Základy pravděpodobnosti a matematické statistiky

13. Lineární regrese

FAKULTA STROJNÍ ČVUT V PRAZE

prof. RNDr. Gejza Dohnal, CSc.

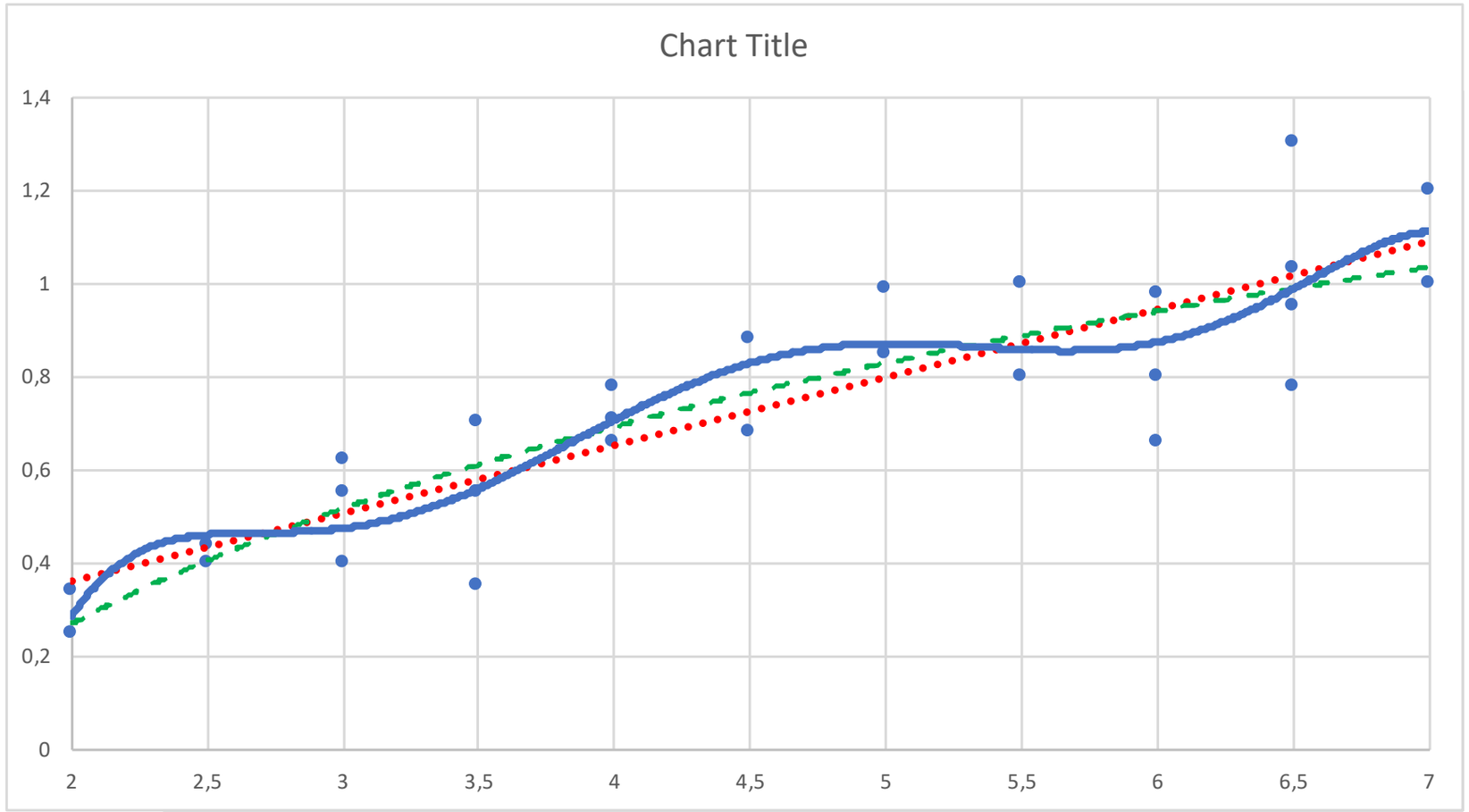
ak. rok 2023/2024

13. Lineární regrese

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

$$\vec{\beta} = \begin{pmatrix} \beta_0 \\ \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix} \quad \vec{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} \quad \Rightarrow \boxed{\mathbf{Y} = \mathbf{X}\vec{\beta} + \vec{e}}$$

Odhad parametrů: (metodou nejmenších čtverců)

Zvolíme ztrátovou funkci: $S(\vec{\beta}) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2, \quad S = (\mathbf{Y} - \mathbf{X}\vec{\beta})'(\mathbf{Y} - \mathbf{X}\vec{\beta})$

kterou minimalizujeme. To vede obecně k řešení $\vec{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{Y}$ $\quad E\vec{b} = \vec{\beta}$

reziduální součet čtverců

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

$$S_R = (\mathbf{Y} - \mathbf{X}\vec{b})'(\mathbf{Y} - \mathbf{X}\vec{b}) = \mathbf{Y}'\mathbf{Y} - \vec{b}'\mathbf{X}'\mathbf{X}$$

$$s^2 = \frac{1}{n-k} S_R$$ reziduální rozptyl

Lineární regresní model

Příklad: Přímková regrese: $Y = \alpha + \beta X + \epsilon$

$$\vec{\beta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \vec{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Odhad parametrů: (metodou nejmenších čtverců) $S(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$

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

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

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

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

normální soustava rovnic

a její řešení:

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

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

rovnice regresní přímky: $y = a + bx$

Lineární regresní model

Příklad: Kvadratická regrese: $Y = \alpha + \beta X + \gamma X^2 + \epsilon$

$$\vec{\beta} = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_2 & x_n^2 \end{pmatrix}, \quad \vec{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Odhad parametrů: (metodou nejmenších čtverců) $S(\alpha, \beta, \gamma) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i - \gamma x_i^2)^2$

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

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

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

$$\begin{pmatrix} n & \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 \\ \sum_{i=1}^n x_i 2 & \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^4 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n y_i x_i \\ \sum_{i=1}^n y_i x_i^2 \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

$$\Rightarrow \quad y = a + bx + cx^2$$

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

Příklad: Závislost délky brzdné dráhy na rychlosti.

Jak závisí délka brzdné dráhy na rychlosti vozidla?

Provedeme experiment s 50 vozidly: měříme délku brzdné dráhy (dist) a rychlost (speed)

```
> library(datasets)
> cars
```

	speed	dist
1	4	2
2	4	10
3	7	4
4	7	22
5	8	16
6	9	10
7	10	18
8	10	26
9	10	34
10	11	17
11	11	28
12	12	14
13	12	20
14	12	24
15	12	28
16	13	26
17	13	34
18	13	34
19	13	46
20	14	26
21	14	36
22	14	60
23	14	80
24	15	20
25	15	26
26	15	54
27	16	32
28	16	40
29	17	32
30	17	40
31	17	50
32	18	42
33	18	56
34	18	76
35	18	84
36	19	36
37	19	46
38	19	68
39	20	32
40	20	48
41	20	52
42	20	56
43	20	64
44	22	66
45	23	54
46	24	70
47	24	92
48	24	93
49	24	120
50	25	85

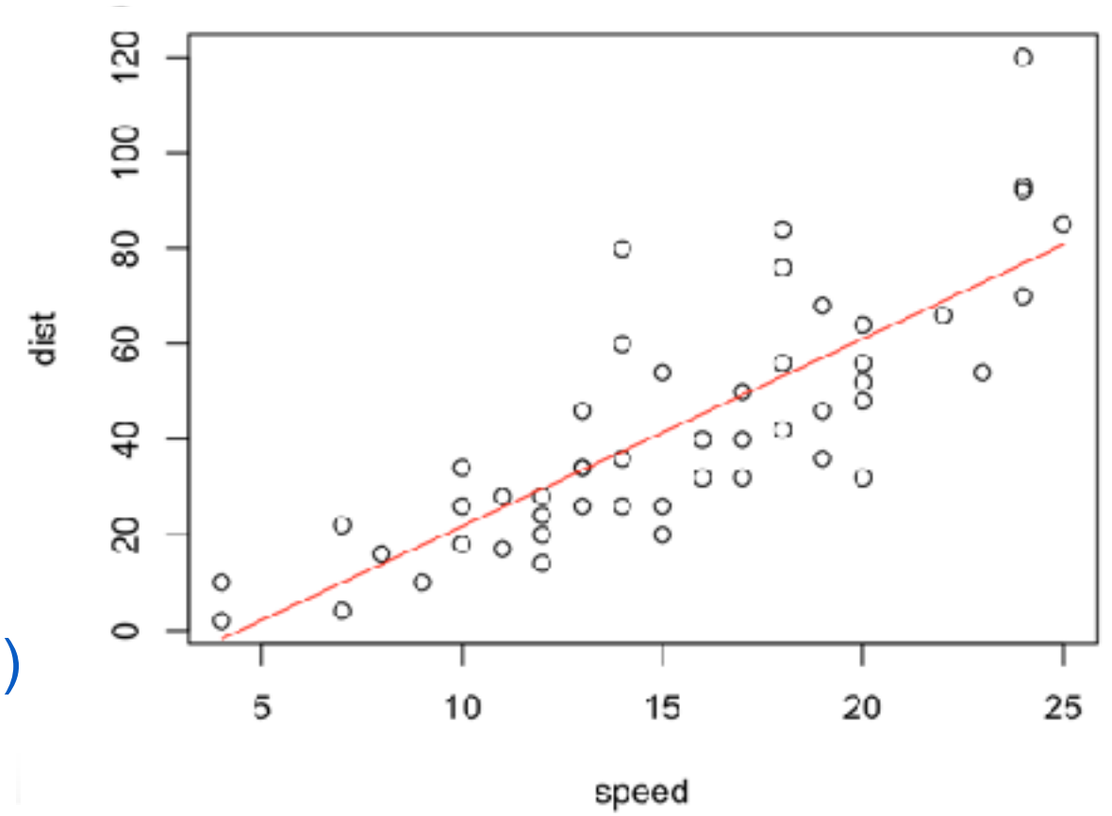
Lineární regresní model

Příklad: Závislost délky brzdné dráhy na rychlosti.

Jak závisí délka brzdné dráhy na rychlosti vozidla?

Provedeme experiment s 50 vozidly: měříme délku brzdné dráhy (dist) a rychlost (speed)

```
> plot(cars)
>  lrc <- lm(dist~speed, data=cars)
>  lrc$coefficients
     (Intercept)        speed
     -17.579095     3.932409
>  a <- lrc$coefficients[1]
>  b <- lrc$coefficients[2]
> lines(cars$speed, a+b*cars$speed, col = "red")
```



Výsledkem je přímka: y=-17,58 + 3,93*x

To je evidentně nesmysl, neboť při nulové rychlosti by byla brzdná dráha -17,58 m!

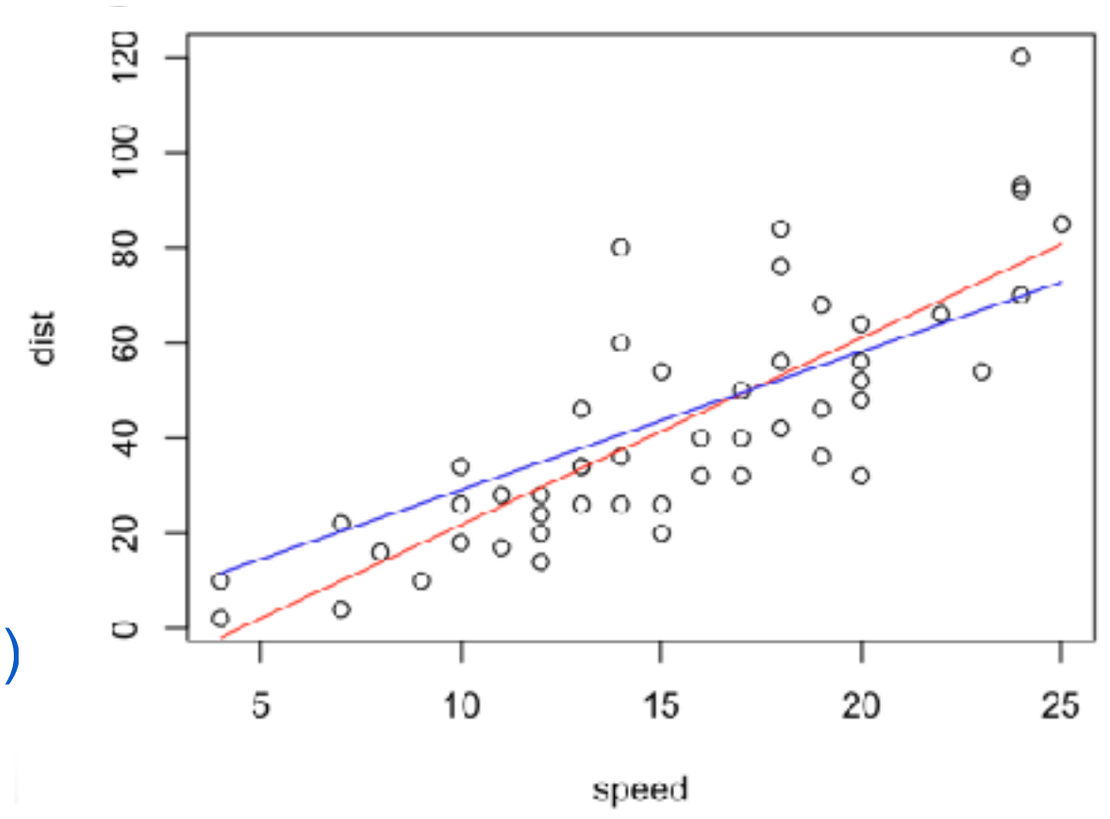
Lineární regresní model

Příklad: Závislost délky brzdné dráhy na rychlosti.

Jak závisí délka brzdné dráhy na rychlosti vozidla?

Provedeme experiment s 50 vozidly: měříme délku brzdné dráhy (dist) a rychlost (speed)

```
> plot(cars)
>  lrc <- lm(dist~speed, data=cars)
>  lrc$coefficients
     (Intercept)        speed
     -17.579095     3.932409
>  a <- lrc$coefficients[1]
>  b <- lrc$coefficients[2]
> lines(cars$speed, a+b*cars$speed, col = "red")
```



regrese procházející počátkem: $Y=\beta^*X + \epsilon$

```
> lrc0  <-  lm(dist~speed - 1, data=cars)
> lrc0$coefficients
     speed
  2.909132
> lines(cars$speed, lrc0$coefficients*cars$speed, col = "blue")
```

Lineární regresní model

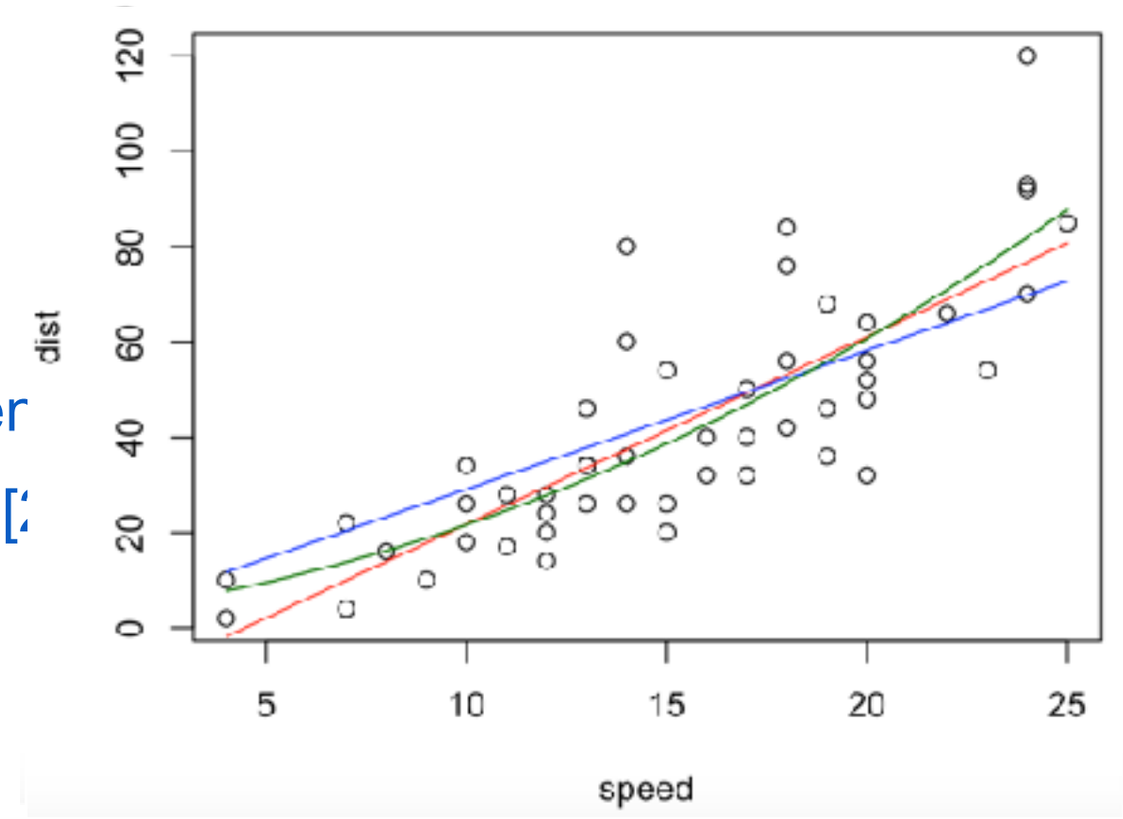
Příklad: Závislost délky brzdné dráhy na rychlosti.

Jak závisí délka brzdné dráhy na rychlosti vozidla?

Provedeme experiment s 50 vozidly: měříme délku brzdné dráhy (dist) a rychlost (speed)

```
> qspeed  <-  cars$speed^2
> qrc  <-  lm(dist~speed + qspeed, data=cars)
> qrc$coefficients
     (Intercept)        speed       qspeed
       2.4701378    0.9132876    0.0999593
```

```
> xfit<-seq(min(cars$speed),max(cars$speed),ler
> yfit = qrc$coefficients[1]*one + qrc$coefficients[
> lines(xfit, yfit, col = "darkgreen")
```



Lineární regresní model

Testování významnosti koeficientů (přímkové) lineární regrese

- reziduální součet čtverců: $S_R = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$ $\qquad s^2 = \frac{1}{n-2} S_R$

- intervalové odhady koeficientů:

$$a - s_a t_{1-\gamma/2}(n-2) \leq \alpha \leq a + s_a t_{1-\gamma/2}(n-2) \qquad s_a^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} s_b^2$$

$$b - s_b . t_{1-\gamma/2}(n-2) \leq \beta \leq b + s_b . t_{1-\gamma/2}(n-2) \qquad s_b^2 = \frac{s^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

- 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:

$$R^2 = \frac{\sum_{i=1}^{n} (a + bx_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{S_R}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\text{část variability vysvětlená modelem}}{\text{celková variabilita dat}}$$

Lineární regresní model

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

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

kde

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

Lineární regresní model

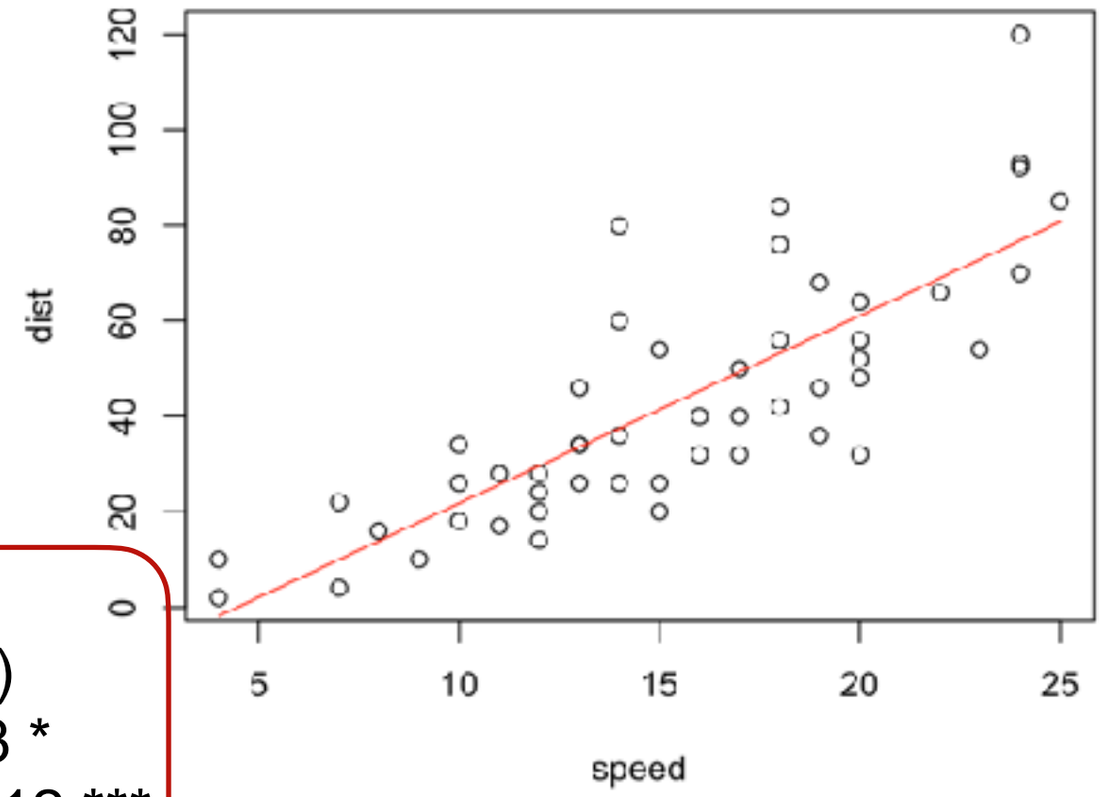
Příklad: Závislost délky brzdné dráhy na rychlosti.

Jak závisí délka brzdné dráhy na rychlosti vozidla?

Provedli jsme experiment s 50 vozidly: máme délku brzdné dráhy (dist) a rychlost (speed)

```
> lrc <- lm(dist~speed, data=cars)
> summary(lrc)
```

Call: lm(formula = dist ~ speed, data = cars)



Residuals:

Min	1Q	Median	3Q	Max
-29.069	-9.525	-2.272	9.215	43.201

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-17.5791	6.7584	-2.601	0.0123	*
speed	3.9324	0.4155	9.464	1.49e-12	***

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438
F-statistic: 89.57 on 1 and 48 DF, p-value: 1.49e-12

Lineární regresní model

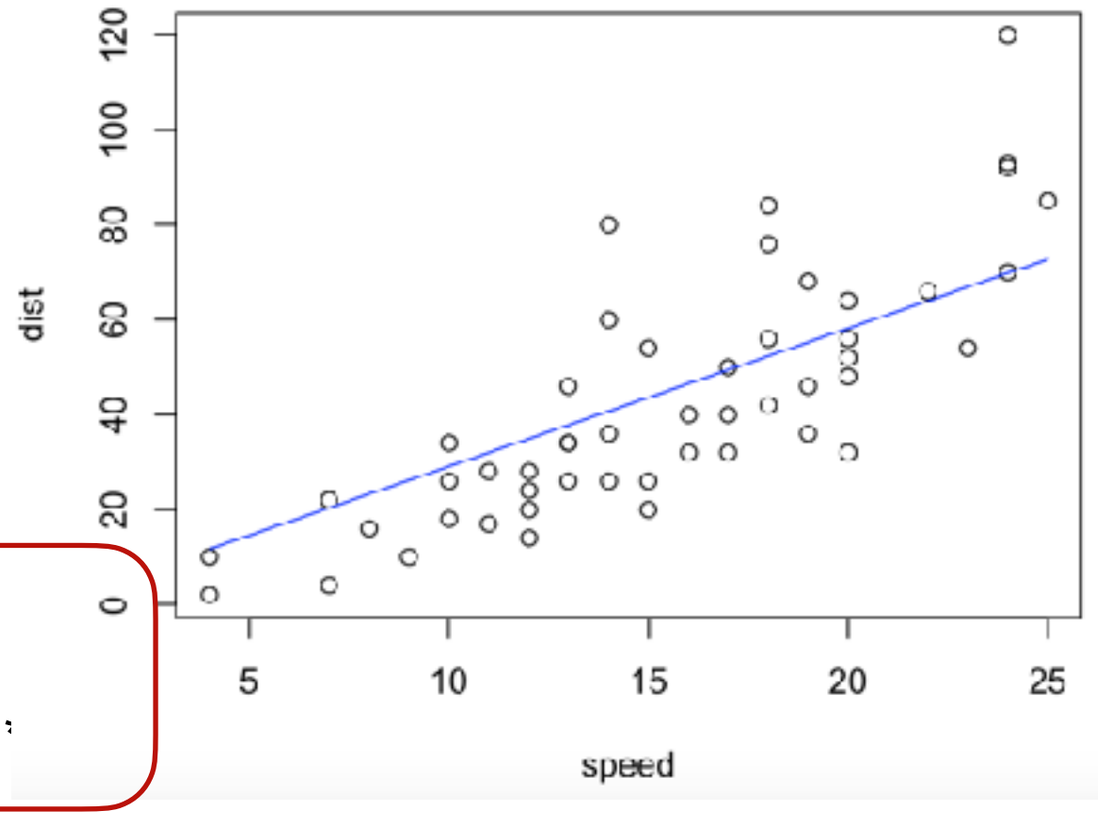
Příklad: Závislost délky brzdné dráhy na rychlosti.

Jak závisí délka brzdné dráhy na rychlosti vozidla?

Provedeme experiment s 50 vozidly: měříme délku brzdné dráhy (dist) a rychlost (speed)

```
> lrc0 <- lm(dist~speed-1, data=cars)
> summary(lrc0)
```

Call: lm(formula = dist ~ speed - 1, data = cars)



Residuals:

Min	1Q	Median	3Q	Max
-26.183	-12.637	-5.455	4.590	50.181

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
speed	2.9091	0.1414	20.58	<2e-16

Residual standard error: 16.26 on 49 degrees of freedom

Multiple R-squared: 0.8963, Adjusted R-squared: 0.8942

F-statistic: 423.5 on 1 and 49 DF, p-value: < 2.2e-16

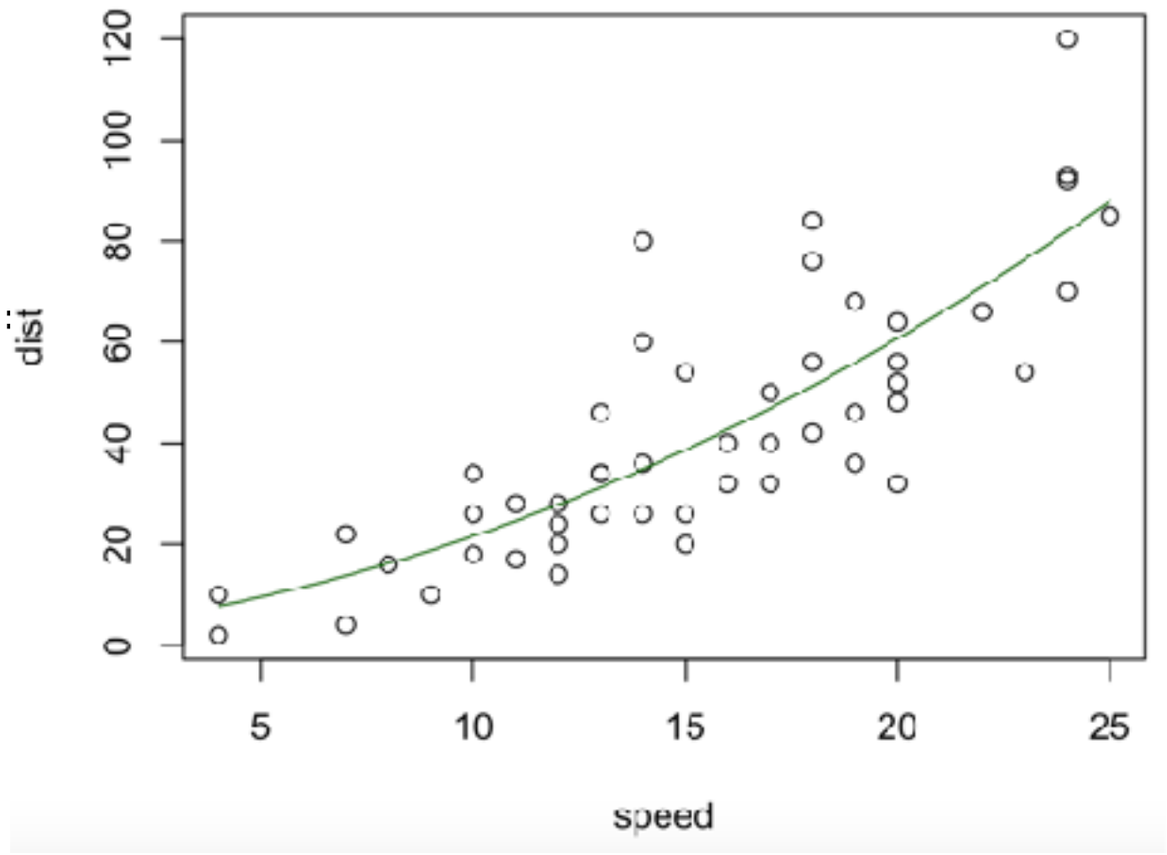
Lineární regresní model

Příklad: Závislost délky brzdné dráhy na rychlosti.

Jak závisí délka brzdné dráhy na rychlosti vozidla?

Provedeme experiment s 50 vozidly: měříme délku brzdné dráhy (dist) a rychlost (speed)

```
> qspeed  <-  cars$speed^2
> qrc  <-  lm(dist~speed + qspeed, data=cars)
> qrc$coefficients
> summary(qrc)
Call: lm(formula = dist ~ speed + qspeed, data
Residuals:
    Min      1Q  Median      3Q     Max
-28.720  -9.184  -3.188   4.628  45.152
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.47014   14.81716   0.167    0.868
speed        0.91329    2.03422   0.449    0.656
qspeed       0.09996    0.06597   1.515    0.136

Residual standard error: 15.18 on 47 degrees of freedom
Multiple R-squared:  0.6673,   Adjusted R-squared:  0.6532
F-statistic: 47.14 on 2 and 47 DF,  p-value: 5.852e-12
```



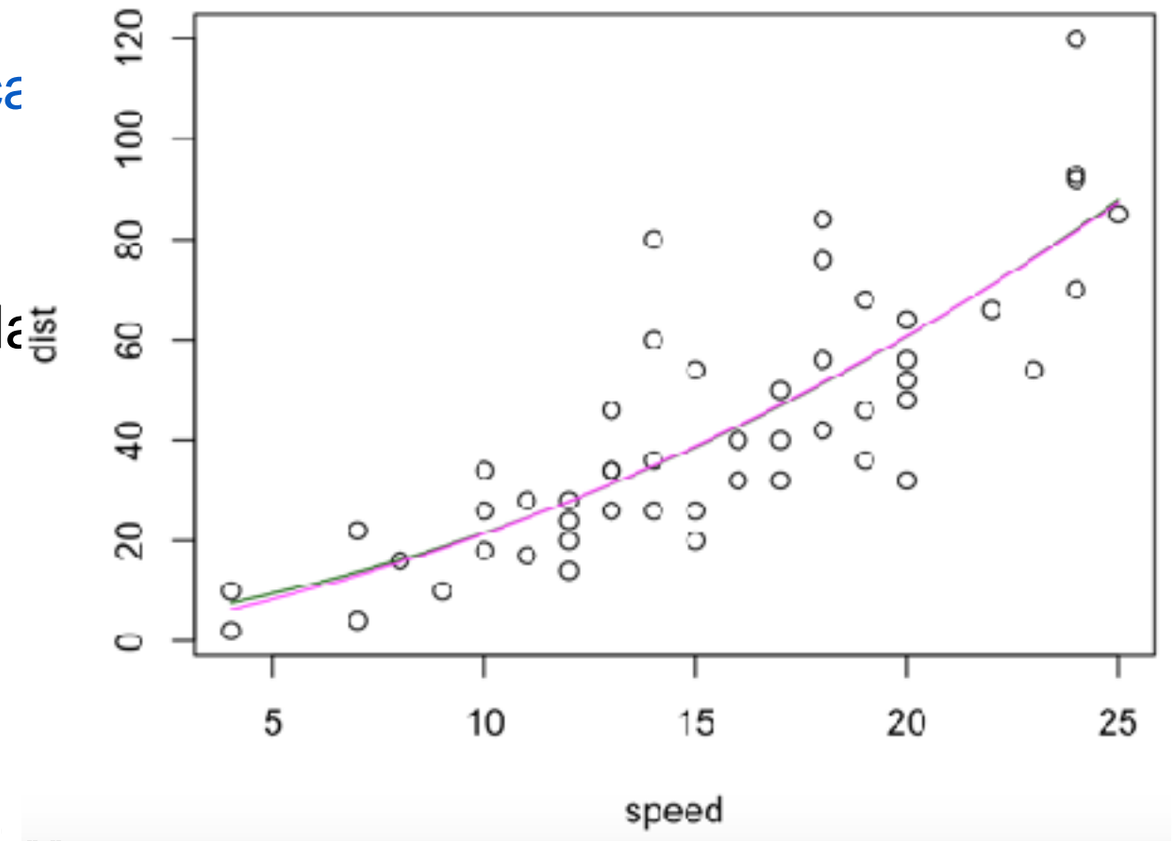
Lineární regresní model

Příklad: Závislost délky brzdné dráhy na rychlosti.

Jak závisí délka brzdné dráhy na rychlosti vozidla?

Provedeme experiment s 50 vozidly: měříme délku brzdné dráhy (dist) a rychlost (speed)

```
> qspeed  <-  cars$speed^2
> qrc0  <-  lm(dist~speed + qspeed - 1, data=ca

> summary(qrc0)
Call: lm(formula = dist ~ speed + qspeed - 1, da
Residuals:
    Min      1Q  Median      3Q     Max
-28.836  -9.071  -3.152   4.570  44.986
Coefficients:
            Estimate  Std. Error  t value    Pr(>|t|)
speed    1.23903    0.55997    2.213    0.03171
qspeed  0.09014    0.02939    3.067    0.00355
---
Residual standard error: 15.02 on 48 degrees of freedom
Multiple R-squared:  0.9133,   Adjusted R-squared:  0.9097
F-statistic: 252.8 on 2 and 48 DF,  p-value: < 2.2e-16
```



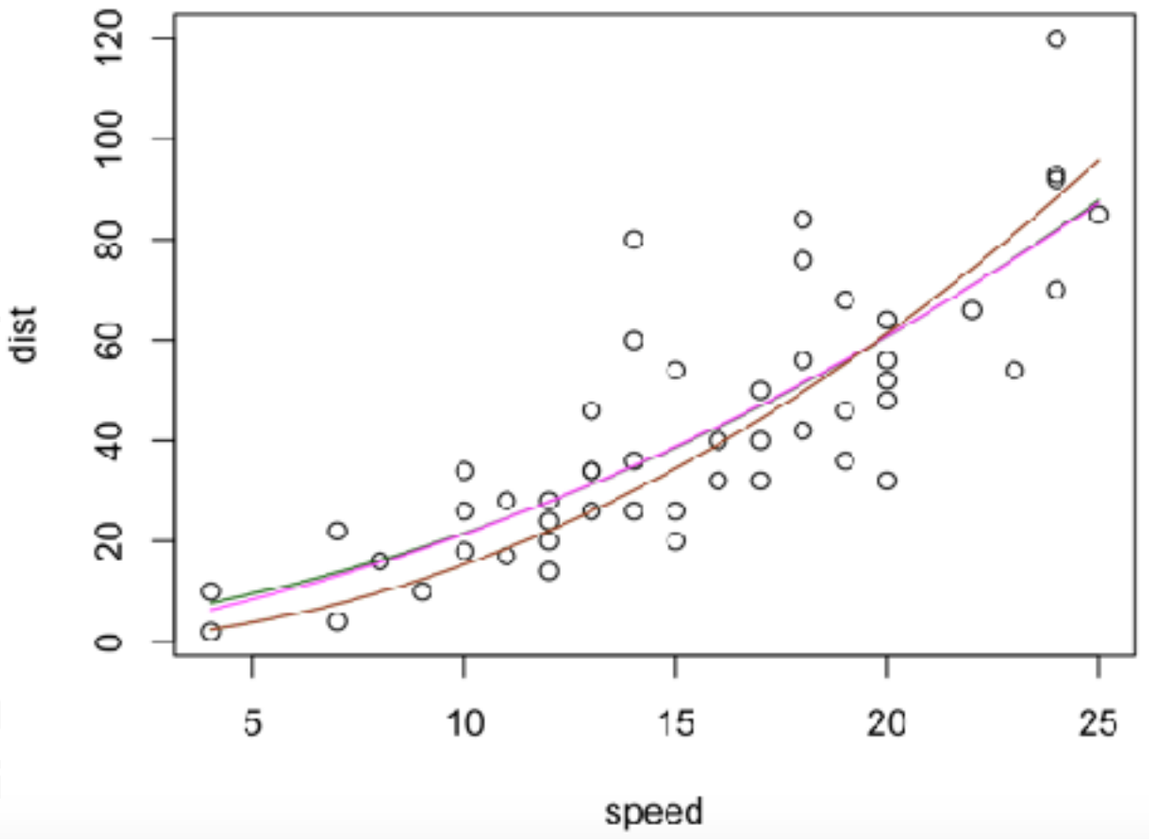
Lineární regresní model

Příklad: Závislost délky brzdné dráhy na rychlosti.

Jak závisí délka brzdné dráhy na rychlosti vozidla?

Provedeme experiment s 50 vozidly: měříme délku brzdné dráhy (dist) a rychlost (speed)

```
> qspeed  <-  cars$speed^2
> qrc2  <-  lm(dist~qspeed - 1, data=cars)

> summary(qrc2)

Call:  lm(formula = dist ~ qspeed - 1, data = car

Residuals:
    Min      1Q  Median      3Q     Max
-29.350  -7.988   1.325   8.080  49.939

Coefficients:
        Estimate  Std. Error  t value  Pr(>|t|)
qspeed  0.153374  0.007122    21.54    <2e-16

Residual standard error: 15.61 on 49 degrees of freedom
Multiple R-squared:  0.9044,  Adjusted R-squared:  0.9025
F-statistic: 463.8 on 1 and 49 DF,  p-value: < 2.2e-16
```



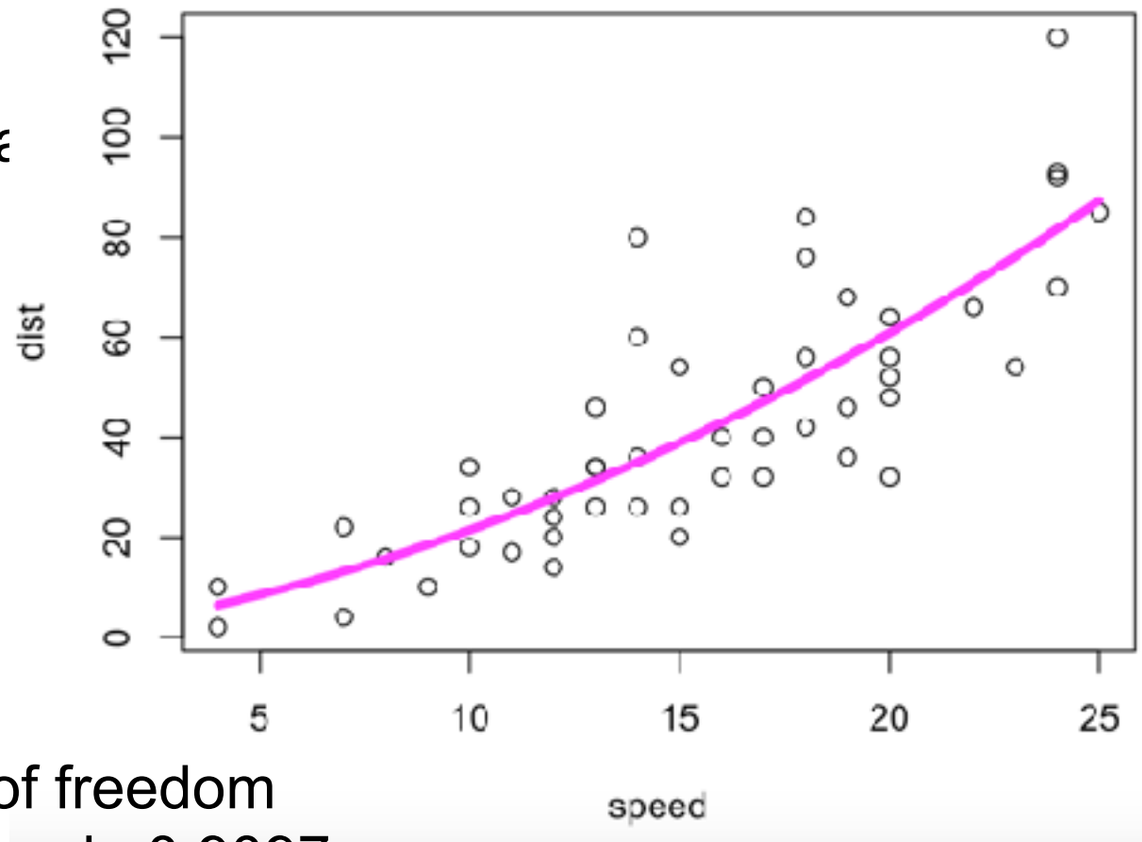
Lineární regresní model

Příklad: Závislost délky brzdné dráhy na rychlosti.

Jak závisí délka brzdné dráhy na rychlosti vozidla?

Provedeme experiment s 50 vozidly: měříme délku brzdné dráhy (dist) a rychlost (speed)

```
> summary(qrc0)
Call: lm(formula = dist ~ speed + qspeed - 1, da
Residuals:
    Min      1Q  Median      3Q     Max
-28.836  -9.071  -3.152   4.570  44.986
Coefficients:
        Estimate  Std. Error  t value   Pr(>|t|)
speed   1.23903    0.55997    2.213    0.03171
qspeed  0.09014    0.02939    3.067    0.00355
```




```
Residual standard error: 15.02 on 48 degrees of freedom
Multiple R-squared:  0.9133,   Adjusted R-squared:  0.9097
F-statistic: 252.8 on 2 and 48 DF,  p-value: < 2.2e-16
```

Závěr: **Závislost délky brzdné dráhy (y) na rychlosti (x) je kvadratická.**
Na základě měření jsme odhadli tuto závislost rovnicí: **$y = 1{,}24x + 0{,}09x^2$**

Lineární regresní model

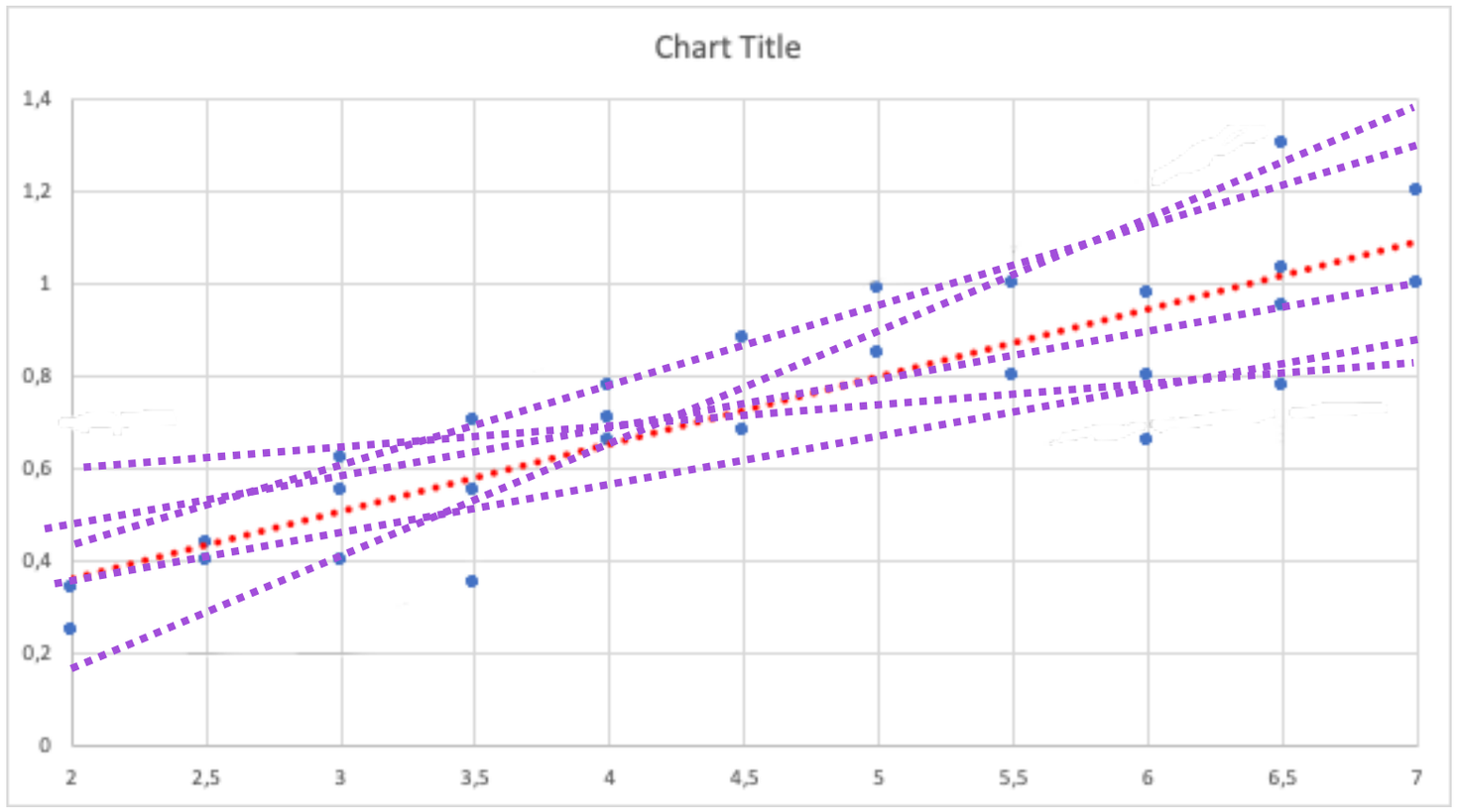
Vyhodnocení významnosti regresního modelu

- **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,** je to 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

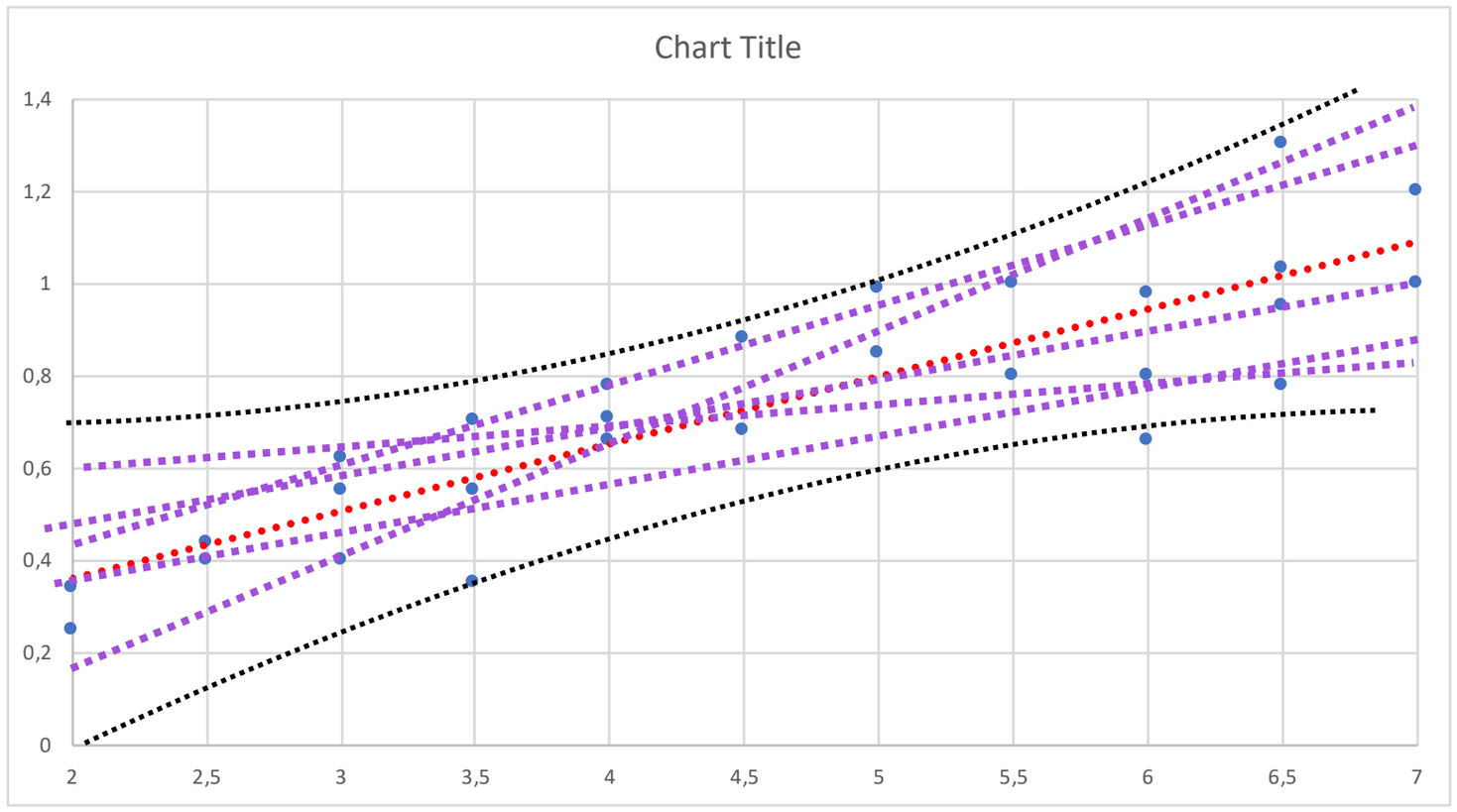


Lineární regresní model

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

pás spolehlivosti pro regresní přímku



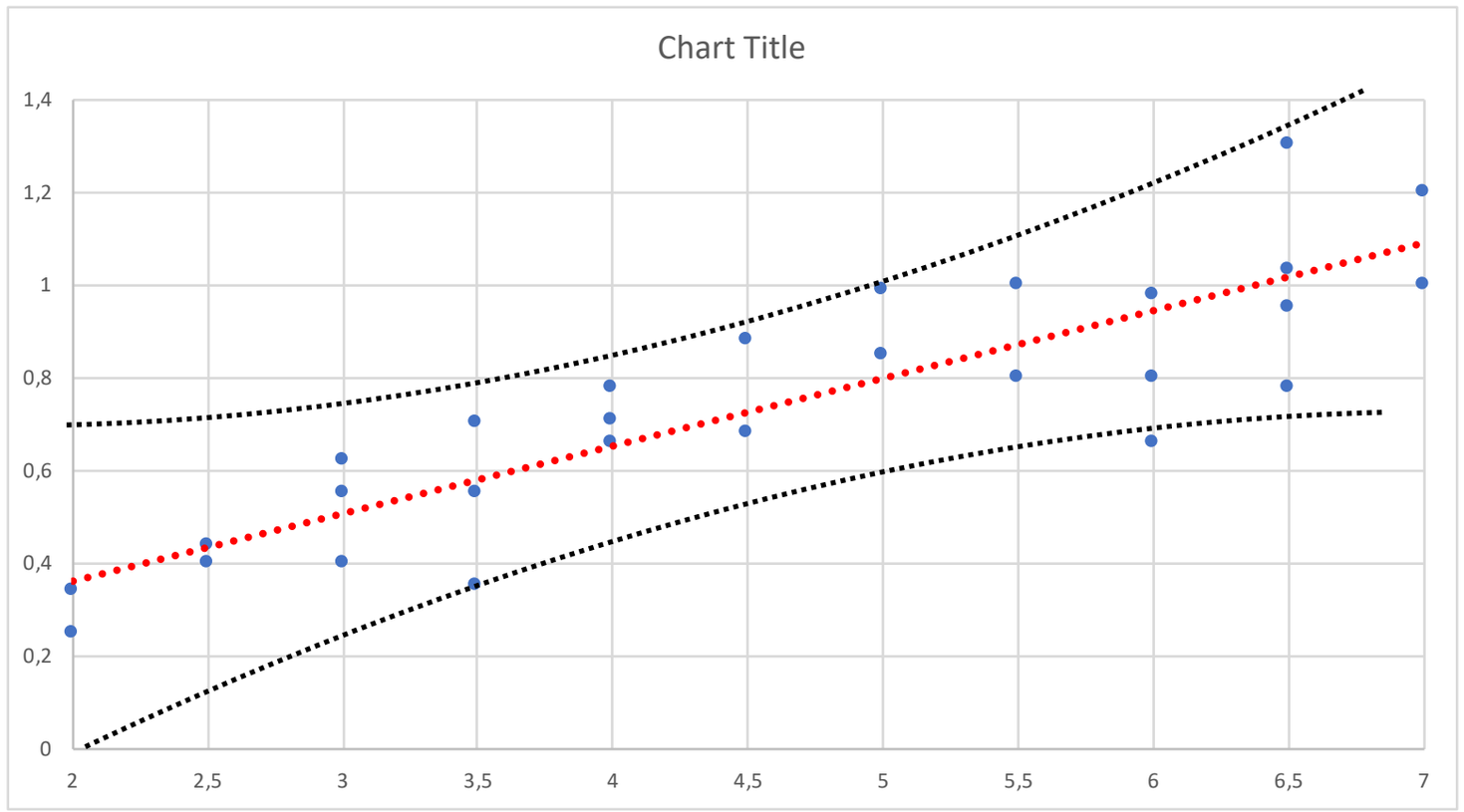
Lineární regresní model

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

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

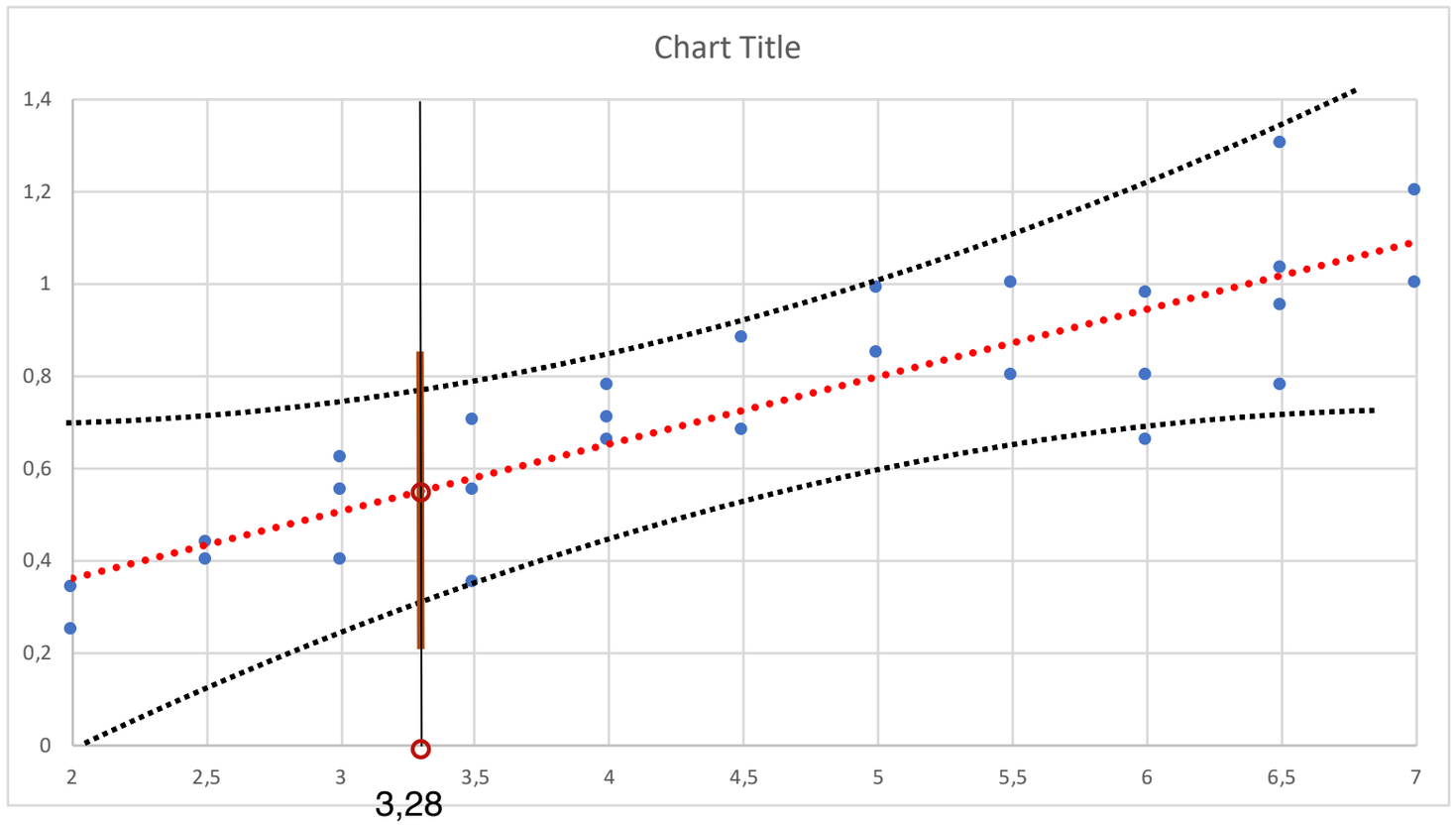
kde

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



Lineární regresní model

Pás spolehlivosti pro predikci $Y(x_0) = a + bx_0$ **:**



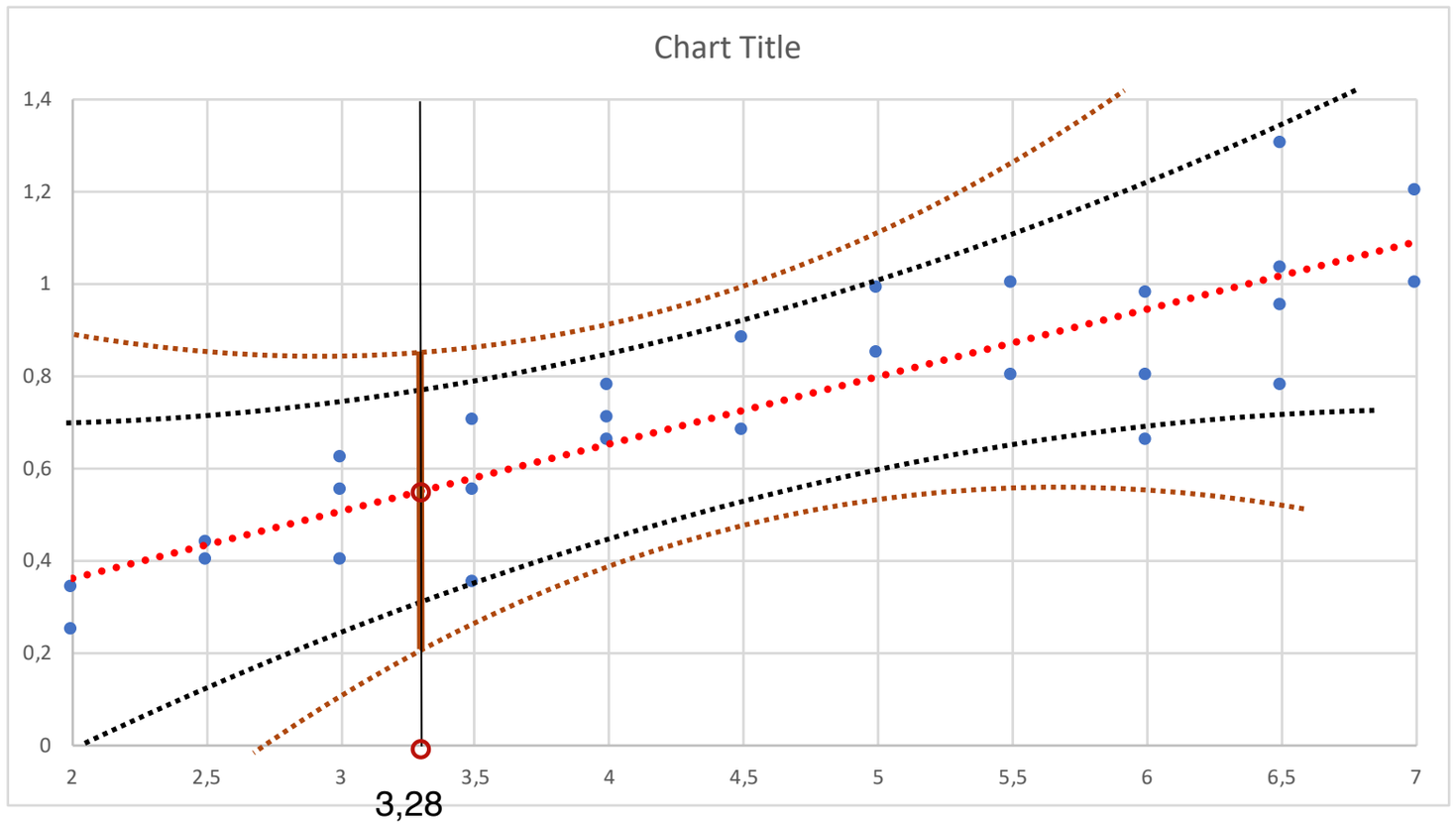
Lineární regresní model

Pás spolehlivosti pro predikci $Y(x_0) = a + bx_0$ **:**

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

kde

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



Lineární regresní model

Lineární model je „lineární“ proto, že je lineární v parametrech (lze jej zapsat maticově).

=>
- Model lineární regrese lze použít i v případech, kdy závislost mezi veličinami X a Y není lineární (viz třeba kvadratická regrese)
- Někdy lze použít tzv. „linearizaci modelu“, kterou původně nelineární model převedeme na lineární s transformovanými veličinami X a Y
- Při linearizaci musíme být opatrní: vše, co je odvozeno pro linearizovaný model za předpokladu normality chybového členu ϵ, platí pouze pro něj; nikoli pro model původní. A to opět za předpokladu, že transformovaná náhodná veličina v linearizovaném modelu má opět normální rozdělení.

Úloha: Pokuste se linearizovat následující nelineární modely:

$$Y = \ln(\alpha + \beta.X + \epsilon)$$

$$Y = \frac{1}{\alpha + \beta.X + \epsilon}$$

Lineární regresní model

Příklad: Závislost mezi teplotou nástroje θ a rychlostí posuvu ν při obrábění lze považovat za regresní závislost ve tvaru $\theta = \alpha.\nu^{\beta}.\epsilon$, kde α a β jsou regresní koeficienty a ϵ je náhodná veličina se střední hodnotou 1. Na základě měření odhadněte neznámé regresní koeficienty.

Předpokládaná závislost není lineární v parametrech => použijeme linearizaci:

zlogaritmováním vztahu mezi teplotou a posuvem dostaneme

$$\ln\theta = \ln\alpha + \beta.\ln\nu + \ln\epsilon.$$

Položíme-li $Y = \ln\theta, \quad X = \ln\nu, \quad a = \ln\alpha, \quad \varepsilon = \ln\epsilon, \quad b = \beta$, dostáváme lineární vztah

$$Y = a + b.X + \varepsilon$$

Dále postupujeme jako v případě přímkové regrese.

Poznámka: Pokud by mělo mít ε normální rozdělení, musí veličina ϵ v původním modelu mít logaritmicko-normální rozdělení LN(0, σ^2).

Vícerozměrný lineární regresní model

Lineární model pro dvě vysvětlující proměnné: $Y = a + bX + cZ + dXZ + \varepsilon$

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

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