GENERAL ASSEMBLY

LINEAR REGRESSION

Brian Chung

DATA SCIENCE PREDICTIONS FOR 2016

"In 2016, the world of big data will focus more on smart data, regardless of size. Smart data are wide data (high variety), not necessarily deep data (high volume). Data are "smart" when they consist of feature-rich content and context (time, location, associations, links, interdependencies, etc.) that enable intelligent and even autonomous data-driven processes, discoveries, decisions, and applications."

- Kirk Borne, Principal Data Scientist at Booze Allen Hamilton and founder of RocketDataScience.org

"2016 will be the year of deep learning. Data will move from experimental to deployed technology in image recognition, language understanding, and exceed human performance in many areas."

- Gregory Piatetsky, President of KDNuggets

"In 2015 we learned that 90% of the world's data had been created in the previous 12 months. In the middle of this BigData explosion, I watched many executives desperate to get themselves up-to-speed as quickly as possible, in order to understand the commercial opportunities that these vast quantities of information will offer their business.

In 2016 - I hope to see those same executives not just looking towards how they can capture as much commercial value from that information as possible, but how they can create the best experiences for their customers. The bigdata motto for 2016 therefore needs to be "We must create more value from data than we capture."

- Jeremy Waite, Head of Digital Strategy at EMEA Salesforce Marketing Cloud

http://blog.import.io/post/22-data-experts-share-their-predictions-for-2016

LINEAR REGRESSION AGENDA

I. Linear Regression
II. Use Cases
III. Pros and Cons

I. LINEAR REGRESSION

TYPES OF ML SOLUTIONS

	Continuous	Categorical
Supervised	Regression	Classification
Unsupervised	Dimension Reduction	Clustering

INTRO TO REGRESSION

Q: What is a **regression** model?

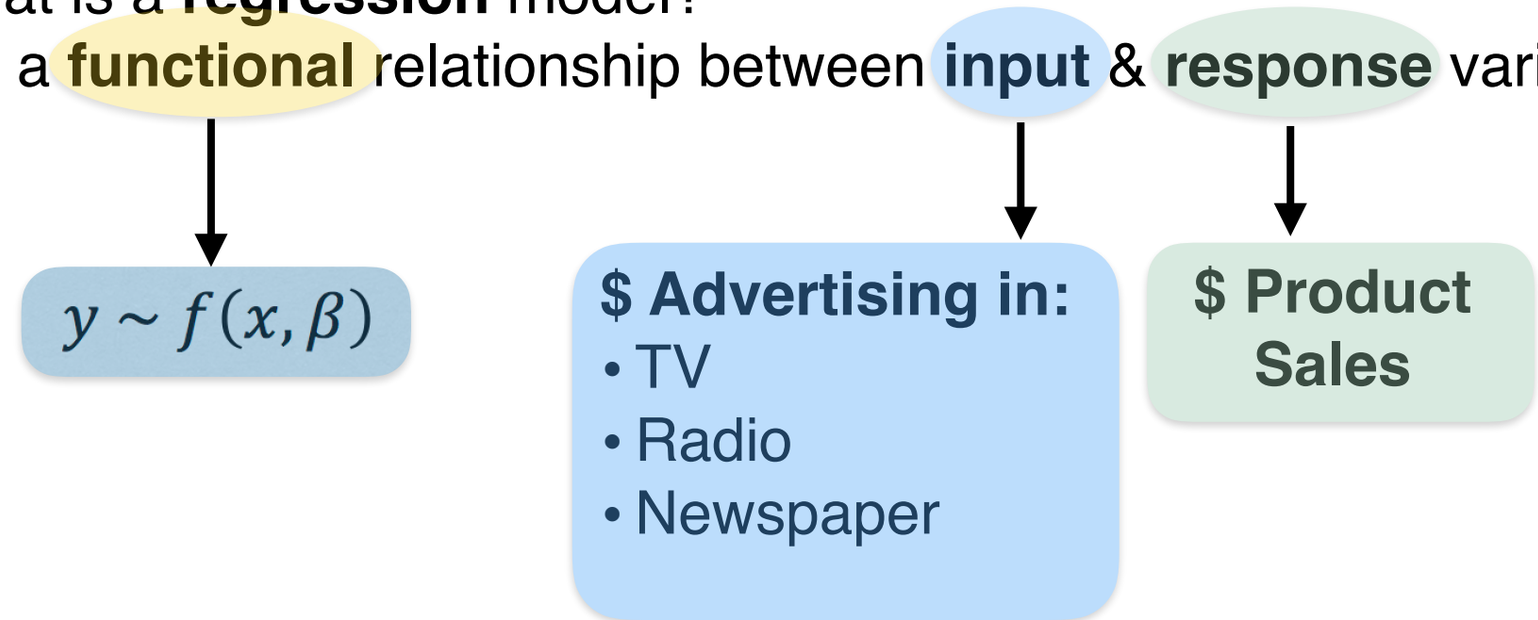
INTRO TO REGRESSION

Q: What is a **regression** model?
A: It is a functional relationship between input & response variables

INTRO TO REGRESSION

Q: What is a **regression** model?

A: It is a **functional** relationship between **input** & **response** variables



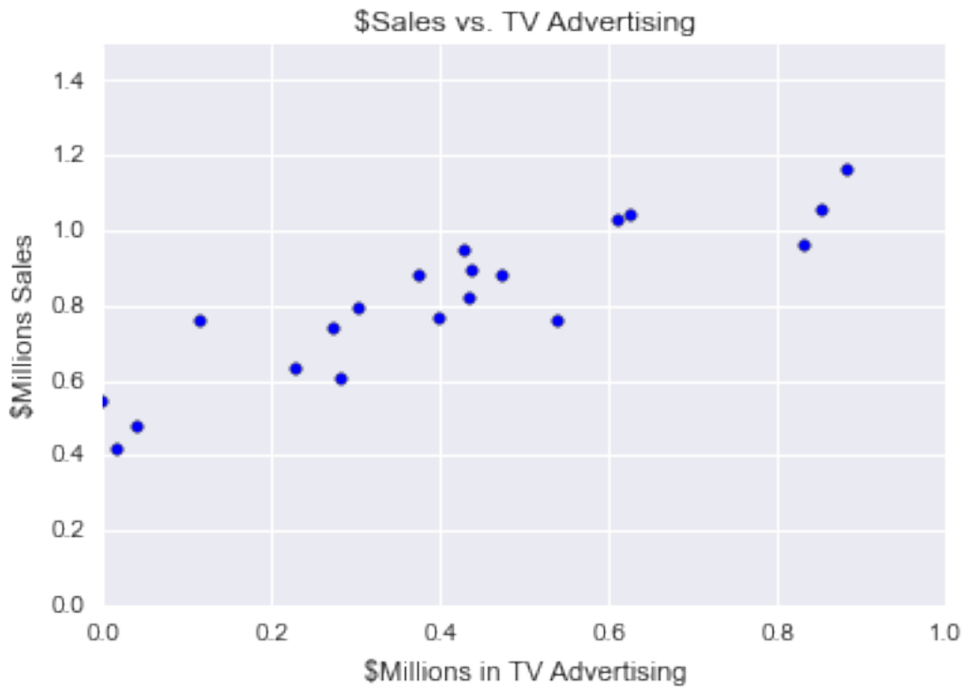
INTRO TO REGRESSION

Q: What is a **regression** model?

A: It is a functional relationship between input and response variables

The **simple linear regression** model captures a linear relationship between a single input variable ***x*** and a response variable ***y***

$$y = \alpha + \beta x + \varepsilon$$



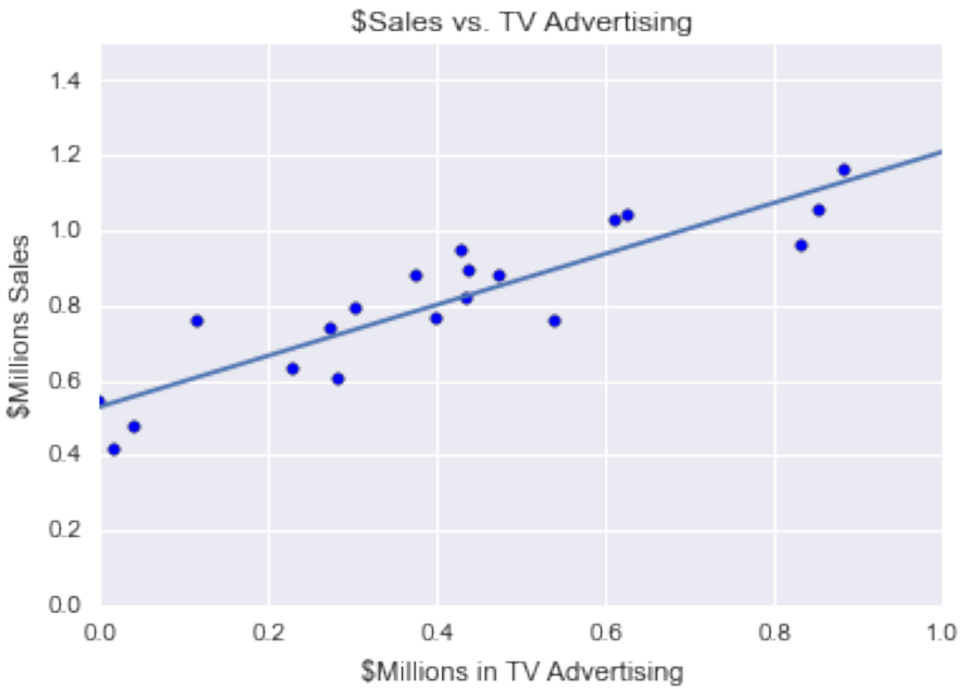
INTRO TO REGRESSION

Q: What is a **regression** model?

A: It is a functional relationship between input and response variables

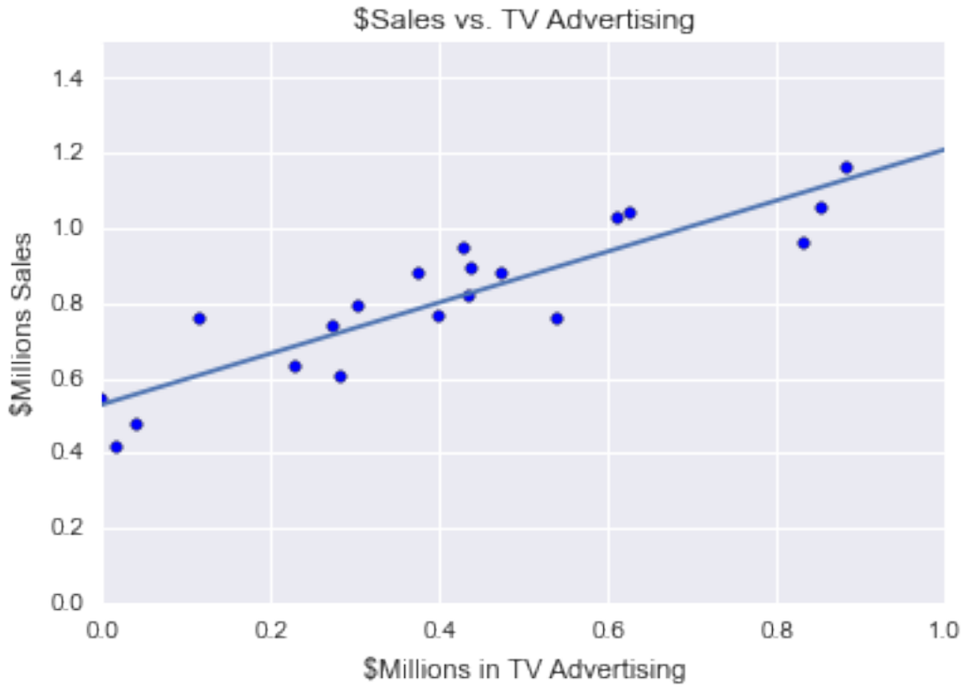
The **simple linear regression** model captures a linear relationship between a single input variable ***x*** and a response variable ***y***

$$y = \alpha + \beta x + \varepsilon$$



INTRO TO REGRESSION

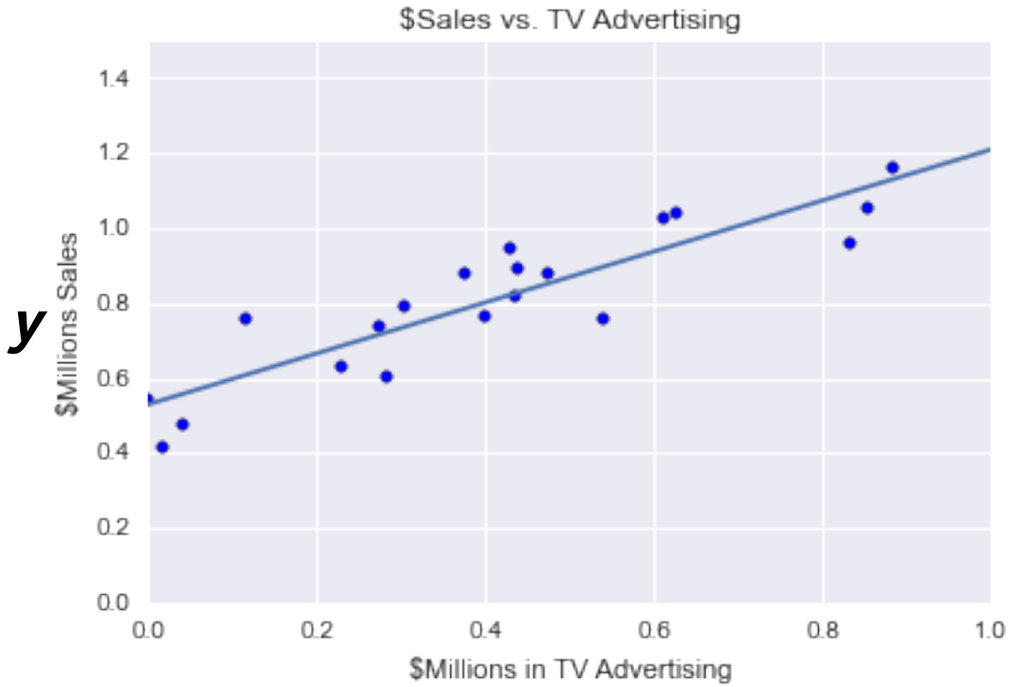
$$y = \alpha + \beta x + \varepsilon$$



INTRO TO REGRESSION

$$y = \alpha + \beta x + \varepsilon$$

y : Predicted Variable ($Millions Sales)

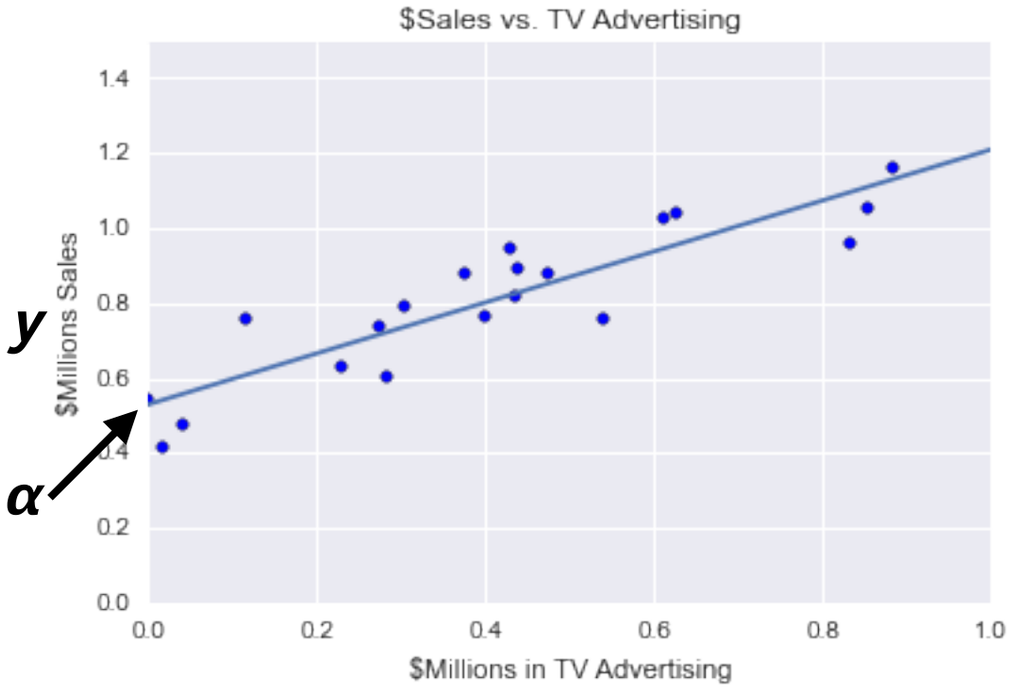


INTRO TO REGRESSION

$$y = \alpha + \beta x + \varepsilon$$

***y*: Predicted Variable ($Millions Sales)**

α: Intercept Value



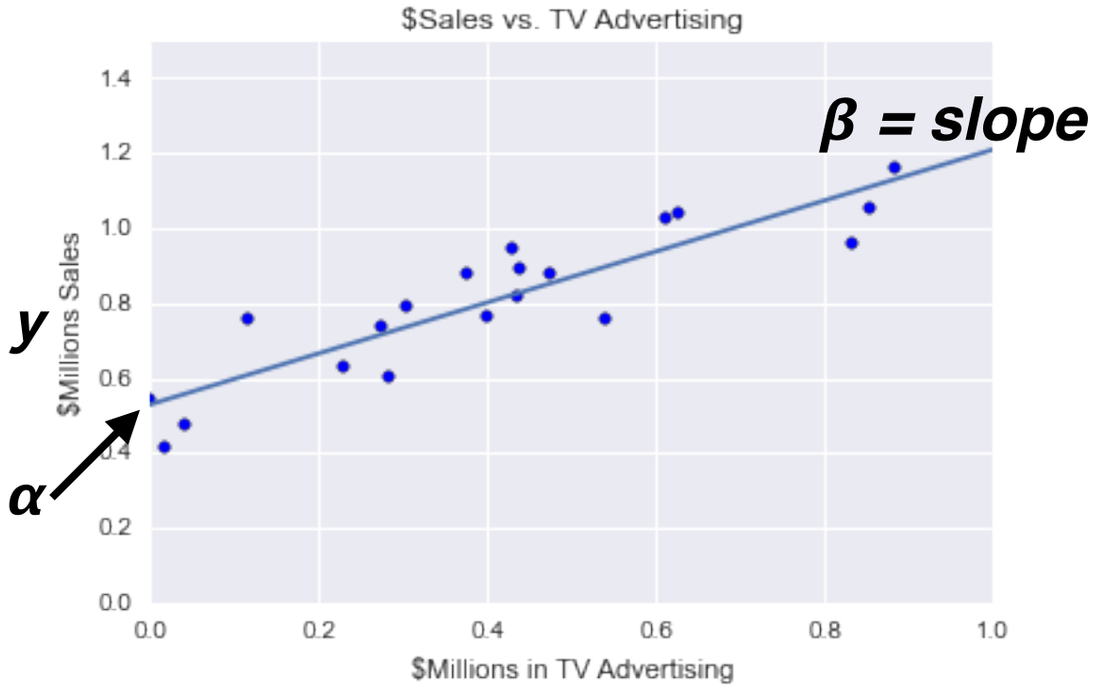
INTRO TO REGRESSION

$$y = \alpha + \beta x + \varepsilon$$

***y*: Predicted Variable ($Millions Sales)**

α: Intercept Value

β: Regression Coefficient



INTRO TO REGRESSION

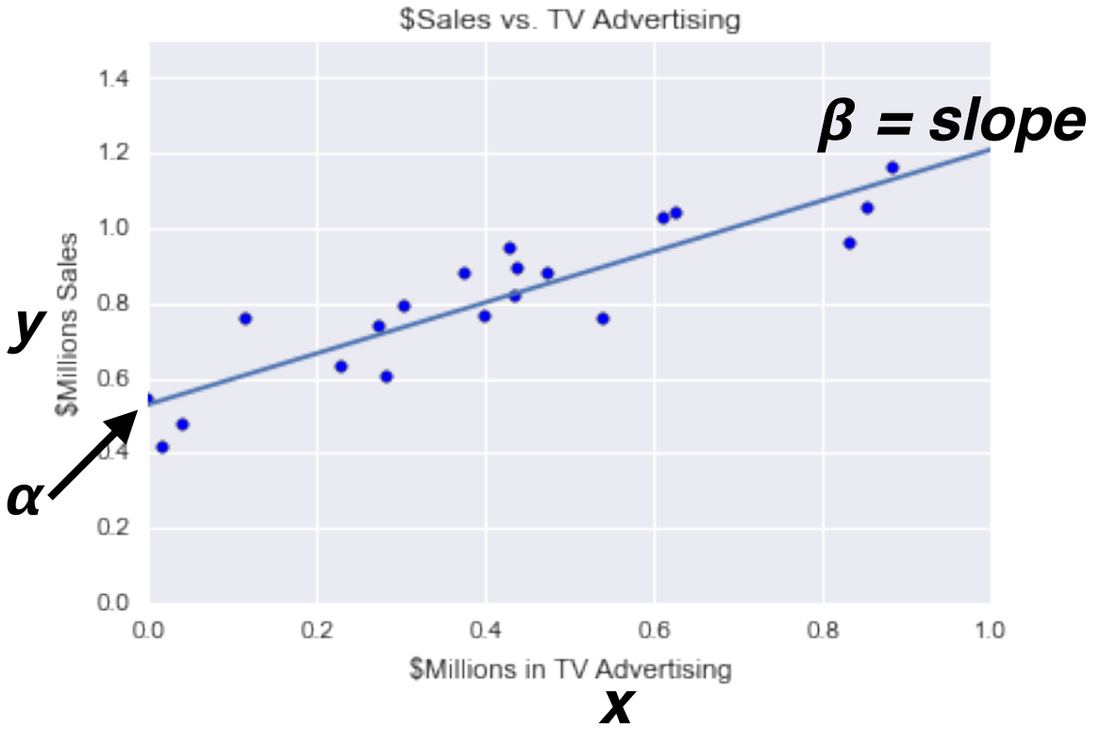
$$y = \alpha + \beta x + \varepsilon$$

y : Predicted Variable ($Millions Sales)

α : Intercept Value

β : Regression Coefficient

x: Input / Feature ($Millions TV Advertising)



INTRO TO REGRESSION

$$y = \alpha + \beta x + \varepsilon$$

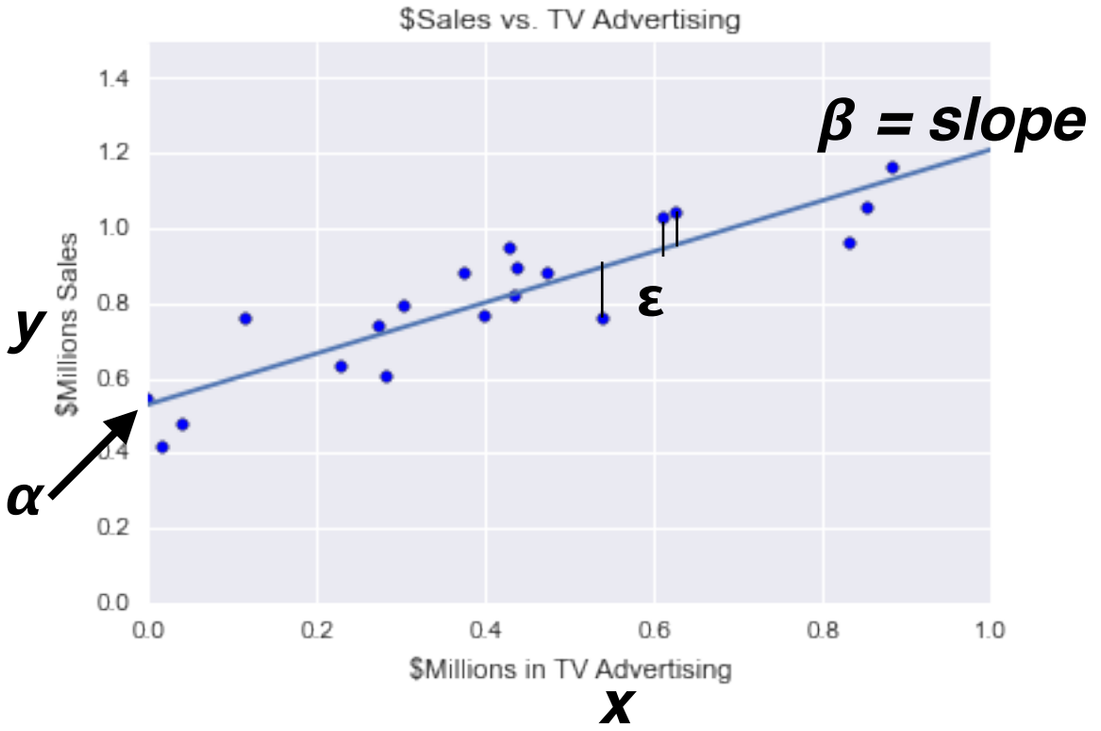
y : Predicted Variable ($Millions Sales)

α : Intercept Value

β : Regression Coefficient

x: Input / Feature ($Millions TV Advertising)

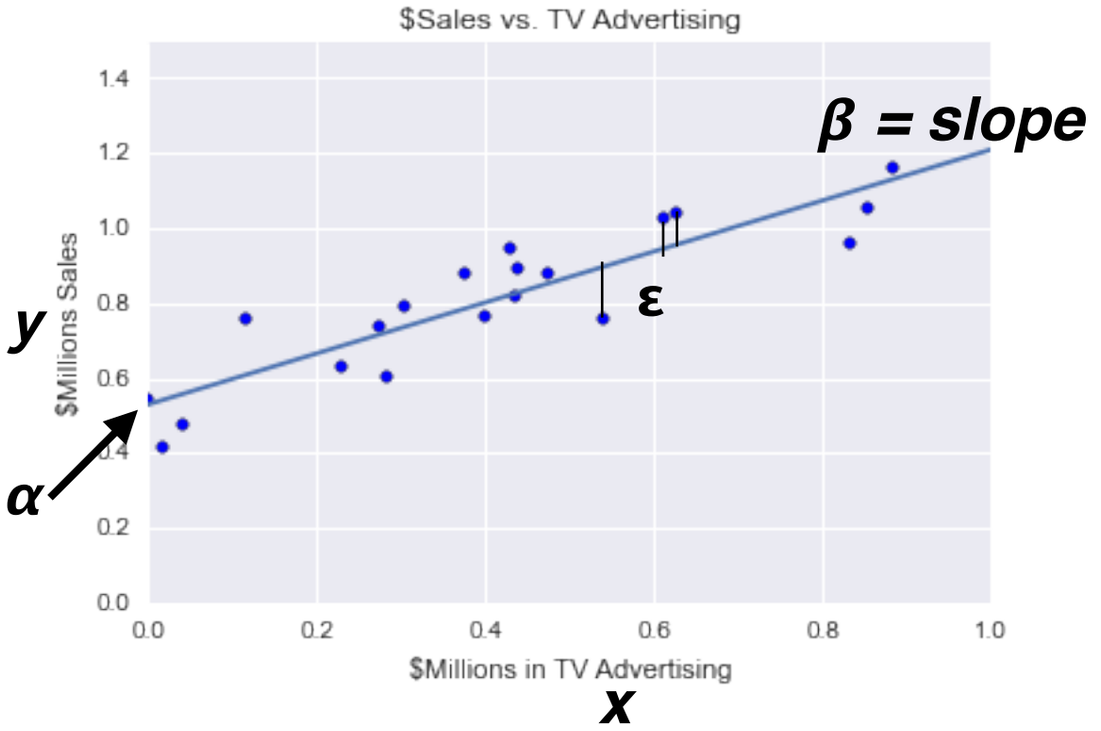
ε : Residual (Error)



INTRO TO REGRESSION

$$y = \alpha + \beta x + \varepsilon$$

y : Predicted Variable ($Millions Sales)
α : Intercept Value
β : Regression Coefficient
x: Input / Feature ($Millions TV Advertising)
ε : Residual (Error)

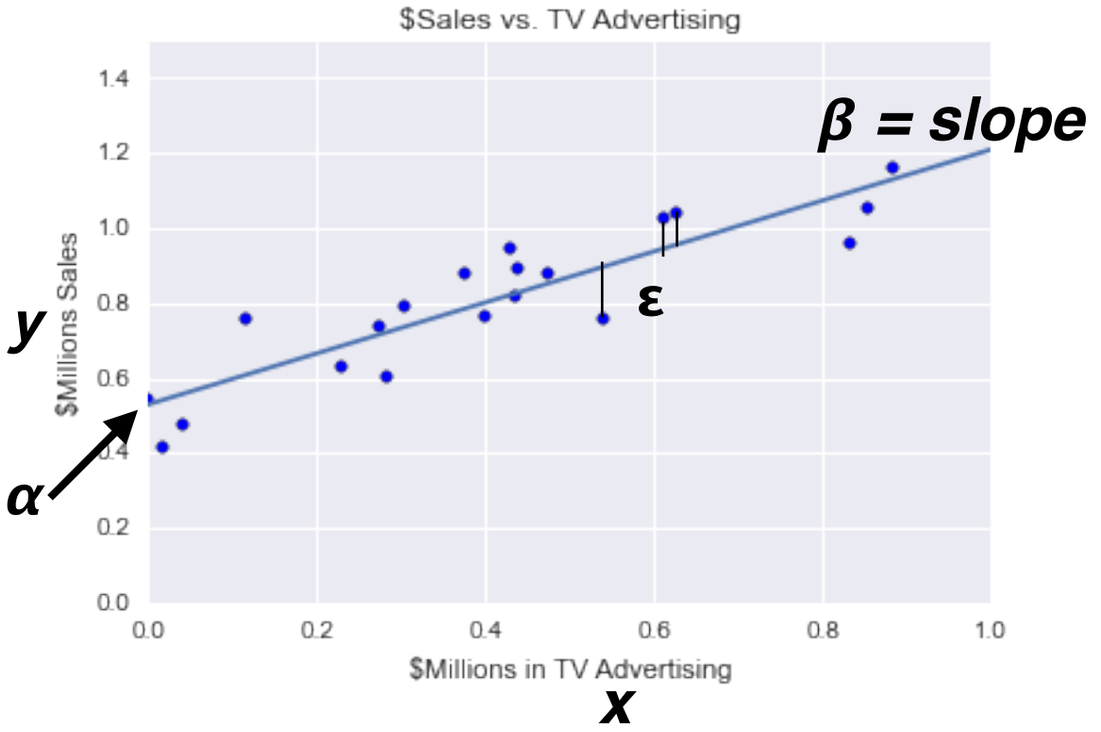


$$y = \alpha + \beta x + \varepsilon \longleftrightarrow \$Sales = 0.5 + 0.75 * (\$TV\ Advertising)$$

INTRO TO REGRESSION

$$y = \alpha + \beta x + \varepsilon$$

y: Predicted Variable ($Millions Sales)
α: Intercept Value
β: Regression Coefficient
x: Input / Feature ($Millions TV Advertising)
ε: Residual (Error)



Model: $Sales = 0.5 + 0.75 * ($TV Advertising)

*Prediction: "I've spent $0.75 in TV Advertising. My prediction for $Sales is: .5 + 0.75 * (.75) = $1.0625 million*

INTRO TO REGRESSION

Naturally, we can extend this to multiple input variables, giving us the **multiple linear regression** model

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \ldots + \beta_n x_n + \varepsilon$$

INTRO TO REGRESSION

Naturally, we can extend this to multiple input variables, giving us the **multiple linear regression** model

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \ldots + \beta_n x_n + \varepsilon$$

***y*: Predicted Sales**

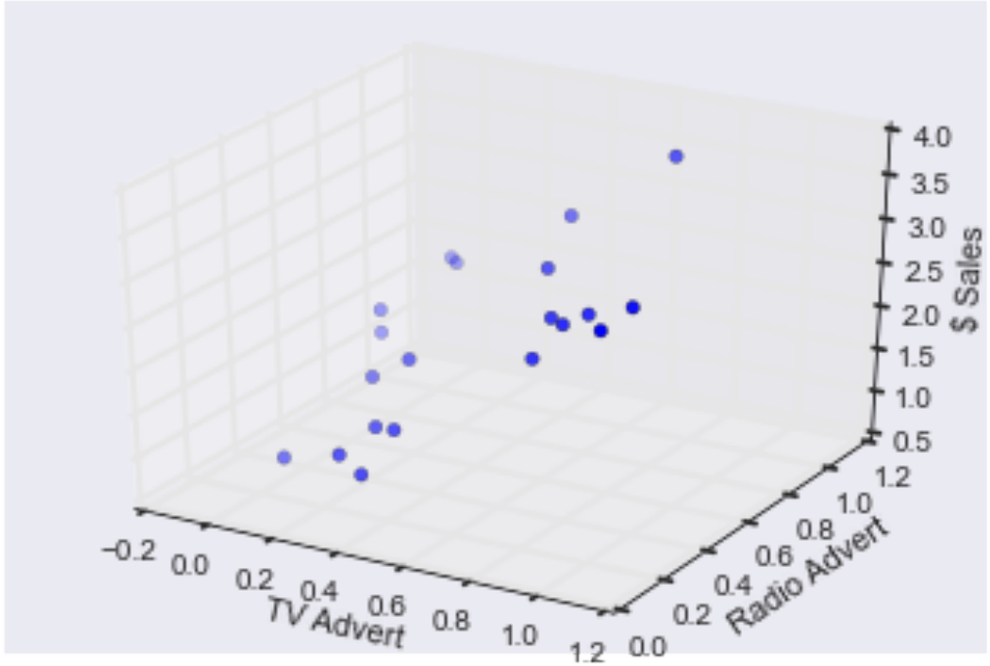
α: Intercept Value

β_1: Regression Coefficient (Beta1)

β_2: Regression Coefficient (Beta2)

...

ε: Residual (Error)

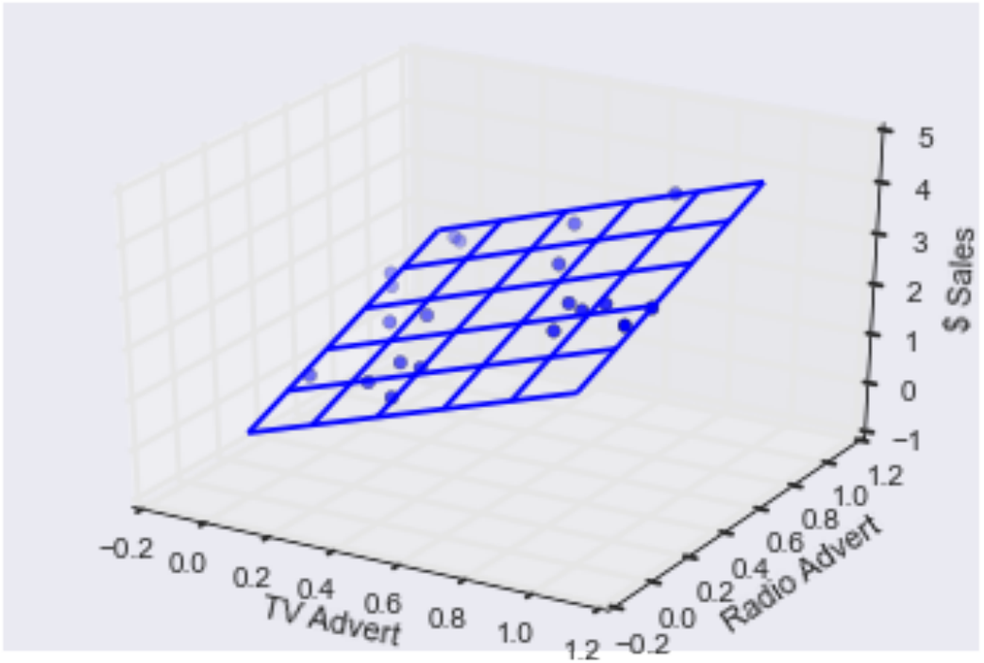


INTRO TO REGRESSION

Naturally, we can extend this to multiple input variables, giving us the **multiple linear regression** model

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \ldots + \beta_n x_n + \varepsilon$$

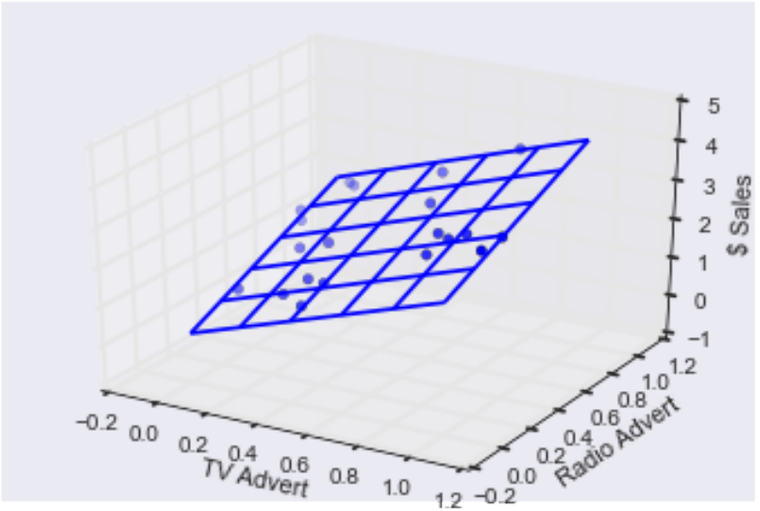
***y*: Predicted Sales**
***α*: Intercept Value**
***β_1*: Regression Coefficient (Beta1)**
***β_2*: Regression Coefficient (Beta2)**
…
***ε*: Residual (Error)**



INTRO TO REGRESSION

Naturally, we can extend this to multiple input variables, giving us the **multiple linear regression** model

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \ldots + \beta_n x_n + \varepsilon$$



Model: ***$Sales = 0.5 + 0.85 * ($TV Advert.) + 0.25 * ($Radio Advert.)***

Prediction: "I've spent $1 million in TV advertising and $.5 million in radio advertising
My prediction for $Sales is:

0.85*(1) + 0.25 * (.5) = $1.475 million

INTRO TO REGRESSION

Q: Great! So, how do we solve for the α and $\beta_1, \beta_2, \ldots, \beta_n$, etc. coefficients?

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \ldots + \beta_n x_n + \varepsilon$$

We have data points for **y** (Sales), as well as all the **x** (TV, Radio, Newspaper)

INTRO TO REGRESSION

Q: Great! So, how do we solve for the α and $\beta_1, \beta_2, \ldots, \beta_n,$ etc. coefficients?

A: In practice, any decent software package (sklearn, statsmodel, R, etc.) will calculate this for you through **OLS** (Ordinary Least Squares)

SOLVING FOR REGRESSION COEFFICIENTS

Q: Great! So, how do we solve for the α and $\beta_1, \beta_2, \ldots, \beta_n,$ etc. coefficients?

A: In practice, any decent software package (sklearn, statsmodel, R, etc.) will calculate this for you through **OLS** (Ordinary Least Squares)

BUT… knowing how linear regressions are solved are important for any data science interview and job!!

Let's go over the basics…

SOLVING FOR REGRESSION COEFFICIENTS

Let's go back to our multiple regression formula:

For a single data point…

$$truth: y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$$

$$prediction: \hat{y} = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

SOLVING FOR REGRESSION COEFFICIENTS

The residual (error) is equal to the observed y minus the predicted y

$$truth : y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$$
$$prediction : \hat{y} = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

$$error : \varepsilon = (y - \hat{y})$$

SOLVING FOR REGRESSION COEFFICIENTS

The **Ordinary Least Squares** solution first defines a cost function “J” that aggregates the total error over all the points. For a given **alpha** and **beta** parameters, the cost could be higher (i.e. alpha and beta are not great at minimizing the total error), or lower (i.e. alpha and beta are good parameters that overall reduce the total error).

$$J(\alpha,\beta) = \sum_{i} (y - \hat{y})^2$$

SOLVING FOR REGRESSION COEFFICIENTS

The **Ordinary Least Squares** solution first defines a cost function “J” that aggregates the total error over all the points. For a given **alpha** and **beta** parameters, the cost could be higher (i.e. alpha and beta are not great at minimizing the total error), or lower (i.e. alpha and beta are good parameters that overall reduce the total error).

$$J(\alpha,\beta) = \sum_{i} (y - \alpha - \beta_1 x_1 \beta_2 x_2 - \ldots - \beta_n x_n)^2$$

SOLVING FOR REGRESSION COEFFICIENTS

Let’s simplify this using linear algebra**…**

SOLVING FOR REGRESSION COEFFICIENTS

Given the **multiple linear regression model**…

$$y_1 = \alpha + \beta_1 x_{11} + \beta_2 x_{12} + ... + \beta_n x_{1n} + \varepsilon_1$$

$$y_2 = \alpha + \beta_1 x_{21} + \beta_2 x_{22} + ... + \beta_n x_{2n} + \varepsilon_2$$

$$y_3 = \alpha + \beta_1 x_{31} + \beta_2 x_{32} + ... + \beta_n x_{3n} + \varepsilon_3$$

$$...$$

$$y_m = \alpha + \beta_1 x_{m1} + \beta_2 x_{m2} + ... + \beta_n x_{mn} + \varepsilon_m$$

SOLVING FOR REGRESSION COEFFICIENTS

Given the **multiple linear regression model**…

$$y_1 = \alpha + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_n x_{1n} + \varepsilon_1$$
$$y_2 = \alpha + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_n x_{2n} + \varepsilon_2$$
$$y_3 = \alpha + \beta_1 x_{31} + \beta_2 x_{32} + \ldots + \beta_n x_{3n} + \varepsilon_3$$
$$\ldots$$
$$y_m = \alpha + \beta_1 x_{m1} + \beta_2 x_{m2} + \ldots + \beta_n x_{mn} + \varepsilon_m$$

m samples

SOLVING FOR REGRESSION COEFFICIENTS

Given the **multiple linear regression model**…

n features

$$y_1 = \alpha + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_n x_{1n} + \varepsilon_1$$

$$y_2 = \alpha + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_n x_{2n} + \varepsilon_2$$

$$y_3 = \alpha + \beta_1 x_{31} + \beta_2 x_{32} + \ldots + \beta_n x_{3n} + \varepsilon_3$$

$$\ldots$$

$$y_m = \alpha + \beta_1 x_{m1} + \beta_2 x_{m2} + \ldots + \beta_n x_{mn} + \varepsilon_m$$

m samples

SOLVING FOR REGRESSION COEFFICIENTS

Given the **multiple linear regression model**…

n features

$$
\begin{aligned}
y_1 &= \alpha + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_n x_{1n} + \varepsilon_1 \\
y_2 &= \alpha + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_n x_{2n} + \varepsilon_2 \\
y_3 &= \alpha + \beta_1 x_{31} + \beta_2 x_{32} + \ldots + \beta_n x_{3n} + \varepsilon_3 \\
&\ldots \\
y_m &= \alpha + \beta_1 x_{m1} + \beta_2 x_{m2} + \ldots + \beta_n x_{mn} + \varepsilon_m
\end{aligned}
$$

m samples

residual errors

SOLVING FOR REGRESSION COEFFICIENTS

Q: This looks familiar… Can we reshape this into vectors and matrices?

$$y_1 = \alpha + \beta_1 x_{11} + \beta_2 x_{12} + ... + \beta_n x_{1n} + \varepsilon_1$$

$$y_2 = \alpha + \beta_1 x_{21} + \beta_2 x_{22} + ... + \beta_n x_{2n} + \varepsilon_2$$

$$y_3 = \alpha + \beta_1 x_{31} + \beta_2 x_{32} + ... + \beta_n x_{3n} + \varepsilon_3$$

$$...$$

$$y_m = \alpha + \beta_1 x_{m1} + \beta_2 x_{m2} + ... + \beta_n x_{mn} + \varepsilon_m$$

SOLVING FOR REGRESSION COEFFICIENTS

Q: This looks familiar… Can we reshape this into vectors and matrices?

A: Yes. Yes we can.

$$Y = \alpha + X\beta + \varepsilon$$

Y: Predicted Variable (m x 1)

α: Intercept Value (m x 1, same value repeating)

X: Matrix of input features (m x n)

β: Vector of regression coefficients (n x 1)

ε: Residuals per sample (m x 1)

SOLVING FOR REGRESSION COEFFICIENTS

Often, the *α is brought INTO the X and β matrix and vector.*

i.e. the first column of X is all "1" and the first row of β becomes the α value we wish to estimate

$$Y = X\beta + \varepsilon$$

Y: Predicted Variable (m x 1)

X: Matrix of input features + initial column of "1" (m+1 x n)

β: Vector of regression coefficients + initial row for intercept (n+1 x 1)

ε: Residuals per sample (m x 1)

SOLVING FOR REGRESSION COEFFICIENTS

Now we are ready. Remember our cost function “J”?
Let’s see it in matrix form…

$$Y = X\beta + \varepsilon$$

$$J(\beta) = \| (Y - X\beta) \|^2$$

SOLVING FOR REGRESSION COEFFICIENTS (ADVANCED)

In class solving for *"β"* that minimizes the cost function

$$Y = X\beta + \varepsilon$$

$$J(\beta) = \| (Y - X\beta) \|^2$$

SOLVING FOR REGRESSION COEFFICIENTS

The **ordinary least squares** solution for our coefficients

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

SOLVING FOR REGRESSION COEFFICIENTS

The **ordinary least squares** solution for our coefficients

There is a **closed form solution** as you've seen, however, many machine learning problems do not necessarily have a closed form solution

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

PREDICTING VALUES

We've solved for the regression coefficients for $\boldsymbol{\beta}$. How do we predict new data points?

PREDICTING VALUES

We've solved for the regression coefficients for $\boldsymbol{\beta}$. How do we predict new data points?

Remember the form of the linear equations!

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_n x_n$$

Take the beta values you have estimated, multiply by their corresponding input values, and take the sum to get your predicted **y** for that datapoint

PREDICTING VALUES

Alternatively, if you store your input data as a matrix **x**, then you can batch predict for multiple points…

$$\hat{Y} = X\hat{\beta}$$

HOW GOOD IS A FIT?

Pop Quiz: How did KNN measure goodness or quality of the model?

HOW GOOD IS A FIT?

Pop Quiz: How did KNN measure goodness or quality of the model?

Answer: KNN quality was measured by scoring: (# Correct) / (# Total)

We decided on the quality of a KNN model by understanding how well the model fit to the data

HOW GOOD IS A FIT?

Similarly, our measure of "goodness of fit" for linear regression will be directly related to how good our attempt to minimize the errors is

One such measure is the **Coefficient of Determination**, or **R squared**

$$MSE = \frac{1}{N}\sum (y - \hat{y})^2$$ **Mean Squared Error**

$$Var(y) = \frac{1}{N}\sum (y - \bar{y})^2$$ **Variance of y**

$$R^2 = 1 - \frac{MSE}{Var(y)}$$ **R squared**

HOW GOOD IS A FIT?

R squared ranges from 0 (poor fit—None of the natural variance of the data can be captured by the regression) to 1 (perfect fit—100% of the natural variance of the data can be captured by the regression**)

R squared can be related to the "fraction of explained variance"

- **Be careful! R square is not the end all of statistics in measuring a model**
- **Check if the linear model fits your needs for intuition and ease**
- **R squared can be gamed (See this later in exercise)**

$$R^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

I. LINEAR REGRESSION

II. USE CASES

WHEN TO USE A LINEAR REGRESSION?

- If you have a **supervised**, **quantitative** problem
- You think the response variable is some additive combination of variables (i.e. if the space of the response variable is highly jagged or irregular, it may be difficult to get good results with a linear regression)
- Generally, it's still always a good idea to try starting with a linear regression. It's pretty much free in terms of building models

WHEN TO USE A LINEAR REGRESSION?

In addition, knowing the assumptions behind a linear regression are helpful in deciding if your problem can be solved with a linear regression

- **Y** can be described as a **linear combination** of your **x** variables

WHEN TO USE A LINEAR REGRESSION?

In addition, knowing the assumptions behind a linear regression are helpful in deciding if your problem can be solved with a linear regression

- **Y** can be described as a **linear combination** of your **x** variables
- **E[error | x] = 0 -** Given x, your prediction error has mean zero (Not variance = 0!)

WHEN TO USE A LINEAR REGRESSION?

In addition, knowing the assumptions behind a linear regression are helpful in deciding if your problem can be solved with a linear regression

- **Y** can be described as a **linear combination** of your **x** variables
- **E[error | x] = 0 -** Given x, your prediction error has mean zero (Not variance = 0!)
- Errors have no linear dependence on x (i.e. homoskedasticity)

WHEN TO USE A LINEAR REGRESSION?

In addition, knowing the assumptions behind a linear regression are helpful in deciding if your problem can be solved with a linear regression

- **Y** can be described as a **linear combination** of your **x** variables
- **E[error | x] = 0 -** Given x, your prediction error has mean zero (Not variance = 0!)
- Variance of errors have no dependence on x (i.e. homoscedasticity)
- The features or input variables in x are linearly independent

WHEN TO USE A LINEAR REGRESSION?

Real life examples:

- Stock Return Prediction https://www.kaggle.com/c/the-winton-stock-market-challenge/data
- Predicting the cost of college education https://www.kaggle.com/c/us-dept-of-education-college-scorecard
- Predicting $ Sales of a coke-like-product in a midwestern region
- Estimating oil supply levels based on production quotas, ship movements, etc.
- Do body weight, calorie intake, fat intake, and age have influence on blood cholesterol level?
- Do customer satisfaction, brand perception, and price influence loyalty?
- By how much will 5 additional weeks of sunshine and 100mm of rainfall raise the sugar concentration in vine grapes?

I. LINEAR REGRESSION

II. USE CASES

III. PROS AND CONS

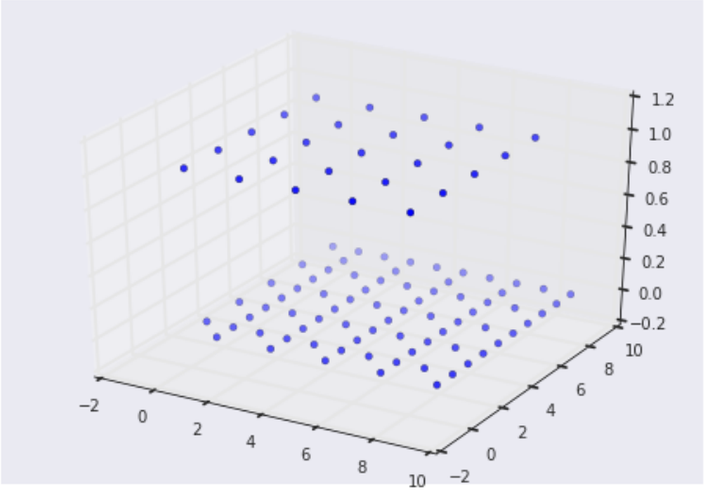
PROS AND CONS

PROS:

- Simple! It’s very intuitive to explain, predict, and use for inference
- Computations are very quick. Even 2000x200 matrices can be computed quickly
- Just do it

CONS:

- Not robust to outliers
- Can quickly overfit when there are a lot of features
  - (Will learn how to solve this issue in next lesson)
- Cannot really handle sharp inflections in responses.
  - (Will also learn how to handle these later)
- Kitchen Sink regression R^2



THAT'S IT!

- Exit Tickets: DAT1 - Lesson 5 - Linear
- Break until Jan 4, 2016
- Homework 3 is due Jan 4, 2016. Will be released by this Wednesday.
- Milestone 2 is due Jan 6, 2016.
- Project Proposals will be reviewed and emailed back
- Additional Resources for today's lesson
  - Regression Analysis By Example (Book)
  - Statistical Models: Theory and Practice (Book)
  - Elements of Statistical Learning (Book)