Bias vs variance. Section 3.2

Hastie, ESL, Chapter 3

$$\mathbb{E}[\vec{\beta}^*] = \mathbb{E}\left[(X^TX)^{-1}X^T\vec{y}\right] = \vec{\beta} + \mathbb{E}\left[(X^TX)^{-1}X^T\vec{\varepsilon}\right] = \vec{\beta}$$

$$\vec{y} = X\vec{\beta} + \vec{\varepsilon} \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

① **Bias** : $\|\mathbb{E}[\vec{\beta}^*] - \vec{\beta}\| = 0$

② $\mathbb{E}\left[(\vec{\beta}-\vec{\beta}^*)(\vec{\beta}-\vec{\beta}^*)^T\right] = \mathbb{E}\left[(X^TX)^{-1}X^T\underbrace{\vec{\varepsilon}\vec{\varepsilon}^T}_{\sigma^2 I}X(X^TX)^{-1}\right]$

|| variance of estimator

$$= (X^TX)^{-1}\sigma^2$$

$$\hat{\sigma}^2 = \frac{1}{N-(p+1)}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$\mathbb{E}[\hat{\sigma}^2] = \sigma^2$$

For an estimator θ^* of θ we have in general that:

$$\mathbb{E}\left[(\theta-\theta^*)^2\right] = \mathbb{E}\left[\theta^2\right] + \mathbb{E}(\theta^{*2}) - 2\,\mathbb{E}\left[\theta\theta^*\right] \overset{①}{=}$$

θ is the true value to be estimated ($\Rightarrow$ not a random variable)

$$\overset{①}{=} \theta^2 + \underbrace{var(\theta^*) + \mathbb{E}(\theta^*)^2}_{} - 2\,\theta\,\mathbb{E}(\theta^*)$$

$$= \underbrace{var(\theta^*)}_{\text{variance of the estimator}} + \Big[\underbrace{\theta - \mathbb{E}(\theta^*)}_{\text{Bias}^2}\Big]^2$$

How good is an estimator is a combination of as low bias and as low variance as possible.

Least squares is best unbiased linear estimator

Based on Gauss-Markov theorem:

Let $\vec{\beta}^*_{LS}$ define our LS estimator

such that $\hat{f}_{LS}(\vec{x}_0) = \vec{x}_0^T \cdot \vec{\beta}^*_{LS}$,

which is unbiased: $\mathbb{E}\left[\hat{f}(\vec{x}_0)\right] = \vec{x}_0^T \cdot \vec{\beta}$

where $\vec{\beta}$ is the true value.

Let $\hat{\hat{f}}$ be another <u>linear estimator</u>.

Gauss-Markov says:

$$\mathrm{var}\left(\hat{f}_{LS}(\vec{x}_0)\right) \leq \mathrm{var}\left(\hat{\hat{f}}(\vec{x}_0)\right)$$

As a consequence we have the following:

Since $\mathrm{MSE} = \mathbb{E}\left[\hat{f} - y\right] = \mathrm{var}(\hat{f}) + \left[\overbrace{y - \mathbb{E}[\hat{f}]}^{\text{bias}}\right]^2$

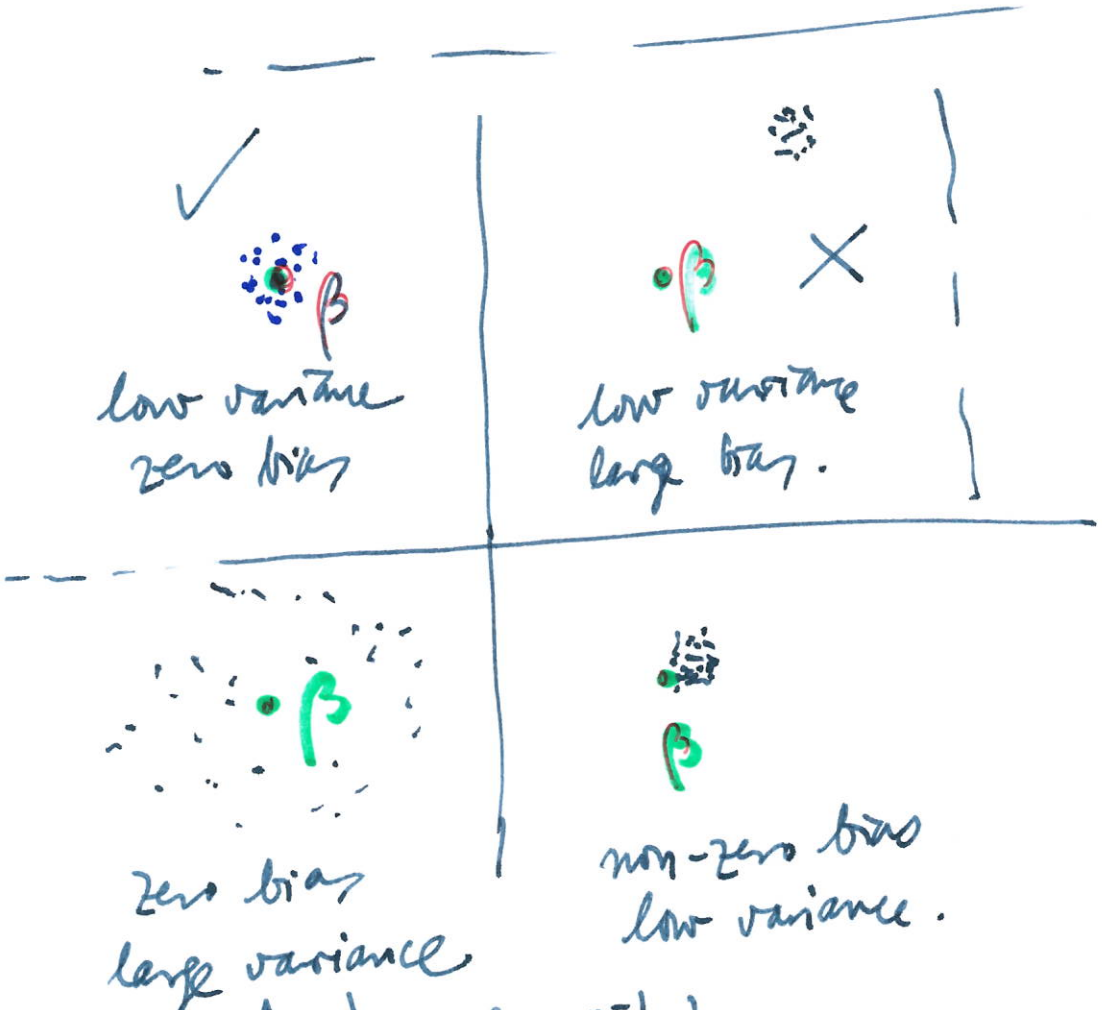
$\Rightarrow$ LS has lowest MSE of all linear unbiased estimators

[If bias $= 0$, $\mathrm{MSE} = \mathrm{var}(\hat{f})$, etc]

Quick set of pointers (sections 3.2, 3.3 ESL)

↳ Gauss-Markov

If estimator is linear and unbiased, one cannot do better than $\vec{\beta}^{*}_{LS}$



LS

$\left\{ \begin{array}{l} (X^T X)^{-1} \\ p \gg \end{array} \right\}$

Motivated by reducing variance
and increasing interpretability
~~we~~ some methods attempt to
reduce the number of descriptors, p

① Find a subset of descriptors that
are "good".

$$X = (\vec{1} \; \vec{x}_1 \; \dots \; \vec{x}_p)$$

1.1. Find the <u>best subset</u> of the
p parameters.

Declare the ~~~~ size of the subset: $\underline{\underline{k}}$

"Leaps and bounds"
works for up to $\underline{p \simeq 50}$

1.2 Sequential approaches.

Forward / Backward.

Forward: Add x_i one at a time choosing the descriptor that reduces the error maximally.

Backward: Starting from the full LS model with p parameters, chop off one by one picking the one that increases the error minimally.

Implemented thorough QR decompositi

Gram-Schmidt.

$$X = QR$$

$$Q^T Q = I$$

R upper triangular.

[Always project in orthogonal directions]

(2) Change the loss function.

Remember our optimisation formulation:

Linear regression is least squares

given $X, \vec{y}$ (our data), we find $\vec{\beta}^*$

from $\min_{\vec{\beta}} \|\vec{y} - X\vec{\beta}\|^2 = \min_{\vec{\beta}} L_{LS}(\vec{\beta})$

The solution is:

$$\vec{\beta}^* = \arg\min L_{LS}(\vec{\beta}) = X^+ \vec{y} = \underbrace{(X^TX)^{-1}}X^T\vec{y}$$

Let us introduce ~~another~~ other estimators which will be biased but with potentially less variance.

2.1 Ridge regression:

Same as above but loss is:

$$L_{ridge}(\vec{\beta}) = \|\vec{y} - X\vec{\beta}\|^2 + \lambda\|\vec{\beta}\|^2$$

The ridge solution is obtained from:

$$\boxed{\min_{\vec{\beta}} L_{ridge}(\vec{\beta}) = \min_{\vec{\beta}} \|\vec{y} - X\vec{\beta}\|^2 + \lambda \|\vec{\beta}\|^2}$$

which is equivalent to:

$$\boxed{\begin{array}{l} \min_{\vec{\beta}} \|\vec{y} - X\vec{\beta}\|^2 \\ \text{subject to } \|\vec{\beta}\|^2 \leq t \end{array}}$$

This solution can be obtained explicitly: