ENVX2001 Practical Topic 9 Variable Selection

Willem Vervoort

Document generated on 2017-05-19

Objectives

- Learn to perform PARTIAL F-TESTS in R;
- Understand the difference between adjusted r^2 and r^2;
- Learn to perform FORWARD SELECTION and BACKWARD ELIMINATION in R.

DATA: *Data_Topic9_2017.xls*

EXERCISE 1 – MODELLING BIRD ABUNDANCE

Data: *Loyn2* worksheet, *2017_Loyn2.csv*

This worksheet has 6 predictor variables but note the 3 new names [L10AREA, L10DIST, L10LDIST]. They are the log10 transformed versions of AREA, DIST and LDIST. Each of these variables was log transformed as they had a small number of larger observations with high leverage. Import into R using `read.csv()`

```
Loyn2 <- read.csv("2017_Loyn2.csv")
names(Loyn2)
```

```
## [1] "ABUND"    "YR.ISOL"  "GRAZE"    "ALT"      "L10DIST"  "L10LDIST"
## [7] "L10AREA"
```

Use R for the following analysis

1. Obtain the correlation between ABUND and all of the 6 predictor variables using `cor()`. Based on these, what would you expect to be the best single predictor of ABUND?
2. Use multiple regression to see whether ABUND can be predicted from L10AREA and GRAZE. Is there a significant relationship? Are the assumptions met? We are using these 2 predictors as they have the largest absolute correlations. Use `lm()` and specify the model as `ABUND~L10AREA + GRAZE`.
3. How good is the model based on the (i) r^2 (ii) adjusted r^2? Use `summary()`.

4. From the variables which one is the most significant? (Refer to the t probabilities and p-values in the table of predictors in the output of `summary()`). Interpret the p-values in terms of dropping predictor variables. Interpret the impact of the variables on the predictant (ABUND).
5. Repeat the multiple regression, but this time include YRS.ISOL as a predictor variable (it has the 3rd largest absolute correlation). This will allow you to assess the effect of YRS.ISOL with the other variables taken into account. Compare the r^2 and adjusted r^2 values with those you calculated for the 2 predictor model.
6. Which is the better model? Why?

EXERCISE 2 - PARTIAL F-TESTS IN R

Data: *California streamflow worksheet*, *2017_Californiastreamflow.csv*

The following dataset contains 43 years of annual precipitation measurements (in mm) taken at (originally) 6 sites in the Owens Valley in California. I have reduced this to three variables labelled `L10APSAB` (Lake Sabrina), `L10OBPC` (Big Pine Creek), `L10OPRC` (Rock Creek), and the dependent variable stream runoff volume (measured in ML/year) at a site near Bishop, California (labelled `L10BSAAM`).

Note that I have made all the data log10 transformed to increase normality of the residuals in the regressions.

The purpose is to manually step through FORWARD SELECTION. Remember from the lectures that you basically have two methods of selecting variables: forward and backward. The philosophy behind this is that to fine the best model you can work from these two different directions:

1. start with nothing and slowly add more and more variables, checking everytime whether the addition of a variable actually improves the model.

2. start with a full model and slowly remove more and more variables, checking everytime whether removal of a variable actually improves the model.

(1) FORWARD SELECTION

This starts with nothing in the model so the 1st step is to add the most significant predictor variable. Create a simple linear regression model using `lm()` and `summary()` for each predictor variable and identify the most significant predictor. What measures

can you use to determine the most significant predictor? An alternative is to use the correlation matrix, `cor()`, and use the predictor with largest absolute correlation as the starting point. What does the regression tell you that the correlation matrix (using `cor()`) does not?

So the models to test are:

Table 1: Models to test

Model	Formula	Comments
Mod1	lm(L10BSAAM~L10APSAB,data=...)	test quality
Mod2	lm(L10BSAAM~L10OBPC,data=...)	test quality
Mod3	lm(L10BSAAM~L10OPRC,data=...)	test quality
ML.Mod1	lm(L10BSAAM~L10OPRC + L10OBPC,data=...)	F-test with Mod3, why?
ML.Mod2	lm(L10BSAAM~L10OPRC + L10OBPC + L10APSAB,data=...)	F-test with ML.Mod1, why?

First read in the data.

```
# read in the data
s.data <- read.csv("2017_Californiastreamflow.csv")
names(s.data)
```

```
## [1] "L10APSAB" "L10OBPC"  "L10OPRC"  "L10BSAAM"
```

(2) Partial F-Tests

The above analysis should tell that you need `L10OPRC` in the model, what should we add next? This involves performing **PARTIAL F-TESTS** as discussed in the lecture. This can be done in **R** by using `anova()` on two model objects. So you have to make objects of all the possible model combinations.

a. The Regression analysis output gives the test for adding L10OBPC.

b. The last row gives the results of the partial F-test. Should we add L10OBPC to the model?

c. Write out the hypotheses you are testing.

Perform a PARTIAL F-TEST to work out if the addition of L10APSAB improves upon the model with just L10OPRC.

d. Which variable should be added to the model containing L10OPRC?

(3) 3 variable model

Based on (2) we either added or did not add an additional predictor to the model with L10OPRC. If we did not find one of the variables to be significantly improve the single variable model then we stop now. If you did add a 2nd predictor, then perform a PARTIAL F-TEST to test whether the 3 variable model is superior.

e. What is your optimal model?

EXERCISE 3 - FORWARD SELECTION AND BACKWARD ELIMINATION IN R

Data: *Dippers* worksheet, *2017_Dippers.csv*

This is dataset we used in the tutorial in Topic 8. The data has been transformed for some of the variables.
The file, Breeding density of dippers, gives data from a biological survey which examined the nature of the variables thought to influence the breeding of British dippers (thrush-sized birds living mainly in the upper reaches of rivers; they feed on benthic invertebrates by probing the river beds with their beaks). Twenty two sites were included in the survey. The variables measured were:

- site altitude
- water hardness
- river-bed slope
- the numbers of caddis fly larvae
- the numbers of stonefly larvae
- the numbers of mayfly larvae
- the numbers of all other invertebrates collected
- the number of breeding pairs of dippers per 10 km of river

In the analyses, the four invertebrate variables were transformed using a Log(Number+1) transformation.

```
Dippers <- read.csv("2017_Dippers.csv")
names(Dippers)
```

```
## [1] "Altitude"   "Hardness"   "RiverSlope" "Br_Dens"    "LogCadd"
## [6] "LogStone"   "LogMay"     "LogOther"
```

Using the information on the lecture slides, perform a backward and forward selection starting from a maximal model:
`MaxMod <- lm(Br_Dens ~ ., data=Dippers)`
and defining the minimum model as:
`MinMod <- lm(Br_Dens~1,data=Dippers)`

Do both approaches identify the same model? If not, which model would you choose? And how would you choose this? Think about AIC and adj r-squared.

To make this easier, assign both the backwards and forwards step analysis to an R object. You should then be able to run `summary()` on the two objects to compare the final models.

END OF PRACTICAL