# Logistic_Regression_Binomial

October 27, 2019

## 1 Logistic Regression for a Binomially-distributed R.V.

### 1.1 Goal

Our goal is to learn how to apply logistic regression to count data where the number of "1"s is restricted. Like logistic regression for a Bernoulli distributed variable, we'll explore Binomially-distributed random variables and how to model the probability of 'success' conditional on covariates.

The logistic function will map an inner product of coefficients and covariates to the interval between 0 and 1.

We will also find the likelihood for this model and show how it generalizes our model of logistic regression for a Bernoulli distributed random variable.

### 1.2 Binomial distributed random variable

A random variable $Y$ is Binomially distributed if, given a set of independent **trials** the probability of observing $s$ successes (sometimes denoted 1s) is

$$p(Y = s; N, p) = \binom{N}{s} p^s (1-p)^{N-s}$$

The expression

$$p^s$$

is the probability of observing a string of $s$ 1s with probability $p$.

The expression

$$(1-p)^{N-s}$$

is the probability of observing a string of $N - s$ 0s with probability $1 - p$.

Then, the probability of observing a string of $s$ 1s and $N - s$ 0s equals

$$p^s(1-p)^{N-s}$$

But the binomial distribution does not consider the order of 1s and 0s. Instead, the binomial distribution computes the probability of observing $s$ 1s among $N$ trials in any order.

The Binomial operator counts the number of ways $s$ 1s can occur among $N$ trials–$\binom{N}{s}$. The Binomial operator equals

$$\binom{N}{s} = \frac{N!}{(N-s)!} \times \frac{1}{s!}$$

The first term counts the number of ways $s$ 1s can occupy any of $N$ different positions. For example, given $N = 10$ different possibilites and $s = 4$ 1s, the first 1 can occupy the 4th position, secon 1 can occupy the 1st position, third 1 can occupy the 8th position, and fourth 1 can occupy the 7th position. In general, the first 1 has N different options. After the first 1 occupies one of the N different spaces, the second one now has only N-1 spaces to choose from. Finally, the s$^{th}$ 1 has $N - s$ spaces to choose from. We can write this as

$$N \times N-1 \times \cdots \times N-(s-1)$$

or

$$\frac{N!}{(N-s)!}$$

where ! is the factorial. The factorial of an integer $N! = N \times N-1 \times \cdots \times 1$.

But the above ratio of factorial counts the number of positions each distinct 1 can occupy. We are not interested in the distinction between the first, second, or fourth 1. For every way to occupy spaces above, we could have chosen any of the $s$ 1s.

If we denote $C$ to be the number of indistinct ways to occupy spaces with 1s, then

$$C(N,s) \times s! = \frac{N!}{(N-s)!}$$

We could have chosen any one of the $s$ 1s to occupy a position. Then we can choose any of the $s-1$ remaining 1s to occupy the next position.

There are $s!$ ways to places the ordered 1s into the same sequence of spaces. Then the number of ways to place 1s into these spaces without caring about the ordering of 1s equals

$$C(N,s) = \binom{N}{s} = \frac{N!}{(N-s)!s!}$$

The expectation of a Binomial r.v. is

$$
\begin{aligned}
E(Y) &= \sum_{i=1}^{N} i \times p(Y = i; N, p) && (1)\\
&= \sum_{i=1}^{N} i \times \binom{N}{i} p^i (1-p)^{N-i} && (2)\\
&= \sum_{i=1}^{N} i \times \frac{N!}{(N-i)!i!} p^i (1-p)^{N-i} && (3)\\
&= \sum_{i=1}^{N} \frac{N!}{(N-i)!(i-1)!} p^i (1-p)^{N-i} && (4)\\
&= \sum_{i=1}^{N} N \frac{(N-1)!}{(N-i)!(i-1)!} p^i (1-p)^{N-i} && (5)\\
&= \sum_{i=1}^{N} Np \frac{(N-1)!}{(N-i)!(i-1)!} p^{i-1} (1-p)^{N-i} && (6)\\
&= Np \sum_{i=1}^{N} \frac{(N-1)!}{(N-i)!(i-1)!} p^{i-1} (1-p)^{N-i} && (7)\\
&= Np \times 1 \text{ Why?} && (8)\\
&= Np && (9)
\end{aligned}
$$

Similar manipulations show that the variance of a Binomial random variable equals

$$Var(Y) = Np(1-p)$$

We can already see a similarity between Bernoulli and Binomial random variables. A binomial distribution with a single trial is the same as a Bernoulli random variable.

If Y is a bionomial r.v.

$$
\begin{aligned}
p(Y|N = 1, p) &= \binom{N}{s} p^s (1-p)^{N-s} && (10)\\
&= \binom{1}{s} p^s (1-p)^{1-s} && (11)\\
&= \frac{1!}{(1-s)!s!} p^s (1-p)^{1-s} && (12)\\
& && (13)
\end{aligned}
$$

If $s = 0$ then $\frac{1!}{(1)!0!} = 1$ and if $s = 1$ then $\frac{1!}{(1-1)!0!} = 1$ so

$$
\begin{aligned}
p(Y|N = 1, p) = p(Y|p) &= p^s (1-p)^{1-s} && (14)\\
& && (15)
\end{aligned}
$$

$Y$ equals 1 with probability $p^1(1-p)^{1-1} = p$ and zero with probability $p^0(1-p)^{1-0} = (1-p)$. These probabilities correspond exactly to a Bernoulli distributed random variable.

A Bernoulli distributed random variable is equivalent to a Binomial distributed random variable with a single trial.

### 1.3 Binomial data and covariates

Throughout, our goal in regression is to estimate the conditional probability of one random variable given a set of fixed covariates, $p(Y|X)$.

Consider as our data a set of Binomial random variables. Every data point we receive has the number of trials, the number of 1s, and a corresponding vector of $x$ data

| N | Number of 1s | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|---|
| 10 | 0 | 2.2 | 1 | 0.3 |
| 4 | 4 | 3 | 0 | 0.001 |
| 6 | 1 | 1/4 | 1 | 0.5 |
| 21 | 4 | 4 | 1 | 0.6 |
| 14 | 10 | 5.6 | 0 | 0.9 |

Covariates will only influence the probability $p$ shared by every observation. We treat the number of trials $N$ as a fixed covariate and do not need to estimate a general $N$. The number of trials will be assumed given.

### 1.4 Likelihood

The probability of a set of $N$ independent Binomial random variables is

$$P(y_1, y_2, y_3, \cdots, y_n|p) = \prod_{i=1}^{N} \binom{N_i}{s_i} p^{s_i}(1-p)^{N_i - s_i}$$

where $N_i$ is the number of trials for the i$^{th}$ observation and $s_i$ the number of 1s for the i$^{th}$ observation.

We take the same approach to estimating $p$ as we did for a sequence of Bernoulli random variables. Define the logistic function on $p$

$$f(p|\beta, x) = \frac{e^{\beta' x}}{1 + e^{\beta' x}}$$

and substitute this function into the above probability model.

$$P(y_1, y_2, y_3, \cdots, y_n|\beta, x) = \prod_{i=1}^{N} \binom{N_i}{s_i} \left(\frac{e^{\beta' x}}{1 + e^{\beta' x}}\right)^{s_i} \left(\frac{1}{1 + e^{\beta' x}}\right)^{N_i - s_i}$$

The likelihood considers the $y$ and $x$ data fixed and treats the above model as a function of $\beta$.

$$\ell(\beta) = P(\beta|y,x) = \prod_{i=1}^{N} \binom{N_i}{s_i} \left(\frac{e^{\beta' x}}{1+e^{\beta' x}}\right)^{s_i} \left(\frac{1}{1+e^{\beta' x}}\right)^{N_i - s_i}$$

The above likelihood generalizes our previous Bernoulli-distributed random variables. Assume every Binomial variable in the above likelihood had only one trial. The likelihood reduces to

$$\ell(\beta) = \prod_{i=1}^{N} \left(\frac{e^{\beta' x}}{1+e^{\beta' x}}\right)^{s_i} \left(\frac{1}{1+e^{\beta' x}}\right)^{1 - s_i}$$

where $s_i$ equals either 0 or 1. This is the exact likelihood we derived for a sequence of Bernoulli-distributed random variables. The Binomial model generalizes logistic regression for a Bernoulli-distributed random variable.

### 1.5 Example Data

The Example comes from the titanic data set. This data set counted the total number of passengers and number of survivors.

[28]:
```
require(plyr)

d = read.csv("https://vincentarelbundock.github.io/Rdatasets/csv/COUNT/titanic.
 ↪csv")
d$survive = ifelse(d$survive=='yes',1,0)

print(head(d))
```

```
  X     class    age sex survived survive
1 1 1st class adults man      yes       1
2 2 1st class adults man      yes       1
3 3 1st class adults man      yes       1
4 4 1st class adults man      yes       1
5 5 1st class adults man      yes       1
6 6 1st class adults man      yes       1
```

We can run a logistic regression to find an association between sex and survival. Our first type of regression considers survival a 0-1 binary variable.

[29]:
```
bernoulliLogR = glm(survive~sex , data = d, family = binomial)
print(summary(bernoulliLogR))
```

```
Call:
glm(formula = survive ~ sex, family = binomial, data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
```

```
-1.6065  -0.6706  -0.6706   0.8023   1.7903

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.37769    0.08459  -16.29   <2e-16 ***
sexwomen     2.34625    0.13554   17.31   <2e-16 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1746.8  on 1315  degrees of freedom
Residual deviance: 1399.0  on 1314  degrees of freedom
AIC: 1403

Number of Fisher Scoring iterations: 4
```

We see that women had a higher chance of surviving compared to men (why?). The "Women and children" first mantra could be the cause.

The same data could be considered to come from a binomial random variable. The Number of trial would be the number of women on board and successes the number who survived. A second observation would consider the number of men on board and successes the number who survived.

[30]:
```
dGrouped = ddply( d, .(sex)
                  , function(x){
                      nRow = nrow(x)
                      numSurvived = sum(x$survive)
                      return(c("N"=nRow,"S"=numSurvived))
                  })
print(dGrouped)
```

```
    sex   N   S
1   man 869 175
2 women 447 324
```

Our regression now has no covariates. The grouping is the covariate we are intersted in.

We can run logistic regression for a Binomial variable similar to LR for a 0-1 variable. Our target is two columns: the first column of successes (survivors) and second column of failure (non-survivors).

[31]:
```
binomialLogR = glm(cbind(S,N-S)~. , data = dGrouped, family = binomial)
print(summary(binomialLogR))
```

```
Call:
glm(formula = cbind(S, N - S) ~ ., family = binomial, data = dGrouped)
```

```
Deviance Residuals:
[1]  0  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.37769    0.08459  -16.29   <2e-16 ***
sexwomen     2.34625    0.13554   17.31   <2e-16 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance:  3.4777e+02  on 1  degrees of freedom
Residual deviance: -1.8607e-13  on 0  degrees of freedom
AIC: 17.109

Number of Fisher Scoring iterations: 2
```

These two models are equivalent and so give identical estimate of survival based on sex. This may not always be the case, but it is clear that logistic regression can be generalized to any binomial distributed variable. Bernoulli, or 0-1, data is a special case of Logistic regression for a Binomially distributed target.