Applied Statistical Regression
AS 2013 – Week 11

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 25, 2013

Applied Statistical Regression
AS 2013 – Week 11

Cross Validation

Definition:

Cross Validation is a *technique for estimating the performance of a predictive model*, i.e. assessing how the prediction error will generalize to a new, independent dataset.

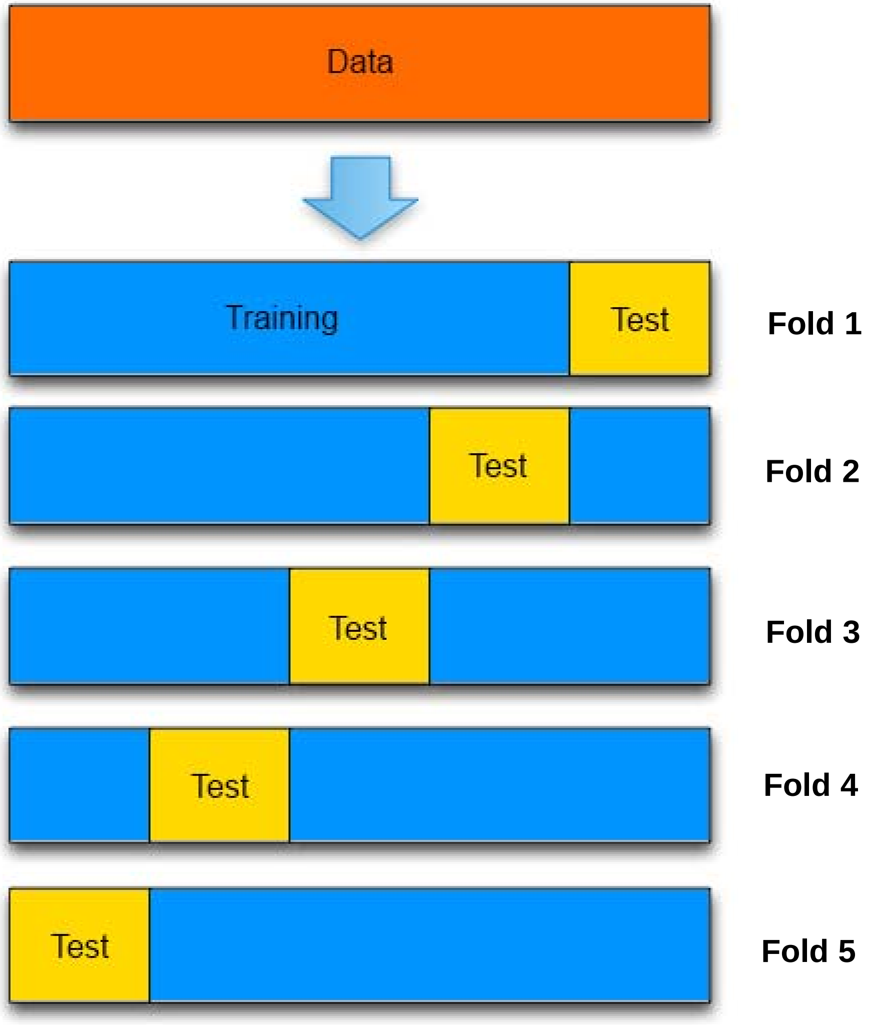
Rationale:

Cross Validation serves to prevent *overfitting*. On a given data set, a *bigger model* always yields a *better fit*, i.e. *smaller RSS*, *higher R-squared*, *less error variance*, et cetera.

While *AIC/BIC* and the *adjusted R-squared* try to overcome this problem by penalizing for model size, its use is limited in reality.

Applied Statistical Regression
AS 2013 – Week 11

Cross Validation: How It Works…



- In this schematic example, 5-fold CV is illustrated.
- Each observations belongs to exactly 1 test set. The test sets are of roughly equal size.
- Also, each data point belongs to exactly 4 training sets.
- In each fold, the test RSS is recorded. The CV-RSS is the summed result over all folds.

When to Use Cross Validation

- If the *ultimate goal* is *predicting new data points*, then it is self suggesting to identify a model which does well at this. We can mimic the prediction task on our training sample with *cross validation.*

AIC/BIC and the adjusted R-squared do not work in case of:

- *Response variable transformation:* if one wants to find out whether a model with logged or non-logged response is better for prediction, cross validation is the only option.
- *Non-identical data:* if we want to evaluate the quality of a fitted with or without removing sketchy data points, then also cross validation is our only option.

Cross Validation

Further remarks:

- Cross validation evaluates the out-of-sample performance, i.e. how precisely a model can forecast observations that were not used for fitting the model.

- There are alternatives to 5-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, …)
```

Cross Validation

Using `for()` to program cross validation loops:

```
> rss   <- c()
> fo    <- 5
> sb    <- round(seq(0,nrow(dat),length=(fo+1)))
> for (i in 1:folds)
> {
>   test   <- (sb[((fo+1)-i)]+1):(sb[((fo+2)-i)])
>   train  <- (1:nrow(dat))[-test]
>   fit    <- lm(res ~ p1+..., data=dat[train,])
>   pred   <- predict(fit, newdata=dat[test,])
>   rss[i] <- sum((dat$response[test] - pred)^2)
> }
```

Applied Statistical Regression
AS 2013 – Week 11

Modelling Strategies

We have learnt a number of technical details about multiple linear regression. The often asked question is in which order the tools need to be applied:

Transformation → Estimation → Model Diagnostics → Variable Refinement & Selection → Evaluation

This is a good generic solution, but not an always-optimal strategy.

Professional regression analysis is the search for structure in the data. It requires technical skill, flexibility and intuition. The analyst must be alert to the obvious as well as to the non-obvious, and needs the flair to find the unexpected.

Modelling Strategies

0) Data Screening & Processing

- learn 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) Transformations

- bring all variables to a suitable scale
- use statistical and specific knowledge
- routinely apply the log-transformation
- break obvious collinearity

Applied Statistical Regression

AS 2013 – Week 11

Modelling Strategies

2) Fitting a Big Model

Fit a big model with potentially too many predictors

- use all if $p < n/5$!!!

- *or* preselect manually according to previous knowledge

- *or* preselect with forward search and a p-value of 0.2

3) Model Diagnostics

- generate the 4 standard plots in R

- a systematic error is non-tolerable, improve the model!!!

- be aware to influential data points, try to understand them

- take care with non-constant variance & long-tailed errors

- think about potential correlation in the residuals

Modelling Strategies

4) Variable Selection

- try to reduce the model to what is utterly required
- run a stepwise search from the full model with AIC/BIC
- if feasible, an all-subset-search with AIC/BIC is even better
- the residual plots must not (substantially) degrade in quality!

5) Refining the Model

- use partial residual plots or plots against other variables
- think about potential non-linearities/factorization in predictors
- interaction terms can improve the fit drastically
- are there still any collinearities that disturb?

Modelling Strategies

6) **Plausibility**
   - implausible predictors, wrong signs, results against theory?
   - remove if (appropriate) and no drastic change in outcome

7) **Evaluation**
   - cross validation for model comparison & performance
   - derive test results, confidence and prediction intervals

8) **Reporting**
   - be honest and openly report manipulations & decisions
   - regression models are descriptive, but not causal!
   - do not confuse significance and relevance!

Significance vs. Relevance

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

With large datasets, we can have:

- statistically significant results which are practically useless
- e.g. high evidence that the response value is lowered by 0.1% which is often a practically totally meaningless result.

Bear in mind that generally:

- most predictors have influence, thus $\beta_j = 0$ hardly ever holds
- the point null hypothesis is thus usually wrong in practice
- we just need enough data so that we are able to reject it

Significance vs. Relevance

Absence of Evidence $\neq$ Evidence of Absence

- if one fails to reject a null hypothesis $\beta_j = 0$ we do not have a proof that the predictor does not influence the response.

- things may change if we have more data, or even if the data remain the same, but the set of predictors is altered.

Measuring the Relevance of Predictors:

- maximum effect of a predictor variable on the response: $\beta_j \cdot (\max_i x_{ij} - \min_i x_{ij})$
- this can be compared to the total span in the response, or it can be plotted vs. the (logarithmic) p-value.

Mortality: Which Predictors Are Relevant?

```
> summary(fit.step)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1031.9491    80.2930  12.852  < 2e-16 ***
JanTemp       -2.0235     0.5145  -3.933  0.00025 ***
Rain           1.8117     0.5305   3.415  0.00125 **
Educ         -10.7463     7.0797  -1.518  0.13510
NonWhite       4.0401     0.6216   6.500  3.1e-08 ***
WhiteCollar   -1.4514     1.0451  -1.389  0.17082
log(Nox)      19.2481     4.5220   4.257  8.7e-05 ***
---
Residual standard error: 33.72 on 52 degrees of freedom
Multiple R-squared: 0.7383, Adjusted R-squared: 0.7081
F-statistic: 24.45 on 6 and 52 DF,  p-value: 1.543e-13
```

Mortality: Which Predictors Are Relevant?

Implementing the idea of maximum predictor effect:

```
> mami   <- function(col) max(col)-min(col)
> ranges <- apply(mort,2,mami)[c(2,5,6,8,9,14)]
> ranges
JanTemp    Rain  Educ  NonWhite  WhiteCollar  log.NOx
  55.00   55.00  3.30     37.70        28.40     5.77
>
> rele   <- abs(ranges*coef(fit.step)[-1])
> rele
JanTemp    Rain  Educ  NonWhite  WhiteCollar  log.NOx
 111.29   99.64 35.46    152.31        41.22   110.97
```

Predictor contributions are quite evenly distributed here. Maximum span in the response is **322.43**

What is a Good Model?

- The *true model* is a concept that exists in theory & simulation, but whether it does in practice remains unclear. Anyway, it is not realistic to identify the true model in observational studies.

- A **good model** is *useful* for the task at hand, *correct*ly describes the data without any systematical errors, has good *predictive* power and is *practical*/applicable for future use.

- Regression models in observational studies are *always only descriptive, but never causal*. A good model yields an accurate idea which of the observed variables drives the variation in the response, but not necessarily reveals the true mechanisms.