Part II: Generalized Linear Models

Chapter II.2

Theory of Generalized Linear Models

Residuals - Implementation in R

Residuals

Residuals are critical tools for diagnosing lack of model fit and identifying possible underlying patterns. Beyond the raw residuals e_i, $i = 1, \ldots, n$ (see Definition II.2.32), other types of residuals used in GLM analysis are given below.

II.2.45 Definition (Pearson residuals)

For the set–up of Definition II.2.32, the **Pearsonian residuals** are defined by

$$e_i^P = \frac{e_i}{\sqrt{v(\hat{\mu}_i)}} = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)}}, \quad i = 1, \ldots, n,$$

where $v(\mu)$ is the variance function (see remark II.2.18) of the EDF response distribution.

- For a Poisson GLM, $v(\hat{\mu}_i) = \hat{\mu}_i$ and the $(e_i^P)^2$s sum to the Pearson's X^2 statistic:

$$X^2 = \sum_{i=1}^{n} (e_i^P)^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}.$$

II.2.46 Definition (deviance residuals)

For the set–up of Definition II.2.32, the **deviance residuals**, e_i^d, $i = 1, \ldots, n$, decompose the deviance to the individual contributions of each observation i. Thus, they satisfy

$$D(y; \hat{\mu}) = \sum_{i}^{n} \left(e_i^d\right)^2 .$$

For distributions in EDF with $a(\phi; i) = \phi / w_i$ (see Definition II.2.35), they are equal to

$$e_i^d = \text{sign}(y_i - \hat{\mu}_i) \cdot \left[2\omega_i \left| y_i(\tilde{\theta}_i - \hat{\theta}_i) - [b(\tilde{\theta}_i) - b(\hat{\theta}_i)] \right| \right]^{1/2}, \quad i = 1, \ldots, n.$$

When the assumed model holds, Pearsonian and deviance residuals are *less variable* than standard normal, since they compare the observed y_is to the estimated $\hat{\mu}_i$s (instead to the true μ_is).

☞ Pearsonian residuals are the estimated standardized residuals and are not approximately standard normal distributed. For this, the *standardized residuals* are more appropriate.

II.2.47 Definition (standardized residuals)

For the set–up of Definition II.2.32, the **standardized residuals** are defined by

$$e_i^s = \frac{e_i^P}{\sqrt{1-\hat{h}_{ii}}} = \frac{e_i}{\sqrt{v(\hat{\mu}_i)(1-\hat{h}_{ii})}},\ i = 1,\ldots,n,$$

where $\hat{h}_{ii}$ is the estimate of the diagonal element h_{ii} of the *generalized hat matrix*

$$\mathbf{H}_{n\times n} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2}.$$

II.2.48 Remark

The distribution of the standardized residuals under the assumed model can be approximately standard normal (see Remark II.2.38) [a].

[a] see Pierce and Schafer (1986). Residuals in generalized linear models, *Journal of the American Statistical Association*, 81, 977–986.

II.2.49 Remark (detection of model lack-of-fit)

Any type of residuals can be plotted against the component fitted values in $\hat{\mu}$. They can also be plotted against each explanatory variable, to detect possible non-linear effects. Residual diagnostics should accompany any model fitting procedure.

Sofware for GLMs

All general-purpose statistical packages have procedures for GLM analysis, most of them under the name GLM (like SAS, S-Plus, SPSS, STATA and SYSTAT).

Historical Note

GLIM (Generalized Linear Interactive Modelling) was the first package with the ability of fitting a variety of GLMs in a unified manner. It was developed by the GLIM working party of the Royal Statistical Society in 1974. GLIM4, the latest release (1993), had many link functions as standard options and was convenient for GLM fit and model selection.

☞ The corresponding R function is `glm()`.

GLMs in R

`glm()` is called as

```
glm(formula, family = gaussian, data, …)
```

where the argument `formula` defines the model to be fitted, `family` determines the error distribution and link function of the model and `data` specifies the data frame on which the model will be applied.

The `formula` argument

This argument specifies the model formula as, e.g. `y ~ x1 + x2`, where `x1`, `x2` are the names of numeric vectors (continuous variables) or factors (categorical variables).

- ☞ All specified variables in this argument must be in the workspace or in the data frame given in the `data` argument.
- ☞ Other symbols that can be used in the `formula` argument are
  - `a:b`, for an interaction between a and b
  - `a*b`, which expands to a + b + a:b
    (e.g. `y ~ a*b` is equivalent to `y ~ a+b+a:b`)

The `family` argument

This argument specifies the link function and the variance function.

The EDF functions (with canonical link) available in R are

- `binomial(link = "logit")`
- `gaussian(link = "identity")`
- `Gamma(link = "inverse")`
- `inverse.gaussian(link = "1/mu^2")`
- `poisson(link = "log")`

Controlling the Output

Calling `glm()` does not save the analysis results and provides a summary output on the screen.
The analysis results are saved under e.g. `m1`, by
`m1 <- glm(...)`
In this case, there is no output provided automatically but can be controlled by the user.

Among the functions that are available for displaying (or saving) components of a `glm` (or `lm`) object, say `m1`, are

- `residuals(m1)`
- `fitted(m1)`
- `predict(m1)`
- `coef(m1)`
- `deviance(m1)`
- `formula(m1)`
- `summary(m1)`

Confidence Intervals for GLM Parameters in R

The functions

- `confint(m1)` and
- `confint.default(m1)`

compute confidence intervals (CIs) for one or more parameters of a fitted GLM saved under the `glm` object `m1`.

- ☞ The classical asymptotic CIs based on the normal approximation are derived by the `confint.default()` function.
- ☞ The `confint()` function provides the *profile likelihood CIs*, which are preferable in case of multiparameter models.

II.2.50 The profile log-likelihood approach

Consider the set–up of a GLM as in Defintion II.2.32 and let β_k be the parameter of interest. Rearrange and partition the p-dimensional parameter vector $\boldsymbol{\beta}$ as $\tilde{\boldsymbol{\beta}}' = (\beta_k, \boldsymbol{\beta}'_{[-k]})$, where $\boldsymbol{\beta}_{[-k]}$ is the $(p-1)$–dimensional vector with the remaining parameters (*nuisance parameters*). In this case the log–likelihood is

$$\ell(\beta_k,\ \boldsymbol{\beta}_{[-k]}) = \sum_{i=1}^{n} \ell_i(\beta_k, \boldsymbol{\beta}_{[-k]}) \ .$$

Consider further the null hypothesis $H_0:\ \beta_k = \beta_{k0}$ (fixed) for the parameter of interest and estimate the nuisance parameters by maximizing the likelihood under H_0, i.e. maximizing

$$\ell_{\beta_{k0}}(\boldsymbol{\beta}_{[-k]}) = \ell(\beta_{k0},\ \boldsymbol{\beta}_{[-k]}) \ .$$

The derived vector of nuisance parameters estimates is $\hat{\boldsymbol{\beta}}_{[-k]}(\beta_{k0})$ and varies as a function of β_{k0}. The likelihood

$$\ell_p(\beta_k) = \ell\left(\beta_k,\ \hat{\boldsymbol{\beta}}_{[-k]}\right) \ ,$$

is the **profile log-likelihood** of β_k (profiles out the nuisance parameters).

Profile Likelihood CIs

II.2.51 Definition

Consider the set–up of a GLM as in Defintion II.2.32 and let β_k, $k = 1, \ldots, p$, be the component of interest from the p-dimensional parameter vector $\boldsymbol{\beta}$. If $\hat{\beta}_k$ and $\hat{\boldsymbol{\beta}}_{[-k]}$ denote the MLEs of β_k and $\boldsymbol{\beta}_{[-k]}$, respectively, then the $(1-\alpha)100\%$ **profile likelihood CI** for β_k is the set of all β_{k0} (see I.9.50) for which

$$-2\left[\ell\left(\beta_{k0},\ \hat{\boldsymbol{\beta}}_{[-k]}(\beta_{k0})\right) - \ell\left(\hat{\beta}_k,\ \hat{\boldsymbol{\beta}}_{[-k]}(\beta_{k0})\right)\right] < \succ_1^2(1-\alpha)\ ,$$

where $\succ_1^2(1-\alpha)$ is the $(1-\alpha)$–th quantile of the $\succ_1^2$ distribution.