Obserwacje odstające

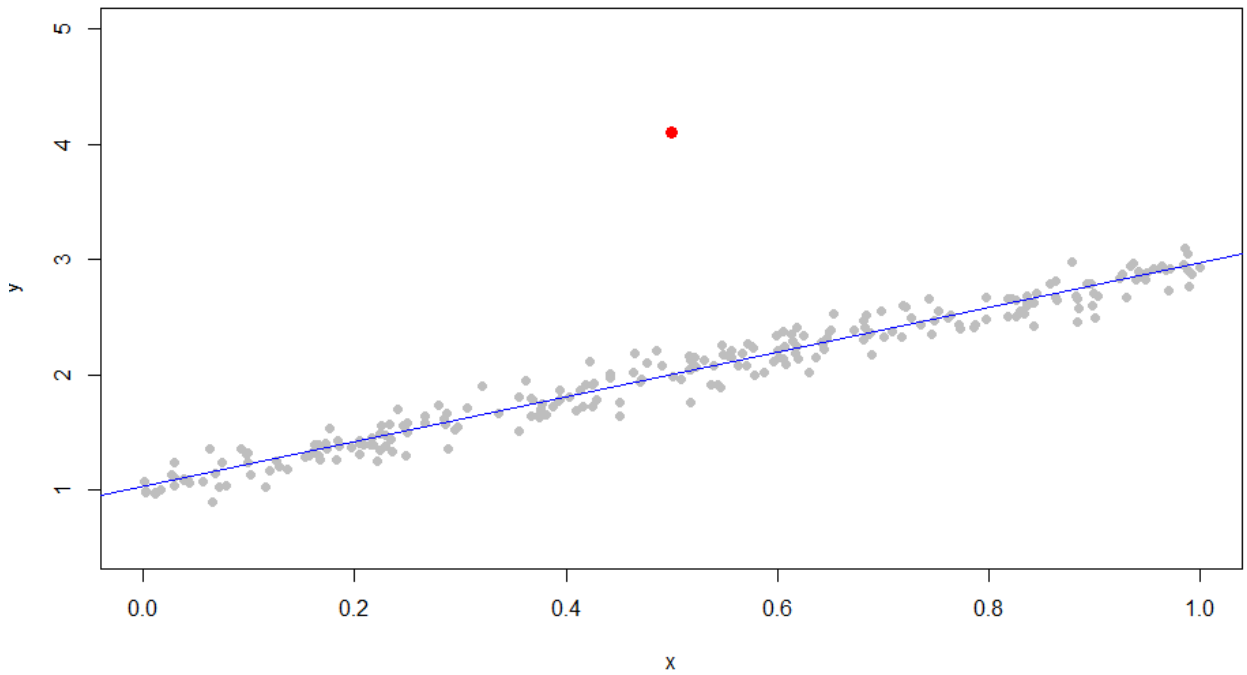
Wykład 1

Piotr Dybka

Ekonometria praktyczna

Czym są obserwacje odstające?

Obserwacja odstająca (nazywana także nietypową lub skrajną – *outlier*) to obserwacja znacząco różniąca się od pozostałych, która może zaburzać wyniki estymacji.



Przyczyny

Błędy w zbieraniu danych lub ich przetwarzaniu

1. Błędy przy wpisywaniu zmiennych (brakujący przecinek, zmiana znaku)
2. Specyfika kodowania zmiennych (np. kodowanie brakującej obserwacji jako 9999)
3. Złe zrozumienie pytania w badaniu ankietowym (np. dochód roczny zamiast miesięcznego)

Czynniki niezależne

1. Nietypowa sytuacja (np. zarobki prezesa spółki notowanej na GPW)
2. Wpływ jednorazowych zaburzeń (np. podwyższona zmienność cen w czasie kryzysu finansowego)

Skutki – wprowadzenie

W prezentacji koncentruję się na estymatorze KMNK, jednak opisane problemy dotyczą też innych estymatorów (np. regresji logistycznej).

W trakcie estymacji modelu możemy zauważyć, że dana obserwacja należy do jednej z grup:

- **Obserwacje odstające** – obserwacje, dla których obserwujemy b. duże reszty
- **Obserwacje wpływowe** – usunięcie obserwacji tego typu prowadzi do znaczących różnic w wartości oszacowanych parametrów nawet przy dużych próbach

Zapamiętaj: obserwacja odstająca $\neq$ obserwacja wpływowa

Skutki

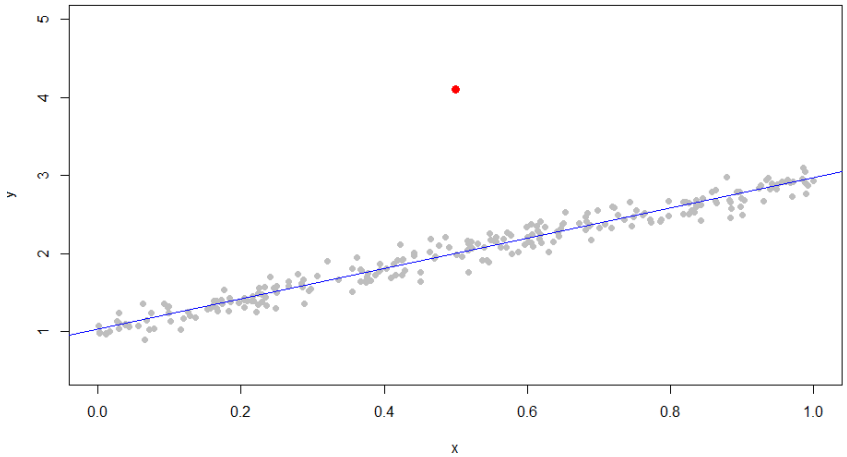
Załóżmy, że mamy do czynienia z obserwacją odstającą, ale nie wpływową:

```
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.28758 -0.08277 -0.00638  0.06456  2.09959

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.03039    0.02188    47.1   <2e-16 ***
x            1.94004    0.03767    51.5   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1728 on 248 DF
Multiple R-squared:  0.9145
Adjusted R-squared:  0.9142
F-statistic: 2653 on 1 and 248 DF, p-value: <2.2e-16
```



Skutki

Zacznijmy od wzoru na macierz wariancji-kowariancji:

$$D^2\left(\hat{\beta}\right) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

która jest estymowana jako

$$\widehat{D^2}\left(\hat{\beta}\right) = S^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

gdzie S^2 to **obserwowana** wariancja reszt z modelu opisana wzorem

$$S^2 = \frac{1}{N-K}\sum_{i=1}^{N}(e_i - \bar{e})^2 = \frac{1}{N-K}\sum_{i=1}^{N}e_i^2$$

N – liczba obserwacji, e_i – reszty z modelu, $\bar{e}$ – średnia reszt (w modelu ze stałą $\bar{e} = 0$), K – liczba parametrów (łącznie ze stałą).

Elementy na diagonali macierzy wariancji-kowariancji informują o wariancji oszacowanych parametrów. Jeżeli mamy **obserwacje odstające** (wysokie wartości e_i), to rośnie wariancja dla oszacowanych parametrów (**niska precyzja oszacowań**).

Skutki

Zatem usuńmy obserwację odstającą. Otrzymujemy wyniki regresji:

Regresja z obserwacją odstającą:

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.28758 -0.08277 -0.00638  0.06456  2.09959

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.03039    0.02188    47.1   <2e-16 ***
x            1.94004    0.03767    51.5   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regresja BEZ obserwacji odstającej:

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.279271 -0.074650  0.001844  0.071560  0.269638

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.02180    0.01392   73.42   <2e-16 ***
x[-125]      1.94035    0.02395   81.02   <2e-16 ***
```

[-125] oznacza usunięcie 125. obserwacji (outliera)

Dla zmiennej x otrzymaliśmy błędy standardowe o **36,4%** niższe!

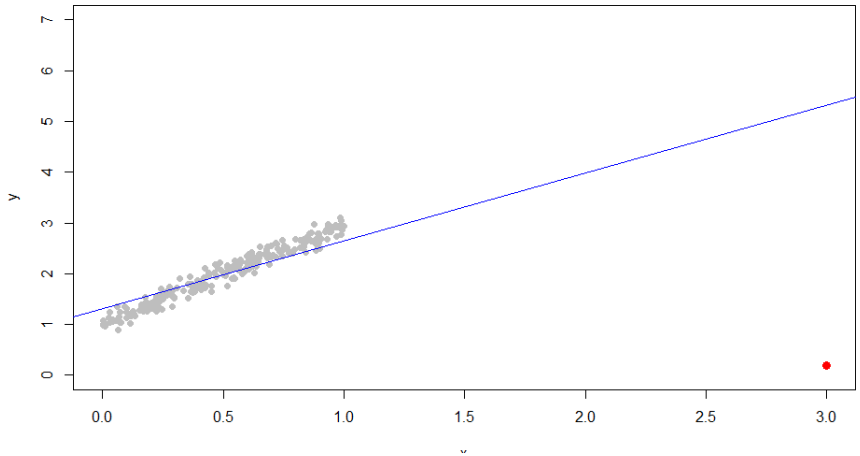
Skutki

Załóżmy, że mamy do czynienia z obserwacją **odstającą i wpływową**:

```
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1095 -0.1499  0.0416  0.1762  0.5004

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.30612    0.04509   28.97   <2e-16 ***
x            1.33445    0.07390   18.06   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

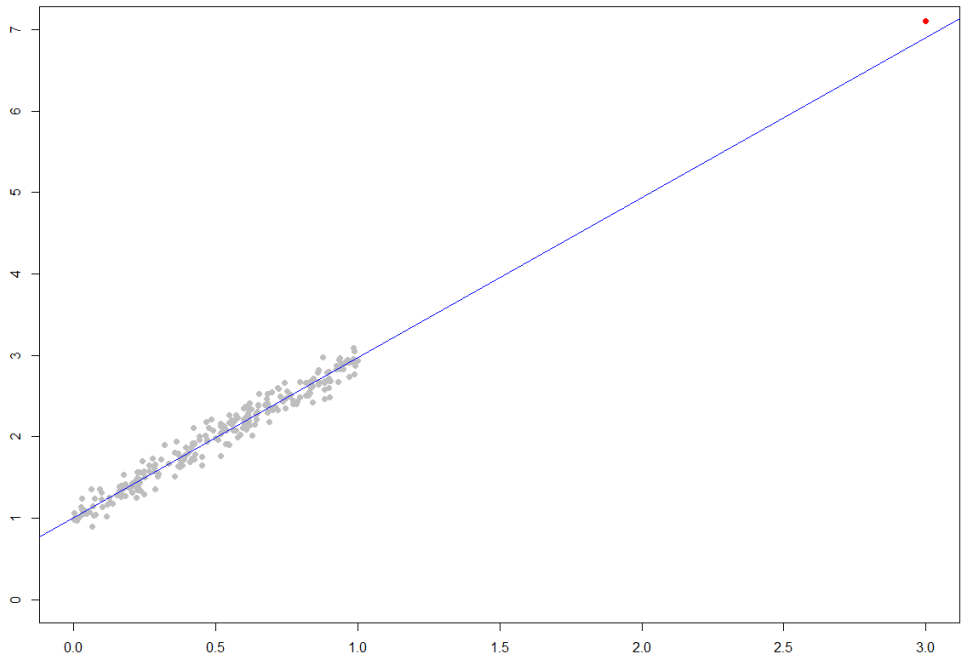


To są te same dane co w poprzednim przykładzie, z wyjątkiem 1 obserwacji.

Obserwacje wpływowe – czy zawsze usuwać?

Na poprzednim slajdzie mogliśmy zauważyć, że wpływowa obserwacja odstająca może prowadzić do błędnego wyniku (obciążenia estymatora). Czy zatem zawsze powinniśmy usuwać obserwacje wpływowe?

Jeżeli obserwacja odstająca powstała w wyniku błędu – zdecydowanie tak. Dlatego istnieją procedury ucinające obserwacje skrajne.

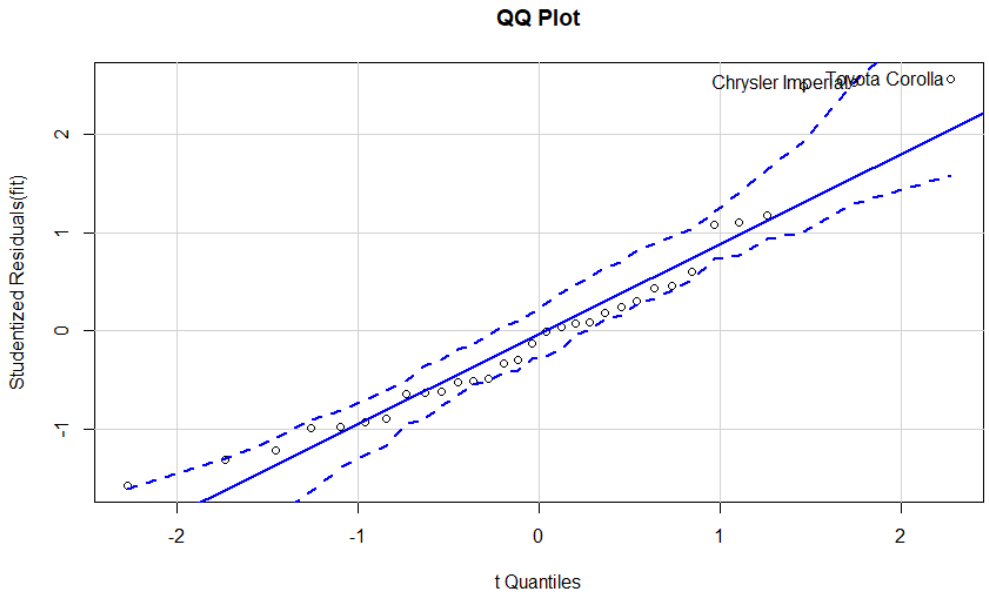


Natomiast jeżeli obserwacja wpływowa ma nietypowy poziom zarówno dla zmiennych objaśniających, jak i objaśnianej, to wówczas stanowi ona weryfikację naszej teorii poza „typową" próbą. Jeżeli nie mamy dużych reszt (nie jest to obs. odstająca), to wnosi ona istotne informacje i prawdopodobnie nie powstała w wyniku błędu.

Diagnostyka – wykres kwantylowy (QQ plot)

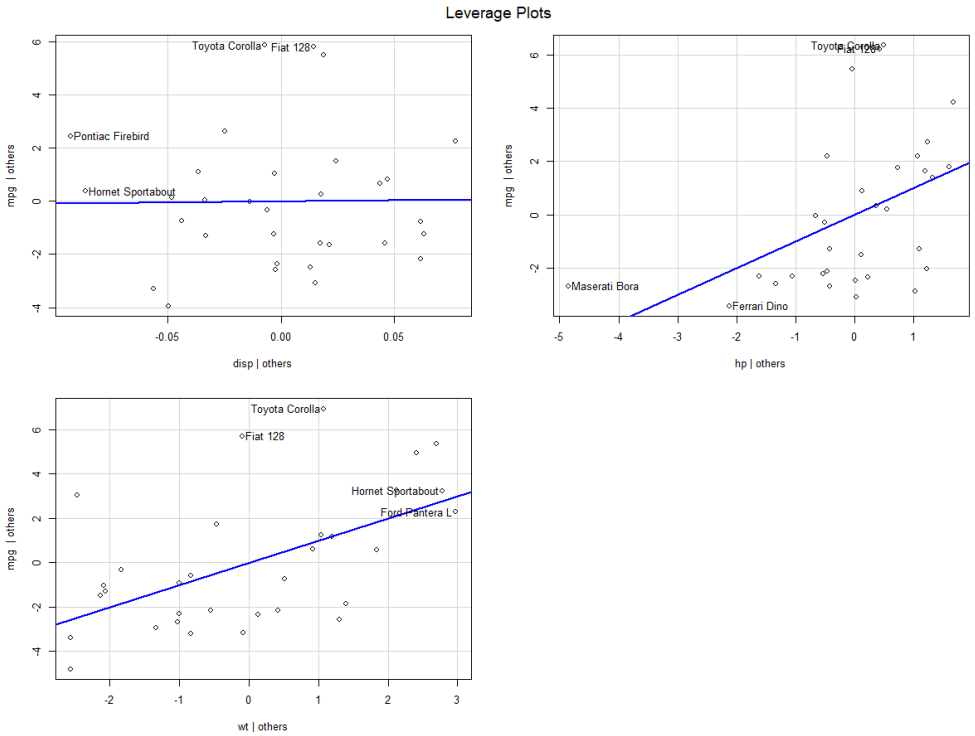
Diagnostyka (wykrywanie obserwacji odstających) opiera się na analizie reszt z modelu. Jednym z narzędzi jest tzw. wykres kwantylowy (ang. *QQ plot*).

1. Wykres ten pokazuje, jak dla poszczególnych kwantyli* rozkładu teoretycznego kształtują się kwantyle empiryczne reszt.
2. Wartości powinny układać się na linii prostej – jeżeli widoczne są istotne odchylenia na krańcach wykresu, to wskazuje to na występowanie obs. odstających.



*kwantyle mierzą pozycję danej wartości w próbie po jej uporządkowaniu. Przykładowo kwantyle rzędu 100, nazywane percentylami, informują o tym, ile % obserwacji w próbie ma wartość równą lub mniejszą.

Wykres dźwigni (*leverage / added-variable plot*)



- Oś Y pokazuje reszty z regresji zmiennej objaśnianej na **pozostałych** zmiennych objaśniających (bez badanej zmiennej)
- Oś X pokazuje reszty z regresji **badanej** zmiennej objaśniającej na pozostałych zmiennych objaśniających
- Nachylenie niebieskiej linii jest równe oszacowaniu parametru przy badanej zmiennej w pełnym modelu – im bardziej płaska, tym mniejsze znaczenie zmiennej (płaska = nieistotna statystycznie)

1. Wykres ten wskazuje, które obserwacje mają największy wpływ na oszacowanie.
2. Wartości na krańcach wykresu mają większe znaczenie dla oszacowania parametrów.

Pomiar wpływu obserwacji (dźwignia)

Do pomiaru wpływu obserwacji możemy wykorzystać miarę odstępstwa obserwacji x_i od $\bar{x}$ (dla regresji prostej):

$$h_i = \frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Ogólnie h_i jest i-tym elementem diagonali macierzy $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, czyli $h_i = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$.

Dla modelu o K parametrach:

$$\sum_{i=1}^{N} h_i = K, \qquad \text{przy czym dla każdego } i: \ \frac{1}{N} \leq h_i \leq 1$$

To oznacza, że przeciętna wartość dźwigni (h_i) wynosi K/N, a wartość powyżej $2K/N$ (dla małych prób $3K/N$) można uznać za potencjalnie wpływową.

Studentyzowane reszty (*studentized residuals*)

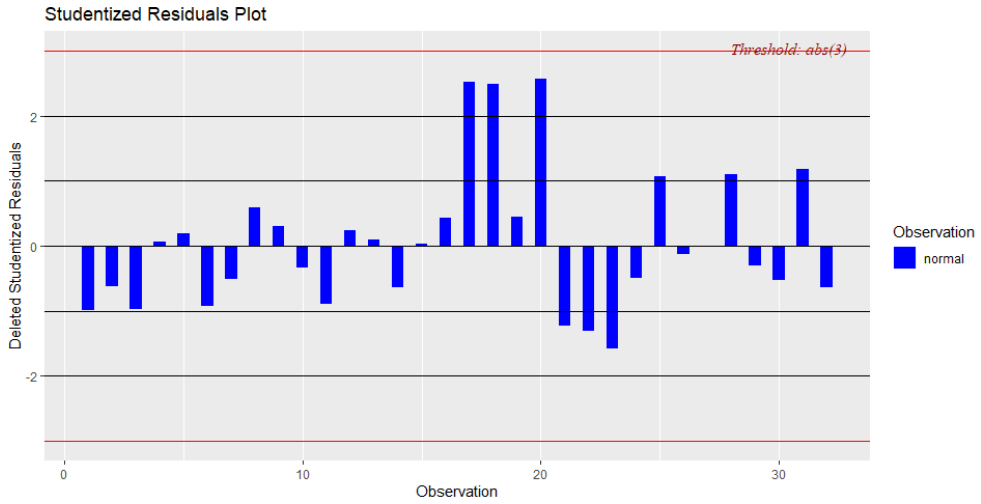
Standaryzowane (wewnętrznie studentyzowane) reszty to reszty podzielone przez oszacowanie odchylenia standardowego reszty

$$r_i = \frac{e_i}{SE_{e_i}}, \qquad SE_{e_i} = S\sqrt{1 - h_i}$$

gdzie

$$h_i = \frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Studentyzowane reszty powinny mieścić się w przedziale $[-2; 2]$.



Jeżeli zamiast S wstawimy $S_{(i)}$ (oszacowane bez i-tej obserwacji), otrzymamy resztę studentyzowaną usuniętą (*deleted studentized residual*, RStudent) – to właśnie ona jest przedstawiona na powyższym wykresie.

Test Bonferroniego (*Bonferroni outlier test*)

Reguła $[-2; 2]$ to heurystyka. Formalnie możemy **przetestować** hipotezę:

H_0 : i-ta obserwacja nie jest odstająca $\qquad$ H_1 : i-ta obserwacja jest odstająca

Statystyką testową jest reszta studentyzowana usunięta, która przy H_0 ma rozkład

$$t_i = \frac{e_i}{S_{(i)}\sqrt{1-h_i}} \sim t(N-K-1)$$

- **Problem**: nie testujemy jednej, z góry wybranej obserwacji – szukamy *największej* spośród N reszt. Przy $\alpha = 0,05$ i $N = 32$ oczekiwalibyśmy ok. 1,6 „odstającej" obserwacji czysto przypadkowo (to jest właśnie *swamping*).
- **Poprawka Bonferroniego**: p-value dla największego $|t_i|$ mnożymy przez liczbę przeprowadzonych testów:

$$p_{Bonf} = \min\left(N \cdot 2P\left(t_{N-K-1} > |t_i|\right),\ 1\right)$$

- W R: `car::outlierTest(fit)`

Test Bonferroniego – przykład

Dla naszego modelu `lm(mpg~disp+hp+wt, data=mtcars)`:

```
> outlierTest(fit)
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
               rstudent unadjusted p-value Bonferroni p
Toyota Corolla 2.564129           0.016225      0.51919
```

Interpretacja:

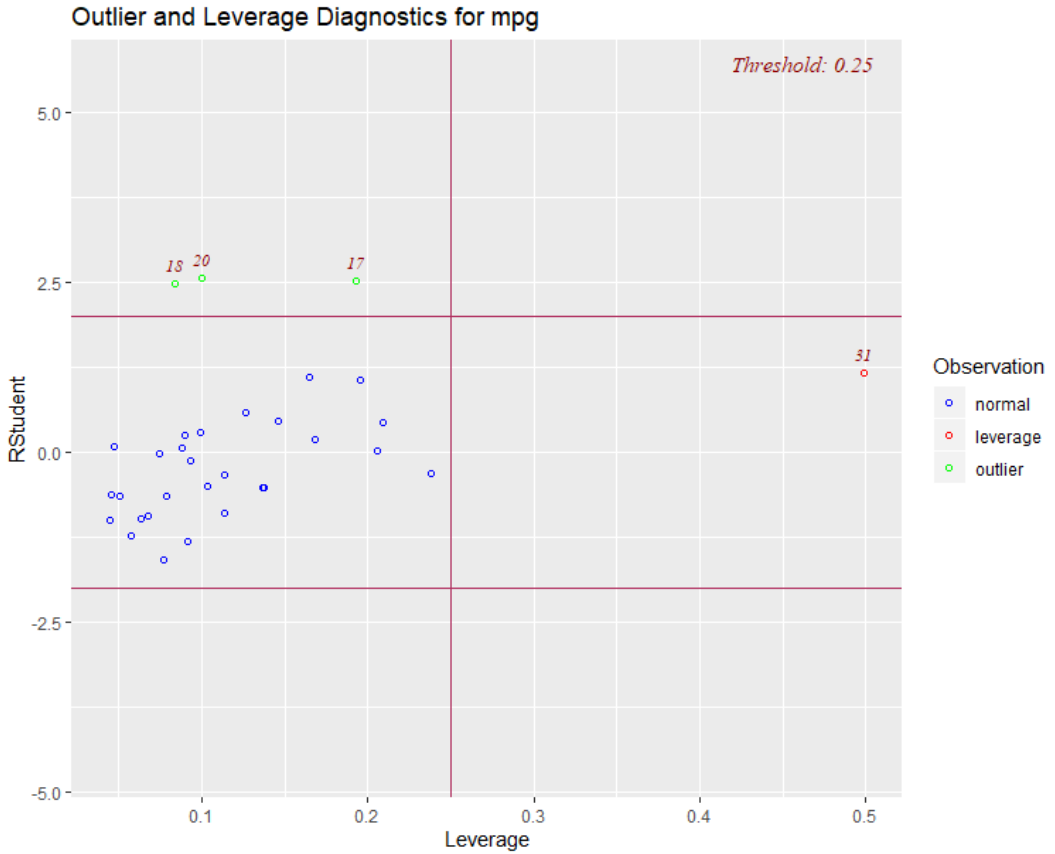
- Nieskorygowane $p = 0,016$ sugerowałoby, że Toyota Corolla jest obserwacją odstającą.
- Po uwzględnieniu, że przejrzeliśmy wszystkie $N = 32$ obserwacje, $p_{Bonf} = 0,52$ – **brak podstaw do odrzucenia** H_0.
- Wartość krytyczna $|t|$ wynosi tu $3,52$, a nie 2 – wykresy z poprzednich slajdów wskazywały obserwacje 17, 18 i 20 jako podejrzane, ale formalny test ich nie potwierdza.

Uwaga: test wskazuje tylko *jedną*, najbardziej skrajną obserwację – przy kilku obserwacjach odstających może zawieść (*masking*). W takiej sytuacji stosujemy go iteracyjnie.

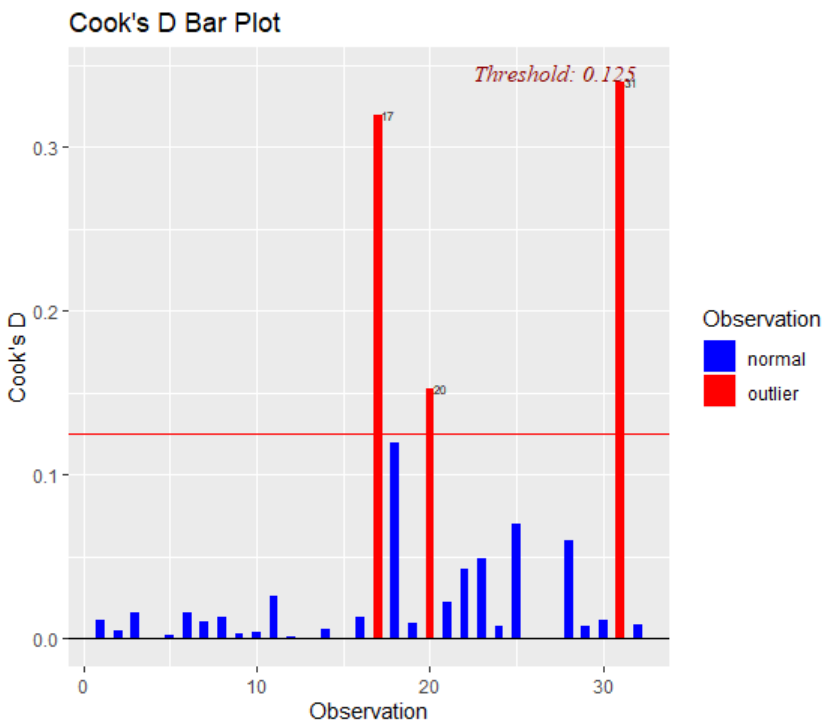
Studentyzowane reszty na tle dźwigni

Możemy jeszcze pokazać wykres studentyzowanych reszt na tle dźwigni – w ten sposób możemy podzielić obserwacje na:

- zwykłe,
- odstające niewpływowe,
- odstające i wpływowe,
- wpływowe nieodstające.



Odległość Cooka (Cook's D, *Cook's distance*)



Odległość Cooka mierzy zmianę w wartości współczynników regresji, gdy pomijamy i-tą obserwację:

$$D_i = \frac{e_i^2}{K \cdot MSE} \left(\frac{h_i}{(1-h_i)^2} \right)$$

gdzie MSE = średni błąd kwadratowy (*mean square error*).

- $\frac{e_i^2}{K \cdot MSE}$ to miara odstawania obserwacji
- $\frac{h_i}{(1-h_i)^2}$ to miara dźwigni (wpływu)

Wartość progową można ustalić jako:

1. $4/N$ (wykr. po lewej)
2. $3\bar{D}$ (trzykrotność średniej odległości Cooka)
3. 1.0
4. $F(0.5; K, N-K)$ (wartości powyżej 50. percentyla rozkł. F)

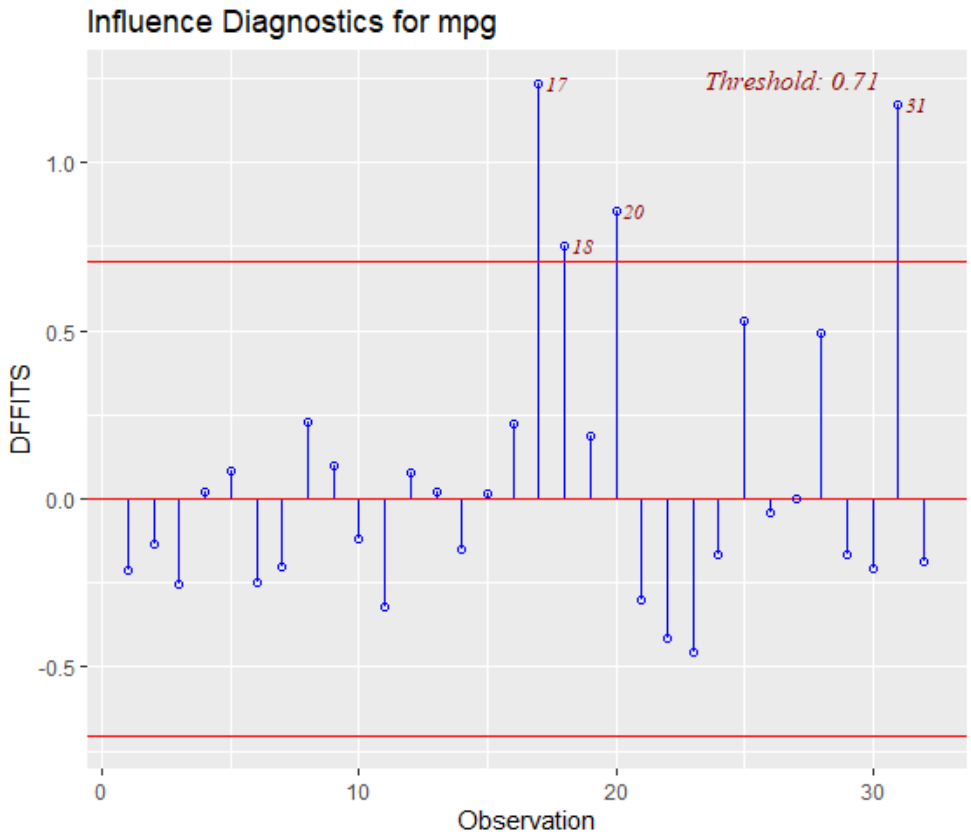
DFFITS (*Difference in Fits*)

DFFITS sprawdza, jak i-ta obserwacja wpływa na wartość teoretyczną wyliczoną z oszacowanego równania regresji:

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{S_{(i)}\sqrt{h_i}} = \frac{e_i}{S_{(i)}} \cdot \frac{\sqrt{h_i}}{1 - h_i}$$

Indeks (i) oznacza wyłączenie i-tej **obserwacji**.

Jako wartość progową przyjmuje się $2\sqrt{\frac{K}{N}}$.



DFBETA(S) (*Difference in Betas*)

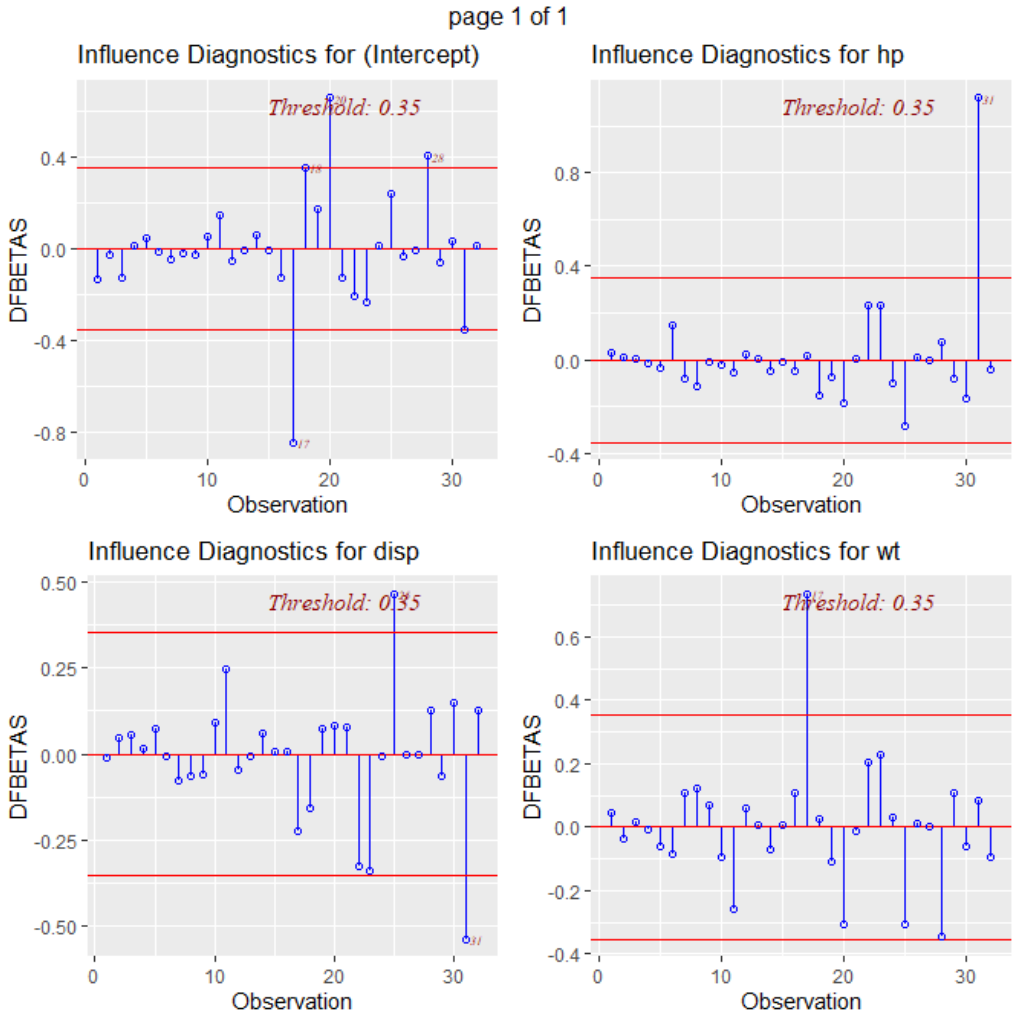
DFBETA analizuje zmiany we współczynnikach regresji (β). Dla każdego $\hat{\beta}_j$ (j – zmiennych) możemy policzyć zmianę wartości $\hat{\beta}_j$, gdy i-ta obserwacja zostanie pominięta:

$$DFBETA_{j(i)} = \hat{\beta}_j - \hat{\beta}_{j(i)}$$

Po standaryzacji otrzymujemy DFBETAS:

$$DFBETAS_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{MSE_{(i)} \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}_{jj}}}$$

Wartość progowa: $\frac{2}{\sqrt{N}}$



Obserwacja odstająca czy błędna specyfikacja?

Zanim usuniesz obserwacje odstające, sprawdź, czy **nie mają one czegoś wspólnego**. W naszym modelu wszystkie trzy podejrzane obserwacje mają reszty **dodatnie** – model je niedoszacowuje:

obserwacja	wt	e_i	RStudent
Toyota Corolla	1,835	$+5,86$	2,56
Fiat 128	2,200	$+5,79$	2,50
Chrysler Imperial	5,345	$+5,49$	2,53

- Zakres zmiennej *wt* to [1,51; 5,42] – dwie obserwacje leżą przy *dolnym*, a jedna przy *górnym* krańcu rozkładu masy.
- Niedoszacowanie na **obu krańcach** zmiennej objaśniającej to typowy objaw zależności **wypukłej**, której człon liniowy nie jest w stanie odwzorować.
- Sprawdzenie: zastąpienie *wt* przez $\ln(wt)$ podnosi R^2 z 0,827 do 0,867, a RStudent dla Chryslera spada z 2,53 do 2,06.

Wniosek: usunięcie tych obserwacji poprawiłoby dopasowanie, ale *ukryłoby* błąd specyfikacji zamiast go naprawić. Wzorzec wśród obserwacji odstających jest sygnałem o postaci modelu, a nie zadaniem z „czyszczenia" danych.

Metody postępowania

1. Usunięcie błędnych informacji
2. Winsoryzacja (modyfikacja obserwacji)
3. Least trimmed squares (LTS)
4. Least absolute deviations (LAD)
5. Regresja kwantylowa

1. Usunięcie błędnych informacji

- W przypadku, gdy mamy pewność, że dana obserwacja jest błędnie zakodowana, powstała w wyniku złego zrozumienia pytania lub powstała w wyjątkowych okolicznościach, których nie chcemy uwzględniać w modelu, to powinniśmy usunąć taką **obserwację** (niezależnie od jej wpływu; nawet jeżeli nie jest wpływowa, to będzie zmniejszała precyzję oszacowań).
- Gdy jednak nie mamy podstaw do stwierdzenia, że jest to obserwacja błędna, to usunięcie jej może wiązać się z utratą cennych informacji (w szczególności gdy jest to tylko obserwacja wpływowa, a nie odstająca).
- W przypadku identyfikacji obserwacji odstających mogą występować dwa zjawiska:
  - *Masking* – przy większej liczbie obserwacji odstających metody mogą ich nie wykrywać (są „ukryte"); w tej sytuacji prezentowane metody mogą zacząć wykrywać obs. odstające dopiero po usunięciu kilku obserwacji tego typu.
  - *Swamping* – błędne zakwalifikowanie obserwacji jako odstająca.

2. Winsoryzacja (modyfikacja obserwacji)

- W przypadku, gdy nie chcemy usuwać obserwacji, możemy wykorzystać tzw. winsoryzację (*winsorizing*).
- Zamiast usuwać obserwacje odstające, możemy podmienić ich wartość na wartość z 1. lub 99. percentyla rozkładu dla danej zmiennej (w zależności z której strony obserwujemy odstawanie zmiennej).
- Idea polega na tym, że skoro mamy obserwację, dla której poziom jednej zmiennej wydaje się źle zmierzony, to szukamy racjonalnej wartości dla tej zmiennej, tak aby nie tracić obserwacji.

3. Least trimmed squares

- Usuwanie obserwacji odstających można też wbudować w procedurę estymacji – LTS (*Least Trimmed Squares*).
- Idea LTS polega na tym, że minimalizujemy sumę *tylko h* najmniejszych kwadratów reszt (pozostałe, największe reszty są odrzucane):

$$S_{LTS}(\tilde{\beta}) = \sum_{i=1}^{h} \left(e^2\right)_{i:N}, \qquad \left(e^2\right)_{1:N} \leq \cdots \leq \left(e^2\right)_{N:N}$$

- W praktyce oznacza to: szacujemy model, porządkujemy obserwacje według wartości absolutnych reszt i odrzucamy zadany % obserwacji z najwyższymi resztami; procedura jest powtarzana, aż do znalezienia podpróby dającej najmniejszą sumę kwadratów reszt.
- Pakiet `robustbase` w R; polecenie `ltsReg()`, parametr $\alpha = h/N$ określa, jaka część obserwacji ma pozostać w próbie.

4. Least absolute deviations

- Innym podejściem jest zastosowanie estymatora LAD (*Least Absolute Deviations*), gdzie minimalizujemy następującą funkcję celu:

$$S_{LAD}(\tilde{\beta}) = \sum_{i=1}^{N} \left| y_i - \boldsymbol{x}_i' \tilde{\beta} \right|$$

- Czyli zamiast kwadratów reszt (jak w przypadku KMNK) mamy wartości absolutne.
- Estymator LAD ma za zadanie dopasować parametry $\tilde{\beta}$ do mediany, a nie średniej rozkładu – mediana jest miarą bardziej odporną na występowanie obserwacji odstających.

5. Regresja kwantylowa

- Regresja kwantylowa jest uogólnieniem LAD, gdzie minimalizujemy następującą funkcję celu:

$$S(\tilde{\boldsymbol{\beta}}_Q) = \sum_{i:\, y_i \geq \boldsymbol{x}_i' \tilde{\boldsymbol{\beta}}_Q} Q \left| y_i - \boldsymbol{x}_i' \tilde{\boldsymbol{\beta}}_Q \right| + \sum_{i:\, y_i < \boldsymbol{x}_i' \tilde{\boldsymbol{\beta}}_Q} (1 - Q) \left| y_i - \boldsymbol{x}_i' \tilde{\boldsymbol{\beta}}_Q \right|$$

- Parametr Q informuje o tym, dla którego kwantyla rozkładu zmiennej objaśnianej szukamy wartości $\boldsymbol{\beta}$.
- Reszty dodatnie (powyżej dopasowanej wartości) są ważone za pomocą Q, a reszty ujemne za pomocą $(1 - Q)$.
- Przykładowo dla $Q = 0.7$ błędy szacunku powyżej dopasowanej wartości będą miały duże znaczenie, a te poniżej – mniejsze; dlatego regresja zostanie dopasowana tak, aby odzwierciedlać zależność dla 70. percentyla rozkładu badanej zmiennej.
- Dla $Q = 0.5$ dopasowujemy się do mediany i otrzymujemy estymator LAD.

5. Regresja kwantylowa

- W ten sposób możemy zobaczyć, czy na krańcach rozkładu zmiennej badana zależność ulega zmianom.
- W szczególności duże zmiany dla krańcowych kwantyli mogą wskazywać na występowanie obs. odstających.