Predicting Life Expectancy of Indians

Capstone Project
for Springboard Intermediate Data Science course

Client

- Fictional Indian government agency
- Aims to increase the average life expectancy of Indians
- Name: Life Expectancy Action Agency (LEAA)

Tasks

- To create a machine learning model that can predict the life expectancy of Indians using the dataset provide by the client

- To give recommendations to my client based on this process

Dataset

- About 770’000 rows and exactly 121 columns
- Each row represents one deceased Indians
- Based on a survey by the LEAA (my client) in which members of the household of the deceased were asked about the deceased

Thematical groups of variables

1) Variables used for creating target variable

2) Variables about the death of the deceased

3) Variables about the deceased himself/herself

4) Variables about the household/family of the deceased

5) Variables without or with unknown meaning

The only group of variables that can be used for fitting the linear regression model is group 4 about the householdof the deceased

Data Cleaning process

- Creating the target variable ‘lifetime’ - the number of years from birth until death
- Deleting unrealistic ‘lifetime’ values
- Deleting all columns which are not about the household of the deceased (besides ‘lifetime’)
- Dealing with missing values
- Deleting columns with mainly the same value
- Converting integer encodings for categories to strings describing the categories
- Deleting rows with undefined categories
- Dividing Data in training and test set

Shape of dataset after Data Cleaning

- 715’550 rows left (training set + test set)
- 1 target variable, ‘lifetime’
- 29 categorical variables (mainly nominal)
- 1 numerical variable (discrete) – the number of rooms in possession of the household

Exploratory Data Analysis (Part 1)

- ‘Lifetime’ target variable is not normally distributed – taking the log didn’t make it Gaussian either
  - ➔ Thus I chose permutation test for evaluating whether there are statistically significant differences within the categories
    - ➔ Within all categories there are statistically significant differences in average ‘lifetime’

Exploratory Data Analysis (Part 2)

- The sole numerical variable (number of rooms) is weakly but statistically significantly correlated to the target variable
- It is not strongly correlated to any of categorical feature variables (judged by Spearman’s rank order correlation)
- Some of the categorical feature variables are strongly associated with each other (Cramer’s V > 0.3)

Fitting the Linear Regression model

The following is True for all iterations of the fitting process:

- I’ve used 3-fold cross validation
- Error metric I’ve used: Mean Absolute Error
- I’ve calculated the statistics such as Pearson’s R between actual and predicted values, MAE, p value for normality of residuals, mean of residuals
- I’ve created plots such as histogram of residuals, quantile-quantile plot of residuals, scatter plot of actual vs predicted values, scatter plot of predicted values vs residuals
- Categorical variables are one-hot encoded, the discrete variable from 0 to 1

Feature Selection for Linear Regression Model

- First I tried to add variables step by step, ordered by the biggest difference of average lifetime variables within the variables
- I left out variables highly associated with any variable already in the model
- Assessing improvement after each added variable

Fitting the Linear Regression model (Part 3)

I’ve tried out the following adjustments (The best option in parentheses):

- Taking the log of the target variable and not (not)
- Removing outliers of the target variables and not (not)
- Fitting with and without intercept (without)

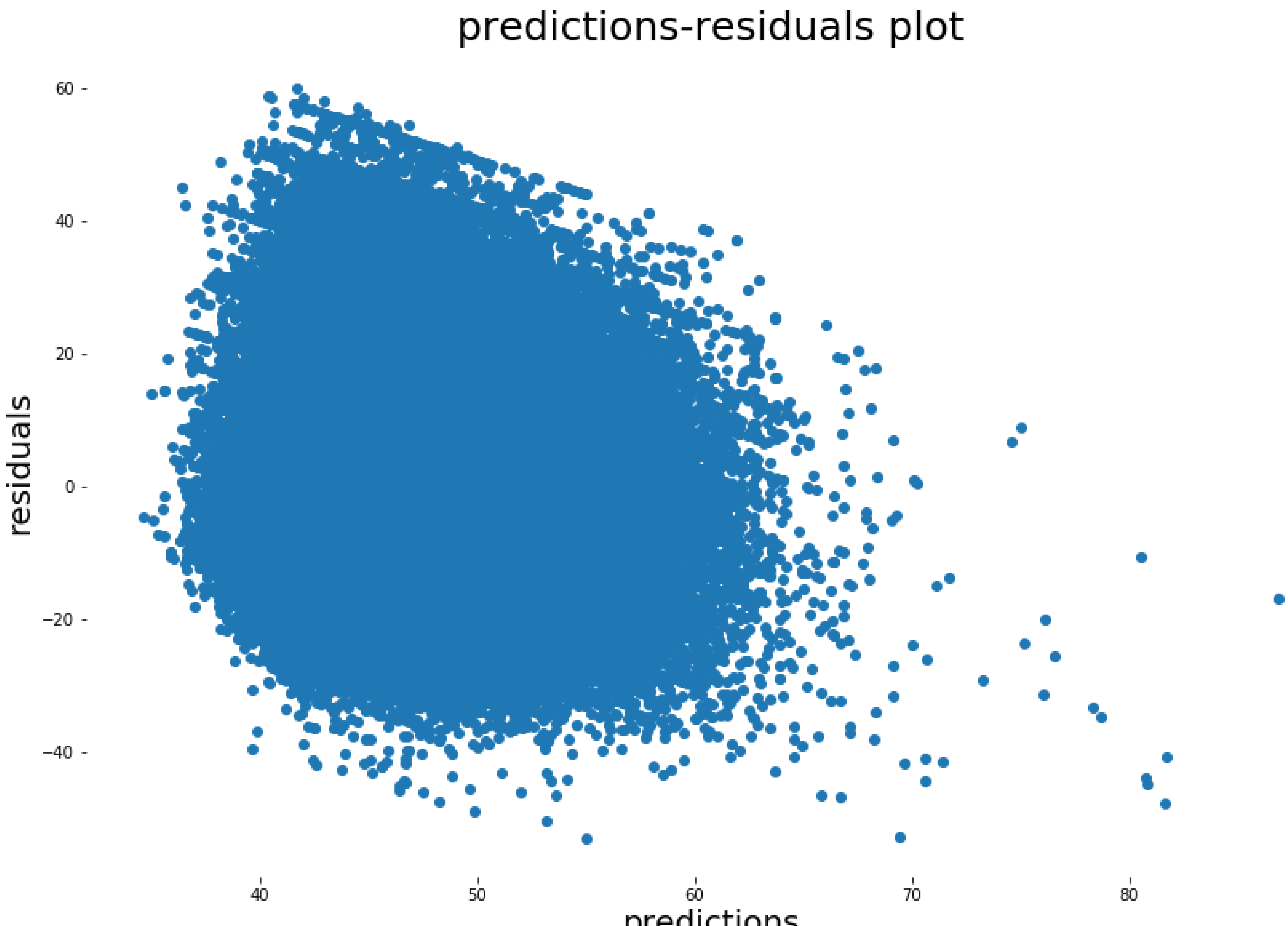
Performance of best model

- Overall the performance was rather weak
- MAE about 11.32
- Heteroscedascidity
- Residuals not normally distributed
- Rather weak correlation of predicted vs. actual values
- no predictions above about 34.6 years and only few above about 70

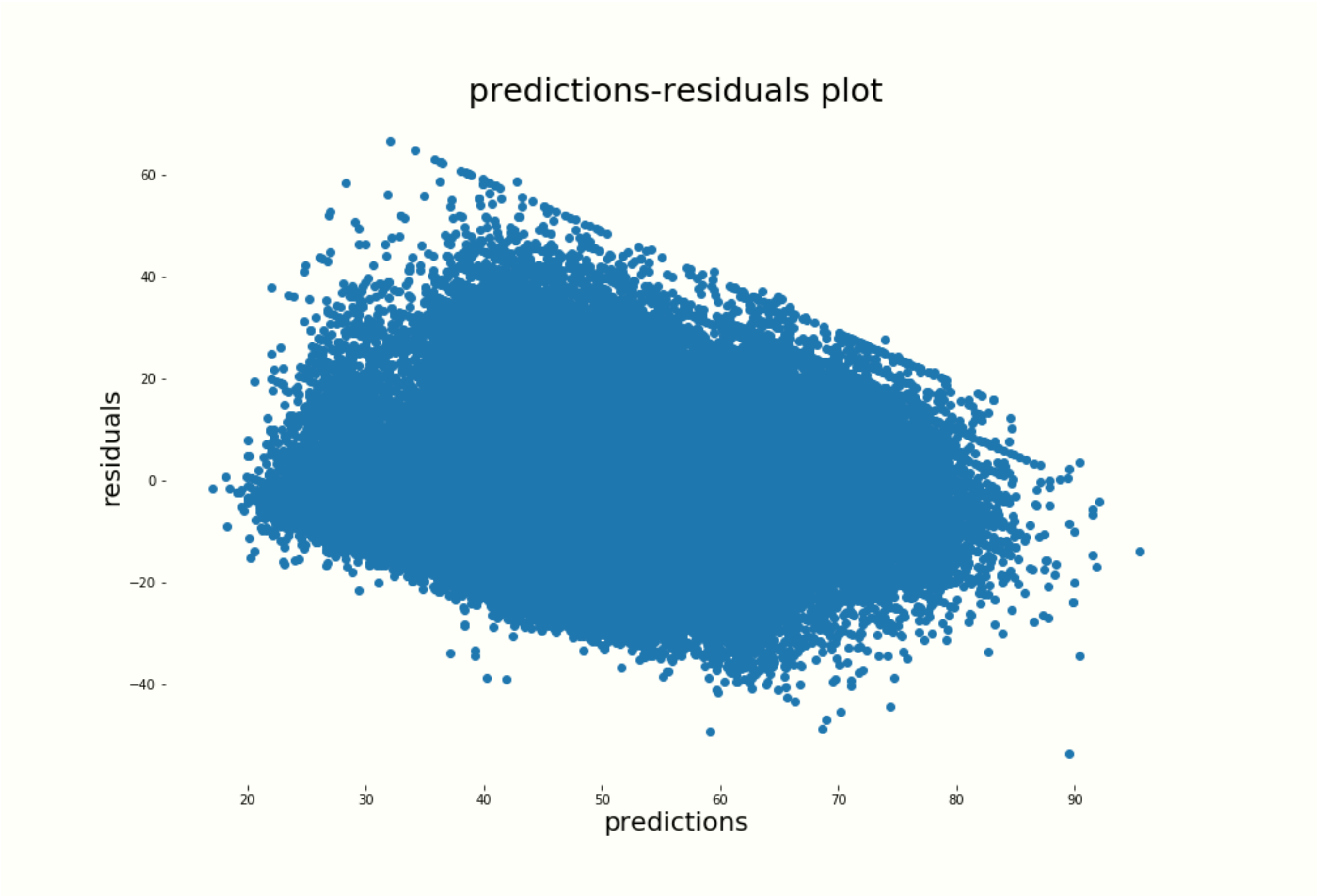
‘Experimental’ model with added personal variables (compared to ‘’real’ model)

- Performed clearly better
- MAE about 2 years better
- Residuals still not gaussian
- Less Heteroscedascidity than
- A far higher range of predicted values
- Clearly stronger correlation between actual and predicted values

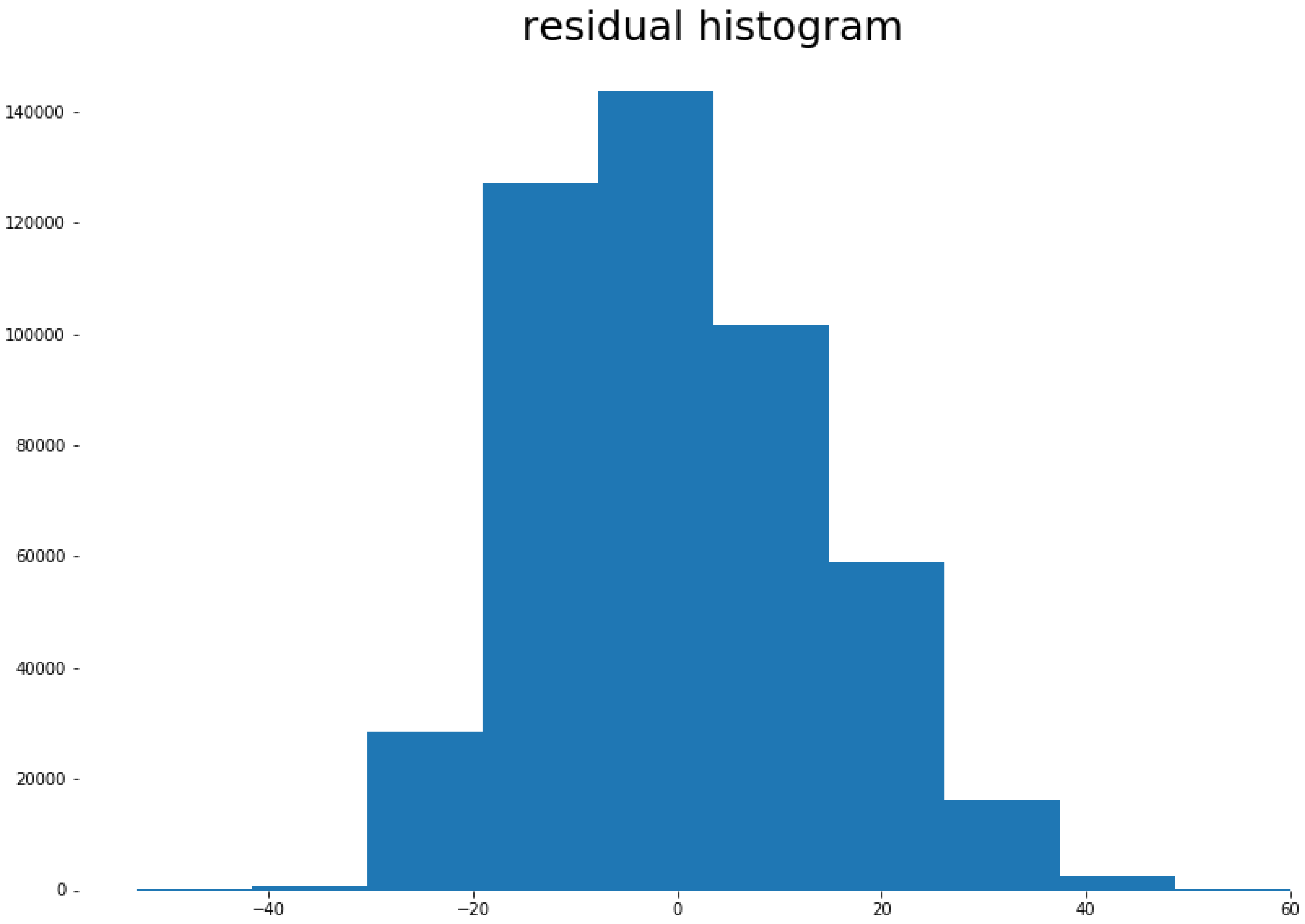
‘Real’


predictions-residuals plot
60
40
20
0
−20
−40
residuals
40
50
60
70
80
predictions

‘Experimental’

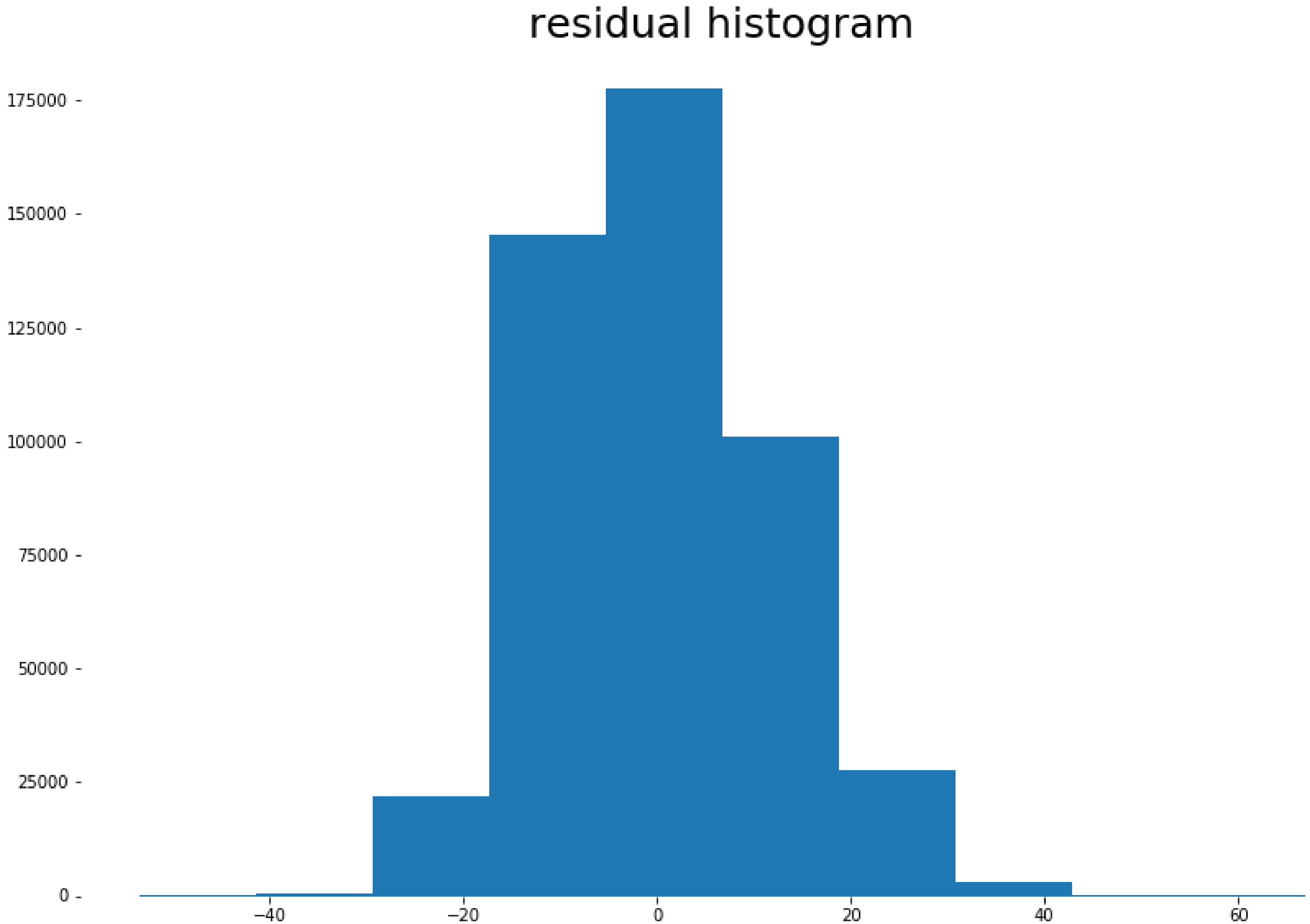


‘Real’

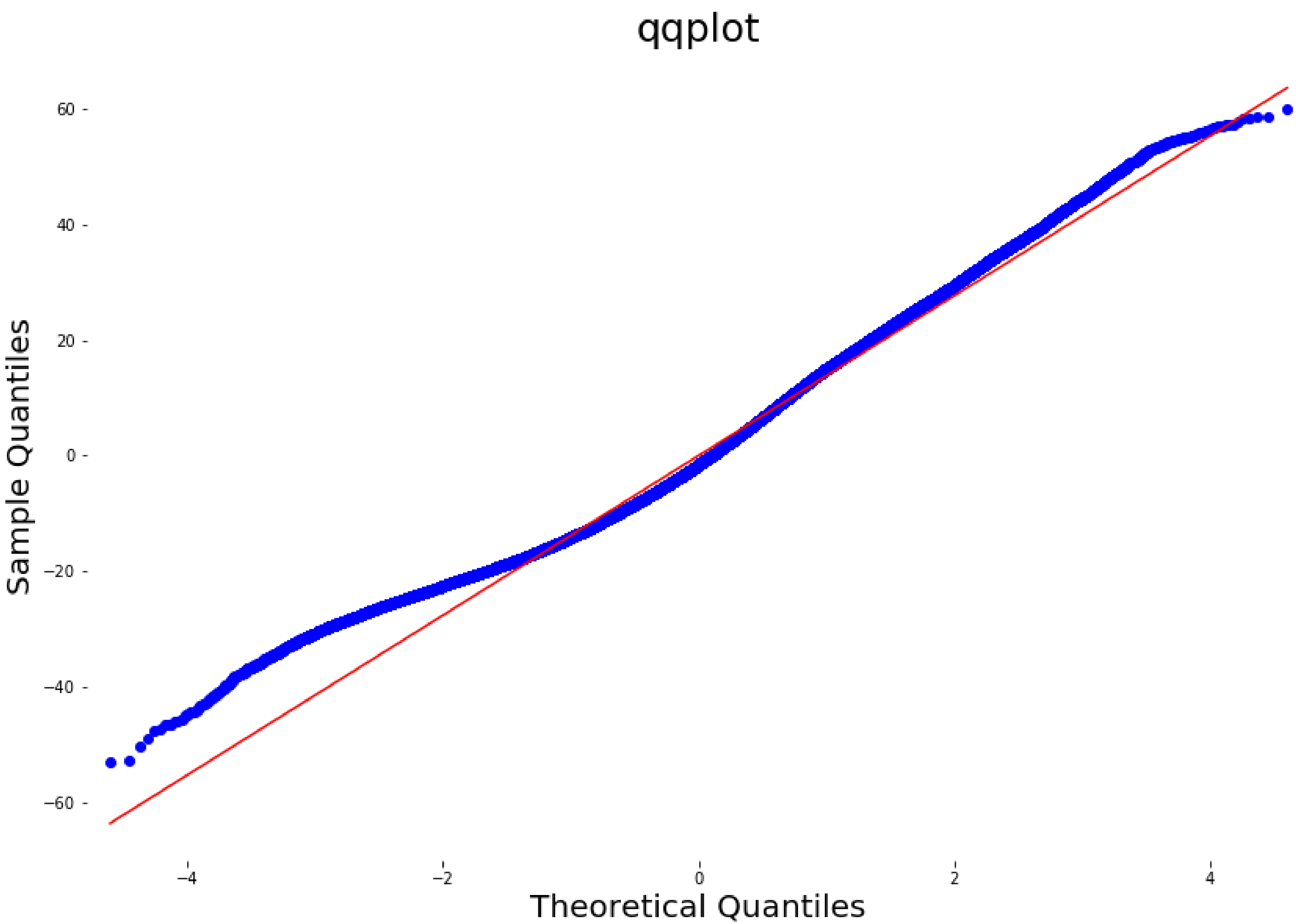


residual histogram
140000
120000
100000
80000
60000
40000
20000
0
−40
−20
0
20
40
60

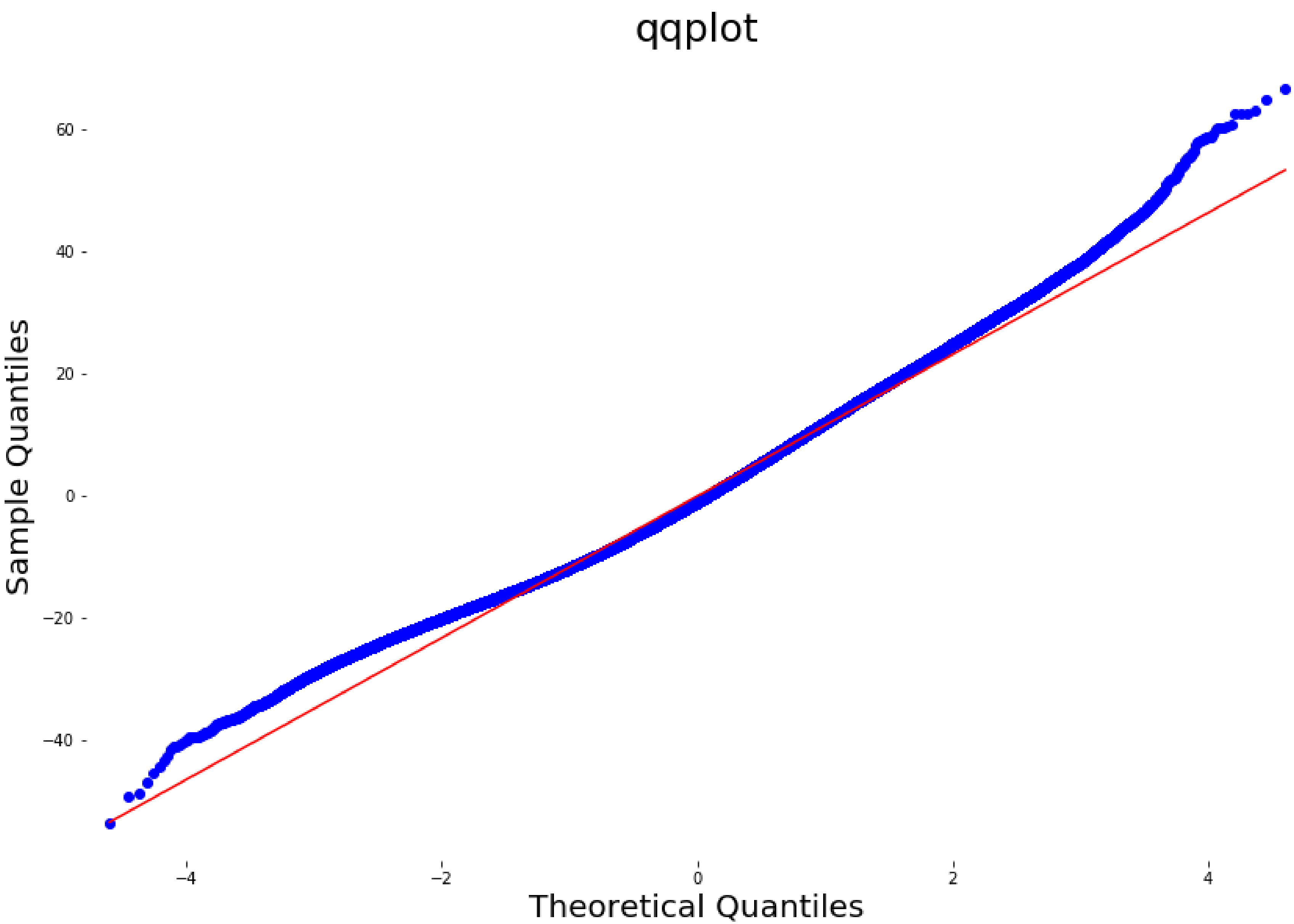
‘Experimental’



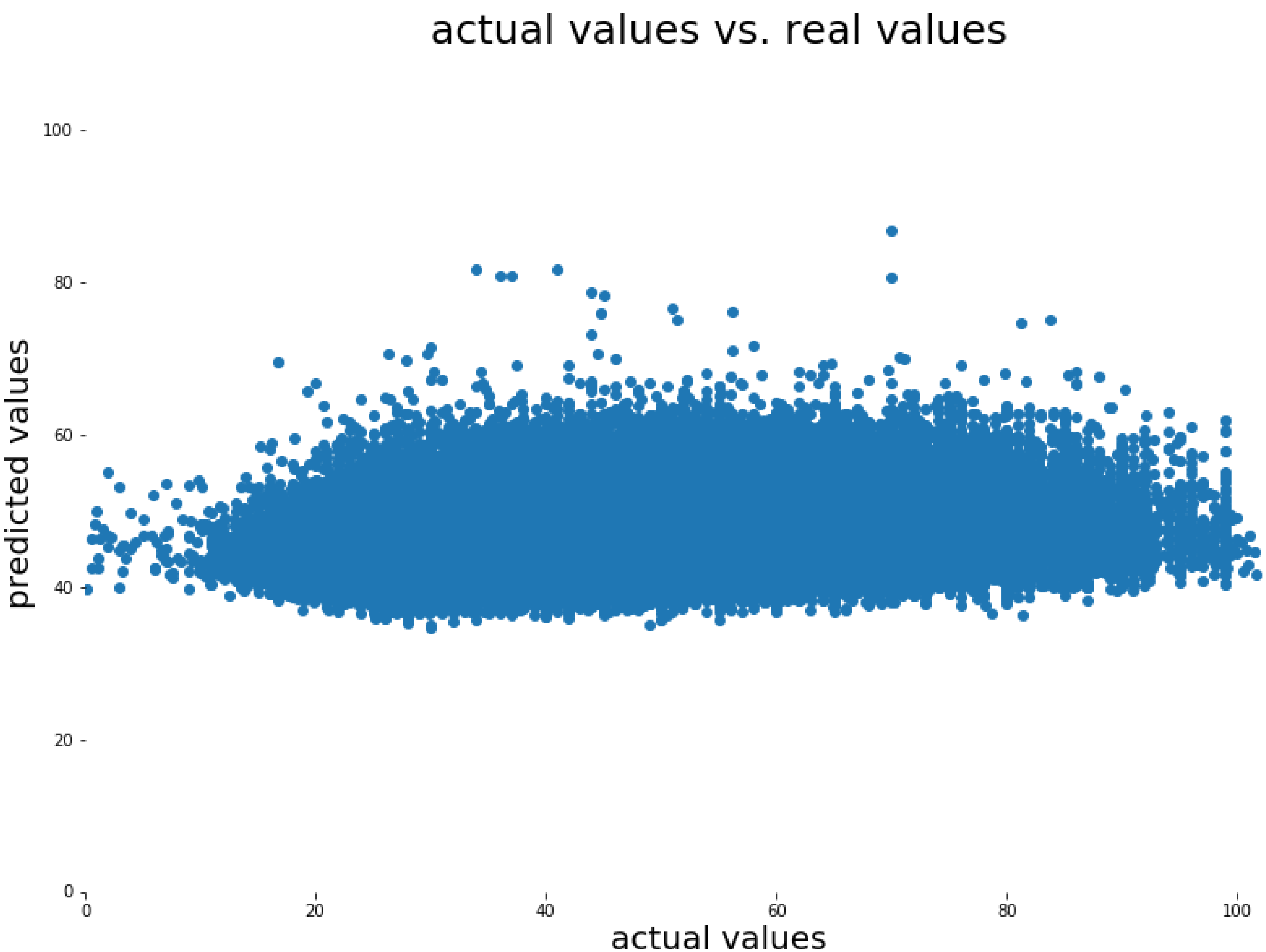
‘Real’


qqplot
Sample Quantiles
Theoretical Quantiles
60
40
20
0
−20
−40
−60
−4
−2
0
2
4

Experimental

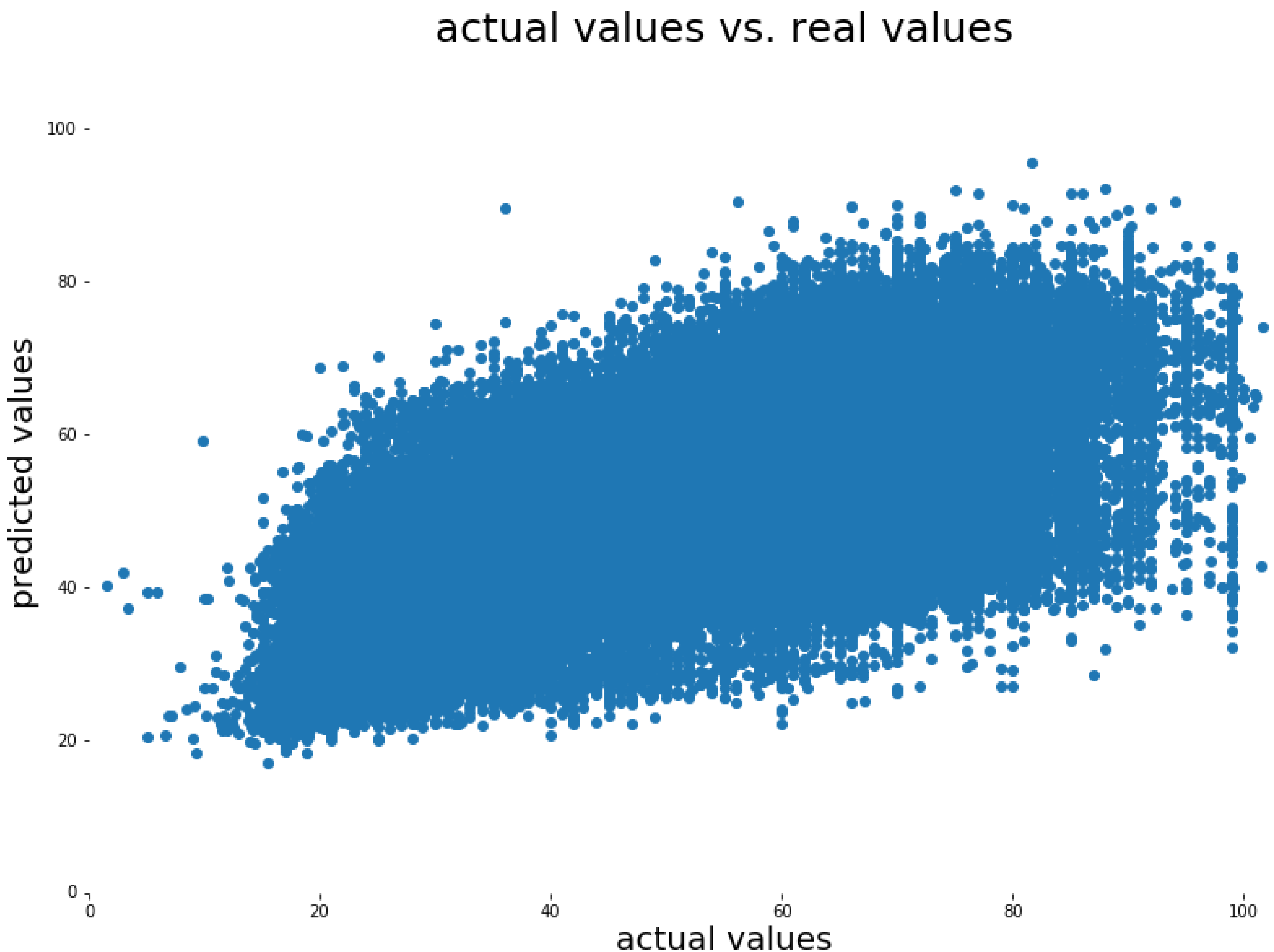


‘Real’



actual values vs. real values
100
80
60
40
20
0
0
20
40
60
80
100
predicted values
actual values

‘Experimental’



Recommendations

- The model is rather weak, so I would use it carefully. It can be used for getting a first impression of a situation however.
- The coefficients of the model can be used in order to educate the staff in the field.
- I would consider to redo the survey in a better way. This might improve the model, because personal variables could be used.