Evaluating Methods: Bias

Prof. Dennis Sun

Data 401

Statistical Framework for Evaluating Methods

Given data, we assume a probability model for the data and estimate the parameters in that model.

\[
\begin{align*}
\theta_0 & \xrightarrow{p_{\theta_0}(x)} \text{Data} \ x \xrightarrow{L(\theta) = p_\theta(x)} \hat{\theta}
\end{align*}
\]

How do we know if the estimate is any good or not?

Statisticians assume that there is some true underlying model that generated the data and study how \( \hat{\theta} \) compares to \( \theta_0 \).

Notice that there are 3 “\( \theta \)”s floating around:

- \( \theta_0 \) is the true value of the parameter
- \( \theta \) is the variable that we optimize over to get the MLE
- \( \hat{\theta} \) is the estimated value of the parameter
Bias

One metric that we can calculate is the bias:

\[ \text{bias} = E[\hat{\theta}] - \theta_0. \]

Example

For example, let’s model data \( X_1, ... X_n \) as i.i.d. Poisson(\( \mu \)).

We saw that the MLE is \( \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i \). Is this a good estimator?

Assume that \( X_1, ..., X_n \) are actually generated from Poisson(\( \mu_0 \)).

\[ E[\hat{\mu}] = \]

The Bias of Linear Regression

Given data \( (x_1, Y_1), ..., (x_n, Y_n) \), we assumed the simple linear regression model (without intercept) \( Y_i \sim \text{Normal}(x_i \beta, \sigma^2) \) and calculated the MLE of \( \beta \) as

\[ \hat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}. \]

What is the bias of \( \hat{\beta} \)?

The argument generalizes to multiple regression.