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.

```
## High_educated_frac        Urban_index        Mean_income
##           1.846691           3.355218           1.576931
##           Non_west        Frac_60plus
```

```
##            3.109682            1.285869
## [1]  74 298
```

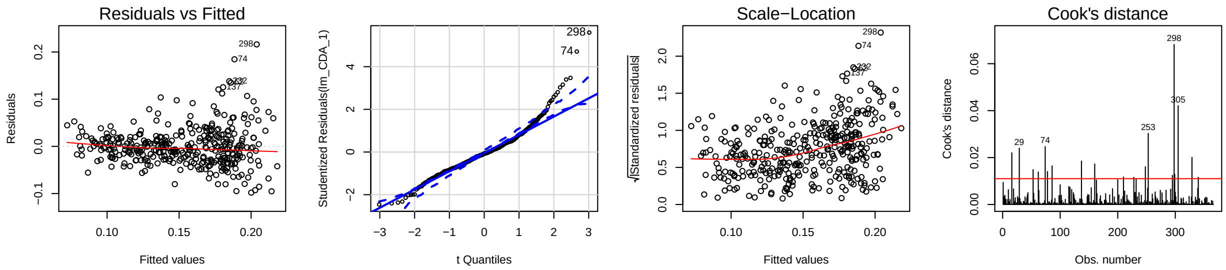

Figure 1: assumptions first model

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.

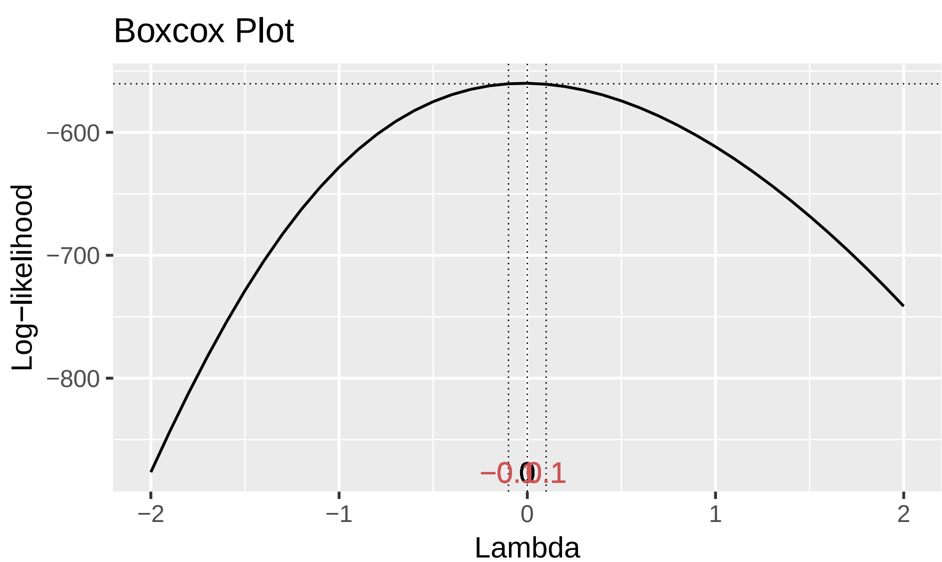

Figure 2: BoxCox first model

Second model

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

```
## [1]  16 298
```

	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

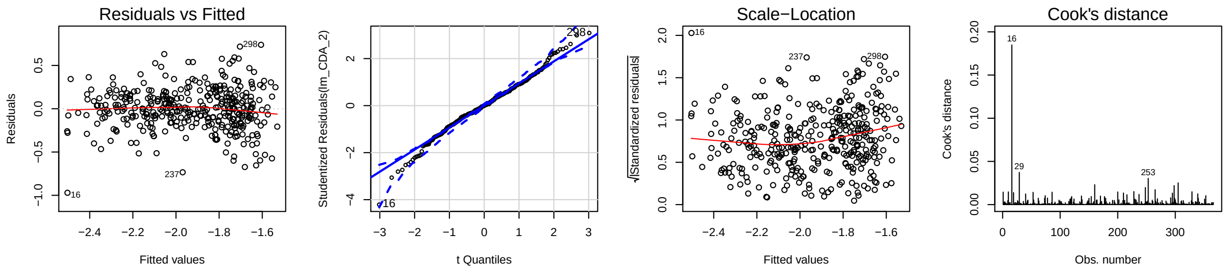

Figure 3: assumptions second model

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=-1042
## log(CDA_frac) ~ High_educated_frac + Urban_index + Mean_income +
##     Non_west + Frac_60plus
##
##                      Df Sum of Sq    RSS     AIC
## - Mean_income         1   0.02142 20.354 -1043.6
## <none>                            20.333 -1042.0
## - High_educated_frac  1   0.40428 20.737 -1036.8
## - Frac_60plus         1   0.64967 20.982 -1032.5
## - Non_west            1   1.45896 21.792 -1018.7
## - Urban_index         1   1.67178 22.005 -1015.2
##
## Step:  AIC=-1043.61
## log(CDA_frac) ~ High_educated_frac + Urban_index + Non_west +
##     Frac_60plus
##
```

```
##                        Df Sum of Sq    RSS     AIC
## <none>                               20.354 -1043.6
## - Frac_60plus          1   0.64626 21.000 -1034.2
## - High_educated_frac   1   0.83927 21.193 -1030.9
## - Non_west             1   1.44891 21.803 -1020.5
## - Urban_index          1   1.65046 22.005 -1017.2
```

```
##
## 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
##            -0.8921             -0.8199             -0.1294
##           Non_west         Frac_60plus
##            -0.1410             -2.9719
```

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
```

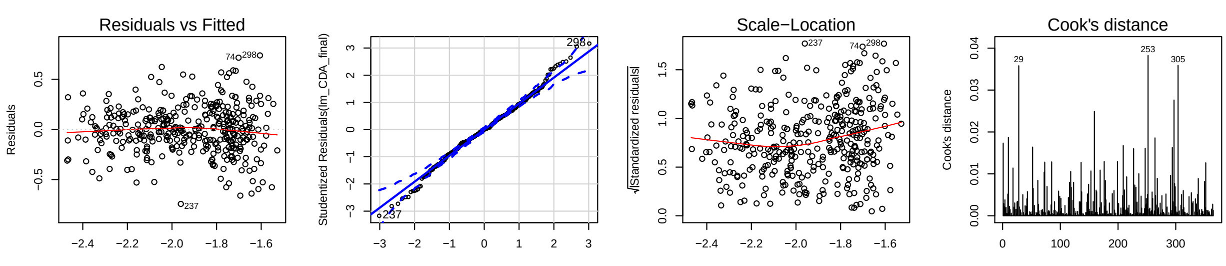

Figure 4: assumptions second model

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