STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee

UNIVERSITY OF TORONTO

January 30- February 1, 2018

STA 303/1002: Class 8-Generalized Linear Models

- Case Study III: The Donner Party Example
- Generalized Linear Models
  - What is a Generalized Linear Model?
  - Common link functions
  - What is a Binary Logistic Regression Model?
  - Maximum likelihood estimation of β's
- Case Study III Example
  - Data and Questions
  - Estimated Model
  - Interpretations

Case Study III: The Donner Party Example

Ramsey & Schafer, 3rd ed.

- Background: (D.K. Grayson, Journal of Anthropological Research, 1990: 223-42)
  - In mid 19th century, a group of 86 American pioneers headed out from Missouri toward California in a wagon train.
  - Due to a combination of harsh weather, unsuitable travel equipment and divisions with the party, the group got stuck in the Sierra Nevada mountain range.
  - They had planned to arrive safe and sound in September but those who survived did not make it there until the following March.
- Question: Who survived?- Men? -Older pioneers?
- Data:
  - age
  - sex
  - outcome: survived or not
- AIM: Study the odds of survival

Case Study III: Model

"success", $y=1$
"failure", $y=0$

- Response: Y_i- a binary variable (eg., survived or died)
- Predictor: X_i- eg., age, sex of ith pioneer
- Model: BINARY LOGISTIC REGRESSION

$$Y_i|X_i = \begin{cases} 1 & \text{if response is in category of interest} \\ 0 & \text{otherwise} \end{cases}$$

$$Y_i|X_i \sim \text{Bernoulli}(\pi_i)$$

Then:

- $E[Y_i|X_i] = \pi_i$ and $\text{Var}(Y_i|X_i) = \pi_i(1-\pi_i)$

$\varepsilon_i = y_i - \hat{y}_i$

obs

Exp.

- A logistic regression model is an example of a **Generalized Linear Model**.

Generalized Linear Models

- Have: · response, Y and
  · a set of explanatory variables $X_1, \ldots, X_p$
- Want: Model $E(Y)$ as a linear function in the parameters, ie.,

$$g(E(Y)) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p = \mathbf{X}\beta$$

- Key idea: Choice of the link function, g such that

$$g(E(Y)) = \mathbf{X}\beta$$

$E(g(y))$ ← transform y

Some Link Functions

Let $E(Y) = \mu$.

General LM →

Link	Function	Usual distribution of $Y \mid X$
Identity	$g(\mu) = \mu$	Normal
Log	$g(\mu) = \log \mu, \mu > 0$	Poisson (count data)
Logit	$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right), 0 < \mu < 1$	Bernoulli (binary), Binomial

$\epsilon_i \sim N(0, \sigma^2 I)$

Note: Link function, $g(\cdot)$ is a function of $\mu = E(Y)$, the mean of Y, and not a transformation of the data.

$\log(E(y))$ $\quad$ $E \log(y)$

$g[E(y)]$ $\quad$ $E[g(y)]$

GLMs vs Transforming the data

- Transform Y so it has an approximate normal distribution with constant variance. Common variance stabilizing transformations (Weisberg, 3rd ed, p. 179):
  - $\sqrt{Y}$: mild transformation; used when $Var(Y|X) \propto E(Y|X)$ as for Poisson data
  - $\log(Y)$: most common; if $Var(Y|X) \propto [E(Y|X)]^2$ or errors behave like percentage of Y.
  - $1/Y$: used when responses are mostly close to 0, but some large values occur.
- As GLM (Agresti, p. 117):
  - distribution of Y not restricted to Normal
  - model parameters describe g[E(Y)] rather than E(g(Y)) as in transformed data approach
  - GLMs provide a unified theory of modelling that encompasses the most important models for continuous and discrete variables.

LOG ODDS, ODDS, ODDS RATIO

1

- Let $\pi = P(\text{“}\mathit{success}\text{”})$, $0 < \pi < 1$.
- The ODDS in favour of “success” is: $(0, \infty)$

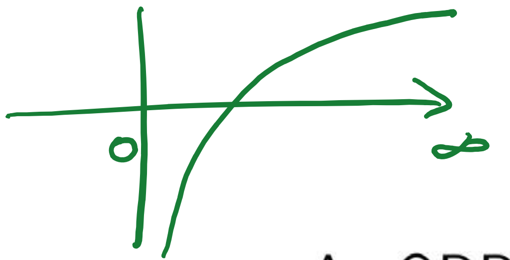
$$\frac{\pi}{1-\pi}$$

$$\frac{P(\text{“success”})}{P(\text{“failure”})}.$$

As $\pi \downarrow 0$, ODDS $\downarrow 0$

As $\pi \uparrow 1$, ODDS $\to \infty$

- Then the LOG ODDS is: $(-\infty, \infty)$



$$\log\left(\frac{\pi}{1-\pi}\right)$$

- An ODDS RATIO is a ratio of ODDS. $(0, \infty)$.

Binary Logistic Regression

- $E(Y|X) = \pi$
- $\mathrm{Var}(Y|X) = \pi(1-\pi)$. *Notice that variance is not constant!*
- Logistic regression model:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \qquad (1)$$

- Linear predictor:

$$\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- LOGISTIC FUNCTION: Find by inverting equation (1)

$$\underline{\pi(\eta)} = \frac{e^{\eta}}{1+e^{\eta}}$$

$$\log \frac{\pi}{1-\pi} = \eta$$

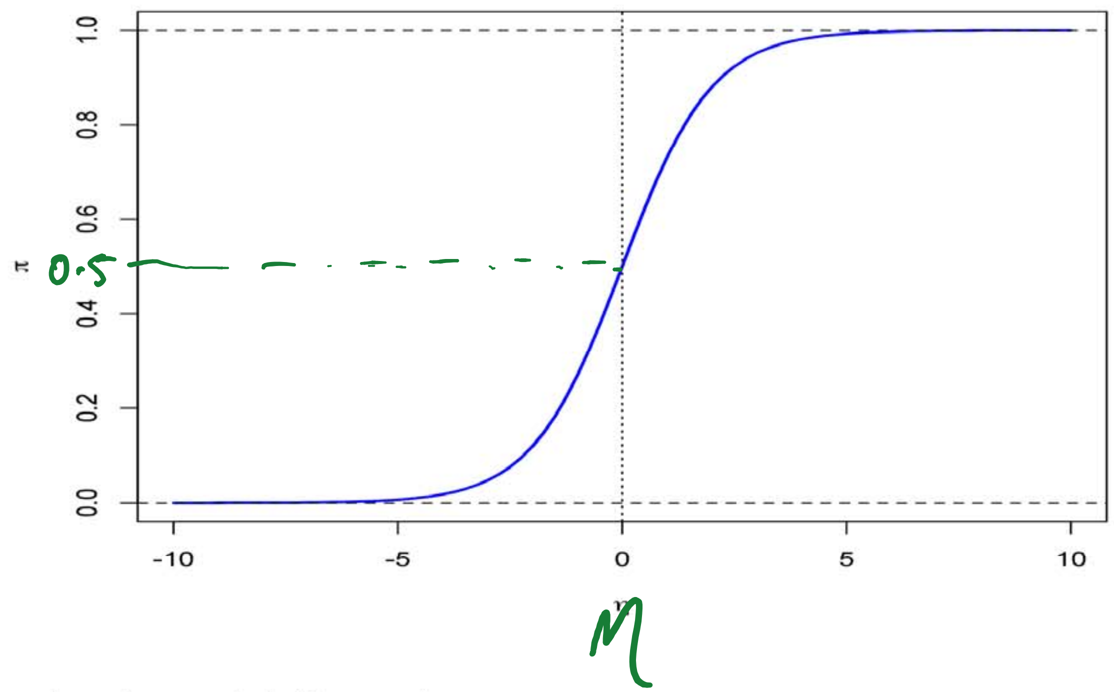
$$\frac{\pi}{1-\pi} = e^{\eta}$$

$$\pi = e^{\eta} - \pi e^{\eta}$$

$$\pi(1+e^{\eta}) = e^{\eta}$$

What does the logistic function look like?

- LOGISTIC FUNCTION: $\pi = \frac{e^{\eta}}{1+e^{\eta}}$



- S-shaped; sigmoid function
- Horizontal asymptotes at 0 and 1; the logistic function, $\pi(\eta)$ varies between 0 and 1

Binary Logistic Regression Model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}, \quad i = 1, \ldots, n$$

- Log-odds, $\log(\pi/(1-\pi))$ are between $-\infty$ and ∞ (good characteristic of a link function)
- As π_i (the probability of "success") increases, odds of success and log-odds increase
- Predicts the natural log of the odds for a subject being in one category or another
- Regression coefficients can be used to estimate odds ratio for each of the independent variables
- Tells which predictors can be used to determine if a subject was in a category of interest

How to estimate the parameter coefficients?

Maximum Likelihood Estimation

- Data: $Y_i = \begin{cases} 1 & \text{if response is in category of interest} \\ 0 & \text{otherwise} \end{cases}$
- Model: $P(Y_i = y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$
- Assume: The $\underline{\underline{n}}$ observations are independent
- Joint density:

$$P(Y_1 = y_1, \ldots, Y_n = y_n) = \Pi_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} = \pi_1^{y_1}(1-\pi_1)^{1-y_1} * \ldots * \pi_n^{y_n}(1-\pi_n)^{1-y_n}$$

where

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip})} = \frac{e^{M_i}}{1+e^{M_i}}$$

and $$1 - \pi_i = 1 - \frac{e^{M_i}}{1+e^{M_i}} = \frac{1+e^{M_i} - e^{M_i}}{1+e^{M_i}} = \frac{1}{1+e^{M_i}}$$

Maximum Likelihood Estimation

- Likelihood function: Plug in observed data and think of the joint density as a function of β's-

$$\mathcal{L}(\beta_0, \ldots, \beta_p) = \Pi_{i=1}^{n} \pi_i(\boldsymbol{\beta})^{y_i} (1 - \pi_i(\boldsymbol{\beta}))^{1-y_i}$$

$$= \prod_{i=1}^{n} \left(\frac{e^{\mu_i}}{1+e^{\mu_i}}\right)^{y_i} \left(\frac{1}{1+e^{\mu_i}}\right)^{1-y_i}$$

- Log-likelihood function:

$$\begin{aligned}\log \mathcal{L}(\beta_0, \ldots, \beta_p) = \sum_{i=1}^{n} [&y_i(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}) \\ &- y_i \log(1 + \exp(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip})) \\ &- (1 - y_i) \log(1 + \exp(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}))]\end{aligned}$$

- Maximize the log-likelihood:

$$(\widehat{\beta}_0, \ldots, \widehat{\beta}_p) = \arg\max \{\log \mathcal{L}(\beta_0, \ldots, \beta_p)\}$$

$\log AB = \log A + \log B.$

$\log c^D = D \log c$

$\log \frac{1}{a} = -\log a$

$\log a^{-1} = -\log a$

MLE solution methods

- No explicit expression exists for the maximum likelihood estimators -$(\widehat{\beta}_0, \ldots, \widehat{\beta}_{\boldsymbol{p}})$.
- Two iterative numerical solution methods are:
  (1) Newton-Raphson algorithm
  (2) Fisher scoring or Iteratively Re-weighted Least Squares (IWLS). This is done in `glm()`.

Large-sample properties of MLEs

If model is correct, and sample size is large enough, as $n \to \infty$

1. MLEs are unbiased
2. MLEs have minimum variance
3. MLEs are Normally distributed — $\hat{\beta}_{MLE} \pm z_{\alpha/2}\, se(\hat{\beta}_{MLE})$
4. Formulas for standard errors of MLEs are well-known. Estimates of standard errors are available as by-product of numerical optimization (maximization) procedures.

Case Study III: The Donner Party Example

Case Study III: The Data

- Data: n=45 pioneers

AGE	SEX	STATUS
23	MALE	DIED
40	FEMALE	SURVIVED
40	MALE	SURVIVED
30	MALE	DIED
28	MALE	DIED
40	MALE	DIED
	...	

  - AGE: Adults, 15-65 yrs old
  - SEX: 15 Females, 30 Males
  - BINARY OUTCOME: 25 Died, 20 Survived
- Questions: What are the odds of survival for a 20-yr old female? Compare the odds of survival to that of a male of the same age.

Case Study III: Binary Logistic Regression Additive Model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 Age_{i1} + \beta_2 Sex_{i2}, \quad i = 1, \ldots, 45$$

- Cannot predict survival ($\pi = 1$) or death ($\pi = 0$) of a pioneer
- Can estimate:
  - π_i (the probability of survival)
  - odds of survival and
  - log-odds of survival based on *Age* and *Sex* of a pioneer
- Can be used to get point and interval estimates of odds ratios
- Can test which predictors are relevant to determine odds of survival

Interpreting coefficients of a Binary Logistic model

For $\pi = P(Y = 1)$, we model

$$\log\left(\frac{\pi}{1-\pi}\right) = \log \text{odds}_{\{Y=1\}} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Let ω be the odds that $Y{=}1$ based on $X_1, \ldots, X_p$, then

$$\omega = \exp\{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\}.$$

Interpretation of β_1: Holding $X_2, \ldots, X_p$ fixed, the ratio of the odds ('ODDS RATIO') that $Y{=}1$ at $X_1{=}a$ to $X_1{=}b$ is

$$\frac{\omega_a}{\omega_b} = \exp\{\beta_1(a-b)\}.$$

$$\frac{\omega_a = e^{\beta_0+\beta_1 a+\beta_2 X_2+\cdots+\beta_p X_p}}{\omega_b = e^{\beta_0+\beta_1 b+\beta_2 x_2+\cdots+\beta_p X_p}}$$

If X_1 increases by 1 unit, holding all other X's constant, the odds that $Y{=}1$ change by a multiplicative factor of e^{β_1}.

Using R for fitting GLMs

$y \sim X$

- fitting function:

  `glm(formula, family, data)`

- family: link function, distribution of Y.
  Examples include `binomial`, `gaussian`, `poisson`, `Gamma`

  logit identity log

- complementary functions:
  - `coefficients()`: coefficient estimates
  - `summary()`: prints a summary of results
  - `anova()`: produces an analysis of variance table
  - `residuals`
  - `deviance`
- Optimization technique: Fisher Scoring / IWLS

Case Study 3: The Data

Get the data (from R library):

```
#load Sleuth3 R data library; see case2001
library(Sleuth3)
#Donner party survival data
donner = case2001
str(donner)
```

```
## 'data.frame':    45 obs. of  3 variables:
##  $ Age   : int  23 40 40 30 28 40 45 62 65 45 ...
##  $ Sex   : Factor w/ 2 levels "Female","Male": 2 1 2 2 2 2 1 2 2 1 ...
##  $ Status: Factor w/ 2 levels "Died","Survived": 1 2 2 1 1 1 1 1 1 1 ...
```

ref. group

```
attach(donner)
head(donner)
```

```
##   Age    Sex   Status
## 1  23   Male     Died
## 2  40 Female Survived
## 3  40   Male Survived
## 4  30   Male     Died
## 5  28   Male     Died
## 6  40   Male     Died
```

Case Study 3: Summarizing the data

```
#two-way contingency table for status by sex
#check that cell counts>0
xtabs(~Status+Sex, data=donner)
```

```
##              Sex
## Status       Female Male
##   Died            5   20
##   Survived       10   10
```

10 Males survived

```
summary(Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    15.0    24.0    28.0    31.8    40.0    65.0
```

Case Study 3: Marginal Mean Ages

```
tapply(Age, Status, mean)
```

Av. age by status

```
##      Died Survived
##     35.48    27.20
```

```
tapply(Age, Sex, mean)
```

Av. age by sex

```
##    Female      Male
## 31.06667 32.16667
```

```
fita<-glm(Status~Age+Sex, family=binomial, data=donner)
```

y x_1 x_2 logit link

Case Study 2: Additve model summary

Summary(fita)

```
##
## Call:
## glm(formula = Status ~ Age + Sex, family = binomial, data = donner)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7445  -1.0441  -0.3029   0.8877   2.0472
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.23041    1.38686   2.329   0.0198 *
## Age         -0.07820    0.03728  -2.097   0.0359 *
## SexMale     -1.59729    0.75547  -2.114   0.0345 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

$\pi = P(\text{"survival"})$

$\hat{\beta}$

$\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$

$X_2 = 1$ for Male

Case Study 3: ANOVA table

```r
anova(fita)
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Status
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev
## NULL                    44     61.827
## Age   1   5.5358        43     56.291
## Sex   1   5.0344        42     51.256
```

Case Study 3: Modelling "Died"

```
status=relevel(Status, ref="Survived")
fitad<-glm(status~Age+Sex, family=binomial, data=donner)
summary(fitad)
```

```
##
## Call:
## glm(formula = status ~ Age + Sex, family = binomial, data = donner)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0472  -0.8877   0.3029   1.0441   1.7445
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.23041    1.38686  -2.329   0.0198 *
## Age          0.07820    0.03728   2.097   0.0359 *
## SexMale      1.59729    0.75547   2.114   0.0345 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
```

$\pi = P(\text{"died"})$

$\log\left(\frac{P(\text{"survived"})}{P(\text{"died"})}\right)$

$\log \frac{P(\text{"died"})}{P(\text{"survived"})}$

$\log \frac{a}{b} = \log a - \log b$

$\log b/a = \log b - \log a$

Case Study 3: Sex Reference group as "Male"

```
sex=relevel(Sex, ref="Male")
fitadf<-glm(status~Age+sex, family=binomial, data=donner)
summary(fitadf)
```

```
##
## Call:
## glm(formula = status ~ Age + sex, family = binomial, data = donner)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0472  -0.8877   0.3029   1.0441   1.7445
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.63312    1.11018  -1.471   0.1413
## Age          0.07820    0.03728   2.097   0.0359 *
## sexFemale   -1.59729    0.75547  -2.114   0.0345 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
```

$\pi = P(\text{"died"})$

$X_2 = 1$ if Female

Case Study 3: Sex Reference group as "Male"

```
fitasf<-glm(Status~Age+sex, family=binomial, data=donner)
summary(fitasf)
```

```
##
## Call:
## glm(formula = Status ~ Age + sex, family = binomial, data = donner)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7445  -1.0441  -0.3029   0.8877   2.0472
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.63312    1.11018   1.471   0.1413
## Age         -0.07820    0.03728  -2.097   0.0359 *
## sexFemale    1.59729    0.75547   2.114   0.0345 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

$\pi = P(\text{"Survived"})$

$x_2 = 1$ if Female

Case Study III: Fitted equations

Using *defaults*, $\pi = P(SURVIVED)$:

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = 3.23 - 0.078 Age_i - 1.60 \mathbb{1}_{Male,i}$$

Using other reference status, $\pi = P(DIED)$:

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -3.23 + 0.078 Age_i + 1.60 \mathbb{1}_{Male,i}$$

Using sex reference group as Males, $\pi = P(SURVIVED)$:

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = 1.63 - 0.078 Age_i + 1.60 \mathbb{1}_{Female,i}$$

Using sex reference group as Males, $\pi = P(DIED)$:

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -1.63 + 0.078 Age_i - 1.60 \mathbb{1}_{Female,i}$$

Case Study III: Using Fitted equation

Using the fitted equation for $\pi = P(SURVIVED)$:

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = 1.63 - 0.078 Age_i + 1.60 \mathbb{1}_{Female,i},$$

Q: Estimate the log odds, odds and probability of survival for a:

	Log odds, $\log(\frac{\hat{\pi}}{1-\hat{\pi}})$	Odds, $\frac{\hat{\pi}}{1-\hat{\pi}}$	$\hat{\pi} = \frac{Odds}{1+Odds}$
(i) 20-yr old Female	$1.63 - 0.078(20) + 1.6$		
(ii) 40-yr old Female	$1.63 - 0.078(40) + 1.6$		
(iii) 20-yr old Male	$1.63 - 0.078(20)$		
(iv) 40-yr old Male	$1.63 - 0.078(40)$	$e^{1.63 - 0.078(40)}$	

Case Study III: Using Fitted equation

Using the fitted equation for $\pi = P(SURVIVED)$:

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = 1.63 - 0.078 Age_i + 1.60 \mathbb{1}_{Female,i},$$

Q: Estimate the log odds, odds and probability of survival for a:

	Log odds, $\log(\frac{\hat{\pi}}{1-\hat{\pi}})$	Odds, $\frac{\hat{\pi}}{1-\hat{\pi}}$	$\hat{\pi}$
(i) 20-yr old Female	1.67	5.31	0.84 $= \frac{5.31}{1+5.31}$
(ii) 40-yr old Female	0.11	1.12	0.53
(iii) 20-yr old Male	0.07	1.07	0.52
(iv) 40-yr old Male	-1.49	0.225	0.18

Qs: Compare the odds of survival for a 40-yr old Female to that of a 20-yr old Female. Compare the odds of survival for a 20-yr old Female to that of a Male of the same age.

Case Study III: Using coefficients to find Odds Ratios

$$\boxed{\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2 = 1.63 - \underline{0.078} Age + 1.60 \mathbb{1}_{Female}}$$

1. (Fixed Sex) Compare the odds of survival for a 40-yr old (Female/Male) to that of a 20-yr old (Female/Male).

$$\exp\{\overset{-0.078}{\cancel{-0.78}}(40-20)\} = 0.21 \approx \frac{1}{5}$$

Hence, the odds of survival for a 20-yr old are about 5 times the odds for a 40-yr old of the same sex.

2. (Fixed Age) Compare the odds of survival for a 20-yr old Female to that of a Male of the same age.

$$\exp\{1.60(1-0)\} = 4.95 \approx 5$$

– nutrition
– Social norm

Hence, the odds of survival for a Female are about 5 times the odds for a Male of the same age.

Case Study III: Odds Ratios

1. Compare the odds of survival for a 40-yr old Female to that of a 20-yr old Female.

$$\frac{1.12}{5.31} = 0.21 \approx \frac{1}{5}$$

See page 31.

Hence, the odds of survival for a 20-yr old Female are about 5 times the odds for a 40-yr old Female.

2. Compare the odds of survival for a 20-yr old Female to that of a Male of the same age.

$$\frac{5.31}{1.07} = 4.96 \approx 5$$

Hence, the odds of survival for a 20-yr old Female are about 5 times the odds for a Male of the same age.

Next Class

- Confidence interval for Odds Ratio
- Testing β's $\rightarrow$ Higher-order Models