

Unit 9: Logistic Regression

Statistics 102 Teaching Team

April 24, 2020

Introduction to logistic regression

Simple logistic regression

Multiple logistic regression

Introduction to logistic regression

LOGISTIC REGRESSION

Logistic regression generalizes methods for two-way tables, allowing for the joint association between a (binary) categorical response and several predictors to be studied. It also allows for numeric predictors.

Similar in intent to linear regression, but details are different. . .

- the response variable is categorical (specifically, binary)
- the model is not estimated via minimizing least squares
- the model coefficients have a different interpretation

SURVIVAL TO DISCHARGE IN THE ICU

Patients admitted to intensive care units (ICUs) are very ill, either from a serious medical event (e.g. respiratory failure from asthma) or from trauma (e.g, traffic accident).

Can patient features available at admission be used to estimate the probability of survival to hospital discharge?

The `icu` dataset in the `ap10re3` package is from a study conducted at Baystate Medical Center in Springfield, MA.

- The dataset contains information about patient characteristics at admission, such as heart rate, diagnosis, and kidney function.
- The variable `sta` is a factor variable with labels `Died` and `Lived`, corresponding to the levels 0 for death before discharge and 1 survival to discharge.
- Information on other variables measured are in the `lab`.

SURVIVAL AND CPR

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

#relevel survival
icu$sta = factor(icu$sta, levels = rev(levels(icu$sta)))

#two-way table of survival and cpr
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
```

The odds of survival for those who did not receive CPR are

$$154/33 = 4.67.$$

The probability of survival for those who did not receive CPR is

$$154/187 = 0.824.$$

ODDS AND PROBABILITIES

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

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

With some algebra, it is possible to show the following relationship:

$$\text{odds} = \frac{p}{1-p} \quad p = \frac{\text{odds}}{1 + \text{odds}}$$

From the previous example,

- $\text{odds} = 4.67, p = 0.824$
- $\frac{p}{1-p} = \frac{0.824}{1-0.824} = 4.67$
- $\frac{\text{odds}}{1 + \text{odds}} = \frac{4.67}{1 + 4.67} = 0.824$

ODDS AND PROBABILITIES...

Probability	Odds = $p/(1 - p)$	Odds
0	$0/1 = 0$	0
$1/100 = 0.01$	$1/99 = 0.0101$	1 : 99
$1/10 = 0.10$	$1/9 = 0.11$	1 : 9
$1/4$	$1/3$	1 : 3
$1/3$	$1/2$	1 : 2
$1/2$	$(\frac{1}{2})/(\frac{1}{2}) = 1$	1 : 1
$2/3$	$(2/3)/(1/3) = 2$	2 : 1
$3/4$	3	3 : 1
1	$1/0 \approx \infty$	∞

Simple logistic regression

THE MODEL FOR LOGISTIC REGRESSION

Suppose Y is a binary variable and X is a predictor variable.

- In the ICU example, let Y be survival to discharge and X represent prior CPR.
- Let $p(x) = P(Y = 1|X = x)$, where $Y = 1$ denotes survival.

The model for a single variable logistic regression is

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

WHY LOG(ODDS) IN REGRESSION MODELS?

Logistic regression is based on modeling the association between the probability p of the event of interest occurring and the values of the predictor variables.

Since a probability only takes values from 0 to 1, it is not ideal as a response.

- The odds, $p/(1 - p)$, ranges from 0 to ∞ .
- The natural log of the odds (log odds) ranges from $-\infty$ to ∞ .

LOGISTIC VERSUS LINEAR REGRESSION

Similarities:

- The right hand side of the model looks the same, but there is no residual error term in the logistic model.

Differences:

- Logistic regression is used to calculate predicted values of $\log(\text{odds})$, rather than the predicted mean.
- The function `glm` is used to estimate a logistic regression model, and requires an additional specification in the argument: `family = binomial(link = "logit")`

CPR STATUS AND SURVIVAL

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

```
##  
## Call:  glm(formula = sta ~ cpr, family = binomial(link = "logit"), data = icu)  
##  
## Coefficients:  
## (Intercept)      cprYes  
##      1.540      -1.695  
##  
## Degrees of Freedom: 199 Total (i.e. Null);  198 Residual  
## Null Deviance:      200.2  
## Residual Deviance: 192.2      AIC: 196.2
```

INTERPRETING THE OUTPUT

Recall that $p(x) = P(Y = 1|X = x) = p(\text{survival} = 1|\text{cpr})$.

The model equation:

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

The intercept represents the $\log(\widehat{\text{odds of survival}})$ for patients who did not receive CPR ($\text{cpr}_{\text{yes}} = 0$)

- estimated odds of survival = $\exp(1.540) = 4.66$
- estimated probability of survival = $\frac{\text{odds}}{1 + \text{odds}} = \frac{4.66}{1 + 4.66} = 0.823$.

INTERPRETING THE OUTPUT...

The slope coefficient of cpr_{yes} represents the change in $\log(\text{odds of survival})$, from the no previous CPR group to the previous CPR group.

The odds of survival in patients who previously received CPR:

- $\log \left[\frac{\hat{p}(\text{status} = \text{lived} | \text{cpr})}{1 - \hat{p}(\text{status} = \text{lived} | \text{cpr})} \right] = 1.540 - 1.695(1)$
- $\widehat{\text{odds of survival}} = \exp(1.540 - 1.695) = 0.856$

The slope coefficient is the log of the estimated *odds ratio* for survival, comparing those who received CPR to those who did not:

- $\widehat{OR} = \exp(-1.695) = 0.184$
- $\widehat{OR} = \frac{\text{odds}_{\text{cpr} = \text{yes}}}{\text{odds}_{\text{cpr} = \text{no}}} = \frac{0.856}{4.66} = 0.184$

INFERENCE FOR SIMPLE LOGISTIC REGRESSION

As with linear regression, the model slope captures information about association between a response and predictor.

- $H_0 : \beta_1 = 0$, the X and Y variables are not associated
- $H_A : \beta_1 \neq 0$, the X and Y variables are associated

These hypotheses can also be written in terms of the odds ratio.

WHAT DOES LOGISTIC REGRESSION ADD?

The χ^2 test does not directly show the direction of a significant association.

- Some information about direction from the residuals or differences between observed and expected values.

Logistic regression gives a numerical estimate of the size of an association.

A two-way table cannot be used with a numerical variable.

Multiple logistic regression

EXTENDING LOGISTIC REGRESSION TO MORE THAN ONE PREDICTOR

Suppose $p(x) = p(x_1, x_2, \dots, x_p) = P(Y = 1 | x_1, x_2, \dots, x_p)$.

With several predictors x_1, x_2, \dots, x_p the model is

$$\log \left[\frac{p(x)}{1 - p(x)} \right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

Each coefficient estimates the change in $\log(\text{odds})$ for a one unit change in that variable, if the other variables do not change.

SURVIVAL VERSUS CPR AND AGE

```
##
## Call:
## glm(formula = sta ~ cpr + cre + age, family = binomial(link = "logit"),
##      data = icu)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.3890   0.3446   0.5580   0.6784   1.2868
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.32901    0.74884   4.446 8.77e-06 ***
## cprYes        -1.69680    0.62145  -2.730  0.00633 **
## cre> 2.0      -1.13328    0.70191  -1.615  0.10641
## age           -0.02814    0.01125  -2.502  0.01235 *
## ---
## 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: 181.47  on 196  degrees of freedom
## AIC: 189.47
##
```

MODEL COMPARISON

The AIC (Akaike's Information Criterion) can be used to compare models.

- It is analogous to the adjusted R^2 for linear regression in that it also penalizes a model for having a larger number of predictors.
- A *lower* AIC is indicative of a more parsimonious model.

INFERENCE FOR MULTIPLE LOGISTIC REGRESSION

Typically, the hypotheses of interest are

- $H_0 : \beta_k = 0$, the variables X_k and Y are not associated
- $H_A : \beta_k \neq 0$, the variables X_k and Y are associated

These hypotheses can also be written in terms of the odds ratio.

SUMMARY OF LOGISTIC REGRESSION

Overall goals similar to linear regression. . .

- estimating the association between a response and several predictors
- assessing statistical significance of the association

However, in logistic regression, association is captured through *odds* and $\log(odds)$, instead of the mean of a response variable.

Logistic regression can be thought of as an extension of two-way tables. . .

- just as linear regression can be thought of as an extension of two-sample t -tests and ANOVA.