Problem Set 2 – Linear Regression and Extensions

2016-02-29

Loading and Exploring the Data Set

For this problem set we will analyze the data set *Boston* which is contained in the library *MASS*. This data set records the median house value (*medv*) for 506 neighbourhoods around Boston. The goal is to predict the variable *medv* using 13 predictors.

- Load the data set.
- Make yourself familiar with the data. Hint: *str()*, *names()*, *help()*
- Generate Descriptive statistics. Hint: *summary*, *mean*, *sd*, *var*, *min*, *max*, *median*, *range*, *quantile*, *fivenum*
- Plot the data, especially the outcome variable *medv* and the variable *lstat*. Hint: *plot*, *hist*, *boxplot*

Univariate Linear Regression

- Analyse the relation between *medv* and *lstat* with a linear regression. Hint: *lm()*
- Interpret the results. Hint: *summary*
- Plot the regression line in a graph with the original data points.
- What is the predicted value of *medv* for a region with a *lstat* of 32?

Multivariate Linear Regression

- Fit now a multivariate regression.
- Interpret the results, in particular with a focus on the variable *lstat*.
- Fit a more complex model, e.g. considering interaction effects and higher order polyomials.

Regression Splines

Now we consider again the relation between *lstat* and *medv*. Hint: library *splines*

- Fit a cubic regression spline to the data!
- Plot the fitted line!
- Experiment with different spline specifications! Hint: options *knots* and *df*
- Compare the different specifications!

Smoothing Splines