R YOU READY?

An Introduction to Data Science in RStudio

Andira Putri
GSU 2018 -- Xylem Inc.

Georgia State University
xylem

Why join GSU Student Chapter of ACM?

- Add to your **resume**: membership in a professional society!
  +You will be listed as a member on our chapter website.
- **Members-only** events: field trips, tailgating, movie outings
- **Very COOL-looking t-shirt**! ☺
- Eligibility to be an **ACM officer**. Serving as an officer is an ideal way to meet other students and faculty, and it looks great on a **resume** as a "**leadership** role" !
- **Email notifications** of: ACM upcoming events, scholarships, summer jobs, internships, job postings, hackathons and other special events of interest to ACM members.

GSU Student Chapter
of the
Association for Computing Machinery

georgiastateuacm

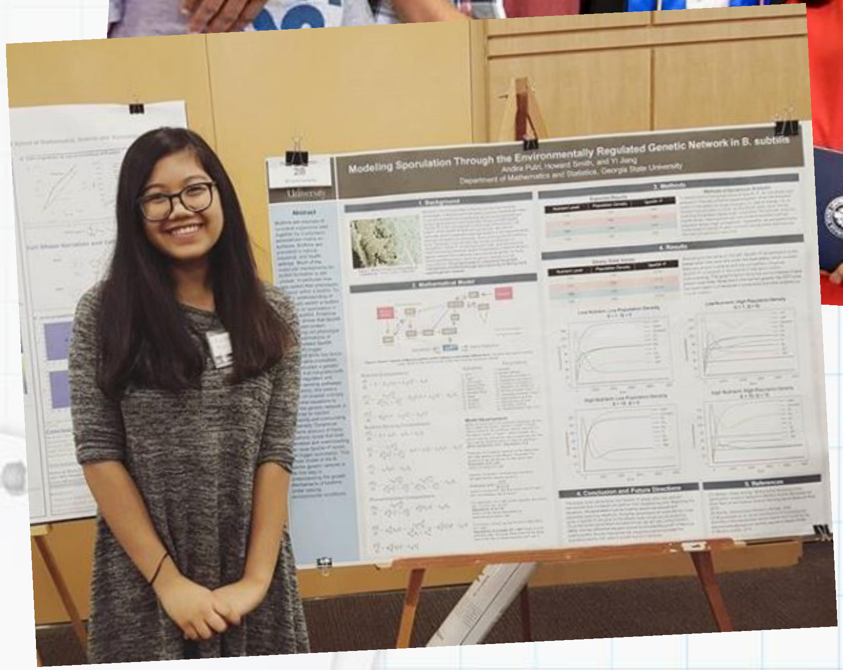
gsuacm

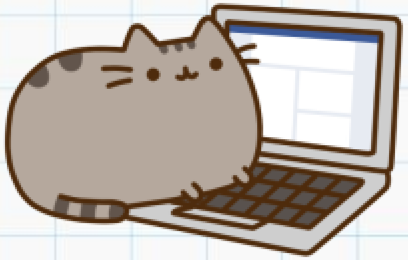
WORKSHOP STRUCTURE

1. Introduction
   - About Me
   - What is Data Science?
2. Linear Regression
3. Logistic Regression
4. Resampling Methods
5. Case Study
6. Questions?

INTRODUCTION

About Me

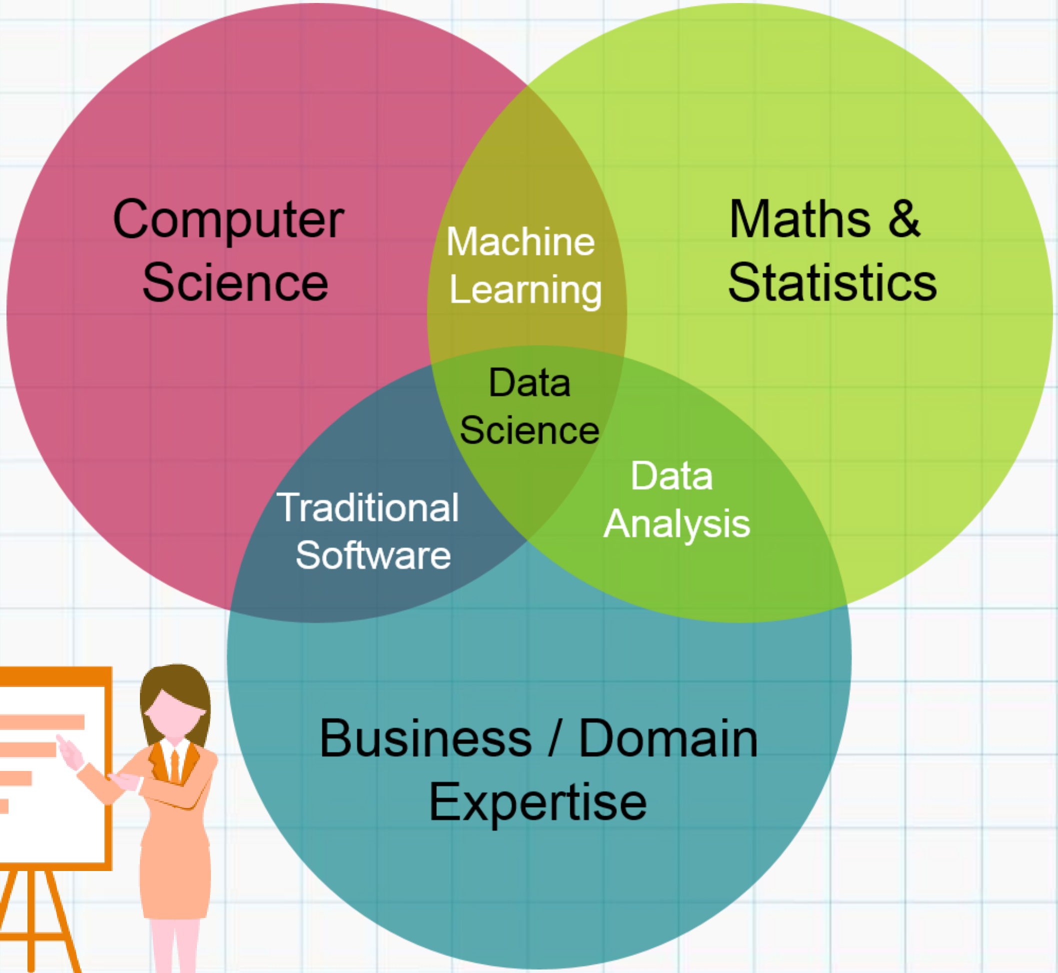


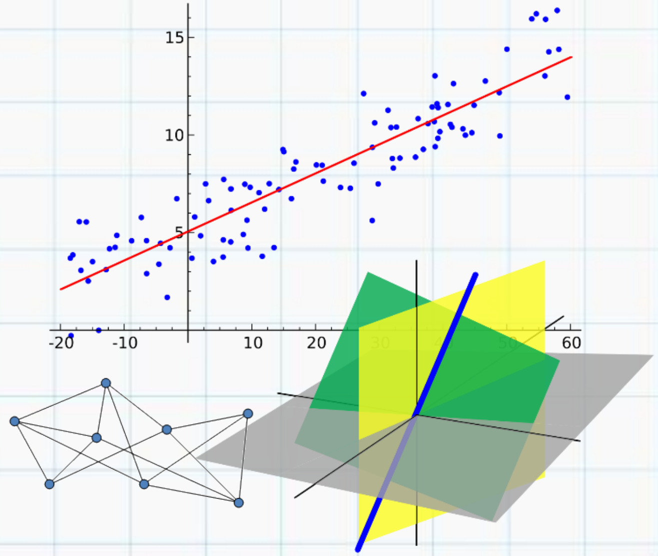


- Graduated from Georgia State in Spring 2018 – B.S. in Mathematics
- Data Scientist at Xylem Inc. in Research Triangle Park, NC
- Will be pursuing an M.S. in Statistics soon at North Carolina State University

INTRODUCTION

What is Data Science?



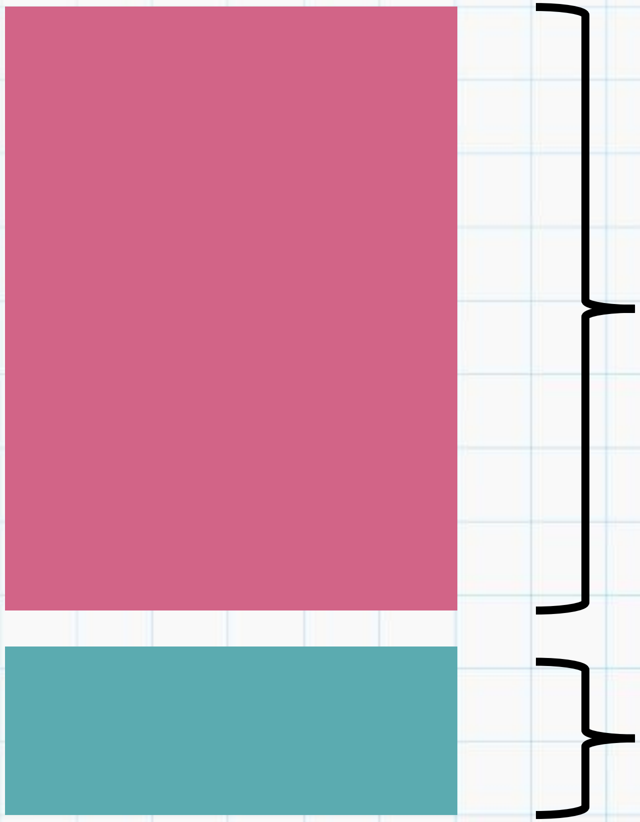


INTRODUCTION

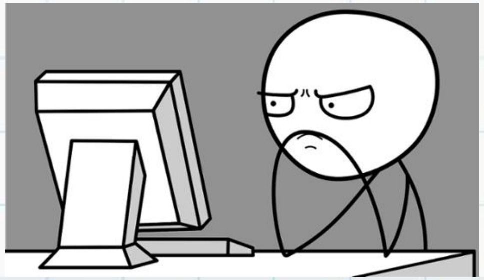
The Perfect Data Scientist

INTRODUCTION

A Typical Project Breakdown



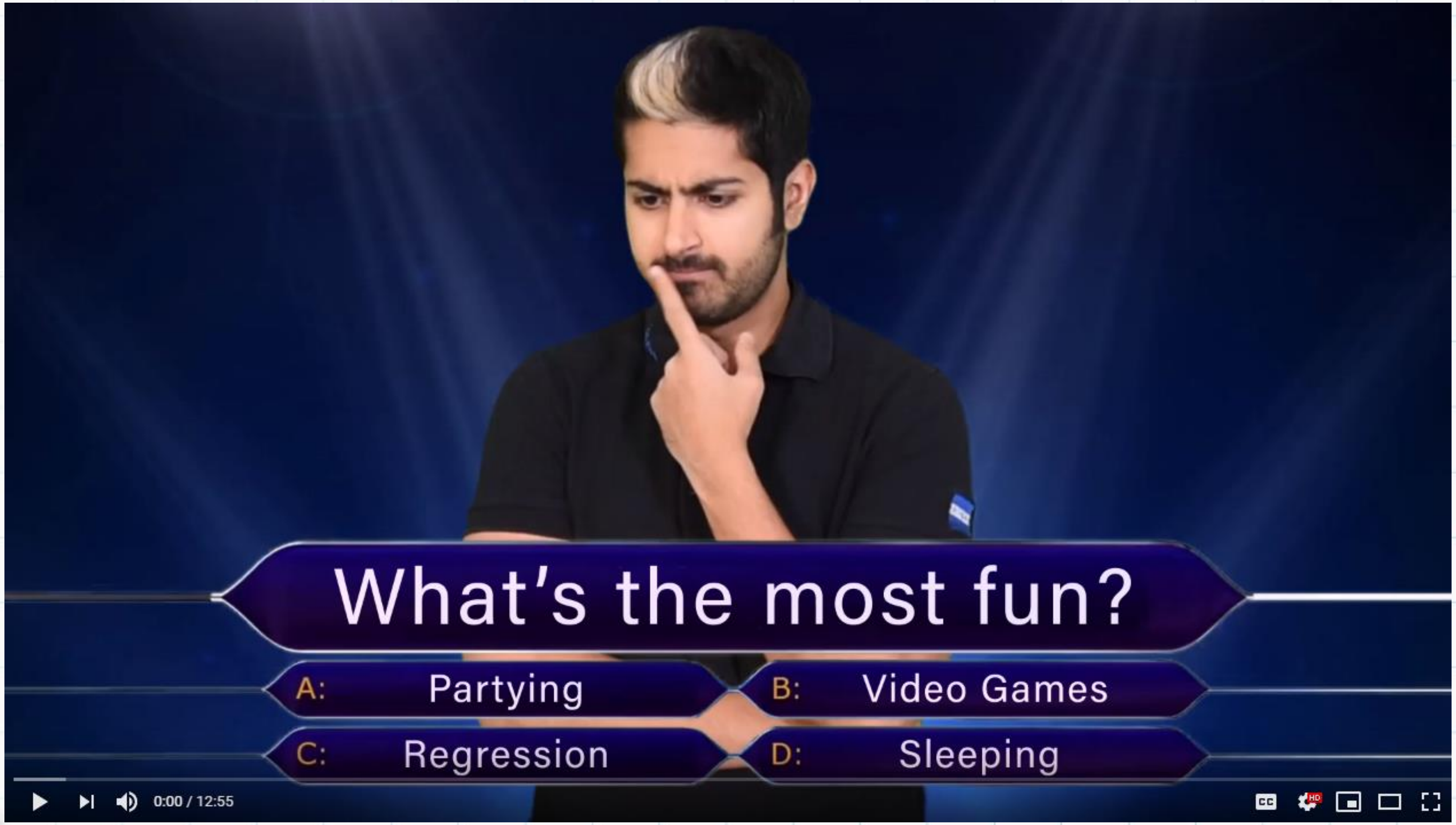
85% Data Mining



15% Applications, Productizing, Writing

REGRESSION

You Must Know It!



Siraj Raval

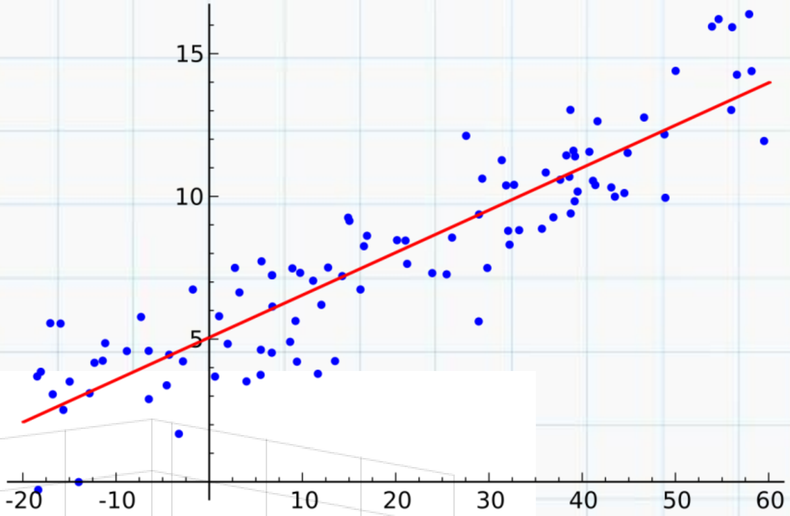
LINEAR REGRESSION

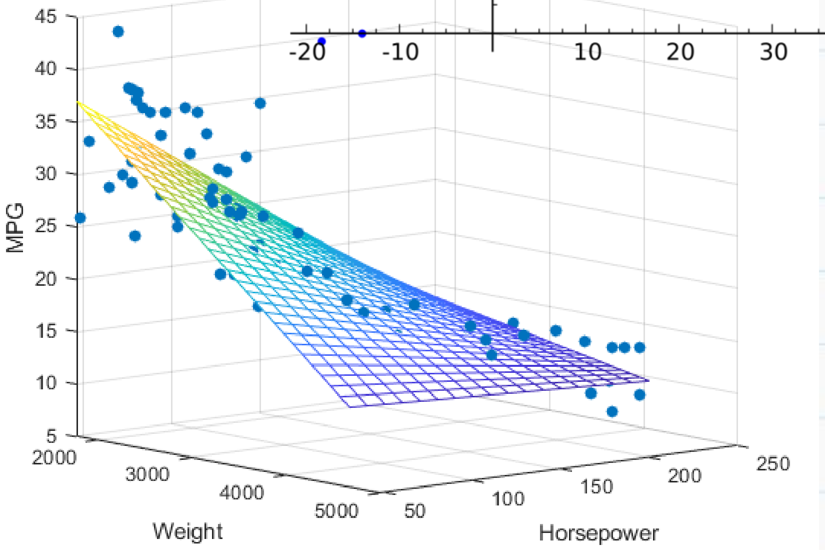
A Foundation of Data Science

- Assumes a linear relationship between predictors $X_{1 \ldots n}$ and a response Y

$$\underbrace{\underbrace{Y = \beta_0 + \beta_1 X_1 + \varepsilon}_{\text{Simple}} + \beta_2 X_2 + \ldots + \beta_n X_n}_{\text{Multiple}}$$

- β_0, $\beta_1 \ldots \beta_n$ are known as regression coefficients
  - $\beta_0 \rightarrow$ expected mean when predictors = 0
  - $\beta_1 \ldots \beta_n \rightarrow$ association between predictor and response
  - $\varepsilon \rightarrow$ error (the residual)





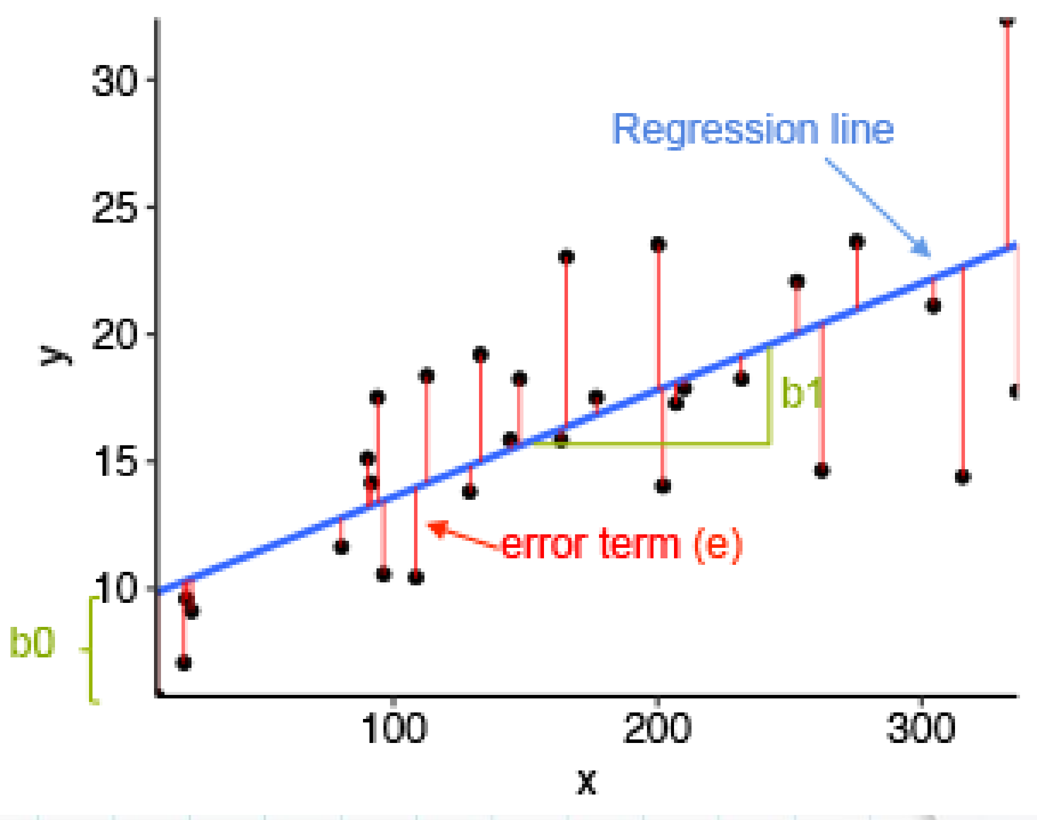
LINEAR REGRESSION

A Foundation of Data Science

- Coefficients are estimated by least squares method

- How accurate is our model? We can assess our model using:
  - R-Squared and Adjusted R-Squared
  - Residual Standard Error (RSE)
  - F-statistics

- All values can be called using the summary() function in R

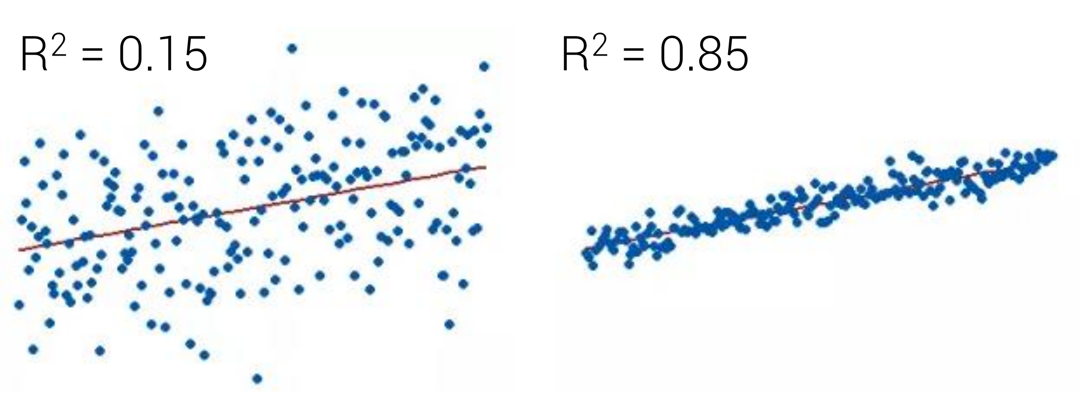


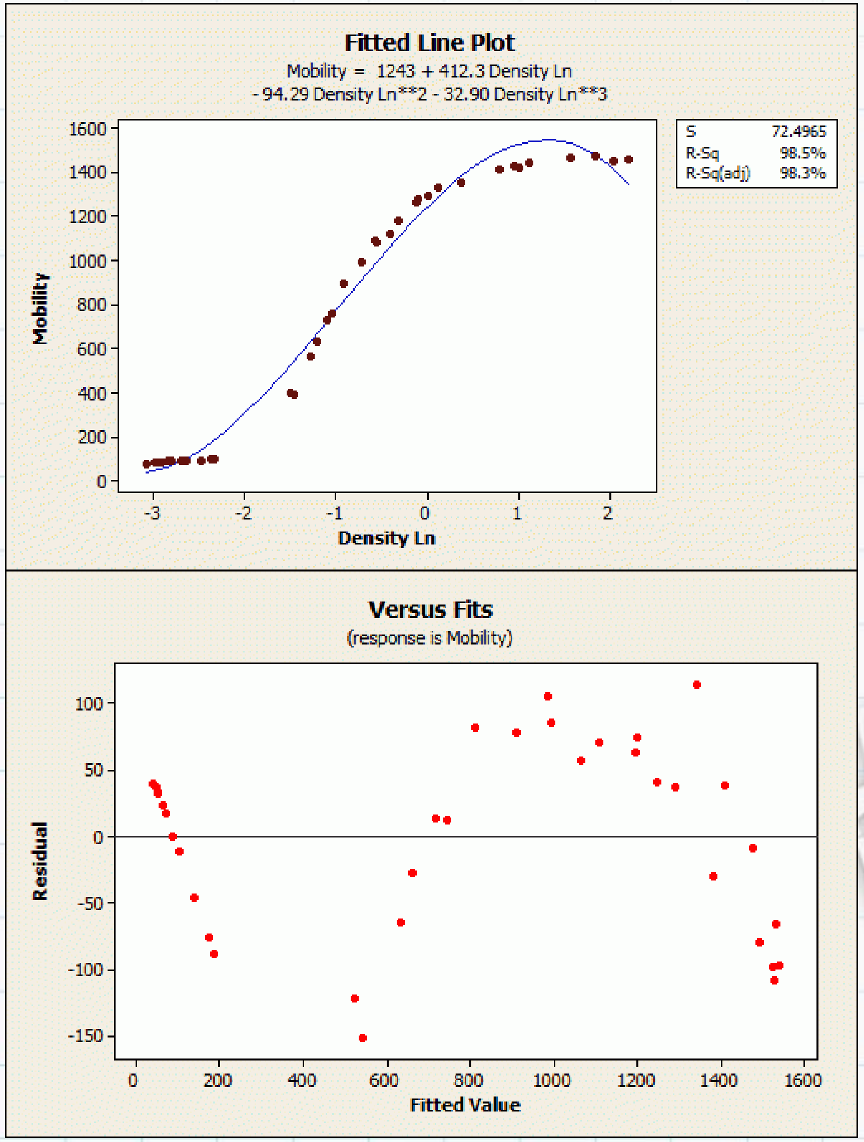
LINEAR REGRESSION

A Foundation of Data Science

R^2 and Adjusted-R^2 Values

- Represents the proportion of variation in the response that can be explained by the predictors
- $R^2 \in [0,1]$; the closer to 1, the better
- R^2 IS NOT ALWAYS RELIABLE
  It increases when you add more predictors – predictors could be weakly related to the response
- Adjusted- R^2 corrects for the additional predictors





$R^2 = 0.99$

Bias!

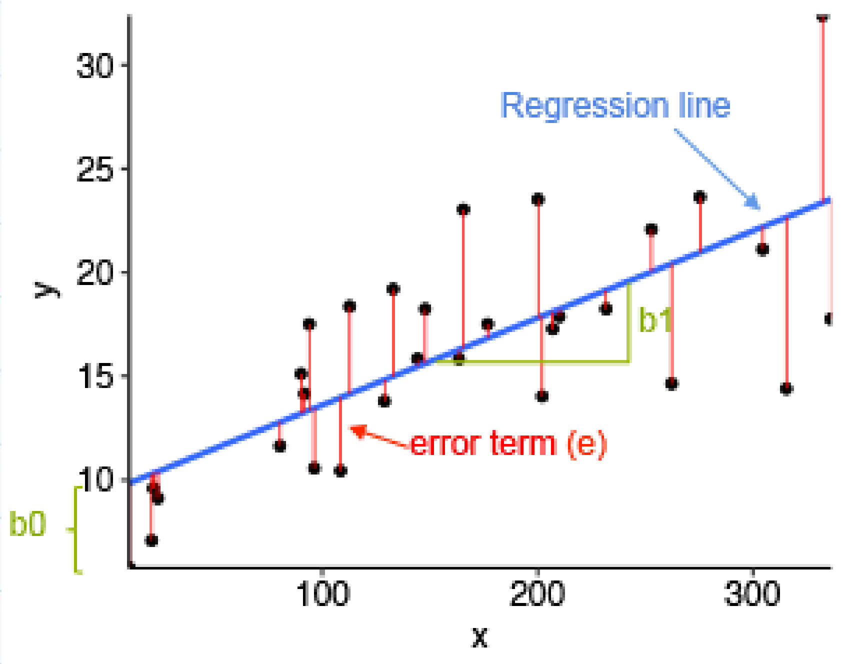
LINEAR REGRESSION
A Foundation of Data Science

Residual Standard Error (RSE)

- Represents roughly the average difference between the true response values and the predicted values by the model

- The smaller, the better!

- Dividing the RSE by the average value of the response will give you the prediction error rate

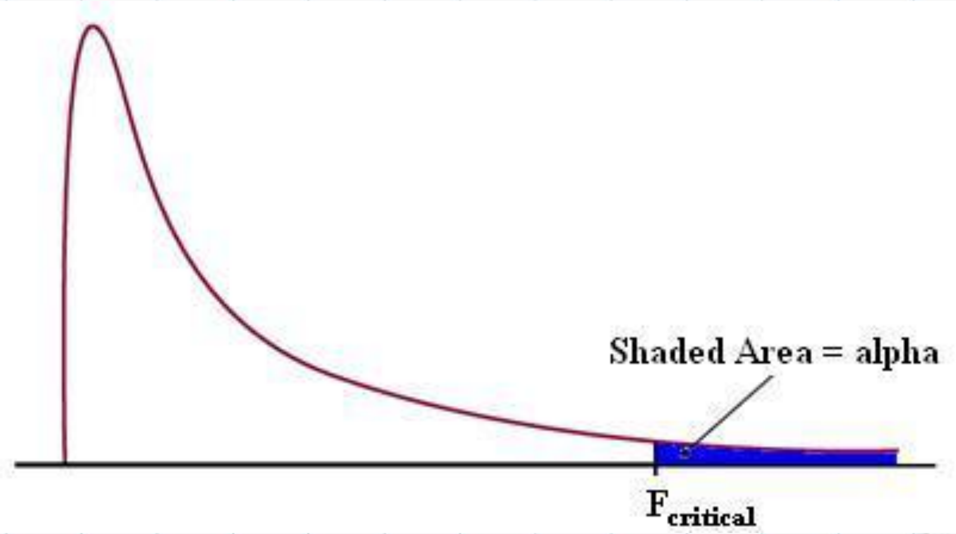


LINEAR REGRESSION

A Foundation of Data Science

F-statistic

- Gives the overall significance of the model - assesses whether at least one predictor variable has a non-zero coefficient
- Useful for multiple regression
- We want a large F-statistic, which gives a statistically significant p-value for a confidence level α (e.g. at α = 95%, we want p < 0.05)



CASE STUDY

In the Market

Let's predict the prices of homes in Boston.

RESAMPLING METHODS

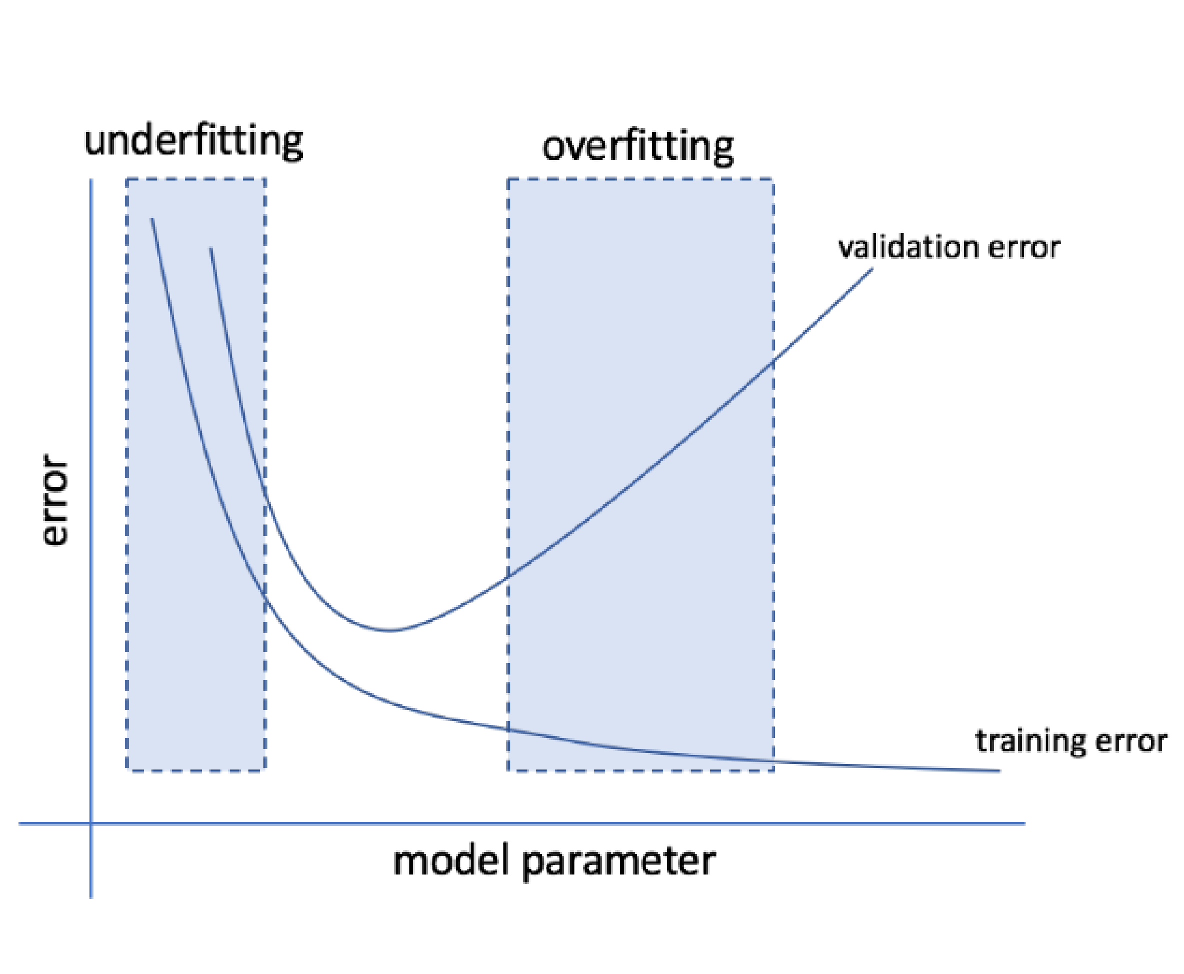
- Involves repeatedly drawing samples from a training set and refitting a model of interest on each sample in order to obtain additional information about the fitted model

- Can our model generalize to new data?

- Two ways: cross-validation and bootstrap

Cross-validation is used to assess model performance

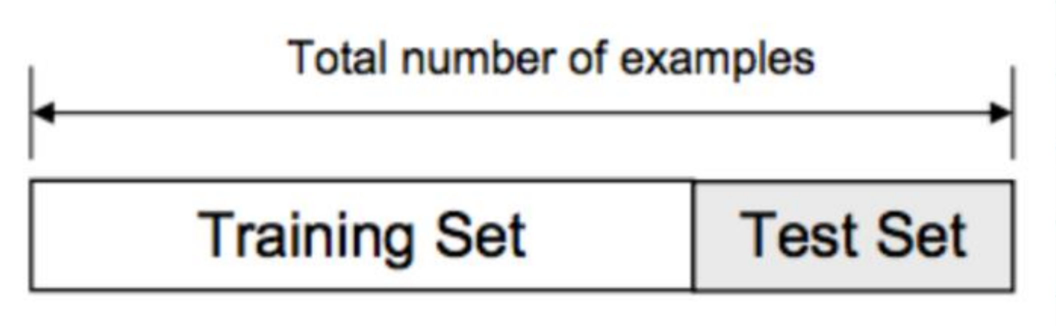
Bootstrap can estimate confidence intervals, accuracy of parameter estimate, etc.



TRAIN / TEST SPLIT

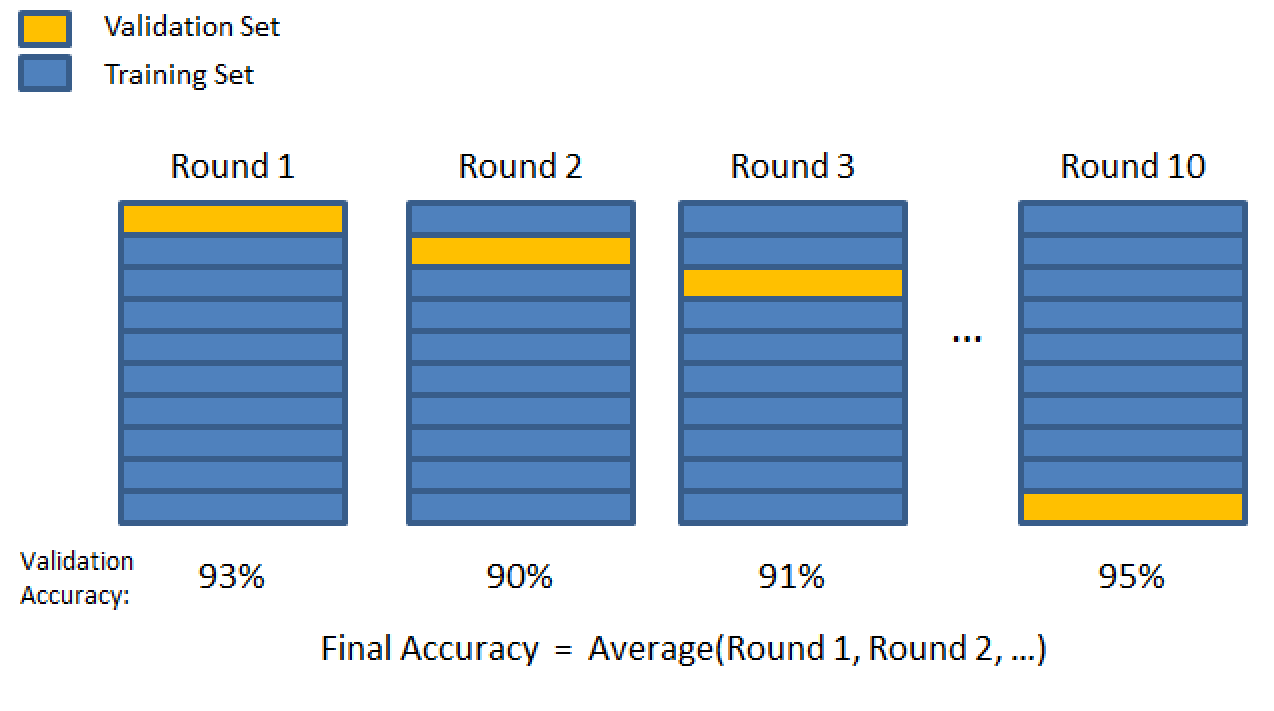
- The simplest resampling method
- You just need to make sure the sets are not overlapping
- NOT THE BEST
  Subsets may not be completely randomized
  Training set could have people with same age, people who live in the same state, etc.
  OVERFITTING!

  Also, it causes a reduction in training data size
  UNDERFITTING!



K-FOLD CROSS VALIDATION

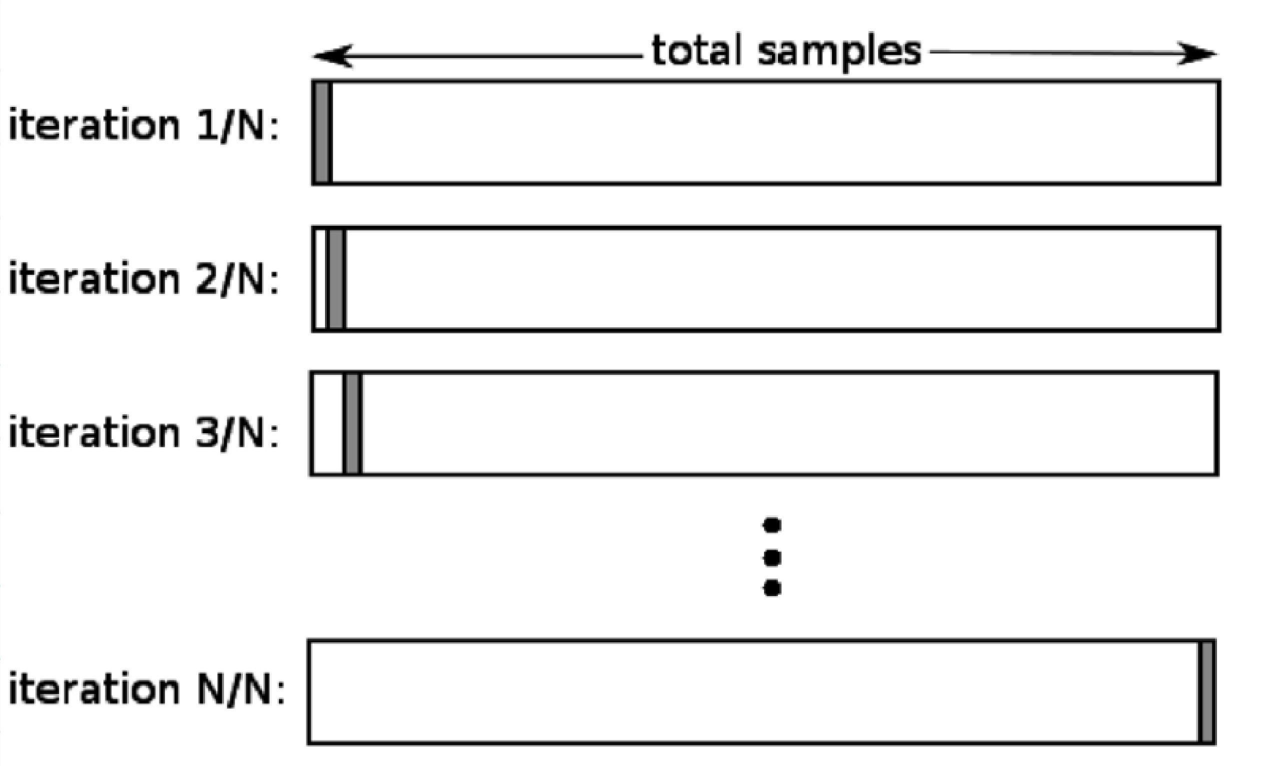
- Train/test split → 2 groups
  k-fold cross validation → k # of groups
- Essentially an iterative / repeated train/test split
- Average the resulting accuracies after repeating k times
- Reduces OVERFITTING
  Most data is being used in the test set
- Reduces UNDERFITTING
  Most data is being used in the training set



General guideline is k=5 or k=10

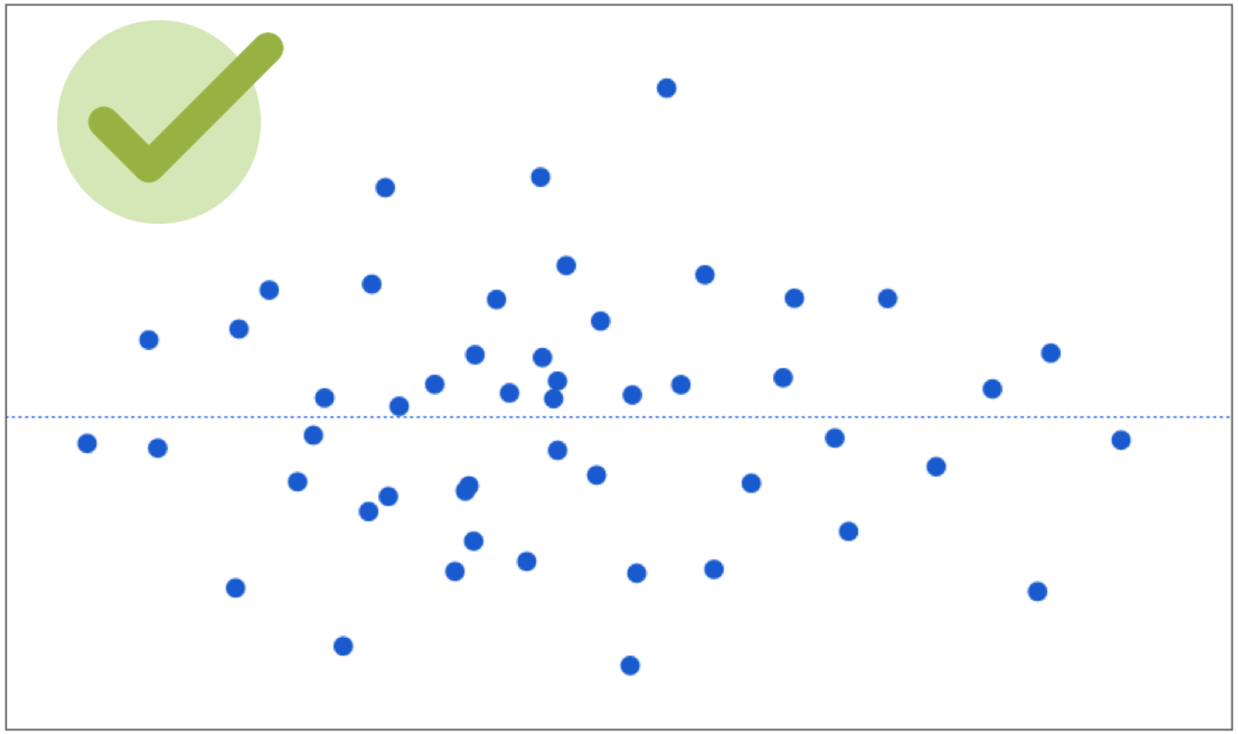
LEAVE ONE OUT

- A special case of k-fold cross validation
- We have n # of observations
  Repeat train/test split n times
  Only one element in the test set
- Great if you have small data sets
- Computationally eXpEnSiVe



ASSUMPTIONS

1. The true relationship is linear
2. Errors are normally distributed
3. Homoscedasticity of errors
4. Observations are independent
5. Variables are not correlated with one another



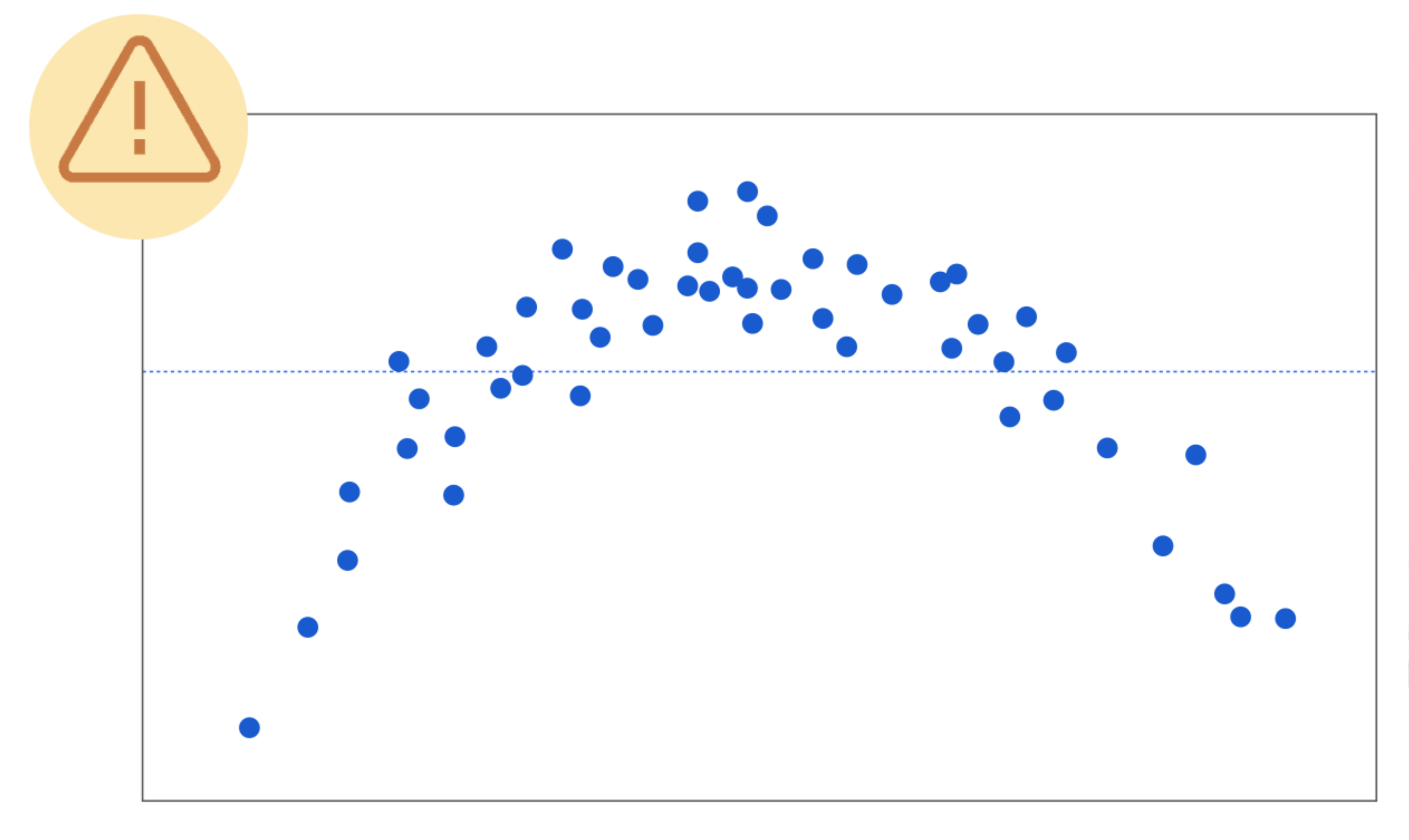
ASSUMPTIONS

VIOLATION in this residual plot!

There is an obvious curvature in the plot → non-linear relationship

A nonlinear regression would be more informative than a linear one

Under-predicts and over-predicts values → model bias



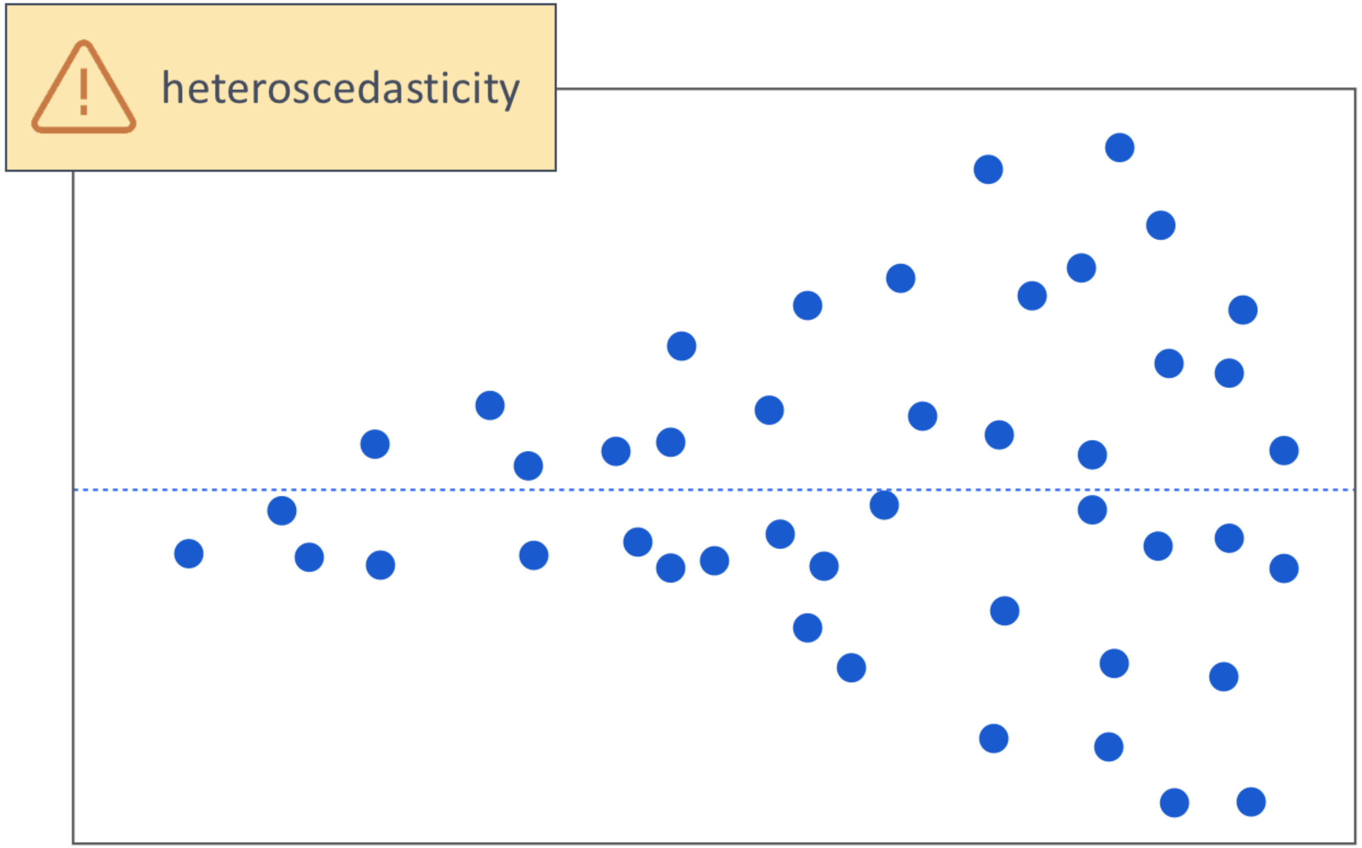
ASSUMPTIONS

VIOLATION in this residual plot!

We have heteroscedasticity

Variance is unequal across the range of values

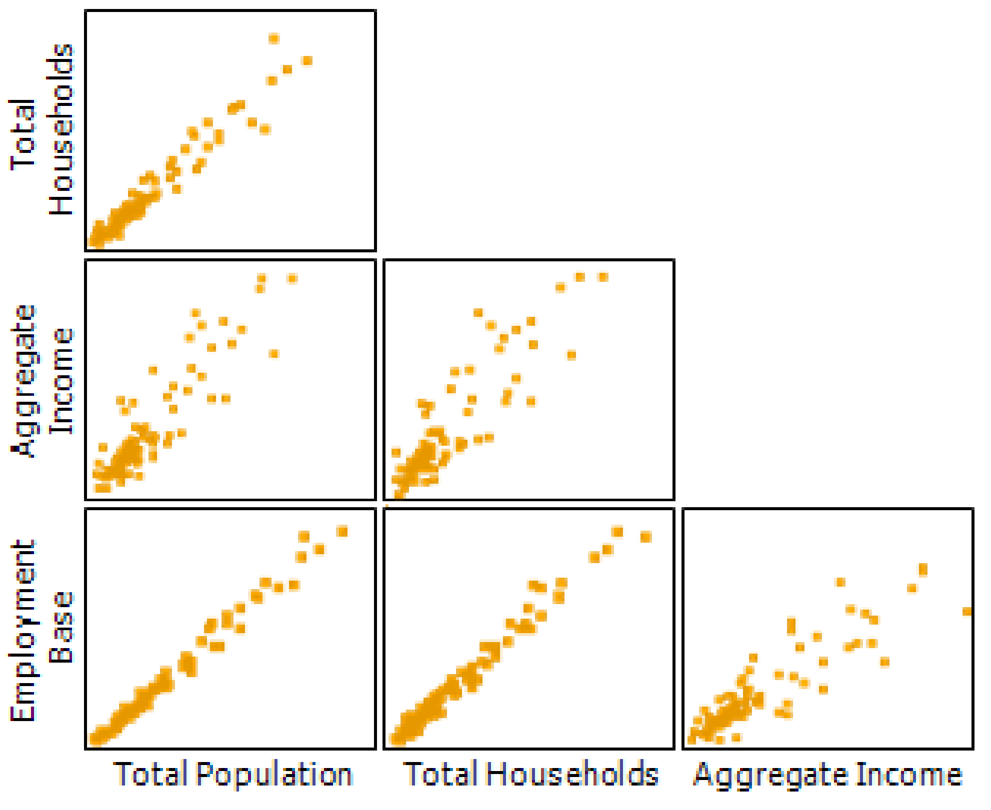
Predicting low values = high accuracy
High values = low accuracy
Model should not be trusted!



ASSUMPTIONS

Collinearity (violation)

- Occurs when predictors are highly correlated with each other
- Classic symptom:
  Small change in data can cause big change in coefficients
  OVERFITTING
- Tell-tale signs:
  High R^2 but low t-scores
  Variation Inflation Factor (VIF) > 5
- You may be able to ignore it if you are just worried about prediction accuracy



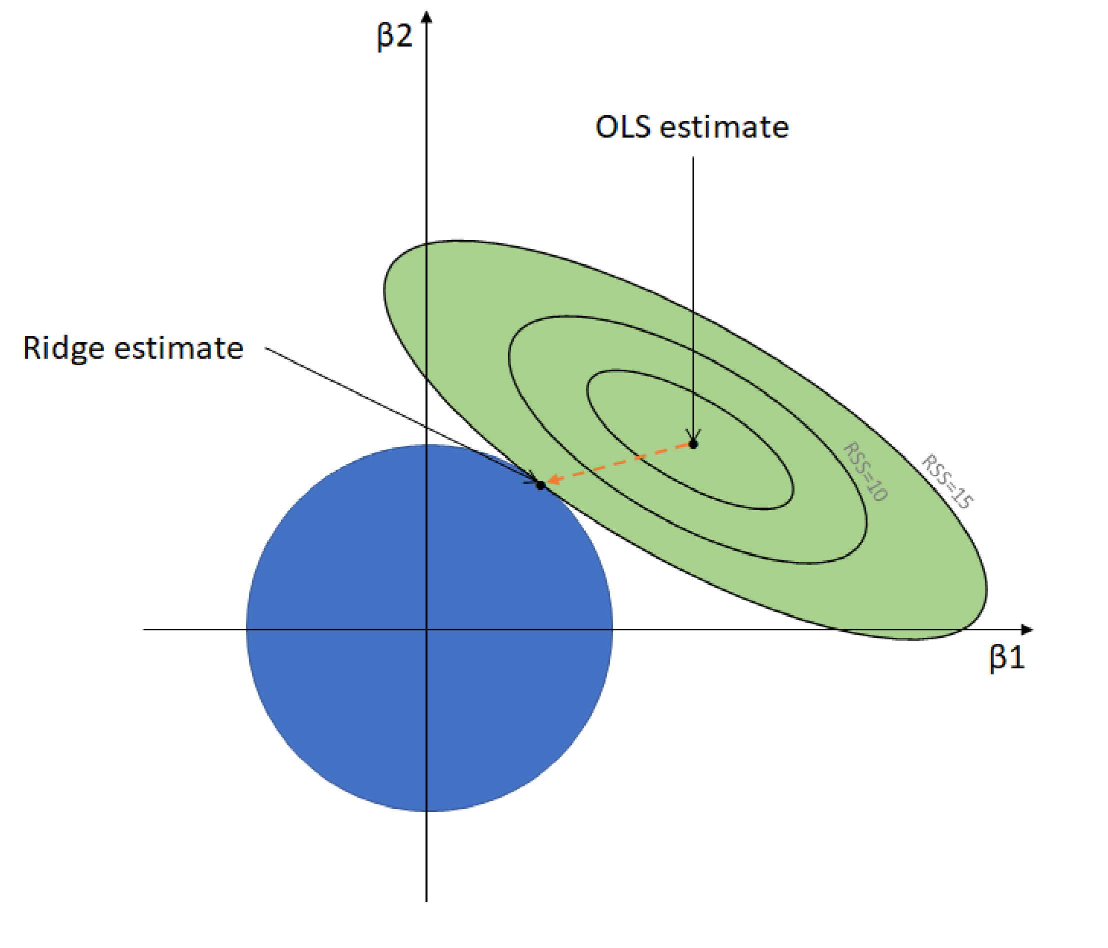
RIDGE REGRESSION

- To address collinearity, ridge regression adds a bias term to our standard equation
  Reduces variance
  - Least squares will want to minimize residual sum of squares (RSS)

$$\mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

The summation is called the shrinkage penalty

  - λ is the tuning parameter
    Higher λ → more shrinkage, and coefficients shrink towards 0
  - λ is determined through cross-validation

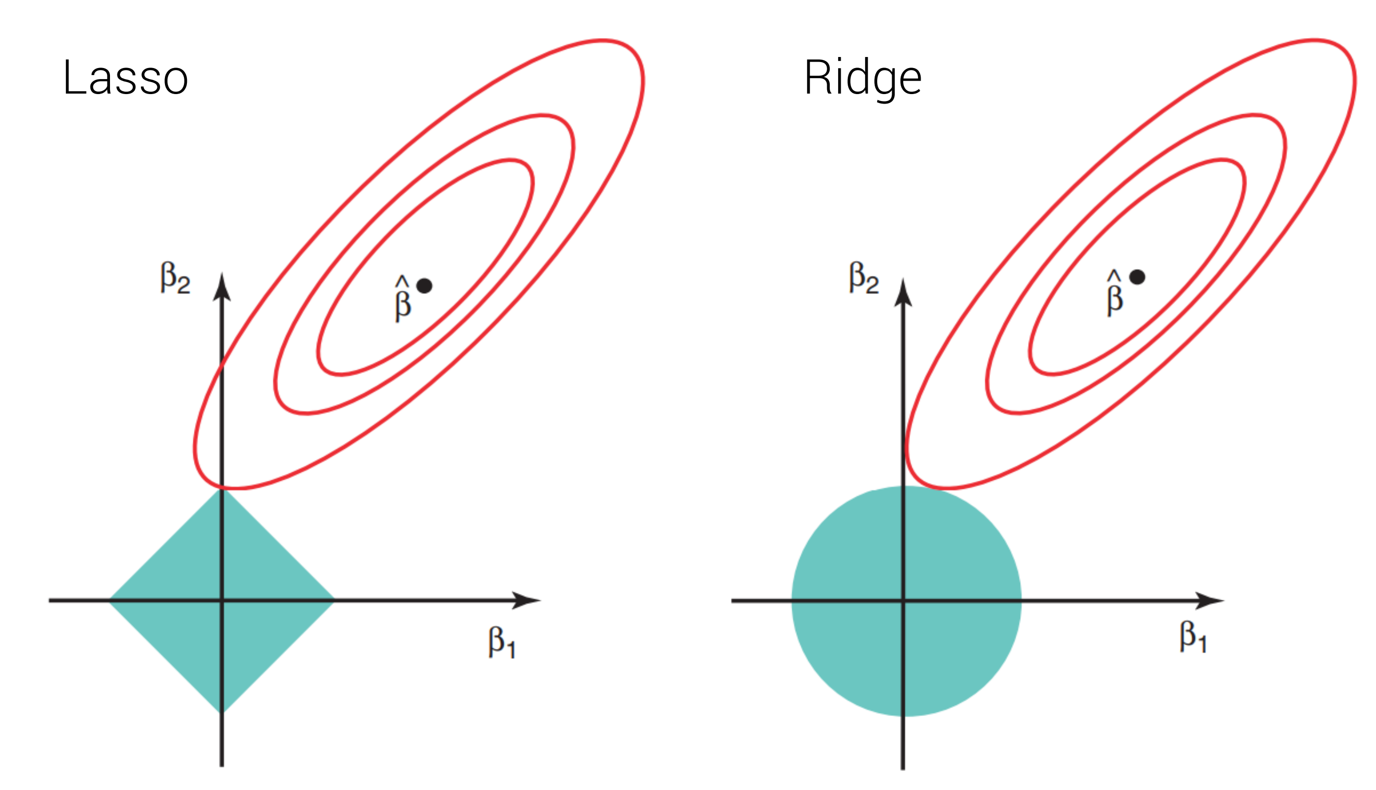


THE LASSO

- Very similar to ridge regression except we have the ability to set coefficients to zero if irrelevant – HUGE effect on variance
- Bias term:

$$\mathrm{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$

- This is my preferred shrinkage method



CASE STUDY

In the Market

Revisiting our Boston home price prediction....

CASE STUDY

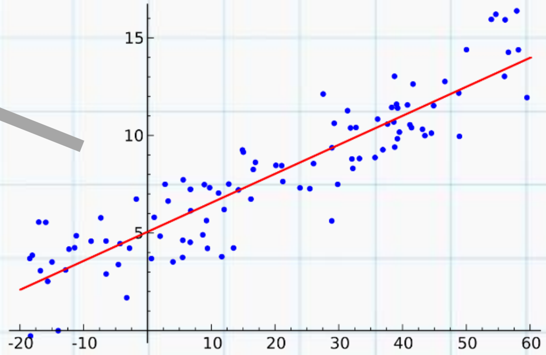
Ka-Chow!

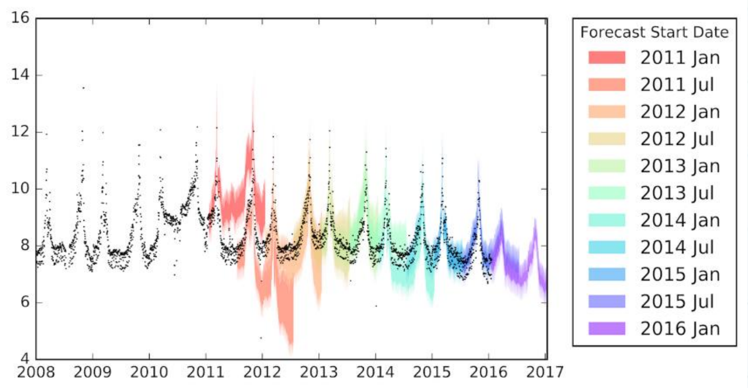
Let's fit some regression models to predict MPG for different cars

LINEAR REGRESSION

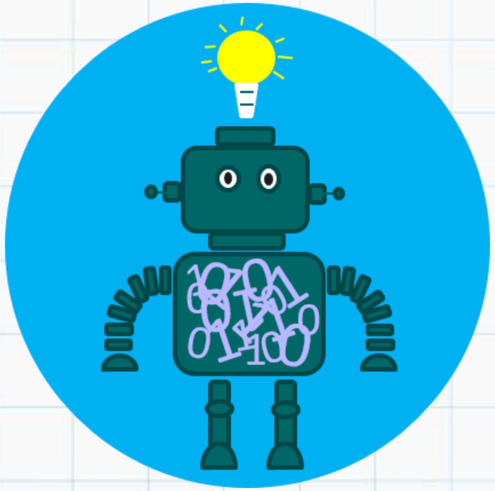
A Foundation of Data Science

Linear Regression



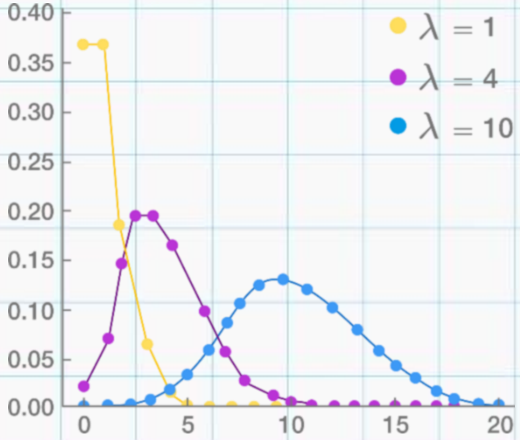


Time Series Analytics



Machine Learning

Neural Networks



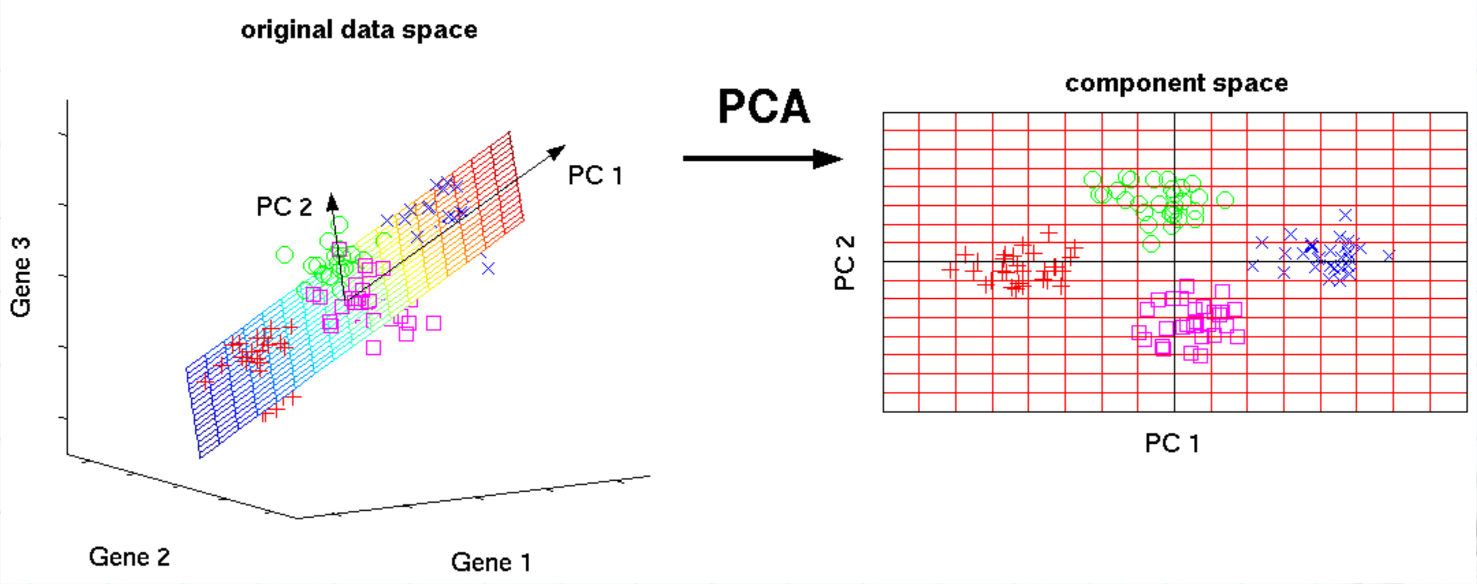
Count Data Analysis

LINEAR REGRESSION

A Foundation of Data Science

Additional *Important* Topics:

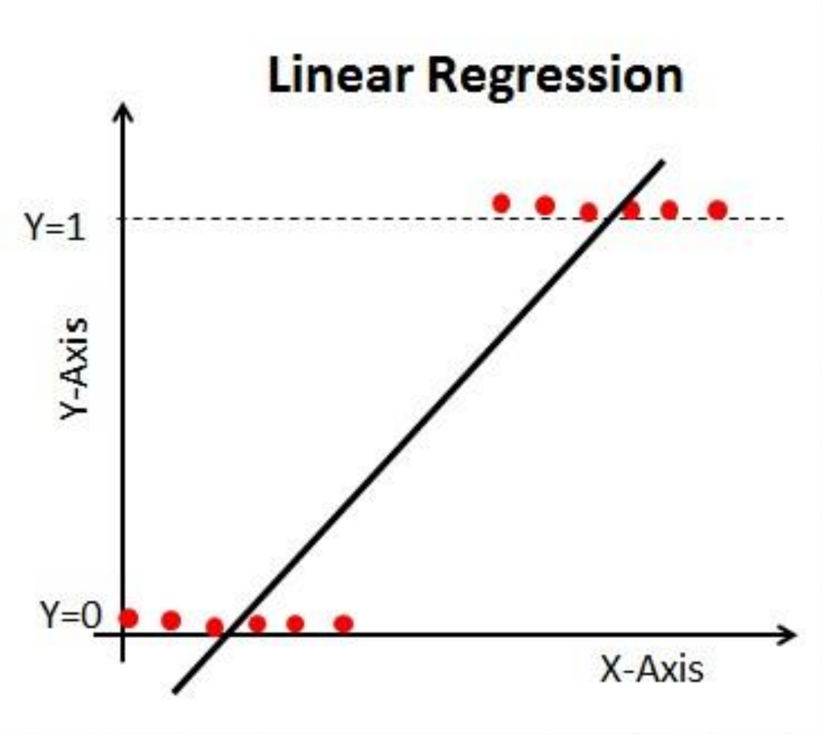
- Residual Analysis
  - Normal Q-Q
  - Leverage
  - Cook's Distance
- Nonlinear Regression
  - Regression Splines
- Model/Variable Selection
  - AIC, BIC, C_p
  - Subsetting
- Dimensionality Reduction
  - Principal Components Analysis



What about qualitative responses?

I want to perform a linear regression to predict if I'll get accepted into a college or not. Y=1 if YES and Y=0 if NO.

Dummy Variables

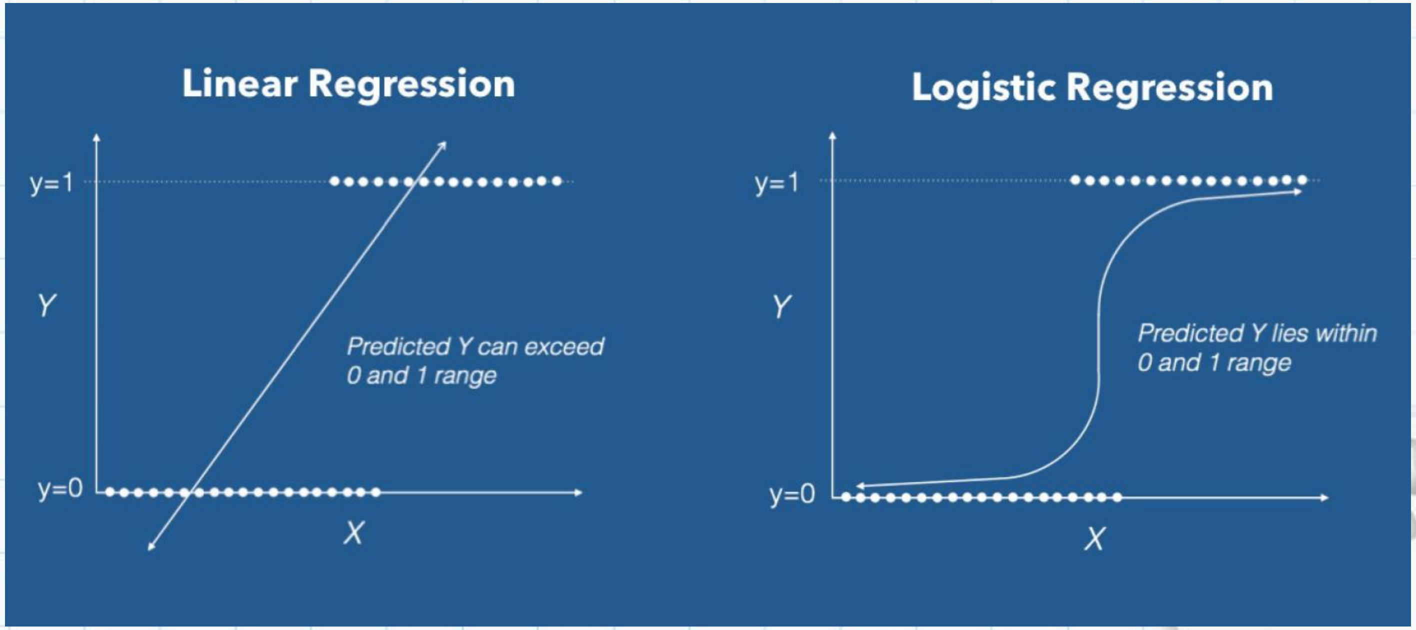


What could be wrong with this approach?

LOGISTIC REGRESSION

Classification

- A simple classification method for binary categorical responses
- Predicts the probability of an outcome that can only have two values
- Linear regression is unbounded; logistic regression only falls within [0,1]
- Typically for predicting a TRUE/FALSE or YES/NO problem
  - An email is spam (1) or not spam (0)
  - Will I get accepted into a college (1) or not (0)?



LOGISTIC REGRESSION

Odds and Log-Odds

- I will derive the logistic regression equations on the board

- The Odds
  - Value of odds close to 0 = very low probabilities
    ∞ = very high probabilities

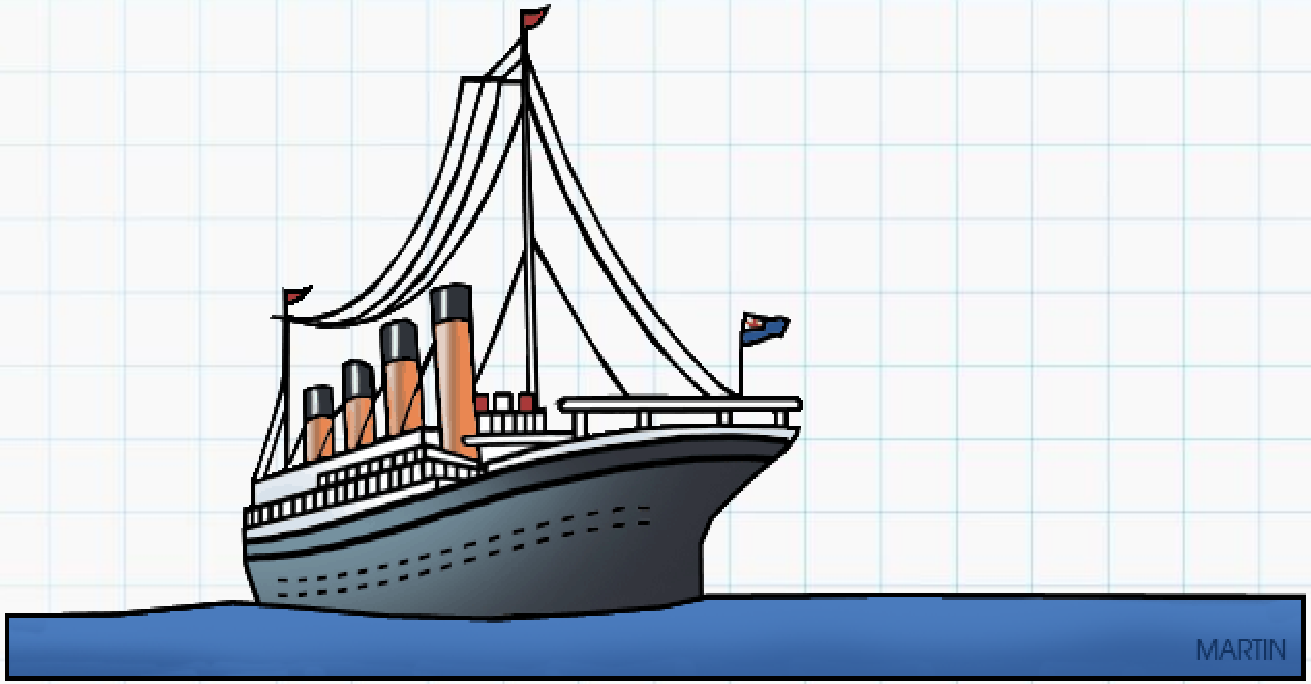
- The Log-Odds / Logit
  - Coefficients are estimated by the maximum likelihood estimation
  - Essentially, β_0, β_1 are determined such that p(x) = 1 when an observation belongs to a class
    p(x) = 0 when it does NOT belong to a class

- Not all values of p(x) = 0 or 1; we need to define a threshold. Usually, it's 0.5
  - If p(x) < 0.5, we round down to p(x) = 0 and state that the observation does NOT belong to a class

CASE STUDY

Sinking Ship

Would you survive the Titanic?



CONNECT WITH ME

www.linkedin.com/in/andira-putri/

@diramputri

diramputri@gmail.com

THAT'S ALL FOLKS!

Thank You

QUESTIONS?