Lec 08: Logistic Regression

MATH 456 - Spring 2016

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

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 probabilty of an event occuring: $P(Y = 1|X)$.

Uses of Logistic Regression

- Assess the effect covariates have on the probability of an outcome occuring.
  - Interpreting Coefficients
- Predict the likelihood / chance / probability of an event occuring given a certain covariate pattern.

We will start by learning how to do the first.

Assigned Reading and additional references

- Open Intro Section 8.4
- Afifi Ch 12

Spam vs Ham

Let's revisit the `email` data set where the `spam` variable is our binary outcome variable.

```
email <- read.delim("C:/GitHub/MATH456/data/email.txt", header=TRUE, sep="\t")
```

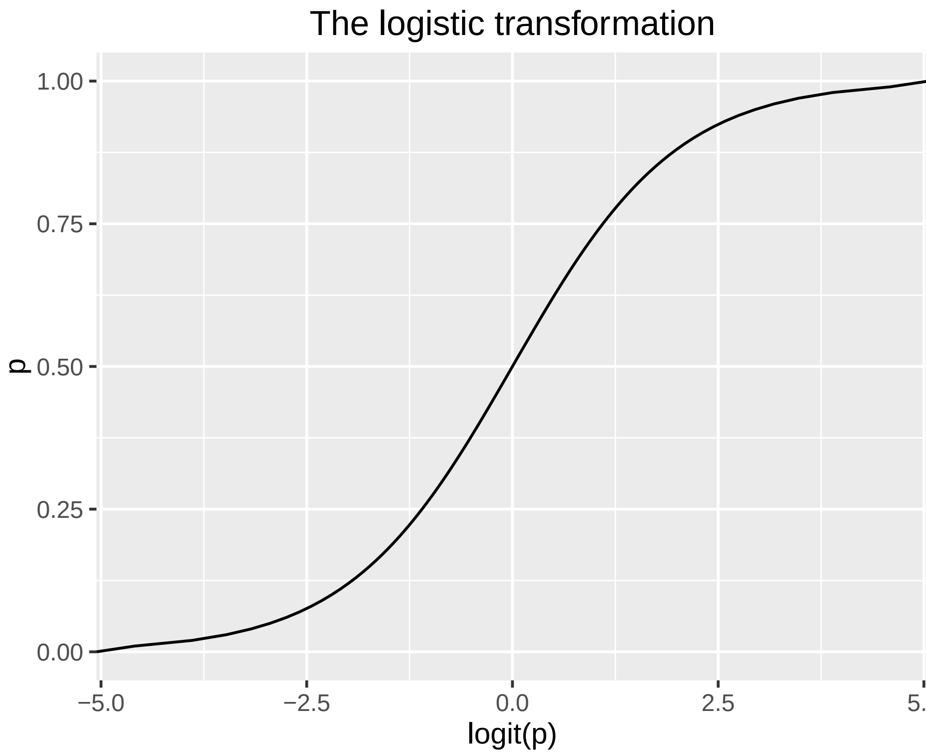
The Logistic Regression Model

Let $p_i = P(y_i = 1)$. Then the logistic model relating the probability of an event based on a set of covariates X is

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

The transformation $log\left(\frac{p_i}{1-p_i}\right)$ is called a ***logit*** transformation.

```
p <- seq(0, 1, by=.01)
logit.p <- log(p/(1-p))
qplot(logit.p, p, geom="line", xlab = "logit(p)", main="The logistic 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), and then turns the logit(p) now has a continuous distribution ranging from $-\infty$ to ∞. This now has 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 $p_i = e^{\beta X}/(1+e^{\beta X})$:

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

Example: Modeling spam based off a single predictor. (Open Intro 8.18)

Here we create a spam filter with a single predictor: `to multiple`. This variable indicates whether more than one email address was listed in the *To* field of the email. We perform logistic regression in R by calling the `glm()` function, where GLM stands for Generalized Linear Models, and specifying that the `family="binomial"`. This is because the sum of the outcome y is a binomial random variable (Sum of Bernoulli RV's).

```
summary(glm(spam ~ to_multiple, data=email, family="binomial"))
```

```
##
## Call:
## glm(formula = spam ~ to_multiple, family = "binomial", data = email)
##
```

```
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -0.477  -0.477  -0.477  -0.477   2.809
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.11609    0.05618 -37.665  < 2e-16 ***
## to_multiple -1.80918    0.29685  -6.095  1.1e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2437.2  on 3920  degrees of freedom
## Residual deviance: 2372.0  on 3919  degrees of freedom
## AIC: 2376
##
## Number of Fisher Scoring iterations: 6
```

The regression equation then is:

$$log\left(\frac{p_i}{1-p_i}\right) = -2.12 - 1.8to\_multiple$$

If an email is randomly selected and it has just one address in the *To* field, what is the probability it is spam?

When `to_multiple = 0` then $log(p/(1-p)) = -2.12$.

Solving for $\hat{p} = e^{-2.12}/(1+e^{-2.12}) = .11$.

```r
exp(-2.12)/(1+exp(-2.12))
```

```
## [1] 0.1071681
```

What if more than one address is listed in the To field?

Odds Ratios Revisited

Recall the section on odds ratios from [Lec07] lecture notes. Table 12.1: Classification of individuals by depression level and gender.

```r
depress <- read.delim("C:/GitHub/MATH456/data/depress_030816.txt")
depress$SEX <- depress$SEX -1 # Refactor to match book table.
table(depress$SEX, depress$CASES, dnn = c("Gender", "Depression"))
```

```
  Depression
```

Gender 0 1 0 101 10 1 143 40 The odds are defined as P / (1-P).

The Odds Ratio (OR) = Odds(Depressed | Male) / Odds(Depressed | Female)

The logistic function is P(Depressed | X)

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