Statistical analysis of JEM data

Esben Budtz-Jørgensen

Section of Biostatistics, University of Copenhagen

Statistical questions

Aim: relate exposure matrix data to health outcomes.

Main challenge: For each subject the exposure variable is calculated from an exposure matrix and may not be equal to the true exposure.

- How should the analysis of exposure effect on health be conducted?
- How should the exposure distribution be summarized in the JEM?

Variance components in exposures

Data are grouped.

Response model: $Y_{gi} = \alpha + \beta X_{gi} + \zeta_{gi}$

True exposure of subject i in group g: $X_{gi} = \mu_g + \gamma_{gi}$

$var(\mu_g) = \sigma_b$: measures variation between groups

$var(\gamma_{gi}) = \sigma_w$: measures variation with-in groups

Measured exposure of subject i in group g: $W_{gi} = X_{gi} + \epsilon_{gi}$

$var(\epsilon_{gi}) = \sigma_e$: measures day-to-day variation and device error

Simplification: analyze data on this form

Group				
	$X_{1,1}$	$X_{1,2}$	...	$X_{1,n}$
1	$W_{1,1}$	$W_{1,2}$	...	$W_{1,n}$
	$Y_{1,1}$	$Y_{1,2}$	...	$Y_{1,n}$
.				
.				
.				
	$X_{G,1}$	$X_{G,2}$	...	$X_{G,n}$
G	$W_{G,1}$	$W_{G,2}$	...	$W_{G,n}$
	$Y_{G,1}$	$Y_{G,2}$	...	$Y_{G,n}$

Individual data methods

1. Regress Y_{gi} on X_{gi}

2. Regress Y_{gi} on W_{gi}

Simulation study:

Simulate data. Estimate exposure effect. Compare to truth $\beta = 0.5$ (y=0.5 x +error)

method	mean	theoretical mean	se	relative se
2. Y_{gi} on W_{gi}	0.361	$\frac{\sigma_b+\sigma_w}{\sigma_b+\sigma_w+\sigma_e} \cdot \beta$	0.027	1.6
1. Y_{gi} on X_{gi}	0.502	β	0.017	1

2: Strong underestimation when day-to-day variation (σ_e) is high

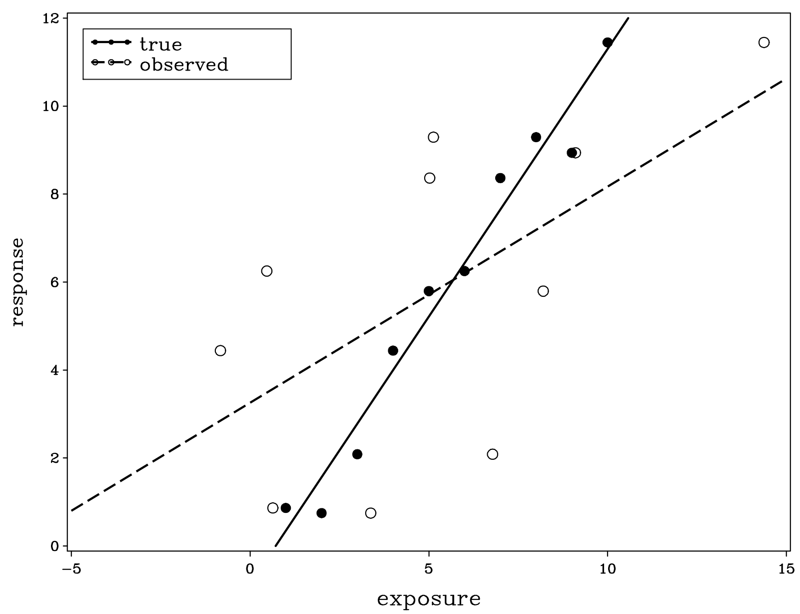
Classical error in linear regression

In approach 2, I replace X_{gi} by $W_{gi} = X_{gi} + \epsilon_{gi}$

X: true exposure, M: exposure marker, U: error

Classical additive error: $M = X + U$ with U independent of X

$Y = \alpha + \beta \cdot X + \epsilon$,   Naive Analysis: replace X by M $\rightarrow$ underestimation



Grouped data methods

3. Regress Y_{gi} on $\bar{X}_g = \frac{1}{n}\sum_{j=1}^{n} X_{gi}$

I create exposure measurement error: instead of X_{gi} I use $\bar{X}_g$

Does this error introduce bias?

method	mean	theoretical mean	se	relative se
3. Y_{gi} on $\bar{X}_g$	0.503	β	0.023	1.4
2. Y_{gi} on W_{gi}	0.361	$\frac{\sigma_b+\sigma_w}{\sigma_b+\sigma_w+\sigma_e} \cdot \beta$	0.027	1.6
1. Y_{gi} on X_{gi}	0.502	β	0.017	1

No bias? - different type of error

Berkson error

X: true exposure, M: exposure marker, U: error

Classical error

$$M = X + U$$

U independent of X. More variation in M than X.

Berkson error

$$X = M + U$$

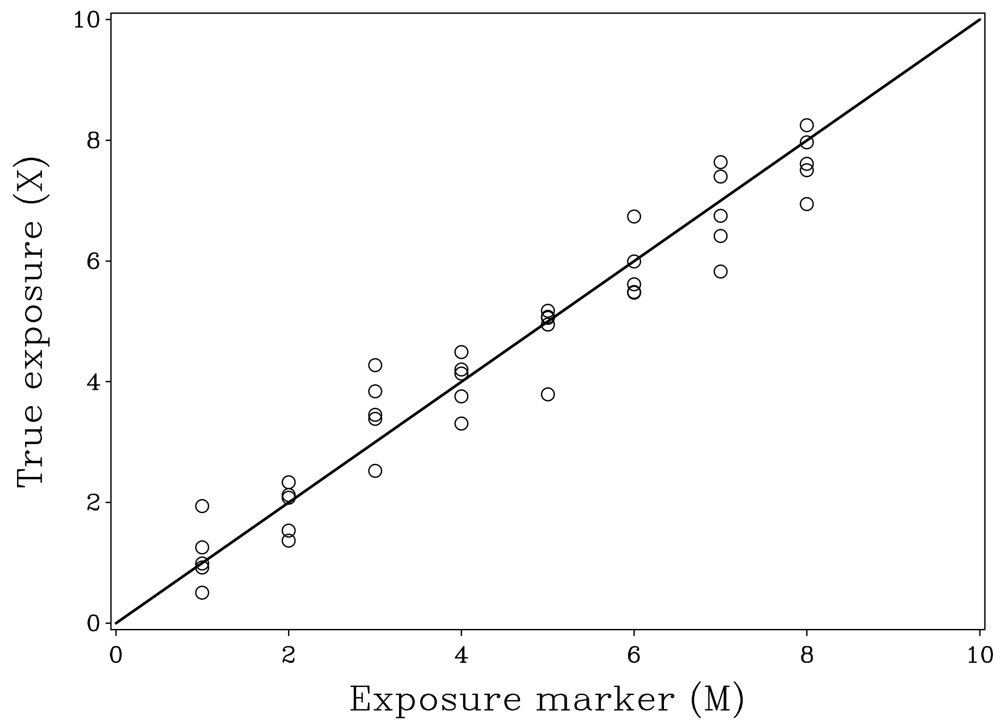
U independent of M. More variation in X than M.

Grouped data method

$$X_{gi} = \bar{X}_g + (X_{gi} - \bar{X}_g)$$

$\bar{X}_g$ and $(X_{gi} - \bar{X}_g)$ uncorrelated -Berkson error

Berkson error: $X = M + U$

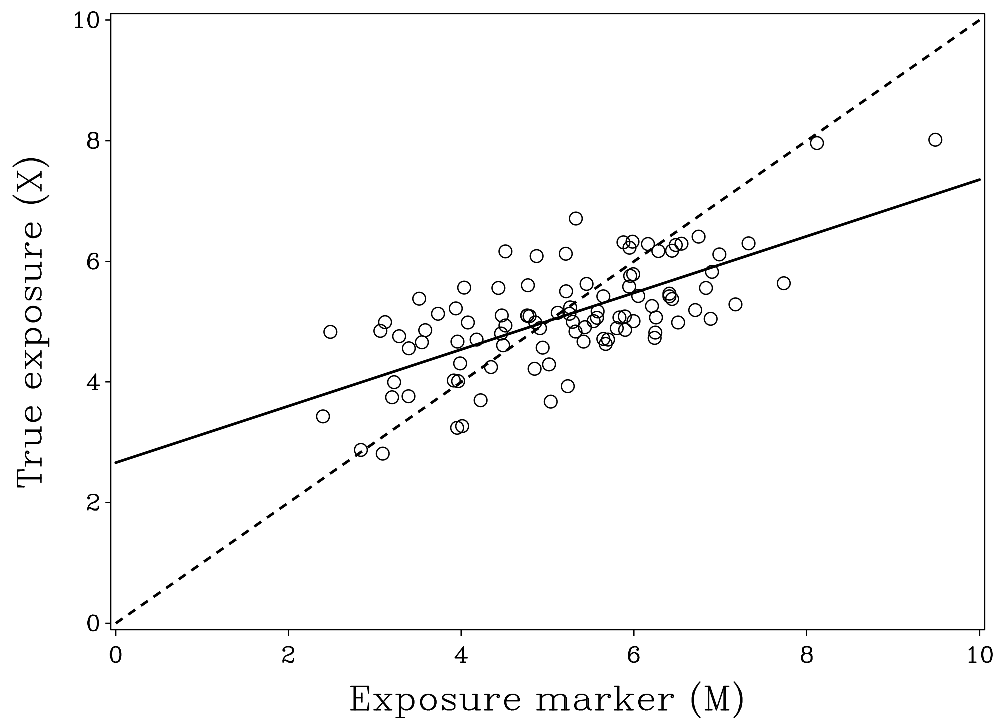


Slope in regression line is 1: if M is increased by 1, X is increased by 1

Regress Y on M: how much is Y increased when M is increased by 1?

If M is increased by 1, X is increased by 1 and Y is increased by β ($Y = \alpha + \beta X$)

Classical error: $M = X + U$



More horizontal variation $\rightarrow$ regression line has slope < 1

Regress Y on M: how much is Y increased when M is increased by 1?

If M is increased by 1, X is increased by less than 1 and Y is increased by *less* than β

Analysis of JEM-data

	Part 1				Part 2			
	.	.	...	.	.	.	...	.
1	$W_{1,1}$	$W_{1,2}$	...	W_{1,n_1}	.	.	...	.
	.	.	...	.	Y_{1,n_1+1}	Y_{1,n_1+2}	...	Y_{1,n_1+n_2}
.								
.								
.								
	.	.	...	.	.	.	...	.
G	$W_{G,1}$	$W_{G,2}$	...	W_{G,n_1}	.	.	...	.
	.	.	...	.	Y_{G,n_1+1}	Y_{G,n_1+2}	...	Y_{G,n_1+n_2}

- No measurements of true exposure.
- Part 1: no measurements of outcome.
- Part 2: no measurements of exposures.
- All subjects: their group is known.

Analysis based exposure group means

1. In group g estimated exposure: $\bar{W}_g = \frac{1}{n_1} \sum_{i=1}^{n_1} W_{gi}$ (part 1)

2. Regress Y_{gi} on $\bar{W}_g$ (part 2)

Error model: Classical and Berkson error

$$\begin{aligned} X_{gi} &= \mu_g + \gamma_{gi} & &= \mu_g + E_1 \\ \bar{W}_g &= \mu_g + \frac{1}{n_1} \sum_{i=1}^{n_1} (\gamma_{gi} + \epsilon_{gi}) & &= \mu_g + E_2 \end{aligned}$$

Response model: $Y_{gi} = \alpha + \beta X_{gi} + \zeta_{gi}$

1. Replace X_{gi} by μ_g - Berkson error

2. Replace μ_g by $\bar{W}_g$ - Classical error

Bias when using exposure group means

Theoretical relationship:

$$E(Y_{gi}|\bar{W}_g) = const. + \lambda\beta\bar{W}_g$$

I estimate $\lambda\beta$ - not β

$$\lambda = \frac{\sigma_b}{\sigma_b + (\sigma_w + \sigma_e)/n_1} \in [0;1]$$

Stronger under-estimation when with-in group variation (σ_w) and day-to-day variation (σ_e) is high and exposure means were calculated based on small sample size (n_1).

A larger between-group variation (σ_b) means smaller bias.

Unbiased analysis with BLUP

Fit Mixed Model to part 1 data. Group is included as a *random effect*

$$W_{gi} = L_g + \text{error}_{gi}$$

where L_g is a latent variable with normal distribution $N(\alpha, \sigma)$

Best linear unbiased predictor (BLUP) is an estimate of L_g

$$\text{BLUP}_g = \lambda \bar{W}_g + (1 - \lambda)\bar{W}, \text{ where } \lambda = \frac{\sigma_b}{\sigma_b + (\sigma_w + \sigma_e)/n_1}$$

From previous slides we have: $E(Y_{gi}|\bar{W}_g) = const. + \beta\lambda\bar{W}_g = co\tilde{n}st. + \beta\text{BLUP}_g$

Group averages are shrunken by the correct factor.

Simulations: Group mean vs BLUP

Simulation study:

Simulate data. Estimate exposure effect. Compare to truth $\beta = 0.5$ (y=0.5 x +error)

method	mean	se
2. Y_{gi} on BLUP_g	0.501	0.091
1. Y_{gi} on $\bar{W}_g$	0.439	0.080

BLUP in unbiased

Representative sampling

If the between job variation (σ_b) is different in the part 1 data from the part 2 data then the BLUP-analysis will be biased.

- Part 1: $\text{BLUP}_g = \lambda_1 \bar{W}_g$

- Part 2: $E(Y_{gi}|\bar{W}_g) = const. + \beta\lambda_2\bar{W}_g \neq const. + \beta\text{BLUP}_g$

$$\lambda_1 = \frac{\sigma_{b1}}{\sigma_{b1} + (\sigma_w + \sigma_e)/n_1} \neq \frac{\sigma_{b2}}{\sigma_{b2} + (\sigma_w + \sigma_e)/n_1} = \lambda_2$$

Modified BLUP

Between job variation: different in part 1 and part 2 data.

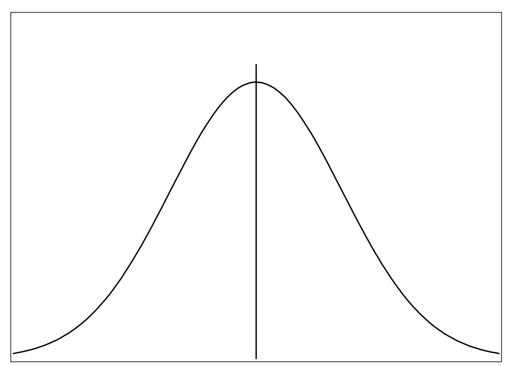
Calculate between group variation based on part 2 data.
Each subject assigned exposure $\bar{W}_g$.
First overall mean exposure in part 2: $\bar{W}$ and $\sum_g n_2(\bar{W}_g - \bar{W})^2/(\text{groups} - 1)$ and obtain: σ_{b2}.

Note: within-group variation is wrong (0) in part 2, but I use the part 1 value

New: $\lambda_2 = \frac{\sigma_{b2}}{\sigma_{b2}+(\sigma_w+\sigma_e)/n_1}$ and modified $\text{BLUP}_g = \lambda_2 \bar{W}_g$

Example: Part 2 data includes only jobs with exposure above the mean

method	mean	se
3. Y_{gi} on modified BLUP_g	0.499	0.057
2. Y_{gi} on BLUP_g	0.461	0.048
1. Y_{gi} on $\bar{W}_g$	0.439	0.046



Complications

- **Sample size varies with job title (part 1 data)**:
  This is taken into account by the BLUP. Jobs with small sample size shrunken more.

$$\lambda_g = \frac{\sigma_b}{\sigma_b + (\sigma_w + \sigma_e)/n_{1g}}$$

- **Confounder adjustment**:
  BLUP must be modified. Include confounder (here: sex) in analysis of part 1 data also.

$$\lambda = \frac{var(\mu_g|\text{sex})}{var(\mu_g|\text{sex}) + (\sigma_w + \sigma_e)/n_1}$$

Optimistic conclusion

Although the JEM design clearly introduces measurement error,

it is possible to get unbiased analysis even in (limited) design considered here.

Non-linear models

$Y_{gi} = H(\beta_0 + \beta X_{gi}) + \zeta_{gi} = H[\beta_0 + \beta\text{BLUP}_g + \beta\text{error}_{gi}] + \zeta_{gi}$

$X_{gi} = \text{BLUP}_g + \text{error}_{gi}$ where $\text{error}_{gi} \sim N[0, (1-\lambda)\sigma_b + \sigma_w]$

Correct analysis: $f(Y_{gi}|\bar{W}_g) = \int f(Y_{gi}|X_{ig})f(X_{ig}|W_g)dX_{gi}$ - non-standard

Ignore error: $Y_{gi} = H[\beta_0 + \beta\text{BLUP}_g] + \zeta_{gi}$

Generally biased. Bias will depend on σ_w (variation within groups)

Need to have information about σ_w. Repeated measurements in the same workers, will give information about σ_e and σ_w can be estimated.

Logistic regression

True model: $logit[P(Y_{gi} = 1|X_{gi})] = \alpha + \beta X_{gi}$

Estimated model: $logit[P(Y = 1|\bar{W}_g)] \approx \tilde{\alpha} + \beta \frac{1}{\sqrt{1+k^2\beta^2 v}}\text{BLUP}_g$

I estimate $\beta \frac{1}{\sqrt{1+k^2\beta^2 v}}$, here: $v = (1 - \lambda)\sigma_b + \sigma_w$

σ_w: variation within groups (of true exposures X)

Unbiased estimation: Need to have information about σ_w. Repeated measurements in the same workers, will give information about σ_e and σ_w can be estimated because I know within group variation in W $(\sigma_w + \sigma_e)$

Logistic regression - rare disease assumption

$$P(Y=1|X_{gi}) = \frac{\exp(\beta_0+\beta X_{gi})}{1+\exp(\beta_0+\beta X_{gi})} \approx \exp(\beta_0+\beta X_{gi})$$

$$\begin{aligned} P(Y=1|\bar{W}_g) &= \int P(Y=1|X_{gi}) f(X_{gi}|\bar{W}_g) dX_{gi} \\ &\approx \int \exp(\beta_0+\beta X_{gi}) N(.,.) dX_{gi} \\ &= \exp[\tilde{\beta}_0 + \beta \text{BLUP}_g + 0.5\beta^2 v] \end{aligned}$$

where $v = (1-\lambda)\sigma_b + \sigma_w$

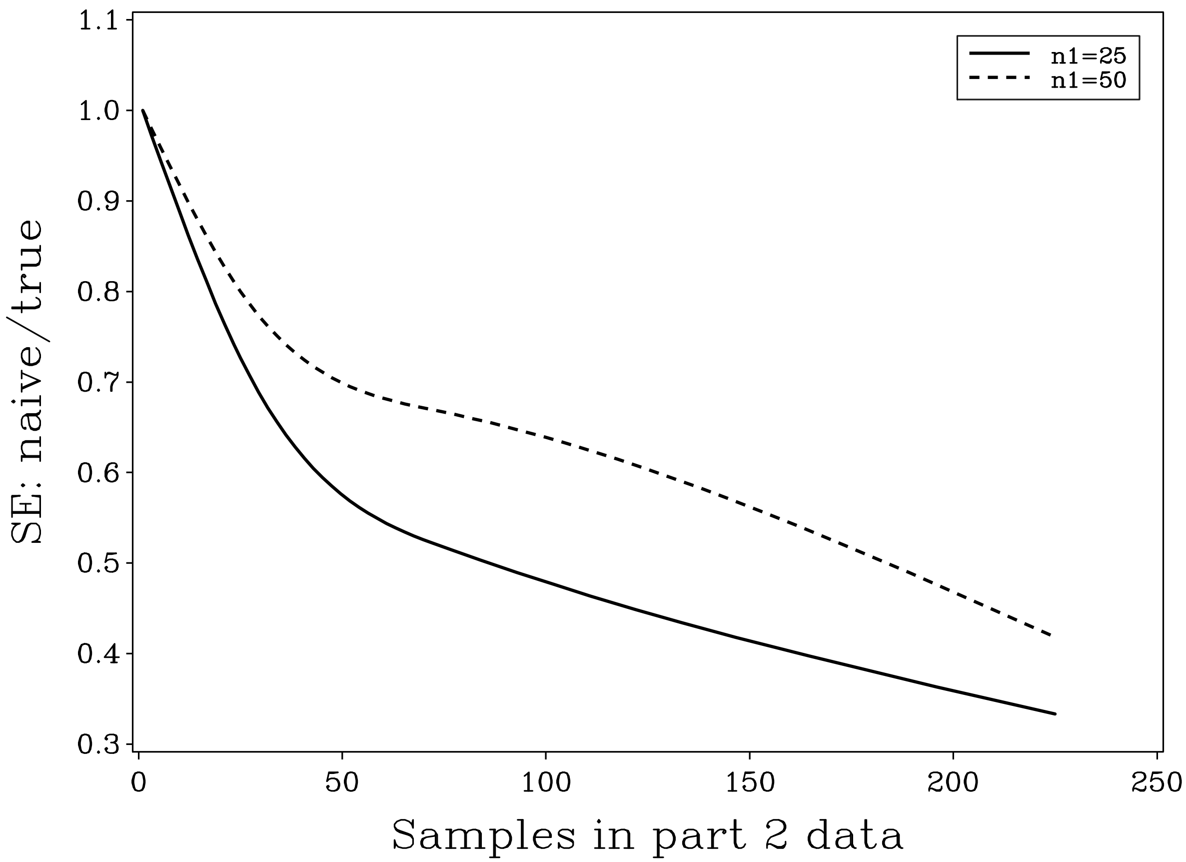
Regression using BLUP works, BUT last term (λ) is not constant if job titles are sampled with different ns, i.e.,

$$\lambda_g = \frac{\sigma_b}{\sigma_b + (\sigma_w+\sigma_e)/n_{1g}}$$

Standard errors? - allow for step 1 uncertainty

Example: 30 groups, $n_1 = 25$ or $n_1 = 50$, n_2 varies. Compare standard errors for $\widehat{\beta}$

n_1: observations in each group in part 1 data. n_2: observations in each group in part 2 data.



Standard errors - sandwich formula

Two-step procedure. Stacked estimating equations...

$$\sum_{i=1}^{n} S_i(Y_i, X_i, \theta) \;=\; \begin{pmatrix} \sum_{i=1}^{n} \phi_i(W_i, G, \theta_1) \\ \sum_{i=1}^{n} \psi_i(Y_i, G, \theta_1, \theta_2) \end{pmatrix} \;=\; \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Standard error: use standard GEE-theory (sandwich formula)

IV-setup: Step 1 parameters are μ_g, but what about λ...

Need to find the right “model” for applying this...

$var(\widehat{\theta_2}) =$ naive variance $+$ variance due to $\widehat{\theta}_1$

Sandwich good, but easier solutions?

Problem: Naive standard errors for $\widehat{\beta}$ are too low

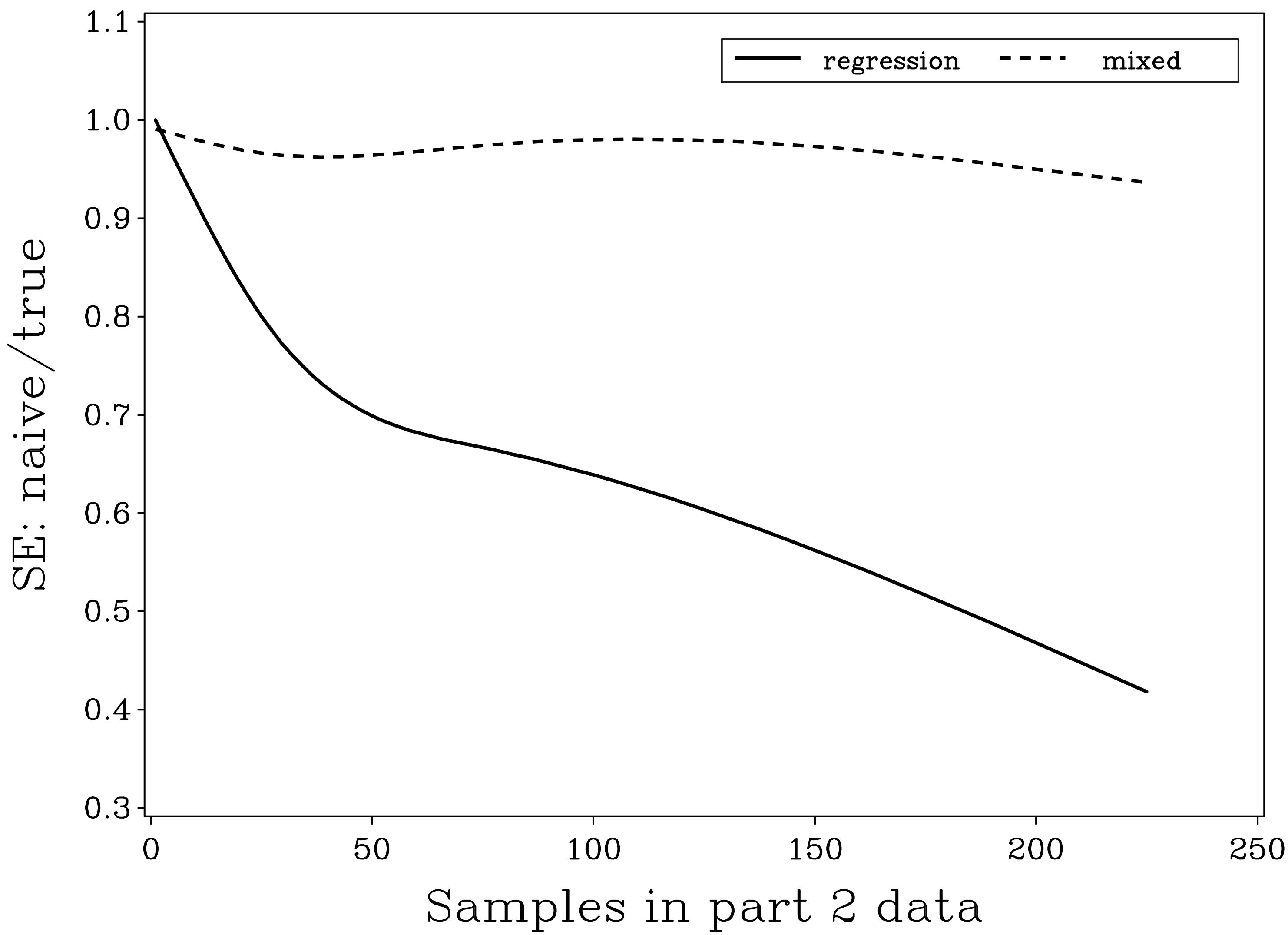
Step II: regress Y_{gi} on $\bar{W}_g$ or BLUP

Model assumes independent observations, but actually responses from the same group are positively correlated.

$$cov(Y_{gi}, Y_{gj}|\bar{W}_g) = \beta^2 \sigma_b (1 - \lambda) > 0$$

Solution: Fit mixed model (random group effect) also in part 2.
Probably a good idea also without this argument. Also it's easy.

Random effect of job-title also in step II



Challenges: misclassification of job titles

Not a problem in part 1 data. In part 2, some subjects may get the wrong job title.

Complicated: job title is categorical. Would need to know misclassification matrix...

M_{gi} : exposure assigned to subject i group g.

New error model:

$$\begin{aligned} X_{gi} &= \mu_g + \gamma_{gi} & &= \mu_g + E_1 \\ M_{gi} &= \mu_g + \tfrac{1}{n_1}\textstyle\sum_{i=1}^{n_1}(\gamma_{gi} + \epsilon_{gi}) + \delta_{gi} & &= \mu_g + E_2 \end{aligned}$$

$\delta_{gi} = M_{gi} - \bar{W}_g$ classical error, does not go away if more people are included in part 1.

Variance of δ_{gi} estimated if some subjects present in both part 1 and part 2. Compare assigned mean to correct mean: $\frac{1}{m-1}\sum_{i=1}^{m}(M_{gi} - \bar{W}_g)^2$

Challenges: User-friendly implementation of general adjustment method

Create package in R (or SAS or STATA) to conduct ML-estimation where latent variables are integrated out using numerical integration techniques.

If measurement error variances $(\sigma_b, \sigma_w, \sigma_e)$ are assumed to be known. Then the distribution of true exposure given measured can be calculated and used for the latent variable.

Generalization: where latent variables do not have to be normal, e.g., using mixture models.

Challenges: cumulative exposure history

$Y_i = \beta_0 + \beta \sum_t X_i(t) + \zeta_{gi}$

$X_i(t)$: exposure in year t

Part 1 data must be representative of exposure levels for all years considered.

Calculation of BLUP

$E(Y_i|W,G) = \beta_0 + \beta \sum_t E(X_i(t)|W,G) + \zeta_{gi}$

Do I need a model for the true process $X_i(t)_{t=0,\ldots,T}$?

What happens if I just plug-in the corresponding series of group exposure means?

Measurement errors will have temporal correlation. A subject above the mean one year is likely to be above the mean the following year. But what if he changes job...

Summary

- Linear model

  1. Information in the design is sufficient
  2. BLUP is unbiased
  3. JEM: $(\bar{W}_g, \sigma_w + \sigma_e, \sigma_b, n_1)$

- Non-linear model (logistic regression, cox-regression)

  1. Information in the design is not sufficient
  2. Information about σ_w is required
  3. BLUP may be biased

- Standard errors: adjust for uncertainty in part 1 data.