Assessing Calibration of Logistic Regression Models: Beyond the Hosmer-Lemeshow Goodness-of-Fit Test

Conservatoire National des Arts et Métiers
February 16, 2018

Stan Lemeshow
College of Public Health
The Ohio State University

THE OHIO STATE UNIVERSITY COLLEGE OF PUBLIC HEALTH

Global significance. Local impact.

Some Quotes:

All models should be as simple as possible...but no simpler

- Albert Einstein

All models are wrong...but some are useful

- George Box

When all you have is a hammer everything looks like a nail

- Abraham Maslow

If I had only one hour to live, I'd spend it at a statistics seminar…that way it would seem longer

- Anonymous

What I want to Cover With You Today:

- Binary Logistic Regresion
- Assessing Calibration
  - Hosmer-Lemeshow GOF Test (*g* = 10 groups)

    Problem: Under large sample sizes, the test tends to reject models that deviate only slightly from the true model. Models that deviate slightly from the true model are acceptable in practice and ideally would not be rejected.

    Possible Solution 1: To reduce power, some have proposed applying the test to smaller subsets of data but the method has not been formalized.

    Possible Solution 2:Increase the number of groups when *n* is large
  - Calibration Bands

LOGISTIC REGRESSION ANALYSIS

GOAL: To find the best fitting, simplest, model possible describing the relationship between an outcome (dependent or response) variable and a set of independent (predictor or explanatory) variables.

or "covariates".

What distinguishes a logistic regression model from the linear regression model is that the outcome variable is binary (or dichotomous).

Example:

AGE (yrs) and presence or absence of evidence of significant coronary heart disease (CHD) for 100 subjects selected to participate in a study.

ID	AGE	CHD	ID	AGE	CHD	ID	AGE	CHD	ID	AGE	CHD
1	20	0	26	35	0	51	44	1	76	55	1
2	23	0	27	35	0	52	44	1	77	56	1
3	24	0	28	36	0	53	45	0	78	56	1
4	25	0	29	36	1	54	45	1	79	56	1
5	25	1	30	36	0	55	46	0	80	57	0
6	26	0	31	37	0	56	46	1	81	57	0
7	26	0	32	37	1	57	47	0	82	57	1
8	28	0	33	37	0	58	47	0	83	57	1
9	28	0	34	38	0	59	47	1	84	57	1
10	29	0	35	38	0	60	48	0	85	57	1
11	30	0	36	39	0	61	48	1	86	58	0
12	30	0	37	39	1	62	48	1	87	58	1
13	30	0	38	40	0	63	49	0	88	58	1
14	30	0	39	40	1	64	49	0	89	59	1
15	30	0	40	41	0	65	49	1	90	59	1
16	30	1	41	41	0	66	50	0	91	60	0
17	32	0	42	42	0	67	50	1	92	60	1
18	32	0	43	42	0	68	51	0	93	61	1
19	33	0	44	42	0	69	52	0	94	62	1
20	33	0	45	42	1	70	52	1	95	62	1
21	34	0	46	43	0	71	53	1	96	63	1
22	34	0	47	43	0	72	53	1	97	64	0
23	34	1	48	43	1	73	54	1	98	64	1
24	34	0	49	44	0	74	55	0	99	65	1
25	34	0	50	44	0	75	55	1	100	69	1

The model we will use is the logistic regression model. We choose this because

(1) from a mathematical point of view, it is an extremely flexible and easily used function and

(2) it lends itself to a biologically meaningful interpretation

Let $\pi(x)$ = conditional mean of y given x.

Specifically,

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

A transformation of $\pi(x)$ that will be central to our study of logistic regression is the logit transformation. This is defined as

$$g(x) = \ln\left\{\frac{\pi(x)}{1-\pi(x)}\right\}$$

but $$1-\pi(x) = 1 - \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{1}{1+e^{\beta_0+\beta_1 x}}$$

hence

$$g(x) = \ln\left\{\frac{\dfrac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}{\dfrac{1}{1+e^{\beta_0+\beta_1 x}}}\right\} = \ln\left\{e^{\beta_0+\beta_1 x}\right\}$$

$$= \beta_0 + \beta_1 x$$

Let us now go back to the AGE/CHD data. Use of a logistic regression routine, such as the one in Stata, produces the following output:

```
. logit CHD AGE

Iteration 0:   log likelihood = -68.331491
Iteration 1:   log likelihood = -54.170558
Iteration 2:   log likelihood = -53.681645
Iteration 3:   log likelihood = -53.676547
Iteration 4:   log likelihood = -53.676546

Logit estimates                                   Number of obs   =        100
                                                  LR chi2(1)      =      29.31
                                                  Prob > chi2     =     0.0000
Log likelihood = -53.676546                       Pseudo R2       =     0.2145

------------------------------------------------------------------------------
         CHD |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         AGE |    .1109211   .0240598     4.610   0.000      .0637647    .1580776
       _cons |   -5.309453   1.133655    -4.683   0.000     -7.531376   -3.087531
------------------------------------------------------------------------------
```

Hence $\hat{\beta}_0 = -5.30950$ and $\hat{\beta}_1 = 0.11092$

fitted values are given by $\hat{\pi}(x) = \dfrac{e^{-5.31+0.11x}}{1+e^{-5.31+0.11x}}$ ← AGE

Assessing the Fit of Logistic Regression Models

After estimates of the coefficients have been obtained, an estimate of the probability of development of the outcome may be calculated for each individual in the study.

Now we would like to know how effective the model we have is in describing the outcome variable. This will be accomplished by comparing observed outcomes to predicted outcomes based on the logistic model.

This comparison is referred to as assessing "Goodness-of-Fit".

What does it mean to say that the model "fits"?

let us denote the observed outcomes as $y_1, y_2, \ldots, y_n$

and

let us denote the values predicted by the model as $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$

We will conclude that the model fits if

- The summary measures of the distance between $\underline{y}$ and $\underline{\hat{y}}$ are small

and if

- The contribution of each pair $\left(y_i, \hat{y}_i\right)$, $i = 1, \ldots, n$ to these summary measures is unsystematic and is small relative to the error structure of the model

Let us concentrate on the first point, computation and evaluation of overall measures of fit.

Summary Measures of Goodness-of-Fit

Summary statistics may not be very specific about individual components

i.e.,

- a small value of one of these statistics does not rule out the possibility of some substantial deviation from fit for a few subjects.
- a large value for one of these statistics is a clear indication of a substantial problem with the model.

def: COVARIATE PATTERN - a single set of values for the covariates in a model

- when developing models we assume that each subject is unique in their configuration of the covariates.

i.e., we assume # covariate patterns = n.

e.g.,

if AGE, RACE, SEX, WT were our variables, then the combination of these may well result in a unique set of values for each subject.

- once a final model is obtained there may be relatively few variables in the model, and the number of covariate patterns may be less than n.

e.g.,

if the final model contains only RACE and SEX, each coded at 2 levels, then there are only 4 possible covariate patterns.

The number of covariate patterns is not an issue in model development. The df for tests are based on the difference in the number of variables in competing models, not on the number of covariate patterns. They become an issue when assessing the fit of a model.

Suppose our fitted model contains p independent variables $x_1, x_2, \ldots, x_p$. Let J denote the number of distinct values of $\underline{x}$ observed (i.e., covariate patterns). If some subjects have the same value of $\underline{x}$ then $J < n$.

Denote the number of subjects with $\underline{x} = \underline{x}_j$ by m_j, $j = 1, 2, \ldots, J$.

Clearly, $\sum_{j=1}^{J} m_j = n$.

Let y_j denote the number of positive responses, $y = 1$, among the m_j subjects with $\underline{x} = \underline{x}_j$.

Then $\sum_{j=1}^{J} y_j = n_1 =$ total number of subjects with $y = 1$.

- The distribution of the goodness-of-fit statistics is obtained by letting n get large
- If J, the number of covariate patterns, also increases with n, then each value of m_j will tend to be small.

- Distributional results obtained under the condition that only n becomes large are said to be based on "n- asymptotics".

- If we fix $J < n$ and let n become large, then each value of m_j will tend to become large.

- Distributional results based on each m_j becoming large are said to be based on "m-asymptotics."

Initially we will assume that $J \approx n$ as in the case most frequently occurring. We expect this to be the case whenever we have some continuous covariates in the model.

Let us now review several of the available methods.

Let $\hat{\pi}_i = \dfrac{e^{\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_j}}{1 + e^{\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_j}}$ be computed for all individuals, $i = 1, \ldots, n$.

Given the values $\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_n$, an informally used approach has been to rank order these n values and establish "deciles of risk".

i.e.,

1st decile contains the smallest $n/10$ values of $\hat{\pi}_i$

2nd decile contains the next smallest $n/10$ values of $\hat{\pi}_i$

⋮

10th decile contains the largest $n/10$ values of $\hat{\pi}_i$

If $n/10$ is not an integer, then the 10 groups may have slightly different numbers

Now, if the model holds then those who actually develop the outcome should have high values for $\hat{\pi}_i$. Similarly, those who don't develop the outcome should have low values for $\hat{\pi}_i$

Procedures have been developed for comparing the observed number with the expected number in each decile.

i.e., for the j^{th} decile

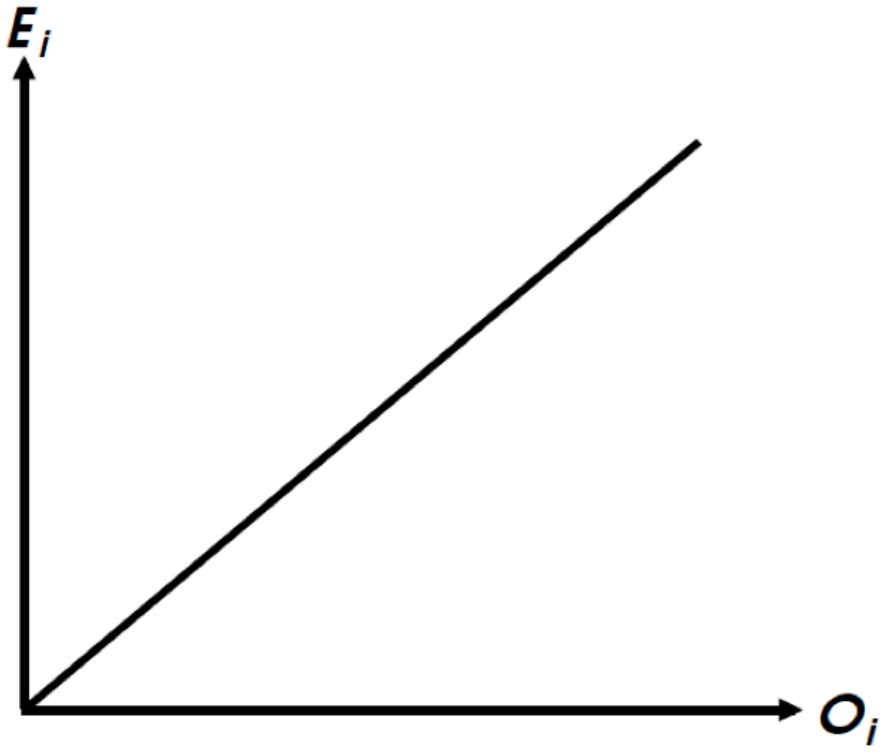
$$O_j = \sum_{i \in D_j} y_i$$

$$E_j = \sum_{i \in D_j} \hat{\pi}_i$$

where $j = 1,\ldots,10$ and where D_j denotes the $n/10$ individuals in the j^{th} decile of risk.

Consider the pairs $(O_1, E_1), \ldots, (O_i, E_i), \ldots, (O_{10}, E_{10})$

One method used has been to plot these pairs



If the observed and expected correspond, then the 10 points should fall on a line with slope = 1, intercept = 0.

This is an eye-ball method as there is no test statistic associated with it.

Pearson Chi-Square Statistic

In linear regression we were concerned with residuals of the form

$$y_i - \hat{y}_i$$

In logistic regression fitted values are calculated for each covariate pattern, and depend on the estimated probability for that covariate pattern

We denote the fitted value, $\hat{y}_i$, as

$$m_j\hat{\pi}_j = m_j\left\{\frac{e^{\hat{g}(\underline{x}_j)}}{1+e^{\hat{g}(\underline{x}_j)}}\right\}$$

where $\hat{g}(\underline{x}_j)$ is the estimated logit.

For a particular covariate pattern the Pearson residual is defined as

$$r\left(y_j\hat{\pi}_j\right) = \frac{\left(y_j - m_j\hat{\pi}_j\right)}{\sqrt{m_j\hat{\pi}_j\left(1-\hat{\pi}_j\right)}}$$

The summary statistic based on these residuals is the Pearson chi-square statistic

$$X^2 = \sum_{j=1}^{J} r\left(y_j, \hat{\pi}_j\right)^2$$

and $X^2 \sim \chi^2\left(J - \left(p+1\right)\right)$ if the model holds

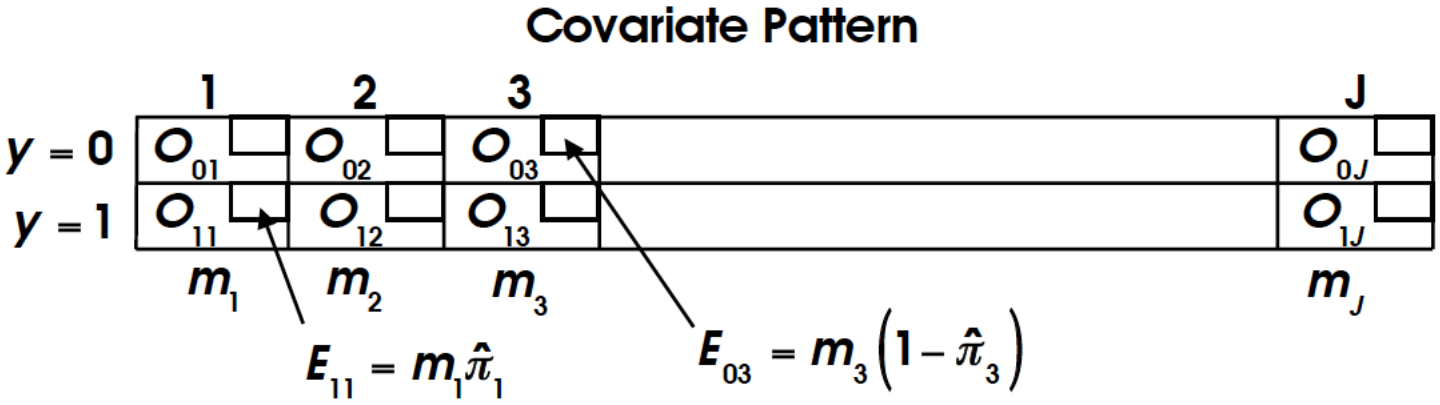
Problem: when $J \approx n$, the distribution is obtained under n - asymptotics, and hence the number of parameters is increasing at the same rate as the sample size.

Hence, $p-$ values calculated for χ^2 are incorrect when $J \approx n$

Although the $p-$value may be slightly off, X^2 is an effective way to compare observed to expected frequencies for each covariate pattern .

This statistic is routinely produced by many software packages.

The Pearson Chi Square Statistic can be thought of as arising from the following 2×J table:



When chi-square tests are computed from a contingency table the *p*-values are correct under the hypothesis when the estimated expected values are "large" in each cell. This condition will hold under *m*-asymptotics.

In this table the expected values will always be quite small since the number of columns, *J*, increases as *n* increases.

One way to avoid these difficulties under *n*-asymptotics is to group the data in such a way that *m*-asymptotics can be used. For example, we may collapse the columns into a fixed number of groups, *g*, and then calculate the observed and expected frequencies.

By fixing the number of columns, the estimated expected frequencies will become large as *n* becomes large. Thus *m*-asymptotics hold.

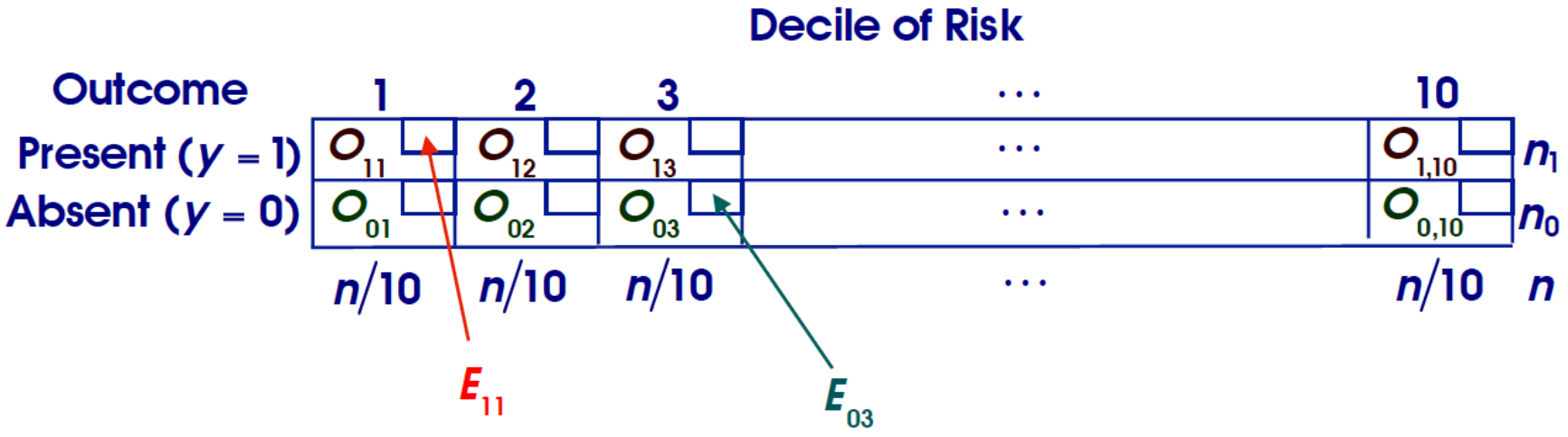
The Hosmer-Lemeshow Tests

Let us suppose that $J = n$. Two grouping strategies are proposed

(1) Collapse the table based on percentiles of the estimated probabilities.

(2) Collapse the table based on fixed values of the estimated probabilities.

With method (1), use of $g = 10$ groups results in the first group containing the $n_1' = n/10$ subjects having the smallest estimated probabilities, and the last group containing the $n_{10}' = n/10$ subjects having the largest estimated probabilities.



Then we compute

$$\hat{C} = \sum_{k=0}^{1} \sum_{j=1}^{10} \frac{\left(O_{kj} - E_{kj}\right)^2}{E_{kj}}$$

This is the Pearson chi-square statistic from the $2 \times g$ table of observed and expected frequencies.

If the 2nd grouping strategy is used, $g = 10$ groups results in cutpoints defined at the values $k/10$, $k = 1,2,\ldots,9$ and the groups contain all subjects with estimated probabilities between adjacent cutpoints

e.g.,

1^{st} group = $0 = \hat{\pi}_i < .1$

2^{nd} group = $.1 \le \hat{\pi}_i < .2$

$\vdots$

10^{th} group = $.9 \le \hat{\pi}_i \le 1.0$

Based on extensive simulations, it has been demonstrated that, when $J = n$ and the fitted logistic model is the correct model, the distribution of $\hat{C}$ is well approximated by $\chi^2(g-2)$

Example: ICU data.

```
. logit STA AGE CAN _ISYSGP_4 TYP LOCD

Iteration 0:   log likelihood = -100.08048
Iteration 1:   log likelihood = -70.385527
Iteration 2:   log likelihood = -67.395341
Iteration 3:   log likelihood = -66.763511
Iteration 4:   log likelihood = -66.758491
Iteration 5:   log likelihood = -66.758489

Logistic regression                               Number of obs   =        200
                                                  LR chi2(5)      =      66.64
                                                  Prob > chi2     =     0.0000
Log likelihood = -66.758489                       Pseudo R2       =     0.3330

------------------------------------------------------------------------------
         STA |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         AGE |    .040628   .0128617     3.16   0.002     .0154196    .0658364
         CAN |   2.078751   .8295749     2.51   0.012     .4528141    3.704688
   _ISYSGP_4 |   -1.51115   .7204683    -2.10   0.036    -2.923242   -.0990585
         TYP |   2.906679   .9257469     3.14   0.002     1.092248     4.72111
        LOCD |   3.965535   .9820316     4.04   0.000     2.040788    5.890281
       _cons |  -6.680532   1.320663    -5.06   0.000    -9.268984    -4.09208
------------------------------------------------------------------------------
```

```
. lfit, group(10) table

Logistic model for STA, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.0105 |     0 |   0.1 |    20 |  19.9 |    20 |
  |     2 | 0.0290 |     0 |   0.4 |    20 |  19.6 |    20 |
  |     3 | 0.0492 |     2 |   1.0 |    21 |  22.0 |    23 |
  |     4 | 0.0666 |     0 |   1.0 |    17 |  16.0 |    17 |
  |     5 | 0.1083 |     2 |   1.8 |    19 |  19.2 |    21 |
  |-------+--------+-------+-------+-------+-------+-------|
  |     6 | 0.1674 |     2 |   2.6 |    17 |  16.4 |    19 |
  |     7 | 0.2254 |     5 |   3.9 |    15 |  16.1 |    20 |
  |     8 | 0.3171 |     4 |   5.5 |    16 |  14.5 |    20 |
  |     9 | 0.4554 |     8 |   7.6 |    12 |  12.4 |    20 |
  |    10 | 0.9623 |    17 |  16.1 |     3 |   3.9 |    20 |
  +--------------------------------------------------------+
      number of observations =       200
            number of groups =        10
    Hosmer-Lemeshow chi2(8) =         4.00
                Prob > chi2 =         0.8570
. lfit
Logistic model for STA, goodness-of-fit test

      number of observations =       200
 number of covariate patterns =       135
           Pearson chi2(129) =        79.23
                Prob > chi2 =         0.9998
```

Because the distribution of $\hat{C}$ depends on *m*-asymptotics, the appropriateness of the *p*-value will depend on the estimated expected frequencies being large enough to employ this theory.

If one is concerned about the magnitude of the expected frequencies, selected adjacent columns may be combined to increase the size of the expected frequencies. Unfortunately, when this is done the power of the test is reduced since the degrees of freedom are reduced.

When $\hat{C}$ is calculated from fewer than 6 groups, it will almost always indicate that the model fits. Thus, try to use with as many groups as possible.

The problem is that, when working with really large data sets, the GOF test may be too powerful, indicating that the model is poorly calibrate when it is not.

"Standardizing The Power Of The Hosmer-Lemeshow Goodness Of Fit Test In Large Data Sets".

Paul, Prabasaj, Michael L. Pennell, and Stanley Lemeshow. *Statistics in Medicine.* 32.1 (2013): 67-80.

In this paper we found that the power of the Hosmer-Lemeshow test increased with sample size and decreased with the number of groups.

Previous work has shown that the Hosmer-Lemeshow test works best when there are at least five observations per group, and when the number of groups is greater than or equal to six.

The test often breaks down as well when the event is rare.

Taking all of these into account, this paper listed recommendations for what group sizes to use in various scenarios.

With sample sizes up to 1000, a group size of ten is recommended. This often keeps the power below 70% which, in some scenarios, may still be too powerful.

For sample sizes between 1,000 and 25,000 observations, we recommend using the following equation to determine the number of groups, *g*, to use:

$$g = \max\left[10, \min\left\{\frac{m}{2}, \frac{n-m}{2}, 2 + 8\left(\frac{n}{1000}\right)^2\right\}\right]$$

where *n* is the sample size and *m* is the number of successes.

This formula is justified by noting that power was kept relatively consistent to a benchmark used with a sample size of 1000 and a group size of 10 in our simulation results when the equation

$$g = 2 + 8\left(\frac{n}{1000}\right)^2$$

was used.

Moreover, the assumption is made that the number of groups taken is never below 10.

It is also noted that this equation breaks down as the sample size becomes smaller, as it is recommended to have at least five observations per group.

Finally, for sample sizes greater than 25,000, this equation breaks down as well, as the equation defaults to the number of successes (m) divided by two.

This results in a test that is too powerful.

Applying the formula for *g* on the previous slide:

- for $n < 1000$, use $g = 10$
- for $n = 2000$, use $g = 34$
- for $n = 4000$, use $g = 130$
- for n > 25,000, we can't apply this rule as the formula breaks down

For large data sets, we have begun to run the H-L test repeatedly using differing numbers of groups to see if good fit is maintained over the range of *g*.

e.g., ICU model with 37,913 patients in developmental data set and 4,212 patients in the validation data set

Developmental dataset

Area under the ROC curve = 0.771

Hosmer-Lemeshow goodness of fit test

Obs (N)	Groups	DoF	p-value
37,913	10	8	0.6599
37,913	20	18	0.1529
37,913	30	28	0.6417
37,913	40	38	0.2924
37,913	50	48	0.6463
37,913	60	58	0.6729
37,913	70	68	0.4528
37,913	80	78	0.4462
37,913	90	88	0.3036
37,913	100	98	0.3119
37,913	150	148	0.1687
37,913	200	198	0.2857
37,913	250	248	0.0580
37,913	300	298	0.5931
37,913	350	348	0.1107
37,913	400	398	0.4498
37,913	450	448	0.1305
37,913	500	498	0.5497
37,913	550	548	0.1334
37,913	600	598	0.4071
37,913	650	648	0.3702
37,913	700	698	0.3172
37,913	750	748	0.5634
37,913	800	798	0.3019
37,913	850	848	0.7065
37,913	900	898	0.6207
37,913	950	948	0.8317
37,913	1000	998	0.1523

Validation dataset

Area under the ROC curve = 0.779

Hosmer-Lemeshow goodness of fit test

Obs (N)	Groups	DoF	p-value
4,212	10	10	0.1615
4,212	20	20	0.4069
4,212	30	30	0.1238
4,212	40	40	0.4082
4,212	50	50	0.0211
4,212	60	60	0.1718
4,212	70	70	0.2039
4,212	80	80	0.2201
4,212	90	90	0.1922
4,212	100	100	0.2597
4,212	110	110	0.7880
4,212	120	120	0.3073
4,212	130	130	0.2000
4,212	140	140	0.7009
4,212	150	150	0.5995

A strategy for evaluating goodness-of-fit for a logistic regression model using the Hosmer-Lemeshow test on samples from a large data set

Adam Bartley, Michael Pennell, Stanley Lemeshow, and Gary Phillips

Purpose of Research

- Evaluate, through a simulation study, a subsampling approach for assessing goodness-of-fit in large data sets.
- Use results of simulations to make recommendations for implementing a subsampling approach.

Simulation Methods

- Data were simulated under 5 different scenarios (Table 1).
- Except for Scenario 1, each data set was analyzed using a model that differed from the truth (Table 1).
- Scenario 2: true and fitted models were virtually identical.
- Scenario 4: small difference in the tails.
- The H-L test was implemented on 100 subsets of size 1,000 and 5,000.
- Number of significant tests (p-value < 0.05) enumerated.
- Process repeated for 100 data sets/scenario.

Table 1. Simulation Scenarios

Scenario	True Model for Log-odds, $g(X)$	Fitted Model
1	$g(\boldsymbol{X}) = X_1 + X_2$	$\hat{g}(\boldsymbol{X}) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$
2	$g(\boldsymbol{X}) = X_1 + X_2 + 0.05X_1X_2$	$\hat{g}(\boldsymbol{X}) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$
3	$g(\boldsymbol{X}) = X_1 + X_2 + 0.5X_1X_2$	$\hat{g}(\boldsymbol{X}) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$
4	$g(\boldsymbol{X}) = X_1 + 0.05X_1^2$	$\hat{g}(\boldsymbol{X}) = \hat{\beta}_0 + \hat{\beta}_1 X_1$
5	$g(\boldsymbol{X}) = X_1 + 0.1X_1^2$	$\hat{g}(\boldsymbol{X}) = \hat{\beta}_0 + \hat{\beta}_1 X_1$

$X_1 \sim Normal\,(0,1)$; $X_2 \sim Bernoulli\,(p = 0.5)$

Results

Table 2. Simulation results: # Significant Subsets out of 100. Frequencies are # data sets (out of 100) with specified number of significant subsets.

		Subset size = 1,000			Subset size = 5,000		
Scenario	N	> 5	> 10	> 20	> 5	> 10	> 20
1	50,000	45	1	0	81	20	1
	100,000	36	3	0	57	3	0
2	50,000	39	0	0	72	18	0
	100,000	35	0	0	63	7	0
3	50,000	71	13	0	100	100	88
	100,000	71	10	0	100	100	67
4	50,000	76	8	0	99	81	20
	100,000	64	5	0	98	78	9
5	50,000	97	47	0	100	100	100
	100,000	92	37	0	100	100	100

Results

- Samples frequently had > 5 significant subsets; even if correct model was fit (Scenario 1).
- > 20 significant subsets was only common when true and fitted models differed greatly (Scenarios 3 and 5).
- > 10 subsets uncommon when true and fitted models were the same or almost identical (Scenarios 1, 2).
- Inadequate power to detect poorly fit models (Scenarios 3 and 5) when subset size < 5,000.
- True model rejected too often when N < 100,000.

Recommendations

- For N ≥ 100,000, draw 100 subsets of size 5,000.
- Conclude lack-of-fit if H-L test is significant in > 10 subsets.

"A new calibration test and a reappraisal of the calibration belt for the assessment of prediction models based on dichotomous outcomes"

Giovanni Nattino, Stefano Finazzi and Guido Bertolini

Statistics in Medicine 2014, 33 2390–2407

Recall:

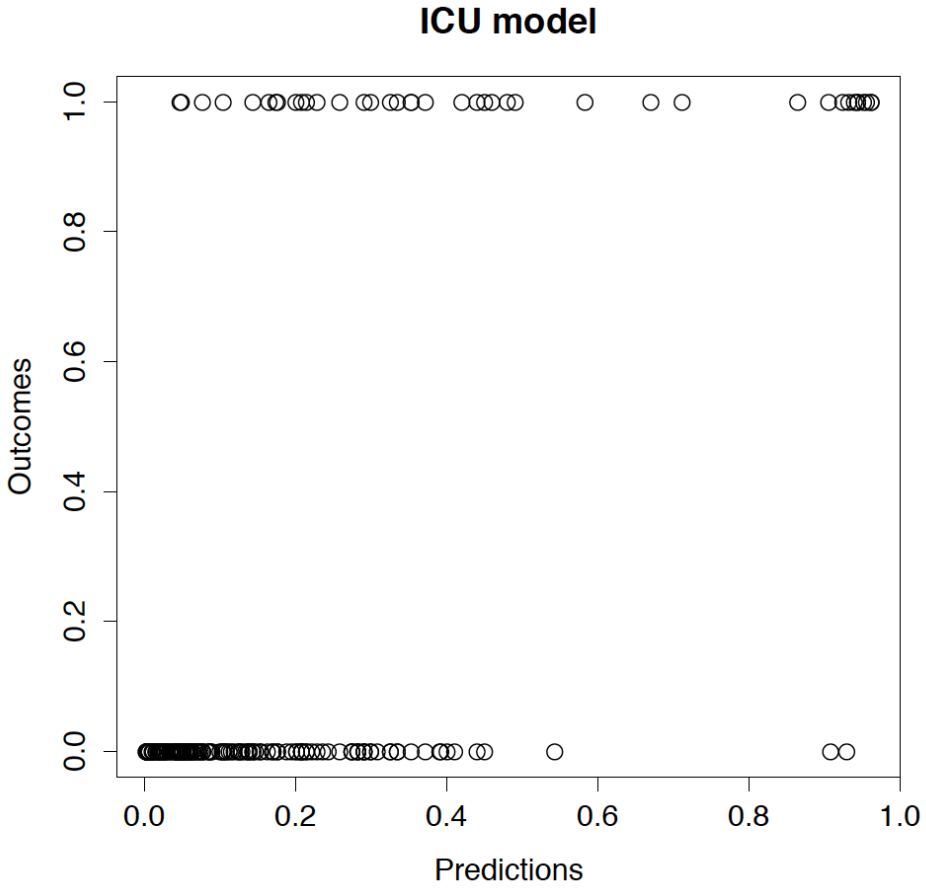
$$\text{logit}\left(\Pr\left(y=1\middle|\underline{x}\right)\right)=\text{logit}\left(\pi\left(\underline{x}\right)\right)=g\left(\underline{x}\right)=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\cdots+\hat{\beta}_px_p$$

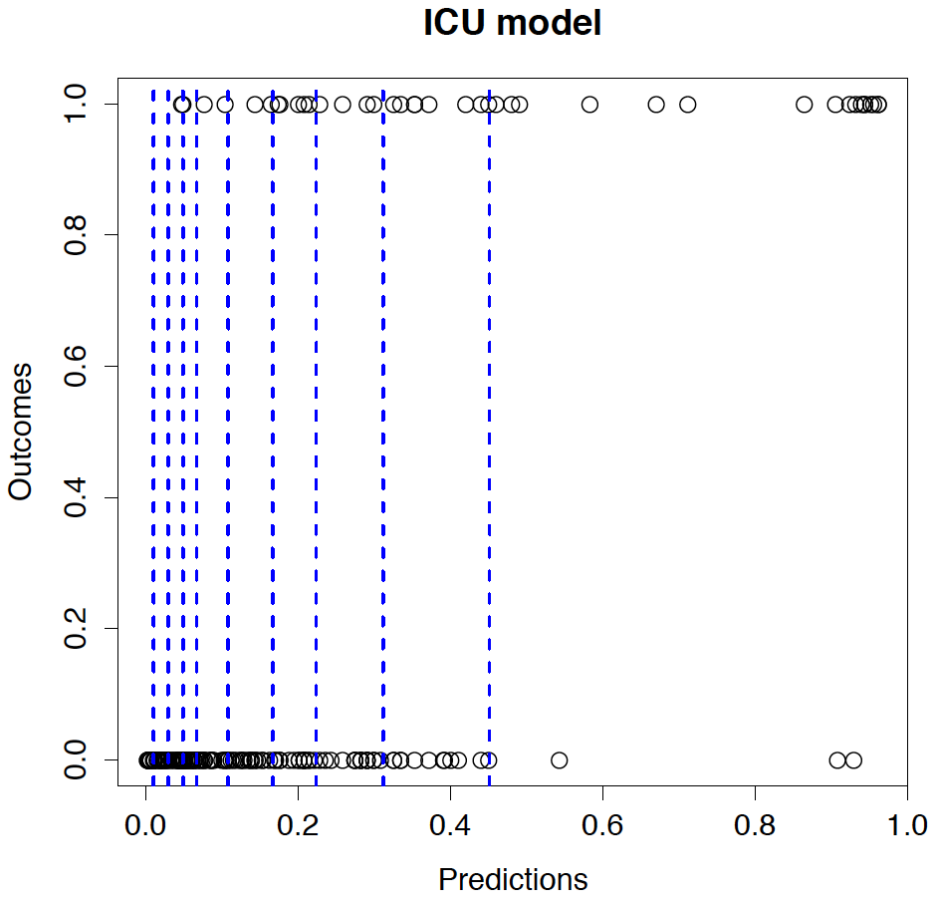
For each subject, $i=1,2,\ldots,n$, we can compute:

- the logit $\hat{g}_i\left(\underline{x}_i\right)$
- the probability $\hat{\pi}_i=\dfrac{e^{\hat{g}_i\left(\underline{x}_i\right)}}{1+e^{\hat{g}_i\left(\underline{x}_i\right)}}$

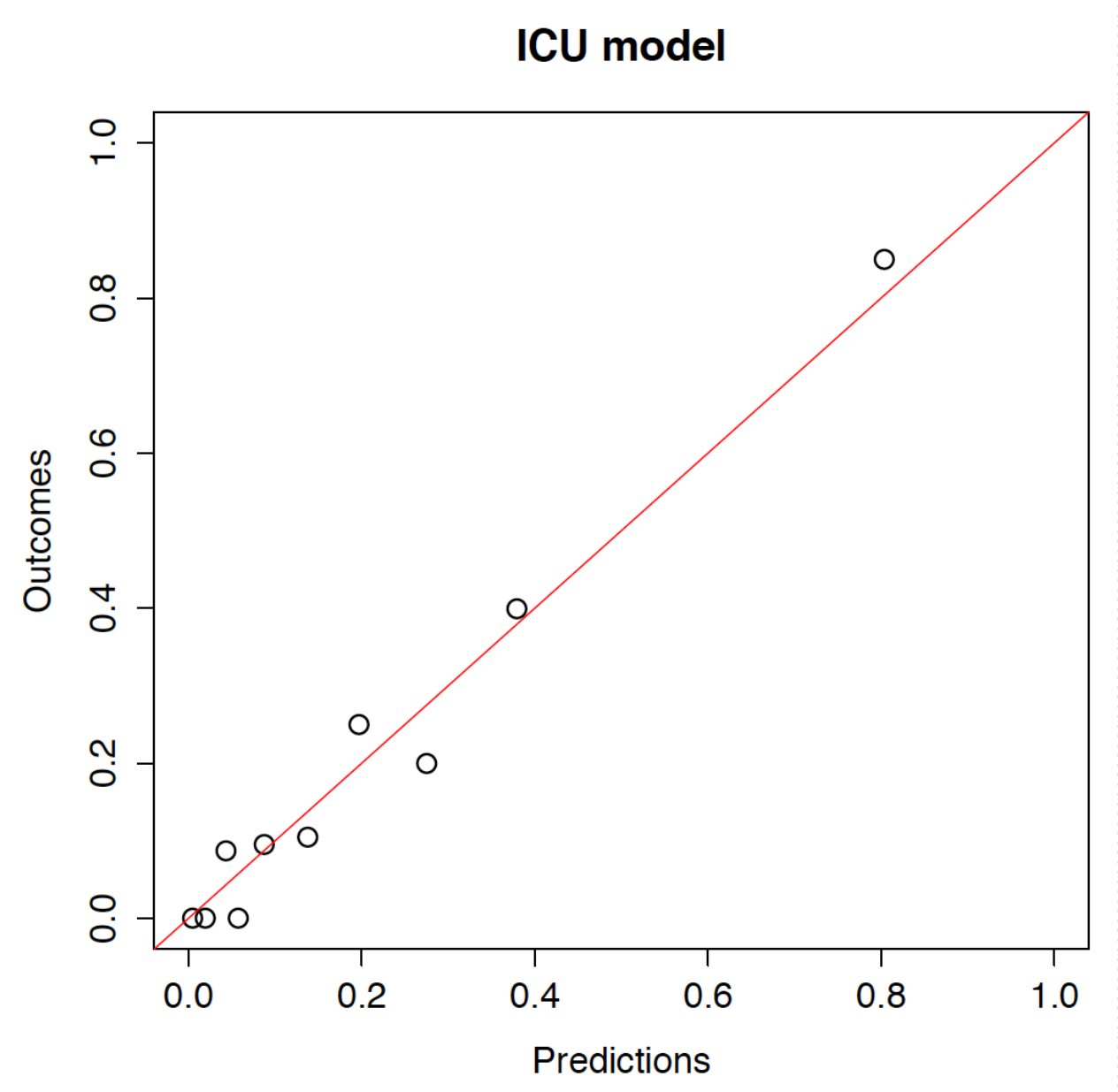
Calibration is the agreement between y_i and $\hat{\pi}_i$

The Calibration Plot





The Calibration Plot



The Calibration Curve:

Now that we've fit our model,

- we have

$$\hat{\pi}_i = \Pr\left(y_i = 1 \middle| \underline{x}_i\right) \text{ and } \hat{g}_i = g\left(x_i\right)$$

for each subject.

This relationship can be expressed as

$$\text{logit}\left\{\Pr\left(Y = 1 \middle| \hat{\pi}\right)\right\} = \alpha_0 + \alpha_1\left\{\text{logit}\left(\hat{\pi}\right)\right\} = \alpha_0 + \alpha_1\left\{\hat{g}\right\} = \text{logit}\left\{\Pr\left(Y = 1 \middle| \hat{g}\right)\right\}$$

so, if $\alpha_0 = 0$ and $\alpha_1 = 1$

$$\text{logit}\left\{\Pr\left(Y = 1 \middle| \hat{\pi}\right)\right\} = 0 + 1\left\{\text{logit}\left(\hat{\pi}\right)\right\} = \text{logit}\left(\hat{\pi}\right)$$

$$\Rightarrow \Pr\left(\left(Y = 1\right) \middle| \hat{\pi}\right) = \hat{\pi}$$

If the data fit perfectly, then $\hat{\alpha}_0 \equiv 0$ and $\hat{\alpha}_1 \equiv 1$

but it certainly doesn't have to be a linear relationship

Why not:

$$\text{logit}\left\{\Pr\left(Y=1\right)\middle|\,\hat{g}\right\}=\alpha_0+\alpha_1\hat{g}+\alpha_2\hat{g}^2+\cdots+\alpha_m\hat{g}^m$$

What should we choose for m?

- if too small $\Rightarrow$ too simplistic
- if too large $\Rightarrow$ estimation of useless parameters

a forward selection algorithm is used

e.g., ICU data

$$m=2:\text{logit}\left\{\Pr\left(Y=1\right)\middle|\,\hat{g}\right\}=0.117+0.917\hat{g}+0.076\hat{g}^2$$

$$\hat{L}_2=-66.22016$$

$$m=3:\text{logit}\left\{\Pr\left(Y=1\right)\middle|\,\hat{g}\right\}=0.116+0.916\hat{g}+0.076\hat{g}^2+0.00019\hat{g}^3$$

$$\hat{L}_3=-66.22015$$

Likelihood Ratio Test: $H_0:\alpha_3=0$ vs $H_a:\alpha_3\neq 0$

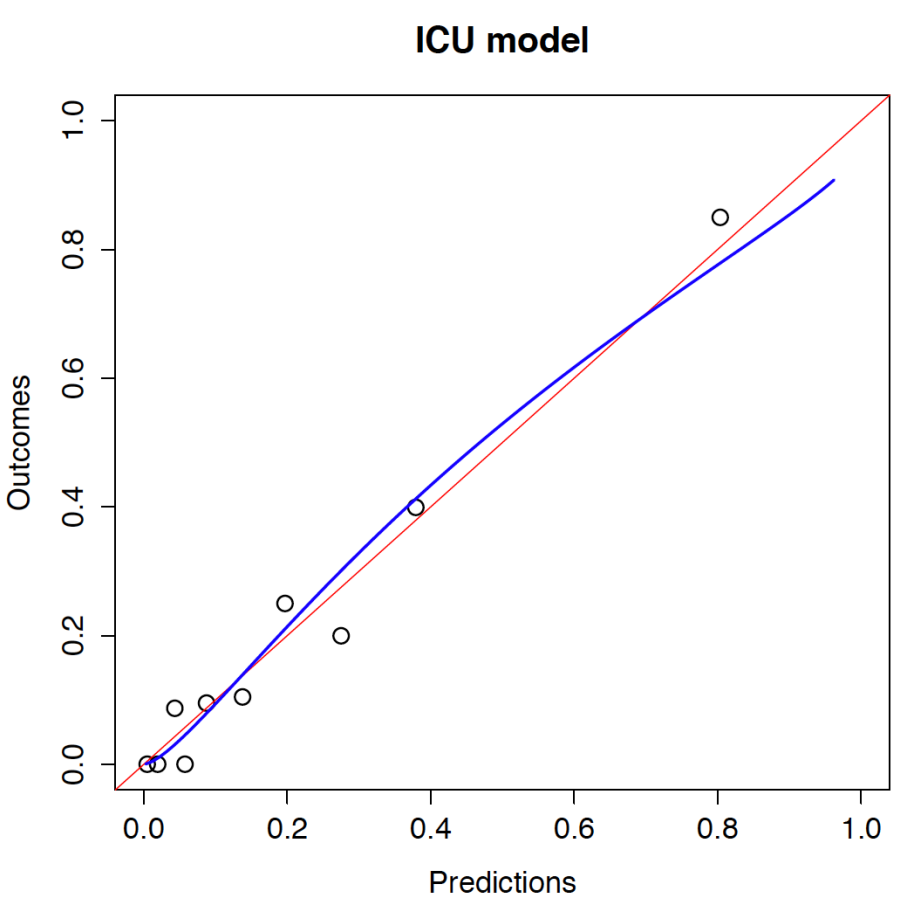
$G=0.00002,\quad p=0.996$ NS $\Rightarrow m=2$

so using the $m = 2$ model:

$$\text{logit}\left\{\Pr\left(Y = 1\middle|\hat{g}\right\}\right) = 0.117 + 0.917\hat{g} + 0.076\hat{g}^2$$

we define the calibration curve as:

$$\Pr\left(Y = 1\middle|\hat{\pi}\right) = f\left(\hat{\pi}\right) = \frac{e^{0.117+0.917\text{logit}\left(\hat{\pi}\right)+0.076\left(\text{logit}\left(\hat{\pi}\right)\right)^2}}{1+e^{0.117+0.917\text{logit}\left(\hat{\pi}\right)+0.076\left(\text{logit}\left(\hat{\pi}\right)\right)^2}}$$



so the best model appears to be the $m = 2$ model:

$$\text{logit}\left\{\Pr\left(Y = 1\right)\middle|\, \hat{g}\right\} = 0.117 + 0.917\hat{g} + 0.076\hat{g}^2$$

If the calibration were perfect: $\alpha_0 = 0,\ \alpha_1 = 1,\ \alpha_2 = 0,$

since then, $\text{logit}\left\{\Pr\left(Y = 1\right)\middle|\, \hat{g}\right\} = 0 + 1 \times \hat{g} + 0 = \hat{g}$

So we would like to test $H_0:\ \alpha_0 = 0$ and $\alpha_1 = 1$ and $\alpha_2 = 0$

vs $H_a:\ \alpha_0 \neq 0$ or $\alpha_1 \neq 1$ or $\alpha_2 \neq 0$

The test:

- is based on a likelihood ratio statistic;
- accounts for the iterative process to define m.

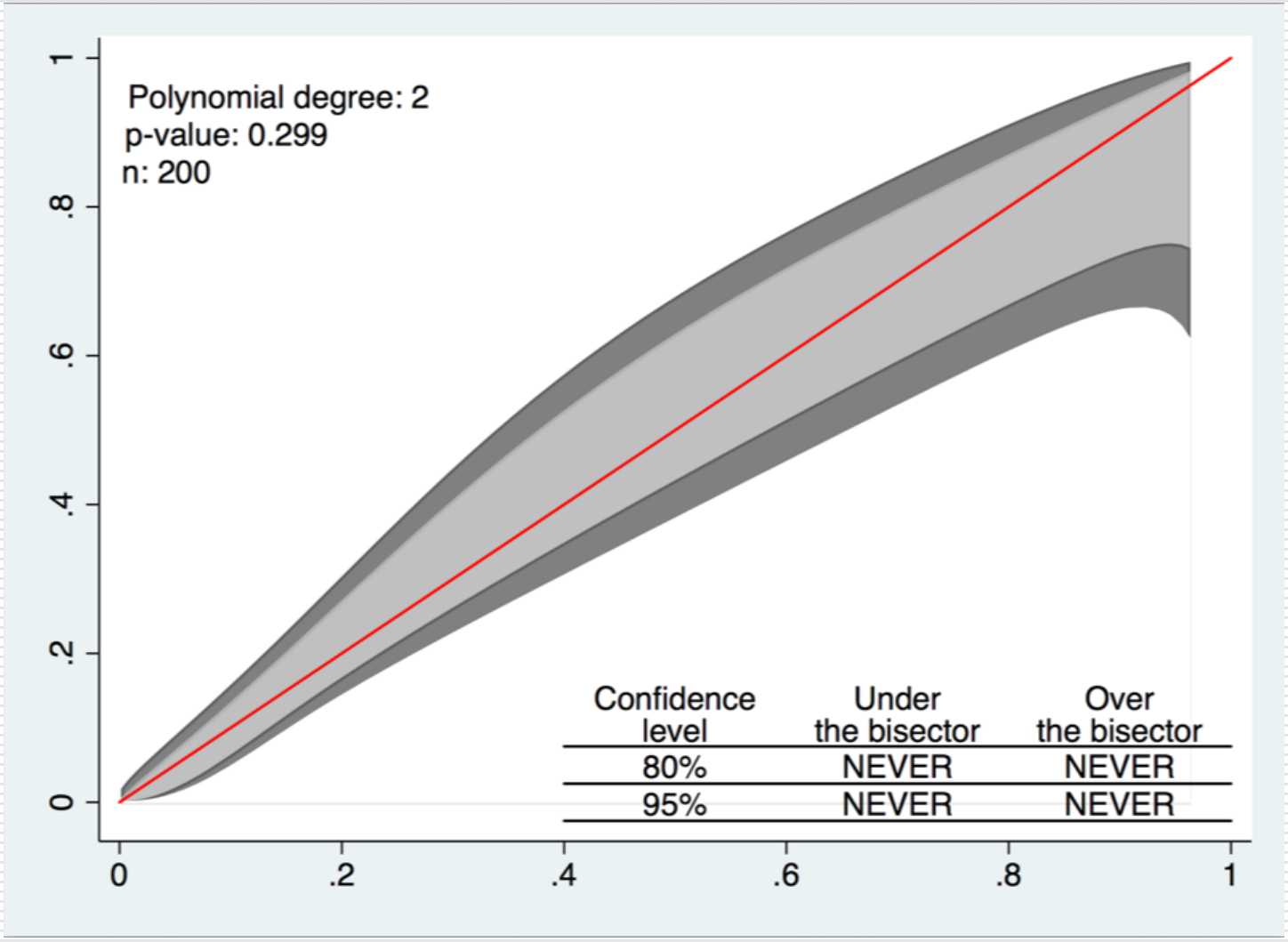
$G = 1.08, \quad p - \text{value} = 0.299$

Recall: Hosmer-Lemeshow p-value = 0.857.

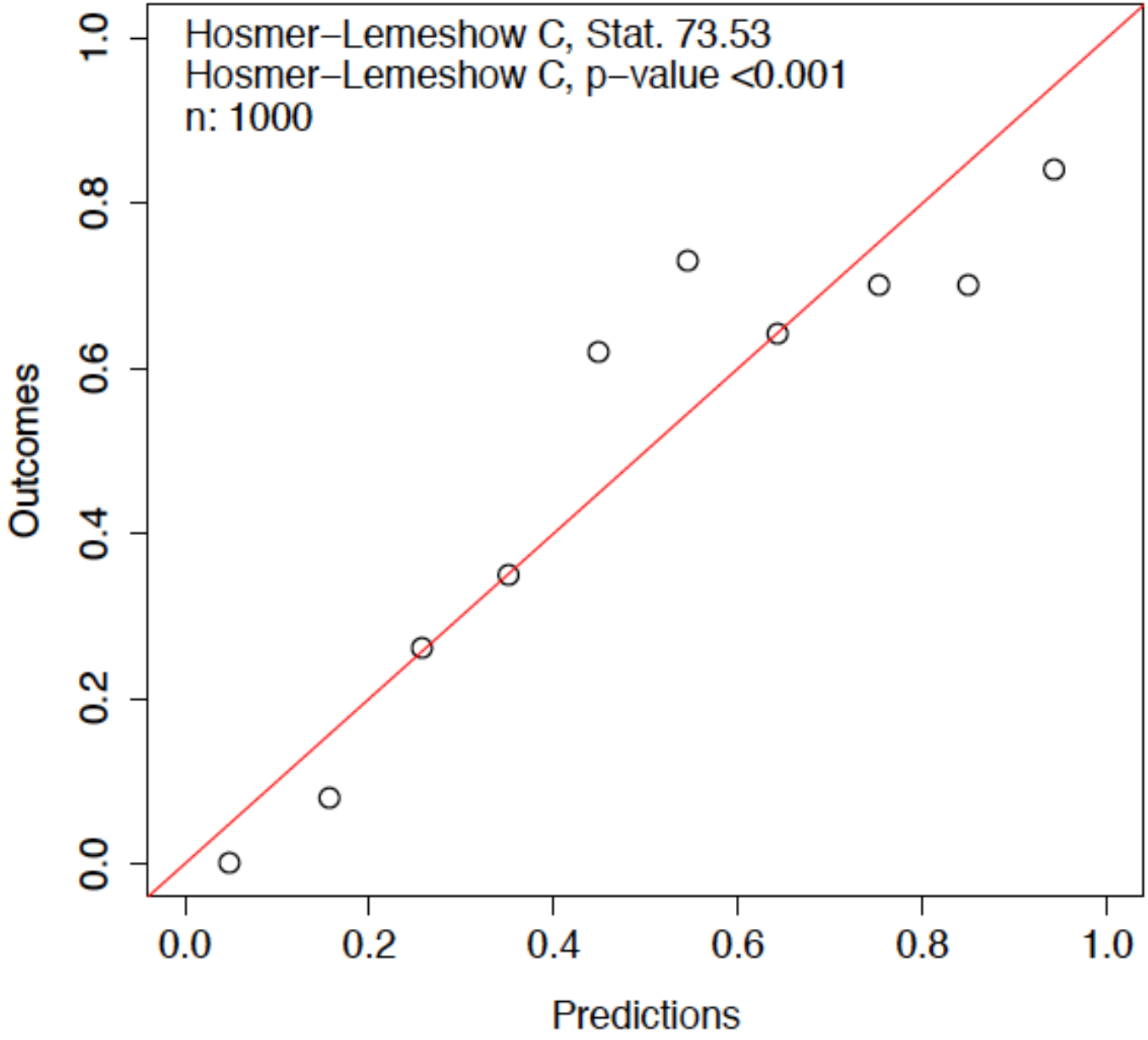
The calibration belt

ICU Data

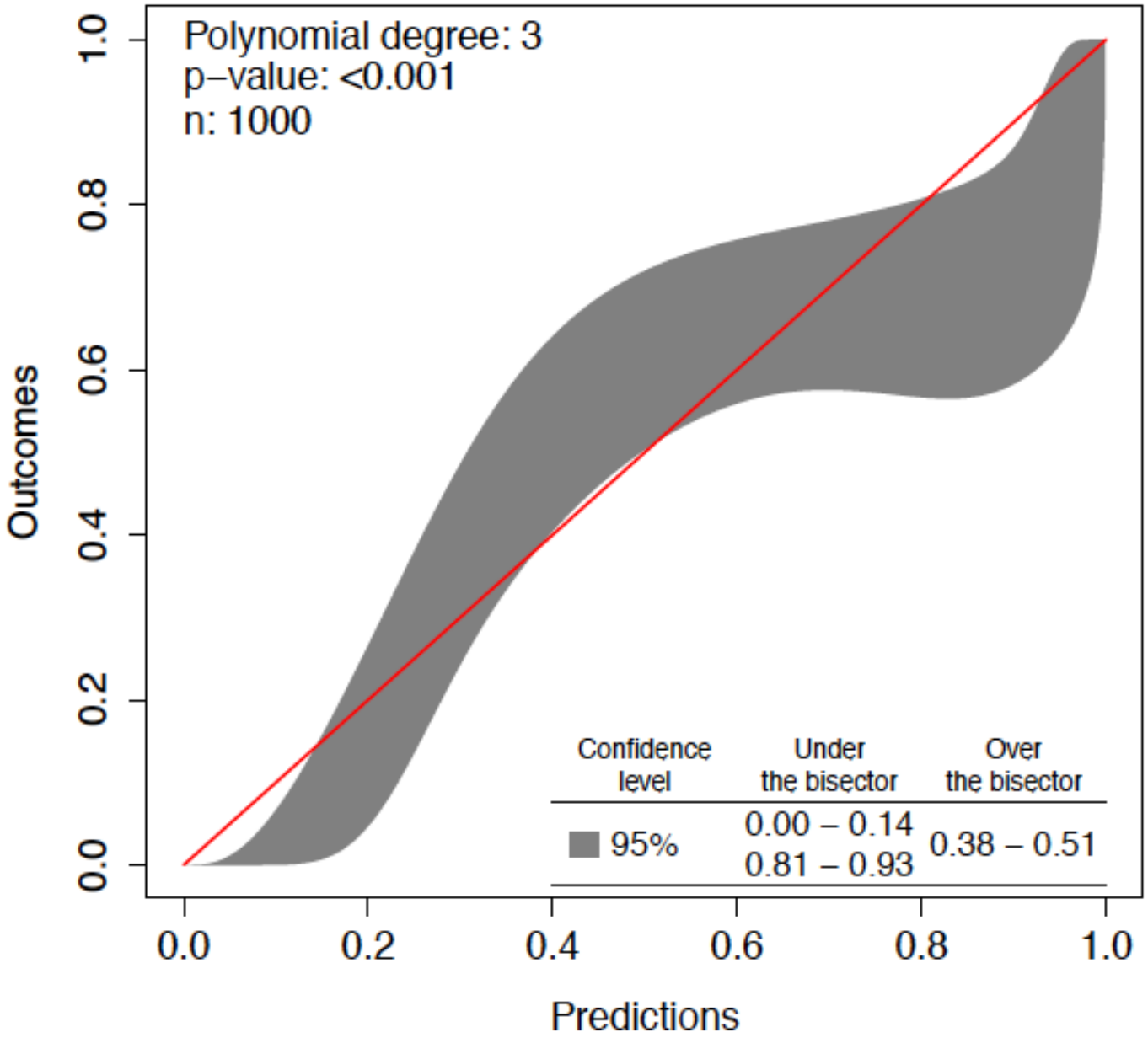
```
. calibrationbelt
```



Example of a Poorly Calibrated Model:



Calibration Belt for this model:



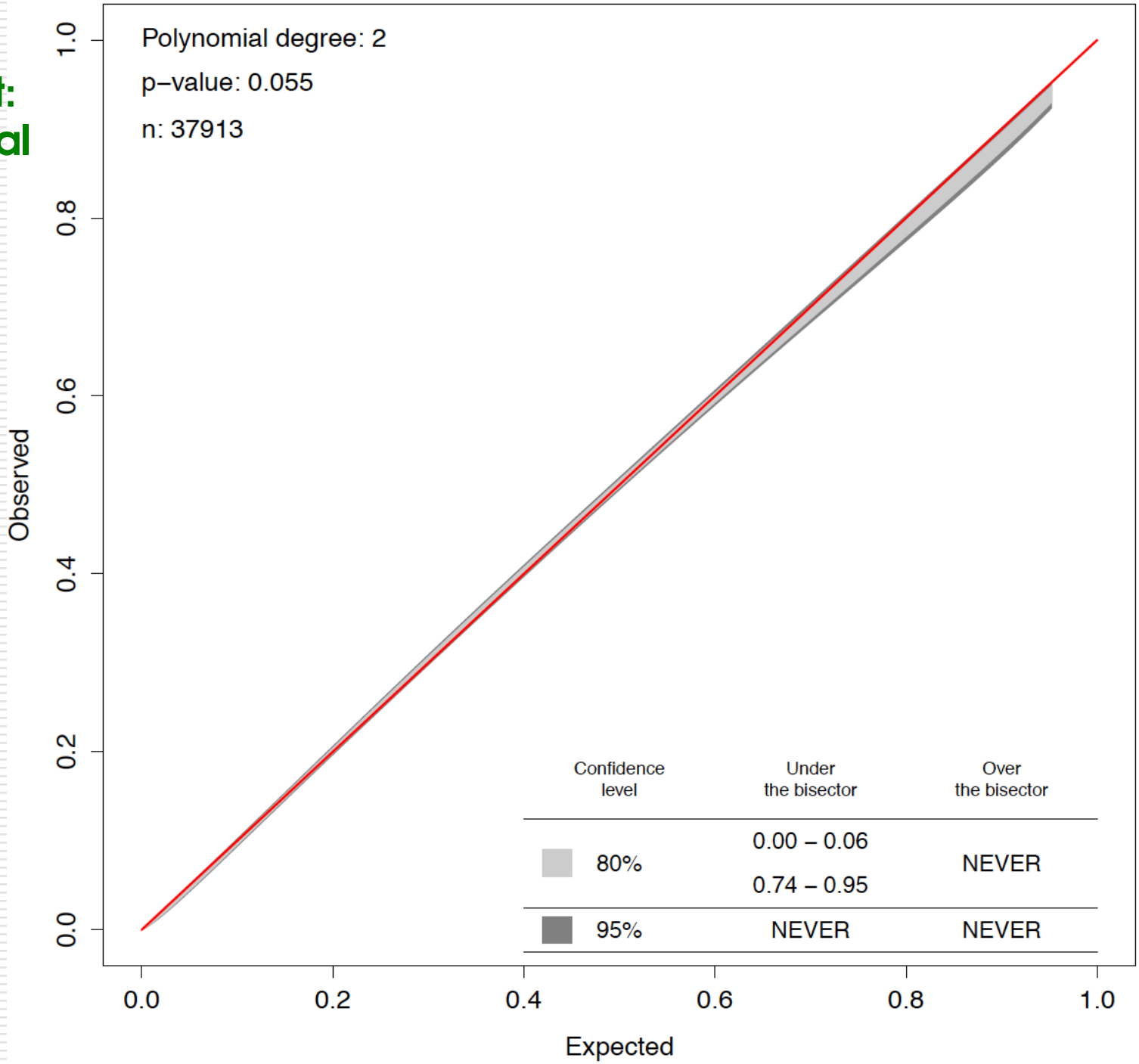
So, as we've seen, the calibration belt can assess the goodness of fit of a model without any categorization.

The calibration belt is an informative tool to detect deviations from the perfect fit of a model.

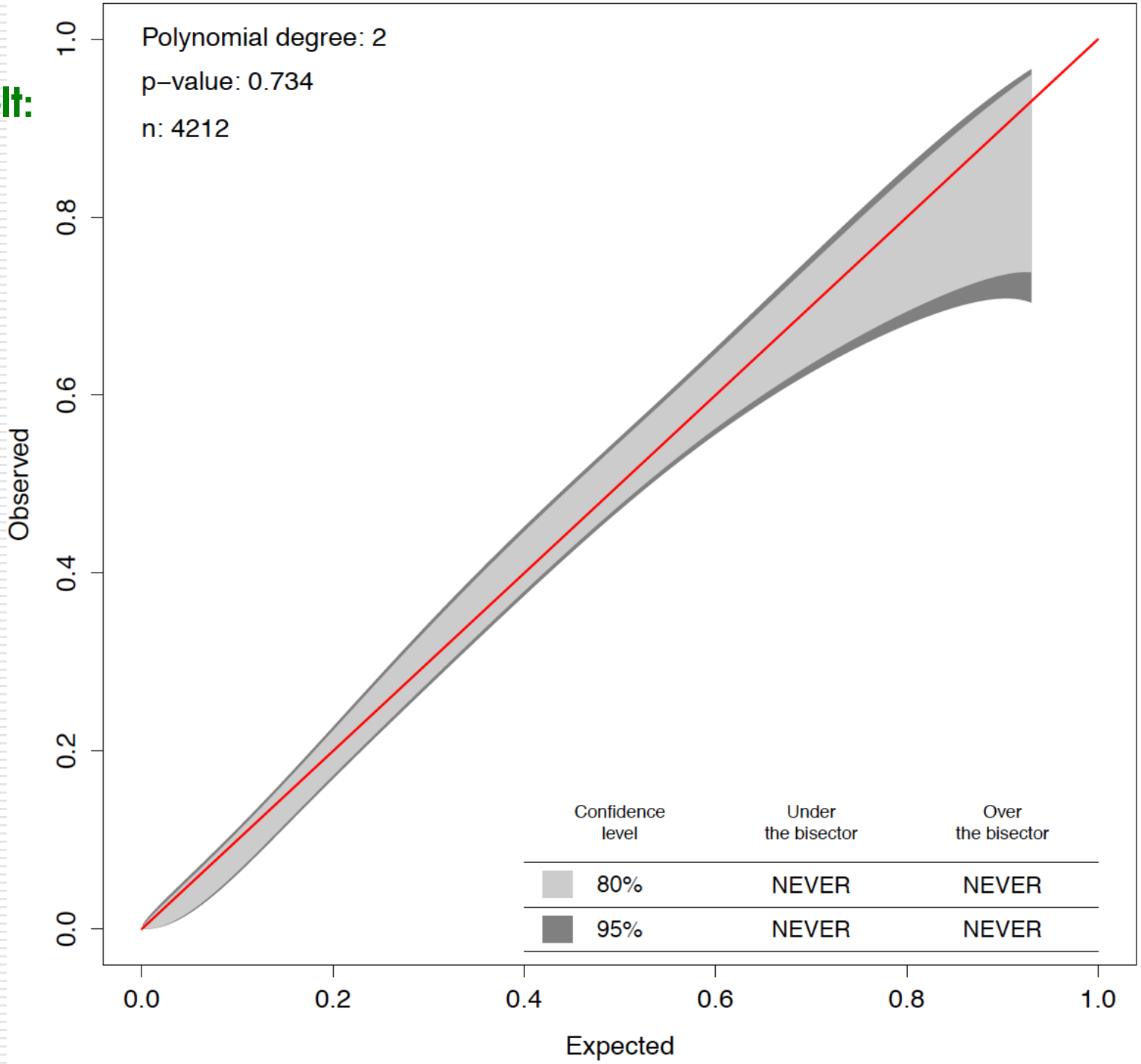
The information provided helps improving the goodness of fit of logistic regression models.

Let us return to our ongoing modeling efforts.

Calibration Belt: Developmental



Calibration Belt: Validation



Assessing the calibration of dichotomous outcome models with the calibration belt

Giovanni Nattino
Division of Biostatistics
College of Public Health
The Ohio State University
Columbus, OH
nattino.1@osu.edu

Stanley Lemeshow
Division of Biostatistics
College of Public Health
The Ohio State University
Columbus, OH

Gary Phillips
Center for Biostatistics
The Department of Biomedical Informatics
The Ohio State University
Columbus, OH

Stefano Finazzi
GiViTI Coordinating Center
Laboratory of Clinical Epidemiology
IRCCS Istituto di Ricerche Farmacologiche 'Mario Negri'
Ranica, Italy

Guido Bertolini
GiViTI Coordinating Center
Laboratory of Clinical Epidemiology
IRCCS Istituto di Ricerche Farmacologiche 'Mario Negri'
Ranica, Italy

Thank you!