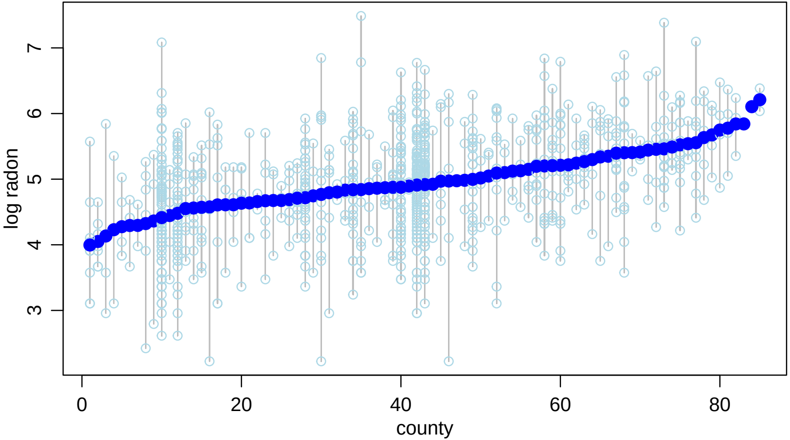
Estimation for group effects

Peter Hoff
Duke STA 610

MN radon data



Different amounts of information

```
y[g=="LACQUIPARLE"]
```

```
## [1] 6.036210 6.383751
```

```
y[g=="WASHINGTON"]
```

```
##  [1] 5.933906 5.653191 4.412045 5.484196 6.112774 5.139915 5.437089 5.484196
##  [9] 4.648416 4.269652 3.834061 4.497065 3.668259 3.834061 4.104487 3.473607
## [17] 4.162503 5.161298 4.162503 4.810531 3.473607 5.893950 5.280842 5.751848
## [25] 4.269652 5.499419 4.950219 5.387661 5.202746 4.537062 5.981707 4.497065
## [33] 4.366735 5.161298 4.923785 6.206521 4.682979 5.072896 4.950219 4.217459
## [41] 4.043070 4.217459 3.908367 5.499419 6.626603 5.404409
```

Linear shrinkage estimator: $\hat{\theta}_j = (1 - w_j)\bar{y}_j + w_j c$

- What should c be?
- What should w_j depend on?

Different amounts of information

```r
y[g=="LACQUIPARLE"]

## [1] 6.036210 6.383751

y[g=="WASHINGTON"]

##  [1] 5.933906 5.653191 4.412045 5.484196 6.112774 5.139915 5.437089 5.484196
##  [9] 4.648416 4.269652 3.834061 4.497065 3.668259 3.834061 4.104487 3.473607
## [17] 4.162503 5.161298 4.162503 4.810531 3.473607 5.893950 5.280842 5.751848
## [25] 4.269652 5.499419 4.950219 5.387661 5.202746 4.537062 5.981707 4.497065
## [33] 4.366735 5.161298 4.923785 6.206521 4.682979 5.072896 4.950219 4.217459
## [41] 4.043070 4.217459 3.908367 5.499419 6.626603 5.404409
```

Linear shrinkage estimator: $\hat{\theta}_j = (1 - w_j)\bar{y}_j + w_j c$

- What should c be?
- What should w_j depend on?

Different amounts of information

```
y[g=="LACQUIPARLE"]

## [1] 6.036210 6.383751

y[g=="WASHINGTON"]

##  [1] 5.933906 5.653191 4.412045 5.484196 6.112774 5.139915 5.437089 5.484196
##  [9] 4.648416 4.269652 3.834061 4.497065 3.668259 3.834061 4.104487 3.473607
## [17] 4.162503 5.161298 4.162503 4.810531 3.473607 5.893950 5.280842 5.751848
## [25] 4.269652 5.499419 4.950219 5.387661 5.202746 4.537062 5.981707 4.497065
## [33] 4.366735 5.161298 4.923785 6.206521 4.682979 5.072896 4.950219 4.217459
## [41] 4.043070 4.217459 3.908367 5.499419 6.626603 5.404409
```

Linear shrinkage estimator: $\hat{\theta}_j = (1 - w_j)\bar{y}_j + w_j c$

- What should c be?
- What should w_j depend on?

Mean squared error

- Let θ be the subpopulation mean of a generic group;
- let $\hat{\theta}$ be an estimator of θ (a function of the data).

The *mean squared error* (MSE) of $\hat{\theta}$ is

$$MSE[\hat{\theta}|\theta] = \mathrm{E}[(\hat{\theta} - \theta)^2|\theta]$$

Bias-variance decomposition: Let $m(\theta) = \mathrm{E}[\hat{\theta}|\theta]$.

$$\begin{aligned} MSE[\hat{\theta}|\theta] &= \mathrm{E}[(\hat{\theta} - m + m - \theta)^2|\theta] \\ &= \mathrm{E}[(\hat{\theta} - m)^2|\theta] + 2\mathrm{E}[(\hat{\theta} - m)(m - \theta)|\theta] + \mathrm{E}[(m - \theta)^2|\theta] \\ &= \mathrm{E}[(\hat{\theta} - m)^2|\theta] + (m - \theta)^2 \\ &= \mathrm{Var}[\hat{\theta}|\theta] + \mathrm{Bias}^2[\hat{\theta}|\theta] \end{aligned}$$

Mean squared error

- Let θ be the subpopulation mean of a generic group;
- let $\hat{\theta}$ be an estimator of θ (a function of the data).

The *mean squared error* (MSE) of $\hat{\theta}$ is

$$MSE[\hat{\theta}|\theta] = \mathrm{E}[(\hat{\theta} - \theta)^2|\theta]$$

Bias-variance decomposition: Let $m(\theta) = \mathrm{E}[\hat{\theta}|\theta]$.

$$\begin{aligned} MSE[\hat{\theta}|\theta] &= \mathrm{E}[(\hat{\theta} - m + m - \theta)^2|\theta] \\ &= \mathrm{E}[(\hat{\theta} - m)^2|\theta] + 2\mathrm{E}[(\hat{\theta} - m)(m - \theta)|\theta] + \mathrm{E}[(m - \theta)^2|\theta] \\ &= \mathrm{E}[(\hat{\theta} - m)^2|\theta] + (m - \theta)^2 \\ &= \mathrm{Var}[\hat{\theta}|\theta] + \mathrm{Bias}^2[\hat{\theta}|\theta] \end{aligned}$$

Mean squared error

- Let θ be the subpopulation mean of a generic group;
- let $\hat{\theta}$ be an estimator of θ (a function of the data).

The *mean squared error* (MSE) of $\hat{\theta}$ is

$$MSE[\hat{\theta}|\theta] = \mathrm{E}[(\hat{\theta} - \theta)^2|\theta]$$

Bias-variance decomposition: Let $m(\theta) = \mathrm{E}[\hat{\theta}|\theta]$.

$$\begin{aligned}
MSE[\hat{\theta}|\theta] &= \mathrm{E}[(\hat{\theta} - m + m - \theta)^2|\theta] \\
&= \mathrm{E}[(\hat{\theta} - m)^2|\theta] + 2\mathrm{E}[(\hat{\theta} - m)(m - \theta)|\theta] + \mathrm{E}[(m - \theta)^2|\theta] \\
&= \mathrm{E}[(\hat{\theta} - m)^2|\theta] + (m - \theta)^2 \\
&= \mathrm{Var}[\hat{\theta}|\theta] + \mathrm{Bias}^2[\hat{\theta}|\theta]
\end{aligned}$$

Mean squared error

- Let θ be the subpopulation mean of a generic group;
- let $\hat{\theta}$ be an estimator of θ (a function of the data).

The *mean squared error* (MSE) of $\hat{\theta}$ is

$$MSE[\hat{\theta}|\theta] = \mathrm{E}[(\hat{\theta} - \theta)^2|\theta]$$

Bias-variance decomposition: Let $m(\theta) = \mathrm{E}[\hat{\theta}|\theta]$.

$$\begin{aligned} MSE[\hat{\theta}|\theta] &= \mathrm{E}[(\hat{\theta} - m + m - \theta)^2|\theta] \\ &= \mathrm{E}[(\hat{\theta} - m)^2|\theta] + 2\mathrm{E}[(\hat{\theta} - m)(m - \theta)|\theta] + \mathrm{E}[(m - \theta)^2|\theta] \\ &= \mathrm{E}[(\hat{\theta} - m)^2|\theta] + (m - \theta)^2 \\ &= \mathrm{Var}[\hat{\theta}|\theta] + \mathrm{Bias}^2[\hat{\theta}|\theta] \end{aligned}$$

Bias-variance tradeoff

In general,

$$MSE(\hat{\theta}|\theta) = \mathrm{Var}[\hat{\theta}|\theta] + \mathrm{bias}(\hat{\theta}|\theta)^2$$

How well an estimator $\hat{\theta}$ does at estimating θ depends on *variance* and *bias*.

In general,

- estimators with low bias have have high variance;
- estimators with low variance have high bias.

Minimizing MSE requires balancing bias and variance.

Bias-variance tradeoff

In general,

$$MSE(\hat{\theta}|\theta) = \mathrm{Var}[\hat{\theta}|\theta] + \mathrm{bias}(\hat{\theta}|\theta)^2$$

How well an estimator $\hat{\theta}$ does at estimating θ depends on *variance* and *bias*.

In general,

- estimators with low bias have have high variance;
- estimators with low variance have high bias.

Minimizing MSE requires balancing bias and variance.

Bias-variance tradeoff

In general,

$$MSE(\hat{\theta}|\theta) = \mathrm{Var}[\hat{\theta}|\theta] + \mathrm{bias}(\hat{\theta}|\theta)^2$$

How well an estimator $\hat{\theta}$ does at estimating θ depends on *variance* and *bias*.

In general,

- estimators with low bias have have high variance;
- estimators with low variance have high bias.

Minimizing MSE requires balancing bias and variance.

Bias-variance tradeoff

In general,

$$MSE(\hat{\theta}|\theta) = \mathrm{Var}[\hat{\theta}|\theta] + \mathrm{bias}(\hat{\theta}|\theta)^2$$

How well an estimator $\hat{\theta}$ does at estimating θ depends on *variance* and *bias*.

In general,

- estimators with low bias have have high variance;
- estimators with low variance have high bias.

Minimizing MSE requires balancing bias and variance.

Bias-variance tradeoff

In general,

$$MSE(\hat{\theta}|\theta) = \mathrm{Var}[\hat{\theta}|\theta] + \mathrm{bias}(\hat{\theta}|\theta)^2$$

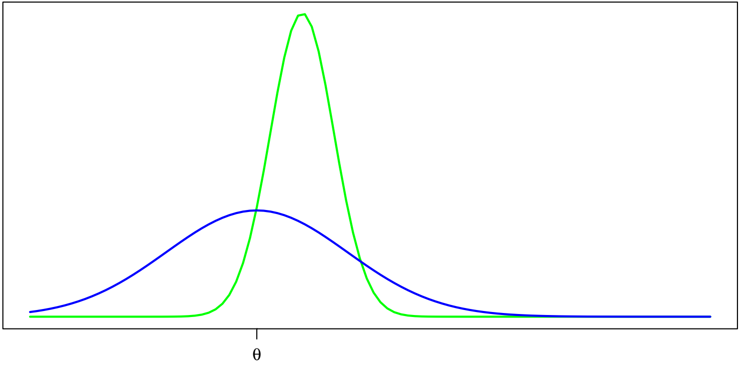
How well an estimator $\hat{\theta}$ does at estimating θ depends on *variance* and *bias*.

In general,

- estimators with low bias have have high variance;
- estimators with low variance have high bias.

Minimizing MSE requires balancing bias and variance.

Bias-variance tradeoff



Sample mean bias and variance

Let $y_1, \ldots, y_n$ be sample from a population with mean θ, variance σ^2.

Sample mean estimator: Let $\hat{\theta} = \bar{y}$

$$\begin{aligned} \mathrm{E}[\bar{y}|\theta] &= \theta \\ \mathrm{Bias}[\bar{y}|\theta] &= 0 \end{aligned}$$

$$\mathrm{Var}[\bar{y}|\theta] = \sigma^2/n$$

$$MSE[\bar{y}|\theta] = \mathrm{Var}[\bar{y}|\theta] = \sigma^2/n$$

Linear shrinkage bias and variance

Linear shrinkage estimator: $\hat{\theta} = (1 - w)\bar{y} + wc$ for some $w \in [0, 1]$.

- w is the amount of shrinkage;
- c is the shrinkage target.

$$\begin{aligned} \mathrm{E}[\hat{\theta}|\theta] &= (1 - w)\theta + wc = \theta + w(c - \theta) \\ \mathrm{Bias}[\bar{y}|\theta] &= w(c - \theta)^2 \geq 0 \end{aligned}$$

$$\mathrm{Var}[\bar{y}|\theta] = (1 - w)^2\sigma^2/n \leq \sigma^2/n$$

$$MSE[\bar{y}|\theta] = \mathrm{Var}[\bar{y}|\theta] = (1 - w)^2\sigma^2/n + w^2(c - \theta)^2$$

Composite MSE

Consider a LSE for $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)$ where $\hat{\theta}_j = (1 - w)\bar{y}_j + wc$

$$
\begin{aligned}
MSE[\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}] &= \mathsf{E}[||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2|\boldsymbol{\theta}] \\
&= \sum_j \mathsf{E}[(\hat{\theta}_j - \theta_j)^2|\boldsymbol{\theta}] \\
&= \frac{\sigma^2}{n} m(1 - w)^2 + w^2 \sum_j (c - \theta_j)^2
\end{aligned}
$$

What should the values of w and c be?

Composite MSE

Consider a LSE for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ where $\hat{\theta}_j = (1 - w)\bar{y}_j + wc$

$$\begin{aligned} MSE[\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}] &= \mathrm{E}[||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2|\boldsymbol{\theta}] \\ &= \sum_j \mathrm{E}[(\hat{\theta}_j - \theta_j)^2|\boldsymbol{\theta}] \\ &= \frac{\sigma^2}{n} m(1-w)^2 + w^2 \sum_j (c - \theta_j)^2 \end{aligned}$$

What should the values of w and c be?

Oracle estimator

Using calculus (homework!) you can show that MSE is optimized by

- $c = \bar{\theta} = \sum_j \theta_j / m$;
- $w = \frac{1/\tau^2}{n/\sigma^2 + 1/\tau^2}$, where
- $\tau^2 = \sum_j (\theta_j - \bar{\theta})^2 / m$.

resulting in the *oracle estimator*

$$\hat{\theta}_j = \frac{n/\sigma^2}{n/\sigma^2 + 1/\tau^2} \bar{y}_j + \frac{1/\tau^2}{n/\sigma^2 + 1/\tau^2} \bar{\theta}.$$

Oracle estimator

Using calculus (homework!) you can show that MSE is optimized by

- $c = \bar{\theta} = \sum_j \theta_j / m$;
- $w = \frac{1/\tau^2}{n/\sigma^2 + 1/\tau^2}$, where
- $\tau^2 = \sum_j (\theta_j - \bar{\theta})^2 / m$.

resulting in the *oracle estimator*

$$\hat{\theta}_j = \frac{n/\sigma^2}{n/\sigma^2 + 1/\tau^2} \bar{y}_j + \frac{1/\tau^2}{n/\sigma^2 + 1/\tau^2} \bar{\theta}.$$