# Cvičení ze statistiky - 3

Filip Děchtěrenko

# Minule bylo..

- Dokončili jsme základní statistiky, typy proměnných a začali analýzu kvalitativních dat
- Tyhle termíny by měly být známé:
  - Histogram, krabicový graf
  - Standardní skóry, z-skór
  - Kvalitativní a kvantitativní proměnné
  - Nominální, ordinální proměnné
  - Diskrétní a spojité
  - Čtyřpolní koeficient korelace
  - Korelace není kauzalita!

# Příklad

- Dělal se výzkum, zda lidé, kteří jsou roztěkaní, stíhají autobus na poslední chvíli
- Otázka: Jak jsou proměnné?
- X - je roztěkaný/není roztěkaný
  Y – stíhá autobus na poslední chvíli/ stíhá v pohodě
- Výsledky:

| | Y=1 | Y=2 | Suma |
|---|---|---|---|
| X=1 | 43 | 56 | 99 |
| X=2 | 12 | 65 | 77 |
| Suma | 55 | 121 | 176 |

- Existuje korelace mezi tím, když je člověk roztěkaný a stíhá/nestíhá autobus na poslední chvíli?

# Příklad pokračování

| | Y=1 | Y=2 | Suma |
|---|---|---|---|
| X=1 | 43 | 56 | 99 |
| X=2 | 12 | 65 | 77 |
| Suma | 55 | 121 | 176 |

- Ze vzorce spočítáme:

$$r_\phi = \frac{43 \cdot 65 - 56 \cdot 12}{\sqrt{99 \cdot 77 \cdot 121 \cdot 55}} = 0.30$$

- Vyšla nám střední korelace mezi roztěkaností a nestíháním autobusu
- Pozor! Korelovat můžeme cokoli s čímkoli, ale ne vždy to má smysl (garbage in, garbage out)

# X i Y kvantitativní

- Např. mám výšku a váhu
- Může nás zajímat, jak se jedna hodnota vyskytuje s druhou ->korelace
- Nebo jak zapsat rovnicí vztah mezi proměnnými -> regrese

# Korelace

- Jde o vyjádření, jak moc se vyskytují hodnoty proměnné spolu (síla vztahu)

$$r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}} = \frac{s_{xy}}{s_x s_y}$$

- Jde o podíl kovariance a součinu směr. odchylek
- Kovariance určuje, jak moc se proměnné mění společně
- Korelace je opět jen normovaná kovariance, aby se to pěkně porovnávalo

$$s_{xy} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{n - 1}$$

# Regrese

- Umožňuje nám zjistit, jak vypadá závislost mezi proměnnými
- Obecně je úkol regrese nalézt funkci, která z X předpovídá Y, tedy $f(X) = \hat{Y}$
- Závislost může být libovolná, ale my budeme uvažovat jen lineární

# Lineární regrese

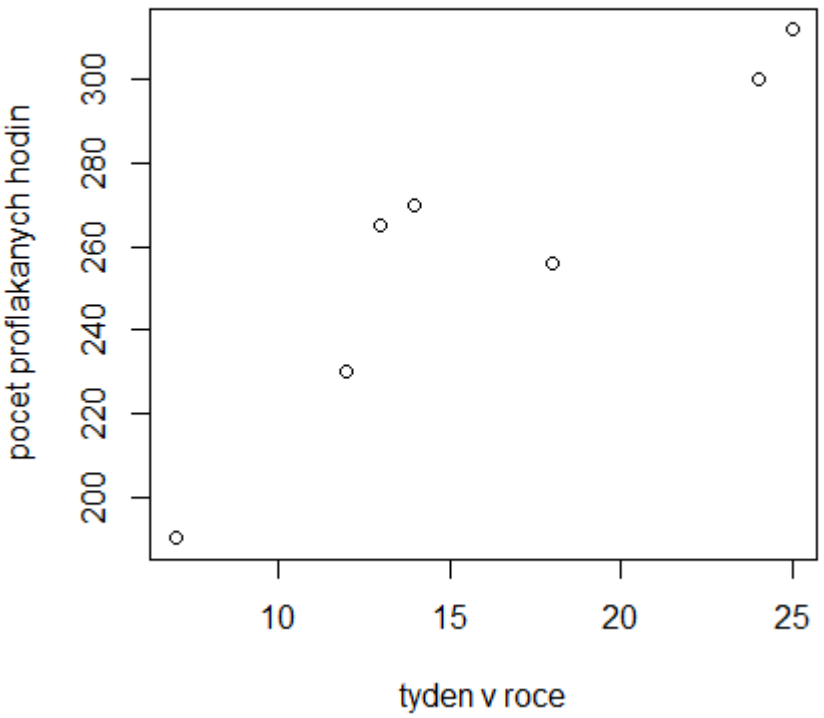
- Jednoduchá varianta – Y dostaneme jako lineární kombinaci X, tedy
$$\hat{Y} = bX + a$$
- Pro každý bod $x_i$ nám tako funkce počítá předpokládanou hodnotu $\hat{y_i}$
- Ta se ale může od skutečné hodnoty $y_i$ lišit!
- Rozdíl mezi skutečnou hodnotou a předpovězenou hodnotou budeme nazývat *residuum (a* značit $\varepsilon_i$)

# Jak to vypadá graficky?

- Mějme data

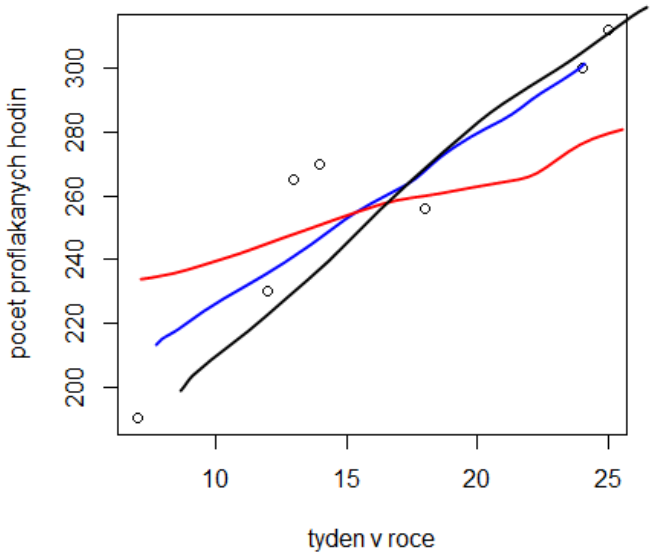
| **Týden v roce** | **Počet prolelkovaných hodin** |
|---|---|
| 12 | 230 |
| 7 | 190 |
| 18 | 256 |
| 25 | 312 |
| 13 | 265 |
| 14 | 270 |
| 24 | 300 |

# Zobrazíme-li si je



# Proložení přímkou

- Máme podezření, že by počet prolelkovaných hodin mohl lineárně záviset (tj. přímka) na týdnu v roce
- Jenže která přímka je nejlepší?



# Metoda nejmenších čtverců

- Idea: budu hledat takovou přímku, která minimalizuje residua $\varepsilon$ (rozdíl mezi naměřenou hodnotou $y$ a předpovězenou hodnotou $\hat{y}$)
- Formálně:

$$\arg\min_{a,b} \sum_i (y_i - (a + bx_i))^2$$

- Mocníme na druhou, abychom se zbavili záporného rozdílu (běžná finta). Proto se to nazývá metoda nejmenších čtverců

# Jak spočítat koeficienty $a$ a $b$

- Koeficient $b$ (směrnice přímky) spočítáme ze vztahu

$$b = \frac{s_{xy}}{s_x^2}$$

kde $s_{xy}$ je kovariance a $s_x^2$ je rozptyl proměnné x

- Pro připomenutí:

$$s_{xy} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{n-1} \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

# Keoficient $a$

- Vypočítáme ho ze vztahu

$$a = \bar{y} - b\bar{x}$$

kde $\bar{y}$ a $\bar{x}$ jsou průměry proměnné X a Y

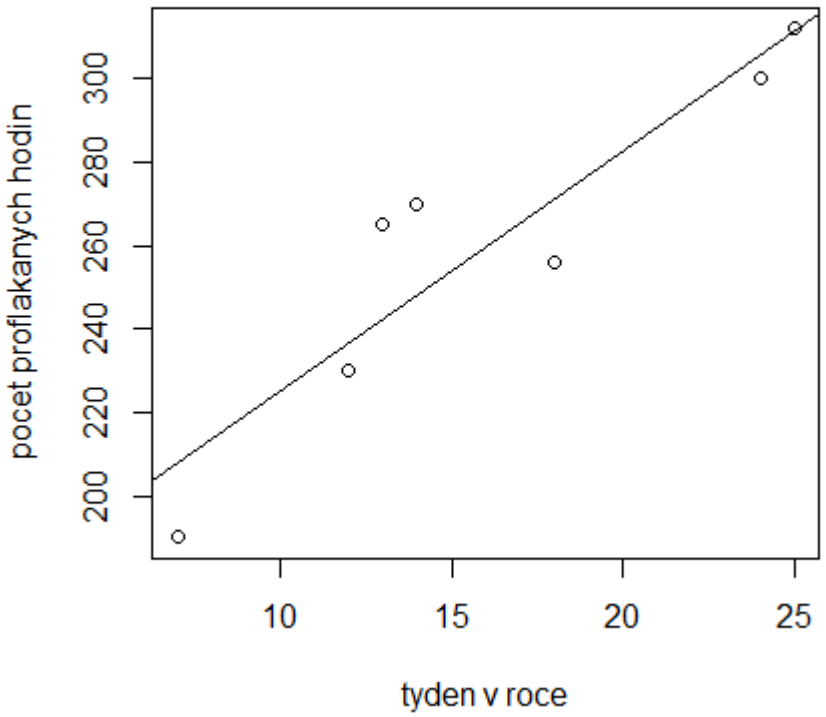
# Příklad

- Týden v roce – X, $m_x = 16.14$, $s_x^2 = 43.14$
  Počet prol. hodin – Y, $m_y = 260.43$, $s_x^2 = 1707.286$

| X | Y | $X - \bar{X}$ | $Y - \bar{Y}$ |
|---|---|---|---|
| 12 | 230 | -4.14 | -30.43 |
| 7 | 190 | -9.14 | -70.43 |
| 18 | 256 | 1.86 | -4.43 |
| 25 | 312 | 8.86 | 51.57 |
| 13 | 265 | -3.14 | 4.57 |
| 14 | 270 | -2.14 | 9.57 |
| 24 | 300 | 7.86 | 39.57 |

- Tedy $s_{xy} = 249.10, \mathrm{b} = 5.77, \mathrm{a} = 167.22$

# Vypočítaná závislost graficky



# Analýza našich naměřených dat

- Ptali jsme se lidí z ročníku na následující věci:
  $X_1$: výška (v cm)
  $X_2$: pohlaví
  $X_3$: Oblíbené zvíře (pes nebo kočka)
  $X_4$: počet stránek oblíbené knihy
- O jaký typ proměnných jde?
- Prvních pár záznamů datové matice:

| | Id | Vyska | Pohlavi | Zvire | Stranky |
|---|---|---|---|---|---|
| 1 | 1 | 156 | Žena | Pes | 126 |
| 2 | 2 | 170 | Žena | Kočka | 394 |
| 3 | 3 | 170 | Žena | Kočka | 436 |
| 4 | 4 | 155 | Žena | Kočka | 94 |
| 5 | 5 | 171 | Žena | Pes | 161 |
| 6 | 6 | 169 | Žena | Pes | 300 |

# Možné vztahy

- Můžeme se ptát na vztahy mezi různými kombinacemi proměnných:
  - Existuje vztah mezi pohlavím a oblíbeným zvířetem
  - Existuje vztah mezi pohlavím a výškou?
  - Existuje vztah mezi výškou a počtem stránek oblíbené knížky?
  - …
- Celkem máme $\binom{4}{2}$=6 možných kombinací (kombinační čísla budou později) pro analýzu

# Vztah mezi pohlavím a zvířetem

- Zjistěte, zda ve vzorku existuje závislost mezi pohlavím a oblíbeným zvířetem
- Počty jednotlivých výskytů (n=25):
  - Muž – kočka: 1
  - Muž – pes: 2
  - Žena – kočka: 8
  - Žena – pes: 14

# Vztah mezi pohl. a zvířetem - řešení

- Uděláme kontingenční tabulku (součty nejsou zobrazeny)

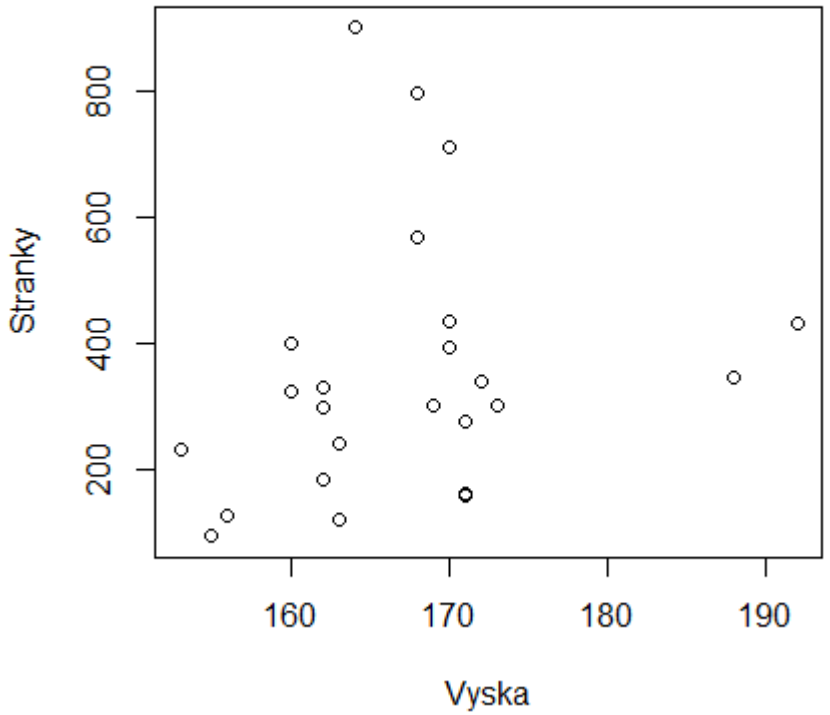
| | Zvire | |
|---|---|---|
| Pohlavi | Kočka | Pes |
| Muž | 1 | 2 |
| Žena | 8 | 14 |

$$r_\phi = \frac{1 \cdot 14 - 2 \cdot 8}{\sqrt{3 \cdot 22 \cdot 9 \cdot 16}} = -0.02$$

- Co znamená výsledek?
- -> ve výběru není závislost mezi pohlavím a oblíbeností zvířete

# Vztah mezi počtem stránek a výškou

- Graficky:




Existuje závislost? Jak je silná a jak vypadá?

# Vztah mezi počtem stránek a výškou

- Několik deskriptivních statistik

| | Id | Vyska | Pohlavi | Zvire | Stranky |
|---|---|---|---|---|---|
| median | 13.00 | 168.000 | NA | NA | 301.00 |
| mean | 13.00 | 167.000 | NA | NA | 350.88 |
| SE.mean | 1.47 | 1.783 | NA | NA | 41.02 |
| CI.mean.0.95 | 3.04 | 3.680 | NA | NA | 84.67 |
| var | 54.17 | 79.500 | NA | NA | 42070.44 |
| std.dev | 7.36 | 8.916 | NA | NA | 205.11 |
| coef.var | 0.57 | 0.053 | NA | NA | 0.58 |

- Kovariační matice:

| | Id | Vyska | Stranky |
|---|---|---|---|
| Id | 54 | 17 | 317 |
| Vyska | 17 | 80 | 417 |
| Stranky | 317 | 417 | 42070 |

- Jaký je koeficient korelace? Jaké jsou regresní koeficienty?

# Řešení

- Všechny potřebné hodnoty máme v tabulce:
- Konkrétně (Výška – X, Stránky – Y):
  - $m_x = 167,\ s_x = 8.92$
  - $m_y = 350.88,\ s_y = 205.11$
  - $s_{xy} = 417$

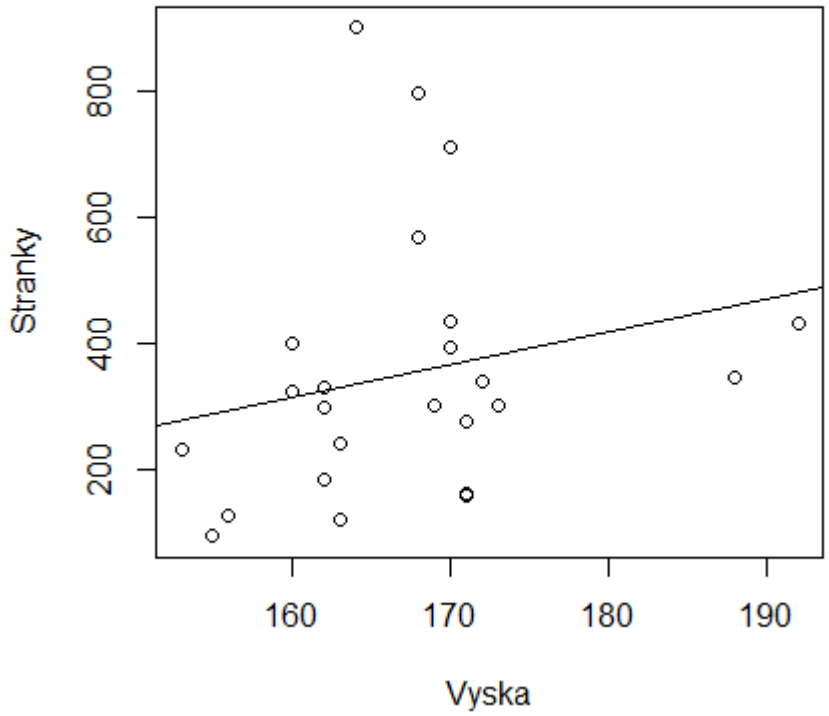
| | Id | Vyska | Pohlavi | Zvire | Stranky |
|---|---|---|---|---|---|
| median | 13.00 | 168.000 | NA | NA | 301.00 |
| mean | 13.00 | 167.000 | NA | NA | 350.88 |
| SE.mean | 1.47 | 1.783 | NA | NA | 41.02 |
| CI.mean.0.95 | 3.04 | 3.680 | NA | NA | 84.67 |
| var | 54.17 | 79.500 | NA | NA | 42070.44 |
| std.dev | 7.36 | 8.916 | NA | NA | 205.11 |
| coef.var | 0.57 | 0.053 | NA | NA | 0.58 |

| | Id | Vyska | Stranky |
|---|---|---|---|
| Id | 54 | 17 | 317 |
| Vyska | 17 | 80 | 417 |
| Stranky | 317 | 417 | 42070 |

- Korelace tedy je 0.23 (slabá závislost) a regresní koeficienty $a = -525.08, b = 5{,}25$

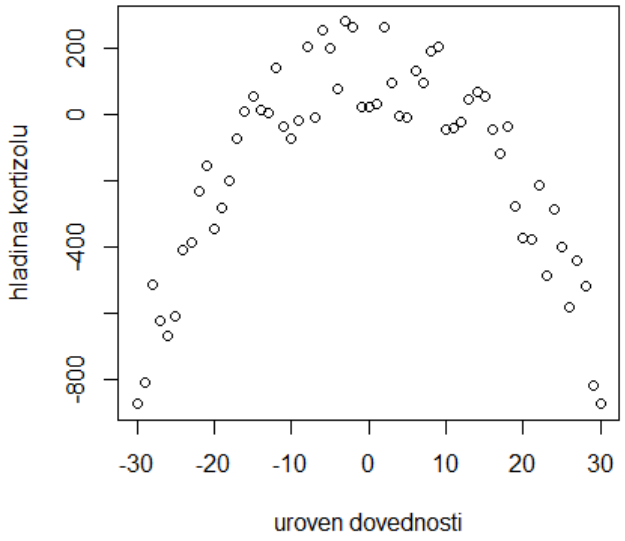
# Regresní přímka v grafu

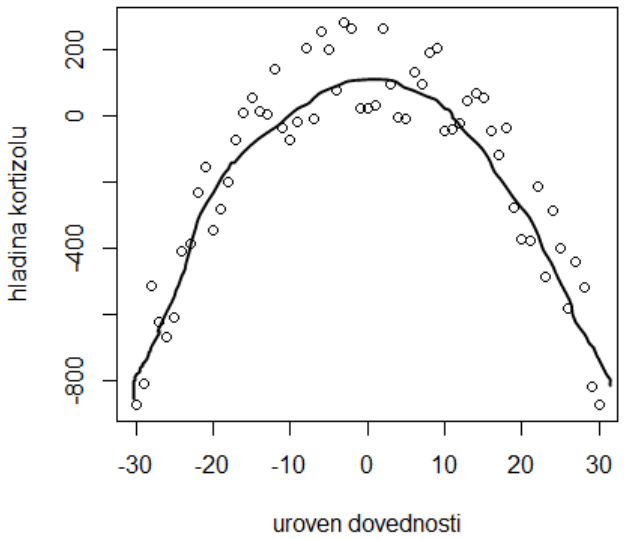
- Vždy se chce zamyslet, zda nepočítáme hlouposti (jako např. korelace počtu stránek a a výšky)



# Jiné typy závislostí

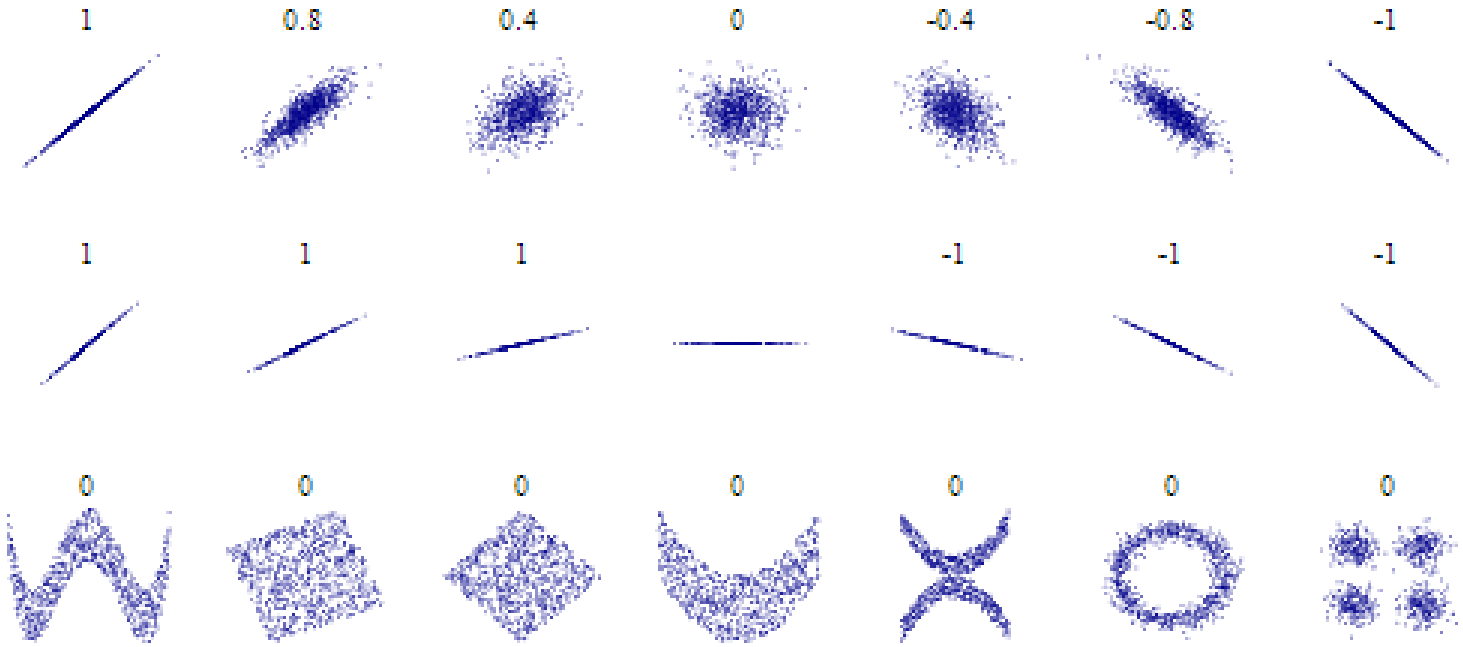
- Měřili jsme množství kortizolu (stresový hormon) při nějaké konkrétní dovednosti
- Jednoduchá regrese není to pravé
- Zkusme jinou křivku $Y = -X^2$
- Dají se opět spočítat parametry (ale to dělat nebudeme)
- Nestačí nám na to Pearsonův korelační koef. (ten je na lin. Závislost)



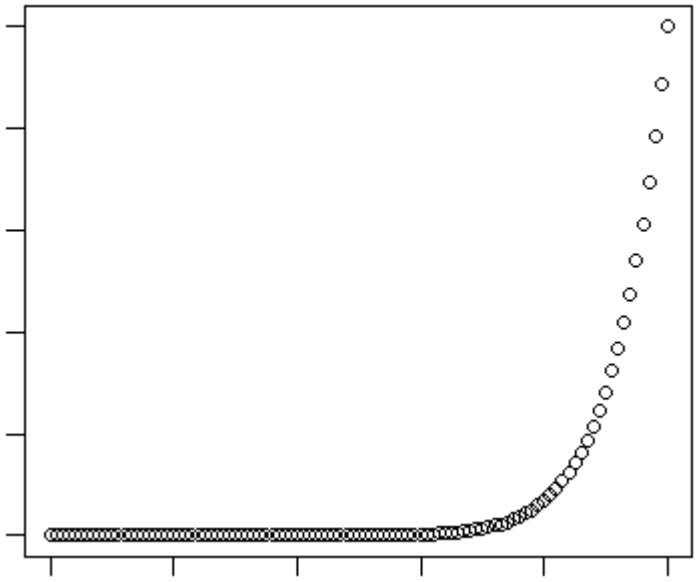
# Jak poznat závislost z grafu?

- Zajímá.li nás lineární závislost



# Co s nelineárními závislostmi?

- Mějme následující data



- Lineární korelace by vycházela kolem 0.6, ale je tam evidentní závislost. Co s tím?

# Spearmanův koeficient korelace

- Spearmanův korelační koeficient se hodí pŕo monotónní (klesající/rostoucí) funkce
- Pracuje s pořadím na rozdíl od skutečných hodnot
- Platí, že pokud je vysoký pearson, je vysoký i spearman (ale obráceně to nemusí platit)

# Výpočet Spearmanova kor. koeficentu

- Spočítáme pořadí hodnot X a Y vzhledem k ostatním (R(X) a R(Y)). Tedy u hodnot Y je nejmenší číslo 11, dostane tedy pořadí 1, atd. V případě rovnosti počítáme průměr pořadí (proto mají některé prvky hodnotu pořadí 1,5)
- Dosadíme do vzorce:

$$r_{sp} = 1 - \frac{6\sum_i (R(x_i) - R(y_i))^2}{n(n^2-1)}$$

- V našem případě $n = 10$

$$\sum_i (R(x_i) - R(y_i))^2 = 84.95$$

- Tedy $r_{sp} = 1 - \frac{6 \cdot 84.95}{10(100-1)} = 0.49$

| $X$ | $Y$ | $R(X)$ | $R(Y)$ | $\lvert R(X) - R(Y) \rvert$ | $\left(R(X) - R(Y)\right)^2$ |
|---|---|---|---|---|---|
| 3 | 12 | 3.5 | 2 | 1.5 | 2.25 |
| 11 | 34 | 5 | 9 | 4 | 16 |
| 2 | 11 | 1.5 | 1 | 0.5 | 2.5 |
| 2 | 16 | 1.5 | 4 | 2.5 | 6.2 |
| 16 | 18 | 7 | 5 | 2 | 4 |
| 8 | 27 | 4 | 8 | 4 | 16 |
| 13 | 25 | 6 | 6 | 0 | 0 |
| 3 | 26 | 3.5 | 7 | 3.5 | 12 |
| 19 | 39 | 9 | 10 | 1 | 1 |
| 17 | 13 | 8 | 3 | 5 | 25 |

# Vícenásobná regrese

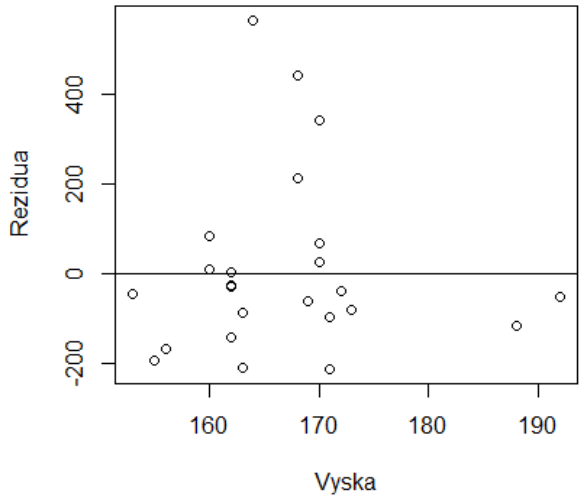
- Můžeme chtít i závislost na více parametrech zároveň (např výkon v testu může záviset na inteligenci a na míře stresu)
- Zapisujeme stejně

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots b_nX_n$$

- A jsou na to nástroje, které zjistí hodnoty parametrů $b_0, b_1, \ldots b_n$
- Křivek vysvětlující variabilitu může být hodně, jak zjistit tu, která vypovídá o datech nejlépe?

# Zobrazení reziduí

- Chceme-li uvažovat, zda použít daný model, můžeme udělat několik kontrol, většina z nich operuje s $\varepsilon_i$
- Nejjednodušší kontrola: podívám se na graf reziduí
- Pro náš příklad s výškou:
- Rezidua by „měla být kolem 0“ (velmi neformálně)



# Analýza dvojice kvalitativní-kvantitavní

- Lineární regrese lze použít i pro dvojici kvalitativní-kvantitativní data
- Je-li proměnná X alternativní, stačí ji překódovat jako 0/1 a tuto hodnotu použít v regresi