Základy biostatistiky

(MD710P09)
ak. rok 2008/2009

Karel Zvára

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

(naposledy upraveno 27. dubna 2009)

Fisherova *z*-transformace

(přibližší 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!**)

$$z = \frac{0{,}286 - 0{,}151}{\sqrt{\frac{1}{50-3} + \frac{1}{49-3}}} = 0{,}650.$$

srovnej s kritickou hodnotou $z(0{,}05/2) = 1{,}960$, $p = 51{,}6\ \%$

Fisherova *z*-transformace

(přibližší chování výběrového korelačního koeficientu *r* normálnímu rozdělení)

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

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!**)

$$z = \frac{0{,}286 - 0{,}151}{\sqrt{\frac{1}{50-3} + \frac{1}{49-3}}} = 0{,}650.$$

srovnej s kritickou hodnotou $z(0{,}05/2) = 1{,}960$, $p = 51{,}6\ \%$

Fisherova z-transformace

(přibližuje 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!**)

$$z = \frac{0{,}286 - 0{,}151}{\sqrt{\frac{1}{50-3} + \frac{1}{49-3}}} = 0{,}650.$$

srovnej s kritickou hodnotou $z(0{,}05/2) = 1{,}960,\ p = 51{,}6\ \%$

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 je jen přibližný!)

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 je jen přibližný!)

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 je jen přibližný!)

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 je jen přibližný!)

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 je jen přibližný!)

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 je jen přibližný!)

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 je jen přibližný!)

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

- **předpokládaná** 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, normalitu nelze ověřovat testováním přímo $Y_1, \ldots, Y_n$!)
- neznámé populační parametry β_0, β_1 odhadujeme metodou **nejmenších čtverců**:

minimalizovat přes β_0, β_1 výraz $$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

- odhady označíme b_0, b_1

regresní přímka

- **předpokládaná** závislost střední hodnoty Y na nenáhodné x:

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

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

  minimalizovat přes β_0, β_1 výraz $$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$
- odhady označíme b_0, b_1

regresní přímka

- **předpokládaná** závislost střední hodnoty Y na nenáhodné x:

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

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

minimalizovat přes β_0, β_1 výraz $\displaystyle\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$

- odhady označíme b_0, b_1

regresní přímka

- **předpokládaná** závislost střední hodnoty Y na nenáhodné x:

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

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

minimalizovat přes β_0, β_1 výraz $$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

- odhady označíme b_0, b_1

regresní přímka

- **předpokládaná** 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, normalitu nelze ověřovat testováním přímo $Y_1, \ldots, Y_n$!)
- neznámé populační parametry β_0, β_1 odhadujeme metodou **nejmenších čtverců**:

minimalizovat přes β_0, β_1 výraz $$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

- odhady označíme b_0, b_1

regresní přímka

- **předpokládaná** závislost střední hodnoty Y na nenáhodné x:

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

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

  minimalizovat přes β_0, β_1 výraz $$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

- odhady označíme b_0, b_1

regresní přímka

- **předpokládaná** závislost střední hodnoty Y na nenáhodné x:

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

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

  minimalizovat přes β_0, β_1 výraz $$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

- odhady označíme b_0, b_1

regresní přímka

- **předpokládaná** závislost střední hodnoty Y na nenáhodné x:

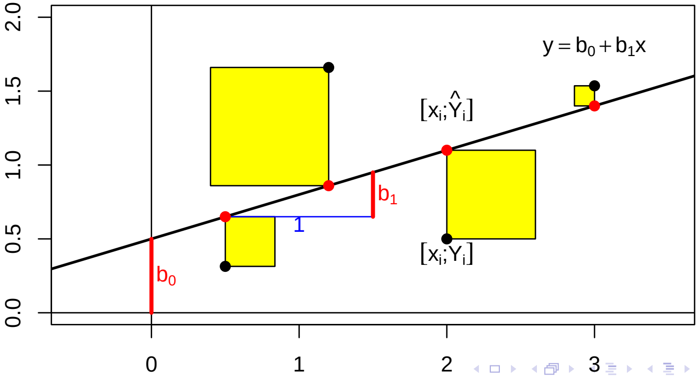
$$\mathrm{E}\, Y = \beta_0 + \beta_1 x$$

- 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, normalitu nelze ověřovat testováním přímo $Y_1, \ldots, Y_n$!)
- neznámé populační parametry β_0, β_1 odhadujeme metodou **nejmenších čtverců**:

  minimalizovat přes β_0, β_1 výraz $$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2$$
- 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

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

- 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

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

- 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

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

- 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

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

- 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, $E_i \sim \mathrm{N}\left(0, \sigma^2\right)$
- 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

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

- 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, $E_i \sim \mathrm{N}\left(0, \sigma^2\right)$
- 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 $\mathrm{E}\,Y$ 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 $\mathrm{H}_0 : \beta_1 = 0$ 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**

koeficient determinace

[coefficient of determination]

- podíl variability Y vysvětlené uvažovanou závislostí (jakou část variability Y se podařilo závislostí na x vysvětlit)
- $$R^2 = \frac{\text{variabilita vysvětlená}}{\text{variabilita vysvětlovaná}} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$$
$$= 1 - \frac{\text{variabilita nevysvětlená}}{\text{variabilita vysvětlovaná}} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$
$$= 1 - \frac{S_e}{\sum(Y_i - \bar{Y})^2}$$
- 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)
-

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

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

$$= 1 - \frac{S_e}{\sum(Y_i - \bar{Y})^2}$$

- 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)
- $$\begin{aligned} R^2 &= \frac{\text{variabilita vysvětlená}}{\text{variabilita vysvětlovaná}} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} \\ &= 1 - \frac{\text{variabilita nevysvětlená}}{\text{variabilita vysvětlovaná}} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2} \\ &= 1 - \frac{S_e}{\sum(Y_i - \bar{Y})^2} \end{aligned}$$
- 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)
- $$\begin{aligned} R^2 &= \frac{\text{variabilita vysvětlená}}{\text{variabilita vysvětlovaná}} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} \\ &= 1 - \frac{\text{variabilita nevysvětlená}}{\text{variabilita vysvětlovaná}} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2} \\ &= 1 - \frac{S_e}{\sum(Y_i - \bar{Y})^2} \end{aligned}$$
- 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: $\hat{Y}_i = -53{,}870 + 0{,}379 x_i$
- $\widehat{\text{fat}} = -\ 53{,}870\ +\ 0{,}379\ \cdot$ 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))]

mnohonásobná lineární regrese

- závislost na dvou (nebo více) nezávisle proměnných
- pozorování $(x_1, v_1, Y_1), \dots, (x_n, v_n, Y_n)$
- představa (model)

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i + \beta_2 v_i}_{\mathrm{E}\, Y_i} + E_i$$

- střední hodnota Y_i (tj. systematická, nenáhodná složka Y_i) vysvětlena pomocí x_i, v_i jako $\beta_0 + \beta_1 x_i + \beta_2 v_i$
- $E_1, \dots, E_n$ (také $Y_1, \dots, Y_n$) jsou **nezávislé** náhodné veličiny
- $E_i \sim \mathrm{N}(0, \sigma^2)$ (normální rozdělení se stejným rozptylem)
- b_0, b_1, b_2 – odhady parametrů $\beta_0, \beta_1, \beta_2$

mnohonásobná lineární regrese

- závislost na dvou (nebo více) nezávisle proměnných
- pozorování $(x_1, v_1, Y_1), \ldots, (x_n, v_n, Y_n)$
- představa (model)

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i + \beta_2 v_i}_{\mathrm{E}\, Y_i} + E_i$$

- střední hodnota Y_i (tj. systematická, nenáhodná složka Y_i) vysvětlena pomocí x_i, v_i jako $\beta_0 + \beta_1 x_i + \beta_2 v_i$
- $E_1, \ldots, E_n$ (také $Y_1, \ldots, Y_n$) jsou **nezávislé** náhodné veličiny
- $E_i \sim \mathrm{N}(0, \sigma^2)$ (normální rozdělení se stejným rozptylem)
- b_0, b_1, b_2 – odhady parametrů $\beta_0, \beta_1, \beta_2$

mnohonásobná lineární regrese

- závislost na dvou (nebo více) nezávisle proměnných
- pozorování $(x_1, v_1, Y_1), \ldots, (x_n, v_n, Y_n)$
- představa (model)

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i + \beta_2 v_i}_{\mathrm{E}\, Y_i} + E_i$$

- střední hodnota Y_i (tj. systematická, nenáhodná složka Y_i) vysvětlena pomocí x_i, v_i jako $\beta_0 + \beta_1 x_i + \beta_2 v_i$
- $E_1, \ldots, E_n$ (také $Y_1, \ldots, Y_n$) jsou **nezávislé** náhodné veličiny
- $E_i \sim \mathrm{N}(0, \sigma^2)$ (normální rozdělení se stejným rozptylem)
- b_0, b_1, b_2 – odhady parametrů $\beta_0, \beta_1, \beta_2$

mnohonásobná lineární regrese

- závislost na dvou (nebo více) nezávisle proměnných
- pozorování $(x_1, v_1, Y_1), \ldots, (x_n, v_n, Y_n)$
- představa (model)

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i + \beta_2 v_i}_{\mathrm{E}\, Y_i} + E_i$$

- střední hodnota Y_i (tj. systematická, nenáhodná složka Y_i) vysvětlena pomocí x_i, v_i jako $\beta_0 + \beta_1 x_i + \beta_2 v_i$
- $E_1, \ldots, E_n$ (také $Y_1, \ldots, Y_n$) jsou **nezávislé** náhodné veličiny
- $E_i \sim \mathrm{N}(0, \sigma^2)$ (normální rozdělení se stejným rozptylem)
- b_0, b_1, b_2 – odhady parametrů $\beta_0, \beta_1, \beta_2$

mnohonásobná lineární regrese

- závislost na dvou (nebo více) nezávisle proměnných
- pozorování $(x_1, v_1, Y_1), \dots, (x_n, v_n, Y_n)$
- představa (model)

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i + \beta_2 v_i}_{\mathrm{E}\, Y_i} + E_i$$

- střední hodnota Y_i (tj. systematická, nenáhodná složka Y_i) vysvětlena pomocí x_i, v_i jako $\beta_0 + \beta_1 x_i + \beta_2 v_i$
- $E_1, \dots, E_n$ (také $Y_1, \dots, Y_n$) jsou **nezávislé** náhodné veličiny
- $E_i \sim \mathrm{N}(0, \sigma^2)$ (normální rozdělení se stejným rozptylem)
- b_0, b_1, b_2 – odhady parametrů $\beta_0, \beta_1, \beta_2$

mnohonásobná lineární regrese

- závislost na dvou (nebo více) nezávisle proměnných
- pozorování $(x_1, v_1, Y_1), \dots, (x_n, v_n, Y_n)$
- představa (model)

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i + \beta_2 v_i}_{\mathrm{E}\, Y_i} + E_i$$

- střední hodnota Y_i (tj. systematická, nenáhodná složka Y_i) vysvětlena pomocí x_i, v_i jako $\beta_0 + \beta_1 x_i + \beta_2 v_i$
- $E_1, \dots, E_n$ (také $Y_1, \dots, Y_n$) jsou **nezávislé** náhodné veličiny
- $E_i \sim \mathrm{N}(0, \sigma^2)$ (normální rozdělení se stejným rozptylem)
- b_0, b_1, b_2 – odhady parametrů $\beta_0, \beta_1, \beta_2$

mnohonásobná lineární regrese

- závislost na dvou (nebo více) nezávisle proměnných
- pozorování $(x_1, v_1, Y_1), \dots, (x_n, v_n, Y_n)$
- představa (model)

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i + \beta_2 v_i}_{\mathrm{E}\, Y_i} + E_i$$

- střední hodnota Y_i (tj. systematická, nenáhodná složka Y_i) vysvětlena pomocí x_i, v_i jako $\beta_0 + \beta_1 x_i + \beta_2 v_i$
- $E_1, \dots, E_n$ (také $Y_1, \dots, Y_n$) jsou **nezávislé** náhodné veličiny
- $E_i \sim \mathrm{N}\left(0, \sigma^2\right)$ (normální rozdělení se stejným rozptylem)
- b_0, b_1, b_2 – odhady parametrů $\beta_0, \beta_1, \beta_2$

interpretace

- b_1 – odhad změny střední hodnoty Y při **jednotkové** změně x a **nezměněné** hodnotě v
- b_2 – odhad změny střední hodnoty Y při **jednotkové** změně v a **nezměněné** hodnotě x
- U_i – **reziduum**

$$U_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 x_i + b_2 v_i)$$

- **rozklad variability** $S_T = S_R + S_e$

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

interpretace

- b_1 – odhad změny střední hodnoty Y při **jednotkové** změně x a **nezměněné** hodnotě v
- b_2 – odhad změny střední hodnoty Y při **jednotkové** změně v a **nezměněné** hodnotě x
- U_i – **reziduum**

$$U_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 x_i + b_2 v_i)$$

- **rozklad variability** $S_T = S_R + S_e$

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

interpretace

- b_1 – odhad změny střední hodnoty Y při **jednotkové** změně x a **nezměněné** hodnotě v
- b_2 – odhad změny střední hodnoty Y při **jednotkové** změně v a **nezměněné** hodnotě x
- U_i – **reziduum**

$$U_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 x_i + b_2 v_i)$$

- **rozklad variability** $S_T = S_R + S_e$

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

interpretace

- b_1 – odhad změny střední hodnoty Y při **jednotkové** změně x a **nezměněné** hodnotě v
- b_2 – odhad změny střední hodnoty Y při **jednotkové** změně v a **nezměněné** hodnotě x
- U_i – **reziduum**

$$U_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 x_i + b_2 v_i)$$

- **rozklad variability** $S_T = S_R + S_e$

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- **koeficient determinace R^2**
  podíl celkové variability, který se podařilo vysvětlit závislostí Y na x, v (jakou část variability Y se podařilo vysvětlit)

$$R^2 = \frac{S_R}{S_T} = 1 - \frac{S_e}{S_T}$$

- $\mathrm{H}_0 : \beta_1 = \beta_2 = 0$ (chování Y nezávisí ani na x ani na v)

$$F = \frac{S_R/2}{S_e/(n-3)} \geq F_{2,n-3}(\alpha)$$

- p-hodnota tohoto testu bývá uváděna spolu s R^2

- **koeficient determinace** R^2
  podíl celkové variability, který se podařilo vysvětlit závislostí Y na x, v (jakou část variability Y se podařilo vysvětlit)

$$R^2 = \frac{S_R}{S_T} = 1 - \frac{S_e}{S_T}$$

- $\mathrm{H}_0 : \beta_1 = \beta_2 = 0$ (chování Y nezávisí ani na x ani na v)

$$F = \frac{S_R/2}{S_e/(n-3)} \geq F_{2,n-3}(\alpha)$$

- p-hodnota tohoto testu bývá uváděna spolu s R^2

- **koeficient determinace** R^2
  podíl celkové variability, který se podařilo vysvětlit závislostí Y na x, v (jakou část variability Y se podařilo vysvětlit)

$$R^2 = \frac{S_R}{S_T} = 1 - \frac{S_e}{S_T}$$

- $\mathrm{H}_0 : \beta_1 = \beta_2 = 0$ (chování Y nezávisí ani na x ani na v)

$$F = \frac{S_R/2}{S_e/(n-3)} \geq F_{2,n-3}(\alpha)$$

- p-hodnota tohoto testu bývá uváděna spolu s R^2

testy o přínosu jednotlivých regresorů

- model $y = \beta_0 + \beta_1 x + \beta_2 v$
- $H_0 : \beta_2 = 0$
  k vysvětlení chování Y stačí x, tj. $y = \beta_0 + \beta_1 x$
  $$T_2 = \frac{b_2}{\mathrm{S.E.}(b_2)}, \qquad \text{zamítat pro } |T_2| \geq t_{n-3}(\alpha)$$
- $H_0 : \beta_1 = 0$
  k vysvětlení chování Y stačí v, tj. $y = \beta_0 + \beta_2 v$
  $$T_1 = \frac{b_1}{\mathrm{S.E.}(b_1)}, \qquad \text{zamítat pro } |T_1| \geq t_{n-3}(\alpha)$$
- $H_0 : \beta_0 = 0$ zpravidla nemá reálný smysl

testy o přínosu jednotlivých regresorů

- model $\quad y = \beta_0 + \beta_1 x + \beta_2 v$
- $H_0 : \beta_2 = 0$
  k vysvětlení chování Y stačí x, tj. $y = \beta_0 + \beta_1 x$

$$T_2 = \frac{b_2}{\text{S.E.}(b_2)}, \qquad \text{zamítat pro } |T_2| \geq t_{n-3}(\alpha)$$

- $H_0 : \beta_1 = 0$
  k vysvětlení chování Y stačí v, tj. $y = \beta_0 + \beta_2 v$

$$T_1 = \frac{b_1}{\text{S.E.}(b_1)}, \qquad \text{zamítat pro } |T_1| \geq t_{n-3}(\alpha)$$

- $H_0 : \beta_0 = 0$ zpravidla nemá reálný smysl

testy o přínosu jednotlivých regresorů

- model $y = \beta_0 + \beta_1 x + \beta_2 v$
- $\mathrm{H}_0 : \beta_2 = 0$
  k vysvětlení chování Y stačí x, tj. $y = \beta_0 + \beta_1 x$

$$T_2 = \frac{b_2}{\mathrm{S.E.}(b_2)}, \qquad \text{zamítat pro } |T_2| \geq t_{n-3}(\alpha)$$

- $\mathrm{H}_0 : \beta_1 = 0$
  k vysvětlení chování Y stačí v, tj. $y = \beta_0 + \beta_2 v$

$$T_1 = \frac{b_1}{\mathrm{S.E.}(b_1)}, \qquad \text{zamítat pro } |T_1| \geq t_{n-3}(\alpha)$$

- $\mathrm{H}_0 : \beta_0 = 0$ zpravidla nemá reálný smysl

testy o přínosu jednotlivých regresorů

- model $\quad y = \beta_0 + \beta_1 x + \beta_2 v$
- $H_0 : \beta_2 = 0$
  k vysvětlení chování Y stačí x, tj. $y = \beta_0 + \beta_1 x$

$$T_2 = \frac{b_2}{\text{S.E.}(b_2)}, \qquad \text{zamítat pro } |T_2| \geq t_{n-3}(\alpha)$$

- $H_0 : \beta_1 = 0$
  k vysvětlení chování Y stačí v, tj. $y = \beta_0 + \beta_2 v$

$$T_1 = \frac{b_1}{\text{S.E.}(b_1)}, \qquad \text{zamítat pro } |T_1| \geq t_{n-3}(\alpha)$$

- $H_0 : \beta_0 = 0$ zpravidla nemá reálný smysl

příklad: závislost procenta tuku na výšce a váze

data: Policie

regresor	b_j	S.E.(b_j)	t	p
abs. člen	11,327	16,682	0,679	0,5005
height	–0,262	0,110	–2,376	0,0216
weight	0,624	0,0690	9,050	<0,0001

- [summary(lm(fat∼height+weight))]
- při **stejné výšce** očekáváme na každý kg hmotnosti o 0,6 proc. bodu více tuku
- u mužů, kteří se liší výškou o 10 cm a **mají stejnou hmotnost** očekáváme, že ti vyšší mají v průměru o 2,6 proc. bodu **méně** tuku
- na 5% hladině nelze vyloučit výšku, průkazně přispívá k vysvětlení pomocí váhy
- na 1% hladině nelze vyloučit váhu, průkazně přispívá k vysvětlení pomocí výšky

příklad: závislost procenta tuku na výšce a váze

data: Policie

regresor	b_j	S.E.(b_j)	t	p
abs. člen	11,327	16,682	0,679	0,5005
height	–0,262	0,110	–2,376	0,0216
weight	0,624	0,0690	9,050	<0,0001

- [summary(lm(fat∼height+weight))]
- při **stejné výšce** očekáváme na každý kg hmotnosti o 0,6 proc. bodu více tuku
- u mužů, kteří se liší výškou o 10 cm a **mají stejnou hmotnost** očekáváme, že ti vyšší mají v průměru o 2,6 proc. bodu **méně** tuku
- na 5% hladině nelze vyloučit výšku, průkazně přispívá k vysvětlení pomocí váhy
- na 1% hladině nelze vyloučit váhu, průkazně přispívá k vysvětlení pomocí výšky

příklad: závislost procenta tuku na výšce a váze

data: Policie

regresor	b_j	S.E.(b_j)	t	p
abs. člen	11,327	16,682	0,679	0,5005
height	–0,262	0,110	–2,376	0,0216
weight	0,624	0,0690	9,050	<0,0001

- [summary(lm(fat∼height+weight))]
- při **stejné výšce** očekáváme na každý kg hmotnosti o 0,6 proc. bodu více tuku
- u mužů, kteří se liší výškou o 10 cm a **mají stejnou hmotnost** očekáváme, že ti vyšší mají v průměru o 2,6 proc. bodu **méně** tuku
- na 5% hladině nelze vyloučit výšku, průkazně přispívá k vysvětlení pomocí váhy
- na 1% hladině nelze vyloučit váhu, průkazně přispívá k vysvětlení pomocí výšky

příklad: závislost procenta tuku na výšce a váze

data: Policie

regresor	b_j	S.E.(b_j)	t	p
abs. člen	11,327	16,682	0,679	0,5005
height	–0,262	0,110	–2,376	0,0216
weight	0,624	0,0690	9,050	<0,0001

- [summary(lm(fat∼height+weight))]
- při **stejné výšce** očekáváme na každý kg hmotnosti o 0,6 proc. bodu více tuku
- u mužů, kteří se liší výškou o 10 cm a **mají stejnou hmotnost** očekáváme, že ti vyšší mají v průměru o 2,6 proc. bodu **méně** tuku
- na 5% hladině nelze vyloučit výšku, průkazně přispívá k vysvětlení pomocí váhy
- na 1% hladině nelze vyloučit váhu, průkazně přispívá k vysvětlení pomocí výšky

příklad: závislost procenta tuku na výšce a váze

data: Policie

regresor	b_j	S.E.(b_j)	t	p
abs. člen	11,327	16,682	0,679	0,5005
height	–0,262	0,110	–2,376	0,0216
weight	0,624	0,0690	9,050	<0,0001

- [summary(lm(fat∼height+weight))]
- při **stejné výšce** očekáváme na každý kg hmotnosti o 0,6 proc. bodu více tuku
- u mužů, kteří se liší výškou o 10 cm a **mají stejnou hmotnost** očekáváme, že ti vyšší mají v průměru o 2,6 proc. bodu **méně** tuku
- na 5% hladině nelze vyloučit výšku, průkazně přispívá k vysvětlení pomocí váhy
- na 1% hladině nelze vyloučit váhu, průkazně přispívá k vysvětlení pomocí výšky

tabulka analýzy rozptylu

variabilita	souč. čtv,	st. vol.	prům. čtv.	F	p
regrese	1833,11	2	916,55	51,050	<0,001
rezid.	843,85	47	17,95		
celk.	2676,95	49	(54,63)		

- $R^2 = 1833{,}11/2676{,}95 = 1 - 843{,}85/2676{,}95 = 0{,}685$
- závislostí na výšce a váze jsme vysvětlili 68,5 % variability procenta tuku
- $s^2 = 17{,}95$
- na každé rozumné hladině zamítáme hypotézu, podle které procento tuku nezávisí ani na výšce ani na váze

tabulka analýzy rozptylu

variabilita	souč. čtv,	st. vol.	prům. čtv.	F	p
regrese	1833,11	2	916,55	51,050	<0,001
rezid.	843,85	47	17,95		
celk.	2676,95	49	(54,63)		

- $R^2 = 1833{,}11/2676{,}95 = 1 - 843{,}85/2676{,}95 = 0{,}685$
- závislostí na výšce a váze jsme vysvětlili 68,5 % variability procenta tuku
- $s^2 = 17{,}95$
- na každé rozumné hladině zamítáme hypotézu, podle které procento tuku nezávisí ani na výšce ani na váze

tabulka analýzy rozptylu

variabilita	souč. čtv,	st. vol.	prům. čtv.	F	p
regrese	1833,11	2	916,55	51,050	<0,001
rezid.	843,85	47	17,95		
celk.	2676,95	49	(54,63)		

- $R^2 = 1833{,}11/2676{,}95 = 1 - 843{,}85/2676{,}95 = 0{,}685$
- závislostí na výšce a váze jsme vysvětlili 68,5 % variability procenta tuku
- $s^2 = 17{,}95$
- na každé rozumné hladině zamítáme hypotézu, podle které procento tuku nezávisí ani na výšce ani na váze

tabulka analýzy rozptylu

variabilita	souč. čtv,	st. vol.	prům. čtv.	F	p
regrese	1833,11	2	916,55	51,050	<0,001
rezid.	843,85	47	17,95		
celk.	2676,95	49	(54,63)		

- $R^2 = 1833{,}11/2676{,}95 = 1 - 843{,}85/2676{,}95 = 0{,}685$
- závislostí na výšce a váze jsme vysvětlili 68,5 % variability procenta tuku
- $s^2 = 17{,}95$
- na každé rozumné hladině zamítáme hypotézu, podle které procento tuku nezávisí ani na výšce ani na váze

regresní diagnostika

zda byly splněny předpoklady

a) zvolili jsme správně **tvar závislosti**?

b) je **rozptyl** všude **stejný**?

c) je přiměřeně splněn předpoklad o **normálním rozdělení**?

d) jsou opravdu pozorování **nezávislá**?
problém často tam, kde působí čas

- často pomůže transformace (a), b), c), např. logaritmování závisle proměnné
- [plot(lm(fat∼height+weight))]

regresní diagnostika

zda byly splněny předpoklady

a) zvolili jsme správně **tvar závislosti**?

b) je **rozptyl** všude **stejný**?

c) je přiměřeně splněn předpoklad o **normálním rozdělení**?

d) jsou opravdu pozorování **nezávislá**?
problém často tam, kde působí čas

- často pomůže transformace (a), b), c), např. logaritmování závisle proměnné
- [plot(lm(fat~height+weight))]

regresní diagnostika

zda byly splněny předpoklady

a) zvolili jsme správně **tvar závislosti**?

b) je **rozptyl** všude **stejný**?

c) je přiměřeně splněn předpoklad o **normálním rozdělení**?

d) jsou opravdu pozorování **nezávislá**?
problém často tam, kde působí čas

- často pomůže transformace (a), b), c), např. logaritmování závisle proměnné
- [plot(lm(fat∼height+weight))]

regresní diagnostika

zda byly splněny předpoklady

a) zvolili jsme správně **tvar závislosti**?

b) je **rozptyl** všude **stejný**?

c) je přiměřeně splněn předpoklad o **normálním rozdělení**?

d) jsou opravdu pozorování **nezávislá**?
problém často tam, kde působí čas

- často pomůže transformace (a), b), c), např. logaritmování závisle proměnné
- [plot(lm(fat∼height+weight))]

regresní diagnostika

zda byly splněny předpoklady

a) zvolili jsme správně **tvar závislosti**?

b) je **rozptyl** všude **stejný**?

c) je přiměřeně splněn předpoklad o **normálním rozdělení**?

d) jsou opravdu pozorování **nezávislá**?
problém často tam, kde působí čas

- často pomůže transformace (a), b), c), např. logaritmování závisle proměnné
- [plot(lm(fat~height+weight))]

regresní diagnostika
zda byly splněny předpoklady

a) zvolili jsme správně **tvar závislosti**?

b) je **rozptyl** všude **stejný**?

c) je přiměřeně splněn předpoklad o **normálním rozdělení**?

d) jsou opravdu pozorování **nezávislá**?
problém často tam, kde působí čas

- často pomůže transformace (a), b), c), např. logaritmování závisle proměnné
- [plot(lm(fat∼height+weight))]