Generalized Linear Models: Distributions and Link Functions

Ashley I Naimi

Spring 2024

Contents

1 Generalized Linear Models

Generalized linear models consist of a family of regression models that are fully characterized by a selected distribution and a link function. That is, to fully specify a GLM, one must select a distribution (which determines the form of the conditional mean and variance of the outcome) and a link function (which determines how the conditional mean of the outcome relates to the covariates).

There are a wide variety of distributions and link functions available in standard statistical software programs that fit GLMs. Here, we'll consider a binary outcome Y with probability $P(Y = 1)$, and focus attention on three link functions:

1. Logit, or the log-odds: $\log P(Y = 1)/[1 - P(Y = 1)]$
2. Log: $\log[P(Y = 1)]$
3. Identity: $P(Y = 1)$.

A common misconception is that to use GLMs correctly, one must choose the distribution that best characterizes the data (the "correct" distribution), as well as the canonical link function corresponding to this distribution. For example, if the outcome is binary, one "must" choose the binomial distribution with the logit link. While the binomial distribution and logit link work well together for binary outcomes, they do not easily provide contrasts like the risk difference or risk ratio, because of the selected link function. Alternative specification of the distribution and link function for GLMs can address this limitation.

2 Link Functions and Effect Measures

There is an important relation between the chosen link function, and the interpretation of the coefficients from a GLM. For models of a binary outcome and the logit or log link, this relation stems from the properties and rules governing the natural logarithm. The quotient rule states that $\log(X/Y) = \log(X) - \log(Y)$.

Because of this relation, the natural exponent of the coefficient in a logistic regression model yields an estimate of the odds ratio. Too see why, we can evaluate the logit link function in a regression model:

$$
\begin{aligned}
&\log\left[\frac{P(Y=1 \mid X)}{P(Y=0 \mid X)}\right] = \beta_0 + \beta_1 X \\
\implies &\beta_1 = \log\left[\frac{P(Y=1 \mid X=1)}{P(Y=0 \mid X=1)}\right] - \log\left[\frac{P(Y=1 \mid X=0)}{P(Y=0 \mid X=0)}\right] \\
\implies &\beta_1 = \log\left[\frac{P(Y=1|X=1)}{P(Y=0|X=1)} \Big/ \frac{P(Y=1|X=0)}{P(Y=0|X=0)}\right] \\
\implies &\exp(\beta_1) = \frac{P(Y=1|X=1)}{P(Y=0|X=1)} \Big/ \frac{P(Y=1|X=0)}{P(Y=0|X=0)}
\end{aligned}
$$

However, by the same reasoning, exponentiating the coefficient from a GLM with a log link function and a binomial distribution (i.e., log-binomial regression) yields an estimate of the risk ratio:

$$
\begin{aligned}
&\log\left[P(Y=1 \mid X)\right] = \beta_0 + \beta_1 X \\
\implies &\beta_1 = \log\left[P(Y=1 \mid X=1)\right] - \log\left[P(Y=1 \mid X=0)\right] \\
\implies &\beta_1 = \log\left[\frac{P(Y=1 \mid X=1)}{P(Y=1 \mid X=0)}\right] \\
\implies &\exp(\beta_1) = \frac{P(Y=1 \mid X=1)}{P(Y=1 \mid X=0)}
\end{aligned}
$$

Alternately, for GLM models with a binomial distribution and identity link function, because logarithms are not used, the unexponentiated coefficient yields an estimate of the risk difference:

$$
\begin{aligned}
&\left[P(Y=1 \mid X)\right] = \beta_0 + \beta_1 X \\
\implies &\beta_1 = \left[P(Y=1 \mid X=1)\right] - \left[P(Y=1 \mid X=0)\right]
\end{aligned}
$$

Unfortunately, using a binomial distribution can lead to convergence problems with the $\log()$ or identity link functions (Zou, 2004). This will occur when, for example, the combined numerical value of all the independent variables in the model is large enough to cause the estimated probabilities to exceed 1, which violates the very definition of a probability (binomial) model (probabilities can only lie between zero and one) and hence, convergence problems. Let's see how these problems can be overcome.

3 A Data Example

We use data from the National Health and Nutrition Examination Survey (NHEFS). We are interested primarily in the covariate adjusted association (on the risk difference and risk ratio scales) between quitting smoking and a greater than median weight change between 1971 and 1982.

In our analyses, we regress an indicator of greater than median weight change against an indicator of whether the person quit smoking. We adjust for exercise status, sex, age, race, income, marital status, education, and indicators of whether the person was asthmatic or had bronchitis. We start by loading the data:

```
#' Load relevant packages
packages <- c("broom", "here", "tidyverse",
    "skimr", "rlang", "sandwich", "boot",
    "kableExtra")

for (package in packages) {
    if (!require(package, character.only = T,
        quietly = T)) {
        install.packages(package, repos = "http://lib.stat.cmu.edu/R/CRAN")
    }
}

for (package in packages) {
    library(package, character.only = T)
}

#' Define where the data are
file_loc <- url("https://bit.ly/47ECRcs")

#' This begins the process of cleaning and formatting the data
nhefs <- read_csv(file_loc) %>%
    select(qsmk, wt82_71, wt82, wt71, exercise,
        sex, age, race, income, marital,
```

```
        school, asthma, bronch, starts_with("alcohol"),
        -alcoholpy, starts_with("price"),
        starts_with("tax"), starts_with("smoke"),
        smkintensity82_71) %>%
    mutate(income = as.numeric(income > 15),
        marital = as.numeric(marital > 2),
        alcoholfreq = as.numeric(alcoholfreq >
            1)) %>%
    na.omit(.)

factor_names <- c("exercise", "income", "marital",
    "sex", "race", "asthma", "bronch")
nhefs[, factor_names] <- lapply(nhefs[, factor_names],
    factor)

#' Define outcome
nhefs <- nhefs %>%
    mutate(id = row_number(), wt_delta = as.numeric(wt82_71 >
        median(wt82_71)), .before = qsmk)

#' Quick summary of data
nhefs %>%
    print(n = 5)
```

```
## # A tibble: 1,055 x 27
##      id wt_delta  qsmk wt82_71  wt82  wt71 exercise sex     age race  income
##   <int>    <dbl> <dbl>   <dbl> <dbl> <dbl> <fct>    <fct> <dbl> <fct> <fct>
## 1     1        0     0  -10.1   68.9  79.0 2        0        42 1     1
## 2     2        0     0    2.60  61.2  58.6 0        0        36 0     1
## 3     3        1     0    4.99  64.4  59.4 2        0        68 1     0
## 4     4        1     0    4.99  92.1  87.1 1        0        40 0     1
## 5     5        1     0    4.42 103.   99   1        1        43 1     0
## # i 1,050 more rows
## # i 16 more variables: marital <fct>, school <dbl>, asthma <fct>, bronch <fct>,
## #   alcoholfreq <dbl>, alcoholtype <dbl>, alcoholhowmuch <dbl>, price71 <dbl>,
```

```
## #   price82 <dbl>, price71_82 <dbl>, tax71 <dbl>, tax82 <dbl>, tax71_82 <dbl>,
## #   smokeintensity <dbl>, smokeyrs <dbl>, smkintensity82_71 <dbl>
```

4 GLMs for Risk Differences and Ratios

For our analyses of the data described above using GLM with a binomial distributed outcome with a log link function to estimate the risk ratio and identity link function to estimate risk difference, an error is returned:

```
#' Here, we start fitting relevant regression models to the data.
#' modelForm is a regression argument that one can use to regress the
#' outcome (wt_delta) against the exposure (qsmk) and selected confounders.

formulaVars <- paste(names(nhefs)[c(3,7:16)],collapse = "+")
modelForm <- as.formula(paste0("wt_delta ~", formulaVars))
modelForm
```

```
## wt_delta ~ qsmk + exercise + sex + age + race + income + marital +
##     school + asthma + bronch + alcoholfreq
```

```
#' This model can be used to quantify a conditionally adjusted
#' odds ratio with correct standard error
modelOR <- glm(modelForm,data=nhefs,family = binomial("logit"))
tidy(modelOR)[2,]
```

```
## # A tibble: 1 x 5
##   term  estimate std.error statistic   p.value
##   <chr>    <dbl>     <dbl>     <dbl>     <dbl>
## 1 qsmk     0.623     0.153      4.07 0.0000471
```

```
#' This model can be used to quantify a conditionally adjusted risk
#' ratio with with correct standard error
#' However, error it returns an error and thus does not provide any results.
modelRR_binom <- glm(modelForm,data=nhefs,family = binomial("log"))
```

```
## Error: no valid set of coefficients has been found: please supply starting values
```

Why is this error returned? We are modeling $P(Y = 1 \mid X) = \exp\{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p\}$. In this context, there may be *no set of values* for the parameters in the model that yield $P(Y = 1 \mid X) < 1$ for every observation in the sample. Because R's glm function (under a binomial distribution) correctly recognizes this as a problem, it returns an error.

The most commonly proposed solution is to follow the error's advice and provide starting values to initiate the algorithm, but there are two problems with this. The first is that it often doesn't work:

```
#' This model can be used to quantify a conditionally adjusted risk
#' ratio with with correct standard error
#' It adds starting values for the parameters in an attempt to
#' get the model to converge.
#' However, it still returns an error.
modelRR_binom <- glm(modelForm,
                     data=nhefs,
                     family = binomial("log"),
                     start = rep(0, 13))
```

```
## Error: cannot find valid starting values: please specify some
```

The second problem is that, even if it does work, the results may be strongly dependent on the starting values of the algorithm. While this is not a fundamental problem, it can create some challenges in interpreting the results we obtain.

Instead, one may resort to using different distributions that are more compatible with the link functions that return the association measures of interest. For the risk ratio, one may use a GLM with a Poisson distribution and log link function. Doing so will return an exposure coefficient whose natural exponent can be interpreted as a risk ratio.

```
#' This model can be used to quantify a conditionally risk ratio
#' using the Poisson distribuiton and log link function.
#' However, because the Poisson distribution is used, the model
#' provides incorrect standard error estimates.
```

```
modelRR <- glm(modelForm, data = nhefs, family = poisson("log"))
tidy(modelRR)[2, ]
```

```
## # A tibble: 1 x 5
##   term  estimate std.error statistic p.value
##   <chr>    <dbl>     <dbl>     <dbl>   <dbl>
## 1 qsmk     0.277    0.0982      2.82 0.00482
```

It's important to recognize what we're doing here. We are using this model as a tool to quantify the log mean ratio contrasting $P(Y = 1 \mid X_{qsmk} = 1)$ to $P(Y = 1 \mid X_{qsmk} = 0)$ (all other things being equal). However, we should not generally assume that ever aspect of this model is correct. In particular, note that the max predicted probability from this model is 1.087:

```
summary(modelRR$fitted.values)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2214  0.3961  0.4865  0.4995  0.5857  1.0873
```

We can use the `augment` function in the `broom` package to evaluate the distribution of these probabilities (among other things):

```
fitted_dat <- augment(modelRR, type.predict = "response")

fitted_dat
```

```
## # A tibble: 1,055 x 18
##    wt_delta  qsmk exercise sex     age race  income marital school asthma bronch
##       <dbl> <dbl> <fct>    <fct> <dbl> <fct> <fct>  <fct>    <dbl> <fct>  <fct>
##  1        0     0 2        0        42 1     1      0            7 0      0
##  2        0     0 0        0        36 0     1      0            9 0      0
##  3        1     0 2        0        68 1     0      1            5 0      0
##  4        1     0 1        0        40 0     1      0           11 0      0
##  5        1     0 1        1        43 1     0      1            9 0      0
##  6        0     0 2        0        51 0     1      0           10 0      0
##  7        1     0 2        0        43 0     1      0           11 0      0
```

```
##  8        1     1 1        0        43 0     1      0           12 0      0
##  9        0     0 2        0        34 0     1      0           12 0      0
## 10        1     0 0        1        47 0     1      0           12 0      0
## # i 1,045 more rows
## # i 7 more variables: alcoholfreq <dbl>, .fitted <dbl>, .resid <dbl>,
## #   .hat <dbl>, .sigma <dbl>, .cooksd <dbl>, .std.resid <dbl>
```

```
plot_hist <- ggplot(fitted_dat) + geom_histogram(aes(.fitted)) +
    scale_y_continuous(expand = c(0, 0)) +
    scale_x_continuous(expand = c(0, 0))

ggsave(here("figures", "2022_02_21-rr_hist_plot.pdf"),
    plot = plot_hist)
```

This distribution is shown in margin Figure 1. We can also see that there are only two observations in the sample with predicted risks greater than 1.

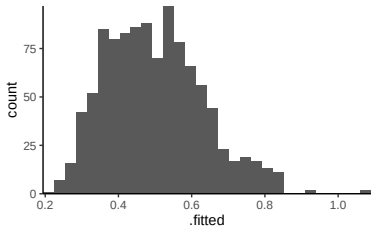

Figure 1: Distribution of fitted values from the Poisson GLM with log link function to obtain an estimate of the adjusted risk ratio for the association between quitting smoking and greater than median weight gain in the NHEFS.

```
fitted_dat %>%
    filter(.fitted >= 1) %>%
    select(wt_delta, qsmk, age, .fitted)
```

```
## # A tibble: 2 x 4
##   wt_delta  qsmk   age .fitted
##      <dbl> <dbl> <dbl>   <dbl>
## 1        1     1    32    1.06
## 2        1     1    25    1.09
```

For these reasons, we are not particularly concerned about the fact that the model predicts risks that are slightly large than 1. However, the model-based standard errors (i.e., the SEs that one typically obtains directly from the GLM output) are no longer valid. Instead, one should use the robust (or sandwich) variance estimator to obtain valid SEs (the bootstrap can also be used) (Zou, 2004). The easiest way to do this in R is to use the `lmtest` and the `sandwich` packages:

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(sandwich)
#' To obtain the correct variance, we use the "sandwich"
#' function to obtain correct sandwich (robust) standard
#' error estimates.
coeftest(modelRR, vcov = sandwich(modelRR))
```

```
##
## z test of coefficients:
##
##               Estimate Std. Error z value  Pr(>|z|)
## (Intercept)  0.1750741  0.2192840  0.7984   0.42464
## qsmk         0.2768027  0.0642420  4.3088 1.642e-05 ***
## exercise1   -0.0965659  0.0746981 -1.2927   0.19610
## exercise2   -0.1729715  0.0836936 -2.0667   0.03876 *
## sex1         0.0257665  0.0626612  0.4112   0.68092
## age         -0.0187890  0.0028121 -6.6814 2.366e-11 ***
## race1       -0.1086352  0.0972719 -1.1168   0.26407
## income1     -0.0661111  0.0915444 -0.7222   0.47019
## marital1     0.0378850  0.0749873  0.5052   0.61340
## school      -0.0056906  0.0119209 -0.4774   0.63310
## asthma1      0.2150948  0.1205733  1.7839   0.07443 .
## bronch1     -0.0632391  0.1271851 -0.4972   0.61903
## alcoholfreq  0.0643882  0.0637096  1.0107   0.31218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For the risk difference, one may use a GLM with a Gaussian (i.e., normal) distribution and identity link function, or, equivalently, an ordinary least squares estimator. Doing so will return an exposure coefficient that can be interpreted as a risk difference. However, once again the robust variance estimator (or bootstrap) should be used to obtain valid SEs.

```
#' This model can be used to obtain a risk difference
#' with the gaussian distribiton or using ordinary least
#' squares (OLS, via the lm function). Again, the model
#' based standard error estimates are incorrect.
modelRD <- glm(modelForm, data = nhefs, family = gaussian("identity"))
modelRD <- lm(modelForm, data = nhefs)
tidy(modelRD)[2, ]
```

```
## # A tibble: 1 x 5
##   term  estimate std.error statistic   p.value
##   <chr>    <dbl>     <dbl>     <dbl>     <dbl>
## 1 qsmk     0.145    0.0353      4.10 0.0000438
```

```
#' To obtain the correct variance, we use the 'sandwich' function
#' to obtain correct sandwich (robust) standard error estimates.
coeftest(modelRD, vcov = sandwich(modelRD))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  0.9277492  0.1072168  8.6530 < 2.2e-16 ***
## qsmk         0.1450221  0.0347495  4.1734 3.252e-05 ***
## exercise1   -0.0513160  0.0402869 -1.2738   0.20303
## exercise2   -0.0875533  0.0427215 -2.0494   0.04067 *
## sex1         0.0132564  0.0312989  0.4235   0.67199
## age         -0.0089792  0.0012601 -7.1255 1.932e-12 ***
## race1       -0.0526963  0.0439398 -1.1993   0.23069
## income1     -0.0356118  0.0451456 -0.7888   0.43040
## marital1     0.0193736  0.0382798  0.5061   0.61289
```

```
## school      -0.0021571  0.0056937 -0.3789   0.70487
## asthma1      0.1108577  0.0683715  1.6214   0.10523
## bronch1     -0.0285204  0.0591734 -0.4820   0.62992
## alcoholfreq  0.0319382  0.0311041  1.0268   0.30475
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The risk ratio and difference, as well as the 95% sandwich variance confidence intervals, obtained for the relation between quitting smoking and greater than median weight change are provided Table 1.

```
knitr::kable(table1_data)
```

Method	Risk Difference	Risk Ratio
GLM	0.14 (0.09, 0.20)	1.32 (1.19, 1.46)
Marginal Standardization	0.14 (0.09, 0.21)	1.31 (1.18, 1.46)

Results in this table obtained using a conditionally adjusted regression model without interactions. Gaussian distribution and identity link was used to obtain the risk difference. A Poisson distribution and log link was used to obtain the risk ratio. 95% CIs obtained via the sandwich variance estimator. 95% CIs obtained using the bias-corrected and accelerated bootstrap CI estimator.

Unfortunately, use of a Poisson or Gaussian distribution for GLMs for a binomial outcome can introduce different problems. For one, while not entirely worrysome in our setting, a model that predicts probabilities greater than one should not instill confidence in the user. Second, performance of the robust variance estimator is notoriously poor with small sample sizes. Finally, the interpretation of the risk differences and ratios becomes more complex when the exposure interacts with other variables in the model.

Table 1: Methods to use for quantifying conditionally adjusted odds ratios, risk ratios, and risk differences.

Odds Ratio	Risk Ratio	Risk Difference
GLM Family = Binomial	GLM Family = Binomial	GLM Family = Binomial
GLM Link = Logistic	GLM Link = Log	GLM Link = Identity
Standard Errors = Model Based	Standard Errors = Model Based	Standard Errors = Model Based
	GLM Family = Poisson	GLM Family = Gaussian
	GLM Link = Log	GLM Link = Identity
	Standard Errors = Sandwich	Standard Errors = Sandwich
		Least Squares Regression
		Standard Errors = Sandwich

For these reasons, marginal standardization should be generally considered as a first line estimator of contrasts of interest. When predicted risks are estimated using a logistic model, relying on marginal standardization will not result in probability estimates outside the bounds [0, 1]. And because the robust variance estimator is not required, model-based standardization will not be as affected by small sample sizes. However, the bootstrap is more computationally demanding than alternative variance estimators, which may pose problems in larger datasets.

References

Guangyong Zou. A modified poisson regression approach to prospective studies with binary data. *Am J Epidemiol*, 159(7):702–706, Apr 2004.