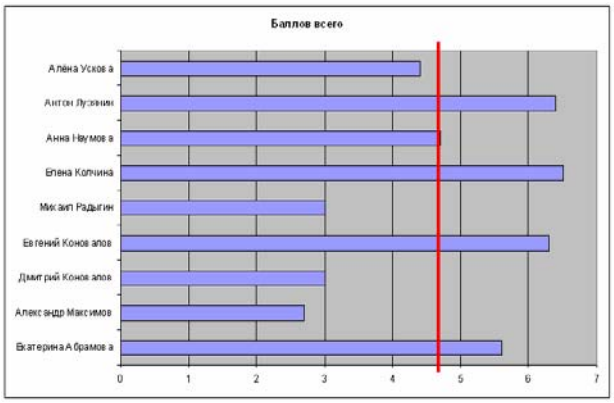


Рис. 1. Диаграмма количества баллов, набранных студентами по типовому расчету на тему «Производная функции нескольких переменных»

Для определения сложности заданий построим диаграмму (рис. 2).

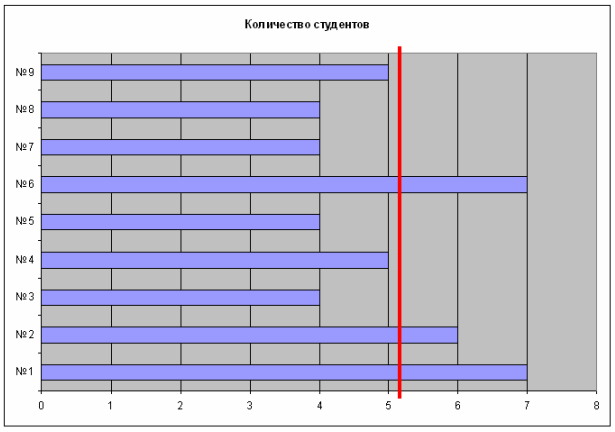


Рис. 2. Диаграмма количества студентов, решивших задачи по типовому расчету на тему «Производная функции одной переменной»

Получено 20.06.14

Среднее значение этого параметра $\bar{x}$ = 5,1. Это говорит о том, что примерно половина студентов справляется с заданиями. Найдем среднеквадратичное отклонение: σ = 1,3; такое отклонение является приемлемым. Для того чтобы студенты лучше справлялись с данным типовым расчетом, нужно заменить задачи, по которым были получены низкие результаты, на более легкие. Это задачи №№ 3, 4, 7, 8. Если заменить их на задачи со сложностью, как у задач № 4 или № 5, то получим $\bar{x}$ = 5,8 и σ = 0,9. Таким образом, замена задач приведет к значительному повышению среднего значения.

Выводы

Была создана информационная технология дистанционного обучения высшей математике и проведено ее экспериментальное исследование на студентах. Был проведен статистический анализ результатов экспериментального исследования, в результате которого были выявлены те задачи, которые необходимо заменить.

Библиографические ссылки

1. *Якушин А. В.* Анализ технологий и систем управления электронным обучением. – М. : Диалектика, 2008. – 25 с.

2. *Леонтьев Б. Е.* Введение в проблематику дистанционного обучения. – М. : Новый Издательский Дом, 2010. – 54 с.

3. Moodle. Домашняя страница // Moodle [Сайт] (дата публикации: 15.05.2008). – URL: https://moodle.org/ (дата обращения: 20.04.2014).

4. Universal Math Solverю. Домашняя страница // Universal Math Solver [Сайт] (дата публикации: 02.09.2005). – URL: http://www.umsolver.com/ (дата обращения: 20.04.2014).

5. *Анисимов А. М.* Работа в системе дистанционного обучения Moodle : учеб. пособие. – 2-е издание. – Харьков : ХНАГХ, 2009. – 40 с.

УДК 658.382

Р. О. Шадрин, кандидат технических наук, доцент, Ижевский государственный технический университет имени М. Т. Калашникова
Б. В. Севастьянов, доктор технических наук, профессор, Ижевский государственный технический университет имени М. Т. Калашникова
Е. Б. Лисина, кандидат технических наук, доцент, Ижевский государственный технический университет имени М. Т. Калашникова
К. В. Гасников, кандидат медицинских наук, доцент, Ижевский государственный технический университет имени М. Т. Калашникова

РЕГРЕССИОННЫЙ АНАЛИЗ В ПРОГНОЗИРОВАНИИ КОЭФФИЦИЕНТА ЧАСТОТЫ НЕСЧАСТНЫХ СЛУЧАЕВ ПО УДМУРТСКОЙ РЕСПУБЛИКЕ

Уравнение регрессии, построенное по выборочным совокупностям статистических данных, позволяет с определенной вероятностью прогнозировать поведение генеральных совокупностей исследуемых величин (показатели производственного травматизма) в рамках некоторого

горизонта прогноза и может быть использовано для расчетов с целью принятия решений на основе установленных закономерностей [1].

В условиях данной задачи будет рассматриваться уравнение множественной регрессии. Для обнаружения значимой статистической зависимости между случайными величинами Y, X_1, X_2, …, X_n ставится задача поиска вида этой зависимости.

В общем случае зависимость ищется в виде функции n переменных: $y = f(x_1, x_2, \ldots, x_n)$. Здесь $\vec{x} = (x_1, x_2, \ldots, x_n)$ – n-мерная случайная величина; y – значение функции $f(x_1, x_2, \ldots, x_n)$. Функцию $y = f(x_1, x_2, \ldots, x_n)$ требуется определить так, чтобы при каждом из значений аргумента $\vec{x} = (x_1, x_2, \ldots, x_n)$ значение функции $f(x_1, x_2, \ldots, x_n)$ было максимально приближено к соответствующему значению случайной величины Y.

Функция f предполагается линейно зависящей от своих аргументов, и уравнение регрессии ищется в виде

$$y = a_0 + a_1x_1 + \ldots + a_jx_j + \ldots + a_nx_n.$$

Для нахождения неизвестных параметров a_0, a_1, …, a_n функции необходимо решить следующую систему уравнений. В условиях данной задачи была выбрана система уравнений с тремя неизвестными:

$$\begin{cases} \sum y = a_0 m + a_1 \sum x_1 + a_2 \sum x_2 + a_3 \sum x_3 , \\ \sum yx_1 = a_0 \sum x_1 + a_1 \sum x_1x_1 + a_2 \sum x_1x_2 + a_3 \sum x_1x_3 , \\ \sum yx_2 = a_0 \sum x_2 + a_1 \sum x_2x_1 + a_2 \sum x_2x_2 + a_3 \sum x_2x_3 , \\ \sum yx_3 = a_0 \sum x_3 + a_1 \sum x_3x_1 + a_2 \sum x_3x_2 + a_3 \sum x_3x_3 , \end{cases}$$

где y – значения коэффициента частоты несчастных случаев; m – объем выборки, для данного случая $m = 13$; x_1, x_2, x_3 – показатели, которые были выбраны исходя из результатов проверки значимости парных коэффициентов корреляции.

В уравнении для коэффициента частоты травматизма $K_ч$: x_1 – средства, израсходованные на мероприятия по охране труда в расчете на одного работающего (S); x_2 – инвестиции в основной капитал в фактически действовавших ценах (I); x_3 – валовой региональный продукт (V).

Система уравнений для коэффициента частоты производственного травматизма $K_ч$:

$$\begin{cases} 57,1 = 13a_0 + 53904,4a_1 + 420,9a_2 + 2282,7a_3, \\ 196427,4 = 53904,4a_0 + 329874061,6a_1 + \\ + 2265829,8a_2 + 12478095,7a_3, \\ 1574,1 = 420,9a_0 + 2265829,8a_1 + 17673,1a_2 + \\ + 94742,4a_3, \\ 8532,3 = 2282,7a_0 + 12478095,7a_1 + 94742,4a_2 + \\ + 513421,6a_3. \end{cases}$$

Значения неизвестных параметров системы находятся методом Крамера. По найденным параметрам составляется уравнение регрессии:

$$K_ч = 6,72 - 0,008 \cdot 10^{-4}S + 0,9325 \cdot 10^{-4}I - 149,6 \cdot 10^{-4}V.$$

Если частные коэффициенты корреляции модели множественной регрессии оказались значимыми, т. е. между результативной переменной и факторными модельными переменными действительно существует корреляционная взаимосвязь, то в этом случае построение множественного коэффициента корреляции считается целесообразным.

С помощью множественного коэффициента корреляции характеризуется совокупное влияние всех факторных переменных на результативную переменную в модели множественной регрессии.

Формула для определения коэффициента корреляции уравнения множественной регрессии через матрицу парных коэффициентов корреляции

$$r_{yx1x2x3} = \sqrt{1 - \frac{\Delta r}{\Delta r_{11}}},$$

где $\Delta r = \begin{vmatrix} 1 & r_{yx1} & r_{yx2} & r_{yx3} \\ r_{yx1} & 1 & r_{x1x2} & r_{x1x3} \\ r_{yx2} & r_{x2x1} & 1 & r_{x2x3} \\ r_{yx3} & r_{x3x1} & r_{x3x2} & 1 \end{vmatrix}$ – определитель матрицы парных коэффициентов корреляции;

$\Delta r_{11} = \begin{vmatrix} 1 & r_{x1x2} & r_{x1x3} \\ r_{x2x1} & 1 & r_{x2x3} \\ r_{x3x1} & r_{x2x3} & 1 \end{vmatrix}$ – определитель матрицы межфакторной корреляции.

Как видно из формул, величина множественного коэффициента корреляции зависит не только от корреляции результата с каждым из факторов, но и от межфакторной корреляции. Рассмотренная формула позволяет определять совокупный коэффициент корреляции, не обращаясь при этом к уравнению множественной регрессии, а используя лишь парные коэффициенты корреляции (табл. 1).

Таблица 1. **Результаты расчетов множественного коэффициента корреляции**

Показатель травмирования	Δr	Δr_{11}	r
$K_ч$	0,000749	0,00819	0,953

Коэффициентом множественной детерминации R^2 называется квадрат множественного коэффициента корреляции. Он характеризует, на сколько процентов построенная модель регрессии объясняет вариацию значений результативной переменной относительно своего среднего уровня, т. е. показывает долю общей дисперсии результативной переменной, объясненной вариацией факторных переменных, включенных в модель регрессии. Чем больше значение коэффициента множественной детерминации, тем лучше построенная модель регрессии характеризует взаимосвязь между переменными.

Для коэффициента множественной детерминации всегда выполняется неравенство

$$R^2\left(y, x_1, \ldots, x_{n-1}\right) \le R^2\left(y, x_1, \ldots, x_n\right).$$

Следовательно, включение в линейную модель регрессии дополнительной факторной переменной не снижает значения коэффициента множественной детерминации. Для $K_ч$: $R^2 = 0{,}909$.

Для того чтобы не допустить преувеличения тесноты связи, применяется скорректированный индекс множественной детерминации, который содержит поправку на число степеней свободы и рассчитывается по формуле

$$R^2_{ск} = 1 - \left(1 - R^2\right)\frac{(n-1)}{(n-m-1)},$$

где n – объем выборки; m – число переменных в уравнении множественной регрессии. При небольшом числе наблюдений нескорректированная величина коэффициента множественной детерминации R^2 имеет тенденцию переоценивать долю вариации результативного признака, связанную с влиянием факторов, включенных в регрессионную модель. Скорректированный индекс множественной детерминации для $K_ч$: $R^2_{ск} = 0{,}879$.

Высокие величины коэффициентов детерминации R^2 указывают на то, что модели регрессии хорошо аппроксимируют исходные данные, и такими регрессионными моделями можно воспользоваться для прогноза значений результативного показателя.

Проверить **значимость** (качество) уравнения регрессии – значит установить, соответствует ли математическая модель, выражающая зависимость между переменными, экспериментальным данным, достаточно ли включенных в уравнение объясняющих переменных для описания зависимой переменной [2]. Чтобы иметь **общее суждение** о качестве модели, по каждому наблюдению из относительных отклонений определяют среднюю ошибку аппроксимации. Проверка **адекватности** уравнения регрессии (модели) осуществляется с помощью **средней ошибки аппроксимации**, величина которой не должна превышать 12-15 % (максимально допустимое значение).

Формула для расчета средней ошибки аппроксимации

$$\overline{\varepsilon} = \frac{1}{n}\frac{\left|y_i - f\left(x_{i1}, x_{i2}, \ldots, x_{in}\right)\right|}{y_i}\cdot 100\ \%,$$

где n – число переменных в уравнении множественной регрессии; $f(x_{i1}, x_{i2}, \ldots, x_{in})$ – i-е расчетное значение переменной y; y_i – i-е опытное значение переменной y. Средняя ошибка аппроксимации для $K_ч$: $\overline{\varepsilon}(\%) = 7{,}63$.

Как видно из результата расчетов, средние ошибки аппроксимации не превышают допустимые значения в 12-15 %, что говорит об адекватности полученных моделей.

Проверка значимости отдельных коэффициентов уравнения означает, что если коэффициент при некоторой переменной незначим, то доверять влиянию этой переменной на значения результирующей функции y нельзя. Незначимый коэффициент следует положить равным нулю, т. е. соответствующую переменную следует исключить из дальнейшего рассмотрения.

Для проверки значимости каждого из коэффициентов a_0, a_1, …, a_n используется t-статистика Стьюдента, опытное значение которой вычисляется по формуле

$$t^{оп}_{a_i} = \frac{a_i}{m_{a_i}}, \quad i = 0, 1, \ldots, n,$$

где a_i – коэффициент при переменной x_i; m_{a_i} – среднеквадратическая ошибка этого коэффициента,

$$m_{a_i} = \frac{\sigma_y\sqrt{1 - R^2_{yx_1,\ldots,x_n}}}{\sigma_{x_i}\sqrt{1 - R^2_{x_i x_1,\ldots,x_n}}}\frac{1}{\sqrt{m-n-1}},$$

где σ_y – среднее квадратичное отклонение для значений переменной y; σ_{x_i} – среднее квадратичное отклонение для значений x_i; $R^2_{yx_1,\ldots,x_n}$ – коэффициент множественной детерминации для уравнения регрессии в целом; $R^2_{x_i x_1,\ldots,x_n}$ – коэффициент множественной детерминации, характеризующий зависимость между фактором x_i и остальными факторами (x_1, x_2, …, x_{i-1}, x_{i+1}, …, x_n) уравнения регрессии.

Каждое из опытных значений статистики $t^{оп}_{a_i}$ сравнивают с критическим значением $t^{кр}_{a_i} = t(\alpha; k)$ (i = 1, 2, …, n), которое ищется по таблице распределения Стьюдента при заданном уровне значимости α и числе степеней свободы k, равном $k = m - n - 1$. В данном случае при уровне значимости α = 0,05 и k = 13 – 3 – 1 = 9 $t^{кр}_{a_i}$ = 2,26 (табл. 2).

Таблица 2. **Рассчитанные опытные значения *t*-статистики Стьюдента**

Показатель травмирования	$t^{оп}_{a_i}$		
	a_1	a_2	a_3
$K_ч$	7,32	7,35	6,45

Если $t^{оп}_{a_i} > t^{кр}_{a_i}$, то гипотеза о значимости коэффициента a_i не отвергается, и соответствующая переменная x_i остается в уравнении. В противном случае коэффициент a_i считается незначимым, и соответствующую ему переменную следует исключить из уравнения регрессии. Таким образом, сравнив полученные опытные значения $t^{оп}_{a_i}$ с критическим $t^{кр}_{a_i}$, можно сделать вывод, что незначимых коэффициентов в уравнении нет.

Если окажется, что при заданном уровне значимости α уравнение незначимо, то пользоваться им нельзя, а найденной зависимостью следует пренебречь.

Для проверки значимости уравнения регрессии используется опытная F-статистика Фишера:

$$F_{\text{оп}} = \frac{\sum_{i=1}^{m}\left[f\left(x_{i1}, x_{i2}, ..., x_{in}\right) - \overline{y}\right]^2}{\sum_{i=1}^{m}\left[y_i - f\left(x_{i1}, x_{i2}, ..., x_{in}\right)\right]^2} \frac{\left(m-n-1\right)}{n},$$

где m – объем выборки; n – число переменных в уравнении множественной регрессии; $f(x_{i1}, x_{i2}, \ldots, x_{in})$ – i-е расчетное значение переменной y; $\overline{y}$ – среднее опытных значений случайной величины Y.

Полученные опытные значения $F_{\text{оп}}$ критерия Фишера (табл. 3) сравниваются с критическими значениями $F_{\text{кр}} = F(\alpha; k_1; k_2)$ при выбранном уровне значимости α. Число степеней свободы $k_1 = m - n - 1$, $k_2 = n$.

При выбранном уровне значимости $\alpha = 0{,}05$ и числе степеней свободы $k_1 = 13 - 3 - 1 = 9$, $k_2 = 3$, $F_{\text{кр}} = 8{,}81$.

При сравнении опытных значений критериев Фишера с критическим (при уровне значимости $\alpha = 0{,}05 F_{\text{кр}} = 8{,}81$), все они удовлетворяют неравенству $F_{\text{оп}} > F_{\text{кр}}$ и делается вывод, что с вероятностью $p = 1 - \alpha = 0{,}95$ уравнение значимо, и мы получаем определенные основания доверять построенным уравнениям регрессии.

Таблица 3. **Рассчитанные опытные значения критерия Фишера**

Показатель травмирования	$\sum_{i=1}^{m}\left[f\left(x_{i1}, x_{i2}, ..., x_{in}\right) - \overline{y}\right]^2$	$\sum_{i=1}^{m}\left[y_i - f\left(x_{i1}, x_{i2}, ..., x_{in}\right)\right]^2$	$F_{\text{оп}}$
$K_{\text{ч}}$	19,63	1,99	29,49

Заключительная статистическая процедура – оценка точности построенных уравнений регрессии.

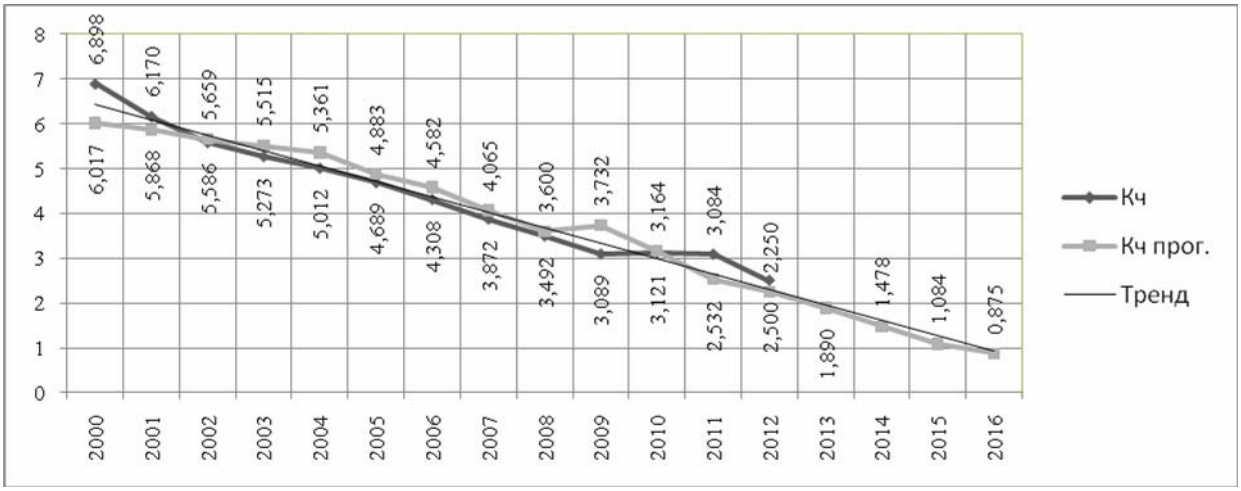
Оценка близости опытных значений y_i случайной величины Y и ее расчетных значений $f(x_i)$, получаемых с помощью уравнения линейной регрессии, выполняется с помощью среднеквадратической погрешности δ по следующей формуле:

$$\delta = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}\left[\left(y_i - \left(a_0 + a_1 x_{i1} + \ldots + a_j x_{ij} + \ldots + a_n x_{in}\right)\right)\right]^2}.$$

Таблица 4. **Результаты расчета среднеквадратичной погрешности уравнений**

	$\sum_{i=1}^{m}\left[y_i - \left(a_0 + a_1 x_{i1} + \ldots + a_j x_{ij} + \ldots + a_n x_{in}\right)\right]^2$	δ
$K_{\text{ч}}$	1,9965	0,408

Для вычисления прогнозных значений воспользуемся данными из Программы социально-экономического развития республики. Результаты представлены на рисунке.



Результаты прогнозирования коэффициента частоты несчастных случаев на 2013–2016 гг.

Библиографические ссылки

1. Отчет по НИР по контракту с Министерством труда УР от 23 августа 2010 № 28/МТ-10 на тему «Разработка модели прогнозирования и управления рисками повреждения здоровья работающими» / исп.: Б. В. Севастьянов, А. П. Тюрин, Р. О. Шадрин, И. Г. Русяк, В. Г. Суфиянов, И. В. Васильева.

2. *Лялькина Г. Б.*, *Бердышев О. В.* Математическая обработка результатов эксперимента : учеб. пособие. – Пермь : Изд-во Перм. нац. иссл. политех. ун-та, 2013. – 78 с.

Получено 20.11.2014