Machine Learning: Linear Regression

Elliott Saslow

Help from Moses & Chyld

galvanize

- What is a Model?
- How can we evaluate our Model?
- Why is it called Linear Regression?
- How can we interpret our Model Output?
- What are the Assumptions of Linear Regression?
- How can we verify these assumptions are met?

High Level: What is a model?

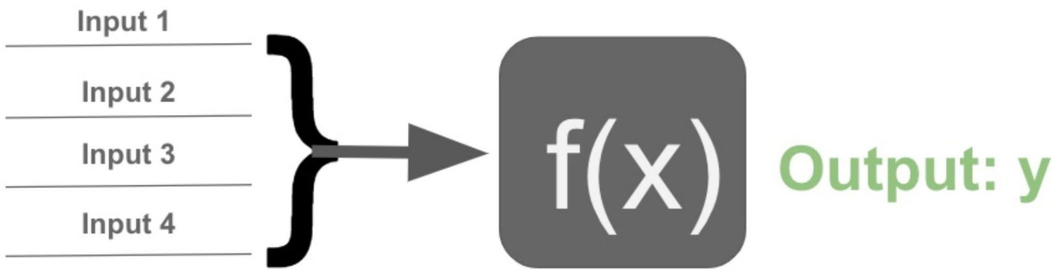
- Machine Learning is a set of tools to learn a very good approximation of the relation between features and a label.

True Models:

$$y = f(x) + \epsilon$$

- Machine Learning learns an approximation of $\hat{f}(x)$ of $f(x)$
- Use $\hat{f}(x)$ to predict y from new values of x

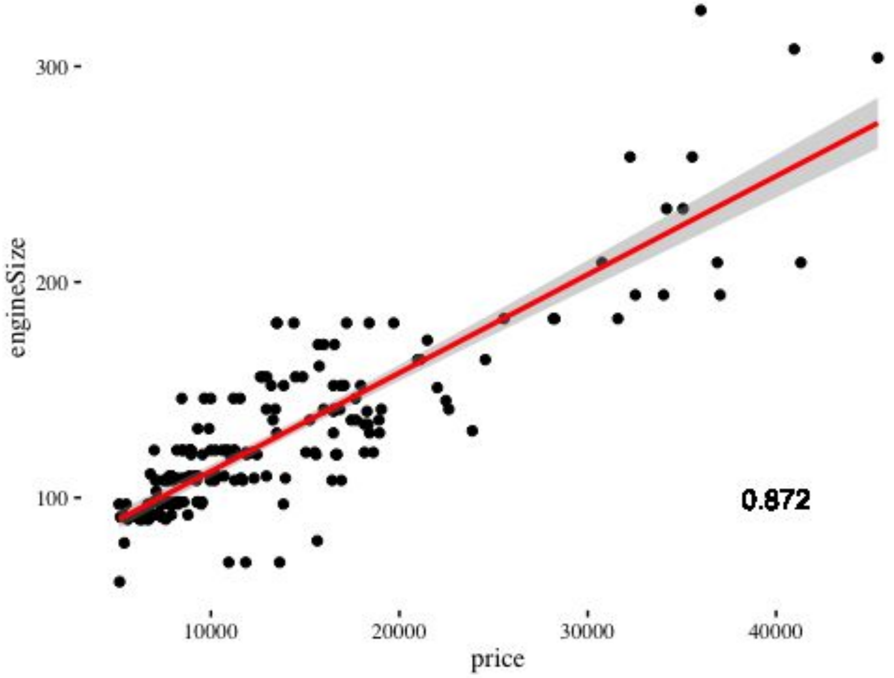
Example of a function



Supervised Machine Learning Models:

Supervised: Models a label with a feature

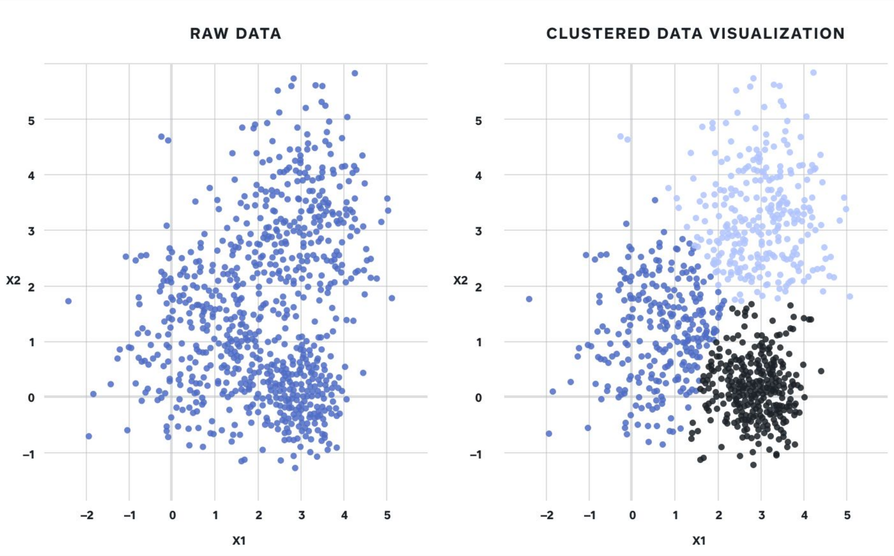
- Regression (Continuous Outcome)
    - Price
    - Demand
- Classification (Categorical or Discrete Outcome)
    - Fraud
    - Churn



Un - Supervised Machine Learning Models:

Unsupervised: Models without a label with a feature

- Clustering
  - Housing Prices
  - kmeans
- Dimensionality Reduction
  - Image Compression / Audio Compression
  - PCA, SVD, NMF, ect



Types of Data

Cross Section: x_i

- One observation per individual or cross-sectional unit
- Computed at one point in time
- Many i , One t

Time Series: x_t

- Multiple observations of a quantity over time (Ex. GDP)
- Computed at multiple instants
- One i , Many t

Panel Data: x_{it}

- Observe units over time, e.g. securities
- Many i at many t

And Many More….

Types of Features

Continuous:

- Price, Quantity, Sales, Tenure
- May bin using quantiles to model non-linearities better

Categorical:

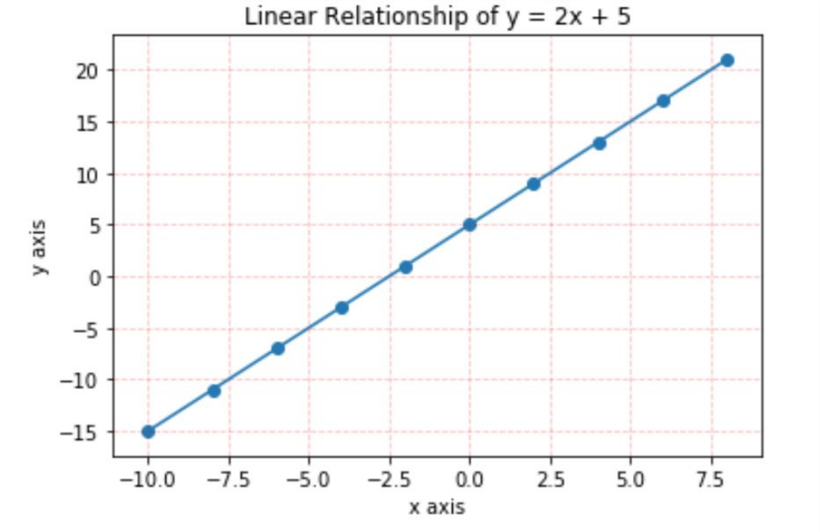
- Takes Discretes Levels
- Also called a factor
- e.g. 1/0, Yes/No, Treated/Control, High, Medium, Low

Text/Audio/Image:

- Need to Engineer Features

What are some of the issues we could run into with Continuous Features? What about Categorical Features?

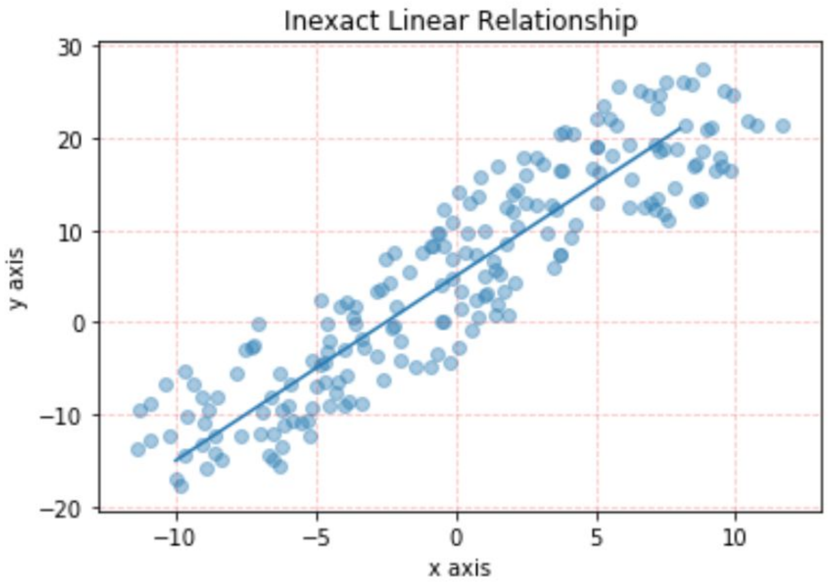
Exact Linear Relationship



Are we always going to have a Perfect Linear Relationship?

If not a Perfect Linear Relationship, what should we do?

In-Exact Linear Relationship



Add an Error Term!

$$y = 2 * x + 5 + \epsilon$$

Linear Regression Model

$$E[y|x] = X^T\beta$$

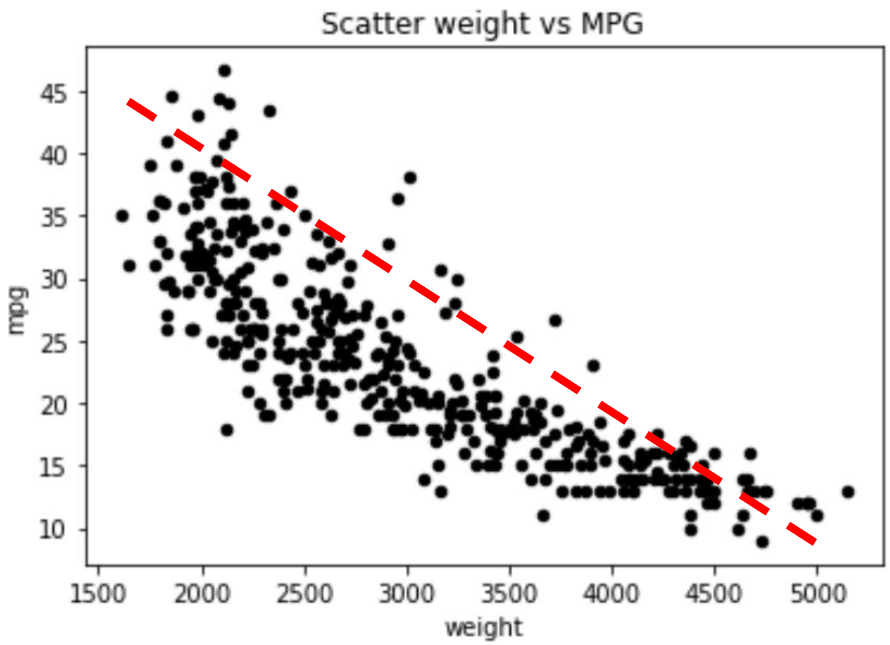
$$y_i = x_i^T\beta + \epsilon_i$$

In the 1-D case this becomes:

$$y_i = \beta_0 + x_1 * \beta_1 + \epsilon$$

Linear Regression models the **Expected value** of the outcome, conditional on features

Example 1 Dimensional!



The form of our equation is:

$$y_i = \beta_0 + x_1 * \beta_1$$

Here we have x1 = weight

Based on this visually, we could come up with a Beta_0 and Beta_1

y = m*x + b

Example Multidimensional:

Assume we have 3 features: Engine Size, Number of tires, and Car Weight

Given the best Beta values for each, can you calculate the MPG for each car?

$$E[y|x] = X^T \beta$$

$$X^T =$$

Engine Size	Number of Tires	Car Weight
1500	4	3000
2500	6	5000
1000	2	1200
3000	4	6000

$$\beta = \begin{array}{l} \beta\ EngineSize = -.000025 \\ \beta\ NumberTires = 3.113 \\ \beta\ Carweight = -.0003 \end{array}$$

Example Multidimensional Continued

$$E[y|x] = X^T\beta$$

Engine Size	Number of Tires	Car Weight
1500	4	3000
2500	6	5000
1000	2	1200
3000	4	6000

$$* \begin{bmatrix} -0.000025 \\ 3.113 \\ -0.0003 \end{bmatrix} =$$

Example Multidimensional Continued

- Linear Regression Predicts the mean value of y holding x constant

- Model is **linear** in parameters *Beta* but features may be **non-linear** functions of the date (e.g. polynomials)

https://phet.colorado.edu/sims/html/least-squares-regression/latest/least-squares-regression_en.html