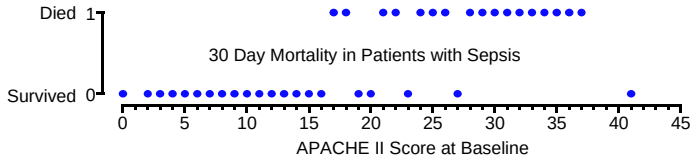
III. INTRODUCTION TO LOGISTIC REGRESSION

- Simple logistic regression: Assessing the effect of a continuous variable on a dichotomous outcome
- How logistic regression parameters affect the probability of an event
- Probability, odds and odds ratios
- Generalized linear models: The relationship between linear and logistic regression
- Confidence intervals for proportions
- Plotting probability of death with 95% confidence bands as a function of a continuous risk factor
- Review of classic 2x2 case-control studies
- Analyzing case-control studies with logistic regression

1. Simple Logistic Regression

a) Example: APACHE II Score and Mortality in Sepsis

The following figure shows 30 day mortality in a sample of septic patients as a function of their baseline APACHE II Score. Patients are coded as 1 or 0 depending on whether they are dead or alive in 30 days, respectively.



We wish to predict death from baseline APACHE II score in these patients.

Let $\pi(x)$ be the probability that a patient with score x will die.

Note that linear regression would not work well here since it could produce probabilities less than zero or greater than one.

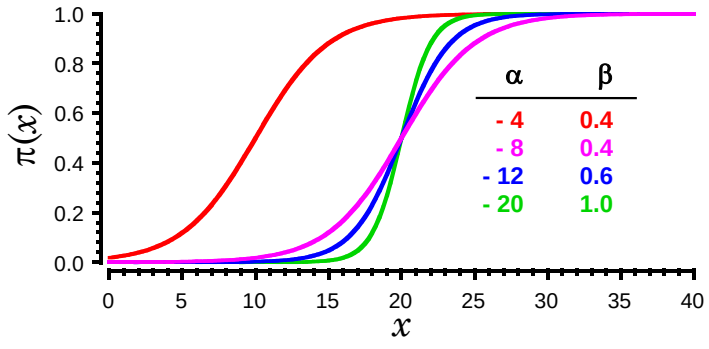
b) Sigmoidal family of logistic regression curves

Logistic regression fits probability functions of the following form:

$$\pi(x) = \exp(\alpha + \beta x) / (1 + \exp(\alpha + \beta x)) \qquad \{3.1\}$$

This equation describes a family of sigmoidal curves, three examples of which are given below.

$$\pi(x) = \exp(\alpha + \beta x) / (1 + \exp(\alpha + \beta x))$$



c) Parameter values and the shape of the regression curve

For now assume that $\beta > 0$.

For negative values of x, $\exp(\alpha+\beta x) \to 0$ as $x \to -\infty$
and hence $\pi(x) \to 0/(1+0) = 0$

For very large values of x, $\exp(\alpha+\beta x) \to \infty$ and hence
$\pi(x) \to \infty/(1+\infty) = 1$

When $x = -\alpha/\beta$, $\alpha+\beta x = 0$ and hence $\pi(x) = 1/(1+1) = 0.5$

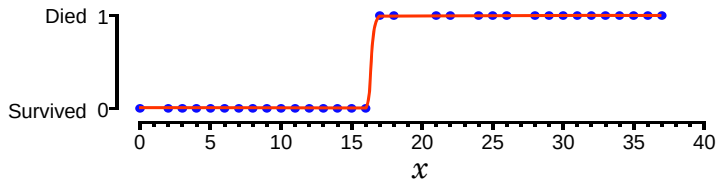
The slope of $\pi(x)$ when $\pi(x)$=.5 is $\beta/4$.

Thus β controls how fast $\pi(x)$ rises from 0 to 1.

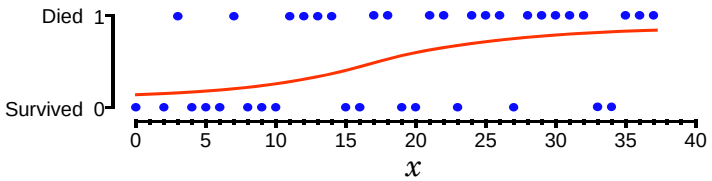
For given β, α controls were the 50% survival point is located.

We wish to choose the best curve to fit the data.

Data that has a sharp survival cut off point between patients who live or die should have a large value of β.



Data with a lengthy transition from survival to death should have a low value of β.



d) The probability of death under the logistic model

This probability is

$$\pi(x) = \exp(\alpha + \beta x) / (1 + \exp(\alpha + \beta x))$$

Hence $1 - \pi(x) =$ probability of survival

$$= \frac{1 + \exp(\alpha + \beta x) - \exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

$$= 1/(1 + \exp(\alpha + \beta x))$$, and the odds of death is

$$\pi(x)/(1 - \pi(x)) = \exp(\alpha + \beta x)$$

The log odds of death equals

$$\log(\pi(x)/(1 - \pi(x)) = \alpha + \beta x \qquad \textbf{\{3.2\}}$$

e) The logit function

For any number π between 0 and 1 the logit function is defined by

$$\text{logit}(\pi) = \log(\pi / (1 - \pi))$$

Let $d_i = \begin{cases} 1\text{: } i^{\text{th}} \text{ patient dies} \\ 0\text{: } i^{\text{th}} \text{ patient lives} \end{cases}$

x_i be the APACHE II score of the i^{th} patient

Then the expected value of d_i is

$$E(d_i) = \pi(x_i)$$

Thus we can rewrite the **logistic regression equation** {3.1} as

$$\text{logit}(E(d_i)) = \alpha + \beta x_i \qquad \textbf{\{3.3\}}$$

2. The Binomial Distribution

Let

m be the number of people at risk of death

d be the number of deaths

π be the probability that any patient dies.

The death of one patient has no effect on any other.

Then d has a **binomial distribution** with

parameters m and π,

mean $m\pi$, and

variance $m\pi(1-\pi)$.

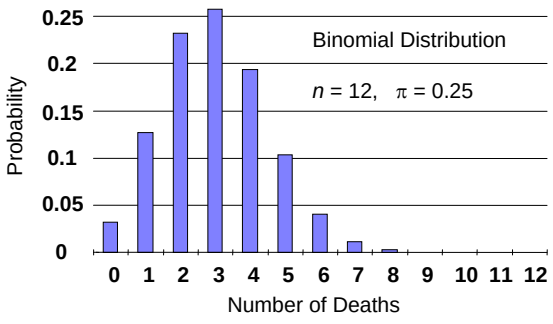
Pr[d deaths]

$$= \frac{m!}{(m-d)!d!} \quad \pi^d(1-\pi)^{(m-d)} \quad : d = 0,1,\cdots,m \qquad \{3.4\}$$

The population mean of any random variable x is also equal to its expected value and is written $E(x)$. Hence

$E(d) = \pi m$ and $E(d/m) = \pi$

For $m = 12$ and $\pi = 0.25$ this distribution is as follows.



A special case of the binomial distribution is when $m = 1$, which is called a Bernoulli distribution.

In this case we can have 0 or 1 deaths with probability 1-π and π, respectively.

The complete logistic regression model for the sepsis data is specified as follows

d_i has a binomial distribution with 0 or 1 failures and probability of failure $\pi(x_i) = E(d_i)$

$E(d_i)$ is determined by $\text{logit}(E(d_i)) = \alpha + \beta x_i$

3. Generalized Linear Models

Logistic regression is an example of a generalized linear model. These models are defined by three attributes: The distribution of the model's random component, its linear predictor, and its link function. For logistic regression these are defined as follows.

a) The random component

d_i *is* the **random component** of the model. In logistic regression, d_i has a binomial distribution obtained from m_i trials with mean $E(d_i)$. (In the sepsis example, $m_i = 1$ for all i.)

Stata refers to the distribution of the random component as the distributional family.

b) The linear predictor

$\alpha + x_i\beta$ is called the **linear predictor**

c) The link function

$E(d_i)$ is related to the linear predictor through a **link function**. Logistic regression uses a logit link function

$$\text{logit}(E(d_i)) = \alpha + x_i\beta$$

4. Contrast Between Logistic and Linear Regression

In linear regression, the expected value of y_i given x_i is

$$E(y_i) = \alpha + \beta x_i \text{ for } i = 1,2,\ldots,n$$

y_i	has a normal distribution with standard deviation σ.
y_i	is the random component of the model, which has a normal distribution.
$\alpha + \beta x_i$	is the linear predictor.

The link function is the identity function $E(y_i) = I(E(y_i))$.

5. Maximum Likelihood Estimation

In linear regression we used the method of least squares to estimate regression coefficients.

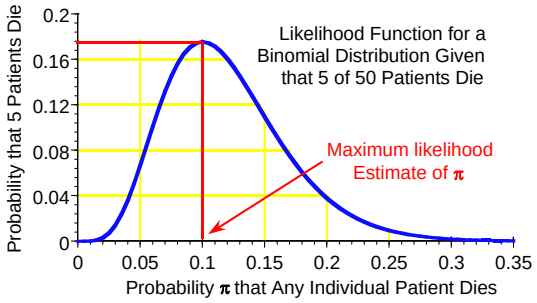
In generalized linear models we use another approach called maximum likelihood estimation.

Suppose that 5 of 50 AIDS patients die in one year. We wish to estimate π, the probability of death for these patients.

We assume that the number of deaths has a binomial distribution obtained from m= 50 patients with probability of death π for each patient.

Let $L(\pi \mid d = 5)$ be the probability of the observed outcome (5) given different values of π.

$L(\pi \mid d = 5)$ is called a likelihood function and looks like this.



The maximum likelihood estimate of π is the value of π that assigns the greatest probability to the observed outcome.

In this example, $\hat{\pi} = 0.1$

Note that if $\pi = \hat{\pi} = 0.1$ that $E(d) = 50 \times 0.1 = 5 = d$

Thus, in this example, the maximum likelihood estimate of π is that value that sets the expected number of deaths equal to the observed number of deaths.

In general, maximum likelihood estimates do not have simple closed solutions, but must be solved interactively using numerical methods. This, however, is not a serious drawback given ubiquitous and powerful desktop computers.

a) Variance of maximum likelihood parameter estimates

It can be shown that when a maximum likelihood estimate is based on large number of patients, its variance is approximately equal to

-1/C, where C is the expected curvature of the likelihood surface at $\hat{\pi}$

6. Logistic Regression with glm

a) Example: APACHE II score and mortal outcome

```
. *  4.11.Sepsis.log
. *
. *  Simple logistic regression of mortal status at 30 days (fate) against
. *  baseline APACHE II score (apache) in a random sample of septic patients
. *
. use C:\\WDDtext\4.11.Sepsis.dta, clear
. summarize fate apache

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        fate |        38    .4473684    .5038966          0          1
      apache |        38    19.55263    11.30343          0         41
```

```
. codebook
apache ---------------------------------------------- APACHE II Score at Baseline
                  type:  numeric (byte)

                 range:  [0,41]                       units:  1
         unique values:  38                   coded missing:  0 / 38

                  mean:   19.5526
             std. dev:   11.3034

           percentiles:        10%       25%       50%       75%       90%
                                4        10      19.5        29        35



fate ------------------------------------------------- Mortal Status at 30 Days
                  type:  numeric (byte)
                 label:  fate

                 range:  [0,1]                        units:  1
         unique values:  2                    coded missing:  0 / 38

            tabulation:  Freq.   Numeric  Label
                           21         0  Alive
                           17         1  Dead
```

```
. glm fate apache, family(binomial) link(logit)                          {1}

Iteration 0:   log likelihood = -15.398485
Iteration 1:   log likelihood =   -14.9578
Iteration 2:   log likelihood = -14.956086
Iteration 3:   log likelihood = -14.956085

Generalized linear models                         No. of obs      =         38
Optimization     : ML: Newton-Raphson             Residual df     =         36
                                                  Scale param     =          1
Deviance         =  29.91217061                   (1/df) Deviance =  .8308936
Pearson          =  66.34190718                   (1/df) Pearson  =  1.842831

Variance function: V(u) = u*(1-u)                 [Bernoulli]              {2}
Link function    : g(u) = ln(u/(1-u))             [Logit]
Standard errors  : OIM

Log likelihood   = -14.95608531                   AIC             =  .8924255
BIC              = -101.0409311

------------------------------------------------------------------------------
        fate |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
      apache |   .2012365   .0608998      3.304   0.001       .0818752    .3205979
       _cons |  -4.347807   1.371609     -3.170   0.002      -7.036111   -1.659503
------------------------------------------------------------------------------
. predict logodds, xb                                                    {3}
. generate prob = exp(logodds)/(1 + exp(logodds))                        {4}
. list apache fate logodds prob in 1/3                                   {5}
```

{1} This *glm* command regresses ***fate*** against ***apache*** using a generalized linear model. The *family* and *link* options specify that the **random component** of the model is **binomial** and the **link function** is **logit**. In other words, a **logisitic** model is to be used.

{2} When there is only one patient per record Stata refers to the binomial distribution as a **Bernoulli** distribution. Along with the **logit** link function this implies a **logisitc** regression model.

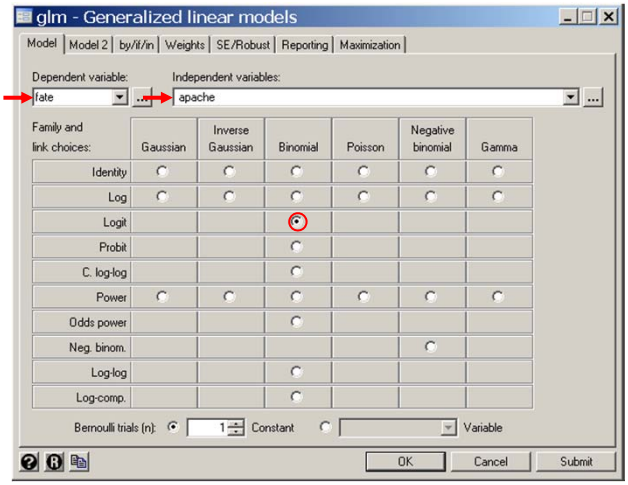
{3} The *xb* option of the ***predict*** command specifies that the **linear predictor** will be evaluated for each patient and stored in a variable named ***logodds.***

Recall that *predict* is a **post estimation** command whose meaning is determined by the latest estimation command, which in this example is *glm*.

{4} ***prob*** equals the estimated **probability** that a patient will **die**. It is calculated using the equation

$$\pi(x) = \exp(\alpha + \beta x) / (1 + \exp(\alpha + \beta x))$$

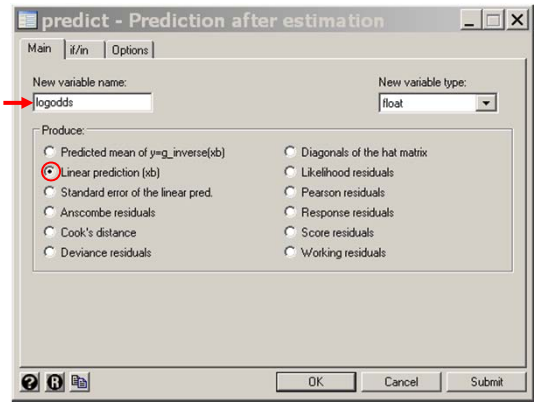
{5} The ***in*** modifier specifies that the first through third record are to be listed.




```
. predict logodds, xb                                                {3}
```

{3} The *xb* option of the ***predict*** command specifies that the **linear predictor** will be evaluated for each patient and stored in a variable named ***logodds.***

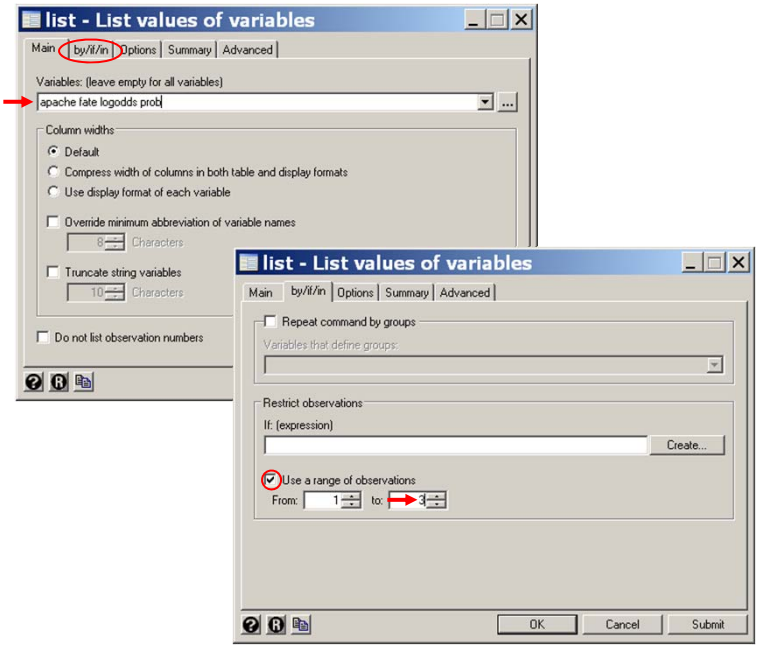
Recall that *predict* is a **post estimation** command whose meaning is determined by the latest estimation command, which in this example is *glm*.



```
. generate prob = exp(logodds)/(1 + exp(logodds))                    {4}
. * Data > Describe data > List data
. list apache fate logodds prob in 1/3                               {5}
```

{4} ***prob*** equals the estimated **probability** that a patient will **die**. It is calculated using equation 3.1.

{5} The ***in*** modifier specifies that the first through third record are to be listed.



	apache	fate	logodds	prob
1.	16	Alive	-1.128022	.2445263
2.	25	Dead	.6831065	.6644317
3.	19	Alive	-.5243126	.3718444

{6}

```
. sort apache
. * Variables Manager
. label variable prob "Probability of Death"
```

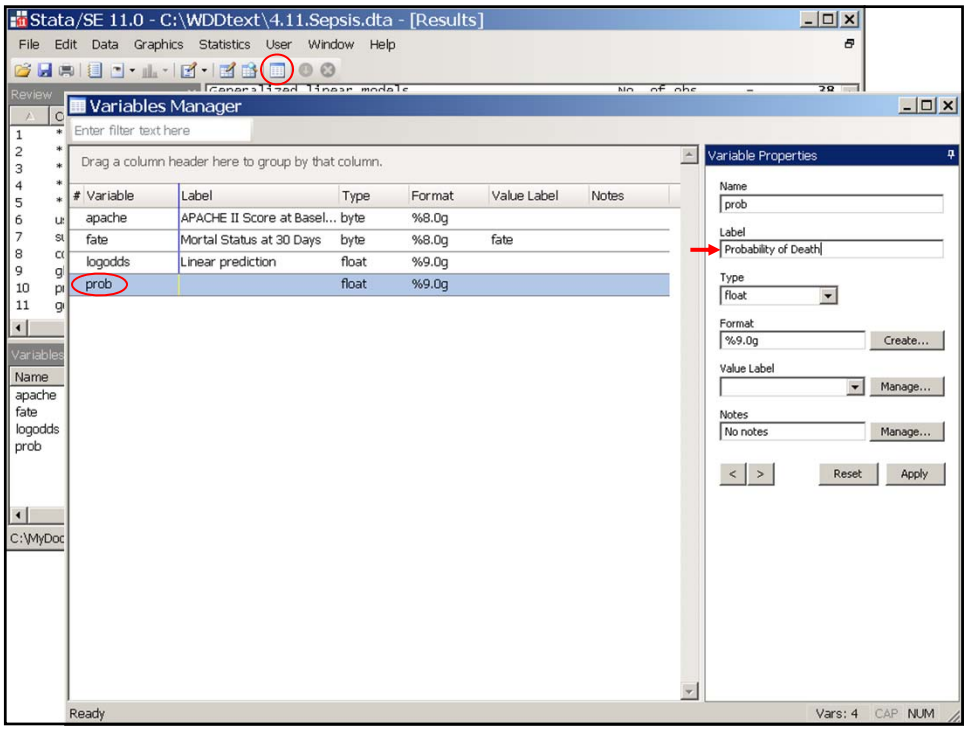
{7}

{7} Assign the label *Probability of Death* to the variable *prob*.

{6} The first patient has an APACHE II score of 16. Hence the estimated linear predictor for this patient is $logodds = \alpha + x_i\beta =$ *_cons* + 16×*apache* = -4.3478 + 16×0.2012 = -1.1286. The second patient has apache = 25 giving logodds = -4.3478 + 25×0.2012 = 0.6831.

For the first patient

$$\text{prob} = \exp(\alpha+\beta x)/(1+\exp(\alpha+\beta x))$$
$$= \exp(logodds)/(1+\exp(logodds))$$
$$= \exp(-1.128)/(1+\exp(-1.128)) = 0.2445$$



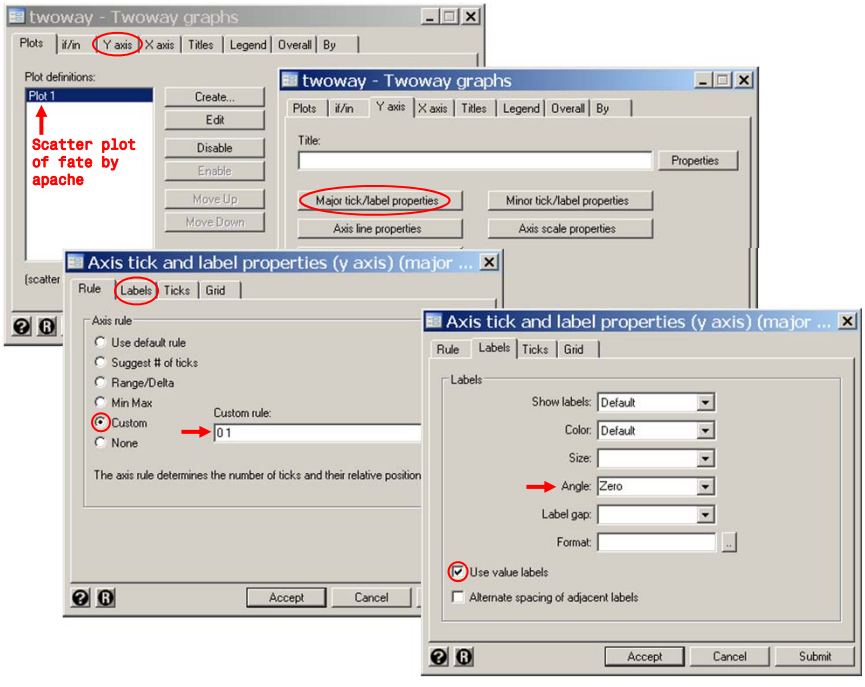
```
. scatter fate apache                                              ///
>     , ylabel(0 1, valuelabel angle(0)) yscale(titlegap(-8)) ///     {8,9}
>     || line prob apache, yaxis(2) xlabel(0(10)40)                    {10}
```

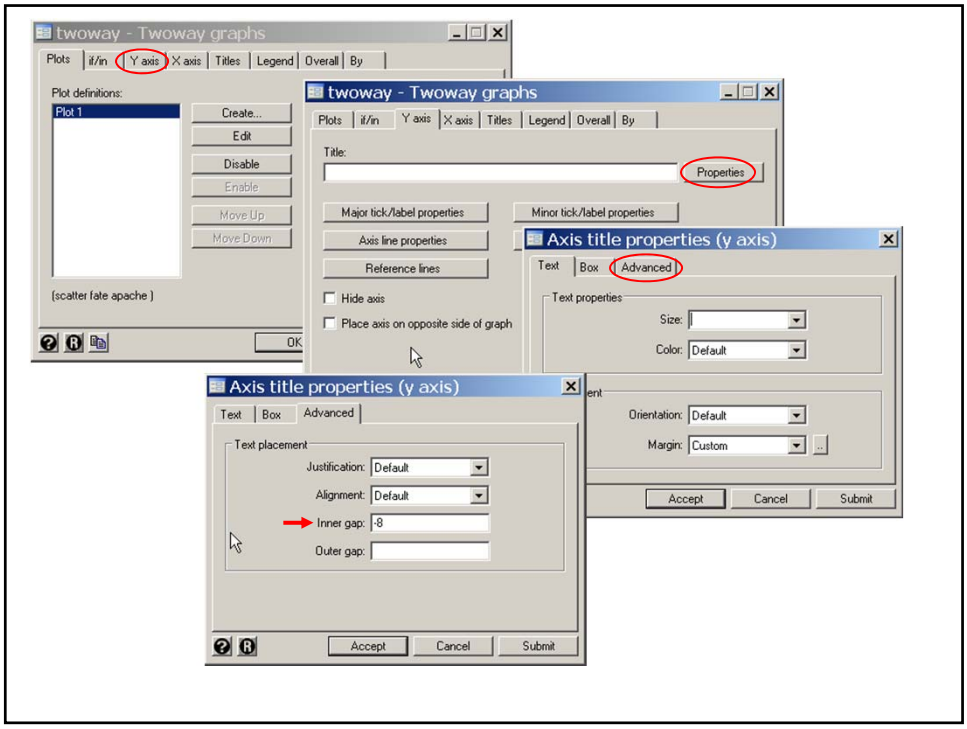
{8} **valuelabel** and **angle** are suboptions of the *ylabel* option. The labels for the *y*-axis are at 0 and 1. *valuelabel* indicates that the value labels of *fate* are to be used. That is, *Alive* and *Dead*.

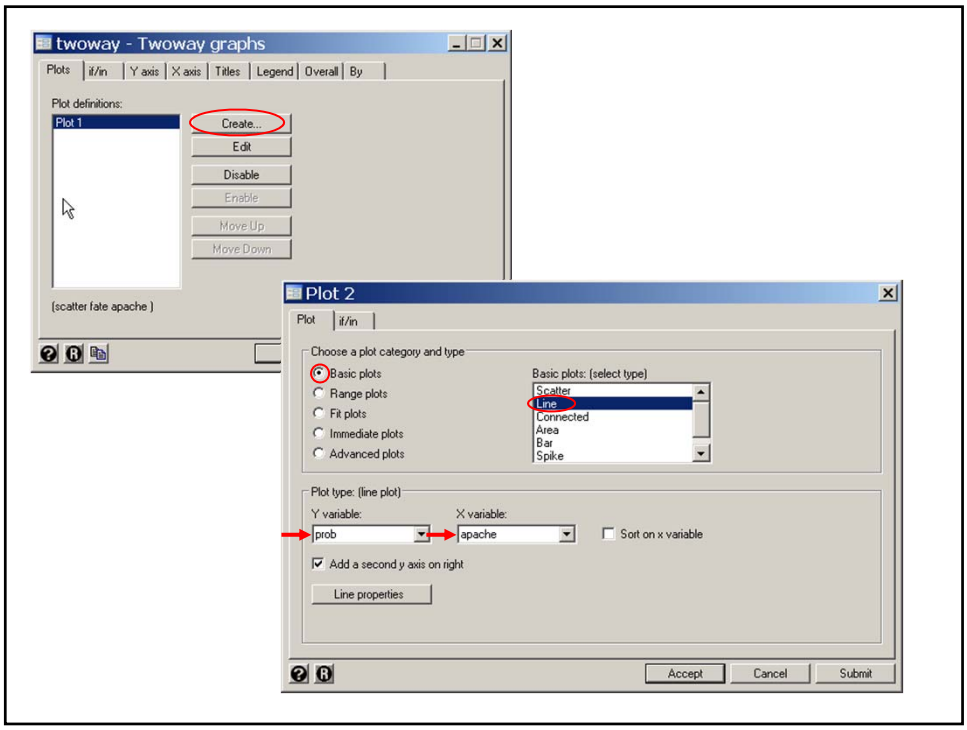
angle specifies the angle at which the labels are written; an angle of 0 means that the labels will be written parallel to the *x*-axis.

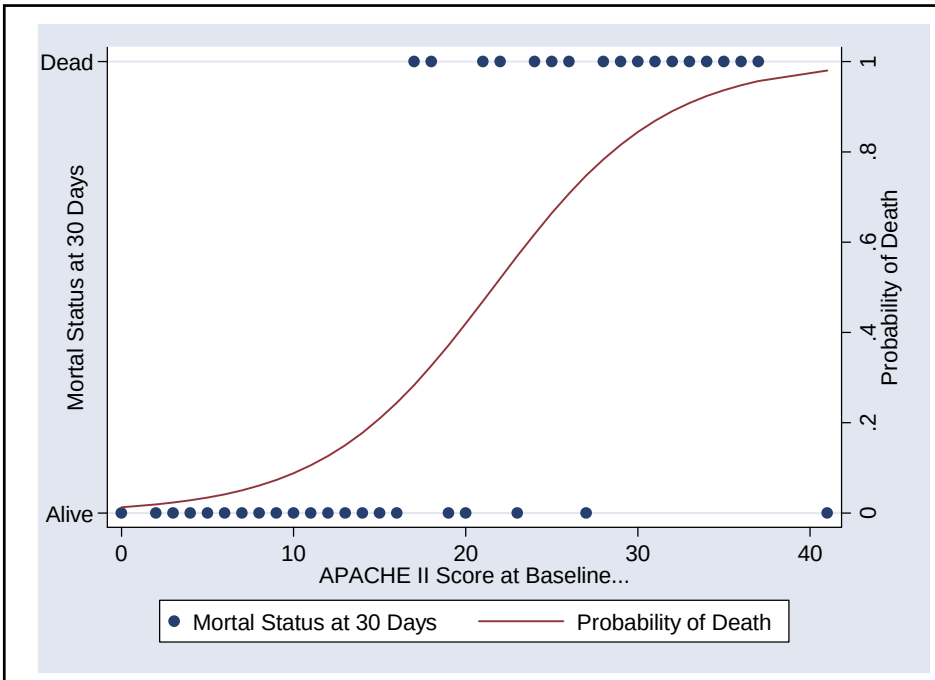
{9} *yscale(titlegap(-8))* specifies how close the title of the *y*-axis is to the axis itself. The default, *titlegap(0)* would place the title just to the left of the labels *Dead* and *Alive*.

{10} **yaxis(2)** indicates that the y-axis for the graph of *prob* vs. *apache* is to be drawn on the right.




twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Create...
Edit
Disable
Enable
Move Up
Move Down
(scatter fate apache)
OK
twoway - Twoway graphs
Title:
Properties
Major tick/label properties
Minor tick/label properties
Axis line properties
Reference lines
Hide axis
Place axis on opposite side of graph
Axis title properties (y axis)
Text
Box
Advanced
Text properties
Size:
Color:
Default
Orientation:
Margin:
Custom
Accept
Cancel
Submit
Axis title properties (y axis)
Text
Box
Advanced
Text placement
Justification:
Default
Alignment:
Default
Inner gap:
-8
Outer gap:
Accept
Cancel
Submit


twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Create...
Edit
Disable
Enable
Move Up
Move Down
(scatter fate apache)
Plot 2
Plot
if/in
Choose a plot category and type
Basic plots
Range plots
Fit plots
Immediate plots
Advanced plots
Basic plots: (select type)
Scatter
Line
Connected
Area
Bar
Spike
Plot type: (line plot)
Y variable:
X variable:
prob
apache
Sort on x variable
Add a second y axis on right
Line properties
Accept
Cancel
Submit



7. Odds Ratios and the Logistic Regression Model

a) Odds ratio associated with a unit increase in *x*

The log odds that patients with APACHE II scores of x and $x + 1$ will die are

$$\text{logit}(\pi(x)) = \alpha + \beta x \qquad \textbf{\{3.5\}}$$

and

$$\text{logit}(\pi(x+1)) = \alpha + \beta(x+1) = \alpha + \beta x + \beta \qquad \textbf{\{3.6\}}$$

respectively.

subtracting {3.5} from {3.6} gives $\beta = \text{logit}(\pi(x+1)) - \text{logit}(\pi(x))$

$$\beta = \text{logit}\,(\pi(x+1)) - \text{logit}(\pi(x))$$

$$= \log\left(\frac{\pi(x+1)}{1-\pi(x+1)}\right) - \log\left(\frac{\pi(x)}{1-\pi(x)}\right)$$

$$= \log\left(\frac{\pi(x+1)/(1-\pi(x+1))}{\pi(x)/(1-\pi(x))}\right)$$

and hence

exp(β) is the odds ratio for death associated with a unit increase in *x*.

A property of logistic regression is that this ratio remains constant for all values of *x*.

8. 95% Confidence Intervals for Odds Ratio Estimates

In our sepsis example the parameter estimate for *apache* (β) was 0.2012 with a standard error or 0.0609. Therefore, the odds ratio for death associated with a unit rise in APACHE II score is

exp(0.2012) = 1.223

with a 95% confidence interval of $\exp(0.2012 \pm 1.96 * 0.0609)$

(1.223exp(-1.96×0.0609), 1.223exp(1.96×0.0609))

= (1.09, 1.38).

9. 95% Confidence Interval for $\pi[x]$

Let $\sigma^2_{\hat{\alpha}}$ and $\sigma^2_{\hat{\beta}}$ denote the variance of $\hat{\alpha}$ and $\hat{\beta}$.
Let $\sigma_{\hat{\alpha}\hat{\beta}}$ denote the covariance between $\hat{\alpha}$ and $\hat{\beta}$.

Then it can be shown that the standard error of is

$$\text{se}\left[\hat{\alpha}+\hat{\beta}x\right] = \sqrt{\sigma^2_{\hat{\alpha}}+2x\sigma_{\hat{\alpha}\hat{\beta}}+x^2\sigma^2_{\hat{\beta}}}$$

A 95% confidence interval for $\alpha+\beta x$ is

$$\hat{\alpha}+\hat{\beta}x \pm 1.96\times\text{se}\left[\hat{\alpha}+\hat{\beta}x\right]$$

A 95% confidence interval for $\alpha+\beta x$ is

$$\hat{\alpha}+\hat{\beta}x \pm 1.96\times\text{se}\left[\hat{\alpha}+\hat{\beta}x\right]$$

Hence, a 95% confidence interval for $\pi[x]$ is $\left(\hat{\pi}_L[x],\hat{\pi}_U[x]\right)$, where

$$\hat{\pi}_L[x]=\frac{\exp\left[\hat{\alpha}+\hat{\beta}x-1.96\times\text{se}\left[\hat{\alpha}+\hat{\beta}x\right]\right]}{1+\exp\left[\hat{\alpha}+\hat{\beta}x-1.96\times\text{se}\left[\hat{\alpha}+\hat{\beta}x\right]\right]}$$

and

$$\hat{\pi}_U[x]=\frac{\exp\left[\hat{\alpha}+\hat{\beta}x+1.96\times\text{se}\left[\hat{\alpha}+\hat{\beta}x\right]\right]}{1+\exp\left[\hat{\alpha}+\hat{\beta}x+1.96\times\text{se}\left[\hat{\alpha}+\hat{\beta}x\right]\right]}$$

10. 95% Confidence Intervals for Proportions

It is useful to be able to estimate a 95% confidence interval for the proportion d_i/m_i in the sepsis study.

Let d be the number of deaths that occur in m patients,

π be the probability that an individual dies..

Then d/m has mean π and standard error $s(\pi) = \sqrt{\pi(1-\pi)/m}$

Estimating π by $\hat{\pi} = d/m$ gives $s(\hat{\pi}) = \sqrt{\hat{\pi}(1-\hat{\pi})/m}$

as the estimated standard error of $\hat{\pi}$

The distribution of $\hat{\pi}$ is approximately normal as long as $\hat{\pi}$ is not too close to 0 or 1 and m is fairly large. This approximation gives a Wald 95% confidence interval for π of

$$\hat{\pi} \pm 1.96 s(\hat{\pi})$$

This estimate works poorly when $\hat{\pi}$ is near 0 or 1. Note that it is possible for this confidence interval to contain values that are less than 0 or greater than 1.

The 100(1-α)% Wald Confidence interval is

$$\hat{\pi} \pm z_{\alpha/2} s(\hat{\pi}) \qquad \text{(recall that } z_{.025} = 1.96\text{)}$$

This interval consists of all π for which

$$-z_{\alpha/2} \le (\hat{\pi} - \pi)/s(\hat{\pi}) \le z_{\alpha/2}$$

Wilson Confidence Interval for a Proportion.

A better 100(1-α)% confidence interval due to Wilson is given by all values of π for which

$$-z_{\alpha/2} \le (\hat{\pi} - \pi)/s(\pi) \le z_{\alpha/2} \qquad \{3.7\}$$

This interval differs from the Wald Interval in that the denominator is $s(\pi)$ rather than $s(\hat{\pi})$. This makes a big difference when π is near 0 or 1.

Equation {3.7} can be rewritten as a complex but unedifying function of d, m and $z_{\alpha/2}$

```
. * proportions.log
. *
. * Illustrate Wald, Wilson and exact confidence intervals
. *
. use proportions.dta
. list

     +-----------------+
     | fate   patients |
     |-----------------|
  1. |    0         10 |
  2. |    1         10 |
     +-----------------+
```

Here is data on 20 patients grouped into two records with 10 patients per record.

Half of these patients live (fate = 0) and the other half die (fate = 1).

```
* Statistics > Summaries, tables ... > Summary ... > Confidence intervals
. ci fate [freq = patients], binomial wald                                  {1}

                                                      -- Binomial Wald ---
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+------------------------------------------------------------------
        fate |         20          .5    .1118034        .2808694    .7191306

. ci fate [freq = patients], binomial wilson                                {2}

                                                      ------ Wilson ------
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+------------------------------------------------------------------
        fate |         20          .5    .1118034         .299298     .700702
```

{1} This **ci** command calculated confidence intervals for the proportion of patients who die (fate = 1). **[freq=patients]** ensures that each record contributes the number of patients indicated by the *patients* variable. (Without this command modifier, each record would count as a single observation.)

binomial specifies that fate is a dichotomous variable. It must be specified whenever Wald or Wilson confidence intervals are required. **wald** indicates that Wald confidence intervals are to be calculated.

{2} **wilson** indicates that Wilson confidence intervals are to be calculated.

These confidence intervals are quite close to each other.

```
. replace patients = 18 in 1
(1 real change made)

. replace patients = 2 in 2
(1 real change made)

. list

     +-----------------+
     | fate   patients |
     |-----------------|
  1. |    0         18 |
  2. |    1          2 |
     +-----------------+
```

Suppose that the mortality rate is 10%

```
. ci fate [freq = patients], binomial wald

                                                      -- Binomial Wald ---
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
        fate |         20          .1     .067082                0    .2314784*

(*) The Wald interval was clipped at the lower endpoint

. ci fate [freq = patients], binomial wilson

                                                      ------ Wilson ------
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
        fate |         20          .1     .067082        .0278665    .3010336
```

The Wald interval is much narrower than the Wilson and would extend below zero if Stata did not clip it at zero.

```
. return list                                                    {3}

scalars:
                  r(ub) =  .3010336452284873
                  r(lb) =  .0278664812137682
                  r(se) =  .0670820393249937
                r(mean) =  .1
                   r(N) =  20

. display r(ub)                                                  {4}
.30103365
```

{3} Stata commands store most of their output were they can be used by other commands. This feature greatly extends the power and flexibility of this software. The **return list** command lists some of these values.

{4} This **display** command displays the upper bound of the confidence interval calculated by the last **ci** command.

Baseline APACHE II Score	Number of Patients	Number of Deaths	Baseline APACHE II Score	Number of Patients	Number of Deaths
0	1	0	20	13	6
2	1	0	21	17	9
3	4	1	22	14	12
4	11	0	23	13	7
5	9	3	24	11	8
6	14	3	25	12	8
7	12	4	26	6	2
8	22	5	27	7	5
9	33	3	28	3	1
10	19	6	29	7	4
11	31	5	30	5	4
12	17	5	31	3	3
13	32	13	32	3	3
14	25	7	33	1	1
15	18	7	34	1	1
16	24	8	35	1	1
17	27	8	36	1	1
18	19	13	37	1	1
19	15	7	41	1	0

Example: APACHE II Score & Mortality in Sepsis

The Ibuprofen and Sepsis Trial contained 454 patients with known baseline APACHE II scores (Bernard et al. 1997). The 30 day mortal outcome for these patients is summarized on the right.

11. Logistic Regression with Grouped Response Data

Suppose that there are m_i patients with covariate x_i.

Let d_i be the number of deaths in these m_i patients.

Then d_i has a binomial distribution with mean $m_i\pi(x_i)$ and hence $E(d_i/m_i) = \pi(x_i)$.

Thus the logistic model becomes

$\text{logit}(E(d_i/m_i)) = \alpha + \beta x_i$

```
. *  4.18.Sepsis.Wilson.log
. *
. *  Simple logistic regression of mortality against APACE score in the
. *  Ibuprofen in Sepsis study (Bernard et al. 1997).  There are two
. *  records in 4.18.Sepsis.Weighted.dta for each observed APACE score.
. *    apache = an APACHE II score at baseline
. *    fate   = 0 or 1 indicating patients who were alive or dead after
. *                    30 days, respectively
. *    n      = number of study subjects with a given fate and APACE score.
. *
. use 4.18.Sepsis.Weighted.dta, clear

. list if apache==6 | apache==7

     +--------------------+
     | apache   fate    n |
     |--------------------|
 11. |      6      0   11 |
 12. |      6      1    3 |
 13. |      7      0    8 |
 14. |      7      1    4 |
     +--------------------+
```

We need to calculate the observed mortality rate and its associated confidence interval for each APACHE score.

There were 37 distinct scores.

We could issue 47 distinct **ci** commands and transcribe the confidence intervals back into Stata.

This would be tedious. Fortunately it is unnecessary.

```
. *
. *  Collapse data to one record per APACHE score.
. *  Calculate observed mortality rate for each score and its
. *  Wilson 95% confidence interval.
. *
. * Statistics > Other > Collect statistics for a command across a by list
. statsby, by(apache): ci fate [freq=n], binomial wilson                  {1}
(running ci on estimation sample)

      command:  ci fate [fweight= n], binomial wilson
           ub:  r(ub)
           lb:  r(lb)
           se:  r(se)
         mean:  r(mean)
            N:  r(N)
           by:  apache

Statsby groups
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
.....................................
```

{1} The **statsby** command can be used in combination with most other Stata commands. It executes the command to the right of the colon for each unique combination of values of the variable(s) specified by the **by** option. This command executes

```
ci fate [freq=n], binomial wilson
```

separately for each unique value of **apache**. The data in memory is replaced by new data with one record for each distinct value of apache. Output from each command is also stored with the indicated variable names.

```
. list if apache==6 | apache==7

     +-------------------------------------------------------+
     | apache         ub         lb         se       mean    N |
     |-------------------------------------------------------|
  6. |      6   .4758923   .0757139   .1096642   .2142857   14 |
  7. |      7   .6093779   .1381201   .1360828   .3333333   12 |          {2}
     +-------------------------------------------------------+
. generate patients = N

. generate p = mean

. generate deaths = p*patients                                          {3}
```

{2} There is now only one record for each value of **apache**. The variables **N** and **mean** store the number of patients with the specified value of **apache** and their associated mortality rate, respectively. **ub** and **lb** give the Wilson 95% confidence interval for this rate.

N.B. All other variables that are not specified by the **by** option are lost. Do not use this command with data that you value and have not saved!

{3} **deaths** give the number of patients with the indicated value of **apache** who die.

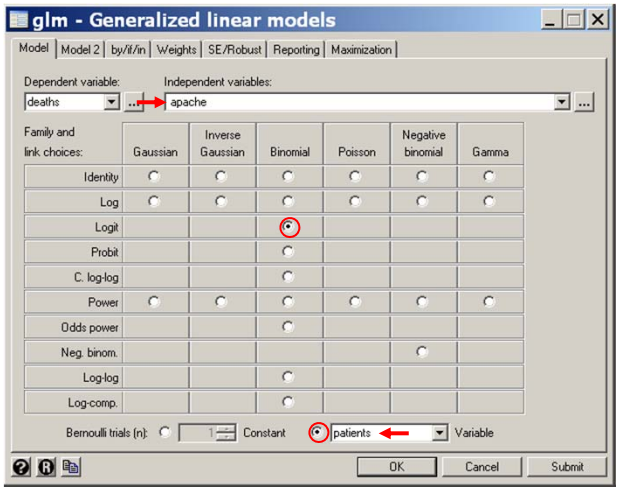
```
. * Statistics > Generalized linear models > Generalized linear models (GLM)
. glm deaths apache, family(binomial patients) link(logit)              {1}
    .
    .
    .
Generalized linear models                         No. of obs      =         38
Optimization     : ML: Newton-Raphson             Residual df     =         36
                                                  Scale param     =          1
Deviance         =  84.36705142                   (1/df) Deviance =  2.343529
Pearson          =  46.72842945                   (1/df) Pearson  =  1.298012

Variance function: V(u) = u*(1-u/patients)        [Binomial]
Link function    : g(u) = ln(u/(patients-u))      [Logit]
Standard errors  : OIM

Log likelihood   = -60.93390578                   AIC             =  3.312311
BIC              = -46.58605033
------------------------------------------------------------------------------
  deaths |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  apache |   .1156272   .0159997      7.23   0.000       .0842684     .146986
   _cons |  -2.290327   .2765283     -8.28   0.000      -2.832313   -1.748342
------------------------------------------------------------------------------
```

{1} Regress **deaths** against **apache** score. The **bionomial** random component and **logit** link function specify that **logistic** regression is to be used.

family(binomial patients) indicates that each observation describes the outcomes of multiple patients with the same apache score; **patients** records the number of subjects with each score; **deaths** records the number of deaths observed in these subjects.




```
. predict logodds, xb     {2}
. generate e_prob = exp(logodds)/(1+exp(logodds))
. label variable e_prob "Expected Mortality at 30 Days"
. *
. *  Calculate 95% confidence region for e_prob
. *
. predict stderr, stdp
. generate lodds_lb = logodds - 1.96*stderr
. generate lodds_ub = logodds + 1.96*stderr
. generate prob_lb = exp(lodds_lb)/(1+exp(lodds_lb))
. generate prob_ub = exp(lodds_ub)/(1+exp(lodds_ub))
. label variable p "Observed Mortality Rate"
. * Data > Describe data > List data
. list p e_prob prob_lb prob_ub ci95lb ci95ub apache if apache == 20

     +-------------------------------------------------------------------------+
     |        p    e_prob    prob_lb   prob_ub         lb        ub   apache |
     |-------------------------------------------------------------------------|
 20. | .4615385   .505554   .4462291   .564723   .2320607   .708562       20 |
     +-------------------------------------------------------------------------+
```

{2} The **linear predictor** is *logodds* = -2.2903 + .115627**apache*.

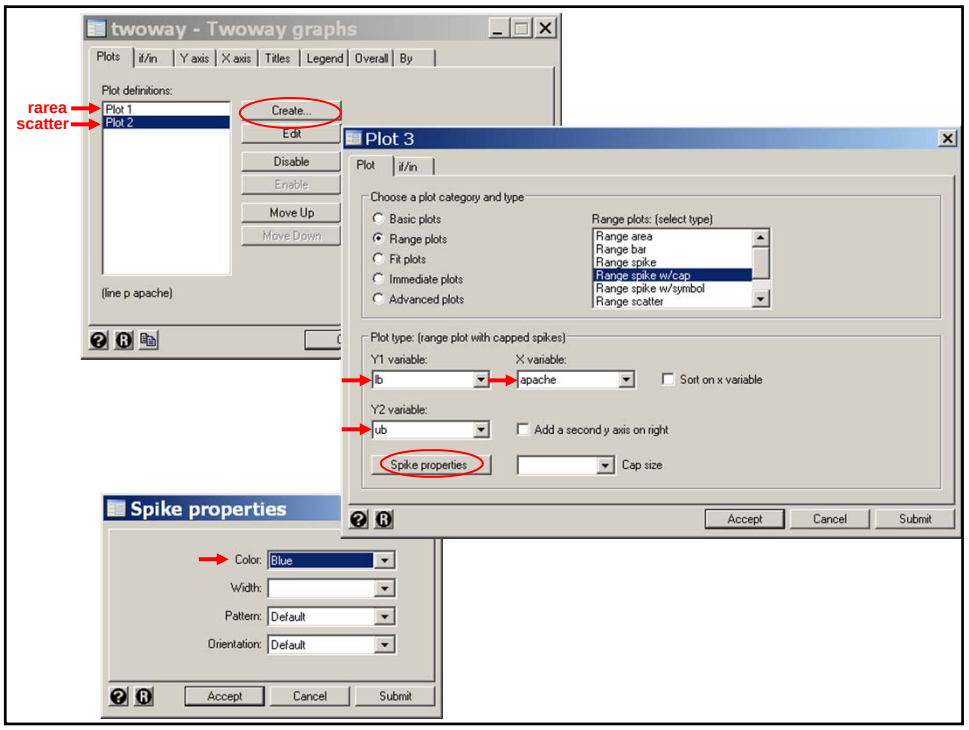
```
. twoway rarea prob_ub prob_lb apache, color(yellow)                    ///
>     || scatter p apache, color(blue)                                  ///
>     || rcap ub lb apache, color(blue)                                 /// {3}
>     || line e_prob apache, yaxis(2)  clwidth(medthick) color(red)     ///
>     ,  ylabel(0(.2)1,labcolor(blue) angle(0))                         /// {4}
>        ytick(0(.1)1, tlcolor(blue))                                   /// {5}
>        ylabel(0(.2)1, axis(2) labcolor(red) angle(0))                 /// {6}
>        ytick(0(.1)1, axis(2) tlcolor(red))                            ///
>        xlabel(0(5)40) xtick(0(1)40)                                   ///
>        ytitle(,axis(2) color(red))                                    ///
>        ytitle(Observed Mortality Rate, color(blue))                   ///
>        legend(order(1 "95% CI from model" 2 3 "95% CI from this obs." 4))
```

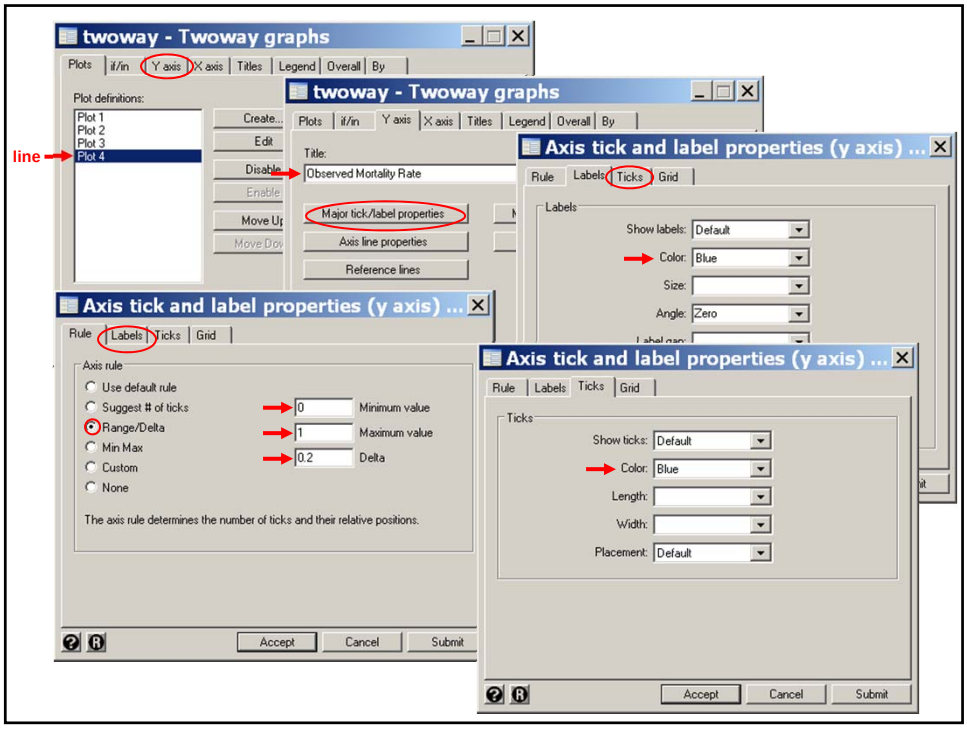
{3} **rcap** plots capped rods (error bars) joining the values of **ub** and **lb** for each value of **apache**.

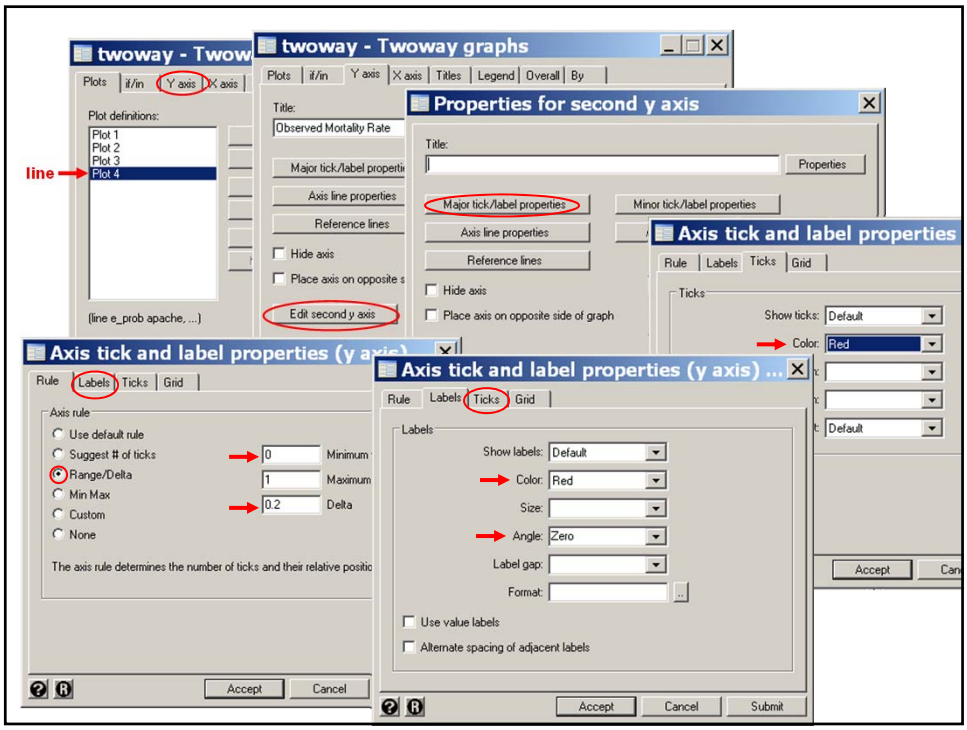
{4} This graph will have two *y*-axes: a left-axis for the observed mortality rate and a right-axis for the expected morbidity rate. Here we color the default (left) axis blue to match the blue scatterplot of observed mortality rates.

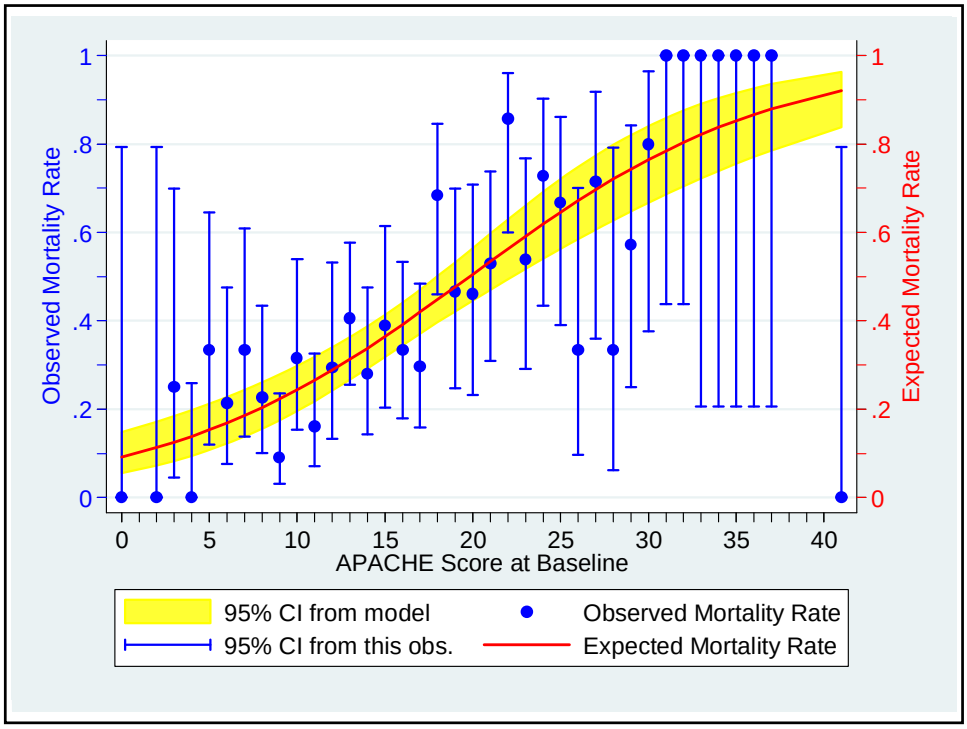
{5} Also, color the tick lines blue on the left axis.

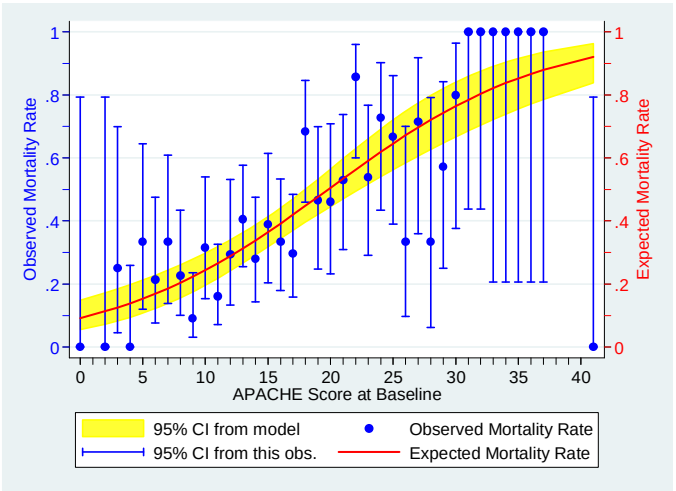
{6} The **axis(2)** suboption indicates that this **ylabel** option refers to the right axis. It is colored red to match the expected mortality curve.




twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Plot 2
Plot 3
Plot 4
line
Create...
Edit
Disable
Enable
Move Up
Move Down
Title:
Observed Mortality Rate
Major tick/label properties
Axis line properties
Reference lines
Axis tick and label properties (y axis) ...
Rule
Labels
Ticks
Grid
Axis rule
Use default rule
Suggest # of ticks
Range/Delta
Min Max
Custom
None
0
1
0.2
Minimum value
Maximum value
Delta
The axis rule determines the number of ticks and their relative positions.
Labels
Show labels:
Default
Color:
Blue
Size:
Angle:
Zero
Ticks
Show ticks:
Length:
Width:
Placement:
Accept
Cancel
Submit


twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Plot 2
Plot 3
Plot 4
line
(line e_prob apache, ...)
Title:
Observed Mortality Rate
Major tick/label properti
Axis line properties
Reference lines
Hide axis
Place axis on opposite s
Edit second y axis
Properties for second y axis
Title:
Properties
Major tick/label properties
Minor tick/label properties
Axis line properties
Reference lines
Hide axis
Place axis on opposite side of graph
Axis tick and label properties
Rule
Labels
Ticks
Grid
Show ticks:
Default
Color:
Red
Axis tick and label properties (y axis) ...
Axis rule
Use default rule
Suggest # of ticks
Range/Delta
Min Max
Custom
None
0
1
0.2
Minimum
Maximum
Delta
The axis rule determines the number of ticks and their relative positic
Labels
Show labels:
Default
Color:
Red
Size:
Angle:
Zero
Label gap:
Format:
Use value labels
Alternate spacing of adjacent labels
Accept
Cancel
Submit





Note that the width of these intervals depends on the number of patients with a given score and the observed mortality rate.

The **blue** error bars in the regression graph give 95% confidence intervals that are derived from the observed mortality rates at each separate APACHE II score. These confidence intervals are not given for scores with zero or 100% mortality. The **yellow shaded region** gives 95% confidence intervals for the expected mortality that are derived from the entire logistic regression.

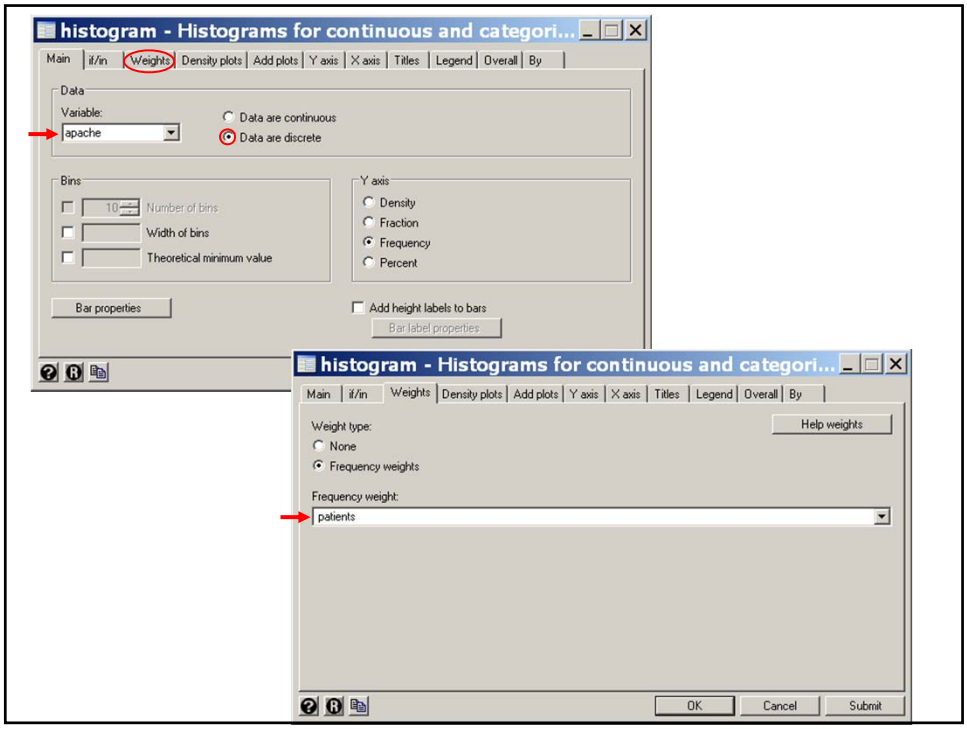
Overall, the fit appears quite good, although the regression curve comes close to the ends of the confidence intervals for some scores and is just outside when the APACHE score equals 18.

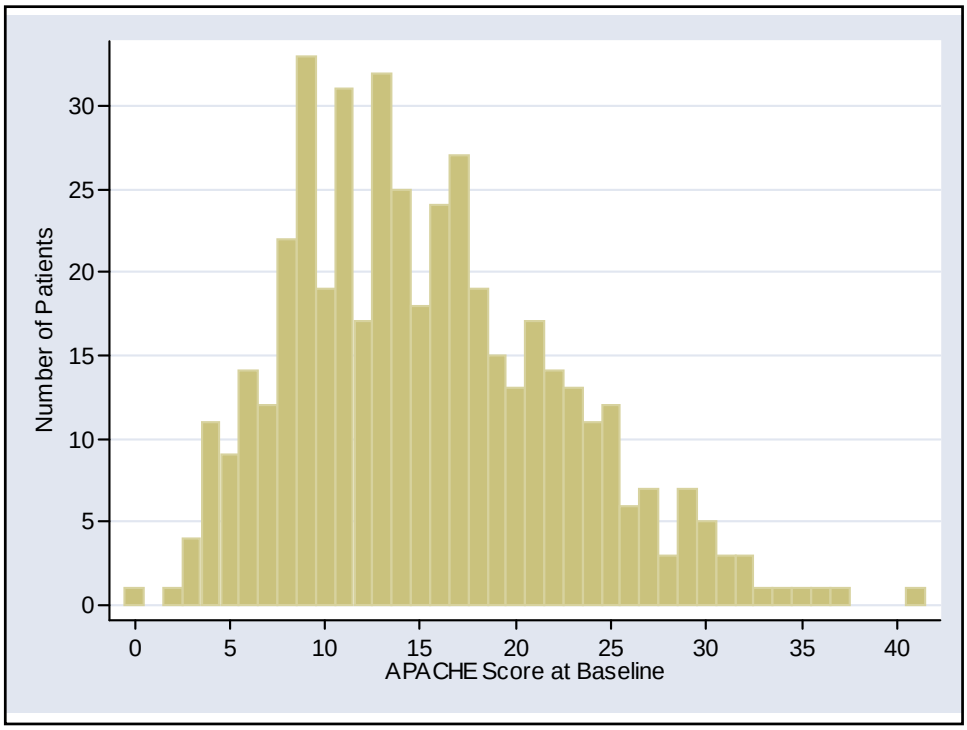
```
*
*  Graph number of patients with different APACHE II scores
*
. * Graphics > Histogram
. histogram apache [freq=patients], discrete frequency xlabel(0(5)40) /// {4}
>     ylabel(0(5)30, angle(0)) ytitle(Number of Patients)
(start=0, width=1)
```

{4} This command produces a histogram of the patient APACHE II scores.

discrete specifies that the data have a discrete number of values. It produces one bar per value unless **width** is also specified.

frequency specifies that the *y*-axis is to be **number of patients** rather than proportion of patients.


histogram - Histograms for continuous and categori...
Main
if/in
Weights
Density plots
Add plots
Y axis
X axis
Titles
Legend
Overall
By
Data
Variable:
apache
Data are continuous
Data are discrete
Bins
Number of bins
Width of bins
Theoretical minimum value
Y axis
Density
Fraction
Frequency
Percent
Bar properties
Add height labels to bars
Bar label properties
histogram - Histograms for continuous and categori...
Weight type:
None
Frequency weights
Help weights
Frequency weight:
patients
OK
Cancel
Submit


Number of Patients
0
5
10
15
20
25
30
APACHE Score at Baseline
0
5
10
15
20
25
30
35
40

12. Simple 2x2 Case-Control Studies

a) Example: Esophageal Cancer and Alcohol

Breslow & Day, Vol. I (1980) give the following results from the Ille-et-Vilaine case-control study of esophageal cancer and alcohol (Tuyns et al. 1977) .

Cases were 200 men diagnosed with esophageal cancer in regional hospitals between 1/1/1972 and 4/30/1974.

Controls were 775 men drawn from electoral lists in each commune.

Esophageal	Daily Alcohol Consumption		
Cancer	≥ 80g	< 80g	Total
Yes (Cases)	96	104	200
No (Controls)	109	666	775
Total	205	770	975

b) Review of Classical Case-Control Theory

Let m_i = number of cases ($i = 1$) or controls ($i = 0$)

d_i = number of cases ($i = 1$) or controls ($i = 0$) who are heavy drinkers.

Then the observed prevalence of heavy drinkers is

d_0/m_0 = 109/775 for controls and

d_1/m_1 = 96/200 for cases.

The observed prevalence of moderate or non-drinkers is

$(m_0 - d_0)/m_0$ = 666/775 for controls and

$(m_1 - d_1)/m_1$ = 104/200 for cases.

The observed odds that a case or control will be a heavy drinker is

$$(d_i / m_i) / [(m_i - d_i) / m_i] = d_i / (m_i - d_i)$$

= 109/666 and 96/104 for controls and cases, respectively.

The observed odds ratio for heavy drinking in cases relative to controls is

$$\hat{\psi} = \frac{d_1 / (m_1 - d_1)}{d_0 / (m_0 - d_0)} = \frac{96/104}{109/666} = 5.64$$

If the cases and controls are a representative sample from their respective underlying populations then

1. $\hat{\psi}$ is an unbiased estimate of the true odds ratio for heavy drinking in cases relative to controls in the underlying population.
2. This true odds ratio also equals the true odds ratio for esophageal cancer in heavy drinkers relative to moderate drinkers.

   Case-control studies would be pointless if this were not true.

Since esophageal cancer is rare $\hat{\psi}$ also estimates the **relative risk** of esophageal cancer in heavy drinkers relative to moderate drinkers.

Woolf's estimate of the standard error of the log odds ratio is

$$\text{se}_{\log(\hat{\psi})} = \sqrt{\frac{1}{d_0} + \frac{1}{m_0 - d_0} + \frac{1}{d_1} + \frac{1}{m_1 - d_1}}$$

and the distribution of $\log(\hat{\psi})$ is approximately normal.

Hence, if we let

$$\hat{\psi}_L = \hat{\psi} \exp\left[-1.96\,\text{se}_{\log(\hat{\psi})}\right]$$

and

$$\hat{\psi}_U = \hat{\psi} \exp\left[1.96\,\text{se}_{\log(\hat{\psi})}\right]$$

then $(\hat{\psi}_L, \hat{\psi}_U)$ is a 95% confidence interval for ψ .

13. Logistic Regression Models for 2x2 Contingency Tables

Consider the logistic regression model

$$\text{logit}(E(d_i / m_i)) = \alpha + \beta x_i \qquad \{3.9\}$$

where $E(d_i / m_i) = \pi_i =$ Probability of being a heavy drinker for cases ($i = 1$) and controls ($i = 0$).

and $x_i = \begin{cases} 1 = \text{ cases} \\ 0 = \text{ for controls} \end{cases}$

Then {3.9} can be rewritten

$$\text{logit}(\pi_i) = \log(\pi_i / (1 - \pi_i)) = \alpha + \beta x_i$$

Hence

$$\log(\pi_1 / (1 - \pi_1)) = \alpha + \beta x_1 = \alpha + \beta$$
$$\log(\pi_0 / (1 - \pi_0)) = \alpha + \beta x_0 = \alpha$$

since $x_1 = 1$ and $x_0 = 0$.

Subtracting these two equations gives

$$\log(\pi_1 / (1 - \pi_1)) - \log(\pi_0 / (1 - \pi_0)) = \beta$$

$$\log\left[\frac{\pi_1 / (1 - \pi_1)}{\pi_0 / (1 - \pi_0)}\right] = \log(\psi) = \beta$$ and hence the true odds ratio $\psi = e^{\beta}$

a) Estimating relative risks from the model coefficients

Our primary interest is in β. Given an estimate $\hat{\beta}$ of β then $\hat{\psi} = e^{\hat{\beta}}$

b) Nuisance parameters

α is called a **nuisance parameter**. This is one that is required by the model but is not used to calculate interesting statistics.

14. Analyzing Case-Control Data with Stata

The Ille-et-Vilaine data may be analyzed as follows:

```
. *  esoph_ca_cc1.log
. *
. *  Logistic regression analysis of Illes-et-Vilaine
. *  2x2 case-control data.
. *
. *  Enter 2x2 table by hand with editor
. *
```

```
. edit                                                          {1}

. list

     +-----------------------------+
     | cancer   alcohol  patients |
     |-----------------------------|
  1. |      0         0       666 |
  2. |      1         0       104 |
  3. |      0         1       109 |
  4. |      1         1        96 |
     +-----------------------------+

. label define yesno 0 "No" 1 "Yes"                             {2}

. label values cancer yesno                                     {3}

. label define dose 0 "< 80g" 1 ">= 80g"

. label values alcohol dose

. list                                                          {4}

     +-----------------------------+
     | cancer   alcohol  patients |
     |-----------------------------|
  1. |     No     < 80g       666 |
  2. |    Yes     < 80g       104 |
  3. |     No    >= 80g       109 |
  4. |    Yes    >= 80g        96 |
     +-----------------------------+
```

{1} Press the **Editor** button to access Stata's spreadsheet-like editor. **Enter** three variables named ***cancer, alcohol*** and ***patients*** as shown in the following *list* command.

{2} The *cancer* variable takes the value **0** for **controls** and **1** for **cases**. To define these values we first define a value **label** called *yesno* that links **0** with "**No**" and **1** with "**Yes**".

{3} We then use the ***label values*** command to link the variable ***cancer*** with the values label ***yesno***. Multiple variables can be assigned to the same values label.

{4} The ***list*** command now gives the value labels of the *cancer* and *alcohol* variables instead of their numeric values. The **numeric values** are still **available** for use in estimation commands.

```
. *
. *  Calculate the odds ratio for esophageal cancer
. *  associated with heavy drinking.
. *
. * Statistics > Epidemiology... > Tables... > Case-control odds ratio
. cc cancer alcohol [freq=patients], woolf
                                                                        {5}

                 | alcohol                 |             Proportion
                 |   Exposed   Unexposed   |     Total     Exposed
-----------------+-------------------------+----------------------
           Cases |        96         104   |       200      0.4800
        Controls |       109         666   |       775      0.1406
-----------------+-------------------------+----------------------
           Total |       205         770   |       975      0.2103
                 |                         |
                 |      Point estimate     |  [95% Conf. Interval]
                 |-------------------------+----------------------
      Odds ratio |          5.640085       |  4.000589    7.951467  (Woolf)    {6}
 Attr. frac. ex. |           .8226977      |  .7500368    .8742371  (Woolf)
 Attr. frac. pop |           .3948949      |
                 +-------------------------------------------------
                             chi2(1) =   110.26  Pr>chi2 = 0.0000
```

{5} Perform a **classical** case-control **analysis** of the data in the 2x2 table defined by *cancer* and *alcohol*. [***freq=patients***] gives the number of patients who have the specified values of *cancer* and *alcohol*. The **woolf** option specifies that the 95% confidence interval for the odds ratio is to be calculated using Woolf's method.

We could have entered one record per patient giving

666 records with cancer = 0 and alcohol = 0,
104 records with cancer = 1 and alcohol = 0,
109 records with cancer = 0 and alcohol = 1, and
96 records with cancer = 1 and alcohol = 1.

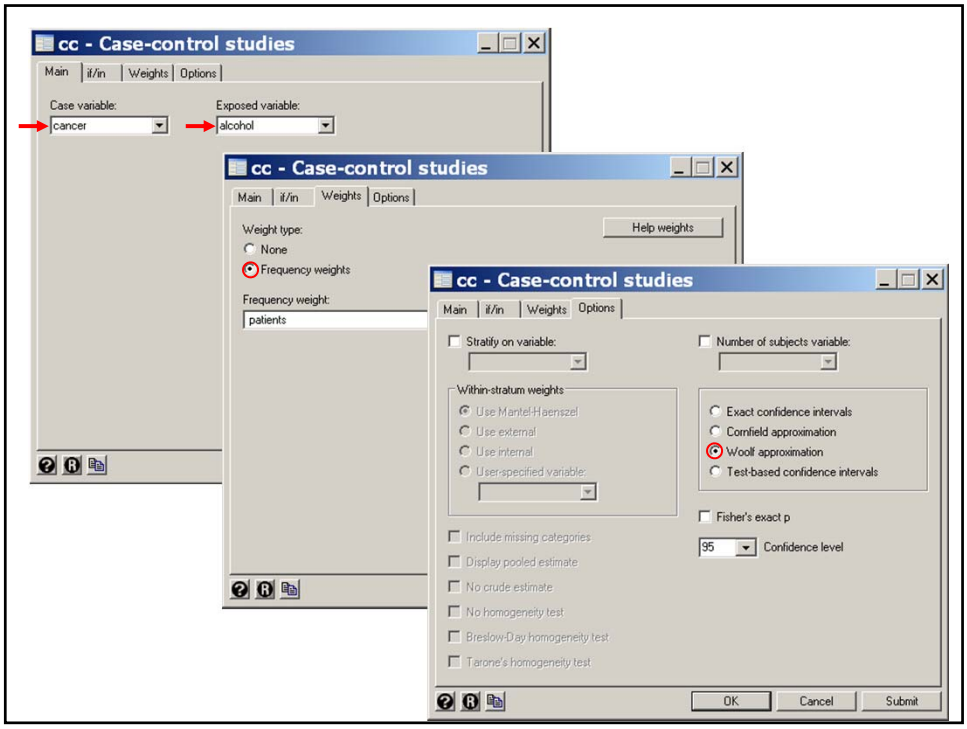
Then the command

```
cc cancer alcohol, woolf
```

would have given exactly the same results as those shown in this example.

N.B. We need to use the **[freq=patients]** command modifier whenever each record represents multiple patients. This will also be true in logistic regression and other commands.

{6} The estimated **odds ratio** is $\frac{96/104}{109/666} = 5.64$




```
. *
. * Now calculate the same odds ratio using logistic regression
. *
. * Statistics > Binary outcomes > Logistic regression
. logit alcohol cancer [freq=patients]                          {7}

Logistic regression                               Number of obs   =        975
                                                  LR chi2(1)      =      96.43
                                                  Prob > chi2     =     0.0000
Log likelihood =  -453.2224                       Pseudo R2       =     0.0962

------------------------------------------------------------------------------
     alcohol |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      cancer |   1.729899   .1752366     9.87   0.000     1.386442    2.073356
       _cons |  -1.809942   .1033238   -17.52   0.000    -2.012453   -1.607431
------------------------------------------------------------------------------
```

{7} This is the analogous ***logit*** command for simple logistic regression. If we had entered the data as

cancer	heavy	patients
0	109	775
1	96	200

Then we would have achieved the same analysis with the command

```
glm heavy cancer, family(binomial patients) link(logit)
```

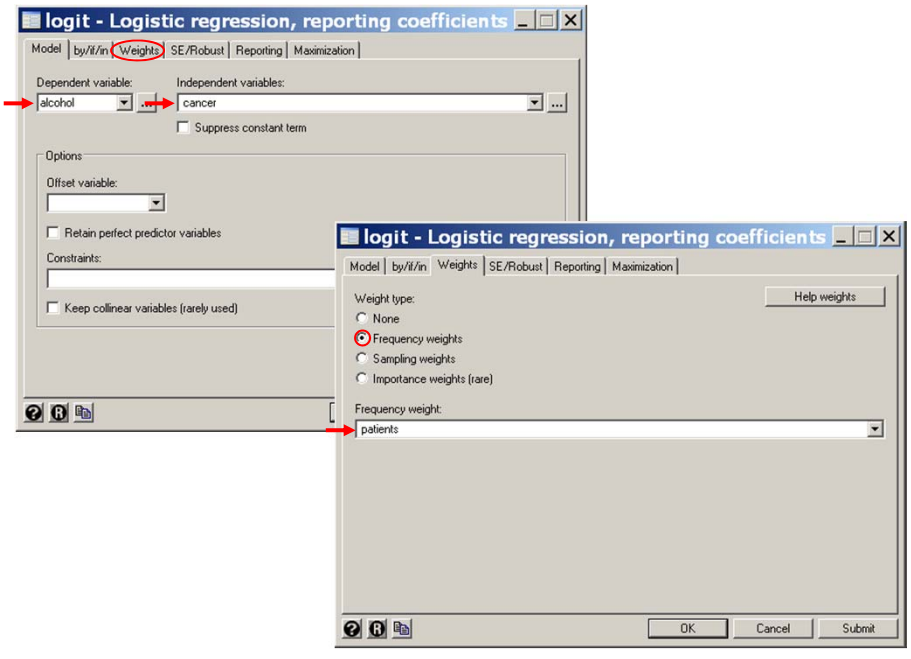
Both of these commands fit the model

$$\text{logit}(E(\text{alcohol})) = \alpha + \text{cancer}*\beta$$

giving **β = 1.73** = the **log odds ratio** of being a heavy drinker in cancer patients relative to controls. The **standard error** of **β** is **0.1752**

The **odds ratio** is exp(1.73) = 5.64.

The **95% confidence interval** for the odds ratio is exp(1.73 ±1.96*0.1752) = (4.00, 7.95)

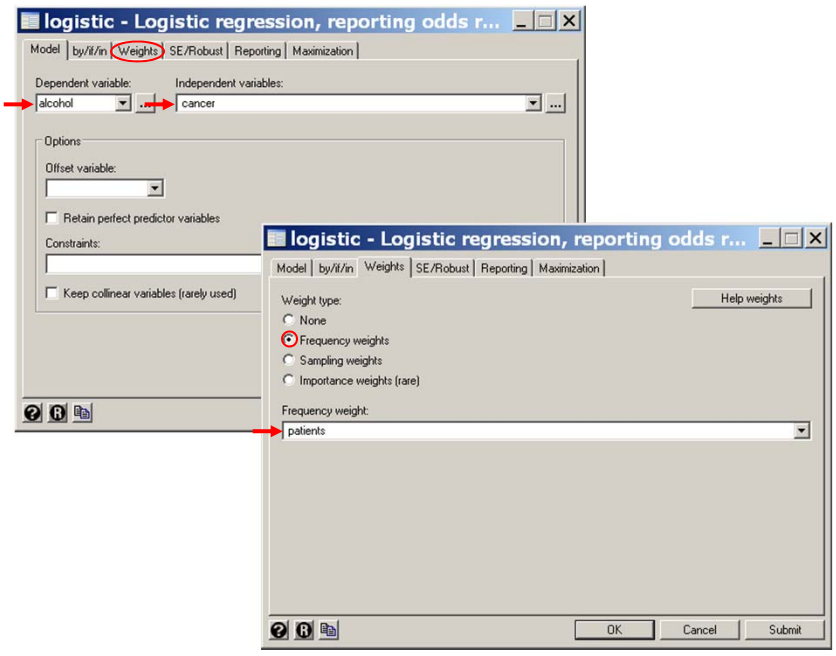


```
. * Statistics > Binary outcomes > Logistic regression (reporting odds ratios)
.  logistic alcohol cancer [freq=patients]                {8}

Logistic regression                                     No. of obs =
975
                  LR chi2(1) =     96.43
                  Prob > chi2=    0.0000
Log likelihood =   -453.2224      Pseudo R2  =    0.0962

------------------------------------------------------------------------------
 alcohol | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  cancer |   5.640085   .9883491      9.87    0.000       4.000589    7.951467
------------------------------------------------------------------------------
```

{8} The ***logistic*** command calculates the odds ratio and its confidence interval directly.



a) Logistic and classical estimates of the 95% CI of the OR

The 95% confidence interval is

$$(5.64\exp(-1.96\times 0.1752),\ 5.64\exp(1.96\times 0.1752)) = (4.00,\ 7.95).$$

The classical limits using Woolf's method is

$$(5.64\exp(-1.96\times s),\ 5.64\exp(1.96\times s)) = (4.00,\ 7.95),$$

where $s^2 = 1/96 + 1/109 + 1/104 + 1/666 = 0.0307 = (0.1752)^2$.

Hence Logistic regression is in exact agreement with classical methods in this simple case.

In Stata the command

```
cc cancer alcohol [freq=patients], woolf
```

gives us Woolf's 95% confidence interval for the odds ratio. We will cover how to calculate confidence intervals using *glm* in the next chapter.

15. Regressing Disease Against Exposure

The simplest explanation of simple logistic regression is the one given above. Unfortunately, it does not generalize to multiple logistic regression where we are considering several risk factors at once. In order to make the next chapter easier to understand, let us return to simple logistic regression one more time.

Suppose we have a population who either are or are not exposed to some risk factor.

Let π'_j denote the true probability of disease in exposed (j = 1) and unexposed (j = 0) people.

We conduct a case-control study in which we select a representative sample of diseased (case) and healthy (control) subjects from the underlying population. That is, the selection is done in such a way that the probability that an individual is selected is unaffected by her exposure status.

Let m_j be the number of study subject who are (j = 1) or are not (j = 0) exposed,

d_j be the number of cases who are (j = 1) or are not (j = 0) exposed,

$x_j =$ j denote exposure status, and

π_j be the probability that a study subject is a case given that she is (j=1) or is not (j=0) exposed.

Consider the model

$$\text{logit}\,(E(d_j / m_j)) = \alpha + \beta x_j$$

This is a legitimate logistic regression model with $E(d_j / m_j) = \pi_j$. It can be shown, however, that this model can be rewritten as

$$\text{logit}(\pi'_j) = \alpha' + \beta x_j$$

where α' is a different constant. However, since α' cancels out in the odds ratio calculation, β estimates the log odds ratio for disease in exposed vs. unexposed members of the population as well as in our case-control sample.

Thus in building logistic regression models it makes sense to regress disease against exposure even though we have no estimate of the probability of disease in the underlying population.

16. What we have covered

- Simple logistic regression: Assessing the effect of a continuous variable on a dichotomous outcome
- How logistic regression parameters affect the probability of an event
  - $\pi(x) = \exp(\alpha + \beta x) / (1 + \exp(\alpha + \beta x))$
  - exp(β) is the odds ratio for death associated with a unit increase in x.
- Probability, odds and odds ratios
- Generalized linear models: The relationship between linear and logistic regression $\text{logit}(E(d_i)) = \alpha + x_i \beta$
- Wald and Wilson confidence intervals for proportions
- Plotting probability of death with 95% confidence bands as a function of a continuous risk factor
- Review of classic 2x2 case-control studies
- Analyzing case-control studies with logistic regression

Cited References

Bernard, G. R., et al. (1997). The effects of ibuprofen on the physiology and survival of patients with sepsis. The Ibuprofen in Sepsis Study Group. *N Engl J Med* 336: 912-8.

Breslow, N. E. and N. E. Day (1980). Statistical Methods in Cancer Research: Vol. 1 - The Analysis of Case-Control Studies. Lyon, France, IARC Scientific Publications.

Tuyns, A. J., G. Pequignot, et al. (1977). Le cancer de L'oesophage en Ille-et-Vilaine en fonction des niveau de consommation d'alcool et de tabac. Des risques qui se multiplient. *Bull Cancer* 64: 45-60.

For additional references on these notes see.

Dupont WD. *Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data. 2nd ed.* Cambridge, U.K.: Cambridge University Press; 2009.