Model Selection

Peter von Rohr

2022-04-11

Why Model Selection

- Start with results of Problem 1 of Exercise 4
- Two models with variables that show a significant effect

```
Call:
lm(formula = `Body Weight` ~ `Breast Circumference`, data = tbl_ex04p01_data)

Residuals:
     Min       1Q   Median       3Q      Max
-17.3941  -6.5525  -0.0673   9.3707  13.2594

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           -1065.115    255.483  -4.169 0.003126 **
`Breast Circumference`     8.673      1.420   6.108 0.000287 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.08 on 8 degrees of freedom
Multiple R-squared:  0.8234,    Adjusted R-squared:  0.8014
F-statistic: 37.31 on 1 and 8 DF,  p-value: 0.000287
```

```
Call:
lm(formula = `Body Weight` ~ Breed, data = tbl_ex04p01_data)

Residuals:
     Min       1Q   Median       3Q      Max
-10.0000  -7.5000  -0.1667   2.7500  21.0000

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     468.000      6.097  76.758 1.68e-11 ***
BreedLimousin    52.000      8.066   6.447 0.000351 ***
BreedSimmental   21.333      8.623   2.474 0.042575 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.56 on 7 degrees of freedom
Multiple R-squared:  0.8597,    Adjusted R-squared:  0.8196
F-statistic: 21.44 on 2 and 7 DF,  p-value: 0.001035
```

- Why not combining them to get an even better model?

Full Model

- All variables included

```
Call:
lm(formula = `Body Weight` ~ `Breast Circumference` + BCS + HEI +
    Breed, data = tbl_ex04p01_data)

Residuals:
      1       2       3       4       5       6       7       8       9      10
 1.8327 -0.5208  2.8604 -1.3120 -5.5552  2.6947  5.2055 -7.2432 -5.7525  7.7902

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             -859.4523   513.6852  -1.673   0.1696
`Breast Circumference`     7.1560     2.7705   2.583   0.0611 .
BCS                        9.9056     3.8258   2.589   0.0607 .
HEI                        0.1220     0.1822   0.669   0.5399
BreedLimousin             13.5466    15.5227   0.873   0.4321
BreedSimmental            -3.8614    10.1592  -0.380   0.7232
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.5 on 4 degrees of freedom
Multiple R-squared:  0.9596,    Adjusted R-squared:  0.909
F-statistic: 18.98 on 5 and 4 DF,  p-value: 0.006868
```

Check pairs plot for dependencies among variables

Best Model

- Including all variables does not always lead to the best model
- Best model aims at explaining a maximum of variation in responses
- Measured by

$$R^2 = \frac{||\hat{y} - \bar{y}||^2}{||y - \bar{y}||^2}$$

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

Finding the Best Model

- Full search over all possible combinations of predictors is too expensive
- Use practical approximations
  - Forward selection
  - Backward elimination

Alternative Model Selection Criteria

- Mallows C_p Statistic

$$C_p(\mathcal{M}) = \frac{SSE(\mathcal{M})}{\hat{\sigma}^2} - n + 2|\mathcal{M}|$$

- Akaike Information Criterion (AIC)
- Bayes Information Criterion (BIC)

Forward Selection

1. Start with the smallest model $\mathcal{M}_0$
2. Include the predictor variable which reduces the residual sum of squares the most.
3. Continue with step 2 until all predictor variables have been chosen
4. Choose the model with the smallest C_p value.

Backward Elimination

1. Start with the full model
2. Exclude the predictor variable increases the residual sum of squares the least.
3. Continue with step 2 until all predictor values have been deleted
4. Choose the model which has the smallest C_p value.

Example

- In R use:
  - package `olsrr` - no spaces in variable names
  - function `MASS::stepAIC()`