EMAT30007 Applied Statistics

Lab 8: Model Building

Nikolai Bode

This lab investigates model buiding for Linear Models (LMs) in statistics. In the first part, we will look a how to construct models with different types of predictors and in the second part, we will explore common pitfalls in statistical analysis with LMs.

There is no data associated with this lab. Instead, you will simulate your own data. The idea behind this is to develop your understanding of how certain features in data arise.

(1) Different types of predictors

In this part of the lab, we will work through different types of predictors that can be used in constructing LMs. For each of the predictors, your task is to:

- simulate data including the response and predictor(s).

- fit a LM with only quantitative predictors and perform model checking using residual plots to see what goes wrong when the predictors in your data are not quantitative.

- fit an appropriate LM to this data to verify the data you constructed behaves as it should do.

- perform model checking using residual plots we have encountered before.

1.1. Quantitative predictor

This is the predictor type we have been looking at in lectures 6 and 7. Simulate data for one quantitative predictor. To do so, think about the structure of LMs, $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma)$.

```
% One solution:

x = random('Uniform',0,10,[1000,1]); % predictor
beta0 = 2.3; % intercept
beta1 = 1.5; % effect of predictor
epsilon = random('Normal',0,2.1,[1000,1]); % error vector N(0,2.1)
% define response
y = beta0 + beta1.*x + epsilon;

%.. then you can fit a Linear Model to this simulated data...
```

1.2. Qualitative predictor

Construct data for a response with one quantitative and one qualitative predictor. You can simply extend the model from section 1.1 and you can choose how many levels you want the qualitative predictor to have. Some tips:

- the first level of the qualitative predictor is often absorbed into the intercept.

- parameters for qualitative predictor levels measure effects relative to the intercept.

- use dummy variables!

- Use the structure of LMs with qualitative predictors when simulating your data.

- before fitting a LM with qualitative predictors in Matlab, you need to let it know that some variables are qualitative. If `data` is a table with data and `x` is one entry, use `data.x = nominal(data.x)` in Matlab.

1.3. Interaction terms

Construct data with one quantitative and one qualitative predictor (2 levels). Include an interaction term between the two predictors so that the slope of the quantitative predictor changes for the levels of the qualitative predictor (cf try to create a plot like the one in the lecture notes). Tip: in Matlab, in `fitlm()`, use `*` to include an interaction between two predictors.

1.4. Polynomials of predictors

Now let's construct data with a response that is a polynomial of one quantitative predictor, e.g. a second-order polynomial: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$.

1.5. Data transformations

Consider a response that is a function of one quantitative predictor. However, instead of being a linear function of the predictor, the response is a linear function of the logarithm of the predictor, e.g. $Y_i = \beta_0 + \beta_1 log(x_i^2) + \varepsilon_i$.

(2) Common pitfalls

In the second part of the lab, we will look at a few common pitfalls in statistical analysis using LMs. In the following, construct data sets that lead to the specified pitfalls. Similary to the first part, your task is to:

- simulate data including the response and predictor(s).

- fit a LM to see what goes wrong for the pitfall you have implemented.

- perform model checking using residual plots we have encountered before.

2.1. Multicollinearity

We have already seen an example for multicollinearity in lab 7. So we will not repeat this here.

2.2. Violated model assumptions

Perhaps the most common issue with LM analyses is that the assumptions of the model are not satisfied. Typically, this relates to the error distribution, which we assume to be normal, with constant variance and comprised of independent samples in LM. The following examples show what the residual plots of models fitted to data for which certain assumtions do not hold look like. In general, when model assumptions are violated, we cannot use the standard framework of hypothesis tests and model selection.

2.2.1. Non-normal error distribution

Construct data for which the errors do not follow a norma distribution (e.g. exponential or gamma distributed).

2.2.2. Non-constant variance of predictors

Construct data for which the errors follow a normal distribution, but the

variance of this distribution is not constant (heteroscedasticity).

2.2.3. Data points not independent

We won't go through this example here. Data points that are not independent arise commonly in time-series data (data points close in time are strongly correlated).

2.3. Extrapolating beyond model scope

As discussed in lectures, we need to be careful when interpreting model fits beyond the range of the data the model was fitted to. A nice way to illustrate this issue, is to construct a response that is a non-linear function of a predictor which can be approximated linearly for part of the range of the predictor

(e.g. consider $y = \sqrt{x}$).

2.4. Overfitting

We won't go through an example here. The extreme case of overfitting occurs when we fit a model with as many parameters as there are data points.