Pokročilé statistické metody
8. cvičení

ROC analýza
Regresní modelování

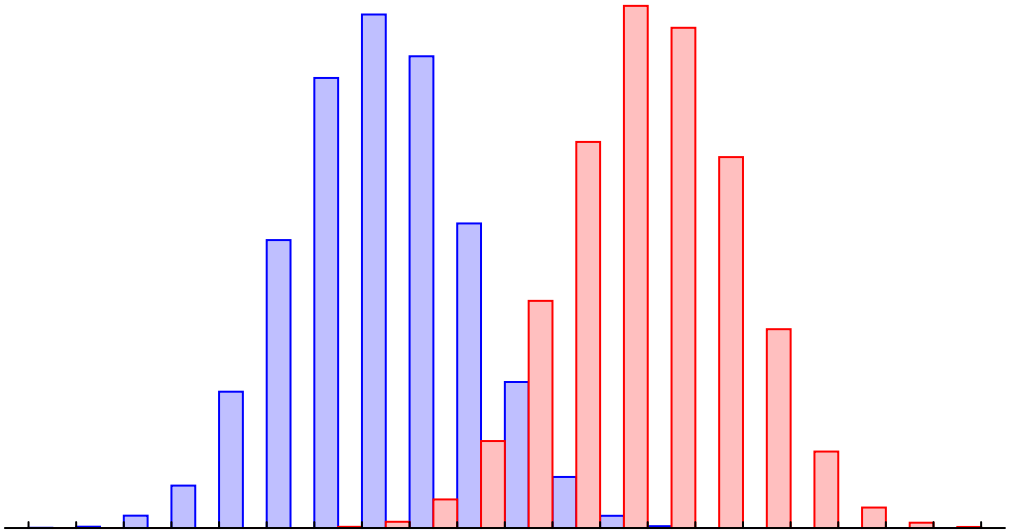
ROC analýza – PROČ?

ROC analýza

- Vyhodnocení prediktivních schopností parametrů
- Identifikace hranice (cut-off) spojitých proměnných, aby při jejich užití v modelech byla maximalizována schopnost klasifikace endpointu na základě nově vytvořené binární proměnné (z původně spojitého parametru).

Odlišení dvou skupin objektů
(modří = zdraví; červení = nemocní)

Kde leží optimální hranice mezi skupinami?



Spojitý parametr, který chceme binarizovat

Sensitivita a specificita - teoreticky

Skutečnost = pacient je zdravý/nemocný

	0 – zdravý	1 – nemocný
0 – neriziková skupina	Skutečně negativní (TN)	Falešně negativní (FN)
1 – riziková skupina	Falešně pozitivní (FP)	Skutečně pozitivní (TP)

Proměnná predikující skutečný stav – např. výsledek laboratorního testu, věk, BMI

$$specificita = \frac{TN}{TN + FP}$$

→ Podíl zdravých jedinců, u kterých vyšel test negativně.

$$senzitivita = \frac{TP}{TP + FP}$$

→ Podíl nemocných jedinců, u kterých vyšel test pozitivně.

Sensitivita a specificita – příklad

- ROC dle každé unikátní hodnoty spojitého parametru vytváří novou binární proměnnou (neriziková vs. riziková skupina = kategorie věku v příkladu).

	Výskyt infarktu (ne)	Výskyt infarktu (ano)
Věk (do 30 let)	20	1
Věk (nad 30 let)	120	69
	$specificita = 14{,}3\ \%$	$senzitivita = 98{,}6\ \%$

	Výskyt infarktu (ne)	Výskyt infarktu (ano)
Věk (do 50 let)	70	5
Věk (nad 50 let)	70	65
	$specificita = 50{,}0\%$	$senzitivita = 92{,}9\ \%$

	Výskyt infarktu (ne)	Výskyt infarktu (ano)
Věk (do 70 let)	110	10
Věk (nad 70 let)	30	60
	$specificita = 78{,}6\%$	$senzitivita = 85{,}7\ \%$

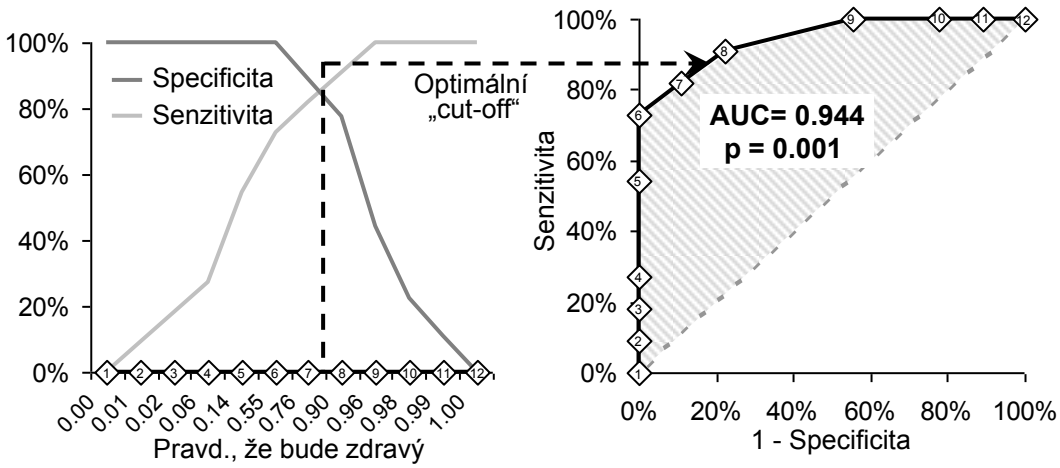
cut-off ⇧

specificita ⇧

senzitivita ⇩

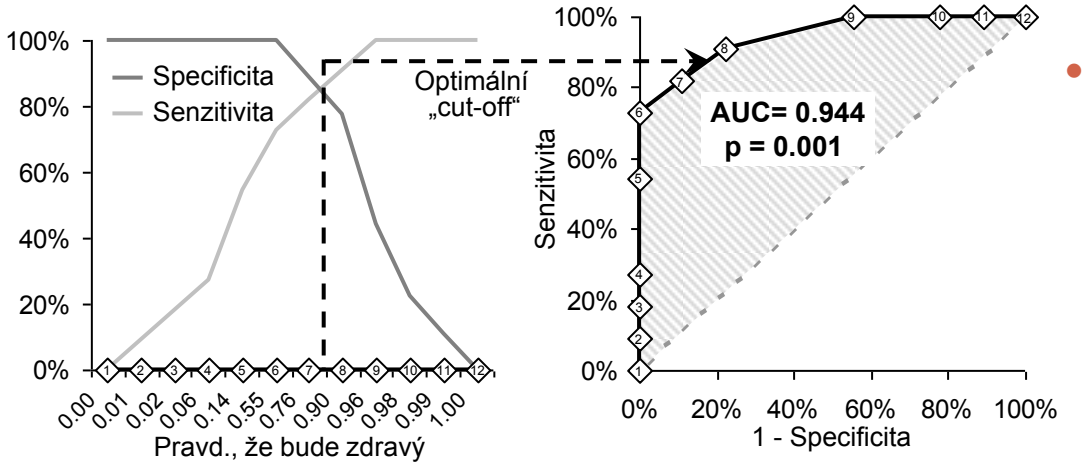
Jaká hranice (cut-off) je nejlepší?

Výběr cut-off I



Výběr cut-off II

- Na základě vyhodnocení hodnot specificit a senzitivit pro každou z unikátních hodnot spojitého parametru vybíráme hranici (cut-off), na základě které rozdělíme spojitý parametr do nové binární proměnné.
- Upřednostnění sensitivity nebo specificity je do určité míry subjektivní dle reálného cíle analýzy:
  - Vysoká senzitivita – screeningový test, kdy je třeba zachytit všechny možné nemocné (např. závažné onemocnění, které je třeba zachytit v počátečním stadiu).
  - Vysoká specificita – pokud je nezbytné odchytit pouze skutečně nemocné pacienty (např. nechceme vystavovat pacienty zbytečné léčbě málo závažného onemocnění).
  - V praxi většinou dobré rozdělení souboru poskytne cut-off, pro který součet specificity a senzitivity dosahuje maximální hodnoty.



- **AUC (plocha pod křivkou) s intervalem spolehlivosti**
  - Čím odlišnější od 0.5, tím lepší predikce

Regresní modelování – PROČ?

Regresní modelování – PROČ?

- Cílem je **vysvětlit variabilitu závislé proměnné** (endpoint, outcome, response, Y) pomocí **prediktorů** (nezávislá, vysvětlující proměnná, kovariáta, X).

- Regresní model kvantifikuje vliv prediktorů a poskytuje regresní rovnici, čímž umožňuje následnou predikci závisle proměnné na nových datech.

Regresní modelování – výběr metody

- Kombinace datového typu predikované proměnné určuje použitou metodu analýzy:

Typ Y	Metoda
spojitá	Lineární regrese
Dvě a více spojitých proměnných	Vícenásobná lineární regrese
Korelovaná data	Smíšené modely
binární	Logistická regrese
ordinální	Ordinální logistická regrese
nominální	Multinomická logistická regrese
Časově závislá proměnná (výskyt události v čase)	Coxův model proporcionálních rizik
Opakované měření v čase	Longitudiální modely

Jednorozměrné vs. vícerozměrné modelování

- Bez ohledu na typ modelu můžeme obecně provádět jednorozměrnou nebo vícerozměrnou analýzu.

- **Jednorozměrné hodnocení (univariate):** do modelu vstupuje vždy jeden prediktor.
  - Hodnotíme vliv jednotlivých prediktorů bez ohledu na ostatní proměnné.

- **Vícerozměrné hodnocení (multivariate):** do modelu vstupuje více proměnných současně.
  - Hodnotíme vliv prediktorů adjustovaný na ostatní proměnné v modelu (unikátní příspěvek proměnné k vysvětlení závislé proměnné).
    - → Umožňuje odstranění vlivu zavádějících faktorů.
    - → Výběr proměnných, které nezávisle na sobě přispívají k vysvětlení závislé proměnné.

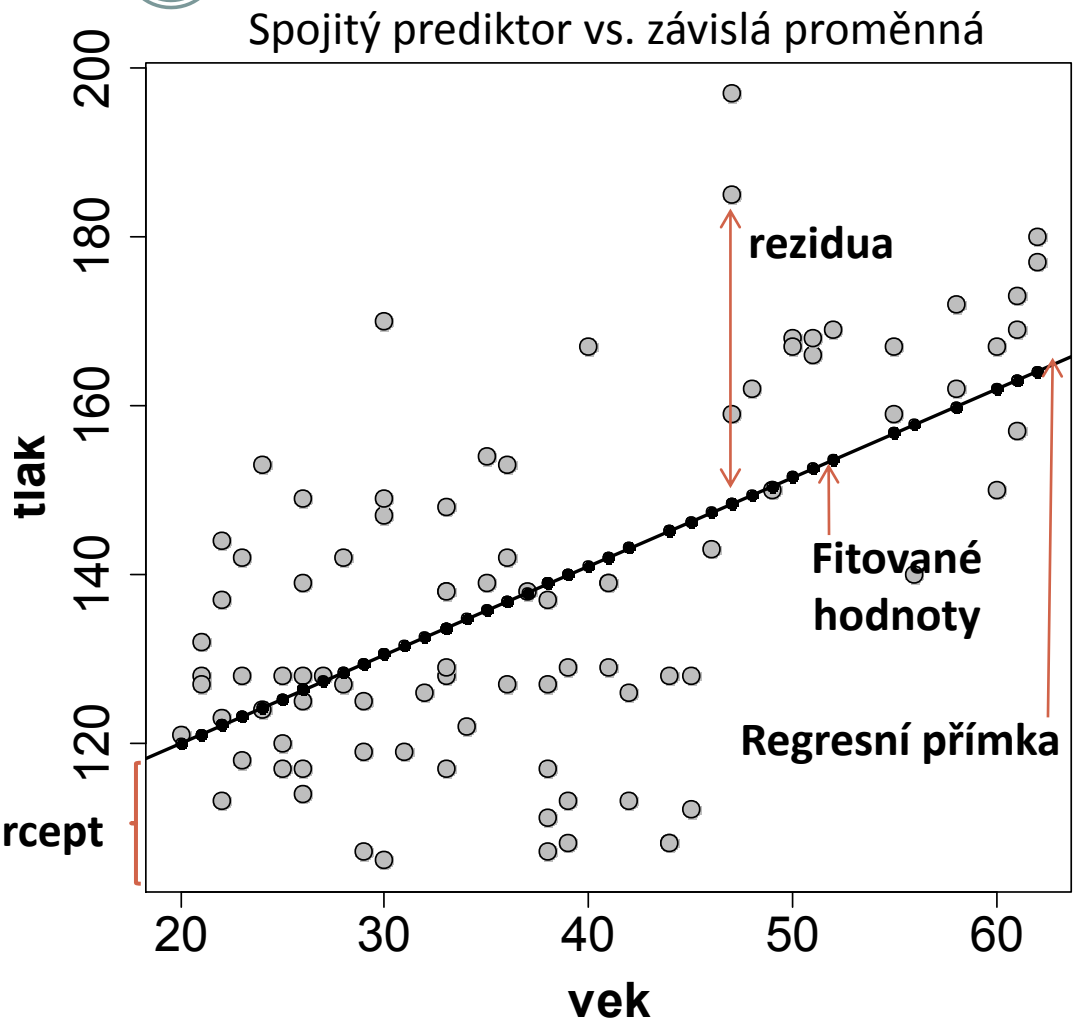
Lineární regrese – model s jednou proměnnou

intercept prediktor rezidua

Závislá proměnná spojitého typu

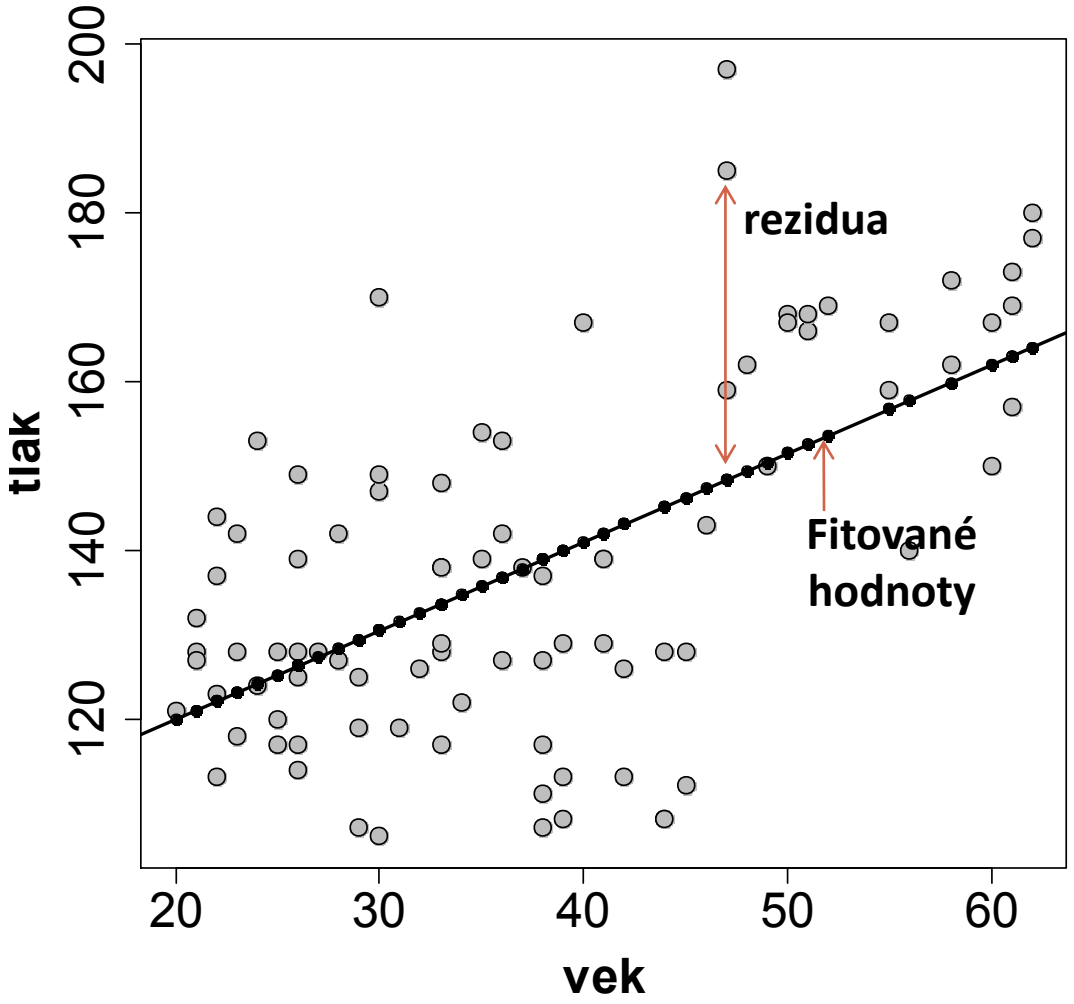
Koeficient pro daný prediktor

Regresní rovnice



Předpoklady

- Zaměřují se na rozložení **reziduí** – rozdíl mezi pozorovanými a odhadnutými (očekávanými) hodnotami závisle proměnné. Variabilita, kterou nevysvětlíme modelem.
- **Předpoklad:** Normální rozdělení reziduí s nulovou střední hodnotou a konstantním rozptylem. Nezávislost jednotlivých pozorování.
- **Multikolinearita** – vysoká korelace parametrů znemožňuje odhad koeficientů.



Výstupy a jejich interpretace I

- **Regresní koeficient** – počet koeficientů odpovídá počtu prediktorů + 1 (intercept). Kvantifikuje, jaká je průměrná změna hodnoty závislé proměnné při změně hodnoty prediktoru.

  - Spojitý prediktor: Jak se změní hodnota závislé proměnné při jednotkovém navýšení nezávislé proměnné.

  - Kategoriální prediktor: Jak se změní hodnota závislé proměnné pro objekty v rizikové kategorii prediktoru ve srovnání s kategorií referenční (v softwaru je potřeba nadefinovat, kterou kategorii bereme jako referenční).

  - Spojitý prediktor ve vícerozměrném modelu: Jak se změní hodnota závislé proměnné při jednotkovém nárůstu prediktoru, zatímco ostatní prediktory zůstávají konstantní. Pokud se výrazně změní hodnota koeficientu po přidání dalšího prediktoru do modelu, lze očekávat korelaci mezi prediktory.

Výstupy a jejich interpretace II

- **Test významnosti jednotlivých parametrů**

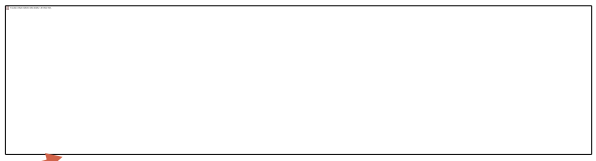
- **Test významnosti modelu** – F test.

- **Koeficient determinace (R^2)** – podíl celkové variability závislé proměnné, kterou vysvětlíme modelem = podíl vyčerpané variability (POZOR - můžeme mít významnou asociaci se závislou proměnnou, ale nízké % popsané variability).

- **AIC = Akaikeho informační kritérium**. Čím je hodnota AIC menší, tím považujeme model za lepší. AIC penalizuje modely s vysokým počtem použitých parametrů a tak zamezuje přeučení statistického modelu.

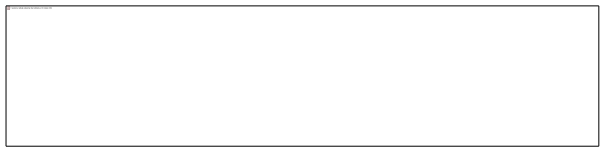
Logistická regrese

- Závislá proměnná binárního typu (bez zahrnutí časové složky).
- Patří mezi zobecněné lineární modely, kde linkovací funkce převádí problém nelineární závislosti y na x na lineární model



Logit linkovací funkce

- Cílem analýzy je:
  - Identifikace vztahů mezi prediktory a endpointem a jejich popis.
  - Vytvoření predikčního modelu umožňujícího zařazení pacientů do hodnocených skupin (obdoba diskriminační analýzy pro 2 skupiny).



Odds ratio (OR)

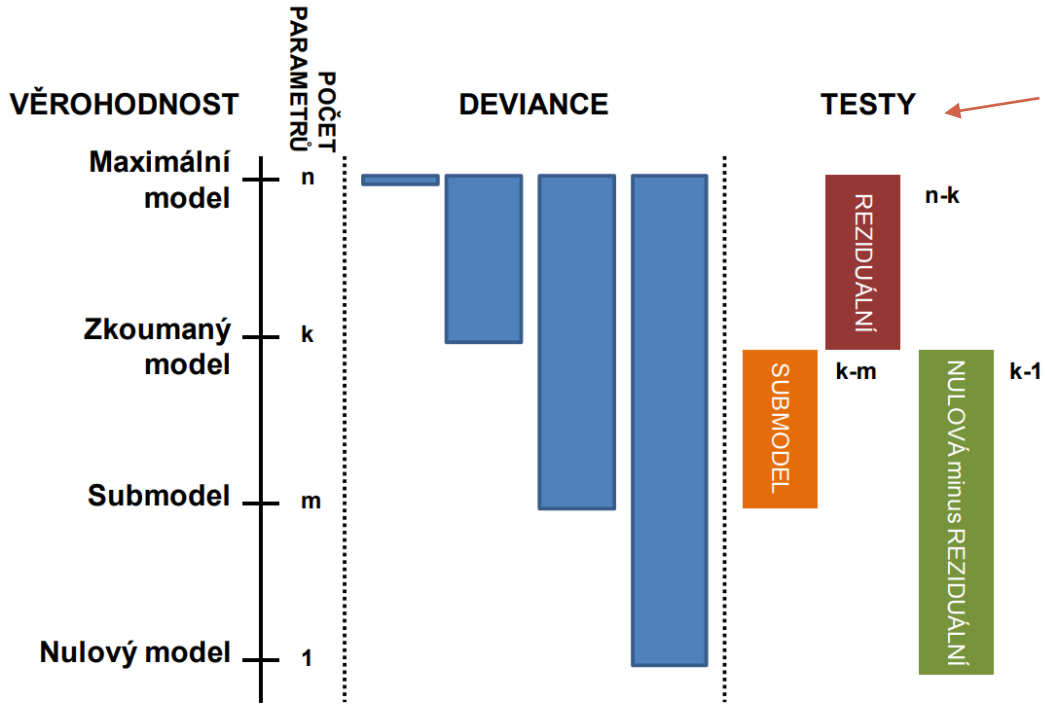
- Koeficient logistické regrese vyjadřuje změnu logaritmu šance výskytu události při jednotkovém nárůstu prediktoru → exponenciální hodnota tohoto koeficientu je interpretována jako poměr šancí = odds ratio (OR), které u:
  - **spojitých proměnných** interpretujeme jako změnu šance na výskyt události při jednotkovém nárůstu prediktoru. Z tohoto důvodu se spojité proměnné často převádí na interpretovatelné jednotky – např. věk po desetiletích, koncentrace po stovkách jednotek).
  - **binárních proměnných** interpretujeme jako změnu šance na výskyt události pro rizikovou kategorii ve srovnání s kategorií referenční.

OR menší než 1: nárůst hodnoty prediktoru značí pokles šance na výskyt události.
OR blízké 1 nárůst hodnoty prediktoru nemění šanci na výskyt události.
OR větší než 1 nárůst hodnoty prediktoru značí nárůst šance na výskyt události.

- **Test významnosti jednotlivých koeficientů**: Waldův test testuje H_0, že koeficient je roven nule (OR = 1) proti H_1, že koeficient je různý od nuly.

Kvalita modelu - vyhodnocení deviancí

- **Vyhodnocení deviancí** = odchylek pozorovaných od predikovaných hodnot

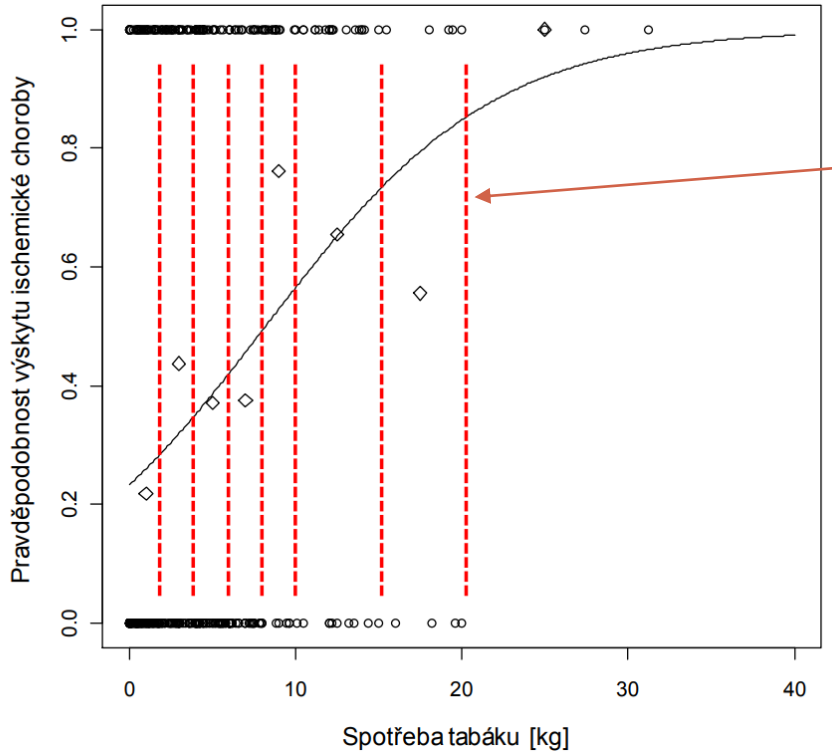


Testujeme, zda se od sebe modely ve svých predikčních schopnostech statisticky významně liší.

http://portal.matematickabiologie.cz/index.php?pg=analyza-a-hodnoceni-biologickych-dat--regresni-modelovani--logisticky-regresni-model-a-jine-zobecnene-linearni-modely--analyza-deviance

Kvalita modelu - Hosmer-Lemeshow test

- test shody pozorovaných a predikovaných frekvencí. Pokud vyjde významně, model fituje špatně.



Pozorování rozdělíme do skupin dle hodnoty predikované pravděpodobnosti výskytu události (např. decily).

Testujeme, zda se liší pozorovaný a očekávaný počet pozorování s událostí v daných kategoriích.

https://is.muni.cz/www/98951/41610771/43823411/43823458/Analyza_a_hodnoc/46097316/textVJ05_logist.pdf