Nonparametric Methods in Statistics

Bartłomiej Jamiołkowski

Comparison of the prognostic properties of the model estimated by OLS and regression nonparametric

1. Introduction

The topic of the project is to compare the prognostic properties of the model estimated with the OLS and non-parametric regression. The study uses available data of poviats from the Local Data Bank from 2019. The chosen dependent variable is the number of new registered entities from the private sector (variable entities). The explanatory variables in the project are:

• roads - index of poviat roads with a hard surface per 100 $km2$;

• population - population index at $1km2$;

• crime - the number of identified crimes.

The prognostic properties are compared using the k-fold cross-validation. The data set consists of 380 rows and 5 columns: the name of the poviat, the dependent variable and 3 potential explanatory variables. The significance level of 0.05 applies to all tests performed.

2. Data description

2.1 Number of new registered entities from the private sector - explained variable

In the first part of the variable analysis, the presence of outliers is checked using a box plot.

The presented chart shows that there are no outliers in the data. The visualization indicates a right-hand skew of the distribution of the number of new registered entities from the private

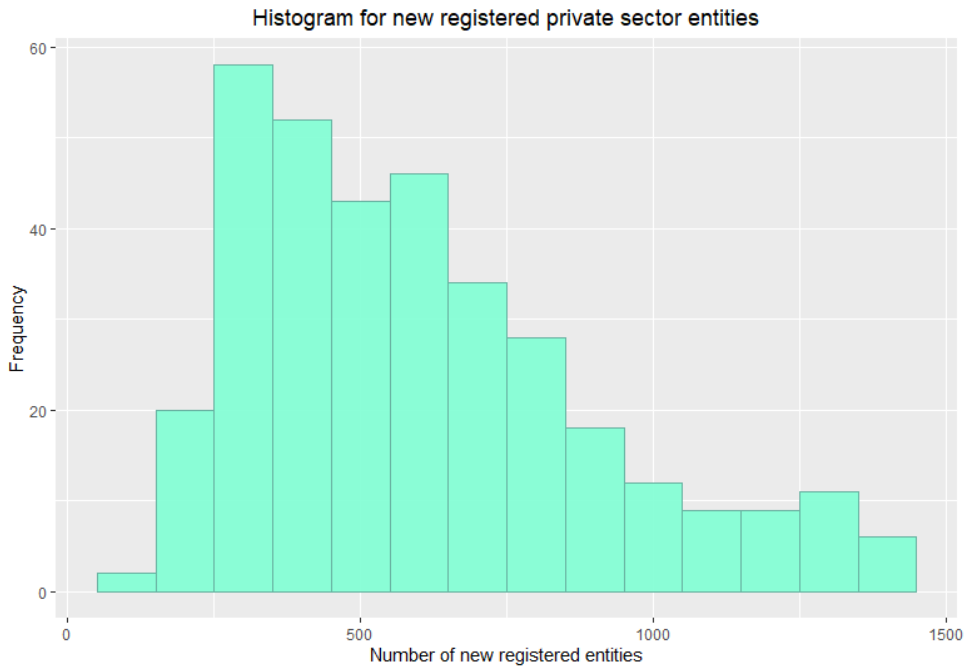
sector. This means that there are more poviats in which the number of new registered business entities is higher than the average for poviats in Poland.

The next step in the analysis is the calculation of the basic variable statistics.

entities	
Min	124.0
Q1	335.8
Median	472.0
Average	529.2
Q3	662.2
Max	1384.0
Skewness	1.1012
Kurtosis	1.0358

Among the presented values, one can observe a kurtosis greater than 0, which means that there is a leptokurtic distribution.

In the further part of the variable analysis, the normality of the distribution is examined. The test uses a histogram and a Shapiro-Wilk normality test.



The following hypotheses are tested in the Shapiro-Wilk test:

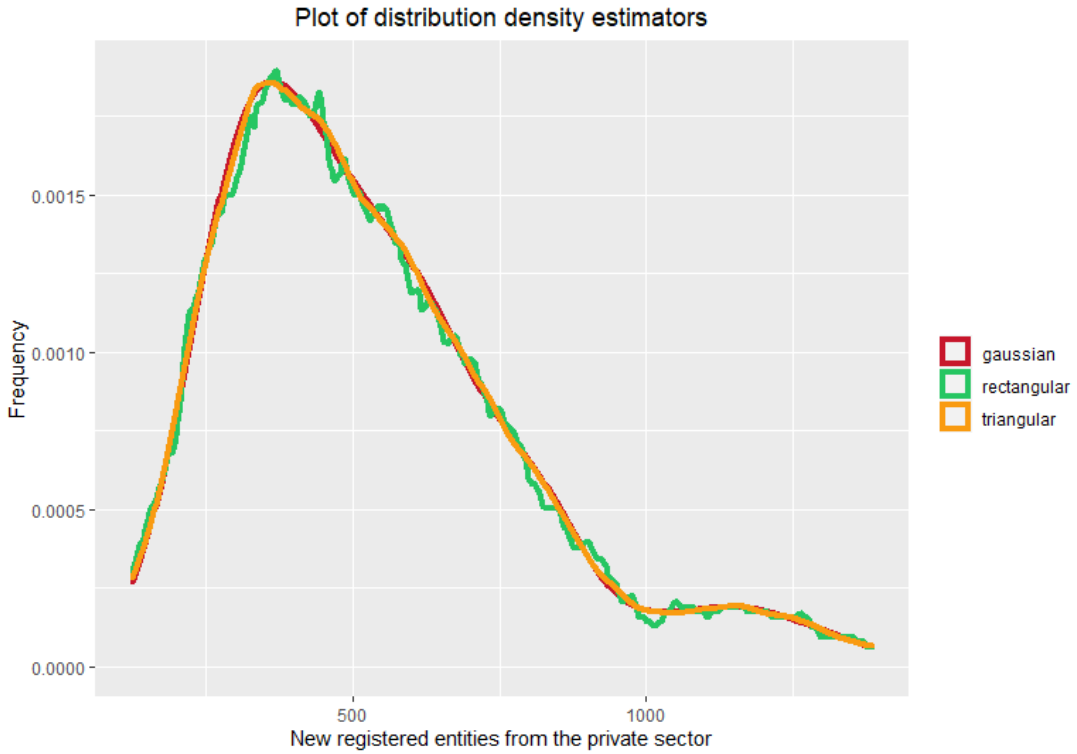
H0: The sample is from a normally distributed population

H1: The sample is not from a normally distributed population

```
        Shapiro-Wilk normality test

data:  my_data$entities
W = 0.91763, p-value = 1.395e-10
```

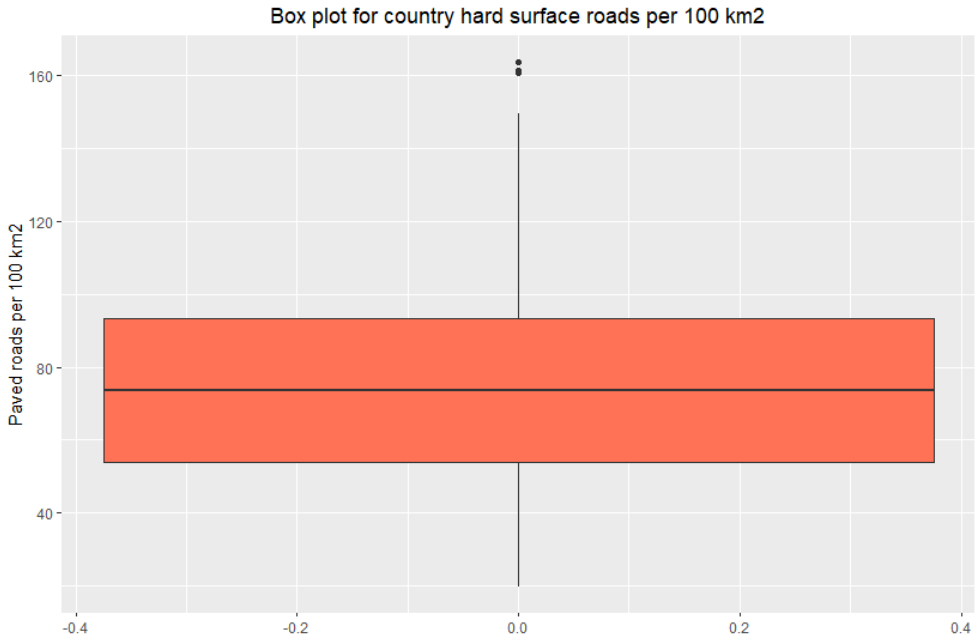
The obtained p-value is lower than the significance level $\alpha = 0.05$, therefore the null hypothesis about the normality of the analyzed distribution is rejected. The histogram also shows that the studied variable is not normally distributed.

The last stage of the explained variable analysis is generating a plot of the distribution density estimators for 3 selected nuclei.



Based on the diagram, it can be concluded that in the case of a rectangular nucleus, the distribution clearly differs from the others.

2.2 Index of poviat roads with hard surface per 100 km2 – explanatory variable



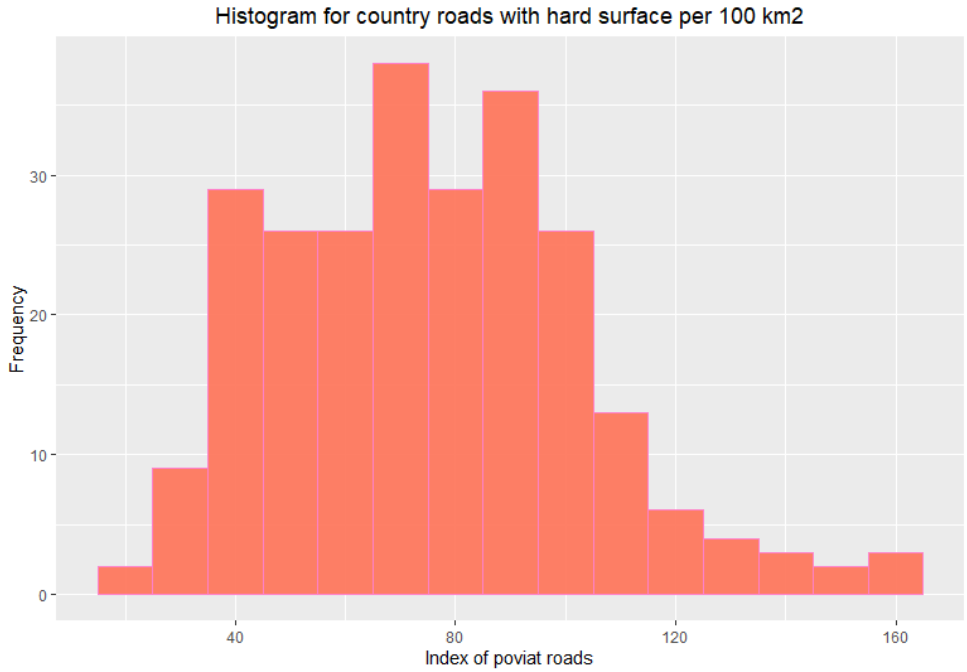
The presented graph shows that there are no outliers in the data. The visualization shows a slight right-hand skewness of the distribution of the index of poviat roads with a hard surface per 100 $km2$. This means that there are more poviats where the index of paved roads per 100 $km2$ is higher than the average for poviats in Poland.

The next step in the analysis is the calculation of the basic variable statistics.

roads	
Min	19.90
Q1	54.00
Median	73.60
Average	75.58
Q3	93.30
Max	163.50
Skewness	0.4979
Kurtosis	0.1632

Among the presented values, one can observe a kurtosis greater than 0, which means that there is a leptokurtic distribution.

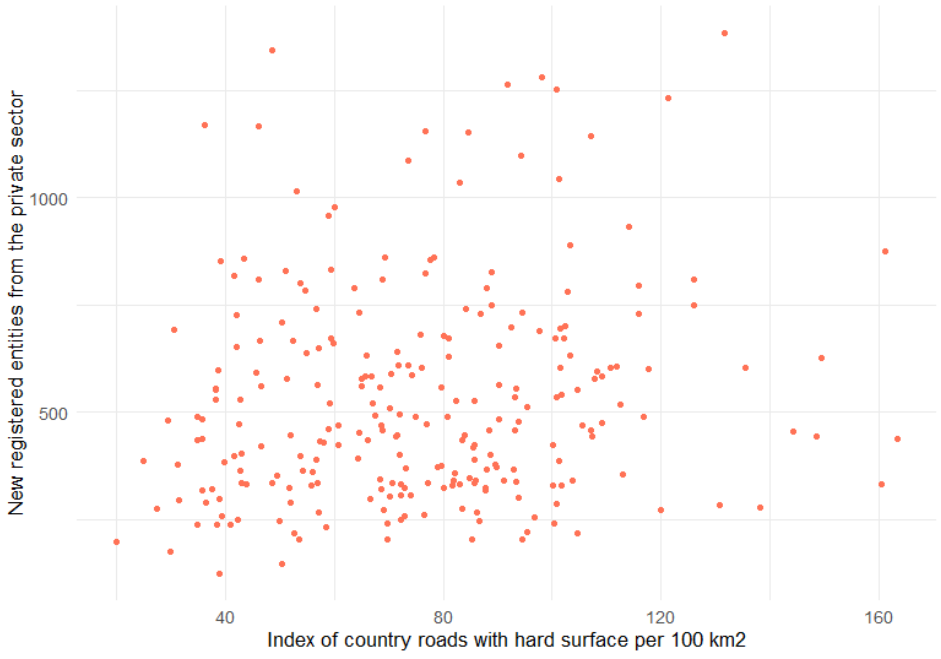
In the further part of the variable analysis, the normality of the distribution is examined. The test uses a histogram and a Shapiro-Wilk normality test.




```
        Shapiro-Wilk normality test

data:  my_data$roads
W = 0.97632, p-value = 0.0003268
```

The obtained p-value is lower than the significance level $\alpha = 0.05$, therefore the null hypothesis about the normality of the analyzed distribution is rejected. The histogram also shows that the studied variable is not normally distributed.

The last stage of the explanatory variable analysis is the study of the relationship between the dependent and the explanatory variable.



The presented graph and the Pearson's linear correlation coefficient for these variables with a value of 0.15 indicate a very weak correlation.

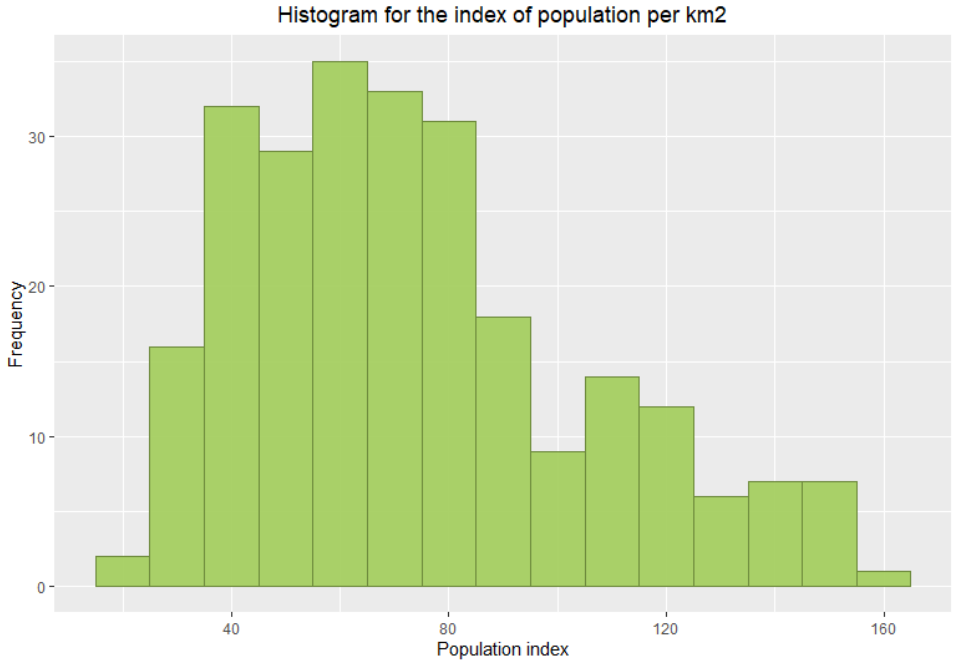
2.3 Population rate per 1 km2 - explanatory variable

The presented chart shows that there are no outliers in the data. The visualization indicates the right-hand skew of the distribution of the population index on $1km2$. This means that there are more poviats in which the population ratio per $1km2$ is higher than the average for poviats in Poland.
The next step in the analysis is the calculation of the basic variable statistics.

population	
Min	19.00
Q1	50.00
Median	69.00
Average	74.38
Q3	93.00
Max	158.00
Skewness	0.6849
Kurtosis	-0,2385

Among the presented values, one can observe a kurtosis lower than 0, which means that there is a platykurtic distribution.

In the further part of the variable analysis, the normality of the distribution is examined. The test uses a histogram and a Shapiro-Wilk normality test.



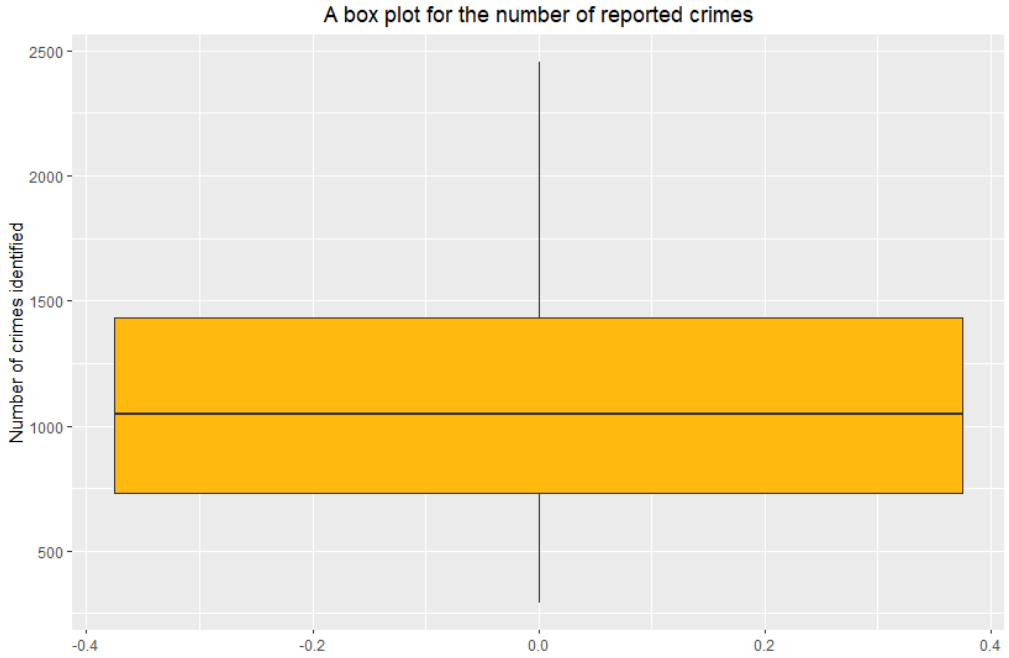

```
        Shapiro-Wilk normality test

data:  my_data$population
W = 0.95026, p-value = 1.404e-07
```

The obtained p-value is lower than the significance level $\alpha = 0.05$, therefore the null hypothesis about the normality of the analyzed distribution is rejected. The histogram also shows that the studied variable is not normally distributed.

The last stage of the explanatory variable analysis is the study of the relationship between the dependent and the explanatory variable.

The presented graph and the Pearson's linear correlation coefficient for these variables with a value of 0.44 indicate a moderate linear relationship.

2.4 Liczba stwierdzonych przestępstw – zmienna objaśniająca



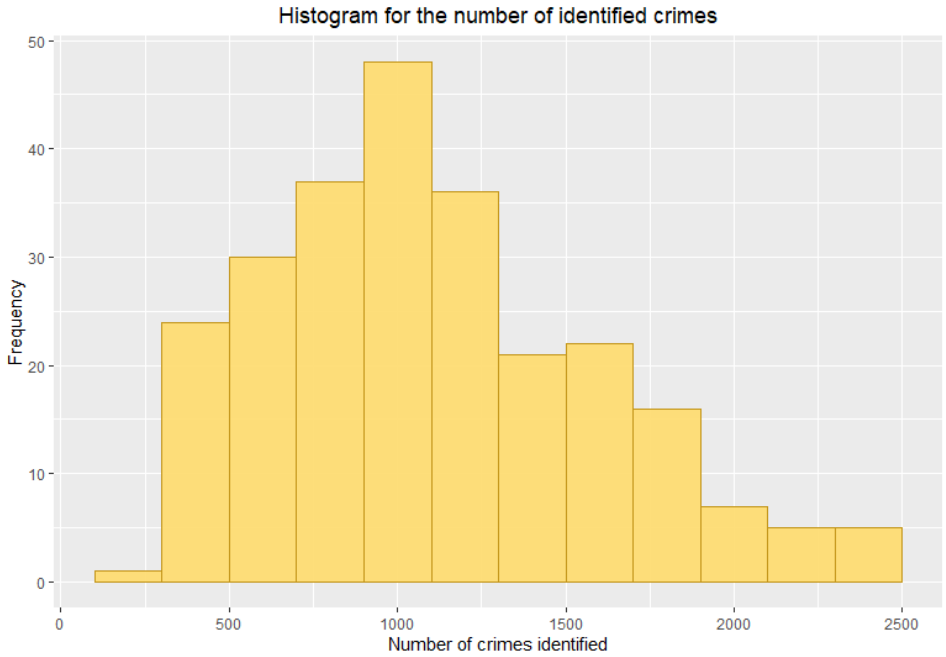
The presented graph shows that there are no outliers in the data. The visualization shows the right-hand skew of the distribution of the number of identified crimes. This means that there are more poviats where the number of crimes is higher than the average for poviats in Poland.

The next step in the analysis is the calculation of the basic variable statistics.

crimes	
Min	293
Q1	733
Median	1048
Average	1109
Q3	1431
Max	2454
Skewness	0.6093
Kurtosis	-0.1560

Among the presented values, one can observe a kurtosis lower than 0, which means that there is a platykurtic distribution.

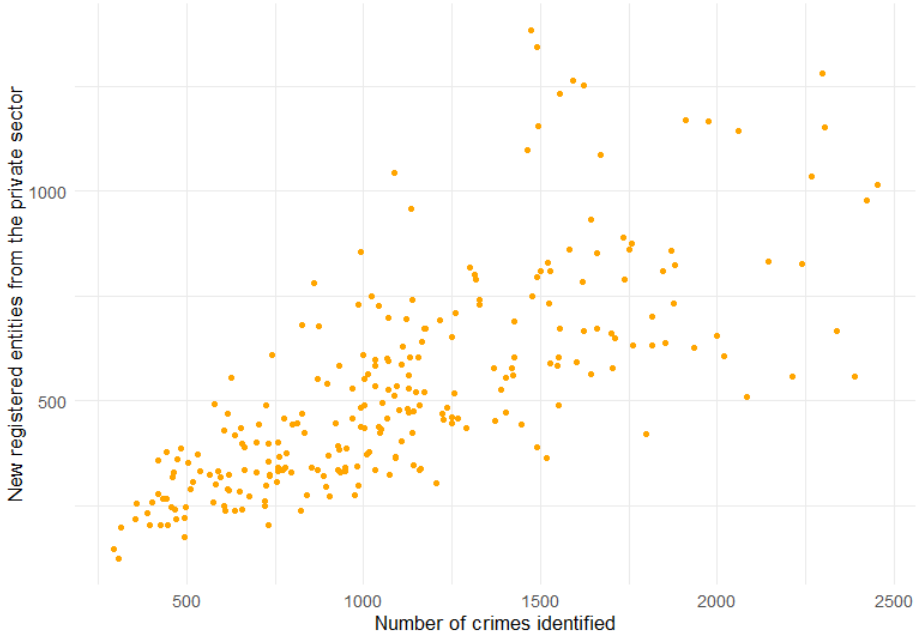
In the further part of the variable analysis, the normality of the distribution is examined. The test uses a histogram and a Shapiro-Wilk normality test.




```
        Shapiro-Wilk normality test

data:  my_data$crimes
W = 0.96281, p-value = 4.114e-06
```

The obtained p-value is lower than the significance level $\alpha = 0.05$, therefore the null hypothesis about the normality of the analyzed distribution is rejected. The histogram also shows that the studied variable is not normally distributed.

The last stage of the explanatory variable analysis is the study of the relationship between the dependent and the explanatory variable.



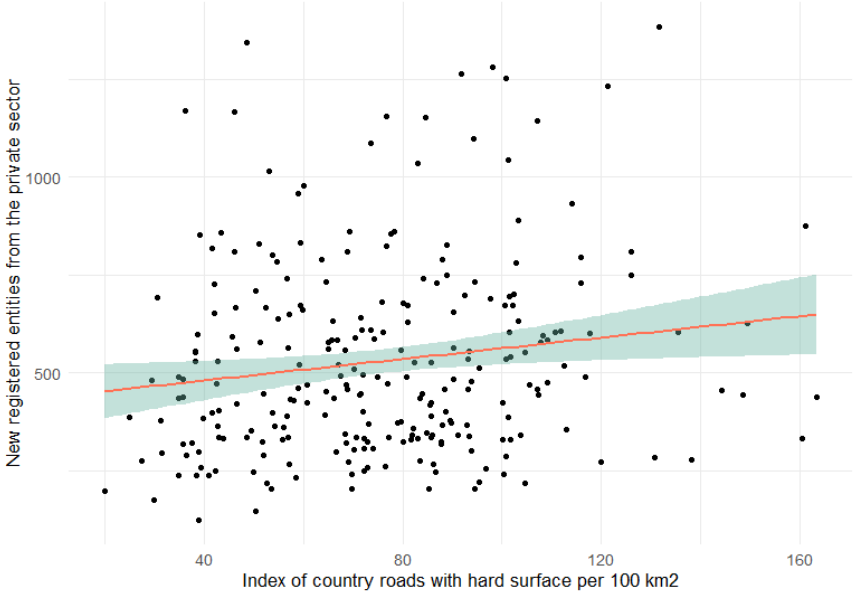
The presented graph and the Pearson's linear correlation coefficient for these variables with a value of 0.73 indicate a fairly strong linear relationship.

3. Linear regression - estimation using the LSM method

In this part of the project, models are estimated using the NMK. These are not dream models and variables, but judging by the topic, it is not the goal of the project.By analyzing subsequent models, a visible effect of heteroscedasticity can be observed.

As the value of the explanatory variable increases, the variance increases. In this case, an attempt should be made to transform the variables by, for example, logarithmizing them to a form that would allow further investigation of parametric and nonparametric regression.

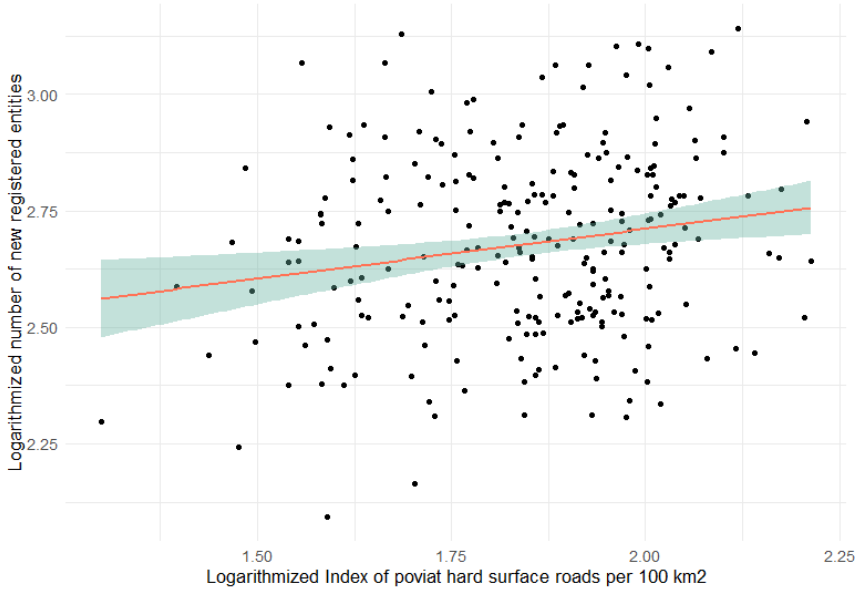
3.1 Index of country hard surface roads per 100 km2



The obtained results are quite surprising, as a stronger positive correlation could be expected. One of the main requirements for the presence of investors and business entities in a given region is a well-developed road infrastructure, expressed, for example, in the index of poviat roads with a hard surface per $100 km2$.

```
Call:
lm(formula = entities ~ roads, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max
-355.02 -185.06  -50.66  132.15  851.85

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 425.8180    45.3056   9.399   <2e-16 ***
roads         1.3677     0.5629   2.430   0.0158 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 247.1 on 250 degrees of freedom
Multiple R-squared:  0.02307,   Adjusted R-squared:  0.01916
F-statistic: 5.903 on 1 and 250 DF,  p-value: 0.01582
```

The printout shows a very low coefficient of determination at the level of about 2%, which means that the potential explanatory variable in this form explains the dependent variable in 2%. The fact that the correlation is at a very low level makes it possible to try to save the model by logarithmizing the variable.



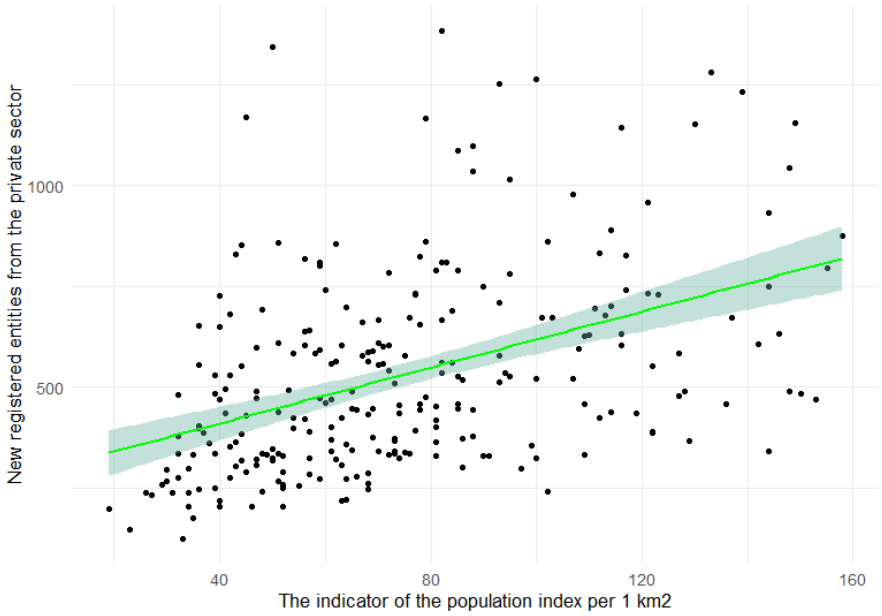
The transformation of the variable does not bring any significant improvement. The correlation between the variables is at the level of 0.18, which means a weak correlation. Also, the coefficient of determination is close to 3%.

```
Call:
lm(formula = log10(entities) ~ log10(roads), data = my_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.53023 -0.14275 -0.00092  0.13443  0.48431

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.28443    0.13515  16.903  < 2e-16 ***
log10(roads)  0.21336    0.07285   2.929  0.00372 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1956 on 250 degrees of freedom
Multiple R-squared:  0.03317,   Adjusted R-squared:  0.0293
F-statistic: 8.577 on 1 and 250 DF,  p-value: 0.003719
```

MAE	MAPE	RMSE	R^2
0.1594	0.0600	0.1948	0.0332

Mean Absolute Percentage Error (MAPE) indicates that the model is forecast incorrectly by an average of 6%. The mean absolute error (MAE) is not significantly different from the root mean square error (RMSE).

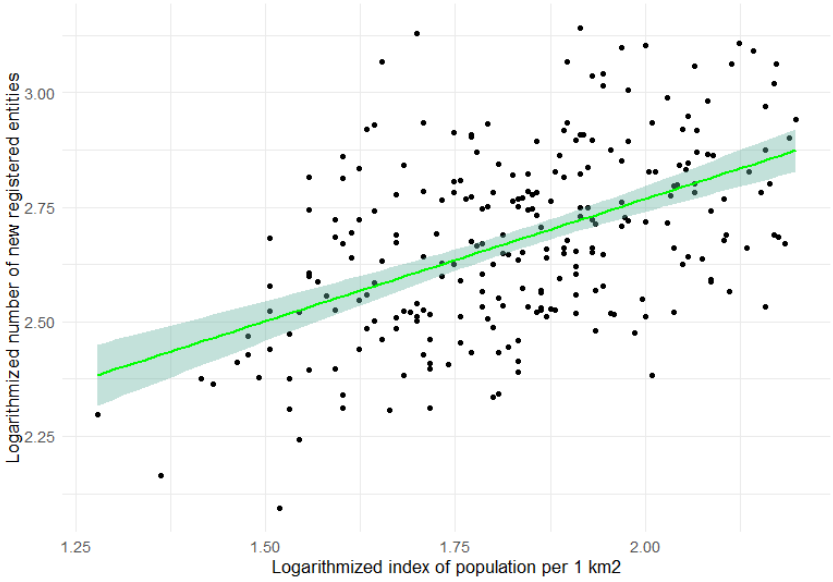
3.2 Population rate per 1 km2



The presented chart shows that the number of new registered entities from the private sector increases with the increase of the population index to $1km2$. The correlation between the variables is 0.44. The visualization shows heteroscedasticity, which makes it necessary to transform the potential explanatory variable.

```
Call:
lm(formula = entities ~ population, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max
-429.75 -151.74  -51.19  114.50  899.38

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 271.1492    36.4615   7.437 1.65e-12 ***
population    3.4695     0.4517   7.680 3.59e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 224.8 on 250 degrees of freedom
Multiple R-squared:  0.1909,    Adjusted R-squared:  0.1877
F-statistic: 58.99 on 1 and 250 DF,  p-value: 3.587e-13
```

The determination coefficient amounts to 19%, i.e. the dependent variable is explained in 19%.



Logging the variable improves curve fit to the observation. The correlation coefficient is 0.50, which means a moderate linear relationship.

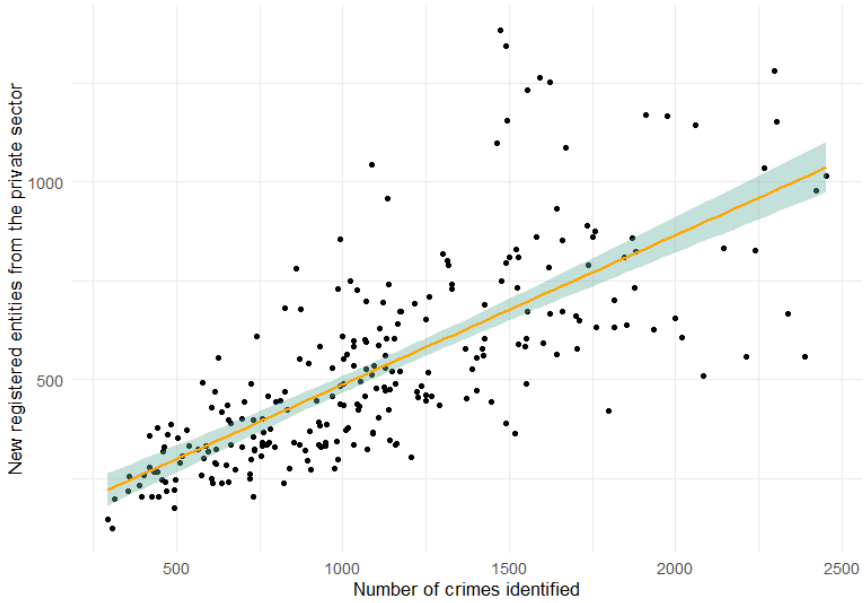
```
Call:
lm(formula = log10(entities) ~ log10(population), data = my_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41796 -0.12665 -0.01186  0.12360  0.52091

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.70265    0.10678  15.945   <2e-16 ***
log10(population)  0.53259    0.05797   9.187   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.172 on 250 degrees of freedom
Multiple R-squared:  0.2524,    Adjusted R-squared:  0.2494
F-statistic:  84.4 on 1 and 250 DF,  p-value: < 2.2e-16
```

The coefficient of determination is approximately 0.25, which means that the dependent variable is explained by the analyzed variable in 25%.

MAE	MAPE	RMSE	R^2
0.1403	0.0526	0.1713	0.2524

Mean Absolute Percentage Error (MAPE) indicates that the model is forecast by an average of 5% wrong. The mean absolute error (MAE) and the root mean square error (RMSE) are similar.

3.3 Number of reported crimes

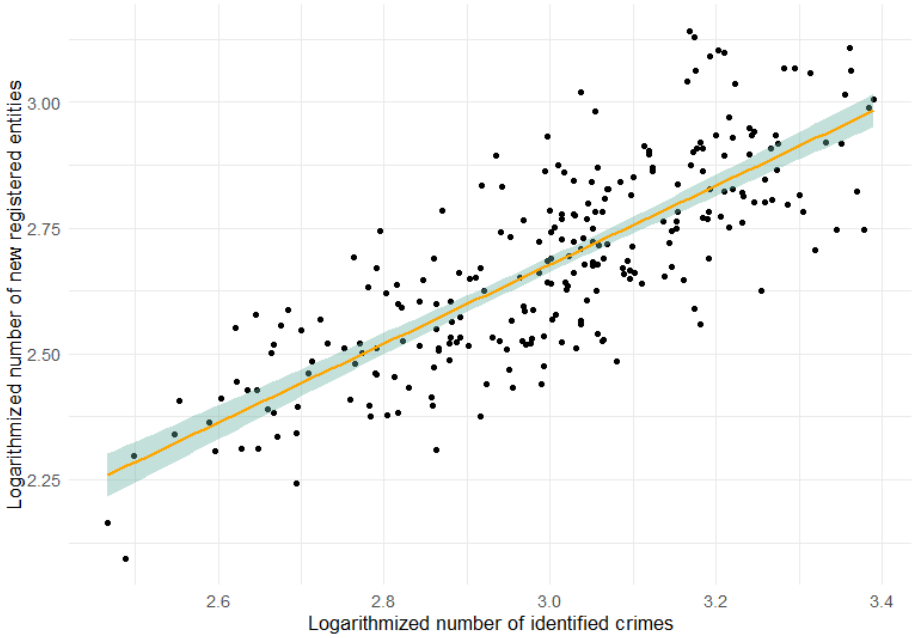


The presented graph shows interesting results. The increase in the number of new registered entities from the private sector is accompanied by an increase in the number of identified crimes. On the surface, this relationship might seem illogical, but it should be noted that the most data is clustered with smaller numbers of offenses. Economic entities may take into account the poviat's safety level, as this may affect their activities or people associated with them. A positive linear relationship may result from the occurrence of poviats with a large population, where the number of new entities and committed offenses is greater than the numbers in smaller poviats.

The correlation between the variables is 0.73. The visualization shows heteroscedasticity, which makes it necessary to transform the potential explanatory variable.

```
Call:
lm(formula = entities ~ crimes, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max
-455.03  -99.51  -27.38   80.17  717.31

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 110.73968   26.94444    4.11 5.37e-05 ***
crimes        0.37742    0.02229   16.93  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 170.6 on 250 degrees of freedom
Multiple R-squared:  0.5343,    Adjusted R-squared:  0.5324
F-statistic: 286.8 on 1 and 250 DF,  p-value: < 2.2e-16
```

The coefficient of determination reaches the level of about 53%, which means that the dependent variable is explained by this variable in 53%.



Logging the variable improves curve fit to the observation. The correlation coefficient is 0.79, which means a fairly strong linear relationship.

```
Call:
lm(formula = log10(entities) ~ log10(crimes), data = my_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.26058 -0.08281 -0.00740  0.08352  0.33160

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.32253    0.11467   2.813  0.00531 **
log10(crimes) 0.78499    0.03812  20.591  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1212 on 250 degrees of freedom
Multiple R-squared:  0.6291,    Adjusted R-squared:  0.6276
F-statistic:   424 on 1 and 250 DF,  p-value: < 2.2e-16
```

The coefficient of determination is approximately 0.63, which means that the dependent variable is explained by the analyzed variable in 63%.

MAE	MAPE	RMSE	R^2
0.0978	0.0365	0.1207	0.6291

Mean Absolute Percentage Error (MAPE) indicates that the model is forecast incorrectly by an average of about 4%. The mean absolute error (MAE) is not significantly different from the root mean square error (RMSE).

3.4 Comparison of OLS models

	MAE	MAPE	RMSE	R^2
roads	0.1594	0.0600	0.1948	0.0332
population	0.1403	0.0526	0.1713	0.2524
crimes	0.0978	0.0365	0.1207	0.6291

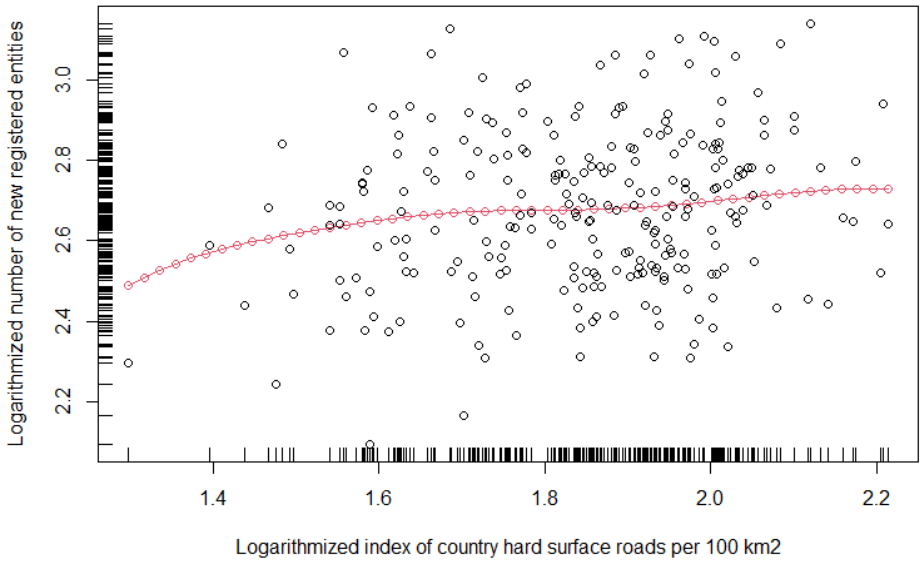
By analyzing the table, it can be concluded that the model containing the crime variable best explains the number of new registered entities, but also has the best prognostic properties. For this model, the errors have the lowest values, while the coefficient of determination R^2 is the highest.

4. Nonparametric regression

The next part of the project concerns model estimation using nonparametric regression. Logarithmized variables are used for the subsequent comparison of the prognostic properties of the selected model with the model obtained using the LSM method.

Nuclear estimation of regression is obtained using the Nadaray-Watson kernel. The nonparametric regression line is marked in red.

4.1 Index of country roads with hard surface per 100 km2

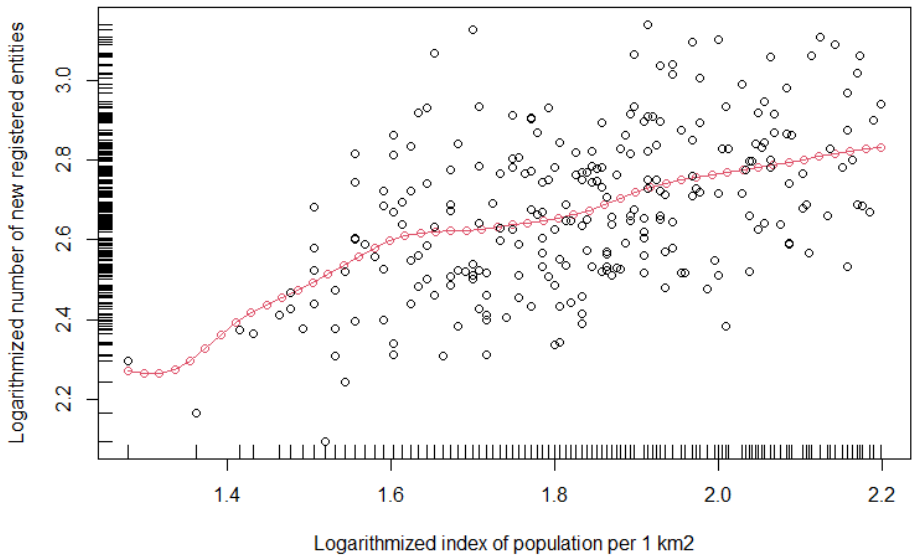


The presented graph shows that the nonparametric regression curve is similar to the linear regression curve but covers more observations.

MAE	MAPE	RMSE	R^2
0.1590	0.0599	0.1940	0.0411

Mean Absolute Percentage Error (MAPE) indicates that the model is forecast incorrectly by an average of about 6%. The mean absolute error (MAE) is not significantly different from the root mean square error (RMSE). The R^2 coefficient of determination indicates that the explained variable is explained in about 4%.

4.2 Population rate per 1 km2

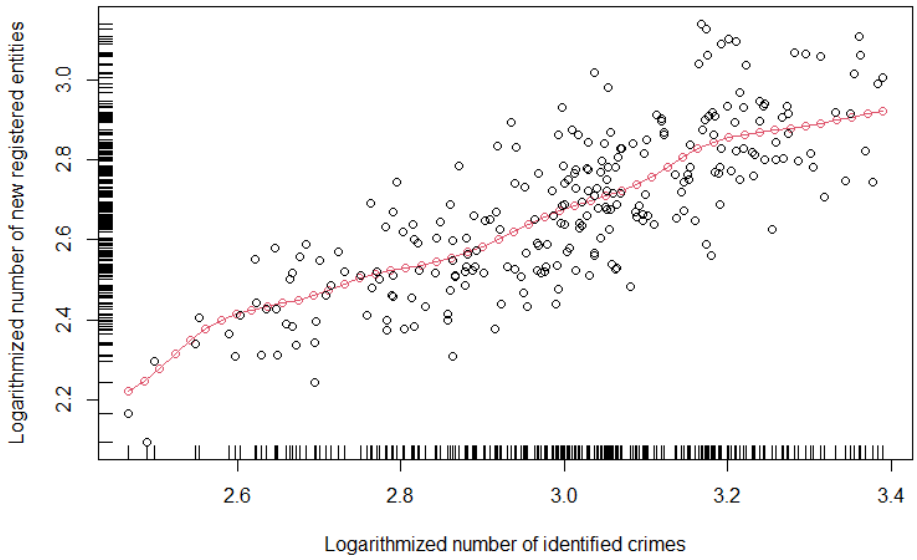


The presented graph shows that the nonparametric regression curve is similar to the linear regression curve with larger deviations.

MAE	MAPE	RMSE	R^2
0.1386	0.0519	0.1687	0.2752

Mean Absolute Percentage Error (MAPE) indicates that the model is forecast incorrectly by an average of about 5%. The mean absolute error (MAE) is not significantly different from the root mean square error (RMSE). The R^2 coefficient of determination indicates that the explained variable is explained in approximately 28%.

4.3 Number of crimes committed



The presented graph shows that the nonparametric regression curve is similar to the linear regression curve but covers more observations.

MAE	MAPE	RMSE	R^2
0.0966	0.0360	0.1188	0.6407

Mean Absolute Percentage Error (MAPE) indicates that the model is forecast incorrectly by an average of about 4%. The mean absolute error (MAE) is not significantly different from the root mean square error (RMSE). The R^2 coefficient of determination indicates that the explained variable is explained in approximately 64%.

4.4 Comparison of nonparametric models

	MAE	MAPE	RMSE	R^2
roads	0.1590	0.0599	0.1940	0.0411
population	0.1386	0.0519	0.1687	0.2752
crimes	0.0966	0.0360	0.1188	0.6407

By analyzing the table, it can be concluded that the model containing the crime variable best explains the number of new registered entities, but also has the best prognostic properties. For this model, the errors have the lowest values, while the coefficient of determination R^2 is the highest.

5. Summary - comparison of prognostic properties

Based on the comparisons of the prognostic properties of the LSM models and non-parametric regression, and the highest values of the R^2 determination coefficients, the crime variable is selected for the cross-test.

Comparing the prognostic properties through a 10-fold cross-validation begins by dividing the set of shuffled data into 10 subsets. For each cross-validation, prognostic properties and coefficients of determination are collected and then averaged.

	MAE	MAPE	RMSE	R^2
Linear regression	0.0986	0.0368	0.1199	0.5884
Nonparametric regression	0.1209	0.0372	0.1209	0.6052

The presented comparison shows that parametric and non-parametric regressions do not differ significantly in terms of prognostic properties and the degree of explanation of the variable number of new registered entities from the private sector.