Bi7491 Regresní modelování

Lineární regresní model II

Co byste vědět a umět z minula?

- Vědět, jak se definuje lineární regresní model
- Vysvětlit předpoklady regresních modelů
- Umět použít v lineárním regresním modelu různé typy prediktorů
- Vědět, co je multikolinearita, jak ji zjistit a jak se s ní vypořádat

Co byste měli vědět a umět po dnešní hodině ?

- Umět se vypořádat s chybějícími daty
- Vědět, co je interakce, jak ji poznat, a jak ji zohlednit v konstruovaném modelu
- Znát možnosti kauzálního působení různých faktorů, umět popsat rozdíl mezi zkreslující proměnnou a mediátorem, popisovat jednoduché vztahy pomocí modelových diagramů
- Znát základní pravidla pro zařazování proměnných do modelu
- Umět posoudit splnění modelových předpokladů pomocí grafických nástrojů

Lineární regresní model II

Chybějící data

Chybějící měření prediktorů

- Chybějící měření z různých důvodů je velmi časté
- U víceprediktorové regrese se problém zvýrazňuje
- Snižuje se síla analýzy
- Může dojít ke zkreslení

Co s tím?

- smazat řádky s chybějícími daty
  - ztráta síly testu
  - riziko zavedení zkreslení
    - ne, pokud chybí měření zcela náhodně – vhodné srovnat subjekty
- vytvořit dummy proměnnou pro chybějící údaj
- snažit se získat data
- vypustit proměnnou s chybějícími daty
  - můžeme ztratit klíčové informace
  - ale taky nemusíme - vzpomeňte na kapitolu o multikolinearitě
- odhadnout chybějící hodnoty

Typy chybějících dat

- **Data chybějící zcela náhodně (*Missing completely at random,* MCAR)**
  - Žádný systematický rozdíl mezi chybějícími a pozorovanými hodnotami. Například výpadek pozorování z důvodu poruchy tlakoměru.
- **Data chybějící náhodně (*Missing at random,* MCAR)**
  - Systematický rozdíl mezi chybějícími a pozorovanými hodnotami je vysvětlitelný pozorovanými hodnotami jiné proměnné. Například chybějící hodnoty krevního tlaku budou nižší než pozorované, pokud mladí lidé spíše propásnou měření.
- **Data chybějící nenáhodně (*Missing not at random,* MNAR)**
  - Systematický rozdíl mezi chybějícími a pozorovanými hodnotami není vysvětlitelný ani pozorovanými hodnotami jiné proměnné. Například pokud lidé s vyšším krevním tlakem propásnou návštěvu ambulance z důvodů bolesti hlavy (což nenaměříme).

Odhad chybějících hodnot

- velmi lákavé
  - neztratíme žádné subjekty
  - smysluplné jen pokud u subjektů chybí málo proměnných
- velmi nebezpečné
  - každý odhad je nevyhnutelně špatně

- přiřadit průměr/medián
- totéž po skupinách subjektů
- regrese na ostatních prediktorech – imputace
  - to ale určitě podhodnotí rozptyl proměnné – přidáváme jen očekávané hodnoty
- vícenásobná imputace – složitější metoda, která pomocí simulace nepodhodnotí chyby

Rekapitulace a doporučení

- snažit se dosbírat data
- prohlédnout charakter chybějících dat
- zvážit vyhození proměnných s velkým podílem chybějících dat
- pokud zbývá jen několik subjektů s velkým podílem chybějících dat, zvážit jejich vyloučení
- prohlédnout, zda se subjekty s chybějícími daty liší od ostatních (chybí data náhodně???)
- pokud chybí náhodně, snažit se odhadnout
- pokud ne, máme problém...

Chybějící měření výsledků

- z výše uvedeného má smysl pouze
  - smazat subjekty
  - snažit se získat data

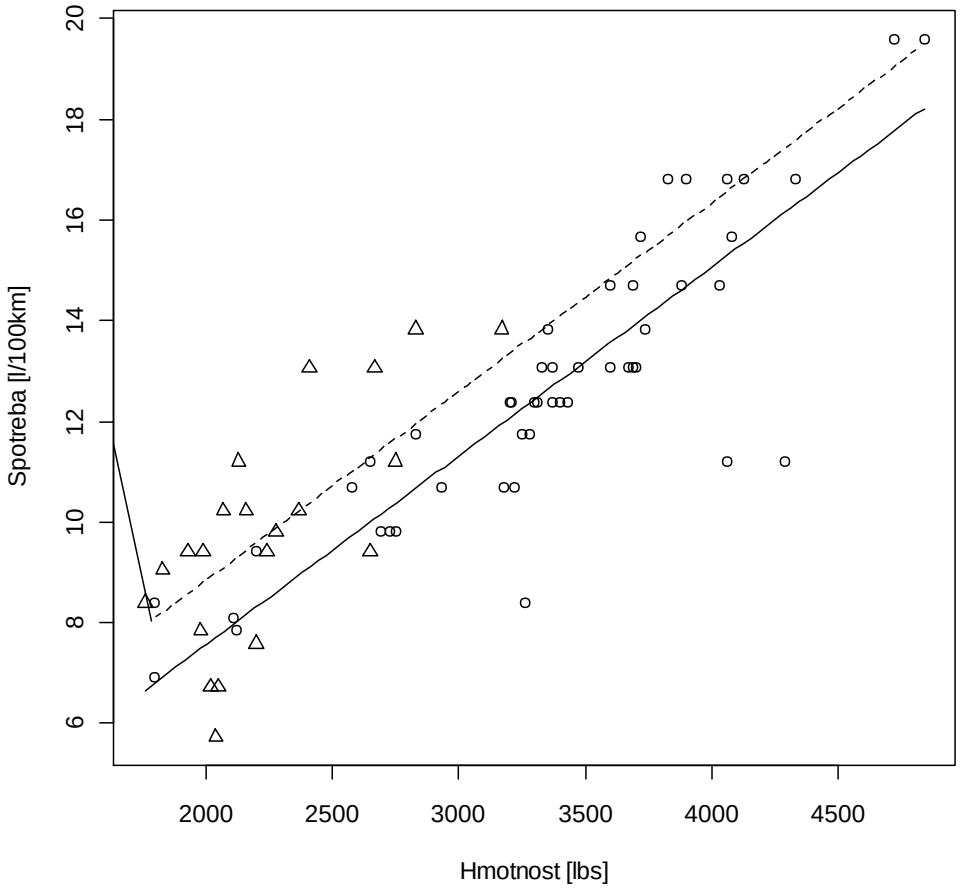
Lineární regresní model II

Interakce proměnných

Aditivita

- Předpokladem regresního modelu je **aditivita**
- to znamená, že účinky prediktorů se nezávisle na sobě sčítají
  - za každou jednotku BMI ubydou dvě jednotky koncentrace vitaminu D
  - každá libra hmotnosti auta přidá 0,004 l na spotřebě
  - americká auta spotřebují o 1,3 litru méně...

Bez interakcí



Americká auta

očekávání

pozorování

Zahraniční auta

očekávání

pozorování

Bez interakcí

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Spotřeba **Hmotnost** **Zahraniční (0/1)**

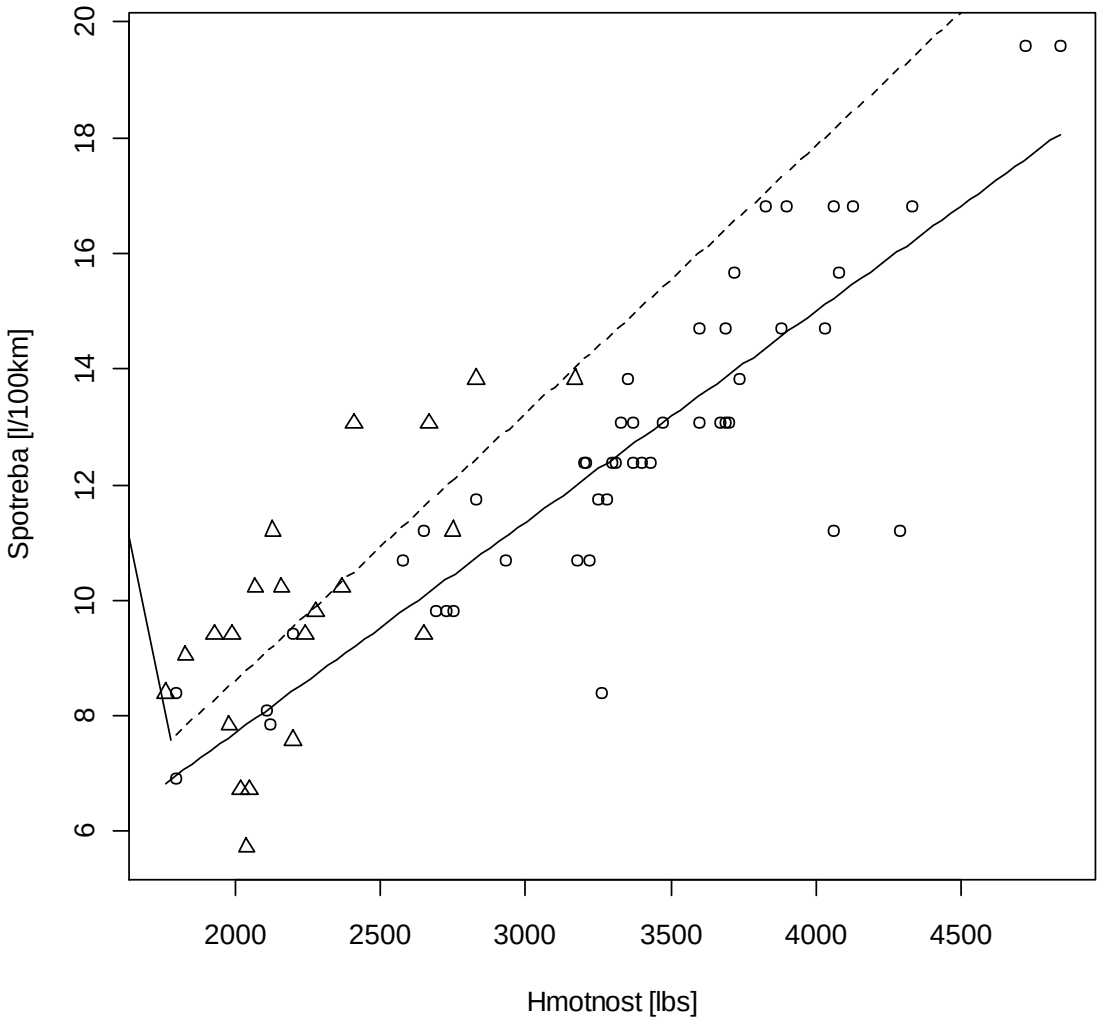
Americká auta $EY_i = \beta_0 + \beta_1 x_{i1}$

na 1 libru hmotnosti poroste o β_1

Zahraniční auta $EY_i = \beta_0 + \beta_1 x_{i1} + \beta_2$

na 1 libru hmotnosti poroste o β_1

S interakcí



Americká auta

očekávání

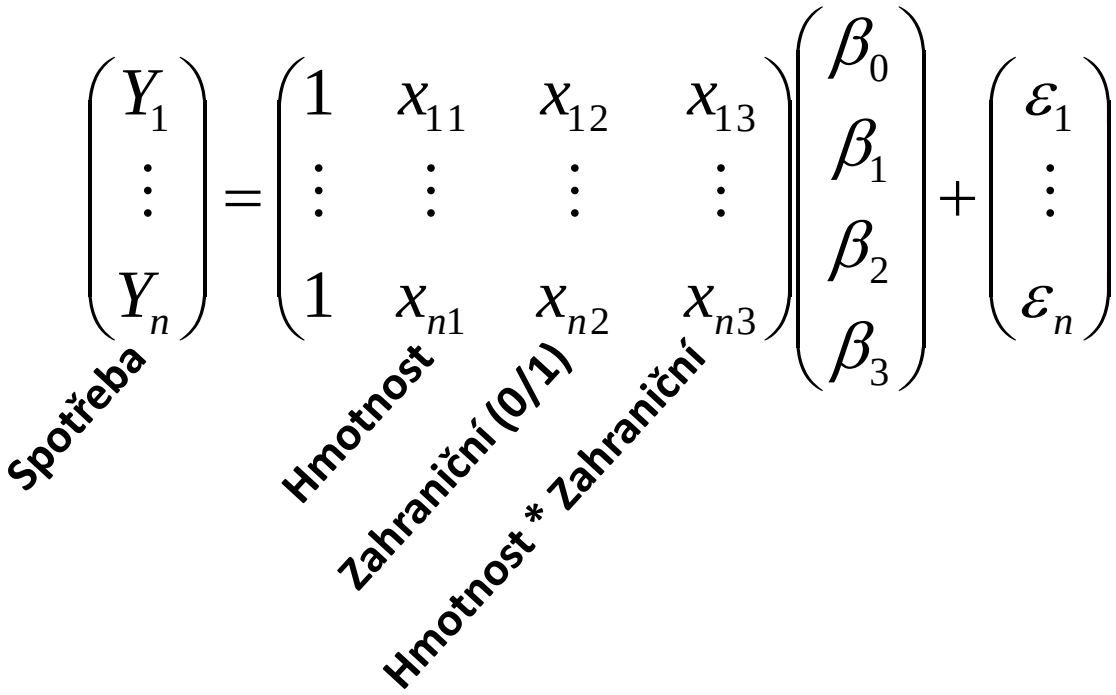
pozorování

Zahraniční auta

očekávání

pozorování

S interakcí



$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

S interakcí

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \left(\begin{array}{cccc} 1 & x_{11} & 0 & 0 \\ \vdots & & \text{americká} & \\ \hline 1 & x_{i1} & 1 & x_{i1} \\ \vdots & & \text{zahraniční} & \end{array}\right) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Spotřeba

Hmotnost

Zahraniční (0/1)

Hmotnost * Zahraniční

S interakcí

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \left(\begin{array}{cccc} 1 & x_{11} & 0 & 0 \\ \vdots & & \text{americká} & \\ \hline 1 & x_{i1} & 1 & x_{i1} \\ \vdots & & \text{zahraniční} & \end{array}\right) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Americká auta $EY_i = \beta_0 + \beta_1 x_{i1}$

na 1 libru hmotnosti poroste o β_1

Zahraniční auta

$$EY_i = \beta_0 + \beta_1 x_{i1} + \beta_2 + \beta_3 x_{i1}$$

$$EY_i = \beta_0 + (\beta_1 + \beta_3) x_{i1} + \beta_2$$

na 1 libru hmotnosti poroste o $\beta_1+\beta_3$

Shrnutí

- Interakce umožňují v prostředí klasického regresního modelu situaci, kdy **vliv** některého prediktoru se mění v závislosti na jiném prediktoru

- spojitá a kategoriální proměnná
  - zahraniční auta mají vyšší spotřebu na jednotku hmotnosti
- 2 kategoriální proměnné
  - mutace genu pro fenylalaninhydroxylasu – OK
  - konzumace mateřského mléka – OK
  - konzumace mateřského mléka postiženým kojencem – POSTIŽENÍ
- 2 spojité proměnné
  - model porodní váhy – závislost na BPD, AC
  - přírůstek váhy [g/cm] pro BPD se liší na každé úrovni AC

Shrnutí

- Interakce umožňují v prostředí klasického regresního modelu situaci, kdy **vliv** některého prediktoru se mění v závislosti na jiném prediktoru

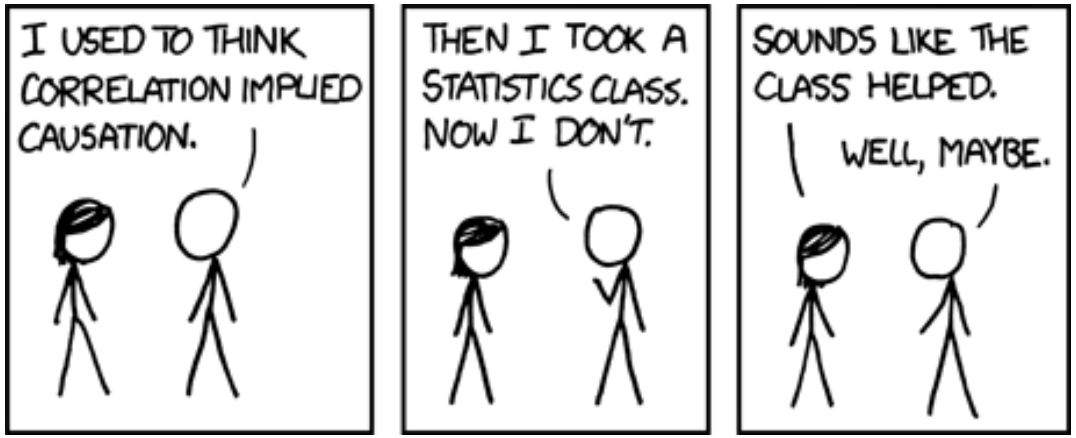
- spojitá a kategoriální proměnná
  - zahraniční auta mají vyšší spotřebu na jednotku hmotnosti
- 2 kategoriální proměnné
  - matematik – nehne s problémem
  - biolog – nehne s problémem
  - matematický biolog – vyřeší problém ☺
- 2 spojité proměnné
  - model porodní váhy – závislost na BPD, AC
  - přírůstek váhy [g/cm] pro BPD se liší na každé úrovni AC

Lineární regresní model II

Kauzalita

(aneb Proč má smysl studovat matematickou biologii?)

Korelace neznamená kauzalitu...



http://xkcd.com/552/

- regrese je statistický nástroj a jako takový zkoumá pouze **asociaci proměnných**
- reálně nás ale zajímá právě ta **kauzalita**
- je nezbytné zapojit své vědomosti a zkušenosti o zkoumaném problému

Není proměnná jako proměnná...

- Závisle proměnná
  - výsledková proměnná (outcome)

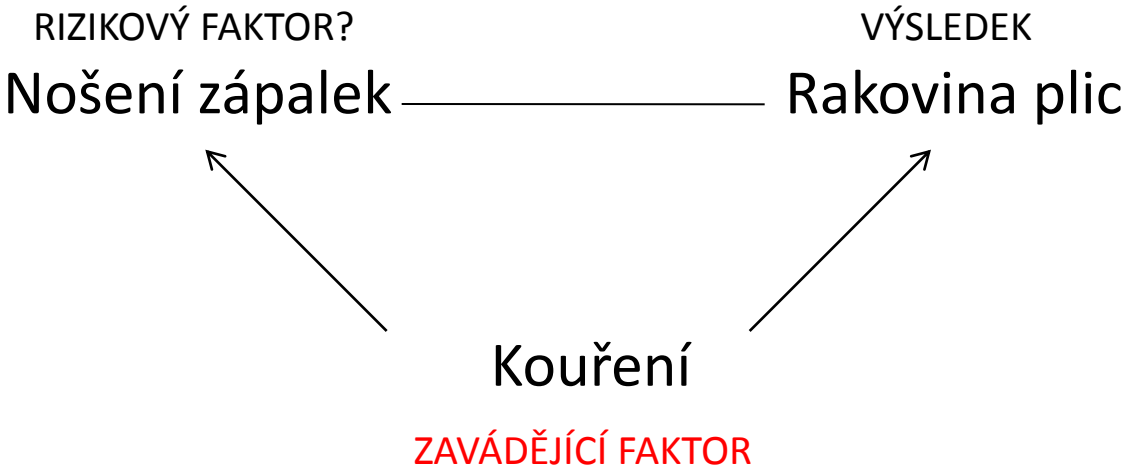
- Nezávisle proměnné (kovariáty)
  - zájmové proměnné
    - ošetření (treatment)
    - expozice (exposure)
  - „rušivé“ proměnné
    - zavádějící faktory (confounder)

Zavádějící faktor (confounder)

- Proměnná asociovaná s rizikovým faktorem a kauzálně spojená s výsledkem

Zavádějící faktor (confounder)

- Proměnná asociovaná s rizikovým faktorem a kauzálně spojená s výsledkem



- může zcela zatemnit skutečný vztah mezi rizikovým faktorem a výsledkem

Supresor

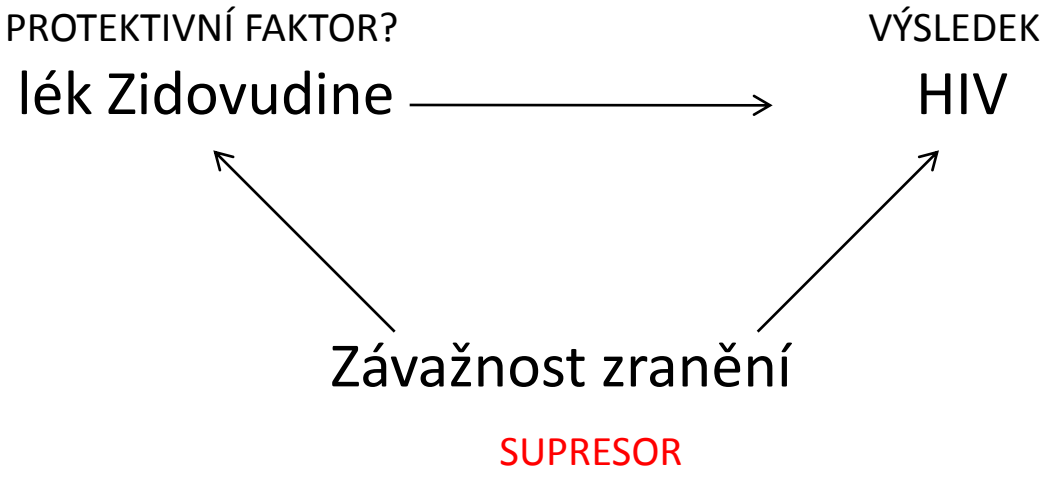
- Zvláštní případ zavádejícího faktoru, který zabrání detekci účinku...
- Uvažujme studii, která zkoumá, zda lék Zidovudine ochrání zraněné zdravotníky před infekcí HIV

PROTEKTIVNÍ FAKTOR?		VÝSLEDEK
lék Zidovudine	⟶	HIV

- Observační studie neukázala významný účinek, ukázalo se však, že lék si brali spíše vážněji zranění zdravotníci

Supresor

- Zvláštní případ zavádejícího faktoru, který zabrání detekci účinku...
- Uvažujme studii, která zkoumá, zda lék Zidovudine ochrání zraněné zdravotníky před infekcí HIV



Supresor

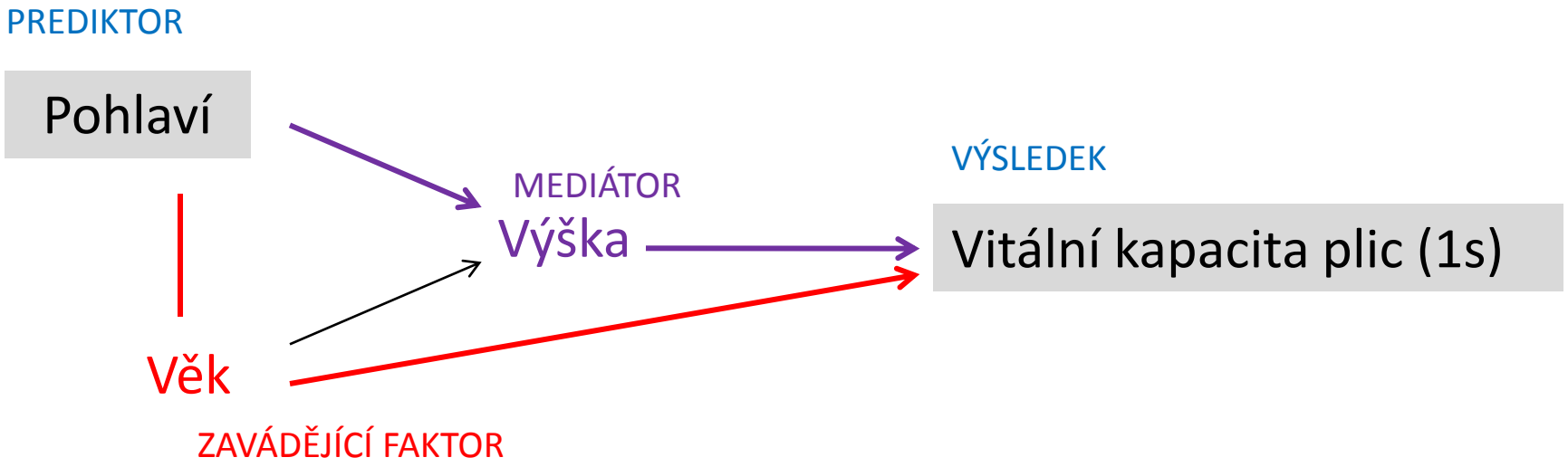
- Zvláštní případ zavádejícího faktoru, který zabrání detekci účinku...
- Uvažujme studii, která zkoumá, zda lék Zidovudine ochrání zraněné zdravotníky před infekcí HIV
- Model doplněný o supresor (závažnost zranění) ukázal významný protektivní efekt Zidovudinu

Mediátor

- stejně jako zkreslující proměnná je asociována s výsledkem
- na rozdíl od ní ale víme, že je kauzálně ovlivněna zájmovým prediktorem (zprostředkovává účinek nějakého prediktoru)
- můžeme ji zařadit do regresního modelu – vysvětluje účinek zájmového prediktoru na výsledek

Modelové diagramy

Příklad



Lineární regresní model II

Které proměnné zařadíme do modelu?

Mimochodem...

Pro tyto hodiny budeme uvažovat:

- Cílem modelování je **pochopení vztahů mezi proměnnými**, ne přesná predikce

- Prediktivní modelování s sebou nese drobné posuny ve filosofii a metodice

Základní pravidla

- pečlivě zformulujeme vědeckou otázku
- studium literatury – prediktory a závisle proměnné
- pečlivé plánování před sběrem dat, aby mohlo zodpovědět danou otázku
- začínáme popisnou analýzou (bivariátní)
- přemýšlení o mechanismu účinku – modelový diagram
- model nesmí obsahovat ani málo, ani moc proměnných
- dostatečná variabilita subjektů ve zkoumaném faktoru (hubení i tlustí)
- spíše nepoužívat automatický výběr prediktorů

Není proměnná jako proměnná...

1. Cílové proměnné studie (léčba, rizikový faktor) – teoretické opodstatnění, neznámý skutečný účinek a jeho forma
2. Proměnné ovlivňující cílovou proměnnou – měly by být v iniciálním modelu a měly by tam zůstat, pokud nenajdeme nějakou velmi korelovanou vysvětlující lépe
3. „Nepostradatelné“ proměnné (pohlaví, věk)
4. Další možné vysvětlující proměnné – *fishing expedition* – jen screening proměnných, nutno ověřit novou studií

Bias x Variance tradeoff

- Pokud do modelu zařadíme příliš mnoho prediktorů – nepřesné výsledky (**VARIANCE**)
- Pokud do modelu zařadíme příliš málo prediktorů, můžeme například opomenout zavádějící faktor – zkreslení (**BIAS**)

- Jednoduché pravidlo říká, že na každou proměnnou zařazenou do modelu, bychom měli mít k dispozici deset pozorování (Events Per Variable) – jednoduše brání *overfittingu*
- V praxi zřejmě může být „beztrestně“ nižší – ale v takovém případě je na místě opatrná interpretace výsledků

Stavba lineárního prediktoru

- Binární proměnné
  - jediná možnost

Stavba lineárního prediktoru

- Kategoriální proměnné (více než dvě hodnoty)
  - dummy proměnné (ale zvážit shluknutí málo zastoupených – jen smysluplně)

Stavba lineárního prediktoru

- Spojité proměnné (záleží na důvod zájmu o prediktor)
  - neměnit tvar užívaný v předchozích studiích, pokud se neví, že to je špatně
  - u těch důležitých většinou známe znaménko a hrubě velikost, ale ne přesný tvar toho vztahu – namalovat graf (scatter, nebo popis v kategoriích) – to je **marginální** vztah, pro **podmíněný** zkoumáme rezidua
  - transformace kovariáty
    - dle znalosti nás něco smyslupného napadne
    - kategorizace – podle počtu pozorování, můžeme nějaké běžně užívané (podváha, norm, nadváha, obezita)
    - teď nás třeba napadne něco lepšího
    - logaritmus, odmocnina, reciproční, exponenciální
    - užitečné jsou vyhlazovací metody – ale pro zájmové proměnné se příliš nehodí, neboť nám neumožní kvantifikovat účinek

Strategie analýzy

- dát do modelu s důležitými proměnnými ty diskutabilní jednu po druhé
  - zjistíme vliv na závisle proměnnou
  - prozkoumáme změny vlivu ošetření, odhalíme zkreslující faktor
  - nechat zájmové proměnné, známé faktory a nalezené významné
- zjednodušovat model???
  - na základě **významnosti**
  - opatrně, možná vůbec ne (to ale zvýší rozptyl odhadů)
- kompromis – vyhodit nevýznamné proměnné neovlivňující ty ostatní (hlavně zájmové)
- **průběžně kontrolovat model – viz dále**

„Významnost“ proměnných

- diskutována v minulé lekci
- **t-test**
  - testování významnosti jediné proměnné (resp. sloupce v matici plánu)
- **F-test**
  - pro jedinou proměnnou totožné výsledky jako t-test
  - možné testovat více proměnných (resp. sloupců v matici plánu)
  - potřeba, pokud jednomu prediktoru odpovídá např. více dummy proměnných

Automatický výběr proměnných

- pokud už použijeme **automatický výběr modelu**, nechat si část dat na ověření – krosvalidace
- HLAVNÍM PROBLÉMEM JE, ŽE NEROZLIŠUJE RŮZNÉ TYPY PROMĚNNÝCH – tj. zájmovou proměnnou (ošetření, expozici), známé zkreslující faktory, potenciální zkreslující faktory, balast...

Lineární regresní model II

Ověření správné volby modelu

Analýza reziduí

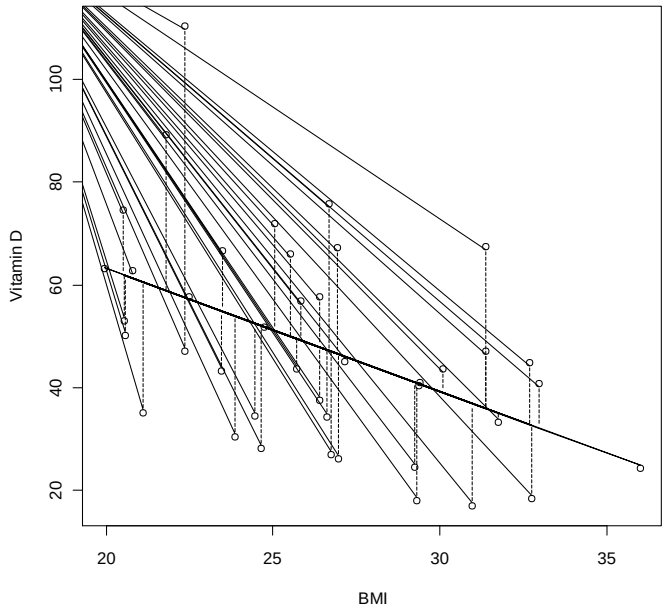
- analýza reziduí je důležitou součástí ověřování vhodnosti modelu. Můžeme tak zjistit, zda výchozí **předpoklad o rozdělení náhodných chyb** a konstrukce lineárního prediktoru byly správné
- pomocí reziduí zjistíme **body, jejichž reziduum je velmi odlišné** od ostatních pozorování. Pokud se v grafu objeví závislost reziduí na prediktorech nebo variabilita reziduí roste v závislosti na veličinách modelu, musíme celý model znovu **přehodnotit**, popř. jej vytvořit od začátku

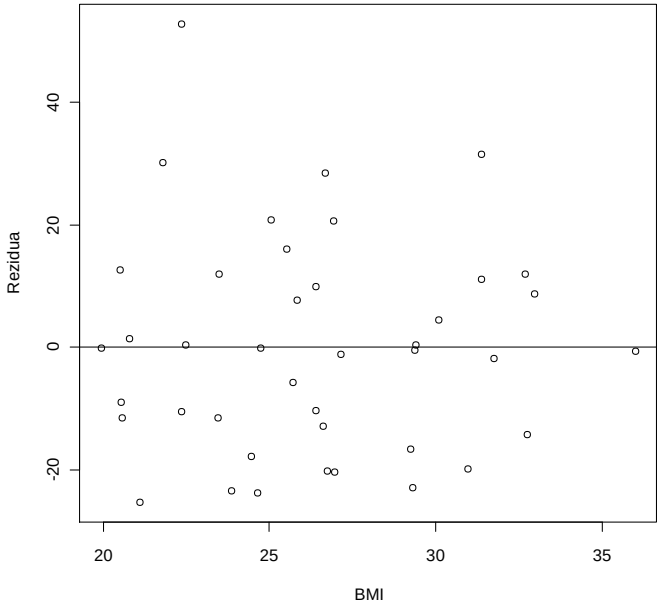
Analýza reziduí

- V lineárním modelu jsou rezidua rozdíly mezi pozorovanými a odhadnutými (očekávanými) hodnotami závisle proměnné:

$$\mathbf{r} = \hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}}$$

- Hodnocení reziduí je nesmírně důležité pro posouzení splnění předpokladů modelu





Předpoklady

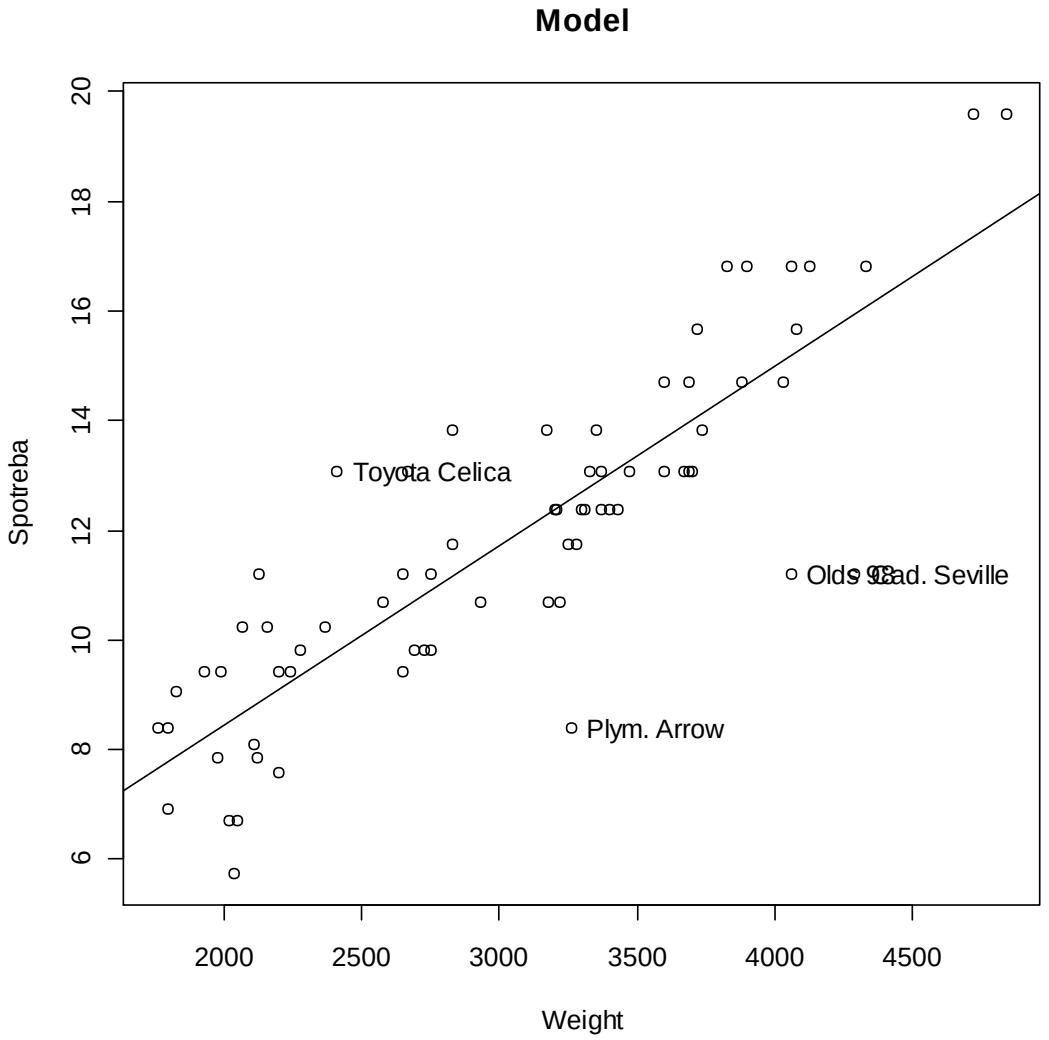
- linearita
- normální rozložení chyb
- homogenní rozptyl

Předpoklady

- linearita
  - graf **rezidua vs. jednotlivé nezávisle proměnné**
  - body musí být symetrické podle nulové hodnoty
- normální rozložení chyb
  - **NP plot reziduí**
  - měla by vycházet přímka
- homogenní rozptyl
  - graf **rezidua vs. jednotlivé nezávisle proměnné**
  - graf **rezidua vs. predikovaný výsledek**
  - rezidua blízko nulové hodnoty
  - rozptýlení hodnot okolo nuly konstantní

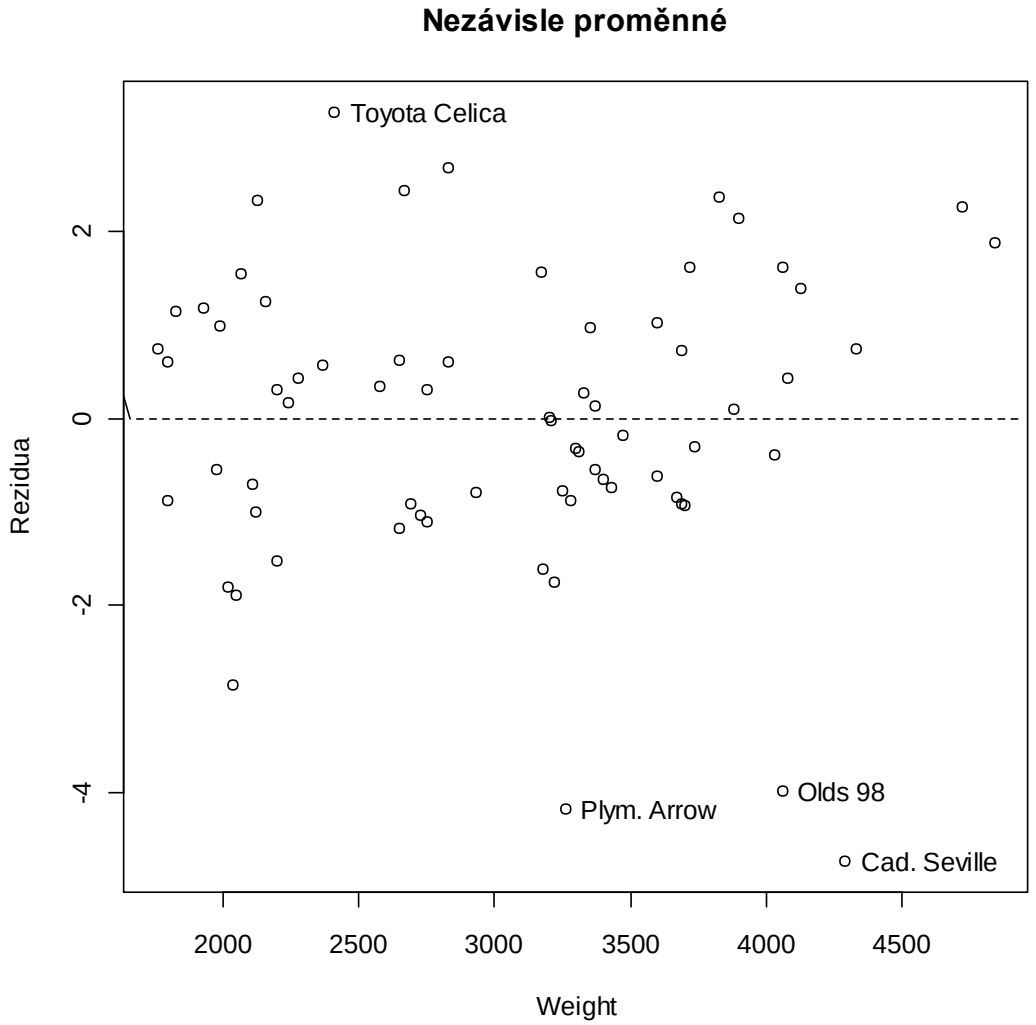
Příklad

spotřeba ~ hmotnost



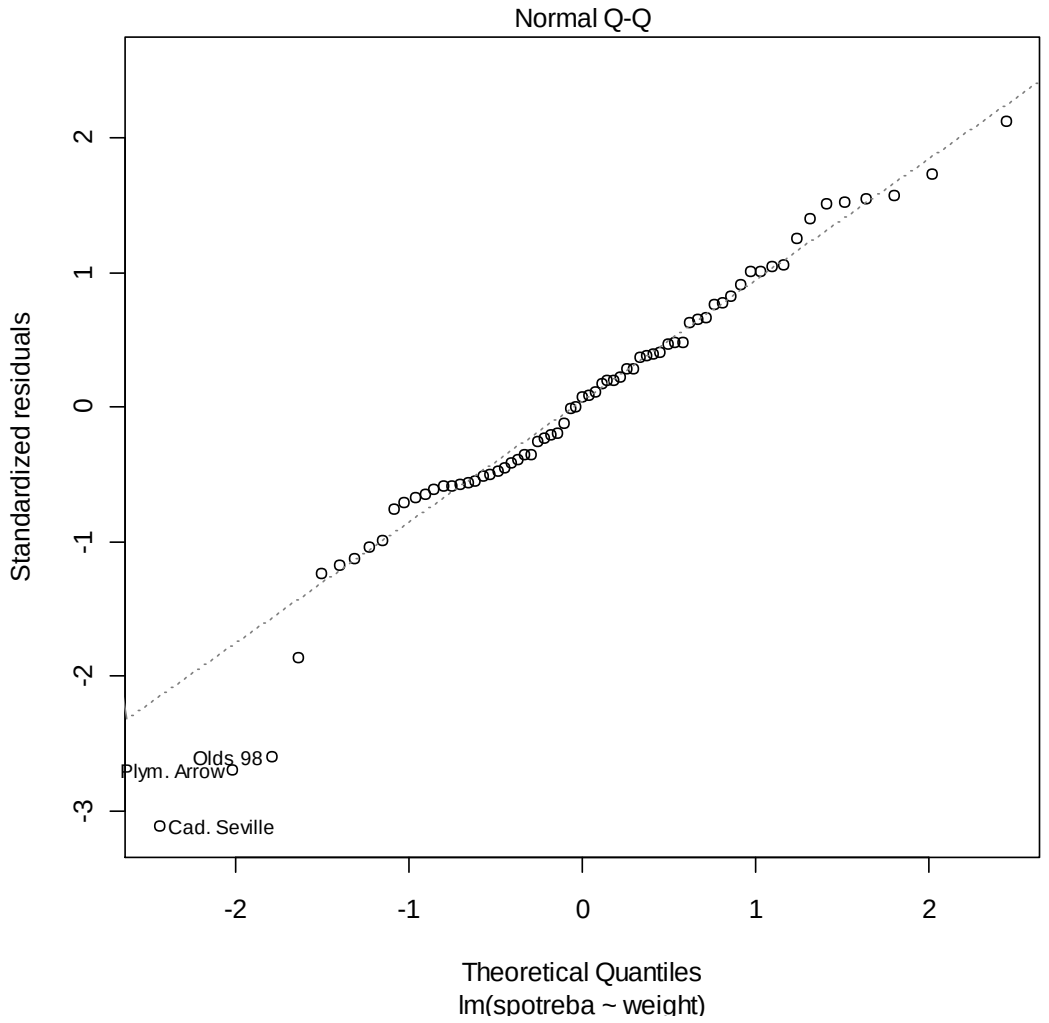
Příklad

spotřeba ~ hmotnost



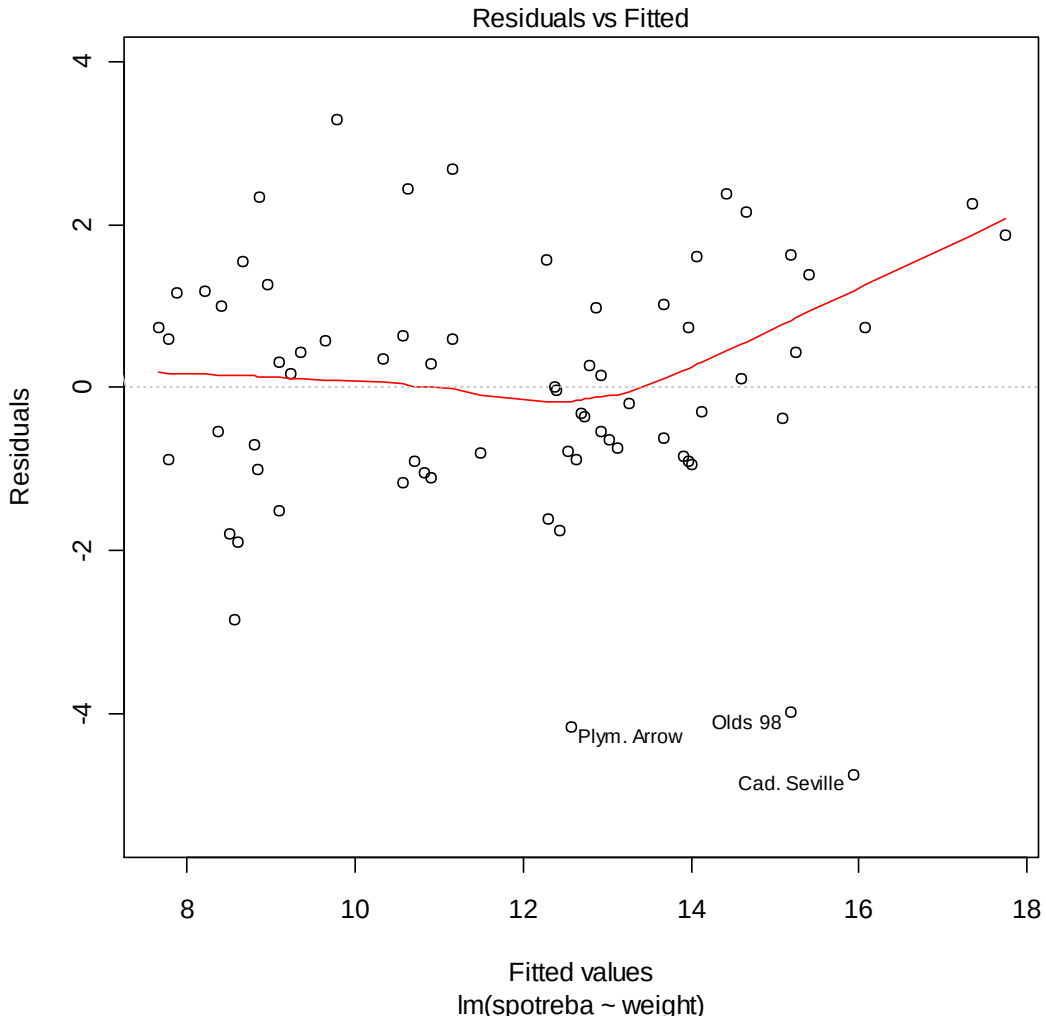
Příklad

spotřeba ~ hmotnost



Příklad

spotřeba ~ hmotnost



Řešení

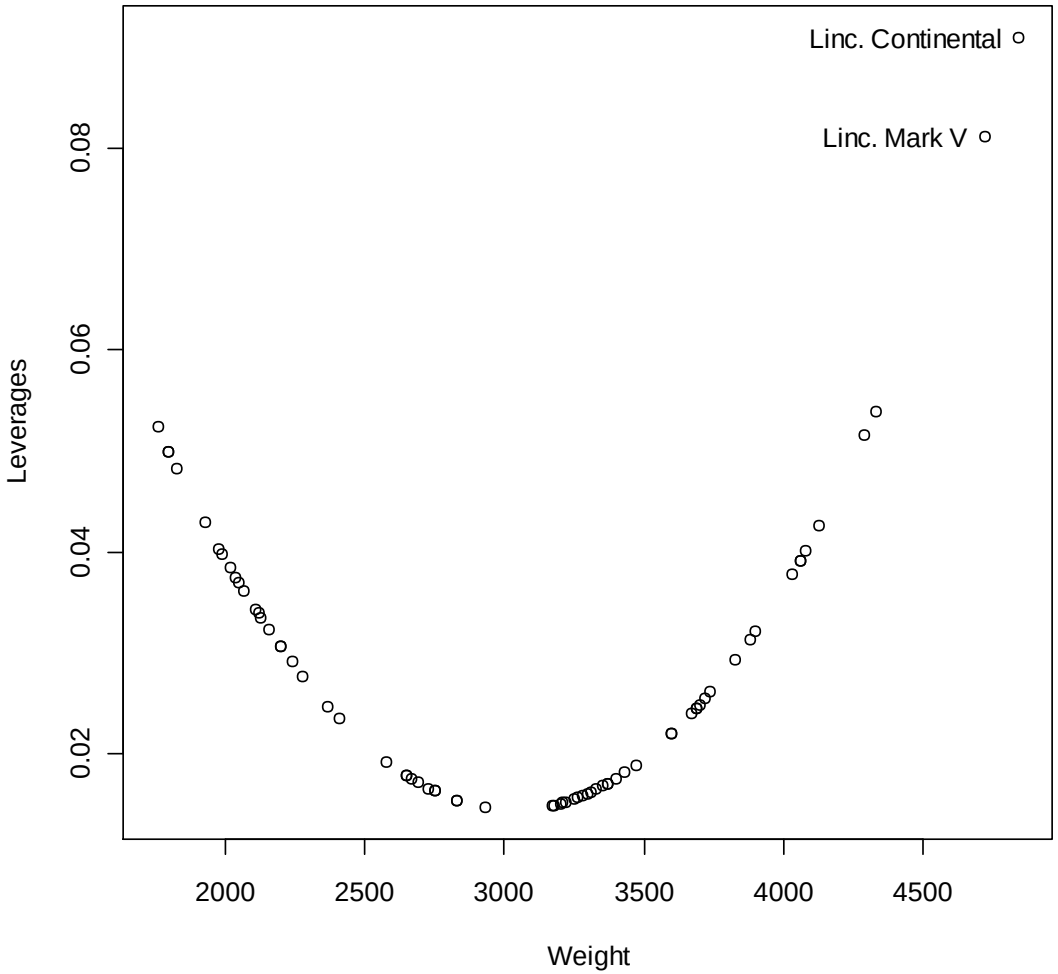
- transformace dat
  - přirozený logaritmus
  - odmocnina
  - převrácená hodnota
  - mocnina
  - arcsin
- prohlídka zvláštních pozorování

Hledání zvláštních bodů

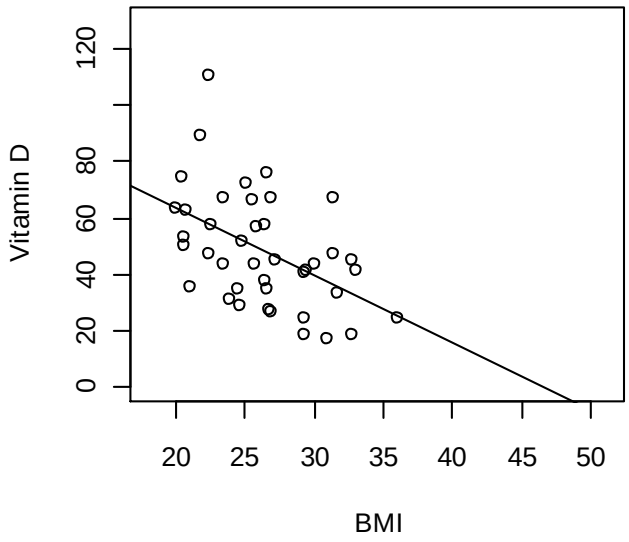
- odlehlé pozorování (outlier)
  - velké reziduum, vzdálené pozorování od očekávané hodnoty
  - extrémní hodnoty **závisle** proměnné
- vlivné pozorování
  - dokáže změnit výsledný model
  - záleží na velikosti vzorku a umístění v prostoru prediktorů
    - veličina LEVERAGE (pákový efekt)
  - extrémní hodnoty **nezávisle** proměnné + atypické hodnoty závisle proměnné

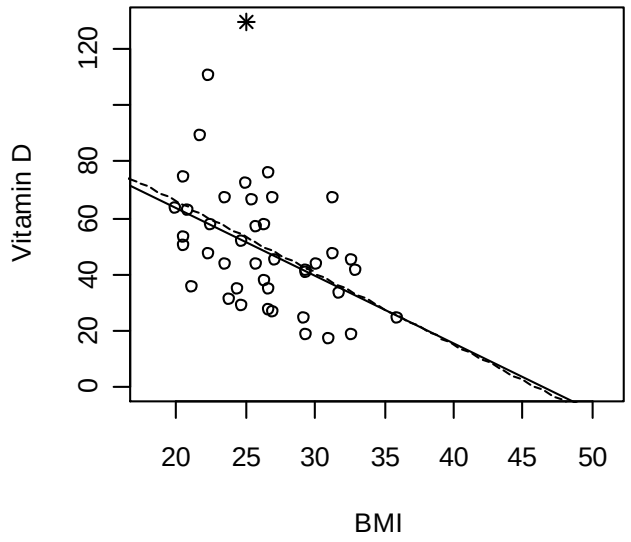
Pákové body

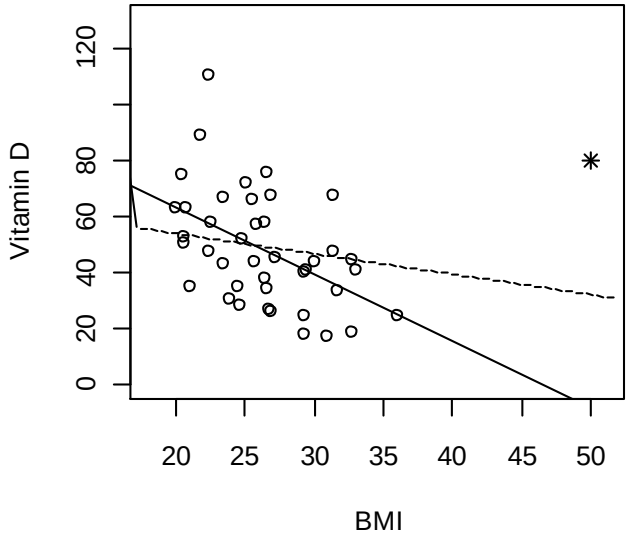
- extrémnost v prostoru prediktorů
- potenciál pro velký vliv na výsledný model

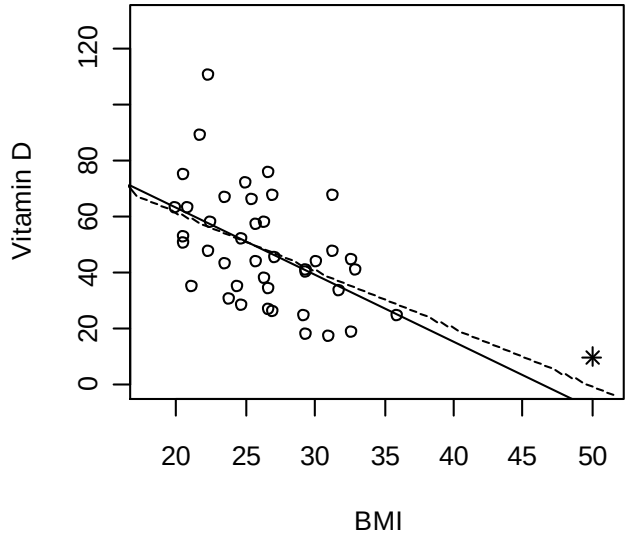


Příklad


Vitamin D
0
20
40
60
80
120
20
25
30
35
40
45
50
BMI


Vitamin D
0
20
40
60
80
120
20
25
30
35
40
45
50
BMI


Vitamin D
0
20
40
60
80
120
20
25
30
35
40
45
50
BMI


Vitamin D
0
20
40
60
80
120
20
25
30
35
40
45
50
BMI

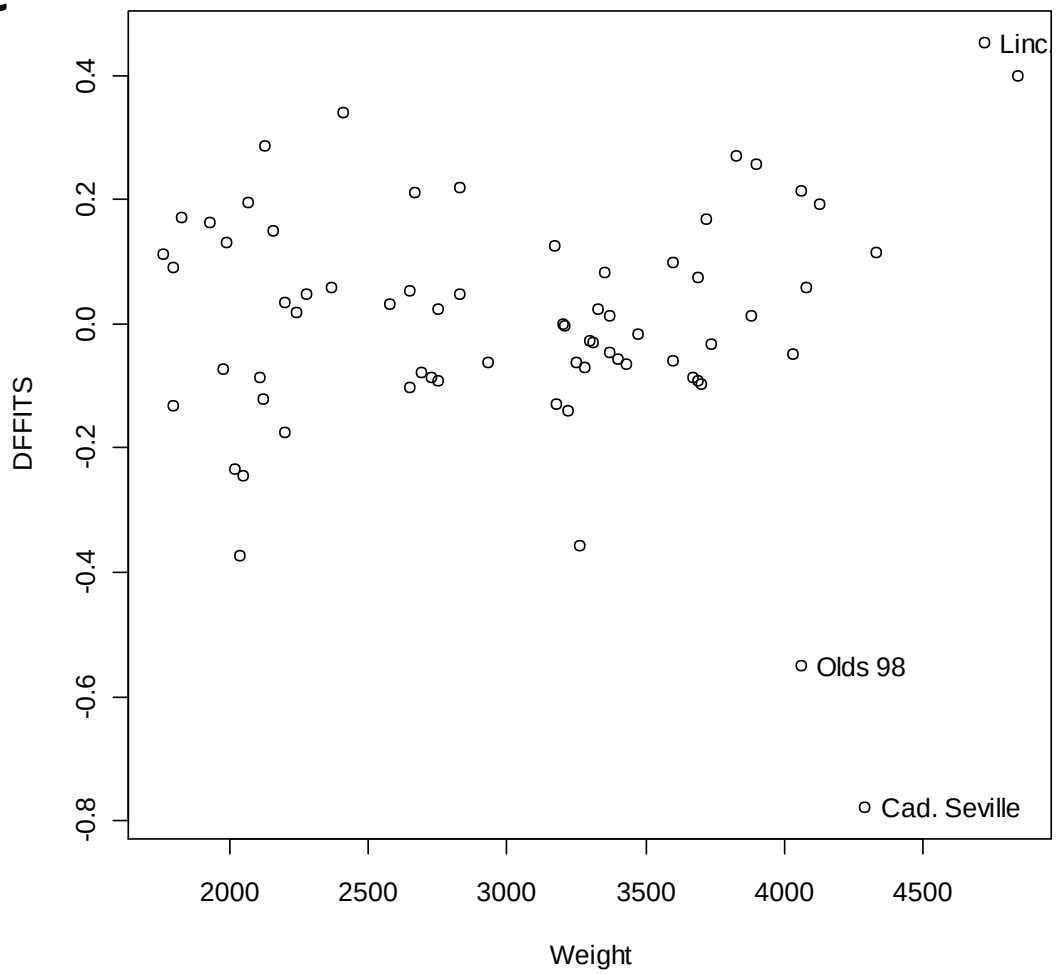
Hledání zvláštních bodů

- hledání **odlehlých pozorování**
  - rezidua vs. jednotlivé nezávisle proměnné
  - NP plot reziduí
  - rezidua vs. predikovaný výsledek

- hledání **vlivných pozorování** – DELEČNÍ DIAGNOSTIKY
  - DFFITS – změna predikovaných hodnot
  - DFBETAS – změna odhadu parametrů
  - Cookova vzdálenost – souhrnná změna odhadu parametrů

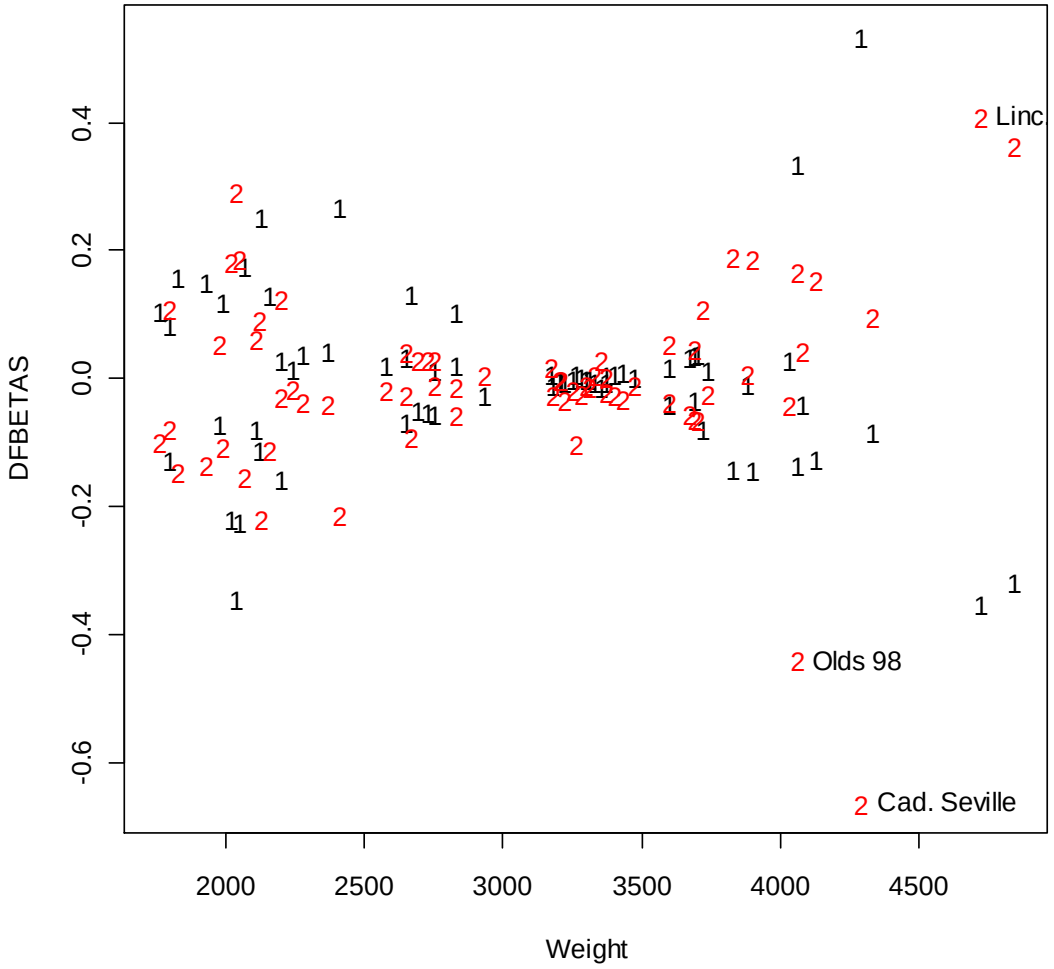
DFFITS

- jak se změní **predikce** pro dané pozorování ve srovnání s jeho nepřítomností



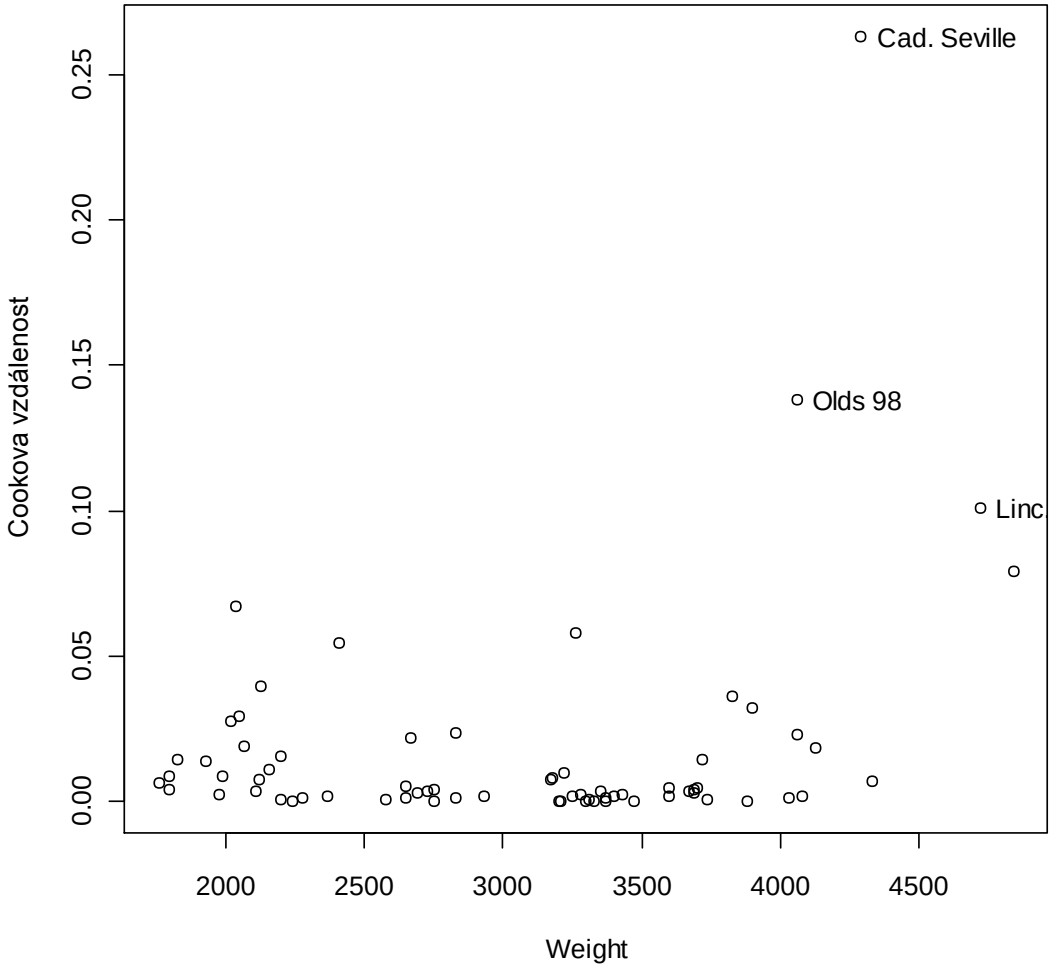
DFBETAS

- jak se změní **odhad parametrů**, ve srovnání s nepřítomností daného pozorování

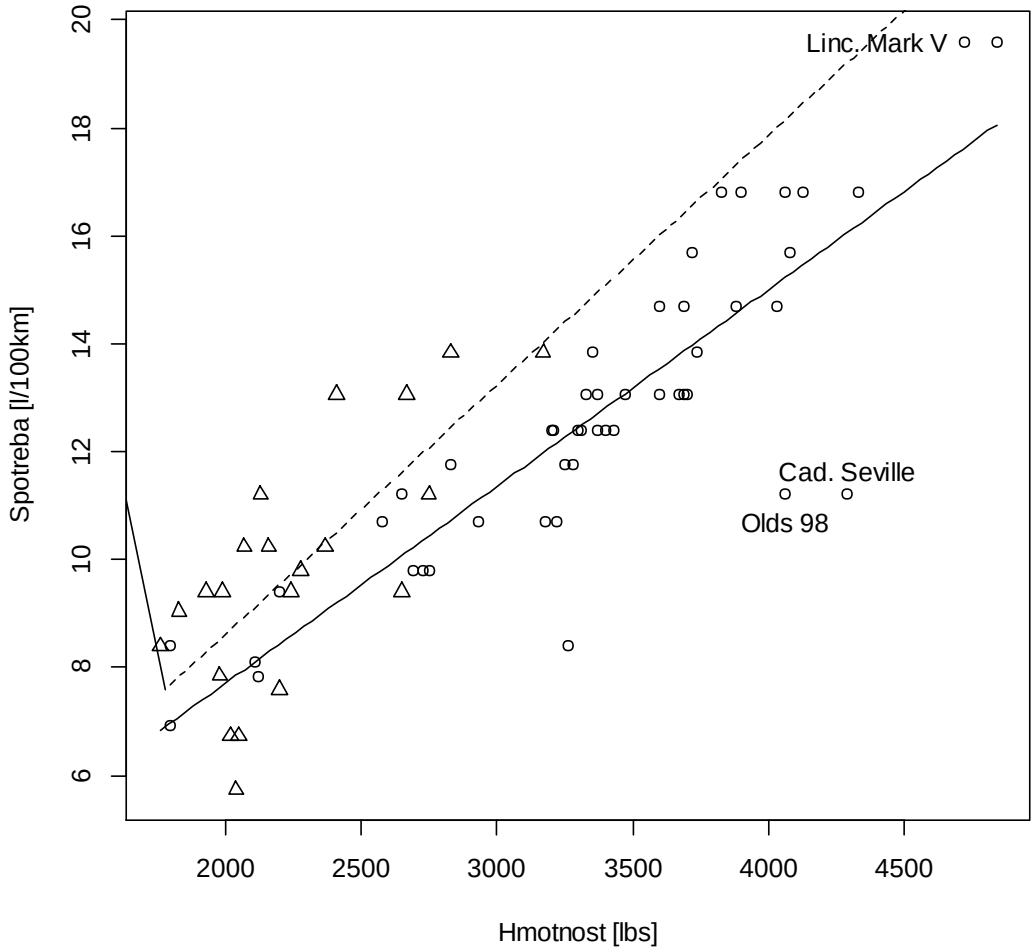


Cookova vzdálenost

- kombinace DFBETAS do jediné hodnoty pro dané pozorování



Vlivná pozorování



Vlivná pozorování

Olds 98 - těžké auto, ale s nízkou spotřebou

Lincoln - těžké auto
s VYSOKOU spotřebou

Cadillac - těžké auto, ale s nízkou spotřebou

Co s nimi?

- podrobněji prozkoumat – třeba je tam důvod
- chybné záznamy – vyhodit z analýzy
- extrémní hodnota prediktoru
  - zformulovat vylučovací kritérium a odstranit rovněž další vyhovující pozorování, obdobně je ale třeba upravit interpretaci výsledného modelu (to je ale lépe dělat předem)
- extrémní hodnota výsledku
  - podívat se, čím jsou pozorování významná, opět se pokusit zformulovat univerzální vylučovací kritérium
- třeba je možné přidat další vysvětlující kovariátu a obohatit tak celkový model

Lineární regresní model II

Závěr

Co byste měli vědět a umět po dnešní hodině ?

- Umět se vypořádat s chybějícími daty
- Vědět, co je interakce, jak ji poznat, a jak ji zohlednit v konstruovaném modelu
- Znát možnosti kauzálního působení různých faktorů, umět popsat rozdíl mezi zkreslující proměnnou a mediátorem, popisovat jednoduché vztahy pomocí modelových diagramů
- Znát základní pravidla pro zařazování proměnných do modelu
- Umět posoudit splnění modelových předpokladů pomocí grafických nástrojů