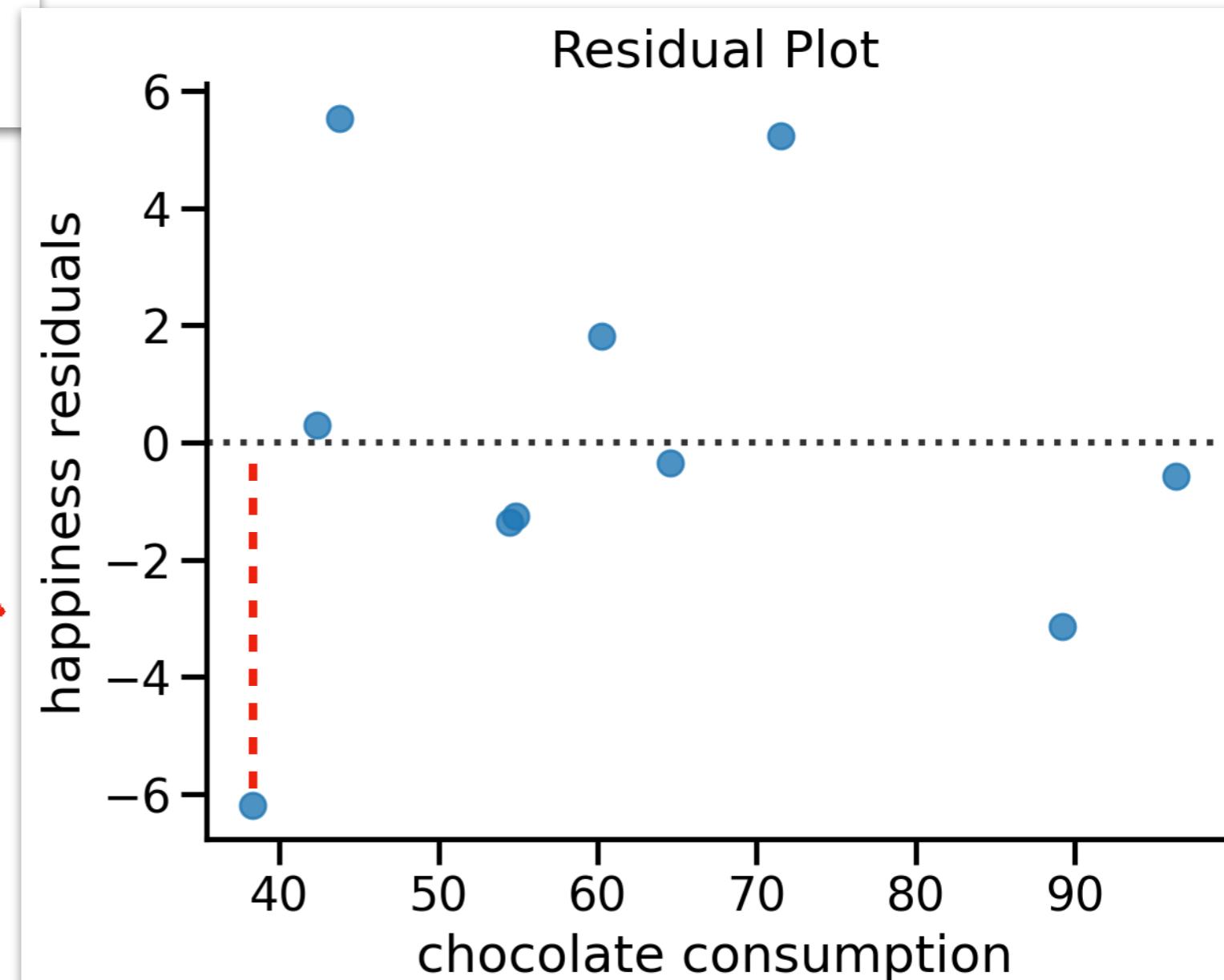
Model errors

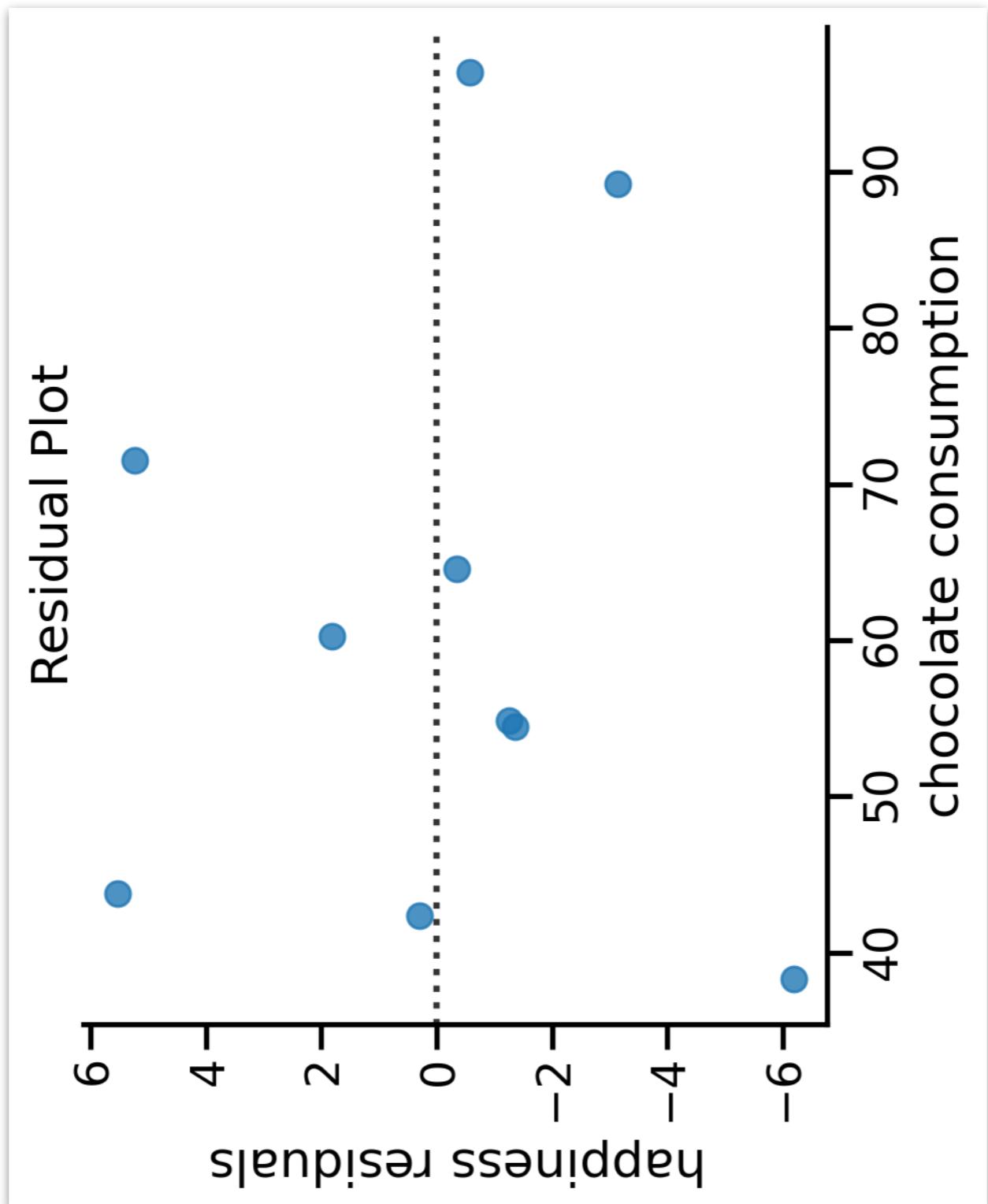
seaborn.residplot

```
seaborn.residplot(data=None, *, x=None, y=None, x_partial=None, y_partial=None,  
lowess=False, order=1, robust=False, dropna=True, label=None, color=None,  
scatter_kws=None, line_kws=None, ax=None) #
```

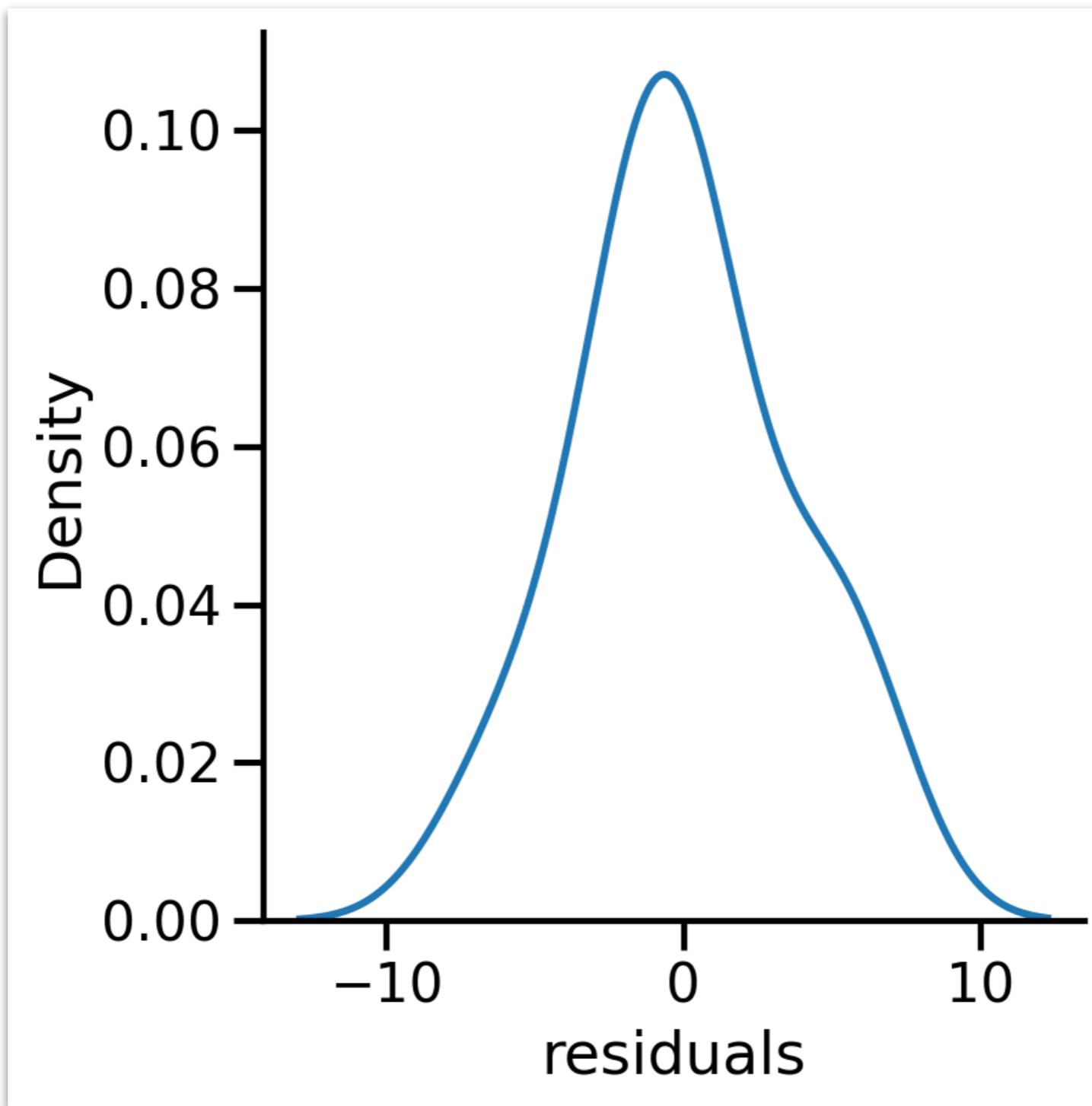
Plot the residuals of a linear regression.

This function will regress y on x (possibly as a robust or polynomial regression) and then draw a scatterplot of the residuals. You can optionally fit a lowess smoother to the residual plot, which can help in determining if there is structure to the residuals.





# Normally Distributed Errors!



# Look for **structure** in your residuals

Should be independent & identically **distributed** (i.i.d)



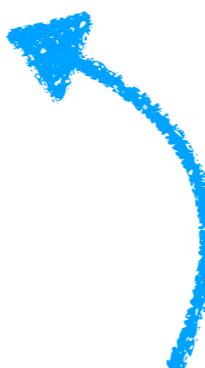
Compact way of stating the assumptions regression makes

# Look for **structure** in your residuals

Should be independent & identically **distributed** (i.i.d)

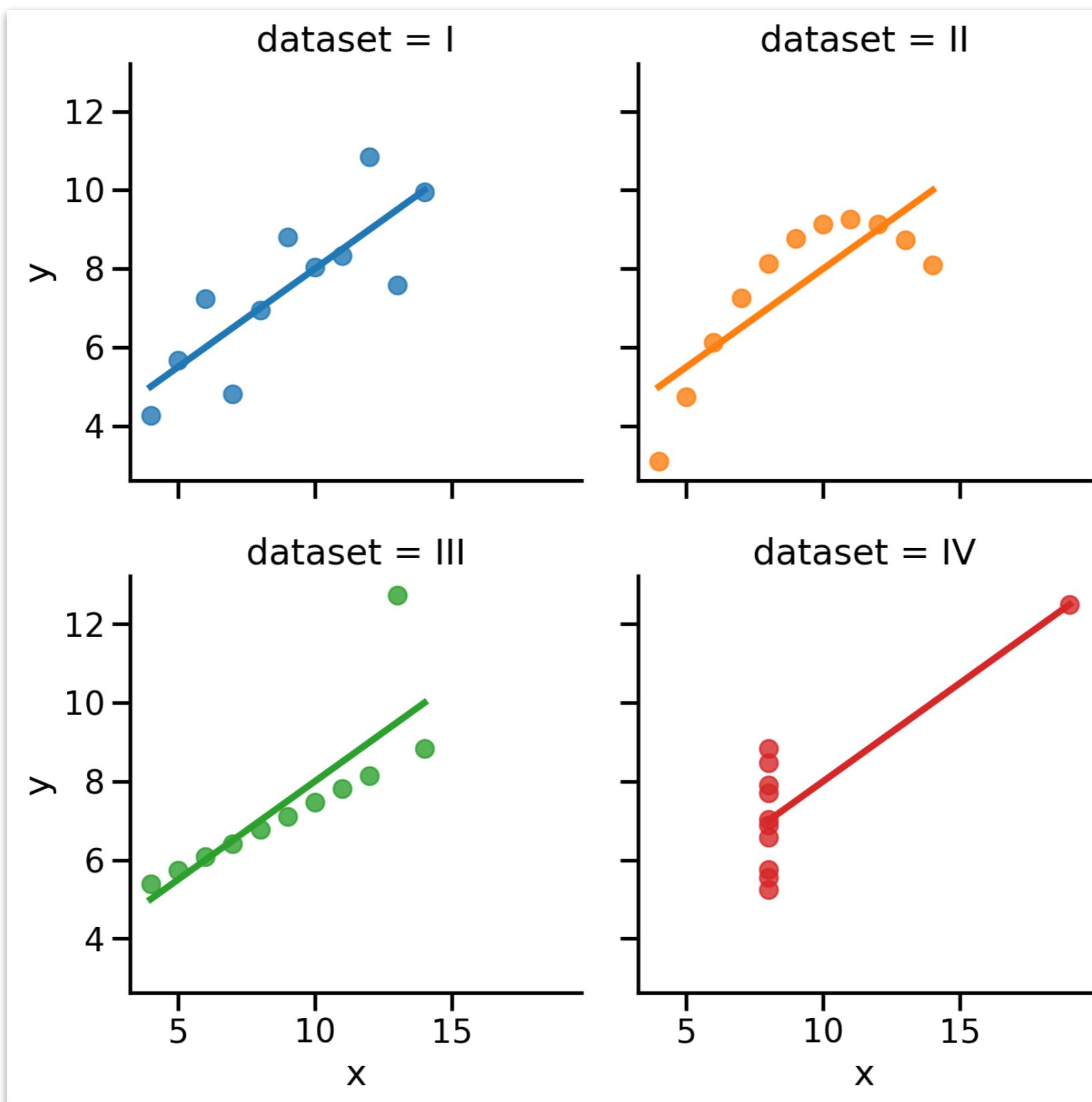
- Errors shouldn't depend on each other
- Or depend on values of Xs
- Approximately Normally distributed
- **No perfect multi-collinearity** - perfect correlations between IVs

Compact way of stating the assumptions regression makes

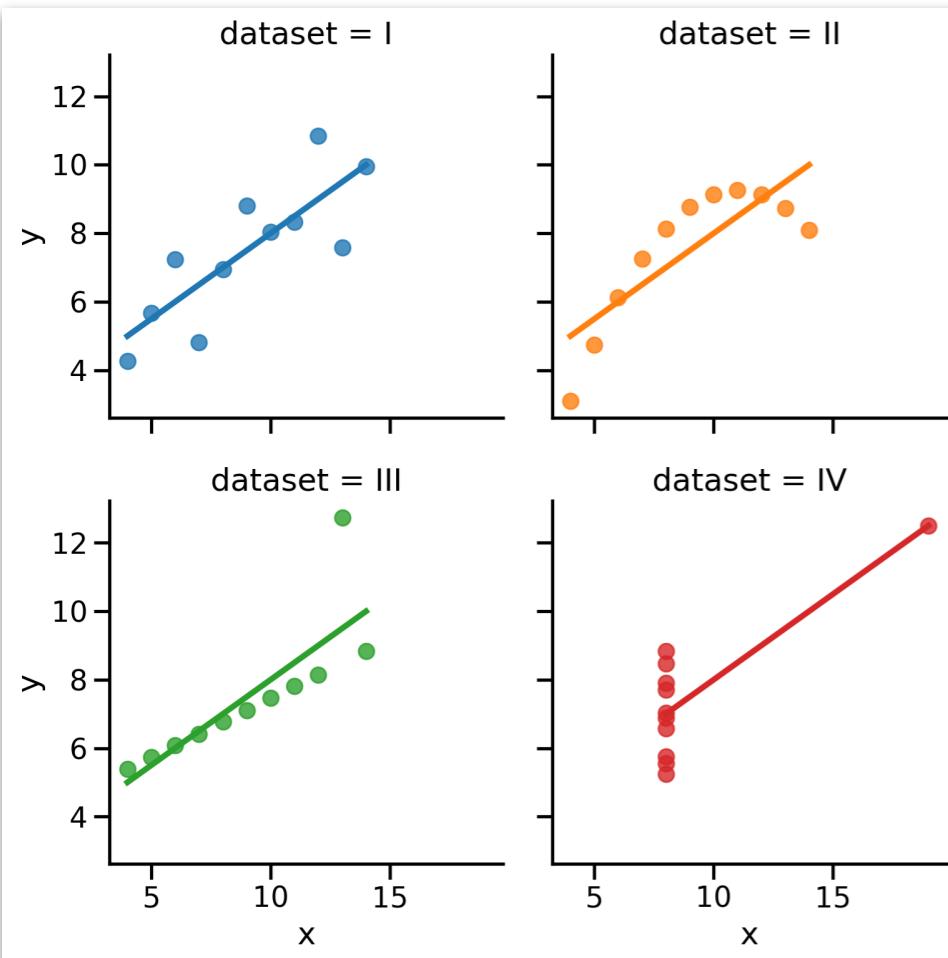


We'll explore this more next week

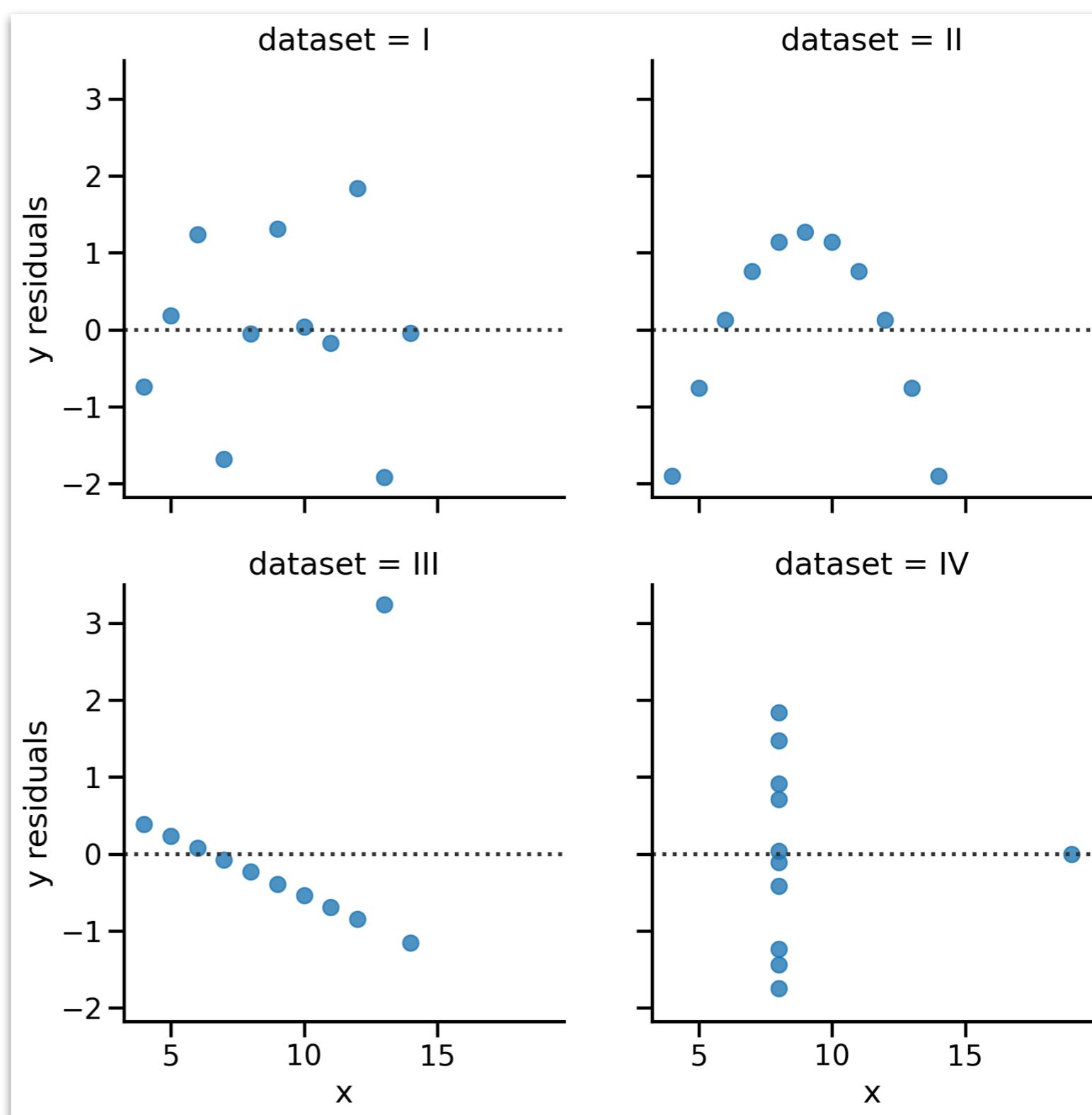
# Remember Anscombe's Quartet?



# Remember Anscombe's Quartet?

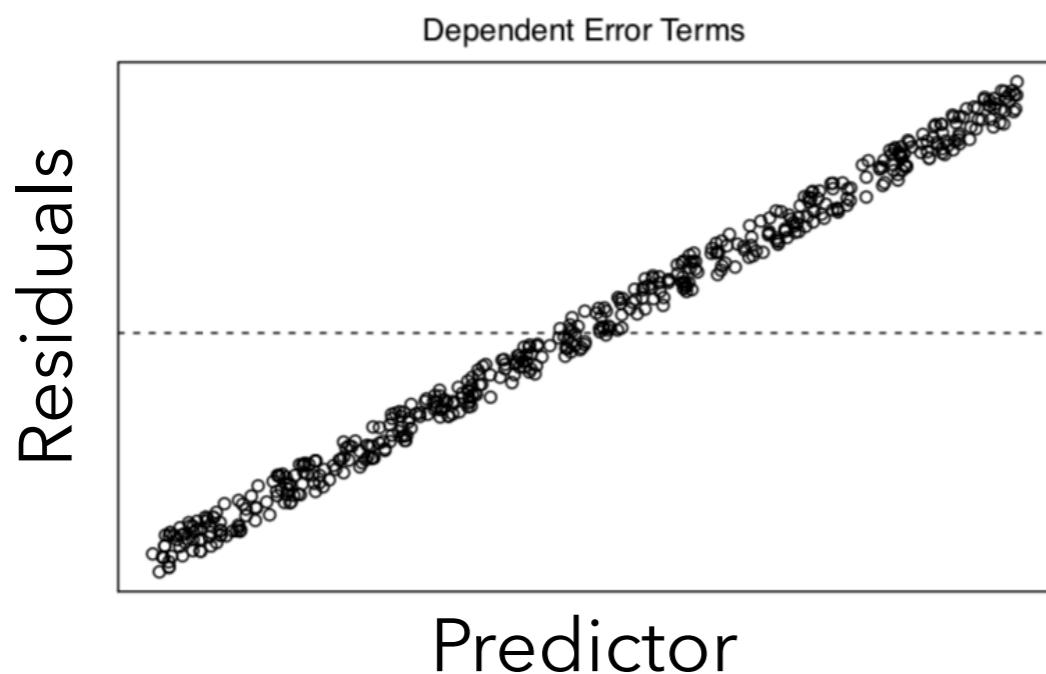


*Are these normal?*

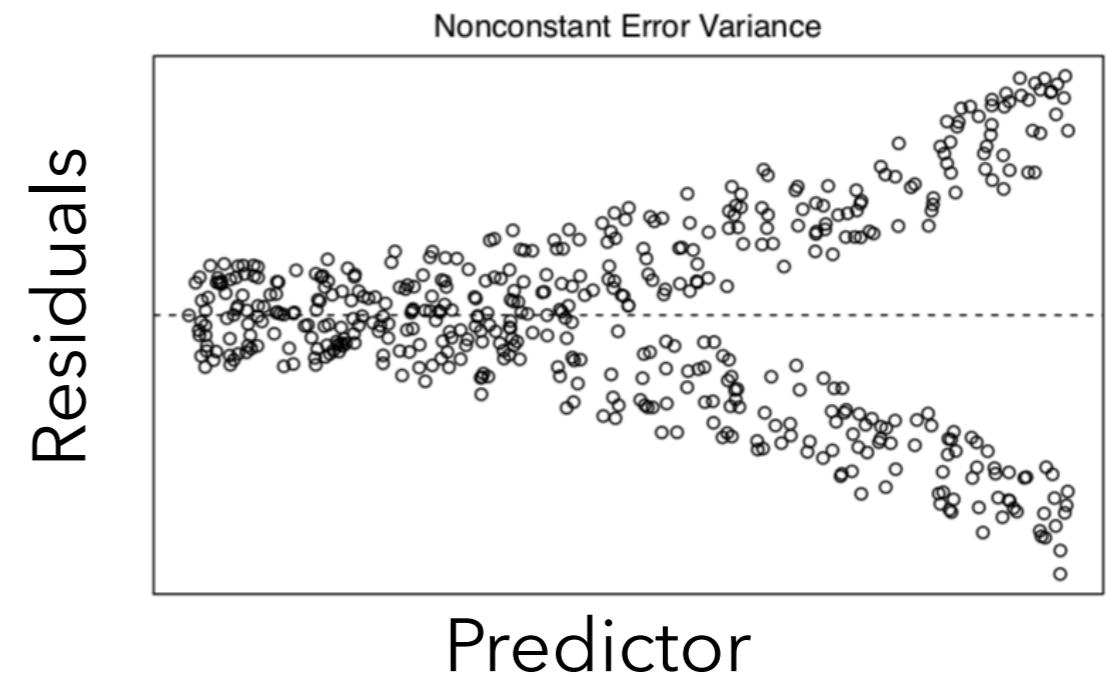


# Wrap-up - Brief Discussion

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



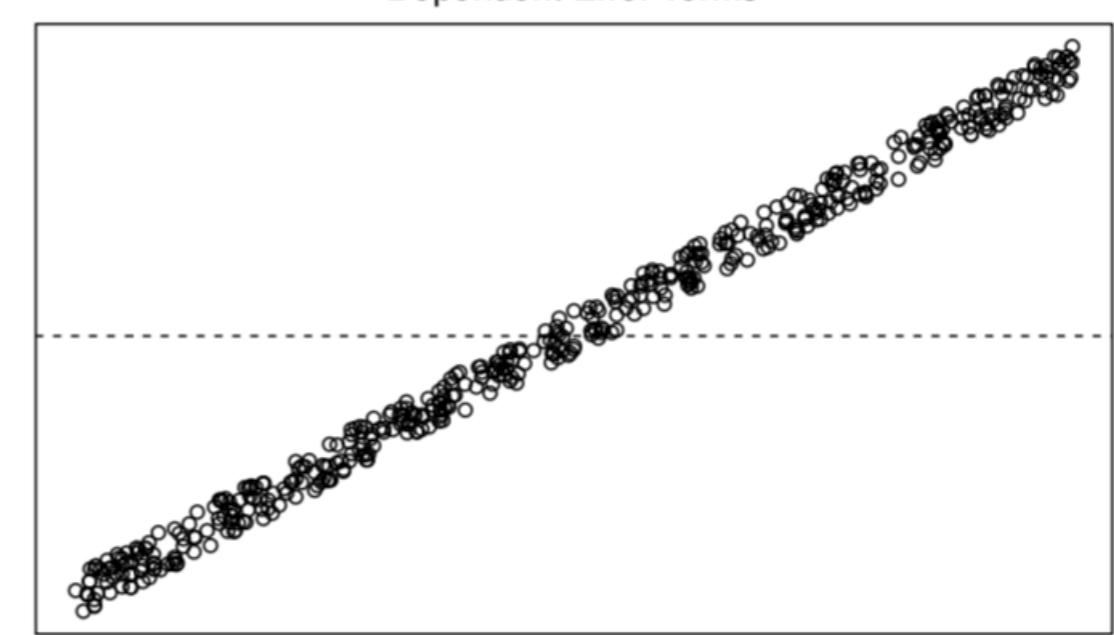
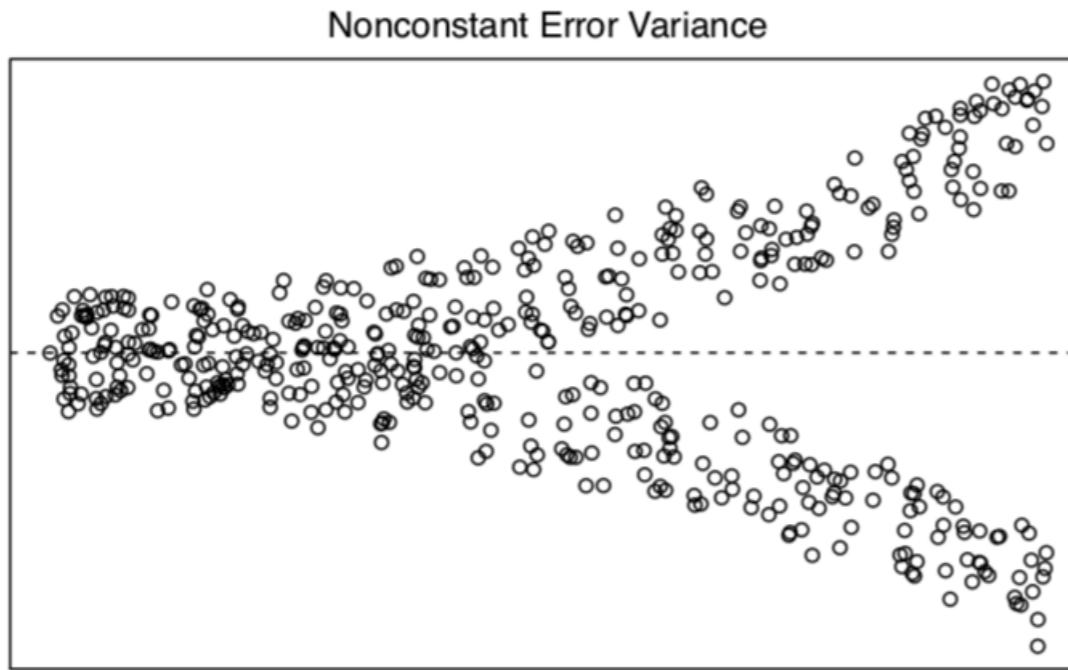
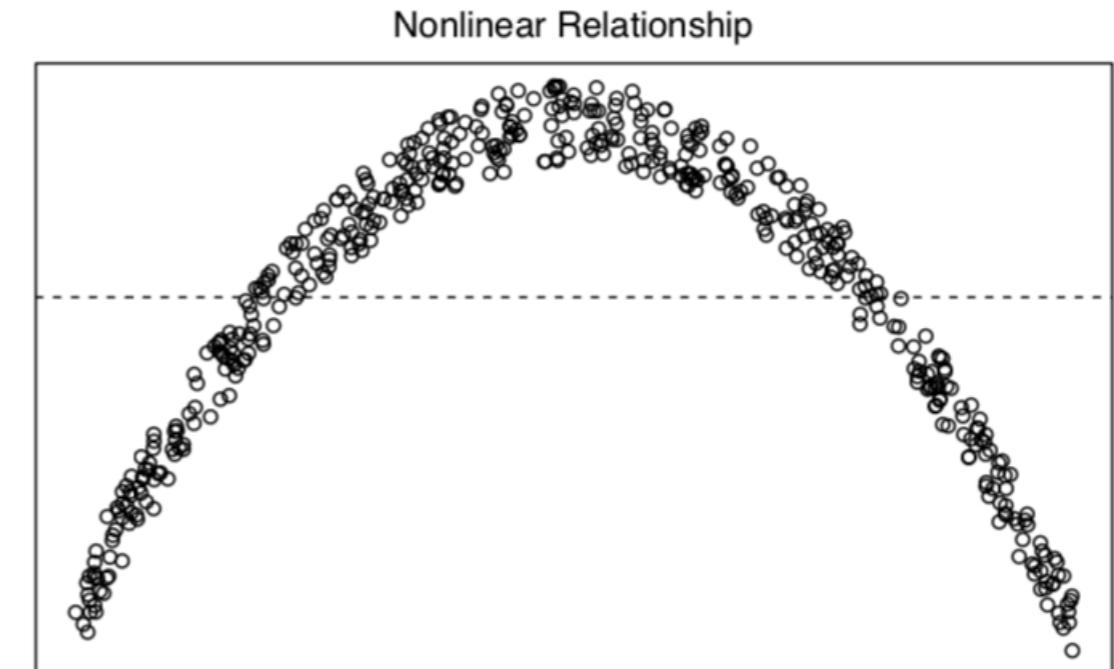
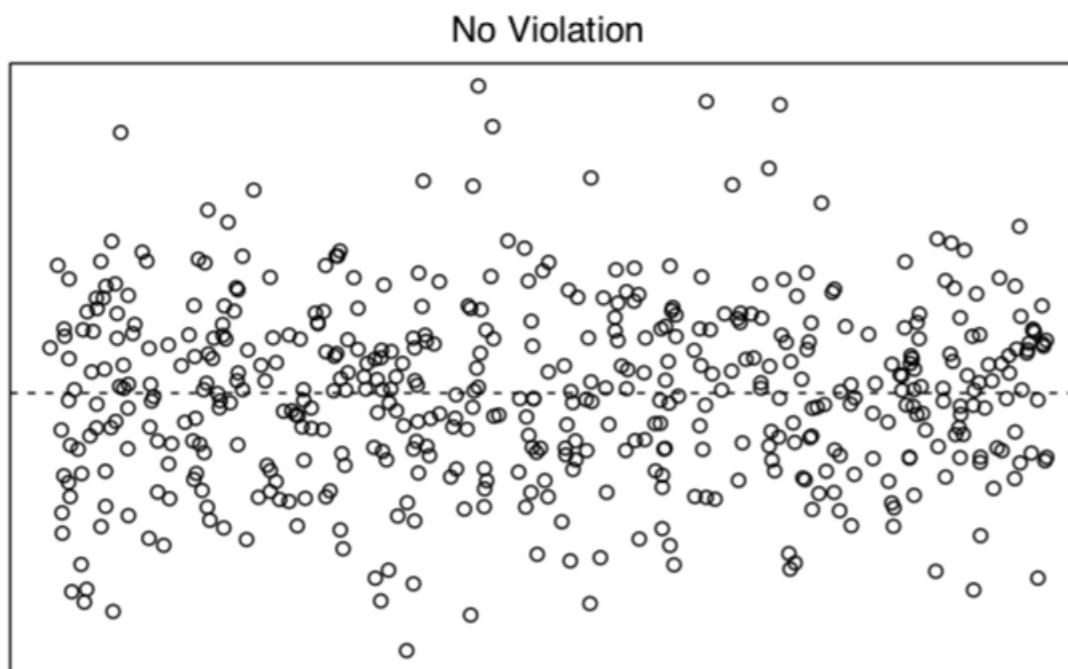
2. What do you think happens to our estimates or model comparisons if our errors have **non-constant variance**?



3. What do you think happens to our estimates or model comparisons if our **predictors are perfectly correlated**?

# Violating assumptions: examples

Residuals



Predictor

# Next time

- Interpreting model **parameters**
- Performing **parameter inference**
- **Out-of-sample** model evaluation
- Linear Regression in Python...

**HW 2 Due Monday @ midnight**

Thanks for being accommodating!

