Multilevel linear models и свитчбэк-тестирование

Структура документа

Когда мы используем MLM

MLM используется для анализа иерархических данных, когда наблюдения на нижнем уровне вложены в верхние уровни. Например, наблюдения по школьникам вложены в классы, сотрудникам – в фирмы, регионы – в страны и т.д. Пример из АБ тестирования: при свитчбэк-тестировании юниты верхнего уровня – юниты рандомизации, географические юниты - временной отрезок. Наблюдения нижнего уровня – метрики на уровне заказов, магазино-часов, магазинов.

В общем виде условия следующие:

- В наших данных есть вложенность
- Мы хотим генерализировать наши выводы на группы, схожие с присутствующими в данных
- Мы хотим объяснить различия во взаимосвязи таргета и предиктора между подгруппами
- Наблюдения внутри юнита скоррелированы (обычно для этого считается intra-class correlation coefficient, но мы будем считать, что это соблюдается)

Корреляция может появляться из-за повторяемых наблюдений или вложенности данных. Наблюдения по магазинам, магазино-часам или покупкам внутри юнита похожи друг на друга. Оценка линейной регрессии в случае скоррелированных данных приведет к некорректным стандартным ошибкам. MLM снимает допущение о независимости наблюдений, предполагая, что наблюдения зависимы друг от друга из-за принадлежности к группе.

Основные допущения

Отсюда можно легко понять допущения использования модели, схожие с допущениями линейной регрессии за исключением допущения о независимости наблюдений:

- Линейность взаимосвязи тагрета и зависимой переменной
- Нормальность распределения остатков
- Гомоскедастичность остатков
- Отсутствие мультиколлинеарности, т.е. слишком сильной корреляции между независимыми переменными

Когда это используется в АБ?

Условие валидности АБ теста – stable under treatment assumption. **SUTVA** говорит, что воздействие, применяемое к тестовым юнитам, не должно воздействовать на юниты из других групп. Когда это предположение нарушается? Например, если мы корректируем объем спроса поднятием минимального заказа и рандомизируем по магазинам, пользователь может решить сделать заказ в магазине из контрольной группы, что внесет сдвиг в результаты теста.

Для смягчения этого эффекта можно использовать свитчбэк-тестирование. Метод предполагает рандомизацию не по магазинам, а по географическому региону и временному промежутку. Такая рандомизация помещает все доставки внутри одного юнита рандомизации, а следовательно для всех них либо присутствует, либо отсутствует тестируемая механика, что снижает сетевой эффект. Все наблюдения (заказы) распределяются между тестом и контролем в соответствии с их временем и принадлежностью к региону. В результате мы получаем вложенную структуру данных: внутри одного юнита есть несколько заказов, относящихся к тесту и контролю в зависимости от юнита.

Можно выбирать различные юниты для разбиения, наиболее универсальным является город.

Хорошо, но причем здесь MLM? Почему мы не можем использовать стандартные методы оценки эффекта?

Если мы будем оценивать эффект стандартным оценщиком с агрегацией до свитчбэк-юнита:

- ограничиваем количество наблюдений и как следствие мощность;

- не учитываем, что количество наблюдений в юните разное и учитываем их с одинаковыми весами в оценке. Так, для юнитов Москва 09-00 и Москва 06-09 разное количество наблюдений (метрик на уровне заказов, магазина в час) и мы хотим это учитывать в оценке, чтобы оценка на уровне наблюдения в юните и юнита не сильно различалась.

Почему просто не включить дамми-переменную на свитчбэк-юнит?

- сложно оценить общий тренд;
- теряем возможность генерализации результатов на другие юниты;
- группы могут быть несбалансированы по объему наблюдений, overfit.

В итоге, MLM – хороший вариант для оценки эффекта в дизайне со свитчбэками:

- наши данные естественно вложены (заказы внутри юнитов);
- метрики внутри групп коррелируют, нарушается стандартное предположение о независимости наблюдений;
- можем ожидать повышение мощности, т.к. MLM использует группы и наблюдения на низшем уровне в оценке эффекта и его дисперсии.

Спецификация

$$Y_{ij} = \beta_{0j} + \beta_{1j} * X_{ij} + r_{ij} (1)$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01} * W_j + u_{0j} (2)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} * W_j + u_{1j} (3)$$

(1) относится к индивидуальному уровню, i; (2) и (3) к уровню юнита, j. В (1) Y_ij – таргет, X_ij – предиктор на индивидуальном уровне, B_0j и B_1j – интерсепт и наклон, которые могут изменяться в зависимости от групп, r_ij – ошибки на индивидуальном уровне;

(2) моделирует изменчивость интерсепта B_0j в зависимости от юнита через предиктор на уровне юнита W_j (т.е. здесь мы моделируем разные "стартовые условия" для юнитов). Здесь Gamma_00 интерпретируется как ожидаемое значение B_0 при Wj = 0, u_ij показывает оставшуюся изменчивость после контроля W_j;

(3) моделирует изменчивость наклона B_1j в зависимости от юнита через предиктор на уровне юнита W_j (т.е. мы дополнительно моделируем разную взаимосвязь между таргетом и предиктором на уровне наблюдений внутри юнитов).

Все вместе это можно записать в следующем виде:

$$Y_{ij} = \gamma_{00} + \gamma_{01} * W_j + \gamma_{10} * X_{ij} + \gamma_{11} * X_{ij} * W_j + u_{0j} + u_{1j} * X_{ij} + r_{ij} (4).$$

Мы можем моделировать отдельно как разные стартовые условия и одинаковую взаимосвязь по группам (random intercept model), так и разные стартовые условия и разную взаимосвязь по группам (random intercept and slope model). Если мы моделируем изменчивость наклона по группам, внутри групп предиктор должен изменяться, то есть нам нужно подобрать такие группы, для которых юнит относится или к тесту или к контролю. По опыту, именно моделирование наклона делает MLM таким мощным инструментом. Вот здесь хорошая визуализация.

Random intercept model:

blocked URL

Random intercept and slope model:

blocked URL

Часто эту модель называют моделью со смешанными эффектами. Фиксированные эффекты являются общими для всей выборки и показывают общую взаимосвязь, в спецификации (4) это $\gamma_{00} + \gamma_{01} * W_j + \gamma_{10} * X_{ij} + \gamma_{11} * Xij * Wj$, где γ_{00} – интерсепт, γ_{10} и γ_{01} -- наклоны для предикторов на индивидуальном и групповом уровнях, γ_{11} -- взаимодействие между группами. Случайные эффекты отражают дополнительную изменчивость фиксированных эффектов в зависимости от принадлежности к группе: $u_{0j} + u_{1j} * X_{ij} + r_{ij}$.То есть мы моделируем общий тренд с поправкой на индивидуальную изменчивость, при этом не теряя в мощности как в случае построения отдельных регрессий по группам.

Допустим, наши данные представлены двумя магазинами и двумя днями, метрики агрегируем по магазину в день. В первый день оба магазины были в группе контроля, во второй – в тритменте. Мы хотим оценить эффект воздействия, моделируя разную взаимосвязь и наклон по магазинам.

blocked URL

Что на этом графике?

- B1 и B2 – это средние в группе контроля и теста соответственно (фиксированные эффекты)
- u1 и u2 – ошибки на индивидуальном уровне (случайные эффекты)
- e11, e12, e21, e22 – residual error, то, что мы не можем моделировать

Воздействие моделируется как фиксированный эффект, тк уровни в данных (контроль-тест) представляют все возможные варианты, в то время как магазины могут не отражать всю генеральную совокупность и поэтому моделируются как случайные эффекты (то есть с ними ассоциирована ошибка ui ~ N(0, σs²)). В анализе свитчбэков такая постановка логична, но важно понимать, что может быть по-другому 🙂

blocked URL

В (4) K моделируются связи между наблюдениями: связанные группы, та же группа и так далее.

(6) var(Ku)=K*var(u)*K^T and var() = σ^2*I and var(u)= σs^2*I, где σ^2 residual variance, σs^2 – random Intercept effects (shared across data points) variance.

В (7) моделируется близость между наблюдениями

(8) Variance-covariance matrix

blocked URL

Имя все это, вспомниаем про Maximum Likelihood или Restricted Maximum Likelihood.

Пример реализации

Реализация оценки эффекта можно посмотреть в дизайне теста суржа.

Random intercept model

```
    import statsmodels.formula.api as smf
md = smf.mixedlm("metric ~ C(group, Treatment('control'))",
                         data,
                         groups=data['unit_interest'],
                         re_formula="1").fit()
md.summary()
```

Что значит запись:

- group – принадлежность наблюдения к группе теста или контроля
- unit_interest – принадлежность наблюдения к группе, для которых полагаем разные стартовые условия/углы наклоны
- Формула записывается как "метрика ~ зависимая переменная 1 +зависимая переменная 2 + ... + зависимая переменная n "
- C(group, Treatment('control')) – обозначаем, что переменная категориальная с базовой категорией "принадлежность к контрольной группе". Это нужно, чтобы интерпретировать коэффициент при группе как изменение таргета при переходе от группы контроля к группе теста.
- re_formula – указываем переменные, по которым будет варьироваться угол наклона

В summary можно увидеть значение Group Var – это отражает дисперсию стартовых условий по группам.

Random slope model

```
    md = smf.mixedlm("metric ~ C(group, Treatment('control'))",
                 data,
                 groups=data['unit_interest'],
                 re_formula = "~ C(group, Treatment('control'))").fit()
md.summary()
```

Что значит запись:

- аналогично random intercept model
- re_formula – указываем переменные, по которым будет варьироваться угол наклона

В summary появилось две новых переменных:

- ses Var – дисперсия углов наклона
- Group x ses Cov – ковариация между случайным углом наклона и интерспетом, то есть какая связь между эффектом и стартовыми условиями

Практические рекомендации

- Нужно большое количество групп (>50), для MLM это важно для оценки всех параметров
- MLM чувствителен к размеру кластеров: при кол-ве наблюдений < 5 RE компонент переоценен (корреляция внутри кластера переоценена, ненадежные RE оценки)
- MLM не чувствительна к не сбалансированным кластерам
- Нужно убедиться, что внутри группа есть изменчивость предиктора при моделировании наклона

Вложенные эффекты

Случайные эффекты бывают разными: перекрывающими (не зависят друг от друга), вложенными (одни эффекты зависят от других). При анализе теста со свитчбэками может пригодиться моделирование вложенных эффектов.

Допустим, мы агрегируем метрики по заказам или по магазину в час. У нас есть подгруппы более высокого порядка (города) и им принадлежат подгруппы более низкого порядка (магазины). При этом одна и та же группа более низкого порядка не может принадлежать сразу нескольким высшим подгруппам.

blocked URL

Давайте это смоделируем!

```
    vc = {'store_id': '0 + C(store_id)'}
model_nested = smf.mixedlm('metric ~ C(group, Treatment('control')',
                                                data = data,
                                                groups = data['unit_interest'],
                                      vc_formula = vc,
                                                re_formula = '1').fit()

model_nested.summary()
```

Что здесь важно:

- vc_formula принимает на вход словарь вида {‘название_эффекта_в_модели_1’ : ‘0 + название_эффекта_в_данных_1’, ‘название_эффекта_в_модели_2’ : ‘0 + название_эффекта_в_данных_2’}, 0 указываем для расчета еще одного интерспета
- нужно указать аргумент **re_formula = '1',** чтобы строился случайный интерсепт, связанный со всей структурой

В summary появляется дополнительный параметр store_id Var, отражающий дисперсию интерспета по юнитам более низкого порядка.

Как это можно сделать по-другому (и почему мы так не делаем)

В статьях по свитчбэкам описывается еще две модели: CRSE-линейная регрессия и GEE. При этом doordash пишут, что они применяют CRSE-регрессию для новых экспериментов, о которых у нас мало априорных знаний. В наших дизайнах MLM показывала себя лучше этих моделей, но их точно нужно иметь в виду.

CRSE-регрессия

Подробнее можно почитать здесь.

GEE

В случае, если мы не хотим моделировать случайные эффекты (это наше теоретическое предположение о связи), можем использовать Generalized Estimating Equations.

- Это полупараметрический метод и более устойчив к нарушением допущений по сравнению с MLM
- Можно самостоятельно указывать внутригрупповую ковариационную структуру. Это упростит жизнь, если у нас уже есть теоретические предположения о характере связи наблюдений внутри групп.

Ковариационные структуры:

- **Неограниченная:** все наблюдения связаны друг с другом по-разному (в этом случае параметров очень много, не рекомендуется)
- **Независимая:** наблюдения не связаны друг с другом, несмотря на то, что они в одной группе
- **Взаимозаменямая:** все наблюдения связаны друг с другом одинаково
- и другие

```
    ind = sm.cov_struct.Independence() #
ex = sm.cov_struct.Exchangeable() #
nest = sm.cov_struct.Nested() #

gee_model = smf.gee('metric ~ C(group, Treatment('control')',
                                  data = classes,
                  groups = data['unit_interest'],
                  cov_struct = ind).fit()

gee_model.summary()
```

[DRAFT] Симуляции свитчбэк-тестов без MLM

TBA

Что почитать

Про MLM в контексте свитчбэков:

- https://doordash.engineering/2019/02/20/experiment-rigor-for-switchback-experiment-analysis/
- https://habr.com/ru/company/citymobil/blog/560426/
- https://habr.com/ru/company/citymobil/blog/575218/

Про MLM эконометрически:

- Сравнение с другими моделями: https://www.hse.ru/data/2020/06/29/1610354484/233EC2020.pdf
- Как оно все работает: https://towardsdatascience.com/how-linear-mixed-model-works-350950a82911

Визуализация свитчбэков на дашбордах:

- Мониторинг бизнес-метрик в процессе проведения switchback-тестирования