Chapter 4

Multiple Linear Regression

Flow of study

- Multiple Linear Regression Model

Now we add more independent variables into our model. While the estimation method remains the same, the formulae to calculate the estimators are totally different.

- Individual Testing

Like what we studied in the previous chapter, can we still test each coefficients' significance?

- Analysis of Variance (ANOVA)

When we try to jointly test multiple variables at the same time, the F-test is a lot more useful and easier, with some caveats in many situation. This part is to showcase concepts underlying F-test mainly and how we can apply on so many useful test.

(1) *Function and estimation*

If we add more independent variable(s) into our model to increase fitness to the model, the specification of the stochastic form and the SRF become

- $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{u}_i$: Stochastic form
- $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$: SRF

Note that we are not going to use the notation X_{1i} to make our estimators number corresponded with the variable names.

Now $\hat{\beta}_2$ and $\hat{\beta}_3$ are known as '**partial slope coefficients**' since when we consider

- $\hat{\beta}_2$ represents the change in $\hat{Y}_i$ when X_{2i} increases for 1 unit, holding everything else constant.
- $\hat{\beta}_3$ represents the change in $\hat{Y}_i$ when X_{3i} increases for 1 unit, holding everything else constant.

(1) *Function and estimation*

If we add more independent variables into this model, the specification becomes

- $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_k X_{ki} + \hat{u}_i$

Let's focus on 2 independent variables model first, we follow the same procedure as when we did, minimizing the term $\sum \hat{u}_i^2$

- $\min_{\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3} \sum \hat{u}_i^2 = \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i}\right)^2$

Skipping all the prove because we utilize the same logic of minimization of a function with calculus, we get

(1) *Function and estimation*

Coefficients

- $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X}_2 - \hat{\beta}_3\bar{X}_3$
- $\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\sum x_{2i}x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2}$
- $\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i}x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2}$

Variance

- $\text{var}(\hat{\beta}_1) = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}_2^2\sum x_{3i}^2 + \bar{X}_3^2\sum x_{2i}^2 - 2\bar{X}_2^2\bar{X}_3^2\sum x_{2i}x_{3i}}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2}\right]$
- $\text{var}(\hat{\beta}_2) = \sigma^2\left[\frac{\sum x_{3i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2}\right] = \frac{\sigma^2}{\sum x_{2i}^2(1-r_{23}^2)}$
- $\text{var}(\hat{\beta}_3) = \sigma^2\left[\frac{\sum x_{2i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2}\right] = \frac{\sigma^2}{\sum x_{3i}^2(1-r_{23}^2)}$

where r_{23} is the coefficient of correlation between X_2 and X_3.

The estimator of σ^2 is $\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-3}$

(2) *Assumptions*

We need to add another assumption here, which is **no multicollinearity**, and still hold prior classical linear regression model assumptions true.

The assumption states that there is no significant linear relationship between each regressor. For instance,

- $X_{2i} = 2X_{3i}$

We then can turn this model into

- $Y_i = \hat{\beta}_1 + \hat{\beta}_2 2X_{3i} + \hat{\beta}_3 X_{3i} + \hat{u}_i$ and then $\hat{\beta}_2 = \hat{\beta}_3$ so
- $Y_i = \hat{\beta}_1 + (\hat{\beta}_2 + 2\hat{\beta}_3)X_{3i} + \hat{u}_i$

which means that either X_{2i} or X_{3i} does not add any more information to this model.

(3) *Adjusted Coefficient of Determination*

When we add more and more independent variables into the model, the coefficient of determination is increasing (at least it will not decrease) from decreasing $\sum \hat{u}_i^2$.

Recall that when we define R^2 as

- $R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2}$ where $\sum y_i^2 = \sum (Y_i - \bar{Y})^2$

Now we define **adjusted R^2**, denoted as

- $\bar{R}^2 = 1 - \frac{\sum \hat{u}_i^2/(n-k)}{\sum y_i^2/(n-1)} = 1 - (1 - R^2)\frac{n-1}{n-k}$

The word **adjusted** means that the R^2 is adjusted by the degrees of freedom associated with the sum of the squares entering the specification.

(3) *Adjusted Coefficient of Determination*

Comparison between R^2 and $\bar{R}^2$

(1) $0 \leq R^2 \leq 1$ but $\bar{R}^2$ can be negative (interpreted as 0).

(2) As k increases, R^2 is increasing but $\bar{R}^2$ may not.

(3) $\bar{R}^2 < R^2$

(4) *Examples*

Cobb-Douglas Production Function

Usual form of the Cobb-Douglas function is

- $Y = AK^{\alpha}L^{\beta}$

where Y is the value of output of an economy,

A is total factor productivity (TFP or sometimes simplified as production technology),

K is number of capital input,

L in number of labor input,

α and β is the output elasticity.

The stochastic form can be expressed as

- $Y_i = AK_i^{\alpha}L_i^{\beta}e^{u_i}$

Taking natural logarithm to enable linear estimation yields

- $\ln Y_i = \ln A + \alpha \ln K_i + \beta \ln L_i + u_i$

(4) *Examples*

Cobb-Douglas Production Function

The data obtained from manufacturing sector of all states in the US, represented for each observation i. The results of linear regression is as follows.

- $\ln \hat{Y}_i = 3.8876 + 0.5213 \ln K_i + 0.4683 \ln L_i$

  $t = (9.8115) \quad (5.3803) \qquad (4.7342)$

  $n = 51 \qquad R^2 = 0.9642 \qquad \bar{R}^2 = 0.9627$

We can test each estimator for its significance or we can also jointly test $\alpha + \beta$ to check returns to scale such as

› H_0: $\alpha + \beta = 1$ or constant returns to scale

› H_a: otherwise.

(4) *Examples*

Polynomial models

There are multiple economic models incorporating polynomial form, such as total cost and marginal cost, an example here is the effect of age to wage in the quadratic form. Given that wage is a function of age as follows (in the stochastic form).

- $w_i = \hat{\beta}_1 + \hat{\beta}_2 age_i + \hat{\beta}_3 age_i^2 + u_i$

where w_i is the value of output of an economy,
age_i is straightforward.
age_i^2 is the squares of age

An example of the estimation is

- $\widehat{w_i} = 3.73 + 0.298 age_i - 0.0061 age_i^2$

(4) *Examples*

Polynomial models

As age_i^2 becomes higher, the square (negative) effect will be dominant.

(1) *The t-test*

With the normality assumption, we find that estimators are normally distributed with an unknown variance. Therefore, to test statistical significance, we need to rely on t statistics as follows.

- $t_{cal}(\beta_i) = \frac{\hat{\beta}_i - \beta_i}{\text{se}_{\hat{\beta}_i}} \sim t_{n-3}$

where $\hat{\beta}_i$ is the value of estimator,
β_i is the value that we would test against and
$\text{se}_{\hat{\beta}_i}$ is the standard error of the estimator.

Now that the d.f. is $n - 3$ if we have two independent variables in our estimation (3 estimators or 3 unknown). Testing procedures are the same.

(2) *Example*

Given the data collected at the beginning of this class, we hypothesize that weight (wei_i) is determined by both height (hei_i) and natural logarithmic household expenditure ($\ln exp_i$) as show here.

- $\widehat{wei}_i = \hat{\beta}_1 + \hat{\beta}_2 hei_i + \hat{\beta}_3 \ln exp_i$

Two results are listed here: the first one we exclude $\ln hhexp_i$ and the second we include it.

(1)	$\widehat{wei}_i =$	-127.3434	$+1.1150 hei_i$	
$n = 52;$ $R^2 = 0.5144; \overline{R}^2 = 0.5047$		(25.2103)	(0.1532)	
(2)	$\widehat{wei}_i =$	-131.5308	$+1.1103 hei_i$	$+0.4429 \ln exp_i$
$n = 52;$ $R^2 = 0.5154; \overline{R}^2 = 0.4956$		(28.7746)	(0.1553)	(1.422)

(2) *Example*

Let's perform the individual tests. We are going to test all the coefficients, one by one.

<u>*Step 1*</u>: State your hypothesis

Stating three hypotheses separately, which are

- For β_1 $H_0: \beta_1 = 0$
  $H_a: \beta_1 \neq 0$
- For β_2 $H_0: \beta_2 = 0$
  $H_a: \beta_2 \neq 0$
- For β_3 $H_0: \beta_3 = 0$
  $H_a: \beta_3 \neq 0$

In this case, we are testing if each coefficient is significantly different from zero or not.

(2) *Example*

<u>*Step 2:*</u> Calculate test statistics

- For β_1 $\quad t_{cal}(\beta_1) = \frac{\widehat{\beta}_1 - \beta_1}{\text{se}_{\widehat{\beta}_1}} = \frac{-131.5308 - 0}{28.7746} =$
- For β_2 $\quad t_{cal}(\beta_2) = \frac{\widehat{\beta}_2 - \beta_2}{\text{se}_{\widehat{\beta}_2}} = \frac{1.1103 - 0}{0.1553} =$
- For β_3 $\quad t_{cal}(\beta_3) = \frac{\widehat{\beta}_3 - \beta_3}{\text{se}_{\widehat{\beta}_3}} = \frac{0.4429 - 0}{1.422} =$

<u>*Step 3:*</u> Pick an α and state decision rules

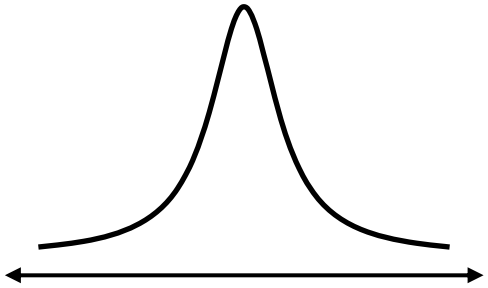
Supposed that we pick $\alpha = 0.05$, now we are testing against zero, we can directly at the t-table for the critical values which are

- $t_{lower} =$
- $t_{upper} =$

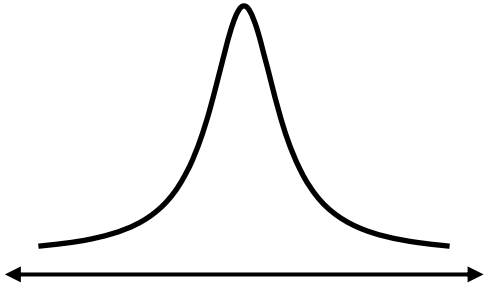
(2) *Example*

<u>*Step 4:*</u> Concluding the test results

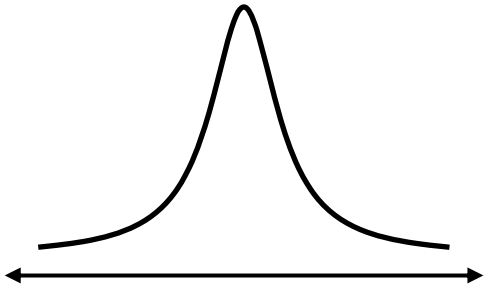
- For β_1



- For β_2



- For β_3



(1) *Introduction*

Now consider if we want to test parameters jointly under the same hypothesis, for instance,

- $H_0: \beta_2 = \beta_3 = 0$
- H_a: Not all the slope coefficients are simultaneously zero (otherwise).

This hypothesis sates that "β_2 and β_3 are jointly or simultaneously equal to zero". We **cannot** utilize ordinary t-test anymore, but rely on specific joint test or the **Analysis of Variance (ANOVA)**

The ANOVA test is a different approach. The focus is on dispersion of each part, namely the ESS and RSS. According to the study of variation from the mean in the coefficient of determination, we have

- $\sum \hat{y}_i^2 = \hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i} + \sum \hat{u}_i^2$

  TSS = ESS + RSS

(1) *Introduction*

From the estimation of weight model, we now look at another information printed from regression table to make use of our joint test.

```
      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(2, 49)        =     26.06
       Model |  3737.27891         2  1868.63945   Prob > F        =    0.0000
    Residual |   3513.8509        49  71.7112429   R-squared       =    0.5154
-------------+----------------------------------   Adj R-squared   =    0.4956
       Total |  7251.12981        51  142.179016   Root MSE        =    8.4682
```

When we perform the ANOVA, we are comparing between a part that the model can explain (ESS) versus a part it cannot (RSS), therefore the comparison is

- $\frac{\text{ESS}/df}{\text{RSS}/df} = \frac{\widehat{\beta}_2 \sum y_i x_{2i} + \widehat{\beta}_3 \sum y_i x_{3i}/(k-1)}{\sum \widehat{u}_i^2/(n-k)} \sim$

(1) *Introduction*

Consider the ratio, from the assumption of $u_i \sim N(0, \sigma^2)$,

- $E\left(\frac{\sum \hat{u}_i^2}{n-k}\right) = \sigma^2$

and if we assume that our stated hypothesis $H_0: \beta_2 = \beta_3 = 0$ is true, then

- $\sum \hat{y}_i^2 = \sum \hat{u}_i^2$ or
- $\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i} = 0$

or the effect of X_{2i} and X_{3i} is **trivial to the variation in Y_i, the only source of variation in Y_i is the error term u_i.**

Therefore, the ratio becomes smaller if $\beta_2 = \beta_3 = 0$. The decision rule for this test is

- $F_{cal} > F_\alpha(k-1, n-k)$ then we can reject H_0

(2) *Overall significance of a model*

Let's perform the individual tests. We are going to test all the coefficients, one by one.

<u>*Step 1*</u>: State your hypothesis

- $H_0: \beta_2 = \beta_3 = 0$
- H_a: otherwise.

(2) Calculate test statistics

```
      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(2, 49)        =     26.06
       Model |  3737.27891         2  1868.63945   Prob > F        =    0.0000
    Residual |   3513.8509        49  71.7112429   R-squared       =    0.5154
-------------+----------------------------------   Adj R-squared   =    0.4956
       Total |  7251.12981        51  142.179016   Root MSE        =    8.4682
```

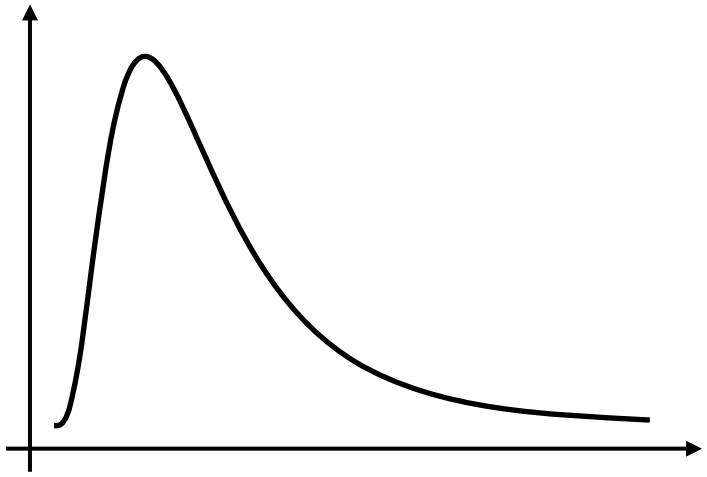
- $F_{cal} = \frac{\text{ESS}/df}{\text{RSS}/df} = \frac{\text{ESS}/k-1}{\text{RSS}/n-k} =$

(2) *Overall significance of a model*

(3) Pick an α and state decision rules

- $\alpha =$
- $F_{upper,\alpha}(2,50) =$

(4) Conclude the test result



(2) *Overall significance of a model*

We then can compute the F statistics directly from R^2.

Given that our model has the $R^2 = 0.5154$,

- $F_{cal} = \frac{\frac{ESS}{TSS}/(k-1)}{\frac{RSS}{TSS}/(n-k)} = \frac{R^2/(k-1)}{1-R^2/(n-k)} =$

which yields the same result compared to computing from the mean squares.

(3) *The marginal contribution*

If we estimate SRF with the same context, but we are not sure if we should add X_{3i} into the model or not, we have two different models.

- Excluding : $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i}$
- Including : $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$

The following test is to measure increment or contribution of added independent variable(s), in this case X_{3i}. Hence, the test highlights a comparison of whole models, instead of just a particular variable's significance.

The answer we are looking for from this test is that "**should the incremental variable(s) be added into the model?**"

(3) *The marginal contribution*

The F-statistics that we use becomes

- $F_{cal} = \frac{ESS_{new} - ESS_{old}/(number\ of\ new\ regressors)}{RSS_{new}/(n-k_{new})}$

or if we measure from the R^2, it would be

- $F_{cal} = \frac{R^2_{new} - R^2_{old}/(number\ of\ new\ regressors)}{1-R^2_{new}/(n-k_{new})}$

Note that the R^2 method is applicable only when both models have the same Y_i, while the mean squares method can apply to other types of test, more on that later.

(3) *The marginal contribution*

Comparing between two models that we estimated, the first one with only height variable.

- $\widehat{wei}_i = \hat{\beta}_1 + \hat{\beta}_2 hei_i$

```
      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(1, 50)        =     52.98
       Model |   3730.3217         1   3730.3217   Prob > F        =    0.0000
    Residual |   3520.8081        50  70.4161621   R-squared       =    0.5144
-------------+----------------------------------   Adj R-squared   =    0.5047
       Total |  7251.12981        51  142.179016   Root MSE        =    8.3914
```

The second model, log of household expenditure is added.

- $\widehat{wei}_i = \hat{\beta}_1 + \hat{\beta}_2 hei_i + \hat{\beta}_3 \ln exp_i$

```
      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(2, 49)        =     26.06
       Model |  3737.27891         2  1868.63945   Prob > F        =    0.0000
    Residual |   3513.8509        49  71.7112429   R-squared       =    0.5154
-------------+----------------------------------   Adj R-squared   =    0.4956
       Total |  7251.12981        51  142.179016   Root MSE        =    8.4682
```

(3) *The marginal contribution*

Step 1: State your hypothesis

- H_0: log of HH expenditure has no marginal contribution to the model.
- H_a: otherwise.

Step 2: Calculate the test statistics

- $F_{cal} = \frac{ESS_{new} - ESS_{old}/(number\ of\ new\ regressors)}{RSS_{new}/(n-k_{new})} =$
- $F_{cal} = \frac{3{,}737.2789 - 3{,}730.3217/(1)}{3{,}513.8509/(52-1)} =$

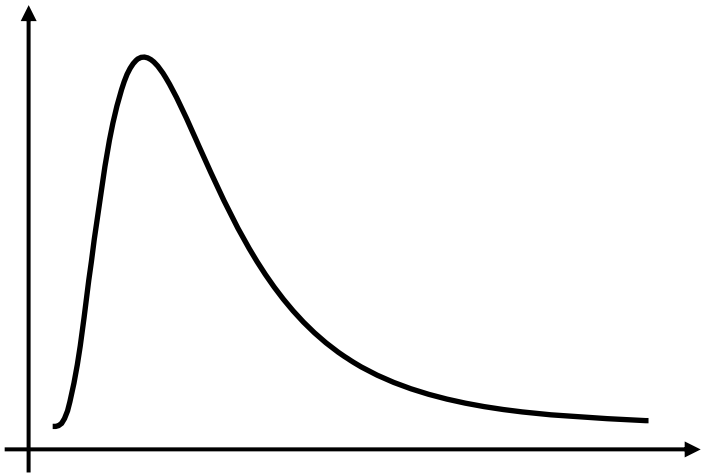
Step 3: Pick an α and state decision rules

- $\alpha =$
- $F_{upper,\alpha}(1,51) =$

(3) *The marginal contribution*

<u>*Step 4*</u>: Conclude the test result

The addition of variable X_{3i} or *log of household expenditure*



Note that the F-test for the contribution is most of the time in line with the individual t-test of those added variable(s) since $t^2 \sim F(1, n)$.

Selecting the most fitted model

- Choose a model with the highest F value which implies highest $\overline{R}^2$.
- Absolute term of t-value of added variable(s) is more than 1, which will also imply higher $\overline{R}^2$.

(4) *Equality of two regression coefficients*

Since the collected data provide interesting insight to this topic, let's shift to another data provided in the book. We have an ordinary demand model represented here.

- $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$

where Y_i is quantity demanded for a commodity
X_{2i} is its price
X_{3i} is consumer income
X_{4i} is consumer wealth

A question arises **if X_{3i} and X_{4i}**, income and wealth, **both are representing affordability in the same way or not**. We can answer this question with two methods

- Using t-test
- Using F-test with Restricted and Unrestricted models

(4) *Equality of two regression coefficients*

(1) Using t-test

- $t_{cal} = \frac{(\hat{\beta}_3 - \hat{\beta}_4) - (\beta_3 - \beta_4)}{\text{se}_{(\hat{\beta}_3 - \hat{\beta}_4)}} \sim t_{n-k}$ where

- $\text{se}_{(\hat{\beta}_3 - \hat{\beta}_4)} = \sqrt{\text{var}(\hat{\beta}_3) + \text{var}(\hat{\beta}_4) - 2\text{cov}(\hat{\beta}_3, \hat{\beta}_4)}$

Example: Given an estimated of cubic cost function below,

- $\hat{Y}_i = 141.7667 + 63.4777X_i - 12.9615X_i^2 + 0.9396X_i^3$

  (6.3753) (4.7789) (0.9857) (0.0591)

  $cov(\hat{\beta}_3, \hat{\beta}_4) = -0.0576 \quad R^2 = 0.9983 \;\; n = 10$

where Y_i is total cost and X_i is output.

(4) *Equality of two regression coefficients*

<u>*Step 1*</u>: State your hypothesis

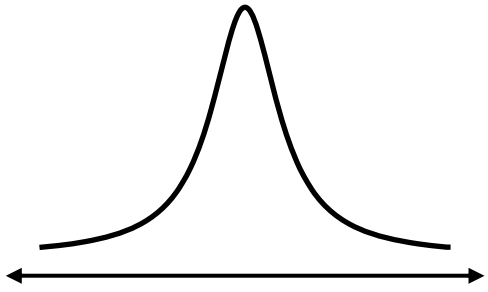
- $H_0: \beta_3 = \beta_4$ or $\beta_3 - \beta_4 = 0$
- $H_a: \beta_3 \neq \beta_4$ or $\beta_3 - \beta_4 \neq 0$

<u>*Step 2*</u>: Calculate test statistics

- $t_{cal} = \frac{(\hat{\beta}_3 - \hat{\beta}_4) - (\beta_3 - \beta_4)}{\text{se}_{(\hat{\beta}_3 - \hat{\beta}_4)}} =$

<u>*Step 3*</u>: Pick an α and state decision rules

- $\alpha =$
- $t_{cri} =$



<u>*Step 4*</u>: Conclude the test result

(4) *Equality of two regression coefficients*

(2) Using F-test with Restricted and Unrestricted models

There are some certain models that economic theory suggest that the coefficients in a regression satisfy some linear equality restrictions. Consider a Cobb-Douglas here, where some notations are altered, which is already linearized

- $\ln Y_i = \ln \beta_1 + \beta_K \ln K_i + \beta_L \ln L_i + u_i$

β_K and β_L are the elasticity of capital and labor input respectively. We also know that addition of these two parameters reveals returns to scale.

Supposed that we want to test if there are constants returns to scale or not, we can hypothesize

- $H_0: \beta_K + \beta_L = 1$
- $H_\mathrm{a}: \beta_K + \beta_L \neq 1$

This is a test to see **if we can make sure that the result of addition of two parameters is 1 or not**.

(4) *Equality of two regression coefficients*

(2) Using F-test with Restricted and Unrestricted models

However, we can "**restrict**" the regression or **force** that returns to scale into the regression directly that either

- $\beta_K + \beta_L = 1 \;/\; \beta_K = 1 - \beta_L \;/\; \beta_L = 1 - \beta_K$

If we rewrite the Cobb-Douglas function with the restriction here, it becomes

- $\ln Y_i = \ln \beta_1 + \beta_K \ln K_i + (1 - \beta_K) \ln L_i + u_i$

where $\frac{Y_i}{L_i}$ is output per labor and $\frac{K_i}{L_i}$ is capital per labor.

You may notice that there is one less coefficient to estimate. This is known as "**Restricted Least Square**" (RLS).

(4) *Equality of two regression coefficients*

(2) Using F-test with Restricted and Unrestricted models

Therefore, the test we are about to perform is a comparison between

- Unrestricted model: $\ln Y_i = \ln \beta_1 + \beta_K \ln K_i + \beta_L \ln L_i + u_i$
- Restricted model: $\ln (\frac{Y_i}{L_i}) = \ln \beta_1 + \beta_K \ln(\frac{K_i}{L_i}) + u_i$

which is a test to see if **the restriction imposed is valid or not**. We can set up the same hypothesis.

- $H_0: \beta_K + \beta_L = 1$ or the restriction is valid
- $H_\mathrm{a}: \beta_K + \beta_L \neq 1$ or the restriction is not valid

(4) *Equality of two regression coefficients*

(2) Using F-test with Restricted and Unrestricted models

After that we can compute the F statistics, which is

- $F_{cal} = \frac{RSS_R - RSS_{UR}/m}{RSS_{UR}/(n-k_{UR})} \sim F_{(m,\, n-k_{UR})}$

where R and UR indicates the value from restricted and unrestricted model respectively, m is the number of linear restriction (1 for this case). We can also derive F_{cal} from R^2 as well.

- $F_{cal} = \frac{(R^2_{UR} - R^2_R)/m}{(1-R^2_{UR})/(n-k_{UR})} \sim F_{(m,\, n-k_{UR})}$

Note that $R^2_{UR} \geq R^2_R$ and the variable on the left-hand side in both models is not the same so **they are not comparable**. In this case, need to rely on calculating F statistics from RSS instead.

(4) *Equality of two regression coefficients*

Example: Supposed we have a result of

- Unrestricted model:

$$\ln \hat{Y}_i = -1.6524 + 0.8460 \ln K_i + 0.3397 \ln L_i$$

$$R^2_{UR} = 0.9951 \qquad RSS_{UR} = 0.0136$$

- Restricted model:

$$\ln \left(\frac{\widehat{Y_i}}{L_i}\right) = -0.4947 + 1.0153 \ln\left(\frac{K_i}{L_i}\right)$$

$$R^2_R = 0.9777 \qquad RSS_R = 0.0166$$

Both models has 20 observations. Perform the test from both RSS and R^2 with the hypothesis here.

(4) *Equality of two regression coefficients*

<u>*Step 1*</u>: State your hypothesis

- $H_0: \beta_K + \beta_L = 1$ or the restriction is valid
- $H_a: \beta_K + \beta_L \neq 1$ or the restriction is not valid

<u>*Step 2*</u>: Calculate test statistics

- $F_{cal} = \frac{RSS_R - RSS_{UR}/m}{RSS_{UR}/(n-k_{UR})} \sim F_{(m,\, n-k_{UR})} =$

Just to prove that R^2 is not comparable, we can also try to calculate the test statistics anyway.

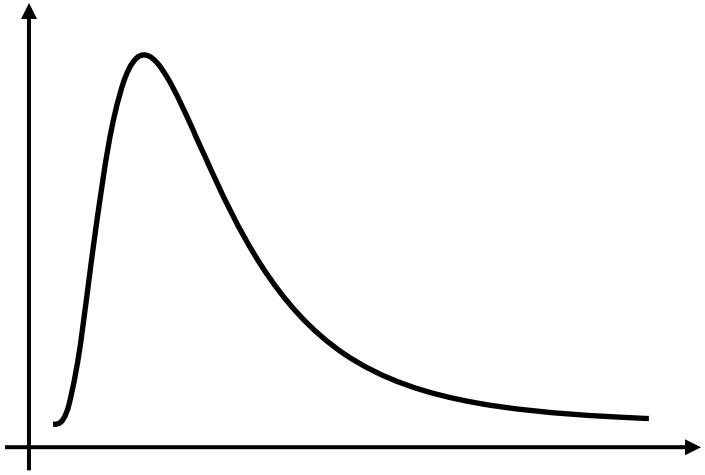
- $F_{cal} = \frac{(R^2_{UR} - R^2_R)/m}{(1-R^2_{UR})/(n-k_{UR})} =$

(4) *Equality of two regression coefficients*

<u>Step 3</u>: Pick an α and state decision rules

- $\alpha =$
- $F_{upper,\alpha}(1,17) =$

<u>Step 4</u>: Conclude the test result



(5) *General F-test*

As we can see that the F test can be utilize in multiple ways of comparing two competing models, we can also set up the test for "**constrained**" and "**unconstrained**" models as seen in the example below. First, we start from the "unconstrained" model

- $\ln Y_i = \ln \beta_1 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + \beta_4 \ln X_{4i} + \beta_5 \ln X_{5i} + u_i$

where Y_i is per capita consumption of chicken

X_{2i} is real disposable per capita income

X_{3i} is real retail price of chicken

X_{4i} is real retail price of pork

X_{5i} is real retail price of beef

β_4 and β_5 refer to cross-price elasticity of chicken and pork and chicken and beef.

(5) *General F-test*

Thus,

- If $\beta_4 \,/\, \beta_5 > 0$, chicken and pork / beef are substitutable products.
- If $\beta_4 \,/\, \beta_5 < 0$, chicken and pork / beef are complementary products.
- If $\beta_4 \,/\, \beta_5 = 0$, chicken and pork / beef are unrelated products.

If we want to test that how chicken and pork / beef are related, we can set up a hypothesis as such.

- $H_0: \beta_4 = \beta_5 = 0$
- H_a: otherwise

If we cannot reject the null hypothesis, the model become "**constrained**" as

- $\ln Y_i = \ln \beta_1 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + u_i$

(5) *General F-test*

The test is similar to the restricted and unrestricted model. Moreover, we can use the R^2 approach here since the Y_i is the same for both the constrained and unconstrained model. Test statistics can be calculated as follows.

- $F_{cal} = \frac{RSS_R - RSS_{UR}/m}{RSS_{UR}/(n-k_{UR})} \sim F_{(m,\, n-k_{UR})}$
- $F_{cal} = \frac{(R^2_{UR} - R^2_R)/m}{(1-R^2_{UR})/(n-k_{UR})} \sim F_{(m,\, n-k_{UR})}$

Note that we use the same notations compared to the restricted model testing so that we don't have to create another complication.

Hence, m is number of coefficient constrained. R and UR denote values from constrained and unconstrained model respectively.

(5) *General F-test*

Example: Given the regression results are as follows (n=23),

- Unconstrained model:

$$\ln \hat{Y}_i = 2.19 + 0.34 \ln X_{2i} - 0.50 \ln X_{3i} + 0.15 \ln X_{4i} + 0.09 \ln X_{5i}$$

$$R^2_{UR} = 0.9823$$

- Constrained model:

$$\ln \hat{Y}_i = 2.03 + 0.45 \ln X_{2i} - 0.38 \ln X_{3i}$$

$$R^2_R = 0.9801$$

Step 1: State your hypothesis

- $H_0: \beta_4 = \beta_5 = 0$
- H_a: otherwise

(5) *General F-test*

<u>*Step 2*</u>: Calculate test statistics

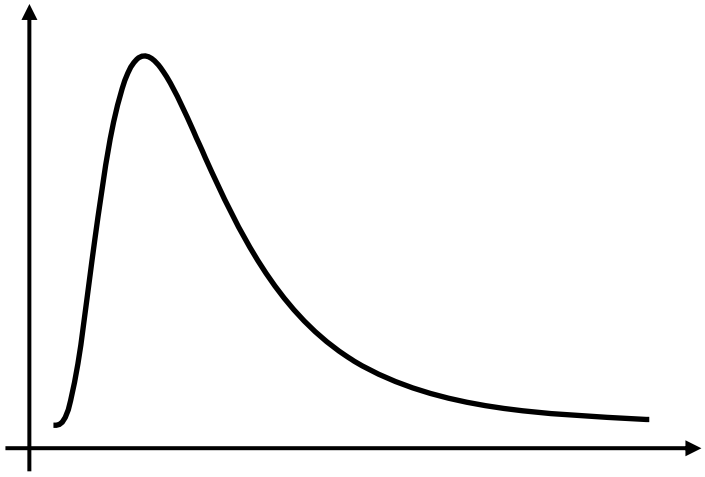
› $F_{cal} = \frac{(R_{UR}^2 - R_R^2)/m}{(1-R_{UR}^2)/(n-k_{UR})} =$

<u>*Step 3*</u>: Pick an α and state decision rules

› $\alpha =$

› $F_{upper,\alpha}(2,18) =$

<u>*Step 4*</u>: Conclude the test result



(6) *Structural change: Chow Test*

Many event in our history, we had sometimes faced with "**structural change**" , a major shift in the basic how market works and how agents interact. The shift can be both internal or external.

In Thailand, we may consider opposing party getting elected and alter policy set. Or major flood in 2011 may alter how business decision of location established.

In econometrics term, it refers to a change in parameter, either the intercept or slopes, due to socio-political shift.

Consider when there is a pool of data, ranging a long periods of time, estimation the whole data can be misleading as it reveals the average values of estimator within that range.

(6) *Structural change: Chow Test*

The Chow Test attempts tackling this problem, to test that **if there is any structural shift sometime in our data or not**.

For this example, we consider saving-income relationship between the second oil shock in the 1980s. Data cover 1970-1995. It is assumed that we can separate between pre and post 1982, when the unemployment rate is at the highest. The setup is as follows.

SRF from 1970-1981 (Eq. 1)

- $Y_t = \lambda_1 + \lambda_2 X_t + u_{1t} \qquad n_1 = 12$

SRF from 1982-1995 (Eq. 2)

- $Y_t = \gamma_1 + \gamma_2 X_t + u_{2t} \qquad n_2 = 14$

SRF from pooled data 1970-1995 (Eq. 3)

- $Y_t = \beta_1 + \beta_2 X_t + u_t \qquad n = (n_1 + n_2) = 26$

where Y is savings and X is income. The slope parameter represents

(6) *Structural change: Chow Test*

Assumptions

(1) $u_{1t} \sim N(0, \sigma^2)$ and $u_{2t} \sim N(0, \sigma^2)$

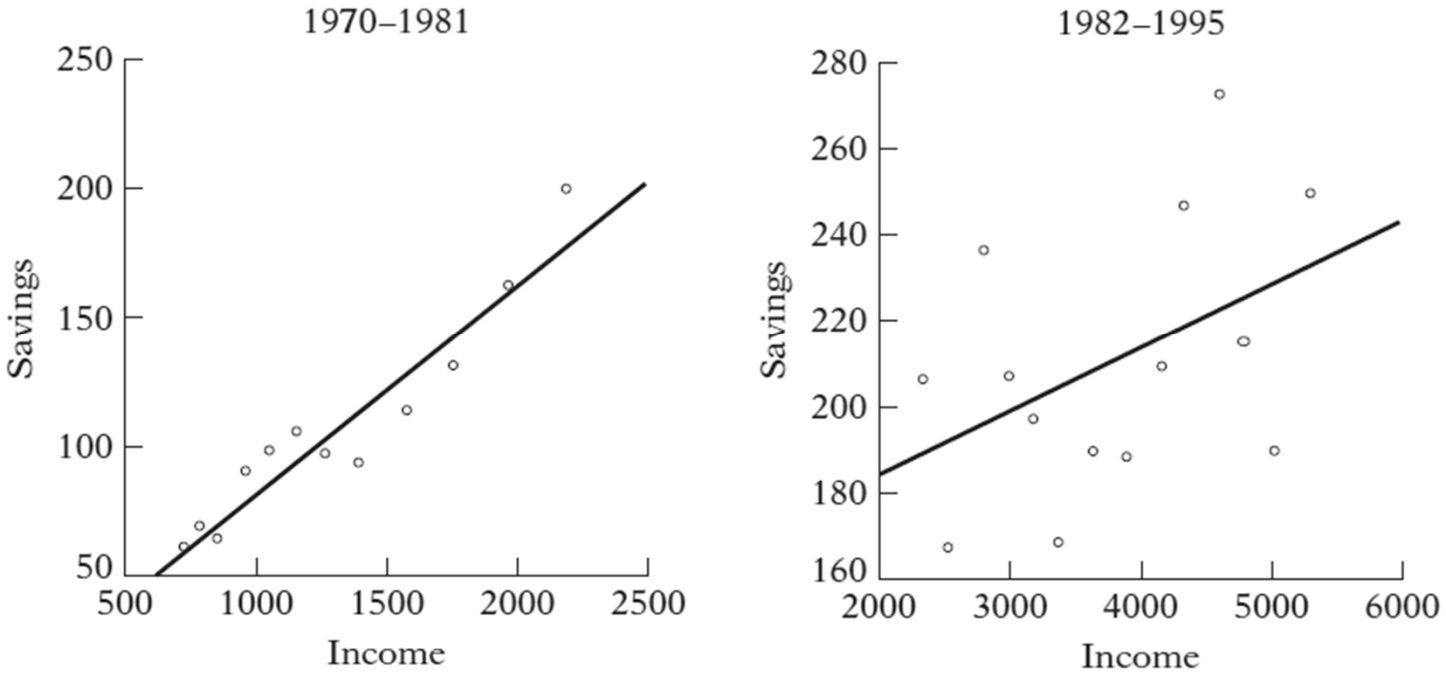
(2) u_{1t} and u_{2t} are independently distributed.

The Chow Test still relies on ANOVA, comparing a restricted model with an unrestricted model.

In this case, the **restricted** one assumes $\lambda_1 = \gamma_1$ and $\lambda_2 = \gamma_2$, or the Eq.3, while the **unrestricted** one allows different values of λ and γ.

(6) *Structural change: Chow Test*

Plotting the graphs above seems to suggest that there is a structural change in people saving behavior, due to the slope change.



(6) *Structural change: Chow Test*

<u>*Step 1*</u>: State a hypothesis

- H_0: $\lambda_1 = \gamma_1$ and $\lambda_2 = \gamma_2$
- H_a: otherwise

<u>*Step 2*</u>: Estimate Eq. 3 or the restricted model. Retrieve

- $RSS_R = RSS_3$ or **restricted sum of squares.**
- d.f. is straightforwardly $n_1 + n_2 - k$

<u>*Step 3*</u>: Estimate Eq. 1 and Eq.2. Retrieve

- $RSS_{UR} = RSS_1 + RSS_2$ or **unrestricted sum of squares.**
- d.f. for this model is $n_1 + n_2 - 2k$

Note that k is the number of parameter estimated.

(6) *Structural change: Chow Test*

<u>*Step 4*</u>: Calculate F statistics by

- $F_{cal} = \frac{RSS_R - RSS_{UR}/k}{RSS_{UR}/(n_1+n_2-2k)} \sim F_{(k,\, n_1+n_2-2k)}$

<u>*Step 5*</u>: Concluding the results,

- We **can** reject the null hypothesis if $F_{cal} > F_{upper}$, meaning that it is either $\lambda_1 \neq \gamma_1$ or $\lambda_2 \neq \gamma_2$ or both. Therefore, it implies a structural change within these two periods.

- We **cannot** reject the null hypothesis if $F_{cal} < F_{upper}$, meaning that $\lambda_1 = \gamma_1$ and $\lambda_2 = \gamma_2$. So, we cannot make sure that there is a structural change within this period.

(6) *Structural change: Chow Test*

Example Given the regression results are,

- **Eq. 1**: $\hat{Y}_t = 1.0161 + 0.0803X_t$

$R_1^2 = 0.9021$ $RSS_1 = 1{,}785.032$ $n_1 = 12$

- **Eq. 2**: $\hat{Y}_t = 153.4947 + 0.0148X_t$

$R_2^2 = 0.2971$ $RSS_2 = 10{,}005.22$ $n_2 = 14$

- **Eq. 3**: $\hat{Y}_t = 62.4226 + 0.0376X_t$

$R_3^2 = 0.7672$ $RSS_3 = 23{,}248.30$ $n_3 = 26$

Step 1: State a hypothesis

- H_0: $\lambda_1 = \gamma_1$ and $\lambda_2 = \gamma_2$
- H_a: otherwise

(6) *Structural change: Chow Test*

<u>*Step 2*</u>: Calculate the test statistics

- $RSS_R =$
- $RSS_{UR} =$

Then, the F statistics is

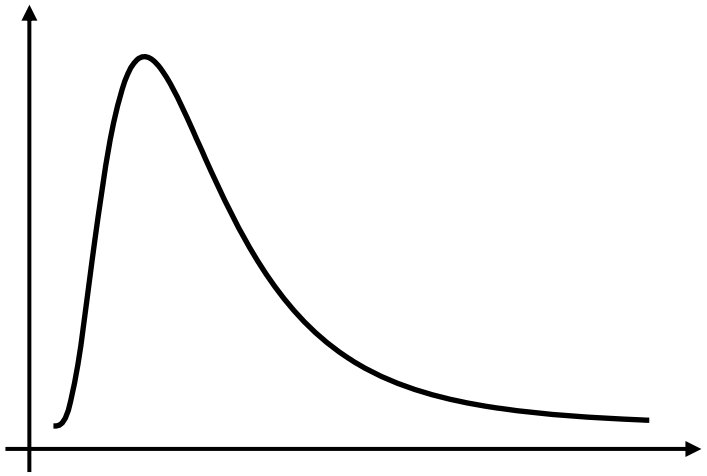
- $F_{cal} = \frac{RSS_R - RSS_{UR}/k}{RSS_{UR}/(n_1+n_2-2k)} =$

<u>*Step 3*</u>: Pick an α and state decision rules

- $\alpha =$
- $F_{upper,\alpha}(2,22) =$

(6) *Structural change: Chow Test*

<u>*Step 4*</u>: Conclude the test result



(6) *Structural change: Chow Test*

Limitations of The Chow Test

- The assumption of u_{1t} and u_{2t} being independently distributed should (or must) be tested, see the further explanation on the test on page 258.
- We need to assume that we know exactly where or when is the structural break point. If the speculation is not completely correct, the result of the test maybe controversial.
- The Chow Test only reveals difference between two models, which means that there might be a difference either in the intercept, slope or both. (The next topic of dummy variables will address this issue)