

Variable Selection

Variable selection is a balance between making the model as realistic as possible (include as many regressors as possible) and as simple as possible (including only the variables needed). Forward Selection will be implemented using p-values with a critical value of $\alpha = 0.05$. All selection criterion will be considered.

```
library('olsrr')
```

```
## Warning: package 'olsrr' was built under R version 4.0.3
```

```
##
```

```
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
```

```
##
```

```
##      rivers
```

```
library(faraway)
```

```
## Registered S3 methods overwritten by 'lme4':
```

```
##   method                      from
```

```
##   cooks.distance.influence.merMod car
```

```
##   influence.merMod              car
```

```
##   dfbeta.influence.merMod       car
```

```
##   dfbetas.influence.merMod      car
```

```
##
```

```
## Attaching package: 'faraway'
```

```
## The following object is masked from 'package:olsrr':
```

```
##
```

```
##      hsb
```

```
data <- read.csv("cleaned_data_4.csv", fileEncoding="UTF-8-BOM")
```

The forward selection function requires the model and a set p-value.

```
model <- lm(Weight ~ ., data = data)
```

```
FWDfit_p <- ols_step_forward_p(model, penter=0.05)
```

```
FWDfit_p
```

```
##
```

```
##                               Selection Summary
```

```
## -----
```

```
##      Variable                      Adj.
```

```
## Step      Entered      R-Square      R-Square      C(p)      AIC      RMSE
```

```
## -----
```

```
##    1    Width      0.9086      0.9081    916.8659    778.8648    2.7671
```

```
##    2   Length1      0.9619      0.9614    294.0794    641.8320    1.7928
```

```
##      3      Height      0.9821      0.9817      59.2495      523.8758      1.2334
##      4      Species_Pike      0.9850      0.9846      26.6765      497.3113      1.1311
##      5      Species_Smelt      0.9867      0.9862      9.4981      480.9309      1.0710
##      6      Species_Whitefish      0.9871      0.9866      6.7601      478.0428      1.0582
## -----
```

The forward selection builds up from no variables in the model. A predictor is added to the model at each iteration based on whether it has the lowest p-value. This continues until there are no more predictors to add, or all the remaining variables have a p-value > 0.05. In this case all selection criteria confirm that Width, Length1, Height, Species_Pike, Species_Smelt, and Species_Whitefish are included in the best model.

```
model <- lm(Weight ~
  Width +
  Length1 +
  Height +
  Species_Pike +
  Species_Smelt +
  Species_Whitefish ,
  data = data)
```

```
summary(model)
```

```
##
## Call:
## lm(formula = Weight ~ Width + Length1 + Height + Species_Pike +
##     Species_Smelt + Species_Whitefish, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.6278 -0.5128 -0.0424  0.4406  3.7161
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -8.16172    0.31660  -25.779  < 2e-16 ***
## Width          1.11711    0.20841   5.360 3.03e-07 ***
## Length1        0.64943    0.04172  15.565  < 2e-16 ***
## Height         0.46026    0.03642  12.636  < 2e-16 ***
## Species_Pike   -3.20450    0.70686  -4.533 1.17e-05 ***
## Species_Smelt   1.63367    0.37477   4.359 2.39e-05 ***
## Species_Whitefish 0.97924    0.44952   2.178  0.0309 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.058 on 152 degrees of freedom
## Multiple R-squared:  0.9871, Adjusted R-squared:  0.9866
## F-statistic: 1933 on 6 and 152 DF, p-value: < 2.2e-16
```

We can go back and see that the VIF is roughly the same as it was before the residual analysis and model selection process.

```
vif(model)
```

```
##              Width              Length1              Height              Species_Pike
```

```
##          17.418522          24.546980          3.439151          6.774753
## Species_Smelt Species_Whitefish
##          1.601499          1.041924
```

Width and Length1 still have high VIF. By removing one, we can see that the VIF's improve.

```
model_reduced <- lm(Weight ~
  Length1 +
  Height +
  Species_Pike +
  Species_Smelt +
  Species_Whitefish ,
  data = data)
vif(model_reduced)
```

```
##          Length1          Height Species_Pike Species_Smelt
##          4.919206          3.426583          2.972007          1.398381
## Species_Whitefish
##          1.025656
```