

# Biased Estimators

Hossein Falaki  
Computer Science Department,  
University of California, Los Angeles  
falaki@cs.ucla.edu

## 1 Introduction

This note introduces Stein's estimator as a biased estimator. We will prove that the testing error of a biased estimator, such as Stein's estimator, can be lower than the least squares estimator which is proved to be the best unbiased linear estimator.

## 2 Estimation Error Analysis

In the previous note we defined estimation error but did not derive it. Using the same nonlinear model assumption:

$$Y = f(X) + \epsilon \quad (1)$$

and uncorrelated zero-mean, and constant variance noise assumption:

$$E(\epsilon_i) = 0 \quad (2)$$

$$Var(\epsilon_i) = \sigma^2 \quad (3)$$

we can derive the estimation error:

$$\begin{aligned} E(\|f - \hat{Y}\|^2) &= E(\|f - \hat{f} + \hat{\epsilon}\|^2) \\ &= E(\|f - \hat{f}\|^2 + \|\hat{\epsilon}\|^2 + 2 \langle f - \hat{f}, \hat{\epsilon} \rangle) \\ &= E(\|f - \hat{f}\|^2) + E(\|\hat{\epsilon}\|^2) \\ &= \underbrace{\|f - \hat{f}\|^2}_{\text{model bias}} + \underbrace{p\sigma^2}_{\text{variance}} \end{aligned} \quad (4)$$

The estimation error consists of two parts: *model bias* and *variance*. Figure 2 illustrates the effect of increasing model complexity on both of these error components. The least squares estimator guarantees the lowest

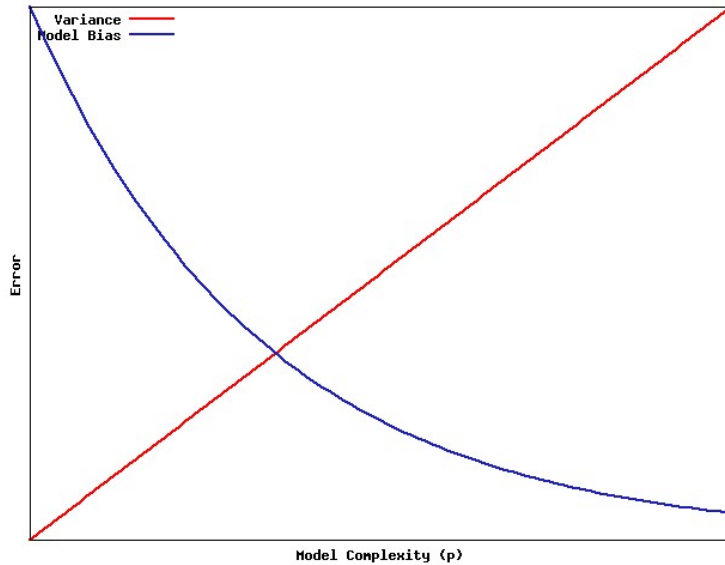


Figure 1: Effect of model complexity on model bias and variance

model bias, but this does not mean it will achieve the best trade-off between model bias and variance.

In fact Stein proved that it is possible to increase the bias of an estimator ( $\tilde{\beta}$ ) and reduce its variance in a way that the sum (estimation error) reduces at the end. Such an estimator is called a *biased estimator*.

We will introduce Stein's estimator and prove its superiority over the least squares estimator for  $p > 2$ .

### 3 Stein's Estimator

Stein used *shrinkage* to improve the least squares estimator. He took the best linear unbiased estimator (BLUE) and shrank it in this way:

$$\tilde{\beta}_{stein} = \left(1 - \frac{(p-2)\sigma^2}{\|\hat{\beta}_{LS}\|^2}\right)\hat{\beta}_{LS} \quad (5)$$

We will prove that for all  $p > 2$ :

$$E(\|\tilde{\beta}_{stein} - \beta_{best}\|^2) \leq E(\|\hat{\beta}_{LS} - \beta_{best}\|^2) \quad (6)$$

Before getting to the proof, consider the following lemma, known as Stein's Lemma.

**Lemma 3.1.** *If  $Z \sim N(\mu, \sigma^2)$ , then:*

$$E((Z - \mu)g(z)) = \sigma^2 E(g'(z)) \quad (7)$$

*Proof.* Knowing that:

$$de^{-\frac{(z-\mu)^2}{2\sigma^2}} = e^{-\frac{(z-\mu)^2}{2\sigma^2}} \cdot \left(\frac{-1}{\sigma^2}(z-\mu)\right) dz$$

and using integration by parts we can prove this lemma:

$$\begin{aligned} E((z-\mu)g(z)) &= \int (z-\mu)g(z) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}} dz \\ &= \sigma^2 g(z) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}} \Big|_{-\infty}^{+\infty} + \sigma^2 \int g'(z) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}} \\ &= \sigma^2 E(g'(z)) \end{aligned}$$

□

Using this lemma, we can now prove 6.

*Proof.* To simplify the proof, lets simplify the terminology in the following way:

$$\begin{aligned} \theta_{p \times 1} &= \beta_{best} \\ \tilde{\theta}_{p \times 1} &= \tilde{\beta}_{stein} \\ X_{p \times 1} &= \hat{\beta}_{LS} \end{aligned}$$

Using this new notation we have the right hand side of 6 (estimation error of the unbiased estimator) is:

$$E(\|X - \theta\|^2) = E\left(\sum_{i=0}^p (x_i - \theta_i)^2\right) = p\sigma^2$$

□

The estimation error of Stein's estimator (left hand side) is:

$$\begin{aligned} E(\|X - \tilde{\theta}\|^2) &= E\left(\|X - \left(\frac{(p-2)\sigma^2}{\|X\|} X - \theta\right)\|^2\right) \\ &= E\left(\|(X - \theta) + \frac{(p-2)\sigma^2}{\|X\|} X\|^2\right) \\ &= E\left(\|X - \theta\|^2 + \frac{(p-2)^2\sigma^4}{\|X\|^2} - 2\langle X - \theta, \frac{(p-2)\sigma^2}{\|X\|^2} X \rangle\right) \end{aligned}$$

We can use Lemma 7 to calculate the last term:

$$\begin{aligned}
E(\langle X - \theta, \frac{(p-2)\sigma^2}{\|X\|^2} X \rangle) &= E\left(\sum_{i=1}^p (x_i - \theta_i) \frac{(p-2)\sigma^2}{x_1^2 + \dots + x_p^2} x_i\right) \\
&= \sum_{i=1}^p E\left((x_i - \theta_i) \underbrace{\frac{(p-2)\sigma^2}{x_i^2 + \sum_{j \neq i} x_j^2} x_i}_{g(x_i)}\right) \\
&= \sigma^2 \sum_{i=1}^p E(g'(x_i)) \\
&= \sigma^2 \sum_{i=1}^p E\left(\frac{(p-2)\sigma^2}{\|X\|^2} - \frac{(p-2)\sigma^2 \cdot 2x_i^2}{\|X\|^4}\right) \\
&= \sigma^2 E\left(\frac{p(p-2)\sigma^2}{\|X\|^2} - \frac{2(p-2)\sigma^2}{\|X\|^2}\right) \\
&= E\left(\frac{(p-2)\sigma^4}{\|X\|^2}\right)
\end{aligned}$$

Therefore:

$$\begin{aligned}
E(\|X - \tilde{\theta}\|^2) &= E\left(\|X - \theta\|^2 + \frac{(p-2)^2\sigma^4}{\|X\|^2} - 2\langle X - \theta, \frac{(p-2)\sigma^2}{\|X\|^2} X \rangle\right) \\
&= E\left(\|X - \theta\|^2 + \frac{(p-2)^2\sigma^4}{\|X\|^2} - 2\frac{(p-2)^2\sigma^4}{\|X\|^2}\right) \\
&= E\left(\|X - \theta\|^2 - \frac{(p-2)^2\sigma^4}{\|X\|^2}\right) \leq p\sigma^2
\end{aligned}$$

## 4 Biased Estimators

It is an interesting question to ask: *Why shrinkage reduces estimation error?*

We can conceive shrinkage as restricting the estimator to some limit. Such a constraint will result in the same shrinking technique.

This idea can be formulated as:

$$\text{Minimize } \|Y - X\beta\|^2$$

subject to  $\|\beta\| < \text{threshold}$ .

This constraint optimization can be translated to the following Lagrange equation:

$$\text{minimize } \|Y - X\beta\|^2 + \underbrace{\lambda\|\beta\|^2}_{\text{penalty}}$$

$$\frac{\partial}{\partial \beta} : -2X^T(Y - X\beta) + 2\lambda\beta = 0$$

$$\Rightarrow (X^T X + \lambda)\beta = X^T Y$$

$$\Rightarrow \beta = (X^T X + \lambda)^{-1} X^T Y$$

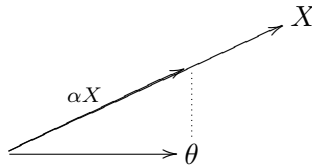
Assuming that  $X^T X = I_p$ :

$$\tilde{\beta}_j = \frac{\langle Y, X_j \rangle}{1 + \lambda}$$

This constrain on  $\tilde{\beta}$  prevents it from chasing noise. Or as we already mentioned, allows us to trade off variance and bias. For a better overall estimation error.

To better understand this consider:

$$\begin{aligned} E(\|X\|^2) &= \sum_{i=1}^p E(x_i^2) = \sum_{i=1}^p (\text{Var}(x_i) + E(x_i)^2) \\ &= \sum_{i=1}^p (\sigma^2 + \theta_i^2) = \sum_{i=1}^p \theta_i^2 + \sum_{i=1}^p \sigma^2 \\ &= \|\theta\|^2 + p\sigma^2 \end{aligned}$$



As you can see in the above diagram, shrinkage essentially reduces the error cause by variance.

#### 4.1 Estimation Bias

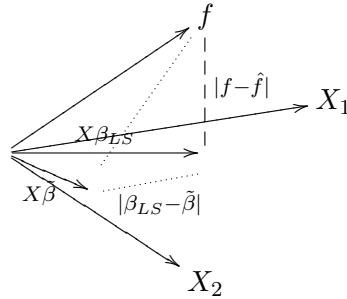
Suppose:

$$\begin{aligned} \hat{f} &= X\beta_{best} \\ \mu &= E(\tilde{\beta}) \end{aligned}$$

We can rewrite estimation error to decompose model bias and extract the estimation bias.

$$\begin{aligned}
E(\|f - \hat{Y}\|^2) &= E(\|f - \hat{f} + X\beta_{best} - X\tilde{\beta}\|^2) \\
&= \|f - \hat{f}\|^2 + E(\|\tilde{\beta} - \beta_{best}\|^2) \\
&= \|f - \hat{f}\|^2 + E(\|\tilde{\beta} - \mu + \mu - \beta_{best}\|^2) \\
&= \underbrace{\|f - \hat{f}\|^2}_{\text{model bias}} + \underbrace{\|\mu - \beta_{best}\|^2}_{\text{estimation bias}} + \underbrace{E(\|\tilde{\beta} - \mu\|^2)}_{\text{estimation variance}} \quad (8)
\end{aligned}$$

*Estimation bias* can graphically be seen in the following diagram:



The distinction between model bias and estimation bias is not very clear. Consider this example: Suppose  $p = 10,000$ . If we set all but 100 of the  $\beta_j$ s to zero, both of the following interpretations are correct:

1. We have changed the model, such that  $p = 100$ , therefore we have increased *model bias*
2. We have not changed the model. The estimation method, estimates most of  $\beta_j$ s to be zero, therefore it has increased *estimation bias*

We can explain (potential) decrease in estimation error with both interpretations. The first interpretation is reducing estimation error through reducing model complexity. The second interpretation is reducing estimation error, by introducing estimation bias to reduce variance.