EDA and Linear Regression

Last updated: September 12, 2016

Overview

- Exploratory Data Analysis
- Simple Linear Regression
- Multiple Linear Regression
- Assessing Fit
- Comparing Model
- Interpretation of Model Output

EDA

High level overview of a new dataset

- How are the data arranged
- What variables do we have: categorical vs. continuous
- Are there missing values
- What do the distributions look like
- How are features related

Data Cleaning/Munging

Most of the we'll have to clean the data we get

- Dirtiness - does the data make sense

Data Cleaning/Munging

Most of the we'll have to clean the data we get

- Dirtiness - does the data make sense
- Missing Data
  - What do we do?

Data Cleaning/Munging

Most of the we'll have to clean the data we get

- Dirtiness - does the data make sense
- Missing Data
  - What do we do?
  - Drop rows with missing values
  - Impute the missing values

Data Cleaning/Munging

Most of the we'll have to clean the data we get

- Dirtiness - does the data make sense
- Missing Data
    - What do we do?
    - Drop rows with missing values
    - Impute the missing values
- Convert data types

Data Cleaning/Munging

Most of the we'll have to clean the data we get

- Dirtiness - does the data make sense
- Missing Data
  - What do we do?
  - Drop rows with missing values
  - Impute the missing values
- Convert data types
- Transform data

Types of Variables

- Qualitative (Categorical)
  - Barcharts
- Quantitative (Continuous)
  - Histogram
  - Scatterplot
  - Boxplot

We want to get an idea of what our variables look like

Simple Linear Regression

The idea is to describe a linear relationship between two variables

- Fuel milage and horsepower
- Income and savings
- On-base percentage and wins
- Etc.

We're going to do that by fitting a line to our data

Linear Regression Model

The basic model is

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

- β_0 and β_1 are unknown constants that represent the intercept and slope of our line
- ε is the error term
  - $\varepsilon \sim i.i.d. N\left(0, \sigma^2\right)$
  - This is the reason not all point are on the line
- Since we don't know β_0 or β_1 we'll estimate them
- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
  - $\hat{\beta}_0, \hat{\beta}_1$ are our estimates
  - $\hat{y}$ is our prediction
- Can think of $Y|_X \sim N\left(\beta_0 + \beta_1 X, \sigma^2\right)$

Estimating Coefficients

We want to find the line that fits our data the "best"
If we define our residual as

$$e_i = y_i - \hat{y}_i$$

Then the best line is the one that minimizes the sum of the squared residuals

$$RSS = \sum_i e^2 = \sum_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

Sovling this equation gives us

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

$$\hat{\beta}_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}$$

Four Assumptions for Linear Regression

There are four assumptions that need to be met in order for our linear regression model to be valid

- Linearity
  - The relationship between X and Y is linear

Four Assumptions for Linear Regression

There are four assumptions that need to be met in order for our linear regression model to be valid

- Linearity
    - The relationship between X and Y is linear
- Constant Standard Deviation
    - The standard deviation of y does not depend on X

Four Assumptions for Linear Regression

There are four assumptions that need to be met in order for our linear regression model to be valid

- Linearity
  - The relationship between X and Y is linear
- Constant Standard Deviation
  - The standard deviation of y does not depend on X
- Independence
  - The residuals are independent of X

Four Assumptions for Linear Regression

There are four assumptions that need to be met in order for our linear regression model to be valid

- Linearity
  - The relationship between X and Y is linear
- Constant Standard Deviation
  - The standard deviation of y does not depend on X
- Independence
  - The residuals are independent of X
- Normality
  - The residuals are normally distributed

Assessing Model Fit

Once we estimate a model we can judge how well it fits our data

- Look at statistical significance of our coefficients
  - $H_0 : \beta_i = 0$
  - Get p-value and CI for $\hat{\beta}_i$
- Look at the significance of the model
  - $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$
  - This is done with an F-test
- Look at fit statistics

Significance of Coefficients

For each of our coefficent estimates we can perform a hypothesis test

- $H_0 : \beta_1 = 0$
- Test statistic is

$$\frac{\hat{\beta}_1 - 0}{SE\left(\hat{\beta}_1\right)}$$

- CI is

$$\hat{\beta}_1 \pm t_{\frac{\alpha}{2}} * SE\left(\hat{\beta}_1\right)$$

If p-value is less than α then coefficient is statitically significant

- The associated X variable has some explanatory power

Significance of Regression

For multiple regression we can test the signficance of the regression as a whole

- Is it even worth doing a regression analysis at all

We do this with a F-test

- $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k$
- Test statistic is

$$F = \frac{(TSS - RSS)/k}{RSS/(n-k-1)} \sim F_{k,n-k-1}$$

- TSS is the Total Sum of Squares $= \sum (y_i - \overline{y}_i)$
- If we reject this null, then at least one of our X variables has some explanatory power

The F-test can also be used to test the significance of a subset of our X variables

Fit Statistics

- RSS is the Residual Sum of Squares
  - The variation in y that is unexplained by X
  - Not very informative (increases with n)
- MSE is $\frac{RSS}{n-k-1}$
  - "Average" unexplained error
- R^2 is $\frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS}$
  - Proportion of variation in y explained by variation in X

Comparing Multiple Models

How do we decide which variables to include in our model?

We could pick the model with the highest R^2

- Turns out not to be such a great idea
- Why?

One solution is to look at the Adjusted R^2

- $Adj.R^2 = 1 - \left(1 - R^2\right) \frac{n-1}{n-k-1}$
- Penalizes R^2 for including extra variables

There are other ways as well

Comparing Multiple Models

How do we decide which variables to include in our model?

We could pick the model with the highest R^2

- Turns out not to be such a great idea
- Why?
- R^2 will never decrease

One solution is to look at the Adjusted R^2

- $Adj.R^2 = 1 - \left(1 - R^2\right) \frac{n-1}{n-k-1}$
- Penalizes R^2 for including extra variables

There are other ways as well

F-test

Suppose we have a model for gas milage

$$Y_{full} = \beta_0 + \beta_1 weight + \beta_2 horsepower + \beta_3 color + \beta_4 height$$

But we suspect height and color might not be important, so we can consider

$$Y_{reduced} = \beta_0 + \beta_1 weight + \beta_2 horsepower$$

we can use an F-test to test $H_0 : \beta_3 = \beta_4 = 0$

$$F = \frac{(RSS_{reduced} - RSS_{full}) / (k_{full} - k_{reduced})}{RSS_{full} / (n - k_{full} - 1)}$$

The idea is that if β_3 and β_4 don't matter, then $(RSS_{reduced} - RSS_{full})$ will be small, so F will be small

AIC and BIC

Additionally we can look at the AIC and BIC for the model

- Akaike Information Criterion $= 2k - 2ln(\mathcal{L})$
- Bayesian Information Criterion $= -2ln(\mathcal{L}) + kln(n)$
- $\mathcal{L}$ is the maximized value of the likelihood function

Both of these scores penalize models with more explanatory variables

- Question: Do we want lower or higher values of AIC/BIC?

Interpretation

Let's interpret some results

EDA Summary

EDA is a first look at the data

- Look at first few rows
- Plot variables to examine distributions/relationships
- What to do with missing data
- What else?

Linear Regression Summary

Steps in Linear Regression

- Fit model
- Examine Residuals
- Examine Results
  - Are all variables significant and make sense?
  - If not, try other models
- Examine Residuals
- Interpret Results