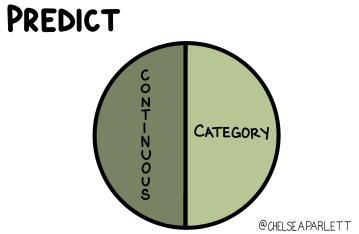


Linear Regression

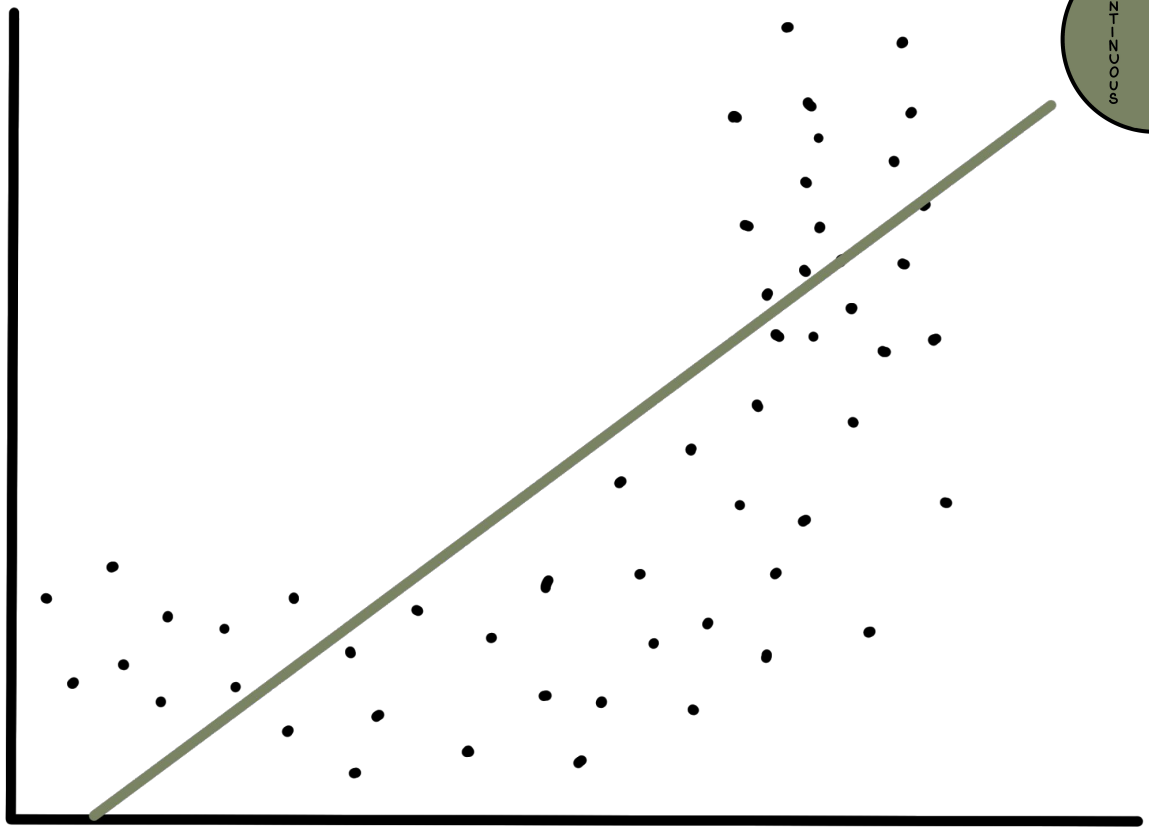
Chelsea Parlett-Pelleriti

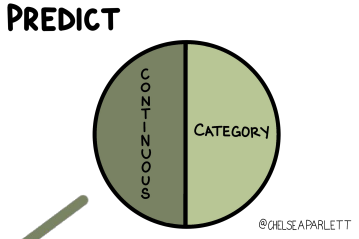
What

- 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

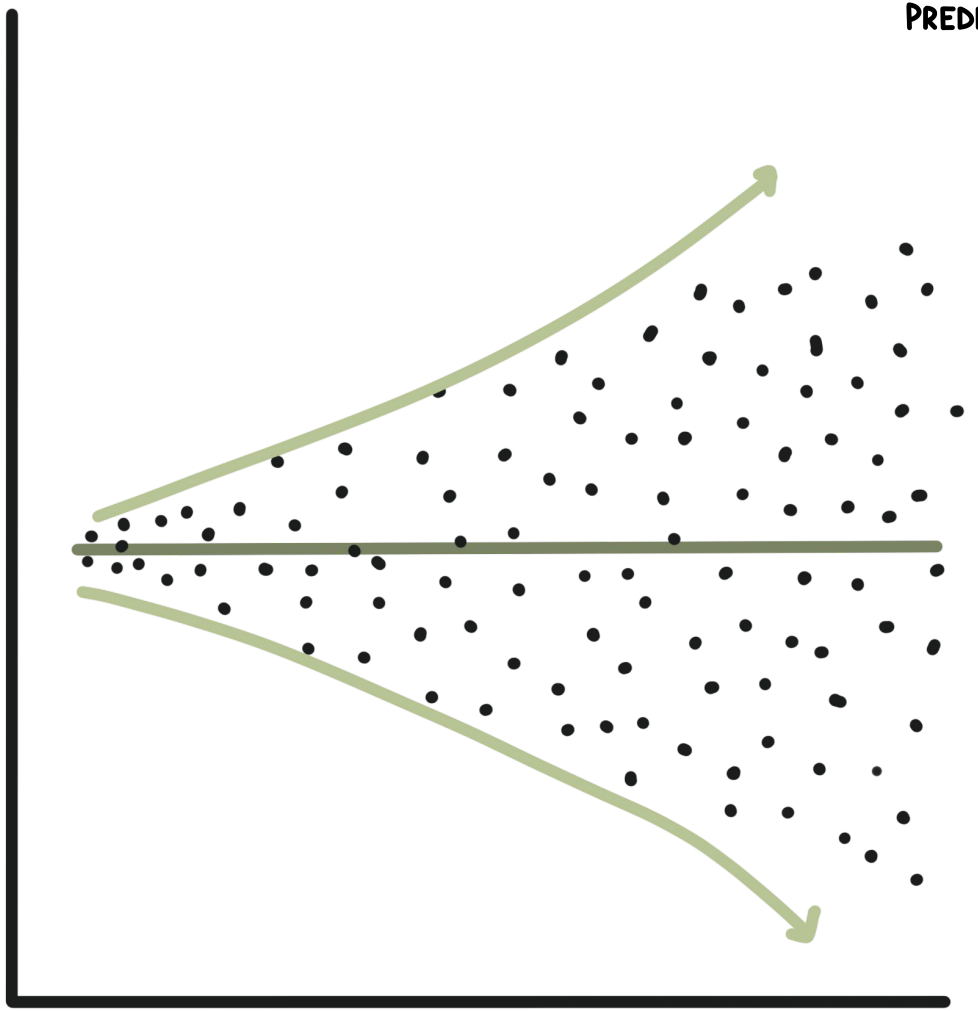
- The relationship between your variables is linear

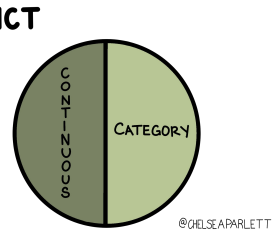




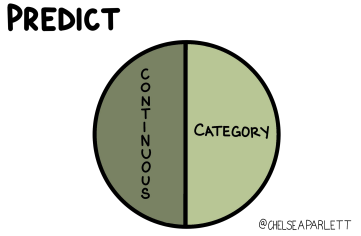
Assumptions

- Homoscedasticity
  - Is the mode worse in some areas than others?
- Normality of Errors

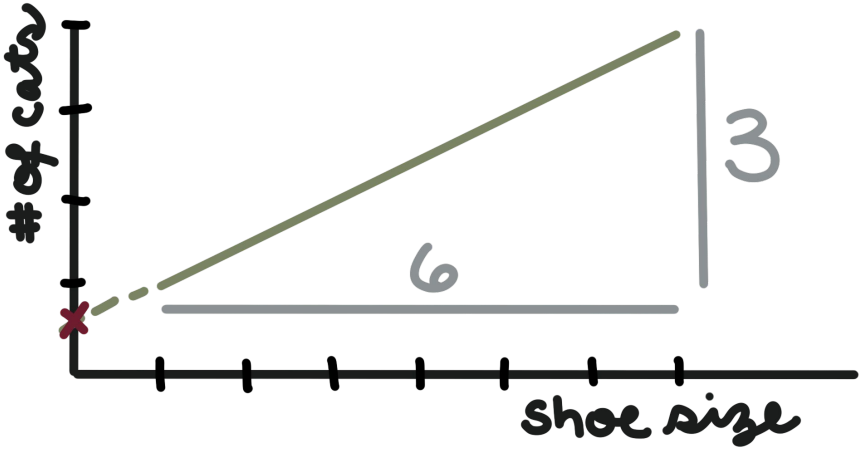




How

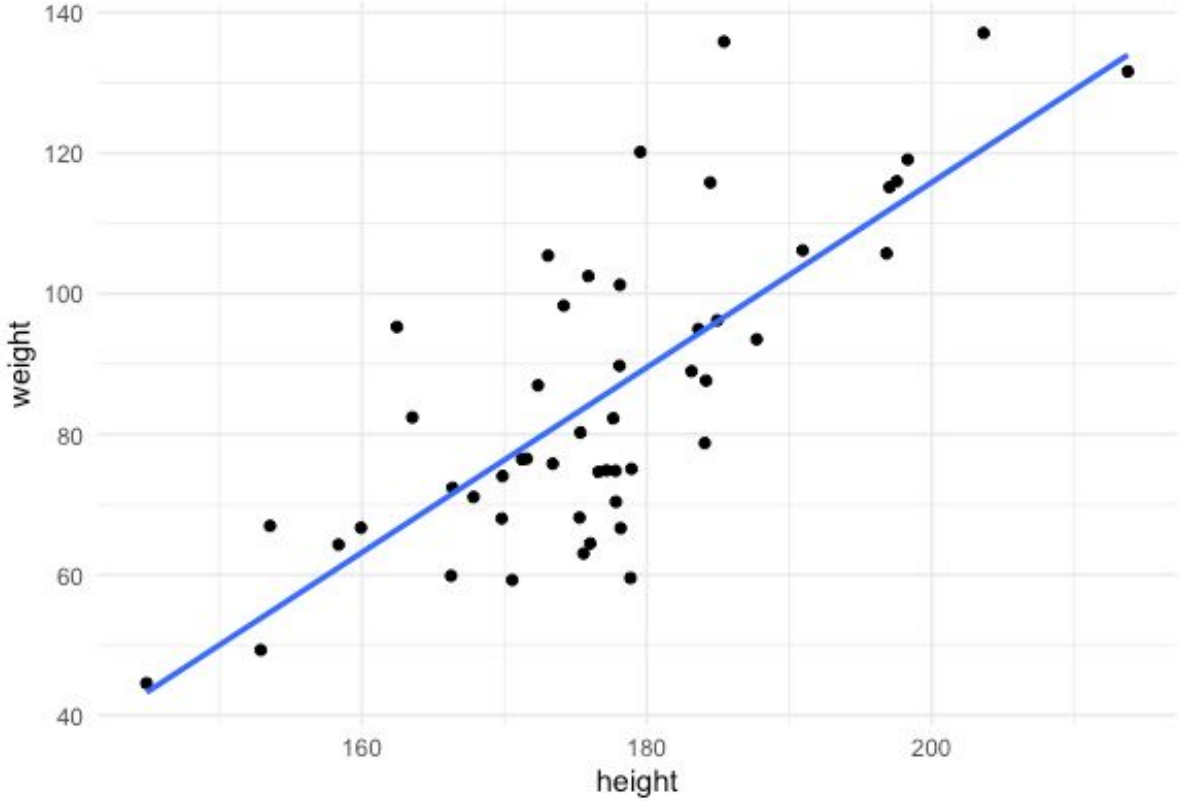


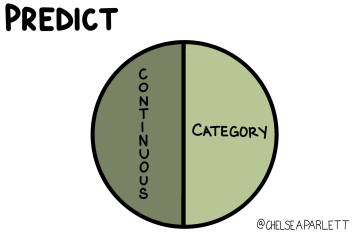
- Y = mx + b
- Y = mx + nz + b
- Slope tells you how variables change together
- Intercept tells you what would happen if all your predictors were 0.



Simple example

Predict weight by height

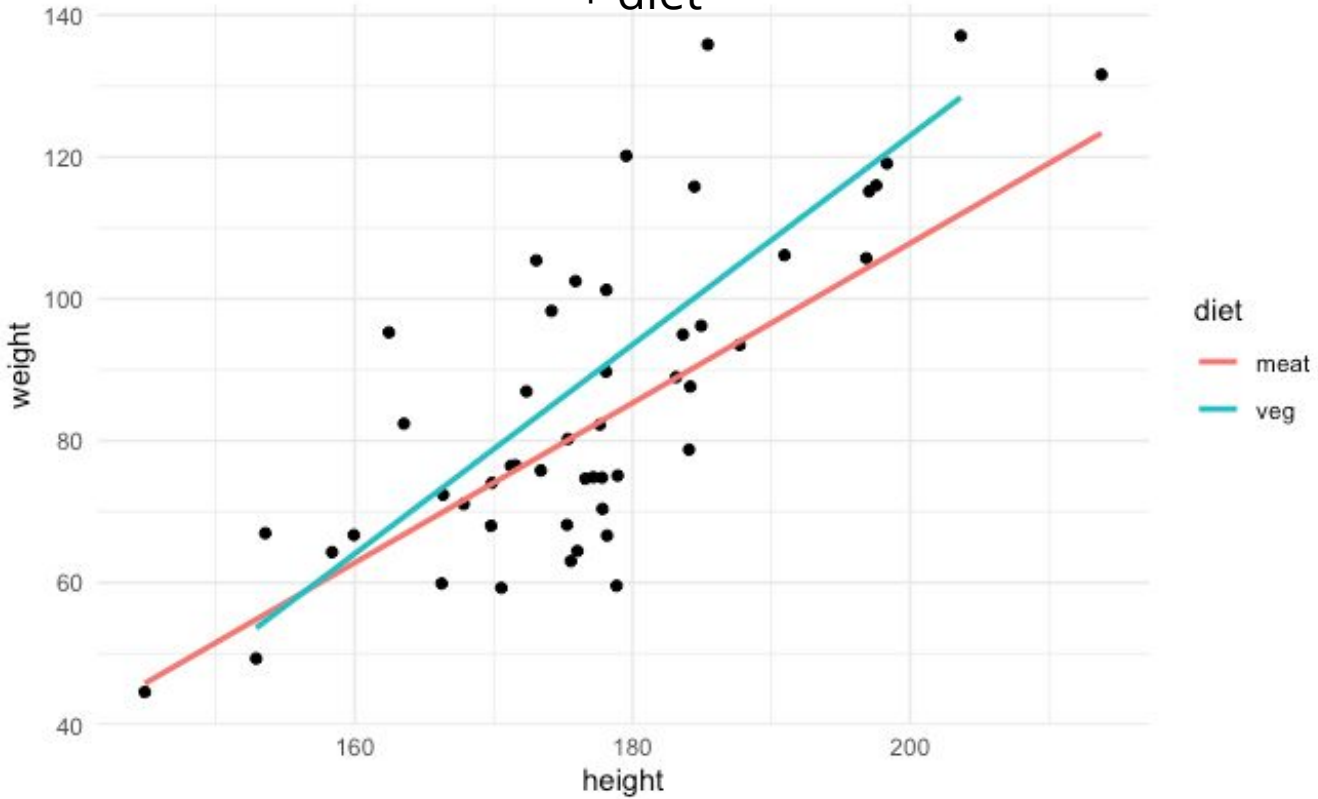


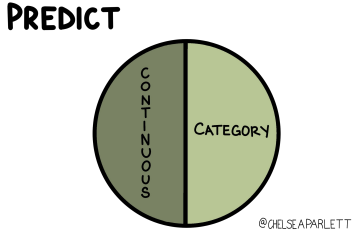


	coef
Intercept	-82.2887
height	0.9786

Simple example

Predict weight by height + diet

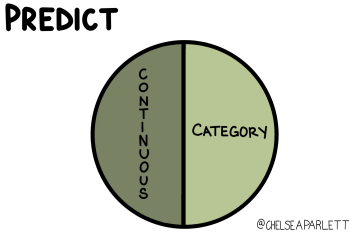




	coef
Intercept	-72.0358
diet[T.veg]	-7.6222
height	0.9420

Simple example

Predict weight by height + diet + age



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

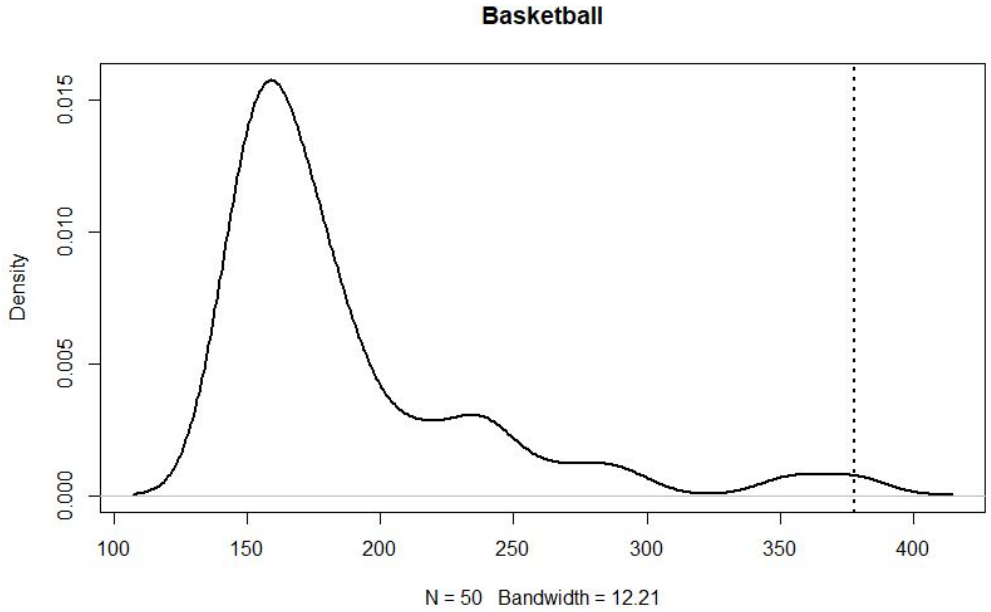
Who is the GOAT? ⚾ 🏀

378 three-pointers

53 home-runs

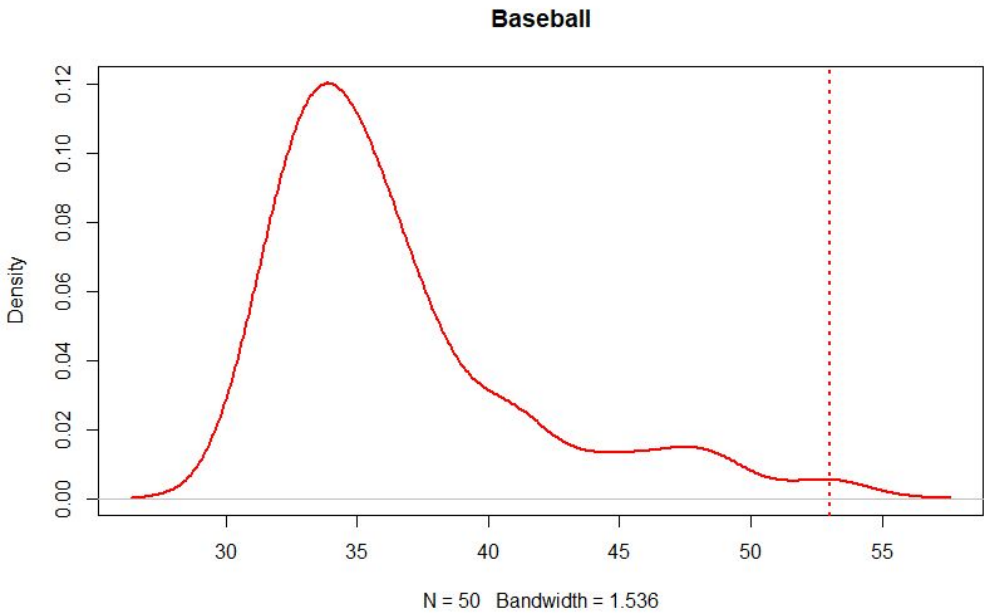
2018-19 NBA Regular Season: Total 3-Pointers Made Leaders
2019 MLB Player Batting Stats | Home Runs

Who is the GOAT?



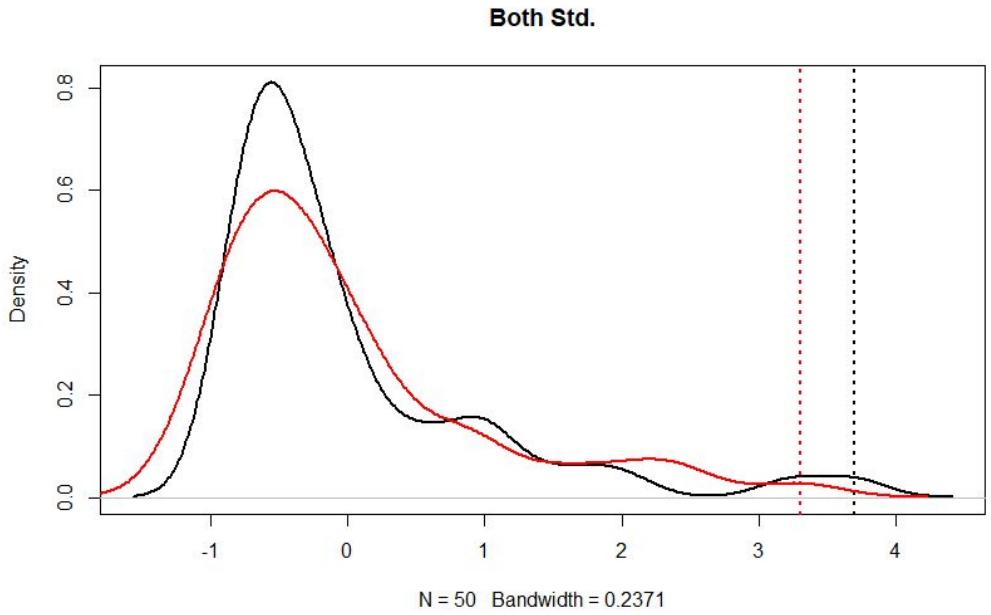
2018-19 NBA Regular Season: Total 3-Pointers Made Leaders
2019 MLB Player Batting Stats | Home Runs

Who is the GOAT?



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

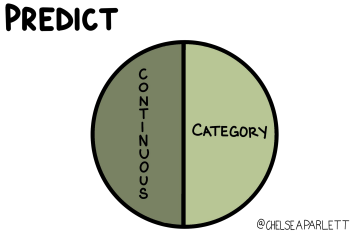
Who is the GOAT?



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

Simple example

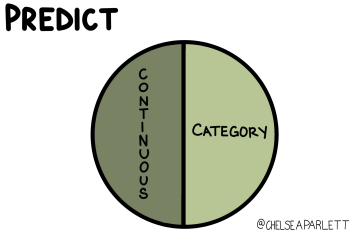
Predict weight by height + diet + age



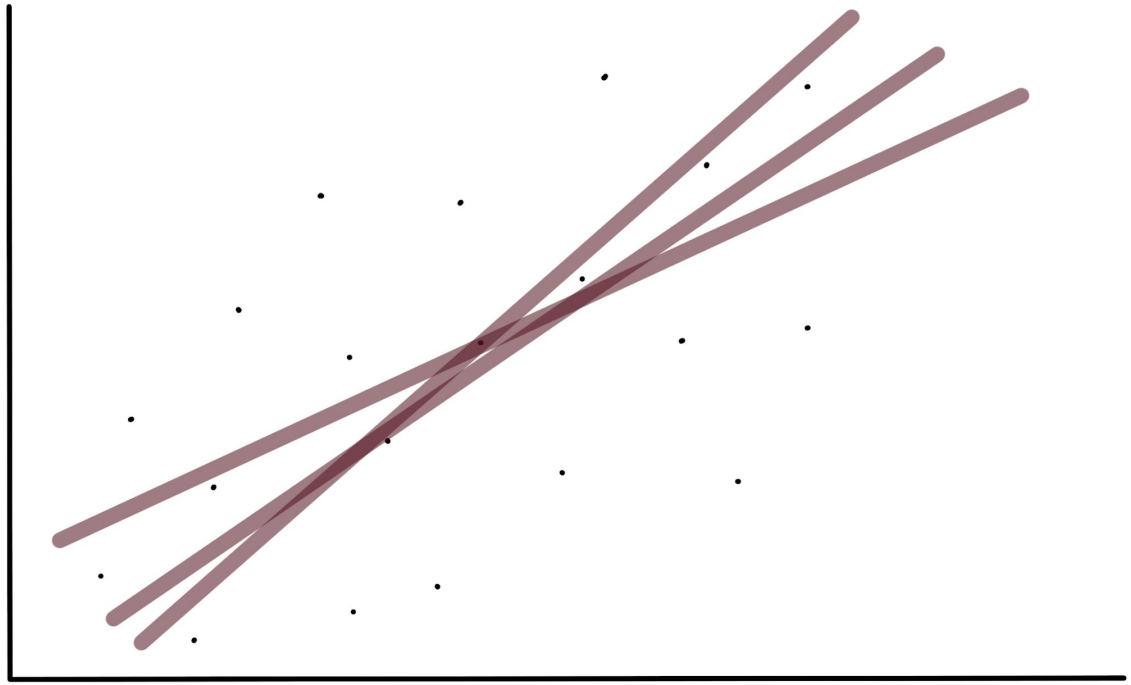
	coef
Intercept	93.6861
diet[T.veg]	-8.2640
height	13.4689
age	-2.5245

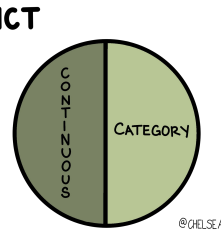
Standardizing variables

For understanding and for model convergence

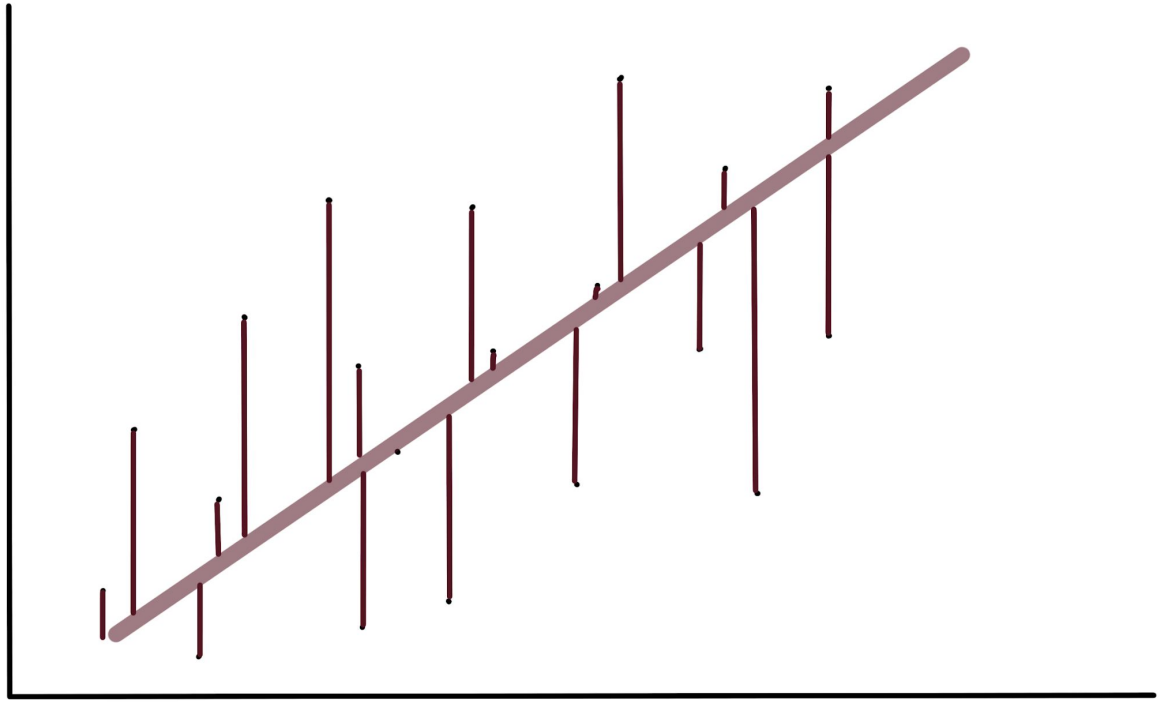


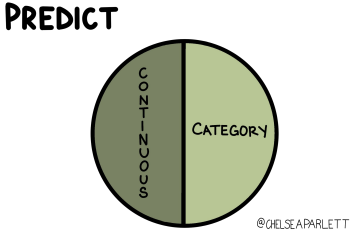
Choosing the line of best fit



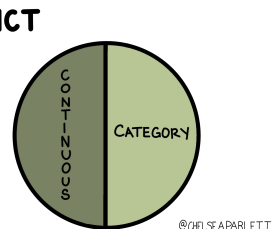


Choosing the line of best fit

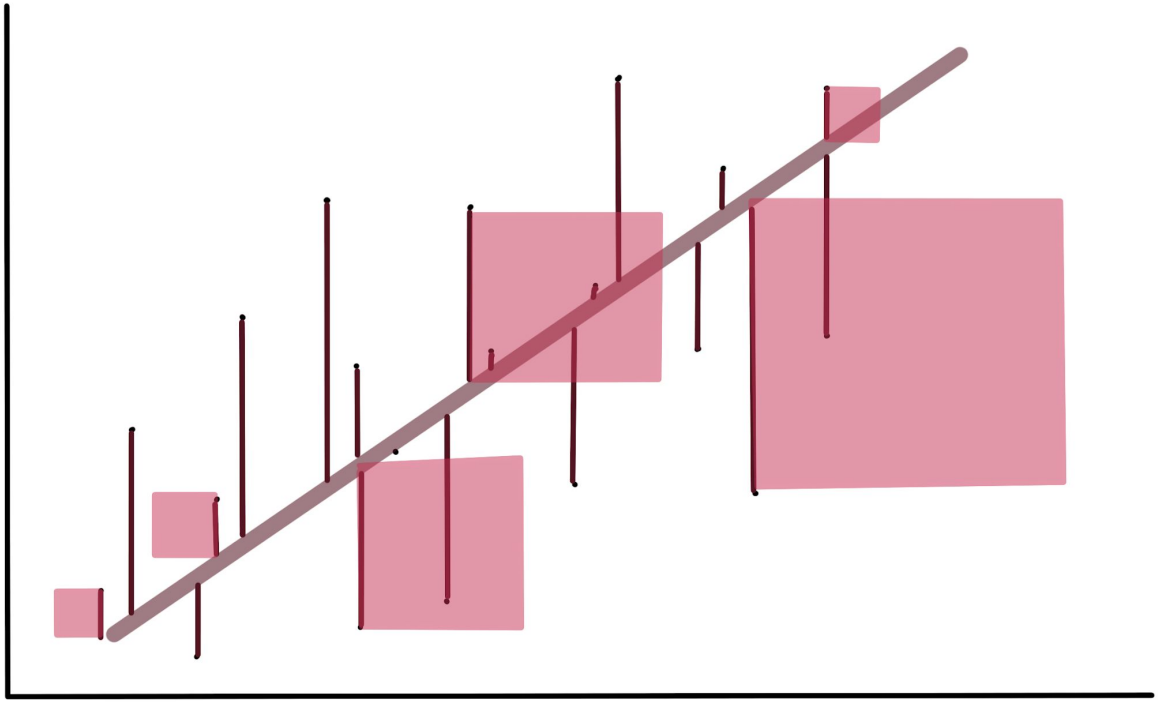




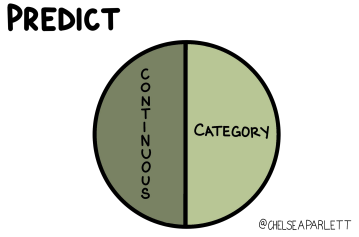
Choosing the line of best fit



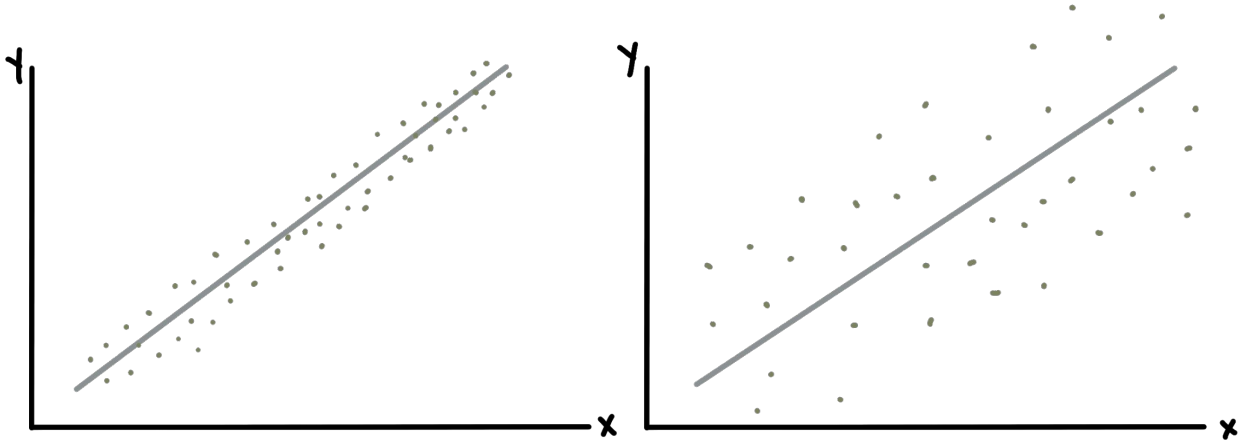
- Sum of Squared Errors
- Mean Squared Error



Choosing the line of best fit



- Sum of Squared Errors
- **Mean Squared Error**

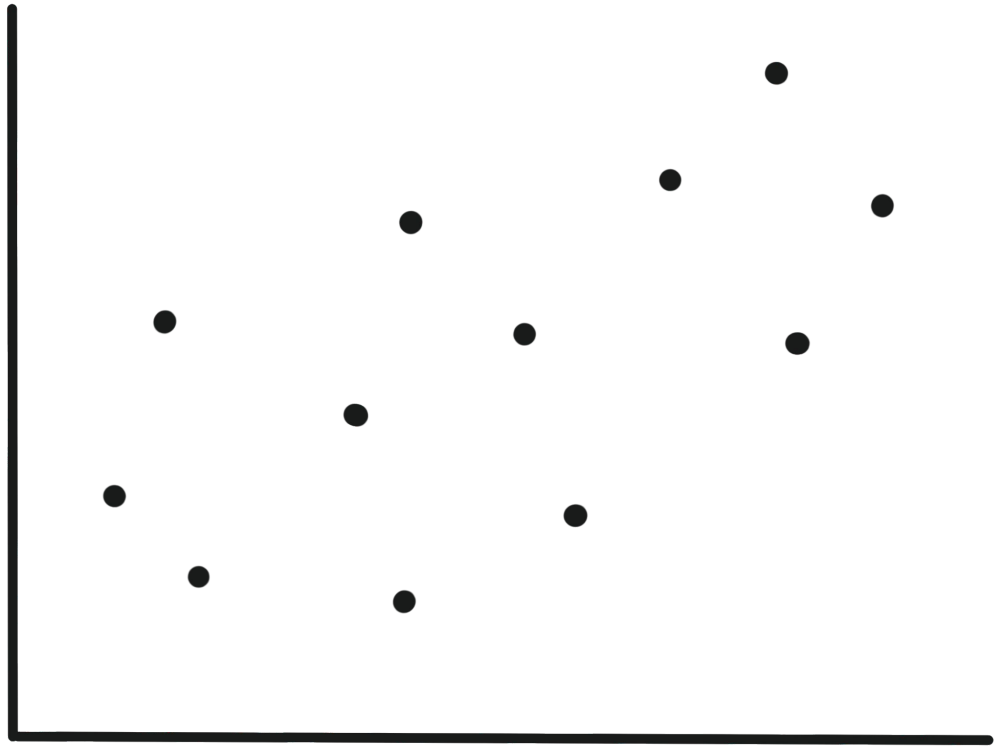


How to Measure Model Success

MSE:

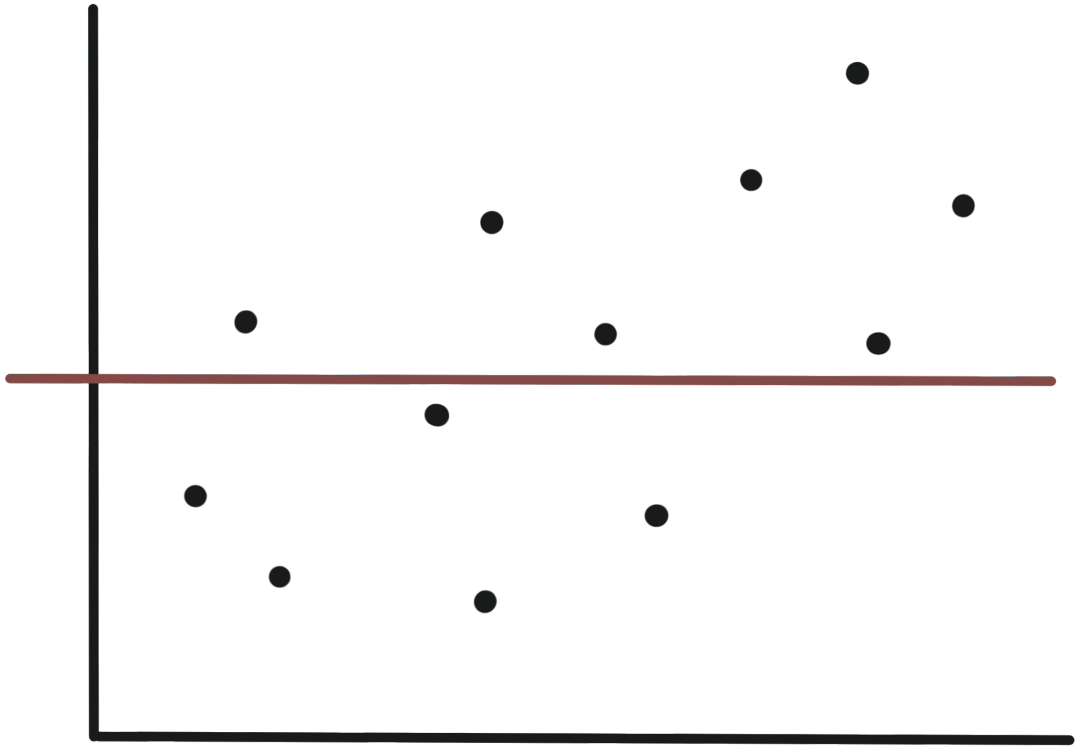
How to Measure Model Success

R^2:



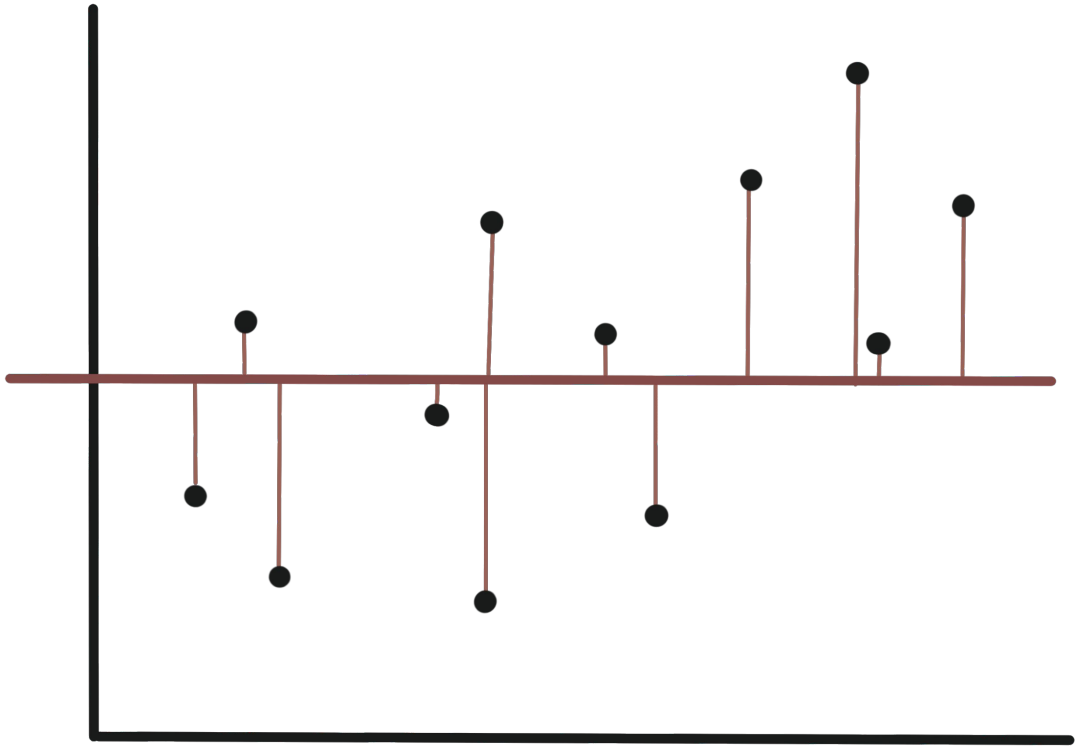
How to Measure Model Success

R^2:



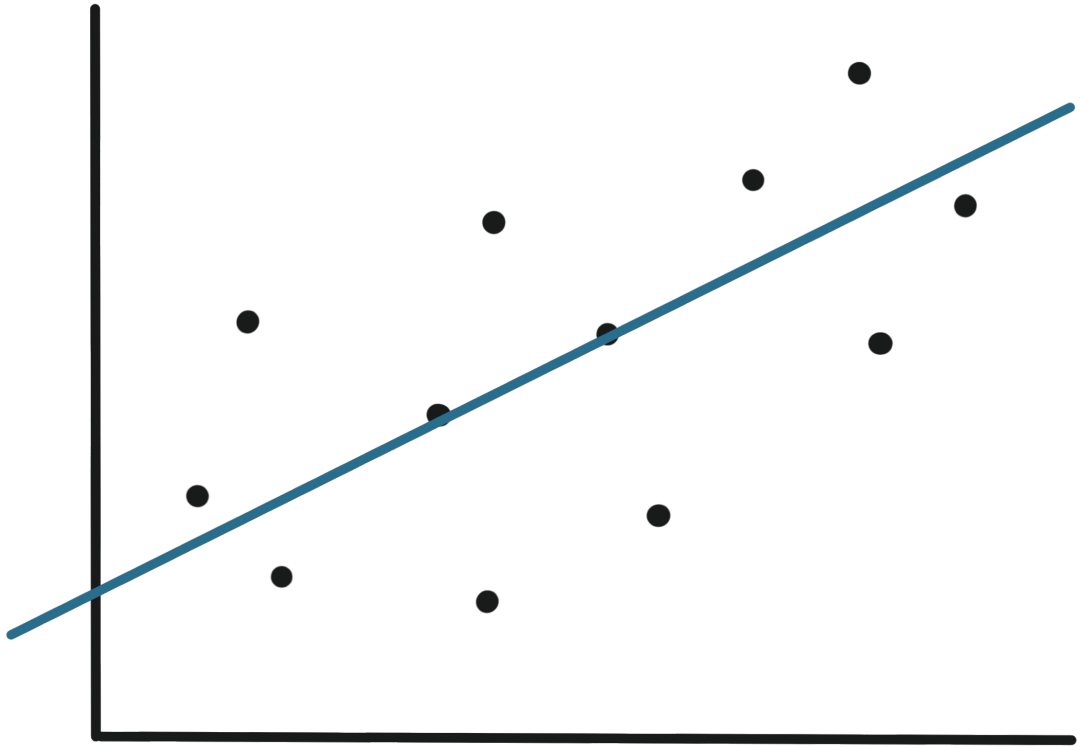
How to Measure Model Success

R^2:



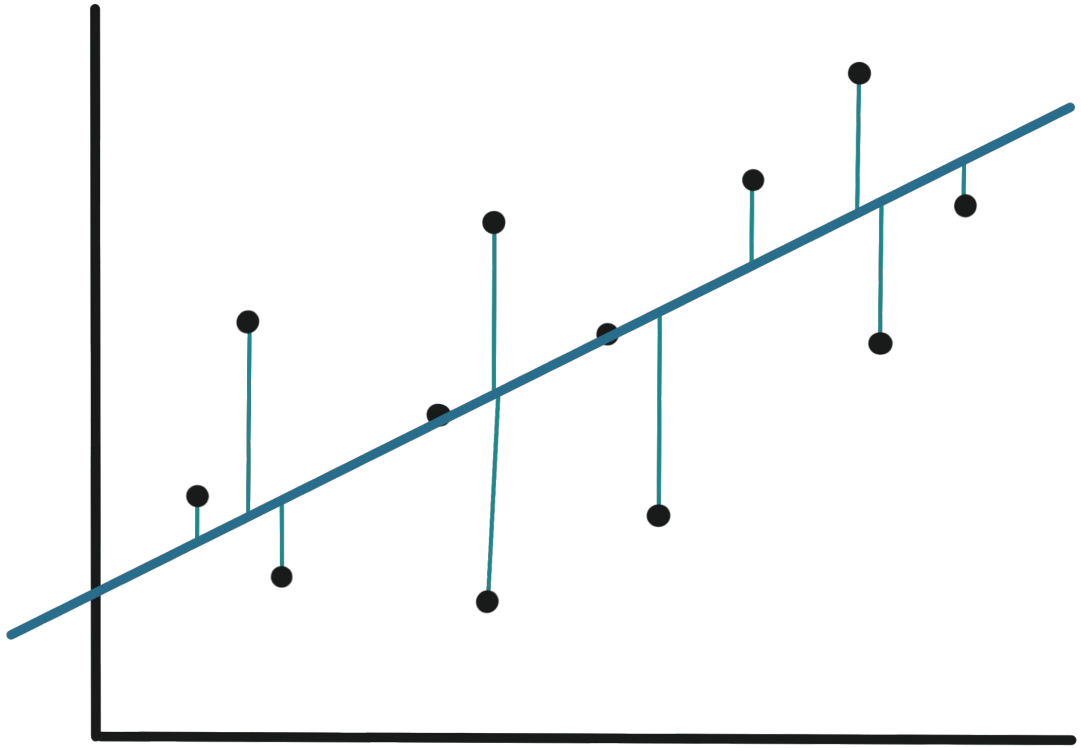
How to Measure Model Success

R^2:



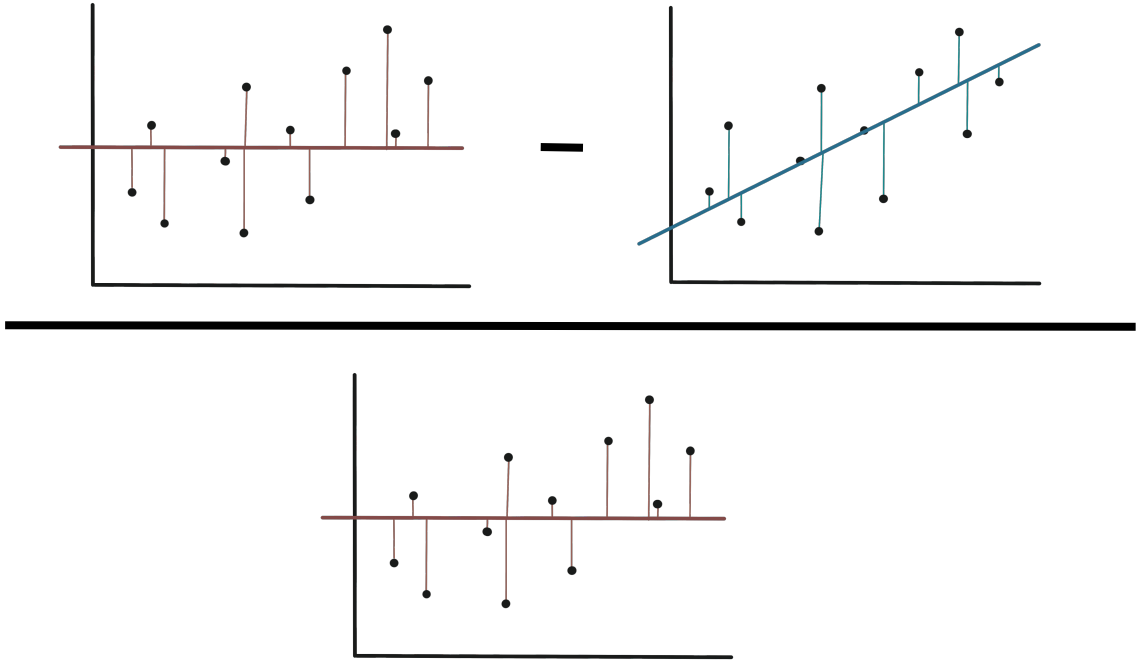
How to Measure Model Success

R^2:

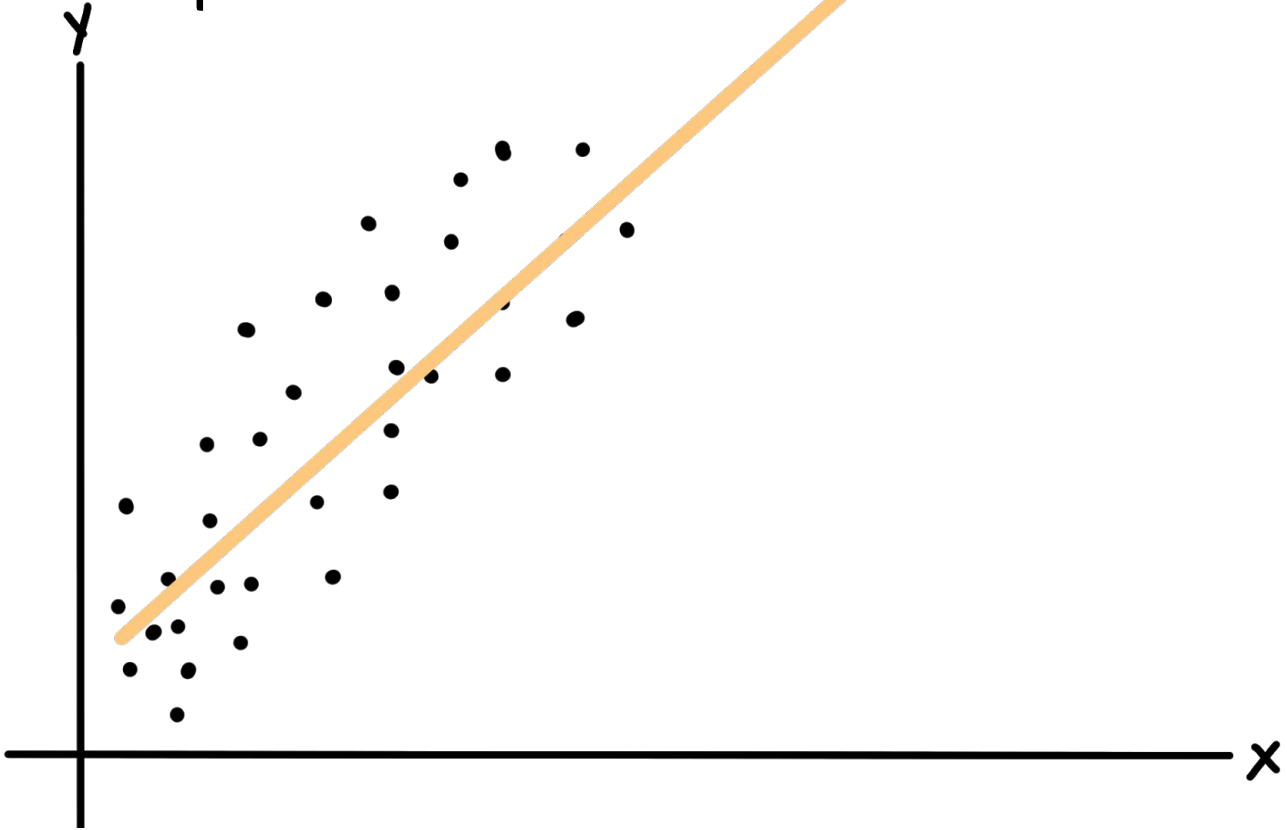


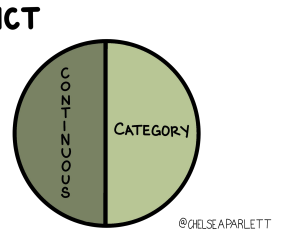
How to Measure Model Success

R^2:

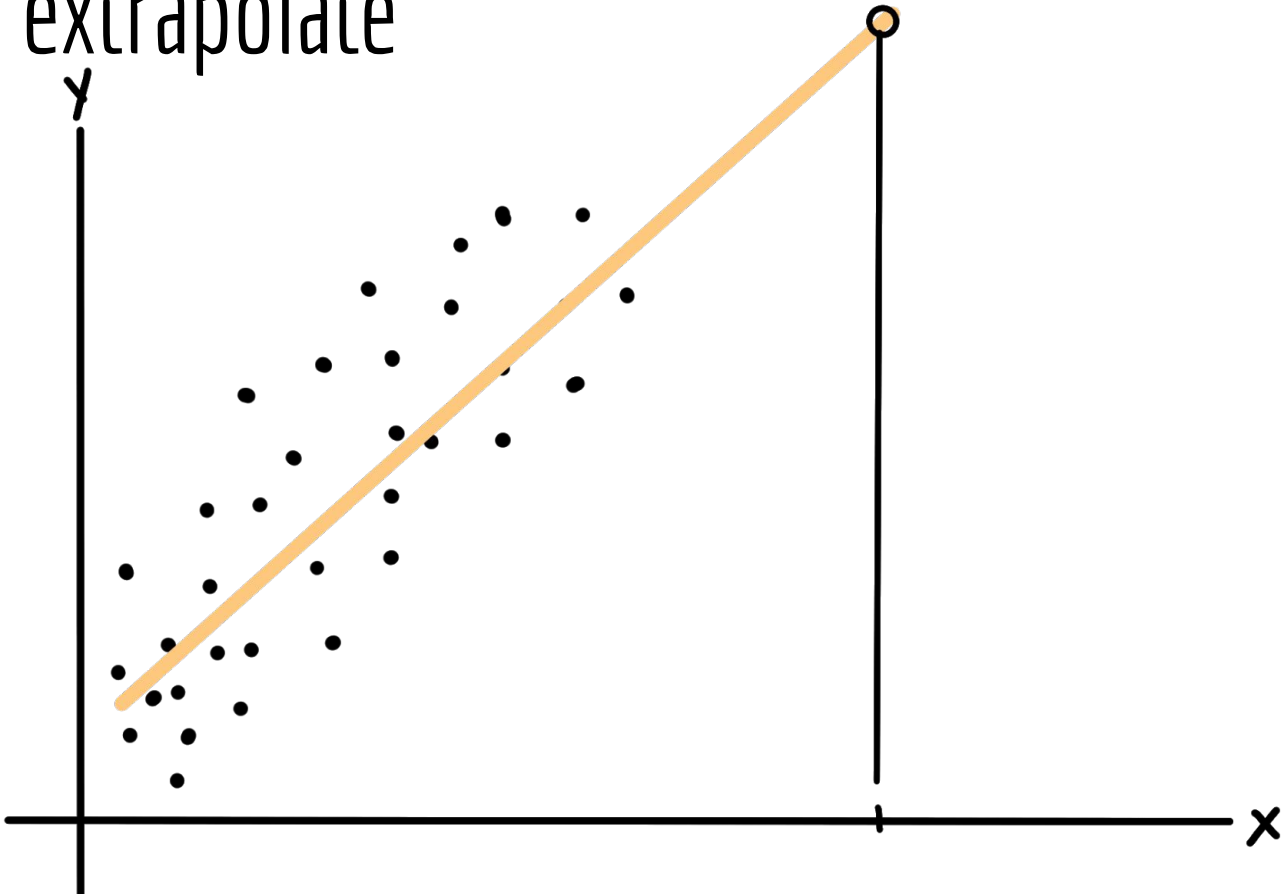


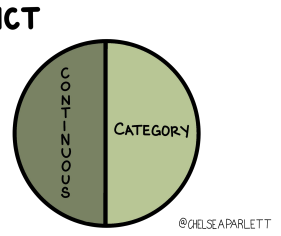
Don't extrapolate



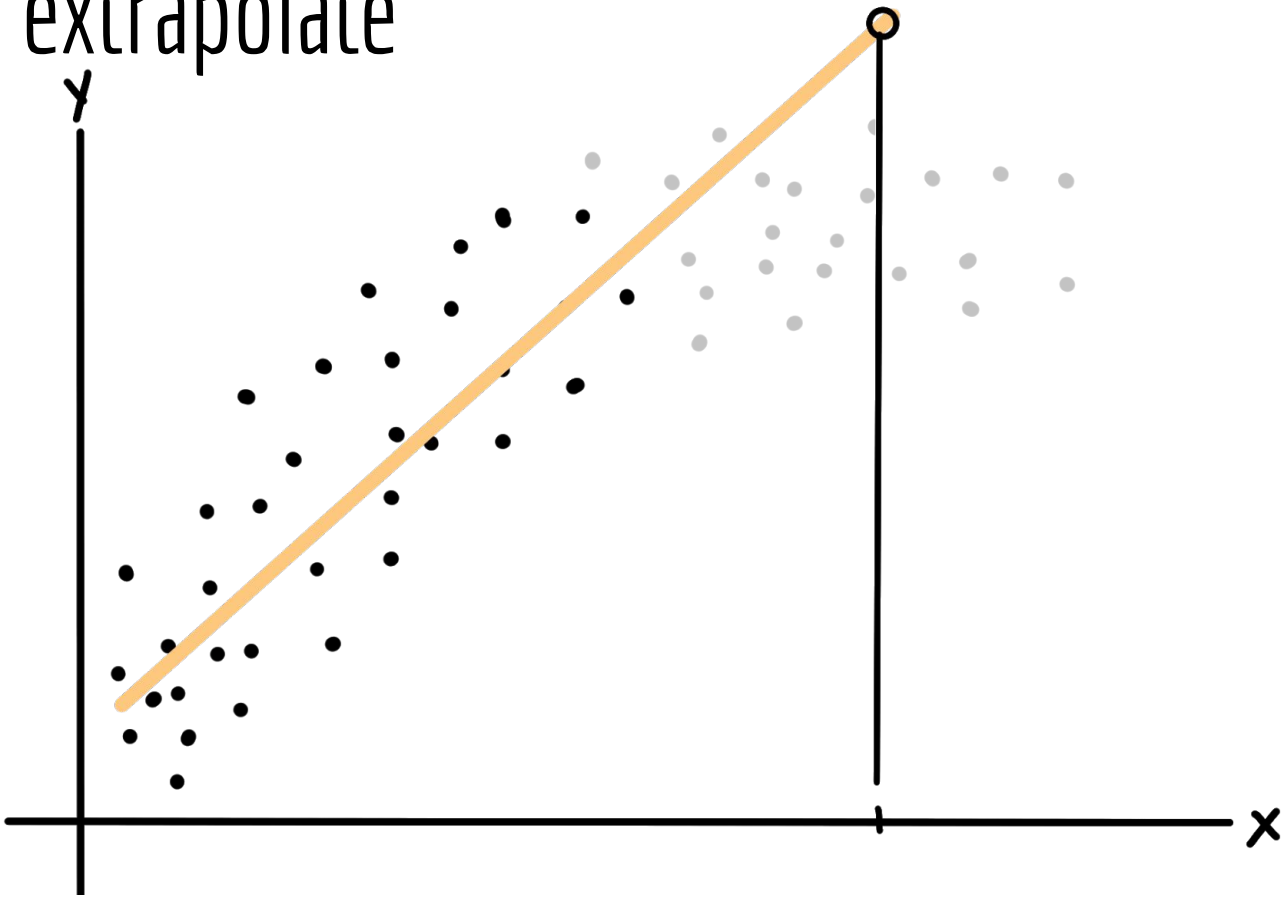


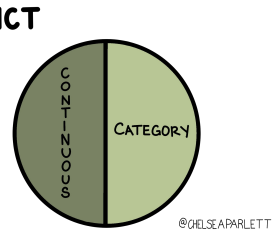
Don't extrapolate



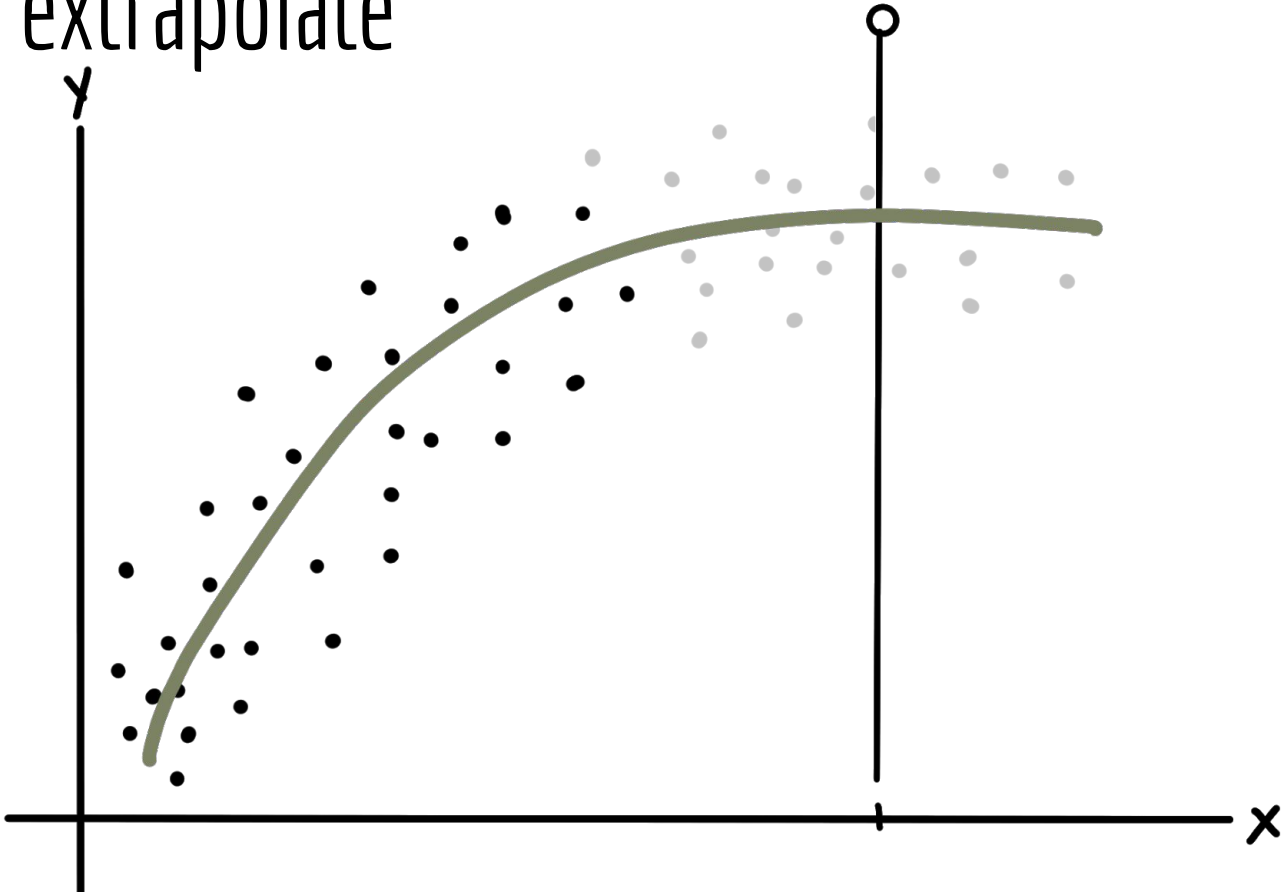


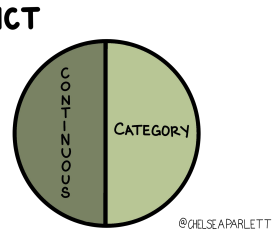
Don't extrapolate



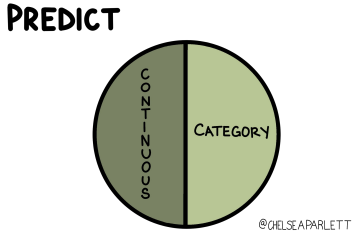


Don't extrapolate





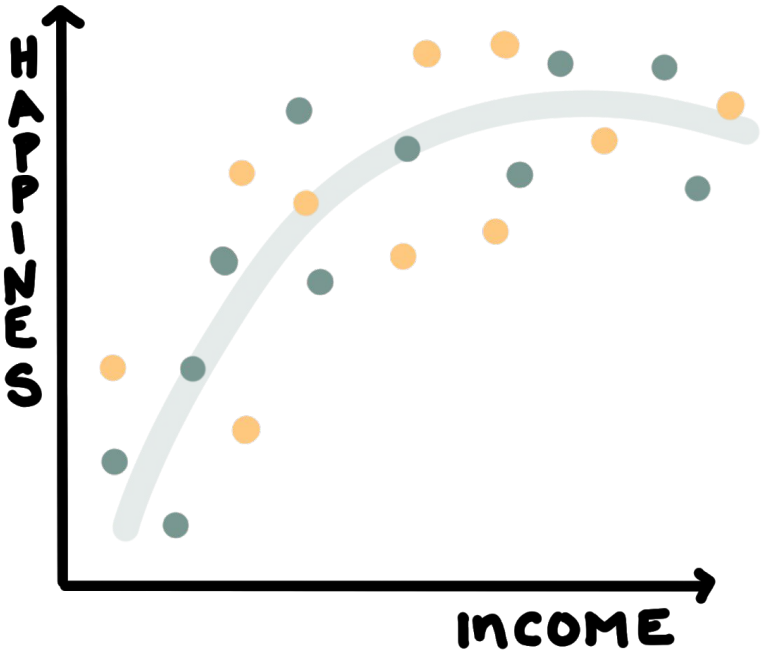
Doing LR for prediction



sklearn:

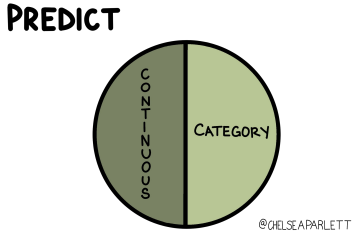
```
x = Model( )

x.fit(X, y)

x.predict(X_new)

x.score(X, y)
```

Bias and Variance



data we have now

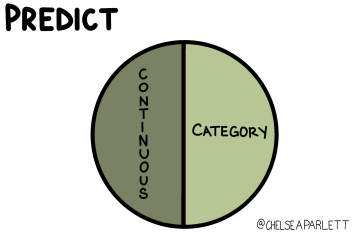
future data



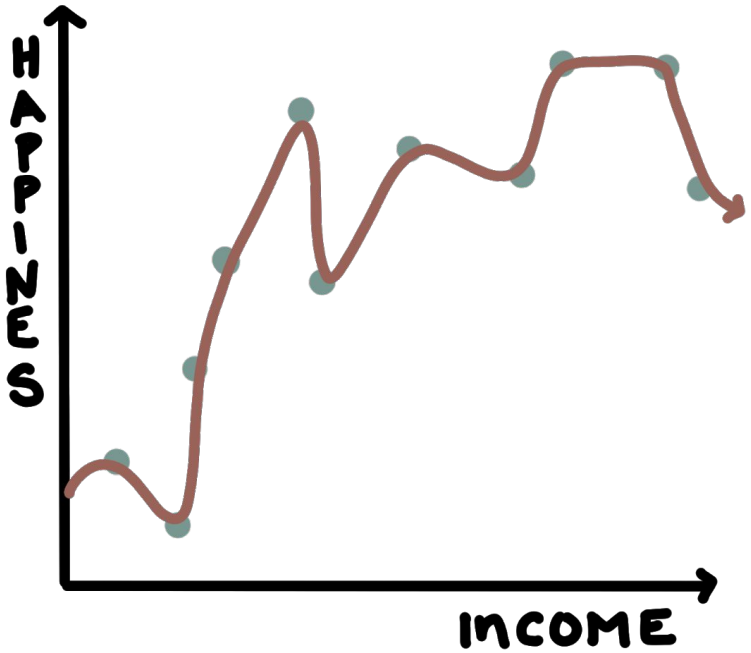
Bias and Variance

data we have now

future data

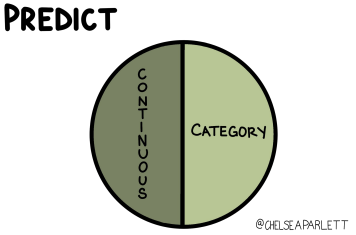


Bias and Variance

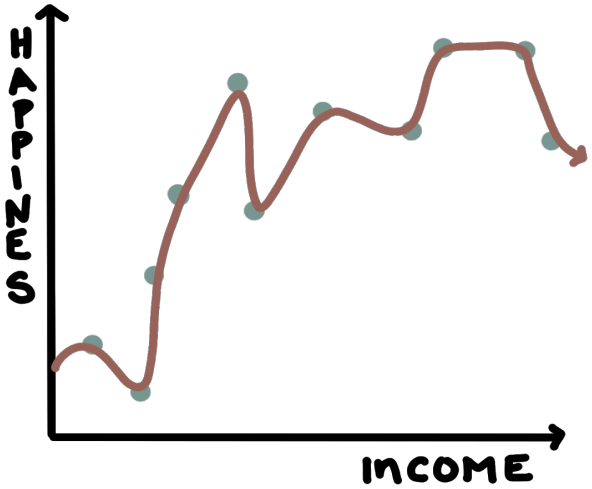


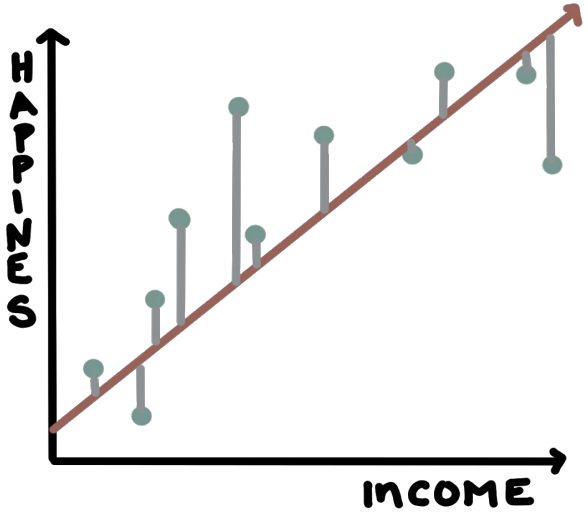
data we have now

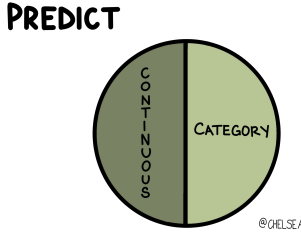
future data



Bias and Variance



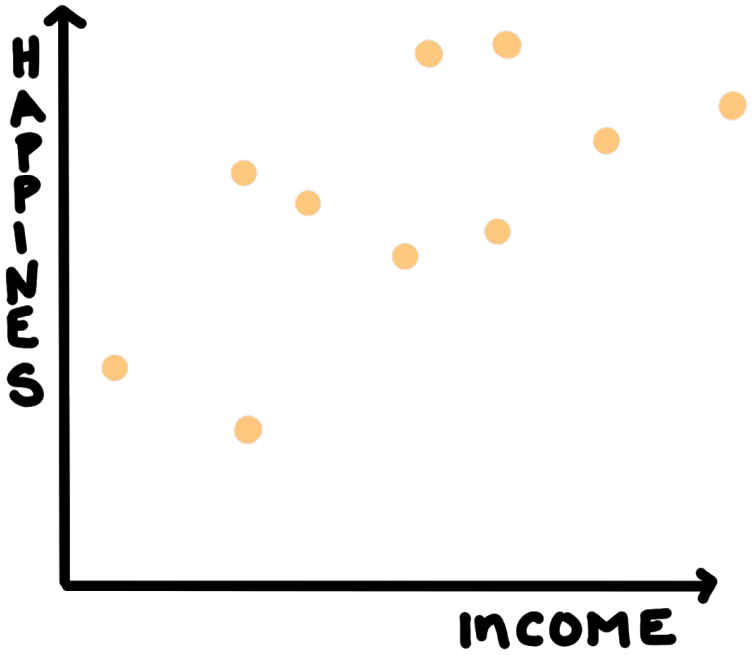




data we have now

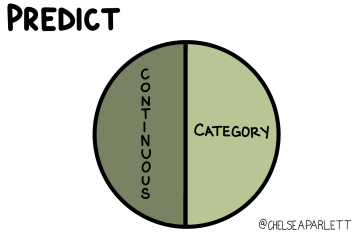
future data

Bias and Variance

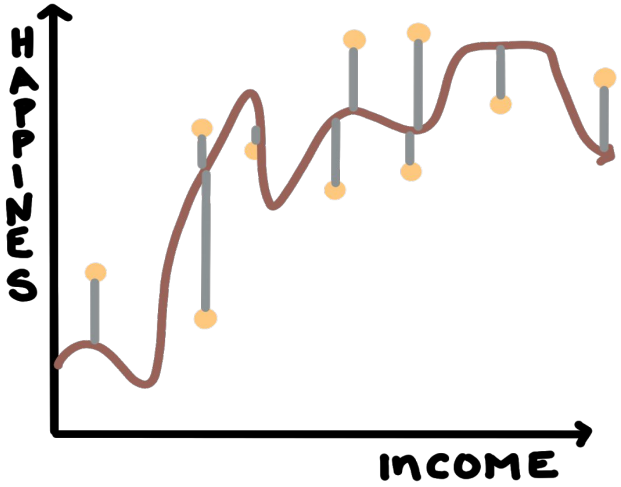


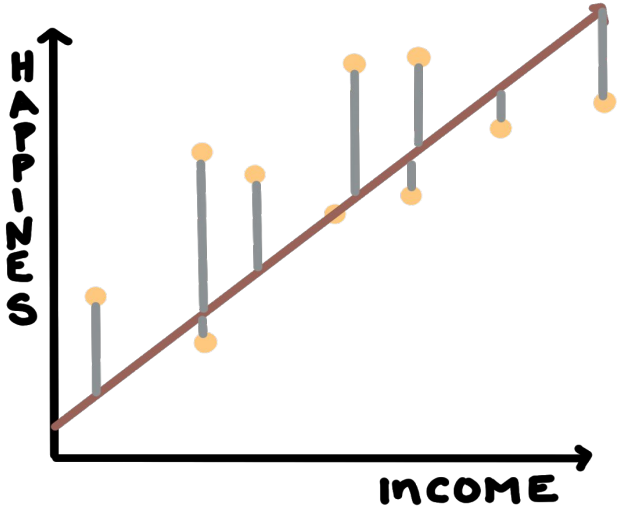
data we have now

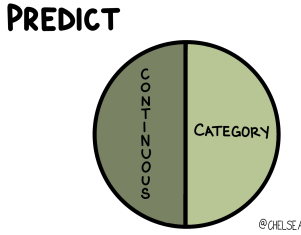
future data



Bias and Variance



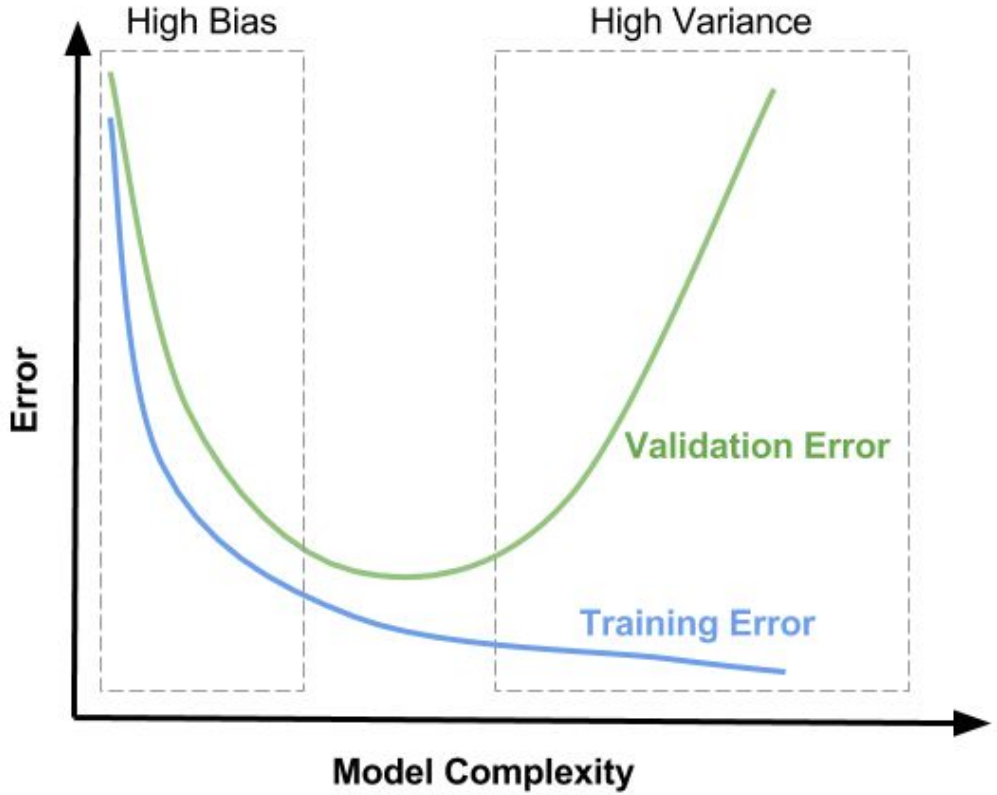




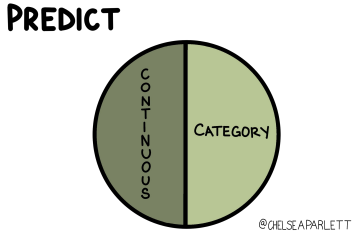
data we have now

future data

Bias and Variance



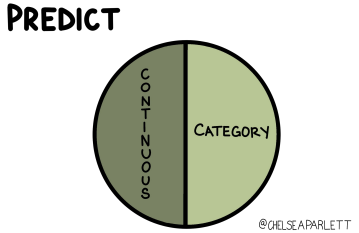
Validation



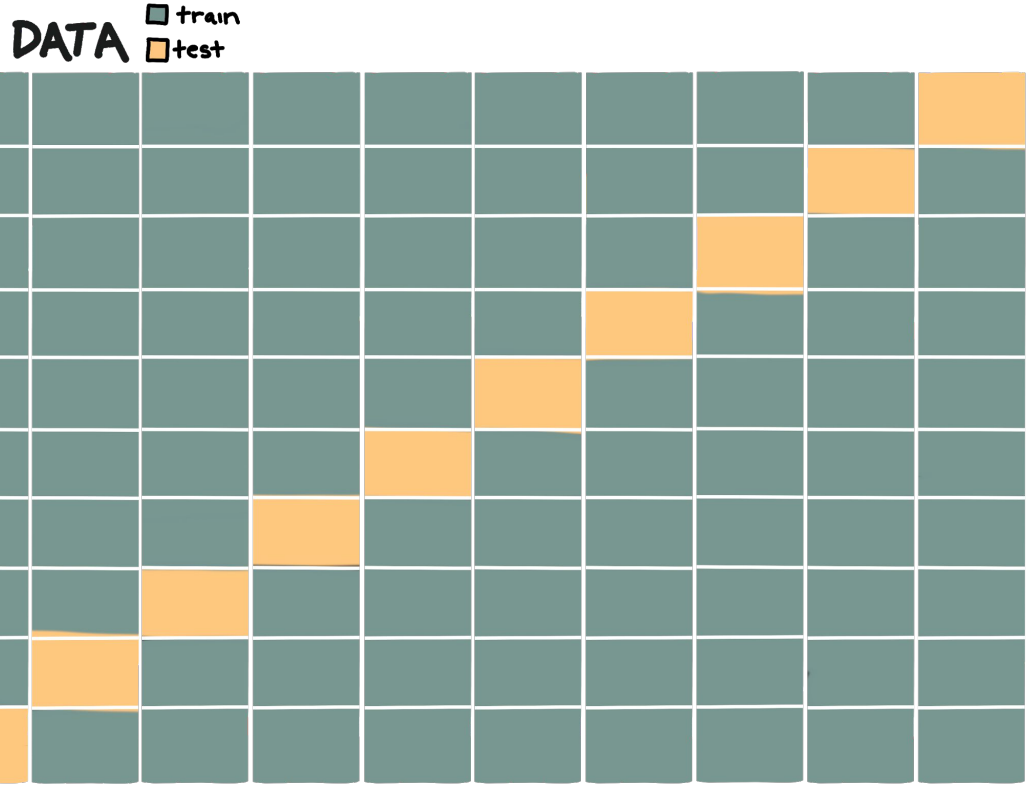
1. **Split data** into train and test, build model on train, test on test
2. Use performance metrics on test data (or average over multiple) to **evaluate model performance**

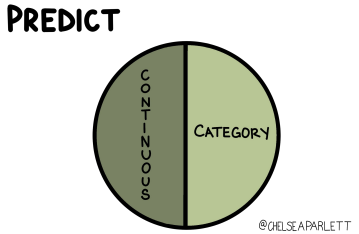
Validation


DATA
train
test

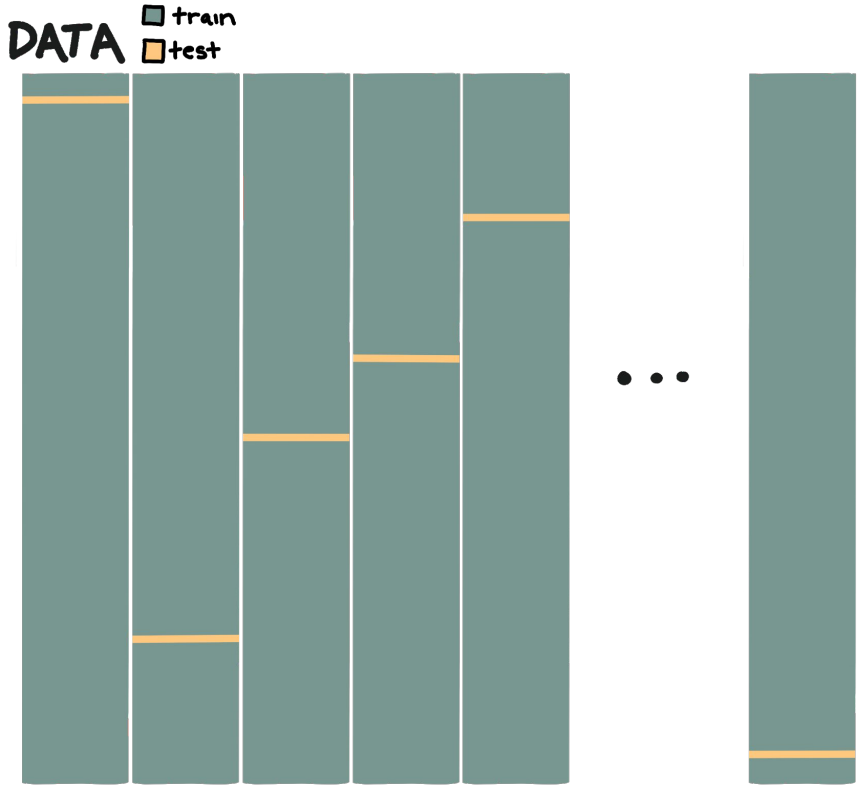

PREDICT
CONTINUOUS
CATEGORY
@CHELSEAPARLETT

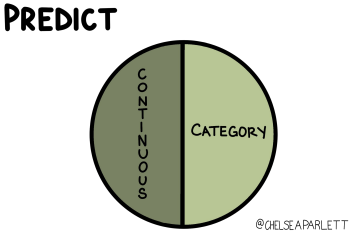
Cross Validation



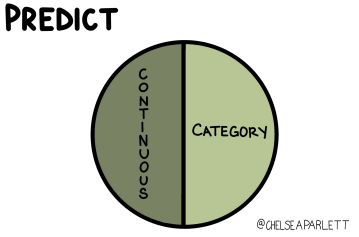


Cross Validation





Cross Validation



How to decide (Things to Think About)

- Size of your Dataset (rows AND columns)
- Computational Expense