Index of Lecture 15: Course wrap-up

Practical Information

Major news:

- projects with comments and mark to be returned to you in Monday's session,
- course Syllabus and updated Instructions for home assignments and exam posted to webpage,
- request for questions and review topics for the last session (15−P),
- no formal course evaluation (your decision!), but I'd be happy to discuss or receive comments per e-mail (maybe after the exam, if you prefer).

Today's lecture:

- Exam: Monday 22/4, 9am-12pm, AVC 287N,
  - exam practical remarks,
  - exam questions (types, calculations),
- a few review slides: new or partly new slides, 15L−4/5/6/7/8,
- review of some questions, suggested choice:
  1. Final exam 2015, Question 1[1],
  3. Multivariate exercise 13 (expanded).

[1] Question 2 of the 2015 exam is not relevant for this year's course, and Question 3 is for regression only.

Exam Practical Remarks

Two versions of the exam:

- "reduced" (Persia, Sitang): 10am−12pm, 2 questions,
- "full" (Emily): 9am−12pm, 3 questions.

Exam rules: open book (all aids are allowed), except a computer-like device, although cell phones and tablets can be used (in flight mode):

* as a calculator (with basic calculation functions, no statistical software),
* to access course notes and electronic textbook material.

Some hints and advices: (to use or not...)

- the questions have equal weight, unless stated otherwise − use your time sensibly!
- layout — essential requirements: readability, and a clear distinction between what is *in* your answer and what is not — don't write first a draft and then a final version,
- conclusions should be part of all analyses,
- statistical model should be part of all data analysis,
- errors: if you realize an error and do not have time to correct it: write what is wrong, what should have been done and how the error would affect the result,
- the exam finishes at 12pm (no extensions of time).

Exam Questions

The plan is for two questions in the "reduced exam", and one extra question on regression (linear and/or logistic) in the "full exam", in both cases with all questions having equal weight (totalling 30%).

Typical elements of exam questions:

- choose between different analyses with results provided (Stata listings per default),
- interpret the results of provided listings,
- supplement the provided listings with extra calculations for specific purposes (e.g., statistical inference),
- outline/sketch analyses based on description of data, specifically:
  - *i)* how the suggested analysis would be done (indicate relevant menus or relevant Stata commands),
  - *ii)* how you would use and interpret the results,
- planning of (simple) experimental designs,
- multiple choice (one or several correct answers).

Suggestions for Your Review

Check (naturally): the course syllabus and the Instructions for home assignments and exam at the website.

Suggested exercises to review:

- exams 2013−2016, 2019 and 2021 (no exam in 2023!),
  - not all questions will apply (e.g. no repeated measures or split-plot models),[2]
  - always the regression question as the third question (when included),
- all home assignments and their solutions (in particular, my comments for your answers),
- all regular exercises listed for the lab sessions,
- all VHM 812 exercises (i.e., VER XX), except for VER 22,
- perhaps also extra exercises listed for the lab sessions,

Review requests for Session 15P are, as earlier mentioned, very welcome.

[2] The listing of exam questions for the lab sessions will include all relevant questions ⇒ those not listed can be skipped!

Choice of (Univariate) Statistical Model

Some useful questions to ask about the data:

- purpose of study?
- response or explanatory variable?
- continuous or discrete/categorical variable?
- particular data structures or experimental designs? – e.g.
  - cross-over design,
  - hierarchical structure,
- random (instead of fixed) effects?
- blocks: do the data include variable(s) of blocking type? (division of experimental units into homogeneous groups, with no intrinsic interest) – or obvious blocking schemes? (Latin square, BIBD etc.), versus "pure" replication,
- interactions between variables? (quantitative or categorical)
- continuous variable (explanatory or response) to be used for prediction of another variable? (regression)
- transformation? (to achieve normal distribution for residuals, homogeneity of variance, linear relation).

Choice of Multivariate Method

Most important consideration = objective of analysis,

- analysis without a clearly stated objective will not count fully (and may be misunderstood),
- suggested to review Manly's datasets, and their use for different objectives.

Features of the data to look for:

- do the data include a grouping variable of interest?,
- what are the units on which multiple measures are taken?
  - are there hypotheses of interest related to those units?,
  - do the units have a relevant structure (e.g., hierarchical or factorial)?,
- are the multiple measures to be considered as outcomes or predictors?
- can distances of interest be defined?, and if so, should variables be standardized for these to be more meaningful?,
- do the data or the problem suggest certain graphical displays to be of particular interest?,
- the usual categorizations of and descriptors for individual variables, perhaps involving the distributional assumptions that can reasonable be made.

Basic Overview of Model Types

Model type[3]	Characteristic	Topics for analysis
basic (VHM 801)	single outcome and explanatory variable	4-step approach for CI and test, ANOVA table, F-statistics, transformation e.g. Box-Cox
multiple linear regression	quantitative explanatory variables	residuals, diagnostics, outlier test, collinearity, test reduced/full model, variable selection
ANOVA models, (general) linear models	categorical explanatory variables ("factors")	replications, blocks, interactions, contrasts, dummy variables, multiple comparisons, margins and least squares means, designs: Latin square, BIBD, cross-over
random effects models	right hand side random variables (in addition to ε)	variance components, extra residuals, more complex SEs, likelihood-based analysis
multivariate analysis	multiple measures on same "subject"	wide range of techniques with different aims: dimension-reduction, classification, relations between subjects and/or variables, distances

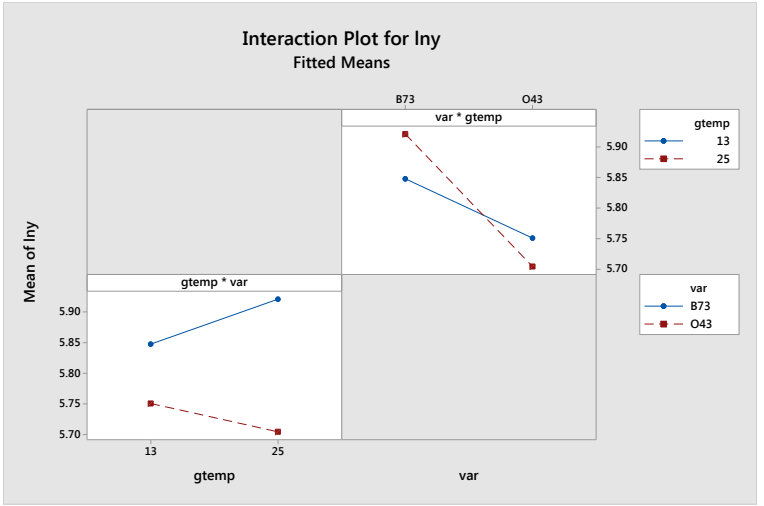
[3] Models for continuous outcomes with normal distribution errors.

Tips for Presenting Interactions

- most helpful tool: interaction plot (2-factor interaction),
- focus on combined factor for estimates (least squares means), confidence intervals and pairwise comparisons, for balanced designs maybe using LSD statistics,
- adjustment for multiple testing may be restricted to subset of all comparisons within combined factor; one major example:
  - within levels of factor A: compare factor B, and vice versa,
- what about "uninteresting" interactions?
  - if non-significant: disregard/drop, and look at main effects[4].

Example: amylase activity in maize
(GO Example 8.10, lecture 6)

var=B73		var=O43		
gtemp=13	gtemp=25	gtemp=13	gtemp=25	SE
5.85	5.92	5.75	5.70	0.015



[4] Beware that the parameter estimates may not represent the main effects when an interaction is present.