Analiza reszt i diagnostyka modelu $(\mathbf{Y}, \mathbf{X\beta}, \sigma^2\mathbf{I})$ -użyteczne wzory

Liniowy model statystyczny $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$, i=1,2,...,n, zapiszmy w postaci wektorowo-macierzowej

1. $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$
2. $E(\boldsymbol{\varepsilon}) = \mathbf{0}$
3. $V(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$

gdzie $\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$, $\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}$, $\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$, $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}$.

Jeżeli model regresji liniowej jest poprawny ciąg rezyduów $\hat{\varepsilon}_i$ powinien zachowywać się w przybliżeniu tak jak ciąg i.i.d $N(0,\sigma^2)$. W szczególności wykres rezyduów względem numeru porządkowego lub wartości przewidywanych $\hat{y}_i$ powinien zachowywać się w przybliżeniu tak jak ciąg i.i.d $N(0,\sigma^2)$. Wiadomo, że rezydua $\hat{\varepsilon}_i$ i=1,...,n nie są niezależne (sumują się do 0!) i nawet w przypadku adekwatnego modelu liniowego nie mają tej samej wariancji. Rzeczywiście

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y} = (\mathbf{I} - \mathbf{H})(\mathbf{X\beta} + \boldsymbol{\varepsilon}) = (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}.$$

Stąd $E(\hat{\boldsymbol{\varepsilon}}) = (\mathbf{I} - \mathbf{H})E(\mathbf{Y}) = (\mathbf{I} - \mathbf{H})\mathbf{X\beta} = \mathbf{X\beta} - \mathbf{HX\beta} = \mathbf{X\beta} - \mathbf{X\beta} = \mathbf{0}$,

$$V(\hat{\boldsymbol{\varepsilon}}) = (\mathbf{I} - \mathbf{H})V(\mathbf{Y})(\mathbf{I} - \mathbf{H})^T = (\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I}(\mathbf{I} - \mathbf{H}) = \sigma^2(\mathbf{I} - \mathbf{H})^2 = \sigma^2(\mathbf{I} - \mathbf{H}).$$

Biorąc pod uwagę postać macierzy $\mathbf{H} = \mathbf{X}[\mathbf{X}^{\mathrm{T}}\mathbf{X}]^{-1}\mathbf{X}^{\mathrm{T}}$

Wariancja i błąd standardowy rezyduum $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ mają postać

$$V(\hat{\varepsilon}_i) = \sigma^2(1 - H_{ii}), \quad SE_{\hat{\varepsilon}_i} = S\sqrt{1 - H_{ii}}, \text{ gdzie } S^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(Y_i - \overline{Y}_i)^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

Stąd mamy standaryzowane rezydua $\dfrac{\hat{\varepsilon}_i}{\sqrt{V(\hat{\varepsilon}_i)}} = \dfrac{\hat{\varepsilon}_i}{\sigma\sqrt{1 - H_{ii}}} = \dfrac{\hat{\varepsilon}_i}{\sigma\sqrt{1 - h_i}}$ (nie obserwujemy ich bo nie znamy σ) o jednostkowej wariancji i

i studentyzowane rezydua $r_i = \dfrac{\hat{\varepsilon}_i}{SE_{\hat{\varepsilon}_i}} = \dfrac{\hat{\varepsilon}_i}{S\sqrt{1 - H_{ii}}} = \dfrac{\hat{\varepsilon}_i}{S\sqrt{1 - h_i}}$, gdzie $h_i = H_{ii}$.

(uwaga: w programie R funkcja rstandard() zwraca reszty studentyzowane)

Uwaga. W przypadku dużej liczności próby wariancja $V(\hat{\varepsilon}_i) = \sigma^2(1 - H_{ii})$ jest w przybliżeniu równa wariancji błędów σ^2. W takim przypadku nie ma znaczenia, czy rozpatrujemy wykres rezyduów czy

rezyduów studentyzowanych. Jednakże w przypadku małych prób dla których wartości zmiennej objaśniającej nie są rozłożone równomiernie, niektóre błędy $SE_{\hat{\varepsilon}_i}$ mogą znacznie odbiegać od błędu *S*. W takim przypadku warto w analizie rezyduów użyć reszt studentyzowanych.

Identyfikacja obserwacji nietypowych-odstających i wpływowych.

Obserwacje odstające, to takie obserwacje, które nie spełniają modelowego równania $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$. Obserwacją wpływową nazywamy taką obserwację, której usunięcie ze zbiory danych powoduje dużą zmianę wektora estymatorów MNK. Powszechną praktyką jest uznawanie, że obserwacja jest odstająca jeżeli jej rezydyum studentyzowane jest co do wartości bezwzględnej większe od 2 (dla bardzo dużych zbiorów danych od 4). Rozpatrując studentyzowane rezydua r_i zamiast $e_i = \hat{\varepsilon}_i$ uwzględniamy różną zmienność rozkładów rezyduów, która może powodować, że niektóre wartości $\hat{\varepsilon}_i$ są pozornie odstające. Sporządzenie wykresu studentyzowanych rezyduów względem ich indeksu umożliwia zidentyfikowanie dużych wartości, które przypuszczalnie odpowiadają obserwacjom odstającym. Metoda ta jednak zawozi w przypadku wpływowej obserwacji odstającej Y_i dla których różnica $\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$ jest mała ($\mathbf{x}_i^T$ oznacza *i*-ty wiersz macierzy planu eksperymentu **X**). W celu poprawnej interpretacji również tych obserwacji rozpatruje się następującą modyfikację *i*-tego rezyduum

$$d_i = \hat{\varepsilon}_{(i)} = Y_i - \hat{Y}_{i(i)},$$

gdzie $\hat{Y}_{i(i)}$ jest wartością przewidywaną zmiennej objaśnianej w modelu regresji $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$, dla $\mathbf{x}^T = \mathbf{x}_i^T$ na podstawie zbioru danych

$$\mathbf{J}_i = \{(Y_1, \mathbf{x}_1^T), ..., (Y_{i-1}, \mathbf{x}_{i-1}^T), (Y_{i+1}, \mathbf{x}_{i+1}^T), ..., (Y_n, \mathbf{x}_n^T)\},$$

który powstaje z całego zbioru przez pominiecie *i*-tej obserwacji. Zauważmy, że dla wpływowej obserwacji odstającej wartość d_i w odróżnieniu od wartości $\hat{\varepsilon}_i$ nie będzie bliska 0. Wartość d_i nazywamy rezyduum modyfikowanym a jego studentyzowaną wersję

$$t_i = \frac{d_i}{SE_{d_i}} = \frac{\hat{\varepsilon}_{(i)}}{SE_{\hat{\varepsilon}_{(i)}}}$$

studentyzowanym rezyduum modyfikowanym.

W języku R funkcja rstudent() wylicza te reszty. W mongrafi Farawaya dla t_i używa się nazw: jacknife (externally studentized or crossvalidated) residuals. Wykorzystując algebrę macierzy w postaci blokowej można pokazać, że do wyznaczenia studentyzowanego rezyduum zmodyfikowanego

nie trzeba powtarzać procedury estymacji parametrów metodą MNK ze zredukowanym zbiorem danych, gdyż prawdziwy jest związek t_i z $\hat{\varepsilon}_i$ oraz $RSS = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ poprzez równość

$$t_i = \hat{\varepsilon}_i \sqrt{\frac{n-k-2}{RSS(1-h_i) - \hat{\varepsilon}_i^2}}, \text{ lub } t_i = r_i \sqrt{\frac{n-k-1}{n-k-r_i^2}}, \quad (p=k+1=\text{rank}(\mathbf{X})).$$

```
#reszty studentyzowane  (w R standaryzowane)  dla modelu
g <-lm (sr ~ pop15+pop75+dpi+ddpi,savings)
sg<-summary(g)
(sg$sig)
sqrt(deviance(g)/df.residual(g))#inny sposób wyznaczenia estymatora sig
r <- rstandard(g)
r1<-residuals(g)/(sg$sig*sqrt(1-hatvalues(g))) # inny sposób wyznaczenia reszt r_i
qqnorm (r) # wykres Q-Q dla reszt studentyzowanych r
abline (0,1)

#reszty studentyzowane zmodyfikowane (w R studentyzowane)
t<-rstudent(g)#reszty studentyzowane zmodyfikowane t_i
r<- rstandard(g) #reszty studentyzowane
t1<-r/sqrt((g$df.residual-r^2))*sqrt((g$df.residual-1)) # reszty studentyzowane zmodyfikowane
t2<-residuals(g)*sqrt(g$df.residual-1)/sqrt((deviance(g)*(1-hatvalues(g))-residuals(g)^2)) # reszty studentyzowane zmodyfikowane ver.2
```

Niektóre obserwacje nie pasują dobrze do modelu - nazywane są obserwacjami odstającymi. Inne obserwacje zmieniają dopasowanie modelu w sposób istotny - nazywane są obserwacjami wpływowymi. Punkt może być żadnym, jednym lub oboma z nich. Punkt dźwigni jest nietypowy w przestrzeni predyktorów - ma potencjał, aby wpłynąć na dopasowanie.

Obserwacje wpływowe, czyli takie których usunięcie ze zbioru danych powoduje dużą zmianę wektora estymatorów MNK mogą, ale nie muszą być obserwacjami odstającymi. Inną grupą punktów wśród których mogą znajdować się obserwacje wpływowe stanowią te, dla których wektor wartości zmiennych objaśniających jest znacznie oddalony od typowego wektora wartości zmiennych objaśniających (tzw.*punkty dźwigni*). Nie nazywamy ich obserwacjami odstającymi, gdyż fakt przyjmowania nietypowych wartości przez zmienne objaśniane nie ma nic wspólnego z zachodzeniem lub nie równości $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$. W przypadku regresji jednokrotnej wyróżnienie tych punktów jest proste, choćby na podstawie histogramu wartości zmiennej objaśniającej. W przypadku wielu zmiennych znaczne odbieganie wektora $\mathbf{x}$ od wektora średnich $\bar{\mathbf{x}} = (1, \bar{x}_1, ..., \bar{x}_k)$ wcale nie musi oznaczać, że któraś ze współrzędnych wektora $\mathbf{x}$ będzie znacznie odstawać od odpowiadającej jej współrzędnej wektora średnich. Pewna globalna miara odstępstwa obserwacji $\mathbf{x}$ od wektora średnich $\bar{\mathbf{x}}$ jest zadana przez i-ty diagonalny $h_i=H_{ii}$ element macierzy daszkowej $\mathbf{H}$ (czasami zwany *wpływem* (*influence*) lub *dźwignią* (*leverage*)). Ponieważ wiadomo, że

$$\sum_{i=1}^{n} h_i = \text{trace}(\mathbf{H}) = p = k+1 \quad \text{i} \quad \tfrac{1}{n} \le h_i \le 1,$$

$h_i = H_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$; (dla regresji jednokrotnej i wielokrotnej $h_i = H_{ii} = \frac{1}{n} + \frac{1}{n-1} d_M^2(\mathbf{x}_i, \bar{\mathbf{x}})$)

(z drugiej postaci macierzy daszkowej uzyskanej z modelu scentrowanego (zob. ML2)).

Oczywiście stąd $h_i \geq \frac{1}{n}$.

Ponieważ $\mathbf{H} = \mathbf{H}^2$ możemy napisać $h_i = H_{ii} = \sum_{j=1}^{n} H_{ij}^2 = H_{ii}^2 + \sum_{j \neq i}^{n} H_{ij}^2$. Stąd po podzieleniu przez h_i

$$1 = h_i + \frac{\sum_{j=1}^{n} H_{ij}^2}{h_i} \Rightarrow h_i \leq 1$$

Możemy więc przyjąć, że typowa wartość h_i nie przekracza znacznie wartości $\frac{k+1}{n} = \frac{p}{n}$. Przyjmuje sie że obserwacja dla której

$$h_i \geq \frac{2(k+1)}{n} = \frac{2p}{n}$$

jest potencjalną obserwacją wpływową. Dodatkowym uzasadnieniem wnikliwego rozpatrzenia obserwacji o dużym wpływie h_i jest fakt, że $V(e_i) = \sigma^2(1-h_i)$. Zatem dla obserwacji o dużym wpływie h_i wariancja odpowiadającego rezyduum jest mała gdyż wartość przewidywana jest "zmuszana" do pozostawania blisko zaobserwowanej wartości Y_i.

Każdą potencjalnie wpływową obserwację (spełniającą $h_i \geq \frac{2(k+1)}{n}$) usuwamy kolejno ze zbioru danych i sprawdzamy na ile zmienił się wektor współczynników i cały model. Alternatywnie proces identyfikacji obserwacji wpływowych można oprzeć na obliczeniu tzw. odległości Cooka D_i , która wykorzystuje koncepcję zmodyfikowanych rezyduów $d_i = \hat{\varepsilon}_{(i)} = Y_i - \hat{Y}_{i(i)}$:

$$D_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{(k+1)S^2} = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{(k+1)S^2} = \frac{\hat{\varepsilon}_i^2}{(k+1)S^2} \frac{h_i}{(1-h_i)^2} = \frac{r_i^2}{(k+1)} \frac{h_i}{(1-h_i)}.$$

```
#obserwacje wpływowe
cook <- cooks.distance(g)  #funkcja R wyznaczająca odległości Cooka
#sprawdzenie wzorów na odległości Cooka
cook1<-(residuals(g)^2*hatvalues(g))/(1-hatvalues(g))^2/sg$df[1]/sg$sigma^2
cook2<-rstandard(g)^2*hatvalues(g)/(1-hatvalues(g))/sg$df[1]

inflm.g <- influence.measures(g)#funkcja R  wyznaczająca wpływy usunięcia obserwacji  na model
inflm.g$infmat
```

Information for influence.measures() function. k = # (predictor)

Function	Description	Rough Cut-off
dffits()	the change in the fitted values (with appropriately scaled)	> 2*sqrt{(k+1)/n}
dfbetas()	the changes in the coefficients (with appropriately scaled)	> 2/sqrt(n)
covratio()	the change in the estimate of OLS covariance matrix (trace of cov matrix)	outside 1+/- 3*(k+1)/n
hatvalues()	standardized distance to mean of predictors used to measure the leverage of observation	> 2*(k+1)/n
cooks.distance()	standardized distance change for how far the estimate vector	> 4/n

Wartość D_i odpowiada wpływowi jaki na prognozę znanych wartości zmiennej objaśnianej ma usunięcie ze zbioru danych i-tej obserwacji. Mierzy ona też kwadrat odległości miedzy wektorem współczynników w pełnym modelu regresji i wektorem współczynników w modelu z usuniętą i-tą obserwacją. Duża wartość D_i wskazuje na znaczny wpływ usunięcia i-tej obserwacji, czyli i-ta obserwacja jest obserwacją wpływową. Zauważmy, że duża wartość h_i jest tylko jedną z przyczyn dużej wartości D_i. Obserwacja może być wpływowa (mieć dużą wartość D_i) przy umiarkowanej wartości h_i ale mając dużą wartość resztową $\hat{\varepsilon}_i$. Ta ostatnia uwag tłumaczy dlaczego często preferuje się analizę odległości Cooka zamiast analizy wartości wpływów.

Wykresy półnormalne half-normal plots

Wykresy półnormalne są przeznaczone do oceny danych dodatnich. Mogą być używane do $|\hat{\varepsilon}_i|$, ale są bardziej przydatne do wielkości diagnostycznych, takich jak dźwignie czy odległości Cooka. Pomysł polega na wykreśleniu danych względem dodatnich kwantyli normalnych. Abu to zrobić trzeba :

- posortować dane $x_{[1]} \leq x_{[2]} \leq \ldots \leq x_{[n]}$
- wyznaczyć $u_i = F^{-1}_{N(0,1)}(\frac{n+i}{2n+i})$
- wykonać wykres rozrzutu $(x_{[i]},u_i)$,i=1,…,n.

Co powinno być zrobione w związku z nietypowymi obserwacjami?

1. Najpierw sprawdź, czy nie wystąpił błąd wprowadzania danych. Są to stosunkowo powszechne sytuacje. Niestety pierwotne źródło danych mogło zostać utracone.
2. Zbadaj kontekst fizyczny - dlaczego tak się stało? Czasami odkrycie wartości odstającej może być szczególnie interesujące. Niektóre odkrycia naukowe wynikają z zauważenia nieoczekiwanych aberracji. Innym przykładem znaczenia wartości odstających jest analiza statystyczna transakcji kartami kredytowymi. W tym przypadku obserwacje odstające mogą stanowić nieuczciwe użycie.
3. Wyklucz punkt z analizy, ale spróbuj ponownie włączyć go później, jeśli model zostanie zmieniony. Wykluczenie jednej lub więcej obserwacji może zadecydować o uzyskaniu statystycznie istotnego wyniku lub niektórych niepublikowalnych badań. Może to prowadzić do trudnej decyzji o tym, jakie wyłączenia są uzasadnione. Aby uniknąć sugestii nieuczciwości, zawsze zgłaszaj istnienie wartości odstających, nawet jeśli nie uwzględniłeś ich w ostatecznym modelu.
4. Załóżmy, że znajdujesz wartości odstające, których nie można racjonalnie zidentyfikować jako błędy lub aberracje, ale są postrzegane jako występujące naturalnie. Zamiast wykluczać

te punkty, a następnie stosować metodę najmniejszych kwadratów, bardziej wydajne i niezawodne jest stosowanie odpornej regresji. Preferencja odpornej regresji staje się silniejsza, gdy istnieje wiele wartości odstających. Odrzucenie wartości odstających w połączeniu z metodą najmniejszych kwadratów nie jest dobrą metodą estymacji.

5. Automatyczne wykluczanie wartości odstających jest niebezpieczne. Narodowa Administracja Aeronautyki i Przestrzeni Kosmicznej (NASA) wystrzeliła satelitę Nimbus 7 w celu rejestrowania informacji o atmosferze. Po kilku latach działalności w 1985 r. British Antarctic Survey zaobserwował duży spadek ozonu atmosferycznego nad Antarktydą. Podczas dalszej analizy danych NASA stwierdzono, że program przetwarzający dane automatycznie odrzucał obserwacje, które były bardzo niskie i uważane za błędy. W ten sposób odkrycie dziury ozonowej na Antarktydzie zostało opóźnione o kilka lat. Być może, gdyby było to znane wcześniej, wycofanie chlorofluorowęglowodorów (CFC) zostałoby uzgodnione wcześniej i szkody mogłyby być ograniczone.

Współliniowość zmiennych objaśnianych

Pakiety statystyczne różnie obsługują brak możliwości identyfikacji macierzy (osobliwość $\mathbf{X}^T\mathbf{X}$). W przypadku regresji niektóre mogą zwracać komunikaty o błędach, a niektóre mogą dopasować do danych model, ponieważ błąd zaokrąglenia może umożliwić dokładną identyfikację, ale mamy wówczas złe uwarunkowanie modelu. W innych przypadkach mogą być zastosowane ograniczenia, ale mogą one różnić się od oczekiwanych. Domyślnie R dopasowuje największy możliwy do zidentyfikowania model, usuwając zmienne w odwrotnej kolejności występowania we wzorze modelu.

Jak wykryć fakt występowania współliniowości zmiennych objaśniających?

Jedną z możliwości jest analiza korelacji tych zmiennych. Jeżeli wartość bezwzględna współczynnika korelacji $|r_{x_1,x_2}|$ jest bliska 1, to wskazuje to na przybliżoną liniową zależność pomiędzy zmiennymi x_1 i x_2. Używając tej metody nie wykryjemy ewentualnych związków liniowych wiążących więcej niż 2 zmienne jednocześnie. W tym celu używa się współczynnika determinacji wielokrotnej R_i^2 dla hipotetycznego modelu liniowego w którym x_i jest zmienną objaśnianą a x_1 ,…, x_{i-1}, x_{i+1},…, x_p są zmiennymi objaśniającymi. Wartość R_i^2 bliska 1 wskazuje na współliniowość zmiennej x_i i pozostałych zmiennych objaśniających. Alternatywnie analizuje się tzw. współczynniki podbicia wariancji (*variance inflation factor*) $VIF_i = (1-R_i^2)^{-1}$. Duża wartość VIF_i dla pewnego i wskazuje na potencjalną liniową zależność i-tej zmiennej objaśniającej od pozostałych zmiennych. Dodatkowym

uzasadnieniem użycia współczynników VIF_i jest związek pomiędzy wariancją estymatora b_i a VIF_i. Można udowodnić, że

$$\sigma_{b_i}^2 = \frac{\sigma^2 VIF_i}{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)^2},$$

zatem duża wartość współczynnika VIF_i z reguły pociąga za sobą dużą zmienność estymatora b_i, co może być spowodowane współliniowością w danych. Przyjmuje się, że wielkość VIF_i większa od 5 wymaga dalszych badań, a powyżej 10 oznacza już współliniowość pomiędzy badanymi zmiennymi.

Jednym ze sposobów radzenia sobie ze zjawiskiem współliniowości jest użycie zamiast estymatorów MNK estymatorów otrzymanych metodą tzw. regresji grzbietowej (*ridge regression*), która mówiąc z grubsza polega na nałożeniu ograniczenia na ograniczenia na normę wektora parametrów $\sum_i b_i^2 \le t$.

Warunek ten jest równoważny zmodyfikowaniu układu równań normalnych $\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{Y}$ poprzez zwiększenie przekątnej macierzy $\mathbf{X}^T\mathbf{X}$ (regularyzacja Tichonowa) i rozważaniu układu

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{b} = \mathbf{X}^T\mathbf{Y}.$$

Tego samego typu metodą jest zaproponowana przez Tibshiraniego(1996) metoda *lasso* w której ograniczenie ma postać $\sum_i |b_i| \le t$.

Obie metody redukują wariancję estymatorów kosztem ich obciążenia. Inne metody atakowania problemu współliniowości, to regresja składowych głównych (*principal component regression*- PCR) i regresja częściowych najmniejszych kwadratów (*partial least squares regression PLSR*)

Jak sprawdzić, czy systematyczna część modelu $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$ jest poprawna?

Możemy spojrzeć wykresy

1. $\hat{\boldsymbol{\varepsilon}}$ względem $\hat{\mathbf{Y}}$ i predyktorów $\mathbf{X}_i$
2. $\hat{\mathbf{Y}}$ względem każdego predyktora $\mathbf{X}_i$

Możemy też sporządzić tzw. **wykresy zmiennej dodanej** (Partial *regression plots or Added Variable plots*). Na wykresie zmiennej dodanej zarówno zmienna odpowiedzi Y, jak i badana zmienna predykcyjna (powiedzmy X_1) są regresowane względem innych zmiennych predykcyjnych już znajdujących się w modelu regresji, a reszty są uzyskiwane dla każdej z nich. Te dwa zestawy reszt odzwierciedlają część każdej (Y i X_1), która nie jest liniowo powiązana z innymi zmiennymi predykcyjnymi. Wykres jednego zestawu reszt względem drugiego zestawu pokazałby, jaki jest marginalny wkład kandydata na predyktora w zmniejszaniu zmienności resztowej, a także informacje o naturze jego marginalnego wkładu.

Częściowe wykresy resztkowe (*Partial Residual plots*) są konkurentem dla wykresów zmiennych dodanych. Wykresy te przedstawiają $\hat{\varepsilon}+\beta_i X_i$ w stosunku do X_i. Aby zobaczyć, skąd to się bierze, spójrz na odpowiedź z usuniętym przewidywanym efektem pozostałych predyktorów X:

$$Y-\sum_{j\neq i}\beta_j X_j = Y-\sum_{j}\beta_j X_j+\beta_i X_i = \hat{\varepsilon}+\beta_i X_i$$

Ponownie nachylenie wykresu będzie wynosić β_i, a interpretacja jest taka sama. Częściowe wykresy resztkowe są uważane za lepsze do wykrywania nieliniowości, podczas gdy wykresy zmiennych dodanych są lepsze do wykrywania wartości odstających/wpływowych.

```
plot(savings$pop15,residuals(g)+coef(g)['pop15']*savings$pop15,xlab="pop'n under 15",ylab="Savings(Adjusted)") #Partial residual plot
abline (0, coef (g) ['pop15']) #współcztnnik przy pop 15 nie zmienił się
prplot(g,1) #funkcja Farawaya Partial residual plot wywołana dla pierwszego predyktora
```

Wybrane własności obiekty klasy lm

$coefficients	Dopasowane współczynniki $\hat{\boldsymbol{\beta}}$ modelu liniowego
$residuals	Wektor reszt $\hat{\boldsymbol{\varepsilon}}$
$fitted.values	Wektor wartości dopasowanych $\hat{\mathbf{Y}}$
$df.residual	Liczba stopni swobody dla reszt $(n\text{-}p)=(n\text{-}k\text{-}1)$
$model	Ramka danych użyta do oceny współczynników

Wybrane własności obiekty klasy summary.lm

$residuals	Wektor reszt $\hat{\boldsymbol{\varepsilon}}$
$coefficients	Macierz wymiaru $p\times 4$. Dla każdego współczynnika w macierzy znajdują się informacje o ocenie tego współczynnika , błędzie standardowym tej oceny, wartość statystyki Walda dla tego współczynnika i p-wartość testu dwustronnego Walda
$sigma	Ocena odchylenia standardowego $\hat{S}=\sqrt{\frac{1}{n-p}\sum(Y_i-\hat{Y}_i)^2}$
$df	Stopnie swobody –wektor(p,n-p,p*) p* oznacza liczbę parametrów w modelu (modelu pełnego rzędu p= p*
$fstatistic	Wektor z wartością statystyki F dla modelu, liczba stopni swobody w liczniku i mianowniku
$r.squared	R^2 proporcja wyjaśnionej wariancji
$adj.r.squared	Skorygowany R^2_{adj}
$cov.unscaled	Macierz kowariancji dla ocen współczynników
$correlation	Macierz korelacji dla ocen współczynników

Wybrane funkcje do operacji na obiektach klasy lm

summary()	Wynikiem jest opis dopasowania modelu liniowego
coeff()	Wynikiem jest wektor ocen współczynników $\hat{\boldsymbol{\beta}}$ modelu liniowego
resid()	Wynikiem jest wektor reszt $\hat{\boldsymbol{\varepsilon}}$
fitted()	Wynikiem jest wektor wartości dopasowanych $\hat{\mathbf{Y}}$
deviance()	Wynikiem jest *RSS* suma kwadratów reszt
anova()	Podsumowanie analizy wariancji dla modelu liniowego
predict()	Predykcja zmiennej objaśnianej na nowym zbiorze danych
plot()	Wykresy diagnostyczne dla modelu liniowego
cooks.distance()	Odległości Cooka
hatvalues()	Przekątna macierzy daszkowej $\hat{\mathbf{H}}$ - dźwignie
influence()	Zwraca ramkę z dźwigniami, zmiany współczynników modelu po usunięciu obserwacji, nowe oceny sigma po usunięciu obserwacji `help("influence")`
influence.measures	Podobne działanie dotyczy także glm zob. `help("influence.measures")`

Uupełnienia szczegółów rachunkowych

Oznaczmy

- $\mathbf{X}_{(i)}$ macierz $\mathbf{X}$ z usuniętym *i*-tym wierszem który oznaczamy przez $\mathbf{x}_i^T$,
- $\mathbf{Y}_{(i)}$ wektor $\mathbf{Y}$ z usuniętym *i*-tym elementem (obserwacją) Y_i,
- $\hat{\boldsymbol{\beta}}_{(i)}$ wektor współczynników uzyskany MNK na podstawie model $(\mathbf{Y}_{(i)}, \mathbf{X}_{(i)}\boldsymbol{\beta}_{(i)}, \sigma^2\mathbf{I})$.

Fakt 1. Pokażemy, że $\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - \frac{\hat{\varepsilon}_i (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i}{1-h_i}$. Zrobimy to w 4 krokach

Krok1. Zauważmy że $\mathbf{X}^T\mathbf{X} = \mathbf{X}_{(i)}^T\mathbf{X}_{(i)} + \mathbf{x}_i\mathbf{x}_i^T$, $\mathbf{X}^T\mathbf{Y} = \mathbf{X}_{(i)}^T\mathbf{Y}_{(i)} + \mathbf{x}_i Y_i$.

Rzeczywiście element (p,q) macierzy $\mathbf{X}^T\mathbf{X}$ jest iloczynem skalarnym *p*-tej i *q*-tej kolumny macierzy $\mathbf{X}$, więc $(\mathbf{X}^T\mathbf{X})_{(p,q)} = \sum_{j=1}^{n} x_{jp}x_{jq} = \sum_{j\neq i}^{n} x_{jp}x_{jq} + x_{ip}x_{iq} = (\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})_{(p,q)} + (\mathbf{x}_i\mathbf{x}_i^T)_{(p,q)}$. Podobnie

$$(\mathbf{X}^T\mathbf{Y})_p = \sum_{j=1}^{n} x_{jp}Y_j = \sum_{j\neq i}^{n} x_{jp}Y_j + x_{ip}Y_i = (\mathbf{X}_{(i)}^T\mathbf{Y}_{(i)})_p + (\mathbf{x}_i Y_i)_p .$$

Krok 2. $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}_{(i)}^T\mathbf{Y}_{(i)} + \mathbf{x}_i Y_i) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{(i)}^T\mathbf{Y}_{(i)} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i Y_i$.

Krok 3. Korzystając ze wzoru Bartletta (B) (zob. uzupełnienie z macierzy)

$$(\mathbf{A} - \mathbf{b}\mathbf{b}^{\mathrm{T}})^{-1} = \mathbf{A}^{-1} + \frac{\mathbf{A}^{-1}\mathbf{b}\mathbf{b}^{\mathrm{T}}\mathbf{A}^{-1}}{1-\mathbf{b}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{b}}$$

pokażemy, że $\hat{\boldsymbol{\beta}}_{(i)} = \left[(\mathbf{X}^T\mathbf{X})^{-1} + \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}}{1-h_i}\right]\mathbf{X}_{(i)}^T\mathbf{Y}_{(i)}$.

$$\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1}\mathbf{X}_{(i)}^T\mathbf{Y}_{(i)} \overset{\text{krok1}}{=} \left[(\mathbf{X}^T\mathbf{X}) - \mathbf{x}_i\mathbf{x}_i^T\right]^{-1}\mathbf{X}_{(i)}^T\mathbf{Y}_{(i)} \overset{\text{B}}{=} \left[(\mathbf{X}^T\mathbf{X})^{-1} + \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}}{1-\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i}\right]\mathbf{X}_{(i)}^T\mathbf{Y}_{(i)} =$$

$$= \left[(\mathbf{X}^T\mathbf{X})^{-1} + \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}}{1-h_i}\right]\mathbf{X}_{(i)}^T\mathbf{Y}_{(i)}$$

Krok 4.

$$\hat{\boldsymbol{\beta}}_{(i)} = \left[(\mathbf{X}^T\mathbf{X})^{-1} + \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}}{1-h_i} \right] \mathbf{X}_{(i)}^T\mathbf{Y}_{(i)} =$$

$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{(i)}^T\mathbf{Y}_{(i)} + \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}}{1-h_i}\mathbf{X}_{(i)}^T\mathbf{Y}_{(i)} \overset{\text{Krok2}}{=}$$

$$= \hat{\boldsymbol{\beta}} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i Y_i + \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}}{1-h_i}\mathbf{X}_{(i)}^T\mathbf{Y}_{(i)} =$$

$$= \hat{\boldsymbol{\beta}} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\left(Y_i - \frac{\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{(i)}^T\mathbf{Y}_{(i)}}{1-h_i}\right) \overset{\text{Krok1}}{=}$$

$$= \hat{\boldsymbol{\beta}} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\left(Y_i - \frac{\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y} - \mathbf{x}_i Y_i)}{1-h_i}\right) =$$

$$= \hat{\boldsymbol{\beta}} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\left(Y_i - \frac{\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} - \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i Y_i}{1-h_i}\right) =$$

$$= \hat{\boldsymbol{\beta}} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\left(Y_i - \frac{\hat{Y}_i - h_i Y_i}{1-h_i}\right) = \hat{\boldsymbol{\beta}} - \frac{\hat{\varepsilon}_i(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i}{1-h_i}$$

Fakt 2. $\hat{\varepsilon}_{(i)} = \dfrac{\hat{\varepsilon}_i}{1-h_i}$

$$\hat{\varepsilon}_{(i)} = Y_i - \hat{Y}_{(i)} = Y_i - \mathbf{x}_i^T\boldsymbol{\beta}_{(i)} \overset{\text{Fakt1}}{=} Y_i - \mathbf{x}_i^T\left[\hat{\boldsymbol{\beta}} - \frac{\hat{\varepsilon}_i(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i}{1-h_i}\right] = Y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}} + \frac{\hat{\varepsilon}_i}{1-h_i}\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i =$$

$$= Y_i - \hat{Y}_i + \frac{\hat{\varepsilon}_i}{1-h_i}h_i = \hat{\varepsilon}_i + \frac{\hat{\varepsilon}_i}{1-h_i}h_i = \frac{\hat{\varepsilon}_i}{1-h_i}$$

Fakt 3. $t_i = \dfrac{d_i}{SE_{d_i}} = \dfrac{\hat{\varepsilon}_{(i)}}{SE_{\hat{\varepsilon}_{(i)}}} = \dfrac{\hat{\varepsilon}_i}{S_{(i)}\sqrt{1-h_i}} = \hat{\varepsilon}_i\sqrt{\dfrac{n-k-2}{SSE(1-h_i)-\hat{\varepsilon}_i^2}}$

Z faktu 2 mamy $V(\hat{\varepsilon}_{(i)}) = V\left(\dfrac{\hat{\varepsilon}_i}{1-h_i}\right) = \dfrac{V(\hat{\varepsilon}_i)}{(1-h_i)^2} = \dfrac{\sigma^2(1-h_i)}{(1-h_i)^2} = \dfrac{\sigma^2}{1-h_i}$.

Przyjmując $\hat{V}(\hat{\varepsilon}_{(i)}) = \dfrac{S_{(i)}^2}{1-h_i}$ jako estymator wariancji $V(\hat{\varepsilon}_{(i)})$ otrzymujemy $\dfrac{\hat{\varepsilon}_{(i)}}{\sqrt{\hat{V}(\hat{\varepsilon}_{(i)})}} = \dfrac{\hat{\varepsilon}_i}{S_{(i)}\sqrt{1-h_i}}$

gdzie $S_{(i)}^2 = \dfrac{SSE_{(i)}}{n-k-2}$.

Uwaga: $SSE = \|\hat{\boldsymbol{\varepsilon}}\|^2 = \|\mathbf{Y}-\hat{\mathbf{Y}}\|^2 = \langle \mathbf{Y}-\hat{\mathbf{Y}}, \mathbf{Y}-\hat{\mathbf{Y}}\rangle = \|\mathbf{Y}\|^2 - \|\hat{\mathbf{Y}}\|^2 = \|\mathbf{Y}\|^2 - \langle \hat{\mathbf{Y}}, \hat{\mathbf{Y}}\rangle = \|\mathbf{Y}\|^2 - \langle \mathbf{Y}, \hat{\mathbf{Y}}\rangle$

$$SSE_{(i)} = \|\mathbf{Y}_{(i)} - \hat{\mathbf{Y}}_{(i)}\|^2 = \|\mathbf{Y}_{(i)}\|^2 - \langle \mathbf{Y}_{(i)}, \hat{\mathbf{Y}}_{(i)}\rangle = \mathbf{Y}_{(i)}^T\mathbf{Y}_{(i)} - \mathbf{Y}_{(i)}^T\hat{\mathbf{Y}}_{(i)} = (\mathbf{Y}^T\mathbf{Y} - Y_i^2) - \mathbf{Y}_{(i)}^T\mathbf{X}_{(i)}\hat{\boldsymbol{\beta}}_{(i)} =$$

$$= (\mathbf{Y}^T\mathbf{Y} - Y_i^2) - \hat{\boldsymbol{\beta}}_{(i)}^T\mathbf{X}_{(i)}^T\mathbf{Y}_{(i)}$$

Równość $\mathbf{X}^T\mathbf{Y} = \mathbf{X}_{(i)}^T\mathbf{Y}_{(i)} + \mathbf{x}_i Y_i$ (Krok 1) po transpozycji przybiera postać $\mathbf{Y}^T\mathbf{X} = \mathbf{Y}_{(i)}^T\mathbf{X}_{(i)} + \mathbf{x}_i^T Y_i$

więc

$$\mathbf{Y}_{(i)}^T\mathbf{X}_{(i)}\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{Y}^T\mathbf{X} - \mathbf{x}_i^T Y_i)(\hat{\boldsymbol{\beta}} - \frac{\hat{\varepsilon}_i}{1-h_i}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i) =$$

$$= \mathbf{Y}^T\mathbf{X}\hat{\boldsymbol{\beta}} - Y_i\mathbf{x}_i^T\hat{\boldsymbol{\beta}} - \mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\frac{\hat{\varepsilon}_i}{1-h_i} + \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\frac{Y_i\hat{\varepsilon}_i}{1-h_i} =$$

$$= \mathbf{Y}^T\mathbf{X}\hat{\boldsymbol{\beta}} - \hat{Y}_i Y_i - \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\frac{\hat{\varepsilon}_i}{1-h_i} + h_i\frac{Y_i\hat{\varepsilon}_i}{1-h_i} =$$

$$= \mathbf{Y}^T\mathbf{X}\hat{\boldsymbol{\beta}} - \hat{Y}_i Y_i - \hat{Y}_i\frac{\hat{\varepsilon}_i}{1-h_i} + h_i\frac{Y_i\hat{\varepsilon}_i}{1-h_i} = \mathbf{Y}^T\mathbf{X}\hat{\boldsymbol{\beta}} - (Y_i - \hat{\varepsilon}_i)Y_i - \hat{Y}_i\frac{\hat{\varepsilon}_i}{1-h_i} + h_i\frac{Y_i\hat{\varepsilon}_i}{1-h_i} =$$

$$= \mathbf{Y}^T\mathbf{X}\hat{\boldsymbol{\beta}} - Y_i^2 + \frac{(1-h_i)\hat{\varepsilon}_i Y_i - \hat{Y}_i\hat{\varepsilon}_i + h_i Y_i\hat{\varepsilon}_i}{1-h_i} = \mathbf{Y}^T\mathbf{X}\hat{\boldsymbol{\beta}} - Y_i^2 + \frac{\hat{\varepsilon}_i Y_i - h_i\hat{\varepsilon}_i Y_i - \hat{Y}_i\hat{\varepsilon}_i + h_i Y_i\hat{\varepsilon}_i}{1-h_i} =$$

$$= \mathbf{Y}^T\mathbf{X}\hat{\boldsymbol{\beta}} - Y_i^2 + \frac{\hat{\varepsilon}_i Y_i - \hat{Y}_i\hat{\varepsilon}_i}{1-h_i} = \mathbf{Y}^T\mathbf{X}\hat{\boldsymbol{\beta}} - Y_i^2 + \frac{\hat{\varepsilon}_i(Y_i - \hat{Y}_i)}{1-h_i} = \mathbf{Y}^T\mathbf{X}\hat{\boldsymbol{\beta}} - Y_i^2 + \frac{\hat{\varepsilon}_i^2}{1-h_i}$$

W powyższych rachunkach wykorzystano związki

$$\mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{Y} = \hat{Y}_i, \qquad \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i = h_i, \qquad Y_i = \hat{Y}_i + \hat{\varepsilon}_i$$

Wobec powyższego

$$SSE_{(i)} = (\mathbf{Y}^T\mathbf{Y} - Y_i^2) - \hat{\boldsymbol{\beta}}_{(i)}^T\mathbf{X}_{(i)}^T\mathbf{Y}_{(i)} = (SSE + \mathbf{Y}^T\hat{\mathbf{Y}} - Y_i^2) - (\mathbf{Y}^T\mathbf{X}\hat{\boldsymbol{\beta}} - Y_i^2 + \frac{\hat{\varepsilon}_i^2}{1-h_i}) =$$

$$= SSE + \mathbf{Y}^T\hat{\mathbf{Y}} - Y_i^2 - \mathbf{Y}^T\hat{\mathbf{Y}} + Y_i^2 - \frac{\hat{\varepsilon}_i^2}{1-h_i} = SSE - \frac{\hat{\varepsilon}_i^2}{1-h_i}$$

Fakt 4. $$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{(k+1)S^2} = \frac{r_i^2 h_i}{(k+1)S^2} = \frac{\hat{\varepsilon}_i^2}{(k+1)S^2}\frac{h_i}{(1-h_i)^2}$$

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{(k+1)S^2} = \frac{(\mathbf{X}\hat{\boldsymbol{\beta}}_{(i)} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{X}\hat{\boldsymbol{\beta}}_{(i)} - \mathbf{X}\hat{\boldsymbol{\beta}})}{(k+1)S^2} = \frac{(\hat{\mathbf{Y}}_{(i)} - \hat{\mathbf{Y}})^T(\hat{\mathbf{Y}}_{(i)} - \hat{\mathbf{Y}})}{(k+1)S^2}$$

Korzystając z zależności $\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}} = -\frac{\hat{\varepsilon}_i(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i}{1-h_i}$ i $r_i = \frac{\hat{\varepsilon}_i}{SE_{\hat{\varepsilon}_i}} = \frac{\hat{\varepsilon}_i}{S\sqrt{1-h_i}}$

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{(k+1)S^2} = \frac{\hat{\varepsilon}_i^2}{(k+1)S^2(1-h_i)^2}\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i =$$

$$\frac{\hat{\varepsilon}_i^2}{(k+1)S^2(1-h_i)^2}\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i = \frac{\hat{\varepsilon}_i^2 h_i}{(k+1)S^2(1-h_i)^2} = \frac{r_i^2 h_i}{(k+1)(1-h_i)}$$