Paper 109-31

Introduction to SAS® Enterprise Guide® 4.1 for Statistical Analysis

Charles Hallahan, Economic Research Service, USDA
Linda Atkinson, Economic Research Service, USDA

ABSTRACT

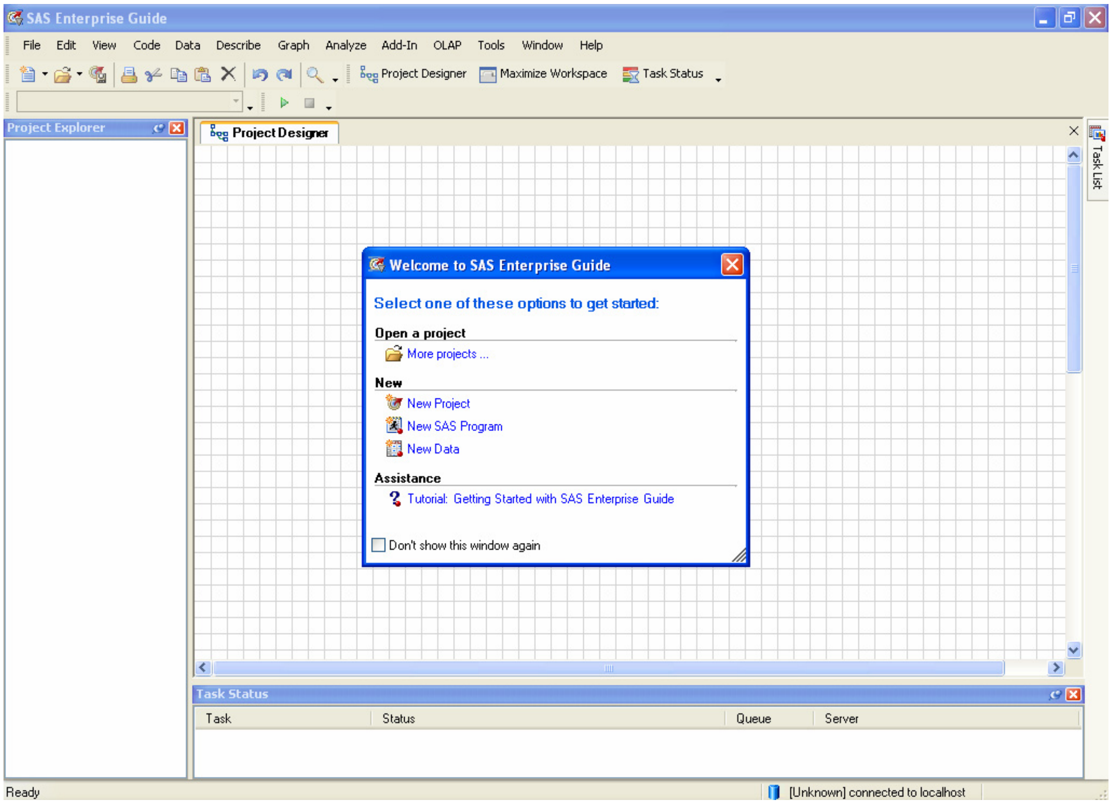
The workshop begins with a tour through Enterprise Guide (EG) as a menu-based interface to SAS® procedures. Emphasis in this workshop will be on the statistical capabilities of EG. A set of data, put together using features of the query builder task, serves as a case study for a statistical analysis. The workshop will show how code that's generated by SAS Enterprise Guide can be customized, stored, and re-run, and custom reports saved with Report Controls Integration.

INTRODUCTION

The SAS System provides a powerful framework for statistical analysis. It has extensive data manipulation capabilities to prepare for analytic and modeling work. It has reporting tools for presenting results. However, for a new user, learning how to write code and run the appropriate procedures can be daunting. Enterprise Guide enables you to get answers without having to write programs, through a point-and-click interface making selections from a series of menus. As a benefit even for experienced SAS programmers, EG provides a framework within which to organize the data, tasks, and results involved in performing a statistical analysis, through the creation and maintenance of "projects". In this workshop, a set of data will serve as a case study for a statistics exercise. Along the way, we will review the code generated automatically by EG, and demonstrate how it can be customized, stored, and rerun. We'll also "put it all together" by collecting results and generating custom reports through a tool called the Document Builder and also with Report Controls Integration.

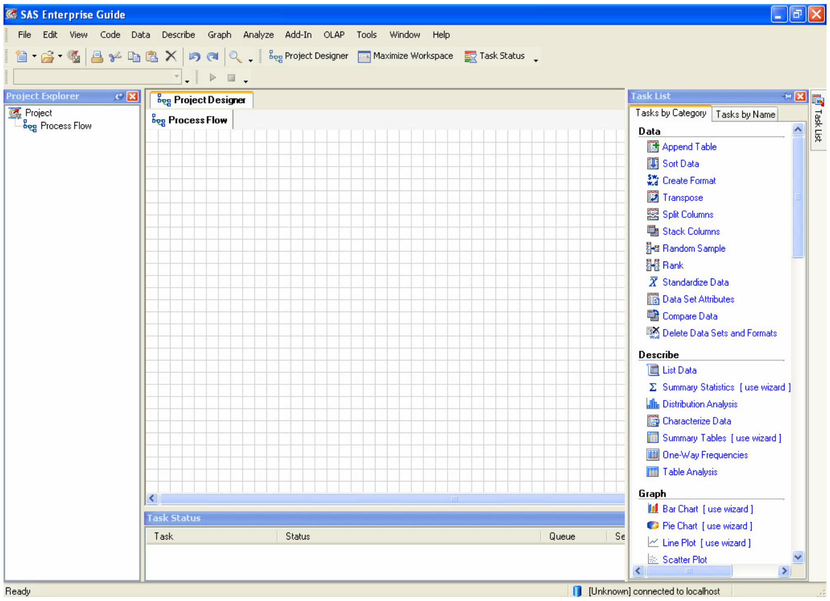
CREATING A PROJECT

When you first bring up Enterprise Guide, you'll be asked whether you'd like to create a new project or open an existing one. Click on **New Project** under **New**.



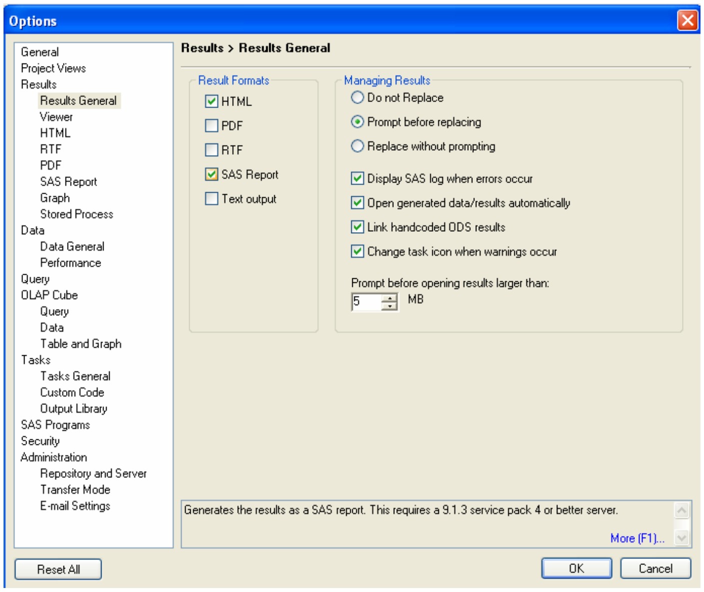
Alternatively, at any time while in Enterprise Guide, you can click on the **File** → **New** pulldown menus to indicate you'd like to open a new project. There's also a **New** button on the toolbar to accomplish the same function.

You'll see a tree view of the Project in the Project Explorer at the left of the EG workspace, a Process Flow view in the middle of the screen (where Results will also be viewed), a Task List window that pulls in from the right when you mouse over it, and a Task Status window at the bottom of the workspace. Any of these windows can be closed to give you more space.



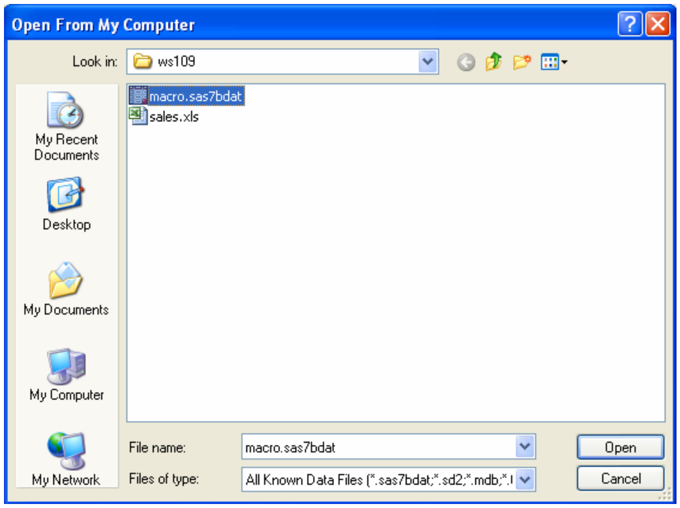
Let's give the Project a better name. Click on **File, Save Project As…**, click on **Local Computer** when asked where to save the project to, and type "SUGI31" in the **File name**: box. Enterprise Guide will append the extension .egp. Navigate to the c:\workshop\ws109 directory and click **Save**.

Besides organizing your work handily, Enterprise Guide also provides opportunity for customizing your work environment. Let's take a look at specifying how we would like our results presented. Click on **Tools** → **Options** from the pulldown menu. Select **Results General**. Be sure that both **HTML** and **SAS Report** are checked, then click on the **OK** button (you can click in the other areas to see what types of things EG gives you control over).

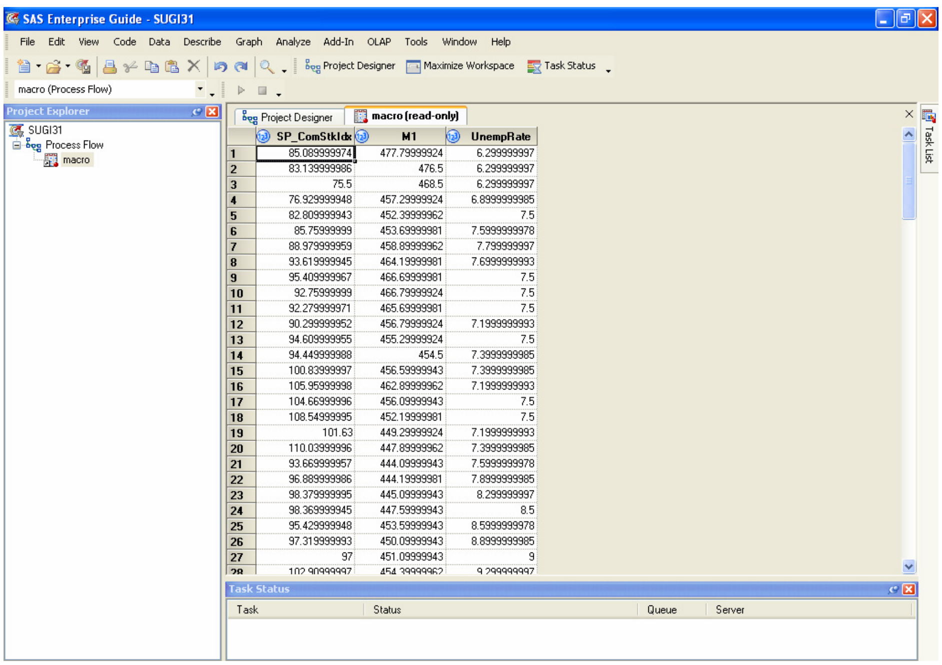


ACCESSING DATA

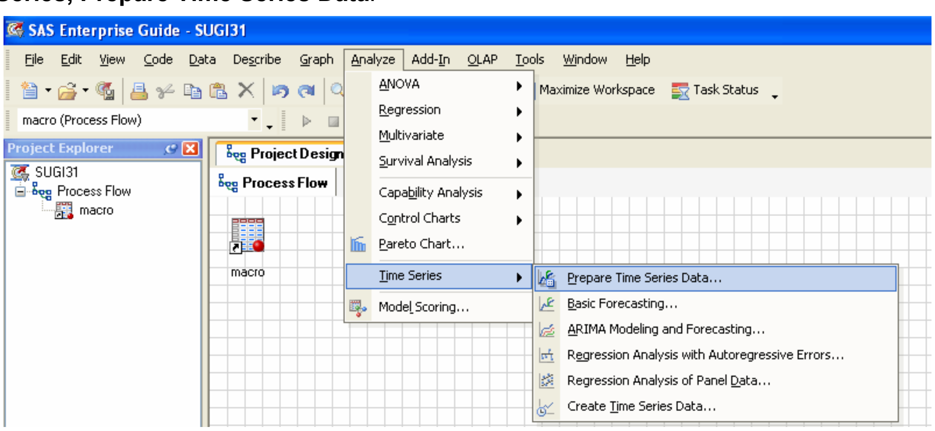
The first step in an analysis is to get some data to work with. Click on **File, Open, Data**. Select **Local Computer** as the location to open the data from. Navigate to the c:\workshop\ws109 directory and select **macro.sas7bdat**, which contains macroeconomic variables for the analysis. Click on **Open**.



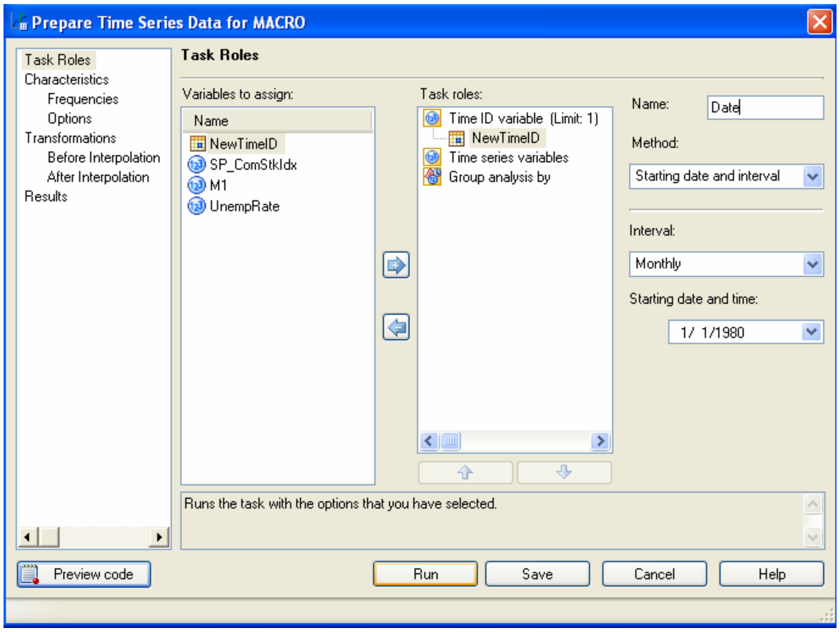
Enterprise Guide reads the data and brings it up in a Data Viewer. You can scroll down to see more observations, or with a larger data set, right and left to view additional variables. If you need to modify a data value, EG will ask you if you'd like to switch to Update mode. Note the entry for **macro** in the Project Explorer window. An icon indicates that this object is a data table. We can close the data viewer by clicking on the X on the right. At any time when you'd like to view the data again, double-click on the data table icon.



These data represent time series of monthly observations beginning in January 1980, but there isn't an explicit date variable on the data set. We'll create one that we can use later in the statistical exercise. Using the pull-down menus, click on **Analyze, Time Series, Prepare Time Series Data**.

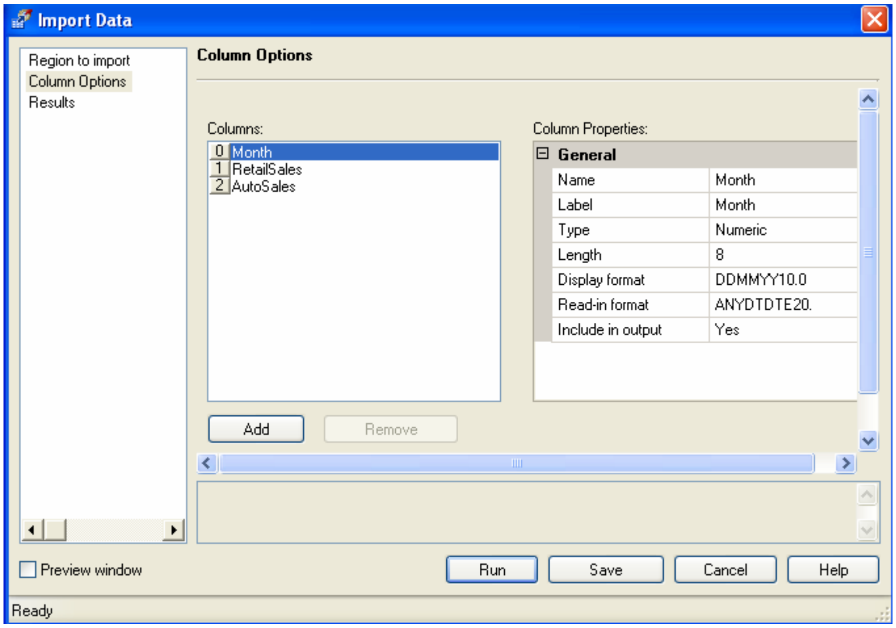


Click and drag **NewTimeID** to the Task role of **Time ID variable**. We'll use the default interval of Monthly, and change the starting date to **1/1/1980**. Then give it a name by typing **Date** in the Name: box. Click on **Run**.



The data viewer shows the modified data table with the new variable that's been created. Closing the views of the original and new tables, we can see in the Process Flow window the steps of the project so far.

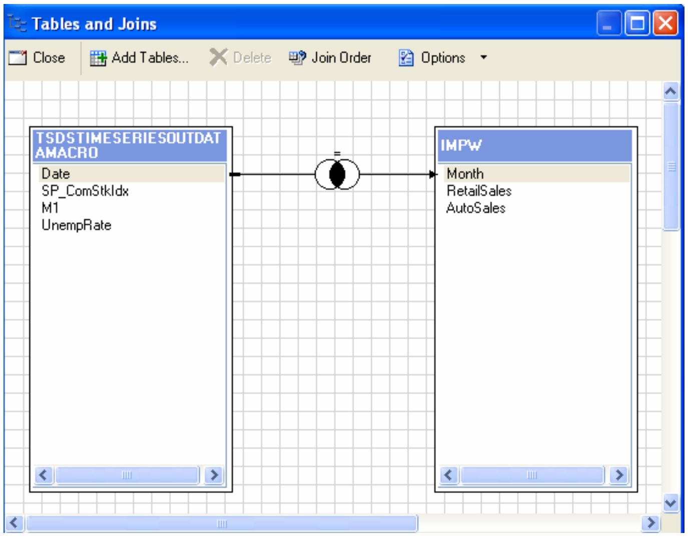
Some of the information we need for this analysis (specific sales data) is stored in a spreadsheet file. Click on **File, Open, Data** again, select **Local Computer,** click on **sales.xls** and **Open**. Select the Sales$ spreadsheet from within the workbook (which represents the entire sheet, whereas the file without the $ at the end represents a range of cells within the sheet). Click on **Open**. You are prompted to choose how you would like to add the file to your project. For this example, choose the second option to create a SAS data set, which will give us more control over making modifications or additions to it. By default, SAS converts the date in the Excel spreadsheet to a datetime variable; we're going to control this so that it is treated as a date variable. Click on **Column options** on the left. Select the Month variable and click in the field for the Read-in format to get further options (click on the box with …). In the Date category, select the Informat ANYDTDTEw. and click on **OK**. Change the display format to DDMMYYw.d (again, in the Date category). Click **OK** and then **Run**.



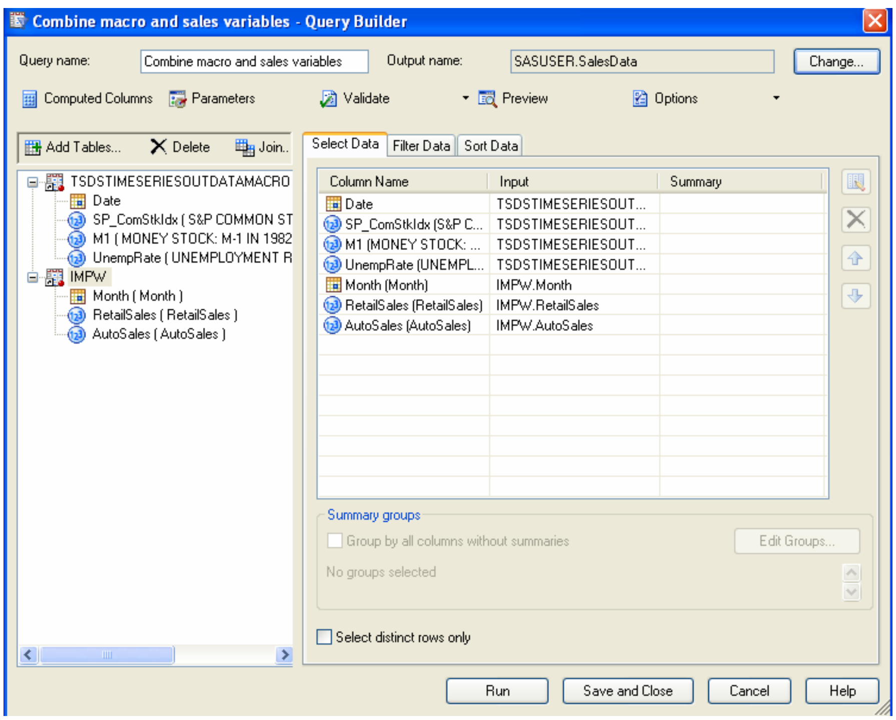
The spreadsheet is added to the project (with a different icon indicating that it is an Excel file), and the data are brought up in a viewer again as the SAS data set was.

We currently have two sources of data in the project: a SAS data set with general macroeconomic variables, and an Excel spreadsheet with the sales variables. For subsequent analysis, we'll want to put them together. Click on the icon for the **Modified Time Series** version of the **macro** data set to make it the active data source. From the pull-down menu, select **Data, Filter and Query**. Click on the **Add Tables…** button. Click on **Project** in response to the question about where to open the data from, select the **SASUSER.IMPW** data created from the **sales** spreadsheet and click on **OK**. We get a message that we'll need to join the tables manually. This is because we don't have a variable name in common to the data sources to match on. We want to join observations by time period, but in the macro data set the identifying variable is called **date**, whereas in the sales data it's called **Month**. We'll have to tell Enterprise Guide how we want this done.

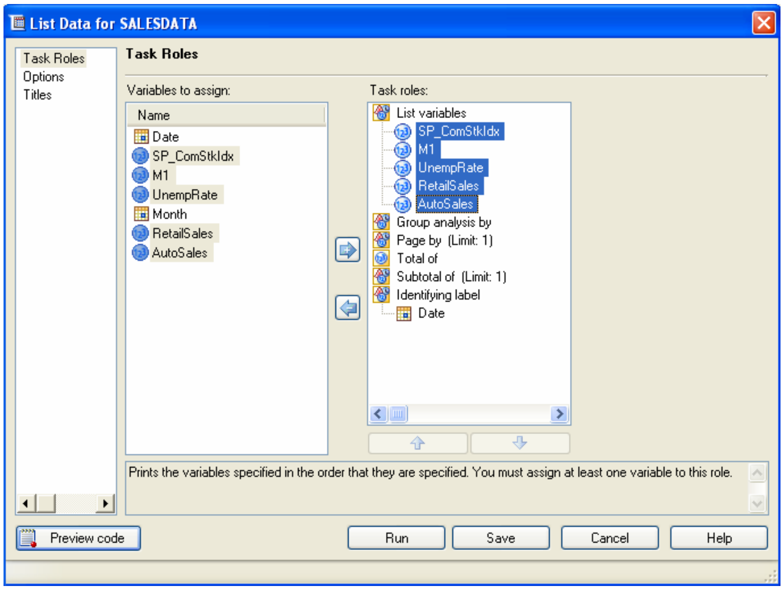
Click on **OK** in the message box. Click on **Date** in the data source created from the macro data set, hold down the left mouse button and drag the cursor to point at **Month** in the IMPW data source created from the sales spreadsheet. Release the mouse button. A symbol indicates how the two sets of data will be joined. Close this window.



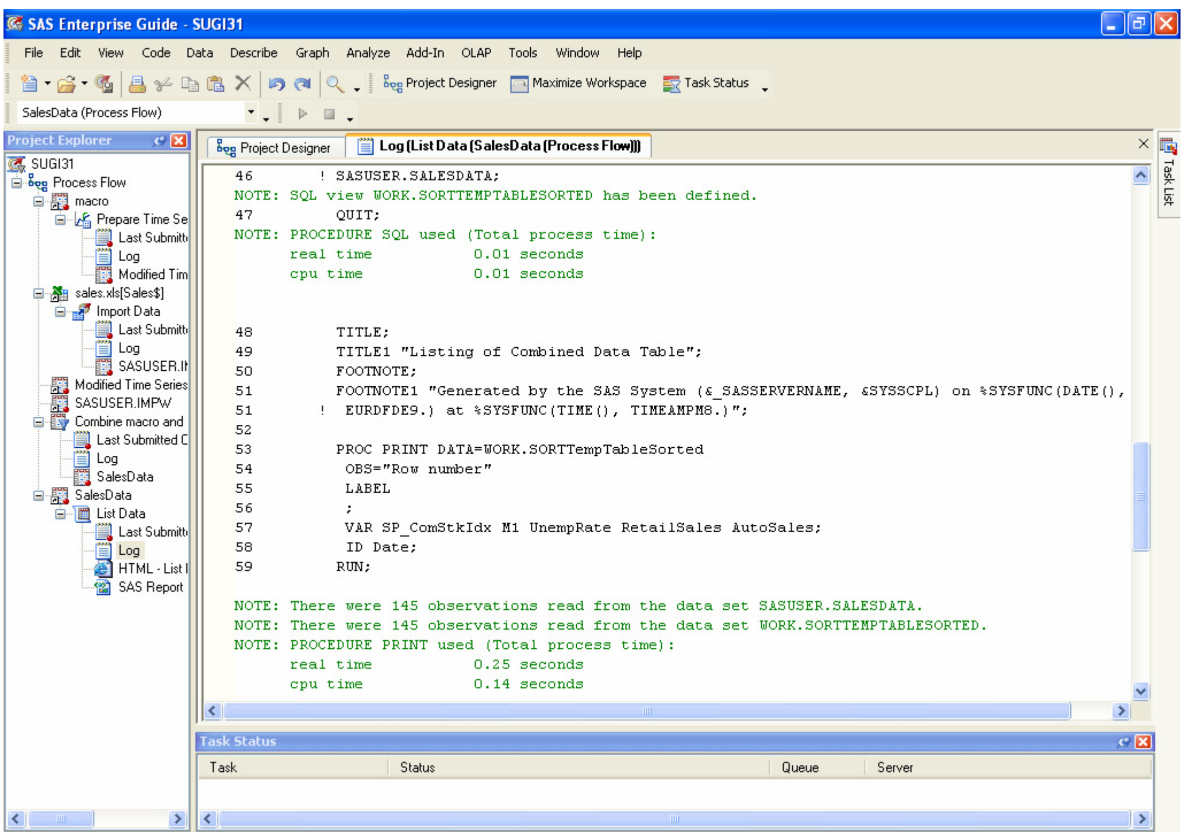
We're going to use all the columns in the two data sources in our combined table, so click on the icon for the **macro** data set, drag it into the **Select Data** area, and do the same for the **sales** spreadsheet. In the box for **Query name:**, type **Combine macro and sales variables**. Click on the **Change…** button to give the resulting data table a better output name. Replace the default name shown with **SalesData**. Click on **Save**, then **Run** to run the query and generate the combined data table.



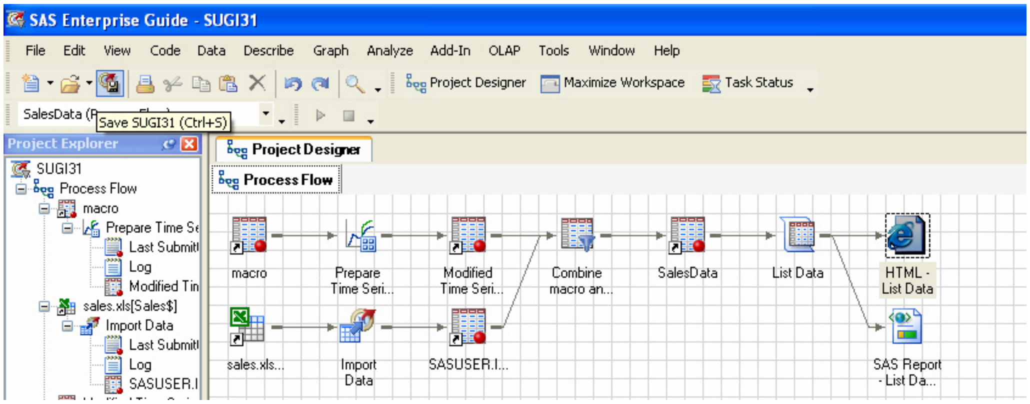
The result of the query, a combined data table, is displayed in the data viewer. An icon for the SalesData data table appears in the Process Flow window. To generate a listing of this new data table, click on **List Data** in the Task List window (under **Describe**). Click and drag **date** to **Identifying label** as a task role. Holding down the **Ctrl** key, click on the five numeric variables to select them. Click and drag them over to the **List variables** role.



Click on **Titles** in the box at the left, deselect **Use default text**, and type **Listing of Combined Data Table**. Click on the **Run** button. The generated results are displayed in the work area in the middle of the screen. In the Project Explorer, there are entries for the **Log** and **Last submitted code** for the task. You can double-click on these to view them, and the code could be edited and re-run.



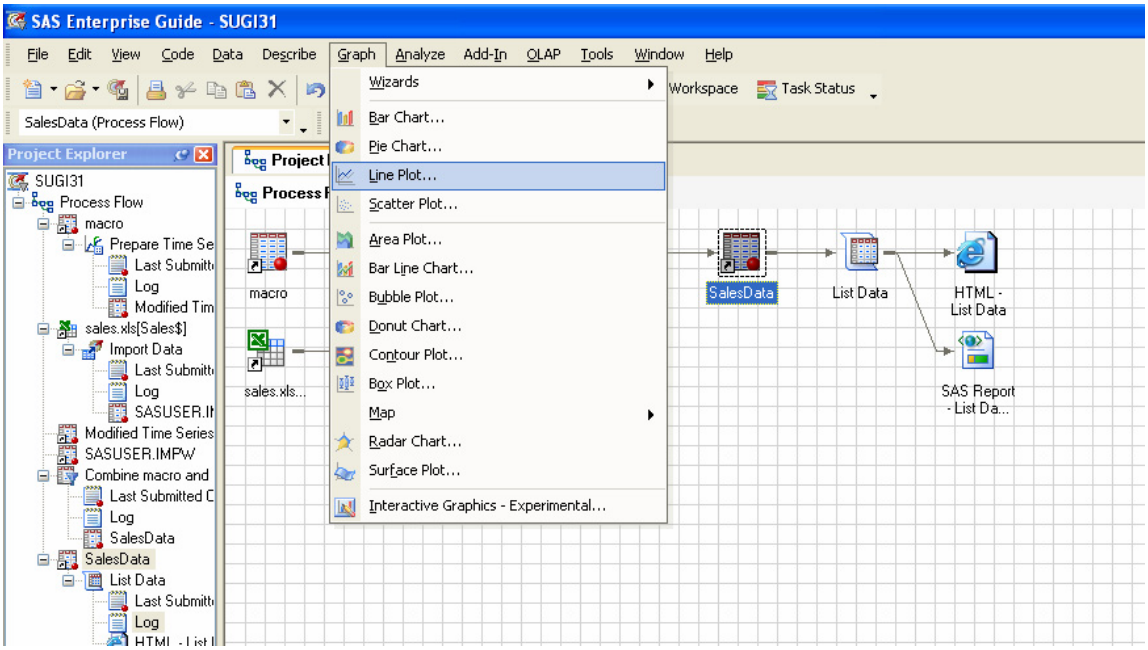
You'll want to periodically click on the **Save project** button on the toolbar. Enterprise Guide will save all of the links listed in your project, so that the next time you access it, everything is ready for you to pick up from there.



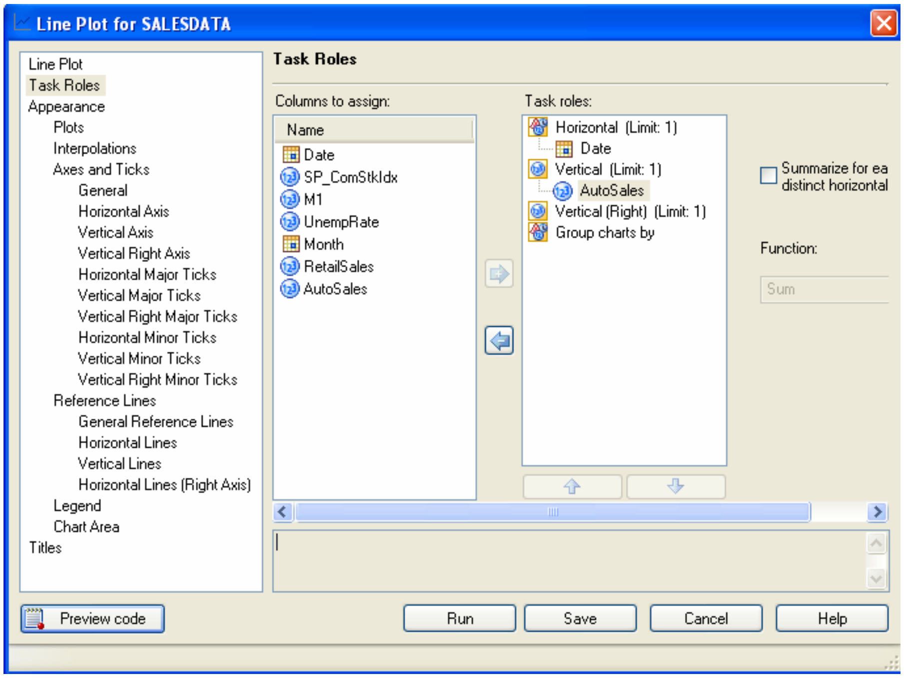
We'll now proceed to data analysis.

EXPLORING THE TIME SERIES

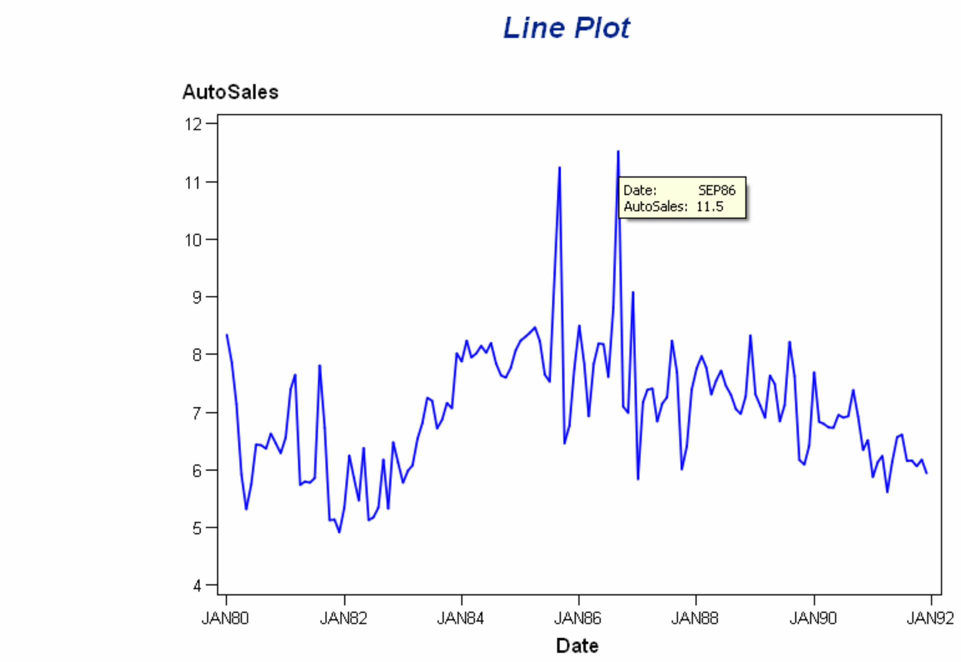
We'll be focusing on the variable **AutoSales** and begin by plotting the series over time. Clicking on the icon for the dataset **SalesData** to make it the active dataset, click on the pulldown menu for **Graph** and select the **Line Plot …** option.



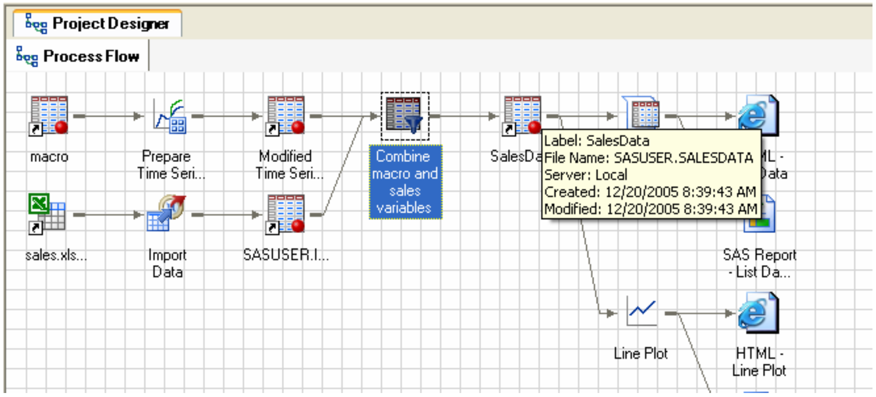
Double-click on the **Line Plot** icon and proceed to the **Task Roles** where the variable **Date** is assigned to the Horizontal axis and **AutoSales** to the Vertical axis.



Click on **Run**. The resulting graph reveals a couple of outliers. The values for these outlying points can be found by moving the mouse pointer over each point.



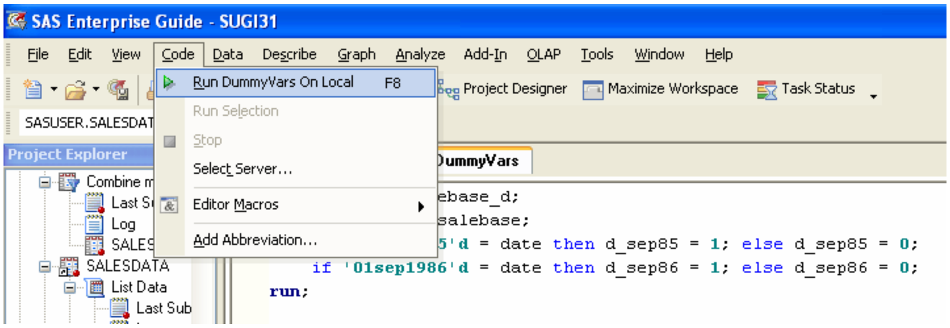
The two extreme values occur in September of 1985 and 1986. To add two dummy variables to our dataset, we can open a **Code** window and write a short **Data step**. We can learn the full two-part name of the SAS dataset **SalesData** by moving the cursor over the icon for the dataset.



Now that we know the full name of the dataset is **SASUSER.SALESDATA**, we can open a **Code** window by clicking on **File, New, Code** and write the following **Data step** in the new **Code** window to create a new SAS dataset **SASUSER.SALESDATA_D**.

```
Data sasuser.salesdata_d;
   set sasuser.salesdata;
   if '01sep1985'd = date then d_sep85 = 1; else d_sep85 = 0;
   if '01sep1986'd = date then d_sep86 = 1; else d_sep86 = 0;
run;
```

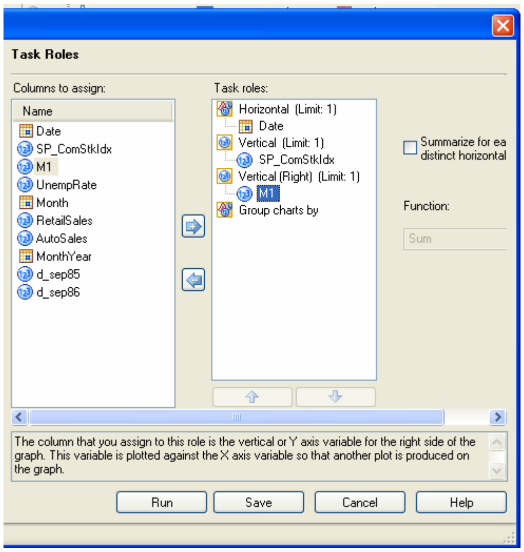
To avoid typing mistakes, the above code has been saved in a file called **DummyVars.sas**. Click on **File, Open, Code…** and navigate to the above file in the folder c:\workshop\ws109. Double-click on the filename and click on **Code**, **Run DummyVars On Local**.

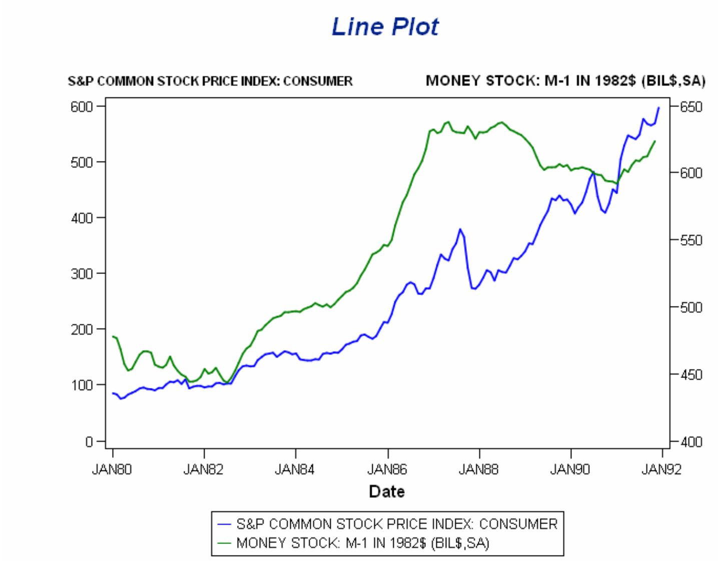


We can verify that everything worked by double-clicking on the icon for the new dataset.

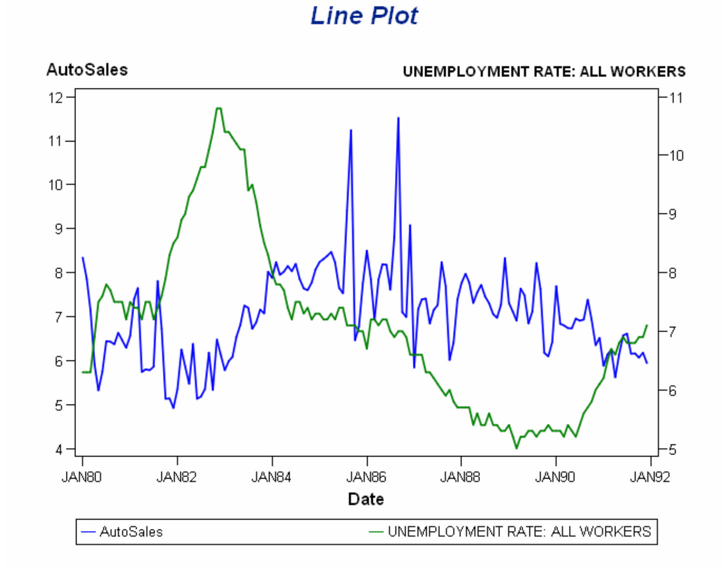
A Regression Model

We'll now fit a regression model for **AutoSales** using some of the macro variables as explanatory variables. We've already examined a plot of **AutoSales**, and it doesn't appear to have a trend. We'll now look at plots of some of the macro variables, first **M1** and **SP_ComStkIdx**. Click on the icon for **SASUSER.SALESDATA_D** and then on **Graph, Line Plot.…** To plot both variables against time on the same graph, assign one variable to the **Vertical** axis and the other to the **Vertical (Right)** axis.

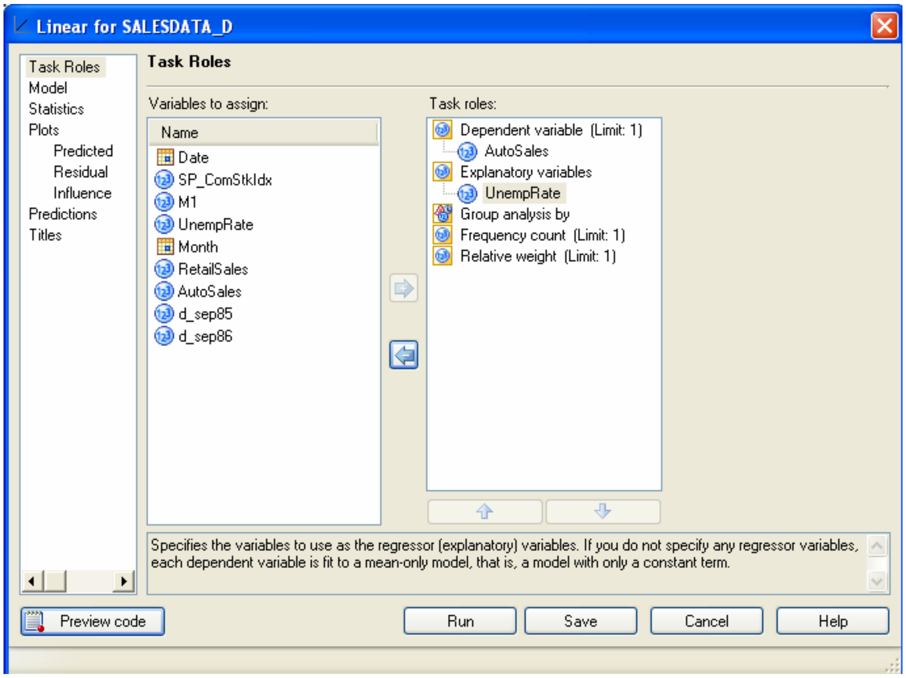




Both variables have an obvious trend. Repeat the above steps to plot **AutoSales** and **UnempRate** on the same graph by double-clicking on the icon for the above **Line Plot** and replying **Yes** to replace the results of the previous run.



UnempRate appears to be negatively correlated with **AutoSales**. A linear regression model is specified by clicking on **Analyze, Regression, Linear…** and assigning **AutoSales** as the dependent variable and **UnempRate** as the explanatory variable. Click on **Run** when done.



The output below shows that **UnempRate** is significant and has a negative sign.

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: AutoSales AutoSales

Number of Observations Read	145
Number of Observations Used	144
Number of Observations with Missing Values	1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	13.98615	13.98615	13.49	0.0003
Error	142	147.27460	1.03715		
Corrected Total	143	161.26075			

Root MSE	1.01840	R-Square	0.0867
Dependent Mean	7.03588	Adj R-Sq	0.0803
Coeff Var	14.47442		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	8.57187	0.42679	20.08	<.0001
UnempRate	UNEMPLOYMENT RATE: ALL WORKERS	1	-0.21685	0.05905	-3.67	0.0003

The R-Square of 0.09 is rather low. We can add the two dummy variables as additional explanatory variables by double-clicking on the icon labeled **Linear.** We see that **UnempRate** remains significantly negative; the two dummy variables are significantly positive, and the R-Square increases to 0.32.

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: AutoSales AutoSales

Number of Observations Read	145
Number of Observations Used	144
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	52.16688	17.38896	22.32	<.0001
Error	140	109.09387	0.77924		
Corrected Total	143	161.26075			

Root MSE	0.88275	R-Square	0.3235
Dependent Mean	7.03588	Adj R-Sq	0.3090
Coeff Var	12.54636		

Parameter Estimates

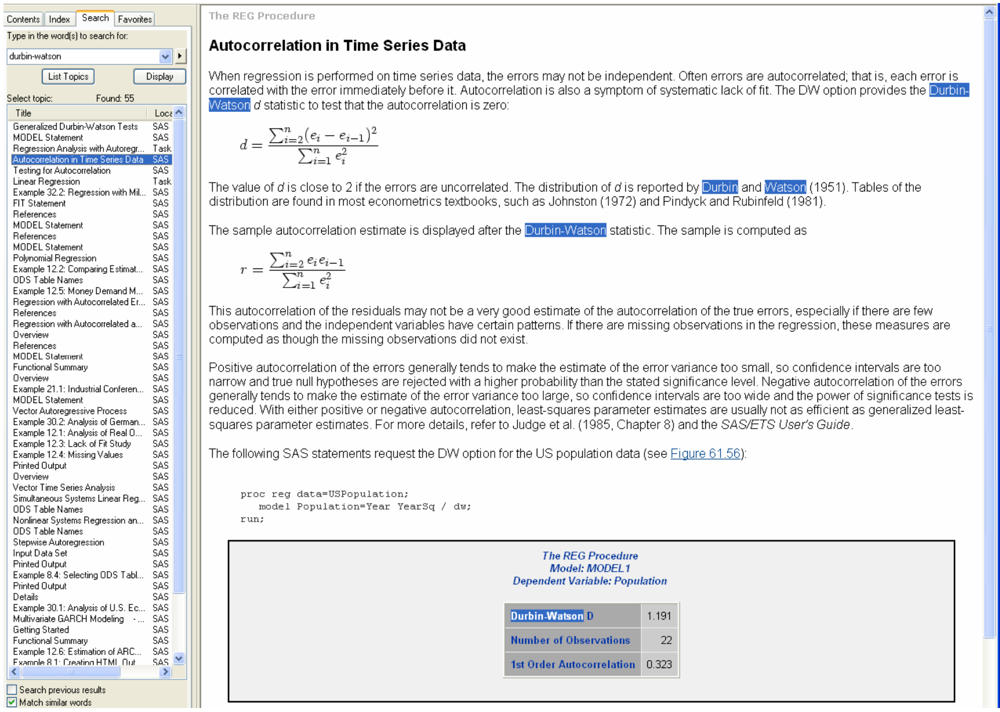
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	8.50349	0.37007	22.98	<.0001
UnempRate	UNEMPLOYMENT RATE: ALL WORKERS	1	-0.21582	0.05118	-4.22	<.0001
d_sep85		1	4.26880	0.88585	4.82	<.0001
d_sep86		1	4.52722	0.88586	5.11	<.0001

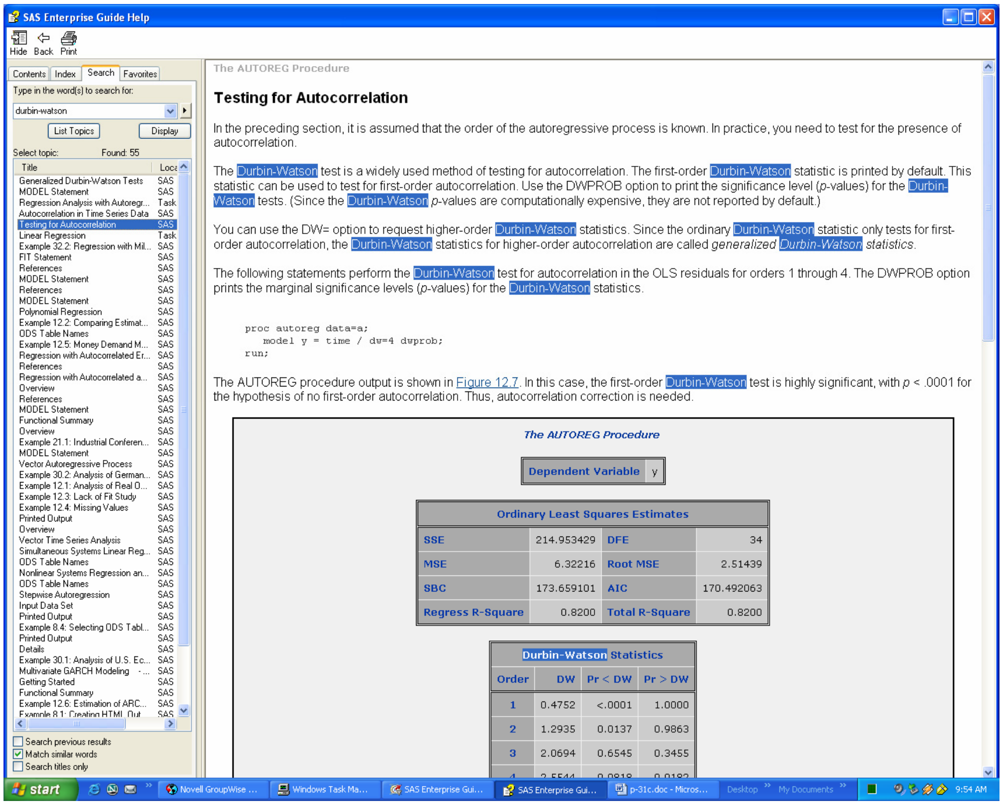
Since we are dealing with time series data, we should check whether or not the residuals exhibit any serial correlation, thus violating one of the OLS regression assumptions. The **Durbin-Watson** statistic can be used to test for 1^{st} order autocorrelation in the residuals.

If you're not familiar with the Durbin-Watson statistic, click on **Help, SAS Enterprise Guide Help** and then click on the **Search** tab, enter durbin-watson, and click on **List Topics**. Finally, click on the fourth topic down labeled **Autocorrelation in Time Series Data**. We see that this Help information is linked to the **REG** procedure.

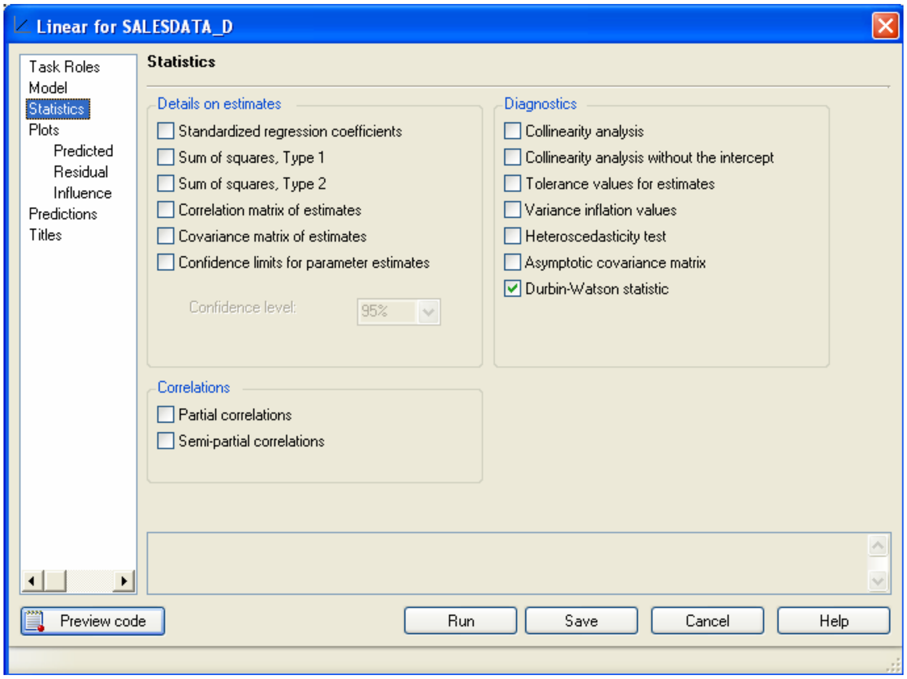
If we then click on the fifth topic labeled **Testing for Autocorrelation**, we get some further information on the DW statistic and see that it is linked to the **AUTOREG** procedure.

The DW statistic d is roughly equal to $2(1 - \rho)$ where ρ is the 1^{st} order autocorrelation coefficient of the residuals. Thus, $0 \leq d \leq 2$ and a value of d between 0 and 1 indicates positive autocorrelation of the residuals.





To request the **DW** statistic, we rerun the **Linear** task and click on the **Statistics** option in the **Task List** in the left window: Check the box labeled **Durbin-Watson statistic** and then click **Run**.



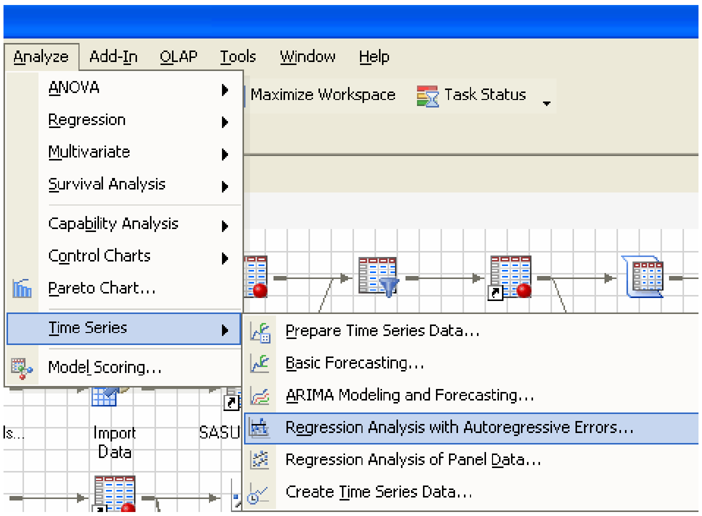
Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: AutoSales AutoSales

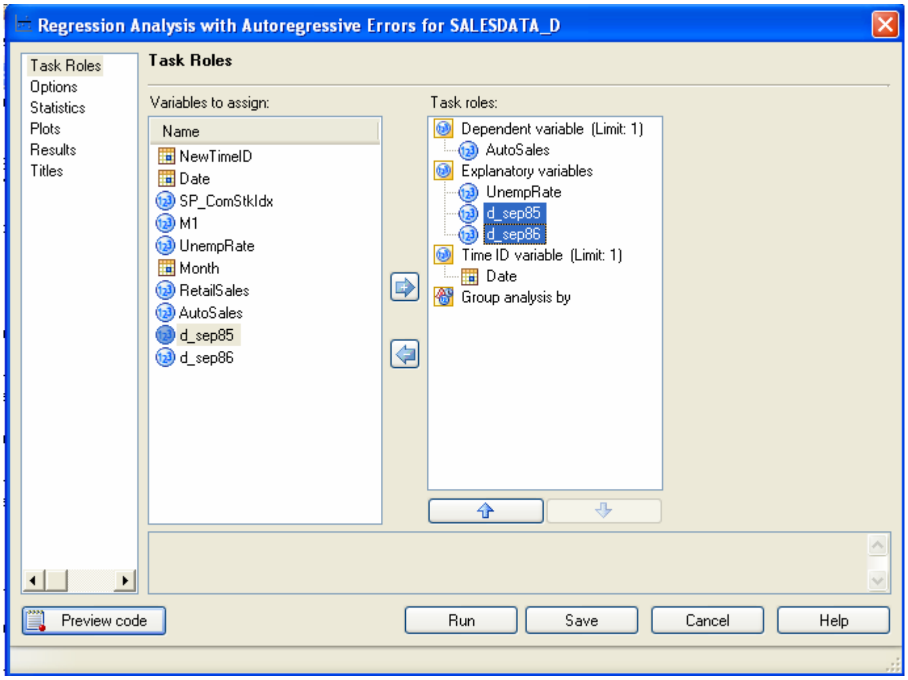
Durbin-Watson D	0.711
Number of Observations	144
1st Order Autocorrelation	0.633

Although a p-value is not reported for d = 0.711, the value of ρ = 0.633 seems high, and we should correct the model for 1st order serial correlation in the residuals. **PROC AUTOREG** is specifically designed for this purpose.

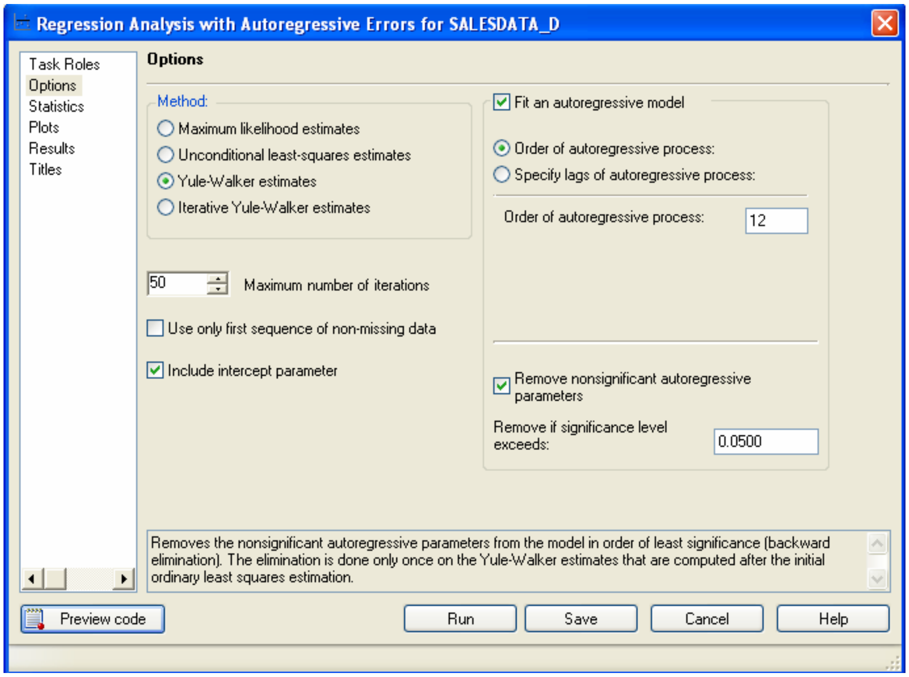
Click on **Analyze, Time Series, Regression Analysis with Autoregressive Errors…** .



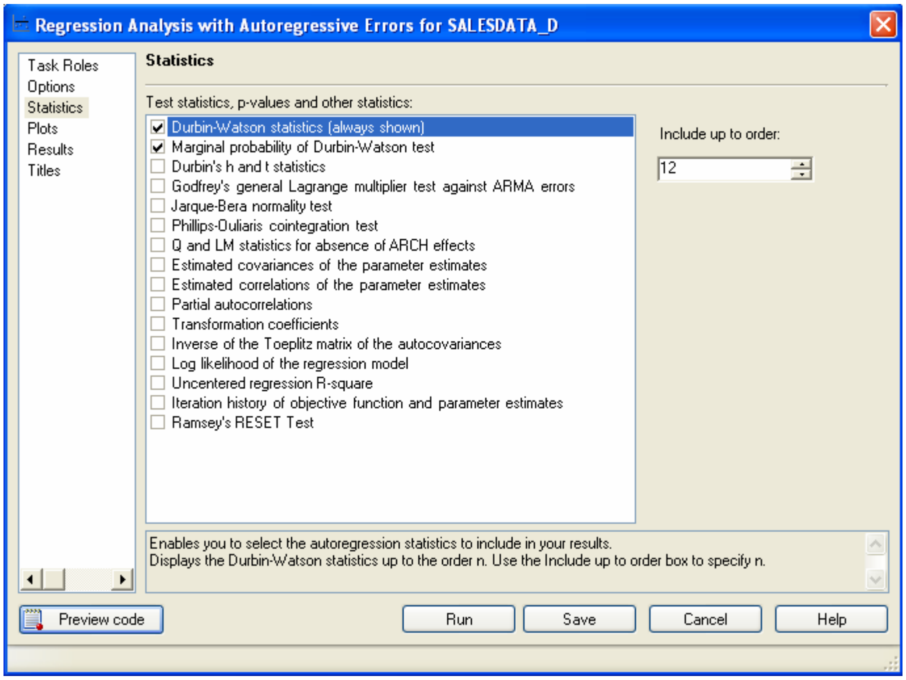
Define the dependent and explanatory variables as before.



Since we have monthly data, we should account for possible autocorrelation up to 12 periods. Click on the **Options** selection in the left window and enter 12 in the box for **Order of autoregressive process**. Also click the box labeled **Remove nonsignificant autoregressive parameters**.



Next click on the **Statistics** option in the left window, note that the Durbin-Watson statistic is already selected as the default, and enter 12 in the box labeled **Include up to order**. Also click on the next box labeled **Marginal probability of Durbin-Watson** test to obtain p-values for the various DW statistics. Then click **Run**.



The first set of results are for the OLS regression model.

The AUTOREG Procedure

Dependent Variable	AutoSales
	AutoSales

Ordinary Least Squares Estimates			
SSE	109.093873	DFE	140
MSE	0.77924	Root MSE	0.88275
SBC	388.558493	AIC	376.67924
Regress R-Square	0.3235	Total R-Square	0.3235

Durbin-Watson Statistics	
Order	DW
1	0.7110
2	1.0466
3	1.1263
4	0.9453
5	0.9253
6	0.8450
7	1.0275
8	1.0217
9	1.0730
10	1.1870
11	1.1678
12	1.0143

We see the same 1st order DW statistic reported by PROC REG.

The OLS parameter estimates are the same as **PROC REG** and the slowly declining Autocorrelation Plot is indicative of serially correlated residuals.

Variable	DF	Estimate	Standard Error	t Value	Approx Pr > |t|	Variable Label
Intercept	1	8.5035	0.3701	22.98	<.0001	
UnempRate	1	-0.2158	0.0512	-4.22	<.0001	UNEMPLOYMENT RATE: ALL WORKERS
d_sep85	1	4.2688	0.8859	4.82	<.0001	
d_sep86	1	4.5272	0.8859	5.11	<.0001	

<table>
<tr><th colspan="4">Estimates of Autocorrelations</th></tr>
<tr><th>Lag</th><th>Covariance</th><th>Correlation</th><th>-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1</th></tr>
<tr><td>0</td><td>0.7576</td><td>1.000000</td><td>| |********************|</td></tr>
<tr><td>1</td><td>0.4796</td><td>0.633096</td><td>| |************* |</td></tr>
<tr><td>2</td><td>0.3483</td><td>0.459743</td><td>| |********* |</td></tr>
<tr><td>3</td><td>0.3150</td><td>0.415806</td><td>| |******** |</td></tr>
<tr><td>4</td><td>0.3767</td><td>0.497274</td><td>| |********** |</td></tr>
<tr><td>5</td><td>0.3731</td><td>0.492530</td><td>| |********** |</td></tr>
<tr><td>6</td><td>0.3986</td><td>0.526191</td><td>| |*********** |</td></tr>
<tr><td>7</td><td>0.3283</td><td>0.433371</td><td>| |********* |</td></tr>
<tr><td>8</td><td>0.3271</td><td>0.431764</td><td>| |********* |</td></tr>
<tr><td>9</td><td>0.2994</td><td>0.395132</td><td>| |******** |</td></tr>
<tr><td>10</td><td>0.2536</td><td>0.334784</td><td>| |******* |</td></tr>
<tr><td>11</td><td>0.2571</td><td>0.339314</td><td>| |******* |</td></tr>
<tr><td>12</td><td>0.3079</td><td>0.406451</td><td>| |******** |</td></tr>
</table>

Preliminary MSE	0.3667

The backward elimination of nonsignificant autoregressive parameters results in significant autocorrelations at lags 1, 4, and 6.

Backward Elimination of Autoregressive Terms			
Lag	**Estimate**	**t Value**	**Pr > |t|**
5	0.002467	0.03	0.9800
3	0.011594	0.12	0.9043
11	-0.016713	-0.18	0.8584
9	-0.041694	-0.45	0.6499
2	0.047958	0.55	0.5805
8	-0.087808	-1.03	0.3065
7	0.057391	0.68	0.4990
12	-0.098392	-1.32	0.1902
10	0.091516	1.20	0.2314

Preliminary MSE	0.3812

Estimates of Autoregressive Parameters			
Lag	**Coefficient**	**Standard Error**	**t Value**
1	-0.439098	0.071825	-6.11
4	-0.218365	0.070389	-3.10
6	-0.209530	0.073556	-2.85

The requested p-values for the DW statistics through order 12 show no autocorrelation left in the residuals through 12 lags. The final parameter estimates show the UnempRate is still significantly negative and the two dummy variables are both significantly positive.

Durbin-Watson Statistics			
Order	**DW**	**Pr < DW**	**Pr > DW**
1	1.9087	0.2854	0.7146
2	2.1741	0.8533	0.1467
3	1.9793	0.4850	0.5150
4	1.8983	0.3492	0.6508
5	1.8886	0.3460	0.6540
6	1.8393	0.2887	0.7113
7	2.0630	0.8022	0.1978
8	1.7671	0.2000	0.8000
9	1.7225	0.1528	0.8472
10	2.0937	0.9040	0.0960
11	1.8752	0.5307	0.4693
12	1.6464	0.1103	0.8897

Note: Pr<DW is the p-value for testing positive autocorrelation, and Pr>DW is the p-value for testing negative autocorrelation.

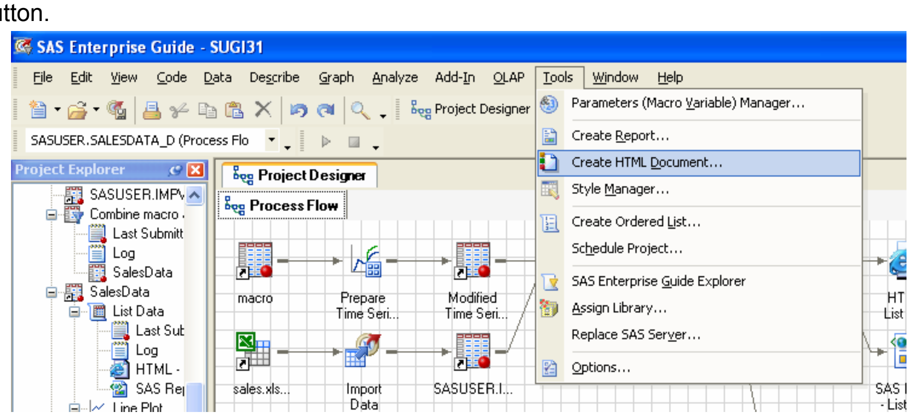
Variable	DF	Estimate	Standard Error	t Value	Approx Pr > |t|	Variable Label
Intercept	1	9.7675	0.9731	10.04	<.0001	
UnempRate	1	-0.4065	0.1300	-3.13	0.0021	UNEMPLOYMENT RATE: ALL WORKERS
d_sep85	1	3.2303	0.5391	5.99	<.0001	
d_sep86	1	4.1726	0.5391	7.74	<.0001	

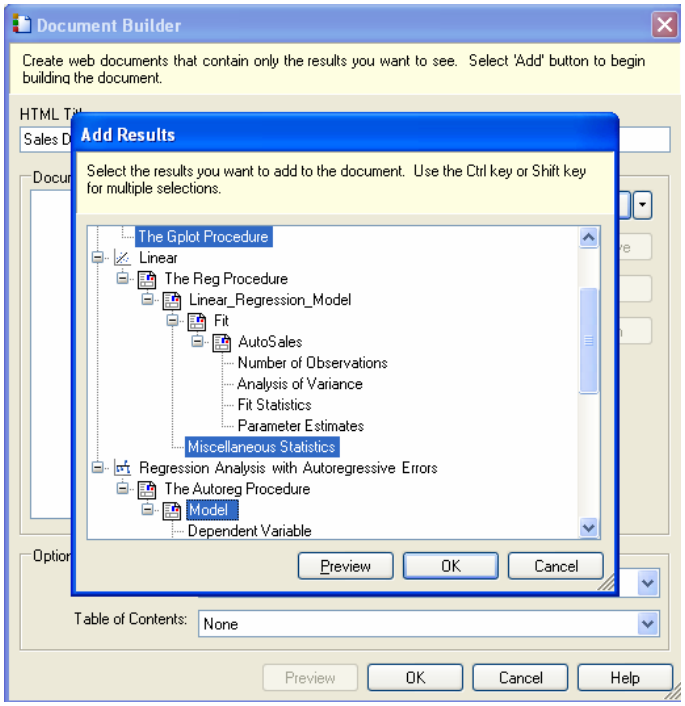
This concludes the modeling part of this workshop.

CREATING A DOCUMENT OF RESULTS

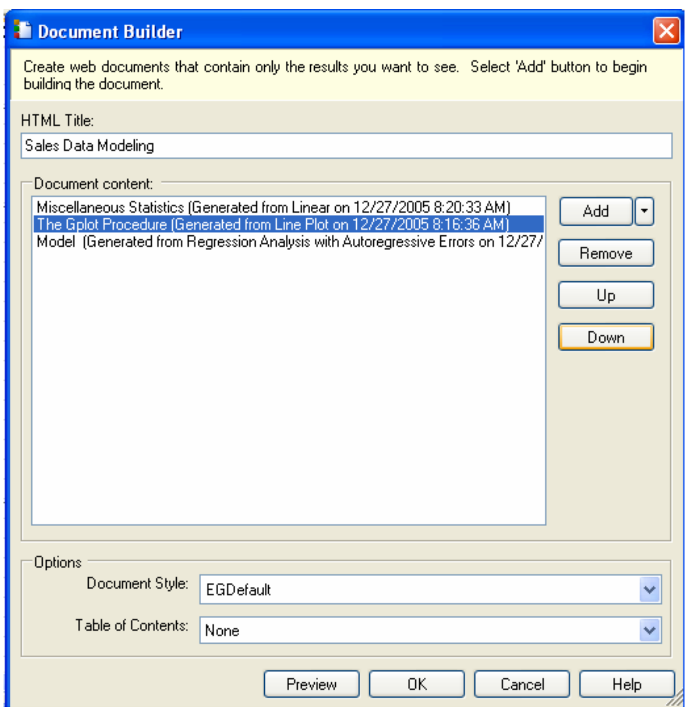
The tasks that we've run have generated a lot of output; some of it for our own exploratory use, other of it suitable for final presentation. We'd like to extract those that fall into the second category and put them together into a report. The Document Builder in Enterprise Guide creates a document definition of instructions for combining HTML results from multiple tasks.

Click on **Tools** → **Create HTML Document...**. Type **Sales Data Modeling** for a document name, and click on **Add**. Holding down the **Ctrl** key, highlight items you'd like to include in a final report. For example, select **The Gplot Procedure** in the Line Plot task, the **Miscellaneous Statistics** table in the Linear_regression_Model task, and **Model** in the Autoreg Procedure task. Click on the **OK** button.

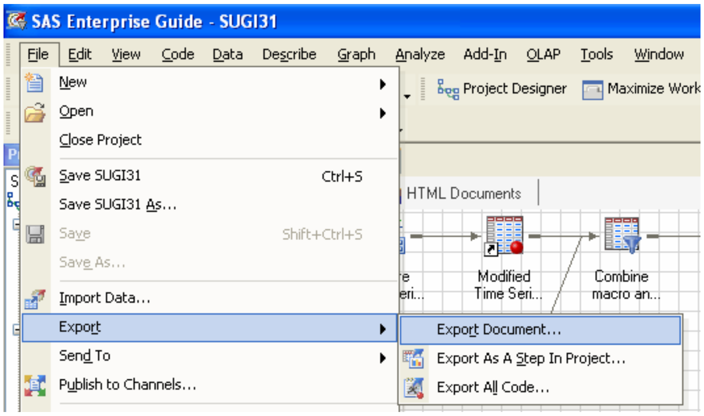




You'll be presented with a list of the items currently in the document and can add further items or remove some of those already included. We can also rearrange the presentation of the output. Highlight the **Gplot Procedure** results and click on the **Down** button to present these after the regression result. Click on the **Preview** button to see what the document looks like so far, scroll through the display and then close the browser.

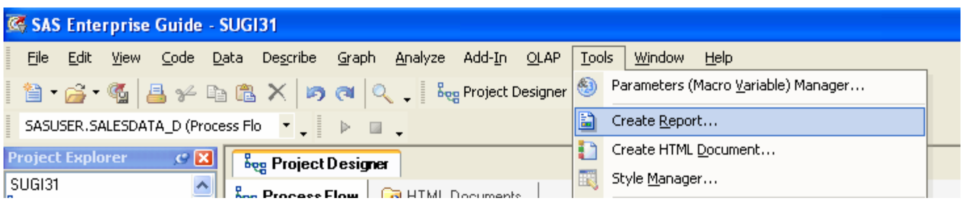


You can reformat the document at other times; double-click on its icon in the Project Window and make further selections to change the definition of the document. To actually save it to a file to make available, for example, on a web site, click on **File→ Export→ Export Document…** and specify the destination you would like.

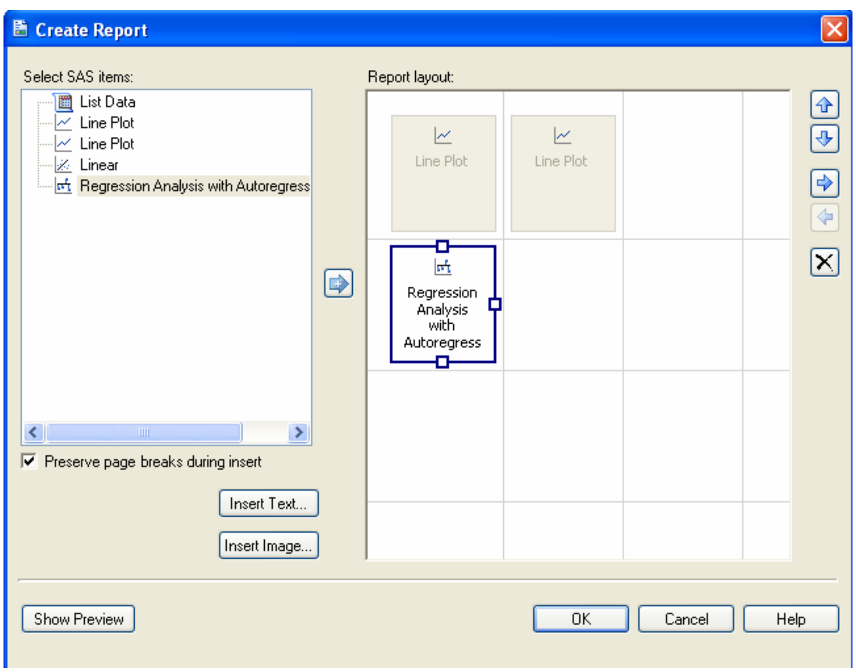


CREATING A SAS REPORT

Results that have been saved in the SAS Report format can be used to create a customized report that can be printed from within Enterprise Guide. Click on **Tools**→ **Create Report…**

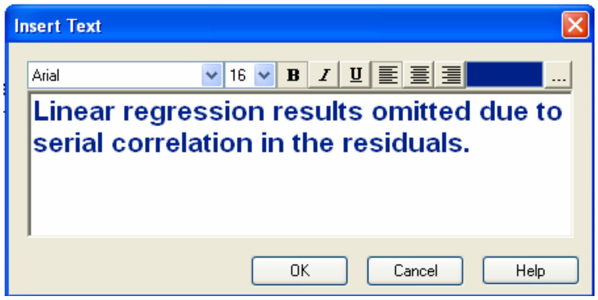


Click and drag items to the cells of the report to indicate where you would like them placed. For example, the two plots followed by the autoregression results:

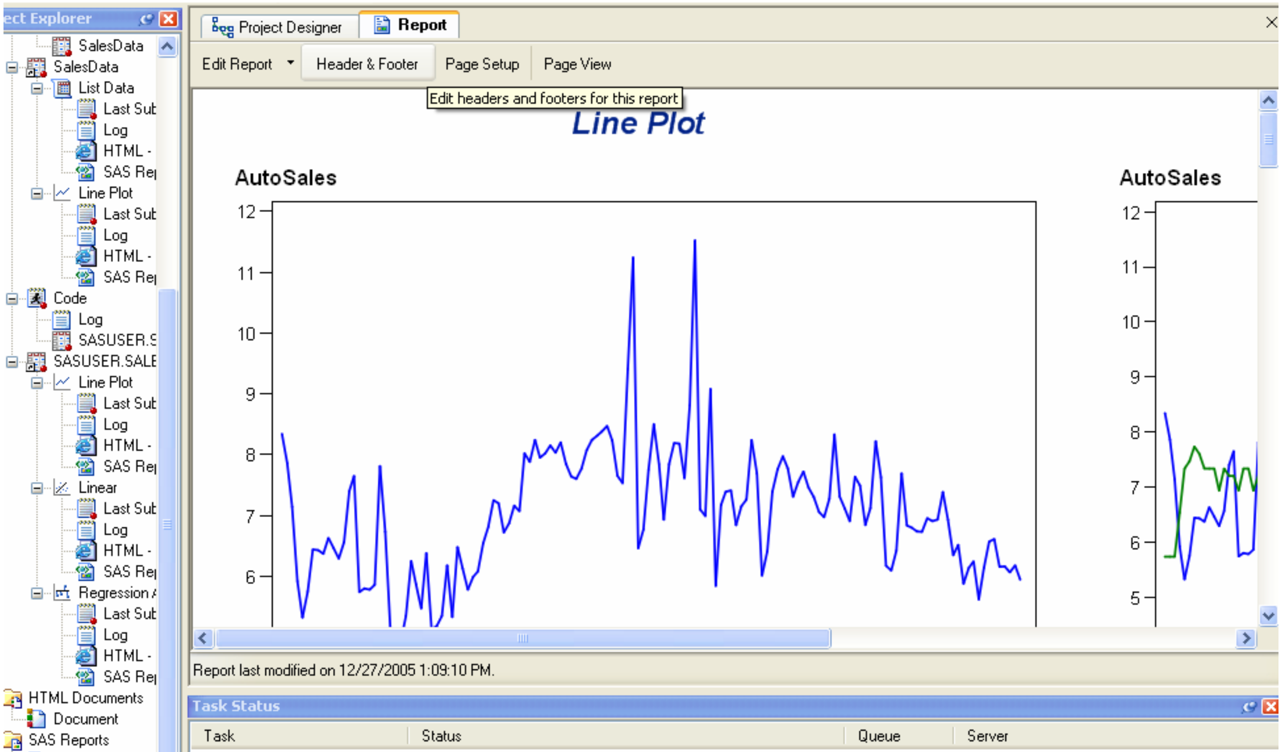


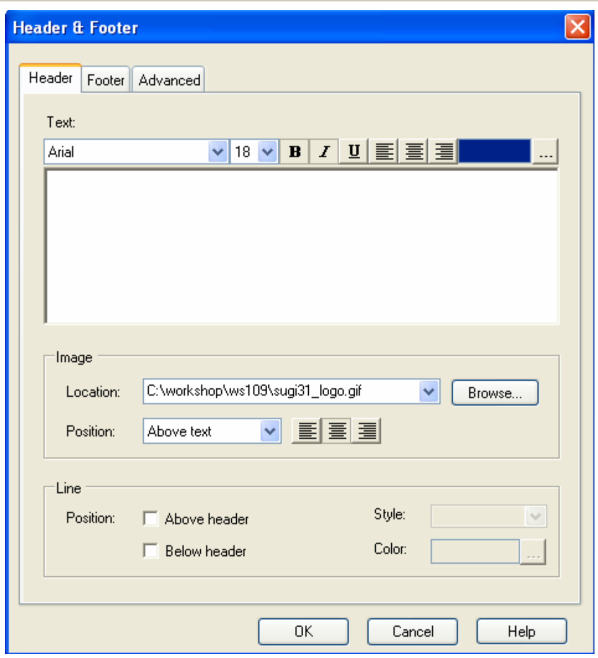
You can use the boxed arrow icons to further move items around. Click on the **Insert Text…** button and type in information about the linear regression results being omitted due to serial correlation (note that you can change font type, size, use bolding, italics, and underlining, and specify text color and justification).

Click on the **OK** button, and then **OK** again to finish defining the report.

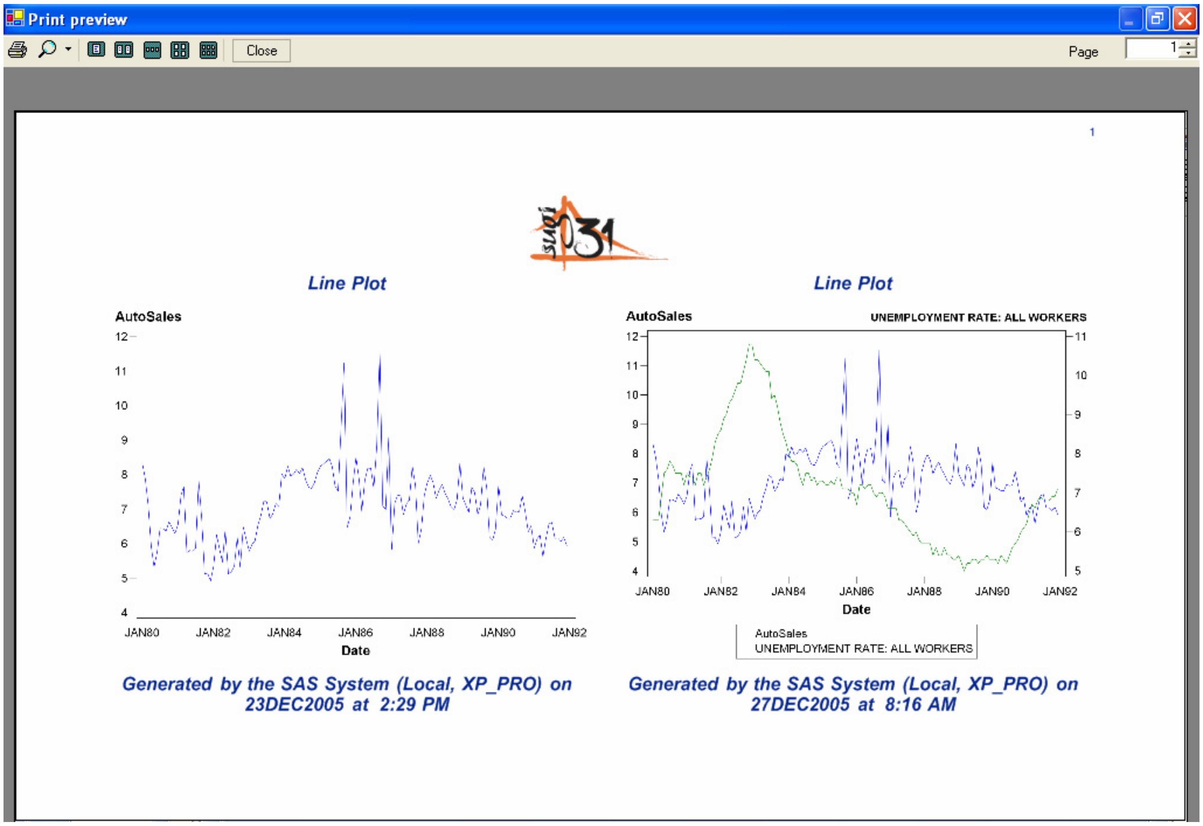


There is now an icon for the report in the project. Double-click on it, and you will have further opportunities to refine it. Click on the **Header & Footer** tab, click on the **Browse** button by **Location** to insert the image **sugi31_logo.gif**. Click **open**, **ok**.





Options in the **Page Setup** tab allow you to change Orientation (for example, from Portrait to Landscape), change Paper Size (for example, to Legal), and so on. Finally, **File→ Print Preview for Report** allows to check what the final report would look like if sent to a printer.



CONCLUSION

Enterprise Guide provides a powerful interface to the suite of tools in the SAS System for conducting statistical analyses. The menus and selection dialogs make it easier to find the correct options available in the analytic procedures. The organization into projects helps group related tasks and the data being examined. EG's facilities for creating HTML documents and SAS reports aid in putting the final results together. An analyst does not need to be a SAS programmer to accomplish these goals.

This workshop has only touched on a few of the capabilities of this software tool. Users are encouraged to explore other options available in the menu system and dialog boxes.

REFERENCES

Accomplishing Tasks in SAS® Using Enterprise Guide® Software Course Notes, SAS Institute Inc., Cary, NC, 2002.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Charles Hallahan
Economic Research Service, USDA
1800 M Street NW, Room S2070
Washington, DC 20036-5831
Work Phone: (202) 694-5051
Fax: (202) 694-5718
Email: hallahan@ers.usda.gov

Linda Atkinson
Economic Research Service, USDA
1800 M Street NW, Room S5015
Washington, DC 20036-5831
Work Phone: (202) 694-5046
Fax: (202) 694-5715
Email: atkinson@ers.usda.gov