HDS Exercise set 3.

YOUR NAME (STUDENT NUMBER)

Return by **10.15** o'clock on **19.11.2019** to the Moodle area of the course. Return the final file in pdf format with name "HDS3_yourname.pdf".

Problem 1.

Read in the data set "HDS_ex3.txt". It has columns `x`, `y` and `z` and 110 rows. Split the data into training set of rows 1,. . . ,10 and test set of rows 11,. . . ,110.

Fit the following three models to the training data via `lm` function.

(1) $y = \beta_0 + x\beta_1 + \varepsilon$

(2) $y = \beta_0 + x\beta_1 + x^2\beta_2 + \varepsilon$

(3) $y = \beta_0 + \sum_{i=1}^{9} x^i\beta_i + \varepsilon$,

that is, the 1st, 2nd and 9th degree polynomials. (You can use `y ~ poly(x, 9, raw = T)` to fit the 9th degree polynomial in `lm`.)

Plot the training data points and draw the curves fitted by each of the three models in the range of training data x-values. Use different colors for different models. (You can make a suitable grid of x-values and then use `predict()` to get the fitted values from each model at those grid points.)

Compute and print out the training MSE and test MSE given by these three models (that were trained only using training data). How do training MSEs compare to test MSEs?

Plot the test data points and draw the curves predicted by each of the three models (trained in training data) in the range of test data x-values. Comment how well each curve fits the test data and how you could have predicted this pattern already from how the models behaved on the small set of 10 training data points.

Which of the three models has the largest bias, which has the largest variance and which finds the best bias-variance tradeoff?

Problem 2.

Let's work with `Boston` data (call `library(MASS)` and you will have `Boston` variable defined). (In all models use the intercept term.)

Fit linear regression model of `medv` on all other 13 variables of the Boston data set. (Use `.` notation in formula rather than typing all variable names.) Print out summary and compute AIC and BIC of the model.

Fit regression of `medv` on all other variables except `age`. (Use `.` notation together with `-` notation in formula.) Compute AIC and BIC.

Fit regression of `medv` on all other variables except `rm`. Compute AIC and BIC.

Which of the three models is to be preferred according to AIC and BIC?

Use your AIC values to compute the likelihood ratio test statistic between model without `age` and the full model. Look for the P-value of likelihood ratio test for variable `age` by comparing the LRT statistic to the appropriate chi-square distribution. Does this P-value agree with P-value given by `summary()` of the full model fitted above?

Do similar LRT calculation but now using BIC values and between the model that misses variable `rm` and the full model.

Problem 3.

Let's continue with `Boston` data set, and still keep on predicting `medv`. Let's split the data to 350 training samples and 156 testing samples.

(a) Use row indexes `tr.ind = 1:350` for training and the rest for testing. Fit the full model with all 13 predictors and intercept to the training data and then use the same model in test data. Print out training MSE and test MSE. What do you expect might be happening based on a comparison between training MSE and test MSE?

(b) Could you have predicted the high test MSE by cross-validation? Apply 10-fold cross-validation (CV) within training data to the model of part (a). You can use `cv.glm()` from `boot` package whence you need to refit the model using `glm()` unless you already did part (a) with `glm`. Use MSE as the measure of fit in CV. After seeing CV MSE in training data and comparing that to test MSE, what do you now assume is happening?

(c) Read in training indexes from file "HDS_ex3.3_tr.txt". It has 350 indexes that were randomly sampled among 1:506. Repeat the fitting of the full model, CV and training and test MSE calculations from parts (a-b) using this new training/testing split. What do you conclude about importance of **randomly** splitting data into training and testing parts.

Problem 4.

Let's continue with the Boston data set from Problem 3 and the same training data indexes from file HDS_ex3.3_tr.txt.

(a) Do forward selection in training data across 13 variables and report the top model according to AIC as well as to BIC. Always use intercept in the model. Use e.g. `step(,trace = 0)` to not put the output from stepwise searches to your final solution (but you should look what comes out from there while you are working on this problem).

(b) Do backward selection in training data across 13 variables and report the top model according to AIC as well as to BIC. Use intercept in the model.

(c) Make two additional models by forward stepwise search through all possible interaction terms in Boston data set, when predicting `medv`. One model for AIC and another for BIC and report the top models from each information criterion.

(d) Print out the predictors used in each of the 6 models. Print out a 6x2 table where the 6 models are represented by rows and columns are 1=Training MSE, 2=Test MSE.

Which model would you prefer? How much variance would you expect it to explain based on the test MSE?

Problem 5.

Read in dataset "HDS_ex3.txt", the same data that we used in Problem 1. It has $n = 110$ rows and 3 columns `x`, `y`, `z` where `z`is a binary variable.

(a) Do forward stepwise selection with AIC starting from the intercept model `z~1` and having `z ~ 1 + x + x^2` as the largest possible model. Use logistic regression, i.e., `glm(formula, family="binomial")` and note that you need to use `I(x^2)` in formula to get the quadratic term in the model. What is the model chosen by AIC? Compute AIC for the model `z~ 1 + x + x^2`. Why didn't the stepwise search give you the model with the lowest AIC out of all 4 models in the search space?

(b) Do logistic regression for models `z ~ 1 + x + x^2` and `z ~ 1 + x + x^2 + y`. What does a comparison of these two models tell about relationship between `z`and `y` and the quadratic term `x+x^2`?

(c) Do backward selection with AIC starting from the model `z ~ 1 + x + x^2 + y`. Based on all the results you have seen in this exercise, what is your model of choice for predicting `z` using logistic regression?