**STAT 40001/STAT 50001 Statistical Computing Fall 2022**

**Lab -23**

**Q.N. 1)** An important challenge in clinical trials is patients who drop out before the trial is completed. This can cost pharmaceutical companies millions of dollars. Can we predict who will drop out of the study early? The link below provides the Age and the Hamilton Rating Depression Scale (HRDS) and whether or not they completed the study (Drop: 1=Yes and 0=No)

[https://media.pearsoncmg.com/aw/aw\_sharpe\_business\_3/datasets/txt/Clinical Trials.txt](https://media.pearsoncmg.com/aw/aw_sharpe_business_3/datasets/txt/Clinical%20Trials.txt)

1. Import the data in R and determine its dimension.
2. The missing values are left blank. Please clean data by removing missing values.
3. Fit a multiple logistic regression model using Age and HDRS as a predictor variables.
4. What is the predicted dropout probability of a 30-year-old patient with HDRS score of 30?

**Q.N. 2)** A dataset concerning the relationship between respiratory function (measured by forced expiratory volume, FEV) and smoking provides a powerful tool for investigating a wide variety of statistical matters. <http://jse.amstat.org/datasets/fev.dat.txt>

Below are the description of the column of the data

Columns Label Description

1-3 age discrete measure, positive integer (years)

5-10 fev continuous measure (liters)

12-15 height continuous measure (inches)

19 sex discrete/nominal (Female coded 0, Male coded 1)

25 smoke discrete/nominal (Nonsmoker coded 0, Smoker coded 1)

1. Import the data in R and identify its dimension.
2. Test whether smoking status differ by gender (Hint: perform two sample proportion test)
3. Fit a multiple linear regression model to study fev using age, height, sex and smoking status as predictor variables.
4. Use your model in (c) to predict the FEV of a 50 inches tall 12-year-old girl who is not a smoker. Please construct a 95% confidence interval as well.
5. A real estate agent used information on 1115 houses. She wants to predict whether a house sold in the first 3 months it was on the market based on other variables. The variables available include:

*Sold :*  1 = Yes—the house sold within the first 3 months it was listed;

0 = No, it did not sell within 3 months.

*Price:*  The price of the house as sold in 2002.

*Living Area:* The size of the living area of the house in square feet

*Bedrooms :* The number of bedrooms

*Bathrooms :* The number of bathrooms (a half bath is a toilet and sink only)

*Age:* Age of the house in years

*Fireplaces:* Number of fireplaces in the house

Fit a multiple logistic regression model to predict whether a house will sell within the first 3 months it’s on the market based on Price ($), Living Area (sq ft), Bedrooms (#), Bathrooms (#), Fireplaces (#), and Age (years).

1. A local health clinic sent fliers to its clients to encourage everyone, but especially older persons at high risk of complications, to get a shot in time for protection against an expected flu epidemic. In a pilot follow-up study, 159 clients were randomly selected and asked whether actually received flu shot. A client who received flu shot is coded Y=1 and client who did not receiveflu shot was coded Y=0. In addition, the data was collected their age (X1), and the their health awareness. The health awareness (X2) is indexed with higher value indicating greater awareness. Also the gender of the client (X3 = 1 for male and X3 = 0 for female) was recorded. The data is provided in the Blackboard. Assuming that a multiple logistic regression best fit the data, answer the following questions:
2. Fit a multiple logistic regression model and check for significance of each variable (X1, X2 and X3)
3. What is the estimate probability that a male client aged 55 with a health awareness index 60 will receive a flu shot?