

Regresní a korelační analýza

Regresní analýza – určování funkční závislosti na základě empirických modelů

Korelační analýza – stanovení míry těsnosti (závislosti, korelace) mezi veličinami

V přírodních vědách velmi často známe souvislosti mezi různými jevy (veličinami), které se řídí přírodními zákony a často je umíme popsat nějakými pravidly, nebo dokonce známe i funkční vztahy mezi nimi, např.: Archimedův zákon, Newtonův gravitační zákon, zákony optiky aj..

V přírodních a technických vědách často známe modelový systém (funkční vztahy), ale **parametry modelu** určujeme až na základě řízeného experimentu (nastavujeme/měříme příčiny a určujeme/měříme následky).

Např.: stanovení indexu lomu pro určitý materiál, určení koeficientu délkové roztažnosti, určení kalibračních hodnot přístrojů či konstanty hranolu aj.

Společenské vědy se vyznačují velmi nepřehlednými vztahy a obtížně definovatelnými a zachytitelnými vlivy. Určování souvislostí na základě experimentů má zde obvykle časově omezenou platnost a takto definované modely lze jen velmi obtížně verifikovat opakovanými pokusy.

Regresní analýza

(určování funkčních závislostí na základě experimentu/měření)

Regrese – systematické změny jedné veličiny při změně veličin jiných a zobrazení těchto změn pomocí matematických funkcí.

$$y = f(x), \quad y = f(x_1, x_2, \dots, x_k)$$

Vysvětlující proměnná/proměnné – též nezávislá proměnná x

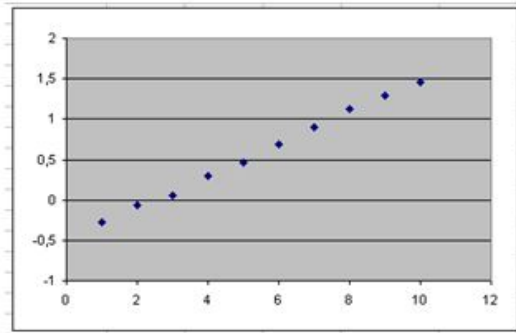
Vysvětlovaná proměnná/proměnné – též závislá proměnná y

Nejprve definujeme matematický model, tj. stanovíme/určíme **tvár funkce** (lineární, nelineární), počet parametrů a jejich rozsah. Potom na základě výsledků **experimentu/měření** určíme **parametry** použité **funkce**. U těchto experimentů máme často možnost měnit (nastavovat) vysvětlující proměnné x . U některých experimentů/modelů vybíráme/registrujeme hodnoty x a měříme/registrujeme hodnoty y . Tvar funkce a její parametry určujeme až jako výsledek následné analýzy. Časté jsou také modely, ve kterých měřené (tj. zatížené měřickými chybami) jsou jak veličiny vysvětlující x , tak i veličiny vysvětlované y .

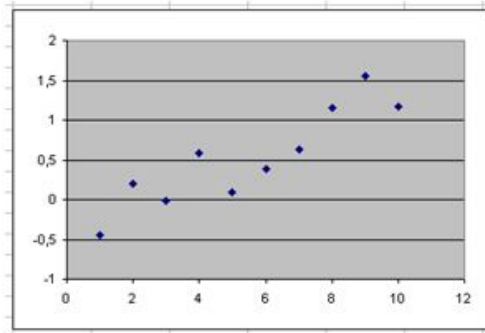
Definujeme-li na základě regresní analýzy vhodný matematický model, bylo by velmi žádoucí také stanovit podmínky jeho použitelnosti a ověřit jej dalším nezávislým experimentem (tzv. verifikace modelu).

Korelační analýza

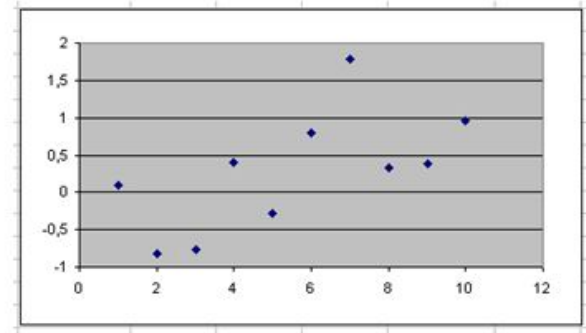
silná závislost



střední závislost

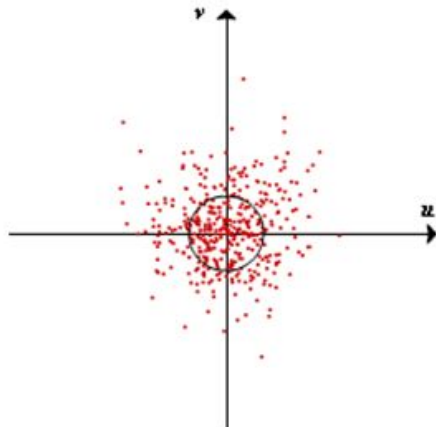


slabá závislost

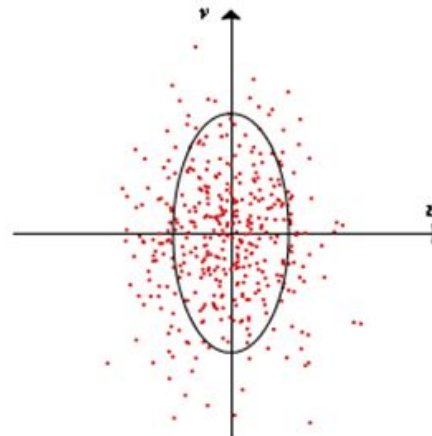


Ve všech třech případech lze určit parametry lineární funkce (budou přibližně stejné), ale s různou mírou těsnosti vztahu (korelace).

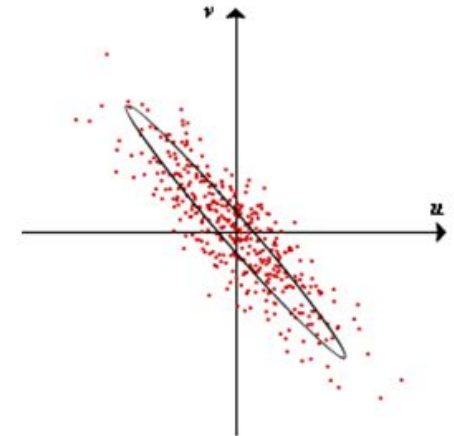
nezávislé veličiny



nezávislé veličiny



závislé veličiny



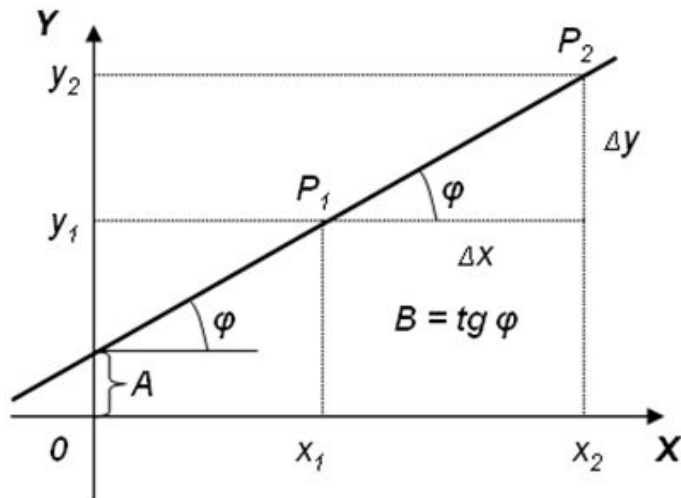
Lineární empirická funkce

(regresní přímka, vyrovnávací přímka)

$$y = A + Bx, \quad y = \beta_0 + \beta_1 x$$

Přímka je dána 2 parametry A – konstantní koeficient, B – násobný koeficient (používá se různé označení těchto parametrů β_0, β_1 ; k_1, k_2 ; p, q aj).

Pro určení dvou parametrů A a B musíme znát minimálně dvě dvojice dat, čili dva různé body v rovině x, y . $P_1 \equiv (x_1, y_1)$ a $P_2 \equiv (x_2, y_2)$



parametr A = úsek na ose Y (pro $x = 0$)
parametr $B = \operatorname{tg} \varphi$ (směrnice přímky)

Výpočet parametrů při zadaných 2 bodech:

$$B = \operatorname{tg} \varphi = \Delta y / \Delta x = (y_2 - y_1) / (x_2 - x_1)$$

$$A = y_1 - B x_1 = y_2 - B x_2$$

V tomto případě přímka prochází právě oběma body.

Pozor při $\Delta x = 0$

Regresní (vyrovnávací) přímka při nadbytečných měřeních

- Chybami budou zatíženy jen hodnoty y – hodnoty x považujeme za bezchybné (standardní případ)
- Chybami budou zatíženy jen hodnoty x – hodnoty y považujeme za bezchybné
- Chybami budou zatíženy obě veličiny x i y

V dalším budeme předpokládat, že se jedná o měřické chyby s normálním rozdělením pravděpodobnosti. Měřené veličiny mohou mít také různou přesnost.

n = počet měření (bodů), jedno měření = jeden bod (dvě souřadnice),
 k = počet nutných měření = počet určovaných parametrů (u přímky $k=2$, A a B),
 $n - k = n - 2$ = počet nadbytečných měření (stupňů volnosti).

<i>Pořadí</i>	1	2	...	n
<i>Veličina x</i>	x_1	x_2	...	x_n
<i>Veličina y</i>	y_1	y_2	...	y_n

A) Chybami jsou zatíženy jen hodnoty y_i

Aplikujeme MNČ ($\sum p v v = \min$) a sice **Vyrovňání zprostředkujících měření**:

(měření y je funkcí neznámých A, B): $y = f(A, B) = f(\beta_0, \beta_1)$

Zprostředkující funkce

$$\tilde{y}_1 = \tilde{A} + \tilde{B}\tilde{x}_1$$

$$\tilde{y}_2 = \tilde{A} + \tilde{B}\tilde{x}_2$$

\vdots

$$\tilde{y}_n = \tilde{A} + \tilde{B}\tilde{x}_n$$

Rovnice oprav

$$v_1 = \hat{A} + \hat{B}x_1 - y_1$$

$$v_2 = \hat{A} + \hat{B}x_2 - y_2$$

\vdots

$$v_n = \hat{A} + \hat{B}x_n - y_n$$

Maticový zápis
rovníc oprav

$$\mathbf{v} = \mathbf{A}\boldsymbol{\beta} - \mathbf{l}$$

Vektor měření \mathbf{l}

Vektor oprav \mathbf{v}

Matice plánu \mathbf{A}

Vektor neznámých $\boldsymbol{\beta}$

$$\mathbf{l} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}, \quad \mathbf{A}_{(n,2)} = \begin{pmatrix} \frac{\partial y_1}{\partial A} & \frac{\partial y_1}{\partial B} \\ \frac{\partial y_2}{\partial A} & \frac{\partial y_2}{\partial B} \\ \vdots & \vdots \\ \frac{\partial y_n}{\partial A} & \frac{\partial y_n}{\partial B} \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \hat{A} \\ \hat{B} \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

Normální rovnice a jejich řešení

Normální rovnice maticově

$$\begin{aligned} \mathbf{A}^T \mathbf{A} \boldsymbol{\beta} - \mathbf{A}^T \mathbf{l} &= \mathbf{0} & \begin{pmatrix} n & \sum x \\ \sum x & \sum x^2 \end{pmatrix} \begin{pmatrix} \hat{A} \\ \hat{B} \end{pmatrix} - \begin{pmatrix} \sum y \\ \sum xy \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\ \mathbf{N} \boldsymbol{\beta} - \mathbf{n} &= \mathbf{0} \end{aligned}$$

$$\mathbf{N} = \begin{pmatrix} n & \sum x \\ \sum x & \sum x^2 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \hat{A} \\ \hat{B} \end{pmatrix}, \quad \mathbf{n} = \begin{pmatrix} \sum y \\ \sum xy \end{pmatrix}, \quad \mathbf{0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Řešení normálních rovnic

$$\boldsymbol{\beta} = \mathbf{N}^{-1} \mathbf{n}$$

$$\begin{pmatrix} \hat{A} \\ \hat{B} \end{pmatrix} = \begin{pmatrix} \frac{\sum x^2}{\det N} & -\frac{\sum x}{\det N} \\ -\frac{\sum x}{\det N} & \frac{n}{\det N} \end{pmatrix} \begin{pmatrix} \sum y \\ \sum xy \end{pmatrix}$$

$$\det N = n \sum x^2 - (\sum x)^2$$

Výpočet určovaných parametrů klasicky

$$\hat{A} = \frac{\sum x^2 \sum y - \sum x \sum xy}{n \sum x^2 - (\sum x)^2}$$

$$\hat{B} = \frac{-\sum x \sum y + n \sum xy}{n \sum x^2 - (\sum x)^2}$$

Vážená varianta určení regresní přímky

Zavedeme matici vah

$$P = \begin{pmatrix} p_1 & 0 & \dots & 0 \\ 0 & p_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & p_n \end{pmatrix}, \text{ kde } p_i = \frac{\sigma_0^2}{\sigma_i^2}$$

Řešení normálních rovnic

$$\boldsymbol{\beta} = N^{-1} \mathbf{n}$$

$$\begin{pmatrix} \hat{A} \\ \hat{B} \end{pmatrix} = \begin{pmatrix} \frac{\sum px^2}{\det N} & -\frac{\sum px}{\det N} \\ -\frac{\sum px}{\det N} & \frac{\sum p}{\det N} \end{pmatrix} \begin{pmatrix} \sum py \\ \sum pxy \end{pmatrix}$$

$$\det N = \sum p \sum px^2 - (\sum px)^2$$

Normální rovnice

$$A^T P A \boldsymbol{\beta} - A^T P \mathbf{l} = \mathbf{0}$$

$$N \boldsymbol{\beta} - \mathbf{n} = \mathbf{0}$$

$$\begin{pmatrix} \sum p & \sum px \\ \sum px & \sum px^2 \end{pmatrix} \begin{pmatrix} \hat{A} \\ \hat{B} \end{pmatrix} - \begin{pmatrix} \sum py \\ \sum pxy \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Klasický postup výpočtu neznámých

$$\hat{A} = \frac{\sum px^2 \sum py - \sum px \sum pxy}{\sum p \sum px^2 - (\sum px)^2}$$

$$\hat{B} = \frac{-\sum px \sum py + \sum p \sum pxy}{\sum p \sum px^2 - (\sum px)^2}$$

Charakteristiky přesnosti

1. Výpočet oprav z rovnic oprav

$$v_i = \hat{A} + \hat{B}x_i - y_i$$

2. Výpočet vyrovnaných měření

$$\hat{y}_i = \hat{A} + \hat{B}x_i = y_i + v_i$$

3. Aposteriorní jednotková variance

$$\hat{\sigma}_0^2 = \frac{\mathbf{v}^T P \mathbf{v}}{n-2} = \frac{\sum p v^2}{n-2}$$

4. Aposteriorní směrodatné odchylky měřených veličin

$$\sigma_i = \hat{\sigma}_0 \sqrt{q_i}$$

5. Směrodatné odchylky neznámých

$$Q = N^{-1} = \begin{pmatrix} Q_{\hat{A}\hat{A}} & Q_{\hat{A}\hat{B}} \\ Q_{\hat{A}\hat{B}} & Q_{\hat{B}\hat{B}} \end{pmatrix}, \quad \sigma_{\hat{A}} = \hat{\sigma}_0 \sqrt{Q_{\hat{A}\hat{A}}}, \quad \sigma_{\hat{B}} = \hat{\sigma}_0 \sqrt{Q_{\hat{B}\hat{B}}}$$

Pozn: V případě menšího počtu nadbytečných měření použijeme apriorní varianci či apriorní váhy

Použití přibližných hodnot A^0 , B^0

V mnoha případech je vhodné volit přibližné hodnoty neznámých

$$\begin{aligned}v_i &= (A^0 + \delta A) + (B^0 + \delta B)x_i - y_i \\v_i &= \delta A + \delta B x_i + (A^0 + B^0 x_i - y_i) \\v_i &= \delta A + \delta B x_i + y'_i\end{aligned} \quad \boldsymbol{l}' = \begin{pmatrix} y'_1 \\ y'_2 \\ \vdots \\ y'_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \hat{A} \\ \hat{B} \end{pmatrix} = \boldsymbol{\beta}^0 + \delta \boldsymbol{\beta} = \begin{pmatrix} A^0 \\ B^0 \end{pmatrix} + \begin{pmatrix} \delta A \\ \delta B \end{pmatrix}.$$

Rovnice oprav

$$\boldsymbol{v} = \boldsymbol{A} \delta \boldsymbol{\beta} + \boldsymbol{l}'$$

Normální rovnice

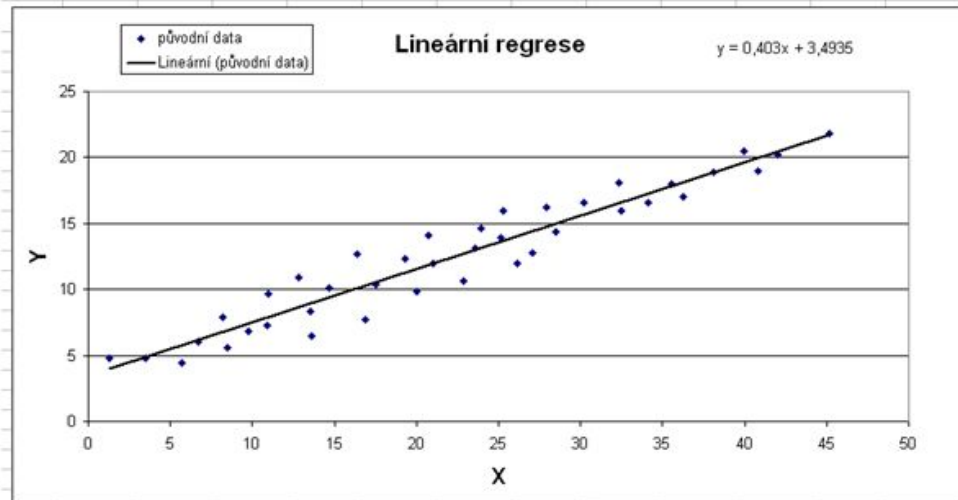
$$\boldsymbol{A}^T \boldsymbol{A} \delta \boldsymbol{\beta} + \boldsymbol{A}^T \boldsymbol{l}' = \mathbf{0}$$

Řešení normálních rovnic

$$\delta \boldsymbol{\beta} = \boldsymbol{N}^{-1} \boldsymbol{n}'$$

Další postup je obdobný i pro váženou variantu

Zavedení redukováných souřadnic (redukce na těžiště)

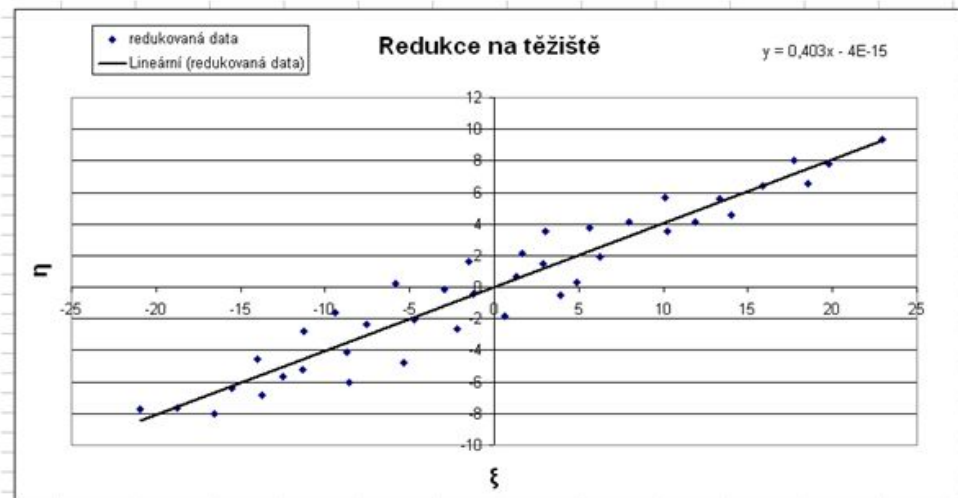


$$\xi_i = x_i - x_T, \quad \eta_i = y_i - y_T,$$

$$x_T = \frac{\sum x}{n}, \quad y_T = \frac{\sum y}{n}.$$

$$v_i = \hat{B}\xi_i - \eta_i,$$

$$\sum \xi^2 \hat{B} - \sum \xi \eta = 0.$$



$$\hat{B} = \frac{\sum \xi \eta}{\sum \xi^2}$$

$$y_T = \hat{A} + \hat{B}x_T$$

$$\hat{A} = y_T - \hat{B}x_T$$

$$\hat{\sigma}_0^2 = \frac{\sum v^2}{n-2}$$

B) Chybami jsou zatíženy jen hodnoty x_i

Tento případ není obvykle uvažován, protože regresní analýza řeší vztah závislé veličiny (y) na nezávislé veličině (x). Kterou veličinu prohlásíme za závislou (zatíženou chybami) a kterou za nezávislou (bezchybnou) je jen otázka případné záměny proměnných. Samozřejmě obdržíme jiné parametry přímky.

Provedeme tedy záměnu proměnných, rovnice přímky bude:

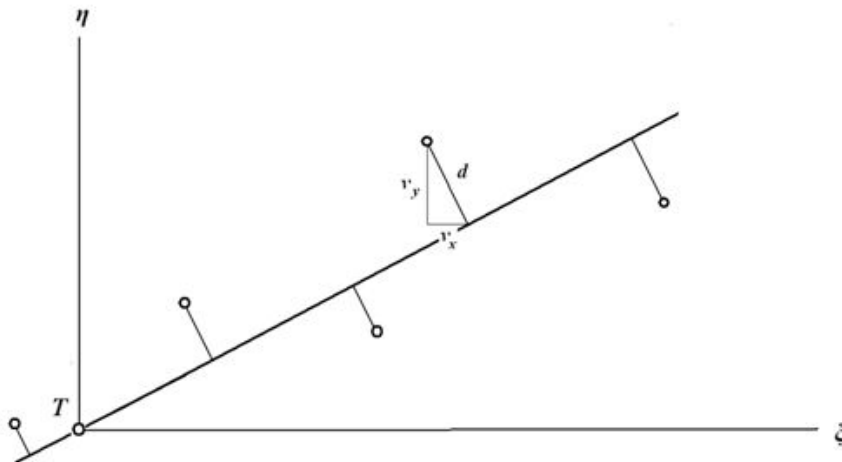
$$x_i = A^* + B^* y_i$$

Parametr A^* je nyní úsek na ose x (souřadnice x_i pro $y_i = 0$) a parametr B^* je směrnice přímky (tangenta úhlu tentokrát vzhledem k ose y).

Chceme-li vyjádřit tuto přímku ve tvaru $y_i = A' + B' x_i$, musíme provést transformaci koeficientů (vyjádřit z původní rovnice veličinu y jako funkci x)

C) Chybami jsou zatíženy hodnoty x_i , y_i

Tento případ je běžný v geodetických aplikacích, kde jsou obě souřadnice zatíženy měřickými chybami. Parametry A'' , B'' určíme podle uvedených vztahů z redukovaných souřadnic na těžiště T . Minimalizuje nyní součet čtverců kolmých vzdáleností vzdálenosti d , tj. $\Sigma d^2 = \Sigma (v_x^2 + v_y^2) = \min.$



$$\xi_i = x_i - x_T, \quad \eta_i = y_i - y_T,$$

$$x_T = \frac{\sum x}{n}, \quad y_T = \frac{\sum y}{n}.$$

$$\operatorname{tg} 2\varphi = \frac{2\sum \xi\eta}{(\sum \xi^2) - (\sum \eta^2)},$$

$$B'' = \operatorname{tg} \varphi$$

$$A'' = y_T - B'' x_T$$

