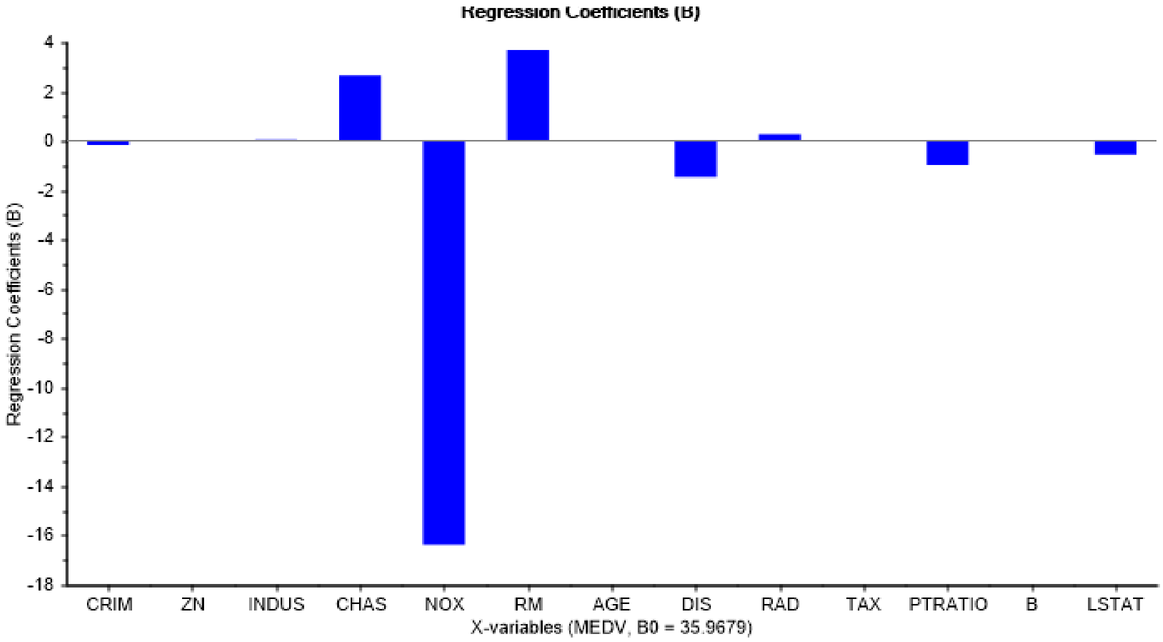
Assignment 6 – Unscrambler part

We see INDUS and NOX are highly positively correlated which is expected. More industry can mean larger emission of NOX gasses. Somewhat surprisingly NOX and AGE are highly positively correlated.

The predicted vs reference plot shows that many of the predicted data points are close to the reference value, meaning that the model works well. We also see the residuals are around zero, with some outliers, and most are within -10 to 10 range. Considering the different scale on the variables this could be acceptable. In the ANOVA table we have Rsquared of 0.74. This is decent, although ideally we would like to be closer to 1.

The p-values for INDUS and AGE are a lot larger than the others, and are way above normal limits like 0.05 or 0.1. This means that the effect of these variables are not certain.



The regression coefficients plot tells us that the most important variables for house value is the number of rooms RM and the CHAS variable (I don't understand what this is supposed to represent), and if the area is highly polluted by NOX gasses the price drops significantly. DIS pulls down some which can make some sense because houses close to employment centers are probably not inhabited by the richest people. RAD pulls up a little bit which also makes some sense since access to to highways makes commutes easier and is a bonus for most people.

Recalculating with only CHAS, NOX and RM gave a much smaller R square of 0.52 meaning that only these variables are not enough to have a good prediction. The residuals are a little more spread out than before.

The RMSE of the training set was 4.72 and for the testing set it was 4.88. They are similar which indicates that the models predicts the values well.