Case Study: Predicting the outcomes of the 2017 Dutch General Elections

Ilse van Beelen, Floor Komen

January 21, 2019

Abstract

For this report demographics of Dutch municipalities are compared with the results of the general Dutch election of 2017. The demographic variables are the *Urban index*, the *fraction of highly educated residents*, *Mean income*, *fraction of 60 plus residents* in a municipality and the factor *Non western*. This factor devides municipalities in the once with less than 5%, 5-10% and more than 10% Non-western residents.The research focust on the results for the party CDA. The goal is to find voting trends per demographic group and to make future predictions for CDA. This goal is approached with two different models. The first is a linear model with a log transformation of the respons. The cross validation resulted in a Mean Prediction Squared Error (MPSE) of 0.058. The second, is a quasi-binomial model with the logit as link function. The found MPSE is 0.0017. It is difficult to say which model fits the data better, because they have different significant variables. However, the logistic model fits the underlying pattern of the data better and has a smaller MPSE. Nonetheless, both models have their limitations and the data displayed a large overdispersion. It is not likely that predictions for future elections can be made with these models.

1. Introduction

1.1 Motivation

For this research, the outcome from the Dutch elections of 2017 and demographical data are combined. All data is collected per municipality and are well maintained and reliable. This will hopefully result in observing voting trends per demographic group. The final goal is to validate the model for making future prediction.

Dutch political parties

The figure above displays the diffences between the political parties in the Netherlands. The Netherlands has a total of 13 parties. This research focusses on only one party. The chosen party should not be to extreme left/right/conservative/progressive and should also be one of the bigger parties. Otherwise, there is not enough data available, making the results less reliable. Therefore, the party Christen-Democratisch Appèl (CDA) is chosen.

In this research the demographics are chosen because of their influence on a municipality level. The expectation is that a municipality with more non-western residents for example votes different than a municipality with less non-western residents. This is the same for the other three demographics. Other demographics are also researched, for example gender, but on a municipality level there is no large difference between the amount of men and women per municipality. Gender is a more interesting demographic to research on an individual level. *The standardized income per municipality* are given in thousands. *Urban index* of a municipality is a database with five categories (0 to 4) per municipality. These five categories are:

- No urbanity (less than 500 addresses per km^2), score 0

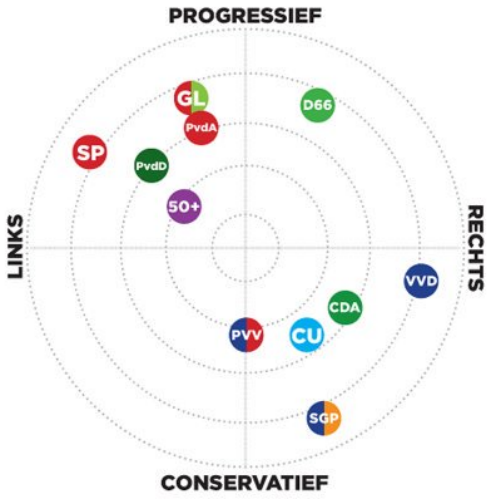

Figure 1: Party landscape

- Little urbanity (500-1000 addresses per km^2), score 1
- Moderate urbanity (1000- 1500 addresses per km^2), score 2
- Strong urbanity (1500-2500 addresses per km^2), score 3
- Really strong urbanity (more than 2500 addresses per km^2), score 4

Per municipality the amount of km^2 per category is given. In the raw dataset the *Non-western residents* is given as an absolute amount per municipality. This is transformed to a fraction and later during the data cleaning divided in three groups.

1.2 Data sources

Electoral data

For the electoral data, the results of the 2017 general election are used. This is the most recent national election and is of the most important election type in the Netherlands. Furthermore, it had a turnup of 81.9%[1]. Therefore, it seems plausible that the data for this election is representative of the political makeup of different municipalities. The raw data is directly downloaded from the official government website[2] This raw dataset is a .csv file with the absolute number of votes for every party in every municipality.

Demographical data

The demographical data is obtained from the CBS, the official Dutch statistical agency.[3]. From the wealth of demographical information available a handful of attributes are picked that are suspected (based on prior research and some gut feeling) to be useful as predictor variables. Five demographical attributes were landed: education grade, average income, age, urbanization and the amount of people with a non-western background. The data downloaded from the CBS site

[1]https://www.kiesraad.nl/actueel/nieuws/2017/03/20/officiele-uitslag-tweede-kamerverkiezing-15-maart-2017
[2]https://data.overheid.nl/data/dataset/verkiezingsuitslag-tweede-kamer-2017
[3]https://opendata.cbs.nl/statline/#/CBS/nl/dataset/70072ned/table?ts=1544803364892

usually had to be transformed to get it in a useful predictor variable format. The specifics of these are described in the next section.

1.3 Data cleaning

An extensive amount of data cleaning had to be done. Below these steps are describes.

Electoral data

The absolute amount of votes for CDA and the total amount of votes per municipalities are kept in the dataset, `CDA_abs` and `Total_abs`, respectively. Information is removed from the other 12 parties. The variable `CDA_frac` is created, this the fraction of votes for CDA per munipicality.

Demographical data

The demographics are downloaded from CBS in multiple csv files. All files are in a long format and are transformed to a wide format using R. The variables *highly educated*, *Non western residens* and *60 plus residents* are in absolute amounts. All are transformed to fractions. These are called, `High_educated_frac`, `Non_west_frac` and `Frac_60plus`, respectively.

`Non_west_frac` does not have a large spread (for most municipalities between 0 to 10%). It is decided to ceate a new variable `Non_west` that divides the variable in three groups:

- Municipalities with less than 5 % non-western residents
- Municipalities with 5-10 % non-western resident
- Municipalities with mre than 10 % non-western residents

Furthermore, the variables `Urban_index` and `Mean_income` did not need to be transformed. All municipalities are in `Muni`. At last, the electoral data and demographic data are combined again. Only the municipality Boxmeer is removed, due to a mistake not all the votes are reported here[4].

```
##      Muni              CDA_frac       Urban_index     High_educated_frac
##  Length:366         Min.   :0.0310   Min.   :0.0000   Min.   :0.1200
##  Class :character   1st Qu.:0.1170   1st Qu.:0.6623   1st Qu.:0.2200
##  Mode  :character   Median :0.1420   Median :1.2305   Median :0.2600
##                     Mean   :0.1528   Mean   :1.4280   Mean   :0.2662
##                     3rd Qu.:0.1820   3rd Qu.:2.1750   3rd Qu.:0.3000
##                     Max.   :0.4200   Max.   :3.7890   Max.   :0.4700
##   Mean_income    Non_west_frac        CDA_abs        Total_abs
##  Min.   :20.80   Min.   :0.01000   Min.   :  421   Min.   :  2727
##  1st Qu.:24.30   1st Qu.:0.03000   1st Qu.: 1737   1st Qu.: 11516
##  Median :25.60   Median :0.05000   Median : 2510   Median : 16915
##  Mean   :25.91   Mean   :0.06574   Mean   : 3254   Mean   : 25162
##  3rd Qu.:27.00   3rd Qu.:0.08000   3rd Qu.: 4023   3rd Qu.: 27087
##  Max.   :41.80   Max.   :0.38000   Max.   :18813   Max.   :440854
##  Non_west  Frac_60plus
##  1:178    Min.   :0.0700
##  2:111    1st Qu.:0.1200
##  3: 77    Median :0.1300
```

[4] https://www.gelderlander.nl/boxmeer/7-600-stemmen-in-boxmeer-niet-meegenomen-in-uitslag-verkiezingen~a063ee9e/

```
##            Mean   :0.1327
##            3rd Qu.:0.1400
##            Max.   :0.1800
```

The summary shows that `CDA_frac` ranges from 0.03 to 0.42. The three groups in `Non_west` are not of equal size. Also `Mean_income` has a large range, from 20.80 to 41.80 times 1000 euros. The final dataset has no NAs.

1.3 Data visualisation

In this part the cleaned data is visualized, so that a good picture can be obtained of the current data. First of all some demographics of data will be showed. In figure 2 of the *parties*, *the urban index*, *the percentage of highly educated residents*, *the mean income*, *The non west residents factor* and *the percentage 60 plus* are plotted.

As visualized in figure 2, most variables are normally distributed. Because of the low values at the x-axis, the CDA, 60 plus percentage and the highly educated densities are above 1. The area beneath the curve sums up to 1, so they are correct. However, the variable `Non_west_frac` shows a large peak around 5%.

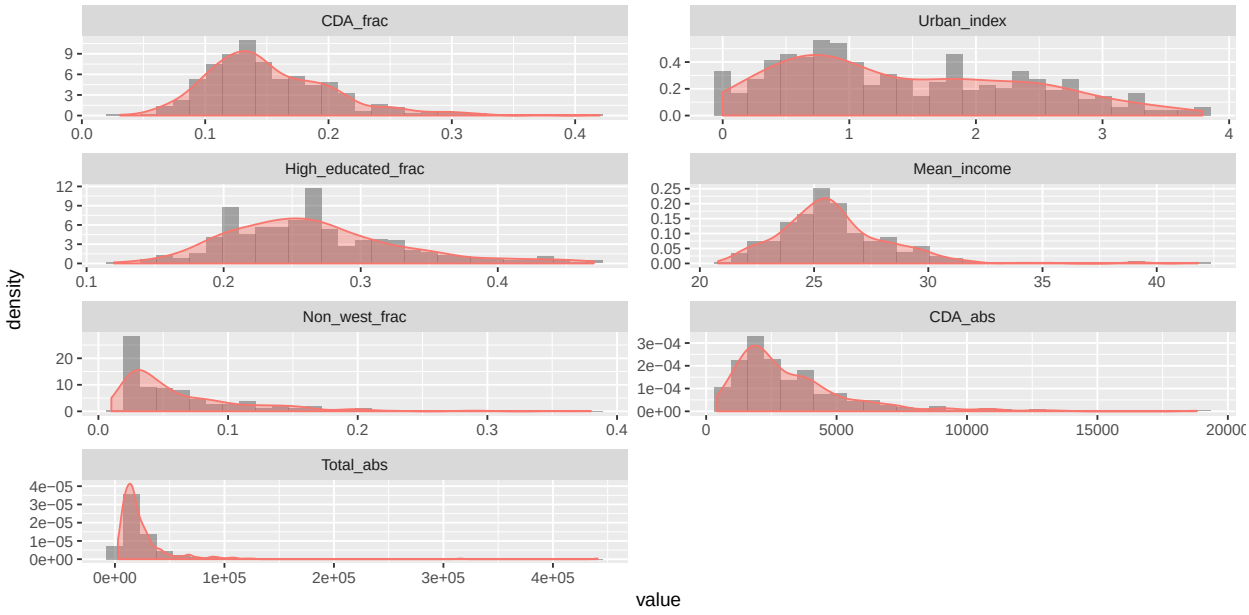

Figure 2: Density plot

Correlation heatmap

In this heatmap (figure 3) the correlation between explanatory and respons variable are showed. The red color means a positive relation, the purple color means a negative relation. The *non_west* variable is not taken into account, because it is a factor and the other variables are continous. `Mean_income` and `High_educated_frac` have the strongest positive correlation and `Urban_index` and `CDA_frac` have the strongest negative correlation. None of the correlations are above 0.8, which is a good sign.

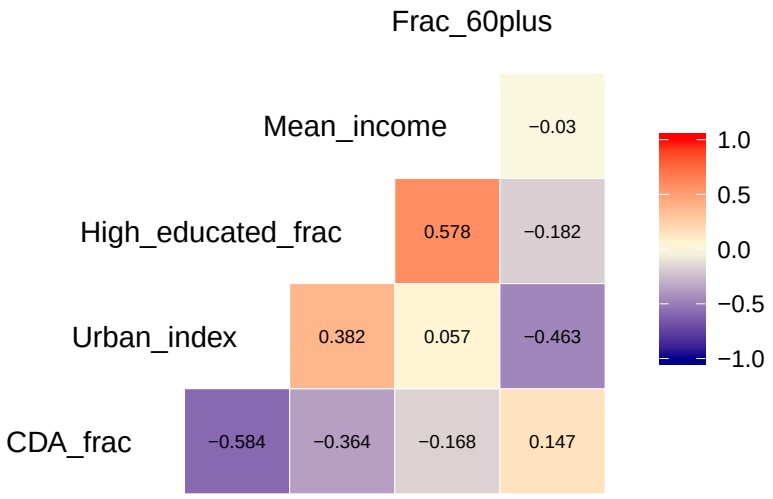

Figure 3: Correlation between explanatory and respons variables

Scatter plots of correlations

The strongest correlations in figure 3 are again visualized in scatterplots. The visible trend is that when the the urbanity index goes up, the votes for CDA goes down. A similar trend is visible for 60 plus residents. The observations of the 60 plus residents seem to follow a horizontal line pattern. This is due to rounding.

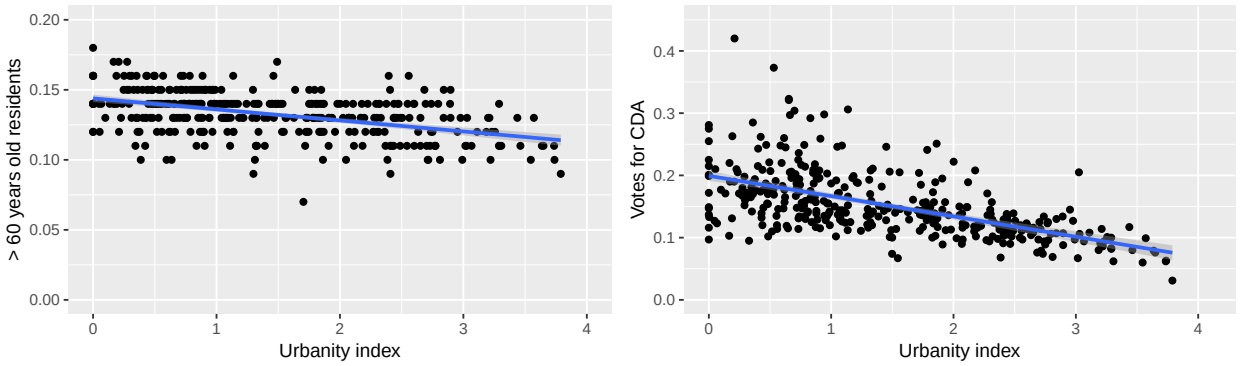

Figure 4: Scatterplots strongest correlations

Figure 5 visualized that when the mean income goes up, the fraction highly educated also goes up. Most municipalities are scattered around an income of 20 to 30 thousand euros, but three municipalities stand out with a mean income around 40 thousand. Also, when the urbanity index increases the fraction highly educated residents increases, but here none of the municipalities stand out.

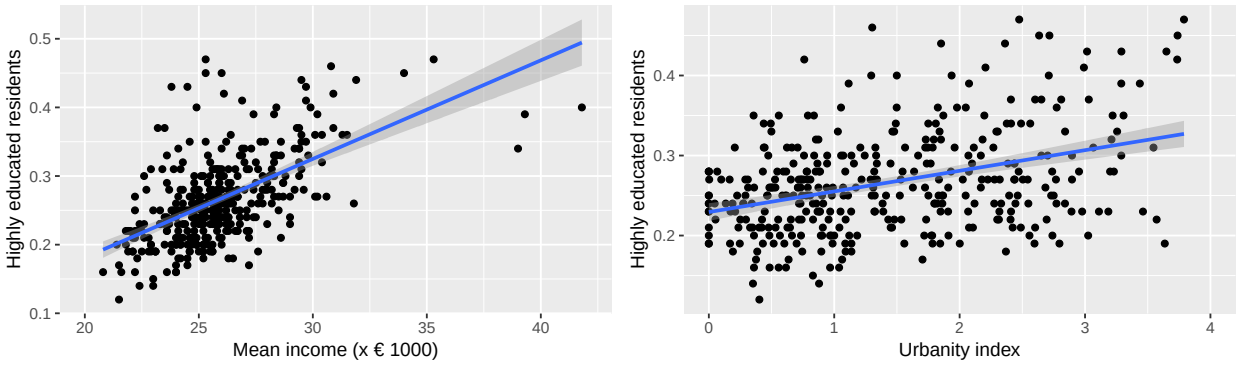

Figure 5: Scatterplot strongest correlations

Multiple boxplots

In this graph boxplots are made, to compare the variable `Non_west` with `Urbanity_index` and `CDA_frac`. A boxplot is a standardized way to display the distribution of data. It gives the minimum, first quartile, median, third quartile and the maximum. If there are any outliers, the boxplot is extended with those. The line within the box is the median, the first and third quartile are the down- and upside of the box, respectively. The length of the box is the Inter Quartile Range (IQR). The minimum and maximum are 1.5X Inter Quartile Range (IQR). Observations further away can be considered outliers.

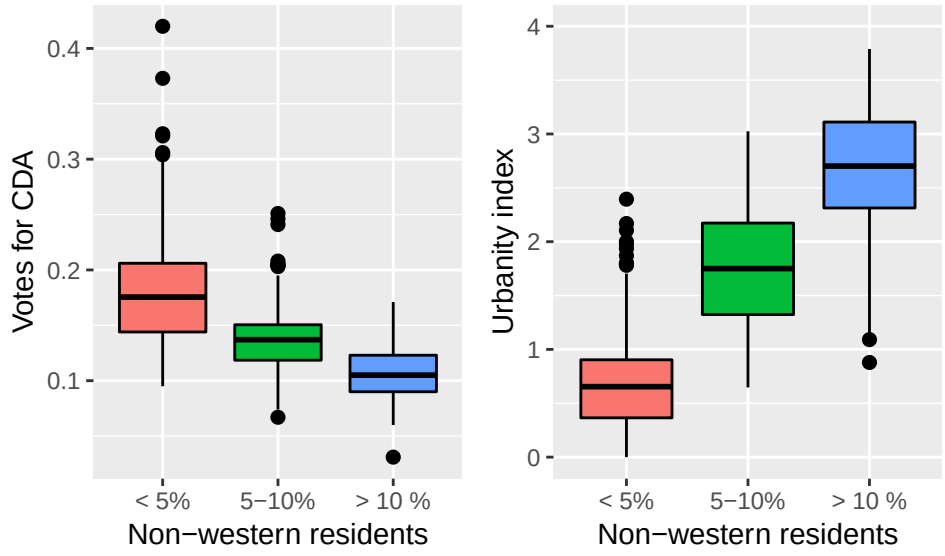

Figure 6: Three boxplots: Boxplot of Non-west against votes for CDA and Urbanity index

Figure 6 visualizes that the Non-western residents tend to live in municipalities with a high Urban Index and vote less for CDA. A few possible outliers are seen in the boxplot for municipalities with less than 5% Non-western residents.

2. Multiple linear regression

In this chapter multiple linear models are generated. The demographics tested in this model are the highly educated fraction in a municipality `High_educated_frac`, the urban index of a municipality `Urban_index`, the mean income of the municipality `Mean_income`, the non-west factor `Non_west` and the fraction that is 60 plus in the municipality `Frac_60plus`. The error assumptions are also discussed. This are assumptions made for the residuals, to check if meet the requirements for correct linear regressions. These assumptions are:

- Linearity: The expected value of the error is zero, $E(\epsilon) = 0$
- Constant variance: The variance of the error is constant
- Normality: The errors are normally distributed
- Indepence: The observations are sampled independently

2.1 First model

The first model will be the model with all the demographics:
$Y_i = \beta_0 + \beta_1 * highEducatedFraction + \beta_2 * UrbanIndex + \beta_3 * MeanIncome + \beta_4 * NonWest2 + \beta_5 * NonWest3 + \beta_6 * Frac60plus + \epsilon_i$

With $i = 1, 2, .., N$ for the number of observations. The outcome of this model is shown below:

	Estimate	Std. Error	t value	Pr()
(Intercept)	0.3381	0.0314	10.78	0.0000
High_educated_frac	-0.0864	0.0454	-1.90	0.0576
Urban_index	-0.0193	0.0041	-4.69	0.0000
Mean_income	-0.0015	0.0011	-1.46	0.1453
Non_west2	-0.0223	0.0065	-3.45	0.0006
Non_west3	-0.0455	0.0095	-4.77	0.0000
Frac_60plus	-0.5904	0.1494	-3.95	0.0001

The first model is the full model, `High_educated_frac` and `Mean_income` do not have a significant t-value. Before any conclusions are made, the assumptions are checked via plots and the VIF is checked. The VIF is the Variation Inflation Factor, it implies if there is multicollinearity between variables. The formula for VIF is $1/(1 - R^2)$ and the thresholdvalue is 10. Meaning that values above 10 give signs of multicollinearity. As shown below none of the values are above 10, so no signs of collinearity.

```
## High_educated_frac        Urban_index        Mean_income
##           1.871032           3.383149           1.658015
##          Non_west2          Non_west3        Frac_60plus
##           1.974537           3.361734           1.289979
```

```
## [1]  74 298
```

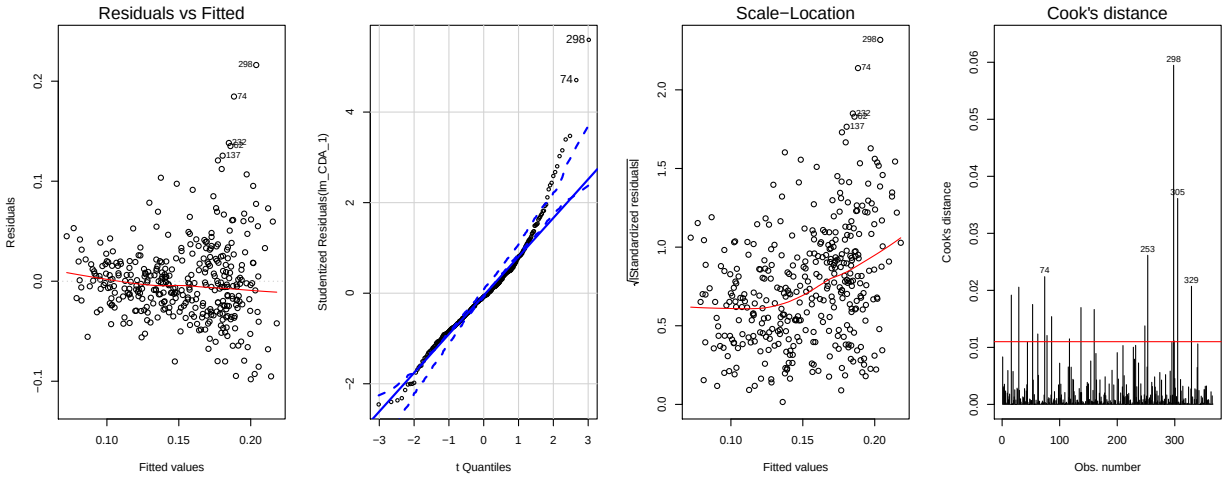

Figure 7: assumptions first model

In figure 7 four diagnostics plots are shown. The first plot (Residuals vs Fitted) shows that the residuals have a 'loudspeaker pattern', the variance of the residuals tends to increase with an increase of the fitted value. Because of this, a BoxCox graph is consulted. This graph suggests a transformation for the response. The BoxCox in figure 8 has a 95% Confidence interval located around the 0. Hence, a log transformation is suggested.

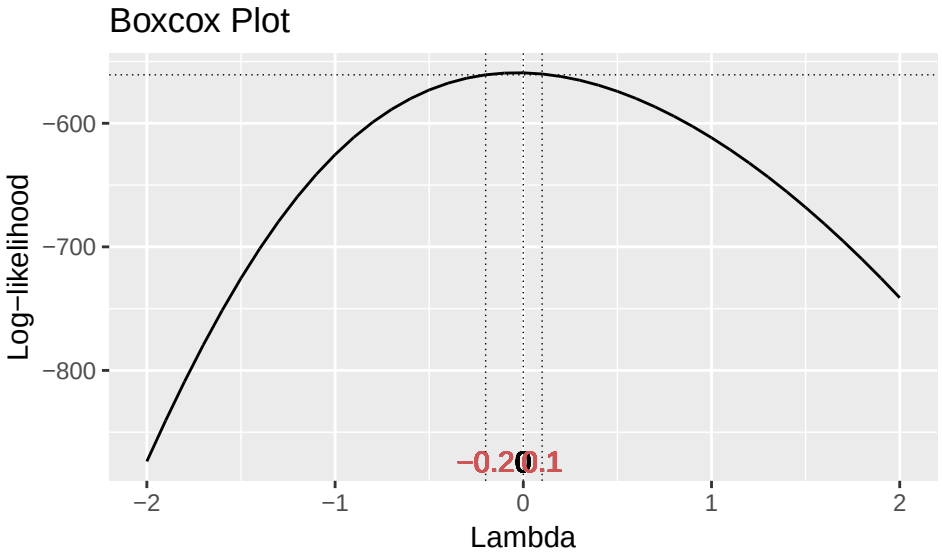

Figure 8: BoxCox first model

2.2 Second model

In the second model the response variable will be ln transformed. So the new model will be:
$ln(Y_i) = \beta_0 + \beta_1 \cdot highheducatedfraction + \beta_2 \cdot Urbanindex + \beta_3 \cdot Meanincome + \beta_4 \cdot Nonwest2 + \beta_5 \cdot Nonwest3 + \beta_6 \cdot Frac60plus + \epsilon_i$

	Estimate	Std. Error	t value	Pr()
(Intercept)	-0.9944	0.1882	-5.28	0.0000
High_educated_frac	-0.8808	0.2723	-3.24	0.0013
Urban_index	-0.1388	0.0247	-5.62	0.0000
Mean_income	-0.0024	0.0064	-0.38	0.7042
Non_west2	-0.0991	0.0389	-2.55	0.0112
Non_west3	-0.2763	0.0572	-4.83	0.0000
Frac_60plus	-2.6940	0.8965	-3.01	0.0028

```
## [1]  16 237
```

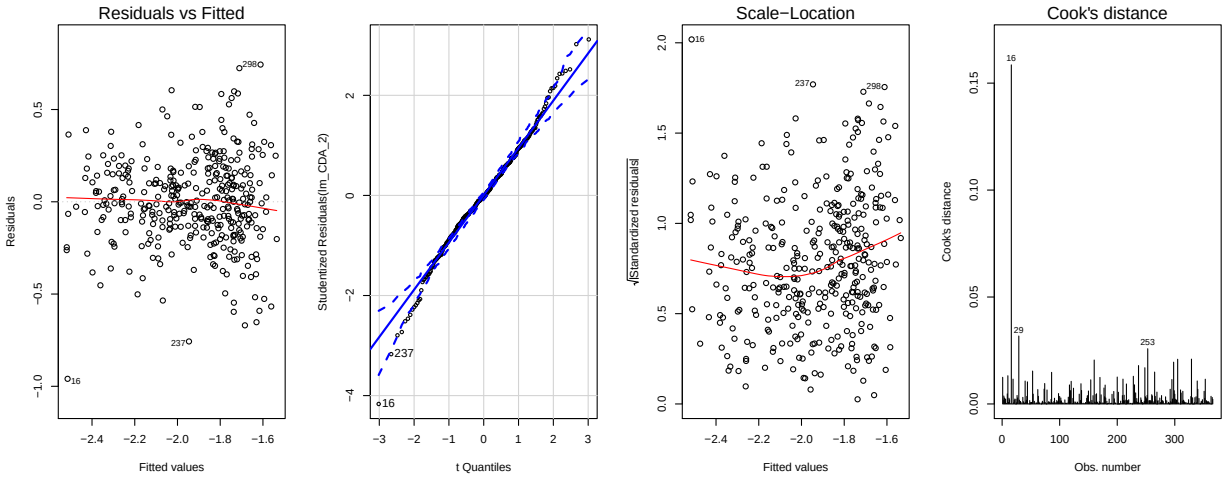

Figure 9: Diagnostics plot second model

The plots in figure 9 show one large outlier, the municipality Amsterdam (obs 16). Amsterdams value for the cooks distance goes far above the cutoff value, $4/(369-5-1) = 0.011$. It is also outside the (-3,3) range with the studentized residuals. This is strong evidence that this municipality is an outlier and should be removed.

For the second model without Amsterdam, a step function is used. This step function uses the Aikaike Information Criterion (AIC) for backward elimination and shows which variable should be removed to decrease the AIC. The formula for AIC is $AIC = -2log(likelihood) + 2p$, p is the number of parameters in the model. The variables that are left are the variables used in the final model.

```
## Start:  AIC=-1041.5
## log(CDA_frac) ~ High_educated_frac + Urban_index + Mean_income +
##     Non_west + Frac_60plus
##
##                      Df Sum of Sq    RSS     AIC
## - Mean_income         1   0.04208 20.291 -1042.7
## <none>                            20.249 -1041.5
## - High_educated_frac  1   0.36195 20.611 -1037.0
```

```
## - Frac_60plus         1   0.67266 20.922 -1031.6
## - Non_west            2   1.54236 21.792 -1018.7
## - Urban_index         1   1.72696 21.976 -1013.6
##
## Step:  AIC=-1042.74
## log(CDA_frac) ~ High_educated_frac + Urban_index + Non_west +
##     Frac_60plus
##
##                       Df Sum of Sq    RSS     AIC
## <none>                             20.291 -1042.7
## - Frac_60plus         1   0.66435 20.956 -1033.0
## - High_educated_frac  1   0.85427 21.146 -1029.7
## - Non_west            2   1.51164 21.803 -1020.5
## - Urban_index         1   1.68687 21.978 -1015.6
##
## Call:
## lm(formula = log(CDA_frac) ~ High_educated_frac + Urban_index +
##     Non_west + Frac_60plus, data = Data_CDA[-16, ])
##
## Coefficients:
##        (Intercept)  High_educated_frac         Urban_index
##            -1.0298             -0.8277             -0.1311
##          Non_west2           Non_west3         Frac_60plus
##            -0.1141             -0.2871             -3.0168
```

2.3 Final model

The backward elimination resulted in the final model.
$ln(Y_i) = \beta_0 + \beta_1 \cdot higheducatedfraction + \beta_2 \cdot Urbanindex + \beta_4 \cdot Nonwest2 + \beta_5 \cdot Nonwest3 + \beta_6 \cdot Frac60plus + \epsilon_i$

The coëfficients are given in the table below

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.0298	0.1365	-7.54	0.0000
High_educated_frac	-0.8277	0.2129	-3.89	0.0001
Urban_index	-0.1311	0.0240	-5.46	0.0000
Non_west2	-0.1141	0.0378	-3.02	0.0027
Non_west3	-0.2871	0.0559	-5.13	0.0000
Frac_60plus	-3.0168	0.8799	-3.43	0.0007

```
## 237 298
## 236 297
```

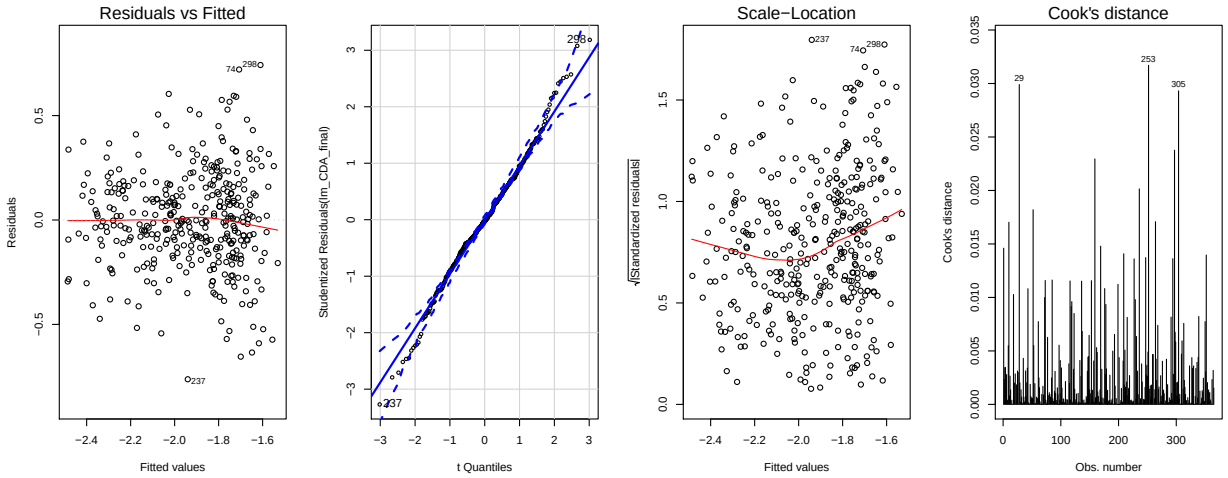

Figure 10: Assumptions second model

The estimates for the predictors are filled in the model and the following results are obtained:

$ln(Y_i) = -1.0298 - 0.8277 \cdot HighEducatedfraction - 0.1311 \cdot UrbanIndex - 0.1141 \cdot NonWest2 - 0.2871 \cdot NonWest3 - 3.0168 \cdot Frac60plus + \epsilon_i$

All the coëfficients are negative, but because the fitted value is a log value the response will be positive.

2.4 Cross validation

To tell something about the prediction possibilities of the model, cross validation is done. Cross validation helps to discover how well the model predicts on average. Cross validation estimates the Mean Squared Prediction Error (MPSE) of a model.

The following steps are done. First 5k-folds are made, meaning that the data is divided in five folds. Next, a loss function is created This function finds the square of the difference between the observed value Y_i predicted value $\widehat{Y}_i$. Afterwards the sum is taken and divided by the length of the folds, to correct for the length of each fold. The formula below shows the loss function:

$$\frac{(Y_i - \widehat{Y}_i)^2}{length(fold)}$$

Four of the five k-folds are used to train the data, the other one is the test data. The model is fitted on the training data and afterwards it tries to fit on the test data, to see if it predicts closely. This is done 5 times, every time another fold is the test data. As is shown below, the MSPE is 0.0582.

```
lm_CDA_final <- lm(log(CDA_frac) ~ High_educated_frac + Urban_index + Non_west +
    Frac_60plus, data = Data_CDA[-16, ])
K <- 5
index <- rep(1:K, floor(nrow(Data_CDA)/K) + 1)[1:nrow(Data_CDA)]
fold.index <- sample(index)
Loss <- function(x, y) {
    sum((x - y)^2)/length(x)
```

```
}
loss <- numeric(K)
for (k in 1:K) {
    training <- Data_CDA[fold.index != k, ]
    validation <- Data_CDA[fold.index == k, ]
    training.fit <- lm_CDA_final
    validation.predict <- predict(training.fit, newdata = validation, type = "response")
    loss[k] <- Loss(log(validation$CDA_frac), validation.predict)
}
mean(loss)
```

```
## [1] 0.0581487
```

3. Logistic regression

The raw respons variable is the absolute amount of residents per municipality that voted for CDA. For linear regresssion, this variable is transformed to a fraction. However, the absolute total amount of votes per municipality is also available. Therefore, a binomial model would be a better fit to the data. A second model is developed that uses the logit as link function to transform the range of the respons. The choice for the logit was easily made. Because the inverse of the logit is directly interpretable as the log-odds ratio and this link displays the underlaying pattern of the data best. Below, the formula for the link function:

$\eta = log(\frac{\theta}{1-\theta}) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_n \cdot x_n$

The ratio is the log odds that $Y_i = 1$ (the odds of voting for CDA).

Also for logistics regression diagnostic plots are needed to visualise the deviance/pearson residuals and search for outliers. Most of the diagnostics from the linear model extend relatively straighforward to logistic regression. However, leverages are no longer just a function of the explanatory variable, but also depend on the respons due to the iterated weighted least squares. Furthermore, θ can never be zero or one. Fortunately, this was not the case for any of the observations in this dataset.

3.1 First model

Again, the first model is the full model. Stepwise backward elimination is used to find the optimal model. Below, the formula for the full model:

$log(\frac{\theta_i}{1-\theta_i}) = \beta_0 + \beta_1 \cdot UrbanIndex + \beta_2 \cdot HighlyEducatedFraction + \beta_3 \cdot MeanIncome + \beta_4 \cdot NonWest + \beta_5 \cdot Fraction60Plus + \epsilon_i$

With $i = 1, 2, .., N$ for the number of observations.

Below the summary of this model:

The summary shows that the all the variables are very significant and have small standard errors. The full model has 359 degrees of freedom and it is expected that the residual deviance is roughly equivalent. However, the residual deviance is far above this value. These are strong indications that this model suffers from overdispersion. This assumption seems reasonable, because there is a very large variance in how many residents per municapality voted for CDA. In some municipalities only

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.0560	0.0157	-67.29	0.0000
Urban_index	-0.1934	0.0020	-94.91	0.0000
High_educated_frac	-2.1028	0.0200	-105.38	0.0000
Mean_income	0.0156	0.0005	29.32	0.0000
Non_west2	-0.0563	0.0032	-17.81	0.0000
Non_west3	-0.2593	0.0046	-56.26	0.0000
Frac_60plus	-1.3424	0.0737	-18.20	0.0000

3% voted for CDA, while it others nearly 50% voted for CDA. It is concluded that a quasi-binomial would fit the data better.

3.2 Second model

The second model has still all the variables, but is fitted to a quasi-binomial. Below the output of the summary is visualized:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.0560	0.2501	-4.22	0.0000
Urban_index	-0.1934	0.0325	-5.95	0.0000
High_educated_frac	-2.1028	0.3180	-6.61	0.0000
Mean_income	0.0156	0.0085	1.84	0.0667
Non_west2	-0.0563	0.0504	-1.12	0.2647
Non_west3	-0.2593	0.0735	-3.53	0.0005
Frac_60plus	-1.3424	1.1753	-1.14	0.2541

By applying a quasi binomial model, a dispersion parameter ϕ is included, resulting in larger standard errors and less significant p-values. ϕ is estimated on the data at 254.0441. Furthermore, the null deviance is estimated at 247,550 with 365 df and the residual at 89969 with 359 df. The variables `Frac_60plus`, `Mean_income` and factor level `Non_west2` are no longer significant.

No goodness of fit test is possible because of the free dispersion parameter. The decision to remove variables is done based on the lowest F-test.

```
##        Urban_index High_educated_frac        Mean_income
##       0.0013605627       0.0005943712       0.0006876039
##          Non_west2          Non_west3        Frac_60plus
##       0.0007725368       0.0012914703       0.0005161910
```

According to the F-test `Frac_60plus` should be removed. This variable has a F-value of 1.32 and a corresponding p-value of 0.25. The values for the VIF are all very low, meaning there is barely collinearity between the explanatory variables.

At last, the residuals and cook's distance are visualized

	Df	Deviance	F value	Pr(>F)
<none>		89968.85		
Urban_index	1	99052.41	36.25	0.0000
High_educated_frac	1	101150.95	44.62	0.0000
Mean_income	1	90820.59	3.40	0.0661
Non_west	2	94111.95	8.27	0.0003
Frac_60plus	1	90300.07	1.32	0.2511

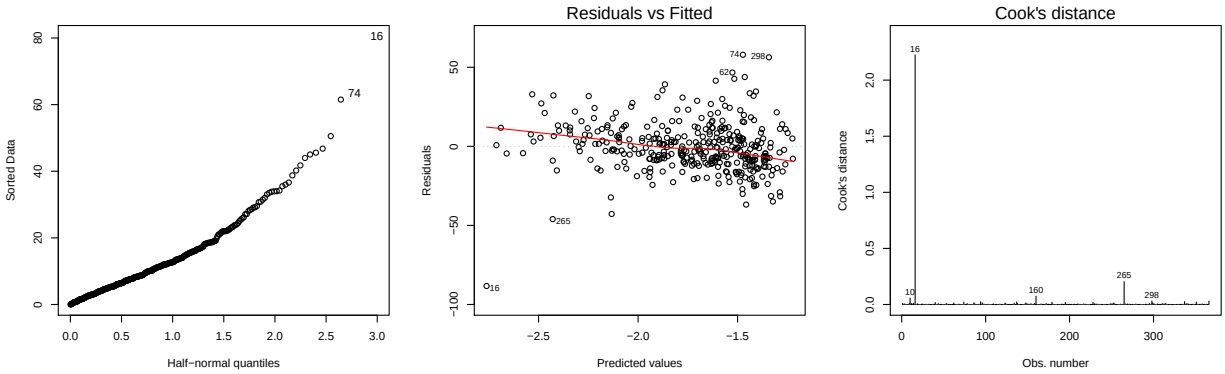

Figure 11: Diagnostics first quasi-binomial model

The left plot of figure 11 visualizes the half-normal quantiles against the pearson residuals. Ideally, these residuals would not be greater than 3. However, this plot shows residuals even up to 80. The middle plot displays the predicted values against the deviance resiuals. Also here a large spread of the residuals is observed and the variance tends to increase with an increase of the fitted value. The right plots visualizes the cook's distance, which can identify influential observations. Observation 16 is an outlier, because it is very influential and stands out from any pattern in the residual plots. Furthermore, Dinkelland (obs 74) and Rotterdam (obs 265) are also influential. Amsterdam is the municipality with the lowest percentage of CDA votes and Dinkelland has the highest percentage of CDA votes.

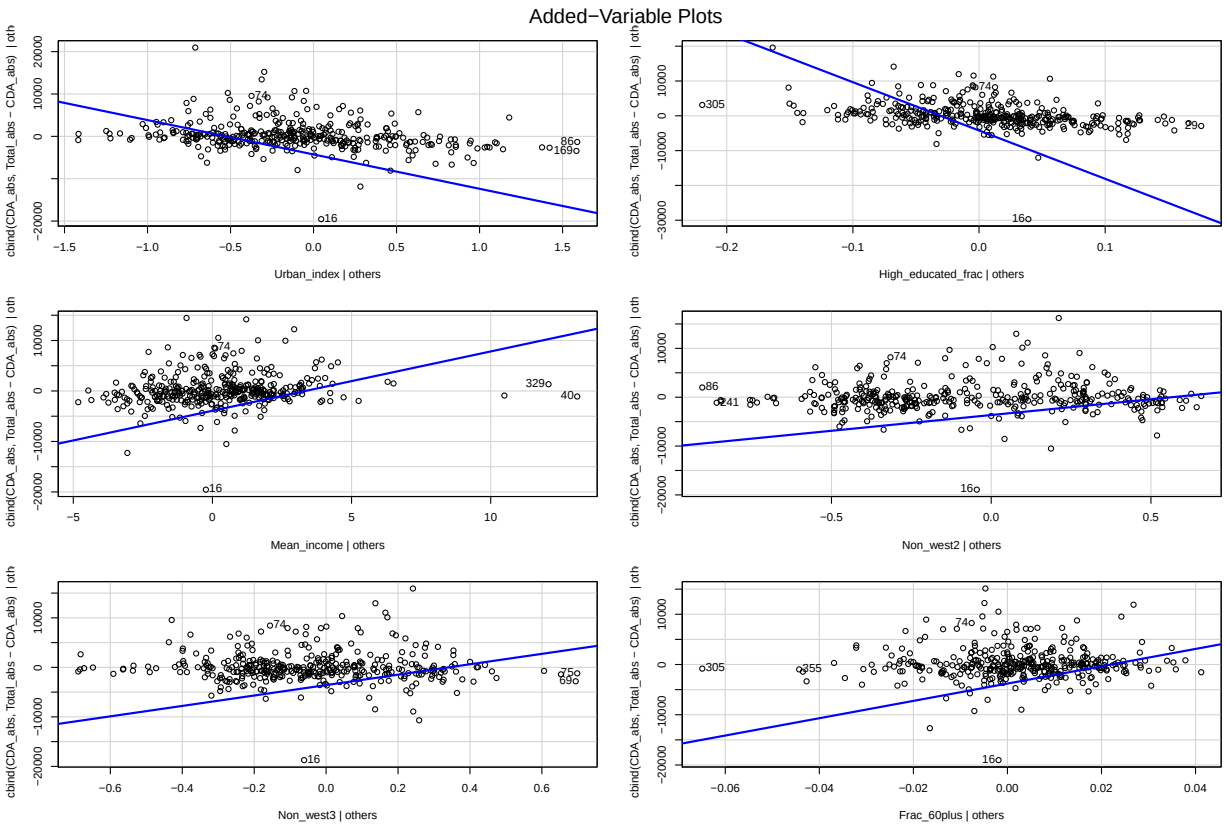

Figure 12: AvPlots first quasi-binomial model

Figure 12 help to interpret the partial regression coefficients in a model when the other variables are held constant. The partial regression line is highly influenced by observation 16 again. The blue lines do not represent the data well at the moment.

3.3 Third model

For this model the variable `Frac_60plus` is removed, because it had the lowest F-value. Furthermore, the observations 16 (Amsterdam) and 265 (Dinkelland) are removed. These influence the partial regression coefficients greatly and have large residuals and cook's distances. These steps were originally done in two, but are combined for this report.

Below the summary output from this third model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.0878	0.1802	-6.04	0.0000
Urban_index	-0.1352	0.0303	-4.46	0.0000
High_educated_frac	-1.2202	0.3126	-3.90	0.0001
Mean_income	-0.0004	0.0082	-0.05	0.9583
Non_west2	-0.1232	0.0479	-2.57	0.0105
Non_west3	-0.3447	0.0691	-4.99	0.0000

By removing observation 16 and 265, the factor `Non_west2` has become significant. The null

deviance has dropped to 173,842 with 363 df and the residual deviance has decreased slightly to 76,966 with 358 df. ϕ has increased slightly to 221.709.

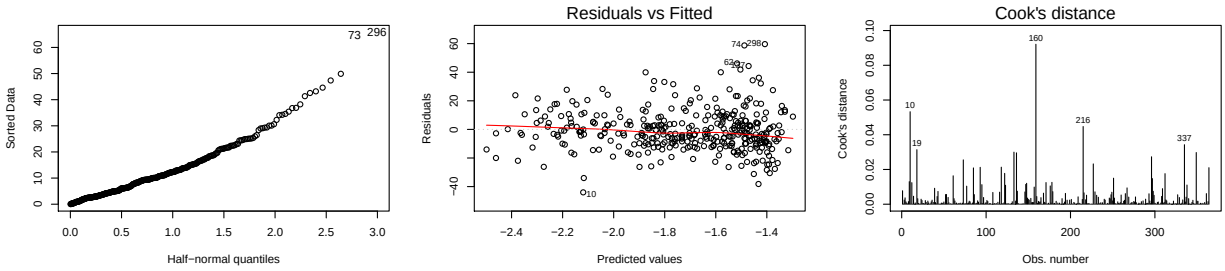

Figure 13: Diagnostics third quasi-binomial model

The plot left still displays very large pearson residuals. The plot in the middle still visualized that the deviance residuals tend to increase when the predicted values increase. The cook's distance no longer displays highly influential observations.

	Df	Deviance	F value	Pr(>F)
<none>		76965.83		
Urban_index	1	81395.63	20.60	0.0000
High_educated_frac	1	80361.41	15.79	0.0001
Mean_income	1	76966.44	0.00	0.9577
Non_west	2	82994.45	14.02	0.0000

According to the F-test `Mean_income` should be removed as well, because the F-value is below 1 and the corresponding p-value is 0.96.

3.4 Final model

The final model is reached after dropping the variable `Mean_income`. It's formula is as follows:

$logit(\theta_i) = -1.09 - 0.13 \cdot UrbanIndex - 1.23 \cdot HighlyEducatedFraction - 0.12 \cdot NonWest2 - 0.34 \cdot NonWest3 + \epsilon_i$

The summary output is as follows:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.0965	0.0653	-16.80	0.0000
Urban_index	-0.1350	0.0300	-4.50	0.0000
High_educated_frac	-1.2297	0.2541	-4.84	0.0000
Non_west2	-0.1235	0.0477	-2.59	0.0100
Non_west3	-0.3444	0.0688	-5.01	0.0000

According to the F-test all variables now significantly contribute to the model. The null deviance is still 173,842 with 363 df and the residual deviance has slightly decreased to 76,966 with 359 df. ϕ is estimated at 221.1.

	Df	Deviance	F value	Pr(>F)
<none>		76966.44		
Urban_index	1	81470.68	21.01	0.0000
High_educated_frac	1	82189.93	24.36	0.0000
Non_west	2	83076.91	14.25	0.0000

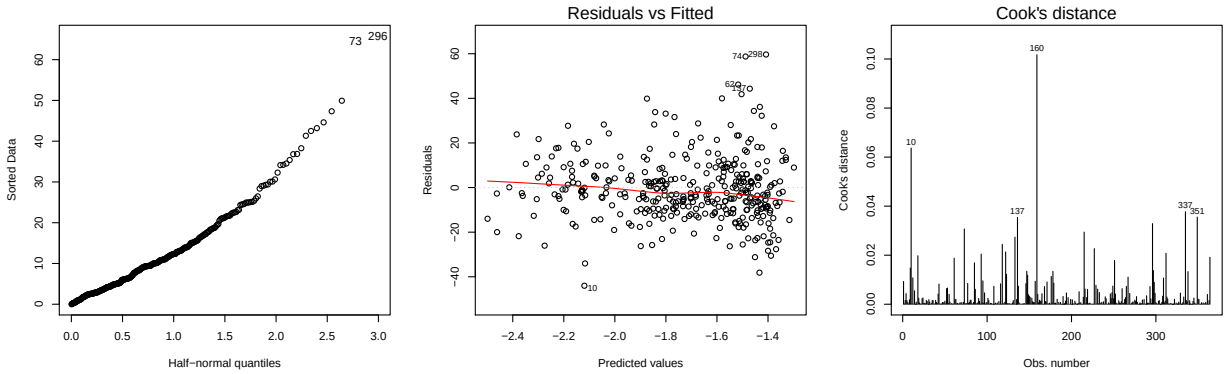

Figure 14: Diagnostics final quasi-binomial model

Figure 14 displays that there is still a large spread of both the pearson (left plot) and deviance (middle plot) residuals. Furthermore, there is non-constant errror variance. The cook's distance does not display very influential observations.

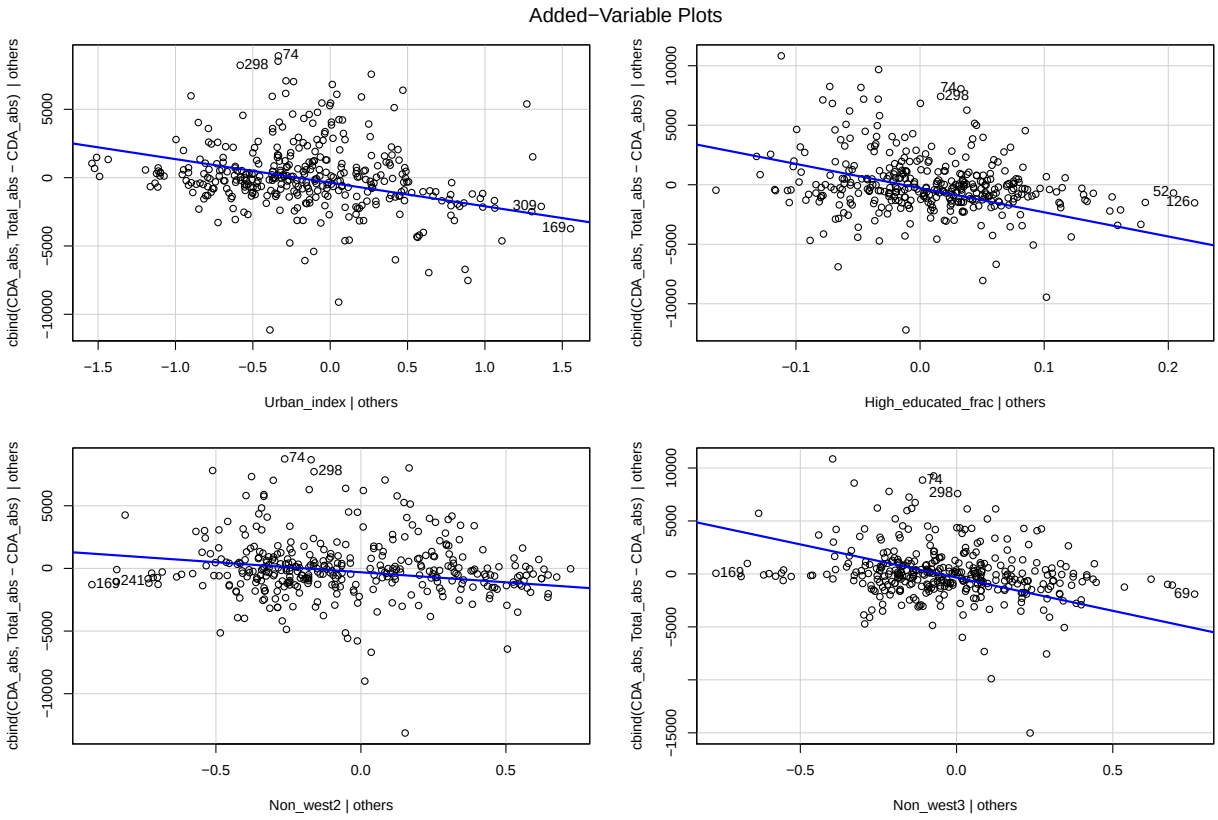

Figure 15: Diagnostics final quasi-binomial model

```
##         Urban_index High_educated_frac          Non_west2
##        0.0012896670       0.0004230904       0.0007928046
##          Non_west3
##        0.0012734699
```

After removing two outliers, the partial regression coefficients represent the data much better. There are no strong correlations between the explanatory variables and the respons. The low VIF values also indicate this.

3.5 Cross validation

At last, the logistic model is validated by k-fold validation. This dataset has 366 observations, therefore K = 5 is enough. The code for cross validation is similar to the one used for the linear model. Therefore, only the output is presented here and not the code.

```
## [1] 0.001707315
```

The cross validation results in a Prediction Means Squared Error (PMSE) of 0.0017.

4. Discussion

4.1 Multiple Linear regression

Because the fitted values are transformed to a log form, it is also possible to raise the coëfficients to a exponential power. The final model obtained then is:

$$Y_i = e^{-1.0298 - 0.8277 \cdot HighEducatedFraction - 0.1311 \cdot UrbanIndex - 0.1141 \cdot Nonwest2 - 0.2871 \cdot Nonwest3 - 3.0168 \cdot Frac60plus + \epsilon_i}$$

Per variable the influence will be discussed. The slope will start at point $exp(-1.0298)$, this is equal to 0.357. This means if all the other demographic variables are zero, the fitted value will be equal to the intercept, so equal to 0.357. This not a possible outcome, because a Municipality with all these demographics equal to zero is not a reality.

For the other coefficients it is a bit harder to predict there influence, because of the log transformation and the different range the variables have. For example, the *Urban Index* has a 0-3.8 range and *Highly educated* has a 0.12-0.47 range in this data set. But still some remarks can be made about the slope of the model. The *Fraction 60 plus* has the lowest marginal impact on the slope, if all variables are hold constant and *Frac 60 plus* changes with one unit, then the exponent changes with -3.02. The *Non west2* has the highest impact on the slope, because the coefficient is the lowest.

The outcome of the cross validation for this model is 0.0582. Hence, the predicted mean squared difference between the fitted and predicted value is 0.0582, which is very close to 0.

There are some limitations for this model, because the response variable is a fraction and will never be larger than one, theoratically a logistic regression would fit the data better. Also some assumptions are violated. In the fitted vs residual graph it is visible that the variance is not equally spread, there is a small "loudspeaker pattern". But because the fitted values are log transformed, it is not really possible to adapt this any further. Also there are two municipalities that fall outside the [-3,3] range in the normality plot, but because they are still in the 95% envelope the decision is made to not delete these municipalities.

4.2 Logistic regression

For the final model two variables and two outliers are removed. The coefficients of the model are on the log-odds scale and need to be transformed before interpretating them. Each estimated coefficient is the expected change in the log odds of voting for CDA for a unit increase in the corresponding explanatory variable holding the other explanatory variables constant.

The coefficient for *Urban Index* is the difference in the log odds. In other words, the expected change in log odds of voting for CDA is -0.13. This can be transformed to the odds: $exp(-0.13) = 0.88$. The odds of voting for CDA decrease with roughly 12% if the Urbanity increases with one unit. The Urbanity index has a range from 0 to 4, so this variable has a large influence. For example, when comparing Terschelling (Urbanity index of 0) and Leiden (Urbanity index of 3.7), the expected decrease is 40.7%.

An similar calculation can be done for one unit increase of *Non west*. When *Non west* increases from 1 (the reference level) to 2 and the other explanatory variables are held constant, then the log odds of voting for CDA is -0.12. Transforming this to the odds results in a decrease of 11%, roughly. And when when comparing factor level 1 to 3, the log odds is -0.34. This results in the odds of $exp(-0.34) = 0.71$, a decrease of 29%. This is not a simple duplication of level 2.

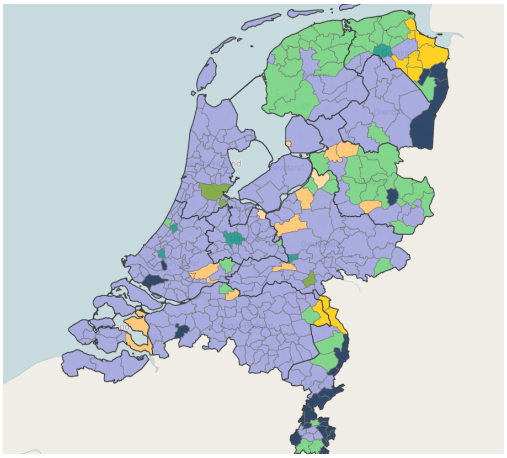

Figure 16: Results elections with biggest parties. CDA is displayed in light green

According to the model, holding *Highly educated* and *Non west* constant, the odds of voting for CDA if the Urban index increases with 1 unit is $exp(-0.16) = 0.87$. In percentage change does this mean that the odds decrease with 12.6%.

At last, the odds of voting for CDA if *Highly educated* increases with 1 unit is $exp(-1.23) = 0.29$. In percentage change does this mean that the odds decrease with 70.8%. This is a very large decrease, but can be explained. This variable is a fraction and has a range from 0 - 1. An increase of one unit is not likely to happen.

In summary, municipalities in rural areas, with smaller amounts of Non-western and Highly educated residents tend to vote for CDA. Below the formula with the odds ratio is shown:

$(\frac{\theta_i}{1-\theta_i}) = 0.33 + 0.88 \cdot UrbanIndex + 0.29 \cdot HighlyEducatedFraction - 0.89 \cdot NonWest2 + 0.71 \cdot NonWest3 + \epsilon_i$

As already said, there is still dispersion, even after using a quasi binomial model. A possible explanation can be clustering of observations. Municipalities close to each other will probably vote similar. Figure 16 shows that the municipalites where CDA (shown in light green) is the biggest party are clustered together. CDA is the biggest party in large parts of Friesland, Groningen and Drente. The municipalities located here probably have a similar population.

Another explanation for the dispersion is the large variation. In some municipalities only 3% voted for CDA, while in other almost 50%. This large variation is hard to modeled in a binomial. By using the log odds, votes are distributed in two groups: voted for CDA or not voted for CDA. However, with 13 political parties is it hard to distinguish two groups. Because it is not possible take to into account which parties are similar to CDA.

Even though, the cross validation resulted in a PMSE of 0.0017, it is not likely that this model can make future predictions. This because of the large overdispersion and non-constant error variance.

4.3 Further research

Both of the models have different significant variables. This makes it even more difficult to determine which model is better fitting. However, the logistic regression fits the underlying pattern of the data better and has a smaller MSPE. Further research into the correctness of the models should

be done. Another topic that can be researched in further research is the influence of demographics on districts of municipalities instead of the whole municipialities. Right now the developed models nullified the differences in demographics between areas in a municipality.