

## 1. Что этот пакет делает?

Этот пакет может посчитать измерения под влиянием конкретного события, такого как выпуск нового продукта. Каузальное воздействие события представляет собой разницу между наблюдаемой величиной ответа и (не наблюдаемой) величиной, которая была бы получена при отсутствии воздействия

### Какие есть проблемы при простом решении этой задачи?

Чтобы посчитать влияние события, которое в течение какого-то времени повлияло на измеряемую величину. Можно было бы просто сравнить показатели перед событием и после события, или сравнить показатели двух групп, например A/B тестирование. Но на самом деле трудно измерить такое воздействие в реальном мире, потому что есть много атрибутов, которые могут влиять на результат.

### Как примерно работает?

CausalImpact делается предположение, что существует три компоненты – априорное знание о параметрах модели, контрольные значения до изменения и значения таргета до изменения. CausalImpact строит байесовскую структурную модель временных рядов на основе нескольких сопоставимых контрольных групп, и поведение самого временного ряда до вмешательства, и использует модель для прогнозирования ряда базовых значений за период времени после события.

Он улучшает существующие методы в двух отношениях: он обеспечивает полностью байесовскую оценку временного ряда, и он использует усреднение модели для построения наиболее подходящего синтетического управления для моделирования контрафакта.

### Как именно работает?

Структурные модели временных рядов полезны на практике, потому что они гибкие и модульные. Очень большой класс моделей, включая все модели ARIMA, можно записать в форме состояния, заданной формулами:

$$y_t = Z_t^T \alpha_t + \varepsilon_t, \text{ -- уравнение наблюдения;}$$

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t, \text{ -- уравнение состояния}$$

где  $\varepsilon_t$  и  $\eta_t$  -- стандартный шум с дисперсией  $\sigma^2$  и  $Q_t$  соответственно,  $y_t$  -- скалярное наблюдение,  $Z_t$  -- выходной вектор,  $T_t$  - матрица перехода,  $R_t$  - матрица управления..

Общий подход заключается в том, чтобы предположить, что ошибки различных моделей состояний-компонентов независимы (т. Е.  $Q_t$  является блочно-диагональным).

Тогда вектор  $\alpha_t$  может быть образован путем конкатенации отдельных компонент состояния,  $T_t$  и  $R_t$  становятся блочно-диагональными матрицами.

Также в модель можно включать модели:

- Локальная линейная тренд.
- Сезонность.
- Современные ковариаты со статическими коэффициентами.

Естественным способом включения контрольных серий в модель является линейная регрессия. Его коэффициенты могут быть статическими или изменяющимися во времени. Используются когда соотношение между элементами управления и обработанной единицей до события было стабильным.

- Современные ковариации с динамическими коэффициентами. Альтернативой вышесказанному является компонент регрессии с коэффициентами динамической регрессии для учета изменяющихся во времени соотношений.

## 2. Сезонность: использовать стоит в тех случаях, когда в ряде-предикторе есть сезонная компонента.

Однако, если в таргет-ряде есть сезонная компонента, и логично было бы, поставить сезонность, но если в ряде-предикторе -- нет, то использование только ухудшает результат.

**Динамические коэффициенты** полезны, когда считается, что линейная зависимость между рядом-предиктором и элементами таргет-ряда изменяется со временем. Лучше не использовать когда соотношение между элементами таргет-ряда и рядом-предиктором до события было стабильным.

3. Можно рассматривать разные ряды-таргеты -- активности пользователей, соответственно с ними меняются и ряды-предикторы:
  - a. Во-первых можно считать количество запросов за день от пользователей, соответственно в качестве ряда-предиктора в таком случае можно рассматривать количество запросов у конкурентов: потому что снижение/повышение запросов вызванное выходными/праздниками будет примерно одинаково. Однако, если наше улучшение будет работать очень хорошо -- то количество запросов у конкурентов может уменьшиться, поэтому стоит рассматривать количество запросов у конкурентов как на русско язычном сегменте, так и на англоязычном.
  - b. Во-вторых, в качестве таргет-ряда, можно посмотреть, количество кликов в сутки, опять же в качестве ряда-предиктора можно рассматривать тот же показатель у конкурентов. Однако, стоит следить за несколькими конкурентами, потому что какой-нибудь конкретный конкурент мог тоже обновить поиск картинок.
  - c. В-третьих можно считать время поиска: от введения текста запроса, до перехода по какой-нибудь ссылке и закрытия поисковика. Аналогично можно использовать в качестве ряда-предиктора конкурентов. Однако, эта самая спорная метрика: потому что если человек не смог найти совсем ничего полезного -- то он тоже закроет поисковик. Таким образом график, полученный по этой метрике будет сложнее всего анализировать.
  - d. Также можно рассматривать количество новых пользователей. В качестве ряда-предиктора можно использовать эмпирические знания о притоке новых пользователей или же те же показатели у конкурентов (снова у конкурентов могут быть тоже вводиться новые фиши, которые увеличат приток пользователей к ним, поэтому лучше смотреть на нескольких).

Также во всех этих пунктах можно провести A/B тестирования: если только у половины пользователей будет улучшенный поиск. Но в этом случае будет проблема с тем, что разная выборка пользователей.

4. Для начала переведем все в одинаковые временные рамки, считая, что в CausalImpact\_homework\_data2 даты считаются в днях, начиная с 2013-12-18.

Выберем ряды-предикторы: они должны коррелировать с target до времени события -- августа 2015, и также должны коррелировать, но возможно чуть меньше, после ноября 2015. Но при этом не должны также сильно коррелировать в период проведения рекламной компании.

\*результат корреляции разных market с target\*

	до рекламы	во время	после
market_1	0.9343916	0.3420212	0.5452291
market_2	0.8996213	0.1718658	0.8236455
market_3	0.8224976	0.2177092	0.8048995
market_4	0.9570632	0.8777875	0.7736828

market_5	0.9349624	0.4058306	0.791723
----------	-----------	-----------	----------

То есть лучше всего брать в качестве ряда-предиктора market\_2, потом market\_1, market\_5. Остальные не прошли проверку.

Теперь исследуем влияние рекламных компаний:

№ рекламной компании	период	влияние		market_2	market_1	market_5
25	623,637	1,5	1	0,01454 98.546%	0.01377 98.623%	0.02331 97.669%
26	623,668	0,1	1	0.00112 99.88814%	0.00106 99.89407%	0.00106 99.89407 %
27	630,668	0	0	0.00106, 99.89407%	0.00106 99.89407%	0.00106 99.89407 %
28	633,644	0,7	0	0.04575 95.425%	0.04767 95.233%	0.06144 94%
29	651,667	0,4	0	0.06754 93%	0.04555 95.445%	0.03814 96.186%
30	690,720	0,2	1	0.05338 95%	0.03072 96.928%	0.02492 97.508%
31	694,707	0	1	0.00436 99.56427%	0.00103 99.89733%	0.00308 99.69199 %

Posterior tail-area probability p and Posterior prob. of a causal effect

Проблема заключается в том, что рекламные компании 25-29 все входят в рекламную кампанию 26. Аналогично в 30 входят 31, 33-35. 33-35 совсем невозможно отличить, но они закончились после ноября 2015, поэтому я их не рассматриваю.

В таблице везде значение p-value ниже 0.05, я также выделила значения, большие 0.01 -- как те, которым не совсем можно доверять

Таким образом, если считать, что влияние рекламы идет равномерно (что конечно же не так), тогда можно разделить влияние реклам, период действия которых

накладывались. Таким образом лучший результат достигается у 27 и 31 реклам -- что очень странно, потому что именно их влияние было оценено 0.