Projekt Zaliczeniowy 1

Ania Macioszek, Dorota Celińska-Kopczyńska, Piotr Pokarowski

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

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; "T" - zamężna/żonaty, "F" - "panna/kawaler"`), liczbę dzieci (`liczba_dzieci`), typ budynku, w którym osoba mieszka (`budynek`), wydatki w badanym miesiącu w zł (`wydatki`) 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: 12 maja 2022
Suma punktów do zdobycia: 15

1. Wczytaj dane, obejrzyj je i podsumuj w dwóch-trzech zdaniach. Pytania pomocnicze: ile jest obserwacji, ile zmiennych ilościowych, a ile jakościowych? Czy są zależności w zmiennych objaśniających (policz i zaprezentuj na wykresach korelacje pomiędzy zmiennymi ilościowymi, a także zbadaj zależność zmiennych jakościowych). Skomentuj wyniki. Czy występują jakieś braki danych? **(2 pkt)** Git

2. 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. Git
b) Wykresy typu pudełkowy (boxplot) dla jednej wybranej zmiennej ilościowej w podziale na płeć respondentów. Git
c) Wykres kołowy (pie chart) dla jednej wybranej zmiennej jakościowej (wykres ma zawierać etykiety z procentami wystąpień danych kategorii). Git

Mile widziane dodatkowe wykresy wg własnej inwencji (np histogram, punktowy, liniowy, mapa ciepła...). **(2 pkt)**

3. Policz p-wartości dla hipotez o wartości średniej m = 170 i medianie me = 165 (cm) dla zmiennej wzrost. Wybierz statystykę testową dla alternatywy lewostronnej, podaj założenia, z jakich korzystałeś i skomentuj czy wydają Ci się uprawnione. **(2 pkt)** Spytac o co chodzi z alternat

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

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

O co chodzi z roznica pomied

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/.

5. Przetestuj na poziomie istotności 0.01 trzy hipotezy:

1. średnie wartości wybranej zmiennej pomiędzy osobami zamężnymi/żonatymi a pannami/kawalerami są równe;
2. dwie wybrane zmienne ilościowe są niezależne;
3. dwie wybrane zmienne jakościowe są niezależne.

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.
Każda hipoteza po **1 punkcie** (w sumie **4**). Punktowane jest sformułowanie hipotezy zerowej, wybranie właściwego testu, przeprowadzenie testu i podjęcie decyzji czy odrzucamy hipotezę zerową.

6. 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 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. (**3 pkt**).