

Generalized Linear Models

Contents

1 The model fitting process	1
1.1 Step 0: Exploratory data analysis	1
2 Choosing a model	2
References	3

This section is primarily adapted starting from the textbook “An Introduction to Generalized Linear Models” (4th edition, 2018) by Annette J. Dobson and Adrian G. Barnett:

<https://doi.org/10.1201/9781315182780>

1 The model fitting process

Dobson and Barnett (2018, sec. 2.1) describes the model fitting process as follows:

The model fitting process described in this book involves four steps:

1. Model specification—a model is specified in two parts: an equation linking the response and explanatory variables and the probability distribution of the response variable.
2. Estimation of the parameters of the model.
3. Checking the adequacy of the model—how well it fits or summarizes the data.
4. Inference—for classical or frequentist inference this involves calculating confidence intervals, testing hypotheses about the parameters in the model and interpreting the results.

Note that Dobson and Barnett (2018) omits exploratory data analysis (EDA) from this initial list, but describes it later as an important preliminary step (Dobson and Barnett (2018, sec. 2.3.1)). We will treat EDA as a “step 0” that precedes these four steps.

1.1 Step 0: Exploratory data analysis

Dobson and Barnett (2018, sec. 2.3.1):

Any analysis of data should begin with a consideration of each variable separately, both to check on data quality (for example, are the values plausible?) and to help with model formulation.

1. What is the scale of measurement? Is it continuous or categorical? If it is categorical, how many categories does it have and are they nominal or ordinal?
2. What is the shape of the distribution? This shape can be examined using frequency tables, dot plots, histograms and other graphical methods.
3. How is it associated with other variables? Cross tabulations for categorical variables, scatter plots for continuous variables, side-by-side boxplots for continuous scale measurements grouped according to the factor levels of a categorical variable, and

other such summaries can help to identify patterns of association. For example, do the points on a scatter plot suggest linear or non-linear associations? Do the group means increase or decrease consistently with an ordinal variable defining the groups?

2 Choosing a model

The choice of statistical model depends on several considerations, primary among which is the type of response variable.

- Measured continuous values such as protein level, age, or weight can usually be modeled using linear regression, possibly after applying a logarithmic transformation.
- Patient survival times, which may be right-censored, call for time-to-event methods such as Kaplan-Meier estimation or Cox proportional hazards regression.
- If the outcome is binary (0/1), we can use logistic regression models.
- If the outcome is a count, we can use Poisson regression.
- If count data exhibit greater variance than assumed under the Poisson model, we can use negative binomial regression or overdispersed Poisson models.
- Other outcome structures correspond to other families within the generalized linear model framework.

In generalized linear modeling, we specify a linear predictor $\eta \stackrel{\text{def}}{=} \tilde{x} \cdot \tilde{\beta}$ of the same form as in linear regression. In principle, a linear predictor can yield any real value, whether positive, negative, or zero.

We then select an appropriate outcome distribution matched to the data type (for example, Gaussian for unbounded continuous values, gamma for positive continuous quantities, binomial for binary outcomes, and Poisson for counts).

Finally, we specify a link function $g(\cdot)$ that maps the expected outcome $\mu \stackrel{\text{def}}{=} E[Y \mid \tilde{X} = \tilde{x}]$ onto the real line $(-\infty, \infty)$ of the linear predictor. The inverse link function $g^{-1}(\cdot)$ transforms the linear predictor back to the scale of the expected outcome.

- Standard linear regression uses the identity link function $g(\mu) \stackrel{\text{def}}{=} \mu$ and assumes a Gaussian response distribution.
- When predicting an inherently positive outcome, a log link function $g(\mu) \stackrel{\text{def}}{=} \log\{\mu\}$ is often appropriate, since $\exp\{\eta\} > 0$ for all real η .
- An alternative to fitting a generalized linear model with a log link is to apply a log transformation directly to the outcome. While log-transforming the outcome works well for positive continuous measurements, it cannot be applied directly to binary outcomes or to count outcomes that can equal zero.

Table 1: R glm() Families

Family	Links
gaussian	identity , log, inverse
binomial	logit , probit, cauchit, log, cloglog
gamma	inverse , identity, log
inverse.gaussian	1/mu^2 , inverse, identity, log
Poisson	log , identity, sqrt
quasi	identity , logit, probit, cloglog, inverse, log, 1/mu^2 and sqrt
quasibinomial	logit , probit, identity, cloglog, inverse, log, 1/mu^2 and sqrt

Family	Links
quasipoisson	log , identity, logit, probit, cloglog, inverse, 1/ μ^2 and sqrt

Table 2: R `glm()` Link Functions; $\eta = X\beta = g(\mu)$

Name	Domain	Range	Link Function	Inverse Link Function
iden- tity	$(-\infty, \infty)$	$(-\infty, \infty)$	$\eta = \mu$	$\mu = \eta$
log	$(0, \infty)$	$(-\infty, \infty)$	$\eta = \log \mu$	$\mu = \exp\{\eta\}$
inverse	$(0, \infty)$	$(0, \infty)$	$\eta = 1/\mu$	$\mu = 1/\eta$
logit	$(0, 1)$	$(-\infty, \infty)$	$\eta = \log \mu / (1 - \mu)$	$\mu = \exp\{\eta\} / (1 + \exp\{\eta\})$
probit	$(0, 1)$	$(-\infty, \infty)$	$\eta = \Phi^{-1}(\mu)$	$\mu = \Phi(\eta)$
cloglog	$(0, 1)$	$(-\infty, \infty)$	$\eta = \log - \log 1 - \mu$	$\mu = 1 - \exp\{-\exp\{\eta\}\}$
1/ μ^2	$(0, \infty)$	$(0, \infty)$	$\eta = 1/\mu^2$	$\mu = 1/\sqrt{\eta}$
sqrt	$(0, \infty)$	$(0, \infty)$	$\eta = \sqrt{\mu}$	$\mu = \eta^2$

References

Dobson, Annette J, and Adrian G Barnett. 2018. *An Introduction to Generalized Linear Models*. 4th ed. CRC press. <https://doi.org/10.1201/9781315182780>.