

Ćwiczenie nr 5 – Obserwacje nietypowe w analizie regresji

1. Wczytaj dane *savings* z biblioteki *faraway*. Dane te zawierają nast. wielkości uśrednione za lata 1960 - 1970: *sr* - wielkość oszczędności w stosunku do dochodu netto, *pop15* - procent populacji poniżej 15 roku życia, *pop75* - procent populacji powyżej 75 roku życia, *dpi* - dochód netto w dolarach na głowę, *ddpi* - procentowy wzrost "dpi".

Przeprowadź regresję ze zmienną zależną *sr* oraz zmiennymi niezależnymi: *pop15*, *pop75*, *dpi*, *ddpi*.

- Narysuj wykres reszt w tym modelu. Zidentyfikuj, którym krajom odpowiada najmniejsza i największa wartość reszt.

```
g<-lm(sr~pop15+pop75+dpi+ddpi,savings);sort(g$res)[c(1,50)]
countries<-row.names(savings);identify(1:50,g$res,countries)
```

- Odczytaj i narysuj wartości dźwigni (leverage). Dla których krajów wartość dźwigni są duże?

```
lev<-hatvalues(g)
plot(lev);identify(1:50,lev,countries)
```

- Wyznacz reszty studentyzowane. Które z nich są duże?

```
stud<-rstandard(g)
```

- Oblicz usunięte wartości resztowe. Wyciągnij wnioski z otrzymanych wyników.

```
jack<-rstudent(g);
qt() - funkcja wyznaczająca kwantyle rozkładu t-studenta
```

- Wyznacz wartości miar DFFITS, DFBETAS. Zinterpretuj uzyskane wyniki.

```
dffits() - funkcja służąca do obliczania wartości DFFITS
dfbeta() - funkcja służąca do obliczania wartości DFBETAS
```

- wyznacz odległości Cooka. Które z nich można uważać za duże?

```
cooks.distance() - funkcja służąca do obliczania odległości Cooka
```

- Wskaż wszystkie obserwacje wpływowe.

```
inf<-influence.measures(g);inf
summary(inf)
```

Przeprowadź regresję dla danych z wyłączoną obserwacją o największej wartości odległości Cooka. Porównaj nowy model z poprzednio rozważanym.

- Narysuj wykres zmian wartości współczynników przy zmiennych *pop15* oraz *pop75* w modelu z usuniętą obserwacją. Który kraj ma największy wpływ?

```
ginf<-lm.influence(g)
plot(ginf$coef[,2],ginf$coef[,3],xlab="pop15 coef",ylab="pop75 coef")
identify(ginf$coef[,2],ginf$coef[,3],countries)
```

Przeprowadź analogiczną analizę dla pozostałych współczynników.

- Rozważ model z usuniętą Japonią. Porównaj go z modelem pełnym.

```
gj<-lm(sr~pop15+pop75+dpi+ddpi,savings,subset=(countries!="Japan"))
```

- Narysuj wykresy "half-normal plot" dla wartości dźwigni oraz odległości Cooka w modelu pełnym. Zinterpretuj wyniki.

```
halfnorm(lm.influence(g)$hat,labs=countries,ylab="Leverages")
halfnorm(cooks.distance(g),labs=countries,ylab="Cook Statistics")
```