Теория вероятностей и математическая статистика

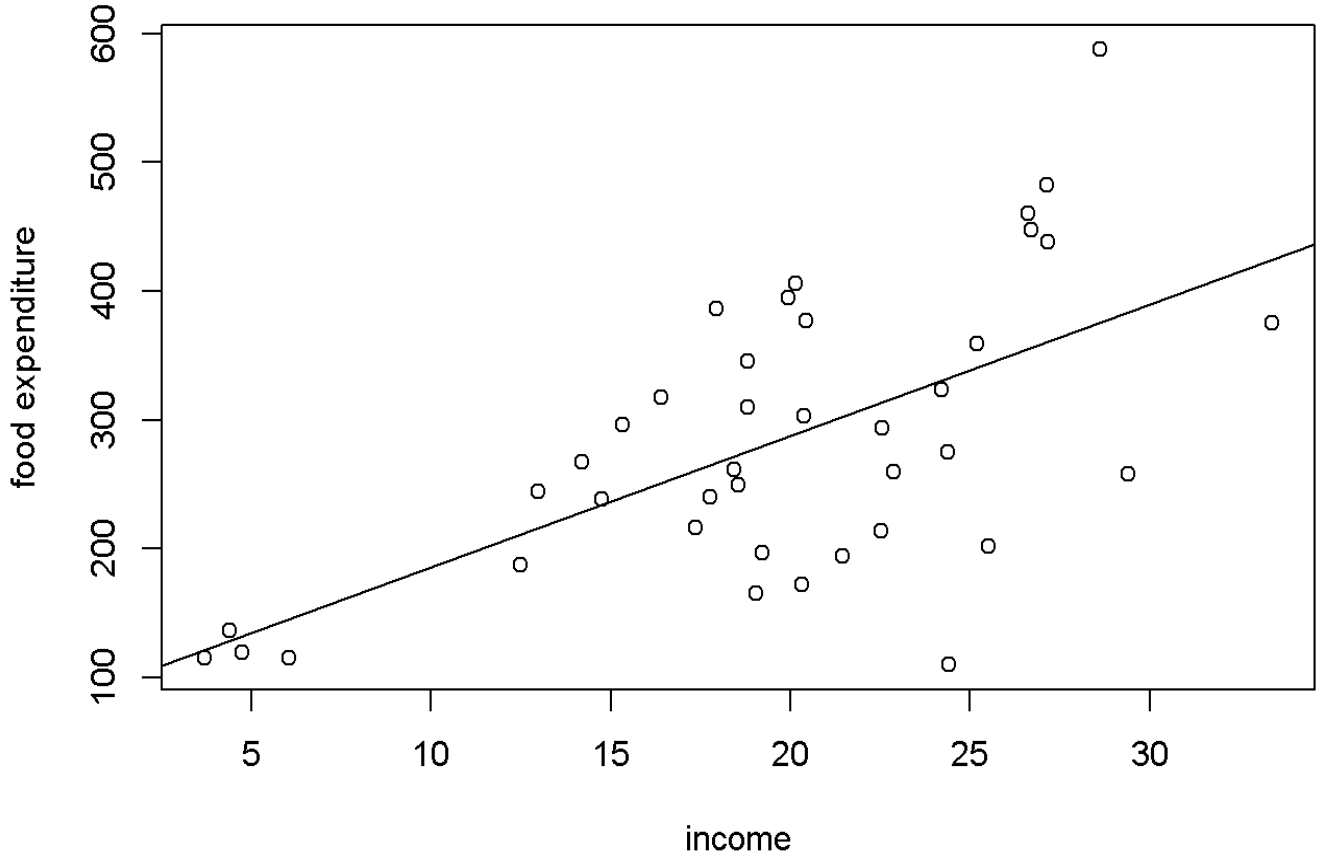
Гетероскедастичность в классической линейной регрессии

26 февраля 2021

Планы:

- определить «проблему» – особенность данных
- почему заслуживает внимание
- источники гетероскедастичности
- диагностика: как выявить?
- корректировки: что делать?
- реализация в R

Иллюстрация гетероскедастичности



Why should we care?

Последствия гетероскедастичности

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

Why should we care?

Последствия гетероскедастичности

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

Итог: главная проблема

Эти последствия делают проверку гипотез о незначимости коэффициентов проблематичной.

Откуда берется гетероскедастичность?

Источники гетероскедастичности

Откуда берется гетероскедастичность?

Источники гетероскедастичности

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 Вашим теоретическим предпосылкам, они и будут самыми важными для того, чтобы принять решение о том, как работать далее с оценками модели
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

- большая выборка
- отсутствуют требования о нормальности распределения ошибок

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: $nR^2 \sim \chi^2_{k-1}$, где k – количество параметров в дополнительной модели

Диагностики: Goldfeld-Quandt test

Предпосылки

- можно на маленьких выборках
- нормальное распределение ошибок

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:
$$\frac{RSS_1/(n_1-k)}{RSS_2/(n_2-k)} \sim F(n_1-k, n_2-k)$$

Что делать?

2 способа корректировки:

- поправить стандартные ошибки (heteroskedasticity-consistent standard errors)

Что делать?

2 способа корректировки:

- поправить стандартные ошибки (heteroskedasticity-consistent standard errors)
- поправить формулу МНК-оценки GLS – generalized least squares (будем изучать на 3 курсе)