Exercises Linear Model Selection & Regularization

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: 1) Multiply the predictor variable values with the estimated coefficients yourself. You can extract the coefficients using the `coef` method. When factor variables with > 2 levels have been selected, you have to manually set up dummy coded indicator variables, and it may be more convenient to: 2) Refit a model with function `lm()`, using the selected variables only, and use the `predict` method. (Indeed, generating predictions like this is tedious.)
- 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:

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

- How many variables did 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.

4. Predicting Depressive Disorder using Questionnaire Items

```r
train <- readRDS("masq_train.Rda")
test <- readRDS("masq_test.Rda")
```

We use a real dataset on prediction of depressive disorder for this exercise. Possible predictor variables are item scores on the Mood and Anxiety Symptom Questionnaire (`MASQ01 - MASQ90`) and socio-demographic characteristics (`GENDER`, `Leeftijd`, `DEMOG1 - DEMOG8`). The response variable is `D_DEPDYS`, whether the respondent has a current depressive or dysthymic disorder (0 = no, 1 = yes), as evaluated by a mental-health professional through a structured interview.

In this assignment, you will fit a range of penalized logistic regression models on the training dataset and compare their performance on a test dataset. You will do this using the cross-validation + test set approach.

- Pick three candidate procedures from ridge, elastic net (with any $0 \leq \alpha \leq 1$), lasso, relaxed lasso. Ideally, this should be done by considering the training set (e.g., multicollinearity), and/or thinking about what would constitute a useful result (e.g., would a (non)-sparse solution be useful for decision making in clinical practice?), but you can also just make a random choice.
- Select the most accurate model through 10-fold cross-validation on the training set.
- Estimate the misclassification rate (MCR) and the Brier score on the test set.
    - The Brier score is the mean squared error of predicted probabilities: $\frac{(y-\hat{p})^2}{n}$
    - By default, `predict.cv.glmnet` returns predicted values on the scale of the linear predictor ($\hat{\eta}$), so to obtain predicted probabilities, compute $\hat{p} = \frac{1}{1+e^{-\hat{\eta}}}$.
- Use the `coef` method to extract the selected variables and their coefficients from the best-performing model. From which MASQ subscale were most items selected?
    - Anhedonic Depression: Items 1, 14, 18, 21, 23, 26, 27, 30, 33, 35, 36, 39, 40, 44, 49, 53, 58, 66, 72, 78, 86 and 89.

- Anxious Arousal: Items 3, 19, 25, 45, 48, 52, 55, 57, 61, 67, 69, 73, 75, 79, 85, 87 and 88.
- General Distress Depression: Items 6, 8, 10, 13, 16, 22, 24, 42, 47, 56, 64 and 74.
- General Distress Anxiety: Items 2, 9, 12, 15, 20, 59, 63, 65, 77, 81 and 82.
- General Distress Mixed: Items 4, 5, 17, 29, 31, 34, 37, 50, 51, 70, 76, 80, 83, 84 and 90.