

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.csv"  
> 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

```
> ## load a few packages  
> #install.packages("xtable")  
> library(Hmisc)  
> library(xtable)  
> library(lattice)
```

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 `glm` function with `binomial` error and `logit` link.

```
> titanic.logistic.class=glm(SURVIVED~CLASS,
+ family=binomial(logit),data=titanic)
> summary(titanic.logistic.class)
```

Call:

```
glm(formula = SURVIVED ~ CLASS, family = binomial(logit), data = titanic)
```

Deviance Residuals:

```
      Min       1Q   Median       3Q      Max
-1.3999  -0.7623  -0.7401   0.9702   1.6906
```

Coefficients:

```
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.15516    0.07876 -14.667 < 2e-16 ***
CLASS1       1.66434    0.13902  11.972 < 2e-16 ***
CLASS2       0.80785    0.14375   5.620 1.91e-08 ***
CLASS3       0.06785    0.11711   0.579  0.562
```

```
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
```

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 2769.5 on 2200 degrees of freedom
Residual deviance: 2588.6 on 2197 degrees of freedom
AIC: 2596.6
```

Number of Fisher Scoring iterations: 4

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.

```
> coefs=summary(titanic.logistic.class)$coef
> est=exp(coefs[,1])
> upper.ci=exp(coefs[,1]+1.96*coefs[,2])
> lower.ci<-exp(coefs[,1]-1.96*coefs[,2])
> cbind(est,lower.ci,upper.ci)
```

```
              est  lower.ci  upper.ci
(Intercept) 0.3150074 0.2699482 0.3675878
CLASS1      5.2822069 4.0223687 6.9366366
CLASS2      2.2430799 1.6923031 2.9731124
CLASS3      1.0702008 0.8507054 1.3463295
```

```
> rm(coefs)
```

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.

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 `glm` function with binomial error and logit link.

```
> titanic.logistic.agesex=glm(SURVIVED~AGE*SEX,  
+ family=binomial(logit),data=titanic)  
> summary(titanic.logistic.agesex)
```

Call:

```
glm(formula = SURVIVED ~ AGE * SEX, family = binomial(logit),  
     data = titanic)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.6497	-0.6732	-0.6732	0.7699	1.7865

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.4990	0.3075	1.623	0.1046
AGEAdult	0.5654	0.3269	1.729	0.0837 .
SEXMale	-0.6870	0.3970	-1.731	0.0835 .
AGEAdult:SEXMale	-1.7465	0.4167	-4.191	2.77e-05 ***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2769.5 on 2200 degrees of freedom
Residual deviance: 2312.8 on 2197 degrees of freedom
AIC: 2320.8

Number of Fisher Scoring iterations: 4

This model may also be expressed as

```
> titanic.logistic.agesex2=glm(SURVIVED~AGE+SEX+AGE:SEX,  
+ family=binomial(logit),data=titanic)  
> summary(titanic.logistic.agesex2)
```

Call:

```
glm(formula = SURVIVED ~ AGE + SEX + AGE:SEX, family = binomial(logit),  
     data = titanic)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.6497	-0.6732	-0.6732	0.7699	1.7865

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.4990	0.3075	1.623	0.1046
AGEAdult	0.5654	0.3269	1.729	0.0837 .
SEXMale	-0.6870	0.3970	-1.731	0.0835 .
AGEAdult:SEXMale	-1.7465	0.4167	-4.191	2.77e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2769.5 on 2200 degrees of freedom
Residual deviance: 2312.8 on 2197 degrees of freedom
AIC: 2320.8

Number of Fisher Scoring iterations: 4

The odds associated with the model are

```
> coefs=summary(titanic.logistic.agesex2)$coef
> est=exp(coefs[,1])
> upper.ci=exp(coefs[,1]+1.96*coefs[,2])
> lower.ci<-exp(coefs[,1]-1.96*coefs[,2])
> cbind(est,lower.ci,upper.ci)
```

	est	lower.ci	upper.ci
(Intercept)	1.6470588	0.90154072	3.0090740
AGEAdult	1.7601573	0.92740960	3.3406529
SEXMale	0.5030612	0.23104993	1.0953069
AGEAdult:SEXMale	0.1743855	0.07705575	0.3946531

```
> rm(coefs)
```

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 `glm` function with binomial error and logit link.

```
> titanic.logistic.classagesex=glm(SURVIVED~AGE*SEX+CLASS*SEX+CLASS:AGE,
+ family=binomial(logit),data=titanic)
> summary(titanic.logistic.classagesex)
```

Call:

```
glm(formula = SURVIVED ~ AGE * SEX + CLASS * SEX + CLASS:AGE,
     family = binomial(logit), data = titanic)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.6771	-0.7099	-0.5952	0.2374	2.2293

Coefficients: (1 not defined because of singularities)

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.86087	0.73347	2.537	0.01118 *
AGEAdult	0.03625	0.39325	0.092	0.92655
SEXMale	-2.46011	0.81614	-3.014	0.00258 **
CLASS1	17.99982	920.38674	0.020	0.98440
CLASS2	17.11036	405.66288	0.042	0.96636
CLASS3	-2.05502	0.63854	-3.218	0.00129 **
AGEAdult:SEXMale	-0.68679	0.52541	-1.307	0.19116
SEXMale:CLASS1	-1.13608	0.82048	-1.385	0.16616
SEXMale:CLASS2	-1.06807	0.74658	-1.431	0.15254
SEXMale:CLASS3	1.66387	0.65601	2.536	0.01120 *
AGEAdult:CLASS1	-16.34159	920.38639	-0.018	0.98583
AGEAdult:CLASS2	-17.19040	405.66231	-0.042	0.96620
AGEAdult:CLASS3	NA	NA	NA	NA

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2769.5 on 2200 degrees of freedom
Residual deviance: 2097.5 on 2189 degrees of freedom
AIC: 2121.5

Number of Fisher Scoring iterations: 15

The odds associated with the model are

```
> coefs=summary(titanic.logistic.classagesex)$coef
> est=exp(coefs[,1])
> upper.ci=exp(coefs[,1]+1.96*coefs[,2])
> lower.ci<-exp(coefs[,1]-1.96*coefs[,2])
> cbind(est,lower.ci,upper.ci)
```

	est	lower.ci	upper.ci
(Intercept)	6.429309e+00	1.52693798	27.0711771
AGEAdult	1.036918e+00	0.47973665	2.2412280
SEXMale	8.542543e-02	0.01725325	0.4229640
CLASS1	6.564805e+07	0.00000000	Inf
CLASS2	2.697330e+07	0.00000000	Inf
CLASS3	1.280899e-01	0.03664230	0.4477617

AGEAdult:SEXMale	5.031883e-01	0.17967631	1.4091922	
SEXMale:CLASS1	3.210755e-01	0.06429782	1.6033118	
SEXMale:CLASS2	3.436711e-01	0.07955044	1.4847160	
SEXMale:CLASS3	5.279697e+00	1.45950127	19.0991297	
AGEAdult:CLASS1	7.997191e-08	0.00000000		Inf
AGEAdult:CLASS2	3.422187e-08	0.00000000		Inf

> rm(coefs)