Test of Goodness of Fit and Correlation

About the components of an econometric model

Given the econometric model, $Y_i = \beta_1 + \beta_2 X_i + U_i$, where i= 1, 2, 3,...,n.

- A Simple Linear Regression model of two variables X_i and Y_i, for a sample of size n
- Y_i= dependent variable, X_i= independent or explanatory variable, U_i = error or disturbance or residual term,
- The parameters β_1 and β_2 are the coefficients of the regression. They are known as the intercept and slope coefficients respectively. The values of β_1 and β_2 are unknown. These coefficient are also called the Least square estimators.
- The objective of regression analysis is to estimate the unknowns, , on the basis of observed values of (X_i, Y_i) and also to draw inferences about the true β_1 and β_2
- The estimated coefficients [say $\beta_1 = \hat{\beta}_1$ (pronounced beta 1 hat) and $\beta_2 = \hat{\beta}_2$ (pronounced beta 2 hat)] help us to determine the estimated regression equation for a given sample.

***Coefficient of determination (R^2)**

Given Y_i = observed Y, $\hat{Y}_i$ = estimated Y and $\overline{Y}$ = mean of Y_i, the total variation in Y is equal to the explained variation in Y and residual variation in Y.

This means, Total variation in Y = Explained variation in Y + Residual variation in Y

i.e., $$\Sigma(Y_i - \overline{Y})^2 = \Sigma(\hat{Y}_i - \overline{Y})^2 + \Sigma(Y_i - \hat{Y}_i)^2$$

i.e., Total sum of squares (TSS) = Regression sum of squares (RSS) + Error sum of squares (ESS)

i.e., $$TSS = RSS + ESS$$

Dividing both sides by TSS gives

$$1 = \frac{RSS}{TSS} + \frac{ESS}{TSS}$$

i.e. we can write $$1 = R^2 + \frac{ESS}{TSS}$$

Hence we define, $$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}$$

R^2 is called the coefficient of determination and described as the proportion of explained variation in Y to total variation in Y, which means in the total variation in Y, it explains the estimated variation in Y.

R^2 is a measure of goodness of fit of the estimated regression line. The closer the observations in the data fall to the regression line, i.e., smaller the residuals, the greater is the variation in Y explained by estimated regression equation. It also helps to determine the correlation coefficient. We can calculate R^2 as follows:

$$R^2 = \frac{\Sigma(\hat{Y}_i - \overline{Y})^2}{\Sigma(Y_i - \overline{Y})^2} = 1 - \frac{\Sigma(Y_i - \hat{Y}_i)^2}{\Sigma(Y_i - \overline{Y})^2}$$

$$\text{i.e. } R^2 = 1 - \frac{\Sigma U_i^2}{\Sigma y_i^2}$$

Where $U_i = (Y_i - \hat{Y}_i)$ & $y_i = (Y_i - \overline{Y})$

***Properties of R^2**

1. The value of R^2 is nonnegative
2. The limits of the value of R^2 are as follows: $0 \leq R^2 \leq 1$. This means R^2 lies in the range 0 and 1. When the value is 1 (one), all the points lie on the regression line and the line is a perfect fit. When the value is 0 (zero), the estimated regression explains none of the variation in Y which means there is no relationship between the dependent and independent variable. In such a situation $\beta_2^{\wedge} = 0$.
3. R^2 is unit free.

***Correlation coefficient (r)**

The **correlation coefficient r** is computed from **R^2**. It is determined as follows:

$$r = \pm \sqrt{R^2}$$

By definition,

$$r = \frac{\text{cov}(x_i, y_i)}{\sigma_x \sigma_y} = \frac{\Sigma x_i y_i}{\sqrt{(\Sigma x_i^2)}\sqrt{(\Sigma y_i^2)}}$$

where, $x_i = (X_i - X)$ and $y_i = (Y_i - Y)$

***Properties of correlation coefficient r**

1. The range of values of r is given as $-1 \leq r \leq +1$
2. If $r < 0$ means X and Y move in opposite direction and $r > 0$ means X and Y change in the same direction. If r= 1, it implies a perfect positive correlation between the variables and if r = -1, it implies perfect negative correlation. Note that $r = \pm 1$ is rarely found.

3. A zero correlation coefficient means there exists no linear relationship between the variables X and Y which means values of X and Y may change without any connection.
4. The sign of r is the same sign as $\beta_2^{\wedge}$
5. It is symmetrical in nature. This means the correlation coefficient between X and Y, i.e. r_{XY} is the same as the correlation coefficient between Y and X, i.e. r_{YX}.
6. It is independent of the origin and scale.

***Difference between regression analysis and correlation analysis**

Regression analysis states the causal relationship between the independent variable X and the dependent variable Y. It implies a relation of dependence of Y on X though it does not prove it. On the other hand correlation analysis explains the degree of association between two variables involved and does not state anything about dependence. Thus it is claimed that correlation analysis is not as powerful tool as regression analysis in econometrics.

Determination of R^2 and r (using example)

Let us take the help of the example in the study material on **Simple Linear Regression**. Some of the important results from Simple Linear Regression were as follows:

1. $\hat{Y}_i = 1.2 + 0.969X_i$
2. $\beta_1^{\wedge} = 1.2$, $\beta_2^{\wedge} = 0.969$
3. $U_i = Y_i - \hat{Y}_i$
4. $x_i = (X_i - \bar{X})$ and $y_i = (Y_i - \bar{Y})$

Table 2.1: Determination of estimated Y_i and U_i

n	(X_i)	(Y_i)	x_i	x_i^2	y_i	y_i^2	x_iy_i	$\hat{Y}_i$	U_i	U_i^2
1	100	96	-100	10000	-99	9801	9900	98.1	-2.1	4.41
2	140	137	-60	3600	-58	3364	3480	136.86	0.14	0.0196
3	150	148	-50	2500	-47	2209	2350	146.55	1.45	2.1025
4	180	175	-20	400	-20	400	400	175.62	-0.62	0.3844
5	190	184	-10	100	-11	121	110	185.31	-1.31	1.7161
6	200	195	0	0	0	0	0	195	0	0
7	230	227	30	900	32	1024	960	224.07	2.93	8.5849
8	250	244	50	2500	49	2401	2450	243.45	0.55	0.3025
9	260	257	60	3600	62	3844	3720	253.14	3.86	14.8996
10	300	287	100	10000	92	8464	9200	291.9	-4.9	24.01
n =10	ΣX_i = 2000 $\bar{X}$ = 200	ΣY_i = 1950 $\bar{Y}$ = 195	Σx_i =0	Σx_i^2 = 33600	Σy_i = 0	Σy_i^2 = 31628	Σx_iy_i = 32570		ΣU_i =0	ΣU_i^2 =56.4296

Procedure for estimating $\hat{Y}_i$ (i=1,2,3,…,10)

Say when $X_i = 100$, putting this value in estimated regression equation in 1. We have,

$\hat{Y}_i = 1.2 + 0.969 \times 100 = 98.1$

In this manner we can find the values of all $\hat{Y}_i$ (i=1,2,3,…,10)

Procedure for estimating U_i (i=1,2,3,…,10)

From 2. We know $U_i = Y_i - \hat{Y}_i$

When $\hat{Y}_i = 98.1$, the corresponding value of $Y_i = 96$, therefore $U_i = 96 - 98.1 = -2.1$

In this manner we can find the values of all U_i (i=1,2,3,…,10). After which finding the value of U_i^2 is not difficult.

$$\text{We know } R^2 = 1 - \frac{\Sigma U_i^2}{\Sigma y_i^2}$$

$$\text{So in our example } R^2 = 1 - \frac{56.4296}{31628} = 1 - 0.00178 = 0.99822$$

We know $r = \pm \sqrt{R^2}$

Then here $r = \pm 0.999$

As $\hat{\beta}_2$ carries a positive sign, $r = 0.999$

Coefficient of Determination = $R^2 = 0.99822$
This implies there is a causal relationship between the dependent variable Y and independent variable X

Correlation Coefficient = r =0.999
This implies there is a positive correlation between the variables X and Y.

*N.B. Explanations of all the theories have been taken from the following references:

References:

Gujarati, D. N. Porter, D.C., Gunasekar, S. (2009), *Basic econometrics*. (Fifth ed.) McGraw-Hill Education (India).

Salvatore, D., & Reagle, D. (2011). *Schaum's outline of statistics and econometrics* (2nd ed.). McGraw-Hill Education.