Lec 08: Logistic Regression

MATH 456 - Spring 2016

Navbar: [Home] [Schedule] [Data] [Week 11 Overview] [HW Info] [Google Group]

Assigned Reading and additional references

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

– Additional References

- http://www.ats.ucla.edu/stat/sas/faq/oratio.htm

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 (Afifi 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 (Afifi 12.4)

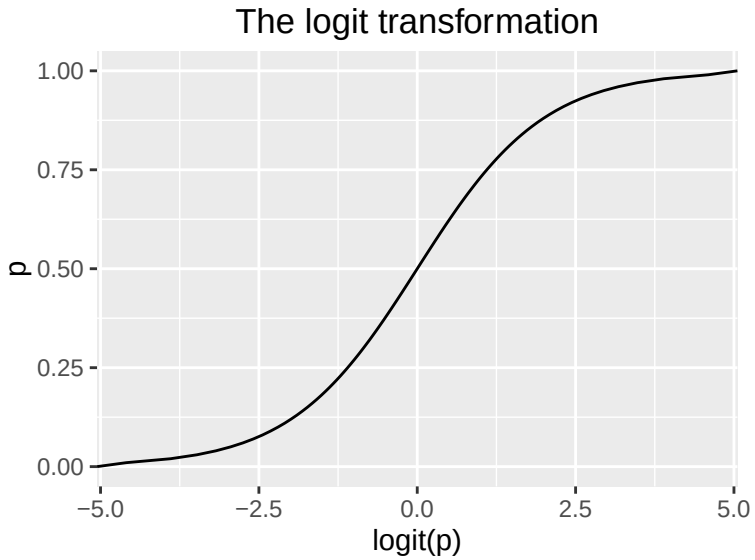
Let $p_i = P(y_i = 1)$.

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 in R

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 a "logit" link function. See `?family` for more information on this.

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

Example: Gender effects on Depression

Read in the depression data and recode sex to be an indicator of being male.

```
depress <- read.delim("C:/GitHub/MATH456/data/depress_030816.txt")
names(depress) <- tolower(names(depress)) # make all variable names lower case.
depress$sex <- depress$sex -1 # Refactor to match book table.
```

Using a two-way table.

Examine the two-way table of gender by depression and calculate the Odds Ratio for depression and gender.

```
table(depress$sex, depress$cases, dnn = c("Gender", "Depression"))
```

```
##        Depression
## Gender   0   1
##      0 101  10
##      1 143  40
```

Recall that the `epi.2by2` function in the `epiR` package required the (1,1) cell to be in the upper left corner. That is not default table orientation for R. So here is a helper function `rotate()` that I found on StackOverflow that will rotate the matrix to the proper orientation.

```
rotate <- function(x) t(apply(t(apply(x, 2, rev)), 2, rev))
```

Create the table object, rotate it (to confirm it works), and call `epi.2by2` to calculate the OR and corresponding CI.

```
library(epiR)
dep_sex_xtab <- table(depress$sex, depress$cases)
rotate(dep_sex_xtab)
```

```
##
##      1   0
##   1 40 143
##   0 10 101
```

```
epi.2by2(rotate(dep_sex_xtab))
```

```
##              Outcome +    Outcome -      Total        Inc risk *
## Exposed +           40          143        183             21.86
## Exposed -           10          101        111              9.01
## Total               50          244        294             17.01
##                  Odds
## Exposed +       0.280
## Exposed -       0.099
## Total           0.205
##
## Point estimates and 95 % CIs:
## -------------------------------------------------------------------
## Inc risk ratio                                 2.43 (1.26, 4.65)
## Odds ratio                                     2.83 (1.35, 5.91)
## Attrib risk *                                  12.85 (4.83, 20.86)
## Attrib risk in population *                    8.00 (1.16, 14.84)
```

```
## Attrib fraction in exposed (%)               58.78 (20.92, 78.52)
## Attrib fraction in population (%)            47.03 (10.63, 68.60)
## ---------------------------------------------------------------------
##  X2 test statistic: 8.082 p-value: 0.004
##  Wald confidence limits
##  * Outcomes per 100 population units
```

Females have 2.83 times the odds of being depressed compared to males (95% CI 1.35, 5.91).

Using Logistic Regression

We will come to the same conclusion by running a logistic regression model,

```
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
```

and exponentiating the coefficients.

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

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

The Odds Ratio for depression among Females compared to males is 2.83.

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:

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

In R this looks like:

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

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

Multiple Logistic Regression (Afifi 12.5, 12.6)

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 respondant 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.

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 probabilty of depression for a male of equal age and income is 0.86.

Logistic models with interaction terms (Afifi 12.7)

This section follows the book very closely so minimal notes are presented

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

Example: The relationsihp 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.

```
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")
```

```
##
##          1   2   3   4   5   6   7 <NA>
##   0    167   0   0  38  27   2   0    0
##   1      0  42  14   0   0   0   0    0
##   <NA>   0   0   0   0   0   0   4    0
```

The **Main Effects** model assumes that the effect of income on depression is indpendent 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
## -0.9227  -0.5894  -0.5195  -0.5195   2.0345
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.9345     0.2259  -8.563  < 2e-16 ***
## lowincome       0.2723     0.3377   0.806  0.42004
## underemployed   1.0285     0.3487   2.949  0.00318 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 263.46  on 289  degrees of freedom
## Residual deviance: 254.86  on 287  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 260.86
##
## 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
## -1.3537  -0.5701  -0.5701  -0.4783   2.1093
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.7346     0.2214  -7.835  4.7e-15 ***
## lowincome                -0.3756     0.4349  -0.864  0.38780
## underemployed             0.3175     0.4520   0.702  0.48238
## lowincome:underemployed   2.1981     0.7888   2.787  0.00533 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 263.46  on 289  degrees of freedom
## Residual deviance: 246.63  on 286  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 254.63
##
## Number of Fisher Scoring iterations: 4
```

Confouding and Effect Modification

- **Confounder**: A covariate that is associated with both the outcome and the risk factor.
- **Effect Modifier**: A covariate that modifies the effect a second covariate has on the outcome.

Refining and evaluating logistic regression

Going further

When your outcome has more than one level and you want to build a regression model to assess the impact a specific variable (or set of variables) has on the levels of this outcome variable, you would need to turn to more generalized linear models such as:

- Multinomial distribution for a nominal outcome
  - http://www.ats.ucla.edu/stat/r/dae/mlogit.htm
- Ordinal logistic regression
- http://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/

[top]

On Your Own

On Your Own

1. What does an Odds Ratio of 1 signify? What if the OR < 1? What about when OR > 1? You can use a pair of example variables such as X=gender and y=death if it helps you explain.
2. Afifi 12.9 (a-c)
3. Afifi 12.14
4. Afifi 12.15
5. Afifi 12.16
6. Afifi 12.17
7. Afifi 12.18
8. Afifi 12.23