Time Series Analytics Notes
Quantitative Validation of the L,H,I,N Assumptions of Regression

In this section of the notes, I present quantitative tests of the regression specifications for time series.[1] You should recall that a regression model is valid if the **L**, **H**, **I**, **N** specifications are true. *Qualitative* tests of LHIN are based on graphical displays of the residuals. Judging those graphs involves subjective interpretation of the displays. The *quantitative* tests remove much of the subjectivity inherent in judging the graphical displays, although they are based on those graphical displays. The quantitative methods simply quantify the intuition behind the graphical methods. The quantitative tests are not intended to replace visual inspection of residual plots and other qualitative procedures for validating a time series model. Rather, quantitative tests should be used as companions to qualitative tests. They should not replace common sense. Common sense should reserve the right to over-rule a quantitative result when necessary.

My plan for this Topic Note proceeds step by step:

- First, I will lay out a precise statement of the LHIN specifications.
- Second, I will review the qualitative tests of LHIN through an example.
- Third, I will present the quantitative tests of LHIN and show how they are inspired by the corresponding qualitative tests.

The LHIN Regression Specifications

First, here are the specifications for a general regression model. The response variable is Y_t. The predictor variables may include other time series, and/or their lags of various orders, as well as lags of Y_t of various orders, and powers and other functions of any of these. To avoid a clumsy notation, I will call any of these predictors "X_t's" – other series, lags, powers, etc. – the whole lot of them. Suppose that there are *p* of these predictor variables. The following box states the linear multiple regression time series model and its specifications.

Suppose that $Y_t, X_{t1}, X_{t2}, ..., X_{tp}$ are time series, and $\alpha, \beta_1, \beta_2, ..., \beta_p$ are unknown constants (parameters), such that for all $t = 1, \ldots, n$

L : $Y_t = \alpha + \beta_1 X_{t1} + \beta_2 x_{t2} + \beta_3 x_{t3} + \cdots + \beta_p x_{tp} + \varepsilon_t$ and $E(\varepsilon_t) = 0$

H : $Var(\varepsilon_t) = \sigma^2$ (the same for all *t*)

I : $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ are independent

N : $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ are normally distributed.

If all of these specifications hold, then the multiple linear time series regression model is said to be valid.

[1] Except for the "I" specification, these tests also apply to cross-sectional regressions. There are no general-purpose quantitative tests for "I" outside time series – instead, one must rely on scrupulous adherence to sound principles of sample design.

Qualitative tests – an example: Austin apartment rents

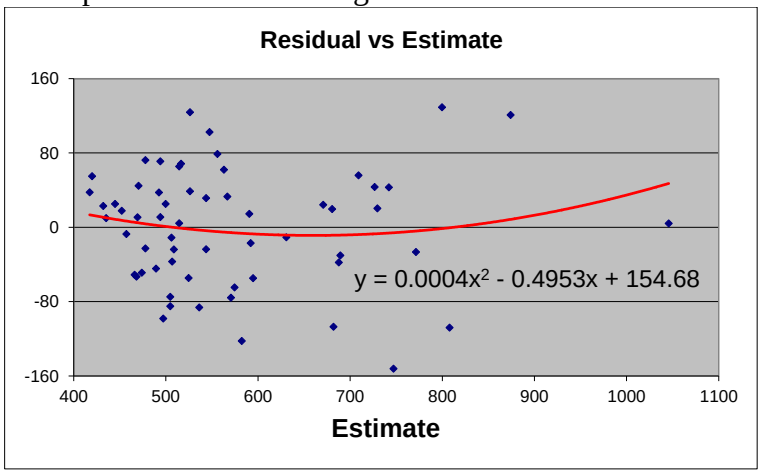
Second, I will review the qualitative tests of LHIN through an example.

Recall the multiple regression of 60 Austin apartment rents on both their area and number of bathrooms. The estimated regression is *Rent* = 143.6693 + 0.3875 *Area* + 89.9290 *Bathrooms*. Let us look at validating the LHIN assumptions by qualitative/graphical means.

L. The graphical test of L is to look at a **plot of residuals vs estimates** and judge whether the plot is roughly level across all estimates from left to right. Figure 1 shows that plot for the Austin apartment multiple regression of rent on area and bathrooms. I recommend adding a curve to the plot to see if there is any appreciable bend. In Figure 1, I have added a quadratic trendline. If the quadratic trendline curves substantially, then the plot is not level and L is violated. If the quadratic is more like a flat line, then the plot is level and L is satisfied.

To my eyes, the quadratic does not appear to curve substantially. So I would pass L. However, I admit that there is a bit of a curve, and subjective opinion may differ as to whether the curve is "substantial".

Figure 1. Graphical test of L for regression of rent on area and bathrooms



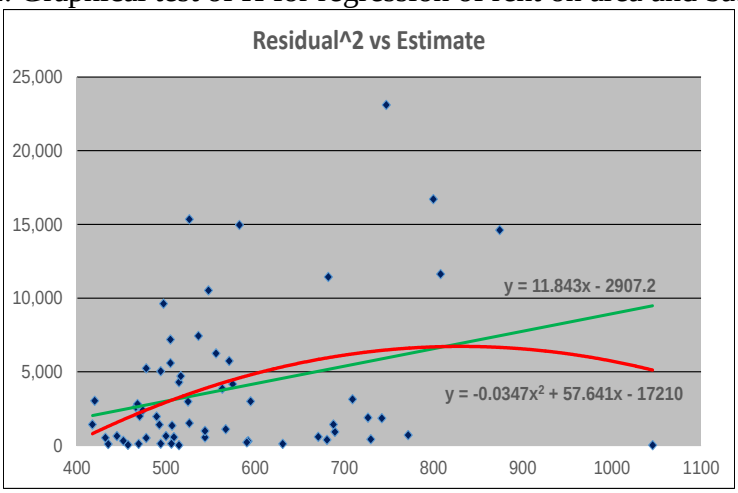
H. The graphical test of homoscedasticity looks at the same plot (Figure 1) of residuals vs estimates that was used for the graphical test of L. However, for H you try to judge whether the average magnitudes of the residuals – the distances up and down from the horizontal line at zero – are roughly the same across all estimates from left to right.

It is often difficult to judge that using the same graph as for L because the magnitudes are on both sides of the central zero line. It appears to my eyes that the magnitudes – the vertical distances of the points from the zero line – may grow somewhat, on average, from left to right. However, I can easily imagine that others may differ with my opinion. The judgment is difficult because the plot shows positive residuals on the opposite side of the center zero line from negative residuals. Consequently, one's eyes must continually flick back and forth, up and down, across the center line at zero. The judgment could be made easier if the graph could show the magnitudes of the residuals vs estimates – either the absolute values of the residuals or their

squares – all on one side. Figure 2 shows the **plot of squared residuals vs estimates** for the multiple regression.

I have added red (linear) and green (quadratic) trendlines to show the trend of squared residuals as estimates increase from left to right. To my eyes, both trendlines suggest that the magnitude of residuals increases substantially as estimates increase. But some may question whether the apparent uptrend is real in the population of all Austin apartments, or just a fluke of these 60 apartments.

Figure 2. Graphical test of H for regression of rent on area and bathrooms



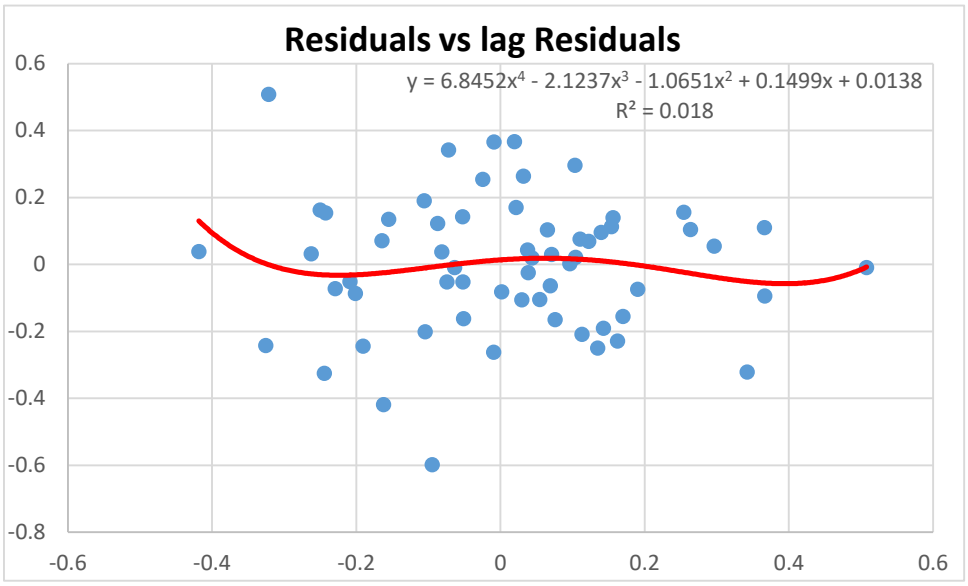
I. For cross-sectional data like the Austin apartment dataset, one should look to the method of sample selection to validate I: If the data were drawn as a random sample, then the I condition automatically holds.

For the Austin apartment data, the method of selection is given that they are a Random Sample. Therefore, the I condition is satisfied automatically for the residuals in the regression of rent on area and bathrooms.

For time series data, graphical validation of I is often difficult to see in the plot of residuals vs estimates. However, if I holds, then the current residual ε_t should be independent of any of its predecessors $\varepsilon_{t-1}, \varepsilon_{t-2}, \varepsilon_{t-3}, \ldots$ Therefore, a simple graphical test of I for time series can be provided by plotting ε_t against ε_{t-1} (and other lags of the residuals, if necessary).

Since the Austin apartment dataset is cross-sectional, it cannot provide an example of the I test for time series data. As an example of the graphical test of I, here is a residual plot for a different regression. The daily closing price of Dell computer stock was regressed on the lag of the daily closing price of Dell. The following plot shows the residuals (vertical) from that regression plotted against the lag residuals (horizontal) from that regression. The I specification requires that there be no significant relationship between successive residual values (between the residuals and their lags). To help make that judgment, I have added a trendline in red that represents the residuals as a 4th power (quartic) polynomial function of lag residuals. That trendline should be essentially flat at the horizontal line at zero. To my eyes, it is; others may differ. Other polynomials could also be used – e.g., quadratic or cubic polynomials – or more exotic functions. Whatever the functional form, it should be essentially flat.

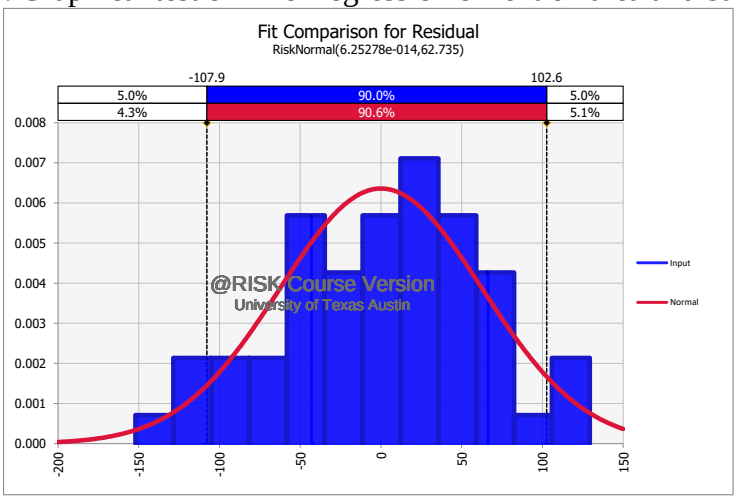
Figure 3. Graphical test of I for residuals of autoregression of Dell stock price on the lag of those residuals



Negative autocorrelation would be indicated by a negative slope in the plot. Positive autocorrelation would be indicated by a positive slope in the plot.

N. The graphical test of normality is to look at a histogram of the residuals and judge whether they appear to be normal. For the regression of rent on area and bathrooms for the Austin apartment data set, Figure 4 shows a histogram of the residuals (produced by @Risk, one of the Decision Tools Suite) with a superimposed normal histogram.

Figure 4. Graphical test of N for regression of rent on area and bathrooms



To my eyes, there does not appear to be substantial separation of the data histogram from the smooth normal histogram. I am willing to pass N. Other people may differ.

Third, I will now present quantitative tests of L,H,I,N.

Quantitative Validation of L

Figure 1 (plotting residuals vs estimates for the regression of rent on area and bathrooms) illustrates the intuition for the quantitative test of L. The L specification is actually the specification that there is no bias – no underestimation and no overestimation. If L holds then the red trendline in Figure 1 should be flat. Any significant curve in the red trendline would suggest that the mean level of the residuals in some places rises above or sinks below the horizontal line at zero.

- If the mean residual is positive at some place, then the Actuals tend to exceed the Estimates at that place. This is because for any datum, Actual = Estimate + Residual. So positive Residuals imply Actual > Estimate. This means that the model underestimates the actual data at that place. ☹
- Similarly, if the mean residual is negative at some place, then we tend to have Actual < Estimate at that place. This is overestimation. ☹
- However, if the red trendline is flat at zero everywhere, then the mean residual averages zero everywhere. From Actual = Estimate + Residual, this means that Actual = Estimate, on average, so there is neither underestimation nor overestimation. It is just right! ☺

We do not expect the red trendline to be exactly flat. How can we tell if it is *sufficiently* flat? The equation of the red trendline is $y = 0.0004\ x^2 - 0.4953\ x + 154.68$. "$y$" is the Residual and "$x$" is the Estimate. If both coefficients 0.0004 and -0.4953 are close to zero, then the equation becomes $y = constant$, which is flat. So testing both coefficients being close to zero tests flatness, which tests L. How to test if both coefficients are close to zero? The T-test for each coefficient can do this. To get the T-Test, run a regression of the Residuals (as Y) on two predictors, the Estimates (as x) and the Estimates^2 (as x^2). This is a *second* and *separate* regression from the original regression of Rent on Area and Bathrooms. Run the original regression and get its Residuals and Estimates. Then create Estimates^2. Then run the second regression of Residuals on Estimates and Estimates^2. The output for that second regression is shown in Figure 5:

Figure 5. Regress residuals of rent = area + baths on estimates and estimates^2

Multiple Regression for Residual Summary	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
	0.1359	0.0185	0.0000	63.23382094	0	0
ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F	p-Value	
Explained	2	4287.331633	2143.665817	0.536115338	0.5879	
Unexplained	57	227915.4183	3998.51611			
Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Confidence Interval 95% Upper
Constant	154.6828187	154.2247512	1.00297013	0.3201	-154.1469185	463.5125559
Estimate^2	0.000375333	0.000362471	1.03548572	0.3048	-0.000350502	0.001101168
Estimate	-0.495286643	0.482773086	-1.025920162	0.3093	-1.462023072	0.471449787

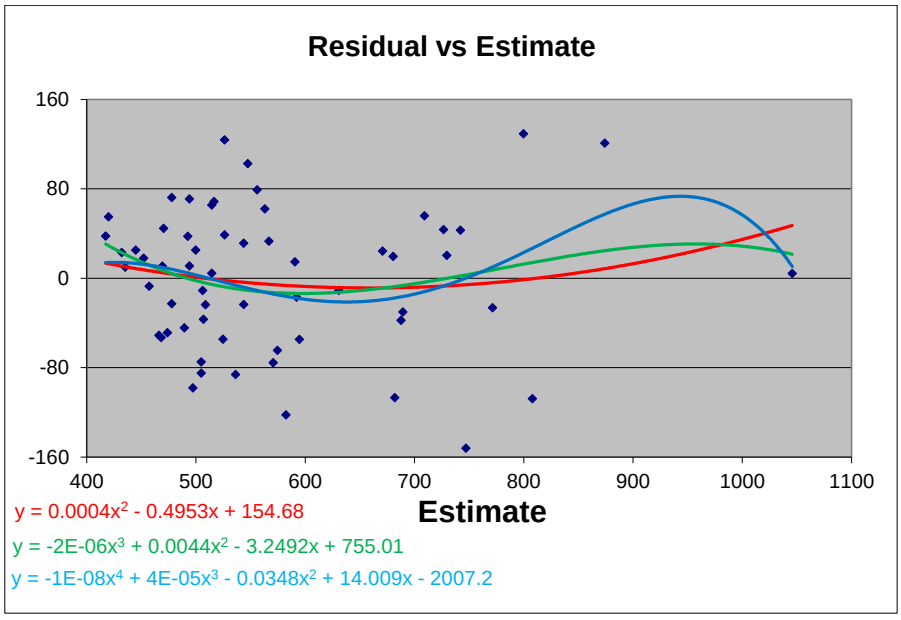
The t-Values (-1.0259 and 1.03548) and p-Values (0.3093 and 0.3048 – in yellow) for the Estimate and Estimate^2, respectively, show that the coefficients of Estimate and Estimate^2 can be considered close to zero. Therefore, the red trendline in Figure 1 is sufficiently flat to pass the "L" specification.

There is another test shown in Figure 5 that also tests whether the coefficients of Estimate and Estimate^2 are zero. That is the F-test with F-value = 0.5361 and p-Value = 0.5879 (in green). The high p-Value indicates that the two coefficients can be considered close to zero. The F-test is a little better than the separate T-tests.[2]

The reasonable thought may have occurred to you that fitting a quadratic equation to the plot of residuals vs estimates is not the only way to search for a violation of L. You could also fit a cubic equation, or a quartic , or … . A violation would be indicated by one or more of the polynomial coefficients differing significantly from zero. Conformity with L would be indicated by all polynomial coefficients being close to zero. Nonpolynomial functions could also be tried: sinusoidal, logarithmic, exponential, etc. The idea is to search for patterns of non-flatness in the plot. However, let's not go overboard! Usually, a non-flat residual pattern will match enough of some polynomial to make it unnecessary to search more exotic functions. Indeed – usually – the first 2-4 polynomial powers are sufficient.

The preceding intuition lies behind a formal, quantitative test of L called Ramsey's **RESET** (**RE**gression **S**pecification **E**rror **T**est) procedure, available in SAS PROC AUTOREG (but not in PROC REG). RESET fits three equations to the plot of residuals vs estimates: quadratic, cubic, and quartic, and tests the significance of each of them with the F-test. For Figure 1, the quadratic procedure is essentially as discussed above. Figure 6 replicates Figure 1 but adds the cubic and quartic equations that RESET also tests.

Figure 6. Graphical basis for Ramsey's RESET test of L for regression of rent on area and bathrooms



[2] The reason for the superiority of the F-test is technical. Usually the F-test and the T-tests agree. However, there are situations in which the F-test is significant (low p-Value) but the T-tests are not (high p-Values). This can occur if the Estimate and Estimate^2 are too highly correlated with each other. Then they share too much explanatory power, leaving little explanatory power for either separately, resulting in high separate p-Values. This is called *multicollinearity*. The F-test avoids this situation by measuring the combined effect of Estimate and Estimate^2.

RESET tests whether the red, green, and blue lines in Figure 6 are flat.

The SAS code is

```
proc autoreg data=apts;
  model rent = area bathrooms / reset;
run;
```

The SAS output is

SSE	232202.75	DFE	57
MSE	4074	Root MSE	63.82580
SBC	678.216955	AIC	671.933922
MAE	50.8951313	AICC	672.362493
MAPE	9.17915706	HQC	674.391562
Durbin-Watson	1.9596	Total R-Square	0.8007

Ramsey's RESET Test

Power	RESET	Pr > F
2	1.1159	0.2953
3	1.1361	0.3285
4	0.9340	0.4307

Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	143.6693	29.5135	4.87	<.0001
Area	1	0.3875	0.0498	7.78	<.0001
Bathrooms	1	89.9290	27.7507	3.24	0.0020

Ramsey's RESET output is highlighted in yellow. The large p-Values for each of the three powers indicate that L is accepted under all three forms of the test.

In summary, Ramsey's RESET procedure for validating L:

1. Run your regression of interest and get the residuals and the estimates ($\hat{Y}_t$) to use in step 2:
2. Run a second regression: $residual_t = \alpha + \beta_2\hat{Y}_t^2 + \beta_3\hat{Y}_t^3 + \beta_4\hat{Y}_t^4 + error_t$. (This is the quartic model.)
3. Use the F test of the second regression to test $H_0: \beta_2 = 0 \; and \; \beta_3 = 0 \; and \; \beta_4 = 0$.
4. Reject L if the F test rejects $H_0: \beta_2 = 0 \; and \; \beta_3 = 0 \; and \; \beta_4 = 0$.
5. Actually, as implemented in SAS, Ramsey's RESET procedure is done three times, for three successively more flexible models:[3]
   - (Quadratic) $residual_t = \alpha + \beta_2\hat{Y}_t^2 + error_t$
   - (Cubic) $residual_t = \alpha + \beta_2\hat{Y}_t^2 + \beta_3\hat{Y}_t^3 + error_t$
   - (Quartic) $residual_t = \alpha + \beta_2\hat{Y}_t^2 + \beta_3\hat{Y}_t^3 + \beta_4\hat{Y}_t^4 + error_t$

[3] There is no linear form ($residual_t = \alpha + \beta_2\hat{Y}_t + error_t$) and the linear term is not used in any of the quadratic, cubic, or quartic forms of RESET – although I used it in Figures 1, 5 and 6 and in my explanation. The reason that there is no linear term is that it is unnecessary – the linear coefficient is always zero because the correlation between residuals and estimates is always zero by construction. It does not hurt to add the linear term as I did, but it is more efficient not to.

6. RESET may be applied both to time series data and to cross-sectional data.
7. In SAS, RESET is invoked as an option in the MODEL statement of PROC AUTOREG for the *original* regression – so you do not actually need to run two separate regressions.

Quantitative Validation of H

If H is true, then the regression residuals can be expected to deviate from zero (their mean) by a magnitude equal to their standard deviation, on average. So the magnitude of the residual, on average, can be expected to equal the error standard deviation σ_e that is posited by the H specification. So the magnitude of the squared residuals can be expected to approximate the common error variance σ_e^2. So a plot of the squared residuals against the X-predictors or against the estimates ($\hat{Y}_t$) should be constant (flat) if H holds.

Figure 2 (above) shows this idea in action. It is a plot of the squared residuals from the regression of apartment rents on their area and number of bathrooms. Instead of being flat, the green linear line extends from about 2,000 when the estimate is about $400 rent to nearly 10,000 when the estimate is about $1050 rent. Neither does the red quadratic curve appear to be flat.

More generally, one could look for other patterns in the plot of squared residuals (Figure 2) – to see if the plot of squared residuals deviates from level. One could carry out a procedure that parallels Ramsey's, but with the squared residuals instead of the residuals. Nothing wrong with that idea! But the most famous and most commonly used test of H uses a somewhat different procedure. The idea is the same, but the details are different. Instead of using as predictors the second, third and fourth powers of the fits, **White's specification test for homoscedasticity** uses first, second, and all cross-product terms of all predictors. Such a set of predictors yields what is termed a "response surface" regression.[4] If there are *p* predictors, then White's test has $\binom{p+2}{2}$ coefficients, including the intercept.[5] For example, if the original specification is $Y_t = \alpha + \beta_1 X_{t1} + \beta_2 X_{t2} + \varepsilon_t$, then White's test fits the model

$$residual_t^2 = a + b_1 X_{t1} + b_2 X_{t1}^2 + b_3 X_{2t} + b_4 X_{2t}^2 + a_5 X_{t1} X_{t2} + \phi_t .$$

White also differs from Ramsey in using a chi-square distribution instead of an F distribution. White's test rejects **H** if $nR^2 > \chi_m^2(\alpha)$, where *n* = sample size, R^2 = R-square for the auxiliary (response surface) regression, *m* = number of predictor coefficients in the auxiliary regression, and $\chi_m^2(\alpha)$ is the 100(1 - α) percentile of the chi-square distribution with *m* degrees of freedom. The intuition is that **H** cannot hold unless all of the predictor coefficients in the auxiliary regression are zero (not counting the intercept). The auxiliary predictor coefficients are all zero if

[4] So-called because such models are often used for three-dimensional displays of a *Y* that varies across a two-dimensional surface grid, like oil deposits across the surface of the earth. The predictors are the geographic coordinates X_1 and X_2.

[5] I leave it as an interesting counting exercise to verify this.

and only if the auxiliary R-square = 0. So a sample R-square much larger than zero (taking the sample size into account) signals violation of **H**. The chi-square distribution, however, is approximate and holds if *n* is sufficiently large.

White's test is available in SAS as the SPEC option in PROC REG (but not in PROC AUTOREG). SPEC is invoked in the *original* regression – so you do not actually need to run two separate regressions.

The following code runs SPEC for the apartment rent example. Output for White's test of H is shown in yellow highlighting. The relatively high p-value (0.1365) indicates that H would not be rejected.[6]

The SAS code is

```
proc reg data=apts;
  model rent = area bathrooms / spec;
RUN;
```

The SAS output is

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	932883	466441	114.50	<.0001
Error	57	232203	4073.73245		
Corrected Total	59	1165086			

Root MSE	63.82580	R-Square	0.8007
Dependent Mean	572.26667	Adj R-Sq	0.7937
Coeff Var	11.15316		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	143.66927	29.51345	4.87	<.0001
Area	1	0.38746	0.04982	7.78	<.0001
Bathrooms	1	89.92902	27.75071	3.24	0.0020

Test of First and Second Moment Specification

DF	Chi-Square	Pr > ChiSq
5	8.38	0.1365

Quantitative Validation of I

If the I specification is true, then the regression errors are independent and hence have zero autocorrelation at all lags. That is, $corr(e_t, e_{t-k}) = 0$ for all $k = 1, 2, \ldots$. $Corr(e_t, e_{t-k})$ is called the **lag *k* autocorrelation** of the residuals.[7] If $corr(e_t, e_{t-k}) \neq 0$ for any *k*, then specification I is violated. So a test of I can be based on the autocorrelations of the regression residuals: Calculate a relevant number of residual autocorrelation lags and test whether they are all zero.

[6] White's test has only 5 degrees of freedom instead of 6 because bathrooms and bathrooms^2 are collinear, since there are only 1 and 2-bathroom apartments in this dataset.

[7] Another term that is synonymous with *autocorrelation* is *serial correlation*.

Sidebar. Two cautions:
1. Testing autocorrelations is not a definitive test for independence of the residuals. That is because zero autocorrelation is necessary for independence, but not sufficient. However, it is a reasonable *practical* equivalence in most cases.
2. How many lags to test? Generally, if the past has an effect on the present, the effect is manifest within a few time periods. The larger the lag, the weaker the autocorrelation – *generally*. However, periodic data provide many exceptions. For example, with quarterly data that are seasonal, like sales of a toy store, the strongest autocorrelation is often at lag 4 – which indicates strong dependence on what happened a year ago. Generally, with periodic data, it is safe to test a number of lags equal to twice the length of the period. Example: 8 lags for quarterly sales of a toy store; 22 for annual sunspot cycles; 14 for daily restaurant sales; etc. Often just half the number (4, 11, 7, etc.) are tested.

The caution over this point arises because the more tests you perform, the more likely you are to get a *false positive*. A false positive occurs when the null hypothesis is falsely rejected. That is, the autocorrelation of lag *k* is really zero, but the test rejects it anyway. This happens because statistical tests are uncertain. They sometimes make mistakes. You can control the probability of false positives, but not eliminate them. If you test whether each of lags 1-20 of the residual autocorrelation is zero and use a cut-off of 0.05 for the p-value of each test, then you should expect at least one of the 20 tests to reject, even if all 20 lags truly have zero autocorrelation. When you choose a cut-off for the p-value, you have chosen the false positive rate.[8]

The more tests you run, the more likely you are to have a false positive among the results. Statisticians call this issue the *multiple inference* problem. Statisticians have long been aware of it, but opinion differs over how to deal with it. There are many proposals and several books. Multiple inference is a big problem in Big Data, especially in Big-*p* versions. For example, an online site might track scores or hundreds of behavioral features of visitors in an attempt to predict online sales. If all features are tested separately, not as a group, then chances are high that there will be some false positives among the features most correlated with purchases. For example, if there are 100 features and each feature is tested separately at the 0.05 significance level, then it is likely that about 5 of the 100 features will test as significant – *even if none of them really is!* The practical consequence will be that the analyst thinks she has discovered a way to predict sales, but it is really just a false alarm.
End sidebar.

Three more-or-less equivalent methods for testing H_0: $corr(e_t, e_{t-k}) = 0$ for a given lag *k*:

- First run the regression *as though it is valid*. Get the residuals u_t from that regression. Then regress the residuals upon their k^{th} lag: That is, run the regression model $u_t = \beta u_{t-k} + \varepsilon_t$.[9] Then use the standard T-test for H_0: $\beta = 0$.

[8] You can reduce the probability of a false positive by choosing a lower cut-off for the p-value. However, lowering the probability of a false positive *raises* the probability of a false negative. A false negative occurs when you accept the null hypothesis falsely: the lagged autocorrelation really is not zero, but your test says it is zero.
[9] This is a *no-intercept* regression model. In SAS, use the option NOINT in the MODEL statement of PROC REG or PROC AUTOREG. It probably would not hurt to use the intercept model, but the intercept will likely be around zero, since the mean residual is zero.

- Second, calculate the residual autocorrelation $corr(e_t, e_{t-k})$. Reject H_0: $corr(e_t, e_{t-k}) = 0$ if the calculated $|corr(e_t, e_{t-k})| > 2\frac{1}{\sqrt{n-k-1}}$.[10,11]
- Third, calculate the Durbin-Watson statistic

$$d_k = \frac{\sum_{t=k+2}^{n}(u_t - u_{t-k-1})^2}{\sum_{t=1}^{n}(u_t)^2}$$

Reject H_0: $\rho = 0$ if d is either too big or too small. Values of d_k close to zero suggest high positive autocorrelation; values of d_k close to 4 suggest high negative autocorrelation; values of d_k around 2 suggest zero autocorrelation. Software will print the p-value, or special tables can be used.

Example. Since the Austin apartment dataset is cross-sectional, it cannot provide an example of the I test for time series data. As an example of the quantitative test of I, I regressed the daily closing price of Dell computer stock on its lag.

First. The following code and output illustrate the method for testing I of the first bullet above (autoregression):

The SAS code is

```
data Dell;
  set Dell;
  lagDell = lag(Dell);
run;
proc reg data=Dell;
  model Dell = lagDell;
  output out=Dell r=resDell;
run;
data Dell;
  set Dell;
  lagResDell = lag(resDell);
run;
proc reg data=Dell;
  model resDell = lagResDell / noint;
run;
```

[10] This formula is a refinement upon the simple $2/\sqrt{n}$ rule. The formula takes into account the loss of data from lagging.

[11] An even better test uses Fisher's *z* transformation of the (auto)correlation *r*: The hyperbolic arctangent of *r* is $z = \frac{1}{2}\log_e\frac{1+r}{1-r}$, which has an approximate normal distribution with mean of $\frac{1}{2}\log\frac{1+\rho}{1-\rho}$ and variance of $1/(n-3)$, where ρ is the true (auto)correlation. Reject zero autocorrelation if the calculated $|z| > 2/\sqrt{n-3}$.

The relevant part of the output of the second PROC REG above is

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
lagResDell		1	-0.01978	0.13830	-0.14	0.8867

The output highlighted in yellow suggests that the coefficient of lagResDell may be regarded as zero. Notice the NOINT option in the code.

Second. The following code and output illustrates the method for testing I of the second bullet above (autocorrelation), and continues the code of the first bullet, in sequence:

The SAS code is

```
proc corr data=Dell;
  var resDell lagResDell;
run;
```

The relevant part of the SAS output is

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	resDell	lagResDell
resDell	1.00000	-0.01806
Residual		0.8892
	63	62
lagResDell	-0.01806	1.00000
	0.8892	
	62	62

The approximate correlation between ResDell and lagResDell is -0.01806, which has a (two-tailed) p-value of 0.8892. This indicates that the autocorrelation can be regarded as close to zero.

Third. The following code and output illustrates the method for testing I of the third bullet above, and continues the code of the second bullet, in sequence:

The SAS code is

```
proc autoreg data=Dell;
  model Dell = lagDell / dw=4 dwprob;
run;
```

The SAS output is

Ordinary Least Squares Estimates

SSE	2.47696874	DFE	61
MSE	0.04061	Root MSE	0.20151
SBC	-16.801725	AIC	-21.087994
MAE	0.15422815	AICC	-20.887994
MAPE	1.27342909	HQC	-19.402183
		Total R-Square	0.4890

Durbin-Watson Statistics

Order	DW	Pr < DW	Pr > DW
1	1.8875	0.2920	0.7080
2	1.6496	0.0881	0.9119
3	1.9419	0.4929	0.5071
4	2.0675	0.7367	0.2633

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

Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	3.5295	1.1299	3.12	0.0027
lagDell	1	0.7090	0.0928	7.64	<.0001

The values and p-Values of the Durbin-Watson statistics for 4 lags are highlighted in yellow in the output. The p-Values for testing positive autocorrelation (`Pr < DW`) are all > 0.05[12]; the p-Values for testing negative autocorrelation (`Pr > DW`) also are all > 0.05. We conclude that the autocorrelation is satisfactorily close to zero. Therefore, I is satisfied.

* Note: The option `dw=4` requests the value of the Durbin-Watson statistic for lags 1-4. This option is not available in PROC REG. However, the option `dw` is available. This produces the value of the Durbin-Watson for one lag. In PROC REG, only the first lag of the Durbin-Watson statistic can be requested. In PROC AUTOREG, the Durbin-Watson statistic is automatically computed and printed out (but not its p-Value), whether you ask for it or not, even if the data are not time series! DW does not make sense for cross-sectional data, so be careful!

* Note: The option `dwprob` requests p-Values for all computed lags of the Durbin-Watson statistic. Without this option, the value of the statistic would be printed but not its statistical significance. This option is also available in PROC REG.

* Note: DW does not test autocorrelation in Dell's price. DW tests autocorrelation in the residuals of the regression of Dell on its lag.

Quantitative Validation of N

The key idea for validating N (normality of the residuals) is to look at the distribution of the residuals to see if it appears normal. We did that earlier in a qualitative manner in Figure 4 for the residuals from the regression of apartment rent on the area and number of bathrooms.

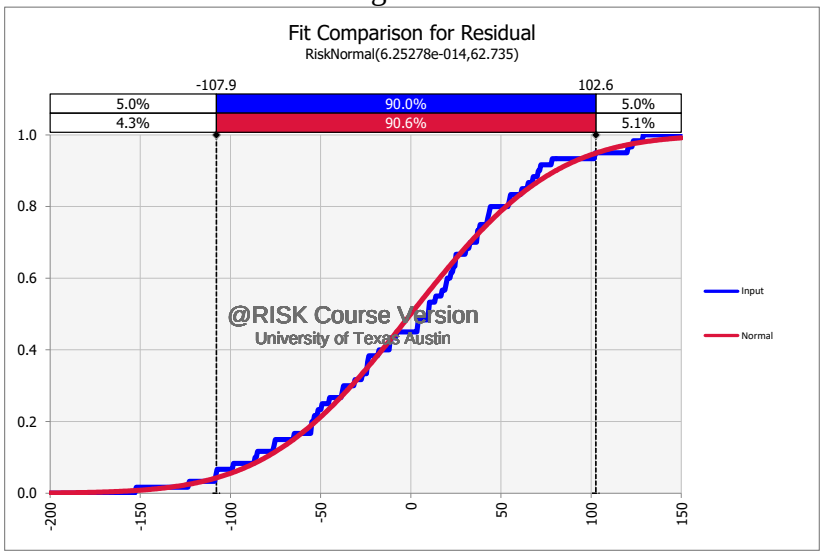
The key to making a quantitative test for N is to figure out how to quantify the idea that the residuals may not look much like the normal curve. To do that, we need to figure out how far apart the residual histogram and the normal curve are. That turns out to be easier to do for the cumulative curves than for the histograms.

[12] In this context, *testing positive autocorrelation* means that positive autocorrelation is the alternative hypothesis, so rejecting the null hypothesis (small p-value) means that the residuals have positive autocorrelation. The 0.0881 p-value for order 2 comes close to rejecting the null hypothesis.

Figure 7 shows the cumulative distribution functions for the residuals (blue curve) and for the normal distribution that has the same mean and standard deviation as the residuals. For each value on the horizontal axis, the cumulative distribution function shows the proportion of data (residuals – in blue) or probability (normal – in red) below that value. For example, at 0 on the horizontal axis, 50% of the probability is below 0, but about 45% of the residuals are. The blue and red curves in Figure 7 are made from the corresponding blue and red curves in Figure 4 by making running totals of the blue and red curves in Figure 4.

If the residuals were perfectly normal, then the blue curve should overlay the red normal curve perfectly. It does not. But it looks close. To the extent that the blue curve deviates from the red normal curve in Figure 7 doubt is cast upon the putative normality of the residuals. So if we can quantitatively measure how far apart the blue and red curves are, we will have a statistic for testing N. We will reject N for large values of that statistic (large differences between the curves.).

Figure 7. The cumulative distribution of residuals (blue) and normal distribution (red) with same mean and standard deviation for regression of rent on area and bathrooms [13]



It turns out that there are at least four good ways to measure how far apart the blue and red curves are. Each way leads to a different test for N. All four tests are based on the idea of computing the distance between an ideal normal distribution (red) and the residuals (blue). If that distance is large, then reject normality; if the distance is small, then accept normality. The tests differ in the way they measure the distance from "ideal" normality:

- The Kolmogorov-Smirnov test is based upon the maximum magnitude of vertical separation between the blue curve and ideal normal red curve with the same mean and standard deviation. The larger the maximum deviation, the less likely N is true.
- The Cramer-von Mises test is based upon the integrated squared vertical deviation between the two curves. The larger the integral, the less likely N is true.
- The Anderson-Darling statistic is a modification of the Kolmogorov-Smirnov test that gives more weight to the tails of the distribution. The larger the weighted deviation, the less likely N is true.

[13] Plot produced by software program @Risk.

- The Shapiro-Wilk statistic is essentially the squared correlation (R-square) between the actual residual values and the ideal values that the residuals "should" have if they were exactly normal. To find the "ideal" value for a given residual, locate the residual value on the horizontal axis, then read straight up to the blue curve, then horizontally over to the red curve, then straight down to the horizontal axis. For example, the residual value 0 actually has about 45% of data values below it (reading up to the blue curve); but the residual value that should have 45% below it (reading over to red curve and down to horizontal axis) is about -10. The closer the blue and red curves are to each other, the closer the actual and ideal values will be. The closer the correlation of actual and ideal values is to 1, the more likely the residuals are normal. The smaller the correlation, the less likely the residuals are normal. Whereas the other three tests reject N for large values of their statistics, the Shapiro-Wilk test rejects normality for small values of the Shapiro-Wilk statistic – i.e., S-W rejects N for small correlations.

Studies indicate that the best of these four tests for most data is probably the Shapiro-Wilk, followed closely by the Anderson-Darling. The Kolmogorov-Smirnov test seems to lack power, especially in small to medium-sized datasets – that is, K-S often accepts normality when it should reject..

Example. Here is how these four tests of N work out in SAS for the residuals from the regression of rent on area and bathrooms.

The first step in checking for non-normality is to compute the residuals. In SAS the residuals can be computed by PROC REG and put into a SAS dataset for diagnosis by another procedure.

The SAS code

```
proc reg data=apts;
  model rent = area bathrooms;
  output out=res_apts r=resid_rent;
run;
proc univariate data=res_apts normal;
  var resid_rent;
run;
```

The **OUTPUT** statement begins the creation of a SAS dataset. **OUT=** names the SAS dataset to be **WORK.res_apts**. **R=** creates a variable called **resid_rent**, which equals the residual for each observation, i.e., resid_rent = rent – predicted rent. **WORK.res_apts** contains the original **WORK.apts** plus one new variable, the residuals. Although the predicted values are not used in diagnosing N, they are useful for other diagnostics (such as L), so it often saves effort to create both predicted values (by the additional option **P=**pred_rent) and residuals at the same time.

The **NORMAL** option on the **proc univariate** line invokes four formal tests of normality for the variables named in the **VAR** statement. Since the output of **proc univariate** can be long, only the output for the four tests is shown here.

The SAS output

Tests for Normality

Test	--Statistic---		-----p Value------
Shapiro-Wilk	W	0.988535	Pr < W 0.8460
Kolmogorov-Smirnov	D	0.079506	Pr > D >0.1500
Cramer-von Mises	W-Sq	0.040556	Pr > W-Sq >0.2500
Anderson-Darling	A-Sq	0.229003	Pr > A-Sq >0.2500

Generally, Shapiro-Wilk is best, followed by Anderson-Darling
Kolmogorov- Smirnov is found to be lacking power

The null hypothesis for all four tests is H_0 : *the distribution of the residuals is normal*. Thus, the benefit of the doubt is given to normality. That is, when you use any of these tests, you are assuming that the distribution is normal unless the evidence is compelling that it is not. The results for all four of the tests is to fail to reject normality, albeit at different significance levels. We conclude that the residuals are close enough to being normal and N is accepted.

SUMMARY

A regression is valid if the residuals satisfy LHIN – both time series and cross-sectional regressions.

Qualitative tests of LHIN are provided by plots of the residuals:

- L – plot residuals vs estimates. The residuals should scatter roughly evenly above and below the horizontal line at zero from left to right across the plot. This indicates that the mean value of the residuals is roughly zero in every region, so estimates are not biased in any region.
- H – plot squared residuals vs estimates. The squared residuals should average approximately the same value from left to right across the plot. This indicates that the mean deviation of the data from the estimates is about the same in every region.
- I – For cross-sectional data, look to the method of selection of the data: If the data are drawn as a random sample, then I is satisfied. For time series data, plot residuals vs lag residuals (maybe more than one lag). The plot should be roughly flat – indicating no relationship between residuals and their lag(s).
- N – make a histogram of the residuals and see if the shape of the histogram appears normal (bell-shaped).

Figures 1 – 4 provide examples of the qualitative tests of LHIN.

Quantitative tests of LHIN remove some of the subjectivity in judging the qualitative tests. But the quantitative tests build on the ideas behind the qualitative tests.

- L – Ramsey's RESET procedure fits quadratic, cubic, and quartic curves to the plot of residuals vs estimates and tests whether any of the coefficients of the power terms differ from zero. The idea is that if the coefficients of the power terms are zeroes, then the curves collapse into flat lines at the horizontal line at zero. This indicates that the mean residual (which is the mean actual – mean estimate) is roughly zero everywhere.
- H – White's test of homoscedasticity fits a second-order curve (first, second, and cross-product terms of all original predictors) to the squared residuals and tests whether any of the coefficients differ from zero. The idea is similar to that of RESET: If the coefficients are all zeroes, then the response surface collapses into a flat (constant) surface. Then the magnitudes of the residuals are roughly the same for any combination of predictor values.
- I – For time series (no quantitative test for cross-sectional data): The Durbin-Watson test for autocorrelation scales the sum of squared changes from one residual to the next in time order. Small DW indicates strong positive relationship between successive residuals; large DW indicates strong negative relationship between successive residuals; medium DW indicates no consistent relationship between successive residuals.
- N – The Shapiro-Wilk, Anderson-Darling, Cramer-von Mises, and Kolmogorov-Smirnov tests all measure the closeness of the residuals to an ideal normal distribution and reject normality if the residuals are not close.

All of the quantitative tests are implemented in SAS.