Generalized Linear Models

Contents

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 the four steps above.

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 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 type of predictive model one uses depends on several issues; one is the type of response.

- Measured values such as quantity of a protein, age, weight usually can be handled in an ordinary linear regression model, possibly after a log transformation.
- Patient survival, which may be censored, calls for a different method (survival analysis, Cox regression).
- If the response is binary, then can we use logistic regression models
- If the response is a count, we can use Poisson regression
- If the count has a higher variance than is consistent with the Poisson, we can use a negative binomial or over-dispersed Poisson
- Other forms of response can generate other types of generalized linear models

We need a linear predictor of the same form as in linear regression βx. In theory, such a linear predictor can generate any type of number as a prediction, positive, negative, or zero

We choose a suitable distribution for the type of data we are predicting (normal for any number, gamma for positive numbers, binomial for binary responses, Poisson for counts)

We create a link function which maps the mean of the distribution onto the set of all possible linear prediction results, which is the whole real line $(-\infty, \infty)$. The inverse of the link function takes the linear predictor to the actual prediction.

- Ordinary linear regression has identity link (no transformation by the link function) and uses the normal distribution
- If one is predicting an inherently positive quantity, one may want to use the log link since ex is always positive.
- An alternative to using a generalized linear model with a log link, is to transform the data using the log. This is a device that works well with measurement data and may be usable in other cases, but it cannot be used for 0/1 data or for count data that may be 0.

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
quasipoisson	**log**, identity, logit, probit, cloglog, inverse, 1/mu^2 and sqrt

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

Name	Domain	Range	Link Function	Inverse Link Function
identity	$(-\infty, \infty)$	$(-\infty, \infty)$	$\eta = \mu.$	$\mu = \eta$
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/mu^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.