Statistical Learning and Analysis
Instructor: Prof. Varun Rai

Assignment 1

There are three questions. Each has several parts. You can either use R language or python (jupyter notebook) to do your assignments

---

Question 1. Interaction Effects in a Linear Regression

Suppose you are modeling the relationship between income as the response variable and race (A/B) and sex (male/female) and their interaction.

a) Write down the equation describing a linear regression model for this problem. Note: You will need to create/define the necessary predictor variables.
b) Interpret the coefficients in the equation in (b).
c) Now drop the "main effects" terms in the equation in (b) and reinterpret the coefficients in the new model. Note: For two predictor variables X1 and X2 the "main effects" terms are the terms $\hat{\beta}1$ X1 and $\hat{\beta}2$ X2.
d) The hierarchical principle for including interactions in a MLR requires inclusion of the main effects in the model whenever interaction effects are included. Given your interpretation in part (d), can you briefly describe the practical relevance of the hierarchical principle?

Question 2. This question involves the use of multiple linear regression on the provided data set

(a) Produce a scatterplot matrix which includes all of the variables in the data set.
(b) *Compute the matrix of correlations between the variables using the function. If one can compute the correlation for qualitative variable.*
(c) Use Python/R (sklearn in python and lm() function in R) to perform a multiple linear regression with SalePrice as the response and all other variables as the predictors. you can use all or less variables if you think it can improve the model. Also, answer the following question.
    - Is there a relationship between the predictors and the response?
    - What does the coefficient for the age variable suggest?

(d) Try a few different transformations of the variables, such as log(X), √X, X2. Comment on your findings.

Question 3. In this exercise you will create some simulated data and will fit simple linear regression models to it. Make sure to set seed prior to starting to ensure consistent results.

a) Create a vector, x, containing 100 observations drawn from a N (0, 1) distribution. This represents a feature, X
b) Create another vector, eps, containing 100 observations drawn from a N(0,0.25) distribution i.e. a normal distribution with mean zero and variance 0.25
c) Using x and eps, generate a vector y according to the model $Y = -1 + 0.5X + \varepsilon$
    - What is the length of the vector y? What are the values of $\beta 0$ and $\beta 1$ in this linear model?

d) Create a scatterplot displaying the relationship between x and y. Comment on what you observe.
e) Fit a least squares linear model to predict y using x. Comment on the model obtained. How do βˆ0 and βˆ1 compare to β0 and β1?
f) Now fit a polynomial regression model that predicts y using x and x2. Is there evidence that the quadratic term improves the model fit? Explain your answer.
g) Repeat (a)–(f) after modifying the data generation process in such a way that there is less noise in the data. The model should remain the same. You can do this by decreasing the variance of the normal distribution used to generate the error term ε in Describe your results.

Code-Book for the provided dataset

Data fields

Here's a brief version of the provided data.

- **SalePrice** - the property's sale price in dollars. This is the target variable that you're trying to predict.
- **MSSubClass**: The building class
- **LotFrontage**: Linear feet of street connected to property
- **LotArea**: Lot size in square feet
- **OverallQual**: Overall material and finish quality
- **OverallCond**: Overall condition rating
- **YearBuilt**: Original construction date
- **BsmtFinSF1**: Type 1 finished square feet
- **BsmtFinSF2**: Type 2 finished square feet
- **BsmtUnfSF**: Unfinished square feet of basement area
- **TotalBsmtSF**: Total square feet of basement area
- **1stFlrSF**: First Floor square feet
- **2ndFlrSF**: Second floor square feet
- **GrLivArea**: Above grade (ground) living area square feet
- **BsmtFullBath**: Basement full bathrooms
- **BsmtHalfBath**: Basement half bathrooms
- **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
- **PoolArea**: Pool area in square feet
- **YrSold**: Year Sold