Ocean Data Analysis with R Programming for Early Career Ocean Professionals (ECOPs) (Asia)

Mohamad Lukman Aidid bin Mohd Yusoff

2023-11-02

Assignment. Lesson 4: Multiple regression

1. Use the glm() function to fit a multiple linear regression model to the data, with Rajidae individual counts as the dependent variable and shore distance as well as minimum depth as the independent variables. What is the R- squared value of the regression model?
2. Use the summary() function to view the regression results, including the coefficients, p-values, and model fit statistics. What is the p-value of the minimum depth variable in the regression model? What can you conclude? Use the plot() function to visualize the regression results, including the fitted values and residuals. What do the residuals indicate about the fit of the regression model?

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(dplyr)
library(ggplot2)
library(stats)
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```
setwd('C:/Users/Administrator/Desktop/R/')
```

```
obis <- read.csv("C:/Users/Administrator/Desktop/R/obis_red_list_filtered_1000.csv")
head(obis)
```

```
##          scientificName date_year          family minimumDepthInMeters
## 1 Balaenoptera physalus      2003 Balaenopteridae                    0
## 2 Balaenoptera physalus      2003 Balaenopteridae                    0
## 3 Balaenoptera physalus      2003 Balaenopteridae                    0
## 4 Balaenoptera physalus      2003 Balaenopteridae                    0
## 5 Balaenoptera physalus      2003 Balaenopteridae                    0
## 6 Balaenoptera physalus      2002 Balaenopteridae                    0
##   shoredistance   sst   sss individualCount    country status
## 1        182964 -1.47 34.03               2 Antarctica     VU
## 2        135623 -1.58 34.01               2 Antarctica     VU
## 3        138638 -1.58 34.01               9 Antarctica     VU
## 4         77966 -1.57 34.06               4 Antarctica     VU
## 5        141441 -1.59 34.02               3 Antarctica     VU
## 6        -14124 -1.43 33.71               3 Antarctica     VU
```

```
obis_rajidae <- obis %>%
  filter(family == "Rajidae")

str(obis_rajidae)
```

```
## 'data.frame':    47 obs. of  10 variables:
##  $ scientificName      : chr  "Dipturus canutus" "Dipturus canutus" "Dipturus canutus" "Dipturus can
##  $ date_year           : int  2001 2000 2001 2000 2000 2001 2000 2001 2001 2001 ...
##  $ family              : chr  "Rajidae" "Rajidae" "Rajidae" "Rajidae" ...
##  $ minimumDepthInMeters: num  618 598 618 579 598 386 598 502 386 425 ...
##  $ shoredistance       : int  15941 15468 22168 19877 19877 26421 20634 20634 16617 25785 ...
##  $ sst                 : num  19.3 19.3 19.2 19.1 19.1 ...
##  $ sss                 : num  35.6 35.6 35.6 35.6 35.6 ...
##  $ individualCount     : num  4 2 2 2 2 11 2 2 3 3 ...
##  $ country             : chr  "Australia" "Australia" "Australia" "Australia" ...
##  $ status              : chr  "EN" "EN" "EN" "EN" ...
```

Question 1 & 2:

Modeling is done using lm() instead of glm() as glm() does not give R-squared value.

```
model3 = lm(individualCount~shoredistance + minimumDepthInMeters, data = obis_rajidae)
model3
```

```
##
## Call:
## lm(formula = individualCount ~ shoredistance + minimumDepthInMeters,
##     data = obis_rajidae)
##
## Coefficients:
##          (Intercept)         shoredistance  minimumDepthInMeters
##            94.125027             -0.002817             -0.069335
```

```
summary(model3)
```

```
##
## Call:
## lm(formula = individualCount ~ shoredistance + minimumDepthInMeters,
##     data = obis_rajidae)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.070 -23.686  -5.764   9.711 312.764
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          94.125027  14.928127   6.305  1.2e-07 ***
## shoredistance        -0.002817   0.002103  -1.340    0.187
## minimumDepthInMeters -0.069335   0.075089  -0.923    0.361
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.93 on 44 degrees of freedom
## Multiple R-squared:  0.3448, Adjusted R-squared:  0.315
## F-statistic: 11.58 on 2 and 44 DF,  p-value: 9.131e-05
```

```
summary(model3)$r.squared
```

```
## [1] 0.3447797
```

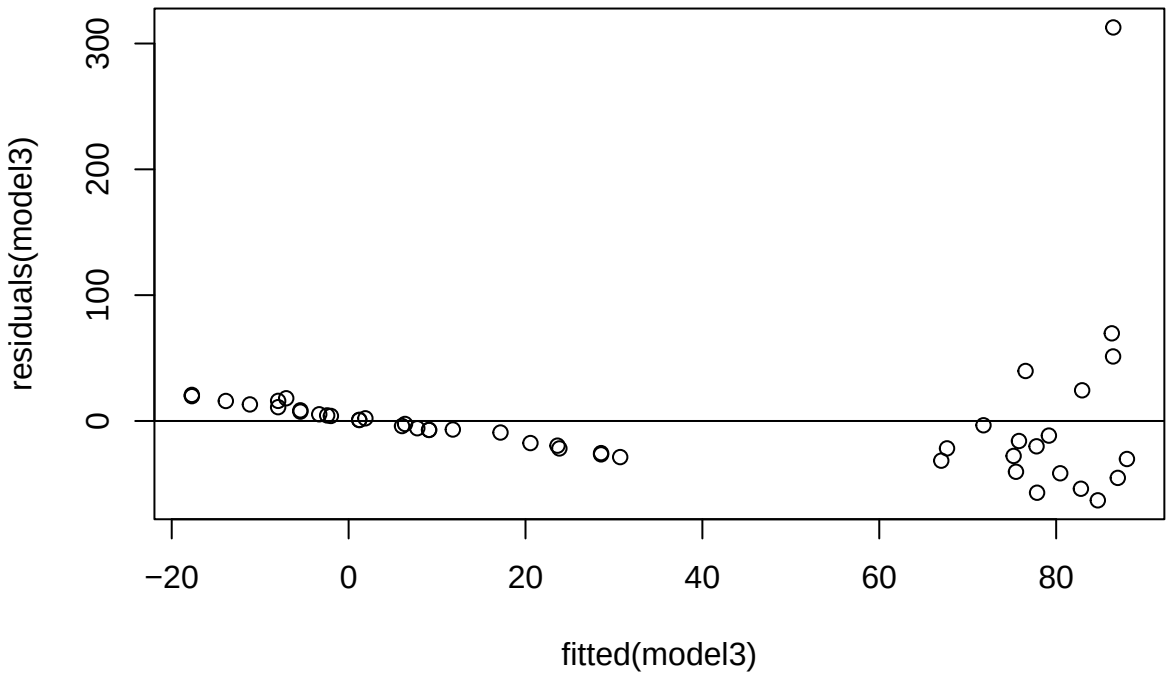
p-value for minimum depth is greater than 0.05. This indicate that there is no significant relationship between minimum depth and individual count.

```
shapiro.test(residuals(model3))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model3)
## W = 0.59219, p-value = 3.347e-10
```

Question 3:

```
plot(residuals(model3)~fitted(model3))
abline(h=0)
```



The plot seems to show a discernible pattern in one grouping of the points. This indicate that the regression model might not be a good fit.

ADDITIONAL TESTING

Testing with gamma distribution with lognormal form for model

```
model4 = glm(individualCount~shoredistance + minimumDepthInMeters, data = obis_rajidae, family = Gamma(
model4
```

```
##
## Call:  glm(formula = individualCount ~ shoredistance + minimumDepthInMeters,
##     family = Gamma(link = log), data = obis_rajidae)
##
## Coefficients:
##          (Intercept)         shoredistance  minimumDepthInMeters
##            4.749e+00            -4.212e-05            -5.270e-03
##
## Degrees of Freedom: 46 Total (i.e. Null);  44 Residual
## Null Deviance:       120.4
## Residual Deviance: 20.8  AIC: 319.8
```

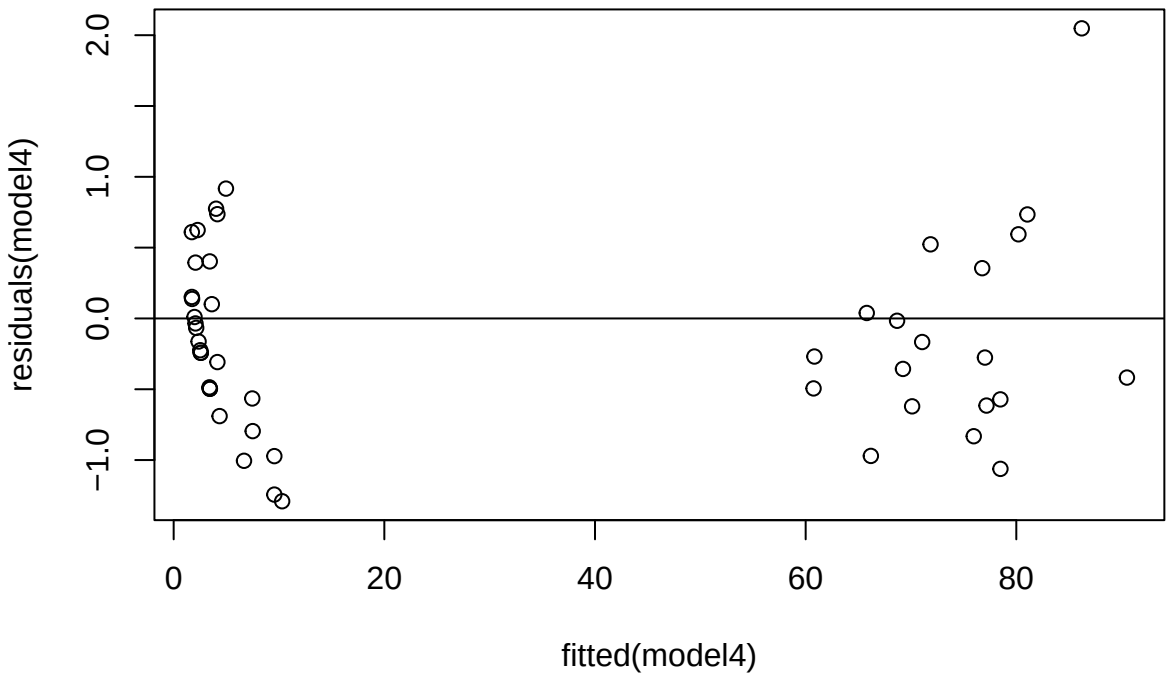
```
summary(model4)
```

```
##
## Call:
## glm(formula = individualCount ~ shoredistance + minimumDepthInMeters,
##     family = Gamma(link = log), data = obis_rajidae)
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4.749e+00  2.110e-01  22.513  < 2e-16 ***
## shoredistance         -4.212e-05  2.971e-05  -1.417    0.163
## minimumDepthInMeters -5.270e-03  1.061e-03  -4.967 1.08e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.6025143)
##
##     Null deviance: 120.36  on 46  degrees of freedom
## Residual deviance:  20.80  on 44  degrees of freedom
## AIC: 319.76
##
## Number of Fisher Scoring iterations: 10
```

```
shapiro.test(residuals(model4))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model4)
## W = 0.96028, p-value = 0.1104
```

```
plot(residuals(model4)~fitted(model4))
abline(h=0)
```




```
AIC(model3, model4)
```

```
##        df      AIC
## model3  4 514.8449
## model4  4 319.7615
```