Dear all, this is an open book exam. Please submit your RMD output file in PDF or Word Format, to the class Brightspace site. Academic dishonesty will result in a grade of F for the class.

Linear Regression with the Ames Housing Data – Regression Task

The **Ames_Housing_Data.csv** file contains data on various aspects of houses

sold. **Our goal is to predict the SalePrice using all the other variables**. The

variables are:

- SalePrice - the property's sale price in dollars. This is the target variable that you're trying
- to predict.
- ID – the ID number we assigned to each house. There are a total of 1460 houses in this
- data set.
- LotArea: Lot size in square feet
- OverallQual: Overall material and finish quality
- OverallCond: Overall condition rating
- YearBuilt: Original construction date
- YearRemodAdd: Remodel date
- CentralAir: Central air conditioning
- 1stFlrSF: First Floor square feet
- 2ndFlrSF: Second floor square feet
- GrLivArea: Above grade (ground) living area square feet
- FullBath: Full bathrooms above grade
- HalfBath: Half baths above grade
- Bedroom: Number of bedrooms above basement level
- Kitchen: Number of kitchens
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
- Fireplaces: Number of fireplaces
- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage in square feet
- YrSold: Year Sold

1. Please use the random seed 123 to divide the data into 75% training and 25% testing.
2. Please find the best model using the stepwise variable selection method (based on the BIC criterion) using the training data.
   Please (a) display the coefficients of the fitted model; (b) make prediction on the testing data, and report the RMSE and the Coefficient of Determination R^2.
3. Please find the best model using the best subset variable selection method (based on the SSE criterion) using the training data. Please (a) display the coefficients of the fitted model; (b) make prediction on the testing data, and report the RMSE and the Coefficient of Determination R^2.
4. Which model selection method among the 2 we have used above is the best? (a) Please compare the BIC of these models using the training data, as well as display

these two models so we can see the parameter estimators and model goodness of fit measures. (b) Furthermore, please compare the RMSE and R^2 of these models using the test data. (c) Please discuss any modifications you can do to further improve your model(s).