Homework 2
CSC 424
Dr. John McDonald

Deliverables: Turn in your answers in a single PDF file. Copy any R commands and output relevant to your answer into your document and explain your answer thoroughly and include a copy of the full analysis in your report along with your conclusions.

1. (20 points) Perform, by hand, the following calculations from linear algebra. For the following matrices and vectors. Submit a copy of your answers either in a high-quality scan or photo (unreadable or clipped answers will not receive credit) or you may format it carefully in Word with equation editor showing all work. Note that *Z* has changed from last week, so re-do a) and b) here. **Note: these numbers are different from homework 1!**

$$Z = \begin{bmatrix} 1 & 2 \\ 1 & -3 \\ 1 & 4 \\ 1 & -1 \end{bmatrix}, Y = \begin{bmatrix} 0 \\ 1 \\ 4 \\ -3 \end{bmatrix}$$

   a. Z^T
   b. $Z^T Z$ (Make sure you get the right dimensions on this matrix)
   c. $(Z^T Z)^{-1}$ (You do not have to perform Gaussian-Elimination here … i.e. this is a hint on what the dimensions of the matrix should be ☺)
   d. $Z^T Y$
   e. $\beta = (Z^T Z)^{-1} Z^T Y$
   f. $\det(Z^T Z)$

2. (10 points) In R, write a script to compute each of the parts in problem 1 to check your answers. Submit both the .r commands and the output. Then, create a dataset with x = <2, -3, 4, -1> and y = <0, 1, 4, -3> and run a regression analysis on the data with "lm". Compare your value for β in part *e* of the last problem, with the coefficients calculated by R's lm function.

3. **(10 pts)** Use the dataset "mtcars" which is built-in RStudio. You can see the structure of the data by the command "head(mtcars)". Use the "mpg" column as your dependent variable Y, and do the following:

   a. Create a copy of the dataset called *A* with **only** the columns {cyl, disp, hp, wt, carb}. Use the column selection mechanism we covered in class to select these columns from the dataset.
   b. Add a column of "ones" to *A* called "count".
   c. Use the "as.matrix" function to convert it to a matrix and assign it back to the variable *A* (so you are overwriting the data.frame here and converting it to a matrix)
   d. Compute the following multiple regression by manually computing the matrix operations: $(A^T A)^{-1} A^T Y$.
   e. Compute the regression with the RStudio "lm" command and compare with your results from d). Note any differences.

4. **(20 pts)** Return to the housing dataset. In the last homework you conducted a regression with various methods of feature selection. In this problem you will turn to evaluating regression's predictive power on this dataset, whether it is overfitting and investigate the performance of regularized regression. For this, I have provided you with the data divided into test and training sets.

   Note that R's glmnet handles scaling internally, so you do not need to scale manually and then undo the scaling when computing betas so that the reported coefficients are applicable to the original data. All of this happens transparently. One other note here: **Do not try to calculate an R^2 on a test set. This measure applies only to a training set and has no meaning for out-of-sample prediction. On a test set, the so-called R^2 value could be > 1 and could even be negative! This is one reason we use RMSE.**

   One good way to compare the test and training RMSE's is to take their quotient (testRMSE / trainingRMSE) and that will tell you how much bigger test is than training %-wise.

   a. Rerun your full (all predictors) linear regression model of MEDV, using the training set and then calculate and report the R^2 and RMSE of the residuals. Then predict the y-values in the test set using this model. What is the RMSE for the test set? Is there evidence of overfitting here?

   b. Use cross-validated ridge regression on the training set and plot the relationship between the cross-validated error and log-lambda. Then use the model for predicting again the y's in the test set using the "lambda.1se". Report the R^2 for the training set (how do you get this?) and the RMSE for both the training and test set. How do these compare to the OLS regression model? Are they improving prediction? If so, how?

   c. Do the same as in b) for Lasso using "lambda.1se". How do the results compare to Ridge and OLS regression? Also, for Lasso, evaluate how the number of variables changes with lambda. How many variables are selected at lambda.1se? Finally, how do the variables selected, and their betas computed compare with the model you got last week with stepwise regression?

   d. Revisit the cross-validated error graphs for both Ridge and Lasso, and evaluate how well regularization is working for this set? Is there a strong indication that cross-validated error can be reduced by regularization in this case? What do you think this mean for overfitting in the original model?

5. (20 points) The data in the files *insurTest.csv & insurTrain.csv* are collected from 47 zip-code areas in the Illinois area. There are 8 columns in the data file but not all are relevant here. The response variable of interest is the number of new home insurance policies (NEWPOL) (minus canceled policies) per 100 housing units. The predictor variables are the percent minority population living in the area (PCTMINOR), the number of fires per 1000 housing units (FIRES), the number of thefts per 1000 in population (THEFTS), the percent of housing units built before 1940 (PCTOLD), and the median income (INCOME).

   a. Run a multiple regression of NEWPOL on the variables: PCT-MINOR FIRES THEFTS PCTOLD INCOME NEWPOL.
      i. What is the overall fit (F-score, R^2) and what is the RMSE on the training set?
      ii. Which predictors have coefficients significantly different from zero at the .05 level?
      iii. Do any of the predictors have signs that are different than suggested by their simple correlations? If so, explain what may be happening.
      iv. Examine a plot of residuals versus predicted values. Do you see any problems?

   b. Use the model from a) to predict the test set. What is the RMSE here? Is there evidence of overfitting?

c. Run a ridge regression on the training set, produce the lambda plot, and report the RMSE for both the training and test sets using lambda.1se. From the lambda plot, how well is regularization working here? Look at the shape of the plot for a significant dip before it rises.

d. Run a lasso regression on the training set, produce the lambda plot, and report the RMSE for both the training and test sets using lambda.1se. Compare them with those you got in c). Are they closer together, more stable, or have they worsened overall?

e. Plot the residuals vs the fitted (predicted values), you can do this with plot(fitted, residuals) since you've calculated both in order to get the RMSE (remember, residuals are the difference between actual and predicted). Does the lasso add any significant bias into the model? Note that it may or may not show up as a slope difference. It could also appear as a mean residual different from 0 as well.

f. Use Elastic Net regression and determine if there is an alpha between 0 and 1 that will give a better result with this test and training set. Try alphas of .25, .5 and .75. Does it appear that there might be an alpha that does better than either ridge or lasso?

g. Is there a practical reason to try and mix lasso and ridge here? Explain your answer.

6. (20 points) Read and review the posted paper "Adding bias to reduce variance in psychological results." Most of the mathematics here will be similar to what we've gone over in class but pay particular attention to Section 3: Examples. Answer the following questions, in detail. You should be able to write a detailed paragraph about each of at least three or four complete sentences with

    a. What is the size of the dataset relative to the number of independent variables?
    b. Is there evidence of overfitting in their dataset?
    c. How are regularized regression techniques being used in their examples?
    d. How do the results of regularized regression differ from the OLS model?
    e. How do they evaluate the performance of each of the regularized regression techniques?
    f. Are there any issues that you can identify with the way that they are evaluating this performance?

7. (5 Points) Post a comment to the "Lecture 1 & 2" discussion forum regarding a topic from lectures 1 & 2 and homeworks 1 & 2. In your post, describe what you found easiest to understand and also what you found most difficult. Think about topics that you found most interesting, topics that you would like to hear more about, or topics that you found confusing and you would like more clarification. Please also take the time to respond to your classmates' questions and comments (respectfully of course ☺).