METODYKA

January Weiner
Zakład Ekologii Zwierząt
Instytut Biologii Środowiskowej
Uniwersytet Jagielloński
ul. Karasia 6
30-060 Kraków

Korelacja i regresja, czyli o szkodliwości kalkulatorów

Correlation and regression or on mischievousness of calculators

1. Wstęp

Odkąd ekologia istnieje, uporczywie powtarzają się nawoływania do precyzyjnej kwantyfikacji, formalizacji matematycznej (modelowanie matematyczne), a przynajmniej eleganckiej analizy statystycznej danych. Nic więc dziwnego, że statystyczne metody analizy korelacji i regresji znajdują szerokie zastosowanie w ekologii, która jest wszak nauką o współzależnościach w przyrodzie. Obliczenia statystyczne są jednak dość zawiłe, pracochłonne, podatne na pomyłki. Trudności te są wystarczające, aby każdą próbę zastosowania żmudnych rachunków poprzedzić głębokim namysłem. Zawiłość trudnych do zapamiętania wzorów każe sięgać do podręczników, w których obok przepisu wykonania obliczeń znajduje się zazwyczaj rozdział teoretyczny, zawierający różne zastrzeżenia i deklaracje koniecznych do spełnienia założeń, zachęcający do namysłu jeszcze głębszego.

Tak było dawniej. Dziś mamy komputery i kalkulatory programowalne. Obliczenia statystyczne wykonuje się przyciskając guzik, a całą teorię zastępuje instrukcja wczytywania danych i naciskania klawiszy. W pracach ekologicznych i pokrewnych (na przykład ewolucyjnych, zwłaszcza dotyczących tzw. strategii życiowych) roi się więc od regresji i współczynników korelacji, po których następuje zazwyczaj sążnista dyskusja i daleko idące wnioski (albo nic nie następuje, jeżeli obliczenia miały charakter rytualno-dekoracyjny). Autor dobrze wie co mówi, bo sam posiada taki kalkulator i z upodobaniem i niezmąconym spokojem wiele takich rachunków sam popełnił, lub, co gorsza, doradził kolegom. Nie bez znaczenia jest też fakt, że najczęściej używane przez biologów podręczniki statystyki koncentrują się na doświadczalnictwie, pobieżnie traktując zagadnienia statystyki opisowej. Jak to jest z eksperymentem w ekologii i ewolucjonizmie nie trzeba nikomu wyjaśniać.

Tymczasem jednak podnoszą się głosy protestu. W ostatnich latach ukazało się wiele prac zwracających uwagę na bezkrytyczne, a nawet błędne stosowanie analizy statystycznej do badań współzależności biologicznych. Spośród grzechów

różnych (np. nieumiejętnego stosowania transformacji logarytmicznej, por Smith 1980, Heusner 1982) krytycy zwracają szczególną uwagę na nierozróżnianie analizy regresji od analizy korelacji. Zalecają też poprawne procedury, które trudno znaleźć w podręcznikach (Harvey i Mace 1982, Seim i Saether 1983, II wydanie „Biometry" Sokala i Rohlfa 1981).

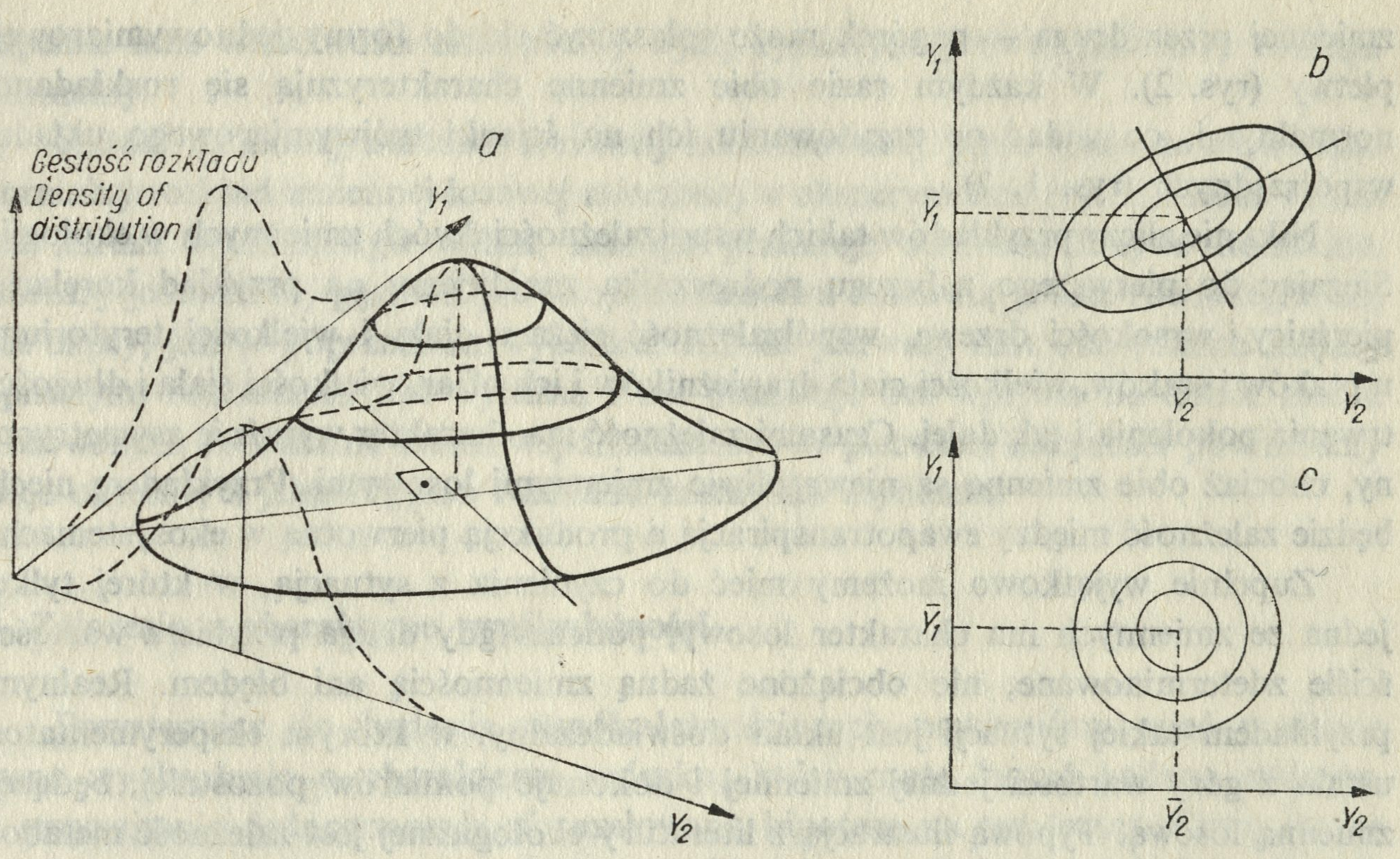
Wobec braku nowoczesnych podręczników statystyki dla biologów w języku polskim (jeżeli nie liczyć książek o zakresie elementarnym) oraz znacznego rozproszenia prac metodycznych w literaturze, zdecydowałem się na napisanie tego artykułu. Moim celem nie jest wyjaśnienie wszystkich niuansów statystycznej analizy współzależności biologicznych, bowiem zdaję sobie sprawę z własnego dyletantyzmu w tej materii. Z tego samego powodu wybrałem metodę obrazkową raczej niż formalizm matematyczny, któremu i tak bym nie sprostał. Jeżeli jednak uda mi się pogorszyć samopoczucie czytelników, stosujących dotąd beztrosko analizy regresji i korelacji, choćby w tym stopniu, w jakim pogorszyło się moje samopoczucie po przeczytaniu wyżej cytowanych tekstów krytycznych, zadanie swoje uznam za wypełnione.

2. Charakter rozkładu dwuwymiarowego

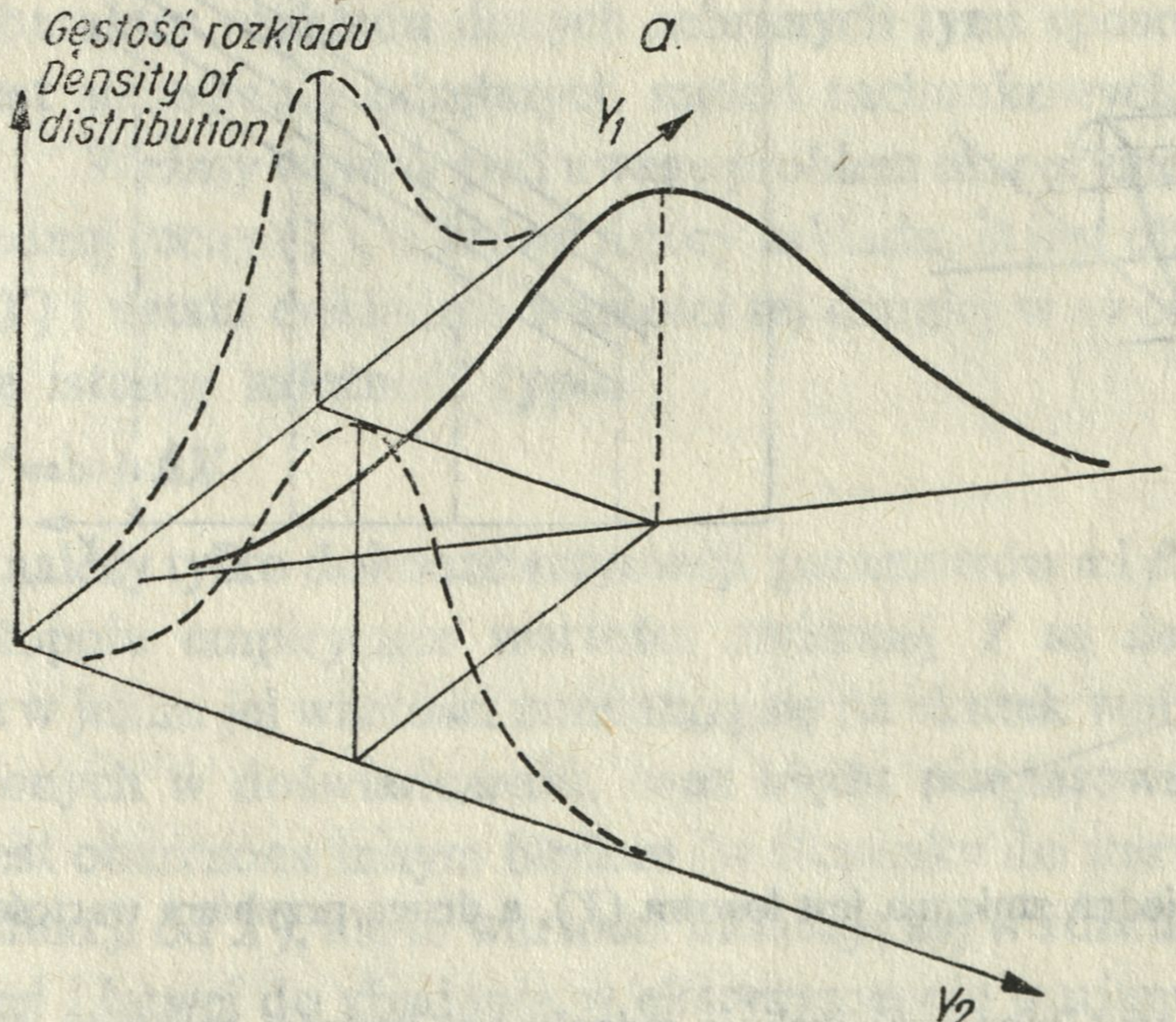
Zwykły stereotyp postępowania przy badaniu współzależności zmiennych (jak tego dowodzą setki prac w różnych czasopismach, z najbardziej renomowanymi włącznie) sprowadza się do obliczenia współczynników korelacji Pearsona (r) i równania regresji liniowej metodą najmniejszych kwadratów, zazwyczaj z podaniem istotności regresji, rzadziej z oceną błędu standardowego współczynnika nachylenia regresji. Stereotyp ten czasem stanowi postępowanie poprawne, a przynajmniej nieszkodliwe, czasem jednak prowadzi do fałszywych wniosków merytorycznych i dowodzi zignorowania różnicy między analizą regersji a analizą korelacyjną. Aby wyjaśnić o co tu chodzi, trzeba niestety zagłębić się w nieco trywialne rozważania o naturze rozkładu dwuwymiarowego.

W ogromnej większości wypadków obie badane cechy są zmiennymi losowymi, a całkowita wariancja, która charakteryzuje ich rozkłady, składa się z naturalnej zmienności i błędu pomiarowego. Podstawowym warunkiem stosowalności regresji i korelacji jest normalny rozkład zmiennych i tylko takimi zmiennymi będziemy się zajmowali. Pominiemy przy tym zagadnienie normalizacji rozkładów poprzez transformację (np. logarytmiczną) danych, chociaż i te procedury nie są wolne od pułapek godnych uwagi.

Mamy więc do czynienia z dwuwymiarowym rozkładem normalnym. Na trójwymiarowym modelu (rys. 1a) rozkład taki wygląda jak symetryczny pagórek, którego prostopadłe przekroje stanowią krzywe normalne, a „poziomice" oznaczające określone gęstości prawdopodobieństwa mają kształt koncentrycznych elips, co jeszcze lepiej widać, gdy spojrzymy na ten rozkład „z góry" (rys. 1b). W szczególnym wypadku (brak korelacji) owe poziome przekroje mogą przybrać kształt koła (rys. 1c), bądź też — w przypadku zupełnego zdeterminowania wartości jednej



Rys. 1. Dwuwymiarowy rozkład normalny
a — przestrzenny obraz dwuwymiarowego rozkładu normalnego dla skorelowanych zmiennych Y_1 i Y_2, *b* — rzut pionowy rozkładu dwuwymiarowego; izolinie jednakowej gęstości rozkładu mają kształt koncentrycznych elips, *c* — przy braku korelacji izolinie mają kształt koncentrycznych kół
Bivariate normal distribution
a — three-dimensional model of the distribution of two correlated normal variables Y_1 and Y_2, *b* — its vertical projection. The isolines of equal density of distribution are concentrically eliptic, *c* — when there is no correlation the isolines form concentric circles

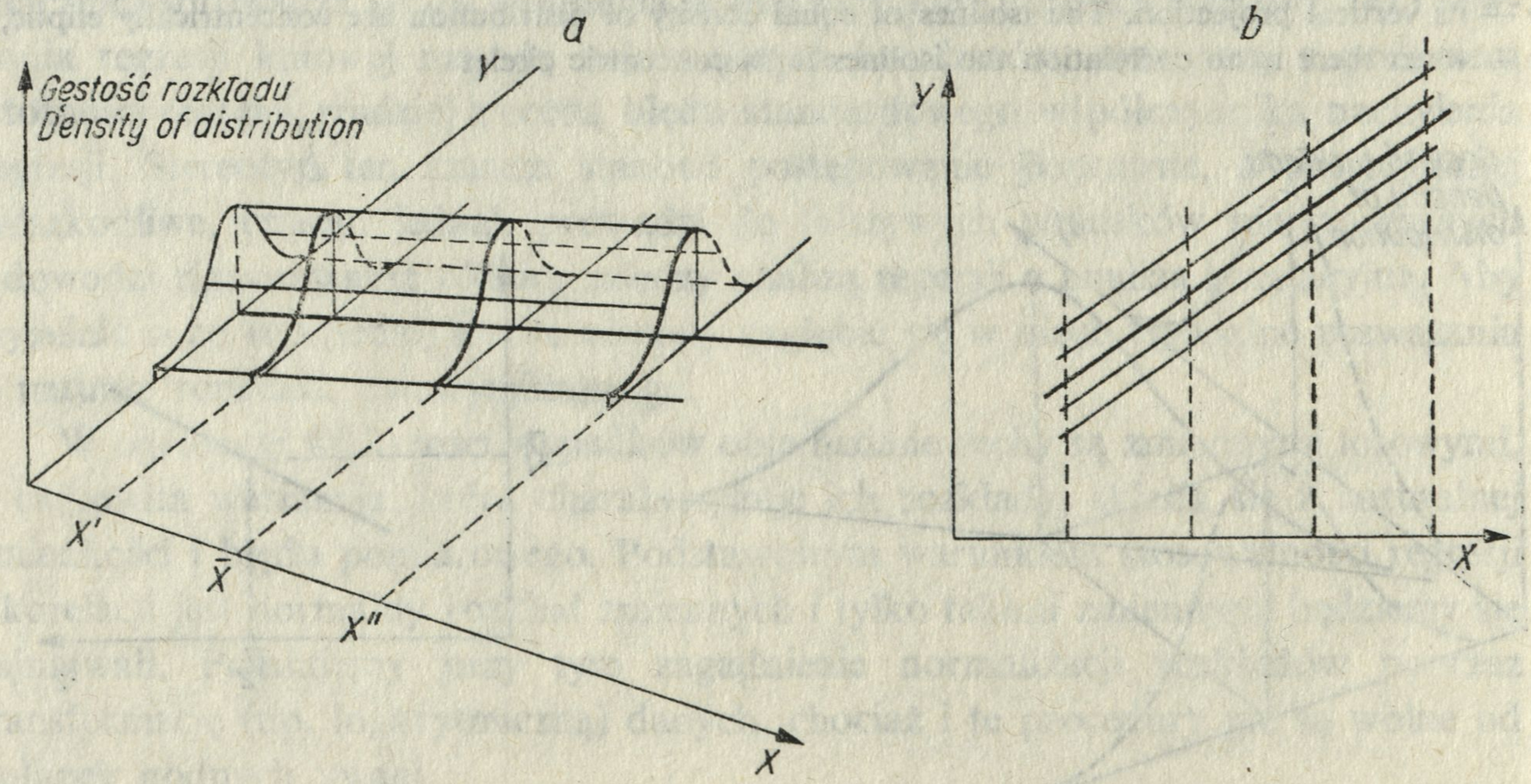


Rys. 2. Dwuwymiarowy rozkład normalny, gdy występuje zupełna korelacja między zmiennymi ($r=1$)
a — przestrzenny obraz takiego rozkładu, *b* — rzut pionowy
Bivariate normal distribution of two variables perfectly correlated ($r=1$)
a — three-dimensional model of such distribution, *b* — its vertical projection

zmiennej przez drugą — pagórek może spłaszczyć się do formy jednowymiarowej płetwy (rys. 2). W każdym razie obie zmienne charakteryzują się rozkładami normalnymi, co widać po zrzutowaniu ich na ścianki trójwymiarowego układu współrzędnych (rys. 1, 2).

Nikt nie zliczy przykładów takich współzależności dwóch zmiennych w ekologii. Sięgając do pierwszego z brzegu podręcznika znajdujemy na przykład korelację pierśnicy i wysokości drzewa, współzależność ciężaru ciała i wielkości terytorium u ptaków i ssaków, wielkości ciała drapieżników i ich ofiar, wielkości ciała i długości trwania pokolenia i tak dalej. Czasami zależność ma charakter wyraźnie asymetryczny, chociaż obie zmienne są niewątpliwie zmiennymi losowymi. Przykładem niech będzie zależność między ewapotranspiracją a produkcją pierwotną w ekosystemach.

Zupełnie wyjątkowo możemy mieć do czynienia z sytuacją, w której tylko jedna ze zmiennych ma charakter losowy, podczas gdy druga przybiera wartości ściśle zdeterminowane, nie obciążone żadną zmiennością ani błędem. Realnym przykładem takiej sytuacji jest układ doświadczalny, w którym eksperymentator ustala z góry wartości jednej zmiennej i dokonuje pomiarów pozostałej, będącej zmienną losową. Typową ilustracją z literatury ekologicznej jest zależność metabolizmu od temperatury otoczenia u zwierząt stałocieplnych. Oczywiście i w tym wypadku pierwsza zmienna może być obciążona błędem pomiarowym, lecz względna wielkość i charakter rozkładu tego błędu pozwalają na pominięcie jego wpływu w tych rozważaniach. W takim wypadku rozkład, z jakim mamy do czynienia, ma



Rys. 3. Dwuwymiarowy rozkład, kiedy jedna zmienna jest losowa (Y), a druga przybiera wartości ustalone (X)

a — przestrzenny obraz rozkładu, *b* — rzut pionowy; izolinie jednakowej gęstości rozkładu dla zmiennej Y przebiegają równolegle, a maksima gęstości rozkładów Y wyznaczają linię regresji

The two-dimensional distribution of one random variable (Y) when the other (X) attains only fixed values

a — three-dimensional model, *b* — its vertical projection. The isolines of equal density for the variable Y are parallel, the maxima of the densites of Y-distributions determine the regression line

zupełnie inne właściwości niż opisany wyżej symetryczny dwuwymiarowy rozkład normalny.

Każdej ustalonej wartości zmiennej kontrolowanej przez badacza odpowiada normalny rozkład zmiennej losowej mierzonej w eksperymencie (rys. 3). Nasz model ma kształt wydłużonego wałka, którego przekroje stanowią krzywe normalne. Obszary jednakowej gęstości prawdopodobieństwa stanowią proste równoległe, nie zaś elipsy, jak w poprzednim wypadku. Już na pierwszy rzut oka różnica między opisanymi rozkładami jest wyraźna i nie pozostaje bez wpływu na dobór metod stosowanych do badania takich współzależności. W pierwszej kolejności powinniśmy więc określić, z jakim typem rozkładu mamy do czynienia.

3. Założenia o charakterze współzależności

Przystępując do badania współzależności cech powinniśmy mieć możliwie jasne wyobrażenie o charakterze związku, który może łączyć badane zmienne, a przynajmniej jednoznacznie sformułowaną hipotezę na ten temat. Czym innym jest bowiem sprawdzanie, czy w badanym układzie w ogóle istnieje jakaś współzależność mierzonych cech, a czym innym testowanie hipotezy o istnieniu ściśle określonej w swojej formie zależności (np. funkcjonalnej), czym innym wreszcie dopasowywanie parametrów do równania opisującego tę zależność.

Hipoteza o istnieniu określonego typu zależności funkcjonalnej może być postawiona zarówno w przypadku układu eksperymentalnego, w którym jedna ze zmiennych („zmienna niezależna") jest kontrolowana przez badacza, jak i w przypadku pobierania losowych próbek z rozkładu dwuwymiarowego. Z uwagi na różny charakter rozkładu danych zebranych tymi sposobami w dalszej analizie konieczne jest stosowanie odrębnych metod rachunkowych.

Weźmy wpierw pod uwagę problem eksperymentatora, który dokonuje pomiarów jednej cechy (Y), o której z góry zakłada, iż jest ona liniową funkcją drugiej zmiennej (X) i ustala dokładnie wartości tej drugiej w swoim doświadczeniu. Wiadomo więc, że istnieje zależność typu:

$$Y=\alpha+\beta X \qquad (1)$$

i należy tylko dokonać estymacji parametrów α i β oraz ewentualnie ocenić, w jakim stopniu empiryczne wartości zmiennej Y są determinowane przez zależność (*1*), a w jakim jej wartości zmieniają się na skutek wpływu innych czynników, nie uchwyconych w doświadczeniu, oraz błędu pomiarowego. Każda zmierzona wartość Y_i jest obarczona innym błędem (w stosunku do wartości Y wynikającej z postulowanej funkcji od X), ale te wartości układają się w rozkład, zazwyczaj normalny, o określonej i łatwej do zbadania w eksperymencie wariancji. Możemy to zapisać w ten sposób:

$$Y_i=\alpha+\beta X_i+\varepsilon_i \qquad (2)$$

gdzie ε_i oznacza błąd zmiennej Y, odpowiedzialny za istnienie zmienności Y dla każdego X.

Wartość oczekiwana rozkładu błędu ε wynosi 0, a więc średnia wszystkich pomiarów Y_i dla każdego X powinna wynosić dokładnie $\alpha+\beta X_i$. Powtarzając wielokrotnie pomiary Y przy ustalonych wartościach X możemy oszacować wariancję tego błędu i dzięki temu wyciągnąć wiele interesujących wniosków: o przedziale ufności oszacowanych parametrów równania (*1*), o pewności z jaką możemy przewidywać nieznane wartości Y dla dowolnie wybranych wartości X, wreszcie o sile badanego związku w stosunku do interferujących czynników przypadkowych, nie ujętych w doświadczeniu.

Inaczej rzecz się ma przy badaniu współzależności dwóch zmiennych losowych, Y_1 i Y_2. Na początku nie możemy poczynić innych założeń niż to, że badane zmienne mają rozkłady normalne. W tej sytuacji możemy postawić tylko hipotezę o istnieniu między nimi współzależności i następnie dokonać oceny współczynnika korelacji (postępując według recepty podanej w każdym podręczniku statystyki). Możemy następnie badać istotność otrzymanego współczynnika korelacji, porównywać współczynniki uzyskane w różnych badanych populacjach i tak dalej, stosując powszechnie znane testy. Bez wprowadzenia dodatkowych założeń żadne inne operacje rachunkowe nie wniosą już nic nowego. Wysoka wartość współczynnika korelacji może jednak dać asumpt do poczynienia założenia, iż pomiędzy badanymi zmiennymi istnieje zależność liniowa, zgodna z równaniem:

$$Y_1=\alpha+\beta Y_2 \qquad (3)$$

Zbliża to nas do poprzedniej sytuacji, kiedy z góry zakładaliśmy istnienie funkcjonalnej zależności. W tym wypadku jednak przeważnie nie mamy powodu wyróżniać a priori jednej ze zmiennych jako zmiennej niezależnej i możemy równie dobrze napisać:

$$Y_2=1/\beta\ Y_1-\alpha/\beta$$

Może nas jednak interesować estymacja parametrów równania (*3*) tak, aby można było ilościowo określić, jak zmienia się wartość oczekiwana jednej ze zmiennych przy zmianach wartości drugiej. W tym wypadku jednak zmierzone wartości obu zmiennych, Y_1 i Y_2, obarczone są błędami, wynikającymi zarówno z błędów pomiarowych jak i z wpływu innych czynników niż określona równaniem (*3*) współzależność. Można tę sytuację zapisać równaniem:

$$(Y_1+\varepsilon_1)=\alpha+\beta\,(Y_2+\varepsilon_2) \qquad (4)$$

Również i te błędy, ε_1 i ε_2, mają rozkład normalny (z założenia!) z wartością oczekiwaną 0 i określoną wariancję, ale o tej ostatniej nie dowiemy się wiele powtarzając pomiary czy obserwacje obu zmiennych, inaczej niż to było w przypadku równania (*2*), bowiem mamy zbyt wiele niewiadomych.

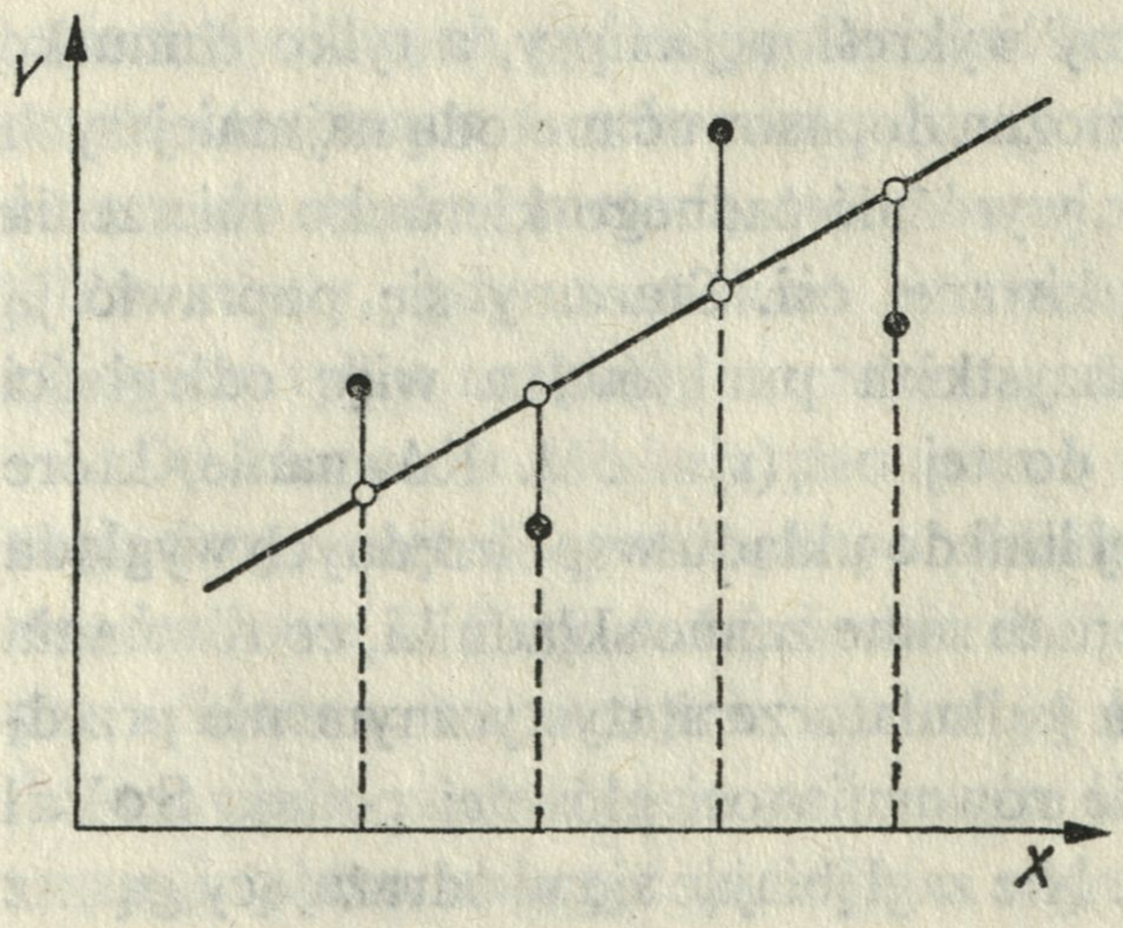
Może się zdarzyć, iż o badanym rozkładzie dwuwymiarowym wiemy z góry, iż kryje się za nim asymetryczna zależność funkcjonalna (a więc jedna ze zmiennych jest wyróżniona jako zmienna niezależna), ale jednak sposób pobierania próby nie spełnia warunku ustalenia jej wartości bez błędu. Otóż sytuacja ta nie różni się w gruncie rzeczy od tej, w której mamy dwie symetryczne zmienne losowe (jak w równaniu (*4*)).

Mamy więc dwa różne modele i do każdego z nich powinniśmy stosować odpowiednie metody analizy statystycznej. Zaczniemy od najczęściej stosowanej, ale nie zawsze wybranej prawidłowo regresji liniowej.

4. Regresja liniowa według najmniejszych kwadratów

Niemal wszystkie podręczniki statystyki podają przepis na obliczanie regresji liniowej i wiele z nich formułuje mniej lub bardziej jednoznacznie listę założeń i ograniczeń tej metody (z podręczników dla biologów w języku polskim regresja zdefiniowana jest najprecyzyjniej w książce Parkera 1978). Mimo to jednak zastrzeżenia te są często ignorowane. Przypomnijmy więc, że analiza regresji dotyczy wyłącznie sytuacji, kiedy tylko jedna ze zmiennych jest zmienną losową, a pozostała zmienna przybiera wartości ustalone przez eksperymentatora (rozkład jak na rys. 3). Założenia o stosowalności regresji, które muszą być spełnione przed przystąpieniem do analizy, są następujące: (1) Obie zmienne związane są funkcjonalną zależnością o charakterze liniowym, opisaną równaniem (*1*): $Y=\alpha+\beta X$, gdzie Y stanowi zmienną zależną i jest zmienną losową, zaś X — zmienną niezależną, kontrolowaną w eksperymencie. Zależność jest więc asymetryczna. (2) Rozkład zmiennej zależnej Y jest normalny dla każdej wartości X, przy czym średnia z tego rozkładu przypada dokładnie na linii regresji danej powyższym równaniem. (3) Wariancja rozkładu zmiennej Y jest stała i nie zależy od wartości X.

Jak więc widać, zastosowanie regresji liniowej ogranicza się do specjalnego przypadku układów doświadczalnych. Metoda najmniejszych kwadratów pozwala na oszacowanie wartości parametrów równania (*1*), dzięki czemu otrzymuje się równanie empiryczne ($Y=a+bX$). Istotą tej metody jest takie dopasowanie współczynników równania, aby suma kwadratów różnic pomiędzy zmierzonymi wartościami zmiennej zależnej Y a wartościami oszacowanymi z równania, przy wyznaczonych wartościach zmiennej niezależnej X, była najmniejsza. Graficzna interpretacja tej procedury wskazuje, iż odległość pomiędzy empiryczną a oszacowaną wartością Y mierzy się w kierunku pionowym (rys. 4).



Rys. 4. Odchylenia wartości zmiennej Y od wartości oczekiwanych z regresji mierzymy pionowo
The deviations of empirical values of the variable Y from the regression are measured vertically

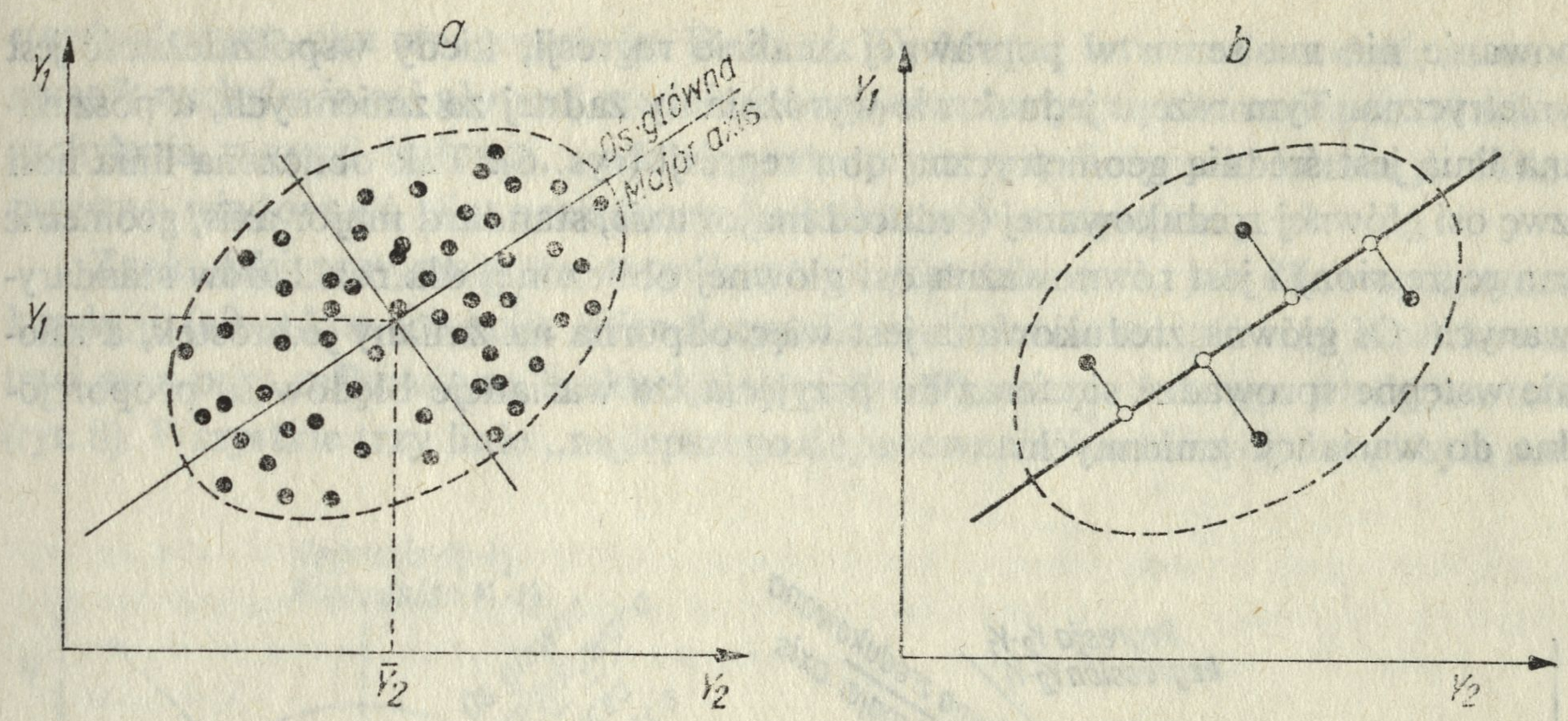
Dyskutowanie istotności współczynnika korelacji (r) w układzie, w którym poprawne było zastosowanie analizy regresji jest pozbawione znaczenia, gdyż fakt istnienia istotnego związku pomiędzy badanymi zmiennymi stanowi założenie a priori tej metody. Ważną interpretację ma jednak w tym wypadku tzw. współczynnik determinacji (r^2). Łatwo dowieść, iż współczynnik ten stanowi stosunek kwadratów odchyleń od średniej Y, wynikający z istnienia regresji, do całkowitej sumy kwadratów. Informuje więc o tym, jaką część całkowitej zmienności Y można przypisać jej związkowi ze zmienną X (Sokal i Rohlf 1981).

Ważną zaletą metody regresji jest to, że pozwala ona wnioskować nie tylko o sile związku pomiędzy zmiennymi. Można dodatkowo porównywać nachylenia linii regresji uzyskanych z różnych populacji oraz stosować równanie regresji do szacowania wartości oczekiwanych zmiennej Y na podstawie wartości X (a nawet odwrotnie), wraz z oszacowaniem granic ufności dla takich przewidywań. Odpowiednie procedury są opisywane szeroko w licznych podręcznikach, ale prawidłowe metody obliczania wartości oczekiwanej X na podstawie Y oraz przedziału ufności są podane tylko w niektórych książkach (np. Sokal i Rohlf 1981).

Z tego co powiedziano wyżej wynika, iż metoda regresji nie powinna być stosowana, jeżeli nie są spełnione założenia (1—3), a więc we wszystkich tych wypadkach, kiedy mamy do czynienia ze zwykłym rozkładem normalnym dwuwymiarowym (obie zmienne są losowe), nawet jeżeli wiemy, że za badanym rozkładem kryje się zależność funkcjonalna.

5. Dopasowywanie parametrów równania liniowego przy korelacji dwóch zmiennych losowych

Co jednak można uczynić, jeżeli istnieje potrzeba porównania nachyleń elips otaczających rozkłady dwuwymiarowe, a warunki używania regresji nie są spełnione? W tym wypadku mamy do czynienia z symetryczną zależnością zmiennych Y_1 i Y_2. Rzut oka na rozkład empirycznych danych na układzie współrzędnych przekonuje nas, że najlepiej dopasowana linia pokrywa się z osią długą (osią główną) elips opisujących topografię rozkładu (rys. 5a). Linia ta bywa też nazywana prostą regresji ortogonalnej (synonimy angielskie: major axis, principal axis, principal component). Oczywiście przystępując do analizy nie mamy wykreślonej elipsy, a tylko chmurkę punktów na układzie współrzędnych. Linię można dopasować metodą najmniejszych kwadratów. Tym razem nie możemy jednak wyróżnić żadnego kierunku mierzenia odchyleń punktów empirycznych od poszukiwanej osi. Staramy się poprawić ją tak, aby przebiegała możliwie najbliżej wszystkich punktów, a więc odległości między nimi a osią mierzymy prostopadle do tej osi (rys. 5b). Równanie, które pozwala obliczyć współczynnik nachylenia tej linii do układu współrzędnych wygląda dość nieprzyjemnie (p. Dodatek 1), ale zawiera te same znane składniki, co równania regresji liniowej czy korelacji. Obliczenia na kalkulatorze statystycznym nie przedstawiają żadnych trudności. Wyprowadzenie równania osi głównej podają Sokal i Rohlf (1981) oraz Seim i Saether (1983). Nie zagłębiając się w odrażający gąszcz



Rys. 5. *a* — Oś główna (długa) i oś krótka elipsy opisującej dwuwymiarowy rozkład zmiennych losowych Y_1 i Y_2, *b* — odchylenia wartości empirycznych od osi głównej mierzymy prostopadle do tej osi

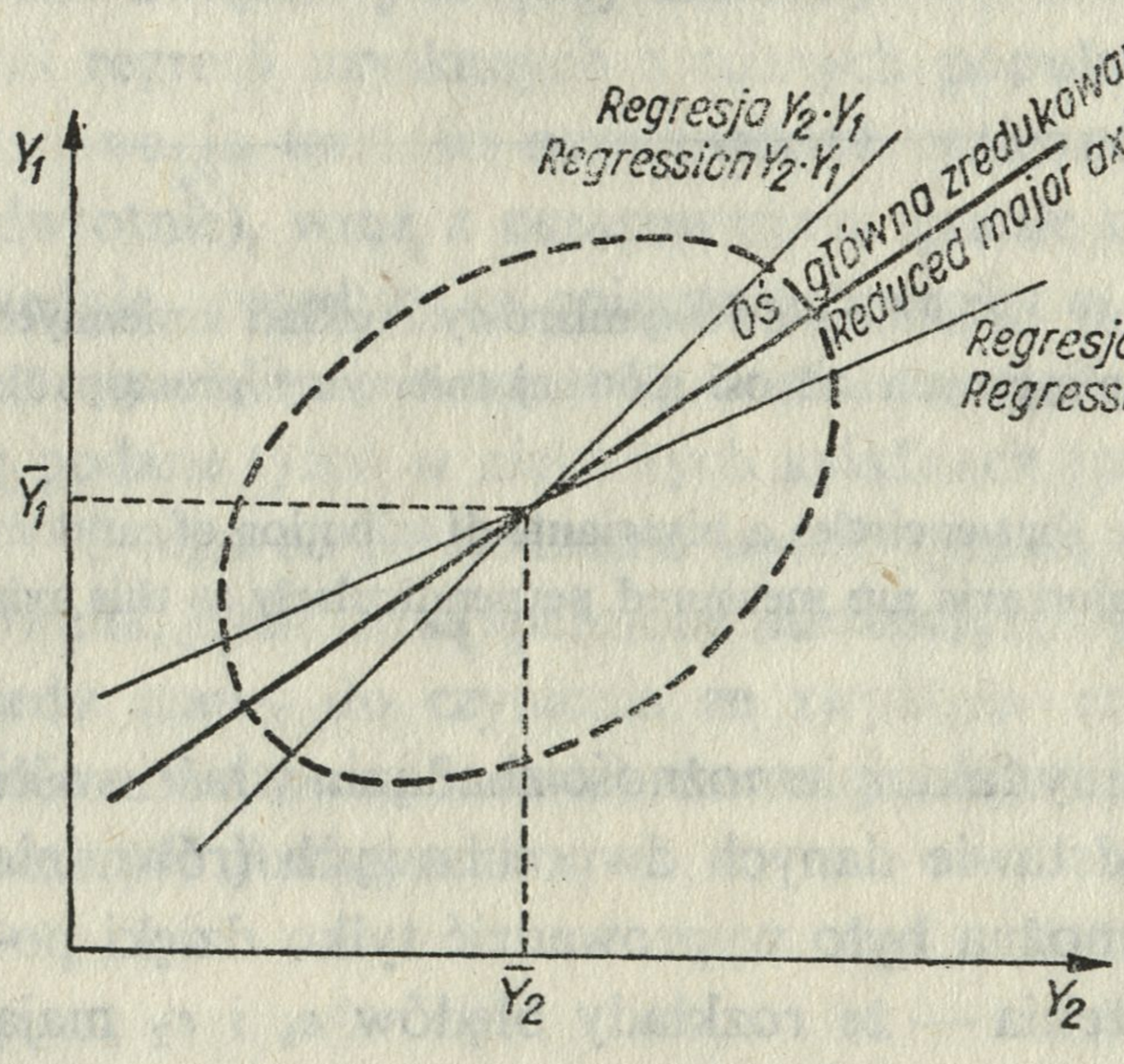
a — Major (principal) and minor axes of the elipse that encircles a bivariante distribution of random variables Y_1 and Y_2, *b* — deviations from the major axis are measured perpendicularly to this axis

równań warto zwrócić uwagę na jeden ważny fakt: niemożność zbadania właściwości dwóch na raz rozkładów błędów na podstawie danych dwucechowych (równanie (*4*)) powoduje, iż równanie osi głównej można było wyprowadzić tylko dzięki poczynieniu dodatkowego, wstępnego założenia — że rozkłady błędów ε_1 i ε_2 mają identyczną wariancję.

Nie miałoby to może znaczenia dla biologów, gdyby nie wynikająca stąd dość przykra, a bynajmniej nieoczywista intuicyjnie właściwość osi głównej. Postulat jednakowej wariancji błędów jest trudny do zrealizowania. Wystarczy np. zmienić skalę pomiarową jednej ze zmiennych (powiedzmy milimetry na centymetry), aby wariancje błędów różniły się wydatnie. W rezultacie współczynnik nachylenia osi głównej obliczony dla danych w centymetrach wcale nie będzie 10 razy mniejszy od współczynnika wyznaczonego dla tych samych danych mierzonych w milimetrach. Trudność ta zwykle nie występuje wtedy, gdy mamy do czynienia z danymi transformowanymi logarytmicznie. W ekologii, fizjologii porównawczej oraz nauce o strategiach życiowych (gdzie ją umieścić?) logarytmiczne transformowanie danych jest często stosowane w celu normalizacji skośnych rozkładów. Takie dane nadają się dobrze do analizy metodą osi głównej. Oczywiście transformację wolno stosować tylko wtedy, gdy w efekcie otrzymuje się rozkład zbliżony do normalnego, albowiem tylko rozkłady normalne nadają się do analizy opisywanymi metodami.

Od kłopotów ze skalą pomiarową można się uwolnić także poprzez standaryzację danych w taki sposób, aby rozkłady obu zmiennych miały średnią 0 i odchylenie standardowe 1. Przy dużej liczbie danych procedura standaryzacji może być dość pracochłonna, na szczęście można jej jednak uniknąć. Symetryczną linię dobrego dopasowania można otrzymać jeszcze innym sposobem. Posługując się rachunkiem regresji można obliczyć dwie linie: regresję Y do X i odwrotną, X do Y. Takie pos-

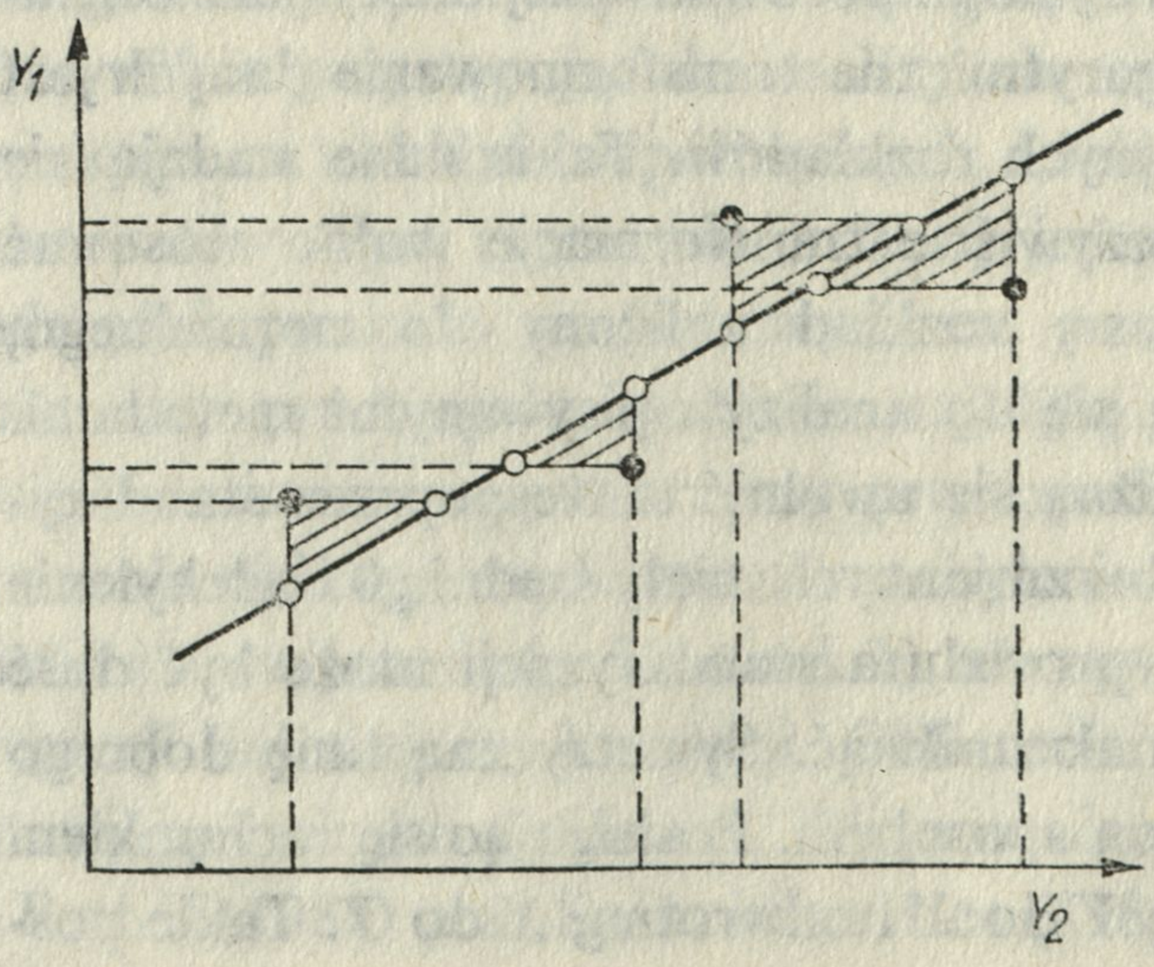
tępowanie nie ma sensu w poprawnej analizie regresji, kiedy współzależność jest asymetryczna. Tym razem jednak nie wyróżniamy żadnej ze zmiennych, a poszukiwana linia jest średnią geometryczną obu regresji (rys. 6). Tak obliczona linia nosi nazwę osi głównej zredukowanej (reduced major axis, standard major axis, geometric mean regression) i jest równoważna osi głównej obliczonej dla rozkładów standaryzowanych. Oś główna zredukowana jest więc odporna na zmiany jednostek, a założenie wstępne sprowadza się teraz do przyjęcia, że wariancje błędów są proporcjonalne do wariancji zmiennych.



Rys. 6. Oś główna zredukowana jest średnią geometryczną regresji $Y \cdot X$ i regresji $X \cdot Y$

The reduced major axis is a geometric mean of the two regressions: $Y \cdot X$ and $X \cdot Y$

Ponieważ w regresji $Y \cdot X$ odległości między wartościami empirycznymi a oczekiwanymi mierzone są równolegle do osi Y, a dla regresji $X \cdot Y$ — równoległe do osi X, zatem odcinki łączące na wykresie punkty empiryczne z poszukiwaną osią wytyczają prostokątne trójkąty (rys. 7). Najlepsze dopasowanie polega na takim dopasowaniu linii, aby pole powierzchni tych trójkątów było w sumie jak najmniejsze (rys. 7). Brzmi to dość zawile, ale rachunkowo wartość bezwzględna współczynnika nachylenia osi głównej zredukowanej równa jest po prostu ilorazowi odchyleń

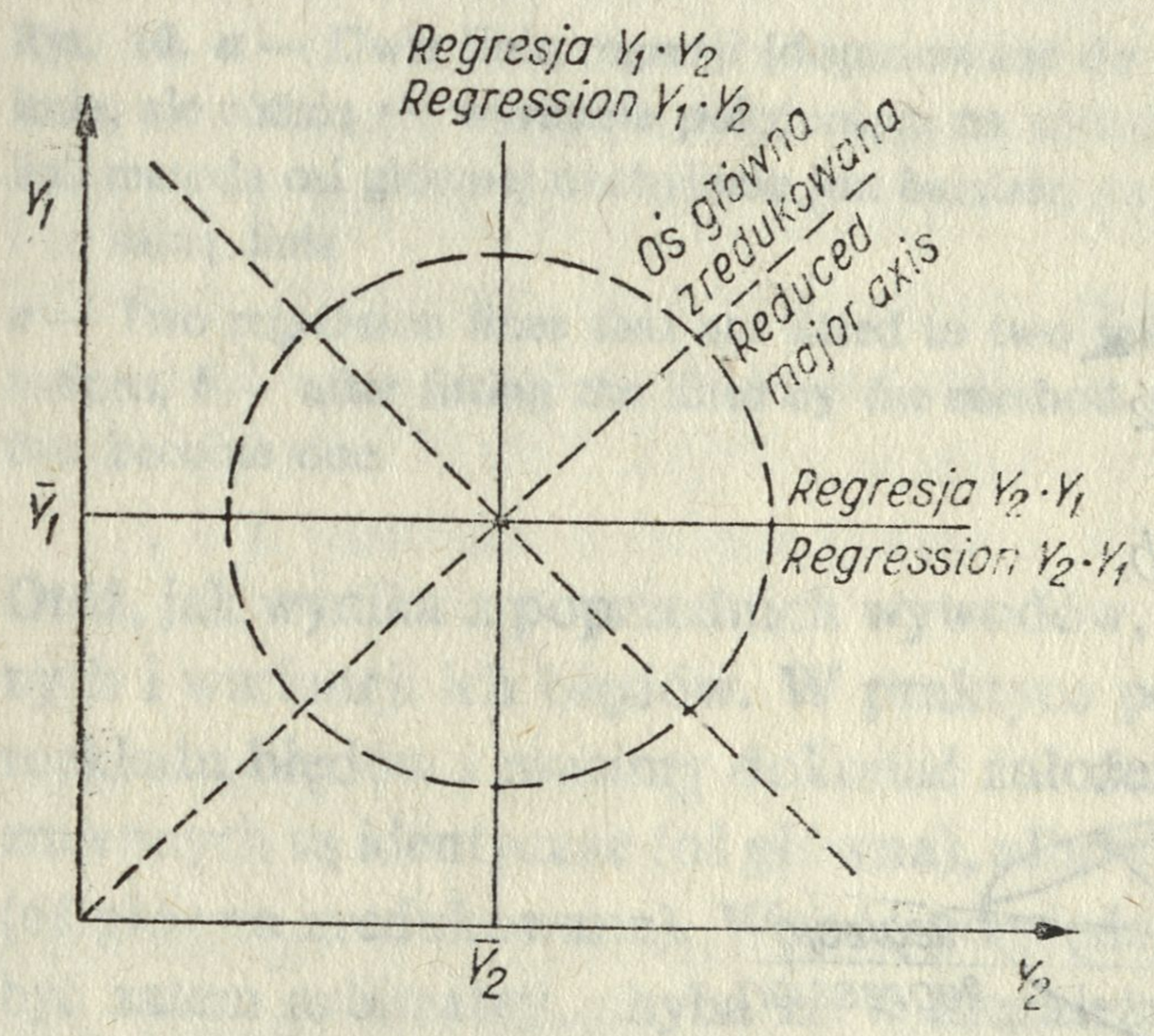


Rys. 7. Dopasowanie osi głównej zredukowanej polega na takim przeprowadzeniu linii, aby suma powierzchni zakreskowanych trójkątów wytyczonych przez tę oś i proste równoległe do osi współrzędnych, łączące punkty empiryczne z osią, była najmniejsza

The best fit of the reduced major axis minimizes the sum of areas of the triangles delimited by this axis and the lines parallel to the coordinates, connecting the empirical points with the axis

standardowych obu zmiennych (p. Dodatek 1). Można łatwo udowodnić, że współczynnik nachylenia osi głównej zredukowanej jest równy zwykłemu współczynnikowi, nachylenia regresji liniowej, podzielonemu przez współczynnik korelacji. Miła to zapewne wiadomość dla posiadaczy kalkulatorów statystycznych!

Znak, jaki przyjmuje ten współczynnik jest taki sam, jak dla współczynnika korelacji (albo kowariancji). Współczynnik nachylenia osi głównej zredukowanej traci sens przy całkowitym braku korelacji ($r=0$), mimo iż osiąga wtedy wartość 1 (rys. 8). Wszystkie trzy linie „najlepszego dopasowania", o których dotąd była mowa



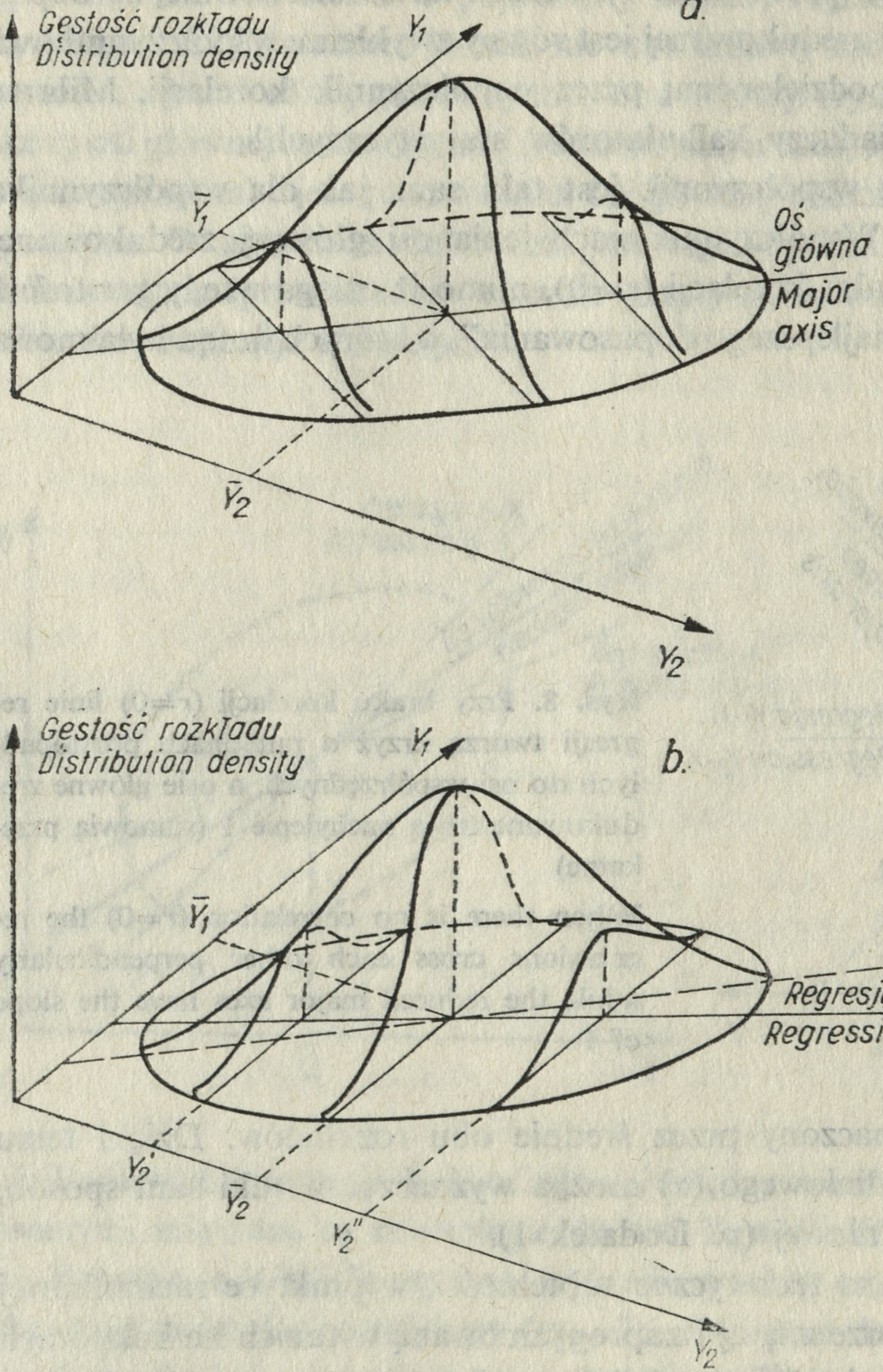
Rys. 8. Przy braku korelacji ($r=0$) linie regresji tworzą krzyż o ramionach prostopadłych do osi współrzędnych, a osie główne zredukowane mają nachylenie 1 (stanowią przekątne)

When there is no correlation ($r=0$) the regressions cross each other perpendicularly while the reduced major axes have the slope of 1

przechodzą przez punkt wyznaczony przez średnie obu rozkładów. Dzięki temu pozostały parametr równania liniowego (α) można wyznaczyć w taki sam sposób, jak to robimy przy regresji liniowej (p. Dodatek 1).

Czy jednak, bez względu na teoretyczne subtelności, w praktyce rachunkowej nie wystarczy posłużyć się dobrze znaną i zaprogramowaną w tanich kalkulatorach metodą regresji? Otóż okazuje się iż prowadziłoby to do sztucznego obniżenia wartości współczynnika nachylenia linii najlepszego dopasowania. Przyjrzyjmy się raz jeszcze trójwymiarowemu modelowi rozkładu normalnego dwuwymiarowego (rys. 9). Oś główna elipsy przecina jego podstawę dokładnie symetrycznie, a płaszczyzna przekroju przechodzi ściśle przez maksima gęstości rozkładu dla każdej krzywej normalnej, wyznaczonej przez cięcie prostopadłe do osi głównej (rys. 9a). Tymczasem przy zastosowaniu metody regresji linia najlepszego dopasowania wyznaczona jest przez wierzchołki rozkładów „przekrojowych", przeprowadzonych pod innym kątem do osi głównej niż kąt prosty (rys. 9b). Prowadzi to ewidentnie do otrzymywania mniejszego kąta nachylenia tak wyznaczonej linii. Autor nie ukrywa, że aby dokładnie zrozumieć o co tu chodzi, posługiwał się intensywnie plasteliną.

Dla identycznych zestawów danych linia regresji zawsze jest najmniej nachylona w stosunku do osi rzędnych, natomiast to, czy oś główna, czy oś zredukowana jest bardziej stroma, zależy od stosunku wariancji i współczynnika korelacji w danym rozkładzie dwuwymiarowym. Różnica między nachyleniami tych linii maleje ze wzrostem korelacji, a dla $r=1$ wszystkie trzy linie są identyczne.

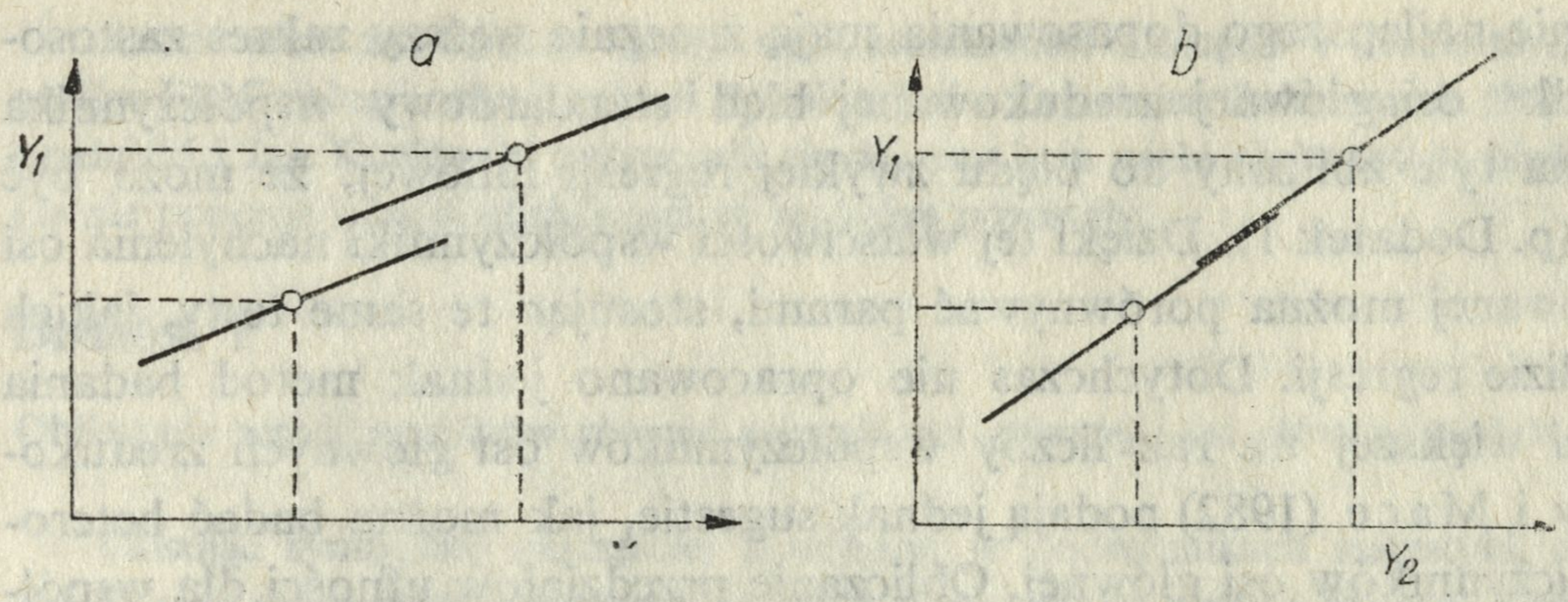


Rys. 9. Dopasowanie linii opisującej współzależność dwóch zmiennych losowych Y_1 i Y_2: *a* — metodą osi głównej, *b* — metodą regresji
Objaśnienia w tekście

Fitting the lines that describe a relation of two random variables Y_1 and Y_2
a — by the method of major axis, *b* — by the method of regression.
See text for explanations

W literaturze ekologicznej i ewolucyjnej nietrudno o przykłady dyskutowania różnic pomiędzy regresjami (na przykład metabolizmu i ciężaru ciała) u różnych grup zwierząt. Zastosowanie regresji tam, gdzie powinna być użyta oś główna, prowadzi nie tylko do sztucznego obniżenia liczbowej wartości współczynnika regresji, ale może też zupełnie zmienić interpretację biologiczną. Harvey i Mace (1982) ilustrują to sugestywnym rysunkiem, który tu z niewielkimi zmianami przytaczamy (rys. 10). Inny charakterystyczny przykład przedstawiono w Dodatku 2.

Skoro jednak ten sam zestaw danych może dostarczyć trzech różnych linii „najlepszego" dopasowania, to jak odgadnąć, która z nich jest najlepsza z najlepszych?



Rys. 10. *a* — Dwie linie regresji (dopasowane do dwóch zestawów danych) mają podobne nachylenie, ale różnią się wyraźnie położeniem na układzie współrzędnych, *b* — na skutek dopasowania linii metodą osi głównej nachylenie jest bardziej strome i okazuje się, że mamy do czynienia z jedną i tą samą linią

a — Two regression lines that are fitted to two sets of data have a similar slope but differe in elevations, *b* — after fitting the lines by the method of major axis the slope becomes steeper and the two become one

Otóż, jak wynika z poprzednich wywodów, zależy to od charakteru rozkładu zmiennych i wariancji ich błędów. W praktyce poza układem regresji liniowej nie znamy rozkładu błędów i musimy dokonać założenia, że albo wariancje błędów przy obu zmiennych są identyczne (oś główna), albo są proporcjonalne do wariancji zmiennych (oś główna zredukowana). Wybór pomiędzy osią główną a osią zredukowaną musi być zatem arbitralny, chyba że w niezależnych pomiarach zdołamy się czegoś dowiedzieć o tych rozkładach. Decyzję o niestosowaniu regresji liniowej można jednak podjąć w sposób obiektywny.

Niektóre podręczniki podają jeszcze jeden sposób obliczania linii najlepszego dopasowania do danych skorelowanych (zupełnie odrębny od metody najmniejszych kwadratów) — tzw. metodę trzech grup Bartletta. Metoda ta ma jednak szereg wad (Sokal i Rohlf 1981), a ponadto jej opis podaje jeden z polskich podręczników statystyki dla biologów (Bogucki 1979), dlatego tutaj ją pomijamy.

6. Jak posługiwać się równaniami linii najlepszego dopasowania?

Równanie regresji liniowej, jeżeli tylko jest prawidłowo zastosowane, daje najbardziej uniwersalne możliwości różnorodnego wykorzystania wyników w dyskusji. Obliczone współczynniki można porównywać, stawiając hipotezę o jednakowych nachyleniach regresji, a jeżeli ta zostanie przyjęta, można badać, czy różnią się wartością współczynnika *a*, to jest usytuowaniem na układzie współrzędnych. Podręczniki zawierają opisy testów przeznaczonych zarówno do porównań regresji parami, jak też do badania heterogeniczności współczynników większej na raz liczby równań. Można łatwo wyznaczyć błąd standardowy współczynnika regresji czy też wyznaczyć przedziały ufności na dowolnym poziomie. Co więcej, regresja pozwala na przewidywanie wartości zmiennej zależnej na podstawie zmiennej niezależnej, a nawet odwrotnie, z oceną błędu takich oszacowań.

Pozostałe linie najlepszego dopasowania mają znacznie węższy zakres zastosowań. W przypadku osi głównej zredukowanej błąd standardowy współczynnika nachylenia jest na tyle zbliżony do błędu zwykłej regresji liniowej, że może być nim zastąpiony (p. Dodatek 1). Dzięki tej właściwości współczynniki nachylenia osi głównej zredukowanej można porównywać parami, stosując te same testy, jakich używamy w analizie regresji. Dotychczas nie opracowano jednak metod badania heterogeniczności większej na raz liczby współczynników osi głównych zredukowanych. Harvey i Mace (1982) podają jednak sugestie, jak można badać heterogeniczność współczynników osi głównej. Obliczanie przedziałów ufności dla współczynnika nachylenia osi głównej jest z kolei dość trudne i nadaje się tylko dla danych bardzo licznych i silnie skorelowanych (p. Dodatek 1; Sokal i Rohlf 1981).

Osie główne nie powinny być, niestety, używane do przewidywania wartości jednej ze zmiennych na podstawie drugiej, gdyż nie sposób wyznaczyć przedział ufności takiej oceny. Ze względu na to, że obie zmienne są zmiennymi losowymi, możemy mówić tylko o dwuwymiarowym obszarze ufności. Sokal i Rohlf (1981) podają procedurę obliczania takiego obszaru wokół średnich rozkładu X i Y.

7. Wnioski końcowe

Któryś z autorów podręcznika statystyki powiedział mądrze, iż najważniejszą wskazówką dla wszystkich użytkowników metod statystycznych jest to, aby znać się dobrze na meritum badanego zagadnienia. Na przykład współzależność aktywności enzymu i temperatury powinien badać fizjolog, zależność wielkości ciała i wielkości terytorium u ptaków — ekolog-ornitolog itd. Tylko to bowiem gwarantuje sensowne sformułowanie hipotez, odróżnianie przyczyn od skutków i — co w naszym problemie najważniejsze — odróżnienie regresji od korelacji oraz trafne intuicje co do charakteru badanych rozkładów. Najbardziej wykwalifikowany statystyk i jego ezoteryczna wiedza nie zastąpią zdrowego rozsądku biologa.

Jeśli zaś idzie o konkretne rady dla tego ostatniego, to autorzy uczonych prac o metodach statystycznych wzdragają się przed podaniem gotowych recept, ale zgodni są w kilku punktach:

1. Regresję liniową należy stosować tylko do układów eksperymentalnych (wyjątkowo opisowych), kiedy jedna ze zmiennych jest kontrolowana przez badacza, a jej wartości nie są (praktycznie) obciążone błędami.

2. Oś główna nadaje się do opisu współzależności dwóch zmiennych losowych o rozkładach normalnych, z podobnymi wariancjami błędów, albo o rozkładach normalnych, standaryzowanych. W szczególności oś główna nadaje się dobrze do badania współzależności zmiennych uprzednio transformowanych logarytmicznie.

3. Oś główna zredukowana może zastąpić regresję liniową tam, gdzie zmienna niezależna jest obciążona godnym uwagi błędem pomiarowym lub losowym. Można jej używać w podobnych sytuacjach jak osi głównej, o ile możemy przyjąć, że wariancje błędów są proporcjonalne do wariancji zmiennych.

Autor wiele zawdzięcza agresywnej dociekliwości Kolegów i Studentów, uczestników seminariów Ekofizjologii oraz Ekologii Populacyjnej i Zagadnień Pokrewnych IBŚ UJ. Panowie Adam Łomnicki i Jan Kozłowski przyczynili się do usunięcia wielu niejasności i błędów z tego artykułu, ale nie ponoszą odpowiedzialności za te, które pozostały.

Dodatek 1

Obliczanie współczynników równań regresji, osi głównej i osi głównej zredukowanej

Stosując symbolikę najczęściej stosowaną w podręcznikach statystyki, zmienne oznaczymy X i Y, przy czym należy pamiętać, że tylko w przypadku regresji liniowej X oznacza zmienną niezależną, a Y zmienną losową zależną od X. W pozostałych przypadkach obie zmienne są losowe, a ich współzależność symetryczna. W celu uproszczenia zapisu nie wprowadzamy jednak osobnych oznaczeń (na przykład Y_1 i Y_2). Przez n oznaczamy liczbę par zmiennych X i Y. Dalej oznaczamy:

$$x=(X-\bar{X}) \text{ oraz } y=(Y-\bar{Y})$$

$$\Sigma x^2=\Sigma (X-\bar{X})^2=\Sigma X^2-\frac{(\Sigma X)^2}{n}$$

$$\Sigma y^2=\Sigma (Y-\bar{Y})^2=\Sigma Y^2-\frac{(\Sigma Y)^2}{n}$$

$$\Sigma xy=\Sigma (X-\bar{X})(Y-\bar{Y})=\Sigma XY-\frac{\Sigma X \Sigma Y}{n}$$

a stąd:

$$s_x^2 \text{ (wariancja } X)=\frac{\Sigma x^2}{n-1}$$

$$s_y^2 \text{ (wariancja Y)}=\frac{\Sigma y^2}{n-1}$$

$$s_{xy} \text{ (kowariancja)}=\frac{\Sigma xy}{n-1}$$

Zadanie polega na oszacowaniu parametrów równania liniowego: $Y=\alpha+\beta X$, przy czym (wg Sokala i Rohlfa 1981) najlepsze oszacowanie parametru β dla regresji oznaczamy symbolem $b_{Y\,.\,X}$, dla osi głównej b_1, a dla osi głównej zredukowanej $v_{Y\,.\,X}$ Oszacowanie parametru α oznaczamy zawsze przez a.

1. Regresja liniowa

$$b_{Y\,.\,X}=\frac{\Sigma xy}{\Sigma x^2}$$

$$a=\bar{Y}-b_{Y\,.\,X}\bar{X}=\frac{\Sigma Y-b_{Y\,.\,X}\Sigma X}{n}$$

Błąd standardowy współczynnika regresji:

$$s_b=\sqrt{\frac{\Sigma y^2-\frac{(\Sigma xy)^2}{\Sigma x^2}}{\Sigma x^2\,(n-2)}}$$

95% przedział ufności dla $b_{Y\,.\,X}$:

$L_1 = b_{Y \cdot X} - t_{0,05\,[n-2]}\, s_b$

$L_2 = b_{Y \cdot X} + t_{0,05\,[n-2]}\, s_b$

gdzie $t_{\alpha,\nu}$ — wartość krytyczna rozkładu Studenta dla testu dwustronnego przy $\alpha = 5\%$ i $\nu = n-2$.

2. Oś główna zredukowana

Wpierw obliczamy współczynnik korelacji:

$$r_{XY} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \, \Sigma y^2}}$$

a następnie współczynniki osi głównej zredukowanej:

$$v_{Y \cdot X} = \pm \frac{\sqrt{s_y}}{\sqrt{s_x}} = \pm \sqrt{\frac{\Sigma y^2}{\Sigma x^2}} = \pm \frac{b_{Y \cdot X}}{r_{XY}}$$

Znak przy $v_{Y \cdot X}$ przyjmujemy taki, jaki ma współczynnik korelacji.

$$a = \frac{\Sigma Y - v_{Y \cdot X} \Sigma X}{n}$$

Błąd standardowy jest taki sam jak dla współczynnika regresji liniowej ($s_v = s_b$), a więc granice przedziału ufności dla $v_{Y \cdot X}$ obliczamy identycznie jak dla $b_{Y \cdot X}$.

3. Oś główna

Obliczenia są nieco bardziej zawiłe. Wprowadźmy dodatkowe oznaczenia:

$$\lambda_1 = \frac{s_x^2 + s_y^2 + D}{2}$$

$\lambda_2 = s_x^2 + s_y^2 - \lambda_1$

gdzie $D = \sqrt{(s_x^2 + s_y^2)^2 - 4\,(s_x^2\, s_y^2 - s_{xy}^2)}$

Wówczas:

$$b_1 = \frac{s_{xy}}{\lambda_1 - s_y^2}$$

albo

$$b_1 = \frac{1}{2\, s_{xy}} \left[s_y^2 - s_x^2 + \sqrt{(s_y^2 + s_x^2)^2 + 4\, s_{xy}} \right]$$

Posiadacze kalkulatorów statystycznych mogą obliczyć kowariancję potrzebną do obliczenia współczynników osi głównej jako:

$s_{xy} = b_{Y \cdot X}\, s_x^2$

95% przedział ufności dla b_1:

$$H = \frac{F_{0,05\,[1,\,n-2]}}{(\lambda_1/\lambda_2 + \lambda_2/\lambda_1 - 2)\ (n-2)}$$

gdzie F_α, ν_1, ν_2 — wartość krytyczna z rozkładu F dla $\alpha = 95\%$, $\nu_1 = 1$ i $\nu_2 = n-2$.

$L_1 = \mathrm{tg}\,(\mathrm{arc\ tg}\, b_1 - 1/2\ \mathrm{arc\ sin}\ 2\sqrt{H})$

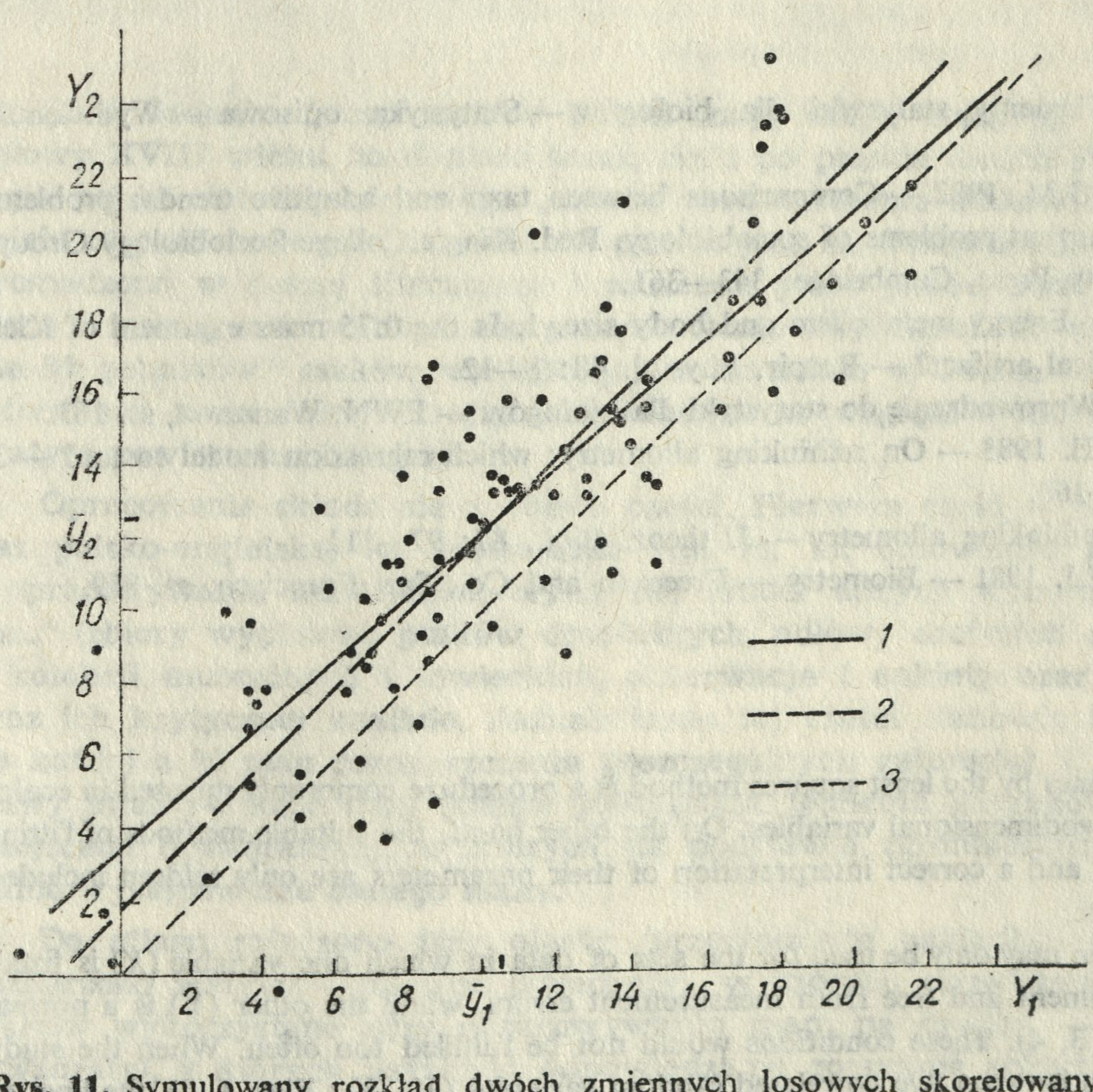
$L_2 = \mathrm{tg}\,(\mathrm{arc\ tg}\, b_1 + 1/2\ \mathrm{arc\ sin}\ 2\sqrt{H})$

Dodatek 2

Przykład liczbowy

Często stosowana weryfikacja matematycznych modeli systemów ekologicznych polega na badaniu korelacji pomiędzy wartościami zmierzonymi w eksperymencie a przewidywanymi przez model. Analiza takiej korelacji pozwala wykryć i zidentyfikować ewentualne błędy systematyczne zawarte w modelu. W idealnej sytuacji wartości modelowe pokrywają się z eksperymentalnymi, a więc na układzie współrzędnych układają się wzdłuż przekątnej, czyli prostej, pod kątem 45°. Jeżeli korelacja nie jest całkowita, a linia dopasowana do rozkładu badanych zmiennych przebiega ukośnie do przekątnej, to może to wskazywać na istnienie w modelu systematycznego błędu, proporcjonalnego do wartości badanych zmiennych. Jeżeli natomiast badana linia jest równoległa do przekątnej, ale przesunięta o stałą wartość, sugeruje to, że systematyczny błąd zawarty w modelu ma charakter addytywny i niezależny od wartości zmiennych.

W celu zilustrowania, jaki wpływ wywiera dobór metody dopasowania linii w takiej analizie, przeprowadzono symulację kalkulatorową w myśl następujących założeń: zmienna Y_1 imituje wartości eksperymentalne, a Y_2 – wartości „modelowe". Wartość zmiennej Y_1 losowana jest



Rys. 11. Symulowany rozkład dwóch zmiennych losowych skorelowanych. Dopasowana linia regresji ($Y_2=3{,}43+0{,}87\ Y_1$) przebiega ukośnie do przekątnej ($Y_1=Y_2$), podczas gdy oś główna zredukowana ($Y_2=1{,}76+1{,}03\ Y_1$) jest do niej prawie równoległa, zgodnie z założeniami symulacji ($Y_2=2+Y_1$)

1 — regresja, *2* — oś główna zredukowana, *3* — przekątna

The simulated distribution of two correlated random variables. The regression line fitted to the data ($Y_2=3.43+0.87\ Y_1$) goes obliquely to the diagonal ($Y_1=Y_2$) while the reduced major axis ($Y_2=1.76+1.03\ Y_1$) is almost parallel to the diagonal and suit the conditions of the simulation ($Y_2=2+Y_1$)

1 — regression, *2* — reduced major axis, *3* — diagonal

z rozkładu normalnego o parametrach $\bar{Y}_1=10$, $SD_1=5$, i obciążona jest błędem ε_1, również losowanym z rozkładu normalnego ($\bar{\varepsilon}_1=0$, $SD_{\varepsilon 1}=2$). Wartości Y_2 równe są wylosowanej wartości Y_1 bez błędu, a do nich dodawany jest błąd ε_2 losowany niezależnie z rozkładu o takich samych parametrach jak błąd dla Y_1. Aby zasymulować istnienie w „modelu" błędu systematycznego, do wszystkich wartości Y_2 dodawana jest jeszcze stała (+2). A zatem gdyby zmienne Y_1 i Y_2 były całkowicie skorelowane ($\varepsilon_1=\varepsilon_2=0$, $r=1$), wówczas dopasowana linia miałaby postać: $Y_2=2+1Y_1$, zaś gdyby nie było założonego błędu systematycznego, to $Y_2=Y_1$.

Rozkład 100 par tych zmiennych przedstawia rys. 11. W przeprowadzonej symulacji współczynnik korelacji wyniósł $r=0{,}85$, a obliczona linia regresji ma postać: $Y_2=3{,}43+0{,}87\ Y_1$. Współczynnik nachylenia regersji różni się więc wyraźnie od spodziewanej wartości 1. Tymczasem równanie osi głównej zredukowanej ma współczynniki: $Y_2=1{,}76+1{,}03\ Y_1$ ($s_v=s_b=1{,}03$).

Gdyby symulowany przykład dotyczył prawdziwego badania zgodności modelu z rzeczywistością, wówczas zastosowanie regresji (nachylenie linii mniejsze od 1) kazałoby szukać w modelu błędu systematycznego o wartościach odwrotnie proporcjonalnych do wartości zmiennych. Natomiast oś główna zredukowana, przebiegająca równolegle do przekątnej, ujawnia błąd addytywny (taki właśnie, jaki był założony w symulacji), niezależny od wartości zmiennych.

Piśmiennictwo

Bogucki Z. 1979 — Elementy statystyki dla biologów — Statystyka opisowa — Wyd. nauk. UAM, Poznań, ss. 120.

Harvey P.H., Mace G.M. 1982 — Comparisons between taxa and adaptive trends: problems of methodology (W: Current problems of sociobiology. Red. King's College Sociobiology Group) — Cambridge University Press, Cambridge, 343—361.

Heusner A.A. 1982 — Energy metabolism and body size. 1. Is the 0.75 mass exponent of Kleiber's equation a statistical artifact? — Respir. Physiol. 48: 1—12.

Parker R.E. 1978 — Wprowadzenie do statystyki dla biologów — PWN, Warszawa, ss. 163.

Seim E., Saether B.-E. 1983 — On rethinking allometry: which regression model to use? — J. theor. Biol. 104: 161—168.

Smith R.J. 1980 — Rethinking allometry — J. theor. Biol., 87: 97—111.

Sokal R.R., Rohlf F.J. 1981 — Biometry — Freeman and Co., San Francisco, ss. 859.

Summary

Fitting regression lines by the least squares method is a procedure commonly misused in ecological studies applying twodimensional variables. On the other hand, the suitable methods of fitting lines to correlated data and a correct interpretation of their parameters are only seldom included in statistical handbooks.

The linear regression may only be used for the sets of data in which one variable (X) is fixed, controlled in the experiment and free from measurement errors, while the other (Y) is a normal random variable (Figs. 3, 4). These conditions would not be fulfilled too often. When the study involves the relationship between two correlated random variables (Y_1, Y_2; Figs. 1, 2), the lines of the best fit should be calculated by the method of major axis (Fig. 5) or reduced major axis (Figs. 6, 7, 8), depending upon the distribution of errors at the variables under study. If the error variances are proportional to the variances of the variables, then the reduced major axis represent perfectly the line of the best fit. When the assumption that the error variances of the both variables are equal can be made, the major axis should be applied. Using an illegal method of line fitting may result in an erroneous estimate of the slope coefficient (Fig. 9) and consequently, invalid conclusions (Fig. 10). The computation formulae for the parameters of all three lines of the best fit, standard errors of slope coefficients, and 95% confidence intervals are given in Appendix 1. A numerical example is presented in Appendix 2 (Fig. 11).