Multivariat øving 6, del 2

Siawash Naqibi

April 2, 2021

Correlation matrix

We can see that this is a correlaiton matrix. The diagonal elements are one because that represents the correlation between a variable with itself which is a perfect correlation of 1. Evertyhing above |corr(A,B)| > 0.7 can be interpreted as a strong correlation between variables A and B.
From this we can see that NOX and INDUS are strongly correlated. We can also see that TAX and INDUS are strongly correlated. AGE and NOX are strongly correlated. AGE and DIS are strongly correlated. TAX and RAD are strongly correlated.
Som of the correlation doesn't make much sense. Why is NOX and AGE correlated? And I don't see why TAX and INDUS would be correlated either.
These are some of the variables that are correlated with each other. Looking at the correlation matrix from a broader perspective, most of the variables have a strong correlation with at least one other variable.

interpret ANOVA table, predicted vs. reference, residuals

Lets look at the ANOVA table first. Since we have 337 examples and we are estimating the mean, that means we will have 336 degrees of freedom. We are estimating over 13 feautures, that leaves us with 323 degrees of freedom for error.
The F ratio is the ratio of two mean square values. If the null hypothesis is true then you expect to see an F ratio close to 1. On the other hand, a large F ratio means that the variation among the group means is more than what you expect to see by chance. We have an F ratio of 174, therefore there is very little chance to see the variation of the data to have appeared just by chance.
The p value is the probability of obtaining results at least as extreme as the observed results. We have a p value of 0.00000, therefore the is in practical sense zero probability of the results occurring just by chance.

Lets look at the residuals table now. I cannot see a structure between the Y-residuals an Y-predicted which is good because that means that our linear model accurately represents our data.

Lets look at the predicted vs. reference plot. Here we can see that our model is able to make accurate predictions over the training data and we can see that we get RMSE error of about 3.

p values and correlation table

The p value for a feature tells us how well that feature rejects the null hypothesis, meaning that there is no relationship between the feature and the target. Assuming there is no relationship between the feature and target, what is the probability of observing the data that we have. A low p value means that that there is very little probability of observing the data that you have observed given that there is no relationship between the variables.
The correlation is a measure of how strongly two variables are coupled together. In a sense the p value and correlation is connected, in the sense that if one has a strong correlation between two variables, then one will also experience a lower p value.
In our case we see that almost all of our variables have very low p values

regression coefficients

The regression coefficients shows us that we can make the MLR model with a constant B_0 = -12.9990 plus 9.24*RM plus -9.15*NOX and so on. the highest weights are the weights for RM and NOX. Since the variables are not scaled, their relative sizes cannot be used as a comparison between variables. The variable NOX has also high correlation with several other variables such as INDUS and AGE and therefore its coefficient can actually not be trusted.

make new model with reduced variables

We reduced from 14 to 9 variables now and took out the variables that seemed to have the highest correlation with other variables. The variables taken out are NOX, RAD, TAX, LSTAT and INDUS.

compare the results between the two models

We can see that although we have greatly reduced the amount of variables, there new model is still doing pretty well. The p value is still at 0.0000, The F ratio has increased from 174 to 256 wich is drastically better. THe RMSE value increased slightly from about 3 to 3.27.

prediction

I have made prediction on the test set for both my MLR models. One can clearly see that the MLR model with more variables do have better prediction, however this is not that much better than the MLR with reduced variables. I am not however able to find the RMSE values for the predictions.