Generalized Linear Models

Lecture 3. Hypothesis testing. Goodness of fit. Model diagnostics

Models

Let $M(\mathbf{X}_r)$ be a model with design matrix $\mathbf{X}_r$ (with r columns)
$r \leq n$ (n – number of rows), design matrix must be nonsingular

Full model or saturated model: $r = n$
Null model or constant model: $r = 1$, design matrix is vector of 1s, $\mathbf{X}_1 = \mathbf{1}$.

We get sequences of models (from constant to saturated), where the null model is the smallest model and the saturated model is the most complex:

$$M(\mathbf{1}),\ M(\mathbf{X}_2),\ \ldots,\ M(\mathbf{X}_r)$$

Let us fix a design matrix $\mathbf{X} = \mathbf{X}_r$. To obtain some (simpler) design matrix $\mathbf{X}_l$ and corresponding model $M(\mathbf{X}_l)$, one can think of two ways:

1. select some columns $\{j_1, j_2, \ldots, j_l\}$ from $\mathbf{X}$
2. delete s columns $\{j'_1, j'_2, \ldots, j'_s\}$ from design matrix $\mathbf{X}$, so $\mathbf{X}_l$ has $l = r - s$ columns.

Models. Restrictions

Let us have a model $M(\mathbf{X}_l)$ with design matrix $\mathbf{X}_l$ (obtained as described on previous slide).

Then we say that:

- $M(\mathbf{X}_l)$ is lower model (restricted), $M(\mathbf{X})$ is upper model (non-restricted)
- $M(\mathbf{X}_l)$ is submodel of $M(\mathbf{X})$
- $M(\mathbf{X}_l)$ is nested in $M(\mathbf{X})$, some of the parameters in $M(\mathbf{X})$ are set equal to zero in $M(\mathbf{X}_l)$

Restrictions on $M(\mathbf{X})$: $\beta_{j'_1} = \ldots = \beta_{j'_s} = 0$, or, in general $\mathbf{C}\beta = \boldsymbol{d}$,

where

- $\mathbf{C}$ is $s \times r$ restriction matrix (with linearly independent columns)
- $\boldsymbol{d}$ is a constant $s \times 1$ vector (often $\boldsymbol{d} = \mathbf{0}$)

Hypothesis testing

Look at two arbitrary models, say $M(\mathbf{X}_1)$ and $M(\mathbf{X}_2)$, with r_1 and r_2 parameters, respectively
Assume that $M(\mathbf{X}_1)$ is nested in $M(\mathbf{X}_2)$, $r_2 > r_1$

$$\begin{cases} H_0: & M(\mathbf{X}_1) \sim M(\mathbf{X}_2), \text{ models are as good} \\ H_1: & M(\mathbf{X}_2) \text{ is better} \end{cases}$$

Let us assume that $\mathbf{X}_1$ is obtained from $\mathbf{X}_2$ by deleting columns $j_1, \ldots, j_q$. Then

$$H_0: \beta_{j_1} = \ldots = \beta_{j_q} = 0; \quad H_1: \ \exists i, \ \beta_{j_i} \neq 0, i = 1, \ldots, q$$

General hypotheses

$$H_0: \mathbf{C}\beta = \boldsymbol{d}; \quad H_1: \mathbf{C}\beta \neq \boldsymbol{d}$$

$\mathbf{C}$ – known deterministic full rank matrix $(q \times r_2, \ q = r_2 - r_1)$
$\boldsymbol{d}$ – known deterministic vector $(q \times 1)$

Statistics for testing general hypothesis

- Likelihood ratio statistic
- Wald statistic
- Score statistic

All mentioned statistics have asymptotically χ^2-distribution with degrees of freedom equal to the difference in the number of parameters estimated by the two models $(r_2 - r_1)$.

Likelihood ratio statistic

The likelihood ratio test is performed by estimating two models and comparing the fit of one model to the fit of the other (comparing maximum values of likelihood).

Let us have upper model $M_2 = M(\mathbf{X}_2)$ and lower model $M_1 = M(\mathbf{X}_1)$, and their maximum values of likelihood, respectively:

$$L(\hat{\beta}_{M_1}; \boldsymbol{y}) = \max_{\beta_{M_1}} L(\beta; \boldsymbol{y}), \quad L(\hat{\beta}_{M_2}; \boldsymbol{y}) = \max_{\beta_{M_2}} L(\beta; \boldsymbol{y})$$

Definition. Likelihood ratio (λ^*)

$$\lambda^* = \frac{L(\hat{\beta}_{M_1}; \boldsymbol{y})}{L(\hat{\beta}_{M_2}; \boldsymbol{y})}$$

$0 \leq \lambda^* \leq 1$ Under regularity conditions $\lambda = -2 \ln \lambda^* \sim \chi^2_{r_2 - r_1}$

Common form of likelihood ratio statistic: $\lambda = -2 \ln \lambda^* = -2(l(\hat{\beta}_{M_1}) - l(\hat{\beta}_{M_2}))$

Wald statistic

The Wald test is a parametric statistical test named after Hungarian statistician Abraham Wald (1943).

Statistic is based on the asymptotic normality property of MLE (under regularity conditions):

$$\hat{\beta} \overset{a}{\sim} N(\beta, \mathbf{F}^{-1}(\hat{\beta}))$$

Definition. Wald statistic (w)

Wald statistic w is defined as

$$w = (\mathbf{C}\hat{\beta} - \boldsymbol{d})^T [\mathbf{C}\mathbf{F}^{-1}(\hat{\beta})\mathbf{C}^T]^{-1}(\mathbf{C}\hat{\beta} - \boldsymbol{d})$$

and used for testing hypotheses $H_0 : \mathbf{C}\beta = \boldsymbol{d}$ vs $H_1 : \mathbf{C}\beta \neq \boldsymbol{d}$

Wald statistic measures certain weighted distance between $\mathbf{C}\hat{\beta}$ and $\boldsymbol{d}$.

The Wald statistic uses the behaviour of the likelihood function at the ML estimate $\hat{\beta}$.

A special case of Wald statistic

Let us have $M_2 = M(\mathbf{X}_2)$ (upper model) and $M_1 = M(\mathbf{X}_1)$ (lower model)
Let us partition the vector of coefficients into two components:

$$\beta_{M_2} = (\beta_{M_1}^T, \beta^T)^T,$$

where β are parameters which are not in M_1, but are in M_2 (q – number of extra parameters)
Consider the hypothesis ($\boldsymbol{d} = 0$)

$$H_0 : \beta = 0, \;\; H_1 : \beta \neq 0$$

Wald statistic: $w = \beta^T \boldsymbol{\Sigma}_\beta^{-1} \beta \sim \chi_q^2$
If $q = 1$, we get $w = \frac{\beta^2}{s_\beta^2}$ and we recognize that its square root is t-**statistic**:

$$t = \frac{\beta}{s_\beta}$$

Score statistic

The score statistic is also known as the Rao efficient score statistic or Lagrange multiplier statistic (Rao, 1948)

Based on the asymptotic normality property of score function (under regularity conditions)

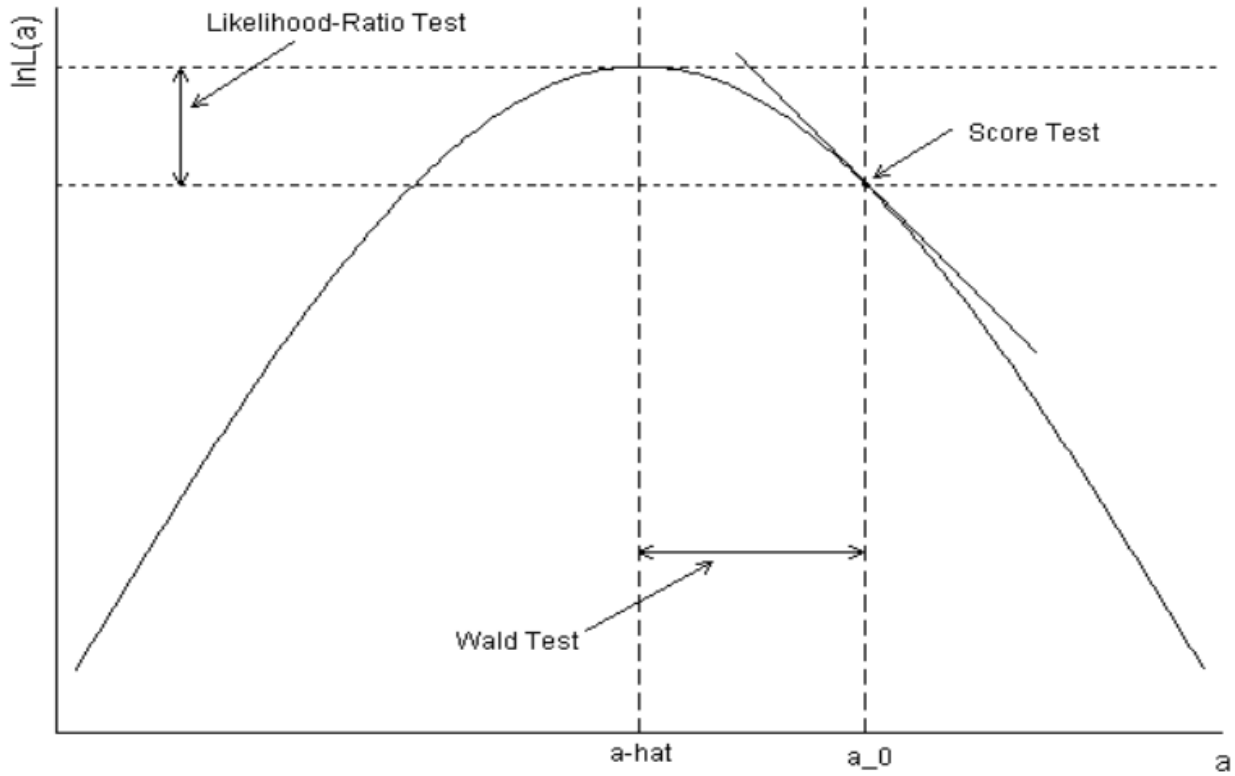
$$s(\hat{\beta}) \overset{a}{\sim} N(0, \mathbf{F}(\hat{\beta}))$$

Definition. Score statistic (u)

$$u = s^T(\hat{\beta})\mathbf{F}^{-1}(\hat{\beta})s(\hat{\beta})$$

The score statistic is based on behaviour of the likelihood function close to the value stated by H_0.

Comparison of statistics $(w \geq \lambda \geq u)$



The likelihood ratio statistic uses more information than the Wald and score statistic. For this reason LR statistic is suggested as most reliable of the three.

Source: J. Fox (1997). Applied regression analysis.

Goodness of Fit. Deviance

- The fitted values produced by the model are most likely not to match the values of the observations perfectly.
  The size of the discrepancy between the model and the data is a measure of the inadequacy of the model (*Goodness of Fit*).
- The saturated model fits the data exactly by assuming as many parameters as observations.
  Deviance compares the current model to the saturated model.

Deviance is a measure for discrepancy that is based on the likelihood ratio statistic for comparing nested models.

To assess goodness of fit for a model, we compare the model by the saturated model.

Deviance

- M_1 – our current model (estimates are denoted by hat)
- $\hat{\mu}_i$ – estimate of conditional mean,
  $\hat{\boldsymbol{\mu}}$ – corresponding vector
- M – saturated model (estimates are denoted by tilde)
- $\tilde{\mu}_i$ – estimate of conditional mean, equals to observation y_i,
  $\boldsymbol{y}$ – corresponding vector

Deviance $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$

Deviance is -2φ times the difference in log-likelihood between the current model and a saturated model: $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = -2\varphi(l(\hat{\boldsymbol{\mu}}; \boldsymbol{y}) - l(\boldsymbol{y}; \boldsymbol{y}))$

Note that deviance is related likelihood ratio statistic λ where we also compare the current model by the saturated model: $D = -2\varphi \ln \lambda$
The ratio $\frac{D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})}{\varphi}$ is called **scaled deviance**

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} n_i [y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]$$

Goodness of Fit (GOF)

The idea of GOF tests is to test the following hypotheses:

$$\begin{cases} H_0: & \text{Model } M_1 \text{ fits as good as saturated model } M, \text{ is adequate} \\ H_1: & \text{Model } M_1 \text{ is not adequate} \end{cases}$$

Statistics

- Deviance: $D = -2\varphi \ln \lambda$
- Pearson χ^2–statistic: $\chi^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\nu(\hat{\mu}_i)}$,
  where $\nu(\cdot)$ is the variance function corresponding to model M_1

Both statistics are asymptotically $\varphi\chi^2_{n-p}$ distributed
(n – sample size, p – number of parameters in M_1):

$$D \overset{a}{\sim} \varphi\chi^2_{n-p}, \quad \chi^2 \overset{a}{\sim} \varphi\chi^2_{n-p}$$

Difference of deviances

Deviance is asymptotically $\varphi\chi^2_{n-p}$ distributed, $D \overset{a}{\sim} \varphi\chi^2_{n-p}$
(n – sample size, p – number of parameters in model)

If the distribution of deviance is not χ^2, it is recommended to use difference of deviances

Let us have M_q – lower model (q parameters), M_r – upper model (r parameters), $r > q$

If φ is known then

$$\frac{D_{M_q} - D_{M_r}}{\varphi} \overset{a}{\sim} \chi^2_{r-q}$$

If φ is not known, normal approximation is used

$$\frac{D_{M_q} - D_{M_r}}{\hat{\varphi}(r-q)} \overset{a}{\sim} F_{r-q,n-r}$$

GOF statistics

An important question

How should we compare two candidate models (especially when one is not a special case of the other)?

(One possible) answer

We can apply certain penalty function to the model that has more parameters

Let us recall that likelihood ratio test can be interpreted as follows:
we should choose more complicated model (with $r - q$ extra parameters) if

$$\ln L_r > \ln L_q + \frac{c_{r-q,\alpha}}{2},$$

where

- L_r is the max. value of likelihood function for the more complicated model
- L_q is the max. value of likelihood function for the simpler model
- $c_{r-q,\alpha}$ is the 1-α-quantile of the distribution $\chi^2(r-q)$

GOF statistics. Akaike information criterion (AIC)

Goodness of fit of a model can be measured by different information criteria (which are also based on log-likelihood)

The most well-known of them is Akaike information criterion AIC (Akaike, 1974):

$$AIC = -2 \ln L + 2p,$$

where

- L is the maximized value of the likelihood function
- p is the number of parameters in a model under consideration

So, AIC measures the goodness of fit as certain tradeoff between two components:

- goodness of fit (measured by the log-likelihood),
- model complexity (measured by the number of parameters p)

One should also note that when the number of parameters is equal then the decision is essentially based on the likelihood function value.

GOF statistics. Bayesian IC and corrected AIC

Information criteria can also take the sample size into account.

Bayes information criterion BIC (Schwarz criterion, SC (Schwarz,1978))

$$BIC = -2 \ln L + p \ln n,$$

One can interpret this as:

- when there are $r - q$ extra parameters and the sample size is n then we should reduce the log-likelihood by $\frac{r-q}{2} \ln n$
- each additional parameter is deemed worthy when the log-likelihood is increased by $0.5 \ln n$

When this is not the case then a simpler model should be preferred

The corrected AIC takes the sample size into account as well:

$$AICc = AIC + \frac{2p(p+1)}{n-p-1} = -2 \ln L + 2p + \frac{2p(p+1)}{n-p-1}$$

Remarks about AIC and BIC

- Does not require the assumption that one of the candidate models is the 'true' or 'correct' model
- Can be used to compare nested as well as non-nested models
- Can be considered as a generalization of the idea of likelihood ratio
- Can be used to compare models based on different families of probability distributions (**same data, same response**, *NB! missing values!*)
- AIC penalizes the extra parameters less than BIC

BIC in comparing non-nested models

Let M_1 and M_2 be two non-nested models with calculated BIC_1 and BIC_2, respectively.

Comparison of non-nested models is based on the difference between the BICs for the two models:

- If $BIC_1 - BIC_2 < 0$, then model M_1 is preferred
- If $BIC_1 - BIC_2 > 0$, then model M_2 is preferred

The scale given by Raftery for determining the relative preference is:

Abs. difference	Degree of preference
0–2	weak
2–8	positive
6–10	strong
> 10	very strong

Source: J.W. Hardin, J.M. Hilbe (2007), Raftery (1996)

Generalized R^2

Recall that in case of linear models, one of the most useful quantities to measure the suitability of the model was the coefficient of determination (R^2).

In case of a GLM, there are several attempts to generalize the classical R^2, but none of them has such nice and clear interpretation.

One of the most known generalizations of R^2 is the following:

Pseudo-R^2 statistic or likelihood ratio index (McFadden, 1974)

$$R^2_{McF} = 1 - \frac{l(M_\beta)}{l(M_\alpha)}$$

$l(M_\beta)$ - log-likelihood for current model, $l(M_\alpha)$ - log-likelihood for null model

It is usually scaled so that $R^2_{McF} \in [0; 1]$

Others: Cox-Snell-R^2, Nagelkerke-R^2, deviance-R^2

Summary of GOF

- The asymptotic distribution of deviance is usually, but not always χ^2-distributed
- Deviance is additive, the difference of deviances is asymptotically χ^2–distributed
- Pearson χ^2–statistic is preferred and more often used than deviance (but the difference of Pearson statistics is not χ^2-distributed)
- BIC may be preferred as compared to AIC
- Generalized R^2 has no reasonable interpretation

Model checking/diagnostics

The fit of the model to the data can be explored by diagnostics tools

The purpose of the model diagnostics is to examine whether the model provides a reasonable approximation to the data

If there are indications of systematic deviations between data and model, the model should be modified

The diagnostics tools are the following:

- Residuals plots, to detect outliers, problems with distribution, etc
- Identifying influential observations, which have unusual impact to results
- Detecting overdispersion and modifying the model

Source: Olsson (2002)

Generalized Hat matrix, 1

Let us first recall the classical linear model $\boldsymbol{y} = \mathbf{X}\beta + \varepsilon$, where

- $\boldsymbol{y} = (y_1, \ldots, y_n)$ – response
- $\mathbf{X}$ – design matrix
- $\beta = (\beta_0, \ldots, \beta_k)$ – unknown parameters
- $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ – random errors

Then the parameters β are found using the normal equation $(\mathbf{X}^T\mathbf{X})\beta = \mathbf{X}^T\boldsymbol{y}$, which yields $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$ and $\hat{\boldsymbol{y}} = \mathbf{X}\hat{\beta} = \mathbf{H}\boldsymbol{y}$, where $\mathbf{H}$ is the hat matrix, $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, i.e. the matrix that projects the data onto the fitted values.

The diagonal elements h_{ii} of $\mathbf{H}$ are called **leverages**. A large value of h_{ii} indicates that the fit may be sensitive to the response in case i.

Generalized Hat matrix, 2

One important difference in GLM model is that the leverages now depend on the response through the weights $\mathbf{W} = \mathbf{W}(\beta)$ and through the parameters itself, so $\hat{\mathbf{H}} = \mathbf{H}(\hat{\beta})$.

The generalized hat matrix $\mathbf{H}$ is defined in terms of the design matrix $\mathbf{X}$ and a diagonal weight matrix $\mathbf{W}$

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{1/2},$$

where $\mathbf{W}^{1/2}$ is the diagonal matrix with diagonal elements $\sqrt{w_{ii}}$

Source: Faraway (2006)

Generalized residuals (1)

Residuals r_i are estimates of random error ε_i and are usually defined as observed minus fitted values $r_i = y_i - \hat{y}_i$ (raw residuals).

We would like the residuals for GLMs to be defined so that they can be used in a similar way as in the classical linear model.

Residuals measure the agreement between single observations and their fitted values and help to identify poorly fitting observations that may have a strong impact on the overall fit of the model.

Pearson residuals

$$r_{iP} = \frac{r_i}{\sqrt{\nu_i(\hat{\mu}_i)}}$$

Pearson residuals are raw residuals scaled by the estimated standard error.
Pearson residuals are related to the Pearson χ^2-statistic: $\chi^2 = \sum r_{iP}^2$

Generalized residuals (2)

Anscombe residuals

$$r_{iA} = \frac{A(y_i) - A(\hat{\mu}_i)}{A'(\hat{\mu}_i)\sqrt{\nu(A(\hat{\mu}_i))}}$$

Anscombe proposed to define a residual using a function $A(y)$ in place of y, where $A(\cdot)$ is chosen to make the distribution of $A(y)$ as normal as possible

Barndorff-Nielsen (1978): $A(\cdot) = \int \frac{d\mu}{\nu^{1/3}(\mu)}$

Deviance residuals

$$r_{iD} = sign(r_i)\sqrt{d_i}$$

d_i – ith component of model deviance. Sign of the raw residual indicates the direction of deviance

Sum of squares of deviance residuals is the model deviance, $D = \sum d_i^2$

Generalized standardized residuals

Standardized residuals

Standardized residuals are simply the residuals divided by their standard deviation

Generalized standardized deviance residuals

$$r_{iDS} = \frac{r_{iD}}{\sqrt{\hat{\varphi} \cdot (1 - h_{ii})}}$$

$\hat{\varphi}$ – estimated scale parameter
h_{ii} – ith diagonal element of generalized hat matrix

Generalized standardized Pearson residuals

$$r_{iPS} = \frac{r_{iP}}{\sqrt{\hat{\varphi} \cdot (1 - h_{ii})}}$$

Generalized Studentized residuals

Generalized studentized residuals

Studentized residuals are obtained by dividing the residual by its standard deviation calculated without ith observation

In a linear model, Studentized residuals can be calculated without actually applying the model n more times, in a GLM this is not possible.
Thus, approximations are used, e.g. likelihood residuals:

Likelihood residuals

$$r_{iL} = sign(r_i)\sqrt{(1 - h_{ii})r_{iDS}^2 + h_{ii}r_{iPS}^2}$$

where h_{ii} is ith diagonal element of generalized hat matrix

Likelihood residuals are a combination of Pearson and deviance residuals, and can also be useful when using software that does not produce Anscombe residuals.

Remarks

- McCullagh and Nelder (1989) showed that Anscombe and deviance residuals are numerically similar, even though they are mathematically quite different. This means that the deviance residuals are also approximately normally distributed, if the response distribution has been correctly specified
- Typically residuals are visualized on a graph. In an index plot, the residuals are plotted against the observation number, or index. It shows which observations have large values and may be considered outliers
- For finding systematic deviations from the model it is often more informative to plot the residuals against the fitted linear predictor
- An alternative graph compares the standardized residuals to the corresponding quantiles of a normal distribution. If the model is correct and residuals can be expected to be approximately normally distributed (depending on sample size), the plot should show approximately a straight line as long as outliers are absent

Model diagnostics (1)

Regression diagnostics aim to identify observations of outlier, leverage and influence. These observations may have significant impact on model fitting and should be examined for whether they should be included.

Sometimes, these observations may be the result of typing error and should be corrected.

Outlier

- An observation with large residual

Leverage

- An observation with an extreme value on an argument variable (x_{ij})
- Leverage is a measure of how far an argument variable deviates from its mean

Influence

- An observation is influential if it influences any part of a regression analysis, the estimated parameters, or the hypothesis test results

Model diagnostics (2)

Summary statistics for outlier, leverage and influence are **standardized/studentized residuals**, **hat values** and **Cook's distance**.

They can be easily visualized with graphs and formally tested.

- Outlier. Generalized standardized/studentized residuals, residuals > 3 are too large.
- Leverage. Generalized hat matrix, leverages h_{ii} are given by the diagonal of $\mathbf{H}$ and represent the potential leverage of the point (but not always)
- Influence. Cook's distance (influence to parameter estimates)

$$C_i = (\hat{\beta}^*_{(i)} - \hat{\beta})^T \mathbf{F}(\hat{\beta})(\hat{\beta}^*_{(i)} - \hat{\beta}),$$

where $\hat{\beta}^*_{(i)}$ is the estimate at first iteration step (without ith observation)

Model diagnostics (3)

Single Case Deletion Diagnostics

- Delta χ^2 statistic (observation's influence to Pearson χ^2 statistic)

$$\Delta\chi_i^2 = \chi^2 - \chi_{(i)}^2,$$

  where $\chi_{(i)}^2$ is the estimate without ith observation

- Delta deviance statistic (observation's influence to deviance)

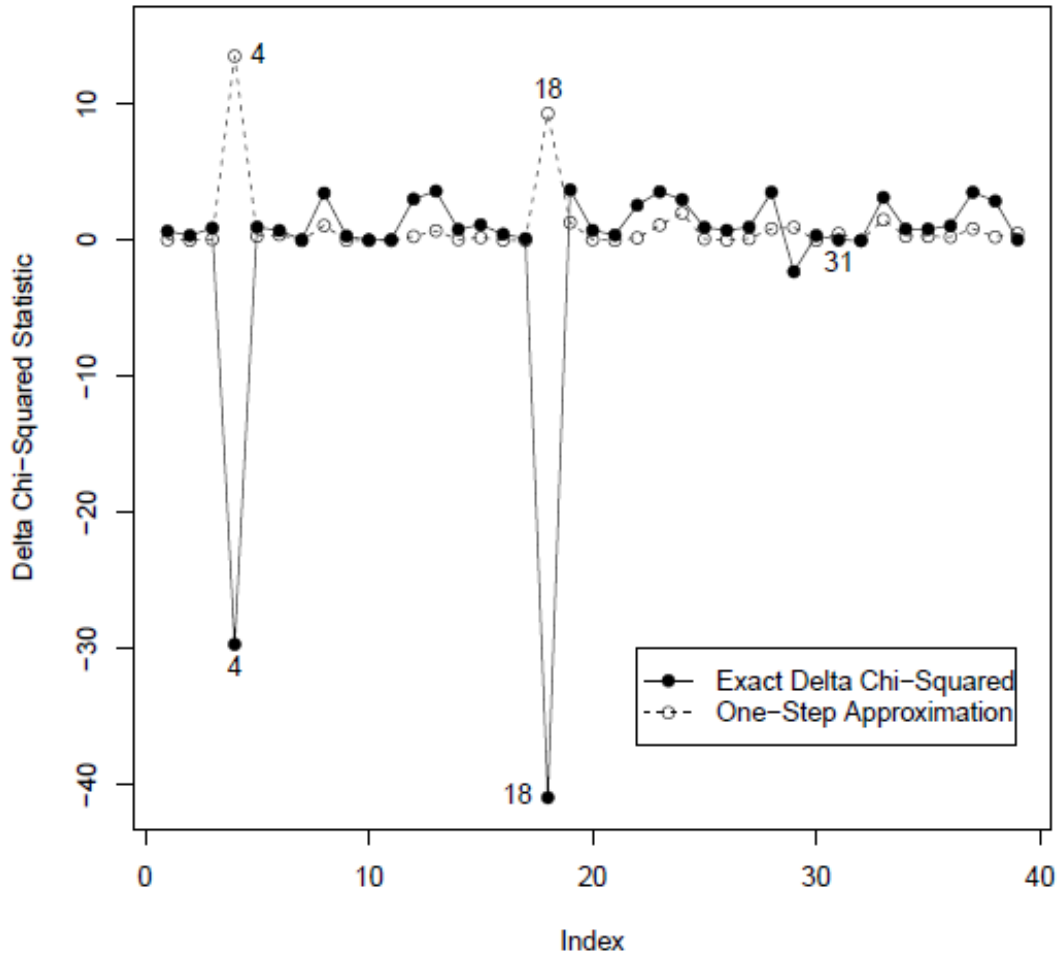
$$\Delta D_i = D - D_{(i)},$$

  where D – deviance, $D_{(i)}$ – deviance without ith observation

NB! Note that influential observations need not be outliers in the sense of having large residuals, nor need they be leverages. Similarly, the outliers need not be always highly influential.

Example. Delta χ^2 statistic

Vaso-constriction data



Source: A. Powne (2011). Diagnostic Measures for Generalized Linear Models (2011). Department of Mathematical Sciences, University of Durham

Summary. Steps of statistical modelling

1. Specifying the model:
   - the probability distribution of the response variable
   - equation linking the response and explanatory variables
2. Estimating parameters used in the model
3. Making inference about the parameters
   - calculating confidence intervals
   - testing statistical hypothesis
4. Checking how well the model fits the real data (general fit)
   - Goodness of Fit statistics
5. Model diagnostics (fit in each point)
   - generalized residuals
   - leverages, influential points
6. Interpreting the final model