Homework 2
Due September 16th at 11:59 pm

Directions: Write out the solution to the problems in the assignment. Additional practice problems have been provided, but you do not need to turn them in. ***Save your work as a pdf file before submitting it under the Assignments tab in Blackboard.***

Answer all questions at the 95% confidence level.

1. City planners are interested in determining the amount that families spend each week on food. They created a random sample of 150 households which included the annual household income in thousands of dollars and the average weekly amount spent on food during the past year. The data is contained in the file food_income.xlsx.

   (a) What is the estimate for the amount of money spent each week on households within the community?

   (b) What is the estimate of the variable coefficient?

   (c) What is the prediction interval for the amount of money spent each week on food for a household whose annual income is $40,000?

2. A realty company in a large city that specializes in apartment sales wants to build a model that helps them predict the costs for apartments in areas of the city that are several miles away from downtown. They collect a random sample of recent apartment sales in the file apt_sales.xlsx that lists the cost of the apartment, the size of the apartment in square feet, the apartment's distance from downtown, and the age of the apartment.

   (a) Is the data appropriately modeled with a linear model? Explain your answer.

   (b) Create a model for the cost of an apartment using all the variables included in the dataset and interpret the model.

   (c) What is your estimate for each variable coefficient? Which of them are statistically significant?

   (d) Would the best model be one that included all the variables you were given? Determine the best possible model and calculate the general linear F-statistic to confirm that adding additional variables does not improve the model.

3. Researchers are interested in testing whether a college student's brain size and body size are predictive of his or her score on an IQ test. The researchers collected the following data in the file iq_physical.xlsx on a sample of 38 students:

   - Performance IQ scores (PIQ) from the revised Wechsler Adult Intelligence Scale
   - Brain size based on the count obtained from MRI scans (given as count/10,000)
   - Height in inches
   - Weight in pounds

   (a) Would you want to include both height and weight in your model? How can you justify your answer using both the correlation coefficient and the coefficient of determination?

   (b) At the 95% confidence level, what is the estimate of the variable coefficients if you only use brain size and height as independent variables in your model? How should you interpret the coefficients?

Additional Problems:

1. The human resources department at a large company wants to develop a model to predict an employee's job satisfaction from the number of hours of unpaid work per week the employee does, the employee's age, and the employee's income. A sample of 120 employees at the company is taken and each employee's job satisfaction score is out of 10 is recorded, with higher values indicating greater job satisfaction. The data is in the job_satisfaction2.xlsx file, with the employee's income recorded in units of thousands of dollars.

   (a) Would the best model be one that included all the variables you were given? Determine the best possible model and calculate the general linear F-statistic to confirm that adding additional variables does not improve the model.

   (b) Interpret the variable coefficients in your preferred model.

   (c) What job satisfaction rating would you expect for an employee with 4 hours of unpaid work who makes $40,000 in income?

2. A national used car dealership is interested in determining whether there is a difference in the resale value of different editions (Premium versus Limited) of Subaru Outbacks. Use the file used_subaru.xlsx to create a multiple regression model that uses the mileage on the odometer and the edition of the car to predict the price they are able to sell the car for.

   (a) Interpret the variable coefficients in terms of how the independent variable effects the dependent variable.

   (b) How much of the variation in the dependent variable is explained by the model? Explain your answer.

   (c) What price would you estimate for the sale of a used Outback that has 15,000 miles on it for each edition?

   (d) Would you want to include the Year variable in your model? Explain why or why not.

3. A local company collects employee data from its HR department for a random sample of employees. The dataset, in the file employee.xlsx, contains each employee's current salary, the salary they were hired at, and a variable listed as jobtime which measures how long the employee has been with the company, in addition to various demographic measures. The company is interested in building a model that predicts current employee salaries.

   (a) Starting with a model that includes only starting salaries and jobtime, describe the results of your model and estimate the salary of an employee who started with a salary of $20,000 and has a jobtime value of 90.

   (b) Does your initial model improve when adding in both gender and minority status as variables? Run the general linear F-test to draw your conclusions.

   (c) Does your initial model improve when adding an interaction term between beginning salary and jobtime? Run the general linear F-test to draw your conclusions.