ST346 Week 7

Contents

Preface

These slides are a slight adaptation from the original slides developed by Prof Martyn Plummer for the module.

If you find any typos, please inform the module leader.

These materials are solely for your own use and you must not distribute these in any format. **Do not upload these materials to the internet or any filesharing sites nor provide them to any third party or forum.**

Chapter 7 Diagnostics for GLMs

7.1 Residuals for GLMs

For linear models we used the notation

$$Y_i \quad = \quad \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i$$

which clearly separates the structural part of the model $\boldsymbol{x}_i^T \boldsymbol{\beta}$ from the random error part (ϵ_i). The residual

$$r_i \quad = \quad y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}$$

can be thought of as an estimate of ϵ_i.

For GLMs there is no clear separation of structural and error components.

Residuals are calculated after model fitting. There are three basic requirements for residuals:

1. A residual measures the discrepancy between the observed value y_i and the fitted value $\widehat{\mu}_i = \widehat{\mu}_i(\widehat{\boldsymbol{\beta}})$.
2. A residual is zero if $y_i = \widehat{\mu}_i$.
3. A residual is signed:
   - positive if $y_i > \widehat{\mu}_i$;
   - negative if $y_i < \widehat{\mu}_i$.

For GLMs there are (at least) four definitions of residuals that satisfy these requirements.

7.1.1 Response residuals

The **response residual** is the difference between the observed and fitted values

$$r_i^{(r)} \quad = \quad y_i - \widehat{\mu}_i.$$

For GLMs we need additional information about prior weight w_i and variance function $V(\mu_i)$ to interpret the response residuals.

7.1.2 Pearson residuals

The **Pearson residual** adjusts for the variance function and prior weight.

$$r_i^{(p)} \quad = \quad \frac{y_i - \widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i)/w_i}}.$$

As sample size $n \to \infty$,

$$\begin{aligned} \mathbb{E}\left(r_i^{(p)}\right) \quad &\longrightarrow \quad 0, \\ \mathbb{V}ar\left(r_i^{(p)}\right) \quad &\longrightarrow \quad \phi(1 - h_i) \end{aligned}$$

where h_i is the leverage of the ith observation (see later).

7.1.3 Deviance residuals

The **deviance residual** is

$$r_i^{(d)} \quad = \quad \text{sign}(y_i - \widehat{\mu}_i)\sqrt{w_i d(y_i, \widehat{\mu}_i)}$$

where $d(y, \mu)$ is the unit deviance and

$$\text{sign}(T) \quad = \quad \begin{cases} 1 & \text{if } T > 0 \\ 0 & \text{if } T = 0 \\ -1 & \text{if } T < 0. \end{cases}$$

Recall from Chapter 4 that the total deviance from a fitted model is

$$D(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}) \quad = \quad \sum_{i=1}^{n} w_i \; d(y_i, \widehat{\mu}_i)$$

This can be rewritten in terms of the deviance residuals

$$D(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}) \quad = \quad \sum_{i=1}^{n} \left[r_i^{(d)} \right]^2.$$

This generalizes the result for (unweighted) linear models that the deviance is the sum of squares of the residuals.

In general, the deviance residual is a non-linear function of the outcome variable Y_i and is not guaranteed to have expectation zero:

$$\mathbb{E}(r_i^{(d)}) \quad \neq \quad 0.$$

This limits the usefulness of deviance residuals for diagnostic plots.

7.1.4 Working residuals

The **working residuals** are the residuals from the last iteration of the IWLS algorithm

$$r_i^{(w)} \quad = \quad \widehat{z}_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}$$

where $\widehat{z}_i$ is the working observation

$$\widehat{z}_i \quad = \quad (y_i - \widehat{\mu}_i) g'(\widehat{\mu}_i) + \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}$$

The working residual can be simplified to

$$r_i^{(w)} \quad = \quad (y_i - \widehat{\mu}_i) g'(\widehat{\mu}_i)$$

Like the response residuals, the working residuals do not account for difference in variance between observations. You need the working weights $\widehat{W}_i$ to interpret them.

Exercise 18 - Residuals in a linear model

Show that for an unweighted normal linear model

$$Y_i \quad \sim \quad \mathcal{N}(\boldsymbol{x}_i^T\boldsymbol{\beta},\ \phi)$$

all residuals are the same

$$r_i^{(r)} \quad = \quad r_i^{(p)} \quad = \quad r_i^{(d)} \quad = \quad r_i^{(w)}.$$

Hint: for a unweighted normal linear model we have

$$\begin{aligned} V(\mu) &= 1, \\ g(\mu) &= \mu, \\ d(y,\mu) &= (y-\mu)^2. \end{aligned}$$

7.1.5 Residuals and weights in R

The `residuals()` function will extract the various residuals from a fitted model.

Use the type argument to specify which residuals you want.

- `residuals(glm.out, type="response")`
- `residuals(glm.out, type="pearson")`
- `residuals(glm.out, type="deviance")`
- `residuals(glm.out, type="working")`

If you omit the type argument, then you get the deviance residuals by default. I recommend you do always specify type.

The `weights()` function will extract the weights from a fitted model.

Use the type argument to specify which weights you want.

- `weights(glm.out, type="prior")`
- `weights(glm.out, type="working")`

If you omit the type argument, then you get the prior weights by default.

7.1.6 Which residuals to use?

- Use Pearson residuals primarily for diagnostics.
- Deviance residuals may be more useful than Pearson residuals for QQ plots (see later).
- Pearson and deviance residuals are often the same when the amount of information in one observation is large. This is not the case for Poisson and binomial data with small counts.
- Response and working residuals exist but are not useful for model criticism. However, we use them to calculate other quantities.

7.2 Leverage and standardized residuals

7.2.1 Leverage in normal linear models - Revision

Consider the weighted linear model

$$Y_i \quad \sim \quad \mathcal{N}(\boldsymbol{x}_i^T\boldsymbol{\beta},\ \phi/w_i).$$

The residual $R_i = Y_i - \widehat{\mu}_i$ has mean and variance

$$\begin{aligned} \mathbb{E}(R_i) &= 0, \\ \mathbb{V}ar(R_i) &= \frac{(1-h_i)\phi}{w_i}, \end{aligned}$$

where $h_i = \boldsymbol{H}_{ii}$ is the ith diagonal element of the hat matrix

$$\boldsymbol{H} \quad = \quad \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}$$

and $\boldsymbol{W} = \text{diag}(w_1, \ldots, w_n)$. See Chapter 1.

The **hat value** $h_i \in [0, 1]$ is a measure of **leverage** for observation i.

Observations with high leverage force the fitted value from the model to be close to the observed value.

If observation i has maximum leverage $h_i = 1$, then

$$\mathbb{V}ar(R_i) \quad = \quad 0.$$

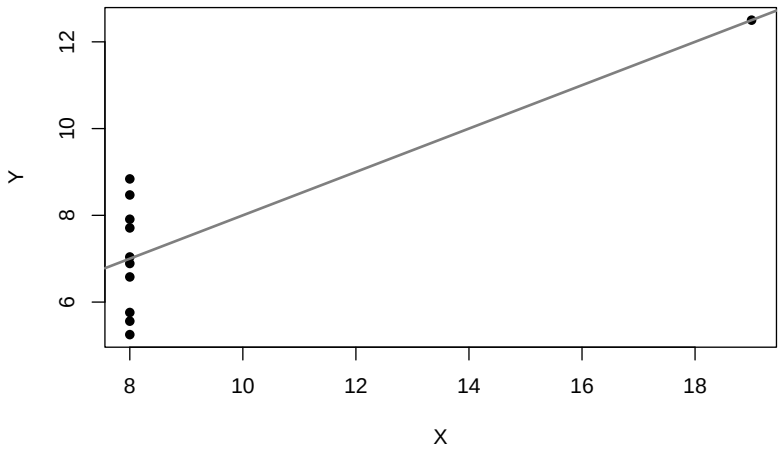
Hence $r_i = \mathbb{E}(R_i) = 0$ and

$$\widehat{\mu}_i \quad = \quad y_i.$$

So observed and fitted values are identical.

Example: Anscombe's Quartet

In this example from Anscombe's quartet, the observation on the right has maximum leverage $h_i = 1$ forcing the fitted line through the observed point.



Example: Simple linear regression

Suppose $\mu_i = \alpha + \beta x_i$. Then

$$h_i \quad = \quad \frac{1}{n} + \frac{(x_i - \overline{x}\,)^2}{\sum_{j=1}^{n}(x_j - \overline{x}\,)^2}.$$

So the hat value h_i measures how far away the predictor variable x_i is from the mean $\overline{x}$.

Suppose $x \in \{0, 1\}$ is a binary predictor with

- n_0 observations with $x_i = 0$ and
- n_1 observations with $x_i = 1$.

Then

$$h_i \quad = \quad \begin{cases} \frac{1}{n_0} & \text{if } x_i = 0, \\ \frac{1}{n_1} & \text{if } x_i = 1. \end{cases}$$

You should expect to see observations with high leverage if they have unusual values for the predictor variables:

- for continuous variables if they have values far from mean;
- for categorical variables if they belong to a sparse category.

7.2.2 Standardized residuals in normal linear models - Revision

The standardized residual for a normal linear model is the residual divided by an estimate of its standard deviation.

$$r_i^{(s)} \quad = \quad \frac{\sqrt{w_i}(y_i - \widehat{\mu}_i)}{\sqrt{(1 - h_i)\widehat{\phi}}}$$

where $\widehat{\phi}$ is the unbiased estimator of the dispersion parameter ϕ, that is

$$\widehat{\phi} \quad = \quad \frac{1}{n - p} \sum_{i=1}^{n} w_i \, (y_i - \widehat{\mu}_i)^2.$$

The standardized residual then has **asymptotic variance** 1 as $n \to \infty$.

7.2.3 Leverage for GLMs

The final iteration of the IWLS algorithm is based on the approximation

$$\widehat{z}_i \quad \sim \quad \mathcal{N}\left(\boldsymbol{x}_i^T\boldsymbol{\beta},\ \frac{\phi}{\widehat{W}_i}\right)$$

for working observation $\widehat{z}_i$ and working weight $\widehat{W}_i$ evaluated at $\widehat{\boldsymbol{\beta}}$.

Hence for the working residuals

$$\begin{aligned}\mathbb{E}(r_i^{(w)}) &\approx 0,\\ \mathbb{V}ar(r_i^{(w)}) &\approx (1-\widehat{h}_i)\phi/\widehat{W}_i\end{aligned}$$

where $\widehat{h}_i = \widehat{\boldsymbol{H}}_{ii}$ for the working hat matrix

$$\widehat{\boldsymbol{H}} \quad = \quad \boldsymbol{X}\left(\boldsymbol{X}^T\widehat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\widehat{\boldsymbol{W}}$$

and $\widehat{\boldsymbol{W}} = \text{diag}(\widehat{W}_1,\ldots,\widehat{W}_n)$ is a diagonal matrix of working weights.

If we standardize the working residuals, then we get

$$r_i^{(sw)} \quad = \quad \frac{\sqrt{\widehat{W}_i}(\widehat{z}_i-\widehat{\mu}_i)}{\sqrt{(1-\widehat{h}_i)\widehat{\phi}}}.$$

Substituting the definitions of $\widehat{z}_i$ and $\widehat{W}_i$ (after some cancellations) gives

$$r_i^{(sw)} \quad = \quad \frac{\sqrt{w_i}(\widehat{y}_i-\widehat{\mu}_i)}{\sqrt{(1-\widehat{h}_i)V(\widehat{\mu}_i)\widehat{\phi}}} \quad = \quad \frac{r_i^{(p)}}{\sqrt{(1-\widehat{h}_i)\widehat{\phi}}} \quad = \quad r_i^{(sp)}.$$

This is also the **standardized Person residual** $r_i^{(sp)}$.

We can also standardize the deviance residual using the working hat matrix

$$r_i^{(sd)} \quad = \quad \text{sign}(y_i-\widehat{\mu}_i)\frac{\sqrt{w_i d(y_i,\widehat{\mu}_i)}}{\sqrt{(1-\widehat{h}_i)\widehat{\phi}}}.$$

The `rstandard()` function extracts standardized residuals from a fitted model

- `rstandard(glm.out, type="deviance")`
- `rstandard(glm.out, type="pearson")`

7.3 Influence

Model fitting should not depend on a handful of observations. We therefore need to consider what happens if we drop an observation from the data set and refit the model. If the estimates or predicted values change substantially, we say the observation is **influential**.

7.3.1 Influence in normal linear models

Consider the weighted normal linear model:

$$\begin{aligned} Y_i &\sim \mathcal{N}\left(\mu_i, \frac{\phi}{w_i}\right), \\ \mu_i &= \boldsymbol{x}_i \boldsymbol{\beta}^T. \end{aligned}$$

We will use a superscript (i) to denote a quantity estimated after deleting observation i. Then, the effect on the estimated parameters can be measured as

$$\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{(i)} = (\boldsymbol{X}\boldsymbol{W}\boldsymbol{X})^{-1} \boldsymbol{x}_i w_i \left[\frac{r_i}{1-h_i}\right]$$

where $\boldsymbol{W} = \text{diag}(w_1, \ldots, w_n)$ and $r_i = y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}$ (Proof omitted).

This quantity is known as **dfbeta**. It can be calculated without refitting the model.

The effect on fitted values is given by

$$\begin{aligned} \widehat{\mu}_j - \widehat{\mu}_j^{(i)} &= \boldsymbol{x}_j^T \left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{(i)}\right) \\ &= \boldsymbol{x}_j^T (\boldsymbol{X}\boldsymbol{W}\boldsymbol{X})^{-1} \boldsymbol{x}_i w_i \left[\frac{r_i}{1-hi}\right] \\ &= \boldsymbol{H}_{ji} \left[\frac{r_i}{1-hi}\right]. \end{aligned}$$

Note that dropping the observation i perturbs the fitted values for $j \neq i$.

The **Cook's distance** for a linear model summarizes the changes to the fitted values across all observations.

$$D_i^{\text{Cook}} = \frac{1}{p\widehat{\phi}} \sum_{j=1}^{n} w_j \left(\widehat{\mu}_j - \widehat{\mu}_j^{(i)}\right)^2$$

where $\widehat{\phi}$ is the unbiased estimator of ϕ.

We may simplify the expression for the Cook's distance as follows. Substituting $\widehat{\mu}_j - \widehat{\mu}_j^{(i)}$ gives

$$\begin{aligned} D_i^{\text{Cook}} &= \frac{1}{p\widehat{\phi}} \sum_{j=1}^{n} w_j \boldsymbol{H}_{ji}^2 \left[\frac{r_i}{1-hi}\right]^2 \\ &= \frac{r_i^2}{(1-h_i)^2 p\widehat{\phi}} \sum_{j=1}^{n} w_j \boldsymbol{H}_{ji}^2 \\ &= \frac{r_i^2}{(1-h_i)^2 p\widehat{\phi}} \left(\boldsymbol{H}^T \boldsymbol{W} \boldsymbol{H}\right)_{ii} \\ &= \frac{r_i^2}{(1-h_i)^2 p\widehat{\phi}} \left(\boldsymbol{W} \boldsymbol{H}\right)_{ii} \\ &= \frac{r_i^2}{(1-h_i)^2 p\widehat{\phi}} h_i w_i \\ &= \frac{1}{p} \frac{w_i\ r_i^2}{\widehat{\phi}(1-h_i)} \frac{h_i}{(1-h_i)} \end{aligned}$$

The above uses the result $\boldsymbol{H}^T \boldsymbol{W} \boldsymbol{H} = \boldsymbol{W} \boldsymbol{H}$ (proof omitted).

Recall the standardized residual for a linear model

$$r_i^{(s)} = \frac{\sqrt{w_i} r_i}{\sqrt{\widehat{\phi}(1-h_i)}}.$$

and so the Cook's distance is given by

$$D_i^{\text{Cook}} = \frac{1}{p} \left[r_i^{(s)}\right]^2 \frac{h_i}{(1-h_i)}.$$

7.3.2 Cook's distance for GLMs

We take the values from the last iteration of the IWLS algorithm

$$D_i^{\text{Cook}} = \frac{1}{p} \left[r_i^{(sw)}\right]^2 \frac{\hat{h}_i}{(1-\hat{h}_i)}$$

where $r_i^{(sw)}$ is the standardized working residual.

Recall that $r_i^{(sw)} = r_i^{(sp)}$, the standardized Pearson residual and so

$$D_i^{\text{Cook}} = \frac{1}{p} \left[r_i^{(sp)}\right]^2 \frac{\widehat{h}_i}{(1-\widehat{h}_i)}.$$

Thus the Cook's distance of observation i is large if

- the observation has high leverage ($\widehat{h}_i$ close to 1) and/or
- the absolute value of the standardized Pearson residual $|r_i^{(sp)}|$ is large.

As sample size $n \to \infty$

$$\mathbb{E}\Big(\sum_{i=1}^{n} D_i^{\text{Cook}}\Big) \quad \longrightarrow \quad 1.$$

(Proof omitted.)

Therefore we expect the influence of individual observations to diminish at the rate $O(n^{-1})$.

There is no commonly accepted threshold for when an observation should be considered **influential**. Instead look at the distribution of D_i^{Cook} across observations and look for values that are much larger than the others.

Question: If we do identify influential observations, should we delete them?

Not necessarily. We should only delete data if we have a very strong justification for doing so!

Investigate first, there may be good reasons why an observation is influential.

7.3.3 Summary

1. Model fitting should not depend on a handful of observations.
2. Look for observations with high influence:
   - large absolute residuals and/or
   - high leverage.
3. Investigate observations with much larger influence than the rest (if any).
4. Sensitivity analysis to check stability of results if you do delete an observation.

7.4 Diagnostic plots in R

The `plot()` method for fitted GLM objects in R produces a sequence of diagnostic plots, using

- Pearson residuals,
- fitted values,
- standardized Pearson residuals,
- hat values,
- Cook's distance.

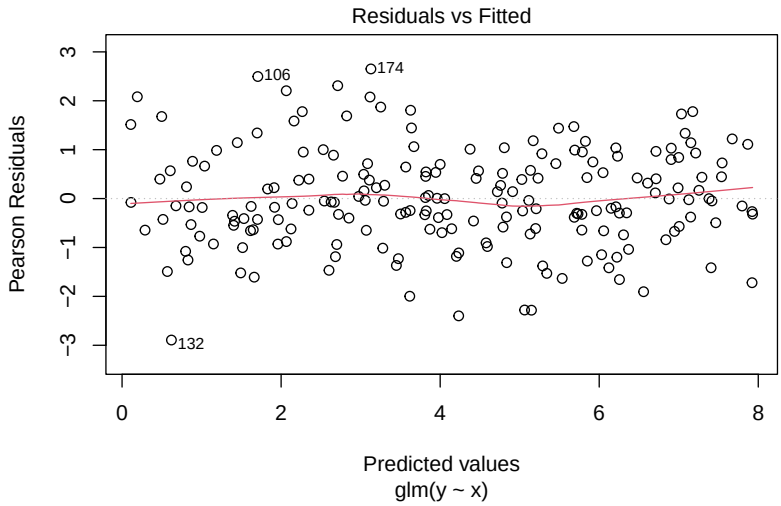
Not all of these plots are appropriate all the time. You can select a single plot using the `which` argument.

7.4.1 Residuals vs fitted values

With `which=1` you get a scatter plot of:

- Pearson residuals on the y-axis ($r_i^{(p)}$);
- fitted values on the scale of the linear predictor on the x-axis ($\widehat{\eta}_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$).

A red line shows a smooth curve fitted to the residuals. If the structural part of the model is correct (link function and predictor variables), the smooth curve should remain close to the value zero on the y-axis.



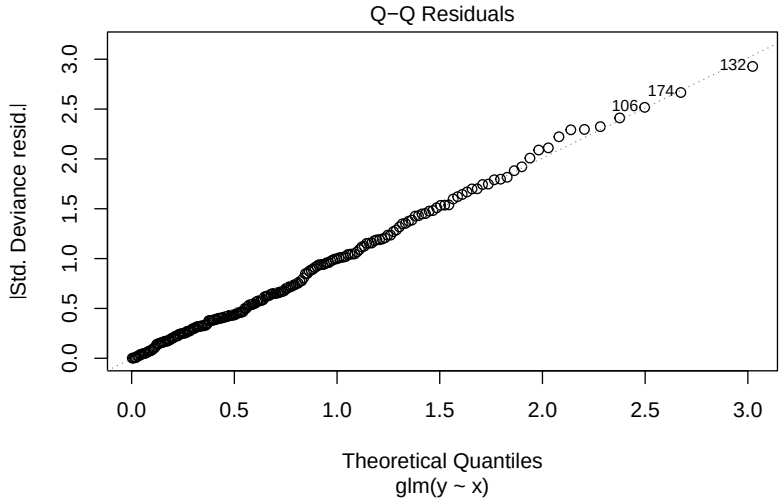
7.4.2 QQ plot

With `which=2` you get a quantile-quantile (QQ) plot with

- sorted absolute values of the standardized Pearson residuals on the y-axis;
- theoretical quantiles of the standard half-normal distribution on the x-axis.

Ideally, the points should line up in a straight line through the origin and with slope 1.

The observations with the 3 largest values of $r_i^{(sp)}$ are labelled with their row numbers.



The behaviour of the QQ plots was recently changed (by Prof M. Plummer) in `R` version 4.3.0. This is one of the few diagnostic uses of the deviance residuals.

The absolute value of the standardized deviance residuals has a half-normal distribution when the **saddlepoint approximation** applies:

- exact for normal and inverse Gaussian,
- approximate for
  - Gamma for shape > 3 ($\phi < 1/3$),
  - Poisson for $y \geq 3$,
  - Binomial for $3 \leq y \leq m - 3$.

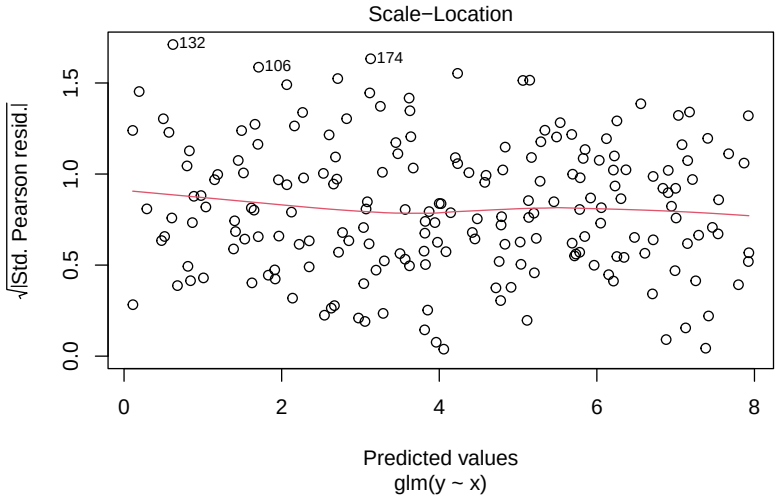
7.4.3 Scale Location plot

With `which=3` you get a scatterplot very similar to the plot of residuals vs fitted values (`which=1`), but on the y-axis is $\sqrt{|r_i^{(sp)}|}$ instead of $r_i^{(p)}$.

A red line shows a smooth curve fitted to $\sqrt{|r_i^{(sp)}|}$.

If the variance function is correct, the red line should show no upward or downward trend.

Observations with the largest standardized Pearson residuals (in absolute value) are labelled with the row numbers.



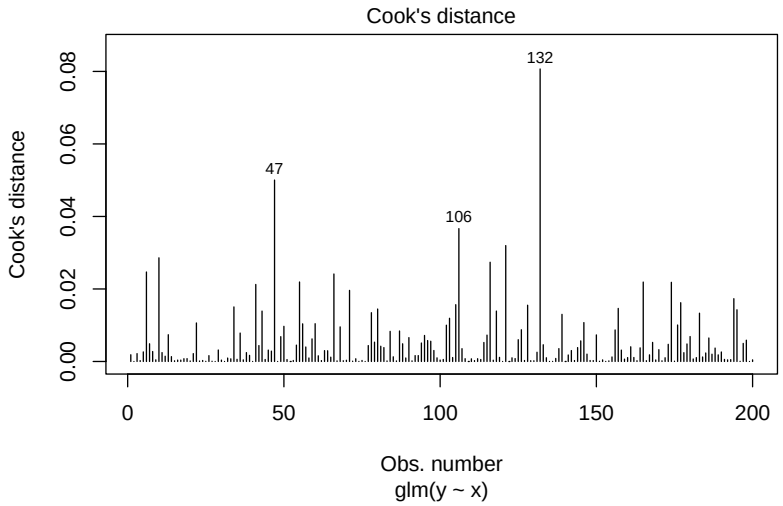
7.4.4 Cook's distance

With `which=4` you get a plot of Cook's distance against observation number.

The plot is an index plot with vertical lines rising from zero on the y-axis to the Cook's distance value.

The observations with the 3 largest values of Cook's distance are labelled with the row number.

This plot is normally omitted if you do not specify the `which` argument.



7.4.5 Standardized residuals versus leverage

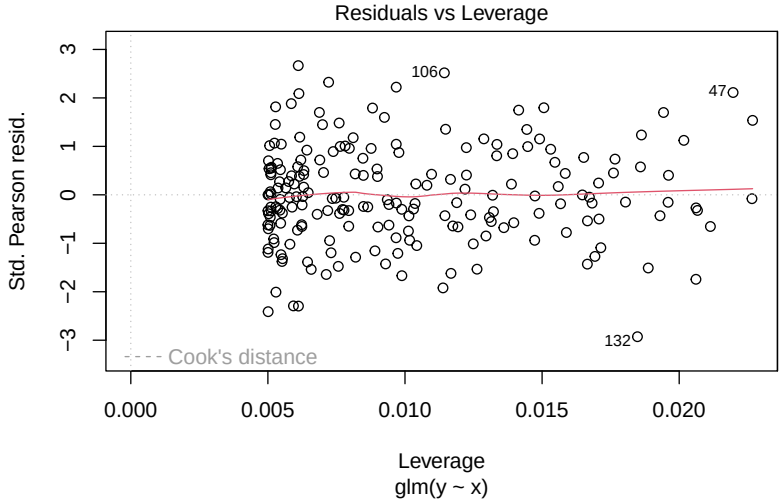
With `which=5` you get a scatter plot of:

- standardized Pearson residuals on the y-axis ($r_i^{(sp)}$);
- hat values on the x-axis ($\widehat{h}_i$).

A red line shows a smooth curve fitted to the residuals.

Note that Cook's distance can be calculated from $r_i^{(sp)}$ and $\widehat{h}_i$.

- The observations with the 3 largest values of D_i^{Cook} are labelled with their row numbers.
- Dashed grey lines show contours of the Cook's statistic for $D^{\text{Cook}} = 0.5$ and $D^{\text{Cook}} = 1$ to highlight highly influential observations.
- Contours may not be shown if they lie outside the boundaries of the plot, but a legend is still produced.



7.4.6 When diagnostics break down

For discrete outcomes (Poisson and binomial) with small counts, the diagnostic plots do not work well.

For binary data, only the first diagnostic plot is of interest. Do not try to interpret the other plots (except leverage plot `which=4`).

Binned residual plots may be more visually appealing for discrete data.

Exercise 19 - Diagnostic plots

Perform a residual analysis for the examples encountered in the computer practicals.