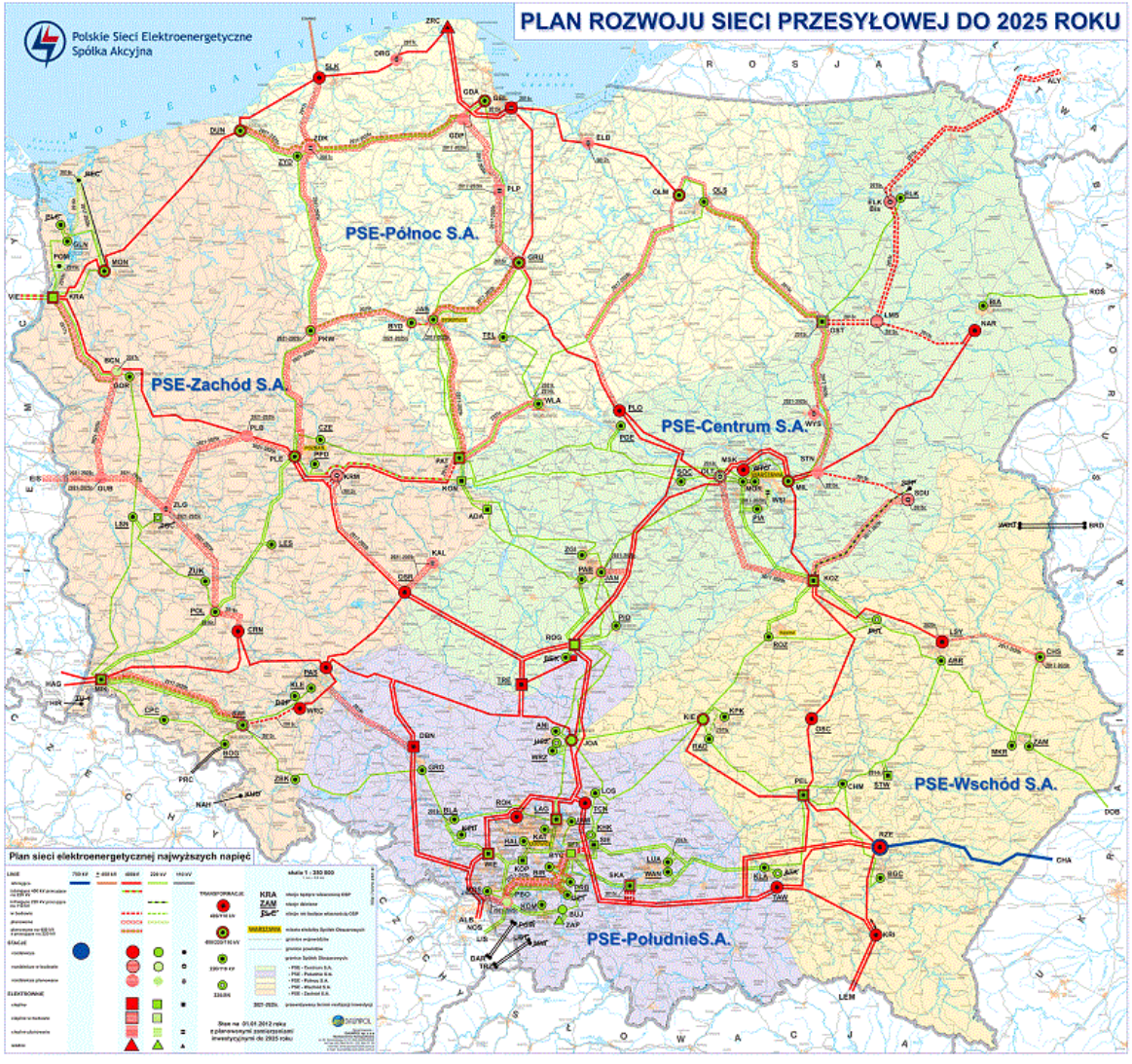
Teoria uczenia statystycznego z perspektywy ekonometryka

Bogumił Kamiński

Zakład Wspomagania i Analizy Decyzji
Instytut Ekonometrii
Kolegium Analiz Ekonomicznych
Szkoła Główna Handlowa

7 marca 2017

Rynek energii elektrycznej



Źródło: Polskie Sieci Elektroenergetyczne, 2013

Przejście na metodę cen węzłowych:

- do 5400 modeli predykcyjnych
- 5-minutowe dane podażowe, popytowe i systemowe, dane pogodowe
- prognozowanie w tej samej granulacji
- automatyczna kontrola jakości modeli

System przesyłu paliwa gazowego

Ograniczenia fizyczne w przesyle (ciśnienie gazu)

Dzienne nominacje dla 63 punktów wejścia i 966 punktów wyjścia

Źródło: GAZ-SYSTEM, 2013

Wspólna charakterystyka problemów

Typowy zbiór danych poddawany analizie:

- liczba obserwacji rzędu kilku do kilkudziesięciu tysięcy
- tysiące zmiennych objaśniających

Oczekiwania od modeli:

- maksymalizacja jakości prognoz
- krótki czas na przygotowanie modelu (automatyzacja procesu)

Uczenie statystyczne

Pierwotna definicja uczenia statystycznego (Vapnik, 1999)

Dla zadanej klasy funkcji $\mathcal{F} = \{\alpha \in \Lambda : f(x, \alpha)\}$, procesu generującego dane (X, Y) oraz funkcji straty $L(y, \hat{y})$ rozwiązać problem:

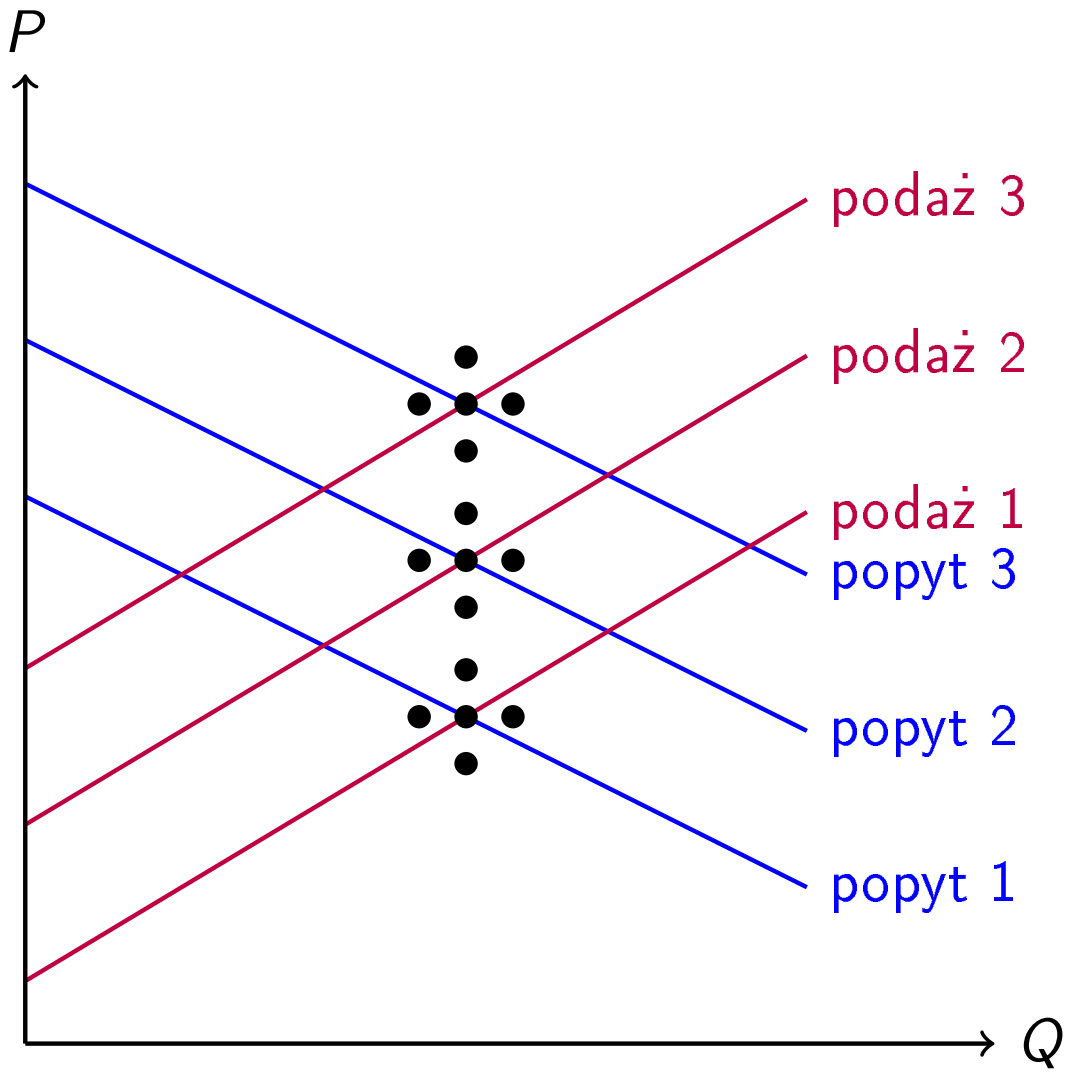
$$\hat{\alpha} = \arg\min_{\alpha} E(L(Y, f(X, \alpha))$$

na podstawie próby $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Aktualna definicja „operacyjna" (James et. al, 2013)

Zestaw narzędzi pozwalających na modelowanie i rozumienie złożonych zbiorów danych.

Kiedy zawodzi założenie, że proces (X, Y) jest stały?



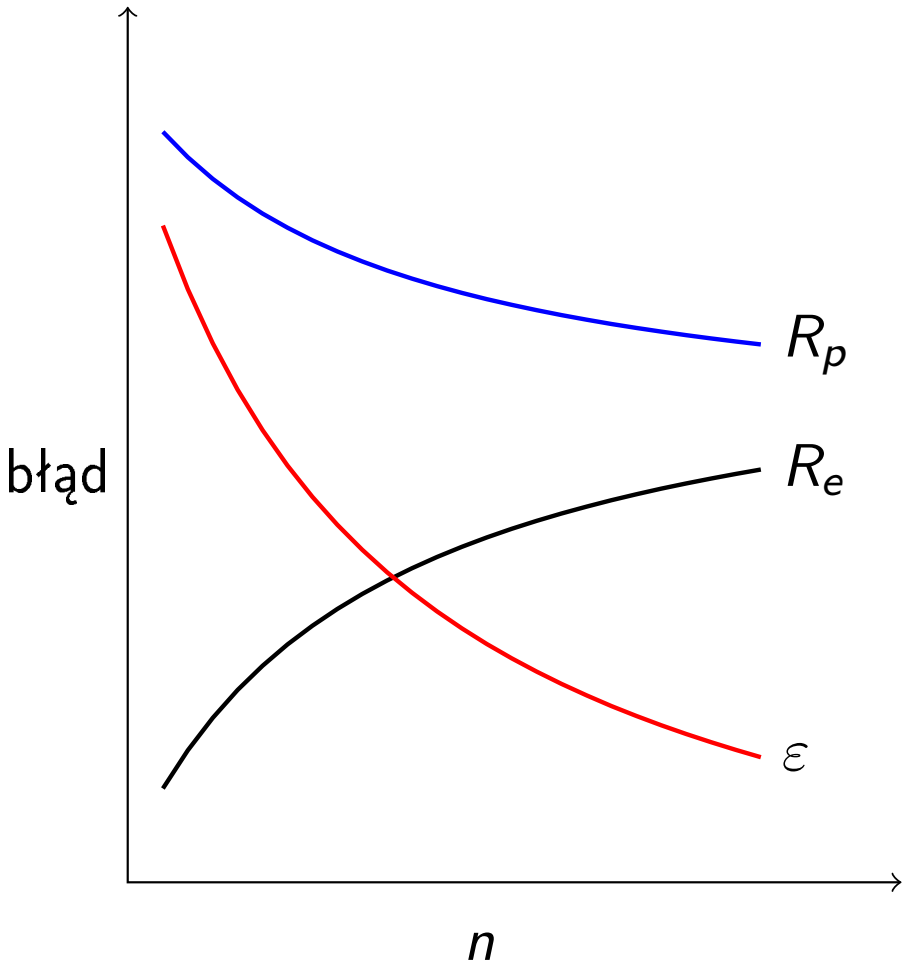
Twierdzenie Vapnika (dla problemu klasyfikacji)

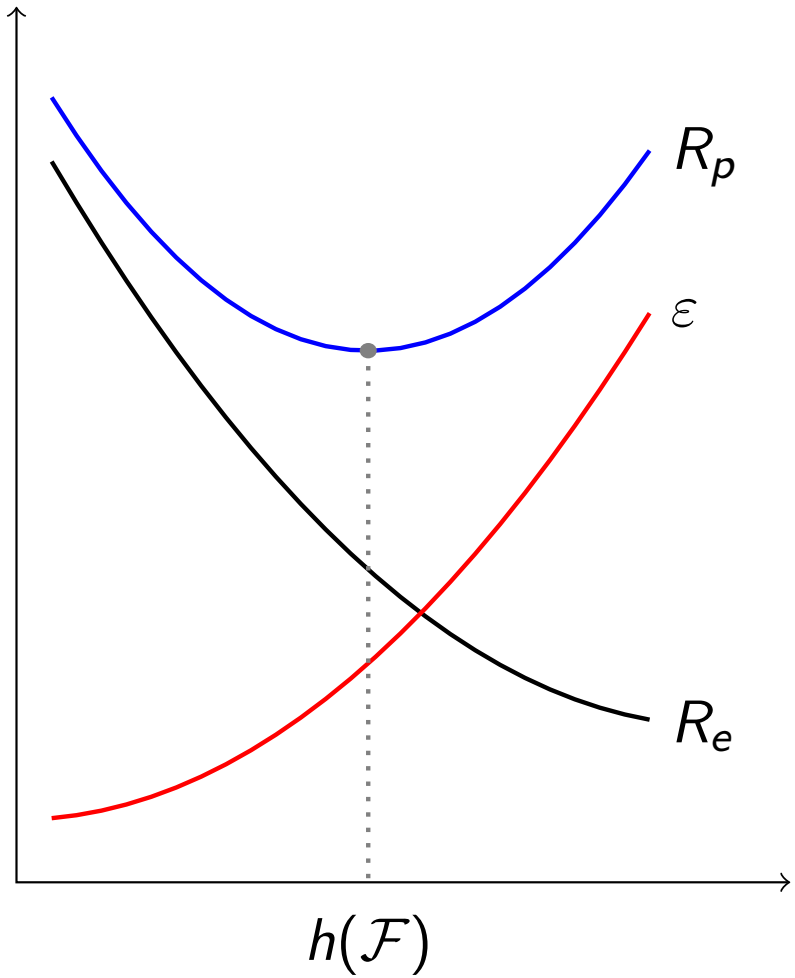
- zadana klasa funkcji dopuszczalnych $\mathcal{F}$
- dla $\mathcal{F}$ można wyznaczyć
  tzw. wymiar Vapnika-Chervonenkisa $h(\mathcal{F})$
  mierzący jej zdolność do dopasowywania się do danych
- dysponujemy n-elementową próbą estymacyjną
- wybieramy funkcję $f \in \mathcal{F}$ minimalizującą błąd na danych estymacyjnych R_e
- chcemy oszacować oczekiwany błąd prognozy R_p

Twierdzenie (Vapnik, 1995)
Dla dowolnego łącznego rozkładu (X, Y) *z prawdopodobieństwem* $1 - q$ *zachodzi zależność:*

$$R_p \leq R_e + \underbrace{\sqrt{\frac{h(\mathcal{F})\left(1 + \ln(2n/h(\mathcal{F}))\right) - \ln(q/4)}{n}}}_{\varepsilon}$$

Twierdzenie Vapnika: ilustracja





Twierdzenie Vapnika: procedura

- wybieramy rodzinę zagnieżdżonych klas funkcji

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \ldots$$

$$\Downarrow$$

$$h(\mathcal{F}_1) \leq h(\mathcal{F}_2) \leq h(\mathcal{F}_3) \leq \ldots$$

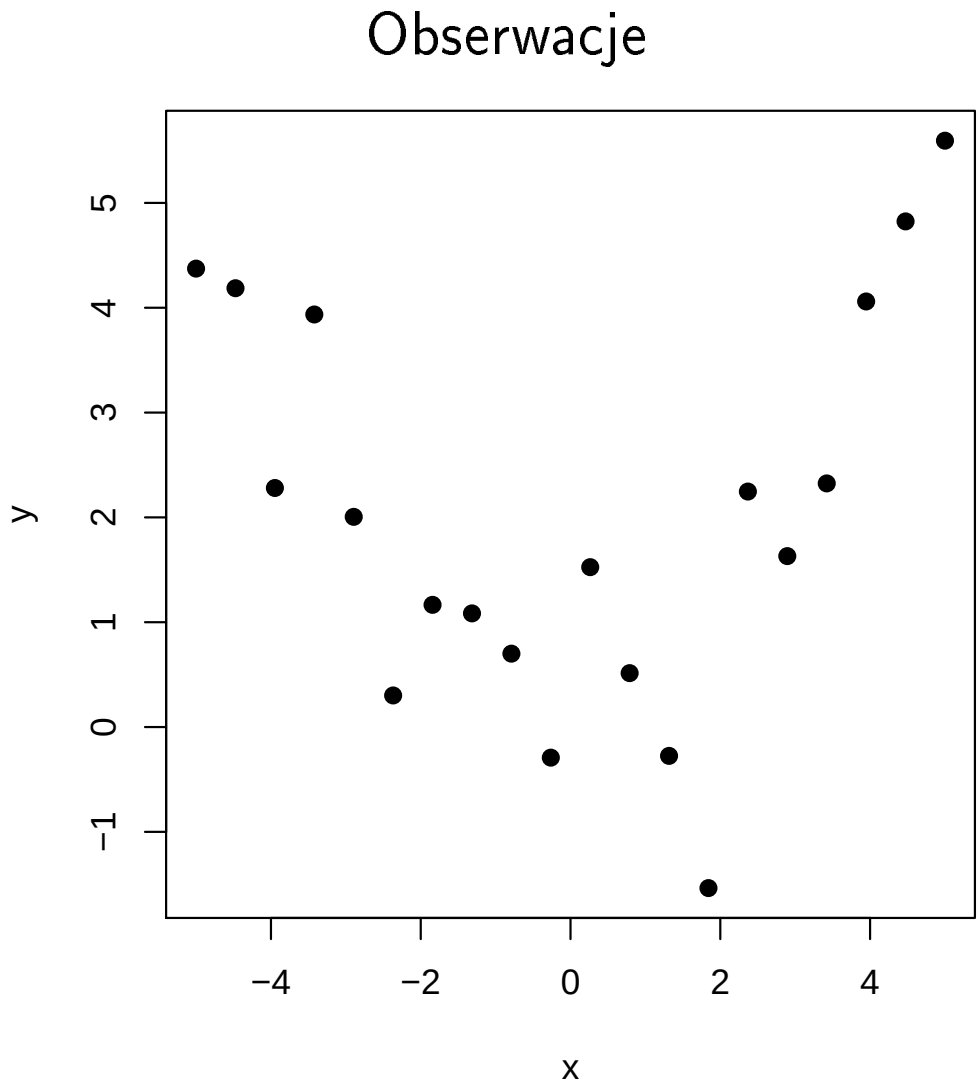
- wyznaczamy

$$R_e(\mathcal{F}_1) \geq R_e(\mathcal{F}_2) \geq R_e(\mathcal{F}_3) \geq \ldots$$

$$\varepsilon(\mathcal{F}_1) \leq \varepsilon(\mathcal{F}_2) \leq \varepsilon(\mathcal{F}_3) \leq \ldots$$

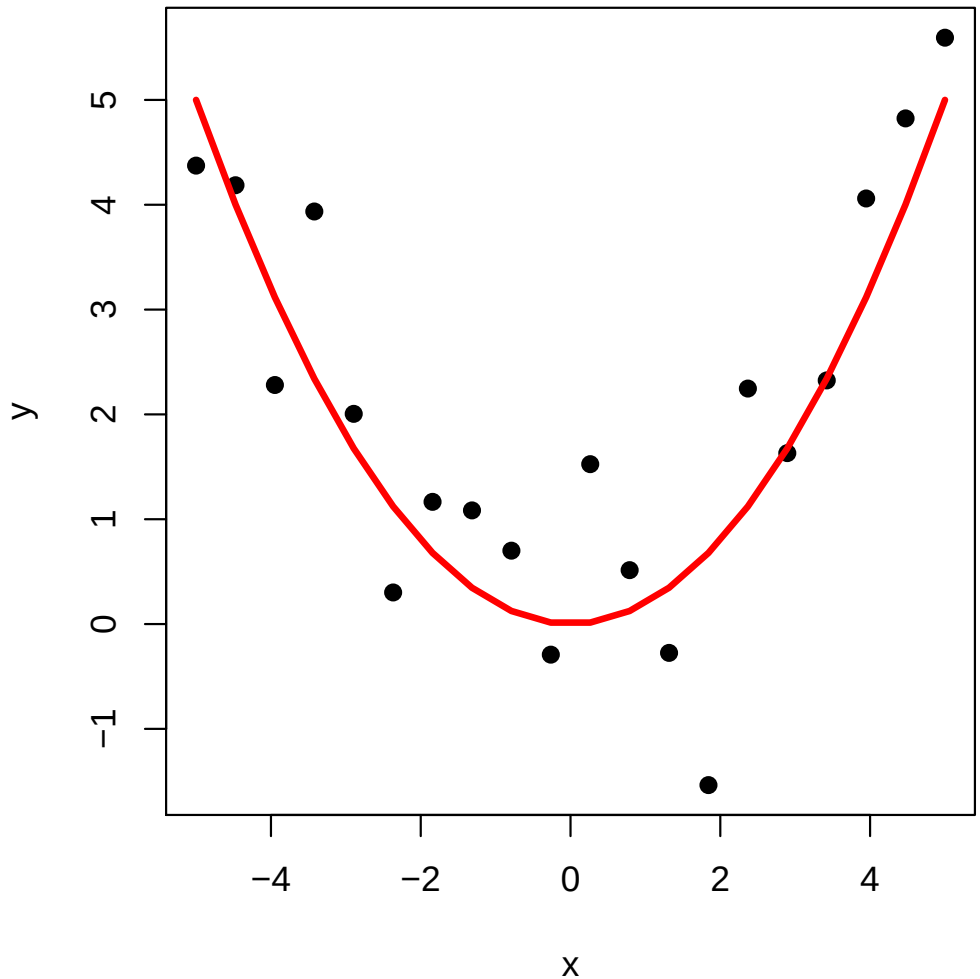
- wybieramy model oszacowany na podstawie klasy funkcji $\mathcal{F}_i$ minimalizującej oszacowanie R_p

Przykład regularyzacji (1)



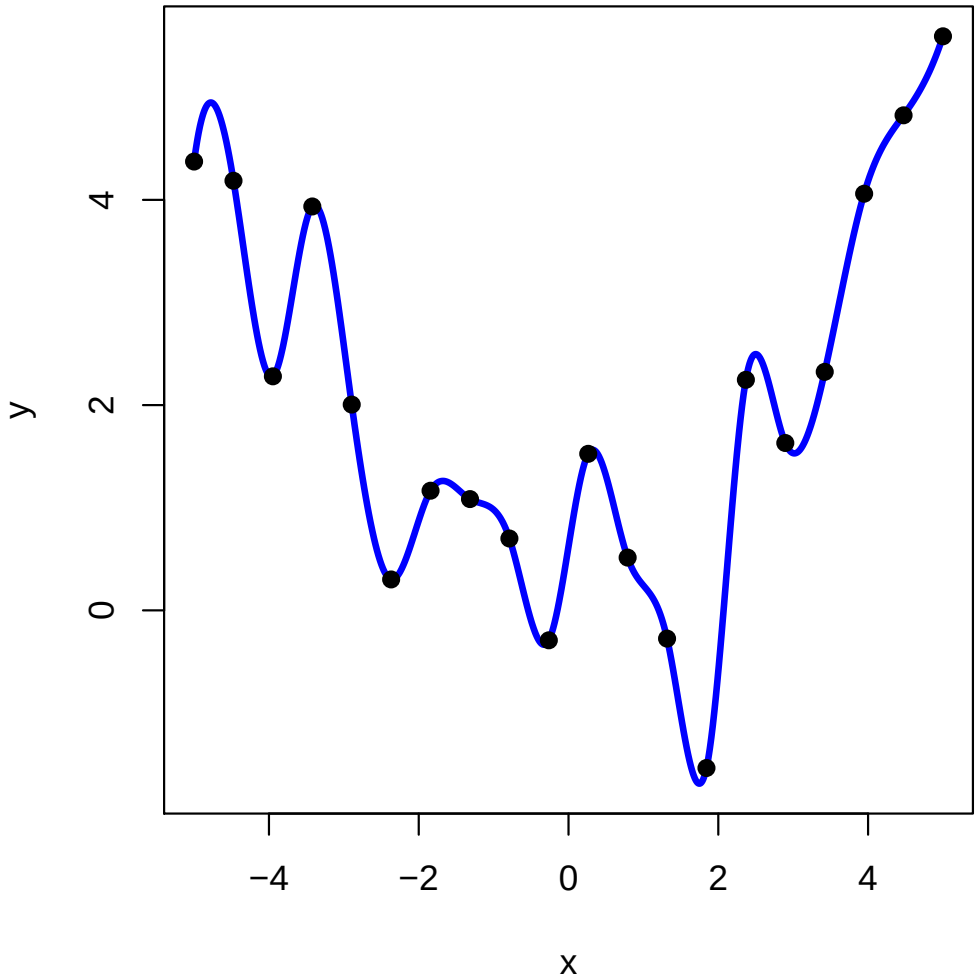
Przykład regularyzacji (2)

proces generujący dane: $y = x^2/5 + \varepsilon$, gdzie $\varepsilon \sim N(0, 1)$



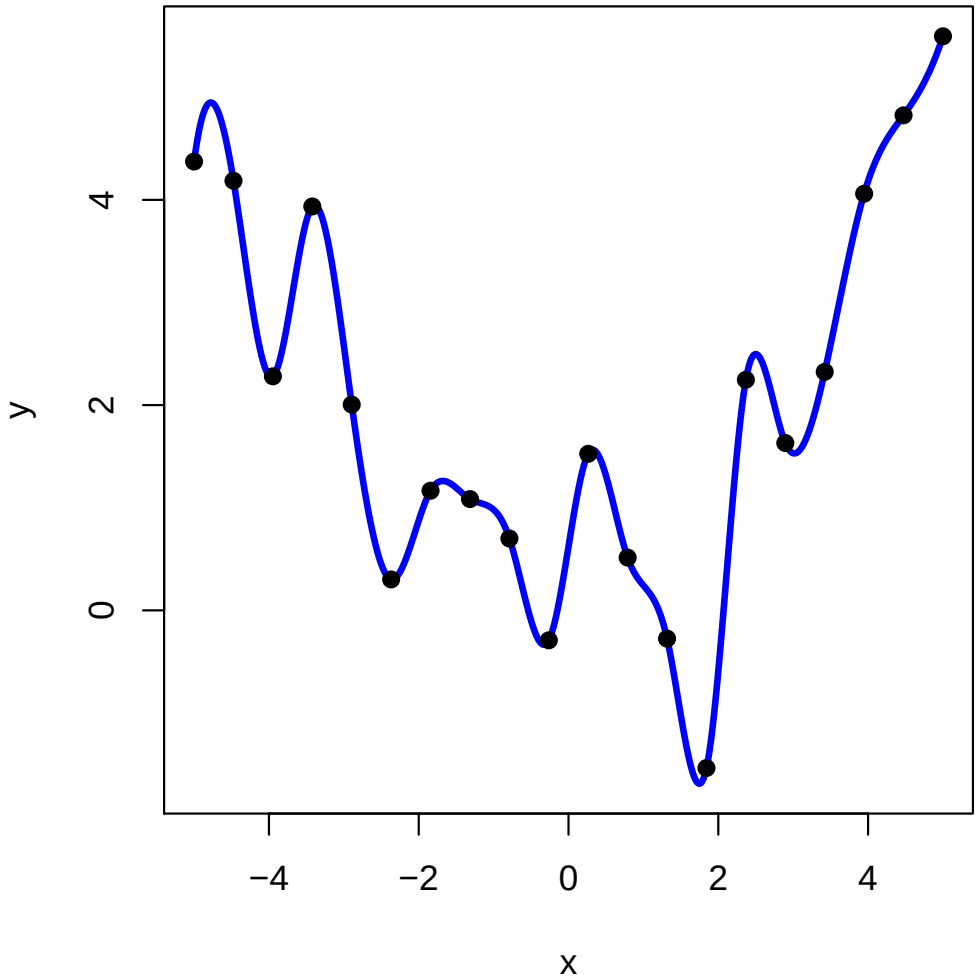
Przykład regularyzacji (3)

Dwukrotnie różniczkowalna funkcja f: $\sum_{i=1}^{n}(f(x_i) - y_i)^2 \to \min$



Przykład regularyzacji (3)

Dwukrotnie różniczkowalna funkcja f: $\sum_{i=1}^{n}(f(x_i) - y_i)^2 \rightarrow \min$

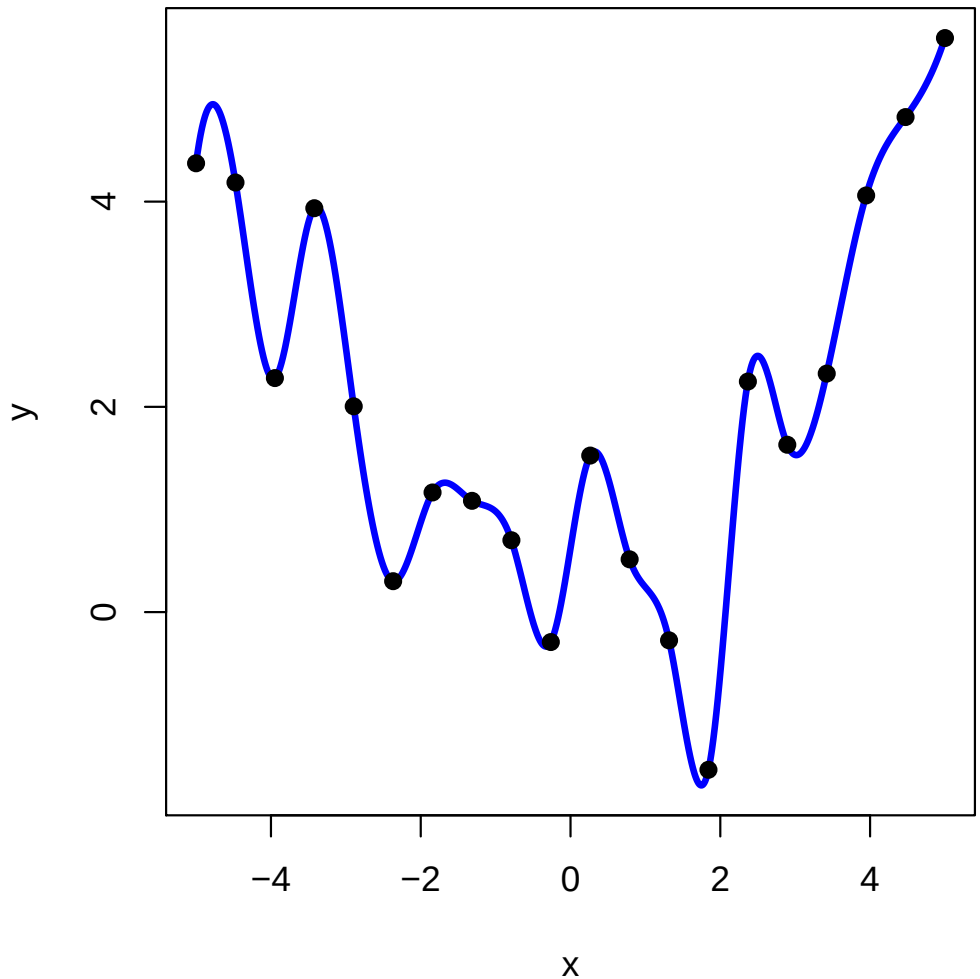


zagnieżdżona klasa funkcji:
wygładzane funkcje sklejane (Hastie et al., 2001)

Przykład regularyzacji (4)

Dwukrotnie różniczkowalna funkcja f:

$\sum_{i=1}^{n}(f(x_i) - y_i)^2 \to \min$, p.w. $\int_D [f''(x)]^2 dx \leq \delta$

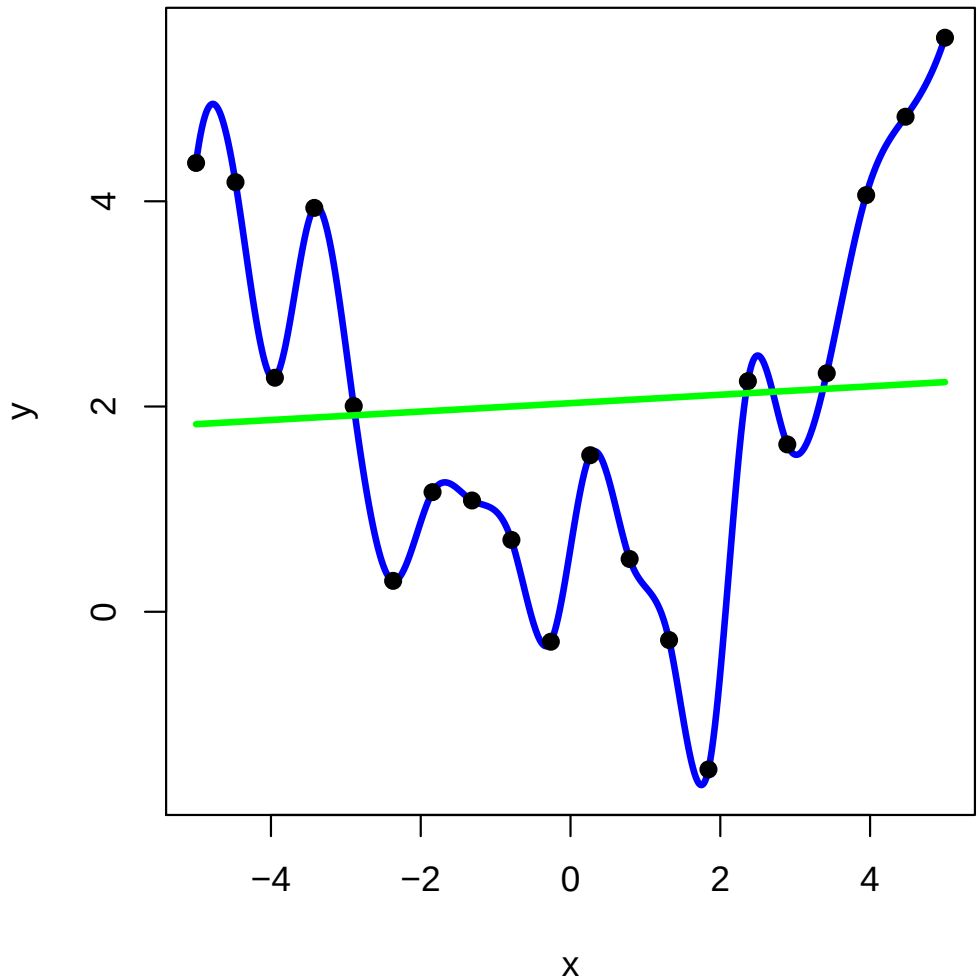


niebieski: $\delta \to +\infty$

Przykład regularyzacji (5)

Dwukrotnie różniczkowalna funkcja f:

$\sum_{i=1}^{n}(f(x_i) - y_i)^2 \to \min$, p.w. $\int_D [f''(x)]^2 dx \leq \delta$

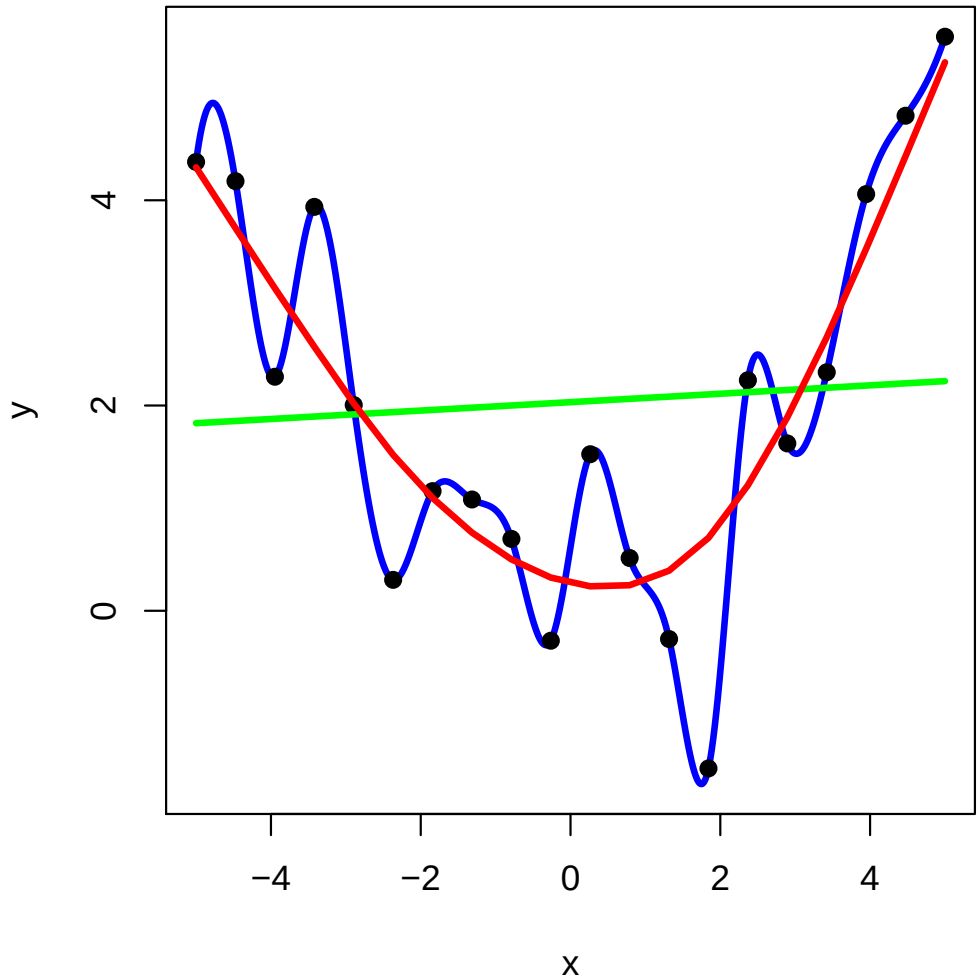


niebieski: $\delta \to +\infty$, zielony: $\delta = 0$

Przykład regularyzacji (6)

Dwukrotnie różniczkowalna funkcja f:

$\sum_{i=1}^{n}(f(x_i)-y_i)^2 \to \min$, p.w. $\int_D [f''(x)]^2 dx \leq \delta$



niebieski: $\delta \to +\infty$, zielony: $\delta = 0$, czerwony: δ optymalne

Sytuacja praktyczna

Ograniczenia twierdzenia Vapnika:

- trudność z wyznaczeniem wartości $h(\mathcal{F})$ dla złożonych klas funkcji
- nierówność z twierdzenia jest bardzo konserwatywna

W praktyce stosujemy zwykle procedury alternatywne:

- kryteria informacyjne (AIC, BIC, ...)
- zbiór walidacyjny
- walidacja krzyżowa
- bootstrapping

Klasyczna ekonometria: model liniowy

- Dysponujemy n obserwacjami i k zmiennymi objaśniającymi
- W modelu liniowym

$$f(x) = \alpha_0 + \sum_{i=1}^{k} \alpha_k x_k$$

zagnieżdżanie klas modeli to wprowadzanie restrykcji na α_k

- Procedury selekcji zmiennych:

$$\min \sum_{i=1}^{n} (f(x_i) - y_i)^2 \qquad p.w. \sum_{j=1}^{k} \mathbf{1}_{\{0\}}(\alpha_j) \leq \delta$$

- tradycyjne kryteria (AIC, BIC, ...) rekomendują wartość δ przy różnych założeniach asymptotycznych
- nie jest możliwe efektywne numerycznie wyznaczanie rozwiązań optymalnych powyższego zadania dla dużych k

Tradycyjne kryteria: porównanie

- AIC: asymptotycznie efektywny, ale nie asymptotycznie zgodny
- BIC: asymptotycznie zgodny, ale nie asymptotycznie efektywny

Standardowe modyfikacje:

wielkość próby	dobra specyfikacja	zła specyfikacja
duża	AIC Akaike (1974)	TIC (Takeuchi, 1978)
mała	AICc (Hurvich i Tsai, 1989)	MAIC (Fujikoshi i Satoh, 1997)

LASSO (Tibshirani, 1996)

Przykładowy alternatywny sposób nakładania restrykcji na parametry:

$$\min \sum_{i=1}^{n}(f(x_i) - y_i)^2 \qquad p.w. \sum_{j=1}^{k} |\alpha_j| \leq \delta$$

- Procedura efektywna numerycznie (Osborne et. al, 2000)
- Metoda prawie prawidłowo identyfikuje niezerowe zmienne (Candes i Plan, 2009)
- Interpretacja w języku optymalizacji odpornej (Fertis, 2009):

$$\max_{||\Delta x||_{1,2} \leq \lambda} \sum_{i=1}^{n}(f(x_i + \Delta x) - y_i)^2 \rightarrow \min$$

Błąd: estymacja a prognoza

Generujemy 20 obserwacji zgodnie z zależnością:

$$Y = 1 + \sum_{j=1}^{10} X_j + \varepsilon, \quad \text{gdzie } \varepsilon \sim N(0,1)$$

Potrzebujemy oszacować:

1) wyraz wolny modelu α_0
2) parametry $\alpha_1, \alpha_2, \ldots, \alpha_{10}$ przy zmiennych X_j

Chcemy ocenić:

1) oczekiwany błąd na danych estymacyjnych (R_e)
2) oczekiwany błąd prognozy (R_p)

Metody estymacji

Tradycyjna:
MNK (metoda najmniejszych kwadratów):

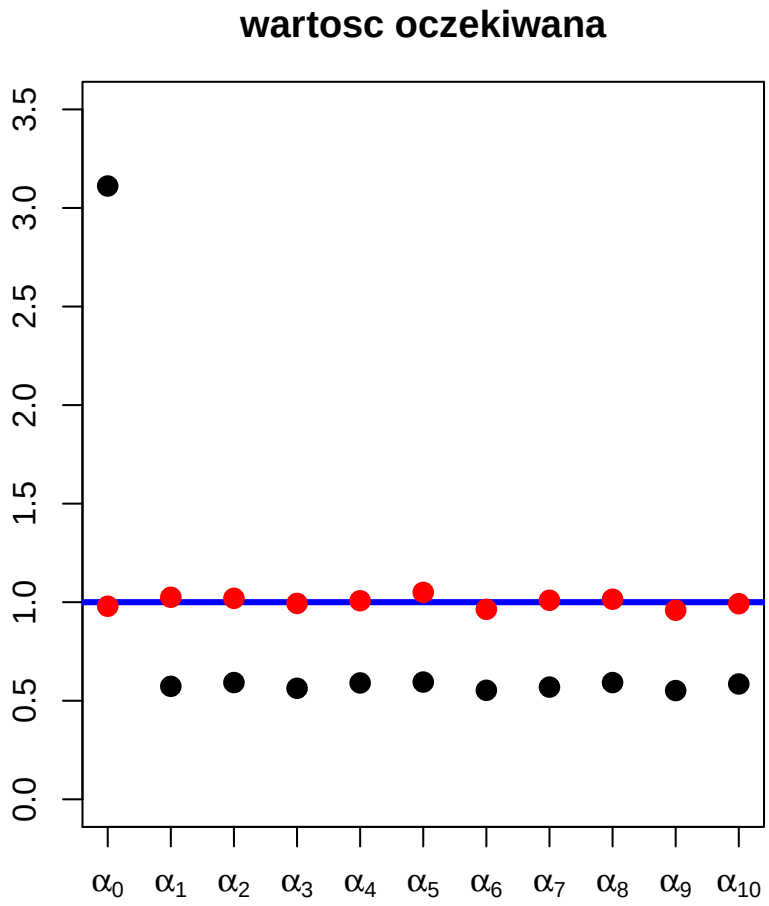
$$\sum_{i=1}^{20} \left(y_i - \left(\alpha_0 + \sum_{j=1}^{10} \alpha_j x_{i,j} \right) \right)^2 \rightarrow \min$$

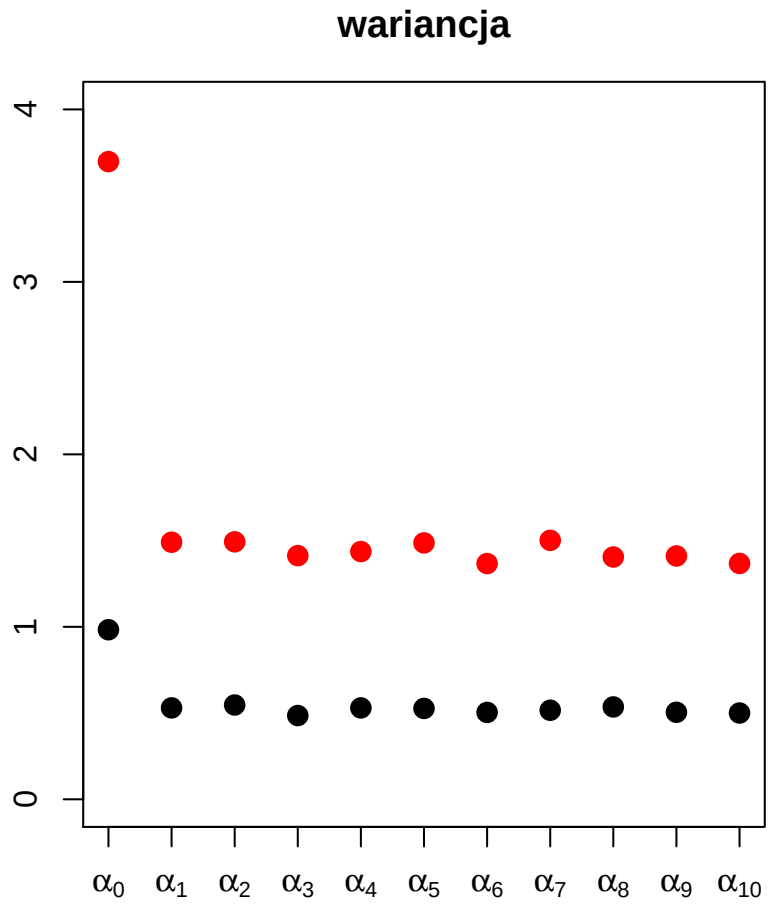
Nowoczesna alternatywa (Tibshirani, 1996):
LASSO (ang. *least absolute shrinkage and selection operator*):

$$\sum_{i=1}^{20} \left(y_i - \left(\alpha_0 + \sum_{j=1}^{10} \alpha_j x_{i,j} \right) \right)^2 \rightarrow \min$$

$$\text{p.w. } \sum_{j=1}^{10} |\alpha_j| \leq \delta$$

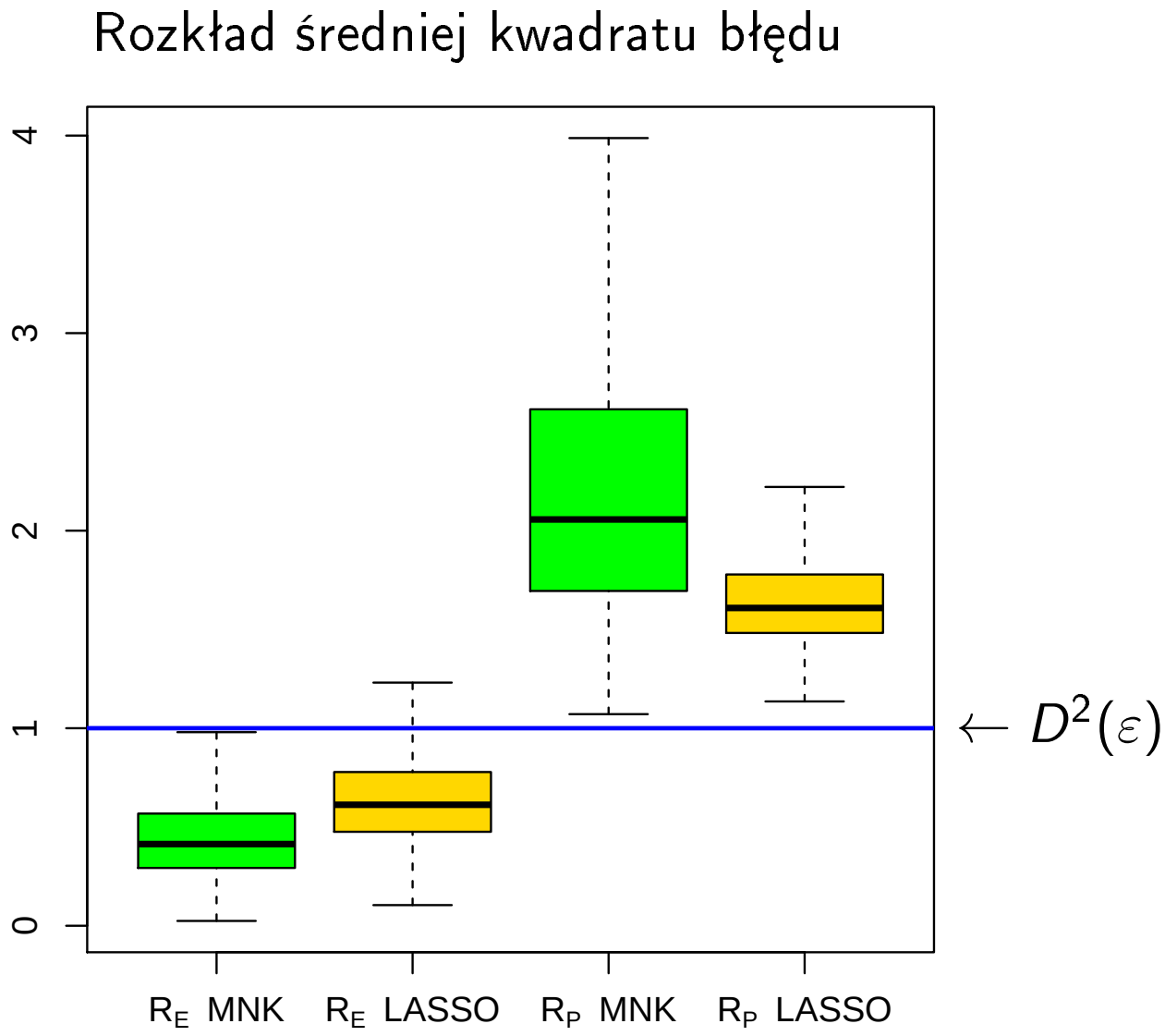
Rozkład estymatorów parametrów





MNK: czerwone LASSO: czarne

Błąd: estymacja a prognoza



Selekcja zmiennych jeśli $k \gg n$ (Belloni et. al, 2014a)

Najprostszy model:

$$y_t = d_t + 0.2x_t + \varepsilon_t$$

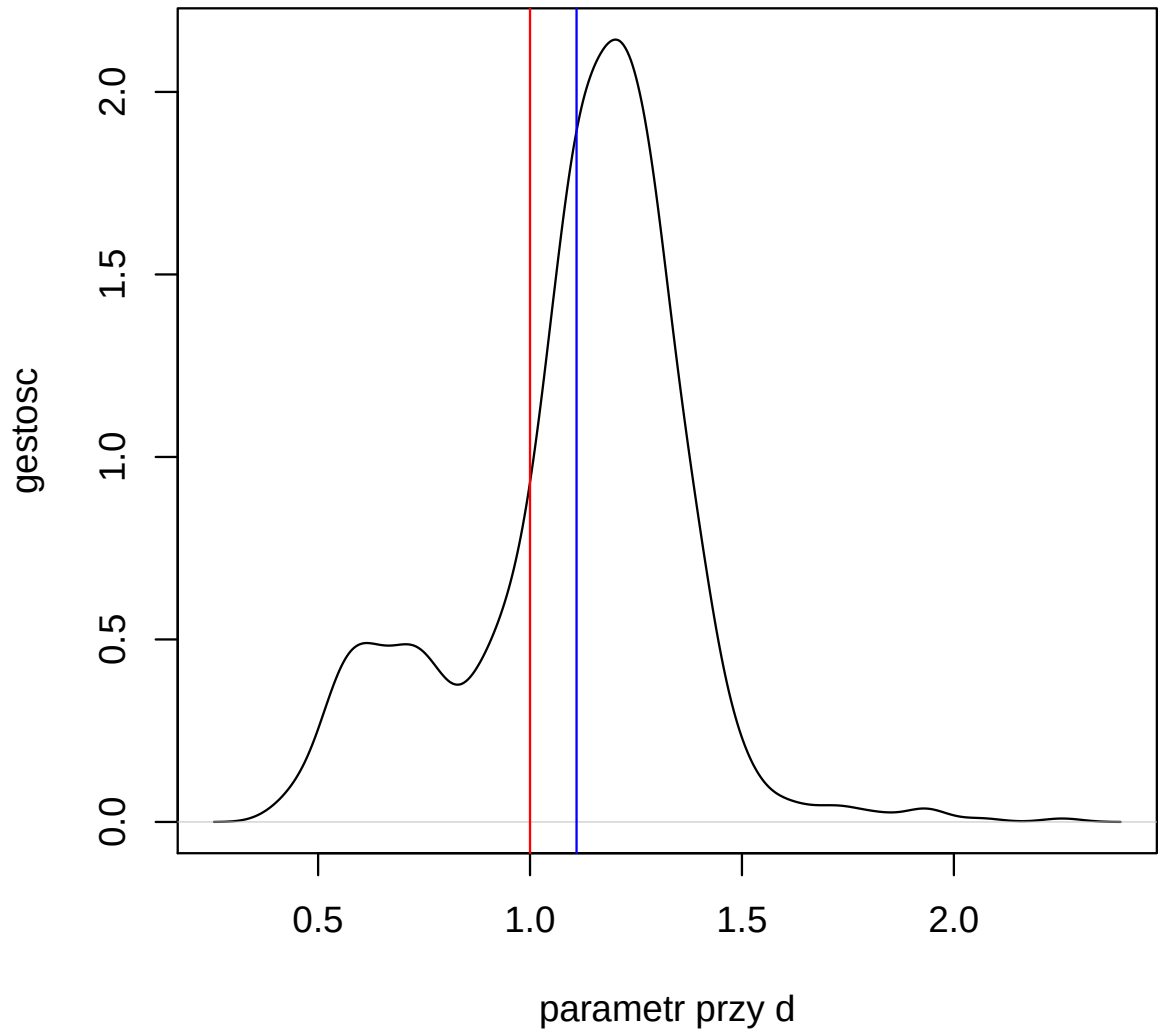
$$d_t = 0.9x_t + \sqrt{1 - 0.9^2}\xi_t$$

gdzie: $\varepsilon_t, \xi_t \sim N(0, 1)$; zakładamy próbę o wielkości $n = 100$.

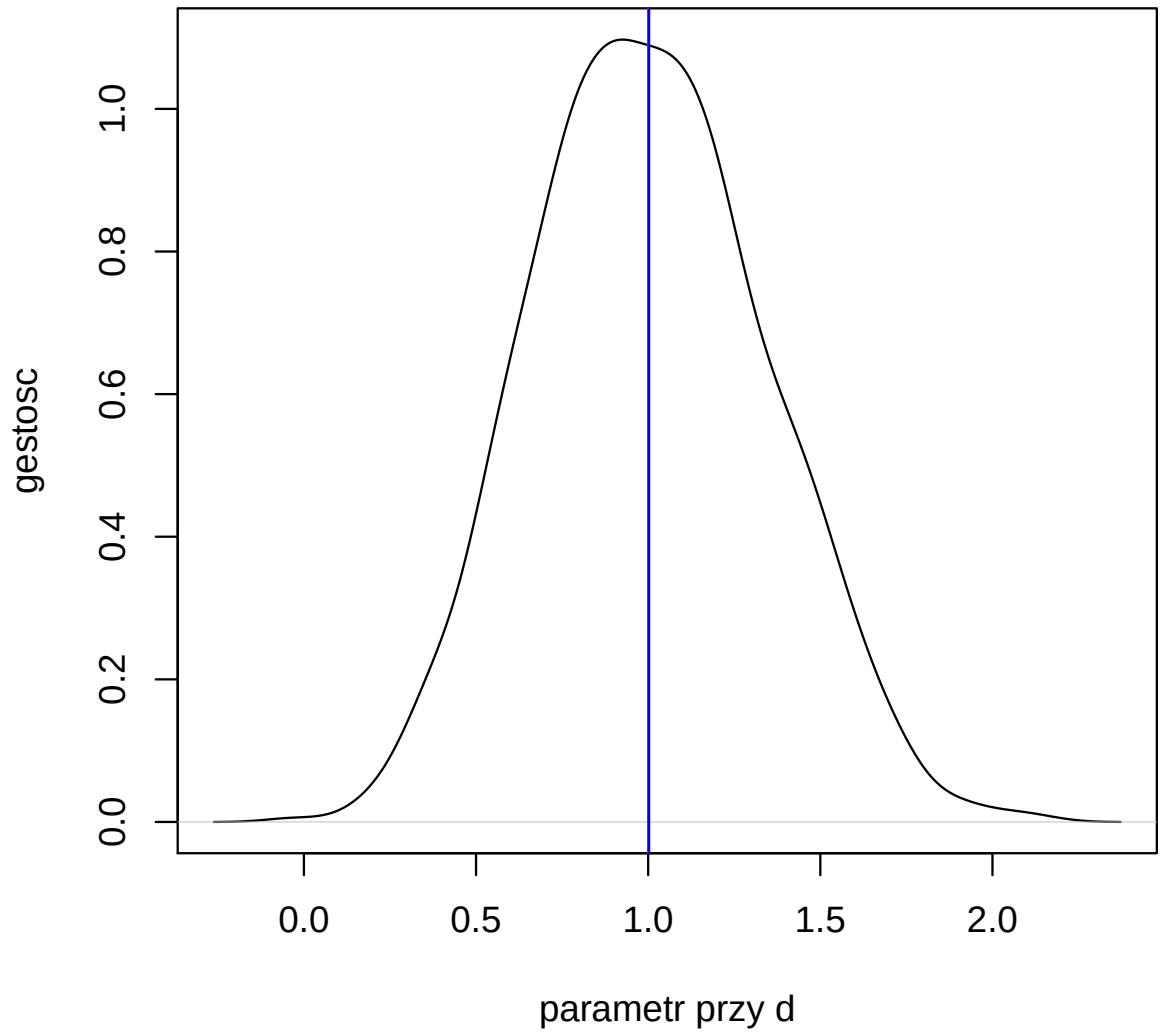
Procedury:

1. pojedyncza selekcja (w modelu na y_t)
2. podwójna selekcja (w modelu na y_t i w modelu na d_t)

Pojedyncza selekcja



Podwójna selekcja



Podwójna metoda post-Lasso (Belloni et. al, 2014b)

Dla modelu:

$$y_t = \alpha d_t + \beta_0 + \beta \mathbf{x}_i + \varepsilon_t$$

$$d_t = \gamma_0 + \gamma \mathbf{x}_t + \xi_t$$

o ile liczba k zmiennych $\mathbf{x}$ spełnia warunek $\log(k) = o\left(n^{1/3}\right)$ wtedy przy niezbyt restrykcyjnych warunkach procedura:

1. dokonaj selekcji zmiennych w obu równaniach za pomocą metody Lasso,
2. dokonaj estymacji pierwszego równania wykorzystując sumę zbiorów zmiennych wybranych w poprzednim kroku,

gwarantuje, że

$$\hat{\sigma}_n^{-1}\sqrt{n}\left(\hat{\alpha} - \alpha\right) \rightsquigarrow N(0,1)$$

Przykładowe zastosowanie

D. Acemoglu, S. Johnson, J.A. Robinson, The Colonial Origins of Comparative Development: An Empirical Investigation, The American Economic Review, 91(5), s. 1369–1401, 2001

śmiertelność osadników
↓
początkowe instytucje
↓
obecne instytucje
↓
obecny PKB per capita

Zmienne kontrolne

- pełny zestaw: efekt instytucji nieistotny
- ograniczony zestaw: efekt instytucji istotny
- podwójna selekcja: efekt instytucji istotny

Uwagi końcowe

1. Konwergencja *kasycznej ekonometrii* i *data-mining*
2. Zagadnienia z bardzo dużą liczbą potencjalnych zmiennych objaśniających
3. Nauczanie: kluczowe zrozumienie *założeń* stosowanych metod

Literatura

[1] Akaike H., A new look at the statistical model identification, IEEE Transactions on Automatic Control, 19(6), s. 716–723, 1974

[2] Belloni A., Chernozhukov V., Hansen Ch., High-Dimensional Methods and Inference on Structural and Treatment Effects, Journal of Economic Perspectives, 28(2), 2014

[3] Belloni A., Chernozhukov V., Hansen Ch., Inference on Treatment Effects after Selection among High-Dimensional Controls, The Review of Economic Studies, 81(2), s. 608–650, 2014

[4] Candes E.J., Plan Y., Near-ideal model selection by ℓ_1 minimization, The Annals of Statistics, 37, s. 2145–2177, 2009

[5] Fertis A.G., A Robust Optimization Approach to Statistical Estimation Problems, rozprawa doktorska, 2009

[6] Fujikoshi Y. and Satoh K., Modified AIC and C_p in multivariate linear regression, Biometrika, 84, s. 07–716, 1997

[7] Hurvich C. M. and Tsai C. L., Regression and time series model selection in small samples, Biometrika, 76, 297–307, 1989

[8] James G., Witten D., Hastie T., and Tibsirani R., An Introduction to Statistical Learning, 2013

[9] Osborne M.R., Presnell B., Turlach B.A., On the LASSO and its Dual, Journal of Computational and Graphical Statistics, 9, s. 319–337, 2000

[10] Takeuchi K., Distribution of information Statistics and Criteria for Adequacy of Models, Mathematical Science, 153, s. 12—18, 1976

[11] Tibshirani R.: Regression shrinkage and selection via the lasso, J. Royal. Statist. Soc B., 58(1), s. 267–288, 1996

[12] Vapnik V., The Nature of Statistical Learning Theory, Springer, New York, 1995

[13] Vapnik V., An Overview of Satistical Learning Theory, IEEE Transactions on Neural Networks, 10(5), s. 988–999, 1999