

Econ 140 - Spring 2016

Section 7

GSI: Fenella Carpena

March 10, 2016

1 Hypothesis Testing in MRM: Overview

| Type | Example | Test Statistic |
|--|--|---|
| 1. Individual One restriction involving one parameter | $H_0 : \beta_1 = 4$ $H_1 : \beta_1 \neq 4$ | t -stat |
| 2. Joint Multiple restrictions | $H_0 : \beta_1 = 0, \beta_2 = 0$ $H_1 : \beta_1 \neq 0$ and/or $\beta_2 \neq 0$ | F -stat (cannot use t-stat) |
| 3. Linear Linear combination of coefficients | $H_0 : 4\beta_1 + 2\beta_2 = 5$ $H_1 : 4\beta_1 + 2\beta_2 \neq 5$ or $H_0 : \beta_1 - \beta_2 = 0$ $H_1 : \beta_1 - \beta_2 \neq 0$ | (1) F -stat, (2) t -stat (by first transforming the regression) |

2 Hypothesis Testing in MRM: Single Coefficients and Joint Tests

Suppose we have the following model that explains baseball players' salaries.

$$salary_i = \beta_0 + \beta_1 years_i + \beta_2 gamesyr_i + \beta_3 bavg_i + \beta_4 hrunsyr_i + \beta_5 rbisyr_i + u_i \quad (1)$$

where for each player i , $salary$ is the salary in 1993, $years$ is years in the league, $gamesyr$ is average games played per year, $bavg$ is the career batting average, $hrunsyr$ is the number of home runs per year, $rbisyr$ is runs batted in per year. Further, suppose that we estimated the above equation using data we have on hand, and that we obtained the following regression results

$$\widehat{salary} = 11.10 + 0.0689 \cdot years + 0.0126 \cdot gamesyr + 0.00098 \cdot bavg + 0.0144 \cdot hrunsyr + 0.0108 \cdot rbisyr$$

(0.29)
(0.0121)
(0.0026)
(0.0010)
(0.0161)
(0.0072)

$$N = 353, SSR = 183.186, R^2 = 0.6278$$

Example 2.1. What test statistic would we use to test the hypothesis $H_0 : \beta_4 = 0, H_1 : \beta_4 \neq 0$? Carry out this test at the 5% level?

We would calculate the t -statistic, as we have learned before in this class. So t -stat = $(\hat{\beta}_4 - 0)/SE(\hat{\beta}_4) = 0.0144/0.0161 \approx 0.894$. Since $|0.894| < 1.96$; we fail to reject the null hypothesis.

Example 2.2. What test statistic would we use to test the hypothesis $H_0 : \beta_2 = 0.1, H_1 : \beta_2 \neq 0.1$? Carry out this test at the 10% level.

Again, we would calculate the t -statistic. That is, t -stat = $(\hat{\beta}_2 - 0.1)/SE(\hat{\beta}_2) = (0.0126 - 0.1)/0.0026$. Since $|t - stat| > 1.96$, we reject the null.

Example 2.3. A sports analyst hypothesizes that once years in the league and games per year have been controlled for, the variables *bavg*, *hrunsyr*, and *rbisyr* (which we can think of as measure of performance) have no effect on salary.

(a) What is the null and alternative hypothesis?

$H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$ vs. $H_1 : \text{at least one of } \beta_3, \beta_4, \beta_5 \text{ is not equal to } 0$

(b) What test statistic would we use to test the hypothesis in part (a)? Assuming that the population errors are homoskedastic, what is the formula for this test statistic?

We would use the F -stat. The formula is:

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

or alternatively,

$$F - \text{stat} = \frac{(R_{ur}^2 - R_r^2)/q}{(1 - R_{ur}^2)/(n - k - 1)}$$

(c) What is the restricted regression?

$\text{salary}_i = \beta_0 + \beta_1 \text{years}_i + \beta_2 \text{gamesyr}_i + u_i$

(d) What is the unrestricted regression?

$\text{salary}_i = \beta_0 + \beta_1 \text{years}_i + \beta_2 \text{gamesyr}_i + \beta_3 \text{bavg}_i + \beta_4 \text{hrunsyr}_i + \beta_5 \text{rbisyr}_i + u_i$

(e) What is q ?

$q = 3$

(f) What is n ?

$n = 353$

(g) What is k ?

$k = 5$

(h) Suppose that a regression of *salary* on *years* and *gamesyr* yielded an SSR of 198.311. Calculate the F -statistic.

$$F - \text{stat} = \frac{198.311 - 183.186}{183.186} \cdot \frac{353 - 5 - 1}{3} \approx 9.55.$$

(i) Find the critical value from the F -distribution.

critical value = 2.60

(j) What is the conclusion of your hypothesis test?

Reject H_0 since $9.55 > 2.60$

Example 2.4. Let us consider again the joint hypothesis

$H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$ vs. $H_1 : \text{at least one of } \beta_3, \beta_4, \beta_5 \text{ is not equal to } 0$

that we tested in Example 2.3 using an F -test. Is it possible to carry out this joint hypothesis test using the 3 t -statistics from the following 3 individual tests: (1) $H_0 : \beta_3 = 0, H_1 : \beta_3 \neq 0$; (2) $H_0 : \beta_4 = 0, H_1 : \beta_4 \neq 0$; (3) $H_0 : \beta_5 = 0, H_1 : \beta_5 \neq 0$?

Testing each coefficient individually is **not appropriate**, because we want to look at all 3 coefficients simultaneously. Therefore, what we need is the **joint distribution** of $\hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5$. If we looked at each of these one at a time by looking at the t -statistic, we will not be putting any restriction on the other parameters. Note that if you looked at the t -stat of $\hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5$ individually, you will see that each has a t -stat that is less than 1.96, which might lead you to conclude that we would fail to reject the joint hypothesis test. But as we saw in Example 2.3, this conclusion turns out to be wrong.

Example 2.5. Looking back at Example 2.3, we found that we rejected the joint hypothesis that *bavg*, *hrunsyr*, *rbisy* have no effect on salary. But if we had looked at each of these variables individually, we would have failed to reject each null hypothesis separately because the individual t-stats are less than 1.96. What might explain the difference in these results?

One possibility is that there is imperfect multicollinearity between these three variables. Imperfect multicollinearity makes it difficult to estimate their coefficients precisely (why? recall again from last section's material) resulting in a low t-stat. Since the F-stat tests whether *bavg*, *hrunsyr* and *rbisy* are jointly different from zero, the high correlation between the three variables becomes less relevant.

3 Hypothesis Testing in MRM: Linear Restrictions

Example 3.1. (Adapted from Stock and Watson, Exercise 7.9) Consider the regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, and the hypothesis test $H_0 : \beta_1 = \beta_2$, $H_1 : \beta_1 \neq \beta_2$.

(a) What test statistics can we use to carry out the above hypothesis test?

We can use either an *F*-stat or a *t*-stat

(b) Describe how you would calculate the *F*-statistic (under the assumption of homoskedasticity). What is the restricted and unrestricted regression?

We would use the *F*-statistic formula using the SSR of the restricted regression and the unrestricted regression, n , k , and q . The restricted regression is $Y_i = \beta_0 + \beta_1(X_{1i} + X_{2i}) + u_i$. The unrestricted regression is $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. Here, $q = 1$ and $k = 2$.

(c) Describe how you can use a *t*-statistic to test $H_0 : \beta_1 = \beta_2$, $H_1 : \beta_1 \neq \beta_2$. Specifically, transform the regression so that you can use a *t*-statistic to carry out the test.

Note that the regression can be re-written as $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i + [\beta_2 X_{1i} - \beta_2 X_{1i}] \implies Y_i = \beta_0 + (\beta_1 - \beta_2) X_{1i} + \beta_2 (X_{2i} + X_{1i}) + u_i$. So we can regress Y on X_1 and W , where $W = X_2 + X_1$, and use a t-stat to test that the coefficient on X_1 is zero.

Example 3.2. In the same regression as in in Example 3.1, transform the regression so that you can use a *t*-statistic to test $\beta_1 + 2\beta_2 = 0$.

Note that the regression can be re-written as $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i + [2\beta_2 X_{1i} - 2\beta_2 X_{1i}] \implies Y_i = \beta_0 + (\beta_1 + 2\beta_2) X_{1i} + \beta_2 (X_{2i} - 2X_{1i}) + u_i$. So we can regress Y on X_1 and Z , where $Z = X_2 - 2X_1$, and use a t-stat to test that the coefficient on X_1 is zero.

4 Hypothesis Testing in MRM: Stata

Example 4.1. Suppose we have 1980 census data on the 50 states recording the population size in each state (`pop`), the median age (`medage`), the number of deaths (`death`), the number of marriages, (`marriage`), and the number of divorces (`divorce`). We estimate the following regression:

```
. reg pop medage death marriage divorce
```

| Source | SS | df | MS | | | |
|----------|------------|----|------------|-----------------|---------|--|
| Model | 1.0800e+15 | 4 | 2.7000e+14 | Number of obs = | 50 | |
| Residual | 9.3500e+12 | 45 | 2.0778e+11 | F(4, 45) = | 1299.46 | |
| | | | | Prob > F = | 0.0000 | |
| | | | | R-squared = | 0.9914 | |
| | | | | Adj R-squared = | 0.9907 | |
| Total | 1.0893e+15 | 49 | 2.2232e+13 | Root MSE = | 4.6e+05 | |

| | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] | |
|----------|-----------|-----------|-------|-------|----------------------|-----------|
| pop | | | | | | |
| medage | -181303.8 | 43749.97 | -4.14 | 0.000 | -269420.8 | -93186.87 |
| death | 91.30243 | 4.137673 | 22.07 | 0.000 | 82.96873 | 99.63613 |
| marriage | 1.80206 | 4.303597 | 0.42 | 0.677 | -6.865829 | 10.46995 |
| divorce | 39.80303 | 8.146704 | 4.89 | 0.000 | 23.39473 | 56.21134 |
| _cons | 5241295 | 1272002 | 4.12 | 0.000 | 2679350 | 7803239 |

What Stata commands would you use to test the following hypothesis?

- Test the individual hypothesis that the coefficient on `medage` is zero.
`test medage = 0`. If you did this command in Stata, you would get an F -statistic of 17.17. Note that in the case of an individual hypothesis test, the F -stat = $(t\text{-stat})^2$. From the regression table above, the t -stat is -4.14 , and you can verify that -4.14^2 is about 17.17 (the differences are due to rounding error).
- Test the joint hypothesis that the coefficients on all four regressors are simultaneously zero.
`test (medage = 0) (death = 0) (marriage = 0) (divorce = 0)`. If you did this command in Stata, you would get an F -stat of 1299.46. Notice that this is the same F -stat that is in the upper right of the regression table, which says $F(4, 45) = 1299.46$. This is not a coincidence, because the F -stat in the upper right of the regression table is for the joint hypothesis that all regressors are zero.
- Test the linear hypothesis that the coefficient on `death` minus the coefficient on `marriage` is zero.
`test death - marriage = 0`. As an aside, note that we could also test any linear combination of coefficients, for example, `test 2*death + 4*marriage = 6`.

5 Additional Exercise: Spring 2014 MT2, Question 4

In this exercise we use a dataset containing information on 269 NBA basketball players including their salaries. Table 1 below (see next page) shows the results from 3 OLS regressions where heteroskedasticity-robust standard errors are given in square brackets. The dependent variable is `salary` (annual salary in thousands \$), and the explanatory variables are `points` (average points per game), `rebounds` (average rebounds per game), and `assists` (average assists per game).

- [6] What is the interpretation of the coefficient on `points` in all three regressions? Give the meaning of the OLSE of the points coefficient from the third regression.
 The coefficient for the multivariate regressions is $\beta_1 = \Delta wage / \Delta points$ under the assumption that all other variables are held constant. An additional point per game (on average for all games played) will fetch the NBA player, on average, over \$80,000 in annual salary – holding rebounds and assists constant.

Table 1: Regressions on NBA Salaries

| Variables | Model 1 | Model 2 | Model 3 |
|-----------|---------------------|--------------------|--------------------|
| points | 111.67 [8.18]** | 87.72 [9.43]** | 80.67 [11.20]** |
| rebounds | | 86.44 [20.62]** | 93.36 [21.25]** |
| assists | | | 26.08 [21.56] |
| _cons | 278.10 [83.62]** | 137.56 [83.09] | 115.02 [86.09] |
| R2 | 0.43 | 0.47 | 0.48 |
| N | 269 | 269 | 269 |

Notes: * p-value<0.05; ** p-value<0.01.

- (b) [8] When the **rebounds** variable is added, both the R2 and the Adjusted R2 increase whereas when additionally **assists** variable is added, only the R2 increases. Does this indicate that **rebounds** should be in the regression and that **assists** should not? Explain.

R2 tells us the percent of variation in our dependent variable that is explained by the regression model. However, this measure should not solely determine whether or not to include a variable in a regression - particularly when we are concerned with determining causal effects or are attempting to control for omitted variables. R2 necessarily increases each time a variable is added so that a small increase in R2 conveys no information. Adjusted R2 does dip slightly when assists was added to the regression and so adjusted R2 is doing its job of penalizing for adding a regressor. Note that the coefficient on rebounds was significantly different from zero whenever it appeared in a regression, whereas the coefficient on assists was not. Now, assists could be correlated with other measures of on-court performance. Apparently it is not correlated enough to impact standard errors of coefficients on points and rebounds. However, the criterion for including a variable should be based on the theory. Here that would be performance on the court translates into wins which translates into ticket sales and broadcast rights.

- (c) [5] We wonder whether a player's position matters, thinking that different positions may be more valuable to a team than others. To investigate, we regress **salary** on dummy indicators of each of the three possible court positions a player can have, **center**, **forward**, and **guard** (and a constant), with no other explanatory variables. Stata refuses to report OLSEs for all three regressor. Why does this happen? How would you solve this problem?

A player is classified as having one of the three positions and so the three dummy variables add up to one, i.e., equal to the constant "variable". Hence, we have perfect multicollinearity and an example of "the dummy variable trap". To escape this trap, exclude one of the three positions, e.g., center. Stata will do this by itself, excluding the first of the offending variables in alphabetical order. Then the coefficients on the other two positions give the increment in salary relative to the average center.

- (d) [3] Based on the results, it seems that the marginal effect of **rebounds** on salaries is higher than the marginal effect of **points**. So immediately after running the regression for model 3 we issue the command: **test rebounds=points**. What does this Stata command do?

This command performs a test of the hypothesis that coefficients on **points** and **rebounds** are equal.

- (e) [7] Stata generates output from the command in (d) that includes an F-statistic. Suppose that it takes the value: 4.60. Decide whether you reject the null at the 1% level of significance.

The 1% critical value of the F-stat with 1 and 265 degrees of freedom is very close to $F(1, \infty) = 6.6349$ from the F-table. Since our F-stat of 4.60 is less than 6.6349 we cannot reject the null that the 2 coefficients are equal at 1% significance level.

