Multiple linear regression II: Modelling strategies and methods

Luïc Damian

Rheinland-Pfälzische
Technische Universität
Kaiserslautern
Landau

Explain and apply strategies to identify the bestfit model

- How does the research goal influence the selection of the best-fit model?
- Which goodness of fit measures can be used to compare models?
- List the model selection strategies and criticise step-wise model selection.
- Categorise and evaluate approaches to improve and replace stepwise model selection.

Explanation

Aim: Identify most important explanatory variables for diversity of marine ostracods.

→ For explanation search for most parsimonious model

"It is futile to do with more things that which can be done with fewer"

Prediction

The full model (including all possible predictors) typically provides meaningful *p*-values, confidence intervals and parameter estimates and has the highest predictive power (Harrell 2015: 70, 95ff; Heinze & Dunkler 2017). Thus, model parsimony is primarily relevant when we aim to identify the most important variables. Notwithstanding, when building models for prediction, we also prefer the model with fewer variables to one with more variables for a similar predictive power. See also Matloff (2017): 339ff.

RPTU Rheinland-Pfälzische Technische Universität Kaiserslautern Landau

Explain and apply strategies to identify the bestfit model

- How does the research goal influence the selection of the best-fit model?
- Which goodness of fit measures can be used to compare models?
- List the model selection strategies and criticise step-wise model selection.
- Categorise and evaluate approaches to improve and replace stepwise model selection.

- **(Adjusted) R squared**

R^2 or adj. R^2

- R^2 increases with each additional variable in model (also noise)
- adj. R^2 should be preferred for model comparison, because it penalises for additional variables

$$R^2 = r^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \quad \text{where } \mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$\mathrm{adj}.R^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

- **AIC**
- **BIC**

$$\mathrm{AIC} = n\log\left(\frac{RSS}{n}\right) + 2p + const.$$

n = sample size

p = parameters in model

$$\mathrm{AIC_c} = \mathrm{AIC} + \frac{2p(p+1)}{n-p-1} \qquad \mathrm{BIC} = n\log\left(\frac{RSS}{n}\right) + \ln(n)p + const.$$

- The lower the value, the better the model

- **Cross-validation with MSPE**

For prediction

Explain and apply strategies to identify the bestfit model

- How does the research goal influence the selection of the best-fit model?
- Which goodness of fit measures can be used to compare models?
- List the model selection strategies and criticise step-wise model selection.
- Categorise and evaluate approaches to improve and replace stepwise model selection.

1) Best subset: Compute all 2^p (p = number of parameters) models (w/o interactions) $\rightarrow$ computationally demanding

- **Computationally demanding**
- Possible **without prior scientific knowledge**

2) Stepwise model selection

- Procedures: **forward, backward, both**
- **Consecutive adding/elimination** of parameters **until model fit decreases** (only **„useful“ parameters** are kept)
- **Backward** procedure = **best approach**
  - Better with **collinear variables**
  - Full model provides **accurate p-values, standard errors, etc**.

Problems include (see Harrell 2015: 68):

- R^2 values biased high
- Standard errors and confidence intervals too low/narrow
- Regression coefficients biased high, require shrinkage
- Collinearity renders variable selection arbitrary
- Allows to not think about the problem

Explain and apply strategies to identify the bestfit model

- How does the research goal influence the selection of the best-fit model?
- Which goodness of fit measures can be used to compare models?
- List the model selection strategies and criticise step-wise model selection.
- Categorise and evaluate approaches to improve and replace stepwise model selection.

(Partial) fixes

- Modify stepwise approach or related results:
  - correction of *p*-values for sequential testing (Fithian 2015 *ArXiv e-prints*)
  - employ bootstrapping or cross-validation on all steps of model selection
    (but see Harrell 2015: 70f, Austin 2008 *J Clin Epidem*)
  - apply shrinkage factor(s) *c* to regression coefficients, which is/are estimated via CV:

Global shrinkage factor

$$b_0^s = (1-\hat{c})\bar{y} + \hat{c}\, b_0$$
$$b_j^s = \hat{c}\, b_j;\; j = 1, \dots, p$$

Parameterwise shrinkage factor

$$b_0^s = (1-\hat{c}_0)\bar{y} + \hat{c}_0 b_0$$
$$b_j^s = \hat{c}_j b_j;\; j = 1, \dots, p$$

- Use shrinkage method such as the LASSO (Least Absolute Shrinkage and Selection Operator)

13

More likely to **find the best model** (with most useful parameters)

Improve performance of bestfit model, **better performance on new data (= less variance)**

Austin (2008) found no improved performance of bootstrapping model selection compared to backward stepwise selection. Harrell (2015: 70f) discusses several drawbacks of the bootstrap approach.
Cross-Validatio

In most cases **backwards model selection performs equally well** as bootstrapping and LASSO but is simpler

A simulation s equally well as the LASSO in t tion with parameter-wise shrinkage approach performed best in all scenarios. Interestingly, backward stepwise elimination yielded often to more parsimonious (sparser) models than the LASSO (see next slides).

Interpret models and apply variable-importance measures

- Which types of model diagnostics are required for multiple regression models?
- Outline methods to diagnose and to deal with collinearity.
- Discuss methods to check the relative importance of variables.

- **Linearity**
- **Homoscedasticity** (constant variance of residuals)
- **Normal distribution** of residuals
- **Independence** of residuals
- Identify leverage/influential points, **outliers**
- No **multicollinearity** present
- (**Cross-validation**, if goal = prediction)

- **Simple** linear regression
- **Multiple** linear regression

Interpret models and apply variable-importance measures

- Which types of model diagnostics are required for multiple regression models?
- Outline methods to diagnose and to deal with collinearity.
- Discuss methods to check the relative importance of variables.

- Check for multicollinearity:
  - Definition: Strong correlation between explanatory variables
  - Can lead to incorrect estimates of regression coefficients and related p-values of relevant predictors in the model
  - Inspect visually and using correlation analysis or variance inflation factors (VIF):

$$\mathrm{VIF} = \frac{1}{1 - R_j^2}$$

R_j is the explained variance for the linear model where the (explanatory) variable X_j is explained by all other variables in the model

VIF > 4 to 5 ⟶ Problem

Dealing with multicollinearity

- Select explanatory variables based on scientific knowledge
- Scatterplots and VIFs can aid in identifying variables with high multicollinearity, but can not suggest what to do
- Do not automatically remove the variable with the highest VIF! Check relevance of variables based on current scientific understanding
- Approaches to deal with multicollinearity:
  - Omit variables from model based on scientific knowledge
  - Select alternative model (e.g. ridge regression, elastic net, principal component regression). If priors can be specified for regression coefficients, use Bayesian regression.

Interpret models and apply variable-importance measures

- Which types of model diagnostics are required for multiple regression models?
- Outline methods to diagnose and to deal with collinearity.
- Discuss methods to check the relative importance of variables.

Measures for relative importance of variables

- Standardized betas, explained variance or both
- Standardized betas are scaled regression coefficients:

$$b_{k,standardized} = b_k \frac{s_k}{s_y}$$

s_k = standard variation of predictor k

s_y = standard variation of response y

- Hierarchical partitioning (Chevan & Sutherland 1991) and PMVD (Feldman 2005) more suitable

22

- **Differentiates between unique and shared effects** on response variable (how much unique predictive information a variable provides and how much is redundant)

- **Easier to compare** the impact of predictors, even with **different units** (e.g. % vs. °C)
- Higher stand. beta = **more impactful**
- **Impact** of a predictor (high standardized beta) **does not necessarily imply a large proportion of explained variance** (R squared), due to correlation between variables

Modelling steps

Describe the modelling steps in multiple linear regression.

Brief tutorial for multiple regression

1. Transform variables if necessary (check range, distribution)
2. Check for multicollinearity, if present, omit variables or adjust/change model

Data preparation

3. Choose modelling strategy (e.g. specify models *a priori*, LASSO) in line with research goal
4. Identify best-fit model by applying modelling strategy

Modelling

5. Run diagnostics for best-fit model
6. Validate model using cross-validation or validation sample
7. Determine variable importance if of interest

Model diagnosis and analysis

24