Logistic Regression

Robin Donatello

Last Updated 2017-11-10 08:37:48

Contents

References

- Open Intro Section 8.4
- Article: When can odds ratios mislead? http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1112884/
- Odds Ratios: http://www.ats.ucla.edu/stat/sas/faq/oratio.htm
- Marin Stats Lecture on OR and RR: https://www.youtube.com/watch?v=V_YNPQoAyCc

Introduction

- Logistic regression is a tool used to model a categorical outcome variable with two levels: Y = 1 if event, = 0 if no event.
- Instead of modeling the outcome directly $E(Y|X)$ as with linear regression, we model the probability of an event occurring: $P(Y = 1|X)$.

Uses of Logistic Regression

- Assess the impact selected covariates have on the probability of an outcome occurring.
- Predict the likelihood / chance / probability of an event occurring given a certain covariate pattern.

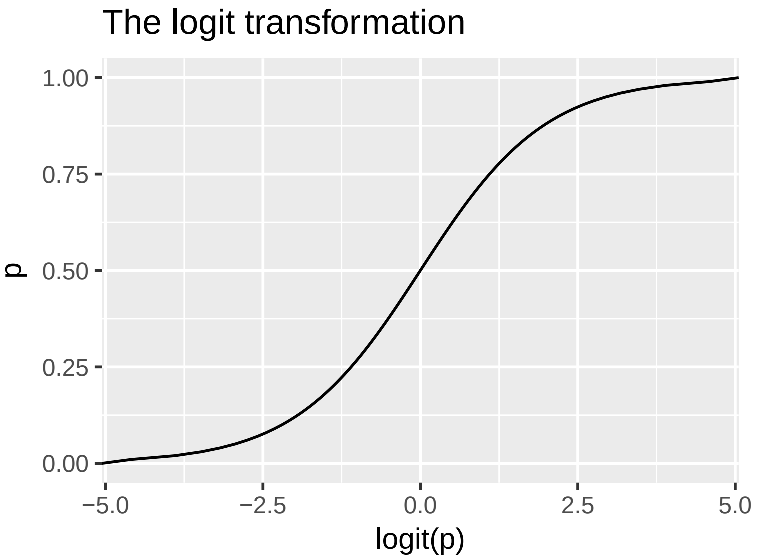
The Logistic Regression Model

Let $p_i = P(y_i = 1)$.

The logistic model relates the probability of an event based on a linear combination of X's.

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$$

Since the *odds* are defined as the probability an event occurs divided by the probability it does not occur: $(p/(1-p))$, the function $log\left(\frac{p_i}{1-p_i}\right)$ is also known as the *log odds*, or more commonly called the ***logit***.



This in essence takes a binary outcome 0/1 variable, turns it into a continuous probability (which only has a range from 0 to 1) Then the logit(p) has a continuous distribution ranging from $-\infty$ to ∞, which is the same form as a Multiple Linear Regression (continuous outcome modeled on a set of covariates)

Logistic Regression via GLM

A logistic regression model can be fit in R using the `glm()` function. GLM stands for *Generalized Linear Model*. GLM's can fit an entire *family* of distributions and can be thought of as $E(Y|X) = C(X)$ where C is a **link** function that relates Y to X.

- Linear regression: C = Identity function (no change)
- Logistic regression: C = logit function
- Poisson regression: C = log function

The outcome y is a 0/1 Bernoulli random variable. The sum of a vector of Bernoulli's $(\sum_{i=1}^{n} y_i)$ has a Binomial distribution. When we specify that `family = "binomial"` the `glm()` function auto-assigns "logit" link function. See `?family` for more information on this.

```r
glm(y ~ x1 + x2 + x3, data=DATA, family="binomial")
```

Example: Gender effects on Depression

Is gender associated with depression? Read in the `depression` data and recode sex to be an indicator of being male.

```r
depress <- read.delim("https://norcalbiostat.netlify.com/data/depress_081217.txt")
names(depress) <- tolower(names(depress)) # make all variable names lower case.
```

- Binary outcome variable: Symptoms of Depression (`cases`)

- Binary predictor variable: Gender (`sex`) as an indicator of being female

We fit the logistic regression model using a *generalized linear model*, specifying that the `family=binomial`. This tells `R` to use a *logit* link on the linear combination. SPSS users will choose the LOGISTIC function.

```
dep_sex_model <- glm(cases ~ sex, data=depress, family="binomial")
summary(dep_sex_model)
```

```
##
## Call:
## glm(formula = cases ~ sex, family = "binomial", data = depress)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7023  -0.7023  -0.4345  -0.4345   2.1941
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.3125     0.3315  -6.976 3.04e-12 ***
## sex           1.0386     0.3767   2.757  0.00583 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 268.12  on 293  degrees of freedom
## Residual deviance: 259.40  on 292  degrees of freedom
## AIC: 263.4
##
## Number of Fisher Scoring iterations: 5
```

We exponentiate the coefficients to back transform the β estimates into Odds Ratios

```
exp(coef(dep_sex_model))
```

```
## (Intercept)         sex
##   0.0990099   2.8251748
```

Females have 2.8 times the odds of showing signs of depression compared to males.

Confidence Intervals

The OR is **not** a linear function of the $x's$, but β is. This means that a CI for the OR is created by calculating a CI for β, and then exponentiating the endpoints. A 95% CI for the OR can be calculated as:

$$e^{\hat{\beta} \pm 1.96 SE_{\beta}}$$

```
exp(confint(dep_sex_model))
```

```
##                  2.5 %    97.5 %
## (Intercept) 0.04843014 0.1801265
## sex         1.39911056 6.2142384
```

Multiple Logistic Regression

Just like multiple linear regression, additional predictors are simply included in the model using a `+` symbol.

```
mvmodel <- glm(cases ~ age + income + sex, data=depress, family="binomial")
summary(mvmodel)
```

```
##
## Call:
## glm(formula = cases ~ age + income + sex, family = "binomial",
##     data = depress)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0249  -0.6524  -0.5050  -0.3179   2.5305
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.67646    0.57881  -1.169  0.24253
## age         -0.02096    0.00904  -2.318  0.02043 *
## income      -0.03656    0.01409  -2.595  0.00946 **
## sex          0.92945    0.38582   2.409  0.01600 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 268.12  on 293  degrees of freedom
## Residual deviance: 247.54  on 290  degrees of freedom
## AIC: 255.54
##
## Number of Fisher Scoring iterations: 5
```

- The sign of the β coefficients can be interpreted in the same manner as with linear regression.
- The odds of being depressed are less if the respondent has a higher income and is older, and higher if the respondent is female.

OR interpretation

- The OR provides a directly understandable statistic for the relationship between y and a specific x given all other x's in the model are fixed.
- For a continuous variable X with slope coefficient β, the quantity e^b is interpreted as the ratio of the odds for a person with value (X+1) relative to the odds for a person with value X.
- $exp(kb)$ is the incremental odds ratio corresponding to an increase of k units in the variable X, assuming that the values of all other X variables remain unchanged.

Binary variables Calculate the Odds Ratio of depression for women compared to men.

WRITE OUT THE MODEL

$$log(odds) = -0.676 - 0.02096 * age - .03656 * income + 0.92945 * gender$$

$$OR = \frac{Odds(Y=1|F)}{Odds(Y=1|M)}$$

Write out the equations for men and women separately.

$$= \frac{e^{-0.676-0.02096*age-.03656*income+0.92945(1)}}{e^{-0.676-0.02096*age-.03656*income+0.92945(0)}}$$

Applying rules of exponents to simplify.

$$= \frac{e^{-0.676}e^{-0.02096*age}e^{-.03656*income}e^{0.92945(1)}}{e^{-0.676}e^{-0.02096*age}e^{-.03656*income}e^{0.92945(0)}}$$

$$= \frac{e^{0.92945(1)}}{e^{0.92945(0)}}$$

$$= e^{0.92945}$$

```
exp(.92945)
```

```
## [1] 2.533116
```

```
exp(coef(mvmodel)[4])
```

```
##      sex
## 2.533112
```

The odds of a female being depressed are 2.53 times greater than the odds for Males after adjusting for the linear effects of age and income (p=.016).

Continuous variables

```
exp(coef(mvmodel))
```

```
## (Intercept)         age      income         sex
##   0.5084157   0.9792605   0.9640969   2.5331122
```

```
exp(confint(mvmodel))
```

```
##                 2.5 %    97.5 %
## (Intercept) 0.1585110 1.5491849
## age         0.9615593 0.9964037
## income      0.9357319 0.9891872
## sex         1.2293435 5.6586150
```

- The Adjusted odds ratio (AOR) for increase of 1 year of age is 0.98 (95%CI .96, 1.0)
- How about a 10 year increase in age? $e^{10*\beta_{age}} = e^{-.21} = .81$

```
exp(10*coef(mvmodel)[2])
```

```
##       age
## 0.8109285
```

with a confidence interval of

```
round(exp(10*confint(mvmodel)[2,]),3)
```

```
##  2.5 % 97.5 %
##  0.676  0.965
```

Controlling for gender and income, an individual has 0.81 (95% CI 0.68, 0.97) times the odds of being depressed compared to someone who is 10 years younger than them.

CAUTION

Consider a hypothetical example where the probability of death is .4 for males and .6 for females.

The odds of death for males is `.4/(1-.4)` = 0.67. The odds of death for females is `.6/(1-.6)` = 1.5.

The Odds Ratio of death for females compared to males is `1.5/.66` = 2.27.

- If you were to say that females were 2.3 times as likely to die compare to males, you wouldn't necessarily translate that to a 40% vs 60% chance.

Modeling the probability of an event.

Sometimes it is more explanatory, or easier to explain the results of a logistic model by providing a probability interpretation instead of an Odds Ratio. For a **specific** set of covariate values, you can calculate the probability of an event for one group, and compare it to the probability of an event for another group.

Back solving the logistic model for $p_i = e^{\beta X}/(1+e^{\beta X})$:

$$p_i = \frac{e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\ldots+\beta_p x_{pi}}}{1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\ldots+\beta_p x_{pi}}}$$

For the above model of depression on age, income and gender the probability of depression is:

$$P(depressed) = \frac{e^{-0.676-0.02096*age-.03656*income+0.92945*gender}}{1+e^{-0.676-0.02096*age-.03656*income+0.92945*gender}}$$

Let's compare the probability of being depressed for males and females separately, while holding age and income constant at their average value.

```
depress %>% summarize(age=mean(age), income=mean(income))
```

```
##        age   income
## 1 44.41497 20.57483
```

Plug the coefficient estimates and the values of the variables into the equation and calculate.

$$P(depressed|Female) = \frac{e^{-0.676-0.02096(44.4)-.03656(20.6)+0.92945(1)}}{1+e^{-0.676-0.02096(44.4)-.03656(20.6)+0.92945(1)}}$$

```
XB.f <- -0.676 - 0.02096*(44.4) - .03656*(20.6) + 0.92945
exp(XB.f) / (1+exp(XB.f))
```

```
## [1] 0.1930504
```

$$P(depressed|Male) = \frac{e^{-0.676-0.02096(44.4)-.03656(20.6)+0.92945(0)}}{1+e^{-0.676-0.02096(44.4)-.03656(20.6)+0.92945(0)}}$$

```
XB.m <- -0.676 - 0.02096*(44.4) - .03656*(20.6)
exp(XB.m) / (1+exp(XB.m))
```

```
## [1] 0.08629312
```

The probability for a 44.4 year old female who makes $20.6k annual income has a 0.19 probability of being depressed. The probability of depression for a male of equal age and income is 0.86.