Index of Lecture 3b: Model building II

Practical Information

Today's session:

- catch-up from previous lectures,
- Linear Regression Exercise 14.3 discussion,
- new material on (linear) model building:[1]
  - new tools:
    - exploring and modelling linearity (categorization, fractional polynomials),
    - reliability to assess internal validity of estimates,
    - measures of model fit (that can be used for variable selection),
    - automated variable selection procedures,
  - principles for presentation of categorical and continuous predictor effects,
  - principles for variable selection.
- textbook reading: VER Chapter 15[2], and also take a look at VER Chapter 30.

News/Schedule:

- homework for Wednesday: VER 15 model building exercise, same setup as before:
  - prepare for discussion and review upon request (Stata or Minitab),
  - first (real!) home assignment due next Friday (31/1).

[1] Many of the procedures and principles apply to logistic (and generalised linear) model building as well.
[2] Omitted parts: 15.4.3 (Example 15.1), 15.4.5 (Example 15.2), 15.5 (see VHM 801, 12L−18).

Exploring Linearity for Continuous Predictors

Methods already discussed:

- graphical tools: scatterplots, possibly with lowess, of outcome or residuals against x,
- model expansions: added quadratic (polynomial) terms, goodness-of-fit test against categorical predictor model (if x has replication).

Categorisation to explore linearity:

- divide x's range into small number of intervals $\rightarrow$ groups for categorical modelling,
- roughly assess shape from categorical predictor estimates,
- `lintrend`, a Stata add-on command, automates the process.

Alternative functional forms, within the linear model class[3]:

- transformation of x: Box-Cox analysis for x (limited to simple regression),
- splines: piecewise smooth polynomials based on chosen points ("knots") where the pieces are adjoined (with specified smoothness),[4]
- fractional polynomials (next slide): a flexible class of curves created from a few polynomial-type terms.

[3] Also possible to have non-linear relations that are not linear models, e.g. the logistic growth curve over time t: $y = \alpha(1 + \exp[-(t - \beta)/\gamma)])^{-1}$.

[4] Stata's `makespline` command offers a variety of spline implementations (advanced techniques).

Fractional Polynomials[5]

Idea: combine polynomial regression (multiple terms of integer powers) with non-integer powers from Box-Cox analysis to create functions of the form:

$$f(x) = \beta_0 + \beta_1 \cdot x^{p_1} + \beta_2 \cdot x^{p_2},$$

- more power terms can be included, but two terms (dimension 2) often suffice,
- the powers p_1 and p_2 are determined as giving optimal fit among a restricted class of powers considered:
  - common set of powers: {-2, -1, $-\frac{1}{2}$, 0, $\frac{1}{2}$, 1, 2, 3}, where power 0 $\sim$ log-transform and power $< 0 \sim$ inverse terms (as for Box-Cox),
  - repeated powers are allowed, with the following meaning: powers (p p) $\sim$ terms x^p and $x^p \ln(x)$.
- (technical) significance testing accounts for determining the dimension by assigning two degrees of freedom when the model dimension is estimated,
- (technical) in practice, the predictor x often needs to be rescaled to yield interesting shapes; this is handled by the software implementation,
- interpretation of β coefficients: difficult, in particular when x is being rescaled,
  - focus on graphical representation of resulting curve,
  - software implementation (Stata) offers confidence bands.

[5] Royston & Sauerbrei (2008): *Multivariable Model-building*, is the authoritative reference.

Fractional Polynomials Example (VER 15.4)

Illustration: fractional polynomials of dimension 2 for `milk120k` as predictor for `cf`:

```
. fp <milk120k>: regress cf <milk120k>
Fractional polynomial comparisons:
------------------------------------------------------------------------------
             | Test             Residual   Deviance
    milk120k |   df   Deviance   std. dev.      diff.       P   Powers
-------------+----------------------------------------------------------------
     omitted |    4   73543.99     28.342     45.690    0.000
      linear |    3   73543.04     28.342     44.738    0.000   1
       m = 1 |    2   73525.42     28.310     27.118    0.000   -2
       m = 2 |    0   73498.30     28.262      0.000       --   -.5 0
------------------------------------------------------------------------------
```

- deviance $\sim$ model fit (lower values are better[6]),
- tests for model comparisons are based on differences in deviance $\sim$ chi-square distributions ("Test df"),
- the *P*-values in the table are for comparisons with the last model.
- conclusions: the dimension 2 model much improves the fit from all other models; the dimension 1 model ($p = -2$) much improves the fit of the linear model.

[6] The deviance is proportional to ln(SSE) and to ln(MSE), except for constants.

Reliability[7]

Objective: quantify how well the model will perform for future predictions.

Why not use R^2?: R^2 does represent predictive ability, but is estimated on the data itself $\Rightarrow$ too high value for predictive ability beyond the data.

Idea: separate estimation and assessment of predictive ability onto different data subsets:

- split-sample approach: divide data into two parts, build model on first part and estimate R^2 on second part,[8]

  $+$ simple to do, generalizes to many other models/settings,

  $-$ requires a large dataset, introduces an arbitrariness (how to split?),

- cross-validation: repeatedly split sample by omitting a small part for estimation, e.g., leave-one-out cross-validation with just one observation left out at a time,

- in both approaches, the shrinkage on cross-validation (the drop in R^2 from full data analysis to R^2 based on split samples) indicates the robustness of predictions.

Special case for linear models: possible to calculate analytically (without doing split-sample analyses) statistics for leave-one-out crossvalidation,

- " PRESS statistic " = predicted residual sum of squares ($= \sum_i [\frac{\hat{\varepsilon}_i}{1-h_i}]^2$, h_i = leverage),
- " predictive R^2 " = R^2 from PRESS statistic.

[7] The term reliability has many other uses, and its use in this context is specific to VER.

[8] The R^2 value is the squared correlation between predictions from the part 1 model with the observations in part 2.

Measures of Model Fit

Main drawback of R^2: can never go down when predictors are added $\Rightarrow$

- useless as criterion for model choice ("model selection"),
- does not take account desirability of simpler models over more complex ones.

Alternatives that attempt to circumvent this problem:

- adjusted $R^2 = 1 - (\text{MSE}/\text{MST})$, can decrease for noise predictors (impact of df),
- predictive R^2 (see previous slide), can decrease with less reliable predictors,
- information criteria of the general form: "fit" + "penalty" for added parameters, where lower value is better:
  - AIC (Akaike's information criterion) $= n \ln(\text{SSE}/n) + 2k$,[9]
  - BIC (Bayesian information criterion) $= n \ln(\text{SSE}/n) + k \cdot \ln(n)$,[9]
  - Mallow's $C_p = \text{SSE}/\hat{\sigma}^2_{\text{full}} + 2k - n$, where $\hat{\sigma}^2_{\text{full}} \sim$ model with all predictors,

  where $n =$ # observations, and $k =$ # parameters (including intercept).

Coleman	Model	SSE	k	R^2	Adj.R^2	Pred.R^2	AIC	BIC	C_p
data:	$x_1 - x_5$	60.24	6	90.6%	87.3%	81.7%	90.81	96.78	6.0
model fit	x_3, x_4	72.43	3	88.7%	87.4%	84.3%	88.49	91.48	2.8

[9] Both AIC and BIC additionally include the constant $(2.8379\,n)$; also, several other versions of AIC exist.

Presentation of Continuous Predictor Effects

Methods already discussed:

- scaling predictor values to get coefficients on a sensible scale,
- centring predictor values to make the intercept more interesting,
- illustrating predictor effects graphically, possibly using the `margins` command.

Extra consideration: to facilitate the comparison of effects between different predictors, in terms of the magnitude of their effect — problem is their different scales.

Possible approches:

- use standardised coefficients $\sim$ change of one sd in distribution of the predictor[10],[11]
- use interquartile range (IQR) of predictor to assess effect of a change from the 25% to the 75% percentile.[11]

The IQR method has the advantage of being robust to outliers of the predictors and to also apply to dichotomous and ordinal categorical predictors.

[10] To also standardize by the sd of the outcome (i.e., to consider $\hat{\beta} \cdot \sigma_x/\sigma_y$), in order to compare effects between studies, is not recommended.

[11] This can be by multiplying the coefficient by the magnitude of interest (sd, IQR), or by rescaling the predictor values by dividing with the magnitude of interest.

Presentation of Categorical Predictor Effects

Things to consider in linear/logistic models when presenting results for a categorical (> 2 categories) predictor:

- unless pre-determined hypotheses exist, the overall P-value (for H_0 : equal effects across all categories) should be presented and used for discussion,
- for significant or "interesting" predictors, estimates with SE or CI should be shown in one of two possible layouts:
  - differences to a "reference" (or "baseline") category: the typical format of software packages when a particular parametrization has been selected (e.g., `regress` and `logit`),
  - estimates for all categories with suitably defined (and reported) averaging across or values set for all other predictors,
- estimates should be backtransformed as needed (e.g., values may be given both as coefficients on logit scale and as ORs),
- unless pre-determined hypotheses exist, pairwise comparisons should be conducted between all category pairs[12],
  - adjusted for multiple comparisons if many categories are present and strict overall significance is required, otherwise unadjusted,
  - many methods for adjustment for multiple comparisons exist[13]; the simplest and most flexible is the Bonferroni method: divide the significance level (0.05) by the number of comparisons (m) or multiply each P-value by m,
- results of pairwise comparisons may be reported (P-values) or indicated by letter coding[14].

[12] Not only for comparisons with a reference category.
[13] To be discussed in detail in Session 6 of VHM 802.
[14] Most common system: two categories with the same letter indicated are not significantly different.

Model Building (VER)

General remarks on model building:

- two main goals of model building: prediction or quantification of relationships/effects,
- need to balance parsimony with model fit (depending on goal),
- need to include biological understanding as much as possible,
- need to consider causal roles of predictors (depending on goal),

⇒ recommended steps of model building:

1) specify maximal model: maximal set of predictors (+ outcome and outcome scale), possibly after excluding predictors of limited value[15] from a larger set of variables,
2) specify criteria[16] for excluding/retaining predictors in the maximal set, possibly involving an initial statistical screening based on univariate (VER: unconditional) associations between predictor and outcome,[17]
3) in a stepwise manner, manually[18] reduce the model until no further reductions are of interest.

[15] Examples of features that reduce the value of variables: missing values, limited variability, concerns about validity.
[16] Criteria can be related to the study objective, the hypothesized causal structure (causal diagram!), the interest in exploring interactions, and the model's fit to the data (guided by, among other things, statistical testing).
[17] Such a statistical screening step may be necessary when the number of variables is too large to be manageable; it is recommended to use a liberal (e.g. $P < 0.20$) significance level to avoid excluding strong predictors in this step.
[18] Caution! against using any automated procedures (slide 3bL−13) for model building.

Model Building — Developments since 2000

The "reproducibility crisis": concern in the scientific community that established (published) findings often cannot not be reproduced (by other studies)[19]; contributing reasons:

- publication bias (tendency of only the significant findings becoming published),
- analyses biased towards significant results (as a natural part of the model building process, or in extreme versions where all options are tried to achieve significance),
- misinterpretation of significance testing outcomes[20]:

  multiple testing and testing of hypotheses suggested by the data, both tend to produce "evidence" that appears much stronger than it really is.

Implications: increased emphasis on conduct and reporting of analyses, and in particular on P-values:

- one journal has banned all P-values and assessment of evidence by confidence intervals — the problem is, what to do instead?
- the New England Journal of Medicine is requiring studies to have a protocol and a statistical analysis plan (SAP), and restricts the use of P-values to confirmatory analysis for which the SAP details how P-values are controlled for multiplicity,
- the ASA officially recommends to avoid using the term "statistically significant".

[19] Some resources for the discussion can be found at the media page, in lecture 12 of VHM 801 and in Section 2.2 of the second edition of the GO textbook.

[20] The Greenland et al. (2016) paper on misinterpretations is (still) much recommended, see media page.

Strobe-Vet Reporting Guidelines

Reporting guidelines[21] are posted at the Equator Network (link at media page), e.g.:

- Strobe (Strengthening the Reporting of Observational Studies in Epidemiology) and its veterinary version Strobe-Vet,
- Sampl (Statistical Analyses and Methods in the Published Literature).

As a way of managing multiplicity of analyses and testing, it is recommended to:

- distinguish between primary and secondary objectives, outcomes and hypotheses,
- state explicitly any primary or secondary prespecified hypothesis or their absence,
- "when multiple outcomes are observed, provide the reader with a rationale for the outcomes presented in the abstract, for example, only statistically significant results or the outcome of the primary hypothesis is presented".

Other Strobe-Vet items related to model building:

7) Variables and 8) Data sources,

9) Bias (specifically confounding, selection and information bias),

11) Quantitative variables (including their potential grouping),

12) Statistical methods — of course!

[21] Documents typically include a checklist and an accompanying document with detailed explanations and examples for the different items.

Model Building . . . Going Forward

Caution: Statements on this slide reflect Henrik's views, and not everyone will agree.

Fact: the VER approach does not guarantee that the resulting P-values are realistic.[22]

It does not solve anything to eliminate P-values[23], but we should (much) reduce the focus on P-values in the presentation.

It is helpful to distinguish between primary and secondary objectives, and between confirmatory and exploratory analyses.

Primary, confirmatory analyses should:

- have a pre-decided, specific analysis plan (with decisions about predictors being made in advance),
- have a number of pre-decided hypotheses of interest (possibly only one),
- be based on a causal diagram (for observational studies).

For all other analyses, wording about significance and P-values should be restricted to a minimum, and any findings reported should be acknowledged to have resulted from exploratory analysis in the Study Limitations.

[22] The VER approach makes the process transparent and less subjective, but does not control P-values for multiplicity.
[23] Nor do data science and machine learning disciplines solve the problem, but they pay less attention to it. . .

Automated Variable Selection

Synthesis: VER recommends against automated variable selection for primary analysis, but notes its potential for quick data exploration.

Two main approaches:

- search through a (possibly large) pool of candidate models and select the one with best model fit (slide 3bL−6),[24]
- in a stepwise manner, successively either expand or reduce (by adding or removing predictors, resp.) a starting model until no improvements are possible, according to model comparisons (typically significance tests at some significance level α),
  - forward selection: start from a null model, add predictors as long as they are significant (i.e., $P < \alpha$) add-ons to current model,
  - backward elimination: start from maximal model, remove predictors with highest P-values $> \alpha$ until none exist in current model,
  - stepwise method: allows for opposite steps for either forward or backwards method; e.g., with forward selection, allow predictors with $P > \alpha$ to be removed.

  note: the stepwise method is most flexible and generally preferred.[25]

[24] Minitab's Best Subsets menu displays the models with highest R^2 for each model size, across all desired model sizes; similar tools exist as Stata add-on commands and in R libraries.

[25] Most statistical software packages have built-in routines for forward, backward and/or stepwise procedures.

Recommendations for Model building

Assorted list of comments related to model building:

- in the maximal model, beware of any error message or strange looking estimates (e.g., missing values, missing estimates or "crazy" estimates or SEs) — and explore the reasons (and adjust accordingly) before continuing analysis,
- model validation/checking should always be for the maximal model (i.e., before eliminating variables based on lack of impact); it can be repeated for the final model (but no new findings are expected),
- assess all predictors included in the maximal model for any issues:
  - continuous: skewness and extreme values (note: normality not required!),
  - categorical: sparse categories (combine categories as needed),
- for model comparisons between nested models, statistical tests are preferred,
- interactions can be crucial for study objectives, but undiscriminatory testing for (many) interactions is at best exploratory,
- in a manual model building, monitor in each step changes in number of observations included, as well as in the estimates, standard errors and significance for each model term; investigate any anomalies occurring,
- proper reporting of epidemiological/statistical methods should never be reduced to copy and paste of text from other sources (e.g., papers).