Applied Econometrics

Multiple Regression Analysis

Hypothesis Testing Part II

Text: Wooldridge, Chapter 4

June 25, 2011

Hypotheses about a Single Linear Combination of the Parameters

- In applications, we are often interested in testing hypothesis involving more than just a single population parameter.

- To illustrate, we consider a model to compare returns of education at junior (community, two–year) colleges and four–year colleges, which will be referred to as *universities*.

- The model is[1]

$$\log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u, \tag{1}$$

  where *jc* is years attending two–year college, *univ* is number of years at four–year college, and *exper* is month in the workforce. Any combination of *jc* and *univ* is possible.

- The hypothesis of interest is whether one year at a university is worth more than one year at a junior college.

[1] See Wooldridge, p. 147.

- Thus a one–sided alternative appears reasonable,

$$H_0 : \beta_1 = \beta_2, \quad \text{and} \quad H_1 : \beta_1 < \beta_2. \tag{2}$$

- We estimate

$$\widehat{\log(wage)} = \underset{(0.0211)}{1.4723} + \underset{(0.0068)}{0.0667}\, jc + \underset{(0.0023)}{0.0769}\, univ + \underset{(0.0002)}{0.0049}\, exper, \tag{3}$$

  where standard errors are given in parentheses.

- In this case, $n = 6763$, $k + 1 = 4$, so $n - k - 1$ is very large, and we can use the critical values implied by the normal distribution.

- Both *jc* and *univ* have economically and statistically significant partial effects on wage.

- The estimated difference is $\widehat{\beta}_2 - \widehat{\beta}_1 = 0.0102$, i.e., the return to a year at a four–year college is about one percentage point higher than that of a year at a community college.

- To test (2), we can consider the test statistic

$$t = \frac{\widehat{\beta}_1 - \widehat{\beta}_2}{\mathsf{se}(\widehat{\beta}_1 - \widehat{\beta}_2)}, \tag{4}$$

  which has a t distribution with $n - k - 1$ degrees of freedom (i.e., in this particular example, essentially a normal distribution), where se is standard error (estimated standard deviation).

- Now we can use the variance of linear combinations of random variables to determine $\mathsf{se}(\widehat{\beta}_1 - \widehat{\beta}_2)$ as

$$\begin{aligned} \mathsf{se}(\widehat{\beta}_1 - \widehat{\beta}_2) &= \sqrt{\widehat{\sigma}^2 \boldsymbol{r}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{r}} \\ &= \sqrt{\widehat{\mathsf{Var}}(\widehat{\beta}_1) + \widehat{\mathsf{Var}}(\widehat{\beta}_2) - 2\widehat{\mathsf{Cov}}(\widehat{\beta}_1, \widehat{\beta}_2)}, \end{aligned}$$

  where $\boldsymbol{r} = [0, 1, -1, 0]'$.

- This gives

$$\mathsf{se}(\widehat{\beta}_1 - \widehat{\beta}_2) = 0.0069 \Rightarrow t = \frac{-0.0102}{0.0069} = -1.468, \tag{5}$$

with p–value 0.0711, and hence some weak evidence against H_0.

More Straightforward Approach for a Single Linear Combination of the Parameters

- An alternative approach that works for single linear restrictions is as follows:

- Define a new parameter $\theta = \beta_1 - \beta_2$ and then test

$$H_0 : \theta = 0 \quad \text{against} \quad H_1 : \theta < 0. \tag{6}$$

- Equation (1) becomes, with $\beta_1 = \theta + \beta_2$

$$\begin{aligned} \log(wage) &= \beta_0 + (\theta + \beta_2)jc + \beta_2 univ + \beta_3 exper + u \\ &= \beta_0 + \theta jc + \beta_2(jc + univ) + \beta_3 exper + u \\ &= \beta_0 + \theta jc + \beta_2 totcoll + \beta_3 exper + u, \end{aligned} \tag{7}$$

  where *totcoll* = *jc* + *univ* is total years in college.

- This equation can be estimated and a standard test of (6) be performed.

- Estimation of (7) gives

$$\widehat{\log(wage)} = \underset{(0.0211)}{1.4723} - \underset{(0.0069)}{0.0102}\ jc + \underset{(0.0023)}{0.0769}\ totcoll + \underset{(0.0002)}{0.0049}\ exper, \quad (8)$$

  which gives rise to the same conclusion as above.

- Note that β_0, β_2, and β_3 and their standard errors remain unaffected (as it should be), and the only thing that we could not extract directly from (1) is the standard error of $\widehat{\theta}$.

- The general procedure is

  1. Use the linear restriction to solve for one of the original parameters, and then
  2. plug the result into the regression equation and rearrange to get a new regression involving constructed variables.

Testing Multiple Linear Restrictions: The F Test

- Suppose we want to test whether a set of independent variables has no effect on y.

- As an example, consider a model for major league baseball players' salaries,[2]

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + u, \tag{9}$$

  where

  - *salary* = 1993 total salary
  - *years* = years in the league
  - *gamesyr* = average games played per year
  - *bavg* = career batting average
  - *hrunsyr* = home runs per year
  - *rbisyr* = runs batted in per year.

- Variables *bavg*, *hrunsyr*, and rbisyr measure the performance of a player.

[2] Cf. Wooldridge, p. 151.

- We want to test the hypothesis that, once years in the league and games per year have been accounted for, performance has no effect on salary.

- That is, we want to test

$$H_0 : \beta_3 = 0, \quad \beta_4 = 0, \quad \beta_5 = 0, \tag{10}$$

  which is a set of **multiple restrictions** because we are putting more than one restriction on the parameters of (9).

- Note that the alternative

$$H_1 : (10) \text{ is not true.} \tag{11}$$

  is valid if *at least* one of β_3, β_4, or β_5 is not zero.

- We might thus be tempted to test any of the parameters separately, i.e., test

$$H_{0,j} : \beta_j = 0, \quad j = 3, 4, 5, \tag{12}$$

  and reject H_0 in (10) at level α if any of the hypotheses in (12) is rejected at level α.

- However, the type I error of such a test would be larger than 5%:

- If the null (10) is true, then we would *not* reject the three tests (12) with probability $1 - \alpha$.

- Suppose for simplicity that the tests are statistically independent (unrealistic, but sufficient to make the point).

- Then the probability that none of the hypothesis is (12) is rejected is

$$\begin{aligned} \Pr(\text{no } H_{0,j} \text{ rejected} | H_0) &= \Pr(H_{0,1} \text{ not rejected} | H_0) \\ &\times \Pr(H_{0,2} \text{ not rejected} | H_0) \times \Pr(H_{0,3} \text{ not rejected} | H_0) \\ &= (1-\alpha)^3. \end{aligned}$$

- Thus the overall type I error is $1 - (1-\alpha)^3$.

- E.g., if $\alpha = 0.05$, then $1 - (1-\alpha)^3 = 1 - 0.95^3 = 0.14$.

- Using separate t statistics to test a multiple hypothesis can be misleading also in other respects.

- To illustrate, estimation of model (9) gives

$$\begin{aligned} \widehat{\log(salary)} &= \underset{(0.2888)}{11.19} + \underset{(0.0121)}{0.0689}\, years + \underset{(0.0026)}{0.0126}\, gamesyr \qquad (13) \\ &\quad + \underset{(0.0011)}{0.0010}\, bavg + \underset{(0.0161)}{0.0144}\, hrunsyr + \underset{(0.0072)}{0.0108}\, rbisyr, \\ n &= 353, \quad \text{SSR} = \sum_i \widehat{u}_i^2 = 183.1863, \quad R^2 = 0.6278, \end{aligned}$$

  where SSR is the sum of squared residuals.

- The t statistics for the performance variables are

$$t_{\widehat{\beta}_3} = \frac{0.0010}{0.0011} = 0.8868, \quad t_{\widehat{\beta}_4} = \frac{0.0144}{0.0161} = 0.8986, \quad t_{\widehat{\beta}_5} = \frac{0.0108}{0.0072} = 1.5005,$$

  so that none of these is significant at the 5% level.

- Thus, we might be tempted to conclude that we cannot reject (10) at the 5% level.

- However, this turns out to be wrong.

- To indicate why, recall the formula

$$\text{Var}(\widehat{\beta}_j) = \frac{\sigma^2}{ns_{x_j}^2(1 - R_j^2)},$$

  where R_j^2 is the coefficient of determination of a regression of x_j on the other independent variables.

- In our example, we have

$$R_4^2 = 0.874, \quad R_5^2 = 0.944,$$

  and in particular x_4 and x_5 are highly correlated (with correlation 0.891), which makes it difficult to identify the *partial effect* of each of these variables.

- Thus, we need to test whether the variables are *jointly* significant.

- The F **test** can be used to do so.

- The F test can be written in terms of the residuals sum of squares of the **unrestricted model** and the **restricted model**.

- In the **restricted model**, the restrictions associated with the null hypothesis have been imposed in the estimation.

- In the case of **exclusion restrictions**, where several coefficients are set equal to zero, estimation of the restricted model is rather straightforward by just excluding the respective variables.

- For example, in our baseball example, we estimate the equation with only two independent variables, *years* and *gamesyr*, to get

$$\begin{aligned} \widehat{\log(salary)} &= \underset{(0.1083)}{11.22} + \underset{(0.0125)}{0.0713}\, years + \underset{(0.0013)}{0.0202}\, gamesyr \qquad (14) \\ n &= 353, \quad \text{SSR} = \sum_i \widehat{u}_i^2 = 198.3115, \quad R^2 = 0.5971. \end{aligned}$$

- Clearly the SSR of the restricted model is larger than the SSR of the unrestricted model.

- We need a rule to decide whether the difference between the sums of squares is large enough to be statistically significant.

- To state the general result, suppose that, under H_0, there are q independent linear restrictions on the parameters of the model, which can be written as

$$H_0 : \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}, \tag{15}$$

  where $\boldsymbol{R}$ is a $q \times (k+1)$ matrix of rank q.

- For example, in the baseball example,

$$H_0 : \underbrace{\begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}}_{=\boldsymbol{R}} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}}_{=\boldsymbol{r}}. \tag{16}$$

- As another example, we may want to test a hypothesis of the form

$$H_0 : \beta_3 - 2\beta_1 = 1, \quad \beta_2 = 2\beta_1 \tag{17}$$

about the coefficients of the model

$$y = \beta_0\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u, \tag{18}$$

where then $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$, and

$$\boldsymbol{R} = \begin{bmatrix} 0 & -2 & 0 & 1 \\ 0 & 2 & -1 & 0 \end{bmatrix}, \quad \boldsymbol{r} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{19}$$

- The $\mathbf{F}$ **statistic** for this null hypothesis is

$$F = \frac{(\text{SSR}_r - \text{SSR}_{ur})/q}{\text{SSR}_{ur}/(n-k-1)} = \frac{\text{SSR}_r - \text{SSR}_{ur}}{\text{SSR}_{ur}} \frac{n-k-1}{q}, \qquad (20)$$

  where

  - SSR_{ur} is the sum of squared residuals from the **unrestricted regression**,
  - SSR_r is the sum of squared residuals from the **restricted regression**,
  - q is the number of restrictions imposed by the restricted model.

- $n-k-1$ and q are also referred to as

  - $q =$ **numerator degrees of freedom**
  - $n-k-1 =$ **denominator degrees of freedom**.

- H_0 will be rejected if F is "large".

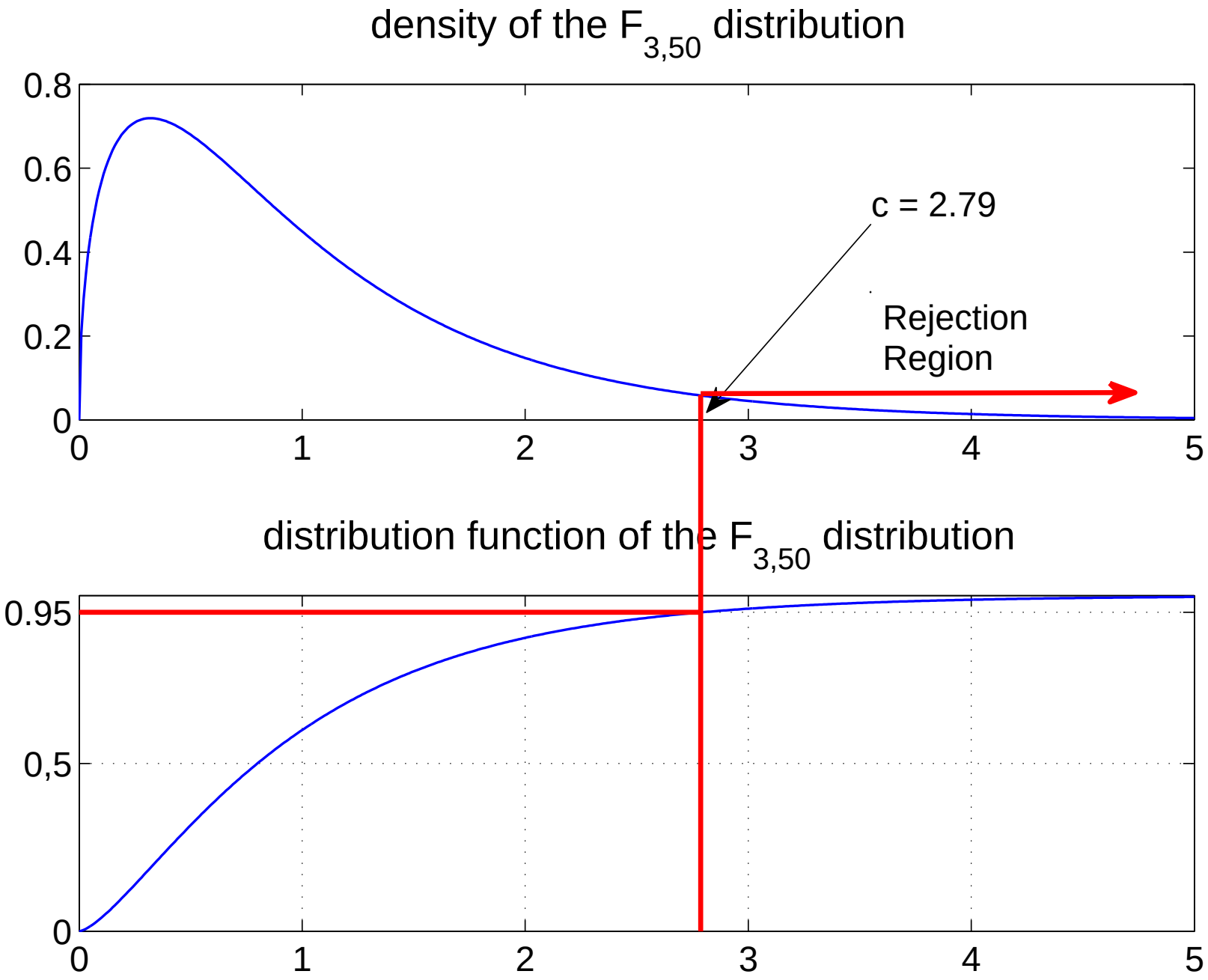
- Under H_0 and the CLM assumptions, the F statistic has an F distribution with q numerator and $n-k-1$ denominator degrees of freedom, written as

$$F \sim F_{q,n-k-1}. \tag{21}$$

The F distribution

Table 1: 95% Quantiles of the F distribution (= 5% critical values) (ν_1 numerator degrees of freedom; ν_2 denominator degrees of freedom)

ν_2/ν_1	2	3	4	5	6	7	8	9	10
10	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782
15	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876	2.5437
20	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928	2.3479
25	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821	2.2365
30	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107	2.1646
35	3.2674	2.8742	2.6415	2.4851	2.3718	2.2852	2.2167	2.1608	2.1143
40	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240	2.0772
45	3.2043	2.8115	2.5787	2.4221	2.3083	2.2212	2.1521	2.0958	2.0487
50	3.1826	2.7900	2.5572	2.4004	2.2864	2.1992	2.1299	2.0734	2.0261
60	3.1504	2.7581	2.5252	2.3683	2.2541	2.1665	2.0970	2.0401	1.9926
70	3.1277	2.7355	2.5027	2.3456	2.2312	2.1435	2.0737	2.0166	1.9689
80	3.1108	2.7188	2.4859	2.3287	2.2142	2.1263	2.0564	1.9991	1.9512
90	3.0977	2.7058	2.4729	2.3157	2.2011	2.1131	2.0430	1.9856	1.9376
100	3.0873	2.6955	2.4626	2.3053	2.1906	2.1025	2.0323	1.9748	1.9267
∞	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799	1.8307


density of the $F_{3,50}$ distribution
0.8
0.6
0.4
0.2
0
c = 2.79
Rejection
Region
0
1
2
3
4
5
distribution function of the $F_{3,50}$ distribution
0.95
0,5
0
0
1
2
3
4
5

The F distribution

- For $\nu_1 = 1$, the critical values are equal to the squares of a two–sided t–test with ν_2 degrees of freedom.
- This is because the square of a t_ν random variable has an $F_{1,\nu}$ distribution.
- Thus, for a single linear restriction, the F test gives rise to the same results as a two–sided t test.
- Clearly t tests are more flexible in such situations, as they allow for one–sided alternatives also.
- Now we return to the baseball problem.
- We have $q = 3$ and $n - k - 1 = 347$.
- So the critical values at the 5% and 1% levels are $c_{0.05} = 2.6049$ and $c_{0.01} = 3.782$, respectively.

- The F statistic is given by

$$F = \frac{\text{SSR}_r - \text{SSR}_{ur}}{\text{SSR}_{ur}} \frac{n-k-1}{q} = \frac{198.3115 - 183.1863}{183.1863} \times \frac{347}{3} = 9.5503. \tag{22}$$

- Obviously,

$$F > c_{0.01},$$

  so we can reject the null hypothesis that the performance variables have no effect on salary.

- We can say that x_3, x_4, x_5 are **jointly statistically significant**.

- As in this example, the F statistic is often useful for testing exclusion of a group of variables when these variables are highly correlated, and the partial effects are difficult to identify.

Calculating the p–value

- Calculating and reporting p–values for F tests may be extra useful since the critical values, depending on two parameters, are less well–known than those of the Gaussian.

- In the current context, the p–value is the probability of observing, under the null, a value at least as large the actually observed value of the F statistic, i.e.,

$$p - \text{value} = \Pr(F_{q,n-k-1} \geq F), \tag{23}$$

where $F_{q,n-k-1}$ is an F random variable with q and $n-k-1$ degrees of freedom and F is the observed value of the F ratio.

R^2 Form of the F Statistic

- Recall that the coefficient of determination

$$R^2 = 1 - \frac{\text{SSR}}{\text{SST}}, \tag{24}$$

where $\text{SST} = \sum_i (y_i - \bar{y})^2$ is total sum of squares.

Thus,

$$\begin{aligned} \text{SSR}_r &= SST \times (1 - R_r^2) && (25) \\ \text{SSR}_{ur} &= SST \times (1 - R_{ur}^2), && (26) \end{aligned}$$

and we can write

$$F = \frac{\text{SSR}_r - \text{SSR}_{ur}}{\text{SSR}_{ur}} \frac{n-k-1}{q} = \frac{R_{ur}^2 - R_r^2}{1 - R_{ur}^2} \times \frac{n-k-1}{q},$$

which is referred to as the R^2 form of the F statistic.

- For the baseball example,

$$\frac{R^2_{ur} - R^2_r}{1 - R^2_{ur}} \times \frac{n-k-1}{q} = \frac{0.6278 - 0.5971}{1 - 0.6278} \times \frac{347}{3} = 9.5405,$$

where the difference to (22) is due to rounding error.

- If the null hypothesis states that

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0, \tag{27}$$

then the restricted model is simply

$$\begin{aligned} y &= \beta_0 + u, \quad \widehat{\beta}_0 = \frac{1}{n}\sum_i y_i = \bar{y}, \quad \sum_i \widehat{u}_i^2 = \sum_i (y_i - \bar{y})^2, \\ &\Rightarrow R^2_r = 0, \end{aligned}$$

and the corresponding F statistic is

$$F = \frac{R^2}{1 - R^2} \times \frac{n-k-1}{k}. \tag{28}$$

- This is referred to as the F **statistic for the overall significance of a regression**, since it tests whether any of the independent variables helps to explain y.

General Linear Restrictions

- So far we have concentrated on exclusion restrictions, where it is tested whether a set of variables has no impact.

- More general linear hypothesis can also be tested.

- For example, suppose that, in the baseball example, we are interested in testing

$$H_0 : \beta_3 = 0, \quad \beta_3 = \beta_4 = 0.01.$$

- Then we can write the restricted model as

$$\begin{aligned} \log(salary) &= \beta_0 + \beta_1 years + \beta_2 gamesyr \quad (29) \\ &\quad +0.01(hrunsyr + rbisyr) + u, \\ \underbrace{\log(salary) - 0.01(hrunsyr + rbisyr)}_{=\tilde{y}} &= \beta_0 + \beta_1 years + \beta_2 gamesyr + u, \end{aligned}$$

  which can be estimated by regressing the left–hand side on the right–hand side in the third line in (29).

- Doing so results in

$$\begin{aligned} \widehat{\widetilde{y}} &= \underset{(0.1043)}{11.3906} + \underset{(0.0120)}{0.0696}\, years + \underset{(0.0013)}{0.0138}\, gamesyr \qquad (30) \\ n &= 353, \quad \text{SSR} = \sum_i \widehat{u}_i^2 = 184.0202, \quad R^2 = 0.4740. \end{aligned}$$

- We calculate the two forms of the F statistic,

$$F_1 = \frac{184.0202 - 183.1863}{183.1863} \times \frac{347}{3} = 0.5265, \qquad (31)$$

and

$$F_2 = \frac{0.6278 - 0.4740}{1 - 0.6278} \times \frac{347}{3} = 47.7956. \qquad (32)$$

- Now the answers in (31) and (32) are radically different. What has happened?

- Obviously, the reason is that the dependent variable $\tilde{y}$ in (30) is not the same as $\log(salary)$.

- So if we estimate the restricted model by substituting our restrictions, we have to keep in mind that regression output refers to the transformed data.

- If we calculate the R^2 of the restricted regression properly by calculating

$$R_r^2 = 1 - \frac{184.0202}{\underbrace{492.1755}_{=n \times s^2_{\log(salary)}}} = 0.6261, \tag{33}$$

we get

$$F_2 = \frac{0.6278 - 0.6261}{1 - 0.6278} \times \frac{347}{3} = 0.5283. \tag{34}$$

- Thus, we cannot use the regression output (the reported R^2 of the restricted regression) directly.