Дисперсионен, корелационен и регресионен анализ

С помощта на статистически хипотези може да се провери влиянието на един или друг фактор върху показателите на качеството. *Дисперсионният анализ* се използва за определянето на значимите влияещи фактори.

Всяка случайна величина **x** може да се разглежда като случайна функция на много и различни по характер фактори. Част от тези фактори оказват незначително влияние върху **x** и по същество обуславят нейния случаен характер. Друга част от факторите обаче оказват съществено влияние върху математическото очакване на случайната величина **x**. Изследването на влиянието на различни по брой и характер фактори по отношение изменението на математическото очакване представлява основна задача на дисперсионния анализ.

За провеждането на такъв анализ е необходимо да бъдат изпълнени следните предпоставки: нормален закон на разпределение; постоянство на дисперсията; независимост на извадките. Нарушаването на тези предпоставки води до известни отклонения на направените статистически изводи.

Обикновено проверката на изпълнението на тези предпоставки изисква голям брой наблюдения и изчисления и практически рядко се извършва такава проверка. Но редица проведени изследвания за устойчивостта на статистическите изводи от дисперсионния анализ показват, че той е малко чувствителен към отклоненията от нормално разпределение на наблюдаваната случайна величина, а при еднакъв обем на извадките и към непостоянство на дисперсията. Това обуславя широкото и ефективно приложение на дисперсионния анализ в практиката.

В зависимост от броя на изучаваните влияещи фактори дисперсионният анализ бива еднофакторен и многофакторен. При еднофакторния дисперсионен анализ се изследва влиянието върху математическото очакване на случайната величина **x** на един фактор A, който приема **k** различни стойности A_1, A_2, ...,A_k. Тези стойности могат да не се оценяват количествено, а да се разглеждат като **k** условни нива. За всяко от нивата се правят случайни извадки с обем **n** и се подлагат на наблюдение (определят се оценките на математическото очакване). Поставя се задачата да се провери дали влиянието на различните стойности (нива) на фактора A е довело до получаването на различни стойности на математическото очакване на случайната величина **x** или не. По същество тази задача представлява типичен случай на статистическа проверка на хипотезата (обикновенно приемана за нулева), че математическите очаквания при изследваните **k** случая са равни помежду си.

В резултат на осъществените за всяко ниво на фактора A **k** случайни извадки с обем **n** се получават **k** групи по **n** данни x_{ij} за случайната величина **x**, които могат да бъдат представени във вида, показан на табл.1. Средните стойности $\overline{x_i}$ се използват за оценка на математическото очакване.

Всеки резултат x_{ij} може да бъде представен във вид на линеен модел, т.е. като сума от три величини: $x_{ij} = M+\alpha_i+\varepsilon_{ij}$, където M е математическото очакване на x, α_i е константа, отразяваща увеличение или намаление на математическото очакване M при i-то ниво на фактора, а ε_{ij} е грешката на линейния модел. Числените характеристики M, α_i, и ε_{ij}, са неизвестни и могат да бъдат оценени с помощта на статистическите оценки $\overline{X}$, $\overline{x_i} - \overline{X}$, и $x_{ij} - \overline{x_i}$:

$$x_{ij} = \overline{X} + (\overline{x_i} - \overline{X}) + (x_{ij} - \overline{x_i}).$$

Таблица 1

Номер на наблюдението (j)	Стойности (нива) на фактора A (i)						
	A_1	A_2	...	A_i	...	A_k	
1	x_{11}	x_{21}	...	x_{i1}	...	x_{k1}	
2	x_{12}	x_{22}	...	x_{i1}	...	x_{k2}	
.	.	.	.	.	.	.	
.	.	.	.	.	.	.	
.	.	.	.	.	.	.	
n	x_{1n}	x_{2n}	...	x_{in}	...	x_{kn}	
Средна стойност	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_i$	...	$\bar{x}_k$	$\bar{X}$

Основното уравнение на еднофакторния дисперсионен анализ има вида: $Q=Q_A+Q_R$, където Q е обща сума на квадратите, Q_A - сума на квадратите между групите, Q_R - остатъчна сума на квадратите.

Обработването на данните се извършва в следния ред:

- Определя се общата сума на квадратите Q, която представлява сума от квадратите на отклоненията на всички наблюдения x_{ij} от тяхната средна стойност $\bar{X}$:

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{X})^2$$

където $\bar{X} = \frac{1}{k}\sum_{i=1}^{k}\bar{x}_i$

- Определя се сумата на квадратите между групите Q_A като сума от квадратите на отклоненията на средните стойности по групи $\bar{x}_i$ от общата средна стойност $\bar{X}$:

$$Q_A = n\sum_{i=1}^{k}(\bar{X} - \bar{x}_i)^2$$

Тя характеризира влиянието на различните стойности (нива) на фактора A върху изменението на математическото очакване на групите по отношение на общото математическо очакване на случайната величина **x**.

- Определя се остатъчната сума на квадратите Q_R, като сума от квадратите на отклоненията на всяко наблюдение x_{ij} от средната стойност $\bar{x}_i$ на i-тата група:

$$Q_R = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2$$

Тя характеризира остатъчното разсейване на случайната величина **x**, породено от действието на случайната грешка ε_{ij}.

• Определят се степените на свобода v, v_A и v_R, т.е. броят на независимите величини, участващи в образуването на сумите на квадратите Q, Q_A и Q_R:

$v=kn - 1;$
$v_A=k - 1;$
$v_R=k(n - 1).$

• Определят се статистическите оценки на дисперсията на случайната величина **x** чрез отношенията на сумите на квадратите Q, Q_A, Q_R и съответните степени на свобода v, v_A и v_R :

$$s^2 = \frac{Q}{v} = \frac{Q}{kn-1};$$

$$s_A^2 = \frac{Q_A}{v_A} = \frac{Q_A}{k-1};$$

$$s_R^2 = \frac{Q_R}{v_R} = \frac{Q_R}{k(n-1)},$$

които се наричат съответно обща оценка на дисперсията (s^2), оценка на дисперсията по факторите (s_A^2) и остатъчна оценка на дисперсията (s_R^2).

Резултатите от обработката на данните при еднофакторния дисперсионен анализ обикновенно се представят във вида, показан на табл.2.

Таблица 2

Сума на квадратите		Степени на свобода	Оценки на дисперсиите
Обща	Q	$v=kn - 1$	$s^2=Q/v$
По фактори	Q_A	$v_A=k - 1$	$s_A^2=Q_A/v_A$
Остатъчна	Q_R	$v_R=k(n - 1)$	$s_R^2=Q_R/v_R$

За проверка на нулевата хипотеза обикновенно се използва F-критерия, основан на разпределението на Фишер. За целта се определя дисперсионното отношение (F - отношение):

$$F = \frac{s_A^2}{s_R^2},$$

което се сравнява с критичната стойност $F_т$ на разпределението на Фишер в зависимост от степените на свобода v_A и v_R и избрано ниво на значимост α (вероятност за грешка от I род - неправилно отхвърляне на вярна нулева хипотеза). Ако $F<F_т$ няма основание да не се приеме нулевата хипотеза и следователно влиянието на фактора А не може да се счита за съществено. Ако

$F>F_т$ нулевата хипотеза се отхвърля и се приема алтернативната ѝ, т.е. има всички основания да се смята, че факторът А влияе съществено върху математическото очакване на x, което не може да се обясни само със случайния характер на извадките.

Целта на дисперсионния анализ е проверка на статистическата значимост на разликите между средните стойности, получени при различните нива на изследвания фактор. Тази проверка се извършва посредством разделяне на сумата на квадратите на компоненти, т.е. посредством разделяне на общата дисперсия на части, едната от които е обусловена от случайни грешки, а втората е свързана с разликите в средните стойности. Дисперсията на извадка се определя от зависимостта:

$$s^2 = \frac{1}{n-1}\sum_{1}^{n}(x_i - \overline{X})^2 .$$

При фиксиран обем на извадката n дисперсията е функция на сумата на квадратите на отклоненията.

Ако се разгледа следният пример:

	Група A_1	Група A_2
Наблюдение 1 Наблюдение 2 Наблюдение 3	2 3 1	6 7 5
Средна стойност $\overline{x_i}$	2	6
Сума на квадратите	2	2
Обща средна стойност $\overline{X}$	4	
Обща сума на квадратите Q	28	

Средните $\overline{x_i}$ за двете групи A_1 и A_2 се различават. Сумата на квадратите на отклоненията вътре във всяка група са равни на 2. Общата сума на квадратите на отклоненията от общата средна стойност, без да се отчита груповата принадлежност е 28. Сумата на квадратите (дисперсията), основана на вътрешногруповата изменчивост има по-малка стойност отколкото общата сума на квадратите, изчислена на базата на общата изменчивост (спрямо общото средно). Общата сума на квадратите Q=28 се разделя на две съставляващи: сума на квадратите, обусловена от вътрешногруповата изменчивост (Q_R=2+2=4) и сума на квадратите, обусловена от разликата в средните стойности на групите (Q_A=28-(2+2)=24). Сумата на квадратите, обусловена от вътрешногруповата изменчивост обикновенно се нарича остатъчна сума на квадратите и характеризира грешката на експеримента.

За статистическа оценка на дисперсиите s^2, s_A^2 и s_R^2 се използват отношенията на сумите на квадратите към степените на свобода (броят на независимите величини, участващи при определянето на сумите на квадратите). За разгледания пример общата сума Q се определя от 3.2=6 независими наблюдения, но във формулата за Q участва и една линейна зависимост $\overline{X}$ =1/6(x_{11}+x_{12}+x_{13}+x_{21}+x_{22}+x_{23}), поради което броят на независимите величини е v=kn–1=6-1=5.

С помощта на F-критерия се оценява статистическата значимост на разликата в средните стойности на групите.

Случаите в които върху едно явление оказва влияние само един фактор са твърде редки. Обикновено се налага изследването на влиянието на множество фактори. За целта може да се приложи еднофакторният дисперсионен анализ за всеки фактор поотделно, но такъв подход води до ненужно увеличаване на обема на наблюденията и не дава възможност за изучаването на съвместното влияние на група фактори.

Многофакторният дисперсионен анализ е особено ефективен метод за едновременното изучаване на влиянието на няколко фактора A, B, C, ... , всеки поотделно и техни съчетания, върху математическото очакване на случайната величина **x**. При него в повечето случаи не са необходими паралелни наблюдения за всяка комбинация от нива на факторите, което води до значително намаляване на общия обем наблюдения. Съществува възможност така да се планират и проведат експериментите, че при минимален брой наблюдения да се извлече максимална информация за влиянието на всеки фактор и техни взаимодействия.

Задачата на многофакторния дисперсионен анализ се състои в потвърждаването или отхвърлянето на редица от хипотези за влиянието на множество фактори и техни взаимодействия. За нулеви хипотези се приемат твърденията, че факторите (поотделно и в съчетания) не оказват влияние върху математическото очакване. Данните от проведените наблюдения могат да бъдат представени в таблица (или група таблици) от вида:

Таблица 3

		i						
		A_1		A_2			A_k	
I	B_1	$X_{111},x_{112},\ldots,$ x_{11n}		$X_{211},x_{212},\ldots,$ x_{21n}			$X_{k11},x_{k12},\ldots.$ $, x_{k1n}$	$\overline{x}_{01}$
	B_2	$X_{121},x_{122},\ldots,$ x_{12n}		$X_{221},x_{222},\ldots,$ x_{22n}			$X_{k21},x_{k22},\ldots.$ $, x_{k2n}$	$\overline{x}_{02}$
	B_q	$X_{1q1},x_{1q2},\ldots,$ x_{1qn}		$X_{2q1},x_{2q2},\ldots,$ x_{2qn}			$X_{kq1},x_{kq2},\ldots.$ $, x_{kqn}$	$\overline{x}_{0q}$
		$\overline{x}_{10}$		$\overline{x}_{20}$			$\overline{x}_{k0}$	$\overline{X}$

Общият брой наблюдения N при два фактора A и B (к нива на фактора A и q нива на фактора B) ще бъде N=nkq.

При двуфакторният дисперссионене анализ всеки резултат от наблюденията може да бъде представен чрез следния модел:

$$x_{ilj}=M+\alpha_i+\beta_l+\alpha\beta_{il}+\varepsilon_{ilj},$$

в който M е общото математическо очакване на x, α_i отразява изменението на M под влиянието на i-то ниво на фактора A, β_l отразява изменението на M под влиянието на l-то ниво на фактора B, ε_{ilj} отразява случайното изменение на x, а $\alpha\beta_{il}$ отразява съвместното влияние на факторите A и B (по-точно на i-то ниво на A и l-то ниво на B). За да се проведе в пълен вид (с отчитане на взаимодействието) двуфакторният дисперсионен анализ, са необходими

паралелни наблюдения, т.е. n≥2. При повече от два фактора (многофакторен дисперсионен анализ) това изискване не е необходимо и много често наблюденията не се дублират.

Основното уравнение на дисперсионния анализ при два фактора с отчитане на взаимодействието има вида: $Q=Q_A+Q_B+Q_{AB}+Q_R$, където Q, Q_A, Q_B, Q_{AB}, Q_R са суми на квадратите – съответно обща сума на квадратите, сума на квадратите между групите на фактора A, сума на квадратите между групите на фактора B, сума на квадратите между групите на взаимодействието на A и B и остатъчна сума на квадратите:

$$Q=\sum_1^k\sum_1^q\sum_1^n(x_{ilj}-\overline{X})^2$$

$$Q_A=nq\sum_1^k(\overline{x}_{io}-\overline{X})^2$$

$$Q_A=nk\sum_1^q(\overline{x}_{ol}-\overline{X})^2$$

$$Q_{AB}=n\sum_1^k\sum_1^q(\overline{x}_{il}-\overline{x}_i-\overline{x}_{ol}+\overline{X})^2$$

$$Q_R=\sum_1^k\sum_1^q\sum_1^n(x_{ilj}-\overline{x}_{il})^2$$

Броят на степените на свобода на тези суми на квадрати са съответно:

$v=N-1$;

$v_A=k-1$;

$v_B=q-1$;

$v_{AB}=(k-1)(q-1)$;

$v_R=kq(n-1)$.

За оценка на дисперсията се използват отношенията на сумите на квадратите и съответните степени на свобода:

$s^2=Q/v$;

$s^2_A=Q_A/v_A$;

$s^2_B=Q_B/v_B$;

$s^2_{AB}=Q_{AB}/v_{AB}$;

$s^2_R=Q_R/v_R$.

С помощта на F-критерия се проверяват три хипотези (фактора A, фактора B и взаимодействието AB не оказват влияние върху математическото очакване) чрез сравняване на отношенията $F_A=s^2_A/s^2_R$, $F_B=s^2_B/s^2_R$, $F_{AB}=s^2_{AB}/s^2_R$ с табличните стойности F_A^T, F_B^T, F_{AB}^T, определени от таблиците за F-

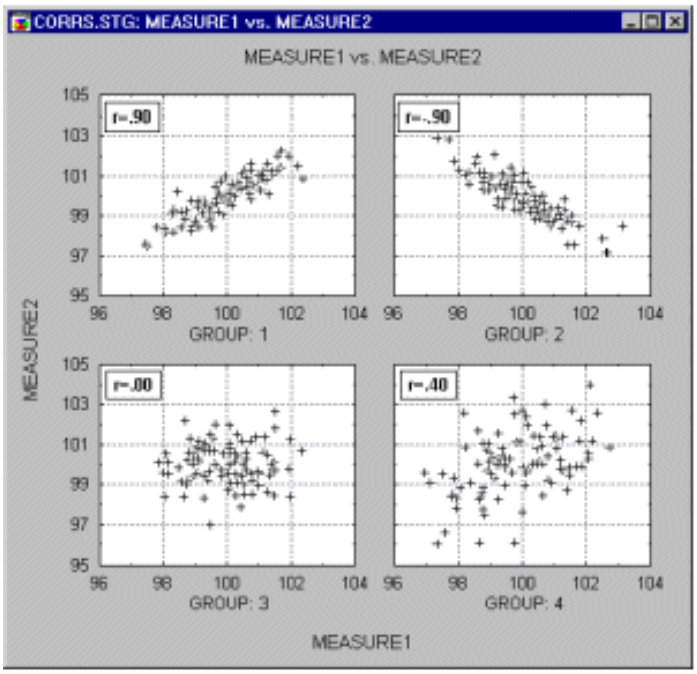
разпределение. Решенията за приемането или отхвърлянето на всяка от трите хипотези се вземат поотделно и независимо.

Чрез дисперсионния анализ се определя влияе ли даден фактор или не върху определен показател на качеството, но не и как влияе, т.е. силата и формата на връзката. Това може да стане с помощта на *корелационния анализ*.

Корелационен анализ

Връзката, изразяваща изменението на средната стойност на една величина при изменение на друга величина се нарича корелационна връзка. Графичното представяне на съвкупността от всички точки (x_i,y_i), получени при наблюдението се нарича диаграма на разсейване (корелационно поле) и дава визуална представа за връзката между величините **x** и **y** (фиг.6).

Съществуването на корелационна връзка между величините **x** и **y**, формата (линейна или нелинейна) и нейната сила се определят с помощта на коефициента на корелация $\mathbf{r_{xy}}$ и корелационното отношение η_y.



Фиг.6 – *Диаграми на разсейване*

Коефициентът на корелация се определя по зависимостта:

$$r_{xy} = \frac{C_{xy}}{s_x s_y},$$

където s_x и s_y са средноквадратичните отклонения съответно за променливите x и y, а C_{xy} е коварация на случайните величини x и y:

$$C_{xy} = \frac{1}{n-1}\sum_{1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Коефициентът на корелация има стойност в границите -1< r_{xy}<+1. Колкото е по-близък да +1, толкова е по-силна положителната линейна корелационна връзка (с нарастването на x нараства и y). Колкото r_{xy} е по-близък до -1, толкова е по-силна отрицателната линейна корелационна връзка (с нарастването на x намалява y). При $r_{xy}=\pm 1$ между величините x и y съществува линейна функционална зависимост. При стойности на коефициента на корелация близки до нула може да се предполага или наличие на нелинейна корелационна връзка, или въобще отсъствие на връзка. В такъв случай r_{xy} се сравнява с корелационното отношение, което се определя от зависимостта [4]:

$$\eta_y = \frac{S(\bar{y}_x)}{S_y},$$

където:

$$S(\bar{y}_x) = \sqrt{\frac{1}{N}\sum_{i=1}^{p}\nu_j(\bar{y}_{xi} - \bar{y}^2)} \quad ; \qquad \bar{y}_{xi} = \frac{1}{\nu_i}\sum_{j=1}^{q} y_j \nu_{ij}$$

Корелационното отношение е в границите $0 \le \eta_y \le 1$, при това $\eta_y \ge |r_{xy}|$. При η_y =0 между величините x и y отсъства връзка, а при η_y =1 между x и y съществува еднозначна функционална връзка, която при r_{xy} =0 е нелинейна.

За изследването на връзката между повече от две величини (например връзката на y с два фактора x_1 и x_2) се използва коефициентът на множествена корелация R_{yx1x2}, който се определя чрез коефициентите на корелация r_{yx1}, r_{yx2} и r_{x1x2}:

$$R_{yx_1x_2} = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1}^2 r_{x_1x_2}}{1 - r_{x_1x_2}}}$$

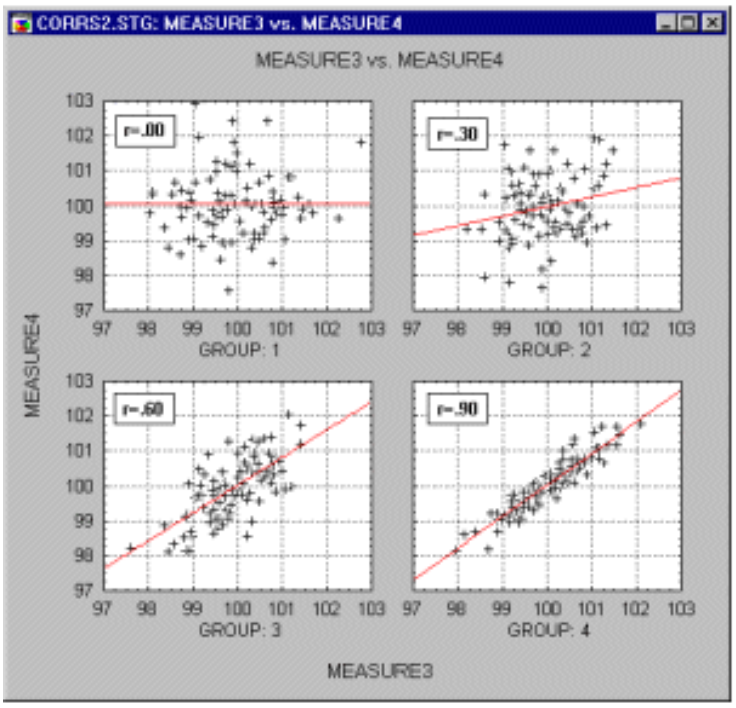
Регресионен анализ

При наличие на функционална зависимост между две величини, математическият модел (уравнението на регресия) на тази зависимост може да бъде определен чрез използване на *регресионен анализ* - метод посредством който се определят коефициентите на уравнението на регресия.

По определение понятието регресия е: регресия на y по отношение на x се нарича условното математическо очакване на случайната величина y при условие, че x приема зададени стойности.

Един от често използваните регресионни модели е полином от m-та степен, който има вида:

$$y = f_m(x) = a_0 + a_1 x + a_2 x^2 + ... + a_m x^m$$



Фиг.7 – *Линии на регресия*

За определянето на коефициентите на уравнението на регресия a_0, a_1, a_2,...,a_m се използва методът на най-малките квадрати. Основно изискване на метода е сумата от квадратите на отклоненията между опитно получените стойности и тези, определени по уравнението на регресия да бъде минимална. Коефициентите a_0, a_1, a_2,...,a_m се определят чрез решаване на системата от (m+1) уравнения:

$$\begin{aligned}
&\alpha_{2m}(x).a_m + \alpha_{2m-1}(x).a_{m-1} + ... + \alpha_{m+1}(x).a_1 + \alpha_m(x).a_0 = \alpha_{m,1}(x,y)\\
&\alpha_{2m-1}(x).a_m + \alpha_{2m-2}(x).a_{m-1} + ... + \alpha_m(x).a_1 + \alpha_{m-1}(x).a_0 = \alpha_{m-1,1}(x,y)\\
&.....\\
&.....\\
&\alpha_m(x).a_m + \alpha_{m-1}(x).a_{m-1} + ... + \alpha_1(x).a_1 + \alpha_0(x).a_0 = \alpha_{0,1}(x,y)
\end{aligned}$$

където

$$\alpha_{2m}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i^{2m};$$

$$\alpha_1(x) = \frac{1}{n}\sum_{i=1}^{n} x_i;$$

$$\alpha_0(x) = \frac{1}{n}\sum_{i=1}^{n} x_i^0 = 1$$

$$\alpha_{m,1}(x,y)=\frac{1}{n}\sum_{i=1}^{n}x_i^m y_i;$$

$$\alpha_{1,1}(x,y)=\frac{1}{n}\sum_{i=1}^{n}x_i y_i;$$

$$\alpha_{0,1}(x,y)=\frac{1}{n}\sum_{i=1}^{n}x_i^0 y_i=\frac{1}{n}\sum_{i=1}^{n}y_i.$$

Например за уравнение от първа степен ($y = a_0+a_1x$) системата уравнения ще бъде:

$$\begin{vmatrix}\alpha_2(x).a_1+\alpha_1(x).a_0=\alpha_{1,1}(x,y)\\ \alpha_1(x).a_1+\alpha_0(x).a_0=\alpha_{0,1}(x,y)\end{vmatrix}$$

където:

$$\alpha_2(x)=\frac{1}{n}\sum_{1}^{n}x_i^2$$

$$\alpha_1(x)=\frac{1}{n}\sum_{1}^{n}x_i$$

$$\alpha_0(x)=1$$

$$\alpha_{0,1}(x,y)=\frac{1}{n}\sum_{1}^{n}y_i$$

$$\alpha_{1,1}(x,y)=\frac{1}{n}\sum_{1}^{n}x_i y_i$$

След решаване на системата, за коефициентите на уравнението на регресия се получава:

$$a_1=\frac{n\sum x_i y_i-\sum x_i\sum y_i}{n\sum x_i^2-(\sum x_i)^2}\quad,\qquad a_0=\frac{\sum y_i-a_1\sum x_i}{n}\quad.$$

Оценката на средноквадратичното отклонение на опитните данни от данните, получени по уравнението на регресия се определя от израза:

$$s_m=\sqrt{\frac{\sum_{1}^{n}[y_i-f_m(x_i)]^2}{n-(m+1)}}.$$

Регресионният анализ се прилага за математическо моделиране на процеси при тяхното проектиране, за описание на предавателната характеристика на измервателни средства, за определяне на апроксимиращата права при съставяне на някой видове контролни карти и др.