1 Macierz daszkowa

Definiujemy macierz rzutu na podprzestrzeń rozpiętą na kolumnach macierzy X jako

$$H = X(X^T X)^{-1} X^T$$

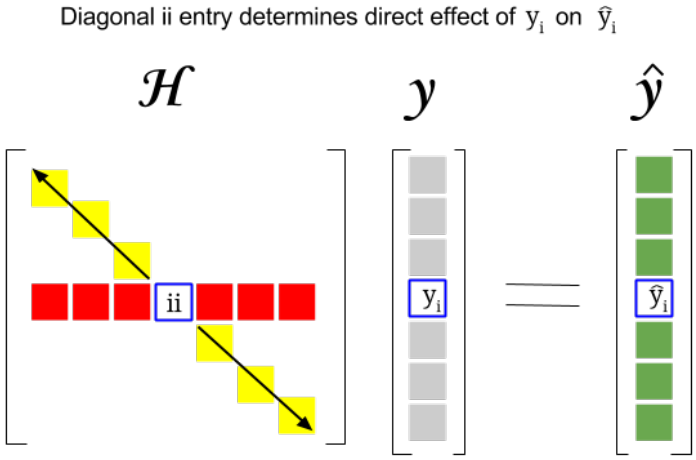
zatem

$$\hat{y} = Hy$$

i wiemy, że $\hat{y} \in C(X)$, zatem $H : y \mapsto C(X)$. Kilka własności

$$tr(H) = \sum h_{ii} = p$$
$$\frac{1}{n} \leq h_{ii} \leq 1$$
idempotentna, czyli rząd = tr

Każdy punkt danych próbuje przyciągnąć linię MNK do siebie. Jednak punkty znajdujące się dalej (w skrajnych wartościach predyktora) będą miały większy wpływ. Intuicyjnie



- $h_{ii} = 1$ oznacza, że obserwacja y_i w pełni determinuje $\hat{y}_i$ (największa dźwignia)
- $h_{ii} \approx 0$ oznacza, że obserwacja y_i ma praktycznie zerowy wpływ na $\hat{y}_i$, która będzie determinowana przez pozostałe obserwacje

Wiemy, że

$$\frac{\partial \hat{y}_i}{\partial y_i} = h_{ii}$$

Dlaczego?

$$\begin{bmatrix} h_{11} & \cdots & h_{1i} & \cdots & h_{1n} \\ & \vdots & & \vdots & \\ h_{i1} & \cdots & h_{ii} & \cdots & h_{in} \\ & \vdots & & \vdots & \\ h_{n1} & \cdots & h_{ni} & \cdots & h_{nn} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_i \\ \vdots \\ \hat{y}_n \end{bmatrix}$$

Zatem

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$$

czyli

$$\frac{\partial \hat{y}_i}{\partial y_i} = \frac{\partial \sum_{j=1}^{n} h_{ij} y_j}{\partial y_i} = \frac{\partial \left[h_{i1} y_1 + \cdots + h_{ii} y_i + \cdots + h_{in} y_n \right]}{\partial y_i} = h_{ii}$$

2 Diagnostyka modelu

Dlaczego model liniowy nie jest adekwatny?

- nie ma zależności liniowej $X\beta$
- nie ma wystarczającej liczby predyktorów
- błędy nie mają rozkładu normalnego
- błędy nie mają tej samej wariancji
- występowanie obserwacji **odstających** lub **wpływowych**
- pominięcie istotnych zmiennych objaśniających
- współliniowość zmiennych objaśniających*

*współliniowość nie obniża mocy predykcyjnej modelu. Jeśli mamy zmienne mocno skorelowane, to jakikolwiek losowy błąd pomiaru, który wchodzi w grę, zostanie również częściowo wyeliminowany przez fakt, że w zasadzie mamy dwa oddzielne pomiary tej samej rzeczy. Problem pojawia się we wnioskowaniu przyczynowym: uniemożliwia nam to stwierdzenie, przynajmniej z pewnością, który z współliniowych predyktorów dokonuje przewidywania, a zatem wyjaśnia i, przypuszczalnie, powoduje. Przy wystarczającej liczbie obserwacji w końcu będziemy w stanie zidentyfikować oddzielne efekty nawet wysoce współliniowych (ale nigdy idealnie współliniowych) zmiennych. **Zawsze** istnieje pewna współliniowość między predyktorami. To jeden z powodów, dla których generalnie potrzebujemy wielu obserwacji.

Rezydua e_i są **estymatami** błędów ε_i. Trudność polega na tym, że

$$e_i \sim \mathcal{N}(0, \sigma^2(1 - h_{ii})$$

zatem rezydua **nie** mają tego samego rozkładu i są od siebie zależne – choćby dlatego, że w modelu z interceptem $X^T e = 0$, zatem mając kolumnę jedynek, dostajemy, że $e_1 + ... + e_n = 0$, czyli jest zależnosć, bo ostatnie rezyduum musi się dopasować.

Cel **normalności** jest taki, że losową zmienność y można rozłożyć na *niezależne* losowe zmiany w projekcji $\hat{y}$ i resztę $y - \hat{y}$.

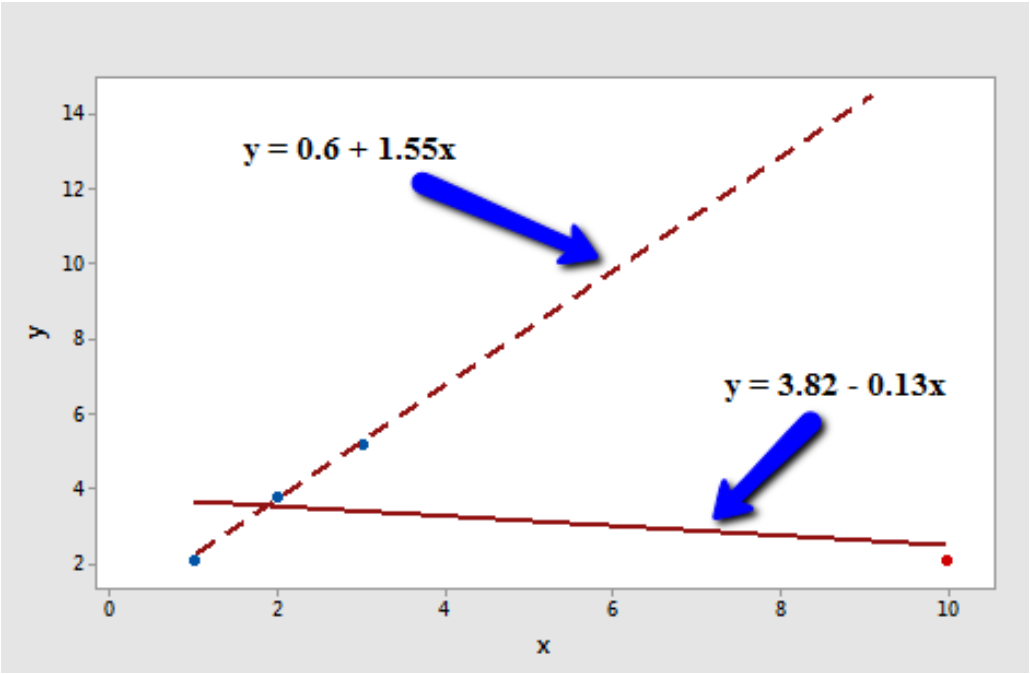
2.1 Rezydua studentyzowane

$$r_i = \frac{e_i}{se(e_i)}$$

gdzie

$$se(e_i) = s^2(1 - h_{ii})$$

Nie odejmujemy średniej, bo ona jest równa 0. Studentyzacja, ponieważ dzielimy przez oszacowanie odchylenia standardowego, a nie odchylenie standardowe. Co ważne, s^2 jest estymatorem nieodpornym na obserwacje odstające. **Problem z tym jest**



taki, że próbując zidentyfikować wartości odstające, może pojawić się sytuacja, w której istnieje potencjalna wartość odstająca, która wpływa na model regresji w takim stopniu, że oszacowana funkcja regresji jest *ciągnięta* w kierunku potencjalnej wartości odstającej, tak że nie jest ona oznaczona jako wartość odstająca przy użyciu studentyzowanego rezyduum. Z tego też powodu używamy **modyfikowanych rezuduów studentyzowanych.**

2.2 Modyfikowane rezydua studentyzowane

Chcąc ocenić, czy $i-$ta obserwacja jest **odstająca**, usuwamy ją ze zbioru danych (u_{-i})– próba z pominięciem $i-$tej obserwacji).

$$e_{i,-i} = y_i - \hat{y}_{i,-i}$$

wtedy

$$t_i = \frac{e_{i,-i}}{se(e_{i,-i})} \sim t_{n-p-1}$$

ale można łatwiej

$$t_i = \frac{e_i}{s_{-i}(1-h_{ii})^{1/2}}$$

gdzie

$$s^2_{-i} = \frac{s^2(n-p) - \frac{e_i^2}{1-h_{ii}}}{n-p-1}$$

Nie musimy dopasowywać nowego modelu, bo

$$e_{i,-i} = \frac{e_i}{1-h_{ii}}$$

Inaczej

$$t_i = r_i \left(\frac{n-p-1}{n-p-r_i^2}\right)^{1/2}$$

2.3 Obserwacje odstające

Mogą, ale nie muszą mieć wpływu na rozwiązanie MNK. Z reguły duże rezyduum. Detekcja

- wykres studentyzowanych rezyduów
- wykres modyfikowanych rezyduów

Mamy pewną **heurystykę**

$$|r_i| = |\frac{e_i}{se(e_i)}| > 2$$
$$|t_i| = |\frac{e_{i,-i}}{se(e_{i,-i})}| > 2$$

podobieństwo z kwantylem rozkładu normalnego, ale r_i nie ma rozkładu normalnego, dlatego więcej sensu ma t_i.

2.4 Obserwacje wpływowe

Są to obserwacje wykorzystujące efekt dźwigni. Te, które wykazują duże odstępstwo od średniej wartości zmiennych objaśniających. Detekcja **heurystyczna**

$$h_{ii} \geq 2\frac{p}{n} \text{ podwojona średnia } h_{ii}$$

Główne narzędzie diagnostyczne: **odległość Cooka**

2.4.1 Odległość Cooka

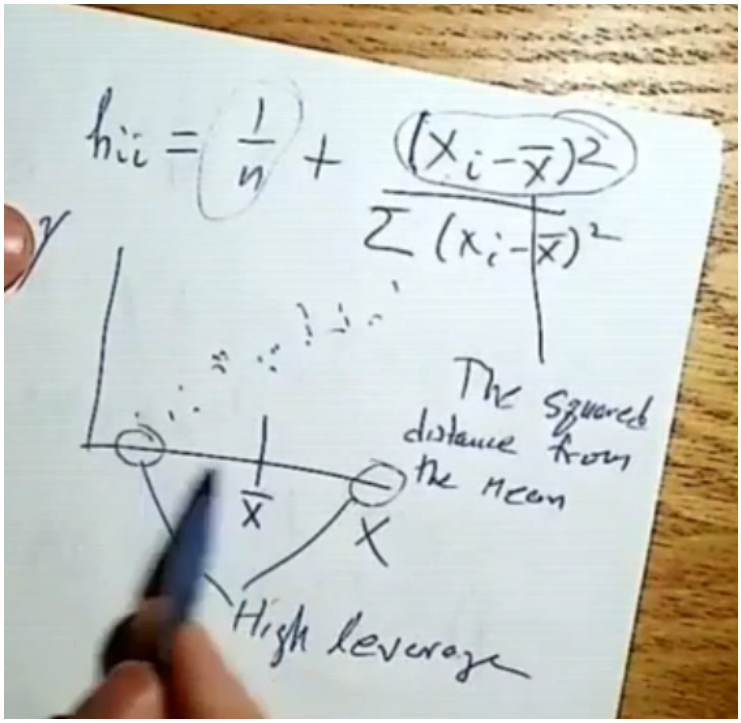
Usuwamy $i-$tą obserwację z danych. Odległość Cooka odnosi się do tego, jak średnio przewidywane wartości y przesuną się, jeśli dana obserwacja zostanie usunięta ze zbioru danych.

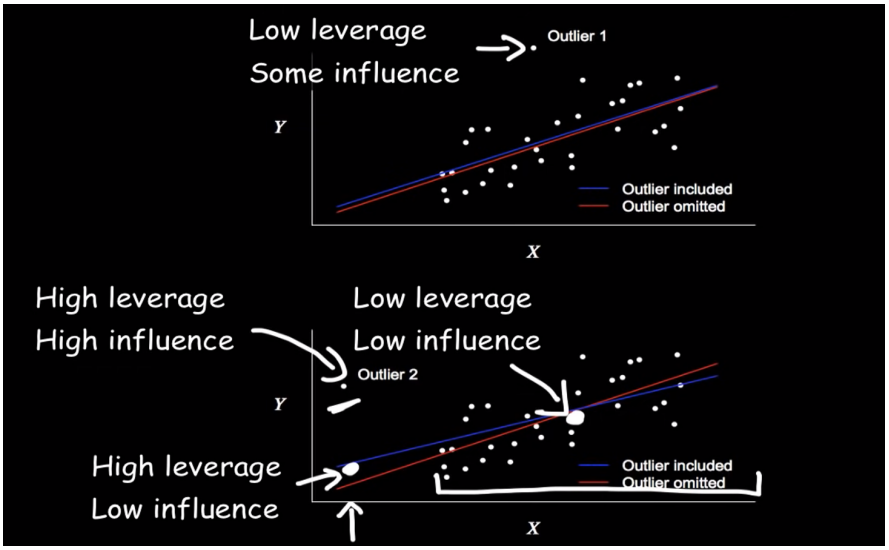
$$D_i = \frac{||\hat{y}_{(i)} - \hat{y}||^2}{ps^2} = \frac{r_i^2 h_{ii}}{p(1-h_{ii})} = \frac{e_i^2}{s^2 p}\frac{h_{ii}}{(1-h_{ii})^2}$$

$$\text{influence on parameters} = \text{f(leverage)} \times \text{g(outlyingness)}$$

Heurystyki

$$D_i > 1 \text{ coś z rozkładem } F_{p,n-p}, \text{ bliskie 1 dla dużych } n$$
$$D_i > \frac{4}{N}$$
$$D_i > \frac{4}{N-p-1}$$





3 lm() output

3.1 Rezydua vs fitted

Wykrywa nieliniowości, heteroskedastyczność, obserwacje odstające. Dla dobrze dopasowanego modelu mamy chmurę punktów wokół prostej $y = 0$, która nie ma żadnej struktury czy tendencji. Czerwona linia to LOWESS (locally weighted scatterplot smoothing), inaczej regresja (wielomianowa) lokalna.

Jeśli rezydua nie są symetryczne, to znaczy że mają skośny rozkład!

3.2 QQ plot

Sprawdzamy normalność rezyduów. One way to approach this question is to look at it in reverse: how could we begin with normally distributed residuals and arrange them to be heteroscedastic? From this point of view the answer becomes obvious: **associate the smaller residuals with the smaller predicted values**.

The moral is that heteroscedasticity characterizes a relationship between residual size and predictions whereas normality tells us nothing about how the residuals relate to anything else.

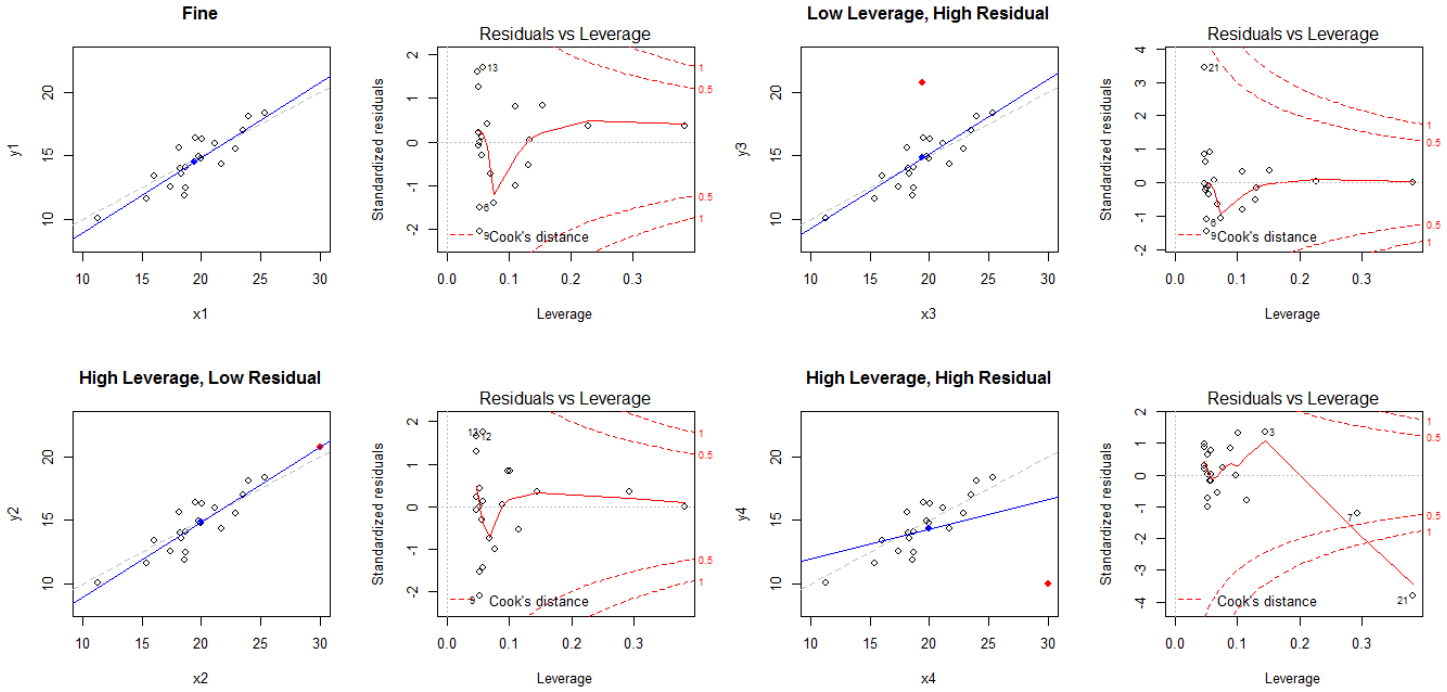
3.3 Scale-location

Ten wykres pokazuje, czy rezydua są równomiernie rozłożone wzdłuż zakresów predyktorów. W ten sposób można sprawdzić założenie równej wariancji (homoskedastyczności). Dobrze, jeśli widać poziomą linię z równo (losowo) rozłożonymi punktami.

3.4 Rezydua vs leverage

Ten wykres pomaga nam znaleźć obserwacje wpływowe, jeśli takie istnieją. Nie wszystkie wartości odstające mają wpływ na analizę regresji liniowej. Nawet jeśli dane mają wartości skrajne, mogą one nie mieć wpływu na wyznaczenie linii regresji. Oznacza to, że wyniki nie będą się znacznie różnić, jeśli uwzględnimy lub wykluczymy je z analizy. W większości przypadków są one zgodne z trendem i nie mają większego znaczenia. Z drugiej strony, niektóre obserwacje mogą być bardzo wpływowe, nawet jeśli wyglądają na mieszczące się w rozsądnym zakresie wartości. Mogą to być skrajne przypadki w stosunku do linii regresji i mogą zmienić wyniki, jeśli wykluczymy je z analizy. Inną interpretacją jest to, że w większości przypadków nie są one zgodne z trendem.

W przeciwieństwie do innych wykresów, tym razem pattern nie jest istotny. Zwracamy uwagę na wartości odstające w prawym górnym rogu lub w prawym dolnym rogu. Są to miejsca, w których obserwacje mogą mieć wpływ na linię regresji. Szukamy obserwacji poza liniami przerywanymi. Gdy obserwacje znajdują się poza liniami przerywanymi (co oznacza, że mają wysokie wyniki odległości Cooka), obserwacje te mają wpływ na wyniki regresji. Rezultaty regresji zostaną zmienione, jeśli wykluczymy te obserwacje.



4 Metoda Ważonych Najmniejszych Kwadratów

4.1 Dlaczego?

Heteroskedastyczność powoduje, że metoda NK jest nieefektywna

- nie ma biasu w parametrach modelu, ale jest w oszacowaniach błędów standardowych tych parametrów
- testy statystyczne dla parametrów nie będą dokładne

Model (mamy heteroskedastyczność)

$$y_i = \beta_0 + \beta_1 x + \varepsilon_i$$

ale teraz mamy

$$\varepsilon_i \sim \mathcal{N}(0, \sigma_i)$$

istnieje też Uogólniona Metoda Najmniejszych Kwadratów, gdzie $Cov(\varepsilon_i, \varepsilon_j) \neq 0$. Definiujemy **wagę**

$$w_i = \frac{1}{\sigma_i^2}$$

dlatego w tej postaci, że wtedy

$$\frac{y - Xb}{\sigma_i} = \frac{\varepsilon_i}{\sigma_i} \sim N(0, 1)$$

Zatem minimalizujemy wyrażenie

$$\sum_{i=1}^{n} w_i(y_i - (\beta_0 + \beta_1 x_i))^2$$

wtedy

$$b_0 = \bar{y}_2 - b_1\bar{x}_2$$
$$b_1 = \frac{\sum w_i(y_i - \bar{y}_w)(x_i - \bar{x}_w)}{\sum w_i(x_i - \bar{x}_w)^2}$$

gdy $w_i = 1$ mamy MNK. Mamy

$$e_{wi} = \sqrt{w_i}e_i$$

Ogólnie, otrzymujemy

$$\hat{\beta}_{WMNK} = (X^T W X)^{-1} X^T W y$$

4.2 Jak estymować wagi?

A rule of thumb for OLS regression is that it isn't too impacted by heteroscedasticity as long as the maximum variance is not greater than 4 times the minimum variance.

Jeśli nie znamy wariancji dla każdej z obserwacji, to możemy np.

- założyć postać funkcyjną, np. proporcjonalność do zmiennej objaśnianej, predyktora
- przeprowadzić MNK, zrobić wykres e_i^2 od $\hat{y}_i$, dopasować linię i wagi będą odwrotnościami tej linii

5 Inne

- stała wariancja jest potrzebna, aby β była $BLUE$
- wykres rezydua vs. obserwacje to **nie** jest dobry pomysł, bo trudny w analizie – rezydua są skorelowane z obserwacjami (wektory nie są ortogonalne z definicji, bez odstających $\approx \sqrt{1 - p/N}$ lub też $\approx \sqrt{1 - R^2}$
- the variance of a residual should be smaller than σ^2 since the fitted line will *pick up* any little linear component that by chance happens to occur in the errors (there's always some). There's a reduction due to the intercept and a reduction due to the slope around the center of the data whose effect is strongest at the ends of the data.
- in a linear regression where the errors are identically distributed, the variability of residuals of inputs in the middle of the domain will be **higher than the variability of residuals at the ends of the domain** linear regressions fit endpoints better than the middle.
- linia regresji **przechodzi** przez punkt $(\bar{x}, \bar{y})$. Zacznijmy od

$$\hat{y}_i = y_i + e_i$$
$$\frac{1}{n}(\hat{y}_1 + \cdots + \hat{y}_n) = \frac{1}{n}(y_1 + e_1 + \cdots + y_n + e_n)$$
$$\frac{1}{n}(\hat{y}_1 + \cdots + \hat{y}_n) = \frac{1}{n}(y_1 + \cdots + y_n)$$
$$\bar{\hat{y}}_i = \bar{y}_i$$

  ponieważ $\sum e_i = 0$. Wtedy

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1}$$
$$\sum_i^n \hat{y}_i = \sum_i^n \left[\hat{\beta}_0 + \hat{\beta}_1 x_{i1}\right]$$
$$\frac{1}{n}\sum_i^n \hat{y}_i = \frac{1}{n}\sum_i^n \left[\hat{\beta}_0 + \hat{\beta}_1 x_{i1}\right]$$
$$\bar{y} = \frac{1}{n}\sum_i^n \left[\hat{\beta}_0 + \hat{\beta}_1 x_{i1}\right]$$
$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1$$

  Zatem linia MNK rzeczywiście przechodzi przez $(\bar{x}, \bar{y})$.

- Współczynnik r korelacji Pearsona to nachylenie prostej dla obu **wystandaryzowanych** zmiennych, zarówno x, jak i y ($\frac{x-\bar{x}}{\sigma_x}$).
- skoro minimalizujemy odległości punktów od prostej, to znaczy że ekstremalne wartości x będą bardziej **naciskać** lub **ciągnąć** naszą dźwignię, czyli linię regresji. A to oznacza, że te punkty dalej od średniej $\bar{x}$ będą bardziej wpływać na równanie prostej, więc błąd standardowy rezyduów będzie większy!
- Cook's distance is presumably more important to you if you are doing predictive modeling, whereas dfbeta is more important in explanatory modeling

$$X'X\hat{\beta} = X'X\left((X'X)^{-1}X'y\right) = \left(X'X(X'X)^{-1}\right)X'y = X'y$$

This does not imply that $X\hat{\beta} = y$ though. In algebra, the statement $CA = CB$ only implies that $A = B$ if C is invertible, and, in this case, $C = X'$ is not even (necessarily) a square matrix.

What it does imply, however, is that $X'y = X'\hat{y}$, which is true in general, since $X'e = X'y - X'X\hat{\beta} = 0$.