Topic 30 - Logistic Regression

STAT 525 - Fall 2013

Outline

- Logistic Regression
  - Background
  - Model
  - Inference
  - Diagnostics

Background

- In many applications, the response variable Y has only two possible outcomes, labeled numerically 0 and 1
  - Not diseased ($Y = 0$) vs Disease ($Y = 1$)
  - Unemployed ($Y = 0$) vs Employed ($Y = 1$)
- Response is *binary* or *dichotomous*
- Can model response using Bernoulli distribution

$$\begin{array}{l} \Pr(\mathrm{Y_i} = 1) = \pi_\mathrm{i} \\ \Pr(\mathrm{Y_i} = 0) = 1 - \pi_\mathrm{i} \end{array} \quad \rightarrow \quad \mathrm{E(Y_i)} = \pi_\mathrm{i}$$

- Goal is to <u>link</u> $E(Y_i) = \pi_i$ with covariates $\mathbf{X_i}$

Just use the linear link?

- Suppose we consider the linear model (with one X_i)

$$Y_i \quad = \quad \beta_0 + \beta_1 X_i + \varepsilon_i$$

- If Y_i only takes values 0 and 1, then
  - We have non-normal error terms
    when $Y_i = 0 : \varepsilon_i = -\beta_0 - \beta_1 X_i$
    when $Y_i = 1 : \varepsilon_i = 1 - \beta_0 - \beta_1 X_i$
  - We have nonconstant variance

$$\mathrm{Var(Y_i)} \quad = \quad \pi_i(1 - \pi_i)$$

  - Need parameter bounds so $0 \leq \pi_i \leq 1$

Logistic Response Function

- Need alternate function to link $E(Y_i) = \pi_i$ and X
- Common to consider the sigmoidal function

$$\begin{aligned} E(Y_i) &= \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} \\ &= (1 + \exp(-\beta_0 - \beta_1 X_i))^{-1} \end{aligned}$$

- Example of a **nonlinear** model
- Other sigmoidal functions (e.g., normal CDF) possible

Properties of Logistic Function

- Monotonic increasing/decreasing function
- Restricts $0 \leq \mathrm{E(Y_i)} \leq 1$
- Can be linearized through the *logit* transformation

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i$$

- Use **logit link** function to relate π_i with $\mathbf{X_i}$
- Other link functions possible (i.e., probit, complementary log-log)

Logistic model

- Assume the Y_i are independent Bernoulli random variables with means π_i
- Could express model as $Y_i = E(Y_i) + \varepsilon_i$ but the error terms ε_i depend on the Bernoulli distribution of Y_i
- Better to express

$$Y_i \sim \mathrm{Bernoulli}(\pi_\mathrm{i})$$
$$\mathrm{logit}(\pi_\mathrm{i}) = \beta_0 + \beta_1 \mathrm{X_i}$$

Estimation

- Given the distributions of Y_i, we can formulate the likelihood function and obtain MLE estimates

$$\begin{aligned} \log(L) &= \log(\prod \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}) \\ &= \sum Y_i \log(\pi_i) + \sum (1-Y_i)\log(1-\pi_i) \\ &= \sum Y_i \log(\frac{\pi_i}{1-\pi_i}) + \sum \log(1-\pi_i) \\ &= \sum Y_i(\beta_0 + \beta_1 X_i) - \sum \log(1+\exp(\beta_0+\beta_1 X_i)) \end{aligned}$$

- MLEs do not have closed forms
- SAS performs iterative reweighted least squares

Interpretation

- Given b_0 and b_1, can calculate $\hat{\pi}_i$

$$\hat{\pi}_i = \exp(b_0 + b_1 X_i)/(1 + \exp(b_0 + b_1 X_i))$$

- $\hat{\pi}_i$ is the estimated probability of individual i having response $Y_i = 1$
- b_1 is now the "slope" of the logit relationship

$$\text{logit}(\hat{\pi}(X_i + 1)) - \text{logit}(\hat{\pi}(X_i)) = b_1$$

- Logit transform is the log of the **odds**
- Thus, $\exp(b_1)$ becomes the **odds ratio**
- Popular summary in epidemiologic studies

Repeat Observations

- Particularly in designed experiments will have repeat observations at certain levels of X
- Allows for assessment of model fit (recall Section 3.7)
- Label X levels as $X_1, X_2, \ldots, X_c$ such that there are n_j replicates at level X_j with Y_{ij} the i^{th} replicate at X_j
- Can simplify log-likelihood function by considering the sums $Y_{\cdot j}$, which are binomially distributed

$$Y_{\cdot j} \sim \text{Binomial}(n_j, \pi_j)$$
$$\text{logit}(\pi_j) = \beta_0 + \beta_1 X_j$$

Example Page 625

- Board of directors interested in estimating the effect of a due increase on membership
- Randomly surveyed $n = 30$ members
- X_i is the due increase posited to the member
- X_i varied between \$30 and \$50
- Repeat observations for certain X_i
- Y_i is whether the member would continue membership

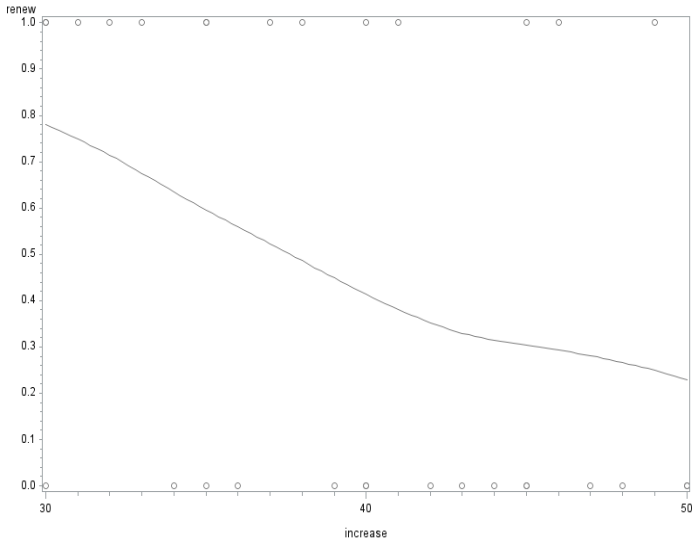
SAS Commands

```
data a1;
   infile 'u:\.www\datasets525\CH14PR07.txt';
   input norenew increase;
   renew=1-norenew;


proc print data=a1;
 var renew increase;


symbol1 v=circle i=sm70;
proc gplot;
 plot renew*increase;
```

The Data - Single Trials

Obs	renew	increase
1	1	30
2	0	30
3	1	30
4	1	31
5	1	32
6	1	33
7	0	34
8	1	35
9	1	35
10	0	35
11	0	36
12	1	37
:	:	:
27	0	48
28	1	49
29	0	50
30	0	50

Scatterplot



SAS Commands

```
***Converting data set to events/trials format****
proc sort data=a1; by renew;
proc freq data=a1; tables increase / out=a1c; by renew;

data renew0;
 set a1c; if renew=0; n0=count; drop count;

data renew1;
 set a1c; if renew=1; n1=count; drop count;

data a1c;
 merge renew0 renew1; by increase;
 if n0=. then n0=0;  if n1=. then n1=0;
 tot=n0 + n1;

proc print data=a1c;
 var increase n1 tot;
run;
```


The Data - Grouped trials

Obs	increase	n1	tot
1	30	2	3
2	31	1	1
3	32	1	1
4	33	1	1
5	34	0	1
6	35	2	3
7	36	0	1
8	37	1	1
9	38	1	1
10	39	0	1
11	40	1	3
:	:	:	:
16	45	1	3
17	46	1	1
18	47	0	1
19	48	0	1
20	49	1	1
21	50	0	2

SAS Commands

```
proc logistic data=a1 descending;
   model renew = increase;                **single trial form
   output out=a2 p=pred;
proc print;

symbol1 v=circle i=none;
symbol2 v=star i=sm30;
proc gplot data=a2;
   plot renew*increase pred*increase /overlay;

proc logistic data=a1c;
   model n1 / tot = increase;             **grouped trial form
run;
```


Output

The LOGISTIC Procedure ***Single trials model***

Response Profile

Ordered Value	renew	Total Frequency
1	1	14
2	0	16

Probability modeled is renew=1. *****Be wary*****

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates	
AIC	43.455	41.465	
SC	44.857	44.267	**Difference in -2 Log L**
-2 Log L	41.455	37.465	3.990

Output

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq	
Likelihood Ratio	3.9906	1	0.0458	**see prev page
Score	3.8265	1	0.0504	
Wald	3.5104	1	0.0610	

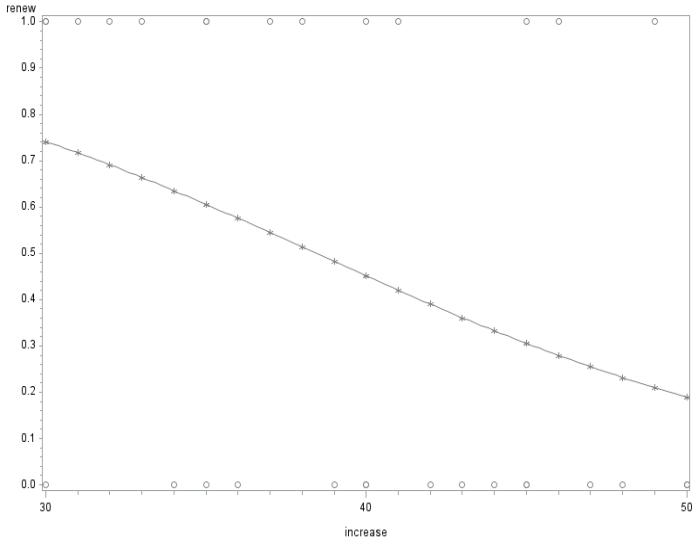
Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	4.8075	2.6558	3.2769	0.0703
increase	1	-0.1251	0.0668	3.5104	0.0610

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits		**Wald test like squared** z test
increase	0.882	0.774	1.006	

Scatterplot/Fit



Output

The LOGISTIC Procedure ***Grouped trials model***

Response Profile

Ordered Value	Binary Outcome	Total Frequency
1	Event	14
2	Nonevent	16

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates Log Likelihood	Full Log Likelihood	
AIC	43.455	41.465	32.676	**Last column considers
SC	44.857	44.267	35.478	binomial likelihood
-2 Log L	41.455	37.465	28.676	

Output

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	3.9906	1	0.0458
Score	3.8265	1	0.0504
Wald	3.5104	1	0.0610

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	4.8075	2.6558	3.2769	0.0703
increase	1	-0.1251	0.0668	3.5104	0.0610

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits		***Same results as before
increase	0.882	0.774	1.006	

Alternate Approaches

Logistic regression is an example of a generalized model so
*** we can consider generalized linear model procedures ***

```
proc genmod data=a1 descending;
   model renew = increase / dist=bin aggregate=increase;
                            **Without aggregate fits single trial model
proc genmod data=a1c descending;
   model  n1 / tot = increase / dist=bin;
run;                        **Fits grouped trial model

proc glimmix data=a1;
   model renew(descending) = increase / chisq dist=binary;
                            **Fits single trial model
proc glimmix data=a1c;
   model n1/ tot = increase / chisq dist=binomial;
run;                        **Fits grouped trial model
```


GENMOD Output

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF	
Deviance	19	22.1885	1.1678	
Scaled Deviance	19	22.1885	1.1678	
Pearson Chi-Square	19	17.9966	0.9472	
Scaled Pearson X2	19	17.9966	0.9472	
Log Likelihood		-18.7324		
Full Log Likelihood		-14.3379	**1st and 3rd columns	
AIC (smaller is better)		32.6759	only provided with	
AICC (smaller is better)		33.1203	grouped model	
BIC (smaller is better)		35.4783		

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	4.8075	2.6558	-0.3977	10.0127	3.28	0.0703
increase	1	-0.1251	0.0668	-0.2559	0.0058	3.51	0.0610
Scale	0	1.0000	0.0000	1.0000	1.0000		

GLIMMIX Output

Fit Statistics		
-2 Log Likelihood	37.46	
AIC (smaller is better)	41.46	**Single trial model
AICC (smaller is better)	41.91	
BIC (smaller is better)	44.27	**Likelihood and model
CAIC (smaller is better)	46.27	selection criteria
HQIC (smaller is better)	42.36	based on single trial
Pearson Chi-Square	30.10	likelihood
Pearson Chi-Square / DF	1.08	

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	Chi-Square	F Value	Pr > ChiSq	Pr > F
increase	1	28	3.51	3.51	0.0610	0.0715

Will not discuss F test but Den df = 30-1-1

GLIMMIX Output

Fit Statistics		
-2 Log Likelihood	28.68	
AIC (smaller is better)	32.68	**Grouped trial model
AICC (smaller is better)	33.34	
BIC (smaller is better)	34.76	**Likelihood and model
CAIC (smaller is better)	36.76	selection criteria
HQIC (smaller is better)	33.13	based on grouped trial
Pearson Chi-Square	18.00	likelihood
Pearson Chi-Square / DF	0.95	

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	Chi-Square	F Value	Pr > ChiSq	Pr > F
increase	1	19	3.51	3.51	0.0610	0.0765

Den df = 21-1-1

Comparison of Global Tests

- **Likelihood ratio test** compares $-2\log(L)$ of two nested models. The df equals the difference in the number of model parameters
- **Wald test** requires fit only to full model. Assesses how far estimated parameters are from hypothesized values in terms of standard errors. Uses asymptotic normality of MLEs. When considering only one parameter, test just the squared-z test.
- **Score test** requires fit only to H_0 model. Test is based on the slope of the log-likelihood function at the values specified by the null hypothesis. Describes expected change in chi-squared statistic if variables were added.

Deviance

- Recall hypothesis testing can be done using a likelihood ratio test, which is similar to the general linear test
- Consider the grouped trials form. The deviance of the fitted model is the difference in the log-likelihood of the fitted model and the most general model, which has a parameter (π_j) for each X_j
- The MLEs for this full model are $p_j = Y_{\cdot j}/n_j$ for $j = 1, 2, \ldots, c$ so the deviance for the fitted model is

$$-2\left(\log(L(\mathrm{R})) - \log(L(\mathrm{F}))\right)$$

$$-2\sum\left(Y_{\cdot j} log\left(\frac{\hat{\pi}_j}{p_j}\right) + (n_j - Y_{\cdot j}) log\left(\frac{1-\hat{\pi}_j}{1-p_j}\right)\right)$$

Using the Deviance

- If the logistic model is correct, then the deviance will be approximately chi-square with $c - p$ df.
- Since the mean of the chi-square is its df, we'd expect the deviance/df to be approximately 1. This is commonly used to assess goodness of fit.
- Can also use deviance to compare two hierarchical models

$$\begin{aligned} \mathrm{DEV}(\mathrm{X_q}, ..., \mathrm{X_{p-1}} | \mathrm{X_0}, ..., \mathrm{X_{q-1}}) &= \mathrm{DEV}(\mathrm{X_0}, ..., \mathrm{X_{q-1}}) \\ &- \mathrm{DEV}(\mathrm{X_0}, ..., \mathrm{X_{p-1}}) \end{aligned}$$

- Partial deviance approx χ^2 with $p - q$ df
- Must compute by hand or with GENMOD
- Will use this in multiple logistic regression

Pearson Goodness of Fit Test

- Alternative goodness of fit test using same principals as deviance goodness of fit test.
- For each observed X_j, can compute the expected number of events and nonevents.

$$X^2 = \sum_c \left(\frac{(Y_{.j} - n_j \hat{\pi}_j)^2}{n_j \hat{\pi}_j} + \frac{(n_j - Y_{.j} - n_j(1 - \hat{\pi}_j))^2}{n_j(1 - \hat{\pi}_j)} \right)$$

- This also will approximately follow a chi-square distribution with $c - p$ df.

Wald's Test

- Can also use Wald's Test for $H_0 : \mathbf{L}'\boldsymbol{\beta} = \mathbf{C}$
- Described on page 578 for single parameter tests

$$S = (\mathbf{L}'\hat{\boldsymbol{\beta}} - \mathbf{C})'(\mathbf{L}'\hat{\boldsymbol{\Sigma}}\mathbf{L})^{-1}(\mathbf{L}'\hat{\boldsymbol{\beta}} - \mathbf{C})$$

where $\hat{\Sigma}$ is the estimate covariance matrix of $\hat{\boldsymbol{\beta}}$

- Under H_0, $S \sim \chi^2_r$ where r is rank of $\mathbf{L}$
- Available in GENMOD and LOGISTIC

Alternative Goodness of Fit Test

- Have previously described goodness of fit tests when there is replication
- For unreplicated studies, can consider the Hosemer-Lemeshow goodness of fit
  - Group observations into classes (usually around 10) according to fitted logit values.
  - Assess overall fit to each class using a Pearson goodness of fit approach.

Hosemer-Lemeshow Goodness of Fit Test

- Divide obs up into ≈ 10 groups of equal size based on percentiles of the estimated probabilities
- Expected # of 1's is $\sum \hat{\pi}_i$
- Expected # of 0's is $n_i - \sum \hat{\pi}_i$
- Compare expected with observed through

  $\chi^2 = \sum \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
- In the example
  - $E_{11} = (.19056 + .19056 + .21060) = 0.59 \rightarrow E_{10} = 2.41$
  - $E_{21} = (.23214 + .25518 + .27967) = 0.77 \rightarrow E_{20} = 2.23$

SAS Commands

```
data a1;
   infile 'u:\.www\datasets525\CH14PR07.txt';
   input norenew increase;
   renew=1-norenew;

proc logistic data=a1 descending;
   model renew = increase / lackfit;     *******NEW******
   output out=a2 p=pred;
run;

proc print;
run;
```


Additional Output

Association of Predicted Probabilities and Observed Responses

Percent Concordant	68.3	Somers' D	0.402
Percent Discordant	28.1	Gamma	0.417
Percent Tied	3.6	Tau-a	0.207
Pairs	224	c	0.701

Partition for the Hosmer and Lemeshow Test

		renew = 1		renew = 0	
Group	Total	Observed	Expected	Observed	Expected
1	3	1	0.59	2	2.41
2	3	1	0.77	2	2.23
3	3	1	0.92	2	2.08
4	3	0	1.08	3	1.92
5	4	2	1.77	2	2.23
6	3	2	1.54	1	1.46
7	4	2	2.39	2	1.61
8	3	2	1.99	1	1.01
9	4	3	2.94	1	1.06

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
2.6526	7	0.9152

Output

Obs	norenew	increase	renew	_LEVEL_	pred
1	1	50	0	1	0.19056
2	1	50	0	1	0.19056
3	0	49	1	1	0.21060
4	1	48	0	1	0.23214
5	1	47	0	1	0.25518
6	0	46	1	1	0.27967
:	:	:	:	:	:
17	1	39	0	1	0.48237
18	0	38	1	1	0.51363
19	0	37	1	1	0.54478
24	1	34	0	1	0.63526
25	0	33	1	1	0.66372
26	0	32	1	1	0.69104
27	0	31	1	1	0.71709
28	0	30	1	1	0.74177
29	1	30	0	1	0.74177
30	0	30	1	1	0.74177

Measures of Agreement

- Have N observations
- Consider all pairs of distinct responses
  - In this example $t = 16 \times 14 = 224$
- Compare predicted probabilities
  - Concordant if $\hat{\pi}_{Y=1} > \hat{\pi}_{Y=0}$
  - Discordant if $\hat{\pi}_{Y=1} < \hat{\pi}_{Y=0}$
  - Tie if $\hat{\pi}_{Y=1} = \hat{\pi}_{Y=0}$
- Measures of agreement
  - Somers' D : (#C - #D)/t
  - Goodman-Kruskal Gamma : (#C - #D)/(#C + #D)
  - Kendall's Tau-a : (#C - #D)/(.5N(N-1))
  - c : (#C +.5(t-#C-#D))/t

Background Reading

- KNNL Chapter 14
- knnl555.sas
- KNNL Chapter 14