Multiple Logistic Regression

Chapter 9, Lab 2: Solutions

OpenIntro Biostatistics

Topics

- Multiple logistic regression
- AIC (Akaike Information Criterion)

This lab introduces multiple logistic regression, a model for the association of a binary response variable with several predictor variable. The use of the Akaike Information Criterion (AIC) for comparing logistic regression models is also discussed.

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

Introduction

Multiple logistic regression

The principles of simple logistic regression can be extended to a multiple logistic regression model with several predictors. Suppose that Y is a binary response variable, where $Y = 1$ represents the particular outcome of interest, and $X_1, X_2, \ldots, X_p$ are predictor variables.

The model for multiple logistic regression, where $p(x) = P(Y = 1|x_1, x_2, ..., x_p)$, is

$$\log\left[\frac{p(x)}{1-p(x)}\right] = \beta_0 + \beta_1 x_2 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

The estimated model equation has the form

$$\log\left[\frac{\hat{p}(x)}{1-\hat{p}(x)}\right] = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p,$$

where $b_0, b_1, b_2, \ldots, b_p$ are estimates of the model parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$.

The coefficient b_j of a predictor x_j is the predicted change in the log of the estimated odds corresponding to a one unit change in x_j, when the values of all other predictors remain constant.

AIC (Akaike Information Criterion)

The **AIC (Akaike Information Criterion)** can be used to compare models. It is analogous to the adjusted R^2 for linear regression in that it penalizes a model for having a larger number of predictors.

A *lower* AIC is indicative of a more parsimonious model.

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
`cre`	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.

1. Previously, age and CPR prior to ICU admission were each found to be statistically significantly associated with survival to discharge; an indicator for creatinine elevated beyond 2.0 mg/dL was also significantly associated with survival to discharge.

   Fit a single model to predict survival to discharge (sta) from age (age), CPR prior to admission (cpr), and an indicator of elevated creatinine level (cre).

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

```
## Warning: package 'aplore3' was built under R version 3.5.3
```

```
data("icu")

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

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

```
##
## Call:  glm(formula = sta ~ age + cpr + cre, family = binomial(link = "logit"),
##     data = icu)
##
## Coefficients:
## (Intercept)          age       cprYes     cre> 2.0
##     3.32901     -0.02814     -1.69680     -1.13328
##
## Degrees of Freedom: 199 Total (i.e. Null);  196 Residual
## Null Deviance:       200.2
## Residual Deviance: 181.5     AIC: 189.5
```

   a) Write the model equation.

$$\log\left[\frac{\hat{p}(\text{status = lived}|\text{age, cpr, cre})}{1-\hat{p}(\text{status = lived}|\text{age, cpr, cre})}\right] = 3.329 - 0.0281(age) - 1.697(cpr_{yes} - 1.133(cre_{>2.0})$$

$$\log(\widehat{\text{odds of lived}} \mid \text{age, cpr, creatinine}) = 3.329 - 0.0281(age) - 1.697(cpr_{yes} - 1.133(cre_{>2.0})$$

   b) Interpret the slope coefficients.

   The slope coefficient for age indicates that an increase in 1 year of age is associated with a decrease of 0.0281 in the estimated log odds of survival to discharge, when the values of the other variables do not change.

   The slope coefficient for prior CPR indicates that the estimated log odds of survival to discharge are 1.697 lower in individuals who receive CPR prior to admission (versus not receiving CPR), when the values of the other variables do not change.

The slope coefficient for creatinine indicates that the estimated log odds of survival to discharge are 1.133 lower in individuals with creatinine higher than 2.0 mg/dL (versus lower than or equal to), when the values of the other variables do not change.

c) Make predictions.

i. Compare the odds of survival for those who did receive CPR prior to admission to those who did not receive CPR prior to admission.

The odds ratio, comparing those who did receive CPR to those who did not, can be directly calculated from the slope coefficient: $\exp(-1.697) = 0.183$. It can be easier to interpret the reciprocal, $\exp(1.697) = 5.46$. The odds of survival to discharge for those who do not receive CPR prior to admission are almost 6 times as large as those for patients who do receive CPR prior to ICU admission.

ii. Compare the odds of survival for those who had elevated creatinine level beyond 2.0 mg/dL to those who did not.

The odds ratio, comparing those who did have creatinine above 2.0 mg/dL to those who did not, can be directly calculated from the slope coefficient: $\exp(-1.133) = 0.322$, or $\exp(1.133) = 3.105$. The odds of survival to discharge for those who have creatinine under or equal to 2.0 mg/dL are about 3 times as large as the odds of survival to discharge for those who have creatinine over 2.0 mg/dL.

iii. Calculate the odds of survival for a 65-year-old individual who did not receive CPR prior to admission and had creatinine level of 1.1 mg/dL. Is this individual more likely to survive than not survive?

The odds of survival for the described individual are $\exp(3.329 - (0.0281)(65) - (1.697)(0) - (1.133)(0)) = 4.480$. Since odds greater than 1 correspond to probability greater than 0.50, this individual is more likely to survive than not survive.

iv. Compare the odds of survival for a 30-year-old individual who received CPR prior to admission and had creatinine level of 1.5 mg/dL to the odds of survival for the individual described in part iii.

The odds of survival for the described 30-year-old individual are $\exp(3.329 - (0.0281)(30) - (1.697)(1) - (1.133)(0)) = 2.199$. The odds ratio, comparing the 65-year-old to the 30-year-old, is 2.04; the odds of the older individual surviving to discharge are about twice as large as the those of the younger individual surviving to discharge.

```
#part i.
exp(-1.697); exp(1.697)
```

```
## [1] 0.1832324
```

```
## [1] 5.45755
```

```
#part ii.
exp(-1.133); exp(1.133)
```

```
## [1] 0.3220656
```

```
## [1] 3.104957
```

```r
#part iii.
model.1 = glm(sta ~ age + cpr + cre, data = icu,
          family = binomial(link = "logit"))
log.odds = predict(model.1, newdata = data.frame(age = 65,
                                                 cpr = "No",
                                                 cre = "<= 2.0"))
exp(log.odds)
```

```
##        1
## 4.479995
```

```r
#part iv.
log.odds.a = predict(model.1, newdata = data.frame(age = 65,
                                                 cpr = "No",
                                                 cre = "<= 2.0"))
exp(log.odds.a)
```

```
##        1
## 4.479995
```

```r
log.odds.b = predict(model.1, newdata = data.frame(age = 30,
                                                 cpr = "Yes",
                                                 cre = "<= 2.0"))
exp(log.odds.b)
```

```
##        1
## 2.198713
```

```r
exp(log.odds.a)/exp(log.odds.b)
```

```
##        1
## 2.037554
```

d) Identify the significant predictors of survival to discharge from this model.

The significant predictors of survival to discharge are age and prior CPR; each have p-value lower than 0.05.

```r
summary(model.1)$coef
```

```
##                Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)  3.32900649 0.74883543  4.445578 8.765577e-06
## age         -0.02814438 0.01124839 -2.502082 1.234655e-02
## cprYes      -1.69680316 0.62144856 -2.730400 6.325752e-03
## cre> 2.0    -1.13327763 0.70191415 -1.614553 1.064075e-01
```

2. Fit a model for predicting survival to discharge from age, CPR prior to admission, and the interaction between age and CPR prior to admission.

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

```
##
```

```
## Call:  glm(formula = sta ~ age * cpr, family = binomial(link = "logit"),
##     data = icu)
##
## Coefficients:
## (Intercept)          age       cprYes   age:cprYes
##     3.04196     -0.02477      3.72294     -0.09419
##
## Degrees of Freedom: 199 Total (i.e. Null);  196 Residual
## Null Deviance:       200.2
## Residual Deviance: 181.1     AIC: 189.1
```

a) Interpret the slope of the interaction term.

The slope of the interaction term indicates the association between estimated log odds of survival to discharge and age is more negative for individuals who received CPR prior to admission compared to those who did not receive CPR prior to admission. In other words, an increase in age is associated with a greater decrease in log odds of survival for patients who received CPR. The slope coefficient for age is -0.0248 in the group who did not receive prior CPR, and $-0.0248 - 0.0942 = -0.119$ in the group who received prior CPR.

b) Assess whether this model is a better parsimonious model than the one from Question 1.

The AIC of this model is 189.1, while the AIC of the earlier model is 189.5. By only a very slim margin, AIC indicates that this model is a better parsimonious model.

```
model.interact = glm(sta ~ age*cpr, data = icu,
                     family = binomial(link = "logit"))

AIC(model.1)
```

```
## [1] 189.4655
```

```
AIC(model.interact)
```

```
## [1] 189.1097
```

3. Consider two additional variables: the level of consciousness at admission (`loc`) and whether infection was probable (`inf`).

a) Explore the relationship of each variable with survival to discharge.

i. Create a two-way table; explore the relationship of level of consciousness at admission and survival to discharge. Summarize your findings.

The proportion of individuals who survive to discharge of those who enter conscious is 158/185 = 0.85. No individuals who entered in a stupor survived to discharge. Of the individuals who entered in a coma, 2/10 = 0.20 survived to discharge.

ii. Create a two-way table; explore the relationship of probable infection at admission and survival to discharge. Summarize your findings.

The proportion of individuals who survived to discharge when infection was not probable is 100/116 = 0.86. The proportion of individuals who survived when infection was probable is 60/84 = 0.71.

```
#table: survival and consciousness
addmargins(table(icu$sta, icu$loc,
       dnn = c("Survival", "Consciousness")))
```

```
##          Consciousness
## Survival Nothing Stupor Coma Sum
##    Died       27      5    8  40
##    Lived     158      0    2 160
##    Sum       185      5   10 200
```

```
#table: survival and probable infection
addmargins(table(icu$sta, icu$inf,
       dnn = c("Survival", "Infection")))
```

```
##          Infection
## Survival  No Yes Sum
##    Died   16  24  40
##    Lived 100  60 160
##    Sum   116  84 200
```

b) Fit a logistic regression model to predict survival to discharge from age, prior CPR, probable infection, and level of consciousness.

i. Write the model equation.

$$\log\left[\frac{\hat{p}(\text{status = lived}|\text{age, cpr, inf, loc})}{1-\hat{p}(\text{status = lived}|\text{age, cpr, inf, loc})}\right] = 2.865 - 0.0277(age) - 1.0945(cpr_{yes}) - 0.705(inf_{yes}) - 18.381(loc_{stupor}) - 2.637(loc_{coma})$$

$$\log(\widehat{\text{odds of lived}} \mid \text{age, cpr, inf, loc}) = 2.865 - 0.0277(age) - 1.0945(cpr_{yes}) - 0.705(inf_{yes}) - 18.381(loc_{stupor}) - 2.637(loc_{coma})$$

ii. Interpret the slope coefficients.

The coefficient for age indicates that an increase of 1 year of age is associated with a decrease in estimated log odds of survival of 0.0277, when the values of the other variables do not change.

The coefficient for prior CPR indicates that the estimated log odds of survival for an individual who received CPR prior to admission are 1.094 lower than those for an individual who did not receive prior CPR, when the values of the other variables do not change.

The coefficient for probable infection indicates that the estimated log odds of survival for an individual for whom infection is probable are 0.705 lower than those

for an individual for whom infection is not probable, when the values of the other variables do not change.

The coefficient for stupor indicates that the estimated log odds of survival for an individual who entered the ICU in a stupor are 18.381 lower than those for an individual who entered conscious, when the values of the other variables do not change.

The coefficient for coma indicates that the estimated log odds of survival for an individual who entered the ICU in a coma are 2.637 lower than those for an individual who entered conscious, when the values of the other variables do not change.

iii. Examine the output of `summary(glm())`. Do you notice anything surprising or unexpected?

The estimate and standard error for stupor are extremely large (-18.381 and 1052.97, respectively). The *p*-value is insignificant, yet the earlier table showed that of the 5 individuals admitted in a stupor, none survived. It seems the model should indicate that the odds of survival for individuals who entered in a stupor are very different from the odds of survival for individuals who entered conscious.

The logistic regression model fails when there are empty or near empty cells for a particular variable; in this case, there were no observations of individuals who entered in a stupor and survived to discharge. The large standard error reflects the lack of precision in the estimate of the slope coefficient and leads to the very large *p*-value. This example illustrates the importance of examining data prior to fitting a model so as to be able to anticipate potential problems.

One possible solution is to collapse the levels of `loc` and create a binary version of the variable that records only whether or not a patient was conscious at admission.

```
#fit a model
model.2 = glm(sta ~ age + cpr + inf + loc, data = icu,
family = binomial(link = "logit"))
coef(model.2)
```

```
##  (Intercept)          age       cprYes       infYes    locStupor
##   3.86520553  -0.02774471  -1.09448036  -0.70460263 -18.38123507
##      locComa
##  -2.63687791
```

```
#model summary
summary(model.2)
```

```
##
## Call:
## glm(formula = sta ~ age + cpr + inf + loc, family = binomial(link = "logit"),
##     data = icu)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5973   0.2678   0.4799   0.5969   1.7291
```

```
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    3.86521    0.86223   4.483 7.37e-06 ***
## age           -0.02774    0.01257  -2.208  0.02725 *
## cprYes        -1.09448    0.75979  -1.441  0.14972
## infYes        -0.70460    0.42046  -1.676  0.09378 .
## locStupor    -18.38124 1052.97990  -0.017  0.98607
## locComa       -2.63688    0.86831  -3.037  0.00239 **
## ---
## 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: 152.20  on 194  degrees of freedom
## AIC: 164.2
##
## Number of Fisher Scoring iterations: 15
```

c) Run the code in the template to create a binary version of `loc` that records whether or not an individual was conscious upon being admitted to the ICU. Fit the model from part b) using `loc.binary`.

i. Interpret the slope coefficient of `loc.binary`.

The estimated log odds of survival for a patient who enters the ICU unconscious (either in stupor or coma) are 3.372 lower than a patient who enters the ICU conscious.

ii. Compare the odds of survival for a 50-year-old versus a 30-year-old, if infection was probable for both, and both received CPR prior to admission and entered the ICU conscious.

Since the values of the other variables in the model remain constant, it is possible to directly compare the odds via the slope coefficient for age. The slope coefficient represents the log of the odds ratio for a change of one year, when other variables are held constant. The odds ratio, comparing the younger individual to the older individual, is $\exp(0.0267 \times (50 - 30)) = 1.71$; the odds of survival to discharge for the 30-year-old are 1.7 times as large as the odds of survival for the 50-year-old.

iii. Suppose that a 70-year-old individual enters the ICU: CPR was administered prior to admission, they entered the ICU conscious, and infection was probable. Suppose that a 40-year-old enters the ICU: CPR was not administered prior to admission, they entered the ICU unconscious, and infection was not probable. Compare the odds of survival for these two individuals.

The estimated odds of survival for the older individual are 1.308; the estimated odds of survival for the younger individual are 2.506. The odds ratio, comparing the older individual to the younger individual, is 2.51; the odds of survival to discharge for the older individual are about 2.5 times as large as the odds of survival

to discharge for the younger individual.

```
#create the loc.binary variable
icu$loc.binary = icu$loc

#redefine the factor levels of loc.binary
levels(icu$loc.binary) = list("Conscious" = c("Nothing"),
                   "Unconscious" = c("Stupor", "Coma"))

#fit the model (part i.)
model.3 = glm(sta ~ age + cpr + inf + loc.binary, data = icu,
family = binomial(link = "logit"))
coef(model.3)
```

```
##           (Intercept)                   age                cprYes
##            3.79141720           -0.02673814           -0.94370358
##                infYes loc.binaryUnconscious
##           -0.70772144           -3.37217448
```

```
#50-year-old vs 30-year-old (part ii.)
exp(model.3$coef[2]*(50-30)); exp(-model.3$coef[2]*(50-30))
```

```
##       age
## 0.5858082
```

```
##      age
## 1.707043
```

```
#70-year-old vs 40-year-old (part iii.)
log.odds.70 = predict(model.3, newdata = data.frame(age = 70,
                                                    cpr = "Yes",
                                                    inf = "Yes",
                                                    loc.binary = "Conscious"))
log.odds.40 = predict(model.3, newdata = data.frame(age = 40,
                                                    cpr = "No",
                                                    inf = "No",
                                                    loc.binary = "Unconscious"))
exp(log.odds.70); exp(log.odds.40) #calculate odds
```

```
##        1
## 1.307768
```

```
##        1
## 0.521898
```

```
exp(log.odds.70)/exp(log.odds.40)  #calculate odds ratio
```

```
##        1
## 2.505793
```

d) Assess whether the model from part c) is a better parsimonious model than the model fit in Question 1.

The model from part c) has a lower AIC than the model from Question 1 (164.8 versus 189.5), so the model from part c) is a better parsimonious model. The addition of probable infection and level of consciousness (and removal of creatinine) result in a better predictive model; i.e., these additional variables are 'worth' adding.

```
AIC(model.1)
```

```
## [1] 189.4655
```

```
AIC(model.3)
```

```
## [1] 164.8071
```