Regrese, zejména lineární.

Petr Pošík

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

Regrese

Regrese je náhrada pozorované závislosti vhodnou funkcí.

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

$$(\boldsymbol{y}, \boldsymbol{x}) = ((y_1, x_1), \ldots, (y_n, x_n)).$$

Předpokládáme, že tato data pocházejí ze statistického modelu

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

kde

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

Funkce $g_{\boldsymbol{\Theta}}$ může mít nejrůznější podobu:

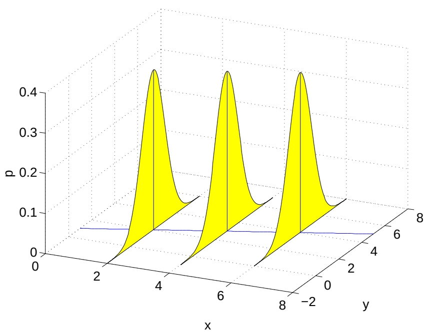
- konstanta: $g_{\boldsymbol{\theta}}(X) = \theta$,
- lineární funkce: $g_{\boldsymbol{\theta}}(X) = \theta_0 + \theta_1 X$,
- nelineární funkce s nejrůznější reprezentací: neuronová síť, regresní strom, ...

Regrese (pokr.)

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

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

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



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

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

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

Stanovení neznámých parametrů

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

$$\Lambda(\boldsymbol{\theta}) = \prod_{j=1}^{n} f_{Y|X}(y_j|x_j) = \prod_{j=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_j - g_{\boldsymbol{\theta}}(x_j))^2}{2\sigma^2}\right),$$

$$\lambda(\boldsymbol{\theta}) = \ln \Lambda(\boldsymbol{\theta}) = \sum_{j=1}^{n} \ln f_{Y|X}(y_j|x_j) = \sum_{j=1}^{n}\left(-\ln\sigma\sqrt{2\pi} - \frac{(y_j - g_{\boldsymbol{\theta}}(x_j))^2}{2\sigma^2}\right) =$$

$$= \underbrace{-n\ln\sigma\sqrt{2\pi}}_{\text{konst.}} - \underbrace{\frac{1}{2\sigma^2}}_{\text{konst.}} \underbrace{\sum_{j=1}^{n}(y_j - g_{\boldsymbol{\theta}}(x_j))^2}_{\kappa(\boldsymbol{\theta})},$$

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \Lambda(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \lambda(\boldsymbol{\theta}) =$$

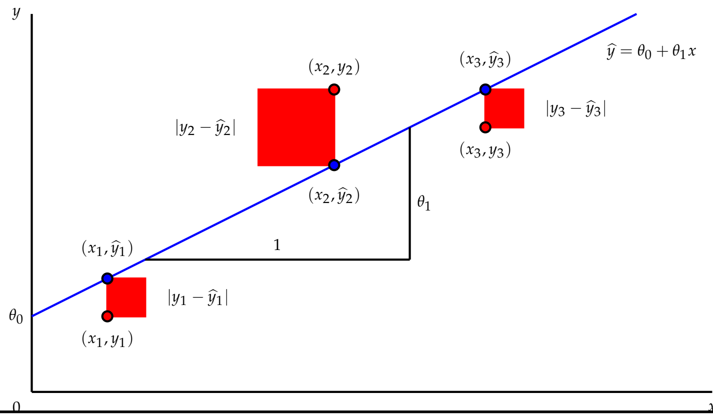
$$= \operatorname*{argmin}_{\boldsymbol{\theta}} \kappa(\boldsymbol{\theta}) = \operatorname{argmin} \sum_{j=1}^{n}(y_j - g_{\boldsymbol{\theta}}(x_j))^2.$$

$\Longrightarrow$ **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 $\boldsymbol{\theta} = (\theta_0, \theta_1)$:

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

Odhad parametrů regresní přímky metodou nejmenších čtverců:

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

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

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

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

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

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

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

Regresní přímka 1

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

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

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

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

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

Odečtením výše uvedených rovnic dostáváme tzv. úsekový tvar regresní přímky

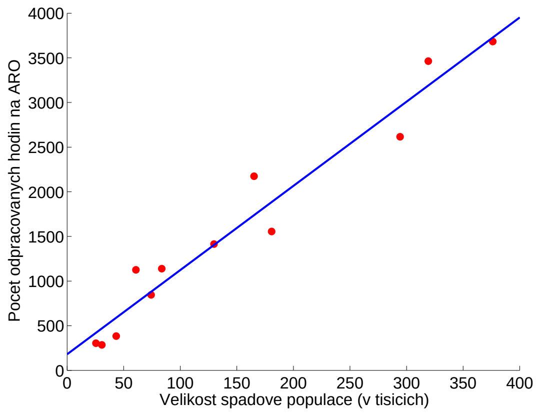
$$y - \bar{y} = \widehat{\theta}_1 (x - \bar{x})$$

Sklon (směrnice) $\widehat{\theta}_1$ 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



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

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

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

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

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

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

$$x = \widehat{\theta}_0^* + \widehat{\theta}_1^* y,$$
$$x - \bar{x} = \widehat{\theta}_1^*(y - \bar{y}),$$
$$y - \bar{y} = \frac{1}{\widehat{\theta}_1^*}(x - \bar{x}).$$

Interpretace regresních koeficientů

Odhady rozptylů a kovariance:

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

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

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

Odhad korelace:

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

Pro závislost Y na X:

$$\widehat{\theta}_1 = \frac{c_{x,y}}{\widehat{\sigma}_x^2} = r_{x,y}\frac{\widehat{\sigma}_y}{\widehat{\sigma}_x},$$
$$y - \bar{y} = \widehat{\theta}_1 (x - \bar{x}),$$
$$y - \bar{y} = \frac{c_{x,y}}{\widehat{\sigma}_x^2}(x - \bar{x}),$$
$$\frac{y - \bar{y}}{\widehat{\sigma}_y} = r_{x,y}\frac{x - \bar{x}}{\widehat{\sigma}_x}.$$

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

$$\widehat{\theta}_1^* = \frac{c_{x,y}}{\widehat{\sigma}_y^2} = r_{x,y}\frac{\widehat{\sigma}_x}{\widehat{\sigma}_y},$$
$$x - \bar{x} = \widehat{\theta}_1^* (y - \bar{y}),$$
$$x - \bar{x} = \frac{c_{x,y}}{\widehat{\sigma}_y^2}(y - \bar{y}),$$
$$\frac{x - \bar{x}}{\widehat{\sigma}_x} = r_{x,y}\frac{y - \bar{y}}{\widehat{\sigma}_y}.$$

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

Chyba lineární regrese

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

$$\widehat{y}_j = g_{\widehat{\boldsymbol{\theta}}}(x) = \widehat{\theta}_0 + \widehat{\theta}_1 x_j,$$
$$\widehat{\boldsymbol{y}} = (\widehat{y}_1, \ldots, \widehat{y}_n)$$

a chyby (**rezidua**)

$$\widehat{e}_j = y_j - \widehat{y}_j = y_j - \widehat{\theta}_0 - \widehat{\theta}_1 x_j = y_j - \bar{y} - \widehat{\theta}_1 (x_j - \bar{x}),$$
$$\widehat{\boldsymbol{e}} = (\widehat{e}_1, \ldots, \widehat{e}_n).$$

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

Důkaz: $\frac{1}{n}\sum_j \widehat{y}_j = \frac{1}{n}\sum_j \left(\widehat{\theta}_0 + \widehat{\theta}_1 x_j\right) = \widehat{\theta}_0 + \widehat{\theta}_1 \bar{x} = \bar{y}$. □

Rozklad rozptylu

Celkový rozptyl (angl. *total variation*):

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

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

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

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

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

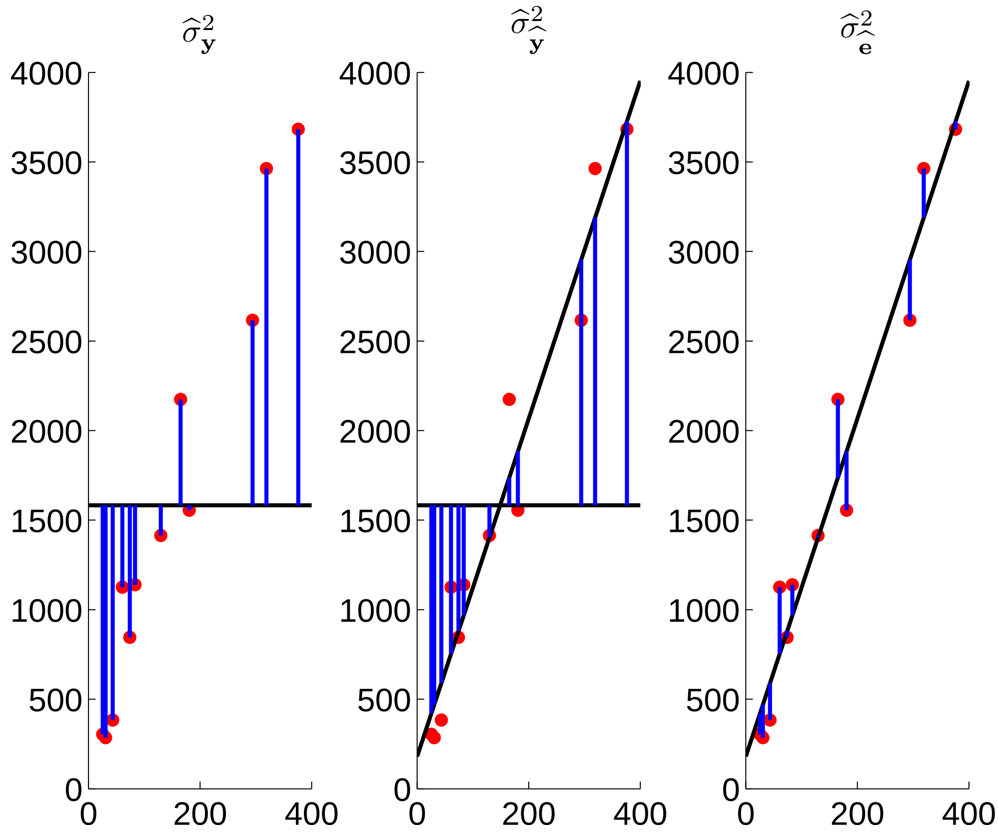
Někdy se ve výše uvedených vzorcích nedělí n; pak se tyto veličiny někdy označují jako **součet čtverců**.

Věta: $\widehat{\sigma}_y^2 = \widehat{\sigma}_{\widehat{y}}^2 + \widehat{\sigma}_{\widehat{e}}^2$.

Věta: **Koeficient determinace**

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

Příklad: rozklad rozptylu



Odhad rozptylu

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

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

$$\lambda(\theta_0, \theta_1, \sigma) = -n \ln \sigma \sqrt{2\pi} - \frac{1}{2\sigma^2} \sum_j (y_j - \theta_0 - \theta_1 x_j)^2 =$$

$$= \underbrace{-n \ln \sqrt{2\pi}}_{\text{konst.}} - n \ln \sigma - \frac{1}{2\sigma^2} \sum_j e_j^2,$$

$$0 = \frac{\partial \lambda(\theta_0, \theta_1, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_j e_j^2.$$

Řešením je

$$\widehat{\sigma}_{\widehat{e}}^2 = \frac{1}{n} \sum_j \widehat{e}_j^2 = \mathrm{D}\,\mathrm{Emp}(\widehat{e}),$$

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

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

Rozdělení odhadů

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

Odhady rozptylů regresních koeficientů

$$\widehat{\sigma}_{\widehat{\theta}_0}^2 = \frac{\widehat{\sigma}^2}{n^2 \widehat{\sigma}_x^2} \sum_j x_j^2 = \frac{\widehat{\sigma}_{\widehat{e}}^2}{n(n-2)\widehat{\sigma}_x^2} \sum_j x_j^2,$$

$$\widehat{\sigma}_{\widehat{\theta}_1}^2 = \frac{\widehat{\sigma}^2}{n \widehat{\sigma}_x^2} = \frac{\widehat{\sigma}_{\widehat{e}}^2}{(n-2)\widehat{\sigma}_x^2}.$$

nejsou nezávislé.

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

$$\text{rozdělení } \widehat{\theta}_0 \text{ se blíží } \mathrm{N}(\theta_0, \sigma_{\widehat{\theta}_0}^2) \approx \mathrm{N}\left(\theta_0, \frac{\widehat{\sigma}_{\widehat{e}}^2}{n(n-2)\widehat{\sigma}_x^2} \sum_j x_j^2\right),$$

$$\text{rozdělení } \widehat{\theta}_1 \text{ se blíží } \mathrm{N}(\theta_1, \sigma_{\widehat{\theta}_1}^2) \approx \mathrm{N}\left(\theta_1, \frac{\widehat{\sigma}_{\widehat{e}}^2}{(n-2)\widehat{\sigma}_x^2}\right).$$

Testy hypotéz o regresních koeficientech

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

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

testujeme na rozdělení $t(n-2)$.

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

$$t = \frac{\widehat{\theta}_1 - c}{\frac{\widehat{\sigma}_{\widehat{e}}}{\widehat{\sigma}_x}}\sqrt{n-2}$$

testujeme na rozdělení $t(n-2)$.

Příklad: typické výstupy stat. programů

- Testy regresních koeficientů (t-testy):

Parametr	Koeficient	SE koeficientu	t	p
Absolutní člen $\widehat{\theta}_0$	180.658	128.31	1.407	0.1897
Směrnice $\widehat{\theta}_1$	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)	$\widehat{\theta}_1^2 \sum(X-\hat{X})^2$	1	$\frac{\widehat{\theta}_1^2 \sum(X-\hat{X})^2}{1}$	$\frac{\widehat{\theta}_1^2 \sum(X-\hat{X})^2}{s^2}$
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			