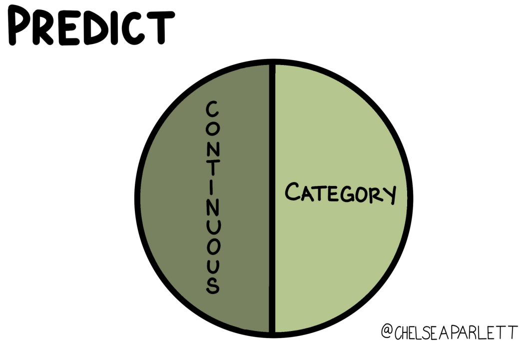


Linear Regression I

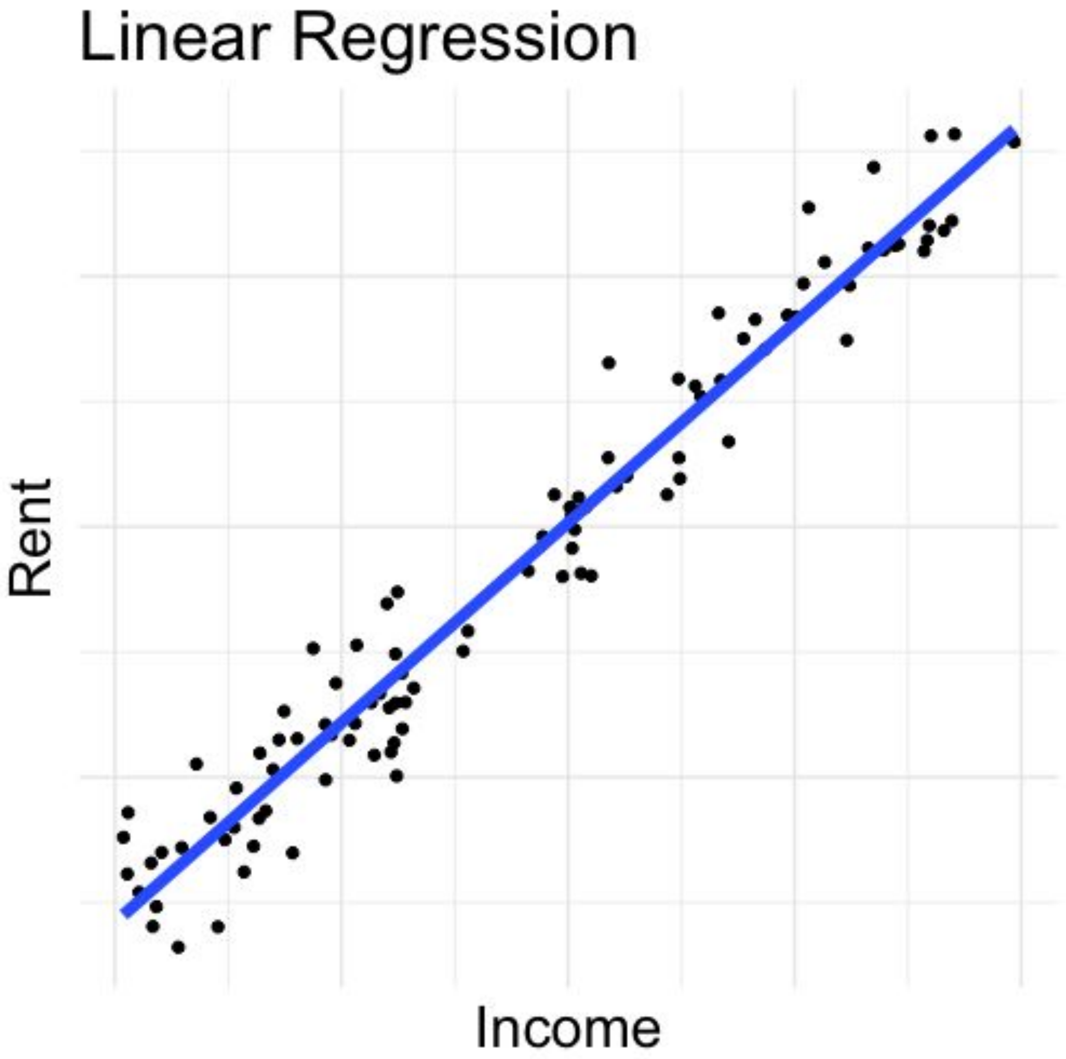
Dr. Chelsea Parlett-Pelleriti

Linear Regression

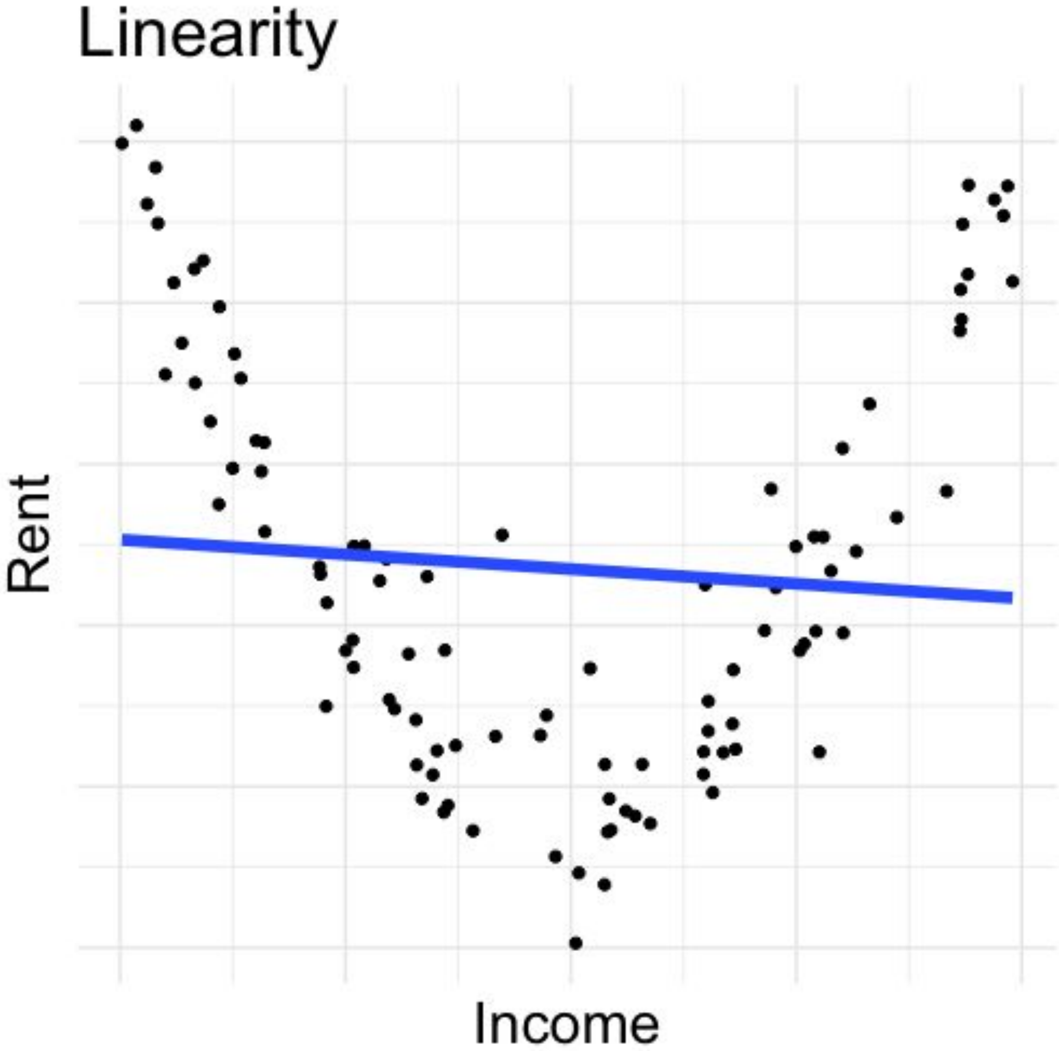
- Linear Regression Basics
- Assumptions
- Coefficients
- Z-Scoring
- Choosing a Line of Best Fit
    - Least Squares
    - Maximum Likelihood Estimation
- Assessing Model Fit

What

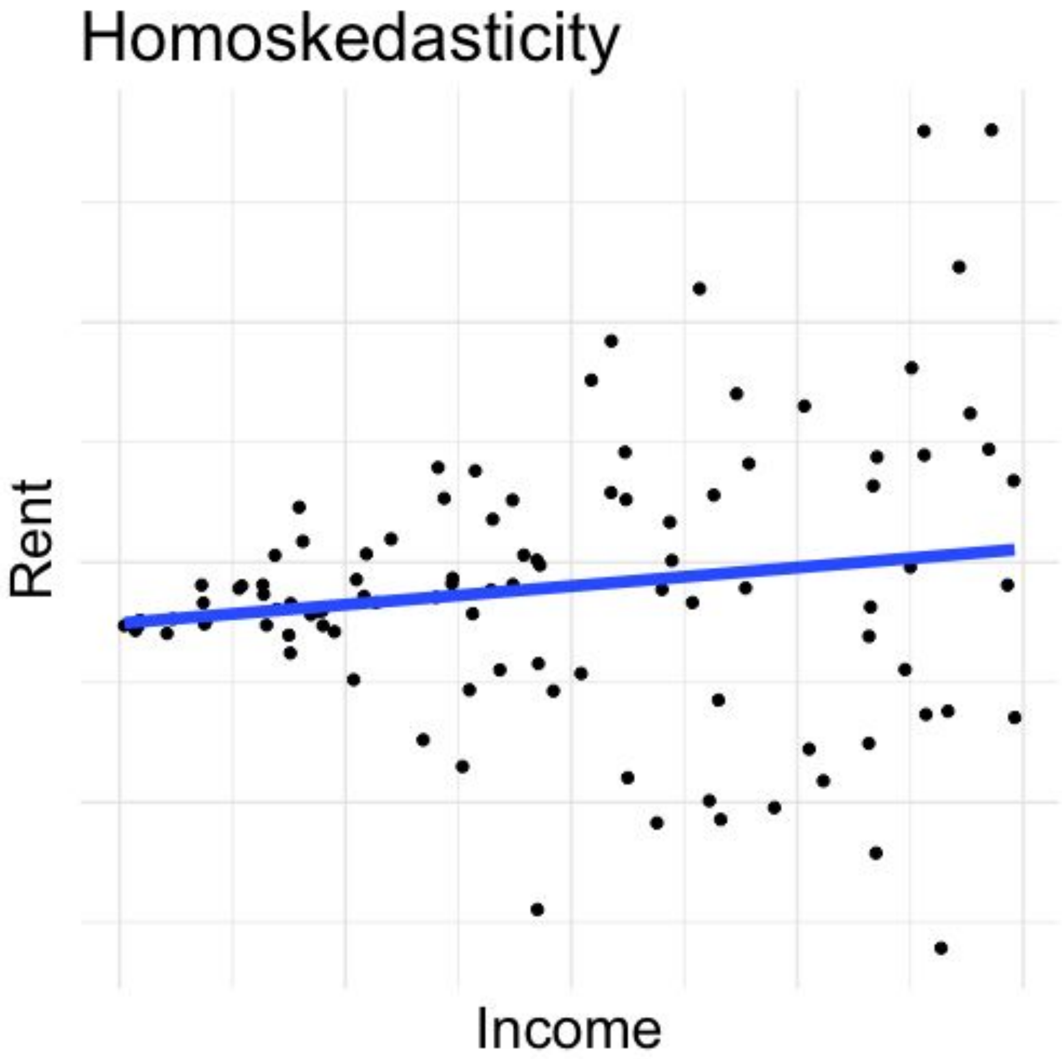
- Use **multiple variables** (can be continuous, categorical, or both) to predict a **continuous variable**.
- Use a line (or a plane) to describe the relationship between these variables.



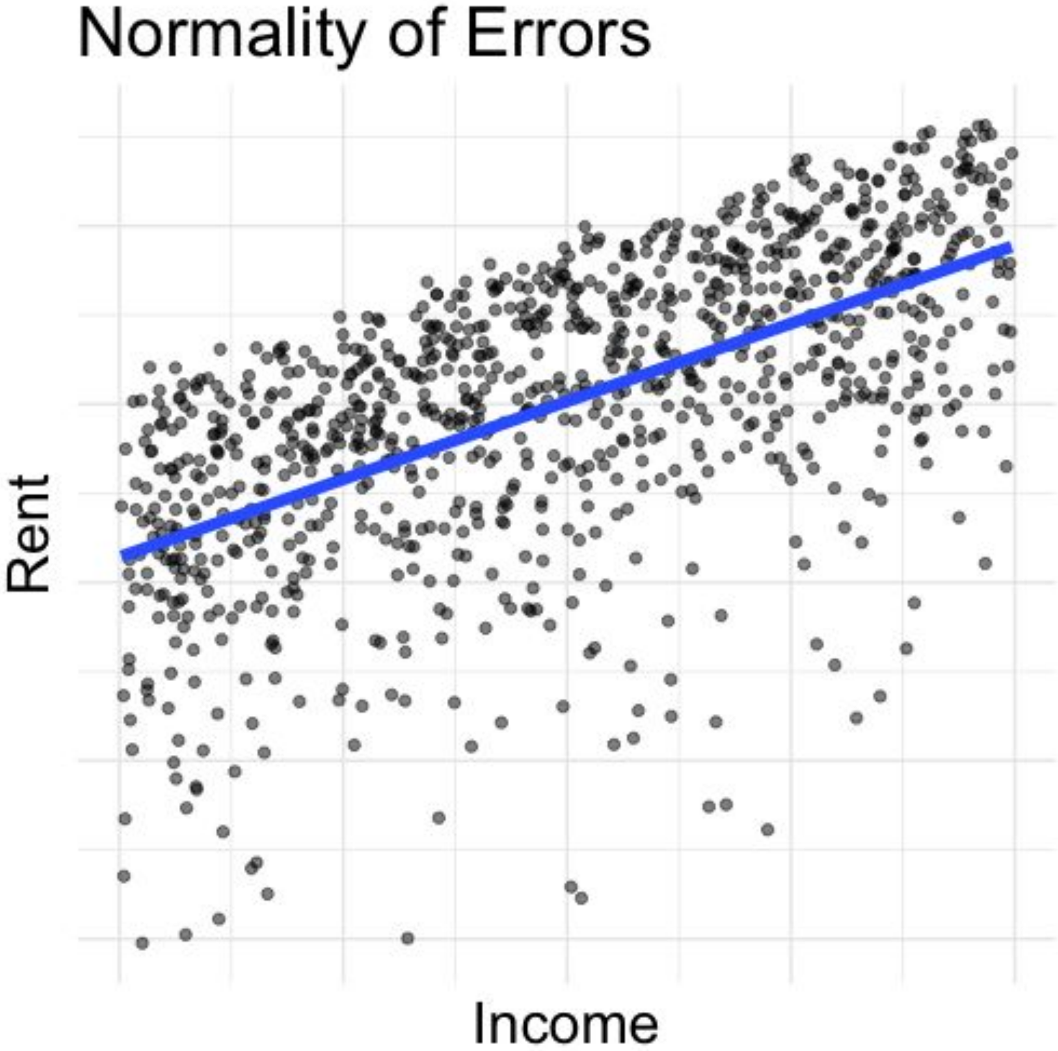
Assumptions



Assumptions



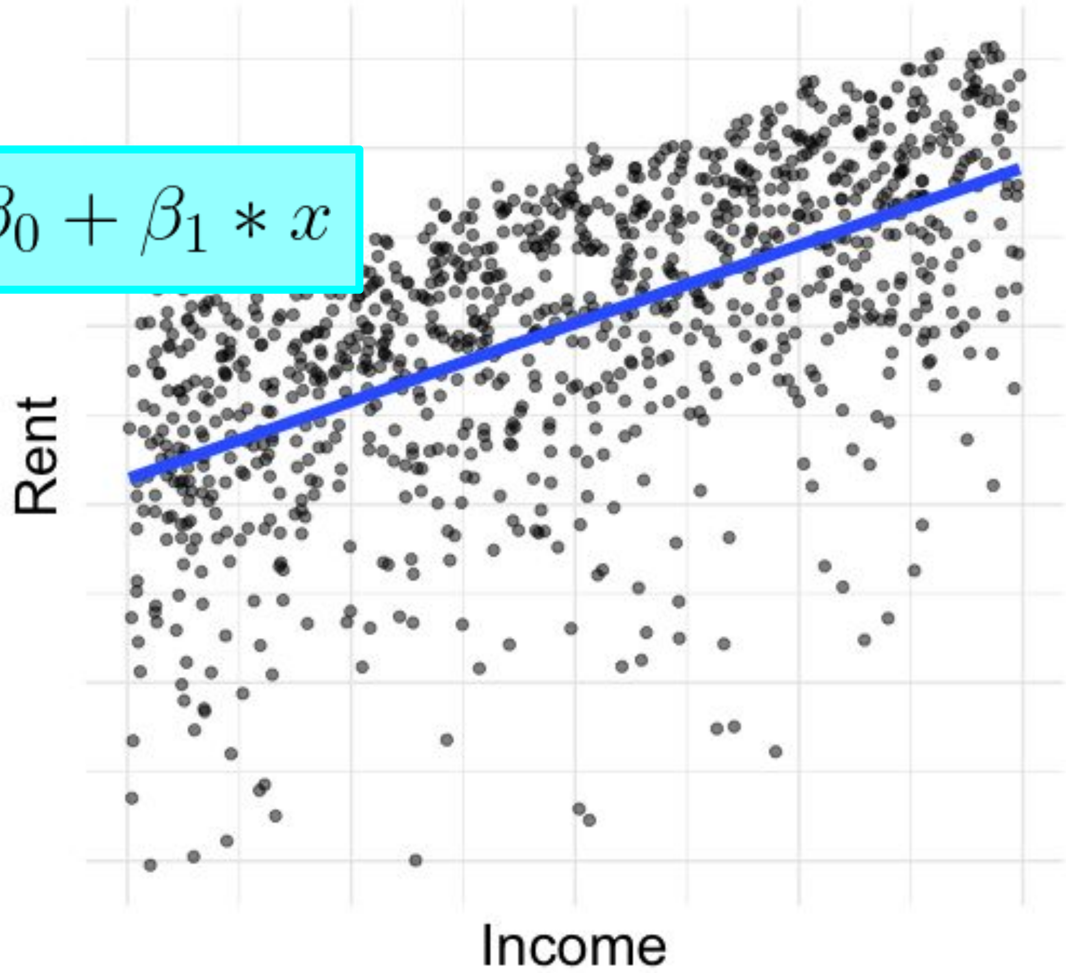
Assumptions



Normality of Errors

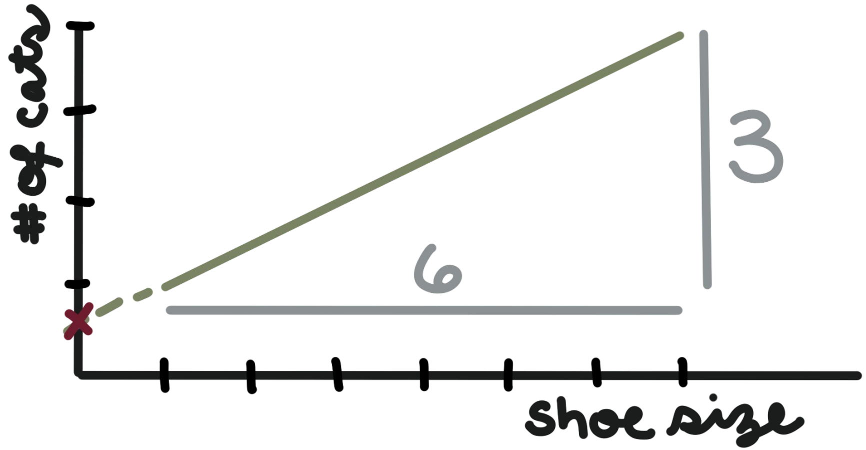
Assumptions

$$E(Y|X) = \beta_0 + \beta_1 * x$$



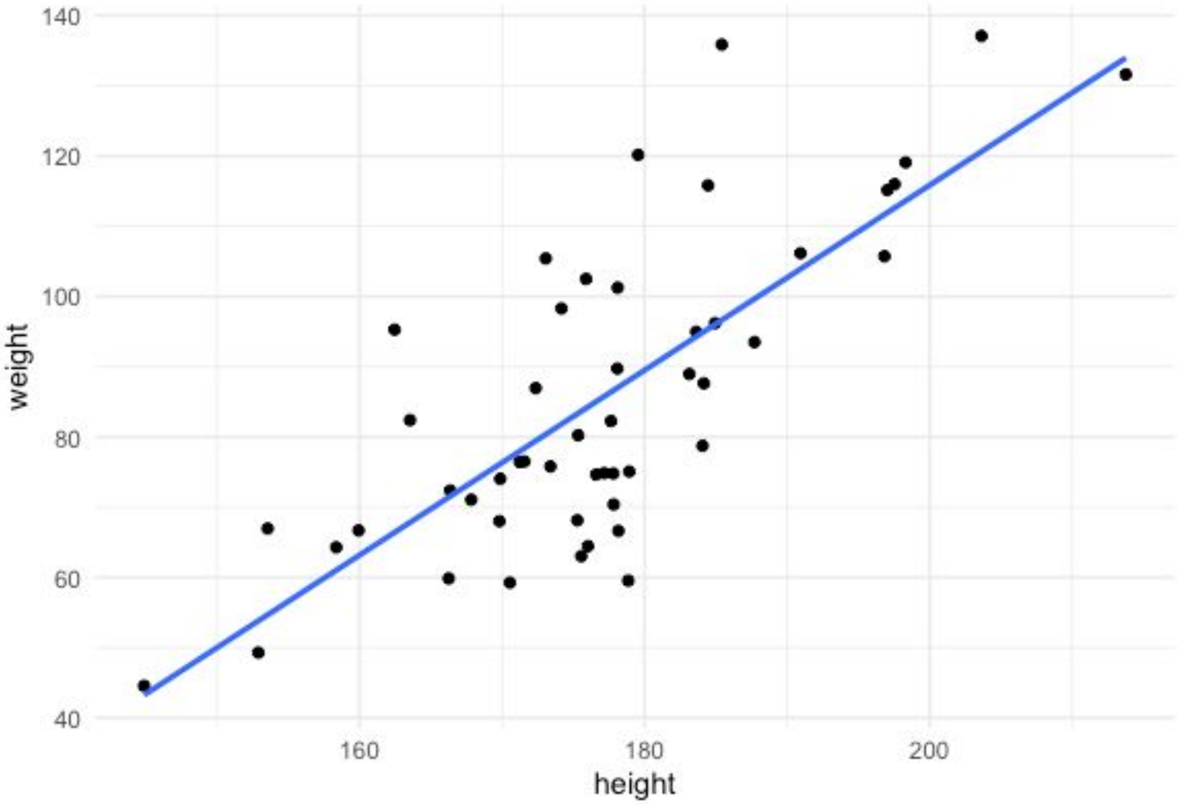
How

- Y = mx + b
- Y = mx + nz + b
- Slope tells you how variables change together
- Intercept tells you what would happen if all your predictors were 0.



Simple example

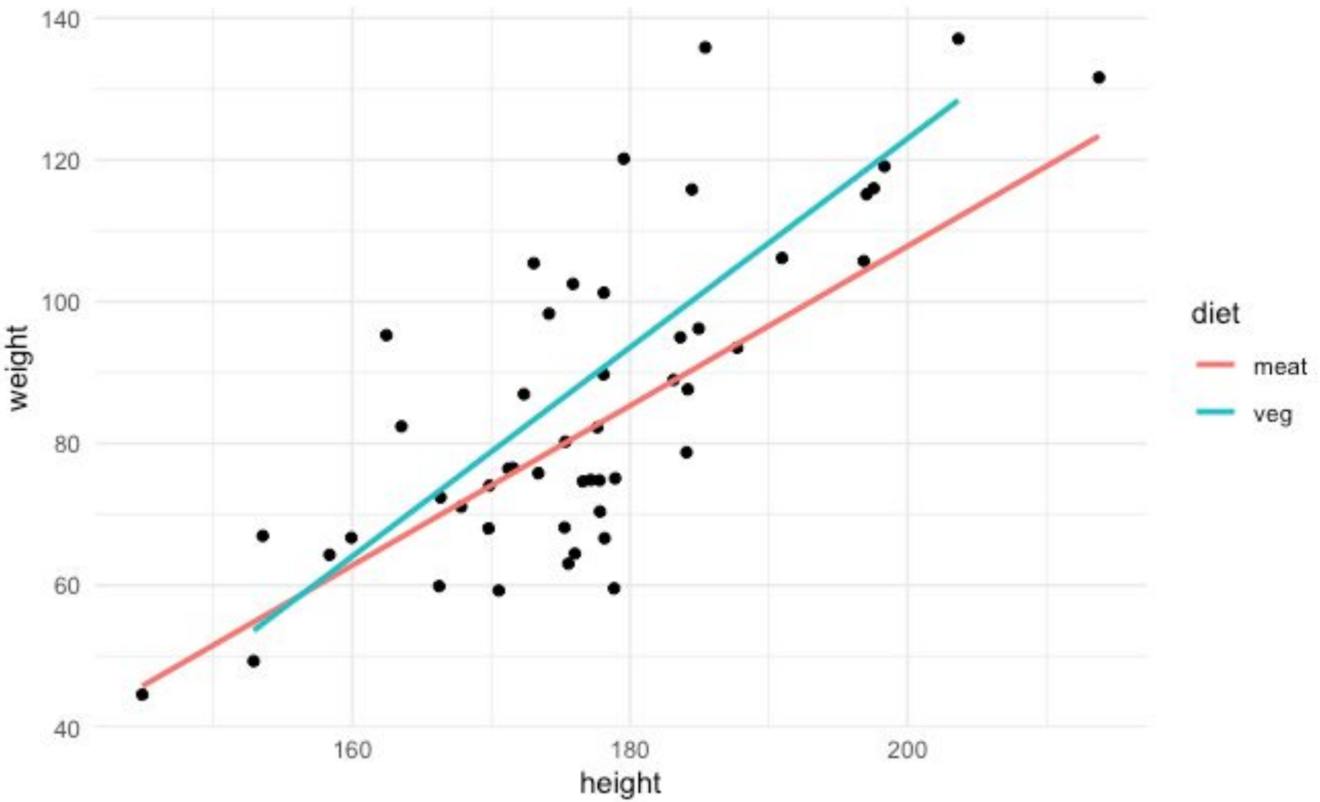
Predict weight by height



	coef
Intercept	-82.2887
height	0.9786

Simple example

Predict weight by height + diet



	coef
Intercept	-72.0358
diet[T.veg]	-7.6222
height	0.9420

Simple example

Predict weight by height + diet + age

A 1-**unit** increase in ____ causes our predicted value to (**increase/decrease**) by _____

	coef
Intercept	-57.4078
diet[T.veg]	-8.2640
height	0.8948
age	-0.1298

Z-Scoring

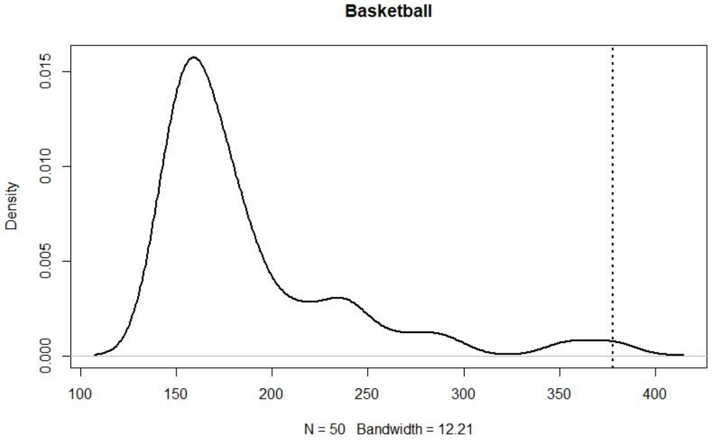
Who is the GOAT? ⚾🏀

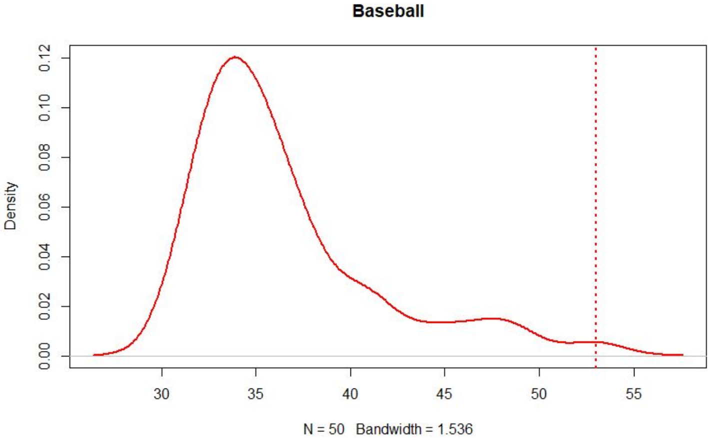
378 three-pointers

53 home-runs

[2018-19 NBA Regular Season: Total 3-Pointers Made Leaders](#)
[2019 MLB Player Batting Stats | Home Runs](#)

Who is the GOAT? ⚾🏀

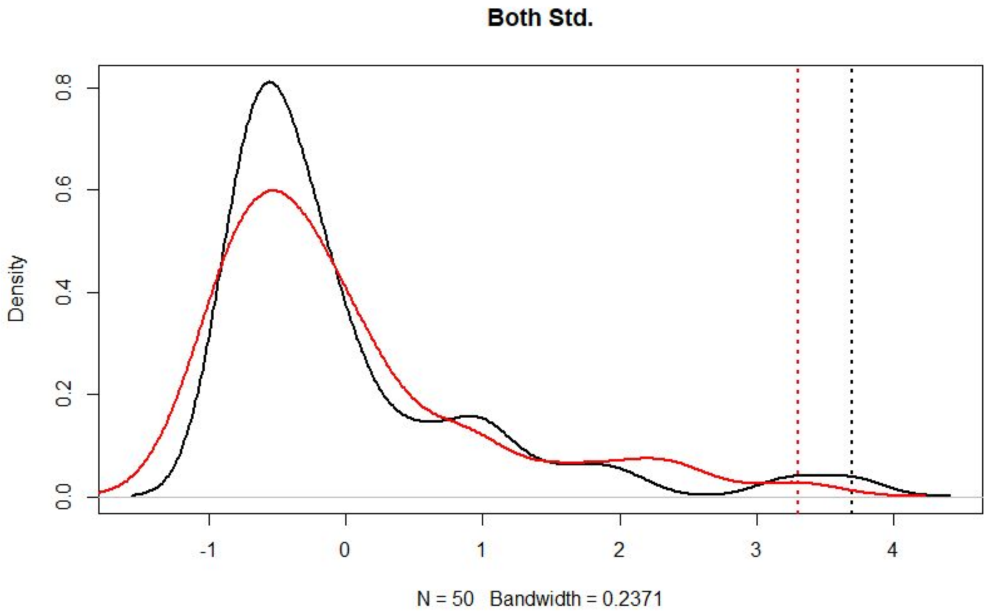




[2018-19 NBA Regular Season: Total 3-Pointers Made Leaders]
[2019 MLB Player Batting Stats | Home Runs]

Who is the GOAT? ⚾🏀

$$z = \frac{x - \bar{x}}{\sigma_x}$$



2018-19 NBA Regular Season: Total 3-Pointers Made Leaders
2019 MLB Player Batting Stats | Home Runs

Z-Score

$$z = \frac{x - \bar{x}}{\sigma_x}$$

Simple example

Predict weight by height + diet + age

A 1-**standard deviation** increase in ____ causes our predicted value to (**increase/decrease**) by _____

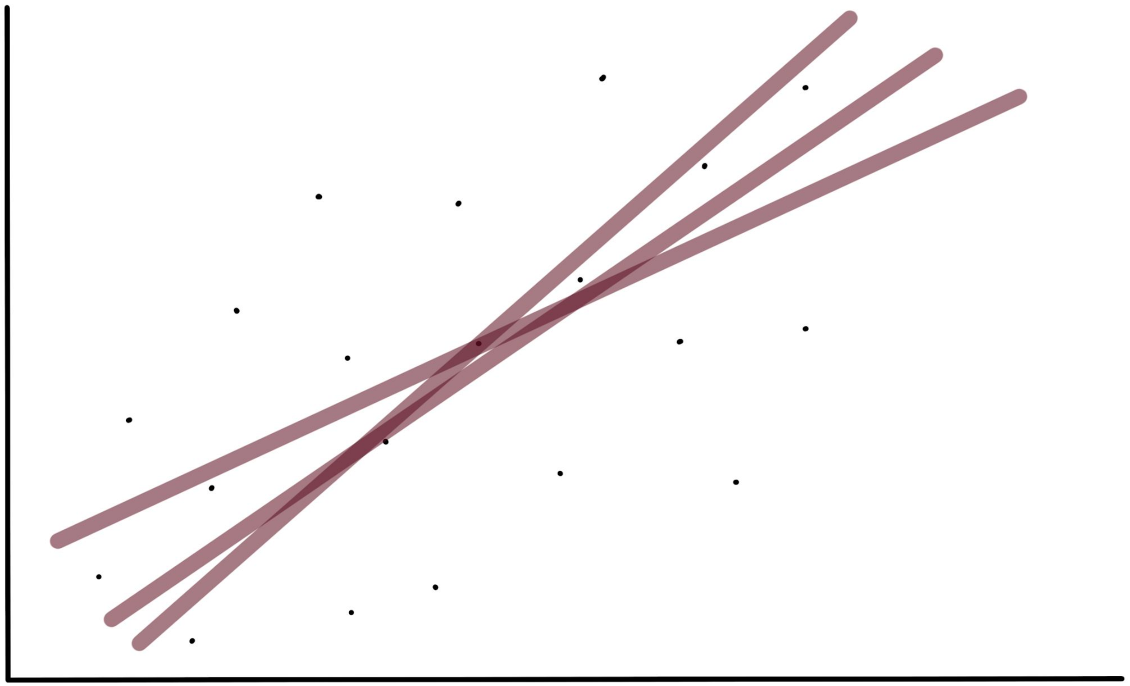
	coef
Intercept	93.6861
diet[T.veg]	-8.2640
height	13.4689
age	-2.5245

Standardizing variables

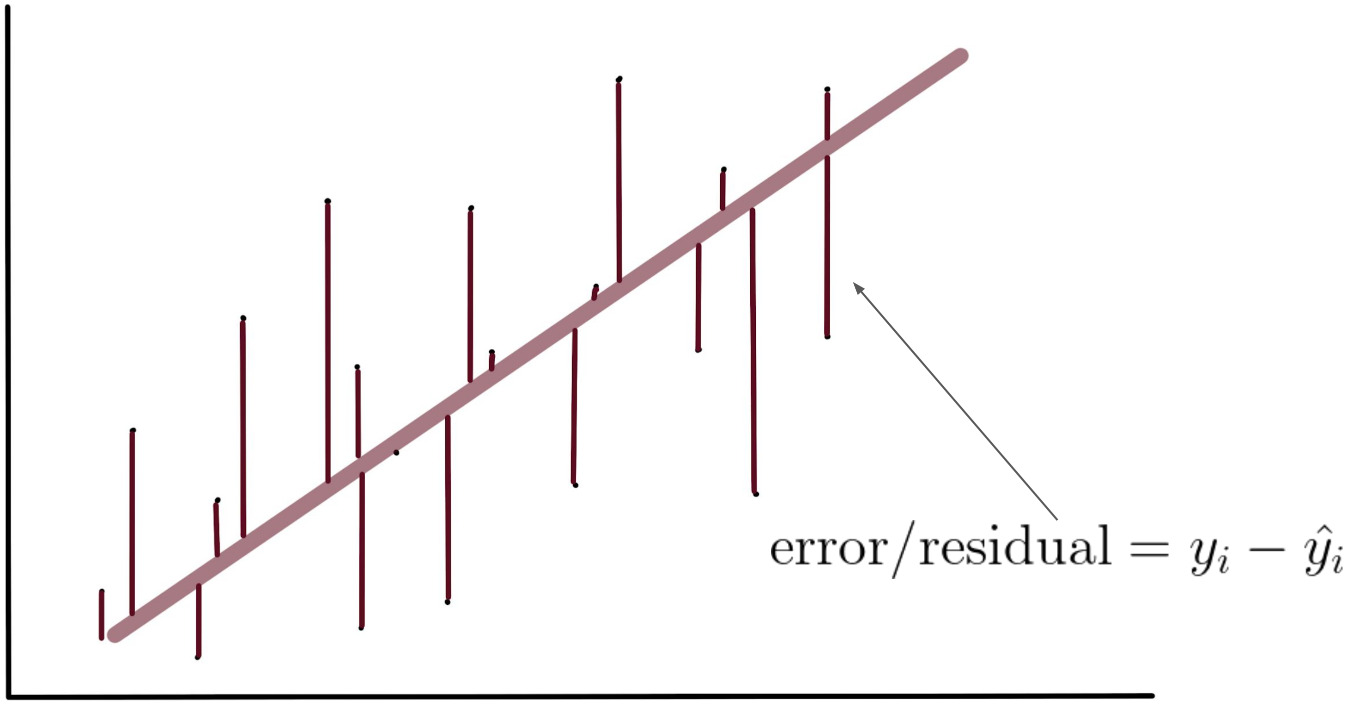
for **understanding** and for model convergence

Choosing A Line/Plane

Choosing the line of best fit

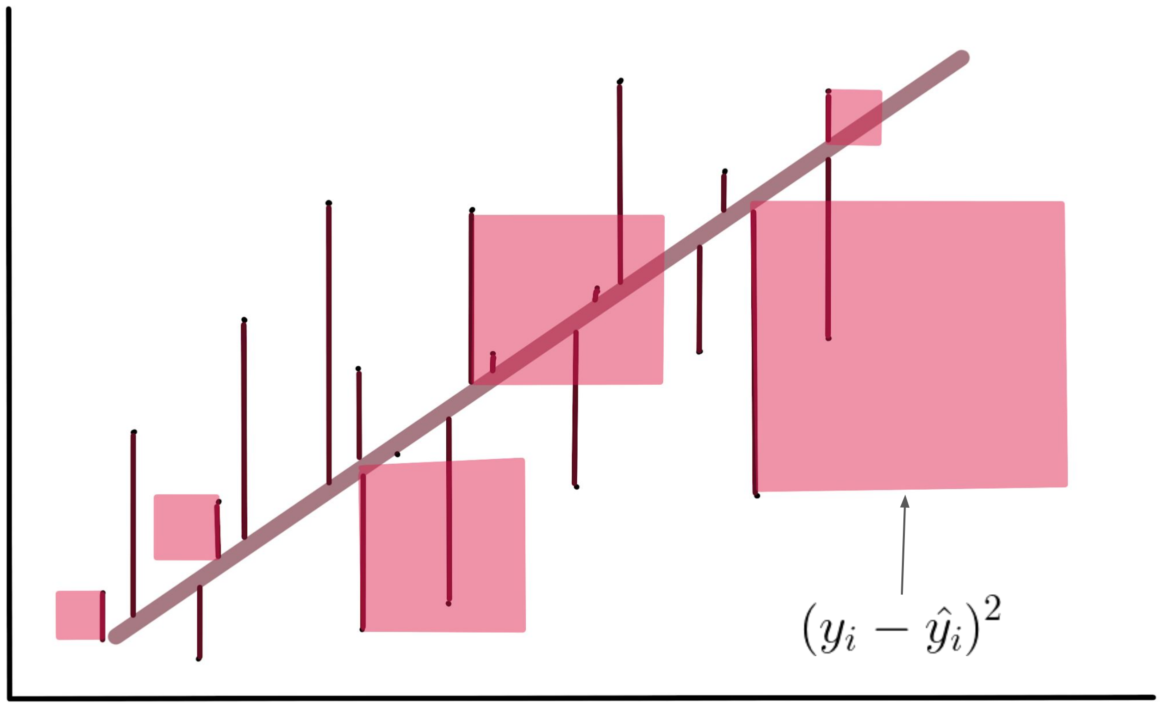


Choosing the line of best fit



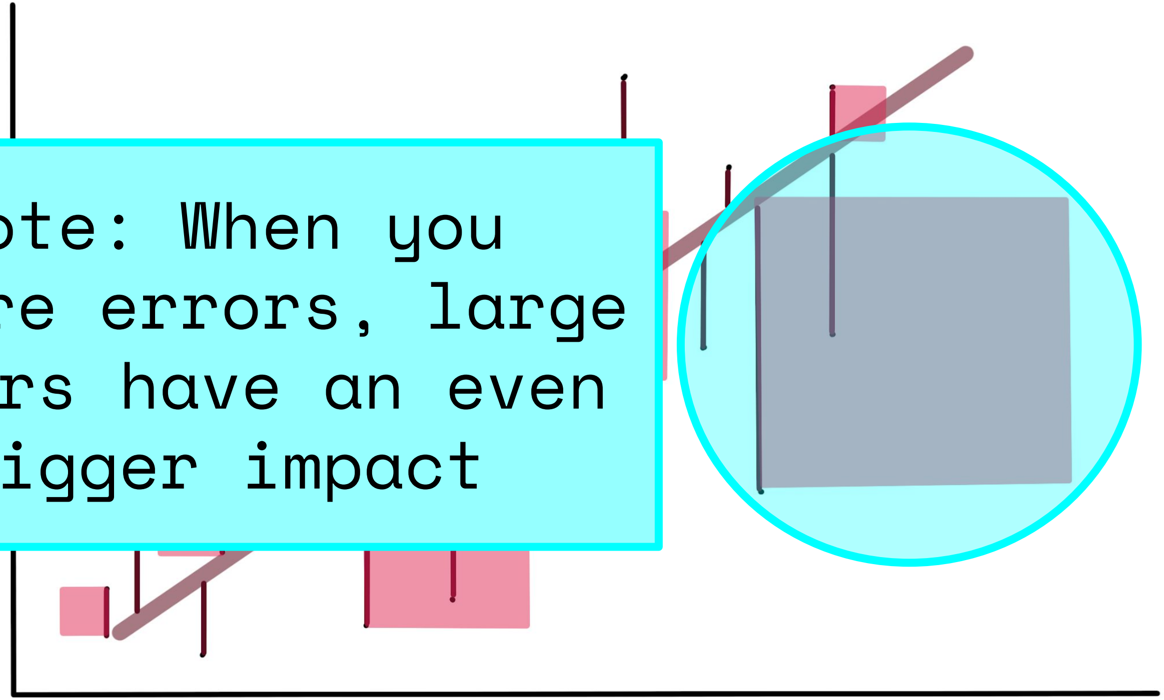
Choosing the line of best fit

- Sum of Squared Errors
- Mean Squared Error



Choosing the line of best fit

- Sum of Squared Errors
- Mean Squared Er



Note: When you square errors, large errors have an even bigger impact

Least Squares and MLE

Least Squares

$$\mathrm{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

For full proof see: https://statproofbook.github.io/P/slr-ols

Least Squares

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 * x_i)^2$$

Least Squares

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 * x_i)^2$$

$$\frac{\partial SSE}{\partial \boxed{\beta_0}} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 * x_i)(-1)$$

$$\frac{\partial SSE}{\partial \boxed{\beta_1}} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 * x_i)(-x_i)$$

For full proof see: https://statproofbook.github.io/P/slr-ols

Least Squares

$$\mathrm{SSE} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 * x_i)^2$$

$$\frac{\partial SSE}{\partial \boxed{\beta_0}} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 * x_i)(-1) = 0$$

$$\frac{\partial SSE}{\partial \boxed{\beta_1}} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 * x_i)(-x_i) = 0$$

For full proof see: https://statproofbook.github.io/P/slr-ols

Least Squares

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 * x_i)^2$$

$$\frac{\partial SSE}{\partial \boxed{\beta_0}} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 * x_i)(-1) = 0$$

$$\frac{\partial SSE}{\partial \boxed{\beta_1}} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 * x_i)(-x_i) = 0$$

$$\beta_0 = \bar{y} - \hat{\beta}_1 * \bar{x}$$

$$\beta_1 = \frac{Cov(x,y)}{Var(x)} = Corr(x,y) * \frac{sd(x)}{sd(y)}$$

For full proof see: https://statproofbook.github.io/P/slr-ols

Maximum Likelihood Estimation (MLE)

$$\theta_{MLE} = \underset{\theta \in \Theta}{\arg\max}\ L(\theta)$$

Pick parameter values that make the data likely

Maximum Likelihood Estimation (MLE)

$$\boxed{\theta_{MLE}} = \arg\max_{\theta \in \Theta} L(\theta)$$

The values we pick for our parameters

Maximum Likelihood Estimation (MLE)

$$\theta_{MLE} = \boxed{\arg\max_{\theta \in \Theta}} L(\theta)$$

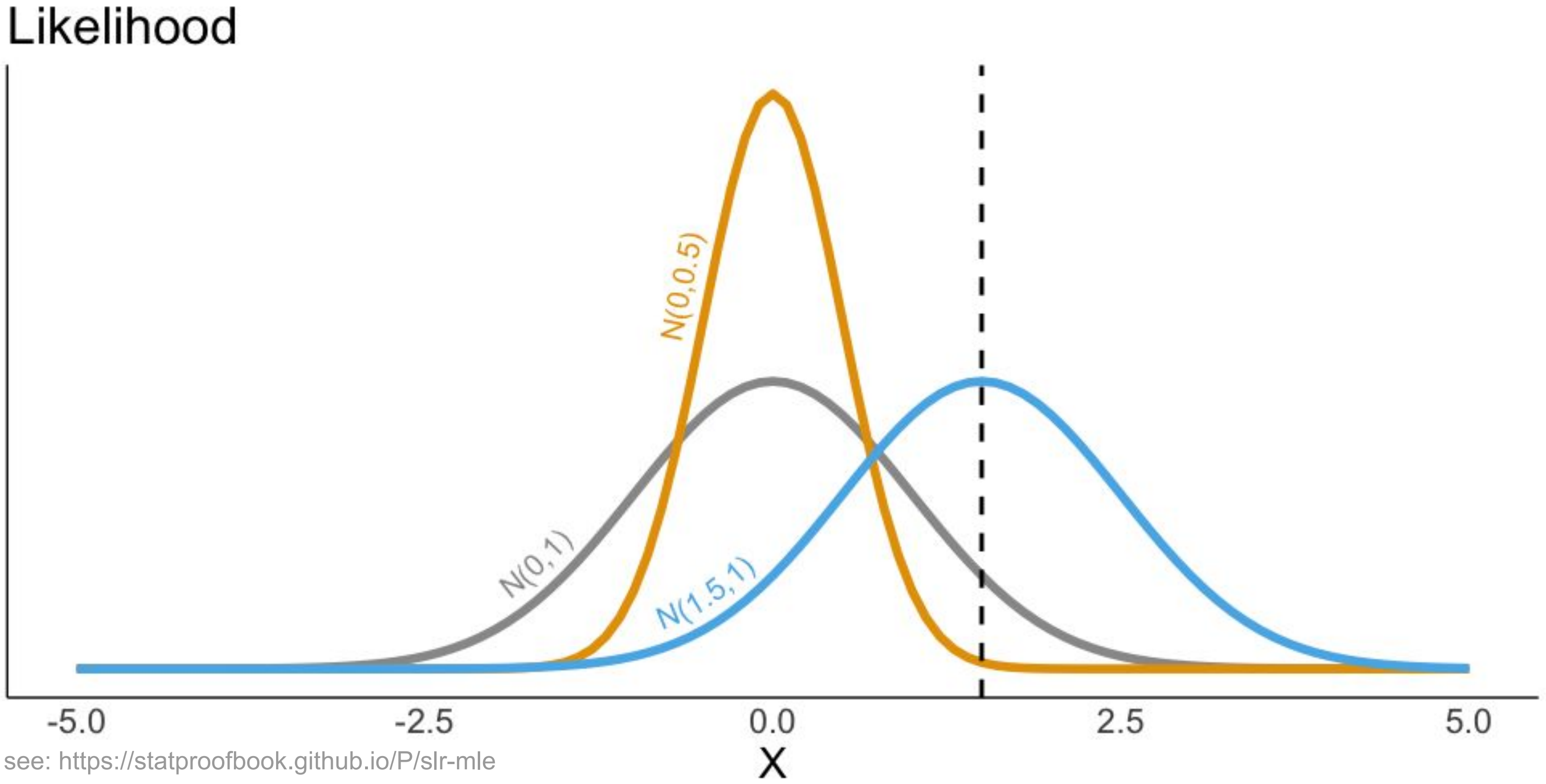
are the parameter
values (out of all
possible parameter
values) that maximize

Maximum Likelihood Estimation (MLE)

$$\theta_{MLE} = \arg\max_{\theta \in \Theta} \boxed{L(\theta)}$$

the likelihood
of the data
using these
parameters

Maximum Likelihood Estimation (MLE)



For full proof see: https://statproofbook.github.io/P/slr-mle

Maximum Likelihood Estimation (MLE)

$$\beta_0 = \bar{y} - \hat{\beta}_1 * \bar{x}$$

$$\beta_1 = \frac{Cov(x, y)}{Var(x)} = Corr(x, y) * \frac{sd(x)}{sd(y)}$$

Assessing Model Fit

How to Measure Model Success

$$MSE = \frac{1}{n}\sum_{i}(actual_i - predicted_i)^2$$

How to Measure Model Success

$$MSE = \frac{1}{n}\sum_{i}(actual_i - predicted_i)^2$$

Loss Function

that assesses performance, smaller is better

How to Measure Model Success

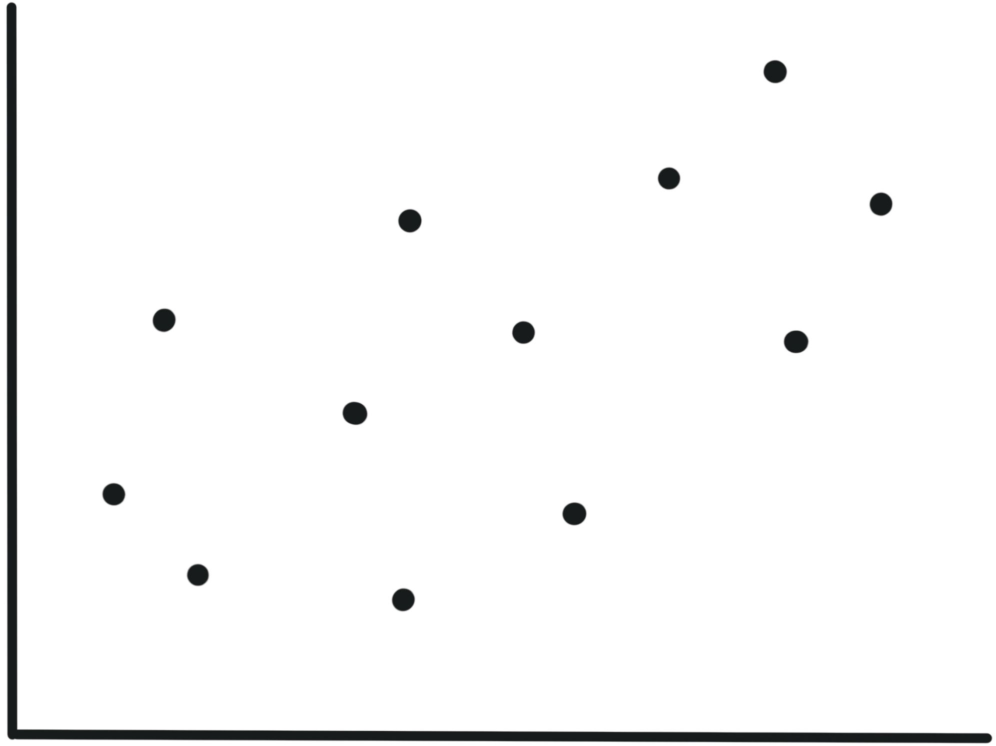
$$MAE = \frac{1}{n}\sum_{i}|actual_i - predicted_i|$$

How to Measure Model Success

$$R^2 = 1 - \frac{\sum_i (actual_i - predicted_i)^2}{\sum_i (actual_i - average)^2}$$

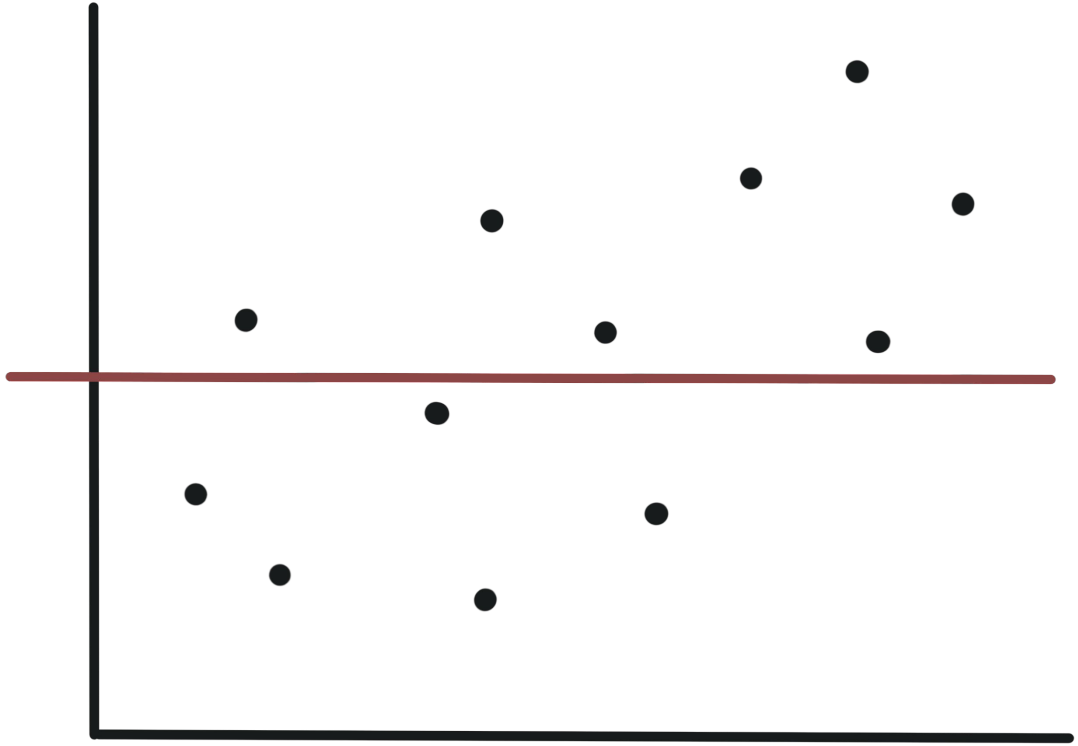
How to Measure Model Success

R^2:



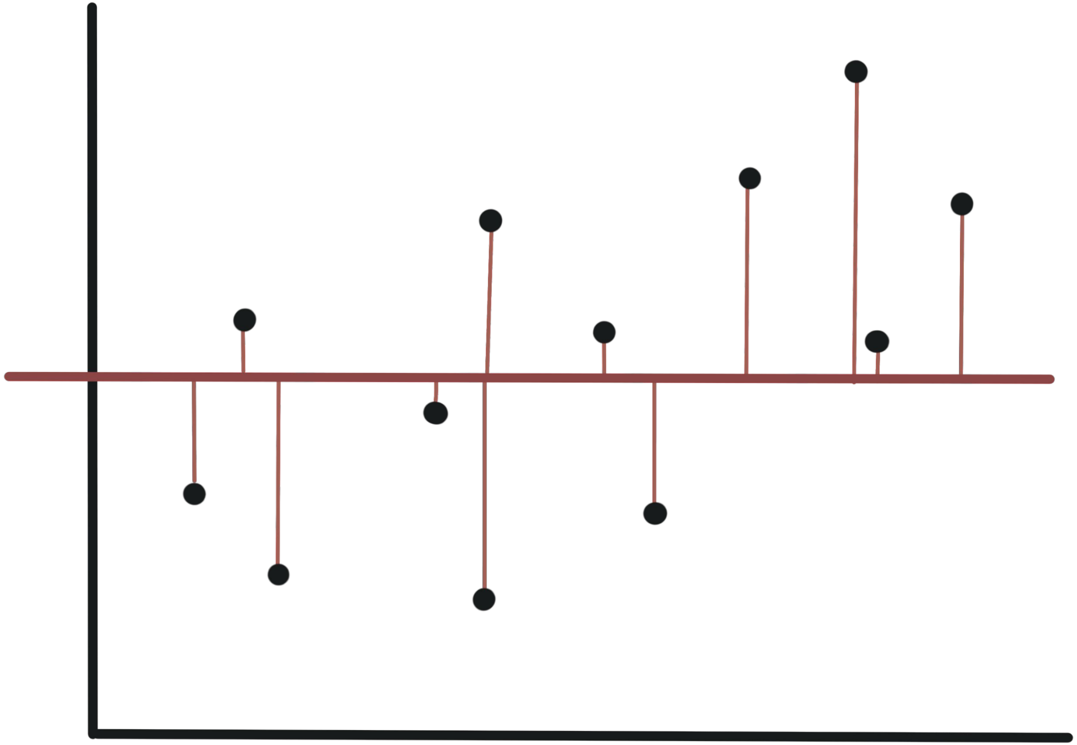
How to Measure Model Success

R^2:



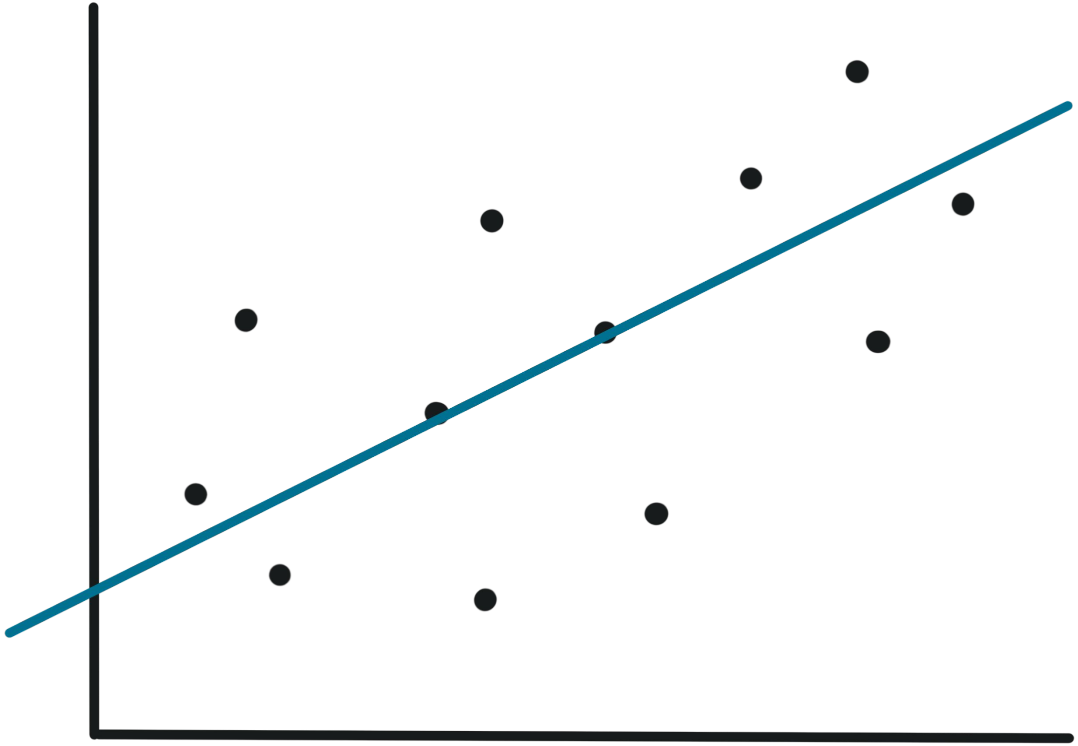
How to Measure Model Success

R^2:



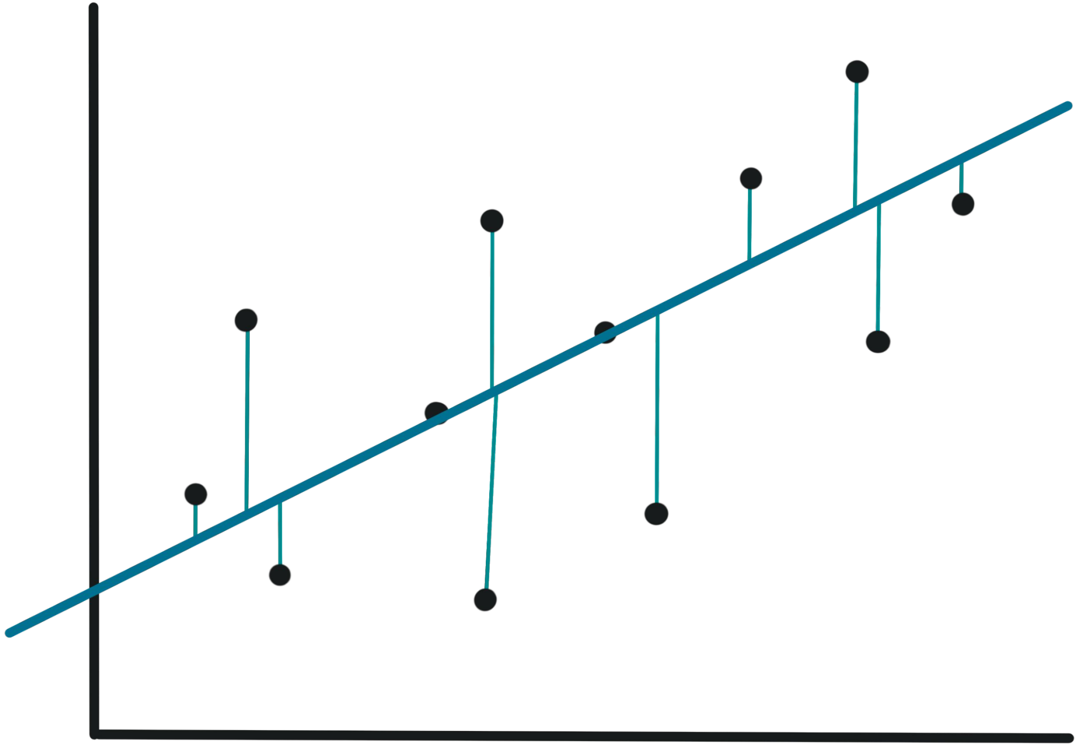
How to Measure Model Success

R^2:



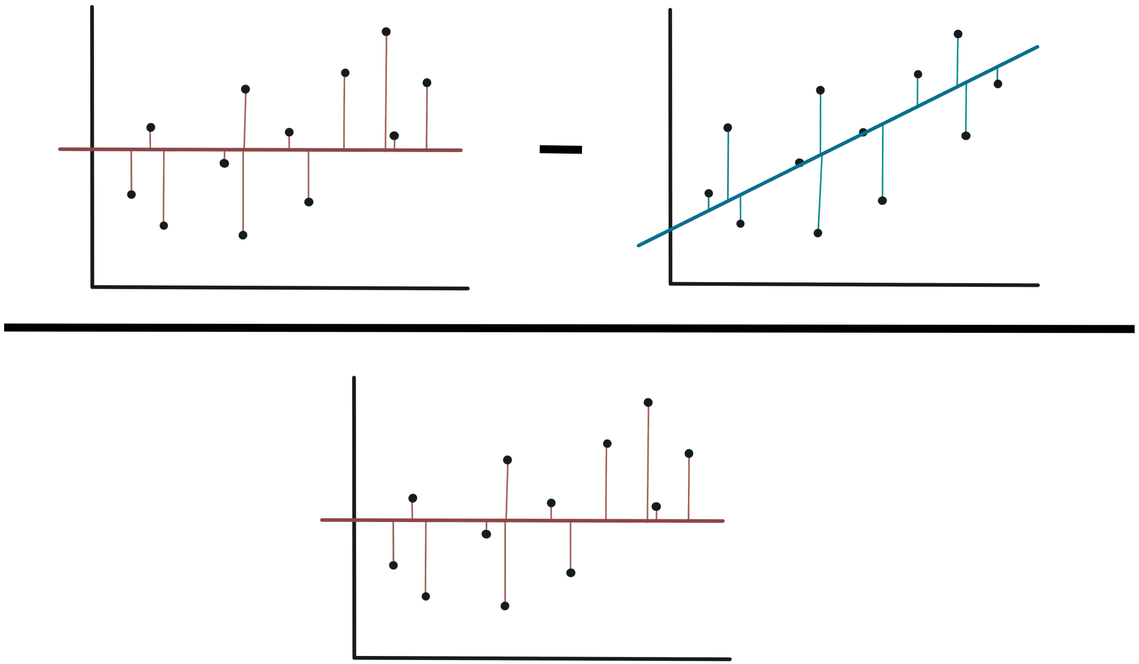
How to Measure Model Success

R^2:



How to Measure Model Success

R^2:



How to Measure Model Success

$$R^2 = 1 - \frac{\sum_i (actual_i - predicted_i)^2}{\sum_i (actual_i - average)^2}$$

How to Measure Model Success

$$MAPE = \frac{1}{n}\sum_{i}\left|\frac{acutal_i - predicted_i}{actual_i}\right|$$

How to Measure Model Success

$$MAPE = \frac{1}{n}\sum_i \left|\frac{acutal_i - predicted_i}{actual_i}\right|$$

$$R^2 = 1 - \frac{\sum_i (actual_i - predicted_i)^2}{\sum_i (actual_i - average)^2}$$

$$MSE = \frac{1}{n}\sum_i (actual_i - predicted_i)^2$$

$$MAE = \frac{1}{n}\sum_i |actual_i - predicted_i|$$