Analiza regresji wielokrotnej - hierarchiczna

Poniżej prezentujemy przykładowe pytania z rozwiązaniami dotyczącymi analizy regresji wielokrotnej wykonanej metodą hierarchiczną. Wszystkie rozwiązania są poprawne, ale pamiętaj, że na całym świecie wykładowcy różnią się pod względem pewnych niuansów przekazywania wiedzy z zakresu statystyki. Niektórzy np. istotność statystyczną zalecają zapisywać tylko jako $p < 0,05$ lub $p > 0,05$. Inni jako $p < 0,05$ lub ni., gdy wynik jest nieistotny statystycznie. Inni z kolei pozwalają na zapisywanie konkretnej wartości istotności jak np. $p = 0,034$ lub $p = 0,255$. Na chwilę obecną każdy z tych sposobów jest poprawny, choć naszym zdaniem tylko jeden z nich jest najlepszy. W miejscach najbardziej problemowych rozwiązanie zadania opatrzyliśmy czerwonym komentarzem.

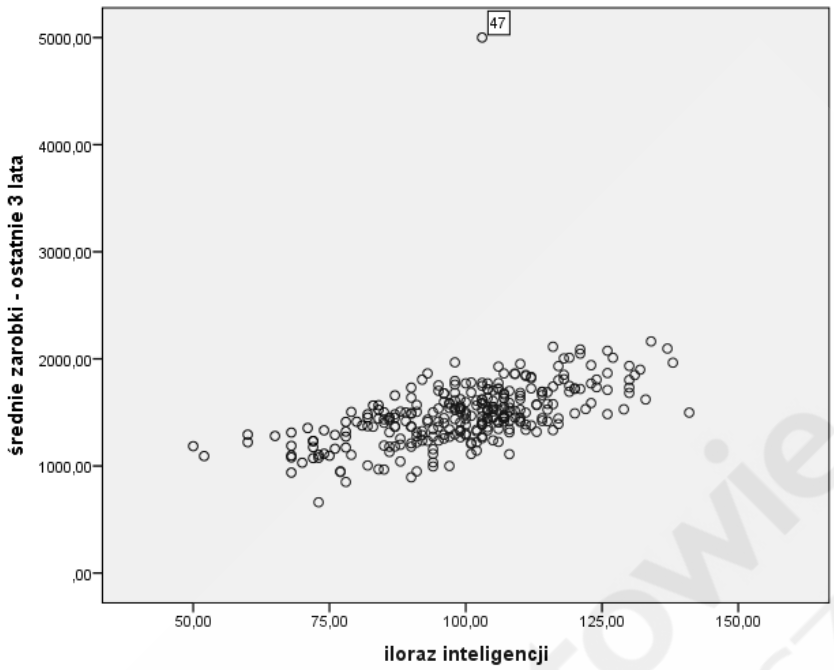
UWAGA: Podsumowanie zbudowanych modeli jest dosyć rozwlekłe i skonstruowane tak, żeby jak najlepiej zdać kolokwium. W taki sposób nie opisuje się wyników do prac dyplomowych i artykułów naukowych. Na kolokwium musisz „popisać się" wiedzą na temat niemal każdego współczynnika wygenerowanego przez SPSS.

➔ **Otwórz bazę REG_H_1a i wykonaj kolejne zadania**

Model 1 *Przetestuj następujące pytanie: czy można przewidywać średnie zarobki (średnia z ostatnich 3 lat) na podstawie znajomości: ilorazu inteligencji (iq) i lat spędzonych w szkole (educ)?*

Zad. 1. Wykonaj wykres rozrzutu zarobków i ilorazu inteligencji. Czy widać tu ewidentne przypadki nietypowe (outlierów)? Jeśli tak, podaj numery przypadków:

Odp.: Tak, na podstawie wykresu rozrzutu dla dwóch wyżej wymienionych zmiennych można wywnioskować, że w danych występuje obserwacja odstająca (tzw. outlier). Jest jedna osoba, która zarabia znacznie więcej niż inni. Jest to badany nr 1 (zmienna nr) występujący w bazie danych na 47 pozycji.



Zad. 2.Wykonaj analizę regresji wielokrotnej, która odpowie na zadane wcześniej pytanie. Czy „diagnostyka obserwacji" wykazała istnienie przypadków nietypowych (outlierów), niezgodnych z przewidywaniami modelu?
Jeśli występuje outlier odfiltruj go i wykonaj kolejne zadania bez obserwacji odstającej

Odp.: Owszem, diagnostyka obserwacji potwierdza wnioski płynące z wykonanego wcześniej wykresu rozrzutu. Osoba będąca w bazie danych w rzędzie nr 47 jest outlierem, którego należy odfiltrować przy użyciu funkcji: $CASENUM ~= 47. Wynik standaryzowanej reszty dla tej osoby nie mieści się w przedziale od -3 do 3.

Zad. 3. Zapisz równanie regresji dla pierwszego modelu (pamiętaj o odpowiednich znakach przed każdym współczynnikiem oraz o nazwach zmiennych /iq; educ/):

Odp.: ZAROBKI' = iq · 6,65 + educ · 58,24 + 115,43

Uwaga Pogotowia Statystycznego nr 1

Niektórzy wykładowcy uważają, że nie należy włączać do modelu zmiennej, której współczynnik nieistotnie statystycznie różni się od zera (istotność testu t Studenta powyżej 0,05). W zadaniu 3 stała w modelu uzyskała istotności p = 0,091 a tym samym niektórzy mogą uznawać takie rozwiązanie za poprawne: ***Odp.: Zarobki' = iq*6,65 + educ * 58,24***

W rozwiązaniu tym na końcu nie ma dodanej stałej ponieważ teoretycznie równa się 0, a nie ma po co dodawać zera. ***Zapytaj o to swojego prowadzącego przed kolokwium!***

Zad. 4. Który z predyktorów ma największy, a który najmniejszy wpływ (ISTOTNY STATYSTYCZNIE) na wartości zmiennej zależnej. Predyktory uszereguj wg wielkości współczynników Beta:

Nazwa predyktora	Korelacja w modelu (zmiennej zależnej i predyktora)	Korelacja poza modelem (*r* Pearsona)
educ	0,47	0,68
iq	0,41	0,66

Uwaga Pogotowia Statystycznego nr 2

Niektórzy wykładowcy uczą, że współczynnik Beta mówi o korelacji predyktora ze zmienną zależną w modelu (np. Kuba Niewiarowski – SWPS Warszawa). Nie jest to błąd, a raczej trochę mylący skrót myślowy (zapewne mający na celu dobro studentów). Współczynnik standaryzowany beta jest miarą mówiącą o stopniu nachylenia linii regresji w stosunku do osi OX. Przyjęło się, że raczej o współczynnikach korelacji mówimy wtedy, gdy na myśli mamy miarę, której wartość może przyjmować wartość w przedziale od -1 do 1. Współczynnik Beta w analizie regresji z wieloma predyktorami może być większy od 1 lub mniejszy od -1. Wiemy jednak, że zadanie zostaje pozytywnie ocenione, gdy jako miarę korelacji predyktora ze zmienną zależną w modelu podaje się wartość beta. Pamiętaj jednak o istnieniu korelacji cząstkowych i semi-cząstkowych. Po prostu pojęcie „korelacja w modelu" jest mało precyzyjne więc dopytaj prowadzącego co ma na myśli.

Zad. 5. Wymień predyktory (jeśli takie są), które nie wnoszą istotnych informacji do modelu (moglibyśmy je pominąć).

Takie predyktory nie istnieją. Każdy z predyktorów w modelu jest istotny statystycznie co najmniej na poziomie $p < 0,05$

Zad. 6. Jaki procent zmienności ZAROBKÓW przewiduje ten model (zapisz wartość odpowiedniego współczynnika, oraz wynik testu istotności dla modelu)

Odp.: $R^2 = 0,583$ co oznacza, że wariancja w zakresie iq i educ wyjaśnia łącznie ponad 58% wariancji (zmienności) w zakresie zarobków.

$F(2, 294) = 207,77$; $p < 0,001$

Uwaga Pogotowia Statystycznego nr 3

W raportowaniu wyników analiz statystycznych w standardzie APA wszystkie miary ***prócz istotności statystycznej*** *podajemy do 2 miejsc po przecinku. Zauważ, że R^2 podałem do 3 miejsc. Niektórzy wykładowcy pozwalają na taki „manewr" po to by z większą dokładnością podać procent wyjaśnionej wariancji. Pamiętaj jednak żeby raczej współczynnik determinacji R^2 podawać do 2 miejsc po przecinku, ale mając na uwadze kolejne cyfry (3 i 4 po przecinku) po to by z większą precyzją przekazać wartość po zamianie na procenty.*

Zad. 7. O ile wzrosną zarobki gdy iloraz inteligencji wzrośnie o 1?

Wzrosną o 6zł i 65gr przy błędzie standardowym równym SE = 0,73

Zad. 8. Podsumuj krótko wyniki modelu 1 (jaki był cel analizy, jakie zmienne użyte, jakie było ostateczne równanie, jak kształtuje się układ predyktorów, jaki był procent wariancji wyjaśnionej).

W celu sprawdzenia możliwości przewidywania zarobków badanych osób na podstawie informacji o ich ilorazie inteligencji i latach nauki szkolnej przeprowadzono analizę regresji liniowej. Do modelu wprowadzono dwa predyktory – IQ i EDUC. Zmienną zależną (objaśnianą) w modelu były średnie zarobki z ostatnich 3 lat. Model ten jest dobrze dopasowany do danych i pozwala na lepsze przewidywanie zarobków badanych osób niż jedynie na podstawie średniej - $F(2, 294) = 207,77$; $p < 0,001$. Równanie regresji dla tego modelu wygląda następująco: Zarobki' = iq · 6,65 + educ · 58,24 + 115,43. Zbudowany model wyjaśnia 58,3% wariancji w zakresie zmiennej objaśnianej. Oznacza to tym samym, że 42% zmienności wyjaśnione jest przez inne, nieuwzględnione w badaniu czynniki. Wartość standardowego błędu oszacowania wynosząca w niniejszym modelu SEE = 165,41 wskazuje na to, że w przewidywaniach zarobków możemy pomylić się o 165 zł i 41 gr. Należy również zaznaczyć, że lata spędzone w szkole są silniejszym predyktorem ($\beta = 0,47$; $p < 0,001$) aniżeli iloraz inteligencji ($\beta = 0,41$; $p < 0,001$).

__Model 2__ Dodatkowo badacz był zainteresowany przewidywaniem zarobków przy kontroli hobby osób badanych. Czy zarobki (poza tym, co zostało sprawdzone w modelu 1) mogą być przewidywane dodatkowo, na podstawie aktywności wykonywanych w wolnym czasie?

Rodzaj hobby jest reprezentowany w zbiorze danych przez 3 (zero-jedynkowe) zmienne: stat – studiowanie wolnym czasie statystyki, sport – uprawianie sportów w wolnym czasie, arts – uprawianie lub podziwianie sztuki w wolnym czasie

Zad. 9. Wprowadź te zmienne jako nowy blok dla poprzedniej analizy. Czy możemy zaobserwować statystycznie istotną zmianę w poziomie wariancji wyjaśnionej po dodaniu tych trzech zmiennych (zapisz wielkość zmiany, oraz odpowiednie wartości testu F)?

Odp.: Tak, możemy dostrzec istotną statystycznie zmianę. Zmiana R^2 wynosi 2,3% i jest to zmiana istotna statystycznie - $F(3, 291) = 5{,}74$; $p = 0{,}001$

Zad. 10. Zapisz równanie regresji dla drugiego modelu (pamiętaj o odpowiednich znakach przed każdym współczynnikiem oraz o nazwach zmiennych /iq; educ; stat; arts; sport/):

ZAROBKI' = iq · 6,88 + educ · 47,93 + stat · 120,1 + sports · 51,54 + arts · 39,18 + 163,94

Pamiętaj o uwadze Pogotowia Statystycznego nr 1.

Zad. 11. Który z predyktorów ma największy, a który najmniejszy wpływ (ISTOTNY STATYSTYCZNIE) na wartości zmiennej zależnej? Predyktory uszereguj wg wielkości współczynników Beta:

Nazwa predyktora	Korelacja w modelu (zmiennej zależnej i predyktora)	Korelacja poza modelem (*r* Pearsona)
iq	0,42	0,65
educ	0,38	0,68
stat	0,20	0,36
sports	0,09	0,01
arts	0,07	-0,05

Uwaga Pogotowia Statystycznego nr 4

W celach szkoleniowych wprowadziliśmy do powyższej tabeli wszystkie predyktory w modelu. Zwróć jednak uwagę, że prowadzący prosi o wprowadzenie tylko tych ***istotnych statystycznie.*** *Możesz dopytać o to na kolokwium żeby doprecyzować ową kwestię. Prawdopodobnie jednak w tabeli powinny zostać zawarte tylko zmienne: iq, educ, stat.*

Zad. 12. Wymień predyktory (jeśli takie są), które nie wnoszą istotnych informacji do modelu (moglibyśmy je pominąć).

Odp.: Istnieją w modelu dwa predyktory, które nie wnoszą istotnych informacji. Usunięciu można poddać zmienną sports oraz arts. Współczynniki dla tych dwóch predyktorów są nieistotne statystycznie na poziomie $p > 0,05$.

Zad. 13. Jaki procent zmienności ZAROBKÓW przewiduje ten model (zapisz wartość odpowiedniego współczynnika, oraz wynik testu istotności dla modelu)

Odp.: $R^2 = 0,602$ co oznacza, że wszystkie predyktory wprowadzone do modelu wyjaśniają łącznie trochę ponad 60% wariancji (zmienności) w zakresie zarobków.

$F(5, 291) = 90,57$; $p < 0,001$

Zad. 14. Podsumuj krótko wyniki modelu 2 (jaki był cel tej analizy – porównaj obydwa modele, co było dodane i jaki był tego efekt)

W kolejnym kroku przeprowadzonej, hierarchicznej analizy regresji wprowadzono dodatkowo do modelu 3 predyktory w postaci zmiennych stat, sports i arts (zamiast nazw można napisać co oznaczają a w nawiasie dopisać nazwy). Okazuje się, że po ich dodaniu dostrzec można istotny statystycznie wzrost procentu wyjaśnionej wariancji o 2,3% - $F_{zmiany}(3, 291) = 5,74$; $p = 0,001$. Łącznie cały zbudowany model wyjaśnia 60,2% zmienności w zakresie średnich zarobków myląc się o SEE = 161,55 zł. Drugi model uznać można tym samym za bardziej precyzyjny ponieważ po pierwsze wyjaśnia większy procent wariancji w zakresie zmiennej objaśnianej, a po drugie, myli się on w mniejszym stopniu (niższa wartość standardowego błędu oszacowania SEE). Jest on dobrze dopasowany do danych i pozwalający na lepsze przewidywanie zarobków niż jedynie na podstawie średniej $F(5, 291) = 90,57$; $p < 0,001$. Najsilniejszym predyktorem w modelu drugim jest iloraz inteligencji ($\beta = 0,42$; $p < 0,001$), słabszym lata spędzone w szkole ($\beta = 0,39$; $p < 0,001$), a najsłabszym i istotnym statystycznie predyktorem jest fakt studiowania statystyki w czasie wolnym ($\beta = 0,20$; $p <$

0,001). Uprawianie sportu (β = 0,09; p = 0,063) oraz uprawianie lub podziwianie sztuki w czasie wolnym (β = 0,07; p = 0,153) to predyktory nieistotne statystycznie. Na tym etapie równanie regresji wygląda następująco:

ZAROBKI' = iq · 6,88 + educ · 47,93 + stat · 120,1 + sports · 51,54 + arts · 39,18 + 163,94

Model 3 *Dodatkowo badacz był zainteresowany przewidywaniem zarobków przy kontroli płci osób badanych. Czy zarobki (poza tym, co zostało sprawdzone w modelu 2) mogą być przewidywane dodatkowo, na podstawie płci (plec)? Wprowadź tą zmienną jako nowy blok dla poprzedniej analizy.*

Zad. 15. Czy możemy zaobserwować statystycznie istotną zmianę w poziomie wariancji wyjaśnionej po dodaniu płci (zapisz wielkość zmiany, oraz odpowiednie wartości testu F)?

Odp.: Nie, po dodaniu nowego proedyktora zmiana procentu wyjaśnionej wariancji wynosi jedynie 0,4% i jest nieistotna statystycznie, $F_{zmiany}(1, 290) = 3,37$; p = 0,067

Zad. 16. Zapisz równanie regresji dla ostatniego modelu:

ZAROBKI' = iq · 6,88 + educ · 47,93 + stat · 120,1 + sports · 51,54 + arts · 39,18 + plec · 48,29 + 36,21

Zad. 17. Jaki procent zmienności ZAROBKÓW przewiduje ten model (zapisz wartość odpowiedniego współczynnika, oraz wynik testu istotności dla modelu)

Odp.: Ostatni, trzeci zbudowany model wyjaśnia łącznie 60,5% wariancji średnich zarobków z ostatnich 3 lat. Wynik ANOVA dla tego modelu to: $F(6, 290) = 76,65$; $p < 0,001$.

Uwaga końcowa

Podczas pisania podsumowania zbudowanego modelu regresji można opisać naprawdę dużą liczbę współczynników i wyników testów uzyskanych w toku przeprowadzonej analizy. Oto lista ważnych aspektów, które należy poruszyć podczas pisania podsumowania modelu. Pogrubione punkty to istne „must be" czyli takie, których nie powinno zabraknąć w opisie jeśli chcesz otrzymać maksymalną liczbę punktów za zadanie. Pozostałe to takie, których brak nie powinien wpłynąć na obniżenie oceny, ale oczywiście dobrze byłoby je zawrzeć w opisie. O niektóre z tych punktów poprosi Cię prowadzący w treści zadania na kolokwium. Może, ale jednak nie

musi tego zrobić. ***Punkty nie są ułożone w kolejności, w której powinny być opisane w podsumowaniu modelu.***

1. ***Czy w modelu znajdowały się obserwacje odstające? Co z nimi zrobiono?***
2. ***Jaki jest procent wariancji wyjaśnionej przez nasz model?***
3. ***Który z kolei jest to model? Czy dodano do poprzedniego jakieś nowe predyktory w wyniku czego otrzymaliśmy kolejny model, który właśnie opisujemy?***
4. ***Czy wprowadzenie nowych predyktorów przyczyniło się do istotnego statystycznie wzrostu wyjaśnionej wariancji? O ile wzrósłten procent?***
5. ***Porównaj obecny model do poprzedniego (szczególnie pod względem R^2 i SEE)***
6. ***Czy model ten jest istotny?***
7. ***O ile myli się średnio nasz model ?***
8. ***Który predyktor jest istotny?***
9. ***Który predyktor jest najlepszy w naszym modelu?***
10. *Jaka jest wartość korelacji tych zmiennych poza modelem (korelacja rzędu zerowego)?*
11. *Czy predyktory korelują ze sobą istotnie statystycznie?*
12. *Jeśli było na zajęciach – czy złamano założenie o braku współliniowości między predyktorami (VIF i Tolerance)?*
13. *Jaki jest unikalny udział danego predyktora w wyjaśnianiu wariancji całkowitej (kwadrat korelacji semicząstkowej)?*
14. *Jak wygląda równanie modelu regresji, dzięki któremu możemy przewidzieć wartość zmiennej objaśnianej?*