ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
"ВЫСШАЯ ШКОЛА ЭКОНОМИКИ"»**

Факультет экономических наук

Программа «Экономика»

Курсовая работа на тему
«Байесовский подход на Лабутенах»

Выполнил: Студент 3 курса группы БЭК 131
Федотова Мирослава Андреевна

Научный руководитель:
Старший преподаватель департамента прикладной экономики
Демешев Борис Борисович

Москва 2016

Содержание

1 Введение

В разные времена великие философы уделяли внимание причинности, как абсолютно нетривиальной категории. В нашем же обыденном сознании это понятие прочно укоренилось, и мы не подвергаем сомнению само существование каузальных зависимостей, как, например, Д. Юм, который считал, что необходимость причинности есть лишь продукт нашего сознания, вымысел. В современных науках встаёт несколько иной вопрос: а как формально установить наличие причинно-следственной связи?

Рассматриваемый нами подход позволяет дать оценку влияния некоторого события на динамику интересующего нас показателя, после чего можно говорить о значимости эффекта в целом. Появляется возможность находить ответы на многие вопросы, например, такие как: был ли эффективен выпуск нового продукта с точки зрения привлечения клиентов, как повлияла реклама на динамику продаж? В действительности эта модель может быть полезна не только в маркетинге и экономике, но и в совершенно отличных областях, таких как биология, политические и социальные науки. В данной работе будет рассмотрена модель состояний наблюдений [state-space model], которая для оценки каузального эффекта предсказывает динамику временного ряда в отсутствии влияния некоторого события. В отличие от классического метода разность разностей [difference-in-differences] модель состояний наблюдений: (1) позволяет сделать вывод об изменяющемся во времени эффекте, (2) включает априорные предпосылки о распределении параметров, (3) способна работать с несколькими источниками изменения динамики, включая локальные тренды, сезонность и непостоянные во времени коэффициенты регрессии. Используя алгоритмы Монте-Карло по схеме Марковской цепи для апостериорного вывода, мы продемонстрируем пользу модели на примере клипа на песню «Экспонат» группы «Ленинград».

2 Теоретическая модель

2.1 *Общий вид модели*

Байесовские структурные модели временных рядов достаточно гибки и позволяют подстраиваться к конкретному набору данных: включать локальный тренд, сезонность, динамические коэффициенты регрессии. В общем виде подобные модели можно описать следующими уравнениями:

$$y_t = Z_t^T \alpha_t + \epsilon_t, \tag{1}$$

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t, \tag{2}$$

где ошибки ϵ_t и η_t распределены нормально с нулевым математическим ожиданием. Уравнение (1) связывает наблюдения y_t с латентным вектором состояний α_t, динамика которого определяется уравнением (2). Подробнее о различных спецификациях можно прочесть в статье Kay H. Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, Steven L. Scott [2015]. Мы же будем строить модель, где $Z_t = \beta^T x_t$, а $\alpha_t = 1$, избавившись таким образом от изменяющихся во времени параметров.

Итак, наша базовая модель будет иметь следующий вид:

$$y = X\beta + \epsilon, \tag{3}$$

где y – вектор наблюдений, X – матрица контрольных переменных, β – вектор коэффициентов, $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{pmatrix} X = \begin{pmatrix} 1 & x_{1,1} & ... & x_{1,k} \\ 1 & x_{2,1} & ... & x_{2,k} \\ ... & ... & ... & ... \\ 1 & x_{n,1} & ... & x_{n,k} \end{pmatrix} \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_{k+1} \end{pmatrix}$$

2.2 *Априорные распередения параметров*

Обозначим за θ все параметры модели. Используемый нами Байесовский подход предполагает наличие априорных распределений параметров $p(\theta)$, описанием которых мы займёмся в этой главе. Стоит обратить внимание: вводимые нами предпосылки об априорных распределениях будут не самыми что ни на есть "чистыми по той простой причине, что будут зависеть в некоторой степени от имеющихся данных.

Сталкиваясь с большим количеством потенциальных регрессоров, мы предпочитаем позволить модели самостоятельно выбрать, какие именно использовать. Это делается с помощью предположения пик-плато [spike-and-slab prior], впервые описанного в статьях George and McCulloch [1997] и Madigan and Raftery [1994].

Введём вектор ϱ, отражающий, будет ли конкретный регрессор включен в модель. Зададим его следующим образом: $\varrho = (\varrho_1, ..., \varrho_J)$, где $\varrho_j = 1$ при $\beta_j \neq 0$ и $\varrho_j = 0$ при $\beta_j = 0$. Назовём β_ϱ вектор, который включает в себя все ненулевые элементы вектора β, и обозначим за $(\Sigma_\varrho)^{-1}$ матрицу точности, строки и столбцы которой соответствуют $(\Sigma)^{-1}$ для ненулевых ϱ. По теореме об умножении вероятностей разложим совместную функцию плотности величин $\varrho, \beta, \sigma^{-2}$ следующим образом:

$$p(\varrho, \beta, \sigma_\epsilon^{-2}) = p(\varrho, \sigma_\epsilon^{-2}) \cdot p(\beta|\varrho, \sigma_\epsilon^{-2}) = p(\varrho) \cdot p(\sigma_\epsilon^{-2}|\varrho) \cdot p(\beta_\varrho|\varrho, \sigma_\epsilon^{-2}) \tag{4}$$

Существует несколко способов определения функциональной формы априорных распределений. На практике чаще всего предполагают, что случайная величина ϱ имеет распределение Бернулли с параметром π_j:

$$p(\varrho) = \Pi_{j=1}^{J} \pi_j^{\varrho_j} (1 - \pi_j)^{1-\varrho_j}, \tag{5}$$

где π_j - априорная вероятность включения в модель регрессора j.

Когда априорной информации недостаточно, удобно полагать, что π_j одинаковы и соответствуют π. Эта вероятность, в свою очередь, определяется с помощью ожидаемого размера модели M. Если M коэффициентов из J ожидаются быть ненулевыми, тогда $\pi = \frac{M}{J}$ априори. Существуют куда более сложные способы задания $p(\varrho)$, например, небернуллиевские модели. Предпочтём оставить их вне зоны нашего рассмотрения.

Вводятся следующие предположения о распределениях:

$$\beta_\varrho|\sigma_\epsilon^2 \sim \mathcal{N}(\mathbf{b}_\varrho, \sigma_\epsilon^2(\Sigma_\varrho^{-1})^{-1}), \tag{6}$$

где вектор $\mathbf{b}$ отражает наши априорные ожидания относительно величины каждого элемента вектора β. На практике обычно полагают $\mathbf{b} = 0$. По предположению Зеллнера [Zellner (1986)] параметр $(\Sigma)^{-1}$ определяет априорную точность коэффициентов β в модели, включающей все исходные регрессоры. Поскольку $X^T X$ – полная информация Фишера по всем данным, параметризовав $(\Sigma)^{-1} = \frac{g}{n} X^T X$, получим среднюю информацию, доступную по g наблюдениям. Предположение Зеллнера становится неверным, когда матрица $X^T X$ не является положительно определённой, поэтому усредним её и получим:

$$(\Sigma)^{-1} = \frac{g}{n}\{wX^T X + (1 - x)diag(X^T X)\} \tag{7}$$

В нашей модели по умолчанию параметр $w = \frac{1}{2}$. Также будем полагать $g = 1$, что в некотором смысле может интерпретироваться, как будто бы мы подглядели одно наблюдение, исходя из которого составили своё априорное мнение.

Обратная величина к диспрерсии ошибки будет распределена следующим образом:

$$\frac{1}{\sigma_\epsilon^2} \sim \mathcal{G}(\frac{v_\epsilon}{2}, \frac{s_\epsilon}{2}), \tag{8}$$

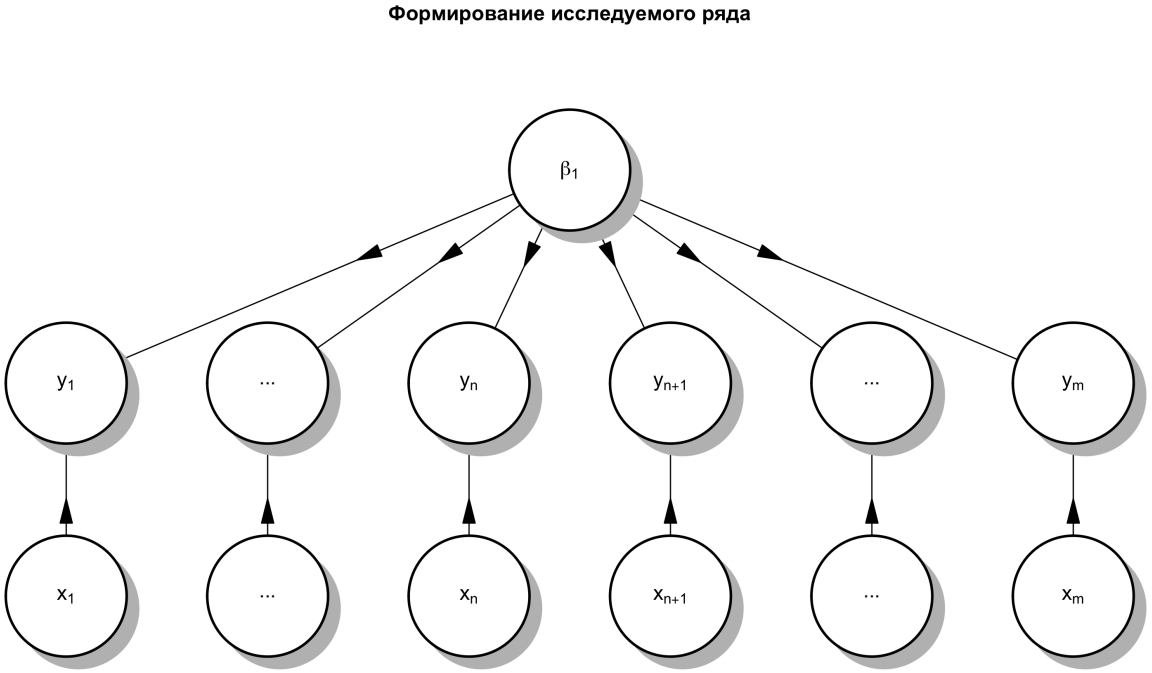
где $\mathcal{G}(.)$ – гамма-распределение с математическим ожиданием $\frac{v}{s}$. Чтобы определить значения v и s, нужно задаться вопросами: какой R^2 мы ожидаем получить, и какой вес v мы хотели бы задать в качестве предположения (разумно взять маленькое значение v). Зная R^2 и v нетрудно вычислить s:

$$s_\epsilon = v_\epsilon \cdot (1 - R^2) \cdot s_y^2, \tag{9}$$

где s_y^2 – оценка стандартного отклонения y в квадрате, иначе говоря, дисперсия.

Обобщим описанное нами выше: каждый из коэффициентов β_j модели (3) принимает либо нулевое значение с вероятностью $1 - \frac{M}{J}$, либо распределён в соответствии с законом (6) с вероятностью $\frac{M}{J}$. Иными словами, коэффициент может быть равен конкретной константе 0 (на графике это выглядело бы как пик) или же случайной величине (плато). Преимущество такого подхода к определению коэффициентов перед классической моделью состоит в том, что у нас появляется возможность напрямую задавать вопрос: чему равна вероятность того, что коэффициент окажется нулевым при заданых наблюдениях?

Изобразим схему формирования y до вмешательства и после. Будем предполагать, что само вмешательство произошло между периодами n и $n + 1$.



2.3 *Апостериорный вывод*

Апостериорный анализ нашей модели можно условно разделить на три шага:

1 шаг: получение параметров модели при наличии наблюдаемого вектора $y_{1:n}$ и матрицы $X_{1:n}$ в период до вмешательства.

2 шаг: использование сгенерированных на первом шаге параметров (не одно значение для каждого параметра, а множество) для моделирования апостериорного распределения $p(\widetilde{y}_{n+1:m}|y_{1:n}, X_{1:m})$

3 шаг: построение апостериорного распределния точечного влияния $y_t - \widetilde{y}_t$ для каждого $t = 1, ..., m$ с использованием апостериорной предиктивной выборки.

Обсудим все шаги более детально.

Постериорное симулирование. Мы используем сэмплирование по Гибсу для создания последовательности $\theta^{(1)}, \theta^{(2)}...$ из цепи Маркова, распределение которой $p(\theta|y_{1:n}, X_{1:n})$.

Получение коэффициентов β статической регрессии происходит следующим образом: для каждого $t = 1, ..., n$ в период до вмешательства рассмотрим вектор $y_{1:n} = (y_1, ..., y_n)$ и матрицу $X_{1:n}$. Наша задача заключается в генерировании параметров модели из распределения $p(\varrho, \beta, \sigma_\epsilon^{-2}|y_{1:n}, X_{1:n})$, которое представимо в виде:

$$p(\varrho, \beta, \sigma_\epsilon^{-2}|y_{1:n}) = p(\varrho, \sigma_\epsilon^{-2}|y_{1:n})p(\beta|\varrho, \sigma_\epsilon^{-2}, y_{1:n}) = \\ = p(\varrho|y_{1:n})p(\sigma_\epsilon^{-2}|\varrho, y_{1:n})p(\beta|\varrho, \sigma_\epsilon^{-2}, y_{1:n}) \tag{10}$$

Выинтегрировав параметры β и σ_ϵ^{-2}, получим:

$$\varrho|y_{1:n} \sim C(y_{1:n}) \frac{|\Sigma_\varrho^{-1}|^{1/2} p(\varrho)}{|V_\varrho^{-1}|^{1/2} S_\varrho^{(N/2)-1}}, \tag{11}$$

где $C(y_{1:n})$ неизвестная константа. Достаточная статистика для (11) выглядит следующим образом:

$$V_\varrho^{-1} = (X^T X)_\varrho + \Sigma_\varrho^{-1}, \tag{12}$$

$$\widetilde{\beta}_\varrho = (V_\varrho^{-1})^{-1}(X_\varrho^T y_{1:n} + \Sigma_\varrho^{-1} b_\varrho), \tag{13}$$

$$N = v_\epsilon + n, \tag{14}$$

$$S_\varrho = s_\epsilon + y_{1:n}^T y_{1:n} + b_\varrho^T \Sigma_\varrho^{-1} b_\varrho - \widetilde{\beta}_\varrho^T V_\varrho^{-1} \widetilde{\beta}_\varrho \tag{15}$$

Для получения выборки из распределения (11) мы используем сэмплирование по Гиббсу. Само условное распределение величины $\varrho|y_{1:n}$ легко оценить, поскольку ϱ может принимать всего 2 значения. Стоит заметить, что все матрицы, определяющие $p(\varrho|y_{1:n})$, имеют одну и ту же размерность, равную $\sum_j \varrho_j$. Несмотря на большое количество матриц, алгоритм работает быстро, поскольку каждая из них мала. А получив из распределения выборку для ϱ, сможем найти параметры из распределения $p(\beta, 1/\sigma_\epsilon^2|\varrho, y_{1:n})$.

Апостериорные предиктивные симуляции. Анализ каузальных зависимостей напрямую связан со следующей величиной:

$$p(\widetilde{y}_{n+1:m}|y_{1:n}, X_{1:m}) \tag{16}$$

Выражение (16) есть распределение исследуемой величины после вмешательства, как будто бы его вовсе и не было. Нас интересует предсказанная динамика временного ряда в отсутствии вмешательства.

Важно отметить, что функция плотности (16) ищется только при условии основного наблюдаемого временного ряда до вмешательства и контрольных рядов до и после вмешательства. Плотность условно **не зависит** от оценок параметров и включения и исключения из модели некоторых регрессоров. Таким образом, с помощью Байесовских моделей мы избегаем привязки как к конкретному набору регрессоров, так и к точечным коэффициентам.

Постериорная предективная функция плотности (16) определяется как совместное распределение всех контрафактных точек после вмешательства, а не как совокупность точечных одномерных распределений. Это гарантирует нам, что мы правильно распространим последовательную структуру, полученную на данных в период до вмешательства, на "контрафактную траекторию". Это имеет решающее значение как для итоговой статистики, так и для кумулятивного эффекта от вмешательства.

Оценка влияния. Выборка, полученная из постериорного предиктивного распределения для контрфактного ряда, используется нами для оценки каузального эффекта. Для каждой симуляции τ для каждого момента времени $t = n+1, ..., m$ зададим :

$$\phi_t^{(\tau)} = y_t - \widetilde{y}_t^{(\tau)} \tag{17}$$

Зачастую полезно знать не только точечный эффект от вмешательства, но и кумулятивный, накапливаемый с течением времени. Одним из основных преимуществ рассматриваемого подхода является гибгость и лёгкость, с которой могут быть получены результаты. Суммируя выражение (17) по t для каждого τ получаем:

$$\sum_{t'=n+1}^{t} \phi_{t'}^{(\tau)} \qquad \forall t = n+1, ..., m \tag{18}$$

3 Приложением модели к реальным данным

Для демонстрирования байесовского подхода мы рассмотрим следующее событие, которое будет в дальнейшем называть вмешательством: 13 января 2016 года в сети появился клип группы «Ленинград» на песню «Экспонат». В связи с этим пользователи начали активно вбивать в поисковую систему Google различные словосочетания, популярнейшим из которых было «На лабутенах». Безусловно, мы никого не удивим, если скажем, что появление клипа вызвало желание у людей увидеть его. Такая причинно-следственная связь не вызывает сомнений и не будет нами рассмотрена. А вот поиск ответа на вопрос о том, стали ли пользователи больше интересоваться самим дизайнером Кристианом Лабутеном или же его продукцией (если да, то в какой мере), на наш взгляд, куда более привлекательное занятие. Поэтому мы будем рассматривать запросы «Louboutin» именно на французском языке, а не на русском, чтобы избежать тех, которые напрямую связаны с желанием посмотреть клип. Для анализа была выбрана только территория Российской Федерации, поскольку мы предполагаем, что в других странах эффект от появления клипа либо был не таким ярко выраженным, либо он вовсе отсутствовал. Итак, цель нашего анализа - ответить на вопрос: можно ли считать значимым влияние появление клипа «Экспонат» на динамику количества запросов «Louboutin». Также нами будет дана количественная оценка каузального эффекта.

3.1 *Базовая спецификация модели*

Данные для анализа были взяты с сервиса Google Trends. Исследуемый ряд, который в описании модели носил название y_t, - количество запросов «Louboutin» [пере-менная louboutin]. В качестве контрольных рядов, которые формируют матрицу X, мы использовали количество запросов «Jimmy Choo» [переменная choo], «Manolo Blahnik» [переменная blahnik] и «Casadei» [переменная casadei] - дизайнерские марки обуви премиум класса той же ценновой категории, что и продукция

от Christian Louboutin. Рассматриваемый период - 1 год (с 31-05-2015 по 28-05-2016), данные еженедельные, поэтому каждый временной ряд включает всего 52 наблюдения, которые соответствует количеству накопленных за неделю запросов. Стоить обратить внимание, что сами данные представляют собой не реальное количество запросов, а нормированную величину, то есть все показатели выражены в процентном соотношении к максимальному количеству запросов.

Итак, подключим необходимые библиотеки:

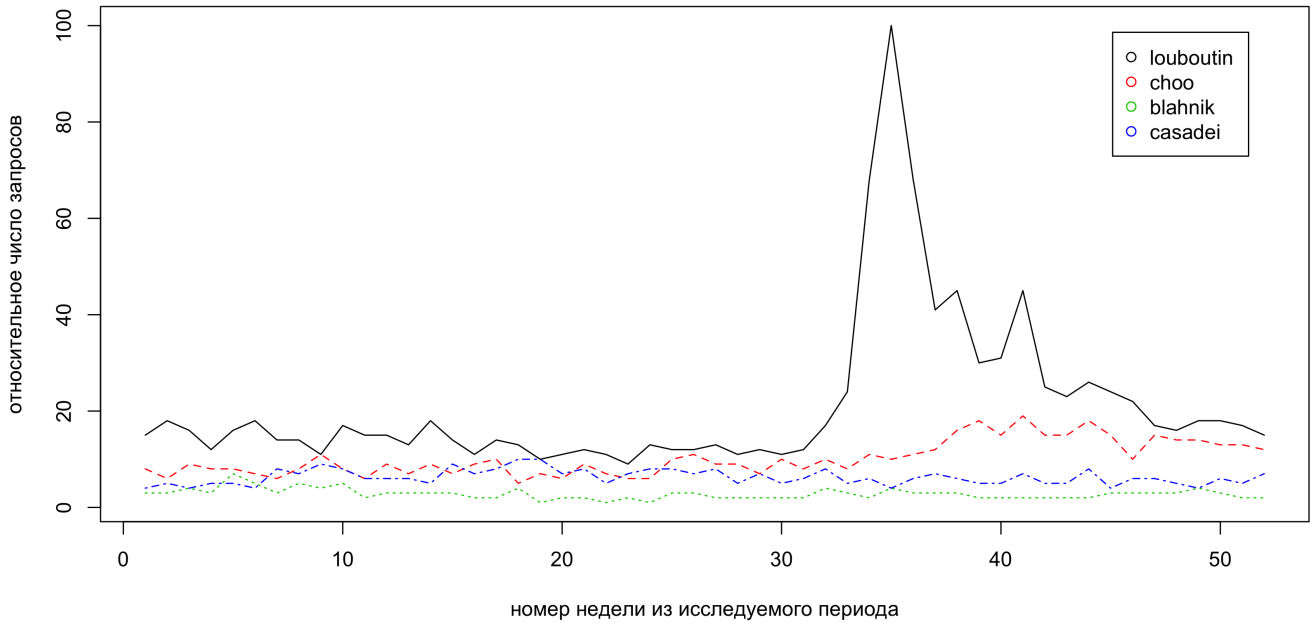
```
library(devtools)
library(CausalImpact)
library(zoo)
```

Выгрузим данные и введём переменные:

```
report <- read.csv("report-4.csv", header = F, stringsAsFactor = F)

louboutin <- report$V2
louboutin  <- louboutin[5:56]
louboutin <- as.numeric(louboutin)

choo <- report$V3
choo <- choo[5:56]
choo <- as.numeric(choo)

blahnik <- report$V4
blahnik <- blahnik[5:56]
blahnik <- as.numeric(blahnik)

casadei <- report$V5
casadei <- casadei[5:56]
casadei <- as.numeric(casadei)
```

Изобразим исследуемый и контрольные временные ряды на одном графике:

```
data <- cbind(louboutin, choo, blahnik, casadei)
matplot(data, type = "l", xlab = "номер недели из исследуемого периода",
        ylab = "относительное число запросов",
        verbose = getOption("verbose") )
legend("topright", inset=.05,
       legend=c("louboutin ", "choo", "blahnik", "casadei"),
       pch=1, col=c(1,2,3,4), horiz=FALSE)
```



Прежде чем строить и оценивать можедь, крайне важно проверить выполнение базовых предпосылок. В нашем случае их всего две. Первая: модель предполагает, что соотношение между контрольными и исследуемым рядами остаётся стабильным не только в период до вмешательства, но и после. К сожалению, эта предпосылка не может быть проверена, поэтому примем её на веру. Вторая: контрольные ряды не должны быть подвержены влиянию вмешательства. Это предпосылку мы проверим с помощью гипотезы о равенстве средних до и после вмешательства.

Разделим наблюдения по запросу «Manolo Blahnik» на две части : до и после вмешательства. Далее проверим гипотезу с помощью t-статистики:

```
blahnik_1 <- blahnik[1:33]
blahnik_2 <- blahnik[34:52]
t.test(blahnik_1, blahnik_2)

##
##  Welch Two Sample t-test
##
## data:  blahnik_1 and blahnik_2
## t = 1.0037, df = 49.69, p-value = 0.3204
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2779252  0.8329491
## sample estimates:
## mean of x mean of y
##  2.909091  2.631579
```

Как мы видим, p-value для ряда blahnik равно 0.3204, следовательно, на уровне значимости 5%, 10% и даже 30% гипотеза о равенстве средних не отвергается.

Проведём аналогичные действия для запросов «Casadei»:

```
casadei_1 <- casadei[1:33]
casadei_2 <- casadei[34:52]
t.test(casadei_1, casadei_2)

##
##  Welch Two Sample t-test
##
## data:  casadei_1 and casadei_2
## t = 2.6381, df = 49.076, p-value = 0.01114
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2466516 1.8235238
## sample estimates:
## mean of x mean of y
##  6.666667  5.631579
```

Для ряда casadei гипотеза не отвергается лишь на 1% уровне значимости, поскольку p-value $= 0.01114$.

И наконец, проверим гипотезу для запросов «Jimmy Choo»:

```
choo_1 <- choo[1:33]
choo_2 <- choo[34:52]
t.test(choo_1, choo_2)

##
##  Welch Two Sample t-test
##
## data:  choo_1 and choo_2
## t = -9.0027, df = 25.556, p-value = 2.102e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.371103 -4.628897
## sample estimates:
## mean of x mean of y
##         8        14
```

Для ряда choo гипотеза будет отвергнута на любом разумном уровне значимости. Тем не менее сперва мы не будем убирать регрессор choo, поскольку его колебание незначительно по сравнению с исследуемым временным рядом. Далее,

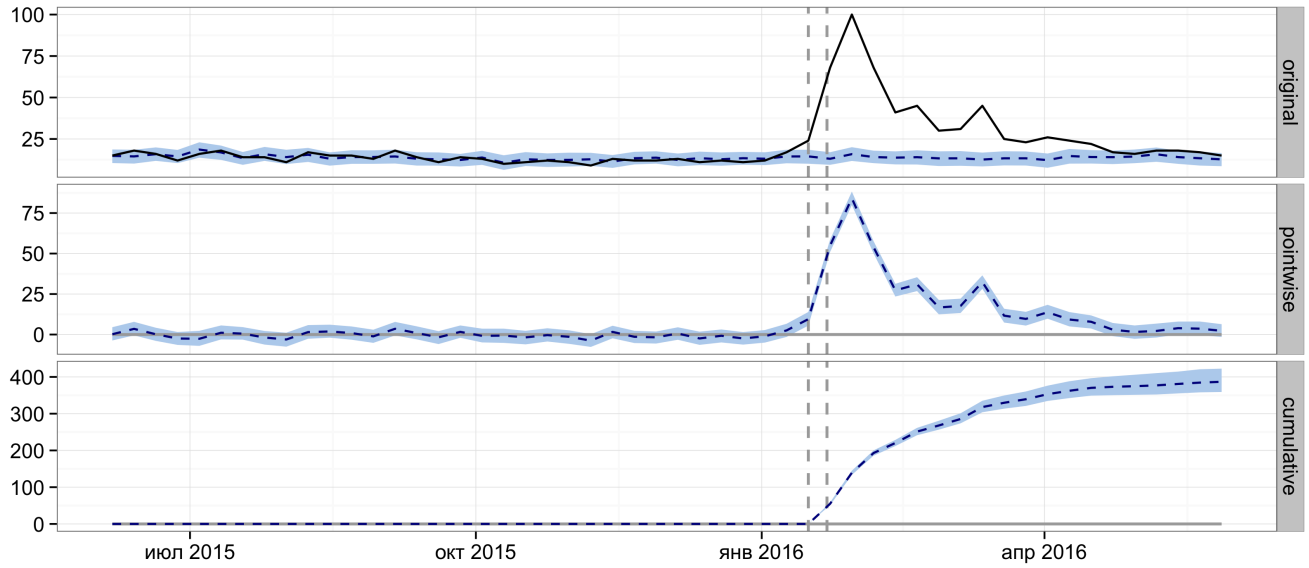
в другой спецификации модели, мы проверим, как отсутствие контрольного ряда choo повлияет на предсказание.

Прежде чем анализировать данные, обозначим наши наблюдения датами и разделим год на два периода: до и после вмешательства (pre.period и post.period соответственно):

```
time.points <- seq.Date(as.Date("2015-06-06"), length.out = 52, by = 7)

new_data <- zoo(data, time.points)

pre.period <- as.Date(c("2015-06-06", "2016-01-16"))
post.period <- as.Date(c("2016-01-23", "2016-05-28"))
```

Теперь проанализируем влияние и построим график:

```
impact <- CausalImpact(new_data, pre.period, post.period)
plot(impact)
```



Выведем статистику:

```
summary(impact)

## Posterior inference {CausalImpact}
##
##                          Average        Cumulative
## Actual                   34             649
## Prediction (s.d.)        14 (0.85)      262 (16.15)
```

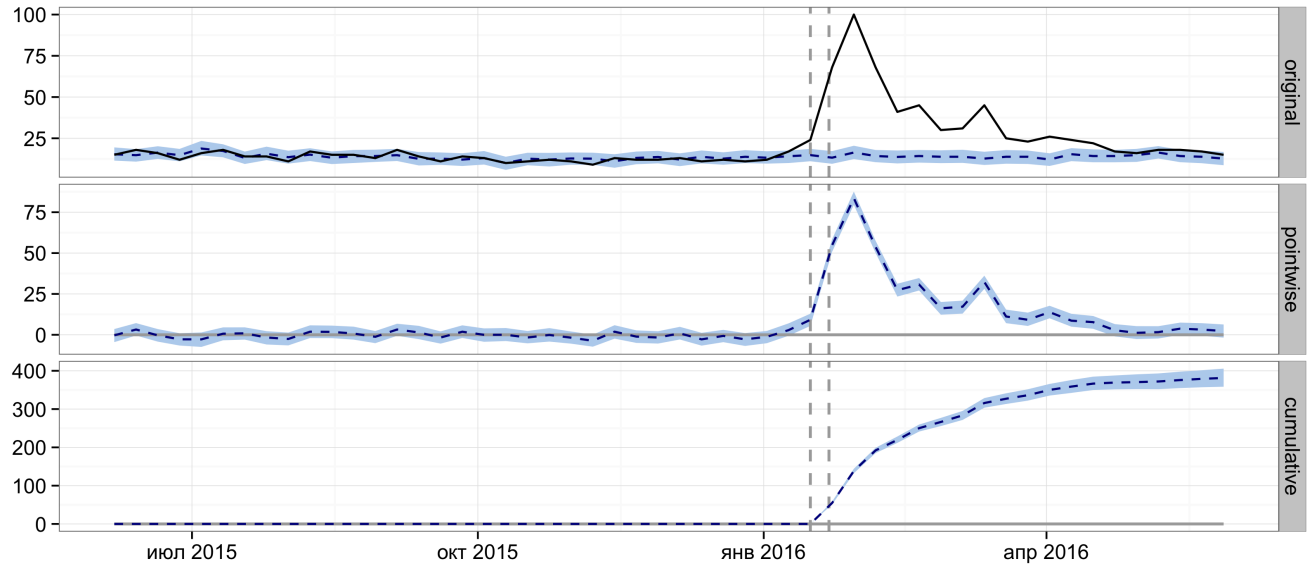
```
## 95% CI                    [12, 15]       [226, 290]
##
## Absolute effect (s.d.)   20 (0.85)      387 (16.15)
## 95% CI                    [19, 22]       [359, 423]
##
## Relative effect (s.d.)   148% (6.2%)    148% (6.2%)
## 95% CI                    [137%, 161%]   [137%, 161%]
##
## Posterior tail-area probability p:   0.00111
## Posterior prob. of a causal effect:  99.88901%
##
## For more details, type: summary(impact, "report")
```

Опишем полученные результаты. Вмешательством будем как и прежде называть появление клипа в сети. Напомним, что исходные данные были выражены в процентах от максимального количества запросов за исследуемый промежуток времени, то есть они отнормированны, в связи с чем не имеют прямой интерпретации как количества запросов. Итак, в течение периода после вмешательства наблюдаемая переменная в среднем имела значение 34. Однако в отсутствии вмешательства мы ожидали бы, что её значение в среднем составит 14, а с вероятностью 95% она попадёт в интервал [12, 14]. Сравнив фактическое и контрафактное значение исследуемой переменной, можем говорить о наличии эффекта от вмешательства, который в числовом выражении равняется 20. Поскольку в 95% доверительный интервал [19, 22] нулевое значение не попадает, мы не можем считать, что эффекта вовсе не было. Теперь обсудим кумулятивный эффект. Суммируя все точечные фактические значения, получаем 649. А если бы вмешательства вовсе не было, кумулятивное значение равнялось бы 262 с доверительным 95% интервалом [227, 291]. Стоит сказать, что в процентном соотношении по причине вмешательства значение наблюдаемой переменной выросло в среднем на 148%. С вероятностью 95% этот рост попал бы в интервал [137, 161]. Крайне мала вероятность того, что описанный выше положительный эффект, наблюдаемый после вмешательства, мог быть следствием случайных колебаний ($p = 0.001$). Приходим к выводу, что причинно-следственная связь может рассматриваться как статистически значимая.

3.2 *Модель с двумя контрольными рядами*

Как было замечено ранее, гипотеза о равенстве средних для ряда choo до и после появления клипа отвергается на любом разумном уровне значимости. В связи

с этим исключение из модели регрессора choo может повлиять на полученный ранее прогноз контрафактного ряда. Оценим новую модель, исключив контрольный ряд choo. Изобразим полученные результаты:



Выведем описание эффекта от вмешательства:

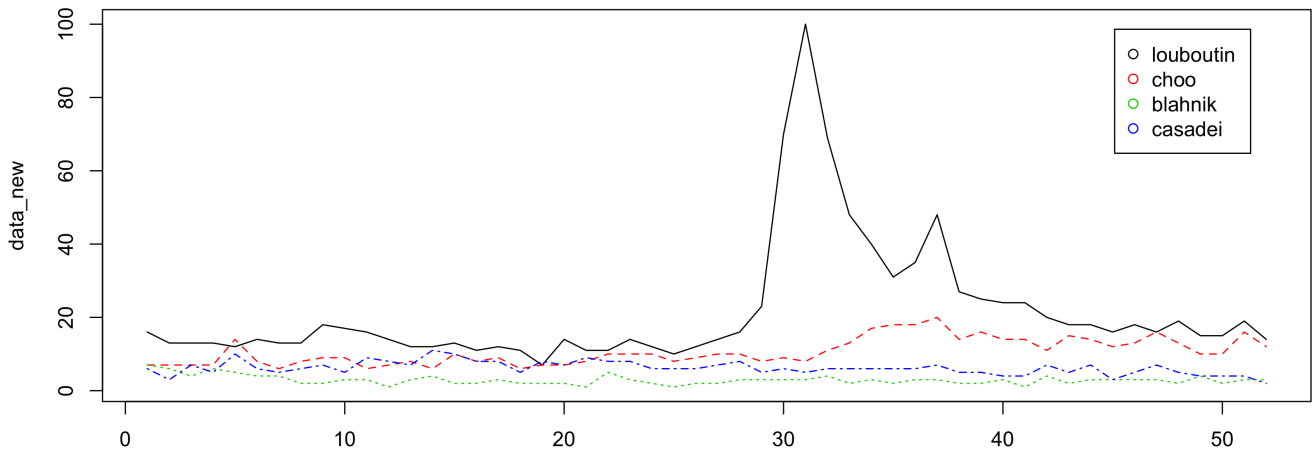
```
## Posterior inference {CausalImpact}
##
##                          Average        Cumulative
## Actual                   34             649
## Prediction (s.d.)        14 (0.63)      268 (11.91)
## 95% CI                   [13, 15]       [243, 291]
##
## Absolute effect (s.d.)   20 (0.63)      381 (11.91)
## 95% CI                   [19, 21]       [358, 406]
##
## Relative effect (s.d.)   142% (4.5%)    142% (4.5%)
## 95% CI                   [134%, 152%]   [134%, 152%]
##
## Posterior tail-area probability p:   0.00111
## Posterior prob. of a causal effect:  99.88901%
##
## For more details, type: summary(impact, "report")
```

Итак, в среднем в отсутствии вмешательства наблюдаемая переменая имела бы то же значение, что и в базовой спецификации модели, - 14 при фактическом значении 34. Её 95% доверительный интервал незначительно сместился: [13, 15]. В числовом выражениии эффект от вмешательства составил как и прежде 20, попадая в 95% доверительный интервал [19, 21], что вновь исключает ноль. Если бы

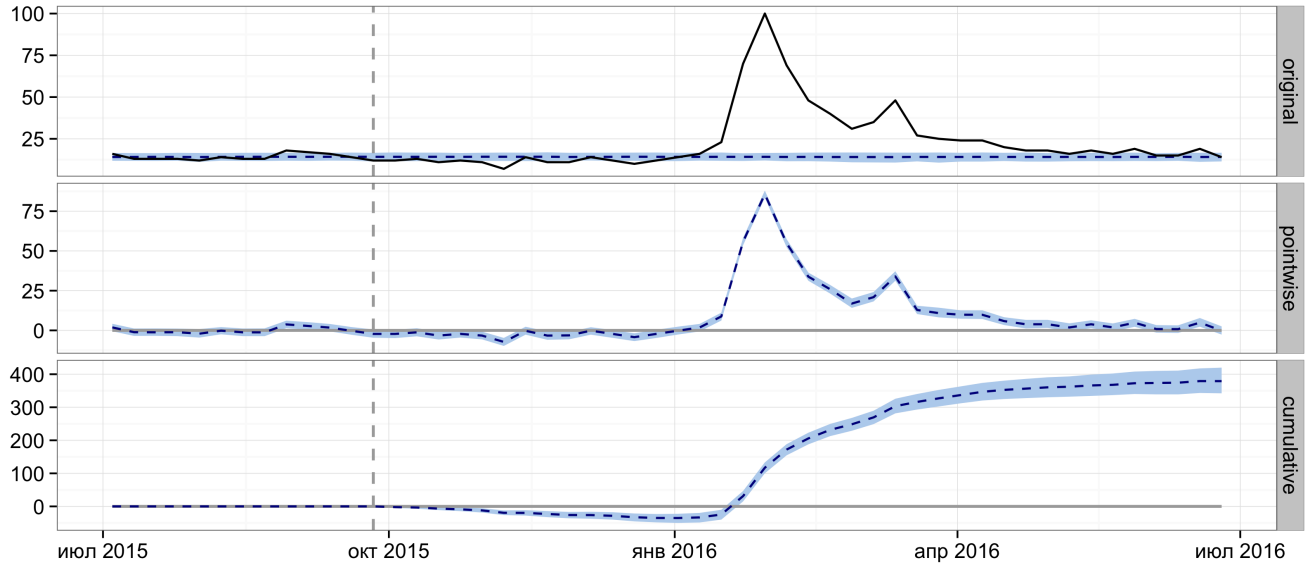
вмешательства не было вовсе, кумулитивный эффект составил бы 268 с 95% доверительным интервалом [246, 289] при фактивческом 649. Напомним, что базавая спецификация модели давала предсказание 262. Эта разница, вероятно, связана с тем фактом, что контрольный ряд choo имел скачок, который модель восприняла как общее колебание запросов, имевшее бы место и в отсутствии вмешательства. Что касается относительного эффекта, из-за появления клипа количество запросов "louboitin"выросло на 142%. По-прежнему крайне мала вероятность того, что положительный эффект, наблюдаемый после вмешательства, мог быть следствием случайных колебаний (модель предполагает, что p = 0.001), значит, причинно-следственная связь может рассматриваться как статистически значимая.

3.3 *Модель с изменённым периодом*

Напомним, что предыдущие две модели были построены на данных с 31-05-2015 по 28-05-2016. При этом клип появился в сети 13 января. Получается, что у нас имелись данные лишь по прошествии 4.5 месяцев с момента вмешательства. Вполне разумным кажется предположение, что чем больше времени проходит, тем менее заметен эффект от появления клипа. Поэтому сейчас, когда появилась возможность хотя бы на месяц увеличить период после вмешательства, мы построем модель базовой спецификации на годовых данных периодом с 28-06-2015 по 25-06-2016. Вновь построим график, наглядно описывающий общую динамику:



Изобразим график предсказания и выведем статистику:




```
## Posterior inference {CausalImpact}
##
##                          Average      Cumulative
## Actual                   24           932
## Prediction (s.d.)        14 (0.5)     553 (19.5)
## 95% CI                   [13, 15]     [512, 590]
##
## Absolute effect (s.d.)   9.7 (0.5)    379.1 (19.5)
## 95% CI                   [8.8, 11]    [342.3, 420]
##
## Relative effect (s.d.)   69% (3.5%)   69% (3.5%)
## 95% CI                   [62%, 76%]   [62%, 76%]
##
## Posterior tail-area probability p:   0.00111
## Posterior prob. of a causal effect:  99.88901%
##
## For more details, type: summary(impact, "report")
```

Итак, что мы видим: после появления клипа среднее количество запросов "louboutin"(напомним, в относительном выражении, где максимальное взято за 100) упало сравнительно с базовой спецификацией модели и составило 32. Однако среднее для предсказания осталось прежним - 14. Сравнив фактическое и контрафактное значение исследуемой переменной, можем говорить о наличии эффекта от вмешательства, который составил 18, попадая в 95% доверительный интервал [16, 20]. Можем считать, что вмешательство действительно оказало влияние, поскольку нулевое значение не попало в этот интервал.

Теперь обсудим кумулятивный эффект. Суммируя все точечные фактические значения, получаем 729, что выше прежнего результата на 80 единиц. А контра-

фактое значение составило 314, в то время как в базовой модели - 262. Однако рост суммарных эффектов - лишь следствие того, что мы увеличили период после вмешательства. В процентном соотношении по причине вмешательства значение наблюдаемой переменной в среднем выросло на 133%. С вероятностью 95% этот рост попал бы в интервал [116, 143]. Крайне мала вероятность того, что эффект, наблюдаемый после вмешательства, мог быть следствием случайных колебаний (p=0.001), поэтому каузальный эффект может рассматриваться как статистически значимый.

Ранее мы рассмотрели самые простые модели, где большинство параметров были заданы по умолчанию. На самом деле пакет CausalImpact предлагает массу возможностей для построения собственной модели, ориентированной на специфику конкретно поставленной задачи и имеющихся данных. Для первого знакомства с такими моделями мы покажем, каким образом можно менять априорные предпосылки. К примеру, по умолчанию $M = 3$ (априорный ожидаемый размер модели, о котором мы говорили в теоретической части). Предположим, нам по некоторой причине захотелось, чтобы модель априорно отбирала лишь 2 регрессора, одним из которых является константа. Тогда сперва нужно убрать наблюдения из исследуемого ряда после вмешательства, предварительно сохранив копию этих наблюдений (для этого мы создали переменную post.period.response):

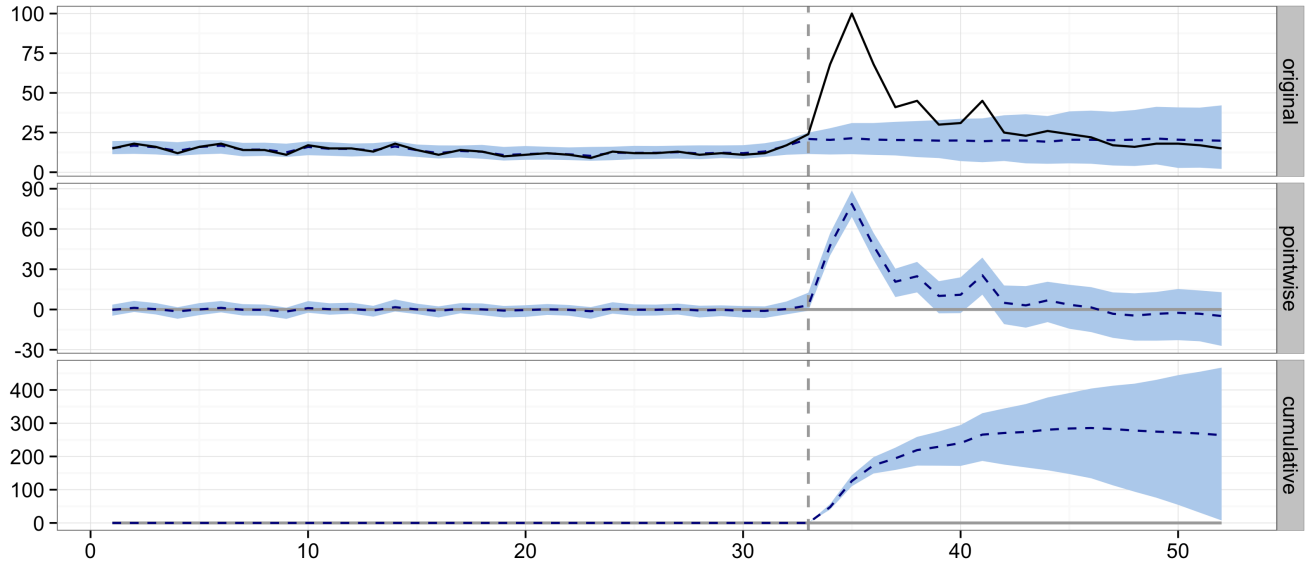
```
post.period_1 <- c(34, 52)
post.period.response <- louboutin[post.period_1[1] : post.period_1[2]]
louboutin_1 <- louboutin
louboutin_1[post.period_1[1] : post.period_1[2]] <- NA
```

Создадим матрицу из регрессоров и оценим байесовскую структурную модель:

```
X <- cbind(1, choo, blahnik, casadei)
ss <- AddLocalLevel(list(), louboutin_1)
myprior <- SpikeSlabPrior(X, louboutin_1,
                          expected.model.size = 2)
bsts.model <- bsts( louboutin_1 ~ choo + blahnik + casadei, ss,
                    niter = 1000, prior = myprior)

## =-=-=-=-= Iteration 0 Sun Jun 26 21:27:41 2016 =-=-=-=-=
## =-=-=-=-= Iteration 100 Sun Jun 26 21:27:41 2016 =-=-=-=-=
## =-=-=-=-= Iteration 200 Sun Jun 26 21:27:41 2016 =-=-=-=-=
## =-=-=-=-= Iteration 300 Sun Jun 26 21:27:41 2016 =-=-=-=-=
## =-=-=-=-= Iteration 400 Sun Jun 26 21:27:41 2016 =-=-=-=-=
## =-=-=-=-= Iteration 500 Sun Jun 26 21:27:41 2016 =-=-=-=-=
## =-=-=-=-= Iteration 600 Sun Jun 26 21:27:41 2016 =-=-=-=-=
## =-=-=-=-= Iteration 700 Sun Jun 26 21:27:41 2016 =-=-=-=-=
## =-=-=-=-= Iteration 800 Sun Jun 26 21:27:42 2016 =-=-=-=-=
## =-=-=-=-= Iteration 900 Sun Jun 26 21:27:42 2016 =-=-=-=-=
```

И наконец, оценим эффект от вмешательства и построим график:

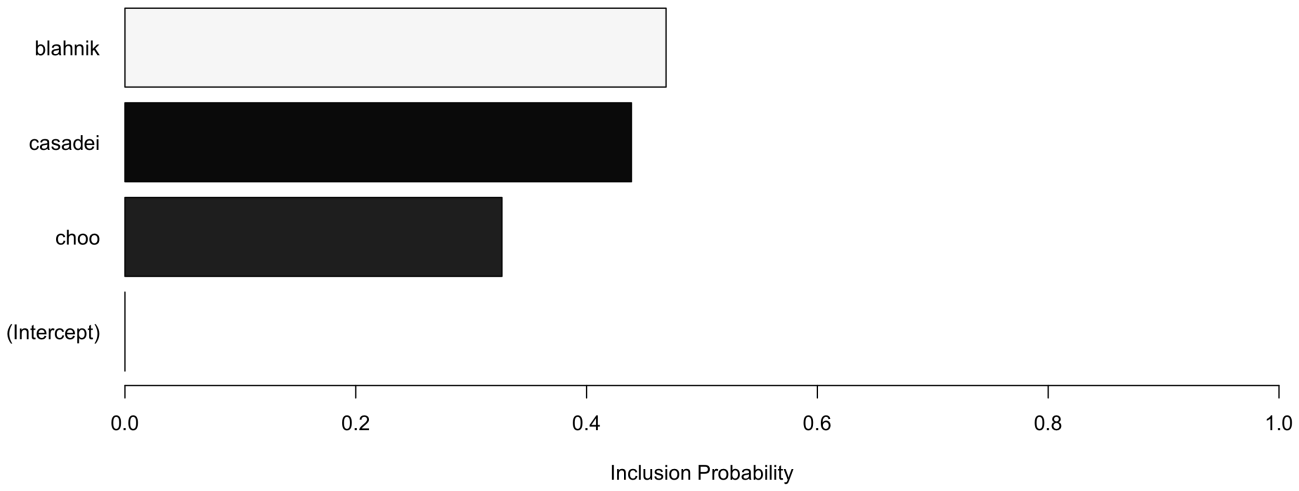
```
impact_1 <- CausalImpact(bsts.model = bsts.model,
                         post.period.response = post.period.response)
plot(impact_1)
```



Выведем результат, чтобы любой желающий имел возможность сравнить его с полученным в ранее описанных моделях:

```
summary(impact_1)

## Posterior inference {CausalImpact}
##
##                          Average      Cumulative
## Actual                   34           649
## Prediction (s.d.)        20 (6.5)     385 (122.7)
## 95% CI                   [9.6, 34]    [181.8, 641]
##
## Absolute effect (s.d.)   14 (6.5)     264 (122.7)
## 95% CI                   [0.4, 25]    [7.6, 467]
##
## Relative effect (s.d.)   69% (32%)    69% (32%)
## 95% CI                   [2%, 121%]   [2%, 121%]
##
## Posterior tail-area probability p:   0.0222
## Posterior prob. of a causal effect:  97.78%
##
## For more details, type: summary(impact, "report")
```

Если у нас возникнет желание, мы всегда можем посмотреть, с какой вероятностью будет включён каждый регрессор в модель:



4 Заключение

Мы ознакомились с моделями, которые предлагают большое количество возможностей для оценки влияния конкретного события на динамику исследуемого показателя. На примере рассмотренных данных мы увидели, что изменение априорных предпосылок влияет на результат намного сильнее, нежели иключение регрессора или изменение периода. Поэтому нужно глубоко разобраться в теории, прежде чем строить собственную модель. Стоит также сказать, что результаты, полученные во всех рассмотренных моделях, интерпретировались бы лучше, будь у нас данные по количеству запросовв в явном виде, а не в нормированных величинах. Например, в нашем случае кумулятивный эффект отражал бы дополнительное число запросов, полученное благодаря появлению клипа. Всё же, несмотря на некоторые недостатки построенных моделей, мы можем смело дать ответ на поставленный вопрос: появление клипа группы «Ленинград» на песню «Экспонат» повлияло на заинтересованность российских граждан к самому Кристиану Лабутену или же продукции его компании. Вероятно, французскому дизайнеру стоит задуматься над наймом нового пиар-менеджера. Но нас это уже не касается.

Итак, байесовский подход позволил нам иначе посмотреть на этот мир, породив множество новых вопросов, на которые так хотелось бы найти ответы. Как говорил Сократ: чем больше мы знаем, тем с большим облаком незнания мы сталкиваемся.

5 Литература

1. Brodersen K. H. et al. Inferring causal impact using Bayesian structural time-series models //The Annals of Applied Statistics. – 2015. – Т. 9. – №. 1. – С. 247-274.

2. Scott S. L., Varian H. R. Bayesian variable selection for nowcasting economic time series. – National Bureau of Economic Research, 2013. – №. w19567.

3. George E. I., McCulloch R. E. Approaches for Bayesian variable selection //Statistica sinica. – 1997. – С. 339-373.

4. CausalImpact : An R package for causal inference in time series.
https://google.github.io/CausalImpact/CausalImpact.html

5. Soetaert K. et al. Package 'diagram'. – 2014.