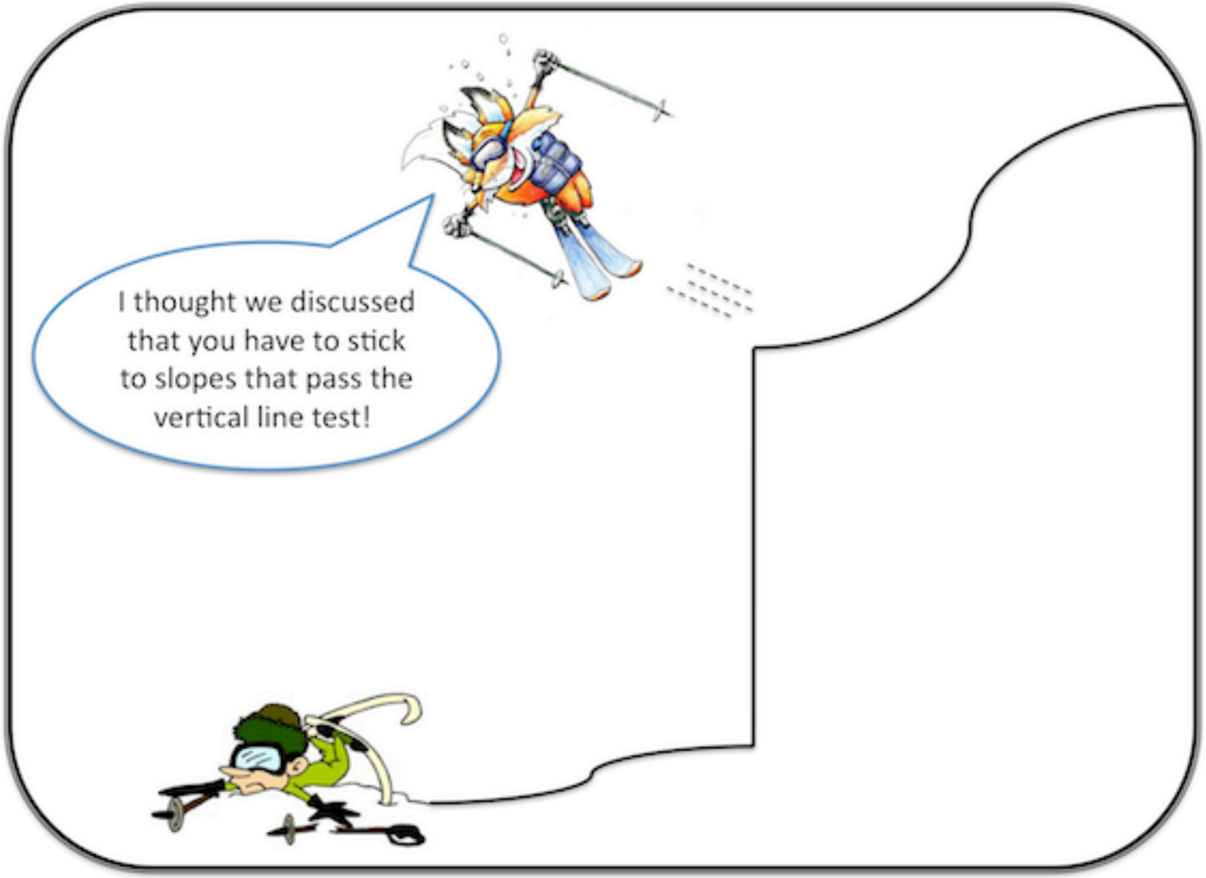
BST 210
Applied Regression Analysis



Images from: http://www.stevenscreekstriders.org/stories/2007/
and http://www.toonpool.com/cartoons/SkiTam%20Fox%20backscratcher_31874

Lecture 14
Plan for Today

- Hypothesis testing in Logistic Regression (continued)
    - Review Wald test
    - LRT (likelihood ratio test)
    - Examples

- Logistic Regression Model building
- Logistic Regression Model Assessment
    - calibration through goodness-of-fit tests

Hypothesis Testing in Logistic Regression

Recall:

Logistic Regression (think 'OR' or 'odds'):

- **<u>Wald tests:</u>**

Formal Hypothesis: $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$

Test Statistic: $Z = \frac{\hat{\beta}_j - 0}{\text{s.d.}(\hat{\beta}_j)} \sim N(0,1)$

Confidence Interval: $\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} \cdot \text{s.d.}(\hat{\beta}_j)$ (log odds ratio scale)

$\exp\{\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} \cdot \text{s.d.}(\hat{\beta}_j)\}$ (odds ratio scale)

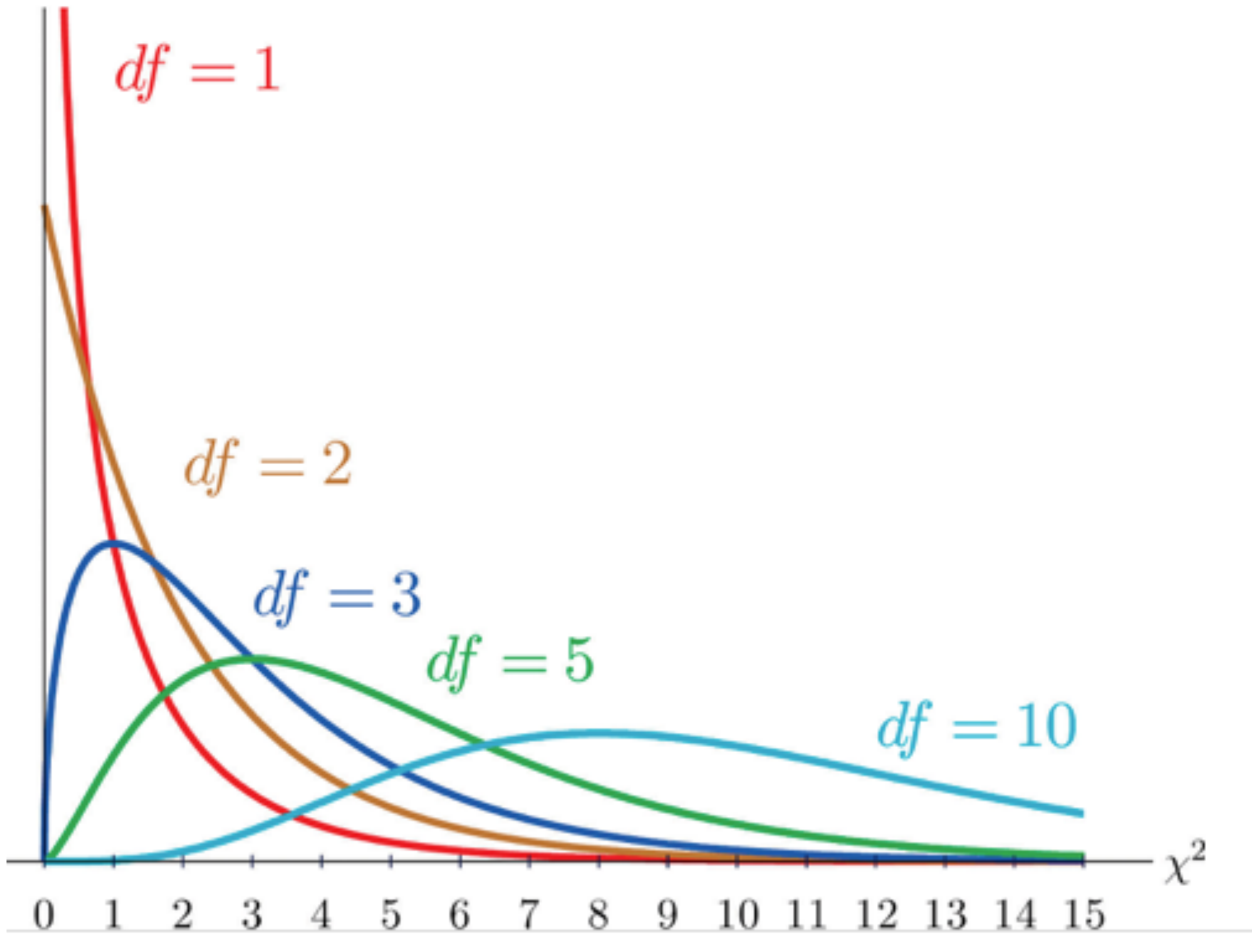
- **<u>Likelihood Ratio Tests:</u>**

The number of predictors in the full model is greater than the number in the reduced model, $p > q$.

Formal Hypothesis: $(H_0 : \text{the reduced model is sufficient})$ versus $(H_1 : \text{the full model is preferred})$

Test Statistic: $-2 \cdot \log(\frac{L_{reduced}}{L_{full}}) = -2 \cdot \log(L_{reduced}) + 2 \cdot \log(L_{full}) \sim \chi^2_{p-q}$

Recall Chi-square Distribution



Likelihood Ratio Test (LRT): Examples

- Let’s now perform Likelihood Ratio Tests on several previous examples using the surfactant study data

- Recall birth weight, which is divided into the following categories:

500-749 g	(group 1)
750-999 g	(group 2)
1000-1249 g	(group 3)
1250-1500 g	(group 4)

LRT: Continuous vs Categorical Birth Weight

- Recall that <u>we first treated birth weight as 'ordinal continuous'</u>. This approach makes the assumption that the association between birth weight and death changes linearly from one birth weight category to the next.

- We fitted the model

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

where a 1 unit increase in x corresponds to a ~250 g increase in birth weight

LRT: Continuous vs Categorical Birth Weight

- Recall that <u>we next treated birth weight as categorical</u>:

  x_2 = 1 if birth weight = group 2, 0 otherwise

  x_3 = 1 if birth weight = group 3, 0 otherwise

  x_4 = 1 if birth weight = group 4, 0 otherwise

- We fitted the model

  $$\text{logit}(p) = \alpha + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

LRT: Continuous vs Categorical Birth Weight

- Is it better to represent birth weight as a 'continuous' (ordinal) variable or as a categorical variable?

- To formally assess this, we note that the linear model is a special case of the categorical model (it is simply the categorical model with constraints on the parameter values), and thus is nested

- That is, if the relationship between logit(p) and birth weight category is linear, then in the categorical birth weight model:

$$\beta = \beta_2 = \beta_3 - \beta_2 = \beta_4 - \beta_3$$

so that

$$\beta_3 - \beta_2 = \beta_2 \quad \rightarrow \quad \beta_3 = 2\beta_2 \text{ and } \beta_4 = 3\beta_2$$

LRT: Continuous vs Categorical Birth Weight

- The linear birth weight model ('reduced') has $q = 2$ parameters (α and β) due to the constraints

- The categorical birth weight model ('full') has $p = 4$ parameters (α, β_2, β_3, β_4)

- We then can compare the two nested models:

  reduced model with $-2 \log$ likelihood $= -2 \log L_{reduced}$

  full model with $-2 \log$ likelihood $= -2 \log L_{full}$

- Note that $-2 \log$ likelihood is often referred to as the 'Deviance'

LRT: Continuous vs Categorical Birth Weight

- Let's perform the test!

```
. logistic death bwt2 bwt3 bwt4

Logistic regression                               Number of obs   =       5629
                                                  LR chi2(3)      =    1114.73
                                                  Prob > chi2     =     0.0000
Log likelihood = -2478.6223                       Pseudo R2       =     0.1836

------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bwt2 |   .2303739   .0203267   -16.64   0.000     .1937889    .2738657
        bwt3 |    .097544   .0097884   -23.19   0.000     .0801279    .1187457
        bwt4 |   .0499619   .0056048   -26.71   0.000     .0401005    .0622485
------------------------------------------------------------------------------

. estimates store A
```

LRT: Continuous vs Categorical Birth Weight

```
. logistic death birthwt

Logistic regression                               Number of obs   =       5629
                                                  LR chi2(1)      =    1079.76
                                                  Prob > chi2     =     0.0000
Log likelihood = -2496.1053                       Pseudo R2       =     0.1778

------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     birthwt |   .3539453   .0126528   -29.05   0.000      .329995    .3796338
------------------------------------------------------------------------------


. estimates store B

. lrtest A B

Likelihood-ratio test                                 LR chi2(2)  =     34.97
(Assumption: B nested in A)                           Prob > chi2 =    0.0000
```

LRT: Continuous vs Categorical Birth Weight

- Specifically, if

  H_0: (linear model is sufficient) is true,

  then

  $$\mathbf{LR = -2[\log L_{reduced} - \log L_{full}] \;\sim\; \chi^2_{p-q}}$$

  follows a chi-square distribution with [full (p) – reduced (q)] df

- We find:

  $$LR = -2(-2496.1053 - (-2478.6223)) = 34.97 \sim \chi_2^2 \quad \text{with } p < 0.001$$

- We reject H_0 -- the linear birth weight model is not adequate; it is better to treat birth weight as a categorical variable

LRT: Continuous vs Categorical Birth Weight

- Note that the odds ratios in the categorical model were 1.0, 0.23, 0.10, and 0.05 for birth weight categories 1, 2, 3, and 4, with group 1 (250-499 g) as the reference group
- There seems to be a big decrease in odds from group 1 to group 2, with smaller decreases after that
- For the linear model to hold, there should be a constant decrease in log odds (or odds) with each increase in birth weight category

LRT: Continuous vs Categorical Birth Weight

- The difference is -1.47 log units [log(0.23/1)] from category 1 to category 2,
- -0.83 log units [log(0.10/0.23)] from category 2 to category 3, and
- -0.69 log units [log(0.05/0.10)] from category 3 to category 4

* Note that the log(OR) is simply the log of the ratio of the odds

LRT: Birth Weight significant?

- Next we'll perform the LRT to assess the significance of categorical birth weight in the model with surfactant-use

- We denote the full model with surfactant and birth weight:

$$\text{logit}(p_i) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i}$$

where

x_{1i} = surfactant use (1 = yes, 0 = no)

x_{2i} = 1 if birth weight = 750-999 g, 0 otherwise

x_{3i} = 1 if birth weight = 1000-1249 g, 0 otherwise

x_{4i} = 1 if birth weight = 1250-1500g, 0 otherwise

LRT: Birth Weight significant?

- To assess the overall effect of birth weight we can perform a likelihood ratio test for the hypothesis:

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0, \beta_1 \neq 0, \text{vs.}$$

$$H_1 : \text{at least one of } \beta_2, \beta_3, \beta_4 \neq 0, \beta_1 \neq 0.$$

LRT: Birth Weight significant?

```
. logistic death surfactant bwt2 bwt3 bwt4
```

```
Logistic regression                               Number of obs   =       5629
                                                  LR chi2(4)      =    1134.94
                                                  Prob > chi2     =     0.0000
Log likelihood =  -2468.517                       Pseudo R2       =     0.1869

------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  surfactant |   .7005633    .056131    -4.44   0.000     .5987518    .8196867
        bwt2 |   .2294394   .0203125   -16.63   0.000     .1928902     .272914
        bwt3 |   .0959767   .0096688   -23.26   0.000     .0787798    .1169275
        bwt4 |   .0494499   .0055618   -26.73   0.000     .0396668    .0616457
------------------------------------------------------------------------------
```

```
. estimates store C
```

LRT: Birth Weight significant?

```
. logistic death surfactant

Logistic regression                               Number of obs   =       5629
                                                  LR chi2(1)      =      16.15
                                                  Prob > chi2     =     0.0001
Log likelihood = -3027.9091                       Pseudo R2       =     0.0027

------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  surfactant |   .7533634   .0537651    -3.97   0.000     .6550237     .866467
------------------------------------------------------------------------------

. estimates store D

. lrtest C D

Likelihood-ratio test                                 LR chi2(3)  =   1118.78
(Assumption: D nested in C)                           Prob > chi2 =    0.0000
```

LRT: Birth Weight significant?

- Specifically, if

  H_0: (surfactant model is sufficient) is true,

  then

  $$\mathbf{LR = -2[\log L_{reduced} - \log L_{full}] \sim \chi^2_{p\text{-}q}}$$

  follows a chi-square distribution with [full (5) – reduced (2)] df

- We find:

  $$LR = -2(-3027.904 - (-2468.517)) = 1118.8 \sim \chi_3^2 \quad \text{with } p < 0.001$$

- We reject H_0 -- the surfactant-only model is not adequate; it is better to include birth weight (as a categorical variable) in the full model.

LRT: Interactions significant?

- Recall previously we looked for effect modification between surfactant and (categorical) birth weight, and fit the full model (p = 8):

$$\text{logit}(p) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_2 + \beta_6 x_1 x_3 + \beta_7 x_1 x_4$$

- The terms x_1, x_2, x_3, x_4 are called *main effects* and the product terms $x_1 x_2$, $x_1 x_3$, and $x_1 x_4$ are called *interaction terms*

- The model without the interactions terms is the reduced model (q = 5):

$$\text{logit}(p) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

LRT: Interactions significant?

- We can employ a LRT to explore whether or not to include these interactions in the model, that is, to assess

H_0 : $\beta 5 = \beta 6 = \beta 7 = 0$ (ie don't need interactions)

H_1: at least one of $\beta 5, \beta 6, \beta 7 \neq 0$ (ie need interactions)

LRT: Interactions significant?

- Recall the fitted full model results:

```
. logistic death i.surfactant##i.birthwt

Logistic regression                             Number of obs   =       5629
                                                LR chi2(7)      =    1141.37
                                                Prob > chi2     =     0.0000
Log likelihood = -2465.2995                     Pseudo R2       =     0.1880

------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.surfactant |   .7003971   .0932387    -2.68   0.007     .5395481     .909198
             |
     birthwt |
    750-999  |   .2408591   .0252328   -13.59   0.000     .1961507    .2957579
  1000-1249  |   .0971421   .0113736   -19.91   0.000     .0772231     .122199
  1250-1500  |   .0435102   .0059352   -22.98   0.000     .0333027    .0568463
             |
surfactant#birthwt |
  1#750-999  |   .8331254   .1644391    -0.92   0.355     .5658528    1.226641
1#1000-1249  |   .9491598   .2196654    -0.23   0.822     .6030379    1.493943
1#1250-1500  |   1.564313   .3759617     1.86   0.063     .9766723    2.505524
             |
       _cons |   1.651724   .1228949     6.74   0.000     1.427594    1.911042
------------------------------------------------------------------------------

. estimates store modelc
```

LRT: Interactions significant?

- And the fitted reduced model results:

```
. logistic death surfactant bwt2 bwt3 bwt4

Logistic regression                               Number of obs   =       5629
                                                  LR chi2(4)      =    1134.94
                                                  Prob > chi2     =     0.0000
Log likelihood =  -2468.517                       Pseudo R2       =     0.1869

------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  surfactant |   .7005633    .056131    -4.44   0.000     .5987518    .8196867
        bwt2 |   .2294394   .0203125   -16.63   0.000     .1928902     .272914
        bwt3 |   .0959767   .0096688   -23.26   0.000     .0787798    .1169275
        bwt4 |   .0494498   .0055618   -26.73   0.000     .0396668    .0616457
       _cons |   1.651602    .109965     7.54   0.000     1.449545    1.881824
------------------------------------------------------------------------------

. estimates store modelb
```

LRT: Interactions significant?

- Then the LRT test uses 3 df, and is not significant:

```
. lrtest modelb modelc

Likelihood-ratio test                                 LR chi2(3)  =      6.44
(Assumption: modelb nested in modelc)                 Prob > chi2 =    0.0923

. test 1.surfactant#2.birthwt 1.surfactant#3.birthwt 1.surfactant#4.birthwt

 ( 1)  [death]1.surfactant#2.birthwt = 0
 ( 2)  [death]1.surfactant#3.birthwt = 0
 ( 3)  [death]1.surfactant#4.birthwt = 0

           chi2(  3) =    6.64
         Prob > chi2 =    0.0844

. * These last two lines perform first the LRT and then the Wald test to see
. * whether or not the three interaction terms are needed. Both tests use
. * 3 d.f. The LRT gives chi square = 6.44, P = 0.0923, while the Wald test
. * gives chi square = 6.64, P = 0.0844. So you get similar results either
. * way, though again the LRT may be preferred here.
. *
```

LRT: Interactions significant?

- Specifically, if

  H_0: (no interaction model is sufficient) is true,

  then

  $$\mathbf{LR = -2[\log L_{reduced} - \log L_{full}] \sim \chi_{p\text{-}q}^2}$$

  follows a chi-square distribution with [full (5) – reduced (2)] df

- We find:

  $LR = -2(-2468.517 - (-2465.2995)) = 6.44 \sim \chi_3^2$ with $p < 0.001$

- We reject H_0 -- the surfactant-only model is not adequate; it is better to include birth weight (as a categorical variable) in the full model.

Next:
Model Building and Assessment in Logistic Regression

- Model building
- Model assessment – calibration through goodness-of-fit tests, discrimination through *c*-statistics and ROC curves

Model Building in Logistic Regression

- As was true for linear regression, in the logistic regression setting, we often want to build a model based on a selection of possible covariates to include

- Decisions are made using both statistical and non-statistical considerations

- Subject matter knowledge is important

- Confounding or effect modifying variables or other adjustment factors may be well known

Model Building in Logistic Regression

- A careful univariable screening of covariates may be useful
- Automated procedures based on p-values may be helpful – forward selection, backward elimination, stepwise selection, as a starting point (each has her/his own approach perhaps)
- Criterion-based comparisons may also be helpful – AIC, BIC (*not* adjusted R^2)
- Consider parsimony/simplicity vs. more complex models

Model Building in Logistic Regression

- Carefully consider the best scale to use for continuous covariates – original value, logged, quadratic, categorized, spline, generalized additive model

- Carefully consider an appropriate number of interactions of interest

- There may be more than one “final model”

Goodness of Fit (GOF)

- The goodness of fit (GOF) of a logistic regression model is usually determined based on:
- (1) Calibration
  (want model predictions to look close to the real data)
- (2) Discrimination
  (has to do with model for cases separating well from model for controls)

Goodness of Fit (GOF)

- *What do these GOF measures tell us*? Let's discuss (1) Calibration, then (2) Discrimination:

- (1) **Calibration**

A model is *well calibrated* if observed probabilities of the outcome (based on the data) and predicted probabilities of the outcome (based on the model) are reasonably close to each other

GOF: Calibration

- There are two common methods used to assess the calibration of a logistic regression model:

  (a) Pearson chi-square goodness-of-fit test

  (for smaller number of covariate patterns)

  (b) Hosmer-Lemeshow goodness-of-fit test

  (for larger number of covariate patterns)

Calibration: (a) Pearson χ^2 GOF Test

- The data are divided into *J* possible covariate patterns (everyone in the same covariate pattern has the same *covariates*)

- Covariate pattern is independent of individual outcome

- For the surfactant data set, there are 4 birth weight groups and 2 treatment groups (surfactant yes/no); therefore, there are 8 different covariate patterns

- n_j = number of subjects with the j^{th} covariate pattern

Calibration: Pearson χ^2 GOF Test

- O_j = observed number of events (successes and failures) for the j^{th} covariate pattern

- Note that O_j should have a binomial distribution with parameters $n = n_j$ and (common) $p = p_j$ where p_j = true probability of an event for the j^{th} covariate pattern

Calibration: Pearson χ^2 GOF Test

- E_j = expected number of events (successes, say) for the j^{th} covariate pattern = $n_j \hat{p}_j$

- Here $\hat{p}_j$ is the estimated probability of an event for the j^{th} covariate pattern, predicted from the logistic regression model

- V_j = the estimated variance of O_j

  = $n_j \hat{p}_j (1 - \hat{p}_j)$

Calibration: Pearson χ^2 GOF Test

- The Pearson chi-square statistic is given by:

$$\chi^2_{\boldsymbol{Pearson}} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{J}} \frac{(\boldsymbol{O_j - E_j})^{\boldsymbol{2}}}{\boldsymbol{V_j}} \sim \chi^2_{\boldsymbol{J-(p+1)}}$$

where

p is the number of parameters in the logistic regression model, excluding the intercept

- The p-value is given by

$$\Pr(\chi^2_{J-(p+1)} > \chi^2_{Pearson}),$$

and we will say the model does not have acceptable fit if the $\chi^2_{Pearson}$ test statistic is too large.

Calibration: Pearson χ^2 GOF Test

- The null hypothesis being tested is that *the model provides a good fit to the data* (meaning that predicted probabilities are close to observed probabilities); the alternative hypothesis is that *the model does not fit well*

- Note situation with saturated model (same # parameters as covariate patterns)

- The number of covariate patterns *J* should be much less than the number of subjects *n* in order for this χ^2 test to be valid (basically visualize a 2 (success/failure) x J (covariate patterns) table of outcomes) → perhaps want 10 per cell

Example: Pearson χ^2 GOF Test

- After running a logistic regression model, we use in Stata for instance, the command:

```
. estat gof, table
```

- The output is a table with the following information for each covariate pattern:

  - The probability of an event (predicted from model)
  - Observed number of events
  - Expected number of events

Example: Surfactant Use

```
. logistic death surfactant bwt2 bwt3 bwt4
```

```
Logistic regression                               Number of obs   =       5629
                                                  LR chi2(4)      =    1134.94
                                                  Prob > chi2     =     0.0000
Log likelihood =  -2468.517                       Pseudo R2       =     0.1869

------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  surfactant |   .7005633    .056131    -4.44   0.000     .5987518    .8196867
        bwt2 |   .2294394   .0203125   -16.63   0.000     .1928902     .272914
        bwt3 |   .0959767   .0096688   -23.26   0.000     .0787798    .1169275
        bwt4 |   .0494499   .0055618   -26.73   0.000     .0396668    .0616457
------------------------------------------------------------------------------
```

- p+1 = 5 parameters are estimated (intercept and 4 slope terms), and there are J = 8 covariate patterns (4 bwt x 2 surfactant)

Example: Pearson χ^2 GOF Test

```
. estat gof, table
```

Logistic model for death, goodness-of-fit test

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
1	0.0541	40	29.7	508	518.3	548
2	0.0755	82	92.3	1141	1130.7	1223
3	0.1000	40	41.5	375	373.5	415
4	0.1368	142	140.5	885	886.5	1027
5	0.2098	78	86.8	336	327.2	414
6	0.2748	257	248.2	646	654.8	903
7	0.5364	177	177.0	153	153.0	330
8	0.6229	479	479.0	290	290.0	769

Example: Pearson χ^2 GOF Test

Group	Prob	surfac~t	bwt2	bwt3	bwt4
1	0.0541	1	0	0	1
2	0.0755	0	0	0	1
3	0.1000	1	0	1	0
4	0.1368	0	0	1	0
5	0.2098	1	1	0	0
6	0.2748	0	1	0	0
7	0.5364	1	0	0	0
8	0.6229	0	0	0	0

```
      number of observations =      5629
number of covariate patterns =         8
            Pearson chi2(3) =         6.72
                Prob > chi2 =         0.0814
```

Example: Pearson χ^2 GOF Test

- The Pearson chi-square test statistic is 6.72, which follows a χ^2 distribution with 8 – 5 = 3 degrees of freedom if H_0 is true

- Since $p = 0.0814 > 0.05$, we cannot reject the null hypothesis that the fit of the model is adequate

- Thus, calibration is considered to be acceptable, though the marginal *p*-value may have us somewhat concerned (perhaps the interaction between surfactant and birth weight group is needed, though the interaction effect is marginal)

Another Calibration χ^2 GOF Test

- If there are continuous explanatory variables (or a lot of categorical predictors) in a logistic regression model, the number of covariate patterns *J* might be large, and not much less than the number of subjects *n*

- In this case, the Hosmer-Lemeshow goodness-of-fit test is preferred over the Pearson chi-square goodness-of-fit test

Calibration: (b) Hosmer-Lemeshow χ^2 GOF Test

- x_{ik} = k^{th} covariate for the i^{th} subject, where i = 1, …, n and k = 1, …, p

- y_i = 1 if the i^{th} subject had an event and 0 otherwise

- Then we know

$$\hat{p}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip})}$$

Hosmer-Lemeshow χ^2 GOF Test

- We order the *n* subjects according to fitted or predicted probabilities of the outcome (based on the model), so that

$$\hat{p}_1 \leq \hat{p}_2 \leq \hat{p}_3 \leq \ldots \leq \hat{p}_n$$

- We then group the data into *G* approximately equal sized groups based on the $\hat{p}_i$ (ordered)

Hosmer-Lemeshow χ^2 GOF Test

- *G* is usually chosen to be 10 to create *deciles* of increasing risk

$$O_j = \sum_{i \in group_j} y_i$$ = observed number of events in group *j*

$$E_j = \sum_{i \in group_j} \hat{p}_i$$ = expected number of events in group *j*

$$\bar{p}_j = E_j / n_j$$ = probability of an event in group *j*

n_j = number of subjects in group *j* where *j* = 1, ..., *G*

Hosmer-Lemeshow χ^2 GOF Test

- The test statistic is then

$$X_{HL}^2 = \sum_{j=1}^{G} \frac{\left(O_j - E_j\right)^2}{n_j \bar{p}_j (1 - \bar{p}_j)} \sim \chi_{G-2}^2$$

under the null hypothesis that the model provides a good fit to the data

- If we reject H_0, then calibration is not adequate for the model

Hosmer-Lemeshow χ^2 GOF Test

- The degrees of freedom are based on simulation results and do not depend on the number of parameters estimated

- In general, G should be greater than or equal to 6 (preferably 10) and n_j should be large

- After running the logistic regression model, we use the Stata command:

```
. estat gof, group(10) table
```

Example: Surfactant Use

```
. estat gof, group(10) table

Logistic model for death, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)
  (There are only 5 distinct quantiles because of ties)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 |  Exp_0 | Total |
  |-------+--------+-------+-------+-------+--------+-------|
  |     3 | 0.0755 |   122 | 122.0 |  1649 | 1649.0 |  1771 |
  |     5 | 0.1368 |   182 | 182.0 |  1260 | 1260.0 |  1442 |
  |     6 | 0.2098 |    78 |  86.8 |   336 |  327.2 |   414 |
  |     8 | 0.2748 |   257 | 248.2 |   646 |  654.8 |   903 |
  |    10 | 0.6229 |   656 | 656.0 |   443 |  443.0 |  1099 |
  +--------------------------------------------------------+

       number of observations =      5629
             number of groups =         5
      Hosmer-Lemeshow chi2(3) =         1.58
                  Prob > chi2 =         0.6648
```

Example: Surfactant Use

- We are unable to reject the null hypothesis that the fit of the model is adequate (*p* = 0.66)

- Once again, the model seems to be well calibrated

- ** Note here that the Pearson GOF test is preferred, because we don't even have 10 different covariate patterns. The Hosmer-Lemeshow GOF test is preferred when you have a *larger* number of covariate patterns.

Criticisms of GOF Tests

- Often there are a number of different models that have adequate fit by some goodness-of-fit test, so this does not help us in selecting between models (unless we can reject some models because of significant *p*-values in the goodness-of-fit test)

- In particular, one model may show adequate fit, but there could be numerous other covariates that could be included into the model with statistically significant *p*-values

- So more focus on model selection is really needed

GOF: Discrimination

→ Next we consider our 2nd approach to GOF testing in logistic regression outside of calibration -

- (2) **Discrimination**

A model has *good discrimination* if the distribution of predicted probabilities of the outcome for cases (with the outcome) and controls (without the outcome) separate out

Cases tend to have higher probabilities of the outcome, controls tend to have lower probabilities, and there is not too much overlap between the two distributions

Ahead:

- Goodness of Fit for Logistic Regression Models continued…

    - Discrimmination

    - ROC/AUC

    - pseudo R^2

- Proportional odds model for *ordinal responses*!

- Multinomial model for *nominal responses*!