Advanced Topics in Applied Statistical Analysis
Haviland Wright, Alexandrina Goeva and Juliane Manitz — Spring 2016

Model Selection and Regularization

Exercise 1 (Conceptual: Model Selection Criteria; to be graded in detail)

(a) It has been stated that for Gaussian models, the AIC statistic is equivalent to Mallow's C_p. Assume the variance to be known $\sigma_\epsilon^2 = \hat{\sigma}_\epsilon^2$ and prove the equality.

(b) Derive the representation of the

$$BIC = \frac{1}{n}(RSS + \log(n) \cdot \hat{\sigma}^2)$$

in dependence of the log-Likelihood. Explain why this representation provides a more general applicability.

Exercise 2 (Conceptual: Lasso Regression)

Exercise 3 (p. 260): Discuss properties of lasso regression.

Exercise 3 (Applied: Model Comparison)

Exercise 9 (p. 263): Compare different methods (least squares, ridge, lasso, PCR, PLS) for `College` data. Perform additionally hybrid stepwise variable selection as well as elastic net.

Exercise 4 (Applied: Bootstrap and Lasso; to be graded in detail)

Here we use the bootstrap as the basis for inference with the lasso.

(a) For the `College` data, apply the bootstrap to estimate the standard errors of the estimated lasso coefficients. Use the nonparametric bootstrap, sampling features and outcome values (x_i, y_i) with replacement from the observed data. Keep the bound s fixed at its estimated value from the original lasso fit.

(b) Repeat part (a), but now re-estimate $\hat{\lambda}$ for each bootstrap replication. Compare the results to those in part (a).

Exercise 5 (Applied: Model Comparison)

Exercise 11 (p. 264): Compare different methods to predict per capita crime rate in the `Boston` data set.

Exercise 6 (Applied: Model Selection and GAMs)

Consider the `bodyfat` data from `TH.data` package, which contains the body fat measurements and several anthropometric measurements of 71 healthy female subjects. Predict the body fat in variable `DEXfat` by the given predictors in the data set.

(a) Get an overview of the data and do simple descriptive analysis including a correlation analysis and a scatterplot matrix.

(b) Fit a linear model assuming normal errors. Are all potential covariates informative? Check the results against a model that underwent AIC-based variable selection.

(c) Check the model assumptions of the final model and give short arguments whether they are fullfilled or not.

(d) Fit a generalized additive model using the results you obtained so far. Use the function `gam()` in the `mgcv` package and fit cubic spline effects using the option `bs='cr'`.

(e) Check again the model assumptions of your final model and compare the results to the ones obtained in the linear model.

This homework is due at the beginning of the discussion section on March 1, 2016 at 3.30pm.

All references refer to the textbook: James, Witten, Hastie and Tibshirani (2013). *An Introduction to Statistical Learning with Applications in R.* Springer. Available online: `http://www-bcf.usc.edu/~gareth/ISL/`.
