Statistika

(MD360P03Z, MD360P03U)
ak. rok 2007/2008

Karel Zvára

karel.zvara@mff.cuni.cz
http://www.karlin.mff.cuni.cz/∼zvara

(naposledy upraveno 3. prosince 2007)

Regrese

- na rozdíl od korelace (síla závislosti) hledáme tvar (způsob) závislosti, zajímá nás také průkaznost závislosti
- snažíme se z daných hodnot **regresorů (nezávisle proměnných)** předpovědět hodnoty **závisle proměnné** (odezvy, vysvětlované proměnné)
- snažíme se variabilitu (kolísání hodnot) odezvy vysvětlit kolísáním regresorů
- prvně v tomto smyslu F. Galton (1886) při vyšetřování závislosti výšky potomků na průměrné výšce rodičů
- Pearson, Lee (1903): potomci otců o dva palce vyšších než průměr všech otců byli v průměru jen o palec vyšší než průměr synů; dvoupalcová odchylka se nereprodukovala celá, byl patrný návrat (**regres**) k průměru

Regrese

- na rozdíl od korelace (síla závislosti) hledáme tvar (způsob) závislosti, zajímá nás také průkaznost závislosti
- snažíme se z daných hodnot **regresorů (nezávisle proměnných)** předpovědět hodnoty **závisle proměnné** (odezvy, vysvětlované proměnné)
- snažíme se variabilitu (kolísání hodnot) odezvy vysvětlit kolísáním regresorů
- prvně v tomto smyslu F. Galton (1886) při vyšetřování závislosti výšky potomků na průměrné výšce rodičů
- Pearson, Lee (1903): potomci otců o dva palce vyšších než průměr všech otců byli v průměru jen o palec vyšší než průměr synů; dvoupalcová odchylka se nereprodukovala celá, byl patrný návrat (**regres**) k průměru

Regrese

- na rozdíl od korelace (síla závislosti) hledáme tvar (způsob) závislosti, zajímá nás také průkaznost závislosti
- snažíme se z daných hodnot **regresorů (nezávisle proměnných)** předpovědět hodnoty **závisle proměnné** (odezvy, vysvětlované proměnné)
- snažíme se variabilitu (kolísání hodnot) odezvy vysvětlit kolísáním regresorů
- prvně v tomto smyslu F. Galton (1886) při vyšetřování závislosti výšky potomků na průměrné výšce rodičů
- Pearson, Lee (1903): potomci otců o dva palce vyšších než průměr všech otců byli v průměru jen o palec vyšší než průměr synů; dvoupalcová odchylka se nereprodukovala celá, byl patrný návrat (**regres**) k průměru

Regrese

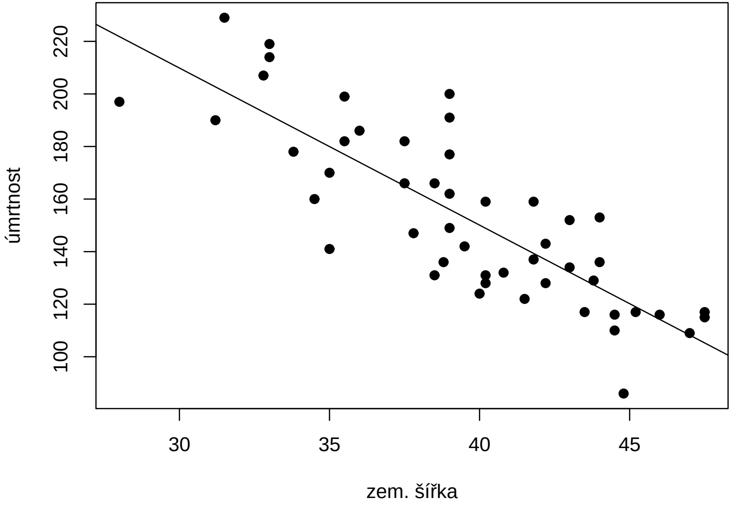
- na rozdíl od korelace (síla závislosti) hledáme tvar (způsob) závislosti, zajímá nás také průkaznost závislosti
- snažíme se z daných hodnot **regresorů (nezávisle proměnných)** předpovědět hodnoty **závisle proměnné** (odezvy, vysvětlované proměnné)
- snažíme se variabilitu (kolísání hodnot) odezvy vysvětlit kolísáním regresorů
- prvně v tomto smyslu F. Galton (1886) při vyšetřování závislosti výšky potomků na průměrné výšce rodičů
- Pearson, Lee (1903): potomci otců o dva palce vyšších než průměr všech otců byli v průměru jen o palec vyšší než průměr synů; dvoupalcová odchylka se nereprodukovala celá, byl patrný návrat (**regres**) k průměru

Regrese

- na rozdíl od korelace (síla závislosti) hledáme tvar (způsob) závislosti, zajímá nás také průkaznost závislosti
- snažíme se z daných hodnot **regresorů (nezávisle proměnných)** předpovědět hodnoty **závisle proměnné** (odezvy, vysvětlované proměnné)
- snažíme se variabilitu (kolísání hodnot) odezvy vysvětlit kolísáním regresorů
- prvně v tomto smyslu F. Galton (1886) při vyšetřování závislosti výšky potomků na průměrné výšce rodičů
- Pearson, Lee (1903): potomci otců o dva palce vyšších než průměr všech otců byli v průměru jen o palec vyšší než průměr synů; dvoupalcová odchylka se nereprodukovala celá, byl patrný návrat (**regres**) k průměru

příklad: souvisí úmrtnost se zeměpisnou šířkou?

úmrtnost na melanom na 10 000 000 obyvatel v státech USA



regresní přímka

- chování Y (úmrtnost, mortality) co nejlépe (nejvíce) vysvětlit lineární závislostí na x (zeměpisná šířka, latitude)
- (naše představa, předpoklad:) každé zem. šířce odpovídá jakási střední úmrtnost, ta závisí na zeměpisné šířce lineárně

$$\mathrm{E}\, Y_i = \beta_0 + \beta_1 x_i, \qquad i = 1, \ldots, n$$

- parametry β_0, β_1 odhadneme **metodou nejmenších čtverců** minimalizací přes β_0, β_1 součtu čtverců „svislých“ odchylek

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

- výsledné minimum (pro $\beta_0 = b_0, \beta_1 = b_1$) nazveme **reziduální součet čtverců**, tj. $S_e = \sum_{i=1}^{n} (Y_i - b_0 - b_1 x_i)^2$

regresní přímka

- chování Y (úmrtnost, mortality) co nejlépe (nejvíce) vysvětlit lineární závislostí na x (zeměpisná šířka, latitude)
- (naše představa, předpoklad:) každé zem. šířce odpovídá jakási střední úmrtnost, ta závisí na zeměpisné šířce lineárně

$$\mathrm{E}\, Y_i = \beta_0 + \beta_1 x_i, \qquad i = 1, \ldots, n$$

- parametry β_0, β_1 odhadneme **metodou nejmenších čtverců** minimalizací přes β_0, β_1 součtu čtverců „svislých“ odchylek

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

- výsledné minimum (pro $\beta_0 = b_0, \beta_1 = b_1$) nazveme **reziduální součet čtverců**, tj. $S_e = \sum_{i=1}^{n} (Y_i - b_0 - b_1 x_i)^2$

regresní přímka

- chování Y (úmrtnost, mortality) co nejlépe (nejvíce) vysvětlit lineární závislostí na x (zeměpisná šířka, latitude)
- (naše představa, předpoklad:) každé zem. šířce odpovídá jakási střední úmrtnost, ta závisí na zeměpisné šířce lineárně

$$\mathrm{E}\, Y_i = \beta_0 + \beta_1 x_i, \qquad i = 1, \ldots, n$$

- parametry β_0, β_1 odhadneme **metodou nejmenších čtverců** minimalizací přes β_0, β_1 součtu čtverců „svislých“ odchylek

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

- výsledné minimum (pro $\beta_0 = b_0, \beta_1 = b_1$) nazveme **reziduální součet čtverců**, tj. $S_e = \sum_{i=1}^{n} (Y_i - b_0 - b_1 x_i)^2$

regresní přímka

- chování Y (úmrtnost, mortality) co nejlépe (nejvíce) vysvětlit lineární závislostí na x (zeměpisná šířka, latitude)
- (naše představa, předpoklad:) každé zem. šířce odpovídá jakási střední úmrtnost, ta závisí na zeměpisné šířce lineárně

$$\mathrm{E}\, Y_i = \beta_0 + \beta_1 x_i, \qquad i = 1, \dots, n$$

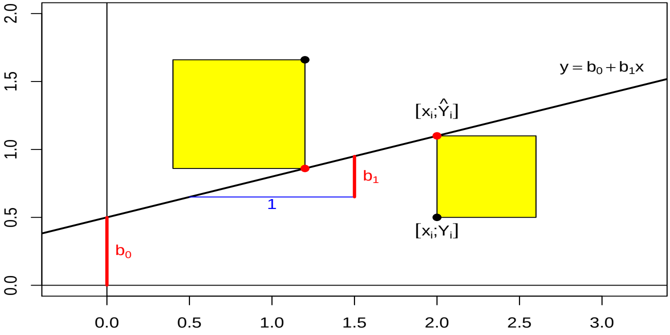
- parametry β_0, β_1 odhadneme **metodou nejmenších čtverců** minimalizací přes β_0, β_1 součtu čtverců „svislých" odchylek

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

- výsledné minimum (pro $\beta_0 = b_0, \beta_1 = b_1$) nazveme **reziduální součet čtverců**, tj. $S_e = \sum_{i=1}^{n} (Y_i - b_0 - b_1 x_i)^2$

metoda nejmenších čtverců

odhadovaná závislost:	$y = \beta_0 + \beta_1 \cdot x$	(populace)
odhad závislosti:	$y = b_0 + b_1 \cdot x$	(výběr)
celková plocha čtverců:	$S_e = \sum_{i=1}^{n} (Y_i - b_0 - b_1 x_i)^2$	(výběr)



náš příklad

[summary(lm(mortality~latitude))]

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	389,19	23,81	16,34	<0,001
latitude	– 5,98	0,60	– 9,99	<0,001

- odhad závislosti: $\widehat{\text{mortality}} = 389{,}19 - 5{,}98\ \text{latitude}$
- s každým stupněm sev. šířky klesá úmrtnost v průměru téměř o 6 osob na 10 000 000 obyvatel
- na rovníku by úmrtnost měla být 389 jednotek, ale je to extrapolace mimo rozmezí známých hodnot – sotva použitelné
- závislost je průkazná, neboť v řádku pro x (latitude) je p <0,001

náš příklad

[summary(lm(mortality∼latitude))]

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	389,19	23,81	16,34	<0,001
latitude	– 5,98	0,60	– 9,99	<0,001

- odhad závislosti: $\widehat{\text{mortality}} = 389{,}19 - 5{,}98$ latitude
- s každým stupněm sev. šířky klesá úmrtnost v průměru téměř o 6 osob na 10 000 000 obyvatel
- na rovníku by úmrtnost měla být 389 jednotek, ale je to extrapolace mimo rozmezí známých hodnot – sotva použitelné
- závislost je průkazná, neboť v řádku pro x (latitude) je p <0,001

náš příklad

[summary(lm(mortality~latitude))]

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	389,19	23,81	16,34	<0,001
latitude	– 5,98	0,60	– 9,99	<0,001

- odhad závislosti: $\widehat{\text{mortality}} = 389{,}19 - 5{,}98$ latitude
- s každým stupněm sev. šířky klesá úmrtnost v průměru téměř o 6 osob na 10 000 000 obyvatel
- na rovníku by úmrtnost měla být 389 jednotek, ale je to extrapolace mimo rozmezí známých hodnot – sotva použitelné
- závislost je průkazná, neboť v řádku pro x (latitude) je p <0,001

náš příklad

[summary(lm(mortality∼latitude))]

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	389,19	23,81	16,34	<0,001
latitude	– 5,98	0,60	– 9,99	<0,001

- odhad závislosti: $\widehat{\text{mortality}} = 389{,}19 - 5{,}98\ \text{latitude}$
- s každým stupněm sev. šířky klesá úmrtnost v průměru téměř o 6 osob na 10 000 000 obyvatel
- na rovníku by úmrtnost měla být 389 jednotek, ale je to extrapolace mimo rozmezí známých hodnot – sotva použitelné
- závislost je průkazná, neboť v řádku pro x (latitude) je $p < 0{,}001$

obecně

- odhadovaná závislost $y = \beta_0 + \beta_1 x$, odhadnutá $y = b_0 + b_1 x$
- závislost na x prokazujeme testováním hypotézy $H_0 : \beta_1 = 0$ (pak je y pro všechna x stejné, tedy $y = \beta_0$) pomocí

$$T = \frac{b_1}{\text{S.E.}(b_1)} = \frac{b_1}{s} \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

- zamítáme H_0 proti oboustr. alternativě, když $|T| \geq t_{n-2}(\alpha)$
- **reziduální součet čtverců – nevysvětlená variabilita Y**
  $S_e = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 x_i))^2$ reziduální součet čtverců
  $s^2 = S_e/(n-2)$ reziduální rozptyl
- **koeficient determinace** ukazuje, jaký **díl variability odezvy** (tj. $\sum_{i=1}^{n} (Y_i - \bar{Y})^2$) jsme závislostí vysvětlili

$$R^2 = 1 - \frac{S_e}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

obecně

- odhadovaná závislost $y = \beta_0 + \beta_1 x$, odhadnutá $y = b_0 + b_1 x$
- závislost na x prokazujeme testováním hypotézy $\mathrm{H}_0 : \beta_1 = 0$ (pak je y pro všechna x stejné, tedy $y = \beta_0$) pomocí

$$T = \frac{b_1}{\mathrm{S.E.}(b_1)} = \frac{b_1}{s}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- zamítáme H_0 proti oboustr. alternativě, když $|T| \geq t_{n-2}(\alpha)$
- **reziduální součet čtverců – nevysvětlená variabilita** Y
  $S_e = \sum_{i=1}^{n}(Y_i - (b_0 + b_1 x_i))^2$ reziduální součet čtverců
  $s^2 = S_e/(n-2)$ reziduální rozptyl
- **koeficient determinace** ukazuje, jaký **díl variability odezvy** (tj. $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$) jsme závislostí vysvětlili

$$R^2 = 1 - \frac{S_e}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

obecně

- odhadovaná závislost $y = \beta_0 + \beta_1 x$, odhadnutá $y = b_0 + b_1 x$
- závislost na x prokazujeme testováním hypotézy $H_0 : \beta_1 = 0$ (pak je y pro všechna x stejné, tedy $y = \beta_0$) pomocí

$$T = \frac{b_1}{\text{S.E.}(b_1)} = \frac{b_1}{s}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- zamítáme H_0 proti oboustr. alternativě, když $|T| \geq t_{n-2}(\alpha)$
- **reziduální součet čtverců – nevysvětlená variabilita Y**
  $S_e = \sum_{i=1}^{n}(Y_i - (b_0 + b_1 x_i))^2$ reziduální součet čtverců
  $s^2 = S_e/(n-2)$ reziduální rozptyl
- **koeficient determinace** ukazuje, jaký **díl variability odezvy** (tj. $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$) jsme závislostí vysvětlili

$$R^2 = 1 - \frac{S_e}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

obecně

- odhadovaná závislost $y = \beta_0 + \beta_1 x$, odhadnutá $y = b_0 + b_1 x$
- závislost na x prokazujeme testováním hypotézy $H_0 : \beta_1 = 0$ (pak je y pro všechna x stejné, tedy $y = \beta_0$) pomocí

$$T = \frac{b_1}{\mathrm{S.E.}(b_1)} = \frac{b_1}{s}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- zamítáme H_0 proti oboustr. alternativě, když $|T| \geq t_{n-2}(\alpha)$
- **reziduální součet čtverců – nevysvětlená variabilita** Y
  $S_e = \sum_{i=1}^{n}(Y_i - (b_0 + b_1 x_i))^2$ reziduální součet čtverců
  $s^2 = S_e/(n-2)$ reziduální rozptyl
- **koeficient determinace** ukazuje, jaký **díl variability odezvy** (tj. $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$) jsme závislostí vysvětlili

$$R^2 = 1 - \frac{S_e}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

obecně

- odhadovaná závislost $y = \beta_0 + \beta_1 x$, odhadnutá $y = b_0 + b_1 x$
- závislost na x prokazujeme testováním hypotézy $\mathrm{H}_0 : \beta_1 = 0$ (pak je y pro všechna x stejné, tedy $y = \beta_0$) pomocí

$$T = \frac{b_1}{\mathrm{S.E.}(b_1)} = \frac{b_1}{s}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- zamítáme H_0 proti oboustr. alternativě, když $|T| \geq t_{n-2}(\alpha)$
- **reziduální součet čtverců – nevysvětlená variabilita** Y
  $S_e = \sum_{i=1}^{n}(Y_i - (b_0 + b_1 x_i))^2$ reziduální součet čtverců
  $s^2 = S_e/(n-2)$ reziduální rozptyl
- **koeficient determinace** ukazuje, jaký **díl variability odezvy** (tj. $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$) jsme závislostí vysvětlili

$$R^2 = 1 - \frac{S_e}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

náš příklad a tabulka analýzy rozptylu

[anova(lm(mortality∼latitude))]

variabilita	st. vol. f	součet čtverců SS	prům. čtverec MS	F	p
model	1	36 464,20	36 464,20	99,797	<0,001
reziduální	47	17 173,07	365,38		
celkem	48	53 637,27			

- kolísání úmrtnosti vysvětlíme závislostí z 68 %, neboť je

$$R^2 = 1 - \frac{17173{,}07}{53637{,}27} = \frac{36464{,}20}{53637{,}27} = 0{,}680$$

interpretace

- odhad byl: $\widehat{\text{úmrtnost}}$=389,19-5,98· šířka
- na 30. stupni očekáváme úmrtnost: $389{,}19 - 5{,}98 \cdot 30 = 209{,}86$
- na 40. stupni očekáváme úmrtnost: $389{,}19 - 5{,}98 \cdot 40 = 150{,}08$
- přechod z 30. stupně na 40. stupeň znamená **v průměru** pokles o $10 \cdot 5{,}98 = 59{,}8$ úmrtí na 10 000 000 obyvatel
- pokusíme se predikci zlepšit přidáním další nezávisle proměnné

interpretace

- odhad byl: $\widehat{\text{úmrtnost}}$=389,19-5,98· šířka
- na 30. stupni očekáváme úmrtnost:
  $389{,}19 - 5{,}98 \cdot 30 = 209{,}86$
- na 40. stupni očekáváme úmrtnost:
  $389{,}19 - 5{,}98 \cdot 40 = 150{,}08$
- přechod z 30. stupně na 40. stupeň znamená **v průměru** pokles o $10 \cdot 5{,}98 = 59{,}8$ úmrtí na 10 000 000 obyvatel
- pokusíme se predikci zlepšit přidáním další nezávisle proměnné

interpretace

- odhad byl: $\widehat{\text{úmrtnost}}$=389,19-5,98· šířka
- na 30. stupni očekáváme úmrtnost:
  389,19 – 5,98 · 30 = 209,86
- na 40. stupni očekáváme úmrtnost:
  389,19 – 5,98 · 40 = 150,08
- přechod z 30. stupně na 40. stupeň znamená **v průměru** pokles o 10 · 5,98 = 59,8 úmrtí na 10 000 000 obyvatel
- pokusíme se predikci zlepšit přidáním další nezávisle proměnné

interpretace

- odhad byl: $\widehat{\text{úmrtnost}}$=389,19-5,98· šířka
- na 30. stupni očekáváme úmrtnost: $389{,}19 - 5{,}98 \cdot 30 = 209{,}86$
- na 40. stupni očekáváme úmrtnost: $389{,}19 - 5{,}98 \cdot 40 = 150{,}08$
- přechod z 30. stupně na 40. stupeň znamená **v průměru** pokles o $10 \cdot 5{,}98 = 59{,}8$ úmrtí na 10 000 000 obyvatel
- pokusíme se predikci zlepšit přidáním další nezávisle proměnné

interpretace

- odhad byl: $\widehat{\text{úmrtnost}}$=389,19-5,98· šířka
- na 30. stupni očekáváme úmrtnost: $389{,}19 - 5{,}98 \cdot 30 = 209{,}86$
- na 40. stupni očekáváme úmrtnost: $389{,}19 - 5{,}98 \cdot 40 = 150{,}08$
- přechod z 30. stupně na 40. stupeň znamená **v průměru** pokles o $10 \cdot 5{,}98 = 59{,}8$ úmrtí na 10 000 000 obyvatel
- pokusíme se predikci zlepšit přidáním další nezávisle proměnné

dva regresory

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	401,17	28,04	14,31	<0,001
latitude	– 5,93	0,60	– 9,82	<0,001
longitude	0,15	0,19	0,82	0,418

- pokusíme se přidat zeměpisnou délku
- není průkazné, že by koeficient u longitude byl nenulový (nezamítneme hypotézu, že koeficient je nulový)
- longitude nepřináší další informaci o mortality, kterou bychom už neměli ze známé hodnoty latitude
- $\Rightarrow$ není vhodné přidávat do modelu s latitude také longitude
- koeficient determinace R^2 =0,684 (původně 0,680)

dva regresory

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	401,17	28,04	14,31	<0,001
latitude	– 5,93	0,60	– 9,82	<0,001
longitude	0,15	0,19	0,82	0,418

- pokusíme se přidat zeměpisnou délku
- není průkazné, že by koeficient u longitude byl nenulový (nezamítneme hypotézu, že koeficient je nulový)
- longitude nepřináší další informaci o mortality, kterou bychom už neměli ze známé hodnoty latitude
- $\Rightarrow$ není vhodné přidávat do modelu s latitude také longitude
- koeficient determinace R^2 =0,684 (původně 0,680)

dva regresory

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	401,17	28,04	14,31	<0,001
latitude	– 5,93	0,60	– 9,82	<0,001
longitude	0,15	0,19	0,82	0,418

- pokusíme se přidat zeměpisnou délku
- není průkazné, že by koeficient u longitude byl nenulový (nezamítneme hypotézu, že koeficient je nulový)
- longitude nepřináší další informaci o mortality, kterou bychom už neměli ze známé hodnoty latitude
- $\Rightarrow$ není vhodné přidávat do modelu s latitude také longitude
- koeficient determinace R^2 =0,684 (původně 0,680)

dva regresory

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	401,17	28,04	14,31	<0,001
latitude	– 5,93	0,60	– 9,82	<0,001
longitude	0,15	0,19	0,82	0,418

- pokusíme se přidat zeměpisnou délku
- není průkazné, že by koeficient u longitude byl nenulový (nezamítneme hypotézu, že koeficient je nulový)
- longitude nepřináší další informaci o mortality, kterou bychom už neměli ze známé hodnoty latitude
- $\Rightarrow$ není vhodné přidávat do modelu s latitude také longitude
- koeficient determinace R^2 =0,684 (původně 0,680)

dva regresory

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	401,17	28,04	14,31	<0,001
latitude	– 5,93	0,60	– 9,82	<0,001
longitude	0,15	0,19	0,82	0,418

- pokusíme se přidat zeměpisnou délku
- není průkazné, že by koeficient u longitude byl nenulový (nezamítneme hypotézu, že koeficient je nulový)
- longitude nepřináší další informaci o mortality, kterou bychom už neměli ze známé hodnoty latitude
- $\Rightarrow$ není vhodné přidávat do modelu s latitude také longitude
- koeficient determinace R^2 =0,684 (původně 0,680)

podrobnější rozbor – vliv oceánu

- závislost jen pro vnitrozemské státy ($R^2 = 59{,}6$ %):
  [lm(mortality∼latitude,subset=Ocean==0)]

koef.	odhad	stř. chyba	*t*-stat.	*p*
abs. člen	360,55	36,70	9,82	<0,001
latitude	– 5,485	0,904	– 6,07	<0,001

- závislost jen pro přímořské státy ($R^2 = 78{,}6$ %):
  [lm(mortality∼latitude,subset=Ocean==1)]

koef.	odhad	stř. chyba	*t*-stat.	*p*
abs. člen	381,20	24,83	15,35	<0,001
latitude	– 5,491	0,640	– 8,58	<0,001

- směrnice jsou téměř stejné, abs. členy rozdílné
- v obou případech s každým stupněm sev. šířky klesá úmrtnost v průměru téměř o 5,5 osob na 10 000 000 obyvatel

podrobnější rozbor – vliv oceánu

- závislost jen pro vnitrozemské státy ($R^2 = 59{,}6\ \%$):
  [lm(mortality~latitude,subset=Ocean==0)]

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	360,55	36,70	9,82	<0,001
latitude	– 5,485	0,904	– 6,07	<0,001

- závislost jen pro přímořské státy ($R^2 = 78{,}6\ \%$):
  [lm(mortality~latitude,subset=Ocean==1)]

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	381,20	24,83	15,35	<0,001
latitude	– 5,491	0,640	– 8,58	<0,001

- směrnice jsou téměř stejné, abs. členy rozdílné
- v obou případech s každým stupněm sev. šířky klesá úmrtnost v průměru téměř o 5,5 osob na 10 000 000 obyvatel

podrobnější rozbor – vliv oceánu

- závislost jen pro vnitrozemské státy ($R^2 = 59{,}6\ \%$):
  [lm(mortality~latitude,subset=Ocean==0)]

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	360,55	36,70	9,82	<0,001
latitude	– 5,485	0,904	– 6,07	<0,001

- závislost jen pro přímořské státy ($R^2 = 78{,}6\ \%$):
  [lm(mortality~latitude,subset=Ocean==1)]

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	381,20	24,83	15,35	<0,001
latitude	– 5,491	0,640	– 8,58	<0,001

- směrnice jsou téměř stejné, abs. členy rozdílné
- v obou případech s každým stupněm sev. šířky klesá úmrtnost v průměru téměř o 5,5 osob na 10 000 000 obyvatel

podrobnější rozbor – vliv oceánu

- závislost jen pro vnitrozemské státy ($R^2 = 59{,}6\ \%$):
  [lm(mortality~latitude,subset=Ocean==0)]

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	360,55	36,70	9,82	<0,001
latitude	– 5,485	0,904	– 6,07	<0,001

- závislost jen pro přímořské státy ($R^2 = 78{,}6\ \%$):
  [lm(mortality~latitude,subset=Ocean==1)]

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	381,20	24,83	15,35	<0,001
latitude	– 5,491	0,640	– 8,58	<0,001

- směrnice jsou téměř stejné, abs. členy rozdílné
- v obou případech s každým stupněm sev. šířky klesá úmrtnost v průměru téměř o 5,5 osob na 10 000 000 obyvatel

společně vnitrozemské i přímořské státy

[summary(lm(mortality∼Ocean+latitude))]

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	360,69	21,50	16,78	<0,001
ocean	20,43	4,83	4,23	<0,001
latitude	– 5,49	0,53	– 10,44	<0,001

- koeficient determinace R^2=0,770
- při „stěhování“ z vnitrozemí k oceánu po rovnoběžce roste úmrtnost v průměru o 20 osob na 10 milionů obyvatel
- je to ekvivalentní vnitrozemskému stěhování o 20,43/5,49 = 3,72 stupňů na jih
- na každý stupeň stěhování na sever klesá úmrtnost o 5,5, pokud se nezmění vztah k oceánu

společně vnitrozemské i přímořské státy

[summary(lm(mortality∼Ocean+latitude))]

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	360,69	21,50	16,78	<0,001
ocean	20,43	4,83	4,23	<0,001
latitude	– 5,49	0,53	– 10,44	<0,001

- koeficient determinace R^2=0,770
- při „stěhování" z vnitrozemí k oceánu po rovnoběžce roste úmrtnost v průměru o 20 osob na 10 milionů obyvatel
- je to ekvivalentní vnitrozemskému stěhování o 20,43/5,49 = 3,72 stupňů na jih
- na každý stupeň stěhování na sever klesá úmrtnost o 5,5, pokud se nezmění vztah k oceánu

společně vnitrozemské i přímořské státy

[summary(lm(mortality~Ocean+latitude))]

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	360,69	21,50	16,78	<0,001
ocean	20,43	4,83	4,23	<0,001
latitude	– 5,49	0,53	– 10,44	<0,001

- koeficient determinace $R^2 = 0{,}770$
- při „stěhování“ z vnitrozemí k oceánu po rovnoběžce roste úmrtnost v průměru o 20 osob na 10 milionů obyvatel
- je to ekvivalentní vnitrozemskému stěhování o $20{,}43/5{,}49 = 3{,}72$ stupňů na jih
- na každý stupeň stěhování na sever klesá úmrtnost o 5,5, pokud se nezmění vztah k oceánu

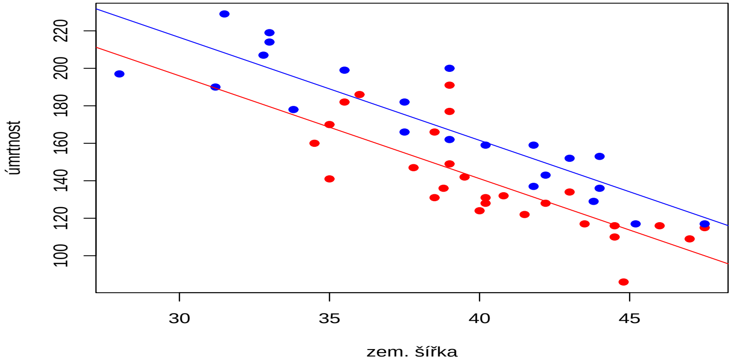
společně vnitrozemské i přímořské státy

[summary(lm(mortality~Ocean+latitude))]

koef.	odhad	stř. chyba	t-stat.	p
abs. člen	360,69	21,50	16,78	<0,001
ocean	20,43	4,83	4,23	<0,001
latitude	– 5,49	0,53	– 10,44	<0,001

- koeficient determinace R^2=0,770
- při „stěhování“ z vnitrozemí k oceánu po rovnoběžce roste úmrtnost v průměru o 20 osob na 10 milionů obyvatel
- je to ekvivalentní vnitrozemskému stěhování o 20,43/5,49 = 3,72 stupňů na jih
- na každý stupeň stěhování na sever klesá úmrtnost o 5,5, pokud se nezmění vztah k oceánu

příklad: souvisí úmrtnost s polohou?



- vnitrozemské státy: $y=360{,}69-5{,}49\ x$
  přímořské státy: $y=(360{,}69+20{,}43)-5{,}49\ x=381{,}12-5{,}49\ x$
- lze ověřit, že přímky mohou být rovnoběžné ($p=$99,6 %)

pozor na interpretaci odhadů (na dalším příkladu)

- závisí procento tuku dospělého muže na jeho výšce? pokud ano, tak s výškou roste nebo klesá?
- závisí na tom, jak se na úlohu díváme, co bereme v úvahu
- $\widehat{\text{fat}} = -47{,}68 + 0{,}341 \text{ height}$ $\quad R^2 = 11{,}8\ \%$
- $\widehat{\text{fat}} = 16{,}55 - 0{,}244 \text{ height} + 0{,}504 \text{ weight}$ $\quad R^2 = 71{,}4\ \%$
- ve všech případech jsou koeficienty u regresorů na 5% hladině průkazně nenulové
- rozdíl je v kvalitě vyrovnání, ale zejména v interpretaci
- průměrná změna procenta tuku při jednotkové změně výšky (a **nezměněné hmotnosti** pro druhý model)

pozor na interpretaci odhadů (na dalším příkladu)

- závisí procento tuku dospělého muže na jeho výšce? pokud ano, tak s výškou roste nebo klesá?
- závisí na tom, jak se na úlohu díváme, co bereme v úvahu
- $\widehat{\text{fat}} = -47{,}68 + 0{,}341\ \text{height}$ $\quad R^2 = 11{,}8\ \%$
- $\widehat{\text{fat}} = 16{,}55 - 0{,}244\ \text{height} + 0{,}504\ \text{weight}$ $\quad R^2 = 71{,}4\ \%$
- ve všech případech jsou koeficienty u regresorů na 5% hladině průkazně nenulové
- rozdíl je v kvalitě vyrovnání, ale zejména v interpretaci
- průměrná změna procenta tuku při jednotkové změně výšky (a **nezměněné hmotnosti** pro druhý model)

pozor na interpretaci odhadů (na dalším příkladu)

- závisí procento tuku dospělého muže na jeho výšce?
  pokud ano, tak s výškou roste nebo klesá?
- závisí na tom, jak se na úlohu díváme, co bereme v úvahu
- $\widehat{\text{fat}} = -47{,}68 + 0{,}341\ \text{height}$ $R^2 = 11{,}8\ \%$
- $\widehat{\text{fat}} = 16{,}55 - 0{,}244\ \text{height} + 0{,}504\ \text{weight}$ $R^2 = 71{,}4\ \%$
- ve všech případech jsou koeficienty u regresorů na 5% hladině průkazně nenulové
- rozdíl je v kvalitě vyrovnání, ale zejména v interpretaci
- průměrná změna procenta tuku při jednotkové změně výšky (a **nezměněné hmotnosti** pro druhý model)

pozor na interpretaci odhadů (na dalším příkladu)

- závisí procento tuku dospělého muže na jeho výšce? pokud ano, tak s výškou roste nebo klesá?
- závisí na tom, jak se na úlohu díváme, co bereme v úvahu
- $\widehat{\text{fat}} = -47{,}68 + 0{,}341\ \text{height}$ $\quad R^2 = 11{,}8\ \%$
- $\widehat{\text{fat}} = 16{,}55 - 0{,}244\ \text{height} + 0{,}504\ \text{weight}$ $\quad R^2 = 71{,}4\ \%$
- ve všech případech jsou koeficienty u regresorů na 5% hladině průkazně nenulové
- rozdíl je v kvalitě vyrovnání, ale zejména v interpretaci
- průměrná změna procenta tuku při jednotkové změně výšky (a **nezměněné hmotnosti** pro druhý model)

pozor na interpretaci odhadů (na dalším příkladu)

- závisí procento tuku dospělého muže na jeho výšce? pokud ano, tak s výškou roste nebo klesá?
- závisí na tom, jak se na úlohu díváme, co bereme v úvahu
- $\widehat{\text{fat}} = -\,47{,}68 + 0{,}341\ \text{height}$ $\qquad R^2 = 11{,}8\ \%$
- $\widehat{\text{fat}} = 16{,}55 - 0{,}244\ \text{height} + 0{,}504\ \text{weight}$ $\qquad R^2 = 71{,}4\ \%$
- ve všech případech jsou koeficienty u regresorů na 5% hladině průkazně nenulové
- rozdíl je v kvalitě vyrovnání, ale zejména v interpretaci
- průměrná změna procenta tuku při jednotkové změně výšky (a **nezměněné hmotnosti** pro druhý model)

pozor na interpretaci odhadů (na dalším příkladu)

- závisí procento tuku dospělého muže na jeho výšce?
  pokud ano, tak s výškou roste nebo klesá?
- závisí na tom, jak se na úlohu díváme, co bereme v úvahu
- $\widehat{\text{fat}} = -47{,}68 + 0{,}341\ \text{height}$ $R^2 = 11{,}8\ \%$
- $\widehat{\text{fat}} = 16{,}55 - 0{,}244\ \text{height} + 0{,}504\ \text{weight}$ $R^2 = 71{,}4\ \%$
- ve všech případech jsou koeficienty u regresorů na 5% hladině průkazně nenulové
- rozdíl je v kvalitě vyrovnání, ale zejména v interpretaci
- průměrná změna procenta tuku při jednotkové změně výšky (a **nezměněné hmotnosti** pro druhý model)

pozor na interpretaci odhadů (na dalším příkladu)

- závisí procento tuku dospělého muže na jeho výšce?
  pokud ano, tak s výškou roste nebo klesá?
- závisí na tom, jak se na úlohu díváme, co bereme v úvahu
- $\widehat{\text{fat}} = -\,47{,}68 + 0{,}341\ \text{height} \qquad R^2 = 11{,}8\ \%$
- $\widehat{\text{fat}} = 16{,}55 - 0{,}244\ \text{height} + 0{,}504\ \text{weight} \qquad R^2 = 71{,}4\ \%$
- ve všech případech jsou koeficienty u regresorů na 5% hladině průkazně nenulové
- rozdíl je v kvalitě vyrovnání, ale zejména v interpretaci
- průměrná změna procenta tuku při jednotkové změně výšky (a **nezměněné hmotnosti** pro druhý model)

regrese v MS Excelu 2000, 2003

	Excel 2000	označení
absolutní člen	Hranice	b_0
odhad	Koeficienty	b_i
střední chyba odhadu	Chyba střední hodnoty	S.E.(b_j)
koeficient (mnohonásobné) korelace	Násobné R	$\sqrt{R^2}$
koeficient determinace	Hodnota spolehlivosti R	R^2
adjustovaný koef. det.	Nastavená hodnota spol. R	R^2_{adj}
resid. směr. odchylka	Chyba střední hodnoty	s
počet pozorování	Pozorování	n
počet st. volnosti	Rozdíl	

regrese v MS Excelu 2000, 2003

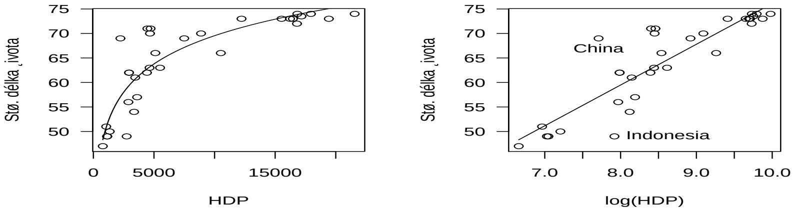
- Pozor na nabízený graf „Graf s rozdělením pravděpodobnosti“: obecně **nevypovídá** o normálním rozdělení, jak by asi chtěl, bylo by třeba použít místo vysvětlované veličiny některá z reziduí
- Nabízená „Normovaná rezidua“ jsou v regresi zcela nestandardní (z-skóry běžných reziduí)

regrese v MS Excelu 2000, 2003

- Pozor na nabízený graf „Graf s rozdělením pravděpodobnosti“: obecně **nevypovídá** o normálním rozdělení, jak by asi chtěl, bylo by třeba použít místo vysvětlované veličiny některá z reziduí
- Nabízená „Normovaná rezidua“ jsou v regresi zcela nestandardní (z-skóry běžných reziduí)

praktické problémy: transformace

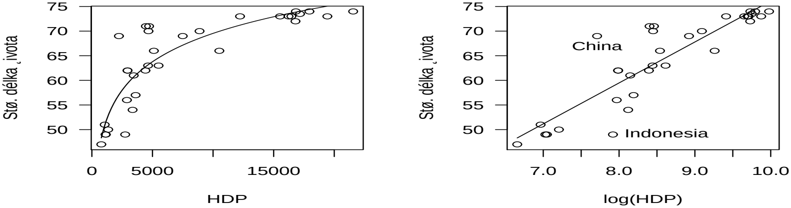
střední délka života ∼ HDP (rok 1992, 33 skupin zemí z celého světa)



- v původním měřítku závislost nelineární
- logaritmování HDP hodně pomohlo, ale ještě jistě jiné vlivy
- log(HDP) vysvětlí téměř 79 % variability střední délky života
- lze identifikovat státy, které se zvlášť vymykají

praktické problémy: transformace

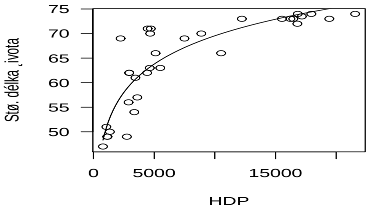
střední délka života $\sim$ HDP (rok 1992, 33 skupin zemí z celého světa)

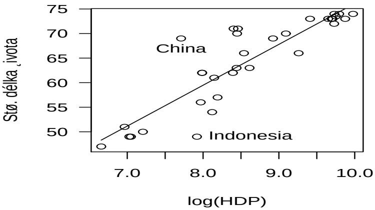


- v původním měřítku závislost nelineární
- logaritmování HDP hodně pomohlo, ale ještě jistě jiné vlivy
- log(HDP) vysvětlí téměř 79 % variability střední délky života
- lze identifikovat státy, které se zvlášť vymykají

praktické problémy: transformace

střední délka života ∼ HDP (rok 1992, 33 skupin zemí z celého světa)

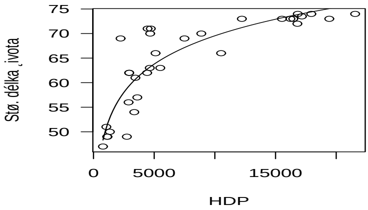


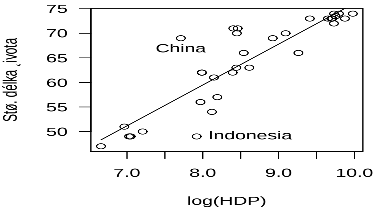


- v původním měřítku závislost nelineární
- logaritmování HDP hodně pomohlo, ale ještě jistě jiné vlivy
- log(HDP) vysvětlí téměř 79 % variability střední délky života
- lze identifikovat státy, které se zvlášť vymykají

praktické problémy: transformace

střední délka života ∼ HDP (rok 1992, 33 skupin zemí z celého světa)

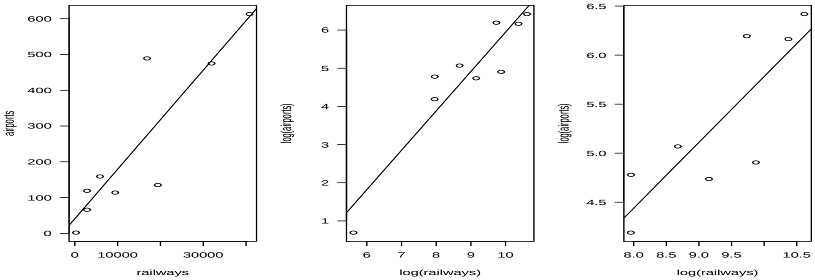




- v původním měřítku závislost nelineární
- logaritmování HDP hodně pomohlo, ale ještě jistě jiné vlivy
- log(HDP) vysvětlí téměř 79 % variability střední délky života
- lze identifikovat státy, které se zvlášť vymykají

praktické problémy: zdánlivá závislost

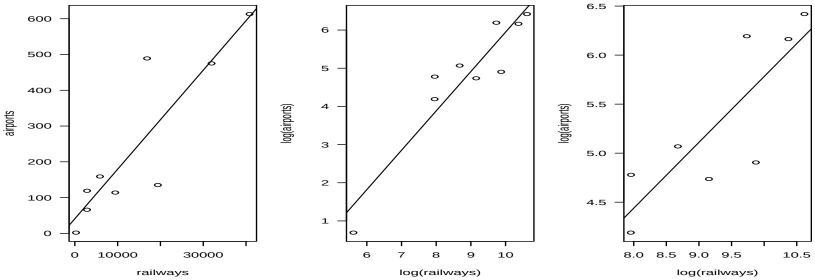
počet letišť ∼ délka železnic v Evropě



- v původním měřítku: $R^2 = 78\ \%$, $p = 0{,}2\ \%$
- v logaritmickém měřítku: $R^2 = 66\ \%$, $p = 0{,}02\ \%$
- logaritmické měřítko, bez Lucemburska: $R^2 = 69\ \%$, $p = 1\ \%$

praktické problémy: zdánlivá závislost

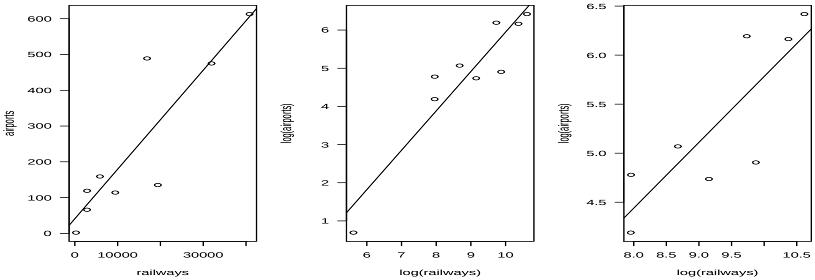
počet letišť ∼ délka železnic v Evropě



- v původním měřítku: $R^2 = 78\ \%$, $p = 0{,}2\ \%$
- v logaritmickém měřítku: $R^2 = 66\ \%$, $p = 0{,}02\ \%$
- logaritmické měřítko, bez Lucemburska: $R^2 = 69\ \%$, $p = 1\ \%$

praktické problémy: zdánlivá závislost

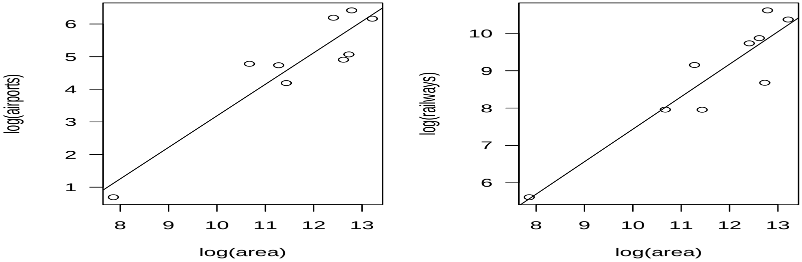
počet letišť $\sim$ délka železnic v Evropě



- v původním měřítku: $R^2 = 78\ \%$, $p = 0{,}2\ \%$
- v logaritmickém měřítku: $R^2 = 66\ \%$, $p = 0{,}02\ \%$
- logaritmické měřítko, bez Lucemburska: $R^2 = 69\ \%$, $p = 1\ \%$

praktické problémy: zdánlivá závislost

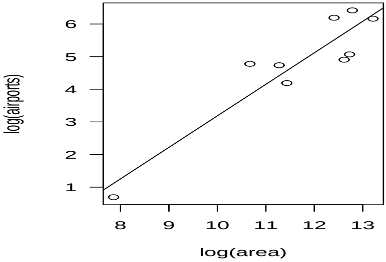
počet letišť ∼ délka železnic v Evropě

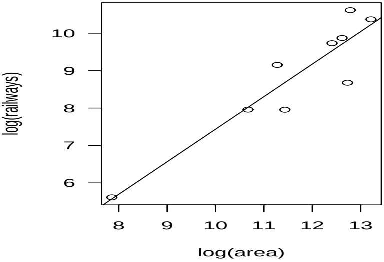


- počet letišť i délka železnic souvisí s velikostí země
- u letišť: $R^2 = 86$ %, $p = 0{,}03$ %
- u železnic: $R^2 = 64$ %, $p = 0{,}03$ %

praktické problémy: zdánlivá závislost

počet letišť ∼ délka železnic v Evropě

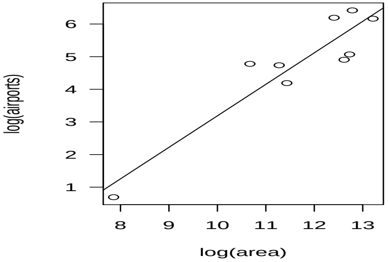


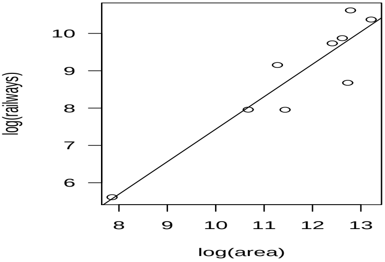


- počet letišť i délka železnic souvisí s velikostí země
- u letišť: $R^2 = 86\ \%$, $p = 0{,}03\ \%$
- u železnic: $R^2 = 64\ \%$, $p = 0{,}03\ \%$

praktické problémy: zdánlivá závislost

počet letišť ∼ délka železnic v Evropě

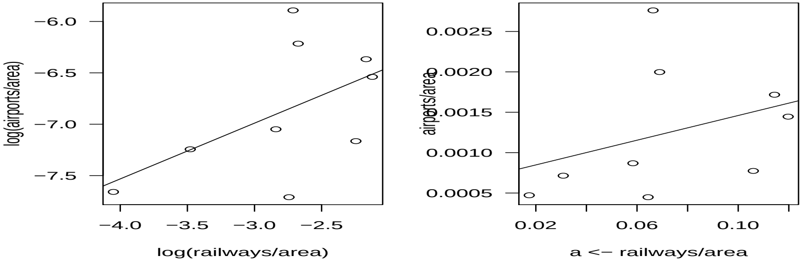




- počet letišť i délka železnic souvisí s velikostí země
- u letišť: $R^2 = 86\ \%$, $p = 0{,}03\ \%$
- u železnic: $R^2 = 64\ \%$, $p = 0{,}03\ \%$

praktické problémy: zdánlivá závislost

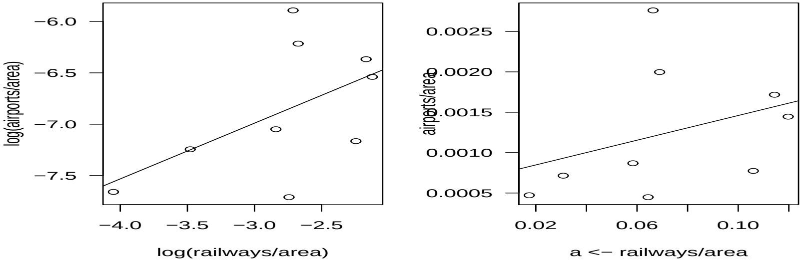
počet letišť a délka železnic ∼ plocha



- závislost v logaritmech: $R^2 = 28\ \%$, $p = 14\ \%$
- závislost v původním měřítku: $R^2 = 12\ \%$, $p = 36\ \%$
- relativní počet letišť nesouvisí s relativní délkou železnic

praktické problémy: zdánlivá závislost

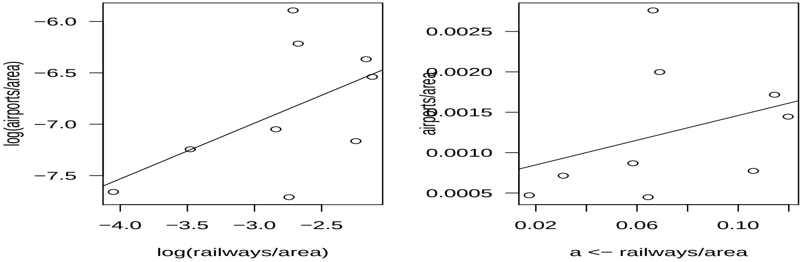
počet letišť a délka železnic $\sim$ plocha



- závislost v logaritmech: $R^2 = 28\ \%$, $p = 14\ \%$
- závislost v původním měřítku: $R^2 = 12\ \%$, $p = 36\ \%$
- relativní počet letišť nesouvisí s relativní délkou železnic

praktické problémy: zdánlivá závislost

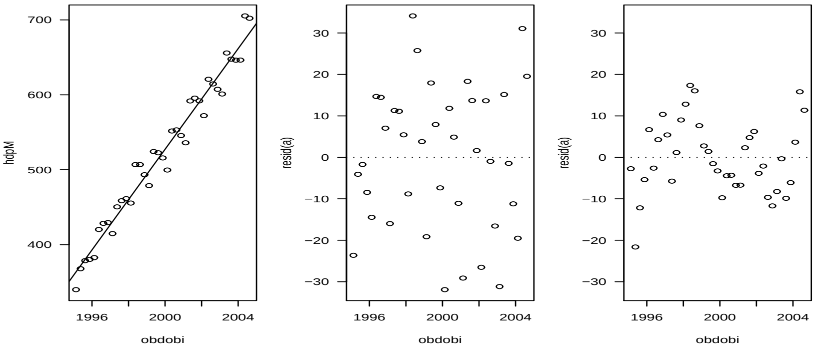
počet letišť a délka železnic ~ plocha



- závislost v logaritmech: $R^2 = 28\ \%$, $p = 14\ \%$
- závislost v původním měřítku: $R^2 = 12\ \%$, $p = 36\ \%$
- relativní počet letišť nesouvisí s relativní délkou železnic

praktické problémy: časová řada

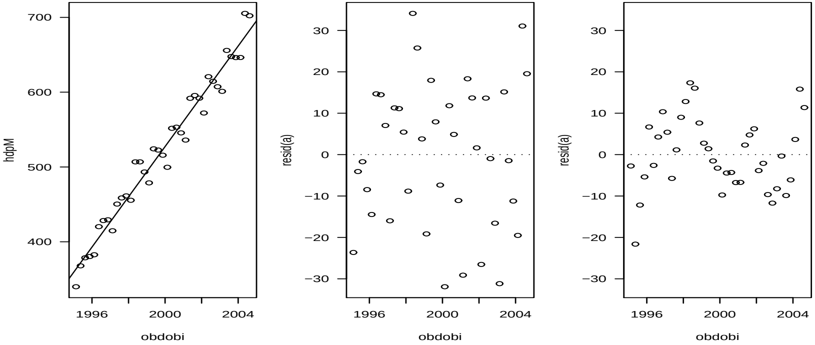
vývoj HDP v ČR – pozorování tvoří časovou řadu



- po sobě jsoucí pozorování nejsou nezávislá
- je patrný vliv čtvrtletí (rezidua vpravo)
- na pravém grafu patrný vliv „balíčku“

praktické problémy: časová řada

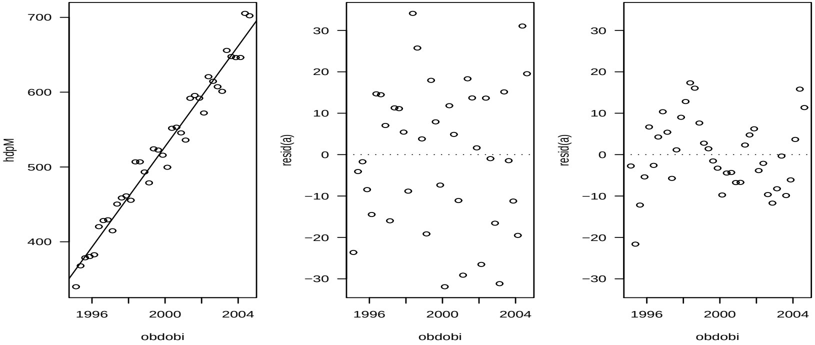
vývoj HDP v ČR – pozorování tvoří časovou řadu



- po sobě jsoucí pozorování nejsou nezávislá
- je patrný vliv čtvrtletí (rezidua vpravo)
- na pravém grafu patrný vliv „balíčku“

praktické problémy: časová řada

vývoj HDP v ČR – pozorování tvoří časovou řadu



- po sobě jsoucí pozorování nejsou nezávislá
- je patrný vliv čtvrtletí (rezidua vpravo)
- na pravém grafu patrný vliv „balíčku“