Variable selection and statistical strategy

Based on a book by Julian J. Faraway

University of Iceland

Data

We will continue to use the `savings` dataset.

```r
library(faraway) # you need to install the package first
data(savings)
```

The dataframe contains the following columns:

sr	savings rate - personal saving divided by disposable income
pop15	percent population under age of 15
pop75	percent population over age of 75
dpi	per-capita disposable income in dollars
ddpi	percent growth rate of dpi

Where are we...

1 Variable selection

2 Statistical strategy

Variable selection

Variable selection is intended to select the ”best” subset of predictors. But why bother?

- We want to explain the data in the simplest way - redundant predictors should be removed.
- Unnecessary predictors will add noise to the estimation of other quantities that we are interested in.
- Degrees of freedom will be wasted.
- Collinearity is caused by having too many variables trying to do the same job.
- Cost: if the model is to be used for prediction, we can save time and/or money by not measuring redundant predictors.

Variable selection

Prior to variable selection:

- Identify outliers and influential points - maybe exclude them at least temporarily.
- Add in any transformations of the variables that seem appropriate.

Hierarchical models

- Some models have a natural hierarchy. For example, in polynomial models, x^2 is a higher order term than x.
- When selecting variables, it is important to respect the hierarchy.
- Lower order terms should not be removed from the model before higher order terms in the same variable.
- There two common situations where this situation arises:
    - Polynomials models.
    - Models with interactions.

Backward elimination

This is the simplest of all variable selection procedures and can be easily implemented without special software. In situations where there is a complex hierarchy, backward elimination can be run manually while taking account of what variables are eligible for removal.

1. Start with all the predictors in the model
2. Remove the predictor with highest p-value greater than α_{crit}
3. Refit the model and goto 2
4. Stop when all p-values are less than α_{crit}.

The α_{crit} is sometimes called the ”p-to-remove” and does not have to be 5%. If prediction performance is the goal, then a 15-20% cut-off may work best, although methods designed more directly for optimal prediction should be preferred.

Backward elimination

```
f1<-lm(sr~pop15+pop75+dpi+ddpi,savings)
summary(f1)


##
## Call:
## lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2422 -2.6857 -0.2488  2.4280  9.7509
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
## pop15       -0.4611931  0.1446422  -3.189 0.002603 **
## pop75       -1.6914977  1.0835989  -1.561 0.125530
## dpi         -0.0003369  0.0009311  -0.362 0.719173
## ddpi         0.4096949  0.1961971   2.088 0.042471 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.803 on 45 degrees of freedom
## Multiple R-squared:  0.3385, Adjusted R-squared:  0.2797
## F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904


f2<-update(f1,~.-dpi)
# and so on...
```

Forward Selection

This just reverses the backward method.

1. Start with no variables in the model.
2. For all predictors not in the model, check their p-value if they are added to the model. Choose the one with lowest p-value less than α_{crit} .
3. Continue until no new predictors can be added.

Stepwise regression

- Stepwise regression is a combination of backward elimination and forward selection.
- This addresses the situation where variables are added or removed early in the process and we want to change our mind about them later.
- At each stage a variable may be added or removed and there are several variations on exactly how this is done.
- Stepwise procedures are relatively cheap computationally but they do have some drawbacks.

Stepwise Regression

Some drawbacks:

- The p-values used should not be treated too literally. There is so much multiple testing occurring that the validity is dubious.
- The procedures are not directly linked to final objectives of prediction or explanation and so may not really help solve the problem of interest.
- Stepwise variable selection tends to pick models that are smaller than desirable for prediction purposes: To give a simple example, consider the simple regression with just one predictor variable. Suppose that the slope for this predictor is not quite statistically significant. We might not have enough evidence to say that it is related to y but it still might be better to use it for predictive purposes.

Criterion-based procedures - AIC and BIC

The Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC) are commonly used criteria:

$$AIC = -2loglikelihood + 2p$$

$$BIC = -2loglikelihood + p\log(n)$$

for linear regression models:

$$-2loglikelihood = n\log(RSS/n)$$

Criterion-based procedures - AIC and BIC

- We want to minimize AIC or BIC.
- Larger models will fit better and so have smaller RSS but use more parameters.
- Thus the best choice of model will balance fit with model size.
- BIC penalizes larger models more heavily and so will tend to prefer smaller models in comparison to AIC.
- AIC and BIC can be used as selection criteria for other types of models too.

Criterion-based procedures - adjusted R^2

- Recall that R^2 = 1-RSS/TSS.
- Adding a variable to a model can only decrease the RSS and therefore only increase the R^2
- Therefore, R^2 by itself is not a good criterion because it would always choose the largest possible model.

Let us define the adjusted R^2 — called R_a^2 as

$$R_a^2 = 1 - \frac{RSS/(n-p)}{TSS/(n-1)} = 1 - \frac{\hat{\sigma}^2_{model}}{\hat{\sigma}^2_{null}}$$

Adding a predictor will only increase R_a^2 a if it has some value.

Criterion-based procedures - PRESS

- Predicted Residual Sum of Squares (PRESS) is defined as $\sum_i \hat{\varepsilon}^2_{(i)}$ where the $\hat{\varepsilon}^2_{(i)}$ are the residuals calculated without using case i in the fit.
- The model with the lowest PRESS criterion is then selected.
- This tends to pick larger models (which may be desirable if prediction is the objective).

Criterion-based procedures - Mallow's C_p Statistic

The C_p statistic is defined as:

$$C_p = \frac{RSS_P}{\hat{\sigma}^2} + 2p - n$$

where $\hat{\sigma}^2$ is from the model with all predictors and RSS_p indicates the RSS from a model with p parameters.

It is usual to plot C_p against p. We desire models with small p and C_p around or less than p.

Summary

- Variable selection is a means to an end and not an end itself.
- The aim is to construct a model that predicts well or explains the relationships in the data.
- Automatic variable selections are not guaranteed to be consistent with these goals - use these methods as a guide only.
- Stepwise methods use a restricted search through the space of potential models and use a dubious hypothesis testing based method for choosing between models.
- Criterion-based methods typically involve a wider search and compare models in a preferable manner - For this reason, they are recommended .

Summary

Accept the possibility that several models may be suggested which fit about as well as each other. If this happens, consider:

- Do the models have similar qualitative consequences?
- Do they make similar predictions?
- What is the cost of measuring the predictors?
- Which has the best diagnostics?

If you find models that seem roughly equally as good but lead to quite different conclusions then it is clear that the data cannot answer the question of interest unambiguously. Be alert to the danger that a model contradictory to the tentative conclusions might be out there.

Cross validation

- Test error: error that results from using a model to predict the response on a new observation, that is a measurement that was not used in the training the model.
- The test error can be easily calculated if a designated test set is available but unfortunately this is usually not the case.
    - Validation set approach
    - Leave on out cross validation
    - K-fold cross validation

Where are we...

1. Variable selection

2. Statistical strategy

Statistical strategy

Thus far we have learnt various tactics

- *Diagnostics:* Checking of assumptions: constant variance, linearity, normality, outliers, influential points, serial correlation and collinearity.
- *Transformation*: Transforming the response — Box-Cox, transforming the predictors - tests and polynomial regression.
- Variable selection: Stepwise and criterion based methods

What order should these be done in? Should procedures be repeated at later stages? When should we stop?

Statistical strategy

- A good strategy: $Diagnostics \rightarrow Transformation \rightarrow Variable\ Selection \rightarrow Diagnostics$
- However, regression analysis is a search for structure in data and there are no hard-and-fast rules about how it should be done.
- Regression analysis requires some skill - you must be alert to unexpected structure in the data.

Statistical strategy

- There is a danger of doing too much analysis.
- The more transformations and permutations of leaving out influential points you do, the better fitting model you will find.
- Torture the data long enough, and sooner or later it will confess.
- Remember that fitting the data well is no guarantee of good predictive performance or that the model is a good representation of the underlying population.

Statistical strategy

So:

- Avoid complex models for small datasets.
- Try to obtain new data to validate your proposed model. Some people set aside some of their existing data for this purpose.
- Use past experience with similar data to guide the choice of model.