Goryainov A.V., Zamyshlyaev A.M., Platonov E.N.

ANALYSIS OF THE INFLUENCE OF FACTORS ON DAMAGE CAUSED BY TRANSPORT ACCIDENTS USING REGRESSION MODELS[1]

The paper analyzes accidents on transport. It offers to use regression analysis methods to analyze the impact of various factors on traffic safety. The authors propose to apply a model of the dependence of damages caused by transport accidents on a set of negative factors that influence its occurrence. A non-linear regression model with discontinuity is applied for construction of the model. The model is constructed to obtain data on actual or predictable levels of train traffic safety. These data are necessary to assess the adequacy of measures aimed at ensuring the standard safety level to minimize the resources allocated to the tasks of safe train operations, including justification of priorities in the allocation of resources. In the paper, we consider two examples. The first example shows the results of calculations based on real data about accidents on the USA railways. The second example is constructed using the results of computational simulation of the degree of damage and factors.

Keywords*: railway transport, traffic safety, regression analysis.*

Introduction

It is known that train traffic safety is one of the priorities for railway transportation. The Federal Law "About Railway Transport in the Russian Federation" defines train operation safety as a condition of stability of the transportation process, wherein there is no unacceptable risk of accidents and their consequences entailing damage to life and health, to environment and to property of individuals and legal entities.

The study [1] offers a method of assessing a risk level of occurrence of railway accidents based on their relation to factors affecting their occurrence. Various methods of analysis of railway safety are presented in [2 – 7]. When analyzing railway safety, in addition to the probability of an accident occurrence, the size of damage that is caused by this accident should be considered. For the analysis of damage, it is offered to use regression analysis methods [8, 9].

1. Damage analysis caused by transportation accidents

The need to consider the size of damage can be seen from the histogram of damage (see Fig. 1) caused by delays of trains on the USA railway transport in 2010 [10].

[1] The research has been made with RFFI's financial support.

It can be seen that the damage is very different from an accident to an accident, so record keeping of traffic accidents as equal observations may lead to the fact that we will take action to reduce the number of accidents, the damage from which is small. This will happen because the number of accidents with insignificant damage is much higher than that of accidents with great damage, while the total damage from accidents with little damage is much less. The picture is similar for other types of transport accidents.

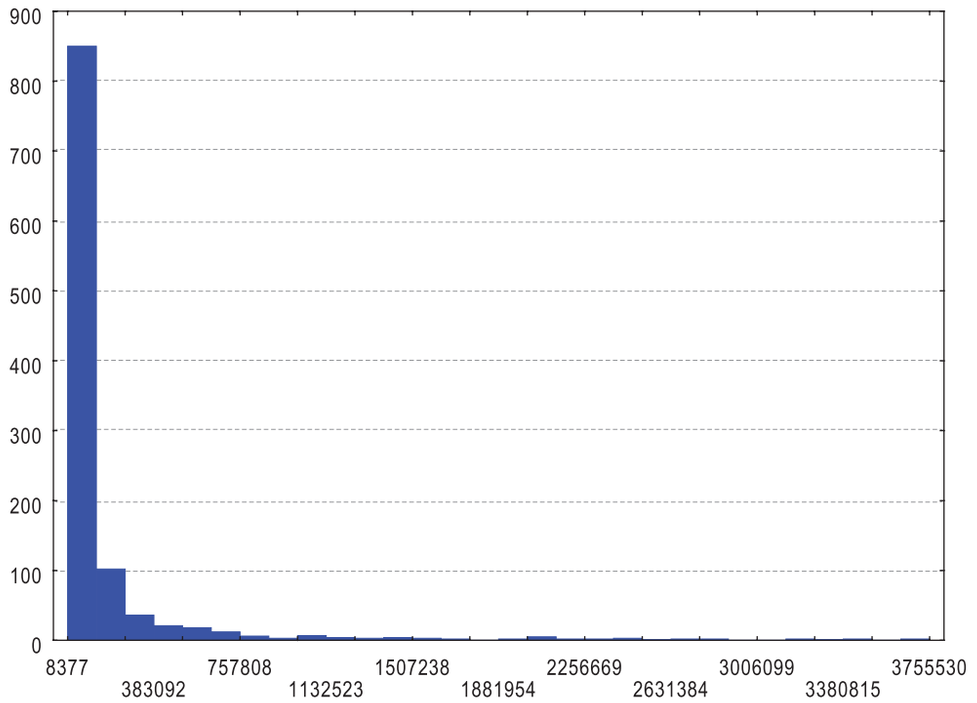

Fig.1. Histogram of damage from train delays

The value of damage can be taken into account if we build a regression dependence on a set of factors. Such models are widely used in the analysis of transport safety [3-7].

Let X be a scalar variable describing the size of damage caused by transport accident of some kind. This may be the damage from a train derailment, a side collision of trains, train arrival delays, etc.

We assume that X depends on m-dimensional vector F of some factors in the sense that the change of the vector F causes some definite changes of the value X.

The latter assumption can be mathematically represented in the form of functional dependence

$$X = f(F), \quad X \in R^1, F \in R^m, \tag{1}$$

where $f(F)$ is a numerical function that depends on m variables.

One of the major tasks is to solve the problem related to definition of the function $f(F)$ that describes the relationship between the variable X and variables F as a result of observations of X and F.

Suppose that we have an opportunity to get simultaneous observations $\{X_k, F_k\}$, $k = 1,\ldots,n$ of X and F in n trials.

Remembering that we introduced the above relationship between X and F, the relation between the results of individual observations can be presented as

$$X_k = f(F_k) + \varepsilon_k, \quad k = 1,\ldots,n,$$

where ε_k is a random error (noise) of the k-th observation.

If we manage to construct a good approximation to the unknown function *f(F)*, then we will have an opportunity to analyze a possible value of accident damage using the value of an observed or supposed vector of factors *F*. We will be able to understand which factors have the greatest effect on the value of damage, and therefore focus on countermeasures against these factors.

Of course, without any prior agreements about a possible form of the function *f(F)*, the problem of its reconstruction based on observations whose number is finite, could be almost impossible. It is usually assumed that the function *f(F)* is linear on variables $F_1,\dots F_m$ [8,9].

Let us consider the linear function $f(F)$ and thus a linear regression model:

$$f(\theta,F)=\theta_0+\theta_1F_1+\theta_2F_2+\cdots+\theta_mF_m,$$

where $\theta=[\theta_0,\theta_1,\dots,\theta_m]^T$ is a vector of unknown parameters.

2. Regression model with discontinuity

Let us consider a piecewise linear model or regression model with discontinuity as an alternative to a linear function [9, 11, 12]:

$$f(\alpha,\beta,b,F)=\begin{cases}\alpha_0+\alpha_1F_1+\alpha_2F_2+\cdots+\alpha_mF_m, & if\ f(\alpha,\beta,b,F)\le b,\\ \beta_0+\beta_1F_1+\beta_2F_2+\cdots+\beta_mF_m, & if\ f(\alpha,\beta,b,F)>b,\end{cases} \tag{2}$$

where $\alpha=[\alpha_0,\alpha_1,\dots\alpha_m]^T$, $\beta=[\beta_0,\beta_1,\dots\beta_m]^T$ are unknown parameters of the model, *b* is an unknown discontinuity point of the function $f(\alpha,\beta,b,F)$.

We'll look for estimators of unknown parameters as the solution for optimization problems with a quadratic loss function:

$$\hat\theta\in\arg\min_{\theta}\sum_{k=1}^{n}\left(X_k-f(\theta,F_k)\right)^2 \quad and \quad \begin{bmatrix}\hat\alpha\\ \hat\beta\\ \hat b\end{bmatrix}\in\arg\min_{\alpha,\beta,b}\sum_{k=1}^{n}\left(X_k-f(\alpha,\beta,b,F_k)\right)^2. \tag{3}$$

The estimator $\hat\theta$ is called as the estimator of the least-squares method and is defined analytically [8]. To construct estimators in a piecewise linear model, iterative numerical methods such as quasi-Newton methods are used [8, 9].

As an example of the application of these models, we shall consider the task of reconstructing damage dependence on train delays using the U.S. railway transport data as of 2010 [13].

Let us consider the following set of factors:

F_1 is Fahrenheit temperature, F_2 is illumination, F_3 is weather conditions, F_4 is train speed at the moment of a transport accident, F_5 is train tonnage.

There are totally 1546 observations.

A linear regression model turns out to be inadequate and poorly corresponds to the observed values. For this model, the coefficient of determination $R^2\in[0;1]$ is set only to 0.612, while the value of the quadratic loss function (3) is 4.05 times greater than that for a non-linear model. The coefficient of determination indicates the extent to which the considered model "better explains" the value of *X* than a trivial model (i.e. *X* is independent of explaining variables and $\theta=\theta_0$). The closer R^2 is to 1, the greater superiority our regression has over trivial regression. And vice versa, if R^2 is close to zero, this means that a linear regression model poorly describes the variable *X* to be explained.

Let us present the results of estimation for a piecewise linear regression. We have applied the quasi-Newton method [11,12] to obtain the following results:

$$\hat{b} = 1006970, \quad \hat{\alpha} = \begin{bmatrix} 5080,6 \\ -4,236 \\ 1374,9 \\ 2963,0 \\ 5814,6 \\ 5,428 \end{bmatrix}, \quad \hat{\beta} = \begin{bmatrix} -466770 \\ -600,2 \\ 361254 \\ -48123 \\ 33206 \\ 26,95 \end{bmatrix}. \tag{4}$$

The estimation accuracy can be characterized by the value of the explained variance $D \in [0;1]$ and the coefficient of determination R^2, and the closer these values are to one, the better the chosen function *f(F)* explains the observations.

For our model, these characteristics are set as $D = 0,845$, $R^2 = 0,919$.

It is evident that the estimates of the regression parameters differ considerably for accidents with damage less than a million dollars and those with damage more than a million dollars.

Unfortunately, the available statistics does not contain detailed information on the importance of factors such as the technical condition of a train, railroad bed and facilities, staff skills, etc. Therefore, it does not seem possible to construct an adequate regression model of the influence of factors on the size of damage caused by train delays for the U.S. railway network.

3. Computational simulation of factors

In order to overcome the shortcomings of the available real statistics, we shall carry out computational simulation of factors affecting the occurrence of transport accidents, as well as the size of damage from these accidents. The simulation can be considered as a basis for future monitoring of the factors. After monitoring, the models of factors are refined, then using these models, the damage model is constructed that can later be used in two ways.

When the actual number of observations is not sufficient for construction of an adequate model of system operation, the "historical" simulation is carried out and the constructed model can be used to increase the sample size.

Construction of predictive values of damage for the current values of factors and predictive values of damage for the theoretical values of factors will allow us to analyze the general safety situation and to make recommendations on the most problematic factors affecting the level of traffic safety.

Factors that may cause a transport accident are simulated as random time functions, independent of each other. For each factor, we identify a critical level, above which the factor becomes a significant hazard. Random lengths of time during which its value above (below) the critical level (i.e. time of waiting for a factor to cross the critical level), have an exponential distribution with its own parameter for each of the factors. In case the value of one or several factors simultaneously exceeds the critical level within a certain time interval, this is supposed to be able to lead to a railway accident with some probability. Expectation time of an accident (also a random variable with an exponential distribution) is modeled for each such time interval. The distribution parameter depends on which factors have exceeded the critical value, in

other words, for each combination of factors there is its own occurrence rate of accidents. The accident is considered as occurred if expectation time was shorter than the time interval under consideration (i.e. a random variable hit the segment on which the values of considered factors exceeded the critical ones).

To simplify the simulation of factors, three models have been proposed.

1) Piecewise-linear «zigzag» function. Local maximums are the middle intervals of time, on which the value of factor exceeds the critical one; the local minimums are the middle intervals of time, on which the value of factor is below than the critical one.

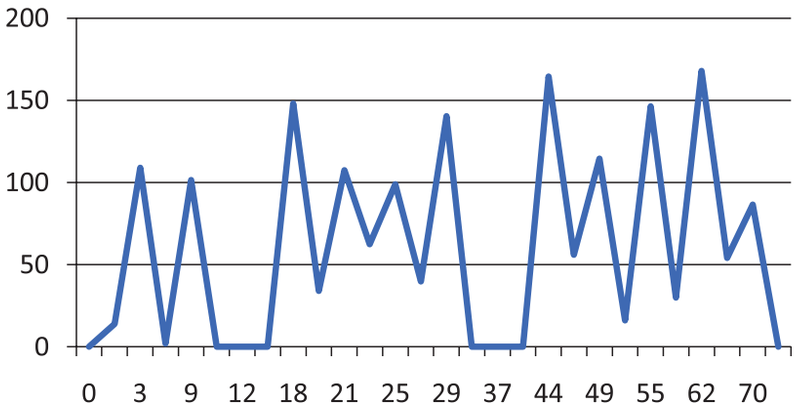

Fig.3. Factor of the first kind

Fig. 3 shows the function graph simulating the behavior of the first type factor. This kind of factor can be used to simulate weather conditions and other natural factors

2) Piecewise constant function. On segments for which the value of a factor is above the critical one, the value of a factor is a random variable identically distributed on the segment from the critical value up to 100. On segments for which the value of factor is below the critical one, the value of factor is a random variable identically distributed on the segment from 0 up to the critical value.

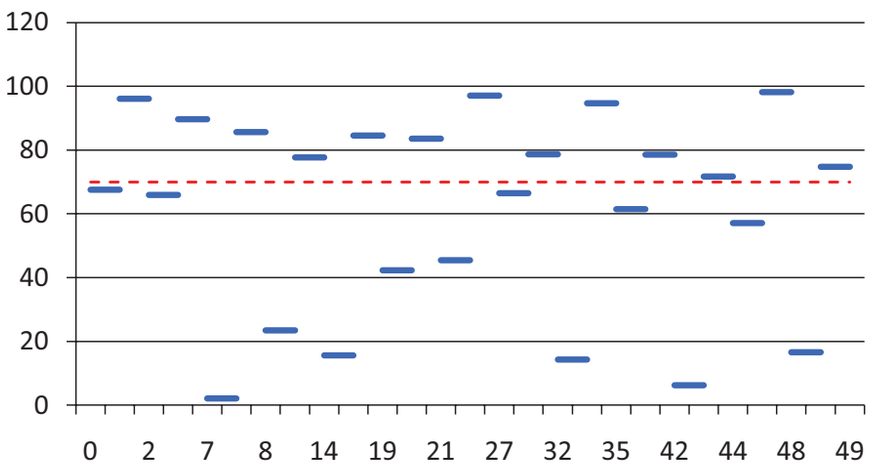

Fig.4. Factor of the second kind

This kind of function can characterize factors responsible for equipment condition. The value of a factor remains consistently below the critical level until some event leads to jump deterioration in its condition. After that the state of the equipment remains the same till the fault is eliminated.

3) The third type of factors are simulated in a similar way, however in this case the dependence is not piecewise constant but is piecewise linear. At the beginning of each interval, as in the previous case, the factor possesses the value identically distributed on the interval from the critical level to 100 or from 0 to the critical level, and then increases linearly with a random angular coefficient (a random value with identical distribution between 0 and π/6).

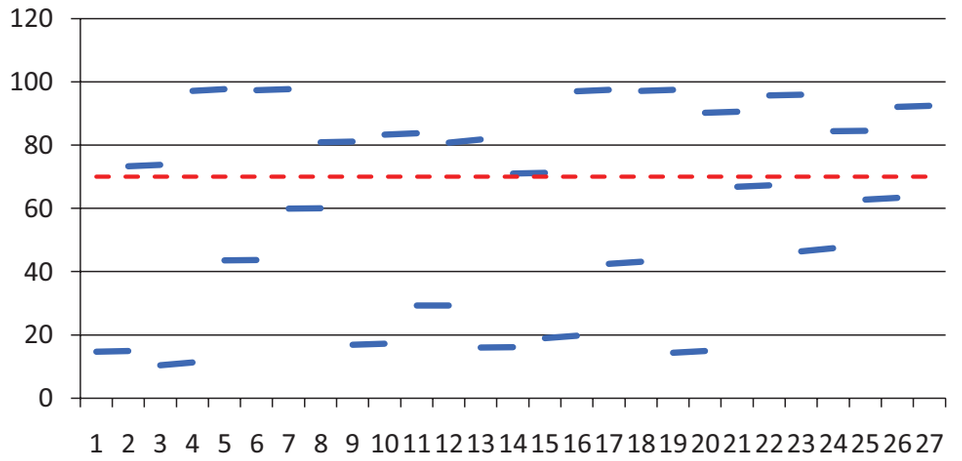

Fig.5. Factor of the third kind

This type of factor can be used for simulating the condition of rolling stock and railroad bed, as well as human factors. Discontinuities correspond to breakdowns and repairs, and the slow growth corresponds to gradual wear and tear of equipment.

4. Simulation of damage caused by a transport accident

Because of the lack of a large array of real data about accidents on Russian railways, the size of damage was calculated based on statistical data on railway accidents in the United States. The damage is modeled as a linear combination of factors with additive interference (noise). The noise is simulated based on an estimator of the distribution law of residuals for construction of the damage model using statistical data on accidents in the United States.

Denote a random variable whose distribution of density *f(x)* we want to estimate, as X. The estimator of the probability density is computed as follows

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right),$$

where X_i is observation of the random variable *X*, *n* is a sample size, *h* is a window width, *K (z)* is a kernel function. This estimator is often called a Rosenblatt-Parzen estimator [16, 17].

A kernel function can be selected at fairly wide ranges. This function is primarily responsible for smoothness and differentiability of the estimator $\hat{f}(x)$. In our case, the kernel function was chosen as a standard Gaussian density

$$K(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}.$$

Unlike selection of a kernel function, selection of the window width h is a key problem in the construction of the kernel estimator $\hat{f}(x)$. There are different ways to choose a window width, and they are widely covered in the literature [18]. In our case h is selected by a method of substitution.

The final distribution for noise was obtained as follows:

- We constructed the estimator $\hat{f}(x)$of damage density distribution according to the U.S. data;
- Then we simulated the sample implementation corresponding to a random value with density $\hat{f}(x)$;
- We constructed linear regression of damage from the values of factors (the implementation of factors was simulated according to the previously described algorithm);
- The residuals for the constructed regression were computed;
- Base on obtained residuals, we constructed a kernel estimator of noise density distribution $\hat{f}_{\varepsilon}(x)$;
- The final damage is modeled as a linear combination of factors with given coefficients plus implementation of a random quantity with density $\hat{f}_{\varepsilon}(x)$.

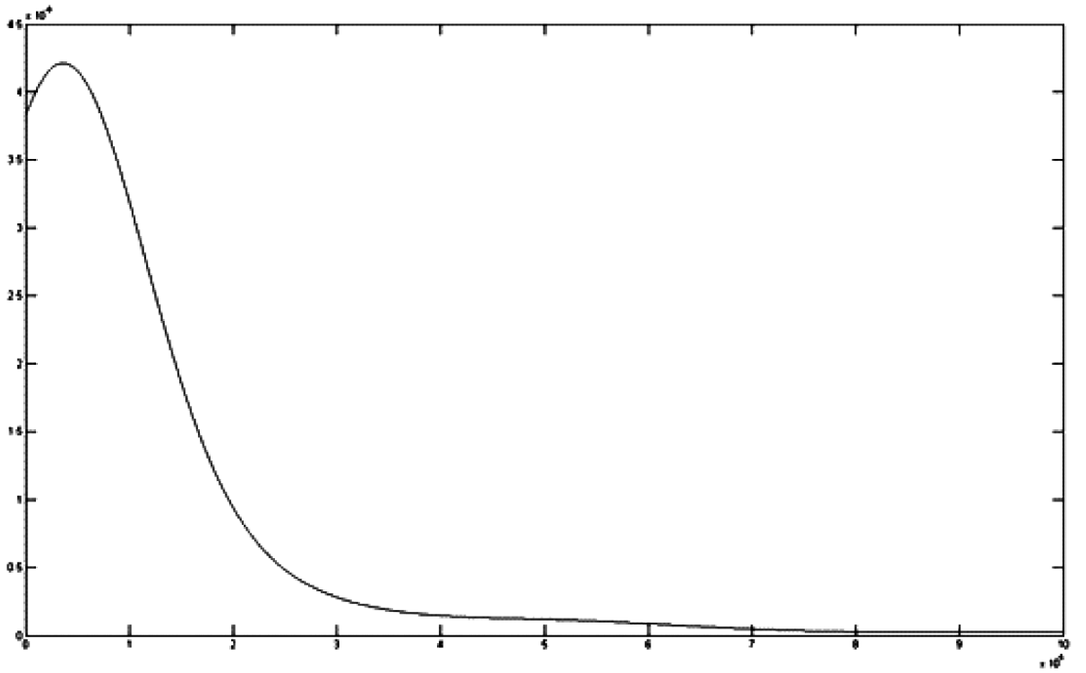

Fig.6. Kernel estimation of damage distribution according to the U.S. data

5. An example of simulation

Let us consider a linear model of observation

$$X_k = \theta_0 + \theta_1 F_1 + \theta_2 F_2 + \cdots + \theta_m F_m + \varepsilon_k \quad k = 1, \ldots, n,$$

where X_k is damage due to the k-th accident, F_i is a factor influencing the occurrence of a transport accident and the size of damage, θ_i is unknown parameters, ε_k is random noise.

Let us simulate 260 observations with parameter values of a linear model as

$$\theta = [0 \quad 300 \quad 350 \quad 400 \quad 2500 \quad 500 \quad 100]^T.$$

Estimation of linear regression parameters using the least squares method is

$$\hat{\theta} = \begin{bmatrix} 133331 & 0,027 & 0,036 & -0,091 & 0,243 & -0,011 & -0,041 \end{bmatrix}^T.$$

The coefficient of determination is $R^2 = 0{,}28$, which means poor accuracy of a linear model.

Now, based on the observations we'll construct a regression model with a discontinuity point:

$$\hat{b} = 237768, \quad \hat{\alpha} = \begin{bmatrix} -19400 \\ 284{,}2 \\ 254{,}8 \\ 338{,}2 \\ 2121{,}1 \\ 423{,}5 \\ 136{,}2 \end{bmatrix}, \quad \hat{\beta} = \begin{bmatrix} 1433733 \\ -1522 \\ -1092 \\ -3921 \\ -5496 \\ -733 \\ -3311 \end{bmatrix}$$

Characteristics of estimation accuracy are $D = 0{,}54$, $R^2 = 0{,}74$.

The sums of squared residuals for an ordinary linear model are 9 times greater than those for a model with discontinuity.

Now we shall model other 40 observations with the same values of coefficients as before and construct prediction of damage using a model with discontinuity.

In prediction using the model with a discontinuity point, it is necessary to decide, which part of the regression the observations relate to, as we do not have the value of damage and we cannot use the value $\hat{b}$ to select a vector $\hat{\alpha}$ or $\hat{\beta}$. To solve this problem, you can take advantage of the following algorithm.

1. Divide the available observations into two groups. The first group includes those sets of values of factors for which the damage Y_k is less than the value $\hat{b}$. The second group includes sets of values of factors for which the damage Y_k is greater than or equal to the value $\hat{b}$.

2. Calculate for each factor a sample mean on the sets from the one group $\overline{F}_i^I = \frac{1}{n_1}\sum_{k=1}^{n_1} F_{i,k}$, $i = 1, \ldots, m$, where n_1 is the number of observations in the first group, $F_{i,k}$ is the values of a factor with the number i in the k-th observation. Sample means $\overline{F}_i^{II}$, $i = 1, \ldots, m$ for the second group are computed in a similar way.

3. When constructing the prediction of damage for observations with the values of factors $F_{1,pr}, \ldots, F_{6,pr}$, estimators $\hat{\alpha}$ are chosen for observations belonging to the first group, and estimators $\hat{\beta}$ are chosen for those belonging to the second group. Observations belong to the first group if $\sum_{i=1}^{6}\left|F_{i,pr} - \overline{F}_i^I\right| \le \sum_{i=1}^{6}\left|F_{i,pr} - \overline{F}_i^{II}\right|$, and to the second group if $\sum_{i=1}^{6}\left|F_{i,pr} - \overline{F}_i^I\right| > \sum_{i=1}^{6}\left|F_{i,pr} - \overline{F}_i^{II}\right|$.

Let us compute a relative prediction error compared to the real values of damage that were also simulated. Denote the prediction as $\widetilde{X}$, then the value of the error for one observation X is defined by the formula $\Delta = \frac{\left|\widetilde{X} - X\right|}{X}$.

Let us now compute the mean relative error of prediction for 40 observations of factors $\overline{\Delta} = \frac{1}{40}\sum_{k=1}^{40}\Delta_k$. This value for the model with a point of discontinuity amounted to 0.15. Percentage of errors in selecting a group, where observations belong to, amounts to about 5%. The linear model loses much as regards a mean relative error. For this model it amounted to 0.52.

It should be noted that the accuracy of prediction values for the group with large sizes of damage is noticeably lower than the accuracy of prediction for observations of the first group. If you build predictions using only the estimators $\hat{\alpha}$, the relative error will make up 0.17. Further research may be focused on the fact that after separation of observations into two groups one should think over the way how to change the method of constructing a model for observations with large sizes of damage.

References

1. **Rosenberg E.N., Zamyshlyaev A.M., Proshin G.B.** Identification of hazard occurrence of railway accidents and events based on control of factors' state influencing their occurrence / / Dependability, 2009, No. 3 (31).p. 37-50.

2. **Zamyshlyaev A.M, Kan U.S., Kibzun A.I., Shubinsky I.B.** Statistical assessment of hazard occurrence of accidents on railways / / Dependability, 2012, No. 2 (41). p. 104-117.

3. **Goryanov A.V., Kan U.S., Platonov E.N.** Regression factor models for assessing risks on railways. Proceedings and plenary lectures of conference participants UKI'12 / Research publication. Electron. text data. -: M RAS' Institute of Control Sciences, 2012 - ISBN 978-5-91450-100-3 - p. 314-320.

4. **Xuedong Y.E., Radwan M. Abdel-Aty.** Characteristics of rear-end accidents at signalized intersections using multiple logistic regression model // Accident Analysis and Prevention. Vol. 37. 2005, P. 983–995.

5. **Joshua S., Garber N.** Estimating truck accident rate and involvement using linear and Poisson regression models // Transportation Planning and Technology. Vol. 15, 1990, P. 41–58.

6. **Jutaek O., Washington S.P., Doohee N.** Accident prediction model for railway-highway interfaces // Accident Analysis and Prevention. Vol. 38(2), 2006. P. 346–356.

7. **Loeb P.D., Clarke W.A.** The determinants of truck accidents // Transportation Research Part E. Vol. 43 (4), 2007. P. 442-452.

8. **Demidenko E.Z.** Linear and nonlinear regression. Moscow: Finance and statistics, 1981

9. **Netter J., Wasserman W., Kutner M.H.** Applied linear statistical model: Regression, analysis of variance, and experimental designs. Homewood, IL: Irwin. 1985.

1O. Railroad Safety Statistics 2010, Rail Equipment Accident Report 6180.54. U.S. Department of Transportation (http://safetydata.fra.dot.gov).

11. **Victor E. McGee, Willard T.C**. Piecewise Regression // Journal of the American Statistical Association. Vol. 65. 1970, № 331. P. 1109–1124.

12. **Martinez-Beneito M.A., García-Donato G., Salmerón D.** A Bayesian Joinpoint regression model with an unknown number of break-points // The Annals of Applied Statistics. Vol. 5, 2011, № 3. P. 2150–2168.

13. **Rosenblatt M.** Remarks on Some Nonparametric Estimates of a Density Function, The Annals of Mathematical Statistics. Volume 27, № 3 (1956). P. 832-837.

14. **Parzen E.** On Estimation of a Probability Density Function and Mode, The Annals of Mathematical Statistics. Volume 33, No. 3 (1962). P. 1065-1076.

15. **Racine J.** Nonparametric econometrics: Introductory Course, Quantile, No.4. P. 7-56.