4 Regression analysis

4.1 Performing an automated regression analysis

To examine and to quantify the effects that different neighbourhood characteristics have on different crime types, we perform linear regressions. More precisely, we run one regression for each of the seven crime types.

While the crime types are always set to be the dependent variable, we have various neighbourhood characteristics available as possible explanatory variables. Certainly, it is not reasonable to include all of them at once: As some of the variables like *youth* and *male.youth* are highly correlated, we would face the problem of imperfect multicollinearity, leading to high standard errors and unstable estimators. Therefore, it is necessary to select those variables that appear to be the most influential on crime and on the same time describe a different neighbourhood characteristic and to exclude all other variables prior to running the regression. After having tried out different combinations of variables and examining their regression results regarding the R squared, standard errors of the estimators, and the correlation among the regressors, the following variables appeared the most reasonable for us to be included: The number of young males between 15 and 24 years living in the respective neighbourhood, the number of people who have less than a high school degree, the number of people who belong to the low-income class by earning less than 40.000 CAD, and the number of immigrants living in the neighbourhood. To better compare the effects among different crime types, we used the same combination of independent variables for each regression. As described in later subchapters, however, we could not completely eliminate the correlations between the regressors as most variables like income and age variables are already correlated despite referring to different neighbourhood characteristics.

After having defined the dependent and independent variables, we are able to perform the regression analysis as done in the (Q) Automated Regression Quantlet. As implied by its name, we automated the seven regressions by implementing a for-loop.

Prior to running the for-loop, however, some preparations had to be made: Besides converting the data table *agg.2016* from the (Q) Merging Quantlet into a data frame called *r*, creating empty lists and data frames for the regression results and their diagnosis tests to be stored in, and defining the vector *crimetypes* which contains the names of the different crime types to loop over, we defined the function *FindBestExponent*:

*FindBestExponent <- function (p, x) {*

*y <- bcPower(x, p)*

*shapiro.test(y)$statistic*

*}*

Listing 1: (Q) Automated Regression Quantlet – rows 5-15 (without comments)

Based on a numeric vector *x* and a value *p*, the function performs a basic power transformation based on p as the exponent and the values of x as bases, and saves the transformed data inside *y*. The transformed data is then tested for normality by using the Shapiro-Wilk test. The null hypothesis in this case states that the assessed variable is normally distributed. Therefore, a high p-value implies that the assumption of normal distribution cannot be rejected. We later use this function to find the best basic power transformation to shift the dependent variable towards the normal distribution. OLS-estimations do not necessarily require this condition to yield unbiased estimates, and we will assess in the next subchapters whether the transformation has led to a qualitative improvement of the regression results or not.

After these preparations, the for-loop is then performed, leading to seven regression results and their diagnosis tests, stored in the previously created lists and data frames:

*for (i in crimetypes){*

*r$tmp <- r[,i]*

*outliers <- r[r$tmp > mean(r$tmp)+2.5\*IQR(r$tmp), ]$obsnumber*

*rtmp <- r[!r$obsnumber %in% outliers, ]*

*rtmp[rtmp$tmp == 0, ] <- 1*

*exponent <- optimize(FindBestExponent, c(-3,3), x=rtmp$tmp)$objective*

*shapiro.test(bcPower(rtmp$tmp,exponent))*

*rtmp$tmp.bp <- bcPower(rtmp$tmp, 0.2734738)*

*model <- lm(tmp.bp~male.youth + less.than.high.school + low.income*

*+ immigrants, data=rtmp)*

*regressionresults[[i]]<-summary(model)*

*ols.ass[i, "means"] <- mean(model$residuals)*

*ols.ass[i, "bptests"] <- bptest(model)$p.value*

*ols.ass[i, "swtests"] <- shapiro.test(residuals(model))$p.value*

*ols.ass[i, "vif1"] <- vif(model)[1]*

*ols.ass[i, "vif2"] <- vif(model)[2]*

*ols.ass[i, "vif3"] <- vif(model)[3]*

*ols.ass[i, "vif4"] <- vif(model)[4]*

*ols.ass[i, "cortest1"] <- cor.test(rtmp$male.youth, model$residuals)$p.value*

*ols.ass[i, "cortest2"] <- cor.test(rtmp$less.than.high.school,*

*model$residuals)$p.value*

*ols.ass[i, "cortest3"] <- cor.test(rtmp$low.income, model$residuals)$p.value*

*ols.ass[i, "cortest4"] <- cor.test(rtmp$immigrants, model$residuals)$p.value*

*crPlots(model)*

*ceresplots[[i]] <- recordPlot()*

*rm(exponent, model, outliers, rtmp)*

*}*

Listing 2: (Q) Automated Regression Quantlet – rows 38-87 (without comments)

To simplify the loop regarding column referencing, the respective dependent variable is first temporarily stored in an extra column in the *r* data frame, *r$tmp*.

To transform the dependent variables towards a normal distribution, outliers of the dependent variables are removed first. It is reasonable to do this step before the actual transformation, as the optimal transformation parameter would be heavily influenced by those extreme values. After having examined the distribution of all seven crime types, it appeared to be the most reasonable to define those neighbourhoods as outliers, which differed from the mean of all neighbourhoods by more than 2.5 interquartile ranges. Between four and seven neighbourhoods were therefore eliminated from each regression. The data frame *rtmp* is then created, which excludes the outlying neighbourhoods.

As we then want to perform a basic power transformation, we need all observed numbers of crimes to be different from zero. Therefore, we artificially convert possibly occurring zeros into ones, which should not affect the validity of our results. We then calculate the exponent which maximizes the p-value of the Shapiro-Wilk test for our basic power transformation. This is done by implementing the previously defined *FindBestExponent*-function inside the optimize-command.

Having obtained the optimal exponent, we then transform the dependent variable and store it in an extra column *tmp.bp* in the *rtmp*-data frame. As we can see in the following graphs, the transformed dependent variables follow much more a normal distribution than the original ones and the p-value of the respective Shapiro-Wilk test is increased in most of the cases. Two exceptions are Theft Over and Total Crime, where the transformation failed to shift the distribution towards a normal distribution. To better compare the distributions with the normal distribution, a normal density line is added to each graph.

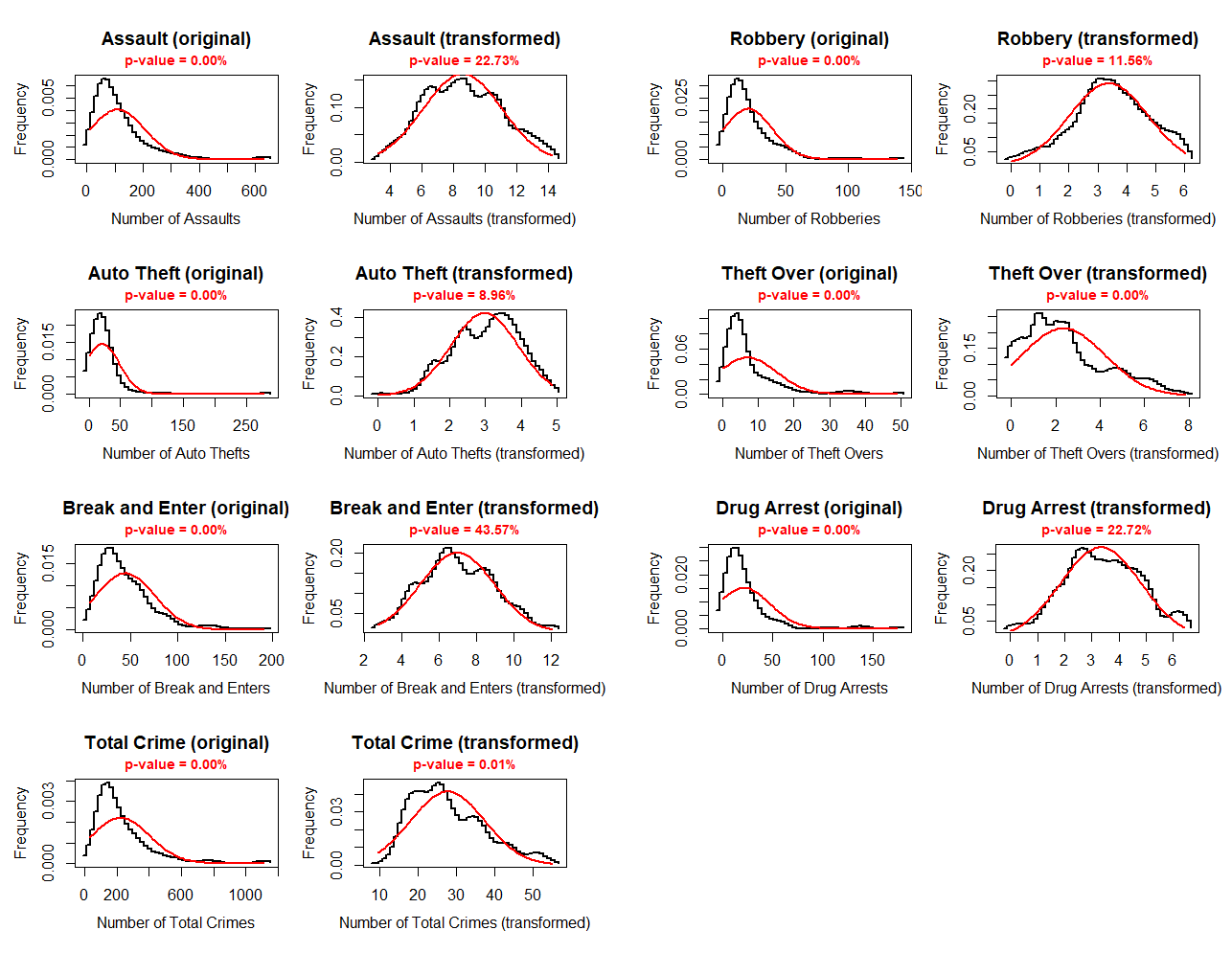


Figure 1: Original and transformed distributions of all crime types.

Using the transformed dependent variable, we then perform the seven regressions again and store them in the *regressionresults*-list.

In the second part of the for-loop, we then examine whether the OLS-assumptions are met for all regressions: Regarding the error term, we test whether it has a population mean of zero, whether it is homoscedastic using the Breusch-Pagan test and whether it is normally distributed using again the Shapiro-Wilk test. The last condition is an optional one as it is not required for the validity of the OLS-assumption, however, normally distributed error terms allow us to perform statistical hypothesis testing of the estimated parameters and generate reliable confidence intervals. All these results are stored in the *ols.ass*-data frame. As the observation order is random, we do not check for serial correlation of the error term.

Regarding the independent variables we need to assess whether there is no multicollinearity among them. This is done by calculating the variance inflation factor. As a rough guideline, a variance inflation factor that is higher than 10 indicates serious multicollinearity. Furthermore, we test whether there is a correlation of the independent variables with the error terms using the correlation test and extracting its p-value. The null hypothesis in this case states that there is no correlation between the variables. Again, the results are stored in the *ols.ass* data frame.

Finally, we check the assumption that the relation between each independent variable and the dependent variable is linear by creating so-called Ceres-plots and saving them inside the *ceresplots* list. In these plots, the terms from each independent variable are added to the residuals and then plotted against the independent variable itself. One can then access via the graphs whether the assumption of linearity is adequate or not.

To be able to assess whether we could achieve a qualitative improvement when transforming the dependent variable, we finally run the for-loop again, this time without excluding outliers and transforming the dependent variable, and saving the results in *regressionresults.first, ols.ass.first*, and *ceresplots.first.* This could have been easily implemented inside the first loop, however, we decided to perform two loops so that each loop is more clearly arranged.

4.2 Regression results

After having performed the automated regressions, we now examine the results that are stored in the data frames *regressionresults* and *regressionresults.first*.

The following two tables contain the estimated parameters, their level of significance, and standard errors. Furthermore, the number of observations, the (Adjusted) R Squared, the Residual Standard Error, and the F-Statistic are reported for each regression. The first one refers to the seven regressions using transformed data while the second table documents the results when we do not transform the dependent variable.

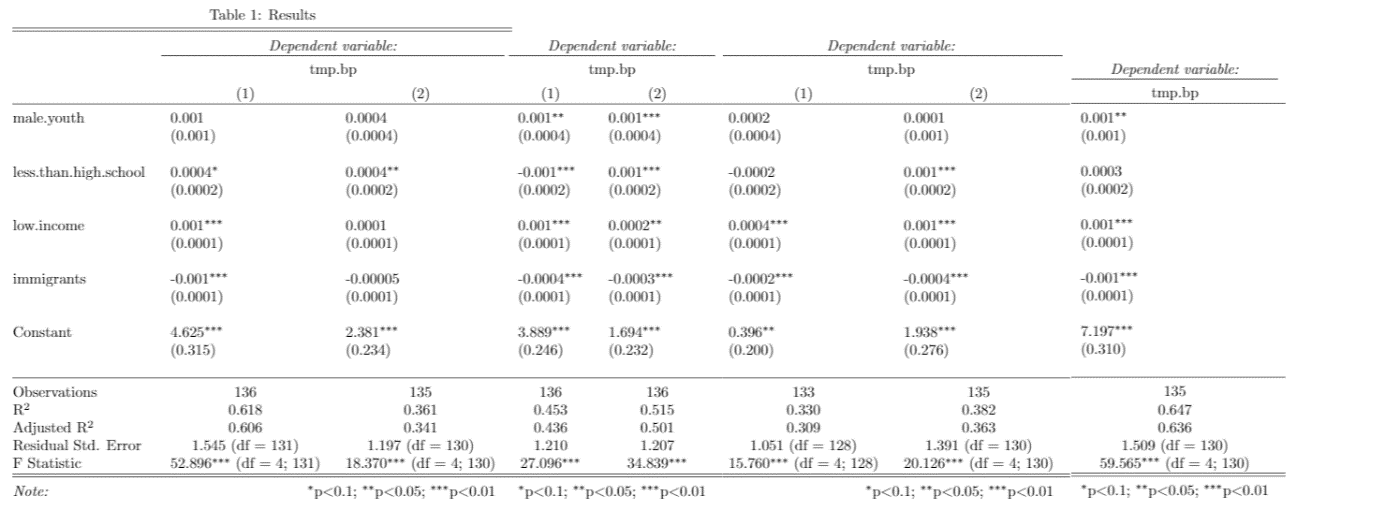


Table 1: Regression results using transformed dependent variables

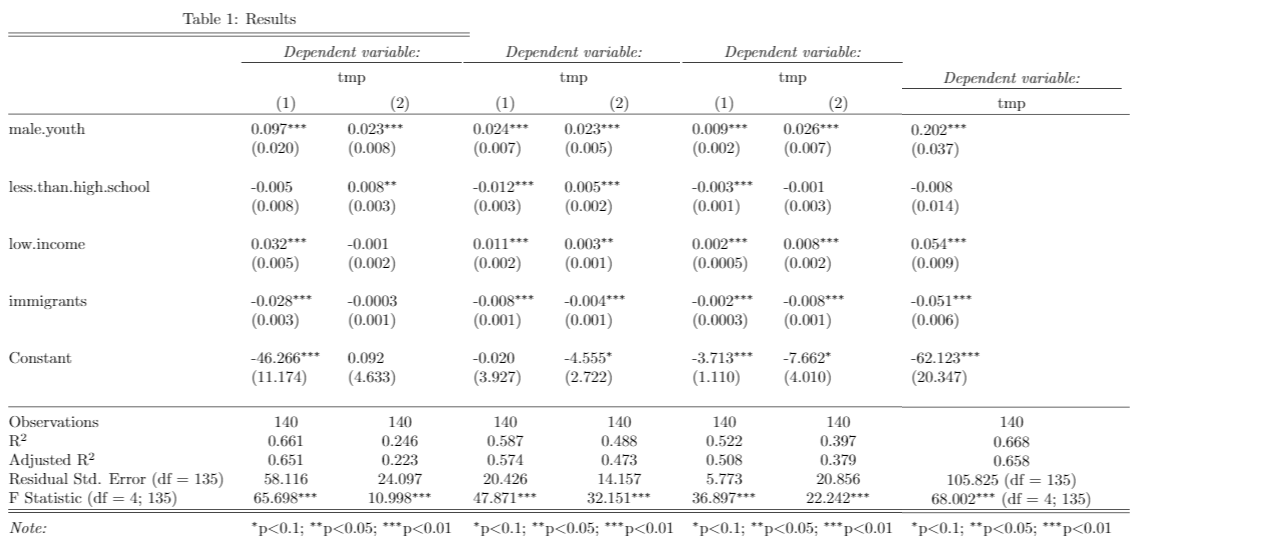


Table 2: Regression results using original dependent variables

Overall, the results do not seem to change significantly between the two tables: While the (Adjusted) R Squared stays roughly the same within each of the seven regressions, we obtain the same amount of significant estimations in each table. Furthermore, with an exception of few cases, the direction of each parameter also stays the same. It therefore does not seem to make a difference whether we use the first or the second table to draw our conclusions about which neighbourhood characteristics affect crime rates in Toronto in which way. As we will see in the next subchapter, some diagnostic tests performed better in the case of transformed data. Therefore, we will now use the first table for our interpretation.

Before interpreting the results, however, we should consider what type of results we are expecting: As all of our four independent variables, the number of young males, the number of people with less than a high school degree, the number of people belonging to the low-income class, and the number of immigrants, are considered as driving forces of crime in general, we should only expect positive estimated parameters.

This is the case for three of our four independent variables: Except of two parameters, *male.youth, less.than.high.school*, and *low.income* all have positive estimates throughout the different crime types. Surprisingly, the variable *immigrants* has statistically significant negative estimates for all regressions. Apparently, immigrants do not increase crime rates as one might expect in the case of the city of Toronto.

Furthermore, some crime types seem to be better described by our regressions than others: While *assault, break.and.enter, robbery, theft.over, drug.arrests* and *total.crime* provide statistically significant parameters a 5% significance level in almost all cases, the parameters for *auto.theft* are not statistically significant in most cases. These differences are also reflected by the (Adjusted) R Squared of each regression.

4.3 Validity of the regression results

Surely, the results of the regressions are only valid if the OLS-assumptions are met. Therefore, we have implemented assumption checks inside the for loop, as described in the former subchapters. The following tables show their results:

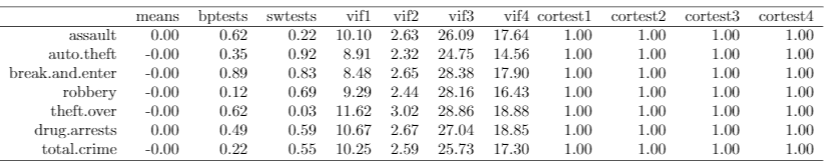


Table 3: OLS-assumptions for transformed dependent variables (*ols.ass*-data frame)

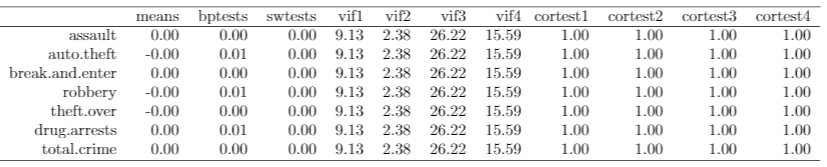


Table 4: OLS-assumptions for original dependent variables (*ols.ass.first*-data frame)

As an intercept was included in all cases, the population mean of the error term is, as expected, equal to zero for all regressions. Regarding the Breusch-Pagan test, the H0 of the error term being homoscedastic cannot be rejected at a level of significance of 5% in any case when using transformed independent variables. Without transformation, however, the H0 is rejected for all regressions. As the Breusch-Pagan-test is sensitive to the assumption of normal distribution and to test the optional OLS-assumption of normally distributed errors, we implemented the Shapiro-Wilk-test. While again the H0 of the error term being normally distributed is rejected in every case for the non-transformed data, it cannot be rejected in the case of transformed data at a level of significance of 5%, except of one case.

Assessing the variance inflation factors which check for imperfect multicollinearity among the independent variables, most of them have a value higher than 10. As mentioned in the former subchapters, this number is a rough guideline to assess the severance of multicollinearity. The assumption of no multicollinearity therefore appears to not be met in this case.

Regarding the correlation test of each independent variable with the error term, the H0 states that the correlation is equal to zero. As we only obtain p-values equal to 1, we cannot reject the H0 for any regression, indicating that the OLS-assumptions are met in these cases.

When looking at the Ceres-plots, we can conclude that the assumption of a linear dependence between the regressors and the dependent variable is valid in all cases. While it is not reasonable to include all Ceres-plots in the paper, we present one plot from the *break.and.enter*-regression with transformed independent variable as an example:

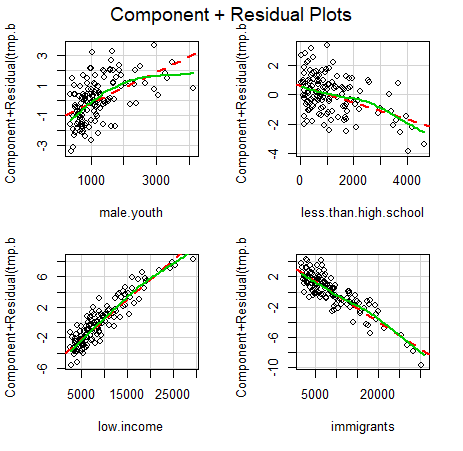


Figure 2: Ceres-plots of the *break.and.enter*-regression with transformed independent variable

While the variables *less.than.high.school, low.income*, and *immigrants* clearly show a linear dependence, there seems to be a slight deviation for male.youth in some areas. However, the assumption of a linear dependence still seems to be reasonable.

All in all, the OLS-assumptions seem to be met in the case of using transformed data. When looking at the Breusch-Pagan test and the Shapiro-Wilk test, we see a qualitative improvement when using transformed data instead of the original data. Although it does not infer with the necessary OLS-assumptions, the existence of imperfect multicollinearity might decrease the quality of regression results as it might lead to unstable parameter estimates and paradoxical results as the overall p-value might be low while all individual p-values being high. However, when we have a look at Table 1 and Table 2 from the previous chapter, these problems do not seem to occur for our regressions. We can therefore conclude that the regressions when using transformed data fulfil the standard OLS-assumptions.