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

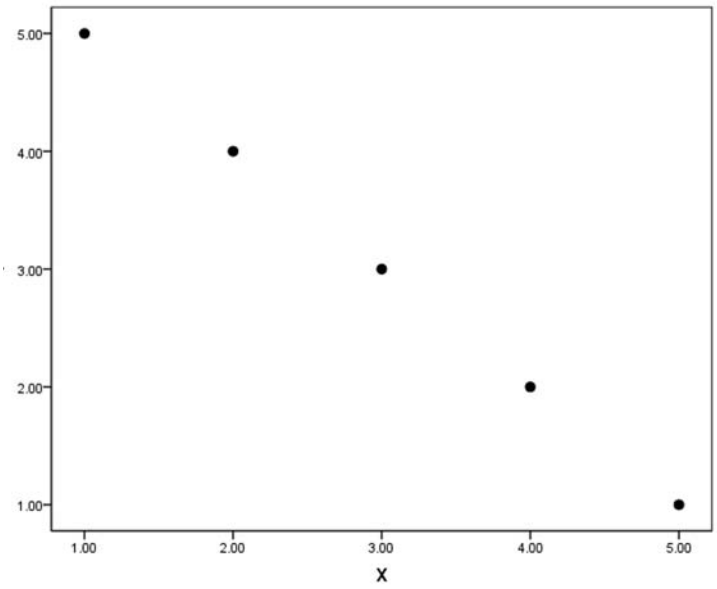
Tabela 1.3. Macierz korelacji dla zmiennych *X* oraz *Y*

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

LINIA NAJLEPSZEGO DOPASOWANIA

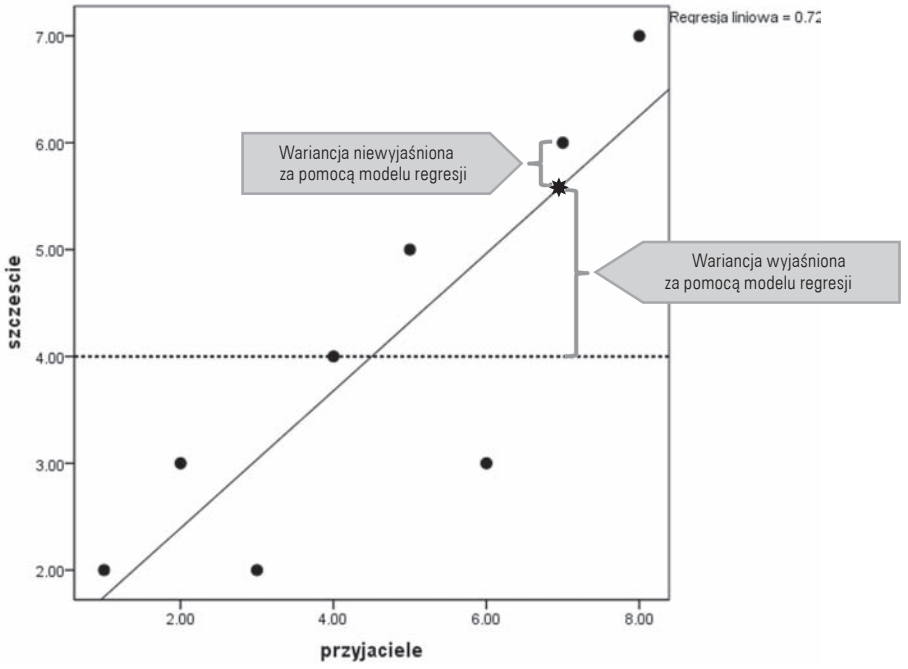
Pierwszym krokiem analizy regresji jest dopasowanie takiej linii prostej, która będzie spełniała jeden ważny warunek: odległości wyników od tej linii będą minimalne. Taka linia prosta może zostać nazwana linią najlepszego dopasowania.

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

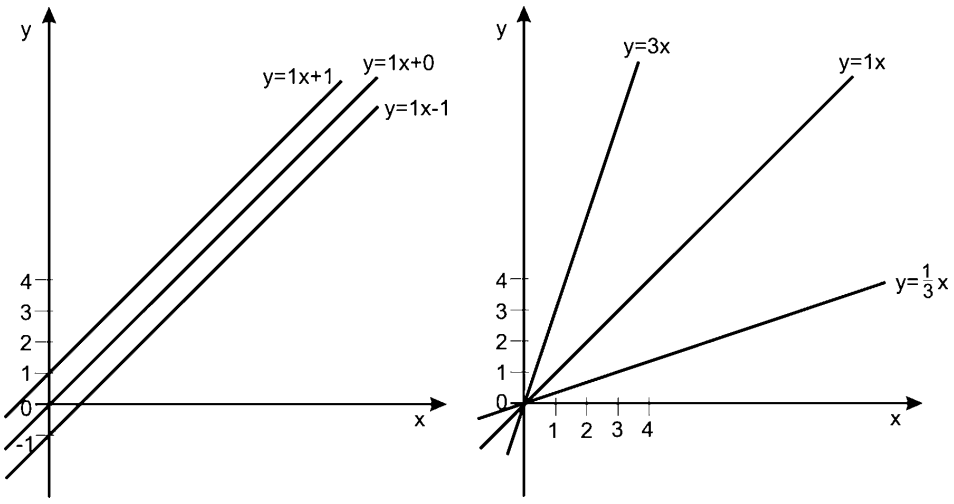
Współczynnik R^2 pomnożony przez 100% wskazuje, ile procent wariancji zmiennej zależnej (jej zmienności) wyjaśnia predyktor. Określa więc bardziej precyzyjnie **dobroć dopasowania** b niż istotność analizy wariancji.

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 $\hat{Y} = B_0 + B_1 \cdot X$, 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.

B_0 (stała) określa punkt przecięcia z osią *Y*, a współczynnik B_1 pozwala stwierdzić, jak mocno nachylona jest linia regresji w stosunku do osi *X*. Poziom dopasowania tej linii do danych określany jest za pomocą analizy wariancji, która sprawdza, czy reszty regresji nie są większe niż odległość wyniku przewidywanego od średniej. Jeśli analiza wariancji jest istotna, to wiemy, że regresja wyjaśnia lepiej wyniki niż zwykła średnia arytmetyczna i wtedy jest sens interpretowania współczynników linii regresji. Współczynniki niestandaryzowane B_0 i B_1 przydatne są wtedy, gdy chcemy obliczyć wartość przewidywaną dla danej osoby ($\hat{Y}$), znając jej wynik w zakresie zmiennej wyjaśniającej (*X*). Wówczas w równaniu w miejsce *X* podstawiamy znaną wartość, mnożymy ją przez wartość współczynnika nachylenia (B_1), dodajemy wartość stałej (B_0) i mamy wynik przewidywany. Jednak do interpretacji wyników ważniejszy jest współczynnik beta, który pozwala określić, jak silna jest relacja między zmiennymi i jaki jest jej kierunek. Tylko ten współczynnik pozwala na porównywanie wyników z różnych modeli. W jednozmiennowej regresji przyjmuje on identyczną wartość jak współczynnik korelacji wielokrotnej *R*. I wreszcie ostatnia statystyka: R^2, która informuje o tym, ile procent wariancji wyników zmiennej zależnej wyjaśnia cały model. Ta statystyka ma szczególne znaczenie w regresji wielozmiennowej, ponieważ określa efektywność całego modelu ze wszystkimi predyktorami łącznie, ale tutaj – w modelu regresji prostej niewiele nowego wnosi.

ANALIZA REGRESJI W PROGRAMIE IBM SPSS STATISTICS

Skoro już wiemy, jakich informacji szukać na wydruku programu IBM SPSS Statistics i jakie jest znaczenie poszczególnych statystyk, możemy przystąpić do zaprezentowania, jak po kolei wykonać obliczenia w programie IBM SPSS Statistics. Analiza regresji powinna być przeprowadzana w kilku krokach. Zanim ją wykonamy, warto przyjrzeć się trochę statystykom opisowym w postaci współczynnika *r* Pearsona i wykresom rozrzutu prezentującym relację między zmienną wyjaśniającą i wyjaśnianą. Dzięki wykresowi możemy choćby wstępnie stwierdzić, czy postulowany przez nas liniowy model relacji jest właściwy oraz czy są szanse na uzyskanie istotnej relacji między zmiennymi. Gdy współczynnik *r* Pearsona nie jest istotny, analiza regresji nie dokona cudu i nie przyniesie informacji o istotnej relacji. Należy wtedy szukać innych zmiennych wyjaśniających zmienną zależną.

Gdy już wykonamy te wstępne kroki i okaże się, że wykres rozrzutu pokazuje smugę punktów układających się w kształt linii prostej, a współczynnik *r* Pearsona jest w miarę wysoki i istotny, możemy przystąpić do wykonania analizy regresji. Spróbujmy przewidzieć poczucie szczęścia na podstawie liczby przyjaciół, posługując się danymi *przyjaciele.sav.* Wchodzimy więc do górnego menu, klikamy na opcje

Tabela 1.4. Fragment wydruku analizy regresji z wynikami analizy wariancji określającej dopasowanie modelu regresji

Anova[b]

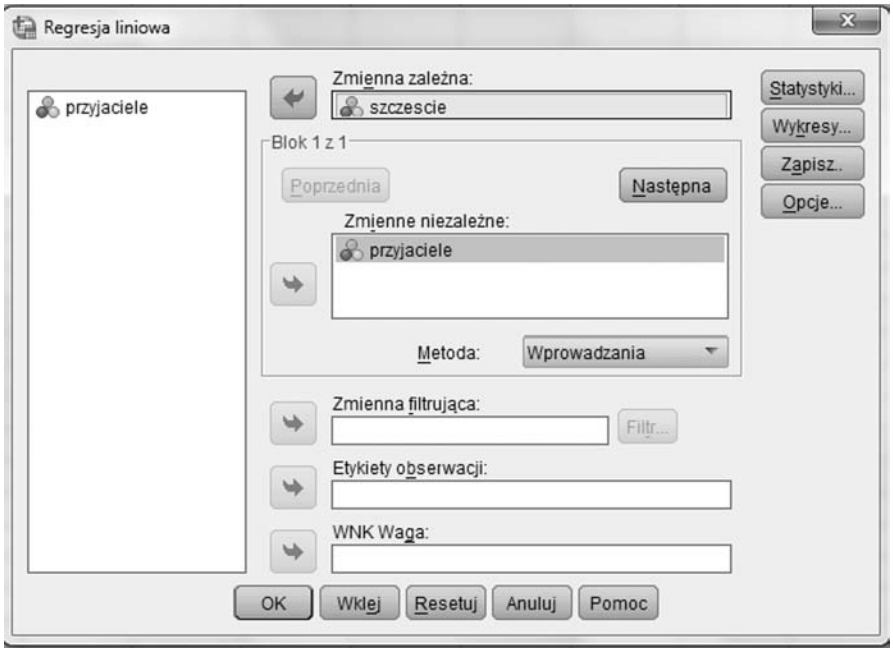
Model		Suma kwadratów	df	Średni kwadrat	F	Istotność
1	Regresja	17.357	1	17.357	15.677	.007[a]
	Reszta	6.643	6	1.107		
	Ogółem	24.000	7			

a. Predyktory: (Stała), przyjaciele
b. Zmienna zależna: szczescie

ANALIZA, a następnie wybieramy z listy REGRESJA–LINIOWA. Teraz umieszczamy zmienną *przyjaciele* jako niezależną, a zmienną *szczęście* jako zależną (zob. rys. 1.4) i potwierdzamy chęć wykonania obliczeń przyciskiem OK.

Po wykonaniu tej operacji w edytorze raportów pojawia się wiele tabel. Niestety, nie są ułożone w odpowiedniej kolejności, więc odczytywanie wyników musimy zacząć od tabeli trzeciej ANOVA prezentującej wyniki analizy wariancji (tab. 1.4).

W tabeli ANOVA zaprezentowane są statystyki pozwalające określić, czy udało się dopasować taką linię prostą – czy bardziej ogólnie: taki model, by wyjaśniał on więcej niż średnia arytmetyczna. Statystyka *F* powstaje z proporcji średniego kwadratu regresji do średniego kwadratu reszty, czyli oszacowań odległości



Rysunek 1.4. Okno dialogowe regresji liniowej pozwalające zdefiniować zmienne w modelu

Tabela 1.5. Współczynniki modelu regresji

Współczynniki[a]

Model		Współczynniki niestandaryzowane		Współczynniki standaryzowane	t	Istotność
		B	Błąd standardowy	Beta		
1	(Stała)	1.107	.820		1.350	.226
	przyjaciele	.643	.162	.850	3.959	.007

a. Zmienna zależna: szczescie

Tabela 1.6. Podsumowanie modelu – wartość współczynnika korelacji wielokrotnej i statystyki R^2

Model - Podsumowanie

Model	R	R-kwadrat	Skorygowane R-kwadrat	Błąd standardowy oszacowania
1	.850[a]	.723	.677	1.05221

a. Predyktory: (Stała), przyjaciele

wyników przewidywanych przez regresję od średniej oraz reszt regresji, czyli odległości wyników rzeczywistych od przewidywanych przez regresję. Tutaj widzimy, że wynik analizy wariancji jest istotny statystycznie i wobec tego możemy uznać, że model jest dobrze dopasowany, bo wyjaśnia więcej zmienności zmiennej zależnej niż jej średnia arytmetyczna. Statystyki tej analizy zapisujemy następująco: $F(1, 6) = 15{,}68$; $p < 0{,}01$. Pamiętajmy, że w nawiasie zamieszczamy dwa rodzaje stopni swobody: jako pierwsze stopnie swobody dla regresji, po przecinku – stopnie swobody dla reszty. Stopnie swobody ogółem można wtedy samodzielnie obliczyć, dodając do siebie te pozostałe dwa rodzaje stopni swobody dla regresji i dla reszty. Skoro udało się dopasować model do danych, to możemy przejść do kolejnej tabeli WSPÓŁCZYNNIKI (tab. 1.5).

W tabeli możemy znaleźć oba typy współczynników: niestandaryzowane i standaryzowane oraz statystyki *t* określające istotność tych współczynników. Zacznijmy od współczynników niestandaryzowanych: stała wynosi w tym modelu 1,11, a współczynnik kierunkowy 0,64. Możemy więc zapisać równanie regresji w postaci: $\hat{Y} = 1{,}11 + 0{,}64 \cdot X$. Dzięki temu obliczamy, ile punktów w skali szczęścia miałaby osoba mająca 10 przyjaciół, podstawiając wartość 10 w miejsce X do równania: $\hat{Y} = 1{,}11 + 0{,}64 \cdot 10 = 7{,}51$. Znając wartość stałej, możemy także łatwo powiedzieć, jaki wynik w skali szczęścia będzie miała osoba, która w ogóle nie ma przyjaciół – będzie to wartość równa stałej, ponieważ po podstawieniu wartości 0 zamiast X obliczony $\hat{Y}$ będzie równy wartości stałej, czyli 1,11. Współczynnik nachylenia informuje nas o tym, o ile wzrośnie poziom szczęścia wraz z każdą

kolejną zaprzyjaźnioną osobą. Gdy liczba przyjaciół wzrośnie o 1, poczucie szczęścia będzie wyższe o 0,64 punktu w skali szczęścia, którą zastosowaliśmy do pomiaru.

Wracamy teraz do tabeli WSPÓŁCZYNNIKI (tab. 1.5). W trzeciej kolumnie zamieszczone są błędy standardowe obu parametrów niestandaryzowanych. Gdy podzielimy wartość współczynnika przez jego błąd standardowy, to uzyskamy wartość statystyki *t* znajdującej się w piątej kolumnie. Istotność znajdująca się w końcowej kolumnie tabeli informuje, czy wartość współczynnika jest różna od 0. A więc gdy istotność jest mniejsza niż 0,05, to możemy zamieszczać ten współczynnik we wzorze linii regresji. Wiemy wtedy również, że współczynnik standaryzowany beta różni się od 0, a więc istnieje istotna relacja między zmiennymi. Tutaj współczynnik beta wynosi 0,85, jego istotność $p < 0{,}05$ (mniejsza od 0,05), więc konkludujemy, że relacja jest istotna, silna i dodatnia. Oznacza to tym samym, że im więcej mamy przyjaciół, tym bardziej jesteśmy szczęśliwi. Pamiętajmy jednak, że regresja nie pozwala określić relacji przyczynowo-skutkowych, więc możemy jedynie wskazać, że relacja między liczbą przyjaciół i poczuciem szczęścia ujawnia się, ale nie wiemy, czy to przyjaciele powodują, że czujemy się szczęśliwi, czy też gdy czujemy się szczęśliwi, to łatwiej się zaprzyjaźniamy i mamy więcej przyjaciół.

I wreszcie ostatnia tabela: MODEL–PODSUMOWANIE prezentująca wartości współczynnika korelacji wielokrotnej *R*, wartość statystyki R^2, jego skorygowaną wersję oraz błąd standardowy oszacowania (zob. tab. 1.6). Współczynnik korelacji wielokrotnej może przyjmować jedynie dodatnie wartości, tutaj w regresji jednozmiennowej jest tożsamy z wartością współczynnika beta i współczynnika *r* Pearsona. Jednak interpretujemy wartość R^2, która przemnożona przez 100 informuje o tym, jaki procent wariancji zmiennej zależnej wyjaśnia zmienna niezależna. W powyższym przykładzie uzyskaliśmy znakomitą moc przewidywania, ponieważ udaje nam się wyjaśnić za pomocą liczby przyjaciół aż 72% zmienności poczucia szczęścia.

Wyniki analizy regresji zapisujemy w takiej kolejności, w jakiej omawiałyśmy je w tym rozdziale. Zbierzmy więc teraz wszystkie informacje w raporcie. Pamiętajmy: w pierwszym zdaniu raportu zwykle wskazuje się rodzaj analiz, jakie zostały wykonane, a każdy wniosek musi zostać poparty odpowiednimi statystykami. Piszemy więc:

> Wykonano analizę regresji jednozmiennowej, w której zmienną wyjaśnianą było poczucie szczęścia, a zmienną wyjaśniającą liczba przyjaciół. Zaproponowany model regresji okazał się dobrze dopasowany do danych $F(1, 6) = 15{,}68$; $p < 0{,}01$. Na podstawie współczynników regresji można stwierdzić, że liczba przyjaciół jest silnie i pozytywnie powiązana z poczuciem szczęścia (beta = 0,85, $p < 0{,}01$). Oznacza to, że osoba mająca dużą liczbę przyjaciół przejawia także wysoki poziom szczęścia. Równanie regresji można zapisać w postaci $Y = 1{,}11 + 0{,}64 \bullet X$. Testowany model wyjaśnia aż 72% zmienności zmiennej zależnej.

PRZYKŁAD: RELACJA TEMPERATURY I SAMOPOCZUCIA

Pokażemy teraz jeszcze jeden przykład regresji, by wskazać możliwe trudności ze stosowaniem tego typu analiz. Poddajmy analizie samopoczucie psychofizyczne osób w zależności od temperatury otoczenia. Badania takie są także domeną psychologów, bo wszelkie czynniki determinujące efektywność pracy i dobrostan psychiczny mogą być uwzględniane w badaniach psychologicznych. Otwórzmy dane *samopoczucie.sav* i wykonajmy próbę przewidywania wyników zmiennej *samopoczucie* na bazie zmiennej *temperatura*.

Obliczmy wartość współczynnika *r* Pearsona (ANALIZA–KORELACJE–PARAMI), żeby sprawdzić, czy istnieje szansa na uzyskanie istotnych wyników w analizie regresji. Wydruk współczynnika *r* Pearsona pokazuje, że korelacja jest wysoka, dodatnia i istotnie różni się od 0 (tab. 1.7).

Zachęceni takim wynikiem wykonujemy analizę regresji ANALIZA–REGRESJA––LINIOWA i oglądamy kolejne tabele wydruku – na początek dopasowanie linii regresji do danych (tab. 1.8). Na podstawie wyników analizy wariancji możemy stwierdzić, że udało się dobrze dopasować linię regresji do danych $F(1, 9) = 17{,}97$; $p < 0{,}01$.

Współczynnik R^2 informuje nas, że wprowadzony predyktor wyjaśnia prawie 67% wariancji, czy inaczej: zmienności zmiennej wyjaśnianej (tab. 1.9).

I wreszcie oglądamy współczynnik standaryzowany beta w tabeli WSPÓŁCZYNNIKI (tab. 1.10). Współczynnik standaryzowany beta wynoszący 0,82 potwierdza całkowicie wynik wcześniejszej analizy eksploracyjnej wykonanej z wykorzystaniem współczynnika *r* Pearsona – zmienna temperatura jest silnie i dodatnio powiązana z samopoczuciem badanych. Oznacza to, że pozytywne samopoczucie jest powiązane z wysoką temperaturą.

Czy na pewno? W całym toku analizy pominęłyśmy jeden istotny, często pomijany etap analizy – wykonanie wykresu rozrzutu. Zrobiłyśmy to celowo, by pokazać, jak ważny jest ten drobiazg. Wspomniałyśmy powyżej, że jednozmiennowa analiza regresji bazuje na modelu linii prostej. Taki właśnie model jest dopasowywany do danych. Jednak nie zawsze zależność musi być prostoliniowa. Niekiedy zależność powinna być modelowana za pomocą krzywej, na przykład U-kształtnej, kiedy zarówno bardzo niskie, jak i bardzo wysokie wyniki w obrębie zmiennej niezależnej są powiązane z wysokimi wynikami w zakresie zmiennej zależnej czy N-kształtnej, gdy wysokie wyniki w zmiennej zależnej odpowiadają przeciętnym wynikom zmiennej niezależnej. Ten ostatni model N-kształtny adekwatnie opisuje prawo Yerkesa-Dodsona wskazujące, że zarówno zbyt niski, jak i zbyt wysoki poziom stresu jest niekorzystny dla efektywności funkcjonowania. Można sądzić, że większość takich zależności, w których jest pewien optymalny poziom (ani zbyt niski, ani zbyt wysoki), to zależności krzywoliniowe (kwadratowe, o kształcie paraboli). Także

Tabela 1.7. Współczynnik *r* Pearsona dla zmiennych *samopoczucie* oraz *temperatura*

Korelacje

		temperatura	samopoczucie
temperatura	Korelacja Pearsona	1	.816**
	Istotność (dwustronna)		.002
	N	11	11
samopoczucie	Korelacja Pearsona	.816**	1
	Istotność (dwustronna)	.002	
	N	11	11

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

Tabela 1.8. Wyniki analizy wariancji testującej istotność dopasowania modelu regresji dla relacji samopoczucia i temperatury powietrza

Anova[b]

Model		Suma kwadratów	df	Średni kwadrat	F	Istotność
1	Regresja	27.500	1	27.500	17.966	.002[a]
	Reszta	13.776	9	1.531		
	Ogółem	41.276	10			

a. Predyktory: (Stała), temperatura
b. Zmienna zależna: samopoczucie

Tabela 1.9. Tabela ze współczynnikami dopasowania regresji

Model - Podsumowanie

Model	R	R-kwadrat	Skorygowane R-kwadrat	Błąd standardowy oszacowania
1	.816[a]	.666	.629	1.23721

a. Predyktory: (Stała), temperatura

Tabela 1.10. Współczynniki regresji dla przewidywania samopoczucia na podstawie temperatury

Współczynniki[a]

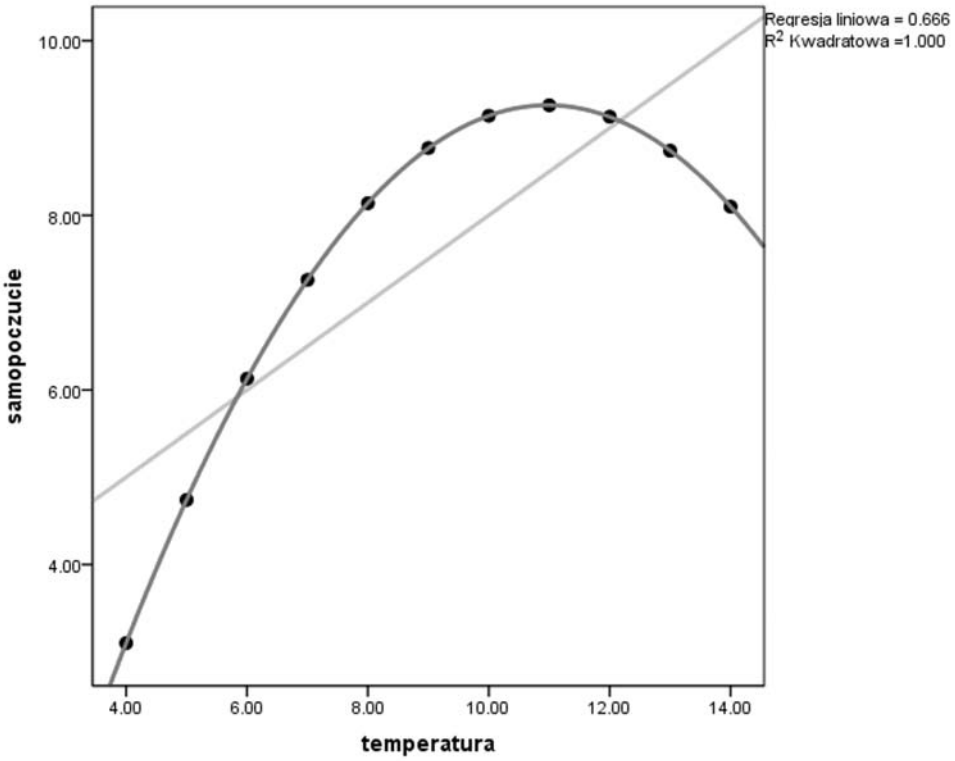
Model		Współczynniki niestandaryzowane		Współczynniki standaryzowane	t	Istotność
		B	Błąd standardowy	Beta		
1	(Stała)	3.001	1.125		2.667	.026
	temperatura	.500	.118	.816	4.239	.002

a. Zmienna zależna: samopoczucie

w tym przypadku – gdy chcemy przewidywać samopoczucie na bazie informacji o temperaturze otoczenia, możemy sugerować, że jest pewne optimum temperatury w obrębie jej przeciętnych wartości, a zarówno zbyt niskie, jak i zbyt wysokie temperatury będą odbierane przez badanych jako niekomfortowe i tym samym będą obniżały ich samopoczucie. Przecież prawie nikt nie lubi marznąć, ani się piec w upale, choć oczywiście optimum jest dla każdego trochę inne.

Wróćmy więc teraz do analizowanego przykładu i wykonajmy wykres rozrzutu, by przekonać się, czy nasze podejrzenia co do kształtu zależności są słuszne (WYKRESY TRADYCYJNE–ROZRZUTU/PUNKTOWY–PROSTY–DEFINIUJ). Na osi *Y* umieszczamy zmienną *samopoczucie*, na osi *X* zmienną wyjaśniającą *temperatura* i uzyskujemy wykres przedstawiony na rysunku 1.5.

Jak widać, po dopasowaniu linii prostej okazuje się, że nie jest to najlepszy model, choć wyjaśnia 66% wariancji; gdy dodamy na wykresie krzywą, uzyskujemy znacznie wyższą wartość R^2 wskazującą na idealne dopasowanie linii do danych. Oczywiście, dane są całkowicie fikcyjne, ale dzięki temu znakomicie obrazują znaczenie krzywoliniowości w analizie regresji.



Rysunek 1.5. Wykres rozrzutu dla zmiennych *temperatura* i *samopoczucie* z dopasowaną linią prostą i funkcją kwadratową

PODSUMOWANIE

Analiza regresji jest techniką analizy danych, która pozwala modelować dane, dopasowując linię prostą. Dzięki temu możliwe staje się opisanie relacji między zmiennymi za pomocą równania regresji z dwoma parametrami: stałą i współczynnikiem nachylenia (kierunkowym). Interpretacja siły i zależności między zmiennymi wykonywana jest na podstawie wartości współczynnika standaryzowanego beta, na bazie którego można określić siłę i kierunek zależności między zmienną wyjaśniającą oraz wyjaśnianą. Poważnym ograniczeniem tej metody jest to, że model prostoliniowy może nie być adekwatnym modelem, dlatego że dane układają się w kształt krzywej. Najprostszą metodą zdiagnozowania takiego stanu jest wykonanie wykresu rozrzutu, na którym widać przebieg rzeczywistych wyników.