CS3130/ECE3530: Probability and Statistics for Engineers **Due Fr 12/12**

Homework 8: Hypothesis Testing and Linear Regression

Instructions: This homework is due ELECTRONICALLY by 3:40PM on 12/12. Submit it via Canvas. You should turn in a single *.r file with your R code, and other answers written using comments, e.g.:

```
## this is a comment
```

Please make sure your name appears at the top of the page.

You may discuss the concepts with your classmates, but write up the answers entirely on your own.

IMPORTANT: To receive credit for this assignment, your code must run without errors! It should be self-contained, not depending on any external files or packages. You should test it finally by starting a fresh R session and running the command: `source("yourfilename.r")`

UPDATE (12/04): Question 3 is now extra credit. We won't cover this material in detail in class, but some notes are available online and in the book. Good luck.

1. **(50 points)** Here we are going to test a couple of hypotheses about the Old Faithful data in R. Remember, this is the `faithful` data frame that is built in to R. First split `faithful` into two separate data frames: (1) those entries with eruption times less than 3 minutes (`faithful$eruptions < 3`) and (2) those entries with eruption times greater than or equal to 3 minutes (`faithful$eruptions >= 3`). Answer the following about the entry **wait time** (`faithful$waiting`):
   (**Hint:** You might try using the R function `t.test` to double-check the answers you get, but you may not use it as the R commands that are asked for below.)

   (a) For the entries with short eruption times, you want to test the hypothesis that the associated waiting last on average less than 60 minutes. What is the null hypothesis? What is the alternative hypothesis?

   (b) Give R commands to compute the t statistic and the resulting p-value (one line of code for t and one line for p). What values did you get? Would you reject the null hypothesis at the $\alpha = 0.05$ level?

   (c) For the entries with long eruption times, you want to test the hypothesis that the associated waiting time last on average shorter than 80 minutes. What is the null hypothesis? What is the alternative hypothesis?

   (d) Give R commands to compute the t statistic and the resulting p-value to test the hypothesis you came up with in part (c) (again, one line of code for t and one line for p). What values did you get? Would you reject the null hypothesis at the $\alpha = 0.05$ level?

2. **(50 points)** In this exercise you will test hypotheses involving the correlation between two variables. We will use the `iris` data set, which is built in to R. Let's assume we have two random variables X and Y that are Gaussian distributed. The null hypothesis will be that these two random variables have zero correlation:

$$H_0 : \rho(X, Y) = 0$$

Under this null hypothesis, the sample correlation coefficient, r, can be transformed into the statistic

$$t = r\sqrt{\frac{n-2}{1-r^2}},$$

which will have a Student's t distribution with $n-2$ degrees of freedom. As usual, n denotes the sample size, which in the case of the iris data is $n = 150$. Remember, you can compute the sample correlation of two vectors in R using the `cor` command. Answer the following: (**Hint:** You might try using the R function `cor.test` to double-check the answers you get, but you may not use it as the R commands that are asked for below.)

(a) Say you hypothesize that irises with wide petals will also have skinny sepals, and skinny petals will coincide with wide sepals (`X= iris$Petal.Width`) and `Y = iris$Sepal.Width`). What is the alternative hypothesis here for $\rho(X, Y)$?

(b) For a significance level of $\alpha = 0.05$ what is the critical value for the t statistic? Give the R command (one line) for computing this. What does it return?

(c) What is the value for the t statistic above?

(d) Give the R command (one line) for computing the p-value. What value does it return?

(e) Now say you hypothesize that longer petals will be wider and shorter petals will be skinnier. What is the alternative hypothesis in terms of petal width and length? (`X = iris$Petal.Width`, and `Y = iris$Petal.Length`)

(f) Repeat steps (b), (c), and (d) for this hypothesis.

3. **(20 points - extra credit)** Write an R function `my.regression(x, y)` that computes the regression of an independent variable `x` and a dependent variable `y`. It should return the estimated slope and intercept.

(a) Run your regression command on the `faithful` data, with `waiting` as the `x` and `eruptions` as the `y` variables. What is the slope and intercept? Draw the resulting regression line over top of a scatterplot of the data.

(b) Say you are watching the Old Faithful geyser, and you time the interval between two eruptions to be 50 minutes. Based on your regression analysis, how long should you expect this next eruption to last?

(c) Run the R command `lm` on the same data and test that your `my.regression` function gives the same results.