Виды анализа и R Markdown

Холмовский Алексей

2025-10-20

Ссылка на репозиторий проекта: https://github.com/SHAITAN228/Rstudio.git

Введение

В данной работе рассматривается 2 вида анализа данных:

1. **Разведочный анализ данных** (Exploratory Data Analysis).
2. **Причинно-следственный анализ данных** (Causal Analysis).

Постановка задачи

Необходимо представить научные статьи, в которых используется указанные виды анализа данных, а также обосновать их применение в исследованиях.

Научные статьи

Разведочный анализ данных:

Особенности:

- Изучает как могут быть связаны различные переменные.
- Полезен для обнаружения новых связей.
- Помогает сформулировать гипотезы и управлять планированием будущих исследований и сбора данных.

Наименование статьи:

Методический подход к оценке Влияния бюджетных расходов на среднюю ожидаемую продолжительность жизни населения (на примере г. Екатеринбурга).

Основная цель статьи:

Статья посвящена исследованию зависимости продолжительности жизни населения от бюджетных расходов на инфраструктуру. В статье используются регрессионные модели для количественной оценки влияния государственных инвестиций в критическую инфраструктуру на показатель СОПЖ.

Обоснование:

В статье присутствуют признаки анализа данных, одним из которых является первичное исследование исходных данных и визуализация для выявления потенциальных взаимосвязей признаков.

На представленном ниже графике отображена динамика расходов бюджета Екатеринбурга на развитие критической инфраструктуры за период 2001-2011 годов, что позволяет выявить тенденции и закономерности в распределении финансовых средств.

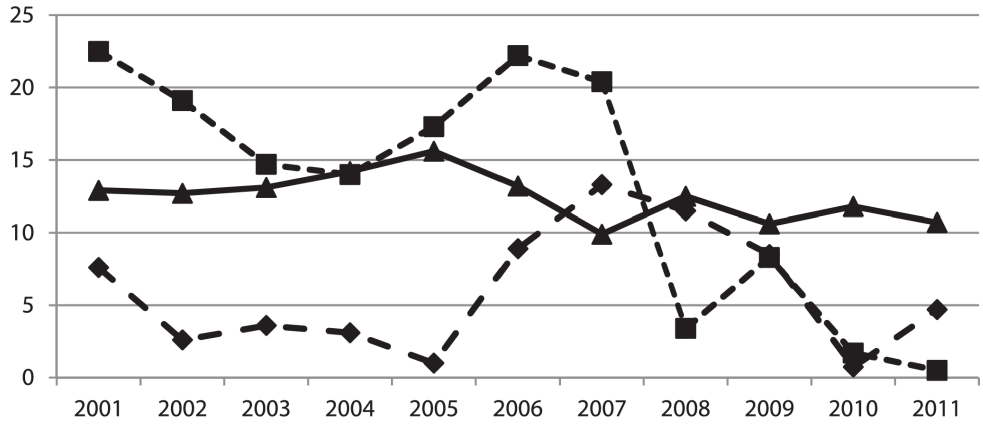

Рис. 1: Динамика бюджетных расходов на критическую инфраструктуру г. Екатеринбурга (2001-2011 гг.)

В таблице ниже представлены данные о средней ожидаемой продолжительности жизни (СОПЖ) - целевой переменной в исследовании. Данная статистика за одиннадцатилетний период позволяет оценить динамику и выявить общие тенденции.

Таблица 1: Динамика средней ожидаемой продолжительности жизни (СОПЖ) по категориям населения, лет

Категория населения	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Оба пола (СОПЖ)	66.1	65.0	65.6	65.9	65.95	67.2	68.6	68.7	69.40	70.3	70.50
Мужчины (СОПЖм)	59.0	58.8	58.6	58.9	59.00	61.4	61.9	62.0	63.10	64.0	64.30
Женщины (СОПЖж)	72.4	72.3	72.5	72.9	73.60	73.9	74.4	74.7	74.85	75.0	75.45

Далее в статье применяется корреляционный анализ, что является характерным признаком разведочного анализа данных. Составленная матрица парных корреляций представлена ниже.

Коэффициенты парной корреляции

Показатель	Временной лаг						
	0 лет	1 год	2 года	3 года	4 года	5 лет	6 лет
Транспорт							
СОПЖ	0,29	0,33	0,58	0,56	0,31	–0,17	–0,69
СОПЖм	0,25	0,28	0,57	0,59	0,32	0,00	–0,67
СОПЖж	0,25	0,28	0,50	0,44	0,42	0,02	–0,86
Коммунальное хозяйство							
СОПЖ	–0,55	–0,64	–0,52	–0,43	–0,14	–0,22	–0,71
СОПЖм	–0,52	–0,65	–0,58	–0,46	–0,05	–0,02	–0,73
СОПЖж	–0,46	–0,64	–0,53	–0,51	–0,10	–0,17	–0,59
Здравоохранение							
СОПЖ	–0,72	–0,52	–0,38	–0,20	–0,13	0,56	0,91
СОПЖм	–0,72	–0,46	–0,37	–0,26	–0,19	0,63	0,89
СОПЖж	–0,58	–0,43	–0,35	–0,12	–0,22	0,42	0,92

Рис. 2: Матрица парных корреляций

Авторы отмечают:

> “Корреляционный анализ показывает наличие некоторой связи между бюджетными расходами на содержание объектов критических инфраструктур и СОПЖ. Однако следует отметить, что пока-

затель средней ожидаемой продолжительности жизни является очень инертным, поэтому анализ дополнен поиском корреляции с учетом временного лага.”

Причинно-следственный анализ:

Особенности:

- Эталонный метод в анализе данных.
- Часто применяется к результатам случайных исследования, которые были разработаны для выявления причинно-следственной связи.
- Обычно анализируются совокупности, а наблюдаемые взаимосвязи обычно являются средним эффектами.

Наименование статьи:

Как уровень заболеваемости влияет на показатель склонности к совершению преступлений в регионах РФ?

DOI:

https://doi.org/10.24866/2311-2271/2023-1/32-46

Основная цель статьи:

Статья посвящена исследованию причинно-следственной связи между уровнем заболеваемости и преступностью в регионах России. В работе используется метод инструментальных переменных для количественной оценки влияния здоровья населения на уровень преступности с учетом проблем эндогенности.

Обоснование:

В статье для оценки влияния уровня здравоохранения на уровень преступности используется несколько различных моделей:

1. Линейная регрессионная модель на основе пространственной вы борки (pooled regression).

$$\log(\text{criminalp})_i = \beta_0 + \beta_1 \cdot \log(\text{ill})_i + \beta_2 \cdot \text{abort}_i + \beta_3 \cdot \text{vodkat}_i + \cdots + \beta_{13} \cdot (gdp\_p)_i + \varepsilon_i,$$

   где β_i — коэффициенты регрессии, i — регион.
2. Модель панельных данных с фиксированными эффектами.

$$\log(\text{criminalp})_{it} = \alpha_i + \beta_1 \cdot \log(\text{ill})_{it} + \beta_2 \cdot \text{abort}_{it} + \beta_3 \cdot \text{vodkat}_{it} + \cdots + \beta_{13} \cdot (gdp\_p)_{it} + \varepsilon_{it},$$

   где α_i - выражает индивидуальный эффект объекта i, не зависящий от времени t, i - регион, t - время.
3. Модель панельных данных с фиксированными эффектами с од новременным применением метода инструментальной переменной (Ко личество онкологов в соседних регионах, Количество онкологов в со седних регионах).
4. Модель панельных данных с фиксированными эффектами с од новременным применением метода инструментальной переменной (От ношение баллов ЕГЭ “платников” к баллам бюджетников по УГН “Здра воохранение”).
5. Модель панельных данных с фиксированными эффектами с од новременным применением метода инструментальной переменной (Ко личество онкологов в соседних регионах, Отношение баллов ЕГЭ “плат ников” к баллам бюджетников по УГН “Здравоохранение”).

Как можно увидеть, в исследовании применяется метод инструментальной переменной, характерный для причинно-следственного анализа. Этот метод позволяет устранить проблему обратной причинности между заболеваемостью и преступностью, а также учесть влияние пропущенных переменных. В результате авторы получают более надежную оценку воздействия уровня заболеваемости на преступность.

В результате применения методов были получены оценки представленных моделей, которые частично приведены в таблице ниже:

Таблица 2: Результаты оцененных моделей

VARIABLES	OLS	FE	FE_iv_1	FE_iv_2	FE_iv_3
Уровень заболеваемости	0.171***(0.046)	-0.012(0.048)	4.152(2.602)	0.573(0.526)	1.260***(0.413)
Аборты	0.013***(0.001)	0.002**(0.001)	0.004(0.003)	0.002(0.001)	0.003(0.002)
Алкоголь	0.093***(0.027)	0.007(0.015)	0.035(0.056)	0.018(0.021)	0.033(0.025)

Сравнительный анализ результатов показал, что наиболее надежной моделью является FE_iv_3 (модель панельных данных с фиксированными эффектами с применением метода инструментальных переменных). В данной модели использованы две инструментальные переменные: количество онкологов в соседних регионах и отношение баллов ЕГЭ “платников” к баллам бюджетников по направлению “Здравоохранение”. Именно эта модель демонстрирует статистически значимое влияние уровня заболеваемости на преступность (коэффициент 1.260***), что подтверждает наличие причинно-следственной связи между изучаемыми переменными.

> “Результаты, полученные при использовании данного метода с учётом рассмотрения региональных фиксированных эффектов, показывают, что увеличение уровня заболеваемости на 1% приводит в среднем в регионах РФ к увеличению уровня преступности на 1,26%.”

Этот результат демонстрирует, что исследование использует причинно-следственный анализ. Авторы не только обнаруживают связь между заболеваемостью и преступностью, но и доказывают, что именно заболеваемость влияет на уровень преступности.