LECTURE 11

SIMPLE REGRESSION MODEL - III

Outline of today's lecture:

I. Variance of the random variable *u*

The formulae of the variance of $\hat{\beta}_0$ and $\hat{\beta}_1$ involve the variance of the random term *u*, which we have denoted by σ^2.

However, the true variance of u_t can not be computed since the values of u_t are not observable. But we may obtain an *unbiased* estimate of σ^2 from the expression

$$S^2 = \hat{\sigma}^2 = \frac{\sum_{t=1}^{T} \hat{u}_t^2}{T-2}$$

where $\hat{u}_t$ is the OLS residual, and hence it is given by $\hat{u}_t = Y_t - \hat{Y}_t$.

II. A Measure of Goodness of Fit: Coefficient of Determination (R^2)

We now consider the goodness of fit of the fitted regression line to a set of data; that is, we shall find out how "well" the sample regression line fits the data.

The coefficient of determination r^2 (two-variable case) or R^2 (multiple regression) is a summary measure that tells how well the sample regression line fits the data.

We need to know how "good" is the fit of the regression line to the sample observations of Y and X, that is to say we need to measure the dispersion around the regression line.

This knowledge is essential, because the closer the observations to the line, the better the goodness of fit, that is better is the explanation of the variations of *Y* by the changes in the explanatory variables.

Below, we will show prove that a measure of the goodness of fit is the square of the correlation coefficient, r^2, which shows the percentage of the total variation of the dependent variable that can be explained by the independent variable X.

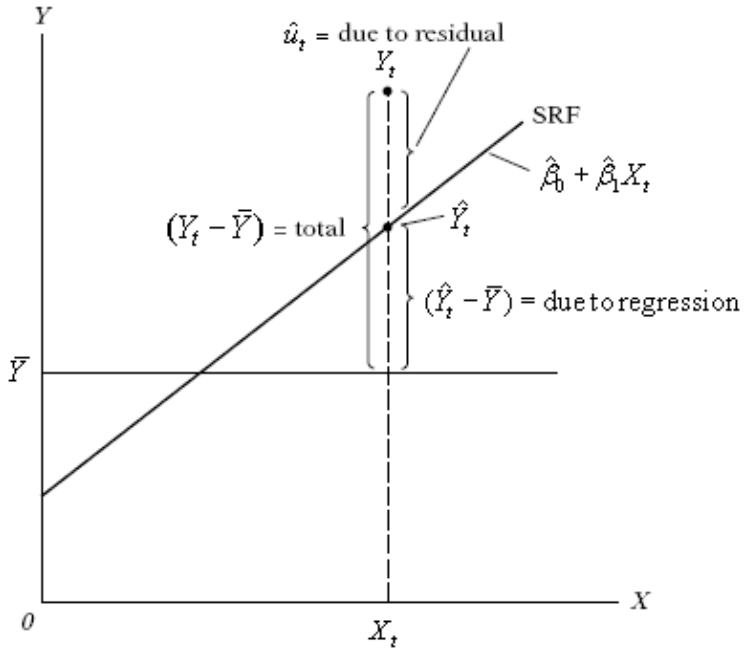

Figure 2. Breakdown of the variation of Y_t into two components.

By fitting the line $Y_t = \hat{\beta}_0 + \hat{\beta}_1 X_t$ we try to obtain the explanation of the variations of the dependent variable *Y* produced by the changes of the explanatory variable *X*. However, the fact that the observations deviate from the estimated line shows that the regression line explain only a part of the total variation of the dependent variable. A part of the variation, defined as $\hat{u}_t = Y_t - \hat{Y}_t$, remains unexplained.

(1) We may compute the total variation of the dependent variable by comparing each value of *Y* to the mean value $\bar{Y}$ and adding all the

resulting deviations.[1] Denoting the deviations of the values Y_t around their mean $\bar{Y}$ by lower case letters we have:

$$\left[\text{Total Variation in Y}\right]=\sum_{t=1}^{T} y_t^2=\sum_{t=1}^{T}\left(Y_t-\bar{Y}\right)^2 \qquad (1)$$

Note that in order to find the total variation of the Y_t's we square the simple deviations, since by definition the sum of the simple deviations of any variable around its mean is identically equal to zero: $\sum_{t=1}^{T}\left(Y_t-\bar{Y}\right)=0$.

(2) In the same way we define the deviation of the regressed values, $\widehat{Y}_t$'s from the mean value, $\hat{y}_t=\widehat{Y}_t-\bar{Y}$. This is the part of the total variation of $\widehat{Y}_t$ which is explained by the regression line. Thus the sum of the squares of these deviations is the total *explained by the regression line variation* of the dependent variable.

$$\left[\text{Explained Variation in } Y\right]=\sum_{t=1}^{T}\hat{y}_t^2=\sum_{t=1}^{T}\left(\hat{Y}_t-\bar{Y}\right)^2$$

[1] When we speak of changes in *Y* we must define the "basis of reference", that is a value of the variable *Y*, to which we compare any other value that may be assumed by this variable. As such reference we may take the origin (*Y*=0) or the mean value ($\bar{Y}$) or any other statistic of *Y* (the median, etc.). However, it is customary and computationally convenient to take the mean as the reference value and express the total variation of *Y*'s as the sum of the deviations of the *Y*'s from their mean.

(3) We have already defined the residual $\hat{u}_t$ as the difference $\hat{u}_t = Y_t - \hat{Y}_t$, that is as the part of the variation of the dependent variable which is not explained by the regression line and is attributed to the existence of the disturbance variable u_t. Thus the sum of the squared residuals gives the total unexplained variation of the dependent variable Y around its mean.

$$\left[\text{Unexplained Variation in } Y\right] = \sum_{t=1}^{T} \hat{u}_t^2 = \sum_{t=1}^{T} \left(Y_t - \hat{Y}_t \right)^2$$

We can write:

$$\sum_{t=1}^{T} y_t^2 = \sum_{t=1}^{T} \hat{y}_t^2 + \sum_{t=1}^{T} \hat{u}_t^2 \qquad (2)$$

or

$$\underbrace{\left[\begin{array}{c}\text{Total} \\ \text{Variation in } Y\end{array}\right]}_{\substack{\text{Total Sum of Squares} \\ \text{(TSS)}}} = \underbrace{\left[\begin{array}{c}\text{Explained} \\ \text{Variation in } Y\end{array}\right]}_{\substack{\text{Explained Sum of Squares} \\ \text{(ESS)}}} + \underbrace{\left[\begin{array}{c}\text{Unexplained (residual)} \\ \text{Variation in } Y\end{array}\right]}_{\substack{\text{Residual Sum of Squares} \\ \text{(SSR)}}}$$

Note that the *explained variation expressed as a percentage of total variation* is given by:

$$\sum_{t=1}^{T} \hat{y}_t^2 \Big/ \sum_{t=1}^{T} y_t^2$$

Dividing both sides of Equation (2) by $\sum_{t=1}^{T} y_t^2$ produces:

$$\frac{\sum_{t=1}^{T} y_t^2}{\sum_{t=1}^{T} y_t^2} = \frac{\sum_{t=1}^{T} \hat{y}_t^2}{\sum_{t=1}^{T} y_t^2} + \frac{\sum_{t=1}^{T} \hat{u}_t^2}{\sum_{t=1}^{T} y_t^2}$$

$$1 = \underbrace{\frac{\sum_{t=1}^{T} \hat{y}_t^2}{\sum_{t=1}^{T} y_t^2}}_{R^2} + \frac{\sum_{t=1}^{T} \hat{u}_t^2}{\sum_{t=1}^{T} y_t^2}$$

or, $1 = R^2 + \dfrac{\sum_{t=1}^{T} \hat{u}_t^2}{\sum_{t=1}^{T} y_t^2}$

Therefore,

$$R^2 = 1 - \frac{\sum_{t=1}^{T} \hat{u}_t^2}{\sum_{t=1}^{T} y_t^2}$$

Thus R^2 determines the proportion of the variation in Y which is explained by variation in X. For this reason R^2 is called the *coefficient of determination*. For example, if R^2=0.90, this means that the regression line gives a good fit to the observed data, since this line explains 90 percent of the total variation of the *Y* around their mean. In other words, 95% of the variation of the dependent variable around its sample mean is explained by the systematic part of the model. The remaining 10 per cent of the total variation in Y is

accounted for the regression line and is attributed to the factors included in the disturbance variable u_t.

Two properties of R^2 may be noted:

1. It is a nonnegative quantity. (Why?)
2. Its limits are $0 \leq R^2 \leq 1$. An R^2 of 1 means a *perfect fit*, that is, $Y_t = \hat{Y}_t$ for each *t*. On the other hand, an R^2 of zero means that there is no relationship between the *regressand* and the *regressor* (i.e., $\hat{\beta}_1 = 0$). In this case, $\hat{Y}_t = \hat{\beta}_1 = \bar{Y}$, that is, the best estimation of any *Y* value is simply its mean value. In this situation therefore the regression line will be horizontal to the *X* axis.

A. Relationship between R^2 and Correlation Coefficient

$$R^2 = \frac{\sum_{t=1}^{T} \hat{y}_t^2}{\sum_{t=1}^{T} y_t^2} = \frac{\sum_{t=1}^{T} (\hat{\beta}_1 x_t)^2}{\sum_{t=1}^{T} y_t^2} = \hat{\beta}_1^2 \frac{\sum_{t=1}^{T} x_t^2}{\sum_{t=1}^{T} y_t^2} = \left[\frac{\sum_{t=1}^{T} y_t x_t}{\sum_{t=1}^{T} x_t^2} \right]^2 \frac{\sum_{t=1}^{T} x_t^2}{\sum_{t=1}^{T} y_t^2}$$

$$R^2 = \frac{\left[\sum_{t=1}^{T} y_t x_t \right]^2}{\left[\sum_{t=1}^{T} x_t^2 \right]^2} \frac{\sum_{t=1}^{T} x_t^2}{\sum_{t=1}^{T} y_t^2} \rightarrow R^2 = \frac{\left[\sum_{t=1}^{T} y_t x_t \right]^2}{\sum_{t=1}^{T} x_t^2 \sum_{t=1}^{T} y_t^2}$$

Recall that the formula of *sample correlation coefficient (r)* is given by:

$$r_{XY} = \frac{\sum_{t=1}^{T} y_t x_t}{\sqrt{\sum_{t=1}^{T} x_t^2 \sum_{t=1}^{T} y_t^2}}$$

Hence we have shown that $R^2 = (r_{XY})^2$. We conclude that the square root of R^2 has the following relationship: $\sqrt{R^2} = r_{XY}$

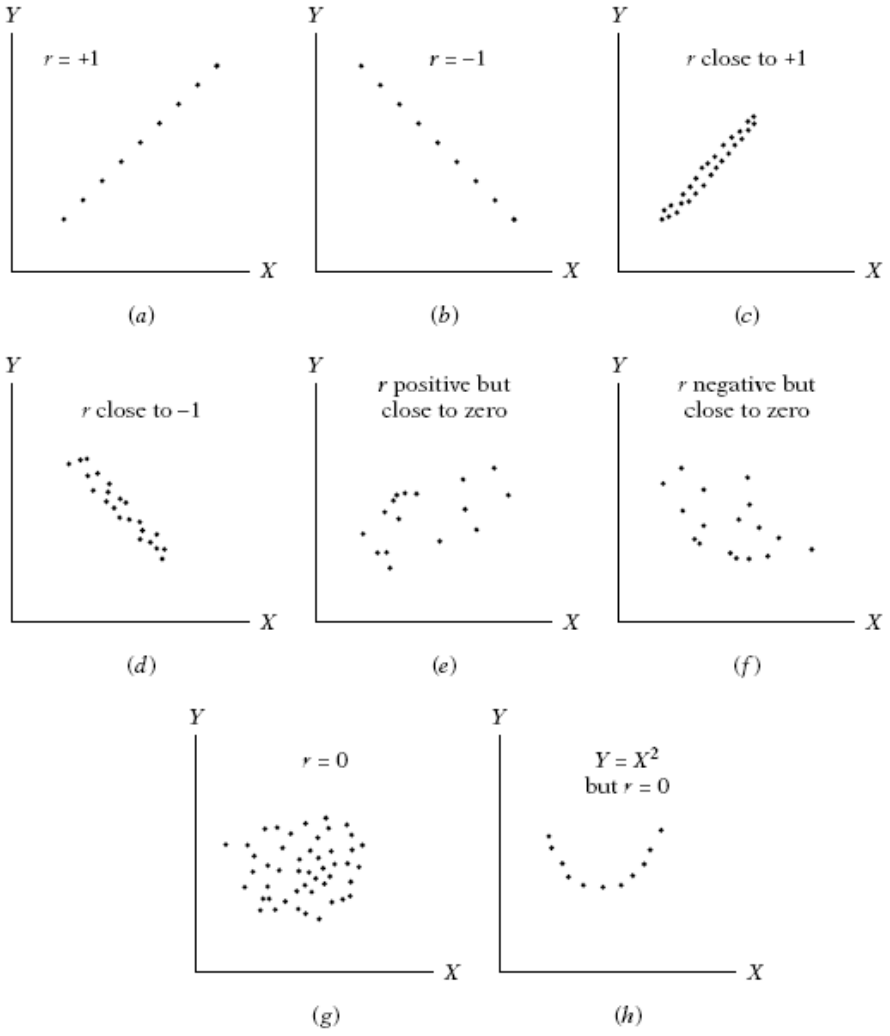

Figure 3. Correlation patterns

III. Sampling Distribution of the Least Squares Estimates

We have found expressions for the mean and variance of the least squares estimates. Given that the random variable u_t is normally distributed, it can be proved that the distribution of the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ is also normal.

These results may be stated in summary form:

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2 \sum X_t^2}{T \sum x_t^2}\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum x_t^2}\right)$$

The *normal distributions* above can be *standardized*, that is they can be transformed into the units of the standard normal variable Z, which has zero mean and unit variance, $Z \sim \text{N}(0, 1)$, through the following transformation formula:

$$Z_t = \frac{\hat{\beta}_0 - \beta_0}{\sigma_{\hat{\beta}_0}} \sim \text{N}(0, 1) \text{ for } \hat{\beta}_0$$

$$Z_t = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \sim \text{N}(0, 1) \text{ for } \hat{\beta}_1$$

where

$$\sigma_{\hat{\beta}_0} = \sqrt{\frac{\sigma^2 \sum X_t^2}{T \sum x_t^2}} \text{ and } \sigma_{\hat{\beta}_1} = \sqrt{\frac{\sigma^2}{\sum x_t^2}}$$

IV. Multiple Regression Analysis

Multiple regression analysis is an extension of simple regression analysis to cover cases in which the dependent variable is hypothesized to depend on more than one explanatory variable. The simplest possible multiple regression model is three-variable regression, with one dependent variable and two explanatory variables.

If we generalize the two- and three-variable linear regression models, the *k*-variable *population regression model* (PRF) involving the dependent variable *Y* and *k−1* explanatory variables $X_1, X_2, \ldots, X_{k-1}$ may be written as

$$PRF: \qquad Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \ldots + \beta_k X_{tk} + u_t \quad t=1,2,3,\ldots,T \qquad (i)$$

where β_0=the intercept term, β_1 to β_k=*partial* slope coefficients, *u*=disturbance term.

V. Two Explanatory Variable Regression

We can write the three variable Population Regression Function (PRF) in stochastic form as follows:

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + u_t$$

In the equation above, β_0 is the intercept term. As usual, it gives the mean or average effect on Y of all the variables excluded from the model, although its mechanical interpretation is the average value of Y when X_1 and X_2 are set equal to zero. The coefficients β_1 and β_2 are called the *partial regression coefficients*.

- β_1 measures the change in Y with respect to X_1, holding other factors (i.e., X_2 here) fixed, and β_2 measures the change in Y with respect to X_2, holding other factors fixed (i.e., X_2 here).

To find the OLS estimators, we will write the sample regression function (SRF) in its stochastic form as follows:

$$Y_t = \hat{\beta}_0 + \hat{\beta}_1 X_{t1} + \hat{\beta}_2 X_{t2} + \hat{u}_t$$

where $\hat{u}_t$ is the residual term. As we know from simple regression model, the OLS procedure consists of minimizing SSR: Symbolically:

$$\min \; SSR = \sum \hat{u}_t^2 = \sum \left(Y_t - \hat{\beta}_0 + \hat{\beta}_1 X_{t1} + \hat{\beta}_2 X_{t2} \right)^2 \qquad (0)$$

The procedure to get the estimators that will minimize the equation (0) is to differentiate it with respect to the unknowns, set the resulting expression to zero, and solve them simultaneously.

VI. Testing Hypotheses

In this section we discuss two types of hypothesis testing: (1) testing the statistical significance of individual coefficients, and (2) testing several regression coefficients jointly.

A. Testing Individual Coefficients

The steps for carrying out tests on an individual coefficient are as follows.

One-Tailed Test

Step 1. $H_0 : \beta_i = \beta_0$ versus $H_A : \beta_i > \beta_0$

Step 2. Construct the t-statistic $t_{\hat{\beta}_i} = \dfrac{\hat{\beta}_i}{se(\hat{\beta}_i)}$, where $\hat{\beta}_i$ is the estimate and $se(\hat{\beta}_i)$ is its standard error[2]. If $\beta_0 = 0$, this t-value

[2] It can be denoted by $\hat{\sigma}_{\hat{\beta}_i}$ or $s_{\hat{\beta}_i}$ instead of $se(\hat{\beta}_i)$.

reduces to the ratio of regression coefficient to its standard error. Under H_0, it has a *t*-distribution with *T-k-1* degrees of freedom, where *T* is total number of observations, *k* is the number of slope terms and *1* is for the intercept term in the regression.

Step 3. Look up in the *t*-table the entry corresponding to *T-k-1* degrees of freedom and find the critical point $t^*_{T-k-1}(\alpha)$ such that the area to the right of it is equal to the level of significance (α).

Step 4. Reject the null hypothesis if $t_{\hat{\beta}_i} > t^*_{T-k-1}(\alpha)$. If the alternative had been $H_A : \beta_i < \beta_0$, H_0 would have been rejected if $t_{\hat{\beta}_i} < -t^*_{T-k-1}(\alpha)$. Equivalently, for either alternative, reject H_0 if $\left|t_{\hat{\beta}_i}\right| > t^*_{T-k-1}(\alpha)$. Using *p*-value approach, reject H_0 if the *p*-value is less than the level of significance.

Two-Tailed Test

Step 1. $H_0 : \beta_i = \beta_0$ versus $H_A : \beta_i \neq \beta_0$

Step 2. Construct the same *t*-statistic $t_{\hat{\beta}_i} = \dfrac{\hat{\beta}_i}{se(\hat{\beta}_i)}$, where $\hat{\beta}_i$ is the estimate and $se(\hat{\beta}_i)$ is its standard error. Under H_0, it has a *t*-distribution with *T-k-1* degrees of freedom, where *T* is total

number of observations, *k* is the number of slope terms and *1* is for the intercept term in the regression.

Step 3. Look up in the *t*-table the entry corresponding to *T-k-1* degrees of freedom and find the critical point $t^*_{T-k-1}(\alpha/2)$ such that the area to the right of it is one-half the level of significance (α).

Step 4. Reject the null hypothesis if $\left|t_{\hat{\beta}_i}\right| > t^*_{T-k-1}(\alpha/2)$. Using *p*-value approach, reject H_0 if the *p*-value is less than the level of significance.

B. Testing Several Coefficients Jointly

As we know from simple regression model, it is also possible to test the joint significance of several regression coefficients.

Step 1. The null hypothesis is $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$. The alternative hypothesis is H_A: at least one of the β's is nonzero. The null hypothesis has thus *p=k* restrictions[3].

Step 2. Estimate the restricted model and unrestricted models and obtain the sum of squared residuals, SSR_R and SSR_U.

Step 3. Compute the Q-statistic using the following equation:

[3] We have *k* slope terms.

$$Q = \frac{(SSR_R - SSR_U)/p}{SSR_U/(T-k-1)}$$

or equivalently,

$$Q = \frac{SSR_R - SSR_U}{SSR_U} . \frac{T-k-1}{p}$$, where:

- SSR_R: *SSR* from restricted model
- SSR_U: *SSR* from unrestricted model
- p: number of restrictions in null hypothesis (number of β's eliminated from the unconstrained model). [4]
- *T-(k+1)*: degrees of freedom in the unrestricted model, where (k+1) parameters in unrestricted model.

Step 4. The Q statistic is distributed with $F^{\alpha}_{p,T-k-1}$, in other words, we can write that: $Q \sim F^{\alpha}_{p,T-k-1}$. Hence, from the F-table obtain the critical point $F^{\alpha}_{p,T-k-1}$ such that the area to the right is equal to the level of significance.

Step 5. Reject H_0 if $Q > F^{\alpha}_{p,T-k-1}$, or if the p-value is less than the level of significance.

[4] In fact, this is the degrees of freedom of a W variable like W= $SSR_R - SSR_U$. Note that if we denote the degrees of freedom for SSR_R and SSR_U by df_R (=*T-k-1+p*) and df_U (=*T-k-1*), then the degrees of freedom of W= $SSR_R - SSR_U$ is given by df_R - df_U which is equal to df_R - df_U = (*T-k-1+p*)-(*T-k-1*)=*p*. That is why we have written *p* for the numerator degrees of freedom in the formula.

VII. Confidence Intervals for β_k

Just as with the simple linear regression, confidence intervals for β_k are constructed using their sampling distribution and the properties of the normal distribution. Since we use an estimate of the variance, the confidence intervals are constructed using the *t*-distribution. The 100.(1-α)% confidence interval for $\hat{\beta}_j$ is as follows:

$$\hat{\beta}_j \pm t_{\alpha/2,df}.se(\hat{\beta}_j)$$

or,

$$P\left\{\hat{\beta}_j - t_{\alpha/2,df}.se(\hat{\beta}_j) \le \beta_j \le \hat{\beta}_j + t_{\alpha/2,df}.se(\hat{\beta}_j)\right\} = 1-\alpha$$

where the degrees of freedom parameter $df = T - k - 1$.

HAVE A NICE HOLIDAY! ☺