USING THE INTERACTIVE FEATURES
OF
PROC REG

by
Rudolf J. Freund
Texas A&M University
April 4, 1990

BACKGROUND

IN THE BEGINNING . . .

we would submit a SAS program to perform a regression analysis. The
program would consist of a PROC REG with options for special
statistics, an OUTP= statement for producing diagnostic statistics,
followed by other procedures such as PROD PLOT to further examine the
results.

If these results suggested further analyses, a new program would be
submitted, and so on.

This process is called the BATCH mode and is still quite popular.

However, regression analyses are often not well structured and
require many iterations before a satisfactory analysis is produced.

Hence a regression using the batch mode can become quite time
consuming, especially if turnaround of individual programs is not
quick.

Increased availability of computing power and full screen data
displays can be used for a more effective interactive mode of data
analysis.

----------

SAS is a registered trademark of SAS Institute.
**EXAMPLE**

The data contains variables related to gasoline consumption in the 48 contiguous (US) states, with data from the District of Columbia incorporated into Maryland. The data are extracted from Frank H. Drysdale and Charles E. Calef, The Energy of the United States: an Atlas. Published by the Departmental National Laboratory, Athens ET 11972, 1977.

The variables are:

- **STATE**, the standard 3 character abbreviation.
- **REGION**, indicating if a state is EAST or WEST of the Mississippi River.
- **GAS**, total gasoline use (including auto diesel) in 10⁶ lb 12 BMU.
- **INC**, estimated number of registered vehicles, in millions.
- **INC**, personal income, in billions of dollars.
- **VAL**, value added by manufacturers, in billions of dollars.
- **SUTA**, population density, in 1000 persons per square mile.

The data are given in **OUTPUT 1**.

**OUTPUT 1**

<table>
<thead>
<tr>
<th>STATE</th>
<th>REGION</th>
<th>GAS 10^6 lb</th>
<th>INC</th>
<th>VAL 10^6</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL</td>
<td>EAST</td>
<td>0.6517</td>
<td>0.9437</td>
<td>0.3924</td>
</tr>
<tr>
<td>AZ</td>
<td>WEST</td>
<td>0.6513</td>
<td>0.9437</td>
<td>0.3924</td>
</tr>
<tr>
<td>AR</td>
<td>EAST</td>
<td>0.7275</td>
<td>0.7067</td>
<td>0.3264</td>
</tr>
<tr>
<td>CA</td>
<td>WEST</td>
<td>0.6259</td>
<td>0.7395</td>
<td>0.3190</td>
</tr>
<tr>
<td>CO</td>
<td>EAST</td>
<td>0.6252</td>
<td>0.7395</td>
<td>0.3190</td>
</tr>
<tr>
<td>CT</td>
<td>EAST</td>
<td>0.5699</td>
<td>0.6008</td>
<td>0.3466</td>
</tr>
<tr>
<td>DE</td>
<td>EAST</td>
<td>0.6746</td>
<td>0.5209</td>
<td>0.3466</td>
</tr>
<tr>
<td>FL</td>
<td>EAST</td>
<td>0.7650</td>
<td>0.7038</td>
<td>0.3599</td>
</tr>
<tr>
<td>GA</td>
<td>EAST</td>
<td>0.6014</td>
<td>0.6602</td>
<td>0.3599</td>
</tr>
<tr>
<td>ID</td>
<td>EAST</td>
<td>0.6447</td>
<td>0.7044</td>
<td>0.3588</td>
</tr>
<tr>
<td>IL</td>
<td>EAST</td>
<td>0.5501</td>
<td>0.5046</td>
<td>0.3464</td>
</tr>
<tr>
<td>IN</td>
<td>EAST</td>
<td>0.7201</td>
<td>0.5863</td>
<td>0.3475</td>
</tr>
<tr>
<td>IA</td>
<td>EAST</td>
<td>0.7005</td>
<td>0.5749</td>
<td>0.3464</td>
</tr>
<tr>
<td>KY</td>
<td>EAST</td>
<td>0.6425</td>
<td>0.6004</td>
<td>0.3599</td>
</tr>
<tr>
<td>LA</td>
<td>EAST</td>
<td>0.1604</td>
<td>0.3263</td>
<td>0.6053</td>
</tr>
<tr>
<td>NH</td>
<td>EAST</td>
<td>0.6962</td>
<td>0.6564</td>
<td>0.3599</td>
</tr>
<tr>
<td>ND</td>
<td>EAST</td>
<td>0.5601</td>
<td>0.5103</td>
<td>0.3464</td>
</tr>
<tr>
<td>MA</td>
<td>EAST</td>
<td>0.5301</td>
<td>0.4944</td>
<td>0.3464</td>
</tr>
<tr>
<td>MI</td>
<td>EAST</td>
<td>0.6104</td>
<td>0.6069</td>
<td>0.3475</td>
</tr>
<tr>
<td>NE</td>
<td>EAST</td>
<td>0.6074</td>
<td>0.6539</td>
<td>0.3464</td>
</tr>
<tr>
<td>NV</td>
<td>EAST</td>
<td>0.7136</td>
<td>0.6262</td>
<td>0.3475</td>
</tr>
<tr>
<td>NJ</td>
<td>EAST</td>
<td>0.7053</td>
<td>0.5155</td>
<td>0.3464</td>
</tr>
<tr>
<td>NY</td>
<td>EAST</td>
<td>0.9201</td>
<td>0.7252</td>
<td>0.3599</td>
</tr>
<tr>
<td>OH</td>
<td>EAST</td>
<td>1.0601</td>
<td>0.8237</td>
<td>0.3599</td>
</tr>
<tr>
<td>OK</td>
<td>EAST</td>
<td>0.6601</td>
<td>0.5944</td>
<td>0.3599</td>
</tr>
<tr>
<td>OR</td>
<td>EAST</td>
<td>0.5978</td>
<td>0.5387</td>
<td>0.3464</td>
</tr>
<tr>
<td>PA</td>
<td>EAST</td>
<td>0.8646</td>
<td>0.6959</td>
<td>0.3599</td>
</tr>
<tr>
<td>RI</td>
<td>EAST</td>
<td>0.4202</td>
<td>0.3818</td>
<td>0.3599</td>
</tr>
<tr>
<td>SC</td>
<td>EAST</td>
<td>0.7064</td>
<td>0.6779</td>
<td>0.3599</td>
</tr>
<tr>
<td>SD</td>
<td>EAST</td>
<td>0.8597</td>
<td>0.7176</td>
<td>0.3599</td>
</tr>
<tr>
<td>TN</td>
<td>EAST</td>
<td>0.8127</td>
<td>0.7376</td>
<td>0.3599</td>
</tr>
<tr>
<td>TX</td>
<td>EAST</td>
<td>0.8127</td>
<td>0.7376</td>
<td>0.3599</td>
</tr>
<tr>
<td>UT</td>
<td>EAST</td>
<td>0.7145</td>
<td>0.5913</td>
<td>0.3464</td>
</tr>
<tr>
<td>VT</td>
<td>EAST</td>
<td>0.7742</td>
<td>0.4903</td>
<td>0.3599</td>
</tr>
<tr>
<td>WA</td>
<td>EAST</td>
<td>0.8372</td>
<td>0.5851</td>
<td>0.3599</td>
</tr>
<tr>
<td>WV</td>
<td>EAST</td>
<td>0.6323</td>
<td>0.5155</td>
<td>0.3464</td>
</tr>
<tr>
<td>WI</td>
<td>EAST</td>
<td>0.4422</td>
<td>0.4510</td>
<td>0.3599</td>
</tr>
<tr>
<td>WY</td>
<td>EAST</td>
<td>0.5669</td>
<td>0.4944</td>
<td>0.3599</td>
</tr>
</tbody>
</table>

**PARAMETER ESTIMATES**

<table>
<thead>
<tr>
<th>PARAMETER</th>
<th>ESTIMATE</th>
<th>STANDARD ERROR</th>
<th>t VALUE</th>
<th>P VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.213846</td>
<td>0.096473</td>
<td>2.2000</td>
<td>0.0263</td>
</tr>
<tr>
<td>DFS</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOF</td>
<td>47</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MS</td>
<td>0.048981</td>
<td>0.015272</td>
<td>3.2159</td>
<td>0.0016</td>
</tr>
<tr>
<td>MSE</td>
<td>0.004891</td>
<td>0.000248</td>
<td>1.9438</td>
<td>0.0583</td>
</tr>
<tr>
<td>R^2</td>
<td>0.7003</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adj R^2</td>
<td>0.6909</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**ANALYSIS OF VARIANCE**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F Value</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODEL</td>
<td>4</td>
<td>0.7003</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RESIDUAL</td>
<td>47</td>
<td>0.004891</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>47</td>
<td>0.7051</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**STANDARD DEVIATION OF RESIDUALS**

- **Mean** 0.004891
- **Standard Deviation** 0.00248

**PREDICTED VALUES**

- **GAS** 0.8127
- **INC** 0.7376
- **VAL** 0.6779

It seems that the number of vehicles is the most important factor, with value added having a marginally negative (11) effect.
EXAMPLE OF PRINT STATEMENT

We can investigate multicollinearity by examining the variance inflation factors and comparing the type I and type II sums of squares. These statistics are obtained by:

```
PRINT SSt 552 VIF;
END;
```

The results are given in OUTPUT 3.

MISCELLANEOUS INTERACTIVE STATEMENTS

A TEST statement can be entered at any stage and will perform the test on the most recently modified model. Test statements are independent of each other.

A RESTRICT statement may be entered at any stage but produces no output. A PRINT statement must be used to provide the output. Restrictions are in effect until cancelled by another RESTRICT or MODEL statement. However, a completely non-restricted model can only be obtained by a new MODEL statement.

Model SELECTION options are not available interactively in that they require a new MODEL statement. However, the last model resulting from any selection option may be subjected to additional interactive statements. (Note: a PRINT statement following a selection must specify ANOVA in addition to any other statistics wanted).

THE PLOT STATEMENT

Scatterplots involving statistics available in the OUTPUT statement and variables in the MODEL or VAR statements can be obtained with the interactive PLOT statement. The PLOT statement is of the form:

```
PLOT variable*variable = symbol ... / OPTIONS;
```

**CONCEPTS**

- The syntax of the PLOT statement is the same as for PROC PLOT.
- OVERLAY's are permitted, but scaling, referencing and plot size options are not available, however, see PLOTS and VPLOTS below.
- Diagnostic statistics available in the OUTPUT statement (and ODS, the observation number) are specified by the corresponding keyword followed by a period.
- Variables in the VAR or MODEL statement are specified by their variable names.
- Plotting symbols may be specified as with PROC PLOT, but may also be specified with a PLOT statement. The default symbol is a '1' for one observation, '2' for two observations, etc., with '1' if there are 10 or more.
- The plot option HORIZON provides for horizontal and vertical plots producing for n^2 graphs per page.
- All specifications for a PLOT statement carry over to subsequent plots unless cancelled by specifications of a subsequent PLOT statement. Thus a simple PLOT will exactly reproduce the previously specified plot for the most recently fitted model.
EXAMPLE OF PLOT STATEMENT
ROW DIAGNOSTICS

we may look for outliers and/or influential observations by plotting
some of the residual or influence statistics. We continue with the
following statements:

```
   PLOT (R. PRESS, R. EFFECTS.)* P. - STATE / PLOTS=2 VPOINTS = 2;
```

The results are given in OUTPUT 4.

**NOTE:** when plotting symbols have been specified, '*' is used to show
any number of multiple plotting points.

The plots show that states beginning with W, N, and R may be outliers
or influential.

![Output 4: ROW DIAGNOSTICS Plots](image)

**OUTPUT 4**

The row diagnostic plots indicated that states beginning with W, N, and R are candidates for outliers or influential observations.

**Obviously 'W' is Rhode Island, but there are more than two such for the other letters. These may be identified by PAINT-ing these observations.**

We enter the statements:

```
   PAINT STATE = "W" / SYMBOL = "*";
   PAINT STATE = "N" / SYMBOL = "*";
   PAINT STATE = "R" / SYMBOL = "*";
   PAINT STATE = "W" / SYMBOL = "*";
   PAINT STATE = "W", "N", "R" / SYMBOL = "*";
   PAINT ALLOBS / RESET;
```

For this example, we do not want all four plots or the state identifications for the PAINTED plots, hence we enter a new PAINT statement:

```
   PLOT R. EFFECTS. * P;
```

This statement will produce only two plots and undo the plotting symbol option. However, the provision for four plots per page has not been cancelled, thus producing two half-page plots.

The result is provided in OUTPUT 5.
We can now see that Wyoming and Nevada are the outliers or influential observations.

## DELETING OBSERVATIONS, THE REWEIGHT STATEMENT

The REWEIGHT statement is used to assign any arbitrary weight to specified observations. Assigning a zero weight causes the observation to be deleted from an analysis.

- The REWEIGHT statement specifies a logical condition identifying observations to be reweighted.
- There may be only one condition per statement.
- The default weight is zero but other weights may be specified as an option.
- There may be more than one REWEIGHT statement. Additional REWEIGHT statements are cumulative and all reweighting stays in effect even through additional MODEL statements.
- REWEIGHT statements create no output. PRINT or PLOT statements must be used to display the results of reweighting.
- Reweighting is cancelled with the statement:

  *REWEIGHT ALLOCS / RESTRT;

### EXAMPLE OF REWEIGHT STATEMENT

Wyoming and Nevada appeared to be influential. It may be useful to see what would happen if these were deleted. The statements are:

```plaintext
REWEIGHT STATE = 'WY' OR STATE = 'NV';
PRINT STATE;
```

The output is given in OUTPUT 6.
ADDITIONAL PLOTS

Since the reweighting has not been cancelled, we can immediately obtain plots of residuals and DFITS values for the models estimated without Wyoming and Nebraska as follows:

PLOT (R, DFITS,)*P, = STATE;
REWEIGHT ALASKA / RESWT;

Remember that a single PLOT statement would have reproduced the last plots for the reweighted analysis. However, if we now want to use the state identifications as plotting symbols, the plot must be respecified. However, the provision for four plots per page is still in effect.

The RESWT option is used to cancel the reweighting for future analyses.

The plot is given in OUTPUT 7.

CHANGING THE MODEL
DELETES AND ADD STATEMENTS

Changing the independent variables in a model is accomplished by the DELETES and ADD statements. In our example, the statement:

DELETES INC VAR;

refits the model that deletes these variables. Equivalently an ADD statement adds variables.

COBENTS

* Variables to be added must have been specified in the first MODEL or VAR statement.
* ADD and DELETE statements automatically refit the model but produce no output. PRINT or PLOT statement must be used to examine model fitting results.
* All subsequent statements reflect results of the altered model.
* A new MODEL statement cancels all ADD and DELETE statements.

EXAMPLE OF DELETE STATEMENT

provides the results in OUTPUT 8 (remember that the deletion of WY and NY has been RESWT).

OUTPUT 8
MODEL WITHOUT INC AND VAL

Model: MODE1
Dependent Variable: GAS

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Mean Squared</th>
<th>F Value</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>0.77100</td>
<td>0.23511</td>
<td>0.00522</td>
</tr>
<tr>
<td>Error</td>
<td>45</td>
<td>0.97211</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C Total</td>
<td>47</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 0.37128
R-Squared 0.7581
Adj R-Squared 0.7474
C.V. 9.97211

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t for H0: Parameter=0 | Prob > |t|
|----------|----|--------------------|----------------|------------------------|--------|
| INC      | 1  | 0.060661           | 0.032681       | 1.83                 | 0.084  |
| NY       | 1  | 0.149890           | 0.0396713      | -3.75                 | 0.0008 |

Note that without INC and VAL, the IN hyp variable becomes quite significant.
COMPARATIVE PLOTS
THE COLLECT OPTION

We have seen that we can interactively change or modify various aspects of a regression analysis. The COLLECT option for PLOT statements allows us to compare plots resulting from different analyses.

For example we may want to:

- compare residual plots from models that use different independent variables.
- compare influence statistic plots when observations have been deleted.
- compare residual (or other) plots for different PRINT specifications.

All of these can be performed with the COLLECT option in the PLOT statement.

HOW THE COLLECT OPTION WORKS

Consider the following statements:

PLOT B.*P. / COLLECT;
DELET VAR: VAR1; DELET; PLOT B.*P. / MODCOLLECT;

* The first PLOT statement produces a single plot of residuals against predicted values from the most recent model, using the most recent plot options. The COLLECT option causes the plot to be saved for 'collecting' with subsequent plots.

* The DELETE statement is used to modify the model. PRINT and REMOVE may also be used.

* The second PLOT statement provides for two plots:

  1. The first overlays the original plot with the plot of (B.*P.) for the altered model resulting from the COLLECT option.
  2. The second plots DIFFP.*P. from the altered model. There is no overlay since there was no second plot in the first PLOT statement. The MODCOLLECT options causes the collecting to stop.

SIDE BY SIDE PLOTS WITHOUT OVERLAY

Overlaid plots are not always useful because:

- the lack of definition in a printer plot.
- the plots are not rescaled hence some data points on the second plot may be out of range.

Producing side by side plots with no overlays is done by producing a blank plot for the same position of subsequent collected plots. This is done by specifying ' ' for the plotting symbol.

EXAMPLE

We want side by side plots of the residuals and DIFFP against predicted values for the full and two variable models. Remember that the last model deleted INC and VAL and the last plot option specified four plots per page. We enter the following statements:

PLOT (B. DIFFP.*P. = STATE / COLLECT;
ADD INC VAL;
PLOT (B. DIFFP.*P. = ' ' (R. DIFFP.*P. = STATE / MODCOLLECT);

* The first PLOT statement produces side by side residual and DIFFP plots for the reduced model (not reproduced here).

* The ADD statement re-enters INC and VAL and automatically refits the model.

* The second PLOT statement leaves the first two plots as they are and adds the second set of plots for the full model. The MODCOLLECT option turns off collecting for future plots.

The plots are reproduced in UMPR 5. The plots for the two models are not very different, which may be expected since the residual sum of squares is not much different for the two models.
OUTPUT 9
SIDE BY SIDE PLOTS FOR DIFFERENT MODELS

0.4
0.2
0.0
-0.2

0.4
0.2
0.0
-0.2

0.4
0.2
0.0
-0.2

0.4
0.2
0.0
-0.2

GENERAL COMMENTS

The interactive analyses now available in PROC REG can be very useful for the regression analyst, but they are not an unaided blessing.

SOME DISADVANTAGES:

* Less flexibility. For example, only scatterplots and data listings are available interactively, while statistics from the OUTPUT statement can be subjected to any and all data summary procedures available in the SAS System.

* The interactive analyses require more computer time since the model is refit for virtually every statement. This is not important for small data sets but could be a factor when the number of observations becomes large.

* The cumulative nature of the individual statements can cause confusion. Frequent use of REST and UNDO options is helpful, but even these can be confusing.

* As noted in the User's Guide and in the SAS User's Guide, interactivity is disabled when a BY statement is used.

A WARNING:

If there is an error, be careful when break out of the analysis, since the 'T' option will exit the procedure hence stopping the interactive session. It is usually better to let the system cancel the offending step.