MIE237

Neil Montgomery
2016-04-01

Practical Model Selection

There is no universally accepted model selection algorithm.

There is a large number of different criteria and strategies to use, some traditional and some modern.

We will focus on one (new) example *criterion* along with the so-called *sequential* strategies (add/remove one variable at a time) for model selection.

We will not discuss some of the more modern, computer-intensive model selection strategies based purely on predictive performance, which also require very large datasets.

A possible new criterion: "Adjusted" R^2

Consider two multiple regression models where one (the "smaller") is nested withing the other (the "larger").

Example: $y = \beta_0 + \beta_1 x_1 + \varepsilon$ versus $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$.

Fact: R^2 for the larger model must be at least as large as R^2 for the smaller model, no matter what.

Even if the extra terms in the larger model are, say, do nothing to predict y at all and don't really belong.

Why does this happen? Because $R^2 = 1 - SSE\big/SST$ is calculated *after* the sum of squared residuals is minimized, and more terms in a model mean there is a larger number of possible sets of residuals over which to minimize.

This is an easy example of why there are no automated model selection algorithms.

Adjusting for the number of model terms

R^2 always increases with more terms because SSE always decreases with more terms.

An adjusted version of R^2 divides SSE by its degrees of freedom as follows:

$$R^2_{adj} = 1 - \frac{SSE\Big/(n-(k+1))}{SST\Big/(n-1)} = 1 - \frac{MSE}{SST/(n-1)} = 1 - \frac{MSE}{s_y^2}$$

The adjusted version is penalized for adding model terms. It can be *smaller* even for larger models if the reduction in SSE can't overcome the decrease in error degrees of freedom.

R^2_{adj} can be used as a model comparison *criterion*, with larger being better. There are many other similar criteria - a few are mentioned in the book.

R^2_{adj} examples

I will simulate from a "true" model: $y = 1 + 2x_1 + 3x_2 + \varepsilon$ with $\varepsilon \sim N(0, 1)$ and $n = 40$.

Then I will simulate from another "true" model: $y = 1 + 2x_1 + 0.1x_2 + \varepsilon$

```
n <- 40
x_1 <- 1:n/10
x_2 <- sample(x_1)
e1 <- data.frame(y = 1 + 2*x_1 + 3*x_2 + rnorm(n, 0, 1))
e2 <- data.frame(y = 1 + 2*x_1 + 0.1*x_2 + rnorm(n, 0, 1))

var(e1$y)
```

```
## [1] 16.46583
```

```
var(e2$y)
```

```
## [1] 6.02302
```

5/13

first example - "smaller" model

```
##
## Call:
## lm(formula = y ~ x_1, data = e1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2869 -3.4702  0.9151  2.8677  6.9535
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.6989     1.2081   7.200 1.32e-08 ***
## x_1           1.4240     0.5135   2.773  0.00856 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.749 on 38 degrees of freedom
## Multiple R-squared:  0.1683, Adjusted R-squared:  0.1464
## F-statistic:  7.69 on 1 and 38 DF,  p-value: 0.008556
```

first example - "larger" model

```
##
## Call:
## lm(formula = y ~ x_1 + x_2, data = e1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5120 -0.8425 -0.1185  0.8838  2.8715
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.3225     0.5766   2.294   0.0276 *
## x_1           1.9781     0.1721  11.493 9.07e-14 ***
## x_2           3.0443     0.1721  17.687  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.236 on 37 degrees of freedom
## Multiple R-squared:  0.912,  Adjusted R-squared:  0.9073
## F-statistic: 191.8 on 2 and 37 DF,  p-value: < 2.2e-16
```

second example - "smaller" model

```
##
## Call:
## lm(formula = y ~ x_1, data = e2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8575 -0.5589  0.1170  0.4689  1.5992
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.2462     0.2880   4.327 0.000106 ***
## x_1           1.9590     0.1224  16.004  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8937 on 38 degrees of freedom
## Multiple R-squared:  0.8708, Adjusted R-squared:  0.8674
## F-statistic: 256.1 on 1 and 38 DF,  p-value: < 2.2e-16
```

second example - "larger" model

```
##
## Call:
## lm(formula = y ~ x_1 + x_2, data = e2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.990 -0.535  0.171  0.483  1.645
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.46136    0.41981   3.481   0.0013 **
## x_1          1.94284    0.12531  15.504   <2e-16 ***
## x_2         -0.08879    0.12531  -0.709   0.4831
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8996 on 37 degrees of freedom
## Multiple R-squared:  0.8725, Adjusted R-squared:  0.8656
## F-statistic: 126.6 on 2 and 37 DF,  p-value: < 2.2e-16
```

Sequential strategies

Sequential strategies involve adding (or removing) one variable at a time in a so-called "greedy" manner until a final model is selected.

Issue 1: usually equivalent to a large number of hypothesis tests performed on the same data.

Issue 2: current choices depend on past choices

Issue 3: multicollinearity can result in good variables omitted/bad variables included/more than one equally good final model

Forward regression - I

Easier to demonstrate than to describe. Start with: $y, x_1, x_2, \ldots, x_k$

Fit the models with one term:

$$y = \beta_0 + \beta_1 x_j + \varepsilon$$

If none give a small F-test p-value, it is unlikely that there willbe any useful model at all.

Either stop, or proceed with the strategy, with great caution (e.g. there is a strong scientific reason to consider interaction terms)

Forward regression - II

Note the model that produces any of the following (equivalent!):

- largest SSR
- smallest SSE
- largest F
- largest $|T|$
- **smallest p-value**

Call x_{j_i} the "winner". (Note the possible arbitriness in this and each subsequent step.)

Forward regression - III

Next: fit *all* the models with two terms:

$$y = \beta_0 + \beta_1 x_{j_1} + \beta_2 x_j + \varepsilon$$

(for $j \neq j_1$)

If no new variable included gets a small enough p-value, stop the procedure.

Otherwise, determine the variable that results in the smallest two-term SSE and call it x_{j_2}

And so on with all three term models…four term models…until you can't add any more variables resulting in small p-values.