

Chapter 306

Multiple Regression with Serial Correlation

Introduction

The regular Multiple Regression routine assumes that the random-error components are independent from one observation to the next. However, this assumption is often not appropriate for business and economic data. Instead, it is more appropriate to assume that the error terms are positively correlated over time. These are called *autocorrelated* or *serially correlated* data.

Consequences of the error terms being serially correlated include inefficient estimation of the regression coefficients, under estimation of the error variance (MSE), under estimation of the variance of the regression coefficients, and inaccurate confidence intervals.

The presence of serial correlation can be detected by the Durbin-Watson test and by plotting the residuals against their lags.

Autoregressive Error Model

When serial correlation is detected, there are several remedies. Since autocorrelation is often caused by leaving important independent variables out of the regression model, an obvious remedy is to add other, appropriate independent variables to the model. When this is not possible, another remedy is to use an autoregressive model. The usual multiple regression model

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_p X_{pt} + \varepsilon_t$$

is modified by adding the equation

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

where

$|\rho| < 1$ is the serial correlation

$$u_t \sim N(0, \sigma^2)$$

The subscript t represents the time period. In econometric work, these u 's are often called the *disturbances*. They are the ultimate error terms. Further details on this model can be found in chapter 12 of Neter, Kutner, Nachtsheim, and Wasserman (1996).

Cochrane-Orcutt Procedure

Several methods have been suggested to estimate the autoregressive error model. We have adopted the Cochrane-Orcutt procedure as given in Neter, Kutner, Nachtsheim, and Wasserman (1996). This is an iterative procedure that involves several steps.

1. *Ordinary least squares.* The regression coefficients are estimated using ordinary least squares. The array of residuals is calculated.
2. *Estimation of ρ .* The serial correlation is estimated from the current residuals ($e_t = Y_t - \hat{Y}_t$) using the formula

$$\hat{\rho} = \frac{\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=2}^n e_{t-1}^2}$$

3. *Obtain transformed data.* A new set of data is created using the formulas.

$$\begin{aligned} Y'_t &= Y_t - \hat{\rho} Y_{t-1} \\ X'_{1t} &= X_{1t} - \hat{\rho} X_{1,t-1} \\ &\vdots \\ X'_{pt} &= X_{pt} - \hat{\rho} X_{p,t-1} \end{aligned}$$

4. *Fit model to transformed data.* Ordinary least squares is used to fit the following multiple regression to the transformed data.

$$Y'_t = b'_0 + b'_1 X'_{1t} + b'_2 X'_{2t} + \cdots + b'_p X'_{pt}$$

5. *Create the regression model for the untransformed data.* The regression equation of the untransformed data is created using the following equations.

$$\begin{aligned} b_0 &= \frac{b'_0}{1 - \hat{\rho}} \\ b_1 &= b'_1 \\ b_2 &= b'_2 \\ &\vdots \\ b_p &= b'_p \end{aligned}$$

The estimated standard errors of the regression coefficients are given by

$$\begin{aligned} s(b_0) &= \frac{s(b'_0)}{1 - \hat{\rho}} \\ s(b_1) &= s(b'_1) \\ s(b_2) &= s(b'_2) \\ &\vdots \\ s(b_p) &= s(b'_p) \end{aligned}$$

6. *Iterate until convergence is reached.* Steps 2 – 4 are then repeated until the value of P stabilizes. Usually, only four or five iterations are necessary.

Multiple Regression with Serial Correlation

7. Calculate Durbin-Watson test on transformed residuals. As a final diagnostic check, the Durbin-Watson test may be run on the residuals $(e'_t = Y'_t - \hat{Y}'_t)$ from the transformed regression model.

Durbin-Watson Test

The Durbin-Watson test is often used to test for positive or negative, first-order, serial correlation. It is calculated as follows

$$DW = \frac{\sum_{j=2}^N (e_j - e_{j-1})^2}{\sum_{j=1}^N e_j^2}$$

The distribution of this test is difficult because it involves the X values. Originally, Durbin-Watson (1950, 1951) gave a pair of bounds to be used. However, there is a large range of 'inclusion' found when using these bounds. Instead of using these bounds, we calculate the exact probability using the beta distribution approximation suggested by Durbin-Watson (1951). This approximation has been shown to be accurate to three decimal places in most cases which is all that are needed for practical work.

Forecasts

The predicted value for a specific set of independent variable values is given by

$$\hat{Y}_t = \hat{b}_0 + \hat{b}_1 X_{1t} + \hat{b}_2 X_{2t} + \cdots + \hat{b}_p X_{pt}$$

For forecasts j periods into the future after the end of the series (period n is the final period on which we have data), the formula is

$$F_{n+j} = \hat{b}_0 + \hat{b}_1 X_{1,n+j} + \hat{b}_2 X_{2,n+j} + \cdots + \hat{b}_p X_{p,n+j} + \hat{\rho}^j e_n$$

where e_n is the residual from the final observation. That is,

$$e_n = Y_n - \hat{Y}_n$$

The approximate $1 - \alpha$ prediction interval for this forecast is

$$F_{n+j} \pm t_{1-\alpha/2, n-3} s_F$$

where s_F is the standard error of the prediction interval based on the transformed data.

Data Structure

The data are entered in two or more variables. An example of data appropriate for this procedure is shown below. These data give the annual values for several economic statistics. Later in this chapter, these data will be used in an example in which Housing is forecast from Mort5Yr and DispInc. These data are stored in a dataset called Housing.NCSS. Note that only two decimal places are displayed here, while on the database, more decimal places are stored.

Multiple Regression with Serial Correlation

Housing dataset (subset)

Year	Housing	Mort5Yr	Displnc	TBill	Unemp_rt
1981	403.34	18.25	27006.90	17.72	7.57
1982	407.92	17.93	26896.58	13.66	10.97
1983	446.87	13.17	26582.63	9.31	11.94
1984	457.22	13.54	27662.41	11.06	11.30
1985	485.25	12.08	28710.13	9.43	10.65
1986	475.87	11.17	29057.02	8.97	9.64
1987	491.30	11.12	29626.58	8.15	8.82
1988	493.23	11.61	31070.52	9.48	7.75
1989	487.14	12.01	32417.38	12.05	7.55
1990	491.00	13.31	32683.10	12.81	8.12
1991	512.39	11.07	31980.30	8.73	10.32
1992	523.07	9.50	32224.67	6.59	11.16
1993	533.20	8.76	32412.84	4.84	11.36
1994	497.75	9.53	32789.41	5.54	10.36
1995	502.59	9.14	33242.99	6.89	9.45
1996	522.73	7.91	33256.65	4.21	9.64
1997	538.72	7.05	33839.28	3.26	9.10
1998	533.61	6.92	34915.04	4.73	8.29
1999	531.89	7.54	35971.46	4.72	7.57
2000	528.09	8.32	37566.34	5.49	6.81
2001	544.91	7.38	38228.92	3.77	7.20
2002	547.70	6.99	38806.22	2.59	7.66
2003	561.19	6.36	38896.05	2.87	7.63
2004	581.54	5.38	39870.12	2.30	7.45
2005		6.00	41000.00		
2006		6.25	42000.00		

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Dependent Variable

Y: Dependent Variable(s)

This option specifies one or more dependent (Y) variables. If more than one variable is specified, a separate analysis is run for each.

Numeric Independent Variables

X's: Numeric Independent Variable(s)

Specify any numeric independent variables in this box. Numeric variables are those whose values are numeric and are at least ordinal. Nominal variables, even when coded with numbers, should be specified as Categorical Independent Variables. Although you may specify binary (0-1) variables here, they are more appropriately analyzed when you specify them as Categorical Independent Variables.

Multiple Regression with Serial Correlation

If you want to create powers and cross-products of these variables, specify an appropriate model in the 'Custom Model' field under the Model tab.

If you want to create predicted values of Y for values of X not in your database, add the X values to the bottom of the database. These rows will not be used during estimation phase, but predicted values will be generated for them on the reports.

Categorical Independent Variables

X's: Categorical Independent Variable(s)

Specify categorical (nominal) independent variables in this box. By categorical we mean that the variable has only a few unique, numeric or text, values like 1, 2, 3 or Yes, No, Maybe. The values are used to identify categories. For further details about categorical variables, see the discussion on this topic in the Multiple Regression chapter.

Default Reference Value

This option specifies the default reference value to be used when automatically generating indicator variables during the processing of selected categorical independent variables. The reference value is often the baseline, and the other values are compared to it. For further details about this value, see the discussion on this topic in the Multiple Regression chapter.

Default Contrast Type

Select the default type of numeric variable that will be generated when processing categorical independent variables. The values in a categorical variable are not used directly in regression analysis. Instead, a set of numeric variables is automatically created and substituted for them. This option allows you to specify what type of numeric variable will be created. For further details about this option, see the discussion on this topic in the Multiple Regression chapter.

Estimation Options

Maximum Cochrane-Orcutt Iterations

This is the maximum number of iterations that the procedure will cycle through. Some authors recommend only one iteration. Others recommend stopping once the Durbin-Watson test is not significant. This option lets you stop after a specific number of iterations. Usually, four or five iterations should be plenty.

Minimum Rho Change

If the change in rho (serial correlation) from one iteration to the next is less than this amount, the algorithm will stop iterating. We suggest you use a small amount such as 0.00001.

Alpha Levels

Alpha of C.I.'s and Tests

The value of alpha for the statistical tests and confidence intervals is specified here. Usually, this number will range from 0.1 to 0.001. A common choice for alpha is 0.05, but this value is a legacy from the age before computers when only printed tables were available. You should determine a value appropriate for your particular study.

Alpha of Assumptions

This value specifies the significance level that must be achieved to reject a preliminary test of an assumption. In regular hypothesis tests, common values of alpha are 0.05 and 0.01. However, most statisticians recommend that preliminary tests use a larger alpha such as 0.10, 0.15, or 0.20.

We recommend 0.20.

Model Tab

These options control the regression model.

Model Specification

Which Model Terms

This option specifies which terms (terms, powers, cross-products, and interactions) are included in the regression model. For a time-series regression model, select *Up to 1-Way*.

The other options on this tab are covered in detail in the Multiple Regression chapter. We refer you to that chapter for further details.

Reports Tab

The following options control which reports and plots are displayed.

Select Reports

Run Summary ... Residuals

Each of these options specifies whether the indicated report is calculated and displayed. Note that since some of these reports provide results for each row, they may be too long for normal use when requested on large databases.

Report Options

Show All Rows

This option makes it possible to display predicted values for only a few designated rows.

When checked predicted values, residuals, and other row-by-row statistics, will be displayed for all rows used in the analysis.

When not checked, predicted values and other row-by-row statistics will be displayed for only those rows in which the dependent variable's value is missing.

Precision

Specifies the precision of numbers in the report. Single precision will display seven-place accuracy, while the double precision will display thirteen-place accuracy.

Variable Names

This option lets you select whether to display variable names, variable labels, or both.

Skip Line After

The names of the indicator variables can be too long to fit in the space provided. If the name contains more characters than the number specified here, only the name is shown on the first line of the report and the rest of the output is placed on the next line.

Enter *1* when you want the each variable's results printed on two lines.

Enter *100* when you want each variable's results printed on a single line.

Multiple Regression with Serial Correlation

Report Options – Decimal Places

Probability ... Mean Square Decimals

Specify the number of digits after the decimal point to display on the output of values of this type. Note that this option in no way influences the accuracy with which the calculations are done.

Enter 'All' to display all digits available. The number of digits displayed by this option is controlled by whether the PRECISION option is SINGLE or DOUBLE.

Plots Tab

These options control the titles and style files used on each of the plots.

Select Plots

Histogram ... Residuals vs X Plot

Indicate whether to display these plots. Click the plot format button to change the plot settings.

Storage Tab

These options let you specify if, and where on the dataset, various statistics are stored.

Warning: Any data already in these columns are replaced by the new data. Be careful not to specify columns that contain important data.

Data Storage Options

Storage Option

This option controls whether the values indicated below are stored on the database when the procedure is run.

- **Do not store data**
No data are stored even if they are checked.
- **Store in empty columns only**
The values are stored in empty columns only. Columns containing data are not used for data storage, so no data can be lost.
- **Store in designated columns**
Beginning at the *Store First Item In* column, the values are stored in this column and those to the right. If a column contains data, the data are replaced by the storage values. Care must be used with this option because it cannot be undone.

Store First Item In

The first item is stored in this column. Each additional item that is checked is stored in the columns immediately to the right of this column.

Leave this value blank if you want the data storage to begin in the first blank column on the right-hand side of the data.

Warning: any existing data in these columns is automatically replaced, so be careful.

Data Storage Options – Select Items to Store

Predicted Y ... Upper C.L. Individual

Indicate whether to store these row-by-row values, beginning at the column indicated by the *Store First Item In* option.

Example 1 – Generating Forecasts (All Reports)

This section presents an example of how to generate forecasts for housing data that was presented earlier in this chapter. This data is stored in the Housing dataset. We suggest that you open it now.

This example will run an adjusted multiple regression of *Housing* on *Mort5Yr* and *DispInc*. The adjustment will use the Cochrane-Orcutt procedure. The data for housing ends in 2004. Forecasts will be generated for the years 2005 and 2006.

You may follow along here by making the appropriate entries or load the completed template **Example 1** by clicking on Open Example Template from the File menu of the Multiple Regression with Serial Correlation window.

1 Open the Housing dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **Housing.NCSS**.
- Click **Open**.

2 Open the Multiple Regression with Serial Correlation window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Multiple Regression with Serial Correlation** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Multiple Regression with Serial Correlation window, select the **Variables tab**.
- Set the **Y: Dependent Variables** box to **Housing**.
- Set the **X's: Numeric Independent Variables** box to **DispInc, Mort5Yr**.
- Set the **Maximum Cochrane-Orcutt Iterations** to **1**.

4 Specify the reports.

- Select the **Reports tab**.
- Make sure all reports and plots are checked.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Multiple Regression with Serial Correlation

Run Summary Section

Parameter	Value	Parameter	Value
Dependent Variable	Housing	Rows Processed	32
Number Ind. Variables	2	Rows Filtered Out	0
Weight Variable	None	Rows with X's Missing	0
R2	0.8941	Rows with Weight Missing	0
Adj R2	0.8860	Rows with Y Missing	2
Coefficient of Variation	0.0361	Rows Used in Estimation	30
Mean Square Error	77.15598	Sum of Weights	29.000
Square Root of MSE	8.783848	Completion Status	Normal Completion
Ave Abs Pct Error	1.801	Autocorrelation (Rho)	0.5121

This report summarizes the multiple regression results. It presents the variables used, the number of rows used, and the basic results. The estimated value of the autocorrelation (rho) has been added to this report. Otherwise, it is identical to the corresponding report in the regular Multiple Regression report.

Note that values such as R2, Mean Square Error, etc., are calculated on the transformed data.

Descriptive Statistics Section

Variable	Count	Mean	Standard Deviation	Minimum	Maximum
Displnc	29	31000.94	5157.438	21780.1	39870.13
Mort5Yr	29	10.53919	3.183494	5.380194	18.25095
Housing	29	491.0214	48.53722	403.3378	581.5398

For each variable, the count, arithmetic mean, standard deviation, minimum, and maximum are computed. This report is particularly useful for checking that the correct variables were selected.

Correlation Matrix Section

	Displnc	Mort5Yr	Housing
Displnc	1.0000	-0.5962	0.7913
Mort5Yr	-0.5962	1.0000	-0.8874
Housing	0.7913	-0.8874	1.0000

Pearson correlations are given for all variables.

Regression Equation Section

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	T-Value to test H0:B(i)=0	Prob Level	Reject H0 at 5%?
Intercept	445.1365	35.4690	12.550	0.0000	Yes
Displnc	0.0044	0.0009	5.118	0.0000	Yes
Mort5Yr	-8.5371	1.0522	-8.113	0.0000	Yes

Estimated Model

445.136489079996+ 4.4434007069797E-03*Displnc-8.53714263704248*Mort5Yr

This section reports the values and significance tests of the regression coefficients. Note that the intercept has been corrected by dividing by 1-rho. Other than this, the report has the same definitions as in regular Multiple Regression.

Multiple Regression with Serial Correlation

Regression Coefficient Section

Independent Variable	Regression Coefficient	Standard Error	Lower 95% C.L.	Upper 95% C.L.	Standardized Coefficient
Intercept	445.1365	35.4690	372.2289	518.0441	0.0000
Displnc	0.0044	0.0009	0.0027	0.0062	0.4068
Mort5Yr	-8.5371	1.0522	-10.7000	-6.3743	-0.6449

Note: The T-Value used to calculate these confidence limits was 2.056.

The report has the same definitions as in regular Multiple Regression.

Analysis of Variance Section

Source	DF	R2	Sum of Squares	Mean Square	F-Ratio	Prob Level
Intercept	1		1720724	1720724		
Model	2	0.8941	16943.61	8471.806	109.801	0.0000
Error	26	0.1059	2006.055	77.15598		
Total(Adjusted)	28	1.0000	18949.67	676.7738		

This section reports the analysis of variance table. Note it was calculated from the transformed data on the last iteration. Other than this, the report has the same definitions as in regular Multiple Regression.

Serial-Correlation and Durbin-Watson Test

Serial Correlation of Residuals from Uncorrected Model

Lag	Serial Correlation	Lag	Serial Correlation	Lag	Serial Correlation
1	0.5090	9	-0.4075	17	-0.1140
2	0.1980	10	-0.5085	18	-0.0147
3	0.0802	11	-0.3018	19	0.1512
4	0.0505	12	-0.1962	20	0.1290
5	0.2072	13	-0.1042	21	0.0519
6	0.2165	14	-0.1067	22	0.0275
7	-0.0649	15	-0.3178	23	0.0457
8	-0.0979	16	-0.2177	24	0.0875

Above serial correlations significant if their absolute values are greater than 0.365148

Serial Correlation of Residuals from Corrected Model

Lag	Serial Correlation	Lag	Serial Correlation	Lag	Serial Correlation
1	0.0261	9	-0.2371	17	0.0817
2	-0.0349	10	-0.3626	18	0.0420
3	0.0972	11	-0.0584	19	0.0314
4	-0.1182	12	0.0042	20	0.0473
5	0.1002	13	-0.0671	21	0.0248
6	0.2071	14	-0.0042	22	0.0761
7	-0.3095	15	-0.2443	23	0.0038
8	0.1301	16	0.0617	24	0.0388

Above serial correlations significant if their absolute values are greater than 0.371391

Durbin-Watson Test For Serial Correlation of Uncorrected Model

Parameter	Value	Did the Test Reject H0: $\rho(1) = 0$?
Durbin-Watson Value	0.9234	
Prob. Level: Positive Serial Correlation	0.0002	Yes
Prob. Level: Negative Serial Correlation	0.9974	No

Multiple Regression with Serial Correlation

Durbin-Watson Test For Serial Correlation of Corrected Model

Parameter	Value	Did the Test Reject H0: $\rho(1) = 0$?
Durbin-Watson Value	1.9221	
Prob. Level: Positive Serial Correlation	0.3273	No
Prob. Level: Negative Serial Correlation	0.4923	No

This section reports the autocorrelation structure of the residuals both before and after the model is corrected for serial correlation. It has the same definitions as in the regular Multiple Regression report.

Predicted Values with Confidence Limits of Means

Row	Actual Housing	Predicted Housing	Standard Error Of Predicted	95% Lower Conf. Limit Of Mean	95% Upper Conf. Limit Of Mean
1	420.722	445.738			
2	431.522	447.504	3.273	440.776	454.232
3	448.085	462.874	4.140	454.364	471.384
.
.
.
26	528.086	541.005	3.211	534.406	547.605
27	544.913	552.012	2.883	546.086	557.938
28	547.703	557.854	3.010	551.667	564.041
29	561.186	563.635	2.950	557.571	569.698
30	581.540	576.364	3.536	569.095	583.633
31		578.744	3.901	570.725	586.762
32		579.760	4.192	571.143	588.377

Confidence intervals for the mean response of Y given specific levels for the IV's are provided here.

Predicted Values with Prediction Limits of Individuals

Row	Actual Housing	Predicted Housing	Standard Error Of Predicted	95% Lower Pred. Limit Of Individual	95% Upper Pred. Limit Of Individual
1	420.722	445.738			
2	431.522	447.504	9.374	428.235	466.772
3	448.085	462.874	9.711	442.914	482.835
.
.
.
26	528.086	541.005	9.352	521.782	560.229
27	544.913	552.012	9.245	533.009	571.015
28	547.703	557.854	9.285	538.768	576.940
29	561.186	563.635	9.266	544.588	582.681
30	581.540	576.364	9.469	556.900	595.828
31		578.744	9.611	558.988	598.500
32		579.760	9.733	559.753	599.766

A prediction interval for the individual response of Y given specific values of the IV's is provided here for each row. Note that the forecasts start where the actual housing values are blank.

Multiple Regression with Serial Correlation

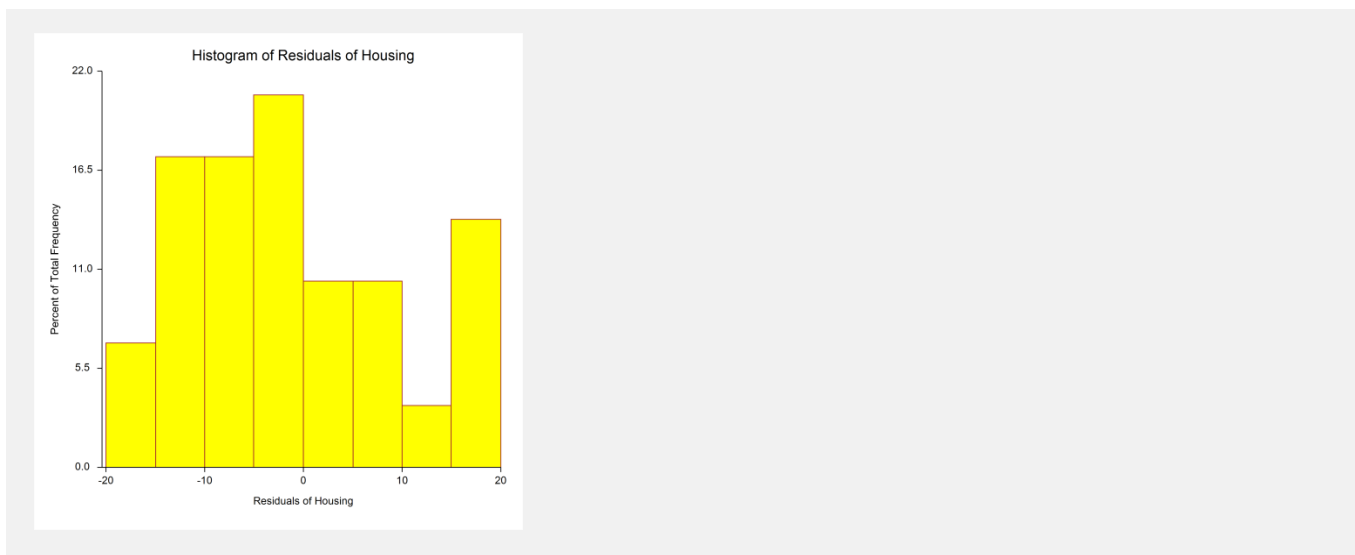
Residual Report

Row	Actual Housing	Predicted Housing	Residual	Absolute Percent Error
1	420.722	445.738		
2	431.522	447.504	-15.982	3.704
3	448.085	462.874	-14.789	3.300
4	447.923	464.696	-16.773	3.745
5	451.401	454.340	-2.939	0.651
6	432.474	438.269	-5.795	1.340
7	403.338	409.328	-5.990	1.485
8	407.922	411.549	-3.627	0.889
.
.
.

This section reports on the sample residuals, or e_i 's.

Histogram

The purpose of the histogram and density trace of the residuals is to evaluate whether they are normally distributed. A dot plot is also given that highlights the distribution of points in each bin of the histogram. Unless you have a large sample size, it is best not to rely on the histogram for visually evaluating normality of the residuals. The better choice would be the normal probability plot.

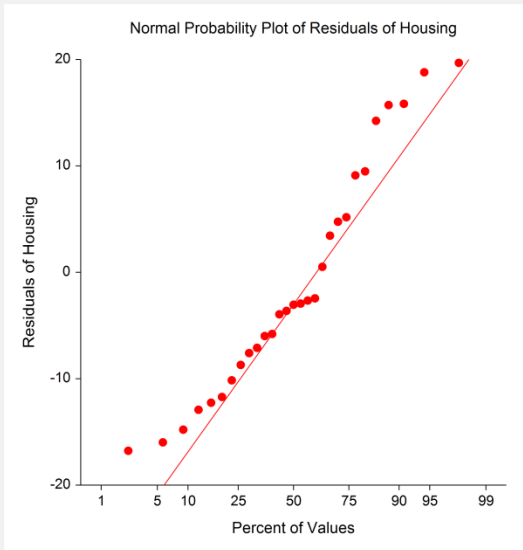


Multiple Regression with Serial Correlation

Probability Plot of Residuals

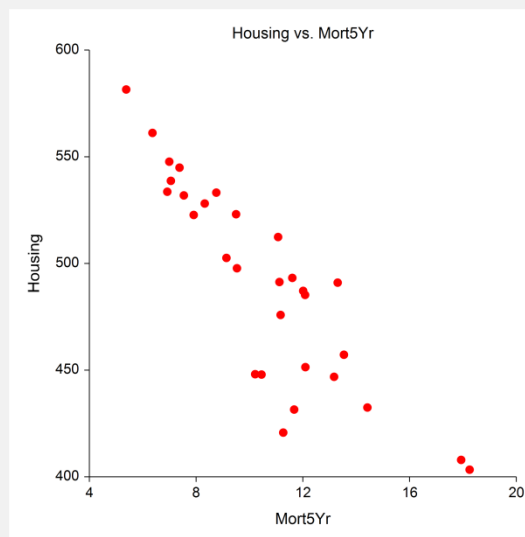
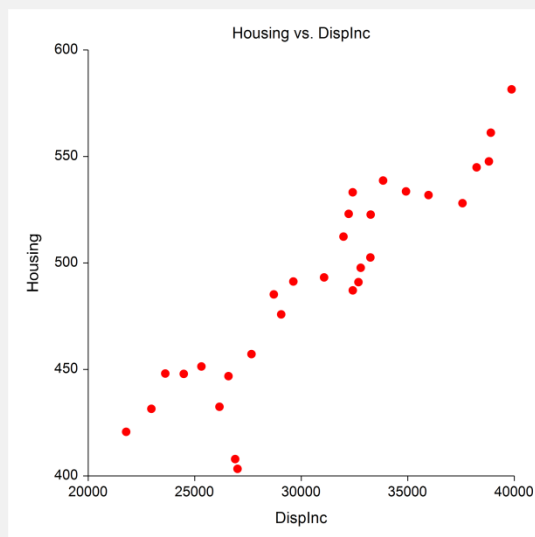
If the residuals are normally distributed, the data points of the normal probability plot will fall along a straight line through the origin with a slope of 1.0. Major deviations from this ideal picture reflect departures from normality. Stragglers at either end of the normal probability plot indicate outliers, curvature at both ends of the plot indicates long or short distributional tails, convex or concave curvature indicates a lack of symmetry, and gaps or plateaus or segmentation in the normal probability plot may require a closer examination of the data or model. Of course, use of this graphic tool with very small sample sizes is not recommended.

If the residuals are not normally distributed, then the t-tests on regression coefficients, the F-tests, and any interval estimates are not valid. This is a critical assumption to check.



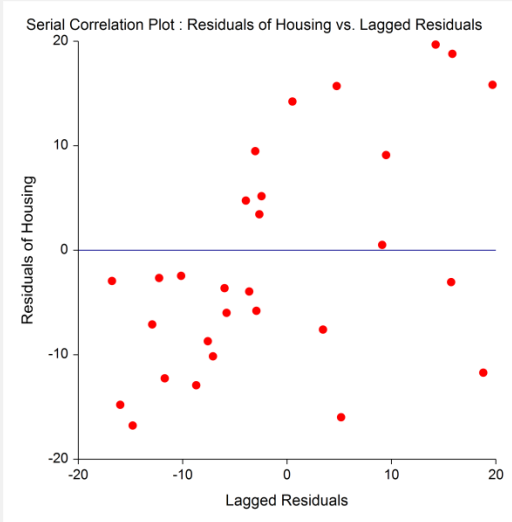
Y vs X's Plots

Actually, a regression analysis should always begin with a plot of Y versus each IV. These plots often show outliers, curvilinear relationships, and other anomalies.



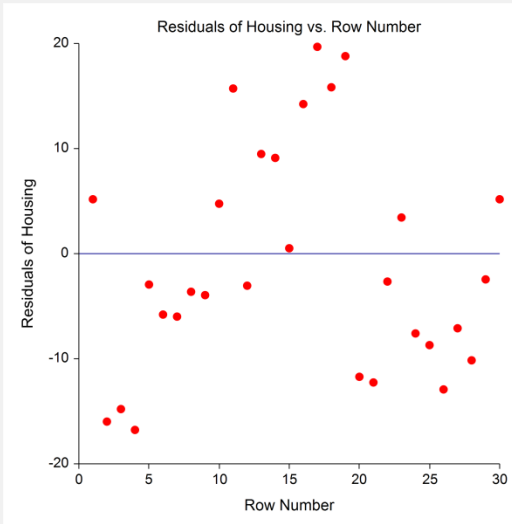
Serial Correlation Plot: Residuals vs Lagged Residuals

This is a scatter plot of the j^{th} residual versus the $j^{\text{th}}-1$ residual. The purpose of this plot is to check for first-order autocorrelation. Positive autocorrelation or serial correlation means that the residual in time period j tends to have the same sign as the residual in time period $(j-1)$. On the other hand, a strong negative autocorrelation means that the residual in time period j tends to have the opposite sign as the residual in time period $(j-1)$.



Sequence Plot: Residuals vs Row

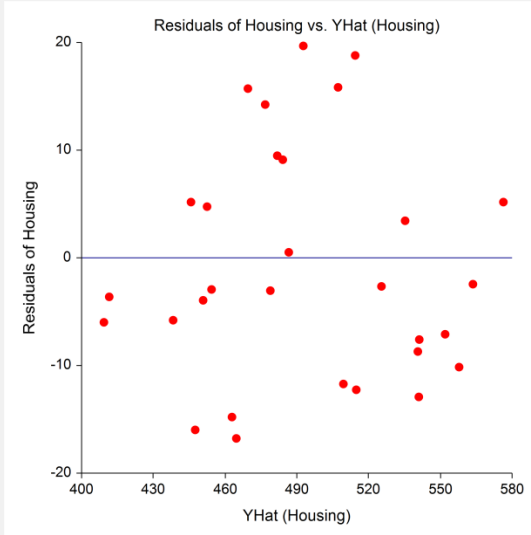
Sequence plots may be useful in finding variables that are not accounted for by the regression equation. They are especially useful if the data were taken over time.



Multiple Regression with Serial Correlation

Residuals vs Yhat Plot

This plot should always be examined. The preferred pattern to look for is a point cloud or a horizontal band. A wedge or bowtie pattern is an indicator of nonconstant variance, a violation of a critical regression assumption. The sloping or curved band signifies inadequate specification of the model. The sloping band with increasing or decreasing variability suggests nonconstant variance and inadequate specification of the model.



Residuals vs X's Plots

This is a scatter plot of the residuals versus each independent variable. Again, the preferred pattern is a rectangular shape or point cloud. Any other nonrandom pattern may require a redefining of the regression model.

