Chapter 6: Building, Checking & Applying Logistic Regression

Dipankar Bandyopadhyay

Department of Biostatistics,
Virginia Commonwealth University

BIOS 625: Categorical Data & GLM

[Acknowledgements to Tim Hanson and Haitao Chu]

- Two competing goals:
  - Model should fit the data well.
  - Model should be simple to interpret (smooth rather than overfit – principle of parsimony).
- Often hypotheses on how the outcome is related to specific predictors will help guide the model building process.
- As a rule of thumb: at least 10 events and 10 non-events should occur for each predictor in the model (including dummies), see Peduzzi et al. 1996. So if $\sum_{i=1}^{N} y_i = 40$ and $\sum_{i=1}^{N} n_i = 830$, you should have no more than $40/10 = 4$ predictors in the model.

- Impacts of over fitting: Severe biased parameter estimates, poor standard error estimates, and error rates from Wald tests and confidence intervals far from the nominal level.
- Certain strategies such as penalized likelihood methods that can shrink many estimates to 0, and it is possible to have many predictors.
- You should not use the guideline to justify overly ambitious. If you have 1000 outcomes of each type, you are not usually well served by a model with 100 predictors.

6.1.2 Horseshoe crab data

- Recall that in all models fit we strongly rejected H_0 : logit $\pi(\mathbf{x}) = \beta_0$ in favor of H_1 : logit $\pi(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$:

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	40.5565	7	<.0001
Score	36.3068	7	<.0001
Wald	29.4763	7	0.0001

- However, it was not until we carved superfluous predictors from the model that we showed significance for the included model effects.
- This is an indication that several covariates may be highly related, or correlated. If one or more predictors are perfectly predicted as a linear combination of other predictors the model is overspecified and unidentifiable. Here's an example:

$$\text{logit } \pi(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 - 3x_2).$$

- The MLE $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ is not unique and the model is said to be unidentifiable. The variable $x_1 - 3x_2$ is totally predicted and redundant given x_1 and x_2.
- Although a perfect linear relationship is usually not met in practice, often variables are *highly* correlated and therefore one or more are redundant. We need to get rid of some!
- Although not ideal, automated model selection is necessary with large numbers of predictors. With $p - 1 = 10$ predictors, there are $2^{10} = 1024$ possible models; with $p - 1 = 20$ there are $1,048,576$ to consider.
- Backwards elimination starts with a large pool of potential predictors and step-by-step eliminates those with (Wald) p-values larger than a cutoff (the default is 0.05 in SAS PROC LOGISTIC).

Backward Elimination

```
proc logistic data=crabs1 descending;
    class color spine / param=ref;
    model y = color spine width weight color*spine color*width color*weight
    spine*width spine*weight width*weight / selection =backward;
    run;
```

When starting from all main effects and two-way interactions, the default *p*-value cutoff 0.05 yields only the model with width as a predictor

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi−Square	Pr > ChiSq
1	color*spine	6	9	0.0837	1.0000
2	width*color	3	8	0.8594	0.8352
3	width*spine	2	7	1.4906	0.4746
4	weight*spine	2	6	3.7334	0.1546
5	spine	2	5	2.0716	0.3549
6	width*weight	1	4	2.2391	0.1346
7	weight*color	3	3	5.3070	0.1507
8	weight	1	2	1.2263	0.2681
9	color	3	1	6.6246	0.0849

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi−Square	Pr > ChiSq
Intercept	1	−12.3508	2.6287	22.0749	<.0001
width	1	0.4972	0.1017	23.8872	<.0001

Change criteria of removal

Let's change the criteria for removing a predictor to p-value ≥ 0.15.

```
model y = color spine width weight color*spine color*width color*weight
  spine*width spine*weight width*weight / selection =backward slstay=0.15;
```

Yielding a more complicated model:

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	color*spine	6	9	0.0837	1.0000
2	width*color	3	8	0.8594	0.8352
3	width*spine	2	7	1.4906	0.4746
4	weight*spine	2	6	3.7334	0.1546
5	spine	2	5	2.0716	0.3549

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	13.8781	14.2883	0.9434	0.3314
color	1	1	1.3633	5.9645	0.0522	0.8192
color	2	1	−0.6736	2.6036	0.0669	0.7958
color	3	1	−7.4329	3.4968	4.5184	0.0335
width		1	−0.4942	0.5546	0.7941	0.3729
weight		1	−10.1908	6.4828	2.4711	0.1160
weight*color	1	1	0.1633	2.3813	0.0047	0.9453
weight*color	2	1	0.9425	1.1573	0.6632	0.4154
weight*color	3	1	3.9283	1.6151	5.9155	0.0150
width*weight		1	0.3597	0.2404	2.2391	0.1346

Drop width and width*weight?

Let's test if we can simultaneously drop width and width*weight from this model. From the (voluminous) output we find:

Criterion	Intercept Only	Intercept and Covariates
AIC	227.759	196.841
SC	230.912	228.374
−2 Log L	225.759	176.841

Fitting the simpler model with color, weight, and color*weight yields

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	227.759	197.656
SC	230.912	222.883
−2 Log L	225.759	181.656

There are 2 more parameters in the larger model (for width and width*weight) and we obtain $-2(L_0 - L_1) = 181.7 - 176.8 = 4.9$ and $P(\chi^2_2 > 4.9) = 0.07$. We barely accept that we can drop width and width*weight at the 5% level.

Forward Selection

- Forward selection starts by fitting each model with one predictor separately and including the model with the smallest p-value under a cutoff (default=0.05 in PROC LOGISTIC). When we instead have SELECTION=FORWARD in the MODEL statement we obtain the model with only width. Changing the cutoff to SLENTRY=0.15 gives the model with width and color.
- Starting from main effects and working backwards by hand, we ended up with width and color in the model. We further simplified color to dark and non dark crabs. Using backwards elimination with a cutoff of 0.05 we ended up with just width. A cutoff of 0.15 and another "by hand" step (at the 0.05 level) yielded weight, color, and weight*color.
- The book considers backwards elimination starting with a three-way interaction model including color, spine condition, and width. The end model is color and width.

Stepwise selection in SAS

- PROC LOGISTIC allows backwards elimination, forwards selection, and something that does both, termed 'stepwise.'
- Stepwise selection checks to see whether one or more effects can be removed from the model after adding a term. Stepwise goes back and forth adding and removing terms until no more can be eliminated at the SLSTAY level and no more can be added at the SLENTRY level. In my opinion, this is the best of the three approaches to variable selection.
- Hierarchical models have interactions and/or quadratic effects only when the main effects comprising them are also in the model (more on this shortly). SAS automatically chooses the default HIERARCHY=SINGLE to force a hierarchical final model. There are other options, e.g. HIER=MULTIPLE or HIER=NONE.

Stepwise selection in SAS

- Recall that default values for SLENTRY and SLSTAY are 0.05. You will get models with more predictors when you increase these.
- For default SLENTRY and SLSTAY, only width is picked using all three selection procedures for the crab data. For SLENTRY=SLSTAY=0.1, all three procedures give the same model: color and width.
- Treating color and spine as continuous also yields an additive model with color and width using all three approaches.

6.1.4 AIC: Minimizing Distance of the Fit from Truth

"No model is correct, but some are more useful than others."
— George Box

- It is often of interest to examine several competing models. In light of underlying biology or science, one or more models may have relevant interpretations within the context of why data were collected in the first place.
- In the absence of scientific input, a widely-used model selection tool is the Akaike information criterion (AIC),

$$\text{AIC} = -2[L(\hat{\boldsymbol{\beta}}; \mathbf{y}) - p].$$

- The $L(\hat{\boldsymbol{\beta}}; \mathbf{y})$ represents model fit. If you add a parameter to a model, $L(\hat{\boldsymbol{\beta}}; \mathbf{y})$ has to increase. If we only used $L(\hat{\boldsymbol{\beta}}; \mathbf{y})$ as a criterion, we'd keep adding predictors until we ran out. The p penalizes for the number of the predictors in the model.

AIC for Crab Data

The AIC has very nice properties in large samples in terms of prediction. The smaller the AIC is, the better the model fit (asymptotically).

Model	AIC
W	198.8
$C+Wt+C*Wt$	197.7
$C+W$	197.5
$D+Wt+D*Wt$	194.7
$D+W$	194.0

If we pick one model, it's $W+D$, the additive model with width and the dark/nondark category.

LASSO for logistic regression

SAS has a new procedure, PROC HPGENSELECT, which can implement the LASSO, a modern variable selection technique. It does not, as of yet, have a HIER=SINGLE option akin to PROC GLMSELECT, but probably will in a future version. SAS will perform forward selection with a very large number of variables in a more principled manner than traditional forward selection in PROC HPGENSELECT with the METHOD=LASSO option. It will star the model with the 'best' selection criterion that you ask for, below the AIC corrected for small sample sizes. Here we try to find a parsimonious model from all main effects and two-way interactions.

```
proc hpgenselect;
class color spine;
model y(event="1") = color spine width weight color*spine color*width
color*weight spine*width spine*weight weight*width / dist=binary link=logit;
selection method=lasso(choose=aicc) details=all;
run;
```

GOFs

GOF tests are global checks for model adequacy.
The data are $(\mathbf{x}_i, Y_i)$ for $i = 1, \ldots, N$. The ith *fitted value* is an estimate of $\mu_i = E(Y_i)$, namely $\widehat{E(Y_i)} = \hat{\mu}_i = n_i\hat{\pi}_i$ where $\pi_i = \frac{e^{\beta'\mathbf{x}_i}}{1+e^{\beta'\mathbf{x}_i}}$ and $\hat{\pi}_i = \frac{e^{\hat{\beta}'\mathbf{x}_i}}{1+e^{\hat{\beta}'\mathbf{x}_i}}$. The raw residual is what we see Y_i minus what we predict $n_i\hat{\pi}_i$. The Pearson residual divides this by an estimate of $\sqrt{\text{var}(Y_i)}$:

$$e_i = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}.$$

The Pearson GOF statistic is

$$X^2 = \sum_{i=1}^{N} e_i^2.$$

The standardized Pearson residual is given by

$$r_i = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)(1-\hat{h}_i)}},$$

where $\hat{h}_i$ is the ith diagonal element of the *hat* matrix $\hat{\mathbf{H}} = \hat{\mathbf{W}}^{1/2}\mathbf{X}(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{W}}^{1/2}$ where $\mathbf{X}$ is the design matrix

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{N,p-1} \end{bmatrix},$$

and

$$\hat{\mathbf{W}} = \begin{bmatrix} n_1\hat{\pi}_1(1-\hat{\pi}_1) & 0 & \cdots & 0 \\ 0 & n_2\hat{\pi}_2(1-\hat{\pi}_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_N\hat{\pi}_N(1-\hat{\pi}_N) \end{bmatrix}.$$

Alternatively, (6.2, p. 220) defines a deviance residual.

Comments

- With good replication (e.g. $n_i \geq 10$), plots of residuals r_j versus one of the $p-1$ predictors x_{ij}, for $j = 1, \ldots, N$ might show systematic lack of fit (i.e. a pattern). Adding nonlinear terms or interactions can improve fit.
- With truly continuous predictors $n_i = 1$ and the residual plots *will* have a distinct pattern. Use the fact that if the model fits, $E(r_i) \approx 0$, and superimpose a loess fit on top of the residuals. The loess line should be *approximately* straight.
- An overall plot is r_j versus the linear predictor $\hat{\eta}_j = \hat{\boldsymbol{\beta}}' \mathbf{x}_j$. This plot will tell you if the model tends to over or underpredict the observed data for ranges of the linear predictor.
- The r_i are approximately $N(0, 1)$ when n_i is not small.

- I usually flag $|r_i| > 3$ as being ill-fit by the model.
- You can look at individual r_i to determine model fit. For the crab data, this might flag some individual crabs as ill-fit or unusual relative to the model.
- The *model* can't tell the difference between, e.g., two nondark crabs with same carapace width 23 *cm*. You can aggregate over same values of the predictors to slightly "improve" the residuals. This way the approximate $N(0, 1)$ may be a bit better. Ill fitting residuals then suggest evidence where the aggregated number of events don't match what we'd expect under the model.

Let's look at $W + D$ for the crab data. We'll consider both width and width truncated to an integer *cm*. The DATA step is

```
data crabs1; set crabs; input color spine width satell weight;
   weight=weight/1000; color=color-1;
   y=0; n=1; if satell >0 then y=1;
   dark=1; if color=4 then dark=2;
   w=int(width);   * round down;
```

Two models fit & r_i plotted:

```
proc logistic data=crabs1 descending;
    * each crab has n_i=1;
    class dark; model y = dark width;
    output out=diag1 reschi=p h=h xbeta=eta;
data diag2; set diag1; r=p/sqrt(1-h);
proc gplot; plot r*width; plot r*dark; plot r*eta;
    * plot r_i vs width, dark, eta_i ;
proc sort data=crabs1; by w dark;
    * aggregate over coarser widths;
proc means data=crabs1 noprint; by w dark; var y n;
    output out=crabs2 sum=sumy sumn;
proc logistic data=crabs2;
    class dark; model sumy/sumn = dark w;
    output out=diag3 reschi=p h=h xbeta=eta;
data diag4; set diag3; r=p/sqrt(1-h);
proc gplot; plot r*w; plot r*dark; plot r*eta; run;
```

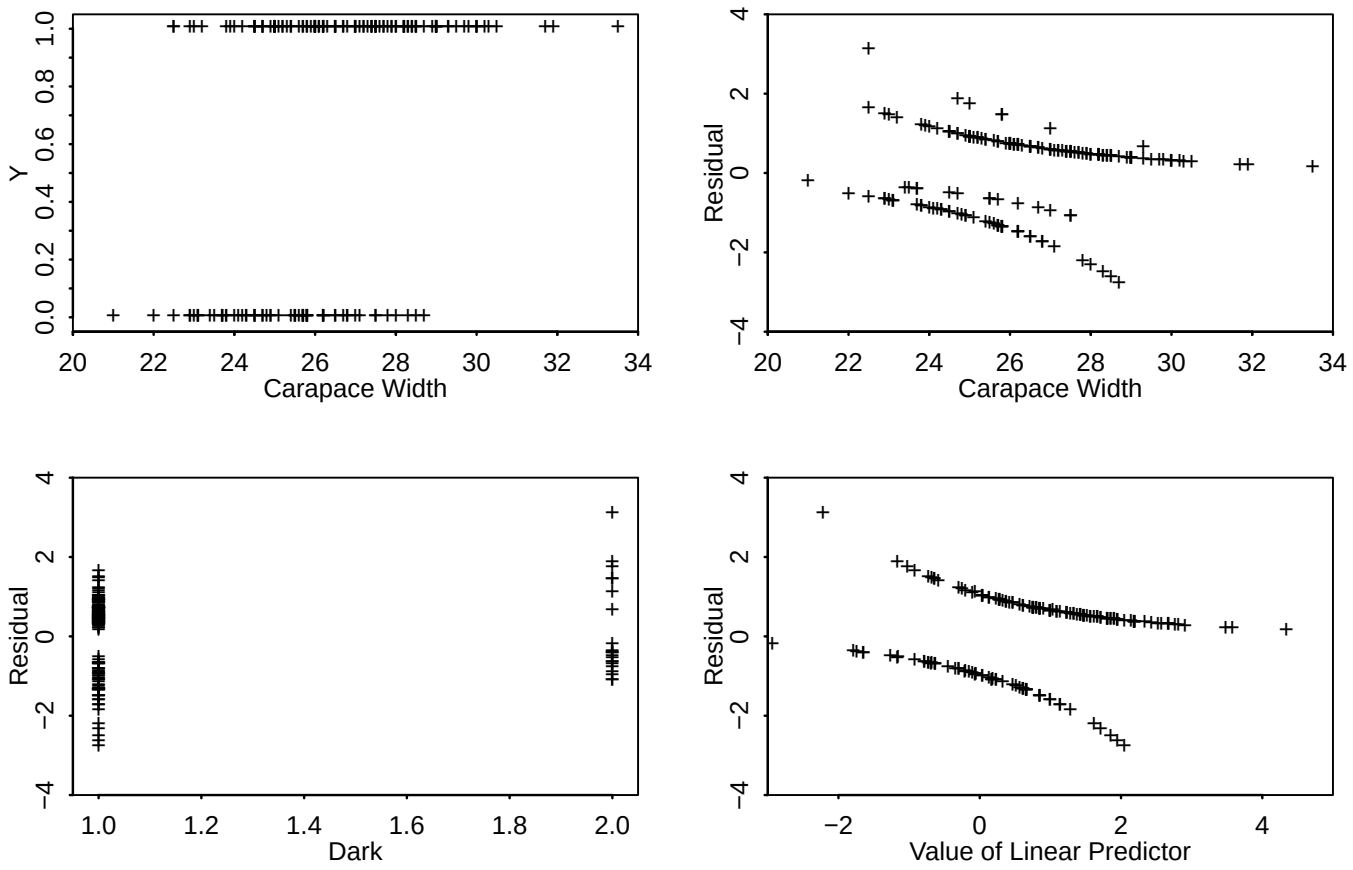

Figure : Raw Data & Residual Plots, $n_i = 1$

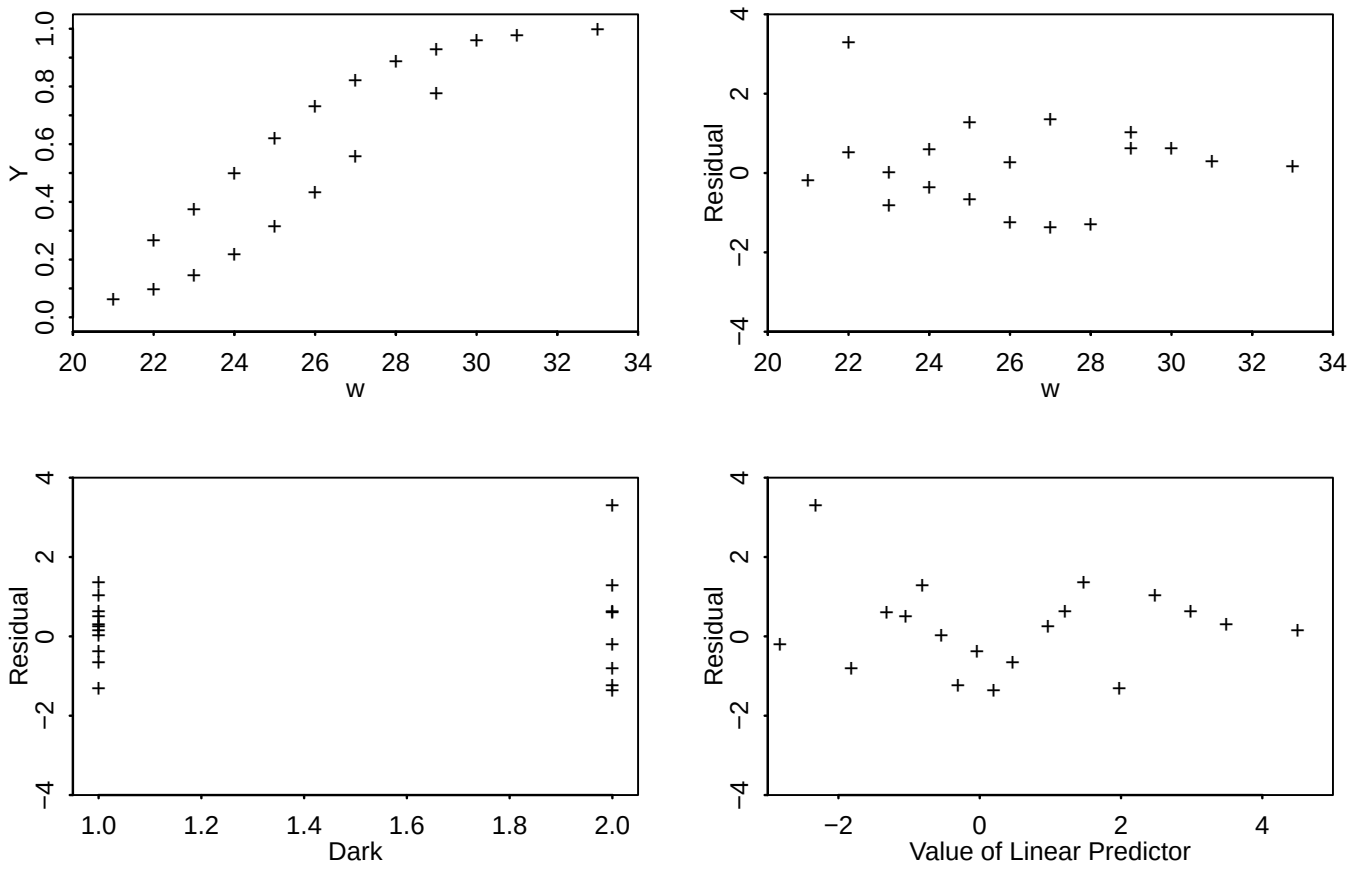

Figure : Raw Data & Residual Plots, Aggregated.

6.2.4 Influence

Unlike linear regression, the leverage $\hat{h}_i$ in logistic regression depends on the model fit $\hat{\boldsymbol{\beta}}$ as well as the covariates $\mathbf{X}$. Points that have extreme predictor values $\mathbf{x}_i$ may not have high leverage $\hat{h}_i$ if $\hat{\pi}_i$ is close to 0 or 1. Here are the influence diagnostics available in PROC LOGISTIC:

- Leverage $\hat{h}_i$. Still may be useful for detecting "extreme" predictor values $\mathbf{x}_i$.
- $c_i = e_i^2 \hat{h}_i/(1-\hat{h}_i)^2$ measures the change in the joint confidence region for $\boldsymbol{\beta}$ when i is left out.
- DFBETA$_{ij}$ is the standardized difference in $\hat{\beta}_j$ when observation i is left out.
- The change in the X^2 GOF statistic when obs. i is left out is DIFCHISQ$_i = e_i^2/(1-\hat{h}_i)$.

I suggest looking at plots of c_i vs. i, and possibly the DFBETA's versus i.

Obs	w	dark	sumy	sumn
1	21	2	0	1
2	22	1	2	6
3	22	2	1	1
4	23	1	4	11
5	23	2	0	4
6	24	1	9	20
7	24	2	1	3
8	25	1	15	27
9	25	2	3	6
10	26	1	20	27
11	26	2	0	2
12	27	1	20	22
13	27	2	1	4
14	28	1	15	19
15	29	1	10	10
16	29	2	1	1
17	30	1	6	6
18	31	1	2	2
19	33	1	1	1

Fitting a logistic regression for the aggregated data:

```
proc logistic data=crabs2;
class dark;
model sumy/sumn = dark w
/aggregate scale=none lackfit influence iplots ;
run;
```

Let's look output from the aggregated crab data:

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	17.3663	16	1.0854	0.3623
Pearson	20.1139	16	1.2571	0.2151

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	41.2125	2	<.0001
Score	36.6705	2	<.0001
Wald	29.0982	2	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
dark	1	5.7191	0.0168
w	1	23.2366	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-12.7834	2.6636	23.0329	<.0001
dark 1	1	0.6374	0.2665	5.7191	0.0168
w	1	0.5044	0.1046	23.2366	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
dark 1 vs 2	3.578	1.259	10.171
w	1.656	1.349	2.033

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
6.5710	6	0.3623

```
                                   Pearson Residual                          Deviance Residual
              Covariates
Case                                   (1 unit =  0.4)                           (1 unit = 0.27)
 #      dark1          w       Value   −8  −4  0 2 4 6 8        Value   −8  −4  0 2 4 6 8

 1    −1.0000     21.0000    −0.2431   |      *|        |     −0.3389   |      *|        |
 2     1.0000     22.0000     0.4131   |       |*       |      0.4021   |       |*       |
 3    −1.0000     22.0000     3.1960   |       |      *|      2.1987   |       |      *|
 4     1.0000     23.0000    −0.0239   |       *        |     −0.0240   |       *        |
 5    −1.0000     23.0000    −0.8053   |     * |        |     −1.0964   |   *   |        |
 6     1.0000     24.0000    −0.3574   |      *|        |     −0.3578   |      *|        |
 7    −1.0000     24.0000     0.5160   |       |*       |      0.4876   |       | *      |
 8     1.0000     25.0000    −0.6239   |     * |        |     −0.6189   |     * |        |
 9    −1.0000     25.0000     1.0202   |       |   *    |      0.9797   |       |    *   |
10     1.0000     26.0000     0.1850   |       *        |      0.1861   |       |*       |
11    −1.0000     26.0000    −1.2135   |    *  |        |     −1.4856   |  *    |        |
12     1.0000     27.0000     1.1509   |       |   *    |      1.2527   |       |     *  |
13    −1.0000     27.0000    −1.2035   |    *  |        |     −1.2174   |   *   |        |
14     1.0000     28.0000    −1.1861   |    *  |        |     −1.0905   |   *   |        |
15     1.0000     29.0000     0.9143   |       | *      |      1.2671   |       |     *  |
16    −1.0000     29.0000     0.5469   |       |*       |      0.7234   |       |   *    |
17     1.0000     30.0000     0.5503   |       |*       |      0.7687   |       |   *    |
18     1.0000     31.0000     0.2469   |       |*       |      0.3466   |       |*       |
19     1.0000     33.0000     0.1054   |       *        |      0.1487   |       |*       |
```

```
                 Hat Matrix Diagonal          Intercept
  Case                (1 unit = 0.02)           DfBeta     (1 unit = 0.07)
Number      Value     0 2 4 6 8  12  16          Value      -8  -4  0 2 4 6 8

     1     0.0237    | *                |       -0.0283    |         *        |
     2     0.1839    |         *        |        0.1929    |         |  *     |
     3     0.0298    |  *               |        0.3672    |         |    *   |
     4     0.2486    |             *    |       -0.0128    |         *        |
     5     0.1467    |       *          |       -0.1877    |     *   |        |
     6     0.2844    |               *  |       -0.1639    |      *  |        |
     7     0.1331    |       *          |        0.0770    |         |*       |
     8     0.2460    |            *     |       -0.0894    |        *|        |
     9     0.3171    |                 *|        0.1331    |         |  *     |
    10     0.2255    |           *      |       -0.0340    |         *        |
    11     0.1232    |      *           |        0.0245    |         *        |
    12     0.2244    |           *      |       -0.4487    |  *      |        |
    13     0.2755    |               *  |        0.2139    |         |  *     |
    14     0.2323    |            *     |        0.5935    |         |       *|
    15     0.1307    |       *          |       -0.3317    |    *    |        |
    16     0.0702    |    *             |       -0.0836    |        *|        |
    17     0.0754    |    *             |       -0.1490    |      *  |        |
    18     0.0222    | *                |       -0.0352    |         *        |
    19    0.00738    |*                 |      -0.00875    |         *        |
```

```
                dark1                                                  w
  Case        DfBeta        (1 unit =  0.1)          DfBeta        (1 unit = 0.08)
Number         Value       −8  −4  0 2 4 6 8          Value       −8  −4  0 2 4 6 8

     1        0.0263      |        *        |        0.0258      |        *        |
     2        0.0390      |        *        |       −0.1901      |      *  |       |
     3       −0.4330      |    *   |        |       −0.3253      |    *    |       |
     4      −0.00363      |        *        |        0.0125      |        *        |
     5        0.3020      |        |  *     |        0.1582      |        |  *     |
     6       −0.0788      |       *|        |        0.1569      |        |  *     |
     7       −0.1947      |      *  |       |       −0.0578      |       *|        |
     8       −0.1464      |       *|        |        0.0751      |        |*       |
     9       −0.7822      |*       |        |       −0.0551      |       *|        |
    10        0.0384      |        *        |        0.0381      |        *        |
    11        0.4486      |        |    *   |       −0.0697      |       *|        |
    12        0.1861      |        | *      |        0.4700      |        |     *  |
    13        0.7676      |        |       *|       −0.2920      |    *   |        |
    14       −0.1497      |      *  |       |       −0.6119      |*       |        |
    15        0.0598      |        |*       |        0.3396      |        |   *    |
    16       −0.1153      |       *|        |        0.0956      |        |*       |
    17        0.0208      |        *        |        0.1519      |        | *      |
    18       0.00400      |        *        |        0.0358      |        *        |
    19      0.000718      |        *        |       0.00887      |        *        |
```

```
      Confidence Interval Displacement C   Confidence Interval Displacement CBar

Case                      (1 unit = 0.05)                      (1 unit = 0.03)
Number        Value      0 2 4 6 8  12  16        Value      0 2 4 6 8  12  16

     1      0.00147     |*                  |     0.00144     |*                  |
     2       0.0471     | *                 |      0.0385     | *                 |
     3       0.3235     |        *          |      0.3139     |          *        |
     4     0.000252     |*                  |    0.000190     |*                  |
     5       0.1306     |    *              |      0.1115     |    *              |
     6       0.0710     | *                 |      0.0508     | *                 |
     7       0.0471     | *                 |      0.0409     | *                 |
     8       0.1685     |     *             |      0.1270     |     *             |
     9       0.7075     |                 * |      0.4832     |                 * |
    10       0.0129     |*                  |     0.00997     |*                  |
    11       0.2359     |      *            |      0.2069     |       *           |
    12       0.4940     |           *       |      0.3831     |            *      |
    13       0.7600     |                  *|      0.5507     |                  *|
    14       0.5544     |             *     |      0.4256     |             *     |
    15       0.1446     |    *              |      0.1257     |     *             |
    16       0.0243     | *                 |      0.0226     | *                 |
    17       0.0267     | *                 |      0.0247     | *                 |
    18      0.00142     |*                  |     0.00139     |*                  |
    19     0.000083     |*                  |    0.000083     |*                  |
```

```
                    Delta Deviance                                  Delta Chi-Square
Case                  (1 unit = 0.32)                                 (1 unit = 0.66)
Number        Value    0 2 4 6 8  12  16          Value      0 2 4 6 8  12  16

     1       0.1163   |*                   |       0.0606   |*                   |
     2       0.2001   | *                  |       0.2091   |*                   |
     3       5.1483   |                   *|      10.5284   |                   *|
     4     0.000764   |*                   |     0.000763   |*                   |
     5       1.3135   |     *              |       0.7600   | *                  |
     6       0.1788   | *                  |       0.1785   |*                   |
     7       0.2787   | *                  |       0.3071   |*                   |
     8       0.5101   |   *                |       0.5163   | *                  |
     9       1.4429   |     *              |       1.5239   |   *                |
    10       0.0446   |*                   |       0.0442   |*                   |
    11       2.4138   |          *         |       1.6794   |    *               |
    12       1.9524   |        *           |       1.7078   |    *               |
    13       2.0328   |        *           |       1.9990   |    *               |
    14       1.6148   |       *            |       1.8326   |    *               |
    15       1.7313   |       *            |       0.9616   | *                  |
    16       0.5459   |   *                |       0.3217   |*                   |
    17       0.6156   |   *                |       0.3276   |*                   |
    18       0.1215   |*                   |       0.0623   |*                   |
    19       0.0222   |*                   |       0.0112   |*                   |
```

- Obs. 3 has a large e_i (and larger r_i) and is flagged as ill-fit. Obs. 3 also has the largest DIFCHISQ. Obs. 3 is $n_3 = 1$ skinny ($22cm$) dark crab that had a satellite. Recall that the probability of having a satellite increases for light crabs and for wider crabs. This observation does not have much of an effect on $\hat{\beta}$ as measured by c_i and the DFBETAs, perhaps because it's only 1 crab.
- Obs. 9 and 13 have the largest c_i. Refining their influence, both 9 and 13 have the largest (in magnitude) DFBETAs for the dark dummy variable. However, with relatively small $|e_i|$, these observations are not ill-fit. Obs. 9 represents $y_9 = 3$ dark crabs out of $n_9 = 6$ that have satellites at width $25cm$. Obs. 13 is $y_{13} = 1$ out of $n_{13} = 4$ dark crabs at $27cm$. These affect the estimate of dark's regression coefficient (adjusting for width) more than the other observations, *but are fit well by the model*.

Let's revisit the output from the aggregated crab data:

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	17.3663	16	1.0854	0.3623
Pearson	20.1139	16	1.2571	0.2151

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	41.2125	2	<.0001
Score	36.6705	2	<.0001
Wald	29.0982	2	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
dark	1	5.7191	0.0168
w	1	23.2366	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-12.7834	2.6636	23.0329	<.0001
dark 1	1	0.6374	0.2665	5.7191	0.0168
w	1	0.5044	0.1046	23.2366	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
dark 1 vs 2	3.578	1.259	10.171
w	1.656	1.349	2.033

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
6.5710	6	0.3623

Fitting a logistic regression for the aggregated data:

```
data crabs2; set crabs2 ; sid = _n_; run;
proc logistic data=crabs2 descending;
 class dark; model sumy/sumn = dark w/aggregate scale=none lackfit;
 where sid <>3;
 run;
```

Let's look output from the aggregated crab data with observation 3 deleted:

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	12.2066	15	0.8138	0.6633
Pearson	10.3209	15	0.6881	0.7991

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	45.4814	2	<.0001
Score	39.9703	2	<.0001
Wald	30.9792	2	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
dark	1	7.3422	0.0067
w	1	24.6940	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-13.8155	2.7781	24.7308	<.0001
dark 1	1	0.7595	0.2803	7.3422	0.0067
w	1	0.5403	0.1087	24.6940	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
dark 1 vs 2	4.568	1.522	13.705
w	1.717	1.387	2.124

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
6.6918	6	0.3503

Fitting a logistic regression for the aggregated data:

```
data crabs2; set crabs2 ; sid = _n_; run;
proc logistic data=crabs2 descending;
 class dark; model sumy/sumn = dark w/aggregate scale=none lackfit;
 where sid <>9 and sid <>13;
 run;
```

Let's look output from the aggregated crab data with observations 9 and 13 deleted:

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	14.8556	14	1.0611	0.3881
Pearson	18.2483	14	1.3034	0.1957

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	40.5189	2	<.0001
Score	35.9213	2	<.0001
Wald	27.9728	2	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
dark	1	2.4644	0.1165
w	1	23.6339	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-13.4764	2.7614	23.8174	<.0001
dark 1	1	0.6144	0.3914	2.4644	0.1165
w	1	0.5327	0.1096	23.6339	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
dark 1 vs 2	3.417	0.737	15.849
w	1.703	1.374	2.112

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
5.5024	6	0.4812

Assessing a model's predictive ability

- Section 6.3.3: SAS will 'predict' each Bernoulli outcome, say $\tilde{y}_i$ based on a fit of the model *without* observation i with the CTABLE option. You can include the proportion of 'successes' in the population, say it's 30%, using PEVENT=0.3. The default for PEVENT is the proportion of successes in the data set.
- An observation will be classified as a success if $\tilde{\pi}_i > k$ where k is a cutoff, and $\tilde{\pi}_i$ is the predicted probability of success through the model leaving observation i out; use PPROB= k. If PPROB is omitted, SAS will pick a bunch of them and give the correct number of correctly predicted successes (true positives) and the number of correctly predicted failures (true negatives), as well as the sensitivity and specificity for each.
- Section 6.3.4: Sensitivity and specificity for different cutoffs k can be combined into a receiver operator characteristic (ROC) curve; the area under this curve is c. OUTROC=name in the MODEL statement and PLOTS in the PROC LOGISTIC statement gives an ROC curve and estimate of c.

- Every pair of observations with different outcomes, i.e. every pair (i_1, i_2) where $y_{i1} \neq y_{i2}$, is either concordant, discordant, or tied. Assume $y_{i1} = 1$ and $y_{i2} = 0$. This pair is concordant if $\hat{\pi_{i1}} \geq \hat{\pi_{i2}}$, discordant if $\hat{\pi_{i1}} \leq \hat{\pi_{i2}}$, and tied if $\hat{\pi_{i1}} = \hat{\pi_{i2}}$. Let C be the number of concordant pairs, D the number of discordant pairs, T the number of ties. The total number of pairs is $C + D + T$. Then, $\hat{\gamma} = (C - D)/(C + D)$ and Somer's D is $(C - D)/(C + D + T)$. $\hat{\gamma}$ does not penalize for ties.
- c is $(C + 0.5T)/(C + D + T)$: the probability that a randomly drawn 'success' will have a higher $\hat{\pi}$ than a randomly drawn 'failure', also called 'area underneath the ROC curve'. $c \approx 1$ indicates excellent discriminatory ability; $c \approx 0.5$ means you might as well flip a coin rather than use the model to predict success or failure.
- The probabilities $\hat{\pi}_i$ are different than the leave-one-out values $\tilde{\pi}_i$ used in the CTABLE option.

6.3.3 Classification Tables

- A Classification Table cross-classifies the binary response with a prediction of whether $y = 0$ or 1. The prediction for observation i is $\hat{y} = 1$ when $\hat{\pi}_i > \pi_0$ and $\hat{y} = 0$ when $\hat{\pi}_i \leq \pi_0$ for some cutoff π_0.
- The predictive power can summarized as $Sensitivity = P(\hat{y} = 1|y = 1)$ and $Specificity = P(\hat{y} = 0|y = 0)$.
- The proportion of correct classification is P(correct classification)$= P(\hat{y} = 1|y = 1) + P(\hat{y} = 0|y = 0)$.

6.3.4 ROC Curves

- A receiver operating characteristic (ROC) is a plot of sensitivity as a function of (1-specificity) for the possible π_0.
- The greater the area under the ROC curve (AUC), the better the prediction.
- The AUC is identical to the concordance index (Hanley and McNeil, 1982).
- A value of AUC $= 0.5$ means predictions are no better than random guessing, corresponding to a straight line connecting points (0, 0) and (1, 1).

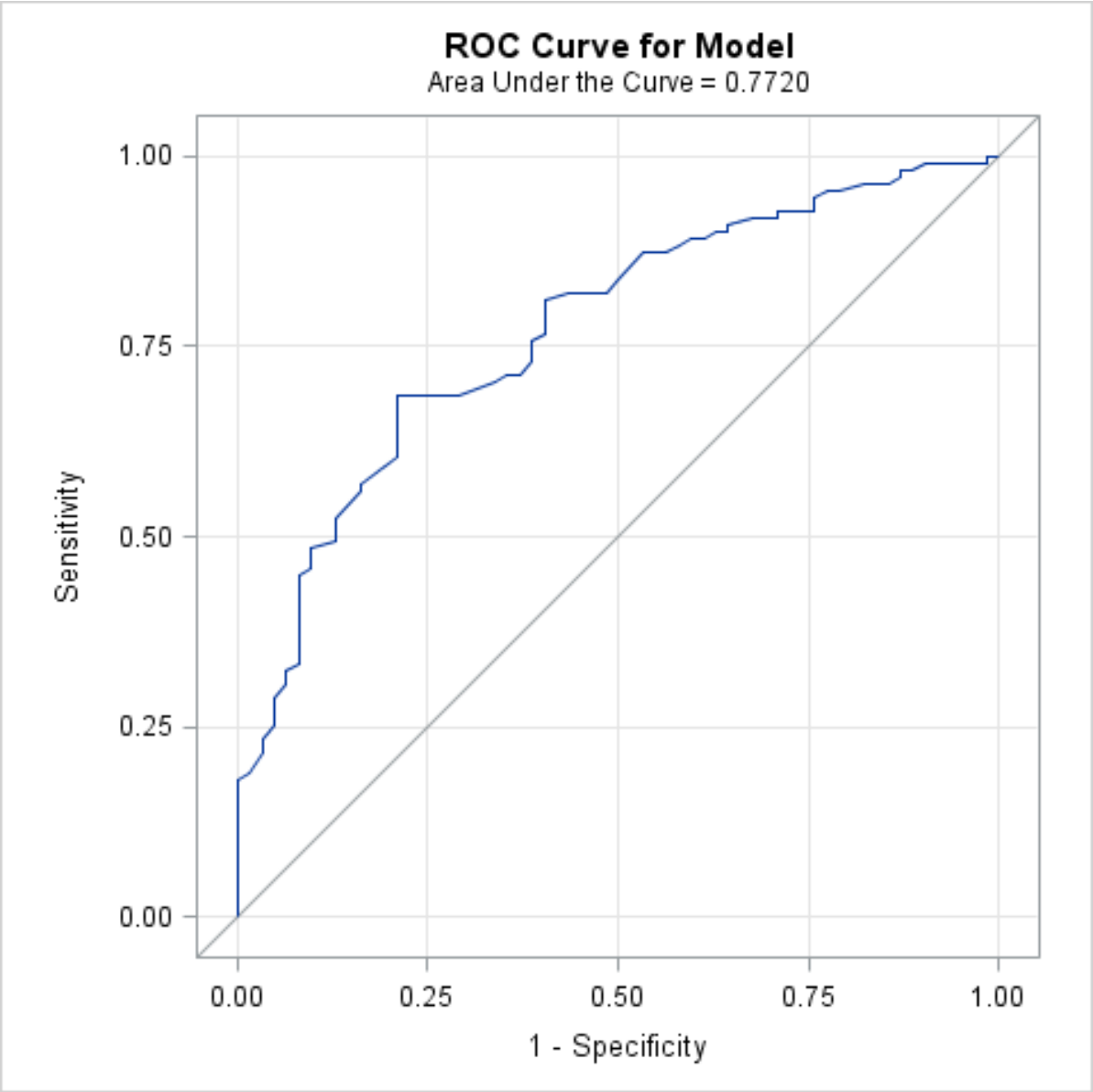

Figure : ROC curve

Let us look at a clinical trial with 8 centers, two creams compared to cure infection.

Center $Z = k$	Treatment X	Response Y Success	Response Y Failure	$\widehat{\theta}_{XY(k)}$
1	Drug	11	25	1.2
	Control	10	27	
2	Drug	16	4	1.8
	Control	22	10	
3	Drug	14	5	4.8
	Control	7	12	
4	Drug	2	14	2.3
	Control	1	16	
5	Drug	6	11	∞
	Control	0	12	
6	Drug	1	10	∞
	Control	0	10	
7	Drug	1	4	2.0
	Control	1	8	
8	Drug	4	2	0.3
	Control	6	1	

Have:

- Y binary outcome (e.g. success/failure of treatment).
- X binary predictor (e.g. treatment).
- Stratum Z (e.g. treatment center).

Want to test $X \perp Y|Z$ versus an alternative.

- Let $\pi_{ik} = P(Y = 1|X = i, Z = k)$ and

$$\theta_{XY(k)} = \frac{P(Y = 1|X = 1, Z = k)/P(Y = 2|X = 1, Z = k)}{P(Y = 1|X = 2, Z = k)/P(Y = 2|X = 2, Z = k)}.$$

  Recall $X \perp Y|Z$ when $\theta_{XY(k)} = 1$. This happens under the model

$$\text{logit } \pi_{ik} = \alpha + \beta_k^Z.$$

- This is an ANOVA-type specification where instead of listing $K - 1$ dummy variables, we concisely include a subscript on Z's effect β_k^Z. So there are K effects for Z, $\beta_1^Z, \beta_2^Z, \ldots, \beta_K^Z$ and they sum to zero.

- An additive alternative model specifies

$$\text{logit } \pi_{ik} = \alpha + \beta I\{X_i = 1\} + \beta_k^Z.$$

  Under this model $\theta_{XY(k)} = e^\beta$ for all k. The odds *ratios* are the same across strata, but the strata-specific probabilities of success change with $Z = k$. $X \perp Y|Z$ if we accept $H_0 : \beta = 0$.

- The most general alternative is

$$\text{logit } \pi_{ik} = \alpha + \beta I\{X_i = 1\} + \beta_k^Z + \beta_k^{XZ} I\{X_i = 1\}.$$

  This is a saturated model and allows $\theta_{XY(1)} \neq \theta_{XY(2)} \neq \cdots \neq \theta_{XY(K)}$. $X \perp Y|Z$ if we accept $H_0 : \beta = 0, \beta_k^{XZ} = 0$ for $k = 1, \ldots, K$.

- Both of these alternatives allow testing $H_0 : X \perp Y|Z$ in PROC LOGISTIC with a Wald test.

Cochran-Mantel-Haenszel Statistic

$$\text{CMH} = \frac{\left[\sum_{k=1}^{K}(n_{11k} - \hat{\mu}_{11k})\right]^2}{\sum_{k=1}^{K} \text{var}(n_{11k})},$$

where $\hat{\mu}_{11k} = n_{1+k}n_{+1k}/n_{++k}$ and
$\text{var}(n_{11k}) = n_{1+k}n_{2+k}n_{+1k}n_{+2k}/n_{++k}^2(n_{++k-1})$.

- Motivated by retrospective studies, e.g. case-control, so response (column) totals are assumed fixed. Then row (treatment) totals are sufficient and conditioned on. Leaves only one free parameter in each table, say n_{11k} which is hypergeometric under H_0:
- Null hypothesis is $H_0 : X \perp Y|Z$.
- $\hat{\mu}_{11k} = E(n_{11k})$ and $\text{var}(n_{11k})$ are under H_0.
- When H_0 true, CMH $\overset{\bullet}{\sim} \chi_1^2$.

A bit more detail why n_{11k} are hypergeometric ...

	$Y=1$	$Y=2$	
$X=1$	n_{11k}	n_{12k}	n_{1+k}
$X=2$	n_{21k}	n_{22k}	n_{2+k}
	n_{+1k}	n_{+2k}	n_{++k}

- There are n_{1+k} "red balls" $X=1$ and n_{2+k} "green balls" $X=2$.
- We choose n_{+1k} balls (controls $Y=1$) from the urn. Under independence one cannot tell the difference between a case and a control. The number n_{11k} out of n_{+1k} that are "red," i.e. exposures $X=1$, is hypergeometric (under H_0).
- See page 91, (3.16) in Section 3.5.1.
- Back to logistic regression formulation...

- The additive alternative looks in a certain direction for deviations from conditional independence $X \perp Y|Z$. It can be more powerful when the additive model truly holds.
- The interaction, saturated model can be more powerful when the additive alternative does not hold.
- The CMH test is equivalent to a score test for testing $H_0 : \beta = 0$ in the additive model; see your book (p. 227). This test can be carried out in PROC FREQ.

```
data cmh;
 input center $ treat response count;
 datalines ;
 a 1 1 11
 a 1 2 25
 a 2 1 10
 a 2 2 27
 b 1 1 16
 b 1 2 4
  ...
 h 1 1 4
 h 1 2 2
 h 2 1 6
 h 2 2 1
;
proc freq ; weight count; tables center*treat*response / cmh;
```

With annotated output:

Cochran–Mantel–Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	6.3841	0.0115
2	Row Mean Scores Differ	1	6.3841	0.0115
3	General Association	1	6.3841	0.0115

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Limits	
Case–Control (Odds Ratio)	Mantel–Haenszel	2.1345	1.1776	3.8692
	Logit **	1.9497	1.0574	3.5949
Cohort (Col1 Risk)	Mantel–Haenszel	1.4245	1.0786	1.8812
	Logit **	1.2194	0.9572	1.5536
Cohort (Col2 Risk)	Mantel–Haenszel	0.8129	0.6914	0.9557
	Logit	0.8730	0.7783	0.9792

** These logit estimators use a correction of 0.5 in every cell of those tables that contain a zero.

We see CMH$=6.384$ with $p = 0.0115$ and so we reject that $X \perp Y|Z$ in favor of a *common odds ratio* estimated as $\hat{\theta}_{XY} = 2.13\ (1.18, 3.87)$.

Alternatively, we can fit the three logit models:

```
data cmh2;
 input center $ treat y n; treat=abs(treat-2);
 datalines ;
 a 1 11 36
 a 2 10 37
 b 1 16 20
 b 2 22 32
...
 h 1 4  6
 h 2 6  7
;
proc logistic data=cmh2; class center; model y/n = center;
proc logistic data=cmh2; class center; model y/n = treat center;
proc logistic data=cmh2; class center; model y/n = treat center treat*center;
```

Label the models (1), (2), and (3) respectively. The fit of (2) corresponds to the alternative in the CMH test:

Type 3 Analysis of Effects

Effect	DF	Wald Chi−Square	Pr > ChiSq
treat	1	6.4174	0.0113
center	7	58.4897	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi−Square	Pr > ChiSq
Intercept		1	−1.2554	0.2692	21.7413	<.0001
treat		1	0.7769	0.3067	6.4174	0.0113
center	a	1	−0.0667	0.3133	0.0453	0.8315
center	b	1	1.9888	0.3556	31.2789	<.0001
center	c	1	1.0862	0.3596	9.1236	0.0025
center	d	1	−1.4851	0.5707	6.7711	0.0093
center	e	1	−0.5866	0.4582	1.6390	0.2005
center	f	1	−2.2136	0.9171	5.8260	0.0158
center	g	1	−0.8644	0.7016	1.5178	0.2180

- We reject $H_0 : \beta = 0$ ($p = 0.0113$) and thus reject $X \perp Y|Z$. We estimate the common odds ratio to be $e^{-0.777} = 2.18$ $(1.19, 3.97)$ (from excised output).
- By adding `/ aggregate scale=none;` to the MODEL statement, we find the Pearson GOF $X^2 = 8.03$ on $df = 16 - (1 + 1 + 7) = 7$ with $p = 0.33$. The additive model does not show gross LOF.

Let's examine the full interaction (saturated) model anyway...

- The `-2 Log L` from (1) is 283.689 (under `Model Fit Statistics`) and from (3) is 267.274. The number of parameters added to (1) to get (3) is 8. The p-value is $P(\chi^2_8 > 16.415) = 0.0368$.
- We reject that $H_0 : \beta = 0, \beta_k^{XY} = 0$ in the saturated model (3) and hence also reject $X \perp Y|Z$. Notice the p-value is about 3 times larger though; we lost some power by considering a *very* general alternative.
- By accepting this more complex alternative we have lost interpretability as well, the estimated odds ratio $\hat{\theta}_{XY(k)}$ changes with center k. From (3)'s fit

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
treat	1	0.0064	0.9362
center	7	24.2036	0.0010
treat*center	7	4.0996	0.7682

- The Type III effects table shows we can drop the treat*center from the model and so we go with the analysis and results from the CMH analysis and/or logit analysis on the previous slide.

Consider an $I \times 2$ table where X is categorical and Y is binary. When the probability of $Y = 2$ is the same for each level of $X = i$, $\pi(i) = P(Y = 2|X = i) = \pi$, we have $X \perp Y$. In terms of log-odds this is

$$\text{logit } \pi(i) = \alpha.$$

1. If X is nominal, allowing a separate probability for each level of X gives

$$\text{logit } \pi(i) = \alpha + \beta_i,$$

for $i = 1, \ldots, I$; the saturated model.

2. When X is ordinal, we can use the above alternative model, or instead use scores $u_1 \leq u_2 \leq \cdots \leq u_I$ in place of X and fit the model

$$\text{logit } \pi(i) = \alpha + \beta u_i.$$

- In the first case a test of $H_0 : \beta_1 = \cdots = \beta_I = 0$ is a test of $H_0 : X \perp Y$ versus the most general possible alternative. The test statistic (score, Wald, or LRT) has a χ^2_{I-1} distribution under H_0.
- In the second case a test of $H_0 : \beta = 0$ tests $X \perp Y$ versus a focused, *linear* alternative. The test statistic has a χ^2_1 distribution under H_0.
- If X is ordinal and the logistic regression model treating X as continuous fits okay, you can increase your power to reject $H_0 : X \perp Y$ by looking in one particular direction (linear log-odds of scores).
- If the model *does not* fit then you can *lose* power by looking in only one place to the exclusion of other alternatives.
- For nominal X we pretty much can only test the saturated model to the intercept model.

6.5: Existence of finite $\hat{\beta}$ [One more time]

- Estimates $\hat{\beta}$ exist, except when data are perfectly separated.
- Complete separation happens when a linear combination of predictors perfectly predicts the outcome. See Figure 6.5 (p. 234). Here, there are an infinite number of perfect fitting curves that have $\alpha = \infty$. Essentially, there is a value of x that perfectly separates the 0's and 1's. In two-dimensions there would be a line separating the 0'1 and 1's.
- Quasi-complete separation happens when there's a line that separates 0's and 1's but there's some 0's and 1's on the line. We'll look at some pictures.
- The end result is that the model will appear to fit but the standard errors will be absurdly large. This is the opposite of what's really happening, that the data can be perfectly predicted.
- A (Bayesian!) fix is hiding in Section 7.4.7 (p. 275). Add FIRTH to the MODEL statement, and quasi and complete separation issues vanish!

Power Settings

Recall:

- $\alpha = P(\text{reject } H_0 | H_0 \text{ true})$
- $\beta = P(\text{accept } H_0 | H_1 \text{ true})$
- Power is $1 - \beta = P(\text{reject } H_0 | H_1 \text{ true})$.

Often we want to find an overall sample size n such that, for example, $1 - \beta = 0.9$ or 0.8 while capping off $\alpha = 0.05$.

One sample proportion

Say we want to test $H_0 : \pi = \pi_0$ for $Y \sim \text{bin}(n, \pi)$.

- The score test statistic is $Z_0 = \frac{\hat{\pi} - \pi_0}{\sigma_0}$ where $\hat{\pi} = Y/n$ and $\sigma_0 = \sqrt{\pi_0(1-\pi_0)/n}$.
- Under $H_0 : \pi = \pi_0$, $Z \overset{\bullet}{\sim} N(0, 1)$; this determines $z_{\alpha/2}$.
- The power $1 - \beta$ is a function of the hypothesized π_0, the true π_1, and the sample size through σ_0 and $\sigma_1 = \sqrt{\pi_1(1-\pi_1)/n}$.

We compute:

$$
\begin{aligned}
1-\beta &= P(\text{reject } H_0 | H_1 \text{ true}) \\
&= P(|Z_0| > z_{\alpha/2} | \pi = \pi_1) \\
&= 1 - P(-z_{\alpha/2} \leq Z_0 \leq z_{\alpha/2} | \pi = \pi_1) \\
&= 1 - P(-z_{\alpha/2}\sigma_0 + \pi_0 \leq \hat{\pi} \leq z_{\alpha/2}\sigma_0 + \pi_0 | \pi = \pi_1) \\
&= 1 - P\left(\frac{-z_{\alpha/2}\sigma_0 + \pi_0 - \pi_1}{\sigma_1} \leq \frac{\hat{\pi} - \pi_1}{\sigma_1} \leq \frac{z_{\alpha/2}\sigma_0 + \pi_0 - \pi_1}{\sigma_1}\right) \\
&= 1 - P\left(\frac{-z_{\alpha/2}\sigma_0 + \pi_0 - \pi_1}{\sigma_1} \leq Z \leq \frac{z_{\alpha/2}\sigma_0 + \pi_0 - \pi_1}{\sigma_1}\right)
\end{aligned}
$$

For a given β, α, π_0, and π_1, we can solve the above equation for the sample size n as

$$n = \frac{\left[Z_{\alpha/2}\sqrt{\pi_0(1-\pi_0)} + Z_{\beta}\sqrt{\pi_1(1-\pi_1)}\right]^2}{(\pi_0 - \pi_1)^2}.$$

Check out http://homepage.cs.uiowa.edu/∼rlenth/Power/

6.6.1 Testing $H_0 : \pi_1 = \pi_2$ from two samples

Recall the two-sample proportion problem. Assume the same number of observations n will be collected in each group $X = 1$ and $X = 2$.

$$Y_1 \sim \text{bin}(n_1, \pi_1) \perp Y_2 \sim \text{bin}(n_2, \pi_2).$$

Let $\hat{\pi}_1 = Y_1/n$ and $\hat{\pi}_2 = Y_2/n$. The CLT gives us

$$\hat{\pi}_1 \overset{\bullet}{\sim} N\left(\pi_1, \frac{\pi_1(1-\pi_1)}{n_1}\right) \perp \hat{\pi}_2 \overset{\bullet}{\sim} N\left(\pi_2, \frac{\pi_2(1-\pi_2)}{n_2}\right),$$

and so

$$\hat{\pi}_1 - \hat{\pi}_2 \overset{\bullet}{\sim} N\left(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right).$$

Under $H_0 : \pi_1 = \pi_2$ and $n_1 = n_2$ the test statistic is

$$Z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{2\hat{\pi}(1-\hat{\pi})/n}},$$

where $\hat{\pi} = (Y_1 + Y_2)/(2n)$ is the pooled estimator (the MLE under H_0). Similar computations as in the one-sample case leads to

$$n_1 = n_2 = (z_{\alpha/2} + z_\beta)^2 \frac{\pi_1(1-\pi_1) + \pi_2(1-\pi_2)}{(\pi_1 - \pi_2)^2}.$$

Note that for $\alpha = 0.05$ and $\beta = 0.1$ we have $z_{0.025} = 1.960$ and $z_{0.1} = 1.282$. $1 - \beta = 0.99$ yields $z_{0.01} = 2.326$.
What happens when $\pi_1 \approx \pi_2$?

6.6.2 Sample size for simple logistic regression

Let

$$\text{logit } \pi(x) = \alpha + \beta X,$$

where $X \sim N(\mu, \sigma^2)$ and

$$\tau = \log \left\{ \frac{\pi(\mu+\sigma)/[1-\pi(\mu+\sigma)]}{\pi(\mu)/[1-\pi(\mu)]} \right\},$$

the log of the ratio of event odds when $x = \mu + \sigma$ and $x = \mu$. Then to test $H_0 : \beta \leq 0$ versus $H_0 : \beta > 0$ (or the other direction) at significance α and power $1 - \beta$ we need sample size

$$n = [z_\alpha + z_\beta e^{-\tau^2/4}]^2 [1 + 2\pi(\mu)\delta]/[\pi(\mu)\tau^2],$$

where

$$\delta = [1 + (1+\tau^2)e^{5\tau^2/4}]/[1 + e^{-\tau^2/4}].$$

Text example.

- X is cholesterol level, Y indicates "severe heart disease."
- Know $\pi(\mu) = 0.08$. Want to be able to detect a 50% increase in probability for a standard deviation increase in cholesterol. 50% increase in probability is $1.5 \times 0.08 = 0.12$.
- $\pi(\mu)/[1 - \pi(\mu)] = 0.08/0.92 = 0.087$.
- $\pi(\mu + \sigma)/[1 - \pi(\mu + \sigma)] = 0.12/0.88 = 0.136$. So the odds ratio is $0.136/0.087 = 1.57$, and $\tau = \log(1.57) = 0.45$.
- Then for $\alpha = 0.05$, $1 - \beta = 0.9$, we have $\delta = 1.306$ and $n = 612$.
- Note: didn't need to know μ and σ, but rather $\pi(\mu)$ and $\pi(\mu + \sigma)$.

6.6.3 Sample size for one effect in multiple logistic regression

Say now that we're interested in X_1 but there's $p-2$ more predictors $X_2, \ldots, X_{p-1}$. Let R denote the multiple correlation between X_1 and the remaining predictors:

$$R = \max_{||\mathbf{a}||=1} \{\text{corr}(X_1, a_2 X_2 + \cdots + a_{p-1} X_{p-1})\}.$$

Let $\pi(\boldsymbol{\mu}) = \pi(\mu_1, \mu_2, \ldots, \mu_{p-1})$ be the probability at the mean of all $p-1$ variables.
τ is the now the log odds ratio comparing $\pi(\mu_1 + \sigma_1, \mu_2, \ldots, \mu_{p-1})$ to $\pi(\mu_1, \mu_2, \ldots, \mu_{p-1})$.

$$n = [z_\alpha + z_\beta e^{-\tau^2/4}]^2 [1 + 2\pi(\boldsymbol{\mu})\delta]/[\pi(\boldsymbol{\mu})\tau^2(1 - R^2)].$$

Text example (continued):

- Say we have another variable X_2 is blood pressure and $R = \text{corr}(X_1, X_2) = 0.4$.
- Then $n = 612/(1 - 0.4^2) = 729$.
- What happens when $\text{corr}(X_1, X_2) \approx 1$. Is this problematic? Hint: think about the interpretation of β_1.
- The formula only provide, at best, very approximate indications of sample sizes. Many applications have only a crude guess for $\hat{\pi}$ and R, and X may be far from normally distributed.

6.6.4, 6.6.5, & 6.6.6 Misc. power and sample size considerations

Read over if interested.