HW 20 – Q2

Nick Huo

2022-11-02

a.

```
fish <- read.csv("Fish.csv")
attach(fish)

mod_full <- lm(Weight ~ ., data=fish)
summary(mod_full)
```

```
##
## Call:
## lm(formula = Weight ~ ., data = fish)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -213.18  -53.19  -12.62   36.49  420.82
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -918.3321   127.0831  -7.226  2.5e-11 ***
## SpeciesParkki      164.7227    75.6995   2.176 0.031152 *
## SpeciesPerch       137.9489   120.3135   1.147 0.253419
## SpeciesPike       -208.4294   135.3064  -1.540 0.125607
## SpeciesRoach       103.0400    91.3084   1.128 0.260954
## SpeciesSmelt       446.0733   119.4303   3.735 0.000268 ***
## SpeciesWhitefish    93.8742    96.6580   0.971 0.333045
## Length1            -80.3030    36.2785  -2.214 0.028403 *
## Length2             79.8886    45.7180   1.747 0.082653 .
## Length3             32.5354    29.3002   1.110 0.268633
## Height               5.2510    13.0560   0.402 0.688128
## Width               -0.5154    23.9130  -0.022 0.982832
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 93.83 on 147 degrees of freedom
## Multiple R-squared:  0.9361, Adjusted R-squared:  0.9313
## F-statistic: 195.7 on 11 and 147 DF,  p-value: < 2.2e-16
```

It appears that `SpeciesParkki`, `SpeciesSmelt`, and `Length1` are useful as they have significant coefficients.

- `SpeciesParkki`: The predicted weighted for Parkki fish is higher than that of Bream fish by 164.7 grams.

- SpeciesSmelt: The predicted weighted for Smelt fish is higher than that of Bream fish by 446.1 grams.
- Length1: The predicted weighted for a fish decrease by 80.3 grams for every 1 cm increase in vertical length.

b.

```
library(MASS)
mod_empty <- lm(Weight ~ 1, data=fish)
stepAIC(mod_empty, scope=list(lower=mod_empty, upper=mod_full),
        direction="both", k=log(length(fish)))
```

```
## Start:  AIC=1870.93
## Weight ~ 1
##
##             Df Sum of Sq      RSS    AIC
## + Length3  1  17251026  2996433 1569.1
## + Length2  1  17085990  3161469 1577.6
## + Length1  1  16978060  3269399 1583.0
## + Width    1  15912356  4335103 1627.8
## + Height   1  10623359  9624100 1754.6
## + Species  6   7515048 12732411 1808.8
## <none>                 20247459 1870.9
##
## Step:  AIC=1569.09
## Weight ~ Length3
##
##             Df Sum of Sq      RSS    AIC
## + Species  6   1654578  1341855 1453.0
## + Width    1    507023  2489410 1541.6
## + Height   1    225843  2770591 1558.6
## <none>                  2996433 1569.1
## + Length2  1      1783  2994650 1570.9
## + Length1  1         1  2996433 1571.0
## - Length3  1  17251026 20247459 1870.9
##
## Step:  AIC=1453.03
## Weight ~ Length3 + Species
##
##             Df Sum of Sq      RSS    AIC
## + Length1  1     18526  1323329 1452.8
## <none>                  1341855 1453.0
## + Height   1      4156  1337699 1454.5
## + Width    1       580  1341275 1454.9
## + Length2  1        40  1341815 1455.0
## - Species  6   1654578  2996433 1569.1
## - Length3  1  11390556 12732411 1808.8
##
## Step:  AIC=1452.77
## Weight ~ Length3 + Species + Length1
##
##             Df Sum of Sq     RSS    AIC
## + Length2  1     27180 1296149 1451.4
## <none>                 1323329 1452.8
## - Length1  1     18526 1341855 1453.0
## + Height   1      1989 1321340 1454.5
## + Width    1       120 1323209 1454.7
## - Length3  1     92501 1415830 1461.6
## - Species  6   1673104 2996433 1571.0
##
## Step:  AIC=1451.41
## Weight ~ Length3 + Species + Length1 + Length2
##
##           Df Sum of Sq     RSS    AIC
## - Length3  1     13742 1309891 1451.1
## <none>                 1296149 1451.4
## - Length2  1     27180 1323329 1452.8
## + Height   1      2026 1294122 1453.1
## + Width    1       606 1295542 1453.3
## - Length1  1     45666 1341815 1455.0
## - Species  6   1664007 2960156 1571.0
##
## Step:  AIC=1451.14
## Weight ~ Species + Length1 + Length2
##
##           Df Sum of Sq     RSS    AIC
## <none>                 1309891 1451.1
## + Length3  1     13742 1296149 1451.4
## + Height   1      4777 1305114 1452.5
## + Width    1      2624 1307268 1452.8
## - Length1  1     46133 1356024 1454.7
## - Length2  1    105939 1415830 1461.6
## - Species  6   1724373 3034264 1573.0
```

```
##
## Call:
## lm(formula = Weight ~ Species + Length1 + Length2, data = fish)
##
## Coefficients:
##      (Intercept)     SpeciesParkki      SpeciesPerch       SpeciesPike
##       -782.11810          85.25108           1.78572        -352.69065
##     SpeciesRoach      SpeciesSmelt  SpeciesWhitefish           Length1
##         13.39046         317.66140           0.03012         -82.46214
##          Length2
##        117.76469
```

Using step-wise selection with BIC, the final model include `Species`, `Length1`, and `Length2` as predictors. This means that using a fish's species, vertical length, and diagonal length to predict its weight is a good approach.

c.

```
mod_final <- lm(Weight ~ Species + Length1 + Length2)
summary(mod_final)
```

```
##
## Call:
## lm(formula = Weight ~ Species + Length1 + Length2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -213.72  -56.55   -9.27   39.29  415.48
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -782.11810   50.96147 -15.347  < 2e-16 ***
## SpeciesParkki      85.25108   38.69264   2.203  0.02910 *
## SpeciesPerch        1.78572   24.29313   0.074  0.94150
## SpeciesPike      -352.69065   35.98226  -9.802  < 2e-16 ***
## SpeciesRoach       13.39046   34.80014   0.385  0.70094
## SpeciesSmelt      317.66140   49.87359   6.369  2.2e-09 ***
## SpeciesWhitefish    0.03012   41.85179   0.001  0.99943
## Length1           -82.46214   35.87738  -2.298  0.02292 *
## Length2           117.76469   33.81109   3.483  0.00065 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 93.45 on 150 degrees of freedom
## Multiple R-squared:  0.9353, Adjusted R-squared:  0.9319
## F-statistic: 271.1 on 8 and 150 DF,  p-value: < 2.2e-16
```

I think the most important predictor in determining the weight of a fish is probably `Length2`. Because the diagonal length has the lower p-value and standard error for its coefficient, it has a stronger correlation with a fish’s weight, compared to the correlation between the vertical length (`Length1`) and weight. And at the same time, the coefficient value of `Length2` has a higher magnitude, which might mean that it is more important, considering that the two lengths predictors values should have also be similar in magnitudes.