Многомерный статистический анализ в политологии

Понятие эндогенности и ее проявления в регрессионном анализе

8 ноября 2019

ЧАСТЬ 1. Эндогенность. Планы:

- кто наш «враг»: дать определение понятию эндогенности
- why should we care? Последствия эндогенности
- где искать корень зла: откуда берется эндогенность?

Регрессионный анализ

К каким оценкам мы стремимся?

Регрессионный анализ

К каким оценкам мы стремимся?

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

В связи с этим мы хотим получить оценки:

Регрессионный анализ

К каким оценкам мы стремимся?

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

В связи с этим мы хотим получить оценки:

- несмещенные (в среднем оценка равна генеральному параметру)

Регрессионный анализ

К каким оценкам мы стремимся?

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

В связи с этим мы хотим получить оценки:

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

Регрессионный анализ

К каким оценкам мы стремимся?

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

Регрессионный анализ: требования к ошибкам

Когда МНК (OLS) 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.

Регрессионный анализ: требования к ошибкам

Когда МНК (OLS) 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

- $E(e_i) = 0$
- $Var(e_i|x_i) = const$ гомоскедастичность
- $Cov(e_i, e_j) = 0$ 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?

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?

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
- латентные переменные

Почему предикторы и ошибки могут быть зависимыми (2)

Post-treatment bias

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

Почему предикторы и ошибки могут быть зависимыми (2)

Post-treatment bias

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

Вопрос для обсуждения

- Невозможность включения неизменяющихся во времени предикторов в модель с фиксированными эффектами часто незаслуженно интерпретирует как ограничение. Почему на самом деле это не является проблемой?

Почему предикторы и ошибки могут быть зависимыми (3)

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: $z_i = x_i + v_i$

Почему предикторы и ошибки могут быть зависимыми (4)

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 (4)

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. К примеру, одна из ключевых переменных – доход респондента. Ходили по домам, опросили преимущественно домохозяек.
- самоотбор. Индивиды выбирают определенное состояние: к примеру, посещать занятия или нет, ходить на выборы или нет и т.д. В результате часть недоступна для анализа.

Почему предикторы и ошибки могут быть зависимыми (5)

Что на что влияет?

Неоднозначность направления причинно-следственной связи предикторов и отклика. Подробно о возможности делать каузальный вывод и инструментах для выявления treatment effect – см. во второй части презентации.

ЧАСТЬ 2. Введение в каузальный анализ. Планы:

- поймем, в чем разница между статистической инференцией VS каузальным выводом
- начнем знакомиться с базовыми понятиями causal inference
- обозначим основную проблему каузального вывода

«Классика» статистической инференции (1)

Какова задача статистической инференции?

«Классика» статистической инференции (1)

Какова задача статистической инференции?

Перенести выводы с выборки на более широкую совокупность.

«Классика» статистической инференции (1)

Какова задача статистической инференции?

Перенести выводы с выборки на более широкую совокупность.

Что мы хотим узнать?

«Классика» статистической инференции (1)

Какова задача статистической инференции?

Перенести выводы с выборки на более широкую совокупность.

Что мы хотим узнать?

- Оценивание параметров. Какой средний уровень математической тревожности среди студентов-политологов?

«Классика» статистической инференции (1)

Какова задача статистической инференции?

Перенести выводы с выборки на более широкую совокупность.

Что мы хотим узнать?

- Оценивание параметров. Какой средний уровень математической тревожности среди студентов-политологов?
- Проверка статистических гипотез. Сравнение. У студентов-магистров выше средний уровень математической тревожности, чем у студентов бакалавриата?

«Классика» статистической инференции (1)

Какова задача статистической инференции?

Перенести выводы с выборки на более широкую совокупность.

Что мы хотим узнать?

- Оценивание параметров. Какой средний уровень математической тревожности среди студентов-политологов?
- Проверка статистических гипотез. Сравнение. У студентов-магистров выше средний уровень математической тревожности, чем у студентов бакалавриата?
- Проверка статистических гипотез об отсутствии взаимосвязи. Связаны ли математическая тревожность и посещение курсов по эконометрике? Если связаны, то

«Классика» статистической инференции (2)

На какие выводы мы можем претендовать?

«Классика» статистической инференции (2)

На какие выводы мы можем претендовать?

- в терминах описания

«Классика» статистической инференции (2)

На какие выводы мы можем претендовать?

- в терминах описания
- на уровне сравнения

«Классика» статистической инференции (2)

На какие выводы мы можем претендовать?

- в терминах описания
- на уровне сравнения
- в терминах совместной изменчивости

«Классика» статистической инференции (2)

На какие выводы мы можем претендовать?

- в терминах описания
- на уровне сравнения
- в терминах совместной изменчивости
  Примечание: Чем больше выборка, тем лучше для свойств оценок.

Известный пример

В солнечную и жаркую погоду растет объем продажи мороженого и увеличивается количество нападений акул на людей. Значит ли это, что продажи мороженого положительно влияют на количество нападений акул?

Каузальный вывод

Логика исследования

1. идентификация причинно-следственного эффекта. Есть ли такая возможность?
2. возвращаемся к статистической инференции как вспомогательному инструментарию
3. каковы свойства полученных оценок?
   Примечание: размер выборки на первом этапе далеко не решающий фактор!

Каузальный вывод

Логика исследования

1. идентификация причинно-следственного эффекта. Есть ли такая возможность?
2. возвращаемся к статистической инференции как вспомогательному инструментарию
3. каковы свойства полученных оценок?
   Примечание: размер выборки на первом этапе далеко не решающий фактор!

Что мы хотим узнать?

Выявить, что на что влияет. Какие факторы ВЛИЯЮТ на математическую тревожность? Претендуем на определение направления причинно-следственной связи.

Почему не все так просто, как хотелось бы?

Сложности идентификации каузального эффекта

- отложенный во времени эффект

Почему не все так просто, как хотелось бы?

Сложности идентификации каузального эффекта

- отложенный во времени эффект
- зависимая и независимая переменные – близкие концепты

Почему не все так просто, как хотелось бы?

Сложности идентификации каузального эффекта

- отложенный во времени эффект
- зависимая и независимая переменные – близкие концепты
- множественные причины

Почему не все так просто, как хотелось бы?

Сложности идентификации каузального эффекта

- отложенный во времени эффект
- зависимая и независимая переменные – близкие концепты
- множественные причины
- spillover effect

Почему не все так просто, как хотелось бы?

Сложности идентификации каузального эффекта

- отложенный во времени эффект
- зависимая и независимая переменные – близкие концепты
- множественные причины
- spillover effect
- неоднородность каузального эффекта

Основные понятия causal inference (1)

Объект изучения

- Treatment (intervention) variable – переменная воздействия. Та переменная, которая оказывает эффект.
- Outcome variable – зависимая переменная (предполагаемый отклик в модели). На эту переменную оказывается воздействие.
- Treatment (intervention) effect

Основные понятия causal inference (2)

Treatment effect и его разновидности

- $TE = Y(T = 1, Z) - Y(T = 0, Z)$ В общем виде, каузальный эффект – разница в значении зависимой переменной при разных значениях treatment variable и при прочих равных условиях.
- Средний эффект воздействия.
  $ATE = E(Y|T = 1, Z) - E(Y|T = 0, Z)$
- Условный средний эффект воздействия (ATE для подвыборки наблюдений, к примеру, эффект только для выборки мужчин):
  $CATE = E(Y|T = 1, G = g1, Z) - E(Y|T = 0, G = g1, Z)$

Основная проблема каузального вывода

Что нам нужно для идентификации каузального эффекта?

- Изменчивость treatment variable + outcome variable
- Нам нужно иметь значения зависимой переменной сразу в нескольких состояниях: при условии того, что есть эффект воздействия, и нет эффекта воздействия
- Дело осложняется тем, что мы еще должны сохранять условие «при прочих равных»

Основная проблема каузального вывода

Проблема пропущенных данных

- Можем ли мы, к примеру, одновременно наблюдать одного кандидата, как выигравшего и проигравшего выборы? Или одного и то же человека, как одновременно имеющего работу и безработного?
- Значения outcome variable при условии гипотетических состояний treatment variable называют counterfactual outcomes
- Проблема каузального вывода как проблема пропущенных данных


ВАШИ ВОПРОСЫ?