Simple Linear Regression

Due: Tuesday 09/07 (11.00PM)
Submission: On Gradescope

The Homework contains two parts:
Part I consists of practice problems that you can work on to practice; you do not need to submit these. Some of these will be discussed during Thursday's office hours. Part II consists of the problems that you need to submit.

Use *R* and *R* Markdown to answer question 2 (Part II). For question 1 (Part II), you can write your answer in a piece of paper, take a picture (or scan it), and include it **inside** your *R* Markdown file with the following code:

```
![text here](path-to-image-here)
```

Part I: Practice Questions

You do not need to submit these questions.

1. Prove the following statements:
   (a) The sum of the y observations is the same as the sum of the fitted values.
   (b) The sum of the residuals weighted by the fitted values is zero.

2. **Muscle Mass**
   A person's muscle mass is expected to decrease with age.To explore this relationship in women, a nutritionist randomly selected 50 women from each 10-year age group, beginning with age 40 and ending with age 79. x is age and y is a measure of muscle mass. Assume that a simple linear regression model is appropriate. The data can be found on Moodle in the data set `muscle.txt`.
   (a) Obtain the estimated regression line. Plot the estimated regression function and the data. Does a linear regression function appear to give a good fit here? Does your plot support the anticipation that muscle mass decreases with age? Explain.
   (b) Obtain the following:
      i. a point estimate of the difference in the mean muscle mass for women differing in age by one year.
      ii. a point estimate of the mean muscle mass for women aged $X = 60$ years.
      iii. the value of the residual for the eighth case.
      iv. a point estimate for σ^2.

3. **Copier Maintenance**
   The Tri-City office Equipment Corporation sells an imported copier on a franchise basis and performs preventive maintenance and repair service on this copier. The data have been collected from

45 recent calls on users to perform routine preventive maintenance service. For each call, x is the number of copiers serviced and y is the total number of minutes spent by the service person. Assume that a simple linear regression model is appropriate. The data can be found on Moodle in the data set `copier.txt`.

(a) Obtain the estimated regression function.

(b) Plot the estimated regression function and the data. How well does the estimated regression function fit the data?

(c) Interpret $\hat{\beta}_0$ in your estimated regression function. Does $\hat{\beta}_0$ provide any relevant information here? Explain.

(d) Obtain a point estimate of the mean service time when $x = 5$ copiers are serviced.

(e) Estimate the change in the mean service time when the number of copiers serviced increases by one. Use and 90% confidence interval. Interpret your confidence interval.

(f) Conduct a test to determine whether or not there is a linear association between x and y here. Control the α risk at 0.01. State the alternatives, decision rule and conclusion. What is the p-value of your test?

(g) The manufacturer has suggested that the mean required time should not increase by more than 14 minutes for each additional copier that is serviced on a service call. Conduct a test to decide whether this standard is being Satisfied by Tri-City. Control the risk of a type I error at 0.05. State the alternatives, decision rule and conclusion.

(h) Obtain a 90% confidence interval for the mean service time on calls in which 6 copiers are serviced. Interpret your confidence interval.

(i) Obtain and 90% prediction interval for the service time on the next call in which six copiers are serviced. Is your prediction interval wider than the corresponding confidence interval? Should it be?

(j) Conduct an F test to determine whether or not there is a linear association between time spent and number of copiers serviced. Use α equal to 0.1. State the alternatives, decision rule and conclusion.

(k) By how much relatively is the total variation in the number of minutes spent on a call reduced when the number of copiers serviced is introduced into the analysis? Is this a relatively small or large reduction? What is the name of this measure?

(l) Calculate r and attach the appropriate sign.

Part II: Homework Questions – to be submitted

1. Consider the Simple Linear Regression model as defined in class:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \text{(Model I)}$$

where the ε random errors have mean zero, are homoscedastic and uncorrelated.

Now, assume that we interchange the response y_i with the predictor x_i and fit the following Simple Linear Regression model:

$$x_i = b_0 + b_1 y_i + \varepsilon_i^* \qquad \text{(Model II)}$$

where the ε^* random errors have mean zero, are homoscedastic and uncorrelated.

(a) **Derive** (i.e. show your work step-by-step) the Least-Squares estimators for b_0 and b_1.

(b) Let R_I^2 be the R^2 of model I and R_{II}^2 the R^2 of model II. Are they the same or not? Discuss.

2. The `StatGrades.csv` data set contains 4 `Quiz Scores` and a `Final Exam` score from an Introductory Statistics course (actual course I taught several years ago!). Our goal in this example is to investigate if the **average** of the `Quizzes` can be used to explain the variation in the `Final Exam` scores by fitting a linear regression model of the `Final Exam` score vs. the `Average Quiz` score.

   (a) Compute the new variable `Quiz Average` and add it to the data frame.

   (b) Obtain the estimated regression line.

   (c) Plot the estimated regression function and the data. How well does the estimated regression function fit the data?

   (d) Interpret $\hat{\beta}_0$ in your estimated regression function. Does $\hat{\beta}_0$ provide any relevant information here? Explain.

   (e) Obtain a point estimate of the mean `Final Exam` score for a student with `Quiz Average` equal to 60.

   (f) Estimate the difference in the mean `Final Exam` score for two students whose `Quiz Average` differs by 1. Use a 90% confidence interval.

   (g) Obtain a 90% confidence interval for the mean `Final Exam` score for students with `Quiz Average` equal to 85.

   (h) Obtain a 90% prediction interval for the mean `Final Exam` score for a new student with `Quiz Average` equal to 85. Is your prediction interval wider than the corresponding confidence interval? Should it be?

   (i) Conduct an F test to determine whether or not there is a linear association between `Final Exam` score and `Quiz Average`. Use α equal to 0.1. State the alternatives, decision rule and conclusion.

   (j) By how much relatively is the total variation in the `Final Exam` score reduced when the `Quiz Average` is introduced into the analysis? Is this a relatively small or large reduction? What is the name of this measure?

   (k) Calculate r (the correlation coefficient) and attach the appropriate sign.