Title

stata.com

brier — Brier score decomposition

Description

`brier` computes the Yates, Sanders, and Murphy decompositions of the Brier mean probability score. The Brier score is a measure of disagreement between the observed outcome and a forecast (prediction).

Quick start

Brier score decomposition for binary outcome `y` and predicted probability `pvar`

```
brier y pvar
```

Same as above, but use 5 groups rather than 10 to compute the decomposition

```
brier y pvar, group(5)
```

Menu

Statistics > Epidemiology and related > Other > Brier score decomposition

Syntax

```
brier outcomevar forecastvar [if] [in] [, group(#)]
```

outcomevar is a binary variable indicating the outcome of the experiment.

forecastvar is the corresponding probability of a positive outcome and must be between 0 and 1.

`by` and `collect` are allowed; see [U] **11.1.10 Prefix commands**.

Option

Main

`group(#)` specifies the number of groups that will be used to compute the decomposition. `group(10)` is the default.

Remarks and examples

stata.com

You have a binary (0/1) response and a formula that predicts the corresponding probabilities of having observed a positive outcome (1). If the probabilities were obtained from logistic regression, there are many methods that assess goodness of fit (see, for instance, [R] **estat gof**). However, the probabilities might be computed from a published formula or from a model fit on another sample, both completely unrelated to the data at hand, or perhaps the forecasts are not from a formula at all. In any case, you now have a *test dataset* consisting of the forecast probabilities and observed outcomes. Your test dataset might, for instance, record predictions made by a meteorologist on the probability of rain along with a variable recording whether it actually rained.

The Brier score is an aggregate measure of disagreement between the observed outcome and a prediction—the average squared error difference. The Brier score decomposition is a partition of the Brier score into components that suggest reasons for discrepancy. These reasons fall roughly into three groups: 1) lack of overall calibration between the average predicted probability and the actual probability of the event in your data, 2) misfit of the data in groups defined within your sample, and 3) inability to match actual 0 and 1 responses.

Problem 1 refers to simply overstating or understating the probabilities.

Problem 2 refers to what is standardly called a goodness-of-fit test: the data are grouped, and the predictions for the group are compared with the outcomes.

Problem 3 refers to an individual-level measure of fit. Imagine that the grouped outcomes are predicted on average correctly but that within the group, the outcomes are poorly predicted.

Using logit or probit analysis to fit your data will guarantee that there is no lack of fit due to problem 1, and a good model fitter will be able to avoid problem 2. Problem 3 is inherent in any prediction exercise.

Example 1

We have data on the outcomes of 20 basketball games (win) and the probability of victory predicted by a local pundit (for).

```
. use https://www.stata-press.com/data/r18/bball
. summarize win for
    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         win |         20         .65    .4893605         0          1
         for |         20       .4785    .2147526       .15         .9

. brier win for, group(5)

Mean probability of outcome   0.6500
                of forecast   0.4785

Correlation                   0.5907
ROC area                      0.8791  p = 0.0030

Brier score                   0.1828
Spiegelhalter's z-statistic  -0.6339  p = 0.7369
Sanders-modified Brier score  0.1861
Sanders resolution            0.1400
Outcome index variance        0.2275
Murphy resolution             0.0875
Reliability-in-the-small      0.0461
Forecast variance             0.0438
Excess forecast variance      0.0285
Minimum forecast variance     0.0153
Reliability-in-the-large      0.0294
2*Forecast-Outcome-Covar      0.1179
```

The mean probabilities of forecast and outcome are simply the mean of the predicted probabilities and the actual outcomes (wins/losses). The correlation is the product-moment correlation between them.

The Brier score measures the total difference between the event (winning) and the forecast probability of that event as an average squared difference. As a benchmark, a perfect forecaster would have a Brier score of 0, a perfect misforecaster (predicts probability of win is 1 when loses and 0 when wins) would have a Brier score of 1, and a fence sitter (forecasts every game as 50/50) would have a Brier score of 0.25. Our pundit is doing reasonably well.

Spiegelhalter's z statistic is a standard normal test statistic for testing whether an individual Brier score is extreme. The ROC area is the area under the receiver operating curve, and the associated test is a test of whether it is greater than 0.5. The more accurate the forecast probabilities, the larger the ROC area.

The Sanders-modified Brier score measures the difference between a grouped forecast measure and the event, where the data are grouped by sorting the sample on the forecast and dividing it into approximately equally sized groups. The difference between the modified and the unmodified score is typically minimal. For this and the other statistics that require grouping—the Sanders and Murphy resolutions and reliability-in-the-small—to be well defined, group boundaries are chosen so as not to allocate observations with the same forecast probability to different groups. This task is done by grouping on the forecast using `xtile, n(#)`, with # being the number of groups; see [D] **pctile**.

The Sanders resolution measures error that arises from statistical considerations in evaluating the forecast for a group. A group with all positive or all negative outcomes would have a Sanders resolution of 0; it would most certainly be feasible to predict exactly what happened to each member of the group. If the group had 40% positive responses, on the other hand, a forecast that assigned $p = 0.4$ to each member of the group would be a good one, and yet, there would be "errors" in

the squared difference sense. The "error" would be $(1 - 0.4)^2$ or $(0 - 0.4)^2$ for each member. The Sanders resolution is the average across groups of such "expected" errors. The 0.1400 value in our data from an overall Brier score of 0.1828 or 0.1861 suggests that a substantial portion of the "error" in our data is inherent.

Outcome index variance is just the variance of the outcome variable. This is the expected value of the Brier score if all the forecast probabilities were merely the average observed outcome. Remember that a fence sitter has an expected Brier score of 0.25; a smarter fence sitter (who would guess $p = 0.65$ for these data) would have a Brier score of 0.2275.

The Murphy resolution measures the variation in the average outcomes across groups. If all groups have the same frequency of positive outcomes, little information in any forecast is possible, and the Murphy resolution is 0. If groups differ markedly, the Murphy resolution is as large as 0.25. The 0.0875 means that there is some variation but not a lot, and 0.0875 is probably higher than in most real cases. If you had groups in your data that varied between 40% and 60% positive outcomes, the Murphy resolution would be 0.01; between 30% and 70%, it would be 0.04.

Reliability-in-the-small measures the error that comes from the average forecast within group not measuring the average outcome within group—a classical goodness-of-fit measure, with 0 meaning a perfect fit and 1 meaning a complete lack of fit. The calculated value of 0.0461 shows some amount of lack of fit. Remember, the number is squared, and we are saying that probabilities could be just more than $\sqrt{0.0461} = 0.215$ or 21.5% off.

Forecast variance measures the amount of discrimination being attempted—that is, the variation in the forecasted probabilities. A small number indicates a fence sitter making constant predictions. If the forecasts were from a logistic regression model, forecast variance would tend to increase with the amount of information available. Our pundit shows considerable forecast variance of 0.0438 (standard deviation $\sqrt{0.0438} = 0.2093$), which is in line with the reliability-in-the-small, suggesting that the forecaster is attempting as much variation as is available in these data.

Excess forecast variance is the amount of actual forecast variance over a theoretical minimum. The theoretical minimum—called the minimum forecast variance—corresponds to forecasts of p_0 for observations ultimately observed to be negative responses and p_1 for observations ultimately observed to be positive outcomes. Moreover, p_0 and p_1 are set to the average forecasts made for the ultimate negative and positive outcomes. These predictions would be just as good as the predictions the forecaster did make, and any variation in the actual forecast probabilities above this is useless. If this number is large, above 1%–2%, then the forecaster may be attempting more than is possible. The 0.0285 in our data suggests this possibility.

Reliability-in-the-large measures the discrepancy between the mean forecast and the observed fraction of positive outcomes. This discrepancy will be 0 for forecasts made by most statistical models—at least when measured on the same sample used for estimation—because they, by design, reproduce sample means. For our human pundit, the 0.0294 says that there is a $\sqrt{0.0294}$, or 17-percentage-point, difference. (This difference can also be found by calculating the difference in the averages of the observed outcomes and forecast probabilities: $0.65 - 0.4785 = 0.17$.) That difference, however, is not significant, as we would see if we typed `ttest win=for`; see [R] **ttest**. If these data were larger and the bias persisted, this difference would be a critical shortcoming of the forecast.

Twice the forecast-outcome covariance is a measure of how accurately the forecast corresponds to the outcome. The concept is similar to that of R^2 in linear regression.

◁

Stored results

`brier` stores the following in `r()`:

Scalars

`r(p_roc)`	one-sided p-value for ROC area test
`r(roc_area)`	ROC area
`r(z)`	Spiegelhalter's z statistic
`r(p)`	one-sided p-value for Spiegelhalter's z test
`r(brier)`	Brier score
`r(brier_s)`	Sanders-modified Brier score
`r(sanders)`	Sanders resolution
`r(oiv)`	outcome index variance
`r(murphy)`	Murphy resolution
`r(relinsm)`	reliability-in-the-small
`r(Var_f)`	forecast variance
`r(Var_fex)`	excess forecast variance
`r(Var_fmin)`	minimum forecast variance
`r(relinla)`	reliability-in-the-large
`r(cov_2f)`	2×forecast-outcome-covariance

Methods and formulas

See Wilks (2019) or Schmidt and Griffith (2005) for a discussion of the Brier score.

Let d_j, $j = 1, \ldots, N$, be the observed outcomes with $d_j = 0$ or $d_j = 1$, and let f_j be the corresponding forecasted probabilities that d_j is 1, $0 \leq f_j \leq 1$. Assume that the data are ordered so that $f_{j+1} \geq f_j$ (`brier` sorts the data to obtain this order). Divide the data into K nearly equally sized groups, with group 1 containing observations 1 through $j_2 - 1$, group 2 containing observations j_2 through $j_3 - 1$, and so on.

Define

$$\overline{f}_0 = \text{average } f_j \text{ among } d_j = 0$$
$$\overline{f}_1 = \text{average } f_j \text{ among } d_j = 1$$
$$\overline{f} = \text{average } f_j$$
$$\overline{d} = \text{average } d_j$$
$$\widetilde{f}_k = \text{average } f_j \text{ in group } k$$
$$\widetilde{d}_k = \text{average } d_j \text{ in group } k$$
$$\widetilde{n}_k = \text{number of observations in group } k$$

The Brier score is $\sum_j (d_j - f_j)^2 / N$.

The Sanders-modified Brier score is $\sum_j (d_j - \widetilde{f}_{k(j)})^2 / N$.

Let p_j denote the true but unknown probability that $d_j = 1$. Under the null hypothesis that $p_j = f_j$ for all j, Spiegelhalter (1986) determined that the expectation and variance of the Brier score is given by the following:

$$E(\text{Brier}) = \frac{1}{N}\sum_{j=1}^{N} f_j(1 - f_j)$$

$$\text{Var}(\text{Brier}) = \frac{1}{N^2}\sum_{j=1}^{N} f_j(1 - f_j)(1 - 2f_j)^2$$

Denoting the observed value of the Brier score by $O(\text{Brier})$, Spiegelhalter's z statistic is given by

$$Z = \frac{O(\text{Brier}) - E(\text{Brier})}{\sqrt{\text{Var}(\text{Brier})}}$$

The corresponding p-value is given by the upper-tail probability of Z under the standard normal distribution.

The area under the ROC curve is estimated by applying the trapezoidal rule to the empirical ROC curve. This area is Wilcoxon's test statistic, so the corresponding p-value is just that of a one-sided Wilcoxon test of the null hypothesis that the distribution of predictions is constant across the two outcomes.

The Sanders resolution is $\sum_k \widetilde{n}_k\{\widetilde{d}_k(1 - \widetilde{d}_k)\}/N$.

The outcome index variance is $\overline{d}(1 - \overline{d})$.

The Murphy resolution is $\sum_k \widetilde{n}_k(\widetilde{d}_k - \overline{d})^2/N$.

Reliability-in-the-small is $\sum_k \widetilde{n}_k(\widetilde{d}_k - \widetilde{f}_k)^2/N$.

The forecast variance is $\sum_j (f_j - \overline{f})^2/N$.

The minimum forecast variance is $\left\{\sum_{j\in F}(f_j - \overline{f}_0)^2 + \sum_{j\in S}(f_j - \overline{f}_1)^2\right\}/N$, where F is the set of observations for which $d_j = 0$ and S is the complement.

The excess forecast variance is the difference between the forecast variance and the minimum forecast variance.

Reliability-in-the-large is $(\overline{f} - \overline{d})^2$.

Twice the outcome covariance is $2(\overline{f}_1 - \overline{f}_0)\overline{d}(1 - \overline{d})$.

Glenn Wilson Brier (1913–1998) was an American meteorological statistician who, after obtaining degrees in physics and statistics, was for many years head of meteorological statistics at the U.S. Weather Bureau in Washington, DC. In the latter part of his career, he was associated with Colorado State University. Brier worked especially on verification and evaluation of predictions and forecasts, statistical decision making, the statistical theory of turbulence, the analysis of weather modification experiments, and the application of permutation techniques.

Acknowledgment

We thank Richard Goldstein for his contributions to this improved version of `brier`.

References

Brier, G. W. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78: 1–3. https://doi.org/10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

Goldstein, R. 1996. sg55: Extensions to the brier command. *Stata Technical Bulletin* 32: 21–22. Reprinted in *Stata Technical Bulletin Reprints*, vol. 6, pp. 133–134. College Station, TX: Stata Press.

Hadorn, D. C., E. B. Keeler, W. H. Rogers, and R. H. Brook. 1993. *Assessing the Performance of Mortality Prediction Models*. Santa Monica, CA: Rand.

Holloway, L., and P. W. Mielke, Jr. 1998. Glenn Wilson Brier 1913–1998. *Bulletin of the American Meteorological Society* 79: 1438–1439.

Jolliffe, I. T., and D. B. Stephenson, ed. 2012. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. 2nd ed. Chichester, UK: Wiley.

Murphy, A. H. 1973. A new vector partition of the probability score. *Journal of Applied Meteorology* 12: 595–600. https://doi.org/10.1175/1520-0450(1973)012⟨0595:ANVPOT⟩2.0.CO;2.

——. 1997. Forecast verification. In *Economic Value of Weather and Climate Forecasts*, ed. R. W. Katz and A. H. Murphy, 19–74. Cambridge: Cambridge University Press.

Redelmeier, D. A., D. A. Bloch, and D. H. Hickam. 1991. Assessing predictive accuracy: How to compare Brier scores. *Journal of Clinical Epidemiology* 44: 1141–1146. https://doi.org/10.1016/0895-4356(91)90146-z.

Sanders, F. 1963. On subjective probability forecasting. *Journal of Applied Meteorology* 2: 191–201. https://doi.org/10.1175/1520-0450(1963)002⟨0191:OSPF⟩2.0.CO;2.

Schmidt, C. H., and J. L. Griffith. 2005. Multivariate classification rules: Calibration and discrimination. In Vol. 2 of *Encyclopedia of Biostatistics*, ed. P. Armitage and T. Colton, 3492–3494. Chichester, UK: Wiley.

Spiegelhalter, D. J. 1986. Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine* 5: 421–433. https://doi.org/10.1002/sim.4780050506.

Von Storch, H., and F. W. Zwiers. 1999. *Statistical Analysis in Climate Research*. Cambridge: Cambridge University Press.

Wilks, D. S. 2019. *Statistical Methods in the Atmospheric Sciences*. 4th ed. Amsterdam: Elsevier.

Yates, J. F. 1982. External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance* 30: 132–156. https://doi.org/10.1016/0030-5073(82)90237-9.

Also see

[R] **logistic** — Logistic regression, reporting odds ratios

[R] **logit** — Logistic regression, reporting coefficients

[R] **probit** — Probit regression