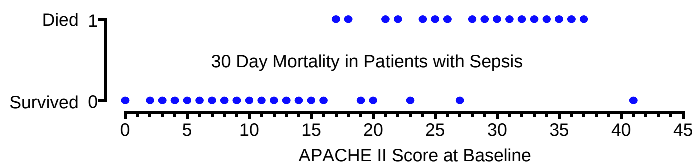
III. INTRODUCTION TO LOGISTIC REGRESSION

1. Simple Logistic Regression

a) Example: APACHE II Score and Mortality in Sepsis

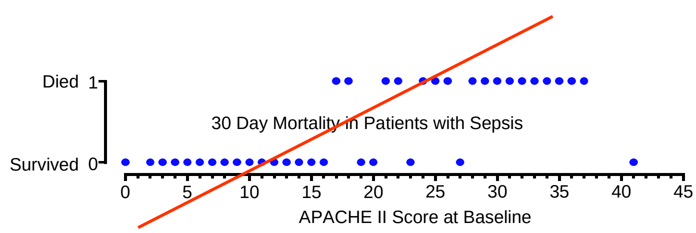
The following figure shows 30 day mortality in a sample of septic patients as a function of their baseline APACHE II Score. Patients are coded as 1 or 0 depending on whether they are dead or alive in 30 days, respectively.



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.



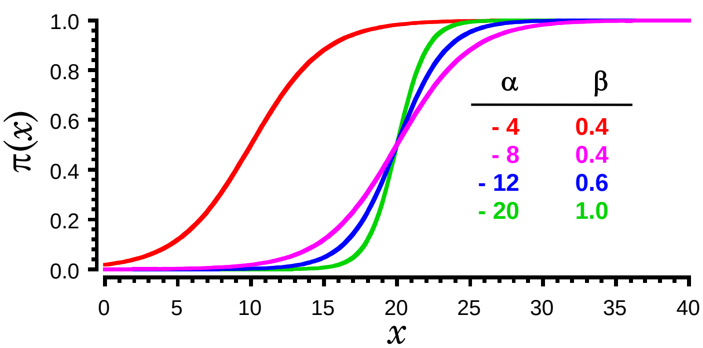
b) Sigmoidal family of logistic regression curves

Logistic regression fits probability functions of the following form:

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

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

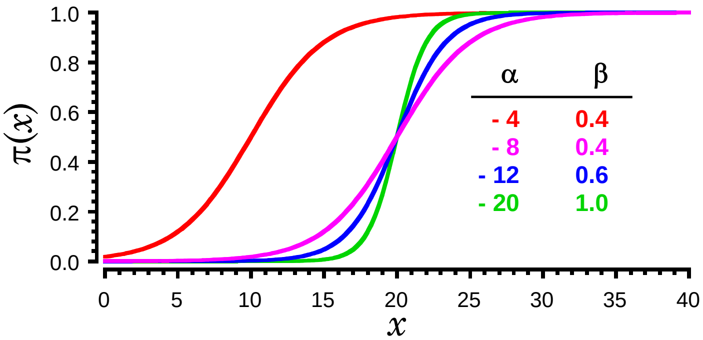
For now assume that $\beta > 0$.

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 $\pi(x) \to \infty / (1 + \infty) = 1$

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



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

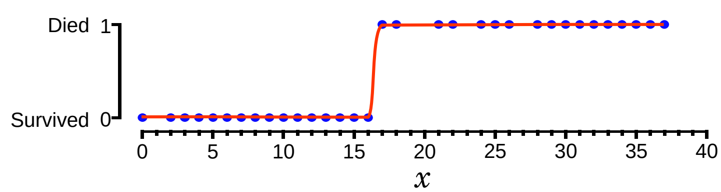
The slope of $\pi(x)$ when $\pi(x)$=.5 is β/4.

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

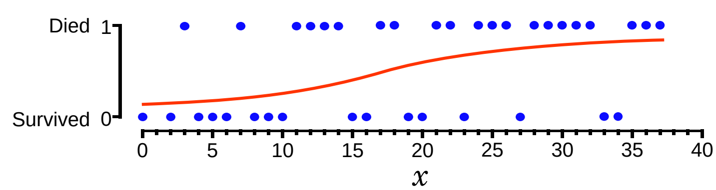
For given β, α controls were the 50% survival point is located.

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

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

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

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

Then the expected value of d_i is

$$E(d_i) = \pi(x_i) = \Pr[d_i = 1]$$

Thus we can rewrite the **logistic regression equation** {5.2} as

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

2. Contrast Between Logistic and Linear Regression

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

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

y_i has a normal distribution with standard deviation σ.
it is the random component of the model, which has a normal distribution.

$\alpha + \beta x_i$ is the linear predictor.

In logistic regression, the expected value of d_i given x_i is $E(d_i) = \pi_i = \pi[x_i]$

logit($E(d_i)$) = $\alpha + x_i \beta$ for $i = 1, 2, \ldots, n$

d_i is dichotomous with probability of event $\pi_i = \pi[x_i]$
it is the random component of the model

logit is the link function that relates the expected value of the random component to the linear predictor.

3. 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.

The maximum likelihood estimate of a parameter is that value that maximizes the probability of the observed data.

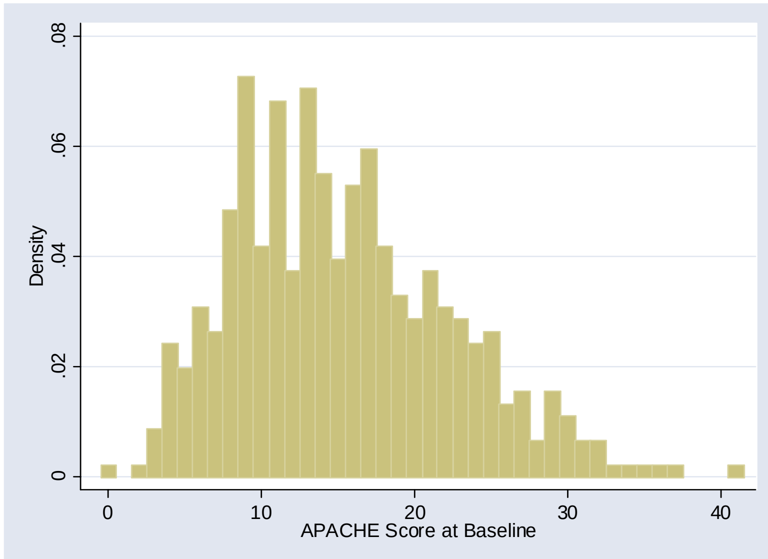
We estimate α and β by those values $\hat{\alpha}$ and $\hat{\beta}$ that maximize the probability of the observed data under the logistic regression model.

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

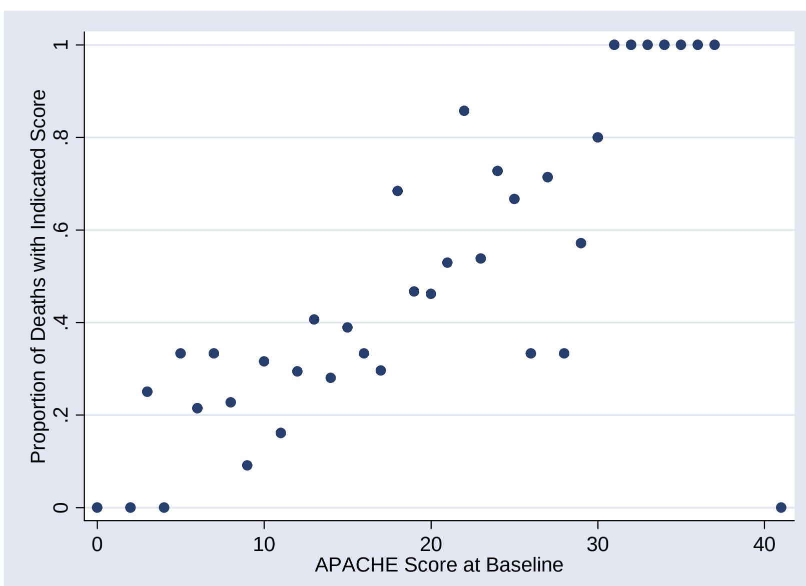
This data is analyzed as follows…

```
. tabulate fate

Death by 30 |
       days |      Freq.     Percent        Cum.
------------+-----------------------------------
      alive |        279       61.45       61.45
       dead |        175       38.55      100.00
------------+-----------------------------------
      Total |        454      100.00
```

```
. histogram apache [fweight=freq ], discrete
(start=0, width=1)
```




```
. scatter proportion apache
```



```
Logit estimates                                   Number of obs   =        454
                                                  LR chi2(1)      =      62.01
                                                  Prob > chi2     =     0.0000
Log likelihood = -271.66534                       Pseudo R2       =     0.1024

------------------------------------------------------------------------------
        fate |      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
------------------------------------------------------------------------------
```

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

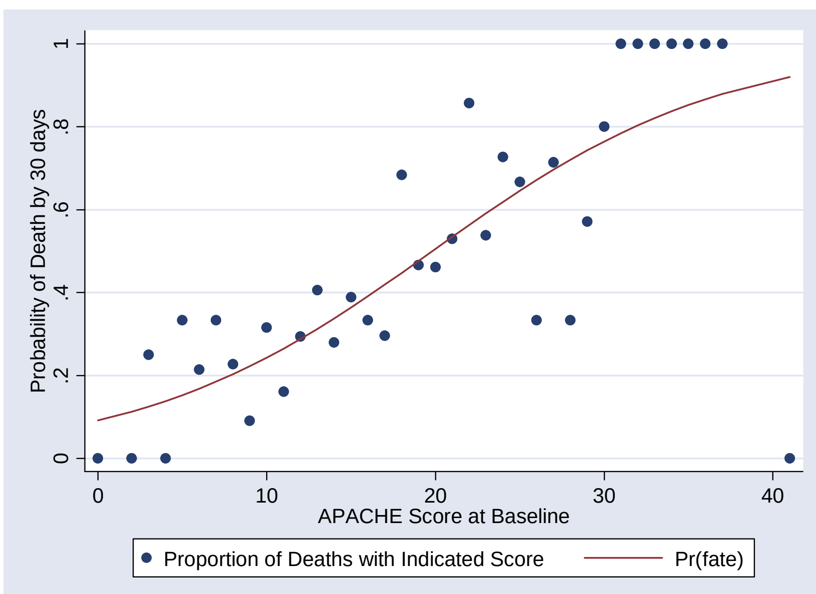
$$\hat{\beta} = .1156272$$

$$\hat{\alpha} = -2.290327$$

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

$$= \frac{\exp(-2.290327 + .1156272x)}{1 + \exp(-2.290327 + .1156272x)}$$

$$\pi(20) = \frac{\exp(-2.290327 + .1156272 \times 20)}{1 + \exp(-2.290327 + .1156272 \times 20)} = 0.50555402$$



4. 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.

5. 95% Confidence Intervals for Odds Ratio Estimates

In our sepsis example the parameter estimate for *apache* (β) was .1156272 with a standard error of .0159997. Therefore, the odds ratio for death associated with a unit rise in APACHE II score is

exp(.1156272) = 1.123

with a 95% confidence interval of

$$(\exp(0.1156 - 1.96 \times 0.0160), \exp(0.1156 + 1.96 \times 0.0160))$$

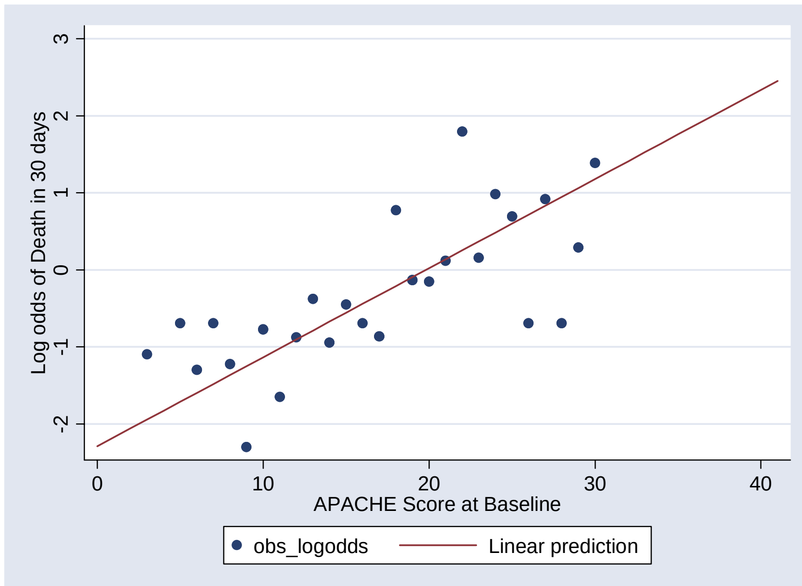
$$= (1.09, 1.15).$$

6. Quality of Model fit

If our model is correct then

$$\text{logit}(\text{observed proportion}) = \hat{\alpha} + \hat{\beta} x_i$$

It can be helpful to plot the observed log odds against $\hat{\alpha} + \hat{\beta} x_i$



7. 95% Confidence Interval for $\pi[x]$

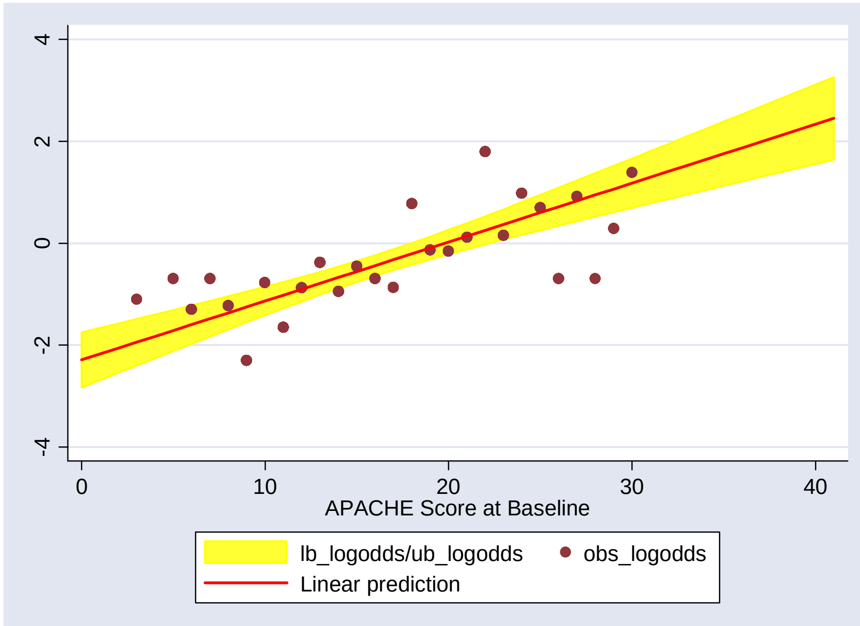
Let $\sigma^2_{\hat{\alpha}}$ and $\sigma^2_{\hat{\beta}}$ denote the variance of $\hat{\alpha}$ and $\hat{\beta}$.
Let $\sigma_{\hat{\alpha}\hat{\beta}}$ denote the covariance between $\hat{\alpha}$ and $\hat{\beta}$.

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

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

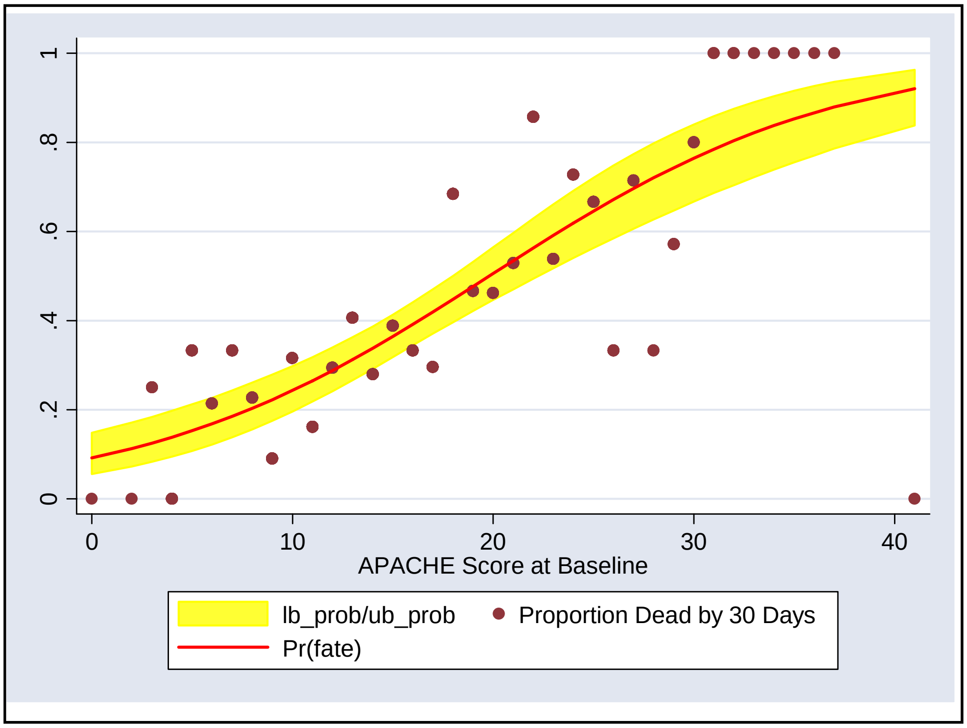
$$\pi(x_i) = \exp(\alpha + \beta x_i) / (1 + \exp(\alpha + \beta x_i))$$

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

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

and

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



It is common to recode continuous variables into categorical variables in order to calculate odds ratios for, say the highest quartile compared to the lowest.

```
. centile apache, centile(25 50 75)

                                                       -- Binom. Interp. --
    Variable |     Obs  Percentile      Centile        [95% Conf. Interval]
-------------+-------------------------------------------------------------
      apache |     454          25           10               9          11
             |                  50           14        13.60845          15
             |                  75           20              19          21

. generate float upper_q= apache >= 20

. tabulate upper_q

    upper_q |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        334       73.57       73.57
          1 |        120       26.43      100.00
------------+-----------------------------------
      Total |        454      100.00
```

```
. cc fate upper_q if apache >= 20 | apache <= 10
                                                         Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+----------------------
           Cases |        77          25  |        102      0.7549
        Controls |        43         101  |        144      0.2986
-----------------+------------------------+----------------------
           Total |       120         126  |        246      0.4878
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |         7.234419       |   3.924571    13.44099  (exact)
 Attr. frac. ex. |         .8617719       |   .7451951    .9256007  (exact)
 Attr. frac. pop |         .6505533       |
                 +-----------------------------------------------
                               chi2(1) =    49.75  Pr>chi2 = 0.0000
```

This approach discards potentially valuable information and may not be as clinically relevant as an odds ratio at two specific values.

Alternately we can calculate the odds ratio for death for patients at the 75th percentile of Apache scores compared to patients at the 25th percentile

$$\text{logit}\,(\pi(20)) = \alpha + \beta \times 20$$

$$\text{logit}\,(\pi(10)) = \alpha + \beta \times 10$$

Subtracting gives

$$\log\left(\frac{\pi(20)/(1-\pi(20))}{\pi(10)/(1-\pi(10))}\right) = \beta \times 10 = \quad 0.1156 \times 10 = 1.156$$

Hence, the odds ratio equals exp(1.156) = 3.18

A problem with this estimate is that it is strongly dependent on the accuracy of the logistic regression model.

Hence, the odds ratio equals exp(1.156) = 3.18

With Stata we can calculate the 95% confidence interval for this odds ratio as follows:

```
. lincom 10*apache, eform

 ( 1)  10 apache = 0

------------------------------------------------------------------------------
        fate |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   3.178064   .5084803     7.23   0.000     2.322593    4.348628
------------------------------------------------------------------------------
```

A problem with this estimate is that it is strongly dependent on the accuracy of the logistic regression model.

Simple logistic regression generalizes to allow multiple covariates

$$\text{logit}(E(d_i)) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$$

where

x_{i1}, x_{I2}, …, x_{ik} are covariates from the i^{th} patient

α and β_1, ...β_k, are known parameters

$$d_i = \begin{cases} 1: & i^{\text{th}} \text{ patient suffers event of interest} \\ 0: & \text{otherwise} \end{cases}$$

Multiple logistic regression can be used for many purposes. One of these is to weaken the logit-linear assumption of simple logistic regression using **restricted cubic splines**.

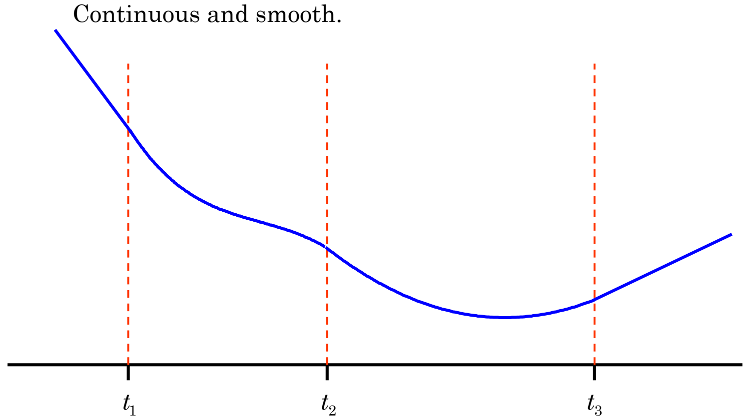
8. Restricted Cubic Splines

These curves have k knots located at $t_1, t_2, \cdots, t_k$. They are:

Linear before t_1 and after t_k.

Piecewise cubic polynomials between adjacent knots (i.e. of the form $ax^3 + bx^2 + cx + d$)

Continuous and smooth.



Given x and k knots a restricted cubic spline can be defined by

$$y = \alpha + x_1\beta_1 + x_2\beta_2 + \cdots + x_{k-1}\beta_{k-1}$$

for suitably defined values of x_i

These covariates are functions of x and the knots but are independent of y.

$x_1 = x$ and hence the hypothesis $\beta_2 = \beta_3 = \cdots = \beta_{k-1}$ tests the linear hypothesis.

In logistic regression we use restricted cubic splines by modeling

$$\text{logit}\left(E(d_i)\right) = \alpha + x_1\beta_1 + x_2\beta_2 + \cdots + x_{k-1}\beta_{k-1}$$

Programs to calculate $x_1, \cdots, x_{k-1}$ are available in Stata, R and other statistical software packages.

We fit a logistic regression model using a three knot restricted cubic spline model with knots at the default locations at the

10th percentile,
50th percentile, and
90th percentile.

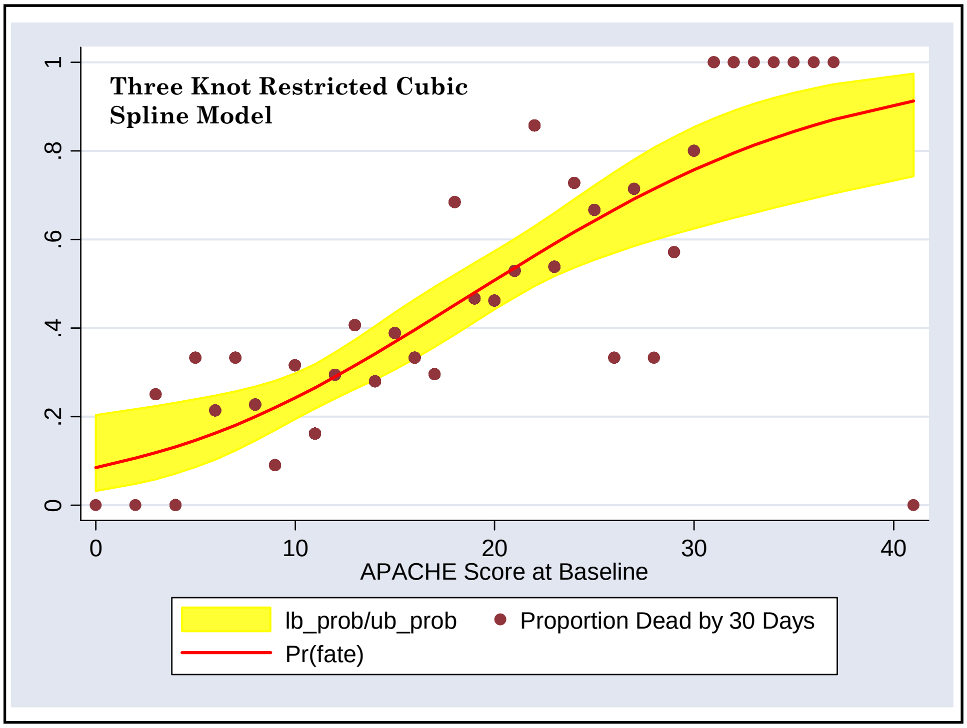
```
. rc_spline apache, nknots(3)
 number of knots = 3
 value of knot 1 = 7
 value of knot 2 = 14
 value of knot 3 = 25

. logit fate _Sapache1 _Sapache2

Logit estimates                                   Number of obs   =        454
                                                  LR chi2(2)      =      62.05
                                                  Prob > chi2     =     0.0000
Log likelihood = -271.64615                       Pseudo R2       =     0.1025

------------------------------------------------------------------------------
        fate |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Sapache1 |   .1237794   .0447174     2.77   0.006      .036135    .2114238
   _Sapache2 |  -.0116944   .0596984    -0.20   0.845     -.128701    .1053123
       _cons |  -2.375381   .5171971    -4.59   0.000    -3.389069   -1.361694
------------------------------------------------------------------------------
```

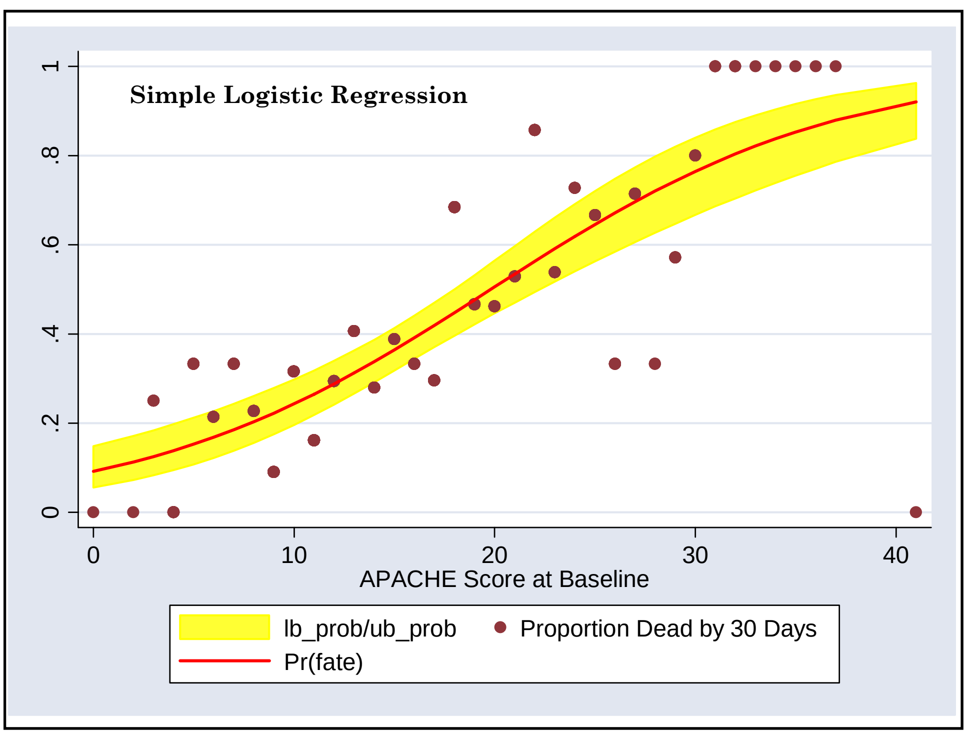
Note that the coefficient for _Sapache2 is small and not significant, indicating an excellent fit for the simple logistic regression model.

Giving analogous commands as for simple logistic regression gives the following plot of predicted mortality given the baseline Apache score.

```
. drop  prob logodds se lb_logodds ub_logodds ub_prob lb_prob

. rename  prob_rcs prob

. rename  logodds_rcs logodds

. predict se, stdp

. generate float lb_logodds= logodds-1.96* se

. generate float ub_logodds= logodds+1.96* se

. generate float ub_prob= exp( ub_logodds)/(1+exp( ub_logodds))

. generate float lb_prob= exp( lb_logodds)/(1+exp( lb_logodds))

. twoway (rarea lb_prob ub_prob apache, blcolor(yellow) bfcolor(yellow))
>       (scatter proportion apache)
>       (line prob apache, clcolor(red) clwidth(medthick))
```

This plot is very similar to the one for simple logistic regression except the 95% confidence band is a little wider.


Three Knot Restricted Cubic
Spline Model
1
.8
.6
.4
.2
0
0
10
20
30
40
APACHE Score at Baseline
lb_prob/ub_prob
Proportion Dead by 30 Days
Pr(fate)


Simple Logistic Regression
1
.8
.6
.4
.2
0
0
10
20
30
40
APACHE Score at Baseline
lb_prob/ub_prob
Proportion Dead by 30 Days
Pr(fate)

This regression gives the following table of values

Percentile	Apache	_Sapache1	_Sapache2
25	10	10	0.083333
75	20	20	5.689955

We calculate the odds ratio of death for patients at the 75th vs. 25th percentile of apache score as follows:

The logodds at the 75th percentile equals

$$\text{logit}(\pi(20)) = \alpha + \beta_1 \times 20 + \beta_2 \times 5.689955$$

The logodds at the 25th percentile equals

$$\text{logit}(\pi(10)) = \alpha + \beta_1 \times 10 + \beta_2 \times 0.083333$$

Subtracting the second from the first equation gives that the log odds ratio for patients at the 75th vs. 25th percentile of apache score is

$$\text{logit}(\pi(20)/\pi(10)) = \beta_1 \times 10 + \beta_2 \times 5.606622$$

Stata calculates this odds ratio to be 3.22 with a 95% confidence interval of 2.3 -- 4.6

```
. lincom _Sapache1*10 + 5.606622* _Sapache2, eform

 ( 1)  10 _Sapache1 + 5.606622 _Sapache2 = 0

------------------------------------------------------------------------------
        fate |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    3.22918    .5815912     6.51   0.000      2.26875    4.596188
------------------------------------------------------------------------------
```

Recall that for the simple model we had the following odds ratio and confidence interval.

```
------------------------------------------------------------------------------
        fate |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   3.178064   .5084803     7.23   0.000     2.322593    4.348628
------------------------------------------------------------------------------
```

The close agreement of these results supports the use of the simple logistic regression model for these data.

9. Simple 2x2 Case-Control Studies

a) Example: Esophageal Cancer and Alcohol

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

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

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 a … rom
their resp …

1. $\hat{\psi}$ is an u … heavy drinking … ing population.

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

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.

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

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

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

Hence, if we let

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

and

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

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

10. Logistic Regression Models for 2x2 Contingency Tables

Consider the logistic regression model

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

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

Then {3.9} can be rewritten

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

Hence

$$\log(\pi_1 / (1-\pi_1)) = \alpha + \beta x_1 = \alpha + \beta \quad \text{and}$$
$$\log(\pi_0 / (1-\pi_0)) = \alpha + \beta x_0 = \alpha$$

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.

11. Analyzing Case-Control Data with Stata

Consider the following data on esophageal cancer and heavy drinking

cancer	alcohol	patients
No	< 80g	666
Yes	< 80g	104
No	>= 80g	109
Yes	>= 80g	96

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

                 | 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)
 Attr. frac. ex. |         .8226977       |  .7500368    .8742371  (Woolf)
 Attr. frac. pop |         .3948949       |
                 +-----------------------------------------------
                               chi2(1) =   110.26  Pr>chi2 = 0.0000

*
* Now calculate the same odds ratio using logistic regression
*
```

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

```
.  logistic alcohol cancer [freq=patients]

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
------------------------------------------------------------------------------
```

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

cancer	heavy	patients
0	109	775
1	96	200

This commands fit the model

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

giving $\boldsymbol{\beta = 1.73}$ = the **log odds ratio** of being a heavy drinker in cancer patients relative to controls.

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

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.