Chapter 14: Logistic regression

- Model for logistic regression
- Fitting and interpreting the model
- Inference for logistic regression
- Multiple logistic regression

14.1 Problems when response variable is binary

We will study methods to model relationships when the response variable has only two possible values.
For example: customer buys or does not buy, patient lives or dies.
We call the two values of the response variable 'success' and 'failure'.

When the response variable is binary, we cannot use what we learned in chapters 1 to 10 due to the following problems:

- Nonnormal error terms
- Nonconstant error variance
- Constraints on response function

14.3 Simple logistic regression

Example 1 (drinking behavior): A survey of 17,096 students in U.S. four-year colleges collected information on drinking behavior and alcohol-related problems. The researchers define "frequent binge drinking" as having five or more drinks in a row, three or more times in the past two weeks. X represents the number of binge drinkers in the sample.

One possible explanatory variable is gender of the student:

Population	n	$\underline{X}$	$\hat{p}$
1 (men)	7,180	1,630	0.227
2 (women)	9,916	1,684	0.170
Total	17,096	3,314	0.194

p = **probability of success**

odds = probability of success / probability of failure = p/(1-p)

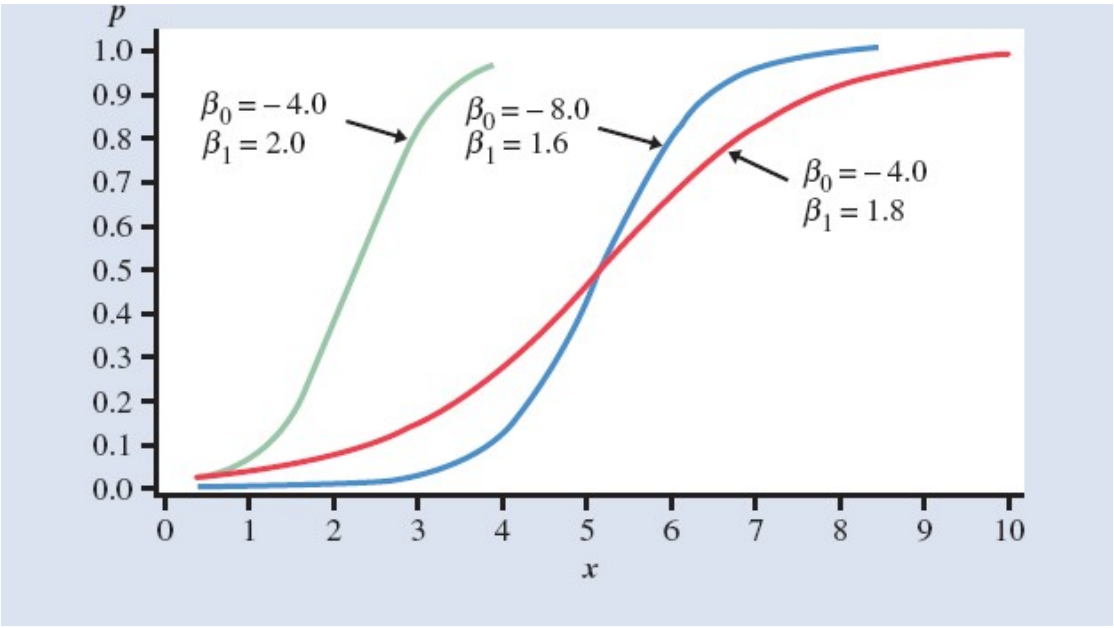
- A similar formula for sample odds is obtained by substituting the sample proportion for p.
- The estimated odds of a male student being a frequent binge drinker are: 0.227/(1 - 0.227) = 0.2937.约等于1/3

- Since 0.29 is approximately 1/3, we could say that the odds that a college male student is a frequent binge drinker are 1 to 3. In a similar way, we could describe the odds that a college male student is *not* a frequent binge drinker as 3 to 1.
- The estimated odds of a female student being a frequent binge drinker are: 0.1698/(1 - 0.1698) = 0.2045.

The logistic regression model works with the natural log of the odds, p/(1 - p).

- The logistic regression model works with the *natural* log of the odds, $p/(1-p)$.
- We use the term log odds for this transformation.
- As p moves from 0 to 1, the log odds moves through all negative and positive numerical values.
- We model the log odds as a linear function of the explanatory variable:

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



Example 1 (drinking behavior; continued):

- The explanatory variable gender can be expressed numerically using an indicator variable: x = I if student is man, 0 if student is women
- Since log odds for men = -1.23, and log odds for women = -1.59, we get the parameter estimates: b_0 = -1.59, and b_1 = -1.23 - (-1.59) = 0.36.
- The fitted logistic model is: log (odds) = -1.59 + 0.36*x*
- In general, the calculations needed to find the parameter estimates are complex and require software.
-

- Most people are not comfortable thinking in the log(odds) scale so we apply a transformation.
- The exponential function (e^x) reverses the natural log transformation. Applying the transformation we get:
  odds = $e^{-1.59 + 0.36x} = (e^{-1.59})(e^{0.36x})$
- From this, the ratio of the odds for men (x =1) and women (x = 0) is
  Odds Ratio = $odds_{men}/odds_{women} = e^{0.36} = 1.43$
- This transformation transforms the logistic regression slope into an odds ratio, i.e. the odds that a man is a frequent binge drinker are 1.43 times the odds for a woman.

Example 2 (honors class): Logistic regression with a single continuous predictor

It describes the relationship between students' math scores and the log odds of being in an honors class:

$\log(p/(1-p)) = \beta_0 + \beta_1$***math.***

Suppose we get the estimated regression function as the following:

$\log(p/(1-p)) = \text{logit}(p) = -9.793942 + .1563404*\text{math}$

Fix two levels of math, i.e., math =54 and math =55, we have

$\log(p/(1-p))(\text{math}=55) - \log(p/(1-p))(\text{math} = 54) = .1563404.$

We can say now that the coefficient for math is the difference in the log odds. In other words, for a one-unit increase in the math score, the expected change in log odds is .1563404.

If we exponentiate both sides of our last equation, we have the following:
odds(math=55)/odds(math=54) = exp(.1563404) = 1.1692241.

So we can say for a one-unit increase in math score, we expect to see about 17% increase in the odds of being in an honors class. This 17% of increase does not depend on the value that math is held at.

14.4 Multiple logistic regression

- In multiple logistic regression, the response variable has two possible values, as in logistic regression, but there can be several explanatory variables.
- As in multiple regression, there is an overall test for all of the explanatory variables.
- The null hypothesis that the coefficients of all the explanatory variables are zero is tested by a statistic that is approximately χ^2 with degrees of freedom = number of explanatory variables.
- Hypotheses about individual coefficients are tested by a statistic that is approximately χ^2 with 1 degree of freedom or z-test.

logistic regression examples using R

http://rstudio-pubs-static.s3.amazonaws.com/5228_1c97747da7f04cf7b18dab73822e8d7c.html

Example 1 A system analyst studied the effect of computer programming experience on ability to complete within a specified time a complex programming task, including debugging. Twenty-five persons were selected with varying amounts of programming experiences (measured in months of experience). The results were coded in binary fashion: Y=1 if the task was completed successfully in the allotted time, Y=0 if the task was not completed successfully.

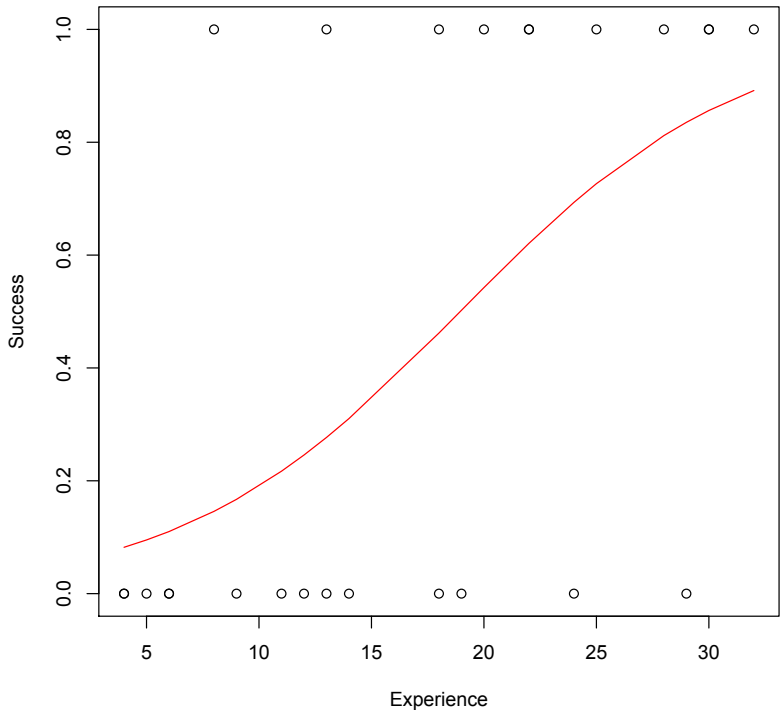
```
> Data <- read.table("CH14TA01.txt",header=FALSE)
> colnames(Data) <- c("Experience", "Success", "Fitted")
> Data
   Experience Success   Fitted
1          14       0 0.310262
2          29       0 0.835263
3           6       0 0.109996
4          25       1 0.726602
5          18       1 0.461837
6           4       0 0.082130
```

generalized linear model

```
> glm.out = glm(Success~Experience, family=binomial(logit),
data=Data)
> summary(glm.out)

Call:
glm(formula = Success ~ Experience, family = binomial(logit),
    data = Data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8992  -0.7509  -0.4140   0.7992   1.9624

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.05970    1.25935  -2.430   0.0151 *
Experience   0.16149    0.06498   2.485   0.0129 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34.296  on 24  degrees of freedom
Residual deviance: 25.425  on 23  degrees of freedom
AIC: 29.425

Number of Fisher Scoring iterations: 4
```

```
>
> plot(Success~Experience, data=Data)
> lines(Data$Experience[order(Data$Experience)],
glm.out$fitted[order(Data$Experience)], type="l", col="red")
> title(main="Table 14.1 Data with Fitted Logistic Regression
Line")
>
```

Table 14.1 Data with Fitted Logistic Regression Line

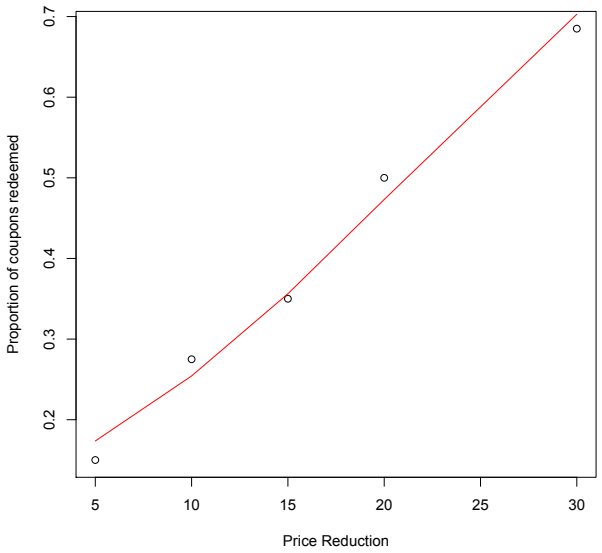


Example 2: repeated observations -- binary outcomes

In a study of the effectiveness of coupons offering a price reduction on a given product, 1000 homes were selected at random. A packet containing advertising material and a coupon for the product were mailed to each home. The coupons were offered different price reductions (5, 10, 15, 20 and 30 dollars), and 200 homes were assigned at random to each of the price reduction categories. The predictor variable X is the amount of price reduction, and Y is a binary variable indicating whether or not the coupon was redeemed within a six-month period.

```
> Data <- read.table("CH14TA02.txt",header=FALSE)
> colnames(Data) <- c("PriceReduction", "NumHouseholds",
 "NumCoupons","PropCoupons")
> Data
  PriceReduction NumHouseholds NumCoupons PropCoupons
1              5           200         30       0.150
2             10           200         55       0.275
3             15           200         70       0.350
4             20           200        100       0.500
5             30           200        137       0.685
> glm.out = glm(cbind(NumCoupons, NumHouseholds-NumCoupons)~
 PriceReduction, family=binomial(logit), data=Data)
> summary(glm.out)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.044348   0.160977  -12.70   <2e-16 ***
PriceReduction  0.096834   0.008549   11.33   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 14.2 Data with Fitted Logistic Regression Line



Chapter 14: logistic regression Example 3 # multiple regression

In a health study to investigate an epidemic outbreak of a disease that is spread by mosquitoes, individuals were randomly sampled within two sectors in a city to determine if the person has recently contracted the disease under study. The response variable Y was coded 1 if this disease was determined to have been present, and 0 if not. The predictors are age, socioeconomic status of household, and sector within city.

```
> Data <- read.table("CH14TA03.txt",header=FALSE)
> colnames(Data) <- c("Case", "Age",
 "Status1","Status2","CitySector","Disease")
> glm.out = glm(Disease ~ Age+Status1+Status2+CitySector,
 family=binomial(logit), data=Data)
> summary(glm.out)

Call:
glm(formula = Disease ~ Age + Status1 + Status2 + CitySector,
    family = binomial(logit), data = Data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6552  -0.7529  -0.4788   0.8558   2.0977

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.31293    0.64259  -3.599 0.000319 ***
Age          0.02975    0.01350   2.203 0.027577 *
Status1      0.40879    0.59900   0.682 0.494954
Status2     -0.30525    0.60413  -0.505 0.613362
CitySector   1.57475    0.50162   3.139 0.001693 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 122.32  on 97  degrees of freedom
Residual deviance: 101.05  on 93  degrees of freedom
AIC: 111.05

Number of Fisher Scoring iterations: 4

>
> anova(glm.out, test='Chisq')
Analysis of Deviance Table

Model: binomial, link: logit

Response: Disease
```

```
Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                          97     122.32
Age         1   7.4050        96     114.91 0.006504 **
Status1     1   1.8040        95     113.11 0.179230
Status2     1   1.6064        94     111.50 0.205003
CitySector  1  10.4481        93     101.05 0.001228 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> ###### prediction for a new observation
> new <- data.frame(Age=33, Status1=0, Status2=0, CitySector=0)
> y.hat <- predict(glm.out, new)  #predict log(p/(1-p))
> p.hat <- exp(y.hat)/(1+exp(y.hat))
> p.hat
       1
0.208964
>
> # or use option of "response" directly
> p.hat <-predict(glm.out, new, type="response")
>
> ###### Compare two models
> glm2.out = glm(Disease ~ Age+CitySector,
 family=binomial(logit), data=Data)
> anova(glm2.out, glm.out, test="Chisq")
Analysis of Deviance Table

Model 1: Disease ~ Age + CitySector
Model 2: Disease ~ Age + Status1 + Status2 + CitySector
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        95     102.26
2        93     101.05  2   1.2052   0.5474
>
>
> ######################
> ## stepwise selection
> fullmod = glm(Disease ~ Age+Status1+Status2+CitySector,
 family=binomial(logit), data=Data)
> nothing <- glm(Disease ~ 1,family=binomial, data=Data)
> #Choose a model by AIC in a Stepwise Algorithm
> bothways = step(nothing,
 list(lower=formula(nothing),upper=formula(fullmod)),
+                 direction="both",trace=1)
Start:  AIC=124.32
```

```
Disease ~ 1

             Df Deviance    AIC
+ CitySector  1   107.53 111.53
+ Age         1   114.91 118.91
+ Status2     1   118.23 122.23
<none>            122.32 124.32
+ Status1     1   120.88 124.88

Step:  AIC=111.53
Disease ~ CitySector

             Df Deviance    AIC
+ Age         1   102.26 108.26
<none>            107.53 111.53
+ Status2     1   106.37 112.37
+ Status1     1   106.88 112.88
- CitySector  1   122.32 124.32

Step:  AIC=108.26
Disease ~ CitySector + Age

             Df Deviance    AIC
<none>            102.26 108.26
+ Status1     1   101.31 109.31
+ Status2     1   101.52 109.52
- Age         1   107.53 111.53
- CitySector  1   114.91 118.91
> formula(bothways)
Disease ~ CitySector + Age
>
>
>
> ### split data into training and testing
> n <- dim(Data)[1]
> library(caTools)
> split <- sample.split(Data$Disease, SplitRatio=3/4)
> training <- subset(Data, split==TRUE)
> testing <- subset(Data, split==FALSE)
>
> fullmod.training = glm(Disease ~
 Age+Status1+Status2+CitySector, family=binomial(logit),
 data=training)
> nothing.training <- glm(Disease ~ 1,family=binomial,
 data=testing)
> #Choose a model by AIC in a Stepwise Algorithm
> bothways.training = step(nothing,
```

```
 list(lower=formula(nothing),upper=formula(fullmod)),
+                  direction="both",trace=0)
> formula(bothways.training)
Disease ~ CitySector + Age
> model <- glm(formula(bothways.training),
 family=binomial(logit), data=training)
> predicted <- predict(model, training, type="response")
> hist(predicted)
>
> # how to assign 1 and 0 according to the predicted
 probabilities
> table(Truth=training$Disease, Prediction=predicted>=0.5)
     Prediction
Truth FALSE TRUE
    0    45    5
    1    15    8
> table(Truth=training$Disease, Prediction=predicted>=0.4)
     Prediction
Truth FALSE TRUE
    0    38   12
    1     8   15
>
> # Use ROC curve
> library(ROCR)

> ROCRpred <- prediction(predicted, training$Disease)
> ROCRpref <- performance(ROCRpred, measure="tpr",x.measure =
 "fpr")
> plot(ROCRpref, colorize=TRUE,
 print.cutoffs.at=seq(0,1,by=0.1))
> abline(0,1)
> ## prediction on testing data
> pred.test <- predict(model, testing, type="response")
> table(Truth=testing$Disease, Prediction=pred.test>=0.4)
     Prediction
Truth FALSE TRUE
    0    13    4
    1     2    6
```