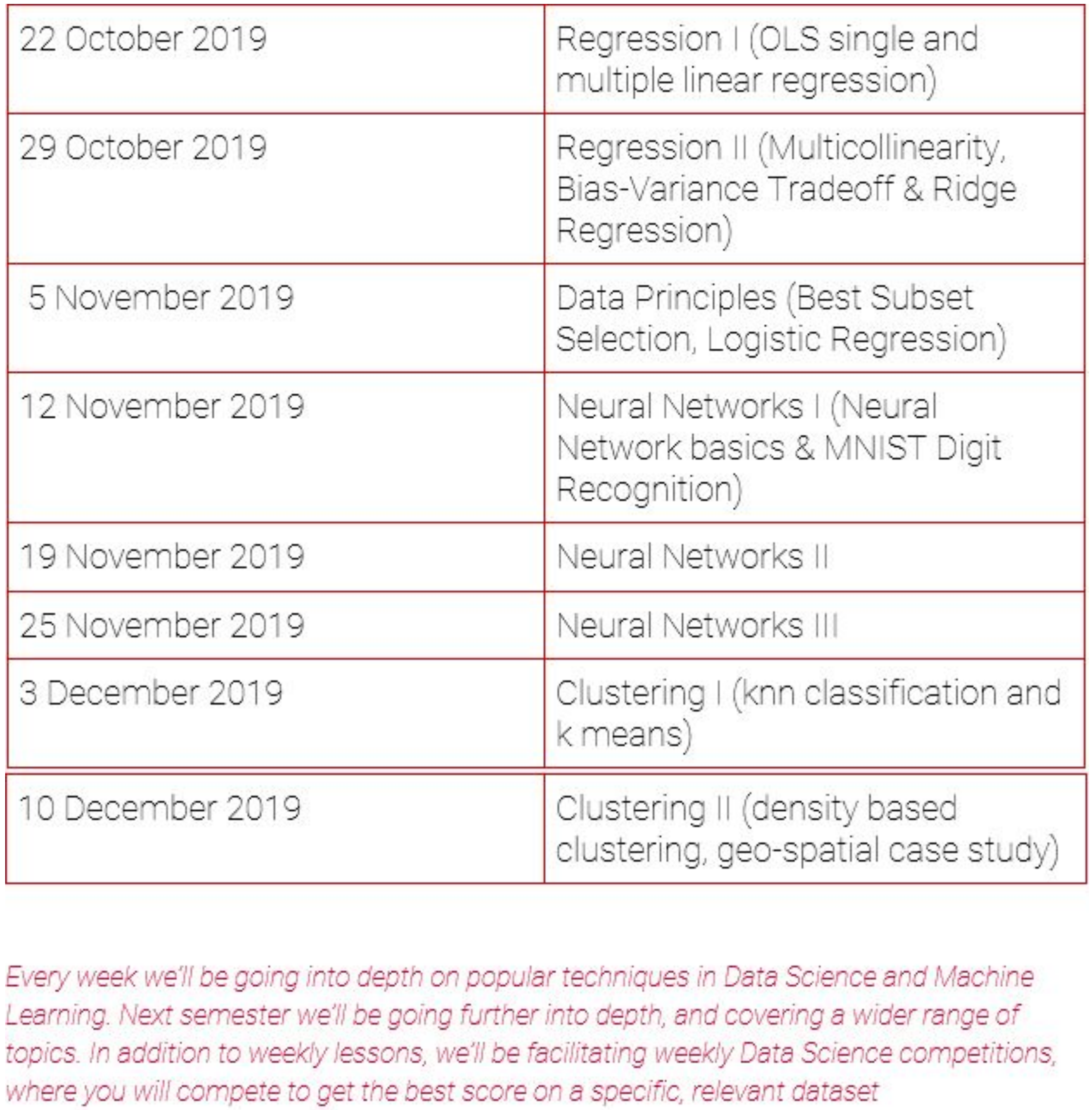
Data Science Bootcamp
1st Meeting

Rutgers Statistics Club x Cognitive Science Club

WELCOME!



22 October 2019	Regression I (OLS single and multiple linear regression)
29 October 2019	Regression II (Multicollinearity, Bias-Variance Tradeoff & Ridge Regression)
5 November 2019	Data Principles (Best Subset Selection, Logistic Regression)
12 November 2019	Neural Networks I (Neural Network basics & MNIST Digit Recognition)
19 November 2019	Neural Networks II
25 November 2019	Neural Networks III
3 December 2019	Clustering I (knn classification and k means)
10 December 2019	Clustering II (density based clustering, geo-spatial case study)

Every week we'll be going into depth on popular techniques in Data Science and Machine Learning. Next semester we'll be going further into depth, and covering a wider range of topics. In addition to weekly lessons, we'll be facilitating weekly Data Science competitions, where you will compete to get the best score on a specific, relevant dataset

IDE info

- You can either use google colab
- Or use your own python IDE

Linear Regression

What, Why & When

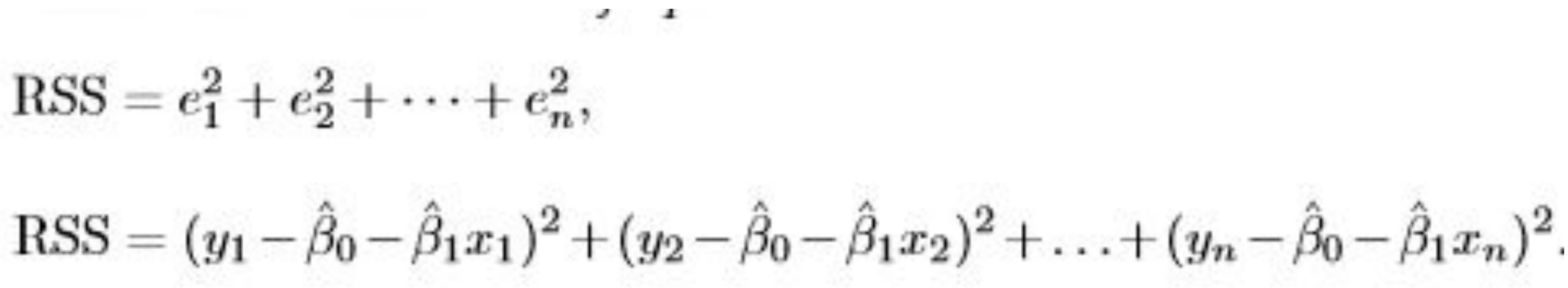
- What:
    - Linear Regression is a supervised learning tool which is useful for predicting a quantitative response.
- Why (are you learning this):
    - Because linear regression is still useful, in fact it’s the most commonly used machine learning tool out there. Furthermore, it serves as a good starter for newer approaches; many statistical learning approaches can be seen as generalizations or extensions of linear regression.
- When (do I use this):
    - Is there a relationship between X&Y
    - How strong is the relationship between X&Y
    - How well can I estimate X&Y
    - Etc, etc

Simple Linear Regression Example

- Simple, single variable linear regression
- We’re going to use X, the *independent variable* to predict Y, the *dependent variable*.
- Dataset: x,y = (0,1),(2,2),(3,3),(4,3.75)
- How well would we fit this if we did….
- Y = x+1?

Simple Linear Regression | Cost functions

- Definitions:
- Residual = y_{act}-y_{hat}
- RSS = sum of residuals squared



$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

Ordinary Least Squares Regression

- Y = β0 + β1x + e
- Where $\beta_1 = \frac{cov(x,y)}{var(x)}$
- $B_0 = y - \beta_1 x$

This minimizes RSS

(I’ll talk about unbiasedness here)

Evaluating our Model

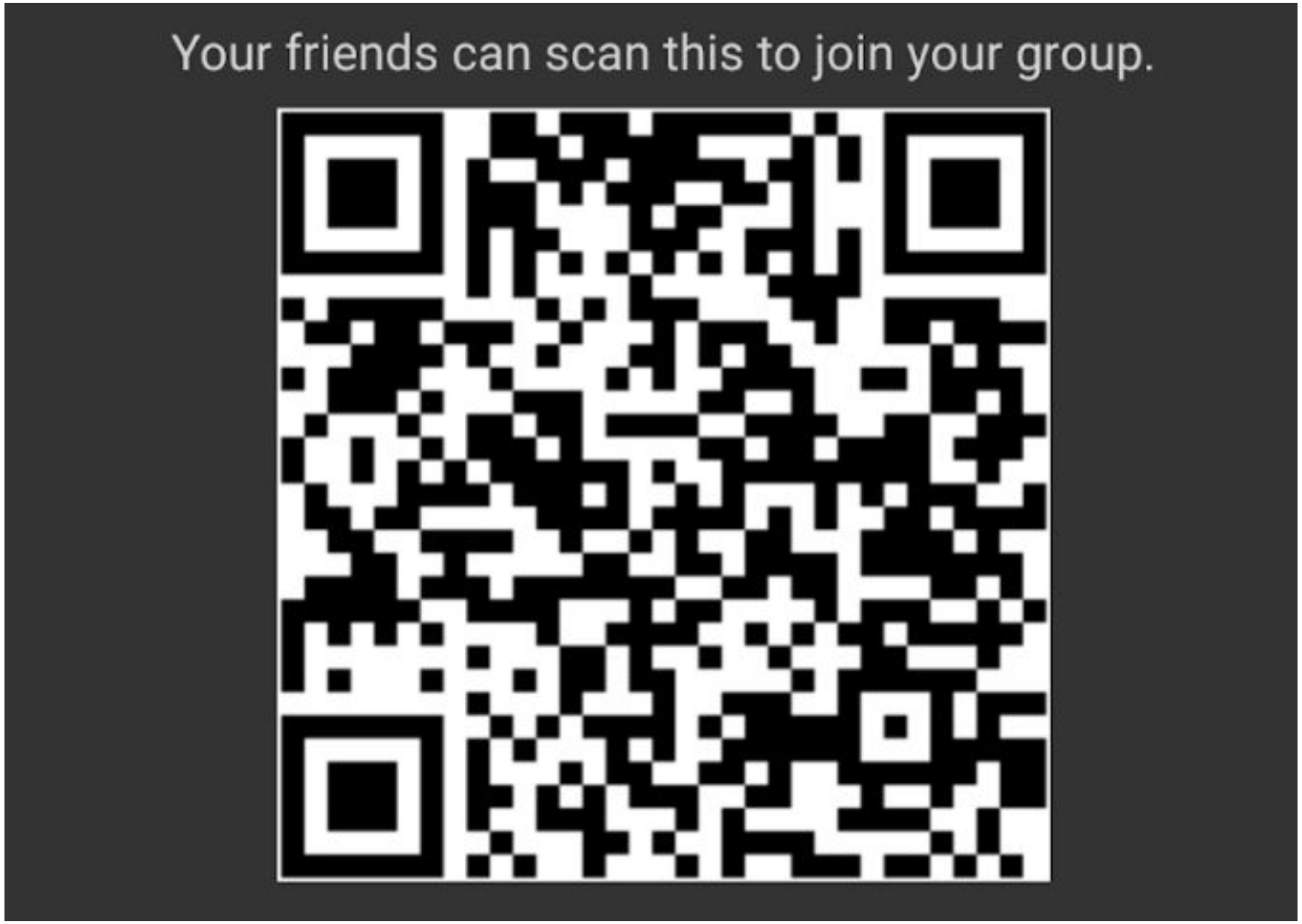
- Going back to our when (do I use this):
    - To determine if there is a relationship between Y and X:
        - We find the pvalue for how significant B1 is.
    - To determine how strong the relationship is:
        - We see the value of B1
    - To determine how well we can predict Y using X:
        - We use R^2, the coefficient of determination
- Important! Intercept may or may not have meaning

Let’s Practice (Boston Housing)

Multiple Linear Regression Model

- We now add multiple Xi’s (i = 1,2…,n) in linear model to get a better fit.
- However now we have to use an Fscore to determine if the model is signficant.

Join the GroupMe



Your friends can scan this to join your group.