Binomial Outcomes

Frequently dichotomous outcomes are used in medical studies, such as presence of a disease or exposure to some factor. The classical model does not apply outcomes that are not continuous, and many other reasons:

1. The variance of the dichotomous response is not constant
2. One could predict probabilities > 1 or smaller than 0
3. Limited interpretation of parameters

In this section we will observe alternatives to traditional linear regression for binomial data.

Generalized Linear Model For Binomial Data

- The outcome is distributed either Bernoulli or Binomial with probability of success θ
- Link function is logit and relates to the probability of success θ to the linear predictor η = β0 +β1 X1 +
  β2X2 + · · · + βk−1Xk−1 + βk Xk as follows:
  $$\text{logit}\,(\theta) = \log\left(\frac{\theta}{1-\theta}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{k-1} X_{k-1} + \beta_k X_k =: \eta$$
- Variance function: V(θ) = θ (1 - θ)

Recall: For ordinary linear regression the distribution of outcomes is normal and the link is the identity function, and the variance is constant. For loglinear models the distribution of the outcome is Poisson, the link function is the LOG function and the variance function is the identity.

The link function is logistic because the logit function takes value between -inf to inf while the logodds is between 0 and 1.

$$\frac{\theta\,(x)}{1-\theta\,(x)} = exp\,(\alpha + \beta x) \implies \begin{matrix} \text{ODDS}_{\text{disease}|\text{Exposure=Yes}} = \exp\{\alpha + \beta\} \\ \text{ODDS}_{\text{disease}|\text{Exposure=No}} = \exp\{\alpha\} \end{matrix} \implies \begin{matrix} OR = \frac{\exp\{\alpha+\beta\}}{\exp\{\alpha\}} \\ \log\,(\text{OR}) = \beta \end{matrix}$$

β represents the increase in logodds for a one-unit increase in X

The coefficients can also be interpreted as probability (risk) difference. However, the change is a function of base probability:

$$p\,(\text{disease}|\text{Exposure} = \text{Yes}) - p\,(\text{disease}|\text{Exposure} = \text{No}) = \frac{\exp\{\alpha + \beta\}}{\exp\{\alpha + \beta\} + 1} - \frac{\exp\{\alpha\}}{\exp\{\alpha\} + 1}$$

Probit, Complementary Log-Log, and Logistic Regression

Probit regression and **Complementary log-log regression** are generalized linear models for the same type of data as the logistic regression, the only difference is in the choice of the link functions.

- Probit uses a link function which is the inverse of the Cumulative Distribution Functions (CDF) of a standard normal distribution
    - Link function:
      $$\mathrm{Probit}\left(x\right) = \Phi^{-1}(x)$$
      Where
      $$\Phi\left(x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\,\mathrm{dy}$$
    - Probit and logistic will in general give similar results in terms of significance; The advantage of logistic regression is interpretability of coefficients as OR's
    - Logit and Probit haave the symmetry property, that is:
      $$g\left(\mu\right) = -g(1-\mu)$$
      $\text{for the logit link } g\left(\mu\right) = \log\left(\frac{\mu}{1-\mu}\right) \text{ and}$
      $$g\left(\mu\right) = \log\left(\frac{\mu}{1-\mu}\right) = -\log\left(\frac{1-\mu}{\mu}\right) = -g(1-\mu)$$
- Complementary log-log is inverse of the CDF for an extreme value distribution
    - The link function:
      $$g\left(p\right) = \log\left(-\log\left(p\right)\right)$$
    - Does not have the symmetry property, a good choice when symmetry property does not hold.

Interpretation of Parameters

Note we've already seen how to interpret logodds in 806 and 852 so I will not being going into great detail here.

One neat thing is that we can express any model using shorthand.

$$\log\left(\frac{\theta}{1-\theta}\right) = \beta_0 + \beta_1 \mathrm{CHO}L^{<200} + \beta_2 \mathrm{CHO}L^{200-219} + \beta_3 \mathrm{CHO}L^{220-259} + \gamma_1 \mathrm{SB}P^{<127} + \gamma_2 \mathrm{SB}P^{127-146} + \gamma_3 \mathrm{SB}P^{147-166}$$

Where CHOL and SBP are binary variables representing levels of these measures. The shorthand equivalent:

$$\log\left(\frac{\theta_{ij}}{1-\theta_{ij}}\right) = \beta_0 + \lambda_i^{\mathrm{CHOL}} + \lambda_j^{\mathrm{SBP}}$$

While not as informative about the levels and coefficients, it is much easier to write.

Testing GoF and Model Selection

As in any other model we can assess the GoF by comparing the observed to the predicted. The deviance statistic (also called the Likelihood Ratio Statistic) is the GoF measure of choice for this class. It can be used to compare any *nested* model, and since every model is nested within the saturated model, we can use:

$$G^2\left(M\right) = 2 \times \left[\log \text{Likelihood}(\text{Saturated Model}) - \log \text{Likelihood}(\text{Model M})\right]$$

Which reduces to the following for binomial models:

$$G^2\left(M\right) = \sum_i y_i \log\left(\frac{y_i}{E_i}\right)$$

Where $E_i = n_i \pi_i$

The Pearson Chi-Squared Test are application for logistic regression when the data can be aggregated, or grouped into unique profiles determined by predictors. Aggregation is **only** possible when **all** predictors are categorical.

With ungrouped data the Pearson chi-square test is:

$$X^2(M) = \sum_i \frac{\left(y_i - \widehat{\pi}_i\right)^2}{\widehat{\pi}_i\left(1 - \widehat{\pi}_i\right)}$$

Where yi is the observed binary outcome, and pi_hat is the model estimate of probability of yi=1. This does NOT have a asymptotic chi-square distribution, but when properly standardized it follows a normal distribution

With grouped data the Pearson chi-square test is:

$$X^2(M) = \sum_i \frac{\left(y_i - n_i\widehat{\pi}_i\right)^2}{n_i\widehat{\pi}_i\left(1 - \widehat{\pi}_i\right)}$$

which follows a chi-square distribution

Note that when all independent variables are categorical, a logistic regression can be estimated using a suitably formulated loglinear model - the converse is NOT always true.

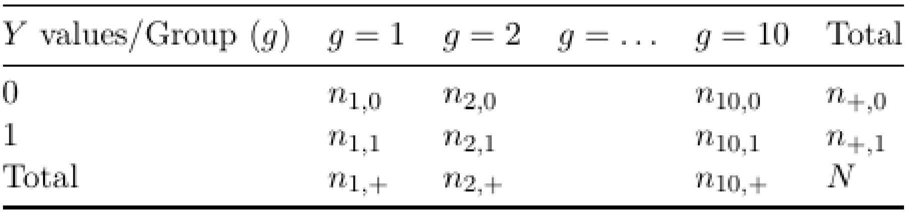
Hosmer-Lemeshow GoF Test

The probabilities of the event are estimated using the model for each unit in the data as:

$$g\left(\widehat{\pi}_i\right) = \widehat{\beta}_0 + \widehat{\beta}_1 X_i^2 + \widehat{\beta}_2 X_i^2 + \widehat{\beta}_3 X_i^3$$ _or $$\widehat{\pi}_i = g^{-1}\left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i^2 + \widehat{\beta}_2 X_i^2 + \widehat{\beta}_3 X_i^3\right)$$

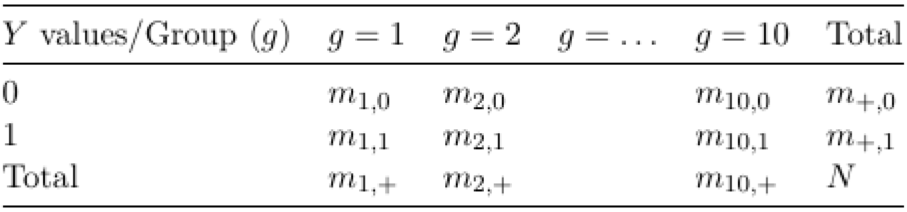
Where β parameteres are estimated by maximum likelihood. These predicted values are then compared with the observed values of Y (which is binary)

The HL test is based on ordering the sample according to the risk or predicted probability. The estimated probabilities are split into 10 groups (by default), the first group consists of the 10% lowest estimated probabilities and so on.Based on the observed Yi's we can calculate the the number of Yi's equal to 0 or 1 in each group.



Y values/Group (g)	$g = 1$	$g = 2$	$g = \ldots$	$g = 10$	Total
0	$n_{1,0}$	$n_{2,0}$		$n_{10,0}$	$n_{+,0}$
1	$n_{1,1}$	$n_{2,1}$		$n_{10,1}$	$n_{+,1}$
Total	$n_{1,+}$	$n_{2,+}$		$n_{10,+}$	N

To estimate the number of 1's and 0's expected base on the model, we calculate the average of the estimated probabilities in each group and multiply by the number of observations per group.



Y values/Group (g)	$g = 1$	$g = 2$	$g = \ldots$	$g = 10$	Total
0	$m_{1,0}$	$m_{2,0}$		$m_{10,0}$	$m_{+,0}$
1	$m_{1,1}$	$m_{2,1}$		$m_{10,1}$	$m_{+,1}$
Total	$m_{1,+}$	$m_{2,+}$		$m_{10,+}$	N

The Hosmer-Lemshow test is a Pearson chi-squared statistic comparing the two tables:

$$HL = \sum_{\text{cell}} \frac{(n_{\text{cell}} - m_{\text{cell}})^2}{m_{\text{cell}}}$$

In the original 1980 paper HL showed the test statistic followed a ~chi-squared distribution with 10 - 2 = 8 degrees of freedom. A small p-value indicates poor fit.

Criticisms

- The HL test could depend on the number of groups/choosing a different number of groups can yield different results
- The suggested number of groups is dependent on the number variables in the model, that is #predictors + 1

Measuring Predictive Ability

There are several measures for quantifying prediction ability in logistic regression. The first set of measures below is based on maximum likelihood values and can be used for any likelihood based models. There are measures specific to models with multi-nomial outcomes.

Maximum Likelihood Based Measures

- L* = -2 * LL ; Minus twice the maximum log-likelihood of the saturated model (lowest possible value)
- L = -2 * LL ; Minus twice the maximum log-likelihood of the current model
- L^0 = -2 * LL ; Minus twice the maximum log-likelihood of the model including only intercept

The deviance is a test for the significance of all predictors in addition to the intercept:

LR = L^0 - L

The largest value that can be explained by a model:
$LR = L^0 - L^*$

The fraction that is explained from what can by explained is:
$R^2 = (L^0 - L) / (L^0 - L^*) = LR / (L^0 - L^*)$
Which is the same as R^2 from linear regression. However, this statistic has some undesirable properties, to circumvent these $R_{LR}{}^2$ has been proposed as:

$$R^2_{\text{LR}} = 1 - \exp\left(-\frac{\text{LR}}{n}\right)$$

Which is how SAS calculates RSQUARE in proc logistic; And adjusted R^2:

$$R^2_N = \frac{R^2_{\text{LR}}}{\left[1 - \exp\left(-\frac{L_0}{n}\right)\right]}$$

Other measures:

- c-statistic or Area Under Curve (AUC) with a value close to 1 indicating good prediction ability while a value close to .5 indicates a poor predictive ability. Closely related to Mann-Whitney, Wilcoxon statistics and the Gini Coefficient
- Somers' D statistic $D_{XY} = 2(c - .5)$ rank correlation with a value close to 1 indicating good prediction ability
- Recent of correctly classified responses

Adjusting for Optimism (Bias) In Measures of Predictive Ability

When the same dataset is used to fit the model and assess its predictive ability it can result in a 'optimistic' estimate of predictive ability. There are a number of approaches to adjust these estimates including: data splitting, cross-validation and bootstrapping.

Using Bootstrap to Estimate Optimism

Bootstrapping is a generic and efficient approach in statistics. It involves repeatedly sampling with replacement from the observed dataset. This process can be repeated to create many (say B) bootstrap datasets. The fundamental principle at work is that the original observed data are a good approximation on what the population on the whole of interest is, and then the bootstrap data set represent instances of data samples from that population.

1. Fit model to original data, and estimate C_0
2. Obtain B bootstrap data as indicated above; For each:
   1. Fit the model to the bootstrap datasets and estimate C_b
   2. Estimate the c-statistic of the model in 2a. on the original data to be $C_{b,0}$
3. Since the model in 2a was estimated on bootstrap data it is expected that $C_{b,0} < C_b$, estimate the optimism as:

$$Optimism = \frac{1}{B}\sum_{b}\left[C_b - C_{b,o}\right]$$

4. Estimate the corrected c-statistic as C_0 - Optimism

This approach for correcting the measure of predictive ability is very general and applies to all the measures presented above. The ideal estimate of predictive ability of a model is obtained on an independent sample.

SAS Code

```
options ps=60 ls=89 pageno=1 nodate;


/* CHD in Framingham Study*/
proc format ;
value sbp 1='<127'  2='127 - 146'  3='147 - 166'  4='167+';
value chl 1='<200'  2='200 - 219'   3='220 - 259'   4='>=260';
run;
data chd;
input Chol SBP CHD Total;
prchd=(chd+0.5)/total;
logitp=log((prchd)/(1- prchd));
format sbp sbp. Chol chl.;
cards;
1 1 2  119
1 2 3  124
1 3 3  50
1 4 4  26
2 1 3  88
2 2 2  100
2 3 0  43
2 4 3  23
3 1 8  127
3 2 11 220
3 3 6  74
3 4 6  49
4 1 7  74
4 2 12 111
4 3 11 57
4 4 11 44
```

```
;
run;



/* First, examine the relationships of incidence of CHD with Chol and SBP graphically. */
goptions reset=all ftext='arial' htext=1.2 hsize=9.5in vsize=7in aspect=1 horigin=1.5in vorigin=0.3in ;
goptions device=emf rotate=landscape gsfname=TempOut1 gsfmode=replace;
  filename TempOut1 "C:\Documents and Settings\doros\My Documents\Projects\BS853\Class 4\Framingham.emf";
axis1 label=(h=1.5 'Cholesterol') minor=none order=(1 to 4 by  1)
value=(tick=1 '< 200' tick=2 '200 - 219' tick=3 '220 - 259' tick=4 '>= 260') offset=(1);
axis2 label=(h=1.5  a=90 'Logit(Probability)') minor=none ;

symbol1 v=none i=j ci=red line=1;
symbol2  v=none i=j ci=green line=2;
symbol3  v=none i=j ci=blue line=3;
symbol4  v=none i=j ci=magenta line=4;
title1 'Incidence of CHD with Chol and SBP';
proc gplot data=chd;
 plot logitp*chol=sbp/haxis=axis1 vaxis=axis2;
 run;
quit;
goptions reset=all;

/* Fit different Models using GENMOD and LOGISTIC*/
options ls=100 ps=60 nodate pageno=1;
title1 'Only Intercept Effect';
proc genmod data=CHD;
class CHOL SBP;
model CHD/Total=/dist=Binomial link=logit;
run;

title1 'Only Chol Effect';
proc genmod data=CHD;
class CHOL SBP;
model CHD/Total=CHOL /dist=Binomial link=logit;
run;

title1 'Only SBP Effect';
proc genmod data=CHD;
class CHOL SBP;
```

```
model CHD/Total=SBP /dist=Binomial link=logit;
run;


title1 'Only main Effects';
ods output obstats=check;
proc genmod data=CHD;
class CHOL SBP;
model CHD/Total= CHOL SBP /dist=Binomial link=logit obstats;
run;

title1 'Only main Effects - Using PROC LOGISTIC';
proc logistic data=CHD;
class CHOL SBP(ref='167+');
model CHD/Total= CHOL SBP ;
run;

title1 'Only main Effects - Continuous predictors';
ods output obstats=check;
proc genmod data=CHD;
model CHD/Total= CHOL SBP /dist=Binomial link=logit obstats ;
run;

title1 ' Saturated Model ';
proc genmod data=CHD;
class CHOL SBP;
model CHD/Total= CHOL|SBP /dist=Binomial link=logit;
run;

title1 'Only main Effects - Hosmer-Lemeshow';
ods select LackFitPartition LackFitChiSq;
proc Logistic data=CHD;
  model CHD/Total= CHOL SBP / LACKFIT;
run;

title1 'Only main Effects - R-Squared and R-Squared Nagelkerke';
ods select RSQUARE;
proc Logistic data=CHD;
  model CHD/Total= CHOL SBP / RSQUARE;
run;
```

```
 /* Logistic regression models as Log-Linear models*/
data chd2;
input chol sbp chd count @@;
datalines;
1 1 1 2  1 1 2 117
1 2 1 3  1 2 2 121
1 3 1 3  1 3 2  47
1 4 1 4  1 4 2  22
2 1 1 3  2 1 2  85
2 2 1 2  2 2 2  98
2 3 1 0  2 3 2  43
2 4 1 3  2 4 2  20
3 1 1 8  3 1 2 119
3 2 1 11 3 2 2 209
3 3 1 6  3 3 2  68
3 4 1 6  3 4 2  43
4 1 1 7  4 1 2  67
4 2 1 12 4 2 2  99
4 3 1 11 4 3 2  46
4 4 1 11 4 4 2  33
;
title1 'Logistic regression models as Loglinear models';
title2 'Model 1' ;
ods select parameterestimates modelfit;
proc genmod data=chd2;
class chol sbp chd;
model count=CHD SBP|CHOL /link=log dist=poisson obstats;
run;

ods select parameterestimates modelfit;
title2 'Model 2' ;
proc genmod data=chd2;
class chol sbp chd;
model count=CHD|SBP SBP|CHOL/link=log dist=poisson obstats;
run;

ods select parameterestimates modelfit;
title2 'Model 3' ;
proc genmod data=chd2;
class chol sbp chd;
model count=CHD|CHOL SBP|CHOL/link=log dist=poisson obstats;
```

```
run;
options ls=90;
ods select parameterestimates;* modelfit;
title2 'Model 4' ;
proc genmod data=chd2;
class chol sbp chd;
model count=CHD|CHOL CHD|SBP SBP|CHOL/link=log dist=poisson obstats;
run;

/**************************************************************/
/* Admission Data                                             */
/**************************************************************/
/* Ignoring Department */
data overall;
input sex $ yes total;
cards;
M 1198  2691
F   557  1835
;
/* Saturated model: with class statement*/;
ODS select modelfit ParameterEstimates;
title 'Differential admission by Gender';
proc genmod data=overall;
class sex;
model yes/total=sex/link=logit dist=bin obstats;
estimate 'overall gender' sex 1 -1/exp;
run;

proc genmod data=overall;
class sex;
model yes/total=/link=logit dist=bin obstats;
run;
/* By Department */
data one;
do Department=1 to 6;
do Sex='M', 'F';;
input yes no;
logitp=log(yes/(yes+no));
total=yes+no;
output;
end;
```

```
end;
cards;
512 313
89 19
353   207
17     8
120   205
202   391
138   279
131  244
53   138
94   299
22   351
24   317
;run;

/* Explain why marginally women seem to be discriminated against
   - More women apply to harder to enter colodges!!!             */

proc sql; create table a as
select department, sum(total) as total , sum(yes)/sum(total) as adminp from one group by department  order by
department;
 create table b as select department, total as fem from one where sex='F' order by department;
 create table c as
 select *, fem/total as pf from a as aa, b as bb where aa.department=bb.department;
quit;

 symbol1 v=circle i=join;
 proc gplot data=c;
 plot pf*adminp;
run;



goptions reset=all ftext='arial' htext=1.2 hsize=9.5in vsize=7in aspect=1 horigin=1.5in vorigin=0.3in ;
goptions device=emf rotate=landscape gsfname=TempOut1 gsfmode=replace;
    filename TempOut1 "C:\Documents and Settings\doros\My Documents\Projects\BS853\Class 4\admission.emf";
  symbol1 v=dot i=j ci=red;
  symbol2 v=circle i=j ci=blue;
  axis1 label=(h=1.5 'Department') minor=none order=(1 to 6 by  1)  offset=(1);
axis2 label=(h=1.5  a=90 'Logit(Rate)') minor=none ;
```

```
title1 'Admitance by Department and Sex';
proc gplot data=one;
plot logitp*Department=sex/vaxis=axis2 haxis=axis1;
run;

title1 'Saturated Model';
proc genmod data=one;
class department sex;
model yes/total=sex|department/dist=b link=logit;
run;

title1 'Only main effects Model';
proc genmod data=one;
class department sex;
model yes/total=sex department/dist=b link=logit;
run;

title1 'Only Sex Model';
proc genmod data=one;
class department sex;
model yes/total=sex/dist=b link=logit;
estimate 'l' sex 1 -1/exp;
run;

title1 'Only Department effects Model';
proc genmod data=one;
class department sex;
model yes/total=department/dist=b link=logit;
run;

ods select estimates;
title1 'Estimated ODDS Ratio (F vs M) in each Department';
proc genmod data=one;
class department sex;
model yes/total=sex|department/link=logit dist = bin covb;
estimate 'sex1'  sex 1 -1  department*sex 1 -1 0  0 0  0 0  0 0  0 0  0    /exp ;
estimate 'sex2'  sex 1 -1  department*sex 0  0 1 -1 0  0 0  0 0  0 0  0    /exp ;
estimate 'sex3'  sex 1 -1  department*sex 0  0 0  0 1 -1 0  0 0  0 0  0    /exp ;
estimate 'sex4'  sex 1 -1  department*sex 0  0 0  0 0  0 1 -1 0  0 0  0    /exp ;
estimate 'sex5'  sex 1 -1  department*sex 0  0 0  0 0  0 0  0 1 -1 0  0    /exp ;
estimate 'sex6'  sex 1 -1  department*sex 0  0 0  0 0  0 0  0 0  0 1 -1    /exp ;
```

```
run;

/* Modeling Trends in Proportions */
data refcanc;
do age=1 to 4;
input cases ref;
total=cases+ref;
output;
end;
cards;
26	20
4	7
3	10
1	11
;
run;

/* only intercept */
title1 'Only Intercept model';
proc genmod data=refcanc;
class age;
model cases/total=;
run;

/* Saturated model */
title1 'Saturated model';
proc logistic data=refcanc;
class age;
model cases/total=age;
run;
title1 'Linear trend model';

proc logistic data=refcanc;
model cases/total=age;
run;
```

Revision #5
Created 14 February 2023 19:17:32 by Elkip
Updated 14 February 2023 21:38:32 by Elkip