Регрессионный анализ социально-экономических процессов

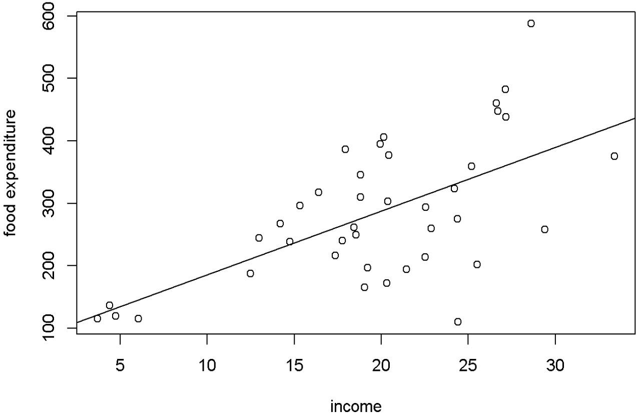
Гетероскедастичность

13 октября 2025

План:

- определить «проблему» – особенность данных
- почему заслуживает внимание
- источники гетероскедастичности
- диагностика: как выявить?
- нетипичные наблюдения как источник гетероскедастичности
- корректировки: что делать?

Иллюстрация гетероскедастичности



Последствия гетероскедастичности

Согласно одному из допущений Гаусса-Маркова, условная вариация ошибок при заданных значениях предикторов является постоянной (гомоскедастичность). Если это допущение нарушается, то мы имеем дело с гетероскедастичностью.

Последствия гетероскедастичности:

1. неэффективность оценок, при этом оценки остаются состоятельными и несмещенными
2. распределение статистик другое

Последствия гетероскедастичности

Согласно одному из допущений Гаусса-Маркова, условная вариация ошибок при заданных значениях предикторов является постоянной (гомоскедастичность). Если это допущение нарушается, то мы имеем дело с гетероскедастичностью.

Последствия гетероскедастичности:

1. неэффективность оценок, при этом оценки остаются состоятельными и несмещенными
2. распределение статистик другое

Итог: главная проблема

Эти последствия делают проверку гипотез о незначимости коэффициентов проблематичной

Откуда берется гетероскедастичность?

Источники гетероскедастичности: примеры

Откуда берется гетероскедастичность?

Источники гетероскедастичности: примеры

1. работаем с объектами разного «размера»
2. нетипичные наблюдения
3. неверно определена функциональная форма взаимосвязи
4. пропущены важные факторы
5. разные методики сбора данных

Как выявить гетероскедастичность?

Диагностики

Как выявить гетероскедастичность?

Диагностики

1. еще до диагностик важно обратиться к самим данным и их структуре
2. визуализация
3. формальные тесты

Диагностики, основанные на визуализации

Графики

- OY – зависимая переменная, OX – предиктор
- OY – зависимая переменная, OX – предсказанное значение ($\hat{y}$)
- OY – остатки в квадрате, OX – предиктор
- OY – остатки в квадрате, OX – предсказанное значение ($\hat{y}$)

Изменяется ли вариация при разных значениях X?

Диагностики: тест Уайта

Предпосылки

- этот тест используется преимущественно на больших выборках (от 50 наблюдений), иначе низкая мощность критерия

Шаги реализации:

- оцениваем модель и сохраняем остатки ($\hat{e}$)
- строим дополнительную модель остатков в квадрате (в качестве зависимой переменной) на все исходные предикторы, их квадраты и попарные произведения
- сохраняем из дополнительной модели R^2
- считаем статистику критерия: $nR^2 \sim \chi^2_k$, где k – количество предикторов в дополнительной модели

Диагностики: тест Голдфелда – Квандта

Предпосылки

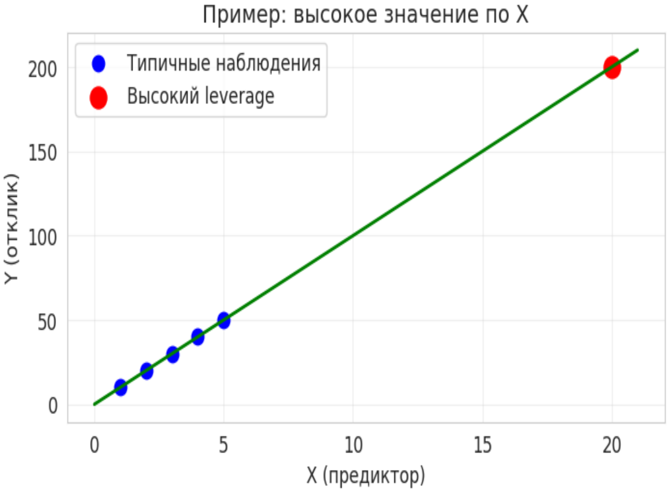
- нормальное распределение ошибок на малой по объему выборке (если n < 30 и асимметричное распределение, не можем доверять результатам)

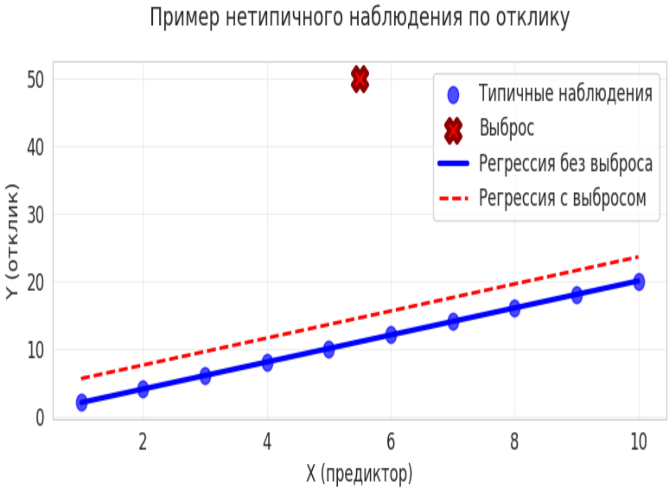
Шаги реализации:

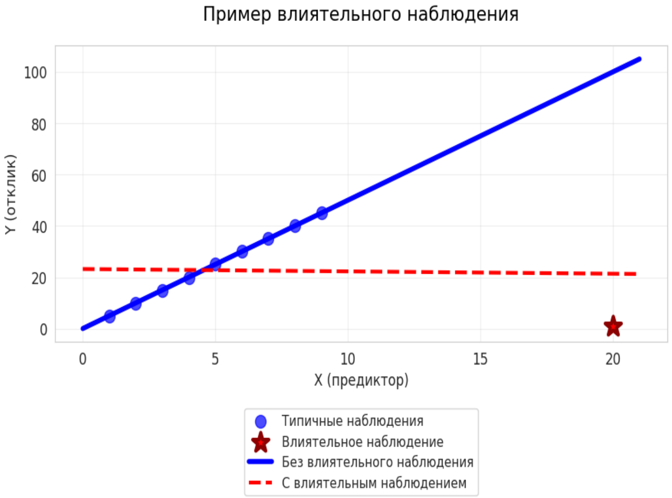
- упорядочиваем наблюдения по X и из середины исключаем часть наблюдений
- оцениваем исходную модель на оставшихся первом и втором сегментах упорядоченной выборки и сохраняем RSS_1 и RSS_2
- считаем статистику критерия: $\frac{RSS_1/(n_1-k-1)}{RSS_2/(n_2-k-1)} \sim F(n_1-k-1, n_2-k-1)$, где k –

Нетипичные наблюдения

1. по предикторам: потенциал влияния (leverage - «рычаг») – экстремальное значение по предиктору, отклоняющееся от среднего по предиктору значимым образом
2. по зависимой переменной: выбросы (outliers) – нетипичное значение отклика для данного значения предиктора
3. по комбинации значений предикторов и зависимой переменной: влиятельные наблюдения (influential observations)


Пример: высокое значение по X
Типичные наблюдения
Высокий leverage
Y (отклик)
200
150
100
50
0
0
5
10
15
20
X (предиктор)


Пример нетипичного наблюдения по отклику
Типичные наблюдения
Выброс
Регрессия без выброса
Регрессия с выбросом
Y (отклик)
X (предиктор)
0
10
20
30
40
50
2
4
6
8
10


Пример влиятельного наблюдения
Y (отклик)
X (предиктор)
100
80
60
40
20
0
0
5
10
15
20
Типичные наблюдения
Влиятельное наблюдение
Без влиятельного наблюдения
С влиятельным наблюдением

Нетипичные наблюдения: последствия

1. по предикторам: изменения в оценках коэффициентов
2. по зависимой переменной: уменьшается R^2, большие значения остатков
3. влиятельные наблюдения: значимым образом меняются оценки коэффициенты (вплоть до перевернутого знака), результаты оценивания модели неустойчивы

Нетипичные наблюдения по предикторам

Для того, чтобы выявить такие наблюдения, нам нужно будет обратиться к матрице проекции (Hat-matrix)

$$\vec{\hat{y}} = X\vec{\hat{\beta}} = X(X^T X)^{-1} X^T \vec{y}$$

$$H = X(X^T X)^{-1} X^T$$

Матрица H:

- Симметричная: $H = H^T$
- Идемпотентная: $H = H^2$
- $tr(H) = p$ (p – количество параметров в модели)
- $0 \leq h_{ii} \leq 1$
- Сумма элементов по строке $= 1$

Нетипичные наблюдения по предикторам

h_{ii} – диагональный элемент матрицы H (показывает потенциал влияния наблюдения на соответствующее предсказанное значение)
$\frac{p}{n}$ – среднее значение диагональных элементов матрицы H

Критерий определения нетипичных наблюдений по предикторам:

$$h_{ii} > \frac{3p}{n}$$

Или используется более низкая граница:

$$h_{ii} > \frac{2p}{n}$$

Нетипичные наблюдения по зависимой переменной: выбросы

Для диагностики выбросов используются стьюдентизированные остатки (studentized residuals):

$$stres_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}_{-i}\sqrt{1 - h_{ii}}}$$

$$\hat{\sigma}_{-i} = \sqrt{\frac{RSS_{-i}}{df}}$$

Если $|stres_i| > 3$, тогда классифицируем i-ое наблюдение как выброс (бывает, что используется более низкая граница 2 вместо 3)

Влиятельные наблюдения

Расстояние Кука:

$$D_i = \frac{\hat{\epsilon_i}^2}{p \cdot \hat{\sigma}^2} \cdot \frac{h_{ii}}{(1 - h_{ii})^2}$$

$p = k + 1$ (количество параметров в регрессионной модели)

Эмпирическое правило:

$$D_i > 4/N$$

трактуется как влиятельное наблюдение

Влиятельные наблюдения

Мера DFBETA показывает, насколько изменится тот или иной коэффициент в регрессионной модели при удалении i-го наблюдения

$$DFBETA_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j(-i)}}{SE(\hat{\beta}_{j(-i)})}$$

$SE(\hat{\beta}_{j(-i)})$ - стандартная ошибка оценки коэффициента, полученная после оценивания модели на массиве без i-го наблюдения Эмпирическое правило:

$$|DFBETA| > \frac{2}{\sqrt{N}}$$

трактуется как влиятельное наблюдение

Что делать в случае гетероскедастичности?

Способы корректировки:

- преобразовать сами переменные (к примеру, логарифмировать)

Что делать в случае гетероскедастичности?

Способы корректировки:

- преобразовать сами переменные (к примеру, логарифмировать)
- скорректировать саму спецификацию модели

Что делать в случае гетероскедастичности?

Способы корректировки:

- преобразовать сами переменные (к примеру, логарифмировать)
- скорректировать саму спецификацию модели
- поправить стандартные ошибки (heteroskedasticity-consistent standard errors - состоятельные в условиях гетероскедастичности стандартные ошибки)

Что делать в случае гетероскедастичности?

Способы корректировки:

- преобразовать сами переменные (к примеру, логарифмировать)
- скорректировать саму спецификацию модели
- поправить стандартные ошибки (heteroskedasticity-consistent standard errors - состоятельные в условиях гетероскедастичности стандартные ошибки)
- поправить формулу МНК-оценки - использовать обобщенный метод наименьших квадратов (GLS – generalized least squares)

Что делать в случае гетероскедастичности?

Способы корректировки:

- преобразовать сами переменные (к примеру, логарифмировать)
- скорректировать саму спецификацию модели
- поправить стандартные ошибки (heteroskedasticity-consistent standard errors - состоятельные в условиях гетероскедастичности стандартные ошибки)
- поправить формулу МНК-оценки - использовать обобщенный метод наименьших квадратов (GLS – generalized least squares)
- проверить результаты оценивания модели на устойчивость к нетипичным наблюдениям

Ковариационная матрица ошибок

Случай гомоскедастичности и отсутствия автокорреляции:

$$\begin{array}{cc} & \begin{matrix} \varepsilon_1 & \varepsilon_2 & \dots & \varepsilon_n \end{matrix} \\ \begin{matrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{matrix} & \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix} \end{array}$$

Случай гетероскедастичности и отсутствия автокорреляции:

$$\begin{array}{cc} & \begin{matrix} \varepsilon_1 & \varepsilon_2 & \dots & \varepsilon_n \end{matrix} \\ \begin{matrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{matrix} & \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix} \end{array}$$

Оценка ковариационной матрицы ошибок

Случай гомоскедастичности и отсутствия автокорреляции:

$$\begin{array}{cc}
 & \begin{array}{cccc} \hat{\varepsilon}_1 & \hat{\varepsilon}_2 & \dots & \hat{\varepsilon}_n \end{array} \\
\begin{array}{c} \hat{\varepsilon}_1 \\ \hat{\varepsilon}_2 \\ \dots \\ \hat{\varepsilon}_n \end{array} & \begin{pmatrix} \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-k-1} & 0 & \dots & 0 \\ 0 & \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-k-1} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-k-1} \end{pmatrix}
\end{array}$$

$$\longrightarrow \quad Var(\hat{\beta}) = (X^T X)^{-1} X^T \sigma^2 I X (X^T X)^{-1} =$$
$$\sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}$$

$$\widehat{Var}(\hat{\beta}) = s^2 (X^T X)^{-1}, \text{ где } s^2 = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{n-k-1} = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-k-1}$$

Вспомогательная матрица весов HC3

HC3 – состоятельные в условиях гетероскедастичности стандартные ошибки (третья версия в последовательности предложенных робастных оценок)

Случай гетероскедастичности и отсутствия автокорреляции:

$$\hat{\Omega}_{HC3} = \begin{pmatrix} \frac{\hat{\varepsilon}_1^2}{(1-h_{11})^2} & 0 & \cdots & 0 \\ 0 & \frac{\hat{\varepsilon}_2^2}{(1-h_{22})^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\hat{\varepsilon}_n^2}{(1-h_{nn})^2} \end{pmatrix}$$

$$\longrightarrow \quad \widehat{Var}_{HC3}(\hat{\beta}) = (X^T X)^{-1} X^T \hat{\Omega}_{HC3} X (X^T X)^{-1}$$

Обобщенный метод наименьших квадратов

Вместо робастной оценки дисперсии предлагаем альтернативный способ оценивания коэффициентов в регрессионной модели – ОМНК (обобщенный МНК)

Σ – ковариационная матрица ошибок (учитывает случай гетероскедастичности)

Матрица P такая, что справедливо следующее: $\Sigma^{-1} = P^T P$

$$\Sigma^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_n^2} \end{pmatrix} \quad P = \begin{pmatrix} \frac{1}{\sigma_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_n} \end{pmatrix}$$

Обобщенный метод наименьших квадратов

Используем матрицу P для предварительного преобразования:

$$PY = PX\beta + P\varepsilon$$

$$Var(P\varepsilon) = PVar(\varepsilon)P^T = I$$

$$\hat{\beta}_{OLS} = (X^TX)^{-1}X^TY$$

$$\hat{\beta}_{GLS} = ((PX)^T(PX))^{-1}(PX)^T(PY) = (X^TP^TPX)^{-1}X^TP^TPY = \\ = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$$

Такой подход (ОМНК / GLS) будет давать наиболее эффективные оценки среди линейных несмещенных оценок, но на практике не все так просто...см. далее идею FGLS

Реализуемый ОМНК

На практике мы имеем дело с реализуемым ОМНК (FGLS – feasible generalized least squares)

$\hat{\beta}_{FGLS} = (X^T \hat{\Sigma}^{-1} X)^{-1} X^T \hat{\Sigma}^{-1} Y$
То есть, здесь нам предстоит оценить (!) ковариационную матрицу ошибок
Оценки FGLS асимптотически состоятельны и эффективны. Однако при малых по объему выборках мы мало что можем сказать о свойствах оценок FGLS. Стоит понимать, что на конечных выборках может быть значимое смещение оценок. Кроме этого, при малых n могут быть даже менее эффективны по сравнению с исходными OLS-оценками