Combined effects of two or more variables, statistical considerations

Niels Keiding & Anne Helby Petersen
Section of Biostatistics

Scientific meeting on the use of composite variables in medical research
Danish Epidemiological Society & Danish Society for Occupational and Environmental Health
September 20, 2018 - Bispebjerg Hospital, Copenhagen

Example: PRISME-data

Data courtesy of Sigurd Mikkelsen, see e.g. Vammen et al. (2016) *JOEM* **58**, 994-1001.

Public employees from Aarhus, Denmark were recruited in 2007 and followed up in 2009. At baseline they filled out a questionnaire and the follow-up focused on diagnosis of clinical depression (individuals with clinical depression at baseline were excluded). Scored according to demands and control (dichotomized at median values) the following results were recorded:

	Depression: yes n = 3035	Depression: no n = 59	Total n = 3094
Low demands, high control	892 (29.4)	13 (22.0)	905 (29.3)
Low demands, low control	668 (22.0)	14 (23.7)	682 (22.0)
High demands, high control	795 (26.2)	7 (11.9)	802 (25.9)
High demands, low control	680 (22.4)	25 (42.4)	705 (22.8)

Logistic regression

We model

$$P(\text{clinical depression}) = \frac{\exp(\alpha + \mu_1 z_1 + \mu_2 z_2 + \cdots + \mu_k z_k)}{1 + \exp(\alpha + \mu_1 z_1 + \mu_2 z_2 + \cdots + \mu_k z_k)}$$

which is called $\quad \text{logit}(\alpha + \mu_1 z_1 + \mu_2 z_2 + \cdots + \mu_k z_k).$

In the beginning, we use the principal exposure variables

$$z_1 = \begin{cases} 1, & \text{if demands} > \text{median} \\ 0, & \text{if demands} < \text{median} \end{cases}$$

$$z_2 = \begin{cases} 1, & \text{if control} < \text{median} \\ 0, & \text{if control} > \text{median} \end{cases}$$

$$z_3 = z_1 \cdot z_2 = \begin{cases} 1, & \text{if demands} > \text{median and control} < \text{median } (\textit{'strain'}) \\ 0, & \text{otherwise} \end{cases}$$

Later additional covariates $z_4, z_5, \ldots$ may be added to handle confounding (examples: sex, age, socio-economic status).

Statistical models for occurrence of depression

Standard full statistical model

		demands	
		low	high
control	high	α	$\alpha + \delta$
	low	$\alpha + \gamma$	$\alpha+\gamma+\delta+\beta$

	Estimates	95% Conf. interv.	P
α: intercept	-4.229	(-4.829, -3.725)	
δ: main effect of demands	-0.504	(-1.487, 0.394)	0.285
γ: main effect of control	0.363	(-0.404, 1.137)	0.350
β: interaction effect of demands and control ('*strain*')	1.066	(-0.046, 2.248)	0.066

Tendency to interaction effect: the combination of high demands and low control increases probability of clinical depression.

Statistical models for occurrence of depression

Standard full statistical model

Net contrast to baseline: High control, low demands

		demands	
		low	high
control	high	*Baseline*	−0.504
	low	0.363	0.925*

	Estimates	95% Conf. interv.	P
δ: main effect of demands	-0.504	(-1.487, 0.394)	0.285
γ: main effect of control	0.363	(-0.404, 1.137)	0.350
β: interaction effect of demands and control ('*strain*')	1.066	(-0.046, 2.248)	0.066

$$(*)\ 0.925 = 0.363 - 0.504 + 1.066$$

Statistical models for occurrence of depression

Model IPD ('strain')

		demands	
		low	high
control	high	α	α
	low	α	$\alpha + \beta$

Main effects of control and demands are assumed to be zero.

	Estimates	95% Conf. interv.	P
α: intercept	-4.229	(-4.829, -3.725)	
β: interaction effect of demands and control ('*strain*')	0.935	(0.402, 1.454)	0.000

Strong 'strain' effect, partly generated by main effects that are not modelled.

The 'hierarchical principle' of regression analysis

If a model contains an interaction between categorical variables, then we must keep lower order interactions and main effects associated with this interaction in the model as well.

Violated by the '*strain*' model.

Statistical models for occurrence of depression

Model IPD ('strain')

Net contrast to baseline: High control and/or low demands

		demands	
		low	high
control	high	*Baseline*	*Baseline*
	low	*Baseline*	0.935

Main effects of control and demands are assumed to be zero.

	Estimates	95% Conf. interv.	P
β: interaction effect of demands and control ('*strain*')	0.935	(0.402, 1.454)	0.000

The contrasts to baseline are heavily driven by the model assumptions of vanishing main effects.

Statistical models for occurrence of depression

Model with only main effects

		demands	
		low	high
control	high	α	$\alpha + \delta$
	low	$\alpha + \gamma$	$\alpha + \gamma + \delta$

Interaction effect of demands and control ('*strain*') assumed to be zero.

	Estimates	95% Conf. interv.	P
α: intercept	-4.530	(-5.081, -4.045)	
δ: main effect of demands	0.193	(-0.324, 0.718)	0.464
γ: main effect of control	0.885	(0.353, 1.448)	0.001

Strong estimated main effect of control (low control related to increased risk of depression).

Statistical models for occurrence of depression

Model with only main effects

Net contrast to baseline: High control, low demands

		demands	
		low	high
control	high	*Baseline*	0.193
	low	0.885	1.078*

Interaction effect of demands and control ('*strain*') assumed to be zero.

	Estimates	95% Conf. interv.	P
δ: main effect of demands	0.193	(-0.324, 0.718)	0.464
γ: main effect of control	0.885	(0.353, 1.448)	0.001

(*) $1.078 = 0.193 + 0.885$

Statistical models for occurrence of depression

Model with freely varying parameters in the four possible combinations of demands and control

		demands	
		low	high
control	high	α	$\alpha + \mu_2$
	low	$\alpha + \mu_3$	$\alpha + \mu_4$

This is just a reparameterization of the standard model:

$$\mu_2 = \delta,\ \mu_3 = \gamma$$

$$\mu_4 = \gamma + \delta + \beta,\ \textit{i.e.}\ \beta = \mu_4 - \mu_2 - \mu_3$$

Statistical models for occurrence of depression

Model with equal main effects of demands and control as well as interaction between them

		demands	
		low	high
control	high	α	$\alpha + \gamma$
	low	$\alpha + \gamma$	$\alpha + 2\gamma + \beta$

	Estimates	95% Conf. interv.	P
α: intercept	-4.229	(-4.829, -3.725)	
γ: main effect of control = main effect of demand	-0.015	(-0.700, 0.707)	0.966
$2\gamma + \beta$: net contrast to baseline for high demands, low control	0.925	(0.265, 1.633)	0.007

Note: β is still the interaction effect of demands and control ('*strain*')

Strong strain effect, but partly driven by arbitrary assumption of equal effects of demands and control.

Statistical models for occurrence of depression

Model with equal main effects of demands and control as well as interaction between them

Net contrast to baseline: High control, low demands

		demands	
		low	high
control	high	*Baseline*	−0.015
	low	−0.015	0.925

	Estimates	95% Conf. interv.	P
γ: main effect of control = main effect of demand	-0.015	(-0.700, 0.707)	0.966
$2\gamma + \beta$: net contrast to baseline for high demands, low control	0.925	(0.265, 1.633)	0.007

Note: β is still the interaction effect of demands and control ('*strain*')

Statistical models for occurrence of depression

Relations between models

Standard full statistical model:

$\alpha, \gamma, \delta, \beta$

$\beta = 0$

Model with only main effects:

α, γ, δ

$\gamma = \delta$

Model with equal main effects and interaction:

α, γ, β

$\gamma = \delta = 0$

Model IPD (strain):

α, β

Statistical models for occurrence of depression

Standard full statistical model

Minimal confounder control (age, sex, marital status)

Modest confounder control (age, sex, marital status, education)

		demands	
		low	high
control	high	α	$\alpha + \delta$
	low	$\alpha + \gamma$	$\alpha + \gamma + \delta + \beta$

	Estimates	Estimates	Estimates
α: intercept	-4.229	-5.298	-5.378
δ: main effect of demands	-0.504	-0.513	-0.508
γ: main effect of control	0.363	0.378	0.395
β: interaction effect of demands and control ('*strain*')	1.066	1.054	1.046

Small effects of correction for possible confounding, no change in general conclusion: Tendency to interaction effect: the combination of high demands and low control increases probability of clinical depression.

Statistical models for occurrence of depression

Model IPD ('strain')

Minimal confounder control (age, sex, marital status)

Modest confounder control (age, sex, marital status, education)

		demands	
		low	high
control	high	α	α
	low	α	$\alpha + \beta$

Main effects of control and demands are assumed to be zero.

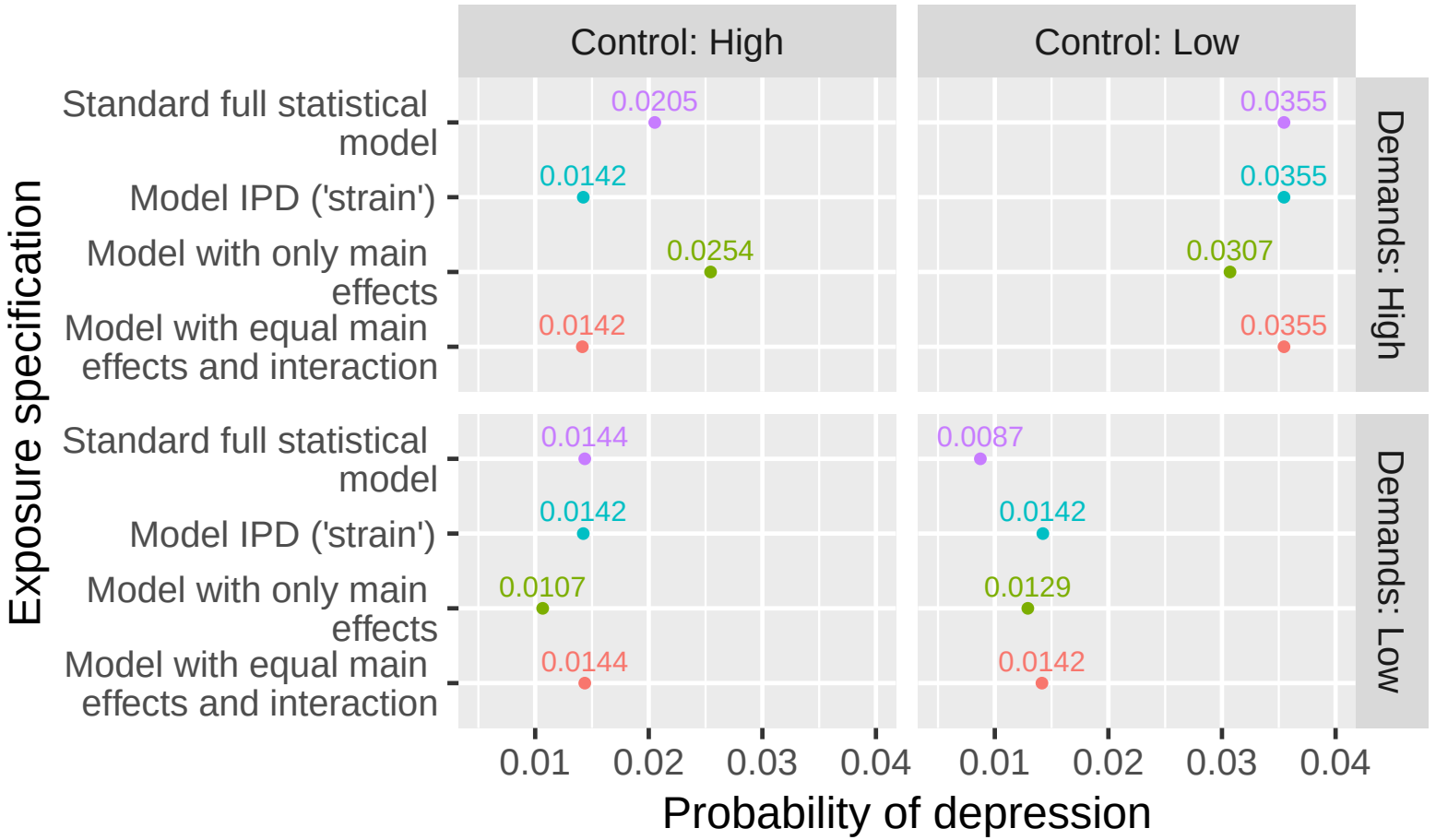
	Estimates	Estimates	Estimates
α: intercept	-4.229	-5.259	-5.230
β: interaction effect of demands and control ('*strain*')	0.935	0.926	0.931

Small effects of correction for possible confounding, no change in general conclusion: Strong 'strain' effect, partly generated by main effects that are not modelled.

An overview of predicted depression probabilites

In models with no confounder adjustment, we estimate

$$P(\text{clinical depression}) = \frac{\exp(\alpha + \dots)}{1 + \exp(\alpha + \dots)}$$



Variable selection

> *"Stepwise variable selection has been a very popular technique for many years, but if this procedure had just been proposed as a statistical method, it would most likely be rejected because it violates every principle of statistical estimation and hypothesis testing."*

F.E. Harrell (2015). *Regression Modeling Strategies*. Second Edition. Springer, New York, p. 67.

There are two issues in model selection: we hope to *minimize bias* and to *minimize variance*. Much variable selection goes too far in focusing on the second target, getting down to very few variables with bias as a result.

Stay with a model that does cover the important structure in the problem and the data and does not violate the hierarchical principle.

Conclusion

As occasional visitors to this interesting area we feel that the standard mainstream statistical approach would be adequate and relevant for these issues.