УДК519.237.5

Антонов Владимир Александрович
доктор технических наук,
главный научный сотрудник,
Институт горного дела УрО РАН,
620219, г. Екатеринбург,
ул. Мамина-Сибиряка, 58
e-mail: Antonov@igduran.ru

ДОСТОВЕРНОСТЬ РЕГРЕССИОННЫХ МОДЕЛЕЙ В ГОРНОТЕХНОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ*

Аннотация:

Изложены методические приемы в оценках достоверности регрессионной модели горнотехнологических объектов и процессов с учетом однократных и многократных измерений, разделяющихся условно по предложенному критерию. Реализация эффекта многократности измерений, как показано на приведенном примере, позволяет повысить достоверность выявления искомой закономерности.

Ключевые слова: экспериментальные измерения, закономерность, случайные отклонения, модель регрессии, коэффициент детерминации

Antonov Vladimir A.
doctor of technical sciences,
chief researcher,
The Institute of Mining UB RAS,
620219, Yekaterinburg,
Mamin-Sibiryak st., 58
e-mail: Antonov@igduran.ru

THE RELIABILITY OF REGRESSION MODELS IN MINING AND TECHNOLOGICAL RESEARCHES

Abstract:

Methodical techniques in the regression model estimates of the reliability of mining and technological facilities and processes are stated including single and multiple dimensions, that are conditionally separated according to the proposed criterion. The realization of measurements multiple effect as shown in the cited example, can improve the reliabilityy of detection the desired regularity.

Key words: experimental measurements, regularity, random deviations, regression model, the coefficient of determination

Введение. Экспериментальные исследования в горном деле проводят на основе измерений физических величин, характеризующих состояние горнотехнологических объектов и процессов. Поиск закономерности в изменении некоторой величины Y, зависящей от величин X_j, где j =1, 2, 3 …, проводят путем их совместных измерений с последующим построением по ряду полученных узловых i-точек (X_{ji}, Y_i,) уравнения регрессии. Условно представим, что результат каждого измерения величины Y_i состоит из компонент значимой для достижения цели исследования (закономерной) и незначимой (случайного отклонения). Здесь принимается, что основной целью исследования является построение модели регрессионной зависимости $Y(X_j)$, представляющей со случайным допустимым отклонением, т.е. погрешностью, значимую закономерность как основную взаимосвязь исследуемого природного явления, повторяющуюся в независимых экспериментах. Достоверность построенной модели проверяется ее адекватностью, т.е. соответствием отображения закономерной и случайной составляющей. Оценки проводятся по методике, описанной в работах [1, 2]. По случайным отклонениям, объясняемым несовершенством (погрешностью) средств измерений зависимой величины и влиянием неучтенных в модели незначимых факторов эксперимента, оценивается интервал адекватного коэффициента ее детерминации.

Часто случайные отклонения зависимой величины в узловых точках априори неизвестны. Тогда они могут определяться по результатам многократных измерений. По рекомендации межгосударственной стандартизации (РМГ 29-99) многократными считаются повторные измерения зависимой величины одного размера, т.е. с одинаковыми аргументами. Однако во многих экспериментах значения аргументов в узловых точках изменяются с малым или большим сдвигом, что приводит к изменению размера зависимой

* Работа выполнена в рамках конкурсного проекта УрО РАН 12-П-5-1028 «Прогноз технологического развития в горнодобывающих отраслях на основе энергосбережения и модернизации геотехники и технологии горного производства».

величины. В таких условиях разделение узловых точек с однократными и многократными измерениями по признаку их повторяемости остается неопределенным. Отмеченное затруднение в разграничении кратности измерений зависимой величины приводит к невозможности оценить и снизить случайные отклонения и тем самым установить требования к достоверности модели регрессии по упомянутому признаку ее адекватности.

В данной работе рассмотрены методические приемы, направленные на решение поставленной проблемы. По ним условно выделяются и учитываются при построении регрессии экспериментальные узловые точки с многократными измерениями зависимой величины.

Оценка случайных отклонений. Регрессия проводится наиболее достоверно при наличии однородности исходной информации, заданной в узловых точках. Под однородностью понимается равное влияние на регрессию всех узловых точек и одинаковые свойства рассеяния измеренных в них значений зависимой величины. При этом каждая узловая точка оказывает существенное влияние лишь на участок регрессии, расположенный в окрестности ее аргументов. Такую окрестность назовем осевым интервалом влияния узловой точки. Очевидно, что чем больше имеется узловых точек, тем меньше на оси j-аргумента размер ΔX_j обозначенного интервала. Выразим его следующим соотношением:

$$\Delta X_j = \frac{X_{jn} - X_{j1}}{n-1},$$

где X_{j1} , X_{jn} – наименьшее и наибольшее значение j-аргумента, соответственно, в первой и последней узловой точке; n – количество узловых точек. При равномерном распределении узловых точек расстояние по оси j-аргумента между соседними точками равно ΔX_j.

Часто в экспериментах однородность информации не выдерживается, т.е. узловые точки распределены по осям аргументов неравномерно. Расстояние по оси j-аргумента между соседними точками существенно меньше или больше ΔX_j. По этому признаку введем следующие допущения в различии узловых точек с однократными и многократными измерениями. Если расстояние по оси хотя бы одного j-аргумента между узловой точкой и смежной с ней соседней точкой равно или больше его ΔX_j, то измерение зависимой величины в узловой точке считаем однократным. Если расстояние по оси каждого j-аргумента между смежными соседними узловыми точками меньше соответствующего ΔX_j, то их количество с таким признаком образует группу узловых точек, в которых измерение зависимой величины принимаем многократным. При этом допускаем, что на малом интервале изменения аргументов групповых точек рельеф соответствующего участка регрессии существенно не изменится.

Положим, что экспериментальные измерения во всех узловых точках проводятся одним средством (прибором, методикой). Отклонения значений зависимой величины, связанные с погрешностью средств измерений и влиянием случайных неучтенных факторов эксперимента, распределены во всех узловых точках одинаково нормально и гомоскедастично. Это означает, что случайные отклонения зависимой величины в однократных и многократных измерениях являются частными реализациями некоторой генеральной совокупности и отличаются лишь количеством точек в выборках. Выделим группы узловых точек с многократными измерениями и рассчитаем экспериментальное среднеквадратичное отклонение $\sigma_э$ зависимой величины в точке как взвешенное внутригрупповое (остаточное) по их совокупности. Расчет проводится по формуле:

$$\sigma_э = \sqrt{\frac{1}{\sum_{\nu=1}^{k} n_\nu} \sum_{\nu=1}^{k} \left(\frac{\sum_{i=1}^{n_\nu} (Y_{\nu i} - Y_\nu)^2}{n_\nu - 1} \right) n_\nu}\ , \qquad (1)$$

где n_ν –количество узловых точек в ν-группе многократных измерений; k – количество групп с многократными измерениями; $Y_{\nu i}$ – значение зависимой величины в узловой i-точке, принадлежащей ν-группе; Y_ν – среднее значение зависимой величины в узловых точках ν-группы. Полученное значение $\sigma_э$ характеризует рассеяние однократного измерения и, согласно принятым допускам по гомоскедастичности, распространяется на все узловые i-точки.

Отметим два случая. Экспериментальная погрешность $\sigma_э$ соизмерима с погрешностью средств измерений $\sigma_п$ ($\sigma_э \approx \sigma_п$). Это означает, что влияние на измерение каких-либо случайных незначимых факторов эксперимента отсутствует. Возможно, что экспериментальная погрешность $\sigma_э$ существенно больше погрешности средств измерений $\sigma_п$ ($\sigma_э >> \sigma_п$). Тогда очевидно, что случайные незначимые факторы эксперимента оказывают влияние на результаты измерений.

Оценим погрешность экспериментальных измерений с учетом их многократности. Узловые точки, содержащиеся в каждой ν-группе, усредним. Таким образом, получим q узловых точек с координатами $X_{j\nu c}$ и $Y_{\nu c}$:

$$X_{j\nu c} = \frac{\sum_{i=1}^{n_{j\nu}} X_{j\nu i}}{n_{j\nu}}, \qquad Y_{\nu c} = \frac{\sum_{i=1}^{n_\nu} Y_{\nu i}}{n_\nu}.$$

Известно, что для выборок, извлеченных с возвращением из нормально распределенной генеральной совокупности, распределение средних значений также является нормальным. С учетом этого определим среднеквадратичное отклонение зависимой величины в ν-узловой точке усреднением по их совокупности следующим образом:

$$\sigma_с = \sqrt{\frac{1}{q}\sum_{\nu=1}^{\nu=q}\frac{\sigma_э^{\,2}}{n_\nu}}\,. \qquad (2)$$

Оценка адекватности регрессии. Достоверность построенных моделей регрессии проверяется по критерию их адекватности случайному среднеквадратичному отклонению зависимой величины, зафиксированной в узловых точках. Для этого рассчитывается интервал допустимых значений адекватного коэффициента детерминации R^2 моделей. В этом интервале они отделяют в зависимой величине с принятой вероятностью P закономерную компоненту от случайной. Нижнее $R^2_н$ и верхнее $R^2_в$ значение адекватного коэффициента детерминации определяется по следующим формулам:

$$R^2_н = 1 - \frac{f\cdot\sigma^2}{\chi^2_{\alpha\,1,f} D_y} \quad \text{и} \quad R^2_в = 1 - \frac{f\cdot\sigma^2}{\chi^2_{\alpha\,2,f} D_y}, \qquad (3)$$

где σ – среднеквадратичное случайное отклонение зависимой величины в узловых точках; $f = \sum_{\nu=1}^{\nu=k}(n_\nu - 1)$ – число степеней свободы в расчете экспериментального среднеквадратичного отклонения $\sigma_э$; $\chi^2_{\alpha 1,f}$ и $\chi^2_{\alpha 2,f}$ – процентные точки распределения Пирсона на соответствующих уровнях значимости α_1 и α_2 ($\alpha_1 = (1+P)/2$, $\alpha_2 = (1-P)/2$); D_y – дисперсия зависимой величины Y в узловых точках. В расчетах (3), проводимых по n узловым точкам с однократными измерениями, или по q усредненным узловым точкам многократных измерений, применяются, соответственно, равенства $\sigma=\sigma_э$, $D_y = D_{yn}$ или $\sigma=\sigma_с$, $D_y = D_{yq}$. В обеих оценках дисперсия закономерной компоненты зависимой величины одинакова. Выразим данное положение следующим равенством:

$$D_{yn} - \sigma_э^2 = D_{yq} - \sigma_с^2.$$

Преобразуем его в соотношение

$$\frac{D_{yq}}{D_{yn}} = \frac{1 - \frac{\sigma_э^2}{D_{yn}}}{1 - \frac{\sigma_с^2}{D_{yq}}}. \qquad (4)$$

После усреднения многократных измерений дисперсия значений зависимой величины, заданных в узловых точках, уменьшается, т. е. $D_{yq} < D_{yn}$. С учетом этого, а также при условиях $\sigma_э^2 < D_{yn}$ и $\sigma_с^2 < D_{yq}$, из (4) получим неравенство

$$\frac{\sigma_с^2}{D_{yq}} < \frac{\sigma_э^2}{D_{yn}},$$

означающее, что при учете эффекта многократности измерений зависимой величины в формулах (3) значения адекватного коэффициента детерминации $R^2_н$ и $R^2_в$ увеличиваются.

После построения и оптимизации регрессионной модели она подвергается испытаниям на достоверность. Адекватной признается модель, коэффициент детерминации которой R^2 удовлетворяет неравенству $R^2_н \le R^2 \le R^2_в$. Если этому неравенству удовлетворяет несколько моделей, то выбирается как наиболее достоверная та из них, коэффициент детерминации которой ближе к середине интервала адекватности. Возможно, что коэффициент детерминации модели окажется меньше нижнего значения интервала адекватности ($R^2 < R^2_н$). Это означает, что отображение искомой закономерности зависимой величины в модели недостаточное и ее следует дополнить с учетом влияния на закономерность ранее упущенных факторов. Если коэффициент детерминации оказался больше верхнего значения интервала адекватности ($R^2 > R^2_в$), то модель содержит избыточную детальную структуру, которая отображает лишь частную реализацию случайных отклонений зависимой величины в данном эксперименте. В повторном эксперименте случайные отклонения зависимой величины в узловых точках перераспределятся с другой реализацией, и, соответственно, изменится избыточная модель уравнения регрессии. Это мешает выявлению искомой закономерности. Следовательно, модель следует упростить, исключив функцию отображения частной реализации случайных факторов.

Результат моделирования регрессионной закономерности представляют ее уравнением $Y_р(X_j)$, ограниченным доверительными интервалами. При наличии лишь однократных измерений в узловых точках по значениям в них зависимой величины и уравнению рассчитывается среднеквадратичное отклонение регрессии $\sigma_{эр}$:

$$\sigma_{эр} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - Y_р)^2}{n - m - 1}},$$

где m – количество коэффициентов в ее уравнении. Результат регрессии с доверительной вероятностью 0,68 представляют в виде $Y_р(X_j) \pm \sigma_{эр}$. При учете эффекта многократности измерений и соответствующем усреднении узловых точек рассчитывается уменьшенное ее среднеквадратичное отклонение $\sigma_{ср}$:

$$\sigma_{ср} = \sqrt{\frac{\sum_{v=1}^{q} (Y_v - Y_р)^2}{q - m - 1}}.$$

Регрессия с учетом погрешности представляется также с доверительной вероятностью 0,68 соотношением $Y_р(X_j) \pm \sigma_{ср}$.

*Пример построения регрессии***.** В исследованиях запыленности воздуха в горной выработке проведена серия измерений поглощенной энергии электромагнитного излучения E, прошедшего через пробу воздушно-пылевой смеси с разным размером частиц d. Результаты совместных измерений величин E и d в виде узловых точек показаны на рис. 1. Полагая, что поглощение электромагнитной энергии зависит от размера пылевых частиц, установим по данным экспериментальных измерений математический вид модели соответствующей регрессионной закономерности $E(d)$.

Погрешность измерения поглощенной энергии электромагнитного излучения, в связи с косвенным методом ее оценки, априори неизвестна. Однако на конечный результат измерений оказывают влияние случайные экспериментальные факторы, связанные с отклонениями состава воздушно-пылевой смеси и колебаниями ее плотности. Определим погрешность по данным эксперимента, принимая во внимание, что координаты узловых точек на оси аргумента d распределены неравномерно. Рассчитаем осевой интервал влияния узловой точки $\Delta X = 9{,}43 \cdot 10^{-7}$ м.

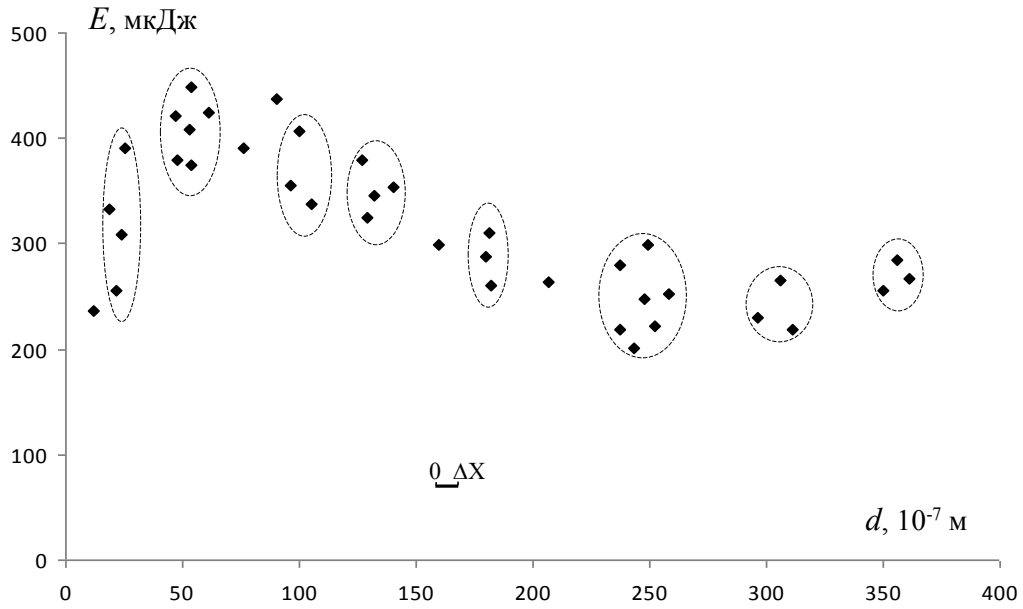

Рис. 1 – Распределение экспериментальных узловых точек с однократными и многократными (в пунктирном овале) измерениями величины E

С учетом его значения на рис. 1 пунктирными овалами выделены группы узловых точек с многократными измерениями. Далее по формуле (1) рассчитано экспериментальное среднеквадратичное отклонение $\sigma_э$=33,25 мкДж в единичном измерении величины E в каждой узловой точке. Данное отклонение $\sigma_э$ обозначено на рис. 2 в виде их вертикальных ограничений.

С учетом отклонения $\sigma_э$ оценен по формуле (3) с вероятностью P=0,95 допустимый интервал адекватного коэффициента детерминации R^2 искомой регрессии в размере от 0,6321 до 0,8521. Соответствующая ее модель построена в виде нелинейного функционально-факторного уравнения, выражающего правостороннее асимметричное распределение. После оптимизации модели методами наименьших квадратов (МНК) и приближений параболической вершины (МППВ) получено ее выражение в следующем конкретном виде:

$$E_1 = 0{,}22498\left(d \cdot 1{,}015135^{-d}\right)^{2{,}079701} + 243{,}82. \qquad (5)$$

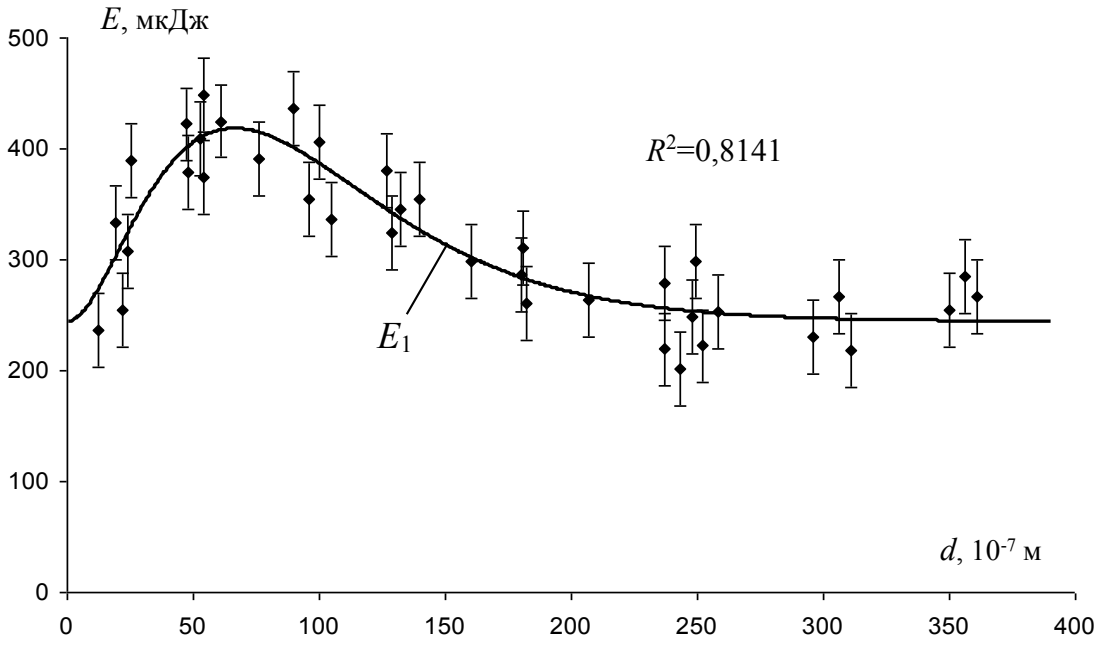

Рис. 2 – Случайные среднеквадратичные отклонения $\sigma_э$ величины E в узловых точках и график ее регрессии E_1

График регрессии показан на рис. 2. Коэффициент ее детерминации R^2=0,8141 и среднеквадратичное отклонение от узловых точек $\sigma_{эр}$=31,95 мкДж соответствует заданному условию адекватности и погрешности $\sigma_э$ измерений энергии E. Это означает, что модель E_1, отсекая или дополняя с вероятностью 0,95 случайные отклонения в значениях энергии E, заданных в узловых точках с однократным измерением, выявляет в них, а также в интервалах интерполяции, закономерную составляющую (5) поглощенной энергии электромагнитного излучения с коэффициентом детерминации 0,8141. Доверительный интервал модели с вероятностью 0,68 выражается соотношением $E_1(d)\pm$ 31,95 мкДж.

Повысим достоверность регрессионной модели, используя эффект многократности измерений в узловых точках. После групповых усреднений их координат количество точек уменьшилось. Среднеквадратичное отклонение измеряемой энергии в точке, рассчитанное по формуле (2), составляет значение σ_c=24,64 мкДж. Расположение узловых точек после усреднения и интервалы их вертикальных отклонений σ_c показаны на рис. 3. По формулам (3) оценен с вероятностью P=0,95 допустимый интервал адекватного коэффициента детерминации R^2 регрессии в размере от 0,7527 до 0,9318. Ее модель так же, как в предыдущем случае, представлена нелинейным функционально-факторным уравнением с правосторонней асимметрией. После оптимизации модели методами МНК и МППВ получено аналогичное уравнение

$$E_2 = 0{,}111491\left(d \cdot 1{,}014545^{-d}\right)^{2{,}271147} + 243{,}148 \qquad (6)$$

с коэффициентом детерминации R^2=0,9253, соответствующим условию его адекватности.

График регрессии показан на рис. 3. Ее среднеквадратичное отклонение от узловых точек снижено и составляет $\sigma_{ср}$=21,89 мкДж, что также соответствует упомянутому значению σ_c. Коэффициенты уравнения (6) мало отличаются от коэффициентов уравнения (5). Уравнения (5) и (6) построены по результатам одного эксперимента, выражают закономерность одного и того же явления. Их графики на рис. 2 и 3 практически одинаковы. Уменьшение головного коэффициента в функциональном слагаемом уравнения (6) компенсируется повышением показателя степени 2,271 вместо 2,0797 так, что их результирующее действие существенно не отличается.

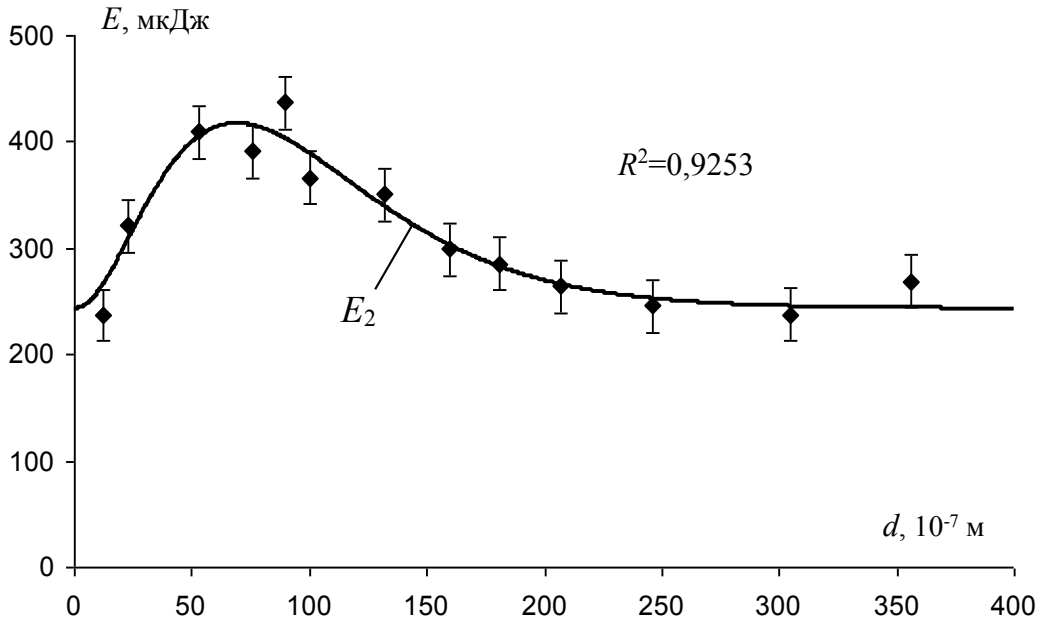

Рис. 3 – Случайные среднеквадратичные отклонения σ_c величины E в узловых точках, усредненных по многократным измерениям, и график ее регрессии E_2

Заключение. Предложенные приемы выделения и учета многократных экспериментальных измерений, как показано на практическом примере, дают возможность оценить адекватность и повысить достоверность регрессионных моделей, отображающих геологические и технологические закономерности в изменении зависимой величины. Практическое применение данной методики приведет к повышению эффективности регрессионного анализа в интерпретации количественных результатов экспериментальных исследований в горном деле и других областях научного знания.

Литература

1. Антонов В. А. Отображение горно-технологических закономерностей функционально факторными уравнениями нелинейной регрессии. / В. А. Антонов, М. В. Яковлев // Горный информационно-аналитический бюллетень. - Проблемы недропользования. - 2011. - С. 571 - 588.

2. Антонов В. А. О достоверности функционально-факторных уравнений регрессии с самоопределяющимися параметрами / В. А. Антонов // Глубинное строение, геодинамика, тепловое поле Земли, интерпретация геофизических полей: шестые научные чтения памяти Ю. П. Булашевича, 12 – 17 сентября: материалы конф. / УрО РАН, Ин-т геофизики. - Екатеринбург: Ин-т геофизики УрО РАН, 2011. - С. 17 - 20.