SAR 2020/2021 Laboratorium 5

3-4.11.2020

5.1 (obserwacje odstające - rezidua)

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) Wczytać zbiór i zwrócić uwagę na fakt, że brakuje części danych (policzyć, jaka część obserwacji ma braki danych). Usunąć obserwacje z brakami danych ze zbioru.

b) Interesuje nas zależność ceny domu od wskaźnika przestępczości w jego okolicy.

c) Zidentyfikować potencjalną obserwację odstającą. Narysować wykresy reziduów (zwykłych).

d) Policzyć rezidua studentyzowane (funkcją `rstandard()` i z definicji) i narysować wykres.

e) Policzyć modyfikowane rezidua studentyzowane (funkcją `rstudent()`, korzystając ze wzoru i dopasowując n modeli liniowych, gdzie n to liczba obserwacji). Narysować wykresy.

f) Usunąć obserwację odstającą ze zbioru i ponownie dopasować model. Czy nowy model dobrze opisuje dane? Narysować te same wykresy i policzyć te same rezidua, co w poprzednich podpunktach.

5.2 (wykresy reziduów)

Wygeneruj dane z poniższych modeli. Następnie sporządź odpowiednie (czyli takie, na których będzie można zaobserwować występujący problem w danych) wykresy reziduów.

a) x_i - wartości od 0 do 10 co 0.01, $Y_i = x_i + \varepsilon_i$, $\varepsilon \sim N(0, \text{diag}(\text{seq}(0.1, 5, \text{n})))$, n - liczba obserwacji.

b) x_i - tak samo, $Y_i = \sin(x_i) + \varepsilon_i$, $\varepsilon \sim N(0, \sigma^2 I)$, $\sigma = 0.5$

c) x_i - tak samo, $Y_i = (x_i + \varepsilon_i)^2$, $\varepsilon \sim N(0, \sigma^2 I)$, $\sigma = 0.5$