

# Stat 328 Lab #5 Summer 2000

## Part #1 (Not to be turned in)

For the final model you arrived at in your odyssey of Lab #4, do the following:

a) Now that Vardeman has discussed what the single variable JMP "leverage plots" are, examine them for the continuous variables in your final model. Do they indicate that any of your continuous predictors are essentially redundant (with nearly uniformly small residuals as predicted from the other  $x$ 's)? (You'll see this in a relatively small "horizontal spread" in the leverage plot.) This would be, of course, indicative of multicollinearity in your predictors. From these plots, do you see a few cases that pretty much "drive the fitting" of the equation with the additional variable (these are data points corresponding to large horizontal distances from the center of the plots). Click on such a point in one of these plots and go to the JMP data table. Note that the corresponding "case" (row of the data table) has been highlighted. Can you discern anything "unusual" about the case, just looking through the data table?

b) Now that Vardeman has discussed the so-called "leverage" values  $h_{ii}$  add them to the data table, find a big value or two and examine the corresponding cases. Can you see any qualitative reason why these cases have "been flagged"? Do these cases correspond in any way to points that were brought to your attention through the JMP "leverage plots" (really a different technical usage of the word "leverage")? (These  $h_{ii}$  values can be saved in JMP-IN under the dollar sign in the lower left corner of the report.)

c) Now that Vardeman has discussed the "Cook's distance/influence" statistic, add those values to your data table. (Again look under the dollar sign in the lower left corner of the report in JMP-IN.) As in b) above find a big value or two and examine the corresponding cases. Do they match perfectly the cases with big  $h_{ii}$ ? Why, in general do they not? Can you see any qualitative reason why the cases you examine have been flagged? Do these cases correspond in any way to points that were brought to your attention through the JMP "leverage plots"?

## Part #2 (Your "Regular" Lab)

Below are some data taken from Vardeman's *Statistics for Engineering Problem Solving*. They are wood joint strengths (in psi stress at failure) recorded in some tests made by ISU engineering students Kotlers, MacFarland and Tomlinson. Three different Joint Types (Factor A) and three different Wood Types (Factor B) were studied. The students intended to test 2 specimens of each type, but circumstances conspired to make 2 samples to be of size 1 only.

		Wood Type		
		1 (Pine)	2 (Oak)	3 (Walnut)
Joint Type	1 (Butt)	829, 569	1169	1263, 1029
	2 (Beveled)	1348, 1207	1518, 1927	2571, 2443
	3 (Lap)	1000, 859	1295, 1561	1489

Enter these data into a JMP data table. You will need a "Joint" column, a "Wood" column and a "y" column. Also invent a 9-level qualitative variable "Cell" naming the 9 different Joint/Wood

combinations (number left to right, top to bottom in the table, beginning with the upper left cell, going across the row and ultimately ending with the lower right cell) and add its values to the data table.

First consider an analysis based on the single qualitative variable "Cell."

(a) After being sure that you have set "Cell" to be a nominal variable, run the JMP "Fit Model" routine with "Cell" as a single factor.

(b) What is the estimate of " $\sigma$ " provided by JMP? Use it to make a 95% two-sided confidence interval for  $\sigma$ . What, in the present context, does  $\sigma$  measure? Look on page 15 of the Stat 328 formula sheet and find the formula for a "pooled sample standard deviation." Compute it and compare it to the JMP estimate of  $\sigma$ .

(c) What is the  $p$ -value for testing whether there are any differences among the 9 mean responses? (Where is it on the JMP report and what hypothesis is being tested in terms of the MLR  $\beta$ 's?)

(d) What are  $\beta_0$  through  $\beta_8$  in a MLR regression analysis here? (What do they represent in terms of 9 mean joint strengths,  $\mu_{ij}$ ?)

(e) Suppose that you want to compare mean strengths for butt joints made with pine and oak woods. The difference in these means is an " $L$ " (a linear combination of the  $\beta$ 's). As such, you can make a confidence interval for it using the "custom test" facility in JMP to get an appropriate estimate and standard error. Make 95% two-sided limits for this difference. How does the  $\hat{L}$  compare to the difference in the sample means?

(f) Go to the "cell" factor part of the JMP report. And look at the "least squares means." What are they in simple terms? Why is it plausible that the ones for cells 2 and 9 are larger than the others? Use the "least squares means" and "standard errors" and make 95% two-sided limits for the mean strength of butt/pine joints. Then do the same for butt/oak joints. Then make 95% prediction limits for the next strength of a joint of each of these two types.

Now begin to consider an analysis of these data that recognizes the 2-Factor structure. We'll start with a "no-interactions" analysis. Use JMP and fit a model to these data using the 2 nominal factors "Joint" and "Wood."

(g) What estimate of " $\sigma$ " do you get for a no-interaction model of joint strength? How does this compare to the estimates in (b) above? What, in qualitative terms, does this suggest to you about a "no-interaction" description of  $y$ ? Open the "Lack of Fit" part of the JMP report. What is the "Mean Square for Pure Error"? What  $p$ -value is provided for testing whether the no-interaction model provides enough flexibility in form of response to describe the 9 different means? What does it suggest about the fit of a no-interactions model?

Now fit a model to these data that includes both main effects and interactions of the factors "Joint" and "Wood." You can do that in "Fit Model" by listing "Joint," "Wood," and "Joint\*Wood" in the "Effects in Model" part of the dialogue box. (So that we're talking about the output in exactly the same format, list the factors in exactly this order.)

- (h) What estimate of " $\sigma$ " do you now get, and how does it compare to the earlier ones?
- (i) What  $p$ -value does the first part of the JMP report provide for comparing the full model to a reduced model that has no interactions? How does this compare to things you've seen earlier in this analysis?
- (j) Using the estimates in the "parameter estimates" part of the JMP report, find a full model "MLR" estimate of the mean strength of butt/pine joints. Then use the parameter estimates part of the report to find an estimate of the average of the 3 butt means,  $\frac{1}{3}(\mu_{11} + \mu_{12} + \mu_{13}) = \mu_1$ . (Find estimates of the 3 different  $\mu$ 's and average them.) You should end up with  $b_0 + b_{A1}$ , which in turn should be the average of the 3 row 1 sample means. Use the JMP "custom test" facility to find a standard error for this average.
- (k) Use the JMP "custom test" facility to again make 95% confidence limits for the difference in mean strengths for butt joints made with pine and oak woods. (Verify that by proper choice of coefficients  $c$  here you get the same result as in part (e).)
- (l) Now look at the "Joint" effect report part of the printout. What are the "least squares means" here? Why is it plausible that two of the standard errors listed there are larger than the other one? What is the "beveled – lap" estimated  $\beta$  (in the parameter estimates table) in terms of the 3 "least squares means" printed out there? (Note, by the way, that since these data are "unbalanced" in the sense that the sample sizes are not all the same, there a "means" printed that are simply averages of  $y$ 's in a given row, and differ from the "least squares means.")
- (m) How do the "least squares means" and their standard errors in the "Joint\*Wood" effect part of the report compare to things you've seen before in this lab?
- (n) By clicking on the triangle in the "Joint\*Wood" effect part of the report, add the "plot effect" diagram to your report. What is plotted here and what about that plot (that, in all truth, should have been one of the very first things looked at in a real world analysis of these data) suggests strongly that a no-interactions analysis of these data is possibly ill-advised?
- (o) The plot in (n) should suggest to you that although a no-interactions analysis of all 3 Joint types and all 3 Wood types in this problem is probably not a good idea, there are smaller parts of the data set where such might well make sense. Consider, for example, limiting attention to Pine and Oak. Go into the JMP data table and exclude all walnut data points from the analysis. Make a "no-interactions" fit to the remaining 6 cells worth of data. Do you see anything in the JMP report that suggests that this is a bad description of the data *for only Pine and Oak*? What would be the practical advantage of using a no-interactions description of strength for these woods? (For example, can I say what "the effect" of changing from Lap to Beveled joints would be? Can that be sensibly done for the whole data set?)