# Final Project

## Academic Honesty

We expect each team to work together to finish the project. Teamwork is highly encouraged but any form of inter-team interaction is prohibited. If any doubts, come forward to me or refer to Rutgers code http://academicintegrity.rutgers.edu/.

## General Guidline

- **Idea**

The final project is to apply several basic idea and methods in statistical modeling to a moderate dataset, in the software environment of SAS. Therefore both good understanding of statistics concepts and proficiency in SAS programming are two necessary elements to a successful project.

- **Grading**

The project is **40pts + 10pts(bonus) = 50 pts** in total.

My grading depends on: quality and clarity of your submission.

**Quality** is a measure of how well you understand real problem and the ability to answer it from a statistician's perspective by applying appropriate statistics modeling methods with virtuosity in SAS coding, graphical presentation is also encouraged.

**Clarity** refers to both readability in compliance with scientific writings forms. Your writings without a consistent and careful indexing manner would be a sign of lacking readability, penalty will be applied to that. Your sas code without enough comment is considered unfriendly. Also including too much irrelevant SAS output in your report is another negative sign that costs you a lot.

- **Submission**

Similar to what we did in homework, only two files be submitted to sakai : **report.pdf** and **code.sas**. If your report and code are not in the required format: .pdf and .sas respectively, **10% penalty** is carried. Make sure all your team members' names show up in **the first page** of your report.

**Problem**

**1. Download Data Set**

The Boston housing data was collected in 1978 and each of the 506 entries represent aggregated data about 14 features for homes from various suburbs in Boston, Massachusetts. Find it here: https://archive.ics.uci.edu/ml/machine-learning-databases/housing/

Note that there are two files: ‘ housing.data ‘ is a data file and ’housing.names’ is a reference file.

**2. Import and Preprocess (10 pts)**

(a)Importing data and Naming variables

when you import data into SAS, please read through the section *7. Attribute Information* in the file ‘housing .names’, and name fourteen variables accordingly. The last column should be named as MEDV, which is the dependent variable in our linear regression analysis. All the rest right-hand-sided variables should be named as these items 1-13 in *7. Attribute Information* ,respectively.

(b) Removing Outlier

Based prior experience , in the dataset, there are two types of outliers to remove, which are any observation of MEDV = 50 or any observation of RM = 8.780.

To be specific, you have to remove any row from the whole data, if the observation of variable MEDV = 50 or RM = 8.780; Report your data size after your removal of outlier. And explain why outlier could be harmful to our modeling.

**3. Two-Way ANOVA (10 pts)**

(a)Does MEDV depend on CHAS , RAD and their interaction?

(b)State these model assumptions and check validity.

**4. Linear Regression (20 pts)**

**(a) Splitting Data into training and test data (5 pts)**

- After outlier removal, make the first 70% to be training date and the rest to be test data.
- To be specific, let N be the total number of rows after removal. First you have to find an integer j* such that j*/N ≈ 70%.

Then, your training data consists of all the rows from the first until j*th row. And your test data consists of all the remaining rows, i.e. from (j*+1)th row to the last row.

- Report the size of your Training data and Test data.

**(b) Full linear Model on Training data (5 pts)**

- Estimate your full linear model using the training data.
- Report F-value, Adj R-square, and Root MSE.
- Do Model Checking on residual and multicollinearity.

**(c)Best Model on Training data (10 pts)**

- Do model selection on linear model, applying these four options for selection = forward, backward, stepwise and rsquare;
- For each option, report your best model's **variable list** (which right-hand-side variables are included in best model) , **F-value**, **Adj R-square**, and **Root MSE**, compare them with full model from (b). Note that different model selection criterion may give us identical model. please report all of them and clearly indicate from which proc options model comes from.
- In terms of minimizing Root MSE, which model is best among all models from last step? If the full model is not the best, explain the intuition why dropping variables may be beneficial to reducing training error.

**(d) Test Error on Test data**

- Report the test Root MSE of your best model from the last step in (c) , compared with that of the full model from (b). Which Root MSE is larger and why?

- If you have different models from the second step in (c), report the test Root MSE. Is minimizing Root MSE in (d) give us the same best model as doing it in (c)?

Bonus 5. **Binary Classification using linear regression(10 pts)**

In hw3, we have seen an example of binary classification in a non-seperable case. Here we have another binary classification problem. Let us start with the dataset after removal.

(a)Find MEDV's median and create a new variable Y: Y = 1 if MEDV >= its median; Y = 0, otherwise. Throw away MEDV and keeps Y in your datase. Using the same spliting rule (as in 4(a)) to create Training and Test data.

(b) Out of all right-hand-side variables in training data, find two variables X1 and X2 with the highest and second highest absolute value of correlation coefficient with Y. To calculate the coefficient, you may treat any string variable as binary or categorical variable.

(c) **Plot** X1 vs X2 in a 2-dim plot and **label** each point in the plot by 'H' if Y =1 and 'L' if Y=0. Are these two class of points separable by a line? show your plot.

(d) If you run a linear regression of Y on X1 and X2, you may notice already that the predicted value from linear regression is not discrete-valued. In order to make prediction of Y, you have to come up with a reasonable rule to assign {0,1} to Y. Please **state explicitly** what your rule is. Under your prediction rule, what is the number of **classification errors using training data**. And what is the number of **classification errors using test data**. Report both of them and explain intuitively why one type of error is larger than the other?