Likelihood Theory and Extensions (BIOS:7110)
Breheny

<div align="center">

Assignment 11
Due: Monday, December 2

</div>

1. *Conditional score.* As we derived in class, given two independent binomial distributions $X \sim$ Binom(n_1, π_1) and $Y \sim$ Binom(n_2, π_2), the conditional likelihood given $T = X + Y$ is

$$L(\theta) = \frac{\binom{n_1}{x}\binom{n_2}{t-x}\theta^x}{\sum_{s=0}^{t}\binom{n_1}{s}\binom{n_2}{t-s}\theta^s},$$

where

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

denotes the odds ratio. Show that the score test statistic of $H_0 : \theta = 1$ is

$$z = \frac{x - \mu}{\sigma},$$

where μ and σ denote the mean and standard deviation of the (n_1, n_2, t) hypergeometric distribution, and $z \stackrel{\cdot}{\sim} N(0, 1)$ under the null. Note: *Vandermonde's identity* states that

$$\binom{m+n}{r} = \sum_{k=0}^{r} \binom{m}{k}\binom{n}{r-k}.$$

2. *Conditional logistic regression.* Suppose we have pairs of binary outcomes Y_{i1}, Y_{i2} and we wish to fit the probability model

$$\log\frac{\pi_{ij}}{1-\pi_{ij}} = \alpha_i + \mathbf{x}_{ij}^\top\boldsymbol{\beta},$$

where $\pi_{ij} = \mathbb{P}(Y_{ij} = 1)$; in other words, each pair has its own intercept, but we assume a common effect of the covariates $\mathbf{x}$. Given $Y_{i1} + Y_{i2} = 1$ (other pairs contain no information about $\boldsymbol{\beta}$), the conditional likelihood for the ith pair (this is a slight extension of what we went over in class) is

$$L_i(\beta) = \frac{\exp(\mathbf{v}_i^\top\boldsymbol{\beta})}{\sum_{\mathbf{u}_i:\mathbf{w}_i=1}\exp(\mathbf{u}_i^\top\boldsymbol{\beta})},$$

where $\mathbf{v}_i = \sum_j y_{ij}\mathbf{x}_{ij}$, $\mathbf{u}_i = \sum_j Y_{ij}\mathbf{x}_{ij}$, and $w_i = \sum_j Y_{ij}$. Here, I am using Y_{ij} to denote potential values and y_{ij} to denote the observed values.

   (a) Show that the conditional likelihood for a single pair can be written

$$L_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}},$$

   where $\eta_i = \Delta_i^\top\boldsymbol{\beta}$, with Δ_i denoting the difference in covariate values between the observation with $Y = 1$ and the observation with $Y = 0$.

   (b) Derive the score vector. Hint: Use the chain rule, first taking $\partial\ell/\partial\boldsymbol{\eta}$, then $\partial\boldsymbol{\eta}/\partial\boldsymbol{\beta}$.

   (c) Derive the information matrix and show that it can be written $\widetilde{\mathbf{X}}^\top\mathbf{W}\widetilde{\mathbf{X}}$, where $\widetilde{\mathbf{X}}$ and $\mathbf{W}$ are matrices that you must derive.

   (d) Suppose we have the following (simulated) data:

<div align="center">

1

</div>

```
n <- 400
a <- rnorm(n)
X <- array(rnorm(n*2*3), dim=c(n, 2, 3))
b <- c(1, 0.5, 0)
Y <- cbind(rbinom(n, 1, binomial()$linkinv(a + X[,1,] %*% b)),
           rbinom(n, 1, binomial()$linkinv(a + X[,2,] %*% b)))
```

Note that `Y` is an $n \times 2$ matrix of paired observations, while `X` is a $n \times 2 \times d$ array of covariate values, where d is the number of covariates. Write a function, `paired_logistic()`, that fits the model, returning the coefficients as well as the information matrix (anything else you want to return is optional). The function should return this list as an S3 object of class `paired_logistic`; i.e., the final line of the function should look like:

```
return(structure(list(beta=beta, Info=Info), class='paired_logistic'))
```

The reason for this will become apparent in the next part of the problem.

(e) Write a function, `summary.paired_logistic()` that accepts the output of the function from part (d) and carries out Wald tests. In other words, you should be able to run the code

```
fit <- paired_logistic(X, Y)
summary(fit)
```

with `X` and `Y` organized as in part (d), and see a summary table like one typically sees when fitting models in R, with columns for the coefficient estimates, standard errors, test statistics, and p-values.