Generalized linear models in practice

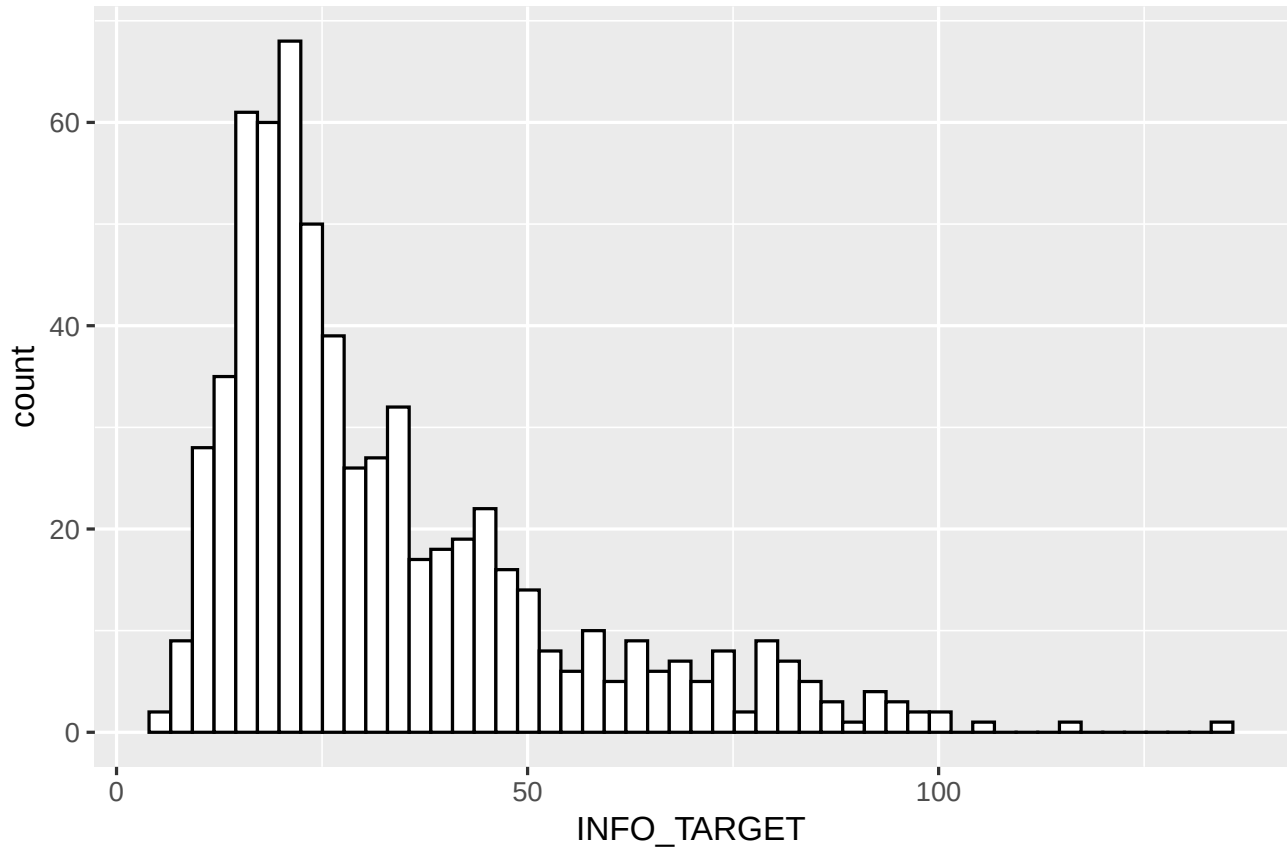
Jana Faganeli Pucer

University of Ljubljana, Faculty of Computer and Information Science

March, 2020

Prediction of PM10 concentrations from Zagorje (2012-2013)

- Dependent variable -> PM10 concentrations from Zagorje
- 9 independent variables -> describing previous PM10 concentrations variables describing the meteorological situation

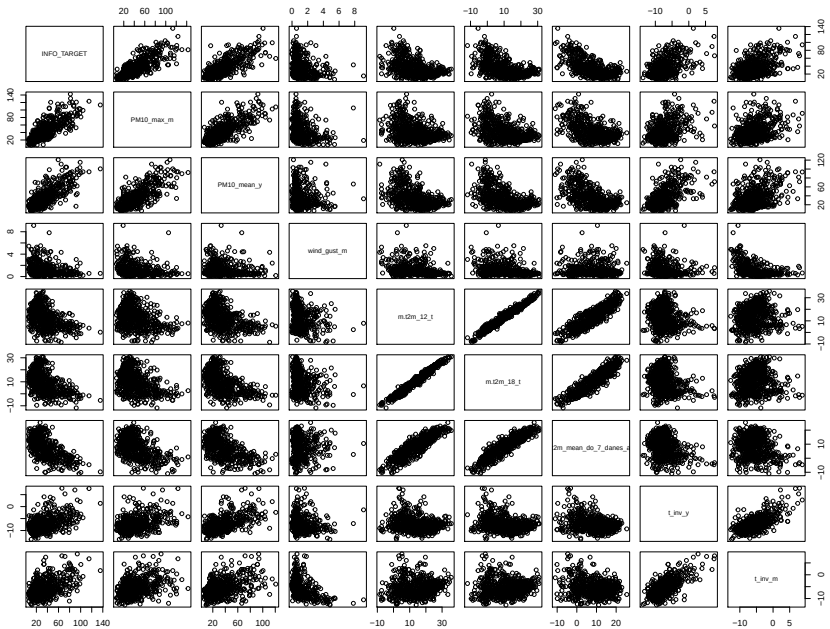


Assumptions of linear regression

Before applying linear regression we make some assumptions about our data:

- Linearity
- Errors are normally distributed
- Homoscedasticity: same error variance for different values of the response variable
- Independence of errors
- Lack of perfect collinearity between independent variables

Collinearity



- Variables should exhibit a linear relationship between the dependent and independent variable.
- The collinearity of m.t2m.12_t and m.t2m.18_t could represent a problem for the interpretation of the regression coefficients, the regression results are still valid.

Fitting an ordinary linear regression

```
lm(formula = INFO_TARGET ~ ., data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-41.198  -5.340  -0.339   4.110  46.920

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.62233    2.39572   8.191 1.42e-15 ***
PM10_max_m   0.34723    0.02744  12.655  < 2e-16 ***
PM10_mean_y  0.28088    0.03462   8.114 2.52e-15 ***
wind_gust_m -0.86431    0.44779  -1.930 0.054029 .
m.t2m_12_t   0.55994    0.24835   2.255 0.024492 *
m.t2m_18_t  -0.52370    0.26454  -1.980 0.048171 *
t2m_mean_m  -0.72412    0.18937  -3.824 0.000144 ***
t_inv_y      0.27100    0.21778   1.244 0.213830
t_inv_m      0.49901    0.22978   2.172 0.030244 *

Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 9.781 on 639 degrees of freedom
Multiple R-squared:  0.7851,    Adjusted R-squared:  0.7824
F-statistic: 291.8 on 8 and 639 DF,  p-value: < 2.2e-16
```

Is this an appropriate model?

Model diagnostics

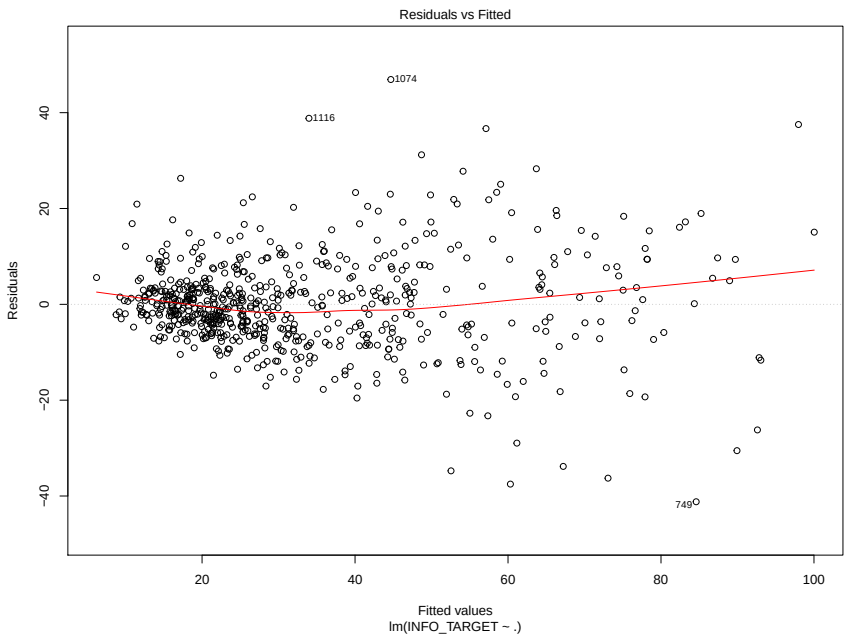
Residuals are the basis of most diagnostic methods. Different residuals:

- **Ordinary residuals** $e_i = y_i - \hat{y}_i$.
  In ordinary least squares (OLS) are uncorrelated with the fitted values.
  If the regressor model is correct than residuals are random variables with mean 0 and with variance $Var(e_i) = \sigma^2(1 - hi)$
  h_i is the leverage.
- **Standardised residuals** $e_{Si} = \frac{e_i}{\hat{\sigma}\sqrt{1-h_i}}$, where $\hat{\sigma}$ is the estimated of σ .
- **Studentised residuals** $e_{Ti} = \frac{e_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_i}}$ where $\hat{\sigma}_{(-i)}$ is the estimate of σ without the i-th observation.

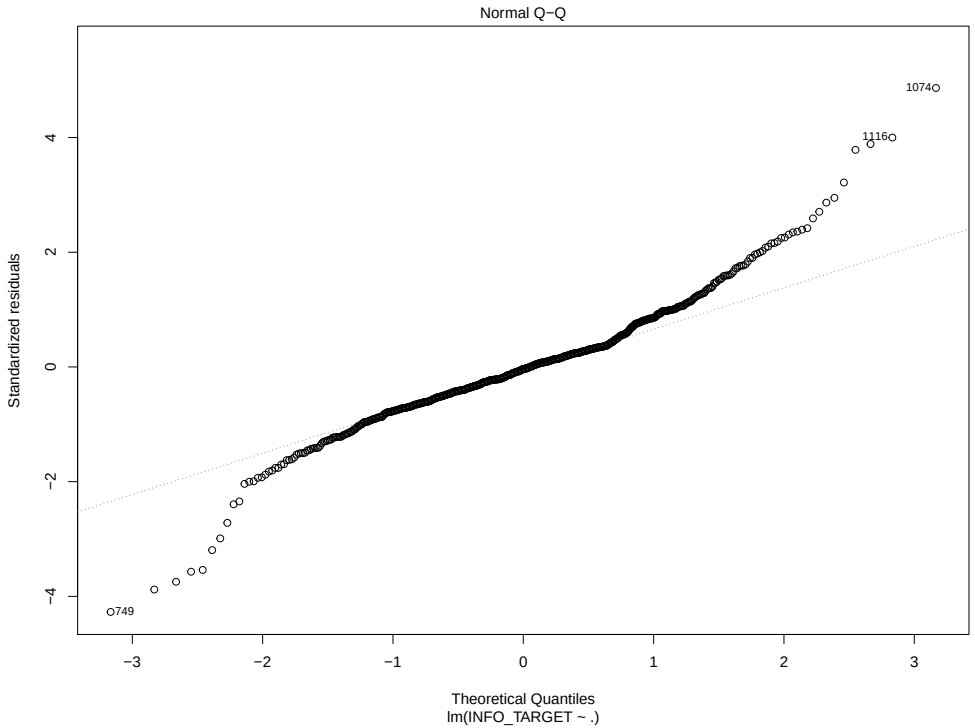
Leverage

Observations that are far from the center of the regressor space have potentially great influence on the least-square regression coefficient estimate.
Leverage Assesses how far away the independent variable values of an observation are from those of the other observations (difference in x-values).
The most common measure of leverage are hat values: The vector of fitted values is given by $\hat{y} = Xb = X(X^T X)^{-1} X^T y = Hy$ where $H = h_{ij} = X(X^T X)^{-1} X^T$.
H projects y into the subspace spanned by the columns of the model matrix X.
h_{ij} are diagonal values

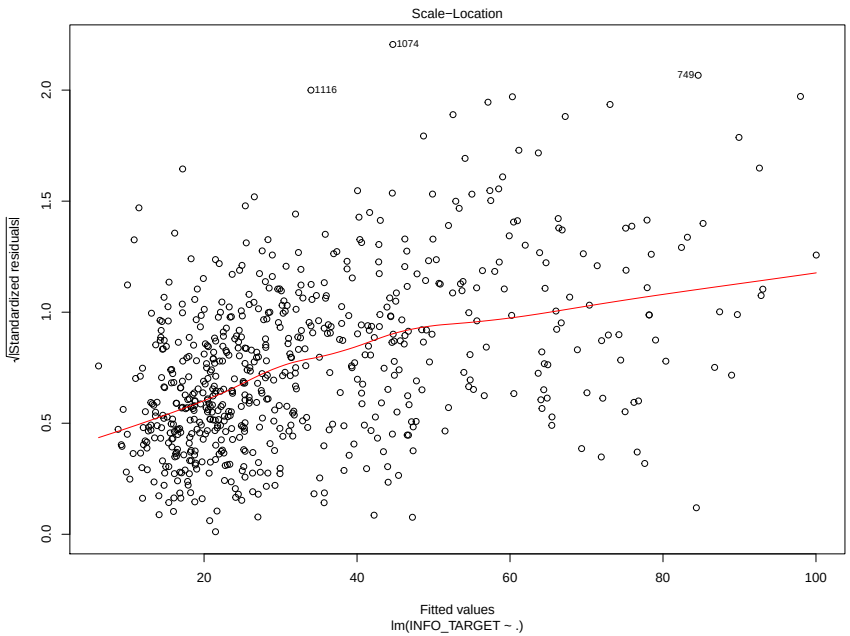
Diagnostic plots



This plot shows if the residuals a non-linear patter (the red approximated line should be straight). It also shows the dispersion of residuals for different fitted values. The values should be evenly distributed around zero (homoscedasticity).



- Shows if the residuals are distributed according to a distribution (in this case the normal distribution).
- On the x-axis the theoretical quantiles of the normal distribution are plotted and on the y-axis the quantiles of our residuals.
- If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.
- Points lying below the straight line are closer to the median value than they should be in the investigated distribution, point lying above are further



Similar to the first plot, but with standardised residuals (and square root). Shows if the residuals are spread equally along the ranges of fitted values (homoscedasticity).

Cook's distance

Is an estimate of the influence of a data point on linear regression.

An observation that is both **outlying** and has high **leverage** exerts high influence on the regression coefficients.
If the observation is removed the regression coefficient change significantly:

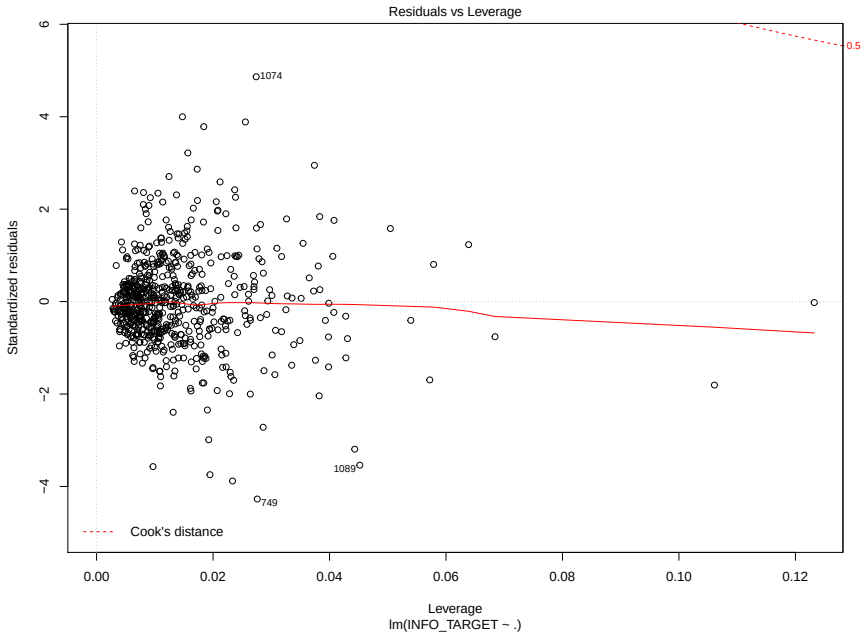
$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{nMSE}, \tag{1}$$

where $\hat{y}_j$ is the fitted response fitted (with all observations) and $\hat{y}_{j(i)}$ is the fitted response without observation i, n is the number of regression coefficients (with the intercept) and MSE is the mean squared error.

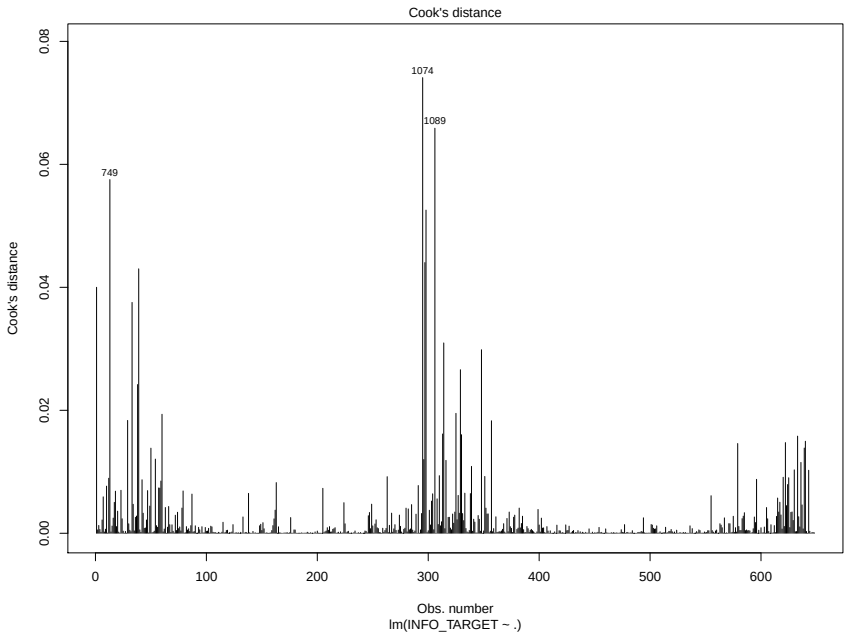
It can also be expressed using the leverage:

$$D_i = \frac{e_{Si}}{nMSE} \frac{h_{ii}}{(1 - h_{ii})^2} \tag{2}$$

The original value proposed for the cut-off value is 1. Values of $4/n$ where n is the sample size, or $8/(n - 2p)$ where p is the number of regressors.



- Helps find influential cases; outlying cases with high leverage
- Points lying in the low and high right corners, the points with high Cook distance.
- The dashed lines represent Cook distance of 1 and 0.5
- In this plot we can observe high leverage points with little leverage and large residuals with small leverage. Nothing really problematic



Cook's distance per observation.

Fitting a generalized linear model (GLM)

```
glm(formula = INFO_TARGET ~ ., family = Gamma(link = "log"),
    data = data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.96601  -0.20973  -0.04454   0.15974   1.08426

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.9487804  0.0747056  39.472  < 2e-16 ***
PM10_max_m   0.0101776  0.0008556  11.895  < 2e-16 ***
PM10_mean_y  0.0067450  0.0010795   6.248 7.58e-10 ***
wind_gust_m -0.0258245  0.0139635  -1.849  0.06486 .
m.t2m_12_t   0.0226227  0.0077442   2.921  0.00361 **
m.t2m_18_t  -0.0259889  0.0082492  -3.150  0.00171 **
t2m_mean_m  -0.0140274  0.0059051  -2.375  0.01782 *
t_inv_y      0.0016731  0.0067912   0.246  0.80548
t_inv_m      0.0152750  0.0071652   2.132  0.03340 *
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

(Dispersion parameter for Gamma family taken to be 0.09301874)

    Null deviance: 226.142  on 647  degrees of freedom
Residual deviance:  59.473  on 639  degrees of freedom
AIC: 4632.4

Number of Fisher Scoring iterations: 6
```

Interpretation of coefficients

- Linear regression -> If we add 1 to a independent variable and keep everything else constant the depended variable will change by the value of the regression coefficients.
- GLM -> The regression coefficients are now transformed by the link function, you need an inverse transform to interpret them

Model diagnosis

In ordinary linear regression the ordinary residual is the difference $\hat{y} - y$, it represents the statistical error $\epsilon = E(y|\eta) - y$.

There is no additive error in the definition of the GLM.

- **Response residuals (raw residuals)**: the difference between the observed value and its estimated expected value $y_i - \hat{\eta}_i$
- **Pearson residuals**: $e_{Pi} = \frac{y_i - \hat{\eta}_i}{\sqrt{(Var(\hat{\eta}_i))}}$; where $Var(\hat{\eta}_i$ is the variance of the estimated value (different for different distributions).
- **Deviance residuals:** $e_{Di} = sign(y_i - \hat{\eta}_i)\sqrt{di}$

Deviance is a generalization of the idea of using the sum of squares of residuals in ordinary least squares to cases where model fitting is achieved by maximum likelihood.

Deviance residuals

Residual deviance is 2 x (loglik(Saturated Model) - loglik(Proposed Model))

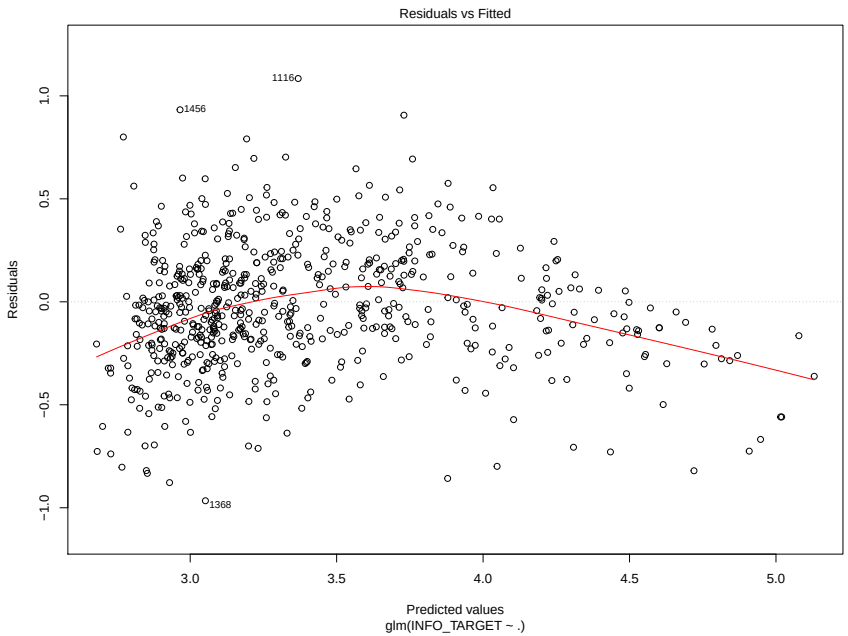
1. The first term of the equation is the likelihood of the data given the saturated model
2. The second term of the equation is the likelihood of the data given the proposed model

A **saturated model** is a model with as many estimated parameters as observations.
From the equation of residual deviance we can get the deviance residuals where.
Deviance residual for observation i is estiamted as:

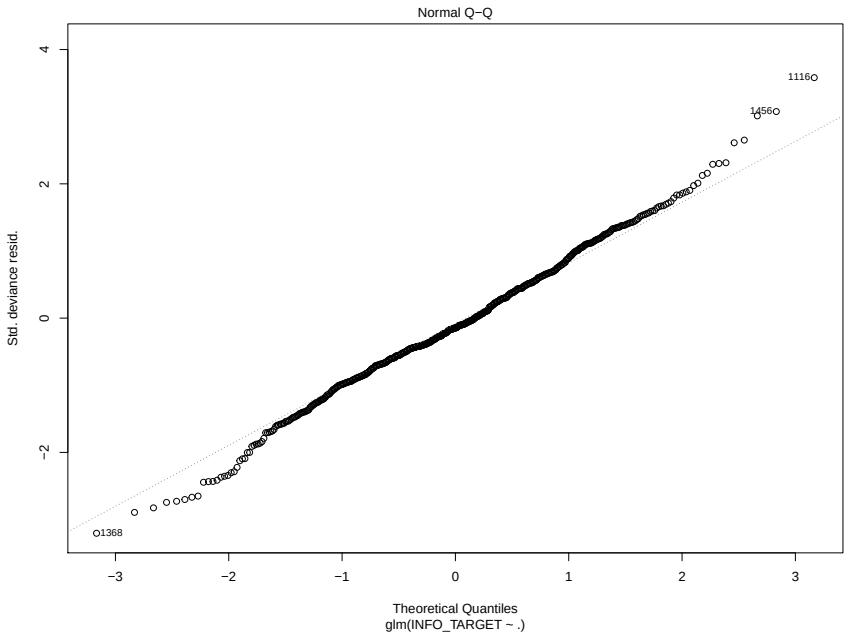
$$\sqrt{2(log(likelihood\ of\ i\ given\ the\ saturated\ model)log(likelihood\ of\ i\ given\ the\ proposed\ model))}$$

They are analogous to ordinary residuals in ordinary linear regression.

Diagnostic plots



Raw residuals are not very informative in GLM analysis (we do not expect a flat red line). Heteroscedasticity is not a problem in GLM.



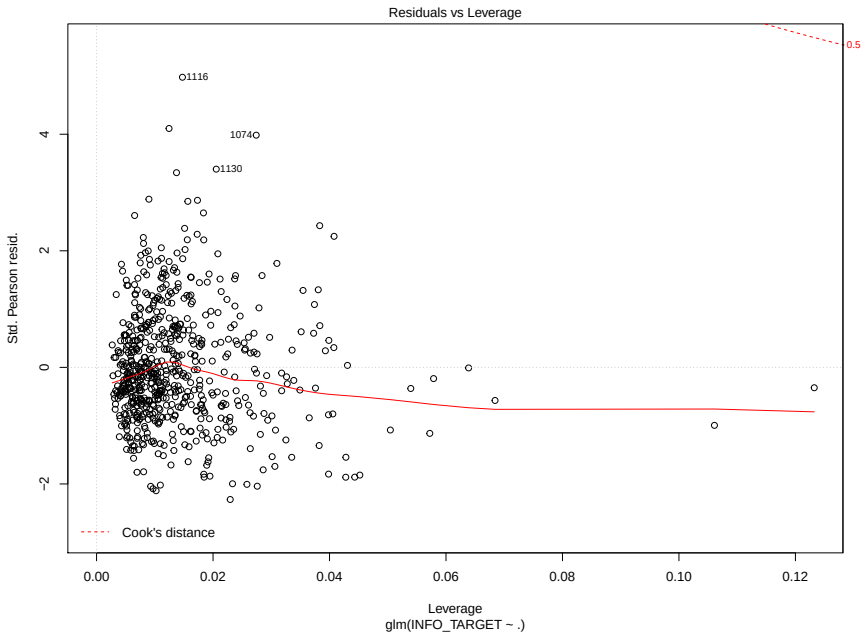
Similar to ordinary residuals of ordinary linear regression, the deviance residuals of GLM should are nearly normally distributed.

Approximation of Cook's distance

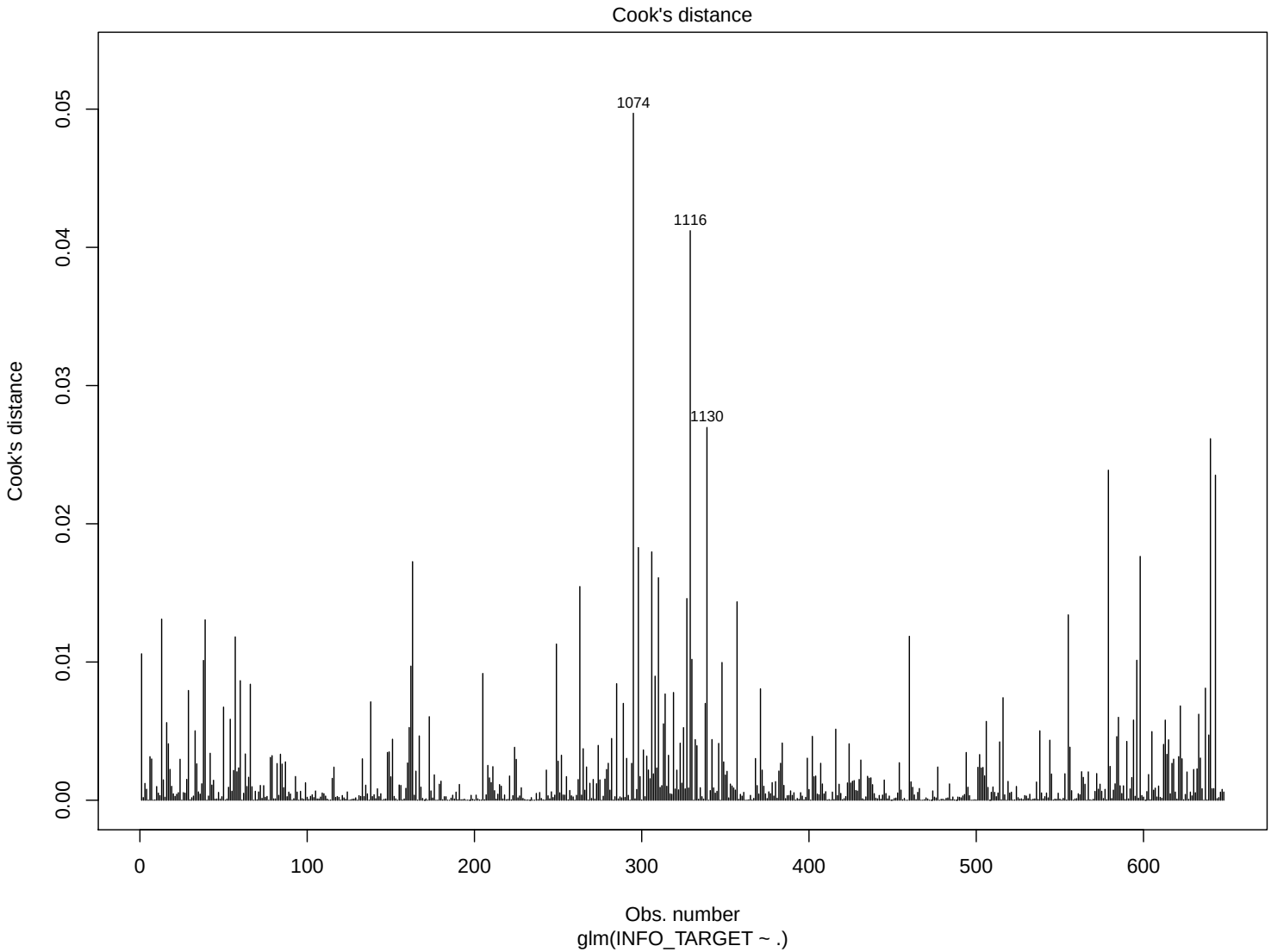
The approximate Cook's distance for GLM is calculated as:

$$D_i = \frac{e_{PSi}}{n} \frac{h_{ii}}{(1 - h_{ii})} \tag{3}$$

Where e_{PSi} are Pearson residuals and h_{ii} is the leverage. As with ordinary linear regression it helps estimate the high influence points.



As with ordinary linear regression this plot enables spotting high influence points.


Cook's distance
1074
1116
1130
Cook's distance
0.00
0.01
0.02
0.03
0.04
0.05
0
100
200
300
400
500
600
Obs. number
glm(INFO_TARGET ~ .)

Logistic regression

We divide the concentrations in two classes, the concentrations below 40 $\mu g/m^3$ are regarded as high others as low.

```
glm(formula = class ~ ., family = binomial, data = data2)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.76215  -0.06772   0.17385   0.32650   3.10577
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.48657    0.90212   3.865 0.000111 ***
PM10_max_m  -0.06029    0.01057  -5.702 1.18e-08 ***
PM10_mean_y -0.04289    0.01277  -3.359 0.000783 ***
wind_gust_m  0.09733    0.18236   0.534 0.593534
m.t2m_12_t  -0.02608    0.08863  -0.294 0.768601
m.t2m_18_t   0.07968    0.09926   0.803 0.422126
t2m_mean_m   0.08174    0.07112   1.149 0.250438
t_inv_y     -0.03925    0.07571  -0.518 0.604151
t_inv_m     -0.07819    0.08121  -0.963 0.335618
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 765.73  on 647  degrees of freedom
Residual deviance: 318.08  on 639  degrees of freedom
AIC: 336.08
Number of Fisher Scoring iterations: 6
```

Diagnostic plots for logistic regression

As in the example with the gamma model, the logistic regression can be to some extent diagnosed with diagnostic plots. On the next slide are the diagnostic plots plotted with the R boot library and the glm.diag.plot().

1. The first plot shows the residuals vs. the fitted values. Logistic regression always exhibits a "double“ response (two separate curves).
2. The Q-Q plot indicate a problem with the fit of the logistic regression model, the deviance residuals do not fit the normal distribution well.
3. The dashed lines in the two plots of Cook's statistics use threshold estimated as $8/(n - 2p)$ which is much lower than 0.5 or 1.

Diagnostic plots