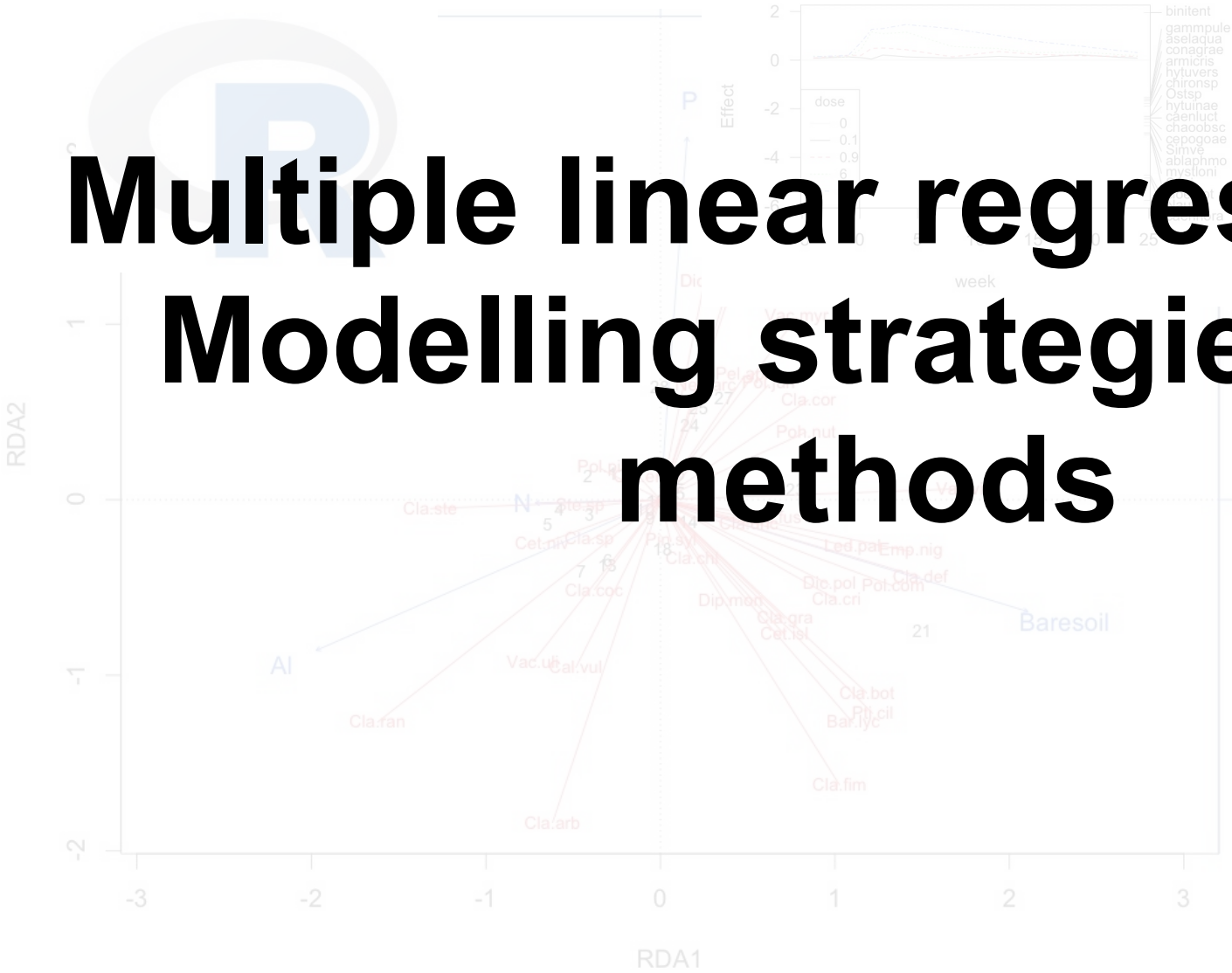
Tools for complex data analysis

University of Koblenz-Landau 2020/21



Multiple linear regression II: Modelling strategies and methods

Ralf B. Schäfer

Learning targets

- Explain and apply strategies to identify the best-fit model
- Interpret models and apply variable-importance measures
- Describe the modelling steps in multiple linear regression.

Learning targets and study questions

- Explain and apply strategies to identify the best-fit 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.

Learning targets and study questions

- 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.
- Describe the modelling steps in multiple linear regression.

Multiple regression analysis

Contents

1. **Modelling scheme and goodness of fit measures**
2. Stepwise model selection
3. The LASSO
4. Data preparation & Multicollinearity
5. Model diagnosis and analysis, small sample sizes and general tutorial

Case study: Ostracods

Which patterns and factors control the diversity of marine arctic ostracods?

136 ostracod samples from different regions
10 explanatory variables

<u>Aim:</u> 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"

http://www.phdcomics.com/comics/archive.php?comicid=1237

Modelling scheme (mainly for explanation)

Which variables should be included in the multiple regression model?

Full: Model 1: *Diversity ~ Var a, Var b, Var c, Var d, Var e*

Reduced: Model 2: *Diversity ~ Var a, Var b, Var c*

Model 3: *Diversity ~ Var a, Var b, Var c, Var d*

⋮

Model n: *Diversity ~ Var b, Var d, Var e*

Strategies

- Compare pre-specified models
- Best subset model selection
- Stepwise model selection
- Shrinkage methods

Quantitative model comparison via goodness of fit measures

Best-fit model

- model diagnostics
- model validation

Goodness of fit (GOF) measures

- 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
- Information theoretic goodness of fit measures for linear model:

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

n = sample size

p = parameters in model

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

  - The lower the value, the better the model
- For prediction: Cross-validation with MSPE

Model selection strategies

How to identify the best-fit model?

- Ideally: Comparison of a limited number of *a priori* specified models (based on knowledge)
- Traditionally used: 1) best subset and 2) stepwise model selection

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

2) Stepwise model selection requires start model, computes all models for next step (inclusion or exclusion of variable) and selects best model. Algorithm is repeated until change of included variables would reduce model fit.

- Stepwise selection procedures: backward (variable elimination), forward (variable inclusion), both (combined)

Multiple regression analysis

Contents

1. Modelling scheme and goodness of fit measures
2. **Stepwise model selection**
3. The LASSO
4. Data preparation & Multicollinearity
5. Model diagnosis and analysis, small sample sizes and general tutorial

Stepwise model selection

- Can be linked to the assessment of hypotheses:
  - Partial F-test for difference in explained variance between models:

$$\frac{\left(RSS_{\text{reduced model}} - RSS_{\text{full model}}\right) / \left(DoF_{\text{reduced model}} - DoF_{\text{full model}}\right)}{RSS_{\text{full model}} / DoF_{\text{full model}}}$$

  - If models nested and differ only by one predictor, partial F-test is equivalent to t-test for this predictor with H_0: $\beta = 0$
    Remove variable if H_0 not rejected/seems likely
  - Multiple inference (e.g. multiple tests on same data or tests on subset of data selected in light of data) leads to inflation of p-values (computed p-values biased low)
    see: Taylor & Tibshirani (2015) PNAS 112: 7629
  - should only be considered for data sets with few variables (< 5) and a high $n{:}p$ ratio (> 20)
- Can be linked to information-theoretic criteria (AIC, BIC)

Problems of stepwise model selection

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

"Let the computer find out" is a poor strategy and usually reflects the fact that the researcher did not bother to think clearly about the problem of interest and its scientific setting"
(Burnham and Anderson, 2002)

Problems generally apply to the stepwise modelling strategy, irrespective of GOF
(Murtaugh 2014 *Ecology* 95: 611; Harrell 2015: 69)

(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

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

Parameterwise shrinkage factor

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

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

Multiple regression analysis

Contents

1. Modelling scheme and goodness of fit measures
2. Stepwise model selection
3. **The LASSO**
4. Data preparation & Multicollinearity
5. Model diagnosis and analysis, small sample sizes and general tutorial

Shrinkage method: LASSO

- Ordinary least square regression:

$$\arg\min_{b_0, b_1} \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{i,j}\right)^2$$

- Linear regression with LASSO:

$$\arg\min_{b_0, b_1} \left\{ \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{i,j}\right)^2 + \lambda \sum_{j=1}^{p} |b_j| \right\}$$

Other formulation:

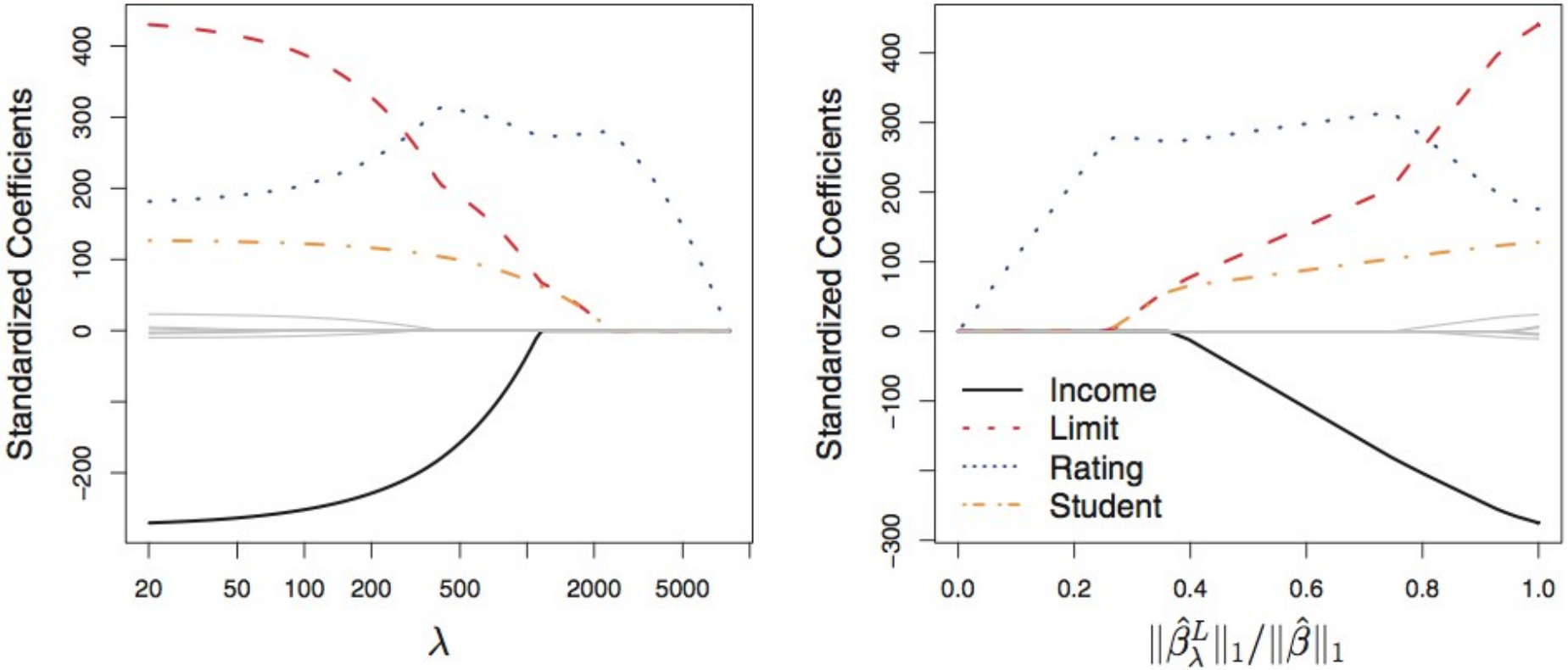
$$\arg\min_{b_0, b_1} \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{i,j}\right)^2 \text{ subject to } \sum_{j=1}^{p} |b_j| \leq s$$

- Simultaneous selection of variables and estimation of (shrinked) regression coefficients

Shrinkage method: LASSO

$$\underset{b_0, b_1}{\arg\min} \left\{ \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |b_j| \right\}$$

Example plots



- How do we identify the optimal λ? → Cross-validation (CV)

Taken from James *et al.* 2013: 220

LASSO extensions

- When does the LASSO not capture the true model?
  - Case 1: Model with several predictors, most or all relevant. → LASSO likely shrinks small regression coefficients to zero (particular of collinear predictor(s)).
  - Case 2: Model with many predictors, only few relevant. → Optimizing λ regarding prediction (in CV) can lead to selection of noise variables.
  - Alternative: Stability selection.
  - Case 3: High correlation among relevant predictors → LASSO likely selects only one. Alternative: Ridge regression or Elastic net.
  - Case 4: High correlation between relevant and irrelevant predictors. → LASSO may select irrelevant predictor(s). Alternative: Adaptive LASSO.
- Sparsity and absence of collinearity as crucial factors

Multiple regression analysis

Contents

1. Modelling scheme and goodness of fit measures
2. Stepwise model selection
3. The LASSO
4. **Data preparation & Multicollinearity**
5. Model diagnosis and analysis, small sample sizes and general tutorial

Data preparation

- Check distribution of predictors and transform if strongly skewed and spanning orders of magnitude
- 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

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.

Multiple regression analysis

Contents

1. Modelling scheme and goodness of fit measures
2. Stepwise model selection
3. The LASSO
4. Data preparation & Multicollinearity
5. **Model diagnosis and analysis, small sample sizes and general tutorial**

Model diagnostics and analysis

1. Check assumptions of simple regression model (normal distribution and homogeneity of variance of residuals, independence of residuals, linearity)
2. Check for leverage points, outliers and influential points
3. Use cross-validation to determine prediction accuracy if goal is prediction (and unless used in model selection)

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

Dealing with small sample sizes

- $\text{sqrt}(n)/p$ > 1; extreme cases (e.g. genetic data): $n < p$
- OLS regression and LASSO unreliable, several modelling approaches not applicable for $n < p$ (e.g. backward elimination)
- Approaches to deal with small sample sizes:
  - Reduce parameters manually: remove variables based on scientific understanding, very low variability or narrow distribution, and missing values
  - Reduce parameters through redundancy techniques: statistical algorithms before modelling that directly reduce number of variables or aid in removal of variables e.g. variable clustering, principal component analysis (PCA)
  - Select alternative model: Elastic net

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

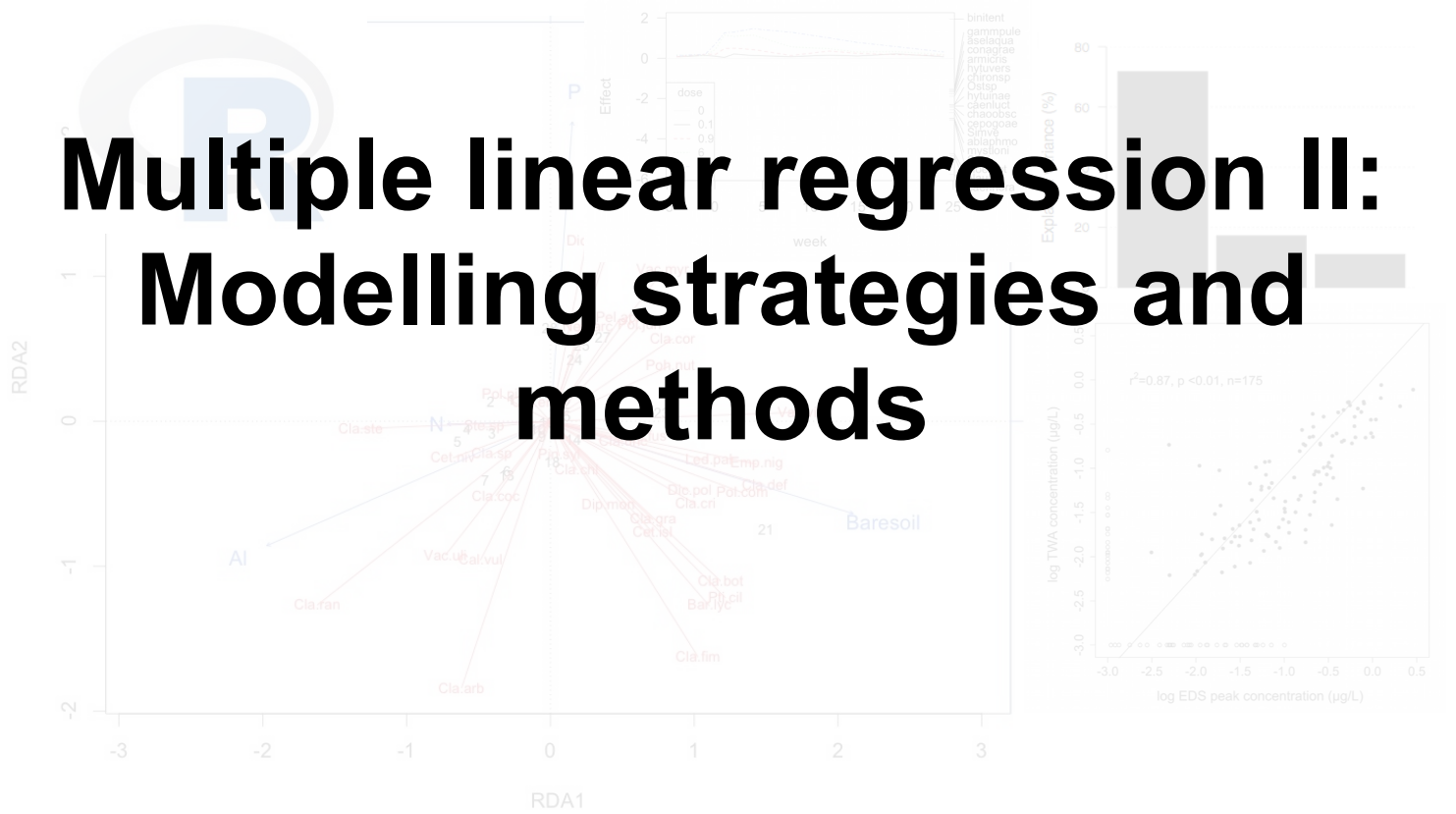
6. Validate model using cross-validation or validation sample

7. Determine variable importance if of interest

Model diagnosis and analysis

Tools for complex data analysis

University of Koblenz-Landau 2020/21



Multiple linear regression II: Modelling strategies and methods

Ralf B. Schäfer

These slides and notes complement the lecture with exercises “Tools for complex data analysis” for ecotoxicologists and environmental scientists. Do not hesitate to contact me if you have any comments or you find any errors (slides, slide notes, or code): schaefer-ralf@uni-landau.de

While I made notes below the slides, some aspects are only mentioned in the R demonstration associated with the lecture.

Learning targets

- Explain and apply strategies to identify the best-fit model
- Interpret models and apply variable-importance measures
- Describe the modelling steps in multiple linear regression.

2

Learning targets and study questions

- Explain and apply strategies to identify the best-fit 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.

3

Learning targets and study questions

- 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.
- Describe the modelling steps in multiple linear regression.

4

Multiple regression analysis

Contents

1. **Modelling scheme and goodness of fit measures**
2. Stepwise model selection
3. The LASSO
4. Data preparation & Multicollinearity
5. Model diagnosis and analysis, small sample sizes and general tutorial

5

Case study: Ostracods

Which patterns and factors control the diversity of marine arctic ostracods?

136 ostracod samples from different regions
10 explanatory variables

<u>Aim:</u> Identify most important explanatory variables for diversity of marine ostracods.
→ For explanation search for most parsimonious model

http://www.phdcomics.com/comics/archive.php?comicid=1237

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

6

The ostracod picture has been taken from:
https://www4.uwm.edu/fieldstation/naturalhistory/bugoftheweek/images/ostracod12-10.jpg

We assume that the relationship between explanatory variables and the species richness is largely linear (or quadratic) in the case study, for details see Yasuhara et al. (2012).

Occam's razor may be translated to our situation as: Given a similar predictive or explanatory power, models with fewer variables are generally better than those with more variables.

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.

Heinze, G., & Dunkler, D. (2017). Five myths about variable selection. *Transplant International*, 30(1), 6–10. https://doi.org/10.1111/tri.12895

Yasuhara M., Hunt G., van Dijken G., Arrigo K.R., Cronin T.M. & Wollenburg J.E. (2012) Patterns and controlling factors of species diversity in the Arctic Ocean. Journal of Biogeography 39, 2081–2088. Freely accessible within our university at: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1365-2699.2012.02758.x

Modelling scheme (mainly for explanation)

Which variables should be included in the multiple regression model?

Full: Model 1: *Diversity ~ Var a, Var b, Var c, Var d, Var e*

Reduced: Model 2: *Diversity ~ Var a, Var b, Var c*

Model 3: *Diversity ~ Var a, Var b, Var c, Var d*

⋮

Model n: *Diversity ~ Var b, Var d, Var e*

Strategies

- Compare pre-specified models
- Best subset model selection
- Stepwise model selection
- Shrinkage methods

Quantitative model comparison via goodness of fit measures

Best-fit model

- model diagnostics
- model validation

7

For prediction, we often can use the full model and do not need to select a modelling strategy (see previous slide), unless this is prohibited by the sample size or too large number of predictors (discussed later). If we aim to determine an effect size or assess a specific hypothesis, we should have a pre-specified model and other strategies are largely irrelevant.

Why do we not assess the importance of variables from multiple simple linear regressions? This is because in simple linear regressions, important variables may be ignored that exert a high explanatory power in the presence of other variables in the model. However, this can only be the case if variables are dependent, i.e. not orthogonal. We have discussed the issue of orthogonality and independence in the context of ANOVA. For further explanation see Sun et al. (1996) and Heinze & Dunkler (2017).

Heinze, G., & Dunkler, D. (2017). Five myths about variable selection. *Transplant International*, 30(1), 6–10. https://doi.org/10.1111/tri.12895

Sun G.-W., Shook T.L. & Kay G.L. (1996) Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. Journal of Clinical Epidemiology 49, 907 – 916.

Goodness of fit (GOF) measures

- 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
- Information theoretic goodness of fit measures for linear model:

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

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
- For prediction: Cross-validation with MSPE

Here, the information criteria are expressed for models subject to least square fitting. Generally, the AIC is given as: AIC = -2 log(L) + 2p, where L is the likelihood function for the parameters in the model. Similarly, the BIC is given as: BIC = -2 log(L) + p log(n). We will discuss the concept of log likelihood in the context of the Generalized Linear Model (GLM).

Note that the R function AIC() calculates a simplified version of AIC without the constant term. This constant term is -n log(n). Hence, when models for the same data are compared, this omission is acceptable because the constant is the same for all models and the addition of a constant does not affect the ranking of the absolute values of the models. The BIC gives higher penalty to more complex models than the AIC for $n \geq 8$ and may thus aid in selection of more parsimonious (sparser) models. The AIC tends towards over-fitting especially for smaller data sets (e.g. $n < 50$). The corrected AIC ($\mathrm{AIC_C}$) is the recommended alternative. In fact, the corrected AIC could always be used as it converges with the AIC for larger sample sizes.

The adjusted (adj.) R^2 should be preferred over the normal R^2 as it takes the number of explanatory variables p into account.

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

Model selection strategies

How to identify the best-fit model?

- Ideally: Comparison of a limited number of *a priori* specified models (based on knowledge)
- Traditionally used: 1) best subset and 2) stepwise model selection

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

2) Stepwise model selection requires start model, computes all models for next step (inclusion or exclusion of variable) and selects best model. Algorithm is repeated until change of included variables would reduce model fit.

- Stepwise selection procedures: backward (variable elimination), forward (variable inclusion), both (combined)

If our aim is to estimate effect sizes or to test hypotheses, we should pre-specify a model (or a few models).

Computing all possible models represents an exhaustive search for the best regression model. This procedure is most useful when no prior knowledge on the ranking of the scientific relevance of variables is available. If the number of possible models is large, different models may have a similar GOF so that the selection of only one model is problematic. Model averaging over all models up to a certain threshold of a GOF measure can be applied in this case (see R demonstration). A review by Grueber et al. (2011) discusses several issues associated with model selection and averaging. Cade (2015) cautions regarding the use of model averaging.

Best subset selection quickly becomes computationally demanding, especially if model validation would be applied to the whole process. For example, for ten-fold CV and 10 explanatory variables, 1024 * 10 models need to be fitted. The number of possible models becomes even higher if interaction terms are included. Moreover, compared to a more constrained search, best subset selection is likely to yield to a model with higher variance in prediction (see Hastie, Tibshirani and Friedman 2017: 59). Kuhn and Johnson (2020, see Chapters 10 to 12) provide guidance on how to evaluate the model via resampling methods to avoid overfitting.

The criticism on stepwise model selection (see next slides) also applies to best subset selection.

Cade B.S. (2015). Model averaging and muddled multimodel inferences. Ecology 96, 2370–2382. Freely accessible within our university at: https://doi.org/10.1890/14-1639.1

Grueber C.E., Nakagawa S., Laws R.J. & Jamieson I.G. (2011). Multimodel inference in ecology and evolution: challenges and solutions. Journal of Evolutionary Biology 24, 699–711. Freely accessible within our university at: http://onlinelibrary.wiley.com/doi/10.1111/j.1420-9101.2010.02210.x/abstract

Multiple regression analysis

Contents

1. Modelling scheme and goodness of fit measures
2. **Stepwise model selection**
3. The LASSO
4. Data preparation & Multicollinearity
5. Model diagnosis and analysis, small sample sizes and general tutorial

10

Stepwise model selection

- Can be linked to the assessment of hypotheses:
  - Partial F-test for difference in explained variance between models:

$$\frac{\left(RSS_{\text{reduced model}} - RSS_{\text{full model}}\right) / \left(DoF_{\text{reduced model}} - DoF_{\text{full model}}\right)}{RSS_{\text{full model}} / DoF_{\text{full model}}}$$

  - If models nested and differ only by one predictor, partial F-test is equivalent to t-test for this predictor with $H_0: \beta = 0$
    Remove variable if H_0 not rejected/seems likely
  - Multiple inference (e.g. multiple tests on same data or tests on subset of data selected in light of data) leads to inflation of p-values (computed p-values biased low)
    see: Taylor & Tibshirani (2015) PNAS 112: 7629
  - should only be considered for data sets with few variables (< 5) and a high $n{:}p$ ratio (> 20)
- Can be linked to information-theoretic criteria (AIC, BIC)

11

n = sample size, p = parameters in model

Heinze and Dunkler (2017) argue that in the case of stepwise selection an $n{:}p$ ratio of up to 50 is needed to achieve stable results.

Murtaugh (2014) pointed out that the p-value and the information-theoretic criteria (AIC, BIC) are intimately linked. Thus, they face similar problems in model comparison and selection. For p-values that correspond to a selection based on the AIC see Heinze & Dunkler (2017).

Heinze, G., & Dunkler, D. (2017). Five myths about variable selection. *Transplant International*, 30(1), 6–10. https://doi.org/10.1111/tri.12895

Murtaugh P.A. (2014). In defense of P values. Ecology 95, 611–617. Freely accessible within our university at: https://doi.org/10.1890/13-0590.1

Taylor J. & Tibshirani R.J. (2015). Statistical learning and selective inference. Proceedings of the National Academy of Sciences 112, 7629–7634. Freely accessible at: https://doi.org/10.1073/pnas.1507583112

Problems of stepwise model selection

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

"Let the computer find out" is a poor strategy and usually reflects the fact that the researcher did not bother to think clearly about the problem of interest and its scientific setting"
(Burnham and Anderson, 2002)

Problems generally apply to the stepwise modelling strategy, irrespective of GOF
(Murtaugh 2014 *Ecology* 95: 611; Harrell 2015: 69)

If stepwise model selection is used, then choose the backward selection approach, because this performs better in the presence of collinear variables and starts with the full model, which is the only model providing accurate p-values, standard errors etc (Harrell 2015: p.70). However, backward selection cannot be used, if $n < p$ (this case is discussed later in this session) and, generally, a high $n{:}p$ ratio (e.g. ~50) may be required to achieve stable results (Heinze & Dunkler 2017). Similarly, Kuhn and Johnson (2020: 252-254) argue that stepwise model selection should be avoided. They also judge backward selection (called Recursive Feature Elimination in their book) as a potentially effective strategy (Kuhn and Johnson 2020: 251). Dunkler et al. (2014) have developed an augmented version of backward elimination that should be used if the $n{:}p$ ratio is below 100.

The issue of collinearity will be discussed in detail later.

Dunkler, D., Plischke, M., Leffondré, K., & Heinze, G. (2014). Augmented backward elimination: A pragmatic and purposeful way to develop statistical models. *PLOS ONE*, 9(11), 1–19. https://doi.org/10.1371/journal.pone.0113677

Heinze, G., & Dunkler, D. (2017). Five myths about variable selection. *Transplant International*, 30(1), 6–10. https://doi.org/10.1111/tri.12895

(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

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

Parameterwise shrinkage factor

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

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

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-Validation is similarly likely to underestimate the true variance.

The application of shrinkage factors after model selection is called post-selection shrinkage.

A simulation study found that backward stepwise elimination performed equally well as the LASSO in the identification of true predictors, particularly in conjunction with parameter-wise shrinkage (Houwelingen & Sauerbrei 2013). However, no approach performed best in all scenarios. Interestingly, backward stepwise elimination yielded often to more parsimonious (sparser) models than the LASSO (see next slides).

Austin P.C. (2008) Bootstrap model selection had similar performance for selecting authentic and noise variables compared to backward variable elimination: a simulation study. Journal of Clinical Epidemiology 61, 1009 – 1017.e1.
Fithian W., Taylor J., Tibshirani R. & Tibshirani R. (2015) Selective Sequential Model Selection. ArXiv e-prints, 1–36. http://adsabs.harvard.edu/abs/2015arXiv151202565F (an updated version can be found here: http://www.stat.cmu.edu/~ryantibs/papers/seqinf.pdf)
Houwelingen H.C. van & Sauerbrei W. (2013) Cross-Validation, Shrinkage and Variable Selection in Linear Regression Revisited. Open Journal of Statistics 03, 79–102. Freely accessible at: https://www.scirp.org/pdf/OJS_2013042410543131.pdf

Multiple regression analysis

Contents

14

Shrinkage method: LASSO

- Ordinary least square regression:

$$\arg\min_{b_0, b_1} \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{i,j} \right)^2$$

- Linear regression with LASSO:

$$\arg\min_{b_0, b_1} \left\{ \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |b_j| \right\}$$

Other formulation:

$$\arg\min_{b_0, b_1} \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{i,j} \right)^2 \text{ subject to } \sum_{j=1}^{p} |b_j| \leq s$$

- Simultaneous selection of variables and estimation of (shrinked) regression coefficients

15

Using the LASSO is motivated by two problems of OLS regression:

1. Regression coefficients are biased high, leading to high variance in prediction. Shrinking of regression coefficients increases the bias (remember the bias-variance tradeoff), but reduces the variance.
2. For a large number of predictors, interpretation of the OLS result becomes tricky. Focusing on a smaller subset improves interpretation.

Through introduction of the penalty term in LASSO, the regression coefficients are shrunk. With increasing penalty (i.e. larger λ and smaller *s*, respectively) some regression coefficients are shrunk to 0, which means that the LASSO performs variable selection.

The penalty term is called ℓ_1-norm in mathematical terms, where norm is a function that assigns a value to a vector (in our case to the vector of regression coefficients) and the ℓ notation is a reference to a specific mathematical space.

A comparison of regression using LASSO and several other techniques can be found in Hastie, Tibshirani & Friedman (2017: 61ff). The LASSO is typically among the methods with the lowest prediction error. However, the optimal λ for prediction does not guarantee the selection of the most parsimonious model for explanation (Zou 2006).

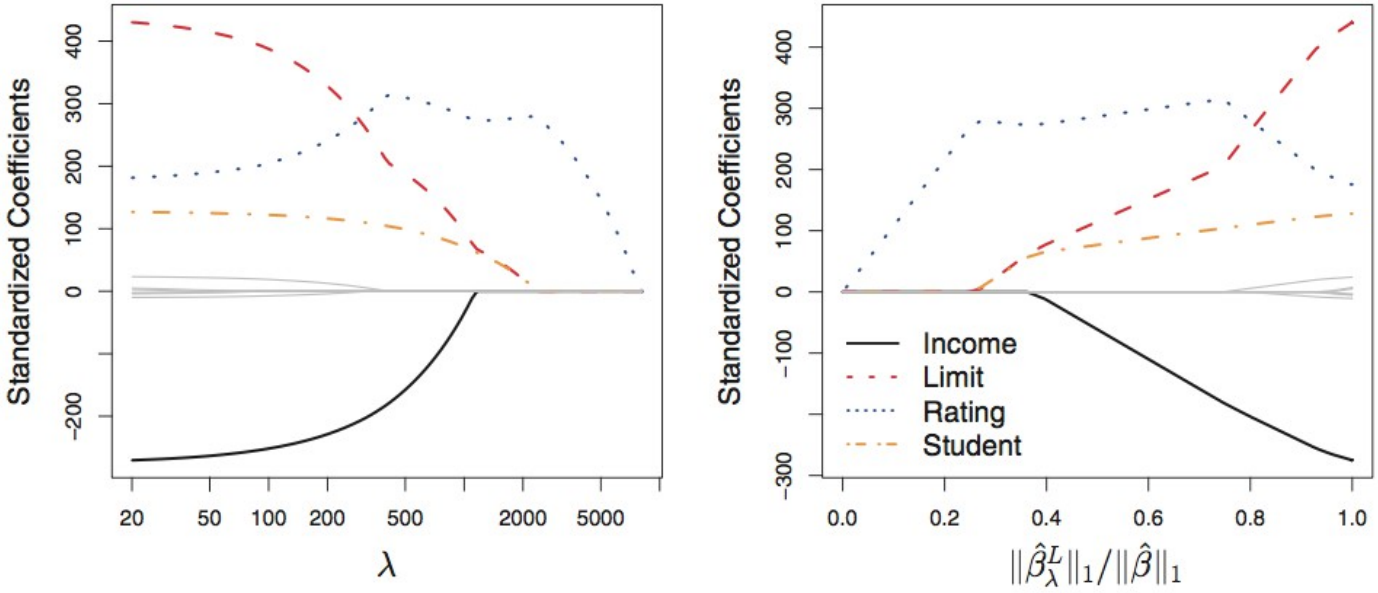
For application in R see James et al. (2013) and for further developments of shrinkage (or more precisely: sparsity) methods see Hastie, Tibshirani & Wainwright (2015). All these books are freely downloadable, the URLs are provided with the literature list.

Zou H. (2006). The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Association 101, 1418–1429. Freely accessible at: http://users.stat.umn.edu/~zouxx019/Papers/adalasso.pdf

Shrinkage method: LASSO

$$\arg\min_{b_0, b_1} \left\{ \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |b_j| \right\}$$

Example plots



- How do we identify the optimal λ? → Cross-validation (CV)

Taken from James *et al.* 2013: 220

For the LASSO analysis, variables are typically standardized to zero mean and standard deviation of one. The advantage is that all variables have the same units after standardization, i.e. are represented on the same scale and their importance can be ranked based on the size of the coefficient. Moreover, the response Y is centred (through subtraction of the mean), which leads to a zero intercept, i.e. b_0 can be removed.

The left plot shows the standardised regression coefficients along increasing λ on the x axis. For very low values of λ, the regression coefficients are the same as for OLS regression. As λ becomes higher, all regression coefficients are shrunk towards zero and for a very high λ, we eventually obtain the null model.

The right plot displays the ratio of the absolute sum of the standardized regression coefficients for the LASSO (i.e. ℓ^1-norm) and the absolute sum of the standardized regression coefficients from OLS. How to choose an optimal λ using the prediction error in cross validation is shown in the R tutorial.

16

LASSO extensions

- When does the LASSO not capture the true model?
  - Case 1: Model with several predictors, most or all relevant. → LASSO likely shrinks small regression coefficients to zero (particular of collinear predictor(s)).
  - Case 2: Model with many predictors, only few relevant. → Optimizing λ regarding prediction (in CV) can lead to selection of noise variables.
  - Alternative: Stability selection.
  - Case 3: High correlation among relevant predictors → LASSO likely selects only one. Alternative: Ridge regression or Elastic net.
  - Case 4: High correlation between relevant and irrelevant predictors. → LASSO may select irrelevant predictor(s). Alternative: Adaptive LASSO.
- Sparsity and absence of collinearity as crucial factors

The LASSO is typically used, if sparsity can be assumed, i.e. that of a large number of predictors only a small subset is relevant. In other words, the true model in the statistical population only contains a few predictors. I outline a few cases where the LASSO may not capture the true model. This may not be that relevant if the research goal is prediction, but it is if the research goal is explanation.

An alternative to the conventional LASSO approach to determine the optimal λ via the prediction error in CV, is to employ a resampling procedure such as bootstrapping, where for each λ several bootstrap samples are used to identify the relevant variables. The related method is called stability selection and identifies the most relevant variables based on their probability of selection in resampling (see Meinhausen & Bühlmann 2010 for details).

Moreover, the true model may not be captured if the relevant predictors exhibit a high intercorrelation (i.e. collinearity) or a high correlation with irrelevant variables. However, for real data where the true regression coefficients are unknown, it is difficult to evaluate whether our data falls into one of these cases outlined. Moreover, simulations by Kuhn and Johnson (2020: 232-235) show that the LASSO is relatively robust against the selection of false positives (i.e. noise variables), but this comes at the cost of false negatives (i.e. not selecting relevant variables).

Alternatives for the case of collinearity among the predictors are ridge regression and the elastic net (Zou & Hastie 2005). Ridge regression represents a shrinkage method similar to the LASSO, but uses a quadratic penalty term called ℓ_2-norm (see James, Witten, Hastie and Tibshirani 2013: 61ff):

$$\arg\min_{b_0, b_1} \left\{ \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} b_j^2 \right\}$$

Compared to the LASSO, ridge regression can better deal with collinear variables, but does not perform variable selection (i.e. regression coefficients are not shrunk to zero). Both ridge regression and the LASSO represent special cases of the elastic net, which combines both ℓ_1 *and* ℓ_2 penalties. Ridge regression should be used if the (relevant) predictors are collinear, and all variables should remain in the model. The elastic net should be used if (relevant) predictors are collinear and variables should be shrunk to zero. Another advantage (besides accounting for collinearity) of the elastic net over the LASSO is that it can deal with low sample sizes (further details later).

The adaptive LASSO relies on individual penalties for each predictor through adaptive weights, which are typically determined via ridge regression (in case of collinearity).

Meinshausen N. & Bühlmann P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72, 417–473. Freely accessible at: https://pdfs.semanticscholar.org/9476/3a504ed7d835051d3e52f288c9d9e4d80e03.pdf

Zou H. & Hastie T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67, 301–320. Freely accessible at: https://web.stanford.edu/~hastie/Papers/B67.2%20%282005%29%20301-320%20Zou%20&%20Hastie.pdf

Zou H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association* 101, 1418–1429. Freely accessible at: http://users.stat.umn.edu/~zouxx019/Papers/adalasso.pdf

Multiple regression analysis

Contents

1. Modelling scheme and goodness of fit measures
2. Stepwise model selection
3. The LASSO
4. **Data preparation & Multicollinearity**
5. Model diagnosis and analysis, small sample sizes and general tutorial

18

Data preparation

- Check distribution of predictors and transform if strongly skewed and spanning orders of magnitude
- 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

Generally, transformation of the explanatory variables is necessary, if they are highly skewed and the data span several orders of magnitude. Especially chemical data often exhibit a skewed distribution (due to detection limits) that should be transformed before multiple regression analysis. See Fox & Weisberg (2019) Chapter 3 for details.

For a mathematical explanation why the regression coefficients are influenced by collinearity see Matloff (2017) p. 268-269.

Multicollinearity does not affect predictions made on the same data set, or new data sets, where variables exhibit a similar collinearity structure as for the original data.

Regarding the VIF there are different rules of thumb as to when one should worry about collinearity. Most textbooks suggest that for VIF values >4 (Kabacoff 2011: 200) or >5 (Sheather 2009: 203) collinearity may represent a problem. However, drawing a sharp line within a continuum of values (e.g. VIFs) remains to some extent arbitrary.

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.

For an overview of methods to deal with collinearity as well as a thorough overview on the issue in general, see Fox (2015) Chapter 13. For a comparison of methods how to deal with collinearity see Dormann et al. (2013). Interestingly, threshold-based pre-selection (i.e. omission of one of the variables if pairwise correlation exceeds a threshold, e.g. Pearson correlation coefficient $|r| > 0.7$.) was not outperformed by more sophisticated methods.

Principal component regression is covered later in the course. A comparison of methods including the LASSO, ridge regression and principle component regression can be found in Hastie, Tibshirani & Friedman (2017: 61ff).

We discussed before that the ridge regression should be used, if all predictors should remain in the model, whereas the elastic net also conducts variable selection by shrinking regression coefficients to zero and is able to deal with low sample sizes.

In cases with hundreds or thousands of predictors or more, collinearity is rather the rule and selection of predictors before modelling may be required. See Kuhn and Johnson (2020: Chapters 10 to 12) for an overview on methods of selecting predictors and advanced methods to deal with a large number of predictors (Chapter 12).

Dormann et al. (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36: 27-46. Freely accessible at: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1600-0587.2012.07348.x

Multiple regression analysis

Contents

1. Modelling scheme and goodness of fit measures
2. Stepwise model selection
3. The LASSO
4. Data preparation & Multicollinearity
5. **Model diagnosis and analysis, small sample sizes and general tutorial**

21

Model diagnostics and analysis

1. Check assumptions of simple regression model (normal distribution and homogeneity of variance of residuals, independence of residuals, linearity)
2. Check for leverage points, outliers and influential points
3. Use cross-validation to determine prediction accuracy if goal is prediction (and unless used in model selection)

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

See previous sessions on how to diagnose violation of assumptions and how to deal with them. As mentioned before, the independence of residuals may not hold true for time series or spatial data. If temporal or spatial autocorrelation exists, this should be included in the model structure for example using a generalised least squares model or using a mixed model, which is beyond the scope of this course (see Zuur et al. 2009).

Standardized betas are contested as they do not consider the partitioning of R^2 (unless the explanatory variables are orthogonal, i.e. their correlation is 0) and do not account for the direct effect of a variable in the model (for example: a high direct effect of a predictor may be assigned to other correlating predictors and result in a low estimate of the beta for that predictor). Nevertheless, they inform on the change of the response variable for one unit change in the predictor when all predictors are standardized to the same scale.

Note that hierarchical partitioning is a general variable importance measure that can be used with any model that provides a goodness of fit metric. For linear models and R^2 as metric, hierarchical partitioning is equivalent to the LMG method that is often mentioned in journal articles.

For a general discussion of variable importance measures refer to Grömping (2015) and Johnson and Lebreton (2004). Wei et al. (2015) give an overview on variable importance measures for different data analysis techniques ranging from regression models to random forests (a machine learning technique discussed later). Grömping (2006) focuses on the implementation in R. For variable importance measures that are applicable across models of different type see Molnar (2020: Chapter 5).

Grömping, U. (2006) Relative importance for linear regression in R: The package relaimpo. *JSS,* 17.

Grömping U. (2015) Variable importance in regression models. *WIREs Comput Stat* 7, 137–152

Johnson & Lebreton (2004). History and Use of Relative Importance Indices in Organizational Research. *Organizat Res Meth*, 7, 238–25

Molnar, C. (2020). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Lean publisher. Freely accessible at: https://christophm.github.io/interpretable-ml-book/

Wei P., Lu Z. & Song J. (2015) Variable importance analysis: A comprehensive review. *Reliability Engineering & System Safety* 142, 399–432.

Zuur A.F., Ieno E.N., Walker N.J., Saveliev A.A. & Smith G.M. (2009). Mixed effects models and extensions in ecology with R. Springer, New York, NY.

Dealing with small sample sizes

- $\text{sqrt}(n)/p$ > 1; extreme cases (e.g. genetic data): $n < p$
- OLS regression and LASSO unreliable, several modelling approaches not applicable for $n < p$ (e.g. backward elimination)
- Approaches to deal with small sample sizes:
  - Reduce parameters manually: remove variables based on scientific understanding, very low variability or narrow distribution, and missing values
  - Reduce parameters through redundancy techniques: statistical algorithms before modelling that directly reduce number of variables or aid in removal of variables e.g. variable clustering, principal component analysis (PCA)
  - Select alternative model: Elastic net

For multiple regression analysis with OLS, the number of parameters p in the model should be smaller than $\text{sqrt}(n)$, see Matloff (2017) p. 441. Harrel (2015) pp. 72-74 suggests a more restrictive rule: the number of parameters p in the model should be smaller or equal than $n/15$. Finally, Heinze and Dunkler (2017) even suggest that the number of parameters p in the model should be around $n/50$ to achieve reliable results. Consult both references for further details.

Running a PCA and then using the principal components in regression analysis is called principal component regression (PCR). We will discuss PCA and cluster analysis later in the course. Briefly, PCA constructs new, non-correlated (orthogonal) gradients from a data set. In case of collinearity, this can help to reduce the number of variables. Cluster analysis identifies similar groups of variables, where group representatives could be subsequently selected for modeling. See Harrell (2015) pp. 79 for further details and techniques. These approaches are particularly powerful if the predictors are strongly correlated, for example, in the case of bioclimatic or water quality variables. For an application of PCR see Bhowmik & Schäfer (2015). For advanced algorithms of selecting predictors see Kuhn and Johnson (2020: Chapters 10 to 12).

The elastic net combines both ℓ^1 *and* ℓ^2 penalties of the LASSO and ridge regression. It can deal with $n < p$ situations and with collinear variables. For details see Zou & Hastie (2005) and Hastie, Tibshirani & Friedman (2017: 661ff).

Another alternative model, but with a completely different modelling approach, is the Random Forest model, which will be discussed later.

Bhowmik & Schäfer (2015) Large Scale Relationship between Aquatic Insect Traits and Climate. PLoS ONE 10, e0130025. Freely accessible at: http://dx.doi.org/10.1371%2Fjournal.pone.0130025

Zou H. & Hastie T. (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 301–320.

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

For further details on regression modelling strategies see Harrell (2015) pp. 63, with more detailed check lists for different modeling objectives (e.g. prediction or effect estimation) on the pages 94ff.

You can find a brief overview on the implementation of standard techniques for multiple regression in R here: http://www.statmethods.net/stats/regression.html.