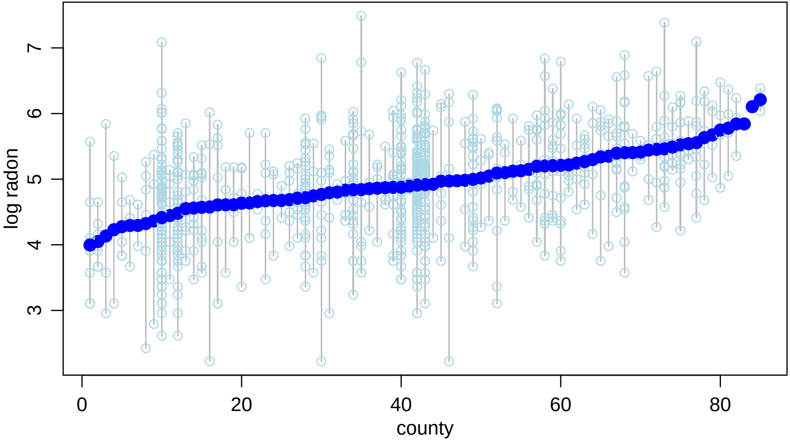
Estimation of group effects

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?

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,

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

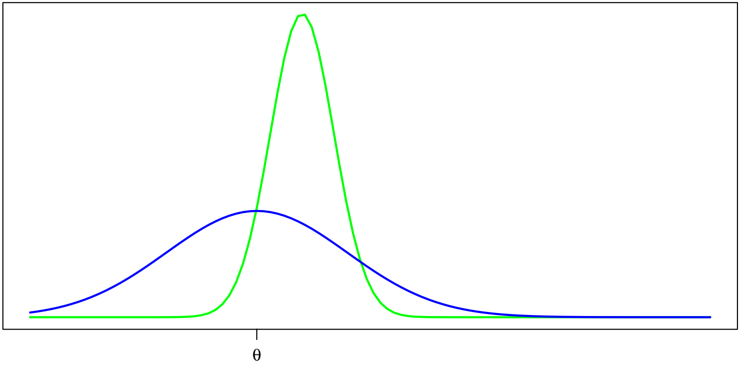
How well an estimator $\hat{\theta}$ does at estimating θ depends on *variance* and *bias*.

In general,

- estimators with low bias have have high variance ($\hat{\theta} = \bar{y}$ but small n);
- estimators with low variance have high bias ($\hat{\theta} = 5$).

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 $\hat{\theta} = \bar{y}$

$$\mathrm{E}[\bar{y}|\theta] = \theta$$
$$\mathrm{Bias}[\bar{y}|\theta] = 0$$

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

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

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 point.

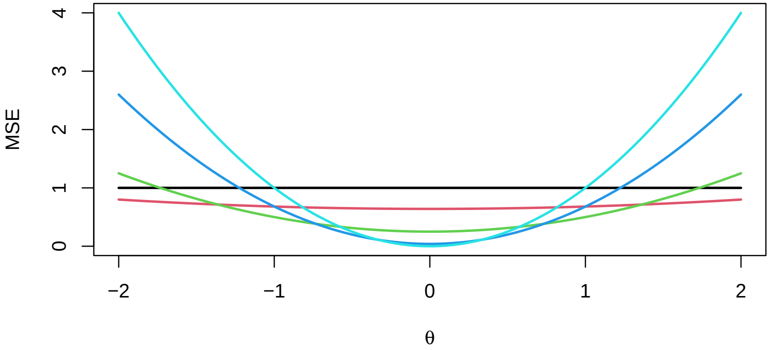
$$\begin{aligned}
\mathrm{E}[\hat{\theta}|\theta] &= (1-w)\theta + wc = \theta + w(c-\theta) \\
\mathrm{Bias}[\hat{\theta}|\theta]^2 &= w^2(c-\theta)^2 \geq 0
\end{aligned}$$

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

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

Mean squared error function

$$\sigma^2/n = 1 \quad c = 0$$



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 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
- $\tau^2 = \sum_j (\theta_j - \bar{\theta})^2 / m$.

This gives the *oracle estimator*

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

This can also be written

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

Composite risk comparison

- $\bar{\mathbf{y}} = (\bar{y}_1, \ldots, \bar{y}_m)$, the vector of sample means;
- $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)$, the vector of oracle estimates.

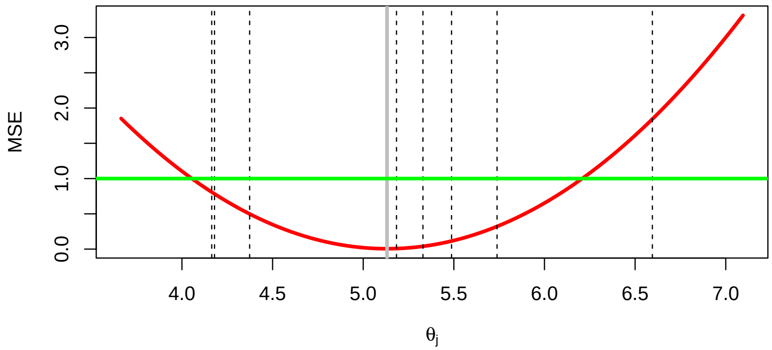
$$MSE[\bar{\mathbf{y}}|\boldsymbol{\theta}] = m\frac{\sigma^2}{n}$$

$$MSE[\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}] = m\frac{\sigma^2}{n} \times \left(\frac{\tau^2}{\sigma^2/n + \tau^2}\right) < MSE[\bar{\mathbf{y}}|\boldsymbol{\theta}].$$

The oracle estimator is better than $\bar{\mathbf{y}}$ *in terms of composite risk* .

Group-level risk of oracle estimator

$$MSE[\hat{\theta}_j|\boldsymbol{\theta}] = (1-w)^2\sigma^2/n + w^2(\theta_j - \bar{\theta})^2.$$



Summary

Composite risk

- $\bar{\mathbf{y}}$ is an unbiased estimator of $\boldsymbol{\theta}$;
- $\hat{\boldsymbol{\theta}}$ is a biased estimator of $\boldsymbol{\theta}$, but has lower variance than $\bar{\mathbf{y}}$.
- $MSE[\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}] \leq MSE[\bar{\mathbf{y}}|\boldsymbol{\theta}]$.

Group-level risk

- $\bar{y}_j$ is an unbiased estimator of θ_j for each $j = 1, \ldots, m$.
- $\hat{\theta}_j$ is a biased estimator of θ_j, but has lower variance than $\bar{y}_j$.
- $MSE[\hat{\theta}_j|\boldsymbol{\theta}] \gtreqless MSE[\bar{y}_j|\boldsymbol{\theta}]$ and you don't know which!

Practical considerations

Typically,

- μ, τ^2, σ^2 are unknown;
- sample sizes may vary across groups.

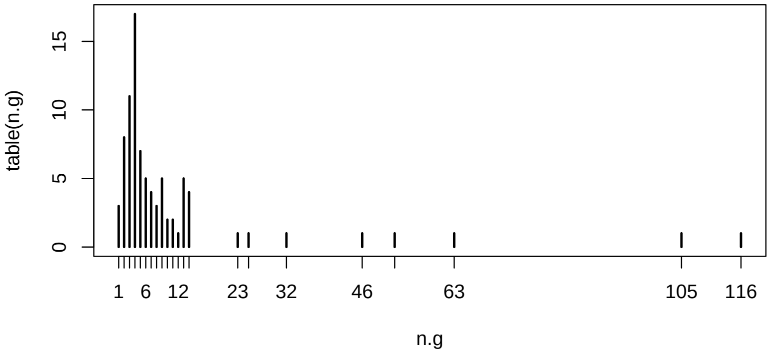
In practice, people use the following *adaptive shrinkage estimator*:

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

- $\hat{\mu}, \hat{\tau}^2, \hat{\sigma}^2$ are obtained from the data (e.g. ANOVA or `lme4`).
- If $n_j = n$, can obtain $\hat{\mu}, \hat{\tau}^2, \hat{\sigma}^2$ so that $\hat{\boldsymbol{\theta}}$ is guaranteed better than $\bar{\mathbf{y}}$ (Stein).
- Otherwise, for large m, $\hat{\theta}$ will be approximately optimal linear estimator (under composite risk).

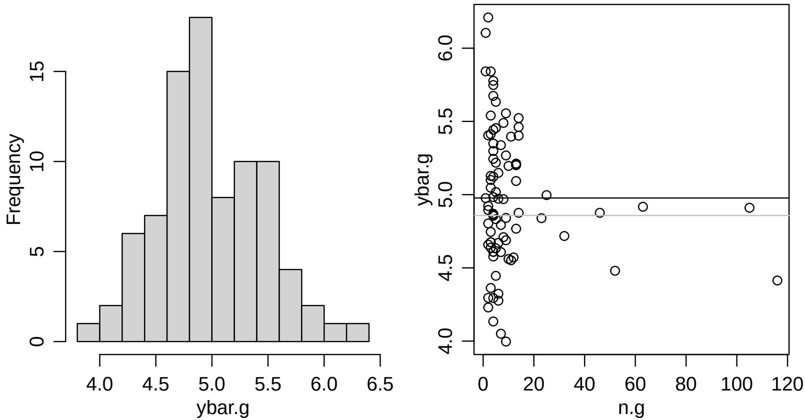
Radon example

```
n.g<-c(table(g) )

plot(table(n.g))
```



Radon example

```r
# county specific radon means
ybar.g<-c(tapply(y,g,"mean"))
```



MLEs

```
library(lme4)
fit.lme<-lmer(y~1+(1|g),REML=FALSE)
summary(fit.lme)

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: y ~ 1 + (1 | g)
##
##      AIC      BIC   logLik deviance df.resid
##   2164.1   2178.5  -1079.0   2158.1      916
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.6165 -0.6141  0.0292  0.6526  3.4932
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  g        (Intercept) 0.08804  0.2967
##  Residual             0.57154  0.7560
## Number of obs: 919, groups:  g, 85
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  4.94656    0.04664   106.1
```

Parameter estimates

```
VarCorr(fit.lme)

##  Groups   Name        Std.Dev.
##  g        (Intercept) 0.29672
##  Residual             0.75600

t2.mle<-as.numeric(VarCorr(fit.lme)$g)

t2.mle

## [1] 0.08804027

sigma(fit.lme)

## [1] 0.7559996

s2.mle<-sigma(fit.lme)^2

s2.mle

## [1] 0.5715354

fixef(fit.lme)

## (Intercept)
##    4.946557

mu.mle<-fixef(fit.lme)
```

Adaptive shrinkage estimates

Replace μ, σ^2, τ^2 with estimates:

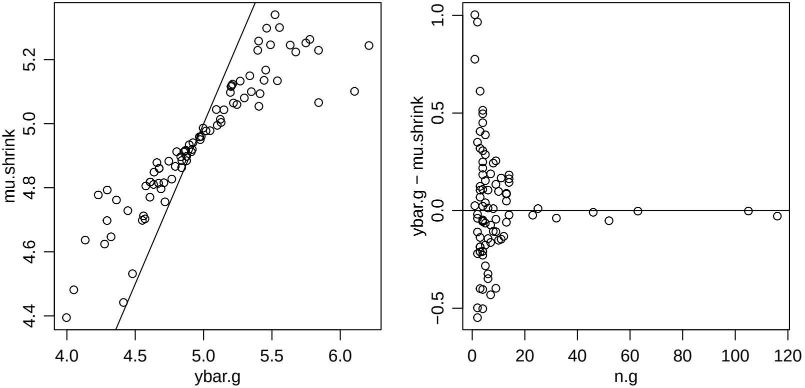
$$\hat{\mu}_j = w_j \bar{y}_j + (1 - w_j)\hat{\mu} \ , \ \text{where } w_j = \frac{n_j/\hat{\sigma}^2}{n_j/\hat{\sigma}^2 + 1/\hat{\tau}^2}.$$

```
w.shrink<- (n.g/s2.mle) /(n.g/s2.mle + 1/t2.mle)

mu.shrink<-w.shrink*ybar.g + (1-w.shrink)*mu.mle

mu.mle

## (Intercept)
##    4.946557

cbind(ybar.g, n.g, mu.shrink)[1:8,]

##             ybar.g n.g mu.shrink
## AITKIN    4.293832   4  4.697704
## ANOKA     4.479973  52  4.531757
## BECKER    4.675008   3  4.860730
## BELTRAMI  4.793035   7  4.866904
## BENTON    4.869503   4  4.917180
## BIGSTONE  5.128199   3  5.003968
## BLUEEARTH 5.522876  14  5.340299
## BROWN     5.244160   4  5.060018
```

Shrinkage

```
topten<-order(ybar.g,decreasing=TRUE)[1:10]
cbind(ybar.g, n.g, mu.shrink)[topten,]

##              ybar.g n.g mu.shrink
## LACQUIPARLE 6.209980   2  5.244122
## MURRAY      6.104550   1  5.101126
## WILKIN      5.841654   1  5.066035
## WATONWAN    5.841041   3  5.229271
## NICOLLET    5.777273   4  5.263269
## LINCOLN     5.748294   4  5.252221
## KANDIYOHI   5.674289   4  5.224006
## JACKSON     5.633758   5  5.245555
## FREEBORN    5.555495   9  5.300322
## NOBLES      5.540083   3  5.134149
```

Shrinkage


mu.shrink
ybar.g
4.4
4.6
4.8
5.0
5.2
4.0
4.5
5.0
5.5
6.0
ybar.g − mu.shrink
n.g
−0.5
0.0
0.5
1.0
0
20
40
60
80
100
120

Shrinkage estimates from lme4

```
mu.shrink[1:10]

##    AITKIN     ANOKA    BECKER  BELTRAMI    BENTON  BIGSTONE BLUEEARTH     BROWN
##  4.697704  4.531757  4.860730  4.866904  4.917180  5.003968  5.340299  5.060018
##   CARLTON    CARVER
##  4.712463  4.958725

a.shrink<-ranef(fit.lme)[[1]][,1]

mu.mle+a.shrink[1:10]

##  [1] 4.697704 4.531757 4.860730 4.866904 4.917180 5.003968 5.340299 5.060018
##  [9] 4.712463 4.958725
```

In `lme4`, `ranef(fit.lme)[[k]][,l]` refers to the

- `l`th random effect for the
- `k`th grouping variable.

Hierarchical normal model

$$\begin{aligned} y_{i,j} &= \mu + a_j + \epsilon_{i,j} \\ \{\epsilon_{1,1},\ldots,\epsilon_{n,1}\},\ldots,\{\epsilon_{1,m},\ldots,\epsilon_{n,m}\} &\sim \text{i.i.d. normal}(0,\sigma^2) \\ a_1,\ldots,a_m &\sim \text{i.i.d. normal}(0,\tau^2) \end{aligned}$$

Equivalently,

$$\begin{aligned} y_{i,j} &= \theta_j + \epsilon_{i,j} \\ \{\epsilon_{1,1},\ldots,\epsilon_{n,1}\},\ldots,\{\epsilon_{1,m},\ldots,\epsilon_{n,m}\} &\sim \text{i.i.d. normal}(0,\sigma^2) \\ \theta_1,\ldots,\theta_m &\sim \text{i.i.d. normal}(\mu,\tau^2) \end{aligned}$$

In this model, we think of

- the groups as being randomly selected from a larger set of possible groups,
- so the means are randomly selected from a set of possible means.
- This interpretation is *not appropriate* for the radon data!

Unbiased predictors

Suppose you will sample a random group with subgroup mean θ, so

$$\theta \sim N(\mu, \tau^2)$$
$$\bar{y}|\theta \sim N(\theta, \sigma^2/n).$$

How should you plan on estimating θ? Consider estimators $\tilde{\theta}$ that are unbiased "on average:"

$$\mathrm{E}[\tilde{\theta} - \theta] = \int_\theta \left(\int_y (\hat{\theta} - \theta) p(y|\theta)\, d\theta \right) \, p(\theta|\mu, \tau^2)\, d\theta$$

- Such estimators are sometimes called "unbiased predictors";
- They might not be unbiased for any particular value of θ!

Unbiased predictors

$$\bar{y} = \sum y_i/n$$
$$\hat{\theta} = \frac{\tau^2}{\sigma^2/n + \tau^2}\bar{y} + \frac{\sigma^2/n}{\sigma^2/n + \tau^2}\mu.$$

Exercises:

1. Show that $\bar{y}$ is an "unbiased predictor";
2. Show that $\hat{\theta}$ is an "unbiased predictor".
3. Identify some other "unbiased predictors".

Best unbiased prediction

Result 1: (Best unbiased predictor). Let $\tilde{\theta}$ be any unbiased predictor, meaning $\mathsf{E}[\tilde{\theta} - \theta] = 0$ where the expectation is averaging over *both* y and θ. Then

$$\mathsf{E}[(\hat{\theta} - \theta)^2] \leq \mathsf{E}[(\tilde{\theta} - \theta)^2]$$

where the expectation is over *both* y and θ.

Best linear unbiased predictor

A similar result holds even if the data are not normal. Suppose

- $\mathsf{E}[\bar{y}|\theta] = \theta$, $\mathsf{Var}[\bar{y}|\theta] = \sigma^2/n$.
- $\mathsf{E}[\theta] = \mu$, $\mathsf{Var}[\theta] = \tau^2$.

Result 2: (Best linear unbiased predictor). Let $\tilde{\theta}$ be any *linear* unbiased predictor, meaning

- $\tilde{\theta} = a\bar{y} + b$ for some fixed a and b;
- $\mathsf{E}[\tilde{\theta} - \theta] = 0$ where the expectation is averaging over *both* y and θ.

Then

$$\mathsf{E}[(\hat{\theta} - \theta)^2] \leq \mathsf{E}[(\tilde{\theta} - \theta)^2]$$

where the expectation is over *both* y and θ.

Practical considerations

As before,

- μ, τ^2, σ^2 are unknown;
- sample sizes may vary across groups.

In practice, people use the following *Empirical BLUP*:

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

where $\hat{\mu}, \hat{\tau}^2, \hat{\sigma}^2$ are estimated from the data (ANOVA or `lme4`)

This is the same estimator as the adaptive shrinkage estimator.

- variabiliy in "random" θ_j's $\approx$ heterogeneity in "fixed" θ_j's.
- integration over θ_j's to get MSE $\approx$ summing over θ_j's to get composite MSE.

BLUPs

The $\hat{\theta}_j$'s are sometimes called the *best unbiased linear predictors (BLUPs)* .

This is confusing, as we have discussed how these estimators are biased:

$$\begin{aligned} \mathrm{E}[\hat{\theta}_j|\theta_j] &= \mathrm{E}[w\bar{y}_j + (1-w)\mu|\theta_j] \\ &= w\theta_j + (1-w)\mu \neq \theta_j \end{aligned}$$

$\hat{\theta}_j$ is *conditionally* biased.

The "U" in BLUP refers to bias only in an unconditional sense:

$$\begin{aligned} \mathrm{E}[\hat{\theta}_j] &= \mathrm{E}[\mathrm{E}[\hat{\theta}_j|\theta_j]] \\ &= \mathrm{E}[w\theta_j + (1-w)\mu] \\ &= w\mu + (1-w)\mu = \mu. \end{aligned}$$

Since $\mathrm{E}[\hat{\theta}_j] = \mathrm{E}[\theta_j] = \mu$ *unconditionally*, people might say $\hat{\theta}_j$ is "unbiased."

Understanding conditional and unconditional expectation

school	A	B	C	D	E	F	G	H	I	J
mean	θ_A	θ_B	θ_C	θ_D	θ_E	θ_F	θ_G	θ_H	θ_I	θ_J

Let $\mu = (\theta_A + \cdots \theta_J)/10$.

Study design:

- sample m schools at random from the population of schools.
- sample n students at random from each of the m schools.

What is the expectation of θ_1, $\bar{y}_1$, $\hat{\theta}_1$?

Expectation of θ_1 : Since each school A through J has equal probability of being selected as unit 1:

$$\begin{aligned} \mathrm{E}[\theta_1] &= \theta_A \times \Pr(\text{unit 1 =A}) + \cdots + \theta_J \times \Pr(\text{unit 1=J}) \\ &= \theta_A \tfrac{1}{10} + \cdots + \theta_J \tfrac{1}{10} = \mu \end{aligned}$$

Understanding conditional expectation

$$\mathrm{E}[\bar{y}_1 - \theta_1 | \text{unit } 1 = \text{D}\ ] = \mathrm{E}[\bar{y}_D - \theta_D] = \theta_D - \theta_D = 0$$

$$\begin{aligned}\mathrm{E}[\hat{\theta}_1 - \theta_1 | \text{unit } 1 = \text{D}] &= \mathrm{E}[w\bar{y}_D + (1-w)\mu - \theta_D] \\ &= w\theta_D + (1-w)\mu - \theta_D = (1-w)(\mu - \theta_D) \neq 0\end{aligned}$$

Conditionally on unit 1=D,

- $\bar{y}_1 = \bar{y}_D$ is unbiased for θ_D,
- $\hat{\theta}_1 = \hat{\theta}_D$ is biased for θ_D.

In English, if your first sampled school is school D, then

- $\bar{y}_1 = \bar{y}_D$ and $\bar{y}_D$ is unbiased for θ_D
- $\hat{\theta}_1 = \hat{\theta}_D$ and $\hat{\theta}_D$ is biased for θ_D.

Understanding unconditional expectation

Before you sample the schools, unit 1 is equally likely to be school A, B, . . . , J.

$$\begin{aligned}
\mathrm{E}[\hat{\theta}_1 - \theta_1] &= \mathrm{E}[\hat{\theta}_A - \theta_A]\,\mathrm{Pr}(\text{unit }1{=}\mathrm{A}) + \cdots + \mathrm{E}[\hat{\theta}_J - \theta_J]\,\mathrm{Pr}(\text{unit }1{=}\mathrm{J}) \\
&= (1-w)(\mu - \theta_A) \times \tfrac{1}{10} + \cdots + (1-w)(\mu - \theta_J) \times \tfrac{1}{10} \\
&= (1-w)\mu - (1-w)(\theta_A + \cdots + \theta_J)\tfrac{1}{10} \\
&= (1-w)\mu - (1-w)\mu = 0.
\end{aligned}$$

This unconditional expectation, and the “U” in BLUP, refers to averaging across the possibilities for the samples:

- $\hat{\theta}_j$ will be a biased estimator of the mean of whatever unit is picked jth.
- on average across studies, $\hat{\theta}_1, \ldots, \hat{\theta}_m$ will be unbiased.

Summary

In many applications interest is more in the conditional expectations.

From this perspective, the shrinkage estimators $\hat{\theta}_1, \ldots, \hat{\theta}_m$

- are biased;
- have conditional MSE given by

$$w^2\sigma^2/n_j + (1-w)^2(\theta_j - \mu)^2,$$

- which is usually lower than the conditional MSE of $\bar{y}_j$.

Review of Bayesian inference

- Prior density: $p(\gamma)$
- Sampling density: $p(y_1, \ldots, y_n|\gamma)$

The prior density describes where you think γ is, before having seen the data.

The sampling density describes where you think the data will be, for each possible value of γ.

Bayes rule:

$$\begin{aligned} p(\gamma|y_1, \ldots, y_n) &= \frac{p(\gamma)p(y_1, \ldots, y_n|\gamma)}{\int p(\gamma')p(y_1, \ldots, y_n|\gamma')\, d\gamma'} \\ &\propto p(\gamma)p(y_1, \ldots, y_n|\gamma) \end{aligned}$$

The *posterior density* $p(\gamma|y_1, \ldots, y_n)$ describes where you think the γ is, after having seen the data.

Bayesian inference for a normal subpopulation

- Prior density: $\theta \sim N(\mu, \tau^2)$
- Sampling density: $y_1, \ldots, y_n | \theta \sim N(\theta, \sigma^2)$.

Bayes rule: $\theta | y_1, \ldots, y_n$ is normal, with

$$\mathrm{E}[\theta | y_1, \ldots, y_n] = \frac{\tau^2}{\sigma^2/n + \tau^2} \bar{y} + \frac{\sigma^2/n}{\sigma^2/n + \tau^2} \mu$$
$$\mathrm{Var}[\theta | y_1, \ldots, y_n] = 1/(1/\sigma^2/n + 1/\tau^2)$$

Bayes estimator: Let $\hat{\theta} = \mathrm{E}[\theta | y_1, \ldots, y_n]$. Then

$$\mathrm{E}[(\hat{\theta} - \theta)^2 | y_1, \ldots, y_n] \leq \mathrm{E}[(\tilde{\theta} - \theta)^2 | y_1, \ldots, y_n]$$

for any estimator $\tilde{\theta}$.

Bayes interpretation

A Bayesian interpretation of $\hat{\theta}$:

- θ_j is some fixed quantity for group j;
- $\theta_j \sim N(\mu, \tau^2)$ describes prior info about θ_j;
- $\theta_j \sim N(\hat{\theta}_j, 1/(n_j/\sigma^2 + 1/\tau^2))$ describes posterior info about θ_j;
- $\hat{\theta}_j$ is "where you think θ_j is".

Practical considerations:

- μ, τ^2, σ^2;
- estimate these parameters with a "fully Bayesian procedures", or
- use plug-in estimates (Empirical Bayes), obtained from data (ANOVA, `lme4`).

Again, the estimator is the same but the justification can be different.