*IRSTI 28.17.27*

*Zh. Rayev[1], A. Aipenova[2], T. Suleizhan[3], D. Zhumabek[4], A. Duman[5]*
[1,2,3,4,5]Suleyman Demirel University, Kaskelen, Kazakhstan

# MULTIVARIATE REGRESSION ANALYSIS AND MODELLING ON CARS DATASET

**Abstract**. The results of the work are based on the construction of a mathematical model for determining unknown parameters using multivariate regression analysis. Structured data are given for the derivation and elimination of significant factors and coefficients. Also, machine learning simple regression models are used for modelling. The results have been evaluated and shown for comparative purposes.

**Keywords**: multivariate regression analysis, modelling, machine learning.

***

**Аннотация**. Результаты работы основаны на построении математической модели для определения неизвестных параметров с использованием многомерного регрессионного анализа. Структурированные данные приведены для выведения и исключения значимых факторов и коэффициентов. Также для моделирования используются простые регрессионные модели машинного обучения. Результаты были оценены и показаны для сравнительных целей.

**Ключевые слова:** многомерный регрессионный анализ, моделирование, машинное обучение.

***

**Аңдатпа.** Жұмыстың нәтижелері көпөлшемді регрессиялық талдауды қолдану арқылы белгісіз параметрлерін анықтау үшін математикалық модельдің құрылысына негізделген. Құрылымдық деректер маңызды факторлармен коэффициенттерді көрсету үшін беріледі. Сондай-ақ модельдеу үшін қарапайым регрессиялық үлгілерді машиналық оқыту модельдері қолданылады. Нәтижелер салыстырмалы мақсаттар үшін бағаланып көрсетілді.

**Түйін сөздер:** көп өлшемді регрессиялық талдау, модельдеу, машиналық оқыту.

*Introduction*

The considered structural data are taken from Kazakhstani sources and not all parameters have beenfully described. During analysis process [3], implicit data were identified and replaced with average values.

| | names | year | volume | fuel_type | transmission | body | drive | mileage | wheel | color | city | customs clearance | price |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Toyota Camry | 2018 | 2.5 | бензин | автомат | седан | передний привод | 1.0 | слева | серебристый металлик | Алматы | Да | 13800000 |
| 1 | Mercedes-Benz G 63 AMG | 2014 | 5.5 | бензин | автомат | внедорожник | NaN | 10300.0 | слева | черный | Алматы | Да | 68000000 |
| 2 | Toyota Corolla | 2013 | 1.6 | бензин | автомат | седан | передний привод | 39258.0 | слева | серый металлик | Алматы | Да | 5499999 |
| 3 | Mercedes-Benz E 220 | 1993 | 2.2 | газ-бензин | автомат | седан | задний привод | NaN | слева | серый металлик | Шымкент | Да | 1850000 |
| 4 | Mitsubishi Outlander | 2013 | 2.4 | бензин | автомат | кроссовер | полный привод | 60000.0 | слева | серый | Актобе | Да | 6100000 |

Table 1.1

The first five rows of dataframehas beenshown in Table 1.1.

```
RangeIndex: 4997 entries, 0 to 4996
Data columns (total 13 columns):
names                4997 non-null object
year                 4997 non-null int64
volume               4995 non-null float64
fuel_type            4995 non-null object
transmission         4997 non-null object
body                 4997 non-null object
drive                4405 non-null object
mileage              3276 non-null float64
wheel                4952 non-null object
color                4596 non-null object
city                 4997 non-null object
customs clearance    4997 non-null object
price                4997 non-null int64
dtypes: float64(2), int64(2), object(9)
```

Table 1.2

General information about dataframecontent (Table 1.2) is quite clear and enough to model and analyze.

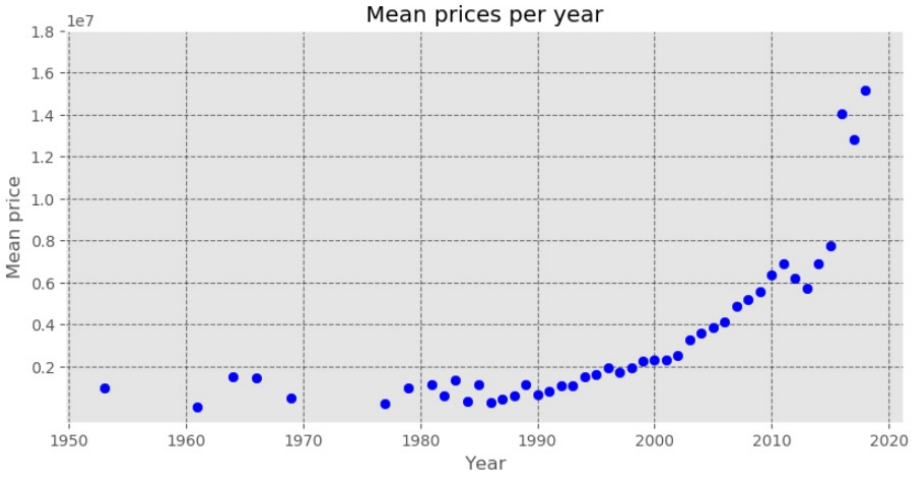


Table 1.3

Mean prices of cars after 2016 has been looked overwhelming (Table 1.3), but it has been up to individual production models of cars as 'Rolls Royce' or so on. Those data has not been counted as outliers.

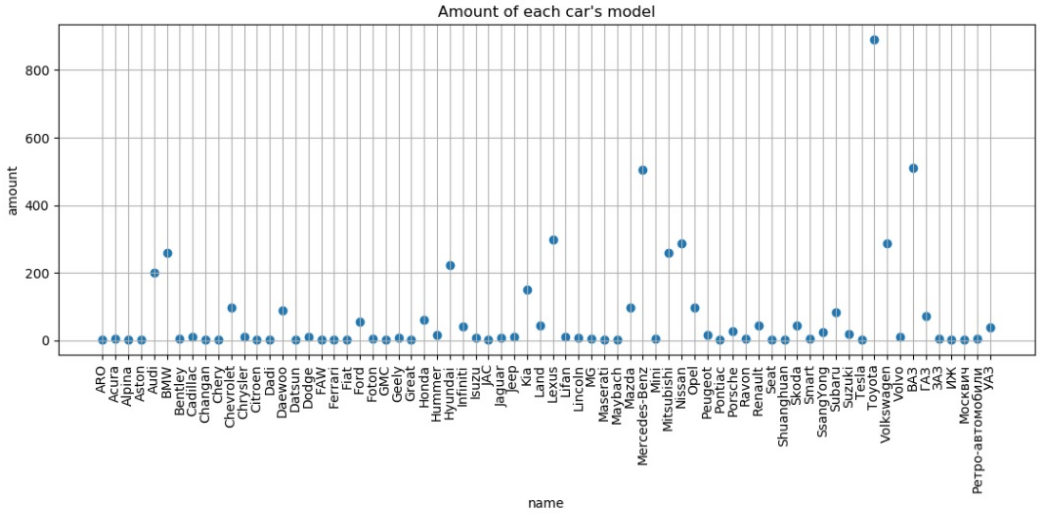


Table 1.4

The minimum point of amount was 1 and it has been grown in certain models of cars which are popular and high sold in Kazakhstan (Table 1.4 and Table 1.5).

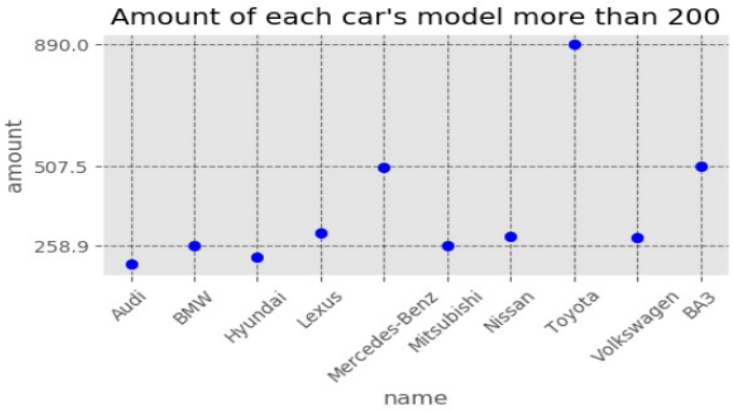


Table 1.5

The total number of unique city is 137 and there were a lot of unimportant suburbs and villages. Table 1.6 has shown the important places in general which amount is more than 30.

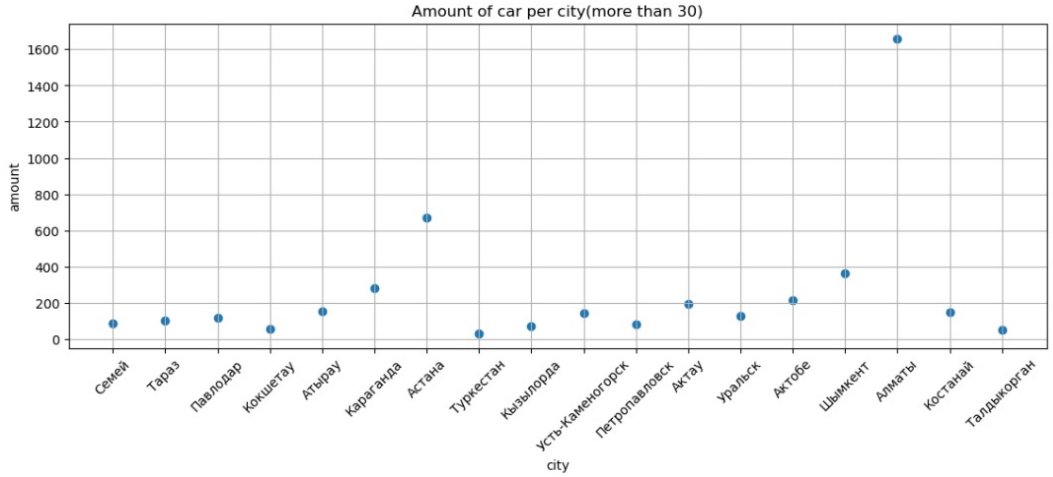


Table 1.6

Table 1.3 and Table 1.4 have fully described the unbalance of data [7]. In the next step we have cut the data by triangle method.

*Modelling*

In modelling process prediction has been resolved in price column. In this case multiple learning algorithms methods, so called ensemble learning methods, have been used. The main principle to use ensembles was to have more flexibility in the function they can represent [1]. This flexibility could enable them to over-fit the training data more than a single model would, but the purpose of modelling is to show that in practice, some ensemble techniques tend to reduce problems related to over-fitting of the training data [5].

*Random forest*

The random forest prediction is the unweighted average over the collection:

$$h(x) = (1/\mathrm{K})\sum_{k=1}^{K} h(\mathrm{x};\theta_k)$$

where x-observed covariate vector of length p, as k approaches to infinity the Law of Large Numbers ensures

$$E_{X,Y}(\mathrm{Y}-\bar{h}(\mathrm{X}))^2 \to E_{X,Y}(\mathrm{Y}-\mathrm{E}_{\theta}\mathrm{h}(\mathrm{X};\theta))^2$$

where the quantity on the right is the generalization error for the random forest [2].

*Adaboost*

AdaBoost (Freund &Schapire, 1997; Schapire& Singer, 1999) is based on the exponential loss

$$\sum_{k=1}^{K} \exp(-\mathrm{y}_k \mathrm{f}_{\lambda}(\mathrm{x}_k))$$

where this equation has upper-bounded

$$\sum_{k=1}^{K} [[\mathrm{y}_k \mathrm{f}_{\lambda}(\mathrm{x}_k) \le 0]]$$

The natural goal is to try to match the sign of function f to y, that is, to attempt to minimize it [6].

*Solution*

Data has been preprocessed to model and has given these correlation results:

```
Kendall rank correlation coefficient(Close to 1 is better): 0.7174682188508716
Spearmans rank corr coefficient(Close to 1 is better): 0.8813390018550825
```

Usually coefficients must be more than 0.90 to make better prediction from model. It has shown that some parameters was not correlated good.

*Random forest*

```
randomf_cv.best_score_ #Result is not good
```

```
0.7082047522846117
```

```
y_pred2 = randomf_cv.best_estimator_.predict(X_test)
```

```
print(np.sqrt(mean_squared_error(randomf_cv.best_estimator_.predict(X_train), y_train)))
print(np.sqrt(mean_squared_error(y_pred2, y_test)))
```

```
1670808.857660936
3018682.8392637796
```

*Adaboost*

As model prediction in [4], a weak prediction model for boosting Decision Tree Regression model has been taken.

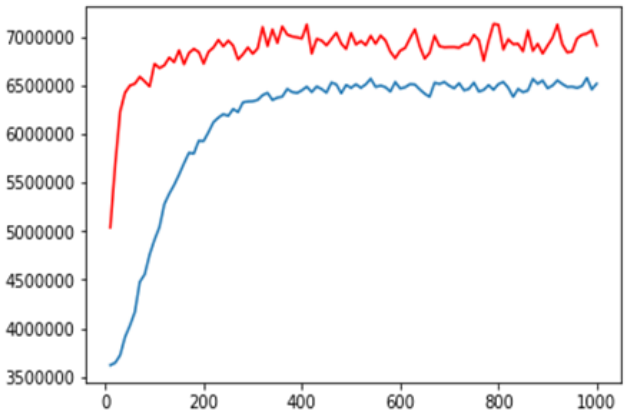


Table 3.1. Max depth is 1

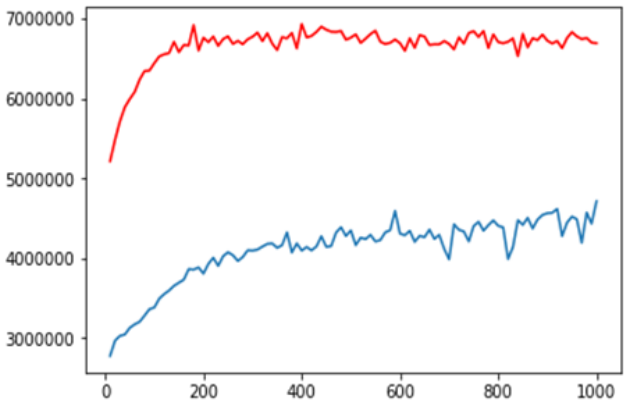


Table 3.2. Max depth is 2

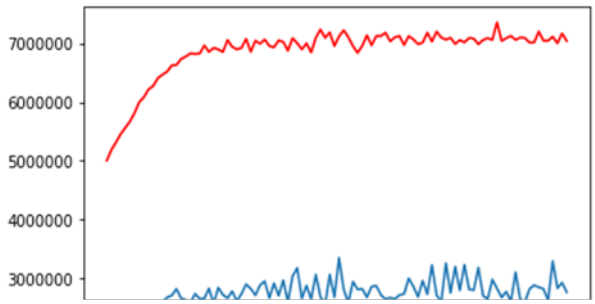


Table 3.3. Max depth is 3

Last results of Adaboost:

```
print('MAE test with GSV: ', mean_absolute_error(y_pred_test_best, y_test))
print('MAE train with GSV: ', mean_absolute_error(y_pred_train_best, y_train))
```

```
MAE test with GSV:  2743987.1198835857
MAE train with GSV:  2829078.748196364
```

*Conclusion*

By the method of reducing the number of outliers and taking the data balanced (outlier detection method) correlation coefficients has shown the better result as usual, but it has affected to the result of first model predictions.Adaboost model has given the better results as random forest, but also we can use another weak models as a base.

## References

1 Wei, J.L., Lin, D.Y., Weissfeld, L., Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. *Journal of the American Statistical Association*, (1995): pp. 5-6.
2 Segal, Mark, R. Machine Learning Benchmarks and Random Forest Regression. Division of Biostatistics, University of California, San Francisco, 2003. – pp. 4-6.
3 Breiman, L., Statistical Modeling: The Two Cultures. University of California, 2003. – pp. 207-208.
4 Liaw, A., Wiener, M. *Classification and Regression by random Forest*, 2002. URL: http://www.stat.berkeley.edu/users/breiman/.

5 Quinlan, J.R., Bagging, Boosting and C4.5, University of Sydney, Sydney, 2006. – pp. 2-3.
6 Solomatine, D.P., Shrestha, D.L., AdaBoost. RT: a boosting algorithm for regression problems. *IEEE International Joint Conference on Neural Networks,* (2004): pp. 2-3.
7 Riccardi, A., Fernández-Navarro, F., Carloni, S., Cost-Sensitive AdaBoost Algorithm for Ordinal Regression Based on Extreme Learning Machine. *IEEE Transactions on Cybernetics,* (2014): pp. 4-6.