PS 4 grading ⇒ final grades up

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Lecture 3

Assessing confounding in logistic regression models
MLE in logistic regression models

Review of Lecture 2

- Regression adjustment in logistic regression

$$Y_i = \begin{cases} 1 & \text{bis expenditure} > \$1000 \\ 0 & \text{o/w} \end{cases}$$

- Model B:

$$mscD_i = \begin{cases} 1 & \text{yes} \\ 0 & \text{no} \end{cases} \qquad older_i = \begin{cases} 1 & >65 \\ 0 & \le 65 \end{cases}$$

$$\text{Logit}[Pr(Y_i = 1 \mid mscD_i)] = \beta_0 + \beta_1 mscD_i$$

age_i

- Model D, binary age:

$$\text{Logit}[Pr(Y_i = 1 \mid mscD_i)] = \alpha_0 + \underline{\alpha_1} mscD_i + \alpha_2 older_i$$

- Model D, continuous age:

$$\text{Logit}[Pr(Y_i = 1 \mid age_i)] = \gamma_0 + \underline{\gamma_1} mscD_i + \gamma_2 age_i$$

$$\text{model B: } \exp(\beta_1) = OR = \frac{Pr(Y=1 \mid msco=1) / Pr(Y=0 \mid msco=1)}{Pr(Y=1 \mid msco=0) / Pr(Y=0 \mid msco=0)}$$

Model D: Parameter interpretation and estimation

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

MSCD = OR = 5.2

```
##                Estimate Std. Error   z value      Pr(>|z|)
## (Intercept) -0.9577826 0.02700779 -35.46321 1.815505e-275
## mscd         1.6549130 0.06803662  24.32386 1.096494e-130
## older        0.5638298 0.04104938  13.73540  6.230701e-43
```

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

Among persons in the same age group, the odds of a big exp. if you have a MSCD are 5.2 times the odds if you don't have a MSCD.

```
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
```

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

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

Among persons in the same disease group, the odds of a big exp among older persons are 76% greater then the odds among younger persons.

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
```

Among persons of the same age, ...

```
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
```

- Interpret both of the coefficients:

Among persons with the same disease status, the odds of a big exp increase by 2.6% per year of age.

Assessing confounding in logistic regression

- Question: Is age a confounder for the big expenditure vs. MSCD relationship?
  - Is age associated with MSCD?
  - Is age associated with having a big expenditure?
  - Is age in the causal pathway between MSCD and having a big expenditure?

- We can answer questions 1 and 2 using statistical analyses
  - Question 3 is not a statistical question

Assessing confounding in logistic regression

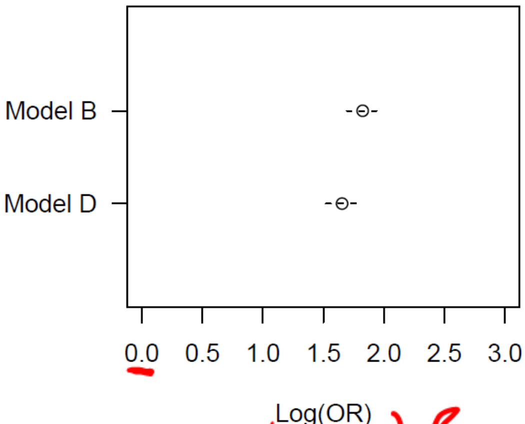
From the analysis of the NMES data, we found:

```
##        Estimate     2.5 %    97.5 %    Chisq     Pr(>Chisq)
## mscd 1.825045 1.694177 1.955913 747.095 1.718138e-164

##        Estimate     2.5 %    97.5 %  Chisq     Pr(>Chisq)
## mscd 1.654913 1.521564 1.788262 591.65 1.096494e-130
```

Model B unadjusted model

Model D adjusted for age



β_1 and α_1

Estimated difference: 1.82 – 1.65 = 0.175

95% CI for the difference: 0.14 to 0.20

bootstrap procedure

What do you think?

Confounding yes or no?

yes!

Assessing confounding in logistic regression

- You can use the process described above for general linear model or generalized linear models with log links
  - However, it gets tricky for other link functions, e.g. the logit link.

Fit unadjusted model
Fit adjusted model
Compare coefficients for primary exposure variable

- For rest of the lecture consider:
  - Y = binary outcome
  - X = primary binary exposure variable
  - Z = potential confounding variable

unadjusted

$$\text{Model M: } logit[Pr(Y=1|X)] = \beta_{0m} + \beta_{1m}X$$

adjusted model

$$\text{Model C: } logit[Pr(Y=1|X,Z)] = \beta_{0c} + \beta_{1c}X + \beta_{2c}Z$$

Non-linearity effect in logistic regression models

Janes 2010

- Assume X and Z are independent, i.e. no confounding
- You can show that $|\beta_{1c}| > |\beta_{1m}|$ and the difference depends on the relationship between X and Z on Y and the variance of Z.

X vs Y
Z vs Y
Var(Z)

- Difference is referred to as the "non-linearity effect"
- This feature of the logistic regression model is known as "non-collapsibility"

Non-linearity effect in logistic regression models

if X and Z are independent

* or example

$|\beta_{1c}| > |\beta_{1m}|$

- Implications for evaluating confounding:

1 $|\beta_{1c}| < |\beta_{1m}|$
  - You have identified "positive confounding" despite the non-linearity effect

* $|\beta_{1c}| > |\beta_{1m}|$
  - You may have "negative confounding" or you may be observing the non-linearity effect

* β_{1c} and β_{1m} have different signs!
  - You may have "qualitative confounding" or you may be observing the non-linearity effect

Linear models: no non-linearity effect

- Y ~ Normal, X and Z independent no confounding
- Marginal model: $E(Y|X) = \beta_{0m} + \beta_{1m} X$ $\quad \beta_{1m} = E(Y|X=1) - E(Y|X=0)$
- Conditional Model: $E(Y|X,Z) = \beta_{0c} + \beta_{1c} X + \beta_{2c} Z$

  $\beta_{1c} = E(Y|X=1,Z) - E(Y|X=0,Z)$

- Marginal model coefficient: law of interated expectation

$$
\begin{aligned}
\underbrace{E(Y|X=1) - E(Y|X=0)}_{\beta_{1m}} &= E_Z[E(Y|X=1,Z) - E(Y|X=0,Z)] \\
&= E_Z[\underbrace{(\beta_{0c} + \beta_{1c} + \beta_{2c} Z)}_{E_Z\{\beta_{1c}\}} - (\beta_{0c} + \beta_{2c} Z)] \\
&= \beta_{1c}
\end{aligned}
$$

Logistic regression: non-linearity effect

- Consider the marginal and conditional odds ratios

$$OR_m \quad exp(\beta_{1m}) = \frac{exp(\beta_{0m}+\beta_{1m})}{exp(\beta_{0m})} = \frac{Pr(Y=1|X=1)/Pr(Y=0|X=1)}{Pr(Y=1|X=0)/Pr(Y=0|X=0)}$$

$$exp(\beta_{1c}) = \frac{exp(\beta_{0c}+\beta_{1c}+\beta_{2c}Z)}{exp(\beta_{0m}+\beta_{2c}Z)} = \frac{Pr(Y=1|X=1,Z)/Pr(Y=0|X=1,Z)}{Pr(Y=1|X=0,Z)/Pr(Y=0|X=0,Z)}$$

OR_c

- When would these be the same?
  - $\beta_{2c} = 0$, Y and Z independent
  - $\beta_{1c} = 0$, $\beta_{1m} = 0$, Y and X independent
  - Var(Z) = 0

→ non linear function of Z

- Why aren't they the same?

$$E(Y|X) = Pr(Y=1|X) = E_Z\left[\frac{exp(\beta_{0c}+\beta_{1c}X+\beta_{2c}Z)}{1+exp(\beta_{0c}+\beta_{1c}X+\beta_{2c}Z)}\right]$$

Simulation: Non-linearity effect

- Assume the following model:

$\beta_{xc} = 1$

$$Logit[\Pr(Y = 1|X, Z)] = \underbrace{-2}_{\beta_{0c}} + 0.4\,X + Z$$

$\beta_{c} = .4$

where *Z ~ N(0,2), and X and Z are independent*

↳ sd

- This model says that regardless of the value of Z, the relative odds of Y = 1 comparing persons with X = 1 to persons with X = 0 are *exp(0.4) =* $\boxed{1.5}$

Conditional OR

OR c

Simulation: Non-linearity effect

This model says that regardless of the value of Z, the relative odds of Y = 1 comparing persons with X = 1 to persons with X = 0 are *exp(0.4) = 1.5*

Logit [Pr(Y=1|X,Z)]
= -2+.4X+Z

Consider a person with Z = 0:

log odds

-2+.4X

$$\underline{Pr(Y=1|X=1,Z=0)} = \frac{exp(\boxed{-2+0.4})}{1+exp(-2+0.4)} = \underline{0.17}$$

$$Pr(Y=1|X=0,Z=0) = \frac{exp(-2)}{1+exp(-2)} = \underline{0.12}$$

$$\underline{OR(Y,X|Z=0)} = \frac{0.17/0.83}{0.12/0.88} = \underline{1.5}$$

Simulation: Non-linearity effect

This model says that regardless of the value of Z, the relative odds of Y = 1 comparing persons with X = 1 to persons with X = 0 are *exp(0.4) = 1.5*

Consider a person with Z = 2:

Log odds = -2 + .4X + 2

$$Pr(Y=1|X=1,Z=2) = \frac{exp(-2+0.4+2)}{1+exp(-2+0.4+2)} = 0.60$$

$$Pr(Y=1|X=0,Z=2) = \frac{exp(-2+2)}{1+exp(-2+2)} = 0.5$$

$$OR(Y,X|Z=2) = \frac{0.6/0.4}{0.5/0.5} = 1.5$$

Simulation: Non-linearity effect

What about the marginal probabilities?

$E_z[Pr(Y=1|X,z)]$

$$Pr(Y = 1|X) = \int_z \underline{Pr(Y = 1|X, z)} f\underline{(z)}dz$$

The marginal probabilities are a weighted average of the conditional probabilities with weights determined by the normal density

X and Z are independent

$$Pr(Y = 1|X = 1) = \underline{0.23}$$

$OR_c = 1.5$

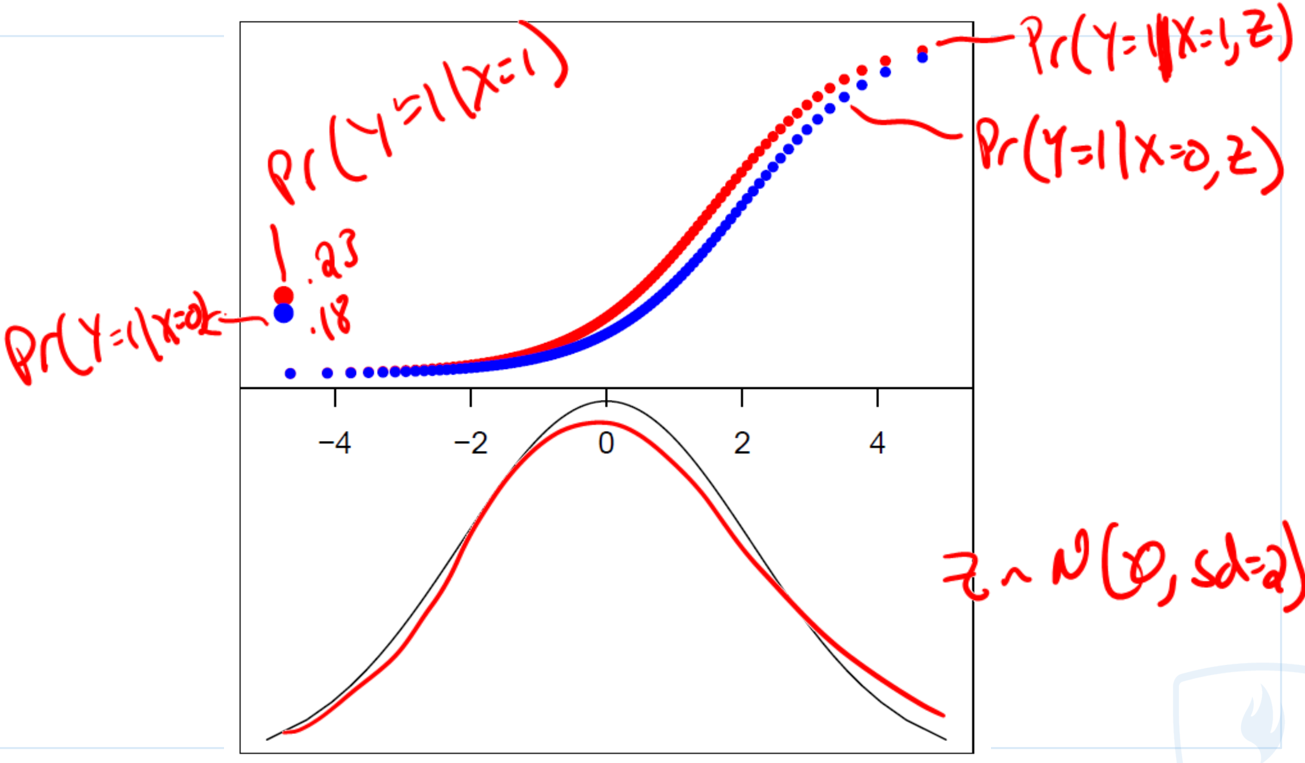
$OR_m = 1.36$

$$Pr(Y = 1|X = 0) = \underline{0.18}$$

marginal $$OR(Y, X) = \frac{0.23/0.77}{0.18/0.82} = \underline{1.36}$$

non-linearity

Simulation: Non-linearity effect



Very important note!

- The non-linearity effect is roughly the same for estimation of β_{1c} and $\text{se}(\hat{\beta}_{1c})$.
- So we would expect the Z statistics for β_{1c} and β_{1m} to be roughly the same if little to no confounding is present (e.g. X and Z are independent).
- So we would expect the Z statistics for β_{1c} and β_{1m} to be different if Z is a confounder.

$$Z_c = \frac{\hat{\beta}_{1c}}{se(\hat{\beta}_{1c})}$$

$$Z_m = \frac{\hat{\beta}_{1m}}{se(\hat{\beta}_{1m})}$$

X and Z independent

Simulation: Confounding present

- Mimic the simulation described in Janes et al (Biostatistics, 2010).

Assume the following:

- We simulate 1000 samples of 250 persons, half with exposure ($X = 1$) and half without ($X = 0$).
- We generate Z as follows: $Z|X \sim N(\alpha_0 + \alpha_1 X, 1)$
- We generate Y from: $logit[Pr(Y = 1|X, Z)] = \beta_0 + \beta_1 X + \beta_2 Z$.
- We set $\alpha_0 = 0$, $\beta_0 = 0$ and $\beta_1 = log(2)$.

conditional est α_1 and β_2

- Simulation scenarios:
  - No confounding / no non-linearity: $\alpha_1 \approx 0, \beta_1 \approx 0$
  - No confounding / non-linearity: $\alpha_1 \approx 0, \beta_1$ large, Y and Z are correlated
  - "Small" confounding: $\alpha_1 > 0, \beta_1 \approx 0$
  - Confounding: α_1, β_1 large

Simulation: Confounding present

β_{1m} β_1

```
##      a1    b1 beta1m beta1 beta1m-beta1   Z1m    Z1 Z1m-Z1
## 1 0.01 0.05  0.701 0.704       -0.003 2.653 2.653  0.000
## 2 0.01 1.50  0.494 0.689       -0.195 1.902 2.183 -0.281
## 3 1.00 0.05  0.752 0.711        0.041 2.839 2.389  0.450
## 4 1.00 1.50  1.617 0.707        0.910 5.312 1.951  3.361
```

- No confounding / no non-linearity: $\alpha_1 \approx 0, \beta_1 \approx 0$
  - Coefficients and test statistics are the same
- No confounding / non-linearity: $\alpha_1 \approx 0, \beta_1$ large
  - Conditional coefficients are different, test statistics are roughly the same
- “Small” confounding: $\alpha_1 > 0, \beta_1 \approx 0$
  - Non-linearity effect is small
  - Marginal coefficient > conditional coefficient -> confounding
  - Test statistics differ
- Confounding: α_1, β_1 large
  - Marginal coefficient > conditional coefficient -> confounding
  - Test statistics differ

Shifting gears to estimation! MLE in linear models

- Define the $(p+1) \times 1$ vector of covariates for subject i as $x_i = (1, x_{1i}, x_{2i}, ..., x_{pi})$.
- Define the $(p+1) \times 1$ vector of association parameters $\beta = (\beta_0, \beta_1, ..., \beta_p)$.

$$Y_i = \mu_i + \epsilon_i, \epsilon_i \text{ iid } N(0, \sigma^2)$$

$$E(Y_i) = \mu_i(\beta) = x_i^{\prime}\beta$$

The score equation, $U(\beta) = \frac{\partial logL(\beta|y_i)}{\partial \beta} = \sum_{i=1}^{n} x_i(y_i - \mu_i(\beta))$.

system of equations
p+1 unknowns
p+1 equations

Setting $U(\beta) = \sum_{i=1}^{n} x_i(y_i - \mu_i(\beta)) = 0$ and solving for β produced:

$$\boxed{\hat{\beta} = (X^{\prime}X)^{-1}X^{\prime}Y}$$

where X is the $n \times p$ matrix of stacked row vectors $x_i^{\prime}$ and Y is the ~~$1 \times n$~~ vector of responses.

nx1

MLE in logistic models

Assume the following model:

- $Y_i \sim Bernoulli(\mu_i)$ for $i = 1, ..., n$ independent observations.
- Define the vector of covariates for subject i as $x_i = (1, x_{1i}, x_{2i}, ..., x_{pi})$.
- Define the vector of association parameters $\beta = (\beta_0, \beta_1, ..., \beta_p)$.
- Assume the logit link such that:

$$log\left(\frac{\mu_i}{1-\mu_i}\right) = x_i'\beta \rightarrow \mu_i = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}$$

NOTE: We should really write $\mu_i(x_i, \beta)$ i.e. μ_i is a function of x_i and β. In this handout, I will simplify this to $\mu_i(\beta)$.

MLE in logistic models

We can express the likelihood function as:

$$
\begin{aligned}
L(\beta|y) &= Pr(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n|\beta) \\
&= \prod_{i=1}^{n} Pr(Y_i = y_i|\beta) \\
&= \prod_{i=1}^{n} \mu_i(\beta)^{y_i}[1 - \mu_i(\beta)]^{1-y_i}
\end{aligned}
$$

The log-likelihood function is:

$$
log[L(\beta|y)] = \sum_{i=1}^{n} y_i log[\mu_i(\beta)] + (1 - y_i)log[1 - \mu_i(\beta)]
$$

MLE in logistic models

The score equation, $U(\beta)$ is the derivative of the log-likelihood function with respect to β.

$$
\begin{aligned}
U(\beta) &= \frac{\partial log[L(\beta|y)]}{\partial\beta} \\
&= \sum_{i=1}^{n} y_i \frac{\partial log[\mu_i(\beta)]}{\partial\beta} + (1-y_i)\frac{\partial log[1-\mu_i(\beta)]}{\partial\beta}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial\beta} log[\mu_i(\beta)] &= \frac{\partial}{\partial\beta} log\left(\frac{e^{x'_i\beta}}{1+e^{x'_i\beta}}\right) \\
&= \frac{\partial}{\partial\beta}[x'_i\beta - log(1+e^{x'_i\beta})] \\
&= x_i - x_i \underbrace{\frac{e^{x'_i\beta}}{1+e^{x'_i\beta}}}_{\mu_i(\beta)} \\
&= x_i[1-\mu_i(\beta)]
\end{aligned}
$$

MLE in logistic models

The score equation, $U(\beta)$ is the derivative of the log-likelihood function with respect to β.

$$
\begin{aligned}
U(\beta) &= \frac{\partial log[L(\beta|y)]}{\partial\beta} \\
&= \sum_{i=1}^{n} y_i \frac{\partial log[\mu_i(\beta)]}{\partial\beta} + (1-y_i)\frac{\partial log[1-\mu_i(\beta)]}{\partial\beta}
\end{aligned}
$$

For the next derivation, note that:

$$\boxed{\frac{\partial log[\mu_i(\beta)]}{\partial\beta} = \frac{1}{\mu_i(\beta)}\frac{\partial\mu_i(\beta)}{\partial\beta}} \rightarrow \frac{\partial\mu_i(\beta)}{\partial\beta} = \mu_i \frac{\partial log[\mu_i(\beta)]}{\partial\beta}$$

$x_i[1-\mu_i(\beta)$

$$
\begin{aligned}
\frac{\partial}{\partial\beta} log[1-\mu_i(\beta)] &= -\frac{1}{1-\mu_i(\beta)}\frac{\partial\mu_i(\beta)}{\partial\beta} \\
&= \frac{-\mu_i(\beta)}{1-\mu_i(\beta)} x_i[1-\mu_i(\beta)] \\
&= -\mu_i(\beta)x_i
\end{aligned}
$$

MLE in logistic models

The score equation, $U(\beta)$ is the derivative of the log-likelihood function with respect to β.

$$
\begin{aligned}
U(\beta) &= \frac{\partial log[L(\beta|y)]}{\partial\beta} \\
&= \sum_{i=1}^{n} y_i \frac{\partial log[\mu_i(\beta)]}{\partial\beta} + (1-y_i)\frac{\partial log[1-\mu_i(\beta)]}{\partial\beta} \\
&= \sum_{i=1}^{n} y_i\left(x_i[1-\mu_i(\beta)]\right) + (1-y_i)[-\mu_i(\beta)x_i] \\
&= \sum_{i=1}^{n} x_i\left(y_i - \cancel{y_i\mu_i(\beta)} + (-\mu_i(\beta)) + \cancel{y_i\mu_i(\beta)}\right) \\
&= \sum_{i=1}^{n} x_i(y_i - \mu_i(\beta)) \\
&= X'(Y-\mu(\beta))
\end{aligned}
$$

MLE in logistic models

NOTE: We will also need to know $U'(\beta) = \frac{\partial U(\beta)}{\partial \beta}$

$$
\begin{aligned}
U'(\beta) &= \frac{\partial U(\beta)}{\partial \beta} \\
&= \frac{\partial}{\partial \beta} X'(Y - \mu(\beta)) \\
&= -X' \frac{\partial \mu_i(\beta)}{\partial \beta} \\
&= -X'VX
\end{aligned}
$$

where we already showed that:

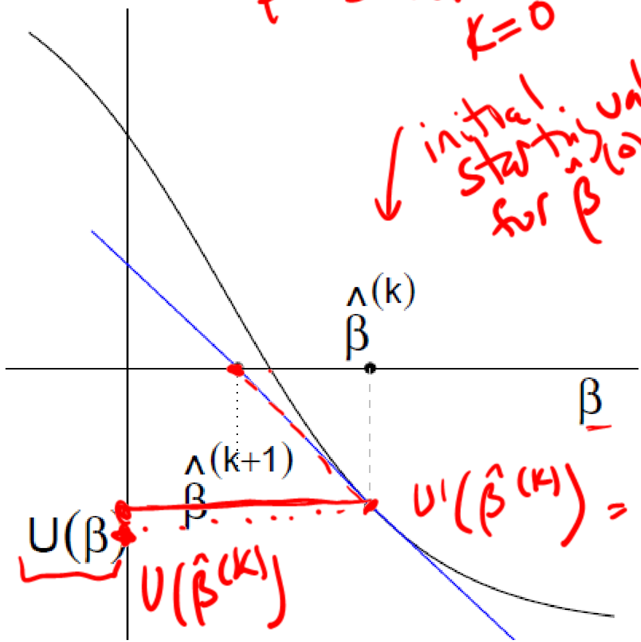
$$\frac{\partial \mu_i(\beta)}{\partial \beta} = \mu_i(\beta) \frac{\partial log[\mu_i(\beta)]}{\partial \beta} = \mu_i(\beta)(1 - \mu_i(\beta))x_i$$

and $V_{n \times n} = diag(\mu_i(\beta)[1 - \mu_i(\beta)])$.

$\Rightarrow Y_i \sim \text{Bernoulli}(\mu_i)$

$Var(Y_i) = \mu_i(1-\mu_i)$

Newton-Raphson Method to find "beta"



β = scalar

$k=0$

initial starting value for $\hat{\beta}^{(0)}$

Find $\hat{\beta}$ such that $U(\hat{\beta})=0$

$(\hat{\beta}^{(k)}, U(\hat{\beta}^{(k)}))$, $(\hat{\beta}^{(k+1)}, 0)$

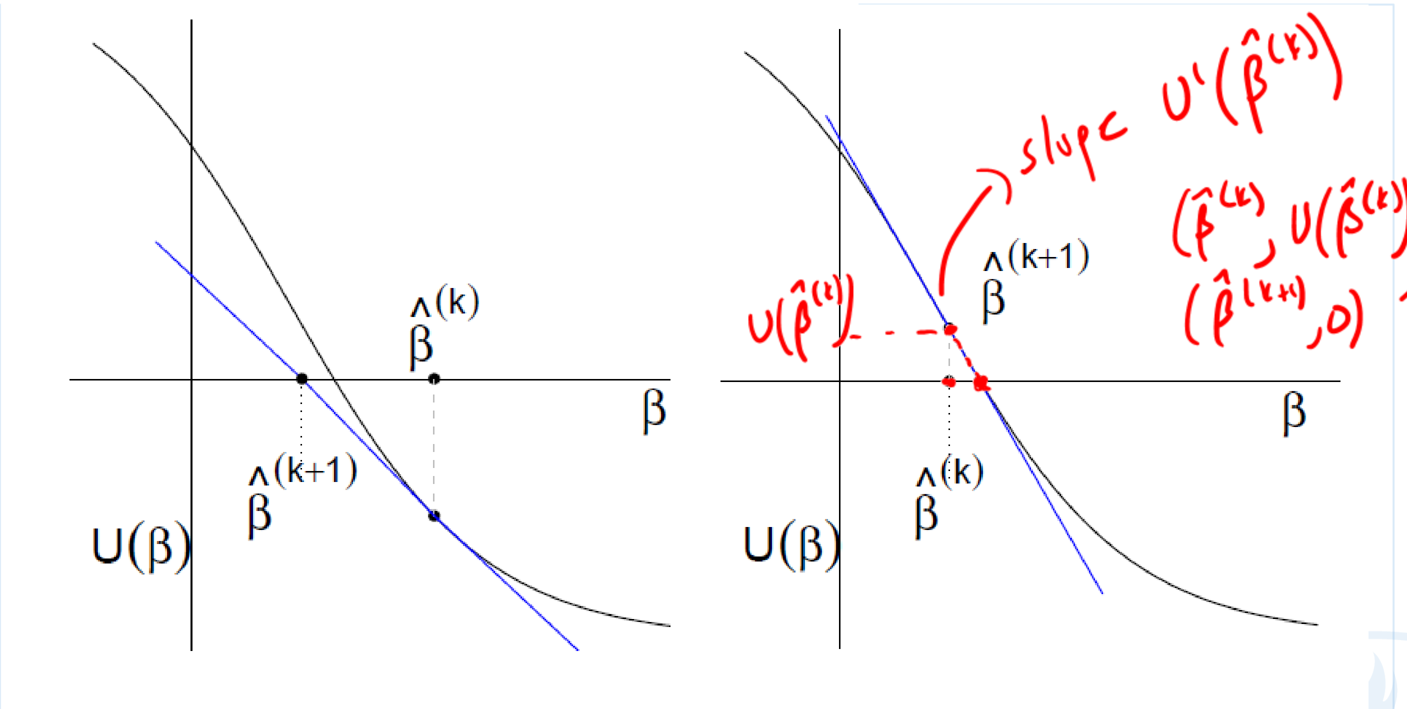
slope in between is $U'(\hat{\beta}^{(k)})$

$U(\hat{\beta}^{(k)})$

$U'(\hat{\beta}^{(k)})$ = slope of the $U(\beta)$ at $\hat{\beta}^{(k)}$

Then compare $\hat{\beta}^{(k)}$ to $\hat{\beta}^{(k+1)}$

Newton-Raphson Method to find “beta”



Newton-Raphson Method to find "beta"

- Step 0: Pick an initial starting value for β, call this $\hat{\beta}^{(k)}$.
- Step 1: Compute the slope of $U(\beta)$ at $\hat{\beta}^{(k)}$, i.e. compute $U'(\hat{\beta}^{(k)})$.
- Step 2: Construct the tangent line, which is a line that passes through the points $(\hat{\beta}^{(k)}, U(\hat{\beta}^{(k)}))$ and $(\hat{\beta}^{(k+1)}, 0)$ and has slope $U'(\hat{\beta}^{(k)})$.
- Step 3: Solve the following for $\hat{\beta}^{(k+1)}$:

$$
\begin{aligned}
U'(\hat{\beta}^{(k)}) &= \frac{U(\hat{\beta}^{(k)})-0}{\hat{\beta}^{(k)}-\hat{\beta}^{(k+1)}} \\
[\hat{\beta}^{(k)} - \hat{\beta}^{(k+1)}]U'(\hat{\beta}^{(k)}) &= U(\hat{\beta}^{(k)}) \\
\hat{\beta}^{(k)} - \hat{\beta}^{(k+1)} &= U'(\hat{\beta}^{(k)})^{-1}U(\hat{\beta}^{(k)}) \\
\hat{\beta}^{(k+1)} &= \hat{\beta}^{(k)} - U'(\hat{\beta}^{(k)})^{-1}U(\hat{\beta}^{(k)}) \\
&= U'(\hat{\beta}^{(k)})^{-1}\left(U'(\hat{\beta}^{(k)})\hat{\beta}^{(k)} - U(\hat{\beta}^{(k)})\right)
\end{aligned}
$$

- Step 4: Stop if $|\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)}|$ is small. If not, let $k = k+1$ and repeat Steps 2 through 4.

Iteratively Re-weighted Least Squares

The general procedure is:

- Step 0: Set an initial value for $\hat{\beta}^{(k)}$, $k = 0$.
- Step 1: Calculate: $V^{(k)}$, $\hat{\mu}(\hat{\beta}^{(k)})$, $Z^{(k)}$.
- Step 2: Update $\hat{\beta}^{(k+1)} = (X'V^{(k)}X)^{-1}(X'V^{(k)}Z^{(k)})$
- Step 3: Stop if $\sum_{j=1}^{p+1}\left(\hat{\beta}_j^{(k+1)} - \hat{\beta}_j^{(k)}\right)^2 < \epsilon$; if not, let $k = k + 1$ and repeat Steps 2 and 3.

Z^k transform Y

reference Lecture 3 handout

Where to next?

review of confounding -> exercise

- inference within logistic regression models
  - Review -> some worked examples