STA 314, Homework
Due date: December 2

This homework will require you to download the files 'trainingdata.csv' and 'testpredictors.csv' from quercus and load both files into R. All questions are related to the corresponding data sets. **Submission: questions 1-4 should be submitted via quercus, a special place for that will be provided. Questions 5-7 should be submitted via kaggle.**

1. (1 mark) Run a linear model for y as response using all predictors (i.e. predictors X1,...,X28) and compute a prediction for the *first predictor vector* from the test set.

2. (1 mark) Which predictors does a model with two predictors selected by best subset selection contain?

3. (2 marks) Run a gam model of the form

$$f(X) = b_0 + b_1 X24 + g_1(X1) + g_2(X12)$$

where g_1 is a natural spline with knots $-1, 0, 1$ and g_2 is a polynomial of degree 2. What is your estimated value for the coefficient b_1?

4. (2 marks) Determine the predictor that has highest variable importance as computed from bagging according to the criterion $\%IncMSE$ function `varImpPlot` as discussed in lectures.

5. (2 marks) Submit a prediction that has a score < 1.8 on the public leaderboard before the competition ends.

6. (1 mark) Submit a prediction that has a score < 0.9 on the public leaderboard before the competition ends.

7. (1 mark) Submit a prediction that has a score < 0.8 on the public leaderboard before the competition ends.