STATYSTYCZNY DROGOWSKAZ 3

Praktyczny przewodnik wykorzystania modeli regresji oraz równań strukturalnych

Sylwia Bedyńska
Monika Książek

SEDNO Wydawnictwo Akademickie

SWPS
SZKOŁA WYŻSZA PSYCHOLOGII SPOŁECZNEJ

STATYSTYCZNY DROGOWSKAZ 3

STATYSTYCZNY DROGOWSKAZ 3

Praktyczny przewodnik wykorzystania modeli regresji oraz równań strukturalnych

Sylwia Bedyńska
Monika Książek

SEDNO Wydawnictwo Akademickie

SWPS SZKOŁA WYŻSZA PSYCHOLOGII SPOŁECZNEJ

Wydawca: **Bożena Kućmierowska**
Recenzenci: **prof. dr hab. Magdalena Marszał-Wiśniewska, prof. dr hab. Grzegorz Sędek**
Redakcja merytoryczna i korekty: **Iwona Witt-Czuprzyńska**
Redakcja techniczna: **Danuta Przymanowska-Boniuk**
Projekt okładki, stron tytułowych i działowych: **Janusz Fajto**
Opracowanie typograficzne: **Wojciech Stukonis**

Publikacja jest wspólną inicjatywą wydawniczą Szkoły Wyższej Psychologii Społecznej
oraz Wydawnictwa Akademickiego Sedno

ISBN 978-83-63354-05-3
ISBN 978-83-62443-24-6
ISBN 978-83-62443-35-0 (tomy 1-3)
ISBN 978-83-63354-97-8 (e-book)

Wydawnictwo Akademickie Sedno Spółka z o.o.
ul. J. Pankiewicza 3
00-696 Warszawa
www.wydawnictwosedno.pl

Spis treści

Przedmowa

Truizmem jest twierdzenie, że znajomość metod statystycznych jest ważna. Ta powtarzana od dawna teza udowadniana jest empirycznie przez pokolenia młodych i starszych badaczy, którzy – niesieni na fali entuzjazmu związanego z planowaniem nowego badania, nie zaprzątają sobie zwykle głowy tak przyziemnymi sprawami jak późniejsza analiza statystyczna zebranych danych. Zdarza się, że badacze biorą po prostu kilka kwestionariuszy lub ankiet mierzących różne właściwości, które mogą być ze sobą skorelowane. Liczą, że skoro te skale pasują do siebie i do ogólnego tematu, to potem coś z tego wyjdzie. Jednak gdy przyjdzie moment liczenia wyników, okazuje się, że sprawa już taka łatwa nie jest. Brak jasnych hipotez i modelu teoretycznego skutkuje tym, że dane są analizowane długo i niestety bez większego efektu, a badacz tonie w morzu wyników. Często okazuje się też, że policzenie prostych korelacji nie wystarczy, by odpowiedzieć na pytania badawcze. Rzeczywistość jest niestety bardziej złożona niż proste korelacje i dopiero po skończeniu badania przychodzi refleksja: „ojej, przecież gdybym inaczej zmierzył tę zmienną, byłoby łatwiej policzyć wyniki".

Dodatkowo brak znajomości metod statystycznych dedykowanych danym ilościowym skutkuje tym, że badacze redukują zmienne ilościowe do zmiennych porządkowych, kategoryzując je według rozmaitych kryteriów. Niestety, zdarza się wciąż, że zamiast uwzględnić dokładny wynik w skali ekstrawersji, dokonuje się podziału medianowego na osoby ekstrawertywne i introwertywne. Czy powszechność tej praktyki przemawia na jej korzyść? Niestety nie. Symulacje wykonane przez S.E. Maxwella i H.D. Delaneya, mające na celu porównanie efektywności poszukiwania interakcji w regresji i w analizie wariancji, wskazują jednoznacznie, że dychotomizacja ciągłych predyktorów i ich późniejsza analiza za pomocą analizy wariancji powodują trudności w ujawnieniu efektów poszczególnych zmiennych (Maxwell, Delaney, 1993). Niestety, brak znajomości sposobu przeprowadzania analizy interakcji

w regresji powoduje, że część badaczy nadużywa dychotomizacji zmiennych ciągłych, by wykorzystać prostszą analizę wariancji. Takie rozwiązanie jest nie tylko niekorzystne pod kątem prawdy naukowej, ale także utrudnia publikację badań w prestiżowych czasopismach. Warto więc zadać sobie pytanie o metodę analizy danych już w momencie planowania badania, a nie po jego zakończeniu. Warto też rozszerzać własny arsenał technik statystycznych, tak by móc stosować odpowiednie narzędzia analityczne zgodne z naturą postawionego pytania badawczego zmiennych i struktury danych.

Konsultując od lat rozmaite projekty badawcze postanowiłyśmy zaproponować remedium na tego typu bolączki – drugi tom prostego podręcznika statystycznego, przybliżającego kolejny zestaw metod statystycznych, tym razem służących do analizy danych korelacyjnych. Naszym celem było rozszerzenie repertuaru technik statystycznych dostępnych dla studentów kierunków humanistycznych, doktorantów oraz młodych badaczy, którzy dzięki temu będą mogli poradzić sobie bardziej efektywnie i efektownie z analizą danych zebranych do prac magisterskich, doktorskich, czy wszelkich prac badawczych przygotowywanych w trakcie studiów.

W realizacji tego celu korzystamy z wcześniejszych doświadczeń zdobytych podczas przygotowywania poprzedniej książki pt. *Statystyczny drogowskaz. Praktyczny poradnik analizy danych w naukach społecznych na przykładach z psychologii* (Bedyńska, Brzezicka, 2007). Staramy się wspierać badacza w analizie danych z wykorzystaniem najbardziej popularnego pakietu statystycznego IBM SPSS Statistics. Dorzucamy informacje o tym, jak zapisać wyniki, choć należy pamiętać, że standardy często się zmieniają, a każdy promotor ma troszkę inne wymagania. To tylko propozycja, a nie nieśmiertelny kanon. Najważniejsze terminy znalazły się na marginesach, by ułatwić nawigowanie po książce. Nie zakładamy bowiem, że książka będzie czytana od deski do deski jednym tchem po nocach. Wiemy, że niekiedy czytelnik będzie po nią sięgał, by znaleźć bardzo konkretne informacje, a nie rozkoszować się narracją. Kolejnym elementem graficznym są ramki podsumowujące najważniejsze treści. Wydaje się, że sprawdziły się znakomicie w poprzedniej książce, więc wykorzystujemy je ponownie.

W tym tomie sięgamy po zagadnienia nieco bardziej złożone i poświęcamy go w całości metodom poszukiwania związku między zmiennymi. Podejmujemy więc temat analiz w tym miejscu, w którym został on przerwany w pierwszym *Statystycznym drogowskazie…* Ponownie omawiamy zatem regresję prostą z jednym predyktorem oraz wielokrotną z wieloma zmiennymi wyjaśnianymi. I poszerzamy znacznie zakres stosowania regresji o możliwość wprowadzania jakościowych predyktorów, zarówno dychotomicznych, jak i tych o większej liczbie kategorii. Wreszcie, pokazujemy jak testować efekty interakcyjne w regresji oraz poszukiwać zmiennych będących mediatorami (zmiennymi pośredniczącymi). Ta ostatnia kwestia jest bardzo ważna pod kątem budowania teorii naukowych. Dzięki informacji o tej klasie

zmiennych możemy powiedzieć, jak dany efekt działa, dlaczego istnieje zależność między pewnymi zmiennymi.

Najciekawszą częścią książki są naszym zdaniem rozdziały wprowadzające zagadnienia modelowania strukturalnego. Ta nowoczesna metoda staje się obecnie standardem, więc jej wprowadzenie wydało nam się bardzo interesujące. By ułatwić jej zrozumienie, w trzech rozdziałach przedstawiamy analizy, które są odpowiednikami wcześniej opisanych technik regresyjnych: regresji wielokrotnej, regresji z kowariancją i mediacji. Uznałyśmy, że taki układ ułatwi zapoznanie się z tą nieco bardziej złożoną metodą. Ze względu na zmianę narzędzia analitycznego z pakietu IBM SPSS Statistics na program AMOS dodałyśmy w ostatnim rozdziale także krótki przewodnik po tym programie. Niestety, nie wyczerpujemy możliwości wykorzystania modelowania równań strukturalnych, ponieważ metoda ta pozwala testować niezwykłe bogactwo układów zależności między zmiennymi ilościowymi i jakościowymi. Sądzimy jednak, że po takim wprowadzeniu, jakie proponujemy, dalsza eksploracja tej problematyki będzie znacznie ułatwiona.

Aby czytelnik mógł samodzielnie powtórzyć analizy pliki z danymi zostały umieszczone na dwóch stronach internetowych: www.wydawnictwosedno.pl oraz www.bedynska.com.pl. Proszę też pamiętać podczas czytania książki, że wartości liczbowe zostały zaokrąglone do drugiego miejsca po przecinku, więc mogą nie być identyczne jak w prezentowanych tabelach.

Na koniec najprzyjemniejsza rzecz. Chcemy podziękować osobom, bez których ta książka by nie powstała. Po pierwsze więc dziękujemy studentom za zadawanie tzw. głupich pytań, bez których nie znalazłybyśmy ciekawych odpowiedzi. Ich pytania zmusiły nas do zastanowienia, jak wyjaśnić przystępnie pozornie oczywiste zagadnienia. Chcemy także podziękować recenzentom książki – prof. Grzegorzowi Sędkowi i prof. Magdalenie Marszał-Wiśniewskiej, którzy czuwali, by nasza tendencja do upraszczania nie stała się karykaturalna. Ich wskazówki były bardzo cenne w naszej pracy nad książką. Podziękowania należą się także naszym „królikom doświadczalnym" – pierwszym czytelnikom książki – Marcie Koć-Januchcie, Rafałowi Albińskiemu, Magdzie Świrkuli. Dzięki nim szanse, że książka będzie zrozumiała, znacznie wzrosły. Nad stroną graficzną czuwali Tomasz Grzelka, Janusz Fajto i Wojciech Stukonis wraz z Pracownikami Wydawnictwa Akademickiego Sedno.

Sylwia Bedyńska
Szkoła Wyższa Psychologii Społecznej w Warszawie
Monika Książek
Szkoła Główna Handlowa w Warszawie

Część I
Modele regresji

ROZDZIAŁ

1

Regresja prosta

W tym rozdziale dowiemy się o tym:

- jaki jest wzór linii prostej – modelu regresji
- jak dopasowywana jest linia regresji oraz jakie jest znaczenie jej poszczególnych parametrów, w tym współczynnika beta
- jak przeprowadzić analizę regresji w programie IBM SPSS Statistics i zinterpretować oraz opisać uzyskane wyniki.

WPROWADZENIE

Poszukiwanie zależności między zmiennymi jest niezwykle ważnym elementem postępowania naukowego. Choć analiza korelacji nie ma takiej mocy jak poszukiwanie przyczyny i skutku w badaniach eksperymentalnych, to jednak pozwalając prześledzić wzajemne zależności dużej liczby zmiennych, przygotowuje podstawy do projektowania eksperymentów. Dzięki tej technice możliwe jest bowiem znaczące zawężenie zmiennych uwzględnianych potem w badaniach eksperymentalnych. Schemat korelacyjny może więc stanowić ważne źródło inspiracji dla eksperymentów, gdzie niemożliwe staje się uwzględnienie zbyt dużej liczby zmiennych jednocześnie. Oczywiście relacje badań eksperymentalnych i korelacyjnych są wzajemne – zidentyfikowane w eksperymencie kluczowe dla danej sfery zmienne mogą zostać następnie uwzględnione w badaniu korelacyjnym, które pozwala prześledzić bardziej skomplikowane relacje między konstruktami, a w konsekwencji – budowanie złożonych teorii naukowych.

ZMIENNA NIEZALEŻNA (OBJAŚNIAJĄCA) ZMIENNA ZALEŻNA (OBJAŚNIANA)

Skoro relacje są takie ważne, to analiza regresji stanowi istotne narzędzie odpowiadania na pytania badawcze o zależności zmiennych. W swej klasycznej postaci wymaga, by zarówno predyktory (zmienne niezależne czy objaśniające), jak i zmienna zależna (czy objaśniana) były ilościowe, ale jak pokażemy w jednym z rozdziałów, możliwe jest także uwzględnienie dychotomicznych predyktorów. Możemy je wprowadzać do regresji, dlatego że metoda ta jest bardziej ogólną techniką analityczną należącą do rodziny metod kryjących się pod nazwą Ogólnego Modelu Liniowego. Do tej samej grupy technik należą także testy *t*-Studenta i analiza wariancji, ale nie są one tak wszechstronne jak regresja. Ograniczenie dla regresji stanowi jednak liczba zmiennych zależnych – nie może ona przekroczyć jednej.

W tym rozdziale przedstawimy szczegółowo najprostszą analizę z wykorzystaniem jednej zmiennej niezależnej i jednej zmiennej zależnej. Dzięki temu, że model będzie tak prosty, możliwy się stanie bardzo szczegółowy i precyzyjny opis podstaw logicznych analizy regresji i sposobu interpretacji jej wyników. Zaczniemy jednak od statystyk opisowych, które pozwalają podsumować współzmienność dwóch zmiennych: kowariancji i korelacji *r* Pearsona. Następnie pokażemy na wykresach rozrzutu, jak wyglądają dane o określonych wartościach współczynnika korelacji *r* Pearsona. Opiszemy także metodę dopasowania linii regresji oraz interpretację parametrów opisujących tę linię. W ostatniej części rozdziału zaprezentujemy sposób wykonania obliczeń w programie IBM SPSS Statistics i zapis wyników w raporcie empirycznym.

KOWARIANCJA I KORELACJA JAKO MIARY WSPÓŁZMIENNOŚCI

By zaprezentować logikę analizy regresji, cofniemy się na chwilę do dwóch statystyk opisowych: **kowariancji i korelacji**. Ta pierwsza nie jest zbyt popularna, ale zrozumienie sensu jej obliczania jest niezbędne, by swobodnie korzystać z niej w znajdującym się w dalszej części książki modelowaniu strukturalnym. Kowariancję można uznać za prekursorkę korelacji, więc to, co teraz będziemy robić, to po trosze archeologiczne wykopaliska.

KOWARIANCJA

Kowariancja wykorzystuje wariancję wyników, czyli odległości wyników od średniej arytmetycznej. Opiera się na obserwacji, że jeśli dwie zmienne mają jakiś specyficzny układ wartości względem siebie, to przykładowo u danej osoby wynik powyżej średniej powinien współwystępować z wynikiem powyżej średniej w drugiej zmiennej. Możliwy jest też taki układ, że wynik poniżej średniej w obrębie jednej zmiennej współwystępuje u danej osoby z wynikiem powyżej średniej w obrębie drugiej zmiennej. A zatem kowariancja to inaczej współzmienność wyników dwóch zmiennych, którą szacujemy, sprawdzając, w jakim kierunku odchylają się wyniki obu zmiennych od odpowiednich średnich. Przykład obliczania kowariancji dla czterech wyników można znaleźć w tabeli 1.1.

Kroki obliczania kowariancji:

1. Obliczamy **średnie** dla obu zmiennych.
2. Odejmujemy wynik osoby w danej zmiennej od średniej dla tej zmiennej. Obliczamy więc **odległości** wyników w danej zmiennej od jej średniej.
3. Dla każdej osoby **mnożymy obie odległości** wyników zmiennych od ich średnich.
4. **Dodajemy do siebie iloczyny odległości** – to jest licznik kowariancji.
5. By uzyskać wartość kowariancji, dzielimy obliczoną w kroku 4 sumę przez liczbę obserwacji pomniejszoną o 1.

Jak w niej widać, obliczamy ją w kilku krokach. Najpierw musimy znaleźć średnie dla obu podsumowywanych zmiennych, następnie odnieść każdy wynik do tej średniej, odejmując wynik od średniej. Mnożymy tak uzyskane odległości dla każdej pary wyników i sumujemy je, uzyskując licznik kowariancji. Teraz już wystarczy tylko podzielić rezultat obliczeń przez liczbę wyników minus 1 i uzyskamy wartość kowariancji. W tym przykładzie będzie to wartość –2,5.

INTERPRETACJA KOWARIANCJI

No dobrze, policzyliśmy kowariancję, ale jak ją teraz zinterpretować? Niestety, poważnym ograniczeniem tej statystyki jest to, że **możemy jedynie określić kierunek zależności**. Ujemna wartość świadczy o tym, że niskie wartości jednej

Tabela 1.1. Kolejne kroki obliczania wielkości kowariancji dla zmiennych *X* oraz *Y*

Wartości zmiennej X	Wartości zmiennej Y	Odległość od średniej dla X	Odległość od średniej dla Y	Iloczyn odległości
1	5	–2	2	–4
2	4	–1	1	–1
3	3	0	0	0
4	2	1	–1	–1
5	1	2	–2	–4
średnia = 3	średnia = 3			suma: –10

zmiennej współwystępują z wysokimi drugiej zmiennej i odwrotnie, a dodatnie, że niskie wartości współwystępują z niskimi, a wysokie z wysokimi. Nie jesteśmy jednak w stanie określić, czy zależność między zmiennymi jest silna czy słaba. Dzieje się tak, dlatego że wielkość kowariancji zależy silnie od jednostek pomiarowych – będzie większa, gdy podamy wartość wzrostu w centymetrach, niż gdy będziemy ją obliczać na podstawie tych samych wartości, ale zapisanych w metrach. By pokonać tę trudność, Robert Pearson zaproponował współczynnik korelacji nazwany później współczynnikiem *r* Pearsona, który ze względu na to, że liczony jest dla wystandaryzowanych wyników, pozwala określić dwa aspekty relacji: siłę i kierunek.

Przyjrzyjmy się zatem **współczynnikowi korelacji *r* Pearsona**. Dla powyższych danych będzie on obliczany następująco: pierwszy krok jest kluczowy, bo zamiast odnosić wyniki obu zmiennych do ich średnich, standaryzujemy je, a więc podajemy odległość od średniej, ale w jednostkach odchylenia standardowego. Następnie postępujemy identycznie jak w przypadku obliczania kowariancji: mnożymy przez siebie pary wartości dla danej osoby, dodajemy te iloczyny do siebie i dzielimy przez liczbę osób badanych pomniejszoną o 1. Efektem tego jest wartość współczynnika *r* Pearsona wynosząca dokładnie –1. Kolejne kroki obliczania korelacji dla przykładowych danych przedstawia tabela 1.2.

Współczynnik korelacji *r* Pearsona może przyjmować wartości od –1 do 1. Znak współczynnika oznacza kierunek zależności – tak jak w przypadku kowariancji.

Kroki obliczania współczynnika korelacji *r* Pearsona:

1. Obliczamy średnie i odchylenia standardowe dla obu zmiennych.
2. Standaryzujemy wyniki każdej zmiennej, odejmując od każdego wyniku średnią i dzieląc tę różnicę przez odchylenie standardowe.
3. Dla każdej osoby mnożymy wystandaryzowane wyniki dla obu zmiennych.
4. Dodajemy do siebie iloczyny wystandaryzowanych wyników – to jest licznik współczynnika korelacji *r* Pearsona.
5. By uzyskać wartość korelacji, dzielimy obliczoną w kroku 4. sumę przez liczbę obserwacji pomniejszoną o 1.

Tabela 1.2. Kolejne kroki obliczania wielkości korelacji dla zmiennych *X* oraz *Y*

Wartości zmiennej *X*	Wartości zmiennej *Y*	Wystandaryzowana odległość od średniej dla *X* (*Xi* – M)/SD	Wystandaryzowana odległość od średniej dla *X* (*Xi* – M)/SD	Iloczyn odległości
1	5	–1,26	1,26	–1,6
2	4	–0,63	0,63	–0,4
3	3	0,00	0,00	0,0
4	2	0,63	–0,63	–0,4
5	1	1,26	–1,26	–1,6
średnia = 3 SD = 1,6	średnia = 3 SD = 1,6			suma: –4

Dodatkowo jednak możemy określić siłę zależności: im wartość współczynnika bliższa wartościom maksymalnym –1 oraz 1, tym silniejsza zależność. Gdy wartość współczynnika znajduje się blisko 0, wówczas mówimy, że nie ma współzależności, przy czym musimy pamiętać, że myślimy wtedy o zależność prostoliniowej – monotonicznej i proporcjonalnej (a więc o zmianie o identyczną liczbę jednostkę jednej zmiennej wraz ze zmianą drugiej zmiennej o jedną jednostkę). Tutaj mamy więc do czynienia z idealną korelacją ujemną, ponieważ współczynnik korelacji $r = –1$.

- **Kowariancja** pozwala określić jedynie kierunek zależności, ale nie siłę relacji. Wielkość kowariancji zależy silnie od jednostek pomiarowych.
- **Korelacja** umożliwia określenie zarówno kierunku, jak i siły zależności. Wielkość korelacji nie zależy od jednostek pomiarowych, bo przed policzeniem korelacji zmienne są standaryzowane.

Operacje w programie IBM SPSS Statistics (ANALIZA–KORELACJE––PARAMI), gdy wpiszemy te dane do edytora danych, potwierdzają poprawność wcześniejszych obliczeń (zob. tab. 1.3).

Zerknijmy teraz, jak taka zależność wygląda na wykresie rozrzutu, na którym na osiach *X* oraz *Y* umieszczone są wartości obu zmiennych. Aby wykonać wykres, wchodzimy do górnego menu programu IBM SPSS Statistics i wybieramy opcję WYKRESY–WYKRESY TRADYCYJNE–ROZRZUTU/PUNKTOWY. Domyślnie w oknie tym zaznaczony jest wykres PROSTY, a taki właśnie chcemy wykonać, więc klikamy przycisk DEFINIUJ, by określić, które zmienne przedstawimy na wykresie. Zmienną *X* umieszczamy na osi *X*, a zmienną *Y* na osi *Y*. Zwykle zmienną, którą traktujemy jako wyjaśnianą, umieszczamy na osi *Y*, a wyjaśniającą na osi *X*. Potwierdzamy chęć wykonania operacji przyciskiem OK i uzyskujemy wykres (zob. rys. 1.1).

Jak widać na rysunku 1.1, punkty układają się dokładnie na linii prostej, ponieważ mamy do czynienia z idealną korelacją ujemną. Biegną od lewego górnego rogu do dolnego prawego, bo korelacja jest ujemna. Dla dodatniej korelacji

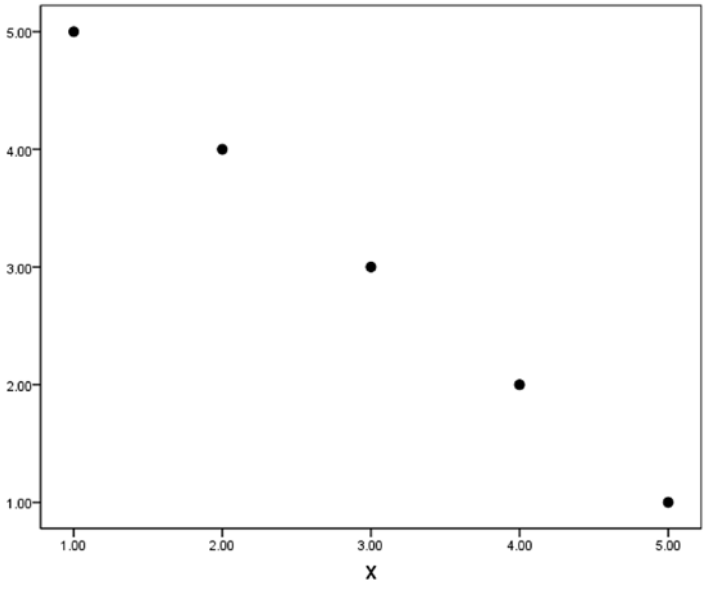
Tabela 1.3. Macierz korelacji dla zmiennych *X* oraz *Y*

Korelacje

		X	Y
X	Korelacja Pearsona	1	-1.000**
	Istotność (dwustronna)		.000
	N	5	5
Y	Korelacja Pearsona	-1.000**	1
	Istotność (dwustronna)	.000	
	N	5	5

**. Korelacja jest istotna na poziomie 0.01 (dwustronnie).

r = 1 punkty przebiegałyby po skosie od lewego dolnego do prawego górnego rogu. Jeśliby natomiast korelacja byłaby słabsza, punkty leżałyby coraz dalej od linii i przypominałyby raczej smugę niż idealny liniowy układ. Im wartość r Pearsona jest bliższa 0, tym bardziej punkty są bezładnie porozrzucane po obszarze wykresu. Pamiętajmy tylko o jednym ważnym zaleceniu: najpierw obejrzyjmy wykres rozrzutu, a potem liczmy współczynnik r Pearsona. Liczenie tej statystyki (a także, jak się zaraz okaże, analizy regresji) wymaga spełnienia założenia o liniowości relacji między zmiennymi. Muszą się więc one układać w linię prostą lub co najmniej smugę, nie mogą natomiast przypominać banana, litery „s" ani przyjmować innych zaokrąglonych kształtów.



Rysunek 1.1. Wykres rozrzutu dla zmiennych *X* oraz *Y*

JEDNOZMIENNOWA ANALIZA REGRESJI

Analiza regresji pozwala przeanalizować zależność między zmiennymi ilościowymi. W tym rozdziale przedstawimy wariant analizy regresji z jednym predyktorem i jedną zmienną zależną, by opisać szczegółowo kolejne kroki analizy i znaczenie parametrów (statystyk regresji). Należy jednak pamiętać, że taki wariant obliczeń jest obecnie rzadkością, ponieważ w większości przypadków badacz dysponuje większą liczbą predyktorów, których znaczenie dla zmiennej zależnej chce uwzględniać. Regresje jednozmiennowa i wielkozmiennowa mają wiele wspólnych elementów. W każdej z nich do danych dopasowywany jest model, ale w regresji jednozmiennowej jest to linia prosta, dwuzmiennowej – płaszczyzna, a trójzmiennowej – przestrzeń trójwymiarowa. Przy większej liczbie predyktorów nie sposób już sobie nawet wyobrazić modelu (choć oczywiście złośliwi twierdzą, że żaden matematyk nie ma problemu z wyobrażeniem sobie przestrzeni *n*-wymiarowej).

Kroki analizy regresji:

1. Dopasowanie modelu (tu: linii) metodą najmniejszych kwadratów.
2. Oszacowanie parametrów linii dla danych surowych (parametry niestandaryzowane: współczynnik nachylenia i stała) i standaryzowanych (współczynnik beta).
3. Określenie dobroci dopasowania modelu.

Zacznijmy więc od najprostszego wariantu, w którym do danych **dopasowujemy linię prostą za pomocą metody najmniejszych kwadratów.** Następnie podajemy parametry tej linii prostej w dwóch wariantach: dla danych surowych i dla danych wystandaryzowanych. Ta ostatnia statystyka, nazywana **współczynnikiem beta**, pozwala na interpretację zależności w kategoriach siły i kierunku, podobnie jak współczynnik *r* Pearsona. Ostatni krok pozwala na określenie, ile procent wariancji zmiennej zależnej wyjaśnia cały model. Dzięki tej informacji możliwe jest porównywanie różnych modeli między sobą, bez względu na liczebność próby, na której zostały obliczone.

DOPASOWANIE LINII REGRESJI METODĄ NAJMNIEJSZYCH KWADRATÓW

Pierwszym krokiem analizy regresji jest dopasowanie takiej linii prostej, która będzie spełniała jeden ważny warunek: odległości wyników od tej linii będą minimalne. Taka linia prosta może zostać nazwana linią najlepszego dopasowania.

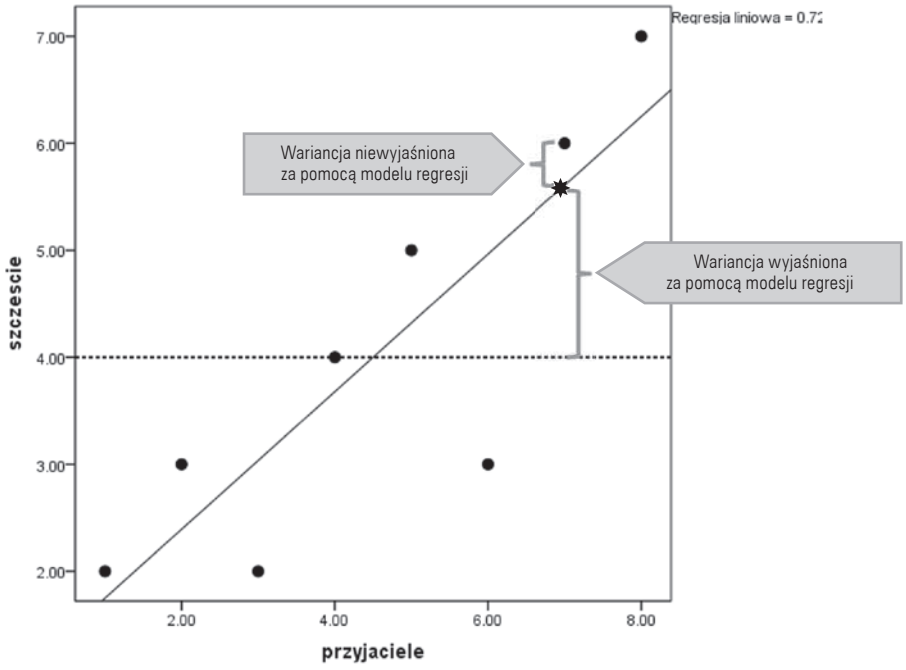
LINIA NAJLEPSZEGO DOPASOWANIA

Jak jednak statystycznie sprawdzić, czy linia jest dobrze dopasowana? Jeśli jesteśmy zainteresowani odległościami wyników od linii, to w sukurs przychodzi nam analiza wariancji, za pomocą której możemy określać wielkość łącznych odległości wyników od linii regresji. Przyjrzyjmy się jednak bardziej szczegółowo procesowi określania, czy linia jest dobrze dopasowana.

ANALIZA WARIANCJI

Punktem wyjścia analizy wariancji, która sprawdza poziom dopasowania linii, jest stwierdzenie, że jeśli nie mamy żadnego predyktora, to próbujemy przewidywać wyniki, posługując się średnią arytmetyczną dla zmiennej zależnej. Ten najprostszy model jest więc punktem odniesienia dla modelu bardziej złożonego – linii prostej. Zerknijmy na wykres rozrzutu na rysunku 1.2.

Będziemy przewidywać poczucie szczęścia na podstawie liczby przyjaciół. Dane do wykonania tego wykresu znajdują się w pliku *przyjaciele.sav.* Zobaczmy, że punkty są nieco oddalone od linii regresji. Te odległości od linii to różnica między wynikiem rzeczywistym a wynikiem przewidywanym przez model liniowy. Gdyby zależność była idealna i punkty *y* leżały dokładnie na linii, wtedy wynik przewidywany równałby się wynikowi rzeczywistemu. Tutaj jednak mamy pewną rozbieżność,



✷ Gwiazdką oznaczono wynik przewidywany.

Rysunek 1.2. Wykres rozrzutu dla zmiennej zależnej poczucie szczęścia (*szczęście*) i predyktora liczba przyjaciół (*przyjaciele*) z dopasowaną linią regresji (linia ciągła) i linią poziomą określającą wartość średniego poczucia szczęścia (linia przerywana)

bo przewidywanie nie jest stuprocentowo precyzyjne. Rozbieżność ta, czyli różnica między wynikiem rzeczywistym a przewidywanym przez model, nazywana jest resztą regresji. Reszty regresji określają wielkość błędu przewidywania, a ich wariancja może być traktowana jako składnik błędu. Czy jednak regresja pozwala lepiej przewidywać niż prostszy model bazujący na średniej arytmetycznej? By to sprawdzić, musimy policzyć, na ile poprawia się przewidywanie, gdy posługujemy się regresją, a więc odniesiemy wynik przewidywany przez regresję do średniej arytmetycznej w postaci wariancji wyników przewidywanych wobec średniej. Jeśli model regresji jest dobrym modelem, to wówczas wynik przewidywany stanowi lepsze przybliżenie rzeczywistego wyniku osoby badanej niż średnia arytmetyczna. Analiza wariancji odnosi do siebie te dwa składniki: wielkość wariancji wyjaśnionej za pomocą modelu regresji do wielkości wariancji niewyjaśnionej przez regresję, czyli wielkości reszt regresji.

RESZTA REGRESJI
BŁĄD PRZEWIDYWANIA
SKŁADNIK BŁĘDU

Analiza wariancji w regresji testuje, czy model jest dobrze dopasowany do danych. Porównuje wielkość wariancji wyjaśnianej przez regresję z prostszym modelem, jakim jest średnia arytmetyczna. Istotna analiza wariancji wskazuje, że model regresji lepiej wyjaśnia dane niż średnia arytmetyczna. Metoda ta nazywana jest metodą najmniejszych kwadratów, bo wariancja to nic innego jak średni kwadrat odległości wyników od średniej (zob. Bedyńska, Brzezicka, 2007: rozdz. 7).

Proporcja tych dwóch wariancji podawana jest w postaci statystyki F wraz ze stopniami swobody dla regresji (liczba wszystkich zmiennych, zależnych i niezależnych, minus 1) i stopniami swobody dla reszt (liczba wszystkich osób badanych pomniejszona o 1) oraz poziomem istotności, który pozwala stwierdzić, czy model regresji jest istotnie statystycznie lepszym sposobem przewidywania wyników niż średnia arytmetyczna. Analiza wariancji podaje także składniki niezbędne do oszacowania, ile procent wariancji (zmienności) zmiennej zależnej udaje się wyjaśnić za pomocą wprowadzonych predyktorów. Możliwe jest to dzięki określeniu proporcji sumy kwadratów dla regresji (oszacowania wariancji wyjaśnionej za pomocą regresji) do sumy kwadratów ogółem (oszacowania całkowitej wariancji). Statystyka, która podaje tę wartość, to statystyka R^2 obliczana poprzez podniesienie do kwadratu współczynnika korelacji wielokrotnej R – miary korelacji wszystkich predyktorów traktowanych łącznie ze zmienną zależną.

STATYSTYKA R^2
KORELACJA WIELOKROTNA R

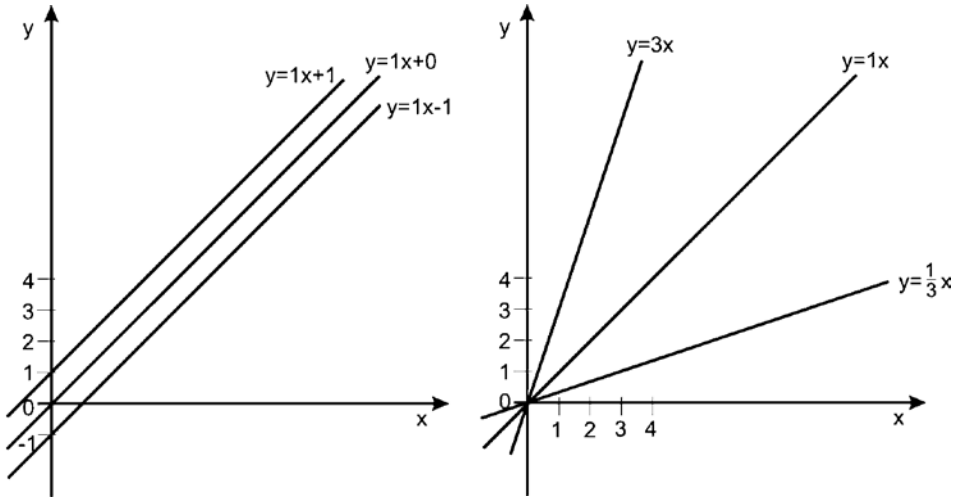
Współczynnik R^2 pomnożony przez 100% wskazuje, ile procent wariancji zmiennej zależnej (jej zmienności) wyjaśnia predyktor. Określa więc bardziej precyzyjnie **dobroć dopasowania** b niż istotność analizy wariancji.

RÓWNANIE LINII PROSTEJ – PARAMETRY MODELU

Skoro na podstawie wyników analizy wariancji zamieszczonych w regresji już wiemy, że udało się dopasować dobry model regresji do danych, to możemy przystąpić do określania dokładnego równania opisującego tę relację. Gdy mamy tylko jeden predyktor, modelem jest linia prosta z jednym *X*, którą można zapisać w postaci równania matematycznego $\hat{Y} = B_0 + B_1 \cdot X$. Taki zapis jest nieco odmienny od tego uczonego w szkole podstawowej, ale celowo podajemy taką właśnie postać linii regresji, ponieważ program IBM SPSS Statistics oznacza kolejne parametry linii prostej kolejno numerowanymi literami B. Opiszmy znaczenie symboli w tym równaniu. Symbol $\hat{Y}$ oznacza przewidywany wynik dla zmiennej zależnej, a *X* – wynik uzyskany dla predyktora. Bardzo ważne jest także, by pamiętać znaczenie obu parametrów równania. **Parametr B_0** zwany jest inaczej **stałą** i wyznacza punkt przecięcia przez linię regresji osi *Y*. Jeśli parametr ten wynosi 1, oznacza to, że linia regresji jest nieco powyżej początku układu współrzędnych; gdy wynosi –1 – to nieco poniżej początku układu współrzędnych (zob. rys. 1.3, wykres z lewej). **Parametr B_1 – współczynnik kierunkowy**, definiuje natomiast stopień nachylenia linii regresji względem osi *X*. Gdy przyjmuje wysoką wartość, to linia przebiega bardziej stromo, gdy niską – bardziej płasko. W sytuacji gdy współczynnik B_1 wynosi dokładnie 0, linia regresji jest równoległa do osi *X*, ponieważ w równaniu pozostaje jedynie stała i tylko ona determinuje przebieg linii (zob. rys. 1.3 z prawej).

PARAMETR B_0

PARAMETR B_1



Rysunek 1.3. Znaczenie parametrów linii: z lewej strony linie różnią się wartością stałej, z prawej wartością współczynnika nachylenia

Parametr B_0, nazywany **stałą**, określa punkt przecięcia linii z osią *Y*, a **parametr B_1**, nazywany **współczynnikiem nachylenia**, określa stopień nachylenia linii względem osi *X*.

Jak pewnie niektórzy zauważyli, zapisane powyżej równanie regresji obliczane jest na podstawie danych surowych, a więc w konsekwencji wielkość obu parametrów tego równania (stałej i współczynnika nachylenia) zależy od jednostek pomiarowych. Pojawia się więc tutaj ten sam problem jak w przypadku współczynnika kowariancji. Załóżmy, że chcemy przykładowo przewidywać wzrost mężczyzny na podstawie wzrostu jego ojca (to taki stary problem badawczy, który interesował między innymi Galtona w XIX wieku i przyczynił się od odkrycia regresji do średniej wskazującej, że synowie niskich ojców są wyżsi, a wysokich – niżsi) i podajemy wzrost za pomocą centymetrów, a drugim przypadku – w metrach. Parametry modelu będą miały wtedy wyższe wartości, gdy wzrost będzie mierzony w centymetrach. To powoduje, że nie możemy porównywać między sobą różnych modeli, posługując się parametrami dla danych surowych. Aby się pozbyć tej niedogodności, potrzebujemy uniwersalnej jednostki i takiej postaci linii regresji, w której będzie podany parametr podobnie uniwersalny co współczynniki *r* Pearsona. By poradzić sobie z tym problemem, musimy więc – podobnie jak podczas obliczania współczynnika *r* Pearsona, wystandaryzować wyniki, a następnie podać wzór linii regresji dla tak przekształconych danych. Konsekwencją tego przekształcenia jest redukcja stałej do 0 i zmiana wartości współczynnika B_1, który w tej postaci jest nazywany **współczynnikiem standaryzowanym beta.** Beta, tak jak współczynnik *r* Pearsona, może przyjmować wartości od –1 do 1; jego interpretacja jest identyczna jak współczynnika *r* Pearsona. Dzięki podanemu na wydruku poziomowi istotności uzyskujemy także informację, czy współczynnik ten jest równy 0, czy też odbiega istotnie od 0. Jeśli odbiega, oznacza to istotną relację między predyktorem a zmienną wyjaśnianą, którą możemy interpretować w kategoriach siły i kierunku zależności.

WSPÓŁCZYNNIK STANDARYZOWANY BETA

Podsumujmy więc kolejne składowe analizy regresji. W analizie regresji jednozmiennowej modelem jest linia prosta, która może zostać opisana za pomocą równania linii regresji. Równanie to ma postać $\hat{Y} = B_0 + B_1 \bullet X$, gdzie współczynnik

Parametry B_0 oraz **B_1** są obliczane dla danych surowych, więc ich wartości zależą od jednostek pomiaru. Pozwalają obliczyć wynik przewidywany dla danej osoby, ale nie nadają się do porównywania różnych modeli. By porównywać modele, posługujemy się bardziej uniwersalnym parametrem beta, który został obliczony dla danych standaryzowanych. Interpretujemy jego wartość tak jak wartość współczynnika *r* Pearsona.