Assessing the Calibration of Dichotomous Outcome Models with the Calibration Belt

Giovanni Nattino

The Ohio Colleges of Medicine Government Resource Center
The Ohio State University

Stata Conference - July 19, 2018

Background: Logistic Regression

- Most popular family of models for binary outcomes ($Y = 1$ or $Y = 0$);
- Models $Pr\,(Y = 1)$, probability of "success" or "event";
- Given predictors $X_1, ..., X_p$, the model is

$$\text{logit}\,\{Pr\,(Y = 1)\} = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p,$$

  where $\text{logit}(\pi) = \log\left(\pi/(1-\pi)\right)$.
- **Does my model fit the data well?**

Goodness of Fit of Logistic Regression Models

Let $\hat{\pi}$ be the model's estimate of $Pr(Y = 1)$ for a given subject.

Two measures of goodness of fit:

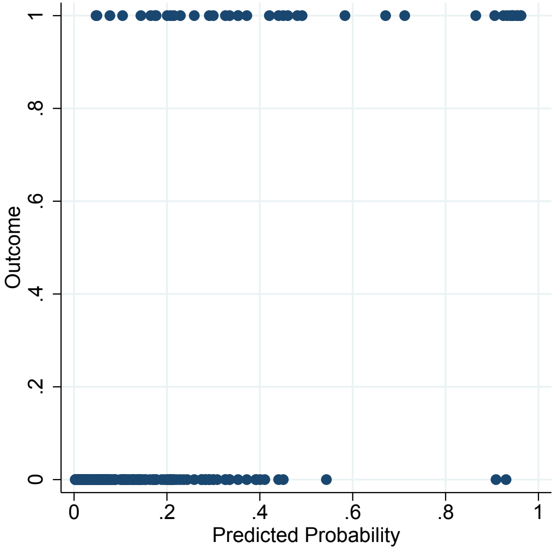
- Discrimination
  - Do subjects with $Y = 1$ have higher $\hat{\pi}$ than subjects with $Y = 0$?
  - Evaluated with area under ROC curve.
- **Calibration**
  - Does $\hat{\pi}$ estimate $Pr(Y = 1)$ accurately?

An Example: ICU Data

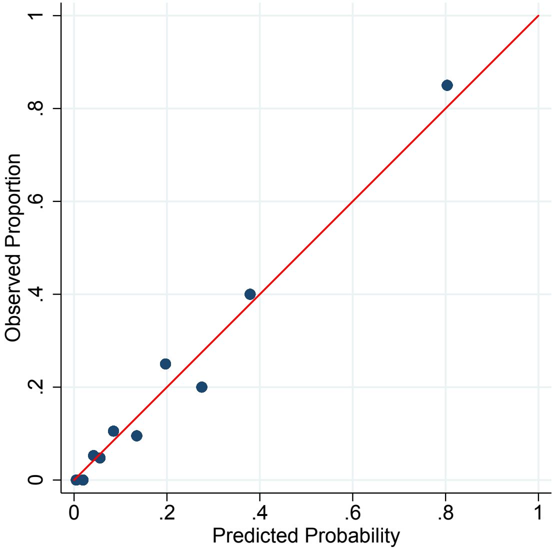
```
. logit sta age can sysgp_4 typ locd

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
sysgp_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

An Example: ICU Data



An Example: ICU Data



The Hosmer-Lemeshow Test

- Divide data into G groups (usually, $G = 10$).
- For each group, define:
  - O_{1g} and E_{1g}: number of observed and expected events ($Y = 1$).
  - O_{0g} and E_{0g}: number of observed and expected non-events ($Y = 0$).
- The Hosmer-Lemeshow statistic is:

$$\widehat{C} = \sum_{g=1}^{G} \left[\frac{(O_{1g} - E_{1g})^2}{E_{1g}} + \frac{(O_{0g} - E_{0g})^2}{E_{0g}} \right]$$

- Under the hypothesis of perfect fit, $\widehat{C} \sim \chi^2_{G-2}$.
- **Problems:**
  - How many groups?
  - Different G, different results.

Hosmer Jr, D. W., Lemeshow, S., Sturdivant, R. X. (2013). *Applied logistic regression*.

The Calibration Curve

- Let $\widehat{g} = \text{logit}(\widehat{\pi})$. What about fitting a new model:

$$\text{logit}\left\{P\left(Y=1\right)\right\} = \alpha_0 + \alpha_1\widehat{g}.$$

- If $\alpha_0 = 0$ and $\alpha_1 = 1$,

$$\text{logit}\left\{P\left(Y=1\right)\right\} = 0 + 1 \times \widehat{g} = \widehat{g}$$
$$\Downarrow$$
$$\text{logit}\left\{P\left(Y=1\right)\right\} = \text{logit}(\widehat{\pi})$$
$$\Downarrow$$
$$P\left(Y=1\right) = \hat{\pi}$$

  If perfect fit, $\widehat{\alpha}_0 = 0$ and $\widehat{\alpha}_1 = 1$.

- **Problems:**
  - Only for external validation of the model.
  - Why linear relationship?

Cox, D. (1958). Two further applications of a model for a method of binary regression. *Biometrika*.

The Calibration Curve

We assume a general polynomial relationship:

$$\text{logit}\left\{P\left(Y=1\right)\right\} = \alpha_0 + \alpha_1 \hat{g} + \alpha_2 \hat{g}^2 + \ldots + \alpha_m \hat{g}^m.$$

m?

- fixed too low ⇒ too simplistic;
- fixed too high ⇒ estimation of useless parameters;

Solution: Forward selection.

Example: ICU Data

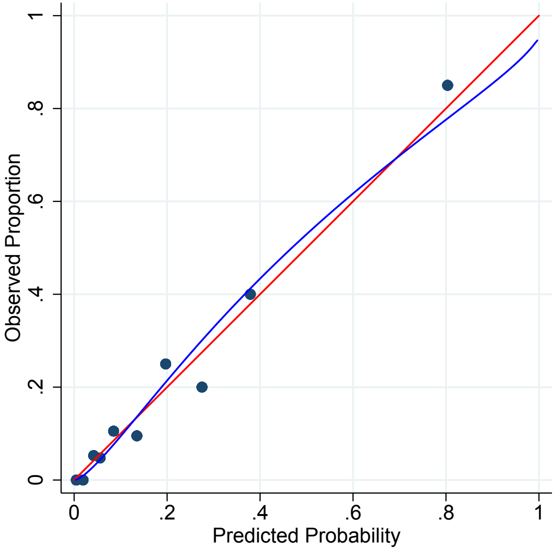
Selected polynomial is $m = 2$:

$$\text{logit}\left\{P\left(Y = 1\right)\right\} = 0.117 + 0.917\hat{g} - 0.076\hat{g}^2.$$

This defines the **calibration curve**

$$P\left(Y = 1\right) = \frac{e^{0.117 + 0.917 \mathrm{logit}(\hat{\pi}) - 0.076 (\mathrm{logit}(\hat{\pi}))^2}}{1 + e^{0.117 + 0.917 \mathrm{logit}(\hat{\pi}) - 0.076 (\mathrm{logit}(\hat{\pi}))^2}}$$

Example: ICU Data



A Goodness of Fit Test

- When m is selected, we can design a goodness of fit test on

$$\text{logit}\left\{P\left(Y=1\right)\right\} = \alpha_0 + \alpha_1 \hat{g} + \alpha_2 \hat{g}^2 + ... + \alpha_m \hat{g}^m.$$

- If perfect fit: $\alpha_1 = 1$, $\alpha_0 = \alpha_2 = ... = \alpha_m = 0$.
- A likelihood ratio test can be used to test the hypothesis

$$H_0 : \alpha_1 = 1,\ \alpha_0 = \alpha_2 = ... = \alpha_m = 0$$

- The distribution of the statistic must account for the forward selection on the same data.
- Inverting the test allows to generate a confidence region around the calibration curve: the **calibration belt**.

Nattino, G., Finazzi, S., Bertolini, G. (2016). A new test and graphical tool to assess the goodness of fit of logistic regression models. *Statistics in medicine*.

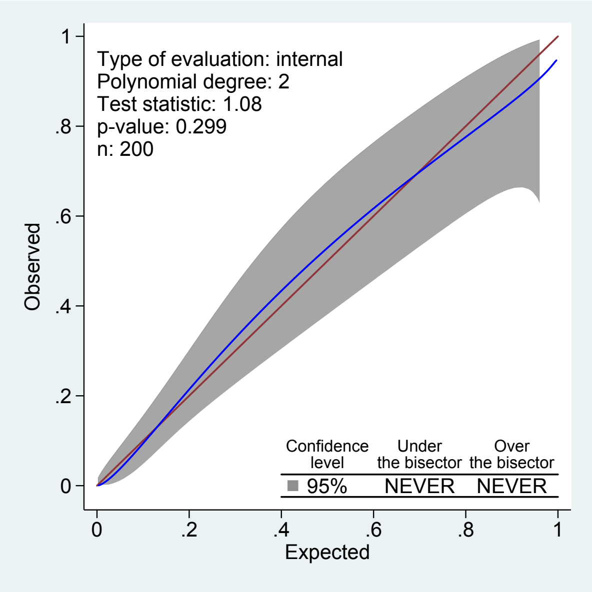
Example: ICU Data

```
. calibrationbelt
------------------------------------------------------------
               GiViTI Calibration Belt
Calibration belt and test for internal validation:
the calibration is evaluated on the training sample.
Sample size: 200
Polynomial degree: 2
Test statistic:  1.08
p-value: 0.2994
------------------------------------------------------------
```

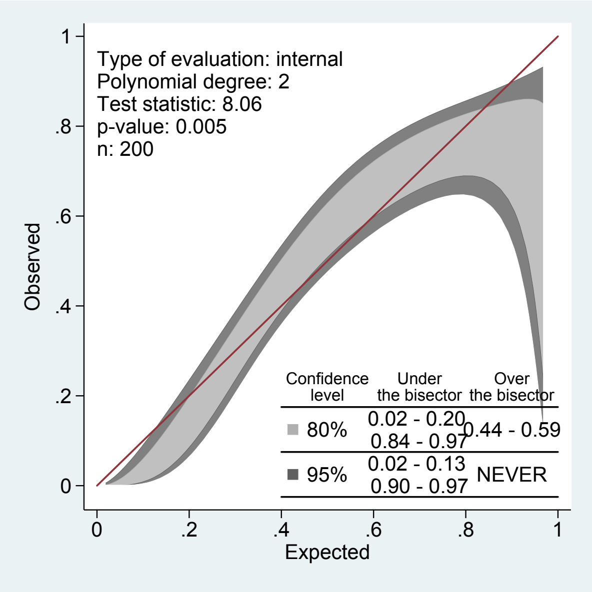
```
. estat gof, group(10)
Logistic model for sta, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
       number of observations =       200
             number of groups =        10
      Hosmer-Lemeshow chi2(8) =         4.00
                  Prob > chi2 =         0.8570
```

Nattino, G., Lemeshow, S., Phillips, G., Finazzi, S., Bertolini, G. (2017). Assessing the calibration of dichotomous outcome models with the calibration belt. *Stata Journal*

Example: ICU Data

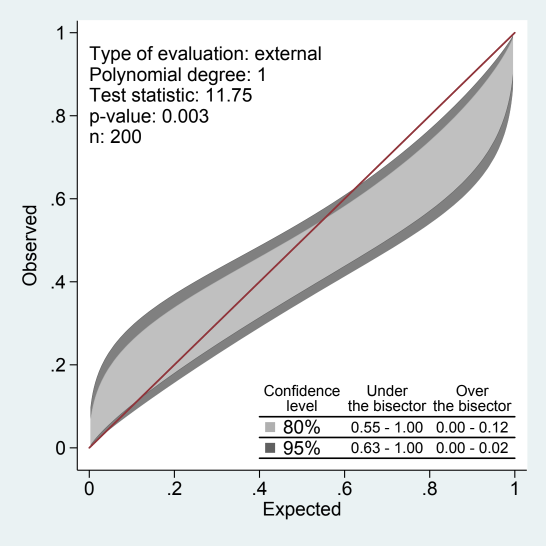


Example 2: Poorly Fitting Model



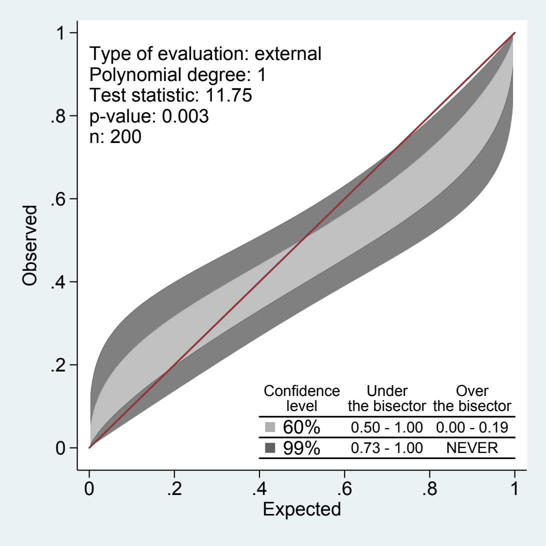
Example 3: External Validation

```
. calibrationbelt y phat, devel("external")
```



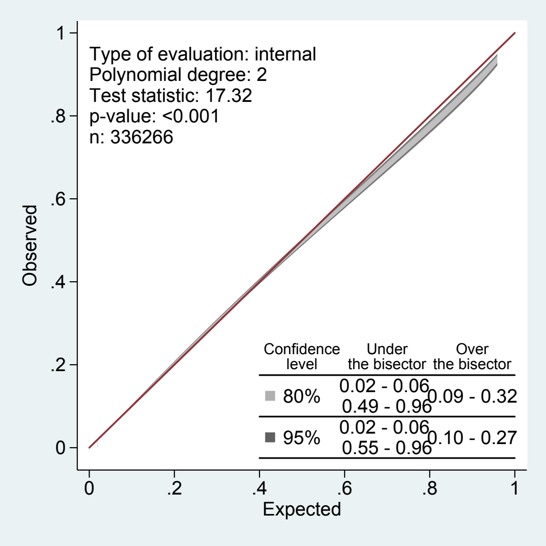
Example 3: External Validation

```
. calibrationbelt y phat, cLevel1(.99) cLevel2(.6) devel("external")
```



Example 4: Goodness of Fit and Large Samples

```
. calibrationbelt
```



Discussion

- The `calibrationbelt` command implements the calibration belt and the related test in Stata.
- Limitation:
  - Assumed polynomial relationship.
- Advantages:
  - No need of data grouping.
  - Informative tool to spot significance of deviations.
- Future work: goodness of fit in very large samples.