CHAPTER 5

SOLUTIONS TO PROBLEMS

5.1 Write $y = \beta_0 + \beta_1 x_1 + u$, and take the expected value: $E(y) = \beta_0 + \beta_1 E(x_1) + E(u)$, or $\mu_y = \beta_0 + \beta_1 \mu_x$ since $E(u) = 0$, where $\mu_y = E(y)$ and $\mu_x = E(x_1)$. We can rewrite this as $\beta_0 = \mu_y - \beta_1 \mu_x$. Now, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1$. Taking the plim of this we have $\text{plim}(\hat{\beta}_0) = \text{plim}(\bar{y} - \hat{\beta}_1 \bar{x}_1) = \text{plim}(\bar{y}) - \text{plim}(\hat{\beta}_1) \cdot \text{plim}(\bar{x}_1) = \mu_y - \beta_1 \mu_x$, where we use the fact that $\text{plim}(\bar{y}) = \mu_y$ and $\text{plim}(\bar{x}_1) = \mu_x$ by the law of large numbers, and $\text{plim}(\hat{\beta}_1) = \beta_1$. We have also used the parts of Property PLIM.2 from Appendix C.

5.3 The variable *cigs* has nothing close to a normal distribution in the population. Most people do not smoke, so *cigs* = 0 for over half of the population. A normally distributed random variable takes on no particular value with positive probability. Further, the distribution of *cigs* is skewed, whereas a normal random variable must be symmetric about its mean.

5.5 (i) Yes, the answer would be zero, because $z = \frac{x-\mu}{\sigma/\sqrt{n}} = \frac{100-72.60}{13.400/\sqrt{856}} = 59.82$ and $P(z > 59.82) = 1 - P(z < 59.82) = 1 - 1 = 0$. Half of the students have scores less than the average value of 72.60 and no student can score more than 100. Therefore, the answer contradict the assumption of a normal distribution for score.

(ii) By observing the histogram, we can say that very small proportion of the students have scored less than 60. No, the normal distribution does not fit well in the left tail.

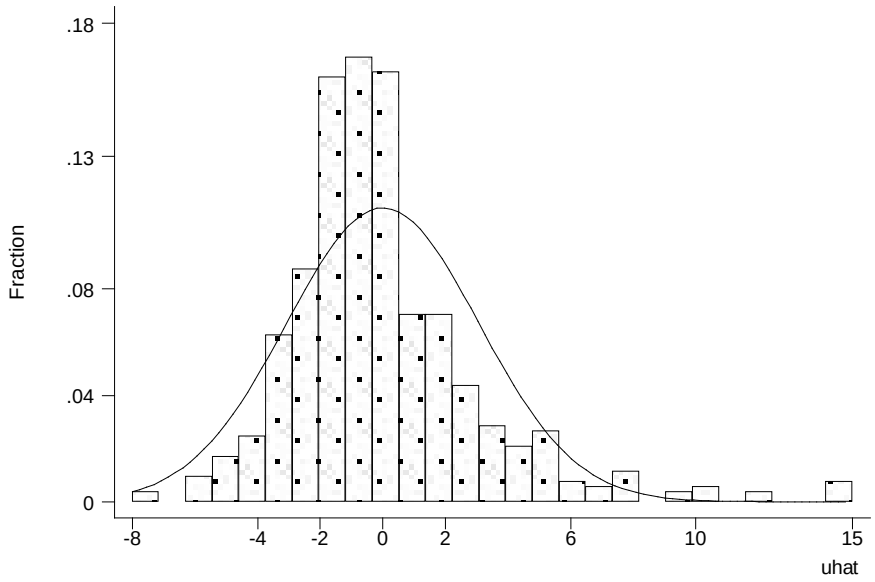
SOLUTIONS TO COMPUTER EXERCISES

C5.1 (i) The estimated equation is

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

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

Seen below is a histogram of the 526 residuals, $\hat{u}_i$, $i = 1, 2, ..., 526$. 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.

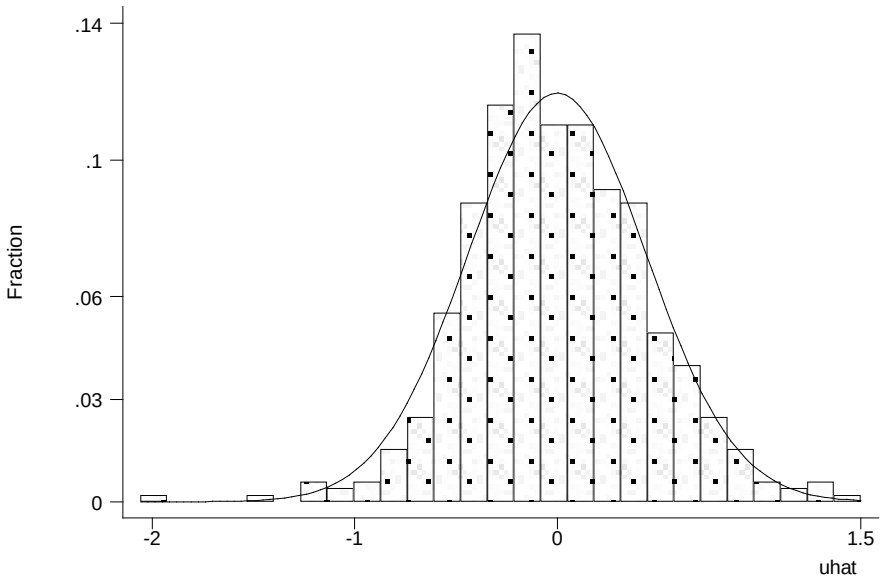


(ii) With log(*wage*) as the dependent variable, the estimated equation is

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

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

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

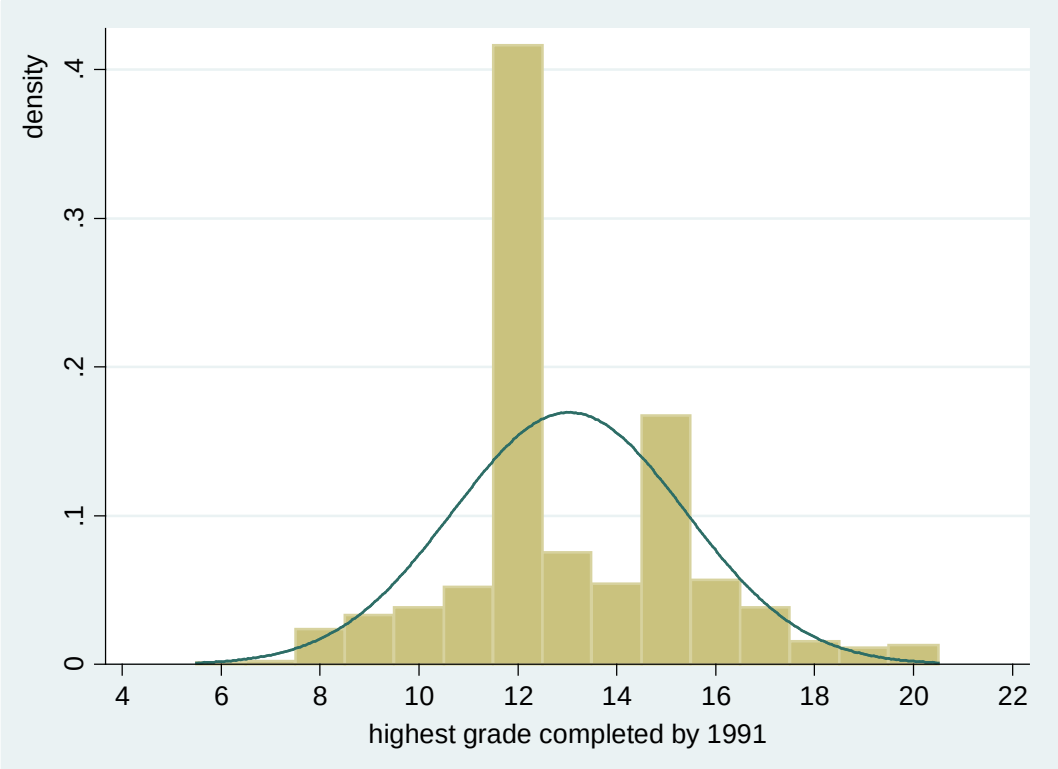


(iii) The residuals from the log(*wage*) regression appear to be more normally distributed. Certainly the histogram in part (ii) fits under its comparable normal density better than the one in part (i), and the histogram for the *wage* residuals is notably skewed to the left. In the *wage* regression, 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 do not appear to be nearly as much of a problem in the log(*wage*) regression.

C5.3 We first run the regression *bwght* on *cigs*, *parity*, and *faminc* using only the 1,191 observations with nonmissing observations on *motheduc* and *fatheduc*. After obtaining these residuals, $\tilde{u}_i$, these are regressed on $cigs_i$, $parity_i$, $faminc_i$, $motheduc_i$, and $fatheduc_i$, where, of course, we can only use the 1,191 observations with nonmissing values for both *motheduc* and *fatheduc*. The *R*-squared from this regression, R_u^2 is about .0024. With 1,191 observations, the chi-square statistic is (1,191)(.0024) $\approx$ 2.86. The *p*-value from the χ_2^2 distribution is about .239, which is very close to .242, the *p*-value for the comparable *F* test.

C5.5 (i) The variable *educ* takes on all integer values from 6 to 20, inclusive. So it takes on 15 distinct values. It is not a continuous random variable, nor does it make sense to think of it as approximately continuous. (Contrast a variable such as hourly wage, which is rounded to two decimal places but takes on so many different values it makes sense to think of it as continuous.)

(ii) With a discrete variable, usually, a histogram has bars centered at each outcome, with the height being the fraction of observations taking on the value. Such a histogram, with a normal distribution overlay, is given below.



Even discounting the discreteness, the best fitting normal distribution (matching the sample mean and variance) fits poorly. The focal point at *educ* = 12 clearly violates the notion of a smooth bell-shaped density.

(iii) Given the findings in part (iii), the error term in the equation

$$educ = \beta_0 + \beta_1 motheduc + \beta_2 fatheduc + \beta_3 abil + \beta_4 abil^2 + u$$

cannot have a normal distribution independent of the explanatory variables. Thus, MLR.6 is violated. In fact, the inequality $educ \geq 0$ means that *u* is not even free to vary over all values given *motheduc*, *fatheduc*, and *abil*. (It is likely that the homoskedasticity assumption fails, too, but this is less clear and does not follow from the nature of *educ*.)

The violation of MLR.6 means that we cannot perform exact statistical inference; we must rely on asymptotic analysis. This in itself does not change how we perform statistical inference: without normality, we use exactly the same methods, but we must be aware that our inference holds only approximately.