university of copenhagen

Combined eects of two or more variables, statistical considerations

Niels Keiding & Anne Helby Petersen
Section of Biostatistics

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 Depression: no Total
n = 3035 n = 59 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

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

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

In the beginning, we use the principal exposure variables

;

$$z_1 = z_2 = \begin{cases} 1, \text{ if demands} > \text{median} \\ 0, \text{ if demands} < \text{median} \end{cases}$$

;

$$\begin{cases} 1, \text{ if control} \\ \text{< median} \end{cases}$$

0, if control > median ;

median (’*strain*’) 0, otherwise

$z_3 = z_1 \cdot z_2 =$

1, if demands > median and control <

Later additional covariates z_4, z_5,... 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 – – + ”
low – + “ –+“+”+—

Estimates 95% Conf. interv. P
–: intercept -4.229 (-4.829, -3.725)
”: main eect of demands -0.504 (-1.487, 0.394) 0.285 “: main eect of control 0.363 (-0.404, 1.137) 0.350 —: interaction eect of demands and control (’*strain*’) 1.066 (-0.046, 2.248) 0.066

Tendency to interaction eect: 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*ú*

Estimates 95% Conf. interv. P
”: main eect of demands -0.504 (-1.487, 0.394) 0.285 *“*: main eect of control 0.363 (-0.404, 1.137) 0.350 —: interaction eect 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 – –
low – – + —

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

Estimates 95% Conf. interv. P
–: intercept -4.229 (-4.829, -3.725)

—: interaction eect of demands and control (’*strain*’) 0.935 (0.402, 1.454) 0.000

Strong ’strain’ eect, partly generated by main eects 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 eects 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 eects of control and demands are assumed to be zero.

Estimates 95% Conf. interv. P
—: interaction eect 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 eects.

Statistical models for occurrence of depression *Model with only main eects*

demands
low high
control high – – + ”

low – + “ – + “ + ”

Interaction eect of demands and control (’*strain*’) assumed to be zero.

Estimates 95% Conf. interv. P
–: intercept -4.530 (-5.081, -4.045)
”: main eect of demands 0.193 (-0.324, 0.718) 0.464
“: main eect of control 0.885 (0.353, 1.448) 0.001

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

Statistical models for occurrence of depression *Model with only main eects*

Net contrast to baseline: High control, low demands

demands
low high
control high *Baseline* 0.193
low 0.885 1.078[ú]

Interaction eect of demands and control (’*strain*’) assumed to be zero.

Estimates 95% Conf. interv. P
”: main eect of demands 0.193 (-0.324, 0.718) 0.464
“: main eect 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 – – + μ_2

low – + μ_3 – + μ_4

This is just a reparameterization of the standard model:

μ_2 = ”, μ_3 = “

$\mu_4 =$ " + " + —, *i.e.* — $= \mu_4 \neq \mu_2 \neq \mu_3$

university of copenhagen section of biostatistics

Statistical models for occurrence of depression *Model with equal main eects of demands and control as well as interaction between them*

demands
low high
control high – – + “
low – + “ – + 2“ + —

Estimates 95% Conf. interv. P
–: intercept -4.229 (-4.829, -3.725)
“: main eect of control = main eect of demand -0.015 (-0.700, 0.707) 0.966 2“ + —: net contrast to baseline for high demands, low control 0.925 (0.265, 1.633) 0.007

Note: — is still the interaction eect of demands and control (’*strain*’)

Strong strain eect, but partly driven by

arbitrary assumption of equal eects of demands and control.

Statistical models for occurrence of depression *Model with equal main eects 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
“: main eect of control = main eect of demand -0.015 (-0.700, 0.707) 0.966 2*“* + —:
net contrast to baseline for high demands, low control 0.925 (0.265, 1.633) 0.007

Note: — is still the interaction eect of demands and control (’*strain*’)

Statistical models for occurrence of depression Relations between models

Standard full statistical model:

–, “, ”, —

— = 0 “ = ”

Model with only main eects: –, “, ”
Model with equal main eects and interaction:
–, “, —

Model IPD (strain):
–, —

“ = ” = 0

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 – – + ”
low – + “ –+“+”+—

Estimates Estimates Estimates
–: intercept -4.229 -5.298 -5.378
”: main eect of demands -0.504 -0.513 -0.508
“: main eect of control 0.363 0.378 0.395
—: interaction eect of demands and control (’*strain*’) 1.066 1.054 1.046

Small eects of correction for possible confounding, no change in general conclusion:

Tendency to interaction eect: 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 – –
low – – + —

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

Estimates Estimates Estimates
–: intercept -4.229 -5.259 -5.230
—: interaction eect of demands and control (’*strain*’) 0.935 0.926 0.931

Small eects of correction for possible

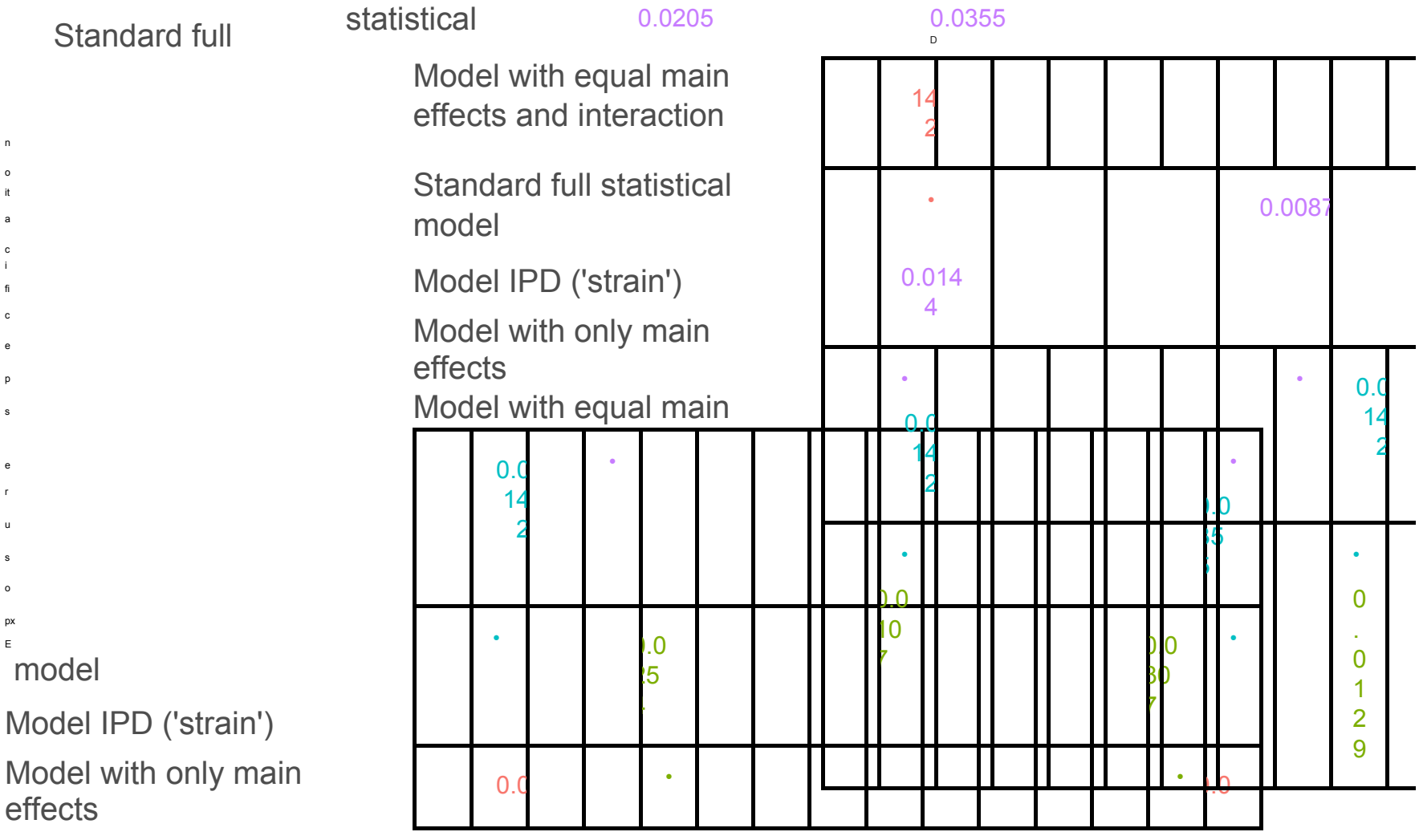
confounding, no change in general conclusion:

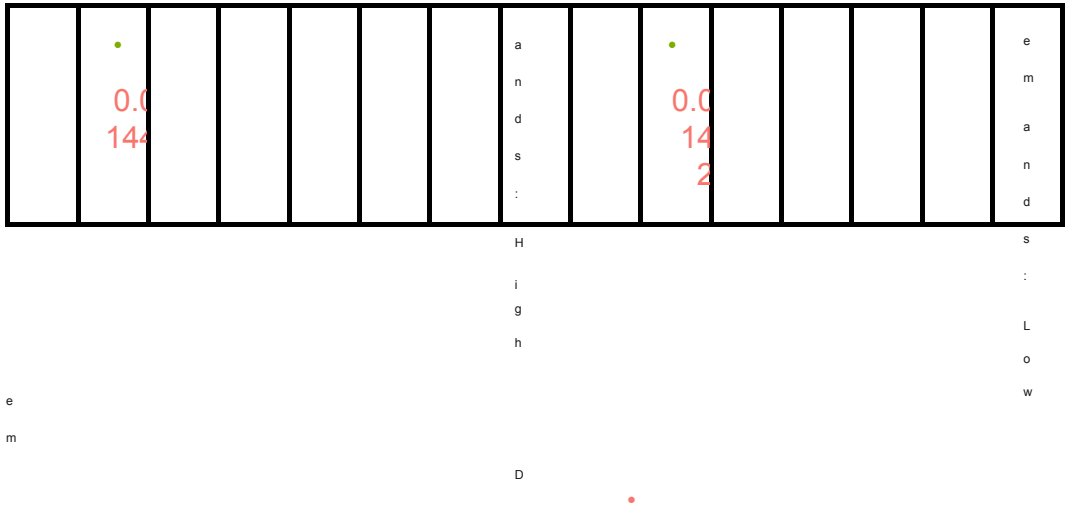
Strong 'strain' eect, partly generated by main eects that are not modelled.

An overview of predicted depression probabilites

In models with no confounder adjustment, we estimate

$$P(\text{clinical depression}) = \frac{\exp(\beta_0 + \beta_1 \cdot \text{Control: High} + \beta_2 \cdot \text{Control: Low})}{1 + \exp(\beta_0 + \beta_1 \cdot \text{Control: High} + \beta_2 \cdot \text{Control: Low})}$$


Standard full
statistical
0.0205
0.0355
D
Model with equal main effects and interaction
Standard full statistical model
Model IPD ('strain')
Model with only main effects
Model with equal main
0.0087
0.014
4
n
o
it
a
c
i
fi
c
e
p
s
e
r
u
s
o
px
E
model
Model IPD ('strain')
Model with only main effects



effects and interaction

0.01 0.02 0.03 0.04 0.01 0.02 0.03 0.04

Probability of depression

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.