

Simple Logistic Regression Analysis

Asst. Prof. Nikom Thanomsien
Department of Biostatistics & Demography
Faculty of Public Health, Khon Kaen University
Email: nikom@kku.ac.th Web: <http://home.kku.ac.th/nikom>

© 2014 Department of Biostatistics & Demography, Faculty of Public Health, Khon Kaen University

Logistic Regression Analysis

Goals: Logistic Regression Analysis

- ❖ เพื่ออธิบายความสัมพันธ์ระหว่างตัวแปรอธิบายหรือตัวแปรอิสระกับตัวแปรตอบสนองหรือตัวแปรตาม
- ❖ เพื่อใช้ในการทำนาย
 - ตัวแปรผลลัพธ์, ตัวแปรตาม, ตัวแปรตอบสนอง (Outcome, Dependent, Response) = discrete (two possible, dichotomous, binary)
 - ตัวแปรอธิบาย, ตัวแปรอิสระ, ตัวแปรทำนาย (explanatory, independent, predictor) = continuous, categorical

Logistic Regression Analysis

Logistic Regression Analysis

Model	ตัวแปรอธิบาย ตัวแปรอิสระ	ตัวแปรตอบสนอง ตัวแปรตาม
Simple Logistic Regression	1 ตัวแปร	Nominal (dichotomous, binary)
Multiple Logistic Regression	> 1 ตัวแปร	Nominal (dichotomous, binary)
Ordinal Logistic Regression	≥ 1 ตัวแปร	Ordinal
Multinomial Logistic Regression	≥ 1 ตัวแปร	Nominal (polychotomous)

การวิเคราะห์ Logistic Regression: การใช้การวิเคราะห์มีเหตุผลหลัก ๆ 2 ประการ
1. Extremely flexible and easily used function
2. It lends itself to clinically meaningful interpretation
(Cox & Snell, 1989)

Simple Regression Analysis

ทบทวนการวิเคราะห์ความสัมพันธ์และปัจจัยเสี่ยง

- ❖ นักวิจัยต้องการศึกษาความสัมพันธ์ระหว่าง อายุ (<55 ปี=0, ≥ 55 ปี=1) กับการเกิดโรคหัวใจ Coronary

id	age	chd	id	age	chd
76	55	1	37	39	1
75	55	1	61	48	1
88	58	1	46	43	0
78	56	1	25	34	0
98	64	1	33	37	0
87	58	1	41	41	0
100	69	1	4	25	0
96	63	1	29	36	1
91	60	0	72	53	1
79	56	1	42	42	0

```
. cc chd age10, w
      age10
      Exposed Unexposed      Total      Proportion
Cases      9         4         13      0.6923
Controls    1         6         7       0.1429
Total      10        10        20      0.5000

      Point estimate      [95% Conf. Interval]
Odds ratio      13.5      1.197348      152.2113 (woolf)
Attr. frac. ex. .9259259      .1648211      .9934302 (woolf)
Attr. frac. pop .6410256
chi2(1) = 5.49 Pr>chi2 = 0.0191

. clear
. do "H:\516707_2556\chd_age10.do"
. input chd age10 freq
      chd      age10      freq
1. 1 1 9
2. 1 0 4
3. 0 1 1
4. 0 0 6
5. end
. logit chd age10 [fweight=freq], nolog or
Logistic regression
Log likelihood = -9.9809464
Number of obs = 20
LR chi2(1) = 5.94
Prob > chi2 = 0.0148
Pseudo R2 = 0.2292

      chd | Odds Ratio      Std. Err.      z      P>|z|      [95% Conf. Interval]
+-----+-----+-----+-----+-----+
age10 |      13.5      16.68644      2.11      0.035      1.19735      152.2111
```

. logit chd age10, or

. logit chd age10 , or

Iteration 0: log likelihood = -12.948933
Iteration 1: log likelihood = -10.105023
Iteration 2: log likelihood = -9.983536
Iteration 3: log likelihood = -9.9809486
Iteration 4: log likelihood = -9.9809464

Logit estimates
Number of obs = 20
LR chi2(1) = 5.94
Prob > chi2 = 0.0148
Pseudo R2 = 0.2292
Log likelihood = -9.9809464

chd Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age10	13.5	16.68644	2.11	0.035	1.19735 152.2111

odds ratio (simple logistic) เท่ากับ crude odds ratio

Odds ratio	13.5	1.197348	152.2113 (woolf)
Attr. frac. ex.	.9259259	.1648211	.9934302 (woolf)
Attr. frac. pop	.6410256	.8333333	

chi2(1) = 5.49 Pr>chi2 = 0.0191

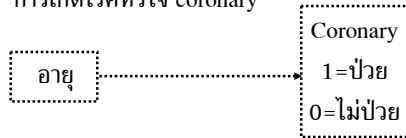
Simple Logistic Regression

ในการวิเคราะห์สมการถดถอยอย่างง่าย $\hat{y} = \beta_0 + \beta_1 X$

สำหรับการวิเคราะห์ logistic $\hat{y} = \ln \frac{\hat{p}}{1-\hat{p}} = \beta_0 + \beta_1 X$

ตัวอย่าง การวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอายุกับ

การเกิดโรคหัวใจ coronary

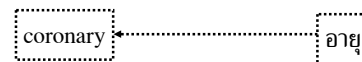


Logistic Regression

ตัวอย่าง การวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอายุกับ

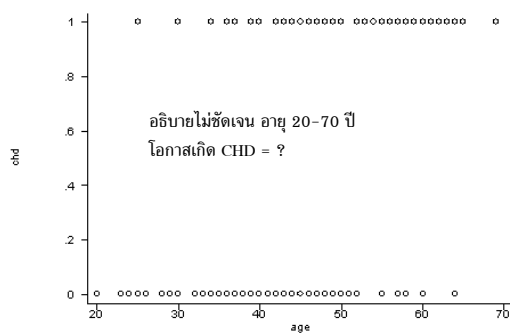
การเกิดโรคหัวใจ coronary

id	age	chd
1	20	0
2	23	0
...	...	...
100	69	1



การวิเคราะห์ Logistic Regression

-plot กราฟ ระหว่างอายุกับ coronary (1,0)

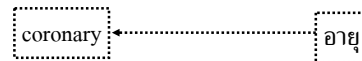


Logistic Regression

การวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอายุกับ

การเกิดโรคหัวใจ coronary

id	age	age_gr	chd
1	20	1	0
2	23	1	0
...	...	...	...
100	69	8	1



plot กราฟ ระหว่างอายุกับ coronary

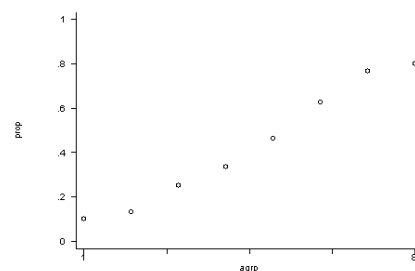
-age_gr vs prob_chd

	age_gr	count	absent	present	prop
1.	1	10	9	1	.1
2.	2	15	13	2	.1333333
3.	3	12	9	3	.25
4.	4	15	10	5	.3333333
5.	5	13	7	6	.4615385
6.	6	8	3	5	.625
7.	7	17	4	13	.7647059
8.	8	10	2	8	.8

การวิเคราะห์ Logistic Regression

-plot กราฟ ระหว่างอายุกับ coronary (age_gr vs prob_chd)

เห็นความสัมพันธ์ อายุมากเกิด CHD มาก (S-shaped)



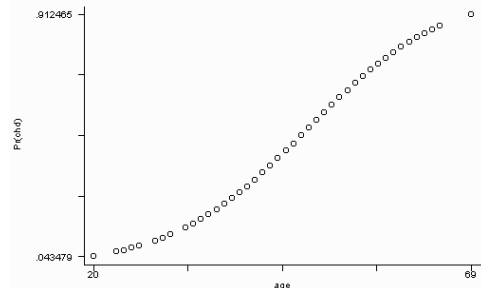
plot กราฟ ระหว่างอายุกับ coronary 100 ราย

-age (continuous) vs prob_chd

	age	chd	prob
1.	69	1	.9124646
2.	54	1	.663803
3.	42	1	.3428171
...			
95.	49	1	.5313793
96.	34	0	.1768066
97.	45	1	.4211628
98.	25	0	.0733438
99.	41	0	.3182802
100.	55	1	.688091

การวิเคราะห์ Logistic Regression

-กรณี plot กราฟ ระหว่างอายุกับ coronary (age vs prob_chd)
เห็นความสัมพันธ์ อายุมากเกิด CHD มาก (S-shaped)



การวิเคราะห์ Logistic Regression

เขียนความสัมพันธ์ได้เป็น $E(y/x) = \beta_0 + \beta_1 x$
x = -∞ ถึง +∞

$$\hat{y} = \ln \frac{\hat{p}}{1-\hat{p}} = \beta_0 + \beta_1 X$$

$$\frac{\hat{p}}{1-\hat{p}} = e^{\beta_0 + \beta_1 X}$$

แก้สมการ คำนวณค่าความน่าจะเป็นในการเกิดเหตุการณ์

$$\hat{p} = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Generalized Linear Model:

- Random component/Family: binomial
- Link Function: logit ดังนั้น $g(\mu) = \ln \left[\frac{\mu}{1-\mu} \right] = \ln \left[\frac{p}{1-p} \right]$
- Systematic component: x_1
- โมเดลเชิงเส้น ได้แก่ $\ln \left[\frac{p}{1-p} \right] = \beta_0 + \beta_1 x_1$

```

. glm chd age, family(binomial) link(logit)
Iteration 0: log likelihood = -9.0660353
...
Iteration 3: log likelihood = -8.9782115
Generalized linear models
Optimization : ML
No. of obs = 20
Residual df = 18
Scale parameter = 1
Deviance = 17.95642306 (1/df) Deviance = .9975791
Pearson = 20.68145326 (1/df) Pearson = 1.14897
Variance function: V(u) = u*(1-u) [Bernoulli]
Link function : g(u) = ln(u/(1-u)) [Logit]
Log likelihood = -8.97821153 AIC = 1.097821 BIC = -35.96676

```

	chd	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
	age	.1425769	.0629518	2.26	0.024	.0191936 .2659599
	_cons	-6.162981	2.962017	-2.08	0.037	-11.96842 -.3575349

Simple Logistic Regression

ตัวอย่าง การวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอายุ (age)
กับการเกิดโรคหัวใจ coronary

id	age	chd	id	age	chd
76	55	1	37	39	1
75	55	1	61	48	1
88	58	1	46	43	0
78	56	1	25	34	0
98	64	1	33	37	0
87	58	1	41	41	0
100	69	1	4	25	0
96	63	1	29	36	1
91	60	0	72	53	1
79	56	1	42	42	0

การวิเคราะห์ Logistic Regression ระหว่าง coronary กับ อายุ

$$\hat{y} = \ln \frac{\hat{p}}{1-\hat{p}} = \beta_0 + \beta_1 X$$

. logit chd age

```

Iteration 0: log likelihood = -12.948933
Iteration 1: log likelihood = -9.1993409
Iteration 2: log likelihood = -8.9848933
Iteration 3: log likelihood = -8.9782217
Iteration 4: log likelihood = -8.9782115

```

Logit estimates

Number of obs = 20
LR chi2(1) = 7.94
Prob > chi2 = 0.0048
Pseudo R2 = 0.3066

	chd	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
	age	.1425769	.0629517	2.26	0.024	.0191939 .2659599
	_cons	-6.162981	2.962012	-2.08	0.037	-11.96842 -.3575447

คำนวณค่าความน่าจะเป็นในการเกิดเหตุการณ์

$$\frac{\hat{p}}{1-\hat{p}} = e^{\beta_0 + \beta_1 X}$$

$$\hat{p} = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

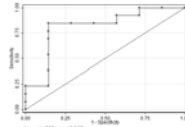
	id	age	chd	phat
1.	76	55	1	.8427389
2.	75	55	1	.8427389
3.	88	58	1	.8915315
...				
19.	72	53	1	.8011657
20.	42	42	0	.4564233

```
. lstat
Logistic model for chd
-----+-----+-----+
Classified | True 0 -0 Total
-----+-----+-----+
+ | 11 1 12
- | 2 6 8
Total | 13 7 20

Classified + if predicted Pr(D) >= .5
True D defined as chd != 0

Sensitivity Pr( + | D) 84.62%
Specificity Pr( - | ~D) 85.71%
Positive predictive value Pr( D + ) 91.67%
Negative predictive value Pr( ~D - ) 75.00%
```

```
. lroc
Logistic model for chd
number of observations = 20
area under ROC curve = 0.8132
```



Simple Logistic Regression

สมมุติฐาน $H_0: \beta_1 = 0; H_A: \beta_1 \neq 0$

การทดสอบนัยสำคัญใช้สถิติ Wald Test (Z)

$$Z = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$$

$$95\%CI = \hat{\beta}_1 \pm Z_{\alpha/2} se(\hat{\beta}_1)$$

. logit chd age

```
Iteration 0: log likelihood = -12.948933
Iteration 1: log likelihood = -9.1993409
Iteration 2: log likelihood = -8.9848933
Iteration 3: log likelihood = -8.9782217
Iteration 4: log likelihood = -8.9782115
```

```
Logit estimates
Number of obs = 20
LR chi2(1) = 7.94
Prob > chi2 = 0.0048
Pseudo R2 = 0.3066
```

chd	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
age	.1425769	.0629517	2.26	0.024	.0191939 .2659599
_cons	-6.162981	2.962012	-2.08	0.037	-11.96842 -.3575447

$$Z = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \quad 95\%CI = \hat{\beta}_1 \pm se(\hat{\beta}_1)$$

chd	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
age	.1425769	.0629517	2.26	0.024	.0191939 .2659599
_cons	-6.162981	2.962012	-2.08	0.037	-11.96842 -.3575447

$$H_0: \beta_1 = 0; H_A: \beta_1 \neq 0$$

ผลการวิเคราะห์

อายุมีความสัมพันธ์กับการเกิดโรคหัวใจ coronary

อย่างมีนัยสำคัญทางสถิติ (z=2.26; p-value=0.024,

95% CI = 0.02, 0.27)

Confidence Interval for the logit

-Estimate Variance of logit

$$Var(\beta_0 + \sum_{i=1}^p \hat{\beta}_i x_i) = Var(\beta_0) + \sum_{i=1}^p x_i^2 Var(\hat{\beta}_i) + \sum_{i=1}^p \sum_{j=1}^p 2x_i x_j Cov(\hat{\beta}_i, \hat{\beta}_j)$$

$$Var(\beta_0 + \hat{\beta}x) = Var(\hat{\beta}_0) + x^2 Var(\hat{\beta}) + 2xCov(\hat{\beta}_0, \hat{\beta})$$

$$100(1-\alpha)5\%CI = \hat{p}(x_i) \pm Z_{\alpha/2} se(\hat{\beta}_0 + \hat{\beta})$$

```
. use "N:\cat2552\logit20case.dta", clear
. logit chd age, noolog

Logistic regression               Number of obs   =        20
                                LR chi2(1)        =         7.94
                                Prob > chi2        =        0.0048
                                Pseudo R2         =        0.3066

                                -----
                                chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
                                -----+-----
                                age |   .1425769   .0629517     2.26   0.024     .0191939   .2659599
                                _cons |  -6.162981   2.962012    -2.08   0.037    -11.96842   -.3575447
                                -----

. vce

Covariance matrix of coefficients of logit model
e(V) |      age      _cons
-----+-----
age |   .00396291   *..... Var(b̂₀)
_cons |  -.18274127   8.7735129 *..... Var(b̂₁)

. prvalue , x(age=50)
logit: Predictions for chd
Confidence intervals by delta method

          Pr(y=1|x):      0.7243   [ 0.4747,   0.9739]
          Pr(y=0|x):      0.2757   [ 0.0261,   0.5253]
          age
          x=      50
```

การแปลผลค่าสัมประสิทธิ์

- ❖ Logistic regression is nonlinear, no approach to Interpretation Can fully describe the relationship between explanatory variables & response.
- ❖ Suggest that try a variety of methods, with goal of finding
- ❖ แปลผลโดยใช้ค่าความน่าจะเป็นจากสมการ
- ❖ แปลผลด้วยค่า odds ratio, percent of change

chd	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
age	.1425769	.0629517	2.26	0.024	.0191939 .2659599
_cons	-6.162981	2.962012	-2.08	0.037	-11.96842 -.3575447

การแปลผลค่าสัมประสิทธิ์ กรณีข้อมูลต่อเนื่อง

-ค่าความน่าจะเป็นจากสมการ

เมื่อ $\beta > 0$, $p(x)$ เพิ่มขึ้น เมื่อ x เพิ่มขึ้น

$\beta < 0$, $p(x)$ ลดลง เมื่อ x เพิ่มขึ้น

ตัวอย่าง เมื่ออายุเพิ่มขึ้น 1 หน่วย ทำให้ความน่าจะเป็นในการเกิดโรคหัวใจ coronary เพิ่มขึ้น

$$\hat{y} = \ln \frac{\hat{p}}{1-\hat{p}} = \hat{\beta}_0 + \hat{\beta}_1 X \Rightarrow \frac{\hat{p}}{1-\hat{p}} = e^{\hat{\beta}_0 + \hat{\beta}_1 X} \Rightarrow \hat{p} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}}$$

```
. use "D:\notebook_acer530\driveE\cate2011\age_coronary.dta", clear
. logit chd age
Iteration 0: log likelihood = -12.948933

-----
chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
age |   .1425769   .0629517     2.26   0.024     .0191939   .2659599
_cons |  -6.162981   2.962012    -2.08   0.037    -11.96842   -.3575447
-----

. predict phat
(option pr assumed; Pr(chd))
. gen phat_cal=(exp(-6.1629808+.14257692*age))/(1+(exp(-6.1629808+.14257692*age)))
. list

+-----+-----+-----+-----+-----+
| id | age | chd | phat | phat_cal |
+-----+-----+-----+-----+-----+
| 1. | 76 | 55 | 1 | .8427389 |
| 2. | 75 | 55 | 1 | .8427389 |
| 3. | 88 | 58 | 1 | .8915315 |
| 19. | 72 | 53 | 1 | .8011657 |
| 20. | 42 | 42 | 0 | .4564233 |
+-----+-----+-----+-----+-----+

. su phat

Variable |      Obs      Mean   Std. Dev.   Min      Max
-----+-----+-----+-----+-----+
phat |      20      .65   .2564946      .4      .9
```

ความน่าจะเป็นของการเกิดโรคหัวใจ coronary มีค่าตั้งแต่ 0.4 ถึง 0.9 โดยมีค่าเฉลี่ยของความน่าจะเป็นในการเกิดโรคเท่ากับ 0.65

การแปลความหมาย odds ratio: กรณีตัวแปรต่อเนื่อง

-การเปลี่ยนแปลง 1 หน่วย is not clinically interesting

เช่นอายุเพิ่มขึ้น 1 ปี หรือความดันโลหิตเพิ่มขึ้น 1 mmHg

-การเปลี่ยนแปลงควรเป็น 5, 10,...

-หรือตรงกันข้าม x มีค่า 0-1 หน่วย การเปลี่ยนแปลง 1 หน่วยเป็นค่ามากไป การเพิ่ม 0.01 อาจมีความเหมาะสมกว่า

-วิธีการคำนวณ odds ratio กรณีตัวแปรต่อเนื่องดังนี้

$$OR(c) = e^{c\hat{\beta}_1}$$

$$95\% CI \text{ of } OR(c) = e^{[c\hat{\beta}_1 \pm Z_{\alpha/2} c[se(\hat{\beta})]}$$

```
. logit chd age
...
chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
_cons |  -6.162981   2.962012    -2.08   0.037    -11.96842   -.3575447
-----

. logit chd age, or
...
chd | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
age |   1.153242   .0725985     2.26   0.024     1.019379   1.304683
-----

. listcoef
logit (N=20): Factor Change in Odds
Odds of: 1 vs 0
-----+-----+-----+-----+-----+
chd |      b      z    P>|z|    e^b    e^bStdX    SdofX
-----+-----+-----+-----+-----+
age |   .1425769   2.26   0.024   1.1532   5.3806   11.8028
-----+-----+-----+-----+-----+

. lincom 10*age or
(1) 10 age = 6
-----+-----+-----+-----+-----+
chd | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----+-----+-----+-----+
(1) |   4.161057   2.619455     2.26   0.024     1.211597   14.29056
-----+-----+-----+-----+-----+

. di exp(10*.1425769)
4.1610565
```

เมื่ออายุเพิ่มขึ้น 10 ปี มีโอกาสเสี่ยงต่อการเกิดโรคหัวใจ coronary 4.16 เท่า

การแปลความหมาย odds ratio: ร้อยละที่เปลี่ยนแปลง

%ที่เปลี่ยนแปลงเท่ากับ $[\exp(c\beta)-1] \times 100$

เช่นอายุเพิ่มขึ้น 1 ปี มีโอกาสเกิดโรคหัวใจ Coronary เพิ่มขึ้น

ร้อยละ 15.3

```

. qui logit chd age
. listcoef, help
logit (N=20): Factor Change in Odds
Odds of: 1 vs 0
-----
chd |      b      z    P>|z|    e^b    e^bStdX    SDofX
age | 0.14258  2.265  0.024  1.1532  5.3886  11.8028
-----
b = raw coefficient
e^b = exp(b) = factor change in odds for unit increase in X
...
. listcoef, percent help
logit (N=20): Percentage Change in Odds
Odds of: 1 vs 0
-----
chd |      b      z    P>|z|      %    %StdX    SDofX
age | 0.14258  2.265  0.024  15.3  438.1  11.8028
-----
% = percent change in odds for unit increase in X

```

```

. logit chd age
Iteration 0: log likelihood = -12.948933
Iteration 1: log likelihood = -9.1993409
Iteration 2: log likelihood = -8.9848933
Iteration 3: log likelihood = -8.9782217
Iteration 4: log likelihood = -8.9782115
Logistic regression
Number of obs = 20
LR chi2(1) = 7.94
Prob > chi2 = 0.0048
Pseudo R2 = 0.3066

Log likelihood = -8.9782115
-----
chd |      Coef.    Std. Err.      z    P>|z|    [95% Conf. Interval]
-----+-----
age | .1425769    .0629517     2.26  0.024    .0191939    .2659599
_cons | -6.162981    2.962012    -2.08  0.037   -11.96842   -.3575447
-----
. di exp(10*.1425769)
4.1610565

. lincom 10*age , or
( 1) 10 age = 0
-----
chd | Odds Ratio    Std. Err.      z    P>|z|    [95% Conf. Interval]
-----+-----
(1) | 4.161057    2.619455     2.26  0.024    1.211597    14.29056

```

เมื่ออายุเพิ่มขึ้น 10 ปี มีโอกาสเสี่ยงต่อการเกิดโรคหัวใจ coronary 4.16 เท่า

การ Fit Model ในการวิเคราะห์ Logistic Regression

-คำนวณค่า coefficient ด้วยวิธี Maximum Likelihood

-การทดสอบระดับนัยสำคัญของ Model

ใช้สถิติ likelihood ratio test (G) ระหว่างโมเดลที่มีเฉพาะ constant กับ fitted Model

$$G = -2 \ln \left[\frac{\text{likelihood without the variable}}{\text{likelihood with g the variable}} \right]$$

```

. logit chd age
Iteration 0: log likelihood = -12.948933
Iteration 1: log likelihood = -9.1993409
Iteration 2: log likelihood = -8.9848933
Iteration 3: log likelihood = -8.9782217
Iteration 4: log likelihood = -8.9782115

Logit estimates
Number of obs = 20
LR chi2(1) = 7.94
Prob > chi2 = 0.0048
Pseudo R2 = 0.3066

Log likelihood = -8.9782115
-----
chd |      Coef.    Std. Err.      z    P>|z|    [95% Conf. Interval]
-----+-----
age | .1425769    .0629517     2.26  0.024    .0191939    .2659599
_cons | -6.162981    2.962012    -2.08  0.037   -11.96842   -.3575447

```

$$G = -2 \ln \left[\frac{\text{likelihood without the variable}}{\text{likelihood with the variable}} \right]$$

$$G = -2 \ln \left[\frac{\text{likelihood without the variable}}{\text{likelihood with the variable}} \right]$$

$$G = (D(\text{model without variable}) - D(\text{model with the variable}))$$

$$D = -2 \ln(\text{likelihood of the fit model})$$

```

Iteration 0: log likelihood = -12.948933
Iteration 1: log likelihood = -9.1993409
Iteration 2: log likelihood = -8.9848933
Iteration 3: log likelihood = -8.9782217
Iteration 4: log likelihood = -8.9782115

Logit estimates
Number of obs = 20
LR chi2(1) = 7.94
Prob > chi2 = 0.0048
Pseudo R2 = 0.3066

Log likelihood = -8.9782115

```

$$G = -2[-12.9489333 - (-8.9782115)] = 7.9414436$$

$$G \cong \chi^2(\nu)$$

Simple Logistic Regression: ตัวแปรอิสระ Dichotomous

-จากข้อมูล การวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอายุกับการเกิดโรคหัวใจ coronary กำหนดให้ อายุ <55 ปี=0, อายุ ≥ 55 ปี=1

id	age	chd	age10	id	age	chd	age10
76	55	1	1	37	39	1	0
75	55	1	1	61	48	1	0
88	58	1	1	46	43	0	0
78	56	1	1	25	34	0	0
98	64	1	1	33	37	0	0
87	58	1	1	41	41	0	0
100	69	1	1	4	25	0	0
96	63	1	1	29	36	1	0
91	60	0	1	72	53	1	0
79	56	1	1	42	42	0	0

วิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอายุ (age10)กับการเกิดโรคหัวใจ coronary หรือศึกษาอายุเป็นปัจจัยเสี่ยงต่อการเกิดโรคหัวใจ coronary

. logit chd age10

```
Iteration 0: log likelihood = -12.948933
Iteration 1: log likelihood = -10.105023
Iteration 2: log likelihood = -9.9809464
Iteration 3: log likelihood = -9.9809464
Iteration 4: log likelihood = -9.9809464
```

```
Logit estimates                                Number of obs   =      20
                                                LR chi2(1)      =      5.94
                                                Prob > chi2     =     0.0148
Log likelihood = -9.9809464                    Pseudo R2      =     0.2292
```

chd	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
age10	2.60269	1.236032	2.11	0.035	.1801107 5.025269
_cons	-.4054651	.6454972	-0.63	0.530	-1.670616 .8596862

การทดสอบสมมติฐาน $H_0: \beta=0$

อายุมีความสัมพันธ์กับการเกิดโรคหัวใจ coronary อย่างมีนัยสำคัญทางสถิติ ($z=2.11$; $p\text{-value}=0.035$, $95\%CI = .18-5.02$)

การแปลความหมาย odds ratio: ตัวแปร dichotomous

$$OR = e^{\hat{\beta}_1}$$

(Hosmer & Lemeshow (2000). Applied Logistic Regression.

Page 48-50.)

chd	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
age10	2.60269	1.236032	2.11	0.035	.1801107 5.025269
_cons	-.4054651	.6454972	-0.63	0.530	-1.670616 .8596862

$$OR = e^{2.60269} = 13.5$$

การแปลความหมาย

กลุ่มอายุ >=50 ปี มีโอกาสเสี่ยงต่อการเกิดโรคหัวใจ coronary เป็น 13.5 เท่า ของ กลุ่มอายุ <50 ปี

คำสั่ง logit & logistic

. logit chd age10,or

```
Iteration 0: log likelihood = -12.948933
```

```
Logit estimates                                Number of obs   =      20
                                                LR chi2(1)      =      5.94
                                                Prob > chi2     =     0.0148
Log likelihood = -9.9809464                    Pseudo R2      =     0.2292
```

chd	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age10	13.5	16.68644	2.11	0.035	1.19735	152.2111

. logistic chd age10

```
Logistic regression                            Number of obs   =      20
                                                LR chi2(1)      =      5.94
                                                Prob > chi2     =     0.0148
Log likelihood = -9.9809464                    Pseudo R2      =     0.2292
```

chd	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age10	13.5	16.68644	2.11	0.035	1.19735	152.2111

คำสั่ง glm

```
. glm chd age10 , f(binomial) link(logit)
```

```
Generalized linear models                    No. of obs   =      20
Optimization : ML                          Residual df   =      18
Deviance     = 19.96189281                  Scale parameter =      1
Pearson      = 20                          (1/df) Deviance = 1.108994
Variance function: V(u) = u*(1-u)          (1/df) Pearson = 1.111111
Link function : g(u) = ln(u/(1-u))         [Bernoulli]
Log likelihood = -9.980946404              [Logit]
AIC = 1.198095
BIC = -33.96129
```

chd	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
age10	2.60269	1.236032	2.11	0.035	.1801107 5.025269
_cons	-.4054651	.6454972	-0.63	0.530	-1.670616 .8596862

```
. glm chd age10 , f(binomial) link(logit) ef
```

```
Generalized linear models                    No. of obs   =      20
Optimization : ML                          Residual df   =      18
Deviance     = 19.96189281                  Scale parameter =      1
Pearson      = 20                          (1/df) Deviance = 1.108994
Variance function: V(u) = u*(1-u)          (1/df) Pearson = 1.111111
Link function : g(u) = ln(u/(1-u))         [Bernoulli]
Log likelihood = -9.980946404              [Logit]
AIC = 1.198095
BIC = -33.96129
```

chd	Odds Ratio	OIM Std. Err.	z	P> z	[95% Conf. Interval]	
age10	13.5	16.68645	2.11	0.035	1.197348	152.2113

คำสั่ง listcoef

```
. logit chd age10
```

```
Logistic regression                            Number of obs   =      20
                                                LR chi2(1)      =      5.94
                                                Prob > chi2     =     0.0148
Log likelihood = -9.9809464                    Pseudo R2      =     0.2292
```

chd	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
age10	2.60269	1.236032	2.11	0.035	.1801107 5.025269
_cons	-.4054651	.6454972	-0.63	0.530	-1.670616 .8596862

```
. listcoef, help
```

```
logit (N=20): Factor Change in Odds
```

```
Odds of: 1 vs 0
```

chd	b	z	P> z	e^b	e^bStdX	SDofX
age10	2.60269	2.106	0.035	13.5000	3.8006	0.5130

b = raw coefficient
z = z-score for test of b=0
P>|z| = p-value for z-test
e^b = exp(b) = factor change in odds for unit increase in X
e^bStdX = exp(b*SD of X) = change in odds for SD increase in X
SDofX = standard deviation of X

การแปลความหมาย odds ratio: ร้อยละที่เปลี่ยนแปลง

%ที่เปลี่ยนแปลงเท่ากับ $[\exp(\beta)-1] \times 100$

เช่นกลุ่มอายุมากกว่าหรือเท่ากับ 55 ปี มีโอกาสเกิดโรคหัวใจ Coronary เพิ่มขึ้น ร้อยละ $(13.5-1) \times 100 = 1250$ เมื่อเปรียบเทียบกับกลุ่มอายุน้อยกว่า 55 ปี

```
. listcoef , percent
```

```
logit (N=20): Percentage Change in Odds
```

```
Odds of: 1 vs 0
```

chd	b	z	P> z	%	%StdX	SDofX
age10	2.60269	2.106	0.035	1250.0	280.1	0.5130

การคำนวณ odds ratio

ตาราง 2x2 การคำนวณ odds ratio จาก ad/bc

หรือ [ความน่าจะเป็นที่การสัมผัสปัจจัยและเกิดโรค/
ความน่าจะเป็นที่ไม่สัมผัสปัจจัยและเกิดโรค]หารด้วย
[ความน่าจะเป็นที่การสัมผัสปัจจัยและไม่เกิดโรค/
ความน่าจะเป็นที่ไม่สัมผัสปัจจัยและไม่เกิดโรค]

$$or = \frac{a/c}{b/d} = \frac{odds1}{odds0} = \frac{ad}{bc} = \frac{p(1)(1-p(0))}{p(0)(1-p(1))}$$

E+	D+		D-	
	$p(1)$	a	b	$1-p(1)$
E-	$p(0)$	c	d	$1-p(0)$

$$\text{จากความน่าจะเป็น} \quad \hat{p} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}}$$

แทนค่า E = 0 และ E = 1 จะได้ค่าความน่าจะเป็นใน 2x2

	D+		D-	
E+	$p(1) = \frac{e^{(\hat{\beta}_0 + \hat{\beta}_1)}}{1 + e^{(\hat{\beta}_0 + \hat{\beta}_1)}}$	a	$1-p(1) = \frac{1}{1 + e^{(\hat{\beta}_0 + \hat{\beta}_1)}}$	b
E-	$p(0) = \frac{e^{\hat{\beta}_0}}{1 + e^{\hat{\beta}_0}}$	c	$1-p(0) = \frac{1}{1 + e^{\hat{\beta}_0}}$	d

เนื่องจาก $or = ad/bc$ ดังนั้นจากตาราง

$$or = \frac{\frac{e^{(\hat{\beta}_0 + \hat{\beta}_1)}}{1 + e^{(\hat{\beta}_0 + \hat{\beta}_1)}} \cdot \frac{1}{e^{\hat{\beta}_0}}}{\frac{e^{\hat{\beta}_0}}{1 + e^{\hat{\beta}_0}} \cdot \frac{1}{e^{(\hat{\beta}_0 + \hat{\beta}_1)}}} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1}}{e^{\hat{\beta}_0}}$$

$$or = e^{\hat{\beta}_0 + \hat{\beta}_1} e^{-\hat{\beta}_0} = e^{\hat{\beta}_1}$$

	D+		D-	
E+	$p(1) = \frac{e^{(\hat{\beta}_0 + \hat{\beta}_1)}}{1 + e^{(\hat{\beta}_0 + \hat{\beta}_1)}}$	a	$1-p(1) = \frac{1}{1 + e^{(\hat{\beta}_0 + \hat{\beta}_1)}}$	b
E-	$p(0) = \frac{e^{\hat{\beta}_0}}{1 + e^{\hat{\beta}_0}}$	c	$1-p(0) = \frac{1}{1 + e^{\hat{\beta}_0}}$	d

การวิเคราะห์ข้อมูลโดยใช้ odds ratio ประมาณค่า risk ratio

-Majority of biostatistician/statisticians use the odds ratios produced from logit model as an estimate of risk ratio or relative risk

-Cornfield (1951) originally proposed the notion of odds ratio as a good approximation to relative risk when the latter could not be determine.

-odds ratio ประมาณ risk ratio กรณี outcome is rare

-กรณี outcome is not rare -> Bias to approximation risk ratio

ตาราง* แสดงค่า Risk ratio เปรียบเทียบกับ Odds ratio ตามความน่าจะเป็นในการเกิดเหตุการณ์

P(event x)	.25	.50	.75	1.00	1.50	2.00	4.00
.2	.2941	.5556	.7895	1.000	1.364	1.667	2.500
.1	.2702	.5263	.7692	1.000	1.429	1.818	3.077
.01	.2519	.5025	.7519	1.000	1.493	1.980	3.883
.001	.2502	.5003	.7502	1.000	1.499	1.998	3.988
.0001	.2500	.5000	.7500	1.000	1.500	2.000	3.999

*Gould (2000). Interpreting logistic regression in all its forms.

Estimating relative risk in Cohort studies of common outcome

❖ Adjusted odds ratio: Convert binary response logistic regression odds ratio prospective cohort studies to estimated risk ratios (Zhang & Yu, 1998)

$$rr = \frac{or}{(1-p_0) + (p_0 * or)} \quad or = \frac{p_1/(1-p_1)}{p_0/(1-p_0)}$$

p_1 = incidence of the outcome in exposed group ($y=1|x=1$)

p_0 = incidence of the outcome in non-exposed group ($y=1|x=0$)

Estimating relative risk in Cohort studies of common outcome

❖ Log-Binomial: Generalized linear regression with a log link and binomial distribution

(Wacholder, 1986; Robbins, Chao, & Fonseca 2002; McNutt et al. 2003; Barros & Hirakata 2003)

Called log-binomial (Blizzard & Hosmer 2006) or

Binomial log-linear regression (Greenland 2004a)

adjusted odds ratio (Zhang & Yu) is incorrect & will produce biased estimate when confounding is present.

ศึกษาโมเดลต่างๆ ประมาณค่า risk ratio: Log-binomial, Poisson etc.

-> less standard error prefer

“Generalized linear regression with a log link & binomial distribution”

-พบปัญหา lack of convergence

Estimating relative risk in Cohort studies of common outcome

-A modified Poisson regression (Zou, 2004) approach to prospective studies with binary data.

โมเดลเชิงเส้นโดยนัยทั่วไป family = Poisson; ฟังก์ชันเชื่อมโยง link=log & robust variance

-Generalized linear models: Extensions to the binomial family

(Hardine & Hilbe, 2007) โมเดลเชิงเส้นโดยนัยทั่วไป family = binomial; ฟังก์ชันเชื่อมโยง link=log & iterative reweight least square (IRLS)

ตัวอย่าง การศึกษาปัจจัยเสี่ยงต่อภาวะ Microalbuminuria (ภาวะที่มีการขับ albumin ทางปัสสาวะ; micro24: 1=microalbuminuria, 0=normal) ตัวอย่าง 172 ราย ข้อมูลส่วนหนึ่งของการศึกษา Diabetes Control & Complication Trial: DCCT)

ตัวแปร Explanatory

intensive = treatment (0=intensive, 1=conventional)

hbae1=the average level of blood glucose control over the preceding 4 to 6 weeks. (Hb_{A1c} at baseline)

duration=the prior duration of diabetes in months

sbp=the level of systolic blood pressure (in mmHg)

female = gender (1=female, 0=male)

ในที่นี้วิเคราะห์เฉพาะตัวแปร micro24 กับ intensive

```
. cs micro24 intensive
+-----+-----+-----+
|      | intensive | Unexposed | Total |
+-----+-----+-----+
| Cases |      31   |      11   |     42 |
| Noncases |      52   |      78   |    130 |
+-----+-----+-----+
| Total |      83   |      89   |    172 |
+-----+-----+-----+
| Risk  | .373494   | .1235955  | .244186 |
+-----+-----+-----+
|      | Point estimate | [95% Conf. Interval] |
+-----+-----+-----+
| Risk difference | 2498985 | .1253782 | .3744188 |
| Risk ratio      | 3.021906 | 1.62654  | 5.614318 |
| Attr. frac. ex. | .669083 | .3851982 | .821884  |
| Attr. frac. pop | .493847 |          |          |
+-----+-----+-----+
| chi2(1) = 14.53 Pr>chi2 = 0.0001
```

```
. cc micro24 intensive, woolf
+-----+-----+-----+-----+
|      | intensive | Unexposed | Total | Proportion Exposed |
+-----+-----+-----+-----+
| Cases |      31   |      11   |     42 |         0.7381     |
| Controls |      52   |      78   |    130 |         0.4000     |
+-----+-----+-----+-----+
| Total |      83   |      89   |    172 |         0.4826     |
+-----+-----+-----+-----+
|      | Point estimate | [95% Conf. Interval] |
+-----+-----+-----+-----+
| Odds ratio | 4.227273 | 1.953016 | 9.149865 (woolf) |
| Attr. frac. ex. | .7634409 | .4879715 | .8907088 (woolf) |
| Attr. frac. pop | .5634921 |          |          |
+-----+-----+-----+-----+
| chi2(1) = 14.53 Pr>chi2 = 0.0001
```

```
. glm micro24 intensive , family(binomial) link(logit) nolog ef
generalized linear models      No. of obs   =    172
Optimization      : ML        Residual df =    170
Scale parameter =    1
Deviance          = 176.2679883 (1/df) Deviance = 1.036871
Pearson          =    172      (1/df) Pearson = 1.011765

Variance function: V(u) = u*(1-u)      [Bernoulli]
Link function      : g(u) = ln(u/(1-u)) [Logit]
AIC                =    1.04807
Log likelihood     = -88.13399417      BIC        = -698.8061
```

	micro24	OIM Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
intensive	1	4.227273	1.665451	3.66	0.000	1.953016 9.149865

Estimating relative risk in Cohort studies of common outcome

❖ Adjusted odds ratio (Zhang & Yu, 1998)

```
. su micro24 if intensive==0
+-----+-----+-----+-----+
| Variable | Obs | Mean | Std. Dev. | Min | Max |
+-----+-----+-----+-----+
| micro24 | 89 | .1235955 | .3309842 | 0 | 1 |
+-----+-----+-----+-----+
. su micro24 if intensive==1
+-----+-----+-----+-----+
| Variable | Obs | Mean | Std. Dev. | Min | Max |
+-----+-----+-----+-----+
| micro24 | 83 | .373494 | .4866722 | 0 | 1 |
+-----+-----+-----+-----+
. di (.373494 / (1-.373494)) / (.1235955 / (1-.1235955))
4.2272734
. di 4.2272734 / ((1-.1235955) + (.1235955 * 4.2272734))
3.0219061
```

Predictor	Odds Ratio	Risk Ratio	[95% Conf. Interval]
intensive	4.2273	3.0219	1.7472 4.5583

Estimating relative risk in Cohort studies of common outcome

❖ Log-Binomial McNutt et al. (2003).

```
. glm micro24 intensive , family(binomial) link(log) nolog ef
Generalized linear models      No. of obs   =    172
Optimization      : ML        Residual df =    170
Scale parameter =    1
Deviance          = 176.2679883 (1/df) Deviance = 1.036871
Pearson          =    172      (1/df) Pearson = 1.011765

Variance function: V(u) = u*(1-u)      [Bernoulli]
Link function      : g(u) = ln(u)      [Log]
AIC                =    1.04807
Log likelihood     = -88.13399417      BIC        = -698.8061
```

	micro24	Risk Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
intensive	1	3.021906	.9550514	3.50	0.000	1.62654 5.614318

Estimating relative risk in Cohort studies of common outcome

❖ Modified Poisson regression (Zou, 2004)

```
. glm micro24 intensive , family(poisson) link(log) nolog vce(robust) ef
```

Generalized linear models
Optimization : ML

No. of obs = 172
Residual df = 170
Scale parameter = 1
(1/df) Deviance = .6297483
(1/df) Pearson = .7647059

Deviance = 107.0572151
Pearson = 129.9999993

Variance function: V(u) = u
Link function : g(u) = ln(u)

[Poisson]
[Log]

AIC = 1.134054
BIC = -.768.0168

Log pseudolikelihood = -95.52860757

micro24	IRR	Robust Std. Err.	z	P> z	[95% Conf. Interval]
intensive	3.021906	.9578398	3.49	0.000	1.623601 5.624481

Estimating relative risk in Cohort studies of common outcome

❖ Generalized linear models: Extensions to the binomial family

(family = binomial; link=log & iterative reweight least square (IRLS)

Hardin & Hilbe (2007)

```
. binreg micro24 intensive, rr nolog
```

Generalized linear models
Optimization : MQL Fisher scoring (IRLS EIM)

No. of obs = 172
Residual df = 170
Scale parameter = 1
(1/df) Deviance = 1.036871
(1/df) Pearson = 1.011763

Deviance = 176.2679883
Pearson = 171.9997346

Variance function: V(u) = u*(1-u)
Link function : g(u) = ln(u)

[Bernoulli]
[Log]

BIC = -698.8061

micro24	Risk Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
intensive	3.021906	.9550499	3.50	0.000	1.626542 5.614312
_cons	.1235955	.0348865	-7.41	0.000	.0710787 .2149146

Simple Logistic Regression: ตัวแปรอิสระ Polychotomous

-ตัวแปรอิสระมีจำนวนกลุ่ม (k) > 2

-เช่น ตัวแปรเชื้อชาติกับการเกิดโรคหัวใจ coronary

$$\ln\left[\frac{p}{1-p}\right] = \hat{\beta}_0 + \sum_{j=1}^{k-1} \beta_{jl} D_{jl}$$

Dummy Variable

ตัวอย่าง โมเดลของ logit กรณีตัวแปรเชื้อชาติ (ขาว=0, ดำ=1
อื่นๆ=2)

$$\ln\left[\frac{p}{1-p}\right] = \hat{\beta}_0 + \beta_1 D_{race=1} + \beta_2 D_{race=2}$$

ตัวแปรเชื้อชาติ (ขาว, ดำ, อื่นๆ) เป็น discrete ให้ทำเป็น
ตัวแปรหุ่น (dummy variables)= k-1 = 3-1 = 2

เชื้อชาติ	D1	D2	
ขาว	0	0	← Reference Cell
ดำ	1	0	
อื่นๆ	0	1	

$$\hat{y} = \alpha + \beta_1 (race_{Black}) + \beta_2 (race_{others})$$

ใช้ software เช่น STATA ระบุ xi: logit y i.race

```
. xi:logit chd i.race
```

i.race _Irace_1-3 (naturally coded; _Irace_1 omitted)

Iteration 0: log likelihood = -12.948933
Iteration 1: log likelihood = -12.787652
Iteration 2: log likelihood = -12.787329
Iteration 3: log likelihood = -12.787329

Logistic regression

Number of obs = 20
LR chi2(2) = 0.32
Prob > chi2 = 0.8508
Pseudo R2 = 0.0125

chd	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
_Irace_2	-.6286087	1.132843	-0.55	0.579	-2.84804 1.591723
_Irace_3	-.2231436	1.204159	-0.19	0.853	-2.583253 2.136966
_cons	.9162907	.83666	1.10	0.273	-.7235328 2.556114

Assessing The fit of The Model

1. Computation and evaluation of overall measures of fit
(summary measures of goodness of fit)

- Pearson Chi-Square/Likelihood Ratio
- Hosmer-Lameshow Test
- classification table
- Area under the receiver operating characteristic curve (ROC curve)
- other summary measure: r^2

2. Logistic regression diagnostics

3 Assessing of fit via external validation (predictive model)

Goodness of Fit

-Pearson Statistics $\chi^2_{Pearson} = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$

-Likelihood ratio Statistics $\chi^2_{LR} = 2 \sum_{i=1}^n y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right)$

-Hosmer-Lemeshow statistics

H* แบ่งข้อมูลเป็น 10 ส่วนเท่า ๆ กัน (ตามความน่าจะเป็น)

$$H^* = \sum_{k=1}^g \frac{(o'_k - e'_k)^2}{e'_k}$$

```
. logit chd age
Iteration 0: log likelihood = -12.948933

Iteration 4: log likelihood = -8.9782115
Logistic regression
```

	Number of obs	LR chi2(1)	Prob > chi2	Pseudo R2
	= 20	= 7.94	= 0.0048	= 0.3066

```
Log likelihood = -8.9782115
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
age	.1425769	.0629517	2.26	0.024	.0191939 .2659599
_cons	-6.162981	2.962012	-2.08	0.037	-11.96842 -.3575447

```
. estat gof
Logistic model for chd, goodness-of-fit test
number of observations = 20
number of covariate patterns = 17
Pearson chi2(15) = 20.68
Prob > chi2 = 0.1473

. estat gof,group(2)
Logistic model for chd, goodness-of-fit test
(Table collapsed on quantiles of estimated probabilities)
number of observations = 20
number of groups = 2
Hosmer-Lemeshow chi2(0) = 0.00
Prob > chi2 = .
```

❖ Classification table

```
. lstat
Logistic model for chd
```

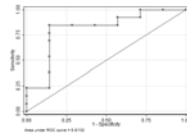
Classified	True	-D	Total
+	11	1	12
-	2	6	8
Total	13	7	20

Classified + if predicted Pr(D) >= .5
True D defined as chd != 0

Sensitivity	Pr(+ D)	84.62%
Specificity	Pr(- D)	85.71%
Positive predictive value	Pr(D +)	91.67%
Negative predictive value	Pr(D -)	75.00%
False + rate for true -D	Pr(+ -D)	14.29%
False - rate for true D	Pr(- D)	15.38%
False + rate for classified +	Pr(D +)	8.33%
False - rate for classified -	Pr(D -)	25.00%
Correctly classified		85.00%

❖ Area under the receiver operating characteristic curve (ROC curve)

```
. lroc
Logistic model for chd
number of observations = 20
area under ROC curve = 0.8132
```



Other Summary Measure

- Measures R^2
- McFadden's Pseudo R^2 , Efron's Pseudo R^2 etc.

$$McFadden's Pseudo R^2 = R_{mf}^2 = 1 - \frac{L_p}{L_0}$$

L_0 = log likelihood for models containing only the intercept

L_p = log likelihood for models containing only the intercept

plus the p covariate

$$Efron's Pseudo R^2 = R_{ef}^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{\pi}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Hosmer & Lemeshow, (2000 p 167)

- Do not Recommend routine publishing of R^2
- However, may be helpful in model building stage.

```
. logit chd age
Iteration 0: log likelihood = -12.948933
Iteration 1: log likelihood = -9.1993409
Iteration 2: log likelihood = -8.9848933
Iteration 3: log likelihood = -8.9782217
Iteration 4: log likelihood = -8.9782115

Logistic regression
```

	Number of obs	LR chi2(1)	Prob > chi2	Pseudo R2
	= 20	= 7.94	= 0.0048	= 0.3066

```
Log likelihood = -8.9782115
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
age	.1425769	.0629517	2.26	0.024	.0191939 .2659599
_cons	-6.162981	2.962012	-2.08	0.037	-11.96842 -.3575447

```
. di 1-((-8.9782115)/(-12.948933))
.30664469
```

$$McFadden's Pseudo R^2 = R_{mf}^2 = 1 - \frac{L_p}{L_0}$$

. fitstat

Measures of Fit for logit of chd

Log-Lik Intercept Only:	-12.949	Log-Lik Full Model:	-8.978
D(18):	17.956	LR(1):	7.941
		Prob > LR:	0.005
McFadden's R2:	0.307	McFadden's Adj R2:	0.152
Maximum Likelihood R2:	0.328	Cragg & Uhler's R2:	0.451
McKelvey and Zavoina's R2:	0.463	Efron's R2:	0.377
Variance of y*:	6.122	Variance of error:	3.290
Count R2:	0.850	Adj Count R2:	0.571
AIC:	1.098	AIC*n:	21.956
BIC:	-35.967	BIC':	-4.946