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 $\left(e_t = Y_t - \hat{Y}_t\right)$ 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.

$$Y_t' = Y_t - \hat{\rho} Y_{t-1}$$

$$X_{1t}' = X_{1t} - \hat{\rho} X_{1,t-1}$$

$$\vdots$$

$$X_{pt}' = X_{pt} - \hat{\rho} X_{p,t-1}$$

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.

$$b_0 = \frac{b_0'}{1 - \hat{\rho}}$$

$$b_1 = b_1'$$

$$b_2 = b_2'$$

$$\vdots$$

$$b_p = b_p'$$

The estimated standard errors of the regression coefficients are given by

$$s(b_0) = \frac{s(b_0')}{1-\hat{\rho}}$$

$$s(b_1) = s(b_1')$$

$$s(b_2) = s(b_2')$$

$$\vdots$$

$$s(b_p) = s(b_p')$$

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.
7. *Calculate Durbin-Watson test on transformed residuals*. As a final diagnostic check, the Durbin-Watson test may be run on the residuals $\left(e_t' = Y_t' - \hat{Y}_t'\right)$ 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

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

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*. Note that only two decimal places are displayed here, while on the database, more decimal places are stored.

Housing Dataset (Subset)

Year	Housing	Mort5Yr	DispInc	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		

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.

Setup

To run this example, complete the following steps:

1 Open the Housing example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Housing** and click **OK**.

2 Specify the Multiple Regression with Serial Correlation procedure options

- Find and open the **Multiple Regression with Serial Correlation** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab	
Y Dependent Variable(s)	**Housing**
X's Numeric Independent Variables	**DispInc, Mort5Yr**
Maximum Cochrane-Orcutt Iterations	**1**
Reports Tab	
All Available Reports	**Checked** (click the *Check All* button)
Plots Tab	
All Available Plots	**Checked** (click the *Check All* button)
Report Options (*in the Toolbar*)	
Variable Labels	**Column Names**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Run Summary

Run Summary

Item	Value	Rows	Value
Dependent Variable (Y)	Housing	Rows Processed	32
Number of Independent Variables (X)	2	Rows Used in Estimation	30
R^2	0.8941	Rows with X's Missing	0
Adjusted R^2	0.8860	Rows with Y Missing	2
Coefficient of Variation	0.0361		
Mean Square Error (MSE)	77.15598		
Square Root of MSE	8.783848		
Average |Percent Error|	1.801		
Autocorrelation (ρ)	0.5121		
Completion Status	Normal Completion		

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

Descriptive Statistics

Variable	Count	Mean	Standard Deviation	Minimum	Maximum
DispInc	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

Correlation Matrix

	DispInc	Mort5Yr	Housing
DispInc	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 Coefficient T-Tests and Estimated Equation

Regression Coefficient T-Tests

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	T-Test of H0: β(i) = 0 T-Statistic	P-Value	Reject H0 at α = 0.05?
Intercept	445.1365	35.46902	12.550	0.0000	Yes
DispInc	0.004443401	0.0008682022	5.118	0.0000	Yes
Mort5Yr	-8.537143	1.052221	-8.113	0.0000	Yes

Estimated Equation

445.136489079996+0.0044434007069797*DispInc-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.

Regression Coefficient Confidence Intervals

Regression Coefficient Confidence Intervals

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	95% Confidence Interval Limits for β(i) Lower	Upper	Standardized Coefficient
Intercept	445.1365	35.46902	372.2289	518.0441	0
DispInc	0.004443401	0.0008682022	0.002658786	0.006228016	0.4067824
Mort5Yr	-8.537143	1.052221	-10.70001	-6.374271	-0.6448714

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

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

Analysis of Variance Summary

Analysis of Variance Summary

Source	DF	R^2 Lost If Term(s) Removed	Sum of Squares	Mean Square	F-Ratio	P-Value
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 are 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 are significant if their absolute values are greater than 0.371391.

Durbin-Watson Test For Serial Correlation of Uncorrected Model

	Test of H0: ρ(1) = 0		
Test Type	Test Statistic Value	P-Value	Reject H0 at α = 0.2?
Positive Serial Correlation Test	0.9234	0.0002	Yes
Negative Serial Correlation Test	0.9234	0.9974	No

Durbin-Watson Test For Serial Correlation of Corrected Model

Test Type	Test Statistic Value to Test H0: ρ(1) = 0	P-Value	Reject H0 at α = 0.2?
Positive Serial Correlation Test	1.9221	0.3273	No
Negative Serial Correlation Test	1.9221	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 for Means

Predicted Values with Confidence Interval Limits for Means

	Housing		Standard Error of	95% Confidence Interval Limits for the Mean	
Row	Actual	Predicted	Predicted	Lower	Upper
1	420.7220	445.7375			
2	431.5217	447.5038	3.273156	440.7757	454.2318
3	448.0854	462.8741	4.140148	454.3640	471.3843
4	447.9233	464.6962	3.128303	458.2659	471.1265
5	451.4012	454.3400	2.580547	449.0356	459.6443
6	432.4736	438.2691	3.157941	431.7779	444.7603
7	403.3378	409.3280	5.545494	397.9290	420.7269
8	407.9224	411.5493	3.422108	404.5150	418.5835
9	446.8660	450.8127	3.589643	443.4341	458.1913
10	457.2156	452.4623	2.224679	447.8894	457.0352
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
30	581.5398	576.3640	3.536395	569.0948	583.6331
31		578.7437	3.901068	570.7249	586.7624
32		579.7596	4.192116	571.1426	588.3766

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

Predicted Values with Prediction Limits for Individuals

Predicted Values with Prediction Interval Limits for Individuals

	Housing			95% Prediction Interval Limits for an Individual	
Row	Actual	Predicted	Standard Error of Predicted	Lower	Upper
1	420.7220	445.7375			
2	431.5217	447.5038	9.373875	428.2355	466.7721
3	448.0854	462.8741	9.710654	442.9136	482.8347
4	447.9233	464.6962	9.324284	445.5299	483.8625
5	451.4012	454.3400	9.155065	435.5215	473.1584
6	432.4736	438.2691	9.334269	419.0822	457.4560
7	403.3378	409.3280	10.387900	387.9753	430.6806
8	407.9224	411.5493	9.426919	392.1720	430.9266
9	446.8660	450.8127	9.489021	431.3078	470.3177
10	457.2156	452.4623	9.061191	433.8368	471.0879
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
30	581.5398	576.3640	9.469006	556.9001	595.8278
31		578.7437	9.611156	558.9877	598.4997
32		579.7596	9.732924	559.7532	599.7659

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.

Residuals

Residuals

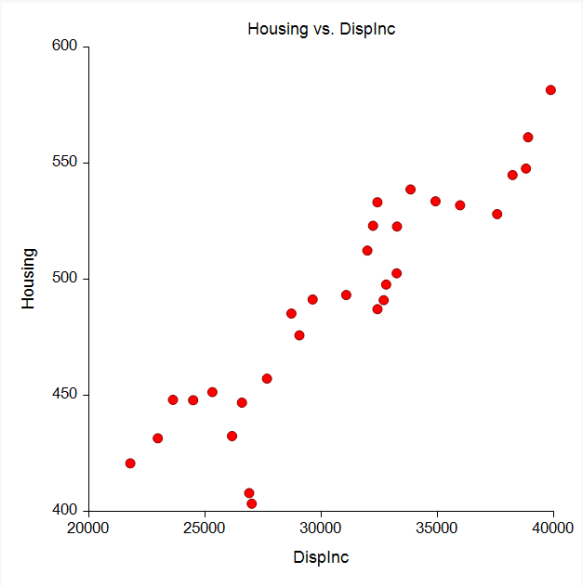
	Housing			Absolute Percent Error	
Row	Actual	Predicted	Residual		
1	420.7220	445.7375			
2	431.5217	447.5038	-15.9820400	3.704	
3	448.0854	462.8741	-14.7887300	3.300	
4	447.9233	464.6962	-16.7729500	3.745	
5	451.4012	454.3400	-2.9388150	0.651	
6	432.4736	438.2691	-5.7954750	1.340	
7	403.3378	409.3280	-5.9901690	1.485	
8	407.9224	411.5493	-3.6269160	0.889	
9	446.8660	450.8127	-3.9467440	0.883	
10	457.2156	452.4623	4.7533190	1.040	
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
30	581.5398	576.3640	5.1758600	0.890	
31		578.7437			
32		579.7596			

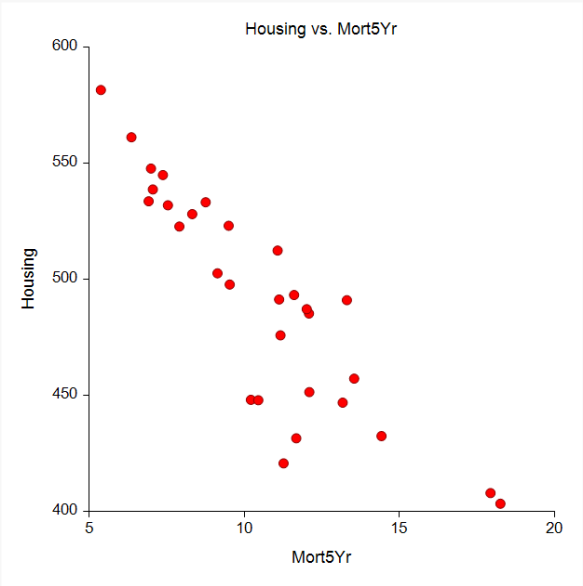
This section reports on the sample residuals, or e_i's.

Y vs X 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.

Y vs X Plots



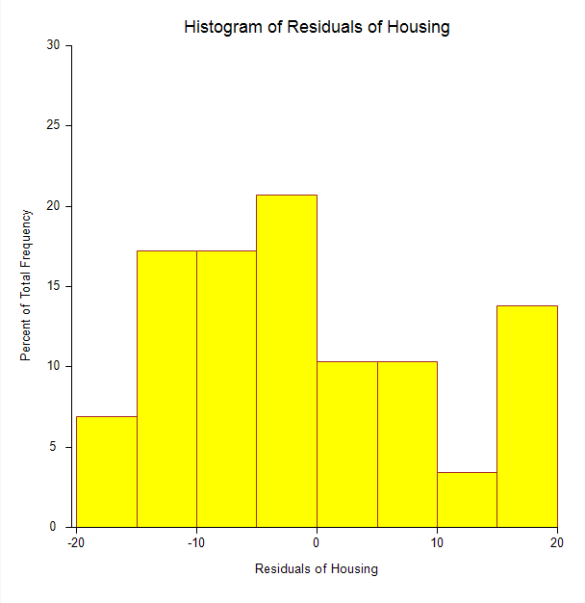


Residual Distribution Plots

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.

Residual Distribution Plots

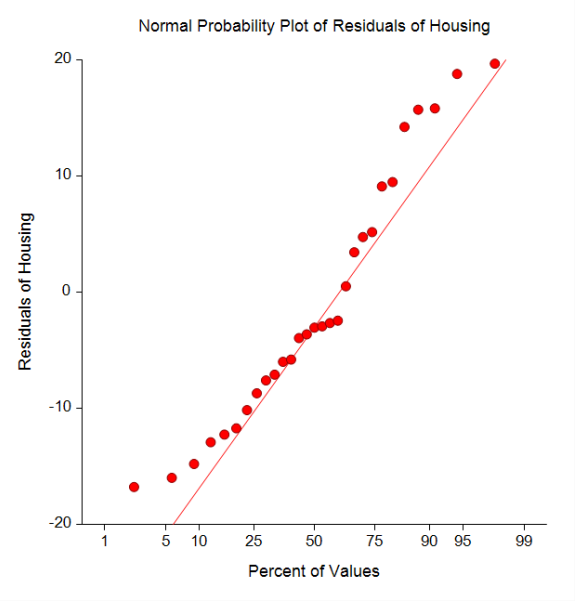


Normal Probability Plot of Residuals

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.

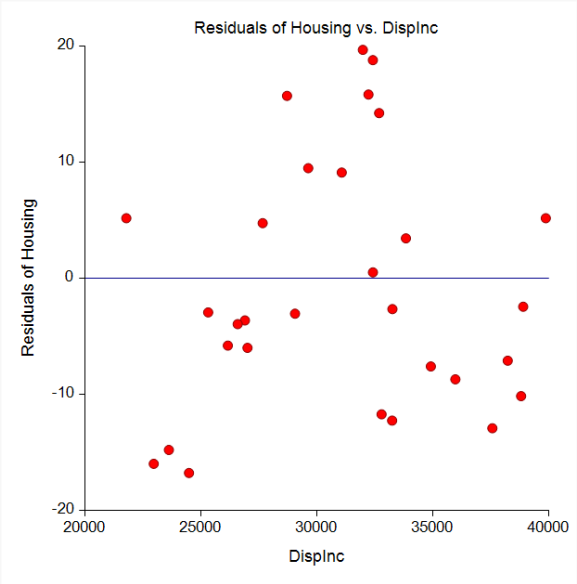
Residual Distribution Plots

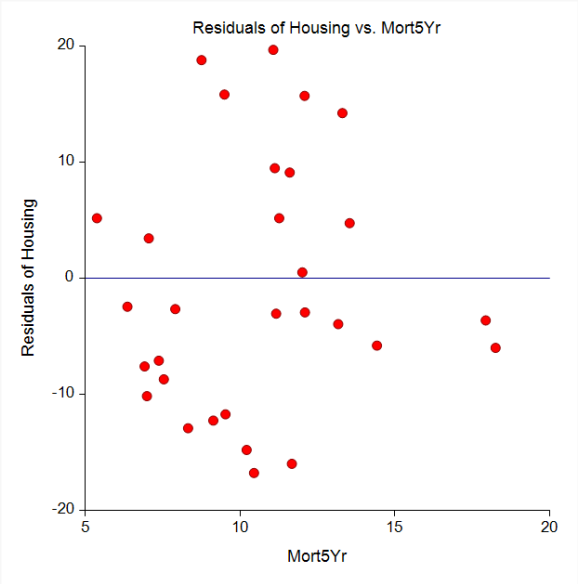


Residuals vs X Plots

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.

Residuals vs X Plots

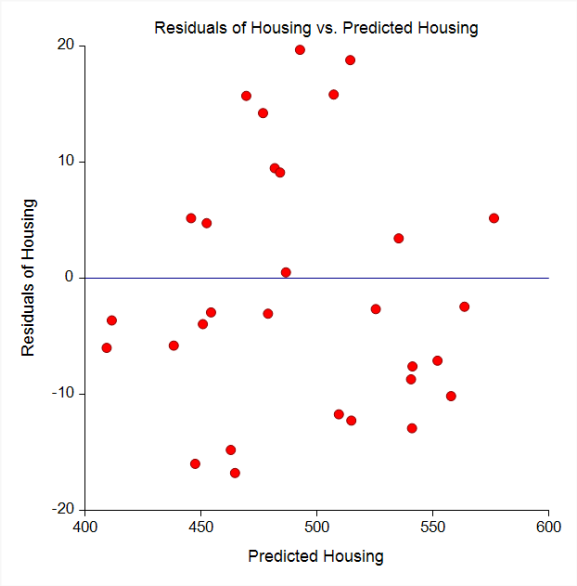




Residuals vs Yhat (Predicted Y) Plot

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.

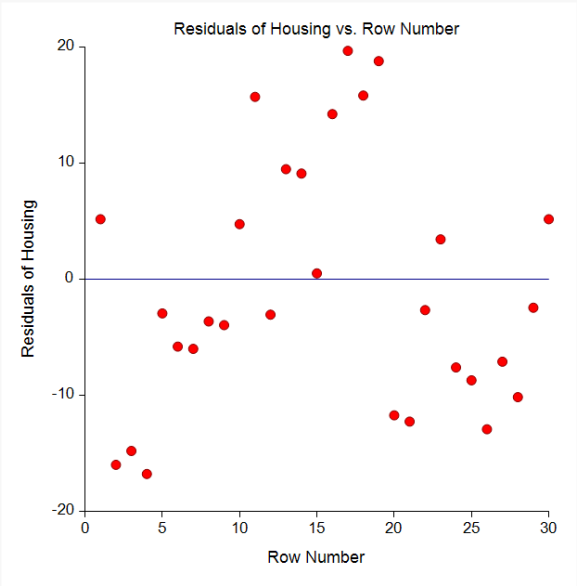
Residuals vs Yhat (Predicted Y) Plot



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.

Sequence Plot: Residuals vs Row Number



Serial Correlation Plot: Residuals vs Lagged Residuals

This is a scatter plot of the j^{th} residual versus the j^{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).

Serial Correlation Plot: Residuals vs Lagged Residuals