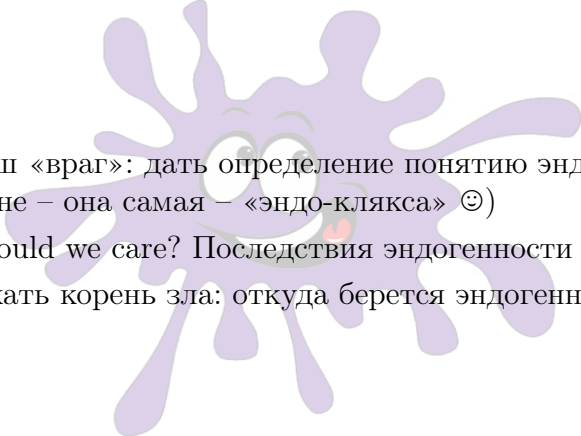


# Введение в многомерный статистический анализ

Эндогенность в регрессионной модели и ее источники

17 февраля 2023

# Планы на сегодня:

- 
- 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?

# Эндогенность: определение

## формальное определение

Эндогенность – это случай нарушения условия экзогенности (нескоррелированности объясняющих переменных и ошибок в регрессионной модели).

# Эндогенность: определение

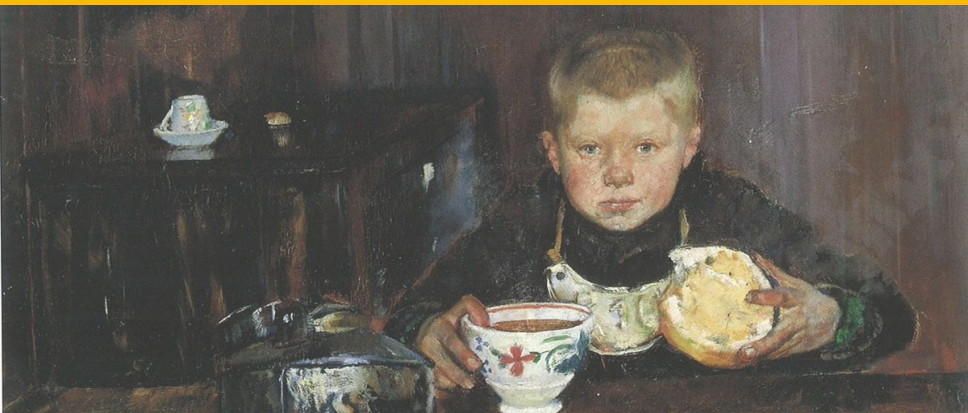
## формальное определение

Эндогенность – это случай нарушения условия экзогенности (нескоррелированности объясняющих переменных и ошибок в регрессионной модели).

## ЧТО ЗА ЭТИМ СТОИТ

В широком смысле эндогенность – проблема пропущенных существенных переменных.

# 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 скоррелированными (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 скоррелированными (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 скоррелированными (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 скоррелированными (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 скоррелированными (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 скоррелированными (4)

## Что на что влияет? Simultaneity problem

Неоднозначность направления причинно-следственной связи предикторов и отклика. Подробно о возможности делать каузальный вывод и способах выявления treatment effect – в следующих лекциях.

# Почему предикторы и ошибки могут быть скоррелированными (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  $x_i$ :  $z_i = x_i + v_i$

