

Correction for measurement error in nutritional epidemiology

A measurement error model allowing for never-consumers

Ruth Keogh^{1,2} & Ian White¹

¹MRC Biostatistics Unit, Cambridge, UK

²MRC Centre for Nutritional Epidemiology in Cancer Prevention and Survival, University of
Cambridge

ISCB 2009 Conference

Background and motivation

Question of interest

What is the association between 'usual' dietary intake and disease?

'Usual' intake of foods and nutrients: Long term average daily intake

EPIC-Norfolk

- ▶ European Prospective Investigation into Cancer and Nutrition
- ▶ Cohort of 25,000 individuals

UK Dietary Cohort Consortium

- ▶ 7 UK cohorts: 153,000 individuals

Measuring dietary intake using diet diaries

- ▶ EPIC-Norfolk: 7-day diet diaries
- ▶ UK Dietary Cohort Consortium: 4-7 day diet diaries

Measurement error in diet diaries

- ▶ A diet diary collects detailed information about dietary intake
- ▶ ...but it's just a 'snapshot' of the diet
- ▶ Measurements are subject to random within-person error

A specific source of measurement error in diet diaries

- ▶ We might not capture consumption of foods which are often not eaten every day, e.g. alcohol, meat
- ▶ Distinguish between **never-consumers** and **episodic-consumers**

Example: Alcohol intake in EPIC-Norfolk

		Measurement 2	
		0	> 0
Measurement 1	0	531 (21%)	248 (10%)
	> 0	261 (10%)	1522 (59%)

Measurement error in diet diaries

- ▶ A diet diary collects detailed information about dietary intake
- ▶ ...but it's just a 'snapshot' of the diet
- ▶ Measurements are subject to random within-person error

A specific source of measurement error in diet diaries

- ▶ We might not capture consumption of foods which are often not eaten every day, e.g. alcohol, meat
- ▶ Distinguish between **never-consumers** and **episodic-consumers**

Example: Alcohol intake in EPIC-Norfolk

		Measurement 2	
		0	> 0
Measurement 1	0	531 (21%)	248 (10%)
	> 0	261 (10%)	1522 (59%)

Measurement error in diet diaries

- ▶ A diet diary collects detailed information about dietary intake
- ▶ ...but it's just a 'snapshot' of the diet
- ▶ Measurements are subject to random within-person error

A specific source of measurement error in diet diaries

- ▶ We might not capture consumption of foods which are often not eaten every day, e.g. alcohol, meat
- ▶ Distinguish between **never-consumers** and **episodic-consumers**

Example: Alcohol intake in EPIC-Norfolk

		Measurement 2	
		0	> 0
Measurement 1	0	531 (21%)	248 (10%)
	> 0	261 (10%)	1522 (59%)

Correcting for measurement error

Measurement error results in biased estimates of the diet-disease association

T_i = True average daily intake
 R_{ij} = j^{th} diet diary measurement, $\mathbf{R}_i = \{R_{i1}, \dots, R_{iJ}\}$
 D_i = disease status (0/1)

True diet-disease association:

$$\Pr(D_i = 1 | T_i) = \frac{\exp(\alpha + \beta T_i)}{1 + \exp(\alpha + \beta T_i)}$$

Estimating β when we can't observe T_i

$$\Pr(D_i = 1 | \mathbf{R}_i) \approx \frac{\exp(\alpha + \beta E(T_i | \mathbf{R}_i))}{1 + \exp(\alpha + \beta E(T_i | \mathbf{R}_i))}$$

This is called regression calibration

Correcting for measurement error

Measurement error results in biased estimates of the diet-disease association

T_i = True average daily intake
 R_{ij} = j^{th} diet diary measurement, $\mathbf{R}_i = \{R_{i1}, \dots, R_{iJ}\}$
 D_i = disease status (0/1)

True diet-disease association:

$$\Pr(D_i = 1 | T_i) = \frac{\exp(\alpha + \beta T_i)}{1 + \exp(\alpha + \beta T_i)}$$

Estimating β when we can't observe T_i

$$\Pr(D_i = 1 | \mathbf{R}_i) \approx \frac{\exp(\alpha + \beta E(T_i | \mathbf{R}_i))}{1 + \exp(\alpha + \beta E(T_i | \mathbf{R}_i))}$$

This is called **regression calibration**

Performing the regression calibration

- ▶ To perform the regression calibration we need to find $E(T_i|\mathbf{R}_i)$

Linear regression calibration model

$$T_i = \lambda_0 + \lambda_1^T \mathbf{R}_i + e_i$$

To fit this model we need

- ▶ to assume $E(R_{ij}|T_i) = T_i$
- ▶ ≥ 2 measurements R_{ij}

Performing the regression calibration

- ▶ To perform the regression calibration we need to find $E(T_i|\mathbf{R}_i)$

Linear regression calibration model

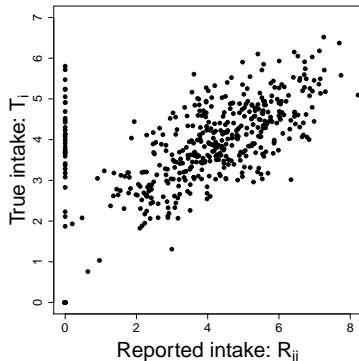
$$T_i = \lambda_0 + \lambda_1^T \mathbf{R}_i + e_i$$

To fit this model we need

- ▶ to assume $E(R_{ij}|T_i) = T_i$
- ▶ ≥ 2 measurements R_{ij}

Never-consumers and episodic-consumers

Is a linear regression calibration model appropriate when we have never-consumers and episodic-consumers?

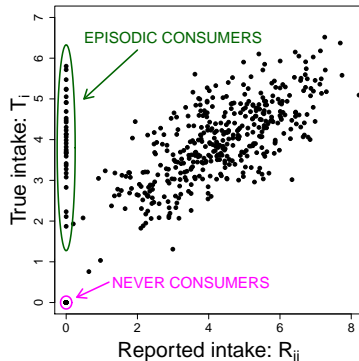


Aims

1. Define a measurement error model which allows never- and episodic-consumers
2. Find $E(T_{ij} | R_{ij})$ so that regression calibration can be performed

Never-consumers and episodic-consumers

Is a linear regression calibration model appropriate when we have never-consumers and episodic-consumers?

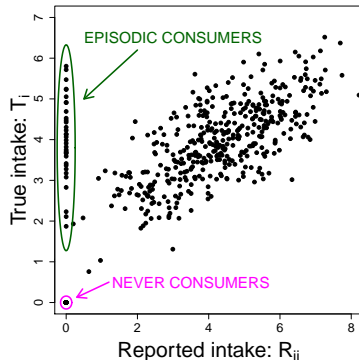


Aims

1. Define a measurement error model which allows never- and episodic-consumers
2. Find $E(T_{ij} | R_{ij})$ so that regression calibration can be performed

Never-consumers and episodic-consumers

Is a linear regression calibration model appropriate when we have never-consumers and episodic-consumers?



Aims

1. Define a measurement error model which allows never- and episodic-consumers
2. Find $E(T_{ij} | R_{ij})$ so that regression calibration can be performed

Never- and episodic-consumers (NEC) model

1. Never-consumers

Assumption: $T_i = 0 \Rightarrow R_{ij} = 0, \forall j$

$$u_{0i} = \begin{cases} 1 & \text{if person } i \text{ a never-consumer} \\ 0 & \text{if person } i \text{ a consumer} \end{cases}, P(u_{0i} = 1) = \frac{1}{1 + e^{\gamma_0}} = H(\gamma_0)$$

2. Episodic-consumers

$$\Pr(R_{ij} = 0 | \mathbf{u}_i) = \begin{cases} 1 & \text{if } u_{0i} = 1 \\ H(\gamma_1 + u_{1i}) & \text{if } u_{0i} = 0 \end{cases}$$

3. Measurement error for consumers

$$R_{ij} | \mathbf{u}_i \sim N(\gamma_2 + u_{2i}, \sigma_\varepsilon^2) \quad \text{if } R_{ij} > 0$$

$$(u_{1i}, u_{2i}) \sim BVN \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{u_1}^2 & \rho_{u_1 u_2} \sigma_{u_1} \sigma_{u_2} \\ \rho_{u_1 u_2} \sigma_{u_1} \sigma_{u_2} & \sigma_{u_2}^2 \end{pmatrix} \right)$$

Never- and episodic-consumers (NEC) model

1. Never-consumers

Assumption: $T_i = 0 \Rightarrow R_{ij} = 0, \forall j$

$$u_{0i} = \begin{cases} 1 & \text{if person } i \text{ a never-consumer} \\ 0 & \text{if person } i \text{ a consumer} \end{cases}, P(u_{0i} = 1) = \frac{1}{1 + e^{\gamma_0}} = H(\gamma_0)$$

2. Episodic-consumers

$$\Pr(R_{ij} = 0 | \mathbf{u}_i) = \begin{cases} 1 & \text{if } u_{0i} = 1 \\ H(\gamma_1 + u_{1i}) & \text{if } u_{0i} = 0 \end{cases}$$

3. Measurement error for consumers

$$R_{ij} | \mathbf{u}_i \sim N(\gamma_2 + u_{2i}, \sigma_\varepsilon^2) \quad \text{if } R_{ij} > 0$$

$$(u_{1i}, u_{2i}) \sim BVN \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{u_1}^2 & \rho_{u_1 u_2} \sigma_{u_1} \sigma_{u_2} \\ \rho_{u_1 u_2} \sigma_{u_1} \sigma_{u_2} & \sigma_{u_2}^2 \end{pmatrix} \right)$$

Never- and episodic-consumers (NEC) model

1. Never-consumers

Assumption: $T_i = 0 \Rightarrow R_{ij} = 0, \forall j$

$$u_{0i} = \begin{cases} 1 & \text{if person } i \text{ a never-consumer} \\ 0 & \text{if person } i \text{ a consumer} \end{cases}, P(u_{0i} = 1) = \frac{1}{1 + e^{\gamma_0}} = H(\gamma_0)$$

2. Episodic-consumers

$$\Pr(R_{ij} = 0 | \mathbf{u}_i) = \begin{cases} 1 & \text{if } u_{0i} = 1 \\ H(\gamma_1 + u_{1i}) & \text{if } u_{0i} = 0 \end{cases}$$

3. Measurement error for consumers

$$R_{ij} | \mathbf{u}_i \sim N(\gamma_2 + u_{2i}, \sigma_\varepsilon^2) \quad \text{if } R_{ij} > 0$$

$$(u_{1i}, u_{2i}) \sim BVN \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{u_1}^2 & \rho_{u_1 u_2} \sigma_{u_1} \sigma_{u_2} \\ \rho_{u_1 u_2} \sigma_{u_1} \sigma_{u_2} & \sigma_{u_2}^2 \end{pmatrix} \right)$$

Fitting the NEC model

- ▶ Reported measurements are modelled using $\mathbf{u}_i = \{u_{0i}, u_{1i}, u_{2i}\}$
- ▶ 7 parameters $\theta = \{\gamma_0, \gamma_1, \gamma_2, \sigma_{u_1}^2, \sigma_{u_2}^2, \rho_{u_1 u_2}, \sigma_\varepsilon^2\}$

Joint distribution of the \mathbf{R}_i

$$f(\mathbf{R}_i) = \underbrace{\{1 - H(\gamma_0)\} \int f(\mathbf{R}_i | \mathbf{u}_i, u_{0i} = 1) f(u_{1i}, u_{2i}) du_{1i} du_{2i}}_{\text{Consumers}} + \underbrace{H(\gamma_0) \prod_{j=1}^J (1 - I_{(R_{ij} > 0)})}_{\text{Never-consumers}}$$

Parameters θ can be estimated by maximum likelihood provided we have ≥ 2 measurements R_{ij}

Finding fitted values $E(T_i|\mathbf{R}_i; \theta)$

Assumption

Reported measurements R_{ij} are **unbiased estimates** of true intake T_i

$$\begin{aligned} T_i &= E(R_{ij}|\mathbf{u}_i) \\ &= \begin{cases} 0 & \text{if } u_{0i} = 1 \\ \{1 - H(\gamma_1 + u_{1i})\}(\gamma_2 + u_{2i}) & \text{if } u_{0i} = 0 \end{cases} \end{aligned}$$

Fitted values for true intake

$$E(T_i|\mathbf{R}_i; \theta) = \frac{\int T_i(\mathbf{u}_i)f(\mathbf{R}_i|\mathbf{u}_i; \theta)f(\mathbf{u}_i; \theta)d\mathbf{u}_i}{\int f(\mathbf{R}_i|\mathbf{u}_i; \theta)f(\mathbf{u}_i; \theta)d\mathbf{u}_i}$$

Finding fitted values $E(T_i | \mathbf{R}_i; \theta)$

Assumption

Reported measurements R_{ij} are **unbiased estimates** of true intake T_i

$$\begin{aligned} T_i &= E(R_{ij} | \mathbf{u}_i) \\ &= \begin{cases} 0 & \text{if } u_{0i} = 1 \\ \{1 - H(\gamma_1 + u_{1i})\} (\gamma_2 + u_{2i}) & \text{if } u_{0i} = 0 \end{cases} \end{aligned}$$

Fitted values for true intake

$$E(T_i | \mathbf{R}_i; \theta) = \frac{\int T_i(\mathbf{u}_i) f(\mathbf{R}_i | \mathbf{u}_i; \theta) f(\mathbf{u}_i; \theta) d\mathbf{u}_i}{\int f(\mathbf{R}_i | \mathbf{u}_i; \theta) f(\mathbf{u}_i; \theta) d\mathbf{u}_i}$$

Some questions

1. How well can we estimate the **parameters** of the NEC model?
2. Is the NEC model successful in allowing us to correct for the effects of measurement error on the **diet-disease association**?
3. How do the results from the NEC model compare with **alternative approaches**?

Simulation study: Alcohol intake in EPIC-Norfolk

- ▶ We fitted the never- and episodic-consumers model for alcohol intake in EPIC-Norfolk
- ▶ 2 reported measurements R_{i1}, R_{i2} for a subset of the study population

Parameter	Estimate (SE)
γ_1	2.29 (0.14)
γ_2	2.55 (0.07)
$\sigma_{u_1}^2$	6.90 (0.79)
$\sigma_{u_2}^2$	3.66 (0.16)
$\rho_{u_1 u_2}$	0.70 (0.01)
σ_ε^2	1.23 (0.06)
$H(\gamma_0)$	0.08 (0.01)

200 simulated data sets

- ▶ 1000 individuals ($i = 1, \dots, 1000$)
- ▶ Obtain true intake T_i
- ▶ Obtain reported measurements $\mathbf{R}_i = \{R_{i1}, R_{i2}, R_{i3}, R_{i4}\}$

We fit the NEC model using 2 measurements $\{R_{i1}, R_{i2}\}$ and 4 measurements $\{R_{i1}, R_{i2}, R_{i3}, R_{i4}\}$

Simulation study: Alcohol intake in EPIC-Norfolk

- ▶ We fitted the never- and episodic-consumers model for alcohol intake in EPIC-Norfolk
- ▶ 2 reported measurements R_{i1}, R_{i2} for a subset of the study population

Parameter	Estimate (SE)
γ_1	2.29 (0.14)
γ_2	2.55 (0.07)
$\sigma_{u_1}^2$	6.90 (0.79)
$\sigma_{u_2}^2$	3.66 (0.16)
$\rho_{u_1 u_2}$	0.70 (0.01)
σ_ε^2	1.23 (0.06)
$H(\gamma_0)$	0.08 (0.01)

200 simulated data sets

- ▶ 1000 individuals ($i = 1, \dots, 1000$)
- ▶ Obtain true intake T_i
- ▶ Obtain reported measurements $\mathbf{R}_i = \{R_{i1}, R_{i2}, R_{i3}, R_{i4}\}$

We fit the NEC model using 2 measurements $\{R_{i1}, R_{i2}\}$ and 4 measurements $\{R_{i1}, R_{i2}, R_{i3}, R_{i4}\}$

Simulation study: Alcohol intake in EPIC-Norfolk

- ▶ We fitted the never- and episodic-consumers model for alcohol intake in EPIC-Norfolk
- ▶ 2 reported measurements R_{i1}, R_{i2} for a subset of the study population

Parameter	Estimate (SE)
γ_1	2.29 (0.14)
γ_2	2.55 (0.07)
$\sigma_{u_1}^2$	6.90 (0.79)
$\sigma_{u_2}^2$	3.66 (0.16)
$\rho_{u_1 u_2}$	0.70 (0.01)
σ_ε^2	1.23 (0.06)
$H(\gamma_0)$	0.08 (0.01)

200 simulated data sets

- ▶ 1000 individuals ($i = 1, \dots, 1000$)
- ▶ Obtain true intake T_i
- ▶ Obtain reported measurements $\mathbf{R}_i = \{R_{i1}, R_{i2}, R_{i3}, R_{i4}\}$

We fit the NEC model using 2 measurements $\{R_{i1}, R_{i2}\}$ and 4 measurements $\{R_{i1}, R_{i2}, R_{i3}, R_{i4}\}$

Simulation study: Alcohol intake in EPIC-Norfolk

- ▶ We fitted the never- and episodic-consumers model for alcohol intake in EPIC-Norfolk
- ▶ 2 reported measurements R_{i1}, R_{i2} for a subset of the study population

Parameter	Estimate (SE)
γ_1	2.29 (0.14)
γ_2	2.55 (0.07)
$\sigma_{u_1}^2$	6.90 (0.79)
$\sigma_{u_2}^2$	3.66 (0.16)
$\rho_{u_1 u_2}$	0.70 (0.01)
σ_ε^2	1.23 (0.06)
$H(\gamma_0)$	0.08 (0.01)

200 simulated data sets

- ▶ 1000 individuals ($i = 1, \dots, 1000$)
- ▶ Obtain true intake T_i
- ▶ Obtain reported measurements $\mathbf{R}_i = \{R_{i1}, R_{i2}, R_{i3}, R_{i4}\}$

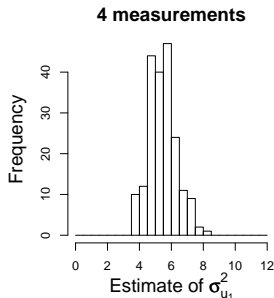
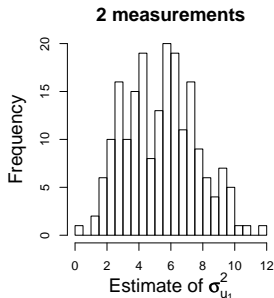
We fit the NEC model using 2 measurements $\{R_{i1}, R_{i2}\}$ and 4 measurements $\{R_{i1}, R_{i2}, R_{i3}, R_{i4}\}$

Simulation results: Parameter estimates

Param	True	Estimates: Mean (SD)	
		2 measurements	4 measurements
γ_1	2.29	2.32 (0.20)	2.39 (0.16)
γ_2	2.55	2.68 (0.13)	2.70 (0.11)
$\sigma_{U_1}^2$	6.90	5.40 (2.23)	5.44 (0.87)
$\sigma_{U_2}^2$	3.66	3.38 (0.23)	3.28 (0.17)
$\rho_{U_1 U_2}$	0.7	0.63 (0.03)	0.68 (0.03)
σ_{ε}^2	1.23	1.29 (0.07)	1.35 (0.04)
$H(\gamma_0)$	0.08	0.11 (0.05)	0.11 (0.02)

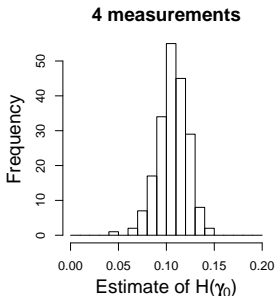
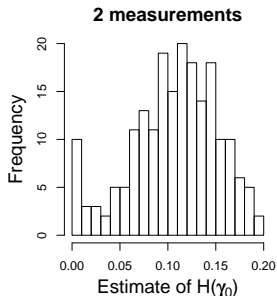
Simulation results: Parameter estimates

Param	True	Estimates: Mean (SD)	
		2 measurements	4 measurements
γ_1	2.29	2.32 (0.20)	2.39 (0.16)
γ_2	2.55	2.68 (0.13)	2.70 (0.11)
$\sigma_{u_1}^2$	6.90	5.40 (2.23)	5.44 (0.87)
$\sigma_{u_2}^2$	3.66	3.38 (0.23)	3.28 (0.17)
ρ_{u_1, u_2}	0.7	0.63 (0.03)	0.68 (0.03)
σ_ε^2	1.23	1.29 (0.07)	1.35 (0.04)
$H(\gamma_0)$	0.08	0.11 (0.05)	0.11 (0.02)



Simulation results: Parameter estimates

Param	True	Estimates: Mean (SD)	
		2 measurements	4 measurements
γ_1	2.29	2.32 (0.20)	2.39 (0.16)
γ_2	2.55	2.68 (0.13)	2.70 (0.11)
$\sigma_{U_1}^2$	6.90	5.40 (2.23)	5.44 (0.87)
$\sigma_{U_2}^2$	3.66	3.38 (0.23)	3.28 (0.17)
$\rho_{U_1 U_2}$	0.7	0.63 (0.03)	0.68 (0.03)
σ_{ϵ}^2	1.23	1.29 (0.07)	1.35 (0.04)
$H(\gamma_0)$	0.08	0.11 (0.05)	0.11 (0.02)



Correcting the diet-disease association

Regression calibration

We replace T_i by $E(T_i|\mathbf{R}_i)$ in the disease model

Simulation study

- ▶ We generated disease status (0/1) according to a logistic model

$$\Pr(D_i = 1 | T_i) = \frac{\exp(\alpha + \beta T_i)}{1 + \exp(\alpha + \beta T_i)}$$

- ▶ ...using $\beta = 0.2$
- ▶ α chosen to give a 10% disease probability

Compare with 3 alternative methods for estimating β

- ▶ 'Naive' method: Use $\text{mean}(R_{i1}, R_{i2})$ in place of T_i
- ▶ Using linear regression calibration to obtain $E(T_i|\mathbf{R}_i)$
- ▶ Using an episodic-consumers model to obtain $E(T_i|\mathbf{R}_i)$

Correcting the diet-disease association

Regression calibration

We replace T_i by $E(T_i|\mathbf{R}_i)$ in the disease model

Simulation study

- ▶ We generated disease status (0/1) according to a logistic model

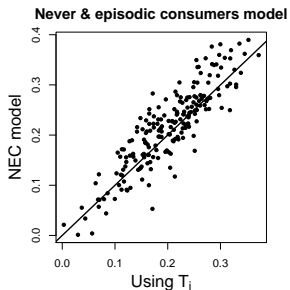
$$\Pr(D_i = 1 | T_i) = \frac{\exp(\alpha + \beta T_i)}{1 + \exp(\alpha + \beta T_i)}$$

- ▶ ...using $\beta = 0.2$
- ▶ α chosen to give a 10% disease probability

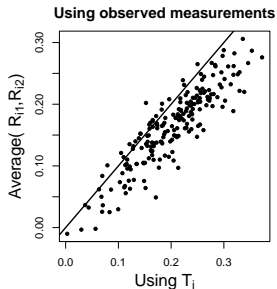
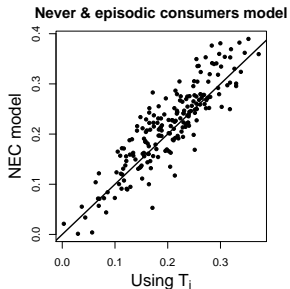
Compare with 3 alternative methods for estimating β

- ▶ 'Naive' method: Use $\text{mean}(R_{i1}, R_{i2})$ in place of T_i
- ▶ Using linear regression calibration to obtain $E(T_i|\mathbf{R}_i)$
- ▶ Using an episodic-consumers model to obtain $E(T_i|\mathbf{R}_i)$

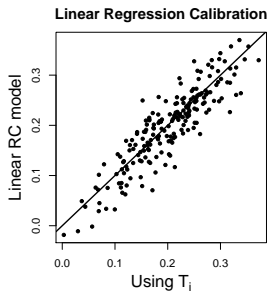
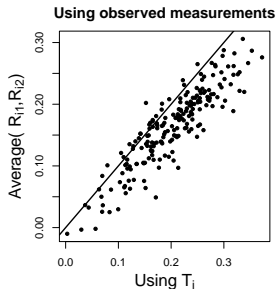
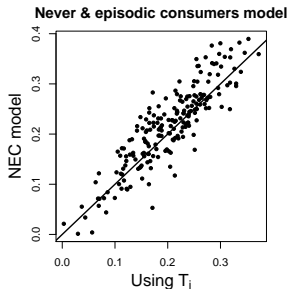
Comparison with alternative methods: log(OR)s



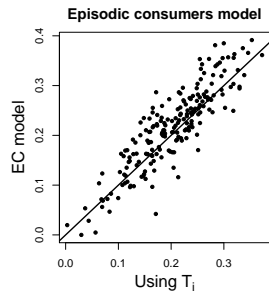
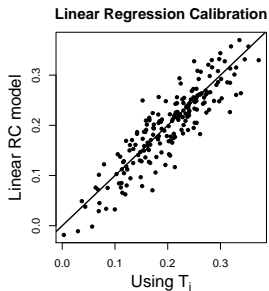
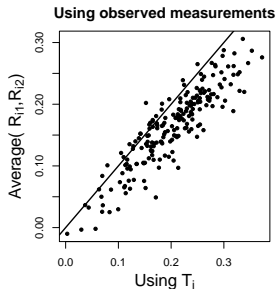
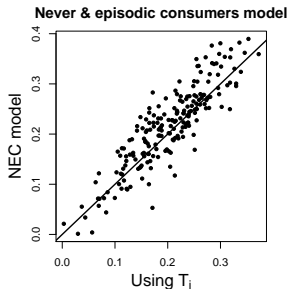
Comparison with alternative methods: log(OR)s



Comparison with alternative methods: log(OR)s



Comparison with alternative methods: log(OR)s



Some comments

- ▶ Some parameters of the model may be badly estimated using only two reported measurements R_{i1}, R_{i2} per person
- ▶ The never- and episodic-consumers model provides a method for correcting the diet-disease association for measurement error
- ▶ ...but we may often be able to achieve similar results using standard linear regression calibration or episodic-consumers model
- ▶ We have not yet looked at other aspects of the different approaches such a coverage probabilities
- ▶ There may be situations in which it is useful to be able to correctly model the association between true intake and reported intake using the never- and episodic-consumers model

Some comments

- ▶ Some parameters of the model may be badly estimated using only two reported measurements R_{i1} , R_{i2} per person
- ▶ The never- and episodic-consumers model provides a method for correcting the diet-disease association for measurement error
- ▶ ...but we may often be able to achieve similar results using standard linear regression calibration or episodic-consumers model
- ▶ We have not yet looked at other aspects of the different approaches such a coverage probabilities
- ▶ There may be situations in which it is useful to be able to correctly model the association between true intake and reported intake using the never- and episodic-consumers model

Some comments

- ▶ Some parameters of the model may be badly estimated using only two reported measurements R_{i1}, R_{i2} per person
- ▶ The never- and episodic-consumers model provides a method for correcting the diet-disease association for measurement error
- ▶ ...but we may often be able to achieve similar results using standard linear regression calibration or episodic-consumers model
- ▶ We have not yet looked at other aspects of the different approaches such a coverage probabilities
- ▶ There may be situations in which it is useful to be able to correctly model the association between true intake and reported intake using the never- and episodic-consumers model

Some comments

- ▶ Some parameters of the model may be badly estimated using only two reported measurements R_{i1}, R_{i2} per person
- ▶ The never- and episodic-consumers model provides a method for correcting the diet-disease association for measurement error
- ▶ ...but we may often be able to achieve similar results using standard linear regression calibration or episodic-consumers model
- ▶ We have not yet looked at other aspects of the different approaches such a coverage probabilities
- ▶ There may be situations in which it is useful to be able to correctly model the association between true intake and reported intake using the never- and episodic-consumers model