

**Wydział Zarządzania**

***Estymacja modeli ANOVA na różnych przykładach***

Autor: *Kacper Prorok*

Kierunek studiów: Informatyka i Ekonometria

Przedmiot: *Statystyczna analiza danych*

Kraków, 2024

Contents

[Wprowadzenie 3](#_Toc183624511)

[ANOVA wieloczynnikowa 3](#_Toc183624512)

[Założenia 4](#_Toc183624513)

[1. Zmienna zależna ma wartości na skali przedziałowej 4](#_Toc183624514)

[2. Założenia o randomizacji (2,3,4) 5](#_Toc183624515)

[3. Założenie o normalności danych w każdej próbie 5](#_Toc183624516)

[**4.** Założenie o równości wariancji w grupach 5](#_Toc183624517)

[Przeprowadzenie ANOVY wieloczynnikowej 6](#_Toc183624518)

[Analiza post hoc 7](#_Toc183624519)

[Podział na dni(Day) 7](#_Toc183624520)

[Podział na Section 7](#_Toc183624521)

[Dla interakcji Day i Section 8](#_Toc183624522)

[ANOVA z powtarzanymi pomiarami 9](#_Toc183624523)

[Założenia 10](#_Toc183624524)

[1. Zmienna zależna ma wartości na skali przedziałowej 10](#_Toc183624525)

[2. Założenia o randomizacji (2,3,4) 10](#_Toc183624526)

[3. Założenie o normalności danych w każdej próbie 10](#_Toc183624527)

[4. Założenie o równości wariancji w grupach 10](#_Toc183624528)

[5. Założenie o sferyczności 11](#_Toc183624529)

[Przeprowadzenie testu ANOVA z powtarzanymi pomiarami 11](#_Toc183624530)

[ANCOVA 12](#_Toc183624531)

[Założenia 13](#_Toc183624532)

[1. Zmienna zależna ma wartości na skali przedziałowej 13](#_Toc183624533)

[2. Założenia o randomizacji (2,3,4) 13](#_Toc183624534)

[3. Założenie o normalności danych w każdej próbie 13](#_Toc183624535)

[**4.** Założenie o równości wariancji w grupach 13](#_Toc183624536)

[1. Liniowość 14](#_Toc183624537)

[2. Rozkład normalny reszt 14](#_Toc183624538)

[3. Brak autokorelacji reszt 14](#_Toc183624539)

[4. Homoskedastyczność 15](#_Toc183624540)

[1. Zmienne niezależne nieskorelowane 15](#_Toc183624541)

[2. Jednorodność lub równoległość regresji w obrębie grup 15](#_Toc183624542)

[Przeprowadzenie testu ANCOVA 16](#_Toc183624543)

[Test post hoc 16](#_Toc183624544)

# Wprowadzenie

ANOVA (Analysis of Variance) to metoda statystyczna, której celem jest analiza różnic pomiędzy średnimi w kilku grupach. Umożliwia ona sprawdzenie, czy co najmniej jedna grupa różni się istotnie od pozostałych, analizując zmienność wyników w obrębie i pomiędzy grupami. Jeśli zmienność pomiędzy grupami jest większa niż wewnątrz grup, sugeruje to, że różnice między średnimi grup są statystycznie istotne.

Metoda ta jest powszechnie stosowana w badaniach eksperymentalnych, gdzie badamy wpływ jednego lub więcej czynników na zmienną zależną. Przykładem może być analiza wpływu różnych rodzajów nawozów na wzrost roślin, gdzie celem jest porównanie średnich wzrostu w grupach nawozowych.

# ANOVA wieloczynnikowa

Test ANOVA z wieloma czynnikami opiera się na tej samej zasadzie co ANOVA jednoczynnikowa. Różnicą jest to, że całkowitą sumę kwadratów SStotal dzieli się na sumy odchyleń wynikające z wpływu działania: czynnika A (SSA) oraz B (SSB), interakcji A i B (SSAB) oraz losowości zjawiska SSE

SStotal=SSA+SSB+SSE

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Day** | **Section** | **Response** |
| **0** | Monday | News | 11 |
| **1** | Monday | News | 8 |
| **2** | Monday | News | 6 |
| **3** | Monday | News | 8 |
| **4** | Tuesday | News | 9 |

Poniżej przedstawiam kilka pierwszych wierszy zestawu danych, który zapożyczyłem z naszego kursu na UPELu:

Tabela : Dane ogłoszenia.csv

Zmienne **Day** (poniedziałek-piątek) oraz **Section** (News, Business, Sports) stanowią zmienne kategoryczne, natomiast zmienna **Response** reprezentuje wynik przypisany do danej kombinacji kategorii.

W ramach analizy **ANOVA wieloczynnikowej** badamy jednoczesny wpływ kilku czynników na zmienną zależną. W tym przypadku zmienna **Response** jest zmienną zależną, natomiast zmienne **Day** i **Section** pełnią rolę czynników, czyli zmiennych niezależnych.

A diagram of a box plot

Description automatically generatedA graph with blue and green squares

Description automatically generated

Rysunek Podział na Section oraz Day

**A graph of different colored boxes

Description automatically generated**

Rysunek Podział na Section i Day

Już na pierwszy rzut oka możemy zobaczyć że istnieją spore różnice między niektórymi podgrupami, ale aby stwierdzić że są one statystycznie istotne, muszę przeprowadzić test ANOVA.

## Założenia

1. **Zmienna zależna ma wartości na skali przedziałowej**
2. **Próbka została wybrana z populacji w sposób losowy**
3. **Elementy próby zostały przypisane do danych podgrup losowo (II zasada randomizacji)**
4. **Wszystkie pomiary są niezależne**
5. Dane w każdej próbie mają rozkład normalny,
6. Wariancje w podgrupach są równe.

### Zmienna zależna ma wartości na skali przedziałowej

Zmienna Response ma wartości na skali przedziałowej, gdyż jesteśmy w stanie określić różnice między dwoma wartościami i możemy ją interpretować (np. różnica wynosząca 10 będzie spora)

### Założenia o randomizacji (2,3,4)

Zakładam, że autor danych poprawnie je przygotował i dane spełniają założenia o randomizacji oraz że pomiary są od siebie niezależne.

### Założenie o normalności danych w każdej próbie

Dla każdej podgrupy przeprowadziłem test normalności Shapiro-Wilka w celu oceny rozkładu danych. Poniżej przedstawiam wyniki testów:

  
**** **Wszystkie rozważane podgrupy spełniają założenie o zgodności z rozkładem normalnym.**

Rysunek Wyniki

Rysunek Wyniki przy podziale na Section

Rysunek Wyniki przy podziale na Section i Day

### Założenie o równości wariancji w grupach

Założenie o równości wariancji w grupach zostanie zweryfikowane za pomocą testu Bartletta, który przeprowadzę zarówno dla interakcji czynników oraz dla każdego osobno. W teście Barletta hipotezy przyjmują następującą postać:

***H0: σ21 = σ22 =... = σ2k  
H1: nie wszystkie σ2i są jednakowe***

Tabela Wyniki testu Bartletta

**Dla każdego z czynników wartość p wynosi powyżej 0,05, co oznacza, że nie ma podstaw do odrzucenia hipotezy zerowej. W związku z tym przyjmujemy, że wariancje w każdej kombinacji są równe.**

## Przeprowadzenie ANOVY wieloczynnikowej

A graph of a business

Description automatically generated with medium confidencePrzed przystąpieniem do przeprowadzenia testu ANOVA należy zweryfikować, czy występuje efekt interakcji. Poniżej przedstawiam wykres średnich, który ilustruje ten efekt:

Rysunek Wykres interakcji

Linie na wykresie przecinają się, co wskazuje na obecność efektu interakcji. Oznacza to, że między czynnikami zachodzą interakcje. W związku z tym, przeprowadzam test ANOVA uwzględniający interakcje. Hipotezy przyjmują następującą formę:

***H0: Średnie we wszystkich grupach są równe, tj. μ₁ = μ₂ = … = μₖ.***

***H1: Przynajmniej jedna średnia różni się od pozostałych.***

Przejdźmy teraz do wyniku testu ANOVA**:**

Rysunek Wyniki testu ANOVA wieloczynnikowej

Tabela Wyniki testu ANOVA wieloczynnikowej

Wyniki są następujące:

* Po podzieleniu danych na grupy według dni (Day) zaobserwowano istotne różnice w średnich
* Po podzieleniu danych na grupy według sekcji (Section) zaobserwowano istotne różnice w średnich
* Gdy uwzględnimy interakcje pomiędzy Section i Day to również obserwujemy istotne różnice w średnich

## Analiza post hoc

Po przeprowadzeniu testu ANOVA, wiemy że różnice między średnimi występują przy podziale danych na czynnik Day, Section oraz interakcji Section:Day. Aby dowiedzieć się, które podgrupy dokładnie spowodowały te różnice musimy przeprowadzić analizę post hoc. Posłużę się do tego testem Tukeya.

### Podział na dni(Day)



Tabela Wyniki testu Tukeya (Dni)

W przypadku podziału danych według dni, różnice w średnich występują pomiędzy grupami: **Friday-Monday, Friday-Thursday oraz Friday-Wednesday**. Piątek szczególnie wyróżnia się na tle pozostałych grup.

### Podział na Section

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **group1** | **group2** | **meandiff** | **p-adj** | **lower** | **upper** | **reject** |
| Business | News | -0,2 | 0,9659 | -2,1167 | 1,7167 | FALSE |
| Business | Sports | -2,1 | 0,0286 | -4,0167 | -0,1833 | TRUE |
| News | Sports | -1,9 | 0,0525 | -3,8167 | 0,0167 | FALSE |

Tabela Wyniki testu Tukeya (Section)

W przypadku podziału danych według sekcji, różnice w średnich występują między grupami **Business i Sports**, a w podgrupie **News-Sports** obserwuje się wynik bliski granicy odrzucenia hipotezy zerowej.

### Dla interakcji Day i Section



Tabela Wyniki testu Tukeya (Section+Dni)

Z uwagi na fakt że powstało bardzo dużo kombinacji grup powyżej ukazuje tylko te, w którym test post hoc wykazał że występują istotne różnice.

# ANOVA z powtarzanymi pomiarami

Procedura ANOVA dla powtarzanych pomiarów analizuje grupy powiązanych zmiennych zależnych, które reprezentują różne pomiary tego samego atrybutu. Szczególnie test ten przydaje się na przykład przy badaniu terapii w czasie lub przy pomiarach wzrostu roślin. Kluczowym czynnikiem w badaniu jest czas, bo tak naprawdę to istotność tego czynnika chcemy zmierzyć (np. czy terapia faktycznie poprawiła wyniki pacjentów).

Do przeprowadzenia badania posłużę się przykładowym zestawem danych, który przedstawia wyniki testu terapeutycznego, do którego przystąpiło pięciu pacjentów. Wyniki są podzielone w czasie – testy były przeprowadzane co miesiąc, co przedstawiają odpowiednie kolumny.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Czas** | **1** | **2** | **3** | **4** |
| **Pacjent** |  |  |  |  |
| **1** | 36 | 38 | 30 | 29 |
| **2** | 34 | 38 | 30 | 29 |
| **3** | 34 | 28 | 38 | 32 |
| **4** | 38 | 34 | 20 | 44 |
| **5** | 26 | 28 | 34 | 50 |

A graph with a line going up

Description automatically generatedTabela Dane o wynikach testu terapeutycznego

Rysunek Wykres średnich

Po wykresie średnich możemy od razu zauważyć, że różnice nie wydają się być wielkie. Będziemy mogli to stwierdzić dopiero po przeprowadzeniu testu.

## Założenia

Założenia są podobne jak w przypadku testu ANOVA jedno-czynnikowej z jednym dodatkowym - *kowariancja pomiarów u tej samej osoby musi być jednorodna (tj. kowariancje są w przybliżeniu równe)*. Nie bierzemy pod uwagę również założenia o niezależności danych, ponieważ dane ze względu na swoją budowę (np. pobierane od jednego pacjenta) są od siebie zależne,

### Zmienna zależna ma wartości na skali przedziałowej

Wynik testu ma wartości na skali przedziałowej, gdyż jesteśmy w stanie określić różnice między dwoma wartościami i możemy ją interpretować (np. różnica wynosząca 30 będzie spora)

### Założenia o randomizacji (2,3,4)

Zakładam, że autor danych poprawnie je przygotował i dane spełniają założenia o randomizacji.

### Założenie o normalności danych w każdej próbie

Należy zbadać zgodność rozkładu każdej z grup (czyli okresów czasowych) z rozkładem normalnym. Poniżej przedstawiam wyniki testu Shapiro-Wilka:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Statistic** | **p-value** |
| Time 1 | 0,859908 | 0,22792 |
| Time 2 | 0,81432 | 0,105475 |
| Time 3 | 0,932799 | 0,615581 |
| Time 4 | 0,833279 | 0,147184 |

Tabela Wyniki testu Shapiro-Wilka

P-value w żadnej grupie nie wyniosło poniżej 0.5, co oznacza że **założenie o zgodności z rozkładem normalnym jest spełnione w każdej grupie.**

### Założenie o równości wariancji w grupach

Levene's test sprawdza, czy wariancje w różnych grupach są jednorodne (homogeniczne). Hipotezy wyglądają następująco:

***H0: Wariancje w różnych grupach są równe (homogeniczne).***

***H1: Wariancje w różnych grupach są różne (brak homogeniczności).***

**P-value wyniosło 0.54**, więc stwierdzam że istnieje brak podstaw do odrzucenia hipotezy zerowej. Oznacza to, że **wariancje w grupach są równe, czyli założenie jest spełnione.**

### Założenie o sferyczności

Założenia o jednorodności kowariancji może być zamienione na nieco mniej wymagające, tzn. równość wariancji dla różnic na każdym poziomie, czyli założenie o sferyczności. Do weryfikacji tego założenia przeprowadzę test Mauchley’a. Hipotezy wyglądają następująco:

***H0: Różnice między parami pomiarów mają równą wariancję, a kowariancje są równe zero***

***H1: Różnice między parami pomiarów mają różne wariancje, a kowariancje są różne od zera***

W przypadku naszego zestawu danych **p-value wyniosło 0.659,** co oznacza że istnieje brak podstaw do odrzucenia hipotezy zerowej – **macierz kowariancji spełnia założenie o sferyczności.**

## Przeprowadzenie testu ANOVA z powtarzanymi pomiarami

W analizie ANOVA z powtarzanymi pomiarami biorą udział następujące hipotezy zerowe i alternatywne:

***H0: µczas1 = µczas2 = µczas3 = µczas4***

***H1: przynajmniej jedna średnia populacji różni się od reszty***

Do przeprowadzenia tego testu używam funkcji *AnovaRM(depvar='Score', subject='Patient', within=['Time'])* z pakietu statsmodels w Pythonie. Wyniki są następujące:

|  |  |  |  |
| --- | --- | --- | --- |
| **F Value** | **Num DF** | **Den DF** | **Pr > F** |
| 0,567884 | 3 | 12 | 0,646647 |

Tabela Wyniki testu ANOVA z powtarzanymi pomiarami

P-value wynosi więcej niż 0.05, więc nie ma podstaw do odrzucenia H0. Oznacza to, że **nie występują istotne różnice między okresami.** Możemy powiedzieć, że działanie terapii na przestrzeni 4 miesięcy nie przyniosło istotnego wpływu na wyniki badań u pacjentów.

# ANCOVA

Analiza kowariancji (ANCOVA) to metoda statystyczna łącząca cechy analizy wariancji (ANOVA) i regresji liniowej. ANCOVA pozwala na badanie wpływu jednego lub więcej czynników (zmiennych niezależnych kategorialnych) na zmienną zależną, jednocześnie kontrolując wpływ zmiennych współzmiennych (covariates), które są ilościowe i mogą wprowadzać dodatkową wariancję.

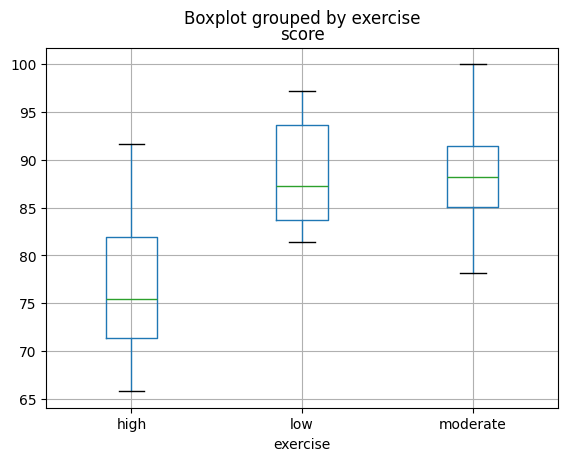
Główne cele ANCOVY to:

1. Usunięcie wpływu współzmiennych na zmienną zależną, aby uzyskać bardziej precyzyjne oszacowanie efektów głównych i interakcji.
2. Sprawdzenie, czy różnice w zmiennej zależnej między grupami są istotne, po uwzględnieniu współzmiennych.

Analizę kowariancji przeprowadzę na danych *stress* z pakietu datarium. Przedstawiają one badanie mające na celu ustalić wpływ nowej terapii oraz ćwiczeń na wynik stresu. Dodatkowo mamy informację o wieku każdego pacjenta. **Celem mojego badania jest sprawdzenie wpływu ilości ćwiczeń na poziom stresu (wynik testu).**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **id** | **score** | **treatment** | **exercise** | **age** |
| **0** | 1 | 95,6 | yes | low | 59 |
| **1** | 2 | 82,2 | yes | low | 65 |
| **2** | 3 | 97,2 | yes | low | 70 |
| **3** | 4 | 96,4 | yes | low | 66 |
| **4** | 5 | 81,4 | yes | low | 61 |

Tabela Zestaw danych stress

**A graph with blue and green squares

Description automatically generated**

Rysunek Rozkład poziomu stresu dla każdej grupy ćwiczeniowej

Rysunek Rozkład wieku dla każdej grupy ćwiczeniowej

Po analizie wykresów, możemy zauważyć że w grupie która ćwiczyła najwięcej poziom stresu był znacząco niższy. Jednocześnie możemy zauważyć, że tej grupie średnia wieku jest najniższa. Może to sugerować, że wiek również istotnie wypływa na poziom stresu – osoby młodsze mniej się stresują.

W celu wyeliminowania wpływu wieku na nasze badanie przeprowadzę test ANCOVA, w którym zmienną zależną będzie ‘score’, czynnikiem będzie zmienna ‘exercise’, a zmienną współzależną (covariant) będzie ‘age’.

## Założenia

Dane muszą spełniać zarówno założenia ANOVY jak i regresji liniowej. Zacznę od sprawdzenia założeń ANOVY:

### Zmienna zależna ma wartości na skali przedziałowej

Zmienna Score ma wartości na skali przedziałowej, gdyż jesteśmy w stanie określić różnice między dwoma wartościami i możemy ją interpretować (np. różnica wynosząca 50 będzie spora)

### Założenia o randomizacji (2,3,4)

Zakładam, że autor danych poprawnie je przygotował i dane spełniają założenia o randomizacji oraz że pomiary są od siebie niezależne.

### Założenie o normalności danych w każdej próbie

Dla każdej podgrupy przeprowadziłem test normalności Shapiro-Wilka w celu oceny rozkładu danych. Poniżej przedstawiam wyniki testów:

|  |  |
| --- | --- |
| **Exercise** | **p-value** |
| Low | 0.07 |
| Moderate | 0.9 |
| High | 0.445 |

Tabela Wyniki testu Shapiro-Wilka

**Wszystkie rozważane podgrupy spełniają założenie o zgodności z rozkładem normalnym.**

### Założenie o równości wariancji w grupach

Założenie o równości wariancji w grupach zostanie zweryfikowane za pomocą testu Barletta. W teście Barletta hipotezy przyjmują następującą postać:

***H0: σ21 = σ22 =... = σ2k  
H1: nie wszystkie σ2i są jednakowe***

**P-value wyniosło 0.397, co oznacza że istnieje brak podstaw do odrzucenia H0, czyli dane spełniają założenie o równości wariancji w grupach**

**Teraz przejdziemy do badania założeń związanych z regresją liniową:**

### Liniowość

A graph of different colored lines

Description automatically generatedWidzimy, że istnieje zależność liniowa między zmienną zależną (score) i zmienną współzaeżną (age) w każdej z podgrup zmiennej niezależnej (exercise).

Rysunek Wykres zależności liniowych

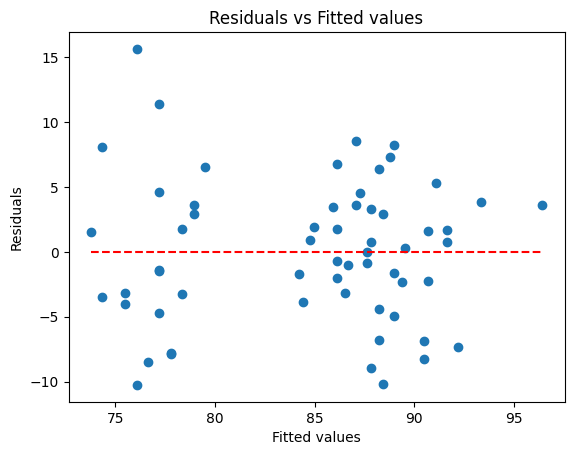
### Rozkład normalny reszt

Stworzyłem model regresji liniowej ('score ~ age + exercise'), aby zbadać czy reszty mają rozkład normalny. P-value w teście Shapiro-Wilka **wyniosło 0.676**, co oznacza że istnieje brak podstaw do odrzucenia H0, czyli **reszty modelu są zgodne z rozkładem normalnym.**

### Brak autokorelacji reszt

Wymagamy, by reszty modelu regresji liniowej były nieskorelowane. W tym celu przeprowadzę test Durbina-Watsona, który sprawdza czy występuje autokorelacja reszt w modelu. Przyjmuje się, że wartość statystyki DW w przedziale od 1.5 do 2.5 oznacza, że autokorelacja nie występuje. W naszym przypadku **statystyka DW wyniosła 1.60, co oznacza że nie występuje autokorelacja reszt.**

### Homoskedastyczność



Rysunek Wykres reszt względem dopasowanych wartości

Aby sprawdzić, czy reszty modelu mają stałą wariancję przeprowadzam dodatkowo test Breusch-Pagana o następujących hipotezach:

***H0: Wariancja reszt jest stała (homoskedastyczność).***

***H1: Wariancja reszt nie jest stała (heteroskedastyczność).***

Wynik p-value : 0.2558. **Oznacza to, że wariancja reszt jest stała.**

**Dodatkowo ANCOVA ma dwa własne założenia:**

### Zmienne niezależne nieskorelowane

Jako że w naszym badaniu występuje tylko jedna zmienna niezależna (exercise), nie istnieje potrzeba sprawdzenia czy zmienne niezależne są skorelowane ze sobą.

### Jednorodność lub równoległość regresji w obrębie grup

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **sum\_sq** | **df** | **F** | **PR(>F)** |
| **exercise** | 980,9121 | 2 | 14,71466 | 7,93E-06 |
| **age** | 298,4212 | 1 | 8,953233 | 0,004169 |
| **age:exercise** | 13,69964 | 2 | 0,205508 | 0,814867 |
| **Residual** | 1799,88 | 54 |  |  |

Nachylenie linii regresji(dla zmiennych age i score) powinno być podobne dla każdej z badanych grup(exercise). Innymi słowy badamy, czy występują istotne interakcje między zmienną niezależną(exercise) oraz zmienną towarzyszącą(age). Przeprowadzę ANOVĘ z interakcjami ('score ~ age \* exercise') i otrzymuje następujące wyniki:

Tabela Wyniki testu ANOVA

Model ANOVA z interakcjami był nieistotny statystyczne (p-value 0.814), co oznacza że **mamy do czynienia z modelem jednakowych odchyleń.**

## Przeprowadzenie testu ANCOVA

Po sprawdzeniu założeń mogę przystąpić do przeprowadzenia testu ANCOVA. Hipotezy tego testu wyglądają następująco:

***H0: Grupy nie różnią się istotnie pod względem średniej zmiennej zależnej po uwzględnieniu kowariatu.***

***H1: Przynajmniej jedna grupa różni się istotnie od pozostałych pod względem średniej zmiennej zależnej po uwzględnieniu kowariatu***

W tym celu użyję funkcji *ancova(dv='score', covar='age', between='exercise')* z pakietu pingouin w Pythonie. Wyniki są następujące:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **SS** | **DF** | **F** | **p-unc** | **np2** |
| exercise | 980,9121 | 2 | 15,14438 | 0.000006 | 0,351016 |
| age | 298,4212 | 1 | 9,214697 | 0,003639 | 0,141298 |
| Residual | 1813,58 | 56 |  |  |  |

Tabela Wyniki testu ANCOVA

Wnioski:

* Po uwzględnieniu wpływu zmiennej towarzyszącej ‘age’ **wystąpiły istotne różnice między grupami zmiennej ‘exercise’.**

## Test post hoc

Aby dowiedzieć się, między którymi grupami wystąpiły istotne różnice przeprowadziłem test post hoc Tukeya:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **group1** | **group2** | **meandiff** | **p-adj** | **lower** | **upper** | **reject** |
| high | low | 11,835 | 0 | 7,2029 | 16,4671 | TRUE |
| high | moderate | 11,225 | 0 | 6,5929 | 15,8571 | TRUE |
| low | moderate | -0,61 | 0,9462 | -5,2421 | 4,0221 | FALSE |

Tabela Wyniki testu Tukeya

Istotne różnice wystąpiły między grupami low-high oraz moderate-high. **Możemy więc stwierdzić, że osoby, które najwięcej ćwiczyły miały znacząco mniejszy poziom stresu (znacząco niższy wynik testu), po uwzględnieniu wpływu wieku.**