

# Regrese, zejména lineární.

Petr Pošík

Části dokumentu jsou převzaty (i doslovně)  
z *Mirko Navara: Matematická statistika 2*,  
s laskavým svolením autora.

<b>Regrese</b>	<b>2</b>
Regrese	3
Stanovení neznámých parametrů	5
Metoda nejmenších čtverců	6
<b>Lineární regrese</b>	<b>7</b>
Lineární regrese	8
Regresní přímka 1	9
Příklad	10
Regresní přímka 2	11
Interpretace regresních koeficientů	12
Chyba lineární regrese	13
Rozklad rozptylu	14
Odhad rozptylu	16
Rozdělení odhadů	17
Testy hypotéz o regresních koeficientech	18

### Regrese

**Regrese** je náhrada pozorované závislosti vhodnou funkcí.

**Vstup:** pozorovaná data — realizace náhodného výběru ze sdruženého rozdělení n. vektoru  $(Y, X)$

$$(y, x) = ((y_1, x_1), \dots, (y_n, x_n)).$$

Předpokládáme, že tato data pocházejí ze statistického modelu

$$Y = g_\theta(X) + \mathcal{E},$$

kde

- $X$  je nezávislá **vysvětlující** náhodná veličina, jejíž hodnoty můžeme měřit (spojitá nebo diskrétní),
- $g_\theta: \mathbb{R} \rightarrow \mathbb{R}$  je funkce závislá na neznámých parametrech  $\theta \in \Theta$ , *kteřé potřebujeme odhadnout z dat*, tj. potřebujeme z množiny kandidátských funkcí  $g_\theta$  vybrat takovou  $g_\theta$ , která bude datům „co nejlépe odpovídat“,
- $\mathcal{E}$  je náhodná veličina s rozdělením  $N(0, \sigma^2)$  (šum) a
- $\sigma^2$  je konstantní (známý nebo neznámý) rozptyl (*homoskedasticita*).

Funkce  $g_\theta$  může mít nejrůznější podobu:

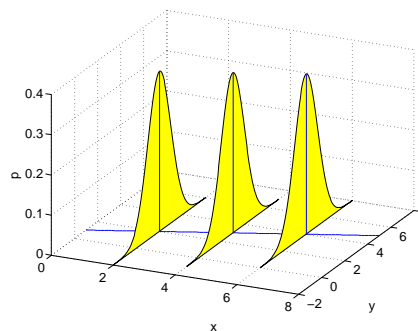
- konstanta:  $g_\theta(X) = \theta$ ,
- lineární funkce:  $g_\theta(X) = \theta_0 + \theta_1 X$ ,
- nelineární funkce s nejrůznějšími reprezentací: neuronová síť, regresní strom, ...

### Regrese (pokr.)

Protože n. veličina  $\mathcal{E}$  má rozdělení  $N(0, \sigma^2)$ , pro hodnoty podmíněné hustoty pravděpodobnosti veličiny  $Y$  platí

$$f_{Y|X}(y|x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y - g_\theta(x))^2}{2\sigma^2}\right)$$

(Vpravo příklad pro lineární  $g_\theta(x)$ .)



Pokud víme, že  $X = x$ , přejdeme k podmíněným pravděpodobnostem a jejich středním hodnotám:

$$E(Y|X = x) = E(g_\theta(X) + \mathcal{E}|X = x) = E(g_\theta(X)|X = x) + E(\mathcal{E}|X = x) = g_\theta(x),$$

tj. hodnoty funkce  $g_\theta(x)$  odhadují střední hodnoty  $E(Y|X = x)$ .

## Stanovení neznámých parametrů

Odhad parametrů  $\theta$  obecné funkce  $g_\theta(x)$  metodou maximální věrohodnosti při známém rozptylu  $\sigma^2$ :

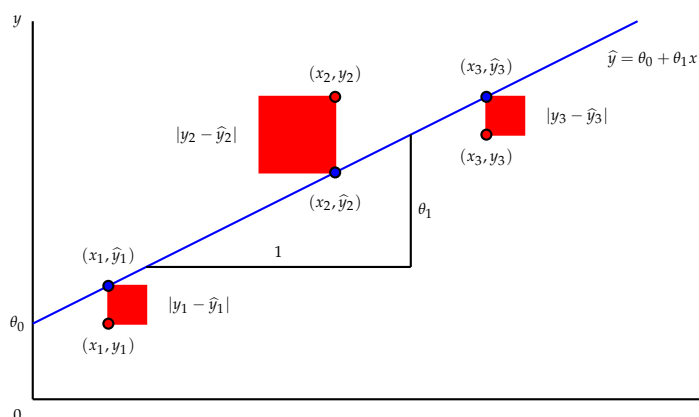
$$\begin{aligned}\Lambda(\theta) &= \prod_{j=1}^n f_{Y|X}(y_j|x_j) = \prod_{j=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_j - g_\theta(x_j))^2}{2\sigma^2}\right), \\ \lambda(\theta) &= \ln \Lambda(\theta) = \sum_{j=1}^n \ln f_{Y|X}(y_j|x_j) = \sum_{j=1}^n \left(-\ln \sigma\sqrt{2\pi} - \frac{(y_j - g_\theta(x_j))^2}{2\sigma^2}\right) = \\ &= \underbrace{-n \ln \sigma\sqrt{2\pi}}_{\text{konst.}} - \frac{1}{2\sigma^2} \sum_{j=1}^n \underbrace{(y_j - g_\theta(x_j))^2}_{\kappa(\theta)}, \\ \hat{\theta} &= \operatorname{argmax}_{\theta} \Lambda(\theta) = \operatorname{argmax}_{\theta} \lambda(\theta) = \\ &= \operatorname{argmin}_{\theta} \kappa(\theta) = \operatorname{argmin}_{\theta} \sum_{j=1}^n (y_j - g_\theta(x_j))^2.\end{aligned}$$

⇒ **Metoda nejmenších čtverců**

Bez předpokladu normality n.v.  $\mathcal{E}$  (ale má-li  $\mathcal{E}$  střední hodnotu), *odhad metodou nejmenších čtverců  $\neq$  max. věrohodný odhad.*

## Metoda nejmenších čtverců

Příklad pro lineární model:



**Lineární regrese**

Lineární model s parametry  $\theta = (\theta_0, \theta_1)$ :

$$Y = g_{\theta}(X) + \mathcal{E} = \theta_0 + \theta_1 X + \mathcal{E}$$

Odhad parametrů regresní přímky metodou nejmenších čtverců:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \kappa(\theta) = \underset{\theta}{\operatorname{argmin}} \sum_{j=1}^n (y_j - g_{\theta}(x_j))^2 = \underset{\theta_0, \theta_1}{\operatorname{argmin}} \sum_{j=1}^n (y_j - \theta_0 - \theta_1 x_j)^2$$

Pro odhad  $\hat{\theta}_0$  musí platit:

$$0 = \frac{\partial \kappa(\hat{\theta}_0, \hat{\theta}_1)}{\partial \hat{\theta}_0} = -2 \sum_{j=1}^n y_j + 2n\hat{\theta}_0 + 2\hat{\theta}_1 \sum_{j=1}^n x_j,$$

$$\hat{\theta}_0 = \frac{1}{n} \sum_{j=1}^n y_j - \hat{\theta}_1 \frac{1}{n} \sum_{j=1}^n x_j = \bar{y} - \hat{\theta}_1 \bar{x}.$$

Pro odhad  $\hat{\theta}_1$ :

$$0 = \frac{\partial \kappa(\hat{\theta}_0, \hat{\theta}_1)}{\partial \hat{\theta}_1} = -2 \sum_{j=1}^n x_j y_j + 2n\hat{\theta}_0 \bar{x} + 2\hat{\theta}_1 \sum_{j=1}^n x_j^2 = -2 \sum_{j=1}^n x_j y_j + 2n\bar{x}(\bar{y} - \hat{\theta}_1 \bar{x}) + 2\hat{\theta}_1 \sum_{j=1}^n x_j^2$$

$$\hat{\theta}_1 = \frac{\sum_{j=1}^n x_j y_j - n\bar{x}\bar{y}}{\sum_{j=1}^n x_j^2 - n\bar{x}^2} = \frac{\sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^n (x_j - \bar{x})^2}.$$

**Regresní přímka 1**

Regresní přímka je tvořena body  $(x, y)$ , které splňují rovnici

$$y = \hat{\theta}_0 + \hat{\theta}_1 x.$$

**Věta:** Bod  $(\bar{x}, \bar{y})$  leží na regresní přímce, tj.

$$\bar{y} = \hat{\theta}_0 + \hat{\theta}_1 \bar{x}.$$

**Důkaz:** Už jsme odvodili  $\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$ .  $\square$

Odečtením výše uvedených rovnic dostáváme tzv. úsekový tvar regresní přímky

$$y - \bar{y} = \hat{\theta}_1 (x - \bar{x})$$

Sklon (směrnice)  $\hat{\theta}_1$  se nezmění, přičteme-li konstantu ke všem hodnotám nezávisle proměnné  $X$ , nebo závisle proměnné  $Y$ . Mohli jsme od nich např. odečíst realizace výběrových průměrů  $\bar{x}$ , resp.  $\bar{y}$ , a zjednodušit si výrazy.

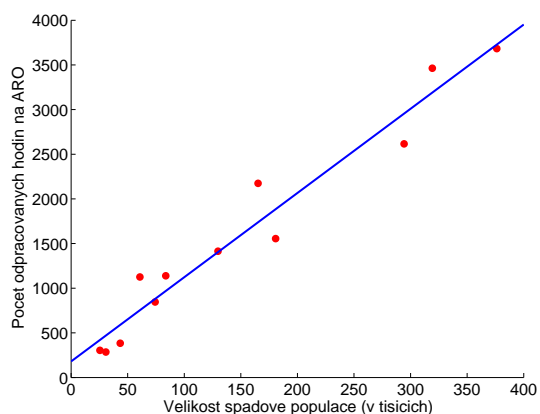
## Příklad

**Zadání:** Primář nově zřizovaného ARO potřebuje odhadnout, kolik personálu bude muset zaměstnat. Pro první přiblížení nasbíral data z dalších nemocnic o tom, jak spolu souvisí velikost spádové oblasti (v tis. obyvatel) a průměrný počet člověkohodin týdně odpracovaných na ARO.

$j$ : Nemocnice	$x_j$ : Populace [tis.]	$y_j$ : Hodin
1	25.5	304.37
2	294.3	2616.32
3	83.7	1139.12
4	30.7	285.43
5	129.8	1413.77
6	180.8	1555.68
7	43.4	383.78
8	165.2	2174.27
9	74.3	845.30
10	60.8	1125.28
11	319.2	3462.60
12	376.2	3682.33

$$\hat{\theta}_1 = \frac{\sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^n (x_j - \bar{x})^2} = 9.429 \text{ hod./tis. lidí}$$

$$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x} = 180.658 \text{ hod.}$$



$$y = \theta_0 + \theta_1 x = 180.658 + 9.429x$$

## Regresní přímka 2

Lze také odhadovat lineární závislost  $X$  na  $Y$  podle modelu

$$X = \theta_0^* + \theta_1^* Y + \mathcal{E}^*,$$

směrnice regresní přímky pak vyjde

$$\hat{\theta}_1^* = \frac{\sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^n (y_j - \bar{y})^2},$$

což je obecně *jiná regresní přímka*, jejíž rovnici lze psát ve tvarech

$$\begin{aligned} x &= \hat{\theta}_0^* + \hat{\theta}_1^* y, \\ x - \bar{x} &= \hat{\theta}_1^* (y - \bar{y}), \\ y - \bar{y} &= \frac{1}{\hat{\theta}_1^*} (x - \bar{x}). \end{aligned}$$

## Interpretace regresních koeficientů

Odhady rozptylů a kovariance:

$$\hat{\sigma}_x^2 := \frac{1}{n} \sum_j (x_j - \bar{x})^2 = \frac{n-1}{n} s_x^2 = D \text{Emp}(x),$$

$$\hat{\sigma}_y^2 := \frac{1}{n} \sum_j (y_j - \bar{y})^2 = \frac{n-1}{n} s_y^2 = D \text{Emp}(y),$$

$$c_{x,y} := \frac{1}{n} \sum_j (x_j - \bar{x})(y_j - \bar{y}) = \text{cov}(\text{Emp}(x, y)).$$

Odhad korelace:

$$r_{x,y} = \frac{\sum_j (x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_j (x_j - \bar{x})^2 \sum_j (y_j - \bar{y})^2}} = \frac{c_{x,y}}{\hat{\sigma}_x \hat{\sigma}_y} = \rho(\text{Emp}(x, y)).$$

Pro závislost  $Y$  na  $X$ :

$$\hat{\theta}_1 = \frac{c_{x,y}}{\hat{\sigma}_x^2} = r_{x,y} \frac{\hat{\sigma}_y}{\hat{\sigma}_x},$$

$$y - \bar{y} = \hat{\theta}_1 (x - \bar{x}),$$

$$y - \bar{y} = \frac{c_{x,y}}{\hat{\sigma}_x^2} (x - \bar{x}),$$

$$\frac{y - \bar{y}}{\hat{\sigma}_y} = r_{x,y} \frac{x - \bar{x}}{\hat{\sigma}_x}.$$

Pro závislost  $X$  na  $Y$  (to je *jiná přímka!*):

$$\hat{\theta}_1^* = \frac{c_{x,y}}{\hat{\sigma}_y^2} = r_{x,y} \frac{\hat{\sigma}_x}{\hat{\sigma}_y},$$

$$x - \bar{x} = \hat{\theta}_1^* (y - \bar{y}),$$

$$x - \bar{x} = \frac{c_{x,y}}{\hat{\sigma}_y^2} (y - \bar{y}),$$

$$\frac{x - \bar{x}}{\hat{\sigma}_x} = r_{x,y} \frac{y - \bar{y}}{\hat{\sigma}_y}.$$

Obě směrnice mají stejná znaménka, součin  $\hat{\theta}_1 \hat{\theta}_1^* = \frac{c_{x,y}^2}{\hat{\sigma}_x^2 \hat{\sigma}_y^2} = r_{x,y}^2$ , takže  $r_{x,y} = \sqrt{\hat{\theta}_1 \hat{\theta}_1^*} \text{sign}(\hat{\theta}_1)$ .

## Chyba lineární regrese

Odhadli jsme lineární regresní funkci  $g_{\hat{\theta}}(x) = \hat{\theta}_0 + \hat{\theta}_1 x$ , pomocí ní odhadneme hodnoty závisle proměnné  $Y$  v jednotlivých realizacích

$$\hat{y}_j = g_{\hat{\theta}}(x) = \hat{\theta}_0 + \hat{\theta}_1 x_j,$$

$$\hat{\mathbf{y}} = (\hat{y}_1, \dots, \hat{y}_n)$$

a chyby (**rezidua**)

$$\hat{e}_j = y_j - \hat{y}_j = y_j - \hat{\theta}_0 - \hat{\theta}_1 x_j = y_j - \bar{y} - \hat{\theta}_1 (x_j - \bar{x}),$$

$$\hat{\mathbf{e}} = (\hat{e}_1, \dots, \hat{e}_n).$$

**Věta:**  $\frac{1}{n} \sum_j \hat{y}_j = \bar{y}$ .

**Důkaz:**  $\frac{1}{n} \sum_j \hat{y}_j = \frac{1}{n} \sum_j (\hat{\theta}_0 + \hat{\theta}_1 x_j) = \hat{\theta}_0 + \hat{\theta}_1 \bar{x} = \bar{y}$ .  $\square$

## Rozklad rozptylu

**Celkový rozptyl** (angl. *total variation*):

$$\hat{\sigma}_y^2 = \frac{1}{n} \sum_j (y_j - \bar{y})^2.$$

**Rozptyl modelu, vysvětlený rozptyl** (angl. *explained variation*):

$$\hat{\sigma}_{\hat{y}}^2 = \frac{1}{n} \sum_j (\hat{y}_j - \bar{y})^2.$$

**Reziduální rozptyl, nevysvětlený rozptyl** (angl. *unexplained variation*):

$$\hat{\sigma}_e^2 = \frac{1}{n} \sum_j \hat{e}_j^2 = \frac{1}{n} \sum_j (y_j - \hat{y}_j)^2.$$

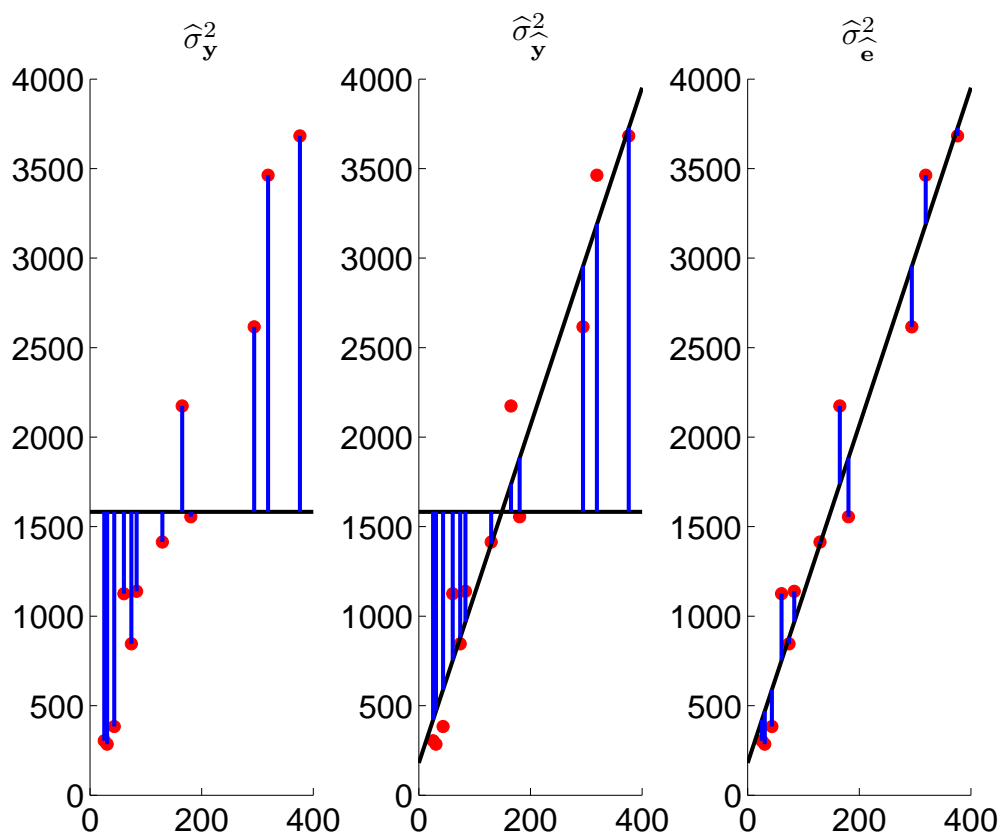
Někdy se ve výše uvedených vzorcích nedělí  $n$ ; pak se tyto veličiny někdy označují jako **součet čtverců**.

**Věta:**  $\hat{\sigma}_y^2 = \hat{\sigma}_{\hat{y}}^2 + \hat{\sigma}_e^2$ .

**Věta: Koefficient determinace**

$$r_{x,y}^2 = \frac{\hat{\sigma}_{\hat{y}}^2}{\hat{\sigma}_y^2} = 1 - \frac{\hat{\sigma}_e^2}{\hat{\sigma}_y^2}.$$

## Příklad: rozklad rozptylu



## Odhad rozptylu

Připomeňme: předpokládáme model  $Y = \theta_0 + \theta_1 X + \mathcal{E}$ , kde  $\mathcal{E} \sim N(0, \sigma^2)$ .

Max. věrohodný odhad  $\sigma^2$  rozdělení  $\mathcal{E}$ :

$$\begin{aligned}\lambda(\theta_0, \theta_1, \sigma) &= -n \ln \sigma \sqrt{2\pi} - \frac{1}{2\sigma^2} \sum_j (y_j - \theta_0 - \theta_1 x_j)^2 = \\ &= \underbrace{-n \ln \sqrt{2\pi}}_{\text{konst.}} - n \ln \sigma - \frac{1}{2\sigma^2} \sum_j e_j^2, \\ 0 &= \frac{\partial \lambda(\theta_0, \theta_1, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_j e_j^2.\end{aligned}$$

Řešením je

$$\hat{\sigma}_e^2 = \frac{1}{n} \sum_j \hat{e}_j^2 = \text{D Emp}(\hat{e}),$$

což je *vychýlený* odhad; *nestranný* odhad je

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_j \hat{e}_j^2 = \frac{1}{n-2} \sum_j (y_j - \hat{\theta}_0 - \hat{\theta}_1 x_j)^2.$$

## Rozdělení odhadů

**Věta:** Odhady  $\hat{\theta}_0, \hat{\theta}_1, \hat{\sigma}^2$  skutečných parametrů  $\theta_0, \theta_1, \sigma^2$  jsou nestranné, konzistentní a asymptoticky normální.

Odhady rozptylů regresních koeficientů

$$\begin{aligned}\hat{\sigma}_{\theta_0}^2 &= \frac{\hat{\sigma}^2}{n^2 \hat{\sigma}_x^2} \sum_j x_j^2 = \frac{\hat{\sigma}_e^2}{n(n-2) \hat{\sigma}_x^2} \sum_j x_j^2, \\ \hat{\sigma}_{\theta_1}^2 &= \frac{\hat{\sigma}^2}{n \hat{\sigma}_x^2} = \frac{\hat{\sigma}_e^2}{(n-2) \hat{\sigma}_x^2}.\end{aligned}$$

*nejsou nezávislé.*

Rozdělení odhadů  $\hat{\theta}_0, \hat{\theta}_1$  regresních koeficientů se blíží normálnímu rozdělení:

$$\begin{aligned}\text{rozdělení } \hat{\theta}_0 &\text{ se blíží } N(\theta_0, \sigma_{\theta_0}^2) \approx N\left(\theta_0, \frac{\hat{\sigma}_e^2}{n(n-2) \hat{\sigma}_x^2} \sum_j x_j^2\right), \\ \text{rozdělení } \hat{\theta}_1 &\text{ se blíží } N(\theta_1, \sigma_{\theta_1}^2) \approx N\left(\theta_1, \frac{\hat{\sigma}_e^2}{(n-2) \hat{\sigma}_x^2}\right).\end{aligned}$$



## Testy hypotéz o regresních koeficientech

**Test absolutního členu**, tj. např. nulové hypotézy  $H_0 : \theta_0 = c$ :  
realizaci testové statistiky

$$t = \frac{\hat{\theta}_0 - c}{\frac{\hat{\sigma}_e}{\hat{\sigma}_x} \sqrt{\frac{1}{n} \sum_j x_j^2}} \sqrt{n-2}$$

testujeme na rozdělení  $t(n-2)$ .

**Test směrnice**, tj. např. nulové hypotézy  $H_0 : \theta_1 = c$ :  
realizaci testové statistiky

$$t = \frac{\hat{\theta}_1 - c}{\frac{\hat{\sigma}_e}{\hat{\sigma}_x}} \sqrt{n-2}$$

testujeme na rozdělení  $t(n-2)$ .

## Příklad: typické výstupy stat. programů

- Testy regresních koeficientů (t-testy):

Parametr	Koeficient	SE koeficientu	t	p
Absolutní člen $\hat{\theta}_0$	180.658	128.31	1.407	0.1897
Směrnice $\hat{\theta}_1$	9.429	0.681	13.846	0.0000

- Test modelu jako celku (F-test)

Zdroj variability	Součet čtverců SS	Stupně volnosti d.f.	Průměrný čtverec $MS = \frac{SS}{d.f.}$	Poměr F
Regrese (vysvětlená)	$\hat{\theta}_1^2 \sum (X - \bar{X})^2$	1	$\frac{\hat{\theta}_1^2 \sum (X - \bar{X})^2}{1}$	$\frac{\hat{\theta}_1^2 \sum (X - \bar{X})^2}{s^2}$
Zbytek (nevysvětlená)	$\sum (Y - \hat{Y})^2$	$n - 2$	$s^2 = \frac{\sum (Y - \hat{Y})^2}{n-2}$	
Celkem	$\sum (Y - \bar{Y})^2$	$n - 1$		

Konkrétně pro náš příklad:

Zdroj variability	Souč. čtverců SS	Stup. volnosti d.f.	Prům. čtverec $MS = \frac{SS}{d.f.}$	Poměr F	Hladina p
Regrese (vysvětlená)	14346071	1	14346071	191.7	0.0000
Zbytek (nevysvětlená)	748192	10	74819.2		
Celkem	15094263	11			