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 `age`), wagę (`weight`), wzrost (`height`), płeć (`gender`), stan cywilny (`married`), liczbę dzieci (`number_of_kids`), posiadane zwierzę domowe (`pet`) oraz miesięczne wydatki (`expenses`) pewnych osób. We wszystkich zadaniach poniżej zmienna `expenses` jest **zmienną objaśnianą** (zależną), a pozostałe zmienne są **zmiennymi objaśniającymi** (niezależnymi).

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

Termin oddania: 9 maja 2021
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)**

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

a) wykres typu scatter-plot (taki jak na wykładzie 6, slajd 3) dla wszystkich zmiennych objaśniających ilościowych i zmiennej objaśnianej.
b) Wykresy typu pudełkowy (boxplot) dla jednej wybranej zmiennej ilościowej.
c) Wykres typu słupkowy (barplot) dla jednej wybranej zmiennej jakościowej.

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)**

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.

Podaj założenia, z jakich korzystałeś i skomentuj czy wydają Ci się uprawnione **(2 pkt).**

5. Przetestuj na poziomie istotności 0.01 trzy hipotezy istotności:

1. różnicy między średnią wartością wybranej zmiennej dla kobiet i dla mężczyzn;
2. zależności między dwiema zmiennymi ilościowymi;
3. zależności między dwiema zmiennymi jakościowymi.

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) wydatki domowe (`expenses`) a jako zmienne niezależne (x) przyjmując pozostałe zmienne. Rozważ, czy konieczne są transformacje zmiennych lub zmiennej 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 `expenses`). 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 i przedstaw na wykresach diagnostycznych: wykresie zależności reszt od zmiennej objaśnianej, na wykresie reszt studentyzowanych i na wykresie dźwigni i przedyskutuj, czy są spełnione. (**3 pkt**).