Simple Logistic Regression

Chapter 9, Lab 1: Solutions

OpenIntro Biostatistics

Topics

- Odds and probabilities
- Simple logistic regression

This lab introduces simple logistic regression, a model for the association of a binary response variable with a single predictor variable. Logistic regression generalizes methods for two-way tables and allows for the use of a numerical predictor.

The material in this lab corresponds to Section 9.xx in *OpenIntro Biostatistics*.

Introduction

Odds and probabilities

If the probability of an event A is p, the odds of the event are

$$\frac{p}{1-p}.$$

Given the odds of an event A, the probability of the event is

$$\frac{\text{odds}}{1+\text{odds}}.$$

Simple logistic regression

Suppose that Y is a binary response variable and X is a predictor variable, where $Y = 1$ represents the particular outcome of interest.

The model for a single variable logistic regression, where $p(x) = P(Y = 1|X = x)$, is

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

The estimated model equation has the form

$$\log\left[\frac{\hat{p}(x)}{1-\hat{p}(x)}\right] = b_0 + b_1 x,$$

where b_0 and b_1 are estimates of the model parameters β_0 and β_1.

Background Information

Patients admitted to an intensive care unit (ICU) are either extremely ill or considered to be at great risk of serious complications. There are no widely accepted criteria for distinguishing between patients who should be admitted to an ICU and those for whom admission to other hospital units would be more appropriate. Thus, among different ICUs, there are wide ranges in a patient's chance of survival. When studies are done to compare effectiveness of ICU care, it is critical to have a reliable means of assessing the comparability of the different patient populations.

One such strategy for doing so involves the use of statistical modeling to relate empirical data for many patient variables to outcomes of interest. The following dataset consists of a sample of 200 subjects who were part of a much larger study on survival of patients following admission to an adult ICU.[1] The major goal of the study was to develop a logistic regression model to predict the probability of survival to hospital discharge.[2]

The following table provides a list of the variables in the dataset and their description. The data are accessible as the `icu` dataset in the `aplore3` package.

Variable	Description
`id`	patient ID number
`sta`	patient status at discharge, either `Lived` or `Died`
`age`	age in years (when admitted)
`gender`	gender, either `Male` or `Female`
`can`	cancer part of current issue, either `No` or `Yes`
`crn`	history of chronic renal failure, either `No` or `Yes`
`inf`	infection probable at admission, either `No` or `Yes`
`cpr`	CPR prior to admission, either `No` or `Yes`
`sys`	systolic blood pressure at admission, in mm Hg
`hra`	heart rate at admission, in beats per minute
`pre`	previous admission to an ICU within 6 months, either `No` or `Yes`
`type`	type of admission, either `Elective` or `Emergency`
`fra`	long bone, multiple, neck, single area, or hip fracture, either `No` or `Yes`
`po2`	PO_2 from initial blood gases, either `60` or `<=60`, in mm Hg
`ph`	pH from initial blood gases, either `>= 7.25` or `< 7.25`
`pco`	PCO_2 from initial blood gases, either `<= 45` or `>45`, in mm Hg
`bic`	HCO_3 (bicarbonate) from initial blood gases, either `>= 18` or `<18`, in mm Hg
`crea`	creatinine from initial blood gases, either `<= 2.0` or `> 2.0`, in mg/dL
`loc`	level of consciousness at admission, either `Nothing`, `Stupor`, or `Coma`

[1] From Hosmer D.W., Lemeshow, S., and Sturdivant, R.X. *Applied Logistic Regression*. 3^{rd} ed., 2013.

[2] Lemeshow S., et al. Predicting the outcome of intensive care unit patients. *Journal of the American Statistical Association* 83.402 (1988): 348-356.

Odds and probabilities

1. Create a two-way table of survival to discharge by whether CPR was administered prior to admission. The template provides code for re-leveling the sta variable such that 0 corresponds to Died and 1 corresponds to Lived.

```
#install the package (only needs to be done once)
install.packages("aplore3")
```

```
#load the data
library(aplore3)
data("icu")

#relevel survival so that 1 corresponds to surviving to discharge
icu$sta = factor(icu$sta, levels = rev(levels(icu$sta)))

#create two-way table
addmargins(table(icu$sta, icu$cpr,
             dnn = c("Survival", "Prior CPR")))
```

```
##         Prior CPR
## Survival  No Yes Sum
##    Died   33   7  40
##    Lived 154   6 160
##    Sum   187  13 200
```

a) Calculate the odds of survival to discharge for those who did not receive CPR prior to ICU admission. Is someone who did not receive CPR prior to admission more likely to survive to discharge than to not survive to discharge?

The odds of survival to discharge for those who did not receive CPR prior to ICU admission are 4.67, which corresponds to a probability of 0.82. Someone who does not receive CPR prior to admission is more likely to survive to discharge than die before discharge. An odds greater than 1 corresponds to probability greater than 50%.

```
#calculate odds
odds = 154/33
odds
```

```
## [1] 4.666667
```

```
#convert to probability
p = (odds)/(1 + odds)
p
```

```
## [1] 0.8235294
```

b) Calculate the odds of survival to discharge for those who received CPR prior to ICU admission. Is someone who received CPR prior to admission more likely to survive to discharge than not?

The odds of survival to discharge for those who receive CPR prior to ICU admission are

0.857, which corresponds to a probability of 0.46. Someone who receives CPR prior to admission is less likely to survive to discharge than survive to discharge. An odds less than 1 corresponds to probability less than 50%.

```
#calculate odds
odds = 6/7
odds
```

```
## [1] 0.8571429
```

```
#convert to probability
p = (odds)/(1 + odds)
p
```

```
## [1] 0.4615385
```

c) Calculate the odds ratio of survival to discharge, comparing patients who receive CPR prior to admission to those who do not receive CPR prior to admission.

   The odds ratio of survival to discharge, comparing patients who receive CPR prior to admission to those who do not receive CPR prior to admission is 0.857/4.667 = 0.184.

2. Creatinine level in the data are recorded as being either less than or equal to 2.0 mg/dL or greater than 2.0 mg/dL. A typical creatinine level is between 0.5 - 1.0 mg/dL, and elevated creatinine may be a sign of renal failure.

```
#create two-way table
addmargins(table(icu$sta, icu$cre,
       dnn = c("Survival", "Creatinine")))
```

```
##          Creatinine
## Survival <= 2.0 > 2.0 Sum
##    Died      35     5  40
##    Lived    155     5 160
##    Sum      190    10 200
```

a) Calculate the odds of survival to discharge for patients who have a creatinine level less than or equal to 2.0 mg/dL. From the odds, calculate the probability of survival to discharge for these patients.

   The odds of survival to discharge for patients with creatinine level less than or equal to 2.0 mg/dL are 4.43, which corresponds to a probability of 0.82.

```
odds = 155/35
odds
```

```
## [1] 4.428571
```

```
p = (odds)/(1 + odds)
p
```

```
## [1] 0.8157895
```

b) Calculate the probability of survival to discharge for patients who have a creatinine level greater than 2.0 mg/dL. From the probability, calculate the odds of survival to discharge for these patients.

The probability of survival to discharge for patients who have a creatinine level greater than 2.0 mg/dL is 0.50, which corresponds to odds of 1. Survival is as equally likely as death.

```
p = 5/10
p
```

```
## [1] 0.5
```

```
odds = p/(1 - p)
odds
```

```
## [1] 1
```

c) Compute and interpret the odds ratio of survival to discharge, comparing patients with creatinine > 2.0 mg/dL to those with creatinine ≤ 2.0 mg/dL.

The odds ratio of survival to discharge, comparing patients with creatinine > 2.0 mg/dL to those with creatinine ≤ 2.0 mg/dL is 4.43. The odds of survival to discharge for patients with relatively lower creatinine level are over 4 times as large as the odds of survival for patients with creatinine elevated past 2.0 mg/dL.

Simple logistic regression

3. Fit a logistic regression model to predict survival to discharge from prior CPR.

```
#fit a model
glm(sta ~ cpr, data = icu, family = binomial(link = "logit"))$coef
```

```
## (Intercept)      cprYes
##    1.540445   -1.694596
```

a) Write the model equation estimated from the data.

$$\log\left[\frac{\hat{p}(\text{status} = \text{lived}|\text{cpr})}{1-\hat{p}(\text{status} = \text{lived}|\text{cpr})}\right] = 1.540 - 1.695(cpr_{yes})$$

$$\log(\widehat{\text{odds}}\text{ of lived} \mid \text{cpr}) = 1.540 - 1.695(cpr_{yes})$$

b) Interpret the intercept. Confirm that the interpretation coheres with the answer to Question 1, part a).

The intercept represents the log of the estimated odds of survival to discharge for individuals who did not receive CPR prior to ICU admission; thus, the estimated odds of survival for this group are exp(1.540) = 4.67.

c) Interpret the slope coefficient. Compute the exponential of the slope coefficient and confirm that this matches the answer to Question 1, part c).

The slope coefficient represents the change in the log of the estimated odds of survival to discharge from the no CPR group to the CPR group; $\exp(-1.695) = 0.184$, which represnts the estimated odds ratio for survival to discharge, comparing those who received CPR prior to admission to those who did not.

d) Compute and interpret an odds ratio that summarizes the association between survival to discharge and prior CPR.

Either the odds ratio from part c) or its inverse $\exp(1.695) = 5.45$ are a summary of the association between survival to discharge and prior CPR. The inverse is the estimated odds ratio for survival to discharge, comparing those who did not receive CPR prior to admission to those who did; this odds ratio indicates that the odds of survival in the no CPR group are over 5 times as large as the odds of survival in the CPR group.

e) Is there evidence of a statistically significant association between survival to discharge and prior CPR at $\alpha = 0.05$?

Yes, the p-value is 0.004, which is less than $\alpha = 0.05$. There is sufficient evidence to reject $H_0 : \beta_1 = 0$ in favor of the alternative. There is evidence of an association between receiving CPR prior to ICU admission and lower probability of survival to discharge.

```r
#use summary(glm())
summary(glm(sta ~ cpr, data = icu, family = binomial(link = "logit")))
```

```
## 
## Call:
## glm(formula = sta ~ cpr, family = binomial(link = "logit"), data = icu)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.8626   0.6231   0.6231   0.6231   1.2435  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)   1.5404     0.1918   8.031 9.71e-16 ***
## cprYes       -1.6946     0.5885  -2.880  0.00398 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 200.16  on 199  degrees of freedom
## Residual deviance: 192.23  on 198  degrees of freedom
## AIC: 196.23
## 
## Number of Fisher Scoring iterations: 4
```

4. Fit a logistic regression model to predict survival to discharge from an indicator of elevated creatinine.

```
#fit the model
glm(sta ~ cre, data = icu, family = binomial(link = "logit"))$coef
```

```
## (Intercept)    cre> 2.0 
##    1.488077   -1.488077
```

a) Write the model equation estimated from the data.

$$\log\left[\frac{\hat{p}(\text{status} = \text{lived}|\text{creatinine})}{1-\hat{p}(\text{status} = \text{lived}|\text{creatinine})}\right] = 1.488 - 1.488(cre_{>2.0})$$

$$\log(\widehat{\text{odds}} \text{ of lived} \mid \text{creatinine}) = 1.488 - 1.488(cre_{>2.0})$$

b) Interpret the intercept and slope coefficient.

The intercept is the log odds of survival to discharge for individuals with creatinine less than or equal to 2.0 mg/dL. The slope coefficient is the difference in the odds of survival to discharge between the groups defined by creatinine; log odds are 1.488 lower in the group with creatinine higher than 2.0 mg/dL

c) Compute and interpret an odds ratio that summarizes the association between survival to discharge and an indicator of elevated creatinine.

The odds ratio of survival to discharge, comparing those with lower creatinine to those with higher creatinine, is 4.43; the odds of survival to discharge are over 4 times as large in the group with creatinine less than 2.0 mg/dL.

d) Is there evidence of a statistically significant association between survival to discharge and an indicator of elevated creatinine at $\alpha = 0.05$?

Yes, the p-value is 0.0024, which is less than $\alpha = 0.05$. There is sufficient evidence to reject $H_0 : \beta_1 = 0$ in favor of the alternative. There is evidence of an association between creatinine level higher than 2.0 mg/dL and lower probability of survival to discharge.

```
#fit the model
summary(glm(sta ~ cre, data = icu, family = binomial(link = "logit")))
```

```
## 
## Call:
## glm(formula = sta ~ cre, family = binomial(link = "logit"), data = icu)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.8394   0.6381   0.6381   0.6381   1.1774  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)   1.4881     0.1871   7.951 1.84e-15 ***
## cre> 2.0     -1.4881     0.6596  -2.256   0.0241 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 200.16  on 199  degrees of freedom
## Residual deviance: 195.40  on 198  degrees of freedom
## AIC: 199.4
##
## Number of Fisher Scoring iterations: 4
```

5. Fit a logistic regression model to predict survival to discharge from age.

```
#fit the model
glm(sta ~ age, data = icu, family = binomial(link = "logit"))$coef
```

```
## (Intercept)         age
##  3.05851323 -0.02754261
```

a) Write the model equation estimated from the data.

$$\log\left[\frac{\hat{p}(\text{status} = \text{lived}|\text{age})}{1 - \hat{p}(\text{status} = \text{lived}|\text{age})}\right] = 3.059 - 0.028(age)$$

$$\log(\widehat{\text{odds}} \text{ of lived} \mid \text{age}) = 3.059 - 0.028(age)$$

b) Does the intercept have a meaningful interpretation in the context of the data?

The intercept represents the log odds of survival to discharge for an individual of age 0 years admitted to the ICU. Since the youngest age observed in the dataset is 16 years, the intercept does not represent a reliable estimate of the odds of survival to discharge for a newborn who needs intensive care.

c) Interpret the slope coefficient.

The intercept represents that an increase in age of 1 year is associated with a decrease of 0.028 in the log odds of survival to discharge.

d) Calculate the odds of survival to discharge for a 70-year-old individual. Is a 70-year-old individual more likely to survive than not?

The log odds of survival to discharge for a 70-year-old individual are $3.059 - 0.028(70) = 1.305$, thus the odds of survival to discharge are $e^{1.305} = 3.10$. A 70-year-old individual is more likely to survive than not, since odds greater than 1 are associated with probability greater than 0.50.

```
#use predict()
icu.model.age = glm(sta ~ age, data = icu, family = binomial(link = "logit"))
log.odds = predict(icu.model.age, newdata = data.frame("age" = 70))

exp(log.odds)
```

```
##        1
## 3.097299
```

e) Calculate the odds ratio of survival to discharge comparing a 45-year-old individual to a 70-year-old individual.

The odds ratio can be calculated directly from the model slope, or from calculating the odds at each age then dividing to obtain the ratio. The odds ratio of survival to discharge comparing a 45-year-old individual to a 70-year-old individual is 1.99; the odds of survival to discharge are almost twice as large for a 45-year-old than a 70-year-old.

```
#use r as a calculator
difference.odds = icu.model.age$coef[2]*(70-45)
exp(difference.odds)
```

```
##       age
## 0.5022962
```

```
exp(-difference.odds)
```

```
##      age
## 1.990857
```

```
#alternatively, use predict()
log.odds.70 = predict(icu.model.age, newdata = data.frame(”age” = 70))
log.odds.45 = predict(icu.model.age, newdata = data.frame(”age” = 45))
exp(log.odds.45)/exp(log.odds.70)
```

```
##        1
## 1.990857
```