Methodological developments: unbiased effect estimation in studies using job exposure matrices

Esben Budtz-Jørgensen and Frederik Sørensen

Section of Biostatistics, University of Copenhagen

How to relate job exposure matrix data to health outcomes?

I will consider the measurement error problem:

For each subject the exposure is calculated from a JEM,

but the calculated exposure may not be equal to the true exposure.

The complete data

Consider a grouped setup. Workers are grouped according to e.g. job title.

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

X: true exposure, W: measured exposure, Y: response

X: not available - but it exists

Health effect model

Y_{gi} health measure for subject i in group g

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

Note, that it is the X-values which affect outcome.

β: increase in response when exposure is increased by 1 unit

Non-linear models Y_{gi} is binary. Logistic regression model

$$P(Y_{gi} = 1|X_{gi}) = \frac{\exp(\alpha + \beta X_{gi})}{1 + \exp(\alpha + \beta X_{gi})}$$

Aim: Data and statistical methods to obtain unbiased estimate of β

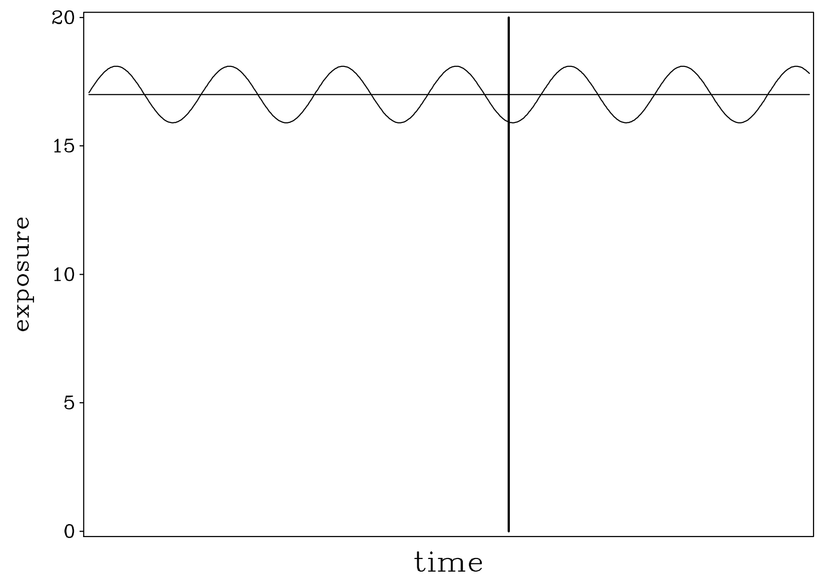
Variance components in exposures

Workers are grouped according to e.g. job title.

This means exposures have three sources of variation:

1. variation within each worker
2. variation between workers in the same group
3. variation between groups

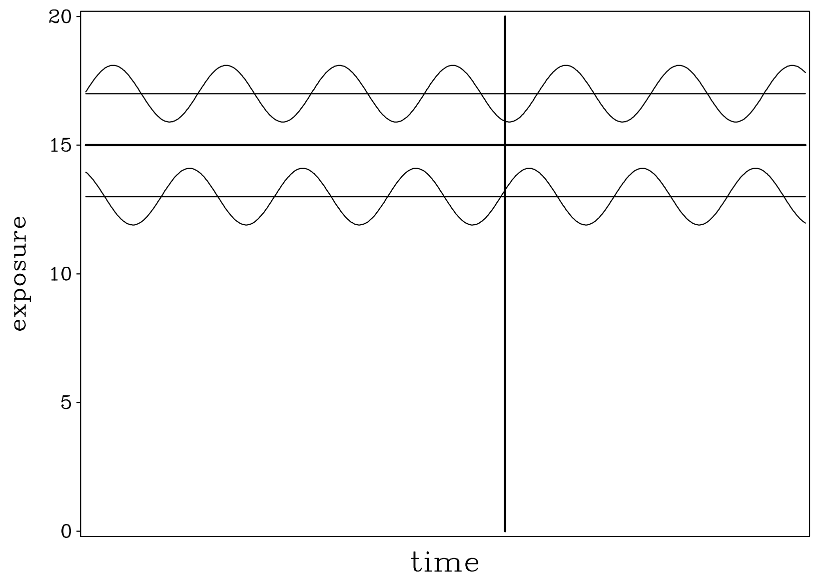
Variation with-in worker



Exposure of subject i, group g: $W_{gi} = X_{gi} + \epsilon_{gi}$

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

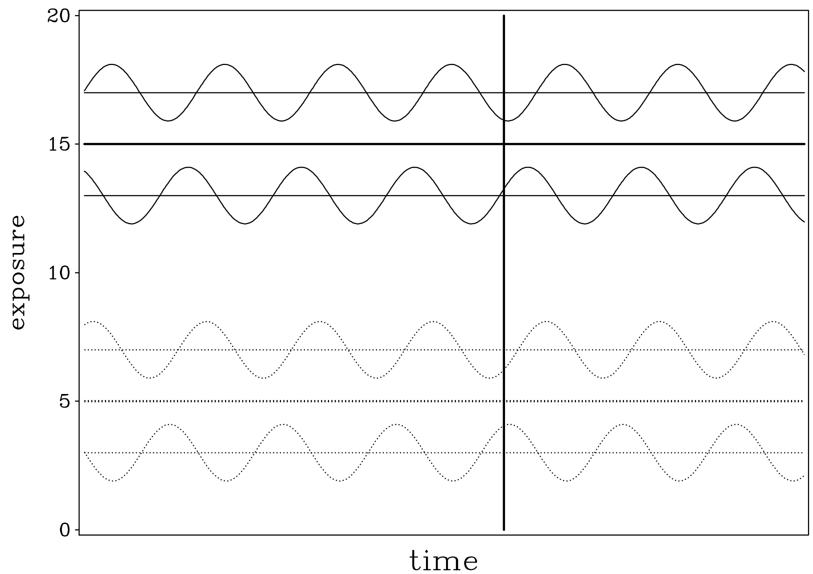
Variation with-in group



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

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

Variation between groups



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

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

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

Regression Rule

General regression problem

$$Y = \alpha + \beta X + \epsilon$$

Estimator: $\widehat{\beta} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$

$$\widehat{\beta} \rightarrow \frac{cov(X,Y)}{var(X)} = \frac{cov(X,\beta X + \epsilon)}{var(X)} = \frac{\beta var(X) + cov(X,\epsilon)}{var(X)} = \frac{\beta var(X) + 0}{var(X)} = \beta$$

Regression Rule

covariate (X) and residual (ϵ) must be uncorrelated $[cov(X, \epsilon) = 0]$ or $\widehat{\beta}$ is biased

Epidemiological study - data

Measures W and Y in all subjects. Forgets about the grouping.

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

Epidemiological study - analysis

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

We do not have X but W: $W_{gi} = X_{gi} + \epsilon_{gi}$ $(X_{gi} = W_{gi} - \epsilon_{gi})$

Naive analysis: Regress Y on W

$$Y_{gi} = \alpha + \beta X_{gi} + \zeta_{gi} = \alpha + \beta W_{gi} - \beta \epsilon_{gi} + \zeta_{gi}$$

Covariate: $W_{gi} = X_{gi} + \epsilon_{gi}$ **Residual:** $-\beta \epsilon_{gi} + \zeta_{gi}$

Correlation! - ϵ_{gi} affects both terms

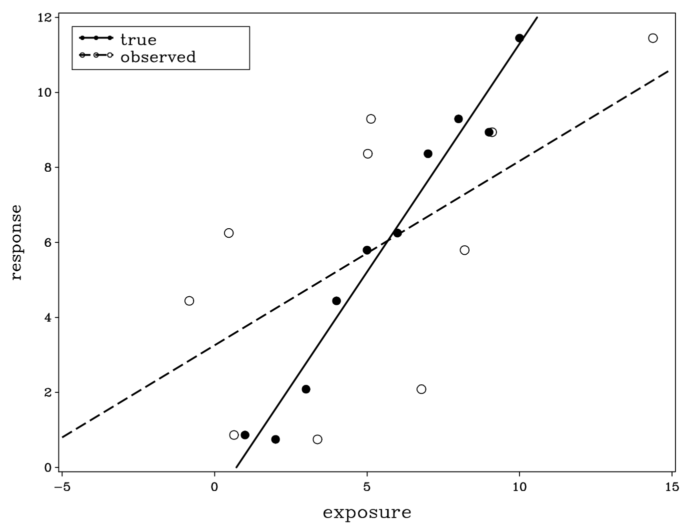
Regression Rule violated - exposure effect biased

Classic error gives bias

I replace X_{gi} by $W_{gi} = X_{gi} + \epsilon_{gi}$

Classical error: marker $=$ truth $+$ error with $cov(\text{error}, \text{truth}) = 0$

$Y = \alpha + \beta \cdot \text{truth} + \epsilon = \alpha + \beta \cdot \text{marker} - \beta \cdot \text{error} + \epsilon$



Exposure effect is underestimated

$$\widehat{\beta} \rightarrow \frac{\sigma_b + \sigma_w}{\sigma_b + \sigma_w + \sigma_e}\beta$$

JEM-data

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

- **Part 1**: no measurements of outcome (X, Y missing)
- **Part 2:** no measurements of exposures (X, W missing)
- **All subjects: group is known**

Analysis based on exposure group averages

1. **Part 1 data:** In group g estimated exposure: $\bar{W}_g = \frac{1}{n_1}\sum_{j=1}^{n_1} W_{gj}$
2. **Part 2 data:** Regress Y_{gi} on $\bar{W}_g$

Measurement error problem: $\bar{W}_g$ is different from X_{gi}

Analysis based on exposure group averages

1. **Part 1 data:** In group g estimated exposure: $\bar{W}_g = \frac{1}{n_1}\sum_{j=1}^{n_1} W_{gj}$

2. **Part 2 data:** Regress Y_{gi} on $\bar{W}_g$

$W_{gi} = X_{gi} + \epsilon_{gi} = \mu_g + \gamma_{gi} + \epsilon_{gi}$

$\bar{W}_g = \frac{1}{n_1}\sum_j W_{gj} = \mu_g + \frac{1}{n_1}\sum_j(\gamma_{gj} + \epsilon_{gj})$

Now pretend that we have μ_g in our data file ($n_1 = \infty$)

Regress Y on μ_g

$$Y_{gi} = \alpha + \beta X_{gi} + \zeta_{gi} = \alpha + \beta\mu_g + \beta\gamma_{gi} + \zeta_{gi}$$

Covariate: μ_g **Residual:** $\beta\gamma_{gi} + \zeta_{gi}$

covariate and residual **not** correlated - analysis unbiased

Berkson error gives no bias

$Y_{gi} = \alpha + \beta X_{gi} + \zeta_{gi}$ and $X_{gi} = \mu_g + \gamma_{gi}$. I replace X_{gi} by μ_g.

Berkson error: truth = marker + error with cov(error, marker) = 0

$Y = \alpha + \beta \cdot \text{truth} + \epsilon = \alpha + \beta \cdot \text{marker} + \beta \cdot \text{error} + \epsilon$

If we use the true group means, then no bias is induced (Regression Rule).

Remember for Classic error: cov(error, truth) = 0

Analysis based on exposure group averages, II

1. **Part 1 data:** In group g estimated exposure: $\bar{W}_g = \frac{1}{n_1}\sum_j W_{gj}$
2. **Part 2 data:** Regress Y_{gi} on $\bar{W}_g$

$\bar{W}_g = \mu_g + \frac{1}{n_1}\sum_j(\gamma_{gj} + \epsilon_{gj})$

marker = truth + error, $cov(\text{error}, \text{truth}) = 0$: $\bar{W}_g$ has classic error

Bias introduced when replacing μ_g with $\bar{W}_g$

$$\widehat{\beta} \to \lambda\beta \text{ with } \lambda = \frac{\sigma_b}{\sigma_b + (\sigma_w + \sigma_e)/n_1}$$

Avoiding bias in JEM analysis

JEM analysis is biased only because of classic error in the relation between $\bar{W}_g$ and μ_g

$$\bar{W}_g = \mu_g + \frac{1}{n_1}\sum_j(\gamma_{gj} + \epsilon_{gj})$$

It is possible to find alternative marker with Berkson error? i.e.

$$\mu_g = \text{marker} + \text{error}, \text{ with } cov(\text{error}, \text{marker}) = 0$$

Best Linear Unbiased Predictor (BLUP):

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

BLUP-calculation: Fit mixed model to part 1 data. Group is random effect

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

L_g latent variable with normal distribution. $\text{BLUP}_g = \widehat{L}_g = E(L_g|\bar{W}_g)$

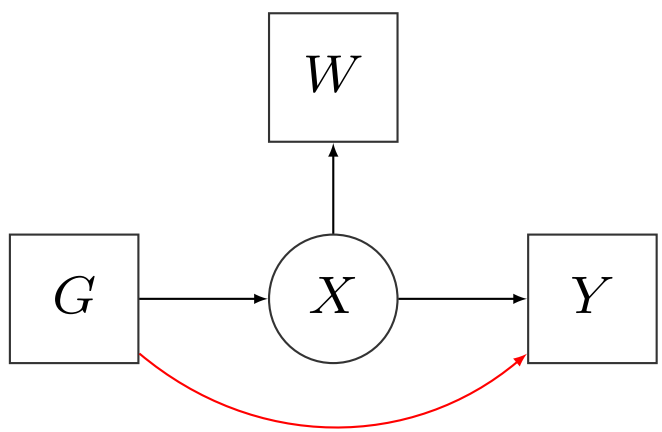
Comparison to epidemiological study

Epidemiological study (W_i, Y_i): $\widehat{\beta} \to \frac{var(X)}{var(W)}\beta$

$var(X)$ not identified - we cannot correct this

JEM study $(G_i, W_i) + (G_i, Y_i)$: $\widehat{\beta} \to \beta$

JEM study exploits assumption that G affects Y only indirectly through X



Red arrow will bias JEM study

Simulation study: Standard errors

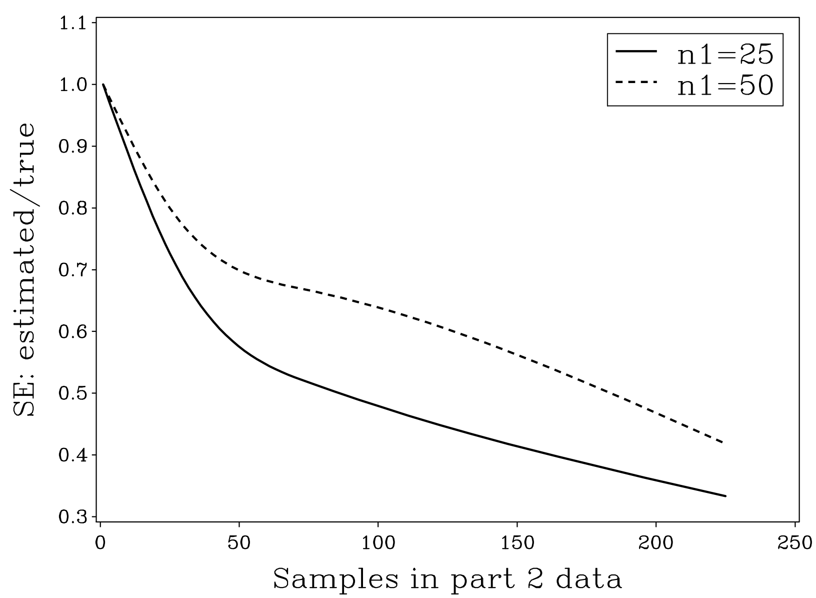
We use BLUP, but how does the corresponding standard errors perform?

1. Simulate data

2. Estimate $\beta + s.e.$ (naive)

3. repeat 1+2 many times

Comparison of naive standard error to true standard error



Standard errors too small?

$$Y_{gi} = \alpha + \beta X_{gi} + \zeta_{gi} = \alpha + \beta\mu_g + \beta\gamma_{gi} + \zeta_{gi}$$

I regress on BLUP:

$$Y_{gi} = \alpha + \beta \text{BLUP}_g + \beta(\mu_g - \text{BLUP}_g) + \beta\gamma_{gi} + \zeta_{gi}$$

Two subjects in the same group will have the same value of $\mu_g - \text{BLUP}_g$
$\longrightarrow$ responses from the same group are positively correlated

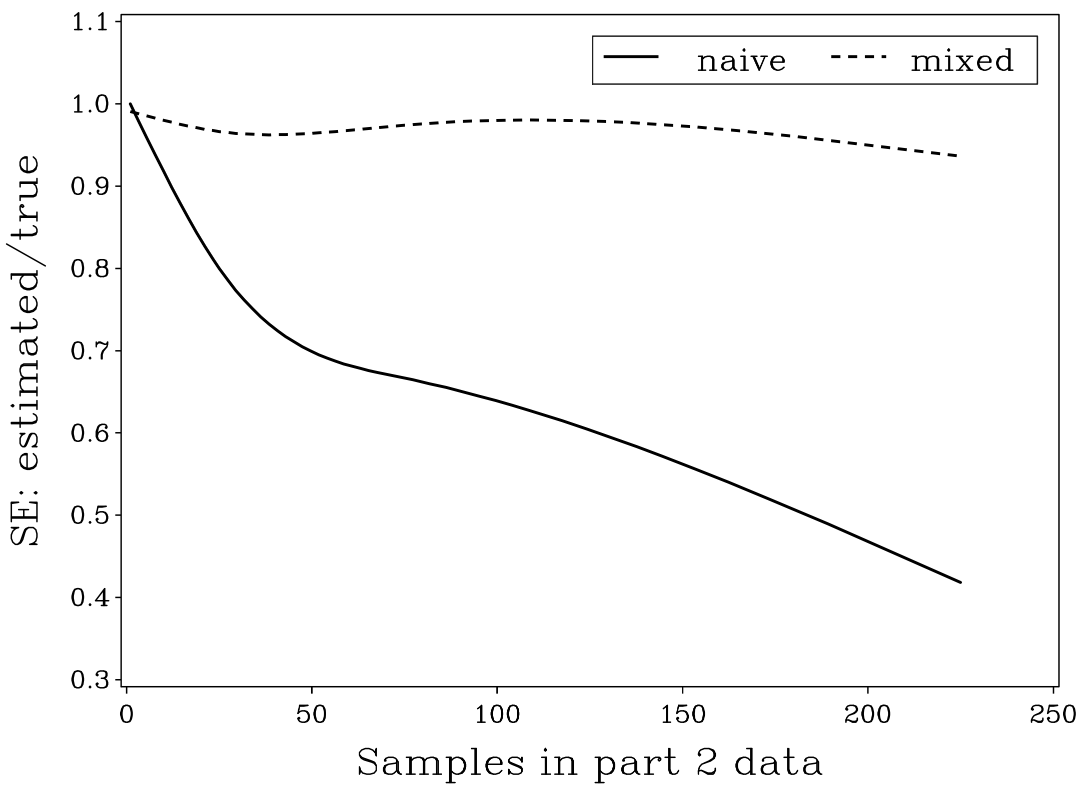
$$cov(Y_{gi}, Y_{gj}|\bar{W}_g) = \beta^2\sigma_b(1-\lambda) > 0$$

Solution: Fit mixed model (random group effect) also in part 2 data.

(Type 1 error correct)

Mixed model in part 2 data

Mixed model: Regress Y_{gi} on BLUP_g with *group* as random effect



Non-linear response models

Logistic regression model is considered.

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

- Estimated model: $logit[P(Y_{gi} = 1|\mathsf{BLUP}_g)] \approx \tilde{\alpha} + \frac{\beta}{\sqrt{1+c\beta^2 v}} \cdot \mathsf{BLUP}_g$

I estimate $\beta \frac{1}{\sqrt{1+c\beta^2 v}}$ not β **Berkson errors give bias - underestimation**

σ_w: variation within groups, $v = (1-\lambda)\sigma_b + \sigma_w$, $c = 0.35$

Need information about true variation within groups σ_w (variation in X).
Easy to estimate total variation with-in groups $\sigma_w + \sigma_e$ (variation in W). Repeated measurements in workers, will give information about σ_e and then σ_w can be estimated.

Simulation study: Bias using BLUP in non-linear model

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

I varied

- **Risk of event**

  rare: $p \approx 6\%$ – common: $p \approx 50\%$

- **Size of exposure effect**

  weak: OR=1.22 – strong OR=2.22

Event	Effect	Bias in BLUP
rare	weak	-0.1%
rare	strong	-3.0%
common	weak	-2.0%
common	strong	-12.5%

When event is rare or effect is weak: BLUP almost unbiased

Summary

- Linear models

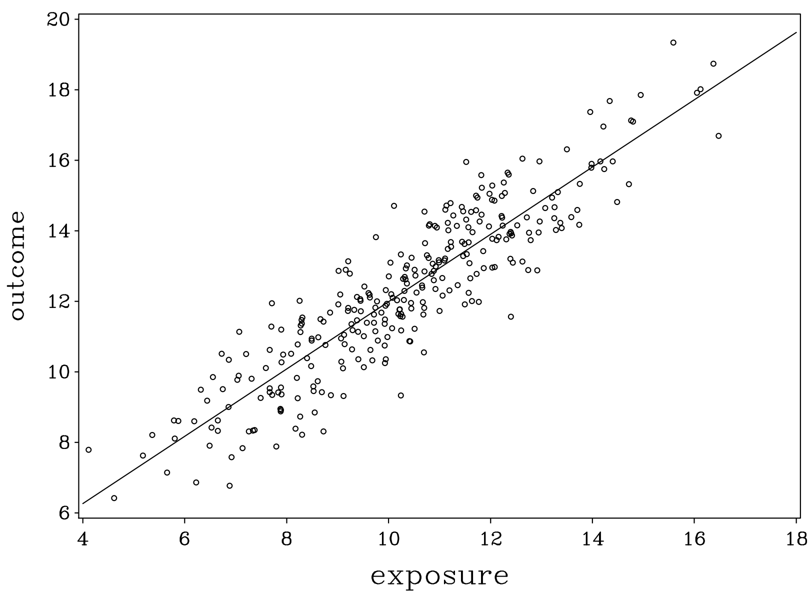
  1. Unbiased analysis based on BLUP
  2. Part 1 data must be representative of part 2
  3. *s.e.* may need adjustment (Type 1 error correct)

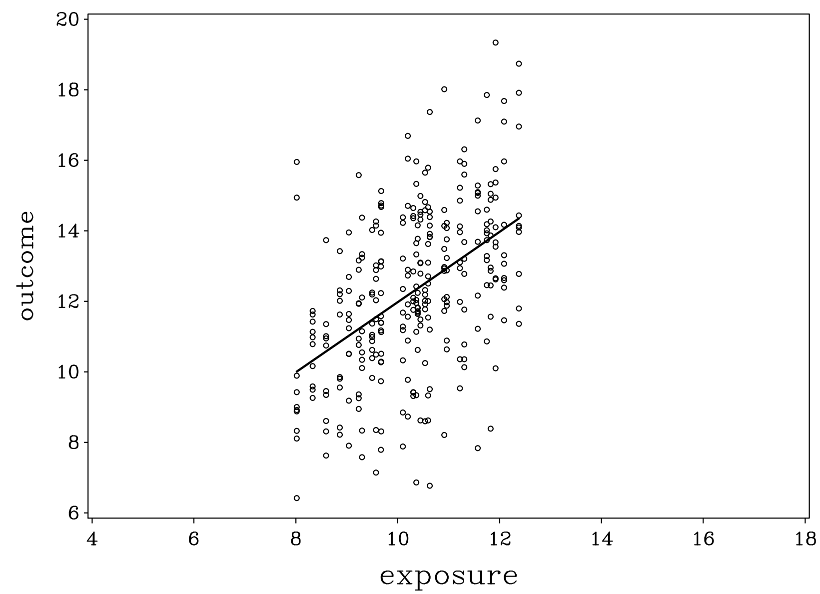
- Non-linear models

  1. BLUP generally biased
  2. but problem is small if event is rare *or* effect small
  3. Consistent analysis more complex (my current project)

Inefficiency of grouped approach

Plot X_{gi} vs Y_{gi} $\qquad\qquad$ μ_g vs Y_{gi}





Grouped analysis leads to weaker variation in exposure, and therefore *s.e.* of regression coefficient is increased.