Goodness of Fit Under Model Misspecification

Zheng Yuan

November 2019

1 Goals

- **Establish an appropriate measurement** to quantify the goodness of model fitting, especially in model misspecification situation
- Hopefully from this measurement we can prove that LOOBIC could work better than BIC in some model misspecification cases, in other words, the model selected by LOOBIC criterion performs better than that selected by BIC criterion **according to this measurement, given certain conditions**.

2 Measurement

The goal is to seek to quantify the discrepancy between the fitting model and the generating model in a manner that (i)being small when the two models are close; (ii) detect when these two models are largely different. A natural starting point is to consider **the expected prediction error**.
Now consider the measurement in more detail. Let the data sample

$$Y = (Y_1, Y_2, ..., Y_n)^T$$

consist of independent random observations and belong to the generating model M_0.

Case 1: Conditional on Y
In this case, we assume that Y from the generating model M_0 is given.
Define

$$d(M, M_0) = E_0[(y^\star - x^{\star T}\hat{\beta}_M)^2|Y] = \int_R (y(x^\star, M_0) - x^{\star T}\hat{\beta}_M)^2 \mu_0(dy),$$

where E_0 denotes the expectation taken with respect to the new data point $y^\star$, which is from generating model M_0, $\hat{\beta}_M$ denotes the OLS estimator under the assumption that Y comes from model M. Specifically, if model M is a linear model

$$Y = X_M\beta_M + \epsilon,\ \epsilon \sim N(0, \sigma^2 I_n)$$

, then

$$\hat{\beta}_M = (X_M^T X_M)^{-1} X_M^T Y.$$

Note that if the generating model M_0 is fixed, then $d(M, M_0)$ only depends on model M, therefore, we denote

$$h(M) = d(M, M_0) = E_0[(y^\star - x^{\star T}\hat{\beta}_M)^2 | Y]$$

Case 2: Unconditional on Y

Now we extend case 1 to the case where Y is not given. Then we just add another layer of expectation based on the discrepancy in case 1 and define

$$\begin{aligned} D(M, M_0) &= E_0(d(M, M_0)) \\ &= E_0[E_0[(y^\star - x^{\star T}\hat{\beta}_M)^2 | Y]] \\ &= E_0[(y^\star - x^{\star T}\hat{\beta}_M)^2], \end{aligned}$$

where E_0 denotes the expectation taken with respect to both $y^\star$ and Y.

Also, for fixed generating model M_0, $D(M, M_0)$ only depends on model M and we let

$$H(M) = D(M, M_0) = E_0[E_0[(y^\star - x^{\star T}\hat{\beta}_M)^2 | Y]]$$

We call this $H(M)$ **the goodness-of-fit measure for model** M .

3 Estimation of $H(M)$

By **the Strong Law of Large Numbers**,

$$\frac{1}{N}\sum_{j=1}^{N}(y_j^\star - x_j^{\star T}\hat{\beta}_M)^2 \to d(M, M_0)$$

almost surely given Y as $n \to \infty$
and

$$\frac{1}{M}\sum_{k=1}^{M}(\frac{1}{N}\sum_{j=1}^{N}(y_j^\star - x_j^{\star T}\hat{\beta}_M^k)^2) \to D(M, M_0)$$

almost surely as $n \to \infty$.
This provides us a mechanism to estimate $H(M)$:

1. For each $k \in \{1, 2, 3..., M\}$, generates Y_K from Model M_0, based on Y_K, OLS estimator $\hat{\beta}_M^k$ is also determined;
2. For each $j \in \{1, 2, 3..., N\}$, generate $y_j^\star$ from Model M_0, calculate its prediction error under Model M, then calculate the average prediction error with respect to each Y_k: $\frac{1}{N}\sum_{j=1}^{N}(y_j^\star - x_j^{\star T}\hat{\beta}_M^k)^2)$
3. Evaluate the overall average prediction error across different Y_k's: $\frac{1}{M}\sum_{k=1}^{M}(\frac{1}{N}\sum_{j=1}^{N}(y_j^\star - x_j^{\star T}\hat{\beta}_M^k)^2)$