ОСНОВЫ ХИМИЧЕСКОЙ МЕТРОЛОГИИ

Лекция 9 Регрессионный и корреляционный анализ

лектор: Образовский Е. Г.

8 апреля 2015 г.

Регрессионный анализ дает возможность построить уравнение, описывающее связь между экспериментальными данными (например, градуировочный график), вид которого задает аналитик, а *корреляционный анализ* позволяет судить о том, насколько хорошо экспериментальные точки согласуются с выбранным уравнением (“ложатся” на кривую).

Поскольку большинство методов анализа являются косвенными, в аналитической химии наиболее часто регрессионный анализ применяется при построении градуировки, т. е. при установлении функциональной связи между аналитическим сигналом и концентрацией определяемого компонента. Для многих методов анализа известны аналитические выражения для этой связи, зависящие от некоторых параметров. Тогда по экспериментальным данным следует наиболее точно оценить эти параметры.

Рассмотрим сначала часто встречающийся случай линейной зависимости аналитического сигнала I от концентрации C:

$$\mathbf{I} = \mathbf{AC} + \mathbf{B} + \varepsilon,$$

где ε – случайная погрешность.

По набору экспериментальных данных (C_n, I_n) для $n = 1, 2, ..., N$ нам нужно получить наилучшую оценку для параметров модели A и B. Зная параметры A и B, мы могли бы для любого значения концентрации C_n вычислить точное (истинное) значение аналитического сигнала $I_t = AC_n + B$. Результат измерения I_n отклоняется от этой величины из-за погрешностей анализа. Обозначим $P_{A,B}(I_n)$ вероятность получить в n-м измерении значение I_n. Тогда, предполагая независимость отдельных измерений, вероятность получения всего набора результатов измерений (I_n, C_n) $n = 1, 2, ..., N$ равна произведению вероятностей

$$\mathbf{P_{A,B}(I_1, ..., I_N) = P_{A,B}(I_1) \cdot ... \cdot P_{A,B}(I_N).}$$

Принято считать, что наилучшая оценка параметров A и B соответствует наибольшей вероятности $P_{A,B}(I_1, ..., I_N)$. Для нормального закона распределения

$$\mathbf{P_{A,B}(I_n) \sim \exp\left[-\frac{(I_n - AC_n - B)^2}{2\sigma^2_{In}}\right]}$$

получаем

$$\mathbf{P_{A,B}(I_1, ..., I_N) \sim e^{-\chi^2/2}},$$

где

$$\mathbf{\chi^2 = \sum_{n=1}^{N} \frac{(I_n - AC_n - B)^2}{\sigma^2_{In}}}.$$

Наибольшей вероятности $P_{A,B}(I_1, ..., I_N)$ соответствует минимум χ^2; отсюда проистекает название – *метод наименьших квадратов*.

Из минимума χ^2 для постоянного значения $\sigma_{In} = \sigma_I$ следует

$$\mathbf{A}\sum_{\mathbf{n}}\mathbf{C_n^2} + \mathbf{B}\sum_{\mathbf{n}}\mathbf{C_n} = \sum_{\mathbf{n}}\mathbf{C_n I_n},$$

$$\mathbf{A}\sum_{\mathbf{n}}\mathbf{C_n} + \mathbf{BN} = \sum_{\mathbf{n}}\mathbf{I_n}.$$

Решением этой системы уравнений являются значения параметров нашей модели

$$\mathbf{A} = \frac{\mathbf{N}\sum_{\mathbf{n}}\mathbf{C_n I_n} - (\sum_{\mathbf{n}}\mathbf{C_n})(\sum_{\mathbf{n}}\mathbf{I_n})}{\mathbf{\Delta}},$$

$$\mathbf{B} = \frac{(\sum_{\mathbf{n}}\mathbf{I_n})(\sum_{\mathbf{n}}\mathbf{C_n^2}) - (\sum_{\mathbf{n}}\mathbf{C_n})(\sum_{\mathbf{n}}\mathbf{C_n I_n})}{\mathbf{\Delta}},$$

где

$$\mathbf{\Delta} = \mathbf{N}\sum_{\mathbf{n}}\mathbf{C_n^2} - (\sum_{\mathbf{n}}\mathbf{C_n})^2 = \mathbf{N}\sum_{\mathbf{n}}(\mathbf{C_n} - \bar{\mathbf{C}})^2, \quad \bar{\mathbf{C}} = \frac{\mathbf{1}}{\mathbf{N}}\sum_{\mathbf{n}}\mathbf{C_n}.$$

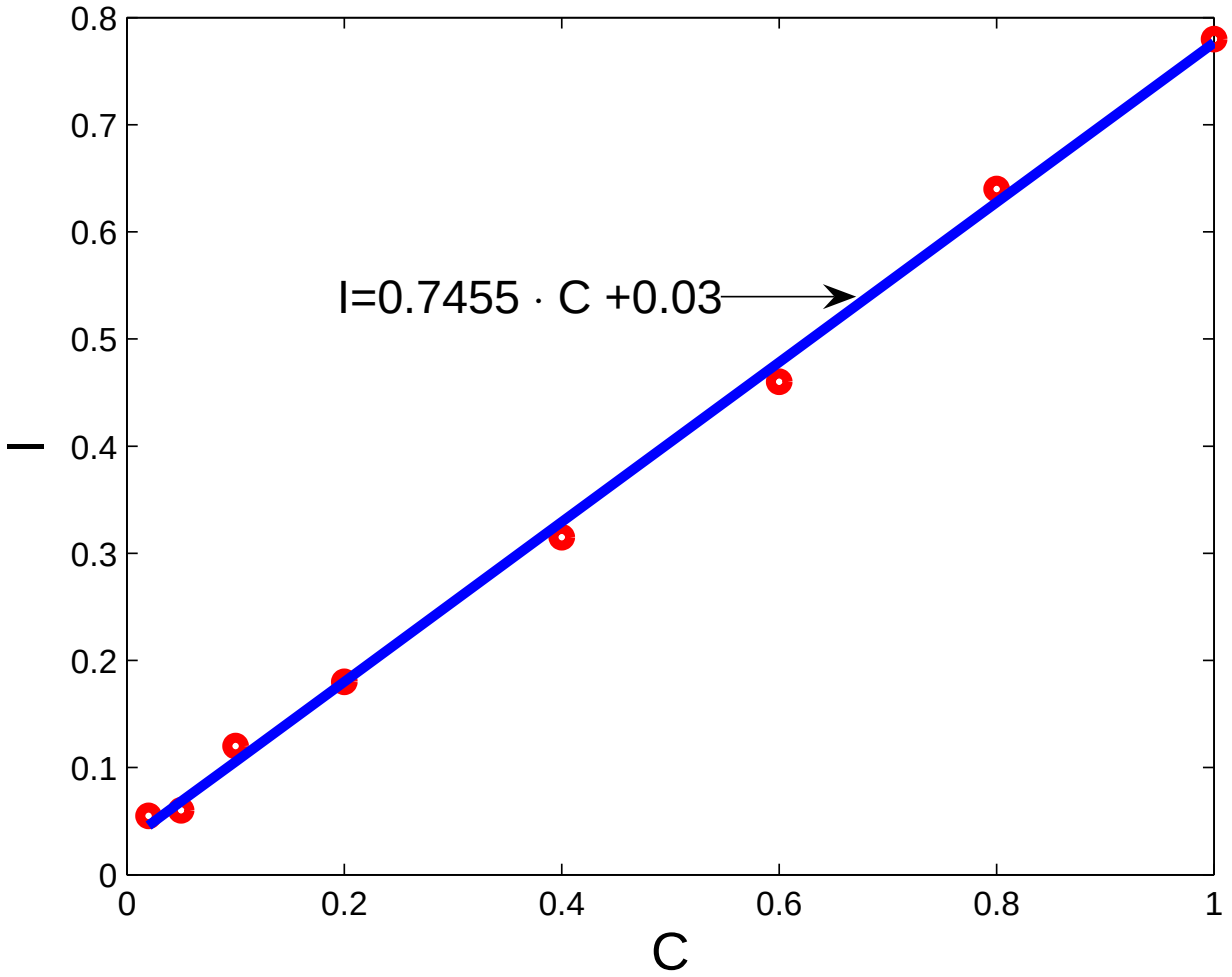


Рис.: Линейная зависимость аналитического сигнала I от концентрации C: $\mathbf{I = AC + B} + \varepsilon$

Параметры A и B, оцениваемые из эксперимента, также являются случайными величинами и нам следует знать погрешность этой оценки.
Из закона сложения ошибок и независимости измерений получаются следующие выражения:

$$\mathbf{S_A^2} = \overline{(\delta \mathbf{A})^2} = \frac{\mathbf{N}\sigma_\mathbf{I}^2}{\mathbf{\Delta}} = \frac{\sigma_\mathbf{I}^2}{\sum_\mathbf{n}(\mathbf{C_n} - \bar{\mathbf{C}})^2}.$$

$$\mathbf{S_B^2} = \overline{(\delta \mathbf{B})^2} = \frac{\sigma_\mathbf{I}^2 \sum_\mathbf{n} \mathbf{C_n^2}}{\mathbf{N}\sum_\mathbf{n}(\mathbf{C_n} - \bar{\mathbf{C}})^2}.$$

Неопределенность в коэффициентах A и B приводит к стандартному отклонению для определения концентрации по градуировочному графику

$$\mathbf{S_C} = \frac{\delta_\mathbf{I}}{\mathbf{A}},$$

где

$$\delta_\mathbf{I} = \sqrt{\overline{[\mathbf{C}\cdot\delta\mathbf{A} + \delta\mathbf{B}]^2}} = \sqrt{\sigma_\mathbf{I}^2\left(\frac{1}{\mathbf{N}} + \frac{1}{\mathbf{m}} + \frac{(\mathbf{C}-\bar{\mathbf{C}})^2}{\sum_\mathbf{n}(\mathbf{C_n}-\bar{\mathbf{C}})^2}\right)},$$

где m – число параллельных определений для анализируемого образца.

Минимальное значение стандартного отклонения погрешности определения концентрации по градуировочному графику достигается вблизи значения $\bar{C}$ и равно

$$S_C = \frac{\sigma_I}{A}\sqrt{\frac{1}{N} + \frac{1}{m}}.$$

Стандартное отклонение σ_I можно оценить также по данным, используемых для построения градуировочного графика,

$$\sigma_I \approx S_I = \sqrt{\frac{1}{N-2}\left[\sum_n (I_n - \bar{I})^2 - A^2 \sum_n (C_n - \bar{C})^2\right]},$$

или

$$\sigma_I \approx S_I = \sqrt{\frac{1}{N-2}\left[\sum_n (I_n - I_{grad})^2\right]},$$

где I_n – экспериментальные данные, I_{grad} – рассчитанные по градуировочному графику.

Ненулевое значение коэффициента B в уравнении градуировочной кривой обусловлено холостым опытом, например, за счет соосаждения элементов матрицы при гравиметрическом определении или за счет наложения на аналитическую линию линий элементов матрицы в спектральном анализе.

Рассмотрим в качестве примера метод стандартных добавок. Если мы можем учесть холостой опыт, то содержание интересующего нас компонента C_x можно определить, используя метод добавок

$$\mathbf{C_x = \bar{C}\frac{I_x}{\bar{I} - I_x}},$$

где I_x – величина аналитического сигнала анализируемой пробы; $\bar{I} = \sum I_n/N$ – среднее значение аналитического сигнала проб с известными добавками C_n; $\bar{C}$ – среднее значение концентрации добавок.

Погрешность определения неизвестного содержания C_x можно оценить, используя приближение для

$$\delta_{\mathbf{I}} \approx \frac{\sigma_{\mathbf{I}}(\mathbf{C} - \bar{\mathbf{C}})}{\sqrt{\sum_{\mathbf{n}}(\mathbf{C_n} - \bar{\mathbf{C}})^2}},$$

получая

$$\delta_{\mathbf{C}} \approx \frac{\mathbf{t}\sigma_{\mathbf{I}}\bar{\mathbf{C}}}{\mathbf{A}\sqrt{\sum_{\mathbf{n}}(\mathbf{C_n} - \bar{\mathbf{C}})^2}}.$$

Это значение превосходит погрешность определения концентрации в середине интервала.

Анализ остатков

Отклонение экспериментального значения величины аналитического сигнала от вычисленного по регрессии называется остатком $\varepsilon_n = I_n - AC_n - B$ и ее анализ позволяет сделать некоторые заключения о пригодности используемой модели.

Если величина остатков колеблется с примерно одинаковым размахом во всей области концентраций, то можно сделать вывод о пригодности используемой модели.

Анализ остатков

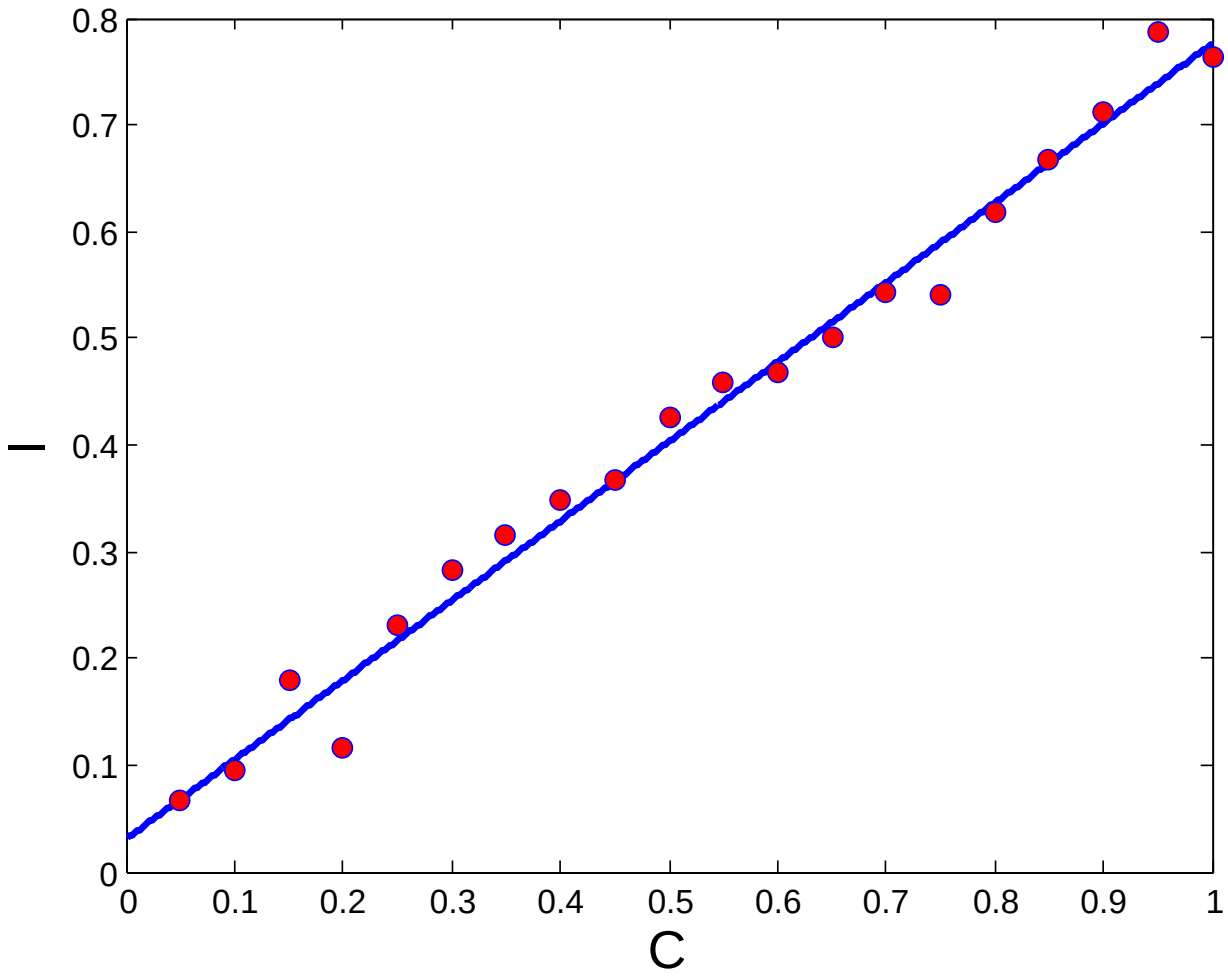


Рис.: Величина остатков случайно колеблется относительно линейной зависимости аналитического сигнала I от концентрации C: $\mathbf{I} = \mathbf{AC} + \mathbf{B} + \varepsilon$

Анализ остатков

Если же колебания величины остатков имеют регулярное отклонение от нулевого значения при изменении концентрации, то можно сделать вывод о непригодности линейной модели и необходимости включения в нее других факторов, например концентрации другого мешающего элемента, т. е. использовать множественную регрессию.

Анализ остатков

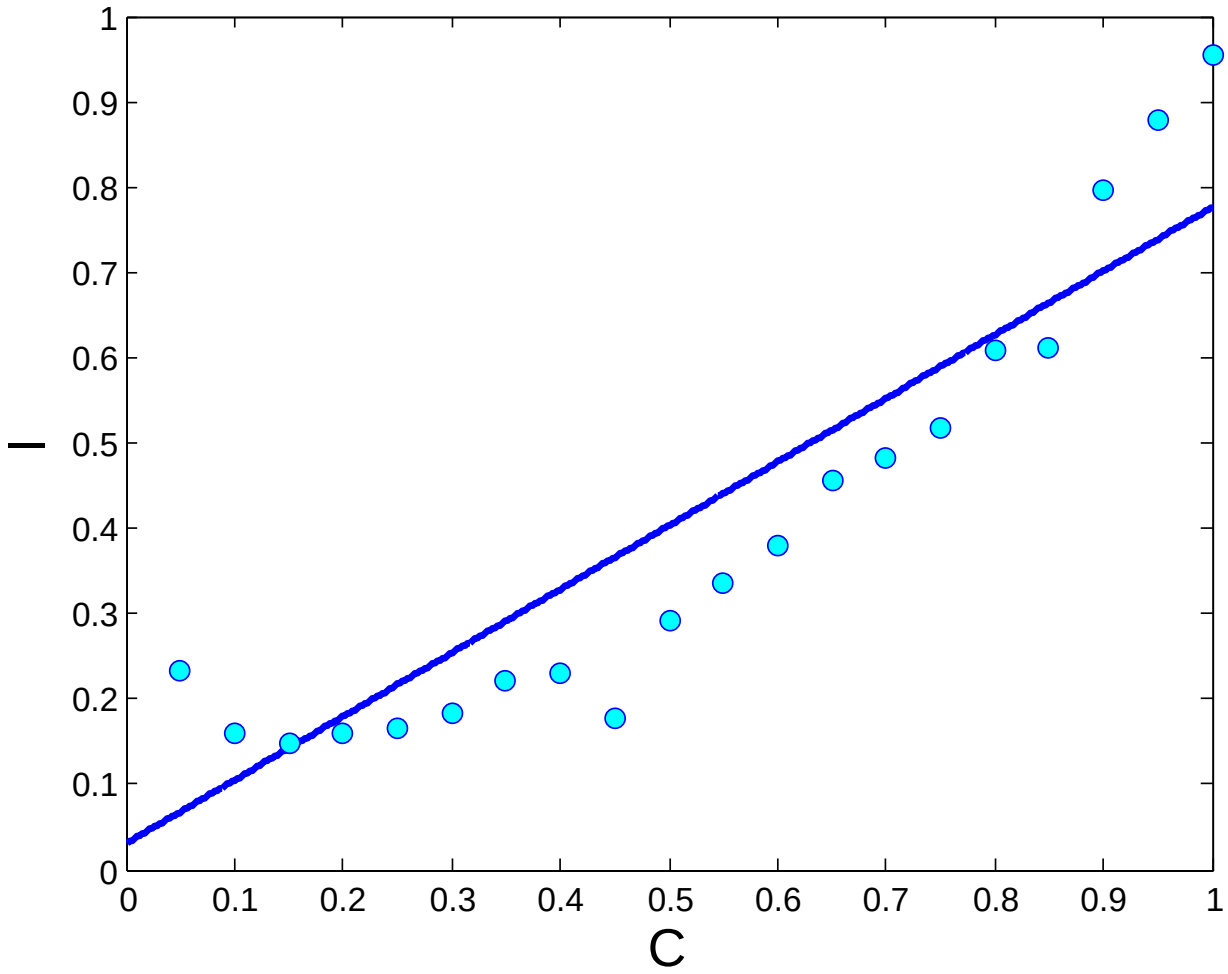


Рис.: Регулярное отклонение величины остатков от линейной зависимости аналитического сигнала I от концентрации C:

Анализ остатков

Если величина колебаний остатков относительно нулевого значения изменяется с изменением концентрации, то это свидетельствует о зависимости погрешности измерения аналитического сигнала от концентрации и следует использовать взвешенный метод наименьших квадратов.

Анализ остатков

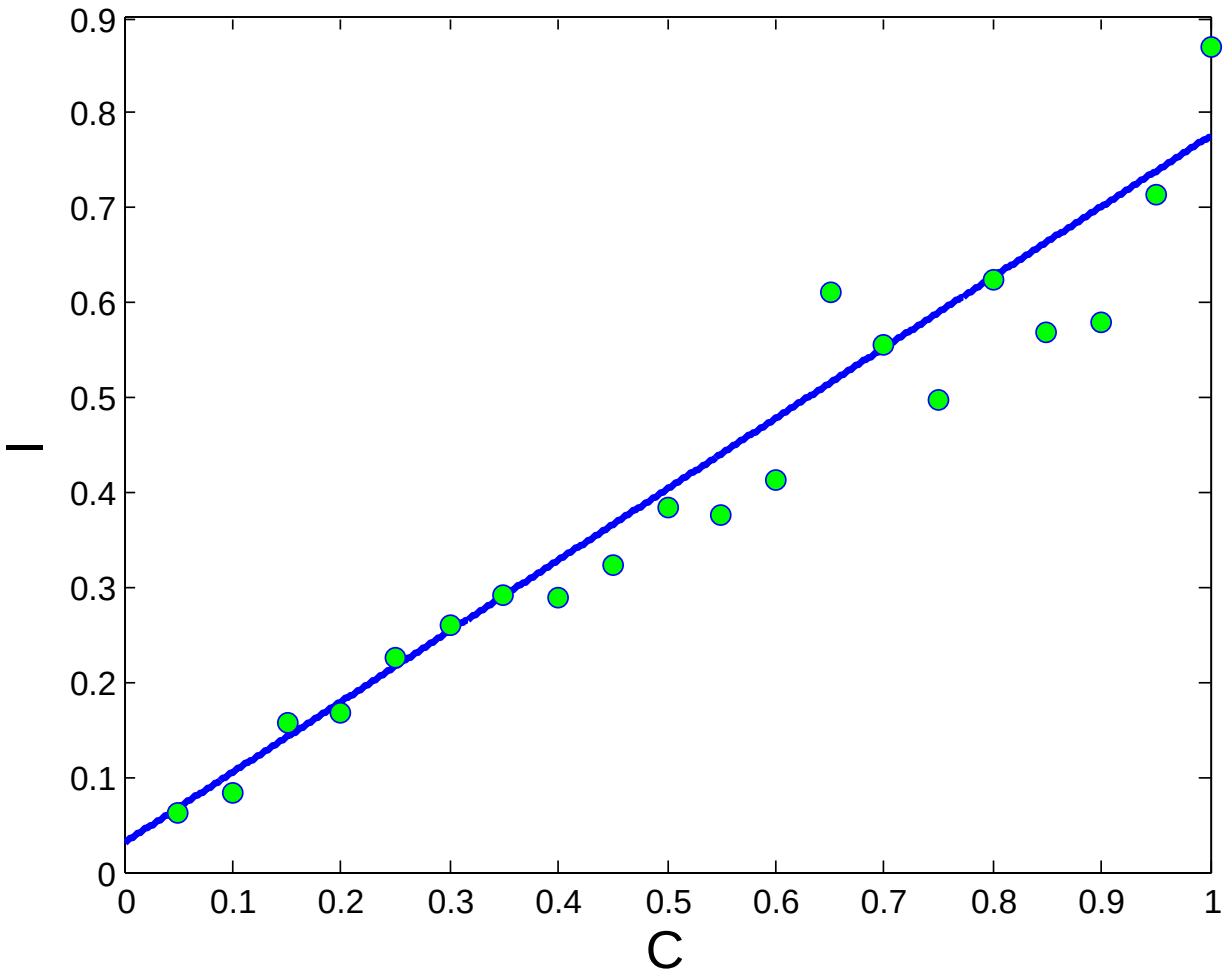


Рис.: Величина колебаний остатков изменяется с изменением концентрации

Взвешенный метод наименьших квадратов

Реально погрешность измерения величины аналитического сигнала σ_I зависит от его величины и в этом случае результаты с меньшей погрешностью более информативны. Коэффициенты A и B теперь оцениваются из условия минимума

$$\chi^2 = \sum_n \frac{(AC_n + B - I_n)^2}{\sigma^2_{In}} = \sum_n w_n (AC_n + B - I_n)^2,$$

где весовую функцию w_n удобно определить как

$$w_n = \frac{N}{\sigma^2_{In} \sum_n (1/\sigma_{In})^2},$$

чтобы выполнялось условие $\sum_n w_n = 1$.

Взвешенный метод наименьших квадратов

Тогда значения коэффициентов

$$A = \frac{N\sum_n w_n C_n I_n - (\sum_n w_n C_n)(\sum_n w_n I_n)}{N\sum_n w_n (C_n - \bar{C})^2},$$

$$B = \frac{(\sum_n w_n I_n)(\sum_n w_n C_n^2) - (\sum_n w_n C_n)(\sum_n w_n C_n I_n)}{N\sum_n w_n (C_n - \bar{C})^2}$$

отличаются от случая постоянного значения σ_I лишь появлением в каждой из сумм весового множителя w_n.

Множественная регрессия

Во многих случаях величина аналитического сигнала определяемого элемента зависит не только от его концентрации, но и концентраций других элементов, присутствующих в анализируемом образце.
В качестве примера можно привести рентгено-флуоресцентный метод анализа, в котором межэлементное влияние особенно заметно.

Множественная регрессия

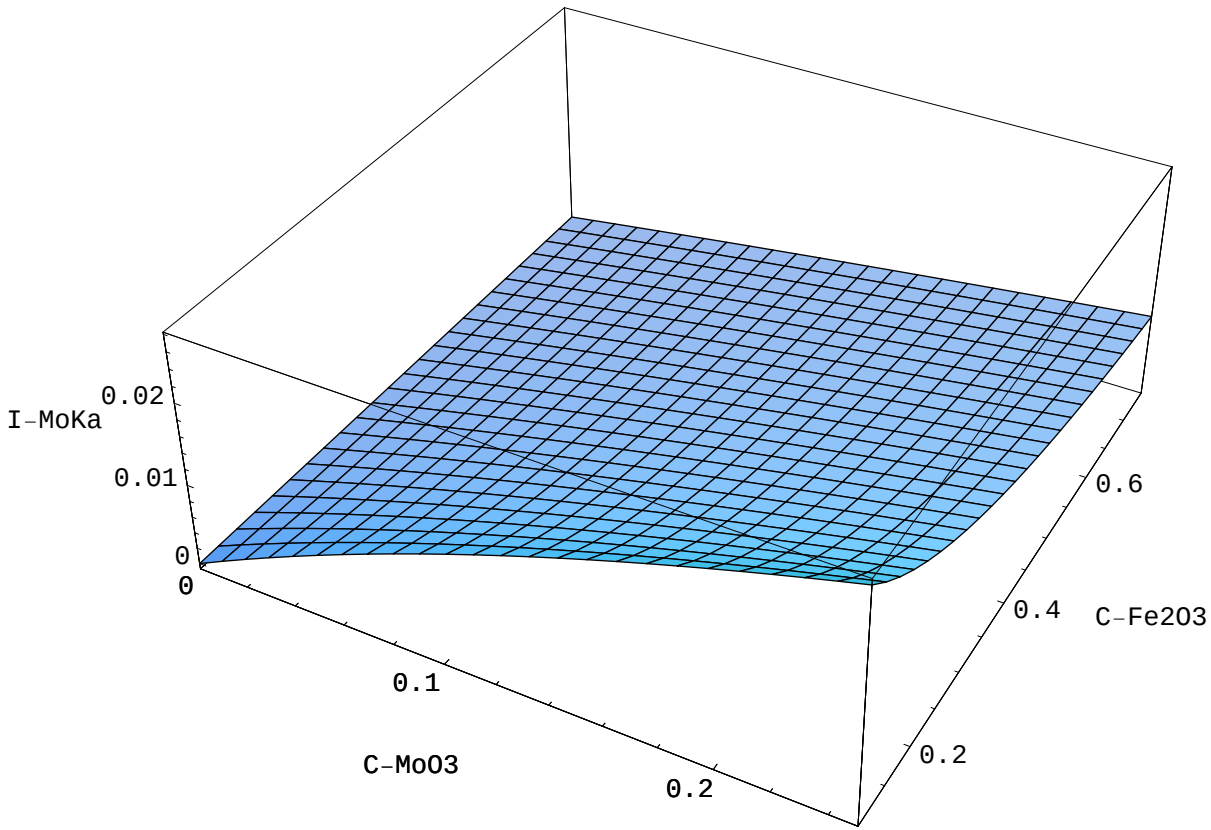


Рис.: Зависимость аналитического сигнала I_{MoK_α} от концентрации C_{Mo} и C_{Fe}.

Множественная регрессия

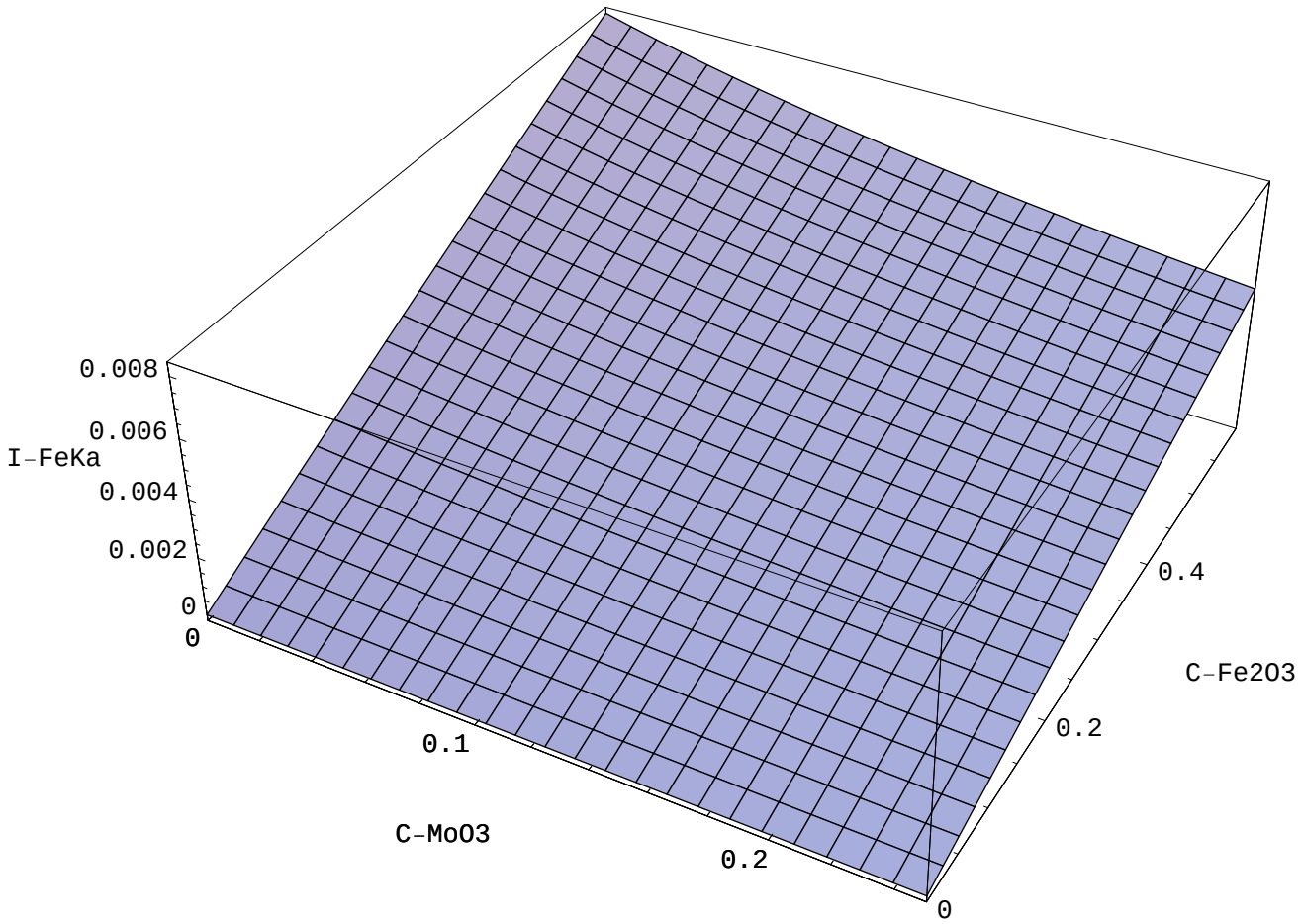


Рис.: Зависимость аналитического сигнала I_{FeK_α} от концентрации C_{Mo} и C_{Fe}.

Множественная регрессия

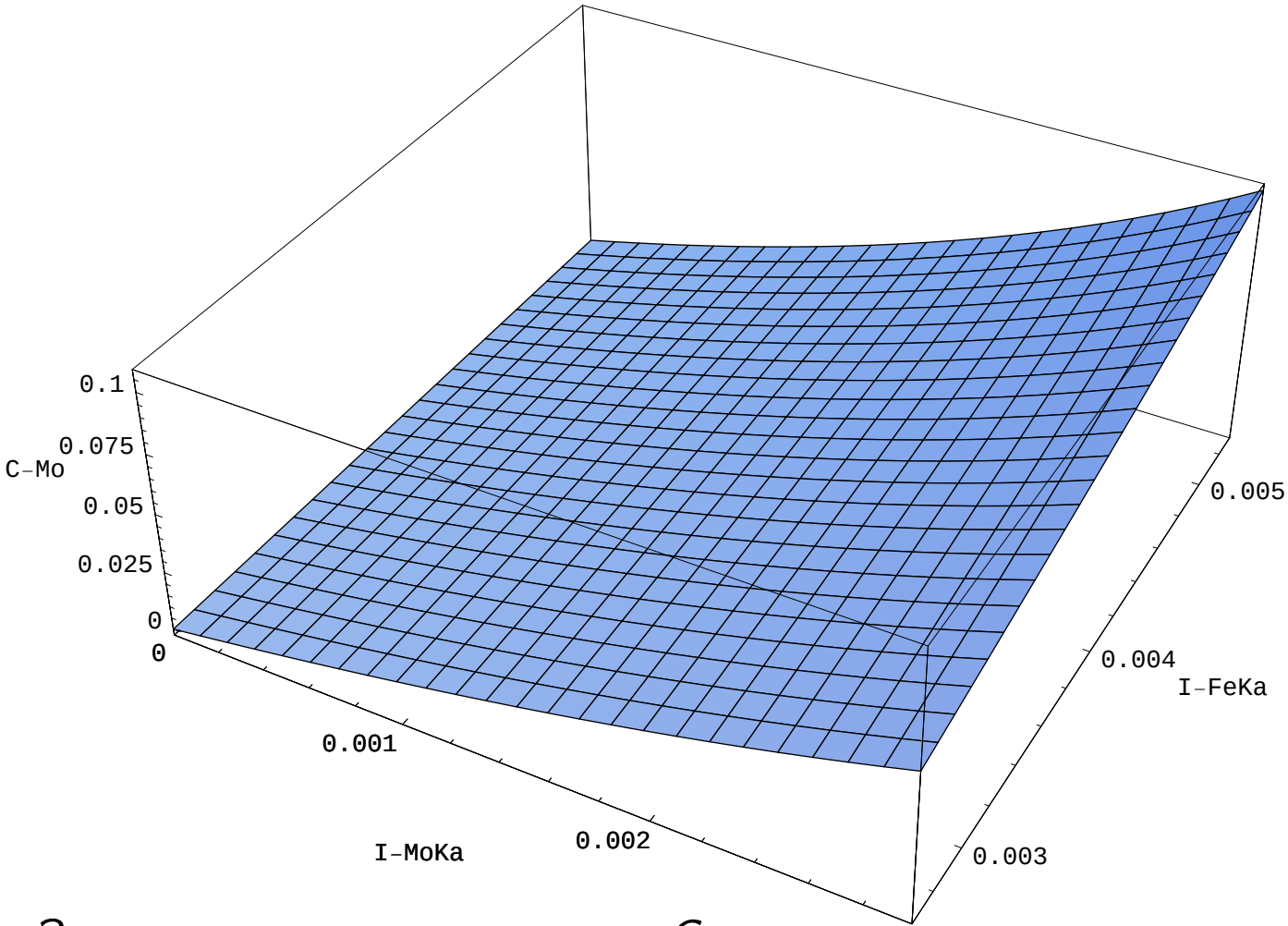


Рис.: Зависимость концентрации C_{Mo} от аналитического сигнала I_{MoK_α} и I_{FeK_α}.

Множественная регрессия

В относительно небольшом диапазоне изменения концентраций можно ограничиться линейной моделью влияния на величину аналитического сигнала I_i концентрации как определяемого компонента C_i, так и концентрации других элементов:

$$\begin{aligned} I_1 &= K_{11}C_1 + K_{12}C_2 + \cdots + K_{1m}C_m \\ I_2 &= K_{21}C_1 + K_{22}C_2 + \cdots + K_{2m}C_m \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ I_n &= K_{n1}C_1 + K_{n2}C_2 + \cdots + K_{nm}C_m. \end{aligned}$$

Это выражение удобно записать в матричной форме

$$\hat{\mathbf{I}} = \hat{\mathbf{K}} \cdot \hat{\mathbf{C}}.$$

Множественная регрессия

Для проведения градуировки используется p $(\geq m)$ образцов с известными концентрациями C_m^o и измеряются аналитические сигналы I_n^o. Тогда можно получить матрицу градуировочных коэффициентов $\hat{K}$:

$$\hat{\mathbf{K}} = \hat{\mathbf{I}}^{\mathbf{o}} \cdot (\hat{\mathbf{C}}^{\mathbf{o}})^{\mathbf{T}} \cdot \left(\hat{\mathbf{C}}^{\mathbf{o}} \cdot (\hat{\mathbf{C}}^{\mathbf{o}})^{\mathbf{T}}\right)^{-\mathbf{1}},$$

где $(\hat{C}^o)^T$ – транспонированная матрица.
Концентрации определяемых элементов C_m по измеренным значениям аналитических сигналов I_n получаются из равенства

$$\hat{\mathbf{C}} = \hat{\mathbf{K}}_{\mathbf{ex}} \cdot \hat{\mathbf{I}},$$

где $\hat{K}_{ex} = \left(\hat{K}^T \cdot \hat{K}\right)^{-1} \cdot \hat{K}^T$.

Пример. Рассмотрим построение градуировки при определении трех элементов по значениям трех аналитических сигналов при использовании пяти образцов с известными содержаниями этих элементов. Концентрации элементов в пяти образцах для градуировки имели следующие значения:

$$\hat{\mathbf{C}}^{\mathbf{o}} = \begin{pmatrix} 1,05 & 0,73 & 0,65 & 0,61 & 1,54 \\ 0,82 & 1,23 & 1,51 & 0,73 & 0,74 \\ 0,32 & 0,76 & 1,44 & 0,91 & 1,45 \end{pmatrix},$$

а измеренные величины аналитических сигналов в этих образцах таковы:

$$\hat{\mathbf{I}}^{\mathbf{o}} = \begin{pmatrix} 1,20 & 1,03 & 1,05 & 0,81 & 1,80 \\ 1,75 & 2,50 & 3,36 & 1,65 & 1,81 \\ 1,05 & 2,31 & 4,43 & 2,79 & 4,50 \end{pmatrix}.$$

Находим

$$\hat{\mathbf{I}}^{\mathbf{o}} \cdot (\hat{\mathbf{C}}^{\mathbf{o}})^{\mathbf{T}} =$$

$$= \begin{pmatrix} 1,20 & 1,03 & 1,05 & 0,81 & 1,80 \\ 1,75 & 2,50 & 3,36 & 1,65 & 1,81 \\ 1,05 & 2,31 & 4,43 & 2,79 & 4,50 \end{pmatrix} \begin{pmatrix} 1,05 & 0,82 & 0,32 \\ 0,73 & 1,23 & 0,76 \\ 0,65 & 1,51 & 1,44 \\ 0,61 & 0,73 & 0,91 \\ 1,54 & 0,74 & 1,45 \end{pmatrix} =$$

$$= \begin{pmatrix} 5,960 & 5,760 & 6,026 \\ 9,640 & 12,128 & 11,424 \\ 14,300 & 15,758 & 17,535 \end{pmatrix},$$

$$\hat{\mathbf{C}}^{\mathbf{o}} \cdot (\hat{\mathbf{C}}^{\mathbf{o}})^{\mathbf{T}} =$$

$$= \begin{pmatrix} 1,05 & 0,73 & 0,65 & 0,61 & 1,54 \\ 0,82 & 1,23 & 1,51 & 0,73 & 0,74 \\ 0,32 & 0,76 & 1,44 & 0,91 & 1,45 \end{pmatrix} \begin{pmatrix} 1,05 & 0,82 & 0,32 \\ 0,73 & 1,23 & 0,76 \\ 0,65 & 1,51 & 1,44 \\ 0,61 & 0,73 & 0,91 \\ 1,54 & 0,74 & 1,45 \end{pmatrix} =$$

$$= \begin{pmatrix} 4,802 & 4,325 & 4,615 \\ 4,325 & 5,546 & 5,109 \\ 4,615 & 5,109 & 5,684 \end{pmatrix}.$$

Тогда матрица $\hat{K}$, позволяющая вычислить ожидаемые значения аналитических сигналов по известной концентрации, равна

$$\hat{\mathbf{K}} = \begin{pmatrix} 0,983 & 0,177 & 0,103 \\ 0,019 & 1,945 & 0,246 \\ 0,063 & -0,014 & 3,046 \end{pmatrix},$$

а матрица, позволяющая определить концентрации по измеренным значениям аналитических сигналов, в данном случае равна просто обратной к матрице $\hat{K}$:

$$\hat{\mathbf{K}}_{\mathbf{ex}} = \begin{pmatrix} 1,021 & -0,093 & -0,027 \\ -0,007 & 0,514 & -0,041 \\ -0,021 & 0,004 & 0,329 \end{pmatrix}.$$

Недиагональные элементы этой матрицы описывают взаимное влияние элементов. Теперь по экспериментальным значениям величин аналитического сигнала для анализируемого образца можно определить содержания элементов. Например, для $I_n = (1,15;\ 2,23;\ 3,39)$ получаем

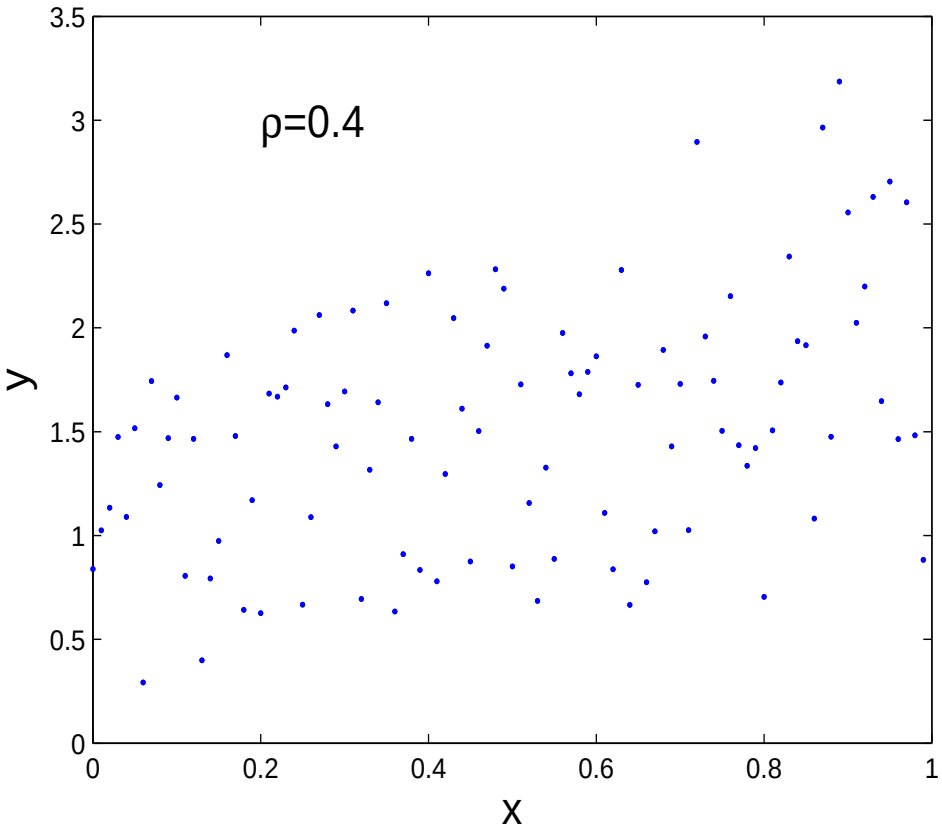
$$\hat{\mathbf{C}} = \begin{pmatrix} 1,022 & -0,093 & -0,028 \\ -0,007 & 0,514 & -0,041 \\ -0,021 & 0,004 & 0,329 \end{pmatrix} \cdot \begin{pmatrix} 1,15 \\ 2,23 \\ 3,39 \end{pmatrix} = \begin{pmatrix} 0,88 \\ 1,00 \\ 1,10 \end{pmatrix}.$$

Степень надежности используемой модели можно оценить с помощью количественного критерия, вычисляя коэффициент корреляции r, определяемый соотношением

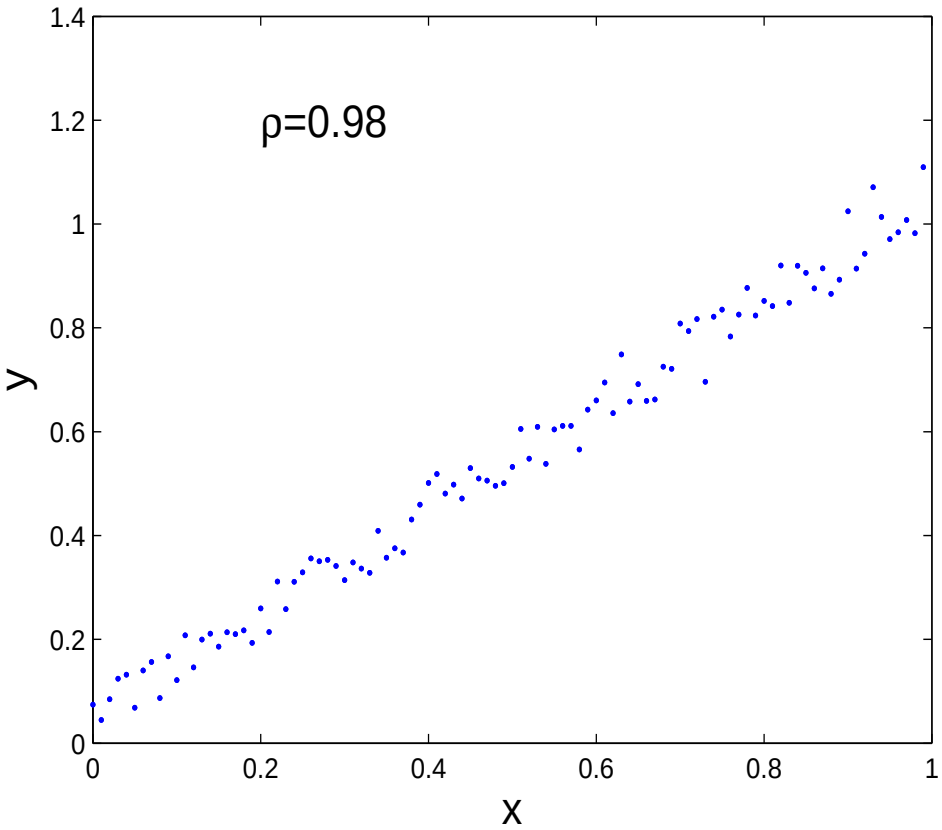
$$\mathbf{r} = \frac{\mathbf{S_{IC}}}{\mathbf{S_I S_C}} = \frac{\sum_n \mathbf{C_n I_n} - \mathbf{N\bar{C}\bar{I}}}{(\mathbf{N}-1)\mathbf{S_C S_I}}.$$

Если число измерений N мало, то возможно, что они случайно выстроятся вдоль прямой. Количественно определить, насколько надежно установлена функциональная зависимость одной случайной величины от другой, можно по значению коэффициента корреляции r.

Корреляционный анализ



Корреляционный анализ



Корреляция считается статистически значимой, если рассчитанная величина r превышает граничное значение $r(P, f = N - 2)$ при заданной доверительной вероятности P и числе измерений N. В таблице приведены граничные значения коэффициентов корреляции для значений доверительной вероятности $P = 0,95$ и $P = 0,99$ при различном числе измерений.

Число измерений	P= 0,95	P= 0,99	Число измерений	P= 0,95	P= 0,99
3	1,00	1,00	14	0,53	0,66
4	0,95	0,99	16	0,50	0,62
5	0,88	0,96	18	0,47	0,59
6	0,81	0,92	20	0,44	0,56
7	0,75	0,87	22	0,42	0,54
8	0,71	0,83	27	0,38	0,49
9	0,67	0,80	32	0,35	0,45
10	0,63	0,77	42	0,30	0,39
11	0,60	0,74	52	0,27	0,35
12	0,58	0,71	72	0,23	0,30

Отметим, что при увеличении числа экспериментальных результатов граничные значения коэффициентов корреляции уменьшаются. Это говорит о том, что случайное возникновение корреляции становится маловероятным.