Lepiej zapobiegać niż leczyć

Diagnostyka regresji

> Anceps remedium melius quam nullum

Na tych zajęciach nauczymy się identyfikować zagrożenia dla naszej analizy regresji. Jednym elementem jest oczywiście niespełnienie założeń. Nie wszystkie założenia są równie istotne i mają równie dramatyczny wpływ na wyniki. Drugim jest fakt, że dane mogą być źle sformatowane, zapisane w złej skali lub po prostu błędne.

Model liniowy:

$$y \sim X\beta + \epsilon$$

Główne założenie jaki robimy to IID o resztach. Diagnostyka będzie się zatem opierać na badaniu, na ile to założenie jest spełnione w naszym modelu.

Anscombe's quartet

Dane spreparowane przez Franka Anscombe w 1973 roku.

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## # A tibble: 4 x 6
##    group mean(x)    sd(x)  mean(y)    sd(y) cor(x, y)
##   <fctr>   <dbl>    <dbl>    <dbl>    <dbl>     <dbl>
## 1      1       9 3.316625 7.500909 2.031568 0.8164205
## 2      2       9 3.316625 7.500909 2.031657 0.8162365
## 3      3       9 3.316625 7.500000 2.030424 0.8162867
## 4      4       9 3.316625 7.500909 2.030579 0.8165214
```

Podstawowe statystyki dla każdego zbioru są takie same. Co więcej, podobnie ma się sprawa ze współczynnikami regresji liniowej.

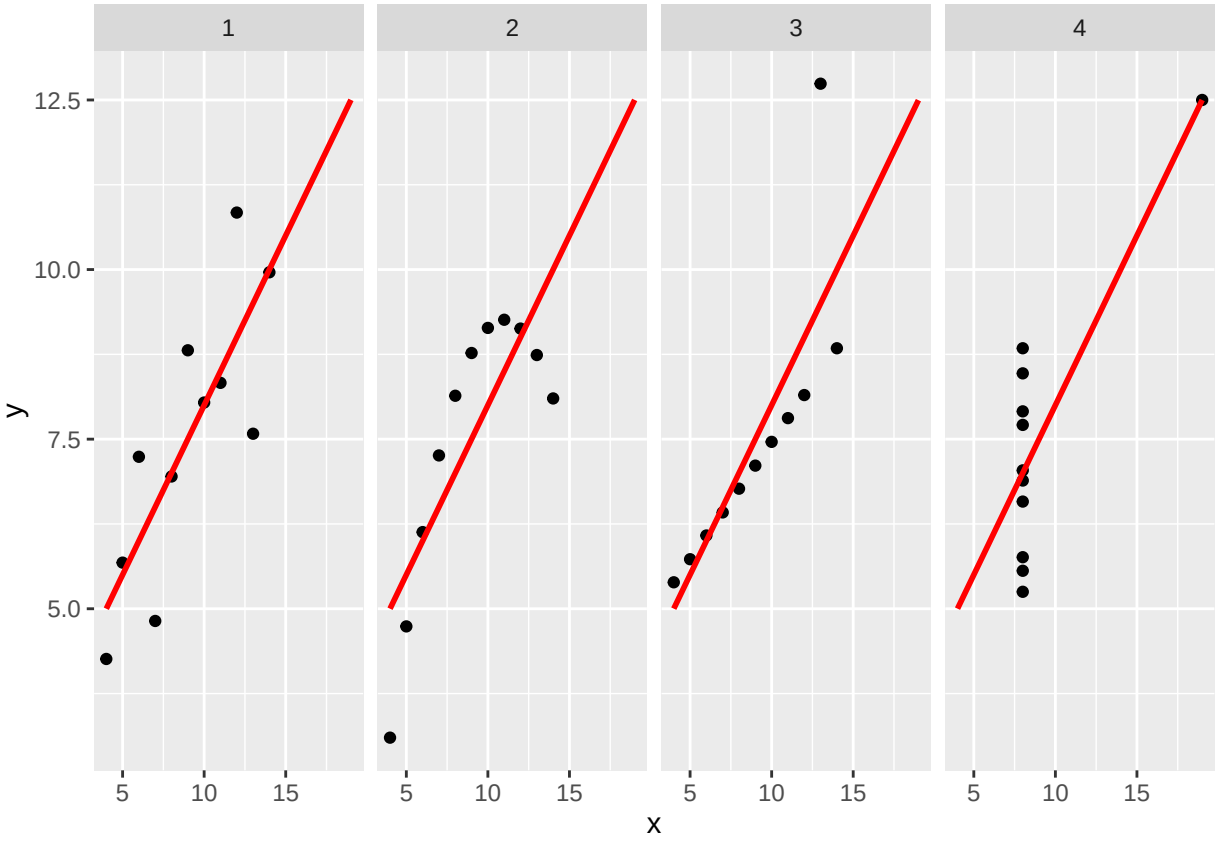
```
for(i in 1:4){
  print(i)
  print(summary(lm(y~x, anscombe.data, anscombe.data$group==i))$coeff)
}
```

```
## [1] 1
##              Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 3.0000909  1.1247468 2.667348 0.025734051
## x           0.5000909  0.1179055 4.241455 0.002169629
## [1] 2
##             Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 3.000909  1.1253024 2.666758 0.025758941
## x           0.500000  0.1179637 4.238590 0.002178816
## [1] 3
##              Estimate Std. Error  t value    Pr(>|t|)
```

```
## (Intercept) 3.0024545  1.1244812 2.670080 0.025619109
## x           0.4997273  0.1178777 4.239372 0.002176305
## [1] 4
##              Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 3.0017273  1.1239211 2.670763 0.025590425
## x           0.4999091  0.1178189 4.243028 0.002164602
```

Okazuje się jednak, że każdy z tych zbiorów jest zupełnie inny. W dodatku tylko jeden z nich „nadaje" się do dopasowania modelu liniowego

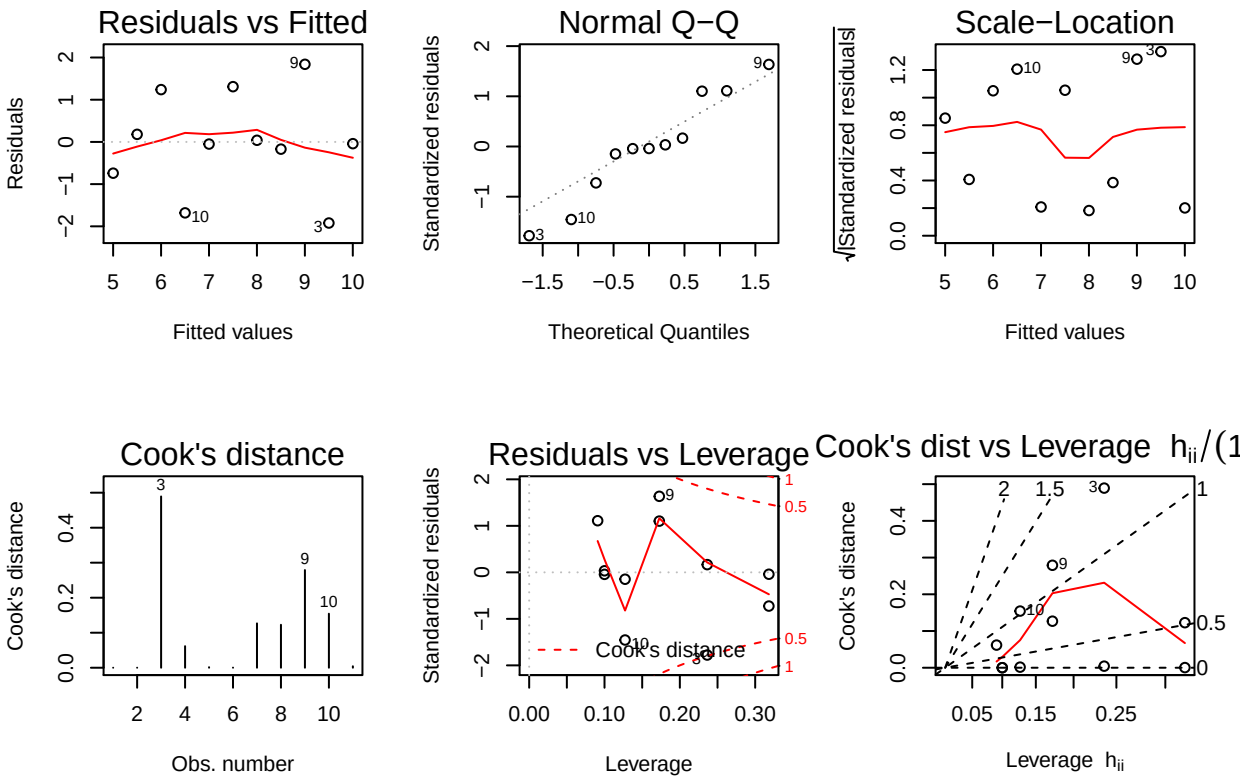
```
ggplot(anscombe.data, aes(x,y)) +
  geom_point() + facet_grid(.~group) +
  stat_smooth(method = "lm", col = "red", se = F, fullrange = T)
```



W tym wypadku jesteśmy w stanie organoleptycznie stwierdzić, że coś jest nie tak. Przy większej liczbie wymiarów mogłoby to być niemożliwe. Z oczywistych względów zależy nam na tym, żeby sprawdzić czy dopasowany model ma sens. Pomoże nam w tym diagnostyka regresji.

Funkcje z pakietu base

```
lmAnscombe1=lm(y~x, anscombe.data, anscombe.data$group==1)
par(mfrow=c(2,3)) # Change the panel layout to 2 x 2
plot(lmAnscombe1, 1:6)
```




```
par(mfrow=c(1,1))
```

Na dzisiejszych zajęciach nauczymy się analizować powyższe wykresy.

Residuals vs Fitted

Residua poznaliśmy już na poprzednich zajęciach.

$$r = \hat{\epsilon} = y - \hat{y} = y - Hy = (I - H)y$$

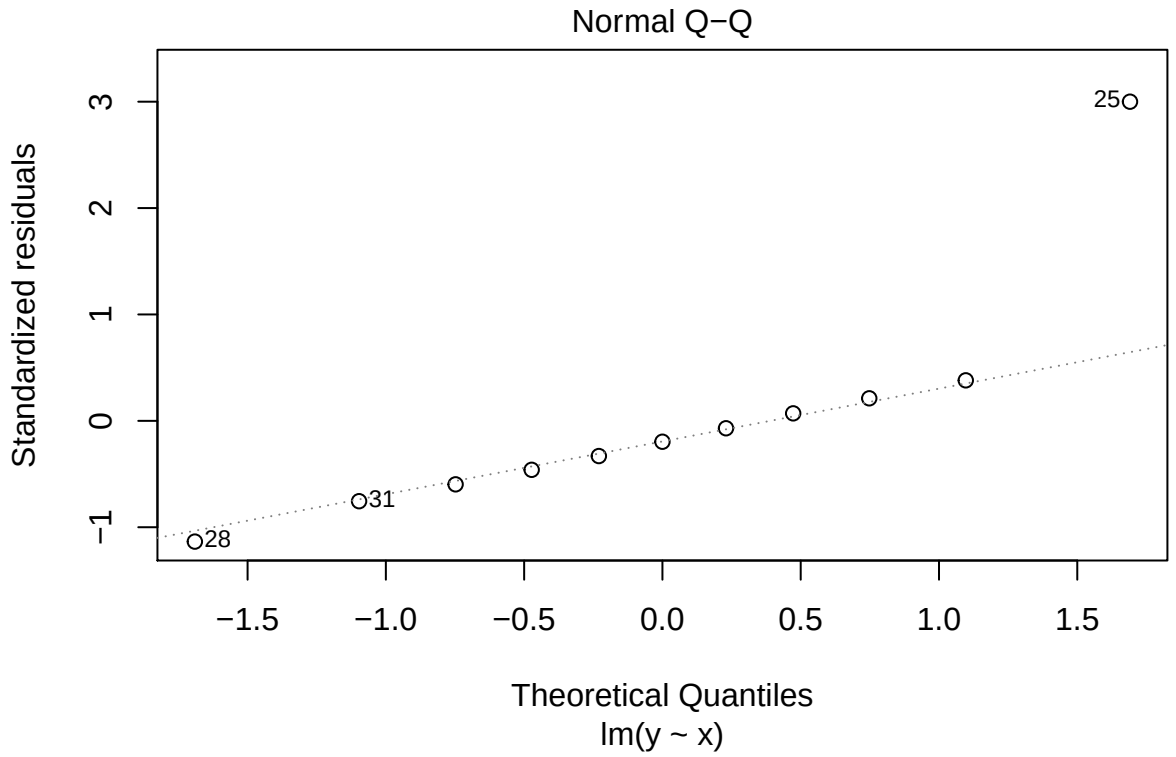
Residua powinny oscylować wokół 0, ponieważ rozkład residuów ma średnią 0 (dlaczego?). W szczególności nie powinno być zależności residuów od $\hat{y}$.

Uwaga! Kiedy mamy więcej niż jedną zmienną objaśniającą warto jest porównać wizualnie reszty i wartości dla każdej ze zmiennych objaśniających. Co nam mówi zależność w tych danych? Jak sprawdzić czy taka zależność występuje?

Normal QQ

Zobaczmy coś, co nie działa

```
lmAnscombe3=lm(y~x, anscombe.data, anscombe.data$group==3)
plot(lmAnscombe3, 2)
```



Czym są standardized residuals? Zauważmy, że:

$$r = \hat{\epsilon} = y - \hat{y} = y - Hy = (I - H)y = (I - H)\epsilon.$$

Ostatnia równość wynika wprost z definicji. Zatem

$$r \sim N(0, \sigma^2(I - H)^T(I - H) = N(0, \sigma^2(I - H)).$$

Ponieważ znamy macierz H, to możemy ustandaryzować r, tak żeby każde residuum miało wariancję równą 1

$$r_i^{std} = \frac{r_i}{\sqrt{\hat{\sigma}^2(1 - h_i)}} \sim N(0, 1)$$

Gdzie h_i jest i-tym elementem na przekątnej macierzy H. Podkreślmy jeszcze raz, że reszty (także standaryzowane) NIE są niezleżne. Ich zalezność dana jest przez macierz H.

Ćwiczenie, z własności macierzy rzutu H

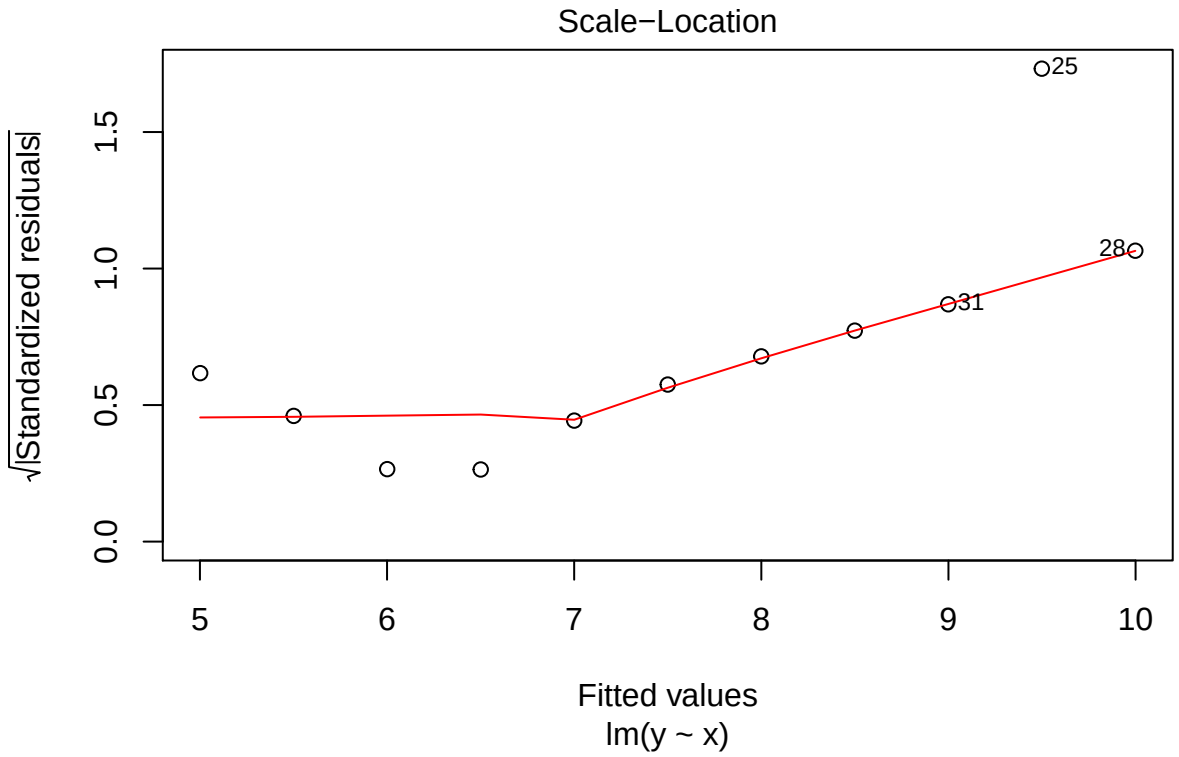
1. $\forall_i \frac{1}{n} < h_i < 1$
2. $\sum_i h_i = p$

Zatem możemy oczekiwać, żeby reszty standaryzowane układały się wzdłuż wykresu qqnorm. Obserwacje, które nie sprawiają, że mamy grubsze ogony niż rozkład normalny są obserwacjami odstającymi.

Co zrobić jeśli jesteśmy bardzo daleko od rozkładu normalnego?

Scale location

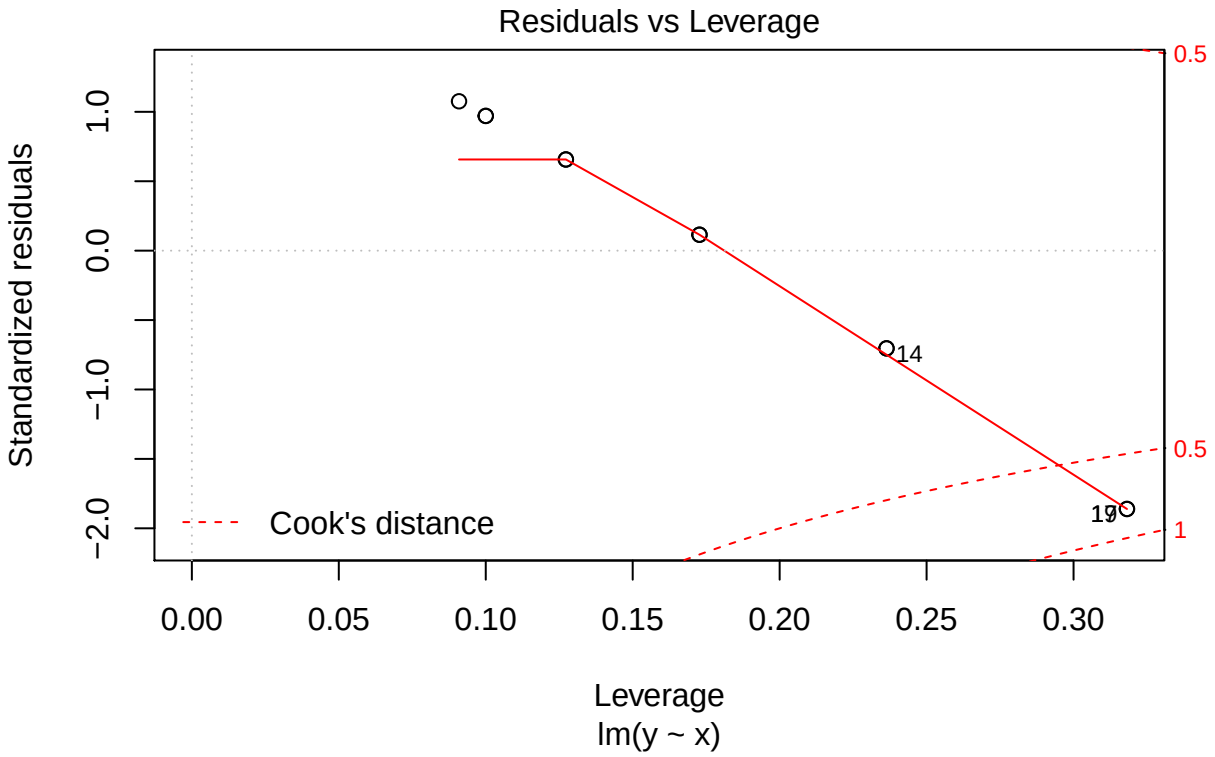
```
plot(lmAnscombe3, 3)
```



Ważnym założeniem jakie czynimy w regresji liniowej jest jednorodność wariancji. W szczególności wariancja nie powinna zależeć funkcyjnie od wartości $\hat{y}_i$.

Residuals vs Leverage

```
lmAnscombe2=lm(y~x, anscombe.data, anscombe.data$group==2)
plot(lmAnscombe2, 5)
abline(v=2*ncol(anscombe.data)/nrow(anscombe), col="blue")
```



Przez dźwignię (ang. leverage) oznaczamy wartości h_i. Skąd nazwa dźwignia i o czym ona mówi? h_i mówi o wpływie jaki na $\hat{y}_i$ ma y_i.

$$h_i = \frac{\partial \hat{y}_i}{\partial y_i}$$

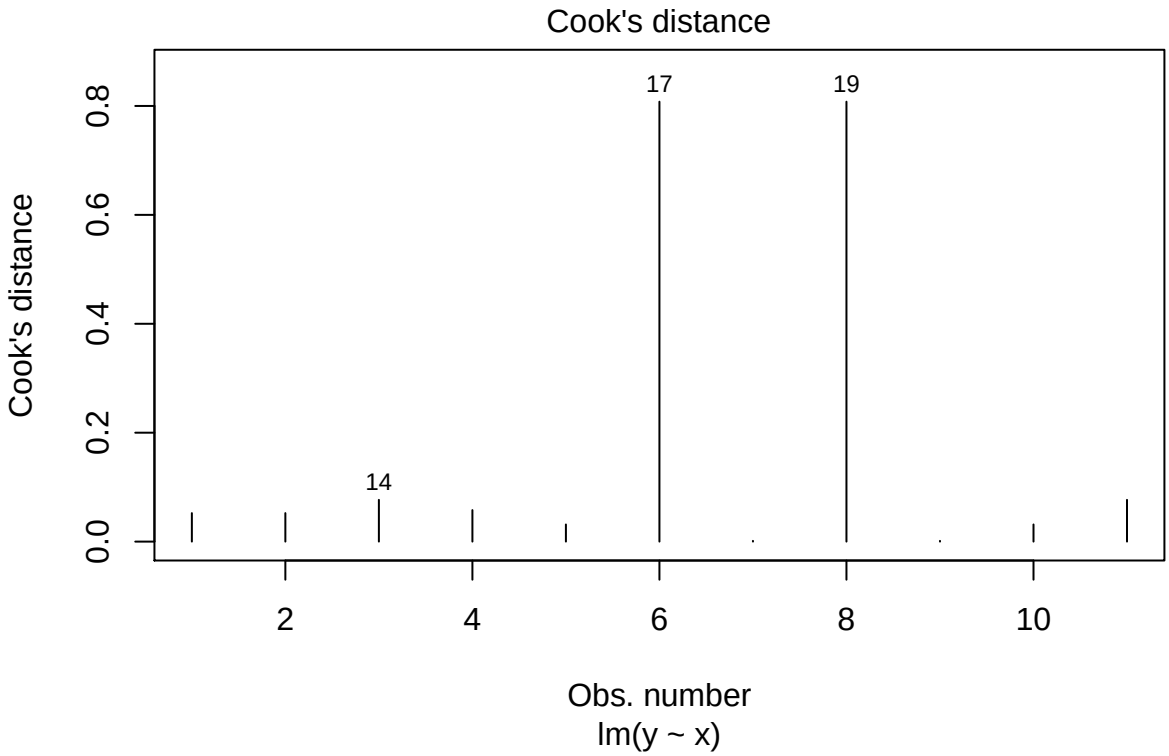
Przeciętna wartość $h_i = \frac{p}{n}$ (dlaczego?). W związku z tym, reguła kciuki mówi, że obserwacjami wpływowymi są te, dla których $h_i \geq \frac{2p}{n}$. p to liczba kolumn macierzy X.

Na powyższym wykresie możemy zidentyfikować obserwacje wpływowe, a także sprawdzić czy wariancja residuów nie zależy od wpływowości obserwacji. Innymi słowy, chcemy, żeby nie było zależności.

Obserwacje o dużym leverage są potencjalnie obserwacjami wpływowymi, to znaczy mogą mieć duży wpływ na wyznaczenie parametrów w modelu i dopasowanie danych do modelu (R^2).

Cook's distance

```
plot(lmAnscombe2, 4)
```



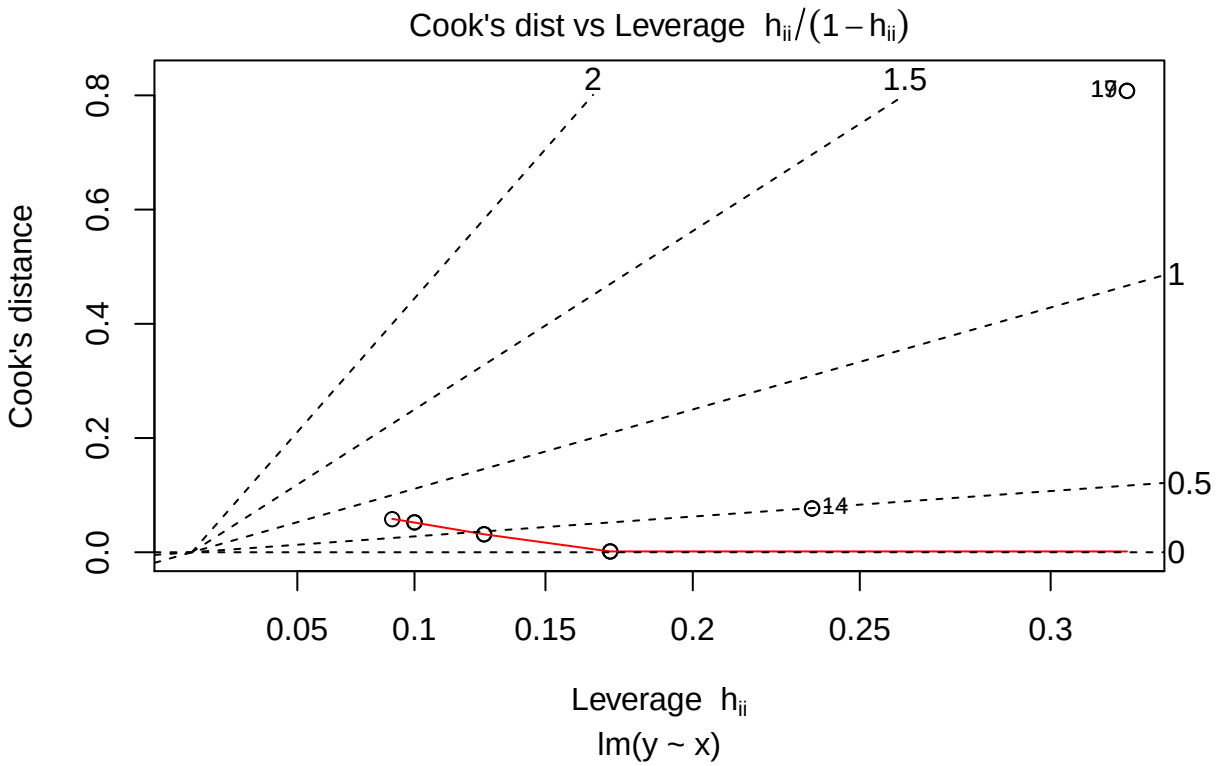
Miara określająca jaki wpływ na dopasowane wartości $\hat{y}$ ma obserwacja x_i. Liczymy dopasowania dla modelu z pełnymi danymi, i z danymi bez i-tej obserwacji.

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j,-(i)})^2}{p\hat{\sigma}^2}$$

Dlaczego takie normowanie? $p\hat{\sigma}^2$ to RSS w naszym modelu. Odjęcie jednej obserwacji, nie powinno mieć większego wpływu. Stąd reguła kciuka mówi, że wpływowa jest obserwacja o mierze Cooka większej niż 1.

Cook's dist vs Leverage

```
plot(lmAnscombe2, 6)
```



Zauważmy, że możemy przepisać z definicji miarę Cooka jako:

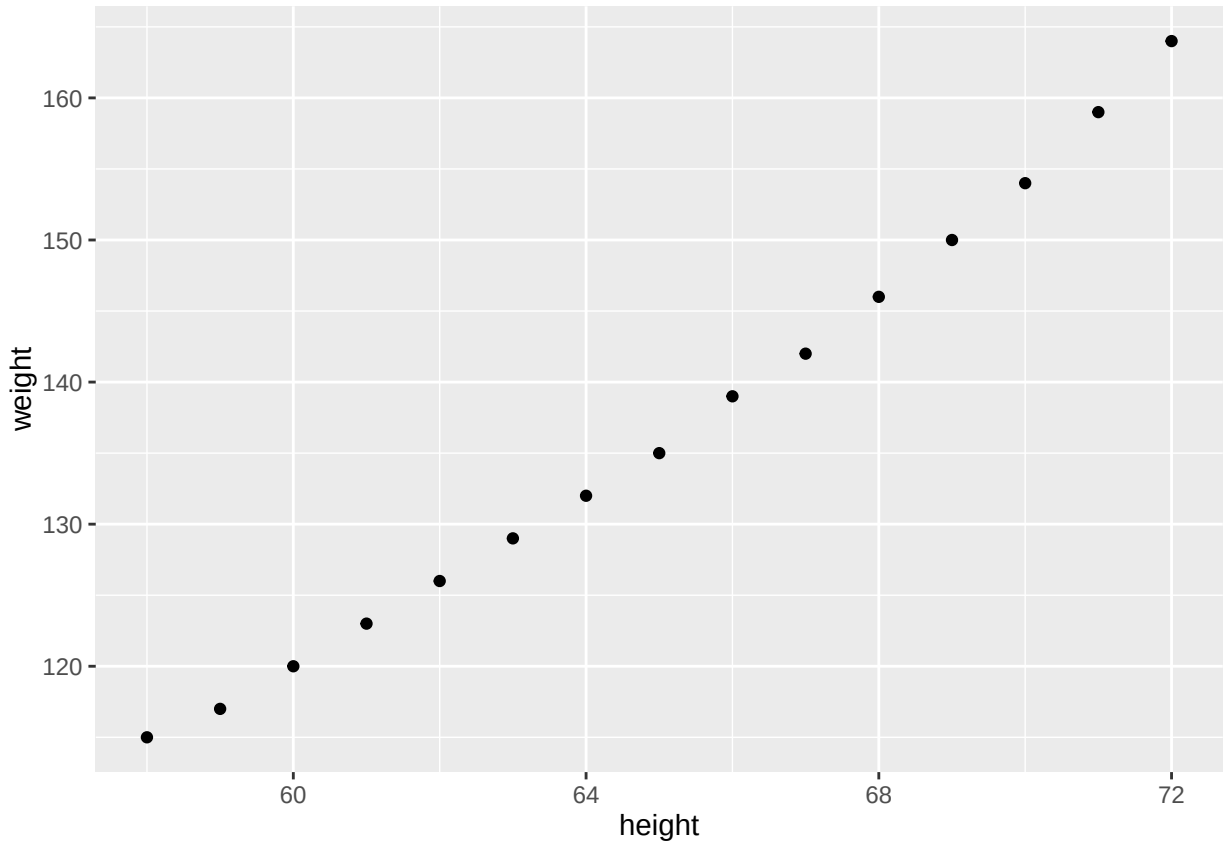
$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{-(i)})^T (X^T X)(\hat{\beta} - \hat{\beta}_{-(i)})}{p\hat{\sigma}^2} = \frac{1}{p} \frac{h_i}{1-h_i} r_i^2$$

czyli zależy od reszty i dźwigni. Obserwacja wpływowa nie musi być niedopasowana do modelu. Obserwacja o dużej reszcie nie musi być wpływowa.

Jak należy rozumieć linie na wykresie? To wartości standardized residuals. Widać w jakich zakresach należy się spodziewać danych residuuów, czyli można je automatycznie rozpoznać.

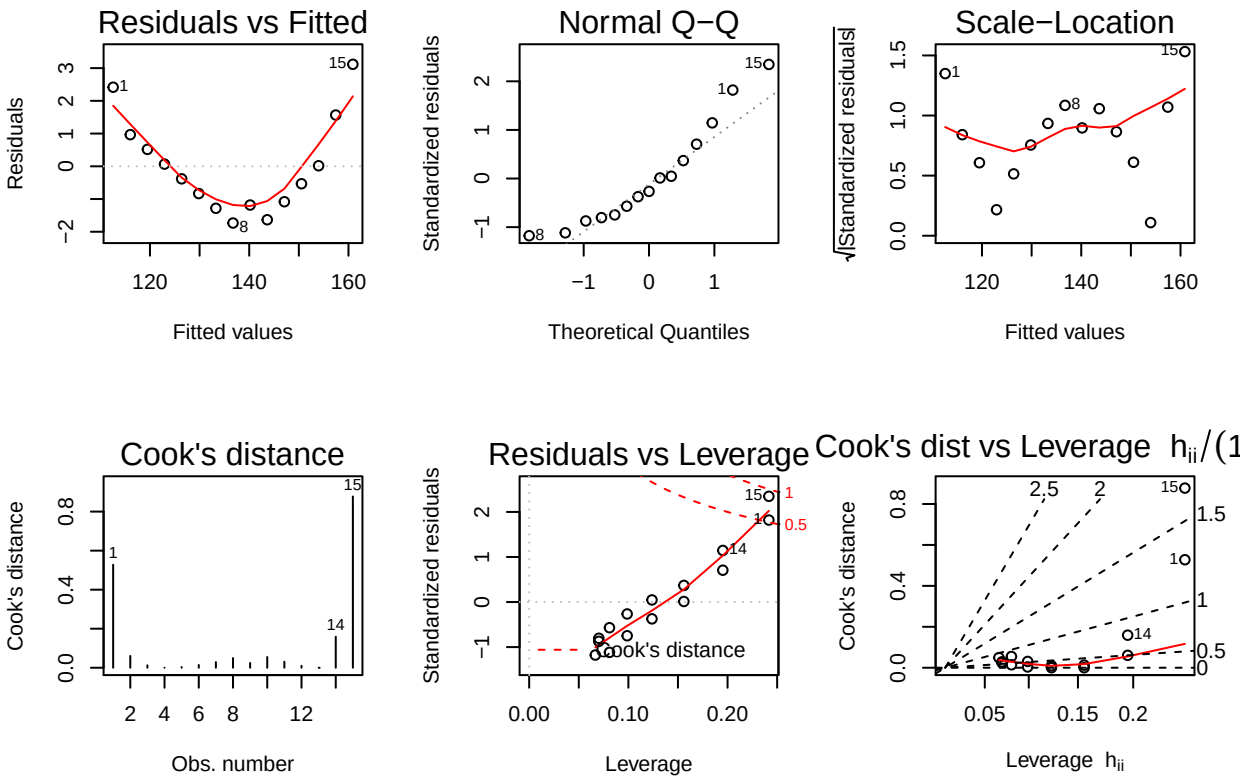
Dane o wzroście kobiet w USA

```
data(women) # Load a built-in data called 'women'
ggplot(women, aes(x=height, y=weight)) +
  geom_point()
```



Mogłoby się wydawać, że regresja powinna hulać, ale...

```
fit = lm(weight ~ height, women)
par(mfrow=c(2,3))
plot(fit, 1:6)
```




```
par(mfrow=c(1,1))
```

Testy diagnostyczne

Będziemy używać pakietu **lmtest**. Ma całkiem dobrą dokumentację

```
library(lmtest)
vignette("lmtest-intro", package="lmtest")
```

Jednorodność wariancji

Najbardziej krytyczne założenie w regresji liniowej jest to dotyczące jednorodności wariancji. Uwaga, nie jest największym problemem rozkład y. Nie musi być normalny. Przy dużej liczbie obserwacji residua mają rozkład normalny z CTG. Natomiast w przypadku niejednorodnej wariancji nic nie działa na naszą korzyść. W jednym z wykresów statystycznych patrzyliśmy na zależność pomiedzy wariancja a dopasowanymi wartościami.

Test Breuscha-Pagana

H0: wariancja jest jednorodnna H1: wariancja zależy od zmiennych objaśniających

```
bptest(weight~height, data = women)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  weight ~ height
## BP = 1.0088, df = 1, p-value = 0.3152
```

Test niezależności reszt

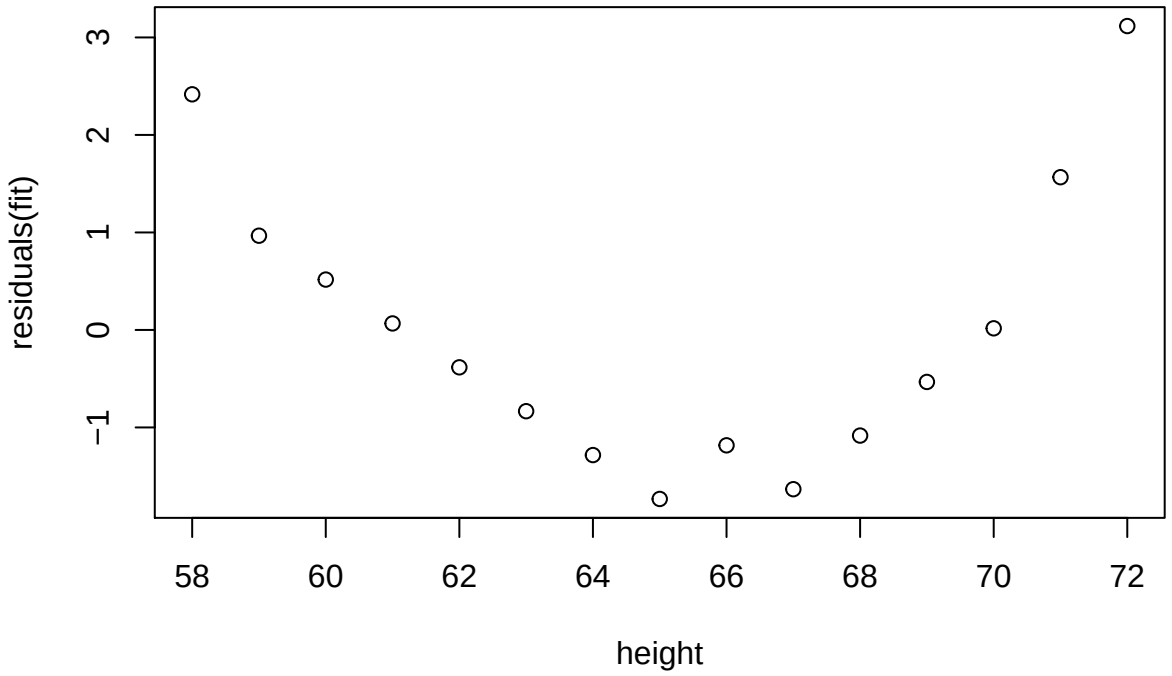
H0: niezależność reszt H1: autokorelacja rzędzu 1 ze względu na daną zmienną objaśniającą

```
dwtest(weight~height, order.by = ~height, data = women)
```

```
##
##  Durbin-Watson test
##
## data:  weight ~ height
## DW = 0.31538, p-value = 1.089e-07
## alternative hypothesis: true autocorrelation is greater than 0
```

Dlaczego to niedobrze? Ten wynik sugeruje, że jest w danych dodatkowa zależność, które nie zdołaliśmy ująć w modelu liniowym. Zobaczmy co się dzieje na wykresie

```
plot(residuals(fit)~height, data=women)
```



Test Harveya-Colliera

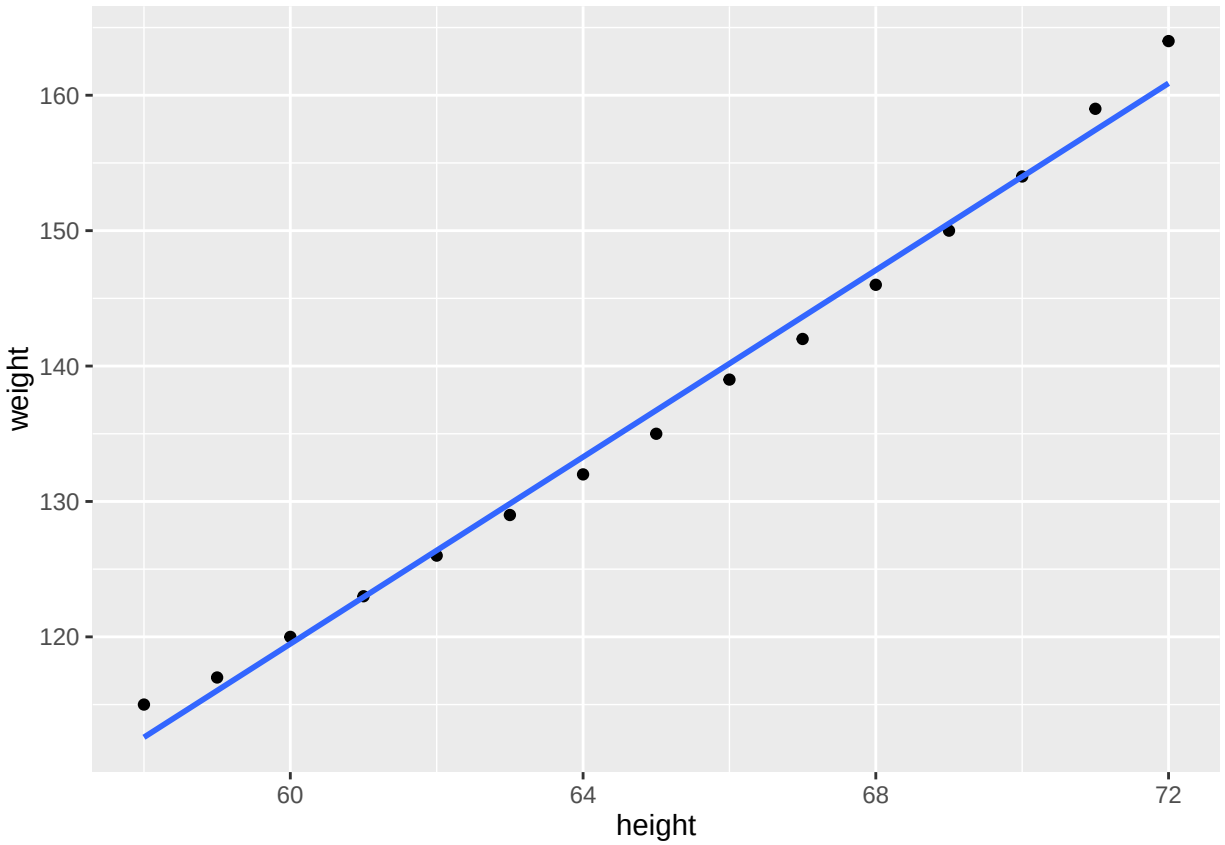
Czy zależność pomiędzy zmiennymi jest liniowa?

```
harvtest(weight~height, order.by = ~height, data = women)
```

```
##
##  Harvey-Collier test
##
## data:  weight ~ height
## HC = 3.7176, df = 12, p-value = 0.00294
```

Aha, bardzo podejrzane! Spójrzmy jeszcze raz na dane.

```
data(women) # Load a built-in data called 'women'
ggplot(women, aes(x=height, y=weight)) +
  geom_point() +
  stat_smooth(method = "lm", se = F, fullrange = T)
```



Alternatywne pakiety

Bardzo dobry zestaw narzędzi do diagnostyki regresji znajduje się w pakiecie **car**.