UNICORN UNIVERSITY

Regresní analýza I

3. seminář k předmětu Statistické metody v analýze dat

12.10.2022

Martina Šimková

simkova.martinka@gmail.com

ZS 2022/2023

Regresní analýza (1)

- slouží k popisu jednostranné závislosti dvou číselných proměnných, kdy proti sobě stojí vysvětlující (nezávislá) proměnná jako „příčina“ a vysvětlovaná (závislá) proměnná jako „následek“
- regresní funkce = „idealizující“ matematická funkce, která co nejlépe vyjadřuje charakter závislosti
- regresní funkce je podmíněnou střední hodnotou náhodné veličiny Y
  - Teoretická regresní funkce: $\eta = \mathrm{E}(\mathrm{Y}|\mathrm{X} = \mathrm{x}) = \mathrm{Y} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \beta_3 \mathrm{X}_3 + \ldots + \beta_n \mathrm{X}_n + \varepsilon$
  - Odhad: $\widehat{\mathrm{Y}} = \mathrm{b}_0 + \mathrm{b}_1 \mathrm{X}_1 + \mathrm{b}_2 \mathrm{X}_2 + \mathrm{b}_3 \mathrm{X}_3 + \mathrm{b}_n \mathrm{X}_n$
- Může být více vysvětlujících proměnných X:
  - Jedna X: jednoduchá regrese
  - Více X: vícenásobná regrese

Náhodná složka

Regresní analýza (2)

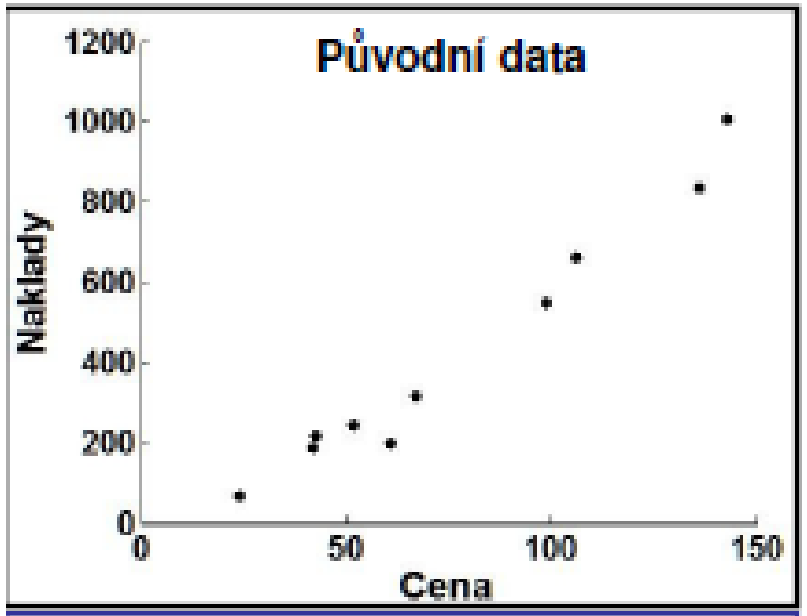
- Náhodná složka = **REZIDUUM** = rozdíl mezi skutečnou a odhadnutou hodnotou
  - Cíl: Minimální reziduální součet čtverců
  - Odhad parametrů modelu: **METODA NEJMENŠÍCH ČTVERCŮ**

$$e_i = y_i - \hat{y}_i; \sqrt{\sum_{i=1}^{n} e_i^2} -> \min$$

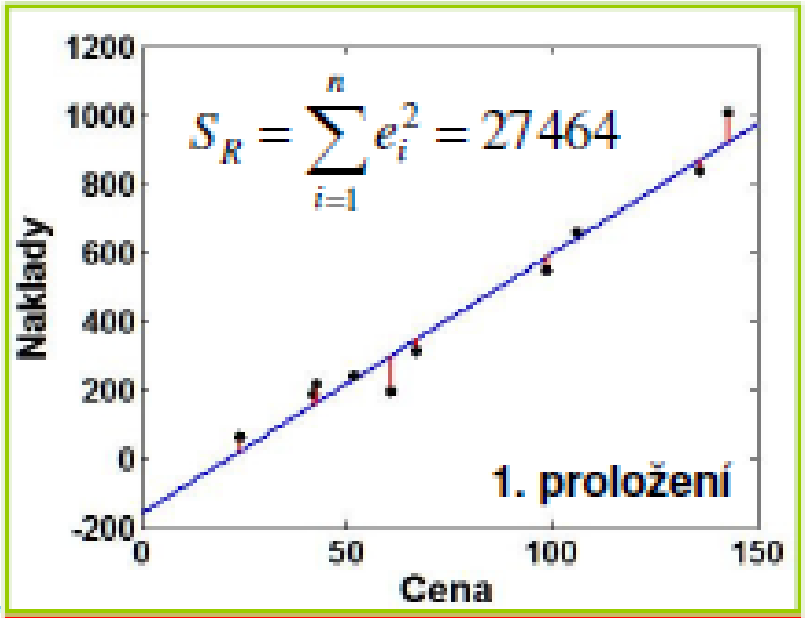
- Náhodná složka má splňovat předpoklady klasického lineárního regresního modelu (KLRM):
  - Nulová střední hodnota reziduí
  - Konstantní rozptyl reziduí
  - Rezidua jsou nekorelovaná
  - Rezidua mají normální rozdělení

… analýza reziduí → významný diagnostický nástroj

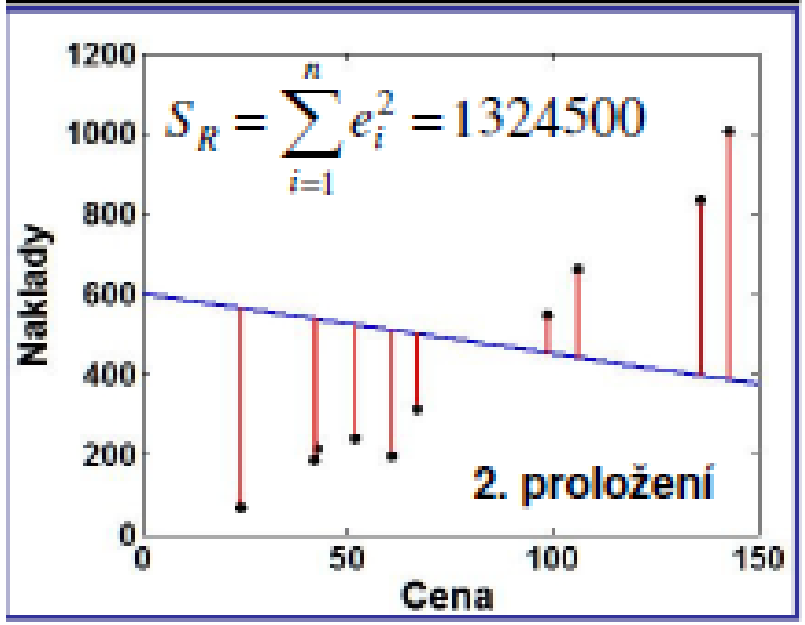
Princip metody nejmenších čtverců

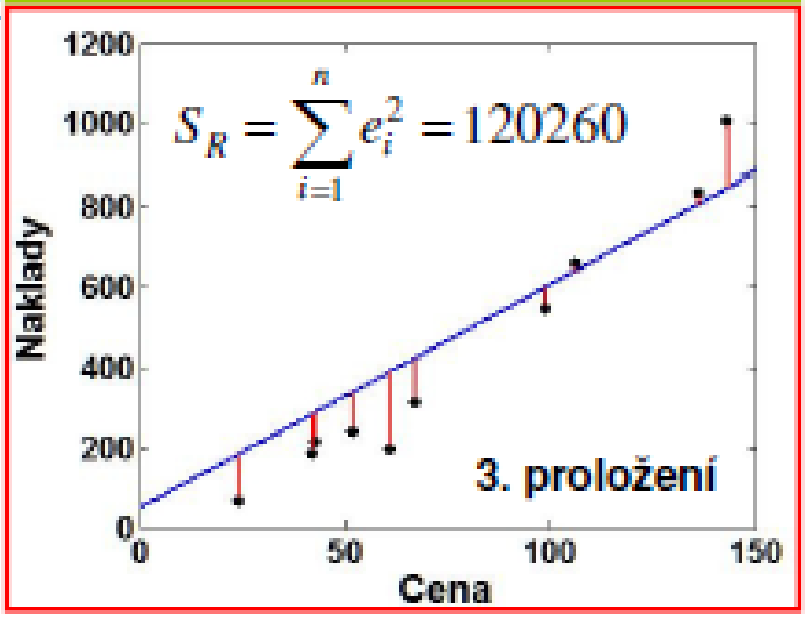


Residua vyznačena červeně



Proložení vpravo nahoře má z daných **tří** proložení **minimální možný součet čtverců** residuí





Postup regresní analýzy

1. nalezení regresního modelu = volba typu regresní funkce
   - regresní funkce = „idealizující“ matematická funkce, která co nejlépe vyjadřuje charakter závislosti
2. odhad parametrů regresního modelu
   - metodou nejmenších čtverců (MNČ) → nalezení takové funkce, pro kterou je reziduální rozptyl nejmenší
3. ověření významnosti regresního modelu a jeho parametrů
   - testování hypotéz: F-test, t-testy
4. posouzení kvality regresního modelu
   - ověření vhodnosti zvoleného regresního modelu pomocí kritérií – např. index determinace, střední čtvercová chyba, apod.
5. odhad střední hodnoty Y pro známé X

Typy regresních funkcí

- Lineární regresní funkce z hlediska parametrů:

přímková regrese $Y = \beta_0 + \beta_1 X,$

hyperbolická regrese $Y = \beta_0 + \frac{\beta_1}{X},$

logaritmická regrese $Y = \beta_0 + \beta_1 \ln X,$

parabolická regrese $Y = \beta_0 + \beta_1 X + \beta_2 X^2$

polynomická regrese $Y = \beta_0 + \beta_1 X + \cdots + \beta_p X^p$

- Pokud regresní funkce není lineární v parametrech, je jednou z možností provést její linearizaci například zlogaritmováním. Pokud to nejde, nelze použít MNČ

- Nelineární regresní funkce:

Funkce	Linearizující transformace
$Y = \beta_0 x^{\beta_1}$	$\ln Y = \ln \beta_0 + \beta_1 \ln x$
$Y = \beta_0 \beta_1^{\frac{1}{x}}$	$\ln Y = \ln \beta_0 + \frac{1}{x} \ln \beta_1$
$Y = \frac{\beta_0}{x^{\beta_1}}$	$\ln Y = \ln \beta_0 - \beta_1 \ln x$
$Y = \beta_0 x^{\beta_1 x}$	$\ln Y = \ln \beta_0 + \beta_1 x \ln x$
$Y = \beta_0 e^{\beta_1 x}$	$\ln Y = \ln \beta_0 + \beta_1 x$
$Y = \frac{1}{\beta_0 + \beta_1 x}$	$\frac{1}{Y} = \beta_0 + \beta_1 x$
$Y = \frac{x}{\beta_0 + \beta_1 x}$	$\frac{x}{Y} = \beta_0 + \beta_1 x$

Testy významnosti

- Test o modelu = Celkový F-test

Test o modelu $p = k + 1$

H_0	H_1	Testové kritérium	Kritický obor
$\beta_0 = c$ $\beta_1 = 0$ ... $\beta_k = 0$	non H_0	$F = \frac{\frac{S_T}{p-1}}{\frac{S_R}{n-p}}$ $F \sim F(p-1, n-p)$	$W_\alpha = \{F; F \geq F_{1-\alpha}\}$

- Testy o parametrech = Individuální t-testy

Test hypotézy o regresním parametru

H_0	H_1	Testové kritérium	Kritický obor
$\beta_j = 0$	$\beta_j \neq 0$	$T = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}}$ $T \sim t(n-p)$	$W_\alpha = \{t; \lvert t \rvert \geq t_{1-\alpha/2}\}$

Posouzení kvality modelu

- vztah je tím silnější a regresní funkce je tím lepší, čím více jsou empirické hodnoty vysvětlované proměnné soustředěné kolem odhadnuté regresní funkce, a naopak tím slabší, čím více jsou vzdálené od odhadnuté regresní funkce
- závislost y a x bude tím silnější, čím větší bude podíl rozptylu vyrovnaných hodnot (S_T) na celkovém rozptylu (S_y)

→ **koeficient (index) determinace**

$$R^2 = I^2 = \frac{S_T}{S_y}$$

... měří se tzv. **těsnost závislosti**
(měří kvalitu modelu)

$$S_Y = \sum_{i=1}^{n} (y_i - \bar{y})^2 \leftarrow \text{Celkový souč. čtv.}$$

$$S_R = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \leftarrow \text{Residuální souč. čtv.}$$

$$S_T = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \leftarrow \text{Teoretický souč. čtv.}$$

$$S_y = S_T + S_R$$

Intervaly spolehlivosti pro střední hodnotu

1. **IS pro podmíněnou střední hodnotu vysvětlované proměnné Y**:

$$P(\hat{y}_i - t_{1-\frac{\alpha}{2}}(n-2) \cdot s_{\hat{y}} < Y_i < \hat{y}_i + t_{1-\frac{\alpha}{2}}(n-2) \cdot s_{\hat{y}}) = 1 - \alpha$$

kde $s_{\hat{y}}$ je směrodatná chyba odhadu: $s_{\hat{y}} = s\sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$

kde s je reziduální směrodatná odchylka: $s = \sqrt{\frac{\sum e_i^2}{n-(p+1)}} = \sqrt{\frac{S_R}{n-(p+1)}}$

2. **IS pro konkrétní střední hodnotu Y:**

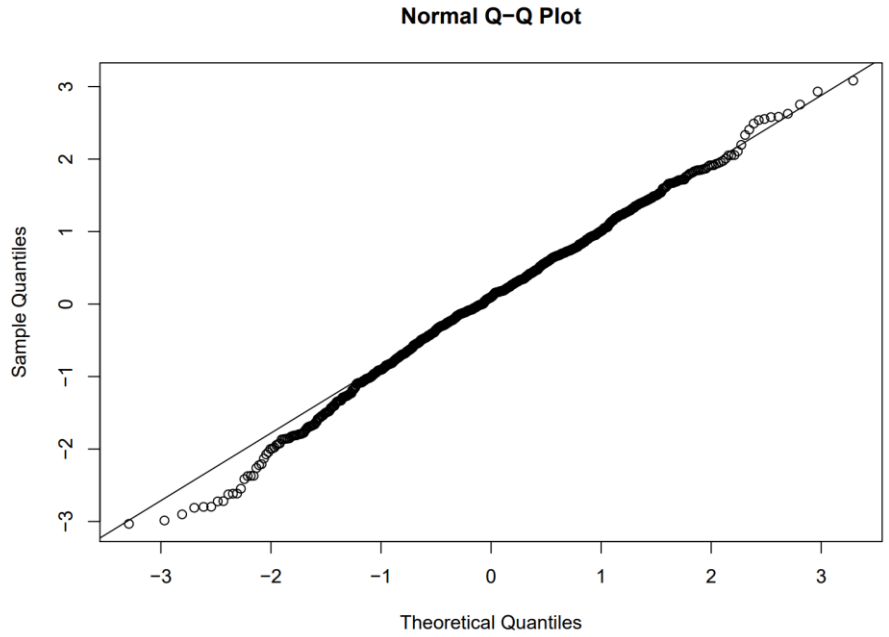
$$P(\hat{y}_i - t_{1-\frac{\alpha}{2}}(n-2) \cdot s_{\hat{y}} < E(Y_i) < \hat{y}_i + t_{1-\frac{\alpha}{2}}(n-2) \cdot s_{\hat{y}}) = 1 - \alpha$$

kde $s_{\hat{y}}$ je: $s_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$

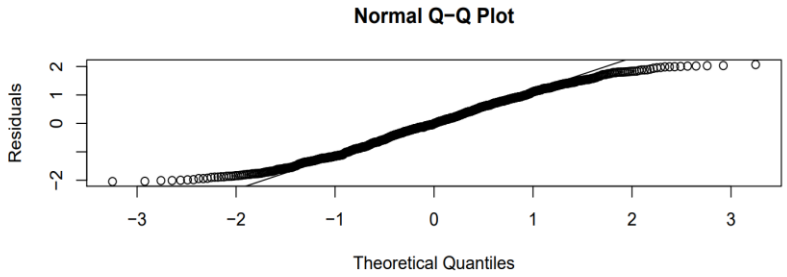
Význam reziduí

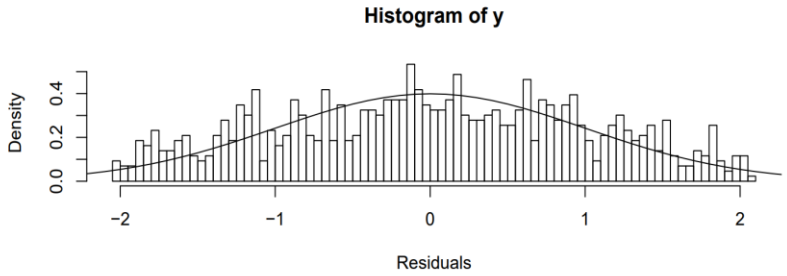
- reziduum = rozdíl mezi skutečnou a vyrovnanou hodnotou
- základní diagnostický nástroj
- požadavek nejmenšího reziduálního součtu čtverců (předpoklad MNČ)
- systematičnost v chování reziduí:
  - vlivné či odlehlé hodnoty
  - porušení předpokladů KLRM (autokorelace, normalita, heteroskedasticita)
- grafická analýza:
  - bodový graf: posouzení předpokladů modelu, identifikace odlehlých pozorování
  - histogram, kvantilový graf: posouzení normality

Samples from $N(0, 1)$ distribution

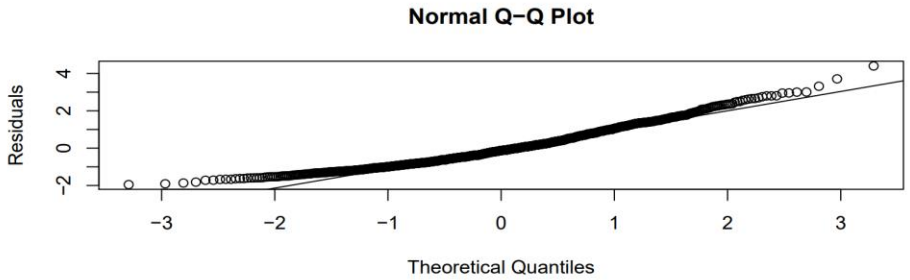


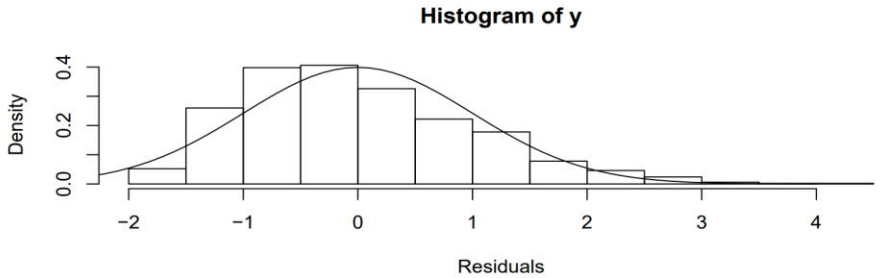
Samples from a light-tailed distribution



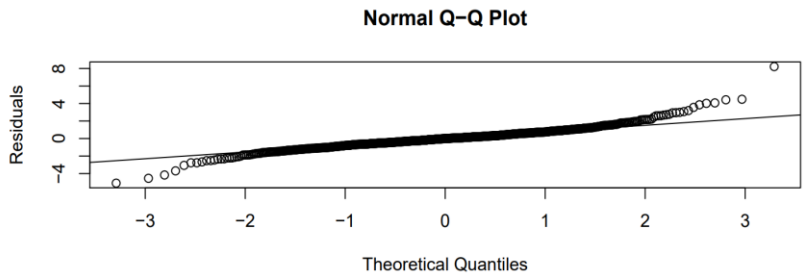


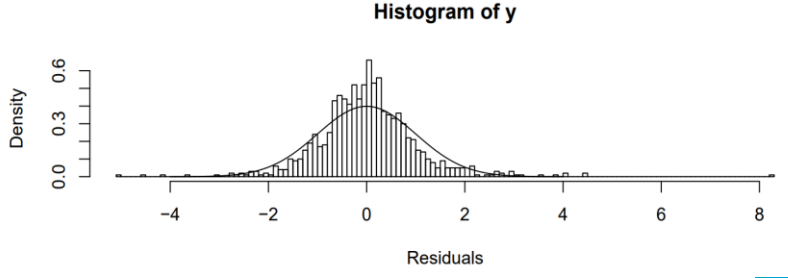
Samples from a skewed distribution

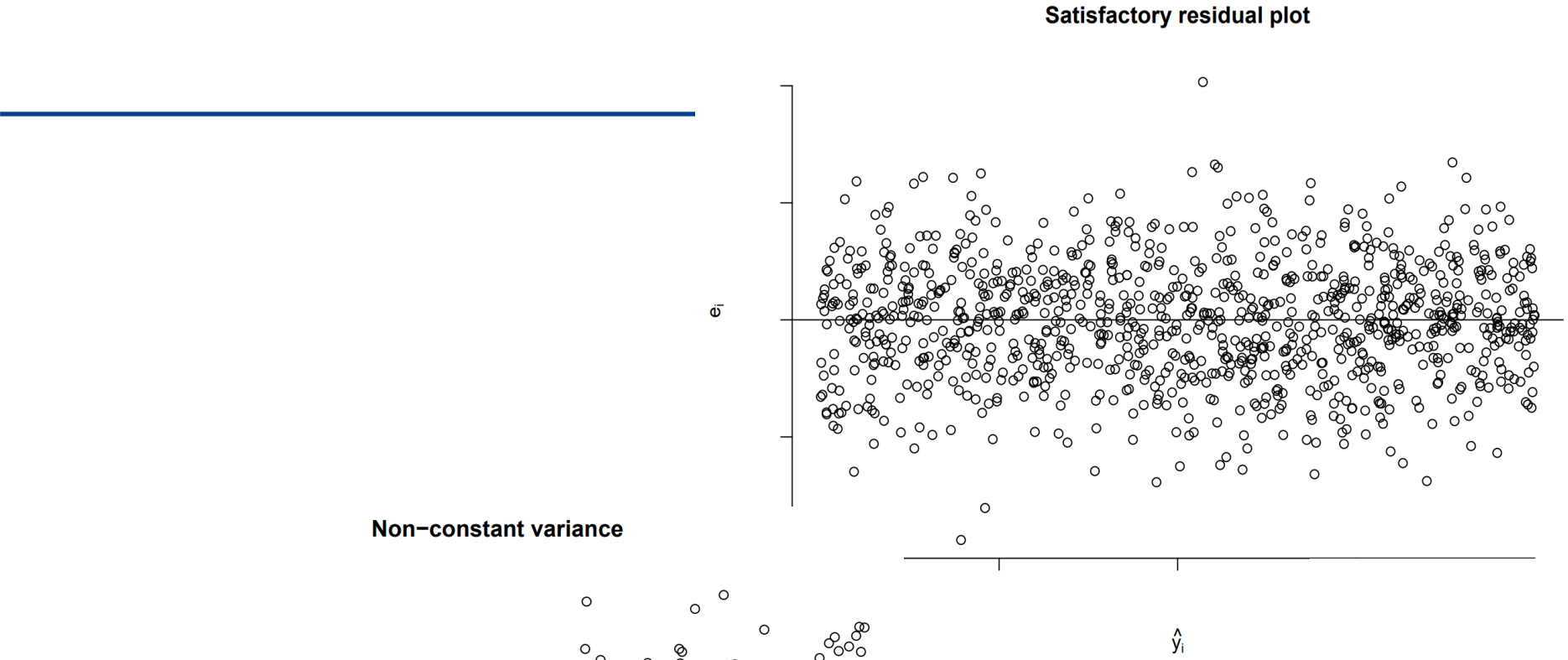


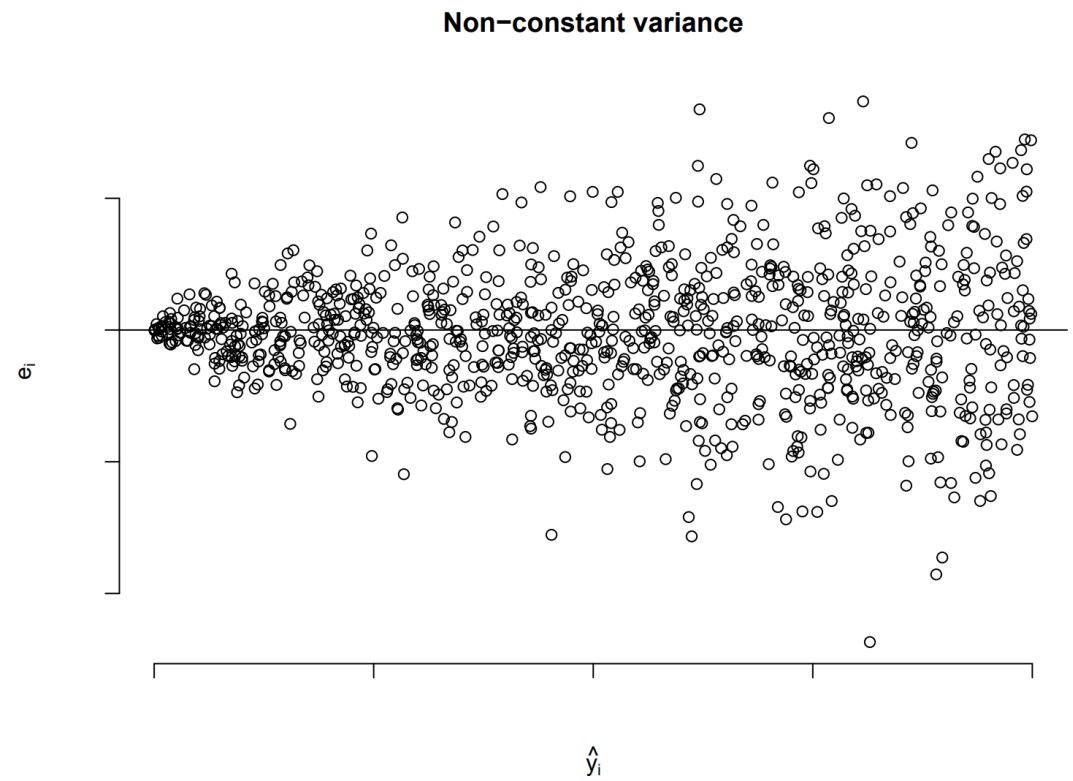


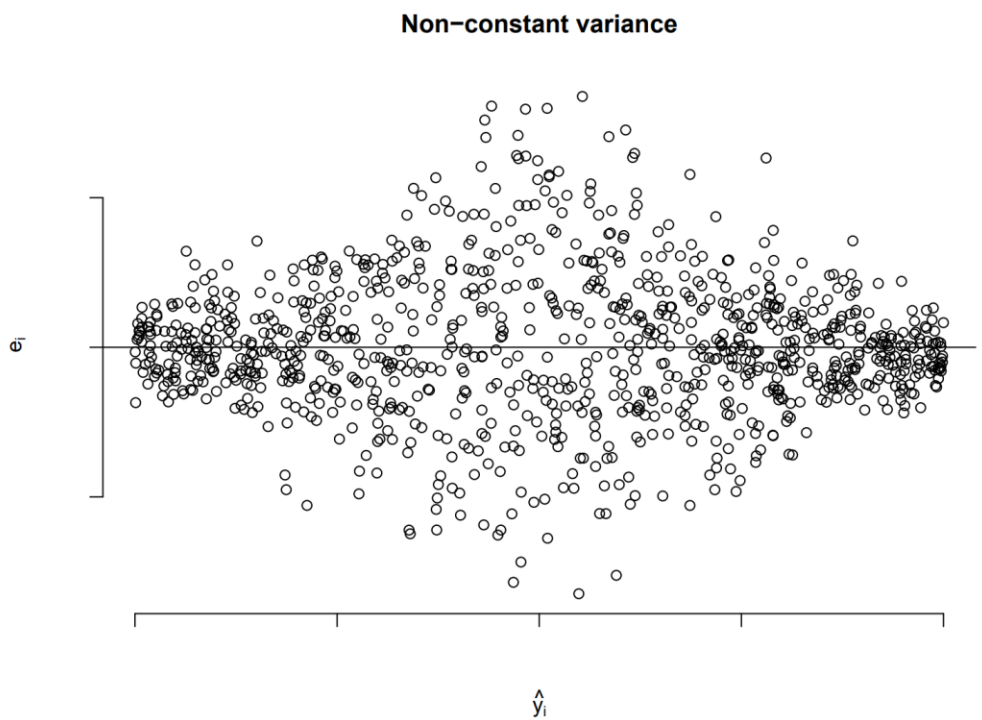
Samples from a heavy-tailed distribution






Satisfactory residual plot
e_i
$\hat{y}_i$


Non−constant variance
e_i
$\hat{y}_i$


Non−constant variance
e_i
$\hat{y}_i$

Regresní přímka

$$Y = b_0 + b_1 \cdot x$$

- odhad parametrů metodou nejmenších čtverců

$$b_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2}$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x}$$

  - **b_1 = výběrový regresní koeficient** = směrnice regresní přímky (udává změnu průměru závisle proměnné y při jednotkové změně nezávislé proměnné x), při lineární nezávislosti je roven 0
  - **b_0** = počáteční hodnota při x=0 (průsečík s osou y)

PŘÍKLAD 1

Data:

MMDA_03_data.xlsx

Firmy.csv

Zadání: V tabulce jsou uvedeny údaje o hodnotě *produkce* (ve 100 000 Kč) a o výši *investic* (v 10 000 Kč) v souboru 12 vybraných firem s počtem zaměstnanců větším než 20.

1. Stanovte rovnici regresní přímky modelující závislost hodnoty produkce na výši investic.
2. Proveďte bodový a intervalový odhad očekávané výše produkce firmy, která investuje 18 000 Kč.

Korelační analýza

- závislost jednostranná (jeden znak vystupuje jako příčina (nezávisle proměnná - X) a druhý znak jako následek (závisle proměnná - Y) → **regrese**
- závislost vzájemná (oboustranná) – lineární závislost → **korelace**
- síla závislosti mezi dvěma proměnnými → **jednoduchý (párový) korelační koeficient**
  - definován jako poměr kovariance s_{yx} a součinu směrodatných odchylek obou proměnných s_x a s_y

$$r_{xy} = r_{yx} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\sqrt{(\overline{x^2} - \overline{x}^2)(\overline{y^2} - \overline{y}^2)}} = \frac{s_{xy}}{s_x s_y}$$

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Kovariance

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n} = \overline{xy} - \overline{x}\,\overline{y}$$

může nabývat kladných i záporných hodnot a její znaménko určuje směr závislosti

Korelační koeficient

r_{xy} = 1 → přímá funkční závislost

r_{xy} = -1 → nepřímá funkční závislost

r_{xy} = 0 → lineární nezávislost *

* !!! Lineární nezávislost ≠ nulová závislost !!!

Pouze v případě regresní přímky = odmocnina z indexu determinace je korelační koeficient

$$r_{xy} = \sqrt{R^2}$$

- KK zachycuje:
  - sílu lineární závislosti mezi dvěma proměnnými (proměnné jsou silně lineárně závislé, pokud je KK v absolutní hodnotě blízký 1)
  - směr lineární závislosti, ve smyslu přímá vs. nepřímá (záporné hodnoty KK představují nepřímou lineární závislost, kladné hodnoty představují lineární závislost přímou)

- Test o významnosti KK:

H_0	H_1	Testové kritérium	Kritický obor
$\rho_{XY} = 0$	$\rho_{XY} \neq 0$	$T = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$ $T \sim t(n-2)$	$W_\alpha = \{t; \lvert t \rvert \geq t_{1-\alpha/2}\}$

Vysoká hodnota výběrového korelačního koeficientu nemusí ještě znamenat silnou závislost v ZS, neboť může být zkreslena v důsledku náhodnosti výběru, zejména v případě malých výběrů.

PŘÍKLAD 2

Data:

MMDA_03_data.xlsx

Firmy.csv

Zadání: V tabulce jsou uvedeny údaje o hodnotě *produkce* (ve 100 000 Kč) a o výši *investic* (v 10 000 Kč) v souboru 12 vybraných firem s počtem zaměstnanců větším než 20.

Vypočítejte korelační koeficient mezi oběma proměnnými a otestujte, zda je statisticky významný.

PŘÍKLAD 3

Data:

MMDA_03_data.xlsx

Cons.csv

Zadání: U třech typů aut se sledovala závislost *spotřeby automobilu* (v l/100 km) na *rychlosti* (v km/h). Stanovte nejvhodnější regresní funkci modelující tuto závislost. Následně odhadněte interval spolehlivosti očekávané spotřebu automobilu, který jede rychlostí 115 km/h.

Regresní funkce v R

- Regresní přímka: lm(Y ~ X, data=???)
- Hyperbolická funkce: lm(Y ~ I(X^-1), data=???)
- Logaritmická funkce: lm(Y ~ log(X), data=???)
- Regresní parabola: lm(Y ~ X+ I(X^2), data=???)
- Polynom 3.stupně: lm(Y ~ X+ I(X^2) + I(X^3), data=???)
- Mocninná funkce: lm(log(Y) ~ log(X), data=???)
- Exponenciální funkce: lm(log(Y) ~ X, data=???)

přímková regrese $Y = \beta_0 + \beta_1 X$,

hyperbolická regrese $Y = \beta_0 + \frac{\beta_1}{X}$,

logaritmická regrese $Y = \beta_0 + \beta_1 \ln X$,

parabolická regrese $Y = \beta_0 + \beta_1 X + \beta_2 X^2$

polynomická regrese $Y = \beta_0 + \beta_1 X + \cdots + \beta_p X^p$

Funkce	Linearizující transformace
$Y = \beta_0 x^{\beta_1}$	$\ln Y = \ln \beta_0 + \beta_1 \ln x$
$Y = \beta_0 \beta_1^{\frac{1}{x}}$	$\ln Y = \ln \beta_0 + \frac{1}{x} \ln \beta_1$
$Y = \frac{\beta_0}{x^{\beta_1}}$	$\ln Y = \ln \beta_0 - \beta_1 \ln x$
$Y = \beta_0 x^{\beta_1 x}$	$\ln Y = \ln \beta_0 + \beta_1 x \ln x$
$Y = \beta_0 e^{\beta_1 x}$	$\ln Y = \ln \beta_0 + \beta_1 x$
$Y = \frac{1}{\beta_0 + \beta_1 x}$	$\frac{1}{Y} = \beta_0 + \beta_1 x$
$Y = \frac{x}{\beta_0 + \beta_1 x}$	$\frac{x}{Y} = \beta_0 + \beta_1 x$

PŘÍKLAD 4

Data:

MMDA_03_data.xlsx

Beer.csv

Zadání: Datový soubor obsahuje údaje o *cenách* (v $) a *prodaných kusech** za 3 velikosti plechovek piva v malém řetězci supermarketů za 52 týdnů. V souboru je tedy 6 proměnných.

Na základě korelační analýzy vyberte nejvhodnější kombinaci vysvětlované a vysvětlující proměnné a podle nich modelujte závislost prodaného množství plechovek piva na ceně pomocí regresní analýzy. Následně odhadněte očekávaný počet prodaných kusů plechovek piva při ceně 15$.

* Proměnné cena a prodané množství jsou převedeny na jednotlivé případy (na 24 plechovek), aby bylo možné přímo porovnávat ceny a množství v grafech a modelových koeficientech.

PŘÍKLAD 5 – AKTIVITA

Data:

MMDA_03_data.xlsx

Cars.csv

Zadání: Zjistěte, zda existuje závislost ceny auta (*cena_prodej*) na výkonu (*horsepower*). Najděte pro tuto závislost nejvhodnější regresní funkci, otestujte její významnost a následně odhadněte (bodově i intervalově), jakou cenu auta můžeme očekávat u auta, které má výkon 400 koní.

Vícenásobná regrese

- Závislost proměnné Y na více vysvětlujících proměnných

$$Y = \hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

- Korelační koeficienty:
  - **Dílčí KK** měří sílu lineární závislosti proměnné *y* na x_p za předpokladu, že všechny ostatní proměnné *x* jsou konstantní
  - **Vícenásobný KK** měří sílu závislosti na všech vysvětlujících proměnných – umožňuje posoudit kvalitu regresního modelu

- Pozor!!! Multikolinearita = závislost mezi vysvětlujícími proměnnými → nežádoucí jev...viz později

PŘÍKLAD 6

Data:

MMDA_03_data.xlsx

Cars.csv

Zadání: Sestrojte vícenásobný regresní model pro závislost ceny auta (*Cena_prodej*) na objemu motoru (*Motor*), výkonu (*Horsepower*) a spotřebě na dálnici (*Dalnice_MPG*).

Následně odhadněte interval spolehlivosti pro očekávanou cenu auta, které má objem motoru 4 litry, výkon 250 koní a spotřebu na dálnici 30 mpg.

Vlastnosti MNČ odhadů za podmínek KLRM

- Za podmínek klasického lineárního regresního modelu (KLRM), si definujme **reziduum** jako rozdíl mezi skutečnou a vyrovnanou hodnotou a následně tzv. reziduální součet čtverců a ten řešíme metodou nejmenších čtverců:

$$\mathbf{e} = \mathbf{y} - \mathbf{Xb}$$

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

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \rightarrow \text{projekční matice H}$$

$$\hat{\mathbf{y}} = \mathbf{Xb} = \mathbf{Hy}$$

- čím vzdálenější je bod x_i od průměru $\bar{x}$, tím větší váhu má odpovídající hodnota y_i na odhad $\hat{y}_i$

Projekční matice H

- Významný diagnostický nástroj pro hodnocení vlivu jednotlivých pozorování na regresní odhady
- Nejdůležitější diagonální prvky = **leverages** (efekty, projekční h-prvky)
  - jejich součet = stopa matice H = počet parametrů modelu

$$\sum h_{ii} = \mathrm{st}(\mathbf{H}) = K + 1 = p$$

- Odhalují vlivná pozorování:

$$h_{ii} > \frac{2p}{n}$$

popř. $h_{ii} > \frac{3p}{n}$ (pro vyšší počet parametrů (cca 6) a n-p větší než 12)

PŘÍKLAD 7

Data:

Cars.csv

Zadání: Sestrojte vícenásobný regresní model pro závislost ceny auta (*Cena_prodej*) na objemu motoru (*Motor*), výkonu (*Horsepower*) a spotřebě na dálnici (*Dalnice_MPG*).

Vypočítejte projekční matici H a identifikujte vlivná pozorování.

- Ukázku maticového výpočtu pro prvních 10 pozorování najdete v *MMDA_03_data.xlsx*

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, …

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

Dotazy?

UNICORN UNIVERSITY

UNICORN
UNIVERSITY