Index of Lecture 2a: Model building I

Practical Information

Today's lecture: follow-up from Lectures 1a−1b, and start of model building,

- interpretation of (more complex) models and parameters,
- more focus on the model equation and its parameters ⇒ more technical material (but important to understand),
- interaction and confounding:
  - recap of discussions in previous courses,
  - new features and recommendations (to be practiced as we go on),
- textbook reading: VER Sections 14.4−7,
- review of Linear Regression Exercise 1 with software demonstrations, as needed.

Other updates:

- homework for Friday: Exercise 2 in "Linear Regression Exercises", same setup as for today's exercise:
  - use your preferred statistical software for the calculations,
  - exercise review with potential Minitab/Stata demo on Friday,
- scheduling of exam: can we move to April 16? (currently April 17).

Predictor Types and Forms

Variable types (review from VHM 801):

- quantitative (continuous, discrete),
- categorical (dichotomous, ordinal, nominal).

Predictor forms (i.e., how predictors enter into models):

- continuous $\sim$ modelled by slope: $\beta \cdot x$, (or other functional form),
- categorical $\sim$ modelled by effects for each category: α_j (for j^{th} category).
- relation to variable types:

	Predictor forms	
Variable types	continuous	categorical
quantitative	$\surd$	$(\surd)$ [a]
dichotomous	$\surd$ [b]	$\surd$
ordinal	$(\surd)$ [c]	$\surd$
nominal	no!	$\surd$

[a] often requires grouping $\sim$ loss of information, when values in a group are not distinguished,

[b] same results as for categorical modelling, when the values are coded as 0 and 1,

[c] implies a non-trivial scale assumption.

Continuous Predictors

Ways to (sometimes) improve interpretation of parameters:

- scaling of x: change the units for x to make the "1 unit change" for the slope more meaningful, e.g.
  - herd sizes are often in the hundreds, so the effect of a 1 unit change can be very small $\Rightarrow$ $x/10$ or $x/100$ may be preferable,
  - if the range of the predictor is very small, a 1 unit change may be too large $\Rightarrow$ $10 \cdot x$ or $100 \cdot x$ may be preferable,
- centring of x: subtract a value from all values of x to make the intercept (corresponding to predictor value 0) more meaningful, e.g.
  - a parity of 0 does not exist $\Rightarrow$ (parity -1) or (parity -3) may be preferable,
  - in dates, a calendar year of 0 is often a wild extrapolation $\Rightarrow$ subtract the first study or a central (mean/median) study year,

  note: centring reduces collinearity, but usually not an important consideration.

Beware: Scaling or centring does not change the model! (the model's fit, inference and validity of assumptions are not affected).

Parametrizing Categorical Predictors (Factors)

Basic problem: too many parameters $\Rightarrow$ cannot meaningfully estimate them all.

Simplest example: 1-way ANOVA, say with factor A and $a(\geq 2)$ groups:

- most natural parameters: the group means μ_j, $j = 1, \ldots, a$,
- can also introduce a mean/intercept μ, so that $\mu_j = \mu + \alpha_j$, $j = 1, \ldots, a$,
  - α_j is the deviation from μ for group j,
  - notation with μ and α_j's extends to multiple factors and slopes,[1]
  - results in $(a+1)$ parameters from a groups — one parameter too many!

Solution for too many parameters: put some restrictions on them, e.g. (again for the 1-way ANOVA example),[2]

- $\alpha_1 = 0$ (μ becomes value of first category, the reference category, Stata/Minitab/R default),
- $\alpha_a = 0$ (μ becomes value of last category, the reference category, SAS default)
- $\alpha_1 + \ldots + \alpha_a = 0$ (μ becomes average value, Minitab/R option).

Beware: Choice of parametrization does not change the model! (the model's fit, inference and validity of assumptions are not affected).

[1] Corresponding to $(\mu + \alpha_j + \beta_l)$ for two factors and to $(\mu + \alpha_j + \beta \cdot x)$ for an added slope.
[2] Section 14.4.3 of VER describes hierarchical restrictions for an ordinal predictor — not part of core curriculum.

Indicator Variables

Fact: A categorical predictor, say A, is represented in the model in its usual multiple regression form, for observation i,

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i,$$

when the corresponding x_j's have a special form: with $\alpha_1 = 0$,[3]

- $x_{ji} = \begin{cases} 1 & \text{if } A_i = j \\ 0 & \text{if } A_i \neq j \end{cases}$, for $j = 2, \ldots, a$
- $x_2, \ldots, x_a$ are called indicator (or "dummy") variables,
- a total of $(a-1)$ variables (none for the reference category).

Therefore, the model output will display (only) the estimates $\hat{\beta}_1, \ldots, \hat{\beta}_a$.

Choice of reference category:

- still does not change the model (fit, inference, assumptions),
- ideally a category for which the comparisons with reference are (most) relevant, because
  - $\beta_j \sim$ difference between j^{th} and reference category,
- avoid a reference category with small sample size (if unbalanced design).

[3] Minitab notation: "(1,0)" coding, as opposed to "(-1,0,1)" coding for $\mu \sim$ average value.

Interaction — Introductory Remarks (Recap)

Interaction (synergism/antagonism, covariation):

- the combined effect of two predictors (often factors, but not necessarily so) on the outcome is not predictable from isolated effects of each of them (on the outcome) when examined separately,
- the effect of one factor depends on the level of the other factor,
- non-parallel lines in interaction plot (i.e., plot of combined means versus one factor), where parallel lines $\sim$ additive effects (and hence no interaction).
- is not the same as collinearity: can occur with both independent and dependent predictors,[4]
- note: interaction is dependent on scale (of outcome), so affected by transformation.

Why explore interactions?:

- biological mechanisms of interest, either in addition to main (separate) effects or as the primary interest (e.g., with time),
- leads to more complex understanding of predictor effects $\Rightarrow$ of potential use/interest,
- improves model fit, but beware of multiple testing issues.

[4] Strong dependence between predictors, e.g. incompleteness of $A \times B$, will make the assessment of interaction difficult and/or little meaningful.

Statistical Modelling of Interactions

Interaction in the model equation:

- additional terms to the main effects (beware to always include main effects),
- specific form depends on types of predictors involved and the chosen parametrization (discussed in the following),
- statistical significance assessed by a suitable F-test.[5]

Interpretation of interaction:

- guided by the interaction plot (strongly recommended),
- quantitative interpretation by estimates and CI — depends on type and form of interaction and often requires additional analysis, e.g. pairwise comparisons.

Simplest example: two categorical predictors (factors, say A and B):

- model equation: $\alpha_j + \beta_l + \gamma_{jl}$, where $j = 1, \ldots, a \sim A$ and $l = 1, \ldots, b \sim B$,
- restrictions on (γ_{jl}) needed — with 1 as the reference category for both A and B:

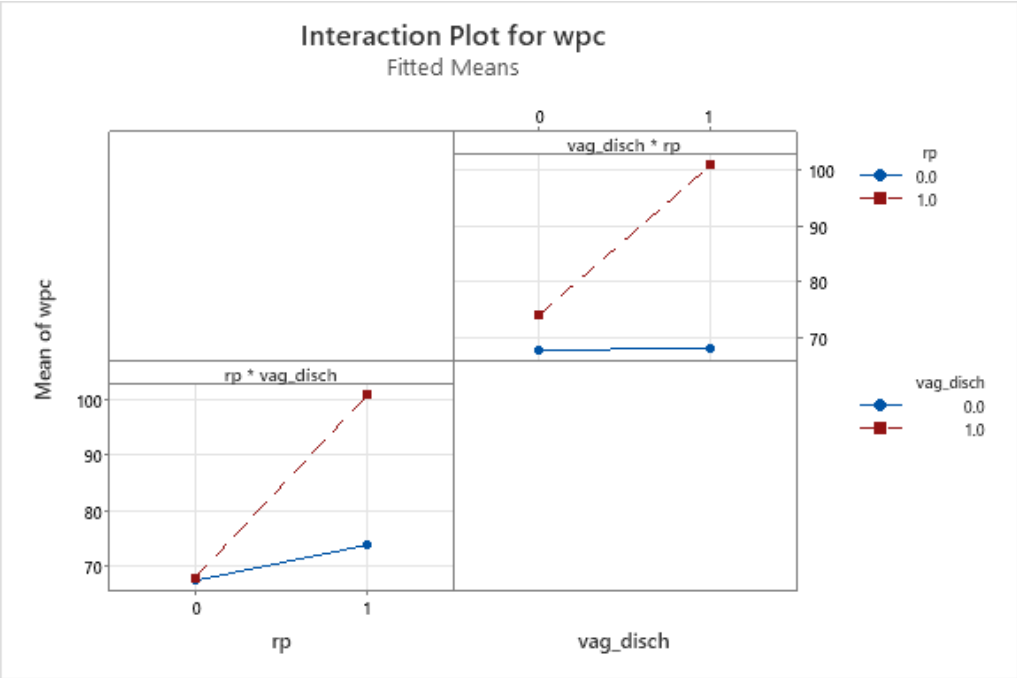
$$\gamma_{11}, \gamma_{21}, \ldots, \gamma_{a1} = 0 \quad \text{and} \quad \gamma_{11}, \gamma_{12}, \ldots, \gamma_{1b} = 0.$$

[5] The numerator degrees of freedom should equal the product of the two main effects, i.e. $\mathrm{DF}_{A*B} = \mathrm{DF}_A \cdot \mathrm{DF}_B$; caution advised if not the case!

Interaction Example (VER 14.9)

Model summary: — **outcome**: `wpc` (time from wait period to conception),
— **model terms**: `rp`, `vag_disch`, and their interaction ($P = 0.039$).

Interaction plot:



Parameter estimates:

- "constant" (67.67): `wpc` for `rp=0` and `vag_disc=0`,
- "`rp 1`" (6.34): `rp` difference (1−0) in `wpc` for `vag_disch=0`,
- "`vag_disch 1`" (0.54): `vag_disch` difference (1−0) in `wpc` for `rp=0`,
- "`(vag_disch,rp)(1,1)`" (26.35): added `rp` difference (1−0) in `wpc` at `vag_disch=1`.[6]

[6] Also: added `vag_disch` difference (1−0) in `wpc` at `rp=1`.

Interactions Involving Continuous Predictors

Essential change (compared to factors only):
interactions involve differences in slopes (instead of differences in means),

- interpretation is often more difficult[7] and technical,
- biologically, non-parallel lines ($\sim$ interaction) is often more natural than parallel lines ($\sim$ no interaction)!
- interaction plots are possible, but not always as easily accessible in software.[8]

Example: one categorical predictor (say x) and one factor (say A):

- model equation: $\alpha_j + \beta \cdot x + \gamma_j \cdot x$, where $j = 1, \ldots, a \sim A$,
- restrictions on (α_j) and (γ_j) needed — with 1 as the reference category for A: $\alpha_1 = 0$ and $\gamma_1 = 0$,
- interpretation: regression line for $A = j$: ("constant" $+ \alpha_j$) $+ (\beta + \gamma_j) \cdot x$,
  - slopes: β for $A = 1$ and $(\beta + \gamma_j)$ for $A = j$,
  - intercepts: "constant" for $A = 1$ and ("constant" $+ \alpha_j$) for $A = j$.
- VER Example 14.10 with $x =$ `milk120` and $A =$ `dyst`; also Example 28.12 in PSLS.

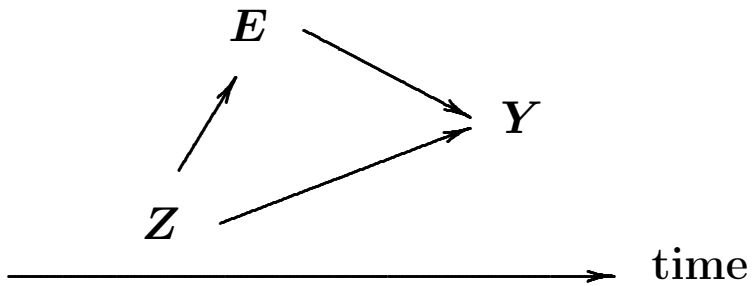
[7] Interaction between two continuous predictors is difficult to visualize and interpret (see VER Example 14.11), and therefore often not attractive in practice.

[8] The most general approach is to construct the desired plot(s) from suitably fitted values from the model.

Confounding — Introductory Remarks (Recap)

Basic notation/terminology and causal diagram (VER 13.5):

- outcome Y (of any type, but continuous for now),
- exposure E (of any type),
- extraneous factor of interest Z (measured or unmeasured; of any type),
- confounder (or lurking variable) Z: extraneous factor that exerts confounding of the relation $E \to Y$.[9]



3 necessary conditions for Z to confound the relation $E \to Y$:

1) Z must be a risk factor for Y; more precisely:
   - ∗ at the reference level of E, i.e. within "exposure-negative subjects" (because the risk must not be caused by a link with E),
2) Z must be associated with E in the source population,[10]
3) Z must not be affected by E (which would make Z an intermediate (mediating) variable[11] between E and Y), and Z must not be an effect of Y.

[9] Note that a confounder is always tied to both outcome and exposure.
[10] See VER 13.5 for specific conditions in cohort and case-control studies.
[11] The VER textbook uses the (non-standard) term intervening variable.

Definition of Confounding

Definition of confounding:

- mathematically not easy; best attempt uses counterfactual arguments, but often infeasible in practice to realistically assume that all its conditions are met,
- literature agrees on necesssary (but not sufficient) conditions for confounding.

Pragmatic solution: define confounding by z for the relation $x \rightarrow y$ as present when both of the conditions (i)$-$(ii) below are met:

(i) z meets the 3 necessary conditions, 1)$-$3) on previous page, to be a confounder,

(ii) the difference between a crude ("total") measure of association/effect[12] and a confounding-adjusted measure of association/effect[13] is "substantial", i.e.[14]

  * a bias above $20-30\%$ (arbitrary cut-off set in VER) measured relative to the crude estimate.

Illustration: confounding by herds (z) for "effect" of `vag_disch` (x) on `wpc` (y):

- in a causal diagram, herds are "before" exposure and outcome — condition 3),
- the associations $z \rightarrow x$ and $z \rightarrow y$ are significant — conditions 1)$-$2),
- the relative change in the `vag_disch` estimate is: $|12.0-17.8|/|12.0| = 0.48 \sim 48\%$.

[12] In regression models: a model with x as predictor for y, but z not included.
[13] In regression models: a model with both x and z as predictors for y.
[14] Rothman *al* (2008), p. 262, note that usually 50% would be considered substantial, and 5% would not...

Multivariable Model Example (VER 14.12)

- outcome of interest: time from wait period to calving (`wpc`) in `daisy2red` dataset,
- exposures of interest (from study objective focusing on diseases/conditions related to calving): `rp`, `vag_disch`, `dyst`,
- findings of causal diagram (VER Figure 14.4):
  - potential confounders: `herd_size`, `parity`, autumn calving (`aut_calv`), `twin`,
  - intermediate variables: `milk120`, calving to first service (`cf`),
- special effects: quadratic term for `herd_size`, interaction `rp*vag_disch`.

Interpretation of model's "effects" ...

`herd_size`	$-36.06 \cdot x + 11.14 \cdot x^2$ (predictor: `herd_size/100`)
`parity`	$1.14 \cdot x$ predictor: (`parity-1`)
`aut_calv`	-8.26 for autumn calving
`twin`	20.68 for twin calving
`dyst`	11.70 for dystocia
`rp`	5.99 for retained placenta
`vag_disch`	1.23 for vaginal discharge
`rp*vag_disch`	22.86 for retained placenta and vaginal discharge
`_cons`	84.66

Additional Notes on Interpretation and Presentation

Categorical predictors:

- assess significance by overall F-test, not individual coefficients,
- focus interpretation on all pairwise comparisons (possibly adjusting for multiple comparisons), not only those with the reference category.

Interactions:

- in presence of an important interaction, the main effects are often of limited interest,
- when a reference category parametrization is used, the coefficients for each predictor in an interaction are not the main effects.

Confounding:

- implies collinearity between x and z, and is often helpful in understanding how to deal with collinearity,
- common practice to include potential confounders in a model, even if insignificant,
- if z is confounder for x, then x is typically intermediate for z,
- confounding and interaction are different things;
  if they exist together, confounding may be explored for each factor level,
- to say z is a confounder without mentioning the exposure, is quite meaningless.

Stata Factor Notation Basics

For predictors x, z and an outcome y, Stata uses the notation:[15]

- `i.x` ∼ categorical effect of x (x must be integer),
- `c.x` ∼ continuous (slope) effect of x (x must be numerical),
- the default depends on the command:
  - `regress y x` ∼ `regress y c.x` (i.e., default is `c.x`),
  - `anova y x` ∼ `anova y i.x` (i.e., default is `i.x`).

Combined effects:

- `x#z` ∼ interaction $x \times z$; in all commands, the default is `x#z` ∼ `i.x#i.z`,
- `c.x#c.z` ∼ multiplication $x \cdot z$,
- always: `x##z` ∼ `x z x#z`
- quadratic regression: `c.x##c.x` ∼ continuous terms x and x^2.

Factor terms can (must) be used in tests, e.g. the Stata commands,

- `testparm i.x`
- `testparm c.x##c.x`

[15] In the Stata help files, the information is listed under "fvvarlist".