

Marcin Gorazda

Przyczynek do krytyki statystyczno-relevantnego modelu wyjaśniania naukowego

Zagadnienia Filozoficzne w Nauce nr 45, 124-139

2009

Artykuł został opracowany do udostępnienia w internecie przez Muzeum Historii Polski w ramach prac podejmowanych na rzecz zapewnienia otwartego, powszechnego i trwałego dostępu do polskiego dorobku naukowego i kulturalnego. Artykuł jest umieszczony w kolekcji cyfrowej bazhum.muzhp.pl, gromadzącej zawartość polskich czasopism humanistycznych i społecznych.

Tekst jest udostępniony do wykorzystania w ramach dozwolonego użytku.

Marcin GORAZDA

PAT, Wydział Filozofii, Katedra Metafizyki

***PRZYCZYNEK DO KRYTYKI
STATYSTYCZNO-RELEWANTNEGO MODELU
WYJAŚNIANIA NAUKOWEGO***

1. MODELE WYJAŚNIANIA NAUKOWEGO

Statystyczno-relevantny model wyjaśniania naukowego został zaproponowany przez Wesley Salmona w 1971 r. w tekście *Statistical Explanation*¹. Model ten stanowił alternatywę dla uprzednio proponowanych modeli wyjaśniania m.in. przez Hempła a wspieranych przez wielu współczesnych filozofów nauki (m.in. przez Poppera). Te poprzedzające modele to: dedukcyjno-nomologiczny (DN), dedukcyjno-statystyczny (DS) oraz indukcyjno-statystyczny (IS). Model dedukcyjno-nomologiczny stanowi pewien schemat wnioskowania gdzie wnioskiem jest *explanandum* (to co wyjaśniamy), a przesłanką jest *explanans* (zbiór zdań wyjaśniających). *Explanans* musi obejmować przynajmniej jedno prawo przyrody (*nomos*), które stanowi istotny element wnioskowania (bez tego prawa wnioskowanie jest zawodne). Ponadto *explanans* musi mieć treść empiryczną, która może być sfalsyfikowana, a zdania należące do niego muszą być prawdziwe. Jeżeli prawo w *explanansie* ma charakter statystyczny, a *explanandum* stanowi prawidłowość statystyczna, wówczas model nazywa

¹ *Statistical Explanation and Statistical Relevance*, W. Salmon, (ed.), 29–87, Pittsburgh: University of Pittsburgh Press.

się dedukcyjno-statystycznym. Jeśli w *explanandum* mamy pojedyncze zdarzenie i określamy prawdopodobieństwo jego wystąpienie ze względu na statystyczne prawo zawarte w *explanansie*, wówczas model nazywamy indukcyjno-statystycznym². Wszystkie te modele odróżnia od modelu SR zaproponowanego przez Salmona konieczność sformułowania w *eksplanansie* (czyli jako element przesłanki wyjaśniającej) jakiegoś *prawa przyrody* czy to o charakterze bezwzględny czy też statystycznym. U podstaw tych modeli tkwi zatem pewien zestaw założeń, który prowadził w konsekwencji do wielu krytycznych uwag. Po pierwsze, bowiem, nie do końca potrafimy sprecyzować co jest, a co nie jest *prawem przyrody*. Po drugie, w wielu przypadkach faktycznego wyjaśniania zjawisk nie posługujemy się jakimkolwiek zgeneralizowanym *prawem przyrody* lecz raczej prostym związkiem, który jawi nam się jako przyczynowy. Po trzecie, co wynika zresztą wprost z poprzedniego zdania, modele DN, DS oraz IS zasadniczo implikują istnienie związku przyczynowego z określonym kierunkiem działania pomiędzy *eksplanans* a *eksplanandum*. Model SR ucieka od pojęcia *prawa przyrody*, zastępując go właśnie statystyczną relewancją dwóch zjawisk. Co zaś do związku przyczynowego, nie tyle go zakłada, co rości sobie pretensje do jego uchwycenia właśnie poprzez ustalenie owej statystycznej relewancji.

2. MODEL SR

Spróbujmy pokrótce przybliżyć istotę wyjaśnienia naukowego z wykorzystaniem modelu SR. Rozważmy jakąś klasę lub populację *A*. W klasie tej atrybut *C* będzie statystycznie relewantny względem innego atrybutu *B* wtedy i tylko wtedy, jeżeli $P(B|A.C) \neq P(B|A)$, a zatem jeżeli prawdopodobieństwo zdarzenia *B* pod warunkiem zaistnienia zdarzeń *A* i *C* jest różne od prawdopodobieństwa zdarzenia *B* uwarunkowanego wyłącznie przez *A*. Zasadę powyższą zilustruję przykładem pochodzącym od autora tego modelu (który to przykład zresztą służył jako narzędzie krytyki modelu DN). Interesująca nas populacja

²Zob. A. Grobler, *Metodologia nauk*, Wydawnictwo Znak, Kraków 2006, s. 104.

to zbiór ludzi oraz dwa podzbiory — mężczyźni i kobiety. W obrębie tej populacji badamy relewancje dwóch atrybutów: stan ciąży oraz zażywanie pigułek antykoncepcyjnych. Stwierdzamy co następuje:

$$\begin{aligned} P(\text{Cięża}|\text{Człowiek.Mężczyzna.Zażywa pigułki}) = \\ P(\text{Cięża}|\text{Człowiek.Mężczyzna}) = 0, \end{aligned}$$

podczas gdy

$$\begin{aligned} P(\text{Cięża}|\text{Człowiek.Kobieta.Zażywa pigułki}) \neq \\ P(\text{Cięża}|\text{Człowiek.Kobieta}), \end{aligned}$$

przy założeniu oczywiście że nie wszystkie kobiety w populacji zażywają pigułki antykoncepcyjne. W takiej sytuacji stwierdzamy, że zażywanie pigułek antykoncepcyjnych jest statystycznie relewante w stosunku do zajścia w ciążę, wyłącznie w odniesieniu do kobiet. W odniesieniu do mężczyzn zażywanie pigułek jest irrelewantne.

Kluczowym pojęciem (i warunkiem prawidłowości zastosowania modelu SR sformułowanym przez Salmona) jest tzw. rozkład jednorodny (*homogenous partition*). O rozkładzie jednorodnym klasy A mówimy wówczas gdy, istnieje zestaw podklas lub elementów C_i należących do A , takich, że są one wzajemnie rozłączne i kompletne gdzie $P(B|A.C_i) \neq P(B|A.C_j)$ dla każdego $C_i \neq C_j$ i gdzie nie istnieje żaden dodatkowy atrybut D_k w A taki, że $P(B|A.C_i) \neq P(B|A.C_i.D_k)$.

W modelu SR zatem wyjaśnienie, dlaczego jakiś element x w danej klasie charakteryzujący się atrybutem A posiada atrybut B , składa się z następującego zestawu informacji:

- (i) Uprzednie prawdopodobieństwo B od A : $P(B|A) = p$.
- (ii) Rozkład jednorodny A ze względu na B , $(A.C_1, \dots, A.C_n)$, razem z prawdopodobieństwem B ze względu na każdy element rozkładu: $P(B|A.C_i) = p_i$ oraz
- (iii) Elementy rozkładu, do których należy element x .

Znowu posłużę się przykładem pochodzącym od autora. Przypuścimy, że chcemy wyjaśnić przy wykorzystaniu modelu SR, dlaczego pacjent x z zakażeniem bakteryjnym (S), szybko powraca do zdrowia (Q). Niech $T(-T)$ oznacza sytuację, w której pacjent x poddany jest (lub nie) leczeniu penicyliną, a $R(-R)$ oznacza, iż dany szczep bakterii jest odporny (lub nie) na działanie penicyliny. Musimy przy tym założyć, że nie istnieją żadne inne czynniki relewantne w stosunku do szybkiego powrotu do zdrowia. W opisanym stanie mamy cztery możliwe kombinacje właściwości: $T.R, -T.R, T.-R, -T.-R$. Przypuścimy, że zachodzi: $P(Q|S.T.R) = P(Q|S.-T.R) = P(Q|S.-T.-R) \neq P(Q|S.T.-R)$, czyli że prawdopodobieństwo szybkiego powrotu do zdrowia przy zakażeniu jest takie samo dla pacjentów którzy zostali zakażeni szczepem odpornym na penicylinę, niezależnie od tego, czy byli oni poddani leczeniu penicyliną, czy też nie oraz takie samo jak i tych, którzy nie byli poddani temu leczeniu. Inaczej w przypadku pacjentów, którzy zakażeni zostali szczepem nie odpornym i poddani byli leczeniu penicylinowemu. Tu prawdopodobieństwo szybkiego powrotu do zdrowia jest różne (jak można się spodziewać znacząco większe). W tym przypadku $[S.(T.R \vee -T.R \vee -R.-T)]$, $[S.T.-R]$ stanowi rozkład jednorodny S ze względu na Q . Wyjaśnienie SR obejmujące szybki powrót do zdrowia x -a będzie składało się ze stwierdzenia prawdopodobieństwa szybkiego powrotu do zdrowia w grupie wszystkich zakażonych (tj. (i) powyżej), ze stwierdzenia prawdopodobieństwa szybkiego powrotu do zdrowia w każdym spośród dwóch elementów powyższego rozkładu jednorodnego (tj. (ii) powyżej), oraz z określenia, do którego zbioru należy x , czyli $S.T.-R$ (tj. (iii) powyżej).

3. OGÓLNA KRYTYKA MODELU SR

Tego typu model wyjaśniania posiada niewątpliwie wiele zalet, którymi nie dysponują modele DN, DS. i IS. O jednej już wspomniałem. Nie wymaga on doszukiwania się jakiegoś *prawa przyrody*, co zawsze budzi spory. Dodatkowo model ten pozwala na wyjaśnienie zjawisk których prawdopodobieństwo zajścia jest niewielkie. Nie jest

istotne dla tego modelu, w odróżnieniu od modelu DS i IS, aby prawdopodobieństwo wystąpienia jakiegoś zdarzenia A ze względu na B było większe od 0,5. Wystarczy, że będzie ono różne od prawdopodobieństwa wystąpienia zdarzenia A ze względu na inne niż B atrybuty rozkładu jednorodnego. Model ten, co wydaje się nieco kontrintuicyjne, prowadzi także do tego, że ten sam *eksplanans* E może wyjaśniać *eksplanandum* M oraz *eksplananda* które nie są spójne z M (np. $\neg M$).

Praktyczne stosowanie tego modelu, nastręcza jednak wiele problemów, które pokazują jego słabości. Był ich świadom autor oraz wielu krytyków jego stosowania. Ogólnie można powiedzieć, że o ile wyjaśnienie zjawisk w obszarze fizyki, właśnie przy zastosowaniu takiego modelu sprawdza się całkiem dobrze (w tym także zjawisk niezdeterminowanych, probabilistycznych, np. z obszaru mechaniki kwantowej), o tyle wyjaśnianie zjawisk z obszaru funkcjonowania organizmów żywych (mikrobiologii), a w szczególności z obszaru zjawisk społecznych albo nie spełnia warunków modelu SR, albo prowadzi do trudnych do zaakceptowania wniosków. Podstawowy problem wiąże się z wymogiem dookreślenia owego rozkładu jednorodnego. W obszarze biologii czy nauk społecznych spełnienie tego warunku w sposób restrykcyjny jest praktycznie niemożliwe. Czy wynika to z jakiejś ontologicznej właściwości organizmów żywych, a w szczególności człowieka, czy też ze słabości naszego aparatu poznawczo-obliczeniowego (ogromna ilość możliwych atrybutów, z których część pozostaje przed nami ukrytych), to zupełnie inny problem. Nie szukając daleko, dość zwrócić uwagę na fakt, że w podanym powyżej przykładzie, przyjęliśmy założenie „...że nie istnieją żadne inne czynniki relewantne względem szybkiego powrotu do zdrowia”. Założenie to jest w sposób oczywisty fałszywe. Takie czynniki istnieją i część z nich jest dostępna naszemu poznaniu. Jednym z nich jest np. ogólny stan odporności organizmu. Przy badaniu zatem tego typu zjawisk, z góry godzimy się na to, iż nie dane nam będzie analizować wszystkich relewantnych czynników, a jedynie wybrane, co skazuje uzyskane wyjaśnienie na ryzyko znaczącego błędu. Praktyka badań farmaceutycznych pokazuje, że błę-

dów tego typu jest niezliczona ilość. Typowy przykład to stosowanie określonej metody leczenia, która na podstawie analizy statystycznej relewancji wydaje się prowadzić do poprawy stanu zdrowia. W wielu przypadkach ten sam skutek osiągnąć jest przy zastosowaniu placebo, co wykazują znacznie późniejsze badania. Dziś już wiadomo, że zażywanie syntetycznie uzyskanego kwasu askorbinowego — witaminy C — nie ma większego sensu. Jednak przez kilkadziesiąt lat był to podstawowy środek leczenia przeziębień.

Drugi problem, który jest częściowo pokrewny z problem rozkładu jednorodnego to wstępne, często intuicyjne typowanie tych elementów *eksplanansa* które mają wykazać się statystyczną relewancją. Naturalną konsekwencją niemożności analizy wszystkich elementów rozkładu jednorodnego jest analiza statystycznej relewancji elementów wybranych. Przypadkowy ich wybór zdeterminuje uzyskane wyniki i może prowadzić do błędnych wniosków co do rzeczywistych, istotnych składników *eksplanansa*.

Trzeci problem to niezależność elementów rozkładu jednorodnego. Przy braku możliwości skonstruowania kompletnego rozkładu, nie mamy pewności czy elementy wytypowane rzeczywiście pozostają względem siebie niezależne.

Czwarty problem, który nie został wskazany przez krytyków tego modelu, to kwestia kwantyfikacji elementów *eksplanansa* i *eksplanandum*. Wyjaśnienie, to nie tylko poszukiwanie statystycznej zależności pomiędzy określonymi atrybutami, ale także próba wyrażenia tej zależności jakimś równaniem. Nie tylko bowiem istotne jest stwierdzenie, że wystąpienie zdarzenia A zwiększa prawdopodobieństwo zdarzenia B ale też o ile je zwiększa. Sytuacja staje się jeszcze bardziej skomplikowana, jeśli zdarzenia A i B są stopniowalne. W obszarze zjawisk z dziedziny fizyki taka kwantyfikacja wydaje się stosunkowo prosta. Większość zmiennych fizycznych posiada bowiem jakiś wymiar liczbowy (prędkość, pęd, okres połowicznego rozpadu, energia itp.). W obszarze zjawisk społecznych kwantyfikacja zmiennych zależnych (przypisanie im jakichś wartości liczbowych) pozostaje często problematyczna i dokonywana jest intuicyjnie, przy arbitralnych zało-

zeniach. Jak np. zmierzyć natężenie szczęśliwości u człowieka albo poziom stresu? Prostsze przykłady też bywają problematyczne. Jak np. określić liczbowo stan techniczny urzędzenia? Jak wykaże poniższa analiza, przyjęcie określonej metody kwantyfikacji może mieć kluczowe znaczenie nie tylko ze względu na znajdowanie określonych zależności liczbowych ale także na samo ustalenie czy dwa atrybuty są względem siebie statystycznie relewantne czy też nie.

Piąty i ostatni problem to kwestia, która pojawia się nieco na boku głównych rozważań dotyczących modelu wyjaśniania SR, i określe ją jako problem nieliniowości społecznych układów złożonych. Rzecz w tym, iż poszukując wyjaśnień statystyczno-relewantnych w obszarze nauk społecznych musimy być świadomi, że zwykle zależności te budowane są w oparciu o wybiórcze dane, na podstawie których konstruuje się modele zależności liniowych. Owa liniowość jednak stanowi li tylko odległą aproksymację rzeczywistych zależności, które liniowe nigdy nie są (o ile w ogóle istnieją i zachowują jakąkolwiek obserwowalną stabilność w dłuższym okresie). W konsekwencji modele te zdają się dobrze wyjaśniać tylko zależności w analizowanym obszarze danych początkowych. Jakikolwiek wyjście poza ten obszar danych i próba tworzenia w oparciu o statystyczny model liniowy jakichkolwiek predykcji zwykle prowadzi do uzyskania wyników nie znajdujących potwierdzenia w rzeczywistości.

4. KRYTYKA MODELU SR NA PRZYKŁADZIE ANALIZY STATYSTYCZNEJ RYNKU

Wszystkie powyższe mankamenty modelu SR postaram się pokazać na przykładzie znanego i wydaje się dobrze sprawdzonego w ekonomii modelu wyceny określonych dóbr z wykorzystaniem tzw. analizy statystycznej rynku³. *Eksplanandum* stanowi tutaj cena określonego dobra na rynku. Elementami *eksplanansu* są wszystkie zidentyfikowane czynniki (zmienne niezależne) które wpływają na ową cenę.

³Zob. J. Hozer, S. Kokot, W. Kuźmiński, *Metody analizy statystycznej rynku w wycenie nieruchomości*, PFSRM, Warszawa 2002.

Warto przy tym nadmienić, że model ten, po pierwsze jest uznawany za jeden z lepszych sposobów wyjaśnienia ceny określonego dobra na rynku, po drugie jest on statystycznie dość skomplikowany, co pozwoli uchwycić wszystkie spośród wskazanych powyżej problemów. Pomimo swojego skomplikowania model wydaje się odpowiadać modelowi SR choć pewne założenia trzeba będzie przeformułować.

Analiza statystyczna rynku pozwala nam w założeniu odpowiedzieć na pytanie: Jakie czynniki (zmiennie niezależne) wpływają na cenę określonego dobra oraz w jaki sposób wpływają? Analiza rozpoczyna się zatem od niemal indukcyjnej analizy danych dotyczących rzeczywistych transakcji występujących na rynku w jakimś zadanym okresie historycznym. Im więcej jest analizowanych transakcji tym większa trafność uzyskanych wyników. Statystycy przyjmują jednak, że przy niewielkich wariancjach (co to są niewielkie wariancje?) wystarczy nawet ok. 10 transakcji aby wypracować w miarę trafny model wyceny. Końcowym elementem analizy jest opracowanie równania regresji liniowej (jeśli zmiennych niezależnych jest więcej niż jedna równanie będzie równaniem wielokrotnym lub inaczej wielorakim). Do takiego równania modelowego możemy podstawić dane dotyczące konkretnego wycenianego dobra aby uzyskać informacje o jego cenie rynkowej. Budowanie modelu statystycznego sprowadza się zatem do trzech podstawowych etapów: Typowanie zmiennych od których może zależeć cena, przypisanie tym zmiennym wartości liczbowych i sprawdzenie jak wpływają na cenę oraz sprawdzenie jakości modelu (w tym istotności zmiennych).

Etap pierwszy polega na tym, że spośród dostępnych danych nt. analizowanych transakcji należy wytypować „kandydatów” na zmiennie zależne, od których może zależeć cena określonego produktu na rynku. Przykładowo, jeśli przedmiotem naszej wyceny (wyjaśnienia, dlaczego cena w poszczególnych transakcjach wynosiła tyle ile wynosiła) będą używane samochody osobowe, to takimi kandydatami mogą być: marka samochodu, rok produkcji, stan licznika, stan techniczny, wyposażenie dodatkowe itp. Przechodząc na język modelu SR szukamy wszystkich podklas/atributów C_i, \dots, C_n charakteryzujących

zbiór A (samochodów osobowych) ze względu na które określać będziemy prawdopodobieństwo, iż zmienna zależna B (cena samochodu) będzie ulegać zmianie odpowiednio do zmiany atrybutów C_1, \dots, C_n . Już w tym miejscu należy zauważyć, że spełnienie warunku rozkładu jednorodnego jest niemożliwe. Zawsze bowiem istnieje możliwość, że w analizowanej klasie będzie występował dodatkowy parametr D wpływający na cenę, który nie został przez nas zidentyfikowany. Ten brak identyfikacji może wynikać bądź z braku wiedzy o wystąpieniu tego parametru (np. samochód kupiła atrakcyjna blondynka co istotnie wpłynęło na obniżenie ceny, ale fakt ten nie jest ujawniony w dostępnych danych), bądź też z błędnego, arbitralnego założenia badacza, że parametr ten nie powinien mieć żadnego wpływu na cenę produktu. Dobór „kandydatów” na zmienne zależne nie jest wynikiem jakiegoś algorytmizowalnego procesu, ale raczej wynikiem naszych przed-założeń dotyczących czynników wpływających lub nie na cenę produktu. Założenia te w istotny sposób ograniczają liczbę „kandydatów”. W przypadku wspomnianego wyjaśnienia ceny używanych samochodów osobowych, rzecz wydaje się mieć umiarkowane znaczenie. Użytkownicy samochodów są w mniejszym lub w większym stopniu zgodni, co do elementów potencjalnie wpływających na wycenę. Sytuacja komplikuje się, kiedy wyjaśnienie ma obejmować produkty, w odniesieniu do których brak jest jakiegokolwiek intuicji, co do elementów wpływających na ich wartość. Przykładem mogą być sieci telekomunikacyjne lub dzieła sztuki.

Odrębnym problemem związanym z doбором kandydatów na zmienne niezależne jest ich „niezależność”. Postulowany rozkład jednorodny powinien być nie tylko kompletny ale i rozłączny co oznacza, że poszczególne czynniki nie powinny być, jak to określa się w statystyce, współzależne. W przypadku wyjaśnienia zjawisk z obszaru nauk społecznych owa współzależność wydaje się być nie do uniknięcia. Czasem będzie ukryta, ale w większości przypadków jest jawna, a mimo to ignorowana z braku lepszego rozwiązania. Powróćmy do przykładu z wyjaśnianiem ceny samochodu. Trzy elementy, które — jak się wydaje — intuicyjnie wpływają na tą cenę i mogą zostać wyty-

powane jako kandydaci na zmienne niezależne to: rok produkcji, stan licznika i stan techniczny. Nie trudno zauważyć, że od strony modelu wyjaśniania SR można by je było zredukować li tylko do stanu technicznego. Wszak rok produkcji oraz stan licznika to tylko dodatkowe elementy, które pozwalają nam lepiej oszacować ów stan techniczny. Jednak pozbycie się tych czynników wydaje się być absurdalne i skąd inąd słusznie. W istocie bowiem pokrywają one techniczną niemożliwość dokładnej analizy stanu technicznego i stopnia zużycia wszystkich podzespołów.

Kolejny etap to przypisanie wytypowanym zmiennym wartości liczbowych. Rzecz wydaje się stosunkowo prosta w przypadku gdy wytypowane zmienne posiadają swoje wartości liczbowe niejako w sposób naturalny. Rok produkcji czy też stan licznika w samochodzie osobowym to są już określone dane liczbowe. Tak jak jednak zaznaczyłem wyżej, sytuacja komplikuje się, gdy czynnik hipotetycznie wpływający na cenę jest niepoliczalny. W przypadku samochodów osobowych może to być np. stan techniczny. Rzeczoznawcy dokonujący wyceny radzą sobie najczęściej z takimi zmiennymi w ten sposób, iż stosując określoną punktację opracowując uprzednio kryteria przydziału punktów. Tu jednak pojawia się problem, którego nie wszyscy stosujący model analizy statystycznej rynku, są świadomi. Wydaje się bowiem pozornie, że o ile tylko zostaną prawidłowo zdefiniowane i zastosowane kryteria punktacji, to przyjęta skala nie powinna mieć znaczenia. Tymczasem tak nie jest. Jeżeli bowiem dokonujemy oceny stanu technicznego urządzenia w skali od 1 do 5, to oznacza to, że urządzenie ocenione na 1 jest pięciokrotnie „gorsze” od urządzenia ocenionego na 5. Jeżeli zaś przyjmujemy skalę od 15 do 20, to zupełnie zmienia wagę, jaką przypisujemy stanowi technicznemu. Przyjęta skala nie będzie miała znaczenia, jeśli tylko wytypowana zmienna przejdzie tzw. test istotności. Przy konstruowaniu bowiem równania liniowego, przyjęta skala zostanie odpowiednio skorygowana współczynnikami równania. Problem jednak w tym, że skala ta może zadecydować o tym, że dana zmienna przy teście istotności okaże się być (fałszywie) nieistotna i tym samym pomijalna.

Na etapie przypisywania zmiennym wartości liczbowych dokonuje się również analiza tego, jak owe zmienne mogłyby wpływać/wyjaśniać cenę produktu. W metodzie analizy statystycznej rynku, badanie owych zależności dokonuje się poprzez zestawienie danych dotyczących ceny produktów oraz danych dotyczących każdej wytypowanej zmiennej i wyliczenie tzw. współczynnika korelacji oraz, o ile współczynnik ten wykracza poza zdefiniowaną wartość graniczną (co wskazuje na istnienie jakiejś korelacji), wykreślenie zależności liniowej tzw. metodą najmniejszych kwadratów. Współczynnik korelacji jest w istocie jednym z elementów określania istotności danego czynnika tzn. na ile wpływa on w istocie na wyjaśnianą wartość. W modelu SR istotność ową mierzy się li tylko badając prawdopodobieństwo w zależności od wystąpienia lub nie określonego atrybutu dodatkowego. Przy analizie statystycznej rynku takie badanie byłoby ułomne, głównie z tego powodu, iż jak wyżej zaznaczono uzyskanie rozkładu jednorodnego jest niemożliwe. Dlatego też, surogatem tego podstawowego braku, są bardzo skomplikowane statystyczne metody analiz istotności. Jedną z nich jest właśnie metoda współczynnika korelacji. Sposobów na wyliczanie tego współczynnika jest wiele, jednym z częściej stosowanych jest tzw. współczynnik korelacji Pearsona.

„Współczynnik korelacji liniowej Pearsona określa poziom zależności liniowej między zmiennymi losowymi. Niech x i y będą zmiennymi losowymi o ciągłych rozkładach. x_i, y_i oznaczają wartości prób losowych tych zmiennych ($i = 1, 2, \dots, n$), natomiast \bar{x}, \bar{y} — wartości średnie z tych prób, tj.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

Wówczas współczynnik korelacji liniowej definiuje się następująco:

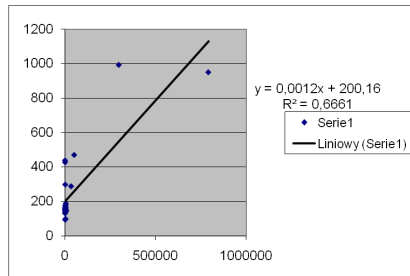
$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}},$$

$$r_{xy} \in [-1, 1].$$

Innymi słowy współczynnik korelacji liniowej dwóch zmiennych jest ilorazem kowariancji i iloczynu odchyleń standardowych tych zmiennych:

$$r_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \text{,}^{4}.$$

Rysunek poniżej obrazuje taką przykładową analizę zależności.



Na osi Y układu współrzędnych zaznaczono cenę produktu. Na osi X zaś wartości liczbowe jednej z wytypowanych zmiennych. Punkty oznaczają faktyczne pary zależności zebrane z rynku. Współczynnik korelacji R^2 jest relatywnie wysoki bo wynosi 0,661 (przy wartościach powyżej 0,5 przyjmuje się istnienie korelacji). Rysunek ten jednak pokazuje jak duża jest tolerancja na błąd, przy tworzeniu de facto sztucznych zależności liniowych. Gdyby połączyć bowiem punkty obrazujące rzeczywiste dane jakąś krzywą to, jeśli już mówić o jakiejś zależności, to raczej miałaby ona kształt odwróconej hiperboli niżli prostej. Jeśli uwzględnimy przy tym fakt, że w standardowym modelu analizy statystycznej wyjaśniającym cenę określonego produktu, takich

⁴Powyższa definicja pochodzi z Wikipedii: <<http://pl.wikipedia.org>>: „Współczynnik korelacji Pearsona”. Autor jest świadom, że nie jest przyjęte powoływanie się na Wikipedię w tekstach, które roszczą sobie pretensję do bycia naukowymi. Jednakże cytowana tu definicja wydaje się po prostu najlepsza i najbardziej przejrzysta dla osób nieobeznanych ze statystyką. Bardziej wymagających odsyłam do podręczników, w tym także powoływanych w niniejszym tekście. Współczynnik korelacji Pearsona wyjaśniany jest m.in. w glosariuszu Elektronicznego Podręcznika Statystyki PL, Krakow, WEB: <<http://www.statsoft.pl/textbook/stathome.html>>.

sztucznych zależności jest co najmniej kilka jeśli nie kilkanaście, to częściowo tłumaczy to, dlaczego dokonywanie jakichkolwiek predykcji, co do wartości zmiennej y w zależności od zmiennej x w oparciu o powyżej wykreśloną zależność liniową tak bardzo odbiega od rzeczywistych obserwacji. Rozbieżność ta jest tym większa, im bardziej wartości zmiennej niezależnej x wykraczają poza obszar obserwowany będący podstawą do wykreślenia przybliżonej zależności (wartości x większe od 1.000.000).

Testowanie jakości uzyskanego modelu wstępnego, jest etapem, który w praktyce jest najchętniej pomijanym. Prawdłowe bowiem zastosowanie w tym celu narzędzi statystycznych często prowadzi do ustaleń podważających arbitralne założenia wstępne. Im więcej użytych zostanie narzędzi do testowania tym, większe prawdopodobieństwo eliminacji kolejnych czynników, które hipotetycznie wyjaśniają cenę. Etap testowania, wraz z omówionym powyżej etapem kreślenia zależności liniowej i badania wartości współczynnika korelacji, odpowiadają temu co w modelu SR określa się jako badanie prawdopodobieństwa określonego zdarzenia B uwarunkowanego atrybutem A i C , gdzie C stanowi kandydata na *eksplanans*. Testowanie pozwala nam na określenie tzw. poziomu istotności stwierdzonej zależności oraz w przypadku jeżeli test jest negatywny, nakazuje odrzucenie określonej hipotezy, na podstawie której zamierzaliśmy wyjaśnić *eksplanandum*. Istnieje wiele narzędzi statystycznych służących badaniu poziomu istotności. Jednym z nich jest tzw. poziom istotności p (p -value). „Statystyczną istotnością wyniku nazywamy miarę stopnia, do jakiego jest on prawdziwy (w sensie jego reprezentatywności dla całej badanej populacji). [...] Poziom- p odpowiada **prawdopodobieństwu popelnienia błędu** polegającego na tym, że przyjmujemy uzyskany rezultat jako prawdziwy, tj. reprezentatywny dla populacji. Na przykład poziom- p równy 0,05 (tzn. 1/20) oznacza, że istnieje 5% szansa, iż odkryta w próbce relacja jest dziełem przypadku. [...] **Decyzja o tym, jaki poziom istotności skłonni jesteśmy uznać za rzeczywiście istotny, jest zawsze podejmowana w sposób arbitralny. Oznacza to, że wy-**

bór poziomu istotności, powyżej którego rezultat będzie odrzucany jako nieistotny, jest wyborem umownym”⁵.

W powyższym cytacie pozwoliłem sobie podkreślić dwa elementy, po pierwsze poziom- p jest w istocie miarą prawdopodobieństwa i z tego punktu widzenia odpowiada modelowi SR. Po drugie jednak, o ile w modelu SR rozstrzygnięcie, czy atrybut C jest czy też nie *eksplanansem* jest zdeterminowane miarą prawdopodobieństwa (wystarczy że miara ta jest różna przy atrybucie C od miary występującej przy pozostałych atrybutach) o tyle w przypadku modeli statystycznych miara poziomu $-p$ jest ściśle umowna, lub jak chce autor cytatu — arbitralna, co oczywiście może być kolejnym źródłem błędu.

5. WNIOSKI KOŃCOWE

Model wyjaśnienia naukowego SR sformułowany przez Salmona miał w założeniu unikać mankamentów modeli poprzednich, a głównie odwołania się do *prawa przyrody*. Jak widać z powyższej analizy, wydaje się on ciekawą propozycją, odpowiadającą z grubsza modelowi wyjaśniania/wnioskowania statystycznego. Praktyczne zastosowanie tego modelu w obszarze nauk społecznych trafia jednak na szereg problemów, które pokazują, że jest on zbyt idealistyczny i tym samym niesprawny, a dodatkowo wbrew pierwotnym założeniom nie unika uwikłania w problematykę *prawa przyrody*. O ile nie stanowi ona w tym modelu, jak i we wnioskowaniu statystycznym koniecznego elementu *eksplanansa*, powraca jednak na etapie konstruowania hipotez dotyczących atrybutów/zmiennych wyjaśniających *eksplanandum*. W istocie bowiem hipotezy te są dobierane według intuicyjnego rozumienia regularności zachodzących w badanym obszarze, które mogą być też interpretowane jako przyczynowo-skutkowe *prawa natury*. Postulat rozkładu jednorodnego jest niemożliwy epistemicznie do osiągnięcia a brak tego rozkładu prowadzi w efekcie do obarczenia wy-

⁵Op. cit, StatSoft (2006). Elektroniczny Podręcznik Statystyki PL, Krakow, WEB: <<http://www.statsoft.pl/textbook/stathome.html>>.

jaśnienia ogromnym ryzykiem błędu. Ryzyko to jest szacowane, ale jego szacunek opiera się na doświadczeniu i intuicji badacza.

LITERATURA

- A. Grobler, *Metodologia nauk*, Znak, Kraków, 2006.
- J. Hozer, S. Kokot, W. Kuźmiński, *Metody analizy statystycznej rynków wycenie nieruchomości*, PFSRM, Warszawa 2002.
- James Woodward, *Scientific explanation*, <<http://plato.stanford.edu/entries/scientific-explanation>>.
- Stanisław M. Kot, Jacek Jakubowski, Andrzej Sokołowski, *Statystyka. Podręcznik dla studiów ekonomicznych*, Difin, Warszawa, 2007.
- StatSoft (2006). *Elektroniczny Podręcznik Statystyki PL*, Kraków, WEB: <<http://www.statsoft.pl/textbook/stathome.html>>.
- Statistical Explanation and Statistical Relevance*, W. Salmon, (ed.), 29–87, Pittsburgh: University of Pittsburgh Press.

SUMMARY

THE CASE FOR CRITIQUE OF STATISTICAL RELEVANCE MODEL OF SCIENTIFIC EXPLANATION

The statistical relevance model of scientific explanation was proposed by Wesley Salmon in 1971 as an interesting alternative to already existed models introduced by Hempel and supported by many other philosophers of science. The most important difference between the nomological models and statistical relevance model is that the latter tries not to use the very dubious term of “law of nature”. The first part of the paper consists of the overview of the Salmon’s model and of the main arguments which were raised by various authors against it. In the main part of the text all of those arguments which were meant to undermine the model are presented on an example taken from the economic practice. It is very popular among the economists and especially among valuation experts the so called “statistical analysis of the market”. The main objective of the analysis is to discover all of the factors which influence the market value of the particular product, in other words

to explain the market value of the product. The example was taken from the social science (economics) for purpose as one of the thesis in the paper is that, the SR model can work quite well in physics or chemistry, but it is dubious whether we can really deploy it in sciences which try to describe and explain the various phenomena of human activity and behavior. The final conclusions are:

The practical deployment of the model in social sciences are problematic, as it is too idealistic and therefore it doesn't work properly.

Against its initial presumption the model doesn't avoid the problem of laws of nature. Although the law of nature is not a required element of the *explanans*, it comes back at the stage of proposing the initial candidates for the relevant variables. The hypothesis on, which variables can be and which cannot be relevant to the explained phenomenon are constructed mostly according to the intuitively understood causal relationship founded on laws of nature.

The important postulate of homogenous partition is in practice unachievable what causes that the explanation is bound with the enormous risk of a mistake. The risk is quantifiable and can be estimated, but the estimation is depended upon experience and intuition of a researcher.