BIULETYN INSTYTUTU HODOWLI I AKLIMATYZACJI ROŚLIN Nr 300 / 2023 : 23–32
E-ISSN: 2657–8913 DOI: 10.37317/biul-2023-0009

Artykuł Przeglądowy

Review Paper

Zmienne towarzyszące jako dodatkowe źródło zmienności w doświadczeniu

Covariates as an additional source of variability in the experiment

Dariusz R. Mańkowski [1)] ✉, Janusz Wątroba [2)]

[1)] Instytut Hodowli i Aklimatyzacji Roślin – Państwowy Instytut Badawczy w Radzikowie
[2)] StatSoft Polska Sp. z o. o.

✉ d.mankowski@ihar.edu.pl

Zmienne towarzyszące charakteryzują i mogą uzupełniać zmienne analizowane w danym doświadczeniu. Mogą one tłumaczyć różne reakcje obserwowanych jednostek doświadczalnych na badane zjawisko lub proces. Podstawowymi narzędziami statystycznymi służących do analizy danych doświadczalnych z uwzględnieniem zmiennych towarzyszących są analiza regresji i analiza kowariancji. O ile analiza funkcji regresji jest metodą często stosowaną w analizie danych doświadczalnych, to z analizą kowariancji już tak nie jest. Prac prezentujących praktyczne wykorzystanie tej metody jest stosunkowo niewiele. Celem niniejszej pracy jest przybliżenie analizy kowariancji i przedstawienie jej praktycznego wykorzystania. W pracy opisano metodę analizy kowariancji na przykładzie klasyfikacji pojedynczej oraz pokazano jej wykorzystanie na dwóch przykładach rolniczych.

Słowa kluczowe: analiza danych, ANCOVA, zmienne towarzyszące, statystyka matematyczna

Covariates characterize and may complement the variables analyzed in the experiment. They can explain various reactions of observed experimental units to the studied phenomenon or process. The basic statistical tools used to analyze experimental data considering covariates are regression analysis and analysis of covariance. While regression function analysis is often used in analyzing experimental data, this is not the case with the analysis of covariance. There are relatively few works presenting the practical use of this method. The aim of this paper is to present the analysis of covariance and its practical use. The paper describes the method of analysis of covariance on the example of a single classification and its use is shown in two agronomical examples.

Keywords: data analysis, ANCOVA, covariates, mathematical statistics

Wprowadzenie

Niepełna informacja o danych, które poddajemy analizie statystycznej w pewnych przypadkach może doprowadzić nas do uzyskania fałszywych wniosków. Jako jeden z często przytaczanych przykładów takiej sytuacji przedstawia się podejrzenia wysunięte w roku 1973 wobec Uniwersytetu Kalifornii w Berkley dotyczące dyskryminacji ze względu na płeć w procesie naboru studentów (Bickel i in., 1975; Kievit i in., 2013). Obserwując łączną liczbę kandydatów (4 321 kobiet oraz 8 442 mężczyzn) oraz liczbę osób przyjętych na studia można odnieść wrażenie, że chętniej przyjmowano mężczyzn niż kobiety (spośród kandydujących dostało się 44% mężczyzn i tylko 35% kobiet). Jednak dopiero analizując odsetki kobiet i mężczyzn przyjętych na poszczególne 101 kierunków studiów okazało się, że nie ma żadnych nieprawidłowości – kobiety chętniej zdawały na kierunki trudniejsze o większej liczbie chętnych na jedno miejsce, natomiast mężczyźni wybierali „bezpieczniejsze" kierunki. Dlatego też obserwowano relatywnie więcej nieprzyjętych kobiet niż mężczyzn. Takie zjawisko, polegające na niewłaściwej lub wręcz błędnej interpretacji danych z powodu braku informacji o właściwościach tych danych (ich strukturze) określane jest mianem Paradoksu Simpsona (Pearson i in., 1899; Yule, 1903; Simpson, 1951).

Czy z taką sytuacją można zetknąć się prowadząc doświadczenia rolnicze i z zakresu hodowli roślin? Jak najbardziej tak. Można tu spotkać błędy wynikające z braku informacji o warunkach panujących na polu podczas prowadzenia doświadczeń polowych (np. zmienność systematyczna), czy też brak dodatkowych informacji charakteryzujących badane obiekty (np. czy w obrębie badanej zbiorowości nie występują subpopulacje). Często by uniknąć błędnej interpretacji i niepoprawnego wnioskowania wystarczy uwzględnić w analizie dodatkowe zmienne, które będą uzupełniały i objaśniały zmienność analizowanych zmiennych. Takie dodatkowe zmienne zwane są często zmiennymi towarzyszącymi (ang. covariates) lub zakłócającymi.

Jedną z najczęściej wykorzystywanych metod statystycznej analizy danych w badaniach rolniczych jest analiza wariancji. Pozwala ona na porównanie kilku obiektów (poziomów czynnika) pod względem badanej zmiennej (zmiennej zależnej), gdy podlega ona dla każdego poziomu czynnika tylko zmienności losowej i nie jest zależna od innych czynników. W przypadku występowania zmienności systematycznej w obrębie doświadczenia, można uwzględnić taką zmienność w mo-

Redaktor Wiodący / Leading Editor
Tomasz Góral

delu analizy wariancji tak by zmienność ta nie rzutowała na ocenę istotności różnic pomiędzy badanymi obiektami. Zdarza się jednak tak, że analizowana zmienna zależna jest uwarunkowana wpływem innym zmiennych mierzalnych – tzw. zmiennych towarzyszących (Cochran i Cox, 1992). Jak podaje Elandt (1964) analiza kowariancji jest metodą, która pozwala uwzględnić w pewien sposób wpływ zmiennej towarzyszącej i porównać średnie obiektowe przez sprowadzenie wspólnego mianownika, to znaczy pozwala tak poprawić średnie obiektowe jak by były wyznaczone dla tej samej wartości zmiennej towarzyszącej.

Analiza kowariancji jest techniką statystyczną, która łączy metody analizy wariancji (ANOVA) i analizy regresji (Timm, 2002). ANCOVA jest stosunkowo rzadko stosowana i opisywana w badaniach z zakresu szeroko pojętej agronomii (Wijesuriya i Thattil, 1996; Yang i Juskiw, 2011; McConnell i in., 2014).

Celem niniejszej pracy było przybliżenie analizy kowariancji i przedstawienie możliwości jej wykorzystania w praktyce doświadczalnictwa rolniczego.

Opis metody

Poniżej przedstawiono opis metody analizy kowariancji dla klasyfikacji pojedynczej. Dla układów wieloczynnikowych stosuje się rozwinięcie przedstawionych obliczeń uwzględniające zastosowany układ doświadczalny.

Założenia

Analizę kowariancji jest uznawana za połączenie analizy wariancji z analizą regresji (Stanisz, 2007), w związku z czym przed przystąpieniem do analizy należy sprawdzić wszystkie założenia leżące u podstaw obydwu tym metod. Dodatkowo należy uwzględnić jeszcze dwa kluczowe dla tej analizy założenia (Elandt, 1964; Stanisz, 2007):

— Założenie o jednorodności lub równoległości regresji w obrębie grup (obiektów, poziomów czynnika).
— Założenie o braku wpływu oddziaływania eksperymentalnego na zmienną towarzyszącą.

Założenie o równoległości regresji jest podstawowym założeniem analizy kowariancji. W celu sprawdzenia tego założenia stawiana jest hipoteza postaci:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p \qquad (1)$$

gdzie β_i to współczynniki regresji (nachylenia funkcji regresji) w *i*-tej populacji (dla *i*-tego poziomu czynnika / obiektu). Tak postawioną hipotezę testuje się za pomocą testu *F* lub za pomocą modelu jednakowych nachyleń (Stanisz, 2007). Model jednakowych nachyleń pozwala na testowanie istotności interakcji zmiennej towarzyszącej ze zmienną grupującą (czynnikiem). Jeśli interakcja ta jest istotna to założenie o równości współczynników regresji (równoległości regresji) nie jest spełnione.

Drugie z wskazanych powyżej założeń odnosi się do niezależności zmiennej towarzyszącej od zmiennej grupującej. Spełnienie tego założenia zapewnia, że obserwowane różnice pomiędzy średnimi obiektowymi wynikają z różnic pomiędzy jednostkami doświadczalnymi (spowodowanymi przez badany czynnik), a nie są efektem oddziaływania zmiennej towarzyszącej (Stanisz, 2007). Założenie to najlepiej sprawdzić za pomocą testu *F* analizy wariancji.

Uwzględnienie zmiennej towarzyszącej w analizie wariancji wymaga dodatkowo wykazania, że występuje zależność pomiędzy zmienną towarzyszącą a zmienną zależną. Można to zrobić za pomocą zwykłej analizy regresji liniowej.

Na koniec pozostaje więc pytanie, kiedy analizę kowariancji można zastosować, a kiedy pozostaje tylko analiza wariancji lub analiza regresji. Przedstawiony na Rysunku 1 schemat jest graficznym posumowaniem tego problemu.

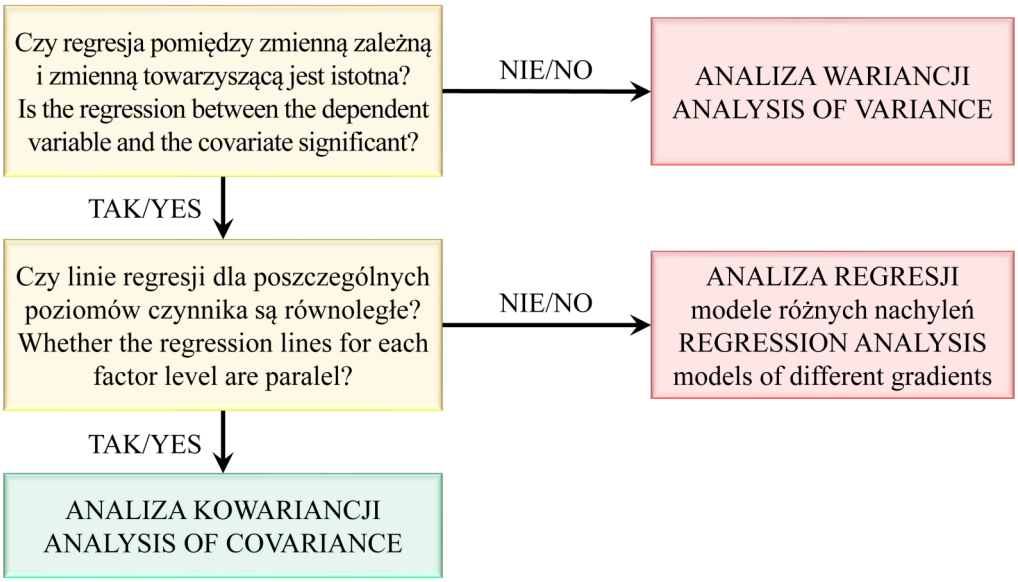
Model obliczeniowy

Rozważając *k* grup (poziomów czynnika / obiektów). Przy liczbie powtórzeń wynoszącej n_i dla *i*-tej grupy (*i*-tego poziomu czynnika / obiektu) oraz łącznej liczbie przypadków (powtórzeń) wynoszącej $N = \sum_{i=1}^{k} n_i$, zmienną zależną o rozkładzie zbieżnym z rozkładem normalnym oznacza się przez *y*, natomiast *x* oznacza zmienną towarzyszącą. Zakładając, że jej zmienność ma charakter losowy a wszystkie założenia analizy kowariancji są spełnione, model liniowy w klasyfikacji pojedynczej można zapisać w postaci (Elandt, 1964; Oktaba, 1971; Szklarska i in., 1978; Wójcik i Laudański, 1989; Cochran i Cox, 1992; Montgomery, 2005):

$$y_{ij} = \mu + \alpha_i + \beta\,(x_{ij} - \overline{x}_{..}) + \varepsilon_{ij} \qquad (2)$$

gdzie y_{ij} to wartość zmiennej zależnej dla *i*-tego poziomu czynnika w *j*-tym powtórzeniu; μ to średnia ogólna wartość zmiennej zależnej dla całej populacji; α_i to efekt *i*-tego poziomu badanego czynnika; $\beta\,(x_{ij} - \overline{x}_{..})$ reprezentuje udział zmiennej towarzyszącej; β to współczynnik regresji wewnątrz grup (dla poszczególnych poziomów czynnika); x_{ij} to obserwacja zmiennej towarzyszącej dla *i*-tego poziomu badanego czynnika w *j*-tym powtórzeniu, $\overline{x}_{..}$ - oznacza średnią wartość zmiennej towarzyszącej w całym doświadczeniu, a ε_{ij} to błąd losowy dla i-tego poziomu czynnika w j-tym powtórzeniu.

Parametry μ, α_i oraz β modelu (2) są nieznane, a oszacować można je za pomocą metody najmniejszych kwadratów (Elandt, 1964; Linnik, 1962). Dąży się więc, by suma kwadratów odchyleń postaci (Elandt, 1964):



Rys. 1. Schemat wyboru analizy (Stanisz, 2007)
Fig. 1. Analysis selection scheme (Stanisz, 2007)

$$Se = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\{y_{ij} - [\mu + \alpha_i + \beta(x_{ij} - \bar{x}_{..})]\}^2 \quad (3)$$

była jak najmniejsza. Analogicznie jak w klasycznej analizie wariancji estymatory poszczególnych parametrów modelu (2) można zapisać w postaci (Elandt, 1964; Wójcik i Laudański, 1989):

$$\begin{gathered}\hat{\mu} = \bar{y}_{..} \\ \hat{a}_i = (\bar{y}_{i.} - \bar{y}_{..}) - b(\bar{x}_{i.} - \bar{x}_{..}) \\ \hat{\beta} = b\end{gathered} \quad (4)$$

przy czym b będące oszacowaniem wspólnego współczynnika regresji wewnątrz grup (w obrębie poszczególnych poziomów czynnika) wyznacza się wg wzoru postaci (Elandt, 1964; Oktaba, 1971):

$$b = \frac{\sum_i^k \sum_j^{n_i}(x_{ij} - \bar{x}_{i.})(y_{ij} - \bar{y}_{i.})}{\sum_i^k \sum_j^{n_i}(x_{ij} - \bar{x}_{i.})^2} \quad (5)$$

A więc szacowana wartość zmiennej zależnej przyjmuje postać (Elandt, 1964):

$$\begin{gathered}\hat{y}_{ij} = \bar{y}_{..} + [(\bar{y}_{i.} - \bar{y}_{..}) - b(\bar{x}_{i.} - \bar{x}_{..})] + \\ + b(x_{ij} - \bar{x}_{..}) = \bar{y}_{i.} + b(x_{ij} - \bar{x}_{i.})\end{gathered} \quad (6)$$

Teraz można wyznaczyć sumy kwadratów odchyleń i iloczynów zmiennych x oraz y odpowiednio dla zmienności ogólnej (T), obiektowej (A) oraz losowej (E) czyli błędu losowego (Elandt, 1964; Oktaba, 1971; Szklarska i in., 1978; Wójcik i Laudański, 1989; Montgomery, 2005):

$$\begin{gathered}\mathrm{var}T_{xx} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}^2 - \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}\right)^2}{n} \\ \mathrm{var}T_{yy} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}y_{ij}^2 - \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}y_{ij}\right)^2}{n} \\ \mathrm{var}T_{xy} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{..})(y_{ij} - \bar{y}_{..}) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}y_{ij} - \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}\sum_{i=1}^{k}\sum_{j=1}^{n_i}y_{ij}}{n}\end{gathered} \quad (7)$$

$$\begin{gathered}\mathrm{var}A_{xx} = \sum_{i=1}^{k}n_i(\bar{x}_{i.} - \bar{x}_{..})^2 = \sum_{i=1}^{k}\frac{\left(\sum_{j=1}^{n_i}x_{ij}\right)^2}{n_i} - \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}\right)^2}{n} \\ \mathrm{var}A_{yy} = \sum_{i=1}^{k}n_i(\bar{y}_{i.} - \bar{y}_{..})^2 = \sum_{i=1}^{k}\frac{\left(\sum_{j=1}^{n_i}y_{ij}\right)^2}{n_i} - \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}y_{ij}\right)^2}{n} \\ \mathrm{var}A_{xy} = \sum_{i=1}^{k}n_i(\bar{x}_{i.} - \bar{x}_{..})(\bar{y}_{i.} - \bar{y}_{..}) = \sum_{i=1}^{k}\frac{\sum_{j=1}^{n_i}x_{ij}\sum_{j=1}^{n_i}y_{ij}}{n_i} - \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}\sum_{i=1}^{k}\sum_{j=1}^{n_i}y_{ij}}{n}\end{gathered} \quad (8)$$

$$\begin{aligned}
\mathrm{var}E_{xx} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{i\cdot})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}^2 - \sum_{i=1}^{k}\frac{\left(\sum_{j=1}^{n_i}x_{ij}\right)^2}{n_i}\\
\mathrm{var}E_{yy} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}y_{ij}^2 - \sum_{i=1}^{k}\frac{\left(\sum_{j=1}^{n_i}y_{ij}\right)^2}{n_i}\\
\mathrm{var}E_{xy} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{i\cdot})(y_{ij}-\bar{y}_{i\cdot}) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}y_{ij} - \sum_{i=1}^{k}\frac{\sum_{j=1}^{n_i}x_{ij}\sum_{j=1}^{n_i}y_{ij}}{n_i}
\end{aligned} \tag{9}$$

Sumy kwadratów odchyleń po uwzględnieniu regresji mają postać (Elandt, 1964):

$$\mathrm{var}T = \mathrm{vat}T_{yy} - \frac{\mathrm{var}T_{xy}^2}{\mathrm{var}T_{xx}} \tag{10}$$

$$\mathrm{var}A = \mathrm{vat}A_{yy} - \frac{\mathrm{var}A_{xy}^2}{\mathrm{var}A_{xx}} \tag{11}$$

$$\mathrm{var}E = \mathrm{vat}E_{yy} - \frac{\mathrm{var}E_{xy}^2}{\mathrm{var}E_{xx}} \tag{12}$$

Ponieważ w tym przypadku nie zgadza się liczba stopni swobody, a różnica wynosi 1 stopień swobody, możemy zapisać, że (Elandt, 1964):

$$\mathrm{var}T - \mathrm{var}A = \mathrm{var}E + D \tag{13}$$

gdzie D odpowiada 1 stopniowi swobody i przedstawia różnicę, którą trzeba dodać by zastosować test F do testowania istotności poprawionych średnich obiektowych (Elandt, 1964):

$$\mathrm{var}A' = \mathrm{var}A + D = \mathrm{var}T - \mathrm{var}E \tag{14}$$

Tabelę analizy kowariancji dla klasyfikacji pojedynczej przedstawiono w Tabeli 1, a statystyka testowa F ma postać (Elandt, 1964; Oktaba, 1971; Szklarska i in., 1978; Wójcik i Laudański, 1989; Montgomery, 2005):

$$F = \frac{s_{A'}^2}{s_E^2} = \frac{\mathrm{var}A'}{\mathrm{var}E}\cdot\frac{n-k-1}{k-1} \tag{15}$$

Tabela 1
Tabele 1

Tabela analizy kowariancji (Elandt, 1964; Oktaba, 1971; Szklarska i in., 1978)
Analysis of covariance table (Elandt, 1964; Oktaba, 1971; Szklarska et al., 1978)

Źródła zmienności Source of variation	Stopnie swobody df	Sumy kwadratów i iloczynów odchyleń Sum of squares and cross products			Odchylenia od regresji Deviations from regression		
		$\sum(x-\bar{x})^2$	$\sum(y-\bar{y})^2$	$\sum(x-\bar{x})\cdot(y-\bar{y})$	Stopnie swobody df	$\sum(y-\hat{y})^2$	Średni kwadrat odchyleń Mean square
Między grup Between groups	$k-1$	$\mathrm{var}A_{xx}$	$\mathrm{var}A_{yy}$	$\mathrm{var}A_{xy}$	$k-2$	$\mathrm{var}A$	
Wewnątrz grup Within groups	$n-k$	$\mathrm{var}E_{xx}$	$\mathrm{var}E_{yy}$	$\mathrm{var}E_{xy}$	$n-k-1$	$\mathrm{var}E$	s_E^2
Całkowita Total	$n-1$	$\mathrm{var}T_{xx}$	$\mathrm{var}T_{yy}$	$\mathrm{var}T_{xy}$	$n-2$	$\mathrm{var}T$	
				Reszta Residual	1	D	
Suma kwadratów odchyleń do testowania poprawionych średnich obiektowych Sum of squared for testing adjusted object means					$k-1$	$\mathrm{var}A' = \mathrm{var}T - \mathrm{var}E$	$s_{A'}^2$

Obserwacji i średnie poprawione

Każdą obserwację y_{ij} można poprawić wg wzoru (Elandt, 1964; Szklarska i in., 1978; Cochran i Cox, 1992;):

$$y'_{ij} = y_{ij} - b_w(x_{ij}-\bar{x}_{..}) \tag{16}$$

gdzie y'_{ij} to wartość zmiennej zależnej poprawiona przy pomocy regresji.

Jeżeli w wyniku przeprowadzonego testowania testem F (wzór (15)) stwierdza się, że występują istotne różnice pomiędzy poprawionymi średnimi obiektowymi, należy takie średnie wyznaczyć zgodnie z wzorem (Elandt, 1964; Wójcik i Laudański, 1989; Montgomery, 2005):

$$\bar{y'}_{i\cdot} = \bar{y}_{i\cdot} - b(\bar{x}_{i\cdot}-\bar{x}_{..}) \tag{17}$$

gdzie $\bar{y'}_{i\cdot}$ to średnia poprawiona dla i-tego obiektu (poziomu czynnika).

Aby możliwe było porównanie średnich poprawionych można wykorzystać jedną z procedur porównań wielokrotnych. Jednak będzie do tego potrzebna ocena błędu różnicy pomiędzy średnimi poprawionymi. Błąd ten określany jest wzorem (Elandt, 1964; Szklarska i in., 1978; Wójcik i Laudański, 1989; Montgomery, 2005):

$$s_{\bar{y'}_{i\cdot}-\bar{y'}_{j\cdot}}^2 = s_E^2\left[\left(\frac{1}{n_i}+\frac{1}{n_j}\right)+\frac{(\bar{x}_{i\cdot}-\bar{x}_{j\cdot})^2}{\mathrm{var}E_{xx}}\right] \tag{18}$$

gdzie $\bar{y}'_{i\cdot}$ to pierwsza poprawiona średni obiektowa, a $\bar{y}'_{j\cdot}$ to druga poprawiona średnia obiektowa.

Przykłady

Poniżej przedstawiono dwa przykłady zastosowania analizy kowariancji w analizie danych pochodzących z doświadczeń rolniczych. Pierwszy przykład przedstawia analizę kowariancji dla klasyfikacji pojedynczej, a drugi przykład – analizę kowariancji dla danych pochodzących z doświadczenia założonego w układzie bloków losowych.

Wszystkie obliczenia wykonano w programie Statistica w wersji 13.3 (TIBCO Software Inc., 2017).

Przykład 1 – klasyfikacja pojedyncza

Przedstawione dane doświadczalne zostały zaprezentowane przed Elandt (1964). W opisywanym eksperymencie porównywano wytrzymałość słomy trzech odmian lnu (S – Śląski, L – LCSD oraz Z – Zwisły) uwzględniając przy tym grubość słomy. Założono, że zmienna grubość słomy w tym przypadku była źródłem dodatkowej zmienności, która mogła rzutować na obserwowane wyniki porównania i je zaburzać (błąd systematyczny). Ocenę wytrzymałości słomy prowadzono na dynamometrze. W każdym powtórzeniu badano 30 łodyg, a analizie poddano wartości średnie. Dane źródłowe zestawiono w Tabeli 2.

Tabela 2
Table 2

Dane źródłowe – Wytrzymałość słomy w g trzech odmian lnu w zależności od grubości łodygi w mm (Elandt, 1964)
Source data – Straw strength in g of three flax varieties depending on the stem thickness in mm (Elandt, 1964)

Powtórzenie Replication	Odmiana S S cultivar		Odmiana L L cultivar		Odmiana Z Z cultivar	
	Wytrzymałość słomy Straw strength	Grubość pędu Stem thickness	Wytrzymałość słomy Straw strength	Grubość pędu Stem thickness	Wytrzymałość słomy Straw strength	Grubość pędu Stem thickness
1	562	10	550	10	420	9
2	810	13	645	11	561	11
3	891	14	868	13	620	12
4	1072	17	1023	15	778	14
5	1227	19	1217	18	820	15

Analizę rozpoczęto od klasycznej jednoczynnikowej analizy wariancji dla wytrzymałości słomy. Czynnikiem różnicującym była odmiana. Tabela analizy wariancji dla takiego układu została przedstawiona w Tabeli 3. Uzyskane wyniki sugerują brak istotnego zróżnicowania średnich wartości wytrzymałości słomy pomiędzy badanymi odmianami.

Tabela 3
Table 3

Analiza wstępna – tabela analizy wariancji dla modelu jednoczynnikowego
Preliminary analysis – variance analysis table for a one-factor model

Źródła zmienności Source of variation	Suma kwadratów odchyleń Sum of squares	Stopnie swobody Degrees of freedom	Średni kwadrat odchyleń Mean square	*F*	*p*
Czynnik: Odmiana Factor: Cultivar	209578	2	104789	1,9022	0,1916
Błąd losowy Random error	661059	12	55088		

Sprawdzono, czy występuje współzależność pomiędzy wytrzymałością słomy a grubością pędu. W tym celu wykonano analizę współczynników korelacji liniowej Pearsona. Wyniki wskazywały na występowanie istotnej statystycznie, bardzo silnej i wprost proporcjonalnej (r = 0,9693, p < 0,0001) relacji pomiędzy badanymi zmiennymi.

Uzyskane wyniki świadczyły o tym, że analizując wyłącznie wytrzymałość słomy nie dało się wykazać różnic odmianowych. Jednakże równocześnie występowała istotna zależność pomiędzy wytrzymałością słomy a grubością łodygi. Dlatego zasadnym wydawało się postawienie hipotezy mówiącej, że różnice pomiędzy odmianami są zacierane przez fakt, że wyniki pomiaru wytrzymałości słomy ściśle zależą od grubości łodygi. Dlatego postanowiono przeprowadzić analizę kowariancji w celu stwierdzenia, czy rzeczywiście grubość słomy wpływa na różnice odmianowe w wytrzymałości słomy.

Kolejnym krokiem było sprawdzenie założeń analizy kowariancji. Tu skupino się na założeniach przedstawionych na Rysunku 1. Ponieważ w analizie współczynników korelacji liniowych Pearsona stwierdzono występowanie istotnej współzależności liniowej pomiędzy zmienną zależną a zmienną towarzyszącą analiza funkcji regresji liniowej opisującej tę relację była formalnością. Wyniki analizy regresji przedstawiono w Tabeli 4. Stwierdzono występowanie istotnej regresji

liniowej pomiędzy wytrzymałością słomy, a grubością łodygi lnu.

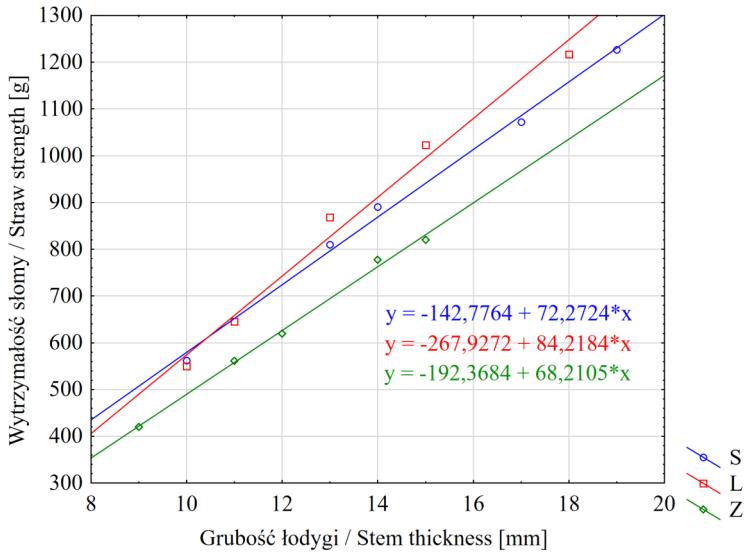
Kolejnym założeniem, które należało sprawdzić było założenie o równości współczynników regresji dla poszczególnych odmian. Założenie to sprawdzono za pomocą analizy modelu jednakowych nachyleń. Wyniki analizy wariancji dla modelu jednakowych nachyleń przedstawiono w Tabeli 5. Wykazano brak istotnego efektu interakcyjnego pomiędzy odmianą a grubością łodygi. Świadczyło to o braku podstaw do odrzucenia hipotezy zakładającej równość współczynników regresji, czyli o spełnieniu ostatniego założenia analizy kowariancji. Dodatkowo sporządzono wykres rozrzutu z naniesionymi prostymi regresji (Rys. 2).

Tabela 4
Table 4
Analiza założeń – wyniki analizy funkcji regresji liniowej pomiędzy zmienną zależną a zmienną towarzyszącą
Analysis of assumptions – results of an analysis of the linear regression function between dependent variable and covariate

Parametr Parameter	Ocena Estimate	Statystyka testowa Test statistic	*p*
Stała regresji Slope	–268,615	t = –3,4767	0,0041
Współczynnik regresji (Grubość łodygi) Coefficient of regression (Stem thickness)	80,066	t = 14,2114	< 0,0001
Model regresji Regression model		F = 201,9624	< 0,0001
R^2	0,9395		

Tabela 5
Table 5
Analiza założeń – wyniki analizy wariancji dla modelu jednakowych nachyleń
Analysis of assumptions – results of variance analysis for model of equal slopes

Źródła zmienności Sourse of variation	Suma kwadratów odchyleń Sum of squares	Stopnie swobody Degrees of freedom	Średni kwadrat odchyleń Mean square	*F*	*p*
Odmiana Cultivar	1741,1	2	870,6	1,3577	0,3053
Grubość łodygi Stem thickness	570795,5	1	570795,5	890,1721	< 0,0001
Odmiana × Grubość łodugi Cultivar × Stem thickness	4845,4	2	2422,7	3,7783	0,0644
Błąd losowy Random error	5771,0	9	641,2		



Rys. 2. Wykres rozrzutu z naniesionymi prostymi regresji wytrzymałości słomy względem grubości łodygi dla trzech badanych odmian lnu
Fig. 2. Scatter plot with simple regression of straw strength in relation to the stem thickness for three analyzed flax cultivars

Założenia zostały spełnione – można więc było przeprowadzić właściwą analizę kowariancji. Tabelę analizy kowariancji przedstawiono w Tabeli 6. Uzyskane wyniki świadczyły o istotności statystycznej różnic średnich wytrzymałości słomy pomiędzy badanymi odmianami. Potwierdzono również, że występowała statystycznie istotna zależność pomiędzy obserwowanymi wytrzymałościami słomy a grubością łodyg. Tu należy zwrócić uwagę na fakt, że gdy nie uwzględnialiśmy w analizie zmiennej towarzyszącej, nie można było stwierdzić, że występowały istotne różnice

Tabela 6
Table 6

Wyniki analizy kowariancji dla wytrzymałości słomy względem badanych odmian lnu z uwzględnieniem grubości łodyg
The results of the analysis of covariance for straw strength in relation to the tested varieties of flax, taking into account the thickness of stems

Źródła zmienności Sourse of variation	Suma kwadratów odchyleń Sum of squares	Stopnie swobody Degrees of freedom	Średni kwadrat odchyleń Mean square	*F*	*p*
Grubość łodygi Stem thickness	650442,8	1	650442,8	673,9472	< 0,0001
Odmiana Cultivar	42036,0	2	21018,0	21,7775	0,0002
Błąd losowy Random error	10616,4	11	965,1		

pomiędzy średnimi wytrzymałościami słomy dla badanych odmian (Tab. 3).

Przedstawiony model analizy kowariancji pozwalał na wyjaśnienie ponad 98% zmienności zmiennej wytrzymałość słomy lnu (współczynnik determinacji dla tego modelu R^2 = 0,9878). Skoro udało się wyjaśnić zmienność zmiennej zależnej za pomocą modelu analizy kowariancji, to należałoby oszacować jeszcze średnie skorygowane dla poszczególnych odmian. Zestawienie średnich surowych i średnich poprawionych wytrzymałości słomy lnu dla badanych odmian pokazano w Tabeli 7.

Tak wyznaczone średnie poprawione obrazują różnicę w wytrzymałości słomy pomiędzy badanymi odmianami lnu przyjmując określoną, prze-

Tabela 7
Table 7

Rzeczywiste średnie brzegowe i średnie skorygowane wytrzymałości słomy dla badanych odmian lnu
Real marginal means and adjusted means straw strengths for tested flax cultivars

Odmiana Cultivar	Rzeczywiste średnie brzegowe Real marginal means				Średnie poprawione* Adjusted means*			
	Ocena Estimate		Błąd standardowy Standard error	Przedział ufności 95% 95% confidence interval	Ocena Estimate		Błąd standardowy Standard error	Przedział ufności 95% 95% confidence interval
S	912,40	a	13,8933	(881,82; 942,98)	821,44	b	14,3284	(789,90; 852,97)
L	860,60	b	13,8933	(830,02; 891,18)	860,60	a	13,8933	(830,02; 891,18)
Z	639,80	c	13,8933	(609,22; 670,38)	730,76	c	14,3284	(699,23; 762,30)

* Oszacowano przyjmując średni poziom grubości łodygi = 13,4 / Estimated assuming an average level of stem thickness = 13.4
Literami oznaczono grupy jednorodne wg testu post-hoc Fishera przy α = 0,05; Letters mark homogeneous groups according to Fisher's post-hoc test at α = 0.05

ciętną grubość łodygi. Porównanie takich średnich poprawionych nie jest więc obciążone wpływem zmiennej towarzyszącej.

Przykład 2 – układ bloków losowych

Drugi przykład stanowią wyniki doświadczenia założonego w układzie bloków losowych. Przedstawione tu dane zostały opisane przez Szklarską i wsp. (1978). W opisywanym doświadczeniu w układzie bloków losowych w czterech blokach badano różnice w plonowaniu pięciu odmian kukurydzy. Oceniano plon ziarna z poletka. Zmienną towarzyszącą była liczba brakujących roślin z poletka. Dane zestawiono w Tabeli 8.

Analizę danych rozpoczęto od klasycznej analizy wariancji dla układu bloków losowych bez uwzględniania zmiennej towarzyszącej (Tab. 9). Wyniki wskazywały na istotne zróżnicowanie pomiędzy średnimi plonami z poletka dla porównywanych odmian.

Jednak mając świadomość, że w doświadczeniu na poszczególnych jednostkach doświadczalnych (poletkach) pomimo wysiana takiej samej liczby roślin, pewnej liczby roślin brakowało, postanowiono sprawdzić, czy obserwowany plon ziarna z poletka był powiązany z liczbą brakujących roślin z poletka. W tym celu przeprowadzono analizę współczynników korelacji liniowej Pearsona. W wyniku tej analizy wyznaczono istotny współczynnik korelacji r = –0,4644 (p = 0,039). Sama wielkość uzyskanego współczynnika korelacji nie była zbyt duża (słaba korelacja odwrotnie proporcjonalna), również wartość prawdopodobieństwa testowego (p) była dość duża (niewiele mniejsza od założonego poziomu istotności α = 0,05), jednak te wyniki wskazywały, że liczba braków roślin z poletka może w istotny sposób wpływać na ocenę przeciętnych plonów z poletka dla porównywanych odmian kukurydzy. By to spraw-

Tabela 8
Table 8

Dane źródłowe – Plon ziarna z poletka pięciu odmian kukurydzy oraz liczba brakujących roślin z poletka (Szklarska i in., 1978)
Source data – Grain yield from a plot of five maize varieties and the number of missing plants from a plot (Szklarska et al., 1978)

Odmiana Cultivar	Blok / Block							
	I		II		III		IV	
	Plon Yield	Liczba brakujących roślin Number of missing plants	Plon Yield	Liczba brakujących roślin Number of missing plants	Plon Yield	Liczba brakujących roślin Number of missing plants	Plon Yield	Liczba brakujących roślin Number of missing plants
1	32	10	30	9	32	13	36	5
2	39	13	35	10	30	12	48	0
3	35	14	33	9	31	24	44	4
4	41	13	36	16	38	22	48	1
5	24	23	32	0	27	11	32	3

Tabela 9
Table 9

Analiza wstępna – tabela analizy wariancji dla modelu jednoczynnikowego w układzie bloków losowych
Preliminary analysis – variance analysis table for a one-factor model in randomized block design

Źródła zmienności Sourse of variation	Suma kwadratów odchyleń Sum of squares	Stopnie swobody Degrees of freedom	Średni kwadrat odchyleń Mean square	*F*	*p*
Blok Block	294,55	3	98,18	9,299	0,0019
Czynnik: Odmiana Factor: Cultivar	351,30	4	87,83	8,318	0,0019
Błąd losowy Random error	126,70	12	10,56		

dzić postanowiono przeprowadzić analizę kowariancji. W kolejnym kroku sprawdzono założenia tej analizy (Rys. 1).

Rozpoczęto od wyznaczenia funkcji regresji liniowej dla zależności plonu ziarna z poletka od liczby brakujących roślin z poletka (Tab. 10). Wyniki analizy regresji liniowej wskazały, że model był istotny. Oznacza to, że pierwsze założenie było spełnione.

Tabela 10
Table 10

Analiza wstępna – tabela analizy wariancji dla modelu jednoczynnikowego w układzie bloków losowych
Preliminary analysis – variance analysis table for a one-factor model in randomized block design

Parametr Parameter	Ocena Estimate	Statystyka testowa Test statistics	*p*
Stała regresji Slope	39,5230	$t = 16,7800$	< 0,0001
Współczynnik regresji (Liczba brakujących roślin) Coefficient of regression (Number of missing plants)		$t = -2,2245$	0,03915
Model regresji Regression model		$F = 4,9484$	0,03915
R^2	0,2156		

Drugie założenie analizy kowariancji dotyczy równości współczynników regresji (jednakowe nachylenia linii regresji) dla poszczególnych poziomów badanego czynnika. Założenie to sprawdzono za pomocą analizy modelu jednakowych nachyleń z uwzględnieniem układu doświadczalnego (Tab. 11). Uzyskane wyniki wskazywały na brak istotnej interakcji pomiędzy czynnikiem (tu odmianą) a zmienną towarzyszącą (liczbą brakujących roślin z poletka). Świadczy to o równoległości linii regresji, czyli równości współczynników regresji.

Obydwa założenia były spełniane, przeprowadzano więc analizę kowariancji z uwzględnieniem układu doświadczenia (Tab. 12). Wszystkie efekty były istotne. Stwierdzono istotne zróżnicowanie odmian co do średnich plonów ziarna z poletka. Stwierdzono również, że występowało istotne oddziaływanie liczby brakujących roślin z poletka na obserwowane średnie plony ziarna z poletka.

Przedstawiony model analizy kowariancji pozwalał na wyjaśnienie ponad 90% zmienności zmiennej plon ziarna kukurydzy z poletka (współczynnik determinacji dla tego modelu $R^2 = 0,9014$). Następnie wyznaczono średnie poprawione plonów ziarna z poletka dla porównywanych odmian pozbawione wpływu liczby brakujących roślin z poletka. Zestawienie średnich suro-

Tabela 11
Table 11

Analiza założeń – wyniki analizy wariancji dla modelu jednakowych nachyleń
Analysis of assumptions – results of variance analysis for model of equal slopes

Źródła zmienności Sourse of variation	Suma kwadratów odchyleń Sum of squares	Stopnie swobody Degrees of freedom	Średni kwadrat odchyleń Mean square	*F*	*p*
Blok Block	66,035	3	22,012	2,8995	0,1112
Odmiana Cultivar	140,708	4	35,177	4,6337	0,0382
Liczba brakujących roślin Number of missing plants	11,550	1	11,550	1,5214	0,2572
Odmiana × Liczba brakujących roślin Cultivar × Number of missing plants	23,005	4	5,751	0,7576	0,5842
Błąd losowy Random error	53,142	7	7,592		

Tabela 12
Table 12

Wyniki analizy kowariancji dla plonu ziarna z poletka dla porównywanych odmian kukurydzy z uwzględnieniem liczby brakujących roślin z poletka
The results of the analysis of covariance for grain yield per plot for the compared maize varieties, taking into account the number of missing plants per plot

Źródła zmienności Sourse of variation	Suma kwadratów odchyleń Sum of squares	Stopnie swobody Degrees of freedom	Średni kwadrat odchyleń Mean square	*F*	*p*
Blok Block	80,428	3	26,809	3,8728	0,0410
Liczba brakujących roślin Number of missing plants	50,553	1	50,553	7,3028	0,0206
Odmiana Cultivar	397,508	4	99,377	14,3558	0,0002
Błąd losowy Random error	76,147	11	6,922		

wych i średnich poprawionych plonu ziarna z poletka dla porównywanych odmian pokazano w Tabeli 13.

Dopiero średnie poprawione stanowiły podstawę do obiektywnego porównania badanych odmian kukurydzy pod względem uzyskanych plonów z poletka. Pomimo, że na początku w przeprowadzonej analizie wariancji uzyskano istotny efekt główny, to uwzględnienie zmiennej towarzyszącej powiązanej ze zmienną zależną sprawiło, że to analiza kowariancji okazała się właściwą metodą oceny różnic pomiędzy badanymi odmianami.

Tabela 13
Table 13

Rzeczywiste średnie brzegowe i średnie skorygowane plonów ziarna z poletka dla porównywanych odmian kukurydzy
Real marginal means and mean corrected grain yields per plot for the compared corn cultivars

Odmiana Cultivar	Rzeczywiste średnie brzegowe Real marginal means				Średnie poprawione* Adjusted means*			
	Ocena Estimate		Błąd standardowy Standard error	Przedział ufności 95% 95% confidence interval	Ocena Estimate		Błąd standardowy Standard error	Przedział ufności 95% 95% confidence interval
1	32,50	ab	1,2583	(28,4955; 36,5045)	31,9672	c	1,3302	(29,0395; 34,8950)
2	38,00	ab	3,8079	(25,8816; 50,1184)	37,2699	b	1,3430	(34,3740; 40,2258)
3	35,75	ab	2,8687	(26,6207; 44,8793)	36,5985	b	1,3525	(33,6217; 39,5753)
4	40,75	a	2,6260	(32,3929; 49,1071)	41,6971	a	1,3614	(38,7007; 44,6936)
5	28,75	b	1,9738	(22,4685; 35,0315)	28,2172	d	1,3302	(25,2895; 31,1450)

* Oszacowano przyjmując średni poziom liczby brakujących roślin = 10,6 / Estimated assuming an average level of number of missing plants = 10.6
Literami oznaczono grupy jednorodne wg testu post-hoc Fishera przy α = 0,05; Letters mark homogeneous groups according to Fisher's post-hoc test at α = 0.05

Podsumowanie

Zdecydowana większość cech w przyrodzie nie jest niezależna. Analizując dowolną zmienną prawie zawsze można wskazać inne zmienne z nią powiązane. Występowanie takich relacji sprawia, że zmienne towarzyszące mogą odgrywać bardzo dużą rolę w objaśnianiu zmienności międzyobiektowej. Należy jednak pamiętać, że nie każda zmienna, nawet spełniająca opisane założenia dla zmiennej towarzyszącej, może być traktowana jako taka zmienna, i być uwzględniana w analizie kowariancji. Przyjmuje się, że zmienne towarzyszące to ciągłe zmienne niezależne, które wpływają na zmienną zależną, ale nie są głównym przedmiotem zainteresowania badania. Ponadto eksperymentatorzy nie kontrolują zmiennych towarzyszących.

W sytuacjach, gdy na pierwszy rzut oka nie da się zidentyfikować różnic międzyobiektowych, zmienne towarzyszące mogą pozwolić na ujawnienie tych różnic. Pomocna może być w tym przypadku analiza kowariancji. Jej zastosowanie pozwala ocenić udział zmiennej towarzyszącej, ale przede wszystkim ocenić różnice pomiędzy badanymi obiektami (średnimi obiektowymi) z uwzględnieniem wpływu zmiennej towarzyszącej.

Jednak w sytuacjach, gdy już na początku stwierdza się występowanie istotnych efektów głównych, ale również stwierdza się występowanie zmiennych towarzyszących istotnie powiązanych ze zmienną zależną, może okazać się, że pominięcie tej zmiennej w analizie i pozostanie wyłącznie przy analizie wariancji może doprowadzić do wyciągnięcia niepełnych lub zafałszowanych wniosków. Dopiero analiza kowariancji pozwala na ocenę istotności efektu głównego z wyłączeniem wpływu zmiennej towarzyszącej i ocenę średnich poprawionych nieobciążonych oddziaływaniem tej zmiennej towarzyszącej.

W opisywanych w pracy przykładach analizowano wpływ jednej zmiennej towarzyszącej na zmienność zmiennej zależnej, dopuszcza się sytuacje w której w doświadczeniu może występować więcej niż jedna zmienna towarzysząca.

Literatura

Bickel, P.J., Hammel, E.A., O'Connell, J.W., 1975. Sex Bias in Graduate Admissions: Data from Berkeley. Science 187, 398–404. DOI: https://doi.org/10.1126/science.187.4175.398

Cochran, W.G., Cox, G.M., 1992. Experimental design. John Wiley & Sons Inc., Hoboken, USA.

Elandt, R., 1964. Statystyka matematyczna w zastosowaniu do doświadczalnictwa rolniczego. PWN, Warszawa.

Kievit, R., Frankenhuis, W., Waldorp, L., Borsboom, D., 2013. Simpson's paradox in psychological science: a practical guide. Front. Psychol. 4.

Linnik, J.W., 1962. Metoda najmniejszych kwadratów i teoria opracowywania obserwacji. PWN, Warszawa.

McConnell, T.E., Shi, S.Q., Chen, H., Wang, G., 2014. Differences Observed in Data Analysis Techniques: An Example Using Natural Fibers' Diameters and Absorption Times. Appl. Eng. Agric. 55–58. DOI: https://doi.org/10.13031/aea.30.10242

Montgomery, D.C., 2005. Design and analysis of experiments. 6th edition. John Wiley & Sons Inc., Hoboken, NJ, USA.

Oktaba, W., 1971. Metody statystyki matematycznej w doświadczalnictwie. PWN.

Paterson, D.D., 1939. Statistical Technique in Agricultural Research: A Simple Exposition of Practice and Procedure in Biometry, McGraw-Hill publications in the agricultural sciences. McGraw-Hill.

Pearson, K., Lee, A., Bramley-Moore, L., 1899. VI. Mathematical contributions to the theory of evolution. —VI. Genetic (reproductive) selection: Inheritance of fertility in man, and of fecundity in thoroughbred racehorses. Philos. Trans. R. Soc. Lond. Ser. Contain. Pap. Math. Phys. Character 192, 257–330. DOI: https://doi.org/10.1098/rsta.1899.0006

Simpson, E.H., 1951. The interpretation of Interaction in Contingency Tables. J. R. Stat. Soc. Ser. B Methodol. 13, 238–241.

Stanisz, A., 2007. Przystępny kurs statystyki z zastosowaniem STATISTICA na przykładach z medycyny. Tom 2. Modele liniowe i nieliniowe., wyd. 3. StatSoft Polska Sp. z o.o., Kraków.

Szklarska, J., Walewski, R., Pielat, H., Radzikowska, A., 1978. Wybrane metody statystyki matematycznej w doświadczalnictwie rolniczym i warzywniczym. Część 1., II. ed PWRiL, Poznań.

TIBCO Software Inc., 2017. Statistica (data analysis software system). Version 13. http://statistica.io.

Timm, N.H., 2002. Applied multivariate analysis. Springer-Verlag Inc., New York, USA.

Wijesuriya, B.W., Thattil, R.O., 1996. Use of Covariates in Improving Precision of Field Experiments in Rubber. Trop. Agric. Res. 20–29.

Wójcik, A.R., Laudański, Z., 1989. Planowanie i wnioskowanie statystyczne w doświadczalnictwie. PWN, Warszawa.

Yang, R.-C., Juskiw, P., 2011. Analysis of covariance in agronomy and crop research. Can. J. Plant Sci. 91, 621–641. DOI: https://doi.org/10.4141/cjps2010-032

Yule, U.G., 1903. Notes on the Theory of Association of Attributes in Statistics. DOI: https://doi.org/10.1093/biomet/2.2.121