Model selection

Module 4 considers 3 issues with linear models:

1) Model selection (choose between different models fitted to our data)
2) Non-linear relationship (between the response variable and independent variables)
3) Relationship between common tests (*t*-test, ANOVA, ANCOVA) and multiple linear regression

Setup

Consider the problem of choosing a suitable model given data

$$(y_1, \overset{(k\times 1)}{\boldsymbol{x}_1}), (y_2, \overset{(k\times 1)}{\boldsymbol{x}_2}), \dots, (y_n, \overset{(k\times 1)}{\boldsymbol{x}_n})$$

where $\boldsymbol{x}$ represents the vector of predictor variables.

not all predictors are informative

We want to find the 'best' subset of predictors for predicting Y.

Selecting predictors

In the process of selecting the smallest 'best-fitting' model, we are confronted with two contradictory criteria:

- Exclusion of important terms clearly leads to an incorrect model, which can lead to misleading conclusions
  (underspecified model)
- Inclusion of unnecessary terms diminishes the value of the model as a simplification of the data, and also reduces the statistical accuracy of parameter estimates and predictions.
  (overspecified model)

Model selection

How do we decide on the best model for our data?

Two parts:

- Choice of procedure
- Choice of criteria

Exhaustive selection:

- search through all subsets of $\{x_1, \ldots, x_k\}$
- evaluate the models constructed from all subsets
- find the model that optimise our chosen criterion

While exhaustive selection will give good results, it is usually too time consuming. If we have k predictors, then we have 2^k possible subsets.

We will consider three commonly used automated methods that will progressively build our optimal subset of predictors:

- forward selection
- backward selection
- stepwise selection

The forward selection algorithm with P-values

1. Begin with the null model. (no predictors, only intercept term)

2. For every term not currently included in the model, calculate a P-value for the inclusion of that term.

3. If the smallest P-value is less than the threshold p_{in} (usually chosen to be $\mathbf{0.05}$), add that term to the model.

4. Iterate (2), (3) until no further terms are significant.

Example 4.1

The marks data contains the assignment and quiz scores (in percentage) of 339 students in a Statistics course.
Suppose we are interested in the following variables:

- `E` (response) exam mark
- `OQ` online quiz
- `A1`
- `A2`
- `A3` assignments
- `A4`
- `A5`
- `A6`

Fit a multiple linear regression to the data using forward selection.

Example 4.1 Solution

```
marks <- read.csv("marks.csv")
null <- lm(E ~ 1, data=marks)
scope <- E ~ OQ + A1 + A2 + A3 + A4 + A5 + A6
add1(null, scope = scope, test = "F")
```

(1) (2)

(scope defines the range of models to be examined in the search)

```
## Single term additions
##
## Model:
## E ~ 1
##        Df Sum of Sq    RSS      AIC F value    Pr(>F)
## <none>               21.155  -938.43
## OQ      1    7.1258 14.029 -1075.67 171.175 < 2.2e-16 ***
## A1      1    1.0852 20.070  -954.28  18.223 2.558e-05 ***
## A2      1    1.7407 19.414  -965.54  30.215 7.644e-08 ***
## A3      1    4.2472 16.908 -1012.40  84.654 < 2.2e-16 ***
## A4      1    7.1621 13.993 -1076.55 172.492 < 2.2e-16 ***
## A5      1    6.9001 14.255 -1070.26 163.129 < 2.2e-16 ***
## A6      1    9.3016 11.853 -1132.80 264.456 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(3)

add A6 to our model

1) Start with null model (i.e. with no predictors)
2) Fit 7 different linear models, each with an intercept and one predictor
3) For each model, perform an F–test to compare it with the null model
4) Find the model with the smallest P-value and add the corresponding predictor to our model

(4) Since OQ, A3, A4, A5, A6 all have p-value $< 2.2 \times 10^{-16}$, we will need to look at the F test statistic instead. A higher F value gives a lower P-value. Hence, A6 is chosen.

Example 4.1 Solution

(1)
(2)

```
fs1 <- update(null, .~. + A6)
add1(fs1, scope = scope, test = "F")
```

(3)

```
## Single term additions
##
## Model:
## E ~ A6
##        Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>               11.853 -1132.8
## OQ      1   1.12811 10.725 -1164.7 35.3421 6.929e-09 ***
## A1      1   0.08001 11.773 -1133.1  2.2834   0.13170
## A2      1   0.09538 11.758 -1133.5  2.7255   0.09969 .
## A3      1   0.56043 11.293 -1147.2 16.6749 5.550e-05 ***
## A4      1   0.55393 11.299 -1147.0 16.4720 6.146e-05 ***
## A5      1   0.32088 11.532 -1140.1  9.3489   0.00241 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1) Update our model with A6 included as a predictor.
2) Fit 6 different linear models, each with an additional predictor
3) For each model, perform an F–test to compare it with the null model
4) Find the model with the smallest P-value and use this model

(4) add OQ to our model

Example 4.1 Solution

```r
fs2 <- update(fs1, .~. + OQ)
add1(fs2, scope = scope, test = "F")
```

```
## Single term additions
##
## Model:
## E ~ A6 + OQ
##        Df Sum of Sq    RSS     AIC F value   Pr(>F)
## <none>              10.725 -1164.7
## A1      1   0.03104 10.694 -1163.7  0.9725 0.324774
## A2      1   0.02419 10.701 -1163.5  0.7573 0.384812
## A3      1   0.33372 10.391 -1173.4 10.7586 0.001147 **
## A4      1   0.18839 10.537 -1168.7  5.9895 0.014904 *
## A5      1   0.09645 10.629 -1165.8  3.0401 0.082150 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

add A3 to our model

Example 4.1 Solution

```
fs3 <- update(fs2, .~. + A3)
add1(fs3, scope = scope, test = "F")
```

```
## Single term additions
##
## Model:
## E ~ A6 + OQ + A3
##        Df Sum of Sq    RSS     AIC F value Pr(>F)
## <none>              10.391 -1173.4
## A1      1  0.000108 10.391 -1171.4  0.0035 0.9530
## A2      1  0.006228 10.385 -1171.6  0.2003 0.6548
## A4      1  0.070882 10.320 -1173.8  2.2939 0.1308
## A5      1  0.039884 10.351 -1172.7  1.2869 0.2574
```

None of the P-values are below our threshold (0.05). We can stop our algorithm.

This becomes our final model.

Example 4.1 Solution

```
Summary(fs3)
```

```
##
## Call:
## lm(formula = E ~ A6 + OQ + A3, data = stats_marks)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.81856 -0.06018  0.02859  0.09063  0.60694
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.13219    0.03273   4.039 6.65e-05 ***
## A6           0.36301    0.04104   8.845  < 2e-16 ***
## OQ           0.20085    0.03726   5.391 1.33e-07 ***
## A3           0.14387    0.04386   3.280  0.00115 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1761 on 335 degrees of freedom
## Multiple R-squared:  0.5088, Adjusted R-squared:  0.5044
## F-statistic: 115.7 on 3 and 335 DF,  p-value: < 2.2e-16
```