# Mod 1: Bias-Variance Trade Off and Ridge Regression

Ruben Zamar
Department of Statistics
UBC

October 3, 2018

# Bias versus Variance

- **Motivation:**

  Measure a quantity $\mu$

- Can either use measurement $X_1$ or measurement $X_2$
- Suppose that $X_1 \sim N\left(\mu, \sigma^2\right)$ and $X_2 \sim N\left(\gamma, \tau^2\right)$
- Assume that $\tau < \sigma$

  Otherwise $X_1$ would be obviously preferred (why?)

- Assume that $\gamma \neq \mu$

  Otherwise $X_2$ would be obviously preferred (why?)

# Loss Functions

- **Quadratic Loss**

$$Q_i = E\left[(X_i - \mu)^2\right], \quad i = 1, 2$$

- **$\mathbf{L}_1$ Loss**

$$L_i = E\left[|X_i - \mu|\right], \quad i = 1, 2$$

# "Canonical" Representation

- Assume (w.l.g.) that $\mu = 0$ :
  In fact,

$$A_i = E\left[(X_i - \mu)^2\right] = E\left(Y_i^2\right) \quad \text{with}$$

$$Y_1 \sim N\left(0, \sigma^2\right), \quad Y_2 \sim N\left(\gamma - \mu, \tau^2\right) = N\left(\delta, \tau^2\right)$$

  we can simply compare variables $Y_1$ and $Y_2$

- The assumption $\gamma \neq \mu$ becomes $|\delta| > 0$.

- This reasoning also applies to the $L_1$ Loss case

# "Canonical" Representation

- For fixed $d > 0$

$$A_1 \quad < \quad A_2 \Leftrightarrow \frac{A_1}{\sigma} < \frac{A_2}{\sigma}$$

# "Canonical" Representation

- Therefore, we can re-define

$$A_i \;=\; E\left[\left(\frac{X_i}{\sigma}\right)^2\right] = E\left[Z_i^2\right]$$

- In summary, we compare variables $Z_1$ and $Z_2$ with

$$Z_1 \;\sim\; N\left(0,1\right), \quad Z_2 \sim N\left(\frac{\delta}{\sigma},\frac{\tau^2}{\sigma^2}\right) = N\left(\Delta,\xi^2\right)$$

where $0 < \xi < 1$ and $|\Delta| > 0$

- Recall that

$$\Delta \;=\; \frac{\gamma-\mu}{\sigma} \quad \text{and} \quad \xi = \frac{\tau}{\sigma}$$

# Quadratic Loss Function

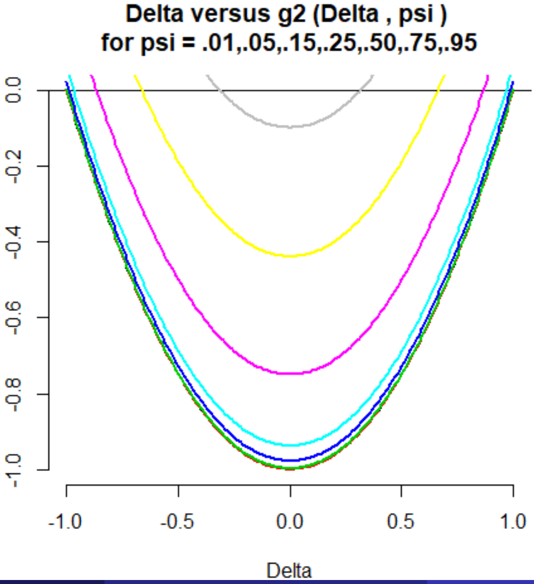
- Measurement $X_2$ is preferred if and only if

$$E\left(Z_2^2\right) \quad < \quad E\left(Z_1^2\right)$$

- That is, if and only if

$$g_2\left(\Delta, \xi\right) \quad = \quad \xi^2 + \Delta^2 - 1 < 0$$

# Quadratic Loss Function

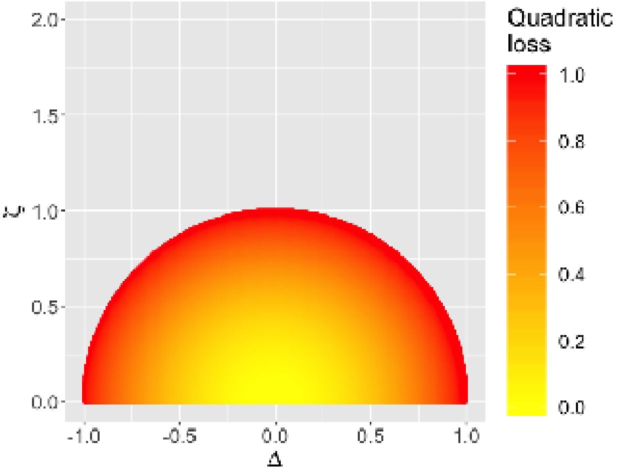


# Quadratic Loss Function

| $\xi$ | $X_2$ is Preferred if |
|---|---|
| 0.15 | $\|\Delta\| < 0.989$ |
| 0.25 | $\|\Delta\| < 0.968$ |
| 0.35 | $\|\Delta\| < 0.937$ |
| 0.45 | $\|\Delta\| < 0.893$ |
| 0.55 | $\|\Delta\| < 0.835$ |

| $\xi$ | $X_2$ is Preferred if |
|---|---|
| 0.65 | $\|\Delta\| < 0.760$ |
| 0.75 | $\|\Delta\| < 0.661$ |
| 0.85 | $\|\Delta\| < 0.527$ |
| 0.95 | $\|\Delta\| < 0.312$ |
| 0.99 | $\|\Delta\| < 0.141$ |

# Quadratic Loss Function



- We will use the formula (students should verify it analytically)

$$E\left|N\left(a, b^2\right)\right| = 2\left(b\varphi\left(a/b\right) + |a|\left[\Phi\left(|a|/b\right) - 1/2\right]\right)$$

- Using this formula

$$E\left|N\left(\Delta, \xi^2\right)\right| = 2\left(\xi\varphi\left(\Delta/\xi\right) + |\Delta|\left[\Phi\left(|\Delta|/\xi\right) - 1/2\right]\right)$$

and

$$E\left|N\left(0, 1\right)\right| = 2\varphi\left(0\right)$$

# L1 Loss Function

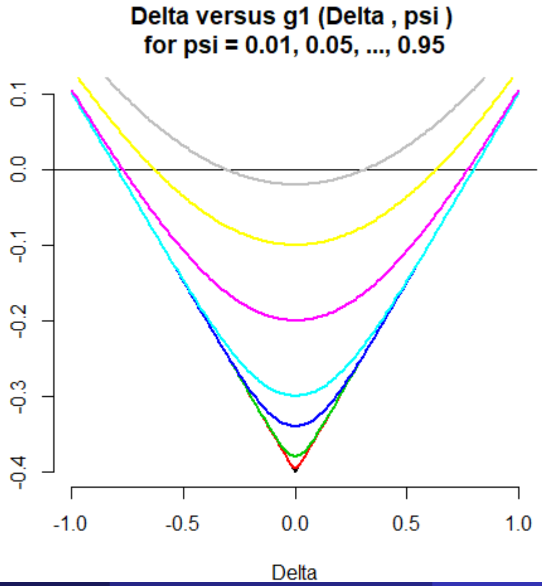
- In this case measurement $X_1$ should be preferred if and only if

$$E\left|N\left(\Delta,\xi^2\right)\right| \quad < \quad E\left|N\left(0,1\right)\right|$$

- That is, if and only if

$$g_1\left(\Delta,\xi\right) \quad = \quad \xi\varphi\left(\Delta/\xi\right) + \left|\Delta\right|\left[\Phi\left(\left|\Delta\right|/\xi\right) - 1/2\right] - \varphi\left(0\right) < 0$$

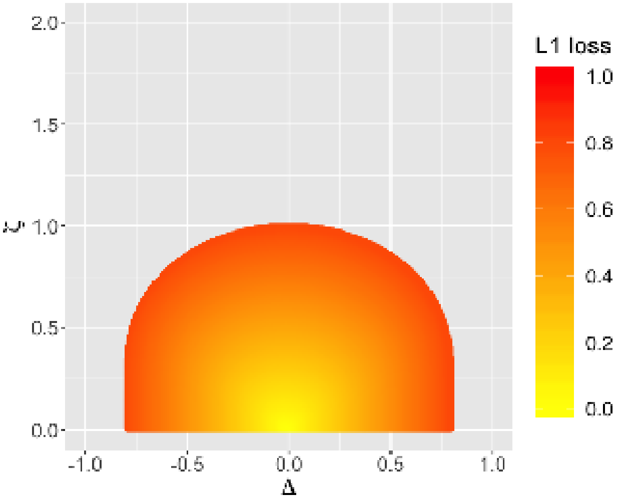
# L1 Loss Function



| $\xi$ | $X_2$ is Preferred if |
|---|---|
| 0.15 | $\lvert\Delta\rvert < 0.798$ |
| 0.25 | $\lvert\Delta\rvert < 0.798$ |
| 0.35 | $\lvert\Delta\rvert < 0.795$ |
| 0.45 | $\lvert\Delta\rvert < 0.783$ |
| 0.55 | $\lvert\Delta\rvert < 0.755$ |

| $\xi$ | $X_2$ is Preferred if |
|---|---|
| 0.65 | $\lvert\Delta\rvert < 0.706$ |
| 0.75 | $\lvert\Delta\rvert < 0.630$ |
| 0.85 | $\lvert\Delta\rvert < 0.513$ |
| 0.95 | $\lvert\Delta\rvert < 0.310$ |
| 0.99 | $\lvert\Delta\rvert < 0.141$ |

# L1 Loss Function



# Comparison of Linear Regression Estimates

Consider the linear regression model

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \quad i = 1, ..., n$$

- The observations $(y_i, x_{i1}, x_{i2})$, $i = 1, ..., n$ are independent
- $\mathbf{x}_i = (x_{i1}, x_{i2})'$ and $\varepsilon_i$ are independent
- $\varepsilon_i \sim N(0, \sigma^2)$ (take $\sigma = 1$ for simplicity)

Some notation:

$$X = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix} = (\mathbf{X}_1, \mathbf{X}_2), \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

# Comparison of Linear Regression Estimates

More notation:

$$\begin{aligned}
\langle \mathbf{X}_1, \mathbf{X}_1 \rangle &= \sum x_{i1}^2, \quad \langle \mathbf{X}_1, \mathbf{y} \rangle = \sum x_{i1} y_i \\
\langle \mathbf{X}_2, \mathbf{X}_2 \rangle &= \sum x_{i2}^2, \quad \langle \mathbf{X}_2, \mathbf{y} \rangle = \sum x_{i2} y_i \\
\langle \mathbf{X}_1, \mathbf{X}_2 \rangle &= \sum x_{i1} x_{i2}
\end{aligned}$$

# Comparison of Linear Regression Estimates

- Analysis will be conditional on the explanatory variables.
- To simplify the notations (and the analysis) we will assume

$$\langle \mathbf{X}_1, \mathbf{X}_1 \rangle = \langle \mathbf{X}_2, \mathbf{X}_2 \rangle = 1$$

- We also set

$$\langle \mathbf{X}_1, \mathbf{X}_2 \rangle = r$$

Clearly, $|r| \leq 1$.

# Comparison of Linear Regression Estimates

We have:

$$B = X'X = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}, \quad X'\,\mathbf{y} = \begin{pmatrix} \langle X_1, y \rangle \\ \langle X_2, y \rangle \end{pmatrix}$$

$$B^{-1} = \frac{1}{1-r^2} \begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix}$$

# Comparison of Linear Regression Estimates

We wish to compare two estimators for $\boldsymbol{\beta} = (\beta_1, \beta_2)'$:

1. The joint LS estimator

$$\widehat{\boldsymbol{\beta}} = \left(X'X\right)^{-1} X' \, \mathbf{y} = B^{-1} X' \mathbf{y},$$

2. The separate LS estimator

$$\widehat{\boldsymbol{\alpha}} = \begin{pmatrix} \langle \mathbf{X}_1, \mathbf{y} \rangle \,/\, \langle \mathbf{X}_1, \mathbf{X}_1 \rangle \\ \langle \mathbf{X}_2, \mathbf{y} \rangle \,/\, \langle \mathbf{X}_2, \mathbf{X}_2 \rangle \end{pmatrix} = \begin{pmatrix} \langle \mathbf{X}_1, \mathbf{y} \rangle \\ \langle \mathbf{X}_2, \mathbf{y} \rangle \end{pmatrix} = X' \mathbf{y}$$

# Comparison of Linear Regression Estimates

For the joint LS estimator we have:

$$E\left(\widehat{\beta}|X\right) = B^{-1}X'X\beta = \beta$$

$$Cov\left(\widehat{\beta}|X\right) = B^{-1}X'XB^{-1} = B^{-1} = \frac{1}{1-r^2}\begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix}$$

# Comparison of Linear Regression Estimates

For the joint LS estimator we have:

$$E\left(\widehat{\boldsymbol{\alpha}}|X\right) = X'X\,\boldsymbol{\beta} = B\,\boldsymbol{\beta} = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \beta_1 + r\beta_2 \\ \beta_2 + r\beta_1 \end{pmatrix}$$

and

$$Cov\left(\widehat{\boldsymbol{\alpha}}|X\right) = X'X = B = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$$

**Example 1:** Compare $\widehat{\beta}_1$ and $\widehat{\alpha}_1 u\sin g$ quadratic loss. Assume that $r = 0.5$.
We have

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{1}{1-r^2}\right)$$

$$\mu \to \beta_1 \qquad \sigma^2 \to \frac{1}{1-r^2}$$

We also have:

$$\widehat{\alpha}_1 \sim N(\beta_1 + r\beta_2, 1)$$

$$\gamma \rightarrow \beta_1 + r\beta_2 \qquad \tau^2 \rightarrow 1$$

Therefore

$$\Delta = \frac{\gamma - \mu}{\sigma} = \beta_2 r\sqrt{1 - r^2} \quad \text{and} \quad \xi = \frac{\tau}{\sigma} = \sqrt{1 - r^2}$$

# Comparison of Linear Regression Estimates

To fix ideas suppose that we wish to use the Quadratic Loss and that

$$r = 0.5$$

Then

$$\xi = \sqrt{1 - r^2} = \sqrt{0.75}$$

The biased estimator $\widehat{\alpha}_1$ is preferred if and only if

$$\begin{aligned} |\Delta| &< \sqrt{1 - \xi^2} \\ \left| \frac{\beta_2}{2\sqrt{0.75}} \right| &< \sqrt{1 - 0.75} = \frac{1}{2} \end{aligned}$$

That is

$$|\beta_2| < \sqrt{0.75} = 0.86603$$

# Comparison of Linear Regression Estimates

Comparre $x_1\widehat{\beta}_1 + x_2\widehat{\beta}_2$ and $x_1\widehat{\alpha}_1 + x_2\widehat{\alpha}_2$. (a) Assume that $r = 0.5$ and take $x_1 = x_2 = 1$. (b) Assume that $\beta_1 = \beta_2 = 1$ and take $x_1 = x_2 = 1$
(a) We have

$$
\begin{aligned}
E\left(x_1\widehat{\beta}_1 + x_2\widehat{\beta}_2\right) &= x_1\beta_1 + x_2\beta_2 \\
Var\left(x_1\widehat{\beta}_1 + x_2\widehat{\beta}_2\right) &= x_1^2 Var\left(\widehat{\beta}_1\right) + x_2^2 Var\left(\widehat{\beta}_2\right) + 2Cov\left(x_1\widehat{\beta}_1, x_2\widehat{\beta}_2\right) \\
&= \left(x_1^2 + x_2^2 + 2x_1x_2r\right) / \left(1 - r^2\right)
\end{aligned}
$$

$$
\begin{aligned}
\mu &\to x_1\beta_1 + x_2\beta_2 \\
\sigma^2 &\to \left(x_1^2 + x_2^2 + 2x_1x_2r\right) / \left(1 - r^2\right)
\end{aligned}
$$

# Comparison of Linear Regression Estimates

On the other hand:

$$\begin{aligned} E\left(x_1\widehat{\alpha}_1 + x_2\widehat{\alpha}_2\right) &= x_1\beta_1 + x_2\beta_2 + r\left(x_1\beta_2 + x_2\beta_1\right) \\ Var\left(x_1\widehat{\alpha}_1 + x_2\widehat{\alpha}_2\right) &= x_1^2 Var\left(\widehat{\alpha}_1\right) + x_2^2 Var\left(\widehat{\alpha}_2\right) + 2Cov\left(x_1\widehat{\alpha}_1, x_2\widehat{\alpha}_2\right) \\ &= x_1^2 + x_2^2 + 2x_1x_2 r \end{aligned}$$

$$\begin{aligned} \gamma &\to x_1\beta_1 + x_2\beta_2 + r\left(x_1\beta_2 + x_2\beta_1\right) \\ \tau^2 &\to x_1^2 + x_2^2 + 2x_1x_2 r \end{aligned}$$

Therefore

$$\Delta = \frac{\gamma - \mu}{\sigma} = r\left(x_1\beta_2 + x_2\beta_1\right) \quad \text{and} \quad \xi = \frac{\tau}{\sigma} = \sqrt{1 - r^2}$$

# Comparison of Linear Regression Estimates

Since

$$\begin{aligned} r &= 0.5 \\ x_1 &= x_2 = 1 \end{aligned}$$

Hence

$$\begin{aligned} \mu &= \beta_1 + \beta_2 \\ \gamma &= \beta_1 + \beta_2 + \frac{\beta_1 + \beta_2}{2} = \frac{3}{2}(\beta_1 + \beta_2) \end{aligned}$$

Then

$$\delta = \frac{\beta_1 + \beta_2}{2}$$

# Comparison of Linear Regression Estimates

Moreover

$$\begin{aligned}\sigma^2 &= \frac{x_1^2 + x_2^2 + 2x_1x_2r}{1-r^2} = \frac{3}{3/4} = 4 \\ \tau^2 &= x_1^2 + x_2^2 + 2x_1x_2r = 3\end{aligned}$$

and

$$\Delta = \frac{|\beta_1 + \beta_2|}{4} = \quad \text{and } \xi = \sqrt{\frac{3}{4}}$$

Therefore, the biased estimator is preferred if and only if

$$\left|\frac{\beta_1 + \beta_2}{4}\right| < \sqrt{1 - \frac{3}{4}} = \frac{1}{2}$$

That is

$$|\beta_1 + \beta_2| < 2$$