

Stat 328 Lab #2 Summer 2000

Below are some data taken from *Statistics and Data Analysis* by Tamhane and Dunlop (originally contributed by Prof. Susan Hughes of the School of Public Health at the University of Illinois, Chicago). These are the lengths of stay, x , and the reimbursed hospital cost, y , for $n = 33$ elderly people.

x	y	x	y
13	13,728	4	2,849
8	8,062	4	2,818
13	4,805	2	2,265
6	5,099	9	1,652
33	14,963	4	1,846
2	4,295	18	25,460
9	4,046	16	4,570
13	3,193	10	12,213
16	15,486	12	5,870
11	9,413	52	24,484
19	9,034	19	4,735
20	8,939	9	13,334
26	17,596	85	35,381
3	1,884	8	5,681
5	1,763	20	7,161
1	1,233	41	10,592
30	6,286		

It would be helpful for insurance executives and government (Medicare) officials to be able to say how y is related to x . This lab investigates the use of simple linear regression in modeling y as a function of x .

(a) To begin, enter the data into JMP (using one column for x and one for y) and (using the **FIT Y by X** procedure) plot y versus x and find the least squares line through the data. What is the sample correlation between y and x ? (Remember that in SLR, the square of the sample correlation between y and x turns out to equal R^2 , the coefficient of determination, and that the sign of the sample correlation between y and x is the same as the sign of the slope of the least squares line.) Now (by clicking on their row labels and going through the **Rows** menu, using the **Exclude/Include** or **Exclude/Unexclude** option) temporarily drop the two longest hospital stays from consideration and recompute the sample correlation. Comments?

Restore all 33 data pairs to consideration.

(b) What is the value of y on the least squares line for $x = 5$?

(c) What about the plot in (a) suggests that the usual SLR model relating y to x may not be such a good one for this situation?

(d) It is sometimes possible to essentially "change scale(s) of measurement" and turn a problem where the SLR model doesn't really fit, into one where it looks better. This is one such problem. Use JMP to create two new variables (2 new columns), $x' = \ln(x)$ and $y' = \ln(y)$. You can do this by going through the **Cols** menu choosing to add a new column via a "formula" and under the "transcendental functions" menu choosing the "natural log" (JMP-IN v. 3.2.6) or "log" (JMP v. 4.0). Plot y' versus x' . Is the problem you noted in part (c) "cured"? Explain.

(e) What is the least squares line through the (x', y') version of the data set? What fraction of the raw variation in y' is accounted for using an equation linear in x' ?

Notice that if

$$\ln y = y' = c_0 + c_1 x' = c_0 + c_1 \ln x$$

then

$$y = e^{\ln y} = e^{c_0 + c_1 \ln x} = e^{c_0} \cdot e^{c_1 \ln x} = e^{c_0} \cdot e^{\ln x^{c_1}} = e^{c_0} \cdot x^{c_1} \quad (*)$$

so that if y' is linearly related to x' there is a predictable (but nonlinear) relationship between x and y , and one can move back and forth from the original measurement scales to the "logged" ones in fairly simple ways.

(f) $x = 5$ has $x' = \ln(x) = 1.6094$. What y' and then y correspond to this value of x through the least squares line fit to the (x', y') version of the data set?

Consider making probability-based inferences based on the SLR for x' and y' ,

$$y'_i = \beta_0 + \beta_1 x'_i + \epsilon_i$$

where the ϵ_i are independent normal random variables with mean 0 and variance σ^2 . Use your JMP report (of the analysis of the (x', y') version of the data set) in what follows questions.

(g) What is a single number estimate of σ ? What does this measure in the context of the problem at hand?

(h) Verify that the "standard error" of b_1 printed out on the JMP report is indeed

$s / \sqrt{\sum_{i=1}^n (x'_i - \bar{x}')^2}$ as expected. (Note that you can get the sample variance of the x' values by running the **Distribution of Y** procedure on your x' column.)

(i) Use b_1 and the "standard error" of b_1 printed out on the JMP report and make 95% confidence limits for β_1 , the increase in mean y' that accompanies a unit increase in x' . Based on your interval, is it plausible that $\beta_1 = 1$? Explain. (Note that in view of equation (*) this is the possibility that reimbursed expenses are simply proportional to length of stay.) (You can actually

get the interval asked for here added to the JMP report by left-clicking on the triangle next to "parameter estimates" in JMP-IN v. 3.2.6 or by right-clicking on the "parameter estimates" table and going through the "columns" option in JMP v. 4.0.)

(j) Use the information on the JMP report and the Stat 328 Formula Sheet to find 95% two-sided limits for the mean value of y' that accompanies $x' = 1.6094$, namely

$\mu_{y'|x'=1.6094} = \beta_0 + \beta_1(1.6094)$. This is the average natural log of reimbursed expenses if $x' = 1.6094$ (i.e. if $x = 5$). Note that if you take these limits and raise e to these powers, you get limits for the center of the distribution of y for this x' (or x). As it turns out, assuming y' is normal makes y nonnormal, and this "center" of the distribution is a median, but not a mean. Do this exponentiation. What are your 95% limits for the median reimbursed expense for $x = 5$? (You can get JMP to show you limits for **all** $\mu_{y'|x'}$ values by left-clicking on the triangle next to "linear fit" and checking "Confidence Curves:Fit". Use this fact to check your calculations of the confidence limits for $\mu_{y'|x'=1.6094}$.)

(k) Use the information on the JMP report and the Stat 328 Formula Sheet to find 95% two-sided prediction limits for the natural logarithm of the reimbursed expenses for an additional hospital stay of $x = 5$ days. If you take these limits and raise e to these powers, you *do* get a prediction interval for the next reimbursed dollar figure for a stay of this length. What are those limits? (You can get JMP to show you prediction limits for **all** x values by left-clicking on the triangle next to "linear fit" and checking "Confidence Curves:Indiv". Use this fact to check your "hand" calculations of the prediction limits for y' .)

(l) If your manager asked you to use your analysis to make a prediction of y for $x = 365$ based on your analysis in this problem, you probably ought to either refuse or proceed with EXTREME caution. Why? Even if you weren't worried about this issue, any prediction limits you provided for $x = 365$ would be of little practical value. Why?

(m) It's worth knowing that some of the pain inflicted by parts (j) and (k) of this lab could have been circumvented (and perhaps additional understanding of the implications of using model (*) obtained in the process). Do the following. Go back to the (x, y) version of the data set and use the **FIT Y by X** procedure. In JMP-In v. 3.2.6 left click the triangle by "Fitting" and choose "Fit Transformed." In JMP 4.0 click on triangle on the "Bivariate Fit of y By x" bar and choose "Fit Special." Then choose the natural log options for both x and y . Add the confidence limits for the median reimbursed expense and prediction limits for an additional expense to the plot. You can adjust the limits on an axis by clicking on a number on that axis and filling in a dialog box. Make the vertical axis cover the range \$0 to \$100,000. Make a printout.

(n) Based on your plot from (m), if you are an insurance claims adjuster, would a new claim for a \$60,000 reimbursement on a 30 day hospital stay merit additional investigation? Explain.