# Lecture 9 Handout

Elizabeth Colantuoni

4/26/2021

## I. Objectives:

Upon completion of this session, you will be able to do the following:

- Check consistency of observed data with assumptions for a logistic regression model
- Compare ordinary and conditional logistic regression models
- Understand and explain the difference between marginal (population-average) and conditional (subject-specific) logistic regression coefficients
- Understand, explain and use conditional logistic regression with application to case-control or longitudinal studies

## II. Logistic Regression Model Assumptions

Recall, the key assumptions from the model by order of importance:

1. We **assume** that $logit[Pr(Y = 1|X)]$ is given by $X_{n\times(p+1)}\beta_{(p+1)\times 1}$. There can be violations of this assumption including missing predictors, wrong functional form (e.g. linear vs. non-linear functions), missing interactions and errors in predictors. **Violations of this assumption affect/bias** $\beta$
2. $Y_i$ and $Y_j$ are independent of each other. There can be violations of this assumption if the data is generated via a clustered or longitudinal design. **Violations of this assumption can affect inference for** $\beta$
3. $Var(Y_i|X) = \mu_i(\beta)(1 - \mu_i(\beta))$ **Violations of this assumption may be addressed via weighted least squares or robust variance estimation**
4. A small fraction of data has high influence on the model fit. **Violations of this assumption can affect estimation and inference on** $\beta$

In this lecture, we will focus on violations of the independence assumption. See Lab 2 for a review of diagnostic procedures and approaches for handling the other assumptions.

## III. Two examples: a longitudinal and clustered design

In this lecture we will discuss two additional extensions to logistic regression models; both relating to "clustered" or "longitudinal" data: **marginal logistic regression models** and **conditional logistic regression models**.

We will discuss the extensions within the context of two examples:

1. A placebo-controlled trial to improve respiratory function. Example extracted from Fitzmaurice, Laird and Ware, Applied Longitudinal Analysis (2nd edition). 111 patients from 2 clinics were randomized to receive active or placebo treatment to treat respiratory illness. The response, respiratory status (1 = good, 0 = bad), was measured at baseline and four follow-up visits during treatment.
2. A matched case-control study was conducted by Mack et al. (1976) to study the effect of exogenous estrogens on the risk of endometrial cancer. It comprises 63 matched sets with one case and 4 controls per set. Controls were matched by being alive in the same community at the time of diagnosis for the case, having age within 1-year, same marital status and entering the community at roughly the same time. Controls could not have had a hysterectomy in which case they would not have been at risk of endometrial cancer. These data were made famous by the groundbreaking two volumes by Breslow and Day entitled Statistical Methods in Cancer Research. Chapters V and VI are excellent overviews of statistical methods for matched case-control studies. The scientific question is whether women who use estrogens, have a history of gall-bladder disease or hypertension were at increased risk of endometrial cancer.

In both examples, the outcome is binary, e.g. respiratory status (1 = good, 0 = bad) and case vs. control (1 = endometrial cancer patient, 0 = matched control without endometrial cancer).

Define $Y_{ij} = 0$ or 1 where $i$ defines the cluster (e.g. individual for the placebo-controlled trial and matched case/control set), $i = 1, ..., m$ and $j$ indexes the units within the cluster (e.g. time for the placebo-controlled trial and individuals for the matched case/control study), $j = 1, ..., n_i$.

# IV. Marginal logistic regression models

## A. Review of generalized linear models

Recall that generalized linear models are a class of regression models for an outcome variable whose distribution belongs to the exponential family of distributions.

To specify a generalized linear model, we are required to define:

- The distribution of $Y$, which includes a definition of $\mu = E(Y)$ and $Var(Y)$. There may be natural bounds on $\mu$.
- A linear model: $g(\mu) = X_i^{\prime} beta$
- A link function $g(\mu)$ and inverse link function, $g^{-1}(X_i^{\prime}\beta)$ that allows us to translate to and from the linear model and the natural bounds for $\mu$.

Generalized linear models assume that observations $Y_i$ are independent and we showed that for logistic regression the score equation is given by:

$$\begin{aligned} U(\beta) &= X^{\prime}(Y - \mu(\beta)) \\ &= \left(\frac{\partial \mu}{\partial \beta}\right)^{\prime} V^{-1}(Y - \mu(\beta)) \\ &= \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta}\right)^{\prime} V_i^{-1}(Y_i - \mu_i(\beta)) \end{aligned}$$

where $\frac{\partial \mu}{\partial \beta} = VX$, $V = diag\left[\mu(\beta)(1 - \mu(\beta))\right]$, $V_i = \mu_i(\beta)(1 - \mu_i(\beta))$.

## B. Marginal generalized linear models

The **marginal** generalized linear model allows us to formulate a similar regression model but account for the correlation of observations nested within clusters. To specify a marginal generalized linear model, we provide/assume:

- $Y_{i\cdot}$ independent of $Y_{k\cdot}$ for all nested values within clusters; i.e. clusters are independent.
- The distribution of $Y_{ij}$, which includes a definition of $\mu_{ij} = E(Y_{ij}|X_{ij})$ and $Var(Y_{ij}) = f(\mu_{ij})$. There may be natural bounds on $\mu$.
- A model for $Corr(Y_{ij}, Y_{ik})$,e.g. an exchangeable model where $Corr(Y_{ij}, Y_{ik}) = \alpha$
- The above two assumptions define a variance matrix for cluster $Y_i$ data, $V_{n_i \times n_i} = V(\beta, \alpha)$. This matrix is NOT a diagonal matrix; it contains the $Var(Y_{ij})$ on the diagonal elements and off diagonal elements are $Cov(Y_{ij}, Y_{ik})$.
- A linear model: $g(\mu_{ij}) = X_{ij}^{\intercal}\beta$
- A link function $g(\mu_{ij})$ and inverse link function, $g^{-1}(X_{ij}^{\intercal}\beta)$ that allows us to translate to and from the linear model and the natural bounds for $\mu$.

With making no additional assumptions, we can utilize the generalized estimating equations (GEE) approach to estimate and make inference for $\beta$. Namely, we solve for $\beta$ using the following estimating equation (i.e. score equation):

$$\sum_{i=1}^{m} \left[\frac{\partial \mu_i}{\partial \beta}\right]^{\intercal} V_i^{-1}(Y_i - \mu_i(\beta)) = 0$$

NOTE: This looks like the score equation for generalized linear models but there is a different $V$.

## C. Interpretation

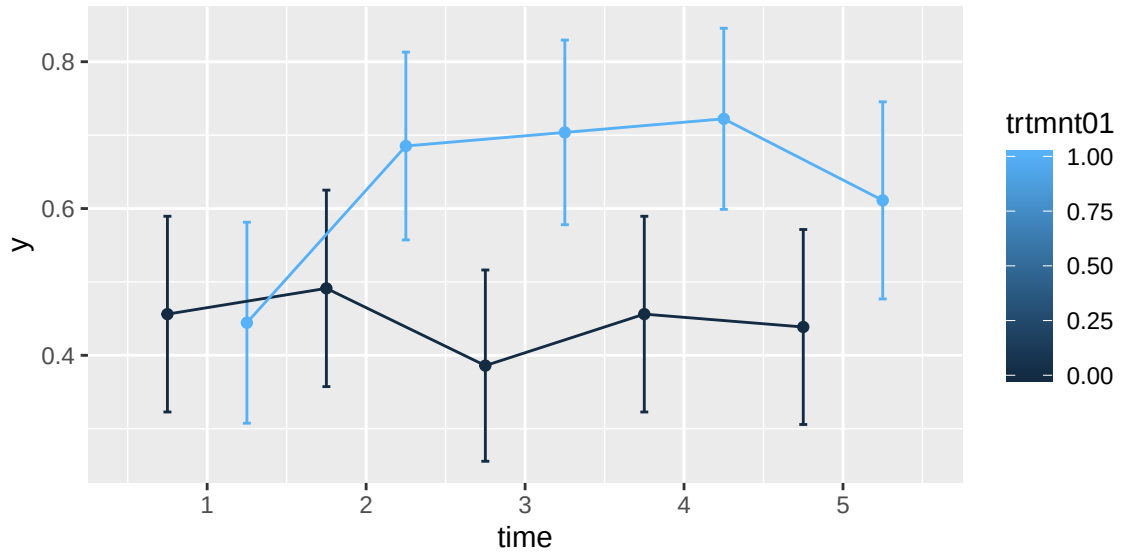
The coefficients from marginal models have "population-average" or "population-level" interpretations. That is, the goal of this analysis is to compare $\mu_{ij}$ across subsets of clusters or units within clusters based on levels of exposures.

## D. Example

We will fit a marginal logistic regression model to the data generated from the placebo-controlled trial for respiratory function.

The data are $Y_{ij} = 1$ or 0 defining a good or bad respiratory response, respectively, for patient $i$ at assessment $j$. Assessments occurred at baseline (prior to randomization, $j = 0$) and then at four follow-up assessments ($j = 1, 2, 3, 4$). The primary covariates are time (i.e. $j$) and treatment (trtmnt01, 1 = active, 0 = placebo).

```
##           1         2         3         4         5
## 0 0.4561404 0.4912281 0.3859649 0.4561404 0.4385965
## 1 0.4444444 0.6851852 0.7037037 0.7222222 0.6111111
```




Based on the exploratory analysis, we propose the following model: Patients are independent and $Y_{ij} \sim Bernoulli(\mu_{ij})$ implying that $Var(Y_{ij}) = \mu_{ij}(1-\mu_{ij})$ and

$$logit[Pr(Y_{ij}=1|post_{ij}, trtmnt01_i)] = \beta_0 + \beta_1 post_{ij} + \beta_2 post_{ij} \times trtmnt01_i$$

where $post_{ij} = I(time_{ij} > 0)$.

Interpretation of the coefficients:

- $\beta_0$: log odds of a good respiratory response at baseline
- $\beta_1$: log odds ratio of a good respiratory response comparing follow-up to baseline among patients receiving the placebo
- $\beta_1 + \beta_2$: log odds ratio of a good respiratory response comparing follow-up to baseline among patients receiving the active treatment
- $\beta_2$: treatment effect! Does the relative improvement in the odds of a good response comparing follow-up to baseline differ for the patients receiving active treatment vs. placebo

We aren't done yet! We need to make an assumption about the within patient correlation!

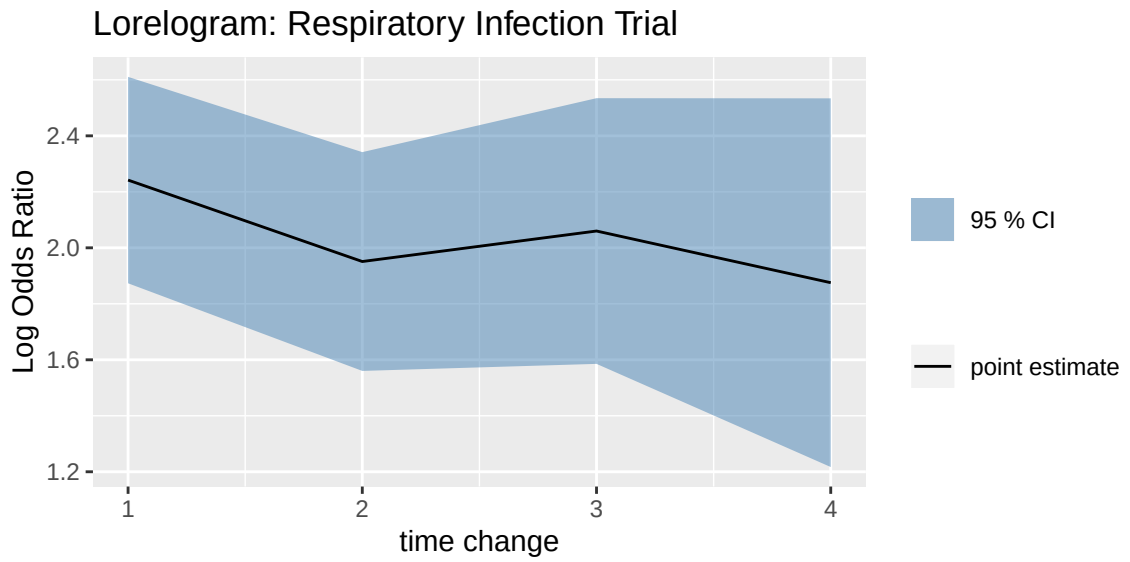
One way to measure association between two binary responses is to compute a paired odds ratio:

$$OR(Y_{ij}, Y_{ik}) = \frac{Pr(Y_{ij}=1, Y_{ik}=1)Pr(Y_{ij}=0, Y_{ik}=0))}{Pr(Y_{ij}=1, Y_{ik}=0)Pr(Y_{ij}=0, Y_{ik}=1))}$$

This can be computed for each pairwise combination of $j$ and $k$ or as a function of the lag, i.e. $|j-k|$.

For more details, interested students should read "Lorelogram: A Regression Approach to Exploring Dependence in LongitudinalCategorical Responses" by Patrick J. Heagerty and Scott L. Zeger (Journal of the American Statistical Association, 1998, Vol. 93, No. 441, pp. 150-162.).

```
# Load the lorelogram function, which was downloaded from
#"https://raw.githubusercontent.com/nstrayer/nviz/master/R/lorelogram.R"
source("lorelogram.R")
lorelogram(data$id,data$time,data$r,title="Lorelogram: Respiratory Infection Trial")
```




Using the lorelogram, it is reasonable to assume an exchangeable correlation structure (you could also try AR1)

```
data$post = ifelse(data$time>1,1,0)
data$postXtrt = data$post * data$trtmnt01
fit.exch = gee(r~post+post:trtmnt01,data=data,
        family="binomial",corstr="exchangeable",id=id)
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27

## running glm to get initial regression estimate

##   (Intercept)          post post:trtmnt01
##   -0.19885086   -0.03021571    0.98539265
```

```
summary(fit.exch)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Logit
##  Variance to Mean Relation: Binomial
##  Correlation Structure:     Exchangeable
##
## Call:
## gee(formula = r ~ post + post:trtmnt01, id = id, data = data,
##     family = "binomial", corstr = "exchangeable")
##
## Summary of Residuals:
##        Min         1Q     Median         3Q        Max
## -0.6831803 -0.4403291  0.3168197  0.5495495  0.5596709
```

```
##
##
## Coefficients:
##                  Estimate Naive S.E.    Naive z Robust S.E.   Robust z
## (Intercept)   -0.19885086  0.1915041 -1.0383635   0.1907707 -1.0423556
## post          -0.04097561  0.1943549 -0.2108288   0.2103911 -0.1947592
## post:trtmnt01  1.00825259  0.2457427  4.1028787   0.2624356  3.8419053
##
## Estimated Scale Parameter:  1.007704
## Number of Iterations:  2
##
## Working Correlation
##           [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] 1.0000000 0.4673692 0.4673692 0.4673692 0.4673692
## [2,] 0.4673692 1.0000000 0.4673692 0.4673692 0.4673692
## [3,] 0.4673692 0.4673692 1.0000000 0.4673692 0.4673692
## [4,] 0.4673692 0.4673692 0.4673692 1.0000000 0.4673692
## [5,] 0.4673692 0.4673692 0.4673692 0.4673692 1.0000000
```

```
fit.ind = gee(r~post+postXtrt,data=data,
              family="binomial",corstr="independence",
              id=id)
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27

## running glm to get initial regression estimate

## (Intercept)        post    postXtrt
## -0.19885086 -0.03021571  0.98539265
```

```
summary(fit.ind)$coefficients
```

```
##                Estimate Naive S.E.   Naive z Robust S.E.   Robust z
## (Intercept) -0.19885086  0.1912884 -1.039535   0.1907707 -1.0423556
## post        -0.03021571  0.2333728 -0.129474   0.2242977 -0.1347126
## postXtrt     0.98539265  0.1981989  4.971735   0.3113723  3.1646767
```

Compare the coefficients and estimates:

```
##                Marg Marg LL Marg UL MargR LL MargR UL   Ind Ind LL Ind UL
## (Intercept)   0.820   0.559   1.202    0.560    1.200 0.820  0.559  1.202
## post          0.960   0.651   1.416    0.630    1.462 0.970  0.608  1.547
## post:trtmnt01 2.741   1.677   4.481    1.622    4.633 2.679  1.802  3.982
##               IndR LL IndR UL
## (Intercept)     0.560   1.200
## post            0.620   1.519
## post:trtmnt01   1.437   4.994
```

# V. Random effects models

## A. Model definition

As an alternative to the marginal model described in Section IV, we could consider a random effects or conditional model.

In conditional models, we define a cluster specific mean $\mu_{ij}^c = E(Y_{ij}|b_i, X_{ij})$, where $b_i$ is a random effect that allows us to link/correlate observations nested within a given cluster.

We define the random effects logistic model as:

$$logit[\mu_{ij}^c] = X_{ij}^{'}\beta^c + Z_{ij}^{'}b_i$$

where $Z_{ij} \in X_{ij}$, $b_i \sim MVN(0, D)$; $b_i$ independent of $X_{ij}$; $Y_{ij} \perp Y_{ik}$ given $b_i$.

## B. Interpretation

Take the placebo-controlled trial for respiratory function and the simplest random effects model, i.e. a random intercept.

$$\begin{aligned} logit[Pr(Y_{ij} = 1|post_{ij}, trtmnt01_i, b_i)] &= \beta_{0i}^c + \beta_1^c I(post_{ij} > 0) + \beta_2^c I(post_{ij} > 0) trtmnt01_i \\ &= \beta_0^c + b_i + \beta_1^c I(post_{ij} > 0) + \beta_2^c I(post_{ij} > 0) trtmnt01_i \end{aligned}$$

where $b_i \sim N(0, \sigma^2)$ and the covariates are independent of $b_i$.

Interpretation:

- $\beta_{0i}^c$: defines a patient specific log-odds of a good respiratory response at baseline
- $\beta_{0i}^c = \beta_0^c + b_i$, where $b_i \sim N(0, \sigma^2)$: $\beta_0^c$ is the log-odds of a good respiratory response for the average patient (i.e. $b_i = 0$)
- $\beta_{0i}^c = \beta_0^c + b_i$, where $b_i \sim N(0, \sigma^2)$: $b_i$ represents the deviation from this average log-odds of a good respiratory response for patient $i$
- For a given patient/time/treatment:

$$\mu_{ij}^c = \frac{exp(\beta_0^c + b_i + \beta_1^c I(post_{ij} > 0) + \beta_2^c I(post_{ij} > 0) trtmnt01_i)}{1 + exp(\beta_0^c + b_i + \beta_1^c I(post_{ij} > 0) + \beta_2^c I(post_{ij} > 0) trtmnt01_i)}$$

- $\beta_1^c$: The difference in the log-odds of a good response comparing follow-up to baseline among patients who received the placebo and whom have the same propensity of a good respiratory response!

$$\begin{aligned} \beta_1^c &= logit[Pr(Y_{ij} = 1|post_{ij} = 1, trtmnt01_i = 0, b_i)] - logit[Pr(Y_{ij} = 1|post_{ij} = 0, trtmnt01_i = 0, b_i)] \\ &= log(exp(\beta_0^c + b_i + \beta_1^c)) - log(exp(\beta_0^c + b_i)) \\ &= log\left[\frac{exp(\beta_0^c + b_i + \beta_1^c)}{exp(\beta_0^c + b_i)}\right] \end{aligned}$$

- $\beta_1^c + \beta_2^c$: The difference in the log-odds of a good response comparing follow-up to baseline among patients who received the treatment and whom have the same propensity of a good respiratory response!

## C. Example

Fit the model we described above.

```
ri.fit = glmer(r~post + postXtrt+(1|id),data=data,family="binomial",nAGQ=7)
summary(ri.fit)
```

```
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 7) [glmerMod]
##  Family: binomial  ( logit )
```

```
## Formula: r ~ post + postXtrt + (1 | id)
##    Data: data
##
##      AIC      BIC   logLik deviance df.resid
##    597.7    614.9   -294.8    589.7      551
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4981 -0.3780  0.1818  0.4003  1.8805
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  id     (Intercept) 6.491    2.548
## Number of obs: 555, groups:  id, 111
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.42120    0.36667  -1.149    0.251
## post        -0.08343    0.36828  -0.227    0.821
## postXtrt     1.94525    0.48502   4.011 6.06e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr) post
## post     -0.511
## postXtrt -0.035 -0.559
```

## D. Marginal vs. Conditional Model

Compare the marginal ($\beta$) and conditional ($\beta^c$) parameter estimates.

```
cbind(summary(fit.exch)$coeff[,1],summary(ri.fit)$coeff[,1])
```

```
##                       [,1]        [,2]
## (Intercept)   -0.19885086 -0.42120330
## post          -0.04097561 -0.08342951
## post:trtmnt01  1.00825259  1.94524720
```

Recall our conversation about assessing confounding in logistic regression models; we showed that $\beta \neq \beta^c$ when $Z$ is independent of $X$. Replace $Z$ with the random effect $b$ and you have:

$$\text{Marginal model:} \quad logit[Pr(Y_{ij}|X_{ij})] = \beta_0 + \beta_1 X_{ij}$$

$$\text{Conditional model:} \quad logit[(Pr(Y_i j|X_{ij}, b_i))] = \beta_0^c + \beta_1^c X_{ij} + b_i$$

and we know that $|\beta| \leq |\beta^c|$.

In general:

- $\beta$ = change in log population odds per unit change in $X$
- $\beta^c$ = change in cluster-specific log odds per unit change in $X$

## E. Estimation

The likelihood function for the observed data $Y$ as a function of $\beta^c$ and $D$ (the variance of the random effects) is:

$$\begin{aligned} L(y|\beta^c, D) &= \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} (\mu_{ij}^c(\beta^c, b_i))^{y_{ij}} (1 - \mu_{ij}^c(\beta^c, b_i))^{1-y_{ij}} f(d_i|D) db_i \\ &= \prod_{i=1}^{m} \int Pr(y_{i1}, ..., y_{in_i}|\beta^c, b_i) Pr(b_i|D) db_i \end{aligned}$$

It can be shown that:

$$\frac{\partial log(L(y|\beta^c, D))}{\partial \beta^c} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} X_{ij}^{'} (y_{ij} - E_{b_i|y}(\mu_{ij}^c(b_i, \beta^c)))$$

The solution requires numerical integration! Typically this is accomplished via gaussian quadrature or adaptive gaussian quadrature. Notice in the *glmer* command the option "nAGQ=7". This option specifies the number of integration points used in the numerical integration. When fitting generalized linear mixed models, you should vary this to be sure your solution has converged!

## F. Random intercept model

Lets further consider the case of the random intercept model.

$$logit[\mu_{ij}^c] = X_{ij}^{'} \beta^c + b_i$$

where $b_i \sim N(0, \sigma^2)$.

The likelihood function is:

$$L(y|\beta^c, \sigma^2) = \prod_{i=1}^{m} \int \frac{exp\left[ (\sum_{j=1}^{n_i} y_{ij} X_{ij})^{'} \beta^c + y_i^{+} b_i \right]}{\prod_{j=1}^{n_i} (1 + exp(X_{ij}^{'} \beta^c + b_i))} f(b_i|\sigma_2) db_i$$

where $y_i^{+} = \sum_{j=1}^{n_i} y_{ij}$ is sufficient for $b_i$, i.e. $Pr(y_{ij}|y_i^{+}, b_i)$ does not depend on $b_i$

### 1. Matched case-control study

Suppose you have a matched case-control study with data:

Control: $(Y_{i0} = 0, X_{i0})$

Case: $(Y_{i1} = 1, X_{i1})$

The model is:

$$Pr(Y_{ij} = 1|X_{ij}, b_i) = \frac{exp(X_{ij}^{'} \beta^c + b_i)}{1 + exp(X_{ij}^{'} \beta^c + b_i)}$$

Seek to estimate $\beta^c$ without assumptions about $b_i$.

You can express the conditional likelihood as:

$$CL(Y_i|\beta^c) = \prod_{i=1}^{m} \left[ Pr(Y_{i0} = 0|X_{i0}, y_i^+ = 1) Pr(Y_{i1} = 1|X_{i1}, y_i^+ = 1) \right]$$

To show that this works, consider:

$$\begin{aligned}
Pr(Y_{i1} = 1|X_{i1}, Y_i^+ = 1, b_i) &= \frac{Pr(Y_{i1}=1 \text{ and } Y_i^+=1|b_i)}{Pr(Y_i^+=1|b_i)} \\
&= \frac{Pr(Y_{i1}=1 \text{ and } Y_{i0}=0|b_i)}{Pr(Y_{i1}=1 \text{ and } Y_{i0}=0|b_i)+Pr(Y_{i1}=0 \text{ and } Y_{i0}=1|b_i)} \\
&= \frac{Pr(Y_{i1}=1|b_i) \times Pr(Y_{i0}=0|b_i)}{Pr(Y_{i1}=1|b_i) \times Pr(Y_{i0}=0|b_i)+Pr(Y_{i1}=0|b_i) \times Pr(Y_{i0}=1|b_i)} \\
&= \frac{\left( \frac{exp(X_{i1}\beta^c+b_i)}{1+exp(X_{i1}\beta^c+b_i)} \times \frac{1}{1+exp(X_{i0}\beta^c+b_i)} \right)}{\frac{exp(X_{i1}\beta^c+b_i)}{1+exp(X_{i1}\beta^c+b_i)} \times \frac{1}{1+exp(X_{i0}\beta^c+b_i)} + \frac{1}{1+exp(X_{i1}\beta^c+b_i)} \times \frac{exp(X_{i0}\beta^c+b_i)}{1+exp(X_{i0}\beta^c+b_i)}} \\
&= \frac{exp(X_{i1}\beta^c+b_i)}{exp(X_{i1}\beta^c+b_i)+exp(X_{i0}\beta^c+b_i)} \\
&= \frac{exp(X_{i1}\beta^c)}{exp(X_{i1}\beta^c)+exp(X_{i0}\beta^c)} \\
&\quad \text{NOTE: Divide the numerator and denominator by } exp(X_{i0}\beta^c) \\
&= \frac{exp((X_{i1}-X_{i0})\beta^c)}{1+exp((X_{i1}-X_{i0})\beta^c)} \\
Pr(Y_{i0} = 0|X_{i0}, Y_i^+ = 1, b_i) &= \frac{Pr(Y_{i0}=0 \text{ and } Y_i^+=1|b_i)}{Pr(Y_i^+=1|b_i)} \\
&= \frac{Pr(Y_{i1}=1 \text{ and } Y_{i0}=0|b_i)}{Pr(Y_{i1}=1 \text{ and } Y_{i0}=0|b_i)+Pr(Y_{i1}=0 \text{ and } Y_{i0}=1|b_i)} \\
&= Pr(Y_{i1} = 1|X_{i1}, Y_i^+ = 1, b_i)
\end{aligned}$$

Therefore, the conditional likelihood can be expressed as:

$$CL(Y|\beta^c) = \prod_{i=1}^{m} \left[ \frac{exp((X_{i1} - X_{i0})\beta^c)}{1 + exp((X_{i1} - X_{i0})\beta^c)} \right]^1$$

A marginal logistic regression of $y = (1, 1, .., )_{mx1}$ on $(X_{11} - X_{10}, X_{21} - X_{20}, ..., X_{m1} - X_{m0})$ with no intercept.

## 2. Example

Consider the matched case-control study of endometrial cancer. The scientific question is whether women who use estrogens, have a history of gall-bladder disease or hypertension were at increased risk of endometrial cancer. There was some prior belief that these risk factors may act synergistically. Use conditional logistic regression with this data set to investigate these questions.

For each model, we will use only the first control in a 1-1 design. NOTE: You will be repeating the analysis with the 1-4 design and comparing the findings in Problem Set 3.

```
dat = read.table("./endometrial.txt")
names(dat) = c("set","case","age","ageg","est","gall","hyp","obesity","nonestdrug")
dat$est = dat$est - 1
dat$gall = dat$gall - 1
```

```
dat$hyp = dat$hyp - 1
dat$obesity[dat$obesity==3] = NA
dat$obesity = dat$obesity - 1
dat$nonestdrug = dat$nonestdrug - 1
dat$firstctrl = unlist(tapply(dat$set,dat$set,FUN=function(x) c(0,1,rep(0,length(x)-2))))

tapply(dat$est,dat$case,mean)
```

```
##         0         1
## 0.5039683 0.8888889
```

```
tapply(dat$gall,dat$case,mean)
```

```
##         0         1
## 0.0952381 0.2698413
```

```
tapply(dat$hyp,dat$case,mean)
```

```
##         0         1
## 0.3253968 0.4126984
```

```
library(survival)
```

```
## Warning: package 'survival' was built under R version 3.6.3
```

```
## Fit the conditional logistic model with
## all three exposures using only 1st control
fit1=clogit(case~est+gall+hyp+ strata(set), data=subset(dat,case==1|firstctrl==1))

## Drop hypertension from the model
fit1=clogit(case~est+gall+strata(set),
            data=subset(dat,case==1|firstctrl==1))

## Add the interactions
fit1.int=clogit(case~est*gall+strata(set),
            data=subset(dat,case==1|firstctrl==1))
coeff.sum = sum(fit1.int$coefficients)
var.sum = t(c(1,1,1)) %*% vcov(fit1.int) %*% c(1,1,1)
exp(coeff.sum)
```

```
## [1] 16.81038
```

```
exp(coeff.sum-1.96*sqrt(var.sum))
```

```
##          [,1]
## [1,] 2.855665
```

```
exp(coeff.sum+1.96*sqrt(var.sum))
```

```
##          [,1]
## [1,] 98.95735
```

In summary, both estrogen use and history of gall bladder disease were found to increase the risk of endometrial cancer. Furthermore, these risk factors were found to be non-additive. That is, on the log odds scale, the risk associated with having both risk factors is only marginally greater than the risk associated with having a single risk factor. However, on the odds scale this translates to a substantive increase in risk. One way to interpret the findings is below.

- The estimated odds of being a case for subjects with only estrogren use are 14.5 (95% CI: 3.1 to 71.4)

times the odds of being a case for subjects with neither estrogen use or history of gallbladder disease.

- The estimated odds of being a case for subjects with only a history of gall bladder disease are 9.9 (95% CI: 0.95 to 104.8) times the odds of being a case for subjects with neither estrogen use or history of gallbladder disease.
- Finally, the estimated odds of being a case for subjects with both estrogen use and gall bladder disease are 16.8 (95% CI: 2.9 to 99.0) times the odds of being a case for subjects with neither estogren use or history of gallbladder disease. This is approximately double the odds ratio from either risk factor alone.