FEDERAL STATE AUTONOMOUS EDUCATIONAL INSTITUTION

OF HIGHER EDUCATION

ITMO UNIVERSITY

Report on learning practice # 2

Analysis of multivariate random variables

Performed by:

Pavel Belenko, J4132c

Yuri Umnov, J4132c

Saint-Petersburg

2021

**Description:**

The data was taken from the [*database*](https://www.kaggle.com/rustydigg918/exploratory-data-analysis-on-car-sales-data/data), describing some physical parameters, costs, and sales volumes of more than 150 models of cars from 30 different manufacturers.

Vehicle type, engine size, horsepower, curb weight, fuel reserve and power factor were taken as a *subsample* of variables for this work. The Price\_in\_thousands variable was chosen as a *target*.

**The main steps:**

1. Изображение выглядит как электроника

   Автоматически созданное описаниеPlotting a non-parametric estimation of PDF in form of a histogram and Kernel density function for MRV

*Figure 1 – Pairwise relations*

1. Изображение выглядит как стол

   Автоматически созданное описаниеEstimation of multivariate mathematical expectation and variance.

*Table 1 - Estimation of multivariate mathematical expectation and variance*

1. Изображение выглядит как доска

   Автоматически созданное описание Non-parametric estimation of conditional distributions, mathematical expectations, and variances

*Figure 2 – Pairwise relations in two ways (blue – passneger cars, orange – trucks)*

*Изображение выглядит как окно, другой

Автоматически созданное описаниеFigure 3 – Non-parametric estimation of conditional distributions*

*Изображение выглядит как стол

Автоматически созданное описание*

*Table 2 - Estimation of mathematical expectations, and variances*

1. Изображение выглядит как стол

   Автоматически созданное описаниеEstimation of pair correlation coefficients, confidence intervals for them and significance levels

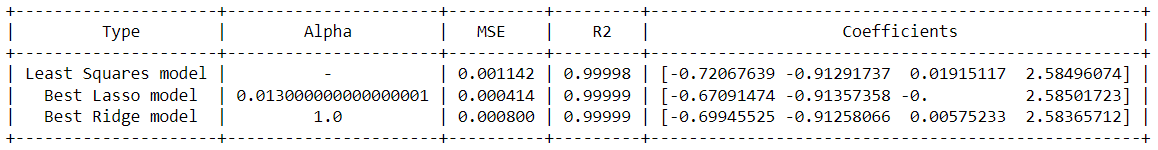
*Table 3 – Estimation of pair correlation coefficients, confidence intervals for them and significance levels*

1. Task formulation for regression, multivariate correlation*.*

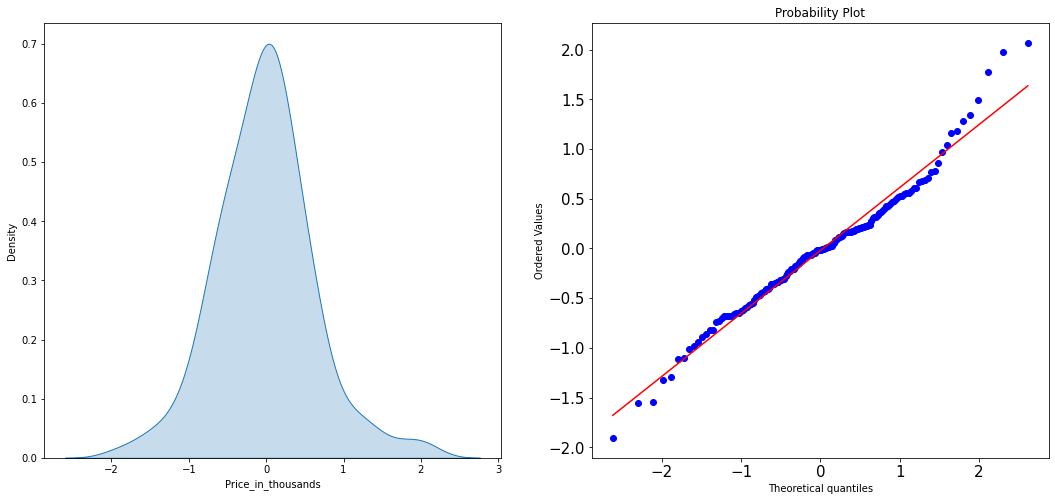
*Figure 4 – Multivariate correlation matrix*

The formulation of the regression: create a model that predicts the cost of a certain model of a car, according to its given physical characteristics.

1. Regression model, multicollinearity, and regularization.

At this step, three linear models were tested, shown in Table number 4. The result was acceptable, regularization was not required

*Table 4 - metrics for assessing the quality of training*

1. Analyze the quality of regression model.

*Figure 5 - Visualization of residuals distribution and results of validation of distribution using quantile biplot for residuals.*

According to the Shapiro-Wilk and Andersen-Darling tests, the null hypothesis that the subsample was taken from a population that corresponds to a normal distribution was rejected.

**Source code:**

<https://github.com/belpablo/MMMSA/blob/main/Lab_2_redo.ipynb>