Chapter 20

Generalized Linear Models I

Today's Menu

Today we introduce Generalized Linear Models (GLM) and look at logistic regression (and some extensions):

- Logistic regression: Y is binary, fixed effects only.
- Mixed-effects logistic regression (GLMM): Y is binary, fixed and random effects.
- Ordinal regression: Y is ordinal (e.g. 5-point Likert).

Generalized Linear Models

Generalized Linear Models

GLMs go back to Nelder & Wedderburn (1972)[1]. Let's start from a linear regression perspective.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i.$$

[1] Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalised Linear Models. Journal of the Royal Statistical Society – Series A, p. 370–384.

Generalized Linear Models

GLMs go back to Nelder & Wedderburn (1972)[1]. Let's start from a linear regression perspective.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i.$$

Let's take the expected value (note that $E(\varepsilon_i) = 0$).

$$E(Y_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi}.$$

and rewrite it as

$$\mu_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi}.$$

[1] Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalised Linear Models. Journal of the Royal Statistical Society – Series A, p. 370–384.

Generalized Linear Models

GLMs go back to Nelder & Wedderburn (1972)[1]. Let's start from a linear regression perspective.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i.$$

Let's take the expected value (note that $E(\varepsilon_i) = 0$).

$$E(Y_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi}.$$

and rewrite it as

$$\mu_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi}.$$

Sometimes it is convenient to wrap a "link function" $g()$ around μ_i:

$$g(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi}.$$

That's the GLM! Now it all comes down to the distribution of the error terms (and which link function we pick).

[1] Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalised Linear Models. Journal of the Royal Statistical Society – Series A, p. 370–384.

Error Distributions, Link Functions

The function to fit GLMs in R is `glm()`[2]:

Error Distribution	Default Link Function
`binomial`	`(link = "logit")`
`gaussian`	`(link = "identity")`
`Gamma`	`(link = "inverse")`
`inverse.gaussian`	`(link = "1/mu^2")`
`poisson`	`(link = "log")`
`quasi`	`(link = "identity", variance = "constant")`
`quasibinomial`	`(link = "logit")`
`quasipoisson`	`(link = "log")`

Two important extensions are negative-binomial regression `glm.nb()` and beta regression `betareg()` which we will cover in the next unit.

[2] `glm(formula, family = familytype(link = "linkfunction"))`

[Generalized Linear Models I]

Logistic Regression

Distributions

Having binary responses (outcome $Y = 0$ or $Y = 1$) at an individual level, this variable Y can be described by a *Bernoulli distribution* with parameter p as the probability of "success".

$$P(Y = k) = p^k (1 - p)^{1-k}$$

That is, for $k = 1 \rightarrow P(k; p) = p$, and for $k = 0 \rightarrow P(k; p) = 1 - p$.

Distributions

Having binary responses (outcome $Y = 0$ or $Y = 1$) at an individual level, this variable Y can be described by a *Bernoulli distribution* with parameter p as the probability of "success".

$$P(Y = k) = p^k (1 - p)^{1-k}$$

That is, for $k = 1 \rightarrow P(k; p) = p$, and for $k = 0 \rightarrow P(k; p) = 1 - p$.

If individuals are exposed to several 0/1 trials, we can aggregate the responses the k "successes" in n trials (k/n would be the corresponding proportion). In this case, the Bernoulli distribution extends to the *Binomial distribution*. The probability of achieving k successes in n trails is

$$P(Y = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

The Bernoulli distribution is a special case of the binomial distribution for $n = 1$. This is why in `R` we always say `family = binomial`.

The classical setup in logistic regression is to operate on the 0/1 level. However, as we will see, for some experimental settings it is terribly helpful to use the aggregated specification (sometimes also called *binomial regression*).

Why Not Using a Normal Linear Model?

Applying normal linear model on a 0/1 response or on proportions (i.e., *k* success in *n* trials) is not a good idea:

- Proportions are bounded between 0 and 1. Obviously a linear model would predict values outside these boundaries. In addition, such data often tend to plateau near 0 and 1 and this would clearly violate normality.
- The variance is not constant across the range of possible proportions: it is larger near the middle and smaller at the ends (heteroscedasticity).

In a Gaussian model, the variance σ^2 is independent from the mean μ, and therefore constant across different response values. A binomial distribution has,

- mean: np,
- variance: $np(1-p)$

We see that the variance changes if the mean changes. This is a super desirable property which applies to all GLM distributions (except for the Gaussian).

Model Formulation

Let us start with the classical logistic regression formulation where the response Y is binary, i.e. it takes values of either 0 or 1. By default we use the logit link and the errors are binomial distributed. The regression model looks as follows:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

with $p = P(Y = 1)$ as the probability of “success”.

Model Formulation

Let us start with the classical logistic regression formulation where the response Y is binary, i.e. it takes values of either 0 or 1. By default we use the logit link and the errors are binomial distributed. The regression model looks as follows:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

with $p = P(Y = 1)$ as the probability of "success".

This equation can be rearranged as

$$\left(\frac{p}{1-p}\right) = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)$$

or,

$$\left(\frac{p}{1-p}\right) = \exp(\beta_0) \times \exp(\beta_1 X_1) \times \exp(\beta_2 X_2) \times \cdots \times \exp(\beta_p X_p)$$

The left hand side of the variable reflects an odds.

Model Formulation

Let us compute an odds ratio involving the odds above and a second odds which reflects a 1-unit change in X_1, while keeping all the other variables constant, that is,

$$
\begin{aligned}
OR(X_1) &= \frac{\exp(\beta_0) \times \exp(\beta_1(X_1+1)) \times \exp(\beta_2 X_2) \times \cdots \times \exp(\beta_p X_p)}{\exp(\beta_0) \times \exp(\beta_1 X_1) \times \exp(\beta_2 X_2) \times \cdots \times \exp(\beta_p X_p)} \\
&= \frac{\exp(\beta_1(X_1+1))}{\exp(\beta_1 X_1)} \\
&= \exp(\beta_1(X_1+1) - \beta_1 X_1) \\
&= \exp(\beta_1 X_1 + \beta_1 - \beta_1 X_1) \\
&= \exp(\beta_1).
\end{aligned}
$$

Interpretation: The odds to fall in category $Y = 1$ multiply by $\exp(\beta_1)$ if we increase X_1 by 1 unit. This representation is important for parameter interpretation.

We can now do the same thing for the remaining β's. The corresponding odds ratios can be used as effect sizes.

Goodness-of-Fit

Logistic regression fits a logistic (S-shaped) curve between the predictors and the response probabilities. In terms of goodness-of fit we have the following options:

- Deviance (LR-test): Compare two models using the `anova()` function.
- Pseudo R^2: various measures have been proposed in the literature but they are pretty much useless.
- ROC curves: Nice tool to get an intuitive measure for goodness-of-fit (very popular in medicine; see packages `pROC` and `ROCR`).

Concluding remarks:
If your predictors are categorical and you include interactions, the interpretation gets a bit trickier compared to an ANOVA setting[3]. If we have one binary predictor only we end up with a χ^2-test.

[3] see Jaccard, J. J. (2001). Interaction Effects in Logistic Regression. Sage

[Generalized Linear Models I]

Mixed-Effects Logistic Regression

Mixed-Effects Logistic Regression

Now we combine what we have learned in the mixed-effects units (right hand side of the equation) with logistic regression (left hand side of the equation). For instance,

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + U$$

In `R` the `glmer()` function in `lme4` does the job.

Nothing changes in terms of fixed-effects interpretation. Again we can play around with random intercepts and random slopes for the random effect.

In the code we show an example of a binomial regression, i.e., we operate on an aggregated “success” level. Important: We can’t use the proportions directly as response but rather we need to provide the number of successes and the number of failures (per participant).

[Generalized Linear Models I]

Ordinal Logistic Regression

Proportional Odds Model

Having an ordinal response Y (e.g. Likert), the most popular model is the *proportional odds model*. It is based on a sequence of logistic regression models (cumulative logits)[4]:

$$\log\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right) = \alpha_j + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Each cumulative logit has its own intercept α_j which increases in j (ordinal response). This intercept determines the horizontal shift of the logistic function. Each cumulative logit gets its own intercept but we only get one set of β parameters.

For instance for $j = 2$ vs. $j = 3$ the curve for $P(Y \leq 2)$ is the curve $P(Y \leq 3)$ shifted by $(\alpha_3 - \alpha_2)/\beta_1$ units in X_1 direction. This applies to all the predictors and all the adjacent category logits; that's why McCullagh (1980) calls it a proportional odds model.

[4]Note that $1 - P(Y \leq j) = P(Y > j)$.

Computation and Extensions

Proportional odds models can be fitted in R using the following functions:

- `polr()` in `MASS`: base implementation, we can produce effects plots.
- `vglm()` in `vgam`: very flexible package, allows for many extensions.
- `clm()` in `ordinal`: it also allows for the inclusion of random effects (see `clmm`).

Other ordinal logistic models can easily be specified (see Hosmer et al., 2013). These extensions can be computed using the `vglm()` in the `vgam` package.

If the response is nominal, the multinomial logit model does the job. See `mlogit` package.

Additional Literature

Two GLM books you must bring with you to a desert island:

Fox, J. (2016). Applied Regression Analysis & Generalized Linear Models (3rd ed.). Sage.
$\rightarrow$ everything you need to know in terms of (generalized) linear modeling, perfect balance between technical details and application, not much details however on mixed-effects specifications.

Stroup, W. W. (2013). Generalized Linear Mixed Models. CRC Press.
$\rightarrow$ extremely profound GLMM book, parts of it are slightly technical, code in SAS unfortunately but that shouldn't be a problem anymore at this point.

If you need more details on logistic and ordinal/multinomial regression, here's the bible (again, perfect technical/applied balance, examples fully worked out in `R`):

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.).Wiley.