Applied Statistical Regression
AS 2012 – 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 26, 2012

Applied Statistical Regression
AS 2012 – Week 10

Cross Validation: Why?

- 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*. AIC/BIC and the F test try 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.*** This is also the only option for testing the performance of variable transformations.

Applied Statistical Regression
AS 2012 – Week 10

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) ...

10-Fold Cross Validation

Summary:

- Each observation is forecasted and compared against the true value 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
AS 2012 – 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
AS 2012 – Week 10

Cross Validation

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

```
> rss   <- c()
> fo    <- 10
> 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 2012 – Week 10

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
AS 2012 – Week 10

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
AS 2012 – Week 10

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
AS 2012 – Week 10

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
AS 2012 – Week 10

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
AS 2012 – Week 10

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 $\beta_j = 0$ never holds
- point null hypothesis is usually wrong in practice
- we just need enough data to be able to reject it

Applied Statistical Regression
AS 2012 – Week 10

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 reponse, or it can be plotted vs. the (logarithmic) p-value.

Applied Statistical Regression
AS 2012 – Week 10

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
```

Applied Statistical Regression
AS 2012 – Week 10

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**

Applied Statistical Regression
AS 2012 – 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!