James Green-Armytage
Econ 229, fall 2015

OLS Regressions with One Independent Variable

1. Setup

A measured variable y is determined by another measured variable x, and random error which we denote as ε. Suppose that the true relationship between the variables can be written as

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

However, we can't observe α, β, or the ε_is directly; instead, we have n observations of y_i values along with corresponding x_i values. Our task is to estimate α and β; we write our estimates of α and β as a and b, respectively. Then, we define $\hat{y}_i = a + bx_i$ as the predicted value of y for each x_i according to our model, and $e_i = y_i - \hat{y}_i$ as the 'residual', i.e. the difference between the observed value and the predicted value. Thus, our estimated model is

$$y_i = a + bx_i + e_i$$

The Ordinary Least Squares (OLS) estimate is designed to minimize the sum of squared residuals (SSR). That is, we will implement OLS by choosing the values of a and b that minimize

$$SSR = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - a - bx_i)^2$$

(Note: all summation signs in this document indicate summation from $i = 1$ to $i = n$, so we will omit this in the notation for visual clarity.)

2. Estimate of the intercept term

To find an expression for the SSR-minimizing value of a, we set the partial derivative of SSR with respect to a equal to zero, and solve for a:

$$\frac{\partial SSR}{\partial a} = \sum 2(y_i - a - bx_i)(-1) \overset{\text{set}}{=} 0$$

$$\sum (y_i - a - bx_i) = 0$$

$$\sum a = \sum y_i - b \sum x_i$$

$$na = \sum y_i - b \sum x_i$$

$$\boxed{a = \bar{y} - b\bar{x}}$$

Here, $\bar{x} = \frac{1}{n}\sum x_i$ and $\bar{y} = \frac{1}{n}\sum y_i$ are the average values of x and y, respectively.

3. Estimate of the slope term

Define $X_i \equiv x_i - \bar{x}$, and $Y_i \equiv y_i - \bar{y}$ as the 'demeaned' versions of the x_is and y_is. We can now write the residual e_i as $e_i = (Y_i + \bar{y}) - a - b(X_i + \bar{x}) = Y_i - bX_i - (a - \bar{y} + b\bar{x})$. Given that we have $a = \bar{y} - b\bar{x}$ (from the previous section), this simplifies to $e_i = Y_i - bX_i$, in which case we can write $SSR = \sum(Y_i - bX_i)^2$.

Next, we take the partial derivate of SSR with respect to b, and solve for b in terms of the data.

$$\frac{\partial SSR}{\partial b} = \sum 2(Y_i - bX_i)(-X_i) \overset{\text{set}}{=} 0$$

$$\sum(-X_iY_i + bX_i^2) = 0$$

$$b\sum X_i^2 = \sum X_iY_i$$

$$b = \frac{\sum X_iY_i}{\sum X_i^2}$$

$$\boxed{b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}}$$

4. Variance of the slope term

Researchers are often interested in determining the probability with which the slope term in the true data-generating process $y_i = \alpha + \beta x_i + \varepsilon_i$ has the sign indicated by the estimated slope term, b. To estimate that probability, we must first estimate the variance of b, as follows:

$$\begin{aligned}
\text{var}(b) = \text{var}\left(\frac{\sum X_iY_i}{\sum X_i^2}\right) &= \text{var}\left(\frac{1}{\sum X_i^2}\sum X_iY_i\right) = \left(\frac{1}{\sum X_i^2}\right)^2 \text{var}\left(\sum X_iY_i\right) \\
&= \left(\frac{1}{\sum X_i^2}\right)^2 \text{var}(X_1Y_1 + \cdots + X_nY_n) = \left(\frac{1}{\sum X_i^2}\right)^2 [\text{var}(X_1Y_1) + \cdots + \text{var}(X_nY_n)] \\
&= \left(\frac{1}{\sum X_i^2}\right)^2 [X_1^2\text{var}(Y_1) + \cdots + X_n^2\text{var}(Y_n)] = \left(\frac{1}{\sum X_i^2}\right)^2 [X_1^2\sigma^2 + \cdots + X_n^2\sigma^2] \\
&= \left(\frac{1}{\sum X_i^2}\right)^2 \sigma^2 \sum X_i^2 = \frac{\sigma^2}{\sum X_i^2} = \boxed{\frac{\sigma^2}{\sum(x_i - \bar{x})^2}}
\end{aligned}$$

In the derivation above, we proceed by treating the X_is as constants, possessing zero variance. On the other hand, the Y_is possess variance because they include the random error terms ε_i. That is, we assume that $\text{var}(X_i) = 0$, and $\text{var}(Y_i) = \text{var}(\beta X_i + \varepsilon_i) = \text{var}(\varepsilon_i)$. We define $\sigma^2 \equiv \text{var}(\varepsilon_i)$ as the variance of the error term, and assume that it is the same for all observations.

We cannot observe σ^2 directly, but we can estimate it based on the data as

$$\boxed{s^2 = \frac{1}{n-2}\sum e_i^2}$$

Substituting s^2 for σ^2 in the expression for $\text{var}(b)$, we obtain the estimated variance of b:

$$\boxed{\widehat{\text{var}}(b) = \frac{s^2}{\sum(x_i - \bar{x})^2} = \frac{\left(\frac{\sum e_i^2}{(n-2)}\right)}{\sum(x_i - \bar{x})^2}}$$

The standard error of b is the estimated standard deviation of b, i.e. the square root of $\widehat{\text{var}}(b)$:

$$\boxed{SE(b) = \sqrt{\widehat{\text{var}}(b)} = \sqrt{\frac{s^2}{\sum(x_i - \bar{x})^2}} = \sqrt{\frac{\left(\frac{\sum e_i^2}{(n-2)}\right)}{\sum(x_i - \bar{x})^2}}}$$

5. Confidence intervals

As n increases, the distribution of b converges toward a normal distribution; that is, b is 'asymptotically normal'. Also, the expected value of b is β; that is, b is a 'consistent' estimator. (Proofs of these two propositions can be found elsewhere.)

Taking these two facts together with our discussion of the variance of b, we believe that (with a sufficiently large sample size) b is distributed approximately normally, with mean β and standard deviation $SE(b)$.

Therefore, with probability $\approx k$, b has fallen in the interval $\beta \pm t^* \cdot SE(b)$, where $t^* = \Phi^{-1}\left(\frac{k+1}{2}\right)$, and $\Phi^{-1}(F)$ is the inverse function of the standard normal distribution. We can find t^* in Excel using the code `tstar = normsinv( (k + 1) / 2 )`.

Therefore we can say that with confidence $\approx k$, β is located in the interval $\boxed{b \pm t^* \cdot SE(b)}$.

6. Hypothesis testing

To argue that the sign of b is 'statistically significant', we need to reject the null hypothesis that $\beta = 0$. So we consider the following question: If β is zero, what is the probability that we observe b being as far from zero as it is?

We define the 't-statistic' as the number of standard errors that separate b from zero, and thus the number of standard deviations separating b from its own mean if the null hypothesis is true:

$$\boxed{t = \frac{b}{SE(b)}}$$

The probability of observing a t-statistic with absolute value $|t|$ or greater, under the null hypothesis that $\beta = 0$, can be approximated by

$$\boxed{p = 2\Phi(-|t|)}$$

In Excel, we can calculate this via `p = 2 * normsdist( -abs(t) )`. We call this the 'p-value'; the closer it is to zero, the less plausible is the null hypothesis, and thus the more convincing is the alternative hypothesis that the true sign of β is equivalent to the sign of b.

7. Refinement: Student's t-distribution

Stata models b as following a Student's t-distribution, which is similar to a normal distribution but not identical. With large samples, the two are functionally equivalent. With smaller samples, the use of a t-distribution leads to slightly wider confidence intervals and slightly higher p-values (and thus, slightly more conservative results overall).

For use in confidence intervals, we can generate t^* values according to a t-distribution with $n - 2$ 'degrees of freedom' with the Excel code `tstar = tinv( 1 - k , n - 2 )`.

We can generate p-values according to a t-distribution using the Excel code `p = tdist( abs( t ) , n - 2 , 2 )`; here we input "2" for the last argument to indicate a 'two-tailed test'.