Lecture 6: Further Inference in the Multiple Regression Model

Shuo Liu

UCLA Summer School Econ 103

July 18, 2017

Outline

1. Joint Hypothesis Testing

2. Model Specification

Joint Hypothesis Testing

Joint Hypothesis Testing tests a null hypothesis with **multiple conjectures**, expressed with more than one "equal sign";

- **Example**: should a group of explanatory variables $\{x_3, x_4, x_5\}$ be included in a particular model?

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + e \tag{1}$$

- **Test Form**:

$$H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0 \tag{2}$$

$$H_1 : \text{“}\beta_3 = 0, \beta_4 = 0, \beta_5 = 0\text{” do not hold simultaneously} \tag{3}$$

- A joint test for whether all the three conjectures hold simultaneously

Joint Hypothesis Testing

SALES example:

- Consider the model:

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + \beta_4 ADVERT^2 + e \tag{4}$$

- Test whether or not advertising has an effect on sales: If advertising does not have effect

$$H_0 : \beta_3 = 0, \beta_4 = 0 \tag{5}$$

- If advertising has effect on sales

$$H_1 : \text{“}\beta_3 \neq 0, \beta_4 = 0\text{” or “}\beta_3 = 0, \beta_4 \neq 0\text{” or “}\beta_3 \neq 0, \beta_4 \neq 0\text{”} \tag{6}$$

- Relative to the null hypothesis H_0, the original model (4) is called **unrestricted model**, where the restrictions in H_0 have not been imposed on the original model (4);
- **Restricted model**: impose restrictions in H_0 to the original model (or assume parameter restrictions in H_0 are true)

$$SALES = \beta_1 + \beta_2 PRICE + e \tag{7}$$

Joint Hypothesis Testing

How to determine which model is better (a little bit like choosing from two specific model forms)?

- We use $SSE = \sum_{i=1}^{n} \hat{e_i}^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, where $\hat{y}_i$ is determined by the model form you choose (either restricted form or unrestricted form);
- We compare sum of squares of error (residuals) from unrestricted model SSE_U with that from restricted model SSE_R. Intuitively, we will choose unrestricted model form (H_1) only when SSE_U is "too smaller" than SSE_R, because:

$$SST = SSR + SSE \tag{8}$$

  and when throwing in more explanatory variables into the model, SSR will always increase, then we always have $SSE_U \leq SSE_R$.
- We use F test to test $H_0 : \beta_3 = 0, \beta_4 = 0$.

Joint Hypothesis Testing

- Remember F test is always one-tail (right-tail) test!
- **First step**: calculate F-statistic

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(n-K)} \tag{9}$$

  where

  1. $J =$ number of restrictions;
  2. n is sample size and K is number of parameters in the original model (unrestricted model).

- **Second step**: determine the distribution of F-statistic above under H_0 is true

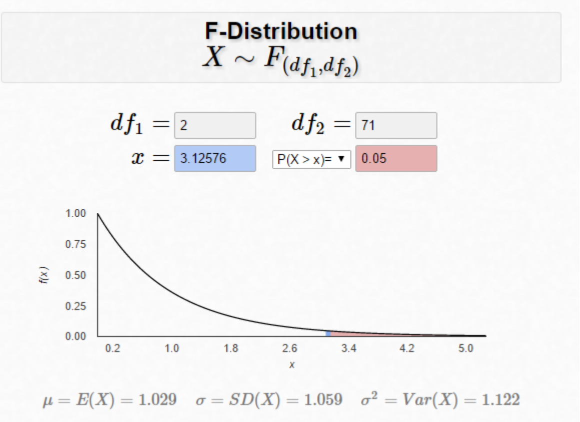
$$F \sim F_{(J,n-K)} \tag{10}$$

  F distribution with J degree of freedom in the numerator and $n-K$ degree of freedom in the denominator;

- Next, given significance level, we will reject H_0 only when F-statistic is "too large", i.e. SSE_U is "too smaller" than SSE_R, which means unrestricted model is much better than restricted model.

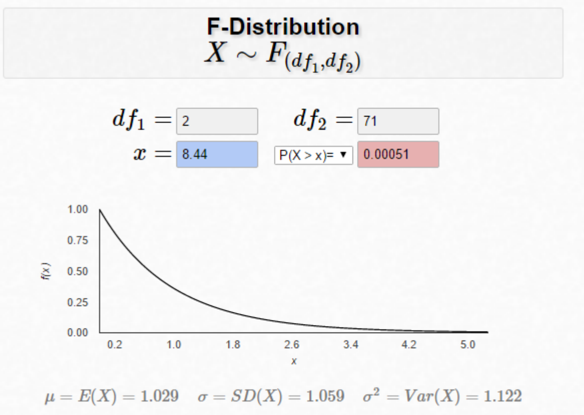
- **Third step**: given significance level α

  Option 1: find critical value with right tail probability equal to α and set rejection region;



- **Third step**: given significance level α

  Option 2: use F-statistic calculated in the first step to calculate corresponding p-value $p = P(F_{(J,n-K)} > F)$, then compare p-value with α.



Joint Hypothesis Testing

- State the conclusion: suppose $F = 8.44$, $J = 2$, $n - K = 71$, $\alpha = 0.05$ then $F_c = 3.126$, since

$$8.44 = F > F_c = F_{(1-\alpha, J, n-K)} = F_{(0.95,2,71)} = 3.126 \tag{11}$$

  we reject the null hypothesis that both $\beta_3 = 0$ and $\beta_4 = 0$, and conclude that at least one of them is not equal to zero.

- Then go back to the example, we conclude that advertising does have a significant effect on sales revenue.

Joint Hypothesis Testing

In STATA output table, consider the general multiple regression model with $K-1$ explanatory variables and K unknown parameters,

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_K x_K + e \tag{12}$$

- **To examine whether we have a viable model,** STATA automatically does the following hypothesis testing:

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_K = 0 \tag{13}$$

$$H_1 : \text{ at least one } \beta_k \neq 0, k = 2, 3, \cdots, K \tag{14}$$

- This is referred to as a test of overall significance of the regression model;
- We use F test to test the above H_0 against H_1.

Joint Hypothesis Testing

- The unrestricted model is equation (12).
- Assuming H_0 is true, the restricted model becomes:

$$y = \beta_1 + e \tag{15}$$

then the OLS estimator of β_1 in the restricted model is:

$$b_1^* = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} \tag{16}$$

and

$$SSE_R = \sum_{i=1}^{n}(y_i - \hat{y_i})^2 = \sum_{i=1}^{n}(y_i - b_1^*)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 = SST \tag{17}$$

- Thus to test the overall significance of a model (not in general, only for multiple regression model), the F-statistic can be modified and written as:

$$F = \frac{(SSE_R - SSE_U)/(K-1)}{SSE_U/(n-K)} = \frac{(SST - SSE)/(K-1)}{SSE/(n-K)} \tag{18}$$

Joint Hypothesis Testing

$$FOODEXP = \beta_1 + \beta_2 INCOME + e \tag{19}$$

For joint hypothesis testing to test the overall significance of the model, what is J?

```
. reg food_exp income
```

Source	SS	df	MS
Model	190626.984	1	190626.984
Residual	304505.176	38	8013.2941
Total	495132.16	39	12695.6964

Number of obs =	40
F(1, 38) =	23.79
Prob > F =	0.0000
R-squared =	0.3850
Adj R-squared =	0.3688
Root MSE =	89.517

food_exp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	10.20964	2.093264	4.88	0.000	5.972052	14.44723
_cons	83.416	43.41016	1.92	0.062	-4.463279	171.2953

$SST = 495132.16$, $SSE = 304505.176$

Joint Hypothesis Testing

- F-statistic:

$$F = \frac{(SST - SSE)/(K-1)}{SSE/(n-K)} = \frac{(495132.16 - 304505.176)/(2-1)}{304505.176/(40-2)} = 23.7888 \quad (20)$$

- **Option 2**: The STATA calculates p-value as very close to zero, given significance level $\alpha = 0.05$, then we reject H_0.
- **Option 1**: How can we check $F_c = F_{(0.95,1,38)}$?

Joint Hypothesis Testing

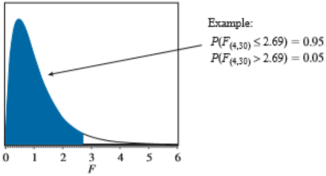


Table 4 **95th Percentile for the *F*-distribution**

v_2/v_1	1	2	3	4	5	6	7
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25

Joint Hypothesis Testing

Connection between F test and t test: when testing single hypothesis (focus on the significance of only one parameter), F test and t test are equivalent.

Suppose now we want to test whether PRICE affects SALES, that is for:

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + \beta_4 ADVERT^2 + e \tag{21}$$

we want to test

$$H_0 : \beta_2 = 0 \tag{22}$$

$$H_1 : \beta_2 \neq 0 \tag{23}$$

or restricted model is:

$$SALES = \beta_1 + \beta_3 ADVERT + \beta_4 ADVERT^2 + e \tag{24}$$

with F-statistic equal to 53.355.

Joint Hypothesis Testing

Fitted model:

$$\underset{(Se)}{\widehat{SALES}} = \underset{(6.8)}{109.72} - \underset{(1.046)}{7.64}\, PRICE + \underset{(5.556)}{12.15}\, ADVERT - \underset{(0.941)}{2.77}\, ADVERT^2 \quad (25)$$

Use t test:

$$t = \frac{-7.64}{1.046} \quad (26)$$

and

$$t^2 = \left(\frac{-7.64}{1.046}\right)^2 = 53.35 = F \quad (27)$$

Then

$$\text{t is either too large or too small} \iff \text{F is too large} \quad (28)$$

So for single hypothesis testing, **two-tail** t test and F test are consistent. Actually in lecture 1, we have: $F_{(1,m)} = t_m^2$, for $\forall m$. In this case, $m = n - 2$.

How about the single hypothesis test on linear combination of parameters?

- Consider testing the following claim: the marginal sales of advertising when advertising expenditure is \$1900/month is equal to \$1 (assume unit of ADVERT is \$1000 and marginal cost of advertising is \$1), which means \$1900/month is the optimal advertising expenditure

$\Longrightarrow$

$$H_0 : \beta_3 + 2\beta_4 ADVERT|_{ADVERT=1.9} = 1 \tag{29}$$

$\Longrightarrow$

$$H_0 : \beta_3 + 3.8\beta_4 = 1 \tag{30}$$

$$H_1 : \beta_3 + 3.8\beta_4 \neq 1 \tag{31}$$

- We already learned how to use **t test** to test the above H_0 against H_1.

Joint Hypothesis Testing

- We can also use **F test** to test the above H_0 (restricted model) against H_1 (unrestricted model).
- **When H_0 applies, the restricted model is:**

$$SALES = \beta_1 + \beta_2 PRICE + (1 - 3.8\beta_4) ADVERT + \beta_4 ADVERT^2 + e \quad (32)$$

$$\Longrightarrow$$

$$SALES - ADVERT = \beta_1 + \beta_2 PRICE + \beta_4 (ADVERT^2 - 3.8 ADVERT) + e \quad (33)$$

- The F-statistic is:

$$F = \frac{(SSE_R - SSE_U)/1}{SSE_U/(n-K)} = 0.9362 \quad (34)$$

suppose $\alpha = 0.05$, $F_c = 3.976$, then $F < F_c$, we cannot reject H_0.

- We conclude an advertising expenditure of \$1900/month is optimal.

Joint Hypothesis Testing

- Suppose now we have the following hypothesis:

$$H_0 : \beta_3 + 3.8\beta_4 \leq 1 \tag{35}$$

$$H_1 : \beta_3 + 3.8\beta_4 > 1 \tag{36}$$

- In this case, we can no longer use F test.
- Because $F_{(1,n-K)} = t^2_{n-K}$ cannot distinguish between the left and right tails as needed for a one-tail test.
- When we have alternative hypothesis H_1 containing inequality signs $\leq$, $\geq$, we restrict to t-test. (test using t statistic and t distribution)

Model Specification

In any econometric investigation, choice of the model is one of the first steps

- What are the important considerations when choosing a model?
- What are the consequences of choosing the wrong model?
- Are there ways of assessing whether a model is adequate?

We have already learned some ways to evaluate a model: significance separately or jointly for parameters, R^2 to measure the goodness-of-fit.

Omitted Variable Bias

- It is possible that a chosen model may have important variables omitted, possibly because the economic theory has overlooked a variable or the lack of data makes us drop a variable even when it is prescribed by economic theory.

$$SALES = \beta_1 + \beta_2 PRICE + e \tag{37}$$

Model Specification

Consider the following model:

$$FAMINC = \beta_1 + \beta_2 HEDUC + \beta_3 WEDUC + e \tag{38}$$

where FAMINC is family income, HEDUC is husband's education, WEDUC is wife's education.

- The estimated model is:

$$\underset{(Se)}{\widehat{FAMINC}} = \underset{(11230)}{-5534} + \underset{(803)}{3132} HEDUC + \underset{(1066)}{4523} WEDUC \tag{39}$$

- If we incorrectly omit WEDUC,

$$\underset{(Se)}{\widehat{FAMINC}} = \underset{(8541)}{-26191} + \underset{(658)}{5155} HEDUC \tag{40}$$

- **Omitting WEDUC leads us to overstate the effect of an extra year of husband's education on family income by about** \$2000
- Then omission of a relevant variable leads to an estimator that is biased, which is called **omitted-variable bias**.

Model Specification

More generally, write a general model as:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + e \tag{41}$$

- Omitting x_3 is equivalent to imposing restriction $\beta_3 = 0$;
- It can be viewed as an example of imposing an incorrect constraint on the parameters;
- Suppose b_2^* is estimator of β_2 in the following model:

$$y = \beta_1 + \beta_2 x_2 + e \tag{42}$$

  then we analyze the bias of b_2^*.
- The bias is:

$$bias(b_2^*) = E(b_2^*) - \beta_2 = \beta_3 \frac{\widehat{Cov}(x_2, x_3)}{\widehat{Var}(x_2)} \tag{43}$$

  why $E(b_2^*) \neq \beta_2$ in this case?

Model Specification

Given

$$bias(b_2^*) = E(b_2^*) - \beta_2 = \beta_3 \frac{\widehat{Cov}(x_2, x_3)}{\widehat{Var}(x_2)} \tag{44}$$

- If wife's education has positive effect on family income: $\beta_3 > 0$;
- If wife's education is positively correlated with husband's education: $\widehat{Cov}(x_2, x_3) > 0$;
- Then we can conclude that the bias is positive, in words, the second estimated regression attributes too much to the husbands education because of the omission of the wife's education.

Model Specification

Table 6.1 **Correlation Matrix for Variables Used in Family Income Example**

	FAMINC	*HEDU*	*WEDU*	*KL6*	X_5	X_6
FAMINC	1.000					
HEDU	0.355	1.000				
WEDU	0.362	0.594	1.000			
KL6	−0.072	0.105	0.129	1.000		
X_5	0.290	0.836	0.518	0.149	1.000	
X_6	0.351	0.821	0.799	0.160	0.900	1.000

- KL6: number of kids lower than six years old;
- X_5 and X_6 are just another two economic variables possibly affecting family income and highly correlated with HEDU and WEDU.

Model Specification

Now consider the model:

$$\underset{(Se)}{\widehat{FAMINC}} = \underset{(11163)}{-7755} + \underset{(797)}{3212}HEDUC + \underset{(1061)}{4777}WEDUC - \underset{(5004)}{14311}KL6 \quad (45)$$

- Note that in this example the coefficient estimators for HEDUC and WEDUC have not changed too much, because $KL6$ is not highly correlated with those two education variables.

The more explanatory variables, the better?

- The presence of many explanatory variables may inflate the variances of the estimators because of multi-collinearity, remember $Var(b_2) = \frac{\sigma^2}{(1-r_{23}^2)\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)^2}$.
- Consider the following fitted model:

$$\underset{(Se)}{\widehat{FAMINC}} = \underset{(11195)}{-7755} + \underset{(1250)}{3340}HEDUC + \underset{(2278)}{5869}WEDUC - \underset{(5044)}{14200}KL6 \quad (46)$$

$$- \underset{(2242)}{889}X_5 + \underset{(1982)}{1067}X_6$$

The inclusion of irrelevant variables (X_5 and X_6) has reduced the precision of the coefficient estimators for other explanatory variables in the regression.

Model Specification

Some important points for choosing a model form:

- Choose explanatory variables and a functional form based on your theoretical and general understanding of the relationship;
- If a fitted model has estimators with unexpected signs, or unrealistic magnitudes, they could be caused by a mis-specification such as the omission of an important explanatory variable;
- One method for assessing whether one or a group of explanatory variables should be included in an equation is to perform significance tests, both separately and jointly.
- We already talked about how to modify the measure of goodness-of-fit to prevent adding as many explanatory variables as possible

$$\bar{R}^2 = 1 - \frac{SSE/(n-K)}{SST/(n-1)} \tag{47}$$

Model Specification

Table 6.2 **Goodness-of-Fit and Information Criteria for Family Income Example**

Included Variables	R^2	$\bar{R}^2$	AIC	SC
HEDU	0.1258	0.1237	21.262	21.281
HEDU, *WEDU*	0.1613	0.1574	21.225	21.253
HEDU, *WEDU*, *KL6*	0.1771	0.1714	21.211	21.248
HEDU, *WEDU*, *KL6*, *X5*, *X6*	0.1778	0.1681	21.219	21.276

- Selecting variables to maximize $\bar{R}^2$ can be viewed as selecting variables to minimize SSE, **subject to a penalty for introducing too many variables.**
- Both the other two information criteria: the **AIC** and the **SC**(BIC) work in a similar way, but with different penalties for introducing too many variables. (Both are positively correlated with SSE and positively correlated with K)

Model Specification

When would a model be mis-specified?

- We have omitted important explanatory variables;
- Included irrelevant ones;
- Chosen a wrong functional form;
- Have a model that violates the assumptions of the multiple regression model, most usually to violate the "no-exact-collinearity" assumption.
- Poor data quality, e.g. from uncontrolled experiment which will generate economic variables that move together in a systematic way

$$Var(b_2) = \frac{\sigma^2}{(1 - r_{23}^2) \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2} \tag{48}$$

Model Specification

Example:

- MPG=miles per gallen;
- CYL=number of cylinders;
- ENG= engine displacement in cubic inches
- WGT=vehicle weight in pounds

Regression of MPG on CYL is:

$$\underset{\substack{(Se)\\(p-value)}}{\widehat{MPG}} = \underset{\substack{(0.83)\\(0.000)}}{42.9} - \underset{\substack{(0.146)\\(0.000)}}{3.558}CYL \tag{49}$$

Now add ENG and WGT:

$$\underset{\substack{(Se)\\(p-value)}}{\widehat{MPG}} = \underset{\substack{(1.5)\\(0.000)}}{44.4} - \underset{\substack{(0.413)\\(0.517)}}{0.268}CYL - \underset{\substack{(0.0083)\\(0.125)}}{0.0127}ENG - \underset{\substack{(0.00071)\\(0.000)}}{0.0057}WGT \tag{50}$$

How to test collinearity?

- One simple way to detect collinear relationships is to use sample correlation coefficients between pairs of explanatory variables;
- However, in some cases, collinear relationships involve more than two of the explanatory variables, the collinearity may not be detected by examining pairwise correlations;
- Try an auxiliary model:

$$x_2 = a_1 x_1 + a_3 x_3 + a_4 x_4 + \cdots + a_K x_K + v \tag{51}$$

- If the R^2 (or adjusted $\bar{R}^2$) from this artificial model is high, e.g. above 0.8, then the implication is that a large portion of the variation in x_2 is explained by variation in the other explanatory variables.