

Семинар 10.

Гетероскедастичность.

1. Рассмотрим модель регрессии

$$y = X\beta + \varepsilon,$$

$$\mathbb{E}(\varepsilon) = 0, \text{Var}(\varepsilon) = \Omega.$$

(a) Проверьте несмещённость оценки

$$\hat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y.$$

Решение.

$$\begin{aligned}\mathbb{E}[\hat{\beta}_{GLS}] &= \mathbb{E}[(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y] = \mathbb{E}[(X'\Omega^{-1}X)^{-1}X'\Omega^{-1})(X\beta + \varepsilon)] = \\ &= \beta + \mathbb{E}[(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\varepsilon] = \beta.\end{aligned}$$

(б) Проверьте равенство

$$\text{Var}(\hat{\beta}_{GLS}) = (X'\Omega^{-1}X)^{-1}.$$

Решение.

Обозначим $\hat{\beta}_{GLS} = Ay$, где $A = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}$. Тогда

$$\begin{aligned}\text{Var}(\hat{\beta}_{GLS}) &= \text{Var}(Ay) = A\text{Var}(y)A' = \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\text{Var}(X\beta + \varepsilon)\Omega^{-1}X(X'\Omega^{-1}X)^{-1} = \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\text{Var}(\varepsilon)\Omega^{-1}X(X'\Omega^{-1}X)^{-1} = \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\Omega\Omega^{-1}X(X'\Omega^{-1}X)^{-1} = (X'\Omega^{-1}X)^{-1}.\end{aligned}$$

2. Найдите наиболее эффективную оценку коэффициента β_1 для модели

$$y_i = \beta_1 + \varepsilon_i,$$

$$\mathbb{E}(\varepsilon_i) = 0, \mathbb{E}(\varepsilon_i\varepsilon_j) = 0, \text{Var}(\varepsilon_i) = \sigma_\varepsilon^2 x_i, x_i > 0$$

в классе линейных несмещённых оценок. Рассчитайте дисперсию этой оценки и сравните её с дисперсией МНК-оценки.

Решение.

Для начала найдём МНК-оценку и её дисперсию:

$$\hat{\beta}_1^{OLS} = \bar{y},$$

$$\text{Var}(\hat{\beta}_1^{OLS}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n y_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(y_i) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(\varepsilon_i) = \frac{\sigma_\varepsilon^2}{n^2} \sum_{i=1}^n \frac{1}{x_i}.$$

Наиболее эффективной оценкой при гетероскедастичных ошибках будут оценки, полученные взвешенным методом наименьших квадратов. Наша цель — сделать одинаковыми дисперсию ошибки для всех наблюдений. Идея такая: стандартизируем ошибки ε_i , тогда дисперсия ошибки станет равной 1 для всех наблюдений. Чтобы добиться этого домножим обе части исходной модели на $\frac{1}{\sqrt{\sigma_\varepsilon^2/x_i}}$:

$$\frac{y_i}{\sqrt{\sigma_\varepsilon^2/x_i}} = \frac{1}{\sqrt{\sigma_\varepsilon^2/x_i}} \beta_1 + \frac{\varepsilon_i}{\sqrt{\sigma_\varepsilon^2/x_i}}.$$

Запись выше эквивалентна следующей записи:

$$\frac{y_i \sqrt{x_i}}{\sigma_\varepsilon} = \frac{\sqrt{x_i}}{\sigma_\varepsilon} \beta_1 + \frac{\sqrt{x_i} \varepsilon_i}{\sigma_\varepsilon}.$$

В новой модели ошибки $\frac{\sqrt{x_i} \varepsilon_i}{\sigma_\varepsilon}$ имеют одинаковую дисперсию для всех наблюдений, равную единице, так как $\text{Var}\left(\frac{\sqrt{x_i} \varepsilon_i}{\sigma_\varepsilon}\right) = \frac{x_i}{\sigma_\varepsilon^2} \frac{\sigma_\varepsilon^2}{x_i} = 1$.

Следовательно, в последней модели ошибки уже являются гомоскедастичными, поэтому можем использовать МНК для получения эффективных в классе линейных по y , несмещённых оценок. Стоит отметить, что теперь модель является моделью парной регрессии без константы. Вспомнив как выглядит оценка коэффициента наклона в модели парной регрессии без константы, получим следующий результат:

$$\hat{\beta}_1^{WLS} = \frac{\sum_{i=1}^n \frac{y_i \sqrt{x_i} \sqrt{x_i}}{\sigma_\varepsilon \sigma_\varepsilon}}{\sum_{i=1}^n \left(\frac{\sqrt{x_i}}{\sigma_\varepsilon}\right)^2} = \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i}.$$

Теперь рассчитаем дисперсию $\hat{\beta}_1^{WLS}$:

$$\begin{aligned} \text{Var}\left(\hat{\beta}_1^{WLS}\right) &= \frac{\sum_{i=1}^n x_i^2 \text{Var}(y_i)}{\left(\sum_{i=1}^n x_i\right)^2} = \frac{\sum_{i=1}^n x_i^2 \text{Var}(\varepsilon_i)}{\left(\sum_{i=1}^n x_i\right)^2} = \frac{\sum_{i=1}^n x_i^2 \frac{\sigma_\varepsilon^2}{x_i}}{\left(\sum_{i=1}^n x_i\right)^2} = \\ &= \frac{\sigma_\varepsilon^2 \sum_{i=1}^n x_i}{\left(\sum_{i=1}^n x_i\right)^2} = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n x_i}. \end{aligned}$$

Сравним дисперсии OLS и WLS оценок:

$$\frac{\text{Var}(\hat{\beta}_1^{OLS})}{\text{Var}(\hat{\beta}_1^{WLS})} = \frac{\sigma_\varepsilon^2 \sum_{i=1}^n \frac{1}{x_i}}{\sum_{i=1}^n x_i} = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n x_i} = \frac{\sum_{i=1}^n \frac{1}{x_i} \sum_{i=1}^n x_i}{n^2}.$$

К последнему равенству применим неравенство Коши–Буняковского, в результате чего получаем, что

$$\sum_{i=1}^n \frac{1}{x_i} \sum_{i=1}^n x_i \geq n^2,$$

то есть $\text{Var}(\hat{\beta}_1^{OLS}) \geq \text{Var}(\hat{\beta}_1^{WLS})$.

3. Рассмотрим следующую регрессионную модель, в которой $2n$ наблюдений разбиты на две равные группы по n наблюдений в каждой:

$$y = X\beta + \varepsilon,$$

$$\mathbb{E}(\varepsilon) = 0; \text{Cov}(\varepsilon_t, \varepsilon_s) = 0, t \neq s$$

$$\text{Var}(\varepsilon_t) = \sigma_1^2, t = 1, \dots, n; \text{Var}(\varepsilon_t) = \sigma_2^2, t = n + 1, \dots, 2n.$$

Введём естественное разбиение матриц на блоки:

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}.$$

- (а) Пусть $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}$ — МНК-оценки вектора коэффициентов β по первой группе наблюдений, по второй группе наблюдений и по всем $2n$ наблюдениям соответственно. Покажите, что $\hat{\beta}$ есть "взвешенное среднее" оценок $\hat{\beta}_1$ и $\hat{\beta}_2$, то есть $\hat{\beta} = L_1\hat{\beta}_1 + L_2\hat{\beta}_2$, где L_1 и L_2 — $k \times k$ матрицы такие, что $L_1 + L_2 = I_k$.
- (б) Выведите следующие формулы для ОМНК-оценок:

$$\hat{\beta}_{GLS} = \left(\frac{X_1'X_1}{\sigma_1^2} + \frac{X_2'X_2}{\sigma_2^2} \right)^{-1} \left(\frac{X_1'y_1}{\sigma_1^2} + \frac{X_2'y_2}{\sigma_2^2} \right),$$

$$\text{Var}(\hat{\beta}_{GLS}) = \left(\frac{X_1'X_1}{\sigma_1^2} + \frac{X_2'X_2}{\sigma_2^2} \right)^{-1}.$$

- (в) Покажите, что $\hat{\beta}_{GLS}$ также является "взвешенным средним" оценок $\hat{\beta}_1$ и $\hat{\beta}_2$ в том смысле, что существуют $k \times k$ матрицы Λ_1 и Λ_2 такие, что $\hat{\beta}_{GLS} = \Lambda_1\hat{\beta}_1 + \Lambda_2\hat{\beta}_2$, $\Lambda_1 + \Lambda_2 = I_k$.
- (г) Опишите процедуру получения FGLS-оценок для данной модели.

4. В файле "Heterosk_5.xlsx" содержатся данные о 150 пользователях некоторого мобильного приложения:

- *Expend* — затраты пользователя на покупки в мобильном приложении;
- *Time* — среднее время, проведенное пользователем в приложении (мин);
- *Age1* — 1 для пользователей от 18 до 21 года, 0 иначе;
- *Age2* — 1 для пользователей от 22 до 25 года, 0 иначе;
- *Age3* — 1 для пользователей от 26 до 29 года, 0 иначе;
- *Age4* — 1 для пользователей от 30 до 34 года, 0 иначе;
- *Age5* — 1 для пользователей от 35 лет и старше, 0 иначе;
- *MPrice* — рыночная стоимость используемой модели смартфона.

Для изучения влияния характеристик, влияющих на затраты пользователя в приложении была рассмотрена следующая модель регрессии:

$$Expend_i = \beta_1 + \beta_2 Time_i + \beta_3 MPrice_i + \beta_4 Age1_i + \beta_5 Age2_i + \beta_6 Age3_i + \beta_7 Age4_i + \varepsilon_i. (1)$$

- Оцените модель регрессии (1) с помощью МНК.
- На основе результатов оценивания из предыдущего пункта проанализируйте наличие гетероскедастичности в данных.
- В случае идентификации гетероскедастичности в данных переоцените модель (1) с помощью ВМНК.
- Используя робастные при гетероскедастичности стандартные ошибки оценок параметров, переоцените модель (1) с помощью МНК. Сравните полученные результаты с моделями из пунктов (а) и (в).