→ No lecture on March 30th
No lab

→ Recorded Lecture 3
Thursdays Lecture
April 1

→ PS1 is posted

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Lecture 2

Review Generalized Linear Models
More on Logistic Regression: Regression adjustment and continuous covariates

Review of Lecture 1: GLMs

- Generalized Linear models
  - Defines a class of regression models for outcomes from the exponential family of distributions
  - Exponential family includes: Normal, Bernoulli/Binomial, Poisson, Gamma, Beta, among others

- Requires specification of three components:

Random component: $Y_i \sim$ distribution
mean $= \mu_i$
variance

Systematic component:
$$g(\mu_i) = \beta_0 + \beta_1 X_{1i} + \dots + \beta_p X_{pi}$$

~~Link~~ Link function: function g that maps the mean μ_i to the linear function of covariates

$$g(\mu_i) = X_i'\beta \qquad g^{-1}(X_i'\beta) = \mu_i$$

Review of Lecture 1: GLMs

- Linear Model

$Y_i \sim N(\mu_i, \sigma^2)$

$g(\mu_i) = \mu_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_p X_{pi}$

identity link $\Rightarrow$ canonical link
but you could use others

- Logistic Model

$Y_i = \begin{cases} 0 \\ 1 \end{cases}$ $\quad Y_i \sim \text{Bernoulli}(\mu_i)$

mean $= \mu_i = Pr(Y_i = 1)$

variance $= \mu_i(1-\mu_i)$

$g(\mu_i) = \log\left[\frac{\mu_i}{1-\mu_i}\right] = \beta_0 + \beta_1 X_{1i} + \dots + \beta_p X_{pi} = X_i'\beta$

$\mu_i = g^{-1}(X_i'\beta) = \frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)}$

Review of Lecture 1: Key quantities for simple logistic regression

- Assume your outcome is Y taking values 0 vs. 1 and your primary exposure variable X is also binary taking values 0 vs. 1.

$$\log\left[\frac{Pr(Y_i=1|X_i)}{1-Pr(Y_i=1|X_i)}\right]$$

- Mean: $\mu_i = E(Y_i) = Pr(Y_i=1|X_i)$
- Odds: $odds[Pr(Y_i=1|X_i)] = \frac{Pr(Y_i=1|X_i)}{Pr(Y_i=0|X_i)}$ $= \beta_0 + \beta_1 X_i$
- Logit: $Logit\{Pr(Y_i=1|X_i)\} = \log[odds\{Pr(Y_i=1|X_i)\}]$
- Odds ratio

$$\frac{odds[Pr(Y_i=1|X_i=1)]}{odds[Pr(Y_i=1|X_i=0)]} = \frac{Pr(Y_i=1|X_i=1)/Pr(Y_i=0|X_i=1)}{Pr(Y_i=1|X_i=0)/Pr(Y_i=0|X_i=0)}$$

Review of Lecture 1 + additional models

Y_i = Bis expenditure
X_i = MSCD
Z_i = Old: Age > 65

- In this lecture we will consider 4 logistic regression models:
- Model A: $\text{Logit}[Pr(Y_i=1)] = \beta_0$
- Model B: $\text{Logit}[Pr(Y_i=1 \mid X_i)] = \beta_0 + \beta_1 X_i$
- Model C: $\text{Logit}[Pr(Y_i=1 \mid X_i, Z_i)] = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i$
- Model D: $\text{Logit}[Pr(Y_i=1 \mid X_i, Z_i)] = \beta_0 + \beta_1 X_i + \beta_2 Z_i$

- We will fit models B through D and a few additional models as well.

Revisit Model B

```
## Create the necessary variables:
data$posexp=ifelse(data$totalexp>0,1,0)
data$mscd=ifelse(data$lc5+data$chd5>0,1,0)
data1=data[!is.na(data$eversmk),]
data1$older=ifelse(data1$lastage<65,0,1)
data1$bigexp=ifelse(data1$totalexp>1000,1,0)
```

$$\text{Logit}[Pr(Bigexp=1|mscd)] = \beta_0+\beta_1\, mscd$$

Y X

```
## Model B
modelB = glm(bigexp~mscd,data=data1,family="binomial")
lincom(modelB,c("(Intercept)","mscd"))
```

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

```
##                Estimate   2.5 %      97.5 %     Chisq    Pr(>Chisq)
## (Intercept) -0.7395315 -0.7806967 -0.6983663 1239.792 1.372226e-271
## mscd         1.825045   1.694177   1.955913   747.095  1.718138e-164
```

```
lincom(modelB,c("(Intercept)","mscd"),eform=TRUE)
```

$\exp(\hat{\beta}_0)$ $\exp(\hat{\beta}_1)$

```
##                Estimate  2.5 %     97.5 %    Chisq    Pr(>Chisq)
## (Intercept) 0.4773375 0.4580868 0.4973973 1239.792 1.372226e-271
## mscd        6.203076  5.442166  7.070374  747.095  1.718138e-164
```

Revisit Model B

```
lincom(modelB,c("(Intercept)","mscd"))
```

Logit [Pr(Big exp = 1 | mscd)] $= \beta_0 + \beta_1$ mscd

```
##               Estimate   2.5 %      97.5 %     Chisq    Pr(>Chisq)
## (Intercept) -0.7395315 -0.7806967 -0.6983663 1239.792 1.372226e-271
## mscd         1.825045   1.694177   1.955913   747.095 1.718138e-164
```

```
lincom(modelB,c("(Intercept)","mscd"),eform=TRUE)
```

```
##              Estimate  2.5 %     97.5 %    Chisq    Pr(>Chisq)
## (Intercept) 0.4773375 0.4580868 0.4973973 1239.792 1.372226e-271
## mscd        6.203076  5.442166  7.070374  747.095  1.718138e-164
```

$\exp(\hat\beta_0)$

$\exp(\hat\beta_1)$

Pr(Big exp = 1 | mscd = 0) $= \frac{.48}{1+.48}$

- Interpret beta_0 and exp(beta_0)

β_0 = log odds of a big expenditure among persons without a major smoking caused disease.

$\exp(\beta_0)$ = odds of a Big exp among mscD = 0 = .48

- Interpret beta_1 and exp(beta_1)

β_1 = Diff in log odds of a Big exp comparing those with and without a mscD = log odds ratio

$\hat\beta_1 = 1.83$

$\exp(\beta_1)$ = relative odds of a Big exp comparing those with and without a mscD

$\exp(\hat\beta_1) = 6.2$

Interpretation of $\exp(\hat{\beta}_{mscd})$

$\exp(\hat{\beta}_1) = 6.2$

- Two interpretations:

1) The odds of a Big expenditure among persons with a MSCD are 6.2 times the odds of a Big expenditure among persons without a MSCD

$$6.2 = \frac{odds[Pr(Big\ exp = 1 \mid MSCD = 1)]}{odds[Pr(Big\ exp = 1 \mid MSCD = 0)]}$$

$$odds[Pr(Big\ exp = 1 \mid \underline{MSCD = 1})] = 6.2 \times odds[Pr(Big\$ = 1 \mid MSCD = 0)]$$

2) The odds of a Big expendite among persons with a MSCD are $\underbrace{520}$% greater than the odds among persons without a MSCD

Revisit Model C

$\text{logit}[Pr(\text{Bigexp}=1 \mid \text{mscd}, \text{older})] = \beta_0 + \beta_1 \text{mscD} + \beta_2 \text{older} + \beta_3 \text{mscd} \times \text{older}$

```
## Model C
modelC = glm(bigexp~mscd+older+mscd:older,data=data1,family="binomial")
lincom(modelC,c("mscd","mscd+mscd:older","mscd:older"))
```

β_1 $\qquad$ $\beta_1+\beta_3$ $\qquad$ β_3

```
##                   Estimate   2.5 %      97.5 %    Chisq    Pr(>Chisq)
## mscd              1.969895   1.735287   2.204503  270.8301 7.481555e-61
## mscd+mscd:older   1.491115   1.329415   1.652815  326.6619 5.126712e-73
## mscd:older        -0.4787796 -0.7637143 -0.193845 10.84618 0.0009899951
```

```
lincom(modelC,c("mscd","mscd+mscd:older","mscd:older"),eform=TRUE)
```

$\exp(\beta_1)$

older = 0

$\beta_0 + \beta_1 \text{mscd}$

```
##                   Estimate 2.5 %     97.5 %    Chisq    Pr(>Chisq)
## mscd              7.169921 5.670554  9.065741  270.8301 7.481555e-61
## mscd+mscd:older   4.442046 3.778832  5.221658  326.6619 5.126712e-73
## mscd:older        0.619539 0.4659326 0.8237856 10.84618 0.0009899951
```

Among persons 65 yrs old or younger, the odds of Big $ for those with a mscD are 7.17 times the odds for those without a mscD.

Revisit Model C

```r
## Model C
modelC = glm(bigexp~mscd+older+mscd:older,data=data1,family="binomial")
lincom(modelC,c("mscd","mscd+mscd:older","mscd:older"))
```

```
##                  Estimate   2.5 %      97.5 %    Chisq    Pr(>Chisq)
## mscd             1.969895   1.735287   2.204503  270.8301 7.481555e-61
## mscd+mscd:older  1.491115   1.329415   1.652815  326.6619 5.126712e-73
## mscd:older       -0.4787796 -0.7637143 -0.193845 10.84618 0.0009899951
```

```r
lincom(modelC,c("mscd","mscd+mscd:older","mscd:older"),eform=TRUE)
```

```
##                  Estimate 2.5 %     97.5 %    Chisq    Pr(>Chisq)
## mscd             7.169921 5.670554  9.065741  270.8301 7.481555e-61
## mscd+mscd:older  4.442046 3.778832  5.221658  326.6619 5.126712e-73
## mscd:older       0.619539 0.4659326 0.8237856 10.84618 0.0009899951
```

Older = 1

$(\beta_0+\beta_2) + (\beta_1+\beta_3)$ mscD

$\dfrac{\text{older OR } 4.44}{\text{younger OR } 7.17} = .62$

$\exp(\hat\beta_1)$

$\exp(\hat\beta_1+\hat\beta_3)$ Among persons older 65 years of age, the odds of a Big $ for those with a mscD are 4.44 times the odds for those without a mscD.

The relative odds of having a Big expenditure for those w and w/o a mscD are 38% smaller among person over 65 compared to person 65 or younger

Model D: Parameter interpretation and estimation

- Model D: $\text{logit}\left[Pr(\text{Big exp}=1 \mid \text{MSCD}, \text{older})\right] = \beta_0 + \beta_1 \text{MSCD} + \beta_2 \text{older}$
- How do you interpret coefficient for MSCD?

  log odds ratio for a big exp comparing those w and w/o a MSCD among persons of the same age group (younger vs older)
- How do we estimate this coefficient?
  - Inverse-variance weighting estimation!
  - Same as linear regression!
  - Need to consider the age (young vs. old) specific 2x2 tables.

model C

	Young		Old	
	MSCD = 1	MSCD = 0	MSCD = 1	MSCD = 0
Bigexp = 1	273	1802	713	1547
Bigexp = 0	101	4780	232	2236

$$\log OR = \log\left[\frac{273 \cdot 4780}{101 \cdot 1802}\right] \qquad \log\left[\frac{713 \times 2236}{232 \times 1547}\right]$$

$$Var(\lg OR) = \frac{1}{273} + \frac{1}{4780} + \frac{1}{101} + \frac{1}{1802} \qquad \frac{1}{713} + \frac{1}{2236} + \frac{1}{232} + \frac{1}{1547}$$

Model D: Parameter interpretation and estimation

Age group	$log\hat{OR}$	$se(log\hat{OR})$	$var(log\hat{OR})$	$\frac{1}{var(log\hat{OR})}$	$w = \frac{\frac{1}{var(log\hat{OR})}}{\Sigma(\frac{1}{var(log\hat{OR})})}$
Younger	1.97	0.12	0.0144	69.4	0.32
Older	1.49	0.083	0.0069	144.9	0.68

214.3

$\frac{69.4}{214.3}$

$\frac{144.9}{214.3}$

$$\hat{\beta}_1 = 1.97 \times 0.32 + 1.49 \times 0.68 = 1.64$$

$$se(\hat{\beta}_1) = \frac{1}{\sqrt{\sum(\frac{1}{var(log\hat{OR})})}} = \frac{1}{\sqrt{214.3}} = 0.068$$

Model D: Parameter interpretation and estimation

```
modelD = glm(bigexp~mscd+older,data=data1,family="binomial")
summary(modelD)$coeff
```

```
##              Estimate Std. Error  z value      Pr(>|z|)
## (Intercept) -0.9577826 0.02700779 -35.46321 1.815575e-275
## mscd         1.6549130 0.06803662  24.32386 1.096494e-130
## older        0.5638298 0.04104938  13.73540  6.230701e-43
```

```
lincom(modelD,c("mscd","older"),eform=TRUE)
```

```
##        Estimate 2.5 %    97.5 %   Chisq    Pr(>Chisq)
## mscd   5.232625 4.57938  5.979054 591.65   1.096494e-130
## older  1.75739  1.621537 1.904625 188.6613 6.230701e-43
```

Among persons of the same age group, the relative odds of a Big expenditure comparing those w and w/o a MSCD is 5.23

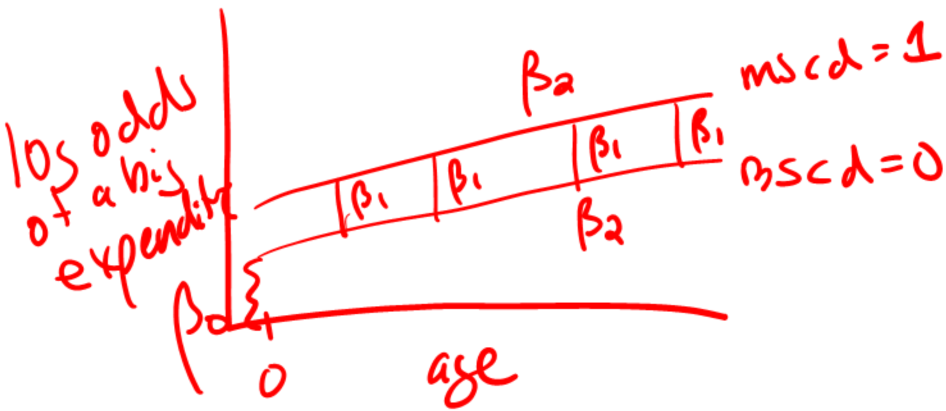
You practice: Use the output above, interpret $exp(\hat{\beta}_2)$.

Among persons with similar disease status, the relative odds of a Big expenditures comparing older to younger persons is 1.76.

", the odds of a Big $ for older persons are 76% greater than the odds for younger person

Model D: Adjustment for continuous covariates

- Now, imagine Model D but where we allow age to be a continuous variable
- Model D with continuous age: $\text{logit}\left[\Pr(\text{Big exp}=1 \mid \text{mscd}, \text{age})\right] = \beta_0 + \beta_1 \text{mscd} + \beta_2 \text{age}$
- Can you draw a picture of this model?



Model D: Adjustment for continuous covariates

```
modelDagecont = glm(bigexp~mscd+lastage,data=data1,family="binomial")
summary(modelDagecont)$coeff
```

```
##              Estimate   Std. Error   z value      Pr(>|z|)
## (Intercept) -2.27990966 0.099135981 -22.99780 4.903428e-117
## mscd         1.60502065 0.068269770  23.50998 3.224831e-122
## lastage      0.02574057 0.001599682  16.09105  2.947835e-58
```

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

Among persons of the same age, the relative odds of a Big $ comparing those w and w/o a MSCD is 4.98

```
lincom(modelDagecont,c("mscd","lastage"),eform=TRUE)
```

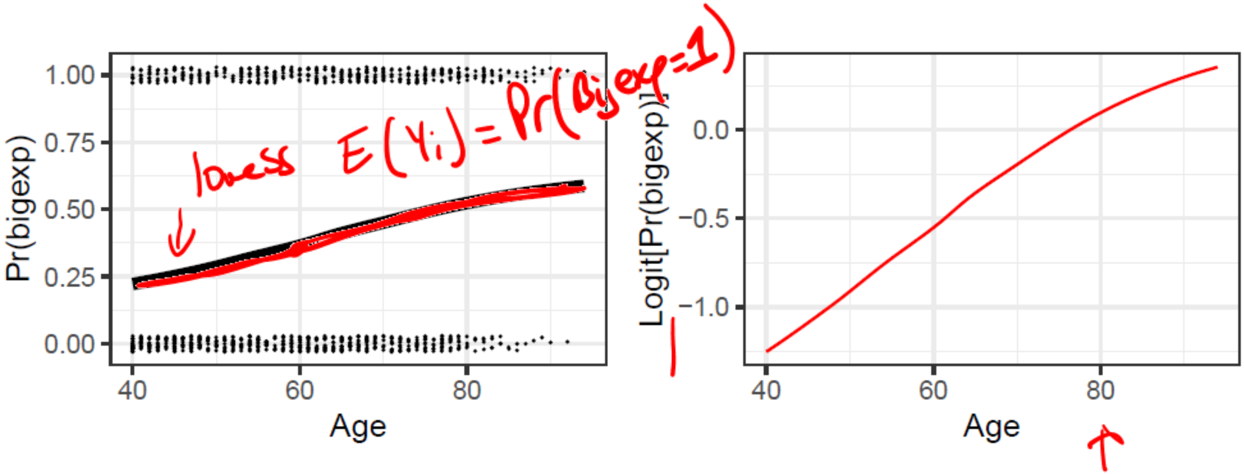
```
##         Estimate 2.5 %    97.5 %   Chisq   Pr(>Chisq)
## mscd    4.977962 4.35452  5.690664 552.719 3.224831e-122
## lastage 1.026075 1.022863 1.029297 258.922 2.947835e-58
```

$\exp(\hat{\beta}_1)$ $\exp(\hat{\beta}_2)$

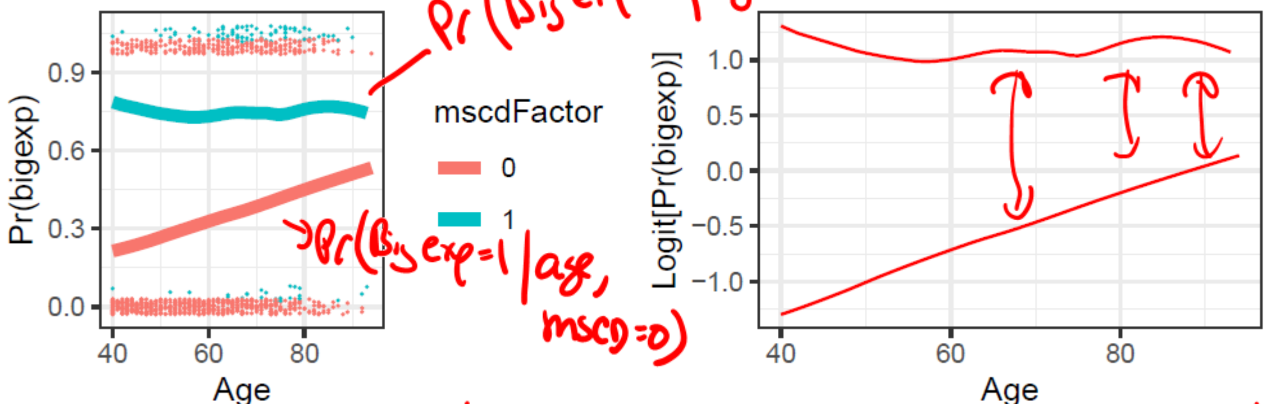
- Interpret both of the coefficients: Among persons with the same disease status, the odds of a Big expenditure increase by 2.6% per additional year of age

Assessing functional form for continuous covariates

- How do we know if the relationship between the logit of a big expenditure and age is linear?



Revisit interaction Model C with continuous age



Pr(Big exp=1 | age, mscd=1)

Pr(Big exp=1 | age, mscd=0)

- What do you think about the MSCD-specific relationship between a big expenditure and age?
  - Linear? Non-linear?

Revision interaction Model C with continuous age

- Assuming the linear assumption is okay!

$$\text{Logit}\left[\Pr(\text{Big exp}=1 \mid \text{mscd}, (\text{age}-60))\right] = \beta_0 + \beta_1 \text{mscd} + \beta_2(\text{age}-60) + \beta_3(\text{age}-60) \times \text{mscd}$$

```
data1$age_c = data1$lastage - 60

modelCcont = glm(bigexp~mscd+age_c+mscd:age_c,data=data1,family="binomial")
lincom(modelCcont,c("mscd","mscd+mscd:age_c","mscd+20*mscd:age_c","mscd:age_c"))
```

```
##                     Estimate   2.5 %       97.5 %      Chisq    Pr(>Chisq)
## mscd                1.792367   1.625218    1.959516    441.7169 4.579093e-98
## mscd+mscd:age_c     1.768144   1.607967    1.928321    468.0905 8.342527e-104
## mscd+20*mscd:age_c  1.307903   1.113342    1.502464    173.595  1.213445e-39
## mscd:age_c          -0.0242232 -0.03631431 -0.01213209 15.41797 8.616514e-05
```

```
lincom(modelCcont,c("mscd","mscd+mscd:age_c","mscd+20*mscd:age_c","mscd:age_c"),eform=TRUE)
```

$\beta_1+\beta_3$

```
##                     Estimate  2.5 %     97.5 %    Chisq    Pr(>Chisq)
## mscd                6.003646  5.079528  7.09589   441.7169 4.579093e-98
## mscd+mscd:age_c     5.859966  4.992649  6.877953  468.0905 8.342527e-104
## mscd+20*mscd:age_c  3.69841   3.044517  4.492744  173.595  1.213445e-39
## mscd:age_c          0.9760678 0.9643371 0.9879412 15.41797 8.616514e-05
```

Among 60 year olds, relative odds of Big exp comparing those w and w/o a MSCD is 6.00 & among 61 year olds, OR is 5.86

$$\frac{5.86}{6.00} = .976$$

Revision interaction Model C with continuous age

- Assuming the linear assumption is okay!

```
data1$age_c = data1$lastage - 60

modelCcont = glm(bigexp~mscd+age_c+mscd:age_c,data=data1,family="binomial")
lincom(modelCcont,c("mscd","mscd+mscd:age_c","mscd+20*mscd:age_c","mscd:age_c"))
```

```
##                    Estimate   2.5 %       97.5 %      Chisq    Pr(>Chisq)
## mscd               1.792367   1.625218    1.959516    441.7169 4.579093e-98
## mscd+mscd:age_c    1.768144   1.607967    1.928321    468.0905 8.342527e-104
## mscd+20*mscd:age_c 1.307903   1.113342    1.502464    173.595  1.213445e-39
## mscd:age_c         -0.0242232 -0.03631431 -0.01213209 15.41797 8.616514e-05
```

```
lincom(modelCcont,c("mscd","mscd+mscd:age_c","mscd+20*mscd:age_c","mscd:age_c"),eform=TRUE)
```

```
##                    Estimate  2.5 %     97.5 %    Chisq    Pr(>Chisq)
## mscd               6.003646  5.079528  7.09589   441.7169 4.579093e-98
## mscd+mscd:age_c    5.859966  4.992649  6.877953  468.0905 8.342527e-104
## mscd+20*mscd:age_c 3.69841   3.044517  4.492744  173.595  1.213445e-39
## mscd:age_c         0.9760678 0.9643371 0.9879412 15.41797 8.616514e-05
```

$\beta_1 + 20\beta_3$ Among 80 yr olds, the OR is 3.70

Revision interaction Model C with continuous age

- How would you rewrite the lincom commands to get estimates of the relationship between having a big expenditure and age, separately for those with and without a MSCD?

```
modelCcont = glm(bigexp~mscd+age_c+mscd:age_c,data=data1,family="binomial")
lincom(modelCcont,c("mscd","mscd+mscd:age_c","mscd+20*mscd:age_c","mscd:age_c"))
```

$$\text{logit}\left[\Pr(\text{Bigexp}=1 \mid \text{mscd}, \text{age}-60)\right]$$
$$= \beta_0 + \beta_1 \text{mscd} + \beta_2(\text{age}-60) + \beta_3(\text{age}-60)\text{mscd}$$

lincom(modelCcont, c("age-c", "age-c + mscd:age-c"), eform=Tre)

Where to next?

- Assessing for confounding in logistic regression models
  - See "Note on confounding and effect modification 2019" by Scott Zeger
  - Additional references are provided in Lecture 2 Handout

- Statistical inference in logistic regression models
  - Maximum likelihood estimation
  - Iteratively reweighted least squares