Projekt Zaliczeniowy 1

Dorota Celińska-Kopczyńska, Piotr Pokarowski, Ania Macioszek

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

Dane: Są to dane symulowane; opisują wiek (zmienna `wiek`), wagę w kg (`waga`), wzrost w cm (`wzrost`), płeć (`plec; "M" – mężczyzna, "K" – kobieta, "NA" – brak danych`), stan cywilny (`stan_cywilny; "1" – zamężna/żonaty, "0" – "panna/kawaler"`), liczbę dzieci (`liczba_dzieci`), typ budynku, w którym osoba mieszka (`budynek`), wydatki ogółem w badanym miesiącu w zł (`wydatki`), w tym wydatki na żywność w zł (`wydatki_zywnosc`) oraz bilans dochodów na koniec badanego miesiąca w zł (`oszczednosci`, ujemne wartości oznaczają, że wydatki przekroczyły dochód) pewnych osób. We wszystkich zadaniach poniżej zmienna `oszczednosci` jest **zmienną objaśnianą** (zależną), a pozostałe zmienne są **zmiennymi objaśniającymi** (niezależnymi).

Wynikiem ma być raport w formacie .Rmd oraz skompilowany do html lub pdf. Raport w obydwu formatach należy przesłać na adres email do prowadzącego laboratorium do sprawdzenia.

Termin oddania: 14 maja 2023
Suma punktów do zdobycia: 15

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

- Ile jest obserwacji, ile zmiennych ilościowych, a ile jakościowych? Czy występują braki danych? **(0,25 pkt.)**
- Przedstaw i skomentuj zasadne tabele częstości lub statystykę opisową dla zmiennych w zbiorze danych (zwróć uwagę na typ zmiennych) **(0,25 pkt.)**

2. Sprawdź, czy występują pomiędzy zmiennymi zależności (policz i zaprezentuj na wykresach zasadne współczynniki korelacji pomiędzy zmiennymi ilościowymi, a także zbadaj zależność zmiennych jakościowych). Skomentuj wyniki ze szczególnym uwzględnieniem kwestii istotności statystycznej. **(1 pkt)**

3. Podsumuj dane przynajmniej trzema różnymi wykresami. Należy przygotować:

a) wykres typu scatter-plot (taki jak na wykładzie 7, slajd 3) dla wszystkich zmiennych objaśniających ilościowych i zmiennej objaśnianej.
b) Wykresy typu pudełkowy (boxplot) dla jednej wybranej zmiennej ilościowej w podziale na stan cywilny respondentów.
c) Wykres kołowy (pie chart) dla jednej wybranej zmiennej jakościowej (wykres ma zawierać etykiety z procentami wystąpień danych kategorii).

Mile widziane dodatkowe wykresy wg własnej inwencji (np histogram, punktowy, liniowy, mapa ciepła...). **(1,5 pkt, każdy wykres z zestawu minimum wart 0,5 pkt.)**

4. Policz p-wartości dla hipotez o wartości średniej m = 70 i medianie me = 65 (kg) dla zmiennej waga, osobno w podpróbach kobiet i mężczyzn. Wybierz statystykę testową dla alternatywy lewostronnej, podaj założenia, z jakich korzystałeś i skomentuj czy wydają Ci się uprawnione. **(2 pkt, po 0,5 pkt. za każdą p-wartość)**

5. Policz dwustronne przedziały ufności na poziomie ufności 0.99 dla zmiennej wiek dla następujących parametrów rozkładu :

1. średnia i odchylenie standardowe;
2. kwantyle 1/4, 2/4 i ¾.

Podaj założenia, z jakich korzystałeś i skomentuj czy wydają Ci się uprawnione **(2 pkt).** *Wskazówka:* o przedziałach ufności dla kwantyli można przeczytać na przykład tu: https://www.r-bloggers.com/2016/10/better-confidence-intervals-for-quantiles/.

6. Odpowiedz na następujące pytania badawcze, przeprowadzając testy statystyczne na poziomie istotności 0,01:

1. Czy istnieją różnice w średnich wartościach wybranej zmiennej pomiędzy osobami zamężnymi/żonatymi a pannami/kawalerami w podpróbie osób w wieku poniżej 40 lat?
2. Czy w podpróbie osób w wieku poniżej 25 lat średnie wydatki ogółem są równe średnim wydatkom na żywność?
3. Czy niższy udział wydatków na żywność w wydatkach ogółem jest skorelowany z wyższymi oszczędnościami?

Ponadto, 4. przetestuj hipotezę o zgodności z konkretnym rozkładem parametrycznym dla wybranej zmiennej (np. "zmienna A ma rozkład wykładniczy z parametrem 10").
Podaj założenia, z jakich korzystałeś i skomentuj czy wydają Ci się uprawnione.
Każdy test statystyczny po **1 punkcie** (w sumie **4**). Punktowane jest sformułowanie hipotezy zerowej, wybranie (i uzasadnienie) właściwego testu, przeprowadzenie testu i podanie konkluzji testu.

7. Oszacuj model regresji liniowej, przyjmując za zmienną zależną (y) bilans dochodów na koniec miesiąca (`oszczednosci`) a jako zmienne niezależne (x) przyjmując pozostałe zmienne. Rozważ, czy konieczne są transformacje zmiennych (objaśniających lub objaśnianej). Podaj RSS, R^2, p-wartości i oszacowania współczynników w pełnym modelu (w modelu zawierającym wszystkie zmienne). Następnie wybierz jedną zmienną objaśniającą, którą można by w pierwszej kolejności z pełnego modelu odrzucić (która najgorzej tłumaczy `oszczednosci`). Aby dokonać wyboru takiej zmiennej, dla każdej ze zmiennych objaśniających sprawdź:

- Jaką ma p-wartość w pełnym modelu?
- O ile zmniejsza się R^2, gdy ją usuniemy z pełnego modelu?
- O ile zwiększa się RSS, gdy ją usuniemy z pełnego modelu?

Opisz wnioski. Oszacuj model ze zbiorem zmiennych objaśniających pomniejszonym o wybraną zmienną. Sprawdź czy w otrzymanym przez Ciebie modelu spełnione są założenia modelu liniowego. Przedstaw (i skomentuj) wykresy diagnostyczne: wykres zależności reszt od zmiennej objaśnianej, wykres reszt studentyzowanych i dźwigni. (**4 pkt**).