

# Modele liniowe

Michał Kos

Uniwersytet Wrocławski

## 1 Problem wyboru modelu

## 2 Diagnostyka

- Partial regression plots
- Macierz H w diagnostyce modelu
- Miara DFFITS
- Odległość Cook'a
- Miara DFBETAS
- Miary Variance inflation factor oraz Tolerance

# Table of Contents

## 1 Problem wyboru modelu

## 2 Diagnostyka

- Partial regression plots
- Macierz  $H$  w diagnostyce modelu
- Miara DFFITS
- Odległość Cook'a
- Miara DFBETAS
- Miary Variance inflation factor oraz Tolerance

Na poprzednim wykładzie poznaliśmy tzw. **"problem wyboru modelu"**, który próbuje odpowiedzieć na pytanie:

**Które zmienne objaśniające  $X_1, \dots, X_{p-1}$  w istotny sposób wpływają na zmienną objaśnianą  $Y$ , a dla których ów wpływ jest pomijalny?**

**Równoważnie możemy pytać o nośnika prawdziwego wektora parametrów  $S = \text{Supp}(\beta) = \{i : \beta_i \neq 0\}$ , gdyż warunek  $\beta_i = 0$  pociąga za sobą brak istotności zmiennej  $X_i$ .**

# Problem wyboru modelu

Poznaliśmy kilka statystycznych narzędzi umożliwiających **częściową** odpowiedź na powyższe pytanie:

- **ogólny test F**, umożliwiający porównanie wyłącznie modeli hierarchicznych (model pełny zawiera wszystkie zmienne znajdujące się w modelu zredukowanym),
- **maksymalizacja współczynnika determinacji  $R^2$** , umożliwiająca porównanie modeli o tej samej liczbie zmiennych objaśniających,
- **maksymalizacja modyfikowanego współczynnika determinacji  $R^2_{adj}$** , (mało stabilne kryterium).

Widzimy, że każda z metod ma swoje ograniczenia. Przy użyciu ogólnego test F nie możemy porównywać modeli niehierarchicznych. Za pomocą  $R^2$  możemy porównywać modele niehierarchiczne, ale muszą one mieć tę samą liczbę zmiennych, a  $R^2_{adj}$  działa w mało stabilny sposób.

Alternatywnymi metodami wyboru modelu są:

- kryterium informacyjne Akaike (AIC – Akaike information criterion),
- Bayesowskie kryterium informacyjne (BIC – Bayesian information criterion; SIC – Schwarz inf. crit.),
- kryterium  $C_p$  Mallows'a.

Kryteria AIC oraz BIC są modyfikacjami metody największej wiarygodności i są konstruowane w taki sposób, by znaleźć balans pomiędzy dopasowaniem modelu do danych i nadmierną złożonością modelu:

$$\hat{\beta} = \operatorname{argmax}_b ( \log(\text{likelihood}(b)) - \text{kara za "duże" } p )$$

Składnik  $\log(\text{likelihood}(b))$  odpowiada za dopasowanie modelu do danych. Z kolei drugi składnik jest karą za wykorzystywanie nieistotnych zmiennych. W ścisły sposób kryterium AIC można zdefiniować w następujący sposób:

- 1 Dla dowolnej podmacierzy  $\tilde{X}$  (o l. kol.  $\tilde{p}$ ) macierzy planu  $X$  wyznacz statystykę:

$$AIC(\tilde{X}) = n \log \left( \frac{SSE(\tilde{X})}{n} \right) + 2\tilde{p}$$

- 2 wybierz model  $\tilde{X}$  o najniższej wartości statystyka AIC.

Kryterium BIC jest zdefiniowana w analogiczny sposób z tą różnicą że szukamy minimum po statystykach postaci:

$$BIC(\tilde{X}) = n \log \left( \frac{SSE(\tilde{X})}{n} \right) + \log(n)\tilde{p}$$

# Kryterium $C_p$ Mallows'a

Widzieliśmy że w przypadku AIC i BIC istotne znaczenie miała statystyka  $SSE(\tilde{\mathbf{X}})$ , będąca miarą dopasowania modelu do danych. W przypadku kryterium  $C_p$  Mallows'a również odgrywa ona istotną rolę. Statystyka  $C_{\tilde{p}}(\tilde{\mathbf{X}})$  Mallows'a stowarzyszona z modelem skonstruowanym na podstawie macierzy  $\tilde{\mathbf{X}}$  ma postać:

$$C_{\tilde{p}}(\tilde{\mathbf{X}}) = \frac{SSE(\tilde{\mathbf{X}})}{MSE(F)} - n + 2\tilde{p}$$

Jednym z kryteriów oceny poprawności modelu jest to, czy nie wprowadza on znaczącego obciążenia w predykcji  $B_i = E(\hat{Y}_i) - E(Y_i)$   $i = 1, \dots, n$ . Można pokazać że statystyka  $C_{\tilde{p}}$  jest estymatorem następującego wyrażenia:

$$\frac{1}{\sigma^2} \sum_{i=1}^n B_i^2(\tilde{\mathbf{X}})$$

w związku z tym opisuje łączne zachowanie obciążeń. Kryterium Mallows'a stwierdza, że model ma dobre własności, gdy statystyka  $C_{\tilde{p}}$  jest bliska lub mniejsza niż  $\tilde{p}$ . Dlatego na jego podstawie należy wybrać najoszczędniejszy model dla którego  $C_{\tilde{p}}$  jest mniejsza niż  $2\tilde{p}$ , lub model o najmniejszym  $C_{\tilde{p}}$ .



# Table of Contents

## 1 Problem wyboru modelu

## 2 Diagnostyka

- Partial regression plots
- Macierz H w diagnostyce modelu
- Miara DFFITS
- Odległość Cook'a
- Miara DFBETAS
- Miary Variance inflation factor oraz Tolerance

# Partial regression plots

W statystyce wykresy typu **partial regression plot (added variable plots, adjusted variable plots, individual coefficient plots)** ukazują wpływ jaki wywiera dodanie nowej zmiennej objaśniającej  $\tilde{X}_i$  do modelu, który już zawiera kilka zmiennych niezależnych.

W regresji liniowej prostej wykres rozrzutu między  $X$  i  $Y$  dobrze opisuje ich wzajemną relację. Jednakże, gdy mamy kilka zmiennych objaśniających sytuacja staje się bardziej skomplikowana. Wciąż można wytworzyć wykresy rozrzutu  $X_i$  vs  $Y$  (dla każdego  $X_i$ ), ale wykresy te nie biorą pod uwagę wpływu pozostałych  $X$ -ów na model.

Tzw. wykresy partial regression plots wypełniają tę lukę, opisując relację  $X_i$  vs  $Y$  z uwzględnieniem wpływu pozostałych  $X$ -ów na model.

# Partial regression plots

Konstrukcja wykresu partial regression plot dla zmiennej  $X_i$ :

- oblicz wektor residuów  $e^{(Y)}$  dla modelu liniowego w którym zmienną objaśnianą jest  $Y$  a zmiennymi objaśniającymi są wszystkie  $X$ -y oprócz  $X_i$ ,
- oblicz wektor residuów  $e^{(X_i)}$  dla modelu liniowego w którym zmienną objaśnianą jest  $X_i$  a zmiennymi objaśniającymi są wszystkie  $X$ -y oprócz  $X_i$ ,
- Stwórz wykres rozrzutu  $e^{(X_i)}$  vs.  $e^{(Y)}$

Ponieważ z definicji wektor residuów opisuje to czego nie wyjaśniły zmienne objaśniające, zatem wykres rozrzutu  $e^{(X_i)}$  vs.  $e^{(Y)}$  opisuje relację między  $X_i$ , a  $Y$  po uwzględnieniu wpływu pozostałych  $X$ -ów.

Zwykle powyższe wykresy tworzone są dla każdej zmiennej  $X_i$ .

Własności wykresu partial regression plot dla zmiennej  $X_i$ :

- Jeżeli na wykresie  $e^{(X_i)}$  vs.  $e^{(Y)}$  nie obserwowana jest żadna wyraźna struktura, wskazuje to na fakt, iż zmienna  $X_i$  nie wnosi do modelu istotnej informacji ponad to co objaśniły pozostałe  $X$ -y,
- jeżeli obserwujemy relację liniową (o wsp. kierunkowym różnym od zera) wskazuje to na fakt, iż zmienna  $X_i$  wnosi dodatkową informację do modelu,
- przy pomocy wykresu  $e^{(X_i)}$  vs.  $e^{(Y)}$  możemy wykrywać odstępstwa od założeń modelu np.: obserwacje odstające, brak liniowej relacji, brak stałości wariancji itp.

# Wektor studentyzowanych residuów

Do analizy założeń modelu liniowego stosowaliśmy do tej pory residua postaci  $e_i = Y_i - \hat{Y}_i$ . Wiemy, że pochodzą one z rozkładu  $e \sim N(0, \sigma^2(\mathbb{I} - H))$ . Łatwo zauważyć, że wariancja każdego elementu wektora  $e$  zależy od własności macierzy  $H$ :

$$\text{var}(e_i) = \sigma^2(1 - H_{ii})$$

i w konsekwencji mogą one być różne. Z drugiej strony błędy losowe  $\epsilon$  (których predyktorem są residua  $e$ ), mają równe wariancje  $\epsilon \sim N(0, \sigma^2\mathbb{I})$ . Z tego względu często w analizie założeń modelu stosuje się tzw. **studentyzowane residua** postaci:

$$\tilde{e}_i = \frac{Y_i - \hat{Y}_i}{\sqrt{\hat{\sigma}^2(1 - H_{ii})}}$$

gdzie  $\hat{\sigma}^2$  jest pewnym estymatorem parametru  $\sigma^2$ .

# Wewnętrzna i zewnętrzna studentyzacja residuów

W literaturze można spotykać dwa rodzaje studentyzowanych residuów:

- residua studentyzowane wewnętrznie (standaryzowane, studentized residuals),
- residua studentyzowane zewnętrznie (studentized deleted residuals).

W residuach studentyzowanych wewnętrznie model konstruowany jest klasycznie, czyli przy użyciu wszystkich obserwacji ("wewnętrzny" – zawierający  $Y_i$  oraz wiersz w mac. planu stowarzyszony z  $Y_i$ ). Na podstawie tak otrzymanego modelu wyznaczana jest wartość  $\hat{Y}_i$  oraz estymator wariancji postaci:  $\hat{\sigma}^2 = s^2$  i w konsekwencji:

$$\tilde{e}_i = \frac{Y_i - \hat{Y}_i}{\sqrt{\hat{\sigma}^2(1 - H_{ii})}} = \frac{e_i}{\sqrt{s^2(1 - H_{ii})}} = \frac{e_i / \sqrt{\sigma^2(1 - H_{ii})}}{\sqrt{s^2/\sigma^2}}$$

Na podstawie wcześniejszych wykładów można by podejrzewać że  $\tilde{e}_i$  ma rozkład studenta ( $e_i / \sqrt{\sigma^2(1 - H_{ii})} \sim N(0, 1)$ ;  $s^2/\sigma^2 \sim \chi^2_{n-p}/(n-p)$ ). Okazuje się jednak, że licznik i mianownik nie są niezależnymi zmiennymi losowymi. Z tego względu rozkład  $\tilde{e}_i$  nie jest rozkładem studenta.

# Wewnętrzna i zewnętrzna studentyzacja residuów

Źródłem zależności między licznikiem i mianownikiem jest wykorzystanie  $Y_i$  zarówno w liczniku ( $e_i = Y_i - \hat{Y}_i$ ) i w mianowniku (w statystyce  $s^2$ ). Z poprzednich wykładów wiemy, że dla **nowej (niezależnej)** obserwacji  $\tilde{Y}_h$  statystyki  $\tilde{Y}_h - \hat{Y}_h$  oraz  $s^2$  są niezależne, co stanowi brakujący element w uzyskaniu rozkładu studenta.

# Wewnętrzna i zewnętrzna studentyzacja residuów

Źródłem zależności między licznikiem i mianownikiem jest wykorzystanie  $Y_i$  zarówno w liczniku ( $e_i = Y_i - \hat{Y}_i$ ) i w mianowniku (w statystyce  $s^2$ ). Z poprzednich wykładów wiemy, że dla **nowej (niezależnej)** obserwacji  $\tilde{Y}_h$  statystyki  $\tilde{Y}_h - \hat{Y}_h$  oraz  $s^2$  są niezależne, co stanowi brakujący element w uzyskaniu rozkładu studenta.

Z tego względu powstała modyfikacja zwana **zewnętrzną studentyzacją residuów**. Modyfikacja ta polega na tym że do wyznaczenia  $\tilde{e}_i$  korzystamy z modelu skonstruowanego z pominięciem w danych wartości  $Y_i$  oraz wiersza w mac. planu stowarzyszonego z  $Y_i$ . Aby podkreślić to, że wyłączona została  $i$ -ta obserwacja z danych będziemy stosować w dolnym indeksie znacznik " $(i)$ ".

Wówczas:

$$\tilde{e}_i = \frac{Y_i - \hat{Y}_{(i)i}}{\sqrt{s_{(i)}^2(1 - H_{(i)ii})}}$$

Ponieważ  $Y_i$  jest "**nową**" (**niezależną**) obserwacją dla modelu zbudowanego z pominięciem  $i$ -tej obserwacji, zatem licznik i mianownik są niezależne i  $\tilde{e}_i \sim t_{n-1-p}$  (dodatkowe " $-1$ " w liczbie stopni swobody, wynika z faktu że model zbudowano na  $n - 1$  obserwacjach). Warto zwrócić uwagę że do wyznaczenia wektora zewnętrznych studentyzowanych residuów musimy stworzyć  $n$  modeli.



Z praktycznego punktu widzenia zwykle obie wersje studentyzowanych residuów mają bardzo podobne własności i mogą być pomocne przy poszukiwaniu:

- obserwacji odstających,
- obserwacji wpływowych,
- odstępstw od założeń dotyczących błędów  $\epsilon$ .

W R są dwie funkcje za pomocą, których można wyznaczyć odpowiednio wewnętrzne i zewnętrzne residua studentyzowane: *rstandard(reg1)*, *rstudent(reg1)* (gdzie *reg1* = *lm(y ~ x, data)* – skonstruowany model).

# Macierz $H$ w diagnostyce modelu

Na jednym z wcześniejszych wykładów wprowadziliśmy oznaczenie na macierz rzutu ortogonalnego na przestrzeń  $Lin(\mathbb{X})$ :

$$H = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$$

Miała ona szczególne znaczenie przy wyznaczaniu predykcji wektora odpowiedzi  $\hat{Y} = HY$  i wektora (niestudentyzowanego) residuów  $e = (\mathbb{I} - H)Y$ .

# Macierz $H$ w diagnostyce modelu

Na jednym z wcześniejszych wykładów wprowadziliśmy oznaczenie na macierz rzutu ortogonalnego na przestrzeń  $Lin(\mathbb{X})$ :

$$H = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$$

Miała ona szczególne znaczenie przy wyznaczaniu predykcji wektora odpowiedzi  $\hat{Y} = HY$  i wektora (niestudentyzowanego) residuów  $e = (\mathbb{I} - H)Y$ .

W kontekście diagnostyki modelu istotne są elementy na diagonalu macierzy  $H$ . Stanowią one miarę wpływu obserwacji  $Y_i$  na predykcję  $\hat{Y}_i$ , np. dla  $Y_1$ :

$$\hat{Y}_1 = \underline{H_{11}}Y_1 + H_{12}Y_2 + H_{13}Y_3 + \dots + H_{1n}Y_n$$

W tym kontekście czasami  $H_{ii}$  określane jest **wagą**  $i$ -tej obserwacji.

# Macierz $H$ w diagnostyce modelu

Z faktu, iż  $H$  jest macierzą rzutu ortogonalnego wynika, że:

- $0 \leq H_{ii} \leq 1$
- $\sum_{i=1}^n H_{ii} = \text{Tr}(H) = p$

Duże wartości  $H_{ii}$  sugerują, że  $i$ -ta obserwacja jest mocno odległa od centrum  $X$ -ów.

Obserwacje o wartości  $H_{ii}$  dalekiej od średniej wartości  $p/n$  należy szczególnie przebadać. Potencjalnie mogą one być obserwacjami odstającymi lub wpływowymi!

Do badania wpływu obserwacji  $Y_i$  na predykcję  $\hat{Y}_i$  można posłużyć się tzw. **miarą DFFITS dla  $i$ -tej obserwacji**, która ma postać:

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{(i)i}}{\sqrt{s_{(i)}^2 H_{ii}}}$$

Z definicji łatwo zauważyć, że DFFITS dla  $i$ -tej obserwacji jest standaryzowaną różnicą pomiędzy predykcjami wartości  $Y_i$  uzyskanymi na podstawie dwóch modeli skonstruowanych na danych, odpowiednio, z/bez obserwacji  $Y_i$ .

Do badania wpływu obserwacji  $Y_i$  na predykcję  $\hat{Y}_i$  można posłużyć się tzw. **miarą DFFITS dla  $i$ -tej obserwacji**, która ma postać:

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{(i)i}}{\sqrt{s_{(i)}^2 H_{ii}}}$$

Z definicji łatwo zauważyć, że DFFITS dla  $i$ -tej obserwacji jest standaryzowaną różnicą pomiędzy predykcjami wartości  $Y_i$  uzyskanymi na podstawie dwóch modeli skonstruowanych na danych, odpowiednio, z/bez obserwacji  $Y_i$ .

Naturalnym jest oczekiwanie, że predykcje  $\hat{Y}_i$  i  $\hat{Y}_{(i)i}$  będą przyjmowały podobne wartości. Odwrotna sytuacja implikuje znaczący wpływ obserwacji  $Y_i$  na obie predykcje i dużą (co do modułu) wartość statystyki  $DFFITS_i$ . Takim punktom warto się dokładnie przyjrzeć!

W literaturze przyjmuje się, że dodatkowej analizie należy poddać te obserwacje dla których  $|DFFITS_i| > 2\sqrt{p/n}$ .

# Odległość Cook'a (Cook's distance)

Do badania wpływu obserwacji  $Y_i$  na cały wektor predykcji  $\hat{Y}$  można posłużyć się tzw. **odległością Cook'a dla  $i$ -tej obserwacji**, która ma postać:

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

Z definicji łatwo zauważyć, że DFFITS dla  $i$ -tej obserwacji jest standaryzowaną różnicą pomiędzy predykcjami wektora  $Y$  uzyskanymi na podstawie dwóch modeli skonstruowanych na danych, odpowiednio, z/bez obserwacji  $Y_i$ .

# Odległość Cook'a (Cook's distance)

Do badania wpływu obserwacji  $Y_i$  na cały wektor predykcji  $\hat{Y}$  można posłużyć się tzw. **odległością Cook'a dla  $i$ -tej obserwacji**, która ma postać:

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

Z definicji łatwo zauważyć, że DFFITS dla  $i$ -tej obserwacji jest standaryzowaną różnicą pomiędzy predykcjami wektora  $Y$  uzyskanymi na podstawie dwóch modeli skonstruowanych na danych, odpowiednio, z/bez obserwacji  $Y_i$ .

Naturalnym jest oczekiwanie, że predykcje  $\hat{Y}$  i  $\hat{Y}_{(i)}$  będą przyjmowały podobne wartości. Odwrotna sytuacja implikuje znaczący wpływ obserwacji  $Y_i$  na obie predykcje i dużą (co do modułu) odległość Cook'a dla  $i$ -tej obserwacji.

W literaturze przyjmuje się, że dodatkowej analizie należy poddać te obserwacje dla których  $|D_i| > 1$ .



Do badania wpływu obserwacji  $Y_i$  na estymację parametru  $\beta_k$  można posłużyć się tzw. **miarą DFBETA dla i-tej obserwacji**, która dla parametru  $\beta_k$  ma postać:

$$DFBETA_k = \frac{\hat{\beta}_k - \hat{\beta}_{(i)k}}{s_{(i)}(\hat{\beta}_{(i)k})}; \quad DFBETAS = (DFBETA_0, \dots, DFBETA_{p-1})$$

gdzie  $\hat{\beta}_k$  i  $\hat{\beta}_{(i)k}$  są estymatorem parametru  $\beta_k$  uzyskanymi na podstawie dwóch modeli skonstruowanych na danych, odpowiednio, z/bez obserwacji  $Y_i$ . Ponadto,  $s_{(i)}(\hat{\beta}_{(i)k})$  jest estymatorem odchylenia standardowego estymatora  $\hat{\beta}_{(i)k}$ .

Do badania wpływu obserwacji  $Y_i$  na estymację parametru  $\beta_k$  można posłużyć się tzw. **miarą DFBETA dla i-tej obserwacji**, która dla parametru  $\beta_k$  ma postać:

$$DFBETA_k = \frac{\hat{\beta}_k - \hat{\beta}_{(i)k}}{s_{(i)}(\hat{\beta}_{(i)k})}; \quad DFBETAS = (DFBETA_0, \dots, DFBETA_{p-1})$$

gdzie  $\hat{\beta}_k$  i  $\hat{\beta}_{(i)k}$  są estymatorem parametru  $\beta_k$  uzyskanymi na podstawie dwóch modeli skonstruowanych na danych, odpowiednio, z/bez obserwacji  $Y_i$ . Ponadto,  $s_{(i)}(\hat{\beta}_{(i)k})$  jest estymatorem odchylenia standardowego estymatora  $\hat{\beta}_{(i)k}$ .

Naturalnym jest oczekiwanie, że estymatory  $\hat{\beta}_k$  i  $\hat{\beta}_{(i)k}$  będą przyjmowały podobne wartości. Odwrotna sytuacja implikuje znaczący wpływ obserwacji  $Y_i$  na oba estymatory, dużą (co do modułu) wartość statystyki  $DFBETA_k$  i znaczący wpływ na estymację parametru  $\beta_k$ .

W literaturze przyjmuje się, że dodatkowej analizie należy poddać te obserwacje, dla których  $|DFBETA_k| > 2/\sqrt{n}$ .

# Miary Variance inflation factor oraz Tolerance

Wspomnieliśmy, że poważnym problemem w regresji liniowej wielorakiej jest zjawisko **multikolinearności**. Do badania wielkości tego zjawiska można posłużyć się tzw. miarą **Variance inflation factor (VIF)**. *VIF* dla  $k$ -tej zmiennej objaśniającej bada, w jakim stopniu zmienna  $X_k$  objaśniana jest przez pozostałe zmienne objaśniające  $X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_{p-1}$ .

# Miary Variance inflation factor oraz Tolerance

Wspomnieliśmy, że poważnym problemem w regresji liniowej wielorakiej jest zjawisko **multikolinearności**. Do badania wielkości tego zjawiska można posłużyć się tzw. miarą **Variance inflation factor (VIF)**.  $VIF$  dla  $k$ -tej zmiennej objaśniającej bada, w jakim stopniu zmienna  $X_k$  objaśniana jest przez pozostałe zmienne objaśniające  $X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_{p-1}$ .

Variance inflation factor można wyznaczyć w następujący sposób:

- 1 Skonstruuj model liniowy, w którym zmienna  $X_k$  jest zmienną objaśnianą ("  $\tilde{Y}$  "), a pozostałe zmienne  $X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_{p-1}$  zmiennymi objaśniającymi:

$$X_{ik} = \beta_0 + \sum_{j \neq k} X_{ij} \beta_j + \epsilon_i \quad i = 1, \dots, n$$

- 2 wyznacz współczynnik determinacji  $R_k^2$  dla powyższego modelu,
- 3 Variance inflation factor dla zmiennej  $X_k$  jest postaci:

$$VIF_k = (1 - R_k^2)^{-1}$$

# Miary Variance inflation factor oraz Tolerance

Duże wartości  $VIF_k$  wskazują na bardzo silną korelację między  $X_k$  i pewną kombinacją liniową pozostałych zmiennych objaśniających. Implikuje to występowanie zjawiska multikolinearności. Zwykle obliczamy  $VIF_k$  dla każdej zmiennej objaśniającej  $X_k$ . Przyjmuje się, że sytuacja, w której  $VIF_k$  ma wartość większą niż 10, wskazuje na poważny problem z multikolinearnością.

# Miary Variance inflation factor oraz Tolerance

Duże wartości  $VIF_k$  wskazują na bardzo silną korelację między  $X_k$  i pewną kombinacją liniową pozostałych zmiennych objaśniających. Implikuje to występowanie zjawiska multikolinearności. Zwykle obliczamy  $VIF_k$  dla każdej zmiennej objaśniającej  $X_k$ . Przyjmuje się, że sytuacja, w której  $VIF_k$  ma wartość większą niż 10, wskazuje na poważny problem z multikolinearnością.

Czasami zamiast Variance inflation factor stosuje się miarę zwaną Tolerancją (Tolerance) zdefiniowaną jako odwrotność VIF:

$$Tol_k = 1/VIF_k$$

Naturalnie w przypadku tolerancji na problem z multikolinearnością wskazywać będą wartości mniejsze od 0.1.