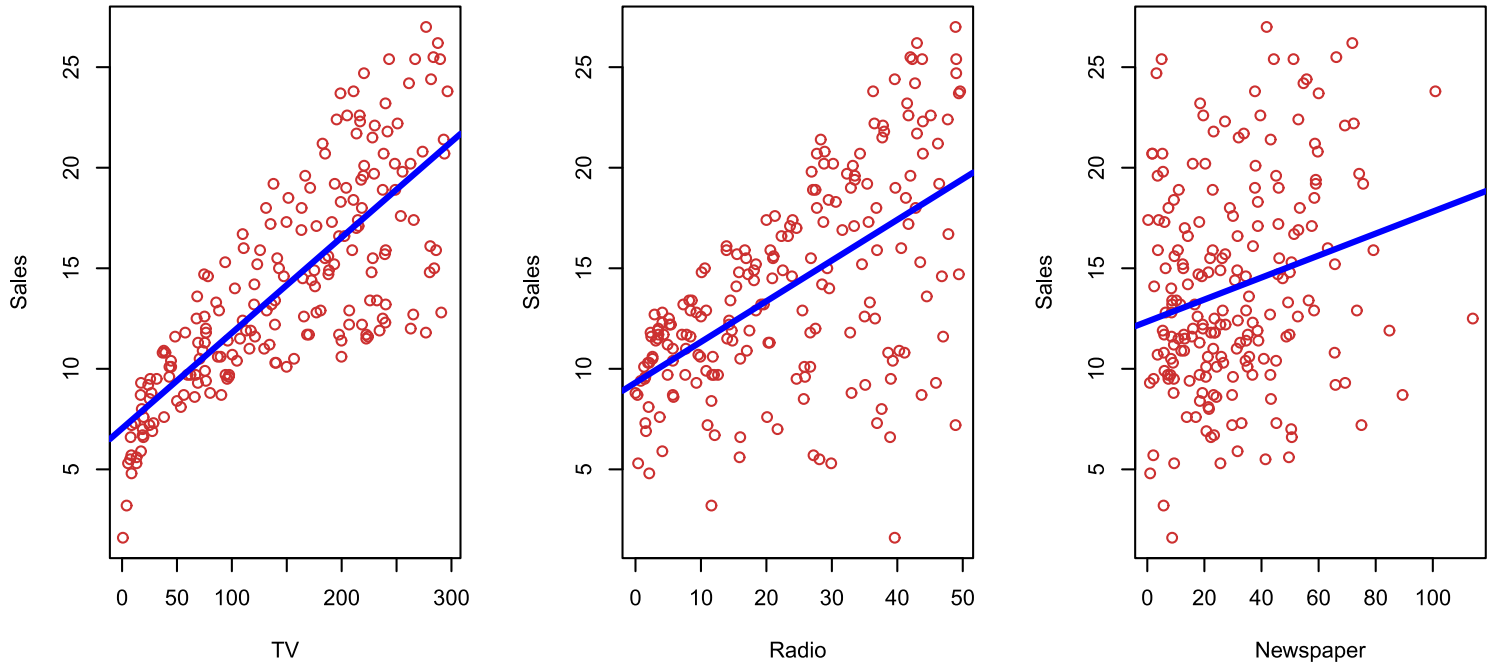
Co je statistické učení?



Zde jsou ukázány `Prodeje (Sales)` versus `TV`, `Rozhlas (Radio)` a `Noviny (Newspaper)` s modrou přímkou lineární regrese pro každý případ jednotlivě. Můžeme předpovídat `Prodeje` pomocí těchto tří diagramů? Možná se nám to povede lépe, použijeme-li nějaký model:

$$\texttt{Prodeje} \approx f(\texttt{TV}, \texttt{Rozhlas}, \texttt{Noviny})$$

Označení

Zde jsou `Prodeje` *odpověď* nebo *cílová hodnota*. Odpověď obvykle značíme Y.
`TV` je *charakteristika*, *vlastnost*, *vstup* nebo *prediktor*, označíme to X_1.
Podobně označíme `Rozhlas` jako X_2 a tak dále.
Můžeme odkazovat na *vstupní vektor* souhrnně jako na

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}.$$

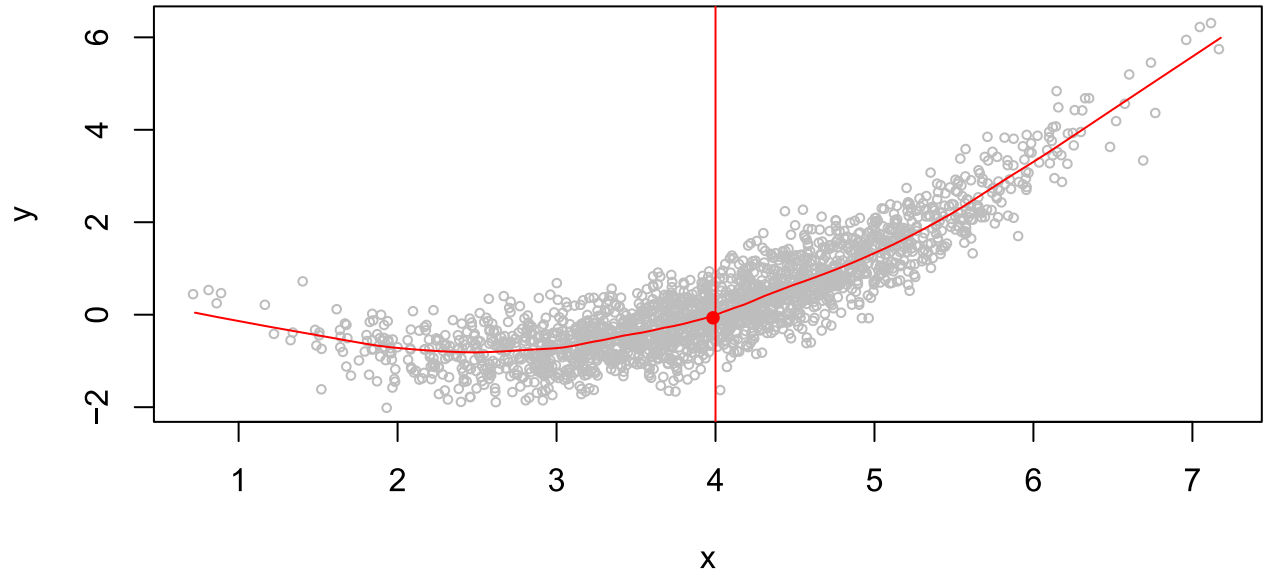
Náš model nyní zapíšeme jako

$$Y = f(X) + \epsilon,$$

kde ϵ zachycuje chyby měření a jiné nepřesnosti.

K čemu je $f(X)$ dobré?

- S dobrým f můžeme dělat předpovědi Y v nových bodech $X = x$.
- Můžeme přijít na to, které složky $X = (X_1, X_2, \ldots, X_p)$ jsou pro pochopení Y důležité a které jsou irelevantní. Tak např. `Stáří` a `Roky vzdělávání` mají velký vliv na `Příjem`, ale `Rodinný stav` typicky ne.
- V závislosti na složitosti funkce f můžeme být schopni pochopit, jak každá složka X_j vektoru X ovlivňuje Y.



Existuje zde ideální $f(X)$? Konkrétně, co je dobrou hodnotou $f(X)$ pro libovolně zvolenou hodnotu X, řekněme $X = 4$? V bodě $X = 4$ může být mnoho hodnot Y. Dobrá hodnota funkce f je

$$f(4) = E(Y|X = 4).$$

$E(Y|X = 4)$ znamená *očekávanou hodnotu* (průměr) hodnot Y pro dané $X = 4$.
Tato ideální funkce $f(x) = E(Y|X = x)$ se nazývá *regresní funkce*.

Regresní funkce $f(x)$

- Je také definována pro vektor X; např. $f(x) = f(x_1, x_2, x_3) = E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$
- Je to *ideální* nebo *optimální* prediktor Y vzhledem ke střední kvadratické chybě: $f(x) = E(Y|X = x)$ je funkce, která minimalizuje $E[(Y - g(x))^2|X = x]$ přes všechny funkce g ve všech bodech $X = x$.
- $\epsilon = Y - f(x)$ je *neredukovatelná* (neodstranitelná) chyba — tj. i kdybychom znali $f(x)$, stejně bychom dělali chyby v předpovídání, neboť v každém bodě $X = x$ typicky existuje rozložení možných hodnot Y.
- Pro každý odhad $\hat{f}(x)$ funkce $f(x)$ máme

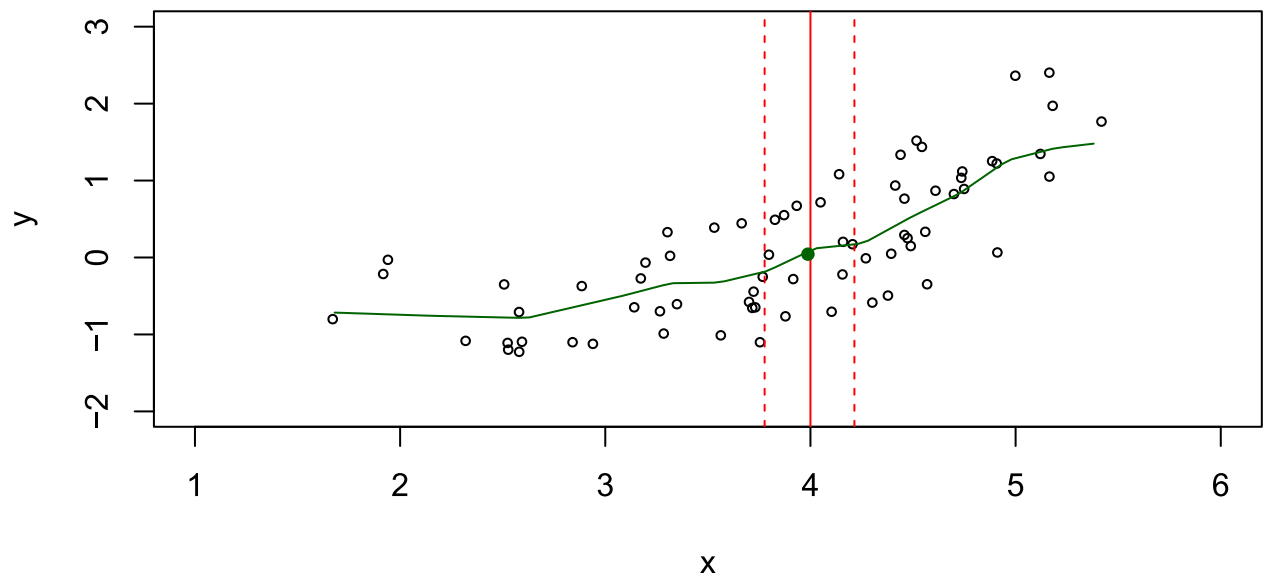
$$E[(Y - \hat{f}(x))^2|X = x] = \underbrace{[f(x) - \hat{f}(x)]^2}_{\textit{redukovatelné}} + \underbrace{\mathrm{Var}(\epsilon)}_{\textit{neredukovatelné}}$$

Jak odhadnout f

- Typicky máme málo bodů pro $X = 4$ přesně (pokud vůbec nějaké).
- Takže nemůžeme spočítat $E(Y|X = x)$!
- Zmírněme definici a položme

$$\hat{f}(x) = \text{Ave}(Y|X \in \mathcal{N}(x))$$

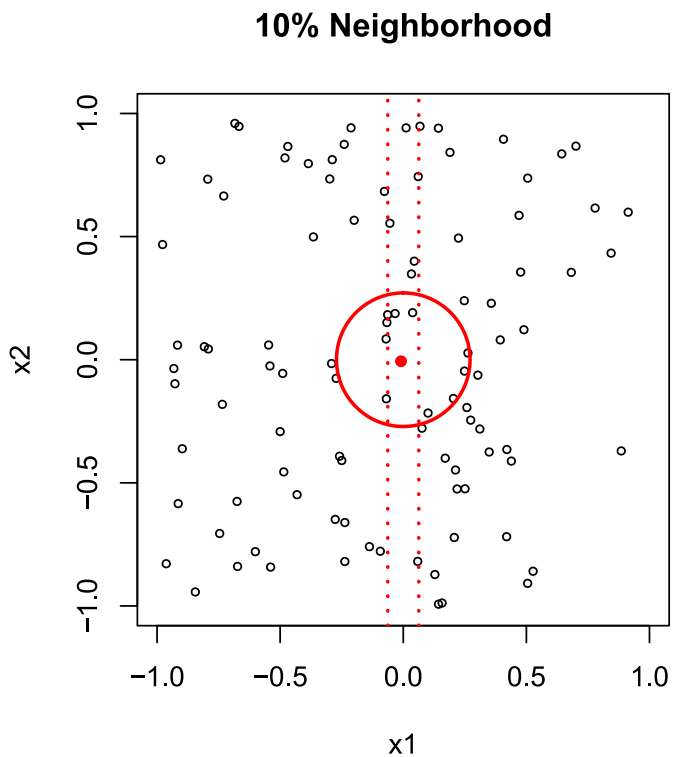
kde $\mathcal{N}(x)$ je nějaké *okolí* bodu x.

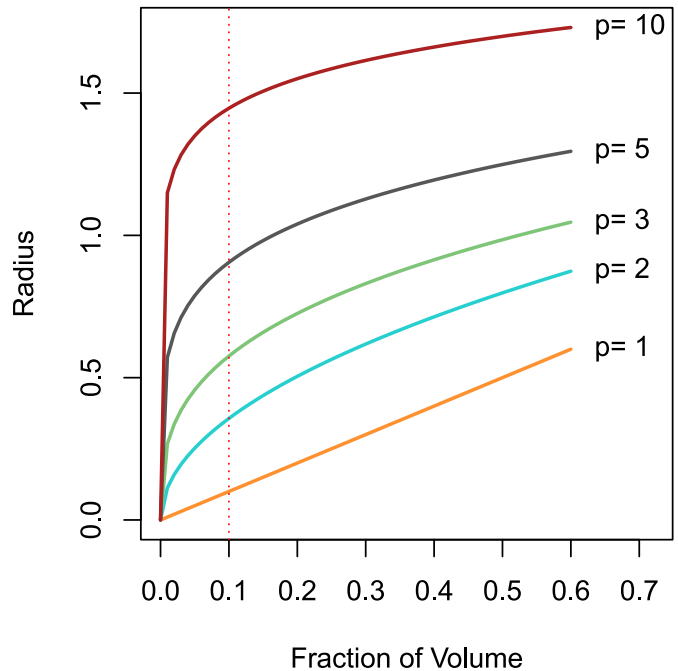


Jak odhadnout f – pokračování

- Metoda nejbližších sousedů může být docela dobrá pro malá p a spíše velká N.
- V tomto kurzu budeme později diskutovat o různých způsobech vyhlazování dat, jako je například jádrové vyhlazování a vyhlazování splajny.
- Metoda nejbližších sousedů může být *velmi špatná*, je-li p velké. Důvod: *prokletí dimensionality*. Ve více dimenzích mají nejbližší sousedé tendenci být hodně daleko.
  - Abychom snížili rozptyl, potřebujeme ke zprůměrování získat rozumný podíl N hodnot y_i, např. 10 %.
  - 10% okolí ve vysokých dimenzích už nemusí být lokální, takže ztrácíme ducha odhadu $E(Y|X = x)$ lokálním průměrováním.

Prokletí dimensionality





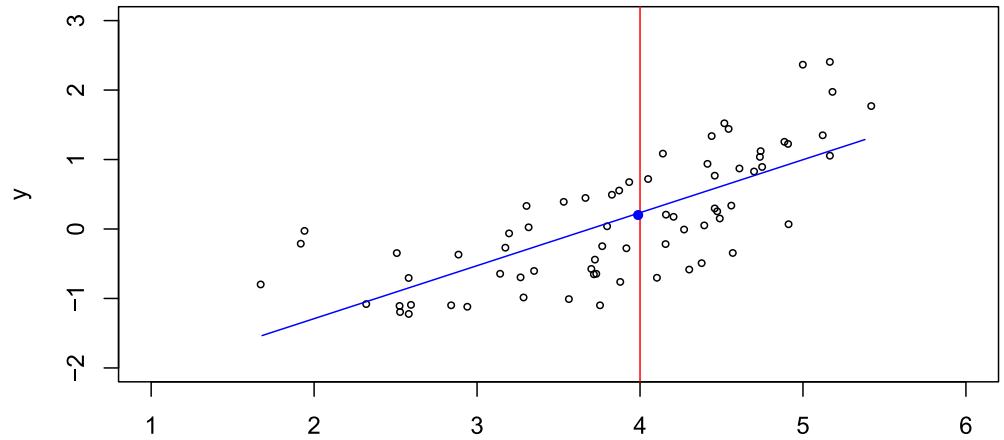
Parametrické a strukturované modely

Lineární model je důležitý příklad parametrického modelu:

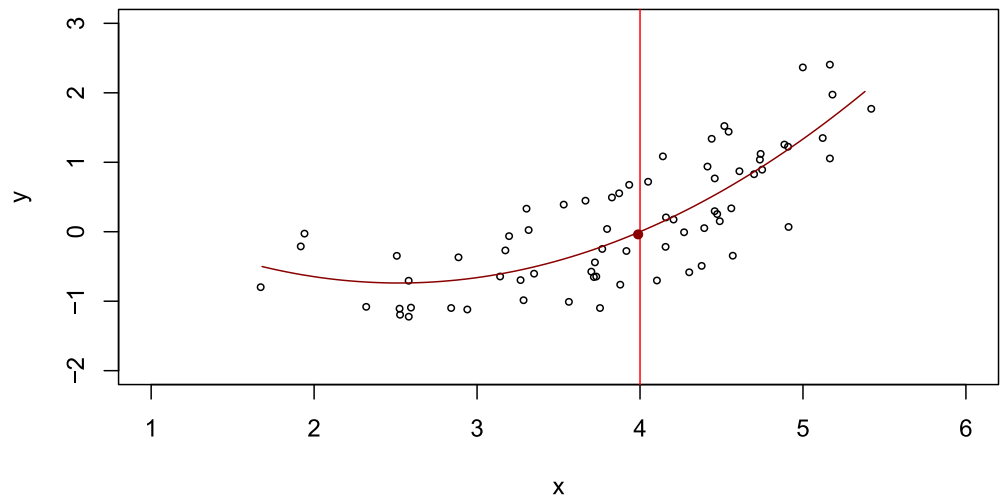
$$f_L(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

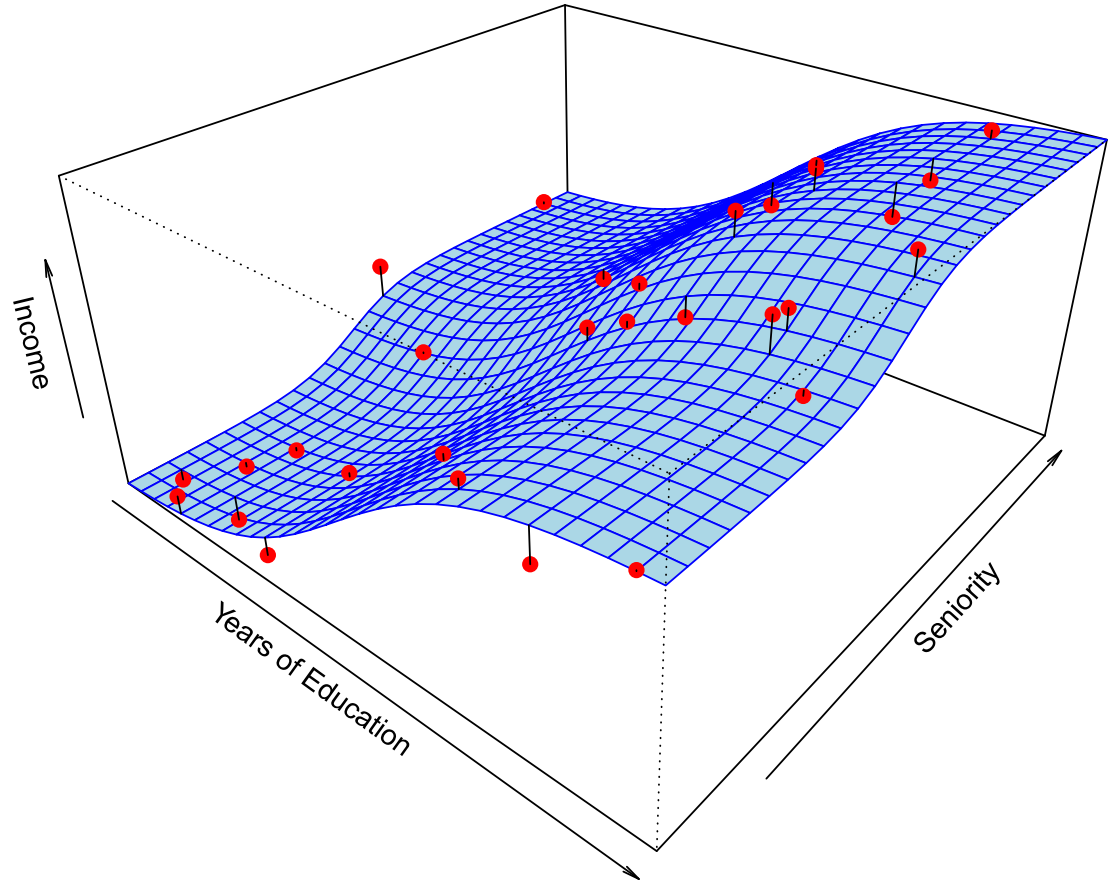
- Lineární model se specifikuje prostřednictvím $p + 1$ parametrů $\beta_0, \beta_1, \ldots, \beta_p$.
- Parametry odhadneme prokládáním modelu trénovacími daty.
- Ačkoli lineární model téměř *nikdy není správný*, často slouží jako dobrá a interpretovatelná aproximace neznámé skutečné funkce $f(X)$.

Lineární model $\hat{f}_L(X) = \hat{\beta}_0 + \hat{\beta}_1 X$ zde dává rozumnou aproximaci:



Kvadratický model $\hat{f}_Q(X) = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$ prochází daty o něco lépe:

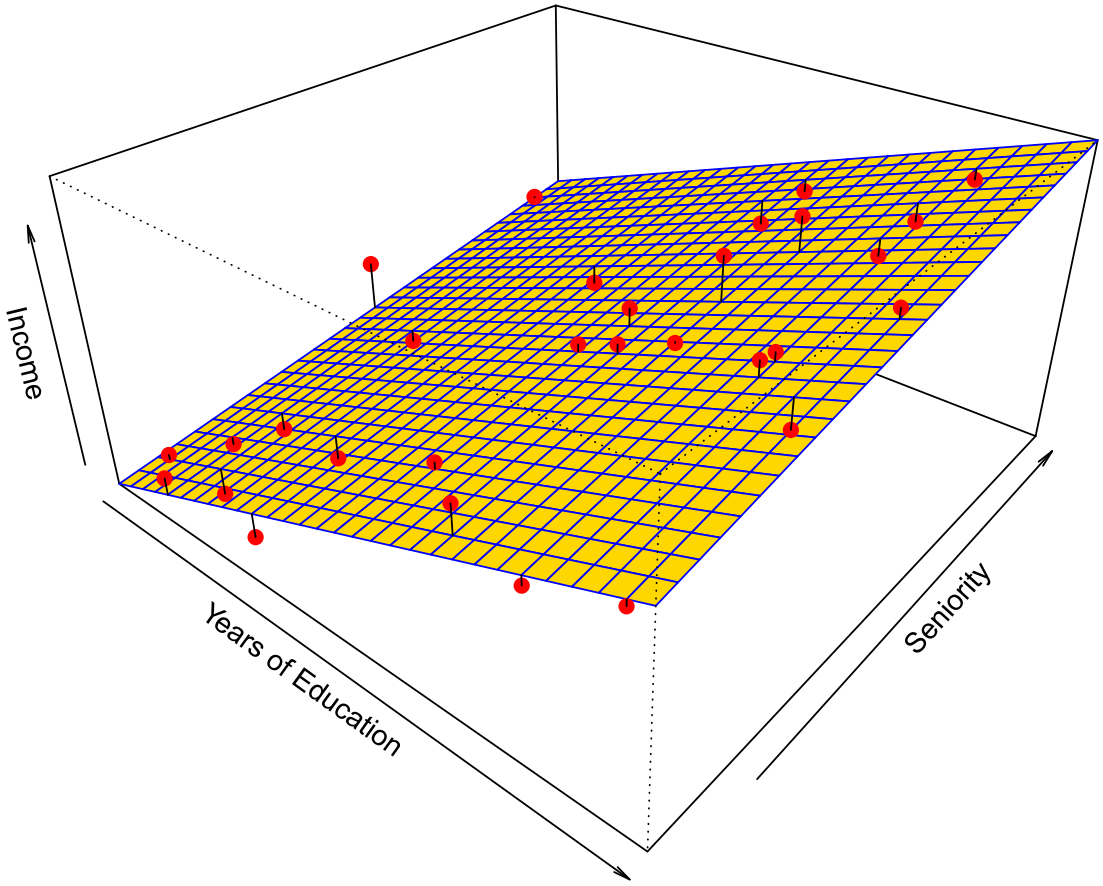




Simulovaný příklad. Červené body jsou simulované hodnoty příjmu z modelu

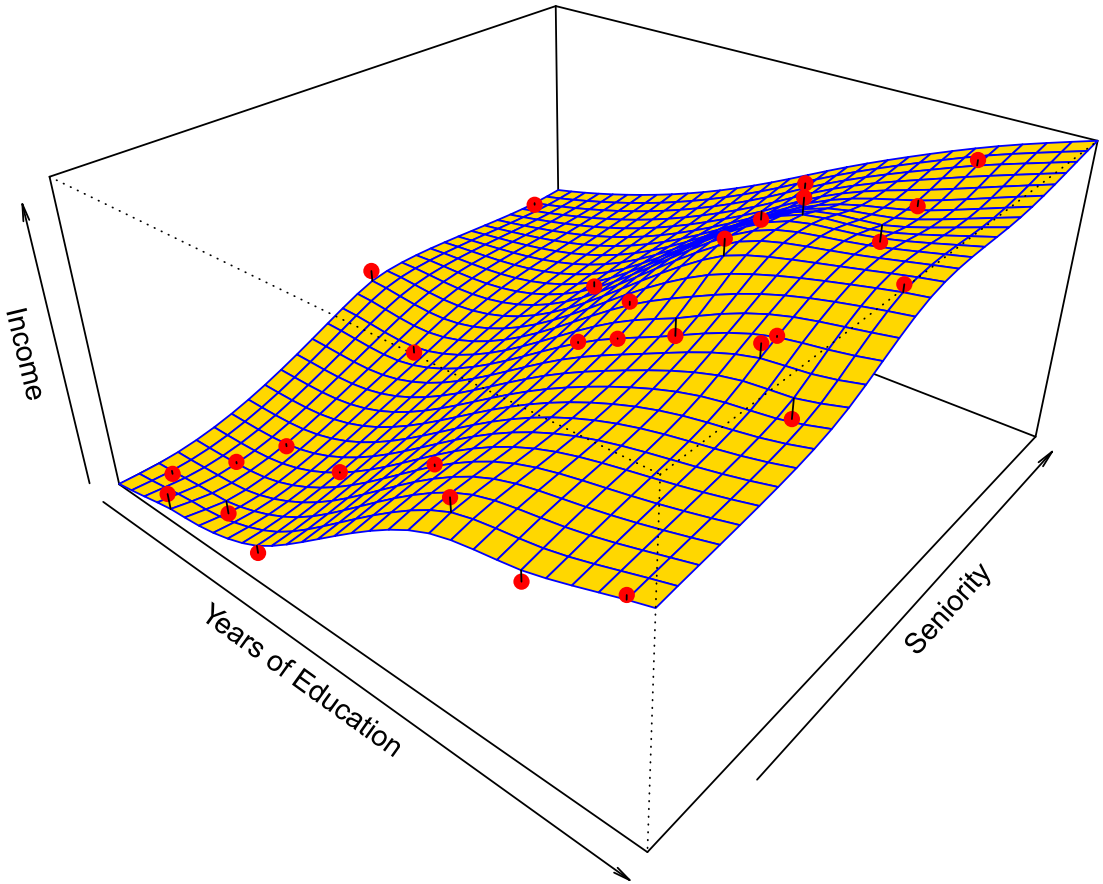
$$\texttt{income} = f(\texttt{education}, \texttt{seniority}) + \epsilon,$$

f je ta modrá plocha.

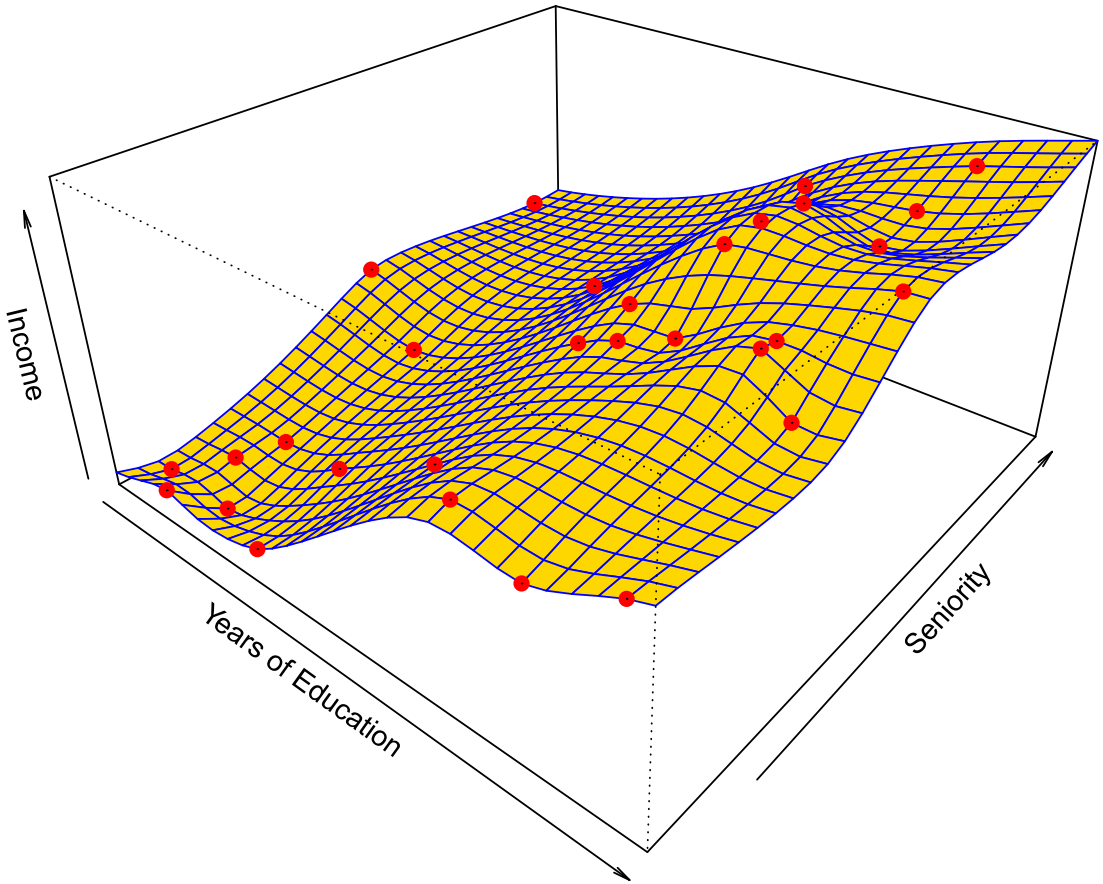


Lineární regresní model proložený nasimulovanými daty:

$$\hat{f}_L(\texttt{education}, \texttt{seniority}) = \hat{\beta}_0 + \hat{\beta}_1 \times \texttt{education} + \hat{\beta}_2 \times \texttt{seniority}$$



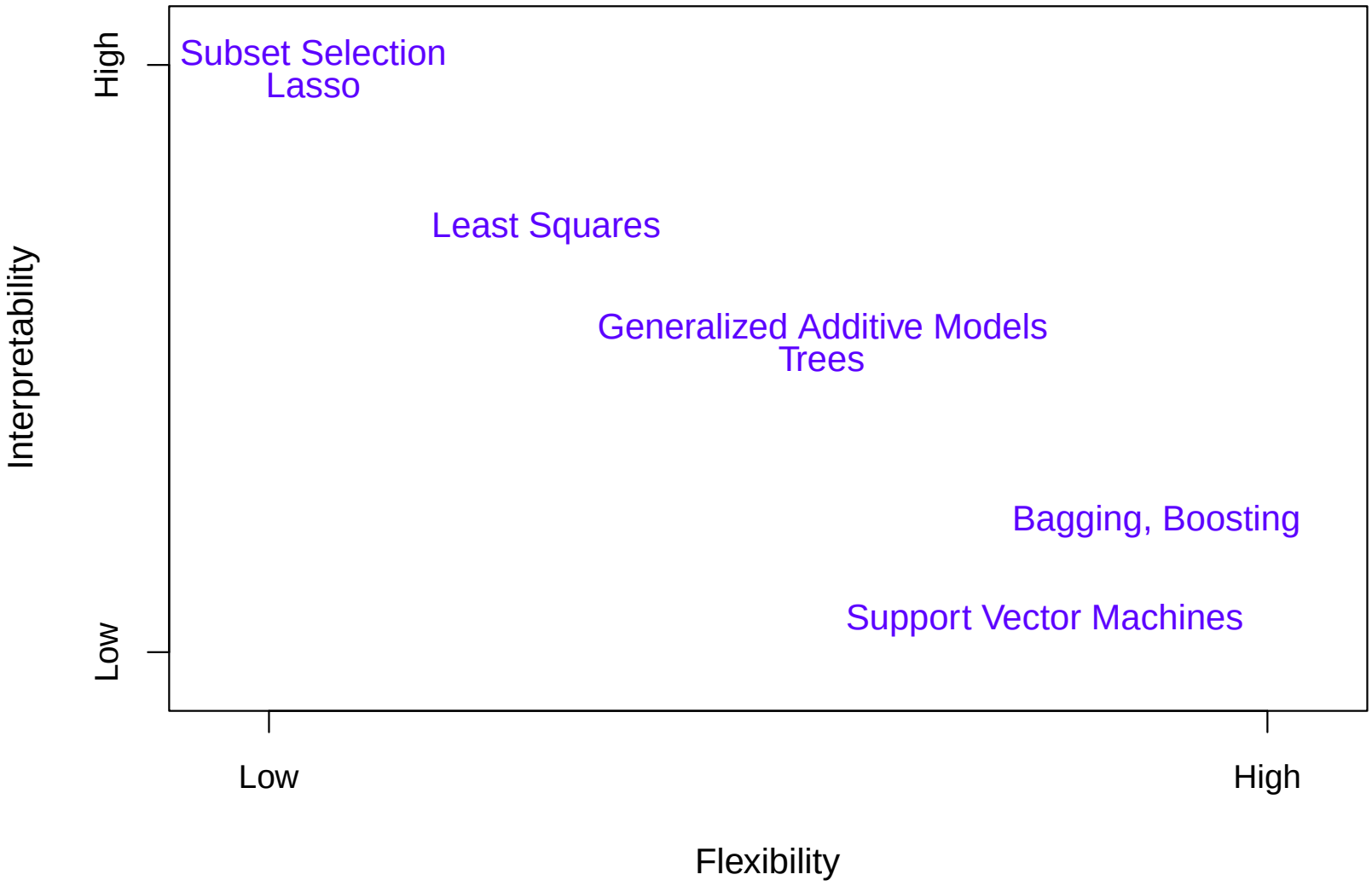
Flexibilnější regresní model $\hat{f}_S(\texttt{education}, \texttt{seniority})$ proložený nasimulovanými daty. Používáme zde postup nazývaný *splajn tenké desky* (thin-plate spline), abychom proložili pružnou plochu. Ovládáme hrbolatost této aproximace (kapitola 7).



Ještě flexibilnější regresní model $\hat{f}_S(\texttt{education}, \texttt{seniority})$ pomocí splajnu proložený nasimulovanými daty. Proložený model zde v trénovacích datech nedělá žádné chyby! Známo také jako *přeurčení* (overfitting).

Některé kompromisy

- Přesnost předpovědi versus interpretovatelnost.
  — Lineární modely se snadno interpretují, splajny tenké desky nikoli.
- Dobrá aproximace versus přeurčení nebo podurčení.
  — Jak poznáme, že aproximace je zrovna ta pravá?
- Úspornost versus černá skřínka.
  — Často dáváme přednost jednoduššímu modelu s méně proměnnými před prediktorem typu černé skřínky, který je zahrnuje všechny.



Překlad názvů metod (zleva doprava, odshora dolů): výběr podmnožiny, Lasso, nejmenší čtverce. zobecněné aditivní modely, stromy, bagging, boosting, metoda podpůrných vektorů.

Posouzení přesnosti modelu

Předpokládejme, že prokládáme nějakými trénovacími daty $\mathtt{Tr} = \{x_i, y_i\}_1^N$ model $\hat{f}(x)$ a chceme vědět, jak dobře si vede.

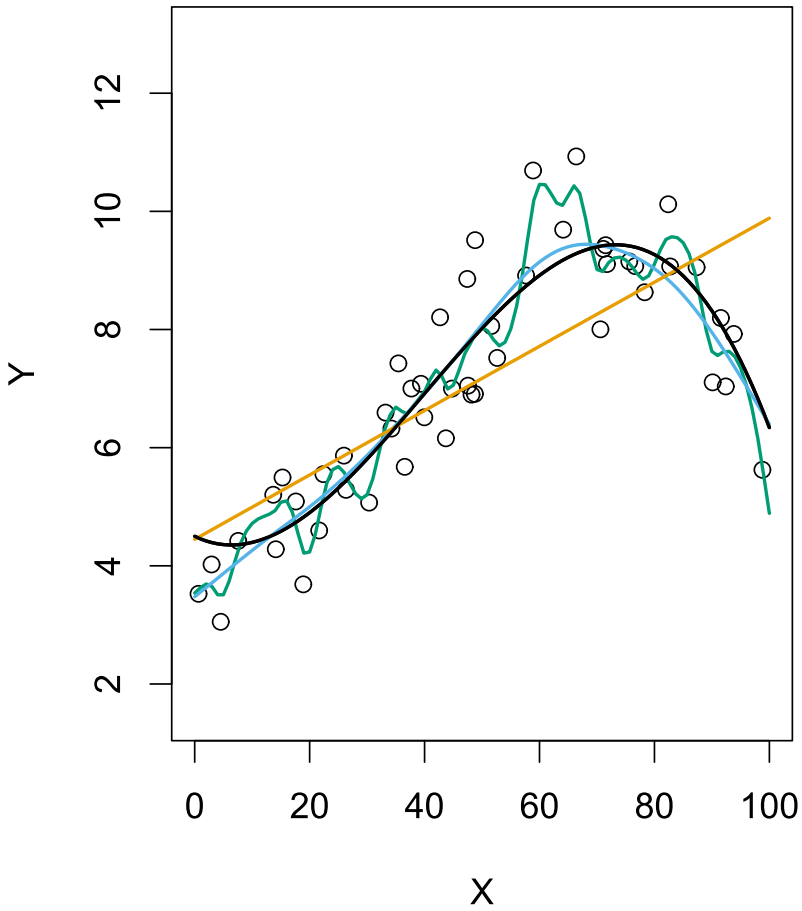
- Mohli bychom vypočítat střední kvadratickou chybu předpovědi přes $\mathtt{Tr}$:

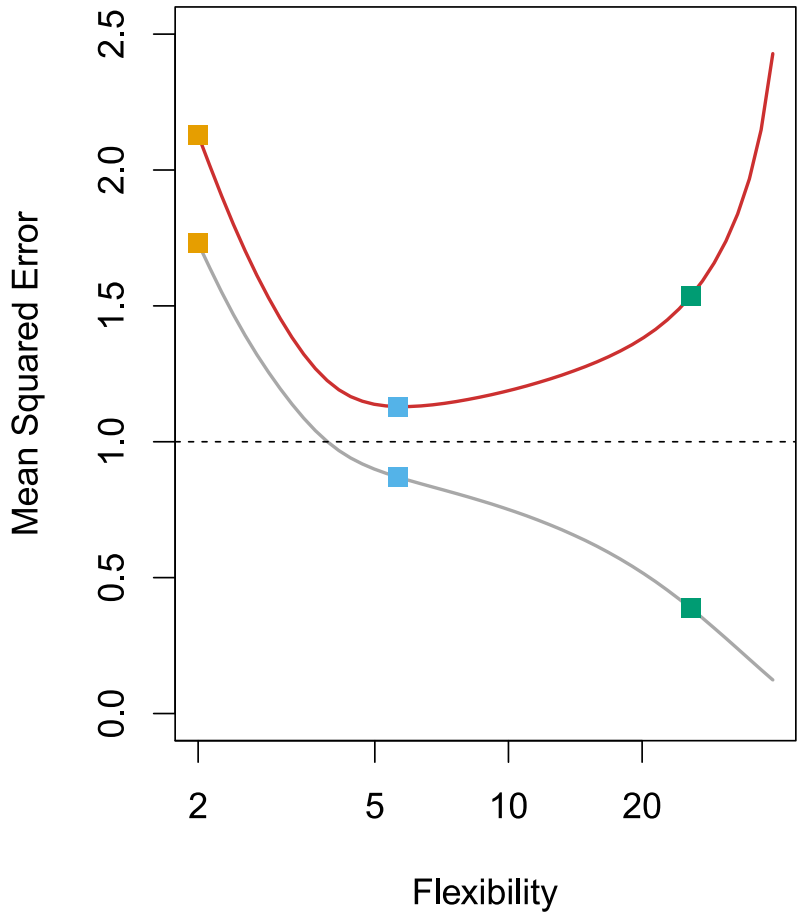
$$\mathrm{MSE}_{\mathtt{Tr}} = \mathrm{Ave}_{i \in \mathtt{Tr}}[y_i - \hat{f}(x_i)]^2$$

To může stranit více přeurčeným modelům.

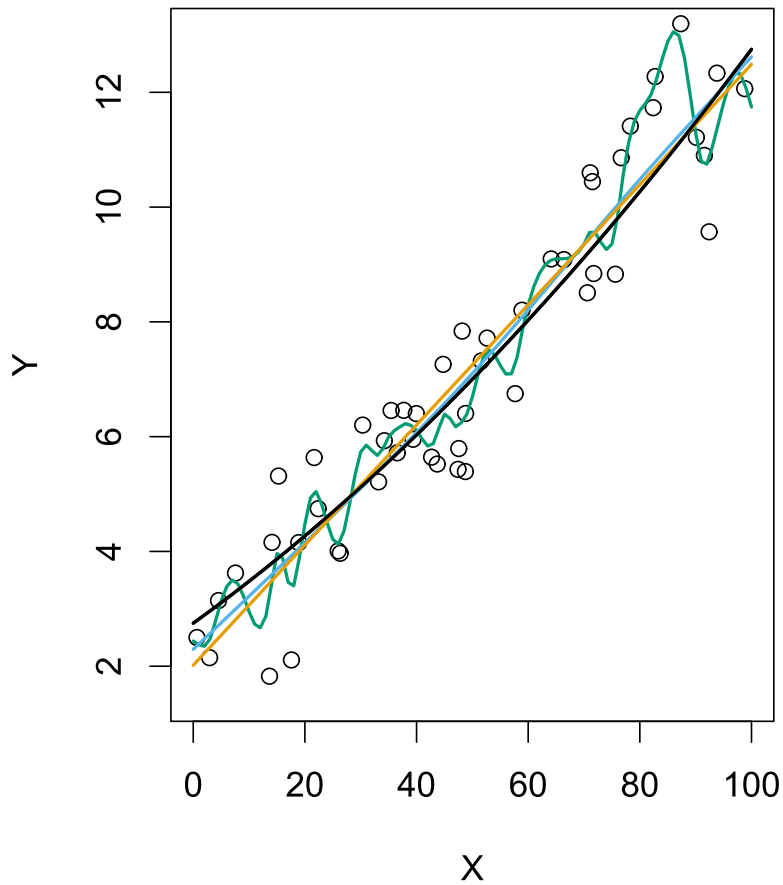
- Místo toho bychom měli, pokud je to možné, vypočítat tu chybu pomocí nových *testovacích* dat $\mathtt{Te} = \{x_i, y_i\}_1^M$:

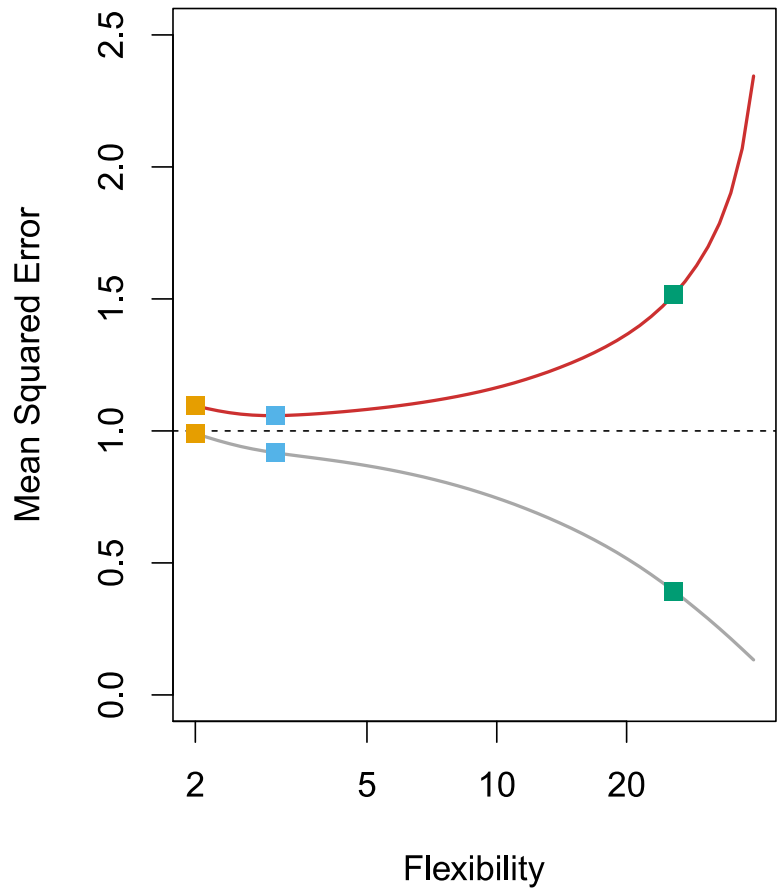
$$\mathrm{MSE}_{\mathtt{Te}} = \mathrm{Ave}_{i \in \mathtt{Te}}[y_i - \hat{f}(x_i)]^2$$



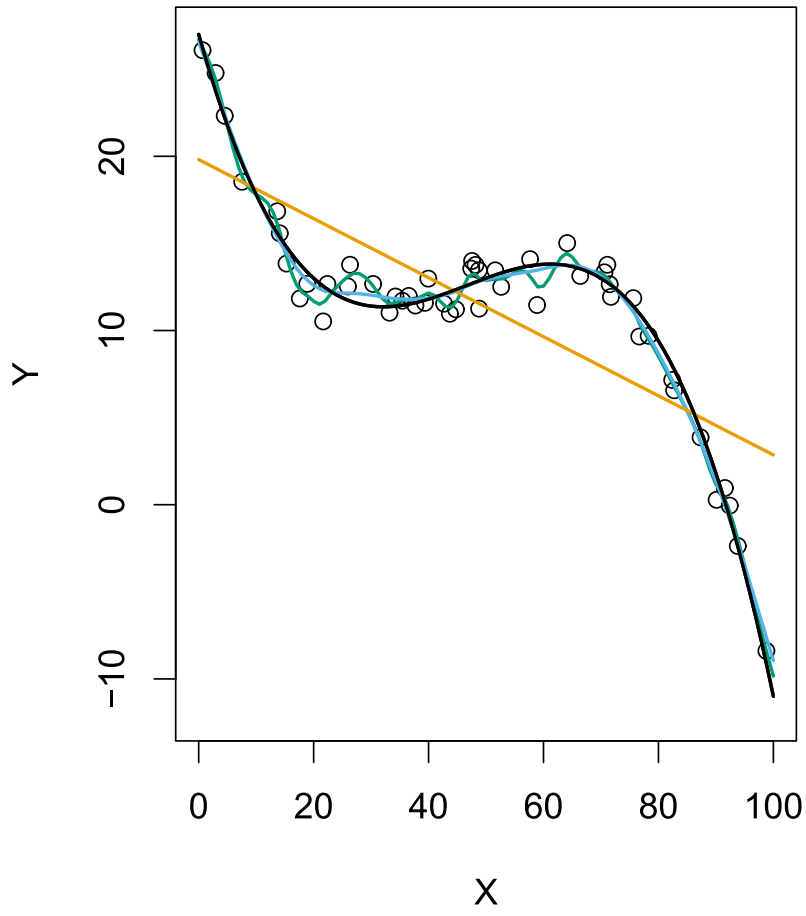


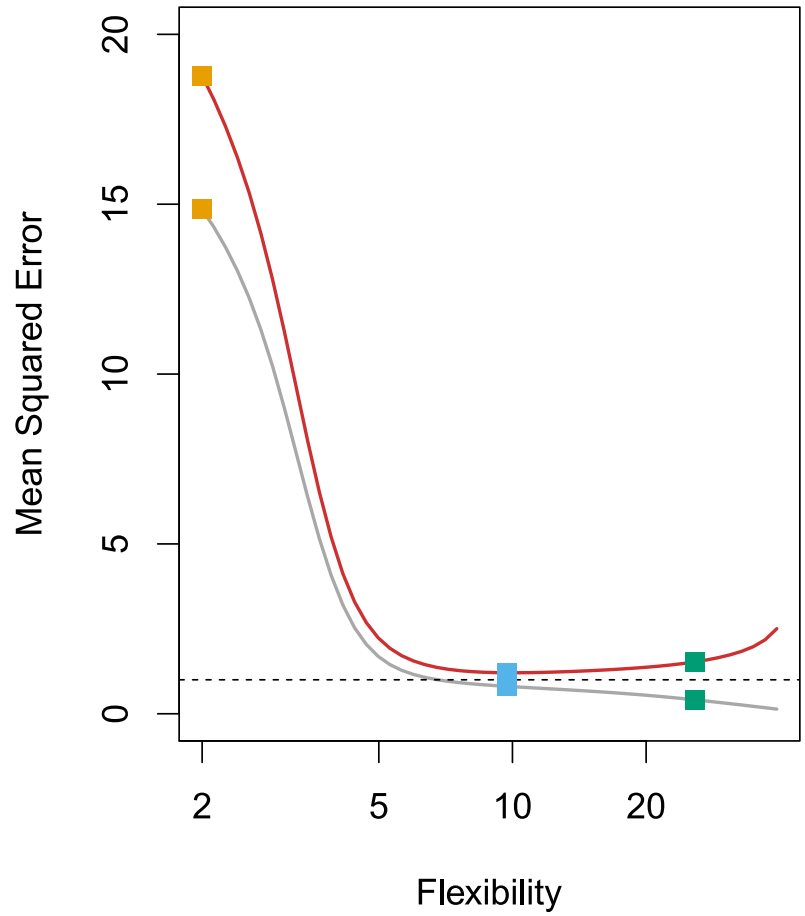
Černá křivka je skutečnost. Červená křivka vpravo je $\mathrm{MSE}_{\mathtt{Te}}$, šedá křivka je $\mathrm{MSE}_{\mathtt{Tr}}$. Oranžová, modrá a zelená křivka (a čtverečky těchto barev) odpovídají aproximacím různé flexibility.





Zde je skutečnost hladší, takže hladší aproximace a lineární model si vedou opravdu dobře.





Zde je skutečnost zvlněná a šum je malý, takže si nejlépe vedou flexibilnější aproximace.

Kompromis mezi zkreslením a rozptylem

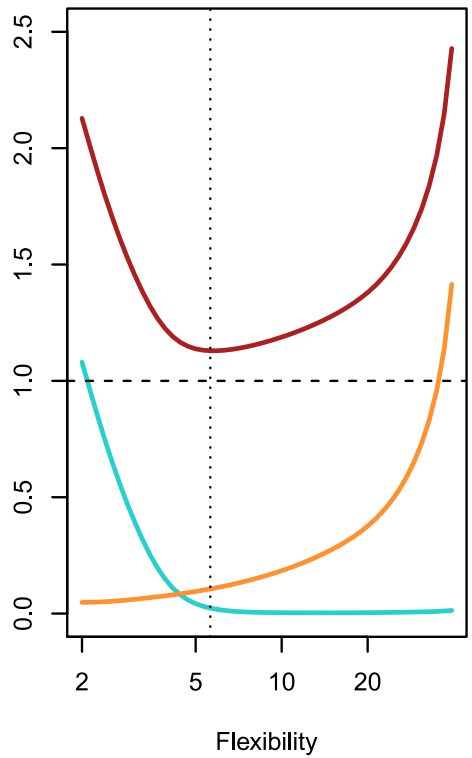
Předpokládejme, že jsme nějakými trénovacím daty `Tr` proložili model $\hat{f}(x)$ a nechť (x_0, y_0) je testovací pozorování vyvozené z této populace. Jestliže skutečný model je $Y = f(X) + \epsilon$ (kde $f(X) = E(Y|X = x)$), pak

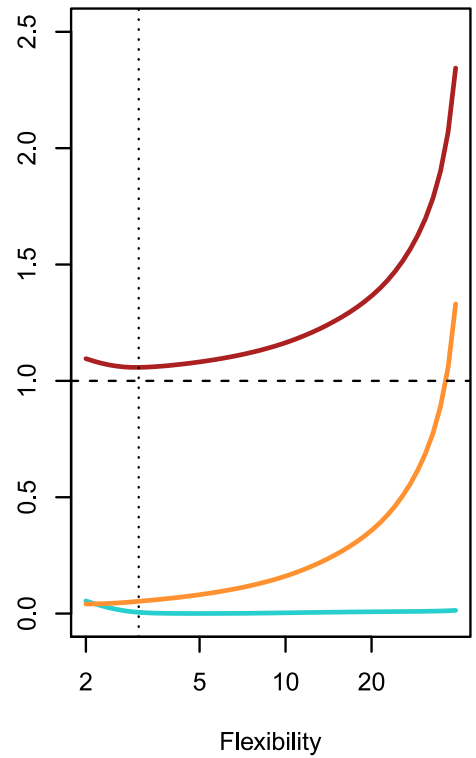
$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\epsilon).$$

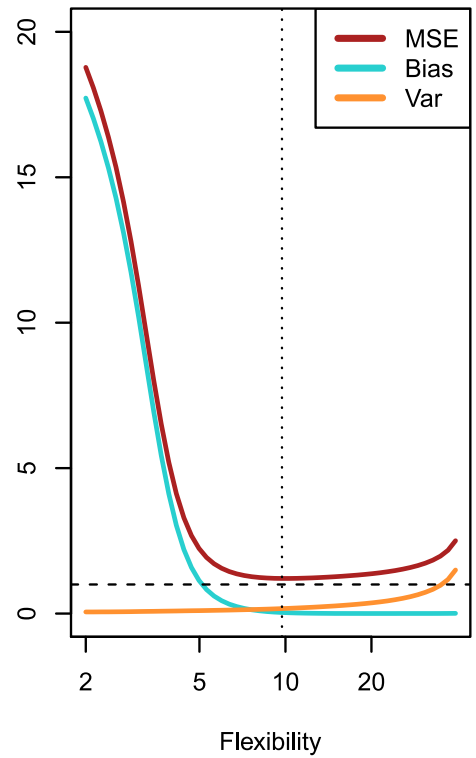
Očekávání počítá průměr přes variabilitu y_0 a rovněž variabilitu v `Tr`. Poznamenáváme, že $\text{Bias}(\hat{f}(x_0)) = E[\hat{f}(x_0)] - f(x_0)$.

Typické je, že jak roste *flexibilita* $\hat{f}$, roste její rozptyl a zkreslení (bias) se snižuje. Takže volba flexibility založená na střední testovací chybě odpovídá *kompromisu mezi zkreslením a rozptylem*.

Kompromis mezi zkreslením a rozptylem na našich třech příkladech



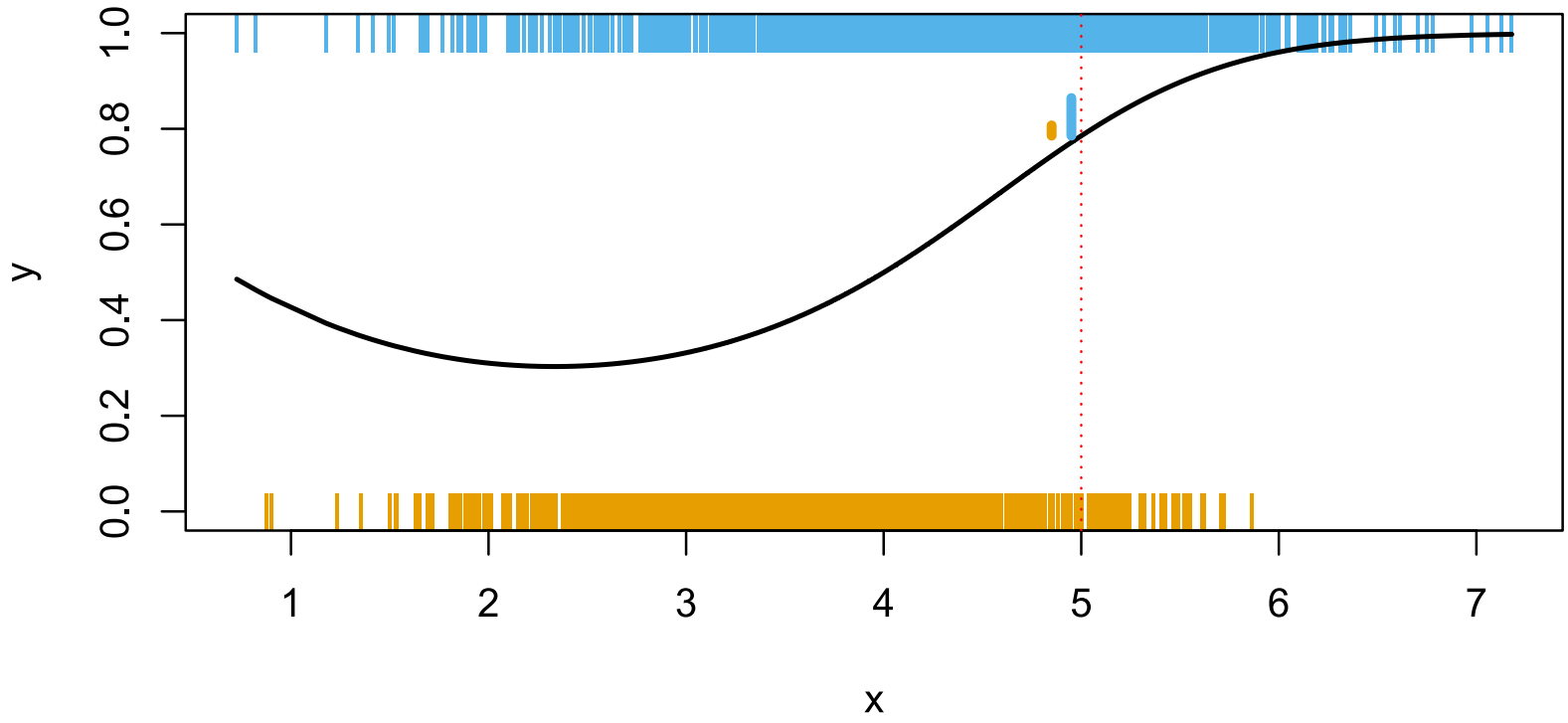




Klasifikační úlohy

Proměnná odpovědi Y je zde *kvalitativní* — např. email je jeden z prvků $\mathcal{C} = (\texttt{spam}, \texttt{ham})$ ($\texttt{ham}$ = dobrý email), třída číslic je jedna z $\mathcal{C} = \{0, 1, \ldots, 9\}$. Naše cíle jsou:

- Vytvořit klasifikátor $C(X)$, který přiřadí značku třídy z $\mathcal{C}$ budoucímu neoznačenému pozorování X.
- Ohodnotit nejistotu v každé klasifikaci.
- Porozumět roli různých prediktorů mezi složkami $X = (X_1, X_2, \ldots, X_p)$.

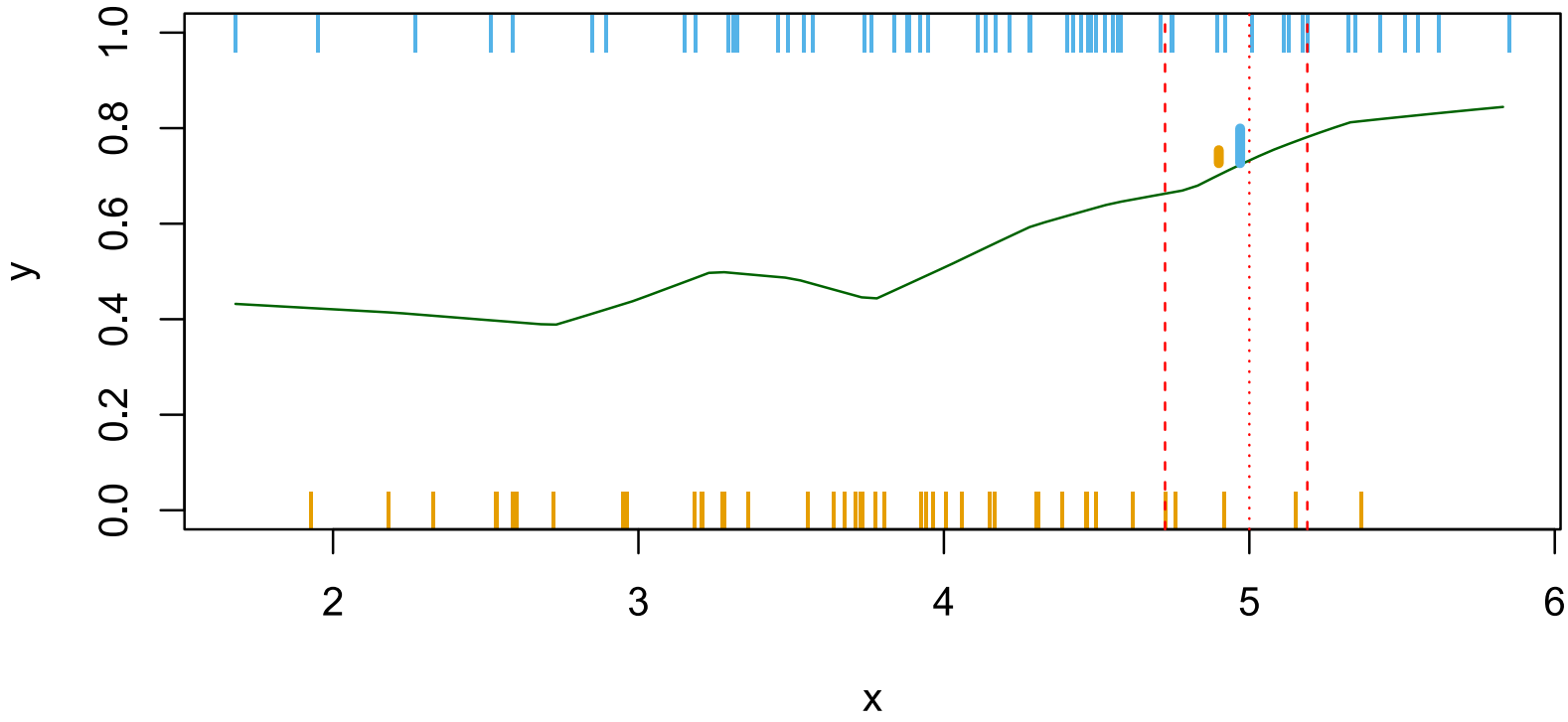


Existuje ideální $C(X)$? Předpokládejme, že K prvků množiny $\mathcal{C}$ je očíslováno $1, 2, \ldots, K$. Položme

$$p_k(x) = \Pr(Y = k|X = x), k = 1, 2, \ldots, K.$$

Toto jsou *podmíněné pravděpodobnosti tříd* pro dané x; viz např. malý sloupcový graf pro $x = 5$. *Bayesův optimální klasifikátor* pro x je pak

$$C(x) = j \quad \text{jestliže} \quad p_j(x) = \max\{p_1(x), p_2(x), \ldots, p_K(x)\}.$$



Dá se zde stejně jako dříve použít průměrování přes nejbližší sousedy.

A pro rostoucí dimenze se to také hroutí. Avšak dopad na $\hat{C}(x)$ je menší než na $\hat{p}_k(x), k = 1, 2, \ldots, K$.

Klasifikace: některé detaily

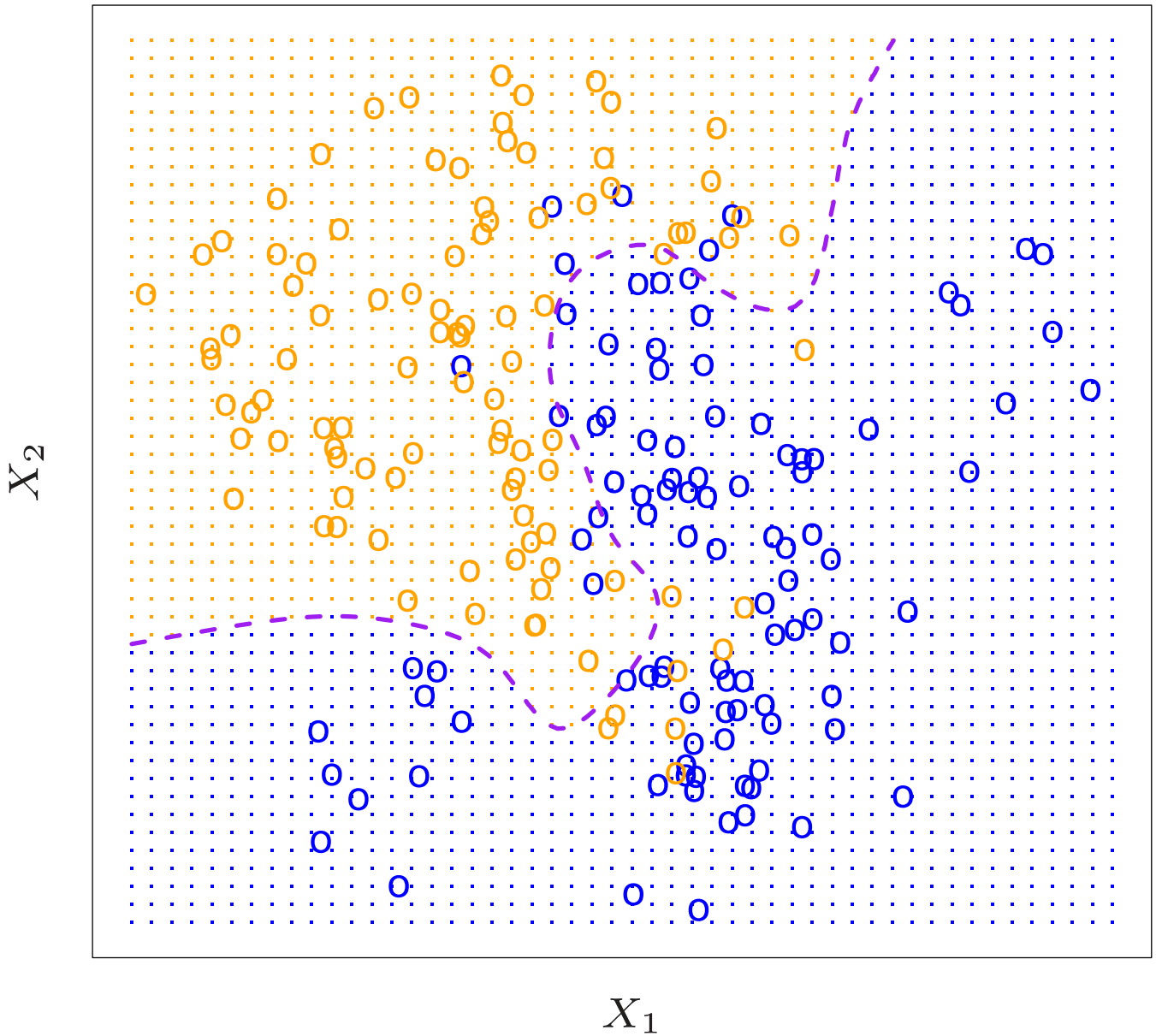
- Výkonnost klasifikátoru $\hat{C}(x)$ typicky měříme pomocí míry chyby nesprávné klasifikace:

$$\mathrm{Err}_{\mathtt{Te}} = \mathrm{Ave}_{i \in \mathtt{Te}} I[y_i \neq \hat{C}(x_i)].$$

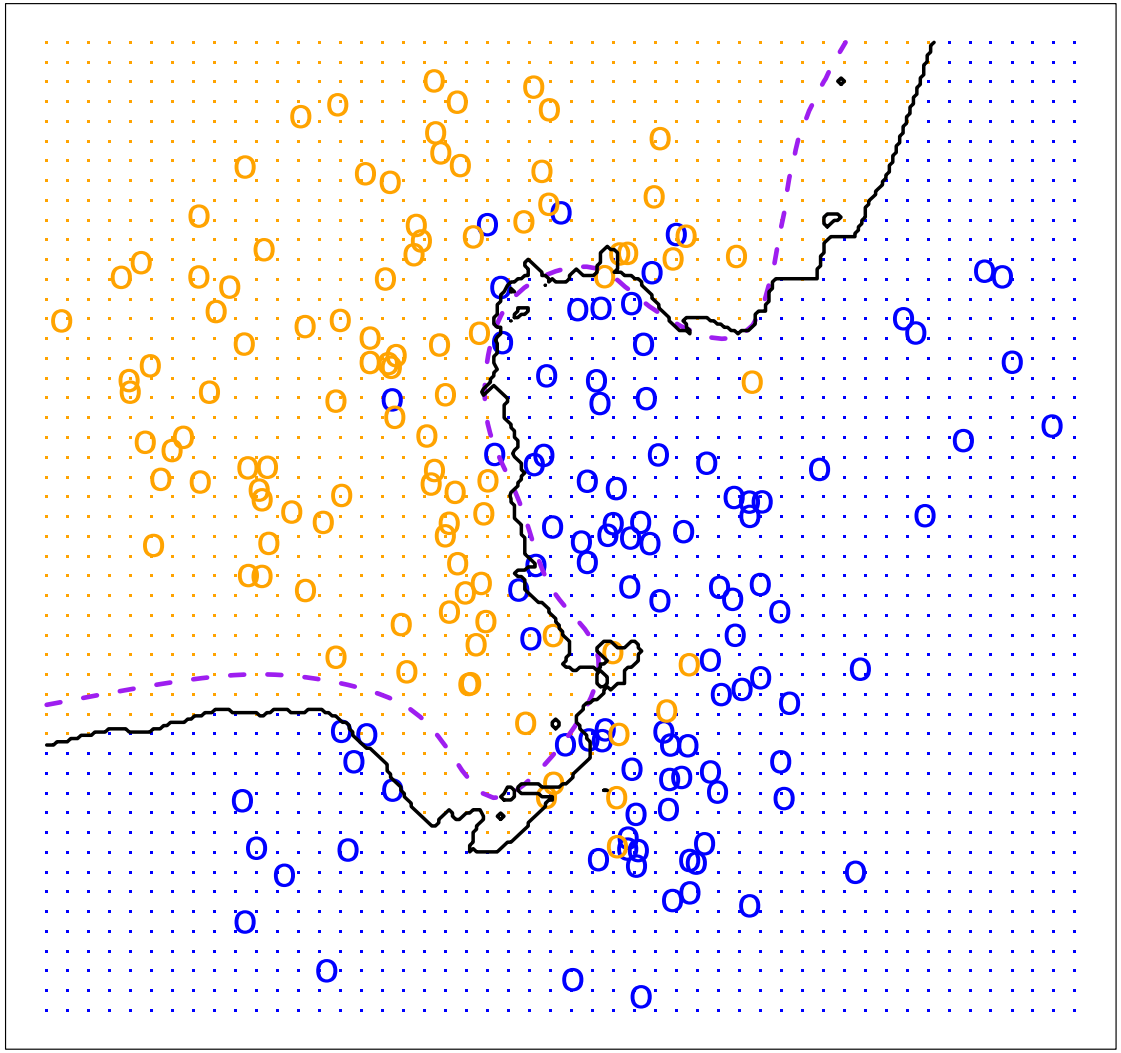
- Bayesův klasifikátor (užívající skutečné hodnoty $p_k(x)$) má nejmenší chybu (v dané populaci).

- Metoda podpůrných vektorů vytváří strukturované modely $C(x)$.
- Budeme také vytvářet strukturované modely pro reprezentaci $p_k(x)$, např. logistickou regresi, zobecněné aditivní modely.

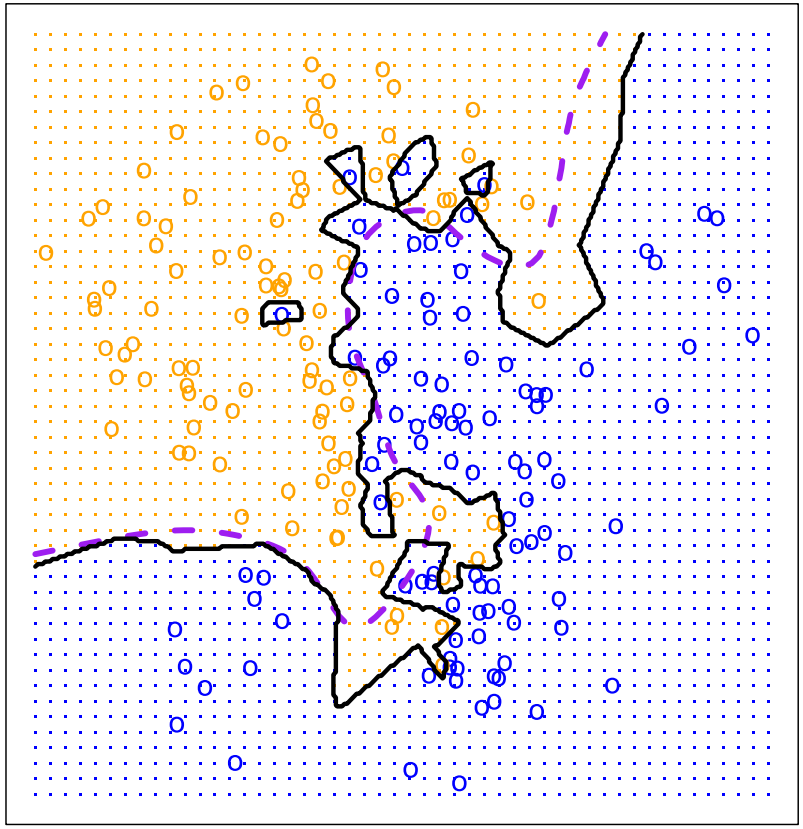
Příklad: K nejbližších sousedů ve dvou dimenzích



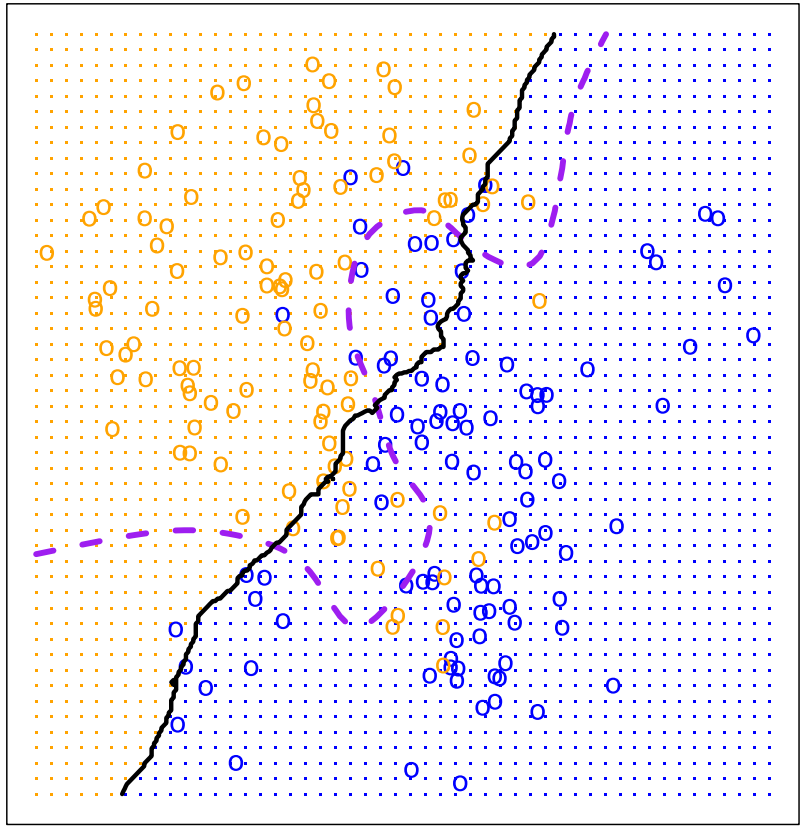
KNN: K=10

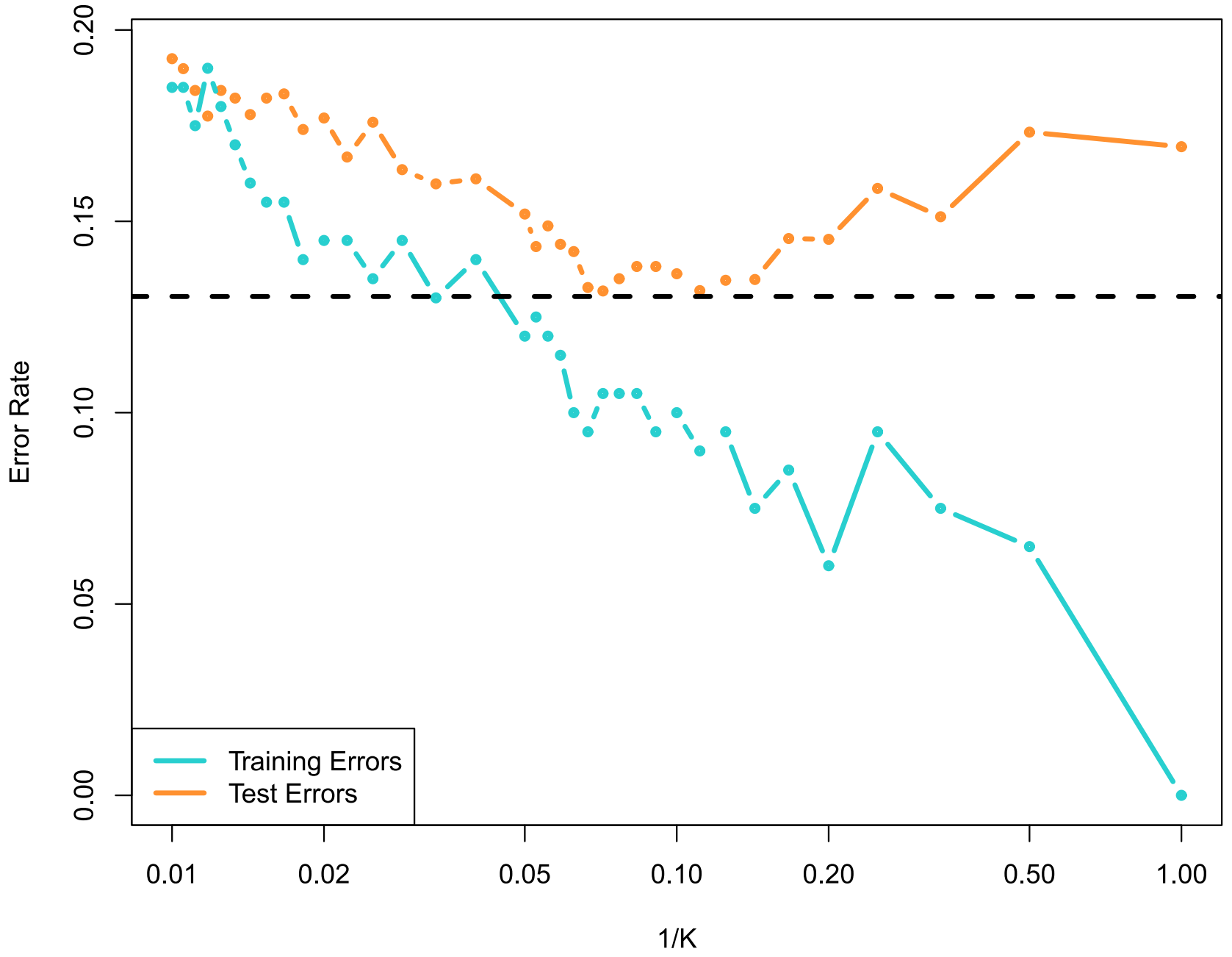


KNS: K=1



KNS: K=100




Error Rate
0.00
0.05
0.10
0.15
0.20
0.01
0.02
0.05
0.10
0.20
0.50
1.00
1/K
Training Errors
Test Errors