Assignment 4

1. Moore (1975) reported the results of an experiment to construct a model for total oxygen demand in dairy wastes as a function of five laboratory measurements (data). Data were collected on samples kept in suspension in water in a laboratory for 220 days. Although all observations reported here were taken on the same sample over time, assume that they are independent. The measured variables are:

   ```
   y : log (oxygen demand, mg oxygen per minute)
   x1 :biological oxygen demand, mg/liter
   x2 : total Kjeldahl nitrogen, mg/liter
   x3 : total solids, mg/liter
   x4 : total volatile solids, a component of x3, mg/liter
   x5 : chemical oxygen demand, mg/liter
   ```

   Fit a multiple regression model $y=\beta_0+\beta_1x1+\beta_2x2+\beta_3x3+\beta_4x4+\beta_5x5+\varepsilon$ using y as the dependent variable and all x_j's as the independent variables.

   a. Form a 95% confidence interval for β_3 and again for β_5.

   b. Form a 95% confidence interval for $\beta_3 + 2\beta_5$.

   c. Show graphically a 95% confidence region for β_3 and β_5. Plot the origin on this display. The location of the origin on the display tells us the outcome of a certain hypothesis test. State that test and its outcome.

   d. If a 95% joint confidence region was computed for (β_1, β_2, β_3, β_4, β_5), would the origin, (0, 0, 0, 0, 0), lie inside or outside the region? Explain.

   e. Suppose it is suspected that non-volatile solids have no linear effect on the response. State a hypothesis in terms of the parameters of the full model that reflects this suspicion, and test it using a confidence interval in your answer to one of the above questions. Explain why the chosen confidence interval can be used to do this work.

2. The data are a random sample of home sales from Spring 1993 in Albuquerque. The variables are:

   Price: Selling price in hundreds of dollars
   SQFT: Square feet of living space.
   Age: Age of home (years)
   Features: Number out of 11 features (dish washer, refrigerator, microwave, disposer, washer, intercom, skylight(s), compactor, dryer, handicap fit, cable TV access)
   NE: Located in northeast sector of city (1) or not (0)
   Corner: Corner location (1) or not (0)
   Taxes: Annual taxes in dollars

   a. There are a large number of missing values (denoted by "NA" in the dataset) in the `Age` variable. We could exclude `Age` from our models for the selling price or we could keep `Age` and exclude the cases that have missing values for `Age`. Which choice is better for this data? Explain your reasoning.
   b. Fit a model with selling price as the response and `SQFT`, `Features`, `NE`, `Corner`, and `Taxes` as predictors. Form 95% confidence intervals for their coefficients. Form 99%

confidence intervals for their coefficients. Explain how the p-value for the parameter for `Corner` relates to whether zero falls in the two corresponding confidence intervals.

c. Predict the selling price of a specific house with `SQFT=2500`, `Features=5`, `NE=1`, `Corner=1`, and `Taxes=1200`. Give an appropriate 95% confidence interval.

d. Suppose you are only told that `SQFT=2500`. Predict the selling price and 95% confidence interval.