Spring 2014
Stat 8004: Statistical Methods II
Lecture 9

1 Review of Stratification

In last lecture, we discussed stratification for observational studies. When the data is not generated by randomization, potential imbalance in the data might cause problems in analysis.

Example: New operation on knee injury.

Direct Injuries

	Response		
Exposure	Success	Partially Success	
New	40	30	70
Old	15	15	30
	55	45	100

Twist Injuries

	Response		
Exposure	Success	Partially Success	
New	15	5	20
Old	55	25	80
	70	30	100

Table 1: Stratiefied Data of Knee Injuries and Operations

To adjust for the imbalances in the data, we should use stratified analysis. Suppose there is a common OR in both strata, to test whether the common $OR = 1$, we can use

- Mantel-Haenszel test
- Cochran's test

All Injuries			
	Response		
Exposure	Success	Partially Success	
New	55	35	90
Old	70	40	110
	125	75	200

Table 2: Marginal Data of Knee Injuries and Operations

Key idea: Suppose $a_j \sim N(\mathbb{E}(a_j), \mathrm{Var}(a_j))$. Then $\sum_{j=1}^{J} a_j \sim N(\sum_j \mathbb{E}(a_j), \sum_j \mathrm{Var}(a_j))$ due the independence between J strata.

And to estimate the corresponding common OR, we can use Cochran-Mantel-Haenszel estimators. We also discussed how to get the confidence intervals.

1.1 Nature of Covariate Adjustment

1.1.1 Confounding and Effect Modification

In the knee injury example, the stratification variable is the type of injuries (direct injury and twist injury). Within each stratum, the new operation has a sligtly higher succefully rate. But if you simply combine the data and do not adjust for the imbanlances in the number of subjects, you will find that the old treatment has a slightly higher successful rate. This phenomenon is called the *Simpson's paradox*.

Confoundings:

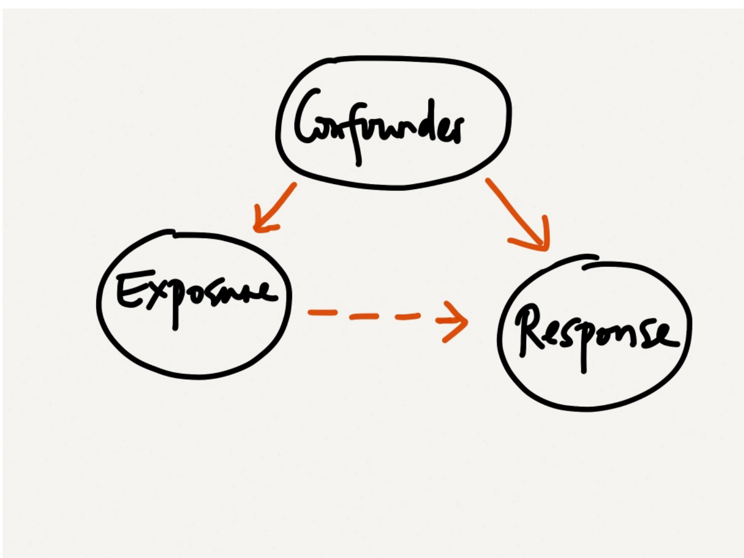

Figure 1: Confounding

Effect Modifications: OR_j are different across stratum. Then we say stratification variable modifies or interacts with the outcome.

2 Logistic Model and Interpretation

2.1 Example 1

Thalidomide is a tranquilizer that was prescribed in the late 1950s and early 1960s to pregnant women, with the devastating result of over 12,000 birth defects in 48 countries, before it was banned in 1962 (it was never sold in the United States). Recently, the drug has reappeared as a possible solution to a very different medical problem. The National Institutes of Health announced on October 31, 1995, the results of 30 hospital study of the effectiveness of thalidomide in healing mouth ulcers in AIDS patients. In the study, which was chaired by Dr. Jeffrey Jacobson of the Bronx Veteran Affairs Medical Center and the Mount SInai School of Medicine in New York, it was found that 14 of 23 patients who received thalidomide had their ulcers heal, compared with 1 of 22 patients who received a placebo. The researchers would like to know whether these results suggest that thalidomide is more effective at healing mouth ulcers than the placebo.

		Ulcers Healed?		
		Yes	No	Total
Exposure	Drug	$a = 14$	$b = 9$	$n_1 = 23$
	Placebo	$c = 1$	$d = 21$	$n_2 = 22$
	Total	$m_1 = 15$	$m_2 = 30$	$N = 46$

Table 3: Contingency Table for Ulcers Heal

Define

- ϖ_1 = the probability of ulcer heal in the drug group.
- ϖ_2 = the probability of ulcer heal in the placebo group.

We can use product binomial distribution to model the data. We also know that for such data, OR is a good measure to compare risks. (Recall the nice properties of OR...)

$$\log(OR) = \log\left\{\frac{\varpi_1/(1-\varpi_1)}{\varpi_2/(1-\varpi_2)}\right\} = \log\{\varpi_1/(1-\varpi_1)\} - \log\{\varpi_2/(1-\varpi_2)\}.$$

Now define Y_i = the indicator of ulcers heal of the ith person, $i = 1, \ldots, N$.

Let π_i = the probability of ulcer heal of the ith person = $\mathbb{P}(Y_i = 1)$. If the ith person belongs to the drug group, then $\pi_i = \varpi_1$; otherwise, $\pi_i = \varpi_2$.

Let $X_i = 1$ if the ith person belongs to the drug group; and $X_i = 0$ if the ith person belongs to the placebo group.

We can reformulate the model as:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i. \tag{1}$$

How to interpretate β_0 and β_1?

$$\begin{aligned} \text{log reference odds} &: \beta_0 = \log(\varpi_2/(1-\varpi_2)) \\ \text{log odds ratio} &: \beta_1 = \log(OR) \end{aligned}$$

Model (1) is called a simple *logistic model*. By estimating the parameters, we make inference on odds ratio. We can use logistic regression to handle contingency table. In addition, logistic regression can handle more complicated data structures. We will discuss examples later.

What's the similarities and differences between simple linear regression and simple linear logistic regression?

Simple linear regression:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \tag{2}$$

where $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

In simple linear regression,

- $\mathbb{E}(Y_i) = \beta_0 + \beta_1 X_i$
- Y_i has normal distribution.
- A separate error term is in the model.

In simple logistic regression,

- $\mathbb{E}(Y_i) = \pi_i$
- $\log(\pi_i/(1-\pi_i)) = \beta_0 + \beta_1 X_i$.
- Y_i has Bernoulli distribution $Ber(\pi_i)$.
- No separate error term, since the random outcome Y_i is not directly written in the model.

Here $\log(\pi/(1-\pi))$ is called the *link function*. It links the expectation of Y_i and the linear expression $\beta_0 + \beta_1 X_i$.

Now let's get back to Model (1). It is easy to see that

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 X_i))}.$$

$$1 - \pi_i = \frac{1}{1 + \exp(\beta_0 + \beta_1 X_i)}$$

Therefore if we can estimate β_0 and β_1, we can estimate OR and π_i.

2.2 Example 2: G-Induced Loss of Consciousness.

Military pilots sometimes black out when their brains are deprived of oxygen due to G-forces during violent maneuvers. Glaister and Miller (1990) produced similar symptoms by exposing volunteers lower bodies to negative air pressure, likewise decreasing oxygen to the brain. The data lists the subjects' ages and whether they showed syncopal blackout related signs (pallor, sweating, slow heartbeat, unconsciousness) during an 18 minute period. The investigator would like to know the relationship between the G-induced loss and age.

Variable	Description
Subject	Initials of the subject's name
Age	Subject's age in years
Signs	Whether subject showed blackout-related signs (0=No, 1=Yes)

For $i = 1, \ldots, n$ $(n = 8)$,

$$Y_i = \begin{cases} 1 & \text{if the } i\text{th person blacked out} \\ 0 & \text{otherwise} \end{cases}$$

Let $\pi_i = \mathbb{P}(Y_i = 1)$ and X_i = age of the ith person.

We can still use Model (1) to model the data. Note that here X_i is continuous rather than binary. Therefore, we cannot display the data by contingency table, but we can still use logistic regression.

2.3 Example 3: Knee Surgery (Stratified Analysis)

Consider the knee surgery example we discussed in the previous lecture (Table 4). How to use logistic regression to solve this problem?

For the ith patient $(i = 1, \ldots, n)$, let

$$Y_i = \begin{cases} 1 & \text{if the } i\text{th patient had completely successful surgery} \\ 0 & \text{otherwise} \end{cases}$$

Direct Injuries			
	Response		
Exposure	Success	Partially Success	
New	40	30	70
Old	15	15	30
	55	45	100

Twist Injuries			
	Response		
Exposure	Success	Partially Success	
New	15	5	20
Old	55	25	80
	70	30	100

Table 4: Stratified Data of Knee Injuries and Operations

Let

$$x_i = \begin{cases} 1 & \text{if the } i\text{th patient had the new surgery} \\ 0 & \text{otherwise} \end{cases},$$

and

$$z_i = \begin{cases} 1 & \text{if the } i\text{th patient had direct knee injury} \\ 0 & \text{otherwise} \end{cases}.$$

Let $\pi_i = \mathbb{E}(Y_i)$. Then we can set up the logistic regression:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_1 x_i + \beta_2 z_i.$$

Note that there are only 4 values the combinations of (x_i, z_i) can take, $(0,0)$, $(1,0)$, $(0,1)$, and $(1,1)$. If we use the notation for the contigency table, we would find that

$$\pi_{00} = \frac{\exp(\alpha)}{1+\exp(\alpha)}, \quad \pi_{10} = \frac{\exp(\alpha+\beta_1)}{1+\exp(\alpha+\beta_1)}$$
$$\pi_{01} = \frac{\exp(\alpha+\beta_2)}{1+\exp(\alpha+\beta_2)}, \quad \pi_{11} = \frac{\exp(\alpha+\beta_1+\beta_2)}{1+\exp(\alpha+\beta_1+\beta_2)},$$

where π_{ij} stands for the complete successful rate in the group where the treatment variable $= i$ and the injury group $= j$.

The Odds ratio in the jth injury group is

$$\frac{\pi_{1j}/(1-\pi_{1j})}{\pi_{0j}/(1-\pi_{0j})} = \exp(\beta_1).$$

So this model assumes two stratification group has the same odds ratio for treatment effect. To assume different treatmetn effect in stratification groups, we need the logistic model with the interaction term. We will discuss in next lecture.

2.4 Example 4: Passengers on the Titanic

The data give the survival status of passengers on the Titanic, together with their names, age, sex and passenger class. About half of the ages for the 3rd Class passengers are missing, although a good many of these could be filled in from the original source below.

Variable	Description
Name	Recorded name of passenger
PClass	Passenger class: 1st, 2nd or 3rd
Age	Age in years
Sex	male or female
Survived	1 = Yes, 0 = No

Suppose there are in total n passengers. For $i = 1, \ldots, n$ $(n = 8)$,

$$Y_i = \begin{cases} 1 & \text{if the } i\text{th passenger survived} \\ 0 & \text{otherwise} \end{cases}$$

And $\pi_i = \mathbb{P}(Y_i = 1) =$ the probability that the ith person survived.

Let

$$X_{1i} = \begin{cases} 1 & \text{if the } i\text{th passenger is at the 2nd class} \\ 0 & \text{otherwise} \end{cases}$$

and

$$X_{2i} = \begin{cases} 1 & \text{if the } i\text{th passenger is at the 3rd class} \\ 0 & \text{otherwise} \end{cases}$$

Note that X_{1i} and X_{2i} are dummy variables for PClass.

Let $X_{3i} =$ the age of the ith passenger. And

$$X_{4i} = \begin{cases} 1 & \text{if the } i\text{th passenger is male} \\ 0 & \text{otherwise} \end{cases}$$

We can build up a multivariate logistic model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i}.$$

Here, β_0 is the log odds of survival for the hypothetic reference group, which is the group of female, age equal to 0 and 1st class. β_j is the coefficient of X_j, it measures how much one unit increase in X_j may affect the log odds of survival.

Similar as multivariate linear model, we can fomulate the logistic model into a matrix form.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \tilde{\boldsymbol{x}}_i^{\mathrm{T}} \boldsymbol{\theta}, \tag{3}$$

Here $\tilde{\boldsymbol{x}}_i = (1, x_{i1}, x_{i2}, x_{i3}, x_{i4})^{\mathrm{T}}$ and $\boldsymbol{\theta} = (\alpha, \beta_1, \beta_2, \beta_3, \beta_4)^{\mathrm{T}}$.

It implies that

$$\pi_i = \frac{\exp(\tilde{\boldsymbol{x}}_i^{\mathrm{T}} \boldsymbol{\theta})}{1 + \exp(\tilde{\boldsymbol{x}}_i{}^{\mathrm{T}} \boldsymbol{\theta})} \tag{4}$$

$$1 - \pi_i = \frac{1}{1 + \exp(\tilde{\boldsymbol{x}}_i^{\mathrm{T}} \boldsymbol{\theta})} \tag{5}$$

3 Estimating Coefficients for Logistic Regression

3.1 Likelihood Function

After we build up the model, we would like to estimate the parameters. One natural choice is to write out the likelihood function of the model and get MLE of the parameters.

Recall the underlying distribution of the logistic model. $Y_i = 1$ or 0 is the response variable. We assume that $Y_i \sim Ber(\pi_i)$.

Now suppose these passengers are independent from each other, the likelihood would be

$$L(\boldsymbol{\pi}) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i}.$$

Binomial regression model:

$$\pi_i = g^{-1}(\alpha + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}),$$

where $g^{-1}(\cdot)$ is the inverse function. In logistic regression, $g(\cdot)$ is the logit and $g^{-1}(\cdot)$ is the logistic function (inverse logit). The link function $g(\cdot)$ can be any twice-differentiable one-to-one function. The Binomial regression helps to model π_i, so that we can reduce the number of parameters in the above likelihood.

Let $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta})^{\mathrm{T}} = (\alpha, \beta_1, \ldots, \beta_p)$. Then the likelihood function is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \{g^{-1}(\alpha + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})\}^{y_i} \{1 - g^{-1}(\alpha + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})\}^{1-y_i}.$$

For logistic regression, we can get the likelihood function of α and $\boldsymbol{\beta}$:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \frac{\exp(\alpha + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})}{1 + \exp(\alpha + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})} \right\}^{y_i} \left\{ \frac{1}{1 + \exp(\alpha + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})} \right\}^{1-y_i} \tag{6}$$

Likewise, the log likelihood in the probilities is

$$l(\boldsymbol{\pi}) = \sum_{i=1}^{n} y_i \log(\pi_i) + \sum_{i=1}^{n} (1 - y_i) \log(1 - \pi_i).$$

and that in the corresponding model parameters is

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} y_i(\alpha + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) - \sum_{i=1}^{n} \log\{1 + \exp(\alpha + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}.$$

Let the score vector

$$U(\boldsymbol{\theta}) = \partial l(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = [U(\boldsymbol{\theta})_\alpha, U(\boldsymbol{\theta})_{\beta_1}, \ldots, U(\boldsymbol{\theta})_{\beta_p}]^{\mathrm{T}}.$$

The score equation for the intercept is

$$\begin{aligned} U(\boldsymbol{\theta})_\alpha = \frac{\partial l}{\partial \alpha} = \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \frac{\exp(\alpha + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{1 + \exp(\alpha + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})} \\ = \sum_{i=1}^{n}(y_i - \pi_i) = \sum_{i=1}^{n}[y_i - \mathbb{E}\,(y_i \mid \boldsymbol{x}_i)] = m_1 - \mathbb{E}\,(m_1 \mid \boldsymbol{\theta}), \end{aligned}$$

where $\mathbb{E}\,(y_i \mid \boldsymbol{x}_i^{\mathrm{T}}) = \pi_i$. The score function for the j-th coefficient is

$$\begin{aligned} U(\boldsymbol{\theta})_{\beta_j} = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij}\left(y_i - \frac{\exp(\alpha + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{1 + \exp(\alpha + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}\right) \\ = \sum_{i=1}^{n} x_{ij}(y_i - \pi_i) = \sum_{i=1}^{n} x_{ij}\{y_i - \mathbb{E}\,(y_i \mid \boldsymbol{x}_i)\}, \end{aligned}$$

which is a weighted sum of the observed value of the response (y_i) minus that expected under the model.

How to obtain MLE? Set $U(\boldsymbol{\theta}) = 0$. Unfortunately, the close-form solutions do not exist. The solution must be obtained by an iterative procedure such as the Newton-Raphson algorithm, which will be discussed in a moment.

Let $U(\hat{\boldsymbol{\theta}})_\alpha = m_1 - \sum_i \hat{\pi}_i = 0$, where $\hat{\pi}_i = \exp(\hat{\alpha} + \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_i)/\{1 + \exp(\hat{\alpha} + \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_i)\}$. The mean estimated probability is $\hat{\bar{\pi}} = \sum_{i=1}^{n} \hat{\pi}_i/n = m_1/n$.

3.2 Hension Matrix and Observed Information

Hension matrix

$$H(\boldsymbol{\theta}) = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial^2 \boldsymbol{\theta}} = \frac{\partial U(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

Observed information matrix $i(\boldsymbol{\theta}) = -H(\boldsymbol{\theta})$.

Consider the logistic regression

$$\pi_i = g^{-1}(\alpha_i + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}).$$

Then

$$
\begin{aligned}
i(\boldsymbol{\theta})_\alpha &= \frac{-\partial U(\boldsymbol{\theta})_\alpha}{\partial \alpha} = \sum_i \pi_i(1-\pi_i) \\
i(\boldsymbol{\theta})_{\beta_j} &= \frac{-\partial U(\boldsymbol{\theta})_{\beta_j}}{\partial \beta_j} = \sum_{i=1}^n x_{ij}\frac{\partial \pi_i}{\partial \beta_j} = \sum_{i=1}^n x_{ij}\frac{\partial \pi_i}{\partial \beta_j} = \sum_{i=1}^n x_{ij}\frac{x_{ij}\exp(\alpha + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{1+\exp(\alpha + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2} \\
&= \sum_{i=1}^n x_{ij}^2 \pi_i(1-\pi_i), \\
i(\boldsymbol{\theta})_{\beta_j,\beta_k} &= \frac{-U(\boldsymbol{\theta})_{\beta_j}}{\partial \beta_k} = \sum_{i=1}^n x_{ij}\frac{\partial \pi_i}{\partial \beta_k} = \sum_{i=1}^n x_{ij}x_{ik}\pi_i(1-\pi_i), \\
i(\boldsymbol{\theta})_{\alpha,\beta_j} &= \frac{-U(\boldsymbol{\theta})(\boldsymbol{\theta})_\alpha}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial \pi_i}{\partial \beta_j} = \sum_{i=1}^n x_{ij}\pi_i(1-\pi_i).
\end{aligned}
$$

3.3 Iteratively Reweighted Least Squares

3.3.1 Newton's Method

Suppose there is a complicated univariate increasing/decreasing function $f(x)$. How to find its root $f(x_0) = 0$.

We get the derivative f'. Let's begin with a first guess x_0 for a root of the function f. A better approximation x_1 is

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

Geometrically, $(x_1, 0)$ is the intersection with the x-axis of a line tangent to f at $(x_0, f(x_0))$. The process is repeated as

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

until a sufficient accurate value is reached. We can prove that $\lim_{n\to\infty} x_n = x_0$.

Now for a concave function univariate function f. If we want to maximize it, we can get its derivative $f'(x)$ and solve $f'(x_0) = 0$. Same as before, we can start with an initial guess x_0, and iterate with

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}$$

until convergence. We can prove that $\lim_{n\to\infty} x_n = x_0$, and the point x_0 is the maximizer of $f(x)$.

Suppose for a multivariate function $f(\mathbf{x})$. To maximize it, we can set $\mathbf{U}(\mathbf{x}) = \partial \mathbf{f}(\mathbf{x})/\partial \mathbf{x} = \mathbf{0}$. we can apply the multivariate version of Newton's method to get the maximizer $\mathbf{x}_0$.

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \left(\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}}\right)^{-1}\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{x}_n - \left(\frac{\partial U(\mathbf{x})}{\partial \mathbf{x}}\right)^{-1} U(\mathbf{x}).$$

3.3.2 Iteratively Reweighted Least Squares

Let's get back to the logistic regression. We would like to maximize $l(\boldsymbol{\theta})$ and equivalently we solve $U(\boldsymbol{\theta}) = \partial l/\partial\boldsymbol{\theta} = \mathbf{0}$. We starting with a guess starting value $\hat{\boldsymbol{\theta}}^0$, and renew it with

$$\hat{\boldsymbol{\theta}}^{l+1} = \hat{\boldsymbol{\theta}}^l - \mathbf{H}(\hat{\boldsymbol{\theta}}^l)^{-1}U(\hat{\boldsymbol{\theta}}^l). \tag{7}$$

Here $\mathbf{H}(\boldsymbol{\beta}) = \partial U(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$.

To simplify notation, let $\tilde{\boldsymbol{x}}_i = (1, \tilde{x}_1, \ldots, \tilde{x}_p)$, and $\tilde{\mathbf{X}} = (\tilde{x}_1, \ldots, \tilde{x}_n)^{\mathrm{T}}$.

$$\begin{aligned}
U(\boldsymbol{\theta}) &= \tilde{\mathbf{X}}^{\mathrm{T}}(\mathbf{Y} - \boldsymbol{\pi}) \\
\mathbf{H}(\boldsymbol{\theta}) &= \frac{\partial U(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = (h_{jk})_{p\times p} \\
h_{jk} &= \frac{\partial u_j}{\partial\boldsymbol{\theta}} = -\sum_{i=1}^{n}\tilde{x}_{ij}\frac{\partial\pi_i}{\partial\theta_k} = -\sum_{i=1}^{n}\tilde{x}_{ij}\tilde{x}_{ik}\frac{\exp(\tilde{\boldsymbol{x}}_i^{\mathrm{T}}\boldsymbol{\theta})}{(1+\exp(\tilde{\boldsymbol{x}}_i^{\mathrm{T}}\boldsymbol{\theta}))^2} = \sum_{i=1}^{n}\tilde{x}_{ij}\tilde{x}_{ik}\pi_i(1-\pi_i)
\end{aligned}$$

Now define $w_i = \pi_i(1-\pi_i)$ and $\mathbf{W} = \mathrm{diag}(w_1, \ldots, w_n)$. Then $H(\boldsymbol{\beta}) = -\tilde{\mathbf{X}}^{\mathrm{T}}\mathbf{W}\tilde{\mathbf{X}}$.

Denote $\hat{\mathbf{W}} = \mathrm{diag}(\hat{w}_1, \ldots, \hat{w}_n)$ to be the estimator of $\mathbf{W}$, where $\hat{w}_i = \hat{\pi}_i(1-\hat{\pi}_i)$. Note that in the iteration procedures, $\hat{\pi}_i$ is updated in each iteration.

For $\hat{\boldsymbol{\theta}}^l$, suppose the corresponding estimator of π is $\hat{\pi}^l$ and the corresponding weight matrix is $\hat{\mathbf{W}}^l$.

Combine these results with (7), we can get

$$\hat{\boldsymbol{\theta}}^{l+1} = \hat{\boldsymbol{\theta}}^l + (\tilde{\mathbf{X}}^{\mathrm{T}}\hat{\mathbf{W}}^l\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^{\mathrm{T}}(\mathbf{Y} - \hat{\boldsymbol{\pi}}^l) \tag{8}$$

Since $\hat{\boldsymbol{\theta}}^l = (\tilde{\mathbf{X}}^{\mathrm{T}}\hat{\mathbf{W}}^l\tilde{\mathbf{X}})^{-1}(\tilde{\mathbf{X}}^{\mathrm{T}}\hat{\mathbf{W}}^l\tilde{\mathbf{X}})\hat{\boldsymbol{\theta}}^l$. We plug it in (8) and get

$$\hat{\boldsymbol{\theta}}^{l+1} = (\tilde{\mathbf{X}}^{\mathrm{T}}\hat{\mathbf{W}}^l\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^{\mathrm{T}}\hat{\mathbf{W}}^l\{\tilde{\mathbf{X}}\hat{\boldsymbol{\theta}}^l + (\hat{\mathbf{W}}^l)^{-1}(\mathbf{Y} - \hat{\boldsymbol{\pi}}^l)\} = (\tilde{\mathbf{X}}^{\mathrm{T}}\hat{\mathbf{W}}^l\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^{\mathrm{T}}\hat{\mathbf{W}}^l\hat{\mathbf{Z}}^l, \tag{9}$$

where $\hat{\mathbf{Z}}^l = \tilde{\mathbf{X}}\hat{\boldsymbol{\theta}}^l + (\hat{\mathbf{W}}^l)^{-1}(\mathbf{Y} - \hat{\boldsymbol{\pi}}^l)$.

Recall for multiple linear regression, we used least square to solve for $\boldsymbol{\beta}$.

$$\hat{\boldsymbol{\theta}} = (\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^{\mathrm{T}}\mathbf{Y}.$$

Here, the method is an iterative algorithm. We stop when convergence is reached. For each step of iteration, we can write the formula into a format of *weighted least squares*. At each step, due to the change of the estimator $\hat{\boldsymbol{\beta}}^l$, the weight matrix changes correspondingly. Therefore, at each step, we need to recalcuate the weight matrix. This is why the algorithm is called iteratively reweighted least squares.

3.4 Data Analysis

In practice, you don't need to program the iteratively reweighted least square to solve for the MLEs. It is build in in many statistical softwares, such as R and SAS.

Example 1. Ulcer Heal

		Ulcers Healed?		
		Yes	No	Total
Exposure	Drug	$a = 14$	$b = 9$	$n_1 = 23$
	Placebo	$c = 1$	$d = 21$	$n_2 = 22$
	Total	$m_1 = 15$	$m_2 = 30$	$N = 46$

Table 5: Contingency Table for Ulcers Heal

```
> fit <- glm(heal~drug, data = ulcer, family = binomial)
> summary(fit)

Call:
glm(formula = heal ~ drug, family = binomial, data = ulcer)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3699  -0.3050  -0.3050   0.9964   2.4864

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.045      1.023  -2.975  0.00293 **
drug           3.486      1.109   3.144  0.00167 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 57.286  on 44  degrees of freedom
Residual deviance: 38.925  on 43  degrees of freedom
AIC: 42.925

Number of Fisher Scoring iterations: 5
```

Note that $\hat{\beta}_1 = 2.49$. Based on the contingency table, we can also get the estimator for OR, $\widehat{OR} = 32.67$. We can see that $\log(\widehat{OR}) = 2.49 = \hat{\beta}_1$.

On the other hand $\hat{\beta}_0 = -3.04 = \log(\hat{\pi}_2)/(1 - \log(\hat{\pi}_2))$

Example 2. G-Induced Loss of Consciousness.

```
> fit <- glm(Signs~Age, data = gloss, family = binomial)
> summary(fit)

Call:
glm(formula = Signs ~ Age, family = binomial, data = gloss)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7112  -0.8686   0.5063   0.6939   1.5336

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.92086    2.64655  -1.104    0.270
Age          0.10569    0.08037   1.315    0.188

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10.5850  on 7  degrees of freedom
Residual deviance:  8.4345  on 6  degrees of freedom
AIC: 12.435

Number of Fisher Scoring iterations: 4
```

Here $\hat{\beta}_1 = 0.1 > 0$, indicating that there might be slightly larger probability to black out under the G-force loss environment for older people.

Example 3. Passengers on the Titanic.

```
> fit <- glm(Survived ~ PClass + Age + Sex, data = titanic,
 family = binomial)
> summary(fit)

Call:
glm(formula = Survived ~ PClass + Age + Sex, family = binomial
  ,
    data = titanic)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7226  -0.7065  -0.3917   0.6495   2.5289

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.759662   0.397567   9.457  < 2e-16 ***
PClass2nd   -1.291962   0.260076  -4.968 6.78e-07 ***
```

```
PClass3rd    -2.521419   0.276657   -9.114  < 2e-16 ***
Age          -0.039177   0.007616   -5.144 2.69e-07 ***
Sexmale      -2.631357   0.201505  -13.058  < 2e-16 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1025.57  on 755  degrees of freedom
Residual deviance:  695.14  on 751  degrees of freedom
AIC: 705.14

Number of Fisher Scoring iterations: 5
```

$\hat{\beta}_{1i} = -1.29$ indicaes that the passengers in the 2nd class has an decreased log odds of survival by -1.29 compared with the passengers in the 1st class. $\hat{\beta}_{2i} = -2.52$ indicaes that the passengers in the 3rd class has an increased log odds of survival by -2.52 compared with the passengers in the 1st class. Also, there are evidence that older passengers and male passengers have a lower survival rate.