SF2930 VT25: Regression Analysis
Notes on model building
For questions or comments, please contact Isaac Ren (isaacren@kth.se) and/or Hanqing Xiang (hanqingx@kth.se).

These notes are based on [MPV]. Note that this is not a replacement for the instructions for Project 1, but simply an aid.

1. Model building [MPV, § 10.3]

[MPV] has a pretty good summary of a general method for model building, see figure 10.11. Here is a more detailed version of it:

1. Fit the full model, that is, the model with all possible regressors. In `R`, this can be done by writing `lm(y ~ ., data = d)`, where `y` is the name of the response column and `d` is the name of the data table.
2. Perform a thorough analysis of this model. This includes a residual analysis and an influence analysis.

   - Evaluate outliers and influential points. Remember that, in order to remove a data point, you should have external justification: that is, knowledge about the actual data, not just its influence measures. For example, an error in the data or a data point that falls outside the scope of the model.
   - If the residuals are not normally distributed, transform the data accordingly.
   - If there are nonlinear relationships identified in residual plots or added-variable plots, transform the data accordingly.

   You may use methods such as Box-Cox or Box-Tidwell to transform data. You may also make some judgements by hand. There may not be an exact answer.
3. If you have transformed the data, refit the full model and repeat the previous steps.
4. Try regression with every subset of regressors. If this is not possible, use a stepwise selection technique (forward or backward). You should end up with a few good models.

   - Usually, the evaluation of models is done with cross-validation.
   - Other techniques for model selection exist: ridge regression and lasso regression.

- Check for multicollinearity.

5. Compare the best models, not only in terms of accuracy, but also in other terms, such as simplicity.

References

[MPV] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to Linear Regression Analysis*. Wiley Series in Probability and Statistics. Wiley, 2012.