Lecture 12: Hypothesis testing in multiple regression

Economics 326 — Introduction to Econometrics II

Vadim Marmer, UBC

The model

- We consider the classical normal linear regression model:
  1. $Y_i = \beta_0 + \beta_1 X_{1,i} + ... + \beta_k X_{k,i} + U_i$.
  2. Conditional on $\mathbf{X}$, $\mathrm{E}[U_i \mid \mathbf{X}] = 0$ for all i's.
  3. Conditional on $\mathbf{X}$, $\mathrm{E}[U_i^2 \mid \mathbf{X}] = \sigma^2$ for all i's.
  4. Conditional on $\mathbf{X}$, $\mathrm{E}[U_i U_j \mid \mathbf{X}] = 0$ for all $i \neq j$.
  5. Conditional on $\mathbf{X}$, U_i's are jointly normally distributed.
- We also continue to assume **no perfect multicollinearity**: the k regressors and constant **do not** form a perfect linear combination, i.e., we **cannot** find constants $c_1, \ldots, c_k, c_{k+1}$ (not all equal to zero) such that for **all i's**:

$$c_1 X_{1,i} + ... + c_k X_{k,i} + c_{k+1} = 0.$$

Testing a single coefficient

- Take the j-th coefficient β_j, $j \in \{0, 1, \ldots, k\}$.
- Under our assumptions, conditional on $\mathbf{X}$, the OLS estimator $\hat{\beta}_j$ satisfies $\hat{\beta}_j \sim N\left(\beta_j, \mathrm{Var}\left(\hat{\beta}_j \mid \mathbf{X}\right)\right)$, where $\mathrm{Var}\left(\hat{\beta}_j \mid \mathbf{X}\right) = \sigma^2 / \sum_{i=1}^n \tilde{X}_{j,i}^2$ (see Lecture 11).
- Therefore, $\left(\hat{\beta}_j - \beta_j\right) / \sqrt{\mathrm{Var}\left(\hat{\beta}_j \mid \mathbf{X}\right)} \sim N(0, 1)$.
- The conditional variance $\mathrm{Var}\left(\hat{\beta}_j \mid \mathbf{X}\right)$ is unknown because σ^2 is unknown. The estimator for $\mathrm{Var}\left(\hat{\beta}_j \mid \mathbf{X}\right)$ is

$$\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) = \frac{s^2}{\sum_{i=1}^n \tilde{X}_{j,i}^2},$$

  where $s^2 = \sum_{i=1}^n \hat{U}_i^2 / (n - k - 1)$ (see Lecture 10).

Testing a single coefficient

- We have that conditional on $\mathbf{X}$,

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)}} \sim t_{n-k-1}.$$

- Standard error: $\mathrm{se}\left(\hat{\beta}_j\right) = \sqrt{\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)} = \sqrt{s^2 / \sum_{i=1}^n \tilde{X}_{j,i}^2}$.

Testing a single coefficient: two-sided

- Consider testing $H_0 : \beta_j = \beta_{j,0}$ against $H_1 : \beta_j \neq \beta_{j,0}$.
- Under H_0, we have that

$$T = \frac{\hat{\beta}_j - \beta_{j,0}}{\sqrt{\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)}} \sim t_{n-k-1}.$$

- Let $t_{df,\tau}$ be the τ-th quantile of the t_{df} distribution.
- **Test:** Reject H_0 when $|T| > t_{n-k-1,1-\alpha/2}$.
- **P-value:** p-value $= 2\left(1 - F_{t_{n-k-1}}(|T|)\right)$, where $F_{t_{n-k-1}}$ is the CDF of the t_{n-k-1} distribution.

Testing a linear combination of coefficients

- Let $c_0, c_1, ..., c_k, r$ be some constants. Consider testing

$$\begin{aligned} &H_0 : c_0\beta_0 + c_1\beta_1 + ... + c_k\beta_k = r \text{ against} \\ &H_1 : c_0\beta_0 + c_1\beta_1 + ... + c_k\beta_k \neq r. \end{aligned}$$

- **Example 1:** Consider $\ln Y_i = \beta_0 + \beta_1 \ln L_i + \beta_2 \ln K_i + U_i$. To test for constant returns to scale, $H_0 : \beta_1 + \beta_2 = 1$, set $c_0 = 0, c_1 = 1, c_2 = 1, r = 1$.
- **Example 2:** Consider $\ln\left(Wage_i\right) = \beta_0 + \beta_1 Experience_i + \beta_2 PrevExperience_i + ... + U_i$. To test that the two experience variables have the same effect on wage, $H_0 : \beta_1 - \beta_2 = 0$, set $c_0 = 0, c_1 = 1, c_2 = -1, c_3 = ... = c_k = 0, r = 0$.
- **Example 3:** Consider $\ln\left(Wage_i\right) = \beta_0 + \beta_1 Exper_i + \beta_2 Exper_i^2 + ... + U_i$. The marginal effect of experience is $\beta_1 + 2\beta_2 Exper_i$. If the wage-experience profile is concave ($\beta_2 < 0$), the marginal effect is smallest at the highest experience level. To test whether the marginal effect equals zero at $Exper = 20$: $H_0 : \beta_1 + 40\beta_2 = 0$, with $c_1 = 1, c_2 = 40, r = 0$.

Testing a linear combination of coefficients

- We have that under $H_0 : c_0\beta_0 + c_1\beta_1 + ... + c_k\beta_k = r$,

$$\begin{aligned} &\frac{c_0\hat{\beta}_0 + c_1\hat{\beta}_1 + ... + c_k\hat{\beta}_k - r}{\sqrt{\mathrm{Var}\left(c_0\hat{\beta}_0 + c_1\hat{\beta}_1 + ... + c_k\hat{\beta}_k \mid \mathbf{X}\right)}} \\ &= \frac{c_0\hat{\beta}_0 + c_1\hat{\beta}_1 + ... + c_k\hat{\beta}_k - (c_0\beta_0 + c_1\beta_1 + ... + c_k\beta_k)}{\sqrt{\mathrm{Var}\left(c_0\hat{\beta}_0 + c_1\hat{\beta}_1 + ... + c_k\hat{\beta}_k \mid \mathbf{X}\right)}} \\ &\sim N(0,1). \end{aligned}$$

- The variance of the linear combination is

$$\begin{aligned} &\mathrm{Var}\left(c_0\hat{\beta}_0 + c_1\hat{\beta}_1 + ... + c_k\hat{\beta}_k \mid \mathbf{X}\right) \\ &= \sum_{j=0}^{k} c_j^2 \mathrm{Var}\left(\hat{\beta}_j \mid \mathbf{X}\right) + \sum_{j=0}^{k}\sum_{l \neq j} c_j c_l \mathrm{Cov}\left(\hat{\beta}_j, \hat{\beta}_l \mid \mathbf{X}\right). \end{aligned}$$

Testing a linear combination of coefficients

- Consider

$$T = \frac{c_0\hat{\beta}_0 + c_1\hat{\beta}_1 + ... + c_k\hat{\beta}_k - r}{\sqrt{\widehat{\mathrm{Var}}\left(c_0\hat{\beta}_0 + c_1\hat{\beta}_1 + ... + c_k\hat{\beta}_k\right)}}.$$

- Under $H_0 : c_0\beta_0 + c_1\beta_1 + ... + c_k\beta_k = r$,

$$T \sim t_{n-k-1}.$$

- **Two-sided test:** Reject H_0 when $|T| > t_{n-k-1,1-\alpha/2}$.

CRS test: details

- Consider the model $\ln Y_i = \beta_0 + \beta_1 \ln L_i + \beta_2 \ln K_i + U_i$.
- We want to test for constant returns to scale: $H_0 : \beta_1 + \beta_2 = 1$.
- The test statistic: $T = \dfrac{\hat{\beta}_1 + \hat{\beta}_2 - 1}{\sqrt{\widehat{\mathrm{Var}}\left(\hat{\beta}_1 + \hat{\beta}_2\right)}}$.
- The estimated variance:

$$\widehat{\mathrm{Var}}\left(\hat{\beta}_1 + \hat{\beta}_2\right) = \widehat{\mathrm{Var}}\left(\hat{\beta}_1\right) + \widehat{\mathrm{Var}}\left(\hat{\beta}_2\right) + 2\widehat{\mathrm{Cov}}\left(\hat{\beta}_1, \hat{\beta}_2\right).$$

  - $\widehat{\mathrm{Var}}\left(\hat{\beta}_1\right)$ and $\widehat{\mathrm{Var}}\left(\hat{\beta}_2\right)$ can be computed from the corresponding standard errors reported by R.
  - In R, $\widehat{\mathrm{Cov}}\left(\hat{\beta}_1, \hat{\beta}_2\right)$ can be obtained (together with the variances) by using the command `vcov(fit)` after running a regression.
- Reject $H_0 : \beta_1 + \beta_2 = 1$ if $|T| > t_{n-3,1-\alpha/2}$.

Example

- 1000 observations were generated using the following model:

$$\left.\begin{array}{l} L_i = e^{l_i} \\ K_i = e^{k_i} \end{array}\right\} \text{ where } l_i, k_i \text{ are iid } N(0,1), \mathrm{Cov}(l_i, k_i) = 0.5,$$
$$U_i \sim \text{iid } N(0,1) \text{ is independent of } l_i, k_i,$$
$$Y_i = L_i^{0.35} K_i^{0.52} e^{U_i}.$$

- The following equation was estimated:

$$\ln Y_i = \beta_0 + \beta_1 \ln L_i + \beta_2 \ln K_i + U_i.$$

- We test $H_0 : \beta_1 + \beta_2 = 1$ against $H_1 : \beta_1 + \beta_2 \neq 1$ at the 5% significance level.

```
set.seed(123)
n <- 1000
lnL <- rnorm(n)
lnK <- 0.5 * lnL + sqrt(1 - 0.5^2) * rnorm(n)
U <- rnorm(n)
lnY <- 0.35 * lnL + 0.52 * lnK + U
```

Example: regression output

- Regression output:

```
fit <- lm(lnY ~ lnL + lnK)
summary(fit)
```

```

Call:
lm(formula = lnY ~ lnL + lnK)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8360 -0.6277 -0.0370  0.6538  3.3787 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.02093    0.03098  -0.676    0.499    
lnL          0.31263    0.03735   8.371   <2e-16 ***
lnK          0.55176    0.03555  15.522   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9788 on 997 degrees of freedom
Multiple R-squared:  0.3942,	Adjusted R-squared:  0.393 
F-statistic: 324.4 on 2 and 997 DF,  p-value: < 2.2e-16
```

- The variance-covariance matrix of the coefficient estimates:

```
vcov(fit)
```

```
              (Intercept)           lnL           lnK
(Intercept)  9.598283e-04  1.015829e-05 -4.491794e-05
lnL          1.015829e-05  1.394680e-03 -7.281792e-04
lnK         -4.491794e-05 -7.281792e-04  1.263649e-03
```

- The critical value $t_{n-3,0.975}$:

```
qt(1 - 0.025, df = fit$df.residual)
```

```
[1] 1.962346
```

Example: manual calculation

- From the regression output:

```
b1 <- coef(fit)["lnL"]
b2 <- coef(fit)["lnK"]
V <- vcov(fit)

cat("b1 =", b1, "\n")
```

```
b1 = 0.3126275
```

```
cat("b2 =", b2, "\n")
```

```
b2 = 0.5517621
```

```
cat("Var(b1) =", V["lnL", "lnL"], "\n")
```

```
Var(b1) = 0.00139468
```

```
cat("Var(b2) =", V["lnK", "lnK"], "\n")
```

```
Var(b2) = 0.001263649
```

```
cat("Cov(b1, b2) =", V["lnL", "lnK"], "\n")
```

```
Cov(b1, b2) = -0.0007281792
```

- The standard error of $\hat{\beta}_1 + \hat{\beta}_2$:

```
se_sum <- sqrt(V["lnL", "lnL"] + V["lnK", "lnK"] + 2 * V["lnL", "lnK"])
cat("se(b1 + b2) =", se_sum, "\n")
```

```
se(b1 + b2) = 0.03466944
```

- The test statistic:

```
T_stat <- (b1 + b2 - 1) / se_sum
cat("T =", T_stat, "\n")
```

```
T = -3.911526
```

- The critical value:

```
cv <- qt(1 - 0.025, df = fit$df.residual)
cat("|T| =", abs(T_stat), ", critical value =", cv, "\n")
```

```
|T| = 3.911526 , critical value = 1.962346
```

- Since $|T| > t_{997,0.975}$, we reject H_0.
- Ignoring the covariance leads to an incorrect result:

```
se_wrong <- sqrt(V["lnL", "lnL"] + V["lnK", "lnK"])
T_wrong <- (b1 + b2 - 1) / se_wrong
cat("T (ignoring covariance) =", T_wrong, "\n")
```

```
T (ignoring covariance) = -2.6302
```

Re-parametrization approach

- We want to test $\beta_1 + \beta_2 = 1$ in $\ln Y_i = \beta_0 + \beta_1 \ln L_i + \beta_2 \ln K_i + U_i$.
- Define $\delta = \beta_1 + \beta_2$, or $\beta_2 = \delta - \beta_1$, so that

$$\begin{aligned}\ln Y_i &= \beta_0 + \beta_1 \ln L_i + \beta_2 \ln K_i + U_i \\ &= \beta_0 + \beta_1 \ln L_i + (\delta - \beta_1) \ln K_i + U_i \\ &= \beta_0 + \beta_1 (\ln L_i - \ln K_i) + \delta \ln K_i + U_i.\end{aligned}$$

- Generate a new variable $D_i = \ln L_i - \ln K_i$.
- Estimate $\ln Y_i = \beta_0 + \beta_1 D_i + \delta \ln K_i + U_i$.
- Test $H_0 : \delta = 1$ against $H_1 : \delta \neq 1$.

Example: reparameterization

- Reparameterized regression output:

```
D <- lnL - lnK
fit2 <- lm(lnY ~ D + lnK)
summary(fit2)
```

```
Call:
lm(formula = lnY ~ D + lnK)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-2.8360 -0.6277 -0.0370  0.6538  3.3787 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.02093    0.03098  -0.676    0.499    
D            0.31263    0.03735   8.371   <2e-16 ***
lnK          0.86439    0.03467  24.932   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9788 on 997 degrees of freedom
Multiple R-squared:  0.3942,	Adjusted R-squared:  0.393 
F-statistic: 324.4 on 2 and 997 DF,  p-value: < 2.2e-16
```

- The 95% confidence interval for the coefficient on $\ln K$:

```r
confint(fit2, "lnK")
```

```
        2.5 %   97.5 %
lnK 0.7963561 0.932423
```

- The interval does not include 1, so we reject H_0.
- In the original equation, $\hat{\beta}_1+\hat{\beta}_2$ equals the coefficient on $\ln K$ in the reparameterized regression, and $\text{se}\left(\hat{\beta}_1+\hat{\beta}_2\right)$ equals its standard error.

Testing with `linearHypothesis()` in R

- The `car` package provides `linearHypothesis()`, which directly tests linear restrictions on regression coefficients.
- Testing for constant returns to scale ($\beta_1 + \beta_2 = 1$):

```r
library(car)
linearHypothesis(fit, "lnL + lnK = 1")
```

```

Linear hypothesis test:
lnL  + lnK = 1

Model 1: restricted model
Model 2: lnY ~ lnL + lnK

  Res.Df    RSS Df Sum of Sq    F    Pr(>F)    
1    998 969.76                                
2    997 955.10  1    14.657 15.3 9.793e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- `linearHypothesis()` reports an F-statistic. If $T \sim t_{n-k-1}$, then $F = T^2 \sim F_{1,n-k-1}$.
- For a single linear restriction, the F-test and the two-sided t-test are equivalent: $F = T^2$ and the p-values are identical.
- Testing for equal effects ($\beta_1 = \beta_2$):

```r
linearHypothesis(fit, "lnL = lnK")
```

```

Linear hypothesis test:
```

```
lnL - lnK = 0

Model 1: restricted model
Model 2: lnY ~ lnL + lnK

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    998 968.42
2    997 955.10  1    13.314 13.898 0.0002039 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```