Logistic Regression

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Logistic Regression

1. Introduction
2. Some Probability Theory Basics
   - The Binomial Distribution
3. Logistic Regression with a Single Predictor
   - Coronary Heart Disease
   - The Logistic Regression Model
   - Fitting with glm
   - Plotting Model Fit
   - Interpreting Model Coefficients
4. Assessing Model Fit in Logistic Regression
   - The Deviance Statistic
   - Comparing Models
   - Test of Model Fit
5. Logistic Regression with Several Predictors
6. Generalized Linear Models

Introduction

- *Logistic Regression* deals with the case where the dependent variable is binary, and the conditional distribution is binomial.
- Recall that, for a random variable Y having a binomial distribution with parameters n (the number of trials), and p (the probability of "success" , the mean of Y is np and the variance of Y is $np(1-p)$.
- Therefore, if the conditional distribution of Y given a predictor X is binomial, then the mean function and variance functions will be necessarily related.
- Moreover, since, for a given value of n, the mean of the conditional distribution is necessarily bounded by 0 and n, it follows that a linear function will generally fail to fit at large values of the predictor.
- So, special methods are called for.

Some Probability Theory Basics

The Binomial Distribution

- This discrete distribution is one of the foundations of modern categorical data analysis
- The binomial random variable X represents the number of "successes" in N outcomes of a *binomial process*
- A binomial process is characterized by
  - N independent trials
  - Only two outcomes, arbitrarily designated "success" and "failure"
  - Probabilities of success and failure remain constant over trials
- Many interesting real world processes only approximately meet the above specifications
- Nevertheless, the binomial is often an excellent approximation

Some Probability Theory Basics

The Binomial Distribution

- The binomial distribution is a two-parameter family, N is the number of trials, p the probability of success
- The binomial has pdf

$$Pr(X = r) = \binom{N}{r} p^r (1 - p)^{N-r}$$

- The mean and variance of the binomial are

$$E(X) = Np$$
$$Var(X) = Np(1 - p)$$

Some Probability Theory Basics

The Binomial Distribution

- The $B(N, p)$ distribution is well approximated by a $N(Np, Np(1-p))$ distribution as long as p is not too far removed from .5 and N is reasonably large
- A good rule of thumb is that both Np and $N(1-p$ must be greater than 5
- The approximation can be further improved by *correcting for continuity*

Logistic Regression with a Single Predictor

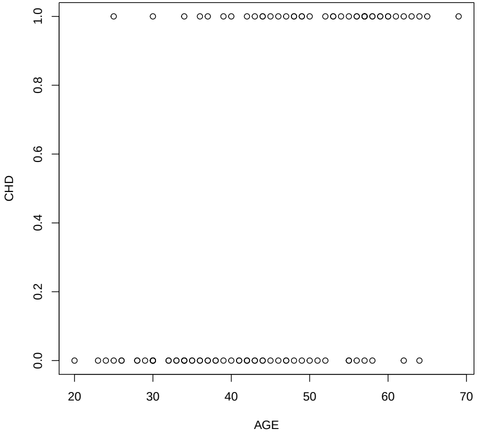
Coronary Heart Disease

- As an example, consider some data relating age to the presence of coronary disease.
- The independent variable is the age of the subject, and the dependent variable is binary, reflecting the presence or absence of coronary heart disease.

```
> chd.data <- read.table(
+ "http://www.statpower.net/R2101/chdage.txt",
+ header=T)
> attach(chd.data)
> plot(AGE,CHD)
```

Logistic Regression with a Single Predictor

Coronary Heart Disease



Logistic Regression with a Single Predictor

Coronary Heart Disease

- The general trend, that age is related to coronary heart disease, seems clear from the plot, but it is difficult to see the precise nature of the relationship.
- We can get a crude but somewhat more revealing picture of the relationship between the two variables by collecting the data in groups of ten observations and plotting mean age against the proportion of individuals with CHD.

Logistic Regression with a Single Predictor

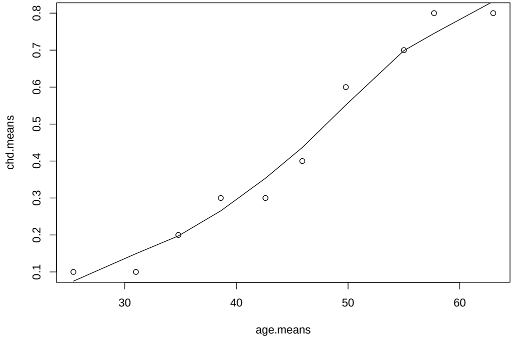
Coronary Heart Disease

```
> age.means <- rep(0,10)
> chd.means <- rep(0,10)
> for(i in 0:9)age.means[i+1]<-mean(
+   chd.data[(10*i+1):(10*i+10),2])
> age.means

 [1] 25.4 31.0 34.8 38.6 42.6 45.9 49.8 55.0 57.7 63.0

> for(i in 0:9)chd.means[i+1]<-mean(
+   chd.data[(10*i+1):(10*i+10),3])
> chd.means

 [1] 0.1 0.1 0.2 0.3 0.3 0.4 0.6 0.7 0.8 0.8
```

Logistic Regression with a Single Predictor

Coronary Heart Disease

```
> plot(age.means,chd.means)
> lines(lowess(age.means,chd.means,iter=1,f=2/3))
```



Logistic Regression with a Single Predictor

The Logistic Regression Model

- For notational simplicity, suppose we have a single predictor, and define $p(x) = \Pr(Y = 1|X = x) = E(Y|X = x)$.
- Suppose that, instead of the probability of heart disease, we consider the *odds* as a function of age.
- Odds range from zero to infinity, so the problem fitting a linear model to the upper asymptote can be eliminated.
- If we go one step further and consider the logarithm of the odds, we now have a dependent variable that ranges from $-\infty$ to $+\infty$.

Logistic Regression with a Single Predictor

The Logistic Regression Model

- Suppose we try to fit a linear regression model to the log-odds variable.
- Our model would now be

$$\operatorname{logit}(p(x)) = \log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x \tag{1}$$

If we can successfully fit this linear model, then we also have successfully fit a nonlinear model for $p(x)$, since the logit function is invertible, so after taking $\operatorname{logit}^{-1}$ of both sides, we obtain

$$p(x) = \operatorname{logit}^{-1}(\beta_0 + \beta_1 x) \tag{2}$$

where

$$\operatorname{logit}^{-1}(w) = \frac{\exp(w)}{1+\exp(w)} = \frac{1}{1+\exp(-w)} \tag{3}$$

Logistic Regression with a Single Predictor

The Logistic Regression Model

- The above system generalizes to more than one predictor, i.e.,

$$p(\mathbf{x}) = E(Y|X = \mathbf{x}) = \text{logit}^{-1}(\beta'\mathbf{x}) \tag{4}$$

Logistic Regression with a Single Predictor

The Logistic Regression Model

- It turns out that the system we have just described is a special case of what is now termed a *generalized linear model*.
- In the context of generalized linear model theory, the logit function that "linearizes" the binomial proportions $p(\mathbf{x})$ is called a *link function*.
- In this module, we shall pursue logistic regression primarily from the practical standpoint of obtaining estimates and interpreting the results.
- Logistic regression is applied very widely in the medical and social sciences, and entire books on applied logistic regression are available.

Logistic Regression with a Single Predictor

Fitting with glm

- Fitting a logistic regression model in R is straightforward.
- You use the `glm` function and specify the binomial distribution family and the logit link function.

Logistic Regression with a Single Predictor

Fitting with glm

```
> fit.chd <- glm(CHD ~AGE, family=binomial(link="logit"))
> summary(fit.chd)

Call:
glm(formula = CHD ~ AGE, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9407  -0.8538  -0.4735   0.8392   2.2518

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.12630    1.11205   -4.61 4.03e-06 ***
AGE          0.10695    0.02361    4.53 5.91e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 136.66  on 99  degrees of freedom
Residual deviance: 108.88  on 98  degrees of freedom
AIC: 112.88

Number of Fisher Scoring iterations: 4
```

Logistic Regression with a Single Predictor

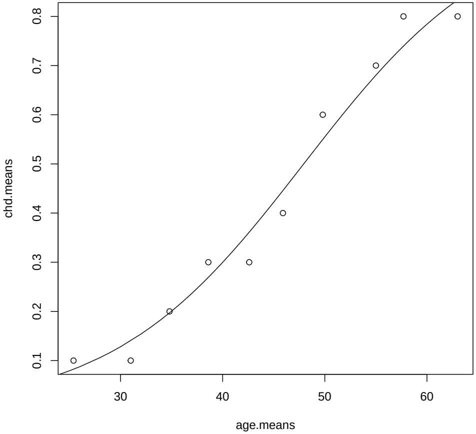
Plotting Model Fit

- Remember that the coefficient estimates are for the transformed model. They provide a linear fit for logit($p(x)$), not for $p(x)$. However, if we define an inverse logit function, we can transform our model back to the original metric.
- Below, we plot the mean `AGE` against the mean `CHD` for groups of 10 observations, then superimpose the logistic regression fit, transformed back into the probability metric.

```
> pdf("Scatterplot02.pdf")
> logit.inverse <- function(x){1/(1+exp(-x))}
> plot(age.means,chd.means)
> lines(AGE,logit.inverse(predict(fit.chd)))
```

Logistic Regression with a Single Predictor

Plotting Model Fit



Logistic Regression with a Single Predictor

Interpreting Model Coefficients

- Suppose there is a single predictor, and it is categorical (0,1). How can one interpret the coefficient β_1?
- Consider the *odds ratio*, the ratio of the odds when $x = 1$ to the odds when $x = 0$.
- According to our model, $\text{logit}(p(x)) = \exp(\beta_0 + \beta_1 x)$, so the log of the odds ratio is given by

$$\begin{aligned}
\log(OR) &= \log\left[\frac{p(1)/(1-p(1))}{p(0)/(1-p(0))}\right] \\
&= \log\left[p(1)/(1-p(1))\right] - \log\left[p(0)/(1-p(0))\right] \\
&= \text{logit}(p(1)) - \text{logit}(p(0)) \\
&= \beta_0 + \beta_1 \times 1 - (\beta_0 + \beta_1 \times 0) \\
&= \beta_1 \qquad (5)
\end{aligned}$$

Logistic Regression with a Single Predictor

Interpreting Model Coefficients

- Exponentiating both sides, we get

$$OR = \exp(\beta_1) \tag{6}$$

- Suppose that X represents the presence or absence of a medical treatment, and $\beta_1 = 2$. This means that the odds ratio is $\exp(2) = 7.389$. If the event is survival, this implies that the odds of surviving are 7.389 times as high when the treatment is present than when it is not.
- You can see why logistic regression is very popular in medical research, and why there is a tradition of working in the "odds metric."

Logistic Regression with a Single Predictor

Interpreting Model Coefficients

- In our coronary heart disease data set, the predictor is continuous.
- Interpreting model coefficients when a predictor is continuous is more difficult.
- Recalling the form of the fitted function for $p(x)$, we see that it does not have a constant slope.
- By taking derivatives, we compute the slope as $\beta_1 p(x)(1 - p(x))$. Hence, the steepest slope is at $p(x) = 1/2$, at which $x = -\beta_0/\beta_1$, and the actual slope is $\beta_1/4$.
- In toxicology, this is called LD_{50}, because it is the dose at which the probability of death is $1/2$.

Logistic Regression with a Single Predictor

Interpreting Model Coefficients

- So a rough "rule of thumb" is that when X is near the middle of its range, a unit change in X results in a change of $\beta_1/4$ units in $p(x)$.
- More precise calculations can be achieved with the aid of R and the logit^{-1} function.

Logistic Regression with a Single Predictor

Interpreting Model Coefficients

Example (CHD vs. AGE)

- We saw that, in our CHD data, the estimated value of β_1 is 0.1069, and the estimated value of β_0 is -5.1263.
- This suggests that, around the age of 45, an increase of 1 year in *AGE* corresponds roughly to an increase of 0.0267 in the probability of coronary heart disease.
- Let's do the calculations by hand, using R.

```
> beta.1 <- coefficients(fit.chd)[2]
> beta.0 <- coefficients(fit.chd)[1]
> predict.45 <- logit.inverse(beta.0 + beta.1 * 45)
> predict.46 <- logit.inverse(beta.0 + beta.1 * 46)
>   change <- predict.46 - predict.45
> results <- data.frame(t(as.numeric(c(predict.45,
+   predict.46,change, beta.1/4))))
> colnames(results) <- c("predict.45","predict.46",
+   "change",".25*beta.1")
> results

  predict.45 predict.46     change .25*beta.1
1   0.422195  0.4484776 0.02628253 0.02673629
```

Logistic Regression with a Single Predictor

Interpreting Model Coefficients

- The numbers demonstrate that, in the "linear zone" near the center of the plot, the rule of thumb works quite well.
- The rule implies that for every increase of 4 units in AGE, there will be roughly a β_1 increase in the probability of coronary heart disease.
- We can simplify the calculations on the preceding slide by using the `predict` function on the fit object.

Logistic Regression with a Single Predictor

Interpreting Model Coefficients

Example (CHD vs. AGE)

- Suppose we wish to obtain predicted probabilities for ages 45 through 50.
- We set up a data frame with the new AGE data. Note that you must use the exact same name as the predictor variable in the data frame you analyzed.

```
> my.data <- data.frame(45:50)
> colnames(my.data) <- c("AGE")
> rownames(my.data) <- as.character(my.data$AGE)
```

- Using the predict function is straightforward.
- However, to obtain the values in the correct (probability) metric, we must remember to use the type = "response" option!

```
> predict(fit.chd,newdata = my.data,type="response")

       45        46        47        48        49        50
0.4221950 0.4484776 0.4750511 0.5017666 0.5284721 0.5550155
```

Assessing Model Fit in Logistic Regression

The Deviance Statistic

- In multiple linear regression, the residual sum of squares provides the basis for tests for comparing mean functions.
- In logistic regression, the residual sum of squares is replaced by the *deviance*, which is often called G^2. Suppose there are k data groupings based on $n_i, i = 1, \ldots, k$ binomial observations. The deviance is defined for logistic regression to be

  $$G^2 = 2 \sum_{i=1}^{k} \left[y_i \log \left(\frac{y_i}{\hat{y}_i} \right) + (n_i - y_i) \log \left(\frac{n_i - y_i}{n_i - \hat{y}_i} \right) \right] \tag{7}$$

  where $\hat{y}_i = n_i \hat{p}(\mathbf{x}_i)$ are the fitted numbers of successes in n_i trials in the ith grouping.
- The degrees of freedom associated with the analysis is the number of groupings n used in the calculation minus the number of free parameters in β that were estimated.

Assessing Model Fit in Logistic Regression

Comparing Models

- Comparing models in logistic regression is similar to regular linear regression.
- For two nested models, the difference in deviances is treated as a chi-square with degrees of freedom equal to the difference in the degrees of freedom for the two models.

Assessing Model Fit in Logistic Regression

Test of Model Fit

- When the number of trials $n_i > 1$, the deviance G^2 can be used to provide a goodness-of-fit test for a logistic regression model.
- The test compares the null hypothesis that the mean function used is adequate versus the alternative that a separate parameter needs to be fit for each value of i (this latter case is called the *saturated model*).
- When all the n_i are large enough, G^2 can be compared with the χ^2_{n-p} distribution to get an approximate p-value.

Assessing Model Fit in Logistic Regression

Test of Model Fit

- An alternative statistic is the Pearson X^2

$$
\begin{aligned}
X^2 &= \sum_{i=1}^{k} \left[(y_i - \hat{y}_i)^2 \left(\frac{1}{\hat{y}_i} + \frac{1}{n_i - \hat{y}_i} \right) \right] \\
&= \sum_{i=1}^{k} \frac{n_i (y_i / n_i - \hat{\theta}(\mathbf{x}_i))^2}{\hat{\theta}(\mathbf{x}_i)(1 - \hat{\theta}(\mathbf{x}_i))} \qquad (8)
\end{aligned}
$$

- According to ALR, X^2 and G^2 have the same large-sample distribution and often give the same inferences. But in small samples, there may be differences, and sometimes X^2 may be preferred for testing goodness-of-fit.

Logistic Regression with Several Predictors

- As an example of logistic predictors, Weisberg presents data from the famous Titanic disaster. (Frank Harrell presents a much more detailed analysis of the Titanic in his superb book *Regression Modeling Strategies*).
- Of 2201 known passengers and crew, only 711 are reported to have survived.
- The data in the file `titanic.txt` from Dawson (1995) classify the people on board the ship according to their *Sex* as Male or Female, *Age*, either child or adult, and *Class*, either first, second, third, or crew.
- Not all combinations of the three factors occur in the data, since no children were members of the crew. For each age/sex/class combination, the number of people *M* and the number surviving *Surv* are also reported.
- The data are shown in Table 12.5.

Logistic Regression with Several Predictors

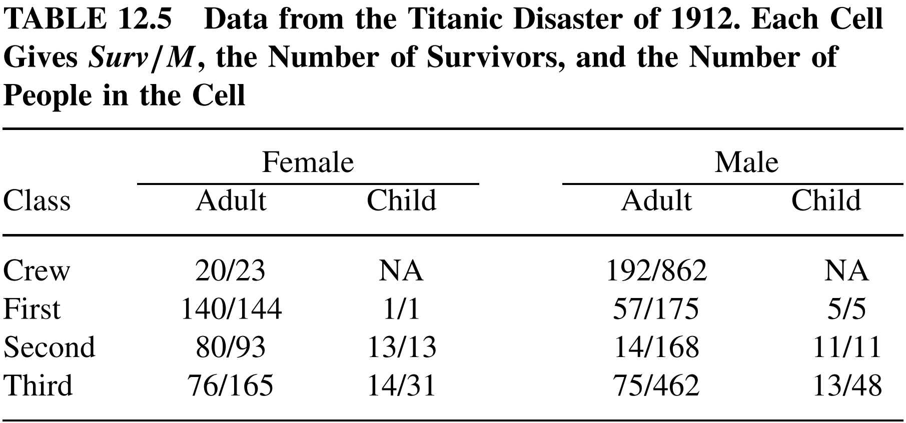


TABLE 12.5 Data from the Titanic Disaster of 1912. Each Cell Gives *Surv*/*M*, the Number of Survivors, and the Number of People in the Cell

	Female		Male	
Class	Adult	Child	Adult	Child
Crew	20/23	NA	192/862	NA
First	140/144	1/1	57/175	5/5
Second	80/93	13/13	14/168	11/11
Third	76/165	14/31	75/462	13/48

Logistic Regression with Several Predictors

- ALR fits a sequence of 5 models to these data.
- Since almost all the m_i exceed 1, we can use either G^2 or X^2 as a goodness-of-fit test for these models.
- The first two mean functions, the main effects only model, and the main effects plus the $Class \times Sex$ interaction, clearly do not fit the data because the values of G^2 and X^2 are both much larger then their df, and the corresponding p-values from the χ^2 distribution are 0 to several decimal places.
- The third model, which adds the $Class \times Age$ interaction, has both G^2 and X^2 smaller than its df, with p-values of about 0.64, so this mean function seems to match the data well.
- Adding more terms can only reduce the value of G^2 and X^2, and adding the third interaction decreases these statistics to 0 to the accuracy shown.
- Adding the three-factor interaction fits one parameter for each cell, effectively estimating the probability of survival by the observed probability of survival in each cell. This will give an exact fit to the data.

Logistic Regression with Several Predictors

```
> mysummary <- function(m){c(df=m$df.residual,G2=m$deviance,
+                            X2=sum(residuals(m,type="pearson")^2) )}
> m1 <- glm(cbind(Surv,N-Surv)~Class+Age+Sex, data=titanic, family=binomial())
> m2 <- update(m1,~.+Class:Sex)
> m3 <- update(m2,~.+Class:Age)
> m4 <- update(m3,~.+Age:Sex)
> m5 <- update(m4,~Class:Age:Sex)
> ans <- mysummary(m1)
> ans <- rbind(ans,mysummary(m2))
> ans <- rbind(ans,mysummary(m3))
> ans <- rbind(ans,mysummary(m4))
> ans <- rbind(ans,mysummary(m5))
>  row.names(ans) <- c( "Main effects only",
+  "Main Effects + Class:Sex",
+  "Main Effects + Class:Sex +  Class:Age",
+  "Main Effects + All 2 Factor Interactions",
+  "Main Effects + All 2 and 3 Factor Interactions")
```

Logistic Regression with Several Predictors

```
> options(scipen=1,digits=3)
> summary(m3)

Call:
glm(formula = cbind(Surv, N - Surv) ~ Class + Age + Sex + Class:Sex + 
    Class:Age, family = binomial(), data = titanic)

Deviance Residuals: 
      1        2        3        4        5        6        7        8  
 0.0000   0.0000   0.0000   0.0001   0.0000   0.0000   0.0000   0.0001  
      9       10       11       12       13       14  
 0.0000   0.0000  -0.8745   0.8265   0.3806  -0.3043  

Coefficients: (1 not defined because of singularities)
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)              1.897      0.619    3.06   0.0022 ** 
ClassFirst               1.658      0.800    2.07   0.0383 *  
ClassSecond             -0.080      0.688   -0.12   0.9073    
ClassThird              -2.115      0.637   -3.32   0.0009 ***
AgeChild                 0.338      0.269    1.26   0.2094    
SexMale                 -3.147      0.625   -5.04  4.7e-07 ***
ClassFirst:SexMale      -1.136      0.821   -1.38   0.1662    
ClassSecond:SexMale     -1.068      0.747   -1.43   0.1525    
ClassThird:SexMale       1.762      0.652    2.70   0.0069 ** 
ClassFirst:AgeChild     22.424  16495.727    0.00   0.9989    
ClassSecond:AgeChild    24.422  13007.888    0.00   0.9985    
ClassThird:AgeChild         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: 671.9622  on 13  degrees of freedom
Residual deviance:   1.6854  on  3  degrees of freedom
AIC: 70.31

Number of Fisher Scoring iterations: 21
```

Logistic Regression with Several Predictors

```
> xtable(ans)
```

	df	G2	X2
Main effects only	8.00	112.57	103.83
Main Effects + Class:Sex	5.00	45.90	42.77
Main Effects + Class:Sex + Class:Age	3.00	1.69	1.72
Main Effects + All 2 Factor Interactions	2.00	0.00	0.00
Main Effects + All 2 and 3 Factor Interactions	0.00	0.00	0.00

Generalized Linear Models

- Both the multiple linear regression model discussed earlier in this book and the logistic regression model discussed in this chapter are particular instances of a *generalized linear model*.
- Generalized linear models all share three basic characteristics:

Generalized Linear Models

1. The distribution of the response Y, given a set of terms X, is distributed according to an exponential family distribution. The important members of this class include the normal and binomial distributions we have already encountered, as well as the Poisson and gamma distributions.
2. The response Y depends on the terms X only through the linear combination $\beta'\mathbf{X}$.
3. The mean $E(Y|X = \mathbf{x}) = \mathbf{m}(\beta'\mathbf{x})$ for some kernel mean function $\mathbf{m}$. For the multiple linear regression model, $\mathbf{m}$ is the identity function, and for logistic regression, it is the logistic function. There is considerable flexibility in selecting the kernel mean function. Most presentations of generalized linear models discuss the link function, which technically is defined as the *inverse* of $\mathbf{m}$ rather than $\mathbf{m}$ itself.