bluh

Floor Komen

15 januari 2019

Abstract

Put the abstract over here

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.

```
##                        GVIF Df GVIF^(1/(2*Df))
## High_educated_frac 1.871032  1        1.367857
## Urban_index        3.383149  1        1.839334
## Mean_income        1.658015  1        1.287640
## Non_west           3.293503  2        1.347146
## Frac_60plus        1.289979  1        1.135772
```

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

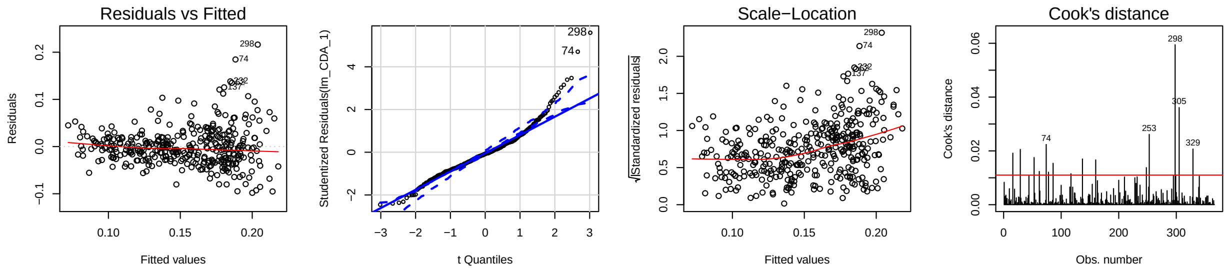

Figure 1: assumptions first model

In figure 1 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 2 in has a 95% Confidence interval located around the 0. So a ln transformation is suggested.

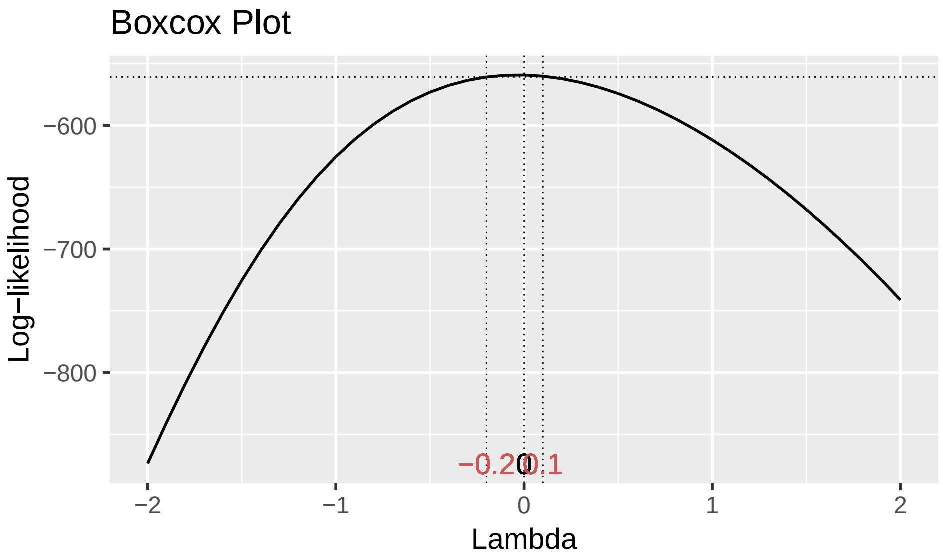

Figure 2: BoxCox first model

Second model

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

	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
```

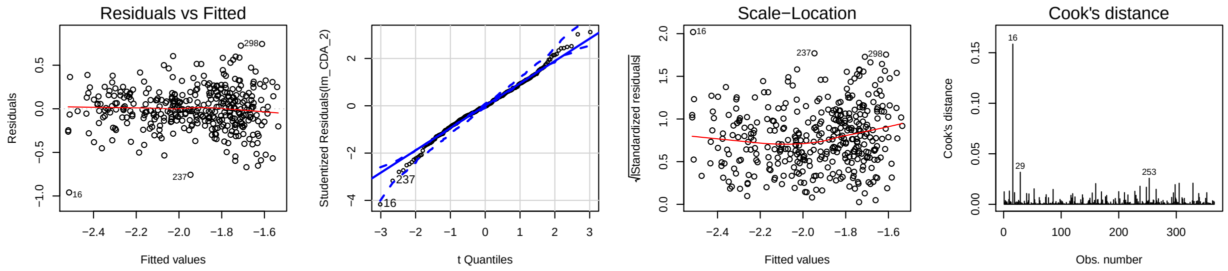

Figure 3: assumptions second model

The plots in figure 4 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

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

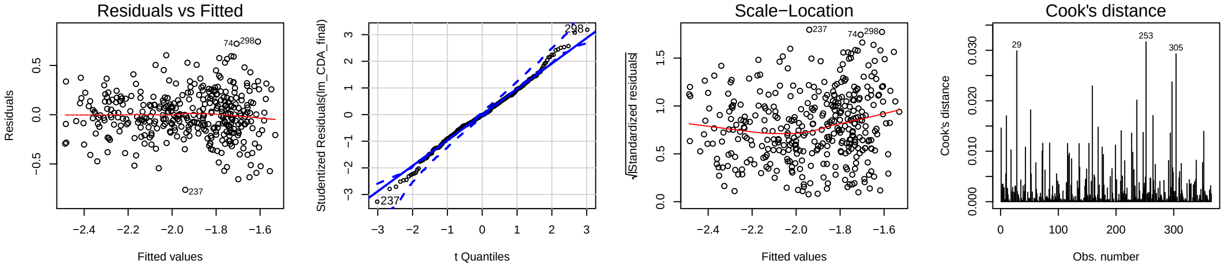

Figure 4: assumptions second model

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

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

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

Cross validation

To tell something about the prediction possibilities of the model, cross validation is done. Cross validation tells something about how well the model predicts on average, it telss nothing about the 'correctness'of the model. Cross validation estimates the expected prediction error of a model. The cross validation works as follows. First 5k-folds are made. This means that the data is divided in five folds. Next, the loss function is made. This function makes the square of the real value minues the predicted value. By taking the sum of these values and the deviation by amount of real values in the k-folds, the mean is taken. 4 of the 5 kfolds are used ase training data, the other one is the test/validation 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 is the testdata. As is shown below, the prediction error for this model 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] 0.05814674
```

Discussion

Linear regression

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

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

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 no reality. For the other coefficients it is a bit harder to predict there influence, because of the log transformation and the different range the different variables have. For example, the urban index has a 0-3.8 range and education ahs 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 everything else stays the same an frac 60 plus changes for example 1 the exponent changes with -3.02. The non west2 has the highest impact on the slope, because the coëfficient is the lowest. Another important point is that or Non-west2 and Non-west3 are both zero, or Non-west2 is one ore Non west 3 is one. The outcome of the crossvalidation for this model is 0.0582. So the mean squared difference between the fitted and predicted value is 0.0582, which is pretty 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 Generalized linear model would be 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.

further research

Both of the models have different significant variables. But it is difficult to say which one is a better fitting model. Because both of them have reasons to choose that kind of model, also both of them have limitations. That is why further research should be done to research which of the model is the best fitting model. Another topic that can be researched in further research is the influence of demographics on areas of municipalities instead of whole municipialities. Because differences between areas are nullified in the demographics of a municipality.

Abstract

In this repport some demographics of dutch municipalities are compared with the general Dutch election results of 2017. The demographic variables in this research are the urban index, the fraction of high educated persons in a municipality, the non west factor, a 1 for the factor means less than 5% non west residents, a 2 means 5 till 10% non west residents and a three means more than 10% residents, the mean income of the municipality and the fraction of persons who are 60 years and older. The response variable is the CDA votes. Two models are made to explain the influence of the demographics on the amount of votes for CDA per municipality. The first model is a multiple linear model, the fraction of CDA votes is the respons variable. Because of error assumptions, the response variable is log transformed. The final model with this transformation is:
$Y_i = e^{(} - 1.0298 - 0.8277 * higheducatedfraction - 0.1311 * Urbanindex - 0.1141 * Nonwest2 - 0.2871 * Nonwest3 - 3.0168 * Frac60plus + \epsilon i)$
The fraction 60 plus has the lowest marginal influence on the slope of the model. The factor

Non west2 has the highest marginal impact on the slope. There are still some limitations, firstly because even though a log transformation has been made there is still a non constant error variance. Secondly because the respons variable is a fraction, theoretically a generalized linear model would be better. That is why the second kind of model described is a generalized linear model. HIER GML SAMENVATTING PLUS DISCUSSION The final linear model has different significant variables than the general linear model. It is difficult to say which model is better, because both have there limitations. In further research the best fitting model of those two can be obtained.