Lecture 4: Logistic Regression (II. Inference and Prediction with More Examples)

Lin Chen

Department of Public Health Sciences
The University of Chicago

Example 1: *Mice exposed to cigarette smoke*

Table 1: Number of mice developing lung tumors when exposed or not exposed to cigarette smoke.

Group	Tumor present	Tumor absent	Total	$\hat{p}_i = y_i/n_i$
Exposed	21	2	23	21/23
Non-exposed	19	13	32	19/32

- Model: $\text{logit}(p) = \beta_0 + \beta_1 X$, where $X = \begin{cases} 1 & \textit{exposed} \\ 0 & \textit{non-exposed} \end{cases}$
- We designate group 1 as the exposed group $(X = 1)$ and group 0 $(X = 0)$ as the non-exposed group. This model says
  - $\text{logit}(p_1) = \beta_0 + \beta_1,\ \text{logit}(p_0) = \beta_0$
  - β_0: log odds of developing lung tumors in the non-exposed group
  - $\beta_1 + \beta_0$: log odds of developing lung tumors in the exposed group

  Taking the difference in log odds gives us log odds ratio.... β_1 exposed/unexposed
  - $\beta_1 = \text{logit}(p_1) - \text{logit}(p_0) = \log(\frac{p_1}{1-p_1} / \frac{p_0}{1-p_0})$: log odds ratio of developing lung tumors under exposure vs. non-exposure $e^{\beta_1} = OR$
  - Testing difference in p $H_0 : p_1 = p_0$ is equivalent to test $H_0 : \beta_1 = 0$

Ways to represent smoked mice data I (*Treat data as Bernoulli*) Binary

- I. List the outcome (y) and group (x) for the entire 55 mice.(long)

outcome group

y	x	
1	1	} 21
⋮	⋮	
1	1	
0	1	} 2
⋮	⋮	
0	1	
1	0	} 19
⋮	⋮	
1	0	
0	0	} 13
⋮	⋮	
0	0	

You can read in the data, or, copy/paste the following code. The code will create the data table. Model estimation follows.

```
. clear
. set obs 55
. generate outcome  = 0
. replace outcome = 1 in 1/21
. replace outcome = 1 in 24/42
. generate exposed = 0
. replace exposed = 1 in 1/23
. logistic outcome exposed, coef
```

```
Logistic regression                               Number of obs     =         55
                                                  LR chi2(1)        =       7.63
                                                  Prob > chi2       =     0.0057
Log likelihood = -28.409968                       Pseudo R2         =     0.1185

------------------------------------------------------------------------------
     outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     exposed |    1.971886   .8229056     2.40   0.017     .3590202    3.584751
       _cons |    .3794896    .359937     1.05   0.292    -.325974     1.084953
------------------------------------------------------------------------------
```

Note: output coefficients (β, log odds ratio) are provided here, not ORs

Inference of smoked mice data

- The estimated log odds in the non-exposed group $\hat{\beta}_0 = 0.3795$, $\log(\frac{\hat{p}_0}{1-\hat{p}_0}) = 0.3795$, $\hat{p}_0 = \exp(0.3795)/(1+\exp(0.3795)) = 19/32 = 0.5938$
- The estimated log odds in the exposed group $\log(\frac{\hat{p}_1}{1-\hat{p}_1}) = \hat{\beta}_0 + \hat{\beta}_1 = 0.3795 + 1.9718 = 2.3514$, $\hat{p}_1 = \exp(2.3514)/(1+\exp(2.3514)) = 21/23 = 0.9130$

- Comparing exposed versus non-exposed group:
  - Method 1: Test log odds ratio being zero $H_0 : \beta_1 = 0$. Take the difference of log odds, $\hat{\beta}_1 = 1.9719$ is the log odds ratio comparing exposed versus non-exposed group.
  - Method 2: Test odds ratio being 1. Test $H_0 : \exp(\beta_1) = 1$.

Prediction in Stata

One may make prediction of probability of outcome, $\hat{p}$. After `logistic` function, by default the probability will be predicted or use the option `pr`.

```
. predict phat
(option pr assumed; Pr(outcome))

. list in 20/28, clean

      outcome   exposed      phat
20.         1         1  .9130435
21.         1         1  .9130435
22.         0         1  .9130435
23.         0         1  .9130435
24.         1         0   .59375
25.         1         0   .59375
26.         1         0   .59375
27.         1         0   .59375
28.         1         0   .59375
```

Estimates and se's for linear predictors

The linear predictor (the linear combination of predictors) equals $\beta_0 + \sum_k \beta_k X_k$ for k covariates ($k = 1$ in our model) and it predicts log odds for different samples.

```
. *create linear predictor (call it lp) and its standard error
. predict lp, xb
. predict lp_se, stdp
. list in 20/28, clean
```

	outcome	exposed	phat	lp	lp_se
20.	1	1	.9130435	2.351375	.7400129
21.	1	1	.9130435	2.351375	.7400129
22.	0	1	.9130435	2.351375	.7400129
23.	0	1	.9130435	2.351375	.7400129
24.	1	0	.59375	.3794896	.359937
25.	1	0	.59375	.3794896	.359937
26.	1	0	.59375	.3794896	.359937
27.	1	0	.59375	.3794896	.359937
28.	1	0	.59375	.3794896	.359937

CI for estimated probabilities

Using the linear predictor, we can produce 'by hand' the probabilities and compute a confidence interval using the equation given in Lecture 3 Slide 11.

```
. generate p_hat = exp(lp)/(1+exp(lp))
. gen lb = lp - invnormal(0.975)*lp_se
. gen ub = lp + invnormal(0.975)*lp_se
. gen plb = exp(lb)/(1+exp(lb))
. gen pub = exp(ub)/(1+exp(ub))

. list in 20/28, clean
```

	outcome	exposed	phat	lp	lp_se	p_hat	lb	ub	plb	pub
20.	1	1	.9130435	2.351375	.7400129	.9130435	.9009767	3.801774	.7111502	.9781567
21.	1	1	.9130435	2.351375	.7400129	.9130435	.9009767	3.801774	.7111502	.9781567
22.	0	1	.9130435	2.351375	.7400129	.9130435	.9009767	3.801774	.7111502	.9781567
23.	0	1	.9130435	2.351375	.7400129	.9130435	.9009767	3.801774	.7111502	.9781567
24.	1	0	.59375	.3794896	.359937	.59375	-.325974	1.084953	.4192205	.7474302
25.	1	0	.59375	.3794896	.359937	.59375	-.325974	1.084953	.4192205	.7474302
26.	1	0	.59375	.3794896	.359937	.59375	-.325974	1.084953	.4192205	.7474302
27.	1	0	.59375	.3794896	.359937	.59375	-.325974	1.084953	.4192205	.7474302
28.	1	0	.59375	.3794896	.359937	.59375	-.325974	1.084953	.4192205	.7474302

Inference of smoke exposed mice data

- We can test whether odds ratio is 1 to examine the difference in tumor development risk comparing smoke exposed and non-exposed group.
  $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$
  z-statistic = $\hat{\beta}_1 / \text{se}(\hat{\beta}_1) = 2.40$, p-value = 0.017
  H_0 is rejected at the $\alpha = 0.05$ level. There is strong evidence that the probabilities of tumor development are different in those two groups.

- OR=7.18. The relative odds of tumor development for smoke exposed group is high.

Ways to represent smoke exposed mice data II (*Treat data as Bernoulli*)

- II. Collapse by outcome-treatment combination (Collapsed)

y	x	count
1	1	21
0	1	2
1	0	19
0	0	13

Here y as the response indicator, x is the treatment variable.

Bernoulli outcome data form

The following code will type in the Bernoulli data into a 2x2 table.

```
clear
input outcome exposure count
1 1 21
0 1 2
1 0 19
0 0 13
end input
```

Repeated counts

To perform logistic regression with binomial data, use the frequency weight `[fweight=count]`.

```
. logistic outcome exposure [fweight=count], coef

Logistic regression                                     Number of obs     =         55
                                                        LR chi2(1)        =       7.63
                                                        Prob > chi2       =     0.0057
Log likelihood = -28.409968                             Pseudo R2         =     0.1185

------------------------------------------------------------------------------
     outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    exposure |    1.971886   .8229056     2.40   0.017     .3590202    3.584751
       _cons |    .3794896    .359937     1.05   0.292    -.325974     1.084953
------------------------------------------------------------------------------
```

Similar as before, we can do prediction after running regression, for each observed case (observed linear predictor values):

```
. predict phat
. predict lp, xb
. predict lp_se, stdp
. generate p_hat = exp(lp)/(1+exp(lp))
. gen lb = lp - invnormal(0.975)*lp_se
. gen ub = lp + invnormal(0.975)*lp_se
. gen plb = exp(lb)/(1+exp(lb))
. gen pub = exp(ub)/(1+exp(ub))

. list
     +-----------------------------------------------------------------------------------------------------------------------+
     | outcome   exposure   count       phat         lp      lp_se      p_hat         lb         ub        plb        pub |
     |-----------------------------------------------------------------------------------------------------------------------|
  1. |       1          1      21   .9130435   2.351375   .7400129   .9130435   .9009767   3.801774   .7111502   .9781567 |
  2. |       0          1       2   .9130435   2.351375   .7400129   .9130435   .9009767   3.801774   .7111502   .9781567 |
  3. |       1          0      19    .59375   .3794896    .359937    .59375   -.325974   1.084953   .4192205   .7474302 |
  4. |       0          0      13    .59375   .3794896    .359937    .59375   -.325974   1.084953   .4192205   .7474302 |
     +-----------------------------------------------------------------------------------------------------------------------+
```

Ways to represent smoke exposed mice data III (*Treat data as Binomial*)

- III. Collapse by exposure group and summarize as responses in n trials

y	n	x
21	23	1
19	32	0

- y = # of successes
  n = # of trials
  x = exposure indicator

$\frac{y}{n} = \hat{p}$

The following code will type in the data in binomial form.

```
clear
input y n x
21 23 1
19 32 0
end input
```

We can perform logistic regression using `blogit`: (Binomial; for "grouped number")

```
. blogit y n x

Logistic regression for grouped data              Number of obs     =         55
                                                  LR chi2(1)        =       7.63
                                                  Prob > chi2       =     0.0057
Log likelihood = -28.409968                       Pseudo R2         =     0.1185

------------------------------------------------------------------------------
     _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
            x |    1.971886   .8229056     2.40   0.017     .3590202    3.584751
        _cons |    .3794896    .359937     1.05   0.292    -.325974    1.084953
------------------------------------------------------------------------------
```

Note that the `glm` function could run the same regression, with slightly different output format.

Generalized Linear Model – Binomial outcome form

outcome predictor

```
. glm y x, family(binomial n) link(logit)

Iteration 0:   log likelihood = -3.2108785
Iteration 1:   log likelihood = -3.2106672
Iteration 2:   log likelihood = -3.2106671

Generalized linear models                         Number of obs   =          2
Optimization     : ML                             Residual df     =          0
                                                  Scale parameter =          1
Deviance         =  2.50027e-17                   (1/df) Deviance =          .
Pearson          =  2.50027e-17                   (1/df) Pearson  =          .

Variance function: V(u) = u*(1-u/n)               [Binomial]
Link function    : g(u) = ln(u/(n-u))             [Logit]

                                                  AIC             =   5.210667
Log likelihood   = -3.210667147                   BIC             =   2.50e-17

------------------------------------------------------------------------------
             |                 OIM
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   1.971886   .8229056     2.40   0.017     .3590202    3.584751
       _cons |   .3794896    .359937     1.05   0.292    -.325974     1.084953
------------------------------------------------------------------------------
```

Prediction

Similar as before, still we can make prediction and calculate CIs for the observations in the data.

```
. predict yhat
. predict lp, xb
. predict lp_se, stdp
. generate p_hat = exp(lp)/(1+exp(lp))
. gen lb = lp - invnormal(0.975)*lp_se
. gen ub = lp + invnormal(0.975)*lp_se
. gen plb = exp(lb)/(1+exp(lb))
. gen pub = exp(ub)/(1+exp(ub))

. list
```

	y	n	x	yhat	lp	lp_se	p_hat	lb	ub	plb	pub
1.	21	23	1	21	2.351375	.7400129	.9130435	.9009767	3.801774	.7111502	.9781567
2.	19	32	0	19	.3794896	.359937	.59375	−.325974	1.084953	.4192205	.7474302

Example 2: *Aircraft fasteners – a model with a continuous predictor*

Table 2: This is a study on the compressive strength of an alloy fastener used in the construction of aircraft. This table displays the number of fasteners failing out of a number subjected to varying pressure loads.

predictor n y Response prop. of failing y/n

Load (psi)	Sample size	Number failing
2500	50	10
2700	70	17
2900	100	30
3100	60	21
3300	40	18
3500	85	43
3700	90	54
3900	50	33
4100	80	60
4300	65	51

Example 2: *Aircraft fasteners*

- Model: $\text{logit}(p) = \beta_0 + \beta_1 X$, where p is the probability of a fastener failing and X is the predictor variable: load.
- β_0 is the log odds of a fastener failing with `load` being zero, $X = 0$.
- β_1 describes the increase in log-odds for a one unit increase in X.

$$
\begin{aligned}
&\text{logit}(p_{x+1}) - \text{logit}(p_x) \\
&= \log\left(\frac{p_{x+1}}{1-p_{x+1}} \Big/ \frac{p_x}{1-p_x}\right) \\
&= (\beta_0 + \beta_1(x+1)) - (\beta_0 + \beta_1 x) \\
&= \beta_1
\end{aligned}
$$

Note that here $p(\cdot)$ is a function, and $p(x)$ is the probability when X takes the value x.

Data is in collapsed binomial form

```
. use "Aircraft_fastener.dta"
. list
```

	load	ntotal	nfail
1.	2500	50	10
2.	2700	70	17
3.	2900	100	30
4.	3100	60	21
5.	3300	40	18
6.	3500	85	43
7.	3700	90	54
8.	3900	50	33
9.	4100	80	60
10.	4300	65	51

Inference of *Aircraft fasteners data*

```
. blogit nfail ntotal load
Logistic regression for grouped data              Number of obs   =        690
                                                  LR chi2(1)      =     112.46
                                                  Prob > chi2     =     0.0000
Log likelihood = -421.85596                       Pseudo R2       =     0.1176

------------------------------------------------------------------------------
    _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        load |   .0015484   .0001575     9.83   0.000     .0012397    .0018572
       _cons |  -5.339711   .5456932    -9.79   0.000    -6.409251   -4.270172
------------------------------------------------------------------------------
```

- $\hat{\beta}_0 = -5.340$ with a 95% CI of (-6.409, -4.270); $\hat{\beta}_1 = 0.00155$ with a 95% CI of (0.00124, 0.00186).
- The model for the relationship between the estimated probability of a fastener failing, $\hat{p}$, and a load of x psi is

$$\text{logit}(\hat{p}) = -5.340 + 0.00155X$$

The fitted probability of a fastener failing at the i^{th} load of $X = x$ is

$$\hat{p}_i = \frac{\exp(-5.340 + 0.00155x)}{1 + \exp(-5.340 + 0.00155x)}$$

Aircraft fasteners data: Predict a new observation

- Suppose we want to predict the probability of a fastener failing at a load of 2600 psi and also obtain the 95% CI for the predicted probability.
- Note psi value of 2600 is not in the dataset. We can use the function `lincom`, followed by probability calculation from log odds.

```
. lincom _cons+load*2600

 ( 1)  2600*[_outcome]load + [_outcome]_cons = 0
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -1.313784   .1539704    -8.53   0.000    -1.61556   -1.012008
------------------------------------------------------------------------------

. di exp(-1.313784)/(1+exp(-1.313784)) " " exp(-1.61556)/(1+exp(-1.61556))
" " exp(-1.012008)/(1+exp(-1.012008))
.21185433 .16581811 .26658707
```

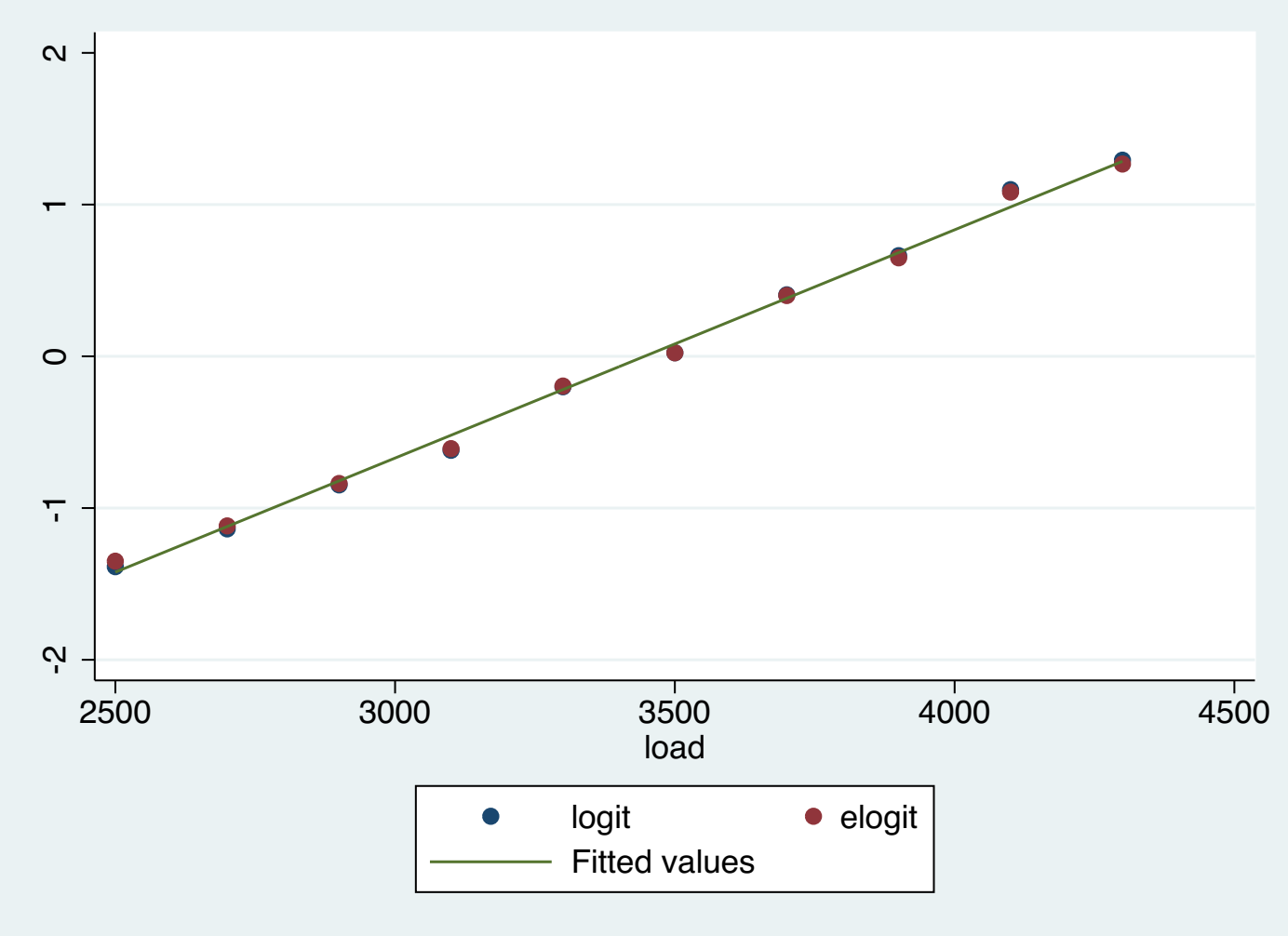
- At a load of 2600 psi, $\text{logit}(\hat{p}) = -5.340 + 0.00155 \times 2600 = -1.314$ with 95%*CI* $(-1.616,\ -1.012)$. Based on log odds, the corresponding $\hat{p} = 0.212$ with a 95% CI of $(0.166,\ 0.267)$.

Plotting empirical logit

- In linear regressions, we often plot Y_i against $\hat{Y}_i$ (i.e., a linear combination of X_i's or X_i if only one predictor) to check model fit. Is there an equivalent plot for logistic models?
- Yes, one can plot the logistic transformation of the observed proportions $\log(\frac{y_i}{n_i - y_i})$ against the fitted linear combination of X_i's to get a visual impression of the adequacy of the model.
- Since the relationship between $\text{logit}(p_i)$ and the values x_i is linear under the assumed logistic regression model, the model is acceptable if the observations follow the fitted straight line relationship.
- In practice, to avoid plotting observations of $0/n_i$ or n_i/n_i, a pragmatic solution is to use the *empirical logit* of the observed proportion y_i/n_i, namely $\log(\frac{y_i+0.5}{n_i - y_i+0.5})$ in constructing the plot.

"observed log OR"

```
. gen logit = log(nfail/(ntotal-nfail))
. gen elogit = log((nfail+0.5)/(ntotal-nfail+0.5))
. twoway (scatter logit load) (scatter elogit load) (lfit elogit load)
```



Looks like a nice fit

Example 3: *Student Smoking* – Multi-level categorical predictor

The table below classifies 5,375 high school students according to the smoking behavior of the student and the smoking behavior of the student's parents.

Student smokes?			
Yes	No	Total number	How many parents smoke?
400	1380	1780	Both $x_2=1, x_1=0$
416	1823	2239	One $x_1=1, x_2=0$
188	1168	1356	Neither $x_1=0, x_2=0$

Model for *Student Smoking data*

- Suppose we pick the baseline category to be "Neither" parent smokes. We will create two dummy variables X_1 and X_2.
- Model: $\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, where

$$X_1 = \begin{cases} 1 & \textit{One parent} \\ 0 & \textit{otherwise} \end{cases}, \quad X_2 = \begin{cases} 1 & \textit{Both parent} \\ 0 & \textit{otherwise} \end{cases}$$

- β_0 = log odds of student smoking given "Neither" parent smokes (Baseline)
  β_1 = log odds ratio of student smoking comparing "One" vs. "Neither" group (Baseline)
  β_2 = log odds ratio of student smoking level "Both" vs. "Neither" (Baseline) group

```
. use "Student_Smoking.dta"
. list

     +-----------------------------+
     |   No   Total   Parents   Yes |
     |-----------------------------|
  1. | 1380    1780      Both   400 |
  2. | 1823    2239       One   416 |
  3. | 1168    1356   Neither   188 |
     +-----------------------------+
```

- Sometimes, you may have a variable that is a string (non-numeric) variable, such as `Parents` in this data. Stata cannot analyze a string variable. We need to create a factor variable for it to conduct analysis.
- Without or with labeling group.

```
. * egen group = group(Parents)
. * or
. egen group = group(Parents), label

. list
     +-----------------------------------------+
     |   No   Total   Parents   Yes     group |
     |-----------------------------------------|
  1. | 1380    1780      Both   400      Both |
  2. | 1823    2239       One   416       One |
  3. | 1168    1356   Neither   188   Neither |
     +-----------------------------------------+
```

- Now group is a factor variable and you may include `i.group` as a predictor variable in the regression
- You may run `blogit Yes Total i.group`. But you cannot choose which one category being the reference group.

Another way is to create dummy variables for all categories.

```
. clear
. use "Student_Smoking.dta"

. tabulate Parents, generate(g)
```

Parents	Freq.	Percent	Cum.
Both	1	33.33	33.33
Neither	1	33.33	66.67
One	1	33.33	100.00
Total	3	100.00	

```
. list
```

Both Neither One

	No	Total	Parents	Yes	g1	g2	g3
1.	1380	1780	Both	400	1	0	0
2.	1823	2239	One	416	0	0	1
3.	1168	1356	Neither	188	0	1	0

We may run `. blogit Yes Total g1 g3` to get the regression output. The dummy variable for the reference group (g2, Neither) is omitted in the predictor list to avoid collinearity. Note that `g1=Both` and `g3=One`.

But the variable names are confusing. Let's change the variable names to avoid confusion, using the `rename` function followed by *oldName* then *newName*.

```
. rename g1 both_parent
. rename g3 one_parent
. rename g2 neither_parent
. rename Yes kid_smoking
. blogit kid_smoking Total one_parent both_parent
```

Now finally the output:

```
. blogit kid_smoking Total one_parent both_parent
Logistic regression for grouped data              Number of obs     =      5,375
                                                  LR chi2(2)        =      38.37
                                                  Prob > chi2       =     0.0000
Log likelihood = -2569.0722                       Pseudo R2         =     0.0074

------------------------------------------------------------------------------
    _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  one_parent |    .3490526    .095539     3.65   0.000     .1617995    .5363056
 both_parent |    .5882319   .0969533     6.07   0.000     .3982068    .7782569
       _cons |   -1.826606   .0785832   -23.24   0.000    -1.980626   -1.672586
------------------------------------------------------------------------------
```

- exp(.3490526) = 1.4177. Having one parent smoke, the odds of a student being a smoker increases by nearly 42% compared to no smoking parents

  relative to neither

- exp(.5882319) = 1.801. Having both parents smoke increases the odds of a student being a smoker by about 80% compared to no smoking parents

  relative to neither

One may also rename the variables to avoid confusion.

Example 4: Toxicity of cypermethrin

Table 3: *Toxicity of cypermethrin:* Mortality/Disability of tobacco budworm moths 72 hours after exposure to cypermethrin

Sex of moth	Dose of cypermethrin	Number affected out of 20
Male	1.0	1
	2.0	4
	4.0	9
	8.0	13
	16.0	18
	32.0	20
Female	1.0	0
	2.0	2
	4.0	6
	8.0	10
	16.0	12
	32.0	16

Note that the predictor dose doubles for each group/experiment. For non-additive increasing variables, skewed variables, variables vary on a relative scale, you may consider a log transformation.

Ex. Toxicity of cypermethrin to moths

```
. use "Budworm.dta"
. list
```

	sex	dose	y	n
1.	1	1	1	20
2.	1	2	4	20
3.	1	4	9	20
4.	1	8	13	20
5.	1	16	18	20
6.	1	32	20	20
7.	2	1	0	20
8.	2	2	2	20
9.	2	4	6	20
10.	2	8	10	20
11.	2	16	12	20
12.	2	32	16	20

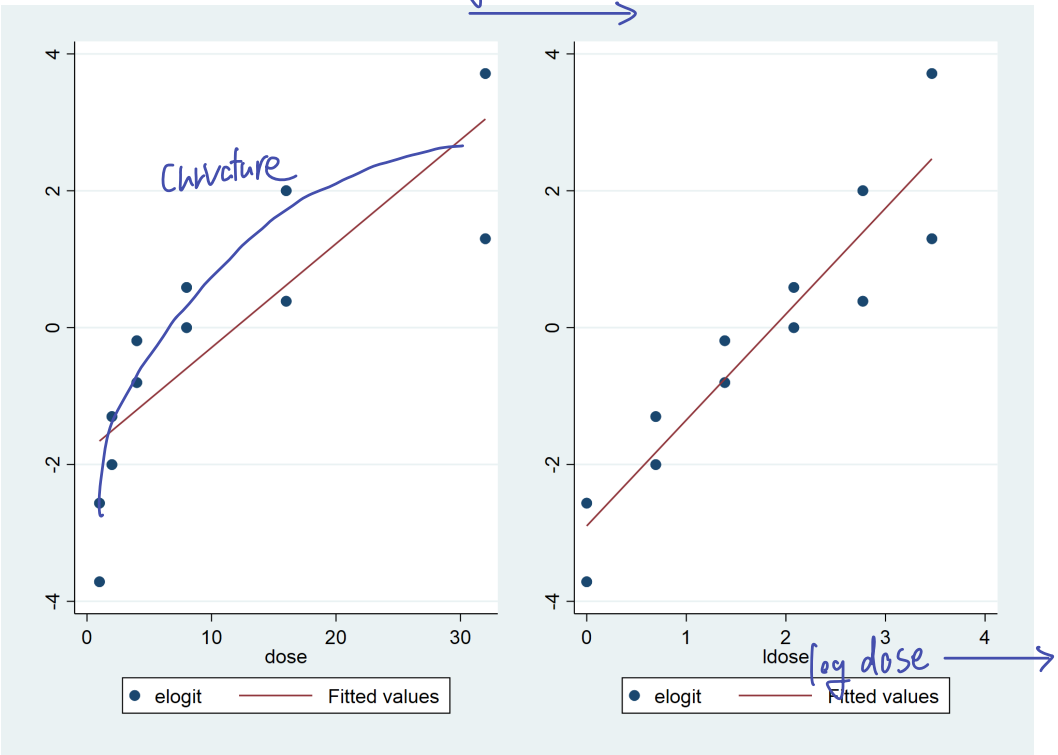
```
. gen female = sex-1
```

Compare the empirical logit of dose or log(dose) as predictor

```
gen ldose = log(dose)

gen elogit = log((y+.5)/(n-y+.5))
graph twoway scatter elogit dose || lfit elogit dose, name(g1) nodraw
graph twoway scatter elogit ldose || lfit elogit ldose, name(g2) nodraw
graph combine g1 g2
```

log transformation



log dose → Better for predictor

Example 4: Toxicity of cypermethrin – The log(dose) model

```
. blogit y n ldose sex

Logistic regression for grouped data                  Number of obs      =       240
                                                      LR chi2(2)         =    118.12
                                                      Prob > chi2        =    0.0000
Log likelihood = -106.62042                           Pseudo R2          =    0.3565

------------------------------------------------------------------------------
     _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
       ldose |    1.535336   .1891048     8.12   0.000     1.164698    1.905975
         sex |   -1.100743   .3558271    -3.09   0.002    -1.798152   -.403335
       _cons |   -1.271669   .5752841    -2.21   0.027    -2.399205   -.1441325
------------------------------------------------------------------------------
```

How to interpret the $\hat{\beta}_{\text{ldose}}$?

Normally, we interpret it as when `ldose` increases by 1 unit, the log odds ratio estimate is $\hat{\beta}_{\text{ldose}} = 1.535$.

What does it mean by `ldose+1`? dose → dose × e ≈ 2.7 dose ⇒ Δ = 1.7 dose ⇒ 170% dose ⇔ (dose + 1)

$\log(\text{dose}) + 1 = \log(\text{dose}) + \log e = \log(\text{dose} \times e)$, where $e \approx 2.7$.

The odds ratio is exp(1.535) when dose increases by about 170%.

Example 4: Toxicity of cypermethrin – The log(dose) model

```
. blogit y n ldose sex
```

_outcome	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ldose	1.535336	.1891048	8.12	0.000	1.164698	1.905975
sex	−1.100743	.3558271	−3.09	0.002	−1.798152	−.403335
_cons	−1.271669	.5752841	−2.21	0.027	−2.399205	−.1441325

Now if we double the dose, $\log(2 \times \text{dose}) = \log(\text{dose}) + \log(2)$.
When log(dose) increases by $\log(2)$, the log odds are
$\text{logit}(\hat{p}_{\text{new}}) = \hat{\beta}_0 + \hat{\beta}_1(\text{ldose} + \log(2)) + \hat{\beta}_2 \text{sex}$, and
$\text{logit}(\hat{p}_{\text{old}}) = \hat{\beta}_0 + \hat{\beta}_1 \text{ldose} + \hat{\beta}_2 \text{sex}$, respectively. The difference (i.e., log odds ratio) is $\text{logit}(\hat{p}_{\text{new}}) - \text{logit}(\hat{p}_{\text{old}}) = \log(2)\hat{\beta}_1$.

The OR is $\exp\left(\log(2)\hat{\beta}_1\right) = \left(\exp(\log(2))\right)^{\hat{\beta}_1} = 2^{\hat{\beta}_1}$. how much the predictor increases by:
When interpreting the coefficient for log-transformed predictor X, calculate the odds ratio when the original predictor X increases by a factor (or %), i.e., when log$X$ increases by certain units. Interpretation of model parameters should be on the original scale (dose).
not log scale

Summary

Logistic Regression

- We discussed logistic regression models for Bernoulli and Binomial outcomes with a single predictor, estimation and interpretations of coefficients, inference, and prediction
- Logistic regression models readily extend to continuous covariates, discrete covariates with more than two levels
- Next, we will discuss model fit, model comparisons in multiple logistic regression