Lecture 5: *Goodness of Fit*

Lin Chen

Department of Public Health Sciences
The University of Chicago

Goodness of Fit Measures: Motivation

- Thus far, we have been looking at models for binary outcomes that reproduce results from the simple 2x2 table. As we move into more complex models (multiple predictors), we want to assess how well a model describes the data.
- To assess the 'goodness-of-fit' of a model, we need a summary statistic that measures the 'closeness' of observed binomial proportions, y_i/n_i to the estimated (fitted) proportions, $\hat{p}_i$
- The likelihood function summarizes the information that the data provide about the unknown parameters in a model of interest.
- Thus it's natural to utilize the likelihood to assess how well the model performs based on the estimated $\hat{p}_i$. model fit v.s. efficiency

· least amount of predictors with the best prediction values.

Deviance: notation and definition

1. $\hat{L}_c$: the maximized likelihood (likelihood given the MLE) under the **c**urrent model of interest
2. reduced
   $\hat{L}_f$: the maximized likelihood under the model fits the data perfectly, which is termed the *full model* or *saturated model* "ruler"
3. "negative #"
   Deviance: $D = -2\log(\hat{L}_c/\hat{L}_f) = -2\left(\log\hat{L}_c - \log\hat{L}_f\right)$ relative to the full model to compare.
   ×2: approx. to χ^2 distribution   Reduced − Full

- Deviance D measures the extent to which the current model deviates from the full model, and it can be used to assess model fit
- Deviance statistic is a likelihood ratio statistic (approx. χ^2 and positive number)
- Large D when $\hat{L}_c$ is small relative to $\hat{L}_f$, indicating the current model is poor. ↓ d between current and full model
- Small D when $\hat{L}_c$ is similar to $\hat{L}_f$, indicating the current model is a good one. $\simeq \mathbb{Q}$

"Deviance" Away from the Original

Deviance: formula

- Recall: the likelihood function for observations $y_i/n_i,\ i = 1, 2, ..., n$ for n groups with unknown p_i is

$$L = \prod_{i=1}^{n} \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i} \tag{1}$$

$$\log L = \sum_{i=1}^{n} \{\log \binom{n_i}{y_i} + y_i \log p_i + (n_i - y_i) \log(1 - p_i)\} \tag{2}$$

- Let $\hat{p}_i,\ i = 1, ..., n$ be the fitted values under *current model*, then

$$\log \hat{L}_c = \sum_{i=1}^{n} \{\log \binom{n_i}{y_i} + y_i \log \hat{p}_i + (n_i - y_i) \log(1 - \hat{p}_i)\} \tag{3}$$

- Define $\tilde{p}_i = y_i/n_i,\ i = 1, ..., n$ which is the fitted probabilities under the *full model*, then

$$\log \hat{L}_f = \sum_{i=1}^{n} \{\log \binom{n_i}{y_i} + y_i \log \tilde{p}_i + (n_i - y_i) \log(1 - \tilde{p}_i)\} \tag{4}$$

Deviance: formula (continued)

- The *Deviance* is then given by

Comparing with "perfect" : Reduced − Full

$$\boxed{D = -2\{\log \hat{L}_c - \log \hat{L}_f\}}$$
$$= 2\sum_{i=1}^{n}\{y_i \log(\frac{\tilde{p}_i^{\text{ full}}}{\hat{p}_i \text{ reduced}}) + (n_i - y_i)\log(\frac{1-\tilde{p}_i^{\text{ full}}}{1-\hat{p}_i \text{ reduced}})\} \tag{5}$$

- Deviance statistics are used to assess the goodness of fit of the current model by comparing the estimated $\hat{p}$ of the current model versus the $\tilde{p}$ from the full model
  - The full model estimated the probability of event at each unique/possible X value or X-combination (if there are multiple X's)
  - The current model of interest could use fewer parameters (fewer X's)
  - We could contrast models by deviance statistics

Deviance: The distribution of the deviance statistic, D

- We use deviance D to evaluate the current model, and so we need to know its distribution.
- Under H_0 : the current model fit is not different from the full model fit — no additional parameters are needed to provide a better fit. As the groups size $n_i \longrightarrow \infty$ (not group number n), D converges to χ^2_{n-p}, where $p = \#$ parameter (including intercept), $n = \#$ group.

  c ≈ f : good
- For grouped binary data with reasonably large-sized groups, the deviance provides a goodness of fit test for the model, and $D \sim \chi^2_{n-p}$ approximately.

  categorical
  - You may have wondered why in calculating deviance, log likelihood difference is multiplied by a factor of 2, $D = -2(\log \hat{L}_c - \log \hat{L}_f)$. This is because it is shown to be χ^2_{n-p} when the size of each group is large.
  - A larger D (small P-value) implies a significant difference between the current model and the full model, i.e., a bad fit.
  - Since mean of a χ^2_{n-p} variable is $\overset{\text{dof}}{n-p}$, a useful rule of thumb is the D is around $n-p$, the model may be satisfactory.

Deviance: The distribution of the deviance statistic, D

When X is a continuous variable,

- For each X value, there is only one response (0/1) for Y. That is, Y is Bernoulli distributed and each sample has a unique Log p_i. In this case, $n_i = 1$ for all i, we call this case as continuous ungrouped binary data.
- The likelihood function is $L = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$. Then $\log \hat{L}_f = 0$, and D depends only on the fitted model $\hat{p}_i$, $D = -2\sum_{i=1}^{n}\{\hat{p}_i \text{logit}(\hat{p}_i) + \log(1-\hat{p}_i)\}$.

  Y Binary Data:

  grouped data $D \sim \chi^2_{n-p}$
  ungrouped data D not $\neq$ χ^2

- The value of a single likelihood is meaningless in isolation, and is only meaningful in comparing likelihoods. Since **the full model likelihood is 0** no additional predictor **for ungrouped binary data , the deviance is uninformative** no reference to compare about the goodness of fit of a model. $D \neq$ not χ^2
- For ungrouped binary data, D is not even approximately χ^2.
- Even when the n_i all exceed unity (1), the χ^2 approximation may not be particularly good when the data are sparse, i.e., some n_i being very small.

We will revisit this case and introduce an alternative test later.

Example 1: *Aircraft fasteners* – Grouped binary data / Binomial response

This is a study on the compressive strength of an alloy fastener used in the construction of aircraft. This table displays the number of fasteners failing out of a number subjected to varying pressure loads.

```
. list

     +-----------------------+
     | load   ntotal   nfail |
     |-----------------------|
  1. | 2500       50      10 |
  2. | 2700       70      17 |
  3. | 2900      100      30 |
  4. | 3100       60      21 |
  5. | 3300       40      18 |
     |-----------------------|
  6. | 3500       85      43 |
  7. | 3700       90      54 |
  8. | 3900       50      33 |
  9. | 4100       80      60 |
 10. | 4300       65      51 |
     +-----------------------+
```

10 groups

Example 1: *Aircraft fasteners* – Checking model fit

- A model with separate parameters for each PSI value could be considered a full model, in that it allows flexibility in how odds of failure increases relative to 2500 (baseline value) PSI.

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \ldots + \beta_9 X_9$$

  This model will reproduce probabilities and odds ratios for each psi level against baseline

- Another model we may consider has a single predictor (PSI) implying linear increase in units of PSI. β_0: log-odds when pressure = 0 psi

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

  This model is restrictive as to the trend in failure risk over psi - must be linear, i.e., equal increment in log odds per PSI increase

- One can compute the deviance statistic and determine if the second model (fewer parameter) is adequate

Example 1: *Aircraft fasteners: full/saturated model*

categorical

```
. blogit nfail ntotal i.load

Logistic regression for grouped data              Number of obs   =        690
                                                  LR chi2(9)      =     112.83
                                                  Prob > chi2     =     0.0000
Log likelihood =    -421.67                       Pseudo R2       =     0.1180

------------------------------------------------------------------------------
    _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        load |
       2700  |   .2492157   .4502127     0.55   0.580    -.6331849    1.131616
       2900  |   .5389965   .4154745     1.30   0.195    -.2753185    1.353311
       3100  |   .7672551    .445264     1.72   0.085    -.1054464    1.639957
       3300  |   1.185624   .4754052     2.49   0.013     .2538465    2.117401
       3500  |   1.409825   .4148076     3.40   0.001     .5968169    2.222833
       3700  |   1.791759   .4138796     4.33   0.000     .9805704    2.602948
       3900  |   2.049588   .4627381     4.43   0.000     1.142638    2.956538
       4100  |   2.484907   .4377975     5.68   0.000     1.626839    3.342974
       4300  |   2.679063   .4647972     5.76   0.000     1.768077    3.590048
             |
       _cons |  -1.386294   .3535534    -3.92   0.000    -2.079246   -.6933424
------------------------------------------------------------------------------
```

relative to 2500

This model has 9 ORs for contrasts with the reference group '2500 PSI'.
There is a monotonic increase in failure odds over increasing PSI.

Example 1: *Aircraft fasteners: full/saturated model* (Alternative way)

If x are non-integers

Generate indicator variables (dummy variables) for each category and choose your own reference group (as the omitted predictor)

```
. tabulate load,generate(g)

       load |      Freq.     Percent        Cum.
------------+-----------------------------------
       2500 |          1       10.00       10.00
       2700 |          1       10.00       20.00
       2900 |          1       10.00       30.00
       3100 |          1       10.00       40.00
       3300 |          1       10.00       50.00
       3500 |          1       10.00       60.00
       3700 |          1       10.00       70.00
       3900 |          1       10.00       80.00
       4100 |          1       10.00       90.00
       4300 |          1       10.00      100.00
------------+-----------------------------------
      Total |         10      100.00
```

Example 1: *Aircraft fasteners: full/saturated model* (an alternative way)

```
. blogit nfail ntotal g2-g10

Logistic regression for grouped data              Number of obs   =        690
                                                  LR chi2(9)      =     112.83
                                                  Prob > chi2     =     0.0000
Log likelihood =    -421.67                       Pseudo R2       =     0.1180

------------------------------------------------------------------------------
    _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          g2 |   .2492157   .4502127     0.55   0.580    -.6331849    1.131616
          g3 |   .5389965   .4154745     1.30   0.195    -.2753185    1.353311
          g4 |   .7672551    .445264     1.72   0.085    -.1054464    1.639957
          g5 |   1.185624   .4754052     2.49   0.013     .2538465    2.117401
          g6 |   1.409825   .4148076     3.40   0.001     .5968169    2.222833
          g7 |   1.791759   .4138796     4.33   0.000     .9805704    2.602948
          g8 |   2.049588   .4627381     4.43   0.000     1.142638    2.956538
          g9 |   2.484907   .4377975     5.68   0.000     1.626839    3.342974
         g10 |   2.679063   .4647972     5.76   0.000     1.768077    3.590048
       _cons |  -1.386294   .3535534    -3.92   0.000    -2.079246   -.6933424
------------------------------------------------------------------------------
```

Same result. The log likelihood value for this model is -421.67

Example 1: *Aircraft fasteners – The current model*

- *The current model of interest:* $\text{logit}(p) = \beta_0 + \beta_1 X$, X is the predictor variable load.

continuous

```
. blogit nfail ntotal load

Logistic regression for grouped data              Number of obs   =        690
                                                  LR chi2(1)      =     112.46
                                                  Prob > chi2     =     0.0000
Log likelihood = -421.85596                       Pseudo R2       =     0.1176

------------------------------------------------------------------------------
    _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        load |   .0015484   .0001575     9.83   0.000     .0012397    .0018572
       _cons |  -5.339711   .5456932    -9.79   0.000    -6.409251   -4.270172
------------------------------------------------------------------------------
```

The log likelihood value for this model is -421.86

Example 1: *Aircraft fasteners* – Contrasting two models

- *Contrast the two models:* calculate the deviance

$$D = -2\{\log \hat{L}_c - \log \hat{L}_{f}\}_{10} = -2(\overset{\text{current}}{-421.86} - (\overset{\text{full}}{-421.67})) = 0.37$$

- The deviance is 0.37, with $(\overbrace{10}^{\beta_1 \sim \beta_9 + \beta_0} - \overbrace{2}^{\beta_1 + \beta_0}) = 8$ degrees of freedom, it's nowhere near significant at any conventional significance level, thus there is no evidence of lack of fit in the current model of interest. The current model has a good fit.

```
. display chi2tail(8, 0.37192)
.99995704
```

p-value

↓

Fail to Reject H_0 : no statistical difference

good!

Some other goodness of fit statistics: *Pearson's χ^2-statistic*

- *Pearson's χ^2-statistic:* $\boxed{\chi^2 = \sum_{i=1}^{n} \frac{(y_i - n_i \hat{p}_i)^2}{n_i \hat{p}_i (1-\hat{p}_i)}}$
- With grouped binary data, the Pearson's χ^2-statistics have the same asymptotic χ^2 distribution under H_0 that the *current model* has a good fit. Those two values will generally differ but with little practical importance.
- Same as deviance, the statistic also cannot be used as a goodness of fit test for ungrouped binary data.

The `glm` function in Stata

The `glm` function in Stata works similarly as `blogit` but provides some different types of output

```
. glm nfail load, family(binomial ntotal)

Iteration 0:   log likelihood = -22.544257
Iteration 1:   log likelihood = -22.544211
Iteration 2:   log likelihood = -22.544211

Generalized linear models                         No. of obs      =         10
Optimization     : ML                             Residual df     =          8
                                                  Scale parameter =          1
Deviance         =  .3719169146                   (1/df) Deviance =   .0464896
Pearson          =  .3706630524                   (1/df) Pearson  =   .0463329

Variance function: V(u) = u*(1-u/ntotal)          [Binomial]
Link function    : g(u) = ln(u/(ntotal-u))        [Logit]

                                                  AIC             =   4.908842
Log likelihood   = -22.54421068                   BIC             =  -18.04876

------------------------------------------------------------------------------
             |                 OIM
       nfail |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        load |   .0015484   .0001575     9.83   0.000     .0012397    .0018572
       _cons |  -5.339712   .5456932    -9.79   0.000    -6.409251   -4.270172
------------------------------------------------------------------------------
```

```
. glm nfail load, family(binomial ntotal)

Iteration 0:   log likelihood = -22.544257
Iteration 1:   log likelihood = -22.544211
Iteration 2:   log likelihood = -22.544211

Generalized linear models                     No. of obs      =         10
Optimization     : ML                         Residual df     =          8
                                              Scale parameter =          1
Deviance         =  .3719169146               (1/df) Deviance =   .0464896
Pearson          =  .3706630524               (1/df) Pearson  =   .0463329

Variance function: V(u) = u*(1-u/ntotal)      [Binomial]
Link function    : g(u) = ln(u/(ntotal-u))    [Logit]

                                              AIC             =   4.908842
Log likelihood   = -22.54421068               BIC             =  -18.04876

------------------------------------------------------------------------------
             |                 OIM
       nfail |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        load |   .0015484   .0001575     9.83   0.000     .0012397    .0018572
       _cons |  -5.339712   .5456932    -9.79   0.000    -6.409251   -4.270172
------------------------------------------------------------------------------
```

Another goodness of fit statistic: the *Hosmer-Lemeshow statistic* for any binary data

- This statistic is a measure of the goodness of fit of a model that can be used in modelling *any* binary outcome data, including **ungrouped binary data**, regardless of the form of X.
- The basic idea is to use *predicted probabilities* to create groups, then compute expected counts of successes for each group by summing the predicted values, and compare these with observed values using Pearson's chi-squared statistic.

  "Artificially" create
- To create groups, the binary observations are first arranged in ascending order of their corresponding fitted probabilities, the ordered values are then formed into groups of similar size based on quantiles of the fitted probabilities, or pre-determined intervals.
- Hosmer and Lemeshow recommend creating groups based on deciles of their predicted probabilities.

Some other goodness of fit statistics: *Hosmer-Lemeshow statistic* (continued)

- Suppose there are m_i observations in the i^{th} of g groups, where the observed number of successes is o_i, and the corresponding estimated expected number is e_i, the Hosmer-Lemeshow statistic is then given by

$$X^2_{HL} = \sum_{i=1}^{g} \frac{(o_i - e_i)^2}{e_i(1 - e_i/m_i)} \tag{6}$$

- Using simulation studies, this statistic has been shown to have an approximate χ^2_{g-2}.

→ # of groups, depends on how many groups one chooses

- This statistic is an informal guide as to the adequacy of the model, not a rigid test.

Example 2: ESR data – ungrouped binary data

Determination of the ESR: the data (ungrouped binary data with a continuous predictor) were obtained in order to study the extent to which the disease state of an individual, reflected in the ESR (the erythrocyte sedimentation rate) reading, is related to the level of a plasma protein, fibrinogen. The outcome is whether each individual has an ESR reading greater than 20 (implying inflammation). There are 32 observations each with a unique fib value.

Cannot use Deviance:

- No ref. to compare
- $L_f = 0$, meaningless
- Not χ^2 distributed

```
. list in 8/15

     +--------------------+
     |  indivi~l   fib  y |
     |--------------------|
  8. |         8  2.21  0 |
  9. |         9  3.15  0 |
 10. |        10   2.6  0 |
 11. |        11  2.29  0 |
 12. |        12  2.35  0 |
     |--------------------|
 13. |        13  5.06  1 |
 14. |        14  3.34  1 |
 15. |        15  2.38  1 |
     +--------------------+
```

A linear logistic odel: $\text{logit}(p) = \beta_0 + \beta_1 X$

This model assumes that the fib value is linearly related to the log odds of event (ESR >20).

```
. logit y fib , nolog

Logistic regression                             Number of obs     =         32
                                                LR chi2(1)        =       6.04
                                                Prob > chi2       =     0.0139
Log likelihood = -12.420178                     Pseudo R2         =     0.1957

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         fib |   1.827081   .9008558     2.03   0.043     .0614358    3.592726
       _cons |  -6.845075   2.770287    -2.47   0.013    -12.27474   -1.415412
------------------------------------------------------------------------------
```

Common Mistake:

$p<0.05$, but may still be bad fit.

Assess "goodness of fit."

Calculation of the H-L statistic

The Stata command `estat` will calculate post-estimation statistics and is used after running regressions).
First, let's try create 10 groups to calculate gof (HL) statistics

"goodness of fit"

```
. estat gof, group(10)
```
(Default is 10 groups)

```
Logistic model for y, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
       number of observations =        32
             number of groups =        10
      Hosmer-Lemeshow chi2(8) =        10.51
                  Prob > chi2 =         0.2307
```
$\rightarrow g-2$

Fail to Reject, good fit.

Now try 8 groups and calculate the gof statistic – similar results

```
. estat gof, group(8)

Logistic model for y, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
       number of observations =        32
             number of groups =         8
      Hosmer-Lemeshow chi2(6) =         8.98
                  Prob > chi2 =         0.1746
```

- The fitted model is satisfactory.

Summary

Assessing the goodness of fit

- Statistics that compare observed to predicted events form a natural evaluation of model fit
  - **Deviance** is used to assess model fit for grouped binary data. It is distributed as χ^2_{n-p} under the null of no difference in fit (current versus full) when group sizes (n_i) are moderate to large.
  - Pearson's χ^2-statistic (idea is similar to deviance)
  - HL statistic can be used on ungrouped binary data with a continuous X, but is less formal
- Model fit must be balanced against complexity, as in all regression models.
- Next: Alternate model comparison, additional diagnostics