Exercises Week 5

Marjolein Fokkema

First, read in the student data from the file "student-mat.csv" as follows:

```
student_full <- read.csv2("student-mat.csv")
```

Use mathematics achievement at moment 3 (`G3`) as the response and all other variables as predictors.

Select a training dataset of 300 observations, and a test dataset of 95 observations.

1. Best subset, forward and backward selection

Use function `regsubset()` from package `leaps` to fit the models to the training data. Consult the documentation (`?regsubsets`), more specifically the **Arguments** and **Values** subsections, to see how different variable selection approaches can be specified.

- Perform best subset regression (take subsets of at most 12 variables) and determine the optimal model by means of Mallows Cp, BIC and adjusted R-squared.
- Use all variables and perform forward and backward selection (go up to 33 predictors).
- Use BIC to select the optimal model for best subset, forward stepwise and backward selection. Did the three methods retain the same or different variable sets? What is the strongest predictor of math achievement at moment 3?
- Generate predictions for the test observations. Since there is no `predict` method for `regsubsets`, so you have to multiply the predictor variable values with the estimated coefficients. You can extract the coefficients using the `coef` method.
- Compute mean squared error (MSE) for the test observations.

2. Ridge, Lasso, Elastic Net

To predict mathematics achievement at moment 3, now fit a Ridge, Lasso, and Elastic Net regression model. Again, use only the training data to fit the model. Determine the optimal value of λ by means of 10-fold cross-validation.

Use function `cv.glmnet` from package `glmnet`. For Elastic Net, you can choose any (set of) values of α (often-tried values are 0.25, 0.5 or 0.75). Note that `cv.glmnet` requires specification of a matrix `x` for the predictors and a vector `y` with responses (it does not accept a `data.frame`), which you can create as follows:

```
x <- model.matrix(G3 ~ ., data = ...)
```

- How many variables does each of the three methods retain? Print the fitted models, and apply the `plot` method to inspect the results.

- Apply the `coef` method to inspect which variables were retained, and what their estimated coefficients are.
- Generate predictions for the test observations for each of the three methods, using the `"lambda.min"` criterion. You can use `predict` for this, make sure to specify arguments `newx` and `s`.
- Compute MSE for the test observations. Which method performed best?

3. Relaxed lasso

Fit a relaxed Lasso model, again using function `cv.glmnet`, but now also specify `relax = TRUE`. Print the fitted model, and use `plot` to inspect the result. Which were the optimal values of λ and γ? Which variables were retained? Again, compute predictions and compare with the models fitted earlier.