STAA577: HW5

- 40 points
- Due: see Canvas for due dates
- Submit your HW by uploading a PDF or DOC file to Canvas. I recommend using the `HW5_yourname.Rmd` file as a template to submit your answers.
- Include all of your R code in an appendix, unless explicitly asked to show it. Your final document may include R output and graphics, but all code should be in the appendix.

---

1. (2 points) Briefly describe the difference between model assessment and model selection.
2. (3 points) In class we discussed methods that provide direct estimates of the test error rate and methods that provide indirect estimates of the test error rate. Describe the relative advantages and disadvantages of these methods.

---

3. The data set `prostate.csv` on Canvas comes from a study by Stamey et al. (1989) that examined the relationship between the level of prostate-specific antigen and a number of clinical measures in men. The clinical measures (predictor variables) are:

   - log cancer volume
   - log prostate weight
   - age
   - log of the amount of benign prostatic hyperplasia
   - seminal vesical invasion
   - log of capsular penetration
   - Gleason score
   - percent of Gleason scores 4 or 5

   The response variable of interest is the log prostate specific antigen (`lpsa`). Use these data to answer the following questions. For reproducibility, be sure to use `set.seed(577)` at the beginning of any code chunk where randomization is being used, e.g., when CV is being used.

   (a) (2 points) Prepare the data for analysis by doing the following:

   - Drop the last column (it's an indicator for training and test, which we will ignore)
   - Standardize the predictor variables (columns 1-8) by using the `scale()` function

   No output needed.

   (b) (3 points) Fit a MLR model with `lpsa` as the response and all other variables as predictors. Examine the `summary()` output and provide a short description of what you see (e.g., which variables appear to be important, variability explained, etc.).

   (c) (2 points) Use the `train()` function to estimate the 10-fold CV **MSE** (note: that RMSE will be reported by default) for predicting `lpsa` with your MLR model from the previous problem. Use `set.seed(577)` to make your results reproducible. (Hint: see the code from HW4, Problem 5(c)).

   (d) (2 points) Use the following code (adapted to your own code) to fit a **ridge regression** model (designated by `alpha = 0`) for a sequence of penalty values λ and make a plot that shows the magnitude of the estimated β coefficients. Include the plot in your output.

   ```
   lpsa <- prostate$lpsa
   xmat <- data.matrix(prostate[, -9])
   ridgeOut <- glmnet(x = xmat, y = lpsa, family = "gaussian", alpha = 0, nlambda = 200)
   plot(ridgeOut)
   ```

   (e) (2 points) We can use CV to determine an optimal value for the penalty parameter λ in our ridge regression model. Use the function `cv.glmnet()` to automatically perform CV for each value of λ in the sequence. Use `set.seed(577)` to make your results reproducible. (Nothing to report for this problem.)

(f) (3 points) Using the results from the previous problem, report the optimal value of λ and the corresponding CV MSE.

(g) (2 points) Fit a **LASSO regression** model for a sequence of penalty values λ and by modifying the code from Problem 6 to set the argument `alpha = 1`. Make a plot that shows the magnitude of the estimated β coefficients. Include the plot in your output.

(h) (3 points) Compare and contrast the plots from Problems 3(d) & 3(g). Describe some of the similarities and differences of these figures and how they relate to ridge regression and LASSO methods. Also note how the patterns in the plots relate to the value of λ.

(i) (2 points) Repeat Problems 3(e) & 3(f) for LASSO. Report the the optimal value of λ and the corresponding CV MSE.

(j) (4 points) Use `glmnet()` to fit the LASSO model using $\lambda = 0.1$. (Note: this value is bigger than the optimal value from the previous problem.) Then, bind together the estimated coefficients from the LASSO model and the MLR model that you fit in Problem 4. You might use some code like this:

```
cbind(coef(lassOut), coef(lmOut))
```

Which variables did the LASSO model select? Briefly describe how the parameter estimates compare for the two methods.

---

4. We will now try to predict per capita crime rate in the ‘Boston’ data set from the ‘MASS’ package.

(a) (4 points) Try out some of the regression methods explored in this chapter such as the lasso, ridge regression, or linear regression. Feel free to tryout methods like Elastic-net, PCR, etc. for extra (3 points). Present and discuss results for the approaches that you consider.

(b) (4 points) Propose a model or a set of models that seem to perform well on this data set and justify your answer. Make sure that you are evaluating performance using validation error, cross-validation error, or some other reasonable alternative (not just training error).

(c) (2 points) Does your chosen model involve all of the features in the data set? Why or why not?