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

Ilse van Beelen, Floor Komen

January 18, 2019

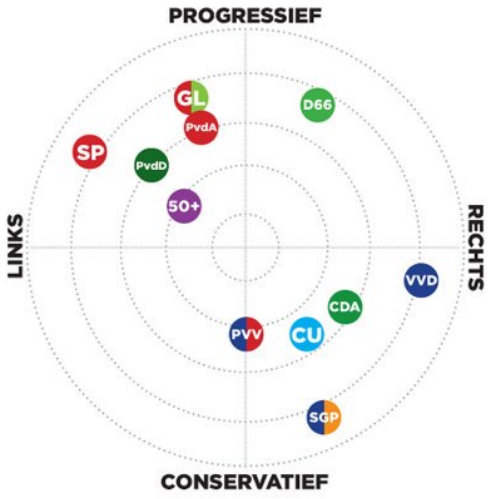
Abstract

Put the abstract over here

1. Introduction

1.1 Motivation

For this case study, it was decided to combine the outcome from the Dutch elections of 2017 and demographic data. Both are collected per municipality and are well maintained and reliable. This makes A lot of information is available for both in the Netherlands. This will hopefully result in observing voting trends per demographic group. The final goal is to validate the model for making futyre predictions.



Dutch political parties

This figure displays the diffences between the political parties in the Netherlands. The Netherlands has a total of 13 parties. This investigation focusses on only one party. This 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, party CDA is chosen.

In this research the above described 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 two 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. So that is a more interesting demographic to research on an individual level. *The*

standardized income per municipality are given in thousands. *the urban index of a municipality* is a database with five categories per municipality. These five categories are:

- Really strong urbanity (more than 2500 addresses per km^2)
- Strong urbanity (1500-2500 addresses per km^2)
- Moderate urbanity (1000- 1500 addresses per km^2)
- Little urbanity (500-1000 addresses per km^2)
- No urbanity (less than 500 addresses per km^2)

Per municipality the amount of km^2 per category is given. The *non-west residents per municipality* is given in an amount per municipality, also the total amount of residents is given per municipality.

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%. Therefore, it seems plausible that the data for this election is representative of the political makeup of different municipalities. We downloaded the raw data directly from the official government source.[1] This contained a .csv file with the raw number of votes for every party in every municipality.

Demographical data
We got our demographical data from the CBS, the official Dutch statistical agency.[2] From the wealth of demographical information available we picked a handful of attributes that we suspected (based on prior research and some gut feeling) to be useful as predictor variables. We landed on five demographical attributes: education grade, average income, age, urbanization and the amount of people with a non-western background. Note that the data we downloaded from the CBS site 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 and a small part of code is displayed.

Electoral data

Demographical data

The variable *non-western residents* are divided in three groups:

- Municipalities with less than 5
- Municipalities with 5-10
- Municipalities with mre than 10

[1] https://data.overheid.nl/data/dataset/verkiezingsuitslag-tweede-kamer-2017
[2] https://opendata.cbs.nl/statline/#/CBS/nl/dataset/70072ned/table?ts=1544803364892

```
Data_CDA$Non_west <- ifelse(Data_CDA$Non_west_frac < 0.05, 1, NA)
Data_CDA$Non_west <- ifelse(Data_CDA$Non_west_frac >= 0.05 & Data_CDA$Non_west_frac <
    0.1, 2, Data_CDA$Non_west)
Data_CDA$Non_west <- ifelse(Data_CDA$Non_west_frac >= 0.1, 3, Data_CDA$Non_west)
Data_CDA$Non_west <- as.factor(Data_CDA$Non_west)
```

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[3]. The final dataset has no NAs

```
summary(Data_CDA)
```

```
##      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
##   Frac_60plus     Non_west
##  Min.   :0.0700   1:178
##  1st Qu.:0.1200   2:111
##  Median :0.1300   3: 77
##  Mean   :0.1327
##  3rd Qu.:0.1400
##  Max.   :0.1800
```

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 1 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 you can see in the plot, they are normal distributed. Because of the low values at the x-axis, the CDA, GroenLinks, 60 plus percentage and the highly educated densities are above 1. The area beneath the curve sums to 1, so it is correct.

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

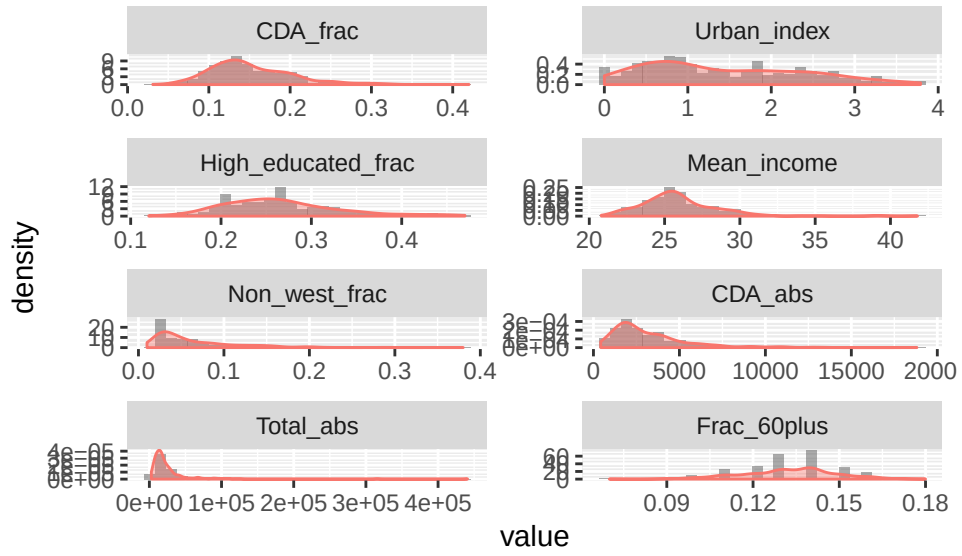


Figure 1: Density plot

Correlation heatmap In this heatmap (figure 2) 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.

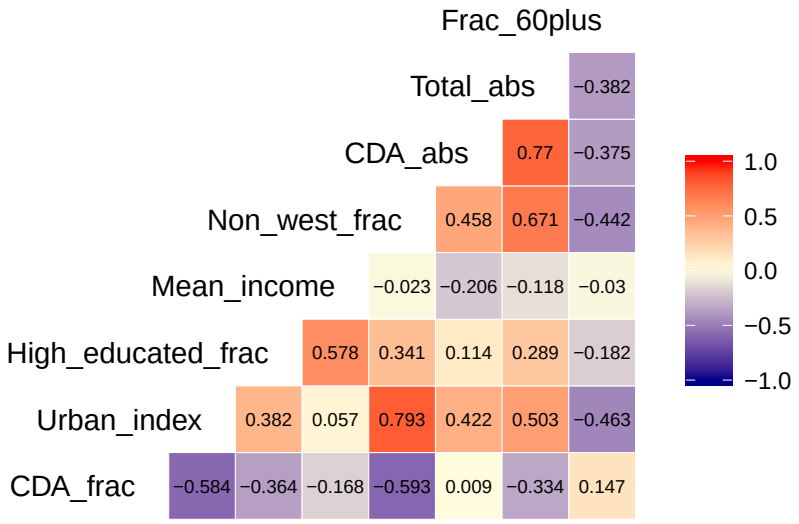


Figure 2: Correlation between explanatory and respons variables

Multilineair plots CDA In these two plots you can see a scatterplot with on the y-axis the votes for CDA in percentages and on the x-axis on the left graph the mean income per municipality in 1000 euro. The right plot has the urbanity index as x-axis. As you can see, the trend is that when the mean income goes up, the votes for CDA goes down. Same with the urbanity index. In the

model formulation graph these trends are checked.

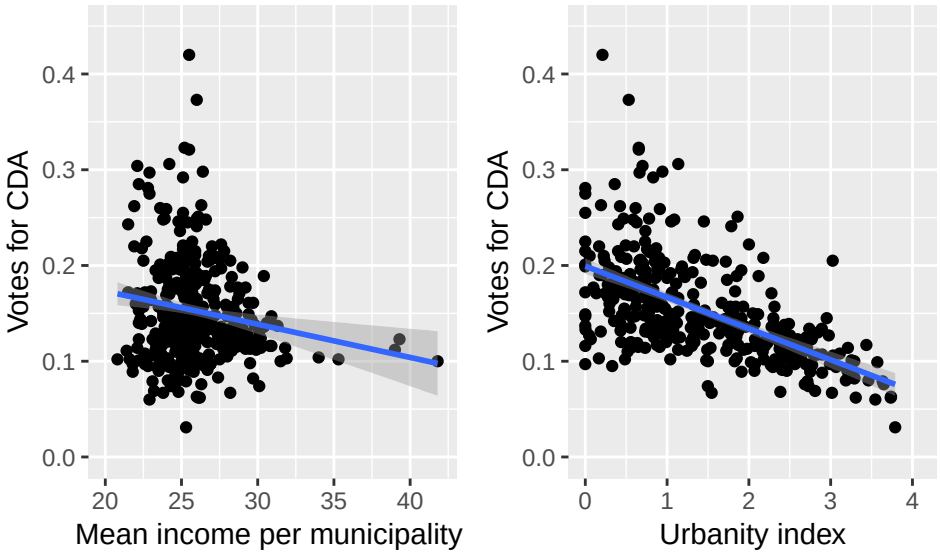


Figure 3: Scatterplots CDA

Exploratory plots of variables These three plots are scatterplots of the explanatory variables.

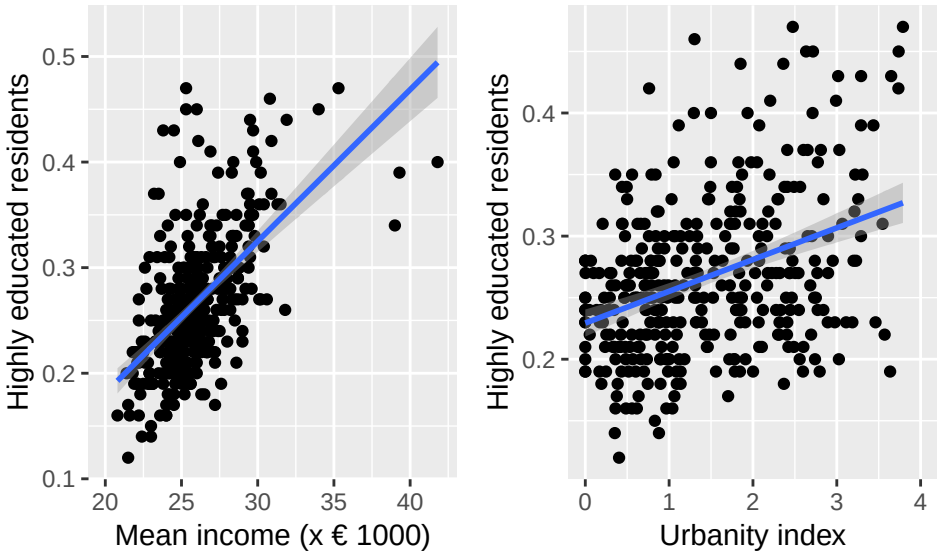


Figure 4: Scatterplot explanatory variables

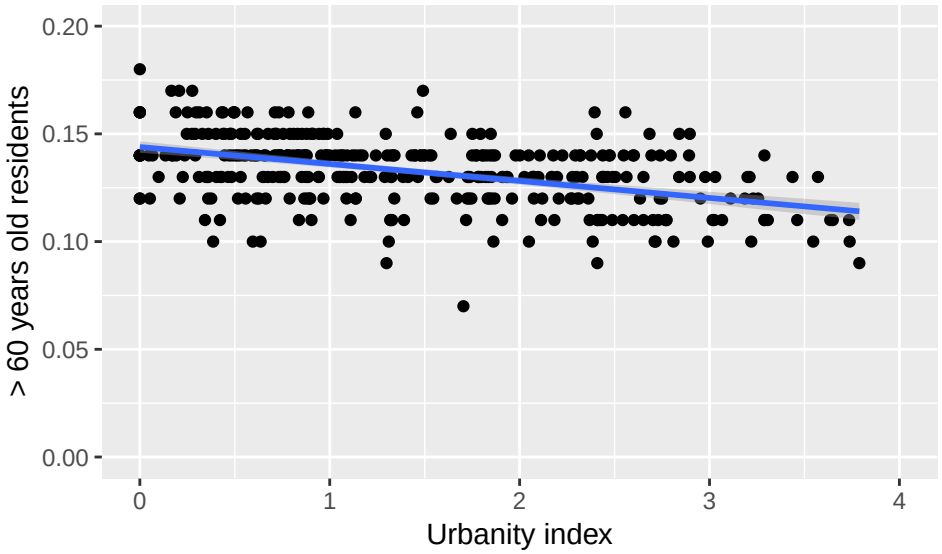


Figure 5: Scatterplot explanatory variables

Multiple boxplots In this graph boxplots are made, to compare some variables. 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.

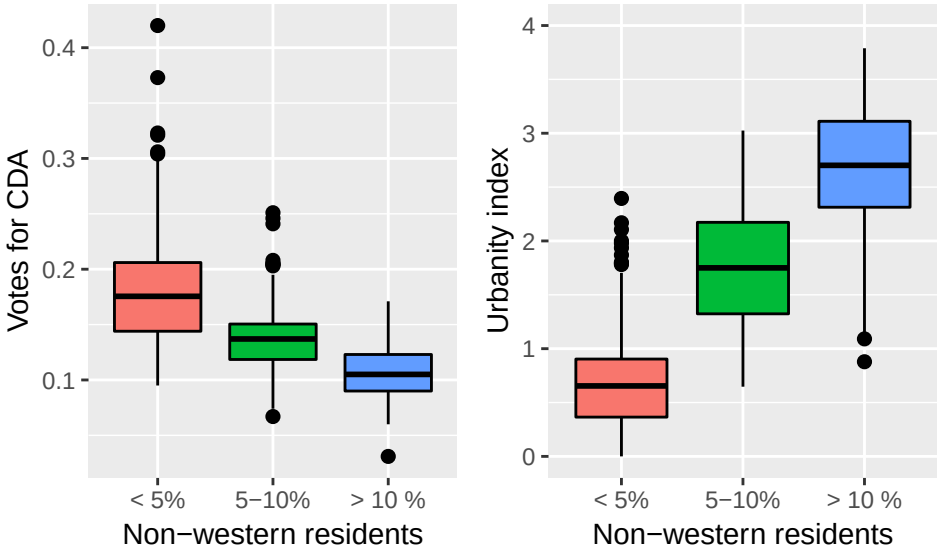


Figure 6: Three boxplots: Votes for CDA, Votes for GroenLinks and Urbanity index

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 * Constant variance: The variance of the error is constant * Normality: The errors are normally distributed * Indepence: The observations are sampled indipendently

First model

The first model will be the model with all the demographics:
$Y_i = \beta_0 + \beta_1 * higheducatedfraction + \beta_2 * Urbanindex + \beta_3 * Meanincome + \beta_4 * Nonwest2 + \beta_5 * Nonwest3 + \beta_6 * Frac60plus + \epsilon i$
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 total 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 two or more variables. The formula for VIF is $1/(1-R^2)$ and the thresholdvalue is 10. So 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
```

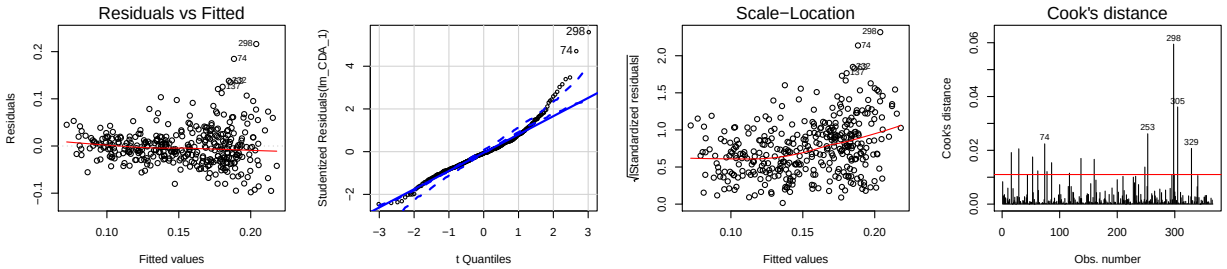


Figure 7: assumptions first model

In figure 10 the four 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 BoxCos figure 8 in has a 95% Confidence interval located around the 0. So a ln transformation is suggested.

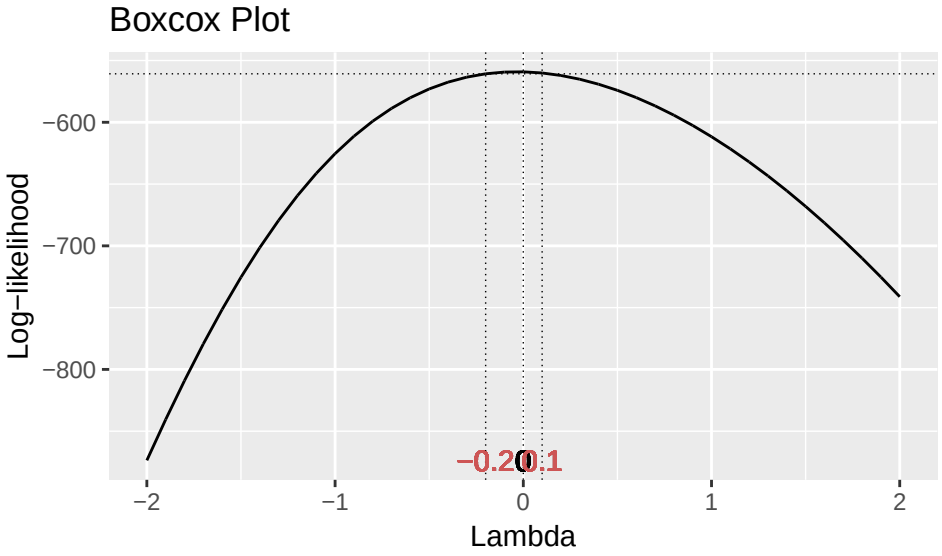


Figure 8: BoxCox first model

Second model

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

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

	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

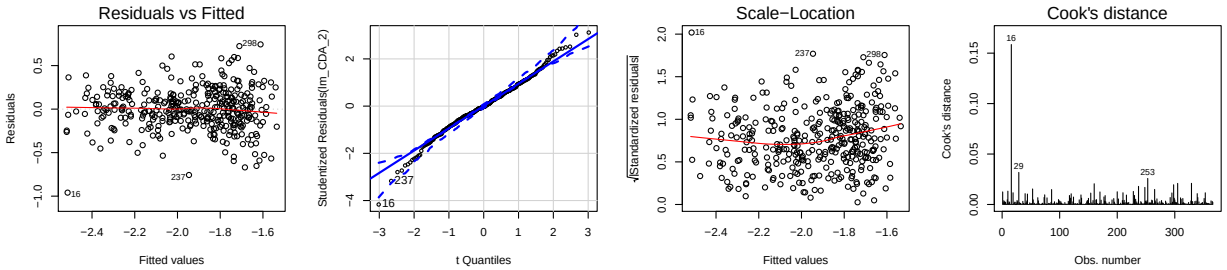


Figure 9: assumptions second model

The plots in figure 9 show one big outlier, the municipality Amsterdam which has number 16. Amsterdams value for the cooks distance goes way above the cutoff value for cooks, $4/(369-5-1) = 0.011$. It is also outside the (-3,3) range with the studentized residuals. That is why this municipality is removed.
For the second model without Amsterdam, a step function is used. This step function uses the AIC for backward elimination. If the AIC can get lower, because a variable is removed that variable will be removed else no variable is removed. 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
```

Final model

The backward elimination gave the final model.
$ln(Y_i) = \beta_0 + \beta_1 * higheducatedfraction + \beta_2 * Urbanindex + \beta_4 * Nonwest2 + \beta_5 * Nonwest3 + \beta_6 * Frac60plus + \epsilon i$ 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

First, the error-assumptions will be checked (figure 10). The 'loudspeaker pattern' almost disappeared, so the variance of the error is almost stable. Because the model is already transformed into a log model, there is not much transformation possible to let the pattern totally disappear. The influence will not be that high, because the pattern is relatively small. In the qq-plot it is visible that almost all municipalities are in the -3,3 range. Only the municipalities Oostzaan and Tubbergen are outside this range. This is because Oostzaan has an extreme low CDA_frac(0.067) and Tubbergen an extreme high CDA_frac(0.42). The decision is made not to delete these values, because they are still in the 95% envelope of the qq-plot. In the third plot the red line is notable, it goes up at the end of the graph. This means that there is some non-constant error variance, but because the scatter is not that big no action is needed.

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

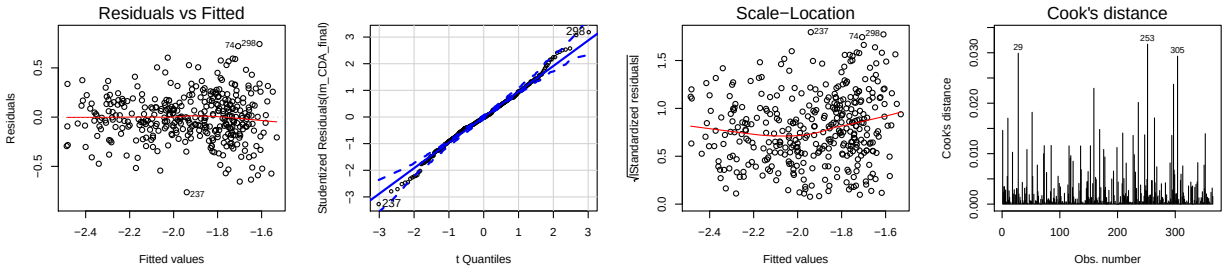


Figure 10: Assumptions final model

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

$ln(Y_i) = -1.0298 - 0.8277 * higheducatedfraction - 0.1311 * Urbanindex - 0.1141 * Nonwest2 - 0.2871 * Nonwest3 - 3.0168 * Frac60plus + \epsilon i$

Something notable is that all coëfficients are negative, but because the fitted values are logaritmic the eventual output will be positive. The added-variable plots are visible below (figure{avf}).

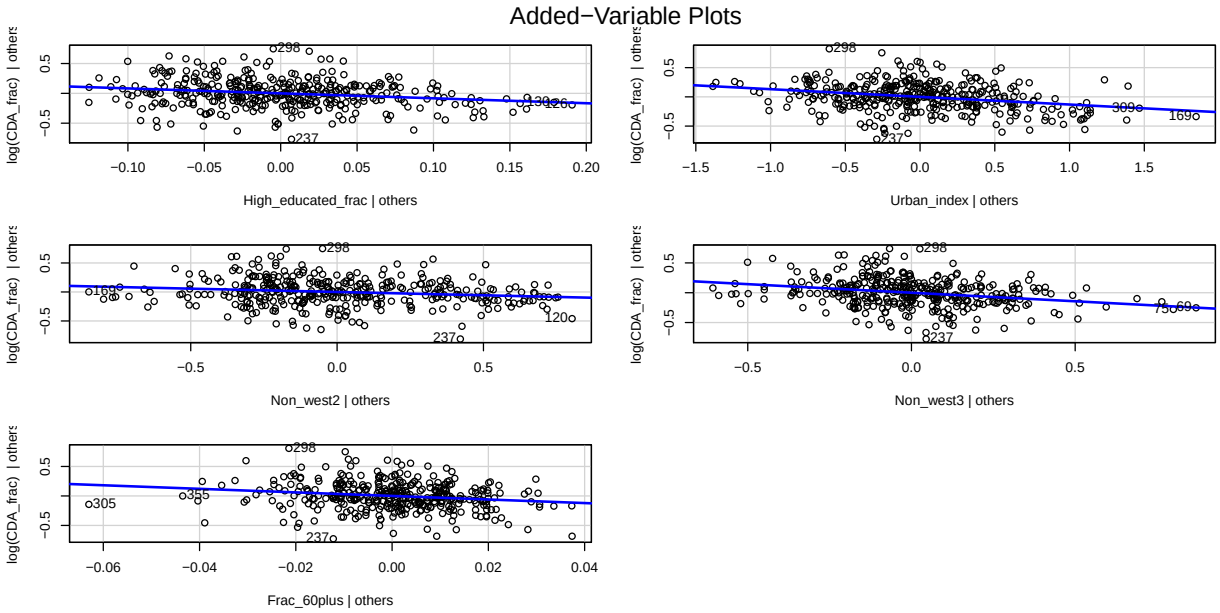


Figure 11: Added variable plots final model

Cross validation

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. This link displays the underlaying pattern of this data best. Below, the formula for the link function:

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

Where θ is the probability of votes for CDA.

Also for logistics regression are diagnostic plots 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 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.

First model

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

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

```
glm_CDA_1 <- glm(cbind(CDA_abs, Total_abs - CDA_abs) ~ Urban_index + High_educated_frac +
    Mean_income + Non_west + Frac_60plus, family = binomial(link = "logit"),
    data = Data_CDA)
summary(glm_CDA_1)
```

```
##
## Call:
## glm(formula = cbind(CDA_abs, Total_abs - CDA_abs) ~ Urban_index +
##     High_educated_frac + Mean_income + Non_west + Frac_60plus,
##     family = binomial(link = "logit"), data = Data_CDA)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -88.318   -8.988   -1.521    7.670   57.920
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -1.0559765  0.0156927  -67.29   <2e-16 ***
## Urban_index        -0.1933926  0.0020376  -94.91   <2e-16 ***
## High_educated_frac -2.1027583  0.0199536 -105.38   <2e-16 ***
## Mean_income         0.0156359  0.0005334   29.32   <2e-16 ***
## Non_west2          -0.0562758  0.0031606  -17.80   <2e-16 ***
## Non_west3          -0.2592768  0.0046088  -56.26   <2e-16 ***
## Frac_60plus        -1.3424322  0.0737404  -18.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 247550  on 365  degrees of freedom
## Residual deviance:  89969  on 359  degrees of freedom
## AIC: 93475
##
## Number of Fisher Scoring iterations: 4
```

The summary visualizes that the all the variables are very significant, with small 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 3% voted for CDA, while it others nearly 50% voted for CDA. It is concluded that a quasi-binomial would fit the data better.

Second model

```
glm_CDA_2 <- glm(cbind(CDA_abs, Total_abs - CDA_abs) ~ Urban_index + High_educated_frac +
    Mean_income + Non_west + Frac_60plus, family = quasibinomial(link = "logit"),
    data = Data_CDA)
summary(glm_CDA_2)
```

```
##
## Call:
## glm(formula = cbind(CDA_abs, Total_abs - CDA_abs) ~ Urban_index +
##     High_educated_frac + Mean_income + Non_west + Frac_60plus,
##     family = quasibinomial(link = "logit"), data = Data_CDA)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -88.318   -8.988   -1.521    7.670   57.920
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -1.055977   0.250122  -4.222 3.07e-05 ***
## Urban_index        -0.193393   0.032478  -5.955 6.21e-09 ***
## High_educated_frac -2.102758   0.318036  -6.612 1.38e-10 ***
## Mean_income         0.015636   0.008501   1.839  0.06670 .
## Non_west2          -0.056276   0.050376  -1.117  0.26469
## Non_west3          -0.259277   0.073459  -3.530  0.00047 ***
## Frac_60plus        -1.342432   1.175330  -1.142  0.25414
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 254.0441)
##
##     Null deviance: 247550  on 365  degrees of freedom
## Residual deviance:  89969  on 359  degrees of freedom
## AIC: NA
```

```
##
## Number of Fisher Scoring iterations: 4
```

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. The variables `Frac_60plus`, `Mean_income` and factor `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.

```
vif(glm_CDA_2)
```

```
##        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
```

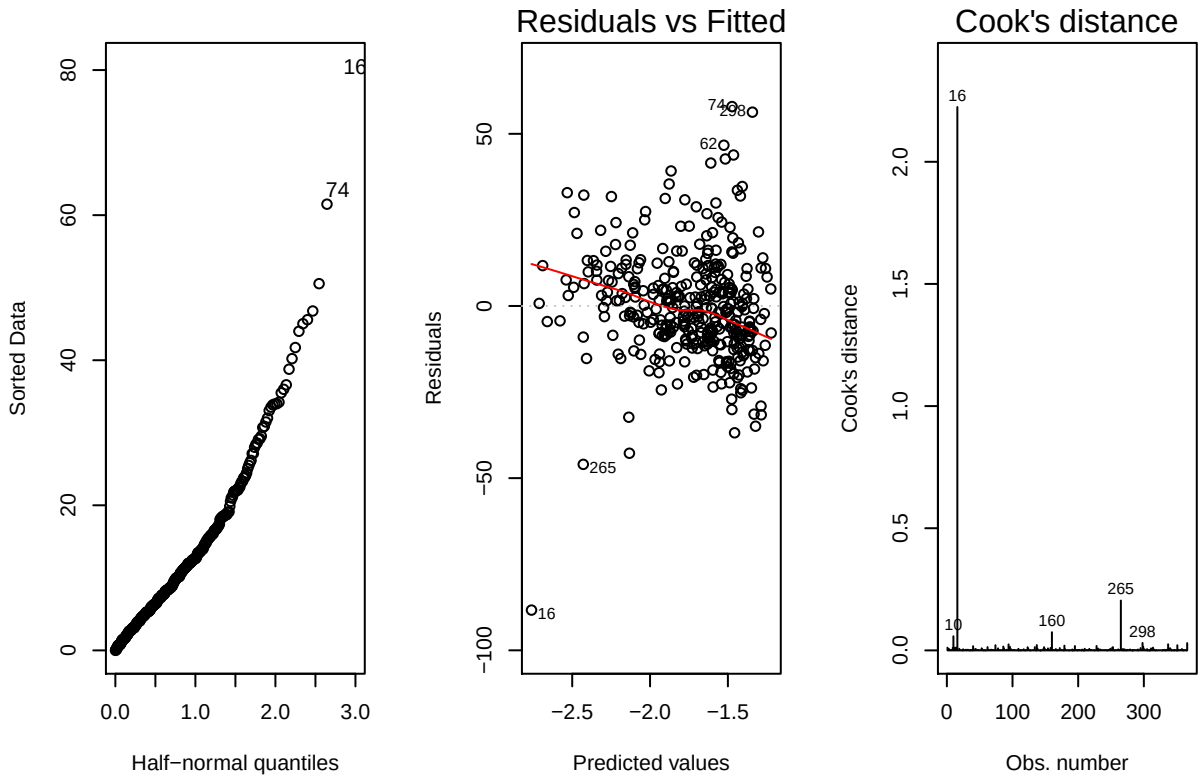
```
drop1(glm_CDA_2, test = "F")
```

```
## Single term deletions
##
## Model:
## cbind(CDA_abs, Total_abs - CDA_abs) ~ Urban_index + High_educated_frac +
##     Mean_income + Non_west + Frac_60plus
##                     Df Deviance F value    Pr(>F)
## <none>                   89969
## Urban_index          1    99052 36.2459 4.300e-09 ***
## High_educated_frac   1   101151 44.6196 9.132e-11 ***
## Mean_income          1    90821  3.3987 0.0660730 .
## Non_west             2    94112  8.2661 0.0003092 ***
## Frac_60plus          1    90300  1.3217 0.2510613
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
par(mfrow = c(1, 3))
halfnorm(residuals(glm_CDA_2, "pearson"))
plot(glm_CDA_2, which = 1, id.n = 5)
plot(glm_CDA_2, which = 4, id.n = 5)
```

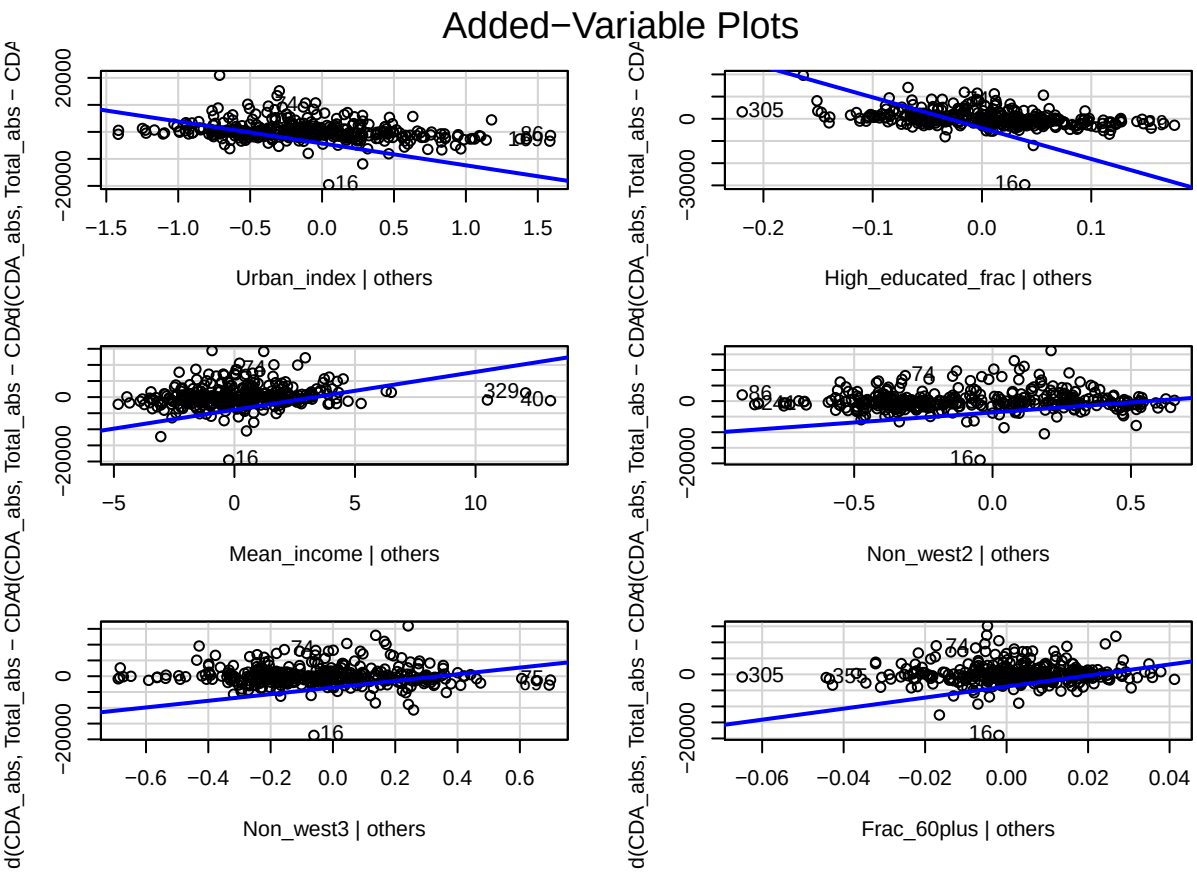



```
Data_CDA[c(16, 74, 265), ]
```

```
##           Muni CDA_frac Urban_index High_educated_frac Mean_income
## 16   Amsterdam    0.031       3.789               0.47        25.3
## 74  Dinkelland    0.373       0.531               0.26        26.0
## 265  Rotterdam    0.060       3.546               0.31        22.9
##     Non_west_frac CDA_abs Total_abs Frac_60plus Non_west
## 16           0.35   13562    440854        0.09        3
## 74           0.02    6560     17586        0.13        1
## 265          0.38   18813    315550        0.10        3
```

The left plot display 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. There seems to be non-constant erro variance, because the spread becomes larger for higher predicted values. The right plots shows 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.

```
avPlots(glm_CDA_2)
```



The avPlots help to interpret the partial regression coefficients 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.

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 influenced the partial regression coefficients greatly and had large residual and cook's distances. These steps were originally done in two, but they are combined for this report.

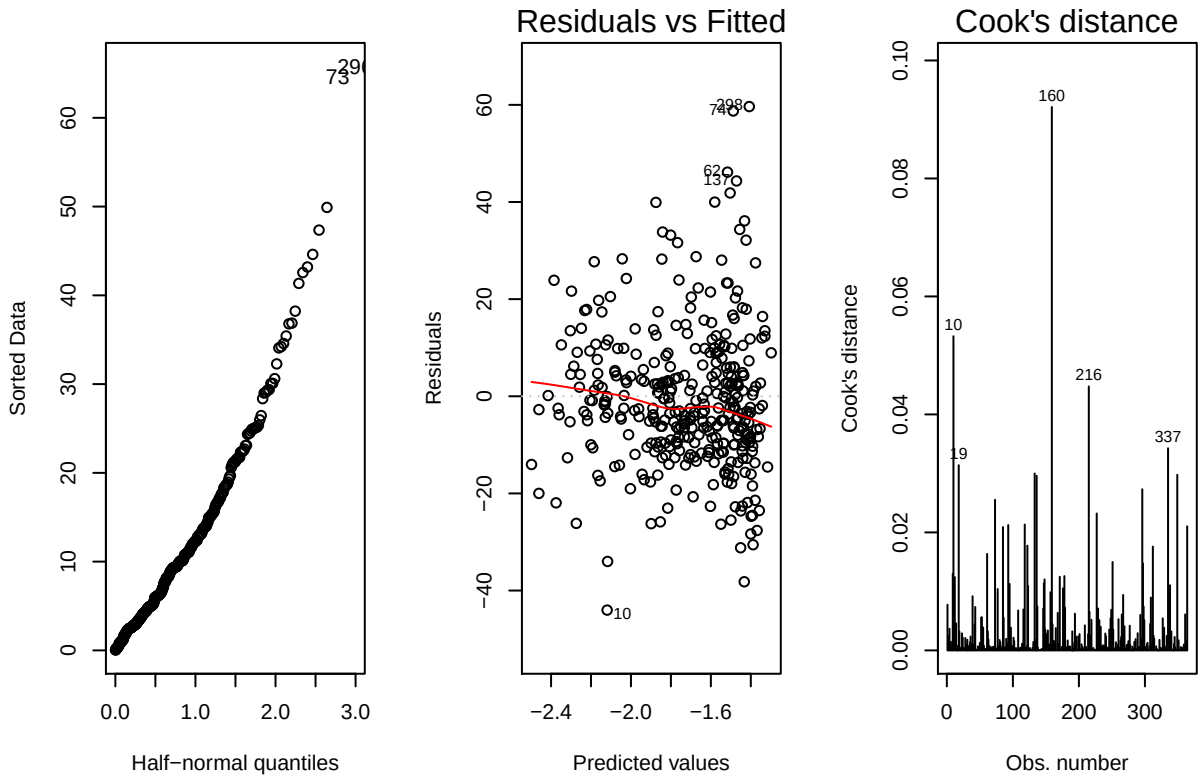
```
glm_CDA_4 <- glm(cbind(CDA_abs, Total_abs - CDA_abs) ~ Urban_index + High_educated_frac +
    Mean_income + Non_west, family = quasibinomial(link = "logit"), data = Data_CDA[-c(16,
    265), ])
summary(glm_CDA_4)
```

```
##
## Call:
## glm(formula = cbind(CDA_abs, Total_abs - CDA_abs) ~ Urban_index +
##     High_educated_frac + Mean_income + Non_west, family = quasibinomial(link = "logit"),
##     data = Data_CDA[-c(16, 265), ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -44.024   -9.619   -2.437    6.074   59.653
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -1.0877540  0.1801672  -6.037 3.92e-09 ***
## Urban_index        -0.1351943  0.0303257  -4.458 1.11e-05 ***
## High_educated_frac -1.2202265  0.3125548  -3.904 0.000113 ***
## Mean_income        -0.0004273  0.0081731  -0.052 0.958331
## Non_west2          -0.1232492  0.0479004  -2.573 0.010483 *
## Non_west3          -0.3447138  0.0690753  -4.990 9.42e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 221.709)
##
##     Null deviance: 173842  on 363  degrees of freedom
## Residual deviance:  76966  on 358  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

By removing observation 16 and 265, the factor `Non_west2` has become significant. Also the residual deviance has decreased slightly, but is still far above the degrees of freedom.

```
par(mfrow = c(1, 3))
halfnorm(residuals(glm_CDA_4, "pearson"))
plot(glm_CDA_4, which = 1, id.n = 5)
plot(glm_CDA_4, which = 4, id.n = 5)
```



The plot left still displays very large pearson residuals. The plot in the middle seems homogenic distributed, which slightly more variance for higher predicted values. And the cook's distance no longer displays highly influential observations.

```
drop1(glm_CDA_4, test = "F")
```

```
## Single term deletions
##
## Model:
## cbind(CDA_abs, Total_abs - CDA_abs) ~ Urban_index + High_educated_frac +
##     Mean_income + Non_west
##                    Df Deviance F value    Pr(>F)
## <none>                  76966
## Urban_index         1    81396 20.6048 7.721e-06 ***
## High_educated_frac  1    80361 15.7942 8.543e-05 ***
## Mean_income         1    76966  0.0028    0.9577
## Non_west            2    82994 14.0208 1.373e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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.

Final model

The final model is reached after dropping the variables `Mean_income` and `Frac_60plus`. It's formula is as follows:

$logit(p_{ij}) = -1.09 - 0.14 \cdot UrbanIndex - 1.23 \cdot HighlyEducatedFraction - 0.12 \cdot NonWest : 2 - 0.34 \cdot NonWest : 3$

```
glm_CDA_5 <- glm(cbind(CDA_abs, Total_abs - CDA_abs) ~ Urban_index + High_educated_frac +
    Non_west, family = quasibinomial(link = "logit"), data = Data_CDA[-c(16,
    265), ])
summary(glm_CDA_5)
```

```
##
## Call:
## glm(formula = cbind(CDA_abs, Total_abs - CDA_abs) ~ Urban_index +
##     High_educated_frac + Non_west, family = quasibinomial(link = "logit"),
##     data = Data_CDA[-c(16, 265), ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -44.014   -9.626   -2.447    6.099   59.654
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -1.09653    0.06526 -16.804  < 2e-16 ***
## Urban_index        -0.13497    0.02996  -4.504 9.02e-06 ***
## High_educated_frac -1.22972    0.25406  -4.840 1.93e-06 ***
## Non_west2          -0.12346    0.04767  -2.590  0.00999 **
## Non_west3          -0.34443    0.06876  -5.009 8.60e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 221.0972)
##
##     Null deviance: 173842  on 363  degrees of freedom
## Residual deviance:  76966  on 359  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

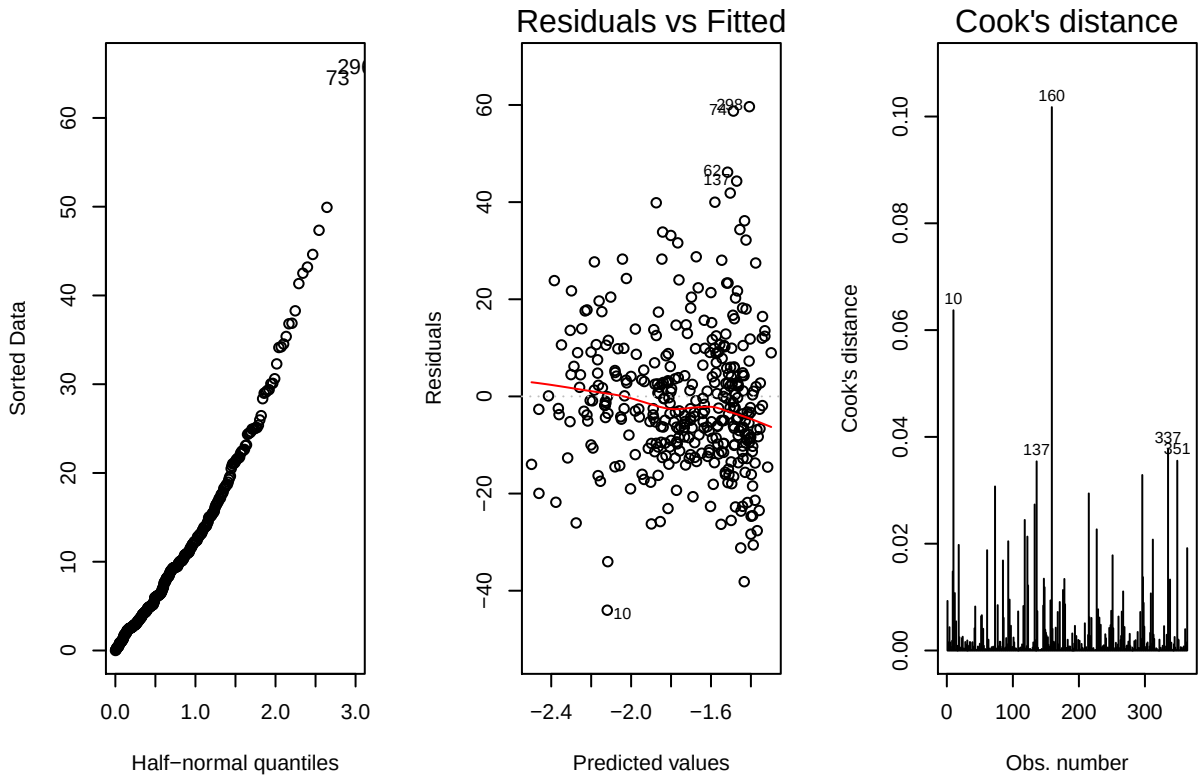
```
drop1(glm_CDA_5, test = "F")
```

```
## Single term deletions
##
## Model:
## cbind(CDA_abs, Total_abs - CDA_abs) ~ Urban_index + High_educated_frac +
##     Non_west
##                   Df Deviance F value    Pr(>F)
## <none>                  76966
```

```
## Urban_index          1    81471  21.009 6.318e-06 ***
## High_educated_frac   1    82190  24.364 1.223e-06 ***
## Non_west             2    83077  14.251 1.108e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-test shows that all variables now significantly contribute to the model. The residual deviance has slightly decreased, in compare to model `glm_CDA_2` with all the variables. However, the residual deviance is still much larger then the degrees of freedom.

```
par(mfrow = c(1, 3))
halfnorm(residuals(glm_CDA_5, "pearson"))
plot(glm_CDA_5, which = 1, id.n = 5)
plot(glm_CDA_5, which = 4, id.n = 5)
```

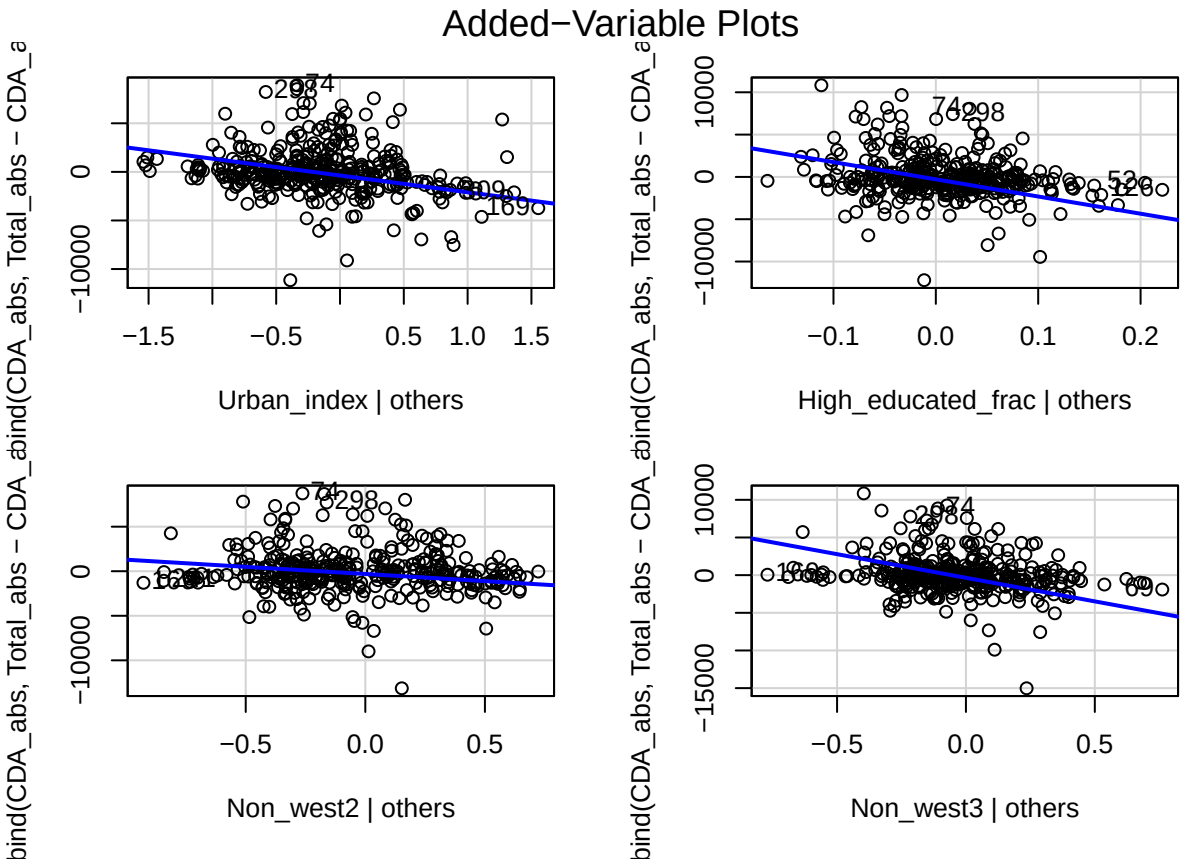


There is still a large spread of both the pearson (left plot) and deviance (middle plot) residuals. Furthermore, there is non-constant errror variance. The cook's distance does not display very influential observations.

```
vif(glm_CDA_5)
```

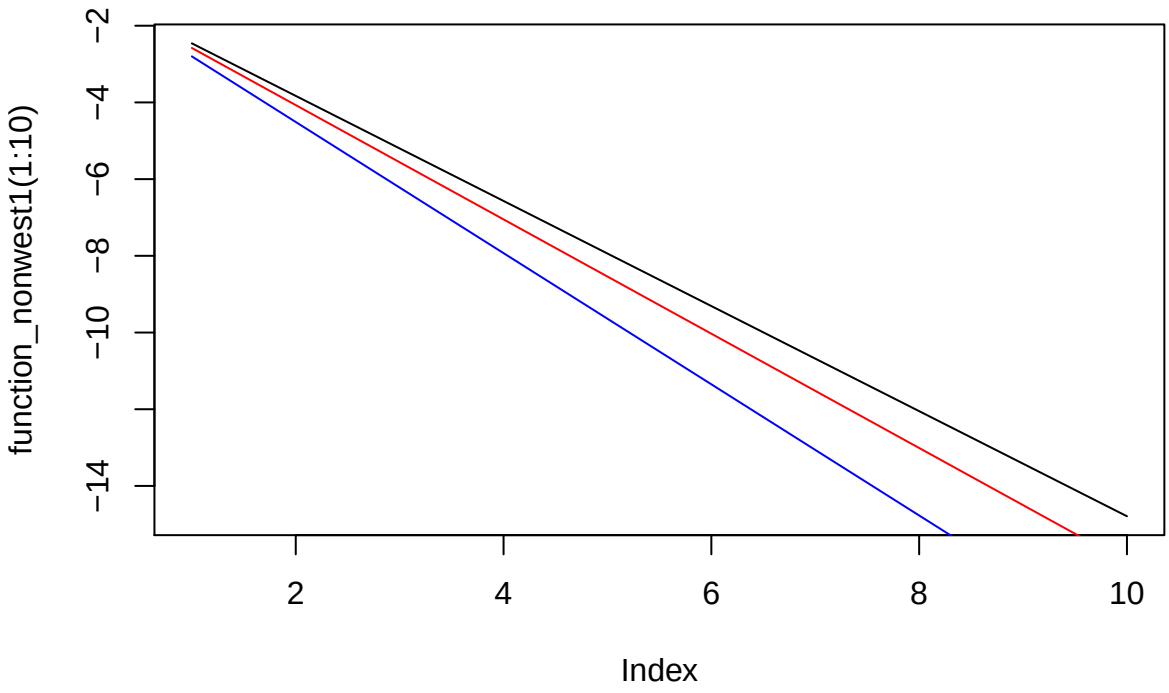
```
##         Urban_index High_educated_frac          Non_west2
##        0.0012896670       0.0004230904       0.0007928046
##          Non_west3
##        0.0012734699
```

```
avPlots(glm_CDA_5)
```



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.

```
function_nonwest1 <- function(x) {
    -1.09 - 0.14 * x - 1.23 * x
}
function_nonwest2 <- function(x) {
    -1.09 - 0.14 * x - 1.23 * x - 0.12 * x
}
function_nonwest3 <- function(x) {
    -1.09 - 0.14 * x - 1.23 * x - 0.34 * x
}
plot(function_nonwest1(1:10), type = "l")
lines(function_nonwest2(1:10), type = "l", col = "red")
lines(function_nonwest3(1:10), type = "l", col = "blue")
```



Cross validation

4. Discussion

Limitations