Lecture 18: Logistic Regression Continued

Dipankar Bandyopadhyay, Ph.D.

BMTRY 711: Analysis of Categorical Data Spring 2011

Division of Biostatistics and Epidemiology

Medical University of South Carolina

Maximum Likelihood Estimation for Logistic Regression

- Consider the general logistic regression model

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK}$$

with

$$Y_i \sim Bern(p_i),$$

$i = 1, ..., n$.

- The likelihood is

$$L(\beta_0, \beta_1, ..., \beta_K) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{(1-y_i)}$$

- Then, we maximize $\log L(\beta_0, \beta_1, ..., \beta_K)$ to get MLE's.
- The log-likelihood is

$$\begin{aligned} &\log[L(\beta_0, \beta_1, ..., \beta_K)] = \\ &\quad \beta_0 \left(\textstyle\sum_{i=1}^{n} y_i\right) + \beta_1 \left(\textstyle\sum_{i=1}^{n} x_{i1} y_i\right) + ... + \beta_K \left(\textstyle\sum_{i=1}^{n} x_{iK} y_i\right) \\ &\qquad - \textstyle\sum_{i=1}^{n} \log(1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK}}) \end{aligned}$$

- With some data, it is possible that the data is so sparse (most people respond 0 or most people respond 1) that there is no solution, but if there is a solution, it is unique and the maximum.
- In practice, if there is no solution, your logistic regression package will say something like 'Convergence not reached after 25 iterations'.
- However, if the iterative approach converges, then the MLE is obtained.
- In the next lectures, we will discuss other methods that may be more appropriate with sparse data (conditional logistic regression, exact methods).
- **Large Sample Distribution**: In large samples, $\hat{\boldsymbol{\beta}}$ has approximate distribution

$$\hat{\boldsymbol{\beta}} \sim N_{K+1}\left(\beta, \left[\sum_{i=1}^{n} p_i(1-p_i)\mathbf{x}_i\mathbf{x}_i'\right]^{-1}\right)$$

Sample Size requirements for approximate normality of the MLE's

- Asymptotics: The sample size needed to get approximately normal estimates using a general logistic regression model depends on the covariates, and the true underlying probabilities, p_i
- As a rule of thumb, for the parameters of a given model to be approximately normal, you would like

$$\frac{\text{total sample size } (n)}{\text{\# parameters in model}} \approx 15$$

- Thus, for a model like

$$p_i = \left(\frac{e^{\beta_0 + \beta_1 x_{i1}}}{1 + e^{\beta_0 + \beta_1 x_{i1}}} \right),$$

you could have $n = 30$ different values of x_i and get approximately normal estimates of (β_0, β_1) since

$$\frac{30}{2} \approx 15$$

Confidence Intervals

- For confidence intervals and test statistics, you can use

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = \left[\sum_{i=1}^{n} \widehat{p}_i(1-\widehat{p}_i)\mathbf{x}_i\mathbf{x}_i'\right]^{-1},$$

- In particular, $\widehat{Var}(\hat{\boldsymbol{\beta}}_k)$ is the $(k+1, k+1)$ element of $\widehat{Var}(\hat{\boldsymbol{\beta}})$ (since there is an intercept, it is not the k^{th}, but $k+1$.)
- Thus, 95% (large sample) confidence intervals are formed via

$$\hat{\boldsymbol{\beta}}_k \pm 1.96\sqrt{\widehat{Var}(\hat{\boldsymbol{\beta}})_{(k+1),(k+1)}}$$

Test Statistics

- For the logistic model, a test of interest is whether x_{ik} affects the probability of success,

$$\mathsf{H}_0 : \beta_k = 0,$$

or equivalently, whether the probability of success is independent of x_{ik}.
Wald Statistic

- The most obvious test statistic is the WALD statistic:

$$Z = \frac{\hat{\boldsymbol{\beta}}_k}{\sqrt{\widehat{Var}(\hat{\boldsymbol{\beta}})_{kk}}} \sim N(0,1)$$

in large samples under the null.

Likelihood Ratio Statistic

- The likelihood ratio statistic is

$$\begin{aligned}
\Delta G^2 &= 2\{\log[L(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1, ..., \hat{\boldsymbol{\beta}}_K)|\mathsf{H}_A] - \\
&\qquad \log[L(\tilde{\beta}_0, \tilde{\beta}_1, ..., \tilde{\beta}_k = 0, ..., \tilde{\beta}_K)|\mathsf{H}_0]\} \\
&= 2\sum_{j=1}^{n}\left[y_i \log\left(\frac{\widehat{p}_i}{\tilde{p}_i}\right) + (1-y_i)\log\left(\frac{1-\widehat{p}_i}{1-\tilde{p}_i}\right)\right] \\
&\sim \chi^2_1
\end{aligned}$$

where $\widehat{p}_j$ is the estimate under the alternative, and $\tilde{p}_j$ is the estimate under the null.

Score test statistic

- The score statistic for

$$H_0 : \beta_k = 0,$$

  is based on the

$$X^2 = \frac{[\sum_{i=1}^n x_{ik}(y_i - \tilde{p}_i)]^2}{\widehat{Var}[\sum_{i=1}^n x_{ik}(y_i - \tilde{p}_i)]},$$

- In general logistic regression, this statistic does not have the simple form that it had earlier, mainly because we have to iterate to get the estimates of β under the null (and thus to get $\tilde{p}_i$).

Goodness-of-Fit

- When we discussed goodness-of-fit statistics earlier, we looked at the likelihood ratio statistic for the given model versus the 'saturated' model.
- However, when the underlying data are bernoulli data,

$$Y_i \sim Bern(p_i),$$

  a 'saturated model', i.e., a model in which we have a different parameter for each individual is not informative, as we now show.
- In particular, since

$$Bern(p_i) = Bin(1, p_i),$$

  our estimate of p_i for the saturated model is

$$\widehat{p_i} = \frac{Y_i}{1} = \begin{cases} 1 \text{ if } Y_i = 1 \\ 0 \text{ if } Y_i = 0 \end{cases} .$$

  which is either 0 or 1.

- Now, the likelihood at the MLE for the saturated model is

$$\prod_{i=1}^{n} \widehat{p}_i^{\,y_i} (1 - \widehat{p}_i)^{1-y_i}.$$

- For individual i, if $y_i = 1$, then $\widehat{p}_i = 1$ and

$$\widehat{p}_i^{\,y_i} (1 - \widehat{p}_i)^{1-y_i} = 1^1 0^0 = 1$$

- Similarly, if $y_i = 0$, then $\widehat{p}_i = 0$ and

$$\widehat{p}_i^{\,y_i} (1 - \widehat{p}_i)^{1-y_i} = 0^0 1^1 = 1.$$

- Then, the likelihood at the MLE for the saturated model is

$$\prod_{i=1}^{n} 1 = 1,$$

and the log-likelihood equals $\log(1) = 0$.

- Suppose, even if the underlying data are coded as bernoulli for your computer program, that, in reality, the underlying data actually arose as product binomial, i.e., many subjects have the same covariate vector $\mathbf{x}_i$.
- In particular, suppose the underlying data are actually made up of J binomials, and there are n_j subjects with the same covariate vector $\mathbf{x}_j$, $j = 1, ..., J$; $n = \sum_{j=1}^{J} n_j$, so we have

$$Y_j \sim Bin(n_j, p_j).$$

- Since the product bernoulli and product binomial likelihoods are proportional (the same except for combinatorial terms not depending on β), we would get the same MLE and standard error estimates; however, for goodness-of-fit, we would use the product binomial likelihood.
- With $\mathbf{x}_j$ the covariate vector associated with (binomial) group j, we want to test the fit of the model

$$\text{logit}(p_j) = \mathbf{x}_j'\beta$$

versus the saturated model (in which we estimate a different p_j for each j.)

The Deviance

- As with logistic models discussed earlier, the likelihood ratio statistic for a given model M_1 with estimates $\widehat{p}_j$ versus a 'saturated' model in which $\widehat{p}_j = \frac{y_j}{n_j}$, is often called the deviance, denoted by D^2,

$$\begin{aligned} D^2(\mathsf{M}_1) &= 2\{\log[L(\widehat{\beta})|\mathsf{Sat}] - \log[L(\tilde{\beta})|\mathsf{M}_1]\} \\ &= \sum_{j=1}^{J}\left[y_j \log\left(\frac{y_j}{n_j\widehat{p}_j}\right) + (n_j - y_j)\log\left(\frac{n_j - y_j}{n_j(1-\widehat{p}_j)}\right)\right] \\ &= \sum_{j=1}^{J}\sum_{k=1}^{2} O_{jk}\log\left(\frac{O_{jk}}{E_{jk}}\right) \\ &\sim \chi^2_P \end{aligned}$$

  under the null, where

$$E_{j1} = n_j\widehat{p}_j \qquad \text{and} \qquad E_{j2} = n_j(1-\widehat{p}_j)$$

  and

$$P = \#\text{ parameters in sat. model} - \#\text{ parameters in } \mathsf{M}_1$$

- Deviance D^2 is often used as measure of 'overall' goodness-of-fit, and is a test statistic from terms **left out** of the model.

Score Statistic

- Again, sometimes the 'Score Statistic' or Pearson's chi-square is used to look at the goodness of fit for a given model versus the saturated model:

$$\begin{aligned} X^2 &= \sum_{j=1}^{J} \left(\frac{[y_j - n_j \widehat{p}_j]^2}{n_j \widehat{p}_j (1-\widehat{p}_j)} \right) \\ &= \sum_{j=1}^{J} \frac{[y_j - n_j \widehat{p}_j]^2}{n_j \widehat{p}_j} + \frac{[(n_j - y_j) - n_j(1-\widehat{p}_j)]^2}{n_j (1-\widehat{p}_j)} \\ &= \sum_{j=1}^{J} \sum_{k=1}^{2} \frac{(O_{jk} - E_{jk})^2}{E_{jk}} \\ &\sim \chi^2_P \end{aligned}$$

- For the deviance D^2 and Pearson's chi-square (X^2) to be asymptotically chi-square, so that they can be used as statistics to measure the fit of the model, you need each n_j to be large (actually, $n_j \to \infty$).

Sample Size for approx chi-square D^2

- For a simple (2×2) table, recall we said that, for the chi-square approximation for the Deviance (likelihood ratio) and Pearson's chi-square to be valid, we should have

$$75\% \text{ of the } E_{jk} \geq 5$$

- In terms of the logistic regression model, this means that 75% of

$$E_{j1} = n_j \widehat{p}_j \geq 5 \qquad \text{and} \qquad E_{j2} = n_j(1 - \widehat{p}_j) \geq 5,$$

or, for stratum 'j',

$$E_{j1} + E_{j2} \geq 5 + 5$$

i.e.,

$$E_{j1} + E_{j2} \geq 10$$

- Note, though,

$$\begin{aligned} E_{j1} + E_{j2} &= n_j \widehat{p}_j + n_j(1 - \widehat{p}_j) \\ &= n_j \\ &\geq 10 \end{aligned}$$

- Thus, as a rough rule-of-thumb, for the chi-square approximation for the Deviance and Pearson's chi-square to be valid, we should have

$$75\% \text{ of the } n_j \geq 10$$

- However, n_j is often small. For example,
  1) the covariates may be continuous, so that $n_j = 1$.
  2) the model may have a lot of covariates (so that very few individuals have the same pattern), and most individuals will have different $\mathbf{x}_j$'s,
- In these cases, the Deviance and Pearson's chi-square will not be approximately chi-square. We will discuss other statistics that can be used in this situation.

Example: The esophageal cancer data

- There are six age levels, four levels of alcohol, and four levels of tobacco. In theory, there are

$$96 = 6 \times 4 \times 4$$

  strata, but 8 strata have no observations, leaving $J = 88$ strata for model fitting. With 975 observations (200 cases and 775 controls), this means we have about 11 observations per stratum, which, on the surface, appears to be a good number.
- However, when looking more closely, we see that only 31 of the 88 strata (35%) have $n_j \geq 10$, so we may want to be a little careful when using D^2.
- Also, with the saturated model having this many degrees of freedom (88), D^2 will have power against a lot of alternatives, but not necessarily high power.
- For example, if we left one marginally significant term out of the model (with 1 df), D^2 probably won't be able to detect a bad fit.
- We will fit models to assess trends in cancer with alcohol and tobacco consumption, and their interaction. We will discuss modelling the data a little later.

GOF with nested Models

- For bernoulli data, the parameter estimates are OK, but since each group j has $n_j = 1$, D^2 will not be asymptotically chi-square, and could give you a very deceptive idea of the fit.
- In this case, you could do as before: fit a broader model than the given model, but not the saturated model.

Likelihood Ratio Statistic for Nested Models

- Suppose
$$Y_i \sim Bern(p_i) = Bin(1, p_i),$$
- Sometimes you can look at a broader model than the one of interest to test for 'Goodness-of-Fit'.
- In general, to test the 'fit' of a given model, you can put interaction terms and/or square terms in the model and test their significance.
- For example, suppose you want to see if Model 1 fits,
Model 1:
$$p_i = \left(\frac{e^{\beta_0 + \beta_1' \mathbf{x}_i}}{1 + e^{\beta_0 + \beta_1' \mathbf{x}_i}} \right),$$
- This model is nested in Model 2:
Model 2:
$$p_i = \left(\frac{e^{\beta_0 + \beta_1' \mathbf{x}_i + \beta_2' \mathbf{z}_i}}{1 + e^{\beta_0 + \beta_1' \mathbf{x}_i + \beta_2' \mathbf{z}_i}} \right),$$
- We want to test
$$\mathsf{H}_0 : \beta_2 = 0$$

- Recall, the deviance D^2 is sort of like a SUMS of SQUARES ERROR (error in the given model versus the saturated), and a smaller model will always have more the same or more error than the bigger model.
- As with log-linear and logistic models discussed earlier, to test for significance of parameters in model 2 versus model 1, you can use

$$\begin{aligned}
\Delta D^2(\mathsf{M}_2|\mathsf{M}_1) &= D^2(\mathsf{M}_1) - D^2(\mathsf{M}_2) \\
&= 2\{\log[L(\widehat{\beta})|\mathsf{Sat}] - \log[L(\tilde{\beta})|\mathsf{M}_1]\} - \\
&\quad 2\{\log[L(\widehat{\beta})|\mathsf{Sat}] - \log[L(\tilde{\beta})|\mathsf{M}_2]\} \\
&= 2\{\log[L(\tilde{\beta})|\mathsf{M}_2] - \log[L(\tilde{\beta})|\mathsf{M}_1]\}
\end{aligned}$$

  which is the 'change in D^2' for model 2 versus model 1.
- If the smaller model fits, in large samples,

$$\Delta D^2(\mathsf{M}_2|\mathsf{M}_1) \sim \chi^2_M,$$

  where M parameters are set to 0 in the smaller model.

- Even though we need binomial data (n_j to be large) for D^2 to be approximately chi-square, the same is not true for ΔD^2.
- Each individual in the study could have a different $\mathbf{x}_i$, (i.e. $n_j = 1$) and ΔD^2 will still be approximately a chi-square test statistic for $\beta_2 = 0$, since $\log[L(\widehat{\beta})|\text{Sat}]$ subtracts out in the difference.
- Using differences in D^2's is just a 'trick' to get the likelihood ratio statistic

$$2\{\log[L(\tilde{\beta})|\mathsf{M}_2] - \log[L(\tilde{\beta})|\mathsf{M}_1]\},$$

in particular, the log-likelihood for the saturated model subtracts out, and does not affect asymptotics.

- Basically, regardless of the size of n_j, if you are comparing
  Model 1:

$$p_i = \left(\frac{e^{\beta_0 + \beta_1' x_i}}{1 + e^{\beta_0 + \beta_1' x_i}} \right),$$

  versus
  Model 2:

$$p_i = \left(\frac{e^{\beta_0 + \beta_1' x_i + \beta_2' z_i}}{1 + e^{\beta_0 + \beta_1' x_i + \beta_2' z_i}} \right),$$

  a Taylor Series approximation can be used to show that,

$$\Delta D^2(\mathsf{M}_2|\mathsf{M}_1) \approx \left[\frac{\hat{\boldsymbol{\beta}}_2}{\sqrt{\widehat{Var}(\hat{\boldsymbol{\beta}}_2)}} \right]^2 \sim \chi_1^2$$

Using G^2

- As before, another popular statistic is G^2, which is the likelihood ratio test statistic for whether the parameters, except the intercept μ, are 0 (i.e., the significance of parameters in the model).
- For G^2, the larger model always has bigger G^2 since it has more parameters (sort of like SUMS of SQUARES REGRESSION)
- Again, to test for significance of parameters in model 2 versus model 1, you can use

$$\Delta G^2(\mathsf{M}_2|\mathsf{M}_1) = G^2(\mathsf{M}_2) - G^2(\mathsf{M}_1)$$

  which is the 'change in G^2' for model 2 versus model 1.
- Thus, the likelihood ratio statistic for two nested models can be calculated using either ΔG^2 or ΔD^2.

Analysis of Esophageal Data

- We will treated AGE, TOBACCO, and ALCOHOL as both ordered and not ordered (quantitative).
- When treating age as ordered, we assigned the values

$$AGE = \begin{cases} 30 \text{ if } 25\text{-}34 \\ 40 \text{ if } 35\text{-}44 \\ 50 \text{ if } 45\text{-}54 \\ 60 \text{ if } 55\text{-}64 \\ 70 \text{ if } 65\text{-}74 \\ 80 \text{ if } 75+ \end{cases} .$$

- Tobacco is given in units of 10g/day,

$$TOB = \begin{cases} 0.5 \\ 1.5 \\ 2.5 \\ 4.0 \end{cases} .$$

- Alcohol is given in units of 10g/day,

$$ALC = \begin{cases} 2 \\ 6 \\ 10 \\ 15 \end{cases} .$$

- And, of course,

$$Y = \begin{cases} \text{1 if CASE} \\ \text{0 if CONTROL} \end{cases} .$$

- We also fit models with dummy variables for AGE (AGEGRP), TOBACCO (TOBGRP), and ALCOHOL (ALCGRP).

Summary of Model Fits

- Since age is considered a possible confounding factor, and there is enough data, all models contain 6 dummy variables for the main effect of AGE (AGEGRP), i.e., the 'basic' model is

$$\text{logit}(p_i) = \beta_0 + \sum_{j=1}^{5} \alpha_j a_{ij}$$

  where the a_{ij} are 5 dummy variables for the six age levels. We then added TOBACCO (TOBGRP), and ALCOHOL (ALCGRP) to this basic model (next page). In some models, TOBACCO and ALCOHOL are treated as ordinal, and in others, non-ordinal.
- When looking at interactions, some models have AGE as ordinal in the interactions, although still non-ordered (dummy variables) for the main effects.

#	COVARIATES FITTED (in addition to AGEGRP)	df	D^2	Hypothesis Tested/ Interpretation
1	TOBGRP + ALCGRP	76	82.34	Non-ordered main effects
2	TOBGRP + ALC	78	87.51	Linear effect of alcohol
3	TOBGRP + ALC + ALC^2	77	87.01	Linear & quad effects of alcohol
4	ALCGRP + TOB	78	84.53	Linear effect of tobacco
5	ALCGRP + TOB + TOB^2	77	83.73	Linear & quad effects of tobacco
6	ALC + TOB	80	89.02	Linear effects of tobacco and alcohol

7	ALC + TOB + ALC*TOB	79	88.05	Linear effects of tobacco and alcohol + Alc/tob interaction
8	ALCGRP + TOBGRP + ALC*TOB	75	81.37	Non-ordered main effects, but ordered alcohol + Alc/tob interacti
9	ALCGRP + TOBGRP + ALC*AGE	75	80.08	Non-ordered main effects, but ordered alcohol slope depends on age
10	ALCGRP + TOBGRP + TOB*AGE	75	82.33	Non-ordered main effects, but ordered tobacco slope depends on age

The p-values for these D^2 are all between .21 and .32

- As stated, the sample size may be large enough to support approximate chi-square distributions for the D^2's if the given model fits.
- However, these D^2s will not necessarily have high power to detect where the model does not fit. None of these D^2's are significant, saying all models are a good fit.
- Comparing models 1 and 2, there is some evidence that the increase in the log-odds ratio with alcohol may not be purely linear,

$$D^2(TOBGRP + ALC) - D^2(TOBGRP + ALCGRP) = 5.07$$

$$df = 78 - 76 = 2, \qquad p = .08$$

This is because, having ALCGRP which includes 3 dummy variables for ALCOHOL, is equivalent to having ALC, ALC^2, and ALC^3. Thus, this likelihood ratio statistic is comparing a model with ALC to one with (ALC, ALC^2, and ALC^3).

- However, when comparing models 2 and 3, which is looking for a quadratic effect (ALC^2), we find that ALC^2 is not significant,

$$D^2(TOBGRP + ALC) - D^2(TOBGRP + ALC + ALC^2) = 0.11$$

$$df = 77 - 76 = 1, \qquad p = .95,$$

  so that the deviation from a straight line alcohol detected in the first test may be due to chance.
- Linearity of trend with tobacco also seems adequate when comparing Model 4 with Models 1 and 5; $(p > .05)$ in both cases.
- Also, none of the models with interactions appear to add anything significant over models with just the main effects of tobacco and alcohol $p > .05$ in all cases
- Also, Model 6, which contains just linear effects for each of alcohol and tobacco, fits the data nearly as well as model 1, which has 4 more parameters for these effects:

$$D^2(TOB + ALC) - D^2(TOBGRP + ALCGRP) = 6.68$$

$$df = 80 - 76 = 4, \qquad p = .154,$$

Final Model

- The model we suggest is model 6,

$$\text{logit}(p_i) = \beta_0 + \sum_{j=1}^{5} \alpha_j a_{ij} + \beta_{\text{TOB}} \text{TOB}_i + \beta_{\text{ALC}} \text{ALC}_i$$

where the a_{ij} are 5 dummy variables for the six age levels and Tobacco is given in units of 10g/day,

$$TOB = \begin{cases} 0.5 \\ 1.5 \\ 2.5 \\ 4.0 \end{cases} ;$$

and Alcohol is given in units of 10g/day,

$$ALC = \begin{cases} 2 \\ 6 \\ 10 \\ 15 \end{cases} .$$

The estimates from the final model are (this is from Proc Logistic):

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	ChiSq
Intercept		1	-2.6068	0.4322	36.3812	<.0001
age	25-34	1	-4.6680	1.1680	15.9720	<.0001
age	35-44	1	-2.8131	0.5539	25.7905	<.0001
age	45-54	1	-1.0542	0.4438	5.6411	0.0175
age	55-64	1	-0.5047	0.4292	1.3827	0.2396
age	65-74	1	0.0317	0.4383	0.0052	0.9424
tob		1	0.4094	0.0921	19.7734	<.0001
alc		1	0.2548	0.0265	92.8054	<.0001

- We estimate that the odds of esophageal cancer increases by a factor of

$$\exp(\hat{\beta}_{\text{ALC}}) = \exp(.255) = 1.29$$

  for every additional 10 grams of alcohol consumed per day.
- We estimate that the odds of esophageal cancer increases by a factor of

$$\exp(\hat{\beta}_{\text{TOB}}) = \exp(.409) = 1.51$$

  for every additional 10 grams of tobacco consumed per day.

SAS Proc Logistic

```
data one;
 input age $ alcohol $ tobacco $ cases controls;

 if alcohol = '0-39'  then alc = 2;
 if alcohol = '40-79' then alc = 6;
 if alcohol = '80-119' then alc =10;
 if alcohol = '120+'   then alc =15;

 if tobacco = '0-9'   then tob = .5;
 if tobacco = '10-19' then tob =1.5;
 if tobacco = '20-29' then tob =2.5;
 if tobacco = '30+'   then tob =4.0;

count=cases;
 y=1;
output;

count=controls;
 y=0;
output;

drop cases controls;
```

```
cards;
25-34   0-39    0-9   0  40 25-34   0-39   10-19  0  10 25-34
0-39   20-29  0   6 25-34   0-39   30+    0   5 25-34  40-79    0-9
0  27 25-34  40-79   10-19  0   7 25-34  40-79   20-29  0   4 25-34
40-79   30+    0   7 25-34 80-119    0-9   0   2 25-34 80-119
10-19  0   1
<<see website for the data>>
```

```
proc print noobs;
 var age alc tob  y count;
run; /* PROC PRINT OUTPUT */
 AGE     ALC    TOB    Y    COUNT

25-34      2    0.5    1        0 25-34      2    0.5    0       40
25-34      2    1.5    1        0 25-34      2    1.5    0       10
25-34      2    2.5    1        0 25-34      2    2.5    0        6
25-34      2    4.0    1        0 25-34      2    4.0    0        5
       MORE

proc logistic descending;
    class age (PARAM=ref) ;
 model y = age tob alc /  aggregate scale=d /* specify for deviance */ ;
 freq count;
run;
```

```
/* SELECTED OUTPUT */
       Deviance and Pearson Goodness-of-Fit Statistics

Criterion        DF           Value     Value/DF     Pr > ChiSq

Deviance         80         89.0166       1.1127         0.2297
Pearson          80         91.6997       1.1462         0.1748

                Analysis of Maximum Likelihood Estimates
                                          Standard          Wald
Parameter            DF    Estimate       Error    Chi-Square    Pr >
ChiSq

Intercept             1     -2.6068      0.4322       36.3812 <.0001
age       25-34       1     -4.6680      1.1680       15.9720 <.0001
age       35-44       1     -2.8131      0.5539       25.7905 <.0001
age       45-54       1     -1.0542      0.4438        5.6411 0.0175
age       55-64       1     -0.5047      0.4292        1.3827 0.2396
age       65-74       1      0.0317      0.4383        0.0052 0.9424
tob                   1      0.4094      0.0921       19.7734 <.0001
alc                   1      0.2548      0.0265       92.8054 <.0001
```

Hosmer-Lemeshow Goodness-of-Fit Statistic when n_j's are small

- Suppose the responses are bernoulli and are not naturally grouped into binomials, either because there is a continuous covariate or many covariates so that the grouping would lead to sparse strata, i.e., suppose that

$$\text{more than } 25\% \text{ of } n_j < 10$$

- Then assume

$$Y_i \sim Bern(p_i) = Bin(1, p_i),$$

- We want to determine if the model,

$$p_i = \left(\frac{e^{\beta_0 + \beta_1' \mathbf{x}_i}}{1 + e^{\beta_0 + \beta_1' \mathbf{x}_i}} \right),$$

  is a good fit (we are thinking of $\mathbf{x}_i$ as a vector).

- One possible way is to fit a broader model (with interactions and/or squared, cubic, etc terms) and see if those ‘extra’ terms are significant.
- Hosmer and Lemeshow suggest forming G (usually 10) ‘Extra’ terms or ‘Extra’ covariates based on combinations of the covariates $\mathbf{x}_i$ in the logistic regression model. You again test if these ‘Extra’ Covariates are significant.

Simple Example

- For example, suppose there is **one continuous covariate** x_i, and we fit the model,

$$p_i = \left(\frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_i x_i}} \right),$$

  where there are $i = 1, ..., n$ observations in the dataset.

- We could then form 10 extra terms or 'groups' based on deciles of the covariate x_i, i.e., We form 10 groups of approximately equal size. The first group contains the $n/10$ subjects with the smallest values of x_i, the second group contains the $n/10$ subjects with the next smallest values of x_i, and ... the last group contains the $n/10$ subjects with the highest values of x_i.

- Suppose we define the 9 indicators (the last one is redundant)

$$I_{ig} = \begin{cases} 1 \text{ if individual } i \text{ is in group } g \\ 0 \text{ if otherwise} \end{cases},$$

- Then, to test goodness-of-fit, we consider the alternative model

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_i + \gamma_1 I_{i1} + ... + \gamma_9 I_{i9}$$

- If model

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_i$$

  is appropriate, then

$$\gamma_1 = ... = \gamma_9 = 0.$$

- Then, to test the fit of the model, we could use a likelihood ratio statistic, a score statistic, or a Wald statistic for

$$H_o : \gamma_1 = ... = \gamma_9 = 0,$$

  and each of these statistics would be approximately chi-square with 9 degrees of freedom if model fits.

General Logistic Model

- Now, we are looking at the fit of the general logistic model,

$$p_i = \left(\frac{e^{\beta_0 + \beta_1' \mathbf{x}_i}}{1 + e^{\beta_0 + \beta_1' \mathbf{x}_i}} \right),$$

  in which the covariates can be continuous, and a method of forming the 'Extra Covariates' is not obvious.

- Hosmer and Lemeshow suggest that we form groups based on 'deciles of risk'; if

$$\widehat{p}_i = \left(\frac{e^{\hat{\boldsymbol{\beta}}_0 + \hat{\boldsymbol{\beta}}_1' \mathbf{x}_i}}{1 + e^{\beta_0 + \hat{\boldsymbol{\beta}}_1' \mathbf{x}_i}} \right),$$

  is the predicted probability of failure for the given model, then Hosmer and Lemeshow suggest forming groups based on deciles of $\widehat{p}_i$, i.e.,

- We form 10 groups of approximately equal size. The first group contains the $n/10$ subjects with the smallest values of $\widehat{p}_i$, the second group contains the $n/10$ subjects with the next smallest values of $\widehat{p}_i$, and ... the last group contains the $n/10$ subjects with the highest values of $\widehat{p}_i$.
- You can show that, if there is just one covariate x_i, then we get equivalent groups whether we base the grouping on x_i or $\widehat{p}_i$; (because $\widehat{p}_i$ is a monotone transformation of x_i.)

A Model for which to look at the fit

- Again, we are looking at the fit of the general model,

$$p_i = \left(\frac{e^{\beta_0 + \beta_1' \mathbf{x}_i}}{1 + e^{\beta_0 + \beta_1' \mathbf{x}_i}} \right),$$

- Suppose we define the G group indicators

$$I_{ig} = \begin{cases} \text{1 if individual } i \ (\widehat{p}_i) \text{ is in group } g \\ \text{0 if otherwise} \end{cases},$$

  where the groups are based on 'deciles of risk'.

- Then, to test goodness-of-fit, we consider the alternative model

$$\text{logit}(p_i) = \beta_0 + \beta_1' \mathbf{x}_i + + \gamma_1 I_{i1} + ... + \gamma_9 I_{i9}$$

- Effectively, we are forming an 'alternative' model used to test the fit of the given model.
- Even though I_{ig} is based on the random quantities $\widehat{p}_i$, Moore and Spruill (1975), showed that, asymptotically, we can treat the partition as based on the true p_i (and thus, we can treat I_{ig} as a 'fixed' covariate).

- If model

$$\text{logit}(p_i) = \beta_0 + \beta_1' \mathbf{x}_i$$

  is appropriate, then

$$\gamma_1 = ... = \gamma_{G-1} = 0.$$

- Then, to test the fit of the model, we could use a likelihood ratio statistic, a score statistic, or a Wald statistic for

$$H_o : \gamma_1 = ... = \gamma_{G-1} = 0,$$

  and each of these statistics would be approximately chi-square with $(G-1)$ degrees of freedom if model fits.
- The score statistic only requires the estimate of (β_0, β_1) under the null, but, both the likelihood ratio and the Wald statistic require the estimates of γ_g from the alternative model.

The Hosmer-Lemeshow statistic

- Hosmer and Lemeshow (1982), suggest using Pearson's chi-square based on the grouping of observations,

$$\begin{aligned} X^2_{HL} &= \sum_{g=1}^{G} \frac{[O_g - E_g]^2}{n_g(E_g/n_g)[1-E_g/n_g]} \\ &= \sum_{g=1}^{G}\sum_{k=1}^{2} \frac{[O_{gk} - E_{gk}]^2}{E_{gk}}, \end{aligned}$$

  where, for failure $(Y = 0)$,

$$O_{g1} = O_g, \qquad E_{g1} = E_g$$

  and, for success $(Y = 1)$,

$$O_{g2} = n_g - O_g, \qquad E_{g2} = n_g - E_g$$

- Basically, this is Pearson's (Hosmer and Lemeshow) chi-square applied to the $(G \times 2)$ table of observed and expected counts in the G groups.

Example–Arthritis Clinical Trial

- Recall the example of an arthritis clinical trial comparing the drug auranofin and placebo therapy for the treatment of rheumatoid arthritis (Bombardier, et al., 1986).
- The response of interest is the self-assessment of arthritis, classified as (0) poor or (1) good.
- Individuals were randomized into one of the two treatment groups after baseline self-assessment of arthritis (with the same 2 levels as the response).
- The dataset contains 293 patients who were observed at both baseline and 13 weeks. The data from 25 cases are shown below:

- We are interested in a pretest-posttest analysis, in which we relate the individual's bernoulli response

$$Y_i = \begin{cases} 1 \text{ if good at 13 weeks} \\ 0 \text{ if poor at 13 weeks} \end{cases} .$$

to the covariates,

1. BASELINE self-assessment (denoted as x)
2. AGE IN YEARS,
3. GENDER
4. TREATMENT

- In particular, the model is

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_i + \beta_{\text{SEX}}\text{SEX}_i + \beta_{\text{AGE}}\text{AGE}_i + \beta_{\text{TRT}}\text{TRT}_i$$

  where the covariates are age in years at baseline (AGE_i), sex (SEX_i, 1=male, 0=female), treatment (TRT_i, 1 = auranofin, 0 = placebo), and x_i is baseline response (1 = good, 0 = poor)

- The main question is whether the treatment increases the odds of a more favorable response, after controlling for baseline response; secondary questions are whether the response differs by age and sex.

- The Hosmer-Lemeshow Statistic can be obtained in both SAS Proc Logistic.
- By default, SAS attempts to form 10 groups of approximately equal size $n/10$. Of course, in this dataset, we have 293 observations, so we cannot get exactly the same number of observations in each group;
- Different software packages (such as STATA) may differ slightly in the partition.

SAS Proc Logistic

The following ascii is in the current directory, and called art.dat

```
 1      54     1      0     1
 0      41     0      1     1
 .       .     .      .     .
 .       .     .      .     .
 .       .     .      .     .
 1      60     0      1     0
 1      63     1      0     0
```

```
/* SAS STATEMENTS */

 DATA ARTH;
  infile 'art.dat';
 input SEX AGE TRT x y;
;

proc logistic descending;
 model y = SEX  AGE TRT  x /lackfit;
run;
```

```
/* SELECTED OUTPUT */

          Response Profile

 Ordered                      Total
   Value             y     Frequency

       1             1           234
       2             0            59

Probability modeled is y=1.

              Analysis of Maximum Likelihood Estimates
                                  Standard          Wald
Parameter    DF    Estimate         Error    Chi-Square    Pr > ChiSq

Intercept     1      0.3327        0.8409        0.1566        0.6923
SEX           1      0.2168        0.3389        0.4095        0.5222
AGE           1    -0.00530        0.0144        0.1361        0.7122
TRT           1      0.7005        0.3136        4.9891        0.0255
x             1      1.4231        0.3102       21.0533        <.0001
```

Partition for the Hosmer and Lemeshow Test

		y = 1		y = 0	
Group	Total	Observed	Expected	Observed	Expected
1	29	13	15.51	16	13.49
2	29	17	18.02	12	10.98
3	29	23	20.99	6	8.01
4	29	25	23.18	4	5.82
5	28	24	23.47	4	4.53
6	31	30	26.21	1	4.79
7	29	23	25.24	6	3.76
8	29	28	26.28	1	2.72
9	30	26	27.45	4	2.55
10	30	25	27.65	5	2.35

Hosmer and Lemeshow Goodness-of-Fit Test	Chi-Square	DF	Pr > ChiSq
	12.9408	8	0.1139

Results

- Recall, the β's have the interpretation as the conditional log-OR for a one unit increase in the covariate, given all of the other covariates are the same.
- We estimate that the odds of good response increases by
  1. a factor of

$$\exp(\hat{\beta}_{\mathsf{TRT}}) = \exp(0.7005) = 2.015$$

  for those on treatment, (this is significant)
  2. a factor of

$$\exp(\hat{\beta}_{\mathsf{SEX}}) = \exp(0.2168) = 1.242$$

  for males, (not significant)
  3. a factor of

$$\exp(\hat{\beta}_{\mathsf{AGE}}) = \exp(-0.00530) = 0.995$$

  for every year older, (not significant)
  4. a factor of

$$\exp(\hat{\beta}_{\mathsf{X}}) = \exp(1.423) = 4.150$$

  for those who also had a good pre-treatment response at baseline, (this is significant).

- In SAS, the Hosmer and Lemeshow Goodness-of-fit Statistic is

$$X^2 = 12.9408$$

with

$$10 - 2 = 8 \; df, \quad p = 0.1139$$

- The H-L test indicates the model fit is OK.

Using D^2

- Suppose, in the previous model, you want to see if there is a treatment by baseline response interaction.
- You use Proc Logistic to do this:
- Reduced Model:

```
proc logistic descending;
 model y = SEX AGE TRT x / aggregate=(SEX AGE TRT x)
                           scale=1   ;
run;

/* SELECTED OUTPUT */

       Deviance and Pearson Goodness-of-Fit Statistics

Criterion        DF           Value     Value/DF     Pr > ChiSq

Deviance        158        180.6392       1.1433         0.1048
Pearson         158        186.7077       1.1817         0.0590
```

```
proc logistic descending;
 model y = SEX AGE TRT x*trt / aggregate=(SEX AGE TRT x)
                               scale=1   ;
run;

/* SELECTED OUTPUT */

       Deviance and Pearson Goodness-of-Fit Statistics

Criterion        DF           Value     Value/DF     Pr > ChiSq

Deviance        157       178.4911       1.1369         0.1153
Pearson         157       183.9152       1.1714         0.0698

Number of unique profiles: 163
```

- To calcluate D^2, Proc Logistic uses the aggregate option to figure out the number of distinct covariate patterns (strata) corresponding to the observed combinations of the covariates (SEX AGE TRT x), and uses that as the saturated model. There are 163 patterns in this dataset:

POPULATION PROFILES

Sample	SEX	AGE	TRT	X	Sample Size
1	0	22	0	1	1
...					
10	0	35	1	1	5
11	0	36	1	1	1
12	0	39	1	1	1
13	0	41	0	1	2
...					
163	1	66	0	0	1

- With 293 observations, and 163 distinct covariate patterns, we only have and average of 293/163=1.8 observations per strata, which is not enough to justify an approximate chi-square for D^2.
- However, if you want to test for no treatment by baseline response interaction, you can use the WALD STAT,

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
TRT*x	1	-0.9213	0.6297	2.1405	0.1435

- Or, the Likelihood ratio statistic,

$$D^2(SEX, AGE, TRT, X) - D^2(SEX, AGE, TRT, X, TRT * X) =$$

$$180.64 - 178.49 = 2.15 \sim \chi^2_1$$

and get almost identical results to the WALD Stat. Again, even though

$$D^2(SEX, AGE, TRT, X)$$

and

$$D^2(SEX, AGE, TRT, X, TRT * X)$$

are not approximate chi-squares, their difference is.

Additional Slides on the Saturated Model

- To see the saturated model, consider the following example based on the esophageal cancer data

```
proc means;
 class age;
  var y;
 freq count;
run;
```

- The above code calculates the average number of "1's" in the database = $\hat{p}$
- Summary

```
            Analysis Variable : y

age       N Obs                N            Mean
-----------------------------------------------
25-34       116              116       0.0086207
35-44       199              199       0.0452261
45-54       213              213       0.2159624
55-64       242              242       0.3140496
65-74       161              161       0.3416149
75+          44               44       0.2954545
```

Model Derived Estimates

- To easily estimate the model predicted probabilities, consider using the “reference” coding

```
proc logistic descending;
 class age (param=ref);
  model y = age / aggregate scale=1;
 freq count;
run;
```

- Then since each AGE category gets a dummy code, then $\hat{p}$ is equal to

$$P(Y = 1|\text{Age Group}) = \frac{e^{\alpha+\beta_{\text{Age Group}}}}{1 + e^{\alpha+\beta_{\text{Age Group}}}}$$

- Parameter Estimates

Parameter		DF	Estimate	Standard Error
Intercept		1	-0.8690	0.3304
age	25-34	1	-3.8759	1.0573
age	35-44	1	-2.1808	0.4749
age	45-54	1	-0.4203	0.3700
age	55-64	1	0.0878	0.3583
age	65-74	1	0.2129	0.3699

- For 25 - 34

$$\begin{aligned} P(Y=1|25-35) &= \frac{e^{-0.8690-3.8759}}{1+e^{-0.8690-3.8759}} \\ &= 0.008620964 \end{aligned}$$

- $\vdots$

$$\begin{aligned} P(Y=1|75+) &= \frac{e^{-0.8690}}{1+e^{-0.8690}} \\ &= 0.295462424 \end{aligned}$$

- Since the model predicted probabilities match the observed probabilities, you have a "zero degrees of freedom" test and you have "perfect fit"
- To illustrate the Goodness of Fit, we need a "non-saturated model"

Introduction of Tobacco and Alcohol Levels

- Using PROC GENMOD

```
proc genmod descending;
 class age ;
  model y = age tob alc /dist=bin link=logit;
 freq count;
run;
```

- Here, we want to observe the Deviance / DF ratio to check for Under/Over-dispersion
- Recall, for properly dispersed data (and hence a good model fit) we want the Deviance to DF ratio to be around 1
- For this model, the ratio is

```
          Criteria For Assessing Goodness Of Fit
Criterion                 DF           Value          Value/DF
Deviance                 967        710.5516            0.7348
```

- Our data appear “under dispersed” meaning that our model is predicting greater variance than observed
- This is in part an artifact since we are treating each observation as a stratum. When we calculate GOF, we need to group the data.

Parameter Estimates

- Using PROC LOGISTIC

```
proc logistic descending;
    class age (PARAM=ref) ;
 model y = age tob alc /  aggregate scale=1;
 freq count;
run;
```

- Here, we have specified SCALE=1 - This means that we are multiplying the variance-covariance matrix by a constant of 1 (i.e., the matrix is not adjusted...Now, are data appear overdispersed)
- GOF statistics

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	89.0166	80	1.1127	0.2297
Pearson	91.6997	80	1.1462	0.1748

Number of unique profiles: 88

Before Scaling, going back to GOF

- We specified “Aggregate” to calculate the change in deviance automatically
- Recall, this defines the J subtables that can be fit as a saturated model
- Thus, you can calculate the Change in deviance
- ΔG^2 is easy to calculate since you can fit a model with only the intercept and look at the increase in likelihood with the added parameters
- SCALE and AGGREGATE must be used together
- Options for SCALE are
  - 1 (or None): No scaling
  - D : Var-Cov matrix scaled by the Deviance
  - P : Var-Cov matrix scaled by Pearson’s χ^2
  - c : Any constant c, 1 is a special case

Quasi-Likelihood Estimation

- Scaling the Var-Cov matrix is a common method of correcting for poor predicted variance (overdispersion or underdispersion)
- Note that the parameter estimates are not changed by this method. However, their standard errors are adjusted for overdispersion, affecting their significance tests.
- Since we are altering the estimating equations, we are no longer calculating true “Maximum Likelihood Estimates”
- The revised estimating equations are sometimes known as “quasi-likelihood”
- Quasi-likelihood estimating equations rely on only the mean and variance specifications (as opposed to the full distribution of the outcome)
- Quasi-likelihood for the basis a variety of estimating approached included Generalized Estimating Equations
- Quasi-likelihood estimating equations perform well given large sample sizes

Scaled Estimates

- You can use either GENMOD or Logistic

```
proc genmod descending;
 class age ;
  model y = age tob alc /dist=bin link=logit aggregate scale=d;
 freq count;
run;


proc logistic descending;
    class age (PARAM=ref) ;
 model y = age tob alc /  aggregate scale=d;
 freq count;
run;
```

GENMOD Results

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF	
Deviance	80	89.0166	1.1127	
Scaled Deviance	80	80.0000	1.0000	*******
Pearson Chi-Square	80	91.7017	1.1463	
Scaled Pearson X2	80	82.4131	1.0302	
Log Likelihood		-319.2895		

Note: Value/DF now equals 1 for the scaled data - i.e., properly dispersed

Note also that V/DF for the unscaled deviance is 1.1127, which indicates, when 88 unique stratum are considered the data are over dispersed meaning the Var(Y) is greater than our predicted model

Parameter Estimates - GENMOD

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error
Intercept		1	-2.6068	0.4322
age	25-34	1	-4.6681	1.1681
age	35-44	1	-2.8131	0.5539
age	45-54	1	-1.0542	0.4438
age	55-64	1	-0.5047	0.4292
age	65-74	1	0.0317	0.4383
age	75+	0	0.0000	0.0000
tob		1	0.4094	0.0921
alc		1	0.2548	0.0265
Scale		0	1.0548	0.0000

NOTE: The scale parameter was estimated by the square root of DEVIANCE/DOF.

USING PROC LOGISTIC

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	89.0166	80	1.1127	0.2297
Pearson	91.6997	80	1.1462	0.1748

Number of unique profiles: 88

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	250.6831	7	<.0001

Note: In GENMOD, you get the “scaled estimates” for GOF, but LOGISTIC doesn’t print them. I guess since they would equal the DF, they are not that exciting to look at.

Parameter Estimates

NOTE: The covariance matrix has been multiplied by the heterogeneity factor (Deviance / DF) 1.11271.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error
Intercept		1	-2.6068	0.4322
age	25-34	1	-4.6680	1.1680
age	35-44	1	-2.8131	0.5539
age	45-54	1	-1.0542	0.4438
age	55-64	1	-0.5047	0.4292
age	65-74	1	0.0317	0.4383
tob		1	0.4094	0.0921
alc		1	0.2548	0.0265

Unscaled Estimates

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	278.9369	7	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error
Intercept		1	-2.6068	0.4097
age	25-34	1	-4.6680	1.1073
age	35-44	1	-2.8131	0.5251
age	45-54	1	-1.0542	0.4208
age	55-64	1	-0.5047	0.4069
age	65-74	1	0.0317	0.4155
tob		1	0.4094	0.0873
alc		1	0.2548	0.0251

Residuals and Regression Diagnostics

- Additional factors to assess when determining “Goodness of Fit”
- Sometimes you can look at residuals to see where the model does not fit well.
- When the responses are binary, but not binomial, a residual can be defined as

$$e_i = \frac{Y_i - \widehat{p}_i}{\sqrt{\widehat{p}_i(1-\widehat{p}_i)}}$$

- In large samples, we can replace $\widehat{p}_i$ in e_i by p_i to give

$$R_i = \frac{Y_i - p_i}{\sqrt{p_i(1-p_i)}}$$

- If the model is correctly specified, in large samples, this residual will have mean 0:

$$\begin{aligned} E(e_i) &= E\left(\frac{Y_i-\widehat{p}_i}{\sqrt{\widehat{p}_i(1-\widehat{p}_i)}}\right) \\ &\approx \left(\frac{E(Y_i)-p_i}{\sqrt{p_i(1-p_i)}}\right) \\ &= \frac{p_i-p_i}{\sqrt{p_i(1-p_i)}} = 0, \end{aligned}$$

and variance:

$$\begin{aligned} Var(e_i) &= Var\left(\frac{Y_i-\widehat{p}_i}{\sqrt{\widehat{p}_i(1-\widehat{p}_i)}}\right) \\ &\approx \frac{Var[Y_i-p_i]}{p_i(1-p_i)} \\ &= \frac{p_i(1-p_i)}{p_i(1-p_i)} = 1, \end{aligned}$$

- If the true model is $E(Y_i) = \pi_i \neq p_i$, then the residual has mean

$$\begin{aligned} E(e_i) &\approx \left(\frac{E(Y_i) - p_i}{\sqrt{p_i(1-p_i)}} \right) \\ &= \frac{\pi_i - p_i}{\sqrt{p_i(1-p_i)}} \neq 0, \end{aligned}$$

  and variance :

$$\begin{aligned} Var(e_i) &\approx \frac{Var[Y_i - p_i]}{p_i(1-p_i)} \\ &= \frac{\pi_i(1-\pi_i)}{p_i(1-p_i)} \neq 1, \end{aligned}$$

- Thus, if we just look at the simple average of these residuals,

$$n^{-1} \sum_{i=1}^{n} e_i$$

  and we do not get around 0, it gives us an idea that the model does not fit well.

- And, if we just look at the average of the squared residuals

$$n^{-1} \sum_{i=1}^{n} e_i^2$$

  and we do not get around 1, then we know that something may be wrong with the model. If the model fits, then

$$Var(e_i) \approx E(e_i^2) \approx 1$$

- However, since Y_i is always 0 or 1, regardless of the sample size (since Y_i is Bernoulli, and does not change with sample size), the observed residuals is not really that informative when the data are binary (Cox).
- Note, the ordinary residuals $Y_i - \widehat{p}_i$ are not as informative since you can show

$$\sum_{i=1}^{n}(Y_i - \widehat{p}_i) = 0$$

  as long as there is an intercept in the model.
- In a moment, we will talk about ways to look at the residuals when the data are not grouped. First we consider the grouped case.

Grouped binomials

- Suppose the data actually arise as J binomials, where y_j is the number of successes, and n_j the sample size, with

$$\widehat{p}_j = \left(\frac{e^{\hat{\boldsymbol{\beta}}' \mathbf{x}_j}}{1 + e^{\hat{\boldsymbol{\beta}}' \mathbf{x}_j}} \right),$$

- The (unadjusted) Pearson residuals

$$e_j = \left(\frac{[y_j - n_j \widehat{p}_j]}{\sqrt{n_j \widehat{p}_j (1 - \widehat{p}_j)}} \right)$$

  In large samples, we can replace $\widehat{p}_j$ by p_j, and

$$e_j = \left(\frac{[y_j - n_j p_j]}{\sqrt{n_j p_j (1 - p_j)}} \right)$$

- If the model is correctly specified, in large samples (n_j large), this residual will have mean 0:

$$E(e_i) \approx \frac{E(Y_j) - n_j p_j}{\sqrt{n_j p_j (1-p_j)}} = \frac{n_j p_i - n_j p_i}{\sqrt{p_i(1-p_i)}} = 0,$$

  and variance:

$$Var(e_i) \approx \frac{Var[Y_j - n_j p_j]}{n_j p_j (1-p_j)} = \frac{n_j p_j (1-p_j)}{\sqrt{n_j p_j (1-p_j)}} = 1,$$

- Also, Y_j will become more normal as n_j gets large, and, if the model fits

$$e_j \sim N(0,1)$$

- Then, if the model fits, we would expect the 95% of the residuals to be between -2 and 2, and to be approximately normal when plotted.

- If the true model is $E(Y_j) = n_j\pi_j \neq n_j p_j$, then the residual has mean

$$\begin{aligned}
\mu_i &= E(e_i) \\
&= \frac{E(Y_j) - n_j p_j}{\sqrt{n_j p_j (1-p_j)}} \\
&= \frac{n_j \pi_i - n_j p_i}{\sqrt{p_i (1-p_i)}} \neq 0,
\end{aligned}$$

and variance

$$\begin{aligned}
\sigma_i^2 &= Var(Y_j) \\
&= \frac{Var[Y_j - n_j p_j]}{n_j p_j (1-p_j)} \\
&= \frac{n_j \pi_j (1-\pi_j)}{n_j p_j (1-p_j)} \neq 1,
\end{aligned}$$

- Still, Y_j will become more normal as n_j gets large, but

$$e_j \sim N(\mu_i, \sigma_i^2)$$

  and plots can sometimes reveal departures from $N(0, 1)$.
- Note that, the Pearson's chi-square is

$$X^2 = \sum_{j=1}^{J} e_j^2.$$

- You can generate reams of output of regression diagnostics using the **INFLUENCE** option in the model statement

Residuals for Esophageal Data

```
data two;
 input age $ alcohol $ tobacco $ cases controls;
 n = cases + controls;
 if alcohol = '0-39'  then alc = 2;
 if alcohol = '40-79' then alc = 6;
 if alcohol = '80-119' then alc =10;
 if alcohol = '120+'   then alc =15;

 if tobacco = '0-9'   then tob = .5;
 if tobacco = '10-19' then tob =1.5;
 if tobacco = '20-29' then tob =2.5;
 if tobacco = '30+'   then tob =4.0;

 if age = '25-34'  then age1 = 1;    else age1=0;
 if age = '35-44'  then age2 = 1;    else age2=0;
 if age = '45-54'  then age3 = 1;    else age3=0;
 if age = '55-64'  then age4 = 1;    else age4=0;
 if age = '65-74'  then age5 = 1;    else age5=0;
cards;
```

```
proc logistic data=two;
 model cases/n = age1-age5 tob alc /  aggregate scale=d INFLUENCE;
 output out=dinf prob=p resdev=dr h=pii reschi=pr difchisq=difchi;
run;
```

Note: We switched to the event trials layout to reduce the output data set. Now, since we have 96 rows of data, we'll get 96 records in the outputted data set **DINF**.

Summary of Pearson Residuals

```
proc univariate data=dinf;
 var pr;
run;
```

Location		Variability	
Mean	-0.00283	Std Deviation	1.02665
Median	-0.18818	Variance	1.05401
Mode	.	Range	6.37022
		Interquartile Range	1.13606

Quantile	Estimate
100% Max	4.133526
99%	4.133526
95%	2.029202
90%	1.168725
75% Q3	0.481946
50% Median	-0.188184
25% Q1	-0.654117
10%	-1.091350
5%	-1.243294
1%	-2.236690
0% Min	-2.236690

New to SAS 9.1

```
ods graphics;
ods pdf file="lec18.pdf"
    style=journal;
proc logistic data=two;
 graphics all;
 model cases/n = age1-age5 tob alc /  aggregate scale=d INFLUENCE;
run;

ods pdf close;
ods graphics off;
```

–For PDF output, see class website (too comprehensive to include in lecture)

Collinearity and Confounding in Logistic Regression

1. Confounding

- Note, confounding can be a problem in logistic regression with many covariates.
- One covariate confounds the relationship between the response and another covariate if it is related to both the response and covariate.
- Suppose you want to control for a possible confounding factor that is not of interest itself.
- If, in the fitted model, the confounding factor is not significant, but it changes the significance and estimated odds ratio for the covariates of interest, then you should always keep the confounding factor in the model (Breslow and Day, Hosmer and Lemeshow).

2. Collinearity

- Strictly speaking, collinearity refers to correlation among the covariates.
- Thus, if there is collinearity in a dataset, there will very often also be confounding, since many of the covariates will be related to both the response and other covariates
- Sometimes, collinearity is unavoidable if we have both x_{ik} and x_{ik}^2 as a covariate (such as age and age^2), which are usually highly correlated.
- In extreme cases, where two covariates are very highly correlated, we get unstable fitted equations, symptomatic of collinearity among the covariates. When a pair of covariates are collinear, estimated coefficients may even change signs when the covariates are in the model together versus in the model separately.

- With logisitic regression, one symptom of multicollinearity is failure of the Newton-Raphson algorithm to converge, because of infinite regression coefficients, very large standard errors and/or coefficients that change wildly under different model specifications
- When there is collinearity, one should consider which of the collinear variables is most important.

Model Building Strategies

- Building logistic regression models when there are many possible covariates can be a bewildering experience.
- There can be many interactions to consider, categorical vs. continuous covariates, data transformations, etc.
- It is often useful to work hierarchically, looking at increasingly more complex structures of nested models, using test statistics (likelihood ratio, score or Wald) in deciding which covariates are important or not important in predicting response

Some helpful steps

- Often, you first look at the relationship between the response and each covariate separately:
  1) With Pearson's chi-square (or Fisher's exact test) for a categorical covariate.
  2) If the covariate is ordered or continuous, you often look at the simple logistic regression

$$\text{logit}(p) = \beta_0 + \beta_1 x$$

  and test for significance of β_1 in the regression (or, use an exact ordinal test).
- These relationships are sometimes called 'univariate' (one covariate) or 'bivariate' (1 response versus 1 covariate) relationships.

Example–Arthritis Clinical Trial

- We are interested in seeing how the binary response

$$Y = \begin{cases} 1 \text{ if good at 13 weeks} \\ 0 \text{ if poor at 13 weeks} \end{cases}.$$

is affected by the covariates,
1. BASELINE self-assessment:

$$X_1 = \begin{cases} 1 \text{ if good at BASELINE} \\ 0 \text{ if poor at BASELINE} \end{cases}.$$

2. AGE IN YEARS,
3. GENDER

$$\text{SEX} = \begin{cases} 1 \text{ if male} \\ 0 \text{ if female} \end{cases}.$$

4. TREATMENT

$$\text{TRT} = \begin{cases} 1 \text{ if auranofin} \\ 0 \text{ if placebo} \end{cases}.$$

Univariate Relationships

Covariate	% GOOD	DF	Pearson's Chi-Square	P-value
Baseline				
GOOD	87.50	1	22.849	0.000
POOR	63.44			
AGE				
21-47	80.61	2	3.636	0.162
48-56	86.08			
57-66	75.00			
GENDER				
MALE	80.75	1	0.382	0.536
FEMALE	77.50			
TREATMENT				
AURANOFIN	84.93	1	4.648	0.031
PLACEBO	74.83			

- Treatment and Baseline response are significant, whereas the other two are not.
- Instead of continuous age, we formed three age groups (based on terciles)

Category	Age Range
0	21-47
1	48-56
2	57-66

- Alternatively, we could have ran a logistic regression with continuous age as a covariate.

Multivariate Models

- After carefully looking at the univariate analyses as a screening tool, you can run a 'multivariate' analyses including all covariates thought to be important from the univariate analyses
- Hosmer and Lemeshow recommend including any covariate in the multivariate analyses which had $p-$value less than .25 in a univariate analysis.
- Hosmer and Lemeshow also recommend including other variables known to be important (treatment, exposure variables, possible confounding variables, etc.) that may not be significant.

- Consider the importance of each covariate in the model, and look to see whether some could be deleted or others need to be added (via test statistics, usually).
- Once you feel ‘close’ to a final model, look more carefully for possible interactions, recoding of variables that might be helpful, addition of quadratic terms, or other ‘transformations’, etc.
- If possible, meet with an investigator on the subject matter to see if the model is biologically plausible.

Notes

- There is more than ONE ‘final model’
- In complex datasets, we often will present the results of several related models.
- What is important is that you write up your model building strategy in a fair and descriptive way.
- ‘Statistical Significance’ is not the only reason to keep a covariate in the model. If a covariate is thought (and shown by others) to be a confounding variable, for the association between the exposure and disease, then it should be kept in.
- You may also be interested in $p-$values for covariates which are not significant, so you often leave them in the model to show they are not significant.

Stepwise regression

- As in linear regression, there is step-up, step-down, and stepwise regression.
- In the **Step-up Procedure:**
  1. Fit the intercept only model, or some other relatively simple model that includes only important covariates.
  2. Fit all models that add an additional covariate to the model in 1. Choose the model with the best fit out of these. If this new model fits significantly better (say, if the $p-$value for the extra covariate is less than .05), keep this covariate in. If not, you might stop with the current model.
  3. Repeat these steps until you want to stop.

- In the **Step-Down Procedure:**
  A step-down procedure starts with a very complex model and then tries to delete covariates that help the least.
- You may also elect to include all of the covariate interactions; however, this may result in many required iterations to get back to a sensible model.

- Hybrid **Stepwise Method**:
  Often, people use a hybrid method, trying to step-up or step down in tandem at each step.
- Other variable selection techniques are available. For example, Hosmer and Lemeshow sugggest 'best subsets' regression as an alternative strategy.
- Although it is done often (including by me), letting the computer select you final model by one of these stepwise procedures can lead to a biological implausbile model just by chance.
- You are doing so many tests in the stepwise procedure that you may blow your α-level out of the water.
- Stepwise regression is most appropriate for 'explaratory analyses' to see what relationships are going on in the data, which are to be proved in a later confirmatory study.
- Using the arthitis dataset, we will look at step-up, step down, and step-wise logistic regression to determine the best predictors. In the SAS output below, any predictor not significant at $\alpha = .05$ (the SAS default) was not kept in the model.

Step-Up Regression

```
proc logistic data=arth descending;
 model y = SEX AGE TRT x   / selection=forward ;  /*forward=step-up */
run;

/*  Selected Output */

Forward Selection Procedure

        Summary of Forward Selection Procedure

        Variable    Number       Score         Pr >
Step    Entered         In    Chi-Square    Chi-Square

   1    X                1       22.8493        0.0001
   2    TRT              2        5.2597        0.0218


NOTE: No (additional) variables met the 0.05 significance level for entry
      into the model.
```

Step-Down Regression

```
proc logistic data=arth descending;
 model y = SEX AGE TRT x   / selection=backward; /*backward=step-down */
run;
```

Summary of Backward Elimination Procedure

Step	Variable Removed	Number In	Wald Chi-Square	Pr > Chi-Square
1	AGE	3	0.1361	0.7122
2	SEX	2	0.4259	0.5140

NOTE: No (additional) variables met the 0.05 significance level for removal from the model.

Step-wise Regression

```
proc logistic data=arth descending;
 model y = SEX AGE TRT x   / selection=stepwise;
run;
```

Summary of Stepwise Procedure

Step	Variable Entered	Removed	Number In	Score Chi-Square	Wald Chi-Square	Pr Chi-Square
1	X		1	22.8493	.	0.0001
2	TRT		2	5.2597	.	0.0218

Goodness-of-Fit

- Once you have come up with your 'best' model, you want to consider the goodness-of-fit of the model you have selected.
- The goodness-of-fit statistics are the same as those discussed earlier:
  1. You can fit a more general model (with interaction terms and quadratics), and see if the extra terms are significant.
  2. The Deviance or Pearson's chi-square can be used if the strata sample sizes are sufficient ($\geq 75\%$ of the $n_j \geq 10$).
  3. Hosmer and Lemeshow's statistic can be used if the strata sample sizes are small ($\geq 25\%$ of the $n_j < 10$).