Základy biostatistiky

(MD710P09)
ak. rok 2007/2008

Karel Zvára

karel.zvara@mff.cuni.cz
http://www.karlin.mff.cuni.cz/∼zvara
katedra pravděpodobnosti a matematické statistiky MFF UK

(naposledy upraveno 16. dubna 2008)

vyšetřování závislosti

nezávisle proměnná(é)	závisle proměnná	
	spojitá	**nominální**
spojitá	regrese korelace	(*logistická regrese*)
nominální	analýza rozptylu	kontingenční tabulky

příklady:

- hmotnost na výšce
- rakovina plic na počtu vykouřených cigaret
- hmotnost obilky na živném roztoku
- barva očí a barva vlasů

vyšetřování závislosti

nezávisle proměnná(é)	závisle proměnná	
	spojitá	**nominální**
spojitá	regrese korelace	(*logistická regrese*)
nominální	analýza rozptylu	kontingenční tabulky

příklady:

- hmotnost na výšce
- rakovina plic na počtu vykouřených cigaret
- hmotnost obilky na živném roztoku
- barva očí a barva vlasů

vyšetřování závislosti

nezávisle proměnná(é)	závisle proměnná	
	spojitá	**nominální**
spojitá	regrese korelace	(*logistická regrese*)
nominální	analýza rozptylu	kontingenční tabulky

příklady:

- hmotnost na výšce
- rakovina plic na počtu vykouřených cigaret
- hmotnost obilky na živném roztoku
- barva očí a barva vlasů

vyšetřování závislosti

nezávisle proměnná(é)	závisle proměnná	
	spojitá	**nominální**
spojitá	regrese korelace	(*logistická regrese*)
nominální	analýza rozptylu	kontingenční tabulky

příklady:

- hmotnost na výšce
- rakovina plic na počtu vykouřených cigaret
- hmotnost obilky na živném roztoku
- barva očí a barva vlasů

korelace a regrese
[correlation, regression]

- **korelace** (dvojice náhodných veličin)
  - měří **sílu** (těsnost) **vzájemné** závislosti **spojitých** veličin
  - lze použít k **prokazování** existence **vzájemné** závislosti X, Y
  - k **porovnávání síly** (těsnosti) závislosti v několika populacích
  - **symetrická** vlastnost veličin X a Y
- **regrese** (náhodná veličina na nenáhodné veličině)
  - udává **jak** závisí střední hodnota **spojité** veličiny Y na nezávisle proměnné (proměnných) x
  - **nesymetrická** vlastnost (závislost Y na $x \neq$ závislost X na y)
  - lze použít k **prokazování** existence závislosti **závisle** proměnné Y na **nezávisle** proměnné x
  - umožňuje **předpovídat** stř. hodnotu Y pro zvolenou hodnotu x

korelace a regrese
[correlation, regression]

- **korelace** (dvojice náhodných veličin)
  - měří **sílu** (těsnost) **vzájemné** závislosti **spojitých** veličin
  - lze použít k **prokazování** existence **vzájemné** závislosti X, Y
  - k **porovnávání síly** (těsnosti) závislosti v několika populacích
  - **symetrická** vlastnost veličin X a Y
- **regrese** (náhodná veličina na nenáhodné veličině)
  - udává **jak** závisí střední hodnota **spojité** veličiny Y na nezávisle proměnné (proměnných) x
  - **nesymetrická** vlastnost (závislost Y na $x \neq$ závislost X na y)
  - lze použít k **prokazování** existence závislosti **závisle** proměnné Y na **nezávisle** proměnné x
  - umožňuje **předpovídat** stř. hodnotu Y pro zvolenou hodnotu x

korelace a regrese
[correlation, regression]

- **korelace** (dvojice náhodných veličin)
  - měří **sílu** (těsnost) **vzájemné** závislosti **spojitých** veličin
  - lze použít k **prokazování** existence **vzájemné** závislosti X, Y
  - k **porovnávání síly** (těsnosti) závislosti v několika populacích
  - **symetrická** vlastnost veličin X a Y
- **regrese** (náhodná veličina na nenáhodné veličině)
  - udává **jak** závisí střední hodnota **spojité** veličiny Y na nezávisle proměnné (proměnných) x
  - **nesymetrická** vlastnost (závislost Y na $x \neq$ závislost X na y)
  - lze použít k **prokazování** existence závislosti **závisle** proměnné Y na **nezávisle** proměnné x
  - umožňuje **předpovídat** stř. hodnotu Y pro zvolenou hodnotu x

korelace a regrese
[correlation, regression]

- **korelace** (dvojice náhodných veličin)
  - měří **sílu** (těsnost) **vzájemné** závislosti **spojitých** veličin
  - lze použít k **prokazování** existence **vzájemné** závislosti X, Y
  - k **porovnávání síly** (těsnosti) závislosti v několika populacích
  - **symetrická** vlastnost veličin X a Y
- **regrese** (náhodná veličina na nenáhodné veličině)
  - udává **jak** závisí střední hodnota **spojité** veličiny Y na nezávisle proměnné (proměnných) x
  - **nesymetrická** vlastnost (závislost Y na $x \neq$ závislost X na y)
  - lze použít k **prokazování** existence závislosti **závisle** proměnné Y na **nezávisle** proměnné x
  - umožňuje **předpovídat** stř. hodnotu Y pro zvolenou hodnotu x

korelace a regrese
[correlation, regression]

- **korelace** (dvojice náhodných veličin)
  - měří **sílu** (těsnost) **vzájemné** závislosti **spojitých** veličin
  - lze použít k **prokazování** existence **vzájemné** závislosti X, Y
  - k **porovnávání síly** (těsnosti) závislosti v několika populacích
  - **symetrická** vlastnost veličin X a Y
- **regrese** (náhodná veličina na nenáhodné veličině)
  - udává **jak** závisí střední hodnota **spojité** veličiny Y na nezávisle proměnné (proměnných) x
  - **nesymetrická** vlastnost (závislost Y na $x \neq$ závislost X na y)
  - lze použít k **prokazování** existence závislosti **závisle** proměnné Y na **nezávisle** proměnné x
  - umožňuje **předpovídat** stř. hodnotu Y pro zvolenou hodnotu x

korelace a regrese
[correlation, regression]

- **korelace** (dvojice náhodných veličin)
  - měří **sílu** (těsnost) **vzájemné** závislosti **spojitých** veličin
  - lze použít k **prokazování** existence **vzájemné** závislosti X, Y
  - k **porovnávání síly** (těsnosti) závislosti v několika populacích
  - **symetrická** vlastnost veličin X a Y
- **regrese** (náhodná veličina na nenáhodné veličině)
  - udává **jak** závisí střední hodnota **spojité** veličiny Y na nezávisle proměnné (proměnných) x
  - **nesymetrická** vlastnost (závislost Y na $x \neq$ závislost X na y)
  - lze použít k **prokazování** existence závislosti **závisle** proměnné Y na **nezávisle** proměnné x
  - umožňuje **předpovídat** stř. hodnotu Y pro zvolenou hodnotu x

korelace a regrese
[correlation, regression]

- **korelace** (dvojice náhodných veličin)
  - měří **sílu** (těsnost) **vzájemné** závislosti **spojitých** veličin
  - lze použít k **prokazování** existence **vzájemné** závislosti X, Y
  - k **porovnávání síly** (těsnosti) závislosti v několika populacích
  - **symetrická** vlastnost veličin X a Y
- **regrese** (náhodná veličina na nenáhodné veličině)
  - udává **jak** závisí střední hodnota **spojité** veličiny Y na nezávisle proměnné (proměnných) x
  - **nesymetrická** vlastnost (závislost Y na $x \neq$ závislost X na y)
  - lze použít k **prokazování** existence závislosti **závisle** proměnné Y na **nezávisle** proměnné x
  - umožňuje **předpovídat** stř. hodnotu Y pro zvolenou hodnotu x

korelace a regrese
[correlation, regression]

- **korelace** (dvojice náhodných veličin)
  - měří **sílu** (těsnost) **vzájemné** závislosti **spojitých** veličin
  - lze použít k **prokazování** existence **vzájemné** závislosti X, Y
  - k **porovnávání síly** (těsnosti) závislosti v několika populacích
  - **symetrická** vlastnost veličin X a Y
- **regrese** (náhodná veličina na nenáhodné veličině)
  - udává **jak** závisí střední hodnota **spojité** veličiny Y na nezávisle proměnné (proměnných) x
  - **nesymetrická** vlastnost (závislost Y na $x \neq$ závislost X na y)
  - lze použít k **prokazování** existence závislosti **závisle** proměnné Y na **nezávisle** proměnné x
  - umožňuje **předpovídat** stř. hodnotu Y pro zvolenou hodnotu x

korelace a regrese
[correlation, regression]

- **korelace** (dvojice náhodných veličin)
  - měří **sílu** (těsnost) **vzájemné** závislosti **spojitých** veličin
  - lze použít k **prokazování** existence **vzájemné** závislosti X, Y
  - k **porovnávání síly** (těsnosti) závislosti v několika populacích
  - **symetrická** vlastnost veličin X a Y
- **regrese** (náhodná veličina na nenáhodné veličině)
  - udává **jak** závisí střední hodnota **spojité** veličiny Y na nezávisle proměnné (proměnných) x
  - **nesymetrická** vlastnost (závislost Y na $x \neq$ závislost X na y)
  - lze použít k **prokazování** existence závislosti **závisle** proměnné Y na **nezávisle** proměnné x
  - umožňuje **předpovídat** stř. hodnotu Y pro zvolenou hodnotu x

korelace a regrese
[correlation, regression]

- **korelace** (dvojice náhodných veličin)
  - měří **sílu** (těsnost) **vzájemné** závislosti **spojitých** veličin
  - lze použít k **prokazování** existence **vzájemné** závislosti X, Y
  - k **porovnávání síly** (těsnosti) závislosti v několika populacích
  - **symetrická** vlastnost veličin X a Y
- **regrese** (náhodná veličina na nenáhodné veličině)
  - udává **jak** závisí střední hodnota **spojité** veličiny Y na nezávisle proměnné (proměnných) x
  - **nesymetrická** vlastnost (závislost Y na $x \neq$ závislost X na y)
  - lze použít k **prokazování** existence závislosti **závisle** proměnné Y na **nezávisle** proměnné x
  - umožňuje **předpovídat** stř. hodnotu Y pro zvolenou hodnotu x

korelační koeficient
(zavedení **výběrového** korelačního koeficientu)

- (populační) korelační koeficient $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$ (zaveden na obr. 73)
  - $|\rho_{XY}| \leq 1$
  - pro nezávislé X, Y je $\rho_{XY} = 0$
  - měří sílu **lineární** závislosti
- (výběrový) korelační koeficient r_{xy} (zaveden na obr. 33)

$$r_{XY} = \frac{s_{xy}}{s_x s_y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

  - odhaduje ρ_{XY}
  - přesnost odhadu závisí na n
- alternativní označení: **Pearsonův** korelační koeficient, momentový korelační koeficient, [correlation coefficient]

korelační koeficient
(zavedení **výběrového** korelačního koeficientu)

- (populační) korelační koeficient $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$ (zaveden na obr. 73)
  - $|\rho_{XY}| \leq 1$
  - pro nezávislé X, Y je $\rho_{XY} = 0$
  - měří sílu **lineární** závislosti
- (výběrový) korelační koeficient r_{xy} (zaveden na obr. 33)

$$r_{XY} = \frac{s_{xy}}{s_x s_y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

  - odhaduje ρ_{XY}
  - přesnost odhadu závisí na n
- alternativní označení: **Pearsonův** korelační koeficient, momentový korelační koeficient, [correlation coefficient]

korelační koeficient
(zavedení **výběrového** korelačního koeficientu)

- (populační) korelační koeficient $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$ (zaveden na obr. 73)
  - $|\rho_{XY}| \leq 1$
  - pro nezávislé X, Y je $\rho_{XY} = 0$
  - měří sílu **lineární** závislosti
- (výběrový) korelační koeficient r_{xy} (zaveden na obr. 33)

$$r_{XY} = \frac{s_{xy}}{s_x s_y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

  - odhaduje ρ_{XY}
  - přesnost odhadu závisí na n
- alternativní označení: **Pearsonův** korelační koeficient, momentový korelační koeficient, [correlation coefficient]

korelační koeficient
(zavedení **výběrového** korelačního koeficientu)

- (populační) korelační koeficient $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$ (zaveden na obr. 73)
  - $|\rho_{XY}| \leq 1$
  - pro nezávislé X, Y je $\rho_{XY} = 0$
  - měří sílu **lineární** závislosti
- (výběrový) korelační koeficient r_{xy} (zaveden na obr. 33)

$$r_{XY} = \frac{s_{xy}}{s_x s_y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

  - odhaduje ρ_{XY}
  - přesnost odhadu závisí na n
- alternativní označení: **Pearsonův** korelační koeficient, momentový korelační koeficient, [correlation coefficient]

korelační koeficient
(zavedení **výběrového** korelačního koeficientu)

- (populační) korelační koeficient $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$ (zaveden na obr. 73)
  - $|\rho_{XY}| \leq 1$
  - pro nezávislé X, Y je $\rho_{XY} = 0$
  - měří sílu **lineární** závislosti
- (výběrový) korelační koeficient r_{xy} (zaveden na obr. 33)

$$r_{XY} = \frac{s_{xy}}{s_x s_y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

  - odhaduje ρ_{XY}
  - přesnost odhadu závisí na n
- alternativní označení: **Pearsonův** korelační koeficient, momentový korelační koeficient, [correlation coefficient]

korelační koeficient

(zavedení **výběrového** korelačního koeficientu)

- (populační) korelační koeficient $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$ (zaveden na obr. 73)
  - $|\rho_{XY}| \leq 1$
  - pro nezávislé X, Y je $\rho_{XY} = 0$
  - měří sílu **lineární** závislosti
- (výběrový) korelační koeficient r_{xy} (zaveden na obr. 33)

$$r_{XY} = \frac{s_{xy}}{s_x s_y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

  - odhaduje ρ_{XY}
  - přesnost odhadu závisí na n
- alternativní označení: **Pearsonův** korelační koeficient, momentový korelační koeficient, [correlation coefficient]

korelační koeficient

(zavedení **výběrového** korelačního koeficientu)

- (populační) korelační koeficient $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$ (zaveden na obr. 73)
  - $|\rho_{XY}| \leq 1$
  - pro nezávislé X, Y je $\rho_{XY} = 0$
  - měří sílu **lineární** závislosti
- (výběrový) korelační koeficient r_{xy} (zaveden na obr. 33)

$$r_{XY} = \frac{s_{xy}}{s_x s_y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

  - odhaduje ρ_{XY}
  - přesnost odhadu závisí na n
- alternativní označení: **Pearsonův** korelační koeficient, momentový korelační koeficient, [correlation coefficient]

korelační koeficient

(zavedení **výběrového** korelačního koeficientu)

- (populační) korelační koeficient $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$ (zaveden na obr. 73)
  - $|\rho_{XY}| \leq 1$
  - pro nezávislé X, Y je $\rho_{XY} = 0$
  - měří sílu **lineární** závislosti
- (výběrový) korelační koeficient r_{xy} (zaveden na obr. 33)

$$r_{XY} = \frac{s_{xy}}{s_x s_y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

  - odhaduje ρ_{XY}
  - přesnost odhadu závisí na n
- alternativní označení: **Pearsonův** korelační koeficient, momentový korelační koeficient, [correlation coefficient]

dokazování závislosti X, Y

- k prokázání závislosti nutno **normální** rozdělení (X, Y)
- $H_0 : \rho_{XY} = 0$ se na hladině α zamítá:

$$T = \frac{r}{\sqrt{1 - r^2}} \sqrt{n - 2}, \qquad |T| \geq t_{n-2}(\alpha)$$

- **Spearmanův** korelační koeficient
  - měří sílu **monotonní** závislosti
  - založen na **pořadích** R_i, Q_i hodnot X_i, Y_i

$$r_{XY}^{(S)} = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^{n} (R_i - Q_i)^2$$

  - k testu nezávislosti nepotřebuje normální rozdělení
  - H_0: (nezávislost) se zamítá, je-li $|r_{XY}^{(S)} \sqrt{n - 1}| \geq z(\alpha/2)$

dokazování závislosti X, Y

- k prokázání závislosti nutno **normální** rozdělení (X, Y)
- $H_0 : \rho_{XY} = 0$ se na hladině α zamítá:

$$T = \frac{r}{\sqrt{1 - r^2}}\sqrt{n-2}, \qquad |T| \geq t_{n-2}(\alpha)$$

- **Spearmanův** korelační koeficient
  - měří sílu **monotonní** závislosti
  - založen na **pořadích** R_i, Q_i hodnot X_i, Y_i

$$r_{XY}^{(S)} = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(R_i - Q_i)^2$$

  - k testu nezávislosti nepotřebuje normální rozdělení
  - H_0: (nezávislost) se zamítá, je-li $|r_{XY}^{(S)}\sqrt{n-1}| \geq z(\alpha/2)$

dokazování závislosti X, Y

- k prokázání závislosti nutno **normální** rozdělení (X, Y)
- $H_0 : \rho_{XY} = 0$ se na hladině α zamítá:

$$T = \frac{r}{\sqrt{1 - r^2}}\sqrt{n-2}, \qquad |T| \geq t_{n-2}(\alpha)$$

- **Spearmanův** korelační koeficient
  - měří sílu **monotonní** závislosti
  - založen na **pořadích** R_i, Q_i hodnot X_i, Y_i

$$r_{XY}^{(S)} = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(R_i - Q_i)^2$$

  - k testu nezávislosti nepotřebuje normální rozdělení
  - H_0: (nezávislost) se zamítá, je-li $|r_{XY}^{(S)}\sqrt{n-1}| \geq z(\alpha/2)$

dokazování závislosti X, Y

- k prokázání závislosti nutno **normální** rozdělení (X, Y)
- $H_0 : \rho_{XY} = 0$ se na hladině α zamítá:

$$T = \frac{r}{\sqrt{1 - r^2}}\sqrt{n-2}, \qquad |T| \geq t_{n-2}(\alpha)$$

- **Spearmanův** korelační koeficient
  - měří sílu **monotonní** závislosti
  - založen na **pořadích** R_i, Q_i hodnot X_i, Y_i

$$r_{XY}^{(S)} = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(R_i - Q_i)^2$$

  - k testu nezávislosti nepotřebuje normální rozdělení
  - H_0: (nezávislost) se zamítá, je-li $|r_{XY}^{(S)}\sqrt{n-1}| \geq z(\alpha/2)$

dokazování závislosti X, Y

- k prokázání závislosti nutno **normální** rozdělení (X, Y)
- $H_0 : \rho_{XY} = 0$ se na hladině α zamítá:

$$T = \frac{r}{\sqrt{1 - r^2}}\sqrt{n-2}, \qquad |T| \geq t_{n-2}(\alpha)$$

- **Spearmanův** korelační koeficient
  - měří sílu **monotonní** závislosti
  - založen na **pořadích** R_i, Q_i hodnot X_i, Y_i

$$r_{XY}^{(S)} = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(R_i - Q_i)^2$$

  - k testu nezávislosti nepotřebuje normální rozdělení
  - H_0: (nezávislost) se zamítá, je-li $|r_{XY}^{(S)}\sqrt{n-1}| \geq z(\alpha/2)$

dokazování závislosti X, Y

- k prokázání závislosti nutno **normální** rozdělení (X, Y)
- $H_0 : \rho_{XY} = 0$ se na hladině α zamítá:

$$T = \frac{r}{\sqrt{1 - r^2}}\sqrt{n-2}, \qquad |T| \geq t_{n-2}(\alpha)$$

- **Spearmanův** korelační koeficient
  - měří sílu **monotonní** závislosti
  - založen na **pořadích** R_i, Q_i hodnot X_i, Y_i

$$r_{XY}^{(S)} = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(R_i - Q_i)^2$$

  - k testu nezávislosti nepotřebuje normální rozdělení
  - H_0: (nezávislost) se zamítá, je-li $|r_{XY}^{(S)}\sqrt{n-1}| \geq z(\alpha/2)$

dokazování závislosti X, Y

- k prokázání závislosti nutno **normální** rozdělení (X, Y)
- $H_0 : \rho_{XY} = 0$ se na hladině α zamítá:

$$T = \frac{r}{\sqrt{1 - r^2}}\sqrt{n-2}, \qquad |T| \geq t_{n-2}(\alpha)$$

- **Spearmanův** korelační koeficient
  - měří sílu **monotonní** závislosti
  - založen na **pořadích** R_i, Q_i hodnot X_i, Y_i

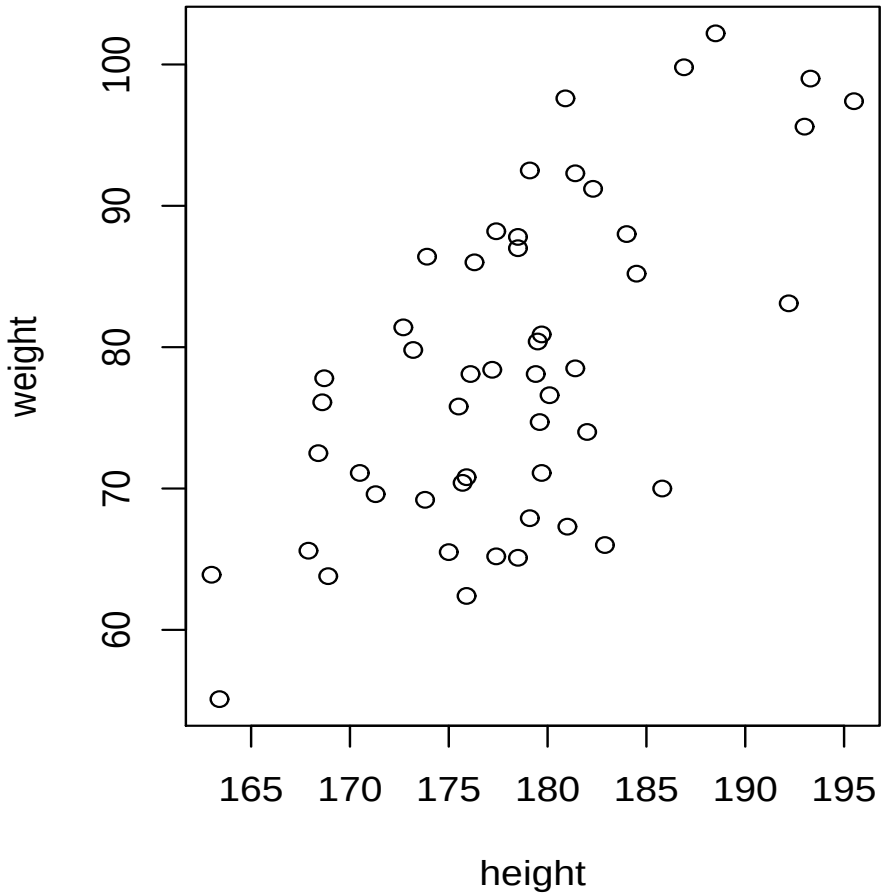
$$r_{XY}^{(S)} = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(R_i - Q_i)^2$$

  - k testu nezávislosti nepotřebuje normální rozdělení
  - H_0: (nezávislost) se zamítá, je-li $|r_{XY}^{(S)}\sqrt{n-1}| \geq z(\alpha/2)$

závislost váhy na výšce u mužů

data: Policie

[plot(weight∼height)]



[cor.test(weight,height)]

$$r = 0{,}648$$

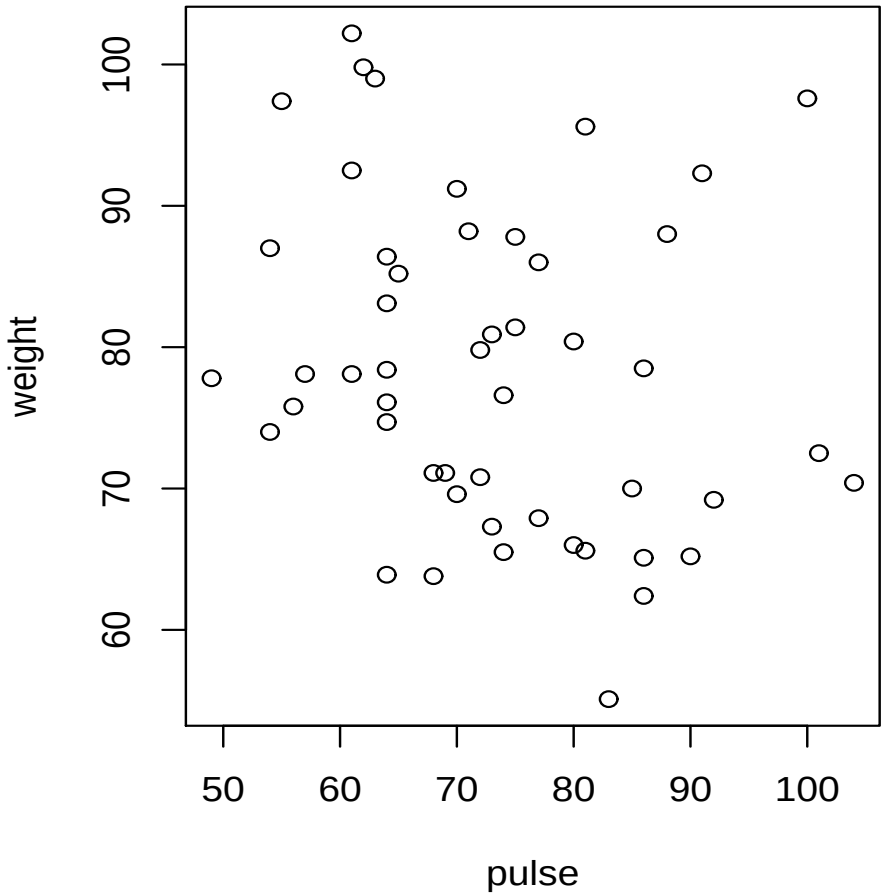
$$t = 5{,}814$$

$$p < 0{,}001$$

závislost váhy na pulsu u mužů

data: Policie

[plot(weight∼pulse]



[cor.test(pulse,weight)]

$$r = -0{,}245$$
$$t = -1{,}752$$
$$p = 8{,}6\ \%$$

Fisherova z-transformace

(přiblíží chování výběrového korelačního koeficientu r normálnímu rozdělení)

$$Z = \frac{1}{2} \ln \frac{1+r}{1-r} \dot{\sim} \mathrm{N}\left(\frac{1}{2} \ln \frac{1+\rho}{1-\rho}, \frac{1}{n-3} \right)$$

test shody dvou nezávisle odhadovaných korel. koeficientů

příklad **Kojeni**: výška rodičů chlapců a dívek

- dívky: $r_1 = 0{,}279$, $n_1 = 50$, $z_1 = \frac{1}{2} \ln \frac{1+0{,}5687}{1-0{,}5687} = 0{,}286$
- hoši: $r_2 = 0{,}150$, $n_2 = 49$, $z_2 = \frac{1}{2} \ln \frac{1+0{,}150}{1-0{,}150} = 0{,}151$
- test $\mathrm{H}_0 : \rho_1 = \rho_2$ (odhady r_1, r_2 jsou **nezávislé!**)

$$z = \frac{0{,}286 - 0{,}151}{\sqrt{\frac{1}{50-3} + \frac{1}{49-3}}} = 0{,}671.$$

srovnej s kritickou hodnotou $z(0{,}05/2) = 1{,}960$, $p = 50{,}2\ \%$

Fisherova z-transformace

(přiblíží chování výběrového korelačního koeficientu r normálnímu rozdělení)

$$Z = \frac{1}{2} \ln \frac{1+r}{1-r} \dot{\sim} \mathrm{N}\left(\frac{1}{2} \ln \frac{1+\rho}{1-\rho}, \frac{1}{n-3} \right)$$

test shody dvou nezávisle odhadovaných korel. koeficientů

příklad **Kojeni**: výška rodičů chlapců a dívek

- dívky: $r_1 = 0{,}279$, $n_1 = 50$, $z_1 = \frac{1}{2} \ln \frac{1+0{,}5687}{1-0{,}5687} = 0{,}286$
- hoši: $r_2 = 0{,}150$, $n_2 = 49$, $z_2 = \frac{1}{2} \ln \frac{1+0{,}150}{1-0{,}150} = 0{,}151$
- test $\mathrm{H}_0 : \rho_1 = \rho_2$ (odhady r_1, r_2 jsou **nezávislé!**)

$$z = \frac{0{,}286 - 0{,}151}{\sqrt{\frac{1}{50-3} + \frac{1}{49-3}}} = 0{,}671.$$

srovnej s kritickou hodnotou $z(0{,}05/2) = 1{,}960$, $p = 50{,}2\ \%$

Fisherova z-transformace

(přiblíží chování výběrového korelačního koeficientu r normálnímu rozdělení)

$$Z = \frac{1}{2} \ln \frac{1+r}{1-r} \dot{\sim} \mathrm{N}\left(\frac{1}{2} \ln \frac{1+\rho}{1-\rho}, \frac{1}{n-3} \right)$$

test shody dvou nezávisle odhadovaných korel. koeficientů

příklad **Kojeni**: výška rodičů chlapců a dívek

- dívky: $r_1 = 0{,}279$, $n_1 = 50$, $z_1 = \frac{1}{2} \ln \frac{1+0{,}5687}{1-0{,}5687} = 0{,}286$
- hoši: $r_2 = 0{,}150$, $n_2 = 49$, $z_2 = \frac{1}{2} \ln \frac{1+0{,}150}{1-0{,}150} = 0{,}151$
- test $\mathrm{H}_0 : \rho_1 = \rho_2$ (odhady r_1, r_2 jsou **nezávislé!**)

$$z = \frac{0{,}286 - 0{,}151}{\sqrt{\frac{1}{50-3} + \frac{1}{49-3}}} = 0{,}671.$$

srovnej s kritickou hodnotou $z(0{,}05/2) = 1{,}960$, $p = 50{,}2\ \%$

interval spolehlivosti pro ρ

opět potřebujeme normální rozdělení (X, Y)

- ve dvou krocích:
  - interval spolehlivosti pro $\zeta = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$
  - pomocí inverzní transformace pak int. spol. pro ρ
- interval spolehlivosti součástí funkce cor.test()
- náš příklad:

skupina	r (bodový odhad)	95% int. spol. pro ρ	p
dívky	0,279	$(0{,}000; 0{,}517)$	5,01 %
hoši	0,150	$(-0{,}137; 0{,}414)$	30,3 %

- u chlapců nelze prokázat na 5% hladině závislost
- u děvčat je závislost na 10% hladině průkazná, na 5% hladině těsně nikoliv

interval spolehlivosti pro ρ

opět potřebujeme normální rozdělení (X, Y)

- ve dvou krocích:
  - interval spolehlivosti pro $\zeta = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$
  - pomocí inverzní transformace pak int. spol. pro ρ
- interval spolehlivosti součástí funkce cor.test()
- náš příklad:

skupina	r (bodový odhad)	95% int. spol. pro ρ	p
dívky	0,279	$(0{,}000; 0{,}517)$	5,01 %
hoši	0,150	$(-0{,}137; 0{,}414)$	30,3 %

- u chlapců nelze prokázat na 5% hladině závislost
- u děvčat je závislost na 10% hladině průkazná, na 5% hladině těsně nikoliv

interval spolehlivosti pro ρ

opět potřebujeme normální rozdělení (X, Y)

- ve dvou krocích:
  - interval spolehlivosti pro $\zeta = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$
  - pomocí inverzní transformace pak int. spol. pro ρ
- interval spolehlivosti součástí funkce cor.test()
- náš příklad:

skupina	r (bodový odhad)	95% int. spol. pro ρ	p
dívky	0,279	$(0{,}000; 0{,}517)$	5,01 %
hoši	0,150	$(-0{,}137; 0{,}414)$	30,3 %

- u chlapců nelze prokázat na 5% hladině závislost
- u děvčat je závislost na 10% hladině průkazná, na 5% hladině těsně nikoliv

interval spolehlivosti pro ρ

opět potřebujeme normální rozdělení (X, Y)

- ve dvou krocích:
  - interval spolehlivosti pro $\zeta = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$
  - pomocí inverzní transformace pak int. spol. pro ρ
- interval spolehlivosti součástí funkce cor.test()
- náš příklad:

skupina	r (bodový odhad)	95% int. spol. pro ρ	p
dívky	0,279	$(0{,}000; 0{,}517)$	5,01 %
hoši	0,150	$(-0{,}137; 0{,}414)$	30,3 %

- u chlapců nelze prokázat na 5% hladině závislost
- u děvčat je závislost na 10% hladině průkazná, na 5% hladině těsně nikoliv

interval spolehlivosti pro ρ

opět potřebujeme normální rozdělení (X, Y)

- ve dvou krocích:
  - interval spolehlivosti pro $\zeta = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$
  - pomocí inverzní transformace pak int. spol. pro ρ
- interval spolehlivosti součástí funkce cor.test()
- náš příklad:

skupina	r (bodový odhad)	95% int. spol. pro ρ	p
dívky	0,279	(0,000; 0,517)	5,01 %
hoši	0,150	(−0,137; 0,414)	30,3 %

- u chlapců nelze prokázat na 5% hladině závislost
- u děvčat je závislost na 10% hladině průkazná, na 5% hladině těsně nikoliv

interval spolehlivosti pro ρ

opět potřebujeme normální rozdělení (X, Y)

- ve dvou krocích:
  - interval spolehlivosti pro $\zeta = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$
  - pomocí inverzní transformace pak int. spol. pro ρ
- interval spolehlivosti součástí funkce cor.test()
- náš příklad:

skupina	r (bodový odhad)	95% int. spol. pro ρ	p
dívky	0,279	$(0{,}000; 0{,}517)$	5,01 %
hoši	0,150	$(-0{,}137; 0{,}414)$	30,3 %

- u chlapců nelze prokázat na 5% hladině závislost
- u děvčat je závislost na 10% hladině průkazná, na 5% hladině těsně nikoliv

interval spolehlivosti pro ρ

opět potřebujeme normální rozdělení (X, Y)

- ve dvou krocích:
  - interval spolehlivosti pro $\zeta = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$
  - pomocí inverzní transformace pak int. spol. pro ρ
- interval spolehlivosti součástí funkce cor.test()
- náš příklad:

skupina	r (bodový odhad)	95% int. spol. pro ρ	p
dívky	0,279	$(0{,}000; 0{,}517)$	5,01 %
hoši	0,150	$(-0{,}137; 0{,}414)$	30,3 %

- u chlapců nelze prokázat na 5% hladině závislost
- u děvčat je závislost na 10% hladině průkazná, na 5% hladině těsně nikoliv

regrese
(původ pojmu)

- tendence (návrat) k průměrnosti
  F. Galton (1886) vyšetřoval dědičnost výšky postavy
- uvažujme otce, jejichž výška je rovna průměrné výšce generace **všech** otců; průměrná výška synů otců této výšky bude rovna průměrné výšce **všech** synů
- uvažujme otce o 10 cm **vyšší**, než je průměrná výška generace otců: průměrná výška synů těchto otců bude jen asi o 5 cm **vyšší**, než průměrná výška generace synů
- uvažujme otce o 10 cm **nižší**, než je průměrná výška generace otců: průměrná výška synů těchto otců bude jen o asi 5 cm **nižší**, než průměrná výška generace synů
- průměrné výšky synů nereprodukují celou odchylku výšky otce od průměru, je tu návrat k průměru (regrese)

regrese
(původ pojmu)

- tendence (návrat) k průměrnosti
  F. Galton (1886) vyšetřoval dědičnost výšky postavy
- uvažujme otce, jejichž výška je rovna průměrné výšce generace **všech** otců; průměrná výška synů otců této výšky bude rovna průměrné výšce **všech** synů
- uvažujme otce o 10 cm **vyšší**, než je průměrná výška generace otců: průměrná výška synů těchto otců bude jen asi o 5 cm **vyšší**, než průměrná výška generace synů
- uvažujme otce o 10 cm **nižší**, než je průměrná výška generace otců: průměrná výška synů těchto otců bude jen o asi 5 cm **nižší**, než průměrná výška generace synů
- průměrné výšky synů nereprodukují celou odchylku výšky otce od průměru, je tu návrat k průměru (regrese)

regrese
(původ pojmu)

- tendence (návrat) k průměrnosti
  F. Galton (1886) vyšetřoval dědičnost výšky postavy
- uvažujme otce, jejichž výška je rovna průměrné výšce generace **všech** otců; průměrná výška synů otců této výšky bude rovna průměrné výšce **všech** synů
- uvažujme otce o 10 cm **vyšší**, než je průměrná výška generace otců: průměrná výška synů těchto otců bude jen asi o 5 cm **vyšší**, než průměrná výška generace synů
- uvažujme otce o 10 cm **nižší**, než je průměrná výška generace otců: průměrná výška synů těchto otců bude jen o asi 5 cm **nižší**, než průměrná výška generace synů
- průměrné výšky synů nereprodukují celou odchylku výšky otce od průměru, je tu návrat k průměru (regrese)

regrese
(původ pojmu)

- tendence (návrat) k průměrnosti
  F. Galton (1886) vyšetřoval dědičnost výšky postavy
- uvažujme otce, jejichž výška je rovna průměrné výšce generace **všech** otců; průměrná výška synů otců této výšky bude rovna průměrné výšce **všech** synů
- uvažujme otce o 10 cm **vyšší**, než je průměrná výška generace otců: průměrná výška synů těchto otců bude jen asi o 5 cm **vyšší**, než průměrná výška generace synů
- uvažujme otce o 10 cm **nižší**, než je průměrná výška generace otců: průměrná výška synů těchto otců bude jen o asi 5 cm **nižší**, než průměrná výška generace synů
- průměrné výšky synů nereprodukují celou odchylku výšky otce od průměru, je tu návrat k průměru (regrese)

regrese
(původ pojmu)

- tendence (návrat) k průměrnosti
  F. Galton (1886) vyšetřoval dědičnost výšky postavy
- uvažujme otce, jejichž výška je rovna průměrné výšce generace **všech** otců; průměrná výška synů otců této výšky bude rovna průměrné výšce **všech** synů
- uvažujme otce o 10 cm **vyšší**, než je průměrná výška generace otců: průměrná výška synů těchto otců bude jen asi o 5 cm **vyšší**, než průměrná výška generace synů
- uvažujme otce o 10 cm **nižší**, než je průměrná výška generace otců: průměrná výška synů těchto otců bude jen o asi 5 cm **nižší**, než průměrná výška generace synů
- průměrné výšky synů nereprodukují celou odchylku výšky otce od průměru, je tu návrat k průměru (regrese)

regresní přímka

- odhadovaná závislost střední hodnoty Y na nenáhodné x:

$$\mathrm{E}\, Y = \beta_0 + \beta_1 x$$

- k daným $x_1, \ldots, x_n$ zjistíme $Y_1, \ldots, Y_n$
- předpoklady:
  - **nezávislá** pozorování $Y_1, \ldots, Y_n$
  - **stejný** rozptyl σ^2
  - **normální** rozdělení (potřebné až pro testy)
- neznámé populační parametry β_0, β_1 odhadujeme metodou **nejmenších čtverců**:

  minimalizovat $\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$
- odhady označíme b_0, b_1

regresní přímka

- odhadovaná závislost střední hodnoty Y na nenáhodné x:

$$\mathrm{E}\, Y = \beta_0 + \beta_1 x$$

- k daným $x_1, \ldots, x_n$ zjistíme $Y_1, \ldots, Y_n$
- předpoklady:
  - **nezávislá** pozorování $Y_1, \ldots, Y_n$
  - **stejný** rozptyl σ^2
  - **normální** rozdělení (potřebné až pro testy)
- neznámé populační parametry β_0, β_1 odhadujeme metodou **nejmenších čtverců**:

  minimalizovat $\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$
- odhady označíme b_0, b_1

regresní přímka

- odhadovaná závislost střední hodnoty Y na nenáhodné x:

$$\mathrm{E}\, Y = \beta_0 + \beta_1 x$$

- k daným $x_1, \ldots, x_n$ zjistíme $Y_1, \ldots, Y_n$
- předpoklady:
  - **nezávislá** pozorování $Y_1, \ldots, Y_n$
  - **stejný** rozptyl σ^2
  - **normální** rozdělení (potřebné až pro testy)
- neznámé populační parametry β_0, β_1 odhadujeme metodou **nejmenších čtverců**:

  minimalizovat $\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$
- odhady označíme b_0, b_1

regresní přímka

- odhadovaná závislost střední hodnoty Y na nenáhodné x:

$$\mathrm{E}\, Y = \beta_0 + \beta_1 x$$

- k daným $x_1, \ldots, x_n$ zjistíme $Y_1, \ldots, Y_n$
- předpoklady:
  - **nezávislá** pozorování $Y_1, \ldots, Y_n$
  - **stejný** rozptyl σ^2
  - **normální** rozdělení (potřebné až pro testy)
- neznámé populační parametry β_0, β_1 odhadujeme metodou **nejmenších čtverců**:

  minimalizovat $\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$

- odhady označíme b_0, b_1

regresní přímka

- odhadovaná závislost střední hodnoty Y na nenáhodné x:

$$\mathrm{E}\, Y = \beta_0 + \beta_1 x$$

- k daným $x_1, \ldots, x_n$ zjistíme $Y_1, \ldots, Y_n$
- předpoklady:
  - **nezávislá** pozorování $Y_1, \ldots, Y_n$
  - **stejný** rozptyl σ^2
  - **normální** rozdělení (potřebné až pro testy)
- neznámé populační parametry β_0, β_1 odhadujeme metodou **nejmenších čtverců**:

  minimalizovat $\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$

- odhady označíme b_0, b_1

regresní přímka

- odhadovaná závislost střední hodnoty Y na nenáhodné x:

$$\mathrm{E}\, Y = \beta_0 + \beta_1 x$$

- k daným $x_1, \ldots, x_n$ zjistíme $Y_1, \ldots, Y_n$
- předpoklady:
  - **nezávislá** pozorování $Y_1, \ldots, Y_n$
  - **stejný** rozptyl σ^2
  - **normální** rozdělení (potřebné až pro testy)
- neznámé populační parametry β_0, β_1 odhadujeme metodou **nejmenších čtverců**:

$$\text{minimalizovat} \quad \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

- odhady označíme b_0, b_1

regresní přímka

- odhadovaná závislost střední hodnoty Y na nenáhodné x:

$$\mathrm{E}\, Y = \beta_0 + \beta_1 x$$

- k daným $x_1, \ldots, x_n$ zjistíme $Y_1, \ldots, Y_n$
- předpoklady:
  - **nezávislá** pozorování $Y_1, \ldots, Y_n$
  - **stejný** rozptyl σ^2
  - **normální** rozdělení (potřebné až pro testy)
- neznámé populační parametry β_0, β_1 odhadujeme metodou **nejmenších čtverců**:

  minimalizovat $\displaystyle\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$

- odhady označíme b_0, b_1

regresní přímka

- odhadovaná závislost střední hodnoty Y na nenáhodné x:

$$\mathrm{E}\, Y = \beta_0 + \beta_1 x$$

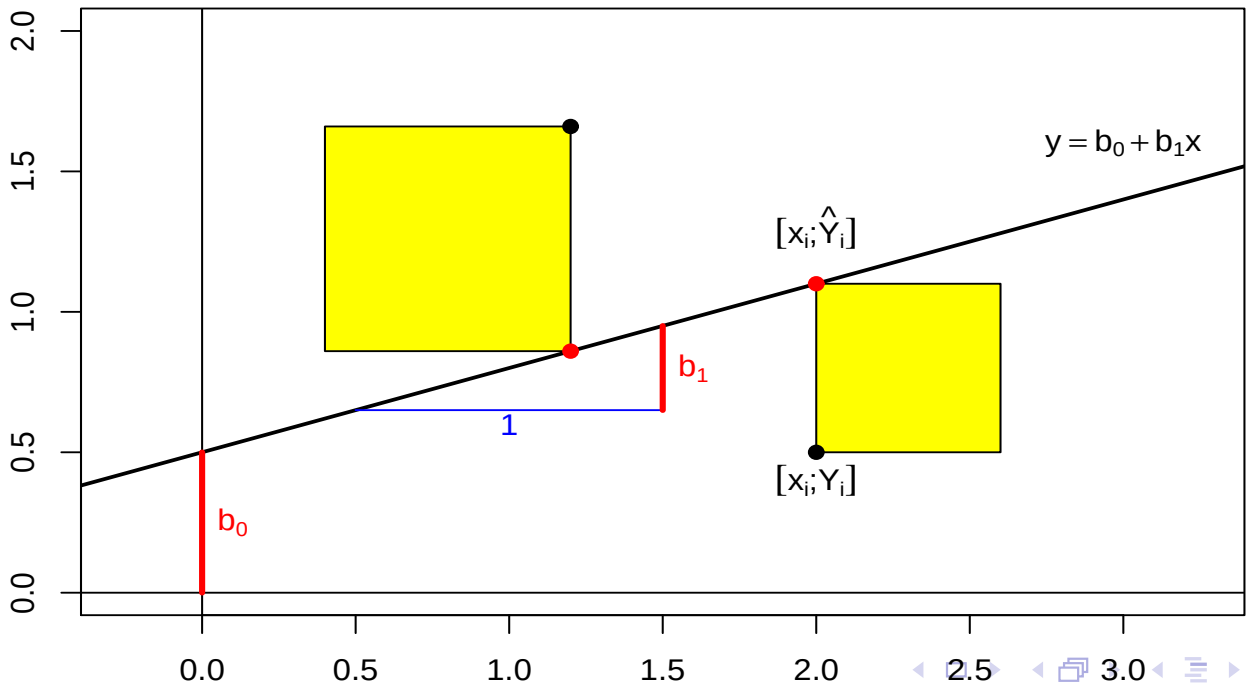
- k daným $x_1, \ldots, x_n$ zjistíme $Y_1, \ldots, Y_n$
- předpoklady:
  - **nezávislá** pozorování $Y_1, \ldots, Y_n$
  - **stejný** rozptyl σ^2
  - **normální** rozdělení (potřebné až pro testy)
- neznámé populační parametry β_0, β_1 odhadujeme metodou **nejmenších čtverců**:

  minimalizovat $\displaystyle\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$

- odhady označíme b_0, b_1

metoda nejmenších čtverců

odhadovaná závislost:	$y = \beta_0 + \beta_1 \cdot x$	(populace)
odhad závislosti:	$y = b_0 + b_1 \cdot x$	(výběr)
i-tá vyrovnaná hodnota	$\hat{Y}_i = b_0 + b_1 x_i$	(výběr)
i-té reziduum	$U_i = Y_i - \hat{Y}_i$	(výběr)
celková plocha čtverců:	$S_e = \sum_{i=1}^{n} U_i^2$	(výběr)



- b_1 – odhad směrnice β_1
- b_1 – odhad změny střední hodnoty závisle proměnné Y při **jednotkové změně** nezávisle proměnné x
- i-té reziduum $U_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 x_i)$
- $Y_i = \hat{Y}_i + U_i$
- (vysvětlováno)=(vysvětleno závislostí)+(nevysvětleno)
- **reziduální součet čtverců** (nevysvětlená variabilita):

$$S_e = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 x_i)^2 = \sum_{i=1}^{n} U_i^2$$

- **reziduální rozptyl**

$$S^2 = \frac{S_e}{n-2}$$

- b_1 – odhad směrnice β_1
- b_1 – odhad změny střední hodnoty závisle proměnné Y při **jednotkové změně** nezávisle proměnné x
- i-té reziduum $U_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 x_i)$
- $Y_i = \hat{Y}_i + U_i$
- (vysvětlováno)=(vysvětleno závislostí)+(nevysvětleno)
- **reziduální součet čtverců** (nevysvětlená variabilita):

$$S_e = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 x_i)^2 = \sum_{i=1}^{n} U_i^2$$

- **reziduální rozptyl**

$$S^2 = \frac{S_e}{n-2}$$

- b_1 – odhad směrnice β_1
- b_1 – odhad změny střední hodnoty závisle proměnné Y při **jednotkové změně** nezávisle proměnné x
- i-té reziduum $U_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 x_i)$
- $Y_i = \hat{Y}_i + U_i$
- (vysvětlováno)=(vysvětleno závislostí)+(nevysvětleno)
- **reziduální součet čtverců** (nevysvětlená variabilita):

$$S_e = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 x_i)^2 = \sum_{i=1}^{n} U_i^2$$

- **reziduální rozptyl**

$$S^2 = \frac{S_e}{n-2}$$

- b_1 – odhad směrnice β_1
- b_1 – odhad změny střední hodnoty závisle proměnné Y při **jednotkové změně** nezávisle proměnné x
- i-té reziduum $U_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 x_i)$
- $Y_i = \hat{Y}_i + U_i$
- (vysvětlováno)=(vysvětleno závislostí)+(nevysvětleno)
- **reziduální součet čtverců** (nevysvětlená variabilita):

$$S_e = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 x_i)^2 = \sum_{i=1}^{n} U_i^2$$

- **reziduální rozptyl**

$$S^2 = \frac{S_e}{n-2}$$

- b_1 – odhad směrnice β_1
- b_1 – odhad změny střední hodnoty závisle proměnné Y při **jednotkové změně** nezávisle proměnné x
- i-té reziduum $U_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 x_i)$
- $Y_i = \hat{Y}_i + U_i$
- (vysvětlováno)=(vysvětleno závislostí)+(nevysvětleno)
- **reziduální součet čtverců** (nevysvětlená variabilita):

$$S_e = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 x_i)^2 = \sum_{i=1}^{n} U_i^2$$

- **reziduální rozptyl**

$$S^2 = \frac{S_e}{n-2}$$

- b_1 – odhad směrnice β_1
- b_1 – odhad změny střední hodnoty závisle proměnné Y při **jednotkové změně** nezávisle proměnné x
- i-té reziduum $U_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 x_i)$
- $Y_i = \hat{Y}_i + U_i$
- (vysvětlováno)=(vysvětleno závislostí)+(nevysvětleno)
- **reziduální součet čtverců** (nevysvětlená variabilita):

$$S_e = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 x_i)^2 = \sum_{i=1}^{n} U_i^2$$

- **reziduální rozptyl**

$$S^2 = \frac{S_e}{n-2}$$

- b_1 – odhad směrnice β_1
- b_1 – odhad změny střední hodnoty závisle proměnné Y při **jednotkové změně** nezávisle proměnné x
- i-té reziduum $U_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 x_i)$
- $Y_i = \hat{Y}_i + U_i$
- (vysvětlováno)=(vysvětleno závislostí)+(nevysvětleno)
- **reziduální součet čtverců** (nevysvětlená variabilita):

$$S_e = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 x_i)^2 = \sum_{i=1}^{n} U_i^2$$

- **reziduální rozptyl**

$$S^2 = \frac{S_e}{n-2}$$

alternativní formulace

- uvažovanou závislost lze psát ve tvaru

$$Y_i = \beta_0^* + \beta_1(x_i - \bar{x}) + E_i$$

- β_0^* vyjadřuje střední úroveň vysvětlované proměnné Y při průměrné hodnotě nezávisle proměnné x
- β_1 vyjadřuje citlivost, s jakou reaguje střední hodnota vysvětlované proměnné Y na jednotkovou odchylku nezávisle proměnné x od jejího průměru $\bar{x}$
- E_i vyjadřuje náhodnou složku i-tého pozorování, $E_i \sim \mathrm{N}(0, \sigma^2)$
- odhadem závislosti je (b_1 je stejné jako při klasickém vyjádření)

$$\hat{Y}_i = \bar{Y} + b_1(x_i - \bar{x})$$

alternativní formulace

- uvažovanou závislost lze psát ve tvaru

$$Y_i = \beta_0^* + \beta_1(x_i - \bar{x}) + E_i$$

- β_0^* vyjadřuje střední úroveň vysvětlované proměnné Y při průměrné hodnotě nezávisle proměnné x
- β_1 vyjadřuje citlivost, s jakou reaguje střední hodnota vysvětlované proměnné Y na jednotkovou odchylku nezávisle proměnné x od jejího průměru $\bar{x}$
- E_i vyjadřuje náhodnou složku i-tého pozorování, $E_i \sim \mathrm{N}(0, \sigma^2)$
- odhadem závislosti je (b_1 je stejné jako při klasickém vyjádření)

$$\hat{Y}_i = \bar{Y} + b_1(x_i - \bar{x})$$

alternativní formulace

- uvažovanou závislost lze psát ve tvaru

$$Y_i = \beta_0^* + \beta_1(x_i - \bar{x}) + E_i$$

- β_0^* vyjadřuje střední úroveň vysvětlované proměnné Y při průměrné hodnotě nezávisle proměnné x
- β_1 vyjadřuje citlivost, s jakou reaguje střední hodnota vysvětlované proměnné Y na jednotkovou odchylku nezávisle proměnné x od jejího průměru $\bar{x}$
- E_i vyjadřuje náhodnou složku i-tého pozorování, $E_i \sim \mathrm{N}(0, \sigma^2)$
- odhadem závislosti je (b_1 je stejné jako při klasickém vyjádření)

$$\hat{Y}_i = \bar{Y} + b_1(x_i - \bar{x})$$

alternativní formulace

- uvažovanou závislost lze psát ve tvaru

$$Y_i = \beta_0^* + \beta_1(x_i - \bar{x}) + E_i$$

- β_0^* vyjadřuje střední úroveň vysvětlované proměnné Y při průměrné hodnotě nezávisle proměnné x
- β_1 vyjadřuje citlivost, s jakou reaguje střední hodnota vysvětlované proměnné Y na jednotkovou odchylku nezávisle proměnné x od jejího průměru $\bar{x}$
- E_i vyjadřuje náhodnou složku i-tého pozorování, $E_i \sim \mathrm{N}(0, \sigma^2)$
- odhadem závislosti je (b_1 je stejné jako při klasickém vyjádření)

$$\hat{Y}_i = \bar{Y} + b_1(x_i - \bar{x})$$

alternativní formulace

- uvažovanou závislost lze psát ve tvaru

$$Y_i = \beta_0^* + \beta_1(x_i - \bar{x}) + E_i$$

- β_0^* vyjadřuje střední úroveň vysvětlované proměnné Y při průměrné hodnotě nezávisle proměnné x
- β_1 vyjadřuje citlivost, s jakou reaguje střední hodnota vysvětlované proměnné Y na jednotkovou odchylku nezávisle proměnné x od jejího průměru $\bar{x}$
- E_i vyjadřuje náhodnou složku i-tého pozorování, $E_i \sim \mathrm{N}(0, \sigma^2)$
- odhadem závislosti je (b_1 je stejné jako při klasickém vyjádření)

$$\hat{Y}_i = \bar{Y} + b_1(x_i - \bar{x})$$

prokazování závislosti

- modelujeme závislost $\mathrm{E}\,Y$ na x pomocí $\mathrm{E}\,Y = \beta_0 + \beta_1 x$
- nezávislost $y = \beta_0 + \beta_1 x$ na x znamená $\beta_1 = 0$
- hypotézu $\mathrm{H}_0 : \beta_1 = 0$ testujeme pomocí statistiky

$$T = \frac{b_1}{\mathrm{S.E.}(b_1)}$$

- hypotézu zamítáme, je-li $|T| \geq t_{n-2}(\alpha)$
  tj. je-li příslušná p-hodnota $\leq \alpha$
- pokud H_0 zamítneme, říkáme, na hladině α je **závislost průkazná**

prokazování závislosti

- modelujeme závislost $\mathrm{E}\,Y$ na x pomocí $\mathrm{E}\,Y = \beta_0 + \beta_1 x$
- nezávislost $y = \beta_0 + \beta_1 x$ na x znamená $\beta_1 = 0$
- hypotézu $\mathrm{H}_0 : \beta_1 = 0$ testujeme pomocí statistiky

$$T = \frac{b_1}{\mathrm{S.E.}(b_1)}$$

- hypotézu zamítáme, je-li $|T| \geq t_{n-2}(\alpha)$
  tj. je-li příslušná p-hodnota $\leq \alpha$
- pokud H_0 zamítneme, říkáme, na hladině α je **závislost průkazná**

prokazování závislosti

- modelujeme závislost E Y na x pomocí E $Y = \beta_0 + \beta_1 x$
- nezávislost $y = \beta_0 + \beta_1 x$ na x znamená $\beta_1 = 0$
- hypotézu $H_0 : \beta_1 = 0$ testujeme pomocí statistiky

$$T = \frac{b_1}{\text{S.E.}(b_1)}$$

- hypotézu zamítáme, je-li $|T| \geq t_{n-2}(\alpha)$
  tj. je-li příslušná p-hodnota $\leq \alpha$
- pokud H_0 zamítneme, říkáme, na hladině α je **závislost průkazná**

prokazování závislosti

- modelujeme závislost E Y na x pomocí E $Y = \beta_0 + \beta_1 x$
- nezávislost $y = \beta_0 + \beta_1 x$ na x znamená $\beta_1 = 0$
- hypotézu $H_0 : \beta_1 = 0$ testujeme pomocí statistiky

$$T = \frac{b_1}{\text{S.E.}(b_1)}$$

- hypotézu zamítáme, je-li $|T| \geq t_{n-2}(\alpha)$
  tj. je-li příslušná p-hodnota $\leq \alpha$
- pokud H_0 zamítneme, říkáme, na hladině α je **závislost průkazná**

prokazování závislosti

- modelujeme závislost E Y na x pomocí E $Y = \beta_0 + \beta_1 x$
- nezávislost $y = \beta_0 + \beta_1 x$ na x znamená $\beta_1 = 0$
- hypotézu $H_0 : \beta_1 = 0$ testujeme pomocí statistiky

$$T = \frac{b_1}{\text{S.E.}(b_1)}$$

- hypotézu zamítáme, je-li $|T| \geq t_{n-2}(\alpha)$
  tj. je-li příslušná p-hodnota $\leq \alpha$
- pokud H_0 zamítneme, říkáme, na hladině α je **závislost průkazná**

koeficient determinace
[coefficient of determination]

- podíl variability Y vysvětlené uvažovanou závislostí
  (jakou část variability Y se podařilo závislostí na x vysvětlit)
-

$$R^2 = \frac{\text{variabilita vysvětlená}}{\text{variabilita vysvětlovaná}}$$
$$= 1 - \frac{\text{variabilita nevysvětlená}}{\text{variabilita vysvětlovaná}}$$
$$= 1 - \frac{S_e}{\sum(Y_i - \bar{Y})^2}$$

- R^2 je bezrozměrné číslo, často vyjádřeno v procentech
- R^2 ukazuje, zda má smysl předpovídat pomocí regrese

koeficient determinace
[coefficient of determination]

- podíl variability Y vysvětlené uvažovanou závislostí
  (jakou část variability Y se podařilo závislostí na x vysvětlit)
- $$R^2 = \frac{\text{variabilita vysvětlená}}{\text{variabilita vysvětlovaná}} = 1 - \frac{\text{variabilita nevysvětlená}}{\text{variabilita vysvětlovaná}} = 1 - \frac{S_e}{\sum(Y_i - \bar{Y})^2}$$
- R^2 je bezrozměrné číslo, často vyjádřeno v procentech
- R^2 ukazuje, zda má smysl předpovídat pomocí regrese

koeficient determinace
[coefficient of determination]

- podíl variability Y vysvětlené uvažovanou závislostí
  (jakou část variability Y se podařilo závislostí na x vysvětlit)
- $$R^2 = \frac{\text{variabilita vysvětlená}}{\text{variabilita vysvětlovaná}} = 1 - \frac{\text{variabilita nevysvětlená}}{\text{variabilita vysvětlovaná}} = 1 - \frac{S_e}{\sum(Y_i - \bar{Y})^2}$$
- R^2 je bezrozměrné číslo, často vyjádřeno v procentech
- R^2 ukazuje, zda má smysl předpovídat pomocí regrese

koeficient determinace
[coefficient of determination]

- podíl variability Y vysvětlené uvažovanou závislostí (jakou část variability Y se podařilo závislostí na x vysvětlit)
-

$$R^2 = \frac{\text{variabilita vysvětlená}}{\text{variabilita vysvětlovaná}}$$
$$= 1 - \frac{\text{variabilita nevysvětlená}}{\text{variabilita vysvětlovaná}}$$
$$= 1 - \frac{S_e}{\sum(Y_i - \bar{Y})^2}$$

- R^2 je bezrozměrné číslo, často vyjádřeno v procentech
- R^2 ukazuje, zda má smysl předpovídat pomocí regrese

příklad závislost procenta tuku na výšce

data: Policie

regresor	b_j	S.E.(b_j)	t	p
abs. člen	–53,870	24,657	–2,185	0,0338
height	0,379	0,138	2,742	0,0086

- předpověď: $\hat{Y}_i = -53{,}870 + 0{,}379 x_i$
- $\widehat{\text{fat}} = -53{,}870 + 0{,}379 \cdot \text{height}$
- závislost procenta tuku na výšce je na 5% hladině průkazná
- na každý centimetr výšky *v průměru* přibude 0,379 procentního bodu tuku
- [summary(lm(fat∼height))]

příklad závislost procenta tuku na výšce

data: Policie

regresor	b_j	S.E.(b_j)	t	p
abs. člen	–53,870	24,657	–2,185	0,0338
height	0,379	0,138	2,742	0,0086

- předpověď: $\hat{Y}_i = -53{,}870 + 0{,}379 x_i$
- $\widehat{\text{fat}} = -53{,}870 + 0{,}379 \cdot \text{height}$
- závislost procenta tuku na výšce je na 5% hladině průkazná
- na každý centimetr výšky *v průměru* přibude 0,379 procentního bodu tuku
- [summary(lm(fat∼height))]

příklad závislost procenta tuku na výšce

data: Policie

regresor	b_j	S.E.(b_j)	t	p
abs. člen	–53,870	24,657	–2,185	0,0338
height	0,379	0,138	2,742	0,0086

- předpověď: $\hat{Y}_i = -53{,}870 + 0{,}379 x_i$
- $\widehat{\text{fat}} = -53{,}870 + 0{,}379 \cdot \text{height}$
- závislost procenta tuku na výšce je na 5% hladině průkazná
- na každý centimetr výšky *v průměru* přibude 0,379 procentního bodu tuku
- [summary(lm(fat∼height))]

příklad závislost procenta tuku na výšce

data: Policie

regresor	b_j	S.E.(b_j)	t	p
abs. člen	–53,870	24,657	–2,185	0,0338
height	0,379	0,138	2,742	0,0086

- předpověď: $\hat{Y}_i = -53{,}870 + 0{,}379 x_i$
- $\widehat{\text{fat}} = -53{,}870 + 0{,}379 \cdot \text{height}$
- závislost procenta tuku na výšce je na 5% hladině průkazná
- na každý centimetr výšky *v průměru* přibude 0,379 procentního bodu tuku
- [summary(lm(fat∼height))]

příklad závislost procenta tuku na výšce

data: Policie

regresor	b_j	S.E.(b_j)	t	p
abs. člen	–53,870	24,657	–2,185	0,0338
height	0,379	0,138	2,742	0,0086

- předpověď: $\hat{Y}_i = -53{,}870 + 0{,}379 x_i$
- $\widehat{\text{fat}} = -53{,}870 + 0{,}379 \cdot \text{height}$
- závislost procenta tuku na výšce je na 5% hladině průkazná
- na každý centimetr výšky *v průměru* přibude 0,379 procentního bodu tuku
- [summary(lm(fat∼height))]

tabulka analýzy rozptylu

variabilita	součet čtverců	st. vol.	prům. čtverec	F	p
regrese	362,54	1	362,54	7,519	0,0086
rezid.	2314,41	48	48,22		
celk.	2676,95	49	(54,63)		

- $s^2 = 48{,}22$
- $$R^2 = \frac{362{,}54}{2676{,}95} = 1 - \frac{2314{,}41}{2676{,}95} = 0{,}135$$
- závislostí na výšce jsme vysvětlili jen 13,5 % variability procenta tuku
- [anova(lm(fat∼height))]

tabulka analýzy rozptylu

variabilita	součet čtverců	st. vol.	prům. čtverec	F	p
regrese	362,54	1	362,54	7,519	0,0086
rezid.	2314,41	48	48,22		
celk.	2676,95	49	(54,63)		

- $s^2 = 48{,}22$
- $$R^2 = \frac{362{,}54}{2676{,}95} = 1 - \frac{2314{,}41}{2676{,}95} = 0{,}135$$
- závislostí na výšce jsme vysvětlili jen 13,5 % variability procenta tuku
- [anova(lm(fat∼height))]

tabulka analýzy rozptylu

variabilita	součet čtverců	st. vol.	prům. čtverec	F	p
regrese	362,54	1	362,54	7,519	0,0086
rezid.	2314,41	48	48,22		
celk.	2676,95	49	(54,63)		

- $s^2 = 48{,}22$
- $$R^2 = \frac{362{,}54}{2676{,}95} = 1 - \frac{2314{,}41}{2676{,}95} = 0{,}135$$
- závislostí na výšce jsme vysvětlili jen 13,5 % variability procenta tuku
- [anova(lm(fat∼height))]

tabulka analýzy rozptylu

variabilita	součet čtverců	st. vol.	prům. čtverec	F	p
regrese	362,54	1	362,54	7,519	0,0086
rezid.	2314,41	48	48,22		
celk.	2676,95	49	(54,63)		

- $s^2 = 48{,}22$
- $$R^2 = \frac{362{,}54}{2676{,}95} = 1 - \frac{2314{,}41}{2676{,}95} = 0{,}135$$
- závislostí na výšce jsme vysvětlili jen 13,5 % variability procenta tuku
- [anova(lm(fat∼height))]