Projekt Zaliczeniowy 1

Ania Macioszek, Dorota Celińska-Kopczyńska

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.

Termin oddania: 10 maja 2020

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 podaj VIF)? Czy występują jakieś braki danych? **(1 pkt)**

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

a) wykres typu scatterplot (taki jak na wykładzie 7 (https://www.mimuw.edu.pl/~szczurek/SAD1/Wyklady/07_W_RegresjaLiniowa2.pdf), 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...). **(3 pkt)**

3. Podaj przedziały ufności dla wartości średniej i wariancji dla zmiennych wiek i wzrost. Jeżeli w celu wyliczenia przedziału ufności musisz poczynić jakieś założenia (np. założyć że zmienna pochodzi z rozkładu normalnego), zaznacz to i skomentuj czy wydaje Ci się to w danym przypadku uprawnione. Opisz wszelkie dodatkowe operacje, jakie zostały wykonane przed testem (takie jak usunięcie obserwacji odstających). Przedyskutuj, dla której ze zmiennych oczekujesz prawidłowych wyników. **(1 pkt)**

4. Sformułuj i zweryfikuj cztery hipotezy:

1. dotyczącą różnicy między średnią wartością wybranej zmiennej dla kobiet i dla mężczyzn
2. dot. niezależności między dwoma zmiennymi ilościowymi
3. jedną dot. niezależności między dwoma zmiennymi jakościowymi

4. jedną dot. rozkładu zmiennej (np. "zmienna A ma rozkład wykładniczy z parametrem 10")

Każda hipoteza po **2 punkty** (w sumie **8 pkt**). Punktowane jest sformułowanie hipotezy zerowej i alternatywnej, wybranie właściwego testu, przeprowadzenie testu i podjęcie decyzji czy odrzucamy hipotezę zerową.

4. Oszacuj model regresji liniowej, przyjmując za zmienną zależną (y) wydatki domowe (`expenses`) a zmienne niezależne (x) wybierając spośród pozostałych zmiennych. Rozważ, czy konieczne są transformacje zmiennych lub zmiennej objaśnianej. Podaj RSS, R^2, p-wartości i oszacowania współczynników i wybierz właściwe zmienne objaśniające, które najlepiej tłumaczą `expenses`. Sprawdź czy w wybranym 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. (**2 pkt**).