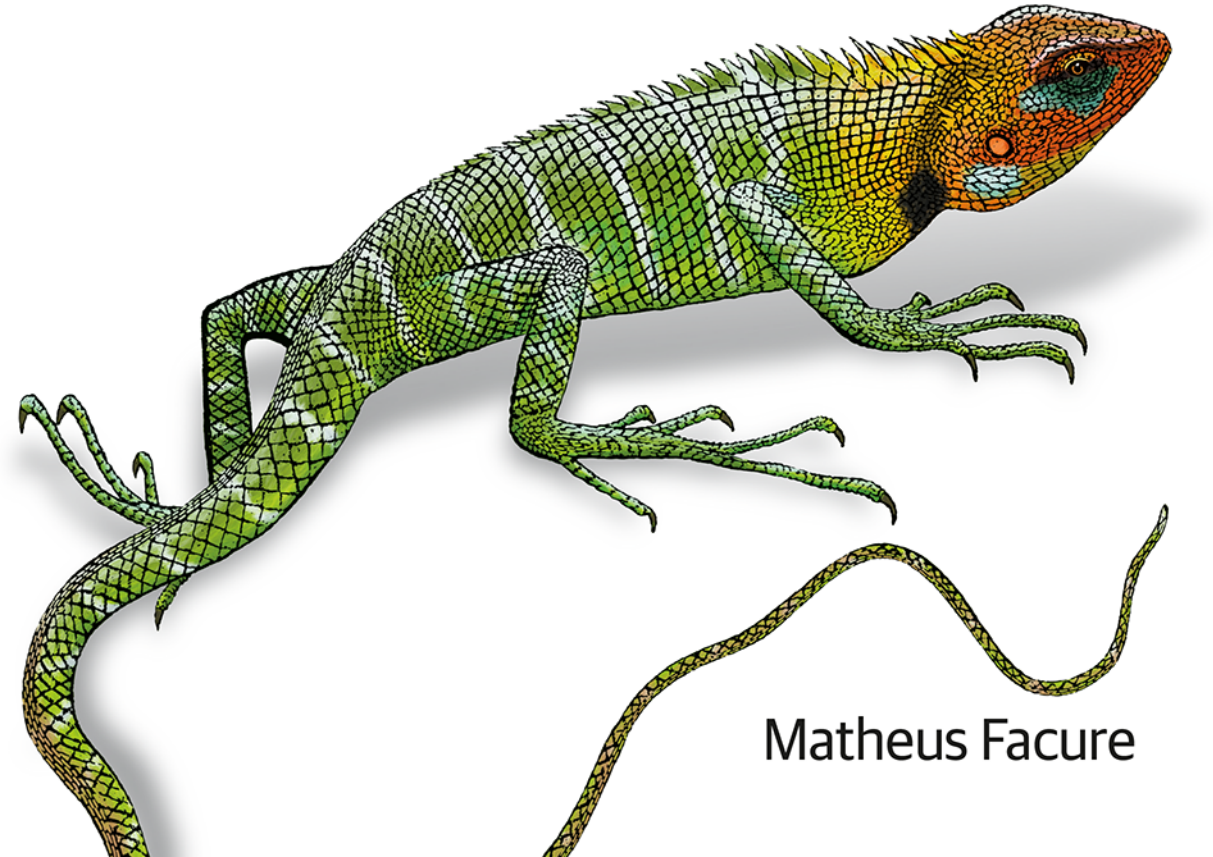
O'REILLY®

Helion

Wnioskowanie przyczynowe w Pythonie

Praktyczne wykorzystanie w branży technologicznej



Matheus Facure

Tytuł oryginału: Causal Inference in Python: Applying Causal Inference in the Tech Industry

Tłumaczenie: Tomasz Walczak

ISBN: 978-83-289-0881-9

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
https://helion.pl/user/opinie/wniprz
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Pliki z przykładami omawianymi w książce można znaleźć pod adresem:
https://ftp.helion.pl/przyklady/wniprz.zip

Helion S.A.
ul. Kościuszki 1c, 44-100 Gliwice
tel. 32 230 98 63
e-mail: *helion@helion.pl*
WWW: *https://helion.pl* (księgarnia internetowa, katalog książek)

Printed in Poland.

Spis treści

ROZDZIAŁ 1.

Wprowadzenie do wnioskowania przyczynowego

W niniejszym pierwszym rozdziale przedstawiam wiele podstawowych zagadnień z obszaru wnioskowania przyczynowego, a także główne wyzwania i zastosowania z nim związane. Poznasz tutaj wiele żargonowych pojęć, które będą używane w dalszej części książki. Zawsze pamiętaj też, dlaczego potrzebujesz wnioskowania przyczynowego i co możesz dzięki niemu uzyskać. Ten rozdział nie jest poświęcony pisaniu kodu, ale bardzo ważnym podstawowym zagadnieniom związanym z wnioskowaniem przyczynowym.

Czym jest wnioskowanie przyczynowe?

Przyczynowość jest czymś, co możesz uważać za niebezpieczny obszar epistemologii, którego należy unikać. Twój nauczyciel statystyki mógł w kółko powtarzać, że „asocjacja nie oznacza związku przyczynowo-skutkowego", a mylenie tych dwóch pojęć naraziłoby Cię na akademicki ostracyzm lub co najmniej na ostrą krytykę. Ale właśnie w tym rzecz, że *czasami asocjacja oznacza przyczynowość*.

My, ludzie, wiemy o tym aż za dobrze, ponieważ najwyraźniej zostaliśmy zaprogramowani do przyjmowania asocjacji za przyczynę. Możliwe, że decydujesz się nie wypić czwartego kieliszka wina, ponieważ prawidłowo wywnioskowałeś, iż zepsuje Ci to następny dzień. Czerpiesz z doświadczeń z przeszłości — z wieczora, kiedy wypiłeś za dużo i obudziłeś się z bólem głowy, a także z innego wieczora, kiedy wypiłeś tylko jeden kieliszek wina lub wcale i czułeś się dobrze. Dzięki temu wiesz, że istnieje coś więcej niż asocjacja między piciem a kacem. Wywnioskowałeś z tego przyczynowość.

Z drugiej strony jest trochę prawdy w ostrzeżeniach nauczyciela statystyki. Związek przyczynowy to skomplikowana sprawa. Kiedy byłem dzieckiem, dwa razy zjadłem smażone kalmary i w obu przypadkach skończyło się to fatalnie, co doprowadziło mnie do wniosku, że jestem na nie uczulony (jak również na małże, ośmiornice i każdy inny rodzaj morskich bezkręgowców). Czekałem ponad dwadzieścia lat, aby ponownie spróbować tego dania. Kiedy to zrobiłem, kalmary okazały się nie tylko pyszne, ale też nie odczułem żadnych negatywnych skutków. W tym przypadku pomyliłem asocjację z przyczynowością. Była to niegroźna pomyłka, ponieważ jedynie pozbawiła

mnie na kilka lat pysznych owoców morza, ale pomylenie asocjacji z przyczynowością może mieć znacznie poważniejsze konsekwencje. Jeśli inwestujesz na giełdzie, prawdopodobnie zdarzały Ci się sytuacje, gdy zdecydowałeś się dokonać zakupu tuż przed gwałtownym wzrostem cen lub wycofać się tuż przed załamaniem. Prawdopodobnie skłoniło Cię to do myślenia, że potrafisz wyczuć właściwy moment na rynku. Jeżeli udało Ci się zignorować tę pokusę, to gratuluję. Jednak wiele osób mylnie uznaje, że ich intuicja jest przyczynowo powiązana z nieregularnymi zmianami cen akcji. Zdarza się, że to przekonanie prowadzi do coraz bardziej ryzykownych transakcji, aż w końcu inwestor traci prawie wszystkie środki.

W skrócie można powiedzieć, że asocjacja zachodzi, gdy dwie wartości lub zmienne losowe poruszają się razem, podczas gdy przyczynowość występuje, gdy modyfikacja jednej zmiennej powoduje zmianę drugiej. Na przykład można powiązać liczbę Nagród Nobla w danym kraju z konsumpcją czekolady na mieszkańca, ale nawet jeśli te zmienne są ze sobą powiązane, głupotą byłoby sądzić, że jeden z tych czynników wpływa na drugi. Łatwo jest zrozumieć, dlaczego asocjacja nie implikuje związku przyczynowego, ale zrównanie tych dwóch rzeczy to zupełnie inna sprawa. *Wnioskowanie przyczynowe to nauka o ustalaniu przyczynowości na podstawie asocjacji i określaniu, kiedy i dlaczego te dwa zjawiska się różnią.*

Po co stosować wnioskowanie przyczynowe?

Wnioskowanie przyczynowe może być przeprowadzane w celu samego zrozumienia rzeczywistości. Często ma ono jednak komponent normatywny. Powodem, dla którego wywnioskowałeś, że zbyt duża ilość alkoholu powoduje ból głowy, jest to, że chcesz zmienić swoje nawyki związane z piciem, aby uniknąć bólu. Firma, w której pracujesz, chce wiedzieć, czy wydatki na marketing powodują wzrost przychodów, ponieważ jeśli tak jest, menedżerowie mogą wykorzystać to do zwiększenia zysków. Na ogólnym poziomie można powiedzieć, że *chcesz poznać relacje przyczynowo-skutkowe, aby móc wpływać na przyczynę w celu uzyskania pożądanego efektu*. W kontekście biznesowym wnioskowanie przyczynowo-skutkowe staje się gałęzią nauk decyzyjnych.

Ponieważ niniejsza książka koncentruje się głównie na biznesie, omawiam tu tę część dziedziny wnioskowania przyczynowego, która zajmuje się zrozumieniem wpływu interwencji. Co by się stało, gdybyś wyznaczył inną cenę zamiast tej, której obecnie żądasz za swój towar? Co by się stało, gdybyś przeszedł z diety o niskiej zawartości cukru na dietę niskotłuszczową? Co stanie się z marżami banku w wyniku zwiększenia linii kredytowych klientów? Czy rząd powinien dać tablety wszystkim dzieciom w szkołach, aby poprawić wyniki w testach czytania, czy też może powinien budować tradycyjne biblioteki? Czy małżeństwo korzystnie wpływa na finanse osobiste, czy też pary małżeńskie są bogatsze tylko dlatego, że bogaci ludzie mają większe szanse na przyciągnięcie partnera? Wszystkie te pytania mają charakter praktyczny i wynikają z chęci zmiany czegoś w biznesie lub w swoim życiu, aby uzyskać lepsze efekty.

Uczenie maszynowe i wnioskowanie przyczynowe

Jeśli przyjrzysz się bliżej rodzajom pytań, na które wnioskowanie przyczynowe ma dawać odpowiedzi, zobaczysz, że są to głównie pytania typu „co by było, gdyby". Przykro mi to przyznać, ale uczenie maszynowe jest wprost beznadziejne w udzielaniu odpowiedzi na pytania tego rodzaju.

Uczenie maszynowe świetnie się sprawdza w generowaniu prognoz. Jak ujęli to Ajay Agrawal, Joshua Gans i Avi Goldfarb w książce *Prediction Machines* (wydawnictwo Harvard Business Review Press), „nowa fala sztucznej inteligencji nie przynosi nam prawdziwej inteligencji, tylko jej niezwykle ważny aspekt — przewidywanie". Dzięki uczeniu maszynowemu można uzyskać wiele wartościowych informacji. Jedynym wymogiem jest umieszczenie problemu w kontekście przewidywania. Chcesz tłumaczyć z angielskiego na portugalski? Zbuduj model uczenia maszynowego, który przewiduje portugalskie zdania na podstawie danego zdania angielskiego. Chcesz rozpoznawać twarze? Opracuj model uczenia maszynowego, który przewiduje obecność twarzy w podsekcji obrazu.

Uczenie maszynowe nie jest jednak uniwersalnym rozwiązaniem. Może zdziałać cuda w sztywno określonych granicach, ale zupełnie zawieść, jeśli dane nieco odbiegają od tego, do czego model jest przyzwyczajony. Oto inny przykład z książki *Prediction Machines*: „w wielu branżach niskie ceny wiążą się z niską sprzedażą. Na przykład w branży hotelarskiej ceny są niskie poza sezonem turystycznym, a wysokie, gdy popyt jest najwyższy, a hotele są pełne. Na podstawie tych danych naiwne wykorzystanie predykcji może sugerować, że podniesienie ceny doprowadzi do większej liczby sprzedanych pokoi".

Uczenie maszynowe wykorzystuje asocjację między zmiennymi, aby przewidzieć jedną z nich na podstawie drugiej. Model będzie działać niezwykle skutecznie, o ile nie zmodyfikujesz zmiennych używanych do generowania prognoz. To sprawia, że w większości scenariuszy podejmowania decyzji, które wymagają interwencji, korzystanie z predykcyjnego uczenia maszynowego nie ma żadnego sensu.

Ponieważ większość danologów wie dużo o uczeniu maszynowym, ale niewiele o wnioskowaniu przyczynowo-skutkowym, modele uczenia maszynowego często stosuje się w scenariuszach, w których nie są przydatne. Jednym z głównych celów firm jest *zwiększenie* sprzedaży lub czasu użytkowania ich produktów. Jednak model uczenia maszynowego, który tylko przewiduje sprzedaż, jest w tym kontekście często bezużyteczny, a nawet szkodliwy. Taki model może nawet generować bezsensowne wnioski, tak jak w przykładzie, w którym wysoki wolumen sprzedaży wiąże się z wysokimi cenami. Zdziwiłbyś się jednak, jak wiele firm wdraża predykcyjne modele uczenia maszynowego, choć cel nie ma nic wspólnego z przewidywaniami.

Nie oznacza to, że uczenie maszynowe jest całkowicie bezużyteczne we wnioskowaniu przyczynowo-skutkowym. Jednak gdy jest stosowane w naiwny sposób, często przynosi więcej szkody niż pożytku. Jeśli jednak potraktujesz uczenie maszynowe jak zestaw wartościowych modeli, a nie wyłącznie maszyn predykcyjnych, zaczniesz dostrzegać, w jaki sposób można je wykorzystać do wnioskowania przyczynowego. W części III pokazuję, na co należy uważać, gdy łączysz uczenie maszynowe z wnioskowaniem przyczynowo-skutkowym, a także jak zmienić przeznaczenie popularnych algorytmów uczenia maszynowego, takich jak drzewa decyzyjne i wzmacnianie gradientowe, na potrzeby wnioskowania przyczynowo-skutkowego.

Asocjacja a związek przyczynowy

Intuicyjnie wiadomo, dlaczego asocjacja nie oznacza przyczynowości. Jeśli ktoś powie Ci, że najwyższej klasy konsultanci spowodują poprawę wyników Twojej firmy, z pewnością uniesiesz brew. Skąd możesz wiedzieć, czy agencja konsultingowa faktycznie prowadzi do poprawy wyników, czy może tylko dobrze prosperujące firmy stać na takie usługi?

Oto bardziej konkretny opis — wyobraź sobie, że pracujesz w firmie prowadzącej internetową platformę sprzedażową. Małe i średnie firmy korzystają z tej platformy do reklamowania i sprzedaży swoich produktów. Firmy te mają pełną samodzielność w kwestiach takich jak ustalanie cen i czasu sprzedaży. Ale w najlepszym interesie Twojej firmy jest to, aby jej klienci rozwijali się i prosperowali. Decydujesz się więc pomóc im przez udzielanie wskazówek dotyczących tego, jak, czy i kiedy organizować kampanie sprzedażowe z okresowymi obniżkami cen dla konsumentów. Aby to zrobić, pierwszą rzeczą, którą musisz wiedzieć, jest *wpływ obniżek na liczbę sprzedanych produktów*. Jeśli zyski ze sprzedaży większej liczby sztuk rekompensują straty wynikające z rabatów, wyprzedaż jest dobrym pomysłem. Jeżeli jeszcze tego nie zauważyłeś, zwracam uwagą na to, że pojawia się tu pytanie przyczynowe. Musisz ustalić, ile dodatkowych sztuk (w porównaniu z brakiem jakichkolwiek działań) sprzeda firma, gdy obniży ceny.

Nie trzeba dodawać, że jest to skomplikowane pytanie. Być może jest ono zbyt skomplikowane jak na początek tej książki. W ramach platformy działają różne firmy. Niektóre sprzedają żywność, inne ubrania, a jeszcze inne nawozy i produkty rolne. Obniżki cen mogą mieć różny wpływ w zależności od rodzaju działalności. Na przykład dla firmy odzieżowej dobrym pomysłem może być rozpoczęcie wyprzedaży na tydzień przed Dniem Ojca. Jednak podobne rabaty prawdopodobnie niewiele dadzą firmom z branży rolniczej. Uprośćmy więc nieco problem. Skupmy uwagę tylko na jednym rodzaju działalności, na sklepach sprzedających zabawki dla dzieci, a także na jednym okresie w roku, na grudniu, czyli okresie przedświątecznym. Na razie postaraj się odpowiedzieć na pytanie, w jaki sposób rabaty w tym okresie zwiększają sprzedaż, abyś mógł przekazać te informacje firmom działającym w branży zabawek dla dzieci i pomóc im w podejmowaniu lepszych decyzji.

Aby zdecydować, czy wyprzedaże są dobrym pomysłem, możesz wykorzystać informacje z wielu firm oferujących zabawki dla dzieci. Dane te są przechowywane w ramce danych biblioteki `pandas`, do której masz dostęp. W tabeli 1.1 znajdziesz kilka pierwszych wierszy, które pozwolą Ci ustalić, jakie dane są dostępne.

Tabela 1.1. Używane dane sprzedażowe

	store	weeks_to_xmas	avg_week_sales	is_on_sale	weekly_amount_sold
0	1	3	12,98	1	219,60
1	1	2	12,98	1	184,70
2	1	1	12,98	1	145,75
3	1	0	12,98	0	102,45
4	2	3	19,92	0	103,22
5	2	2	19,92	0	53,73

Pierwsza kolumna to unikatowy identyfikator sklepu (ID). Dostępne są tygodniowe dane z grudnia dla każdego sklepu. Podane są również informacje o wielkości każdej firmy mierzone średnią tygodniową sprzedażą produktów w danym roku. Kolumna z wartością logiczną (0 lub 1) informuje, czy firma prowadziła wyprzedaż w danym czasie. Ostatnia kolumna pokazuje średnią tygodniową sprzedaż określonego sklepu w danym tygodniu.

Jednostka analizy

Jednostką analizy w badaniach obejmujących wnioskowanie przyczynowe jest zwykle rzecz, której dotyczy interwencja (oddziaływanie). Zazwyczaj jednostką analizy są ludzie, tak jak w scenariuszu, gdy chcesz poznać wpływ wprowadzenia nowego produktu na utrzymanie klientów. Nie jest jednak niczym niezwykłym stosowanie innych typów jednostek analizy. W przykładzie z tego rozdziału jednostką analizy jest biznes. W tym samym przykładzie można również spróbować odpowiedzieć na pytanie, *kiedy* jest najlepszy moment na przeprowadzenie wyprzedaży. Wtedy jednostką analizy będzie okres (w tym scenariuszu tydzień).

Oddziaływanie i wynik

Teraz, gdy masz już dostępne dane, nadszedł czas, aby zapoznać się z pierwszymi aspektami technicznymi. Niech T_i będzie oddziaływaniem dotyczącym jednostki i:

$$T_i = \begin{cases} 1, \text{ jeśli jednostka } i \text{ otrzymała oddziaływanie} \\ 0 \text{ w przeciwnym razie} \end{cases}$$

Oddziaływaniem nie musi być lek lub jakikolwiek inny zabieg medyczny (w języku angielskim używane jest pojęcie *treatment*, czyli dosłownie *leczenie*). Jest to po prostu termin, którego używam do określenia działań, których efekt chcę poznać. W tym przykładzie oddziaływaniem jest obniżenie cen w jednej z firm na platformie internetowej. Taka obniżka jest reprezentowana w kolumnie `is_on_sale`.

Notacja dotycząca oddziaływania

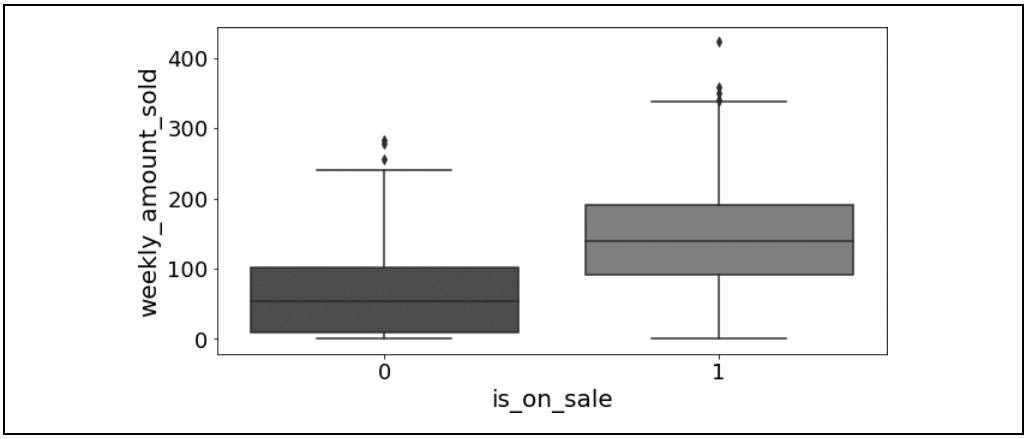
W niektórych tekstach i w dalszej części tej książki czasami zetkniesz się z oznaczaniem oddziaływania literą D zamiast T. Litera D pozwala uniknąć niejasności, gdy w problemach przyczynowych występuje wymiar czasowy.

Ponadto wartości z kolumny `weekly_amount_sold` (jest to zmienna, na którą chcę mieć wpływ) będę nazywał *wynikiem*. Wynik dla jednostki i oznaczam jako Y_i. Dzięki tym dwóm nowym koncepcjom mogę ponownie określić cel wnioskowania przyczynowego jako proces poznawania wpływu T na Y. W omawianym przykładzie oznacza to ustalenie wpływu wartości zmiennej `is_on_sale` na wartość zmiennej `weekly_amount_sold`.

Podstawowy problem z wnioskowaniem przyczynowym

Teraz sytuacja staje się interesująca. *Podstawowym problemem z wnioskowaniem przyczynowym* jest to, że nigdy nie można zaobserwować tej samej jednostki z oddziaływaniem i bez oddziaływania.

To tak, jakby istniały dwie rozchodzące się drogi i można było ustalić tylko to, dokąd prowadzi jedna z nich. Aby w pełni docenić znaczenie tej kwestii, warto wrócić do omawianego przykładu i utworzyć wykres wyników w zależności od oddziaływania, czyli wykres wartości zmiennej `weekly_amount_sold` w zależności od zmiennej `is_on_sale`. Od razu widać, że sklepy, które obniżyły ceny, sprzedają znacznie więcej (zobacz rysunek 1.1).



Rysunek 1.1. Wielkość sprzedaży tygodniowej w trakcie wyprzedaży (1) i bez wyprzedaży (0)

Jest to również zgodne z intuicyjną wiedzą — ludzie kupują więcej, gdy ceny są niskie, a wyprzedaż (zazwyczaj) oznacza niższe ceny. To bardzo dobrze, ponieważ wnioskowanie przyczynowe jest zgodne z wiedzą ekspercką. Należy jednak zachować ostrożność. Prawdopodobne jest, że udzielanie rabatów i ich reklamowanie sprawi, iż klienci będą kupować więcej. Ale o ile więcej? Z wykresu z rysunku 1.1 wynika, że w trakcie wyprzedaży liczba sprzedanych sztuk jest średnio o około 150 wyższa. Jest to podejrzanie duża różnica, ponieważ zakres liczby sprzedanych sztuk w tygodniach bez wyprzedaży wynosi mniej więcej od 0 do 50. Jeśli się zastanowisz, możesz dostrzec, że możesz mylić asocjację z przyczynowością. Może jest tak, że tylko większe firmy, które i tak sprzedają najwięcej, mogą sobie pozwolić na agresywne obniżki cen? Może firmy urządzają wyprzedaże bliżej świąt Bożego Narodzenia, a wtedy klienci i tak kupują najwięcej?

Chodzi o to, że pewność co do rzeczywistego wpływu obniżek na liczbę sprzedanych sztuk można uzyskać tylko wtedy, gdy ta sama firma (jednostka) jest obserwowana w tym samym czasie w dwóch scenariuszach — z wyprzedażą i bez niej. Tylko porównując te dwie sprzeczne sytuacje można mieć pewność co do efektu obniżek cen. Jednakże, jak wspomniałem wcześniej, podstawowym problemem z wnioskowaniem przyczynowym jest to, że opisane rozwiązanie jest niewykonalne. Trzeba więc wymyślić inne podejście.

Modele przyczynowe

Wszystkie te problemy można zrozumieć intuicyjnie, ale jeśli chcesz wyjść poza prostą intuicję, potrzebna będzie formalna notacja. Będzie to codzienny język do opisywania przyczynowości. Potraktuj go jako wspólny język, którym będziesz się posługiwać razem z innymi adeptami sztuki wnioskowania przyczynowego.

Model przyczynowy jest serią mechanizmów przypisania oznaczanych przez symbol ←. W tych mechanizmach będę używał litery u do oznaczania zmiennych spoza modelu, co określa, że nie opisuję w żaden sposób ich generowania. Wszystkie pozostałe zmienne są istotne, dlatego występują w modelu. Istnieją także funkcje f, które odwzorowują jedną zmienną na inną. Przyjrzyj się następującemu przykładowemu modelowi przyczynowemu:

$$T \leftarrow f_t(u_t)$$

$$Y \leftarrow f_y(T, u_y)$$

Pierwsze równanie oznacza, że u_t, zestaw zmiennych, których nie uwzględniam w modelu (zwany również zmiennymi egzogenicznymi), odpowiada za oddziaływanie T za pomocą funkcji f_t. Dalej T wraz z innym zestawem zmiennych u_y (których również nie uwzględniam w modelu) wspólnie powoduje wynik Y na podstawie funkcji f_y. Zmienne u_y występują w równaniu, aby poinformować, że wynik nie jest określony przez samo oddziaływanie. Niektóre inne zmienne również mają na niego wpływ, nawet jeśli decyduję się ich nie uwzględniać w modelu. W przykładzie ze sprzedażą oznaczałoby to, że wartość zmiennej `weekly_amount_sold` zależy od oddziaływania (zmiennej `is_on_sale`) i innych czynników, które nie zostały określone, a są reprezentowane jako u. Celem dodania u jest ujęcie całej zmienności spowodowanej przez czynniki, które nie są uwzględnione w zmiennych zawartych w modelu (zmiennych endogenicznych). W omawianym przykładzie mogę powiedzieć, że obniżki cen są spowodowane czynnikami (może to być wielkość firmy, ale też coś innego), które nie występują w modelu:

$$\begin{aligned} Wyprzedaż &\leftarrow f_t(u_t) \\ \text{PoziomSprzedaży} &\leftarrow f_y(\text{Wyprzedaż}, u_y) \end{aligned}$$

Używam symbolu ← zamiast =, aby bezpośrednio podkreślić nieodwracalność przyczynowości. Gdy używany jest znak równości, wyrażenie $Y = T + X$ jest równoważne wyrażeniu $T = Y - X$, ale nie chcę przecież stwierdzić, że powodowanie Y przez T jest równoważne powodowaniu T przez Y. Mimo to często unikam używania symbolu ←, ale robię to tylko dlatego, że jest nieco kłopotliwy. Należy jednak pamiętać, że ze względu na nieodwracalność przyczyn i skutków w modelach przyczynowych (inaczej niż w tradycyjnej algebrze) nie można dowolnie umieszczać wyrażeń wokół znaku równości.

Jeśli chcesz bezpośrednio uwzględnić w modelu więcej zmiennych, możesz usunąć je z u i uwzględnić w modelu. Pamiętasz na przykład, jak stwierdziłem, że duża różnica w poziomie sprzedaży między tygodniami z obniżkami cen a tygodniami bez obniżek może wynikać z tego, iż większe firmy przeprowadzają bardziej agresywne wyprzedaże? W poprzednim modelu *WielkośćFirmy* nie jest bezpośrednio uwzględniona w modelu. Zamiast tego jej wpływ zostaje zignorowany wraz ze wszystkimi innymi zmiennymi z u. Ale mógłbym ją bezpośrednio uwzględnić w modelu:

$$WielkośćFirmy \leftarrow f_s(u_s)$$

$$Wyprzedaż \leftarrow f_t(WielkośćFirmy, u_t)$$

$$PoziomSprzedaży \leftarrow f_y(Wyprzedaż, WielkośćFirmy, u_y)$$

Aby uwzględnić tę kolejną zmienną endogeniczną, najpierw dodaję kolejne równanie, aby wyjaśnić, w jaki sposób powstaje ta zmienna. Następnie usuwam zmienną *WielkośćFirmy* z u_t. Oznacza to, że nie traktuję jej już jako zmiennej spoza modelu. Bezpośrednio stwierdzam, że *WielkośćFirmy*

powoduje *Wyprzedaż* (wraz z kilkoma innymi czynnikami zewnętrznymi, których nadal nie chcę uwzględniać w modelu). Jest to tylko formalny sposób zapisu przekonania, że większe firmy są bardziej skłonne do obniżania cen. Na koniec mogę również dodać *WielkośćFirmy* do ostatniego równania. W ten sposób zapisuję przekonanie, że większe firmy również sprzedają więcej. Innymi słowy, *WielkośćFirmy* jest wspólną przyczyną zarówno oddziaływania *Wyprzedaż*, jak i wyniku *PoziomSprzedaży*.

Ponieważ ten sposób modelowania prawdopodobnie jest dla Ciebie czymś nowym, warto powiązać go z jakimś znanym podejściem. Jeśli znasz się na ekonomii lub statystyce, możesz być przyzwyczajony do innego sposobu modelowania tego samego problemu:

$$PoziomSprzedaży_i = \alpha + \beta_1 Wyprzedaż_i + \beta_2 WielkośćFirmy_i + e_i$$

Na pozór ten wzór wygląda zupełnie inaczej, ale bliższa analiza pokazuje, że ten model jest bardzo podobny do tego, który przedstawiłem wcześniej. Po pierwsze, zauważ, że po prostu zastąpiłem ostatnie równanie z poprzedniego modelu i rozwinąłem funkcję f_y przez bezpośrednie określenie, iż zmienne endogeniczne *Wyprzedaż* i *WielkośćFirmy* są liniowo i addytywnie łączone w celu uzyskania wyniku *PoziomSprzedaży*. W tym sensie ten *model liniowy* obejmuje więcej założeń niż wcześniej przedstawiony zapis. Można powiedzieć, że nowy model narzuca funkcyjną zależność między zmiennymi. Po drugie, nowy model nie określa w żaden sposób, jak powstają niezależne (endogeniczne) zmienne *Wyprzedaż* i *WielkośćFirmy*. Po trzecie, w tym modelu używany jest znak równości zamiast operatora przypisania, ale już wyjaśniłem, że nie należy przykładać do tego nadmiernej wagi.

Interwencje

Powodem, dla którego poświęcam czas na omówienie modeli przyczynowych, jest to, że po przygotowaniu takiego modelu można zacząć manipulować zmiennymi z nadzieją uzyskania odpowiedzi na pytanie przyczynowe. Formalnym terminem na to jest *interwencja*. Na przykład można wziąć bardzo prosty model przyczynowy i poddać wszystkich oddziaływaniu t_0. Wyeliminuje to naturalne przyczyny T dzięki zastąpieniu ich pojedynczą stałą:

$$T \leftarrow t_0$$

$$Y \leftarrow f_y(T, u_y)$$

To wnioskowanie ma postać eksperymentu myślowego, w którym szukana jest odpowiedź na następujące pytanie: „Co stałoby się z wynikiem Y, gdyby zastosować interwencję t_0?”. W rzeczywistości nie musisz przeprowadzać interwencji (chociaż możesz i będziesz to robić, ale później). W literaturze dotyczącej wnioskowania przyczynowego takie interwencje są opisywane za pomocą operatora $do(.)$. Jeśli chcesz się dowiedzieć, co by się stało w wyniku interwencji w T, możesz zapisać to tak: $do(T = t_0)$.

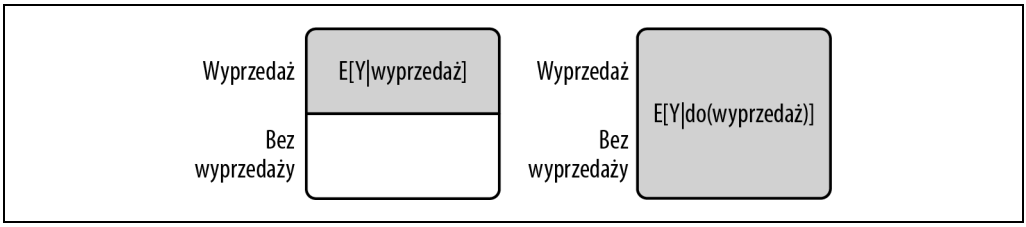
Wartości oczekiwane

Od teraz będę często używał wartości oczekiwanej i warunkowej wartości oczekiwanej. Można traktować wartość oczekiwaną jak dotyczącą populacji wartość, której średnią próbujesz oszacować. Zapis $E[X]$ oznacza wartości oczekiwane zmiennej losowej X. Przybliżenie można ustalić za pomocą średniej wartości X dla próbki. Zapis $E[Y|X = x]$ oznacza wartość oczekiwaną Y dla $X = x$. Przybliżenie można otrzymać za pomocą średniej wartości Y dla $X = x$.

Operator *do*(.) daje również pierwsze wyobrażenie o tym, dlaczego asocjacja różni się od związku przyczynowego. Argumentowałem już, dlaczego wysoki poziom sprzedaży dla firmy urządzającej wyprzedaż, $E[PoziomSprzedaży|Wyprzedaż = 1]$, może prowadzić do przeszacowania średniego poziomu sprzedaży, jaki miałaby firma, gdyby dokonała obniżki ceny, $E[PoziomSprzedaży, do(Wyprzedaż = 1)]$. W pierwszym scenariuszu mamy do czynienia z firmami, które zdecydowały się na obniżkę cen, czyli prawdopodobnie większymi przedsiębiorstwami. Z kolei druga wartość, $E[PoziomSprzedaży, do(Wyprzedaż = 1)]$, określa, co by się stało, gdyby wszystkie firmy (nie tylko te duże) zorganizowały wyprzedaż. Co ważne, na ogólnym poziomie mamy:

$$E[PoziomSprzedaży|Wyprzedaż = 1] \neq E[PoziomSprzedaży, do(Wyprzedaż = 1)]$$

Jednym ze sposobów myślenia o różnicy między tymi wyrażeniami jest uwzględnienie doboru i interwencja. Kiedy warunek dotyczy wyprzedaży, mierzysz poziom sprzedaży w wybranej podpróbie firm, które faktycznie obniżyły ceny. Kiedy warunek dotyczy interwencji *do*(*Wyprzedaż*), każda firma musi obniżyć ceny, a następnie należy zmierzyć poziom sprzedaży w całej próbie (patrz rysunek 1.2).



Rysunek 1.2. Dobór polega na filtrowaniu próby na podstawie oddziaływania. Interwencja wymusza oddziaływanie na całej próbie

Operator *do*(.) służy do definiowania wartości przyczynowych, które nie zawsze są możliwe do otrzymania na podstawie obserwowanych danych. W omawianym przykładzie nie można zaobserwować interwencji *do*(*Wyprzedaż* = 1) dla każdej firmy, ponieważ nie ma wymogu urządzania wyprzedaży. Operator *do*(.) jest najbardziej przydatny jako teoretyczna koncepcja, której można użyć do bezpośredniego określenia badanej wartości przyczynowej. Ponieważ taka wartość nie jest bezpośrednio obserwowalna, wnioskowanie przyczynowe często wymaga wyeliminowania jej z teoretycznego zapisu. Proces ten nazywa się *identyfikacją*.

Indywidualny efekt oddziaływania

Operator *do*(.) pozwala również zapisać *indywidualny efekt oddziaływania*, czyli wpływ oddziaływania na wynik dla konkretnej jednostki *i*. Można to zapisać jako różnicę między dwiema interwencjami:

$$\tau_i = Y_i|do(T = t_1) - Y_i|do(T = t_0)$$

Słownie można odczytać to tak: „efekt, τ_i, zmiany oddziaływania z t_0 na t_1 dla jednostki *i* jest różnicą w wyniku dla tej jednostki przy t_1 w porównaniu z wynikiem przy t_0". Można to wykorzystać do rozwiązania problemu polegającego na ustaleniu wpływu zmiany wartości *Wyprzedaż* z 0 na 1 na wartość *PoziomSprzedaży*:

$$\tau_i = PoziomSprzedaży_i|do(Wyprzedaż = 1) - PoziomSprzedaży_i|do(Wyprzedaż = 0)$$

Ze względu na fundamentalny problem z wnioskowaniem przyczynowym można zaobserwować tylko jeden człon przedstawionego równania. Tak więc choć można teoretycznie zapisać szukaną wartość, niekoniecznie oznacza to, że można ją otrzymać na podstawie danych.

Wyniki potencjalne

Inną rzeczą, jaką można zdefiniować za pomocą operatora *do*(.), jest być może najciekawsza i najczęściej stosowana koncepcja z obszaru wnioskowania przyczynowego — *wyniki potencjalne* (kontrfaktyczne):

$$Y_{ti} = Y_i|do(T_i = t)$$

Należy to czytać tak: „wynik jednostki *i* byłby równy *Y*, gdyby zastosowano oddziaływanie *t*". Czasami używam notacji funkcyjnej do definiowania wyników potencjalnych, ponieważ w indeksach dolnych może szybko pojawić się za dużo elementów:

$$Y_{ti} = Y(t)_i$$

W kontekście oddziaływań binarnych (oddziaływanie lub jego brak) stosuję oznaczenie Y_{0i} jako wynik potencjalny dla jednostki *i* bez oddziaływania i Y_{1i} jako wynik potencjalny dla *tej samej* jednostki *i* z oddziaływaniem. Będę również nazywać jeden wynik potencjalny faktycznym, co oznacza, że można go zaobserwować, a drugi kontrfaktycznym, co oznacza, że nie można go zaobserwować. Na przykład jeśli jednostka *i* jest poddana oddziaływaniu, mogę zobaczyć, co się z nią dzieje pod jego wpływem. To oznacza, że widzę Y_{1i}, czyli faktyczny wynik potencjalny. Nie mogę jednak zaobserwować, co by się stało, gdyby jednostka *i* nie była poddana oddziaływaniu. Oznacza to, że nie mogę zobaczyć Y_{0i}, ponieważ jest to wynik kontrfaktyczny:

$$Y_i = \begin{cases} Y_{1i}, \text{ jeśli jednostka } i \text{ otrzymała oddziaływanie} \\ Y_{0i} \text{ w przeciwnym razie} \end{cases}$$

Tę samą sytuację można zapisać także w następujący sposób:

$$Y_i = T_i Y_{1i} + (1 - T_i) Y_{0i} = Y_{0i} + (Y_{1i} - Y_{0i}) T_i$$

Wróćmy teraz do przykładu. Zapis $PoziomSprzedaży_{0i}$ oznacza wartość sprzedaży w firmie *i* bez obniżki cen, a zapis $PoziomSprzedaży_{1i}$ oznacza wartość sprzedaży po urządzeniu wyprzedaży. Można również zdefiniować efekt w kategoriach tych potencjalnych wyników:

$$\tau_i = Y_{1i} - Y_{0i}$$

Założenia

W tej książce wnioskowaniu przyczynowemu zawsze towarzyszą założenia. Założenia to stwierdzenia, które wyrażają przekonania związane ze sposobem wygenerowania danych. Problem polega na tym, że zwykle takich twierdzeń nie można zweryfikować na podstawie danych. Dlatego trzeba przyjąć założenia. Założenia nie zawsze są łatwe do dostrzeżenia, dlatego dokładam wszelkich starań, aby były one przejrzyste.

Spójność i stabilna wartość oddziaływania jednostkowego

W przedstawionych wcześniej równaniach występują dwa ukryte założenia. Pierwsze z nich dotyczy tego, że wynik potencjalny jest spójny z oddziaływaniem: $Y_i(t) = Y$, gdy $T_i = t$. Innymi słowy, nie ma ukrytych wielu wersji oddziaływania poza tymi określonymi za pomocą T. To założenie może zostać naruszone, jeśli istnieje wiele poziomów oddziaływania, ale uwzględniasz tylko dwa z nich. Na przykład jeśli interesuje Cię wpływ kuponów rabatowych na sprzedaż i traktujesz problem jako binarny (klienci albo otrzymali kupon, albo nie), ale w rzeczywistości wypróbowałeś rabaty o różnej wartości. Niespójność może wystąpić również w sytuacji, gdy oddziaływanie jest źle zdefiniowane. Wyobraźmy sobie na przykład, że próbujemy określić wpływ otrzymania pomocy od doradcy finansowego na finanse. Co oznacza tutaj „pomoc"? Czy jest to jednorazowa konsultacja? A może regularne doradztwo i śledzenie realizacji celów? Połączenie wszystkich tych rodzajów doradztwa finansowego w jedną kategorię również narusza założenie spójności.

Drugim założeniem jest brak interferencji, czyli stabilna wartość oddziaływania jednostkowego (ang. *Stable Unit Treatment Value Assumption* — SUTVA). Oznacza to, że na efekt uzyskany dla jednej jednostki nie ma wpływu oddziaływanie zastosowane do innych jednostek: $Y_i(T_i) = Y_i(T_1, T_2, \ldots, T_i, \ldots, T_n)$. Założenie to może zostać naruszone, jeśli występują efekty uboczne lub sieciowe. Na przykład jeśli chcemy poznać wpływ szczepionek na zapobieganie chorobom zakaźnym — zaszczepiony z mniejszym prawdopodobieństwem będzie przenosił chorobę, co zmniejsza ogólne ryzyko zachorowania. Naruszenie tego założenia zwykle powoduje, że myślimy, iż efekt jest mniejszy niż w rzeczywistości. Jeśli chodzi o efekty uboczne, oddziaływanie wpływa w jakimś stopniu na jednostki z grupy kontrolnej, co z kolei powoduje, że grupy kontrolna i poddana oddziaływaniu różnią się w mniejszym stopniu niż w przypadku braku interferencji.

Naruszenia

Na szczęście często można poradzić sobie z naruszeniami obu założeń. Aby wyeliminować naruszenie spójności, należy uwzględnić w analizie wszystkie wersje oddziaływań. Aby poradzić sobie z efektami ubocznymi, można rozszerzyć definicję efektu oddziaływań, by uwzględnić efekt pochodzący od innych jednostek, a także posłużyć się bardziej elastycznymi modelami do szacowania efektów.

Interesujące wartości przyczynowe

Po omówieniu pojęcia wyniku potencjalnego można ponownie sformułować podstawowy problem z wnioskowaniem przyczynowym: *nigdy nie można poznać indywidualnego efektu oddziaływań, ponieważ obserwowany jest tylko jeden z wyników potencjalnych.* Ale nie wszystko jest stracone. Dzięki omówionym nowym koncepcjom możesz poczynić pewne postępy w pracy nad tym fundamentalnym problemem. Nawet jeśli nigdy nie poznasz indywidualnych efektów τ_i, istnieją inne interesujące wielkości przyczynowe, które można ustalić na podstawie danych. Na przykład zdefiniujmy *średni efekt oddziaływania* (ang. *average treatment effect* — ATE) w następujący sposób:

$$ATE = E[\tau_i]$$

lub

$$ATE = E[Y_{1i} - Y_{0i}]$$

lub nawet

$$ATE = E[Y|do(T = 1)] - E[Y|do(T = 0)]$$

Średni efekt oddziaływania reprezentuje średni wpływ oddziaływania *T*. Dla niektórych jednostek ten wpływ będzie większy, dla innych mniejszy, przy czym niemożliwe jest ustalenie indywidualnego wpływu na poszczególne jednostki. Ponadto jeśli chcesz oszacować średni efekt oddziaływania na podstawie danych, możesz zastąpić wartości oczekiwane średnimi z próby:

$$\frac{1}{N}\Sigma_{i=0}^{N}\tau_i$$

lub

$$\frac{1}{N}\sum_{i=0}^{N}(Y_{1i} - Y_{0i})$$

Oczywiście w praktyce, ze względu na fundamentalny problem z wnioskowaniem przyczynowym, nie można tego zrobić, ponieważ dla każdej jednostki można zaobserwować tylko jeden potencjalny wynik. Na razie nie martw się zbytnio o to, jak oszacować szukaną wartość. Wkrótce się dowiesz, jak to zrobić. Skup się na zrozumieniu, jak zdefiniować wartość przyczynową w kategoriach potencjalnych wyników i dlaczego warto je oszacować.

Innym interesującym efektem grupowym jest *średni efekt oddziaływania na poddanych jego wpływowi* (ang. *average treatment effect on the treated* — ATT):

$$ATT = E[Y_{1i} - Y_{0i}|T = 1]$$

Jest to wpływ oddziaływania na jednostki, które zostały mu poddane. Na przykład jeśli przeprowadziłeś internetową kampanię marketingową w mieście i chcesz ustalić, ilu dodatkowych klientów przyniosła ona w tej miejscowości, używany będzie właśnie ATT, czyli wpływ marketingu na mieszkańców miasta, w którym kampania została zrealizowana. W tym scenariuszu należy zauważyć, że dla tego samego oddziaływania zdefiniowane są oba potencjalne wyniki. W ATT warunek dotyczy oddziaływania, dlatego Y_{0i} nigdy nie jest obserwowalne, ale mimo to dobrze zdefiniowane.

Istnieją też *warunkowe średnie efekty oddziaływania* (ang. *conditional average treatment effect* — CATE):

$$CATE = E[Y_{1i} - Y_{0i}|X = x]$$

Jest to efekt w grupach zdefiniowanych przez zmienne X. Na przykład możesz chcieć poznać wpływ wiadomości e-mail na klientów w wieku powyżej 45 lat i na młodsze osoby. Warunkowy średni efekt oddziaływania jest nieoceniony w kontekście personalizacji, ponieważ pozwala się dowiedzieć, który typ jednostek lepiej reaguje na interwencję.

Można również zdefiniować te wartości dla oddziaływania ciągłego. Wtedy różnicę należy zastąpić pochodną cząstkową:

$$\frac{\partial}{\partial t} E[Y_i]$$

Może się to wydawać skomplikowane, ale jest to po prostu sposób na zapisanie oczekiwanej wielkości zmiany $E[Y_i]$ przy niewielkim wzroście oddziaływania.

Wartości przyczynowe — przykład

Zobaczmy, jak można zdefiniować te wartości w przykładowym problemie biznesowym. Po pierwsze, zauważ, że nigdy nie możesz poznać wpływu obniżek cen na sprzedaż w pojedynczej firmie, ponieważ wymagałoby to ustalenia obu potencjalnych wyników, $PoziomSprzedaży_{0i}$ i $PoziomSprzedaży_{1i}$, z tego samego czasu. Zamiast tego możesz skupić swoją uwagę na czymś, co można oszacować, na przykład na średnim wpływie obniżek cen na poziom sprzedaży:

$$ATE = E[PoziomSprzedaży_{1i} - PoziomSprzedaży_{0i}]$$

lub na tym, jak firmy decydujące się na obniżki cen zwiększyły sprzedaż:

$$ATT = E[PoziomSprzedaży_{1i} - PoziomSprzedaży_{0i}|Wyprzedaż = 1],$$

lub na wpływie urządzania wyprzedaży w tygodniu ze świętami Bożego Narodzenia:

$$CATE = E[PoziomSprzedaży_{1i} - PoziomSprzedaży_{0i}|tygodnieDoŚwiąt = 0].$$

Wiem, że nie można poznać obu potencjalnych wyników, ale dla dobra dyskusji i aby opis był bardziej konkretny, załóżmy, iż jest to możliwe. Przyjmij przez chwilę, że bogowie wnioskowania przyczynowego zostali obłaskawieni wieloma bitwami statystycznymi, które stoczyłeś, i nagrodzili Cię boskimi mocami pozwalającymi zobaczyć alternatywne wszechświaty, w których pojawiają się wszystkie wyniki. Załóżmy, że dzięki tym mocom udało Ci się zebrać dane na temat trzech firm, z których trzy urządziły wyprzedaże, a trzy nie.

W tabeli 1.2 *i* jest identyfikatorem jednostki, *y* jest zaobserwowanym wynikiem, $_{y0}$ i $_{y1}$ są potencjalnymi wynikami z oddziaływaniem i bez oddziaływania, *t* jest indykatorem oddziaływania, a *x* jest zmienną towarzyszącą, która oznacza czas do Bożego Narodzenia. Należy pamiętać, że organizacja wyprzedaży jest oddziaływaniem, a poziom sprzedaży to wynik. Załóżmy ponadto, że dla dwóch z tych firm dane dotyczą tygodnia poprzedzającego tydzień z Bożym Narodzeniem, co jest oznaczone w tabeli jako $x = 1$. Pozostałe obserwacje dotyczą tygodnia z Bożym Narodzeniem.

Tabela 1.2. Dane zebrane dzięki boskiej mocy

	i	y0	y1	t	x	y	te
0	1	200	220	0	0	200	20
1	2	120	140	0	0	120	20
2	3	300	400	0	1	300	100
3	4	450	500	1	0	500	50
4	5	600	600	1	0	600	0
5	6	600	800	1	1	800	200

Dzięki swoim boskim mocom możesz poznać zarówno $PoziomSprzedaży_0$, jak i $PoziomSprzedaży_1$. Dzięki temu obliczanie wszystkich wartości przyczynowych, o których pisałem wcześniej, jest niezwykle łatwe. Na przykład ATE będzie średnią z ostatniej kolumny, czyli efektem oddziaływania:

$$ATE = (20 + 20 + 100 + 50 + 0 + 200)/6 = 65$$

Oznacza to, że wyprzedaż zwiększa poziom sprzedaży średnio o 65 jednostek. Jeśli chodzi o ATT, wystarczy obliczyć średnią z ostatniej kolumny dla $T = 1$:

$$ATT = (50 + 0 + 200)/3 = 83{,}33$$

Innymi słowy, w firmach, które zdecydowały się na wyprzedaż (zostały poddane oddziaływaniu), obniżone ceny zwiększyły poziom sprzedaży średnio o 83,33 jednostki. Wreszcie średni efekt pod warunkiem, że uwzględniany jest tydzień przed tygodniem z Bożym Narodzeniem ($x = 1$), jest po prostu średnią efektu dla jednostek 3 i 6:

$$CATE(x = 1) = (100 + 200)/2 = 150$$

Z kolei średni efekt w tygodniu świątecznym jest średnim efektem oddziaływania dla $x = 0$:

$$CATE(x = 0) = (20 + 20 + 50 + 0)/4 = 22{,}5$$

Oznacza to, że firmy skorzystały na obniżkach cen znacznie bardziej w tygodniu poprzedzającym tydzień z Bożym Narodzeniem (wzrost o 150 jednostek) w porównaniu z wyprzedażami w tygodniu z Bożym Narodzeniem (wzrost o 22,5 jednostki). Tak więc sklepy, które urządziły wyprzedaż wcześniej, skorzystały na tym bardziej niż te, które zrobiły to później.

Teraz, gdy lepiej rozumiesz wartości przyczynowe, które zwykle będą Cię interesować (ATE, ATT i CATE), nadszedł czas, aby opuścić Wyspę Fantazji i wrócić do prawdziwego świata. W tym świecie życie jest brutalne, a dostępne dane są znacznie trudniejsze do wykorzystania. Możesz poznać tylko jeden wynik potencjalny, przez co nie da się ustalić indywidualnego efektu oddziaływania (zobacz tabelę 1.3).

Problem brakujących danych

O wnioskowaniu przyczynowym można myśleć jak o rozwiązywaniu problemu brakujących danych. Aby wywnioskować interesujące wartości przyczynowe, należy obliczyć brakujące potencjalne wyniki.

Tabela 1.3. Dane dostępne w zwykłym świecie

	i	y0	y1	t	x	y	te
0	1	200,0	NaN	0	0	200	NaN
1	2	120,0	NaN	0	0	120	NaN
2	3	300,0	NaN	0	1	300	NaN
3	4	NaN	500,0	1	0	500	NaN
4	5	NaN	600,0	1	0	600	NaN
5	6	NaN	800,0	1	1	800	NaN

Można spojrzeć na te dane i pomyśleć tak: „Z pewnością nie jest to idealne rozwiązanie, ale czy nie mogę po prostu obliczyć średniej dla jednostek poddanych oddziaływaniu i porównać jej ze średnią dla jednostek, które nie zostały poddane oddziaływaniu? Innymi słowy, czy nie mogę po prostu wykonać obliczeń wartości ATE = 500 + 600 + 800/3 – 200 + 120 + 300/3 = 426,67?”. Nie! Właśnie popełniłeś najcięższy grzech polegający na myleniu asocjacji z przyczynowością!

Zauważ, jak różne są otrzymane wyniki. Obliczona wcześniej wartość ATE wynosiła mniej niż 100, a teraz okazuje się, że jest wyższa niż 400. Problem polega na tym, że firmy, które organizują wyprzedaż, różnią się od tych, które tego nie robią. Firmy, które urządziły wyprzedaż, prawdopodobnie sprzedałyby więcej od pozostałych nawet bez obniżki cen. Aby się o tym przekonać, wystarczy cofnąć się do momentu, w którym można było zobaczyć oba potencjalne wyniki. Wtedy Y_0 dla jednostek poddanych oddziaływaniu jest znacznie wyższe niż dla jednostek niepoddanych oddziaływaniu. Ta różnica w Y_0 między grupami znacznie utrudnia ustalenie efektu oddziaływania na podstawie zwykłego porównania obu grup.

Chociaż porównywanie średnich nie jest najlepszym pomysłem, intuicja kieruje Cię we właściwą stronę. Nadszedł czas, aby wykorzystać nowe koncepcje, które właśnie omówiłem. Pozwoli Ci to lepiej rozwinąć intuicję i wreszcie zrozumieć, dlaczego asocjacja nie oznacza przyczynowości. Czas stawić czoła głównemu wrogowi wnioskowania przyczynowego.

Błąd systematyczny

Przejdę od razu do sedna — *to błąd systematyczny jest tym, co odróżnia asocjację od przyczynowości*. Cały problem polega na tym, że to, co szacujesz na podstawie danych, niekoniecznie pasuje do wartości przyczynowych, które chcesz obliczyć. Na szczęście można to łatwo intuicyjnie zrozumieć. Przypomnę badany przykład biznesowy. Gdy usłyszysz twierdzenie, że obniżenie cen zwiększa ilość sprzedanych przez firmę produktów, możesz je zakwestionować, ponieważ możliwe jest, że firmy urządzające wyprzedaż prawdopodobnie i tak sprzedałyby więcej od pozostałych nawet bez obniżek cen. Być może wynika to z tego, że są większe i mogą sobie pozwolić na bardziej agresywną sprzedaż. Innymi słowy, firmy poddane oddziaływaniu (firmy organizujące wyprzedaż) nie są porównywalne z firmami niepoddanymi oddziaływaniu (nieobniżającymi cen).

Aby przedstawić bardziej formalną argumentację, można przekształcić to intuicyjne rozumowanie za pomocą notacji potencjalnego wyniku. Po pierwsze, aby oszacować ATE, należy ustalić, co stałoby się z jednostkami poddanymi oddziaływaniu, gdyby pominąć oddziaływanie (czyli $E[Y_0|T = 1]$),

a także co stałoby się z jednostkami niepoddanymi oddziaływaniu, gdyby zastosować do nich oddziaływanie (czyli $E[Y_1|T = 0]$). Gdy porównujesz średnie wyniki dla grup poddanych i niepoddanych oddziaływaniu, używasz $E[Y|T = 0]$ do oszacowania $E[Y_0]$ i $E[Y|T = 1]$ do oszacowania $E[Y_1]$. Innymi słowy, szacujesz $E[Y|T = t]$ z nadzieją na otrzymanie $E[Y_t]$. Jeśli te wartości nie są zgodne, estymator, który ma przybliżać $E[Y|T = t]$, na przykład średni wynik dla jednostek poddanych oddziaływaniu t, będzie obciążonym błędem systematycznym estymatorem $E[Y_t]$.

Definicja techniczna

Można powiedzieć, że estymator jest obciążony błędem systematycznym, jeśli różni się od parametru, którego ma być oszacowaniem. *Błąd systematyczny* = $E[\hat{\beta} - \beta]$, gdzie $\hat{\beta}$ jest estymatą, a β szacowaną wartością (estymandą). Na przykład estymator średniego efektu oddziaływania jest obciążony błędem systematycznym, jeśli systematycznie zaniża lub zawyża prawdziwą wartość ATE.

Wróćmy do intuicji. Możesz wykorzystać swoje zrozumienie tego, jak działa świat, aby zrobić kolejny krok. Można powiedzieć, że Y_0 firmy poddanej oddziaływaniu jest prawdopodobnie większe niż Y_0 firmy niepoddanej oddziaływaniu. Dzieje się tak, ponieważ firmy, które mogą sobie pozwolić na obniżki cen, sprzedają więcej niezależnie od wyprzedaży. Pozwól, aby te informacje zakorzeniły się w Tobie. Potrzeba trochę czasu, aby przyzwyczaić się do mówienia o potencjalnych wynikach, ponieważ wiąże się to z rozumowaniem o rzeczach, które mogłyby się wydarzyć, ale tak się nie stało. Przeczytaj ten akapit jeszcze raz i upewnij się, że go rozumiesz.

Wzór na błąd systematyczny

Teraz, gdy już rozumiesz, dlaczego średnia z próby może różnić się od średniego potencjalnego wyniku, którego ta średnia ma być oszacowaniem, przyjrzyjmy się bliżej, dlaczego różnice w średnich zwykle nie pozwalają ustalić prawdziwego ATE. Ten podrozdział jest dość techniczny, więc jeśli nie interesują Cię równania matematyczne, możesz przejść do następnego.

W przykładzie dotyczącym wyprzedaży związek między oddziaływaniem a wynikiem jest mierzony na podstawie równania $E[Y|T = 1] - E[Y|T = 0]$. Jest to średni poziom sprzedaży firm, które zorganizowały wyprzedaż, minus średni poziom sprzedaży w firmach, które nie obniżyły cen. Z kolei związek przyczynowy jest mierzony na podstawie równania $E[Y_1 - Y_0]$ (co jest skróconym zapisem wzoru $E[Y|do(t = 1)] - E[Y|do(t = 0)]$).

Aby zrozumieć, dlaczego i jak te równania różnią się między sobą, zastąpmy zaobserwowane wyniki potencjalnymi wynikami w równaniu asocjacji $E[Y|T = 1] - E[Y|T = 0]$. Dla firm poddanych oddziaływaniu zaobserwowany wynik to Y_1, a dla jednostek niepoddanych oddziaływaniu wynik to Y_0:

$$E[Y|T = 1] - E[Y|T = 0] = E[Y_1|T = 1] - E[Y_0|T = 0]$$

Teraz dodajmy i odejmijmy $E[Y_0|T = 1]$, czyli wynik kontrfaktyczny określający, co stałoby się z wynikiem firm poddanych oddziaływaniu bez tego oddziaływania:

$$E[Y|T = 1] - E[Y|T = 0] = E[Y_1|T = 1] - E[Y_0|T = 0] + E[Y_0|T = 1] - E[Y_0|T = 1]$$

Na koniec można zmienić kolejność wyrazów i uwzględnić wartości oczekiwane:

$$E[Y \mid T=1] - E[Y \mid T=0] = \underbrace{E[Y_1 - Y_0 \mid T=1]}_{ATT} + \underbrace{\{E[Y_0 \mid T=1] - E[Y_0 \mid T=0]\}}_{BŁĄD\ SYSTEMATYCZNY}$$

Te proste obliczenia matematyczne ilustrują wszystkie problemy, które napotkasz w trakcie szukania odpowiedzi na pytania przyczynowe. Aby pomóc w lepszym zrozumieniu przedstawionych równań, przeanalizuję, co z nich wynika. Po pierwsze, równanie to pokazuje, dlaczego asocjacja nie jest związkiem przyczynowym. Jak widać, asocjacja jest równa efektowi oddziaływania na poddane mu jednostki plus błąd systematyczny. *Błąd systematyczny zależy od tego, jak grupy poddana oddziaływaniu i kontrolna różnią się od siebie niezależnie od oddziaływania*, co wyraża się różnicą w Y_0. Teraz potrafisz wyjaśnić, dlaczego możesz być podejrzliwy, gdy ktoś mówi, że obniżki cen w tak dużym stopniu zwiększają poziom sprzedaży. W przykładzie z wyprzedażą uważasz, że $E[Y_0|T=0] < E[Y_0|T=1]$, co oznacza, że firmy, które mogą sobie pozwolić na obniżki cen, sprzedają więcej niezależnie od tego, czy organizują wyprzedaż, czy nie.

Dlaczego tak się dzieje? To zagadnienie omawiam w rozdziale 3., w którym analizuję czynniki zakłócające. Na razie można uznać, że błąd systematyczny wynika z tego, iż wiele rzeczy, których nie można zaobserwować, zmienia się wraz z oddziaływaniem. W rezultacie firmy poddane i niepoddane oddziaływaniu różnią się na więcej sposobów niż tylko tym, czy urządzają wyprzedaż. Różnią się także wielkością, lokalizacją, tygodniem, w którym decydują się na wyprzedaż, stylem zarządzania, miastami, w których się znajdują, jak również wieloma innymi czynnikami. Aby móc określić, w jakim stopniu obniżki cen zwiększają poziom sprzedaży, firmy organizujące wyprzedaż i jej nieurządzające muszą być do siebie podobne. Innymi słowy, *możliwa musi być wymiana jednostek poddanych oddziaływaniu i kontrolnych.*

PRAKTYCZNY PRZYKŁAD

Jeden kieliszek wina dziennie trzyma lekarzy z dala ode mnie

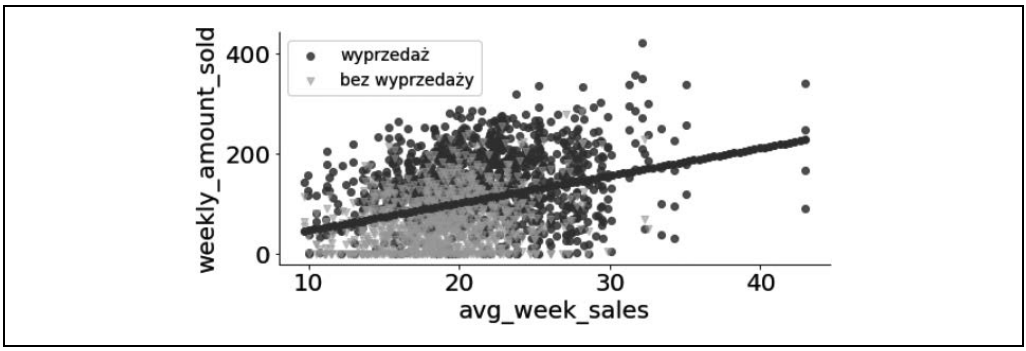
Popularne jest przekonanie, że wino w umiarkowanych ilościach jest dobre dla zdrowia. Argumentem jest to, że mieszkańcy państw śródziemnomorskich, na przykład Włoch i Hiszpanii, są znani z picia kieliszka wina każdego dnia i średnio żyją wiele lat.

Należy jednak podejść podejrzliwie do takiego twierdzenia. Aby przypisać dłuższy wiek życia winu, ci, którzy je piją, i ci, którzy go nie piją, musieliby być wymienialni, a wiadomo, że tak nie jest. Na przykład Włochy i Hiszpania mają rozbudowane systemy opieki zdrowotnej i stosunkowo wysokie wskaźniki rozwoju społecznego. Technicznie można zapisać to tak: $E[DługośćŻycia_0|PicieWina = 1] > E[DługośćŻycia_0|PicieWina = 0]$, więc błąd systematyczny może zaciemniać prawdziwy efekt przyczynowy.

Wizualny przewodnik po błędzie systematycznym

Nie musisz używać tylko matematyki i intuicji, aby mówić o wymienności. W omawianym przykładzie można nawet sprawdzić, czy firmy są wymienne, na podstawie wykresu zależności wyniku od zmiennych dla różnych grup (zobacz rysunek 1.3). Jeśli utworzysz wykres wyniku (`weekly_amount_sold`) według wielkości firmy (mierzonej na podstawie zmiennej `avg_week_sales`) i dodasz do każdego

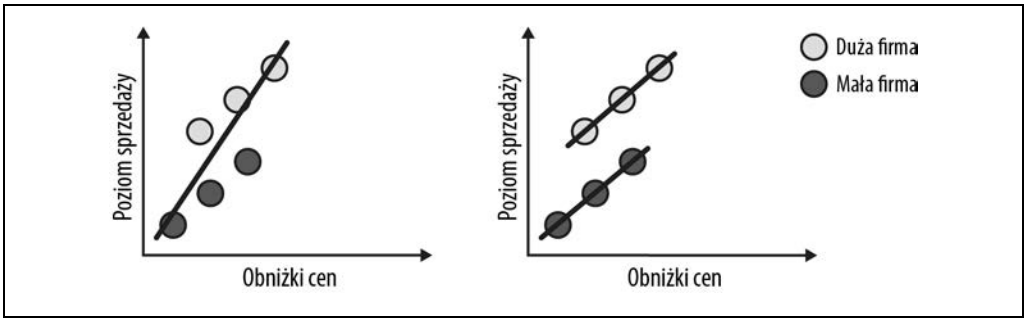
wykresu kolor na podstawie oddziaływania (`is_on_sale`), zobaczysz, że firmy poddane oddziaływaniu (organizujące wyprzedaż) znajdują się głównie po prawej stronie wykresu. To oznacza, że zwykle są to większe firmy, tak więc firmy poddane i niepoddane oddziaływaniu nie są podobne do siebie.



Rysunek 1.3. Na rysunku widać, że firmy poddane oddziaływaniu są średnio większe

Jest to bardzo mocny dowód na to, że hipoteza $E[Y_0|T=1] > E[Y_0|T=0]$ jest poprawna. Występuje tu błąd systematyczny (zawyżający efekt), ponieważ zarówno liczba firm, które obniżyły ceny ($T=1$), jak i wyniki tych firm, gdyby nie zorganizowały wyprzedaży (Y_0 dla tych firm), rosną wraz z wielkością firmy.

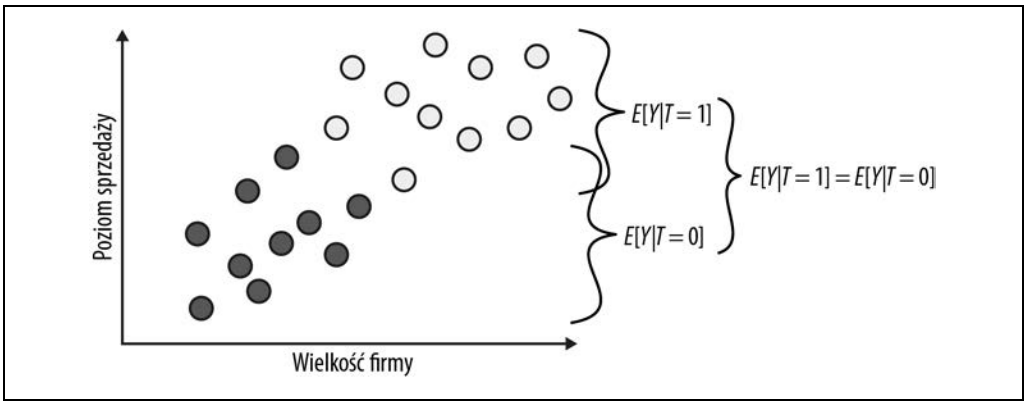
Jeśli kiedykolwiek słyszałeś o paradoksie Simpsona, zauważ, że ten błąd systematyczny jest jego mniej skrajną wersją. W paradoksie Simpsona związek między dwiema zmiennymi jest początkowo dodatni, ale po uwzględnieniu trzeciej zmiennej staje się ujemny. W omawianym przykładzie błąd systematyczny nie jest na tyle skrajny, aby odwrócić znak asocjacji (zobacz rysunek 1.4). W tym przykładzie zaczynamy od sytuacji, w której związek między obniżkami cen a poziomem sprzedaży jest zbyt wysoki, a uwzględnienie trzeciej zmiennej zmniejsza wielkość tego związku. Jeśli przyjrzysz się tylko firmom tej samej wielkości, związek między obniżkami cen a poziomem sprzedaży się zmniejsza, ale pozostaje dodatni.



Rysunek 1.4. Błąd systematyczny a paradoks Simpsona

To zagadnienie jest tak ważne, że myślę, iż warto jeszcze raz je omówić, tam razem z rysunkami. Nie są one realistyczne, ale dobrze wyjaśniają kwestię błędu systematycznego. Załóżmy, że mamy

zmienną określającą wielkość firmy. Jeśli zestawisz poziom sprzedaży z wielkością firmy, zobaczysz trend rosnący, w którym im większa firma, tym wyższa sprzedaż. Następnie pokoloruj kropki zgodnie z oddziaływaniem: białe kropki to firmy, które obniżyły ceny, a czarne kropki to firmy, które tego nie zrobiły. Jeśli porównasz średni poziom sprzedaży między firmami poddanymi i niepoddanymi oddziaływaniu, otrzymasz następujący wynik widoczny na rysunku 1.5.



Rysunek 1.5. Przykładowe dane na temat wielkości firmy i poziomu sprzedaży

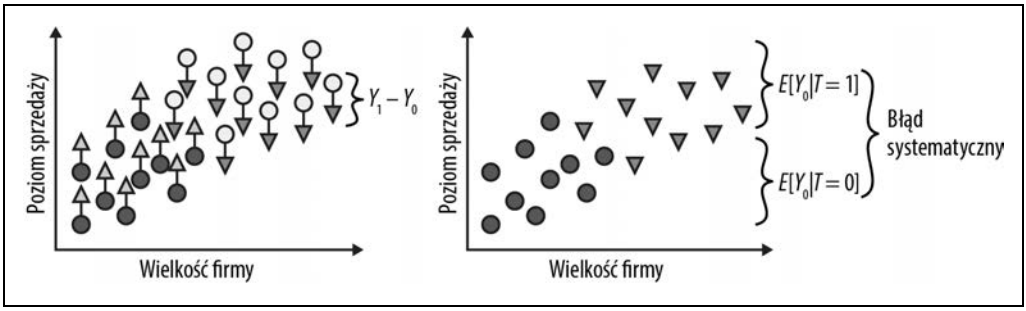
Zauważ, że różnica w poziomie sprzedaży między dwiema grupami może mieć (i prawdopodobnie ma) dwie przyczyny:

1. Efekt oddziaływania. Wzrost poziomu sprzedaży spowodowany obniżką cen.
2. Wielkość firmy. Większe firmy są w stanie zarówno sprzedawać więcej, jak i dokonywać większych obniżek cen. To źródło różnicy między firmami poddanymi i niepoddanymi oddziaływaniu *nie* wynika z obniżki cen.

Wyzwaniem we wnioskowaniu przyczynowym jest oddzielenie od siebie obu tych powodów.

Porównaj to z tym, co zobaczysz po dodaniu obu potencjalnych wyników do rysunku (zobacz rysunek 1.6; wyniki kontrfaktyczne są przedstawione za pomocą trójkątów). Indywidualny efekt oddziaływania jest różnicą między wynikiem jednostki a innym teoretycznym wynikiem, który uzyskałaby ta sama jednostka po otrzymaniu oddziaływania. Średni efekt oddziaływania, który chcesz oszacować, to średnia różnica między potencjalnymi wynikami dla każdej jednostki, $Y_1 - Y_0$. Te indywidualne różnice są znacznie mniejsze niż różnica widoczna na poprzednim wykresie, uzyskana między grupami poddaną i niepoddaną oddziaływaniu. Powodem tego jest błąd systematyczny, który jest przedstawiony na prawym wykresie.

Błąd systematyczny można zilustrować dzięki wyeliminowaniu oddziaływania dla wszystkich jednostek. W takim scenariuszu znany jest tylko potencjalny wynik Y_0. Następnie można sprawdzić, jak grupy poddana i niepoddana oddziaływaniu różnią się pod względem tych potencjalnych wyników w sytuacji braku oddziaływania. Jeśli występują różnice między grupami, oznacza to, że coś innego niż oddziaływanie powoduje, iż grupy poddana i niepoddana mu różnią się od siebie. To jest właśnie błąd systematyczny, o którym piszę. Jest to coś, co przesłania prawdziwy efekt oddziaływania.



Rysunek 1.6. Rzeczywista średnia różnica wynikająca z oddziaływania

Identyfikowanie efektu oddziaływania

Teraz, gdy już rozumiesz problem, nadszedł czas, aby przyjrzeć się rozwiązaniu (przynajmniej jednemu z nich). Identyfikacja jest pierwszym krokiem w każdej analizie mającej na celu wnioskowanie przyczynowe. Więcej technik poznasz w rozdziale 3., a na razie zapoznaj się z tym, czym jest identyfikacja. Pamiętaj, że nie możesz obserwować wartości przyczynowych, ponieważ obserwowany jest tylko jeden potencjalny wynik. Nie możesz bezpośrednio oszacować wartości $E[Y_1 - Y_0]$, ponieważ dla żadnego punktu danych nie da się zaobserwować tej różnicy. Ale być może uda się znaleźć jakąś inną wartość, która jest obserwowalna i może zostać wykorzystana do obliczenia szukanej wartości przyczynowej. Tego właśnie dotyczy proces identyfikacji: *ustalenia, w jaki sposób obliczyć wartości przyczynowe na podstawie obserwowalnych danych*. Na przykład gdyby jakimś cudem na podstawie $E[Y|T = t]$ udało się obliczyć $E[Y_t]$ (zidentyfikować $E[Y_t]$), można by uzyskać $E[Y_1 - Y_0]$ na podstawie prostego oszacowania $E[Y|T = 1] - E[Y|T = 0]$. Można to zrobić przez oszacowanie średniego wyniku dla firm poddanych i niepoddanych oddziaływaniu, które to średnie są obserwowalnymi wartościami.

Zobacz także

W ostatniej dekadzie (2010 – 2020) cała dziedzina identyfikacji przyczynowej została spopularyzowana przez Judeę Pearla i jego zespół w ramach prób ujednolicenia języka wnioskowania przyczynowego. Używam części tego języka w tym rozdziale (choć prawdopodobnie jest to jego „heretycka" wersja), a szczegółowo omawiam go w rozdziale 3. Jeśli chcesz dowiedzieć się więcej na ten temat, krótką, ale naprawdę ciekawą publikacją jest *Causal Inference and Data Fusion in Econometrics* autorstwa Paula Hünermunda i Eliasa Bareinboima.

Identyfikację można również traktować jak proces eliminowania błędu systematycznego. Na podstawie potencjalnych wyników można również stwierdzić, co jest konieczne, aby asocjacja była równa przyczynowości. *Jeśli* $E[Y_0|T = 0] = E[Y_0|T = 1]$*, to asocjacja OZNACZA PRZYCZYNOWOŚĆ!* Aby to zrozumieć, nie wystarczy zapamiętać równania. Istnieje tu silny intuicyjny argument. Stwierdzenie, że $E[Y_0|T = 0] = E[Y_0|T = 1]$, oznacza, iż grupy poddana oddziaływaniu i kontrolna są porównywalne niezależnie od oddziaływania. Matematycznie wyraz błędu systematycznego znika, a pozostaje jedynie wpływ na jednostki poddane oddziaływaniu:

$$E[Y|T=1] - E[Y|T=0] = E[Y_1 - Y_0|T=1] = ATT$$

Ponadto jeśli grupy poddana i niepoddana oddziaływaniu reagują na nie podobnie, czyli $E[Y_1 - Y_0|T=1] = E[Y_1 - Y_0|T=0]$, to (zwróć na to szczególną uwagę), *różnica między średnimi STAJE SIĘ średnim efektem przyczynowym*:

$$E[Y|T=1] - E[Y|T=0] = ATT = ATE = E[Y_1 - Y_0]$$

Choć te równania matematyczne mogą wyglądać skomplikowanie, oznaczają tylko tyle, że *jeśli grupy poddana i niepoddana oddziaływaniu są wymienne*, wyrażenie efektu przyczynowego w kategoriach wartości obserwowalnych w danych staje się banalnie proste. W omawianym przykładzie oznacza to, że jeśli firmy, które obniżają ceny lub ich nie obniżają, są do siebie podobne (czyli wymienne), wówczas różnica w poziomie sprzedaży między tymi, które organizują wyprzedaż, a tymi, które jej nie urządzają, może być w całości przypisana obniżce cen.

Założenie o niezależności

Wymienność jest kluczowym założeniem we wnioskowaniu przyczynowym. Ponieważ jest tak ważna, naukowcy znaleźli różne sposoby na jej określanie. Zacznę od jednego sposobu, prawdopodobnie najbardziej znanego, którym jest *założenie o niezależności*. Ten zapis oznacza, że potencjalne wyniki są niezależne od oddziaływania: $(Y_0, Y_1) \perp T$.

Niezależność oznacza tu, że $E[Y_0, T] = E[Y_0]$. Innymi słowy, oddziaływanie nie daje żadnych informacji o potencjalnych wynikach. Fakt, że jednostka została poddana oddziaływaniu, nie oznacza, iż będzie miała niższy lub wyższy wynik niż w sytuacji, gdyby oddziaływanie nie miało na nią miejsca (Y_0). Jest to po prostu inny sposób powiedzenia, że $E[Y_0|T=1] = E[Y_0|T=0]$. W omawianym przykładzie biznesowym oznacza to po prostu, że jeśli żadna firma nie wprowadza obniżek, nie da się odróżnić firm, które zdecydowały się na wyprzedaż, od tych, które tego nie zrobiły. Z wyjątkiem oddziaływania i jego wpływu na wynik wszystkie firmy są do siebie podobne. W zbliżony sposób $E[Y_1|T] = E[Y_1]$ oznacza, że nie da się odróżnić firm, jeśli wszystkie organizują wyprzedaż. Mówiąc prościej, oznacza to, że grupy poddana i niepoddana oddziaływaniu są podobne i nieodróżnialne (niezależnie od tego, czy wszystkie zostały poddane oddziaływaniu, czy nie).

Identyfikacja przy randomizacji

W tym podejściu niezależność jest traktowana jak założenie. Oznacza to, że wiesz, iż musisz zrównać asocjację z przyczynowością, ale jeszcze nie wiesz, jak spełnić ten warunek. Przypomnijmy, że rozwiązanie problemu z wnioskowaniem przyczynowym często dzieli się na dwa etapy:

1. Identyfikację, kiedy to ustalasz, jak wyrazić interesującą Cię wartość przyczynową w kategoriach obserwowalnych danych.
2. Estymację, kiedy to faktycznie wykorzystuje się dane do oszacowania zidentyfikowanej wcześniej wartości przyczynowej.

Aby zilustrować ten proces na bardzo prostym przykładzie, załóżmy, że można zrandomizować oddziaływanie. To prawda, wcześniej stwierdziłem, że na rynku internetowym, na którym pracujesz, firmy mają pełną autonomię w ustalaniu cen, ale nadal możesz znaleźć sposób na randomizację

oddziaływania *Wyprzedaż*. Przyjmijmy na przykład, że wynegocjowałeś z firmami prawo do wymuszenia na nich obniżek, ale platforma płaci za wymuszoną różnicę cen. W porządku, załóżmy więc, że masz teraz sposób na randomizację wyprzedaży. I co z tego? W rzeczywistości jest to bardzo ważne!

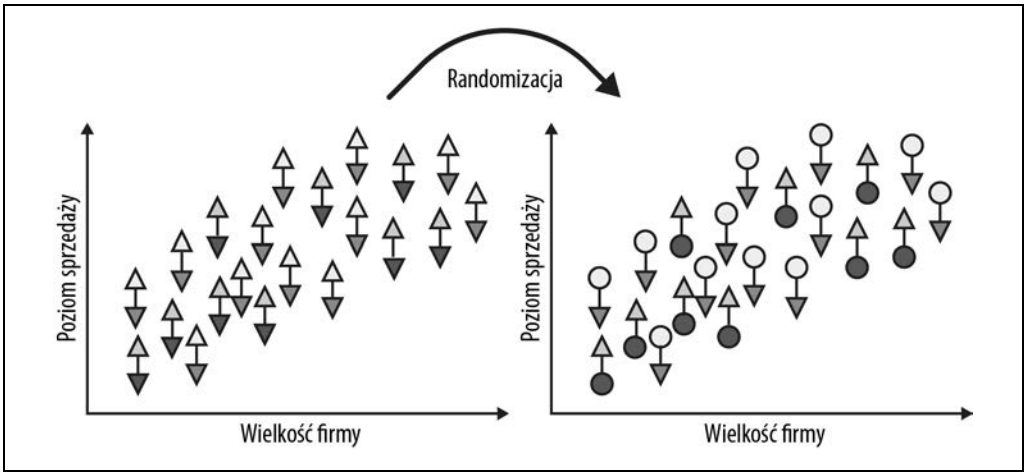
Przede wszystkim randomizacja powoduje, że przypisanie do grupy poddanej oddziaływaniu odbywa się losowo, więc to przypisanie nie jest związane z innymi czynnikami w mechanizmie przyczynowym:

$$Wyprzedaż \leftarrow rand(t)$$

$$PoziomSprzedaży \leftarrow f_y(Wyprzedaż, u_y)$$

Przy randomizacji u_t znika z modelu, ponieważ mechanizm przypisywania oddziaływania do jednostek jest w pełni znany. Co więcej, ponieważ oddziaływanie jest losowe, staje się niezależne od wszelkich innych czynników, w tym od potencjalnych wyników.

Aby to wyjaśnić, zobaczmy, w jaki sposób randomizacja w znacznym stopniu eliminuje błąd systematyczny. Zaczniemy od sytuacji przed przypisaniem oddziaływania. Pierwszy wykres przedstawia świat potencjalnych wyników (trójkątów), które nie zostały jeszcze zrealizowane (rysunek 1.7). Ilustruje to wykres po lewej stronie.



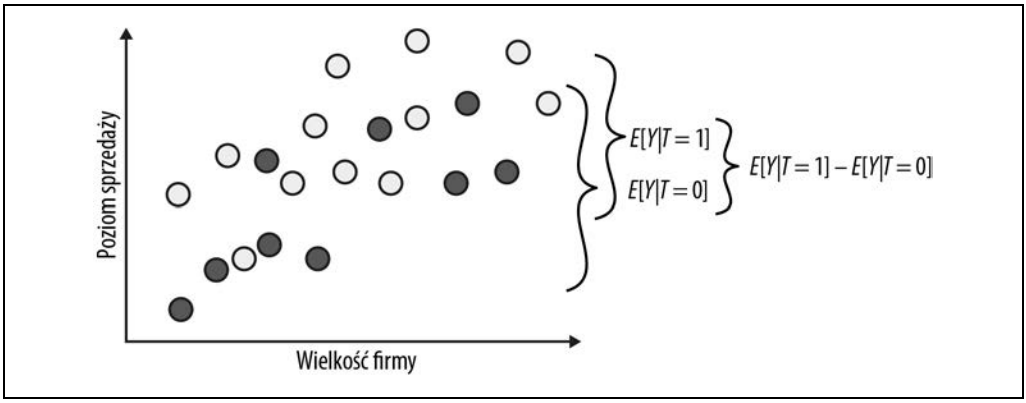
Rysunek 1.7. Randomizacja pozwala wyeliminować błąd systematyczny

Następnie oddziaływanie losowo prowadzi do materializacji jednego lub drugiego potencjalnego wyniku.

Randomizowane a obserwacyjne

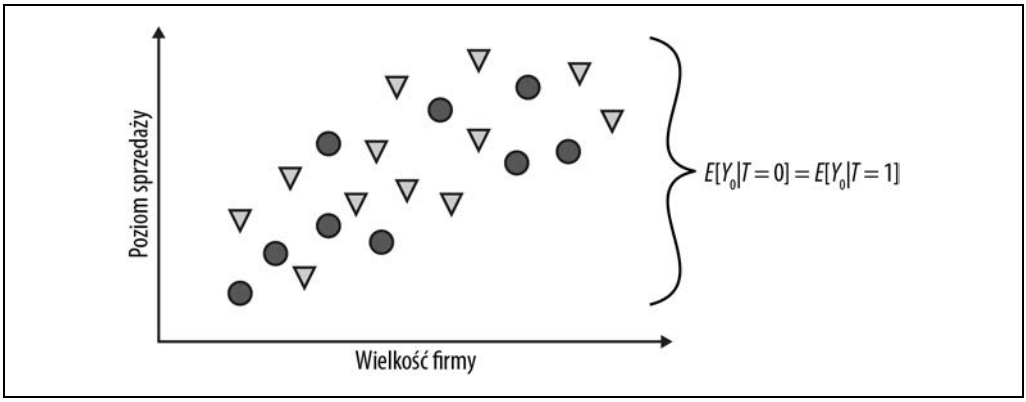
We wnioskowaniu przyczynowym używamy pojęcia *randomizowane* do opisu danych, w których oddziaływanie było randomizowane lub gdy mechanizm przypisania do grup jest w pełni znany i niedeterministyczny. Z kolei pojęcie *obserwacyjne* jest używane do opisu danych, w których można zobaczyć, kto został poddany określonemu oddziaływaniu, ale nie wiadomo, w jaki sposób to oddziaływanie zostało przypisane.

Następnie uprośćmy wykres przez usunięcie niezrealizowanych potencjalnych wyników (trójkątów). Teraz możesz porównać jednostki poddane i niepoddane oddziaływaniu (zobacz rysunek 1.8).



Rysunek 1.8. Wykres po usunięciu niezrealizowanych potencjalnych wyników

W tym przykładzie różnica w wynikach między grupami poddaną a niepoddaną oddziaływaniu jest średnim efektem przyczynowym. Dzieje się tak, ponieważ nie ma innego źródła różnicy między tymi grupami niż samo oddziaływanie. Dlatego wszystkie zaobserwowane różnice należy przypisać właśnie temu czynnikowi. Mówiąc prościej, nie występuje tu błąd systematyczny. Jeśli sprawisz, że jednostki nie zostaną poddane oddziaływaniu (co pozwoli zaobserwować samo Y_0), nie znajdziesz żadnej różnicy między grupami poddaną a niepoddaną oddziaływaniu. Ilustruje to rysunek 1.9.



Rysunek 1.9. Grupy poddana i niepoddana oddziaływaniu są takie same

Na tym właśnie polega herkulesowe zadanie identyfikacji wartości przyczynowych. Chodzi o pomysłowe znalezienie sposobów na wyeliminowanie błędów systematycznych i sprawienie, by grupy poddana i niepoddana oddziaływaniu były porównywalne, tak aby wszystkie widoczne różnice można było przypisać efektowi oddziaływania. Co ważne, *identyfikacja jest możliwa tylko wtedy, gdy wiesz coś na temat procesu generowania danych (lub jesteś skłonny przyjąć określone założenia na ten temat). Zazwyczaj istotna jest wiedza na temat przydziału lub przypisania oddziaływania.* Właśnie dlatego stwierdziłem wcześniej, że same dane nie pozwalają na uzyskanie odpowiedzi

na pytania przyczynowe. Oczywiście, dane są ważne dla oszacowania efektu przyczynowego. Ale oprócz danych zawsze będziesz potrzebować informacji o tym, jak powstały dane, a przede wszystkim jak przypisano oddziaływanie. Te informacje można uzyskać na podstawie wiedzy eksperckiej lub interwencji w rzeczywistym świecie, wpływu na oddziaływanie i zaobserwowania związanych z tym zmian w wynikach.

PRAKTYCZNY PRZYKŁAD

Niesamowity program członkowski

Duży internetowy sklep detaliczny wprowadził program członkowski, którego uczestnicy ponoszą dodatkową opłatę, aby uzyskać dostęp do większej liczby rabatów, szybszych dostaw, zwolnienia z opłat za zwrot i świetnej obsługi klienta. Aby zrozumieć wpływ programu, firma wprowadziła go dla losowej próby klientów, którzy mogli zdecydować się na uiszczenie opłaty w celu uzyskania korzyści wynikających z członkostwa. Po pewnym czasie zauważono, że klienci objęci programem członkowskim przynosili firmie znacznie większe zyski niż osoby z grupy kontrolnej. Klienci z programu nie tylko kupowali produkty od firmy, ale ponadto zajmowali mniej czasu obsłudze klienta. Czy powinniśmy zatem stwierdzić, że program członkowski okazał się wielkim sukcesem w zwiększeniu sprzedaży i skróceniu czasu przeznaczanego na obsługę klientów?

Nie do końca. Chociaż dobór osób mających możliwość zapisania się do programu był losowy, klienci z tej grupy nadal samodzielnie decydowali się na udział w programie. Innymi słowy, losowa kwalifikacja do programu sprawia, że osoby, które mogły się do niego zapisać, są porównywalne z tymi, które nie otrzymały takiej możliwości. Jednak spośród tak wybranych osób tylko część zdecydowała się wziąć udział w programie. Ta decyzja nie była już losowa. Prawdopodobnie tylko bardziej zaangażowani klienci zdecydowali się na uczestnictwo, podczas gdy okazyjni klienci zrezygnowali z udziału w programie. Tak więc mimo że kwalifikacja do programu była losowa, uczestnictwo w nim już takie nie było. W rezultacie klienci, którzy uczestniczyli w programie, nie są porównywalni z tymi, którzy nie wzięli w nim udziału.

Jeśli się nad tym zastanowić, to spośród kwalifikujących się klientów ci, którzy faktycznie zdecydowali się wziąć udział w programie, prawdopodobnie zrobili to właśnie dlatego, że wydali już dużo w danej firmie internetowej. To sprawiło, że dodatkowe rabaty były dla nich czymś, za co warto zapłacić. Oznacza to, że $E[Przychody_0|Udział = 1] > E[Przychody_0|Udział = 0]$, tak więc klienci, którzy zdecydowali się na udział w programie, prawdopodobnie generowali większe przychody niezależnie od niego.

Ostatecznym celem wnioskowania przyczynowego jest ustalenie, jak działa świat, po wyeliminowaniu wszelkich złudzeń i błędnych interpretacji. Teraz, gdy to rozumiesz, możesz przejść do opanowania niektórych z najskuteczniejszych metod eliminowania błędu systematycznego, narzędzi dla uczciwych i odważnych, które pozwalają zidentyfikować efekt przyczynowy.

Najważniejsze zagadnienia

Znasz już język matematyczny, którego będziemy używać do opisu wnioskowania przyczynowego w pozostałej części książki. Co ważne, znasz już definicję potencjalnego wyniku. Jest to wynik, który zaobserwowałbyś dla jednostki, gdyby została ona poddana określonemu oddziaływaniu $T = t$:

$$Y_{ti} = Y_i|do(T_i = t)$$

Potencjalne wyniki pomogły zrozumieć, dlaczego asocjacja różni się od związku przyczynowego. Mianowicie gdy grupy poddana i niepoddana oddziaływaniu różnią się z powodów innych niż samo oddziaływanie, czyli $E[Y_0|T = 1] \neq E[Y_0|T = 0]$, to porównanie obu grup nie daje prawdziwego efektu przyczynowego, ale jedynie obciążone błędem systematycznym oszacowanie. Wykorzystaliśmy również potencjalne wyniki, aby zobaczyć, co jest potrzebne, aby asocjacja oznaczała przyczynowość:

$$(Y_0, Y_1) \perp T$$

Gdy grupy poddana i niepoddana oddziaływaniu są wymienne lub porównywalne, tak jak w przypadku randomizacji oddziaływania, proste porównanie wyników tych grup da efekt oddziaływania:

$$E[Y_1 - Y_0] = E[Y|T = 1] - E[Y|T = 0]$$

Zacząłeś także rozumieć niektóre kluczowe założenia, które należy przyjąć podczas wnioskowania przyczynowego. Na przykład aby uniknąć błędu systematycznego podczas szacowania efektu oddziaływania, przyjąłeś założenie o niezależności między przypisaniem oddziaływania a potencjalnymi wynikami, $T \perp Y_t$.

Założyliśmy również, że oddziaływanie na jedną jednostkę nie wpływa na wynik innej jednostki (SUTVA), a także że wszystkie wersje oddziaływania zostały uwzględnione (jeśli $Y_i(t) = Y$, to $T_i = t$) dla Y będącego wynikiem funkcji wyboru między potencjalnymi wynikami:

$$Y_i = (1 - T_i)Y_{0i} + T_iY_{1i}$$

Zawsze warto pamiętać, że wnioskowanie przyczynowe wymaga założeń. Potrzebujesz ich, aby przejść od wartości przyczynowej, którą chcesz poznać, do estymatora statystycznego, który pozwala uzyskać tę wartość.

Skorowidz

N

O

P

R

S

Ś

T

U

W

X

Z

Poznaj narzędzia najbardziej znanych analityków danych korzystających z Pythona!

prof. Nick Huntington-Klein, autor *The Effect: An Introduction to Research Design and Causality*

Wnioskowanie przyczynowe przydaje się w sytuacji, gdy trzeba określić wpływ decyzji biznesowej na konkretny wynik, na przykład wielkość sprzedaży. Działania te są dobrze znane nauce, ale dopiero od niedawna świat poznaje korzyści z ich zastosowania w branży technologicznej. Przyczyniły się do tego postępy w uczeniu maszynowym, automatyzacji procesów i danologii. Teraz, aby uzyskać wymierne korzyści, wystarczy kilka wierszy kodu w Pythonie.

Tę książkę docenią w szczególności analitycy danych. Wyjaśniono w niej potencjał wnioskowania przyczynowego w zakresie szacowania wpływu i efektów w biznesie. Opisano klasyczne metody wnioskowania przyczynowego, w tym testy A/B, regresję liniową, wskaźnik skłonności, metodę syntetycznej kontroli i metodę różnicy w różnicach, przy czym skoncentrowano się przede wszystkim na praktycznym aspekcie tych technik. Znalazło się tu również omówienie nowoczesnych rozwiązań, takich jak wykorzystanie uczenia maszynowego do szacowania heterogenicznych efektów. Każda metoda została zilustrowana opisem zastosowania w branży technologicznej.

W książce między innymi:

- podstawy wnioskowania przyczynowego
- problemy biznesowe jako zagadnienia z obszaru wnioskowania przyczynowego
- eksperymenty geograficzne i eksperymenty z przełączaniem oddziaływania
- badanie błędu systematycznego
- modele graficzne i wizualizacja związków przyczynowych

Matheus Facure jest ekonomistą i starszym analitykiem danych w Nubank, brazylijskiej firmie z branży FinTech. Z powodzeniem stosował wnioskowanie przyczynowe w rozmaitych scenariuszach biznesowych. Często występuje jako prelegent na konferencjach branżowych i uczestniczy w projektach open source.

Najlepsza książka poświęcona najnowocześniejszym metodom, działaniu na rzeczywistych danych i rozwiązywaniu praktycznych problemów!

Sean J. Taylor, główny badacz w Motif Analytics

Helion
helion.pl
HELION S.A.
ul. Kościuszki 1c
44-100 Gliwice
tel.: 32 230 98 63
helion@helion.pl
KOD KORZYŚCI
Sięgnij po więcej!
ISBN 978-83-289-0881-9
9 788328 908819
Cena: 74,90 zł