

Graphical Methods for the interpretation of Linear Regression Models
Simin Kavoossi & Tim Cooley
Boehringer Mannheim Corporation Research and Development

ABSTRACT

The SAS output provided by REG procedure supplies a great deal of information in the form of summary statistics and estimates. Not only is this information difficult to communicate to non-statisticians it also can obscure important aspects of the data itself. This paper will demonstrate the importance of visualization methods (graphical statistics) in the interpretation and evaluation of linear regression models. Special emphasis will be placed on methods which can reveal the unexpected or which can expose patterns in the data which may not be readily obtained from the standard regression output.

INTRODUCTION

When a regression model, such as the simple linear regression model (see 1.1), is selected for an application, one can usually not be certain in advance that the model is appropriate for that application. Any one, or several of the features of the model, such as linearity of the regression function or normality of the error terms, may not be appropriate for the particular data at hand. Hence, it is important to examine the aptness of the model before further analysis based on that model is undertaken.

There are a number of simple graphical methods that examine the aptness of a model. The Scatter Plot, Normal Probability Plot, Residual Plot, and Partial Residual Plot will be discussed in this paper.

These plots can easily be generated and analyzed using the SAS/STAT™ software package and SAS Macro processing facility. The macro in Appendix A is used to obtain the information needed from the user as well as to call the macros that are used to generate the plots.

WHAT IS LINEAR REGRESSION?

For the purposes of this paper linear regression attempts to find the *line of best fit* where the sum of the squares of the residuals are minimized. An intuitive way of expressing this is:

$$\text{observed values} = \text{Fit} + \text{Residual} \quad (1.1)$$

When there is a linear relationship between two variables, then fit the regression line $Y = \alpha + \beta X_i$ to the data. The random variable $\varepsilon_i = Y_i - (\alpha + \beta X_i)$ is called *residual*.

The assumptions usually associated with least squares regression includes the following:

- 1) Normality of residuals
- 2) Homogeneity of variance of residuals
- 3) Independence of residuals

REGRESSION USING SAS

The procedure in SAS that is used to fit the linear regression model is **PROC REG**. The assumptions above hold in this procedure.

- General form of **PROC REG**:

QUIT;

- Description of selected statements:

- 1) Tells SAS that you want to do regression on the data set selected.
- 2) The **MODEL** statement specifies the model to be used.
- 3) The **RUN** statement is required to complete the SAS regression.

SCATTER PLOT & REGRESSION LINE

General

One of the first things that you should do whenever you perform a regression is to look at a graphical representation of your data and the fitted line. This plot helps to study the relation between the dependent variable Y and the independent variable X. The plot can be used to evaluate the following types of questions:

- Is there a linear relation between the response variable and the independent variable?
- Is simple linear regression model not appropriate? (i.e., it appears that a quadratic term is missing from the model).
- Is there an outlying value?
- Is the slope of the regression line is determined by one observation?

The type of plot which is described here is composed of three main elements: the observed data, the fitted line, and confidence intervals for the mean responses.

How to create in SAS

```
/******  
/*MACRO SCATTER - This macro generates multiple */  
/*"Scatter Plot" where the Y-axis is denoted by dependent */  
/*variable (Y) and X-axis is denoted by individual */  
/*independent variable (X). When there is one */  
/*independent variable (simple linear model) this macro */
```

```
/*in addition to the Scatter Plot generates a regression */  
/*line and 95% confidence limits. */  
/* */  
/* Procedure: GPLOT - produces high resolution plot */  
/*Macro Variable: FILE - name of data set */  
/* COUNT - number of independent variable*/  
/* XVAR - name of independent variable */  
/* YVAR - name of dependent variable */  
/*Local Variable: I - loop counter */  
/* X&I - individual independent variable*/  
/* Misc: SYMBOL - specifies linear regression */  
/* analysis with 95% confidence */  
/* limits and plot symbol */  
/******
```

```
LIBNAME LIBREF "SAS- data-Library";  
OPTIONS MSTORED SASMSTORE= LIBREF;  
OPTION MLOGICAL MPRINT SYMBOLGEN;
```

```
%MACRO SCATTER (FILE,COUNT,XVAR,YVAR)STORE;
```

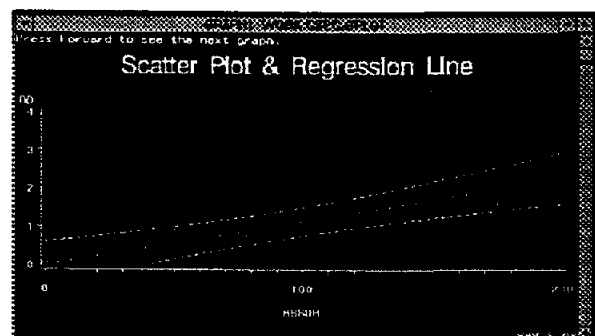
```
%LOCAL I;
```

```
%DO I=1 %TO &COUNT;  
%LOCAL X&I;  
%LET X&I=%SCANT(&XVAR,&I');  
%END;
```

```
%IF &COUNT=1 %THEN  
%DO;  
SYMBOL1 C=DEFAULT I=RLCLM95 L=1  
CI=GREEN V=STAR CV=GREEN;  
PROC GPLOT DATA=&FILE;  
PLOT &YVAR*&XVAR;  
TITLE "SCATTER PLOT & REGRESSION LINE";  
RUN;  
%END;
```

```
%ELSE  
%DO  
I=1 %TO &COUNT;  
PROC GPLOT DATA=&FILE;  
PLOT &YVAR*&X&I;  
TITLE;  
TITLE "SCATTER PLOT";  
RUN;  
%END;  
%MEND SCATTER;
```

Output



Interpretation

The plot above indicates that for OD versus ABSOR, there is an outlying value that affects the parameter estimates. The plot also indicates that all the data points but one fall within the 95% confidence interval.

THE NORMAL PROBABILITY PLOT

General

The Normal Probability Plot (cumulative distribution of the residuals) is used to check the assumption of normality of the residuals. If the residuals are normally distributed, the plot should resemble a straight line. Conversely the normality of the residuals is suspect if the points deviate from a straight line.

Note that Normal Probability Plot complements the R-squared value (correlation coefficient) in the ANOVA table.

One very common defect that often shows up on Normal Probability Plot is one residual that is very much larger than any of the others. Such residual may be due to an outlier. Note that the existence of the outlier is not detected by the R-squared value.

How to create in SAS

```
/******  
/*MACRO NORMAL- This macro generates a "Normal  
/*Probability Plot" where the Y-axis is the residual and the  
/*X-axis is the expected values of the ordered residual  
/*(normal quantile). This plot also contains the regression  
/*line (fitted line).  
/*  
/* Procedure: REG - fits linear regression model  
/* RANK - generates the expected value  
/* of residuals and ranks them  
/* GPLOT - produces high resolution plot  
/*Macro Variable: FILE - name of data set  
/* COUNT - number of independent variable  
/* XVAR - name of dependent variable  
/* YVAR - name of independent variable  
/* Misc: SYMBOL - specifies linear regression  
/* analysis with 95% confidence  
/* limits and plot symbols  
/******
```

```
LIBNAME LIBREF "SAS- data-Library";  
OPTIONS MSTORED SASMSTORE= LIBREF;  
OPTION MLOGICAL MPRINT SYMBOLGEN;
```

```
%MACRO NORMAL (FILE,COUNT,XVAR,YVAR)STORE;
```

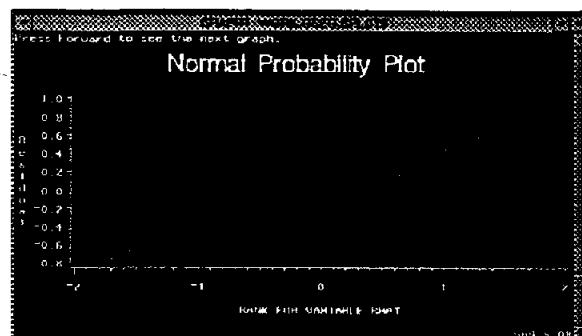
```
PROC REG DATA=&FILE;  
MODEL &YVAR=&XVAR;  
OUTPUT OUT=DEMOOUT P=PHAT R=RHAT;
```

```
PROC RANK NORMAL=BLOM DATA=DEMOOUT;  
VAR RHAT;  
RANKS RESNORM;  
SYMBOL1 INTERPOL=RL CI=GREEN  
VALUE=DOT HEIGHT=1 CV=RED;
```

```
PROC GPLOT;  
PLOT RHAT*RESNORM;  
TITLE "NORMAL PROBABILITY PLOT";  
RUN;
```

```
%MEND NORMAL;
```

Output



Interpretation

The plot above indicates that the residual falls approximately on a diagonal straight line. Hence the assumption of normality of the residuals is accepted.

THE RESIDUAL PLOT

General

Residual plots are helpful to study whether the variance of the residuals are constant (homogenous). The plot is created by plotting residuals against the fitted values (predicted values). If the statistical model does a good job of predicting the observed values, then the residuals should be small and should be evenly distributed without evidence of patterns or trends.

The assumption of homogeneity of variances should be questioned when:

- The variance of the observations increases as the magnitude of the observation increases.
- The data follow a non normal, skewed distribution.
- Existence of an outlier.

If heterogeneity of the variance is suspected, one of the ways of approaching the nonconstant variance is to transform either the response Y, the independent variable X or both.

How to create in SAS

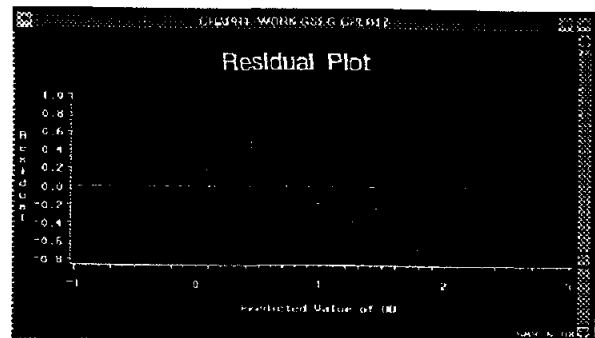
```
/******  
/*MACRO RESIDUAL- This macro generates a "Residual */  
/*Plot" where the Y-axis is the residual and the X-axis is the */  
/*predicted value (fitted value). */  
/* */  
/* Procedure: REG - fits linear regression model */  
/* GPLOT - produces high resolution plot */  
/*Macro Variable: FILE - name of data set */  
/* COUNT - number of independent variable */  
/* XVAR - name of dependent variable */  
/* YVAR - name of independent variable */  
/* Misc: SYMBOL - specifies the plot symbols */  
/******  
  
LIBNAME LIBREF "SAS- data-Library";  
OPTIONS MSTORED SASMSTORE= LIBREF;  
OPTION MLOGICAL MPRINT SYMBOLGEN;  
  
%MACRO RESIDUAL (FILE,COUNT,XVAR,YVAR)STORE;
```

```
PROC REG DATA=&FILE;  
MODEL &YVAR=&XVAR;  
OUTPUT OUT=DEMOOUT P=PHAT R=RHAT;  
RUN;
```

```
SYMBOL1 V=STAR CI=GREEN;  
PROC GPLOT DATA=DEMOOUT;  
PLOT RHAT*PHAT/VREF=0 LVREF=2;  
TITLE "RESIDUAL PLOT";  
RUN;
```

%MEND RESIDUAL;

Output



Interpretation

Using the same data as in the Normal Probability Plot, the plot above indicates that there is no apparent trends or pattern in the residuals. However, it indicates that there is one particular large residual, and it is toward the end of the values.

THE PARTIAL RESIDUAL PLOT

General

When the relation between Y and X_i is nonlinear, the Partial Residual Plot may show both the linear and nonlinear components of the relation controlling for the other predictors. This helps determine if a term of X_i^2 or some other function of X_i needs to be included or omitted. In other words, a Partial Residual Plot shows graphically the nature of the functional relation in which a variable should be added to the regression model.

How to create in SAS

```

/*****
/*MACRO PARTIAL- This macro generates a "Partial
/*Residual Plot" where the Y-axis is the residual of (Y|Xj) and
/*the X-axis is the residual of (Xi|Xj). The Xi is the
/*independent variable under consideration and Xj are the
/*remaining independent variables.
/*
/* Procedure: REG - fits linear regression model
/* GPLOT - produces high resolution plot
/*Macro Variable: FILE - name of data set
/* COUNT - number of independent variable
/* XVAR - name of dependent variable
/* YVAR - name of independent variable
/*Local Variable: I - loop counter
/* J - inner loop
/* NEW - temporary data file
/* X&I - individual independent variable
/* Misc: SYMBOL - specifies the plot symbols
*****/

```

```

LIBNAME LIBREF "SAS-data-Library";
OPTIONS MSTORED SASMSTORE= LIBREF;
OPTION MLOGICAL MPRINT SYMBOLGEN;

```

```
%MACRO PARTIAL (FILE,COUNT,XVAR,YVAR)STORE;
```

```
%LOCAL I J NEW;
```

```
%DO I=1 %TO &COUNT;
```

```
%LOCAL X&I;
```

```
%LET X&I=%SCAN(&XVAR,&I,');)
```

```
%END;
```

```
%DO I=1 %TO &COUNT;
```

```
%LET NEW=;
```

```
%DO J=1 %TO &COUNT;
```

```
%IF &J ^=&I %THEN %LET NEW = &NEW &&X&I;
```

```
%END;
```

```

PROC REG DATA=&FILE;
MODEL &YVAR=&NEW;

```

```

OUTPUT OUT=DEMOOUT P=PHAT R=RHAT;
RUN;

```

```
PROC REG DATA=DEMOOUT;
```

```
MODEL &&X&I=&NEW;
```

```

OUTPUT OUT=DEMOOUT2 P=PHAT2 R=RHAT2;
RUN;

```

```
SYMBOL1 V=STAR CI=GREEN;
```

```
PROC GPLOT DATA=DEMOOUT2;
```

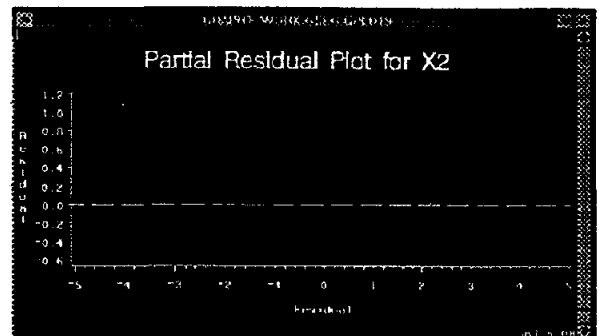
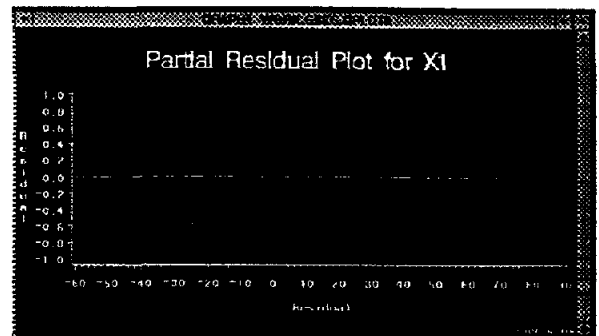
```
PLOT RHAT*RHAT2/VREF=0 LVREF=2;
```

```
TITLE "PARTIAL RESIDUAL PLOT";
```

```
RUN;
```

```
%MEND PARTIAL;
```

Output



Interpretation

Like the Residual Plot, using the same data as in the Normal Probability Plot, the plots above indicate that both X_1 and X_2 should be included in the regression model.

SUMMARY

In terms of interpretation and evaluation of linear regression, we would suggest looking at both statistical graphics and statistical analysis when dealing with checking the model assumptions. Graphical methods not only help to verify whether the data meet the assumptions but it also lead one into taking the correct action if they do not. These plots are also useful for uncovering outlying data points. In summary:

- The Scatter Plot is used to study the relation between the response variable and independent variables.
- The Normal Probability Plot is used to check the assumption of normality .
- The Residual Plot is used to check the assumption of homogeneity of variance.
- The Partial Residual Plot is used to identify the effect of a variable in the regression model.

TRADEMARK NOTICE

SAS/STAT and SAS/GRAPH are registered trademark of SAS Institute Inc., Cary, N.C. USA

AUTHOR

Simin Kavoossi
Tim Cooley
Boehringer Mannheim Company
9115 Hague Rd.
Indianapolis, IN 46256

REFERENCES

- [1] Edwards, L. A. (1976). *An Introduction to Linear Regression and Correlation*. W. H. Freeman and Company.
- [2] Kurner, M., Neter, J. and W. Wasserman (1989). *Applied Linear Regression Models*. Richard D. Irwin, INC.
- [3] Montgomery, C. C. (1991). *Design and Analysis of Experiment*. John Wiley & sons.
- [4] Weisberg, S. (1980). *Applied Linear Regression*. John Wiley & sons.
- [5] Younger, M. S. (1979). *A Handbook for Linear Regression*. Duxbury Press.
- [6] SAS Institute, Inc. (1991). *SAS System for Statistical Graphics, First Edition*, Cary, NC: SAS Institute Inc.
- [7] SAS Institute, Inc. (1990), *SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 2*, Cary, NC: SAS Institute Inc.

APPENDIX A: Macro Visual

```

/*****
/*
/*      Boehringer Mannheim
/*      Research and Development
/*
/*      NAME: VISUAL.SAS
/*      TITLE: Data Analysis and Information Visualization
/*      SYSTEM: VMS environment
/*SUBMACRO: SCATTER, NORMAL, RESIDUAL, PARTIAL*/
/*MAIN MACRO: VISUAL
/* PROCEDURES: REG, RANK, GPLOT
/*      MISC: SAS Version 6.08
/*      AUTHOR: SIMIN KAVOOSI
*****/
LIBNAME LIBREF "SAS- data-Library";
OPTIONS MSTORED SASMSTORE=LIBREF;
OPTION MLOGICAL MPRINT SYMBOLGEN;

%macro visual;

/*****
/* Window INFO: This window collects information from
/*the user that is needed to generate Scatter, Normal
/*Probability, Residual, and Partial Residual plots.
*****/

%window info

#3 @4 "DATA ANALYSIS AND INFORMATION
      VISUALIZATION: &sysday, &sysdate"

#5 @20 'Press' @26 'TAB' attr=highlight @30 'to move
      between fields.'

#9 @10 'Enter the FILENAME containing your set:' +1 file 15
      attr=highlight required=yes

#11 @10 'Enter the NUMBER of independent variables:' +1
      count 2 attr=highlight required=yes

#13 @10 'Enter the names of INDEPENDENT variables:'

#15 @10 xvar 50 attr=highlight required=yes

#17 @10 'Enter the name of your DEPENDENT variable:' +1
      yvar 10 attr=highlight required=yes

#20 @17 'Press ' @23 'RETURN' attr=underline @30
      'to prompt the next screen.';
/*****
/*Window PROCESS: This window is to inform the user
/*that the program is processing the submitted statements.
*****/

%window process

#7 @8 'Standby - I am currently processing the
      submitted statements';

%display info;
%display process noinput;

```

```

/*****
/* The DATA step begins creating observations by reading
/*an external file, assigning a value to the variables XVAR
/*and YVAR, and holding the file for further processing.
*****/
data temp;
    infile "&file..dat";
    input &xvar &yvar;

    run;

    %scatter(temp,&count,&xvar,&yvar);
    %normal(temp,&count,&xvar,&yvar);
    %residual(temp,&count,&xvar,&yvar);
    %partial(temp,&count,&xvar,&yvar);

%mend visual;
%visual;
quit;

```