Introduction to Statistical Modeling
MSDS 598

Simple
Linear Regression

Michael Ruddy

Modeling

- “All models are wrong, but some are useful” - George Box

- Predictive Modeling
  - Based on your data, predict the value of new data points

- Descriptive Modeling
  - Based on your data, describe relationships between variables

- Difference is often subtle
  - Prediction vs. Explanation
  - How well does my model perform on new data? vs. How well does my model explain the relationship in my current data

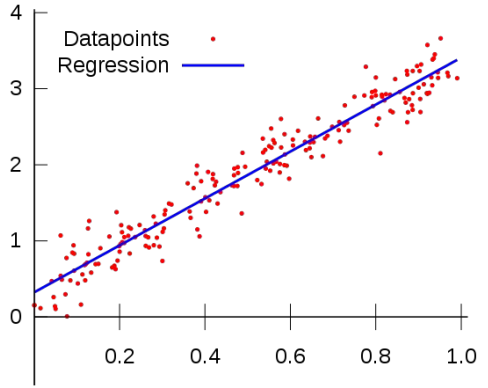
Simple Linear Regression

- *Simple* is often king!
  - Easy to interpret, explain (nice for descriptive modeling)
  - Reduces the possibility you are *overfitting* (more on this later!) to your data which causes poor generalization to new data (nice for predictive modeling)

- One of the simplest models (or relationship between variables) is a line.

- Linear Regression
  - Find the line that best "fits" the data.
  - Do these variables have a linear relationship?

Photo from here



Optimization

- **Problem**: find parameters that minimize some error function.

- Linear Regression by Least Squares:
  - Find β_0, β_1 that minimizes the Residual Sum of Squares

$$\beta_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Optimization

- **Problem**: find parameters that minimize some error function.

- Linear Regression by Least Squares:
  - Find β_0 , β_1 that minimizes the Residual Sum of Squares

- Other set-ups (often) don't always have a closed form solution!
  - Parameters are approximated using various approximation methods (such as Newton's method)

Optimization

- **Problem**: find parameters that minimize some error function.

- Linear Regression by Least Squares:
  - Find β_0 , β_1 that minimizes the Residual Sum of Squares

- Other set-ups (often) don't always have a closed form solution!
  - Parameters are approximated using various approximation methods (such as Newton's method)

- What other reasonable function might we minimize to find the "best" line?

Evaluation

- We can evaluate our model using the R^2 statistic.

$$TSS = \sum_{i=1}^{N} (y_i - \bar{y})$$

Total Sum of Squares

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

- What percentage of the variation is explained by the linear relationship between the variables?
- “How much better is the model than just using the mean?”

Evaluation

- We can evaluate our model using the R^2 statistic.

$$TSS = \sum_{i=1}^{N} (y_i - \bar{y})$$

Total Sum of Squares

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

- What percentage of the variation is explained by the linear relationship between the variables?
- “How much better is the model than just using the mean?”
- What is the range? What does it mean to be positive? zero? negative?

Loss vs. Metrics

- We often use *RSS* is **optimize** the model, while we use R-squared to **evaluate** the model.

- In this sense we can call the *RSS* function our **loss function** and the R-squared as our **metric**.

- Loss functions are used to create the model, while the metric is used for final evaluation.
    - Often metrics are more flexible and tied more explicitly to the application in mind.

Next Time

- Why squared residuals?
- Assumptions behind Linear Regression
- Linear Regression and Hypothesis Testing
- What is that Confidence Interval around your line?
- Bootstrapping
- Multiple Linear Regression!