Linear Regression: Correlations

Dr. Patrick Toche

Textbook:

James H. Stock and Mark W. Watson, *Introduction to Econometrics*, 4th Edition, Pearson.

Other references:

Joshua D. Angrist and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, 1st Edition, Princeton University Press.
Jeffrey M. Wooldridge, *Introductory Econometrics: A Modern Approach*, 7th Edition, Cengage Learning.

The textbook comes with online resources and study guides. Other references will be given from time to time.

Correlations

a. Show that the regression R^2 in the regression of Y on X is the squared value of the sample correlation between X and Y. That is, show that $R^2 = r_{XY}^2$.

Pearson's sample correlation coefficient r_{XY} (population is denoted ρ_{XY}) is

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

The OLS estimator may be written:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2}$$

where s_{XY} denotes the sample covariance and s_X^2 denotes the variance, $s_{XY} = s_X^2$. The proof of this is standard and follows from minimizing the sum of squared residuals of the regression of Y on X.

The coefficient of determination of the regression of Y on X is

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{s_{YY}}$$

Correlations

a. Show that the regression R^2 in the regression of Y on X is the squared value of the sample correlation between X and Y. That is, show that $R^2 = r_{XY}^2$.

Pearson's sample correlation coefficient r_{XY} (population is denoted ρ_{XY}) is

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

The OLS estimator may be written:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2}$$

where s_{XY} denotes the sample covariance and s_X^2 denotes the variance, $s_{XY} = s_X^2$. The proof of this is standard and follows from minimizing the sum of squared residuals of the regression of Y on X.

The coefficient of determination of the regression of Y on X is

$$R^2 = 1 - \frac{\sum_{i=1}^n \hat{u}_i^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{s_{YY}}$$

Correlations

a. Show that the regression R^2 in the regression of Y on X is the squared value of the sample correlation between X and Y. That is, show that $R^2 = r_{XY}^2$.

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \implies \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \sum_{i=1}^{n} [(Y_i - \overline{Y}) - \hat{\beta}_1 (X_i - \overline{X})]^2$$

$$= \sum_{i=1}^{n} (Y_i - \overline{Y})^2 - 2\hat{\beta}_1 \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}) + \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2$$

$$= s_{YY} - 2\hat{\beta}_1 s_{XY} + \hat{\beta}_1^2 s_{XX}$$

Plugging the above into the coefficient of determination:

$$R^2 = 1 - \frac{s_{YY} - 2\hat{\beta}_1 s_{XY} + \hat{\beta}_1^2 s_{XX}}{s_{YY}} = 2\hat{\beta}_1 \frac{s_{XY}}{s_{YY}} - \hat{\beta}_1^2 \frac{s_{XX}}{s_{YY}}$$

$$= 2\frac{s_{XY}}{s_X^2}\frac{s_{XY}}{s_{YY}} - \left(\frac{s_{XY}}{s_X^2}\right)^2 \frac{s_{XX}}{s_{YY}} = \frac{s_{XY}^2}{s_X^2\, s_Y^2} = r_{XY}^2$$

Correlations

a. Show that the regression R^2 in the regression of Y on X is the squared value of the sample correlation between X and Y. That is, show that $R^2 = r^2_{XY}$.

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \implies \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \sum_{i=1}^{n} [(Y_i - \overline{Y}) - \hat{\beta}_1 (X_i - \overline{X})]^2$$

$$= \sum_{i=1}^{n} (Y_i - \overline{Y})^2 - 2\hat{\beta}_1 \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}) + \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2$$

$$= s_{YY} - 2\hat{\beta}_1 s_{XY} + \hat{\beta}_1^2 s_{XX}$$

Plugging the above into the coefficient of determination:

$$R^2 = 1 - \frac{s_{YY} - 2\hat{\beta}_1 s_{XY} + \hat{\beta}_1^2 s_{XX}}{s_{YY}} = 2\hat{\beta}_1 \frac{s_{XY}}{s_{YY}} - \hat{\beta}_1^2 \frac{s_{XX}}{s_{YY}}$$

$$= 2 \frac{s_{XY}}{s_X^2} \frac{s_{XY}}{s_{YY}} - \left(\frac{s_{XY}}{s_X^2} \right)^2 \frac{s_{XX}}{s_{YY}} = \frac{s_{XY}^2}{s_X^2 \, s_Y^2} = r^2_{XY}$$

Correlations

b. Show that the R^2 from the regression of Y on X is the same as the R^2 from the regression of X on Y.

Since we have shown that $R^2 = r_{XY}^2$, it follows from $s_{XY} = s_{YX}$:

$$R^2 \text{ of } Y \text{ on } X = r_{XY}^2 = \frac{s_{XY}}{s_X s_Y} = \frac{s_{YX}}{s_Y s_X} = r_{YX}^2 = R^2 \text{ of } X \text{ on } Y$$

Correlations

b. Show that the R^2 from the regression of Y on X is the same as the R^2 from the regression of X on Y.

Since we have shown that $R^2 = r_{XY}^2$, it follows from $s_{XY} = s_{YX}$:

$$R^2 \text{ of } Y \text{ on } X = r_{XY}^2 = \frac{s_{XY}}{s_X s_Y} = \frac{s_{YX}}{s_Y s_X} = r_{YX}^2 = R^2 \text{ of } X \text{ on } Y$$

Correlations

c. Show that $\hat{\beta}_1 = r_{XY}(s_Y/s_X)$, where r_{XY} is the sample correlation between X and Y, and s_X and s_Y are the sample standard deviations of X and Y.

We've done most of the work already:

$$\hat{\beta}_1 = \frac{s_{XY}}{s_X^2} = \frac{r_{XY}s_Xs_Y}{s_X^2} = \frac{r_{XY}s_Y}{s_X}$$

Correlations

c. Show that $\hat{\beta}_1 = r_{XY}(s_Y/s_X)$, where r_{XY} is the sample correlation between X and Y, and s_X and s_Y are the sample standard deviations of X and Y.

We've done most of the work already:

$$\hat{\beta}_1 = \frac{s_{XY}}{s_X^2} = \frac{r_{XY}s_Xs_Y}{s_X^2} = \frac{r_{XY}s_Y}{s_X}$$