Applied Statistical Regression
HS 2011 – Week 10

Marcel Dettling

Institute for Data Analysis and Process Design

Zurich University of Applied Sciences

marcel.dettling@zhaw.ch

http://stat.ethz.ch/~dettling

ETH Zürich, November 28, 2011

Applied Statistical Regression
HS 2011 – Week 10

Variable Selection: Technical Aspects

We want to keep a model small, because of

1) **Simplicity**
   → *among several explanations, the simplest is the best*

2) **Noise Reduction**
   → *unnecessary predictors leads to less accuracy*

3) **Collinearity**
   → *removing excess predictors facilitates interpretation*

4) **Prediction**
   → *less variables, less effort for data collection*

Applied Statistical Regression
HS 2011 – Week 10

AIC/BIC

Bigger models are not necessarily better than smaller ones!

→ *balance goodness-of-fit with the number of predictors used*

AIC Criterion:

$$\begin{aligned} AIC &= -2\max(\log likelihood) + 2p \\ &= const + n\log(RSS/n) + 2p \end{aligned}$$

BIC Criterion:

$$\begin{aligned} BIC &= -2\max(\log likelihood) + p\log n \\ &= const + n\log(RSS/n) + p\log n \end{aligned}$$

AIC or BIC?

Both can lead to similar decisions, but BIC punishes larger models more heavily:

→ **BIC models tend to be smaller!**

- AIC/BIC is not limited to all subset regression
- Criteria can also be (and are!) applied in the backward, forward or stepwise approaches.
- In R, variable selection is generally done by function `step()`
- Default choice: stepwise regression with AIC as a criterion.

Variable Selection: Final Remark

- Every procedure may yield a different “best” model.
- If we could obtain another sample from the same population, even a fixed procedure might result in another “best” model.
- “Best model”: element of chance, “random variable”

How can we mitigate this in practice?

It is usually advisable to not only consider the “best” model according to a particular procedure, but to check a few more models that did nearly as good, if they exist.

Applied Statistical Regression
HS 2011 – Week 10

Model Selection with Hierarchical Input

→ Some regression models have a natural hierarchy.

I.e. in polynomial models, x^2 is a higher order term than x

Important:

Lower order terms should not be removed from the model before higher order terms in the same variable. As an example, consider the polynomial model:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

→ **see blackboard…**

Applied Statistical Regression
HS 2011 – Week 10

Interactions and Categorical Input

Models with Interactions

Do not remove main effect terms if there are interactions with these predictors contained in the model.

Categorical Input

- If a single dummy coefficient is non-significant, we cannot just kick this term out of the model, but we have to test the entire block of indicator variables.
- When we work manually and testing based, this will be done with a partial F-test. When working criterion based, `step()` does the right thing

Cross Validation: Why?

- We have seen before that on a given dataset, a *bigger model* always yields a *better fit*, i.e. smaller residuals, and thus better RSS, R-squared, error variance, etc.
- Bigger models have an *unfair advantage* because there are *more predictors*. The AIC criterion tries to balance this by penalizing for the number of predictors used.
- If the *ultimate goal* is *predicting new datapoints*, then it is self suggesting to identify a model which does well at this. For making the right choice, we mimic the prediction task on our training sample with ***Cross Validation.***

10-Fold Cross Validation

Idea:

0) Split the (training) data into 10 equally sized folds

1a) Use folds 1-9 for the fit, and use the model to predict fold 10

1b) On fold 10, the forecasting performance is measured by computing the RSS, i.e. the squared difference between the forecasted and true values

2) Use folds 1-8 & 10 for fitting, predict fold 9 and record RSS

3) Use folds 1-7 & 9-10 for fitting, predict fold 8 and record RSS

4) ...

Applied Statistical Regression
HS 2011 – Week 10

10-Fold Cross Validation

Summary:

- Each observation is forecasted and gauged against the true values exactly 1x. On the other hand, it is used 9x during the model fitting process.
- With cross validation, we evaluate the "out-of-sample"-Performance, i.e. how precisely a model can forecast observations that were not used for fitting the model.
- In this regard, bigger and/or more complex models are not necessarily better than smaller/simpler ones.

Applied Statistical Regression
HS 2011 – Week 10

Cross Validation

Further remarks:

- Cross validation is often used for identifying the most predictive model from a few candidate models that were found by stepwise variable selection procedures.
- There are alternatives to 10-fold CV. Popular is n-fold CV, which is known as Leave-One-Out Cross Validation.
- In R, it's easy to code "for-loops" that do the job, but there are also existing functions (that have some limits...):

```
> library(DAAG)
> CVlm(data, formula, fold.number, …)
```

Applied Statistical Regression
HS 2010 – Week 08

Modeling Strategies

- In which order to apply: estimation – diagnostics – transformation – variable selection???

There is no definite answer to this: regression analysis is the search for structure in the data and there are no hard-and-fast rules about how it should be done.

Professional regression analysis can be seen as an art and definitely requires skill an expertise – one must be alert to unexpected structure in the data.

→ **We here provide a rough guideline for regression analysis**

Applied Statistical Regression
HS 2010 – Week 08

Guideline for Regression Analysis

0) Preprocessing the data

- learning the meaning of all variables
- give short and informative names
- check for impossible values, errors
- if they exist: set them to NA
- systematic or random missings?

1) First-aid transformations

- bring all variables to a suitable scale
- use statistical and specific knowledge
- routinely apply the first-aid transformations

Applied Statistical Regression
HS 2010 – Week 08

Guideline for Regression Analysis

2) Fitting a big model

First fit a big model with potentially too many predictors

- use all if $p < n/5$
- preselect manually according to previous knowledge
- preselect with forward search and a p-value of 0.2

3) Model Diagnostics

Check for normality, constant variance, uncorrelated errors:

- transformations
- robust regression
- weighted regression
- dealing with correlation

Applied Statistical Regression
HS 2010 – Week 08

Guideline for Regression Analysis

6) Interactions

- try (two-way) interactions
- do only use predictors that are in the model

7) Influential data points

- attractors for the regression line
- keep them or skip them?
- compare with and without

8) Do model and coefficients make sense?

- implausible predictors, wrong signs, against theory, …
- remove if there are no drastic changes!

Applied Statistical Regression
HS 2010 – Week 08

Guideline for Regression Analysis

If there were substantial changes to the model in steps 4-8), then one should go back to 3) and repeat the diagnostics.

Hypothesis testing:

- proceed similarly
- careful: transformations, selection, collinearity
- question dictates what works and what not!

Prediction:

- guideline is still reasonable
- we are a little less picky here in selection and diagnostics
- check generalization error with test data / cross validation

Applied Statistical Regression
HS 2010 – Week 08

Significance vs. Relevance

The larger a sample, the smaller the p-values for the very same predictor effect. Thus do not confuse a small p-values with an important predictor effect!!!

With large datasets:

- statistically significant results which are practically useless
- we have high evidence that a blood value is lowered by 0.1%

Models are approximative:

- most predictors have influence, thus $\beta_1 = 0$ never holds
- point null hypothesis is usually wrong in practice
- we just need enough data to be able to reject it

Applied Statistical Regression
HS 2011 – Week 10

Final Remarks to Multiple Linear Regression

All models we fit are most likely too simple/wrong…

-

However, some of these turn out to be really useful…

-

And some are more, and other are less useful…

-

Identifying the "good" ones is your job!