Wstęp do uczenia maszynowego

Regresja liniowa (2)

Ewa Szczurek + BW (modyfikacje)

bartek@mimuw.edu.pl
Instytut Informatyki
Uniwersytet Warszawski

8 kwietnia 2024

UNIWERSYTET
WARSZAWSKI

Model regresji

$$f(X) = \beta_0 + \sum_{i=1}^{p} X_i \beta_i,$$

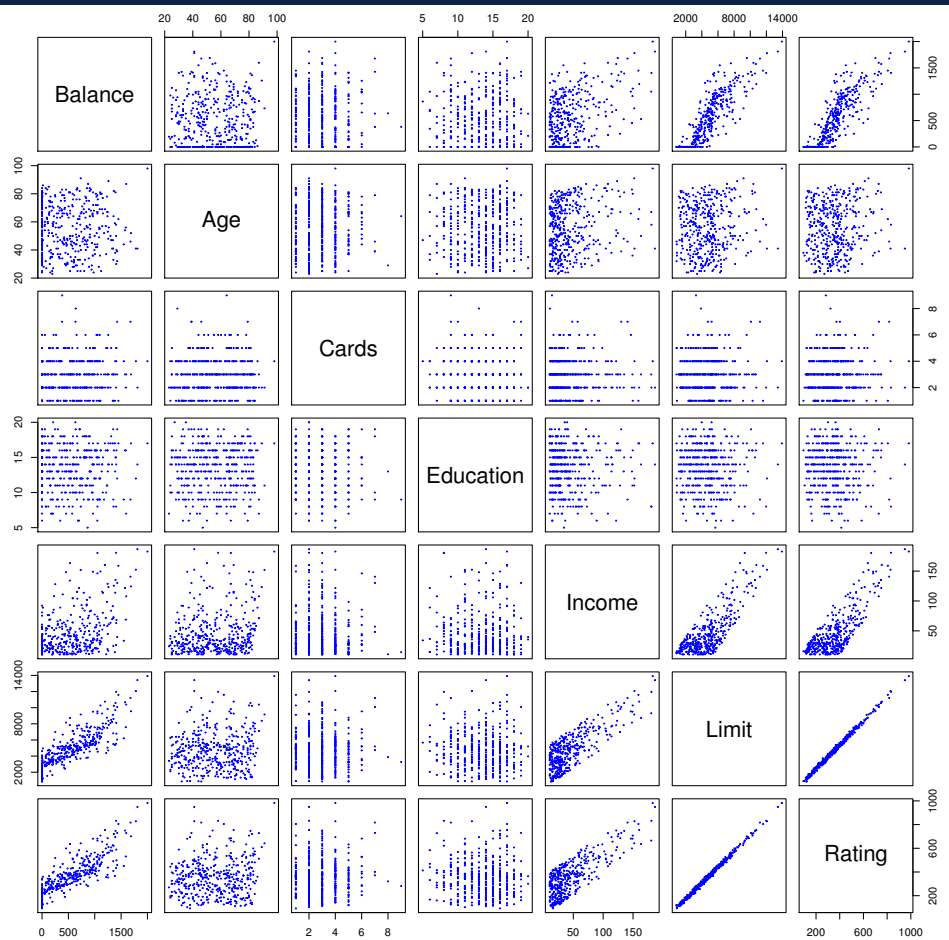
gdzie X_i predyktory, a β_i - nieznane współczynniki.

Zmienne X_i mogą być postaci

- zmiennych ilościowych
- transformacji zmiennych ilościowych, np log
- funkcji wielomianowych zmiennych ilościowych, np $X_2 = X_1^2$, $X_3 = X_1^3$
- zmiennych kodujących (ang. *dummy encoding*) predyktory jakościowe (nominalne), np kolor oczu
- zmiennych odpowiadających interakcjom pomiędzy predyktorami, np $X_3 = X_1 \cdot X_2$

We wszystkich przypadkach f jest liniową funkcją parametrów.

Przykład: Model zadłużenia karty kredytowej (*balance*)



Tu są predyktory ilościowe. Można rozważać również predyktory jakościowe: 'gender' (płeć), 'status' (małżeński), 'ethnicity' (Afroamerykanin, Azjata, rasa kaukaska (biały)).

Dwupoziomowe predyktory jakościowe

$$x_i = \begin{cases} 1, & \text{jeśli } i\text{-ta osoba jest kobietą} \\ 0, & \text{jeśli } i\text{-ta osoba jest mężczyzną} \end{cases}$$

To prowadzi do modelu regresji:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i, & \text{jeśli } i\text{-ta osoba jest kobietą} \\ \beta_0 + \varepsilon_i, & \text{jeśli } i\text{-ta osoba jest mężczyzną} \end{cases}$$

β_0 można interpretować jako średni poziom kredytu wśród mężczyzn, $\beta_0 + \beta_1$ można interpretować jako średni poziom kredytu wśród kobiet, a β_1 jako średnia różnica pomiędzy poziomami kredytów kobiet i mężczyzn.

Możemy użyć innych wartości do oznaczenia poziomów dla zmiennej x_i. Wówczas estymowane parametry $\hat{\beta}_0$ i $\hat{\beta}_1$ zmienią się, ale predykcje $\hat{y}_i$ pozostaną bez zmian.

Różnica w średnim zadłużeniu na karcie nie zależy istotnie od płci

Parametr	Estymacja	Std. błąd	t-statystyka	p-wartość
$\hat{\beta}_0$	509.80	33.13	15.389	< 0.0001
Płeć [kobieta] $\hat{\beta}_1$	19.73	46.05	0.429	0.6690

- Średnie zadłużenie na karcie kredytowej u mężczyzn wynosi 509.80
- Średnie zadłużenie u kobiet jest o 19.73 większe i wynosi 509.80+19.73= 529.53
- Ale różnica ta nie jest statystycznie istotna (duża p-wartość).

Wielopoziomowe predyktory jakościowe

Rozważmy przykład predyktora 'ethinicity' przyjmującego 3 możliwe wartości. Musimy wprowadzić dwie zmienne pomocnicze:

$$x_{i1} = \begin{cases} 1, \text{ jeśli } i\text{-ta osoba jest Azjatą} \\ 0, \text{ w przeciwnym przypadku} \end{cases}$$

$$x_{i2} = \begin{cases} 1, \text{ jeśli } i\text{-ta osoba jest rasy kaukaskiej} \\ 0, \text{ w przeciwnym przypadku} \end{cases}$$

Wówczas model regresji wygląda następująco (rasa kaukaska= rasa K; Afroamerykanin=Afroam):

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i, \text{ jesli } i\text{-ta osoba jest Azjatą} \\ \beta_0 + \beta_2 + \varepsilon_i, \text{ jesli } i\text{-ta osoba jest rasy K} \\ \beta_0 + \varepsilon_i, \text{ jesli } i\text{-ta osoba jest Afroam} \end{cases}$$

Różnica w średnim zadłużeniu na karcie nie zależy istotnie od rasy

Parametr	Estymacja	Std. błąd	t-statystyka	p-wartość
$\hat{\beta}_0$	531.00	46.32	11.464	< 0.0001
Rasa [Azjata] $\hat{\beta}_1$	-18.69	65.02	-0.287	0.7740
Rasa [kaukaska] $\hat{\beta}_2$	-12.50	56.68	-0.221	0.8260

- Średnie zadłużenie na karcie kredytowej u Afroamerykanina wynosi 531.00
- Średnie zadłużenie u Azjaty jest o 18.69 mniejsze.
- Średnie zadłużenie u człowieka rasy kaukaskiej jest o 12.50 mniejsze niż u Afroamarykanina.
- Ale różnice te nie są statystycznie istotne (duże p-wartości).

Uwzględnianie interakcji (synergii) pomiędzy predyktorami

Model regresji liniowej z dwoma predyktorami:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

Model regresji liniowej z interakcją predyktorów:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon.$$

Uwaga: regresja tu jest nadal liniowa.

Przykład: Synergia wydatków na reklamę

$$sales = \beta_0 + \beta_1 \times (TV) + \beta_2 \times (radio) + \beta_3 \times (TV \times radio) + \varepsilon.$$

Parametr	Estymacja	Std. błąd	t-statystyka	p-wartość
$\hat{\beta}_0$	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
$radio$	0.0289	0.009	3.24	0.0014
$TV \times radio$	0.0011	0.0000	20.73	< 0.0001

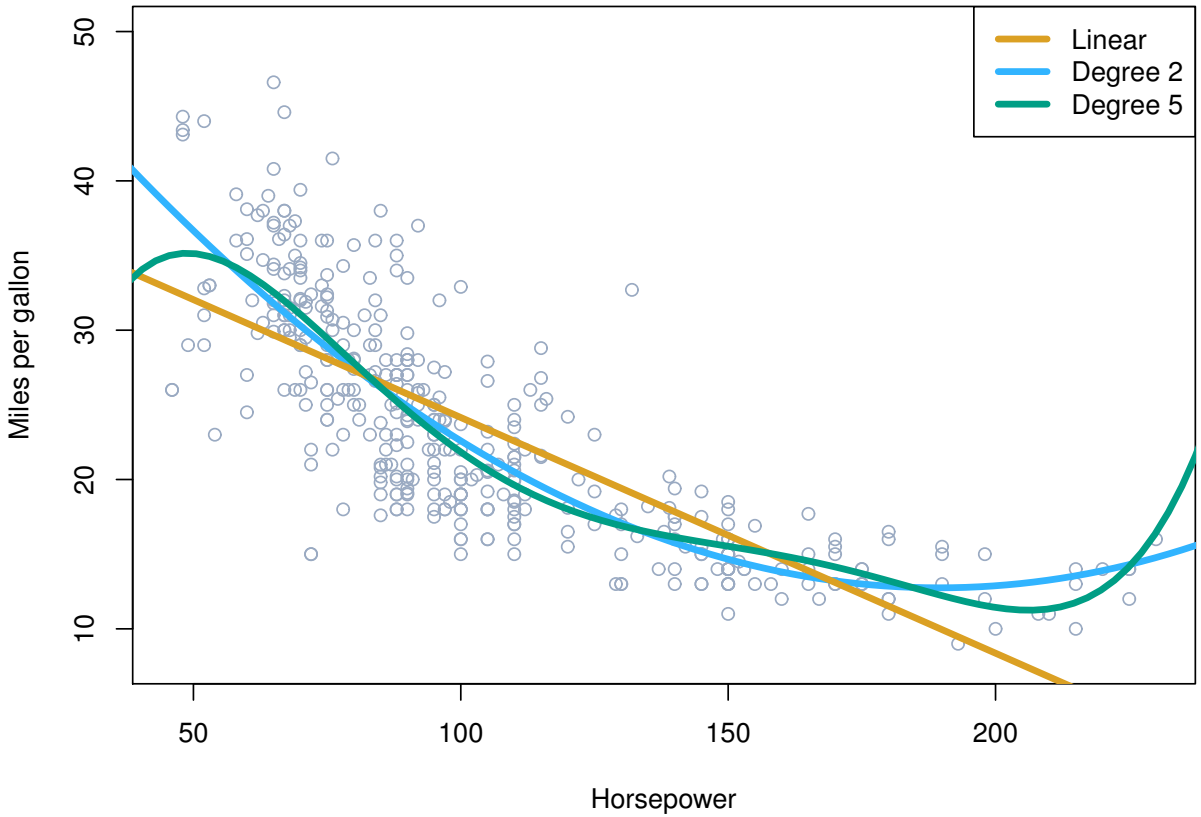
Zasada hierarchiczności: jeśli model zawiera interakcję predyktorów X_i oraz X_j (czyli term X_iX_j), to musi też zawierać predyktory X_i oraz X_j, bez względu na wielkość p-wartości związanej z tymi predyktorami.

Model z interakcją predyktorów lepiej tłumaczy zmienność w danych reklamowych

- p-wartość dla termu interakcyjnego jest bardzo mała, co sugeruje, że rola tego termu w modelu jest bardzo istotna.
- Statystyka R^2 dla modelu bez interakcji wynosi 89.7%.
- Statystyka R^2 dla modelu z interakcją 'TV' oraz 'radio' jest równa 96.8%.
- Większy model ma co do zasady większe R^2, tutaj ten wzrost jest bardzo duży.
- Można wnioskować, że model uwzględniający interakcję pomiedzy predyktorami dużo lepiej tłumaczy zmienność danych reklamowych.

Regresja wielomianowa (dane 'Auto')

Zależność zużycia paliwa od liczby koni mechanicznych dla samochodów



- pomarańczowy: prosta regresja liniowa
- niebieski: regresja z horsepower^2
- zielony: model ze wszystkimi wielomianami horsepower aż do potęgi 5

Regresja do wielomianu kwadratowego

$$mpg = \beta_0 + \beta_1 \times horsepower + \beta_2 \times horsepower^2 + \varepsilon.$$

- **Uwaga:** to jest nadal liniowa regresja, czyli metody estymacji parametrów są te same.
- **Kwadratowe dopasowanie poprawia jakość modelu**: współczynnik ma małą p-wartość, a R^2 mocno wzrasta (przy dopasowaniu liniowym $R^2 = 0.606$, a dla dopasownia kwadratowego $R^2 = 0.688$).

Parametr	Estymacja	Std. błąd	t-statystyka	p-wartość
$\hat{\beta}_0$	56.9001	1.8004	31.6	< 0.0001
horsepower	-0.4662	0.0311	-15.0	< 0.0001
$horsepower^2$	0.0012	0.0001	10.1	< 0.0001

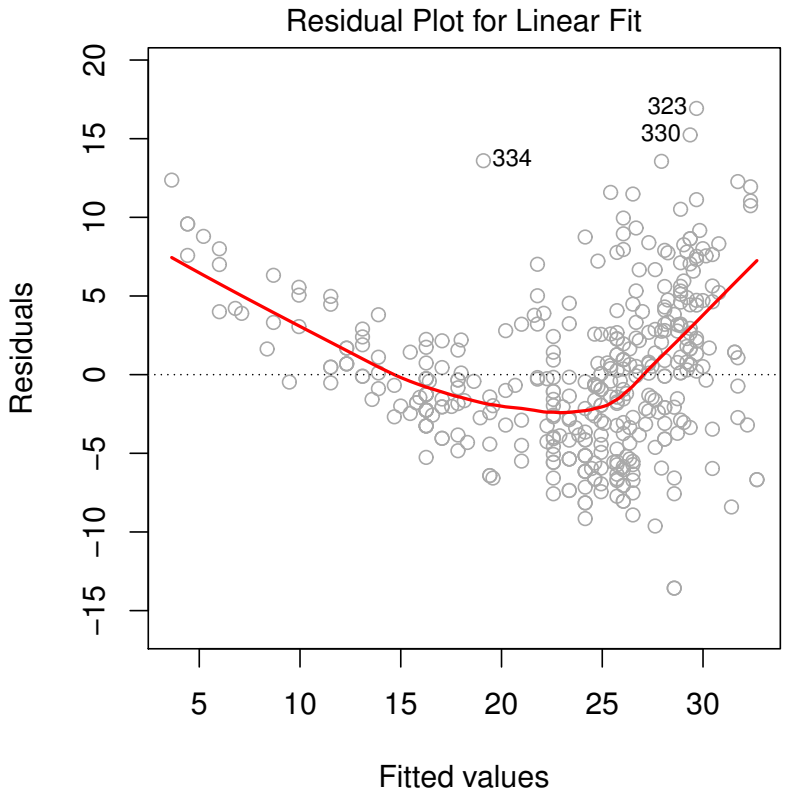
Potencjalne problemy z liniową regresją

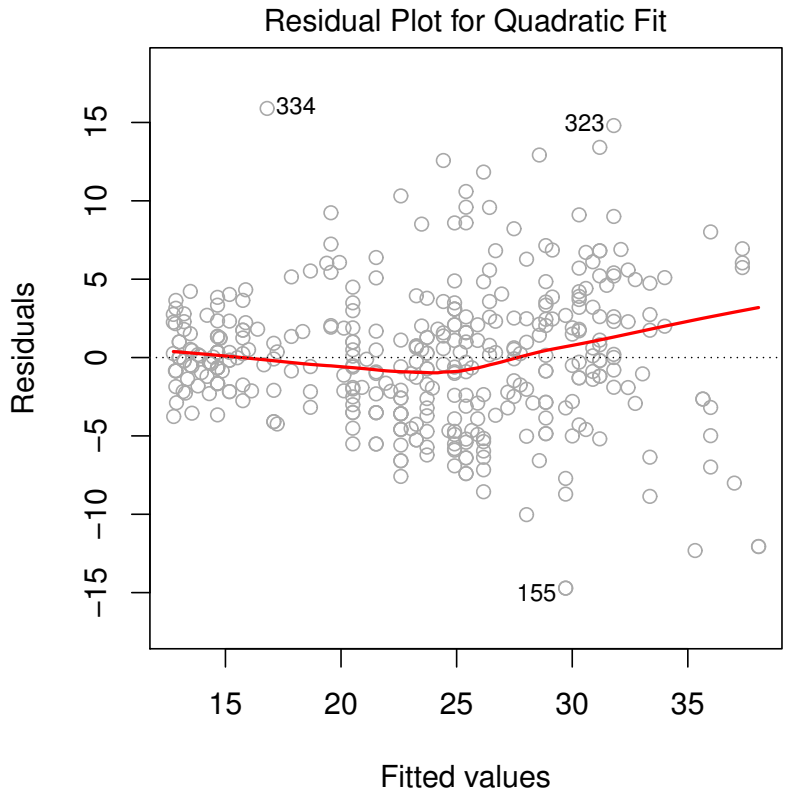
- Nieliniowość związku zmiennych objaśniających i objaśnianych.
- Skorelowane wartości reszt.
- Heteroskedastyczność (zmienność wariancji).
- Obserwacje odstające (outliery).
- Obserwacje o wysokiej dźwigni.
- Współliniowość zmiennych objaśniających

(1) Nieliniowość w danych – rezydualne wykresy

Wykres wartości resztowych to wykres różnic $y_i - \hat{y}_i$ względem przewidzianych wartości $\hat{y}_i$. Takie wykresy mogą sugerować nieliniowe zależności w danych.

Dane 'Auto':





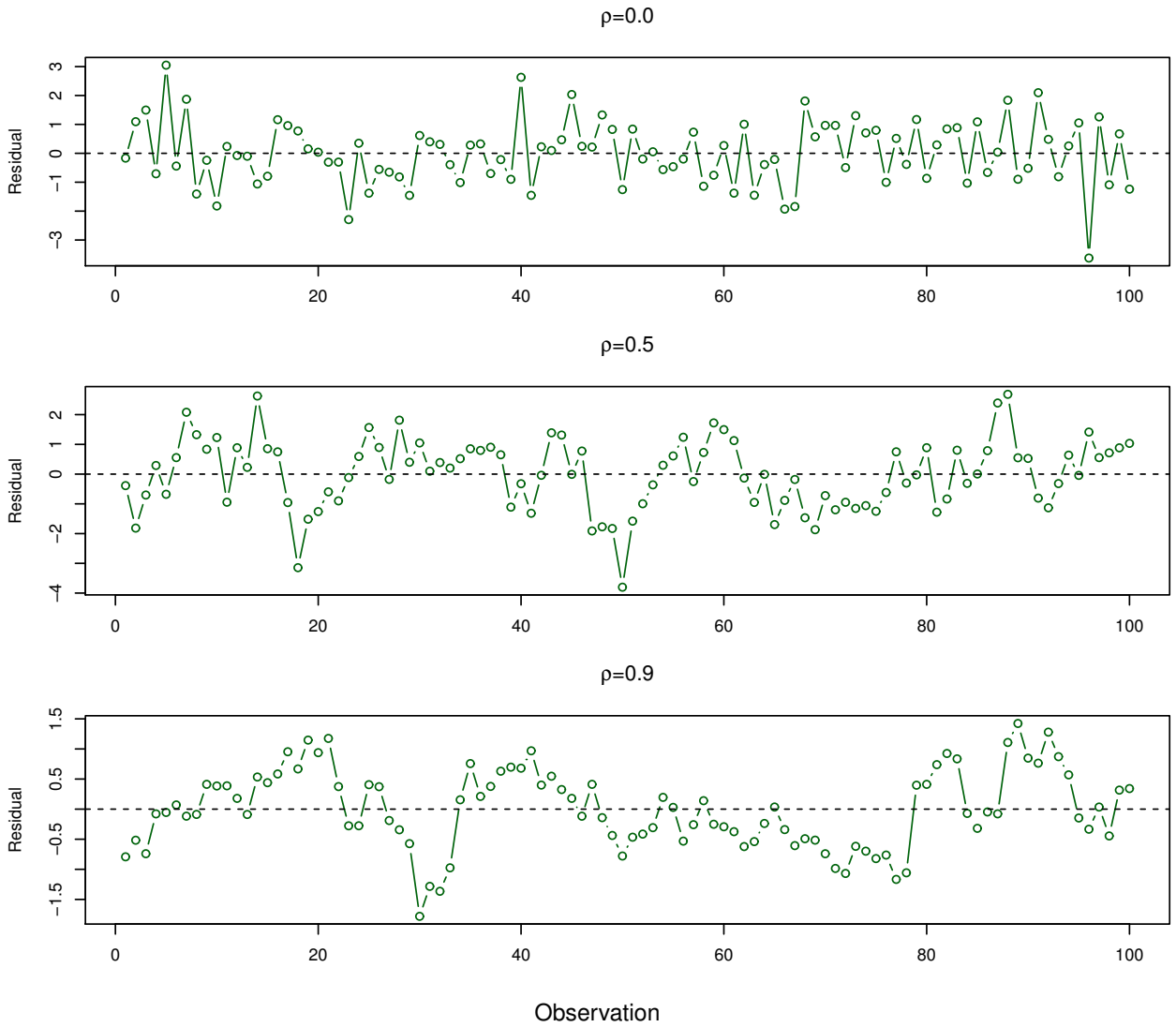
(2) Korelacja pomiędzy resztami obserwacji

- Dla $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \in \mathbb{R}^n$ wektora błędów zakładamy, że
  - Macierz wariancji-kowariancji $\mathrm{Var}[\epsilon] = \sigma^2 I_n$, $\sigma^2 < \infty$
  - czyli błędy ϵ_i pochodzące z różnych obserwacji są nieskorelowane
  - czyli wartość ϵ_i nie daje informacji o wartości ϵ_{i+1}
- Tego założenia wymaga w szczególności, aby błąd standardowy dla estymatora $\hat{\beta}_i$ miał postać $SE(\hat{\beta}_i) = \sqrt{v_i}\hat{\sigma}$, gdzie v_i to i-ty element na diagonali macierzy $(\mathrm{X}^T\mathrm{X})^{-1}$.
- Przy błędach skorelowanych estymator błędu standardowego będzie go zaniżał, przez co
  - przedziały ufności będą zawężone
  - dany przedział np na poziomie 95% ufności będzie miał mniejsze prawdopodobieństwo zawierania prawdziwej wartości parametru niż 0.95
  - p-wartości dla hipotez $H_1 : \beta_i \neq 0$ będą zaniżone

(2) Korelacja pomiędzy resztami obserwacji

- Korelacja pomiędzy błędami pojawia się często w przypadku danych pochodzących z szeregów czasowych – błędy pochodzące z sąsiednich punktów czasowych są często dodatnio skorelowane.
- Takie korelacje widać na wykresach wartości resztowych dla kolejnych obserwacji.
- W przypadku wystąpienia korelacji wartości leżące w sąsiednich punktach czasowych są położone blisko siebie.

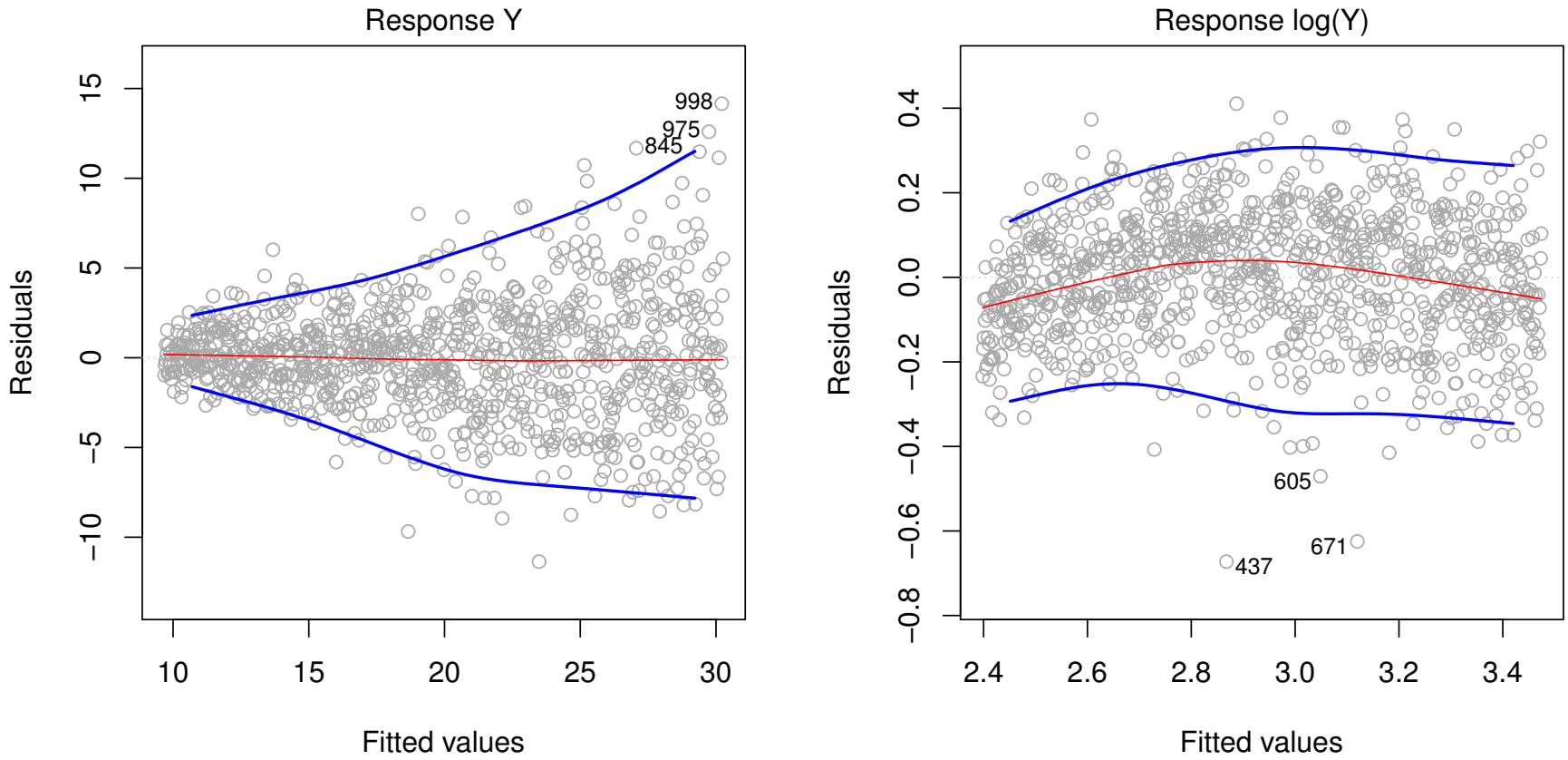
Dane symulowane – różne poziomy korelacji dla błędów pomiędzy sąsiednimi punktami czasowymi



(3) Zmienność wariancji reszt

- Zakłada się *homoskedastyczność*, czyli założenie, że wariancja reszt dla każdej obserwacji jest taka sama: $Var(\varepsilon_i) = \sigma^2$.
- Zdarza się, że wariancja reszt rośnie wraz ze wzrostem obserwowanych wartości.
- Zmienność wariancji (*heteroskedastyczność*) można zidentyfikować na wykresie wartości resztowych – kształt *lejka* tego wykresu sugeruje zmienność wariancji.

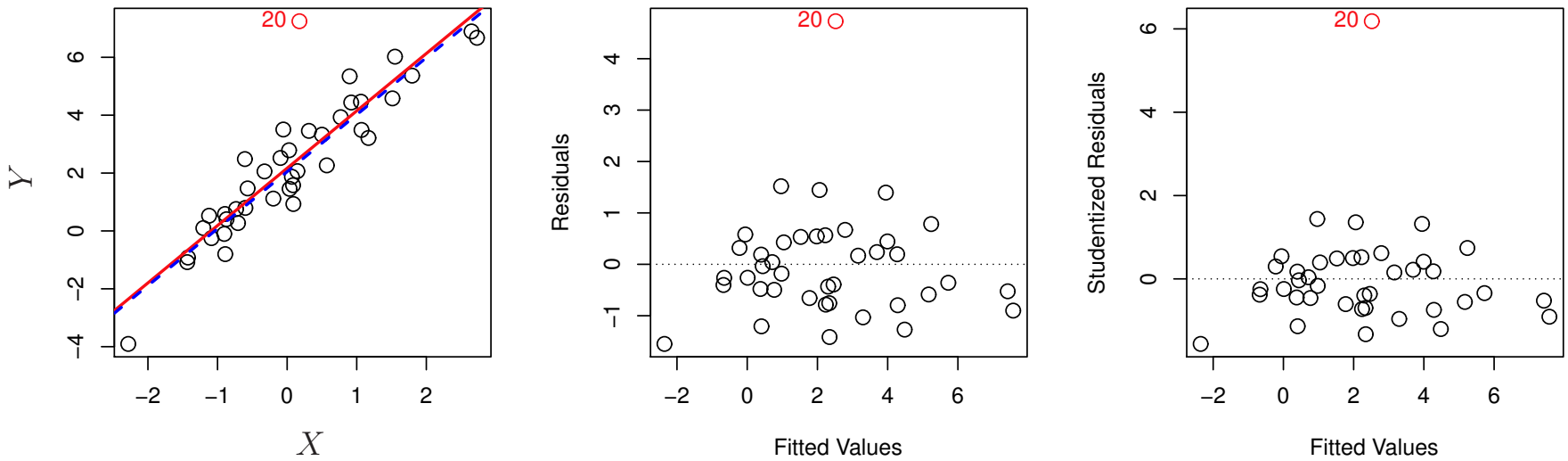
Identyfikacja zmienności wariancji termów błędu



Zmienność wariancji można poprawić poprzez transformację zmiennej objaśnianej Y wklęsłą funcją taką jak $\log Y$ (prawy panel) lub $\sqrt{Y}$.

(4) Obserwacje odstające (outliers)

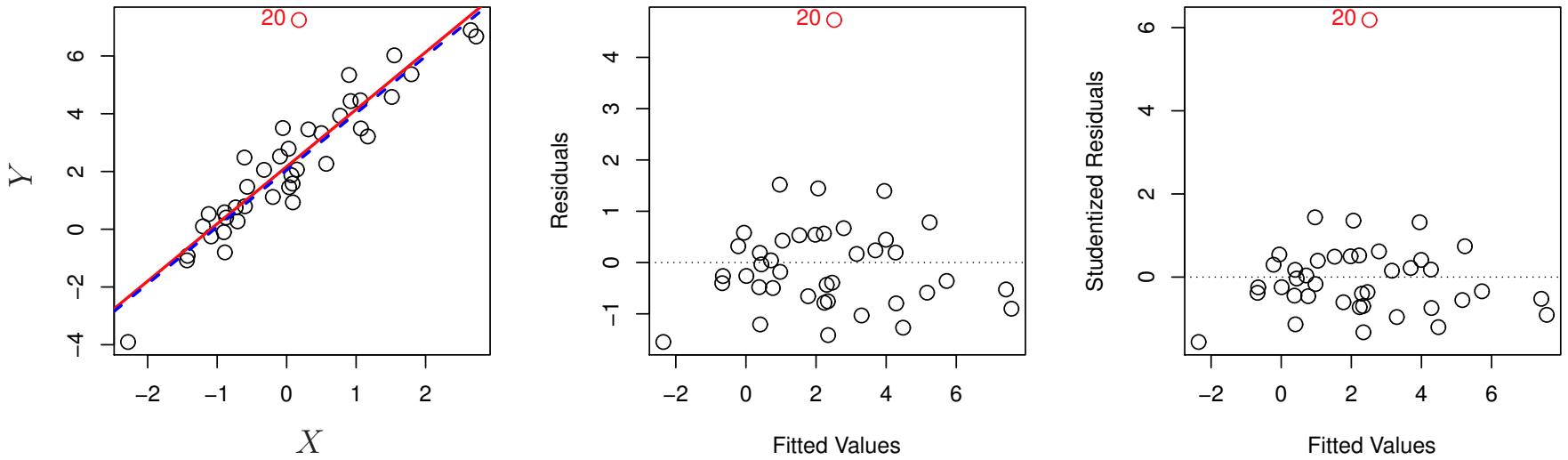
Obserwacje odstające: o nietypowej wartości objaśnianej (Y) przy typowej wartości predyktora (X).



Usunięcie obserwacji odstającej

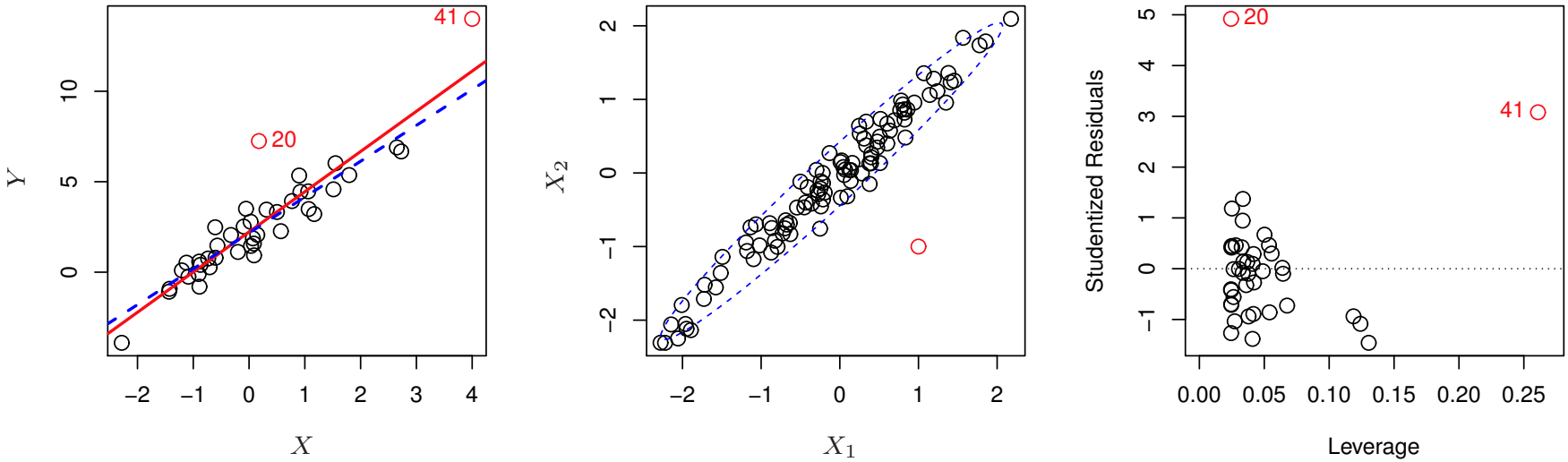
- tylko nieznacznie zmienia wynik regresji liniowej (przerywana linia)
- może znacznie zmienić miarę jakości dopasowania: tu RSE zmienia się z 1.09 do 0.77, a R^2 z 0.805 do 0.892.

Odkrywanie obserwacji odstających



- Przy pomocy wykresów wartości resztowych – środkowy panel, lub
- Przy pomocy reszt *studentyzowanych* (wyjaśnienie w dalszej części wykładu) – prawy panel.
- Przyjmuje się, że dla obserwacji odstających reszty studentyzowane odchylają się od 0 o więcej niż 3

(5) Obserwacje wysokiej dźwigni (high leverage)



- Obserwacje o wysokiej *dźwigni*: obserwacje o nietypowej wartości predyktora.
- Ich usunięcie powoduje istotne zmiany w przebiegu prostej regresji (lewy panel, przerywana linia).
- Odkrycie obserwacji o wysokiej dźwigni na wykresie resztowym jest łatwe dla prostej regresji liniowej, ale może być nieoczywiste przy większej liczbie predyktorów (panel środkowy).

Obliczanie *dźwigni* h_{ii} dla danej obserwacji x_i

- Dźwignia h_{ii} dla obserwacji x_i to pochodna cząstkowa estymowanej i-tej wartości zmiennej objaśnianej po właściwej i-tej wartości zmiennej objaśnianej

$$h_{ii} = \frac{d\hat{y}_i}{dy_i}$$

- Mamy $\hat{\mathrm{y}} = \mathrm{Hy}$, gdzie $\mathrm{H} = \mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T$
- Stąd h_{ii} to i-ty element na diagonali macierzy daszkowej H
- Zachodzi $0 \leq h_{ii} \leq 1$

Własności dźwigni

- Zauważmy, że macierz H jest *idempotentna*: $\mathrm{H}^2 = \mathrm{H}$

$$\mathrm{H}^2 = \mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T = \mathrm{XI}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T = \mathrm{H}$$

- i symetryczna, czyli, $h_{i,j} = h_{j,i}$
- Stąd, przyrównując i-te elementy na diagonali H oraz H^2

$$h_{ii} = h_{ii}^2 + \sum_{i \neq j} h_{i,j}^2 \geq 0$$

a zatem

$$h_{ii} \geq h_{ii}^2 \Rightarrow h_{ii} \leq 1.$$

Własności dźwigni

- Niech $rank(\mathrm{X}) = k$
- Z przemienności operatora śladu $tr(AB) = tr(BA)$ mamy

$$\sum_i h_{ii} = tr(\mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T) = tr((\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{X}) = tr(I_k) = k$$

Stąd średnia dźwignia $\bar{h}$ wynosi

$$\bar{h} = k/n$$

- Zatem
  - w modelu bez wyrazu wolnego $\bar{h} = p/n$ dla p- liczba predyktorów
  - gdy w modelu mamy wyraz wolny, $\bar{h} = (p+1)/n$

Własności dźwigni

- Przy $\text{Var}[\epsilon_i] = \sigma^2$ dla reszty $e_i = y_i - \hat{y}_i$ mamy

$$\text{Var}[e_i] = (1 - h_{ii})\sigma^2$$

(to wynika z faktu że $\hat{y} = Hy$ i $\text{Var}[y] = I\sigma^2$)

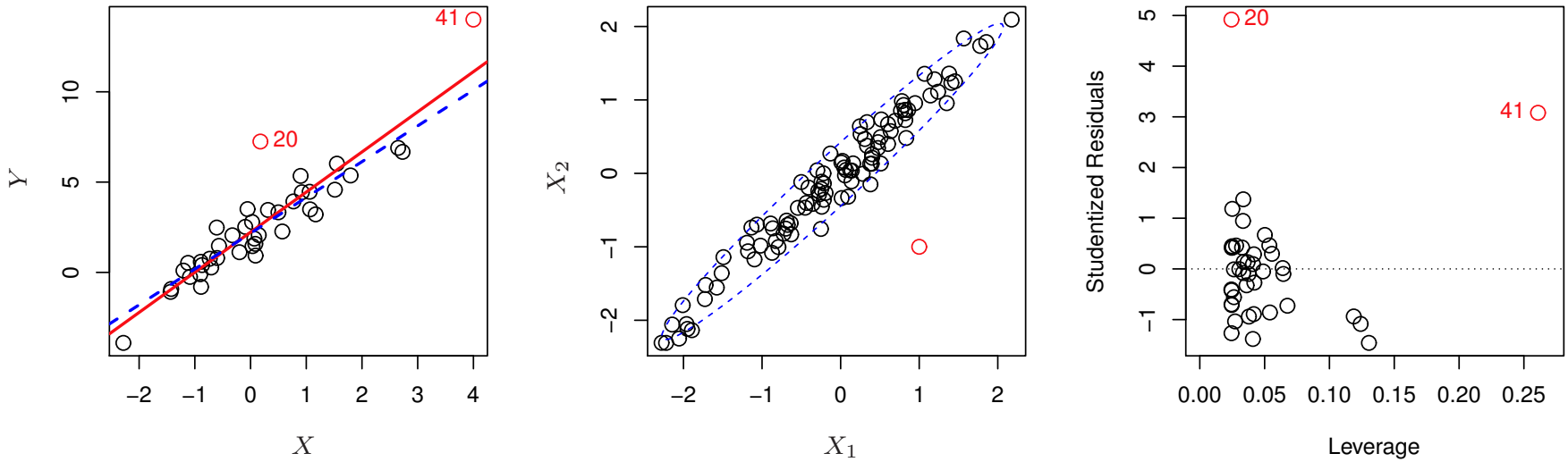
- Stąd dla obserwacji o dużej dźwigni reszty mają małą wariancję

Reszta studentyzowana

Reszta unormowana (podzielona przez estymator swojego odchylenia standardowego)

$$t_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

Odkrywanie obserwacji dźwigniowych

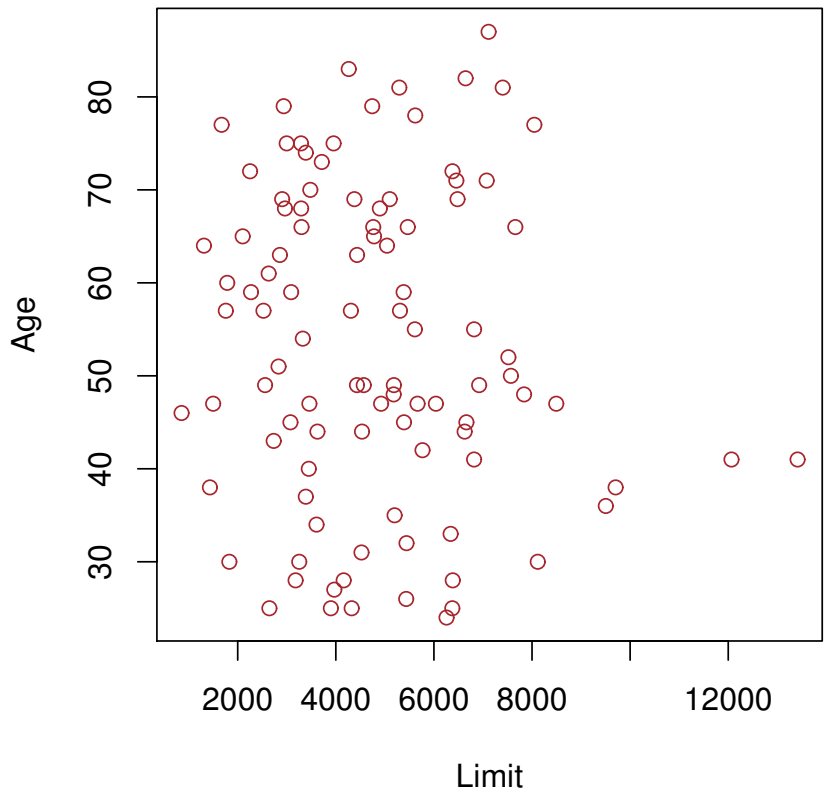


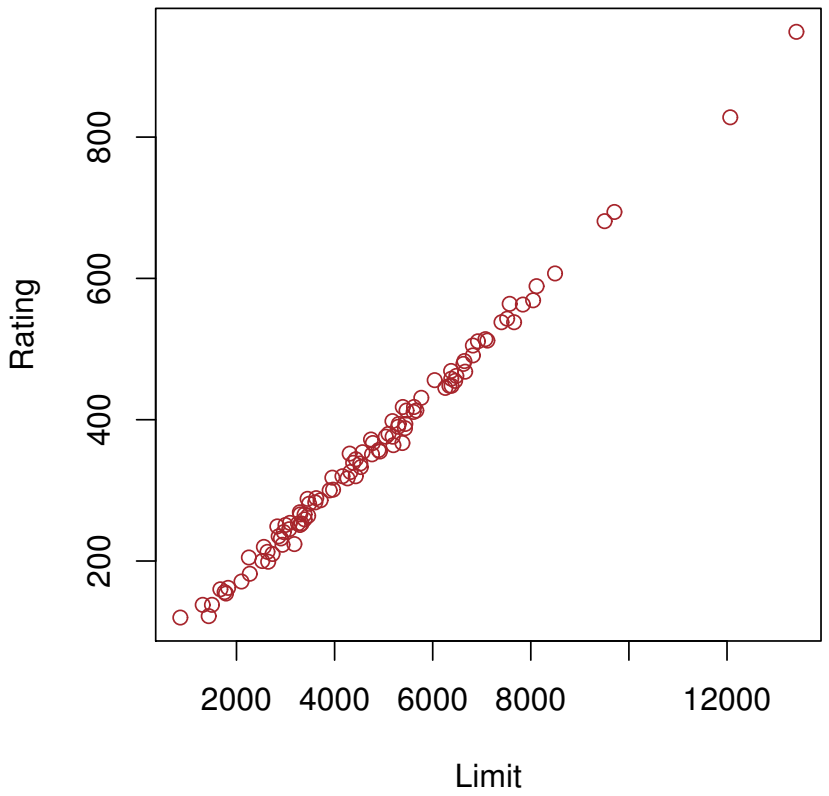
- Obserwacje o dźwigni znacznie przekraczającej średnią wartość ($\geq 2(p+1)/n$) moga być uznane za **obserwacje dźwigniowe**.
- Tu obserwacja 20 jest odstająca, a obserwacja 41 jest zarówno odstająca, jak i dźwigniowa. Obserwacje odstające, ale nie dźwigniowe są zwykle usuwane (nawet jeżeli nie powstały w wyniku błędu, nie wpłyną znacznie na parametry modelu).

(6) Współliniowość predyktorów

- Jest to sytuacja, w której dwa lub więcej predyktory mają wartości nawzajem ze sobą związane liniowo.
- Dla dwóch predyktorów widać to z ich korelacji, ale mogą zdarzać się mniej oczywiste związki liniowe typu $x_i \sim x_j + x_k$

Przykład z poziomem kredytu: predyktory 'limit' oraz 'rating' sa współliniowe

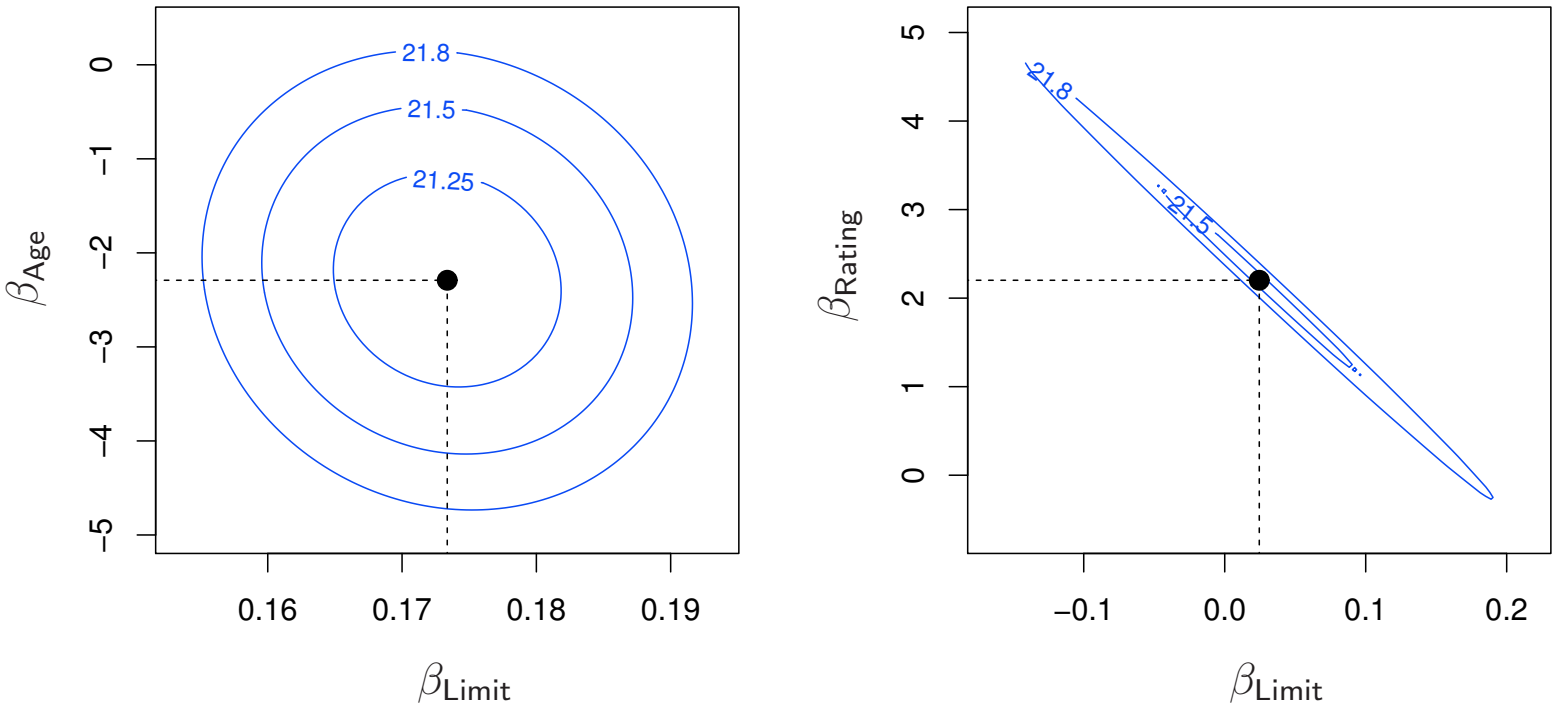




Predyktory: wiek (ang. age), limit kredytu (ang. limit) oraz ocena kredytobiorcy przez bank (rating).

Większa niepewność estymowanych wartości przy współliniowości predyktorów

Wykresy konturowe dla RSS:



- osie tak dobrane, aby prezentować wartości do 4-krotności odchylenia standardowego estymatorów po obu stronach (odpowiednik przedziałów ufności)

Dwa modele regresji liniowej: z współliniowością predyktorów i bez

Model I: regresja ze względu na **niewspółliniowe** predyktory:

Parametr	Estymacja	Std. błąd	t-statystyka	p-wartość
$\hat{\beta}_0$	-173.411	43.828	-3.957	< 0.0001
`age`	-2.292	0.672	-3.407	0.0007
`limit`	0.173	0.005	34.496	< 0.0001

Model II: regresja ze względu na **współliniowe** predyktory:

Parametr	Estymacja	Std. błąd	t-statystyka	p-wartość
$\hat{\beta}_0$	-377.537	45.254	-8.343	< 0.0001
`rating`	2.202	0.952	2.312	0.0213
`limit`	0.025	0.064	0.384	0.7012

Badanie korelacji wartości par predyktorów może nie wystarczyć

Czynnik inflacji wariancji (*variance inflation factor*, VIF):

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}},$$

gdzie $R^2_{X_j|X_{-j}}$ jest statystyką R^2 dla liniowej regresji zbudowanej dla zmiennej X_j, traktowanej teraz jako zmienna objaśniana, względem wszystkich pozostałych predyktorów.

Duża wartość VIF (>5 lub 10) sugeruje współliniowość pewnych predyktorów. Na przykład, wartości VIF dla predyktorów age, `rating` oraz `limit` wynosi odpowiednio: 1.01, 160.67, oraz 160.59.

Podsumowanie

- Predyktory jakościowe
- Interakcje
- Potencjalne problemy z liniową regresją