**IBEHS 4C03: STATISTICAL METHODS IN BIOMEDICAL ENGINEERING**

**ASSIGNMENT #5**

**Topics:**

Simple Linear Regression Transformations

Multiple Linear Regression

Logistic Regression

**Total marks: 100**

**Submission Instructions:** Electronic submission to the IBEHS 4C03 Avenue to Learn Assignment folder.

**Assignment Submission Format**:

File Names: StudentLastNameFirstNameAssignment#.doc/pdf/jpynb etc.

Both your name and student numbers should appear at the top of the document. If separate documents are submitted, then you should submit the answers to the questions in a document including the plots you made in python. Regardless, the python file used to generate

the plots and any provided answers should also be submitted with any submission. Submit a single zip file per assignment.

# Problem 1: [50 points]

This question utilizes the ReactorData.csv file and continues with the SLR example from Assignment 4. The data contains information from a lab-scale bioreactor. The outcome of interest is the yield of the bioreactor (g). Each batch was performed with a different combination of temperature (measured in C), mixing speed (measured in 100 RPM) and reactor type (one with baffles and one without). For the questions where you need to fit and interpret a regression model, you can assume all of the assumptions hold (you don’t need to verify them).

1. For the model for 𝑦𝑖 = 𝑏0 +𝑏1 ∗ 𝑇𝑒𝑚𝑝+𝑒 , what is the predicted value ot a temperature of 70®C? What is the 95% confidence interval of the prediction? What is the 95% prediction interval of the prediction? What is the rough 95% prediction interval of the prediction from this model if you wanted to estimate this?
2. Determine the model coefficients for the model:

𝑦𝑖 = 𝑏0 + 𝑏1 ∗ 𝑇𝑒𝑚𝑝 + 𝑏2 ∗ 𝐵𝑎𝑓𝑓𝑙𝑒𝑠 + 𝑒 . Copy the summary of the results here. Don’t forget to include the units when reporting the values. Construct a two-way scatter plot of temperature versus yield. On the graph, draw the least squares regression lines corresponding to baffles use. Is the Baffles use difference in yield response significant, adjusted for temperature?

1. Determine the model coefficients for the model:

𝑦𝑖 = 𝑏0 + 𝑏1 ∗ 𝑇𝑒𝑚𝑝 + 𝑏2 ∗ 𝑆𝑝𝑒𝑒𝑑 + 𝑏3 ∗ 𝐵𝑎𝑓𝑓𝑙𝑒𝑠 + 𝑒 . Copy the summary of the results here. Don’t forget to include the units when reporting the values.

1. Which factors significantly impact reactor yield? How do you know? Give an interpretation for the variables which are statistically significant (ignore the intercept term).
2. What is the predicted yield for a reactor without baffles operated at a temperature of 70C and a speed of 4000 RPM? Do this question by hand and show your work.

**Problem 2: [25 points/100 points]**

There is an experimental windmill that an engineer is developing to generate electricity. They collect data on the direct current flow output from the windmill and the wind speed corresponding to the output.

Use the WindPower dataset to explore the dependence of current flow on wind speed. Investigate a systematic method to determine if a linear model is the appropriate model to analyze the effect of x on the outcome of y. Explore transforming the variables to find an appropriate linear model.

**Problem 3: [25 points/100 points] Logistic Regression.**

Use the CHD dataset to explore the research question of the impact of sex and age on a binary outcome of 10-year CHD risk based on the following variables.

1. Outline the differences between a least-squares regression model and a logistic regression model.
2. Fit a logistic regression model to predict the characteristics of individuals most likely to have a 10-year risk of coronary heart disease (CHD). First, analyze a logistic regression model with TenYearCHD as the response and sex (Variable name: male) as the predictor. Produce and discuss the model results and comment on the Odds Ratio and significance for sex as a predictor of 10-year risk of CHD.

1. Fit a logistic regression model to predict 10-year risk of CHD as the response and sex and age as the independent variables. Produce and discuss the results and provide conclusions based on the model for the characteristics of those who are deemed at risk for CHD.