МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ

(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)

Институт №8 «Информационные технологии и прикладная математика»
Кафедра 804 «Теория вероятностей и компьютерное моделирование»

Реферат по курсу
“Эконометрика”

“Метод разности разностей”

Выполнил: Дюсекеев А.Е.

Группа: 8О-404Б

Москва, 2020

Разность разностей – это статистический метод, используемый в эконометрике. Является полезным инструментом в анализе данных.

Механизм работы данного метода состоит в следующем: «В простейшей постановке наблюдаются некоторые исходы для двух групп и двух временных периодов. Одна из групп подвержена воздействию, или участвует в некоторой программе, во втором периоде, но не в первом. Вторая группа не подвержена воздействию ни в одном периоде. В случае, когда одни и те же объекты внутри групп наблюдаются в каждом периоде, среднее изменение исхода во второй (контрольной) группе вычитается из среднего изменения исхода в первой (опытной) группе. Это устраняет смещение при сравнении во времени, которое может быть следствием постоянных различий между этими группами, а также смещение при сравнении во времени, которое может быть вызвано временными трендами, никак не связанными с программой»

В постановке наблюдаются некоторые исходы для двух групп и двух временных периодов. Одна из групп подвержена воздействию, или участвует в некоторой программе, в одном из периодов, а вторая группа не подвержена воздействию ни в одной из периодов.

В случае, когда одни и те же объекты внутри групп наблюдаются в каждом периоде, среднее изменение исхода во второй(контрольной) группе вычитается из среднего изменения исхода в первой(опытной) группе. Это устраняет смещение при сравнении исходов в опытной и контрольной группах только во втором периоде, которое может быть следствием постоянных различий между группами, а также смещение при сравнении во времени, которое может быть вызвано временными трендами. Никак не связанными с программой.

Математическая постановка метода

При наличии повторяющихся выборок за два периода времени модель, тестируемая с помощью данного метода, записывается следующим образом:

$$y = \beta_0 + \beta_0 dB + \delta_0 d2 + \delta_1 d2 \cdot dB + u \qquad (1)$$

Где y – представляющий интерес исход,

d2 – фиктивная переменная для второго периода,

dB – фиктивная переменная для опытной группы.

Фиктивная переменная d2 улавливает факторы, которые бы вызвали изменения у даже при отсутствии воздействия или программы. Фиктивная переменная dB улавливает возможные различия между опытной и контрольной группами соответсвенно. Представляющий собой интерес коэффициент δ_1 находится при переменной взаимодействия d2dB, которая совпадает с фиктивной переменной, равной единице для наблюдений в опытной группе во втором периоде. Оценка δ_1 находится методом “разность разностей” (РР-оценка) – это обычная МНК-оценка для уравнения (1) на основе случайной выборке по исследуемым группам. Её можно записать в виде:

$$\hat{\delta}_1 = (\overline{y}_{B,2} - \overline{y}_{B,1}) - (\overline{y}_{A,2} - \overline{y}_{A,1})$$

виде: где A обозначает контрольную группу

Более подробно рассмотрим данный метод на следующем небольшом примере его применения. Предположим, например, что один из штатов США осуществляет программу в области здравоохранения для пожилых людей, скажем, в возрасте 65 лет и старше, и зависимая переменная y - это некий показатель здоровья. Одна из возможностей состоит в том, чтобы использовать данные только по жителям штата, в котором реализуется программа, как до, так и после ее внедрения, и в качестве контрольной группы взять жителей в возрасте до 65 лет (или в возрасте от 55 до 64 лет), а в качестве опытной группы - жителей в возрасте 65 лет и старше. Потенциальная проблема такого PP-анализа состоит в том, что другие факторы, не связанные с новой программой штата, могут повлиять на уровень здоровья пожилых людей по сравнению с молодыми, например изменения политики в области здравоохранения на федеральном уровне. Иная стратегия PP-анализа заключается в использовании другого штата для формирования контрольной группы, то есть в рассмотрении пожилых жителей штата, в котором программа отсутствует, в качестве контрольной группы.

Если обозначить два временных периода за 1 и 2, штат, в котором проводится программа за B, а группу пожилого населения за E, то расширенная версия уравнения (1) принимает вид:

$$y = \beta_0 + \beta_1 dB + \beta_2 dE + \beta_3 dB \cdot dE + \delta_0 d2 + \delta_1 d2 \cdot dB + \delta_2 \cdot d2 \cdot dE + \delta_3 d2 \cdot dB \cdot dE + u \qquad (2)$$

Теперь интерес представляет коэффициент δ_3.

МНК-оценку δ_3 можно записать в виде:

$$\delta_3 = \left[\left(\overline{y}_{B,E,2} - \overline{y}_{B,E,1}\right) - \left(\overline{y}_{B,N,2} - \overline{y}_{B,N,1}\right)\right] - \left[\left(\overline{y}_{A,E,2} - \overline{y}_{A,E,1}\right) - \left(\overline{y}_{A,N,2} - \overline{y}_{A,N,1}\right)\right] \qquad (3)$$

Оценку (3) называют PPP-оценкой

$\left[\left(\overline{y}_{B,E,2} - \overline{y}_{B,E,1}\right) - \left(\overline{y}_{B,N,2} - \overline{y}_{B,N,1}\right)\right]$ называется PP-оценкой полученная для штата B(в качестве контрольной группы использовалось молодое население)

$\left[\left(\overline{y}_{A,E,2} - \overline{y}_{A,E,1}\right) - \left(\overline{y}_{A,N,2} - \overline{y}_{A,N,1}\right)\right]$ называется PP-оценкой полученная для штата A(в качестве контрольной группы использовалось молодое население)

Если различия между пожилым и молодым населением в штате A отсутствуют то PPP-оценка будет близка к PP-оценке полученной только для штата B.

Изложенный стандартный выше подход предполагает, что вся вариация при инференции происходит из-за выборочных ошибок при оценивании средних для каждой комбинации группы и периода. Этот подход эквивалентен дисперсионному анализу (ANOVA). Сначала в соответствия с временными периодами и статусом подверженности воздействию определяются различные подгруппы. Затем для каждой подгруппы берутся случайные выборки и подсчитываются соответствующие разности выборочных средних.

Далее рассмотрим другие подходы, основанные на других источниках вариации. Одним из таких является подход предложенный в Donald & Lang (DL).

Подход Donald & Lang.

Постановка DL применима к общему случаю PP-анализа, когда число групп(контрольных и опытных групп) достаточно мало, и для каждой из них доступны довольно большие выборки.

Рассмотрим случай с единственным регрессором, меняющимся только по группам:

$$y_{gm} = \alpha + \beta_{x_g} + c_g + u_{gm} \quad (4)$$

$$= \delta_g + \beta_{x_g} + u_{gm}, m = 1, \dots, M_g; g = 1, \dots, G \quad (5)$$

В простейшем случае x_g- единственный индикатор программы. Ключевой характеристикой уравнения (4) является наличие переменной c_g, эффекта группы или кластера. Иначе модель можно записать в виде (5) с общим коэффициентом наклона β, но зависящей от группы константой δ_g. DL рассматривают модель в виде (4) предполагая, что c_g не зависит от x_g и имеет нулевое среднее.

Один из способов увидеть проблему при применении стандартной инференции – заметить что при $M_g = M$ для всех $g = 1, \dots, G$ обычная МНК-оценка $\hat{\beta}$ совпадает с межгрупповой оценкой, полученной из регрессии

y_gна 1, $x_g, g = 1, \dots, G$. Наличие c_g означает, что новые наблюдения внутри группы не дают дополнительной информации для оценивания β. По сути имеются только G полезных носителей информации.

Если добавить несколько сильных предположений то у данной проблемы существует решение, но только при G>2. Вдобавок к предпосылке о том, что $M_g = M$ для всех g, предположим, что $c_y|x_g \sim N(0, \sigma_c^2)$ и $u_{gm}|x_g c_g \sim N(0, \sigma_c^2)$. Тогда $\bar{v}_g$ не зависит от x_g и $\bar{v}_g \sim N(0, \sigma_c^2 + \sigma_u^2/M)$ для всех g. Поскольку предполагается независимость по g, уравнение

$$\bar{y}_g = \alpha + \beta x_y + \bar{v}_g, g = 1, \dots, G \quad (6)$$

удовлетворяет предположениям классической линейной модели.

Модель DL находит наиболее широкую сферу применения, когда среднее значение ошибок, $\bar{v}_g$, можно игнорировать, то есть либо σ_u^2 мала по сравнению с σ_c^2, либо M_g велики, либо и то и другое. На самом деле применение подхода DL с различными размерами групп или негауссовыми u_{gm} равносильно игнорированию ошибки оценивания выборочных средних $\bar{y}_g$.

Если $1 \times L$ вектор z_{gm}, меняющийся внутри группы, присутствует в исходной модели, как это имеет место во многих PP-приложениях, можно использовать усреднённое уравнение

$$\bar{y}_g = \alpha + x_y\beta + z_y\gamma + \bar{v}_g, g = 1, ..., \mathrm{G} \quad (7)$$

Уравнение При условии что G > K + L + 1. Если c_g не зависит от $(x_g, \bar{z}_g)$, имеет гомоскедастичное нормальное распределение, и размеры групп велики, инференция может быть реализована на основе $t_{G-K-L-1}$ – распределения. Это довольно стандартный способ проверки устойчивости результатов, полученных по дезагрегированным данным, но часто он осуществляется для несколько больших значений G(G = 50).

Не очевидно следует ли использовать подход DL, даже при достаточно большом числе степеней свободы. Допустим, например, что есть G = 4 группы, две из которых контрольные($x_1 - x_2 = 0$) и две – опытные ($x_3 - x_4 = 1$). Подход DL предполагает подсчёт средних для каждой группы $\bar{y}_g$, и оценивание регрессии $\bar{y}_g$ на 1, $x_g, g = 1, \dots 4$.

Инференция основана на t_2- распределении. Оценку β в данном случае можно записать в виде:

$$\hat{\beta} = \frac{\bar{y}_3 + \bar{y}_4}{2} - \frac{\bar{y}_1 + \bar{y}_2}{2} \quad (8)$$

Если определить объект интереса как:

$$\tau = \hat{\beta} = \frac{\mu_3+\mu_4}{2} - \frac{\mu_1+\mu_2}{2} \quad (9)$$

то есть своего рода средний эффект воздействия, то оценка $\hat{\beta}$ состоятельна для τ и(при должной нормировке) асимптотически нормально по мере роста M_g

При больших размерах групп даже при небольших G можно сформулировать задачу в рамках модели минимального расстояния.

Для каждой группы g запишем:

$$y_{gm} = \delta_g + z_{gm}\gamma_g + u_{gm}, m = 1, \dots, M_g \quad (10)$$

предполагая случайность выборок внутри групп и независимость выборок по группам.

Наличие переменных x_g группового уровня в "структурной" модели можно рассматривать как наложение ограничений на константы δ_g в отдельных моделях для групп в (10). В частности:

$$\delta_g = \alpha + x_g\beta, g = 1, \dots, G \quad (11)$$

где x_g – фиксированные наблюдаемые характеристики разнородных групп.

Если в качестве взвешивающей матрицы I_g взять единичную матрицу $G \times G$, то оценку минимального расстояния можно подсчитать из МНК-регрессии

$$\delta_g \text{ на } 1, x_g, g = 1, \dots, G \quad (12)$$

При асимптотике с $M_g = \rho_g M$, где $0 < \rho_g \leq 1$ и $M \to \infty$, оценка минимального расстояния θ

состоятельна и $\sqrt{M}$ – асимптотически нормальна.