Wstęp do uczenia maszynowego

Metody repróbkowania, wybór modelu

Ewa Szczurek + BW (modyfikacje)

bartek@mimuw.edu.pl
Instytut Informatyki
Uniwersytet Warszawski

23 kwietnia 2024

UNIWERSYTET
WARSZAWSKI

Metody repróbkowania

Metody repróbkowania

- Walidacja krzyżowa (*cross-validation*)
  - do oceny jakości modelu (poprzez szacowanie błędu testowego),
  - do wyboru optymalnego poziomu elastyczności (wariancja vs. obciążenie) modelu (wybór modelu; ang *model selection*)
- Bootstrap
  - do oceny niepewności estymatora parametru

Przypomnienie: podział danych

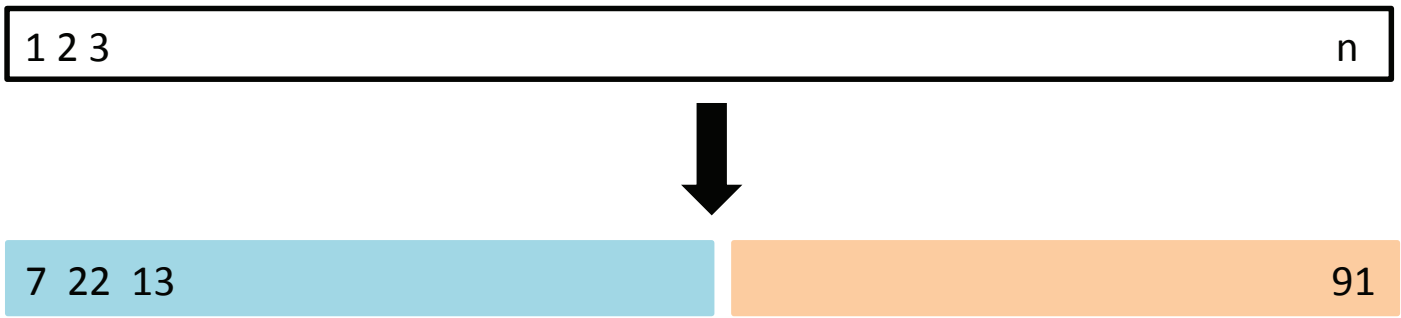
- **Dane treningowe**
  - $(x_{\text{Train}}, y_{\text{Train}})$ użyte do estymacji modelu $\hat{f}$
- **Dane walidacyjne (opcjonalne)**
  - $(x_{\text{Val}}, y_{\text{Val}})$ użyte do dobierania tzw. hiper-parametrów, np. k w kNN
- **Dane testowe**
  - $(x_{\text{Test}}, y_{\text{Test}})$ użyte do ewaluacji jak dobrze dobraliśmy $\hat{f}$
  - mają na celu przybliżać sutuacje, które pojawią się w zastosowaniu modelu, czyli zastosowanie $\hat{f}$ do x dla których nie będziemy znać y
  - uczenie modelu powinno nastąpić wyłącznie w oparciu o dane uczące (i ew. walidacyjne) **bez zaglądania do danych testowych**
- Istotniejsze jest minimalizowanie błędu na danych testowych niż na treningowych

Przypomnienie: podział danych

- W tym wykładzie zakładamy, że mamy już w ręku zbiór testowy
  - Dany dedykowany, niezależny od danych treningowych zbiór
  - Jeśli dedykowanego zbioru testowego nie ma, to wydzielamy go ze zbioru danych treningowych
- Dziś opowiemy, jak wydzielać zbiór walidacyjny, lub kilka zbiorów walidacyjnych, i jak się nimi posługiwać w uczeniu statycznym

Losowy zbiór walidacyjny

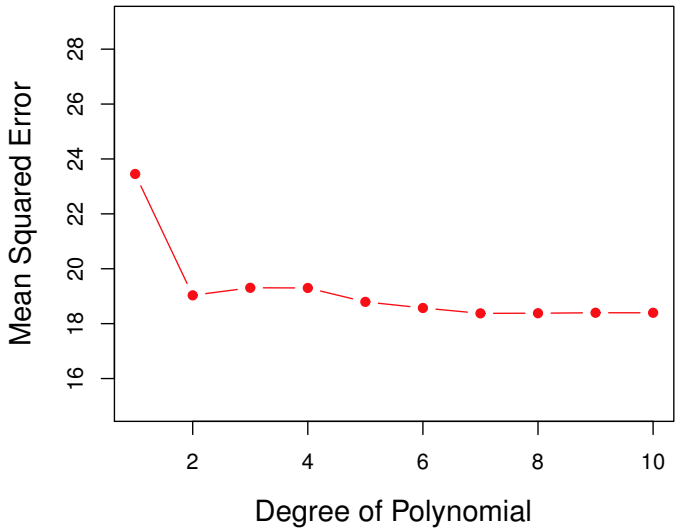
Dzielimy dane losowo na dwie części. Na jednej trenujemy model, a na drugiej szacujemy błąd.

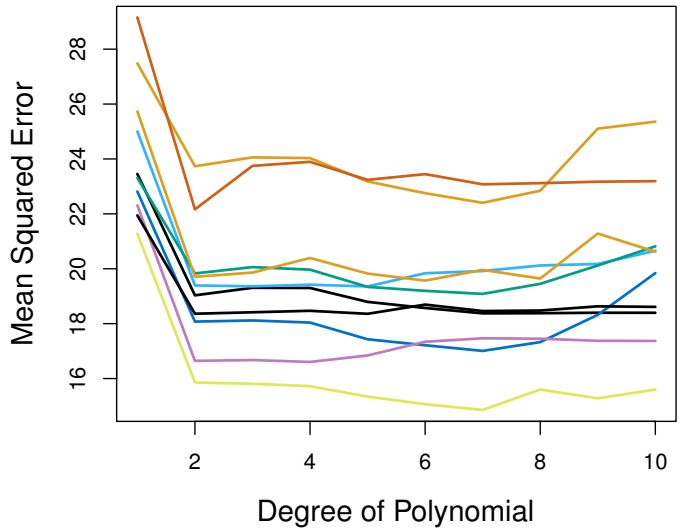


Niebieski: część treningowa; **Beżowy:** część walidacyjna.

Wyestymowany błąd testowy może wykazać dużą zmienność

Dane 'Auto': 'mpg' jako wielomian od 'horsepower'

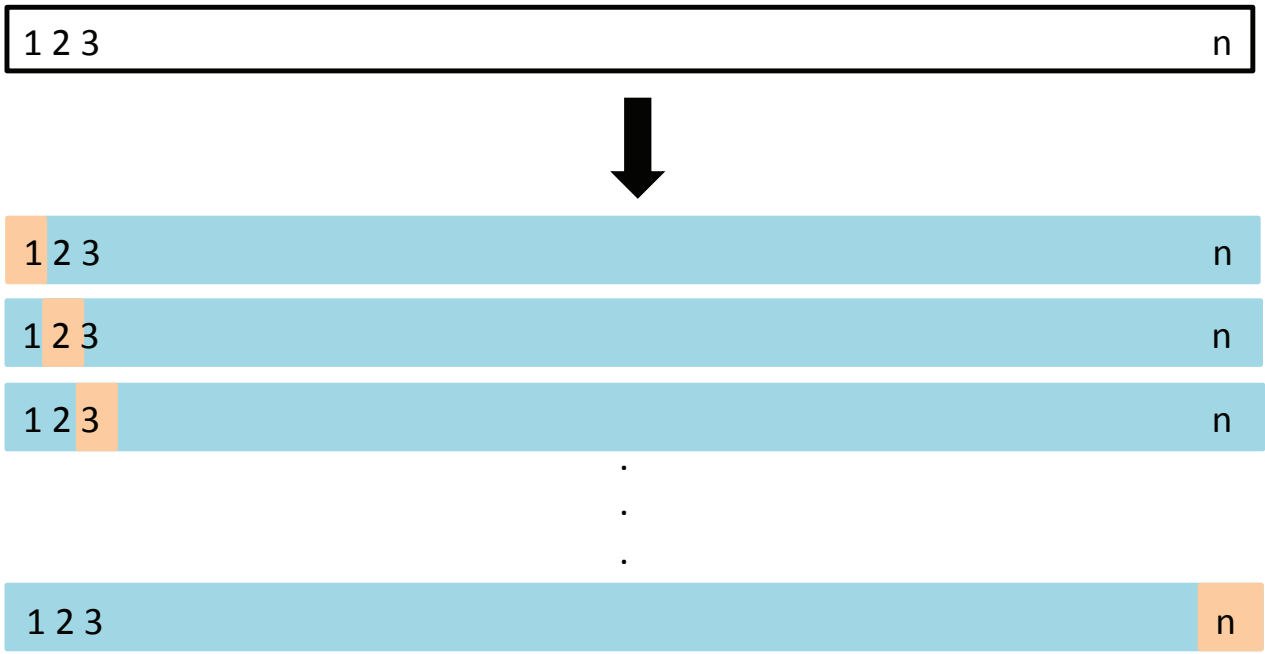




- MSE w zależności od stopnia wielomianu, przy pewnym losowym ale ustalonym podziale danych
- Kiedy zbiór trenujący jest mały (1/2 oryginalnego) - możliwe przeszacowanie błędu

- MSE przy różnych podziałach na zbiór treningowy/testowy
- Wyraźna tendencja – wielomian stopnia 2 daje zauważalną poprawę wielkości błędu.
- Brak konsensusu dla jakiego stopnia MSE jest najmniejszy

Walidacja krzyżowa *Leave-One-Out* (LOOCV)



- Mamy n obserwacji $(x_1, y_1), \ldots, (x_n, y_n)$
- Dla każdego $1 \leq i \leq n$ odkładamy i-tą obserwację (beżowy) i dopasowujemy model do pozostałych $n-1$ obserwacji (niebieski). Wyliczamy błąd dla i-tego modelu $MSE_i = (y_i - \hat{y}_i)^2$.

Walidacja krzyżowa *Leave-One-Out* (LOOCV)

- Estymacja błędu przy LOOCV to średnia

$$CV_{(n)} = (1/n) \sum_{i=1}^{n} MSE_i$$

- Dla regresji liniowej/wielomianowej estymowanej metodą najmniejszych kwadratów możemy też stosować wzór:

$$CV_{(n)} = (1/n) \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - h_i} \right)^2$$

gdzie

- $\hat{y}_i$ to dopasowana wartość zmiennej objaśnianej dla obserwacji i (z regresji ze wszystkimi predyktorami)
- h_i to wartość dźwigni dla obserwacji i (usunięcie obserwacji o wysokiej dźwigni powoduje duże zmiany w przebiegu prostej regresji)

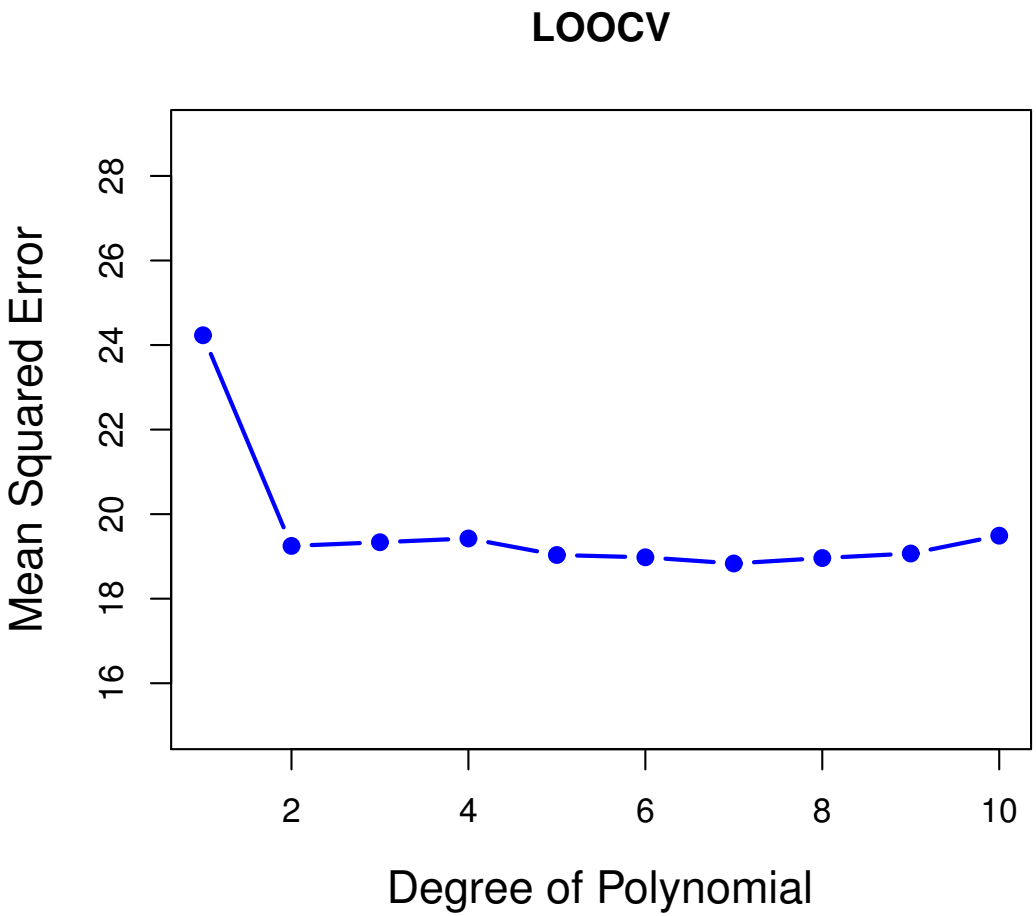
Zalety LOOCV w porównaniu do podejścia jednego zbioru walidacyjnego

- W podejściu zbioru walidacyjnego model trenowany jest na znacznie mniejszych danych i z tego powodu możliwe jest przeszacowanie błędu testowego
- LOOCV opiera się o wielokrotne trenowanie modelu na $n-1$ danych treningowych - prawie tylu co w pełnym zbiorze. Oczekujemy mniejszego obciążenienia oszacowania testowego MSE.
- Oszacowanie błędu w oparciu o podejście zbioru walidacyjnego da inny wynik w zależności od losowego podziału danych na zbiór treningowy i walidacyjny
- Jako średnia z n wartości, LOOCV zawsze da ten sam wynik dla tego samego zbioru danych (jeśli metoda uczenia jest deterministyczna)

Wady LOOCV w porównaniu do podejścia jednego zbioru walidacyjnego

- W ogólności, musimy powtarzać trenowanie modelu n razy (poza trickiem w modelu liniowym).
- Duży koszt obliczeniowy dla dużych n.
- W istocie analizujemy n **różnych** modeli, może być tak, że żaden z modeli nie ma dokładnie takiej jakości predykcji jak zmierzona średnia

Przykład LOOCV dla danych 'Auto' (Lewy panel)



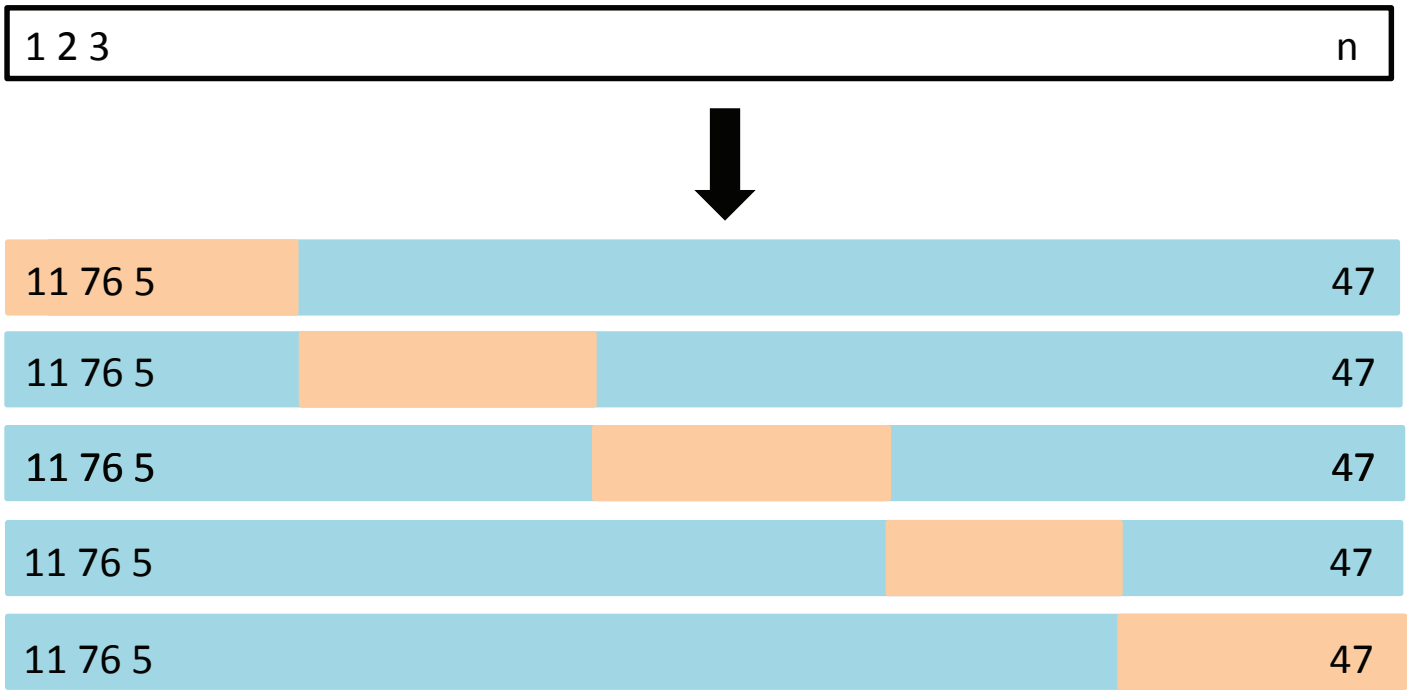
k-krotna walidacja krzyżowa

- Dane obserwowane są dzielone losowo na k części, mniej więcej podobnej wielkości.
- Dla $1 \leq i \leq k$, i-tą część traktujemy jako testową, a model trenujemy na pozostałych $k-1$ częściach. Obliczamy błąd MSE_i.
- Estymacja błędu dla k-krotnej walidacji krzyżowej to

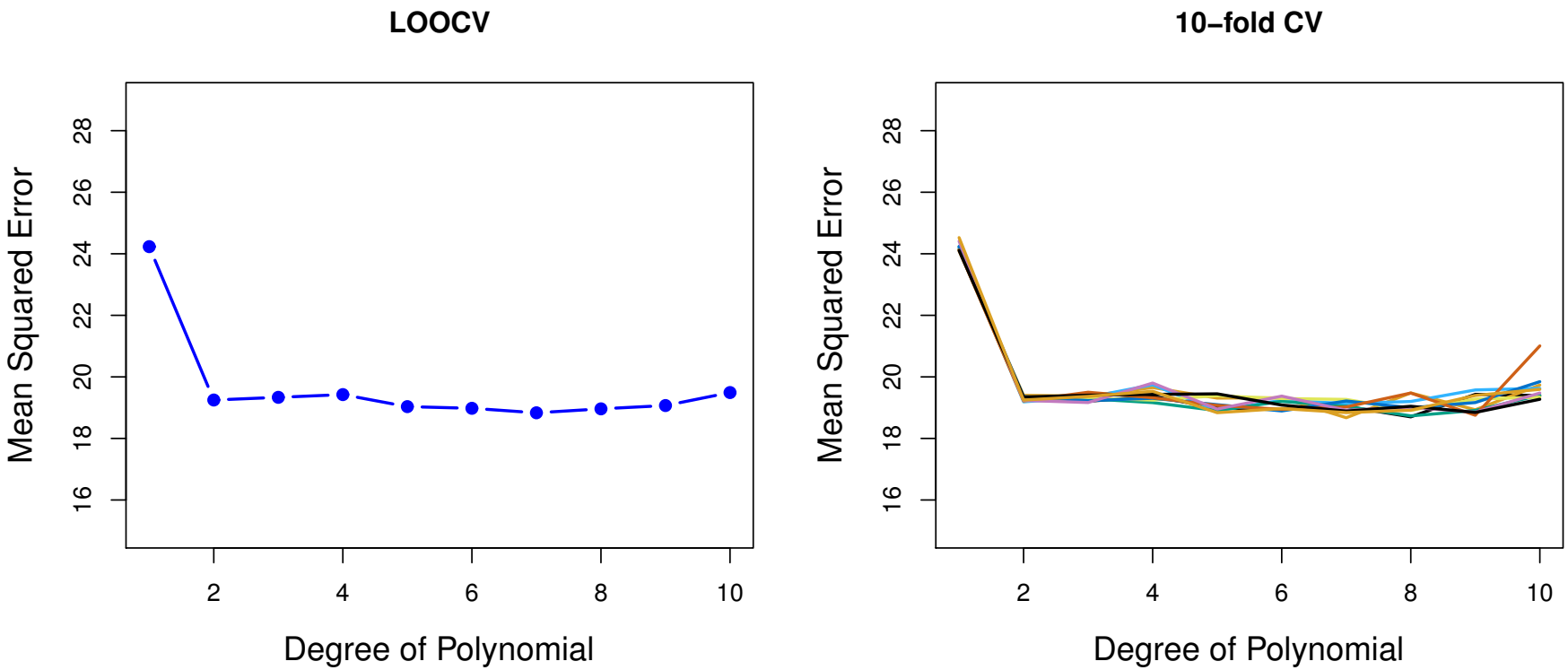
$$CV_{(k)} = (1/k) \sum_{i=1}^{k} MSE_i.$$

- LOOCV jest szczególnym przypadkiem tej metody ($k = n$).
- Typowo przyjmuje się $k = 5$ lub $k = 10$.
- Oszczędność, to n/k krotnie mniej estymacji modelu.
- Koszt, to – znów – mniejszy zbiór treningowy $\rightarrow$ gorsza estymacja jakości modelu

Ilustracja k-krotnej walidacji krzyżowej

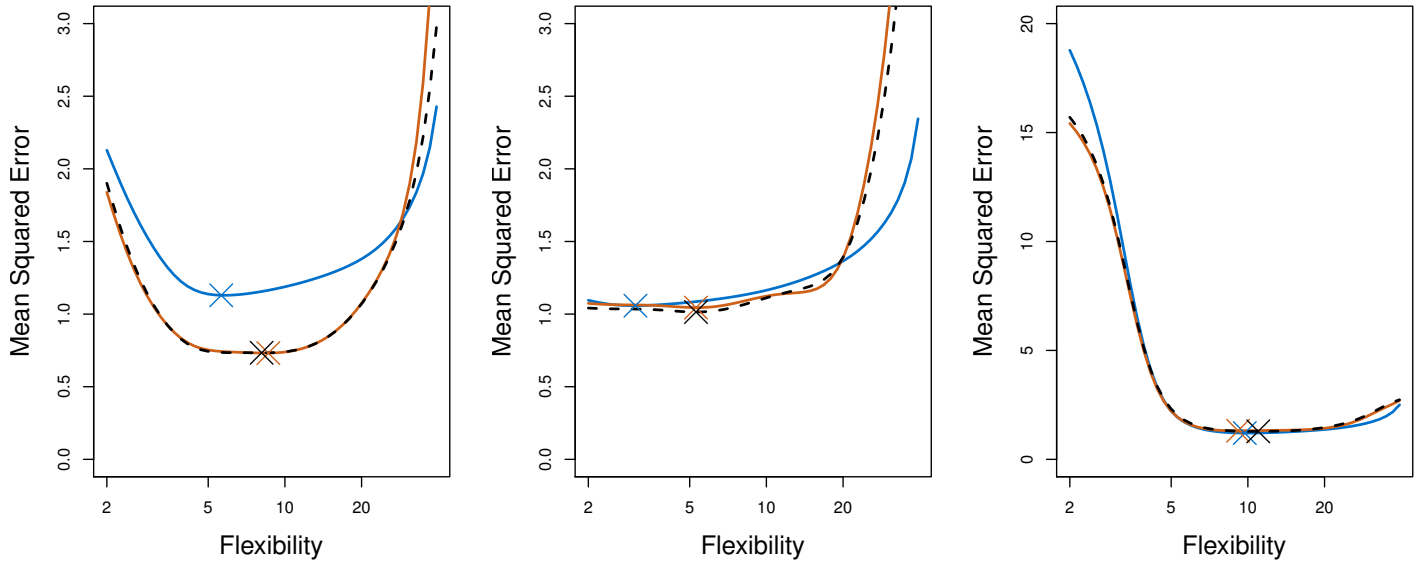


Przykład 10-cio krotnej walidacji krzyżowej wykonanej 9 razy dla danych 'Auto' (prawy panel)



Estymacja błędów w walidacji krzyżowej (dane symulowane)

Dla różnych stopni nieliniowości przy generowaniu danych:



- Niebieski: Prawdziwy błąd testowy
- Kreskowany czarny: Błąd testowy z metody LOOCV
- Pomarańczowy: Błąd testowy dla metody 10-cio krotnej walidacji krzyżowej
- Krzyżyki – punkty minimalne wykresów.
- (LOO)CV może niedo- albo prze-szacować prawdziwe MSE, ale dobrze znajduje punkty minimalne

Walidacja krzyżowa dla problemów klasyfikacji

Podejście podobne do problemów regresji. Dla LOOCV mamy:

$$CV_{(n)} = (1/n) \sum_{i=1}^{n} Err_i,$$

gdzie $Err_i = I(y_i \neq \hat{y}_i)$.

Dla k-krotnej walidacji krzyżowej i dla podejścia zbioru walidacyjnego definicja jest analogiczna.

Przykład: wielomianowa regresja logistyczna ($p = 2$)

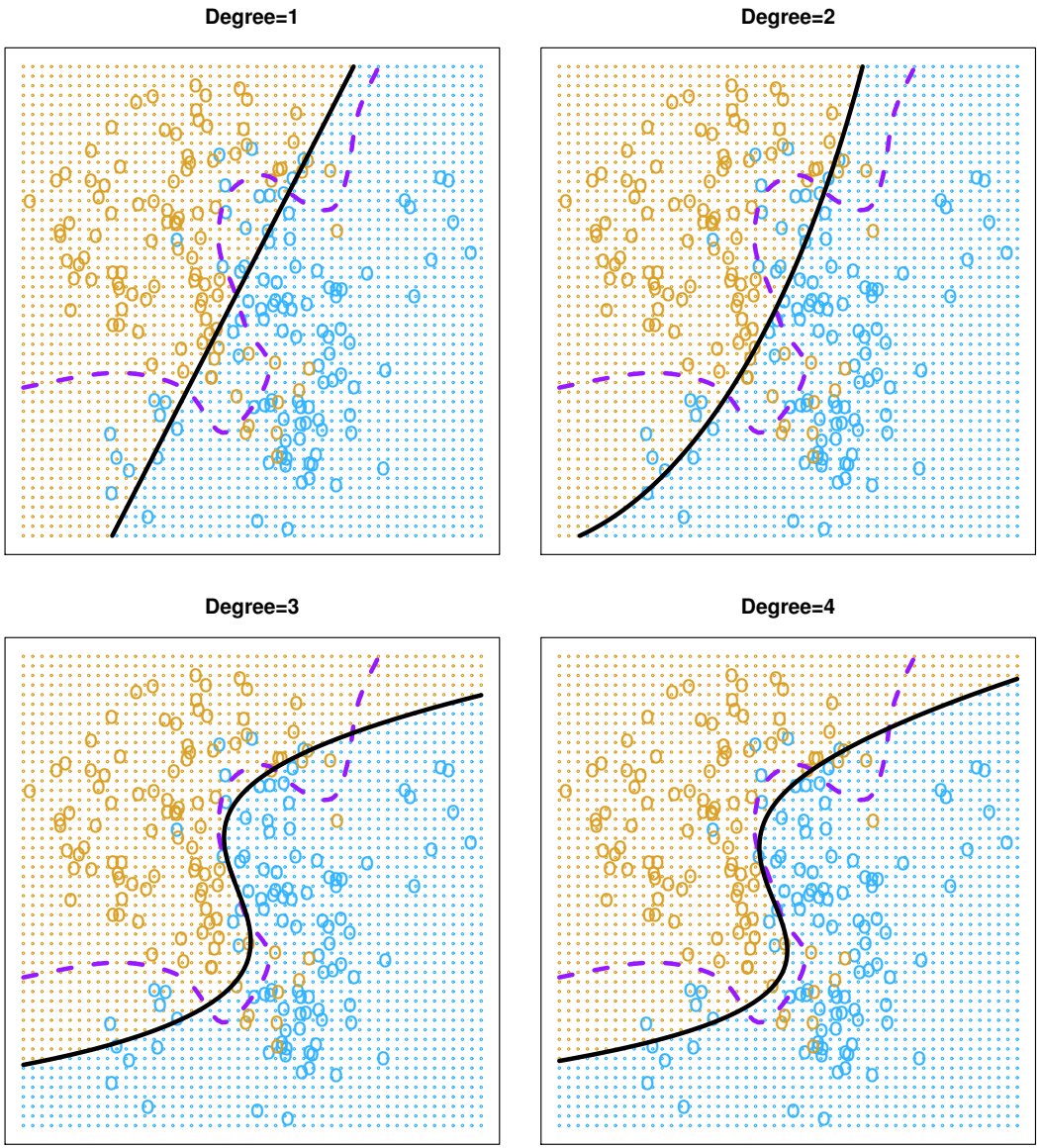
- Standardowa (liniowa) regresja logistyczna: estymowanie parametrów logarytmu ilorazu szans funkcją liniową:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

- Przy kwadratowej regresji logistycznej mamy:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2.$$

Przykład: wielomianowa regresja logistyczna ($p = 2$)

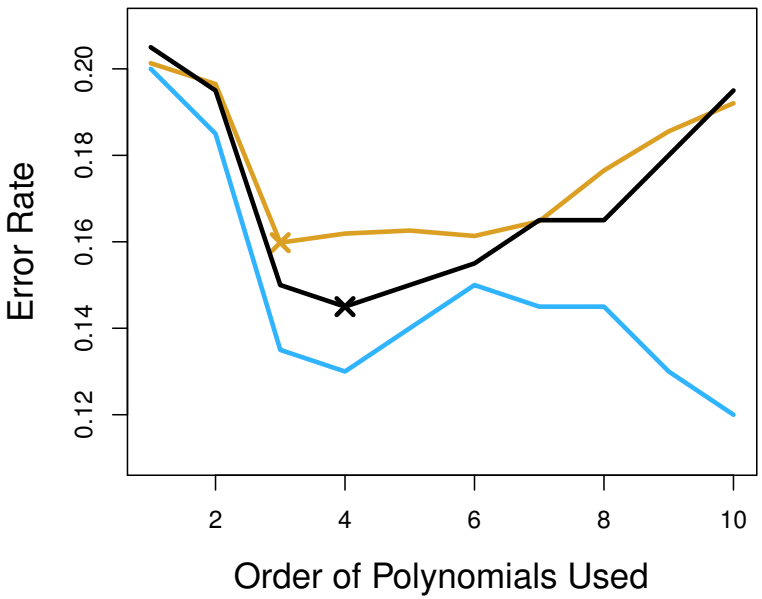


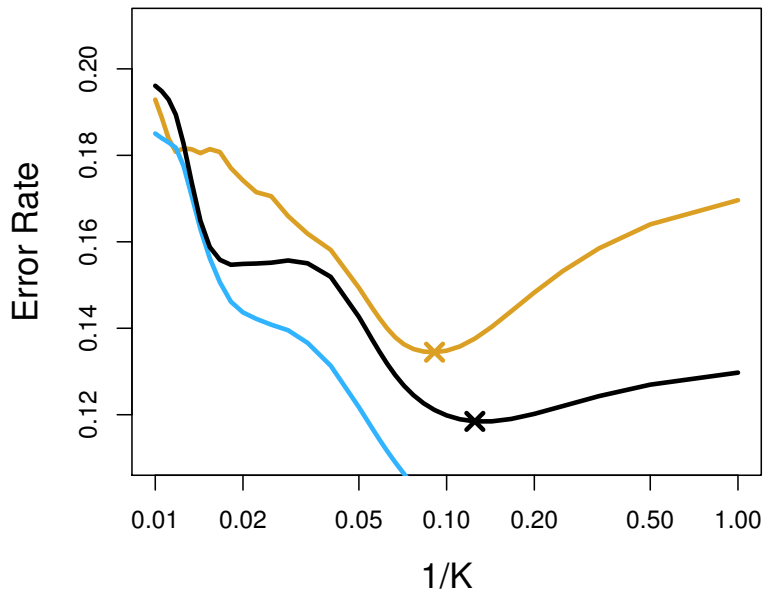
- Dane generowne syntetycznie
- Purpurowy przerywany: Bayesowska granica decyzyjna
- Ciągły czarny: Granica decyzyjna dla danej metody
- Błąd bayesowski: 0.133
- Błędy dla metod:
  - deg 1: 0.201
  - deg 2: 0.197
  - deg 3: 0.160
  - deg 4: 0.162

Walidacja krzyżowa pomaga znaleźć optymalny poziom elastyczności modelu

Logistyczna regresja (w zależności od stopnia wielomianu)

KNN (w zależności od elastyczności $(1/K)$).





Brązowy: błąd testowy; Niebieski: błąd treningowy; Czarny: błąd **10-cio krotnej CV**.

Obserwacje dotyczące poprzedniego wykresu

- Błąd 10-cio krotnej CV niedoszacowuje prawdziwy błąd testowy.
- Minima wykresów błędu testowego i 10-cio krotnej walidacji są położone blisko siebie.

Bootstrap

Do czego służy bootstrap?

Szacowanie wariancji estymatora parametru modelu.

Metoda bootstrap na prostym przykładzie (1)

- Mamy dwie zmienne losowe X i Y, które przedstawiają zwrot przy inwestycji w pierwszy lub drugi instument finansowy.
- Mamy pewną sumę pieniędzy, którą chcemy zainwestować: część α w pierwszy instrument i część $1-\alpha$ w drugi.
- Chcemy dobrać α tak, aby zminimalizować ryzyko inwestycji, co w tym przypadku jest wariancją $\mathrm{Var}(\alpha X + (1-\alpha)Y)$.
- Można pokazać (przyrównanie pochodnej po α do 0), że optymalne α spełnia

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}},$$

gdzie $\sigma_X^2 = \mathrm{Var}(X)$, $\sigma_Y^2 = \mathrm{Var}(Y)$ oraz $\sigma_{XY} = Cov(X, Y)$.

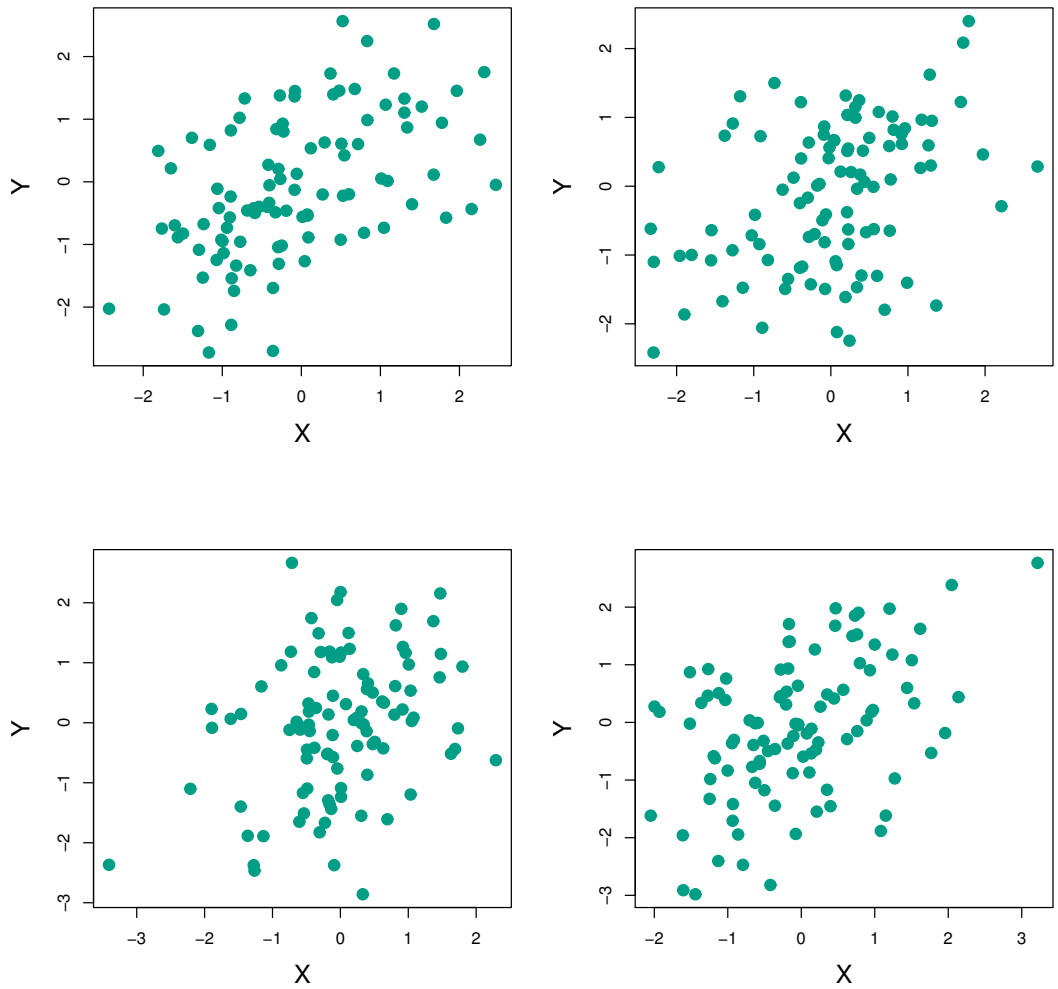
Metoda bootstrap na prostym przykładzie (2)

- W rzeczywistości nie znamy prametrów σ_X^2, σ_Y^2 oraz σ_{XY}. Estymujemy je z obserwowanych danych, co prowadzi do estymacji α

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}.$$

- Dla potrzeb przykładu budujemy syntetyczny model generowania danych: $\sigma_X^2 = 1$, $\sigma_Y^2 = 1.25$ oraz $\sigma_{XY} = 0.5$. Więc $\alpha = 0.6$.

Cztery zestawy symulowanych danych (po 100 par)



Wyestymowane $\hat{\alpha}$ jest równe odpowiednio: 0.576, 0.532, 0.657, 0.651.

Estymacja α

Powtarzając tę procedurę 1000 razy otrzymujemy wartości $\hat{\alpha}_1, \ldots, \hat{\alpha}_{1000}$, dla których wyliczamy średnią

$$\bar{\alpha} = (1/1000) \sum_{r=1}^{1000} \hat{\alpha}_r = 0.5996.$$

oraz odchylenie standardowe

$$\sqrt{\frac{1}{1000-1} \sum_{r=1}^{1000} (\hat{\alpha}_r - \bar{\alpha})^2} = 0.083$$

Czyli błąd standardowy wynosi $SE(\hat{\alpha}) \approx 0.083$.

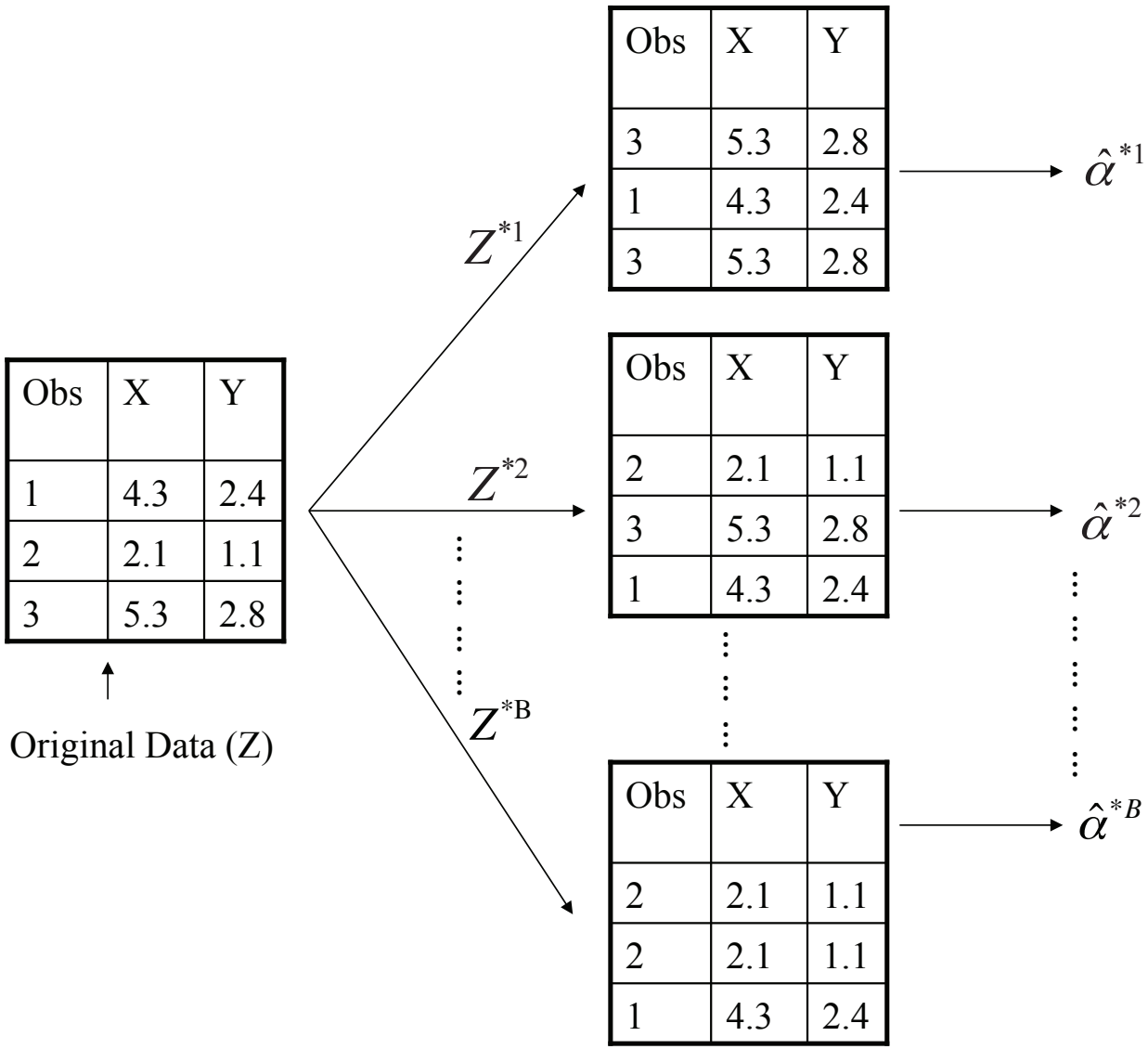
Wniosek: możliwość 1000-krotnego symulowania próbki rozmiaru 100 z modelu pozwala dość dobrze wybrać model 1-parametrowy

Bootstrap jako metoda generowania nowych próbek

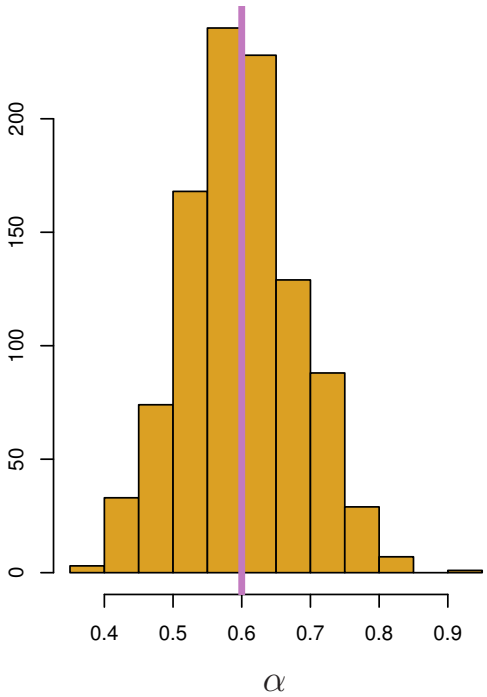
- Przypuśćmy, że mamy N obserwacji stanowiących nasz zbiór danych Z.
- **Próba bootstrap'owa:** Losujemy N obserwacji ze zbioru Z ze zwracaniem, otrzymując w ten sposób nowy zbiór Z^* oraz estymację $\hat{\alpha}^*$.
- Powtarzamy to wiele razy (np. B razy), otrzymując zbiory N-elementowe $Z^{*1}, \ldots, Z^{*B}$ oraz estymacje $\hat{\alpha}^{*1}, \ldots, \hat{\alpha}^{*B}$.
- Obliczamy błąd standardowy

$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - (1/B)\sum_{s=1}^{B}\hat{\alpha}^{*s}\right)^2}$$

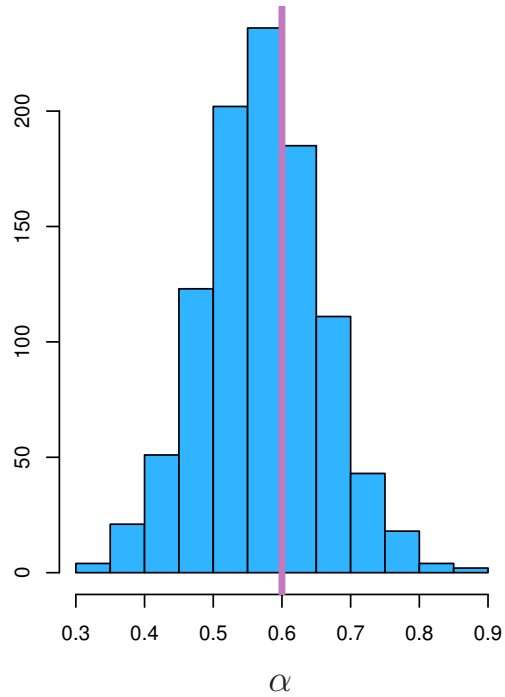
Bootstrap – ilustracja (N=3)



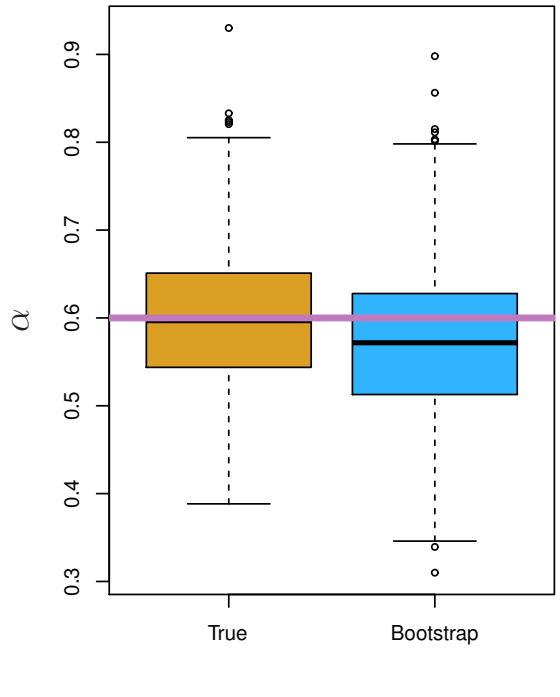
Histogramy losowań estymacji $\hat{\alpha}$



Histogram otrzymanych $\hat{\alpha}$ z losowania 1000 razy X i Y.



Histogram otrzymanych $\hat{\alpha}^*$ z bootstrapowych losowań 1000 razy.



Wykresy pudełkowe (*boxplot*) dla obu populacji.

Wybór najlepszego podzbioru predyktorów

Które predyktory są istotne?

Przypomnienie: statystyki RSS i R^2 jako miary dopasowania

Mamy n odpowiedzi $y_1, \ldots, y_n$ i n predykcji $\hat{y}_1, \ldots, \hat{y}_n$.

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Statystyka R^2:

$$R^2 = 1 - \frac{RSS}{TSS},$$

dla $TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$ mierzącego **całkowitą wariancję** wartości zmiennej objaśnianej y.

Ale: te miary dopasowania **nie** są dobrym kryterium wyboru modelu (model selection) gdy modele mogą różnić się złożonością.

C_p Mallowa (dla regresji liniowej)

- Dla n obserwacji, d predyktorów oraz estymowanej $\hat{\sigma}^2$ wariancji błędu ε związanego z odpowiedzią w modelu liniowym

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2).$$

- Im mniejsze C_p, tym lepiej.
- Polega na dopisaniu **kary** do RSS za liczbę wybranych predyktorów
- Gdy rośnie liczba predyktorów w modelu, to maleje RSS, ale wzrasta kara.
- Metodę C_p stosujemy tylko w przypadku regresji liniowej
  - gdy $\hat{\sigma}^2$ jest nieobciążonym estymatorem σ^2 to C_p jest nieobciążonym estymatorem testowego MSE
  - C_p jest ściśle związane z AIC (Akaike Information Criterion) i zawsze prowadzi do wyboru tego samego modelu.

Akaike information criterion (AIC)

- AIC jest oryginalnie zdefiniowane dla metod estymowanych metodą największej wiarogodności

$$AIC = 2d - 2\log(\hat{L})$$

gdzie

  - d to liczba estymowanych parametrów modelu (stopni swobody)
  - $\hat{L}$ to maksymalna wartość funkcji wiarogodności.

- Im mniejsze AIC, tym lepiej.
- Większy log likelihood oznacza lepsze dopasowanie
- Dodana kara za złożoność modelu
- Ten wzór funkcjonuje w literaturze z dokładnością do stałych, np. też

$$\frac{2}{n}d - \frac{2}{n}\log(\hat{L}).$$

AIC dla modelu regresji liniowej

- Dla regresji liniowej przy założeniu normalnego rozkładu ε, AIC odpowiada C_p.
- Ogólnie, AIC opiera się na dopisaniu **kary** do $\hat{L}$ lub RSS za liczbę wybranych predyktorów.

BIC (Byesian information criterion, lub kryterium informacyjne Schwarza)

- Podobnie jak AIC, wprowadzone dla metod estymowanych metodą największej wiarogodności

$$BIC = \log(n)d - 2\log(\hat{L})$$

- Im mniejsze BIC, tym lepiej.
- Silniejsza kara za liczbę wybranych predyktorów niż w przypadku AIC.

Poprawione R^2

- Poprawiony współczynnik determinacji R^2

$$\textit{Poprawiony } R^2 = 1 - \frac{RSS(n-1)}{TSS(n-d-1)}.$$

- Im większy poprawiony R^2, tym lepiej.
- Poprawka w R^2 działa tak, że ze wzrostem liczby predyktorów ułamek RSS/TSS maleje, ale rośnie ułamek $(n-1)/(n-d-1)$.
- W przeciwieństwie do C_p, AIC i BIC, które mają one solidne podstawy statystyczne, jest to heurystyczna poprawka.
- Działa podobnie, jak $RSE = \hat{\sigma} = \sqrt{\frac{RSS}{n-p-1}}$.

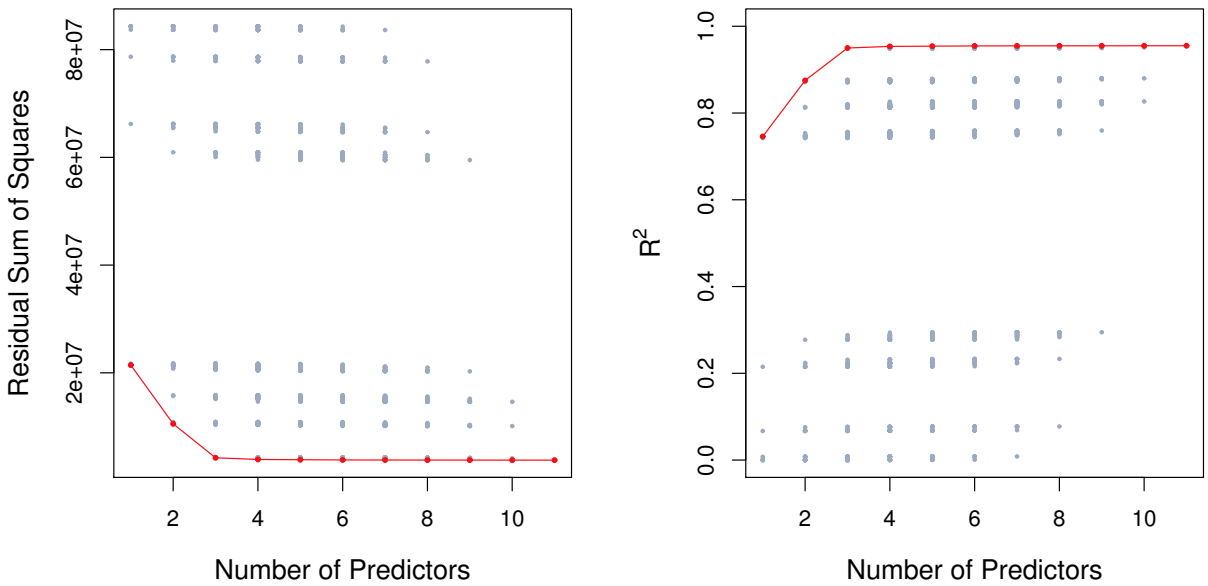
Metoda siłowa: wyczerpujące przeszukiwanie

- Zaczynamy od modelu $\mathcal{M}_0$ z pustym zbiorem predyktorów.
- Dla $k = 1, 2, \ldots, p$:
  - Dopasuj parametry dla każdego z $\binom{p}{k}$ k elementowych podzbiorów predyktorów.
  - Wybierz najlepszy (najmniejsze RSS lub największe R^2) model $\mathcal{M}_k$ z tych $\binom{p}{k}$ modeli.
- Wybierz najlepszy model spośród $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ przy pomocy jednej z metod oceny modeli (walidacja krzyżowa, C_p (AIC), BIC, lub poprawiony R^2).

Uwaga: złożoność wykładnicza ($O(2^p)$).

Ilustracja wyczerpującego przeszukiwania (Dane 'Credit').

Każdy szary punkt - określony model z podzbiorem predyktorów.
Czerwona linia łączy najlepsze modele.



- Jeśli zwiększamy liczbę predyktorów w modelu, to statystyka RSS maleje, a R^2 rośnie.
- Można ich użyć do porównania modeli tego samego rozmiaru.
- **Nie nadają się do wyboru najlepszego podzbioru predyktorów.**

Algorytm zachłanny – wyszukiwanie w przód

- Zaczynamy od modelu $\mathcal{M}_0$ z pustym zbiorem predyktorów.
- Dla $k = 0, 2, \ldots, p-1$:
  - Rozszerzamy model $\mathcal{M}_k$ przez dodanie jednego predyktora (z tych które nie są obecne w $\mathcal{M}_k$)
  - Wybierz najlepszy (najmniejsze RSS lub największe R^2) model spośród tych $p-k$ rozważanych powyżej rozszerzeń. Oznaczmy go $\mathcal{M}_{k+1}$.
- Wybierz najlepszy model spośród $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ przy pomocy jednej z metod oceny modeli (walidacja krzyżowa, C_p (AIC), BIC, lub poprawiony R^2).

Złożoność $1 + \sum_{k=0}^{p-1} p - k = 1 + p(p+1)/2$
Działa nawet dla danych wysokowymiarowych ($n < p$).
Dla regresji liniowej rozważane są tylko modele $\mathcal{M}_0, \ldots, \mathcal{M}_{n-1}$ (MNK nie daje jednoznacznego rozwiązania przy $p > n$).

Algorytm zachłanny może nie znaleźć optymalnego podzbioru – dane 'Credit'

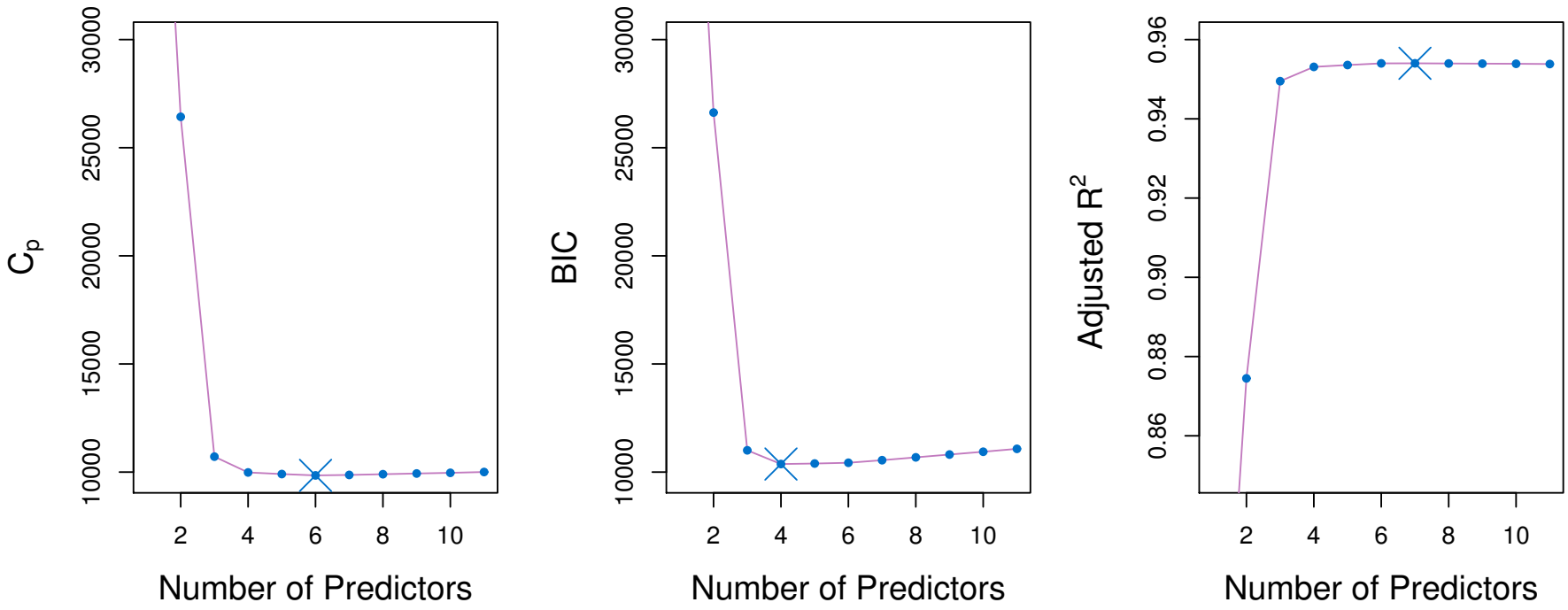
# Variables	Best subset	Forward stepwise
One	rating	rating
Two	rating, income	rating, income
Three	rating, income, student	rating, income, student
Four	cards, income, student, limit	rating, income, student, limit

Algorytm zachłanny – wyszukiwanie wstecz

- Zaczynamy od modelu $\mathcal{M}_p$ z wszystkimi predyktorami.
- Dla $k = p, p-1, \ldots, 1$:
  - Rozważamy każdy z k modeli otrzymanych z $\mathcal{M}_k$ przez usunięcie jednego predyktora.
  - Wybierz najlepszy (najmniejsze RSS lub największe R^2) model spośród tych k rozważanych powyżej rozszerzeń. Oznaczmy go $\mathcal{M}_{k-1}$.
- Wybierz najlepszy model spośród $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ przy pomocy jednej z metod oceny modeli (walidacja krzyżowa, C_p (AIC), BIC, lub poprawiony R^2).

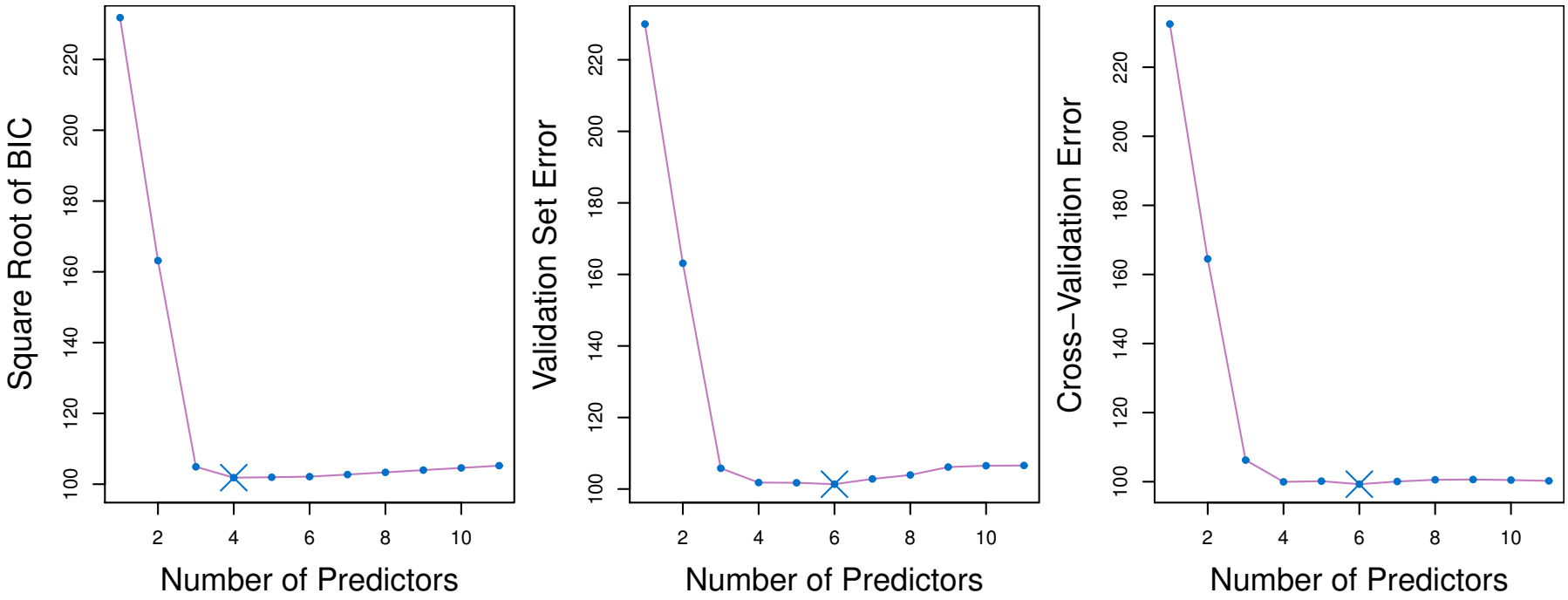
Uwaga: metoda wyszukiwania wstecz nie działa w przypadku $p > n$.

Metody C_p, BIC oraz poprawione R^2 w zastosowaniu do danych 'Credit'



Krzyżykiem zaznaczono minimum dla C_p (6 predyktorów: 'income', 'limit', 'rating', 'cards', 'age' oraz 'student'), dla BIC (4 predyktory: 'income', 'limit', 'cards' oraz 'student') oraz maksimum dla poprawionego R^2 (7 predyktorów).

Metoda BIC a walidacja i walidacja krzyżowa



Dla liczby predyktorów 4, 5, 6 te trzy metody oceny jakości są podobne. Warto wybrać model najprostszy (4 predyktory).

Metoda walidacji krzyżowej jest najbardziej uniwersalną metodą oceny jakości modelu, ale jest też najbardziej kosztowna obliczeniowo.

Podsumowanie

Szacowanie błędu testowego

- Podejście zbioru walidacyjnego
- Walidacja leave one out
- Walidacja krzyżowa

Szacowanie wariancji estymatora parametru

- Bootstrap

Algorytmy selekcji modelu

- Kryteria porównywania modeli o różnej liczbie cech
- Wyczerpujące przeszukiwanie
- Algorytmy zachłanne: przeszukiwanie w przód, wstecz, mieszane