Linear Regression

Joe

Morning Objectives

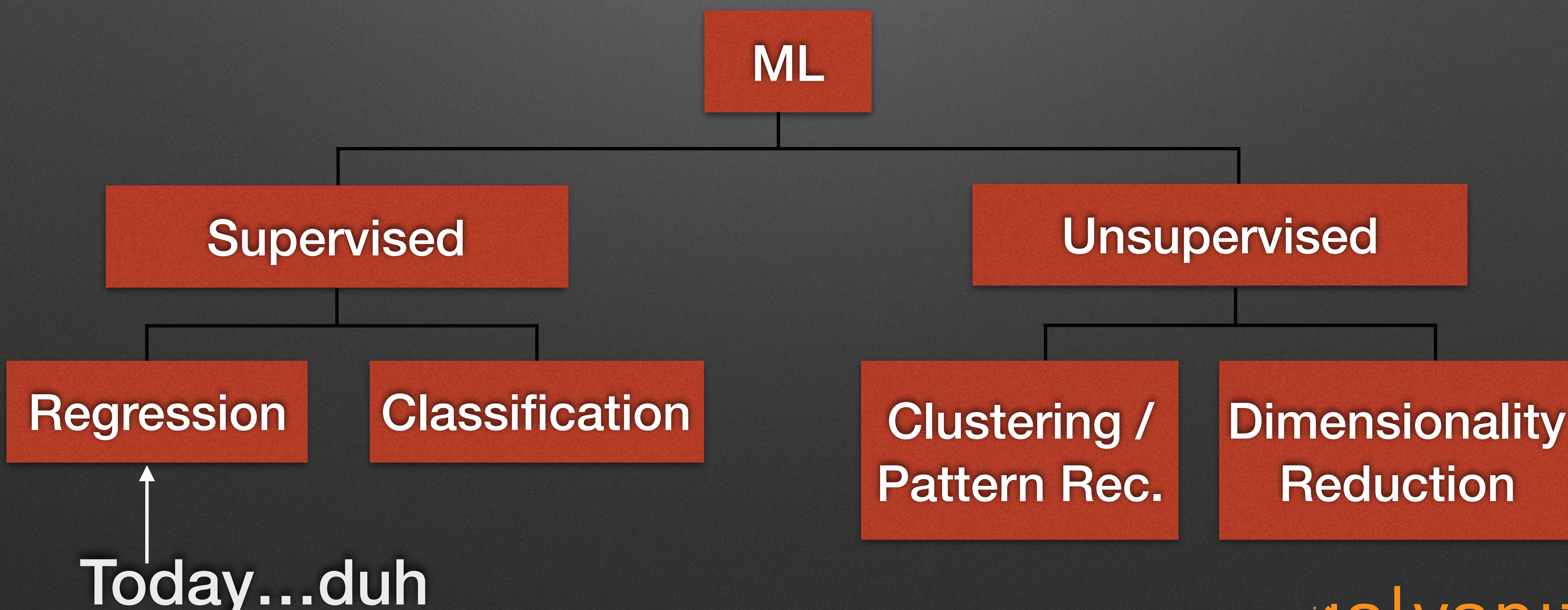
1. Fit Non-linear relationships using OLS
2. Introduce multiple linear-regression
3. Understand the implicit assumptions of linear regression, troubleshoot when these go wrong

Our First Foray Into ML

A casual definition of Machine Learning might be having a computer program do something that is not explicitly instructed by a person. We focus on predictive analytics and statistical learning, which can be roughly categorized as follows

- ML
  - Supervised
    - Regression
    - Classification
  - Unsupervised
    - Clustering / Pattern Rec.
    - Dimensionality Reduction

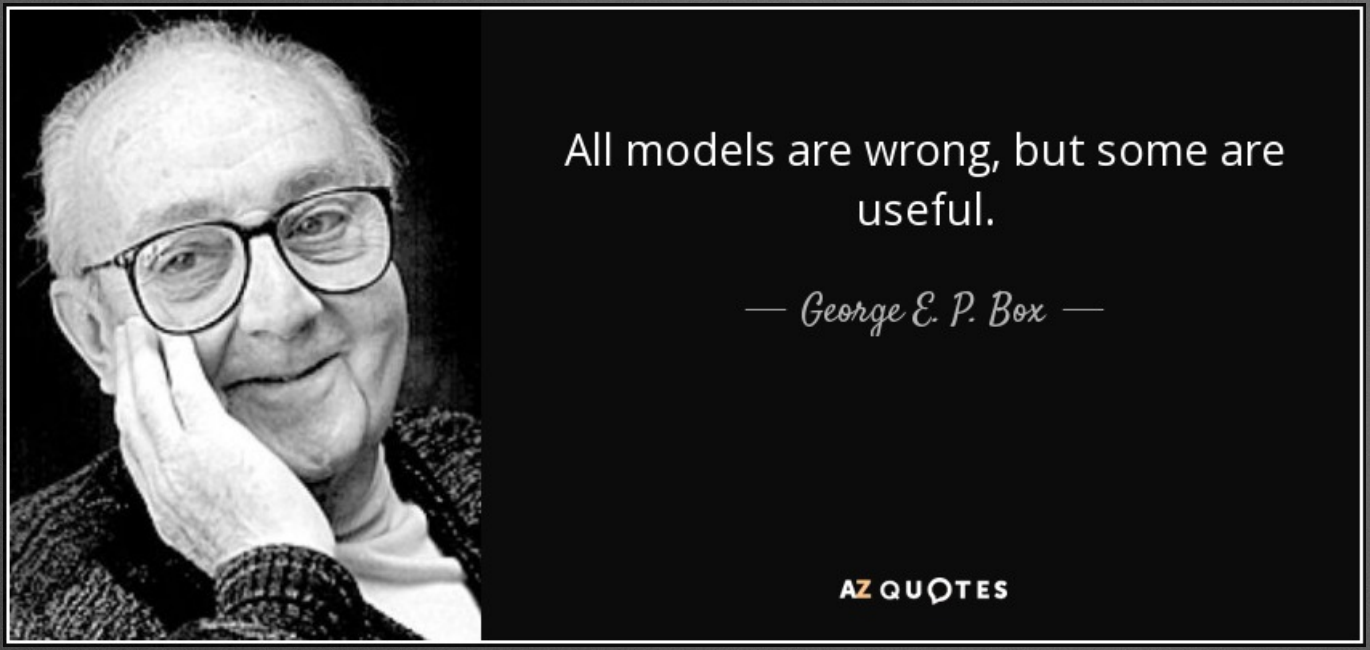
Our First Foray Into ML



Introduction & Review From Yesterday

Questions

- Using Linear Regression, we aspire to answer a series of questions:
  - Does any relationship exist between our target and feature variables?
  - If a relationship does exist, how strong is the relationship?
  - How accurately can we measure this relationship?
  - Are the relationships linear? What type of non-linear relationships should we be able to illustrate?



Yesterday, we introduced several features of OLS, what can you tell me about:

1. How we find the line of best fit?
2. The metrics we use to assess our fit?
3. How we evaluate the parameters of our model?

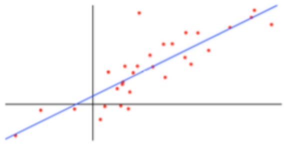
Ordinary Least Squares

- Simple linear regression assumes that a response variable (Y) has a simple relationship w.r.t. a feature (X)
  - β_0 and β_1 are unknown
  - ε is the error term, which is assumed to be i.i.d., and normally distributed

- Our model creates predictions (y-hat) based on estimated parameters ($\beta_{0\&1}$-hat)

Data

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



Model

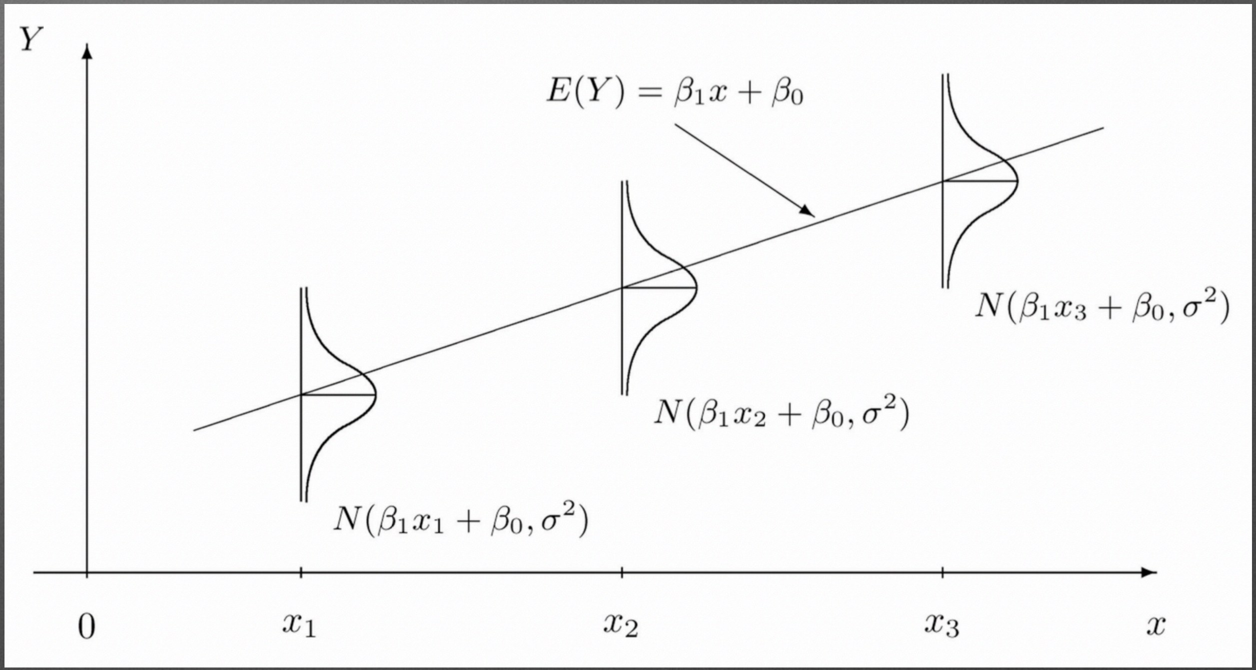
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Model Assumptions

- Recall that our assumption about the world is that the variance in the response variable is attributable to two factors

  1. The <u>linear</u> relationship between the feature and response variable

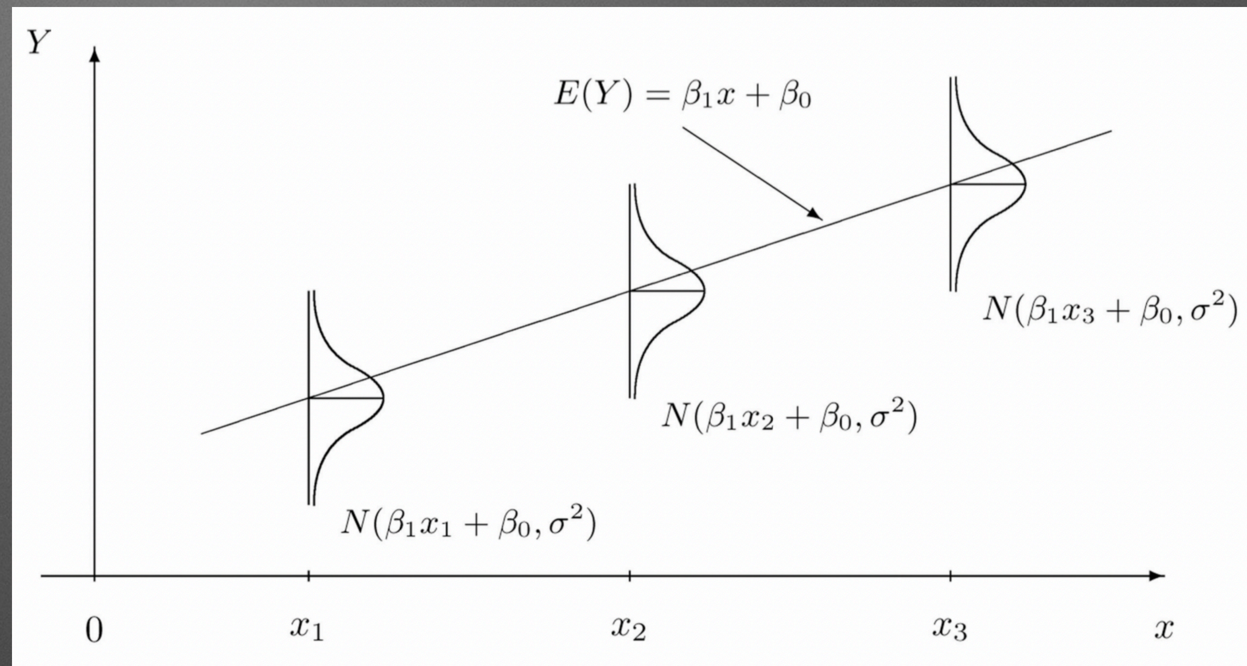
  2. Variance is either attributed to other response variables, or noise that cannot be accounted for in our data



galvanize

Model Assumptions

- The assumptions about the model are cooked into the model
  - Q: How do we find the line of best fit?
- The response feature is our prediction plus residuals
- We assert that the MSE divided by the D.O.F. is constant, for all ranges of the feature space and is normally distributed



Fitted/Predicted value $\hat{Y}_i$ Residual Variance

- $Y_i = \hat{\beta}_0 + x_i\hat{\beta}_1 + \hat{\epsilon}_i, \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{n-p-1}$ $(p = \#of\ coefficients)$

Residual

Model Accuracy

Residual Sum of Squares

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

vs

R-Squared, or "Proportion of Variance Explained"

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where $\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$

☺ Nice interpretation
Independent of scale of y

Q: What are R^2 drawbacks?

galvanize

Troubleshooting Linear Regression

See Notebook

Morning Objectives

1. Introduce multiple linear-regression
2. Understand the implicit assumptions of linear regression, troubleshoot when these go wrong