Exercise 1

Advanced Methods for Regression and Classification

October 24, 2024

Load the data `College` from the package `ISLR`. This means that you first need to install the package with

```
install.packages("ISLR")
```

and then load the data with

```
data(College,package="ISLR")
```

Look at `?College` and at `str(College)` for more detailed information. Remove (if necessary) all observations which contain missings by using the command `na.omit()`.

Our goal is to find a linear regression model which allows to predict the variable `Apps`, i.e. the number of applications received, using the remaining variables except of the variables `Accept` and `Enroll`.

For the following tasks, split the data randomly into training and test data (about 2/3 and 1/3), build the model with the training data, and evaluate the model using the RMSE as a criterion. The RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2},$$

where N is the number of observations to be considered (e.g. only training data, or only test data), y_i are the values of the response variable, and $\hat{y}_i$ are the estimated values of the response. You can report the RMSE always for the training and the test data.

1. Look first at your data. Is any preprocessing necessary or useful? Argue why a log-transformation of the response variable can be useful. Continue with `log(Apps)` as the response.

2. *Full model*: Estimate the full regression model and interpret the results.

   (a) For that purpose, apply the function `lm()` to compute the estimator – for details see course notes. Interpret the outcome of `summary(res)`, where `res` is the output from the `lm()` function. Which variables contribute to explaining the response variable? Look at diagnostics plots with `plot(res)`. Are the model assumptions fulfilled?

   (b) Now we try to manually compute the LS coefficients, in the same way as `lm()`. Thus, replace from the above command `lm()` by `model.matrix()`. This gives you the matrix $\mathbf{X}$ as it is used to estimate the regression coefficients. Now apply the formula to compute the LS estimator. You can do matrix multiplication in R by `%*%`, and the inverse of a matrix is computed with `solve()`. How is R handling binary variables (`Private`), and how can you interpret the corresponding regression coefficient? Compare the resulting coefficients with those obtained from `lm()`. Do you get the same result?

   (c) Compare graphically the observed and the predicted values of the response variable – once only for the training data, and once for the test data. What do you think about the prediction performance of your model?

(d) Compute the RMSE separately for training and test data, and compare the values. What do you conclude?

3. *Reduced model*: Exclude all input variables from the model which were not significant in 2(a), and compute the LS-estimator.

   (a) Are now all input variables significant in the model? Why is this not to be expected in general?

   (b) Visualize the fit and the prediction from the new model, see 2(c).

   (c) Compute the RMSE for the new model, see 2(d). What would we expect?

   (d) Compare the two models with `anova()`. What can you conclude?

4. Perform variable selection based on stepwise regression, using the function `step()`, see help file and course notes. Perform both, forward selection (start from the empty model) and backward selection (start from the full model). Compare the resulting models with the RMSE, and with plots of response versus predicted values.