STATS 2107
Statistical Modelling and Inference II
Solutions
Workshop 2: Bias, MSE, and BLUE

Matt Ryan

Semester 2 2022

Contents

Bias and MSE of Simple Linear Regression Estimates

The model

For data $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$, $x_i, Y_i \in \mathbb{R}$, consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \,,$$

where $\varepsilon_i \sim N(0, \sigma^2)$

The model estimates

Recall that the estimates for β_0 and β_1 are given by

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

where

$$S_{XY} = \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})$$
$$S_{XX} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Question

- What is the bias and MSE of $\hat{\beta}_1$?

First note that $\hat{\beta}_1$ is a *linear estimator* of β_1.

$\hat{\beta}_1$ is linear

You can write:

$$\hat{\beta}_1 = \sum_{i=1}^{n} a_i Y_i \,,$$

where $a_i = \dfrac{(x_i - \bar{x})}{S_{XX}}$.

PROOF:

Expected value and bias of $\hat{\beta}_1$

- $\mathrm{E}\left[\hat{\beta}_1\right] = \beta_1$
- Hence $b_{\hat{\beta}_1}(\beta_1) = 0$

i.e. $\hat{\beta}_1$ is an unbiased estimator of β_1.

PROOF:

The MSE of $\hat{\beta}_1$

Recall that:

$$\mathrm{MSE}_{\hat{\beta}_1}(\beta_1) = \mathrm{Var}(\hat{\beta}_1) + b_{\hat{\beta}_1}(\beta_1)^2 = \mathrm{Var}(\hat{\beta}_1)$$

so

$$\mathrm{MSE}_{\hat{\beta}_1}(\beta_1) = \frac{\sigma^2}{S_{XX}}$$

Your turn

What to do

1. Show that $\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{XX}}$.

Solutions:

$$\begin{aligned}
\text{Var}(\hat{\beta}_1) &= \text{Var}\left(\sum_{i=1}^{n} a_i Y_i\right) \\
&= \sum_{i=1}^{n} a_i^2 \text{Var}(Y_i), \qquad \text{(independence)} \\
&= \sigma^2 \sum_{i=1}^{n} a_i^2 \,.
\end{aligned}$$

Now look at $\sum_{i=1}^{n} a_i^2$. Subbing in a_i we get

$$\begin{aligned}
\sum_{i=1}^{n} a_i^2 &= \sum_{i=1}^{n} \left(\frac{(x_i - \bar{x})}{S_{XX}}\right)^2 \\
&= \frac{1}{S_{XX}^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= \frac{S_{XX}}{S_{XX}^2} \\
&= \frac{1}{S_{XX}} \,.
\end{aligned}$$

Hence, $\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{XX}}$.

---

2. Show that $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$ is a linear estimator, that is, You can write $\hat{\beta}_0 = \sum_{i=1}^{n} b_i Y_i$ for some constants b_i.

Solutions:
Similar to $\hat{\beta}_1$ we have:

$$\begin{aligned}
\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{x} \\
&= \frac{1}{n} \sum_{i=1} Y_i - \sum_{i=1} a_i \bar{x} Y_i \\
&= \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{XX}}\right) Y_i \\
&= \sum_{i=1}^{n} b_i Y_i \,.
\end{aligned}$$

---

3. Derive the bias and MSE of $\hat{\beta}_0$.

Solutions:
Similar to $\hat{\beta}_1$ we have:

$$\begin{aligned}
\mathrm{E}\left[\hat{\beta}_0\right] &= \mathrm{E}\left[\sum_{i=1}^n b_i Y_i\right] \\
&= \sum_{i=1}^n b_i \mathrm{E}\left[Y_i\right] \\
&= \sum_{i=1}^n b_i \left(\beta_0 + \beta_1 x_i\right) \\
&= \sum_{i=1}^n \left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{XX}}\right)\left(\beta_0 + \beta_1 x_i\right) \\
&= \frac{1}{n}\sum_{i=1}^n \beta_0 - \beta_0 \sum_{i=1}^n \frac{(x_i - \bar{x})\bar{x}}{S_{XX}} + \frac{1}{n}\sum_{i=1}^n \beta_1 x_i - \beta_1 \bar{x} \sum_{i=1}^n \frac{(x_i - \bar{x})x_i}{S_{XX}} \\
&= \beta_0 - \beta_0 \frac{\bar{x}}{S_{XX}} \sum_{i=1}^n (x_i - \bar{x}) + \beta_1 \bar{x} - \beta_1 \frac{\bar{x}}{S_{XX}} \sum_{i=1}^n \frac{(x_i - \bar{x})x_i}{S_{XX}} \\
&= \beta_0 - 0 + \beta_1 \bar{x} - \beta_1 \bar{x} \\
&= \beta_0 \,.
\end{aligned}$$

Hence $\hat{\beta}_0$ is unbiased for β_0. Thus we have

$$\begin{aligned}
\mathrm{MSE}_{\hat{\beta}_0}(\beta_0) &= \mathrm{Var}\left(\hat{\beta}_0\right) \\
&= \mathrm{Var}\left(\sum_{i=1}^n b_i Y_i\right) \\
&= \sum_{i=1}^n b_i^2 \mathrm{Var}\left(Y_i\right), \qquad \text{(independence)} \\
&= \sigma^2 \sum_{i=1}^n \left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{XX}}\right)^2 \\
&= \sigma^2 \left(\sum_{i=1}^n \frac{1}{n^2} - 2\frac{\bar{x}}{nS_{XX}} \sum_{i=1}^n (x_i - \bar{x}) + \frac{\bar{x}^2}{S_{XX}^2} \sum_{i=1}^n (x_i - \bar{x})^2\right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right) .
\end{aligned}$$

BLUE

A Theorem

Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \,,$$

where $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. Then $\hat{\beta}_1$ is the BLUE for β_1.

What does this mean?

Recall what BLUE stands for:

- **B**est
- **L**inear
- **U**nbiased
- **E**stimator

What does this mean?

- We have already shown that $\hat{\beta}_1$ is a linear, unbiased estimator of β_1.
- By "Best", we mean that for ANY other linear unbiased estimator $\tilde{\beta}_1$ of β_1, we must have

$$\mathrm{Var}\left(\tilde{\beta}_1\right) \geq \mathrm{Var}\left(\hat{\beta}_1\right)$$

How do we show this:

We break the proof into 3 parts:

1. Use the fact that $\tilde{\beta}_1$ is linear and unbiased to derive some properties.
2. Add 0.
3. Show the cross term (covariance) is 0.

$\tilde{\beta}_1$ is unbiased

Let $\tilde{\beta}_1 = \sum_{i=1}^{n} c_i Y_i$. Then:

$$\sum_{i=1}^{n} c_i = 0\,,$$
$$\sum_{i=1}^{n} c_i x_i = 1\,.$$

PROOF:

Add 0

Let's look at the variance of $\tilde{\beta}_1$:

$$\begin{aligned}\mathrm{Var}\left(\tilde{\beta}_1\right) &= \mathrm{Var}\left(\tilde{\beta}_1 - \hat{\beta}_1 + \hat{\beta}_1\right) \\ &= \mathrm{Var}\left(\hat{\beta}_1\right) + \mathrm{Var}\left(\tilde{\beta}_1 - \hat{\beta}_1\right) + 2\mathrm{cov}\left(\tilde{\beta}_1 - \hat{\beta}_1, \hat{\beta}_1\right)\,.\end{aligned}$$

The cross term is 0

We can show that

$$\mathrm{cov}\left(\tilde{\beta}_1 - \hat{\beta}_1, \hat{\beta}_1\right) = 0\,.$$

PROOF:

Putting it all together

Using these results, we have that

$$\mathrm{Var}\left(\tilde{\beta}_1\right) = \mathrm{Var}\left(\hat{\beta}_1\right) + \mathrm{Var}\left(\tilde{\beta}_1 - \hat{\beta}_1\right) \geq \mathrm{Var}\left(\hat{\beta}_1\right) .$$

Your turn

What to do

1. Show that $\hat{\beta}_0$ is the BLUE for β_0.

Solutions:
Using very similar methods:

Let $\tilde{\beta}_0 = \sum_{i=1}^{n} d_i Y_i$ be a linear, unbiased estimator of β_0. What does this tell us about the d_i?

$$\begin{aligned}
\beta_0 &= \mathrm{E}\left[\tilde{\beta}_0\right] \\
&= \mathrm{E}\left[\sum_{i=1}^{n} d_i Y_i\right] \\
&= \sum_{i=1}^{n} d_i \mathrm{E}\left[Y_i\right] \\
&= \sum_{i=1}^{n} d_i \left(\beta_0 + \beta_1 x_i\right) \\
&= \beta_0 \sum_{i=1}^{n} d_i + \beta_1 \sum_{i=1}^{n} d_i x_i \,.
\end{aligned}$$

Equating coefficients gives

$$\begin{aligned}
\sum_{i=1}^{n} d_i &= 1\,, \\
\sum_{i=1}^{n} d_i x_i &= 0\,.
\end{aligned}$$

These will be useful later. Now consider the variance of $\tilde{\beta}_0$.

$$\begin{aligned}
\mathrm{Var}\left(\tilde{\beta}_0\right) &= \mathrm{Var}\left(\tilde{\beta}_0 - \hat{\beta}_0 + \hat{\beta}_0\right) \\
&= \mathrm{Var}\left(\hat{\beta}_0\right) + \mathrm{Var}\left(\tilde{\beta}_0 - \hat{\beta}_0\right) + 2\mathrm{cov}\left(\tilde{\beta}_0 - \hat{\beta}_0, \hat{\beta}_0\right) .
\end{aligned}$$

Let's look at the covariance:

$$
\begin{aligned}
\operatorname{cov}\left(\tilde{\beta}_0 - \hat{\beta}_0, \hat{\beta}_0\right) &= \operatorname{cov}\left(\sum_{i=1}^{n}(d_i - b_i)Y_i, \sum_{j=1}^{n} b_j Y_j\right) \\
&= \sum_{ij}(d_i - b_i)b_j \operatorname{cov}\left(Y_i, Y_j\right) \\
&= \sum_{i=1}^{n}(d_i - b_i)b_i\sigma^2\,, \qquad \text{(independence)} \\
&= \sigma^2\left(\sum_{i=1}^{n} d_i b_i - \sum_{i=1}^{n} b_i^2\right)\,.
\end{aligned}
$$

Now, $\sum_{i=1}^{n} b_i^2 = \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)$ which you can easily check. Let's look at

$$
\begin{aligned}
\sum_{i=1}^{n} d_i b_i &= \sum_{i=1}^{n} d_i\left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{XX}}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} d_i - \frac{\bar{x}}{S_{XX}}\sum_{i=1}^{n} d_i x_i + \frac{\bar{x}^2}{S_{XX}}\sum_{i=1}^{n} d_i \\
&= \frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\,.
\end{aligned}
$$

Thus

$$
\operatorname{cov}\left(\tilde{\beta}_0 - \hat{\beta}_0, \hat{\beta}_0\right) = 0\,.
$$

Returning to our variance, this gives that:

$$
\begin{aligned}
\operatorname{Var}\left(\tilde{\beta}_0\right) &= \operatorname{Var}\left(\hat{\beta}_0\right) + \operatorname{Var}\left(\tilde{\beta}_0 - \hat{\beta}_0\right) + 2\operatorname{cov}\left(\tilde{\beta}_0 - \hat{\beta}_0, \hat{\beta}_0\right) \\
&= \operatorname{Var}\left(\hat{\beta}_0\right) + \operatorname{Var}\left(\tilde{\beta}_0 - \hat{\beta}_0\right) \\
&\geq \operatorname{Var}\left(\hat{\beta}_0\right)\,,
\end{aligned}
$$

proving the result.