WUM 2024, projekt 1

Celem zadania jest statystyczna analiza danych znajdujących się w pliku dane_projekt1.csv.

Dane: Są to dane symulowane; opisują fragment wyników badania ankietowego dotyczącego zwyczajów konsumenckich mieszkańców fikcyjnej krainy Bajtocji, przeprowadzonego na próbie reprezentatywnej[1]. Dane mogą zawierać losowe błędy. Poniżej objaśnienie nazw zmiennych zastosowanych w badaniu:

- *id* – identyfikator obserwacji, nie zawiera żadnej dodatkowej informacji
- *waga* – waga respondenta (w kg)
- *wzrost* – wzrost respondenta (w cm)
- *plec* – płeć w dokumencie tożsamości respondenta (1 – "kobieta", 2 – "mężczyzna")
- *dzieci* – liczba dzieci na utrzymaniu respondenta (w osobach)
- *wiek* – wiek respondenta (w latach)
- *dochod* – deklarowany dochód respondenta w badanym miesiącu (w bajtalarach)
- *oszczednosci* – deklarowane oszczędności respondenta w badanym miesiącu (w bajtalarach, ujemne wartości oznaczają, że wydatki ogółem przekroczyły dochód)
- *jednoos* – status gospodarstwa domowego (1 – "gospodarstwo jednoosobowe", 0 – "gospodarstwo wieloosobowe")
- *miejsce* – wielkość miejscowości, w której mieszka respondent (1 – "do 10 000 mieszkańców", 2 – "od 10 000 mieszkańców do 100 000 mieszkańców", 3 – "powyżej 100 000 mieszkańców")
- *wydatki_zyw* – deklarowane wydatki na żywność respondenta w badanym miesiącu (w bajtalarach).

Wynikiem ma być raport w notatniku jupyter (.ipynb). **Raport i komentarze muszą być wystarczające do zrozumienia i odtworzenia podejmowanych przez Państwa kroków bez konieczności czytania Państwa kodów.** Każde podjęte działanie modyfikujące w istotny sposób bazę (np. usuwanie rekordów, modyfikacja i wprowadzanie nowych zmiennych) musi być uzasadnione i opisane. W każdym zadaniu można skorzystać z gotowych implementacji. W moodle przedmiotu pojawi się zadanie – miejsce do przesłania raportu. Raport oceniany będzie przez prowadzącego Państwa grupę.

Termin oddania: 8 maja 2024 23:59. Proszę zapoznać się z polityką dotyczącą spóźnień, opisaną w moodle przedmiotu.

Suma punktów do zdobycia: 30

[1] Próba reprezentatywna to próba, której struktura ze względu na badane cechy (zmienne) jest zbliżona do struktury populacji statystycznej, z której pochodzi.

Zadanie 1: Wczytaj dane, obejrzyj je i podsumuj w dwóch-trzech zdaniach. Zadania pomocnicze:

- Ile jest obserwacji? Przedyskutuj strukturę zbioru danych: ile jest zmiennych ilościowych, a ile jakościowych? Czy występują braki danych? (1pkt)
- Przedstaw i skomentuj zasadne tabele częstości lub statystykę opisową dla zmiennych w zbiorze danych (zwróć uwagę na typ zmiennych). (1pkt)
- Przedstaw i skomentuj (tam, gdzie zasadne) rozkłady zmiennych, w szczególności porównując je wizualnie z rozkładem normalnym (np. z wykorzystaniem histogramów, wykresów kwantyl-kwantyl, etc). (2pkt)

Zadanie 2: Sprawdź, czy występują pomiędzy zmiennymi zależności. Policz i zaprezentuj na wykresie typu mapa ciepła (*heatmap*) zasadny współczynnik korelacji pomiędzy zmiennymi ilościowymi, a także zbadaj zależność zmiennych jakościowych. Skomentuj wyniki ze szczególnym uwzględnieniem kwestii istotności statystycznej. (3pkt)

Zadanie 3: Podsumuj dane przynajmniej trzema różnymi wykresami (skomentuj każdy z wykresów). Podstawowy zestaw wykresów zawiera:

- Wykresy rozrzutu (*scatter-plot*) dla wszystkich zmiennych ilościowych względem zmiennej *wydatki_zyw*.
- Wykresy typu pudełkowy (*boxplot*) dla jednej wybranej zmiennej ilościowej w podziale na miejsce zamieszkania respondentów.
- Wykres słupkowy skumulowany (*stacked bar chart*) dla płci respondenta i faktu, czy prowadzi jednoosobowe gospodarstwo domowe.

(3pkt, każdy wykres z podstawowego zestawu wart 1pkt: 0,25 pkt za sam wykres, 0,75pkt za komentarz w kontekście analizy eksploracyjnej. Mile widziane dodatkowe wykresy wg własnej inwencji uzupełniające analizę eksploracyjną, np. słupkowe, liniowe, kołowe... – możliwe dodanie do 1 dodatkowego punktu przez osobę sprawdzającą za interesujące dodatkowe wizualizacje)

Zadanie 4: Policz dwustronne przedziały ufności na poziomie ufności $1 - \alpha = 0.99$ dla zmiennej *wiek* dla następujących parametrów rozkładu:

- średnia i odchylenie standardowe;
- kwartyle 1, 2 i 3.

Podaj wykorzystane założenia i skomentuj, czy wydają Ci się one uprawnione (2pkt: 0,25pkt za średnią, 0,25pkt za wariancję, 0,75pkt za kwartyle, 0,75pkt za podanie i komentarz do przyjętych założeń)

Zadanie 5: Socjologowie bajtoccy dzielą społeczeństwo Bajtocji według czterech klas zamożności:

- klasa niższa (osiągany dochód poniżej 25 centylu rozkładu dochodów)
- klasa średnia (osiągany dochód równy lub wyższy 25 centylowi i niższy niż 75 centyl rozkładu dochodów)
- klasa wyższa średnia (osiągany dochód równy lub wyższy 75 centylowi i niższy niż 90 centyl rozkładu dochodów)
- klasa wyższa (osiągany dochód równy lub wyższy 90 centylowi rozkładu dochodów)

Przedyskutuj i porównaj zróżnicowanie wydatków na żywność w wyżej wymienionych klasach zamożności (2pkt: 0,5pkt za przeprowadzenie podziału, 1pkt za obliczenie właściwej miary zróżnicowania, 0,5pkt. za komentarz i dyskusję wyników).

Zadanie 6: Odpowiedz na następujące pytania badawcze, przeprowadzając najlepiej nadające się do tego testy statystyczne na poziomie istotności $\alpha = 0,01$:

- Czy kobiety cechują się wyższymi wartościami oszczędności niż mężczyzni?
- Czy niższa proporcja wydatków na żywność względem dochodu jest skorelowana z wyższymi oszczędnościami?
- Czy średnia waga kobiet w próbie jest wyższa niż 56 kg?

oraz:

- zweryfikuj dodatkową (sensowną) hipotezę o zgodności z konkretnym rozkładem parametrycznym dla wybranej zmiennej (np. "zmienna A ma rozkład Poissona z parametrem 1").

Podaj wykorzystywane założenia i skomentuj czy wydają Ci się one uprawnione. Każdy test statystyczny po 1 punkcie (w sumie 4pkt). Punktowane jest sformułowanie hipotezy zerowej oraz alternatywnej (0,25pkt), uzasadnienie/zasadność wybranego testu (0,25pkt), przeprowadzenie testu (0,25pkt) i podanie konkluzji testu (0,25pkt).

Zadanie 7: Przeprowadź badanie wysokości wydatków na żywność z wykorzystaniem zmiennych z bazy. Przyjmij poziom istotności $\alpha = 0.01$. W tym celu:

- Oszacuj wstępny model zawierający wszystkie zmienne z oryginalnej bazy (poza id) oraz stałą, gdzie zmienna *wydatki_zyw* jest zmienną objaśnianą. Pamiętaj o rozkodowaniu zmiennych jakościowych. (0,5pkt)
- Skomentuj R^2, testy łącznej i indywidualnej istotności we wstępnym modelu. (1pkt)
- Sprawdź, czy wstępny model spełnia założenia Klasycznego Modelu Regresji Liniowej (KMRL). Zwróć szczególną uwagę na kwestie liniowości formy funkcyjnej, homoskedastyczności i braku autokorelacji składnika losowego oraz rozkładu składnika losowego. (2pkt)
- Sprawdź, czy we wstępnym modelu występuje problem niedokładnej współliniowości (*multicollinearity*) (0,5pkt)
- Korzystając z analizy obserwacji odstających dla wstępnego modelu, sprawdź, czy baza zawiera błędy. Jeśli znajdziesz podejrzane obserwacje, zdecyduj i uzasadnij, co z nimi zrobisz. (1pkt)
- Popraw model tak, aby spełniał jak najwięcej założeń KMRL. Opisz kroki podjęte do otrzymania "najlepszego" modelu (4pkt).

  Wskazówka: Rozważ różne formy funkcyjne oraz transformacje zmiennych.
- Przedstaw ilościową interpretację wybranych dwóch indywidualnie istotnych współczynników w "najlepszym" modelu. Pamiętaj, że stałej nie interpretuje się. Zalecany wybór zmiennych niepoddanych transformacji. (1pkt)
- Jakie są opisowe charakterystyki osób, które cechują wydatki na żywność należące do górnych 10% predykcji wydatków na żywność w Państwa "najlepszym" modelu? Sprawdź i przedyskutuj (2pkt).