Math 2310
Correlation & Regression

In this lab assignment we will be using R to draw scatterplots, calculate correlations, and fit regression lines to data.

A group of researchers from Victoria, Australia studied how the chemical composition of cheddar cheese influenced how people rated the taste of the cheese. Three chemical compounds were measured – hydrogen sulfide, lactic acid, and acetic acid. We will refer to these three variables as the predictor variables, and we will refer to the taste score as the response variable. The data is stored in the file cheese.xls

When analyzing data with two or more numerical variables, we will often start by examining the relationships between the variables using scatterplots and correlations. Then, we will frequently be interested in creating a model that can make predictions for one of the variables based on one or more of the other variables. We can also assess the quality of the regression model through the coefficient of determination, r^2.

Goals for this assignment:

- □ Create scatterplots in R
- □ Calculate correlations in R
- □ Interpret scatterplots and correlations
- □ Use R to find the equation of a regression line
- □ Use a regression equation to make a prediction
- □ Interpret the r^2 value to assess the quality of a regression model
- □ Use R to find a regression equation using more than one predictor variable at once
- □ Use a regression model with transformed data to address a non-linear relationship

The **skill objectives** and **analysis objectives** in each activity specify what you should be submitting in your lab report.

Grading: there are two possible points for each skill objective and for each analysis objective.

Activity 1

Examine the relationships between each of the three predictor variables and the response variable.

Skill Objective: Using R, construct three scatterplots showing the relationship between the response variable and each of the three predictor variables. For each scatterplot, also find the corresponding correlation value, r.

Analysis Objective: Based on your scatterplots, comment briefly on whether there appears to be a relationship between each of the predictor variables and the response variable, and whether those relationships appear to be positive or negative. Which of the three predictor variables has the strongest relationship with the response variable?

Hints:

- R has a built in function cor(x, y) that will calculate the correlation between variables x and y
- You can read in the excel file by putting library(readxl) at the beginning of your file, and then doing something like:
  data <- read_excel(“path/on/your/computer/to/cheese.xls”)

Activity 2

Now we will look at how we could use information about hydrogen sulfide levels to predict taste ratings.

Skill Objective: Find the equation of the regression line predicting taste score based on hydrogen sulfide level.

Analysis Objective: What taste score would you predict for a cheese whose hydrogen sulfide measurement was 5.0?

Hints:

- You can calculate a linear regression model using specific columns of a dataframe (assume it’s called “data”) with the following:
  reg <- lm(formula = y_variable_name ~ x_variable_name, data = data)
- Once you have the regression, you can extract the coefficients using:
  slope = reg$coefficients[[2]]
  intercept = reg$coefficients[[1]]

Activity 3

Now we will assess the quality of this regression model

Skill Objective: Using R, find the value of the coefficient of determination, r^2, for the regression model predicting taste rating based on hydrogen sulfide levels.

Analysis Objective: In one sentence, explain what this r^2 value tells you about the quality of the regression model.

Hint: You can calculate r^2 manually, or you can extract it from a regression object via `summary(reg)$r.squared`

Activity 4

Occasionally in lab we will look at ideas that expand somewhat beyond what is covered in our book. This activity is one such time. In class, and in our book, we have seen how to construct a regression equation using one variable to predict another. But the same basic idea can be used to construct a regression equation using multiple variables to predict a variable. So, for example, we might have an equation of the form Y = a + bX + cW + dZ, where X, W, and Z are all variables we want to use to make our prediction.

While we will not be learning the mathematics to estimate such an equation by hand, it is quite simple to estimate an equation like this using R or other statistical software.

Skill Objective: Using R, Estimate the equation of the regression line predicting taste score based on all three predictor variables in a single equation.

Analysis Objective: Based on your regression equation, what taste score would you predict for a cheese whose hydrogen sulfide measurement was 5.0, whose acetic acid measurement was 6.1, and whose lactic acid measurement was 0.90? Also, based on the output from R, what proportion of the variability in taste scores can be explained by this model using all three predictor variables?

Hint:

- You can carry out a regression using multiple independent variables by doing something like: reg <- lm(formula = y_variable_name ~ x_variable_1 + x_variable_2 +… , data = data)
- Look into using the R function “predict” to make a prediction using your regression model

Activity 5
(adapted from problem 28 in section 3.4 of the textbook)

Polyester fiber ropes are increasingly being used as components of mooring lines for offshore structures in deep water. The authors of the paper “Quantifying the Residual Creep Life of Polyester Mooring Ropes” (*Intl. J. of Offshore and Polar Explor.*, 2005: 223-228) used the data contained in the file “ropes.txt”as a basis for studying how time to failure (hr) depended on load (% of breaking load).

In the data file, the first column is load (as a percentage out of 100), and the second column is the time to failure.

Hint:

- You can read in the data using:

  ```
  df <- read.delim("ropes.txt",  sep = " ", header=FALSE)
  ```

a) We will examine the relationship between the two variables.

Skill Objective: Construct a scatterplot of x = load versus y = time.

Analysis Objective: Would it be reasonable to characterize the relationship between the two variables to be linear?

b) We will try to address the nonlinearity through a transformation.

Skill Objective: Transform the response variable by computing y’ = log(y). Construct a scatterplot of x and y’.

Analysis Objective: Would it be reasonable to characterize the relationship between these two variables to be linear?

c) Finally, we will fit a regression model.

Skill Objective: Fit a straight line to the (x, y’) data.

Analysis Objective: Based on the linear fit, predict the value of failure time from a load of 85%.