# О ВОЗМОЖНОСТИ ПРИМЕНЕНИЯ МЕТОДОВ МОНТЕ-КАРЛО В АНАЛИЗЕ НЕЛИНЕЙНЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ

© 2014 г. Г. И. Рудой*
*(*119333 Москва, ул. Вавилова, 40, ВЦ РАН)*
*e-mail: 0xd34df00d@gmail.com*
*УДК 519.65, 519.245*

**Аннотация**

Предлагается понятие устойчивости коэффициентов существенно нелинейных суперпозиций, а также метод оценки устойчивости решения задачи восстановления регрессионной зависимости. Данный метод иллюстрируется вычислительным экспериментом на данных, полученных при измерении зависимости показателя преломления полимера от длины волны в области 400-1000 нм, соответствующей области прозрачности полимера, а также исследуется сходимость предложенного метода к известному аналитическому решению для линейной зависимости.

**Ключевые слова**: *символьная регрессия, нелинейные модели, устойчивость решений, дисперсия прозрачной среды, методы класса Монте-Карло.*

## 1 Введение

Символьная регрессия часто используется для построения экспертно интерпретируемых моделей [1–5]. В приложении к естественнонаучным экспериментам речь идет о восстановлении функциональной зависимости между измеряемыми и задаваемыми с некоторой точностью параметрами, как то: зависимость термоэмиссионного тока электронной лампы от температуры катода $I_k(T)$ при неизменных геометрии системы и разности потенциалов, зависимость мощности излучения непрерывного лазера от коэффициента отражения выходного зеркала $W_l(R)$ при постоянных модовой структуре излучения и мощности возбуждения активной среды, зависимость показателя преломления материала от длины волны $n(\lambda)$ при постоянной температуре и т. п., далее мы более подробно рассмотрим именно последний случай.

При регрессионном анализе такого рода экспериментов необходимо учитывать следующие обстоятельства:

1. Все измеряемые (и контролируемые) параметры в каждой экспериментальной точке определяются с некоторой (обычно известной) точностью, причем абсолютная погрешность $\sigma_i$ соответствующего параметра может существенно изменяться в исследуемом диапазоне. Например, если в качестве спектрального прибора, выделяющего конкретную длину волны $\lambda_i$ при измерении $n_i(\lambda_i)$, используется дифракционная решетка, то $\frac{\sigma_i}{\lambda_i} \approx \text{const}$, и считать погрешность определения длины волны постоянной некорректно для измерений в достаточно широком спектральном диапазоне.

2. Как правило, эксперимент ставится так, что измеряется функциональная зависимость от одной переменной $x$, то есть, строится зависимость вида $y(x, \boldsymbol{\omega})$, где $\boldsymbol{\omega}$ — набор параметров, которые поддерживаются неизменными. Как отмечалось выше, параметры поддерживаются постоянными с конечной точностью и в ряде случаев при построении модели это обстоятельство необходимо учитывать. Однако обычно

эксперт заранее может оценить влияние вариаций условий эксперимента и обеспечить необходимую стабильность проведения измерений. В противном случае необходимо прямо учитывать зависимость измеряемой характеристики от нескольких переменных, что для целей настоящей работы непринципиально.

3. В большинстве случаев эксперт заранее знает вид искомой функциональной зависимости, или же требуется провести выбор между несколькими возможными вариантами, что упрощает задачу регрессии. В то же время для эксперта важнейшее значение имеют не только оптимальные значения коэффициентов регрессионной формулы, но и дисперсия этих коэффициентов, а также связь их дисперсии с точностью определения измеряемых (контролируемых) в эксперименте величин. Это особенно существенно в тех случаях, когда коэффициенты регрессионной модели прямо связаны с фундаментальными характеристиками исследуемого процесса и по ним рассчитывается, например, эффективная масса электронов в полупроводнике, температура Дебая, резонансная частота и затухание оптического перехода и т. д. — соответственно, точность измерения соответствующих материальных констант определяется точностью вычисления коэффициентов регрессионной модели.

В такой постановке, когда требуется определить не только оптимальные коэффициенты регрессионной модели, но и их погрешность, насколько нам известно, задача нелинейной регрессии не рассматривалась. Известны теоретические результаты для случая линейной регрессии:

$$y = ax + b,$$

в случае, когда дисперсия всех экспериментально измеренных значений $y_i$ зависимой переменной $y$ одна и та же $D(y_i) = \sigma^2$, а значения независимой переменной $x_i$ известны точно: $D(x) = 0$. Тогда при переходе к представлению

$$y_i = a(x_i - \overline{x}) + b + \xi_i \mid i \in \{1, \ldots, \ell\},$$

где $\overline{x} = \frac{\sum_{i=1}^{\ell} x_i}{\ell}$, а $\xi_i \sim \mathcal{N}(0, \sigma^2)$, согласно [6], случайные величины $a$ и $b$ независимы и нормально распределены, и, кроме того, их дисперсии выражаются известными соотношениями:

$$D(a) = \frac{\sigma^2}{\sum_{i=1}^{\ell}(x_i - \overline{x})^2}. \tag{1}$$

$$D(b) = \frac{\sigma^2}{\ell}. \tag{2}$$

В настоящей работе предложен общий метод определения погрешности коэффициентов нелинейной регрессии, и на примере зависимости $n(\lambda)$ для прозрачного полимера определена зависимость погрешности параметров регрессии от точности определения длины волны и показателя преломления. Здесь мы ограничиваемся одной независимой переменной $\lambda$. Обобщение предлагаемого метода на случай нескольких переменных проводится очевидным образом.

## 2 Основная гипотеза

Пусть имеется выборка $D = \{(x_i, y_i)\} \mid i = \{1, \ldots, \ell\}$, причем для каждого значения $x_i, y_i$ известно распределение вероятности отклонения независимой и зависимой переменных

$P_i^x(x - x_i)$ и $P_i^y(y - y_i)$ от их средних значений $x_i$ и $y_i$ соответственно. Вероятности $P_i^x(x-x_i)$ и $P_i^y(y-y_i)$ обычно принимаются гауссовыми, и для них считаются известными значения дисперсий $\sigma_i^x$, $\sigma_i^y$.

Пусть далее с помощью некоторого алгоритма регрессии строится зависимость $y(x, \boldsymbol{\omega})$, минимизирующая некоторый функционал S, например, среднеквадратичное отклонение:

$$S = \sum_{i=1}^{\ell} \big(y(x_i, \boldsymbol{\omega}) - y_i\big)^2 \underset{\boldsymbol{\omega}}{\to} \min. \tag{3}$$

Для таким образом определенного функционала, а также для его модификаций, учитывающих сложность регрессионной модели [5], процедура минимизации эффективно проводится с помощью алгоритма Левенберга-Марквардта (АЛМ) [7, 8].

Далее фиксируем структурный вид полученной зависимости $y(x, \boldsymbol{\omega})$ и многократно повторяем следующую вычислительную процедуру:

1. На $k$-м шаге генерируется случайная выборка $D_k = \{(x_i^k, y_i^k)\} \mid i = \{1, \ldots, \ell\}$. Значение $x_i^k$ получается из соответствующего значения $x_i$ исходной выборки $D$ путем добавления случайного шума, распределенного согласно $P_i^x$:

   $$x_i^k = x_i + \xi_i^k \mid \xi_i^k \sim P_i^x.$$

   Аналогично получается $y_i^k$.

2. Для таким образом построенного набора данных $D_k$ (далее — реализация), используя один и тот же алгоритм оптимизации, находим оптимальный (минимизирующий выбранный функционал) набор $\boldsymbol{\omega}^k$ коэффициентов регрессии $y(x, \boldsymbol{\omega})$ для $k$-й реализации.

   Таким образом, для каждого конкретного коэффициента регрессии $\boldsymbol{\omega}_p$ получаем совокупность его значений в сгенерированных реализациях $\{\boldsymbol{\omega}_p^k\}$.

3. Для достаточно большого числа реализаций $M$ обычным образом определим среднее значение и стандартное отклонение соответствующего коэффициента регрессии $\boldsymbol{\omega}_p$:

   $$\overline{\boldsymbol{\omega}_p} = \frac{\sum_{i=1}^{M} \boldsymbol{\omega}_p^i}{M}, \tag{4}$$

   $$D(\boldsymbol{\omega}_p) = \sigma_{\boldsymbol{\omega}_p}^2 = \frac{1}{M-1} \sum_{i=1}^{M} (\boldsymbol{\omega}_p^i - \overline{\boldsymbol{\omega}_p})^2. \tag{5}$$

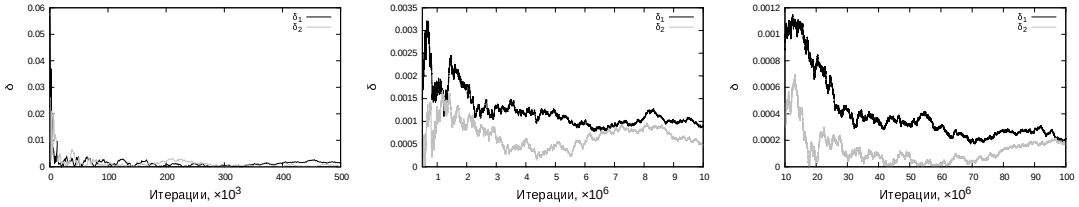
Наша гипотеза состоит в том, что полученные согласно (4) и (5) значения соответствуют реальности. Предлагаемый подход к определению погрешности регрессионных коэффициентов, очевидно, представляет собой фактически применение метода типа Монте-Карло к задаче регрессии.

Из предложенной интерпретации также следует очевидный критерий останова вычислительной процедуры, когда с ростом числа реализаций вариация значений $\overline{\boldsymbol{\omega}_p}$ и $D(\boldsymbol{\omega}_p)$ становится меньше экспертно выбранного значения.

Необходимо заметить, что в общем случае пределы выражений типа (4) и (5) при $M \to \infty$ могут и не существовать, что делает предложенную вычислительную схему некорректной. Однако для достаточно гладких функций, которые собственно и представляют практический интерес, корректность предложенной процедуры можно строго доказать, что, однако, выходит за рамки данной работы.

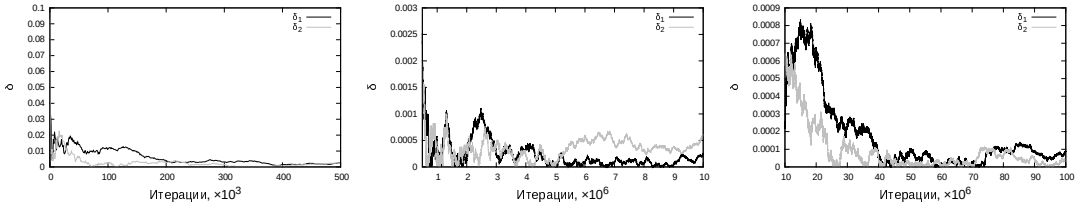
## 3 Модельный случай

Проверим предлагаемый метод на случае с известным аналитическим решением (1)-(2), рассчитав погрешности коэффициентов линейной зависимости с точно известной независимой переменной и гауссовым распределением ошибки зависимой переменной, при этом параметры распределения погрешности зависимой переменной одинаковы для каждой экспериментальной точки. В вычислительном эксперименте моделируемая зависимость имела вид $y = ax + b \mid a = 3, b = 10$. Независимая переменная определена в $\ell = 10$ (в другом эксперименте $\ell = 50$) точках отрезка $[0, 10]$, количество реализаций составляло 100 миллионов. Оптимизация коэффициентов регрессии проводилась с помощью алгоритма Левенберга-Марквардта по аналогии с более общим случаем нелинейной регрессии.



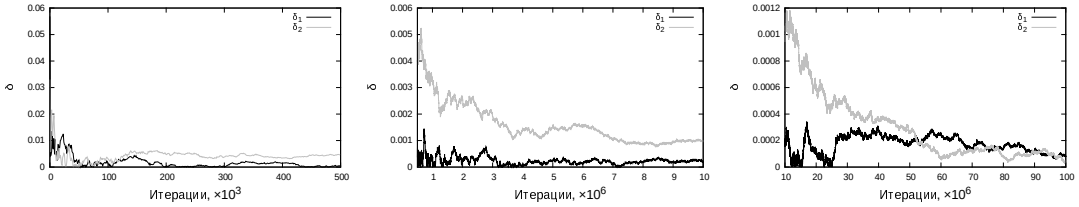

(a) $N \in [0;\ 5 \cdot 10^5]$ (b) $N \in [5 \cdot 10^5;\ 10^7]$ (c) $N \in [10^7;\ 10^8]$

Рис. 1: Зависимость $\delta$ от числа итераций $N$ при $D(\xi) = 10$ и $\ell = 10$.




(a) $N \in [0;\ 5 \cdot 10^5]$ (b) $N \in [5 \cdot 10^5;\ 10^7]$ (c) $N \in [10^7;\ 10^8]$

Рис. 2: Зависимость $\delta$ от числа итераций $N$ при $D(\xi) = 1$ и $\ell = 10$.




(a) $N \in [0;\ 5 \cdot 10^5]$ (b) $N \in [5 \cdot 10^5;\ 10^7]$ (c) $N \in [10^7;\ 10^8]$

Рис. 3: Зависимость $\delta$ от числа итераций $N$ при $D(\xi) = 1$ и $\ell = 50$.

На рис. 1-3 по оси абсцисс указано количество реализаций $N$, а по оси ординат для каждого из коэффициентов $a$, $b$ отношение $\delta = \frac{\sigma_{c.e.} - \sigma_t}{\sigma_t}$, где $\sigma_{c.e.}$ — значение дисперсии, полученное в вычислительном эксперименте, а $\sigma_t$ — точное теоретическое значение дисперсии согласно (1)-(2).

Как видно, уже для $N \approx 2.5 \cdot 10^7$ относительное различие $\frac{\sigma_{c.e.} - \sigma_t}{\sigma_t}$ не превышает 0.05% вне зависимости от количества точек, в которых определена независимая переменная, и значения дисперсии случайной величины. Это, на наш взгляд, является хорошим результатом и свидетельствует в пользу корректности обсуждаемого подхода и основной гипотезы.

# 4 Погрешность регрессионных коэффициентов зависимости $n(\lambda)$

В эксперименте при 17 значениях длины волны был измерен показатель преломления прозрачного (в исследуемом спектральном диапазоне) полимера, результаты представлены в таблице 1.

| $\lambda$, нм | 435.8 | 447.1 | 471.3 | 486.1 | 501.6 | 546.1 | 577.0 | 587.6 | 589.3 |
|---|---|---|---|---|---|---|---|---|---|
| $n$ | 1.35715 | 1.35625 | 1.35449 | 1.35349 | 1.35275 | 1.35083 | 1.34968 | 1.34946 | 1.34938 |
| $\lambda$, нм | 656.3 | 667.8 | 706.5 | 750 | 800 | 850 | 900 | 950 | |
| $n$ | 1.34768 | 1.34740 | 1.34664 | 1.34607 | 1.34544 | 1.34487 | 1.34437 | 1.34407 | |

Таблица 1: Экспериментальные значения коэффициентов преломления.

Абсолютная погрешность измерения показателя преломления $\sigma_n$ составляет обычно $(3 \div 10) \cdot 10^{-5}$ во всем диапазоне длин волн, относительная погрешность определения длины волны $\frac{\sigma_\lambda}{\lambda}$ для используемых дифракционных приборов составляет $(3 \div 30) \cdot 10^{-4}$.

Для предложенной экспертом модели зависимости $n(\lambda)$ требуется определить погрешность коэффициентов регрессии, обусловленную экспериментальными погрешностями измерения длины волны и показателя преломления.

Кроме того, целесообразно найти зависимость погрешности коэффициентов регрессии от точности эксперимента, что имеет не только теоретический, но и практический интерес. Например, улучшение точности измерения показателя преломления $n$ до $2 \cdot 10^{-5}$ требует значительных усилий (и затрат), но если при этом погрешность определения коэффициентов регрессионной модели практически не изменится, то соответствующая работа по модернизации экспериментальной установки не имеет смысла. Точно так же, при улучшении точности определения показателя преломления имеет не только теоретический, но и практический смысл вопрос о том, насколько требуется улучшить точность определения длины волны, чтобы обеспечить значительное снижение погрешности определения коэффициентов регрессии.

Экспертом была предложена функциональная зависимость следующего вида:

$$n(\lambda) = a + \frac{b}{\lambda^2} + \frac{c}{\lambda^4}. \tag{6}$$

Для указанного вида нелинейной регрессии с помощью АЛМ были определены оптимальные коэффициенты регрессии, обеспечивающие среднеквадратичную ошибку $\approx 3.97 \cdot 10^{-9} : a = 1.33344; b = 2841.63; c = 1599.27$; длина волны измеряется в нанометрах.

Вычисление погрешностей коэффициентов регрессионной модели было проведено для ряда комбинаций погрешностей определения длины волны (в относительных единицах) и показателя преломления (также в относительных единицах), в каждом случае проводилась статистическая обработка миллиона реализаций. Некоторые результаты вычислительного эксперимента приведены в таблицах 2-4.

| $\frac{\sigma_\lambda}{\lambda}$ \ $\frac{\sigma_n}{n}$ | $2 \cdot 10^{-5}$ | $1 \cdot 10^{-4}$ | $3 \cdot 10^{-4}$ |
|---|---|---|---|
| $10^{-3}$ | $1.255 \cdot 10^{-5}$ | $5.54 \cdot 10^{-5}$ | $1.656 \cdot 10^{-4}$ |
| $10^{-2}$ | $6.32 \cdot 10^{-5}$ | $8.24 \cdot 10^{-5}$ | $1.753 \cdot 10^{-4}$ |

Таблица 2: Относительная погрешность определения коэффициента регрессии $a$.

| $\frac{\sigma_\lambda}{\lambda}$ \ $\frac{\sigma_n}{n}$ | $2 \cdot 10^{-5}$ | $1 \cdot 10^{-4}$ | $3 \cdot 10^{-4}$ |
|---|---|---|---|
| $10^{-3}$ | $2.14 \cdot 10^{-3}$ | $8.68 \cdot 10^{-3}$ | $2.54 \cdot 10^{-2}$ |
| $10^{-2}$ | $1.36 \cdot 10^{-2}$ | $1.59 \cdot 10^{-2}$ | $2.85 \cdot 10^{-2}$ |

Таблица 3: Относительная погрешность определения коэффициента регрессии $b$.

Для $\frac{\sigma_\lambda}{\lambda} < 10^{-4}$ погрешности первого и второго коэффициентов регрессии очень слабо зависят от дальнейшего увеличения точности определения длины волны и почти линейно зависят от точности измерения показателя преломления. Абсолютная погрешность определения третьего коэффициента регрессии высока и сравнима по порядку величины с самим коэффициентом, что, очевидно, связано с незначительным вкладом третьего слагаемого в (6) в показатель преломления: изменение этого слагаемого слабо влияет на общую величину среднеквадратичного отклонения, следовательно, этот компонент определяется с низкой точностью.

## 5 Заключение

1. Предложенный в настоящей работе метод определения погрешностей коэффициентов регрессии может использоваться при любом распределении вероятности ошибки зависимых и независимых переменных, включая различные виды распределения ошибки для разных переменных, а также для одной переменной в различных экспериментальных точках.

2. Предложенный метод будет давать несколько разные значения погрешностей при использовании различных функционалов ошибки $S$. Такая ситуация не является необычной [5], и, например, коэффициенты регрессии будут разными, если по стандартному методу наименьших квадратов (МНК) минимизируется сумма квадратов расстояний по оси, соответствующей зависимой переменной, от экспериментальных точек до аппроксимирующей регрессионной модели согласно (3) или же минимизируется сумма евклидовых расстояний от экспериментальных точек до аппроксимирующей модели аналогично известной работе Пирсона [9].

   С нашей точки зрения, выбор оптимального функционала G проводится экспертом, исходя, прежде всего, из анализа погрешностей определения зависимых и независимых переменных. В самом деле, если независимая переменная определяется точно, то МНК предпочтителен, а при близких погрешностях переменных подход Пирсона представляется предпочтительным.

3. Для оптимизации коэффициентов регрессии использовался АЛМ в стандартной форме, когда минимизируется функционал типа (3). Как указано выше, этот вари-

| $\frac{\sigma_\lambda}{\lambda}$ \ $\frac{\sigma_n}{n}$ | $2 \cdot 10^{-5}$ | $1 \cdot 10^{-4}$ | $3 \cdot 10^{-4}$ |
|---|---|---|---|
| $10^{-3}$ | 1.15 | 1.426 | 1.454 |
| $10^{-2}$ | 1.45 | 1.456 | 1.47 |

Таблица 4: Относительная погрешность определения коэффициента регрессии $c$.

ант МНК корректен, когда погрешность независимой переменной мала, а в общем случае метод АЛМ следует модифицировать, минимизируя суммы вида $(\frac{\Delta x_i}{\sigma_{x_i}})^2 + (\frac{\Delta y_i}{\sigma_{y_i}})^2$. Однако для конкретных вычислительных экспериментов, результаты которых представлены в настоящей работе, возможно использование АЛМ в стандартной форме.

В самом деле, заранее предполагается, что независимая переменная известна точно: $\forall i : D(x_i) = 0$. Далее, можно показать, что использование функционала (3) в вычислительном эксперименте корректно, если $\frac{\sigma_n}{\sigma_\lambda} > \frac{dn}{d\lambda}$ и для анализируемого полимера это условие выполняется с кратным запасом.

Таким образом, в настоящей работе предложен метод определения погрешностей коэффициентов регрессионной модели, включая случай нелинейной регрессии. В вычислительном эксперименте получено хорошее соответствие теоретически описанному случаю, когда стандартное отклонение коэффициентов регрессии определяется точно. Показана целесообразность применения предложенного вычислительного алгоритма при анализе конкретных физических экспериментов.

## Список литературы


[1] Davidson, J. W., Savic, D. A., and Walters, G. A.: *Symbolic and numerical regression: experiments and applications.* In John, Robert and Birkenhead, Ralph (editors): *Developments in Soft Computing*, pages 175–182, De Montfort University, Leicester, UK, 29-30 6 2000. 2001. Physica Verlag, ISBN 3-7908-1361-3.

[2] Sammut, C. and Webb, G. I.: *Symbolic regression.* In Sammut, Claude and Webb, Geoffrey I. (editors): *Encyclopedia of Machine Learning*, page 954. Springer, 2010, ISBN 978-0-387-30768-8. `http://dx.doi.org/10.1007/978-0-387-30164-8`.

[3] Strijov, V. and Weber, G. W.: *Nonlinear regression model generation using hyperparameter optimization.* Computers & Mathematics with Applications, 60(4):981–988, 2010. `http://dx.doi.org/10.1016/j.camwa.2010.03.021`.

[4] Стрижов, В. В.: *Методы индуктивного порождения регрессионных моделей.* Препринт ВЦ РАН им. А. А. Дородницына. — М., 2008.

[5] Рудой, Г. И. и Стрижов, В. В.: *Алгоритмы индуктивного порождения суперпозиций для аппроксимации измеряемых данных.* Информатика и ее применения, 7(1):44–53, 2013.

[6] Ватутин, В. А., Ивченко, Г. И., Медведев, Ю. И., и Чистяков, В. П.: *Теория вероятностей и математическая статистика в задачах.* Дрофа, 3 редакция, 2005.

[7] Marquardt, D. W.: *An algorithm for least-squares estimation of non-linear parameters.* Journal of the Society of Industrial and Applied Mathematics, 11(2):431–441, 1963.

[8] More, J. J.: *The Levenberg-Marquardt algorithm: Implementation and theory.* In *G.A. Watson*, Lecture Notes in Mathematics 630, pages 105–116. Springer-Verlag, Berlin, 1978. Cited in Åke Björck's bibliography on least squares, which is available by anonymous ftp from `math.liu.se` in `pub/references`.

[9] Pearson, K.: *On lines and planes of closest fit to systems of points in space.* Philosophical Magazine, 2(6):559–572, 1901.