UNICORN UNIVERSITY

Regresní analýza II

4. seminář k předmětu Statistické metody v analýze dat

19.10.2022

Martina Šimková

simkova.martinka@gmail.com

ZS 2022/2023

Shrnutí – možnosti detekce vlivných pozorování

- Grafická analýza – XY bodový graf
- Lineární kombinace jednotlivých proměnných
  - Centrování, normování proměnných apod.
  - **Normované proměnné**: hodnota cca vyšší než -2 / 2
- Vzdálenosti objektů
  - Euklidovská, Normovaná, Mahalanobisova vzdálenost
  - **Mahalanobisova vzdálenost**: lze vypočítat přesné testové kritérium, přibližně jde o hodnotu vyšší než 12
- Analýza reziduí
  - **Jacknife rezidua**: hodnota cca nižší než -3 a vyšší než 3
- Matice H a její diagonální prvky = leverages: $h_{ii} > \frac{2p}{n}$
- Další koeficienty – např. Cookovo D, DFBETA, DFFIT, …

Nesplněné předpoklady KLRM

- Předpoklady KLRM:
  - Rezidua mají normální rozdělení
  - Nulová střední hodnota reziduí
  - Rezidua jsou nekorelovaná
  - Konstantní rozptyl reziduí

- Normalita
- Multikolinearita
- Heteroskedasticita
- Autokorelace

Normalita náhodné složky

- indikuje nevhodný regresní model

H_0: normalita náhodné složky
H_1: nenormalita

- Identifikace a testy normality:
  - např. Shapiro-Wilkův test

- Příčiny nevhodného modelu:
  - přítomnost vlivných pozorování, zvolená špatná regresní funkce, multikolinearita, heteroskedasticita
  - řešení → odstranění odlehlých pozorování, jiná regresní funkce, výběr jiných proměnných
    - je možná také transformace proměnných (např. logaritmizace, transformace Boxe a Coxe atd.)

Multikolinearita

= **závislost v matici X** (vysvětlujících proměnných X mezi sebou)

- zvyšuje rozptyly odhadů (výsledky t-testů mohou mylně ukazovat na nevýznamnost proměnných – včetně absolutního členu)
- věcná interpretace regresních parametrů nemusí dávat smysl
- velká citlivost na změny odhadů při malé změně dat (nerobustnost)
- nadhodnocení regresního součtu čtverců (některé proměnné zdají důležitější než opravdu jsou)

- **Multikolinearitu musíme vždy řešit. Pokud bychom ji neřešili, nesprávně sestavíme model.**

- Správnou identifikaci mohou zkomplikovat vlivná pozorování.

Detekce multikolinearity

- Postupy:
  - prozkoumat párové korelace mezi proměnnými X
  - vypočítat determinant korelační matrice R(X) → nízké hodnoty ukazují na velmi silnou závislost
  - prozkoumat vlastní čísla matice R(X) → malé hodnoty alespoň jednoho vlastního čísla (nenulové číslo až na 3. desetinném místě) identifikují silnou závislost
  - vysoké hodnoty VIF (variance inflation factor) = diagonální prvky matice $\mathbf{D = R(X)^{-1}}$ → cca VIF > 5 je indikována multikolinearita
  - index podmíněnosti matice R(X) – odmocnina z podílu max a min vlastního čísla > 30 indikuje multikolinearitu
  - vícenásobné korelační koeficienty mezi proměnnými v matici X
  - Farrar-Glauber test, založený na statistice X^2

PŘÍKLAD 1

Data:

usacities.csv

Zadání: V datovém souboru jsou některé zajímavé ukazatele za několik amerických měst. Modelujte standardním způsobem závislost úmrtnosti na všech ostatních šesti ukazatelích.

Ověřte pomocí několika kritérií, zda se v modelu vyskytuje multikolinearita vysvětlujících proměnných.

Možnosti odstranění multikolinearity

1. Jiný typ odhadu modelu, např. **HŘEBENOVÁ REGRESE**

- neposkytuje nezkreslené odhady parametrů beta, avšak pomáhá překonat problém multikolinearity pomocí libovolné konstanty k:

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X} + \mathrm{k}\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

kde

....*k* je kladná konstanta (menší než 1, obvykle se udává menší než 0,3)

.... I je jednotková matice

- Kritika na volbu k – subjektivní nebo pomocí nějakého kritéria → zkreslení

2. Vypuštění zbytečných vysvětlujících proměnných z regresního modelu

PŘÍKLAD 2

Data:

Cars.csv

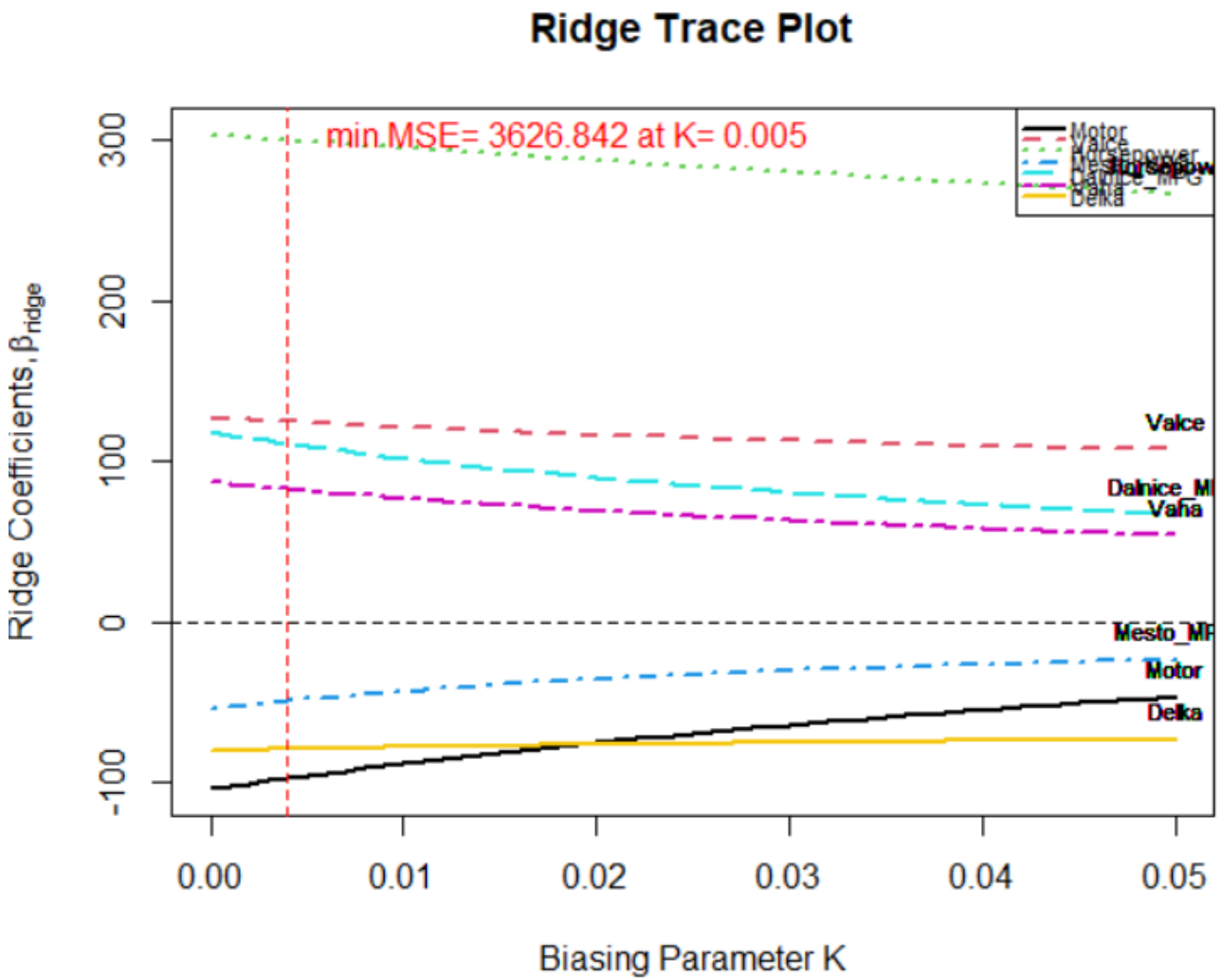
Zadání: Minule jsme si ukázali vícenásobný regresní model pro závislost ceny auta (*Cena_prodej*) na proměnných: objem motoru (*Motor*), počet válců (*Valce*), výkon (*Horsepower*), spotřeba ve městě a na dálnici (*Mesto_MPG, Dalnice_MPG*), váha (*Vaha*) a délka (*Delka*). Nyní pomocí známých kritérií ověřte, zda se v tomto modelu vyskytuje multikolinearita vysvětlujících proměnných.

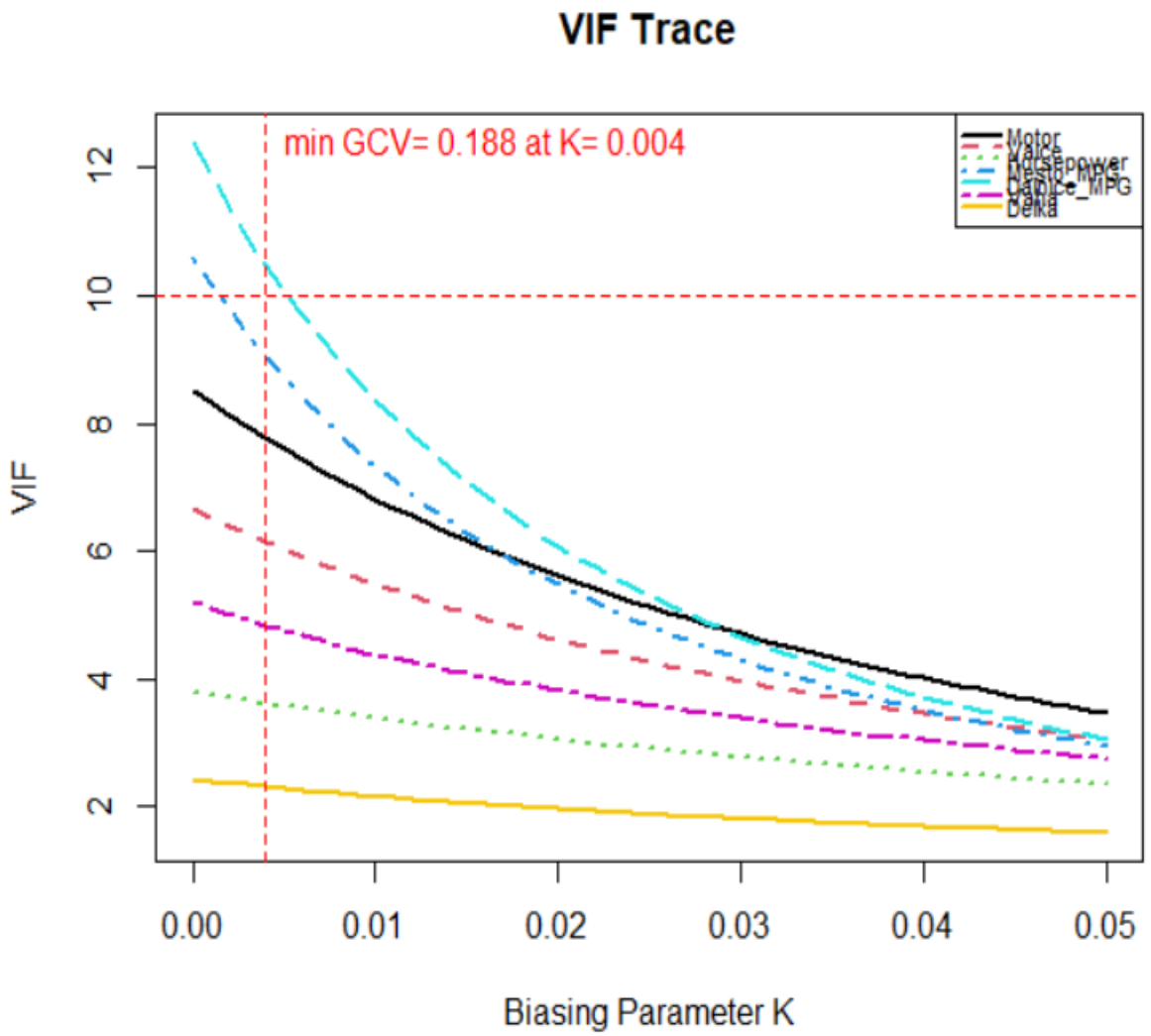
V případě výskytu použijte na tento model jako řešení multikolinearity hřebenovou regresi.

Ověřte kvalitu této regrese rozdělením dat na testovací (400 pozorování) a predikční množinu (zbytek).

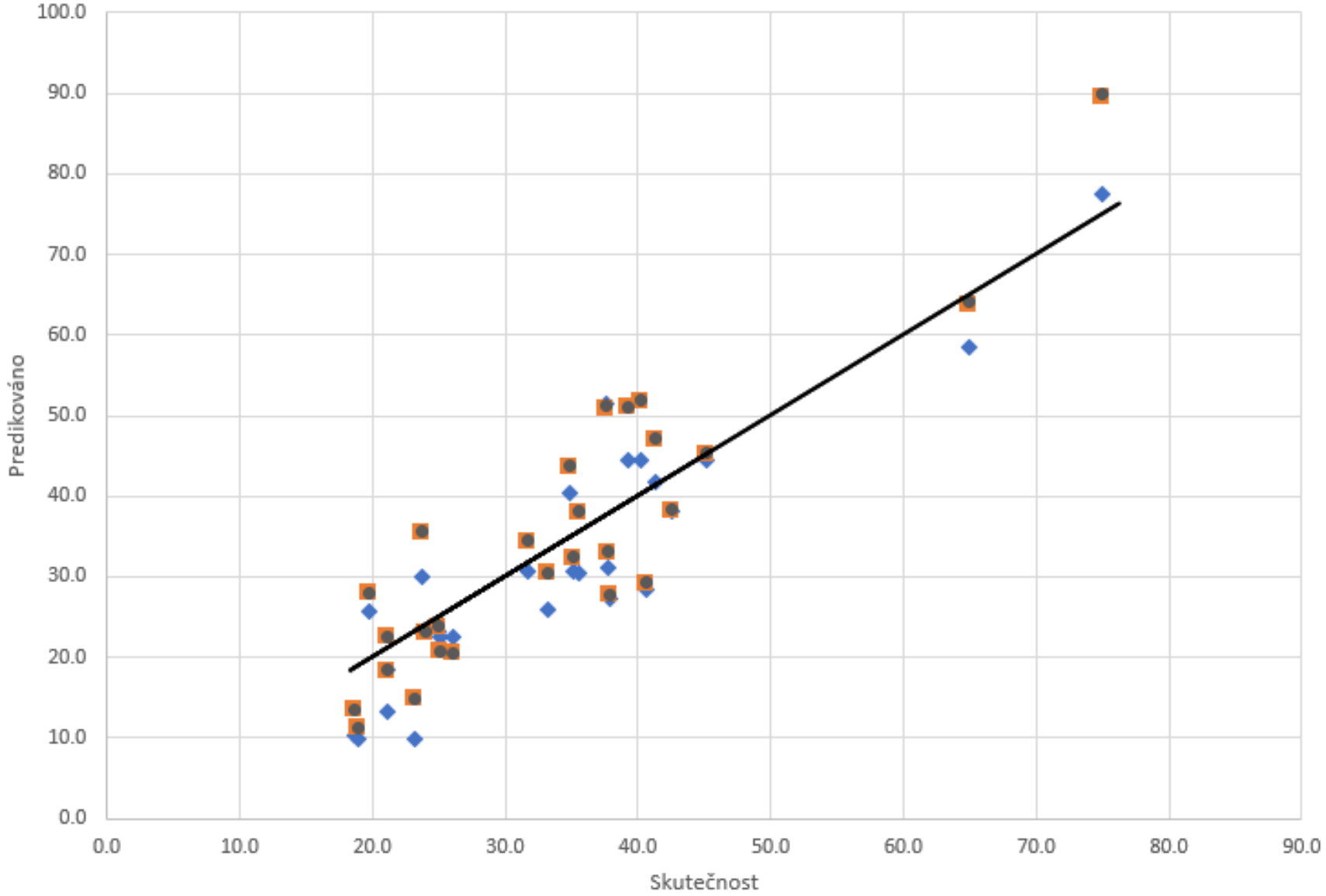
Pokud nebude hřebenová regrese efektivní, zamyslete se nad vyloučením některých vysvětlujících proměnných z modelu.

PŘÍKLAD 2





PŘÍKLAD 2



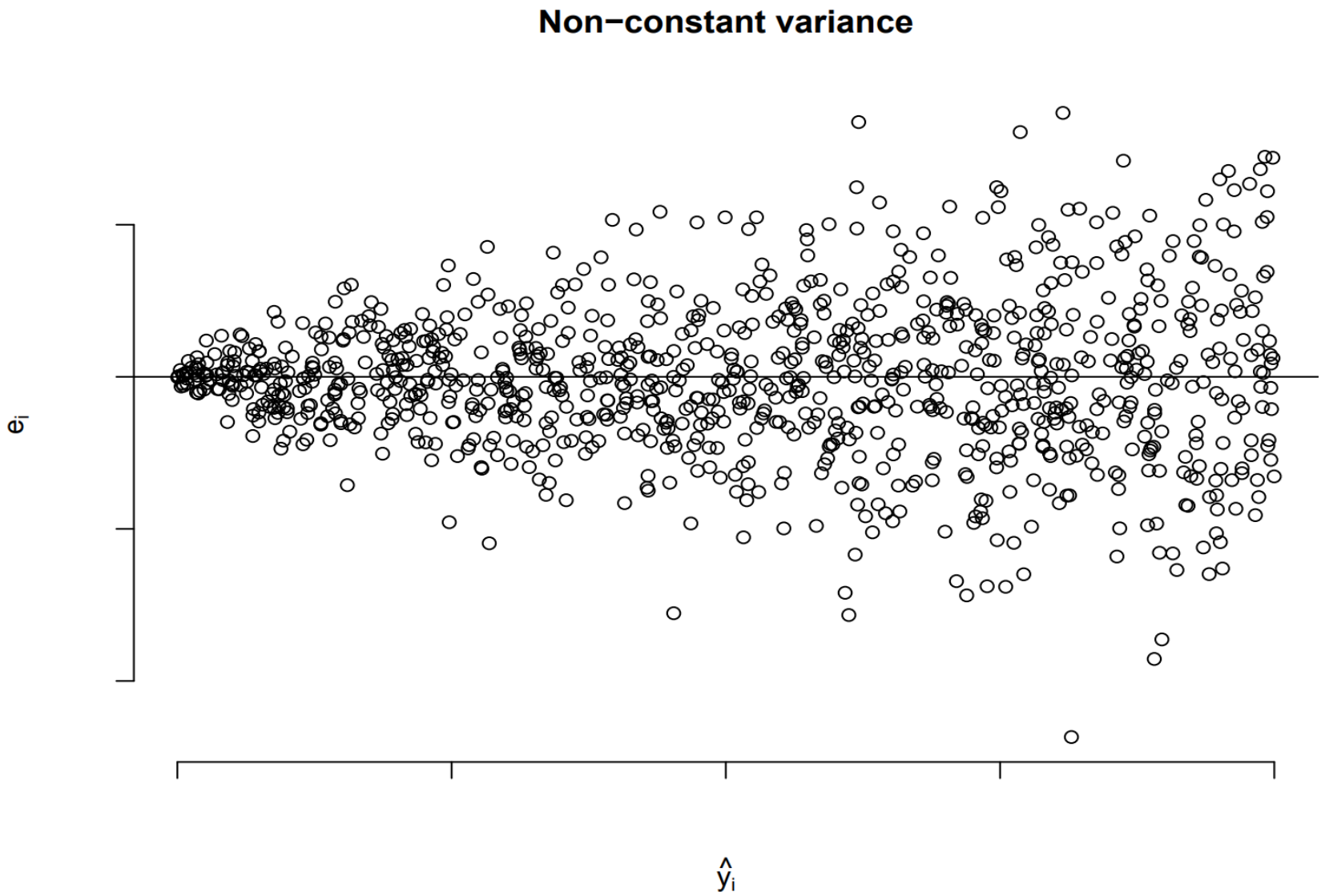
Heteroskedasticita

= **porušení předpokladu o konstantním rozptylu náhodné složky** (homoskedasticita)

- nepříznivě ovlivňuje zejména odhady směrodatných chyb a vypovídají schopnost diagnostických testů
- Testy pro identifikaci:
  - Bartlettův test
  - Levenův test
  - Breusch-Paganův test
  - Score Test
  - F Test

H_0: Homoskedasticita (rozptyl je konstantní)
H_1: Heteroskedasticita (rozptyl není konstantní)

Heteroskedasticita



Řešení heteroskedasticity

- VÁŽENÁ METODA NEJMENŠÍCH ČTVERCŮ
  - podmnožina Zobecněné MNČ
  - řešíme diagonální prvky → stanovíme váhy pro odhad parametrů b

$$\mathbf{b} = (\mathbf{X}^T \hat{\mathbf{\Omega}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{\Omega}}^{-1} \mathbf{y}$$

    - kde $\hat{\Omega}$ je odhad matice rozptylů

  - metod, jak stanovit váhy je několik, např. lze vytvořit regresní model závislosti absolutní hodnoty reziduí na vyrovnaných hodnotách – vahou pak bude převrácená hodnota kvadrátu vyrovnaných hodnot tohoto vytvořeného váhového modelu
    - vahovy.model <- lm(abs(regresni.model$residuals) ~ regresni.model$fitted.values)
    - vaha <- 1/(vahovy.model$fitted.values^2)

PŘÍKLAD 3

Data:

usacities.csv

Zadání: Otestujte pomocí známých testů, zda model z Příkladu 1 je homoskedastický.

V případě že není, použijte váženou metodu nejmenších čtverců k odhadu parametrů modelu.

PŘÍKLAD 4

Data:

Cars.csv

Zadání: Otestujte pomocí známých testů, zda model z předchozího Příkladu 3 (po vyloučení všech korelovaných proměnných) je homoskedastický.

V případě že není, použijte váženou metodu nejmenších čtverců k odhadu parametrů modelu.

Autokorelace

= **korelace náhodné složky v čase** (pouze u časových dat)

- způsobuje:
  - ztrátu vydatnosti odhadu i asymptotickou vydatnost odhadu regresních parametrů → odhady směrodatných chyb jsou vychýlené
  - Index determinace je nadhodnocený
  - t-testy jsou slabé
  - rezidua jsou podhodnocená

- Autokorelace prvního řádu:
  - i-tá hodnota náhodné složky závisí na předchozích hodnotách této složky a na jiné chybě (tzv. bílém šumu)

Testy pro identifikaci autokorelace prvního řádu

- Autokorelační funkce reziduí
- Breusch-Godfreyův test
- Durbin-Watsonův test

H_0: nulová hodnota koeficientu autokorelace = autokorelace není
H_1: autokorelace je

  - Testové kritérium **DW**, která nabývá hodnot 0 až 4, symetrické kolem 2
    - DW = 2 → koeficient autoregrese = 0, jinak je třeba porovnat DW se spočtenou dolní a horní mezí, kdy lze vyloučit či přijmou hypotézu o nekorelovanosti.
    - DW < 2, tak lze čekat kladný koeficient autoregrese
    - DW > 2, koeficient autoregrese bude záporný

- Řešení: vážená metoda nejmenších čtverců

$$\mathbf{b} = (\mathbf{X}^T \hat{\mathbf{\Omega}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{\Omega}}^{-1} \mathbf{y}$$

PŘÍKLAD 5 - AKTIVITA

- **Data**:

USArrests.csv → balíček integrovaný v Rku → data("USArrests")

- **Zadání:** Datový soubor obsahuje statistiky kriminality v USA. Obsahuje čtyři proměnné:
  - Murder: Počet zatčení za vraždu (na 100 000 obyvatel)
  - Assault: Počet zatčení za násilná přepadení (na 100 000 obyvatel)
  - UrbanPop: Procento městské populace
  - Rape: Počet zatčení za znásilnění (na 100 000 obyvatel)

Odhadněte, zda velikost populace ve městech lze vysvětlit pomocí ukazatelů kriminality. Postup jako vždy: sestavení regresního modelu, exploratorní analýza dat, testy normality náhodné složky, multikolinearity vysvětlujících proměnných, heteroskedasticity náhodné složky. Okomentujte postupy, závěry a zapište výsledný model regresní rovnicí.

Shrnutí – základní problémy regresního modelu

- Nevýznamný F-test
  - Řešení → zvolit jiný model
- Nevýznamné t-testy
  - Řešení → zvolit jiný model či zkusit stávající bez nevýznamných proměnných, odstranění odlehlých pozorování
- Nevýznamná konstanta b_0 → hlubší problém, nejspíše multikolinearita
  - Řešení → odstranit některé proměnné, které jsou lineárně závislé na jiných vysvětlujících proměnných či sloučit proměnné (metody shlukové či faktorové analýzy)
- Nenormalita náhodné složky → špatný model (neodstraněn trend), přítomnost odlehlých pozorování
  - Řešení → transformace proměnných (např. logaritmizace, Box a Cox atd.), odstranění odlehlých pozorování, zvolit jiný model

26.10.2022 – Průběžný test 1 – 15 bodů

- Druhá polovina semináře: cca 18:30 – 20:00
- Zadán datový soubor, odpovědi na papír

- Co umět?
  - Deskriptivní statistika
  - Standardizace proměnných
  - Lineární transformace proměnných
  - Vzdálenosti objektů
  - Detekce a vyloučení vlivných pozorování
  - Testy normality
  - Tvorba základních regresních modelů
  - Analýza reziduí
  - Ověření předpokladů KLRM
  - Řešení multikolinearity
  - Řešení heteroskedasticity

Dotazy?

UNICORN UNIVERSITY

UNICORN
UNIVERSITY