# Multivariate logistic distribution

July 2, 2019

## 1 Latent variable notation

We are interested in modeling multivariate binomial outcomes. We will write the observed outcome as $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ip})'$, a vector of binary variables. To keep our notations consistent, we will model $\mathbf{y}_i$ using a latent variable approach:

$$\begin{aligned} \mathbf{y}_i &= 1(\mathbf{z}_i > 0) \\ \mathbf{z}_i &\sim \text{distrib}(\mu, \Sigma^2). \end{aligned} \tag{1}$$

where $\mathbf{z}_i$ is a vector of continuous variable following a probability distribution with mean $\mu$ and covariance $\Sigma^2$. This approach generalizes the univariate logistic regression.

## 2 Models

### 2.1 Logistic regression

First, consider the univariate logistic regression (with intercept only):

$$\text{logit } \Pr(y_i = 1) = \mu. \tag{2}$$

This is equivalent to writing

$$\begin{aligned} \mathbf{y}_i &= 1(\mathbf{z}_i > 0) \\ \mathbf{z}_i &\sim \text{logistic}(\mu, 1), \end{aligned} \tag{3}$$

where $\text{logistic}(\cdot|\mu, s)$ is a logistic *distribution* with location parameter $\mu$ and a scale parameter $s$.

One way to extend this model to account for multivariate observation is to use a mixed model approach:

$$\begin{aligned} \text{logit } \Pr(\mathbf{Y}_i = y_i) &= \boldsymbol{\mu} + \mathbf{r}_i \\ \mathbf{r}_i &\sim \mathcal{N}(0, \sigma^2 R), \end{aligned} \tag{4}$$

where $\mathcal{N}(0, \sigma^2 R)$ is a multivariate normal distribution with covariance $\sigma^2 R$ ($R$ is the correlation matrix). For convenience, we assume that random effects

variance $\sigma^2$ is constant among response variables. Equivalently, we can write this as

$$\begin{aligned} \mathbf{y}_i &= 1(\mathbf{z}_i > 0) \\ \mathbf{z}_i &\sim \text{logistic}(\boldsymbol{\mu} + \mathbf{r}_i, 1) \\ \mathbf{r}_i &\sim \mathcal{N}(0, \sigma^2 R), \end{aligned} \tag{5}$$

which can be further expanded as

$$\begin{aligned} \mathbf{y}_i &= 1(\mathbf{z}_i > 0) \\ \mathbf{z}_i &= \boldsymbol{\mu} + \mathbf{r}_i + \boldsymbol{\epsilon}_i \\ \mathbf{r}_i &\sim \mathcal{N}(0, \sigma^2 R) \\ \epsilon_{ij} &\sim \text{logistic}(0, 1) \end{aligned} \tag{6}$$

Essentially, we have a continuous latent variable $\mathbf{z}_i$ which has a mean $\boldsymbol{\mu}$ and two "error" terms: $\mathbf{r}_i$, which follows a multivariate normal with covariance $\sigma^2 R$, and $\boldsymbol{\epsilon}_i$, which follows an independent logistic (each marginal distribution is an i.i.d. logistic distribution). Due to two levels of uncertainties, it becomes much harder to estimate the correlation structure $R$. I don't have a good analytical argument for this but I hope this is somewhat intuitive... I'll compare this with other models later; this might make things slightly clearer.

**Identifiability of the random effects variance**

It doesn't seem like Jonathan is completely convinced that $\sigma^2$ is not identifiable; he says that it is "practically" unidentifiable. Let's try to do some math. Consider a univariate logistic regression with an underlying normal random effects on the mean:

$$\begin{aligned} y_i &= 1(z_i > 0) \\ z_i &\sim \text{logistic}(\mu + r_i, 1) \\ r_i &\sim \mathcal{N}(0, \sigma^2) \end{aligned} \tag{7}$$

Then, the marginal likelihood of this model can be written as

$$\prod_{i=1}^{n} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left\{ 1(z_i > 0)^{y_i} 1(z_i \leq 0)^{1-y_i} \right\} f(z_i | \mu + r_i, 1) dz_i g(r_i | 0, \sigma^2) dr_i, \tag{8}$$

where $f$ is the pdf of the standard logistic distribution and $g$ is the pdf of the standard normal. This is ugly. When $y_i = 0$, we have

$$\begin{aligned} &\int_{-\infty}^{\infty} \left\{ 1(z_i > 0)^{y_i} 1(z_i \leq 0)^{1-y_i} \right\} f(z_i | \mu + r_i, 1) dz_i \\ &= \int_{-\infty}^{0} f(z_i | \mu + r_i, 1) dz_i \\ &= \frac{1}{1 + \exp(\mu + r_i)} \end{aligned} \tag{9}$$

Then,

$$\begin{aligned}&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left\{1(z_i>0)^{y_i}1(z_i\leq 0)^{1-y_i}\right\}f(z_i|\mu+r_i,1)dz_i g(r_i|0,\sigma^2)dr_i\\&=\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}\frac{1}{1+\exp(\mu+r_i)}\exp\left(-\frac{r_i^2}{2\sigma^2}\right)dr_i\end{aligned}\tag{10}$$

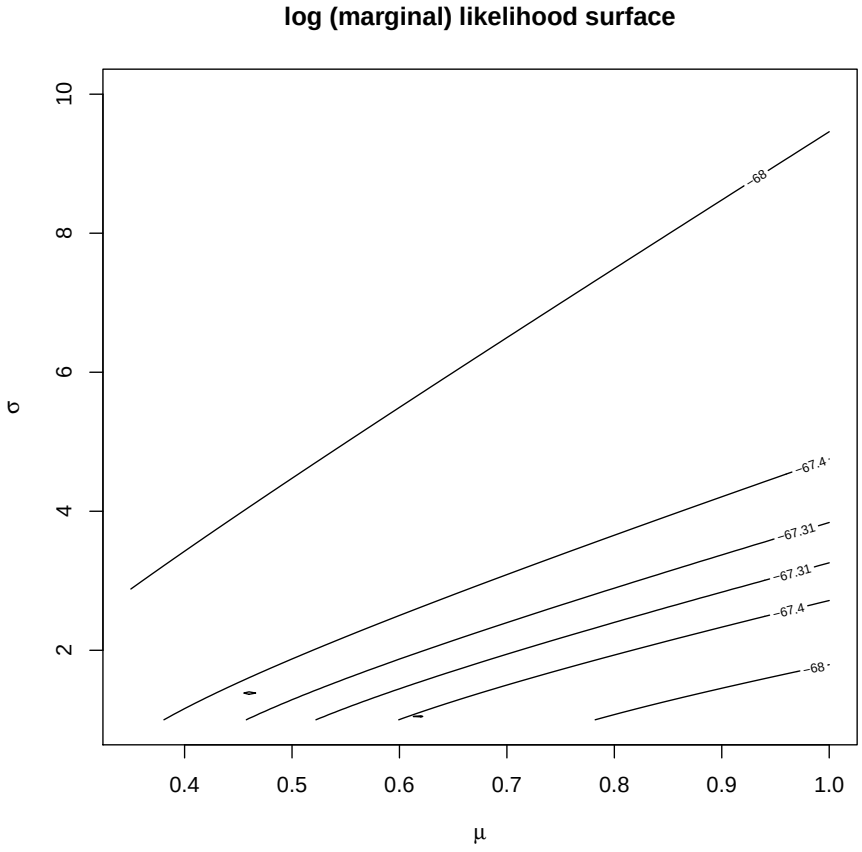
I can't evaluate this integral analytically but eventually the marginal likelihood can be written as

$$\begin{aligned}&\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n\left(\int_{-\infty}^{\infty}\frac{1}{1+\exp(\mu+r_i)}\exp\left(-\frac{r_i^2}{2\sigma^2}\right)dr_i\right)^{n_0}\\&\times\left(\int_{-\infty}^{\infty}\left(1-\frac{1}{1+\exp(\mu+r_i)}\right)\exp\left(-\frac{r_i^2}{2\sigma^2}\right)dr_i\right)^{n_1},\end{aligned}\tag{11}$$

where $n_0$ is the number of 0's and $n_1$ is the number of 1's.

We can work out a numerical example. Assume $n_1 = 60$ and $n_0 = 40$. Then, the MLE of $\mu$ of the logistic regression without random effects is approximately 0.4 (plogis(0.4) is approximately 0.6). As we increase $\sigma$, we see that our estimate of $\mu$ increases. Here, we show the log marginal likelihood surface:

```
library(emdbook)
ifun1 <- function(r, mu, sigma)
        1/(1 + exp(mu + r)) * exp(-r^2/(2 * sigma^2))
ifun2 <- function(r, mu, sigma)
        (1 - 1/(1 + exp(mu + r))) * exp(-r^2/(2 * sigma^2))

llfun <- function(mu, sigma, n1=60, n0=40, log=TRUE) {
        first <- integrate(ifun1, -200, 200, mu=mu, sigma=sigma)
        second <- integrate(ifun2, -200, 200, mu=mu, sigma=sigma)

        ll <- (n0 + n1) * log(1/sqrt(2 * pi * sigma^2)) +
                n0 * log(first[[1]]) + n1 * log(second[[1]])

        if (!log) ll <- exp(ll)

        ll
}

muvec <- seq(0.35, 1, by=0.01)
sigmavec <- exp(seq(log(1), log(10), length.out=100))

contour(muvec, sigmavec, apply2d(llfun, muvec, sigmavec),
                xlab=expression(mu),
                ylab=expression(sigma),
                main="log (marginal) likelihood surface",
                levels=c(-67.31, -67.4, -68, 70))
```




We can see that there's a relatively flat region starting from $\sigma \approx 0$ and $\mu \approx 0.5$ (see line -67.31).

## 2.2 Probit regression

Probit regression provides a natural way of defining the underlying correlation for multivariate binomial response:

$$\begin{aligned} \mathbf{y}_i &= 1(\mathbf{z}_i > 0) \\ \mathbf{z}_i &\sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{R}) \end{aligned} \tag{12}$$

In other words,

$$\begin{aligned} \mathbf{y}_i &= 1(\mathbf{z}_i > 0) \\ \mathbf{z}_i &\sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{R}) \end{aligned} \tag{13}$$

which can be rewritten as

$$\begin{aligned} \mathbf{y}_i &= 1(\mathbf{z}_i > 0) \\ \mathbf{z}_i &= \boldsymbol{\mu} + \boldsymbol{\epsilon}_i \\ \boldsymbol{\epsilon}_i &\sim \mathcal{N}(0, \mathbf{R}). \end{aligned} \tag{14}$$

This model tries to capture the correlation among the latent "residuals". Compare this expression with the mixed model approach:

$$\begin{aligned} \mathbf{y}_i &= 1(\mathbf{z}_i > 0) \\ \mathbf{z}_i &= \boldsymbol{\mu} + \mathbf{r}_i + \boldsymbol{\epsilon}_i \\ \mathbf{r}_i &\sim \mathcal{N}(0, \sigma^2 R) \\ \epsilon_{ij} &\sim \text{logistic}(0, 1) \end{aligned} \tag{15}$$

The mixed model approach seeks to decompose the latent residual into two terms. We're going to have less power to detect the correlation structure.

## 2.3 Multivariate logistic distribution

Multivariate logistic distribution suggested by O'Brien allows us to model residual correlations while preserving the marginal logistic distribution:

$$\begin{aligned} \mathbf{y}_i &= 1(\mathbf{z}_i > 0) \\ \mathbf{z}_i &= \boldsymbol{\mu} + \log\left(\frac{F(\mathbf{e}_i)}{1 - F(\mathbf{e}_i)}\right) \end{aligned} \tag{16}$$

where $\mathbf{e}_i$ comes from a multivariate distribution with mean 0 and some correlation structure and $F$ is the univariate cumulative distribution function of $\mathbf{e}_i$. Then, the resulting distribution of $\mathbf{z}_i$ also has a very similar correlation structure (see code below) as $\mathbf{e}_i$ and each $z_{ij}$ follows a logistic distribution.

```
## very similar correlations
library(mvtnorm)

## Warning:  package 'mvtnorm' was built under R version 3.5.2
##
## Attaching package:  'mvtnorm'
## The following object is masked from 'package:emdbook':
##
##     dmvnorm

corr <- matrix(
        c(1, .1, .3,
          .1, 1, -.3,
          .3, -.3, 1),
        3, 3
)
```

```r
set.seed(101)
rr <- rmvnorm(10000, sigma=corr)
cor(rr)
```

```
##             [,1]         [,2]       [,3]
## [1,] 1.00000000  0.09329252  0.2929604
## [2,] 0.09329252  1.00000000 -0.3104384
## [3,] 0.29296044 -0.31043837  1.0000000
```

```r
cor(log(pnorm(rr)/(1 - pnorm(rr))))
```

```
##           [,1]       [,2]       [,3]
## [1,] 1.0000000  0.0926942  0.2898303
## [2,] 0.0926942  1.0000000 -0.3088695
## [3,] 0.2898303 -0.3088695  1.0000000
```

Regardless of what model one decides to use, there should be a way to convert the estimate into a single, consistent scale...? Predict probability from posterior and convert that into odds ratio...? Not sure yet.