MULTIPLE LINEAR REGRESSION VIEWPOINTS
VOLUME 17, NUMBER 2, FALL 1990

Testing Assumptions in Multiple Regression: Comparison of Procedures Available in SAS and SPSSX

Paula L. Woehlke

Patricia B. Elmore

Southern Illinois University Carbondale

Debra L. Spearing

State University of New York at Albany

That the use of multiple linear regression requires satisfying several assumptions has seldom been disputed. However, assessing whether one has met important assumptions is not always easy, and given the limited time available to instructors in a typical multiple regression course, the techniques available for checking assumptions are often not taught, or mentioned only briefly. The purpose of this paper is to compare the most easily available techniques for checking assumptions from two of the most popular statistics packages in use today, SAS (SAS Institute, 1985) and SPSS^X (SPSS, Inc., 1985). It is hoped that the attached examples will make the multiple regression course instructor's job easier by providing concrete examples of computer input and output that illustrate the testing of assumptions.

A condition that should be met for the use of multiple regression, but which is not, strictly speaking, an assumption, is that there be an absence of multicollinearity. Multicollinearity is defined as the existence of substantial correlation among a set of independent variables, and its presence creates three distinct problems:

- the substantive interpretation of partial regression coefficients,
- the sampling stability of these coefficients

and

- computational accuracy of the regression analysis.

Thus, although absence of multicollinearity is not a regression assumption, failure to assure that predictor variables are not multicollinear can result in faulty interpretations of analyses, regression equations that cannot be used for prediction, or both.

In terms of actual theoretical assumptions for using multiple regression analyses, errors of the prediction or residuals from estimated values of the regression provide the basis for assessing the adequacy of the model (Cohen & Cohen, 1983). Specifically, it is assumed that errors

(1) are normally distributed

(2) are independent of one another (that is, errors associated with one observation Y_i are not correlated with errors associated with any other observation Y_j)

(3) are identically distributed (that is, are sampled from the same distribution and have constant variances, also known as the assumption of homoscedasticity)

(4) have a mean of zero

and

(5) are uncorrelated with the independent variables (X's).

In addition to these assumptions about errors, it is further assumed that

(6) the independent variables, (X's) are fixed and measured without error

(7) the regression of Y on X is linear

and

(8) Y is a random variable composed of two components: a fixed component, a + bX, and a random error e_i.

Two conditions under which these assumptions about residuals fail to be met occur when

- the regression of Y on X (or X's) is curvilinear (so that condition 7 above is not met)

and

- there are one or more extreme residual values, known as "outliers, which not only make relatively large contributions to error or residual variance (thus reducing R^2) but also exert a disproportionately strong pull on the regression.

To illustrate the use of SAS and $SPSS^X$ to test these assumptions, we used the (in)famous Longley data set. This data set has multicollinearity and some cases of univariate outliers through which to illustrate the diagnostic procedures available in both SAS and

SPSS^X. The following pages provide annotated output from these two packages, which we will describe in the next section.

Description of Output

The first assumption about errors is that the residuals are normally distributed. This assumption can be assessed by examining the residual scatterplot in Figure 4.SAS and the normal probability plot and statistical analyses shown in Figure 6.SAS; similar plots and statistics are produced by SPSS^X, as shown in Figure 4.SPSS^X, Figure 5.SPSS^X, and Figure 6.SPSS^X. If residuals are normally distributed, the plus signs (+'s) and the asterisks (*'s) will coincide in the SAS normal probability plot (or the asterisks [*'s] and dots [•'s] in the SPSS^X normal probability plot). Also, a statistical test for normality is provided in SAS in Figure 6.SAS; in this case, W:NORMAL = 0.948682, p=.471. It should be noted that SPSS^X's CONDESCRIPTIVE procedure routinely does not provide a comparable statistical test. All of these plots and tests from both SAS and SPSS^X indicate that the assumption about normally distributed residuals has been met.

That residuals are independent of one another or errors associated with one observation are not correlated with errors associated with any other observation is the second assumption to be tested. The Durbin-Watson D statistic shown in Figure 3.SAS and Figure 3.SPSS^X tests for nonindependence of errors when the order of cases is meaningful. For this data set, the Durbin-Watson D statistic is irrelevant. The residual scatterplots in Figure 4.SAS and Figure 5.SPSS^X show that the residuals are independent.

The third assumption is that residuals are identically distributed. This means that the errors are sampled from the same distribution and have constant variance, also known as homoscedasticity. Examination of the residual scatterplots in Figure 4.SAS and Figure 5.SPSS^X indicates that the assumption of homoscedasticity has been met.

Assumption 4, that the residuals have a mean of zero, can be determined by examining Figure 6.SAS or Figure 6.SPSS^X. For this data set, the mean is -7.421E-10 (Figure 6.SAS), which is considered zero for our purposes, or .000 (Figure 6.SPSS^X).

The correlation matrix showing the correlations between all of the independent variables and the residual should be used to assess assumption 5, that the residuals are uncorrelated with the independent variables. Examination of the correlation matrix for this data set, as found in Figure 7.SAS or Figure 7.SPSS^X indicates a correlation between each of the six independent variables and the residual equal to zero.

That the regression of Y on X is linear, assumption 7, can be determined by creating bivariate scatterplots for all predictors with the criterion. One example is shown in Figure 5.SAS and another in Figure 1.SPSS^X; both show the relation between Y and X_1. All six predictors in this data set are linearly related to the criterion Y.

Figure 1.SAS and Figure 2.SPSS^X show a check for multicollinearity. Low tolerance value and high condition number with large variance proportion for two or more variables may indicate multicollinearity. Variables X_5 and X_6 in this data set may be multicollinear with previous terms in the model.

Figure 2.SAS has two indices to check for outliers. A studentized residual value in excess of ±3.00 may indicate a univariate outlier (Tabachnick & Fidell, 1989, p. 67). Also, a data point with a Cook's Distance value greater than 1.00 is suspected of being an outlier. Cook's Distance is discussed in depth in Tabachnick & Fidell (1989), p. 130, and Kleinbaum, Kupper & Muller (1988), p. 201. Also note that in Figure 3.SPSS^X a similar casewise plot appears, as well as a listing and a histogram of standardized residuals.

Discussion

Although the output of the regression modules and related descriptive statistics procedures for SAS and SPSS^X are quite similar, there are a few differences worth noting. First, SPSS^X includes a histogram of standardized residuals to make the spotting of outliers somewhat easier; the program also has a normal probability plot that is a little easier to read than that provided in SAS. SPSS^X also provides standard errors for the skewness and kurtosis values for the variables analyzed in the CONDESCRIPTIVE module; these values are not printed in the SAS output. On the other hand, SAS provides a statistical test of normality when requested through PROC UNIVARIATE, as well as stem and leaf diagrams and boxplots of distributions, through the same PROC. It is also easy to obtain Cook's D values through SAS's PROC REG; it is somewhat more difficult to get similar statistics from SPSS^X, requiring the use of a RESIDUALS subcommand. In most other respects, output is comparable for the data and regression analyses shown here. For more advanced regression applications, it is somewhat easier to obtain leverage (partial regression residual) plots for general linear hypotheses, used in assessing degree of fit, nonfitting points, and multicollinearity (Sall, 1990) from SAS (via an option in PROC REG) than from SPSS^X, which produces "partial regression plots" through a PARTIALPLOT subcommand. It should be noted, however, that some anomalies recently have been detected in SAS's regression and GLM procedures for models using different types of intercept terms (see Uyar & Erdem, 1990). Finally, although it is somewhat more difficult to obtain several diagnostic statistics from SPSS^X, the package supplements its regression module with an extensive and flexible MANOVA procedure that allows one to easily build advanced regression models. With these advantages and disadvantages in mind, it should be possible for the reader to choose which computer package is most appropriate for a particular regression analysis.

References

Cohen, J. & Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Erlbaum.

Draper, N. R. & Smith, H. (1981). *Applied Regression Analysis*, 2nd ed. New York: Wiley.

Kleinbaum, D. G., Kupper, L. L. & Muller, K. E. (1988). *Applied Regression Analysis and Other Multivariable Methods*, 2nd ed. Boston: PWS—Kent.

Pedhazur, E. J. (1982). *Multiple Regression in Behavioral Research*, 2nd ed. New York: Holt, Rinehart & Winston.

SAS Institute, Inc. (1985). *SAS User's Guide: Statistics, Version 5 Edition*. Cary, NC: Sas Institute, Inc.

Sall, J. (1990). Leverage plots for general linear hypotheses. *The American Statistician*, *44*(4), 308-315.

SPSS, Inc. (1986). *SPSS X User's Guide, Edition* 2. New York: McGraw-Hill.

Tabachnick, B. G. & Fidell, L. S. (1989). *Using Multivariate Statistics*, 2nd ed. New York: Harper & Row.

Uyar, B. & Erdem, O. (1990). Regression procedures in SAS: Problems? *The American Statistician*, *44*(4), 296-301.

```
SAS(R) LOG   CMS SAS 5.18       VM/CMS CMS USER GA0157

NOTE: COPYRIGHT (C) 1984,1986 SAS INSTITUTE INC., CARY, N.C.  27512, U.S.A.
NOTE: CMS SAS RELEASE 5.18 AT SOUTHERN ILLINOIS UNIVERSITY (01150002).

NOTE: CPUIB   VERSION = FF   SERIAL = 023435   MODEL = 3081 .

  you have questions, call the Computing Affairs Help Desk at 453-4361

  1 DATA AERA; INPUT Y X1 X2 X3 X4 X5 X6;
  2 OPTIONS NODATE LS=74;
  3 CARDS;

NOTE: DATA SET WORK.AERA HAS 16 OBSERVATIONS AND 7 VARIABLES.
NOTE: THE DATA STATEMENT USED 0.06 SECONDS AND 24K.

 20 PROC REG;
 21      MODEL Y=X1 X2 X3 X4 X5 X6/STB P R CLI DW TOL VIF COLLIN;
 22      OUTPUT OUT=NEW
 23           P=PRED
 24           R=RES;
NOTE: THE DATA SET WORK.NEW HAS 16 OBSERVATIONS AND 9 VARIABLES.
NOTE: THE PROCEDURE REG USED 0.18 SECONDS AND 40K
      AND PRINTED PAGES 1 TO 2.

 25 PROC PLOT;
 26      PLOT PRED*RES Y*X1 Y*X2 Y*X3 Y*X4 Y*X5 Y*X6;
NOTE: THE PROCEDURE PLOT USED 0.23 SECONDS AND 52K
      AND PRINTED PAGES 3 TO 9.

 27 PROC UNIVARIATE NORMAL PLOT;
NOTE: THE PROCEDURE UNIVARIATE USED 0.23 SECONDS AND 64K
      AND PRINTED PAGES 10 TO 18.

 28 PROC CORR;
NOTE: THE PROCEDURE CORR USED 0.11 SECONDS AND 76K
      AND PRINTED PAGES 19 TO 21.
NOTE: SAS USED 76K MEMORY.
```

SAS

DEP VARIABLE: Y
ANALYSIS OF VARIANCE

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	6	184172402	30695400.32	330.285	0.0001
ERROR	9	836424.06	92936.00616		
C TOTAL	15	185008826			

ROOT MSE	304.8541	R-SQUARE	0.9955
DEP MEAN	65317	ADJ R-SQ	0.9925
C.V.	0.4667301		

PARAMETER ESTIMATES

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
INTERCEP	1	-3482258.63	890420.38	-3.911	0.0036
X1	1	15.06187236	84.91492578	0.177	0.8631
X2	1	-0.03581918	0.03349101	-1.070	0.3127
X3	1	-2.02022980	0.48839968	-4.136	0.0025
X4	1	-1.03322687	0.21427416	-4.822	0.0009
X5	1	-0.05110411	0.22607320	-0.226	0.8262
X6	1	1829.15146	455.47850	4.016	0.0030

VARIABLE	DF	STANDARDIZED ESTIMATE	TOLERANCE	VARIANCE INFLATION
INTERCEP	1	0	.	0
X1	1	0.04628202	0.007378307	135.53244
X2	1	-1.01374635	0.000559124	1788.51348
X3	1	-0.53754258	0.02974518	33.61889060
X4	1	-0.20474069	0.27863456	3.58893019
X5	1	-0.10122111	0.002505317	399.15102
X6	1	2.47966438	0.001317557	758.98060

COLLINEARITY DIAGNOSTICS

NUMBER	EIGENVALUE	CONDITION NUMBER	VAR PROP INTERCEP	VAR PROP X1	VAR PROP X2	VAR PROP X3
1	6.861393	1.000000	0.0000	0.0000	0.0000	0.0000
2	0.082103	9.141721	0.0000	0.0000	0.0000	0.0143
3	0.045681	12.255735	0.0000	0.0000	0.0003	0.0008
4	0.010688	25.336607	0.0000	0.0003	0.0011	0.0646
5	0.0001292	230.424	0.0000	0.4568	0.0157	0.0056
6	6.246E-06	1048.088	0.0001	0.5046	0.3284	0.2253
7	3.664E-09	43275.045	0.9999	0.0383	0.6546	0.6893

FIGURE 1. SAS MULTICOLLINEARITY
Low tolerance value and high condition number with large variance proportion for two or more variables may indicate multicollinearity. Variables X_5 and X_6 may be multicollinear with previous terms in the model.

NUMBER	VAR PROP X4	VAR PROP X5	VAR PROP X6
1	0.0004	0.0000	0.0000
2	0.0919	0.0000	0.0000
3	0.0634	0.0000	0.0000
4	0.4267	0.0000	0.0000
5	0.1154	0.0097	0.0000
6	0.0000	0.8306	0.0002
7	0.3020	0.1597	0.9998

OBS	ACTUAL	PREDICT VALUE	STD ERR PREDICT	LOWER95% PREDICT	UPPER95% PREDICT	RESIDUAL
1	60323.0	60055.7	198.6	59232.6	60878.8	267.3
2	61122.0	61216.0	229.1	60353.3	62078.7	-94.0139
3	60171.0	60124.7	183.4	59319.9	60929.6	46.2872
4	61187.0	61597.1	186.0	60789.3	62405.0	-410.1
5	63221.0	62911.3	239.2	62034.7	63787.8	309.7
6	63639.0	63888.3	185.3	63081.2	64695.4	-249.3
7	64989.0	65153.0	213.7	64310.8	65995.3	-164.0
8	63761.0	63774.2	216.6	62928.2	64620.1	-13.1804
9	66019.0	66004.7	206.1	65172.2	66837.2	14.3048
10	67857.0	67401.6	175.3	66606.1	68197.1	455.4
11	68169.0	68186.3	182.9	67382.1	68990.5	-17.2689
12	66513.0	66552.1	211.9	65712.2	67391.9	-39.0550
13	68655.0	68810.5	186.5	68002.1	69619.0	-155.5
14	69564.0	69649.7	145.7	68885.3	70414.0	-85.6713
15	69331.0	68989.1	186.2	68181.0	69797.1	341.9
16	70551.0	70757.8	253.0	69861.6	71653.9	-206.8

OBS	STD ERR RESIDUAL	STUDENT RESIDUAL	-2 -1 0 1 2	COOK'S D
1	231.3	1.1560	| |** |	0.141
2	201.1	-0.4676	| | |	0.041
3	243.5	0.1901	| | |	0.003
4	241.5	-1.6979	| ***| |	0.244
5	189.0	1.6304	| |*** |	0.614
6	242.1	-1.0300	| **| |	0.009
7	217.4	-0.7547	| *| |	0.079
8	214.6	-0.0614	| | |	0.001
9	224.6	0.0637	| | |	0.000
10	249.4	1.8258	| |*** |	0.235
11	243.9	-0.0708	| | |	0.000
12	219.2	-0.1782	| | |	0.004
13	241.1	-0.6451	| *| |	0.036
14	267.8	-0.3199	| | |	0.004
15	241.4	1.4163	| |** |	0.170
16	170.1	-1.2154	| **| |	0.467

UM OF RESIDUALS	-1.18744E-08
UM OF SQUARED RESIDUALS	836424.1
REDICTED RESID SS (PRESS)	2886893

URBIN-WATSON D	2.559
FOR NUMBER OF OBS.)	16
ST ORDER AUTOCORRELATION	-0.348

FIGURE 2.SAS OUTLIERS

Check STUDENTIZED RESIDUAL for values in excess of ±3.00 for univariate outliers. See Tabachnick & Fidell (1989), p. 67.

COOK'S DISTANCE values greater than 1.00 are suspected of being outliers. See Tabachnick & Fidell (1989), p. 130 and Kleinbaum, Kupper, & Muller (1988), p. 201.

FIGURE 3.SAS INDEPENDENCE OF RESIDUALS

The Durbin-Watson D statistic tests for nonindependence of errors when the order of cases is meaningful. Tables are found in Draper & Smith (1981), pp. 164-166. It is irrelevant for this data set.

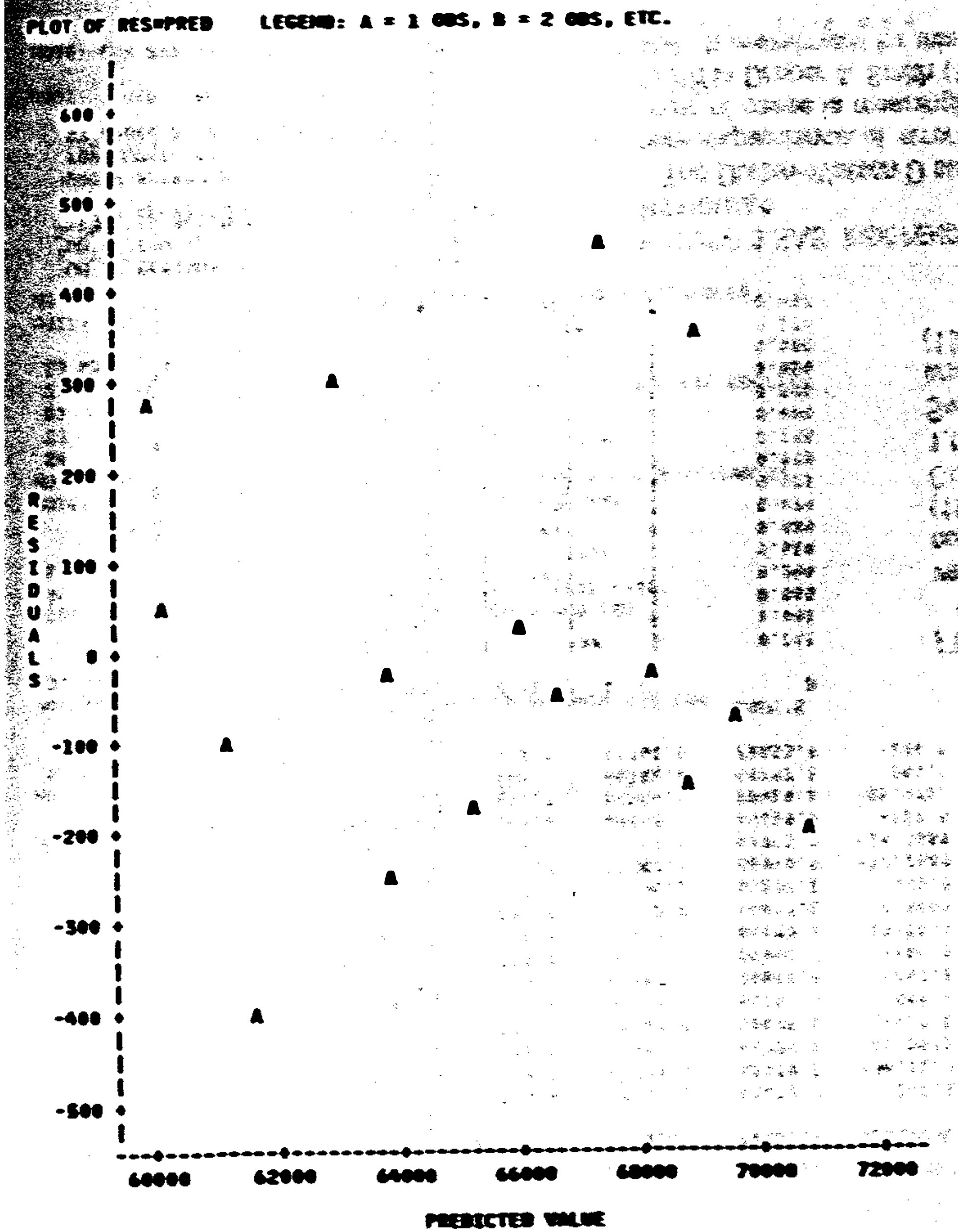


FIGURE 4.SAS NORMALITY, LINEARITY, HOMOSCEDASTICITY, AND INDEPENDENCE OF RESIDUALS

Examine the residual scatterplot to assess all four assumptions. All four assumptions are met in this data set.

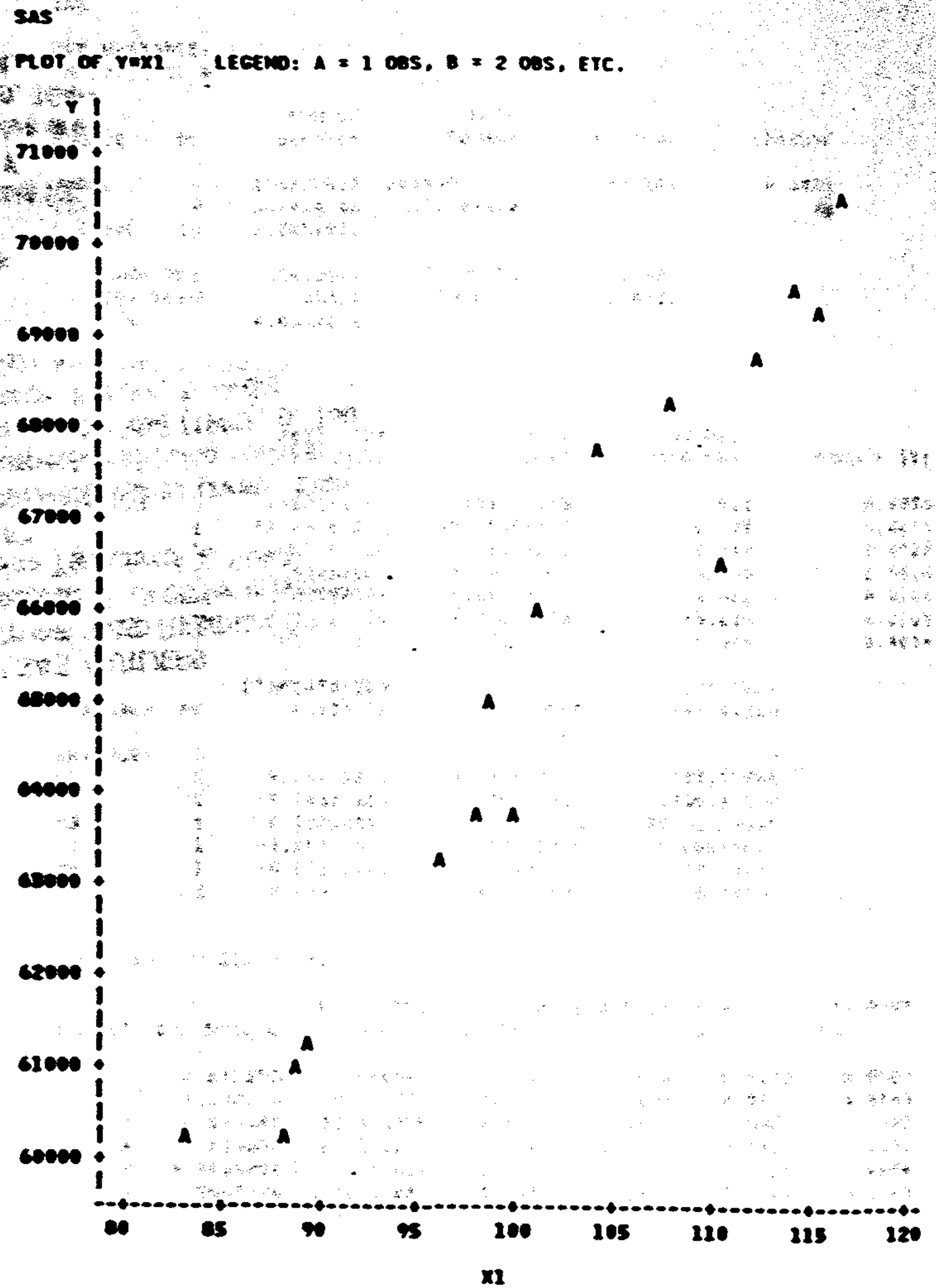


FIGURE 5.SAS LINEARITY

Use bivariate scatterplots to assess linearity of predictor - criterion association.

```
UNIVARIATE

VARIABLE=RES                    RESIDUALS

                    MOMENTS

N                      16  SUM WGTS           16
MEAN          -7.421E-10  SUM         -1.187E-08
STD DEV          236.139  VARIANCE       55761.6
SKEWNESS        0.464739  KURTOSIS     -0.298894
USS               836424  CSS             836424
CV                -99999  STD MEAN       59.0347
T:MEAN=0      -1.257E-11  PROB>|T|             1
SGN RANK              -6  PROB>|S|      0.776105
NUM ^= 0              16
W:NORMAL        0.948682  PROB<W           0.471


               QUANTILES(DEF=4)                        EXTREMES

100% MAX     455.394      99%      455.394       LOWEST     HIGHEST
 75% Q3      212.077      95%      455.394     -410.115     46.2872
 50% MED     -28.162      90%       375.97     -249.311      267.34
 25% Q1     -161.924      10%     -297.552     -206.758     309.715
  0% MIN    -410.115       5%     -410.115     -164.049     341.932
                           1%     -410.115      -155.55     455.394

 RANGE       865.509
 Q3-Q1       374.001
 MODE       -410.115

STEM LEAF                              #   BOXPLOT
   0 5                                 1      |
   0 333                               3   +--+--+
  -0 422221100000                     12   *-----*
     ----+----+----+----+
   MULTIPLY STEM.LEAF BY 10**+03

                        NORMAL PROBABILITY PLOT
     750+                                                          ****
        |                                   ++++++++++++++++++++++
    -250+     ************************* **
         +----+----+----+----+----+----+----+----+----+----+
             -2        -1         0        +1        +2
```

FIGURE 6. SAS NORMALITY OF RESIDUALS

If residuals are normally distributed, the plus signs (+'s) and asterisks (*'s) should coincide in the normal probability plot. A statistical test for normality is also provided. In this case, W:NORMAL = 0.948682, p = .471.

SKEWNESS can be detected by observing the stem and leaf and the boxplot as well as the skewness index of .464739 indicating the residuals are positively skewed. The kurtosis value of -.298894 indicates the residuals are platykurtic. Statistical tests of skewness and kurtosis are discussed in Tabachnick & Fidell (1989), pp. 72-73.

For these data,

$$z_{SKEWNESS} = \frac{S-0}{\sqrt{\frac{6}{N}}} = \frac{.464739}{\sqrt{\frac{6}{16}}} = .76$$

$z_{SKEWNESS} < 1.96$ ∴ not skewed.

$$z_{KURTOSIS} = \frac{K-0}{S_K} = \frac{K-0}{\sqrt{\frac{24}{N}}} = \frac{-.298894}{\sqrt{\frac{24}{16}}} = -.24$$

$z_{KURTOSIS} > -1.96$ ∴ no kurtosis

SAS

VARIABLE	N	MEAN	STD DEV	SUM	MINIMUM	MAXIMUM
Y	16	65317.0	3511.97	1045072	60171.0	70551.0
X1	16	101.7	10.79	1627	83.0	116.9
X2	16	387698.4	99394.94	6203175	234289.0	554894.0
X3	16	3193.3	934.46	51093	1870.0	4896.0
X4	16	2606.7	695.92	41707	1456.0	3594.0
X5	16	117424.0	6956.10	1878784	107600.0	130081.0
X6	16	1954.5	4.76	31272	1947.0	1962.0
PRED	16	65317.0	3504.02	1045072	60055.7	70757.8
RES	16	-7.421E-10	236.14	-1.187E-08	-410.1	455.4

FIGURE 7. SAS CORRELATION OF ERRORS AND INDEPENDENT VARIABLES

Use the correlation matrix to determine association between each independent variable and the residual from the multiple regression equation.

SAS

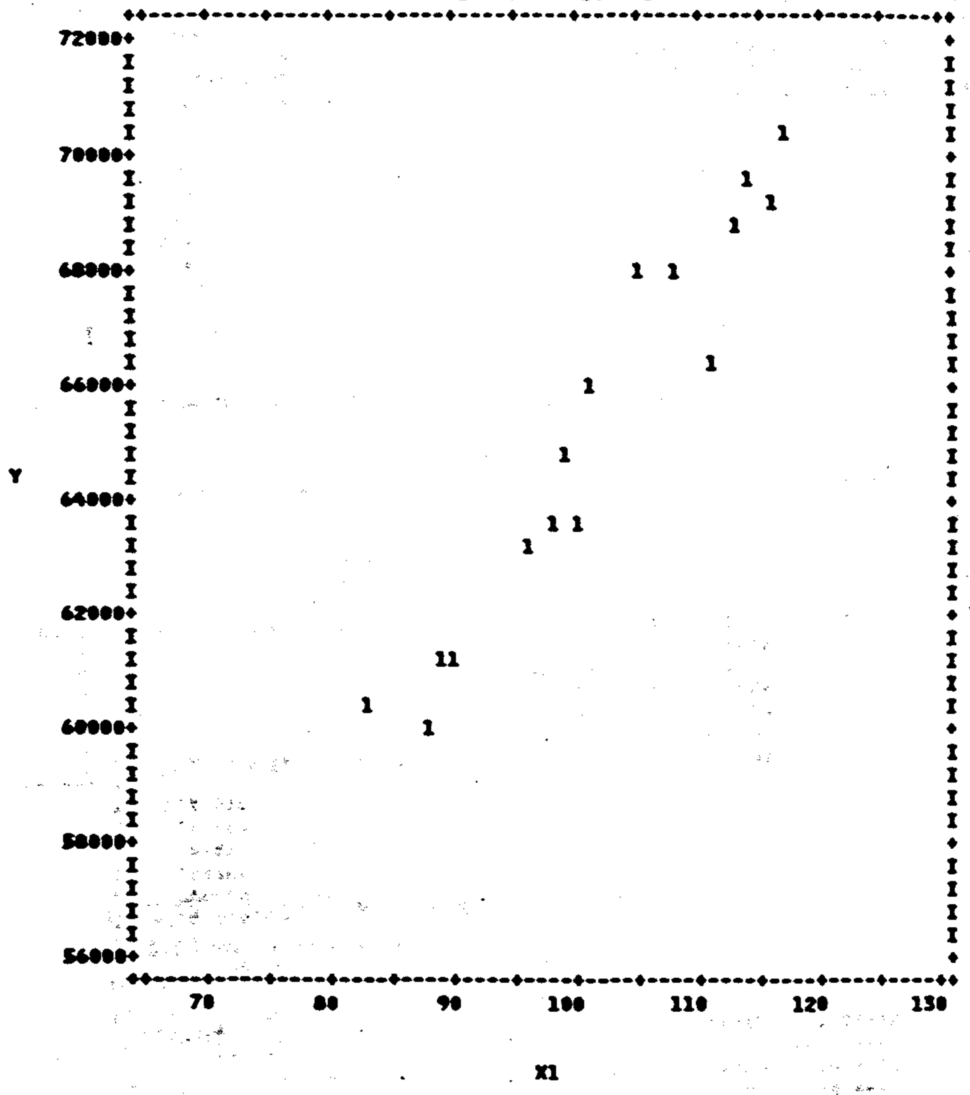
PEARSON CORRELATION COEFFICIENTS / PROB > |R| UNDER H0:RHO=0 / N = 16

	Y	X1	X2	X3	X4	X5
Y	1.00000	0.97090	0.98355	0.50250	0.45731	0.96839
	0.0000	0.0001	0.0001	0.0473	0.0749	0.0001
X1	0.97090	1.00000	0.99159	0.62063	0.46474	0.97916
	0.0001	0.0000	0.0001	0.0103	0.0697	0.0001
X2	0.98355	0.99159	1.00000	0.60426	0.44644	0.99109
	0.0001	0.0001	0.0000	0.0132	0.0830	0.0001
X3	0.50250	0.62063	0.60426	1.00000	-0.17742	0.68655
	0.0473	0.0103	0.0132	0.0000	0.5109	0.0033
X4	0.45731	0.46474	0.44644	-0.17742	1.00000	0.36442
	0.0749	0.0697	0.0830	0.5109	0.0000	0.1652
X5	0.96839	0.97916	0.99109	0.68655	0.36442	1.00000
	0.0001	0.0001	0.0001	0.0033	0.1652	0.0000
X6	0.97133	0.99115	0.99527	0.66826	0.41725	0.99395
	0.0001	0.0001	0.0001	0.0047	0.1078	0.0001
PRED	0.99774	0.97310	0.98578	0.50364	0.45834	0.96257
PREDICTED VALUE	0.0001	0.0001	0.0001	0.0467	0.0742	0.0001
RES	0.06724	-0.00000	-0.00000	-0.00000	-0.00000	-0.00000
RESIDUALS	0.8046	1.0000	1.0000	1.0000	1.0000	1.0000

	X6	PRED	RES
Y	0.97133	0.99774	0.06724
	0.0001	0.0001	0.8046
X1	0.99115	0.97310	-0.00000
	0.0001	0.0001	1.0000
X2	0.99527	0.98578	-0.00000
	0.0001	0.0001	1.0000
X3	0.66826	0.50364	-0.00000
	0.0047	0.0467	1.0000
X4	0.41725	0.45834	-0.00000
	0.1078	0.0742	1.0000
X5	0.99395	0.96257	-0.00000
	0.0001	0.0001	1.0000
X6	1.00000	0.97353	-0.00000
	0.0000	0.0001	1.0000
PRED	0.97353	1.00000	-0.00000
PREDICTED VALUE	0.0001	0.0000	1.0000
RES	-0.00000	-0.00000	1.00000
RESIDUALS	1.0000	1.0000	0.0000

```
For IBM VM/XA           Southern Illinois University      License Number 19077
This software is functional through June 30, 1991.

Try the new SPSS-X Release 3.1 features:

* Interactive SPSS-X command execution          * The new RANK procedure
* Online Help                                   * Improvements in:
* Nonlinear Regression                          *    REPORT and TABLES
* Time Series and Forecasting (TRENDS)          *    Simplified Syntax
* Macro Facility                                *    Matrix I/O

See SPSS-X User's Guide, Third Edition, for more information on these features.

   1  0  DATA LIST FREE / Y X1 X2 X3 X4 X5 X6
   2  0  SET WIDTH =80
   3  BEGIN DATA
  19  END DATA

Preceding task required .02 seconds CPU time;  .03 seconds elapsed.

  20  PLOT PLOT=Y WITH X1/

There are 1,749,224 bytes of memory available.
The largest contiguous area has 1,749,224 bytes.

PLOT requires 15000 bytes of workspace for execution.


                                                                     Page    2
17-Oct-90 SPSS-X RELEASE 3.1 FOR IBM VM/CMS
09:47:37  Southern Illinois University    IBM 3081         IBM VM/XA

* * * * * * * * * * * * * * * * *  P L O T  * * * * * * * * * * * * * * * *

Data  Information

     16 unweighted cases accepted.

Size of the plots

    Horizontal size is 65
      Vertical size is 40

Frequencies and symbols used (not applicable for control or overlay plots)

          1 - 1    11 - B    21 - L    31 - V
          2 - 2    12 - C    22 - M    32 - W
          3 - 3    13 - D    23 - N    33 - X
          4 - 4    14 - E    24 - O    34 - Y
          5 - 5    15 - F    25 - P    35 - Z
          6 - 6    16 - G    26 - Q    36 - *
          7 - 7    17 - H    27 - R
          8 - 8    18 - I    28 - S
          9 - 9    19 - J    29 - T
         10 - A    20 - K    30 - U
```

PLOT OF Y WITH X1



16 cases plotted.

FIGURE 1.SPSS[X]. LINEARITY
Use bivariate scatterplots to assess linearity of predictor-criterion association.

17-Oct-90 SPSS-X RELEASE 3.1 FOR IBM VM/CMS Page
09:47:37 Southern Illinois University IBM 3081 IBM VM/XA

Preceding task required .05 seconds CPU time; .44 seconds elapsed.

```
21  REGRESSION VARS=Y X1 X2 X3 X4 X5 X6/STATISTICS=ALL F/DEP=Y
22           /METHOD=ENTER
23           /RESID=DEFAULTS
24           /CASEWISE=ALL DEFAULTS
25           /SCATTERPLOT=(*ZRESID,*ZPRED)
26           /SAVE PRED (YHAT) RESID (ERR)
```

There are 1,749,184 bytes of memory available.
The largest contiguous area has 1,740,832 bytes.

* * * * M U L T I P L E   R E G R E S S I O N * * * *

Listwise Deletion of Missing Data

Equation Number 1   Dependent Variable.. Y

Beginning Block Number 1. Method: Enter

Variable(s) Entered on Step Number
1.. X6
2.. X4
3.. X3
4.. X1
5.. X5
6.. X2

Multiple R	.99774		
R Square	.99548	R Square Change	.99548
Adjusted R Square	.99247	F Change	330.28534
Standard Error	304.85407	Signif F Change	.0000

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	6	184172401.94449	30695400.32408
Residual	9	836424.05551	92936.00617

F = 330.28534   Signif F = .0000

AIC	187.82004
PC	.01155
CP	7.00000
SBC	193.23696

Var-Covar Matrix of Regression Coefficients (B)
Below Diagonal: Covariance   Above: Correlation

	X6	X4	X3	X1	X5	X2
X6	207460.663	-.54937	-.82410	.10628	.30816	-.00160
X4	-53.61674	.04591	.61857	-.34881	-.18091	.46860
X3	-183.32591	.06473	.23853	-.55500	-.75826	.94561
X1	7204.91263	-6.34671	-23.01719	7210.54462	.65918	-.64942
X5	39.96940	-.00915	-.08372	12.65424	.05111	-.83321
X2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

* * * * M U L T I P L E   R E G R E S S I O N * * * *

Equation Number 1   Dependent Variable.. Y

XTX Matrix

	X6	X4	X3	X1	X5	X2
X6	758.98060	-28.67217	-131.6399	59.74668	213.64593	-934.0346
X4	-28.67217	3.58893	6.79454	-7.69306	-7.15017	37.54356
X3	-131.6399	6.79454	33.61889	-37.46332	-87.83677	231.87218
X1	59.74668	-7.69306	-37.46332	135.53244	153.31803	-319.7367
X5	213.64593	-7.15017	-87.83677	153.31803	399.15102	-703.9908
X2	-934.0346	37.54356	231.87218	-319.7367	-703.9908	1788.5135
Y	2.47966	-.20474	-.53754	.04628	-.10122	-1.01375

	Y
X6	-2.47966
X4	.20474
X3	.53754
X1	-.04628
X5	.10122
X2	1.01375
Y	.00452

Equation Number 1   Dependent Variable..   Y

---------------------- Variables in the Equation ----------------------

Variable	B	SE B	95% Confdnce Intrvl B		Beta
X6	1829.151465	455.470499	798.788692	2859.514237	2.479664
X4	-1.033227	.214274	-1.517948	-.548506	-.204741
X3	-2.020230	.480400	-3.125065	-.915394	-.537543
X1	15.061872	84.914926	-177.028016	207.152560	.046282
X5	-.051104	.226073	-.562517	.460308	-.101221
X2	-.035819	.033491	-.111581	.039943	-1.015746
(Constant)	-3482258.635	890420.3836	-5496527.183	-1467990.086	

------------------------- Variables in the Equation -------------------------

Variable	SE Beta	Correl	Part Cor	Partial	Tolerance	VIF	F
X6	.617463	.971329	.090087	.001140	.001318	758.981	16.127
X4	.042460	.457307	-.100074	-.049004	.278635	3.589	23.252
X3	.129953	.502498	-.092709	-.009509	.029745	33.619	17.110
X1	.260926	.970099	.003975	.059022	.007378	135.532	.031
X5	.447780	.968391	-.005066	-.075137	.002505	399.151	.051
X2	.947855	.983662	-.023971	-.335004	5.591E-04	1788.513	1.144
(Constant)							15.294

------ in -------

Variable	Sig F
X6	.0038
X4	.0009
X3	.0025
X1	.8631
X5	.8262
X2	.3127
(Constant)	.0036

Number	Eigenval	Cond Index	Variance Proportions Constant	X1	X2	X3	X4	[illegible]
1	6.86139	1.000	.00000	.00000	.00000	.00004	.00035	.0[illegible]
2	.08210	9.142	.00000	.00000	.00001	.01428	.09191	.0[illegible]
3	.04568	12.256	.00000	.00000	.00026	.00004	.06357	.0[illegible]
4	.01069	25.337	.00000	.00034	.00107	.06464	.42672	.0[illegible]
5	.00013	230.424	.00000	.45677	.01564	.00559	.11540	.0[illegible]
6	.00001	1048.080	.00015	.50456	.32839	.22534	.00000	.8[illegible]
7	.00000	43275.05	.99985	.03833	.65463	.60926	.30205	.1[illegible]

	X6
1	.00000
2	.00000
3	.00000
4	.00000
5	.00000
6	.00016
7	.99984

End Block Number   1   All requested variables entered.

* *

Summary table

Step	MultR	Rsq	F(Eqn)	SigF		Variable	BetaIn
1					In:	X6	.9713
2					In:	X4	.0630
3					In:	X3	-.3910
4					In:	X1	-.0224
5					In:	X5	-.5002
6	.9977	.9955	330.285	.000	In:	X2	-1.0137

FIGURE 2. SPSS^X. MULTICOLLINEARITY
Low tolerance value and high condition number with large variance proportion for two or more variables may indicate multicollinearity. Variables X_5 and X_6 may be multicollinear with previous terms in the model.

*** * * *   M U L T I P L E   R E G R E S S I O N   * * * ***

Equation Number 1   Dependent Variable..   Y

Casewise Plot of Standardized Residual

*: Selected   M: Missing

```
         -3.0      0.0      3.0
 Case #  O:........:........:O        Y        *PRED       *RESID
     1   .         .   *    .   60323.00  60055.6600   267.3400
     2   .        *.        .   61122.00  61216.0139   -94.0139
     3   .         *        .   60171.00  60124.7128    46.2872
     4   .     *   .        .   61187.00  61597.1146  -410.1146
     5   .         .   *    .   63221.00  62911.2854   309.7146
     6   .       * .        .   63639.00  63888.3112  -249.3112
     7   .       * .        .   64989.00  65153.0490  -164.0490
     8   .         *        .   63761.00  63774.1004   -13.1004
     9   .         *        .   66019.00  66004.6952    14.3048
    10   .         .    *   .   67857.00  67401.6059   455.3941
    11   .         *        .   68169.00  68186.2689   -17.2689
    12   .         *        .   66513.00  66552.0550   -39.0550
    13   .       * .        .   68655.00  68810.5500  -155.5500
    14   .        *.        .   69564.00  69649.6713   -85.6713
    15   .         .   *    .   69331.00  68989.0685   341.9315
    16   .       * .        .   70551.00  70757.7578  -206.7578
 Case #  O:........:........:O        Y        *PRED       *RESID
         -3.0      0.0      3.0
```

*** * * *   M U L T I P L E   R E G R E S S I O N   * * * ***

Equation Number 1   Dependent Variable..   Y

Residuals Statistics:

	Min	Max	Mean	Std Dev	N
*PRED	60055.6563	70757.7500	65317.0000	3504.0206	16
*RESID	-410.1145	455.3940	.0000	256.1309	16
*ZPRED	-1.5015	1.5527	.0000	1.0000	16
*ZRESID	-1.3453	1.4958	.0000	.7746	16

Total Cases =   16

Durbin-Watson Test =   2.59949

* *

Outliers - Standardized Residual

Case #	*ZRESID
10	1.49581
4	-1.34528
15	1.12162
5	1.01594
1	.87694
6	-.81781
16	-.67822
7	-.53812
13	-.51024
2	-.30839

Histogram - Standardized Residual

```
NExp N      (* = 1 Cases,     . : = Normal Curve)
 0   .01   Out
 0   .02  3.00
 0   .06  2.67
 0   .14  2.33
 0   .29  2.00
 0   .53  1.67 .
 1   .88  1.33 :
 3  1.29  1.00 :**
 0  1.70   .67 .
 0  2.00   .33 .
 5  2.12   .00 *:***
 2  2.00  -.33 *:
 4  1.70  -.67 *:**
 0  1.29 -1.00 .
 1   .88 -1.33 :
 0   .53 -1.67 .
 0   .29 -2.00
 0   .14 -2.33
 0   .06 -2.67
 0   .02 -3.00
 0   .01   Out
```

FIGURE 3.SPSS^X. OUTLIERS
Check STUDENTIZED RESIDUAL for values in excess of ±3.00 for univariate outliers. See Tabachnick & Fidell (1989), p. 67. ~~Also check~~ plot of STANDARDIZED RESIDUAL. Also note that the Durbin-Watson D statistic tests for <u>nonindependence</u> of errors when the order of cases is meaningful. Tables are found in Draper & Smith (1981), pp. 164-166. It is irrelevant for this data set.

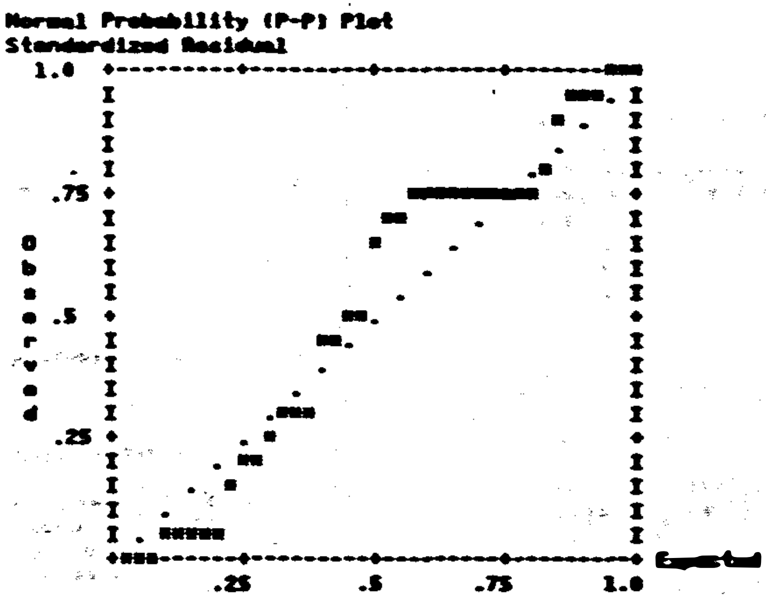


FIGURE 4. SPSS^X. NORMALITY OF RESIDUALS

If residuals are normally distributed, the dots (•'s) and asterisks (*'s) should coincide in the NORMAL PROBABILITY PLOT.

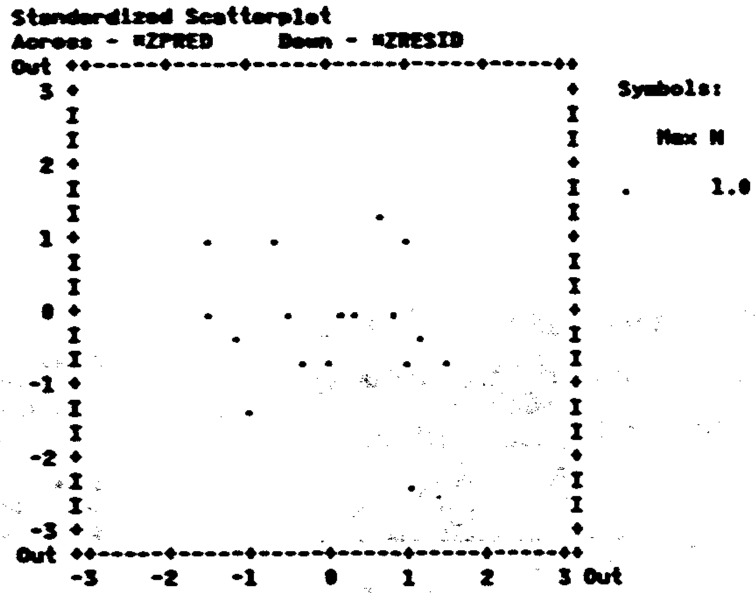


FIGURE 5. SPSS^X NORMALITY, LINEARITY, HOMOSCEDASTICITY, AND INDEPENDENCE OF RESIDUALS

Examine the standardized scatterplot of predicted and residual values to assess all four assumptions. All are met in this data set.

```
17-Oct-90 SPSS-X RELEASE 3.1 FOR IBM VM/CMS                    Page 14
09:47:39  Southern Illinois University   IBM 3081        IBM VM/XA

Preceding task required .19 seconds CPU time;  1.83 seconds elapsed.

  27  CONDESCRIPTIVE YHAT ERR/
  28  STATISTICS ALL

>Warning # 11003
>The new default column-style printing cannot be used for this DESCRIPTIVES, as
>there are too many statistics to print on one line per variable.  Old style
>printing will be used instead.

There are 1,748,296 bytes of memory available.
The largest contiguous area has 1,746,720 bytes.

148 bytes of memory required for the DESCRIPTIVES procedure.
4 bytes have already been acquired.
144 bytes remain to be acquired.

Number of valid observations (listwise) =        16.00

Variable  YHAT        Predicted Value

Mean         65317.000              S.E. Mean       876.005
Std Dev       3504.021              Variance   12278160.130
Kurtosis        -1.290              S.E. Kurt         1.091
Skewness         -.101              S.E. Skew          .564
Range        10702.090              Minimum     60055.65997
Maximum     70757.75783             Sum        1045072.000

Valid observations -      16        Missing observations -        0
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Variable  ERR         Residual

Mean              .000              S.E. Mean        59.035
Std Dev        236.139              Variance      55761.604
Kurtosis         -.299              S.E. Kurt         1.091
Skewness          .465              S.E. Skew          .564
Range          865.509              Minimum      -410.11462
Maximum      455.39409              Sum  -7.61614415908E-09

Valid observations -      16        Missing observations -        0
```

FIGURE 6.SPSS^X. NORMALITY OF RESIDUALS

SKEWNESS index of .465 indicates the residuals are positively skewed. The KURTOSIS value of -.299 indicates the residuals are platykurtic. Statistical tests of skewness and kurtosis are discussed in Tabachnick & Fidell (1989), pp. 72-73. For these data,

$$z_{SKEWNESS} = \frac{S - 0}{\sqrt{\frac{6}{N}}} = \frac{.464739}{\sqrt{\frac{6}{16}}} = .76$$

or, using reported SE_{SKEW},

$$z_{SKEWNESS} = \frac{.465}{.564} = .82;$$

$z_{SKEWNESS} < 1.96$ ∴ not skewed.

$$z_{KURTOSIS} = \frac{K - 0}{s_K} = \frac{K - 0}{\sqrt{\frac{24}{N}}} = \frac{-.298894}{\sqrt{\frac{24}{16}}} = -.24$$

or, using reported $SE_{KURTOSIS}$,

$$z_{KURTOSIS} = \frac{-.299}{1.091} = -.27$$

$z_{KURTOSIS} > -1.96$ ∴ no kurtosis

```
29  PEARSON CORR Y X1 X2 X3 X4 X5 X6 YHAT ERR/
30  OPTIONS 6

PEARSON CORR problem requires 1,872 bytes of workspace.

- - - -  P E A R S O N   C O R R E L A T I O N   C O E F F I C I E N T S  - - -

VARIABLE              VARIABLE              VARIABLE              VARIABLE
PAIR                  PAIR                  PAIR                  PAIR
--------              --------              --------              --------

Y            .9709    Y            .9836    Y            .5025    Y            .4573
WITH     N(    16)    WITH     N(    16)    WITH     N(    16)    WITH     N(    16)
X1       SIG .000     X2       SIG .000     X3       SIG .024     X4       SIG .037

Y            .9604    Y            .9713    Y            .9977    Y            .0672
WITH     N(    16)    WITH     N(    16)    WITH     N(    16)    WITH     N(    16)
X5       SIG .000     X6       SIG .000     YHAT     SIG .000     ERR      SIG .402

X1           .9916    X1           .6206    X1           .4647    X1           .9792
WITH     N(    16)    WITH     N(    16)    WITH     N(    16)    WITH     N(    16)
X2       SIG .000     X3       SIG .005     X4       SIG .035     X5       SIG .000

X1           .9911    X1           .9731    X1           .0000    X2           .6043
WITH     N(    16)    WITH     N(    16)    WITH     N(    16)    WITH     N(    16)
X6       SIG .000     YHAT     SIG .000     ERR      SIG .500     X3       SIG .007

X2           .4464    X2           .9911    X2           .9953    X2           .9858
WITH     N(    16)    WITH     N(    16)    WITH     N(    16)    WITH     N(    16)
X4       SIG .042     X5       SIG .000     X6       SIG .000     YHAT     SIG .000

X2           .0000    X3          -.1774    X3           .6866    X3           .6683
WITH     N(    16)    WITH     N(    16)    WITH     N(    16)    WITH     N(    16)
ERR      SIG .500     X4       SIG .255     X5       SIG .002     X6       SIG .002

X3           .5036    X3           .0000    X4           .3646    X4           .4172
WITH     N(    16)    WITH     N(    16)    WITH     N(    16)    WITH     N(    16)
YHAT     SIG .023     ERR      SIG .500     X5       SIG .083     X6       SIG .054

X4           .4583    X4           .0000    X5           .9948    X5           .9626
WITH     N(    16)    WITH     N(    16)    WITH     N(    16)    WITH     N(    16)
YHAT     SIG .037     ERR      SIG .500     X6       SIG .000     YHAT     SIG .000

X5           .0000    X6           .9735    X6           .0000    YHAT         .0000
WITH     N(    16)    WITH     N(    16)    WITH     N(    16)    WITH     N(    16)
ERR      SIG .500     YHAT     SIG .000     ERR      SIG .500     ERR      SIG .500

SIG IS 1-TAILED, "." IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED.
```

FIGURE 7.SPSS^X. CORRELATION OF ERRORS AND INDEPENDENT VARIABLES
Use the correlation matrix to determine association between each independent variable and the residual from the multiple regression equation.