EMAT30007 Applied Statistics

Lab 7: Linear Models (Multiple Linear Regression)

Nikolai Bode

This lab has two parts. In the first part, you will work through a multiple linear regression example including model selection. The second half is a group exercise, where you will work in groups on different data sets, performing model selection. Each group will then present their approach to the rest of the class.

(1) Multiple Linear Regression Example

In this part of the lab we'll work through an example in some detail. We will use one of the data sets we already encountered in the last lab, but we will consider more variables as possible predictors. As a reminder, the data was collected to investigate the energy use of appliances in low-energy buildings. The data set is from one house. Readings were taken every 10 minutes for about 4.5 months. Reference for the data:

[Candanedo, L. M., Feldheim, V., Deramaix, D. (2017) "Data driven prediction models of energy use of appliances in a low-energy house". Energy and Buildings, Volume 140, 1 April 2017, Pages 81-97]

You can find the data in the file 'APPLIANCES2.csv'.

In this file, the response 'Appliances' is included alongside four variables that have been measured independently of the energy usage of the appliances (outside temperature, pressure outside, humidity outside, windspeed outside). The task is to establish which out of these variables (if any) help to predict the energy usage of the appliances in the house.

1.1 Preliminary analysis

It is always a good idea to have a look at the raw data to see if this shows some trends. Read in the data and plot scatterplots of the possible predictors against the energy usage of appliances.

```
%% read data:
delimiterIn = ',';
A = importdata('APPLIANCES2.csv',delimiterIn);

%% ... then plot scatterplots of data
```

With this many data points, it can be hard to tell what's going on in plots of raw data. So we proceed to statistical analysis using a multiple linear regression model.

1.2. Multiple linear regression

When performing statistical model fitting on data that includes more than one possible predictor, there are different approaches. We could either start with minimal models that only consider one predictor at a time or we could fit a full model that includes all available predictors. Either approach is valid and which one to use typically depends on the type of analysis. For an exploratory analysis as we perform it here, it is often convenient to start with a full model.

Fit a full model to the data. The code in Matlab for multiple linear regression is a straighforward extension of what we do for simple linear regression. In `fitlm()`, simply add more predictors to the model, separating them with a '+'.

```
% fit a full model that includes all predictors to the data
```

This model fit tells us quite a few things already. Consider r-squared. This is quite low (you should find `R-squared=0.0281`), suggesting that only a small proportion of the variance in the responce (appliance energy usage) is explained by the model. However, the F-test suggests that our model is better than a constant model (i.e. one that only includes the intercept, `p-value=1.56e-120`). The t-tests are specific to the parameters linked with the different variables. Before we consider these in more detail, look at the residual plots for this model fit (use the same plots as in lab 6).

```
% Check residual plots for model m1:
```

These plots look pretty catastrophic - very similar to what we saw in lab 6. Really, at this stage, we should reconsider fundamental aspects of our modelling approach or seek additional variables that might explain the features we see in the data.

However, to illustrate model selection, we will continue with this type of model for the data.

1.3. Model selection

Have another look at the table produced by the model fit above. Remember that the t-tests investigate the null hypothesis that the parameter associated with a variable in the regression model equals zero. For temperature it looks like we cannot reject the null hypothesis (interesting, considering that in lab 6 temperature had an effect...!!!).

So let's fit a new model to the data that does not include the predictor temperature.

```
% fit a reduced model excluding temperature
```

We can now assess in two ways whether this model is an improvement compared to the full model we fitted earlier.

First, we can compare quality measures like the AIC or BIC for the models (remember, lower values are better and also recall my dislike of r-squared for model comparison!). You can find quality measures for a fitted model 'm' in Matlab using the command 'm.ModelCriterion'.

```
% e.g. use
```

```
m1.ModelCriterion
```

The AIC and certainly the BIC should be somewhat smaller for the reduced model we fitted.

Second, we can use a Likelihood-ratio test, to formally test the null hypothesis that the parameter associated with temperature is equal to zero (notice how this tests something similar to the t-test, but it requires another model to be fitted).

In Matlab, the command is `lratiotest(logL_full, logL_reduced, df)`, where `logL_full` and `logL_reduced` are the log-likelihood for the full and reduced model, respectively (access these for a model `m` with the command `m.LogLikelihood`). 'df' are the degrees of freedom of the test. In practice this is the difference in the number of parameters between the full and reduced model (i.e. here 'df=1'). Perform the Likelihood-ratio test, making sure you output a p-value (check Matlab help pages).

```
% e.g. use
[h pvalue] = lratiotest(m1.LogLikelihood,m2.LogLikelihood,1)
```

You should find a large p-value (`p-value=0.9755`), suggesting that we cannot reject the null hypothesis and thus that temperature does not need to be included in our model.

Now repeat this analysis using the Likelihood-ratio test for the variable 'pressure'. You should find a low p-value and reject the null hypothesis.

Another important aspect of the data to check is whether any of the predictors are correlated. If two predictors are highly correlated, then it may not make sense to include them as separate predictors into a model. In addition, high correlations between predictors can lead to issues with model fitting and with performing hypothesis tests on model fits (see next lecture).

Find a function in Matlab that allows you to compute the correlation between the variables included in the model.

```
% e.g. use corrcoef():
corrcoef(pressure,humidity)
```

Apart from a slightly negative correlation between temperature and humidity, you shouldn't find any issues.

1.4. Fit multiple linear regression model using matrices

Recall the matrix notation for Linear model from lectures. We also saw a convenient way for estimating the parameters of a Linear models that used this matrix notation. Use this approach to recover the parameter estimates of the reduced model you fitted above.

(2) Group exercise

The second part of the lab is a greoup excercise. Each group will be assigned one of the following four data sets. Your task is to perform an exploratory analysis of this data and to use model selection to formulate a Linear Model for the data that only includes relevant parameters.

The datasets are illustrative and have been created for this exercise based on data sets published online (http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/frames/frame.html.)

Antelope data

The data set 'anteolpe.csv' is based on a data recording the population development of antelopes in the Thunder Basin Grasslands in the USA. The data set includes the fawn count (in hundreds), the adult population (in hundreds) and information on the annual precipitation (inches). Your task is to investigate how the fawn count depends on the other two variables.

Blood pressure data

This data ('blood_pressure.csv') looks at how peoples' systolic blood pressure depends on their age, weight and tea consumption (litres per day). Investigate whether age, weight or tea consumption have an effect on blood pressure. Please do not adjust your lifestyle based on this data, it is for illustration only ;-) .

Exam scores data

The data set on exam scores ('exam_scores.csv') is a hypothetical record of students' performance on a university module. It includes the results from three exams and the final exam. Your task is to investigate if any of the three exams predict students' performance in the final exam.

Hollywood movie data

The file 'hollywood_movies.csv' investigates revenues from book sales after the release of hollywood movies. Three possible predictors are included in the data: first year box office receipts, production costs and promotional costs. Investiate which out of these three variables predicts book sales.