Interpretation of the logistic regression model

HL Chapter 3 – part 1

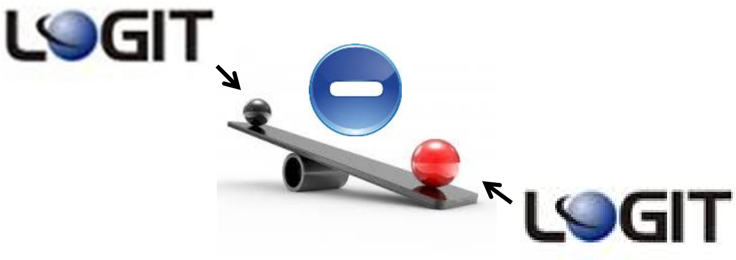
OR calculation

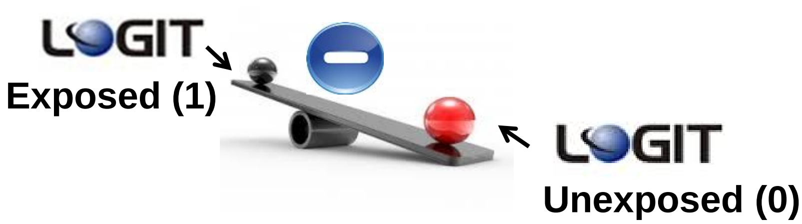
OR =

Exponentiated

LOGIT

difference





Logit = $g(x) = \beta_0 + \beta_1 x$

Logit, Exposed $= g(Exp = 1) = \beta_0 + \beta_1(Exp = 1)$

Logit, Unexposed $= g(Exp = 0) = \beta_0 + \beta_1(Exp = 0)$

Logit difference $= g(Exp = 1) - g(Exp = 0)$

$$= (\beta_0 + \beta_1(1)) - (\beta_0 + \beta_1(0)) = \beta_1$$

$$\rightarrow OR = e^{\beta_1}$$

LOGIT

Male (1)

LOGIT

Female (0)

Logit $= g(x) = \beta_0 + \beta_1 x$

Logit, Male = $g(Gender = 1) = \beta_0 + \beta_1(Gender = 1)$

Logit, Female = $g(Gender = 0) = \beta_0 + \beta_1(Gender = 0)$

Logit difference = $g(Gender = 1) - g(Gender = 0)$

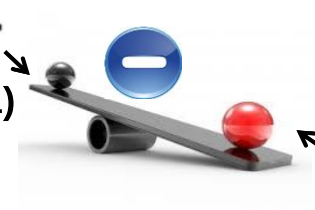
$$= \left(\beta_0 + \beta_1(1)\right) - (\beta_0 + \beta_1(0))$$

$$= \beta_1$$

$$\Rightarrow OR = e^{\beta_1}$$

LOGIT

Exposed (1)
Male (1)



LOGIT

Unexposed (0)
Male (1)

Logit = $g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Logit, Exposed, Male = $g(Exp = 1, Gender = 1)$
$= \beta_0 + \beta_1(Exp = 1) + \beta_2(Gender = 1)$

Logit, Unexposed, Male = $g(Exp = 0, Gender = 1)$
$= \beta_0 + \beta_1(Exp = 0) + \beta_2(Gender = 1)$

Logit difference = $g(Exp = 1, Gender = 1) - g(Exp =$

LOGIT

Exposed (1)
Male (1)

LOGIT

Unexposed (0)
Female (0)

$\boxed{\text{Logit} = g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2}$

Logit, Exposed, Male = $g(Exp=1, Gender=1)$
$= \beta_0 + \beta_1(Exp=1) + \beta_2(Gender=1)$

Logit, Unexposed, Male $g(Exp=0, Gender=0)$
$= \beta_0 + \beta_1(Exp=0) + \beta_2(Gender=0)$

Logit difference = $g(Exp=1, Gender=1) - g(Exp=$

LOGIT

Exposed (1)
Male (1)
Overweight (1)
Black
(R2=1, R3=0)

LOGIT

Unexposed (0)
Female (0)
Normal weight (0)
White
(R2=0, R3=0)

Logit = $g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$

LOGIT

10 years older
Age = a+10

LOGIT

10 years younger
Age = a

Logit = $g(x) = \beta_0 + \beta_1 x_1$

Logit, Older = $g(Age = a + 10) = \beta_0 + \beta_1(Age = a + 10)$
Logit, Younger = $g(Age = a) = \beta_0 + \beta_1(Age = a)$

Logit difference = $g(Age = a + 10) - g(Age = a)$

$$= \big(\beta_0 + \beta_1(a + 10)\big) - (\beta_0 + \beta_1(a))$$
$$= (\beta_0 + \beta_1(a) + \beta_1(10)) - (\beta_0 + \beta_1(a))$$
$$= 10\beta_1$$

$\Rightarrow OR = e^{10\beta_1}$

Goal

- Be able to obtain any OR from a logistic regression model using logit differences

The odds ratio (OR)

- Recall that the logistic regression model is defined as

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

- Further recall that proc logistic provides us with estimates of β_0 and β_1

- e^{β_1} is the OR for an increase in x of 1

The odds ratio (OR)

- Ex. 1: If x=gender (0=female, 1=male), then e^{β_1} is the OR for males vs. females (increase in x of 1)
- Ex. 2: If x=age (continuous), then e^{β_1} is the OR for an increase in age of 1 year

Why can the OR be calculated this way?

$$OR = \frac{ad}{bc} = \frac{ED \times \overline{E}\overline{D}}{E\overline{D} \times \overline{E}D}$$

$$= \frac{\pi(1) \times (1-\pi(0))}{(1-\pi(1)) \times \pi(0)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

What if we want the OR for a 5 year age increase?

- Logit = $g(x) = \beta_0 + \beta_1 x$
- 1 year age increase:

$g(age+1) - g(age) = \left(\beta_0 + \beta_1(age+1)\right) - (\beta_0 + \beta_1(age))$

$= \beta_1 age + \beta_1 - \beta_1 age = \beta_1 \rightarrow OR = e^{\beta_1}$

- 5 year age increase:

$g(age+5) - g(age) = \left(\beta_0 + \beta_1(age+5)\right) - (\beta_0 + \beta_1(age))$

$= \beta_1 age + 5\beta_1 - \beta_1 age = 5\beta_1 \rightarrow OR = e^{5\beta_1}$

What if gender is coded differently?

- Logit = $g(x) = \beta_0 + \beta_1 x$
- Regular coding, male=1, female=0:

$g(gender = 1) - g(gender = 0)$

$= \left(\beta_0 + \beta_1(1)\right) - (\beta_0 + \beta_1(0)) = \beta_1 \rightarrow OR = e^{\beta_1}$

- Different coding, male=3, female=1:

$g(gender = 3) - g(gender = 1)$

$= \left(\beta_0 + \beta_1(3)\right) - (\beta_0 + \beta_1(1)) = 2\beta_1 \rightarrow OR = e^{2\beta_1}$

The logit difference can be used to calculate any OR

What if we have more than 1 variable?

- The logit difference can be used
- Example 1: OR for **MALE SMOKERS** vs. **FEMALE NON-SMOKERS**
- Logit = $g(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
- Male=1, female=0; smoking=1, non-smoking=0:

$$g(\boldsymbol{gender} = \mathbf{1}, \boldsymbol{smo} = \mathbf{1}) - g(\boldsymbol{gender} = \mathbf{0}, \boldsymbol{smo} = \mathbf{0})$$

$$= (\beta_0 + \beta_1(1) + \beta_2\,(1)) - (\beta_0 + \beta_1(0) + \beta_2(0)) = \beta_1 + \beta_2$$

$$\Rightarrow OR = e^{\beta_1 + \beta_2}$$

What if we 3 variables?

- Example 2: OR for **50 YEAR OLD MALE SMOKERS** vs. **30 YEAR OLD FEMALE SMOKERS**

- Logit = $g(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

$$g(\boldsymbol{age} = \mathbf{50}, \boldsymbol{gender} = \mathbf{1}, \boldsymbol{smo} = \mathbf{1})$$
$$- g(\boldsymbol{age} = \mathbf{30}, \boldsymbol{gender} = \mathbf{0}, \boldsymbol{smo} = \mathbf{1})$$
$$= (\beta_0 + \beta_1(50) + \beta_2\,(1) + \beta_3\,(1))$$
$$-(\beta_0 + \beta_1(30) + \beta_2(0) + \beta_3\,(1))$$
$$= 20\beta_1 + \beta_2$$
$$\Rightarrow OR = e^{20\beta_1 + \beta_2}$$

What if we have a continuous and a categorical variable?

- Example 3: OR for **5 years older male smokers** vs. **5 years younger female non-smokers**

- Logit = $g(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

$$g(\boldsymbol{age} + \mathbf{5}, \boldsymbol{gender} = \mathbf{1}, \boldsymbol{smo} = \mathbf{1})$$
$$- g(\boldsymbol{age}, \boldsymbol{gender} = \mathbf{0}, \boldsymbol{smo} = \mathbf{0})$$
$$= (\beta_0 + \beta_1(age + 5) + \beta_2\,(1) + \beta_3\,(1))$$
$$-(\beta_0 + \beta_1(age) + \beta_2(0) + \beta_3\,(0))$$
$$= 5\beta_1 + \beta_2 + \beta_3$$
$$\Rightarrow OR = e^{5\beta_1 + \beta_2 + \beta_3}$$

What about categorical variables with >2 categories?

- The logit difference can still be
- Example: Race 1 (White), 2 (Black), 3 (Asian), 4 (Other)
- Design variables:

Race	r2	r3	r4
1 (White)	0	0	0
2 (Black)	1	0	0
3 (Asian)	0	1	0
4 (Other)	0	0	1

- Logit = $g(r_2, r_3, r_4) = \beta_0 + \beta_1 r_2 + \beta_2 r_3 + \beta_3 r_4$

Black vs. White

Race	r2	r3	r4
1 (White)	0	0	0
2 (Black)	1	0	0
3 (Asian)	0	1	0
4 (Other)	0	0	1

$$g(\boldsymbol{r2}=\mathbf{1}, \boldsymbol{r3}=\mathbf{0}, \boldsymbol{r4}=\mathbf{0}) - g(\boldsymbol{r2}=\mathbf{0}, \boldsymbol{r3}=\mathbf{0}, \boldsymbol{r4}=\mathbf{0})$$

$$= (\beta_0 + \beta_1(1) + \beta_2\,(0) + \beta_3(0)) - (\beta_0 + \beta_1(0) + \beta_2(0) + \beta_3(0)) = \beta_1$$

➔ $OR = e^{\beta_1}$

Asian vs. Black

Race	r2	r3	r4
1 (White)	0	0	0
2 (Black)	1	0	0
3 (Asian)	0	1	0
4 (Other)	0	0	1

$$g(\boldsymbol{r2 = 0, r3 = 1, r4 = 0}) - g(\boldsymbol{r2 = 1, r3 = 0, r4 = 0})$$

$$= (\beta_0 + \beta_1(0) + \beta_2\,(1) + \beta_3(0)) - (\beta_0 + \beta_1(1) + \beta_2(0) + \beta_3(0))$$
$$= -\beta_1 + \beta_2$$

$$\Rightarrow OR = e^{\beta_2 - \beta_1}$$

ORs in SAS

- ORs automatically produced by SAS are the exponentiated coefficients
- If you want a different OR, you must add more statements (see below)

More examples: Dichotomous variable

- GLOW500 data set
- PRIORFRAC=history of prior fracture (1=yes, 0=no)
- Logit = $g(PRIORFRAC) = \beta_0 + \beta_1 PRIORFRAC$
- $g(PRIORFRAC = 1) - g(PRIORFRAC = 0)$

$$= \left(\beta_0 + \beta_1(1)\right) - (\beta_0 + \beta_1(0)) = \beta_1$$

$$\rightarrow OR = e^{\beta_1}$$

More examples: Dichotomous variable

Parameter	Estimate
PRIORFRAC	1.0638

- $OR = e^{1.0638} = 2.897$

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
PRIORFRAC	2.897	1.871	4.486

More examples: Dichotomous variable

- Interpretation
  - First year fracture incidence= 4%
  - OK to interpret OR as relative risk (RR)
  - Subjects with a history of fracture are almost 3 times as likely to have a fracture in the first year than subjects with no history of fracture
  - The increased risk is statistically significant

More examples: Polychotomous variable

RATERISK	r2	r3
1 (Less)	0	0
2 (Same)	1	0
3 (Greater)	0	1

- GLOW500 data set
- RATERISK=self-reported risk of fracture (1=less than, 2=same as, 3=greater than others of the same age)
- Logit = $g(r2, r3) = \beta_0 + \beta_1 r_2 + \beta_2 r_3$
- Same vs. less: $g(\boldsymbol{r2} = \mathbf{1}, \boldsymbol{r3} = \mathbf{0}) - g(\boldsymbol{r2} = \mathbf{0}, \boldsymbol{r3} = \mathbf{0})$

$$= (\beta_0 + \beta_1(1) + \beta_2\,(0)) - (\beta_0 + \beta_1(0) + \beta_2\,(0))$$

$$= \beta_1 \Rightarrow OR = e^{\beta_1}$$

More examples: Polychotomous variable

RATERISK	r2	r3
1 (Less)	0	0
2 (Same)	1	0
3 (Greater)	0	1

- RATERISK=self-reported risk of fracture (1=less than, 2=same as, 3=greater than others of the same age)
- Logit = $g(r2, r3) = \beta_0 + \beta_1 r_2 + \beta_2 r_3$
- Greater vs. less:

$$g(\mathbf{r2} = \mathbf{0}, \mathbf{r3} = \mathbf{1}) - g(\mathbf{r2} = \mathbf{0}, \mathbf{r3} = \mathbf{0})$$
$$= (\beta_0 + \beta_1(0) + \beta_2\,(1)) - (\beta_0 + \beta_1(0) + \beta_2\,(0))$$
$$= \beta_2 \Rightarrow OR = e^{\beta_2}$$

More examples: Polychotomous variable

RATERISK	r2	r3
1 (Less)	0	0
2 (Same)	1	0
3 (Greater)	0	1

- RATERISK=self-reported risk of fracture (1=less than, 2=same as, 3=greater than others of the same age)
- Logit = $g(r2, r3) = \beta_0 + \beta_1 r_2 + \beta_2 r_3$
- Greater vs. same:

$$g(\mathbf{r2} = \mathbf{0}, \mathbf{r3} = \mathbf{1}) - g(\mathbf{r2} = \mathbf{1}, \mathbf{r3} = \mathbf{0})$$
$$= (\beta_0 + \beta_1(0) + \beta_2\,(1)) - (\beta_0 + \beta_1(1) + \beta_2\,(0))$$
$$= -\beta_1 + \beta_2 \Rightarrow OR = e^{\beta_2 - \beta_1}$$

Polychotomous variables in SAS

RATERISK	r2	r3
1 (Less)	0	0
2 (Same)	1	0
3 (Greater)	0	1

- Automatic SAS results

Parameter		Estimate
RATERISK	2	0.5462
RATERISK	3	0.9091

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
RATERISK 2 vs 1	1.727	1.024	2.911
RATERISK 3 vs 1	2.482	1.459	4.223

$OR_{2\ vs.1} = e^{0.5462} = 1.727$

$OR_{3\ vs.1} = e^{0.9091} = 2.482$

Polychotomous variables in SAS

- SAS doesn't automatically calculate greater vs. same
- We need $OR = e^{\beta_2 - \beta_1}$
- We could use the ODDSRATIO statement in SAS but this statement is limited

Polychotomous variables in SAS

- Instead, we use the CONTRAST statement
- Logit differences:

Same vs. Less: $\beta_1 \quad = (\mathbf{1})\ \beta_1 + (\mathbf{0})\ \beta_2$
Greater vs. Less: $\beta_2 \quad = (\mathbf{0})\ \beta_1 + (\mathbf{1})\ \beta_2$
Greater vs. Same: $-\beta_1 + \beta_2 \quad = (-\mathbf{1})\ \beta_1 + (\mathbf{1})\ \beta_2$

Polychotomous variables in SAS

```
proc logistic descending data=glow500;
  class raterisk/param=ref ref=first;
  model fracture=raterisk;

  contrast 'Same vs. Less'     raterisk 1 0/estimate=exp;
  contrast 'Greater vs. Less'  raterisk 0 1/estimate=exp;
  contrast 'Greater vs. Same' raterisk -1 1/estimate=exp;
run;
```

Polychotomous variables in SAS

Contrast	Estimate	Confidence Limits		Pr > ChiSq
Same vs. Less	1.7267	1.0243	2.9108	0.0404
Greater vs. Less	2.4821	1.4589	4.2231	0.0008
Greater vs. Same	1.4375	0.8941	2.3111	0.1341

Polychotomous variables, Interpretation

- Subjects who think their fracture risk is the same as others of the same age
- are 1.7 times as likely (70% more likely)
- to have a fracture in the first year
- than subjects who think their fracture risk is less than others of the same age

- The increased risk is borderline statistically significant

Polychotomous variables, Interpretation

- Subjects who think their fracture risk is the greater than others of the same age
- are about 2.5 times as likely
- to have a fracture in the first year
- than subjects who think their fracture risk is less than others of the same age

- The increased risk is highly statistically significant

Polychotomous variables, Interpretation

- Subjects who think their fracture risk is the greater than others of the same age
- are about 1.4 times as likely (40% more likely)
- to have a fracture in the first year
- than subjects who think their fracture risk is the same as others of the same age

- The increased risk is not statistically significant

More examples: Continuous variable

- Glow500 data set
- AGE=age at enrollment
- Logit = $g(x) = \beta_0 + \beta_1 x$
- 1 year age increase:

$$g(age+1) - g(age)$$
$$= \left(\beta_0 + \beta_1(age+1)\right) - (\beta_0 + \beta_1(age))$$
$$= \beta_1 age + 1\beta_1 - \beta_1 age = \beta_1 \rightarrow OR = e^{\beta_1}$$

Parameter	Estimate
AGE	0.0529

$$OR = e^{0.0529} = 1.054$$

Automatic SAS results

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
AGE	1.054	1.031	1.079

More examples: Continuous variable

- Glow500 data set

- AGE=age at enrollment
- Logit = $g(x) = \beta_0 + \beta_1 x$
- 5 year age increase:

$$g(age + 5) - g(age)$$
$$= (\beta_0 + \beta_1(age + 5)) - (\beta_0 + \beta_1(age))$$
$$= \beta_1 age + 5\beta_1 - \beta_1 age = 5\beta_1 \rightarrow OR = e^{5\beta_1}$$

Parameter	Estimate
AGE	0.0529

$OR = e^{5\times0.0529}$ =1.303

Continuous variables in SAS

- Contrast statement

```
proc logistic descending data=glow500;
  model fracture=age;
  contrast '5 yr increase' age 5/estimate=exp;
run;
```

Contrast	Estimate	Confidence Limits	
5 yr increase	1.3027	1.1624	1.4599

Continuous variables in SAS

- Instead of the contrast statement we can use the units statement:

proc logistic descending data=glow500;
 model fracture=age**/clodds=Wald**;
 units age=5;
run;

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
AGE	5.0000	1.303	1.162	1.460

More examples: Continuous variable

- We can specify more than one unit in the units statement

proc logistic descending data=glow500;
 model fracture=age**/clodds=Wald**;
 units age=5 10;
run;

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
AGE	5.0000	1.303	1.162	1.460
AGE	10.0000	1.697	1.351	2.131

More examples: Continuous variable

- Interpretation
  - A 5-year age increase increases the likelihood of fracture in the first year by about 30%
  - The increased risk is statistically significant
  - A 10-year age increase increases the likelihood of fracture in the first year by about 70%
  - The increased risk is statistically significant