

LINEAR REGRESSION MODELS W4315

HOMWORK 3 QUESTIONS

October 5, 2010

Instructor: Frank Wood

You are not allowed to use build-in regression functions in matlab. Pdf and cdf of common distributions are allowed. Attach relevant and concise code to your homework. Hand in a hard copy.

1. (25 points) ¹ Refer to **Copier maintenance** homework 1 problem 4. Find the data in “copier.txt”.

- Estimate the change in the mean service time when the number of copiers serviced increases by one. Use a 90 percent confidence interval. Interpret your confidence interval.
- Conduct a t test to determine whether or not there is a linear association between X and Y here; control the α risk at .10. State the alternatives, decision rule, and conclusion. What is the p-value of your test?
- Are your results in parts (a) and (b) consistent? Explain.
- 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 .05. State the alternatives, decision rule, and conclusion. What is the p-value of the test?
- Does b_0 give any relevant information here about the “start-up” time on calls-i.e., about the time required before service work is begun on the copiers at a customer location?

2. (25 points) ² The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a student’s grade point average(GPA) at the end of the freshman year(Y) can be predicted from the ACT test score(X). You can find the data in the file “GPA.txt”, the first column is Y and the second is X . Assume a simple linear regression model with normal error terms.

- Obtain a 95 percent interval estimate of the mean freshman GPA for students whose ACT test score is 28. Interpret your confidence interval.
- Mary Jones obtained a score of 28 on the entrance test. Predict her freshman GPA using

¹This is problem 2.5 in “Applied Linear Regression Models(4th edition)” by Kutner etc.

²This is problem 2.13 in “Applied Linear Regression Models(4th edition)” by Kutner etc.

- a 95 percent prediction interval. Interpret your prediction interval.
- c. Is the prediction interval in part (b) wider than the confidence interval in part (a)? Should it be?
- d. Determine the boundary values of the 95 percent confidence band for the regression line when $X_h = 28$. Is your confidence band wider at this point than the confidence interval in part (a)? Should it be?

3. (25 points) ¹ Refer to **Copier maintenance**.

- a. Prepare a dot plot for the number of copiers serviced X_i . What information is provided by this plot? Are there any outlying cases with respect to this variable?
- b. The cases are given in time order. Prepare a time plot for the number of copiers serviced. What does your plot show?
- d. Prepare residual plots of e_i versus \hat{Y}_i and e_i versus X_i on separate graphs. Do these plots provide the same information? What departures from regression model (2.1) can be studied from these plots? State your findings. And the model (2.1) is as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where:

β_0 and β_1 are parameters

X_i are known constants

ϵ_i are independent $N(0, \sigma^2)$

- e. Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality (use the approximation given in (3.6) of textbook). Does the normality assumption appear to be tenable here? Use Table B.6 of the textbook and set $\alpha = .10$.
- h. The file “copier_more.txt” provides two more variables into consideration (the third and fourth column), namely, mean operational age of copiers serviced on the call (X_2 , in months) and years of experience of the service person making the call (X_3). Plot the residuals against X_2 and X_3 on separate graphs to ascertain whether the model can be improved by including either or both of these variables. What do you conclude?

N.B. You don't need to do part c, f and g.

- 4. (25 points)** In normal probability plot (also referred to as Q-Q plot), we draw the expectation of the order statistics from normal distribution against the observed ordered residuals. If the error terms are indeed normal, then we should expect to see a linear

¹This is problem 3.4 in “Applied Linear Regression Models(4th edition)” by Kutner etc.

relationship. Now consider a simplified situation, $X_1, X_2, X_3, \dots, X_{10}$ i.i.d. $\sim N(0, 1)$.

- a. Write out the probability density functions of order statistics $X_{(1)}, X_{(5)}$ and $X_{(10)}$. ($X_{(i)}$ is the i th largest item in $X_1, X_2, X_3, \dots, X_{10}$.)
- b. Draw the pdf of $X_{(1)}, X_{(5)}$ and $X_{(10)}$ on one graph.
- c. What are the expected values of $X_{(1)}, X_{(5)}$ and $X_{(10)}$? (Use the approximation given in (3.6) of textbook). Mark them on the graph.