Multiple Regression Analysis

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$

2. Inference

Ch.4 Multiple Regression: Inference

1. Sampling Distributions of OLSE
2. Testing Hypotheses: The t test
3. Confidence Intervals
4. Testing Hypotheses about a Single Linear Combination of the Parameters
5. Testing Multiple Restrictions: F test
6. Reporting Regression Results

4.1 Sampling Distributions of OLSE

- In Ch.3, we learned that OLSE is BLUE under the Gauss-Markov assumption.
- In order to do classical hypothesis testing, we need to add another assumption (beyond the Gauss-Markov assumptions).
- Assume that u is independent of $x_1, x_2, \ldots, x_k$ and u is ***normally distributed*** with zero mean and variance σ^2:

$$u \sim \mathrm{N}(0, \sigma^2) \qquad \text{MLR.6}$$

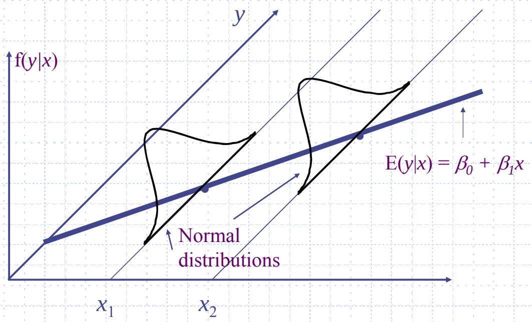
CLM Assumptions

- We can summarize the population assumptions of CLM as follows

$$y|\mathbf{x} \sim \mathrm{N}(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \sigma^2)$$

- Large sample will let us drop normality assumption, because of ***Central Limit Theorem*** **(CLT)**.
  - **CLT**: The sum of independent random variables, when standardized by its standard deviation, has a distribution that tends to standard normal as the sample size grows.

The homoskedastic normal distribution with a single explanatory variable



Normal Sampling Distributions

Theorem 4.1

Under the CLM assumptions, conditional on the sample values of the independent variables

$$\hat{\beta}_j \sim \mathrm{N}\left[\beta_j, Var\left(\hat{\beta}_j\right)\right]. \qquad (4.1)$$

$$\text{Therefore, } \frac{\left(\hat{\beta}_j - \beta_j\right)}{sd\left(\hat{\beta}_j\right)} \sim \mathrm{N}(0,1).$$

$\hat{\beta}_j$ is distributed normally because it is a linear combination of the errors.

4.2 Testing Hypotheses: *t* Test

Theorem 4.2

Under the CLM assumptions,

$$\frac{(\hat{\beta}_j - \beta_j)}{se(\hat{\beta}_j)} \sim t_{n-k-1}. \quad (4.3)$$

Note this is a *t* distribution (not normal dist.) because we have to estimate σ^2 by $\hat{\sigma}^2$.

Note the degrees of freedom: $n-k-1$.

Cont. The *t* Test

- Knowing the sampling distribution for the standardized estimator allows us to carry out hypothesis tests.
- Start with a ***null hypothesis***, for example,
  H_0: $\beta_j = 0$ (4.4)
  - The null hypothesis means that x_j has no effect on *y*, controlling for other independent variables.

Cont. The *t* Test

- The statistic we use to test (4.4) is called "the" *t* statistic of estimate β_j:

$$t_{\hat{\beta}_j} \equiv \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \quad (4.5)$$

- Then, we will use our *t* statistic along with a rejection rule to determine whether to accept the null hypotheses, H_0.

t Test: procedure

- Besides our null, H_0, we need an ***alternative hypothesis***, H_1, and a ***significance level***.
- H_1 may be one-sided, or two-sided.
  - H_1: $\beta_j > 0$ and H_1: $\beta_j < 0$ are one-sided
  - H_1: $\beta_j \neq 0$ is a two-sided alternative
- If we want to have only a 5% probability of rejecting H_0 if it is really true, then we say our significance level is 5%.

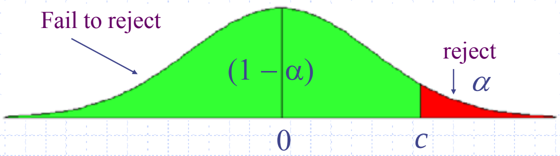
Cont. t Test: procedure

- Having picked a significance level, α, we look up the $(1-\alpha)^{th}$ percentile in a *t* distribution with $n-k-1$ *d.f.* and call this *c*, the critical value.
  - We can reject the null hypothesis if the *t* statistic is greater than the critical value.
  - If the *t* statistic is less than the critical value then we fail to reject the null.

One-Sided Alternatives

$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i$

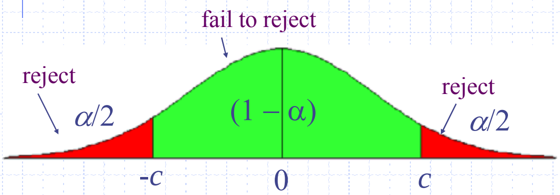
H_0: $\beta_j = 0$, H_1: $\beta_j > 0$



Two-Sided Alternatives

$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i$

$H_0: \beta_j = 0$, $H_1: \beta_j \neq 0$



Summary for $H_0: \beta_j = 0$

- The alternative is usually assumed to be two-sided.
- If we reject the null, we typically say, "x_j is statistically significant at the α % level".
- If we fail to reject the null, we typically say, "x_j is statistically insignificant at the α % level".

Testing other hypotheses

- A more general form of the t statistic recognizes that we may want to test something like

$$H_0: \beta_j = a_j . \qquad (4.12)$$

- In this case, the appropriate t statistic is

$$t = \frac{\left(\hat{\beta}_j - a_j\right)}{se\left(\hat{\beta}_j\right)} \qquad (4.13),$$

where $a_j = 0$ for the standard test

Computing p-values for t tests

- An alternative to the classical approach is to ask, "what is the smallest significance level at which the null would be rejected?"
- So, compute the t statistic, and then look up what percentile it is in the appropriate t distribution – this is the ***p*-value**.
- Most computer packages will compute the p-value, ***assuming a two-sided test*** ($H_0: \beta_j = 0$).
  - If you want a one-sided alternative, just divide the two-sided p-value by 2.

Economic vs. statistical significance

- Check for statistical significance.
- Discuss economic or practical importance.
  - Small t & economic importance may be due to small sample.

  → You can use more "liberal" significance level.
  - Significant, wrong sign & economic importance could result from omitting variable bias or sample selection bias.

  →Rethink the model or the data nature.

4.3 Confidence Intervals

- Another way to use classical statistical testing is to construct a confidence interval using the same critical value as was used for a two-sided test.
- A (1 - α) % confidence interval is defined as

$$\hat{\beta}_j \pm c \times se\left(\hat{\beta}_j\right), \qquad (4.16)$$

where c is the $\left(1 - \frac{\alpha}{2}\right)$ percentile in a t_{n-k-1} distribution

4.4 Testing a Linear Combination

- Suppose instead of testing whether β_1 is equal to a constant, you want to test if it is equal to another parameter, that is

  $H_0 : \beta_1 = \beta_2$. (4.18)
- Use same basic procedure for forming a t statistic; $t = \dfrac{\hat{\beta}_1 - \hat{\beta}_2}{se(\hat{\beta}_1 - \hat{\beta}_2)}$ (4.20)
- Usually, it's hard to calculate it by hand. So, more generally, you can always restate the problem to get the test you want.

Econometrics 19

Example:

- Suppose you are interested in the difference of return by academic record.

  $\ln(w) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u$ (4.17)

  H_0: $\beta_1 = \beta_2$, (4.18) or

  H_0: $\theta_1 = \beta_1 - \beta_2 = 0$ (4.24) $\Rightarrow$ $\beta_1 = \theta_1 + \beta_2$,
- so substitute in and rearrange

  $\ln(w) = \beta_0 + \theta_1 jc + \beta_2 (jc + univ) + \beta_3 exper + u$

Econometrics 20

Cont. Example

- This is the same model as originally, but now you get a standard error for $\beta_1 - \beta_2 = \theta_1$ directly from the basic regression.
- Any linear combination of parameters could be tested in a similar manner.
  - Other examples of hypotheses about a single linear combination of parameters:

    $\beta_1 = 1 + \beta_2$, $\beta_1 = 5\beta_2$, $\beta_1 = -1/2\beta_2$ etc.

Econometrics 21

4.5 Multiple Linear Restrictions

- Everything we've done so far has involved testing a single linear restriction, (e.g. $\beta_1 = 0$ or $\beta_1 = \beta_2$)
- However, we may want to jointly test multiple hypotheses about our parameters.
  - A typical example is testing "exclusion restrictions" – we want to know if ***a group of parameters*** are all equal to zero.

Econometrics 22

Testing Exclusion Restrictions

- Now the null & alternative hypothesis might be something like

  H_0: $\beta_{k-q+1} = 0, \ldots, \beta_k = 0$.

  H_1: H_0 is not true.
  - We can't just check each t statistic separately, because we want to know if the q parameters are ***jointly*** significant at a given level – it is possible for none to be individually significant at that level.

Econometrics 23

Cont. Exclusion Restrictions

- To do the test, we need to estimate the "restricted model" without $x_{k-q+1}, \ldots, x_k$, as well as the "unrestricted model" with all x's.

  (unrestricted model)

  $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-q} x_{k-q} + \ldots + \beta_k x_k + u$ (4.34)

  (restricted model)

  $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-q} x_{k-q} + u$ (4.36)

Econometrics 24

Cont. Exclusion Restrictions

- Intuitively, we want to know if the change in SSR is big enough to warrant inclusion of $x_{k-q+1}, \ldots, x_k$.

$$F \equiv \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} \quad (4.37),$$

where SSR_r is the sum of squared residuals from the restricted model and SSR_{ur} is the sum of squared residuals from the unrestricted model.

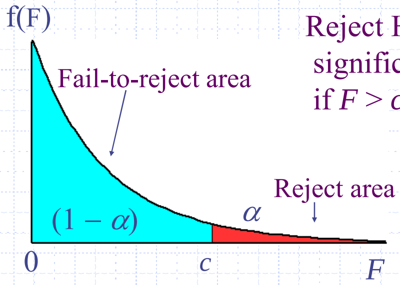
The *F* statistic

- The F statistic is always positive, since the SSR_r can't be less than the SSR_{ur}.
- Essentially the F statistic is measuring the relative increase in SSR when moving from the unrestricted to restricted model.
  - q = number of restrictions = $df_r - df_{ur}$
  - $n - k - 1 = df_{ur}$

Cont. The *F* statistic

- To decide if the increase in SSR when we move to a restricted model is "big enough" to reject the exclusions, we need to know about the sampling distribution of our F stat.
- Not surprisingly, $F \sim F_{q,n-k-1}$, where q is referred to as the numerator degrees of freedom and $n - k - 1$ as the denominator degrees of freedom.

Cont. The *F* statistic



Reject H_0 at α significance level if $F > c$

The R^2 form of the *F* statistic

- Because the SSR's may be large and unwieldy, an alternative form of the formula is useful.
- We use the fact that $SSR = SST(1 - R^2)$ for any regression, so we can substitute in for SSR_u and SSR_{ur}.

$$F \equiv \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/(n-k-1)} \quad (4.41),$$

where again r is restricted and ur is unrestricted.

Overall Significance

- A special case of exclusion restrictions is to test that none of the regressors has an effect on y.

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0 \quad (4.44)$$

- Since the R^2 from a model with only an intercept will be zero, the F statistic is simply,

$$F = \frac{R^2/k}{(1 - R^2)/(n-k-1)} \quad (4.46)$$

F Statistic Summary

- Just as with t statistics, p-values can be calculated by looking up the percentile in the appropriate F distribution.
- If only one exclusion is being tested, then $F = t^2$, and the p-values will be the same.

4.6 Reporting Results

Dependent Variable: log(*salary*)			
Independent Variables	**(1)**	**(2)**	**(3)**
log(*sales*)	.224 (.027)	.158 (.040)	.188 (.040)
log(*mktval*)	——	.112 (.050)	.100 (.049)
profmarg	——	−.0023 (.0022)	−.0022 (.0021)
ceoten	——	——	.0171 (.0055)
comten	——	——	−.0092 (.0033)
intercept	4.94 (0.20)	4.62 0.25)	4.57 (0.25)
Observations *R*-squared	177 .281	177 .304	177 .353