# 4 Modeling

## 4.1 Process of modeling

1. Study design
2. Data collection
3. Selection of model class
4. Estimation
5. Model checking
6. Conclusions
7. Reporting

## 4.2 Residuals

Residuals can be used to check the model fit. For GLMs different kind of residuals can be defined:

**Raw residuals (response residuals)**

$$r_i = y_i - \hat{\mu}_i \tag{46}$$

**Pearson residuals**

$$r_{P,i} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)/a_i}} \tag{47}$$

whose squared sum

$$\sum_{i=1}^{n} r_{P,i}^2 = X^2 \tag{48}$$

is the Pearson chi-squared goodness-of-fit statistic.

**Deviance residuals**

$$r_{D,i} = \operatorname{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}, \tag{49}$$

where

$$d_i = 2a_i\left(y_i\left(\theta_i(y_i) - \theta_i(\hat{\mu}_i)\right) - b\left(\theta_i(y_i)\right) + b\left(\theta_i(\hat{\mu}_i)\right)\right). \tag{50}$$

The deviance is the squared sum of the deviance residuals

$$\sum_{i=1}^{n} r_{D,i}^2 = D(\mathbf{y}; \hat{\boldsymbol{\mu}}) \tag{51}$$

**Anscombe residuals** where $y_i$'s and $\mu_i$'s are transformed so that the residuals become approximately normally distributed.

In R, method `residuals` for class `glm` can compute raw, Pearson and deviance residuals.

Influential observations can be identified, for instance, by calculating differences

$$\Delta_i \hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(i)} \tag{52}$$

where $\hat{\boldsymbol{\beta}}^{(i)}$ is estimated from data without observation $i$.

## 4.3 Nonlinear terms

GLMs allow inclusion of known transformations of the covariates as far as the linear predictor $\eta_i$ can be presented as a sum of transformed covariates. For instance, the design matrix may be defined as

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{11}^2 & x_{11}^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n1}^2 & x_{n1}^3 \end{pmatrix}. \tag{53}$$

to fit a GLM with a third order polynomial for covariate $x_1$.

## 4.4 Interactions

Interaction terms are nonlinear transformations of two or more covariates. The type of interaction can synergistic (the joint effect is stronger than the additive effect) or antagonist (the joint effect is weaker than the additive effect). It is usually a bad idea to include interaction terms without the corresponding main effects.

## 4.5 Hypothesis testing

### 4.5.1 Single test

The test of hypothesis

$$\begin{aligned} H_0 : &\quad \beta_j = 0 \\ H_1 : &\quad \beta_j \neq 0 \end{aligned}$$

for a certain regression coefficient $\beta_j$ of a GLM can be based on the likelihood ratio test

$$\Lambda = \frac{\sup\{L(\beta_1, \ldots, \beta_p; \mathbf{y}) : \beta_j = 0\}}{\sup\{L(\beta_1, \ldots, \beta_p; \mathbf{y})\}} \tag{54}$$

Using log-likelihoods the test statistic can be written

$$-2 \log \Lambda = 2(l(\hat{\boldsymbol{\beta}}; \mathbf{y}) - l(\hat{\boldsymbol{\beta}}_0; \mathbf{y})) \tag{55}$$

where $\hat{\boldsymbol{\beta}} = \hat{\beta}_1, \ldots, \hat{\beta}_p$ and $\hat{\boldsymbol{\beta}}_0 = \hat{\beta}_1, \ldots, \beta_j = 0, \ldots, \hat{\beta}_p$ are the maximum likelihood estimates under the two models. The statistic $-2 \log \Lambda$ follows asymptotically $\chi^2_1$ distribution The test can written also in terms of deviance

$$-2 \log \Lambda = \frac{D(\mathbf{y}; \hat{\boldsymbol{\beta}}_0) - D(\mathbf{y}; \hat{\boldsymbol{\beta}})}{\phi}. \tag{56}$$

The likelihood ratio test for more than one parameter is similar but the test statistic follows asymptotically $\chi^2$ distribution with degrees of freedom equal to the difference in dimensionality of $\beta$ and $\beta_0$. If the dispersion parameter is not known, the test statistics

$$\frac{D(\mathbf{y}; \hat{\boldsymbol{\beta}}_0) - D(\mathbf{y}; \hat{\boldsymbol{\beta}})}{\hat{\phi}(p - q)} \tag{57}$$

where $q$ is the dimensionality of $\beta$ follows asymptotically F-distribution $F_{p-q,n-p}$.

### 4.5.2 Multiple tests

Let $p_1, p_2, \ldots, p_m$ be the nominal p-values from $m$ tests. Family-wise error rate (FWER) is the probability that at least one true null hypothesis is falsely rejected. Several approaches for controlling FWER exist: a simple approach is the Bonferroni correction where the nominal p-values are compared to the $\alpha/m$ where $\alpha$ is the significance level. If the tests are dependent, the Bonferroni correction is too conservative and the actual significance level is smaller than $\alpha$.

False discovery rate (FDR) is the expected proportion of incorrectly rejected null hypothesis in a set of hypotheses. The FDR analysis has been used e.g. in genome wide association (GWA) studies where the number of tests can be one million.

## 4.6 Model selection

**Multiple models.** Competing models are fitted and the estimated model parameters are reported for each model. The properties of the models are discussed. This is actually not a formal model selection method but a commonly used practical approach to the problem. The approach is feasible only if the number of the competing models is small.

**Likelihood ratio test** can be used to compare nested models.

**Stepwise regression.** In forward selection, the procedure starts with a null model and covariates are added one by one. The procedure continues until the newly added covariate does not improve the model. The improvement of the model defined e.g. by the p-value of the likelihood ratio test. In backward elimination, the procedure starts with the full model and covariates are removed one by one. The procedure continues until the removal of a covariate makes the model worse. The lasso (least absolute shrinkage and selection operator, `http://www-stat.stanford.edu/~tibs/lasso.html`) can be understood as a modernized version of stepwise regression (not based on likelihood). Stepwise methods cannot guarantee that the best model will be selected. Automated methods should not replace careful thinking.

**Information criteria:** Akaike information criterion (AIC), Bayesian information criterion (BIC), Bayes factor, crossvalidation, etc. AIC and BIC are straightforward to compute

$$\begin{aligned} AIC &= -2l(\boldsymbol{\beta}; \mathbf{y}) + 2p, \\ BIC &= -2l(\boldsymbol{\beta}; \mathbf{y}) + p\log(n), \end{aligned}$$

where $p$ is the number of parameters in the model. The model with smallest value of AIC (or BIC if that is used) will be considered the best. Both AIC and BIC penalize models for a higher number of parameters. In BIC, the penalty depends also on the number of observations.

## 4.7 Experimental and observational studies

Experimental data origin from data generating mechanism where the experimenter selects the values of some variables. In observational data, all values are recorded as observed. The same GLMs can be used for both types of data. The analysis follows the same lines but the interpretation of the results may differ. In general, only experimental data allows causal inference. With observational data, the possibility of confounders and alternative causal explanations must be accounted.

## 4.8 Missing data

Usually there are missing observations in real world data. A statistician has the following options:

**Ignore** the missing observations and analyze only the complete cases. This is applicable if only few observations are missing.

**Impute** the missing values. Multiple imputation is preferred over single imputation. The challenges lie in the definition of the imputation model.

**Model** the data. The likelihood becomes an integral over the missing values. The results are sensitive to model misspecification and estimation may require a lot of computational resources.

## 4.9 Few words on independence

Term “independence” may have different meanings depending on the context. In statistics, the term refers to independence of events or to independence of random variables. Events $A$ and $B$ are independent if

$$P(A \text{ and } B) = P(A)P(B). \tag{58}$$

or equivalently, using conditional probabilities

$$P(A \mid B) = P(A) \tag{59}$$

or

$$P(B \mid A) = P(B). \tag{60}$$

Random variables $X$ and $Y$ are independent (marked $X \perp\!\!\!\perp Y$) if

$$F_{X,Y}(X, Y) = F_X(X)F_Y(Y). \tag{61}$$

The term “linear independence of random variables” is sometimes used to indicate that the random variables are uncorrelated but this usage is not recommended. In general, zero correlation does not imply independence.

In linear algebra, linear independence of a family of vectors means that none of the vectors can be presented as a linear combination of the other vectors. A matrix whose columns are linearly independent has full rank.

The concept of conditional independence is important when causality is considered. Random variables $X$ and $Y$ are independent on the condition of $Z$ (notation $X \underset{Z}{\perp\!\!\!\perp} Y$ or $X \perp\!\!\!\perp Y \mid Z$ may be used) when

$$F_{X,Y \mid Z}(X, Y \mid Z) = F_{X \mid Z}(X \mid Z)F_{Y \mid Z}(Y \mid Z) \tag{62}$$

or equivalently

$$F_{X, \mid Y,Z}(X \mid Y, Z) = F_{X \mid Z}(X \mid Z). \tag{63}$$

# 5 Binary response

## 5.1 Representations of binary response data

In binary response data, the response $Y_i$ has two possible values, for instance, 0 and 1. Binary response data can be presented in different formats:

Data matrix

| $x$ | $y$ |
|---|---|
| 250 | 0 |
| 250 | 1 |
| 350 | 1 |
| 300 | 0 |
| 250 | 0 |
| 300 | 1 |
| ⋮ | ⋮ |

Weighted data matrix

| $x$ | $y$ | frequency |
|---|---|---|
| 250 | 0 | 23 |
| 250 | 1 | 12 |
| 300 | 1 | 21 |
| 300 | 0 | 19 |
| 350 | 0 | 7 |
| 350 | 1 | 13 |

Frequency table (crosstabulation)

| | $Y = 0$ | $Y = 1$ |
|---|---|---|
| $x = 250$ | 23 | 12 |
| $x = 300$ | 21 | 19 |
| $x = 350$ | 7 | 13 |

The response $Y_i$ can be either a Bernoulli random variable (binary response) or a sum of Bernoulli random variable (binomial response). In the latter case, the observational units with the identical covariate values belong to the same covariate class. For the $i$th covariate class $m_i$ binary responses are recorded and the number of responses 1 is denoted by $K_i$. The binomial response is defined

$$Y_i = \frac{K_i}{m_i}. \tag{64}$$

## 5.2 Link functions for binary data

If the possible values of $Y_i$ are 0 and 1, it holds

$$P(Y_i = 1) = \mathrm{E}(Y_i) = g^{-1}(\eta_i), \tag{65}$$

where the possible values of the inverse link function $g^{-1}()$ belong to the interval $(0, 1)$. Any cumulative distribution function defines the inverse of a link function. The commonly used link functions are the logit link

$$g(\mu_i) = \mathrm{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right), \tag{66}$$

the probit link

$$g(\mu_i) = \mathrm{probit}(\mu_i) = \Phi^{-1}(\mu_i), \tag{67}$$

where $\Phi^{-1}$ is the inverse of cumulative distribution function (cdf) of the standard normal distribution and the complementary log-log link

$$g(\mu_i) = \mathrm{cloglog}(\mu_i) = \log(-\log(1-\mu_i)). \tag{68}$$

## 5.3 Odds and log-odds

It is often interesting to compare the estimated responses for different values of covariates. Denote

$$p_A = P(Y_i = 1 \mid \eta_A) \tag{69}$$

$$p_B = P(Y_i = 1 \mid \eta_B) \tag{70}$$

where $\eta_A$ and $\eta_B$ are the linear predictors for certain values of covariates. Now the odds ratio is defined as

$$\frac{p_A/(1-p_A)}{p_B/(1-p_B} \tag{71}$$

and the logarithm of the odds ratio becomes

$$\log\left(\frac{p_A/(1-p_A)}{p_B/(1-p_B}\right) = \log\left(\frac{p_A}{1-p_A}\right) - \log\left(\frac{p_B}{1-p_B}\right) = \mathrm{logit}(p_A) - \mathrm{logit}(p_B), \tag{72}$$

which in the case of logit link simplifies

$$\text{logit}(p_A) - \text{logit}(p_B) = \eta_A - \eta_B. \tag{73}$$

## 5.4 Latent variables

Consider an example where the effectiveness of an insecticide to mosquitos is studied. Mosquitos have different resistance to the insecticide. A mosquito dies ($Y = 1$) if the amount of insecticide $x$ is higher than a threshold value $T$, which varies in the population. Because $T$ cannot be directly measured, it is called latent variable. If $T$ follows the normal distribution with mean $-\alpha/\beta$ and variance $1/\beta^2$ we obtain for a mosquito randomly chosen from the population

$$P(Y = 1) = P(T \le x) = \Phi\left(\frac{x - (-\alpha/\beta)}{1/\beta}\right) = \Phi(\alpha + \beta x). \tag{74}$$

In other words, the use of normal distributed latent variable led to the probit model. If $T$ follows logistic distribution, we will end up with the logistic model. If $T$ follows Gumbel distribution, we will end up with the GLM with cloglog link.

## 5.5 Overdispersion

Overdispersion means that the variance in the data is greater than the variance assumed in the model. The sum of independent Bernoulli random variables

$$K = Y_1 + Y_2 + \ldots + Y_m \tag{75}$$

follows binomial distribution $K \sim \text{Bin}(m, \mu)$ where $\text{E}(Y_i) = \mu$. It follows that $\text{Var}(K) = m\mu(1 - \mu)$. In real world datasets, however, the assumption of independence is often unrealistic and $\text{Var}(K) > m\mu(1 - \mu)$. This is called overdispersion.

## 5.6 Non-existence of maximum likelihood estimates

Maximum likelihood estimates do not exist if the data can be perfectly separated on the basis of covariate values, for example, response 1 is always obtained if $x > 100$ and response 0 is always obtained if $x < 100$.

## 5.7 Example: Switching measurements

The Josephson junction (JJ) circuits are important non-linear components of superconducting electronics. The strong dependence of the physical parameters of JJ circuits as function of changes in environmental variables, for instance, temperature, electric noise, and magnetic field makes the JJ circuits to have several applications as ultra-sensitive sensors. Moreover, certain JJ circuits are promising candidates for realization of quantum computation. An experiment called switching measurement is a common way to probe the properties of a JJ circuit sample. In the experiment, sequences of current pulses are applied to the sample, while the voltage over the structure is monitored. Switching measurements are ideal applications for design of experiments in sense that the underlying parametric model for the switching dynamics of a single JJ can be derived directly from the laws of physics. With quantum mechanical arguments, it can be shown that the probability of the voltage response can be approximated by

$$\begin{aligned} P(Y=1) &= 1 - e^{-\exp(ax+b)} \\ P(Y=0) &= e^{-\exp(ax+b)}, \end{aligned} \tag{76}$$

where $a$ and $b$ are unknown parameters to be estimated and $x$ is the height of the current pulse. It follows that the measurement data can be modeled by a GLM with cloglog link function.

In an experiment carried out in Low Temperature Laboratory, Helsinki University of Technology in August 2005, a sample consisting of aluminium–aluminium oxide–aluminium Josephson junction circuit in a dilution refrigerator at 20 millikelvin temperature was connected to computer controlled measurement electronics in order to apply the current pulses and record the resulting voltage pulses. The resistance of the sample at room temperature suggested that a pulse of 300 nA always causes a switching (response 1), which gave the upper limit for the initial estimation. The lower limit for the initial estimation, 200 nA was roughly estimated from the dimensions of the Josephson junction by an experienced physicist. The experiment was carried out sequentially so that the height of pulse for stage was determined using the measurement data recorded on the earlier stages.