Outline for Linear Regresion

Data Mining

This linear regression data mining session will focus on the elementary aspects of linear regression. The Housing data and the software R will be used again. The goal will be to get comfortable recognizing the elements in regression, and modeling decision making.

Exercises and terms

This are exercises and associated vocabulary. As you do the exercises, also look at the vocabulary.

- variables: predictors, independent variables, features, variables, response and independent variable. With the housing data, discuss the data and how to label them appropriately.
- scatter plot: What is a scatter plot used for. Construct a variety of scatter plots using ggplot. Describe what the plot is indicating. Try to find the most informative plots.
- Linear model: Create a variety of linear models using the data. Plot the fit line on the model.
- Model fit: What is the model? What is the equation? What is the error.
- Mean prediction vs point prediction: What is the difference. Why is it difficult to predict individual values? Create some predictions and their confidence limits.
- Create some plots that use one continuous predictor and one categorical predictor. Add a regression line.
- Thing about how categorical fields can enter as additive or as an interaction.
- Write the model including the error term.
- Confidence intervals. Create confidence intervals as well as prediction intervals. Explain why they are different.
- Correlation. Use correlation to select a few interesting variables. Why is it difficult to keep improving the model fit by continually selecting more and more variables.
- discuss what model fit is and how it is measured.
- Why is the log of price used instead of the actual value. In what cases is this transformation a good idea. In what cases is it a bad idea?
- What would happen to the model if you had hundreds of variables?
- What is over fitting. How can it be controlled.
- Why do we fit models?: inference. Discuss inference and what that means.
- What would an automated method for selecting a model look like. What would have to be controlled for?
- What is missed when linear regression is used instead of something else?
- How are categorical variables entered mathematically into the model.
- What happens when the variance is not constant.
- What happens when points are bunched together.
- When can categorical variables be used as if they are continuous?
- How does dimensionality affect the model.
- What happens when correlated variables are added to the model?
- What are the three ways predictors can influence the model?