

Stat 328 Lab #5 Summer 2002

1. Below are some data taken from *Statistics* by McClave and Sincich. These are weekly newspaper advertising expenditures (in dollars) x_1 , shelf space allocated (in ft²) x_2 and the resulting sales (in dollars) y , for a supermarket chain's store brand of canned vegetables.

x_1	x_2	y	x_1	x_2	y
201	75	2010	996	75	5005
205	50	1850	625	50	2500
355	75	2400	860	50	3005
208	30	1575	1012	50	3480
590	75	3550	1135	75	5500
397	50	2015	635	30	1995
820	75	3908	837	30	2390
400	30	1870	1200	50	4390
997	75	4877	990	30	2785
515	30	2190	1205	30	2989

(a) Under the JMP **Multivariate** menu select all 3 variables and have a look at the "scatterplot matrix". Ignore the contours on these plots and simply concentrate on the scatterplots. Note that the (x_1, x_2) points (that were under the control of the managers doing this study) are laid out on a nice pattern covering the (x_1, x_2) -region of interest. From the plots of y versus x_1 and x_2 , does it look promising that one may be able to predict y on the basis of these variables?

(b) Fit the model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$

to these data. You will need to look under the **Analyze>Fit Model** menu. y (sales) goes into the dialog box as **Y** and x_1 (advertise) and x_2 (space) go in as "effects." What fraction of the raw variability in sales is accounted for in fitting an equation linear in advertising dollars and shelf space to the data? What is an estimate of the standard deviation of sales for any fixed set of advertising/shelf space conditions? Give a 95% confidence interval for this standard deviation.

(c) If one wishes to do a formal significance test of $H_0 : \beta_1 = \beta_2 = 0$, a summary of an F test is readily on the JMP printout for this purpose. What are the observed value of the test statistic, the associated degrees of freedom, and the corresponding p -value for this test? Interpret the results of this significance test. Does it appear that as a pair, the variables x_1 and x_2 provide important explanatory power for predicting sales?

(d) Consider now the effect of advertising on sales. Go to the "Parameter Estimates" part of the JMP report and find 95% confidence limits for the increase in mean sales that accompanies a \$1 increase in advertising expenditure *if shelf space is held fixed*. (You can right click on the body of the table to add these limits.) Check/show that these limits are indeed given by the "hand" formula provided in class.

(e) If one wishes to do a formal significance test of $H_0 : \beta_1 = 0$, summaries of both a t test and an F test are readily available on the JMP report. (For the first, look under the "Parameter Estimates" part of the report and for the second, look under the "Effect Tests" part of the report.) Find the observed values of the test statistics, name the reference distributions and give the p -values for these tests. Verify by running a SLR of y on the variable x_2 alone (use the **Analyze>Fit Y by X** menu) and using the "Partial F Test" from the class notes, that the F value printed out the JMP report is what you expect it to be. Does it appear that (in the presence of the shelf space variable) the advertising expenditure adds in an "important" way to one's ability to predict/explain sales?

(f) The t and F tests in part (e) are completely equivalent. How are the observed values of the test statistics related, and how are the reference distributions related?

(g) Does it seem that (even after accounting for the advertising variable) the shelf space allocated to these vegetables remains an important determiner of sales? Explain.

(h) JMP will give predicted values and both confidence limits for mean sales at a fixed combination of advertising and space ($\mu_{y|x_1, x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$) and prediction limits for an additional sales value (y) at a fixed combination of x_1 and x_2 . These can be saved into the data table. Click the red triangle on the main "Response" bar, go to the "Save Columns" and check "Predicted Values," "Mean Confidence Interval" and "Indiv Confidence Interval." Give 95% confidence limits for mean sales at $x_1 = 820$ and $x_2 = 75$. Give 95% prediction limits for the next sales value at this set of conditions.

(i) Using steps similar to those in part (h) get s_m at $x_1 = 820$ and $x_2 = 75$ from JMP. Use this standard error, a tabled t value and some "hand calculations" to reproduce the confidence limits and prediction limits from (h). In order to deal with estimation of mean response and prediction at a set of conditions not included in the original data set, one must check the "Prediction Formula" and "StdErr Pred Formula" under the "Save Columns" menu and then add those conditions to the " x " columns of the data table. \hat{y} and s_m will then appear in the data table for the new set of conditions. Use this method to get \hat{y} and s_m when $x_1 = 500$ and $x_2 = 55$ and make 95% confidence limits for $\mu_{y|x_1, x_2}$ prediction limits for an additional sales figure y under these conditions.

(j) The "Profiler/Prediction Profile" and "Contour Profiler" options in JMP help in the understanding of the nature of a fitted equation. Activate these options. (You need to choose them from the "Factor Profiling" menu on the main "Response" bar.) Consider first what can be done with the "Profiler/Prediction Profiler." What you see is a plot of \hat{y} and confidence limits for $\mu_{y|x_1, x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, as a function of one of the variables, with the other(s) held fixed. The red numbers at the bottom of the plots say what value of that variable is being used in the plot(s) for the other variable(s). Let's set up the plots around the conditions $x_1 = 700$ and $x_2 = 50$. Alt-clicking on a variable name brings up a dialog box that allows you to set the value of that "x" variable. (You can also drag the vertical red line(s) around if you wish.) JMP seems to let you have only 3 sets of confidence limits (for $\mu_{y|x_1, x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$). The red y value is the fitted or predicted response \hat{y} at your choice of inputs. What fitted sales do you get for $x_1 = 700$ and $x_2 = 50$? Your profile plots should indicate that if you move away from $x_1 = 700$ and $x_2 = 50$ holding one variable fixed and varying the other, \hat{y} is a linear function of the variable you change. How is that consistent with the basic model we're using here?

(k) Now look at the "Contour Profiler." (Make sure the "Surface Plot" option is activated. It is on the main bar of this part of the JMP report.) Based on the "Surface Plot" graphic, how would you describe the geometry of the surface in 3-dimensional space that has been fit to these data?

(l) Click on the "Current X" for both advertising and space and set them to $x_1 = 700$ and $x_2 = 50$. What is the "Current Y"? How does it compare to your \hat{y} from (j)?

(m) Click on "Contour" and set it to 3000. The red curve produced is the set of (x_1, x_2) pairs that have $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 = 3000$. Clicking on the graph gives a set of cross-hairs that can be moved around to pick out (x_1, x_2) pairs on the graph. Find an advertising budget that combined with a shelf space of 70 produces a predicted sales of \$3000.

(n) Enter 2800 in the "Lo Limit" position and 3200 in the "High Limit" position. Print out the resulting contour plot. This shows those sets of (x_1, x_2) pairs that have predicted/fitted sales of less than \$2800 and those with predicted sales more than \$3200. How might such a plot be useful to a store manager?

(o) Find the "Plot Residual by Predicted" option in JMP and activate it. (It is on the main "Response" bar under "Row Diagnostics.") Under the MLR model, the $e_i = y_i - \hat{y}_i$ which are here plotted against \hat{y}_i are supposed to look like approximate versions of the supposedly "random noise" ϵ_i . In particular, they are not supposed to have any obvious patterns. What does the current plot indicate about how the fitted equation is doing as a description of sales in terms of advertising and shelf space? (Does it over-predict or under-predict "small," "moderate" and finally "large" sales volumes?) Does this cause you concern about the appropriateness of inferences based on the model in (b)?

(p) The plot referred to in (o) raises the possibility that a surface more complicated than the one fit thus far (perhaps allowing for some "curvature") might produce a better description of sales. Try fitting the model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}^2 + \beta_4 x_{2i}^2 + \beta_5 x_{1i} x_{2i} + \epsilon_i$$

to these data. You can do this by putting the advertising and space variables into the "effects" part of the **Fit Model** dialog box, highlighting one of them in both the original list of variables and effects part of the box and hitting the "Cross" button in order to make products. What fraction of the raw variability in y is accounted for by this more complicated MLR model (with now $k = 5$ predictor variables)?

(q) Notice the slight curvature introduced into the fitted (x_1, x_2, y) surface, now evident on the "surface plot" in the contour profiler. From a plot of e_i versus \hat{y}_i does the problem identified in (o) seem to be cured?

(r) Do a partial F test of $H_0: \beta_3 = \beta_4 = \beta_5 = 0$ in the model of (p) (report a p -value as best you can.). On the basis of this test, would you say the increase in R^2 you find moving from the model of (b) to the model of (p) is "statistically significant"?

(r) Redo (k) through (n) using the more complicated model of part (p). Is there much practical difference in the sales predictions you obtain on the basis of the two models used in this lab?

2. As practice with very simple Time Series applications of regression analysis, do Problems 3.16 and 4.5. (Dielman's data can be downloaded from the Duxbury Web site <http://www.duxbury.com/default.htm> under the "Data Library" heading. Lagged variables can be made in JMP by using the "lag" function in the "Row" functions list.)