Pravděpodobnostní metody a matematická statistika

Prof. RNDr. Gejza Dohnal, CSc.

Pravděpodobnostní metody a matematická statistika

Prof. RNDr. Gejza Dohnal, CSc.

XIV. Úvod do lineární regrese

https://sms.nipax.cz/pas

Regresní model

Regresní model

Lineární regresní model

Závislost mezi nezávisle proměnnou a závisle proměnnou: $y=f(x)$

Lineární regresní model

Závislost mezi nezávisle proměnnou a závisle proměnnou: $y=f(x)$

- funkční závislost: x a y jsou nenáhodné, pro jedno x je nejvýše jedna hodnota y

Lineární regresní model

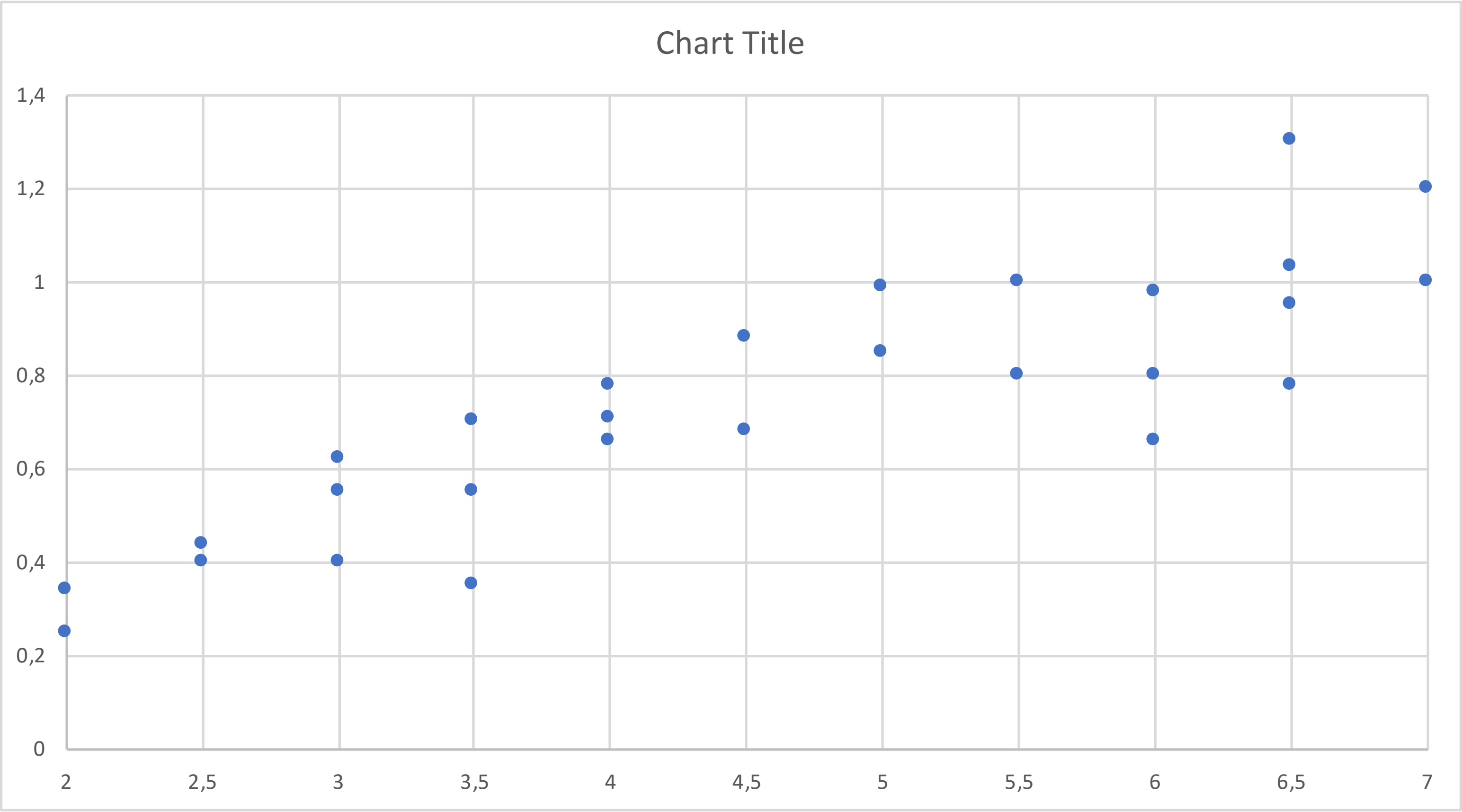
Závislost mezi nezávisle proměnnou a závisle proměnnou: $y=f(x)$

- funkční závislost: x a y jsou nenáhodné, pro jedno x je nejvýše jedna hodnota y
- regresní závislost: y je realizace náhodné veličiny Y při konkrétním x; pro jedno x můžeme pozorovat různé hodnoty Y

Lineární regresní model

Závislost mezi nezávisle proměnnou a závisle proměnnou: $y=f(x)$

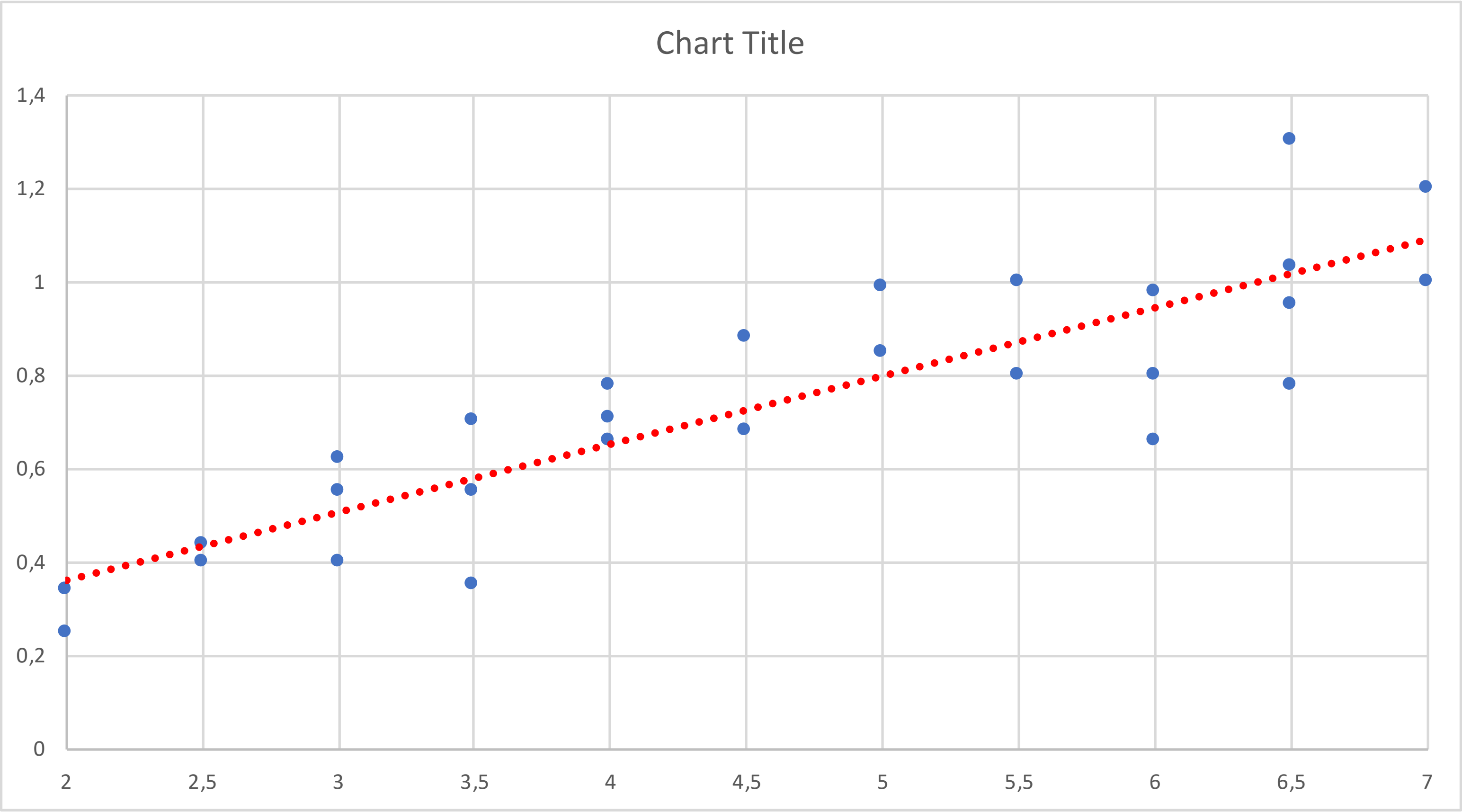
- funkční závislost: x a y jsou nenáhodné, pro jedno x je nejvýše jedna hodnota y
- regresní závislost: y je realizace náhodné veličiny Y při konkrétním x; pro jedno x můžeme pozorovat různé hodnoty Y



Lineární regresní model

Závislost mezi nezávisle proměnnou a závisle proměnnou: $y=f(x)$

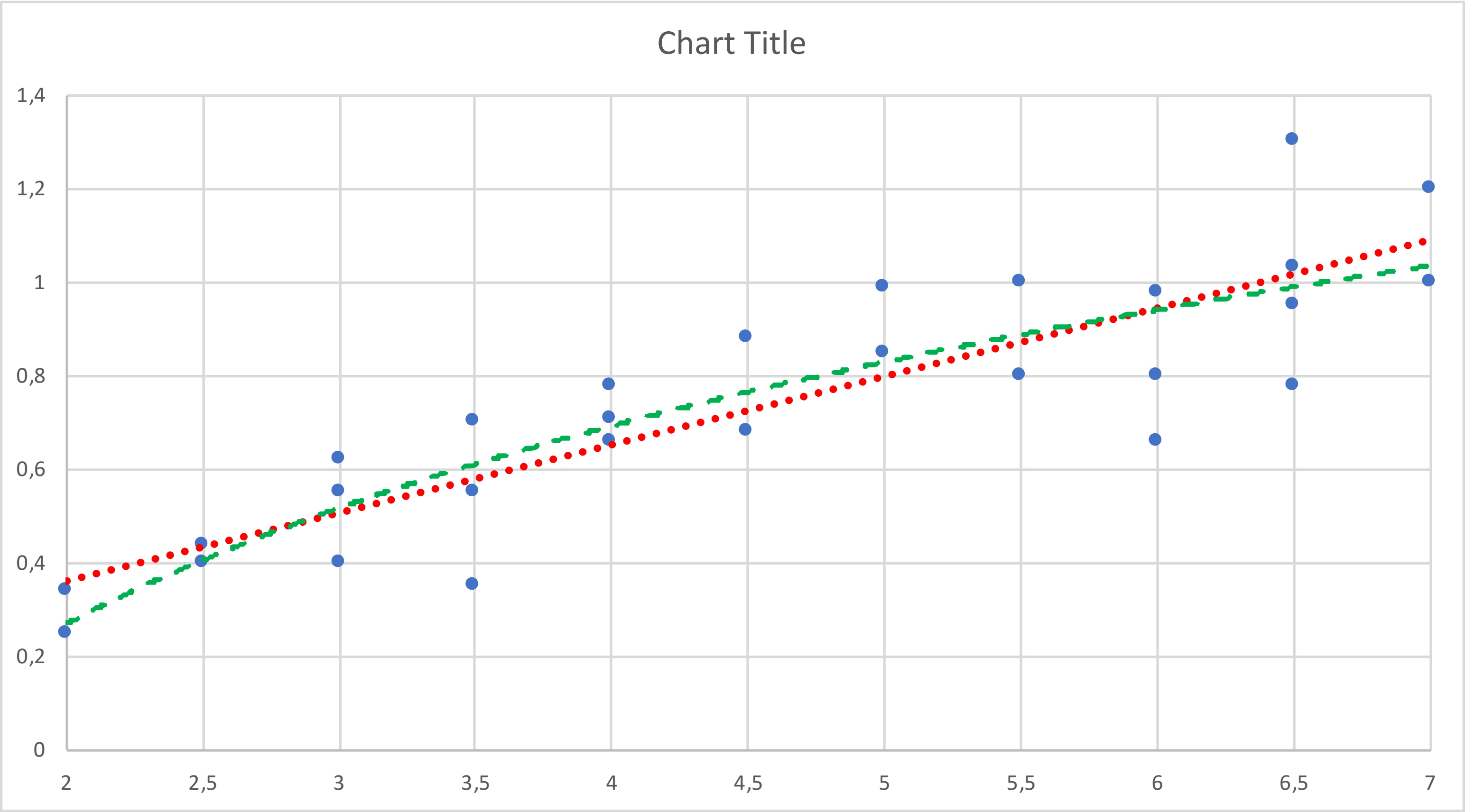
- funkční závislost: x a y jsou nenáhodné, pro jedno x je nejvýše jedna hodnota y
- regresní závislost: y je realizace náhodné veličiny Y při konkrétním x; pro jedno x můžeme pozorovat různé hodnoty Y



Lineární regresní model

Závislost mezi nezávisle proměnnou a závisle proměnnou: y=f(x)

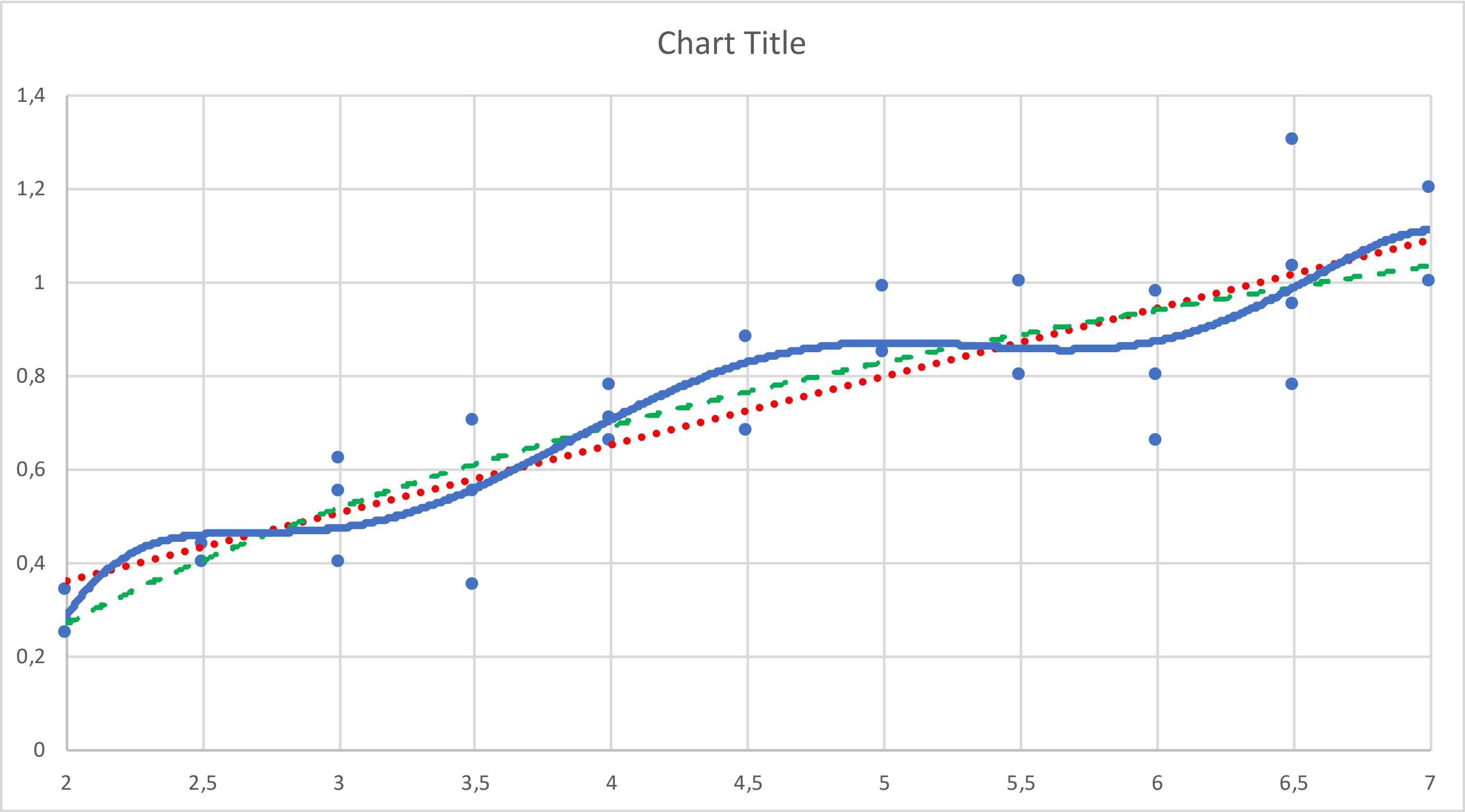
- funkční závislost: x a y jsou nenáhodné, pro jedno x je nejvýše jedna hodnota y
- regresní závislost: y je realizace náhodné veličiny Y při konkrétním x; pro jedno x můžeme pozorovat různé hodnoty Y



Lineární regresní model

Závislost mezi nezávisle proměnnou a závisle proměnnou: y=f(x)

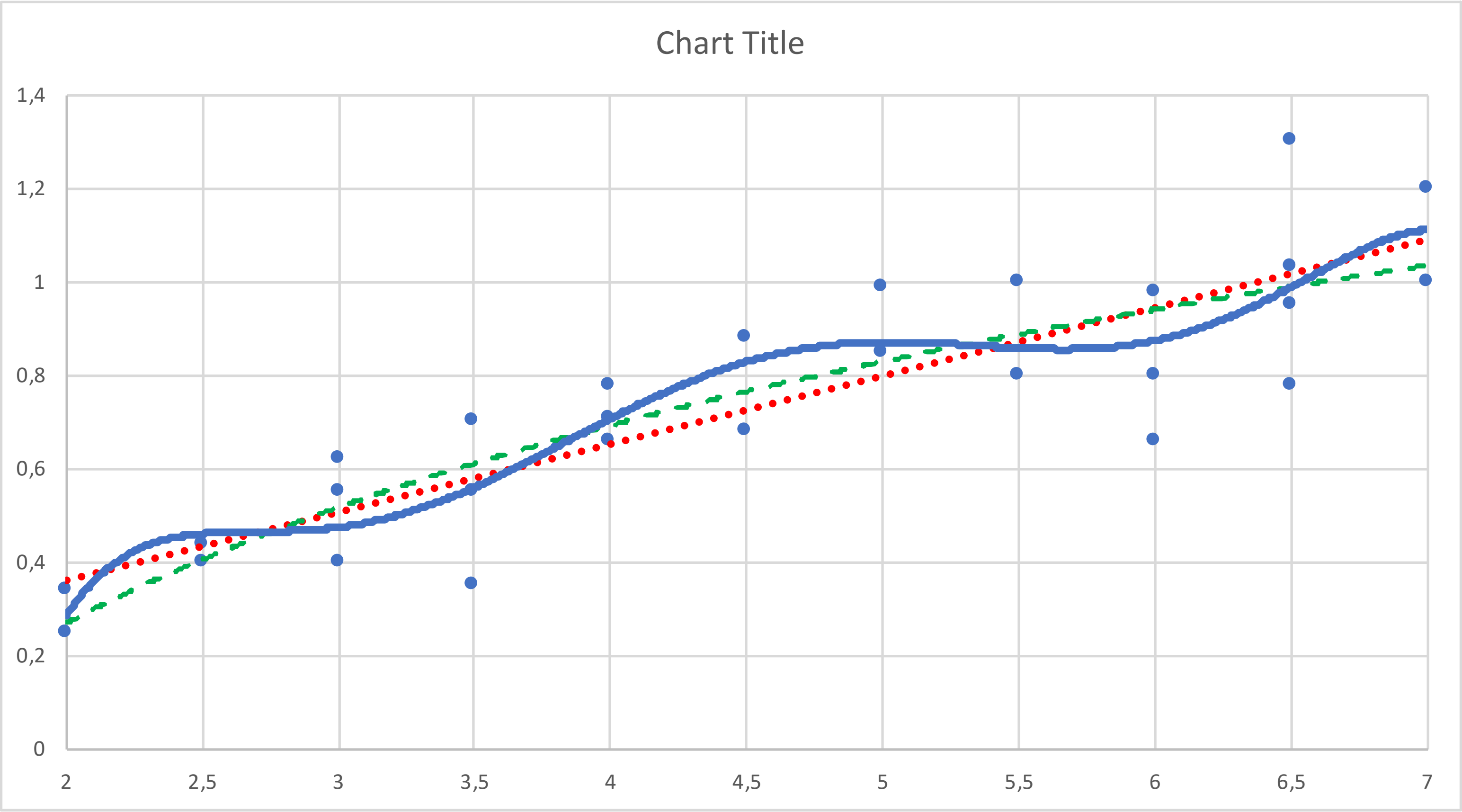
- funkční závislost: x a y jsou nenáhodné, pro jedno x je nejvýše jedna hodnota y
- regresní závislost: y je realizace náhodné veličiny Y při konkrétním x; pro jedno x můžeme pozorovat různé hodnoty Y



Lineární regresní model

Závislost mezi nezávisle proměnnou a závisle proměnnou: $y=f(x)$

- funkční závislost: x a y jsou nenáhodné, pro jedno x je nejvýše jedna hodnota y
- regresní závislost: y je realizace náhodné veličiny Y při konkrétním x; pro jedno x můžeme pozorovat různé hodnoty Y



Lineární regresní model

Lineární model: $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

Lineární regresní model

Lineární model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix} \quad \vec{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

Lineární regresní model

Lineární model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$

$$\vec{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix} \quad \vec{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

$$\mathbf{Y} = \mathbf{X}\vec{\beta} + \vec{e}$$

Lineární regresní model

Lineární model: $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

$$\vec{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix} \quad \vec{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

$$\mathbf{Y} = \mathbf{X}\vec{\beta} + \vec{e}$$

Minimalizujeme ztrátovou funkci: $$S(\alpha, \beta) = \sum_{i=1}^{n} \big(y_i - f(x_i; \alpha, \beta)\big)^2$$

Lineární regresní model

Lineární model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$

$$\vec{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix} \; \vec{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

$$\mathbf{Y} = \mathbf{X}\vec{\beta} + \vec{e}$$

Minimalizujeme ztrátovou funkci: $S(\alpha, \beta) = \sum_{i=1}^{n} \big(y_i - f(x_i; \alpha, \beta)\big)^2$

v příadě $f(x_i; \alpha, \beta) = \alpha + \beta . x$: $S(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$

$$S = (\mathbf{Y} - \mathbf{X}\vec{\beta})'(\mathbf{Y} - \mathbf{X}\vec{\beta})$$

Lineární regresní model

Lineární model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$

$$\vec{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix} \qquad \vec{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

$$\mathbf{Y} = \mathbf{X}\vec{\beta} + \vec{e}$$

Minimalizujeme ztrátovou funkci: $S(\alpha, \beta) = \sum_{i=1}^{n} \big(y_i - f(x_i; \alpha, \beta)\big)^2$

v příadě $f(x_i; \alpha, \beta) = \alpha + \beta . x$: $S(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$

$$S = (\mathbf{Y} - \mathbf{X}\vec{\beta})'(\mathbf{Y} - \mathbf{X}\vec{\beta})$$

Metoda nejmenších čtverců

Minimalizujeme ztrátovou funkci: $$S(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Metoda nejmenších čtverců

Minimalizujeme ztrátovou funkci: $S(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$

$$\frac{\partial S}{\partial \alpha} = -2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0$$

$$\frac{\partial S}{\partial \beta} = -2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) x_i = 0$$

Metoda nejmenších čtverců

Minimalizujeme ztrátovou funkci: $S(\alpha,\beta) = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$

$$\frac{\partial S}{\partial \alpha} = -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0$$

$$\frac{\partial S}{\partial \beta} = -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)x_i = 0$$

což vede k tzv. *normální soustavě rovnic*

$$n\alpha + \beta\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\alpha\sum_{i=1}^{n} x_i + \beta\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Metoda nejmenších čtverců

Minimalizujeme ztrátovou funkci: $S(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$

$$\frac{\partial S}{\partial \alpha} = -2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0$$

$$\frac{\partial S}{\partial \beta} = -2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) x_i = 0$$

což vede k tzv. *normální soustavě rovnic* a k jejímu řešení:

$$n\alpha + \beta \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

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

Metoda nejmenších čtverců

Minimalizujeme ztrátovou funkci: $$S(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

$$\frac{\partial S}{\partial \alpha} = -2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0$$

$$\frac{\partial S}{\partial \beta} = -2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) x_i = 0$$

což vede k tzv. *normální soustavě rovnic* a k jejímu řešení:

$$n\alpha + \beta \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

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

Lineární regresní model

$$\mathbf{Y} = \mathbf{X}\vec{\beta} + \vec{e}$$

$$S = (\mathbf{Y} - \mathbf{X}\vec{\beta})'(\mathbf{Y} - \mathbf{X}\vec{\beta})$$

$$\vec{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{Y} \qquad E\vec{b} = \vec{\beta} \qquad \mathrm{Var}\vec{b} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

$$S_e = (\mathbf{Y} - \mathbf{X}\vec{b})'(\mathbf{Y} - \mathbf{X}\vec{b}) = \mathbf{Y}'\mathbf{Y} - \vec{b}'\mathbf{X}'\mathbf{Y} \qquad s^2 = \frac{1}{n-k} S_e$$

Podmínky pro použití lineárního modelu

- Nezávislost pozorování: náhodné veličiny Y_i jsou navzájem stochasticky nezávislé
- Stejné rozptyly (homoskedasticita): rozptyl vysvětlované náhodné veličiny Y nezávisí na hodnotách vysvětlující veličiny $\boldsymbol{X}$
- Normalita dat: vysvětlovaná náhodná veličina Y má normální rozdělení

Lineární regresní model

$$\mathbf{Y} = \mathbf{X}\vec{\beta} + \vec{e}$$

$$S = (\mathbf{Y} - \mathbf{X}\vec{\beta})'(\mathbf{Y} - \mathbf{X}\vec{\beta})$$

$$\vec{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{Y} \qquad E\vec{b} = \vec{\beta} \qquad \mathrm{Var}\vec{b} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

$$S_e = (\mathbf{Y} - \mathbf{X}\vec{b})'(\mathbf{Y} - \mathbf{X}\vec{b}) = \mathbf{Y}'\mathbf{Y} - \vec{b}'\mathbf{X}'\mathbf{Y} \qquad s^2 = \frac{1}{n-k}S_e$$

Podmínky pro použití lineárního modelu

- Nezávislost pozorování: náhodné veličiny Y_i jsou navzájem stochasticky nezávislé
- Stejné rozptyly (homoskedasticita): rozptyl vysvětlované náhodné veličiny Y nezávisí na hodnotách vysvětlující veličiny $\boldsymbol{X}$
- Normalita dat: vysvětlovaná náhodná veličina Y má normální rozdělení

Testy významnosti koeficientů regrese

Testování významnosti koeficientů (přímkové) lineární regrese:

Testy významnosti koeficientů regrese

Testování významnosti koeficientů (přímkové) lineární regrese:

- reziduální součet čtverců:

$$S_R = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Testy významnosti koeficientů regrese

Testování významnosti koeficientů (přímkové) lineární regrese:

- reziduální součet čtverců:

$$S_R = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 = \sum_{i=1}^{n} y_i^2 - a \sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i y_i$$

Testy významnosti koeficientů regrese

Testování významnosti koeficientů (přímkové) lineární regrese:

- reziduální součet čtverců:

$$S_R = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 = \sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i y_i$$

- intervalový odhad směrnice:

$$b - s_b.t_{1-\gamma/2}(n-2) \leq \beta \leq b + s_b.t_{1-\gamma/2}(n-2)$$

Testy významnosti koeficientů regrese

Testování významnosti koeficientů (přímkové) lineární regrese:

- reziduální součet čtverců:

$$S_R = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - a - bx_i)^2 = \sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i y_i$$

- intervalový odhad směrnice:

$$b - s_b.t_{1-\gamma/2}(n-2) \leq \beta \leq b + s_b.t_{1-\gamma/2}(n-2)$$

$$s^2 = \frac{1}{n-2}S_R \qquad s_b^2 = \frac{s^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \qquad s_a^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} s_b^2$$

Testy významnosti koeficientů regrese

Testování významnosti koeficientů (přímkové) lineární regrese:

- reziduální součet čtverců:

$$S_R = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 = \sum_{i=1}^{n} y_i^2 - a \sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i y_i$$

- intervalový odhad směrnice:

$$b - s_b.t_{1-\gamma/2}(n-2) \leq \beta \leq b + s_b.t_{1-\gamma/2}(n-2)$$

$$s^2 = \frac{1}{n-2} S_R \qquad s_b^2 = \frac{s^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \qquad s_a^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} s_b^2$$

- test významnosti regresních koeficientů:

$$|T_a| = \left|\frac{a}{s_a}\right| \geq t_{1-\gamma/2}(n-2) \qquad |T_b| = \left|\frac{b}{s_b}\right| \geq t_{1-\gamma/2}(n-2)$$

Testy významnosti koeficientů regrese

Testování významnosti koeficientů (přímkové) lineární regrese:

- reziduální součet čtverců:

```
Regression Analysis - Linear model: Y = a + b*X
Regresní analýza - Lineární regrese
-----------------------------------------------------------------------------
Dependent variable (Závislé proměnná): Doba opravy
Independent variable (Nezávisle proměnná): Počet kalkulátorů
-----------------------------------------------------------------------------
                                  Estimate     Standard Error        T Statistic    P-Value
Parameter                          Odhad       Směrodatná odchylka
-----------------------------------------------------------------------------
Intercept (absolutní člen, b0)    -2,32215     2,56435               -0,905549      0,3786
Slope (směrnice, b1)              14,7383      0,519257              28,3834        0,0000
-----------------------------------------------------------------------------
```

- test významnosti regresních koeficientů:

$$|T_a| = |\frac{a}{s_a}| \geq t_{1-\gamma/2}(n-2) \qquad |T_b| = |\frac{b}{s_b}| \geq t_{1-\gamma/2}(n-2)$$

Testy významnosti koeficientů regrese

Testování významnosti koeficientů (přímkové) lineární regrese:

- reziduální součet čtverců:

$$S_R = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - a - bx_i)^2 = \sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i y_i$$

- intervalový odhad směrnice:

$$b - s_b.t_{1-\gamma/2}(n-2) \leq \beta \leq b + s_b.t_{1-\gamma/2}(n-2)$$

$$s^2 = \frac{1}{n-2} S_R \qquad s_b^2 = \frac{s^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \qquad s_a^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} s_b^2$$

- test významnosti regresních koeficientů:

$$|T_a| = \left|\frac{a}{s_a}\right| \geq t_{1-\gamma/2}(n-2) \qquad |T_b| = \left|\frac{b}{s_b}\right| \geq t_{1-\gamma/2}(n-2)$$

- koeficient determinace R^2:

$$R^2 = \frac{\sum_{i=1}^{n}(a + bx_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Testy významnosti koeficientů regrese

Testování významnosti koeficientů (přímkové) lineární regrese:

- reziduální součet čtverců:

$$S_R = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 = \sum_{i=1}^{n} y_i^2 - a \sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i y_i$$

- intervalový odhad směrnice:

$$b - s_b . t_{1-\gamma/2}(n-2) \leq \beta \leq b + s_b . t_{1-\gamma/2}(n-2)$$

$$s^2 = \frac{1}{n-2} S_R \qquad s_b^2 = \frac{s^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \qquad s_a^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} s_b^2$$

- test významnosti regresních koeficientů:

$$|T_a| = \left|\frac{a}{s_a}\right| \geq t_{1-\gamma/2}(n-2) \qquad |T_b| = \left|\frac{b}{s_b}\right| \geq t_{1-\gamma/2}(n-2)$$

- koeficient determinace R^2:

$$R^2 = \frac{\sum_{i=1}^{n} (a + bx_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{S_R}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Testy významnosti koeficientů regrese

Testování významnosti koeficientů (přímkové) lineární regrese:

- reziduální součet čtverců:

$$S_R = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 = \sum_{i=1}^{n} y_i^2 - a \sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i y_i$$

- intervalový odhad směrnice:

$$b - s_b . t_{1-\gamma/2}(n-2) \leq \beta \leq b + s_b . t_{1-\gamma/2}(n-2)$$

$$s^2 = \frac{1}{n-2} S_R \qquad s_b^2 = \frac{s^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \qquad s_a^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} s_b^2$$

- test významnosti regresních koeficientů:

$$|T_a| = \left|\frac{a}{s_a}\right| \geq t_{1-\gamma/2}(n-2) \qquad |T_b| = \left|\frac{b}{s_b}\right| \geq t_{1-\gamma/2}(n-2)$$

- koeficient determinace R^2:

$$R^2 = \frac{\sum_{i=1}^{n} (a + bx_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{S_R}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Testy významnosti regresního modelu

Testování významnosti regresního modelu:

zde se testuje nulová hypotéza H_0: $\beta_1 = \beta_1 = \ldots = \beta_k$ proti alternativní hypotéze H_A: $\beta_j = 0$ pro alespoň jedno $j = 1,2,\ldots k$.

K tomu se používá metoda ANOVA s testovou statistikou

$$F = \frac{\frac{S_T}{k-1}}{\frac{S_R}{n-k}}$$

která má F-rozdělení s k-1 a n-k stupni volnosti.

Zde $$S_R = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \hat{y} = \mathbf{X}.\vec{b}$$

$$S_T = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \qquad \bar{y}_i = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Testy významnosti regresního modelu

Testování významnosti regresního modelu:

zde se testuje nulová hypotéza H_0: $\beta_1 = \beta_1 = \ldots = \beta_k$ proti alternativní hypotéze H_A: $\beta_j = 0$ pro alespoň jedno $j = 1,2,\ldots k$.

Zde

$$S_R = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad \hat{y} = \mathbf{X}.\vec{b}$$

$$S_T = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad \bar{y}_i = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Testy významnosti regresního modelu

Testování významnosti regresního modelu:

zde se testuje nulová hypotéza H_0: $\beta_1 = \beta_1 = \ldots = \beta_k$ proti alternativní hypotéze H_A: $\beta_j = 0$ pro alespoň jedno $j = 1,2,\ldots k$.

Testy významnosti regresního modelu

Testování významnosti regresního modelu:

zde se testuje nulová hypotéza H_0: $\beta_1 = \beta_1 = \ldots = \beta_k$ proti alternativní hypotéze H_A: $\beta_j = 0$ pro alespoň jedno $j = 1,2,\ldots k$.

Testy významnosti regresního modelu

Vyhodnocení významnosti regresního modelu:

Testy významnosti regresního modelu

Vyhodnocení významnosti regresního modelu:

- Jsou-li celkový F-test i všechny t-testy statisticky významné, model se považuje za vhodný k vystižení variability proměnné Y (to však ještě neznamená, že je model správně navržen).

Testy významnosti regresního modelu

Vyhodnocení významnosti regresního modelu:

- Jsou-li celkový F-test i všechny t-testy statisticky významné, model se považuje za vhodný k vystižení variability proměnné Y (to však ještě neznamená, že je model správně navržen).
- Jsou-li celkový F-test i všechny t-testy statisticky nevýznamné, model se považuje za nevhodný, protože nevystihuje variabilitu proměnné Y .

Testy významnosti regresního modelu

Vyhodnocení významnosti regresního modelu:

- Jsou-li celkový F-test i všechny t-testy statisticky významné, model se považuje za vhodný k vystižení variability proměnné Y (to však ještě neznamená, že je model správně navržen).
- Jsou-li celkový F-test i všechny t-testy statisticky nevýznamné, model se považuje za nevhodný, protože nevystihuje variabilitu proměnné Y .
- Je-li celkový F-test statisticky významný, ale některé t-testy vychází nevýznamné, model se považuje za vhodný, ale provádí se zpravidla vypuštění nevýznamných parametrů.

Testy významnosti regresního modelu

Vyhodnocení významnosti regresního modelu:

- Jsou-li celkový F-test i všechny t-testy statisticky významné, model se považuje za vhodný k vystižení variability proměnné Y (to však ještě neznamená, že je model správně navržen).
- Jsou-li celkový F-test i všechny t-testy statisticky nevýznamné, model se považuje za nevhodný, protože nevystihuje variabilitu proměnné Y .
- Je-li celkový F-test statisticky významný, ale některé t-testy vychází nevýznamné, model se považuje za vhodný, ale provádí se zpravidla vypuštění nevýznamných parametrů.
- Je-li celkový F-test statisticky významný, ale všechny t-testy vychází nevýznamné – paradox: formálně model jako celek vyhovuje, ale žádný člen modelu sám o sobě významný není – jde o důsledek tzv. multikolinearity, tj. lineární závislosti mezi jednotlivými regresory.

Testy významnosti regresního modelu

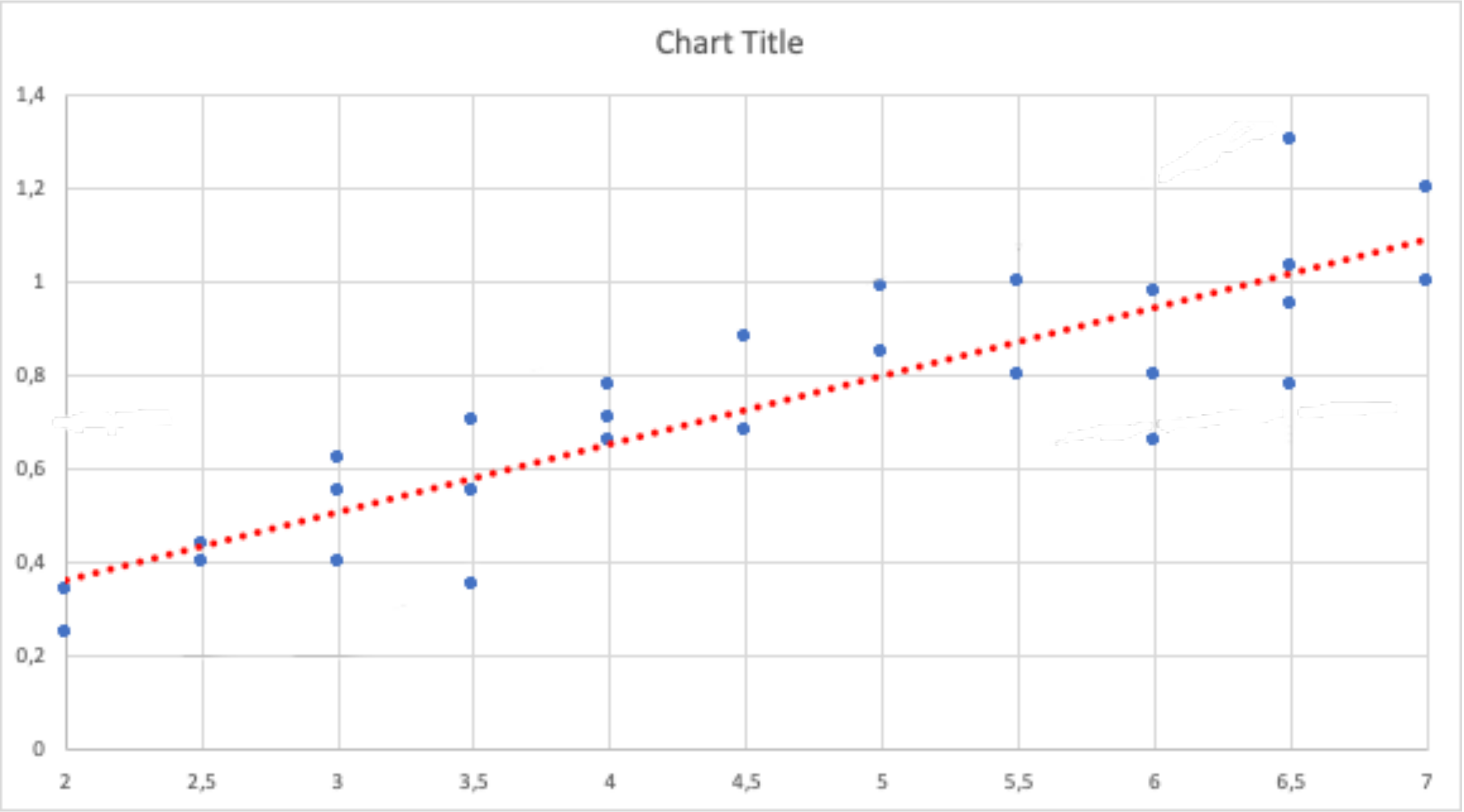
Vyhodnocení významnosti regresního modelu:

- Jsou-li celkový F-test i všechny t-testy statisticky významné, model se považuje za vhodný k vystižení variability proměnné Y (to však ještě neznamená, že je model správně navržen).
- Jsou-li celkový F-test i všechny t-testy statisticky nevýznamné, model se považuje za nevhodný, protože nevystihuje variabilitu proměnné Y .
- Je-li celkový F-test statisticky významný, ale některé t-testy vychází nevýznamné, model se považuje za vhodný, ale provádí se zpravidla vypuštění nevýznamných parametrů.
- Je-li celkový F-test statisticky významný, ale všechny t-testy vychází nevýznamné – paradox: formálně model jako celek vyhovuje, ale žádný člen modelu sám o sobě významný není – jde o důsledek tzv. multikolinearity, tj. lineární závislosti mezi jednotlivými regresory.

Lineární regresní model

Závislost mezi nezávisle proměnnou a závisle proměnnou: y=f(x)

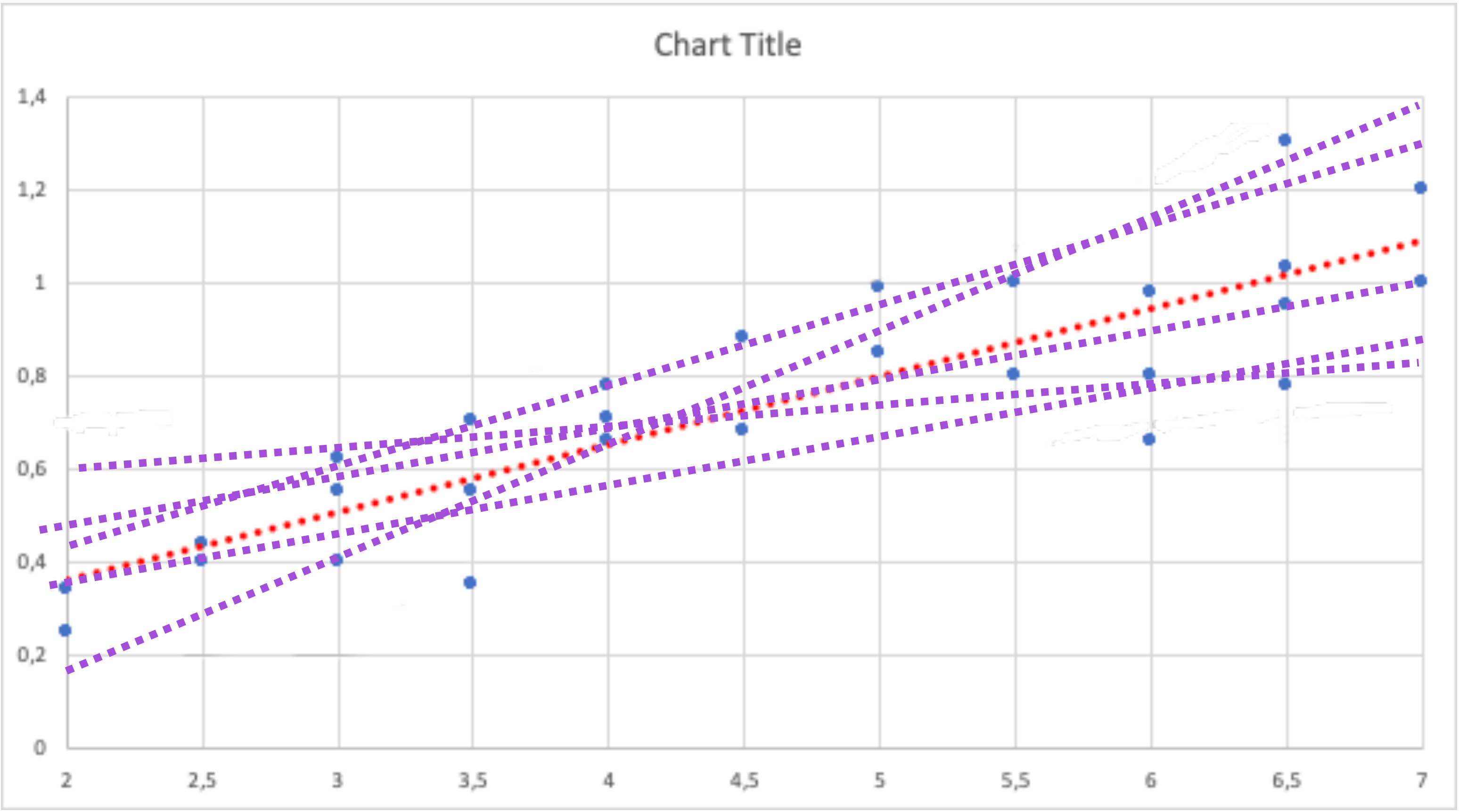
- funkční závislost: x a y jsou nenáhodné, pro jedno x je nejvýše jedna hodnota y
- regresní závislost: y je realizace náhodné veličiny Y při konkrétním x; pro jedno x můžeme pozorovat různé hodnoty Y



Lineární regresní model

Závislost mezi nezávisle proměnnou a závisle proměnnou: $y=f(x)$

- funkční závislost: x a y jsou nenáhodné, pro jedno x je nejvýše jedna hodnota y
- regresní závislost: y je realizace náhodné veličiny Y při konkrétním x; pro jedno x můžeme pozorovat různé hodnoty Y

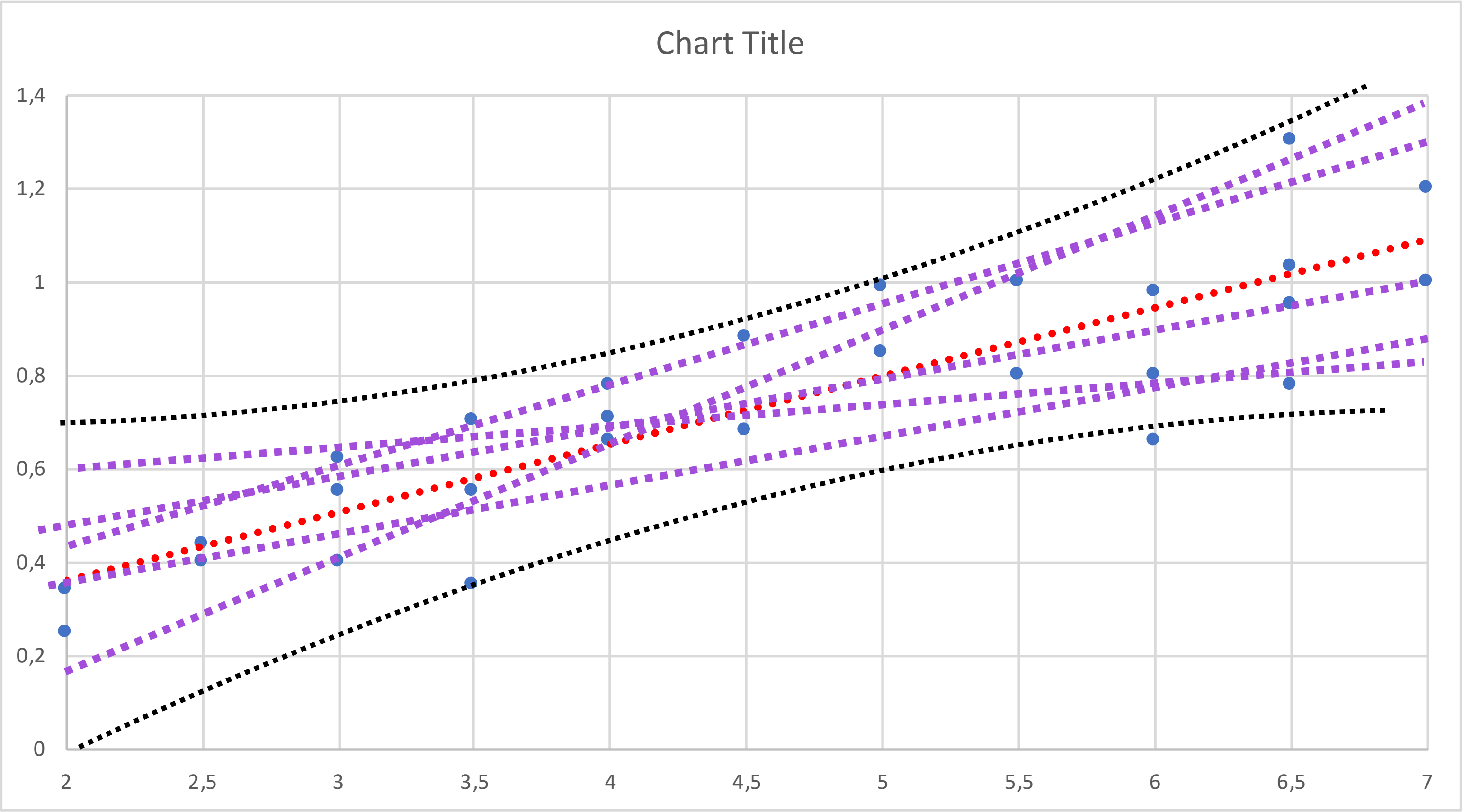


Lineární regresní model

Závislost mezi nezávisle proměnnou a závisle proměnnou: y=f(x)

- funkční závislost: x a y jsou nenáhodné, pro jedno x je nejvýše jedna hodnota y
- regresní závislost: y je realizace náhodné veličiny Y při konkrétním x; pro jedno x můžeme pozorovat různé hodnoty Y

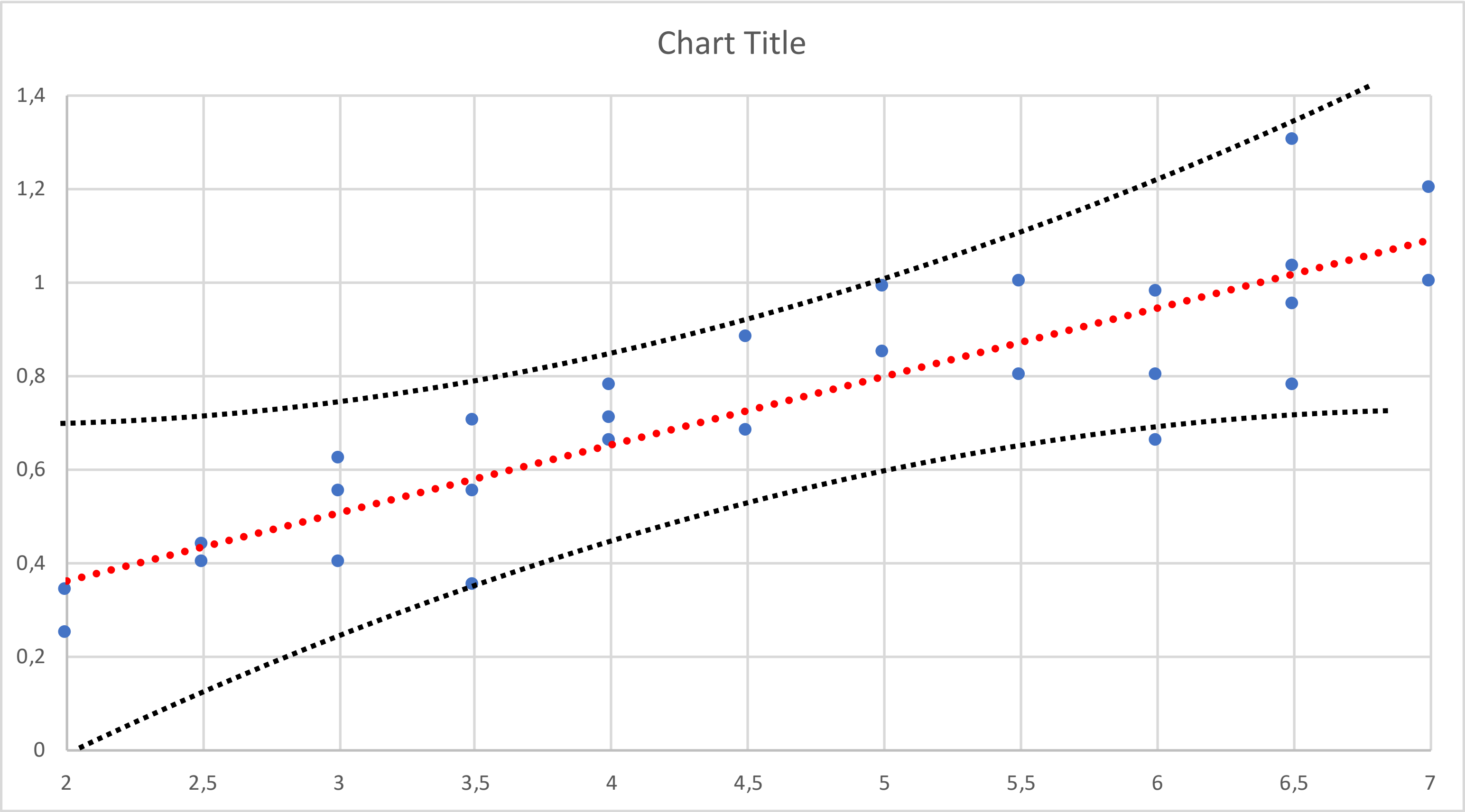
pás spolehlivosti pro regresní přímku



Lineární regresní model

Závislost mezi nezávisle proměnnou a závisle proměnnou: y=f(x)

- funkční závislost: x a y jsou nenáhodné, pro jedno x je nejvýše jedna hodnota y
- regresní závislost: y je realizace náhodné veličiny Y při konkrétním x; pro jedno x můžeme pozorovat různé hodnoty Y



pás spolehlivosti pro regresní přímku

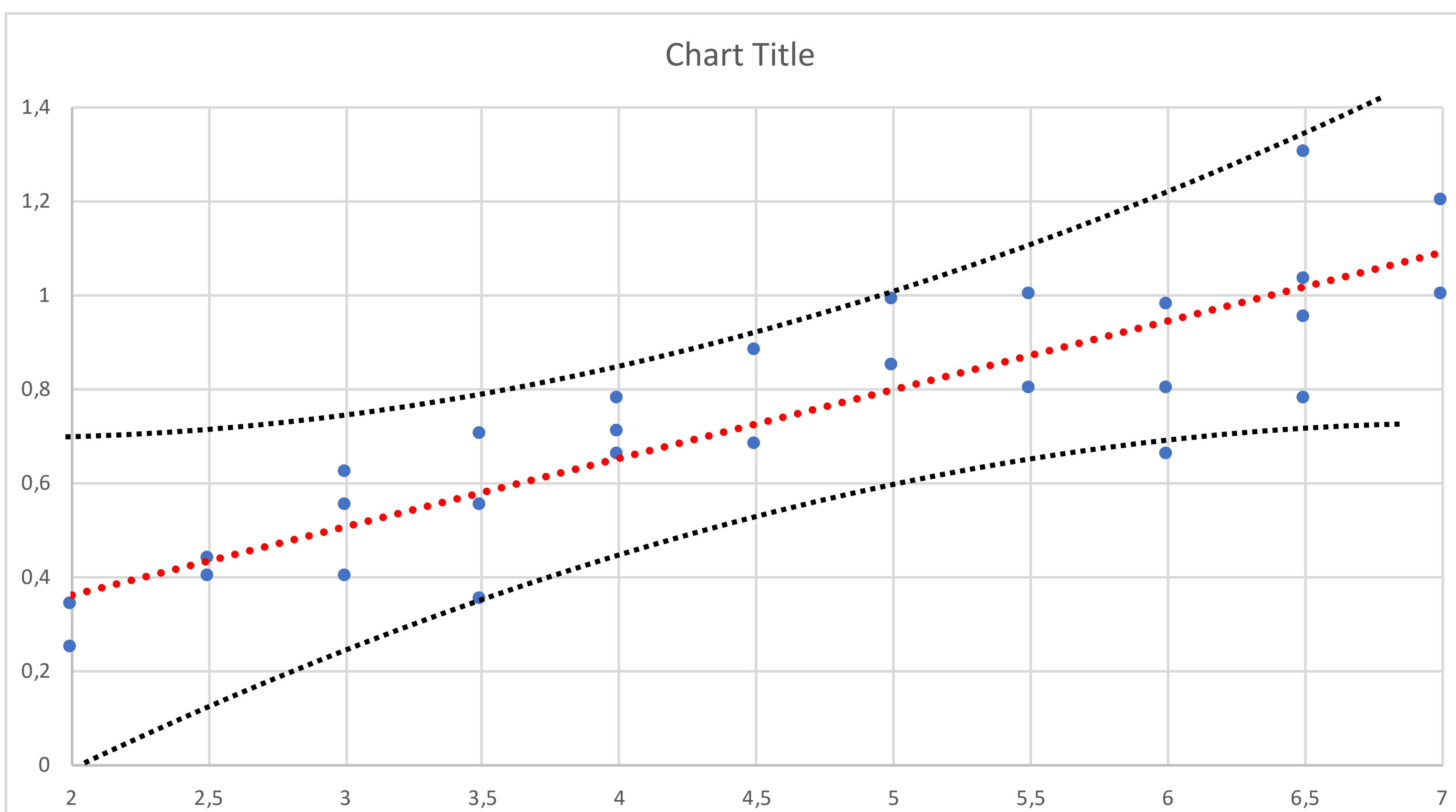
Lineární regresní model

Pás spolehlivosti pro regresní přímku $\hat{Y}(x) = a + bx$ **:**

$$\hat{Y}(x) \pm s_{\hat{y}}(x) t_{1-\gamma/2}(n-2)$$

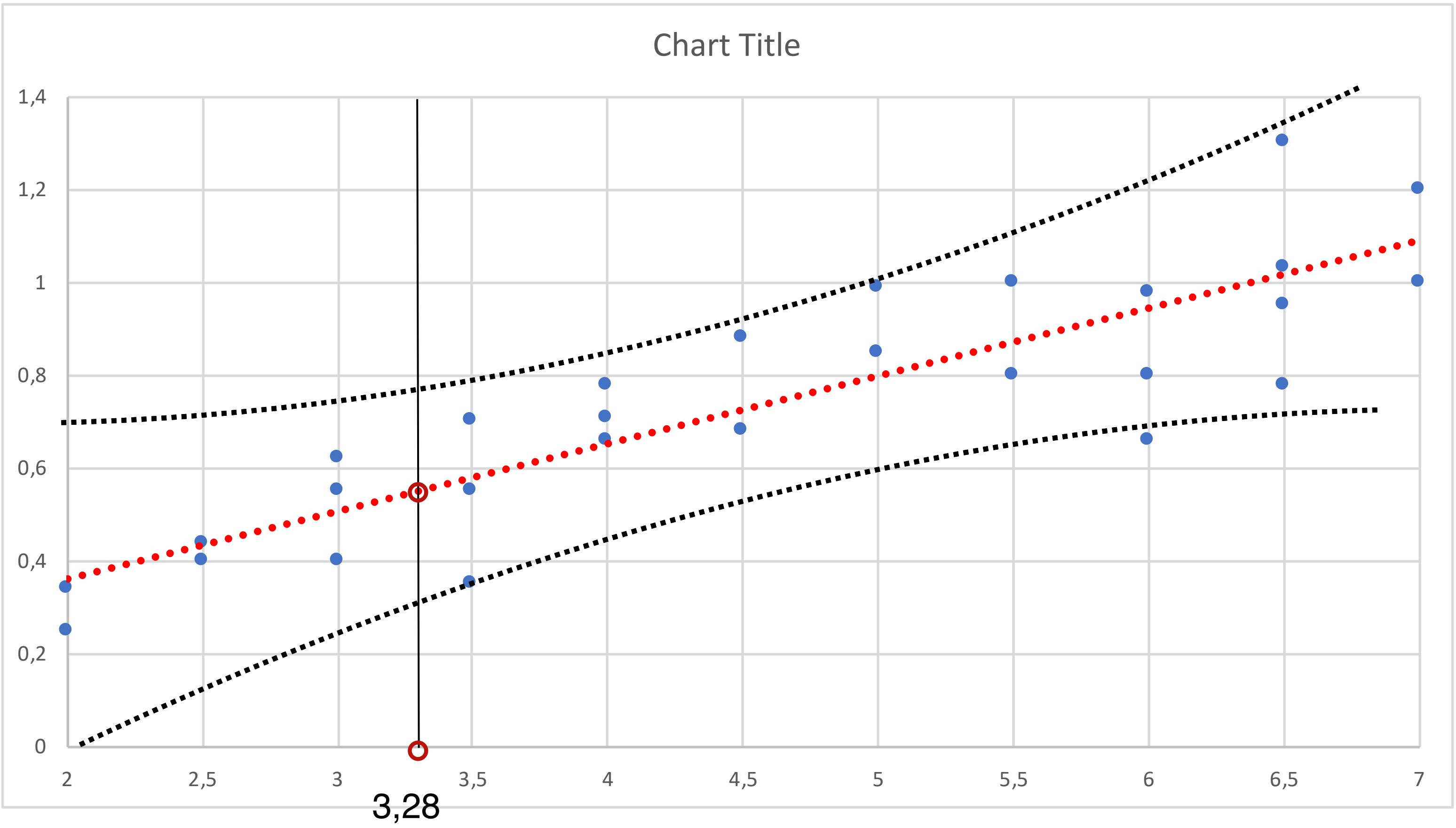
kde

$$s_{\hat{y}}(x) = s \cdot \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$



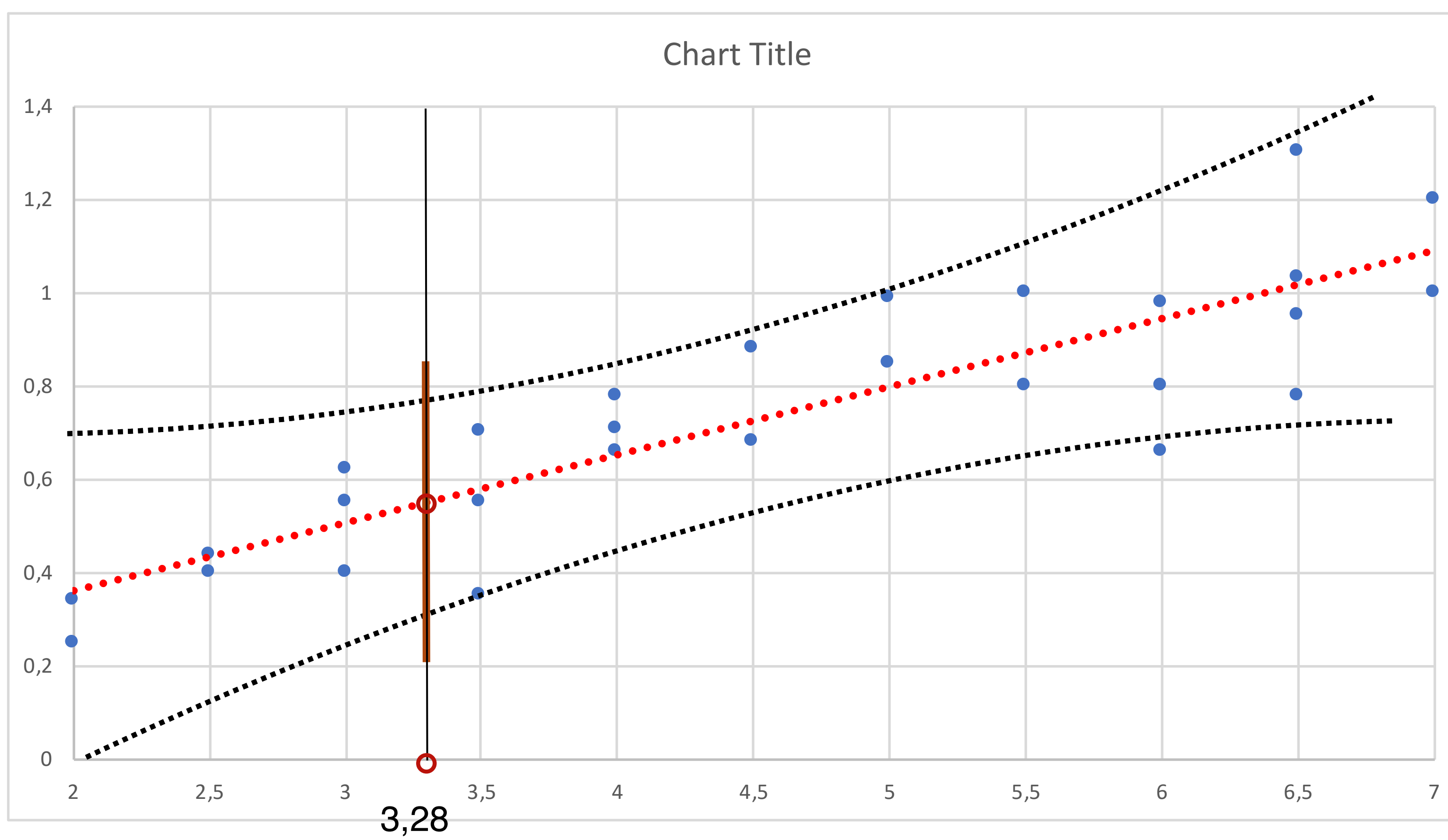
Lineární regresní model

Pás spolehlivosti pro predikci $\hat{Y}(x_0) = a + bx_0$ **:**



Lineární regresní model

Pás spolehlivosti pro predikci $\hat{Y}(x_0) = a + bx_0$ **:**



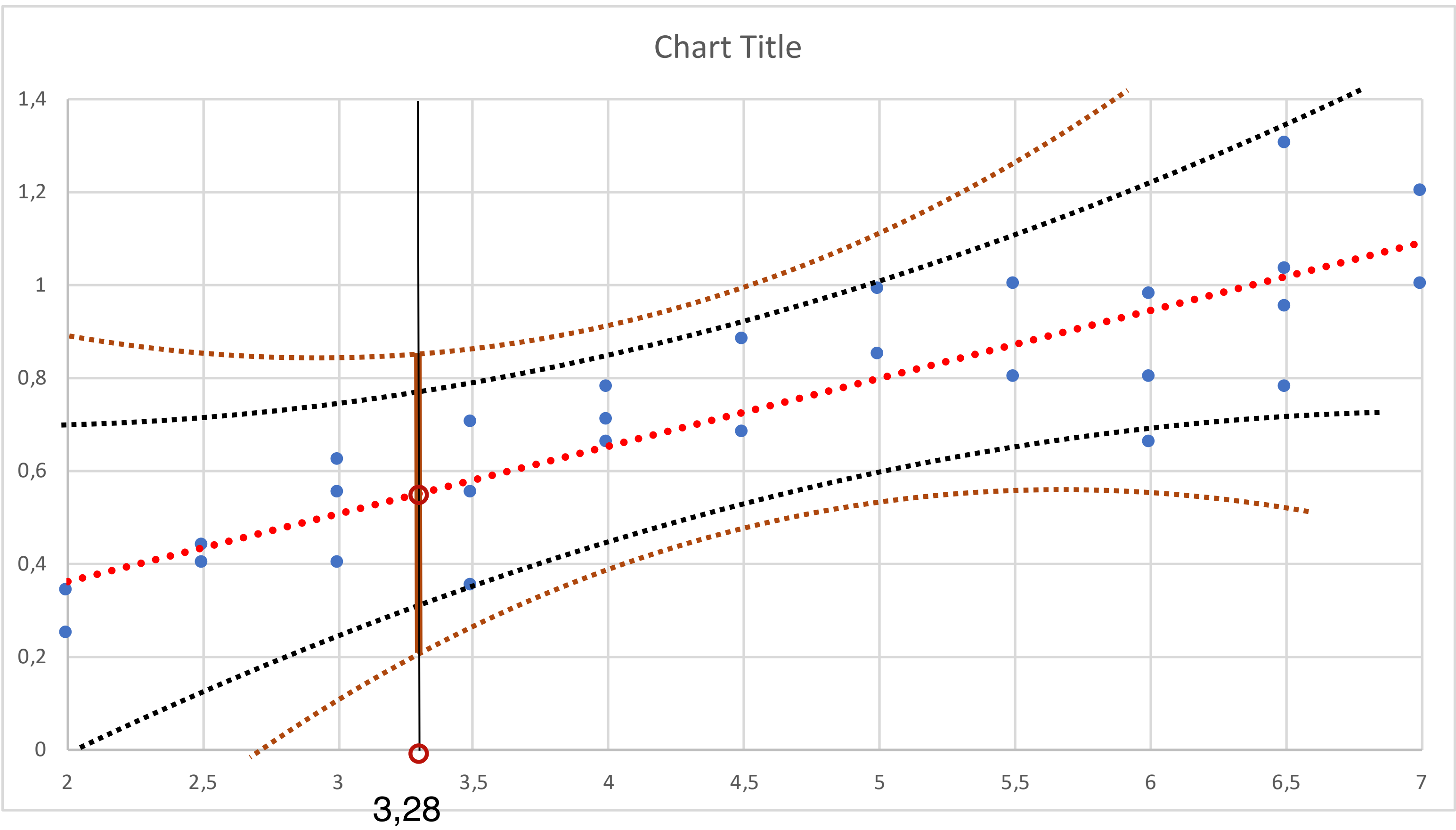
Lineární regresní model

Pás spolehlivosti pro predikci $\hat{Y}(x_0) = a + bx_0$ **:**

$$\hat{Y}(x_0) \pm s_{\hat{y}}(x_0) t_{1-\gamma/2}(n-2)$$

kde

$$s_{\hat{y}}(x_0) = s \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + 1}$$



Metoda nejmenších čtverců

Příklad 2.4.2 *Závislost mezi teplotou* θ *a rychlostí posuvu* v *v příkladu 2.4.1. lze považovat za regresní závislost ve tvaru* $\theta = \alpha.v^{\beta}.\epsilon$, *kde* α *a* β *jsou regresní koeficienty a* ϵ *je náhodná veličina se střední hodnotou 1. Provedeme-li transformaci* $Y = ln\theta, X = lnv, a = ln\alpha, e = ln\epsilon ab = \beta$, *dostaneme* $Y = a + bX + e$, *tedy lineární vztah.*

Podobně jako v předchozím příkladu lze linearizovat i jiné modely, např. logaritmický, tj. $Y = ln(\alpha + \beta.X)$, reciproký $Y = \frac{1}{\alpha+\beta.X}$ a další.

Metoda nejmenších čtverců

Příklad 2.4.2 *Závislost mezi teplotou* θ *a rychlostí posuvu* v *v příkladu 2.4.1. lze považovat za regresní závislost ve tvaru* $\theta = \alpha . v^{\beta} . \epsilon$, *kde* α *a* β *jsou regresní koeficienty a* ϵ *je náhodná veličina se střední hodnotou 1. Provedeme-li transformaci* $Y = ln\theta, X = lnv, a = ln\alpha, e = ln\epsilon ab = \beta$, *dostaneme* $Y = a + bX + e$, *tedy lineární vztah.*

Podobně jako v předchozím příkladu lze linearizovat i jiné modely, např. logaritmický, tj. $Y = ln(\alpha + \beta . X)$, reciproký $Y = \frac{1}{\alpha + \beta . X}$ a další.

Metoda nejmenších čtverců

Model lineární regrese lze použít i v některých případech, kdy závislost mezi veličinami X a Y není lineární. Jsou to případy, kdy lze provést takzvanou **linearizaci modelu**. Vhodnou transformací převedeme nelineární závislost na lineární a použijeme lineární regresní model. Přitom však musíme být velmi opatrní, neboť vše, co bylo odvozeno pro lineární regresní model za předpokladu normality chybového členu ϵ platí pouze pro "linearizovaný model", nikoli pro model původní, a to opět za předpokladu, že náhodná veličina, odpovídající transformovanému chybovému členu v linearizovaném modelu, má normální rozdělení.

Příklad 2.4.2 *Závislost mezi teplotou θ a rychlostí posuvu v v příkladu 2.4.1. lze považovat za regresní závislost ve tvaru $\theta = \alpha.v^{\beta}.\epsilon$, kde α a β jsou regresní koeficienty a ϵ je náhodná veličina se střední hodnotou 1. Provedeme-li transformaci $Y = ln\theta, X = lnv, a = ln\alpha, e = ln\epsilon ab = \beta$, dostaneme $Y = a + bX + e$, tedy lineární vztah.*

Podobně jako v předchozím příkladu lze linearizovat i jiné modely, např. logaritmický, tj. $Y = ln(\alpha + \beta.X)$, reciproký $Y = \frac{1}{\alpha+\beta.X}$ a další.

2.4. Regresní analýza

Regresní model experimentu

lineární model pro dva faktory: $Y = a + bX + cZ + dXZ + \varepsilon$

2.4. Regresní analýza

Regresní model experimentu

lineární model pro dva faktory: $Y = a + bX + cZ + dXZ + \varepsilon$

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \dots \\ Y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & X_1 & Z_1 & X_1 Z_1 \\ \dots & \dots & \dots & \dots \\ 1 & X_n & Z_n & X_n Z_n \end{pmatrix} \qquad \beta = \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \dots \\ \epsilon_n \end{pmatrix}$$

2.4. Regresní analýza

Regresní model experimentu

lineární model pro dva faktory: $Y = a + bX + cZ + dXZ + \varepsilon$

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \dots \\ Y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & X_1 & Z_1 & X_1 Z_1 \\ \dots & \dots & \dots & \dots \\ 1 & X_n & Z_n & X_n Z_n \end{pmatrix} \qquad \beta = \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \dots \\ \epsilon_n \end{pmatrix}$$

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum x_i & \sum z_i & \sum x_i z_i \\ \sum x_i & \sum x_i^2 & \sum x_i z_i & \sum x_i^2 z_i \\ \sum z_i & \sum x_i z_i & \sum z_i^2 & \sum x_i z_i^2 \\ \sum x_i z_i & \sum x_i^2 z_i & \sum x_i z_i^2 & \sum x_i^2 z_i^2 \end{pmatrix} \qquad \mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \\ \sum z_i Y_i \\ \sum x_i z_i Y_i \end{pmatrix}$$

2.4. Regresní analýza

Regresní model experimentu

lineární model pro dva faktory: $Y = a + bX + cZ + dXZ + \varepsilon$

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \dots \\ Y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & X_1 & Z_1 & X_1 Z_1 \\ \dots & \dots & \dots & \dots \\ 1 & X_n & Z_n & X_n Z_n \end{pmatrix} \quad \beta = \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \dots \\ \epsilon_n \end{pmatrix}$$

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum x_i & \sum z_i & \sum x_i z_i \\ \sum x_i & \sum x_i^2 & \sum x_i z_i & \sum x_i^2 z_i \\ \sum z_i & \sum x_i z_i & \sum z_i^2 & \sum x_i z_i^2 \\ \sum x_i z_i & \sum x_i^2 z_i & \sum x_i z_i^2 & \sum x_i^2 z_i^2 \end{pmatrix} \quad \mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \\ \sum z_i Y_i \\ \sum x_i z_i Y_i \end{pmatrix}$$