The stepwise selection algorithm with P-values

a combination of forward and backward selection

1. Begin with the null model.

repeat

2. Perform one step of forward selection using a liberal value of p_{in} such as $\mathbf{0.2}$ or $\mathbf{0.15}$.
3. Perform one step of backward elimination with a value of P_{out} such as $\mathbf{0.05}$.
4. Iterate (2), (3) until no further changes occur or the algorithm cycles.

Example 4.3

Consider again the marks data in Example 4.1.

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

Example 4.3 Solution

1

```
add1(null, scope = scope, test = "F")
```

Start with one forward selection step.
As in Example 4.1, the predictor A6 is the corresponding predictor with the smallest (significant) P-value.

```
## 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
```

add A6 to our model

Example 4.3 Solution

```
ss1 <- update(null, .~. +A6)
drop1(ss1, test = "F")
```

2

Then perform one backward selection step.
The only predictor A6 has a significant P-value.
Hence we cannot remove it from our model.
No change to our model is made in this step.

```
## Single term deletions
##
## Model:
## E ~ A6
##        Df Sum of Sq    RSS      AIC F value    Pr(>F)
## <none>              11.853 -1132.80
## A6      1    9.3016 21.155  -938.43  264.46 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A6 is significant
(p-value < threshold)

keep A6

Example 4.3 Solution

3

```
add1(ss1, scope=scope, test = "F")
```

We now do another forward selection step.
The predictor OQ is the corresponding predictor with the smallest P-value. Hence we add A6 to our model.

```
## 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
```

add OQ to our model

Example 4.3 Solution

4

```
ss2 <- update(ss1, .~. + OQ)
drop1(ss2, test = "F")
```

Next is a backward selection step.

```
## Single term deletions
##
## Model:
## E ~ A6 + OQ
##        Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>               10.725 -1164.7
## A6      1    3.3039 14.029 -1075.7 103.506 < 2.2e-16 ***
## OQ      1    1.1281 11.853 -1132.8  35.342 6.929e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both A6 and OQ are significant, we will keep them in the model.

Example 4.3 Solution

5

```
add1(ss2, scope=scope, test = "F")
```

Another forward selection step.
The predictor A3 is the corresponding predictor with the smallest (significant) P-value.

```
## 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

Example 4.3 Solution

```
ss3 <- update(ss2, .~. + A3)
drop1(ss3, test = "F")
```

6

```
## Single term deletions
##
## Model:
## E ~ A6 + OQ + A3
##        Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>               10.391 -1173.4
## A6      1   2.42664 12.818 -1104.3  78.231 < 2.2e-16 ***
## OQ      1   0.90140 11.293 -1147.2  29.060 1.327e-07 ***
## A3      1   0.33372 10.725 -1164.7  10.759  0.001147 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this backward selection step, again all predictors are significant. So no change is made in this step.

cannot drop any predictors

Example 4.3 Solution

7

```
add1(ss3, scope=scope, test = "F")
```

In this forward selection step, all the remaining predictors are non-significant.

```
## 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
```

no predictors can be added

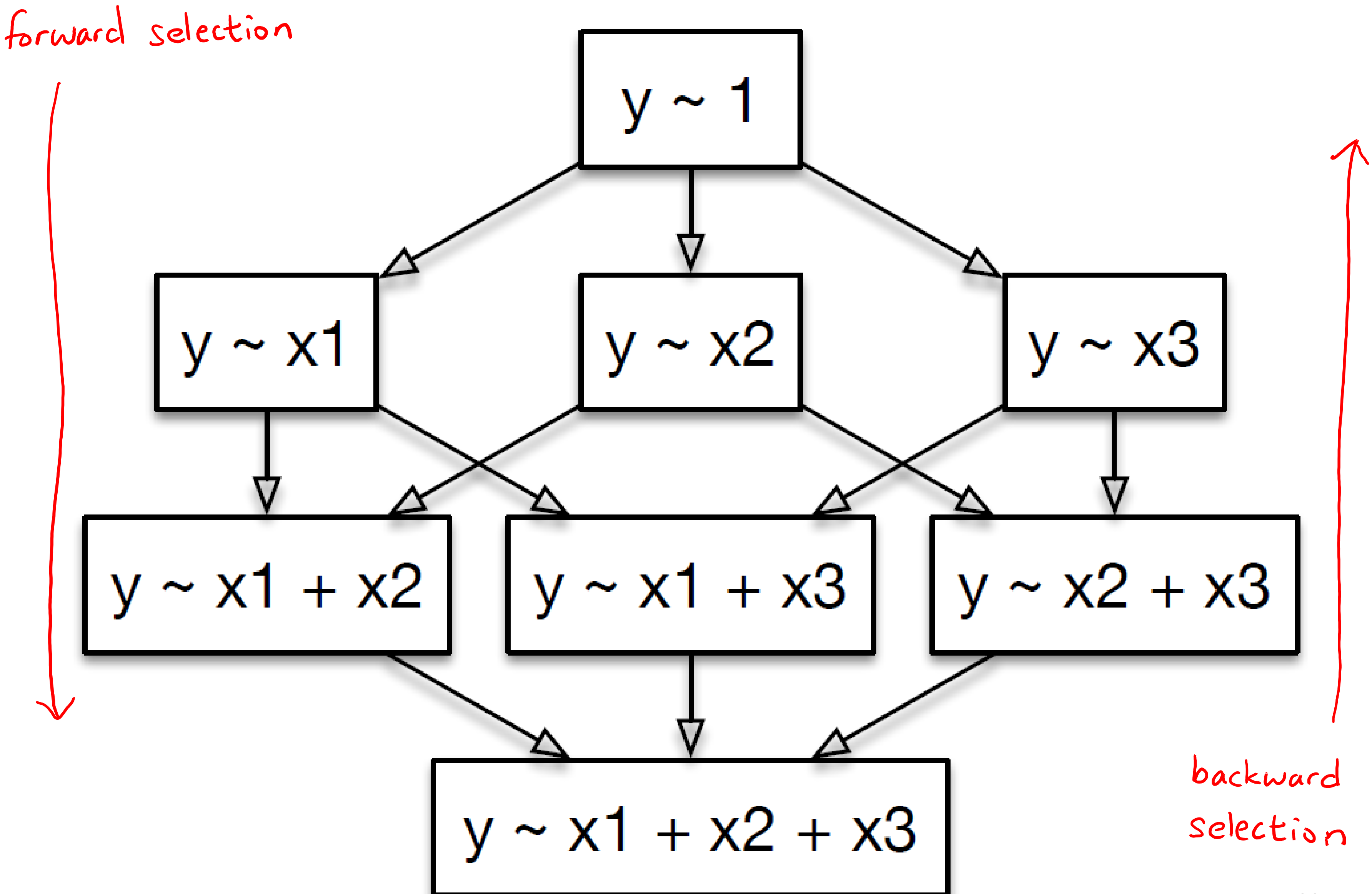
As both the preceding backward step and this forward step make no changes to the model, we can stop our algorithm. This becomes our final model.

Example 4.3 Solution

```
summary(ss3)
```

```
##
## 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
```

Comparison of methods



Principle of marginality

Whenever an interaction term is included in the model, all implied lower order interactions and main effects must also be included.

For example, if we find that we have an interaction term $x_{i1}x_{i2}$ in the model, then we must keep the main effects x_{i1} and x_{i2} in the model.

Some further remarks:

- None of the three methods is guaranteed to produce the "best" model
- They do not necessarily give the same subset of predictors

- In stepwise selection, we can use different p_{in} and p_{out}.
  - If $p_{out} < p_{in}$, we may remove predictors that were added before
  - If $p_{out} > p_{in}$, then this provides some 'protection' for the predictors that were added to the model

- Weakness of forward selection
  - Typically begins with an incorrect model, hence significance calculations may be wrong
  - Terms included in the model that may become insignificant later cannot be removed
  - Can fail to detect predictors that are jointly, but not separately, significant

- Weakness of backward selection
    - Assumes the initial model is correct. If this is not correct, we cannot add additional predictors.
    - Can be impractical if there are many predictors

- Stepwise selection combines the good features of forward and backward selection, but still not fool-proof