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*

It can be concluded that, firstly, passenger and trucks on average have the same cost, power factor and horsepower, but the variance of these values in passenger cars is much greater. Secondly, the weight of the car, the engine volume, and its capacity of trucks (as well as the variance of these values) is greater than that of passenger cars.

1. Изображение выглядит как стол

   Автоматически созданное описаниеEstimation of pair correlation coefficients, confidence intervals for them and significance levels (target - Price\_in\_thousands).

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

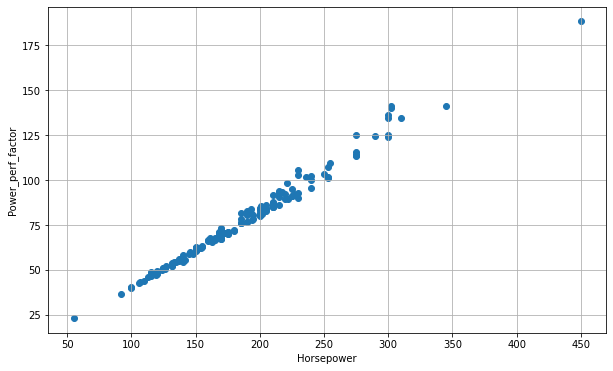
1. Task formulation for regression, multivariate correlation*.*

Task: to create a model that predicts the cost of a certain model of a car (Price\_in\_thousands), according to its given physical characteristics.

*Figure 4 – Multivariate correlation matrix*

The following conclusions can be drawn from the multidimensional correlation matrix (Figure 4): Power\_perf\_factor and Horsepower correlate well with the target variable. Average correlate - Engine\_size and Curb\_weight. The Fuel\_capacity variable is weakly correlated, so we will not consider it in further analysis.

It is also worth mentioning, that the correlation coefficient between the variables Power\_perf\_factor and Horsepower is extremely high. This feature can have a bad effect on the operation of the model.

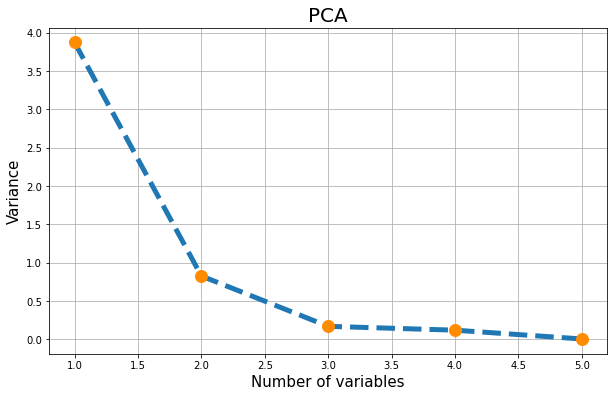


*Figure 5 – Almost linear correlation between Horsepower and Power\_perf\_factor variables*

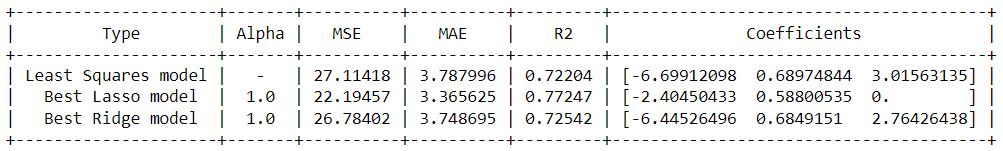
1. Regression model, multicollinearity, and regularization.

In the process of constructing a regression model, a problem was identified related to the high correlation of the two variables mentioned in the last paragraph. The model had a negligible coefficient for one independent variable, but a significant R2 for a simple regression model using the same independent variable.

To generate the best estimates of multiple regression coefficients, the underlying data should be subject to the same assumptions as a simple regression plus one additional one — the absence of complete multicollinearity. This means that independent variables should not be strictly linearly correlated with each other. Therefore, the “Horsepower” variable was not used in the final model.

The PCA algorithm was also used to reduce the dimension of objects. When the number of components increases from 1 to 3, the decrease in variance is significant, and a larger number of variables is not descriptive. Therefore, the number of selected variables for the regression task was chosen to be 3:

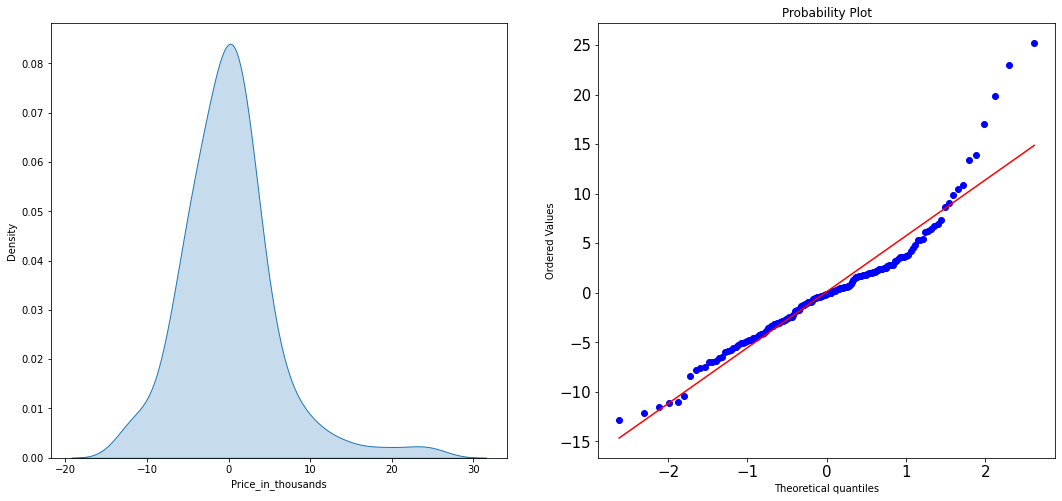
*Figure 6 – PCA analysis*

Thus that, the following variables were taken to build the model: Engine\_size, Power\_perf\_factor, Curb\_weight. Next, three linear models were tested, shown in following table:

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

In the final version, the Lasso regression model was used

1. Analyze the quality of regression model.



*Figure 7 - Visualization of residuals distribution (left plot) and the results of validation of distribution using quantile biplot for residuals (right plot).*

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>