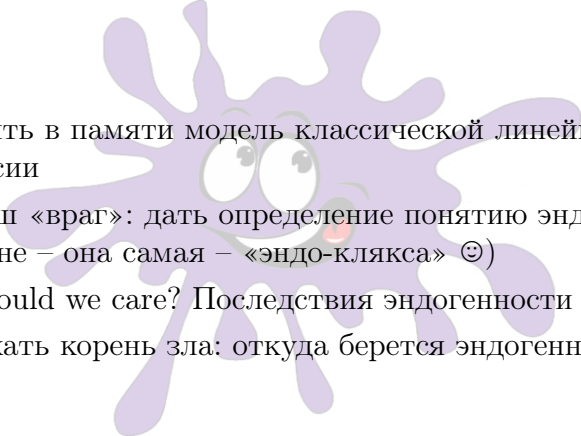


Регрессионный анализ: продолжение

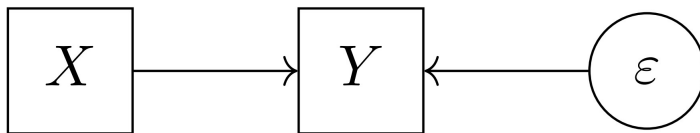
Понятие эндогенности и ее проявления

8 ноября 2024

Планы на сегодня:

- 
- освежить в памяти модель классической линейной регрессии
 - кто наш «враг»: дать определение понятию эндогенности (на фоне – она самая – «эндо-клякса» 😊)
 - why should we care? Последствия эндогенности
 - где искать корень зла: откуда берется эндогенность?

Путевая диаграмма: регрессия



Y – зависимая переменная (отклик);

X – независимая переменная (объясняющая переменная / предиктор);

ε – ошибка

Модель классической линейной регрессии

К каким оценкам мы стремимся?

Модель классической линейной регрессии

К каким оценкам мы стремимся?

Статистическая инференция посредством регрессионного анализа (оценка VS генеральный параметр, задача – перенести выводы на широкую совокупность).

В связи с этим мы хотим получить оценки:

Модель классической линейной регрессии

К каким оценкам мы стремимся?

Статистическая инференция посредством регрессионного анализа (оценка VS генеральный параметр, задача – перенести выводы на широкую совокупность).

В связи с этим мы хотим получить оценки:

- несмещенные (в среднем оценка равна генеральному параметру)

Модель классической линейной регрессии

К каким оценкам мы стремимся?

Статистическая инференция посредством регрессионного анализа (оценка VS генеральный параметр, задача – перенести выводы на широкую совокупность).

В связи с этим мы хотим получить оценки:

- несмещенные (в среднем оценка равна генеральному параметру)
- эффективные (в простом варианте – минимальная вариация оценки)

Модель классической линейной регрессии

К каким оценкам мы стремимся?

Статистическая inferencia посредством регрессионного анализа (оценка VS генеральный параметр, задача – перенести выводы на широкую совокупность).

В связи с этим мы хотим получить оценки:

- несмещенные (в среднем оценка равна генеральному параметру)
- эффективные (в простом варианте – минимальная вариация оценки)
- состоятельные (при увеличении размера выборки оценки, приближающиеся по вероятности к генеральным параметрам)

Когда МНК дает хорошие результаты

Если ошибки в регрессии удовлетворяют особым условиям, то МНК-оценки несмещенные, состоятельные и наиболее эффективные среди линейных оценок. Об этих условиях – см. далее.

Какие должны быть ошибки, чтобы МНК давало желаемые оценки

Требования

- $Var(e_i|x_i) = const$ гомоскедастичность

Какие должны быть ошибки, чтобы МНК давало желаемые оценки

Требования

- $Var(e_i|x_i) = const$ гомоскедастичность
- $Cov(e_i, e_j|x_i) = 0$ отсутствие автокорреляции

Какие должны быть ошибки, чтобы МНК давало желаемые оценки

Требования

- $Var(e_i|x_i) = const$ гомоскедастичность
- $Cov(e_i, e_j|x_i) = 0$ отсутствие автокорреляции
- $Cov(e_i, x_i) = 0$ экзогенность (!)

Эндогенность: определение

формальное определение

Эндогенность – это случай нарушения условия $Cov(e_i, x_i) = 0$

Эндогенность: определение

формальное определение

Эндогенность – это случай нарушения условия $Cov(e_i, x_i) = 0$

ЧТО ЗА ЭТИМ СТОИТ

В широком смысле эндогенность – проблема пропущенных существенных переменных.

В чем проблема?

Последствия эндогенности

Мы получаем смещенные и несостоятельные оценки при применении классического МНК.

В чем проблема?

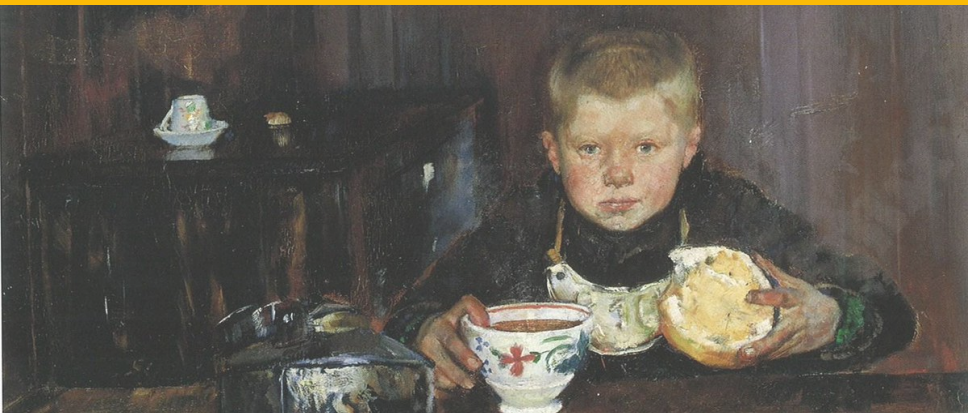
Последствия эндогенности

Мы получаем смещенные и несостоятельные оценки при применении классического МНК.

Вопросы для самопроверки

- Что называется смещенной оценкой?
- Что называется несостоятельной оценкой?
- В чем ключевая идея МНК (OLS)?

История 1. Школьные обеды



Наблюдение: в школах с бесплатными обедами ученики демонстрируют более низкую успеваемость по сравнению со школами, в которых не реализуется программа бесплатного школьного питания.

История 2. Эффект Матфея



Р. Мертон: «Учёные преувеличивают достижения своих коллег, уже заслуживших себе репутацию, а достижения учёных, ещё не получивших известности, они, как правило, преуменьшают или вообще не признают»

История 3. Некоторая IT-компания



Известно, что в некоторой IT-компании мужчины и женщины сотрудники имеют одинаковый уровень заработной платы. Означает ли это, что в данной компании нет дискриминации в заработной плате по гендерному признаку? Опишите разные ситуации.

Почему предикторы и ошибки могут быть зависимыми (1)

Пропущен важный фактор (omitted variable bias)

Не включили значимый показатель, который влияет как на зависимую переменную, так и на те объясняющие переменные, которые уже включены в модель. Значимая зависимость предикторов и пропущенных факторов приводит к смещенности оценок.

Почему предикторы и ошибки могут быть зависимыми (1)

Пропущен важный фактор (omitted variable bias)

Не включили значимый показатель, который влияет как на зависимую переменную, так и на те объясняющие переменные, которые уже включены в модель. Значимая зависимость предикторов и пропущенных факторов приводит к смещенности оценок.

Почему мы можем что-то пропустить?

- недоработка в теории
- отсутствие данных по необходимым показателям
- латентные концентры

Почему предикторы и ошибки могут быть зависимыми (2)

Selection bias

Для анализа доступна только подвыборка с определенными значениями характеристик. Если эти характеристики влияют на изучаемые переменные, то оценки смещенные.

Почему предикторы и ошибки могут быть зависимыми (2)

Selection bias

Для анализа доступна только подвыборка с определенными значениями характеристик. Если эти характеристики влияют на изучаемые переменные, то оценки смещенные.

Почему может возникать selection bias

- проблема дизайна исследования
- самоотбор
- non-response bias

Почему предикторы и ошибки могут быть зависимыми (3)

Post-treatment bias

При отборе контрольных переменных надо помнить, что они должны влиять и на зависимую переменную, и на ключевой предиктор. Если x_i влияет, наоборот, на контрольную переменную, то возникает смещение в оценках при ключевых предикторах (post-treatment bias).

Почему предикторы и ошибки могут быть зависимыми (4)

Что на что влияет? Simultaneity problem

Неоднозначность направления причинно-следственной связи ключевых предикторов и отклика

Почему предикторы и ошибки могут быть зависимыми (5)

Ошибки измерения

Проблема: Включенные предикторы измерены с ошибкой, что может происходить вследствие неверной операционализации, неадекватного инструмента измерения, попытки измерить латентный (ненаблюдаемый) концент.

Формальное представление в спецификации модели: смещение

$$y_i = b_0 + b_1 x_i + e_i$$

$$y_i = a_0 + a_1 (x_i + v_i) + e_i$$

Мы хотим узнать влияние x_i на отклик. Но у нас есть только z_i , который неаккуратно измеряет x_i : $z_i = x_i + v_i$

