SIMPLE LINEAR REGRESSION OF CPS DATA

Using the 1995 CPS data, hourly wages are regressed against years of education. The regression output in Table 4.1 indicates that there are 1003 persons in the CPS sample, allowing a check that the data are intact. The coefficient of variation is R^2=.1708, indicating that about 17% of the variation in hourly wages is due to differences in years of education across wage earners. This is probably lower than expected. It also means that about 83% of the variation is due to other factors, possibly some that were measured by the CPS.

A plot of wages against years of education shows remarkable concordance between the fitted regression line and mean wages at each education level, except at the extremes (Figure 4.1). In fact, tracking mean wages is the primary intent of regression analysis. When mean wages show a linear trend, these two curves will be close. Otherwise, a nonlinear regression curve or a transformation of the data may be required. The plot shows that wages may be higher than predicted on the extremes. It may be thus necessary to modify the initial model, in light of possible nonlinearity and also skewness of the error distribution for each education level.

A test of association between wages and education is performed using the *t*-test for the slope or regression coefficient. For this data, t=14.36 with an observed level of significance or p-value of p<.0001. This means if there were no relationship between wages and education, the chances of obtaining a slope this large or larger are less than 1 in 10,000. This is as rare an event as can be hoped for so there is strong evidence of a positive relationship between education and wages. The relationship between wages and education and is called statistically significant.

The regression coefficient ($\pm$ standard error) for years of education is \$1.47$\pm$ 0.10 with a 95% confidence interval of (\$1.27, \$1.67). This means that the best single estimate for the true increase in average hourly wages for each additional year of education is \$1.47. Because this is based on a random sample and different random samples would give different estimates of return to education, there is uncertainty associated with the point estimate. This uncertainty is reflected in the standard error and thus in the confidence interval. An interval estimate, or a range of values where the true value could lie, is computed based on the underlying variability of wages within each education level and the size of the sample, and is called a confidence interval. For these data, the 95% confidence interval ranges from \$1.27 to \$1.67. The lower end of the interval is far enough from zero to provide confidence that not only more education is associated with higher wages but that the actual dividend to education is between about \$1.30 and \$1.70 per hour for every additional year of education.

The regression output gives other information such as an estimate of the intercept, the root mean square error, and the Analysis of Variance (ANOVA) table, that are often less useful for the practical interpretation of the analysis. For example, the intercept is meaningless in this context. It estimates hourly wages for someone with no education as –\$4.96 per hour, which is nonsensical. Since there are no observations in this region, such an estimate is an extrapolation of the data. For those with less than primary school education, the curve may have a shallower slope than the one that fits the rest of the data. There may also be reporting (measurement) error associated with the lower end of education, since it is difficult not to have attended primary school these days.

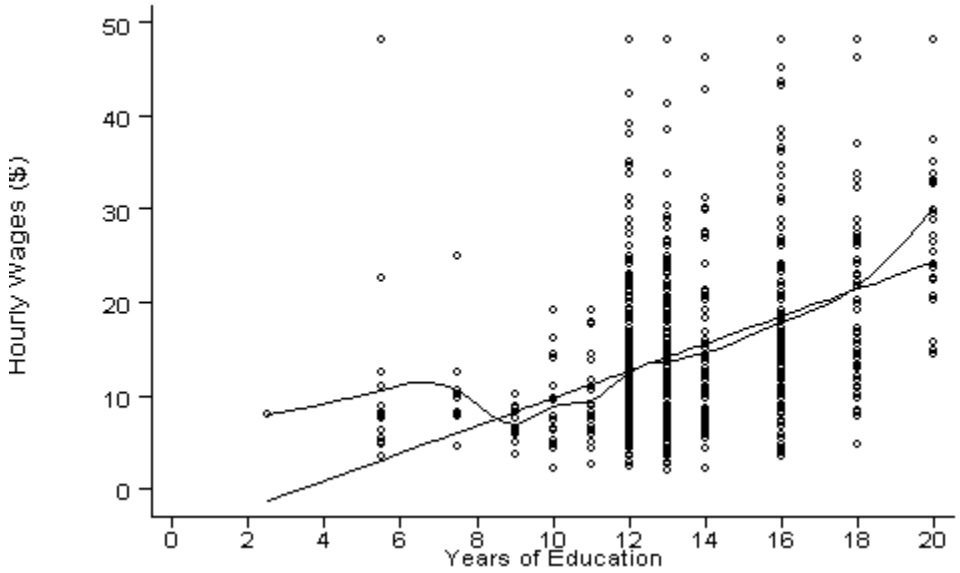

Figure 4.1. Plot of hourly wages against years of education, superimposing the fitted regression line.

Log Transformations

The plot of hourly wages against years of education (Figure 4.1) shows that wages are not symmetrically distributed for a given number of years of education. In fact, they are right skewed, so that for a given number of years of education, a few people have very high wages but most are lower than the average. Log transforming hourly wages results in a plot where wages are generally symmetrically distributed within each education level so as to satisfy the linear model assumptions but where the variable of interest, log wages, is measured in log dollars and is not as easily interpretable (Figure 4.2).

Log hourly wages are now regressed against years of education to obtain the results in Table 4.2 and the fitted regression line in Figure 4.2. Again, there is general concordance between the fitted regression line and mean log wages for every education level, except at the extremes. The coefficient of variation is R^2=.1620, indicating that about 16% of the variation in log hourly wages is due to differences in years of education across wage earners, not very different from before.

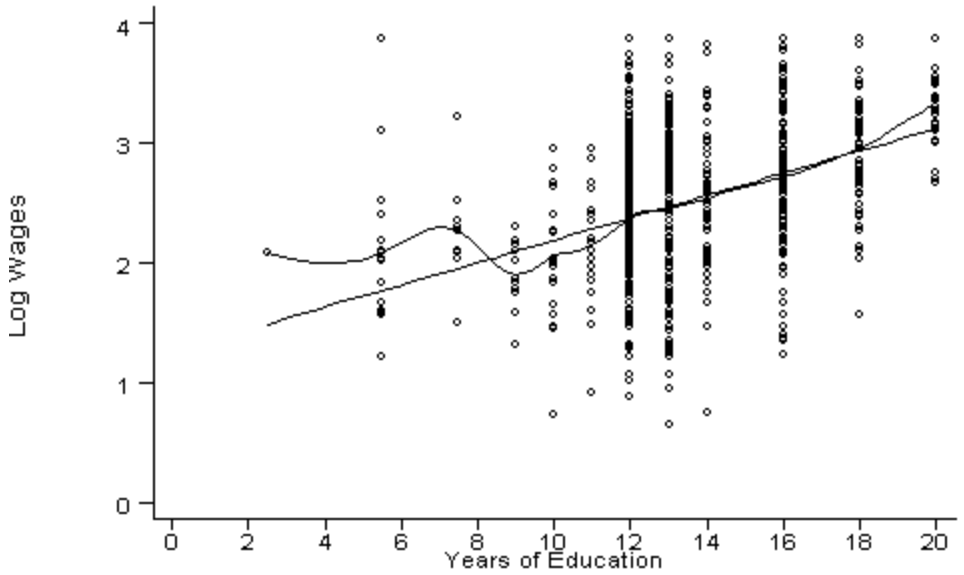

Figure 4.2 Plot of log hourly wages against years of education, superimposing the fitted regression line.

A test of association between hourly wages and education is performed using a *t*-test for the regression coefficient. Since the log transformation is a monotone function, if hourly wages increase according to years of education, the same will be true for any monotone function, such as log hourly wages, and conversely. In this case, t=13.91 (p<.0001) so there is strong evidence of a positive relationship between education and log wages, and therefore also between education and wages. Often, the *t*-statistic and observed level of significance are not the same on the two scales, since the data may be more dispersed on one scale; for these data, they happen to be similar. For scales that are monotone transformations of each other, the direction of effect, either positive or negative, is the same on both scales, so that inference made on one scale can be applied to the other scale.

The regression coefficient ($\pm$ standard error) for years of education is 0.0933 ± 0.0067 log dollars with a 95% confidence interval of (0.0801, 0.1064). This is meaningless to most people but it can be converted to the effect on the original scale. Using the properties of logarithmic and exponential functions, average hourly wages increase by a factor of exp(0.0933)=1.10 or 10% for each additional year of education. A 95% confidence interval for the proportional rate of increase is obtained by exponentiating the upper and lower bounds of the 95% confidence interval for the slope as (exp(0.0801), exp(0.1064)) = (1.08, 1.11). This gives the best interval estimate for the rate of return to wages for every additional year of education. The lower bound is far enough from 1.00 (no return) that not only there is evidence of a positive return to education, but its effect is on the order of 8 to 11% increase in wages for every additional year of schooling. Using a log transformed dependent variable thus requires a different interpretation – proportional or relative rather than additive – for the regression coefficients.

The effect on wages of four years of college or four years of high school can also be analyzed. Average wages increase by a factor of about exp(4*0.0933)= 1.45, with a 95% confidence interval of (exp(4*0.0801), exp(4*0.1064))= (1.38, 1.53) for every additional four years of education. This means that high school or college graduates' wages are an average of 45% (38%, 53%) higher than they would be without this education.

Since log dollars are not easily understandable, the original scale is used for interpretation. When plotted on the original scale, the fitted line from a log-transformed regression is no longer straight but curved (Figure 4.3). The curve indicates that the absolute return to wages from education increases exponentially. The relative rate of return to wages is constant, about 10% per year. The absolute return to education, however, is small for those with less than high school education but is much larger for those with college and post-graduate education. Thus a constant relative rate of return translates to a higher absolute return when wages are higher. Actuaries and those who deal in the stock market are well aware of this effect, called compounding. The 10% relative return to wages per year is in fact being compounded over years of education. For comparison, the fitted line from the usual linear regression is also plotted in Figure 4.3. Note that the curve from the log-transformed regression gives a better fit in the extremes.

Finally, from the intercept of the log regression, an estimate of hourly wages for those with no education is exp(1.2599) = $3.52, a more reasonable estimate of wages earned by those with little education than the negative number obtained from regression on the original scale.

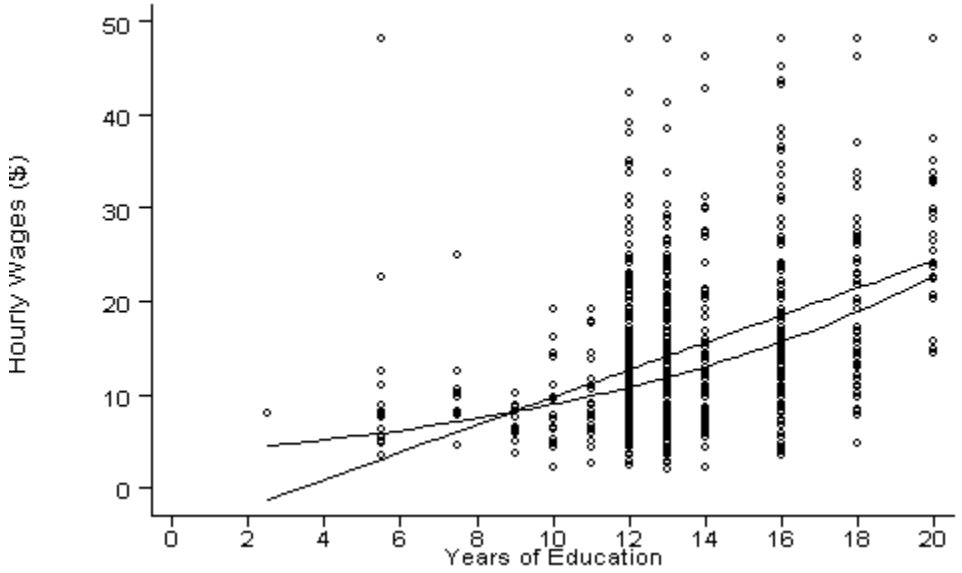

Figure 4.3. Plot of hourly wages against years of education, superimposing the fitted lines from the original (straight line) and log-transformed (curved, exponential line) regressions of wages on education.

The analysis of a log-transformed dependent variable has a more complex interpretation than a non-transformed dependent variable. Further analyses will use log-transformed wages as the dependent variable since the assumptions required to fit a linear model are better satisfied on this scale. Chapter 5 will introduce multiple regression to show how factors other than education contribute to differences in wages.

Table 4.1. Linear regression of hourly wages against years of education.

```
  Source |       SS       df       MS              Number of obs =    1003
---------+------------------------------           F(  1,  1001) =  206.20
   Model |  14666.8401     1  14666.8401           Prob > F      =  0.0000
Residual |  71201.2121  1001   71.130082           R-squared     =  0.1708
---------+------------------------------           Adj R-squared =  0.1700
   Total |  85868.0522  1002  85.6966589           Root MSE      =  8.4339

------------------------------------------------------------------------------
    wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
    educ |   1.466404   .1021203    14.360   0.000       1.26601    1.666799
   _cons |  -4.956578   1.399311    -3.542   0.000      -7.702497   -2.210659
------------------------------------------------------------------------------
```

Table 4.2. Linear regression of log hourly wages against years of education.

```
  Source |       SS       df       MS              Number of obs =    1003
---------+------------------------------           F(  1,  1001) =  193.52
   Model |  59.3347547     1  59.3347547           Prob > F      =  0.0000
Residual |   306.91002  1001  .306603416           R-squared     =  0.1620
---------+------------------------------           Adj R-squared =  0.1612
   Total |  366.244774  1002  .365513747           Root MSE      =  .55372

------------------------------------------------------------------------------
  lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
    educ |   .0932696   .0067046    13.911   0.000       .0801129    .1064263
   _cons |   1.259661   .0918705    13.711   0.000       1.079381    1.439942
------------------------------------------------------------------------------
```