MPulleyM2Q6

Melissa Pulley

2024-04-23

You are tasked on evaluating the growth of grass (kg of dry matter per hectare: KgDMHA).

You want to test three models:

KgDMHA~ Water + Salinity + Nitrogen

KgDMHA~ Pests + Graze

KgDMHA~ Water + Salinity + Nitrogen + Pests + Graze

Before running the models, you wonder whether AICc will truly choose the model that best predicts the data, so, you have decided to do some model validation.

```
grass=read.csv("grasslanddata.csv")
summary(grass)
```

```
##        X              KgDMHA          Water           Salinity
##  Min.   :   1.0   Min.   : 2180   Min.   :442024   Min.   : 0.000
##  1st Qu.: 250.8   1st Qu.: 5118   1st Qu.:644507   1st Qu.: 7.100
##  Median : 500.5   Median : 5955   Median :695289   Median : 9.450
##  Mean   : 500.5   Mean   : 5972   Mean   :698142   Mean   : 9.511
##  3rd Qu.: 750.2   3rd Qu.: 6820   3rd Qu.:752103   3rd Qu.:11.900
##  Max.   :1000.0   Max.   :10473   Max.   :947395   Max.   :19.400
##     Nitrogen         Pests           Graze
##  Min.   :1.300   Min.   : 1.00   Min.   : 0.00
##  1st Qu.:3.100   1st Qu.:35.00   1st Qu.:33.00
##  Median :3.700   Median :45.00   Median :40.00
##  Mean   :3.668   Mean   :44.53   Mean   :40.33
##  3rd Qu.:4.200   3rd Qu.:54.00   3rd Qu.:48.00
##  Max.   :6.100   Max.   :86.00   Max.   :77.00
```

1. Randomly split your dataset, with ~75% of the data for the training and ~25% of the data for the test

```
set.seed(5)
N=length(grass$X)
test_size = 0.25*N
test_data_ind = as.numeric(sample(1:1000,test_size,replace=F))
max(test_data_ind)
```

```
## [1] 1000
```

```
test_dat = grass[test_data_ind,]
summary(test_dat)
```

```
##        X              KgDMHA          Water           Salinity
##  Min.   :   2.0   Min.   :2180   Min.   :471758   Min.   : 0.000
##  1st Qu.: 207.5   1st Qu.:4980   1st Qu.:644962   1st Qu.: 7.400
##  Median : 443.5   Median :5890   Median :700357   Median : 9.400
```

```
##  Mean   : 474.0   Mean   :5883   Mean   :701954   Mean   : 9.544
##  3rd Qu.: 743.0   3rd Qu.:6700   3rd Qu.:764211   3rd Qu.:11.800
##  Max.   :1000.0   Max.   :8912   Max.   :904451   Max.   :19.000
##     Nitrogen         Pests           Graze
##  Min.   :1.400   Min.   : 3.00   Min.   : 8.0
##  1st Qu.:3.100   1st Qu.:36.00   1st Qu.:34.0
##  Median :3.600   Median :46.00   Median :40.5
##  Mean   :3.554   Mean   :45.36   Mean   :40.6
##  3rd Qu.:4.000   3rd Qu.:55.00   3rd Qu.:48.0
##  Max.   :5.300   Max.   :75.00   Max.   :74.0
```

```
training_dat = grass[-test_data_ind,]
summary(training_dat)
```

```
##        X             KgDMHA          Water           Salinity
##  Min.   :  1.0   Min.   : 2348   Min.   :442024   Min.   : 0.0
##  1st Qu.:271.5   1st Qu.: 5163   1st Qu.:643946   1st Qu.: 7.0
##  Median :519.5   Median : 5970   Median :694063   Median : 9.5
##  Mean   :509.3   Mean   : 6002   Mean   :696871   Mean   : 9.5
##  3rd Qu.:751.8   3rd Qu.: 6873   3rd Qu.:749747   3rd Qu.:11.9
##  Max.   :999.0   Max.   :10473   Max.   :947395   Max.   :19.4
##     Nitrogen         Pests           Graze
##  Min.   :1.300   Min.   : 1.00   Min.   : 0.00
##  1st Qu.:3.200   1st Qu.:35.00   1st Qu.:32.00
##  Median :3.700   Median :44.00   Median :40.00
##  Mean   :3.706   Mean   :44.25   Mean   :40.24
##  3rd Qu.:4.200   3rd Qu.:54.00   3rd Qu.:48.00
##  Max.   :6.100   Max.   :86.00   Max.   :77.00
```

2. With the training dataset, run each of the three models. You should obtain an output for each model

```
m1 = lm(KgDMHA~Water + Salinity + Nitrogen, data=training_dat)
m2 = lm(KgDMHA~Pests + Graze, data=training_dat)
m3 = lm(KgDMHA~Water + Salinity + Nitrogen + Pests + Graze, data=training_dat)
```

3. Obtain an AICc table for the models you ran

```
library(MuMIn)
test=AICc(m1,m2,m3)
test
```

```
##    df     AICc
## m1  5 12450.63
## m2  4 12286.94
## m3  7 10512.43
```

4) In the test dataset, add three new columns called predictmodel1, predictmodel2, precitmodel3. Populate the columns with the predicted values based on each of the three models you ran. You can use the coefficients, or the predict function to obtain those values.

```
newdata1 <- data.frame(Water = seq(min(test_dat$Water,na.rm = TRUE),
                                   max(test_dat$Water,na.rm=TRUE),length=250),
                       Salinity = seq(min(test_dat$Salinity,na.rm = TRUE),
                                      max(test_dat$Salinity,na.rm=TRUE), length=250),
                       Nitrogen = seq(min(test_dat$Nitrogen,na.rm = TRUE),
                                      max(test_dat$Nitrogen,na.rm=TRUE), length=250))

newdata2 <- data.frame(Graze = seq(min(test_dat$Graze,na.rm = TRUE),
```

```
                                    max(test_dat$Graze,na.rm=TRUE),length=250),
                          Pests = seq(min(test_dat$Pests,na.rm = TRUE),
                                      max(test_dat$Pests,na.rm=TRUE), length=250))

newdata3 <- data.frame(Water = seq(min(test_dat$Water,na.rm = TRUE),
                                    max(test_dat$Water,na.rm=TRUE),length=250),
                          Salinity = seq(min(test_dat$Salinity,na.rm = TRUE),
                                         max(test_dat$Salinity,na.rm=TRUE), length=250),
                          Nitrogen = seq(min(test_dat$Nitrogen,na.rm = TRUE),
                                         max(test_dat$Nitrogen,na.rm=TRUE), length=250),
                          Graze = seq(min(test_dat$Graze,na.rm = TRUE),
                                      max(test_dat$Graze,na.rm=TRUE),length=250),
                          Pests = seq(min(test_dat$Pests,na.rm = TRUE),
                                      max(test_dat$Pests,na.rm=TRUE), length=250))

test_dat$predictmodel1=predict(m1, newdata = newdata1)
test_dat$predictmodel2=predict(m2, newdata = newdata2)
test_dat$predictmodel3=predict(m3, newdata = newdata3)
```

5. Use the following equation:

$$RSME = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

   And estimate RMSE for EACH of the three models.

NOTE: Do not use a built-in R function to estimate RMSE, you should estimate it using the equation. You can take many steps (e.g., creating new columns, multiple lines of code) if you need to. As a reminder y is the observed data, and ŷ is the predicted value (for each model)

```
rsme1 = sqrt(sum((test_dat$predictmodel1-mean(test_dat$predictmodel1))^2)/length(test_dat$predictmodel1)
rsme2 = sqrt(sum((test_dat$predictmodel2-mean(test_dat$predictmodel2))^2)/length(test_dat$predictmodel2)
rsme3 = sqrt(sum((test_dat$predictmodel3-mean(test_dat$predictmodel3))^2)/length(test_dat$predictmodel3)

c(rsme1, rsme2, rsme3)
```

```
## [1]  676.2303 1944.7004 1127.5202
```

6. According to the RMSE, what model is the best?

   Since model 1 has the lowest RSME, it is the best model by this metric.

7. Compare the RMSE with the AICc and with the R-squared from the models. Do all three metrics agree on the best model?

   The AICc and r^2 indicates that model 3 is the best model since its AICc is lowest and its r^2 is the largest, describing 95.36%. This disagrees with th conclusion from the RMSE.

```
#checking r^2s
summary(m1) #0.3821
```

```
##
## Call:
## lm(formula = KgDMHA ~ Water + Salinity + Nitrogen, data = training_dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2672.74  -693.06    34.95   670.04  3035.75
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.226e+02  3.680e+02   1.963     0.05 *
## Water        6.957e-03  4.447e-04  15.645   <2e-16 ***
## Salinity    -1.256e+02  1.050e+01 -11.961   <2e-16 ***
## Nitrogen     4.384e+02  4.557e+01   9.621   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 970.1 on 746 degrees of freedom
## Multiple R-squared:  0.3821, Adjusted R-squared:  0.3796
## F-statistic: 153.8 on 3 and 746 DF,  p-value: < 2.2e-16
```

```
summary(m2) #0.5019
```

```
##
## Call:
## lm(formula = KgDMHA ~ Pests + Graze, data = training_dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2666.28  -575.37     3.29   600.70  2465.30
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10107.986    153.029   66.05   <2e-16 ***
## Pests         -41.773      2.209  -18.91   <2e-16 ***
## Graze         -56.092      2.644  -21.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 870.4 on 747 degrees of freedom
## Multiple R-squared:  0.5019, Adjusted R-squared:  0.5006
## F-statistic: 376.4 on 2 and 747 DF,  p-value: < 2.2e-16
```

```
summary(m3) #0.9536
```

```
##
## Call:
## lm(formula = KgDMHA ~ Water + Salinity + Nitrogen + Pests + Graze,
##     data = training_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -841.62 -178.61    7.72  173.95  805.37
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.999e+03  1.068e+02   37.44   <2e-16 ***
## Water        7.997e-03  1.225e-04   65.29   <2e-16 ***
## Salinity    -1.145e+02  2.883e+00  -39.71   <2e-16 ***
## Nitrogen     5.186e+02  1.255e+01   41.32   <2e-16 ***
## Pests       -4.277e+01  6.765e-01  -63.22   <2e-16 ***
## Graze       -6.239e+01  8.143e-01  -76.62   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 266.1 on 744 degrees of freedom
## Multiple R-squared:  0.9536, Adjusted R-squared:  0.9533
## F-statistic:  3060 on 5 and 744 DF,  p-value: < 2.2e-16
```