6.2 Model Selection: Testing-based procedures

There are two ways that we work when we do testing-based model selection. We either do *backward elimination* or *forward selection*.

Backward elimination

1. Start with all the predictors in the model.
2. Remove the predictor with highest $p-\text{value} > \alpha_0$ (most insignificant).
3. Refit the model, and repeat the above process.
4. **Stop** when all $p-\text{values} \leq \alpha_0$.

α_0 is often set to 15% or 20% which is higher than usual

Let's try this with the `Birthweight` example:

Birthweight Example

We start by fitting the full model.

```
birthweight.full = lm(Birthweight~., data=birthweight2)
```

We look at the `summary` and remove the variable with the highest p-value. We keep going until all variables in the model are statistically significant. Here we will use α=15%.

```
summary(birthweight.full)
```

```
##
## Call:
## lm(formula = Birthweight ~ ., data = birthweight2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38656 -0.26722 -0.06068  0.18271  0.60295
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.4286417  2.1583202  -1.589  0.12339
## Length       0.0257860  0.0336226   0.767  0.44954
## Headcirc     0.0850933  0.0297843   2.857  0.00798 **
## Gestation    0.0916226  0.0322518   2.841  0.00829 **
## smoker      -0.2198237  0.1728531  -1.272  0.21393
## mage        -0.0158203  0.0191605  -0.826  0.41597
## mnocig       0.0002094  0.0070011   0.030  0.97635
## mheight      0.0056438  0.0143947   0.392  0.69797
## mppwt        0.0084338  0.0116016   0.727  0.47329
## fage         0.0046535  0.0167998   0.277  0.78382
## fedyrs       0.0016448  0.0312497   0.053  0.95840
## fnocig       0.0040531  0.0041251   0.983  0.33424
## fheight     -0.0122665  0.0097854  -1.254  0.22037
## lowbwt      -0.1751779  0.2346936  -0.746  0.46164
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3367 on 28 degrees of freedom
## Multiple R-squared:  0.7877, Adjusted R-squared:  0.6891
## F-statistic: 7.989 on 13 and 28 DF,  p-value: 2.432e-06
```

`mnocig` has the highest p-value, so we remove it first:

```r
model1 = update(birthweight.full,.~.-mnocig)
summary(model1)
```

```
##
## Call:
## lm(formula = Birthweight ~ Length + Headcirc + Gestation + smoker +
##     mage + mheight + mppwt + fage + fedyrs + fnocig + fheight +
##     lowbwt, data = birthweight2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38565 -0.26728 -0.06041  0.18093  0.60331
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.431583   2.118612  -1.620  0.11612
## Length       0.025658   0.032771   0.783  0.44000
## Headcirc     0.084968   0.028976   2.932  0.00651 **
## Gestation    0.091783   0.031249   2.937  0.00643 **
## smoker      -0.216211   0.121501  -1.779  0.08564 .
## mage        -0.015723   0.018554  -0.847  0.40370
## mheight      0.005720   0.013923   0.411  0.68424
## mppwt        0.008423   0.011394   0.739  0.46572
## fage         0.004602   0.016422   0.280  0.78128
## fedyrs       0.001834   0.030069   0.061  0.95178
## fnocig       0.004067   0.004027   1.010  0.32084
## fheight     -0.012310   0.009507  -1.295  0.20558
## lowbwt      -0.176326   0.227510  -0.775  0.44460
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3309 on 29 degrees of freedom
## Multiple R-squared:  0.7876, Adjusted R-squared:  0.6998
## F-statistic: 8.964 on 12 and 29 DF,  p-value: 7.791e-07
```

Remove `fedyrs` next which is the variable with the highest p-value

```
model2 = update(model1,.~.-fedyrs)
summary(model2)
```

```
##
## Call:
## lm(formula = Birthweight ~ Length + Headcirc + Gestation + smoker +
##     mage + mheight + mppwt + fage + fnocig + fheight + lowbwt,
##     data = birthweight2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38561 -0.26811 -0.06193  0.18251  0.60062
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.426270   2.081375  -1.646  0.11017
## Length       0.025542   0.032168   0.794  0.43341
## Headcirc     0.085086   0.028428   2.993  0.00549 **
## Gestation    0.091805   0.030723   2.988  0.00555 **
## smoker      -0.215874   0.119343  -1.809  0.08051 .
## mage        -0.015284   0.016815  -0.909  0.37062
## mheight      0.005613   0.013582   0.413  0.68235
## mppwt        0.008496   0.011141   0.763  0.45163
## fage         0.004495   0.016055   0.280  0.78140
## fnocig       0.003985   0.003730   1.068  0.29389
## fheight     -0.012159   0.009022  -1.348  0.18786
## lowbwt      -0.177772   0.222482  -0.799  0.43055
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3254 on 30 degrees of freedom
## Multiple R-squared:  0.7876, Adjusted R-squared:  0.7097
## F-statistic: 10.11 on 11 and 30 DF,  p-value: 2.351e-07
```

Remove `mheight` next which is the variable with the highest p-value

```r
model3 = update(model2,.~.-mheight)
summary(model3)
```

```
##
## Call:
## lm(formula = Birthweight ~ Length + Headcirc + Gestation + smoker +
##     mage + mppwt + fage + fnocig + fheight + lowbwt, data = birthweight2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39449 -0.26988 -0.07323  0.18049  0.60128
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.971497   1.742901  -1.705  0.09821 .
## Length       0.030125   0.029789   1.011  0.31971
## Headcirc     0.087781   0.027297   3.216  0.00304 **
## Gestation    0.089553   0.029829   3.002  0.00526 **
## smoker      -0.208271   0.116329  -1.790  0.08317 .
## mage        -0.013914   0.016263  -0.856  0.39880
## mppwt        0.011610   0.008097   1.434  0.16162
## fage         0.002093   0.014764   0.142  0.88818
## fnocig       0.003859   0.003667   1.052  0.30082
## fheight     -0.011715   0.008837  -1.326  0.19465
## lowbwt      -0.166312   0.217775  -0.764  0.45083
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.321 on 31 degrees of freedom
## Multiple R-squared:  0.7864, Adjusted R-squared:  0.7175
## F-statistic: 11.41 on 10 and 31 DF,  p-value: 7.206e-08
```

Remove `fage` next which is the variable with the highest p-value

```
model4 = update(model3,.~.-fage)
summary(model4)
```

```
##
## Call:
## lm(formula = Birthweight ~ Length + Headcirc + Gestation + smoker +
##     mage + mppwt + fnocig + fheight + lowbwt, data = birthweight2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38612 -0.25785 -0.06499  0.17795  0.59520
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.917772   1.674957  -1.742  0.09111 .
## Length       0.029189   0.028600   1.021  0.31510
## Headcirc     0.088943   0.025635   3.470  0.00151 **
## Gestation    0.090344   0.028849   3.132  0.00370 **
## smoker      -0.206735   0.114036  -1.813  0.07924 .
## mage        -0.012086   0.009753  -1.239  0.22427
## mppwt        0.011566   0.007966   1.452  0.15626
## fnocig       0.004009   0.003456   1.160  0.25459
## fheight     -0.012064   0.008358  -1.443  0.15861
## lowbwt      -0.173820   0.207978  -0.836  0.40949
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.316 on 32 degrees of freedom
## Multiple R-squared:  0.7863, Adjusted R-squared:  0.7262
## F-statistic: 13.08 on 9 and 32 DF,  p-value: 1.921e-08
```

Remove `lowbwt` next which is the variable with the highest p-value

```
model5 = update(model4,.~.-lowbwt)
summary(model5)
```

```
##
## Call:
## lm(formula = Birthweight ~ Length + Headcirc + Gestation + smoker +
##     mage + mppwt + fnocig + fheight, data = birthweight2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40190 -0.25534 -0.05499  0.22033  0.58437
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.388028   1.570397  -2.157 0.038349 *
## Length       0.035225   0.027546   1.279 0.209902
## Headcirc     0.091217   0.025374   3.595 0.001045 **
## Gestation    0.098331   0.027096   3.629 0.000951 ***
## smoker      -0.216332   0.112937  -1.916 0.064127 .
## mage        -0.012248   0.009706  -1.262 0.215835
## mppwt        0.012894   0.007770   1.659 0.106523
## fnocig       0.003530   0.003393   1.040 0.305674
## fheight     -0.013809   0.008055  -1.714 0.095866 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3146 on 33 degrees of freedom
## Multiple R-squared:  0.7816, Adjusted R-squared:  0.7287
## F-statistic: 14.76 on 8 and 33 DF,  p-value: 6.6e-09
```

Remove `fnocig` next which is the variable with the highest p-value

```
model6 = update(model5,.~.-fnocig)
summary(model6)
```

```
##
## Call:
## lm(formula = Birthweight ~ Length + Headcirc + Gestation + smoker +
##     mage + mppwt + fheight, data = birthweight2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44366 -0.24216 -0.06831  0.17653  0.56236
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.825512   1.514901  -2.525  0.01639 *
## Length       0.039141   0.027321   1.433  0.16110
## Headcirc     0.091350   0.025404   3.596  0.00101 **
## Gestation    0.091819   0.026395   3.479  0.00140 **
## smoker      -0.171334   0.104456  -1.640  0.11017
## mage        -0.011579   0.009696  -1.194  0.24067
## mppwt        0.012943   0.007780   1.664  0.10535
## fheight     -0.011015   0.007604  -1.449  0.15661
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.315 on 34 degrees of freedom
## Multiple R-squared:  0.7744, Adjusted R-squared:  0.728
## F-statistic: 16.68 on 7 and 34 DF,  p-value: 2.546e-09
```

Remove `mage` next which is the variable with the highest p-value

```
model7 = update(model6,.~.-mage)
summary(model7)
```

```
##
## Call:
## lm(formula = Birthweight ~ Length + Headcirc + Gestation + smoker +
##     mppwt + fheight, data = birthweight2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47853 -0.23664 -0.05791  0.21743  0.61059
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.291512   1.472662  -2.914  0.00618 **
## Length       0.038566   0.027483   1.403  0.16934
## Headcirc     0.087763   0.025379   3.458  0.00145 **
## Gestation    0.093041   0.026535   3.506  0.00127 **
## smoker      -0.205576   0.101053  -2.034  0.04955 *
## mppwt        0.010571   0.007567   1.397  0.17124
## fheight     -0.008631   0.007382  -1.169  0.25019
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3169 on 35 degrees of freedom
## Multiple R-squared:  0.765,  Adjusted R-squared:  0.7247
## F-statistic: 18.99 on 6 and 35 DF,  p-value: 1.077e-09
```

Remove `fheight` next which is the variable with the highest p-value

```
model8 = update(model7,.~.-fheight)
summary(model8)
```

```
##
## Call:
## lm(formula = Birthweight ~ Length + Headcirc + Gestation + smoker +
##     mppwt, data = birthweight2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61303 -0.22377 -0.02461  0.21675  0.58663
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.607343   0.954751  -5.873 1.03e-06 ***
## Length       0.034615   0.027413   1.263  0.21480
## Headcirc     0.090151   0.025425   3.546  0.00111 **
## Gestation    0.090323   0.026568   3.400  0.00166 **
## smoker      -0.221552   0.100634  -2.202  0.03419 *
## mppwt        0.010448   0.007605   1.374  0.17800
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3185 on 36 degrees of freedom
## Multiple R-squared:  0.7558, Adjusted R-squared:  0.7219
## F-statistic: 22.28 on 5 and 36 DF,  p-value: 4.084e-10
```

Remove `Length` next which is the variable with the highest p-value

```
model9 = update(model8,.~.-Length)
summary(model9)
```

```
##
## Call:
## lm(formula = Birthweight ~ Headcirc + Gestation + smoker + mppwt,
##     data = birthweight2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.64164 -0.20889 -0.00633  0.22222  0.64842
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.195901   0.904612  -5.744 1.40e-06 ***
## Headcirc     0.102085   0.023793   4.291 0.000123 ***
## Gestation    0.111149   0.020996   5.294 5.67e-06 ***
## smoker      -0.231712   0.101114  -2.292 0.027720 *
## mppwt        0.012913   0.007409   1.743 0.089652 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.321 on 37 degrees of freedom
## Multiple R-squared:  0.745,  Adjusted R-squared:  0.7174
## F-statistic: 27.02 on 4 and 37 DF,  p-value: 1.554e-10
```

Now, we stop because all the p-values are less than the 10% threshold, and the selected model includes:

$$\text{Headcirc, Gestation, smoker, mppwt}$$

The opposite approach here is the forward elimination.

Forward elimination

1. Start with the intercept-only model.

2. For all predictors not in the model, check their p-value if being added to the model. Add the one with the lowest $p-\text{value} \leq \alpha_0$ (most significant).
3. Refit the model, and repeat the above process.
4. **Stop** when no more predictors can be added.

Let's try this with the `Birthweight` example:

Birthweight Example

Step 1 We start by fitting SLR with each one of the variables, and we look at the p-value of the slope of each one.

```
summary(lm(Birthweight ~ Length, birthweight2))$coef[2,4]
```

```
## [1] 5.029346e-08
```

```
summary(lm(Birthweight ~ Headcirc, birthweight2))$coef[2,4]
```

```
## [1] 5.734798e-07
```

```
summary(lm(Birthweight ~ Gestation, birthweight2))$coef[2,4]
```

```
## [1] 1.542295e-07
```

```
summary(lm(Birthweight ~ smoker, birthweight2))$coef[2,4]
```

```
## [1] 0.04269625
```

```r
summary(lm(Birthweight ~ mage, birthweight2))$coef[2,4]
```

```
## [1] 0.9991319
```

```r
summary(lm(Birthweight ~ mnocig, birthweight2))$coef[2,4]
```

```
## [1] 0.3355036
```

```r
summary(lm(Birthweight ~ mheight, birthweight2))$coef[2,4]
```

```
## [1] 0.01812163
```

```r
summary(lm(Birthweight ~ mppwt, birthweight2))$coef[2,4]
```

```
## [1] 0.008513416
```

```r
summary(lm(Birthweight ~ fage, birthweight2))$coef[2,4]
```

```
## [1] 0.2656859
```

```r
summary(lm(Birthweight ~ fedyrs, birthweight2))$coef[2,4]
```

```
## [1] 0.6548055
```

```r
summary(lm(Birthweight ~ fnocig, birthweight2))$coef[2,4]
```

```
## [1] 0.5574383
```

```
summary(lm(Birthweight ~ fheight, birthweight2))$coef[2,4]
```

```
## [1] 0.8453709
```

```
summary(lm(Birthweight ~ lowbwt, birthweight2))$coef[2,4]
```

```
## [1] 2.90685e-06
```

The variable with the lowest p-value is the model with `Length`, so we add it first in the model. Then, we check the model with `Length` with each one of the variables added, and repeat until no significant variables can be added.

```
summary(lm(Birthweight ~ Length+Headcirc, birthweight2))$coef[3,4]
```

```
## [1] 0.001308298
```

```
summary(lm(Birthweight ~ Length+Gestation, birthweight2))$coef[3,4]
```

```
## [1] 0.009167711
```

```
summary(lm(Birthweight ~ Length+smoker, birthweight2))$coef[3,4]
```

```
## [1] 0.05781097
```

```
summary(lm(Birthweight ~ Length+mage, birthweight2))$coef[3,4]
```

```
## [1] 0.6206798
```

```
summary(lm(Birthweight ~ Length+mnocig, birthweight2))$coef[3,4]
```

```
## [1] 0.2607006
```

```
summary(lm(Birthweight ~ Length+mheight, birthweight2))$coef[3,4]
```

```
## [1] 0.9132313
```

```
summary(lm(Birthweight ~ Length+mppwt, birthweight2))$coef[3,4]
```

```
## [1] 0.2684809
```

```
summary(lm(Birthweight ~ Length+fage, birthweight2))$coef[3,4]
```

```
## [1] 0.4868267
```

```
summary(lm(Birthweight ~ Length+fedyrs, birthweight2))$coef[3,4]
```

```
## [1] 0.9042375
```

```
summary(lm(Birthweight ~ Length+fnocig, birthweight2))$coef[3,4]
```

```
## [1] 0.3659651
```

```
summary(lm(Birthweight ~ Length+fheight, birthweight2))$coef[3,4]
```

```
## [1] 0.2620871
```

```
summary(lm(Birthweight ~ Length+lowbwt, birthweight2))$coef[3,4]
```

```
## [1] 0.01335802
```

The model with the lowest p-value is the model with `Length` and `Headcirc`, so `Headcirc` enter the model second.

```
summary(lm(Birthweight ~ Length+Headcirc+Gestation, birthweight2))$coef[4,4]
```

```
## [1] 0.003359556
```

```
summary(lm(Birthweight ~ Length+Headcirc+smoker, birthweight2))$coef[4,4]
```

```
## [1] 0.08240424
```

```
summary(lm(Birthweight ~ Length+Headcirc+mage, birthweight2))$coef[4,4]
```

```
## [1] 0.3180101
```

```
summary(lm(Birthweight ~ Length+Headcirc+mnocig, birthweight2))$coef[4,4]
```

```
## [1] 0.4150132
```

```r
summary(lm(Birthweight ~ Length+Headcirc+mheight, birthweight2))$coef[4,4]
```

```
## [1] 0.8585754
```

```r
summary(lm(Birthweight ~ Length+Headcirc+mppwt, birthweight2))$coef[4,4]
```

```
## [1] 0.3700732
```

```r
summary(lm(Birthweight ~ Length+Headcirc+fage, birthweight2))$coef[4,4]
```

```
## [1] 0.8786758
```

```r
summary(lm(Birthweight ~ Length+Headcirc+fedyrs, birthweight2))$coef[4,4]
```

```
## [1] 0.8481611
```

```r
summary(lm(Birthweight ~ Length+Headcirc+fnocig, birthweight2))$coef[4,4]
```

```
## [1] 0.4196266
```

```r
summary(lm(Birthweight ~ Length+Headcirc+fheight, birthweight2))$coef[4,4]
```

```
## [1] 0.3439151
```

```r
summary(lm(Birthweight ~ Length+Headcirc+lowbwt, birthweight2))$coef[4,4]
```

```
## [1] 0.02446704
```

`Gestation` is next.

```
summary(lm(Birthweight ~ Length+Headcirc+Gestation+smoker, birthweight2))$coef[5,
```

```
## [1] 0.04531073
```

```
summary(lm(Birthweight ~ Length+Headcirc+Gestation+mage, birthweight2))$coef[5,4]
```

```
## [1] 0.3604293
```

```
summary(lm(Birthweight ~ Length+Headcirc+Gestation+mnocig, birthweight2))$coef[5,
```

```
## [1] 0.2172318
```

```
summary(lm(Birthweight ~ Length+Headcirc+Gestation+mheight, birthweight2))$coef[5
```

```
## [1] 0.6377429
```

```
summary(lm(Birthweight ~ Length+Headcirc+Gestation+mppwt, birthweight2))$coef[5,4
```

```
## [1] 0.2614658
```

```
summary(lm(Birthweight ~ Length+Headcirc+Gestation+fage, birthweight2))$coef[5,4]
```

```
## [1] 0.715077
```

```
summary(lm(Birthweight ~ Length+Headcirc+Gestation+fedyrs, birthweight2))$coef[5,
```

```
## [1] 0.5900573
```

```
summary(lm(Birthweight ~ Length+Headcirc+Gestation+fnocig, birthweight2))$coef[5,
```

```
## [1] 0.7061295
```

```
summary(lm(Birthweight ~ Length+Headcirc+Gestation+fheight, birthweight2))$coef[5
```

```
## [1] 0.1816579
```

```
summary(lm(Birthweight ~ Length+Headcirc+Gestation+lowbwt, birthweight2))$coef[5,
```

```
## [1] 0.1190075
```

`smoker` is next.

```
summary(lm(Birthweight ~ Length+Headcirc+Gestation+smoker+mage, birthweight2))$co
```

```
## [1] 0.6457006
```

```
summary(lm(Birthweight ~ Length+Headcirc+Gestation+smoker+mnocig, birthweight2))$
```

```
## [1] 0.7646006
```

```
summary(lm(Birthweight ~ Length+Headcirc+Gestation+smoker+mheight, birthweight2))
```

```
## [1] 0.4782143
```

```
summary(lm(Birthweight ~ Length+Headcirc+Gestation+smoker+mppwt, birthweight2))$c
```

```
## [1] 0.1780036
```

```
summary(lm(Birthweight ~ Length+Headcirc+Gestation+smoker+fage, birthweight2))$cc
```

```
## [1] 0.8667765
```

```
summary(lm(Birthweight ~ Length+Headcirc+Gestation+smoker+fedyrs, birthweight2))$
```

```
## [1] 0.5836636
```

```
summary(lm(Birthweight ~ Length+Headcirc+Gestation+smoker+fnocig, birthweight2))$
```

```
## [1] 0.5937779
```

```
summary(lm(Birthweight ~ Length+Headcirc+Gestation+smoker+fheight, birthweight2))
```

```
## [1] 0.2638458
```

```
summary(lm(Birthweight ~ Length+Headcirc+Gestation+smoker+lowbwt, birthweight2))$
```

```
## [1] 0.2226971
```

All the p-values are above the 10% threshold, so we stop and we do not include any more variables. So, the final model is

$$\text{Length, Headcirc, Gestation, smoker}$$

The main advantage of the testing-based methods is the low computational cost. However, since we test whether to add/drop variables "one-at-a-time", we are not able to compare all possible models. So, it is possible to miss the "optimal" model. In addition, it is not clear how to choose α_0, the cut-off for p-values.