Multiple linear regression

Multiple Linear Regression

- Its Simple Linear Regression, but with more Relationships

$$Y = \beta_0 + \beta_1 X$$

- Becomes

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

- The multiple linear regression explains the relationship between:
  - One continuous dependent variable (y) and
  - Two or more independent variables (x1, x2, x3… etc)

The Data

- Note that in the wild, when you do encounter a dataset, it is going to be ugly.
- It’s going to have
  - missing values,
  - erroneous entries,
  - wrongly formatted columns,
  - irrelevant variables… etc
- You’ll need to
  - Select good features,
  - Try different algorithms

Categorical Variables >>> Continuous Variables

- Due to the nature of the regression equation, your x variables have to be continuous as well.
- Thus, you'll need to look into changing your categorical variables into continuous ones.
- Continuous variables are running numbers.
- Categorical variables are categories.

categorical: Well, I'm tall and smart

continuous: Well, I'm 180 cm and have an IQ of 126

How to deal with Categorical Variables

- Label Encoder

Categorical	Continuous
No	0
Yes	1

- One Hot Encoding

Colour	blue	red
blue	1	0
red	0	1
gray	0	0

Feature Selection

- Having too many variables could potentially cause your model to become less accurate
- Certain variables have no effect on the outcome or have a significant effect on other variables.
- Basic step of Feature Selection:
  - Use your Common and/or Business Sense(s)
- One other way to select features is to use the p-values.
  - P-values tell you how statistically significant the variable is.
- This action of omitting variables is part of stepwise regression. There are 3 ways to do this:
  - Forward Selection
  - Backward Elimination
  - Bidirectional Elimination

Feature Scaling

- Is a method used to standardize the range of independent variables or features of data.

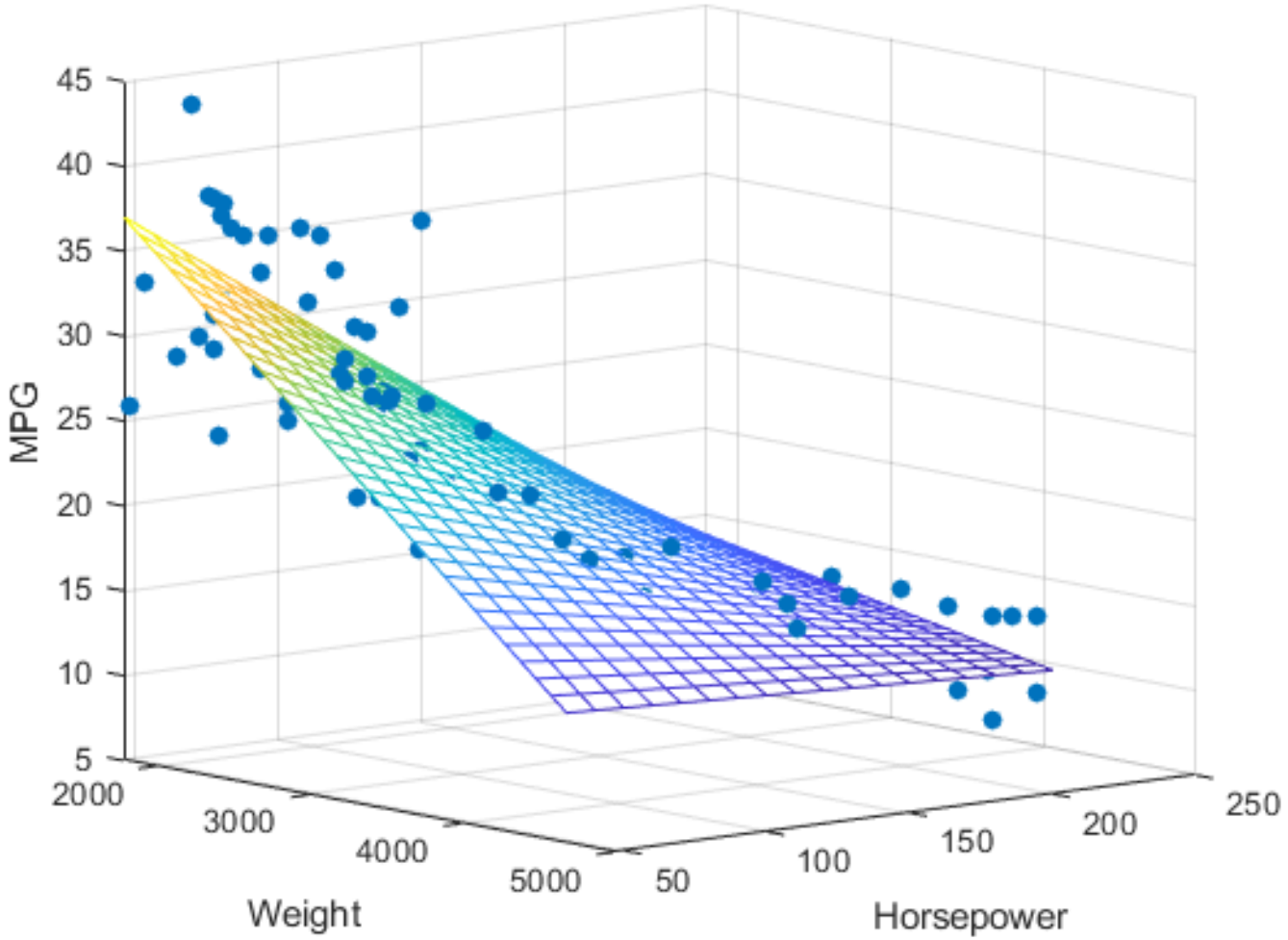
$$x' = \frac{x - \bar{x}}{\sigma}$$

- It is a step of Data Pre Processing which is applied to independent variables.
- It helps to normalize the data within a particular range.
- It also helps in speeding up the calculations in an algorithm.

Multiple Linear Regression

- Finally using the same we will find out the intercept and coefficient.

Here we will have multiple coefficient – one for each input feature.



Thanks