FINM3123 Introduction to Econometrics

Chapter 5 Exercises

Solutions

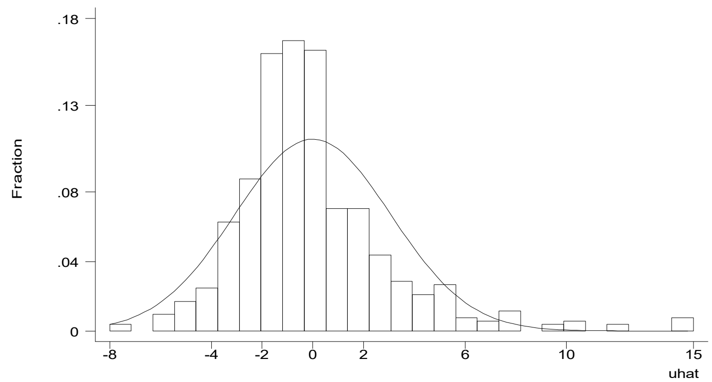
1. Write y = β_0 + $\beta_1 x_1 + u$, and take the expected value: E(y) = β_0 + β_1 E(x_1) + E(u), or μ_y = β_0 + $\beta_1 \mu_x$ since E(u) = 0, where μ_y = E(y) and μ_x = E(x_1). We can rewrite this as β_0 = μ_y – $\beta_1 \mu_x$. Now, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1$. Taking the plim of this we have plim($\hat{\beta}_0$) = plim($\bar{y} - \hat{\beta}_1 \bar{x}_1$) = plim($\bar{y}$) – plim($\hat{\beta}_1$)$\cdot$plim($\bar{x}_1$) = μ_y – $\beta_1 \mu_x$, where we use the fact that plim($\bar{y}$) = μ_y and plim($\bar{x}_1$) = μ_x by the law of large numbers, and plim($\hat{\beta}_1$) = β_1.

2. (a) The estimated equation is

$$\widehat{wage} = \underset{(0.73)}{-2.87} + \underset{(.051)}{.599}\, educ + \underset{(.012)}{.022}\, exper + \underset{(.022)}{.169}\, tenure$$

$n = 526, \quad R^2 = .306, \quad \hat{\sigma} = 3.085.$

Below is a histogram of the 526 residual, $\hat{u}_i$, i = 1, 2 , ..., 526. The histogram uses 27 bins, which is suggested by the formula in the Stata manual for 526 observations. For comparison, the normal distribution that provides the best fit to the histogram is also plotted.

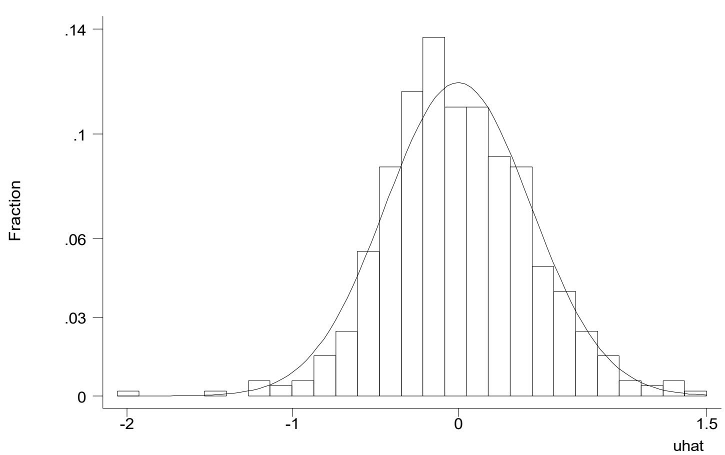


(b) With log(*wage*) as the dependent variable the estimated equation is

$$\widehat{\log(wage)} = \underset{(.104)}{.284} + \underset{(.007)}{.092}\, educ + \underset{(.0017)}{.0041}\, exper + \underset{(.003)}{.022}\, tenure$$

$n = 526, \quad R^2 = .316, \quad \hat{\sigma} = .441.$

The histogram for the residuals from this equation, with the best-fitting normal distribution overlaid, is given below:



(c) The residuals from the log(*wage*) regression appear to be more normally distributed. Certainly the histogram in part (ii) fits under its comparable normal density better than in part (i), and the histogram for the *wage* residuals is notably skewed to the left. In the *wage* regression there are some very large residuals (roughly equal to 15) that lie almost five estimated standard deviations ($\hat{\sigma}$ = 3.085) from the mean of the residuals, which is identically zero, of course. Residuals far from zero does not appear to be nearly as much of a problem in the log(*wage*) regression.

3.

(a) The measure of skewness for *inc* is about 1.86. When we use log(*inc*), the skewness measure is about .360. Therefore, there is much less skewness in log of income, which means *inc* is less likely to be normally distributed. (In fact, the skewness in income distributions is a well-documented fact across many countries and time periods.)

(b) The skewness for *bwght* is about −.60. When we use log(*bwght*), the skewness measure is about −2.95. In this case, there is much more skewness after taking the natural log.

(c) The example in part (ii) clearly shows that this statement cannot hold generally. It is possible to introduce skewness by taking the natural log. As an empirical matter, for many economic variables, particularly dollar values, taking the log often does help to reduce or eliminate skewness. But it does not *have* to.

(d) For the purposes of regression analysis, we should be studying the *conditional* distributions; that is, the distributions of *y* and log(*y*) conditional on the explanatory variables, $x_1, \ldots, x_k$. If we think the mean is linear, as in Assumptions MLR.1 and MLR.3, then this is equivalent to studying the distribution of the population error, *u*. In fact, the skewness measure studied in this question often is applied to the residuals from an OLS regression.