Introductory Econometrics

Lecture 13: Hypothesis testing in the multiple regression model

Instructor: Ma, Jun

Renmin University of China

April 26, 2023

The model

- We consider the classical normal linear regression model:
  1. $Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_k X_{k,i} + U_i$.
  2. Conditional on X's, $\mathrm{E}[U_i] = 0$ for all i's.
  3. Conditional on X's, $\mathrm{E}\left[U_i^2\right] = \sigma^2$ for all i's.
  4. Conditional on X's, $\mathrm{E}\left[U_i U_j\right] = 0$ for all $i \neq j$.
  5. Conditional on X's, U_i's are jointly normally distributed.
- We also continue to assume no perfect multicolinearity: 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:

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

Testing a hypothesis about a single coefficient

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

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

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

- We have that conditional on X's,

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

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

Testing a hypothesis about a single coefficient: Two-sided alternatives

- 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: Find $t_{n-k-1,1-\tau}$ such that $|T| = t_{n-k-1,1-\tau}$. The p-value=$\tau \times 2$.

Testing a hypothesis about a single coefficient: One-sided alternatives

- Consider testing $H_0 : \beta_j \leq \beta_{j,0}$ against $H_1 : \beta_j > \beta_{j,0}$.
- When $\beta_j = \beta_{j,0}$ 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 $T > t_{n-k-1,1-\alpha}$.
- P-value: Find $t_{n-k-1,1-\tau}$ such that $T = t_{n-k-1,1-\tau}$. The p-value=τ.

Testing a hypothesis about a single linear combination of the coefficients

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

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

- Example 1: Consider the model

$$\log(Y_i) = \beta_0 + \beta_1 \log(L_i) + \beta_2 \log(K_i) + U_i.$$

  - We want to test for constant returns to scale $H_0 : \beta_1 + \beta_2 = 1$.
  - In this case: $c_0 = 0, c_1 = 1, c_2 = 1, r = 1$.

- Let $r, c_0, c_1, \ldots, c_k$ are some constants. Consider testing

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

- Example 2: Consider the model

$$\log(Wage_i) = \beta_0 + \beta_1 Experience_i + \beta_2 PrevExperience_i + \beta_3 X_{3,i} + \ldots \beta_k X_{k,i} + U_i.$$

  - We want to test that $Experience$ and $PrevExperience$ have the same effect on wage: $H_0 : \beta_1 = \beta_2$ or $H_0 : \beta_1 - \beta_2 = 0$.
  - In this case: $c_0 = 0, c_1 = 1, c_2 = -1, c_3 = \ldots = c_k = 0, r = 0$.

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

$$\frac{c_0\hat{\beta}_0 + c_1\hat{\beta}_1 + \ldots + c_k\hat{\beta}_k - r}{\sqrt{\mathrm{Var}\left[c_0\hat{\beta}_0 + c_1\hat{\beta}_1 + \ldots + c_k\hat{\beta}_k\right]}} = \frac{c_0\hat{\beta}_0 + c_1\hat{\beta}_1 + \ldots + c_k\hat{\beta}_k - (c_0\beta_0 + c_1\beta_1 + \ldots + c_k\beta_k)}{\sqrt{\mathrm{Var}\left[c_0\hat{\beta}_0 + c_1\hat{\beta}_1 + \ldots + c_k\hat{\beta}_k\right]}} \sim \mathrm{N}(0,1).$$

- Note that

$$\mathrm{Var}\left[c_0\hat{\beta}_0 + c_1\hat{\beta}_1 + \ldots + c_k\hat{\beta}_k\right] = \sum_{j=0}^{k} c_j^2 \mathrm{Var}\left[\hat{\beta}_j\right] + \sum_{j=0}^{k}\sum_{l \neq j} c_j c_l \cdot \mathrm{Cov}\left[\hat{\beta}_j, \hat{\beta}_l\right].$$

- Consider

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

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

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

- Two-sided Test: Reject H_0 when $|T| > t_{n-k-1,1-\alpha/2}$.
- One-sided: When testing $H_0 : c_0\beta_0 + c_1\beta_1 + \ldots + c_k\beta_k \leq r$ against $H_1 : c_0\beta_0 + c_1\beta_1 + \ldots + c_k\beta_k > r$, reject H_0 when $T > t_{n-k-1,1-\alpha}$.

- Consider the model $\log(Y_i) = \beta_0 + \beta_1 \log(L_i) + \beta_2 \log(K_i) + U_i$.
- We want to test for constant returns to scale: $H_0 : \beta_1 + \beta_2 = 1$.
- The test statistic: $T = \frac{\hat{\beta}_1 + \hat{\beta}_2 - 1}{\sqrt{\widehat{\text{Var}}\left[\hat{\beta}_1 + \hat{\beta}_2\right]}}$.
- $\widehat{\text{Var}}\left[\hat{\beta}_1 + \hat{\beta}_2\right] = \widehat{\text{Var}}\left[\hat{\beta}_1\right] + \widehat{\text{Var}}\left[\hat{\beta}_2\right] + 2\widehat{\text{Cov}}\left[\hat{\beta}_1, \hat{\beta}_2\right]$.
  - $\widehat{\text{Var}}\left(\hat{\beta}_1\right)$ and $\widehat{\text{Var}}\left(\hat{\beta}_2\right)$ can be computed from the corresponding standard errors reported by Stata.
  - In Stata, $\widehat{\text{Cov}}\left[\hat{\beta}_1, \hat{\beta}_2\right]$ can be obtained (together with the variances) by using the command "matrix list e(V)" 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:

$$\left.\begin{array}{l} L_i = e^{l_i} \\ K_i = e^{k_i} \end{array}\right\} \text{ where } l_i, k_i \text{ are iid } \mathrm{N}(0,1), \mathrm{Cov}[l_i, k_i] = 0.5,$$

$$U_i \sim \text{iid } \mathrm{N}(0,1) \text{ is independent of } l_i, k_i,$$

$$Y_i = L_i^{0.35} K_i^{0.52} e^{U_i}.$$

- The following equation was estimated:

$$\log(Y_i) = \beta_0 + \beta_1 \log(L_i) + \beta_2 \log(K_i) + U_i.$$

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

```
. regress lnY lnL lnK
      Source |       SS       df       MS              Number of obs =    1000
-------------+------------------------------           F(  2,   997) =  321.51
       Model |  630.003101     2  315.00155           Prob > F      =  0.0000
    Residual |  976.803234   997  .979742461           R-squared     =  0.3921
-------------+------------------------------           Adj R-squared =  0.3909
       Total |  1606.80633   999  1.60841475           Root MSE      =  .98982

------------------------------------------------------------------------------
         lnY |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnL |   .4484374   .0356212    12.59   0.000     .3785364    .5183385
         lnK |    .466826   .0350918    13.30   0.000     .3979636    .5356883
       _cons |  -.0195782   .0313531    -0.62   0.532    -.0811039    .0419476
------------------------------------------------------------------------------
. matrix list e(V)

symmetric e(V)[3,3]
         lnL              lnK        _cons
  lnL   .00126887
  lnK  -.00059823   .00123144
_cons   5.066e-06    -.000058   .00098302
. display invttail(997 ,0.025)
1.9623462
```

- We obtained:
  - $\hat{\beta}_1 = 0.4484374$,
  - $\hat{\beta}_2 = 0.466826$.
  - $\widehat{\mathrm{Var}}\left[\hat{\beta}_1\right] = 0.00126887 = 0.0356212^2$
  - $\widehat{\mathrm{Var}}\left[\hat{\beta}_2\right] = 0.00123144 = 0.0350918^2$.
  - $\widehat{\mathrm{Cov}}\left[\hat{\beta}_1, \hat{\beta}_2\right] = -0.00059823$.
  - $t_{997,0.975}$ =1.9623462.
- $\sqrt{\widehat{\mathrm{Var}}\left[\hat{\beta}_1 + \hat{\beta}_2\right]} =$ $\sqrt{0.00126887 + 0.00123144 - 2 \times 0.00059823}$ =0.036108863.
- $T = (0.4484374 + 0.466826 - 1)/0.036108863 \approx -2.35$,
- $|T| = 2.35 > 1.962 = t_{997,0.975} \Longrightarrow$ We reject H_0.
- Note that ignoring the covariance leads to an incorrect result: $(0.4484374 + 0.466826 - 1)/\sqrt{0.0356212^2 + 0.0350918^2} \approx -1.69$.

An alternative approach

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

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

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

Example

```
. gen D=lnL-lnK

. regress lnY D lnK

      Source |       SS       df       MS              Number of obs =    1000
-------------+------------------------------           F(  2,   997) =  321.51
       Model |  630.003101     2  315.001551           Prob > F      =  0.0000
    Residual |  976.803233   997  .979742461           R-squared     =  0.3921
-------------+------------------------------           Adj R-squared =  0.3909
       Total |  1606.80633   999  1.60841475           Root MSE      =  .98982

------------------------------------------------------------------------------
         lnY |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           D |   .4484374   .0356212    12.59   0.000     .3785364    .5183385
         lnK |   .9152634   .0361088    25.35   0.000     .8444054    .9861213
       _cons |  -.0195782   .0313531    -0.62   0.532    -.0811039    .0419476
------------------------------------------------------------------------------
```

- The 95% CI for the coefficient on $\log(K)$ in the transformed mode does not include 1 $\Longrightarrow$ We reject H_0.
- Note that in the original equation $\hat{\beta}_1 + \hat{\beta}_2 = 0.9152634$ and $\sqrt{\widehat{\text{Var}}\left[\hat{\beta}_1 + \hat{\beta}_2\right]} = 0.0361088$.

Multiple restrictions

- Consider the model:

$$\log\left(Wage_i\right) = \beta_0 + \beta_1 Experience_i + \beta_2 Experience_i^2 + \\ + \beta_3 PrevExperience_i + \beta_4 PrevExperience_i^2 + \beta_5 Education_i + U_i,$$

  where $Experience$ is the experience at current job, and $PrevExperience$ is the previous experience.
- Suppose that we want to test the null hypothesis that, after controlling for the experience at current job and education, the previous experience has no effect on wage:

$$H_0 : \beta_3 = 0, \beta_4 = 0.$$

- We have two restrictions on the model parameters.
- The alternative hypothesis is that at least one of the coefficients, β_3 or β_4, is different from zero:

$$H_1 : \beta_3 \neq 0 \text{ or } \beta_4 \neq 0.$$

t-statistics and multiple restrictions

- Let T_3 and T_4 be the t-statistics associated with the coefficients of $PrevExperience$ and $PrevExperience^2$:

$$T_3 = \frac{\hat{\beta}_3}{SE\left(\hat{\beta}_3\right)} \text{ and } T_4 = \frac{\hat{\beta}_4}{SE\left(\hat{\beta}_4\right)}.$$

- We can use T_3 and T_4 to test significance of β_3 and β_4 separately by using two separate size α tests:
  - Reject $H_{0,3} : \beta_3 = 0$ in favor of $H_{1,3} : \beta_3 \neq 0$ when $|T_3| > t_{n-k-1,1-\alpha/2}$.
  - Reject $H_{0,4} : \beta_4 = 0$ in favor of $H_{1,4} : \beta_4 \neq 0$ when $|T_4| > t_{n-k-1,1-\alpha/2}$.

- Rejecting $H_0 : \beta_3 = 0, \beta_4 = 0$ in favor of $H_1 : \beta_3 \neq 0$ or $\beta_4 \neq 0$ when at least one of the two coefficients is significant at level α , i.e. when

$$|T_3| > t_{n-k-1,1-\alpha/2} \text{ or } |T_4| > t_{n-k-1,1-\alpha/2},$$

is not a size α test!

  - Recall that if A and B are two sets then $(A \cap B) \subseteq A$ and therefore $\Pr(A \cap B) \leq \Pr(A)$.
  - When $\beta_3 = \beta_4 = 0$:

$$\begin{aligned}
&\Pr\left(\text{Reject } H_{0,3} \text{ or } H_{0,4}\right) = \\
&\quad = \Pr\left[|T_3| > t_{n-k-1,1-\alpha/2} \text{ or } |T_4| > t_{n-k-1,1-\alpha/2}\right] \\
&\quad = \Pr\left[|T_3| > t_{n-k-1,1-\alpha/2}\right] + \Pr\left[|T_4| > t_{n-k-1,1-\alpha/2}\right] \\
&\quad - \Pr\left[|T_3| > t_{n-k-1,1-\alpha/2} \text{ and } |T_4| > t_{n-k-1,1-\alpha/2}\right] \\
&\quad = \alpha + \alpha - \Pr\left[|T_3| > t_{n-k-1,1-\alpha/2} \text{ and } |T_4| > t_{n-k-1,1-\alpha/2}\right] \\
&\quad \geq \alpha.
\end{aligned}$$

Testing multiple exclusion restrictions

- Consider the model

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_q X_{q,i} + \beta_{q+1} X_{q+1,i} + \ldots + \beta_k X_{k,i} + U_i.$$

  Suppose that we want to test that the first q regressors have no effect on Y (after controlling for other regressors).
- The null hypothesis has q exclusion restrictions:

$$H_0 : \beta_1 = 0, \beta_2 = 0, \ldots, \beta_q = 0.$$

- The alternative hypothesis is that at least one of the restrictions in H_0 is false:

$$H_1 : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or } \ldots \text{ or } \beta_q \neq 0.$$

F-statistic

- The idea of the test is to compare the fit of the unrestricted model with that of the null-restricted model.
- Let SSR_{ur} denote the Residual Sum-of-Squares of the unrestricted model

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_q X_{q,i} + \beta_{q+1} X_{q+1,i} + \ldots + \beta_k X_{k,i} + U_i.$$

- The restricted model given $H_0 : \beta_1 = 0, \ldots, \beta_q = 0$ is

$$Y_i = \beta_0 + \beta_{q+1} X_{q+1,i} + \ldots + \beta_k X_{k,i} + U_i.$$

- LetSSR_r denote the Residual Sum-of-Squares of the restricted model .
- Consider the following statistic:

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}.$$

  - Note that q = number of restrictions;
  - $n - k - 1$ = unrestricted residual df, where k is the number of regressors in the unrestricted model.

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}.$$

- Since SSR can only increase when you drop some regressors,

$$SSR_r - SSR_{ur} \geq 0,$$

  and therefore $F \geq 0$.
- If the null restrictions are true, the excluded variables do not contribute to explaining Y (in population), and therefore we should expect that $SSR_r - SSR_{ur}$ is small and F is close to zero.
- If the null restriction are false, the imposed restriction should substantially worsen the fit, and we should expect that $SSR_r - SSR_{ur}$ is large and F is far from zero.
- Thus, we should reject H_0 when $F > c$ where c is some positive constant.

F test

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}.$$

- We should reject H_0 when $F > c$.
- There is a probability that $F > c$ even when H_0 is true, thus we need to choose c so that $\Pr[F > c \mid H_0 \text{ is true}] = \alpha$.
- It turns out that when H_0 is true, the F-statistic has F distribution with two parameters: the numerator df (q) and the denominator df ($n-k-1$):

$$F \sim F_{q,n-k-1}.$$

- Similarly to the standard normal and t distributions, the F distribution has been tabulated and its critical values are available in statistical tables and statistical software such as Stata.

When H_0 is true,

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} \sim F_{q,n-k-1}.$$

- Let $F_{q,n-k-1,\tau}$ be the τ-quantile of the $F_{q,n-k-1}$ distribution.
- A size α test $H_0 : \beta_1 = 0, \ldots, \beta_q = 0$ against $H_1 : \beta_1 \neq 0$ or $\ldots$ or $\beta_q \neq 0$ is

$$\text{Reject } H_0 \text{ when } F > F_{q,n-k-1,1-\alpha}.$$

- One can find the p-value by finding τ such that $F = F_{q,n-k-1,1-\tau}$. The p-value is equal to τ.

F distribution in Stata

- To compute F critical values use

$$\text{disp invFtail}(q, n - k - 1, \alpha).$$

- To compute p-values from F distribution use

$$\text{disp Ftail}(q, n - k - 1, F).$$

Example

- Consider the model:

$$\log\left(Wage_i\right) = \beta_0 + \beta_1 Experience_i + \beta_2 Experience_i^2 +$$
$$+\beta_3 PrevExperience_i + \beta_4 PrevExperience_i^2 + \beta_5 Education_i + U_i.$$

- We test

$$H_0 : \beta_3 = 0, \beta_4 = 0 \text{ against } H_1 : \beta_3 \neq 0 \text{ or } \beta_4 \neq 0.$$

- $q = 2$.
- $\alpha = 0.05$.

Example: the unrestricted model

```
. regress lnWage Experience Experience2 PrevExperience PrevExperience2 Education
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  5,   520) =   55.04
       Model |  51.3318741     5  10.2663748           Prob > F      =  0.0000
    Residual |  96.9978773   520  .186534379           R-squared     =  0.3461
-------------+------------------------------           Adj R-squared =  0.3398
       Total |  148.329751   525   .28253286           Root MSE      =   .4319

------------------------------------------------------------------------------
      lnWage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  Experience |   .0471914   .0068074     6.93   0.000     .0338179    .0605649
 Experience2 |  -.0008518   .0002472    -3.45   0.001    -.0013374   -.0003662
PrevExperi~e |   .0168997   .0047331     3.57   0.000     .0076013    .0261981
PrevExperi~2 |  -.0003727   .0001208    -3.09   0.002      -.00061   -.0001354
   Education |   .0887704   .0072131    12.31   0.000     .0745999    .1029408
       _cons |   .2368427     .10287     2.30   0.022     .0347509    .4389346
------------------------------------------------------------------------------
```

- SSR_{ur} =96.9978773.
- $n - k - 1$ =526-5-1=520.

Example: the restricted model

```
. regress lnWage Experience Experience2  Education
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  3,   522) =   85.49
       Model |  48.8668114     3  16.2889371           Prob > F      =  0.0000
    Residual |    99.46294   522  .190542031           R-squared     =  0.3294
-------------+------------------------------           Adj R-squared =  0.3256
       Total |  148.329751   525   .28253286           Root MSE      =  .43651

------------------------------------------------------------------------------
      lnWage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  Experience |   .0510784   .0067937     7.52   0.000      .037732    .0644248
 Experience2 |  -.0009941   .0002463    -4.04   0.000     -.001478   -.0005103
   Education |   .0852822   .0068978    12.36   0.000     .0717313    .0988331
       _cons |   .3688491   .0908138     4.06   0.000     .1904437    .5472544
------------------------------------------------------------------------------
```

- SSR_r =99.46294.

Example: F statistic and test

- To compute the statistic:

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{(99.46294 - 96.9978773)/2}{96.9978773/(526-5-1)} \approx 6.61.$$

- The critical value:

```
. disp invFtail(2,520,0.05)
3.0130572
```

- The test: 6.61 >3.0130572 and at 5% significance level we reject H_0 that previous experience has no effect on wage.
- The p-value:

```
. disp Ftail(2,520,6.61)
.00146284
```

$\Longrightarrow$We reject H_0 for any $\alpha > 0.00146284$.

Example: Stata test command

- Instead of running two models, restricted and unrestricted, one can use the Stata test command after estimation of the unrestricted model.
- To test that previous experience has no effect:

```
. test (PrevExperience=0) (PrevExperience2=0)
```

- The output of this command is:

(1) PrevExperience = 0
(2) PrevExperience2 = 0
F(2, 520) = 6.61
Prob > F = 0.0015

- To test that the coefficient on previous experience equal to the coefficient on experience and the coefficient on previous experience squared is zero:

  ```
  . test (Experience==PrevExperience2) (PrevExperience2=0)
  ```

- The output is:

  ```
   ( 1)  Experience - PrevExperience2 = 0
   ( 2)  PrevExperience2 = 0
         F(  2,   520) =   31.94
              Prob > F =    0.0000
  ```

F and R^2

- Let R^2_{ur} denote the R^2 corresponding to the unrestricted model:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_q X_{q,i} + \beta_{q+1} X_{q+1,i} + \ldots + \beta_k X_{k,i} + U_i.$$

- Let R^2_r denote the R^2 corresponding to the restricted model:

$$Y_i = \beta_0 + \beta_{q+1} X_{q+1,i} + \ldots + \beta_k X_{k,i} + U_i.$$

- The two models have the same dependent variable and therefore the same Total Sum-of-Squares:

$$SST = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = SST_{ur} = SST_r.$$

- In this case, we can write then

$$
\begin{aligned}
F &= \frac{\left(SSR_r - SSR_{ur}\right)/q}{SSR_{ur}/(n-k-1)} \\
&= \frac{\left(\frac{SSR_r}{SST} - \frac{SSR_{ur}}{SST}\right)/q}{\frac{SSR_{ur}}{SST}/(n-k-1)} \\
&= \frac{\left(1 - R_r^2 - \left(1 - R_{ur}^2\right)\right)/q}{\left(1 - R_{ur}^2\right)/(n-k-1)} \\
&= \frac{\left(R_{ur}^2 - R_r^2\right)/q}{\left(1 - R_{ur}^2\right)/(n-k-1)}.
\end{aligned}
$$

F test: more examples

- Suppose that you want to test $H_0 : \beta_1 = 1$ against $H_1 : \beta_1 \neq 1$ in

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + U_i.$$

- The restricted model is

$$Y_i = \beta_0 + X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + U_i,$$

  or

$$Y_i - X_{1,i} = \beta_0 + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + U_i.$$

  1. Generate a new dependent variable $Y_i^* = Y_i - X_{1,i}$.
  2. Regress Y^* against a constant, $X_2, \ldots, X_k$ to obtain SSR_r.
  3. Estimate the unrestricted model to obtain SSR_{ur}.
  4. Compute $F = \frac{(SSR_r - SSR_{ur})/1}{SSR_{ur}/(n-k-1)}$.

- Suppose that you want to test $H_0 : \beta_1 + \beta_2 = 1$ against $H_1 : \beta_1 + \beta_2 \neq 1$ in

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + U_i.$$

- The restricted model is

$$Y_i = \beta_0 + (1 - \beta_2)\, X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + U_i,$$

or

$$Y_i - X_{1,i} = \beta_0 + \beta_2 \left(X_{2,i} - X_{1,i} \right) + \ldots + \beta_k X_{k,i} + U_i.$$

1. Generate a new dependent variable $Y_i^* = Y_i - X_{1,i}$.
2. Generate a new regressor $X_2^* = X_{2,i} - X_{1,i}$.
3. Regress Y^* against a constant, $X_2^*, X_3, \ldots, X_k$ to obtain SSR_r.
4. Estimate the unrestricted model to obtain SSR_{ur}.
5. Compute $F = \frac{(SSR_r - SSR_{ur})/1}{SSR_{ur}/(n-k-1)}$.

Relationship between F and t statistics

- The F statistic can also be used for testing a single restriction.
- In the case of a single restriction, the F test and t test lead to the same outcome because

$$t_{n-k-1}^2 = F_{1,n-k-1}.$$

Test of model significance

- Consider the model

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_k X_{k,i} + U_i.$$

- Suppose that you want to test that none of the regressors explain Y :

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \ldots = \beta_k = 0 \text{ (k restrictions) against} \\ H_1 &: \beta_j \neq 0 \text{ for some } j = 1, \ldots, k. \end{aligned}$$

- The restricted model is given by

$$Y_i = \beta_0 + U_i,$$

and since $\hat{\beta}_0 = \bar{Y}$ in this model,

$$SSR_r = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = SST \text{ and } SSR_{ur} = SSR.$$

- The F statistic for model significance test is

$$\begin{aligned}
F &= \frac{(SSR_r - SSR_{ur})/k}{SSR_{ur}/(n-k-1)} \\
&= \frac{(SST - SSR)/k}{SSR/(n-k-1)} \\
&= \frac{SSE/k}{SSR/(n-k-1)} \\
&= \frac{R^2/k}{\left(1-R^2\right)/(n-k-1)}.
\end{aligned}$$

- The F statistic for the model significance test and its p-value is reported by Stata as in the top part of the regression output.

```
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  5,   520) =   55.04
       Model |  51.3318741     5  10.2663748           Prob > F      =  0.0000
    Residual |  96.9978773   520  .186534379           R-squared     =  0.3461
-------------+------------------------------           Adj R-squared =  0.3398
       Total |  148.329751   525   .28253286           Root MSE      =   .4319
```

Model selection

- If a subset of the coefficients in the linear model

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_k X_{k,i} + U_i$$

  are exactly zero, we wish to find the smallest sub-model consisting of only explanatory variables with nonzero coefficients.
- Estimate the full model with all variables. Let $T_j = \hat{\beta}_j / SE\left(\hat{\beta}_j\right)$ denote the t-statistic for $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$.
- Order $T_1, ..., T_k$ in absolute value:

$$\left|T_{(1)}\right| \geq \left|T_{(2)}\right| \geq \cdots \geq \left|T_{(k)}\right|.$$

- Let $\hat{j}$ be the value of j that minimizes $RSS\,(j) + j \cdot s^2 \log\,(n)$, where $RSS\,(j)$ is the residual sum of squares from the model with j variables corresponding to the j largest absolute t-statistics.
- The selected model is the model with $\hat{j}$ variables corresponding to the $\hat{j}$ largest absolute t-statistics.
- When n is large, with high probability, this selected model is the same as the smallest sub-model with only nonzero coefficients.