A Short Note on Regression Bias

Ioana Boier
ioanaboier@gmail.com

December 8, 2016

Abstract

This short note summarizes results around the question of bias of linear regression coefficients when overfitting or underfitting.

1 Definition, Notations, Assumptions

Definition. The *bias* of a statistical estimator of a stochastic parameter is the difference between its expected value and the true value of the parameter being estimated.
In this note we consider linear regression models parameterized by a coefficient vector β:

$$y = X\beta + \epsilon$$

and we denote by b the least squares estimator:

$$b = (X^T X)^{-1} X^T y$$

We assume that the residuals satisfy: $E[\epsilon_i | X_{j*}] = 0$ for all i, j.

2 Analysis

2.1 General

$$E[b|X] = E[(X^T X)^{-1} X^T y|X] = E[(X^T X)^{-1} X^T (X\beta + \epsilon)|X]$$
$$E[b|X] = E[\beta + (X^T X)^{-1} X^T \epsilon)|X]$$
$$E[b|X] = \beta + E[(X^T X)^{-1} X^T \epsilon)|X]$$

Using the residual assumption and the fact that $E[f(X)\epsilon|X] = f(X)E[\epsilon|X] = 0$, it follows that:

$$E[b|X] = \beta$$

Applying the tower property:

$$E[b] = E[E[b|X]] = E[\beta] = \beta$$

i.e., b is an unbiased estimator of β.

2.2 Overfitting

Let us first consider the case in which some subset X_2 of the independent variables is superfluous. The regression can be reformulated as:

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

In reality, $\beta_2 = 0$. However, this does not change the previous proof that $E[b|X] = \beta$ which remains valid even when some of the components β_2 of the β vector are actually zero. Hence, overfitting does not introduce bias in the regression coefficients.

2.3 Underfitting

This is the case when relevant variables X_2 are left out of the model specification and we regress $y = X_1\beta_1 + \epsilon$ instead of $y = X_1\beta_1 + X_2\beta_2 + \epsilon$. In this case, the estimator is:

$$b_1 = (X_1^T X_1)^{-1} X_1^T y = \beta_1 + (X_1^T X_1)^{-1} X_1^T X_2 \beta_2 + (X_1^T X_1)^{-1} X_1^T \epsilon$$

and its expected value:

$$E[b_1|X] = \beta_1 + (...) * \beta_2$$

which leads to the conclusion that b_1 is biased except for cases when the term $(...) * \beta_2 = 0$.

2.4 What Does It Mean In Practice?

At first glance, from a theoretical standpoint, it would appear that overfitting is "better" since including superfluous variables does not introduce bias in the regression coefficients whereas underfitting does. This is where the free lunch suspicions should kick in: what is the hidden cost of overfitting? Throwing in all manner of variables without regards of how they correlate with other variables and magically finding that they are superfluous sounds too good to be true. Indeed, the price to pay comes in the form of reduced precision of coefficient estimates. Their variances could, in fact, blow up. For an excellent, in depth treatment see Chapter 4 in [1].

References

[1] Greene, W. H. *Econometric Analysis.* Prentice Hall, 5th edition, 2003.