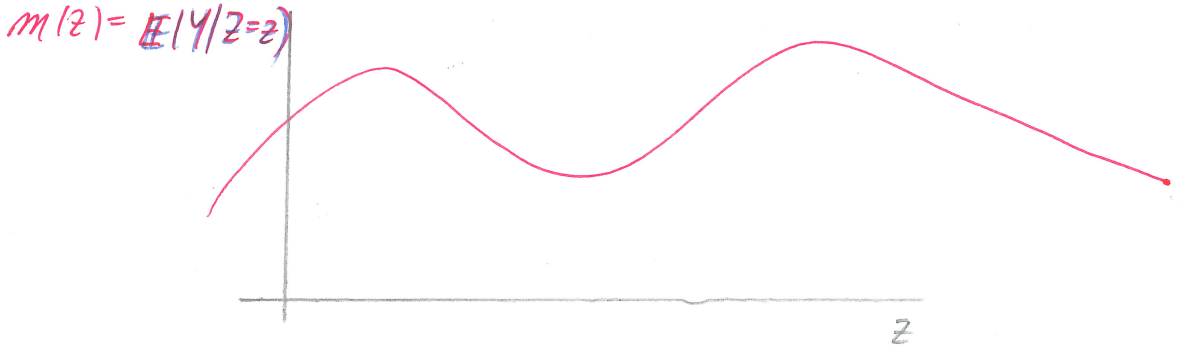
XII. Model Building

Covariate vector $Z = (Z_1, \ldots, Z_p)^T$,
$p > 1$ (usually $p >> 1$)

TASK: Propose a model for $E(Y|Z=z)$, $z \in \mathbb{R}^p$
(= BUILD)

NOT OF INTEREST NOW:

$Z = Z_1 \in \mathbb{R}$ numeric



$m(z)$ = parameterization of a numeric covariate → Section 4.3

12.1 General principles

WITH $p >> 1$

- numeric covariates parameterized mostly by identity ("as they are")
- categorical covariates → (pseudo) contrasts (mostly reference group pseudocontrasts (dummy variables) used)
- HWF models involving at most two-way interactions

→ the most complicated model to consider is

$$\sim 1 + z_1 + \ldots + z_p + z_1 : z_2 + \ldots + z_{p-1} : z_p$$

MAIN PROBLEM:

→ which terms should be kept in the model and which not

- not everything because of multicollinearity and increase of standard errors in general
- "important" terms must stay in the model otherwise bias in estimation of β's and also in prediction

→ see Chapter 10

Primary tool of model building

→ submodel testing (Chapter 8)

→ comparisons smaller model vs. larger model

- submodel test significant

  ≡ terms in the larger model cannot be removed

PROBLEM: • sequence of submodel tests is needed

→ increase in type I error (its probability)

type I error = simple model is rejected even if sufficient

→ "too complex" model is chosen (overfitting)

→ we should be aware of this and interprete the p-values not too dogmatically ...

"Correct" approach to model building

→ depends on many factors:
- nature of the problem
- structure of the data
- questions to be addressed by the analysis
- ...

→ There exist no "correct" universal approach to model building...
→ NO RECIPE can be given...

→ Experience is quite important
→ can only be gain by running (repeatedly) statistical analyzes...

LUCKILY, there exist some typical problems for which a bit more can be said

12.2 Prediction

AIM: Build a model suitable to predict $Y|Z=z$.

- better to consider more complex (rich & flexible) models
  - omitted important predictors (or included with a wrong parameter.)
    → bias in prediction (Section 10.2.3)
  - irrelevant predictors in the model
    → some increase in variability (error) of predictions
    → no bias

→ more fancy parameterizations (splines, ...) of at least the most important predictors might be needed / useful

---

Section 10.2.2: Prediction quality was evaluated using the same dataset as that used to build the model / we compared Y_i with $\hat{Y}_i$

External validity (quality) of prediction: $\hat{Y}_i = X_i^T \hat{\beta}$

- training / validation dataset
- cross-validation

training dataset

= subset of available data $(Y_i, Z_i^T)^T$ (random)

→ build the model → $\hat{\beta}$ (calculated from $i \in$ training)

validation dataset

≡ complement to training dataset

- somehow compare Y_i and $\hat{Y}_i = X_i^T \hat{\beta}$

$i \in$ ~~train~~ validation

---

cross-validation (see also chapter 11)

- comparison of Y_i and $\hat{Y}_{[-i]} = X_i^T \hat{\beta}_{(-i)}$

Section ~~10~~ 10.1: multicollinearity is not really a problem for prediction BUT

multicollinearity → no impact on $\mathrm{var}(\hat{Y}_i | X_i)$

"internal" predict.

but how about $\mathrm{var}(\hat{Y}_{[-i]} | X_i)$

$=$

$X_i^T \hat{\beta}_{(-i)}$

multicollinearity: high $\mathrm{var}(\hat{\beta} | X)$

→ small changes in data may lead to large changes in $\hat{\beta}$

→ perhaps large change $\hat{\beta} \to \hat{\beta}_{(-i)}$ (see also influential obsv.)

→ large variab. in $\hat{Y}_{[-i]}$

Moreover, even if prediction is the primary aim, the analyst usually wants to know which covariates are "significantly" associated with the outcome

→ we still need reasonably low var($\hat{\beta}|X$)

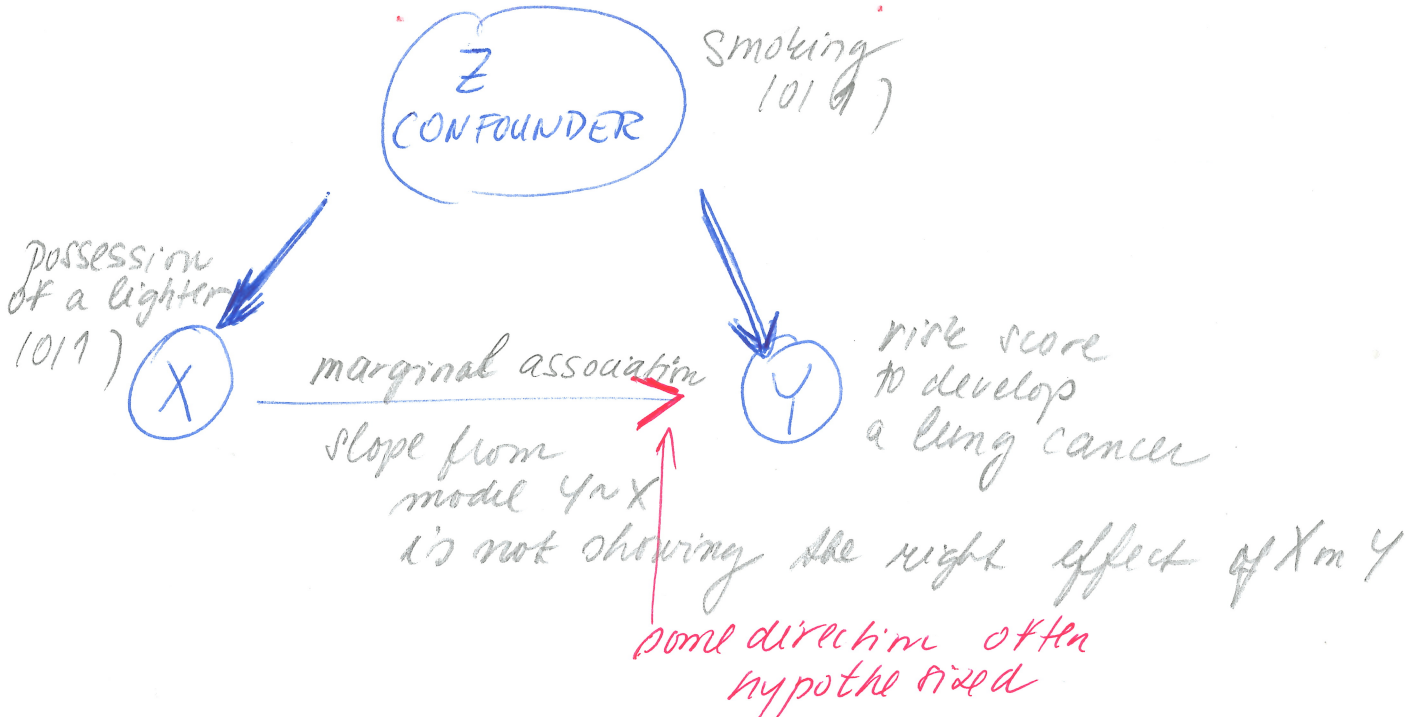
(i.e., no multicollinearity)

12.3 Evaluation of a covariate effect

- one (or few) primary covariate X
- Research question: how (strongly) X affects response mean (is associated with the response mean)

  - multicollinearity is problem
    - must be treated in advance
  - omission of important variables is problem
    → bias in estimation of effect of X (Section 10.2.3)
    → the most important issue here is CONFOUNDING

CONFOUNDING



- to evaluate a true effect of X on Y, all possible confounders should be identified (and then controlled for in the model)

+ all notes : in 12.3

12.4 Model building strategy

Initial model ?

$z_1 + \dots + z_p + \underbrace{z_1 : z_2 + \dots + z_{p-1} : z_p}$

all two-way interactions ?

- usually not good starting point
- easily resid d.f. ≈ 0
  → overfitting

Initial model

$z_1 + \dots + z_p$

→ use a sequence of F-tests to eliminate unimportant terms

→ elimination in blocks
+ checking whether it is not possible / necessary to return any of removed terms back to the model
(anti-multicollinearity measure)

→ see EXAMPLE of model building (additional presentation)

→ if needed, find suitable parameterizations for (important) covariates

Interactions

$z_1 + \dots + z_k$: model that involves "important" covariates in from the first stage

(a) start with $z_1 + \dots + z_k + z_1 : z_2 + \dots + z_{k-1} : z_k$

- try to simplify the model (sequence of F-tests)

(b) $z_1 + \dots + z_k$

→ try to add always the most significant interaction one-by one

---

- only rarely three-way, four-way, ... interactions useful

- unless there is a specific reason for it, keep models HWF

12.5 Conclusion

- There is no such thing as the TRUE model (even though in theory, we always have one...)
- "all models are wrong, but some are useful.

[some are wrong & useless]

The practical question is how wrong do they have to be to not be useful."

George E.P. Box
(1919-2013)