# Regression, Classification and Clustering

## Regression

### Problem 1

Using National Foot Ball League data available as nfl1.csv

1. build a linear regression model with **3** most relevant independent variables relating to the number of games won.
2. Construct the analysis of variance table and test for significance regression
3. Calculate *t* statistics for testing the hypotheses for the three coefficients you have chosen *H*0: *β*v1 = 0 *H*0: *β*v2 = 0, and *H*0: *β*v3 = 0. What conclusions can you draw about the roles the variables *x*v1, *x*v2, and *x*v3 play in the model
4. Calculate R squared and R Adjusted for the model Using the partial F test, determine the contribution of each variable to the model
5. find the 95% CI for one of the coefficients *β*v1, *β*v2 = 0 or *β*v3
6. Find a 95% CI on the mean number of games won by a team for this data set x1 to x9 -> 2110, 2127, 38, 59, 1, 789, 58, 2110, 2127 with your model
7. using the same data, fit a model to these data using only *x*7 and *x*8 as the regressors and test for significance of regression
8. Calculate *R*2 and *R*2Adj. How do these quantities compare to the values computed for the previous model you built which has 3 regressors?
9. Calculate a 95% CI on *β*7. Also find a 95% CI on the mean number of games won by a team when *x*7 = 56.0 and *x*8 = 2100. Compare the lengths of these CIs to the lengths of the corresponding CIs from previous model
10. What conclusions can you draw from this problem about the consequences of omitting a regressor from a model?

For all Standard error calculation, read the article below:

Reference: Reference: <https://onlinecourses.science.psu.edu/stat501/node/297>

To calculate 95% CI <http://www.stat.yale.edu/Courses/1997-98/101/linregin.htm>

Matrix operation <https://datascienceplus.com/linear-regression-from-scratch-in-r/>

### Problem 2

Using the gasoline mileage data - gasmf.csv

1. Fit a multiple linear regression model relatmg gasoline mileage *y* (miles per gallon) to engine displacement *x*1 and the number of carburetor barrels *x*6.
2. Construct the analysis-of-variance table and test for significance of regression.
3. Calculate *R*2 and *R*2Adj for this model
4. Find a 95% CI for *β*1.-----WRONG
5. Compute the *t* statistics for testing *H*0: *β*1 = 0 and *H*0: *β*6 = 0. What conclusions can you draw? – predict from the summary
6. Find a 95% CI on the mean gasoline mileage when *x*1 = 275 in.3 and *x*6 = 2 barrels.
7. Find a 95% prediction interval for a new observation on gasoline mileage when *x*1 = 257 in.3 and *x*6 = 2 barrels.
8. Repeat this with just with x1 and compute a 95% CI on mean gasoline prediction interval on mileage with the engine displacement *x*1 = 275 in.
9. Compare the lengths of these intervals to the lengths of the confidence and prediction intervals from the above model with two variables. Does this tell you anything about the benefits of adding *x*6 to the model?

### Problem 3

Using the house price data houseprice.csv

Y = sales price/10000; x1 = taxes, x2=number of baths, x3 = lot size / 1000 sq.ft, x4= living space /1000 sq. ft, x5 = no. of garage stalls, x6 =number of rooms, x7= no of bed rooms, x8 = age of home, x9 =no. of fire places

1. Fit a multiple regression model relating selling price to all nine regressors.
2. Test for significance of regression. What conclusions can you draw?
3. Use t tests to assess the contribution of each regressor to the model. Discuss your findings.
4. What is the contribution of lot size and living space to the model given that all of the other regressors are included?
5. Is multicollinearity a potential problem in this model? – use vif function to see the multicolinearity

### Problem 4

Use the bhp.csv file for the following exercise

1. Use caret library
2. Read the bhp.csv file
3. Convert outcome variable to a factor 0: failure; 1 ; Good
4. Create 70:30 training / test data set
5. Build a logistic regression model and store it in logit variable
6. Discuss the deviance results and significance of coefficients
7. Compute probabilities of success using train data and store it in a variable
8. Classify the cases using a cutoff probability of .5—plot graph
9. Generate the error/classification-confusion matrix
10. Repeat this for test data – **7 8 9**

### Problem 5 - Clustering

1. Load the following packages -

library(class) #k-nearest neighbors

library(kknn) #weighted k-nearest neighbors

library(e1071) #SVM

library(caret) #select tuning parameters

library(MASS) # contains the data

library(reshape2) #assist in creating boxplots

library(ggplot2) #create boxplots

library(kernlab) #assist with SVM feature selection

1. Combine following data sets - Pima.tr and Pima.te
2. Use the melt function by "type"
3. Create a two column plot with type and value using the melt data
4. Normalize the data and recreate the data frame
5. redo the plot
6. Explain if there are any colinearity issues
7. What is the ratio of Yes and No in the response
8. Set a seed and create a 70 / 30 train / Test data set
9. Train the model and come up with the optimal cluster size
10. Calculate the confusion matrix and the error
11. Use the kknn package and its train.kknn() function to select the optimal weighting scheme
12. Create a new training set using kknn library In the kknn train data set use kmax = 25 and distance 2 for absolute distance and for kernel use c("rectangular", "triangular", "epanechnikov"))
13. Find the best k
14. use the kknn to predict the test data and build the confusion matrix; calculate the accuracy
15. use the e1071 package to build the SVM models