Topic 31 - Multiple Logistic Regression

STAT 525 - Fall 2013

Outline

- Multiple Logistic Regression
  - Model
  - Inference
  - Diagnostics and remedies
- Polytomous Logistic Regression
  - Ordinal
  - Nominal

Multiple Logistic Regression

- Easy extension to multiple predictors using matrix notation
- Same diagnostics used for simple logistic regression
- Likelihood ratio/deviance test to look at collections of predictors
- Similar model building strategies
  - Stepwise
  - Forward
  - Backward
  - Score (best)

Example Page 573

- Want to understand epidemic outbreak of a disease spread by mosquitoes
- Randomly sampled individuals within two sectors of city
- Assessed whether individual had symptoms of disease and obtained other info
  - X_{i1} is age
  - X_{i2} is socioeconomic status
  - X_{i3} is the sector
  - Y_i is whether they had symptoms

SAS Commands

```
data a3;
   infile 'u:\.www\datasets525\APPENC10.txt';
   input case age socioecon sector Y savings;
   if case < 99; drop savings;


proc logistic data=a3 descending;
 class sector socioecon;
 model Y = age sector socioecon;


proc logistic data=a3 descending;
 class sector;
 model Y = age sector;
run;
```


Output - Full Model

Response Profile

Ordered Value	Y	Total Frequency
1	1	31
2	0	67

Probability modeled is Y=1.

Class Level Information

Class	Value	Design Variables	
sector	1	1	
	2	-1	
socioecon	1	1	0
	2	0	1
	3	-1	-1

***Notice the different choice
***of design vectors in Proc
***logistic compared to Proc GLM

Output - Full Model

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	124.318	111.054
SC	126.903	123.979
-2 Log L	122.318	101.054

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	21.2635	4	0.0003
Score	20.4067	4	0.0004
Wald	16.6437	4	0.0023

Output - Full Model

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
age	1	4.8535	0.0276
sector	1	9.8543	0.0017
socioecon	2	1.2053	0.5474

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.4909	0.4411	11.4215	0.0007
age	1	0.0297	0.0135	4.8535	0.0276
sector 1	1	-0.7873	0.2508	9.8543	0.0017
socioecon 1	1	-0.0345	0.3367	0.0105	0.9183
socioecon 2	1	0.3742	0.3662	1.0439	0.3069

Output - Reduced Model

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	124.318	108.259
SC	126.903	116.014
-2 Log L	122.318	102.259

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	20.0583	2	<.0001
Score	19.5250	2	<.0001
Wald	16.1851	2	0.0003

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
age	1	4.9455	0.0262
sector	1	11.7906	0.0006

Output - Reduced Model

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.4984	0.4342	11.9102	0.0006
age		1	0.0293	0.0132	4.9455	0.0262
sector	1	1	-0.8367	0.2437	11.7906	0.0006

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
age	1.030	1.003	1.057
sector 1 vs 2	0.188	0.072	0.488

Association of Predicted Probabilities and Observed Responses

Percent Concordant	77.9	Somers' D	0.562
Percent Discordant	21.6	Gamma	0.565
Percent Tied	0.5	Tau-a	0.246
Pairs	2077	c	0.781

GENMOD Commands

```
proc genmod data=a3 descending;
 class sector socioecon;
 model Y = age socioecon sector / link=logit noscale dist=bin;
 contrast 'age' age 1;
 contrast 'sector' sector 1 -1;
 contrast 'socioecon' socioecon 1 -1 0 , socioecon 1 0 -1;
run;


proc genmod data=a3 descending;
 class sector socioecon;
 model Y = age socioecon sector / link=logit noscale dist=bin aggregate;
 contrast 'age' age 1;
 contrast 'sector' sector 1 -1;
 contrast 'socioecon' socioecon 1 -1 0 , socioecon 1 0 -1;
run;
```


Output - No Aggregate

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Log Likelihood		-50.5271	
Full Log Likelihood		-50.5271	
AIC (smaller is better)		111.0542	
AICC (smaller is better)		111.7063	
BIC (smaller is better)		123.9790	

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-1.0434	0.6524	-2.3221	0.2352	2.56	0.1097
age	1	0.0298	0.0135	0.0033	0.0562	4.85	0.0276
socioecon 1	1	0.3053	0.6041	-0.8788	1.4893	0.26	0.6134
socioecon 2	1	0.7140	0.6537	-0.5672	1.9953	1.19	0.2747
sector 1	1	-1.5747	0.5016	-2.5579	-0.5916	9.86	0.0017

Contrast Results

Contrast	DF	Chi-Square	Pr > ChiSq	Type
age	1	5.15	0.0233	LR
sector	1	10.45	0.0012	LR
socioecon	2	1.21	0.5474	LR

Output - Aggregate

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	80	94.4625	1.1808
Scaled Deviance	80	94.4625	1.1808
Pearson Chi-Square	80	82.6652	1.0333
Scaled Pearson X2	80	82.6652	1.0333
Log Likelihood		-50.5271	
Full Log Likelihood		-48.7353	
AIC (smaller is better)		107.4706	
AICC (smaller is better)		108.1228	
BIC (smaller is better)		120.3955	

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.0434	0.6524	-2.3221	0.2352	2.56	0.1097
age	1	0.0298	0.0135	0.0033	0.0562	4.85	0.0276
socioecon 1	1	0.3053	0.6041	-0.8788	1.4893	0.26	0.6134
socioecon 2	1	0.7140	0.6537	-0.5672	1.9953	1.19	0.2747
sector	1	-1.5747	0.5016	-2.5579	-0.5916	9.86	0.0017

Contrast Results

Contrast	DF	Chi-Square	Pr > ChiSq	Type
age	1	5.15	0.0233	LR
sector	1	10.45	0.0012	LR
socioecon	2	1.21	0.5474	LR

Hypothesis Testing

- Can use deviance to compare models
- Models must be hierarchical (Full/Reduced)

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

- Partial deviance approx χ^2 with $p - q$ df
- Must compute by hand or with GENMOD
- For example: Testing H_O: Socioeconomic=0

```
DEV(Socioecon|Age,Sector) = 102.259-101.054 = 1.205
                  Pvalue = 0.5474
```


GLIMMIX Commands

```
proc glimmix data=a3;
 class sector socioecon;
 model Y(descending) = age socioecon sector / chisq link=logit dist=bin;
 contrast 'age' age 1;
 contrast 'sector' sector 1 -1;
 contrast 'socioecon' socioecon 1 -1 0, socioecon -1 0 1;
run;
```


Output

Fit Statistics

-2 Log Likelihood	101.05
AIC (smaller is better)	111.05
AICC (smaller is better)	111.71
BIC (smaller is better)	123.98
CAIC (smaller is better)	128.98
HQIC (smaller is better)	116.28
Pearson Chi-Square	92.24
Pearson Chi-Square / DF	0.99

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	Chi-Square	F Value	Pr > ChiSq	Pr > F
age	1	93	4.85	4.85	0.0276	0.0300
socioecon	2	93	1.21	0.60	0.5472	0.5493
sector	1	93	9.86	9.86	0.0017	0.0023

Contrasts

Label	Num DF	Den DF	F Value	Pr > F
age	1	93	4.85	0.0300
sector	1	93	9.86	0.0023
socioecon	2	93	0.60	0.5493

Logistic Residuals

- Distribution of residuals under correct model is unknown and thus common residual plot uninformative.
- Pearson residual is the ordinary residual divided by the standard error of Y_i. Sum of squared residuals equals Pearson X^2

$$r_{P_i} = \frac{Y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}$$

- Studentized Pearson residual is similar but divided by its standard error (includes hat matrix diagonal element) so they have unit variance

- Deviance residual is the signed square root of the observation's contribution to the model deviance

$$\begin{aligned} \mathrm{DEV}(\mathrm{X}_0, ..., \mathrm{X}_{\mathrm{p}-1}) &= -2\sum\left(Y_i log\left(\frac{\hat{\pi}_i}{Y_i}\right) + (1-Y_i) log\left(\frac{1-\hat{\pi}_i}{1-Y_i}\right)\right) \\ &= -2\sum\left(Y_i log\left(\hat{\pi}_i\right) + (1-Y_i) log\left(1-\hat{\pi}_i\right)\right) \end{aligned}$$

- Sign depends on $Y_i - \hat{\pi}_i$

$$\sum (dev_i)^2 = \mathrm{DEV}(\mathrm{X}_0, ..., \mathrm{X}_{\mathrm{p}-1})$$

Diagnostics

- "Residual" analysis
  - Can plot residual by predicted value : A flat lowess smooth to this plot suggests the model is correct
  - Can generate half-normal probability plot with simulated envelope to examine the linearity and identifying outliers
    - kth ordered absolute residual plotted against $z\left(\frac{k+n-1/8}{2n+1/2}\right)$
    - Outliers appear at the top right separated from others
    - Simulated envelope created by simulating data using $\hat{pi}_i$
    - Deviations from the mean of many simulations suggest model misfit
- DFFITS, DFBETAS
- SAS contains influence and iplots options

```
proc logistic data=a1 descending;
   model renew = increase / iplots influence lackfit
                 clparm=both clodds=both;
run;
```


Output

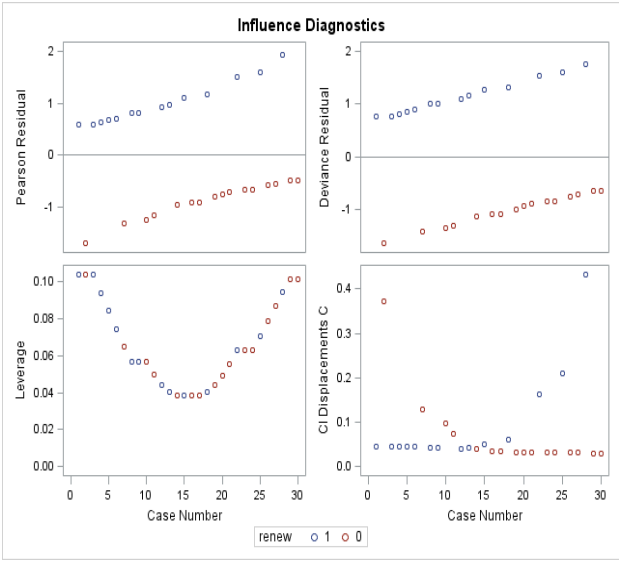
Regression Diagnostics

Case Number	Covariates increase	Pearson Residual	Deviance Residual	Hat Matrix Diagonal	Intercept DfBeta	increase DfBeta	Confidence Interval Displacement C	Confidence Interval Displacement CBar
1	30.0000	0.5900	0.7729	0.1040	0.1945	-0.1798	0.0451	0.0404
2	30.0000	-1.6948	-1.6455	0.1040	-0.5587	0.5165	0.3720	0.3334
3	30.0000	0.5900	0.7729	0.1040	0.1945	-0.1798	0.0451	0.0404
4	31.0000	0.6281	0.8155	0.0941	0.1902	-0.1740	0.0453	0.0410
5	32.0000	0.6686	0.8597	0.0841	0.1831	-0.1654	0.0448	0.0411
6	33.0000	0.7118	0.9054	0.0743	0.1732	-0.1539	0.0440	0.0407
7	34.0000	-1.3197	-1.4203	0.0651	-0.2792	0.2425	0.1297	0.1213
8	35.0000	0.8066	1.0012	0.0568	0.1441	-0.1213	0.0415	0.0392
9	35.0000	0.8066	1.0012	0.0568	0.1441	-0.1213	0.0415	0.0392
10	35.0000	-1.2397	-1.3645	0.0568	-0.2215	0.1864	0.0981	0.0925
11	36.0000	-1.1646	-1.3092	0.0497	-0.1689	0.1353	0.0746	0.0709
12	37.0000	0.9141	1.1021	0.0442	0.1013	-0.0746	0.0404	0.0386
13	38.0000	0.9731	1.1543	0.0404	0.0744	-0.0456	0.0415	0.0398
26	47.0000	-0.5853	-0.7676	0.0788	0.1240	-0.1415	0.0318	0.0293
27	48.0000	-0.5498	-0.7268	0.0869	0.1306	-0.1469	0.0315	0.0288
28	49.0000	1.9361	1.7651	0.0946	-0.5053	0.5619	0.4324	0.3915
29	50.0000	-0.4852	-0.6502	0.1017	0.1370	-0.1509	0.0297	0.0267
30	50.0000	-0.4852	-0.6502	0.1017	0.1370	-0.1509	0.0297	0.0267

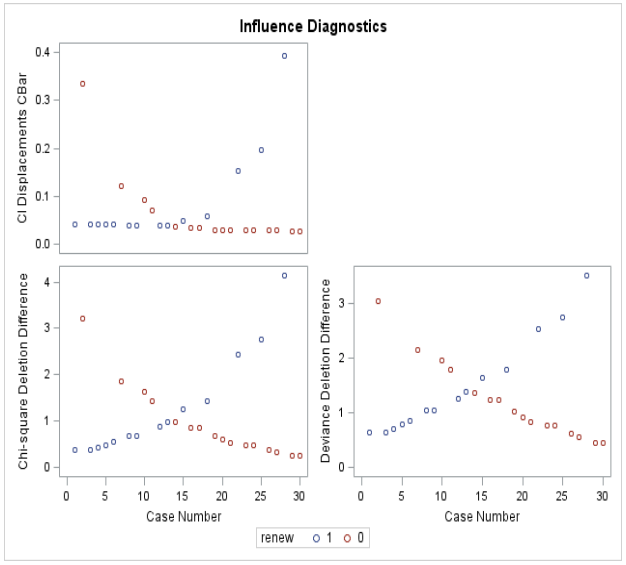
Regression Diagnostics

Case Number	Delta Deviance	Delta Chi-Square
1	0.6378	0.3885
2	3.0411	3.2058
3	0.6378	0.3885
4	0.7061	0.4355
5	0.7802	0.4882
6	0.8605	0.5473
7	2.1384	1.8630
8	1.0415	0.6898
9	1.0415	0.6898
10	1.9544	1.6294
11	1.7850	1.4271
12	1.2533	0.8742
13	1.3723	0.9868
26	0.6185	0.3719
27	0.5571	0.3311
28	3.5071	4.1399
29	0.4495	0.2621
30	0.4495	0.2621

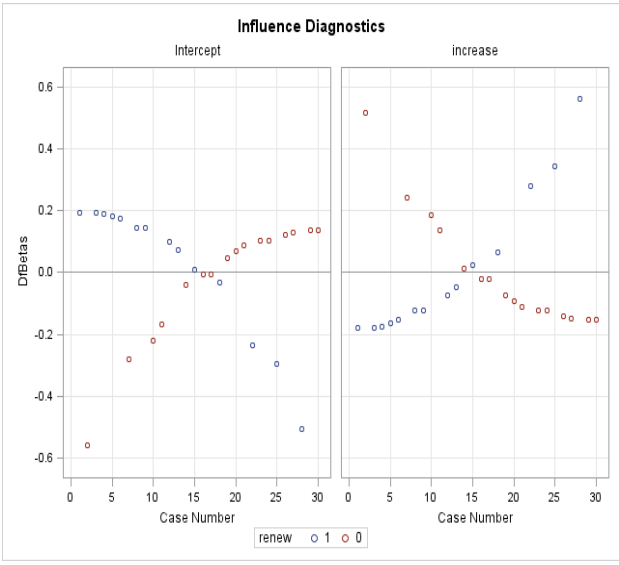
Influence/Residual plots



Influence/Residual plots



Influence/Residual plots



Ordinal Logistic Regression

- Have more than two possible outcomes on ordered scale (e.g., rating scale like Likert 1-5)
- Ordinal often called **proportional odds** model

$$\log\left(\frac{\Pr(\mathrm{Y} \leq \mathrm{j})}{1-\Pr(\mathrm{Y} \leq \mathrm{j})}\right) = \beta_{0j} + \beta_1 X_i$$

- Results in constant β_1 value

Example : SAS Commands

```
DATA;
   INPUT dose symptoms $ n @@;
   ldose = LOG10(dose);
CARDS;
10      None       33  10      Mild        7
10      Severe     10  20      None       17
20      Mild       13  20      Severe     17
30      None       14  30      Mild        3
30      Severe     28  40      None        9
40      Mild        8  40      Severe     32
;

proc logistic order = data;
   class symptoms;
   model symptoms = ldose;
   freq n;
   output out=b2 predprobs=i;
run;
```


Output

```
                Response Profile
 Ordered                      Total           Total
   Value     symptoms     Frequency          Weight
       1     None                 4       73.000000
       2     Mild                 4       31.000000
       3     Severe               4       87.000000
Probs modeled are cumulated over the lower Ordered Values.

Score Test for the Proportional Odds Assumption
Chi-Square       DF     Pr > ChiSq
    0.1674        1         0.6825

         Model Fit Statistics
               Intercept    Intercept &
Criterion        Only        Covariates
AIC              393.986        364.658
SC               394.956        366.113
-2 Log L         389.986        358.658
```


Output

```
       Testing Global Null Hypothesis: BETA=0
Test                 Chi-Square       DF     Pr > ChiSq
Likelihood Ratio        31.3281        1         <.0001
Score                   29.6576        1         <.0001
Wald                    28.5177        1         <.0001

         Analysis of Maximum Likelihood Estimates
                              Standard        Wald
Parameter       DF Estimate   Error  Chi-Square  Pr > ChiSq
Intercept None   1   4.1734  0.8862     22.1759      <.0001
Intercept Mild   1   4.9372  0.9083     29.5473      <.0001
ldose            1  -3.5207  0.6593     28.5177      <.0001

           Odds Ratio Estimates
             Point          95% Wald
Effect    Estimate      Confidence Limits
ldose        0.030       0.008       0.108
```

Output- Predicted Probs

Obs	IP_None2	IP_Mild2	IP_Severe2
1	0.65761	0.14717	0.19522
2	0.65761	0.14717	0.19522
3	0.65761	0.14717	0.19522
4	0.39958	0.18863	0.41179
5	0.39958	0.18863	0.41179
6	0.39958	0.18863	0.41179
7	0.26363	0.17090	0.56547
8	0.26363	0.17090	0.56547
9	0.26363	0.17090	0.56547
10	0.18739	0.14370	0.66891
11	0.18739	0.14370	0.66891
12	0.18739	0.14370	0.66891

Nominal Logistic Regression

- For polytomous or multicategory, consider there are J response categories
- Select one as baseline or reference category (call it J)

$$\log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = X_i'\beta_{jJ}$$

- Results in $J-1$ parameter vectors

SAS Commands

```
proc logistic data=b1 order=data;
   freq n;
   class symptoms(ref=last);
   model symptoms = ldose / link=glogit;
   output out=b1a predprobs=I;
run;

proc catmod data=b1 order=data;
   direct ldose;
   weight n;
   model symptoms = ldose / prob itprint;
run;
```


Output

Response Profile

Ordered Value	symptoms	Total Frequency
1	None	73
2	Mild	31
3	Severe	87

Logits modeled use symptoms='Severe' as the reference category.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	393.986	366.891
SC	400.491	379.900
-2 Log L	389.986	358.891

Output

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	31.0948	2	<.0001
Score	29.6579	2	<.0001
Wald	26.4676	2	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
ldose	2	26.4676	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	symptoms	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	None	1	5.3897	1.0998	24.0165	<.0001
Intercept	Mild	1	2.3076	1.3681	2.8448	0.0917
ldose	None	1	-4.1488	0.8066	26.4600	<.0001
ldose	Mild	1	-2.4128	0.9902	5.9373	0.0148

Output

Odds Ratio Estimates

Effect	symptoms	Point Estimate	95% Wald Confidence Limits	
ldose	None	0.016	0.003	0.077
ldose	Mild	0.090	0.013	0.624

Obs	ldose	IP_None	IP_Mild	IP_Severe
1	1.00000	0.64541	0.16798	0.18661
4	1.30103	0.40866	0.17937	0.41197
7	1.47712	0.27108	0.16153	0.56740
10	1.60206	0.19029	0.14086	0.66885

Overdispersion

- Binomial/Bernoulli Model

  $\mathrm{E}(\mathrm{Y_i}) = \pi_\mathrm{i}$

  $\mathrm{Var}(\mathrm{Y_i}) = \pi_\mathrm{i}(1-\pi_\mathrm{i})$

- Beta-Binomial Model

  $\mathrm{E}(\mathrm{Y_i}) = \pi_\mathrm{i}$

  $\mathrm{Var}(\mathrm{Y_i}) = \sigma^2\pi_\mathrm{i}(1-\pi_\mathrm{i})$

- Mean unaffected but $\mathrm{Cov}(\hat{\beta}) \approx \sigma^2(X'WX)^{-1}$
- Can show $E(\chi^2/(N-p)) \approx \sigma^2$

SAS Commands

```
options nocenter;
goptions colors=('none');

data a1;
 infile 'u:\.www\datasets525\CH14PR06.DAT';
 input norenew increase;
 renew=1-norenew;

proc genmod data=a1 descending;
 model renew = increase / dist=binomial noscale link=logit;

proc genmod data=a1 descending;
 model renew = increase / dist=binomial scale=2 link=logit;
run;
```

Output

```
                Analysis Of Parameter Estimates
                        Standard      Chi-
Parameter  DF  Estimate    Error    Square  Pr > ChiSq
Intercept   1    4.8075   2.6558      3.28      0.0703
increase    1   -0.1251   0.0668      3.51      0.0610
Scale       0    1.0000   0.0000
-----------------------------------------------------

                Analysis Of Parameter Estimates
                        Standard    Chi-
Parameter  DF  Estimate    Error  Square  Pr > ChiSq
Intercept   1    4.8075   5.3115   0.82       0.3654
increase    1   -0.1251   0.1335   0.88       0.3489
Scale       0    2.0000   0.0000
```


SAS Commands

```
data ingots;
 input heat soak r n @@;
 cards;
 7 1.0 0 10  14 1.0 0 31  27 1.0 1 56  51 1.0 3 13
 7 1.7 0 17  14 1.7 0 43  27 1.7 4 44  51 1.7 0  1
 7 2.2 0  7  14 2.2 2 33  27 2.2 0 21  51 2.2 0  1
 7 2.8 0 12  14 2.8 0 31  27 2.8 1 22  51 4.0 0  1
 7 4.0 0  9  14 4.0 0 19  27 4.0 1 16
;

proc genmod data = ingots;
 model r/n = heat soak / dist=binomial link=logit scale=p;

proc genmod data = ingots;
 model r/n = heat soak / dist=binomial link=logit scale=d;
run;
```


Output

```
          Criteria For Assessing Goodness Of Fit
Criterion              DF         Value    Value/DF
Deviance               16       13.7526      0.8595
Scaled Deviance        16       16.2476      1.0155
Pearson Chi-Square     16       13.5431      0.8464
Scaled Pearson X2      16       16.0000      1.0000
Log Likelihood                 -56.3214

                  Analysis Of Parameter Estimates
                      Standard      Chi-
Parameter  DF  Estimate     Error    Square  Pr > ChiSq
Intercept   1   -5.5592    1.0301     29.12      <.0001
heat        1    0.0820    0.0218     14.11      0.0002
soak        1    0.0568    0.3047      0.03      0.8522
Scale       0    0.9200    0.0000
NOTE: The scale parameter was estimated by the square root
      of Pearson's Chi-Square/DOF.
```


Background Reading

- KNNL Chapter 14
- knnl570.sas
- KNNL Chapter 14