Logistic Models in R

Jim Bentley

1 Sample Data

The following code reads the titanic data that we will use in our examples.

```
> titanic = read.csv(
+ "http://bulldog2.redlands.edu/facultyfolder/jim_bentley/downloads/math111/titanic.cs
> titanic$AGE=factor(titanic$AGE,labels=c(Child,Adult))
> titanic$CLASS=factor(titanic$CLASS,labels=c(0,1,2,3))
> titanic$SEX=factor(titanic$SEX, labels=c(Female,Male))
> titanic$SURVIVED=factor(titanic$SURVIVED,labels=c(No,Yes))
```

Note that the plus signs (+) at the beginning of the lines are there to indicate that R is reading from a new line. They should not be entered as part of the code.

We can now check to see if the data frames have been created by entering

```
> ls()

[1] "titanic"
```

2 Loading R Packages

Additional functions necessary for validation and graphical analysis of the quality of logistic models can be found in Frank Harrell's `rms` package. Harrell provides some functions to make pretty output in `Hmisc`.

```
> ## load a few packages
> #install.packages("xtable")
> #install.packages("rms","Hmisc")
> library(Hmisc)
> library(xtable)
> library(lattice)
> library(rms)  ### Modern R replacement for Design package
```

3 Fitting Logistic Models

The models fitted here are the equivalent of those fitted in the SAS documentation.

3.1 CLASS

A model to test for the difference in odds of survival as determined by class may be fitted using the `lrm` function.

```
> dd = datadist(titanic)
> options(datadist="dd")
> attach(titanic)
> titanic.lrm.class=lrm(SURVIVED~CLASS, x=TRUE, y=TRUE)
> titanic.lrm.class

Logistic Regression Model

lrm(formula = SURVIVED ~ CLASS, x = TRUE, y = TRUE)

                      Model Likelihood     Discrimination    Rank Discrim.
                         Ratio Test           Indexes          Indexes
Obs          2201    LR chi2     180.90    R2       0.110    C       0.642
 No          1490    d.f.             3    g        0.545    Dxy     0.283
 Yes          711    Pr(> chi2) <0.0001    gr       1.725    gamma   0.386
max |deriv| 6e-12                          gp       0.124    tau-a   0.124
                                           Brier    0.200

          Coef    S.E.   Wald Z Pr(>|Z|)
Intercept -1.1552 0.0788 -14.67 <0.0001
CLASS=1    1.6643 0.1390  11.97 <0.0001
CLASS=2    0.8078 0.1438   5.62 <0.0001
CLASS=3    0.0678 0.1171   0.58 0.5624

> anova(titanic.lrm.class)

                Wald Statistics          Response: SURVIVED

 Factor     Chi-Square d.f. P
 CLASS      173.23     3    <.0001
 TOTAL      173.23     3    <.0001
```

Note that the (log) odds of survival do not differ for classes 0 (viewed as baseline) and 3. However, classes 1 and 2 differ from 0 (and thus 3) as well as from each other. This can most easily be seen using the odds ratios.

```
> summary(titanic.lrm.class,CLASS=0)

             Effects              Response : SURVIVED

 Factor      Low High Diff. Effect S.E. Lower 0.95 Upper 0.95
 CLASS - 1:0 1   2    NA    1.66   0.14  1.39      1.94
```

```
  Odds Ratio 1   2    NA    5.28     NA  4.02      6.94
 CLASS - 2:0 1   3    NA    0.81   0.14  0.53      1.09
  Odds Ratio 1   3    NA    2.24     NA  1.69      2.97
 CLASS - 3:0 1   4    NA    0.07   0.12 -0.16      0.30
  Odds Ratio 1   4    NA    1.07     NA  0.85      1.35
```

```
> summary(titanic.lrm.class,CLASS=3)
```

```
             Effects              Response : SURVIVED

 Factor      Low High Diff. Effect S.E. Lower 0.95 Upper 0.95
 CLASS - 0:3 4   1    NA    -0.07  0.12 -0.30      0.16
  Odds Ratio 4   1    NA     0.93    NA  0.74      1.18
 CLASS - 1:3 4   2    NA     1.60  0.14  1.31      1.88
  Odds Ratio 4   2    NA     4.94    NA  3.72      6.54
 CLASS - 2:3 4   3    NA     0.74  0.15  0.45      1.03
  Odds Ratio 4   3    NA     2.10    NA  1.57      2.80
```

```
> summary(titanic.lrm.class,CLASS=1)
```

```
             Effects              Response : SURVIVED

 Factor      Low High Diff. Effect S.E. Lower 0.95 Upper 0.95
 CLASS - 0:1 2   1    NA    -1.66  0.14 -1.94      -1.39
  Odds Ratio 2   1    NA     0.19    NA  0.14       0.25
 CLASS - 2:1 2   3    NA    -0.86  0.17 -1.18      -0.53
  Odds Ratio 2   3    NA     0.42    NA  0.31       0.59
 CLASS - 3:1 2   4    NA    -1.60  0.14 -1.88      -1.31
  Odds Ratio 2   4    NA     0.20    NA  0.15       0.27
```

```
> summary(titanic.lrm.class,CLASS=2)
```

```
             Effects              Response : SURVIVED

 Factor      Low High Diff. Effect S.E. Lower 0.95 Upper 0.95
 CLASS - 0:2 3   1    NA    -0.81  0.14 -1.09      -0.53
  Odds Ratio 3   1    NA     0.45    NA  0.34       0.59
 CLASS - 1:2 3   2    NA     0.86  0.17  0.53       1.18
  Odds Ratio 3   2    NA     2.35    NA  1.70       3.26
 CLASS - 3:2 3   4    NA    -0.74  0.15 -1.03      -0.45
  Odds Ratio 3   4    NA     0.48    NA  0.36       0.64
```

While the odds for class 3 relative to class 0 are essentially 1:1, class 1 has a 5.28:1 odds of survival and class 2 has a 2.24:1 odds of survival relative to class 0.

The probability of survival for the different classes may be plotted (Figure 1).

```
> print(plot(Predict(titanic.lrm.class, fun=plogis), ylab=Probability of Survival))
```

And we should validate the model.

```
> validate(titanic.lrm.class, B=80)

          index.orig training    test optimism index.corrected  n
Dxy           0.2833   0.2866  0.2785   0.0080          0.2753 80
R2            0.1102   0.1110  0.1091   0.0019          0.1083 80
Intercept     0.0000   0.0000  0.0017  -0.0017          0.0017 80
Slope         1.0000   1.0000  0.9960   0.0040          0.9960 80
Emax          0.0000   0.0000  0.0012   0.0012          0.0012 80
D             0.0817   0.0823  0.0808   0.0015          0.0803 80
U            -0.0009  -0.0009 -0.0001  -0.0008         -0.0001 80
Q             0.0826   0.0832  0.0810   0.0022          0.0804 80
B             0.1998   0.1993  0.2001  -0.0008          0.2006 80
g             0.5450   0.5528  0.5480   0.0048          0.5402 80
gp            0.1240   0.1252  0.1244   0.0008          0.1232 80
```

A nomogram may be helpful at this point (Figure 2.

```
> nom = nomogram(titanic.lrm.class, fun=plogis)
> print(plot(nom))

NULL
```

3.2 AGE and SEX

A model to test for the difference in odds of survival as determined by age and sex may be fitted using the `lmr` function.

```
> dd = datadist(titanic)
> options(datadist="dd")
> attach(titanic)

The following objects are masked from titanic (position 4):

    AGE, CLASS, SEX, SURVIVED

> titanic.lrm.agesex=lrm(SURVIVED~AGE*SEX, x=TRUE, y=TRUE)
> titanic.lrm.agesex

Logistic Regression Model

lrm(formula = SURVIVED ~ AGE * SEX, x = TRUE, y = TRUE)

                      Model Likelihood     Discrimination    Rank Discrim.
                         Ratio Test            Indexes          Indexes
```

```
Obs          2201    LR chi2     456.68    R2       0.262    C       0.713
 No          1490    d.f.             3    g        0.841    Dxy     0.427
 Yes          711    Pr(> chi2) <0.0001    gr       2.320    gamma   0.787
max |deriv| 1e-10                          gp       0.187    tau-a   0.187
                                           Brier    0.171

                     Coef    S.E.   Wald Z Pr(>|Z|)
Intercept             0.4990 0.3075  1.62  0.1046
AGE=Adult             0.5654 0.3269  1.73  0.0837
SEX=Male             -0.6870 0.3970 -1.73  0.0835
AGE=Adult * SEX=Male -1.7465 0.4167 -4.19  <0.0001
```

```
> anova(titanic.lrm.agesex)

                Wald Statistics          Response: SURVIVED

 Factor                                     Chi-Square d.f. P
 AGE  (Factor+Higher Order Factors)          23.88     2    <.0001
  All Interactions                           17.57     1    <.0001
 SEX  (Factor+Higher Order Factors)         371.97     2    <.0001
  All Interactions                           17.57     1    <.0001
 AGE * SEX  (Factor+Higher Order Factors)    17.57     1    <.0001
 TOTAL                                      391.59     3    <.0001
```

The odds associated with the model are

```
> summary(titanic.lrm.agesex, AGE=Adult, SEX=Male)

             Effects              Response : SURVIVED

 Factor            Low High Diff. Effect S.E. Lower 0.95 Upper 0.95
 AGE - Child:Adult 2   1    NA     1.18  0.26 0.67        1.69
  Odds Ratio       2   1    NA     3.26    NA 1.96        5.41
 SEX - Female:Male 2   1    NA     2.43  0.13 2.19        2.68
  Odds Ratio       2   1    NA    11.40    NA 8.89       14.61

Adjusted to: AGE=Adult SEX=Male
```

```
> summary(titanic.lrm.agesex, AGE=Child, SEX=Female)

             Effects              Response : SURVIVED

 Factor            Low High Diff. Effect S.E. Lower 0.95 Upper 0.95
 AGE - Adult:Child 1   2    NA     0.57  0.33 -0.08      1.21
  Odds Ratio       1   2    NA     1.76    NA  0.93      3.34
 SEX - Male:Female 1   2    NA    -0.69  0.40 -1.47      0.09
```

```
  Odds Ratio       1   2    NA     0.50    NA  0.23      1.10

Adjusted to: AGE=Child SEX=Female
```

The probability of survival for the different combinations of sex and age group may be plotted (Figure 3).

```
> Predict(titanic.lrm.agesex, fun=plogis,
+         AGE=c(Child,Adult), SEX=c(Female,Male))

    AGE    SEX      yhat     lower     upper
1 Child Female 0.6222222 0.4741134 0.7505638
2 Adult Female 0.7435294 0.6998704 0.7828084
3 Child   Male 0.4531250 0.3362143 0.5754460
4 Adult   Male 0.2027594 0.1841419 0.2227454

Response variable (y):

Limits are 0.95 confidence limits

> print(plot(Predict(titanic.lrm.agesex, fun=plogis,
+              AGE=c(Child,Adult), SEX=c(Female,Male))))
```

And we should validate the model.

```
> validate(titanic.lrm.agesex, B=80)

          index.orig training    test optimism index.corrected  n
Dxy           0.4267   0.4240  0.4264  -0.0024          0.4291 80
R2            0.2618   0.2611  0.2606   0.0005          0.2613 80
Intercept     0.0000   0.0000 -0.0052   0.0052         -0.0052 80
Slope         1.0000   1.0000  0.9977   0.0023          0.9977 80
Emax          0.0000   0.0000  0.0015   0.0015          0.0015 80
D             0.2070   0.2065  0.2060   0.0005          0.2066 80
U            -0.0009  -0.0009  0.0001  -0.0010          0.0001 80
Q             0.2079   0.2074  0.2060   0.0014          0.2065 80
B             0.1713   0.1711  0.1717  -0.0005          0.1718 80
g             0.8414   0.8361  0.8377  -0.0016          0.8430 80
gp            0.1867   0.1853  0.1860  -0.0007          0.1874 80
```

A nomogram may be helpful at this point (Figure 4).

```
> nom = nomogram(titanic.lrm.agesex, fun=plogis)
> print(plot(nom))

NULL
```

3.3 CLASS, AGE and SEX

A model to test for the difference in odds of survival as determined by class, age and sex may be fitted using the `lmr` function.

```
> dd = datadist(titanic)
> options(datadist="dd")
> attach(titanic)

The following objects are masked from titanic (position 3):

    AGE, CLASS, SEX, SURVIVED
The following objects are masked from titanic (position 5):

    AGE, CLASS, SEX, SURVIVED

> titanic.lrm.classagesex=lrm(SURVIVED~CLASS*SEX+AGE*SEX, x=TRUE, y=TRUE)
> titanic.lrm.classagesex

Logistic Regression Model

lrm(formula = SURVIVED ~ CLASS * SEX + AGE * SEX, x = TRUE, y = TRUE)

                      Model Likelihood     Discrimination    Rank Discrim.
                         Ratio Test           Indexes          Indexes
Obs           2201    LR chi2     634.70    R2       0.350    C       0.766
 No           1490    d.f.             9    g        1.341    Dxy     0.532
 Yes           711    Pr(> chi2) <0.0001    gr       3.823    gamma   0.638
max |deriv| 5e-10                          gp       0.233    tau-a   0.233
                                           Brier    0.157

                     Coef    S.E.   Wald Z Pr(>|Z|)
Intercept             2.0775 0.7171  2.90  0.0038
CLASS=1               1.6642 0.8003  2.08  0.0376
CLASS=2               0.0497 0.6874  0.07  0.9424
CLASS=3              -2.0894 0.6381 -3.27  0.0011
SEX=Male             -1.7888 0.7728 -2.31  0.0206
AGE=Adult            -0.1803 0.3618 -0.50  0.6182
CLASS=1 * SEX=Male   -1.1033 0.8199 -1.35  0.1784
CLASS=2 * SEX=Male   -0.7647 0.7271 -1.05  0.2929
CLASS=3 * SEX=Male    1.5623 0.6562  2.38  0.0173
SEX=Male * AGE=Adult -1.3581 0.4551 -2.98  0.0028

> anova(titanic.lrm.classagesex)

                Wald Statistics          Response: SURVIVED
```

Factor	Chi-Square	d.f.	P
CLASS (Factor+Higher Order Factors)	124.28	6	<.0001
All Interactions	48.25	3	<.0001
SEX (Factor+Higher Order Factors)	254.38	5	<.0001
All Interactions	63.47	4	<.0001
AGE (Factor+Higher Order Factors)	31.30	2	<.0001
All Interactions	8.91	1	0.0028
CLASS * SEX (Factor+Higher Order Factors)	48.25	3	<.0001
SEX * AGE (Factor+Higher Order Factors)	8.91	1	0.0028
TOTAL INTERACTION	63.47	4	<.0001
TOTAL	311.38	9	<.0001

The odds associated with the model are

```
> summary(titanic.lrm.classagesex, CLASS=0, AGE=Adult, SEX=Male)
```

Effects Response : SURVIVED

Factor	Low	High	Diff.	Effect	S.E.	Lower 0.95	Upper 0.95
CLASS - 1:0	1	2	NA	0.56	0.18	0.21	0.91
Odds Ratio	1	2	NA	1.75	NA	1.24	2.48
CLASS - 2:0	1	3	NA	-0.72	0.24	-1.18	-0.25
Odds Ratio	1	3	NA	0.49	NA	0.31	0.78
CLASS - 3:0	1	4	NA	-0.53	0.15	-0.83	-0.23
Odds Ratio	1	4	NA	0.59	NA	0.44	0.80
SEX - Female:Male	2	1	NA	3.15	0.62	1.92	4.37
Odds Ratio	2	1	NA	23.26	NA	6.84	79.12
AGE - Child:Adult	2	1	NA	1.54	0.28	1.00	2.08
Odds Ratio	2	1	NA	4.66	NA	2.71	8.00

Adjusted to: CLASS=0 SEX=Male AGE=Adult

```
> summary(titanic.lrm.classagesex, CLASS=3, AGE=Child, SEX=Female)
```

Effects Response : SURVIVED

Factor	Low	High	Diff.	Effect	S.E.	Lower 0.95	Upper 0.95
CLASS - 0:3	4	1	NA	2.09	0.64	0.84	3.34
Odds Ratio	4	1	NA	8.08	NA	2.31	28.22
CLASS - 1:3	4	2	NA	3.75	0.53	2.72	4.79
Odds Ratio	4	2	NA	42.67	NA	15.11	120.56
CLASS - 2:3	4	3	NA	2.14	0.33	1.49	2.79
Odds Ratio	4	3	NA	8.49	NA	4.45	16.20
SEX - Male:Female	1	2	NA	-0.23	0.42	-1.06	0.60
Odds Ratio	1	2	NA	0.80	NA	0.35	1.83
AGE - Adult:Child	1	2	NA	-0.18	0.36	-0.89	0.53

```
  Odds Ratio       1   2    NA     0.83    NA  0.41        1.70

Adjusted to: CLASS=3 SEX=Female AGE=Child
```

The probability of survival for the different combinations of sex and age group may be plotted (Figure 5).

```
> Predict(titanic.lrm.classagesex, fun=plogis,
+         CLASS=c(0,1,2,3), AGE=c(Child,Adult), SEX=c(Female,Male))

   CLASS   AGE    SEX      yhat      lower     upper
1      0 Child Female 0.8886935 0.66194638 0.9701987
2      1 Child Female 0.9768348 0.92575376 0.9930367
3      2 Child Female 0.8935160 0.78056016 0.9519104
4      3 Child Female 0.4970148 0.33827964 0.6563539
5      0 Adult Female 0.8695652 0.66454828 0.9573283
6      1 Adult Female 0.9723831 0.92874210 0.9895961
7      2 Adult Female 0.8750999 0.79597143 0.9263784
8      3 Adult Female 0.4520760 0.37865906 0.5276396
9      0 Child   Male 0.5716729 0.43150500 0.7012113
10     1 Child   Male 0.7004700 0.55927805 0.8116614
11     2 Child   Male 0.3949969 0.25626440 0.5529915
12     3 Child   Male 0.4406809 0.32190559 0.5666583
13     0 Adult   Male 0.2227378 0.19619893 0.2517422
14     1 Adult   Male 0.3342723 0.26910345 0.4064468
15     2 Adult   Male 0.1229466 0.08312897 0.1781310
16     3 Adult   Male 0.1446912 0.11604927 0.1789709

Response variable (y):

Limits are 0.95 confidence limits
```

```
> print(plot(Predict(titanic.lrm.classagesex, fun=plogis,
+       CLASS=c(0,1,2,3),SEX=c(Female,Male), AGE=c(Child,Adult)),
+       pch=c(2,1),col=c(1,2),layout=c(1,2)))
```

And we should validate the model.

```
> validate(titanic.lrm.classagesex, B=80)

          index.orig training    test optimism index.corrected  n
Dxy           0.5322   0.5272  0.5281  -0.0009          0.5331 80
R2            0.3500   0.3519  0.3449   0.0071          0.3429 80
Intercept     0.0000   0.0000 -0.0228   0.0228         -0.0228 80
Slope         1.0000   1.0000  0.9745   0.0255          0.9745 80
Emax          0.0000   0.0000  0.0096   0.0096          0.0096 80
```

```
D              0.2879   0.2899  0.2831   0.0068          0.2811 80
U             -0.0009  -0.0009  0.0003  -0.0012          0.0003 80
Q              0.2888   0.2908  0.2828   0.0080          0.2808 80
B              0.1571   0.1566  0.1578  -0.0012          0.1582 80
g              1.3411   1.3710  1.3290   0.0420          1.2991 80
gp             0.2326   0.2315  0.2285   0.0030          0.2296 80
```

A nomogram may be helpful at this point (Figure 6).

```
> nom = nomogram(titanic.lrm.classagesex, fun=plogis)
> print(plot(nom))

NULL
```

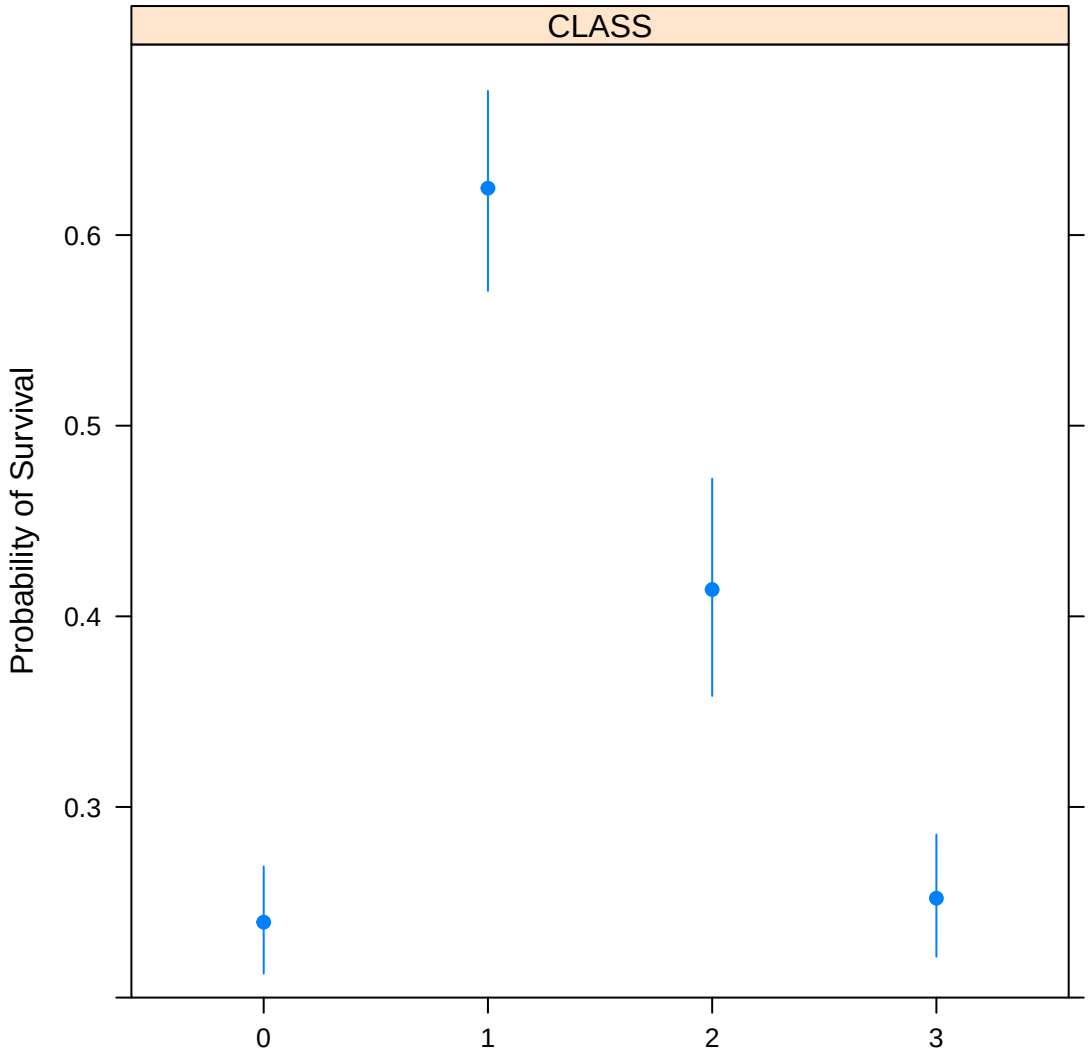

Figure 1: Estimated probability of survival (and 95% CIs) based upon class.

Figure 2: Estimated probability of survival based upon class.

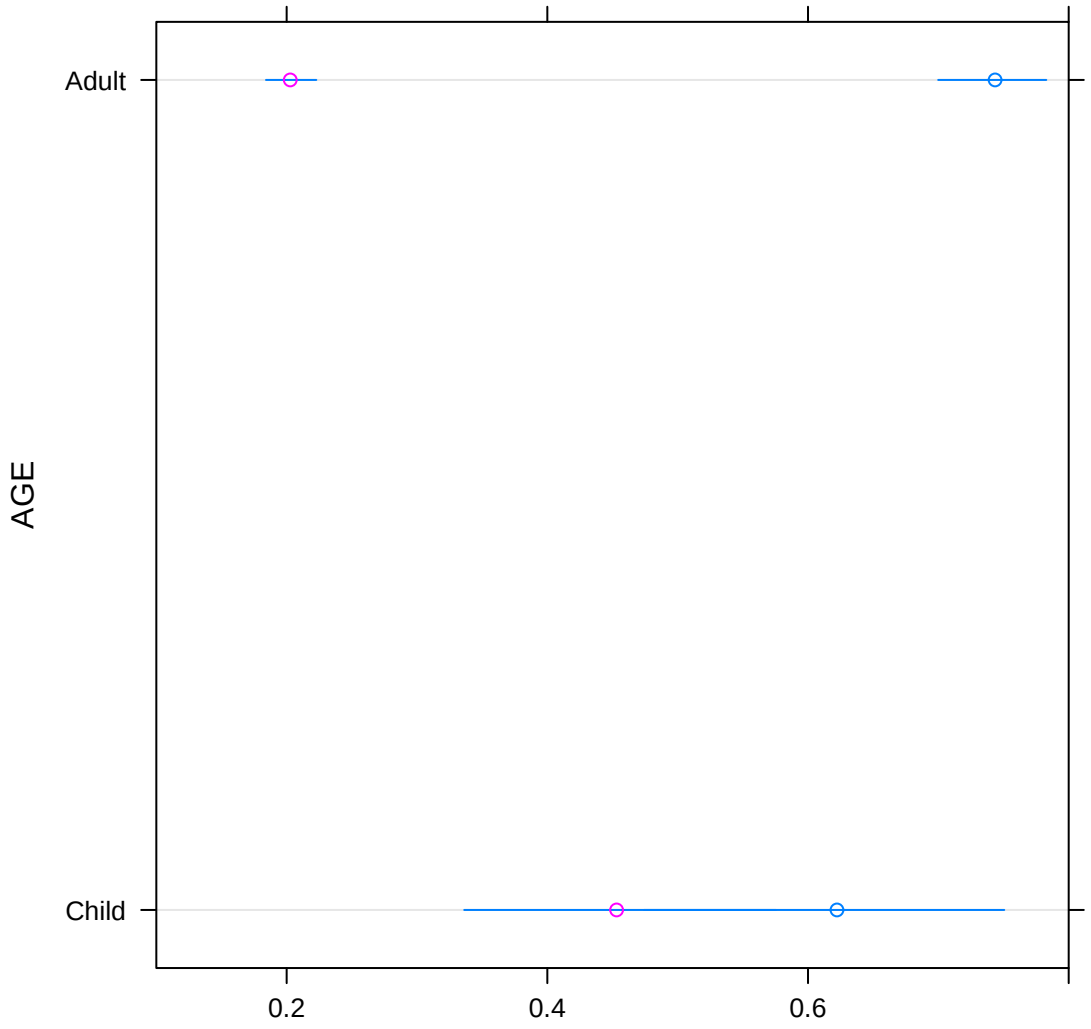

Figure 3: Estimated probability of survival based upon sex and age group.

Figure 4: Nomogram for survival based upon sex and age group.

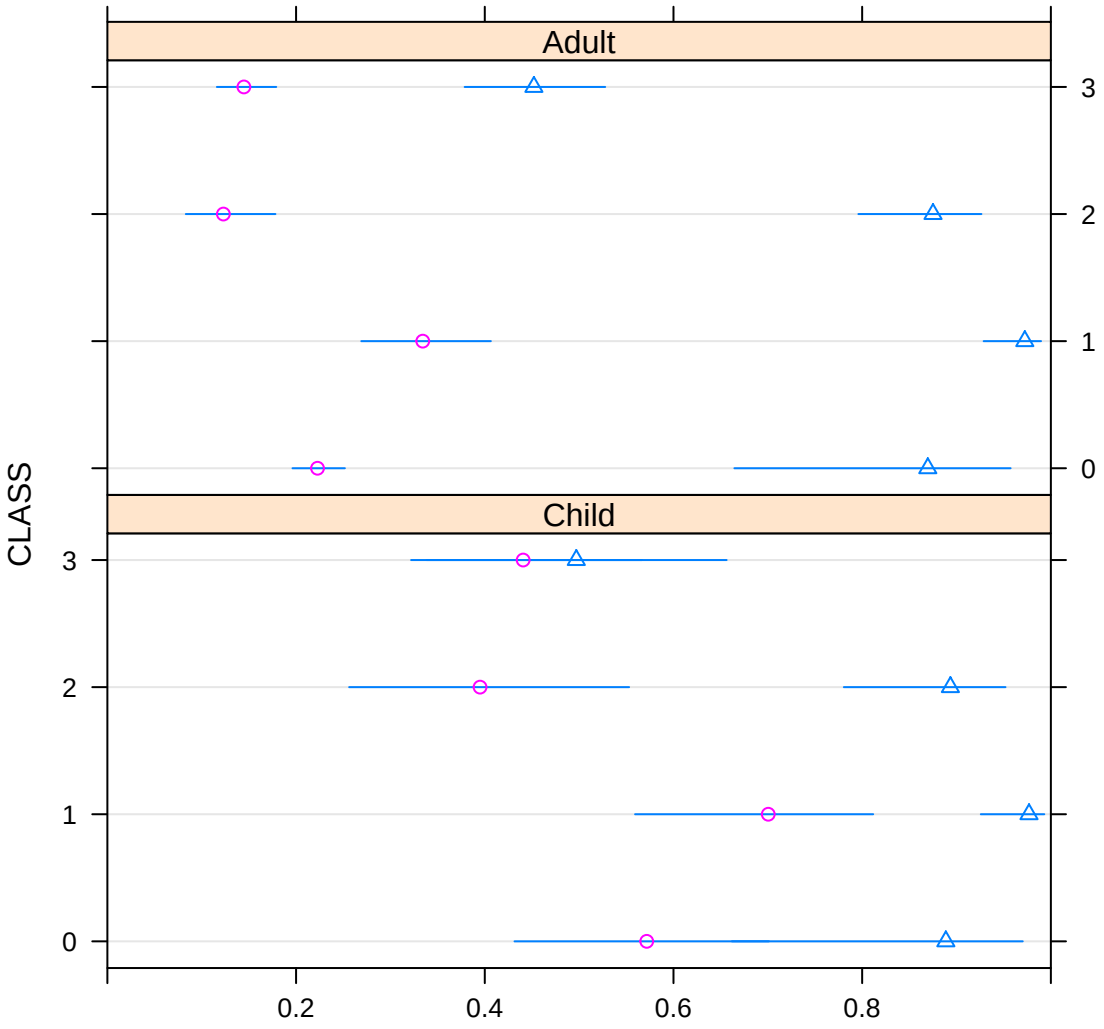

Figure 5: Estimated probability of survival based upon class, sex and age group.

Figure 6: Nomogram for survival based upon class, sex and age group.