Please complete the assigned problems to the best of your abilities. Ensure that the work you do is entirely your own, external resources are only used as permitted by the instructor, and all allowed sources are given proper credit for non-original content.

1 Recitation Exercises

These exercises are to be found in: **Introduction to Statistical Learning, 2nd Edition (Online Edition)** by *Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani.*

1.1 Chapter 6

Exercises: 1,2,3,4,5

1.2 Chapter 7

Exercises: 2,3,4,5

2 Practicum Problems

These problems will primarily reference the *lecture materials and the examples given in class* using **R** and **CRAN**. It is suggested that a *RStudio* session be used for the programmatic components.

2.1 Problem 1

Load the *mtcars* sample dataset from the built-in datasets (**data(mtcars)**) into **R** using a dataframe. Perform a basic 80/20 test-train split on the data (you may use **caret**, the sample method, or manually) and fit a linear model with *mpg* as the target response, and all other variables as predictors/features (you will need to set up a dummy variable for *am*). What features are selected as relevant based on resulting t-statistics? What are the associated coefficient values for relevant features? Perform a *ridge* regression using the **glmnet** package from CRAN, specifying a vector of 100 values of λ for tuning. Use cross-validation (via **cv.glmnet**) to determine the minimum value for λ - what do you obtain? (**Hint**: You can use **doMC** in order to speed-up your cross-validation by specifying **parallel=TRUE** in your **glmnet** calls.). Plot training *MSE* as a function of λ (you may also use $\log \lambda$). What is out-of-sample test set performance (using **predict**), and how do the coefficients differ versus the regular linear model? Has ridge regression performed shrinkage, variable selection, or both?

2.2 Problem 2

Load the *swiss* sample dataset from the built-in datasets (**data(swiss)**) into **R** using a dataframe. Perform a basic 80/20 test-train split on the data (you may use **caret**, the sample method, or manually) and fit a linear model with *Fertility* as the target response, and all other variables as predictors/features. What features are selected as relevant based on resulting t-statistics? What are the associated coefficient values for relevant features? Perform a *lasso* regression using the **glmnet** package from CRAN, specifying a vector of 100 values of λ for tuning. Use cross-validation (via **cv.glmnet**) to determine the minimum value for λ - what do you obtain? (**Hint**: You can use **doMC** in order to speed-up your cross-validation by specifying **parallel=TRUE** in your **glmnet** calls.). Plot training *MSE* as a function of λ (you may also use $\log \lambda$). What is out-of-sample test set performance (using **predict**), and how do the coefficients differ versus the regular linear model? Has lasso regression performed shrinkage, variable selection, or both?

2.3 Problem 3

Load the *Concrete Compressive Strength* sample dataset from the UCI Machine Learning Repository (**Concrete_Data.xls**) into **R** using a dataframe (**Note**: You will need to either use the *xlsx* or *readxl* packages to load Excel data, or manually save as CSV). Use the **mgcv** package to create a generalized additive model (via the **gam** function) to predict the Concrete Compressive Strength (CCS) as a non-linear function of the input components (C1-C6) - compare the R^2 value for a GAM with linear terms as well as smoothed terms (**Hint**: Use the **s()** function to apply smoothing using the default *bs* of *tp*). Visualize the regression using the **visreg** package, showing the fit as a function of each predictor with confidence intervals - comment on the quality of the fit at extreme values of the predictors.