LECTURE / DISCUSSION

Hypothesis Testing

Introduction

Consider the model you estimated in the previous workshop. The output for the first model is:

```
Dependent variable: USAGE
Current sample:  1 to 500
Number of observations:  500

 Mean of dependent variable = 229970.
Std. dev. of dependent var. = 91538.8
   Sum of squared residuals = .362272E+13
      Variance of residuals = .731863E+10
   Std. error of regression = 85549.0
                  R-squared = .133589
         Adjusted R-squared = .126588
    Durbin-Watson statistic = 2.06459
  F-statistic (zero slopes) = 19.0806
 Schwarz Bayes. Info. Crit. = 22.7658
 Log of likelihood function = -6385.38

           Estimated     Standard
Variable  Coefficient      Error       t-statistic
C         127568.        15133.6       8.42949
AC        -20606.7       8116.66       -2.53882
CDD       23.9860        4.80534       4.99152
NEMPLOY   -6.61867       89.4999       -.073952
SQFT      931.616        147.957       6.29652
```

The point estimates are in the first column of numbers. This lecture introduces you to the use of the other columns.

The second column gives numbers that describe the statistical precision of the point estimates. Because the regression relationship includes a disturbance term, one cannot estimate the coefficients on the explanatory variables exactly. If a new data set were collected, and the same regression coefficients re-estimated, the new estimates would not equal the initial ones because the OLS estimators are random variables. The second column contains estimates of the standard deviations of each OLS regression coefficient; these estimates are commonly called ***standard errors***. A common rule of thumb is that the actual regression coefficient is probably within two standard errors of its point estimate. Thus, the actual audit impact is probably between 4,375. and 36,839.

The third column of numbers, labeled ***t-statistics,*** gives the ratio of the point estimates over their standard errors. These ratios describe the likelihood that the actual coefficient is zero. Using the ‘‘two standard errors” rule of thumb, an actual coefficient is unlikely to be zero if the t-statistic is larger than 2 in absolute value. In these estimates, it is likely that the number of employees has no impact on monthly electricity consumption in a building (given its size and the weather).

In the material that follows, we explain the basis for the ‘‘two standard errors” rule and how this rule is generalized to examining several coefficients simultaneously. If we are going to forecast energy consumption with this estimated regression, we will want to compute standard errors for the forecasts that include the variation in a combination of the estimated coefficients:

$$\hat{y} = \sum_{k=1}^{K} \hat{\beta}_k x_k$$

is a forecast of y given the explanatory variable values x. Another way we might be concerned with many coefficients simultaneously occurs in combining distinct data sets: before pooling all observations in a single, larger, data set, we want to check whether there is any evidence that the actual regression coefficients are different in the distinct data sets. This is analogous to asking whether there is evidence for a set of coefficients where the coefficients are not zero.

Confidence Intervals

If the regression model is specified correctly, so that no explanatory variables are missing, then

$$\frac{\hat{\beta}_k - \beta_k}{s_k} \sim t_{N-K}$$

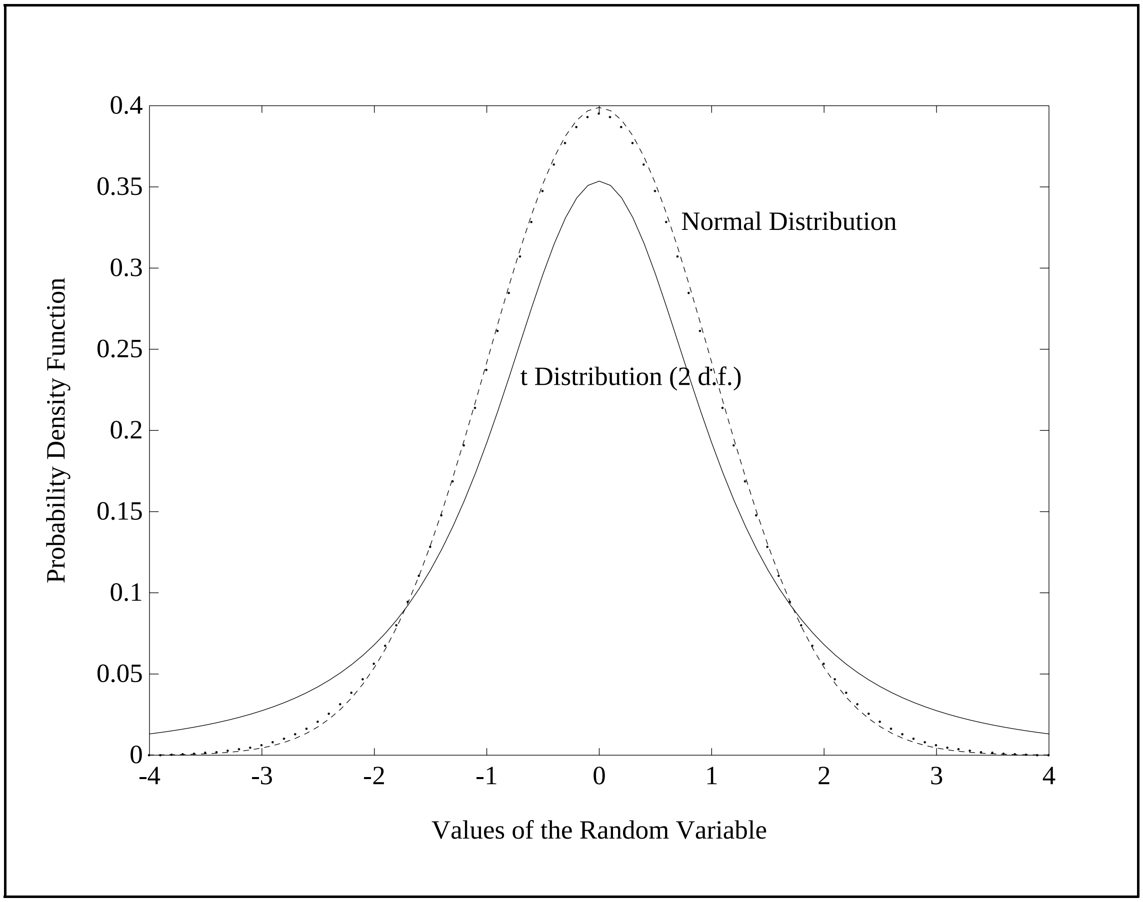
where β_k is the true value of the kth explanatory variable, $\hat{\beta}_k$ is the OLS estimator of this coefficient, and s_k is the estimated standard error of $\hat{\beta}_k$. The symbol t_{N-K} represents a random variable with a particular distribution: Student's t with N - K degrees of freedom. We will use this fact to:

1. Construct confidence intervals for individual regression coefficients.

2. Test hypotheses about the individual regression coefficients.

Student's t Distribution

The Student's t distribution, or simply t distribution, is a generalization of the standard normal distribution. The distribution is characterized by an additional parameter, called the ***degrees of freedom.*** For small degrees of freedom (less than 10), the t distribution has a probability density function that is much 'fatter' than the normal distribution: the tails of the density approach zero more slowly. As the degrees of freedom increase, the t distribution approaches the normal distribution so that in the limit the two distributions coincide. When the degrees of freedom exceed 30, the coincidence is quite close.



Student's t and Normal Probability Density Functions

The statistic $(\hat{\beta}_k - \beta_k)/s_k$ has been ***standardized*** by subtracting from $\hat{\beta}_k$ its expectation and then dividing the result by the estimated standard error. Given the correct value of β_k, this statistic has a t distribution with degrees of freedom equal to the number of observations in the sample minus the total number of explanatory variables in the estimated regression. The distribution of $(\hat{\beta}_k - \beta_k)/s_k$ does not depend on any unknown parameters.

Probability Intervals

Because we can compute the probability density function of any t distribution, ***given any interval*** [a,b], we can compute the probability that $(\hat{\beta}_k - \beta_k)/s_k$ will fall in the interval if we are given a brand new data set of N observations. ***Given a probability*** p, we can also find an interval $[a,b]$ so that

$$\Pr\left\{a \le \frac{\hat{\beta}_k - \beta_k}{s_k} \le b\right\} = p$$

The interval $[a,b]$ is called a ***probability interval.***

Because the t probability density function is symmetric around zero, an important family of intervals consists of those that are also symmetric around zero. These intervals are the shortest ones for a given probability value. Table 3 (page 563) of Pindyck and Rubinfeld tabulates such intervals for different degrees of freedom and various probability values (statistical tables are reprinted at the back of this binder).

For example, according to the table, if N - K = 15 then

$$\Pr\left\{-1.753 \leq \frac{\hat{\beta}_k - \beta_k}{s_k} \leq 1.753\right\} = 1.0 - 0.1 = 0.9 \quad .$$

The value 1.753 is listed under the column labeled .10 and in the row labeled 15.

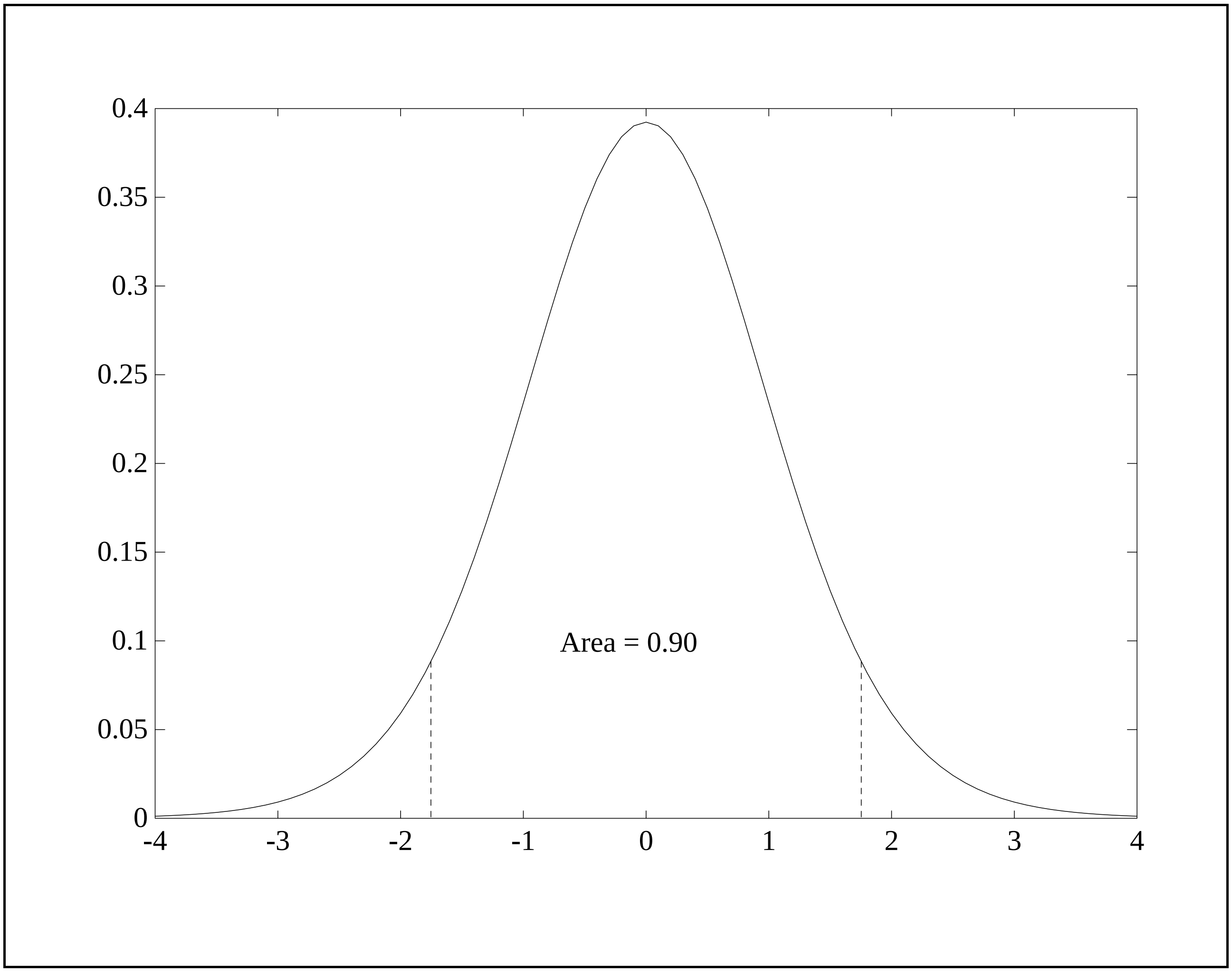
Confidence Intervals

Given a data set, we can compute N - K, $\hat{\beta}_k$, s_k , and symmetric probability intervals for any probability level. The only unknown element of the formula above would be β_k , and if we undo the standardization of $\hat{\beta}_k$, we can compute a ***confidence interval*** for the unknown β_k . For example,

$$\Pr\left\{\hat{\beta}_k - 1.753 \cdot s_k \le \beta_k \le \hat{\beta}_k + 1.753 \cdot s_k\right\} = 0.9 \quad .$$

We say in words that ***the 90% confidence interval*** *for* β_k is $[\hat{\beta}_k - 1.753 \cdot s_k ,\ \hat{\beta}_k + 1.753 \cdot s_k] = [\hat{\beta}_k \pm 1.753 \cdot s_k]$.

This is not a probability interval because, given the data set and the values of $\hat{\beta}_k$ and s_k , the probability that β_k falls in the interval is zero or one.



90% Probability Interval Student's t Distribution

The interval is called a ***confidence*** interval because the probability level expresses our degree of confidence that β_k is inside the computed interval. Because we do not know β_k , we do not know whether this coefficient is inside the interval. But we do know the probability that this procedure will produce an interval containing β_k if we were to repeat our procedure with a new data set.

Confidence intervals are sometimes called ***interval estimators.*** This is a good term to remember because it emphasizes that confidence intervals are random variables. They provide estimates of a range of likely values instead of just a ***point*** estimate. The OLS estimator $\hat{\beta}_k$ is a point estimator.

Now we can see where the "two-standard-errors" rule originates. If you look at Table 3 again, you will see that many of the numbers in the .05 column are near the number 2. Ninety-five percent confidence intervals are a convention in econometric practice and sample sizes often imply degrees of freedom greater than 30 so that confidence intervals are often four standard errors wide.

Hypothesis Tests: The t Test

Model building usually involves an examination of parameter estimates to judge whether estimates make sense. If one finds that an important property of the desired model is not satisfied by the estimates, then one is compelled to rethink the empirical model and attempt to resolve a potential conflict between data and theory. If, for example, consumers' implicit discount rate is estimated to be 500% or demand is estimated to increase with price, then the usefulness of the empirical model is thrown into doubt.

Hypothesis tests formalize a way to judge whether parameter estimates are consistent with theoretical requirements. In the simplest case, we ask whether an estimate of a parameter is contradictory to an hypothesized value: is the empirical evidence consistent with $\beta_k = \beta_k^0$?

One way to describe the standard t test of this hypothesis is that one chooses a confidence level and checks whether the corresponding interval estimator for β_k includes the hypothesized value. If it does, then the empirical evidence is consistent with the hypothesis at the chosen confidence level. Otherwise, there is evidence that the hypothesized value of β_k is not β_k^0. But this is not the usual description.

An equivalent, and more common description is to compute the ***t statistic***

$$t = \frac{\hat{\beta}_k - \beta_k^0}{s_k}$$

and to choose a ***significance level,*** the probability that one is willing to risk rejecting the hypothesis when it actually holds. A popular choice is 0.05. Finally, one checks whether the t statistic falls in the ***acceptance region*** for the test. The acceptance region is simply the symmetric probability interval for the t distribution corresponding to $1 - \alpha$ probability. If the t statistic falls outside the acceptance region, an unlikely event, it suggests that the hypothesis is false. The boundary of the acceptance region is often called the ***critical value***.

Rather than choose a significance level, many researchers report the probability, or ***P-value,*** at which the t statistic would be on the boundary of the acceptance region.

Regression software routinely report t statistics for whether coefficients are different from zero. If a t statistic is larger than its critical value, the estimated coefficient is said to be ***statistically significant***. One should bear in mind the difference between statistically significant and other measures of importance.

Restricted Estimation

If a parameter restriction is true, then estimation that imposes the restriction will provide more precise estimators. To impose a parameter restriction to zero, one simply drops the corresponding explanatory variable from the regression. If $\beta_3 = 0$ is the restriction, then

$$\begin{aligned} y &= \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \ldots + \beta_K x_K + u \\ &= \beta_1 + \beta_2 x_2 + 0 \cdot x_3 + \beta_4 x_4 + \ldots + \beta_K x_K + u \\ &= \beta_1 + \beta_2 x_2 + \beta_4 x_4 + \ldots + \beta_K x_K + u \end{aligned}$$

We will call this ***restricted*** estimation. Dropping an explanatory variable will cause the sum of squared residuals to rise and this rise is related to the t statistic for testing whether the coefficient is zero:

$$|t| = \sqrt{\frac{SSR_R - SSR_U}{s^2}}$$

where SSR_R is the sum of squared residuals from the restricted regression, SSR_U is the sum of squared residuals from the original, unrestricted regression, and s^2 is the estimated variance from the unrestricted regression.

We can extend restricted estimation to nonzero restrictions. If $\beta_3 = 2$ is the restriction, then

$$
\begin{aligned}
y &= \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \ldots + \beta_K x_K + u \Leftrightarrow \\
y &= \beta_1 + \beta_2 x_2 + 2 \cdot x_3 + \beta_4 x_4 + \ldots + \beta_K x_K + u \Leftrightarrow \\
y - 2x_3 &= \beta_1 + \beta_2 x_2 + \beta_4 x_4 + \ldots + \beta_K x_K + u
\end{aligned}
$$

and the restricted estimation is a regression of a new dependent variable, $y - 2x_3$, on the abbreviated list of explanatory variables $1, x_1, x_2, x_4, \ldots, x_K$. The calculation of the t statistic as the ratio of the change in the sum of squared residuals to the unrestricted estimate of the variance is unchanged.

We can also extend restricted estimation to restrictions involving several parameters. If $\beta_3 = \beta_4$ is the restriction, then

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \ldots + \beta_K x_K + u \Leftrightarrow$$
$$y = \beta_1 + \beta_2 x_2 + \beta_4 x_3 + \beta_4 x_4 + \ldots + \beta_K x_K + u \Leftrightarrow$$
$$y = \beta_1 + \beta_2 x_2 + \beta_4 (x_3 + x_4) + \ldots + \beta_K x_K + u$$

and the restricted estimation is a regression of the original dependent variable, y, on the new list of explanatory variables $1, x_1, x_2, (x_3+x_4), x_5, \ldots, x_K$. The calculation of the t statistic as the ratio of the change in the sum of squared residuals to the unrestricted estimate of the variance is unchanged.

Multivariate Restricted Estimation

There are many situations in which we consider several parameter restrictions at once. Perhaps the simplest example is restricting a list of explanatory variable coefficients to be zero. That is, we consider dropping a list of explanatory variables from the right-hand side of the regression. The most common example of this is the list of all explanatory variables in the model, except the constant term. Researchers often examine this ’straw man’ hypothesis as a general diagnostic test of the explanatory power of the regression model.

When all of the constraints set coefficients to zero, imposing several restrictions to obtain the restricted regression is just like imposing a single restriction: one simply drops the corresponding explanatory variables from the regression.

Multivariate Hypothesis Tests: The F Test

In general, we test several restrictions using the F test, a test that generalizes the t test principle:

$$F = \frac{(SSR_R - SSR_U)/M}{s^2}$$

where M is the number of restrictions, SSR_R is the sum of squared residuals from the restricted regression, SSR_U is the sum of squared residuals from the original, unrestricted regression, and s^2 is the estimated variance from the unrestricted regression. The t statistic has a square root in its formula and it only applies to testing one restriction at a time. Clearly, $t^2 = F$ when we are testing one restriction.

When we are testing whether all of the explanatory variables have coefficients equal to zero, $M = K\text{-}1$. The TSP program always prints the F statistic for this test.

The F statistic has a distribution that depends on two degrees of freedom parameters: $K - 1$, the number of restrictions under test, and $N - K$, the degrees of freedom parameter that we encountered in the t test. The critical values of this statistic for the 5% and 1% significance levels are listed in Tables 4a and 4b of Pindyck and Rubinfeld.

Goodness of Fit: R^2 and $\overline{R}^2$

The F statistic for testing the significance of the entire regression model is closely related to the R^2 goodness-of-fit measure:

$$F = \frac{R^2}{1 - R^2} \frac{N - K}{K - 1} .$$

You can see that a high R^2 does not imply a large value of F: if there are many explanatory variables, F may be small. Because R^2 does not take into account the number of explanatory variables in a regression model, some researchers prefer a corrected measure, denoted $\overline{R}^2$, that is similar in spirit to the F statistic. Whereas

$$R^2 = 1 - \frac{SSR_U}{SST} ,$$

the corrected R^2 is given by

$$\overline{R}^2 = 1 - \frac{SSR_U}{SST} \frac{N - 1}{N - K} .$$

Thus, $\overline{R}^2$ is related to R^2 by the formula

$$\overline{R}^2 = 1 - (1 - R^2)\frac{N-1}{N-K} \quad .$$

This measure penalizes a regression for having extra explanatory variables, a way that one can always increase R^2.

Chow Test

Another common set of restrictions arises when the researcher is deciding whether to combine (or 'pool') two or more data sets with the same variables. It is natural to ask whether the regression coefficients are significantly different among the data sets. Consider the special case of two data sets. Without the restrictions, data set 1 and data set 2 possess different regression equations:

$$y_1 = \sum_{k=1}^{K} x_{1k}\beta_{1k} + u_1$$

$$y_2 = \sum_{k=1}^{K} x_{2k}\beta_{2k} + u_2$$

This can also be written as a single equation if we define a dummy variable d such that

$$d = \begin{Bmatrix} 1 & \text{if the observation is from data set 1} \\ 0 & \text{if the observation is from data set 2} \end{Bmatrix}$$

then

$$y = \sum_{k=1}^{K} (d \cdot x_{1k})\beta_{1k} + \sum_{k=1}^{K} \left[(1-d) \cdot x_{2k}\right]\beta_{2k} + u$$

With the restrictions, data sets 1 and 2 have the same coefficients:

$$y_1 = \sum_{k=1}^{K} x_{1k}\beta_{1k} + u_1$$

$$y_2 = \sum_{k=1}^{K} x_{2k}\beta_{2k} + u_2$$

and the restrictions are

$$\beta_{1k} = \beta_{2k} \quad , \quad k = 1,\ldots,K \quad .$$

Written as a single equation,

$$\begin{aligned} y &= \sum_{k=1}^{K} \left[(d \cdot x_{1k}) + (1-d) \cdot x_{2k} \right] \beta_{2k} + u \\ &= \sum_{k=1}^{K} x_k \beta_{2k} + u \end{aligned}$$

Estimating restricted and unrestricted regressions is straightforward. Unrestricted estimation simply runs a large regression with two sets of explanatory variables and coefficients. Restricted estimation is simply a single regression using all of the data with a common set of explanatory variables and coefficients. The F test of the restrictions is

$$F = \frac{(SSR_R - SSR_U)/K}{s^2}$$

where $s^2 = SSR_U \ / \ (N - 2K)$ is the estimated variance for the unrestricted regression, and N is the total number of observations in both data sets. The number of restrictions tested is K in this case, not just $K - 1$. The degrees of freedom are K and $N - 2K$. This statistic has the special name ***Chow*** test.

Example of a Chow Test

Two cross-sections of residential phone users were collected: one in the San Francisco Bay Area and the other in the greater Los Angeles area. Regressions for monthly phone bills were run on various household characteristics. The sum of squared residuals for the LA regression was 1523.76 for 1000 households and the sum of squared residuals for the Bay Area regression was 1205.13 for 750 households. A single regression using both data sets yielded SSR = 2850.81. Given K = 8 explanatory variables, The Chow/F test of equal coefficients in LA and the Bay Area is

$$F = \frac{(2850.81 - (1523.76 + 1205.13))/8}{(1523.76 + 1205.13)/(1750 - 16)}$$

$$= 9.68$$

Three Equivalent Tests

We have already described two equivalent ways to write the test for a single zero restriction. Firstly, we can write

$$F = \frac{\hat{\beta}_k^{\,2}}{s_k^{\,2}}$$

as the F statistic (or squared t-statistic). This version is called the ***Wald*** test statistic.

Secondly, we can also write

$$F = \frac{SSR_R - SSR_U}{s^2}$$

This version is called the ***likelihood ratio*** test.

We can also write the test statistic in the so-called ***score*** or ***Lagrange multiplier*** form

$$F = \frac{\hat{\gamma}^2}{s_\gamma^{\,2}}$$

where γ is the regression coefficient from an artificial regression of the fitted residuals from the restricted regression on the single explanatory variable x_k.

Each of these statistics illustrates a general testing method that we will also apply later. For the moment, note the differences in their calculation.

The ***Wald*** statistic is computed from the estimation of the ***unrestricted*** model only.

The ***Lagrange multiplier*** is computed from the estimation of the ***restricted*** model only.

The ***likelihood ratio*** statistic requires the estimation of *both* models and depends on the difference in estimation function, the sum of squared errors.