Modele liniowe

Michał Kos

Uniwersytet Wrocławski

Plan wykładu

Table of Contents

Standaryzacja danych

W regresji liniowej często pojawia się sytuacja, w której regresory są wyrażone w różnych jednostkach. Taka sytuacja powoduje niejednokrotnie brak możliwości wzajemnego porównania wpływu regresorów.

Przykład

Załóżmy, że chcemy przewidzić obwód w pasie pacjenta (Y) na podstawie wzrostu wyrażonego w centymetrach (X_1) i wagi wyrażonej w kilogramach (X_2). Załóżmy, że uzyskany model jest postaci:

$$\hat{Y} = 1 + 2X_1 + 3X_2$$

Na podstawie modelu można by wysnuć wniosek, że większy wpływ na zmienną Y ma waga ($2 < 3$). Jednakże gdybyśmy wyrazili wzrost w metrach $\tilde{X}_1 = X_1/100$, wówczas ten sam model by miał postać:

$$\hat{Y} = 1 + 200(X_1/100) + 3X_2 = 1 + 200\tilde{X}_1 + 3X_2$$

i wniosek byłby odwrotny ($200 > 3$). Widzimy zatem, że w takiej sytuacji, aby móc porównywać wpływ regresorów potrzebna jest standaryzacja danych.

Standaryzacja danych

W przypadku regresji liniowej standaryzacji dokonuje się wprowadzając następujące przekształcenie zmiennych niezależnych:

$$\tilde{X}_i = s(Y)X_i/s(X_i) \quad i = 1, ..., p-1$$

W wyniku powyższej standaryzacji wariancja każdej z nowych zmiennych $\tilde{X}_i$ jest taka sama. W konsekwencji wzrost o dowolną ustaloną wartość K dla każdej zmiennej $\tilde{X}_i$ jest tak samo prawdopodobny.
Jak wygląda model wyrażony w języku nowych zmiennych:

$$\hat{Y} = ... + \hat{\beta}_i X_i + ... = ... + \left(\hat{\beta}_i s(X_i)/s(Y)\right)(s(Y)X_i/s(X_i)) + ... =$$

$$= ... + \hat{\tilde{\beta}}_i \tilde{X}_i + ...$$

gdzie $\hat{\tilde{\beta}}_i = \hat{\beta}_i s(X_i)/s(Y)$.
Jak interpretować wpływ nowych zmiennych:
Wzrost o wartość 1 dowolnej zmiennej $\tilde{X}_i$ powoduje wzrost $\hat{Y}$ o $\hat{\tilde{\beta}}_i$. Ponieważ wzrost te są tak samo prawdopodobne, zatem $\hat{\tilde{\beta}}_i$ są porównywalne i mówią o wzajemnej sile wpływu poszczególnych zmiennych $\tilde{X}_i$ (X_i).

Table of Contents

Multikolinearność

W kontekście modeli liniowych o zjawisku multikolinearności mówimy, gdy macierz $\mathbb{X}'\mathbb{X}$ jest bliska macierzy singularnej. Miarą tej bliskości może być np. najmniejsza wartość własna macierzy $\mathbb{X}'\mathbb{X}$ (λ_{min}) (dla macierzy singularnej $\lambda_{min} = 0$). Pojawiają się wówczas 2 problemy:

- Obliczeniowy: pomimo tego, że macierz $\mathbb{X}'\mathbb{X}$ formalnie jest odwracalna to ze względu na skończoną dokładność obliczeniową komputerów, wyznaczenie macierzy $(\mathbb{X}'\mathbb{X})^{-1}$ stanowi duże wyzwanie, co wpływa na dokładność np. $\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$
- Statystyczny: ponieważ wartości własne z macierzy A^{-1} są odwrotnościami wartości własnych macierzy A, zatem jeżeli $\lambda_{min}(\mathbb{X}'\mathbb{X})$ jest bliska 0 to $\lambda_{max}((\mathbb{X}'\mathbb{X})^{-1}) = \lambda^{-1}_{min}(\mathbb{X}'\mathbb{X})$ jest bardzo duża. Przekłada się to na duże wartości wariancji estymatorów $\hat{\beta}_i$ ($\hat{\beta} \sim N(\beta, \sigma^2(\mathbb{X}'\mathbb{X})^{-1})$) (ślad macierzy $tr((\mathbb{X}'\mathbb{X})^{-1}) = \sum_{i=1}^{p} \lambda_i((\mathbb{X}'\mathbb{X})^{-1})$).

Okazuje się zatem, że zjawisko multikolinearności wpływa negatywnie zarówno na własności statystyczne estymatora $\hat{\beta}$ jak i dokładność jego wyznaczenia.

Multikolinearność

Problem multikolinearności jest ściśle związany z własnościami macierzy eksperymentu $\mathbb{X}$. Można pokazać, że zjawisko to pojawia się, gdy kolumny macierzy planu są prawie liniowo zależne.

Multikolinearność

Problem multikolinearności jest ściśle związany z własnościami macierzy eksperymentu $\mathbb{X}$. Można pokazać, że zjawisko to pojawia się, gdy kolumny macierzy planu są prawie liniowo zależne.
Jak wiemy z algebry liniowej zbiór wektorów $\{v_1, v_2, ..., v_{p-1}\}$ (kolumn macierzy $\mathbb{X}$) jest liniowo zależny gdy:

$$\exists i \ : \ v_i = \sum_{j \neq i} \gamma_j v_j$$

gdzie $\gamma_j \in \mathbb{R}$ to waga wektora v_j. Innymi słowy wektor v_i jest kombinacją liniową pozostałych wektorów.

Multikolinearność

Problem multikolinearności jest ściśle związany z własnościami macierzy eksperymentu $\mathbb{X}$. Można pokazać, że zjawisko to pojawia się, gdy kolumny macierzy planu są prawie liniowo zależne.
Jak wiemy z algebry liniowej zbiór wektorów $\{v_1, v_2, ..., v_{p-1}\}$ (kolumn macierzy $\mathbb{X}$) jest liniowo zależny gdy:

$$\exists i \ : \ v_i = \sum_{j \neq i} \gamma_j v_j$$

gdzie $\gamma_j \in \mathbb{R}$ to waga wektora v_j. Innymi słowy wektor v_i jest kombinacją liniową pozostałych wektorów.
Powyższa zależność implikuje fakt, iż:

$$cor(v_i, \sum_{j \neq i} \gamma_j v_j) = 1$$

W powyższym języku wyrażana jest prawie liniowa zależność:
Problem multikolinearności występuje, gdy:

$$\exists i \ cor(v_i, \sum_{j \neq i} \gamma_j v_j) \approx 1$$

Multikolinearność

Przykład

Wpływ korelacji pomiędzy dwoma kolumnami w macierzy $\mathbb{X}'\mathbb{X}$ na wielkość maksymalnej wartości własnej macierzy $(\mathbb{X}'\mathbb{X})^{-1}$:

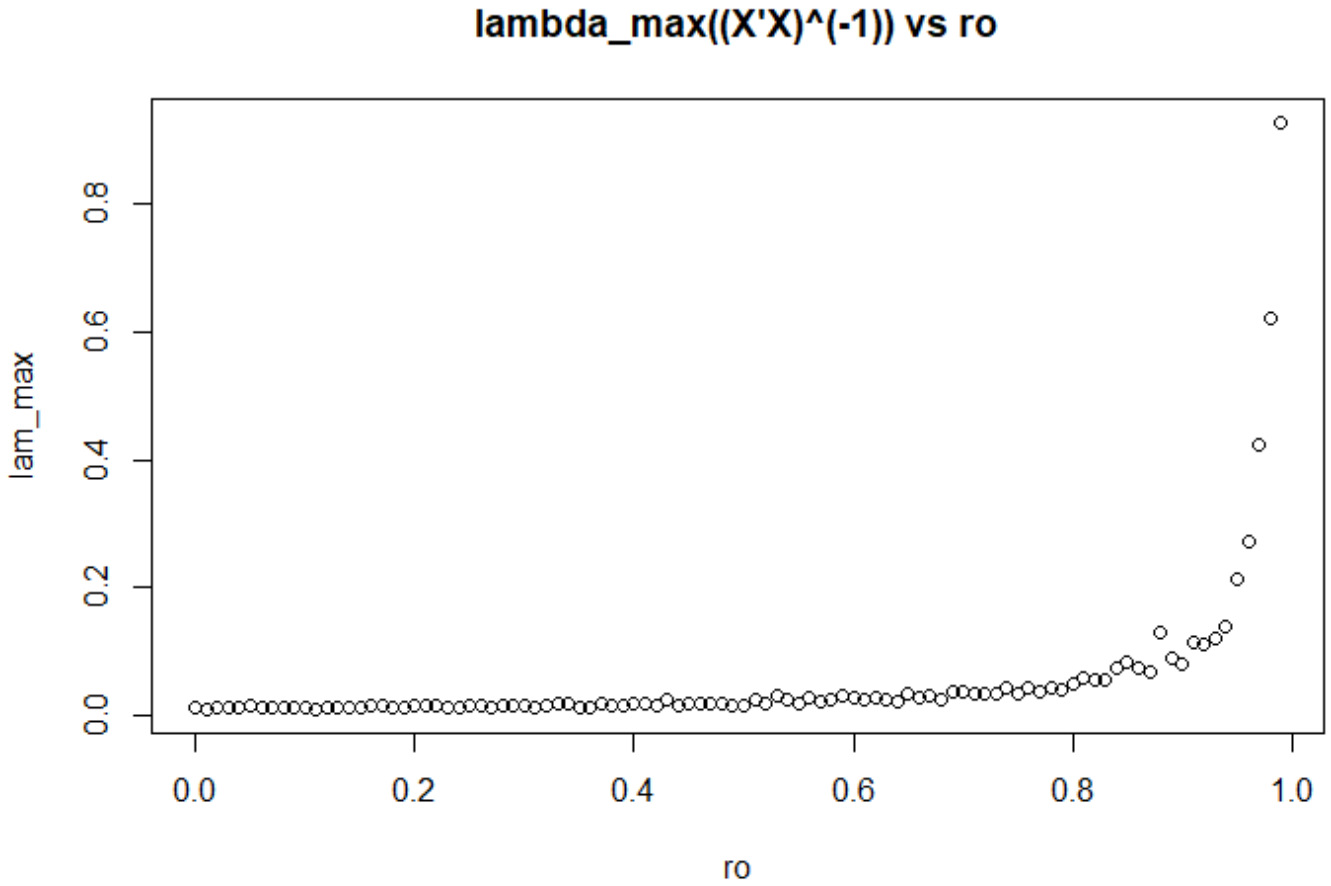


Table of Contents

Modele z interakcją

W sytuacji, w której mamy kilka zmiennych objaśniających, należy rozważyć przypadek, w którym wpływ jednej ze zmiennych objaśniających na Y zależy od wartości innych zmiennych objaśniających.

Modele z interakcją

W sytuacji, w której mamy kilka zmiennych objaśniających, należy rozważyć przypadek, w którym wpływ jednej ze zmiennych objaśniających na Y zależy od wartości innych zmiennych objaśniających.

Do modelowania takich efektów służą tzw. **modele liniowe z interakcją**

Rozważmy dwa najprostsze przypadki:

- dwie zmienne objaśniające: ciągła i binarna,
- dwie ciągłe zmienne objaśniające,

Gruntowne zrozumienie tych przypadków jest kluczowe, gdyż idea modeli z interakcją w sposób naturalny uogólnia się na więcej zmiennych objaśniających.

Modele z interakcją – dwie zmienne objaśniające: ciągła i binarna

Załóżmy że dysponujemy dwiema zmiennymi objaśniającymi:

- binarną X_1; przyjmującą dwie wartości 0 i 1, w zalezności od tego czy występuje pewne zdarzenie,
- ciągłą X_2.

Standardowy model dla dwóch zmiennych objaśniających jest postaci:

$$Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \epsilon_i \quad i = 1, ..., n$$

Modele z interakcją – dwie zmienne objaśniające: ciągła i binarna

Załóżmy że dysponujemy dwiema zmiennymi objaśniającymi:

- binarną X_1; przyjmującą dwie wartości 0 i 1, w zalezności od tego czy występuje pewne zdarzenie,
- ciągłą X_2.

Standardowy model dla dwóch zmiennych objaśniających jest postaci:

$$Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \epsilon_i \quad i = 1, ..., n$$

Rozważamy jednak sytuację, w której podejrzewamy, że wpływ zmiennej X_2 (odp. X_1) na Y zależy od wartości zmiennej X_1 (X_2). Zjawisko takie analizowane jest za pomocą modeli z interakcją poprzez dodanie do modelu dodatkowego składnika odpowiadającego iloczynowi zmiennych objaśnianych: X_1X_2. Model z interakcją ma postać:

$$Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + X_{i1}X_{i2}\beta_3 + \epsilon_i \quad i = 1, ..., n$$

Modele z interakcją - ciągła i binarna zmienna objaśniająca

Bardzo pouczające jest rozpatrzenie wartości oczekiwanych z obu modeli:

$$Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \epsilon_i; \qquad Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + X_{i1}X_{i2}\beta_3 + \epsilon_i$$

dla możliwych wartości zmiennej binarnej X_1:

	$X_1 = 0$	$X_1 = 1$
model bez interakcji	$E(Y_i) = \beta_0 + X_{i2}\beta_2$	$E(Y_i) = (\beta_0 + \beta_1) + X_{i2}\beta_2$
model z interakcją	$E(Y_i) = \beta_0 + X_{i2}\beta_2$	$E(Y_i) = (\beta_0 + \beta_1) + X_{i2}(\beta_2 + \beta_3)$

Modele z interakcją - ciągła i binarna zmienna objaśniająca

Bardzo pouczające jest rozpatrzenie wartości oczekiwanych z obu modeli:

$$Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \epsilon_i; \qquad Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + X_{i1}X_{i2}\beta_3 + \epsilon_i$$

dla możliwych wartości zmiennej binarnej X_1:

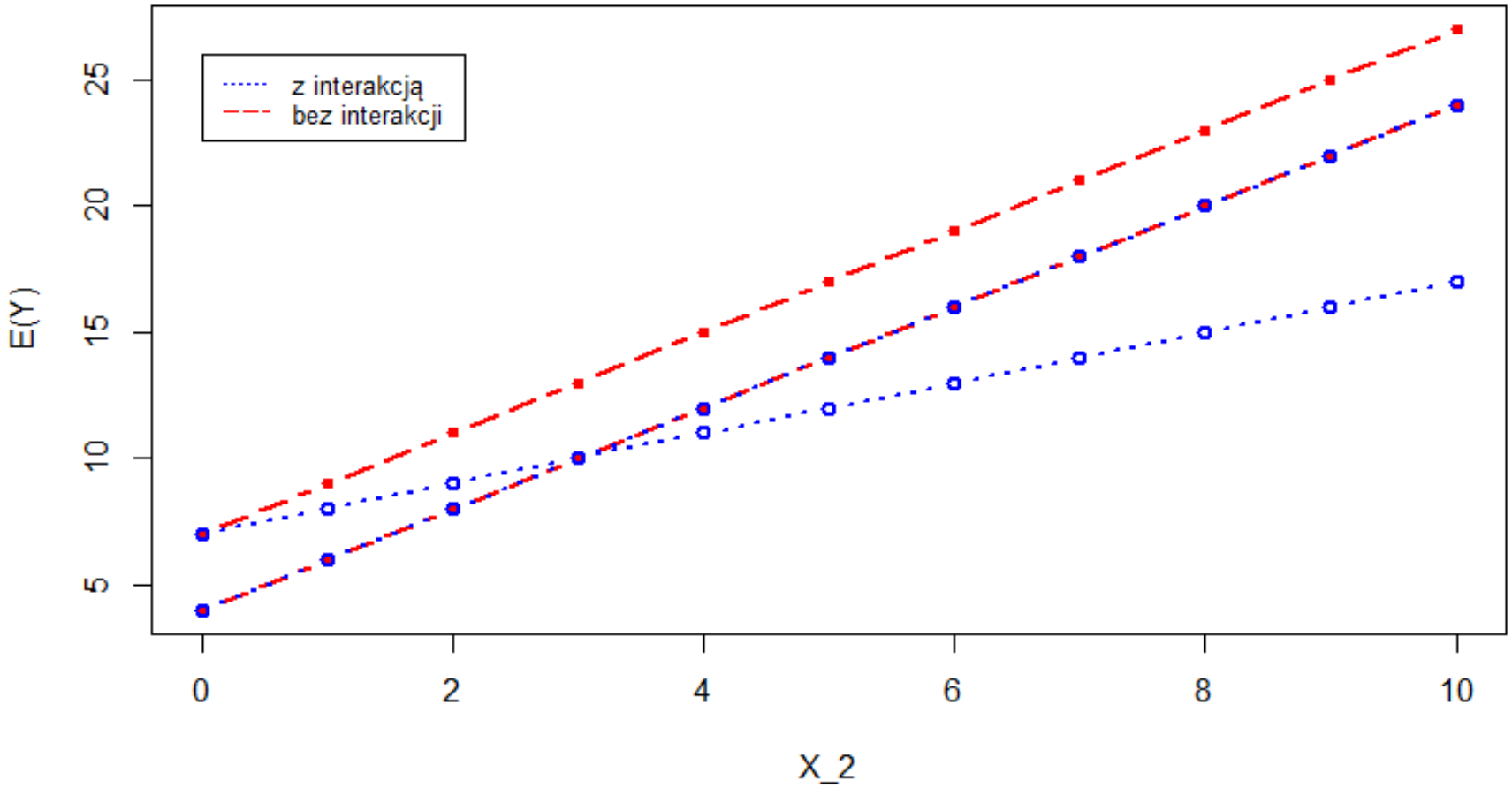
	$X_1 = 0$	$X_1 = 1$
model bez interakcji	$E(Y_i) = \beta_0 + X_{i2}\beta_2$	$E(Y_i) = (\beta_0 + \beta_1) + X_{i2}\beta_2$
model z interakcją	$E(Y_i) = \beta_0 + X_{i2}\beta_2$	$E(Y_i) = (\beta_0 + \beta_1) + X_{i2}(\beta_2 + \beta_3)$

Z powyższej tabeli można odczytać kilka wniosków:

- jeżeli $X_1 = 0$ to oba modele oscylują wokół dokładnie tej samej prostej $E(Y_i) = \beta_0 + X_{i2}\beta_2$,
- w modelu bez interakcji modyfikacji ulega wyłącznie Intercept, a współczynnik kierunkowy prostych $E(Y_i|X_1 = 0)$ i $E(Y_i|X_1 = 1)$ jest taki sam (proste są równoległe). Fakt ten odzwierciedla **niezależność** przyrostów Y dla każdej zm. objaśniającej,
- w modelu z interakcją, gdy $X_1 = 1$, modyfikacji ulega dodatkowo współczynnik kierunkowy. Widzimy zatem, że przyrosty Y dla zmiennej X_2 **zależą** od wartości zmiennej X_1.

Modele z interakcją - ciągła i binarna zmienna objaśniająca

Na poniższym rysunku przedstawiono zachowanie $E(Y)$ od X_2 w zależności od wartości zmiennej objaśniającej $X_1 \in \{0, 1\}$ oraz typu modelu (z/bez interakcji). $[\beta_0 = 4, \beta_1 = 3, \beta_2 = 2, \beta_3 = -1]$



Modele z interakcją - dwie ciągłe zmienne objaśniające

Przykład z jedną zmienną binarną i jedną zmienną ciągłą łatwo uogólnić na sytuację dwóch zmiennych ciągłych. Model z interakcją zdefiniowany jest w taki sam sposób:

$$Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + X_{i1}X_{i2}\beta_3 + \epsilon_i$$

Przy ustalonej wartości zmiennej X_1, zachodzi:

$$E(Y|X_1 = a) = (\beta_0 + a\beta_1) + X_{i2}(\beta_2 + a\beta_3)$$

i symetrycznie dla ustalonej wartości X_2:

$$E(Y|X_2 = b) = (\beta_0 + b\beta_2) + X_{i1}(\beta_1 + b\beta_3)$$

Widzimy zatem że w obu przypadkach przyrosty Y dla określonej zmiennej X_i zależą od drugiej zmiennej objaśniającej.

Modele z interakcjami - wiele zmiennych objaśniających

Jeżeli mamy kilka zmiennych objaśniających wówczas model z interakcjami ma postać:

$$Y_i = \beta_0 + \sum_{j=1}^{p-1} X_{ij}\beta_j + \sum_{j=1}^{p-1}\sum_{k>j}^{p-1} X_{ij}X_{ik}\beta_{jk} + \epsilon_i$$

gdzie składnik $\sum_{j=1}^{p-1}\sum_{k>j}^{p-1} X_{ij}X_{ik}\beta_{jk}$ opisuje iloczyny wszystkich par zmiennych objaśniających, a $\beta_{jk} \in \mathbb{R}$ jest współczynnikiem opisującym wpływ iloczynu $X_{ij}X_{ik}$ na zmienną zależną Y.

Modele z interakcjami - wiele zmiennych objaśniających

Jeżeli mamy kilka zmiennych objaśniających wówczas model z interakcjami ma postać:

$$Y_i = \beta_0 + \sum_{j=1}^{p-1} X_{ij}\beta_j + \sum_{j=1}^{p-1}\sum_{k>j}^{p-1} X_{ij}X_{ik}\beta_{jk} + \epsilon_i$$

gdzie składnik $\sum_{j=1}^{p-1}\sum_{k>j}^{p-1} X_{ij}X_{ik}\beta_{jk}$ opisuje iloczyny wszystkich par zmiennych objaśniających, a $\beta_{jk} \in \mathbb{R}$ jest współczynnikiem opisującym wpływ iloczynu $X_{ij}X_{ik}$ na zmienną zależną Y.

Warto zauważyć, że liczba różnych iloczynów $X_{ij}X_{ik}$ wynosi $(p-1)(p-2)/2$ i przyrasta w tempie kwadratowym wraz ze wzrostem liczby zmiennych w modelu. Taka sytuacja sprzyja zjawisku nadmiernego dopasowania się modelu do danych (over-fitting).

Table of Contents

Problem wyboru modelu

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

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

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

Problem wyboru modelu

Poznaliśmy kilka statystycznych narzędzi umożliwiajacych **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 tą samą liczbę zmiennych, a R^2_{adj} działa w mało stabilny sposób.

Problem wyboru modelu

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.

AIC i BIC

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

$$\hat{\beta} = argmax_b \, (\ \log(\text{likelihood(b)}) - \text{kara za "duże" p})$$

Składnik $log(likelihood(b))$ 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 $\tilde{\mathbb{X}}$ (o l. kol. $\tilde{p}$) macierzy planu $\mathbb{X}$ wyznacz statystykę:
$$AIC(\tilde{\mathbb{X}}) = n \log\left(\frac{SSE(\tilde{\mathbb{X}})}{n}\right) + 2\tilde{p}$$
2. wybierz model $\tilde{\mathbb{X}}$ o najniższej wartości statystyka AIC.

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

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

Kryterium C_p Mallows'a

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

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

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

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

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 $C_{\tilde{p}}$.