Lab Four: Multiple Regression
Geog 301a

Introduction

As the previous lab demonstrated, the general form of the bivariate regression model is expressed as:

$$Y = a + bx$$

where Y is the dependent variable, a is a constant value, b is the slope, and the x term attached to the slope is the independent variable. Thus, the relationship being tested by the equation is a bivariate one between a dependent variable (Y), and a single independent variable (x).

Multiple regression is most easily described as an extension of the bivariate case, with the model expressed as:

$$Y = a + b_1x_1 + b_2x_2 + b_3x_3 \ldots b_kx_k$$

where all of the conditions associated with the bivariate model hold, yet each new b_kx_k term is an additional independent variable. Several assumptions are made:

1) All variables must be measured at the interval level and without error.

2) For each set of values for the k independent variables ($X_{2j}, \ldots, X_{kj}$), $E(\epsilon j) = 0$ (i.e., the mean value of the error term is constant).

3) For each set of values for the k independent variables, $VAR(\epsilon_j) = \sigma^2$ (i.e the variance of the error term is constant).

4) For any two sets of values for the k independent variables, $COV(\epsilon_j, \epsilon_h) = 0$ (i.e., the error terms are uncorrelated; thus there is no autocorrelation).

5) For each X_i, $COV(X_i, \epsilon) = 0$ (i.e., each independent variable is uncorrelated with the error term).

6) There is no perfect collinearity -- no independent variable is perfectly linearly related to one or more of the other independent variables in the model.

7) For each set of values for the k independent variables, ϵj is normally distributed.

Source: Berry, William B. and Feldman, Stanley (1989) "Multiple Regression in Practice." Sage University Paper series on Quantitative Applications in the Social Sciences, 07-050, 5th edition. Beverly Hills and London: Sage Publications.

While the potential predictive power of the multiple regression model can be great, all of its assumptions are very rarely met by researchers. When re-specification of the multiple regression model is not incorporated into the multiple regression model when it is warranted, the resulting analysis is rendered invalid. Fortunately, corrective techniques are available to researchers who have violated an assumption of the multiple regression model. As such, the analysis of multivariate data by this method remains very popular among social scientists today.

Instructions

The following data set shows the relationship between sales volume and the average price of a basket of groceries was examined for 15 grocery retailers. In this lab, the columns correspond to the following variables:

Column 1 - retailing centre ID
Column 2 - volume of sales in 000's, the dependent variable (y)
Column 3 - price of a typical shopping basket (x_1)
Column 4 - number of parking stalls (x_2)
Column 5 - number of customers with cheque cashing cards (x_3)

1	54	98	297	247
2	59	96	388	299
3	31	102	205	187
4	33	102	257	264
5	66	95	352	263
6	45	100	290	256
7	13	108	103	125
8	66	94	371	299
9	27	103	214	187
10	44	100	256	244
11	10	110	138	152
12	81	92	17	320
13	83	92	459	317
14	72	95	399	311
15	44	105	263	218

The syntax to do an extensive regression analysis is listed below. To run this program it has to be inserted into the SYNTAX window in SPSS for Windows and then executed.

```
1. TITLE 'GEOG 301: MULTIVARIATE LINEAR REGRESSION'.
DATA LIST FIXED /SHOPID 1-2 SALES 4-5 PRICE 7-9 PARKING 11-13 CHEQUES 15-17.
VARIABLE LABELS SHOPID 'GROCERY STORE IDENTIFICATION NUMBER'
SALES 'VOLUME OF SALES -- $000S PER YEAR'
PRICE 'PRICE OF AVERAGE ORDER'
PARKING 'NUMBER OF PARKING STALLS'
CHEQUES 'NUMBER OF CUSTOMERS WITH CHEQUE-CASHING CARDS'.
BEGIN DATA.
01 54 098 297 247
02 59 096 388 299
03 31 102 205 187
04 33 102 257 264
05 66 095 352 263
06 45 100 290 256
07 13 108 103 125
08 66 094 371 299
09 27 103 214 187
10 44 100 256 244
```

```
11 10 110 138 152
12 81 092 417 320
13 83 092 459 317
14 72 095 399 311
15 44 105 263 218
END DATA.
GRAPH/SCATTERPLOT(BIVAR)=sales WITH price/MISSING=LISTWISE
/TITLE='SALES AS A FUNCTION OF PRICE'.
GRAPH/SCATTERPLOT(BIVAR)=sales WITH parking/MISSING=LISTWISE
/TITLE='SALES AS A FUNCTION OF PARKING AVAILABILITY'.
GRAPH/SCATTERPLOT(BIVAR)=sales WITH CHEQUES/MISSING=LISTWISE
/TITLE='SALES AS A FUNCTION OF CUSTOMERS WITH CHEQUE CARDS'.
REGRESSION /VARIABLES SALES PRICE PARKING CHEQUES
/CRITERIA = DEFAULTS
/DESCRIPTIVES MEAN STDDEV VARIANCE CORR SIG
/STATISTICS=R CHA COLLIN OUTS TOL COEFF
/DEPENDENT SALES
/METHOD ENTER
/PARTIALPLOT
/RESIDUALS.
REGRESSION /VARIABLES SALES PRICE PARKING CHEQUES
/CRITERIA = DEFAULTS
/DESCRIPTIVES MEAN STDDEV VARIANCE CORR SIG
/STATISTICS=R CHA COLLIN OUTS TOL COEFF
/DEPENDENT SALES
/METHOD STEPWISE
/SCATTERPLOT=(*SRESID,*PRED)
/RESIDUALS.
```

I have already set up the program for you to access.

If you are using Internet Explorer right click on the spss program link and use the 'save target as' function to save the file on disk, then have SPSS use that file. So open SPSS and use the 'File' command to load the file, you will need to set file type to sps which denotes a syntax file. Once it is loaded use the 'Run' command to run all of the commands in the file.

To get an ascii version of the syntax right click on the text file link and save it.

Once you have obtained an error-free output, you are asked to examine it very carefully, and to answer each of the following questions:

1) Briefly discuss the distinction between a `stepwise' and a `saturated' multiple regression model.

2) How many multiple regression models might have been generated by this particular data set?

3) With respect to both the saturated and the stepwise multiple regression model, state the value

of each of the following calculations and briefly describe what each of the statistics means (provide a formula if it helps):

a) multiple r

b) r^2

c) degrees of freedom (df)

4) Write the equations for both the stepwise and saturated models in the general form:

$Y = a + b_1x_1 + b_2x_2 + b_3x_3 \ldots b_kx_k$

State which equation pertains to the saturated model, and which derives from the stepwise model.

5) Assess the strength of the relationship between volume of sales and the three independent variables.

a) Present your argument in the form of a null and research hypotheses, being certain to state whether the hypotheses examined pertains to the saturated or stepwise model.

b) Perform a t-test by reading the computer output for each of the hypotheses, setting α=0 .05. Determine whether or not you would accept or reject H_o for each of the calculated values.

6) Calculate the predicted values of y for store numbers 8 and 11 using both of the equations. The saturated model provides a more accurate prediction for y than does the stepwise model. Reconcile this with the notion that the stepwise method is preferred in most instances to the saturated model.

7) Examine the values of r^2 for each of the models developed by SPSS. What is the percentage gain in explanatory power obtained by adding two independent variables? Is it worth it?

8) In any regression analysis one should check for violations of the underlying assumptions of the regression model. Using the partial residual plots, the histogram of residuals, the normal probability plot and the standardized scatterplot what can you conclude about violations of the assumptions?

9) What does the VIF statistics on the output tell you?