Applied Statistical Regression
HS 2010 – Week 07

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 8, 2010

Applied Statistical Regression
HS 2010 – Week 07

Variable Selection: Why?

We want to fit a regression model…

Case 1: functional form and predictors exactly known
→ *estimation, test, confidence and prediction intervals*

Case 2: neither functional form nor the predictors are known
→ *explorative model search among potential predictors*

Case 3: we are interested in only 1 predictor, but want to correct for the effect of other covariates
→ *which covariates we need to correct for?*

Question in cases 2 & 3: WHICH PREDICTORS TO USE?

Applied Statistical Regression
HS 2010 – Week 07

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 2010 – Week 07

Method or Process?

- Please note that variable selection is not a method. It is a process that cannot even be separated from the rest of the analysis.
- For example, outliers and influential data points will not only change a particular model – they can even have an impact on the model we select. Also variable transformations will have an impact on the model that is selected.
- Some iteration and experimentation is often necessary for variable selection, i.e. to find smaller, but better models.

Applied Statistical Regression
HS 2010 – Week 07

Example

```
> head(state)
```

	Population	Income	Illiteracy	Life.Exp	Murder	HS.Grad	Frost	Area
AL	3615	3624	2.1	69.05	15.1	41.3	20	50708
AK	365	6315	1.5	69.31	11.3	66.7	152	566432
AZ	2212	4530	1.8	70.55	7.8	58.1	15	113417
AR	2110	3378	1.9	70.66	10.1	39.9	65	51945
CA	21198	5114	1.1	71.71	10.3	62.6	20	156361
CO	2541	4884	0.7	72.06	6.8	63.9	166	103766

Applied Statistical Regression
HS 2010 – Week 07

First-Aid Transformations

There are more transformations than the logged response model!

First-Aid Transformations:

→ do always apply these (if no practical reasons against it)

→ to both response and predictors

Absolute values and concentrations:

log-transformation: $y' = \log(y)$

Count data:

square-root transformation: $y' = \sqrt{y}$

Proportions:

arcsine transformation: $y' = \sin^{-1}\left(\sqrt{y}\right)$

Applied Statistical Regression
HS 2010 – Week 07

Example

```
> summary(lm(Life.Exp ~ ., data = state.trsf)

(Intercept)  6.878e+01  2.806e+00  24.511  < 2e-16 ***
Population   2.799e-01  1.238e-01   2.261   0.0290 *
Income      -5.601e-05  2.345e-04  -0.239   0.8124
Illiteracy  -5.885e-01  7.663e+00  -0.077   0.9392
Murder      -1.510e+00  2.188e-01  -6.905 1.99e-08 ***
HS.Grad      5.845e+00  2.458e+00   2.378   0.0220 *
Frost       -9.968e-02  4.821e-02  -2.067   0.0449 *
Area         3.361e-02  1.036e-01   0.325   0.7472
---
```

Applied Statistical Regression
HS 2010 – Week 07

Backward Elimination

→ **Removing more than one variable at a time is problematic**

- Start with the full model, and exclude the predictor with the highest p-value, as long as it exceeds α_{crit}
- Refit the model with the reduced predictor set. Again exclude the least significant predictor if it exceeds α_{crit}
- Repeat until all “non-significant” predictors are removed. Then, we stop and have found the final model.

→ Usually $\alpha_{crit} = 0.05$, for prediction also 0.15 or 0.20

→ **R-demo…**

Applied Statistical Regression
HS 2010 – Week 07

Backward Elimination: Final Result

```
> summary(lm(Life.Exp~Population + Murder + HS.Grad + Frost, data=...)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 68.78767    1.75860  39.115  < 2e-16 ***
Population   0.27663    0.10600   2.610 0.012259 *
Murder      -1.49218    0.17046  -8.754 2.83e-11 ***
HS.Grad      5.83746    1.37130   4.257 0.000104 ***
Frost       -0.09671    0.03669  -2.636 0.011477 *
---
Residual standard error: 0.6888 on 45 degrees of freedom
Multiple R-squared: 0.7582,   Adjusted R-squared: 0.7367
F-statistic: 35.28 on 4 and 45 DF,    p-value: 2.416e-13
```

Applied Statistical Regression
HS 2010 – Week 07

Interpretation of the Result

- The remaining predictors are now “more significant” than before. This is almost always the case…
- Do not overestimate the importance of these predictors!
- Collinearity among the predictors is one of the reasons for this artifact. The job is spread out among several predictors first.
- Important: the removed variables are still related to the response. If we consider them alone, they can even be significant.

Applied Statistical Regression
HS 2010 – Week 07

Forward Selection

→ This is an analogue to the backward elimination.

- Starts with an empty model, i.e. a model where only the intercept, but no predictors are present.
- We try all predictors one after the other and add the one which has the lowest p-value, provided it's below α_{crit}.
- Repeat until either all predictors are included in the model, or no further significant predictors can be found.

→ Feasible in situation where $p > n$

→ Computationally cheap, thus historically popular

Applied Statistical Regression
HS 2010 – Week 07

Stepwise Regression

→ This is mix between forward and backward selection.

- Can either start with an empty or a full model.
- We reconsider terms that were added or removed early.
- Can be based on individual hypothesis tests, too.

→ Often applied in practice
→ Default option in R-function step()
→ In practice, often based on AIC/BIC

Applied Statistical Regression
HS 2010 – Week 07

Testing Based Variable Selection

What are the drawbacks of the forward, backward or stepwise approach if based on hypothesis tests?

1) Missing the „best“ model
→ due to „one-at-a-time“ adding/dropping

2) Multiple testing problem
→ p-values should not be taken too literally

3) Missing link to final objective
→ hypothesis tests != prediction/explanation

4) Too small models
→ for prediction, we usually want bigger models

Applied Statistical Regression
HS 2010 – Week 07

All Subsets Regression

If there are *m* potential predictors, we can build 2^m models.

- complete, exhaustive search
- only feasible if *m* is reasonably small

We need a means of comparison for models of different size!

1) Coefficient of determination R^2
2) Test statistic or p-value of the global F-test
3) Estimated error variance $\hat{\sigma}_\varepsilon^2$

→ *measuring goodness-of-fit, increasing with more predictors*

Applied Statistical Regression
HS 2010 – Week 07

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:

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

Applied Statistical Regression
HS 2010 – Week 07

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.

Applied Statistical Regression
HS 2010 – Week 07

Example

```
> step(lm(Life.Exp ~ ., data=state.trsf))

Start:  AIC=-26.84

Life.Exp ~ Population+Income+Illiteracy+Murder+HS.Grad+Frost+Area

             Df Sum of Sq    RSS      AIC
- Illiteracy  1    0.0030 21.231 -28.8291
- Income      1    0.0288 21.256 -28.7682
- Area        1    0.0532 21.281 -28.7109
<none>                    21.228 -26.8361
- Frost       1    2.1603 23.388 -23.9903
- Population  1    2.5844 23.812 -23.0918
- HS.Grad     1    2.8591 24.087 -22.5183
- Murder      1   24.0982 45.326   9.0927
```

See R-demo

Applied Statistical Regression
HS 2010 – Week 07

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 2010 – Week 07

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 2010 – Week 07

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 mode, 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

Applied Statistical Regression
HS 2010 – Week 07

The Lasso

The result of any variable selection procedure is a subset of predictors that will be included into the model.

→ This is a random set. If the data were only slightly changed, or if we obtained another sample from the same population, that set might be completely different.

- **Arbitrary, non-continuous behavior**
- **Need for doing variable selection in a smoother way**
- The Lasso: a ***penalized regression approach***.

Applied Statistical Regression
HS 2010 – Week 07

The Lasso: Idea Behind

The idea behind is to complement the ordinary least squares criterion for model fitting with a penalty term for the magnitude of the coefficients. Thus, we minimize

$$Q(\beta, \lambda) = \sum_i r_i^2 + \lambda \sum_i |\beta_j|$$

→ Goodness-of-fit versus a penalty term.

Alternative formulation is possible:

Minimization of the sum of squared residuals under the condition that the sum of absolute values of the regression coefficients is: $\sum_i |\hat{\beta}_j| \leq c$

Applied Statistical Regression
HS 2010 – Week 07

The Lasso: Result