Homework 8 **Due Date: See Canvas**

PLEASE COMPLETE YOUR CHOICE OF 4 OF THE 6 PROBLEMS. IF YOU SUBMIT MORE THAN 4 PROBLEMS, ONLY THE FIRST 4 PROBLEMS OF YOUR ASSIGNMENT WILL BE GRADED FOR CREDIT. PLEASE LIST THE NUMBERS OF THE PROBLEMS YOU CHOOSE TO COMPLETE AT THE TOP OF YOUR ASSIGNMENT.

1. The rowan (*Sorbus aucuparia*) is a tree that grows in a wide range of altitudes. To study how the tree adapts to its varying habitats, researchers collected twigs with attached buds from 12 trees growing at various altitudes in North Angus, Scotland. The buds were brought back to the laboratory and measurements were made on the dark respiration rate. The average altitude from the origin for the 12 trees is 433.33 meters. The average respiration rate computed for the 12 trees is 0.21 μl oxygen per hour per mg dry weight of tissue. The standard deviation of the altitude for the 12 trees is 214.62 meters and for respiration rate is 0.077 μl/hr-mg. The correlation between altitude from the origin and dark respiration rate was 0.887.

   (a) We will use linear regression to describe the relationship between altitude and dark respiration. Which of the two variables would most naturally be considered the explanatory variable and which would be the response variable?

   (b) Compute the equation of the least-squares regression line.

   (c) Compute a 95% confidence interval for the slope of the regression line.

   (d) Estimate the mean respiration rate for a tree growing at 300 meters above the origin.

   (e) If the value of the new explanatory variable lies within one s.d. of the mean of the values of explanatory variables, the prediction is called *interpolation*, otherwise it is called *extrapolation*. Do you think your estimate in part (d) involved interpolation or extrapolation? Explain.

   (f) Compute the s.e. of the mean respiration rate at 300 meters above the origin.

2. The data in `eruption.txt` are the interval waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming. Ignore the date variable and consider all observations as one sample of X and Y values. Answer the following questions:

   (a) Fit the regression model that predicts interval from the duration of the last eruption and report the estimated intercept and slope.

   (b) Calculate and report a 95% confidence interval for the slope.

   (c) Test whether the slope equals 0. Report your test statistic, p-value, and a one sentence conclusion.

   (d) Rangers at Yellowstone report an interval describing when they expect the next eruption to occur. Is it more appropriate to provide a confidence interval or a prediction interval? Briefly explain why.

3. It is generally thought that the percentage of fruit attacked by codling moth larvae is greater on apple trees bearing a small crop. Apparently the density of the flying moths is unrelated to the size of the crop on a tree, so the chance of attack for any particular fruit is increased if few fruits are on the tree. Data collected for a random sample of 10 trees gives a sample linear correlation coefficient of -0.77. Other summary statistics obtained from the sample are provided below. The data are available in `wormyfruit.txt`.

Variable	Sample Mean	Sample Standard Deviation
Crop Size (number of fruit)	126.702	54.924
Percentage of Wormy Fruits	41.539	12.070

   (a) Estimate the least-squares regression line for predicting percentage of wormy fruits from crop size. Show all your work in addition to the equation of the estimated model.

   (b) Write down the ANOVA table for the simple linear regression of percentage of wormy fruit on crop size. Compute the F statistic and find a p-value.

(c) Based on the F statistic and p-value computed in part (b), is there statistically significant evidence that the slope of the regression line is significantly different from 0?

(d) What proportion of the variability in the percentage of wormy fruit is explained by the regression of percentage of wormy fruit on crop size?

(e) Estimate the mean percentage of wormy fruits for trees with a crop size of 150 fruit.

(f) Provide a 95% confidence interval for the mean percentage of wormy fruits for trees with a crop size of 150 fruit.

(g) Is there statistically significant evidence from these data that trees with a crop size of 150 have, on average, 50% wormy fruit?

(h) Provide a 90% prediction interval for the average percentage of wormy fruit for the first tree with 150 fruit that was selected to be part of the study.

4. The data in `diversity.txt` are described in Chapter 8, problem 22 (both editions). Ecological theory suggests that the appropriate regression model is diversity = $\beta_0 + \beta_1 \times \log(\text{area})$. Use this model for all the questions below. The data come from an experimental study where the area of the patch was randomly assigned to plots of land.

   (a) Estimate the regression coefficients of the regression model suggested by ecological theory, then test whether the data provide evidence of a positive linear relationship between diversity and log(area). Report the estimates, the test statistic and the p-value for the test.

   (b) Does the mean diversity vary linearly with log(area), or does a model that allows the mean diversity to follow some other pattern appear to fit the data better? Provide the hypotheses, test statistic and a p-value in addition to your conclusion.

   (c) The investigators are interested in using this model to predict the number of species on new patches of forest. What is the area of a patch that has the smallest standard error for predictions of the mean number of species? What is the area of a patch that has the smallest standard error for predictions of number of species in an individual patch?

   (d) The investigators would really like to make predictions of diversity in an individual patch that has standard errors less than 20 species. Is that possible with this model and data? Explain why or why not.

   (e) The study that produced these data has attracted a lot of international attention. Imagine that this study can be repeated with 10 times as many plots (160 instead of the 16 here). All else about the system remains the same: the regression intercept, slope, and error variance (i.e. $\hat{\sigma}$) are the same as here. Will this study be able to predict diversity in an individual patch with a standard error less than 20 species? Briefly explain why or why not.

5. Manatees (a.k.a. sea cows) live off the coast of Florida and unfortunately, many manatees are killed or injured by powerboats every year. The file `manatee.txt` contains data on X = the number of Florida powerboat registrations (in 1000s) and Y = number of manatees killed near Florida. There is one point for each year from 1977 to 1990.

   (a) Use a software to construct a scatter-plot between number of manatees killed and number of Florida power boat registrations. Is the relationship roughly linear?

   (b) Compute the correlation coefficient between the two variables. Describe what this value says about the relationship between X and Y, in the context of the study.

   (c) Give the equation of the least-squares regression line for predicting the number of manatee deaths as a function of powerboat registrations.

   (d) Predict the number of manatee deaths for a year in which 600,000 powerboats are registered in Florida ($X = 600$).

(e) Provide a 95% prediction interval for the number of manatee deaths in a year in which 600,000 powerboats are registered in Florida.

(f) Examine the residual plot and the normal probability plot of the residuals from the simple linear regression model. Is there an indication that the assumptions of the model are violated? Explain why or why not.