Section 3 Inference in Simple Regression

Having derived the (asymptotic or small-sample, depending on the assumptions) probability distribution of the OLS coefficients, we are now in a position to make inferential statements about the population parameters: hypothesis tests and confidence intervals.

Hypothesis tests about single coefficients

- The most common test in econometrics is the "t-test" of the hypothesis that a single coefficient equals zero. (But remember that Stata calculates standard errors based on classical assumptions.)
  - This test is printed out for each regression coefficient in Stata and other statistical packages.
  - Depending on the assumptions of the model (and whether they are valid), the "t-statistic" may or may not follow Student's *t* distribution.
- General form for calculating a t-statistic is $t = \frac{\hat{\beta} - \bar{\beta}}{\text{s.e.}\left(\hat{\beta}\right)}$, where $\bar{\beta}$ is the hypothetical value (usually zero) that we are testing against and s.e. is the standard error of the estimated coefficient $\hat{\beta}$.
- Using the t-statistic to test H_0: $\beta = \bar{\beta}$ against the two-sided alternative H_1: $\beta \neq \bar{\beta}$.
  - Note that hypothesis to be tested is always expressed in terms of the *actual* coefficient, not the estimated one.
  - Use the formula above to calculate the *t* statistic. (Stata will print out $\hat{\beta}$ and its standard error, and also the *t* value corresponding to $\bar{\beta}$ = 0.)
  - Compute the probability (*p*) value associated with the test: the probability that an outcome at least this inconsistent with the null hypothesis would occur if the null is indeed true.
    - $$p = \Pr_{H_0}\left[\left|\hat{\beta}_1 - \bar{\beta}_1\right| > \left|\hat{\beta}_1^{act} - \bar{\beta}_1\right|\right] = \Pr_{H_0}\left[\left|\frac{\hat{\beta}_1 - \bar{\beta}_1}{s.e.\left(\hat{\beta}_1\right)}\right| > \frac{\left|\hat{\beta}_1^{act} - \bar{\beta}_1\right|}{s.e.\left(\hat{\beta}_1\right)}\right] = \Pr_{H_0}\left(\left|t\right| > \left|t^{act}\right|\right).$$
    - If we know the distribution of the *t* statistic, then we can calculate the past probability from tables.

        - Under S&W's assumptions, the t statistic will be asymptotically normal.
        - Under the classical assumptions (including exogenous regressors and scalar error covariance), t statistic follows the t distribution with $n - 2$ degrees of freedom.
        - Stata calculates the p value associated with the null hypothesis $\beta = 0$.
    - Show diagram corresponding to S&W's Figure 5.1 on page 153: For given $|t|$, show how to calculate p value.
    - On same diagram show *critical values* for test at given level of significance, and how to decide the result of the test
        - Note 1.96 as two-tailed 5% critical value for normal distribution.
    - Then show the symmetry: the p value is the smallest significance level at which the null hypothesis can be rejected.
- One-tailed test such as H_0: $\beta_1 = \bar{\beta}$, H_1: $\beta_1 < \bar{\beta}$. (Or H_0: $\beta_1 \geq \bar{\beta}$)
    - Same basic procedure, but in this case we concentrate the entire rejection region in one tail of the distribution.
        - We reject the null if and only if $\Pr[t < t^{act}]$ < critical value (ignoring left tail of distribution) and fail to reject for any positive t value no matter how large.
        - Other direction if H_1 is $\beta_1 > \bar{\beta}$: Fail to reject null for any negative value of t and reject when $\Pr[t < t^{act}]$ > critical value.
- Present some examples of regressions and practice with tests of $\beta = 0$ and β = other values.

Confidence intervals for individual coefficients

- What is a confidence interval?
    - We shouldn't say "The probability is 95% that the true value of the coefficient lies in this interval." The coefficient is a constant which either does or does not lie in the interval with probability one.
    - We can say "In 95% of random samples the computed confidence interval will contain the true value of the coefficient."
- Using 1.96 as the 2.5% cutoff for each tail of the Gaussian distribution, the confidence interval for a coefficient estimate is $\left(\hat{\beta} - 1.96 \cdot s.e.\left(\hat{\beta}\right),\ \ \hat{\beta} + 1.96 \cdot s.e.\left(\hat{\beta}\right)\right)$.

Dummy (binary or indicator) independent variables

- Dummy variables are (yes, no) variables. We traditionally give the value 1 to yes and 0 to no.
- While dummy variables are often very useful in multiple regressions (with more than one regressor), they are limited in simple regression, but have a special interpretation.
    - Suppose that D is a dummy variable for sex with $D = 1$ being male.
    - Consider the model $Y_i = \beta_0 + \beta_1 D_i + u_i$. For females, $D = 0$ and the expected value of Y is β_0. For males, $D = 1$ and the expected value of Y is $\beta_0 + \beta_1$. Thus, β_1 is the difference between the expected Y for males and females.
    - A test of the null hypothesis $\beta_1 = 0$ would be a test of whether males and females have the same average Y.
        - This is equivalent to the *t* test for the equality of means and is a simple application of "analysis of variance."
- We shall encounter dummy variables frequently in multiple regression, where we want to allow the regression line to be shifted upward or downward for the set of observations with $D = 1$ relative to the rest of the sample.
- If we have a dummy variable that is 1 for only a single observation (presumably in multiple regression), then the residual for that observation will be zero and the coefficient of that dummy variable will have the value of the residual of that observation in an otherwise identical regression that excludes the dummy.

Asymptotic properties of OLS bivariate regression estimator

(Based on S&W, Chapter 17.)

- **Convergence in probability (probability limits)**
    - Assume that $S_1, S_2, \ldots, S_n, \ldots$ is a sequence of random variables.
        - In practice, they are going to be estimators based on 1, 2, …, n observations.
    - $S_n \xrightarrow{p} \mu$ if and only if $\lim_{n\to\infty} \Pr\left[\left|S_n - \mu\right| \geq \delta\right] = 0$ for any $\delta > 0$. Thus, for any small value of δ, we can make the probability that S_n is further from μ than δ arbitrarily small by choosing n large enough.
    - If $S_n \xrightarrow{p} \mu$, then we can write plim $S_n = \mu$.
    - This means that the entire probability distribution of S_n converges on the value μ as n gets large.
    - Estimators that converge in probability to the true parameter value are called **consistent estimators**.
- **Convergence in distribution**

- If the sequence of random variables $\{S_n\}$ has cumulative probability distributions $F_1, F_2, \ldots, F_n, \ldots$, then $S_n \xrightarrow{d} S$ if and only if $\lim_{n\to\infty} F_n(t) = F(t)$, for all t at which F is continuous.
- If a sequence of random variables converges in distribution to the normal distribution, it is called **asymptotically normal**.

- Properties of probability limits and convergence in distribution
  - Probability limits are very forgiving: Slutsky's Theorem states that
    - plim $(S_n + R_n)$ = plim S_n + plim R_n
    - plim $(S_n R_n)$ = plim $S_n \cdot$ plim R_n
    - plim (S_n / R_n) = plim S_n / plim R_n
  - The continuous-mapping theorem gives us
    - For continuous functions g, plim $g(S_n) = g$(plim S_n)
    - And if $S_n \xrightarrow{d} S$, then $g(S_n) \xrightarrow{d} g(S)$.
  - Further, we can combine probability limits and convergence in distribution to get
    - If plim $a_n = a$ and $S_n \xrightarrow{d} S$, then
      - $a_n S_n \xrightarrow{d} aS$
      - $a_n \pm S_n \xrightarrow{d} a \pm S$
      - $S_n / a_n \xrightarrow{d} S / a$
    - These are *very* useful since it means that asymptotically we can treat any consistent estimator as a constant equal to the true value.
- **Central limit theorems**
  - There is a variety with slightly different conditions.
  - Basic result: If $\{S_n\}$ is a sequence of estimators of μ, then for a wide variety of underlying distributions, $\sqrt{n}(S_n - \mu) \xrightarrow{d} N(0, \sigma^2)$, where σ^2 is the variance of the underlying statistic.
- Applying asymptotic theory to the OLS model
  - Under the more general conditions (including, specifically, the finite kurtosis assumption), the OLS estimator satisfies the conditions for consistency and asymptotic normality.
  - $\sqrt{n}\left(\hat{\beta}_1 - \beta_1\right) \xrightarrow{d} N\left(0, \frac{\text{var}\left[\left(X_i - E(X)\right)u_i\right]}{\left[\text{var}\left(X_i\right)\right]^2}\right)$.
  - plim $\hat{\sigma}^2_{\hat{\beta}_1} = \sigma^2_{\hat{\beta}_1}$, as proven in Section 17.3.
  - $t = \frac{\hat{\beta}_1 - \bar{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \xrightarrow{d} N(0,1)$.
- Choice for t statistic:
  - If homoskedastic, normal error term, then exact distribution is t_{n-2}.

- If heteroskedastic or non-normal error (with finite 4th moment), then exact distribution is unknown, but asymptotic distribution is normal
- Which is more reasonable for any given application?

Weighted least squares

WLS is an example of a more general strategy known as generalized least squares (GLS). The idea in GLS is to transform the model into one that has a classical error term so that OLS on the transformed equation is BLUE.

- If var(u_i | X_i) is known:
  - Suppose that $\text{var}(u_i \mid X_i) = \lambda h(X_i)$, so that we know the variances up to a constant of proportionality. h is known but λ is not.
    - Note that the pattern could depend on some variable other than X.
  - Transform model into

    $$\tilde{Y}_i \equiv \frac{Y_i}{\sqrt{h(X_i)}},$$
    $$\tilde{X}_{0,i} \equiv \frac{1}{\sqrt{h(X_i)}},$$
    $$\tilde{X}_i \equiv \frac{X_i}{\sqrt{h(X_i)}},$$
    $$\tilde{u}_i \equiv \frac{u_i}{\sqrt{h(X_i)}}.$$
  - $$\text{var}(\tilde{u}_i \mid X_i) = \frac{1}{\left(\sqrt{h(X_i)}\right)^2}\text{var}(u_i \mid X_i) = \frac{1}{h(X_i)}\lambda h(X_i) = \lambda.$$
  - Thus, the transformed system $\tilde{Y}_i = \beta_0 \tilde{X}_{0,i} + \beta_1 \tilde{X}_i + \tilde{u}_i$ has a homoskedastic error and OLS will be BLUE if the other classical assumptions are satisfied.
  - This is difficult because $h(X_i)$ is rarely known.
- If var(u_i | X_i) is a function of known variables:
  - Suppose, for example, that we know that $\text{var}(u_i \mid X_i) = \theta_0 + \theta_1 X_i^2$, but we don't know the θ parameters.
  - If we can get a consistent estimate of the θs, then we can calculate consistent estimates of the variance and use the WLS transformation with estimated variances. Thanks to Slutsky's Theorem, this is good enough for the resulting WLS estimators to be consistent and asymptotically efficient.
  - To estimate θ, run an initial OLS regression of Y on X, capturing the residuals.
    - Then regress $\hat{u}_i^2 = \theta_0 + \theta_1 X_i^2 + \varepsilon_i$.

- Using the estimates from this equation, calculate $\widehat{\text{var}}(u_i \mid X_i) = \hat{\theta}_0 + \hat{\theta}_1 X_i^2$.
- Define the weighted observations as

$$\hat{\tilde{Y}}_i \equiv \frac{Y_i}{\sqrt{\widehat{\text{var}}(u_i \mid X_i)}},$$

$$\hat{\tilde{X}}_{0,i} \equiv \frac{1}{\sqrt{\widehat{\text{var}}(u_i \mid X_i)}},$$

$$\hat{\tilde{X}}_i \equiv \frac{X_i}{\sqrt{\widehat{\text{var}}(u_i \mid X_i)}},$$

$$\hat{\tilde{u}}_i \equiv \frac{u_i}{\sqrt{\widehat{\text{var}}(u_i \mid X_i)}}.$$

- The weighted error term will converge asymptotically to homoskedasticity, so we can run the regression $\hat{\tilde{Y}}_i = \beta_0 \hat{\tilde{X}}_{0,i} + \beta_1 \hat{\tilde{X}}_i + \hat{\tilde{u}}_i$ using OLS and the results will be asymptotically BLUE.
- This procedure is "feasible WLS" because we more often know that the variance is a function of certain variables than that we know it up to a constant of proportionality.

- Is WLS better than OLS with heteroskedasticity-consistent SEs?
  - Not necessarily. There may be sufficient uncertainty about the nature of the variance of the error that you could be making things worse rather than better.
  - Yes, *if* you are very confident of the pattern of heteroskedasticity.
  - But, you can do valid inference either way as long as you use the consistent (robust) standard errors for basic OLS.