

## Эконометрика, 2017-2018, 2 модуль

### Семинар 4

20.11.17 для

Группы Э\_Б2015\_Э\_3

Семинарист О.А.Демидова

## Исследование структурной устойчивости коэффициентов регрессии с помощью теста Чоу (Chow)

Предположим, у нас есть две выборки, объема  $n_1$  и  $n_2$  соответственно для одного и того же набора переменных. По каждой выборке мы оцениваем коэффициенты уравнения регрессии:

$$Y_i = \beta'_1 + \beta'_2 X_{2i} + \dots + \beta'_k X_{ki} + \varepsilon'_i, i = 1, \dots, n_1,$$

$$Y_i = \beta''_1 + \beta''_2 X_{2i} + \dots + \beta''_k X_{ki} + \varepsilon''_i, i = n_1 + 1, \dots, n_1 + n_2.$$

Нас интересует ответ на вопрос: что лучше, оценивать модель по объединенной выборке или по каждой выборке в отдельности?

Для проверки гипотезы

$H_0 : \beta'_i = \beta''_i, i = 1, \dots, k, \sigma_{\varepsilon'}^2 = \sigma_{\varepsilon''}^2$ , (т.е. лучше оценивать модель по объединенной выборке)

при альтернативной гипотезе

$H_1 : \exists i, j : \beta'_i \neq \beta''_i$  или  $\sigma_{\varepsilon'}^2 \neq \sigma_{\varepsilon''}^2$  (т.е. лучше оценивать модели по каждой выборке в отдельности)

Используется тест Чоу.

Тестовая статистика имеет вид:

$$F = \frac{(RSS_p - RSS_1 - RSS_2) / k}{(RSS_1 + RSS_2) / (n_1 + n_2 - 2k)},$$

где  $RSS_1$  – сумма квадратов остатков регрессии, оцененной по  $n_1$  наблюдениям,

$RSS_2$  – сумма квадратов остатков регрессии, оцененной по  $n_2$  наблюдениям,

$RSS_P$  – сумма квадратов остатков регрессии, оцененной по всем наблюдениям.

При выполнении нулевой гипотезы тестовая статистика имеет  $F$  – распределение со степенями свободы  $(k, n_1 + n_2 - 2k)$ .

Если рассчитанное значение  $F$  – статистики не превышает критическое  $F_\alpha(k, n_1 + n_2 - 2k)$ , то основная гипотеза не отвергается, зависимость можно считать единой для двух наборов данных.

### **Различия в потреблении россиянами основных типов продуктов**

Используйте данные файла Chow.dta.

- 1) Оцените зависимость потребления одного из видов товаров  $Y$  от его цены  $P$  и дохода домохозяйства  $I$  с помощью модели:

$$Y = \beta_1 + \beta_2 P + \beta_3 I + \varepsilon$$

$$\text{или } \ln Y = \beta_1 + \beta_2 \ln P + \beta_3 \ln I + \varepsilon.$$

- 2) Оцените коэффициенты регрессии по двум выборкам, например,

А) для жителей двух различных типов населенных пунктов (2 – город, 3 – поселок городского типа),

Б) для жителей двух различных административных образований (например, Москвы и Московской области)

В) Центрального и Северо-Западного округов

и общую.

- 3) С помощью теста Чоу ответьте на вопрос, можно ли считать, что для двух выделенных видов домохозяйств имеет место единая зависимость?

#### **Методические рекомендации по выполнению упражнения**

Предположим, Вы хотите оценить функцию спроса на картофель в виде линейной в логарифмах модели.

- 1) Создайте новую переменную, набрав в командном окне

**`gen lnpotat_c = log(buypotat_c)`**

- 2) Аналогично создайте логарифмы остальных необходимых переменных

- 3) Если Вы хотите оценить регрессию только для домохозяйств Центрального округа, то следует набрать команду:

**`reg lnpotat_c lnpr_potat lninc if fed_okr==1`**

(сохраните  $RSS$  с помощью команды `scalar rss1=e(rss)`),

если для домохозяйств Северо-Западного округа, то:

**reg lnpotat\_c lnpr\_potat lninc if fed\_okr ==2**

(сохраните RSS с помощью команды `scalar rss2=e(rss)`),

для оценки объединенной по двум округам регрессии

**reg lnpotat\_c lnpr\_potat lninc if fed\_okr ==1|fed\_okr ==2**

(сохраните RSS с помощью команды `scalar rssp=e(rss)`),

4) Используя RSS из оцененных регрессий, следует рассчитать тестовую  $F$  – статистику

`scalar F=((rssp-rss1-rss2)/3)/((rss1+rss2)/(306-2*3))`

`display F`

и  $p$ -value для этой статистики:

`display Ftail(3, 300, F)`

Если рассчитанное  $p$ -value не превышает выбранного уровня значимости, то основная гипотеза отвергается, зависимость нельзя считать единой для двух наборов данных.

## Эконометрика, 2017-2018, 2 модуль

### Семинар 4

20.11.17 для

Группы Э\_Б2015\_Э\_3

Семинарист О.А.Демидова

## Исследование структурной устойчивости коэффициентов регрессии с помощью теста Чоу (Chow)

Предположим, у нас есть две выборки, объема  $n_1$  и  $n_2$  соответственно для одного и того же набора переменных. По каждой выборке мы оцениваем коэффициенты уравнения регрессии:

$$Y_i = \beta'_1 + \beta'_2 X_{2i} + \dots + \beta'_k X_{ki} + \varepsilon'_i, i = 1, \dots, n_1,$$

$$Y_i = \beta''_1 + \beta''_2 X_{2i} + \dots + \beta''_k X_{ki} + \varepsilon''_i, i = n_1 + 1, \dots, n_1 + n_2.$$

Нас интересует ответ на вопрос: что лучше, оценивать модель по объединенной выборке или по каждой выборке в отдельности?

Для проверки гипотезы

$H_0 : \beta'_i = \beta''_i, i = 1, \dots, k, \sigma_{\varepsilon'}^2 = \sigma_{\varepsilon''}^2$ , (т.е. лучше оценивать модель по объединенной выборке)

при альтернативной гипотезе

$H_1 : \exists i, j : \beta'_i \neq \beta''_i$  или  $\sigma_{\varepsilon'}^2 \neq \sigma_{\varepsilon''}^2$  (т.е. лучше оценивать модели по каждой выборке в отдельности)

Используется тест Чоу.

Тестовая статистика имеет вид:

$$F = \frac{(RSS_P - RSS_1 - RSS_2) / k}{(RSS_1 + RSS_2) / (n_1 + n_2 - 2k)},$$

где  $RSS_1$  – сумма квадратов остатков регрессии, оцененной по  $n_1$  наблюдениям,

$RSS_2$  – сумма квадратов остатков регрессии, оцененной по  $n_2$  наблюдениям,

$RSS_P$  – сумма квадратов остатков регрессии, оцененной по всем наблюдениям.

При выполнении нулевой гипотезы тестовая статистика имеет  $F$  – распределение со степенями свободы  $(k, n_1 + n_2 - 2k)$ .

Если рассчитанное значение  $F$  – статистики не превышает критическое  $F_\alpha(k, n_1 + n_2 - 2k)$ , то основная гипотеза не отвергается, зависимость можно считать единой для двух наборов данных.

## 1) Различия в потреблении россиянами основных типов продуктов

Используйте данные файла Chow.dta.

- 1) Оцените зависимость потребления одного из видов товаров  $Y$  от его цены  $P$  и дохода домохозяйства  $I$  с помощью модели:

$$Y = \beta_1 + \beta_2 P + \beta_3 I + \varepsilon$$

$$\text{или } \ln Y = \beta_1 + \beta_2 \ln P + \beta_3 \ln I + \varepsilon.$$

- 2) Оцените коэффициенты регрессии по двум выборкам, например,

А) для жителей двух различных типов населенных пунктов (2 – город, 3 – поселок городского типа),

Б) для жителей двух различных административных образований (например, Москвы и Московской области)

В) Центрального и Северо-Западного округов

и общую.

- 3) С помощью теста Чоу ответьте на вопрос, можно ли считать, что для двух выделенных видов домохозяйств имеет место единая зависимость?

### **Методические рекомендации по выполнению упражнения 7.5**

Предположим, Вы хотите оценить функцию спроса на картофель в виде линейной в логарифмах модели.

Воспользуйтесь соответствующими указаниями в разделе «Оценка регрессий в пакете STATA».

- 1) Создайте новую переменную, набрав в командном окне

**`gen lnpotat_c = log(buypotat_c)`**

- 2) Аналогично создайте логарифмы остальных необходимых переменных

- 3) Если Вы хотите оценить регрессию только для домохозяйств Центрального округа, то следует набрать команду:

**`reg lnpotat_c lnpr_potat lninc if fed_okr == 1`**

(сохраните RSS с помощью команды `scalar rss1=e(rss)`),

если для домохозяйств Северо-Западного округа, то:

**`reg lnpotat_c lnpr_potat lninc if fed_okr ==2`**

(сохраните RSS с помощью команды `scalar rss2=e(rss)`),

для оценки объединенной по двум округам регрессии

**`reg lnpotat_c lnpr_potat lninc if fed_okr ==1| fed_okr ==2`**

(сохраните RSS с помощью команды `scalar rssp=e(rss)`),

4) Используя RSS из оцененных регрессий, следует рассчитать тестовую  $F$  – статистику

`scalar F=((rssp-rss1-rss2)/3)/((rss1+rss2)/(306-2*3))`

`display F`

и  $p$ -value для этой статистики:

`display Ftail(3, 300, F)`

Если рассчитанное  $p$ -value не превышает выбранного уровня значимости, то основная гипотеза отвергается, зависимость нельзя считать единой для двух наборов данных.

## 2) Задача

1. По данным для 27 фирм, упорядоченных по выпуску ( $Y_1 < \dots < Y_n$ ) была оценена зависимость выпуска  $Y$  от труда  $L$  и капитала  $K$  с помощью моделей

$$\ln Y_i = b_1 + b_2 \ln L_i + b_3 \ln K_i + \varepsilon \quad (1)$$

$$\ln Y_i = b_1 + b_2 \ln (L_i \cdot K_i) + \varepsilon \quad (2)$$

Результаты оценок приведены в таблицах 1, 2

Табл. 1

Variable	Coefficient	Std. Error	t - statistic	Prob.
C	1.1706	0.326	3.582	0.0015
ln L	0.6029	0.125	4.787	0.0001
ln K	0.375	0.085	4.402	0.0002

R-squared            0.943                      F – statistic            200.24

Sum squared resid   0.851                      Prob (F-statistic)      0.0000

Табл. 2

Variable	Coefficient	Std. Error	t - statistic	Prob.
C	1.2833	0.3117	4.1164	0.0004
ln (L·K)	0.4663	0.0234	19.895	0.0000

R-squared            0.94                      F – statistic            395.81

Sum squared resid   0.894                      Prob (F-statistic)      0.0000

1) Проверить для модели 1) гипотезы: а)  $H_0 : b_2 = 0$  б)  $H_0 : b_3 = 0$  в)  $H_0 : b_2 = b_3 = 0$

- 2) Объяснить, почему вторая модель является ограниченной версией первой и проверить выполнение соответствующих ограничений на коэффициенты регрессии.
- 3) Разделив фирмы на маленькие  $i = 1, \dots, 14$  и большие  $i = 15, \dots, 27$ , для них оценили отдельные регрессии. Результаты приведены в таблицах 3 и 4. Можно ли считать, что производственные функции для больших и маленьких фирм не различаются?

Табл. 3 Included observation: 14

Variable	Coefficient	Std. Error	t - statistic	Prob.
C	0.6998	0.649	1.078	0.3040
ln L	0.9000	0.133	6.764	0.0000
ln K	0.2100	0.056	3.718	0.0034

R-squared            0.896                      F – statistic            47.84  
Sum squared resid   0.119                      Prob (F-statistic)      0.0000

Табл. 4 Included observation: 13

Variable	Coefficient	Std. Error	t - statistic	Prob.
C	1.4082	0.678	2.075	0.0647
ln L	0.0081	0.226	0.036	0.9720
ln K	0.805	0.179	4.492	0.0012

R-squared            0.908                      F – statistic            49.81  
Sum squared resid   0.362                      Prob (F-statistic)      0.0000

### Три формы уравнений МНК для нахождения оценок коэффициентов регрессии

1. Для нахождения оценок МНК коэффициентов регрессии

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon$$

требуется решить систему нормальных уравнений

$$X'X\hat{\beta} = X'Y$$

2. Для регрессии в отклонениях

$$y = \alpha_1 x_1 + \dots + \alpha_k x_k + \varepsilon,$$

где  $y = Y - \bar{Y}I$ ,  $x_i = X_i - \bar{X}_i I$ ,  $i = 1, \dots, k$ ,  $I$  – единичный вектор,

система нормальных уравнений может быть переписана в виде

$$\text{var}[X]\hat{\alpha} = \text{cov}[X, Y]$$

3. Для регрессии в центрированных и нормированных переменных

$$\tilde{y} = \tilde{\beta}_1 \tilde{x}_1 + \dots + \tilde{\beta}_k \tilde{x}_k + \varepsilon,$$

$$\text{где } \tilde{y} = \frac{y}{\hat{\sigma}_y}, \quad \tilde{x}_i = \frac{x_i}{\hat{\sigma}_{x_i}}, \quad i = 1, \dots, k$$

система нормальных уравнений может быть переписана в виде

$$\text{cov}[X]\tilde{\beta} = \text{cov}[X, Y]$$

1. Как связаны коэффициенты первой и второй формы уравнений МНК?
2. Как связаны коэффициенты первой и третьей формы уравнений МНК?
3. Для регрессий  $Y = a_0 + a_1 X_1 + \varepsilon$ ,

$$Y = b_0 + b_1 X_2 + \varepsilon,$$

$$X_2 = c_0 + c_1 X_1 + \varepsilon,$$

$$X_1 = d_0 + d_1 X_2 + \varepsilon$$

известны следующие МНК – оценки коэффициентов:  $\hat{a}_1 = 3$ ,  $\hat{b}_1 = -3$ ,  $\hat{c}_1 = -2$ ,  $\hat{d}_1 = -0.8$ .

Найти оценки МНК коэффициентов наклона в регрессии  $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \varepsilon$ .

2. Для регрессии в отклонениях  $y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ , оцениваемой по 100 наблюдениям, известны следующие суммы:

$$\sum y^2 = \frac{493}{3}, \sum x_1^2 = 30, \sum x_2^2 = 3, \sum x_1 y = 30, \sum x_2 y = 20, \sum x_1 x_2 = 0.$$

4. Найти оценки МНК коэффициентов  $\beta_1, \beta_2$  и коэффициент множественной детерминации  $R^2$ .