Logistic Regression

Robin Donatello

Last Updated 2017-11-07 09:58:25

Contents

References

Assigned reading

- Open Intro Section 8.4
- PMA5 Ch 12 (selected)
- Article: When can odds ratios mislead? http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1112884/

Additional references

- 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 (PMA5 12.10)

- 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 (PMA5 12.4)

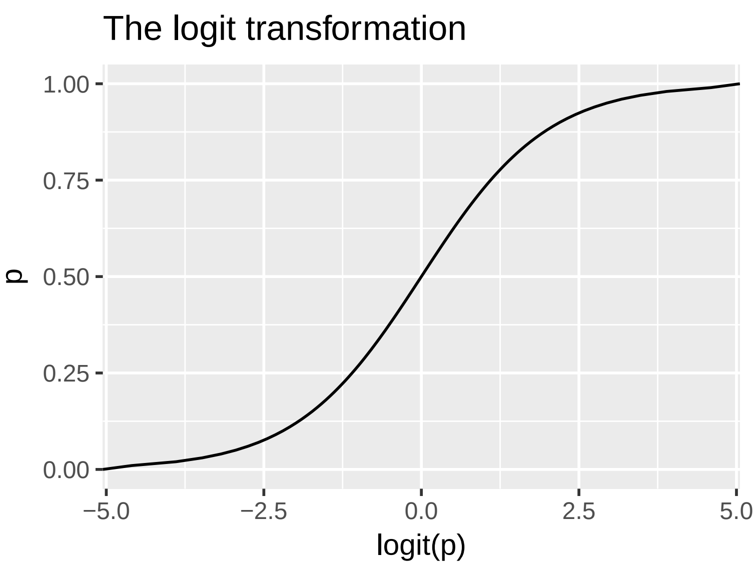
Let $p_i = P(y_i = 1)$.

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***.

```
p <- seq(0, 1, by=.01)
logit.p <- log(p/(1-p))
qplot(logit.p, p, geom="line", xlab = "logit(p)", main="The logit transformation")
```



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)

Modeling the probability of an event.

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

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

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.

```
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.

```
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

```r
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:

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

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

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

Multiple Logistic Regression (PMA5 12.5, 12.6)

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

```r
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

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

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

```r
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? $e^{10*\beta_{age}} = e^{-.21} = .81$

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

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

with a confidence interval of

```r
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.

Interaction terms (PMA5 12.7)

The inclusion of an interaction is necessary if the effect of an independent variable depends on the level of another independent variable.

Example: The relationship between income, employment status and depression.

Here I create the binary indicators of lowincome and underemployed as described in the textbook. In each case I ensure that missing data is retained.

```r
depress$lowincome <- ifelse(depress$income < 10, 1, 0)
depress$lowincome <- ifelse(is.na(depress$income), NA, depress$lowincome)

depress$underemployed <- ifelse(depress$employ %in% c(2,3), 1, 0 )
depress$underemployed <- ifelse(is.na(depress$employ) | depress$employ==7, NA, depress$underemployed)
table(depress$underemployed, depress$employ, useNA="always")
```

```
##
##         FT Houseperson In School Other  PT Retired Unemp <NA>
##   0    167          27         2     4  42      38    14    0
##   <NA>   0           0         0     0   0       0     0    0
```

The **Main Effects** model assumes that the effect of income on depression is independent of employment status, and the effect of employment status on depression is independent of income.

```r
me_model <- glm(cases ~ lowincome + underemployed, data=depress, family="binomial")
summary(me_model)
```

```
##
## Call:
## glm(formula = cases ~ lowincome + underemployed, family = "binomial",
##     data = depress)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6431  -0.6312  -0.5957  -0.5957   1.9062
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.6393     0.1902  -8.618   <2e-16 ***
## lowincome       0.1684     0.3294   0.511    0.609
## underemployed       NA         NA      NA       NA
## ---
## 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: 267.87  on 292  degrees of freedom
## AIC: 271.87
##
## Number of Fisher Scoring iterations: 4
```

To formally test whether an interaction term is necessary, we add the interaction term into the model and assess whether the coefficient for the interaction term is significantly different from zero.

```r
summary(glm(cases ~ lowincome + underemployed + lowincome*underemployed, data=depress, family="binomial
```

```
##
## Call:
## glm(formula = cases ~ lowincome + underemployed + lowincome *
##     underemployed, family = "binomial", data = depress)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6431  -0.6312  -0.5957  -0.5957   1.9062
##
## Coefficients: (2 not defined because of singularities)
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.6393     0.1902  -8.618   <2e-16 ***
## lowincome                 0.1684     0.3294   0.511    0.609
## underemployed                 NA         NA      NA       NA
## lowincome:underemployed       NA         NA      NA       NA
## ---
## 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: 267.87  on 292  degrees of freedom
```

```
## AIC: 271.87
##
## Number of Fisher Scoring iterations: 4
```

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.

Probability Interpretation

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.

Relative Risk and 2x2 tables

An Odds Ratio is one measure of association between two binary variables, the **Relative Risk Ratio** is another measure. Both can be calculated on a 2x2 contingency table. Note that the OR that is generated from a GLM with only a binary categorical predictor will be identical to the one calculated on the 2x2 table. GLM's have more flexibility for further model building, which is why it is introduced first.

Consider a 2x2 contingency table similar to the following:

```
nnnn <- matrix(c("$n_{11}$", "$n_{12}$", "$n_{1.}$",
                 "$n_{21}$", "$n_{22}$", "$n_{2.}$",
                 "$n_{.1}$", "$n_{.2}$", "$n_{..}$"), nrow=3, byrow=TRUE,
        dimnames = list(c("Exposed", "Not-Exposed", "Total"), c("Diseased", "Not-Diseased", "Total")))
#print(xtable(nnnn, align='cccc'), type='html')
```

```
pander(nnnn)
```

	Diseased	Not-Diseased	Total
Exposed	n_{11}	n_{12}	$n_{1.}$
Not-Exposed	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$

Sometimes the cell contents are abbreviated as:

```
abcd <- matrix(c("a", "b", "c", "d"), nrow=2,
        dimnames = list(c("Exposed", "Not-Exposed"), c("Diseased", "Not-Diseased")))
#print(xtable(abcd, align='ccc'), type="html")
```

```
pander(abcd)
```

	Diseased	Not-Diseased
Exposed	a	c
Not-Exposed	b	d

Relative Risk

The **Relative Risk (RR)** or **Risk Ratio** is the ratio of the probability of an event occurring in an exposed group compared to the probability of an event occurring in a non-exposed group.

- Asymptotically approaches the OR for small probabilities.
- Often used in cohort studies and randomized control trials.

Consider sample proportions Diseases within Exposed and Non-exposed groups.

$$\hat{\pi}_1 = \frac{n_{11}}{n_{1.}} \qquad \text{and} \qquad \hat{\pi}_2 = \frac{n_{21}}{n_{2.}}$$

The Relative Risk is calculated as

$$RR = \frac{\hat{\pi}_1}{\hat{\pi}_2} \qquad \text{or} \qquad \frac{a/(a+b)}{c/(c+d)}$$

with variance

$$V = \frac{1-\hat{\pi}_1}{n_{11}} + \frac{1-\hat{\pi}_2}{n_{21}}$$

Odds Ratio

The **Odds Ratio (OR)** is a way to quantify how strongly the presence or absence of a characteristic affects the presence or absence of a second characteristic.

- Often used in case-control studies
- The main interpretable estimate generated from logistic regression

The **Odds of an event** is the probability it occurs divided by the probability it does not occur.

$$odds_1 = \frac{n_{11}/n_{1.}}{n_{12}/n_{1.}} = \frac{n_{11}}{n_{12}}$$
$$odds_2 = \frac{n_{21}/n_{2.}}{n_{22}/n_{2.}} = \frac{n_{21}}{n_{22}}$$

The **Odds Ratio** for group 1 compared to group 2 is the ratio of the two odds written above:

$$OR = \frac{odds_1}{odds_2} = \frac{n_{11}n_{22}}{n_{12}n_{21}} \qquad \text{or} \qquad \frac{ad}{bc}$$

with variance $V = n_{11}^{-1} + n_{12}^{-1} + n_{21}^{-1} + n_{22}^{-1}$.

Confidence Intervals

Neither the Risk Ratio nor the Odds Ratio are linear functions, so a 95% CI for the population estimates are not your typical $\hat{\theta} \pm 1.96\sqrt{Var(\hat{\theta})}$.

Instead they are calculated as the point estimate $\hat{\theta}$ times e raised to the ± 1.96 times the standard deviation of the estimate.

$$(\hat{\theta}e^{-1.96*\sqrt{V}}, \hat{\theta}e^{1.96*\sqrt{V}})$$

Example: Are females more likely to show signs of depression than males?

```
table(depress$sex, depress$cases, dnn=c("Female", "Signs of Depression"))
```

```
##       Signs of Depression
## Female   0   1
##      0 101  10
##      1 143  40
```

Note that both the columns and rows are swapped when compared to the a/b/c/d format. For ease of interpretation I will recreate the table manually.

```
tab_sn <- matrix(c(101, 10, 143, 40), nrow=2, byrow=T,
                 dimnames = list(c("Male", "Female"), c("No signs", "Signs")))
tab_sn
```

```
##        No signs Signs
## Male        101    10
## Female      143    40
```

Now I use the `epi.2by2` function contained in the `epiR` package to calculate the Odds Ratio, the Risk Ratio, and their respective confidence intervals.

```r
library(epiR)
epi.2by2(tab_sn)
```

```
##              Outcome +    Outcome -      Total        Inc risk *
## Exposed +          101           10        111              91.0
## Exposed -          143           40        183              78.1
## Total              244           50        294              83.0
##                   Odds
## Exposed +        10.10
## Exposed -         3.58
## Total             4.88
##
## Point estimates and 95 % CIs:
## -------------------------------------------------------------------
## Inc risk ratio                               1.16 (1.06, 1.28)
## Odds ratio                                   2.83 (1.35, 5.91)
## Attrib risk *                                12.85 (4.83, 20.86)
## Attrib risk in population *                  4.85 (-2.52, 12.22)
## Attrib fraction in exposed (%)               14.12 (5.43, 22.02)
## Attrib fraction in population (%)            5.85 (1.92, 9.61)
## -------------------------------------------------------------------
##  X2 test statistic: 8.082 p-value: 0.004
##  Wald confidence limits
##  * Outcomes per 100 population units
```

- Females are 1.16 (1.06, 1.28) times as likely as men to show signs of depression.
- Females have 2.83 (1.35, 5.91) times the odds of showing signs of depression compared to males.

Both intervals are greater than 1, therefore the event (depressive signs) is statistically more likely to occur in the exposed group (female) than in the control (males) (p=.004).

- Mathematical reference for the Wald test Statistic http://www.statlect.com/Wald_test.htm