Эконометрика 1
осень 2016

Лекция 4

05.10.2016

Гомоскедастичность ошибки u_i

Случайная ошибка u_i называется **гомоскедастичной**, если условная дисперсия u_i относительно X_i постоянна для $i = 1, 2, \dots n$ (т.е. $\mathrm{var}(u_i | X_i = x) = const,\ i = 1, \dots, n$). В частности, условная дисперсия u_i относительно X_i не зависит от X_i.

В противном случае ошибка называется **гетероскедастичной**.

Теорема Гаусса-Маркова

Если для всех $i, j = 1, 2, \ldots, n$ выполняются условия Гаусса-Маркова (1)-(3)

$$(1)\ E(u_i | X_1, \ldots, X_n) = 0;$$
$$(2)\ \mathrm{var}(u_i | X_1, \ldots, X_n) = \sigma_u^2, 0 < \sigma_u^2 < \infty;$$
$$(3)\ E\left(u_i u_j \middle| X_1, \ldots, X_n\right) = 0, i \neq j,$$

то МНК-оценка $\hat{\beta}_1$ является наилучшей (эффективной) линейной условно не смещенной оценкой (BLUE)

Предположения МНК

Предположение №1: условное распределение u_i относительно X_i имеет нулевое среднее: $E(u_i|X_i) = 0$

Предположение №2: (X_i, Y_i), $i = 1, \dots, n$, независимы и одинаково распределены $(\mathrm{i.i.d.})$

Предположение №3: большие выбросы маловероятны: X_i и Y_i имеют ненулевые конечные четвертые моменты

Связь условий Гаусса-Маркова и предположений МНК

УГ-М (1) следует из предположений 1 и 2

УГ-М (2) следует из предположения 2 и предположения о гомоскедастичности ошибок

УГ-М (3) следует из предположения 2

Теорема Гаусса-Маркова

- Линейность:

$$\sum_{i=1}^{n}(X_i - \bar{X}) = 0 \Rightarrow \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n}(X_i - \bar{X})Y_i - \bar{Y}\sum_{i=1}^{n}(X_i - \bar{X}) = \sum_{i=1}^{n}(X_i - \bar{X})Y_i \Rightarrow$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \sum_{i=1}^{n}\hat{a}_i Y_i,$$

где $\hat{a}_i = \frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \Rightarrow$ линейность.

Условная несмещенность: см. предыдущую лекцию.

Теорема Гаусса-Маркова

Эффективность:

Пусть $\tilde{\beta}_1$ - любая линейная условно не смещенная оценка β_1, т.е. $\tilde{\beta}_1 = \sum_{i=1}^{n} a_i Y_i$. Тогда (покажите) $\sum_{i=1}^{n} a_i = 0, \sum_{i=1}^{n} a_i X_i = 1$.

$\tilde{\beta}_1 = \sum_{i=1}^{n} a_i Y_i = \beta_0 \sum_{i=1}^{n} a_i + \beta_1 \sum_{i=1}^{n} a_i X_i + \sum_{i=1}^{n} a_i u_i \Rightarrow$

$\tilde{\beta}_1 - \beta_1 = \sum_{i=1}^{n} a_i Y_i = \sum_{i=1}^{n} a_i u_i \Rightarrow \mathrm{var}\left(\tilde{\beta}_1 \middle| X_1, \dots, X_n\right) =$
$\mathrm{var}(\sum_{i=1}^{n} a_i u_i \,|X_1, \dots, X_n) =$
$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \mathrm{cov}\left(u_i, u_j \middle| X_1, \dots, X_n\right) = {\sigma_u}^2 \sum_{i=1}^{n} {a_i}^2$.

Справочно: $\mathrm{var}\left(\hat{\beta}_1 \middle| X_1, \dots, X_n\right) = {\sigma_u}^2 \sum_{i=1}^{n} {\hat{a}_i}^2 = \frac{{\sigma_u}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

Теорема Гаусса-Маркова

Пусть $a_i = \hat{a}_i + d_i \Rightarrow$

$$\sum_{i=1}^{n} {a_i}^2 = \sum_{i=1}^{n} {\hat{a}_i}^2 + 2\sum_{i=1}^{n} \hat{a}_i d_i + \sum_{i=1}^{n} {d_i}^2$$

По определению $\hat{a}_i$ (см. выше) $\Rightarrow$ (покажите) $\sum_{i=1}^{n} \hat{a}_i d_i = 0$

$\Rightarrow {\sigma_u}^2 \sum_{i=1}^{n} {a_i}^2 = {\sigma_u}^2 \sum_{i=1}^{n} {\hat{a}_i}^2 + {\sigma_u}^2 \sum_{i=1}^{n} {d_i}^2 =$
$\mathrm{var}(\hat{\beta}_1 | X_1, \dots, X_n) + {\sigma_u}^2 \sum_{i=1}^{n} {d_i}^2 \Rightarrow$

$\mathrm{var}(\tilde{\beta}_1 | X_1, \dots, X_n) - \mathrm{var}(\hat{\beta}_1 | X_1, \dots, X_n) = {\sigma_u}^2 \sum_{i=1}^{n} {d_i}^2 > 0 \Rightarrow$

Дисперсию любой линейной условной не смещенной оценки β_1 больше оценки МНК■

Тема 3: Проверка гипотез в модели парной линейной регрессии

- Проверка статистических гипотез о коэффициентах регрессии и доверительные интервалы.
- Нарушения предположений теоремы Гаусса-Маркова, их последствия и методы «борьбы» с ними. Использование оцененной модели для прогнозирования.
- Регрессия без свободного члена

Тестирование двусторонних гипотез относительно β_1

$$\underset{}{\hat{Y}_i} = \underset{\left(SE(\hat{\beta}_0)\right)}{\hat{\beta}_0} + \underset{\left(SE(\hat{\beta}_1)\right)}{\hat{\beta}_1} X_i$$

$$\begin{aligned} H_0: &\quad \beta_1 = \beta_{1,0} \\ H_1: &\quad \beta_1 \neq \beta_{1,0} \end{aligned}$$

Тестирование двусторонних гипотез относительно β_1

1. Вычисляем стандартную ошибку $\hat{\beta}_1$ - $\left(SE\left(\hat{\beta}_1\right)\right)$

2. Вычисляем тестовую статистику t^{act}

3. Отвергаем нулевую гипотезу на уровне значимости 5%, если $|t^{act}| > 1{,}96$. Или, эквивалентно, отвергаем нулевую гипотезу, если p-значение меньше 0,05

1. Вычисление стандартной ошибки $\hat{\beta}_1$

$SE(\hat{\beta}_1)$ - оценка $\sigma_{\hat{\beta}_1}$:

$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}},$$

где

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \hat{u}_i{}^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}$$

! $\sigma^2_{\beta_1} = \frac{1}{n}\frac{\text{var}[(X_i - \mu_X)u_i]}{[\text{var}(X_i)]^2}$!

2. Вычисление тестовой статистики t^{act}

$$t^{act} = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE\left(\hat{\beta}_1\right)}$$

3. Отвержение/ не отвержение нулевой гипотезы

Способ 1: Сравнение t^{act} и $t_{crit,\frac{\alpha}{2}\%}$ - $|t^{act}| > \left|t_{crit,\frac{\alpha}{2}\%}\right|$ - отвергаем нулевую гипотезу на уровне значимости $\alpha\%$

Способ 2: Вычисление p-значения:

$$p - \text{value} = \Pr{}_{H_0}\left[\left|\hat{\beta}_1 - \beta_{1,0}\right| > \left|\hat{\beta}_1{}^{act} - \beta_{1,0}\right|\right] =$$

$$= \Pr{}_{H_0}\left[\left|\frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}\right| > \left|\frac{\hat{\beta}_1{}^{act} - \beta_{1,0}}{SE(\hat{\beta}_1)}\right|\right] = \Pr{}_{H_0}[|t| > |t^{act}|]$$

В больших выборках:

$$p - \text{value} = \Pr[|Z| > |t^{act}|] = 2\Phi[-|t^{act}|]$$

Р-значение

р-значение или вероятность значимости – минимальная вероятность отвержения нулевой гипотезы на основе имеющейся выборки в предположении, что она (нулевая гипотеза) верна , т.е. это вероятность совершения ошибки первого рода

Тестирование односторонних гипотез относительно β_1

Левосторонняя альтернатива:

$$H_0: \qquad \beta_1 = \beta_{1,0}$$
$$H_1: \qquad \beta_1 < \beta_{1,0}$$

Правосторонняя альтернатива:

$$H_0: \qquad \beta_1 = \beta_{1,0}$$
$$H_1: \qquad \beta_1 > \beta_{1,0}$$

Тестирование односторонних гипотез относительно β_1

1. Вычисляем стандартную ошибку $\hat{\beta}_1$ - $\left(SE\left(\hat{\beta}_1\right)\right)$

2. Вычисляем тестовую статистику t^{act}

3. Отвергаем нулевую гипотезу на уровне значимости 5%, если $t^{act} < -1{,}645$ или $t^{act} > 1{,}645$. Или, эквивалентно, отвергаем нулевую гипотезу, если p-значение меньше 0,05 (!!!)

3. Отвержение/ не отвержение нулевой гипотезы

Способ 1: Сравнение t^{act} и $t_{crit,\alpha\%}$

Способ 2: Вычисление p-значения в больших выборках:

Левосторонний тест:

$$p - \text{value} = \Pr[Z < t^{act}] = \Phi[t^{act}]$$

Правосторонний тест:

$$p - \text{value} = \Pr[Z > t^{act}] = \Phi[t^{act}]$$

Тестирование двусторонних гипотез относительно β_0

$$H_0: \qquad \beta_0 = \beta_{0,0}$$
$$H_1: \qquad \beta_0 \neq \beta_{0,0}$$

Далее – аналогично процедуре для β_1

Различие:

$$\hat{\sigma}^2_{\hat{\beta}_0} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n} \widehat{H}_i^{\,2}\, \hat{u}_i^{\,2}}{\left[\frac{1}{n}\sum_{i=1}^{n} \widehat{H}_i^{\,2}\right]^2},$$

где $\widehat{H}_i = 1 - \left[\bar{X} / \frac{1}{n}\sum_{i=1}^{n} X_i^{\,2}\right] X_i$

! $\sigma^2_{\beta_0} = \frac{1}{n}\frac{\mathrm{var}[H_i u_i]}{[\mathrm{E}(H_i^2)]^2}$, где $H_i = 1 - \left[\frac{\mu_X}{E(X_i^2)}\right] X_i$!

Пример: размер класса и результаты тестов в Калифорнии

$$\widehat{TestScore} = \underset{(10,4)}{698,9} - \underset{(0,52)}{2,28} \times STR, \quad R^2 = 0,051, SER = 18,6$$

Построение доверительных интервалов для коэффициентов регрессии

95%-й двухсторонний доверительный интервал (в больших выборках):

- для β_1
$$\left[\hat{\beta}_1 - 1{,}96 \cdot SE\left(\hat{\beta}_1\right); \hat{\beta}_1 + 1{,}96 \cdot SE\left(\hat{\beta}_1\right)\right]$$
- для β_0
$$\left[\hat{\beta}_0 - 1{,}96 \cdot SE\left(\hat{\beta}_0\right); \hat{\beta}_0 + 1{,}96 \cdot SE\left(\hat{\beta}_0\right)\right]$$
- для односторонних гипотез – аналогично (с заменой 1,96 на 1,645)

Доверительные интервалы для оценки влияния изменения X

Пусть X изменяется на Δx. Тогда Y изменится на $\Delta y = \beta_1 \Delta x$.

Тогда 95%-й доверительный интервал для $\beta_1 \Delta x$:

$$\left[\left\{\hat{\beta}_1 - 1{,}96 \cdot SE\left(\hat{\beta}_1\right)\right\}\Delta x; \left\{\hat{\beta}_1 + 1{,}96 \cdot SE\left(\hat{\beta}_1\right)\right\}\Delta x\right]$$

Регрессия с бинарной объясняющей переменной

Предположение №1: условное распределение u_i относительно X_i имеет нулевое среднее: $E(u_i|X_i) = 0$

Предположение №2: $(X_i, Y_i), i = 1, \ldots, n,$ независимы и одинаково распределены $(\mathrm{i.i.d.})$

Предположение №3: большие выбросы маловероятны: X_i и Y_i имеют ненулевые конечные четвертые моменты

Регрессия с бинарной объясняющей переменной

$$E(Y_i|D_i = 0) = \beta_0$$

и

$$E(Y_i|D_i = 1) = \beta_0 + \beta_1$$

Тогда β_1 - коэффициент регрессии - разность между двумя условными средними

Степени свободы

Число степеней свободы – минимальное количество элементов варьирования, которые могут принимать произвольные значения, не изменяющие заданных характеристик.

Пример:

1. Пусть дано 7 чисел со средней, равной 5 (т. е. в сумме 35). Задача: подобрать другие 7 чисел со средней, равной 5. Произвольно можем выбрать только 6 чисел. Число с. с. здесь равно $7 - 1 = 6$, или в общем случае: n .
2. При вычислении дисперсии по выборке из n наблюдений число степеней свободы равно n-1, т.к. 1 степень свободы мы уже использовали при расчете среднего.