Modeling of accident rates involving trucks in order to improve road safety in the Russian Federation

I E Ilina*

Penza State University of Architecture and Construction, Penza, Russia

Abstract. A statistical analysis of accidents on road transport in the Russian Federation involving trucks was carried out. We have studied the legal acts regulating road safety in General and the implementation of cargo transport in particular. The object of the study is the state of road safety. The possibility of using regression for modeling and analyzing the relationship between the number of cargo vehicles and the length of roads is considered. These indicators are taken into account for 85 regions of the Russian Federation. The result of the joint influence of independent variables on the number of road accidents is obtained. A multiple regression equation is given to explain the model parameters. The constructed model is useful, and the selected two variables "length of highways, km"," number of cargo vehicles, units " allow predicting the level of accidents on road transport.

1 Introduction

In the Russian Federation, about 80% of the total volume of cargo transported by all modes of transport (rail, air, sea) is carried out by road. Most of the goods cannot be transported to the consumer without the participation of road transport. The growth of production volumes, changes in the specifics of products, the need and development of the economy reflect the demand for road freight transport. The annual (2-4%) growth of the freight rolling stock fleet determines the need for these services. [1]

The main industries that use road freight transport are related to retail. Road freight transport is economically feasible for use in short-and medium-distance transportation [2-7].

Every year, about 7% of accidents occur due to violations of traffic rules by truck drivers. The severity of accidents involving trucks is the highest of all the committed accidents and is 10.1 deaths per 100 accidents. Reducing accident rates is a priority task of the state.

The authors' works [8-14] are devoted to assessing the impact of such factors as the number of people, the number of traffic violations, the parameters of automatic detection systems, the commissioning of new roads, etc. on the stability and safety of road traffic.

* Corresponding author: iie.1978@yandex.ru

2 Methods and materials

In this paper, preference is given to the study of multi-factor regression, which involves establishing a linear relationship between a set of input independent and one output dependent variables [15].

The output dependent variable Y is "number of road accidents, units", the input independent variables X_1 are "total length of highways, km" and X_2"number of cargo vehicles, units".

The initial data for the study are: statistical data on accidents in road transport, the number of registered trucks for 2019 in accordance with the data of the State road safety Inspectorate, the length of public roads in the subjects of the Russian Federation according to the Federal state statistics service for 2019. The research was conducted in the context of 85 subjects of the Russian Federation.

3 Results and discussion

Thus, to get the model, the number of observations n is 85. The number of independent variables m in the model is 2. The number of regressors taking into account the unit vector will be equal to the number of unknown coefficients.

Using the least squares method and matrix approach, the regression coefficients in the equation are determined.

The vector of regression coefficient estimates will take the form:

$$Y(X) = \begin{pmatrix} 0.0396 & -1.0E\text{-}6 & 0 \\ -1.0E\text{-}6 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 168099 \\ 3994995689.9 \\ 1947486016 \end{pmatrix} = \begin{pmatrix} 416.542 \\ 0.00921 \\ 0.192 \end{pmatrix}$$

Using a matrix scatter plot, which is a top view of the regression plane and two views along the plane, we can say that the spread of points relative to the regression plane is in the range +/-3.

A matrix of Y and $X$and a transposed $X^T X$ matrix are made up. The resulting matrix has the following correspondence:

$\sum n$	$\sum y$	$\sum x_1$	$\sum x_2$
$\sum y$	$\sum y^2$	$\sum x_1 y$	$\sum x_2 y$
$\sum x_1$	$\sum y x_1$	$\sum x_1^2$	$\sum x_2 x_1$
$\sum x_2$	$\sum y x_2$	$\sum x_1 x_2$	$\sum x_2^2$

Based on the available data, the paired correlation coefficients will take the values:

$$r_{yx1} = 0.5717\ ,$$
$$r_{yx2} = 0.7946,$$
$$r_{x1x2} = 0.6746.$$

The matrix of paired correlation coefficients R will take the form:

-	y	x_1	x_2
y	1	0.5717	0.7946
x_1	0.5717	1	0.6746
x_2	0.7946	0.6746	1

The analysis of multicollinearity of factors Y, X_1 and X_2showed that the results of multiple regression are reliable. In the case under study, with the available initial data, all paired correlation coefficients are$|r| < 0.7$, which indicates that there is no multicollinearity of the factors:

$$r_{yx1/x2} = 0.0797,$$
$$r_{yx2/x1} = 0.6750,$$
$$r_{x1x2/y} = 0.4420.$$

Matrix analysis allows you to select factor features that can be included in the multiple correlation model.

Since the results obtained are in the range of $0.3 \leq |r| \leq 0.7$, the relationship between the factors is significant.

The factor x_2 "number of cargo vehicles, units" ($r = 0.7946$) has the greatest influence on the result attribute. This means that it will be the first to enter the regression equation when building the model.

A more objective description of the tightness of the relationship is given by partial correlation coefficients that measure the influence of factor xi on the result at the same level of other factors.

The multiple correlation index evaluates the tightness of the joint influence of factors on the result. If the Rvalue is close to 1, the regression equation better describes the actual data and factors have a stronger influence on the result.

The multiple correlation coefficient (1) can be determined using a matrix of paired correlation coefficients:

$$R = \sqrt{1 - \frac{\Delta_r}{\Delta_{r11}}} \quad (1)$$

where Δ_r – is the determinant of the matrix of paired correlation coefficients;Δ_{r11}– determinant of the interfactor correlation matrix.

$$\Delta_r = \begin{vmatrix} 1 & 0.572 & 0.795 \\ 0.572 & 1 & 0.675 \\ 0.795 & 0.675 & 1 \end{vmatrix} = 0.2$$

$$\Delta_{r11} = \begin{vmatrix} 1 & 0.675 \\ 0.675 & 1 \end{vmatrix} = 0.545$$

The multiple correlation coefficients$R$will be 0.7961, i.e. the relationship between feature Y and factors X_i is strong.

A more objective estimate is the adjusted coefficient of determination (2):

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-m-1} = 0.625 \quad (2)$$

We evaluate the significance of the multiple regression equation. Let's test the hypothesis of General significance, i.e. the hypothesis that all regression coefficients are simultaneously equal to zero for explanatory variables:

$$H_0: R^2 = 0; \beta_1 = \beta_2 = \cdots = \beta_m = 0$$

$$H_1: R^2 \neq 0$$

We can test this hypothesis using the F-statistics of the Fischer distribution (3).

If $F < F_{кр} = F_{a;n-m-1}$,then there is no reason to reject the hypothesis H_0

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m} = 70.935 \quad (3)$$

Since the actual value of $F > F_{кр} = 3.07$for $F_{кр}$ (2; 82), the coefficient of determination is statistically significant and the regression equation is statistically reliable (i.e., the b_i coefficients are jointly significant).

The need to assess the significance of the additional inclusion of a factor is due to the fact that not every factor included in the model can significantly increase the proportion of the explained variation of the effective feature. This may be due to the sequence of input factors, since there is a correlation between the factors themselves.

A measure of evaluating the significance of improving the quality of the model, after the factor $x_j$is included in it, is a special F – criterion-F_{xj} (4):

$$F_{xj} = \frac{R^2 - R^2(x_1, x_n)}{1-R^2}(n - m - 1) \quad (4)$$

where $R^2 - R^2(x_1, x_n)$ – increase in the proportion of variation Y due to an additional factor included in the model x_j.

If the observed value of F_{xj} is greater than $F_{кр}$, then the additional introduction of the x_j factor into the model is statistically justified. A particular F – criterionevaluates the significance of the regression coefficients b_j.

We evaluate the feasibility of including $x_1$factors in the regression model after the introduction of x_j (F_{x1})using a particular F – criterion.

The observed value of the partial F – criterion is $F_{x1} = 0.519$.

$$R^2(x_2, x_n) = r^2(x_2) = 0.7946^2 = 0.631 \quad (5)$$

Comparing the observed value of the partial F – criterion$F_{x1}$with the critical $F_{x1} < F$, we determine that it is not advisable to include factor x_1 in the model after the introduction of factor x_2.

We evaluate the feasibility of including x_2 factors in the regression model after the introduction of x_j (F_{x2}) using a particular F – criterion.

The observed value of the private F – criteria will be $F_{x2} = 68.687$.

$$R^2(x_1, x_n) = r^2(x_1) = 0.5717^2 = 0.327 \quad (6)$$

Let's compare the observed value of a particular F -criterion with the critical one: $F_{x2} > F$, and determine whether it is advisable to include factor x_2 in the model after introducing factor x_1.

As a result of calculations, the multiple regression equation was obtained (7):

$$Y = 416{,}542 + 0{,}00921X_1 + 0{,}1919X_2 \quad (7)$$

4 Conclusion

The model parameters can be interpreted as follows: an increase in X_1"length of highways, km." by 1 unit leads to an increase in Y"number of accidents, units." on average by 0.00921 units; an increase in X_2" number of cargo vehicles, units." by 1 unit. leads to an increase in Yon average by 0.192 units. By the maximum coefficient $\beta_2 = 0.75$, we conclude that the factor X_2has the greatest influence on the result Y. The statistical significance of the equation was verified using the coefficient of determination and the Fisher criterion. It was found that in the studied situation, 63.37% of the total y variability is due to changes in the x_j factors.

Thus, the constructed model is useful, and the selected two variables "length of highways, km"," number of cargo vehicles, units " allow predicting the level of accidents on road transport.

The work was performed under the RGNF grant 21-19-00240.

References

1. A.E. Gorev Cargo transportation: textbook for students. institutions of higher education (Academy, Moscow, 2013).
2. E. Vitvitskii, M. Simul, S. Porkhacheva, Transportation Research Procedure (2017).
3. S. Voytenkov, E. Vitvitskiy, Transportation Research Procedure **36,** 786 (2018).
4. E.E. Vitvitsky, E.S. Fedoseenkova, Bulletin of science and education of the North-West of Russia **4(4)**, 82-88 (2018).
5. E.E. Vitvitsky, E.S. Fedoseenkova, IOP Conference Series: Earth and Environmental Science International **072013** (2018).
6. S. Voytenkov, E. Vitvitskiy, Transportation Research Procedia (2018).
7. E.R. Aitbagina, E.E. Vitvitsky, Advances in Intelligent Systems and Computing **1116,** 968-974 (2020).
8. A. Marusin, I. Danilov, Transportation Research Procedia **36,** 500- 506 (2018). https://doi.org/10.1016/j.trpro.2018.12.136.
9. P. Kravchenko, E. Oleshchenko, Transportation Research Procedia **20,** 367–372 (2017). https://doi.org/10.1016/j.trpro.2017.01.051
10. A.V. Gasnikov, S.L. Klenov, E.A. Nurminski, Y.A. Kholodov, N.B. Shamray Introduction to mathematical modeling of traffic flows (MTSNMO, Moscow, 2013).
11. V.R. Vuchic Transportation for Livable Cities (1999).
12. I.A. Garkina, A.M. Danilov, Vestnik of PGWS: construction, science and education **2(9),** 98-101 (2019).
13. I.A. Garkina, A.M. Danilov, Bulletin of the Tajik Technical University **4(24),** 75-80 (2013).

14. A.M. Danilov, I.A. Garkina, IOP Conf. Ser.: Mater. Sci. Eng. **449,** 012002 (2018).
15. V.B. Shashkov Applied regression analysis. Multivariate regression: Textbook (GOU VPO OSU, Orenburg, 2003).