Bias-Variance Dilemma

Park Sukjin

Department of Economics, Sogang University

October 20, 2021

Error Decomposition

- Consider N observations, $\{(x_n, y_n)\}_{n \in (1,N)}$, and a function $f(x_n)$ that predicts y_n.
- The errors, $\epsilon_n = y - f(x_n)$ are unpredictable, with $\epsilon_n \sim N(0, \sigma_\epsilon^2)$.
- The true model $f(x_n)$ is unknown to us, so instead we propose a statistical model $\hat{f}(x_n)$ which approximates $f(x_n)$ with errors $e = y_n - \hat{f}(x_n)$.

Error Decomposition

- We would like to minimze $E\left[(f(x_n)-\hat{f}(x_n))^2\right]$ (mean squared true error), but since $f(x_n)$ is unknown, the best we can do is to minimize $E\left[(y-\hat{f}(x_n))^2\right]$ (mean squared empirical error).
- $E\left[(y-\hat{f}(x_n))^2\right]$ can be decomposed as

$$\text{MSE} = E\left[e_n^2\right] = \left(E\left[(f(x_n)-\hat{f}(x_n))\right]\right)^2 + V\left[\hat{f}(x_n)\right] + \sigma_\epsilon^2$$

.

- $\left(E\left[(f(x_n)-\hat{f}(x_n))\right]\right)^2$: bias^2
- $V\left[\hat{f}(x_n)\right]$: variance
- σ_ϵ^2: noise

Error Decomposition

- Thus, the error of a prediction model has three components:
  - Bias: error due to wrong assumptions
  - Variance: error due to overfitting
  - Noise: Unpredictable error, due to the random nature of the variable.
- The key difference between Econometrics and ML is how they tackle the Bias-Variance dilemma.

- (Econometrics) Fit the estimator with minimum variance among all unbiased estimators (i.e., MVUE, BLUE).
- (ML) Find a balance between bias and variance to maximize performance (i.e., minimize MSE).

Best Linear Unbiased Estimator

- Classical assumptions
  - $f(x_n)$ is linear
  - $\hat{f}(x_n)$ incorporates all k variables (no omitted variables).
  - e_n is white noise, and uncorrelated to x_n.
- Then, according to the Gauss-Markov theorem the OLS (ordinary least squares) estimator is BLUE (best linear unbiased estimator).
- In order words, when the true model is linear, there are no omitted variables, and errors are white noise, then OLS gives us the MVUE.
- The Gauss-Markov theorem is the fundamental reason why OLS is the standard model in Econometrics.

Best Linear Unbiased Estimator

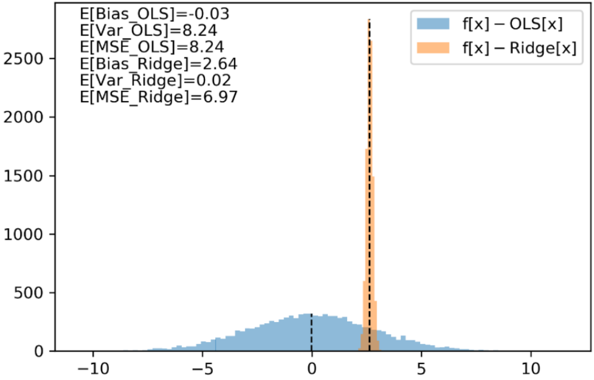
- Classical assumptions
  - $f(x_n)$ is linear
  - $\hat{f}(x_n)$ incorporates all k variables (no omitted variables).
  - e_n is white noise, and uncorrelated to x_n.
- Then, according to the Gauss-Markov theorem the OLS (ordinary least squares) estimator is BLUE (best linear unbiased estimator).
- In order words, when the true model is linear, there are no omitted variables, and errors are white noise, then OLS gives us the MVUE.
- The Gauss-Markov theorem is the fundamental reason why OLS is the standard model in Econometrics.

Best Linear Unbiased Estimator

- There are three common violations of the classical assumptions in financial datasets
  - Specification errors (e.g., non-linear, omitted variables)
  - Stochastic regressors (e.g., measurement errors, multicollinearity, endogenous variables)
  - Non-spherical errors (heteroskedasticity, autocorrelation)
- As a result, econometric models often are inefficient (high variance, bias).

Best Linear Unbiased Estimator

- What if the classical assumptions hold?
- Even if all GMT assumptions are correct, OLS does not necessarily yield the minimum MSE
  - non-linear biased estimators with lower variance
  - linear biased estimators with lower MSE (e.g., Ridge, Lasso)
- Econometrics' reliance on unbiased estimators (including, but not limited to OLS) is not justified by model performance


E[Bias_OLS]=-0.03
E[Var_OLS]=8.24
E[MSE_OLS]=8.24
E[Bias_Ridge]=2.64
E[Var_Ridge]=0.02
E[MSE_Ridge]=6.97
f[x] − OLS[x]
f[x] − Ridge[x]
2500
2000
1500
1000
500
0
−10
−5
0
5
10

Best Linear Unbiased Estimator

- Financial systems are too complex for expecting well-specified (or, unbiased) models.
- Accordingly, minimizing variance subject to zero bias may be the wrong objective.
- Even if all classical assumptions are correct, OLS does not necessarily yield the minimum MSE.
- ML methods can achieve lower MSE than econometric models.

ML methods to solve the Bias-Variance Dilemma

- Cross-validation
  - K-fold
- Complexity reduction
  - Regularization
  - Feature selection, dimensionality reduction
- Ensemble methods
  - Bagging
  - Boosting
  - Stacking