5.1 Zbiór *airpollution.txt* zawiera dane dotyczące związku pomiędzy zanieczyszczeniem powietrza i śmiertelnością w 60 miastach amerykańskich. Między zmiennymi są:

Mortality - skorygowana wiekiem liczba zgonów na 100 000 mieszkańców,

Education - mediana liczby lat kształcenia,

NonWhite - procent tej podpopulacji,

income - mediana zarobków w tys. dolarów,

JanTemp, JulTemp - średnie temperatury w styczniu i lipcu (w stopniach Fahrenheita),

NOx - stężenie tlenku azotanu.

(a) Dopasuj model liniowy ze zmienną objaśnianą *Mortality* i zmienną objaśniającą *NOx*. Podaj współczynnik nachylenia prostej MNK oraz jego błąd standardowy. Sprawdź, czy dopasowany model dobrze opisuje dane.

(b) Dopasuj model liniowy ze zmienną objaśnianą *Mortality* i zmienną objaśniającą *log(NOx)*. Podaj współczynnik nachylenia prostej MNK oraz jego błąd standardowy. Czy model ten dobrze opisuje dane?

(c) W modelu liniowym ze zmienną objaśnianą *Mortality* i zmienną objaśniającą *log(NOx)* znajdź obserwacje o dużych rezyduach studentyzowanych. Sporządź nowy model pomijając te obserwacje. Porównaj wartości współczynnika R^2 dla tych dwóch modeli.

5.2 W zbiorze danych *phila.txt* zebrano informacje dotyczące cen domów położonych w okolicach Filadelfii (zmienna *HousePrice*) i innych ich cech, na przykład wskaźnika przestępczości w okolicy, w której dom jest położony (zmienna *CrimeRate*).

(a) Wczytaj zbiór i zwrócić uwagę na fakt, że brakuje części danych.

(b) Interesuje nas zależność ceny domu od wskaźnika przestępczości w jego okolicy. Sporządź stosowny wykres rozproszenia i dopasuj model liniowy. Zidentyfikuj potencjalną obserwację odstającą i wpływową. Usuń ją ze zbioru i ponownie dopasuj model. Czy nowy model dobrze opisuje dane?

5.3 Zbiór danych *cellular.txt* zawiera informacje dotyczące liczby czytelników pewnego czasopisma (zmienna *Subscribers*).

(a) Sporządź wykres liczby czytelników w funkcji czasu (zmienna *Period*). Dopasuj do tych danych model liniowy i przeprowadź jego diagnostykę. Zwróć uwagę na fakt, że współczynnik β_1 jest istotny, mimo, że model jest źle dopasowany.

(b) Wybierz model najlepszy spośród następujących modeli alternatywnych: $log(Subscribers) \sim Period$, $(Subscribers)^{1/2} \sim Period$, $(Subscribers)^{1/4} \sim Period$.

(c) Używając metody Boxa-Coxa, wyznacz przekształcenie $g_\lambda(\cdot)$ zmiennej *Subscribers* dające najlepszy model liniowy $g_\lambda(Subscribers) \sim Period$. W jakim sensie model $g_\lambda(Subscribers) \sim Period$ jest najlepszy?

5.4 Dane w pliku *savings.txt* dotyczą sytuacji ekonomicznej mieszkańców 50 krajów. Poszczególne kolumny zawierają wartości średnie z lat 1960-1970:

Country - nazwa kraju,

Savings - łączne oszczędności przypadające na osobę podzielone przez dochód netto,

dpi - dochód netto przypadający na jednego mieszkańca, *ddpi* - tempo wzrostu dochodu (w %),

Pop15 , *Pop75* - procent obywateli w wieku, odpowiednio, mniejszym niż 15 lat i powyżej 75 lat.

(a) Dopasuj model liniowy opisujący zależność *Savings* od *dpi* , *ddpi* , *Pop15* i *Pop75* . Zidentyfikuj obserwacje o dużych studentyzowanych rezyduach modyfikowanych i dużych wpływach (h_{ii}). Sporządź

diagram Cooke'a. Usuń z modelu obserwację o największej odległości Cooke'a. Sprawdź, że jest ona wpływowa.

(b) Dlaczego *dpi* nie jest istotne w modelu dopasowanym w punkcie (a)? Sporządź częściowe wykresy regresji dla zmiennych *dpi* i *ddpi* i zinterpretuj je.

(c) Obliczy współczynnik korelacji dla zmiennych *Pop15* i *Pop75*. Porównaj go ze współczynnikiem korelacji dla estymatorów β_{Pop15} i β_{Pop75}. Jak możemy wytłumaczyć zaobserwowane zjawisko?

(d) Sporządź wykres częściowych rezyduów dla zmiennej *Pop15*. Jaką dodatkową relację w danych możemy zauważyć na tym wykresie?

5.5 Wygeneruj trzy zestawy danych o liczności $n = 30$:

$$x_1 \sim U[0, 10], \quad \varepsilon_1 \sim N(0, 1)$$

$$x_2 \sim U[10, 20], \quad \varepsilon_2 \sim N(0, 3)$$

$$x_3 \sim U[20, 30], \quad \varepsilon_3 \sim N(0, 5)$$

(a) Sporządź wektor x powstały z połączenia wektorów $x1, x2, x3$ i wektor ε powstały z połączenia wektorów $\varepsilon 1, \varepsilon 2, \varepsilon 3$. Stwórz wektor $y = x + \varepsilon$ i dopasuj model.

(b) Sporządź wykres rezyduów studentyzowanych modyfikowanych, w zależności od numeru obserwacji. Co można z niego wywnioskować?

(c) Dopasuj metodę ważonych najmniejszych kwadratów. Jak wygląda wykres z poprzedniego podpunktu dla tej metody?

(d) Do danych dodaj punkt $(x, y) = c(5, 10)$. Który z modeli zidentyfikuje obserwację jako odstającą?