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

Linear models

Mohamad Lukman Aidid bin Mohd Yusoff

2023-11-02

Assignment. Lesson 3: Linear models

1. In Module I, you learned how to filter your data with the filter() function. Use it to create a new dataset that includes only lines from the Rajidae family.
2. Use the glm() function to fit a linear model to the data, with the number of Rajidae individuals as the dependent variable and distance to shore as the independent variable. What is the p-value of the distance to shore variable? What can you conclude?
3. Use the plot() function to view the results of the linear model, including fitted values and residuals and normality. What do the residuals indicate about the fit of the linear model?
4. Change the family of the model to a gamma distribution with a lognormal form (family = Gamma(link = log). Compare the normality and homogeneity of the residual values of this model with those of the previous model. Which model is the best fit?

```
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(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
```

```
library(MASS)
```

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

```
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
```

```
str(obis)
```

```
## 'data.frame':    1000 obs. of  10 variables:
##  $ scientificName      : chr  "Balaenoptera physalus" "Balaenoptera physalus" "Balaenoptera physalus
##  $ date_year           : int  2003 2003 2003 2003 2003 2002 2003 2003 2003 2020 ...
##  $ family              : chr  "Balaenopteridae" "Balaenopteridae" "Balaenopteridae" "Balaenopteridae
##  $ minimumDepthInMeters: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ shoredistance       : int  182964 135623 138638 77966 141441 -14124 727065 184171 144748 478287 .
##  $ sst                 : num  -1.47 -1.58 -1.58 -1.57 -1.59 -1.43 -0.51 -1.48 -1.55 0.35 ...
##  $ sss                 : num  34 34 34 34.1 34 ...
##  $ individualCount     : num  2 2 9 4 3 3 3 6 6 8 ...
##  $ country             : chr  "Antarctica" "Antarctica" "Antarctica" "Antarctica" ...
##  $ status              : chr  "VU" "VU" "VU" "VU" ...
```

Question 1:

```
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 canu
##  $ 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" ...
```

```
summary(obis_rajidae)
```

```
##  scientificName       date_year       family          minimumDepthInMeters
##  Length:47          Min.   :2000   Length:47          Min.   : 36.0
##  Class :character   1st Qu.:2000   Class :character   1st Qu.: 57.5
##  Mode  :character   Median :2001   Mode  :character   Median :386.0
##                     Mean   :2002                      Mean   :324.1
##                     3rd Qu.:2004                      3rd Qu.:559.5
##                     Max.   :2004                      Max.   :618.0
##  shoredistance        sst             sss        individualCount
##  Min.   :  916   Min.   :18.96   Min.   :35.59   Min.   :  2.00
##  1st Qu.: 4940   1st Qu.:19.18   1st Qu.:35.59   1st Qu.:  2.00
##  Median :14496   Median :19.27   Median :35.59   Median :  4.00
##  Mean   :13093   Mean   :19.46   Mean   :36.20   Mean   : 34.78
##  3rd Qu.:20634   3rd Qu.:19.86   3rd Qu.:37.02   3rd Qu.: 43.81
##  Max.   :26421   Max.   :19.97   Max.   :37.38   Max.   :399.22
##    country             status
##  Length:47          Length:47
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

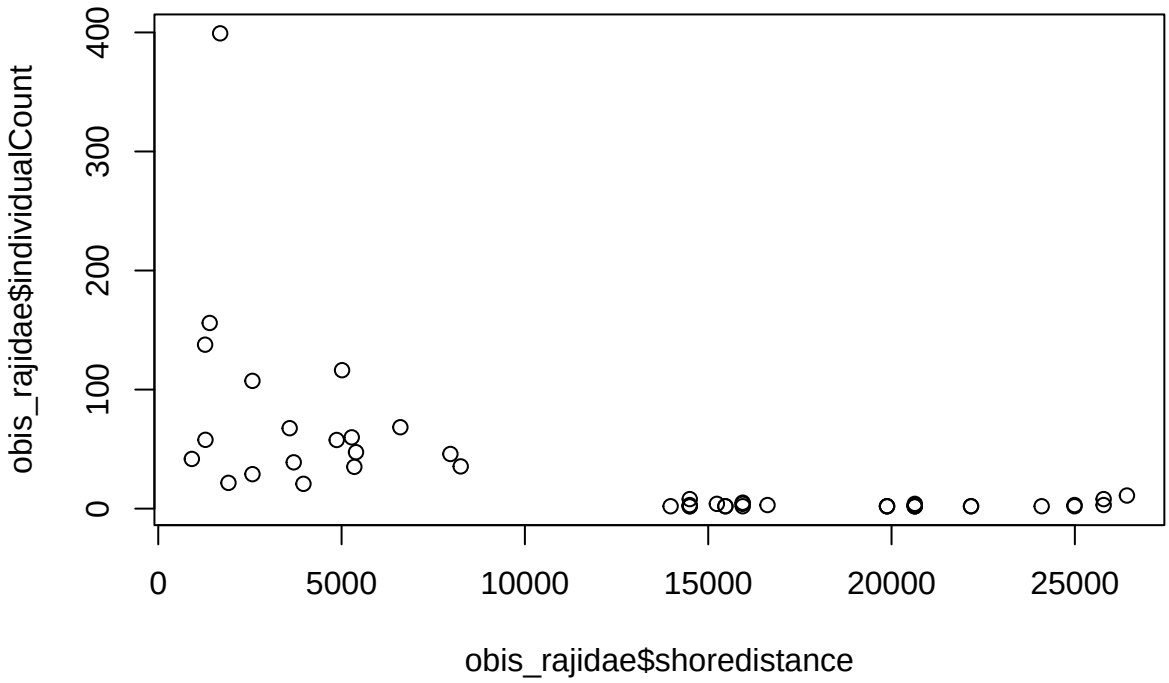
Question 2:

```
model = glm(individualCount~shoredistance, data = obis_rajidae, "gaussian")
summary(model)
```

```
##
## Call:
## glm(formula = individualCount ~ shoredistance, family = "gaussian",
##     data = obis_rajidae)
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   94.2564502 14.9029805   6.325 1.03e-07 ***
## shoredistance -0.0045428  0.0009604  -4.730 2.25e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 3007.11)
##
##     Null deviance: 202600  on 46  degrees of freedom
## Residual deviance: 135320  on 45  degrees of freedom
## AIC: 513.75
##
## Number of Fisher Scoring iterations: 2
```

The p-value < 0.05, thus there is some confidence that the model is fitting the data well.

```
plot(obis_rajidae$shoredistance, obis_rajidae$individualCount)
```

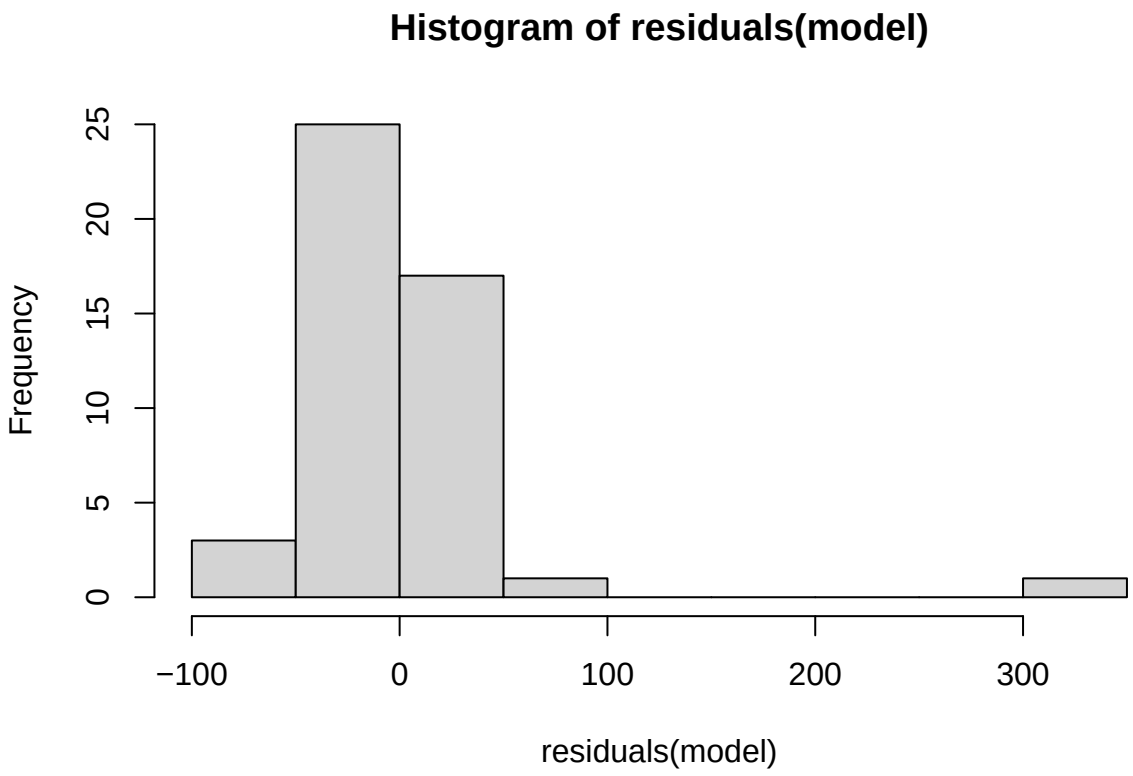



```
confint(model, level=0.95)
```

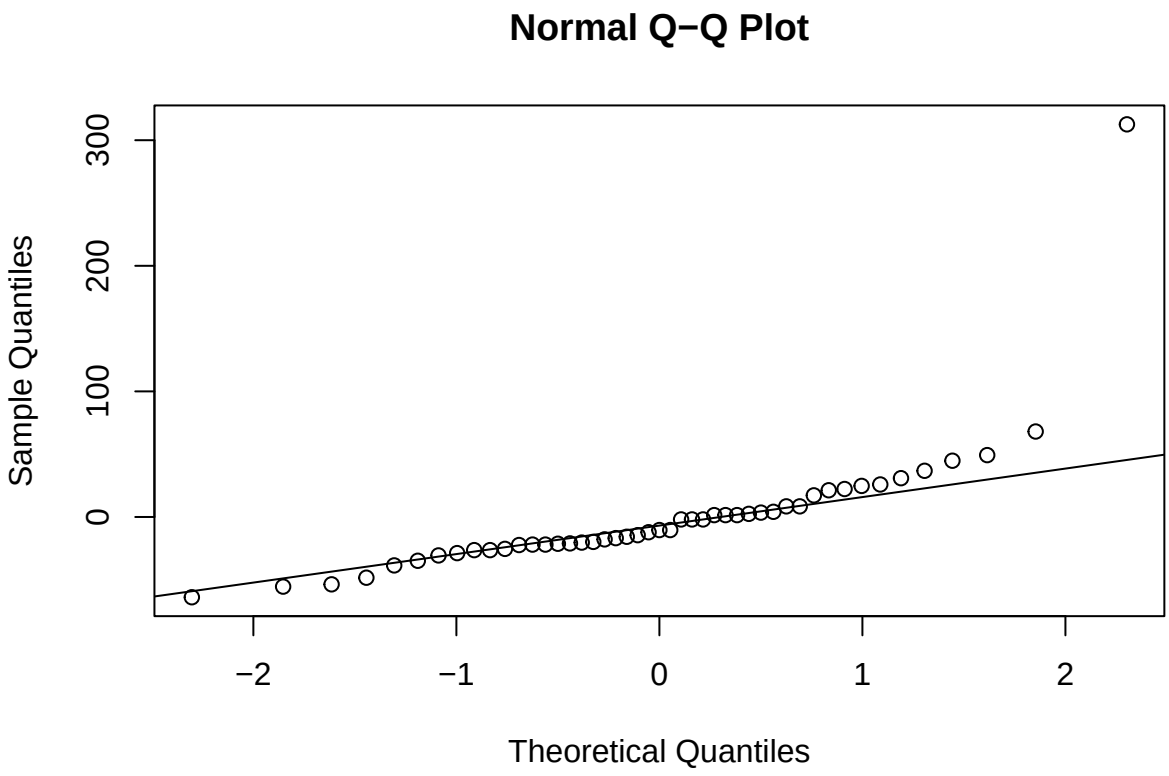
```
## Waiting for profiling to be done...
```

```
##                       2.5 %        97.5 %
## (Intercept)   65.047145107 123.465755343
## shoredistance -0.006425231  -0.002660467
```

```
hist(residuals(model))
```




```
qqnorm(residuals(model))
qqline(residuals(model))
```



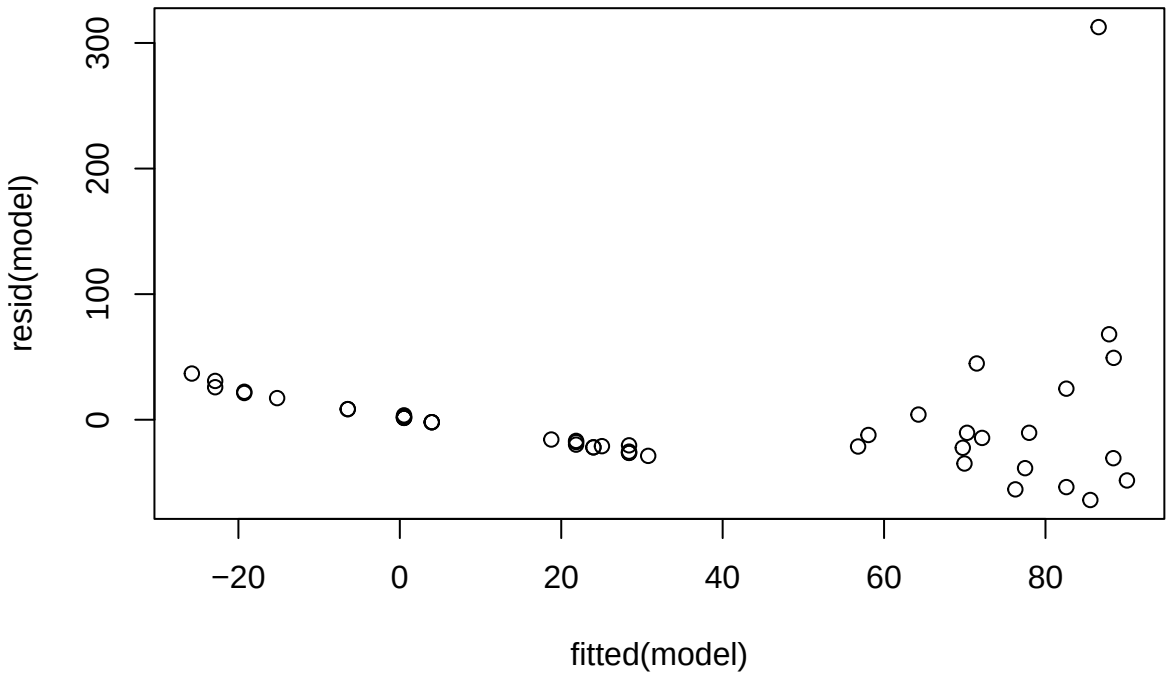

```
shapiro.test(residuals(model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model)
## W = 0.60673, p-value = 5.437e-10
```

P-value is less than 0.05, indicating some likelihood that the residuals are not normally distributed.

Question 3:

```
plot(fitted(model), resid(model))
```



The plot shows that there are two distinct groups, 1) group 1 between -30 and 30 and 2) group 2 between 55 and 100. For group 1, the residuals is not randomly distributed around zero as it decreases when the fitted values increases. **This indicate that the linear model does not fit the data well.**

Question 4:

```
model2 <- glm(individualCount~shoredistance, data = obis_rajidae, family = Gamma(link = log))
summary(model2)
```

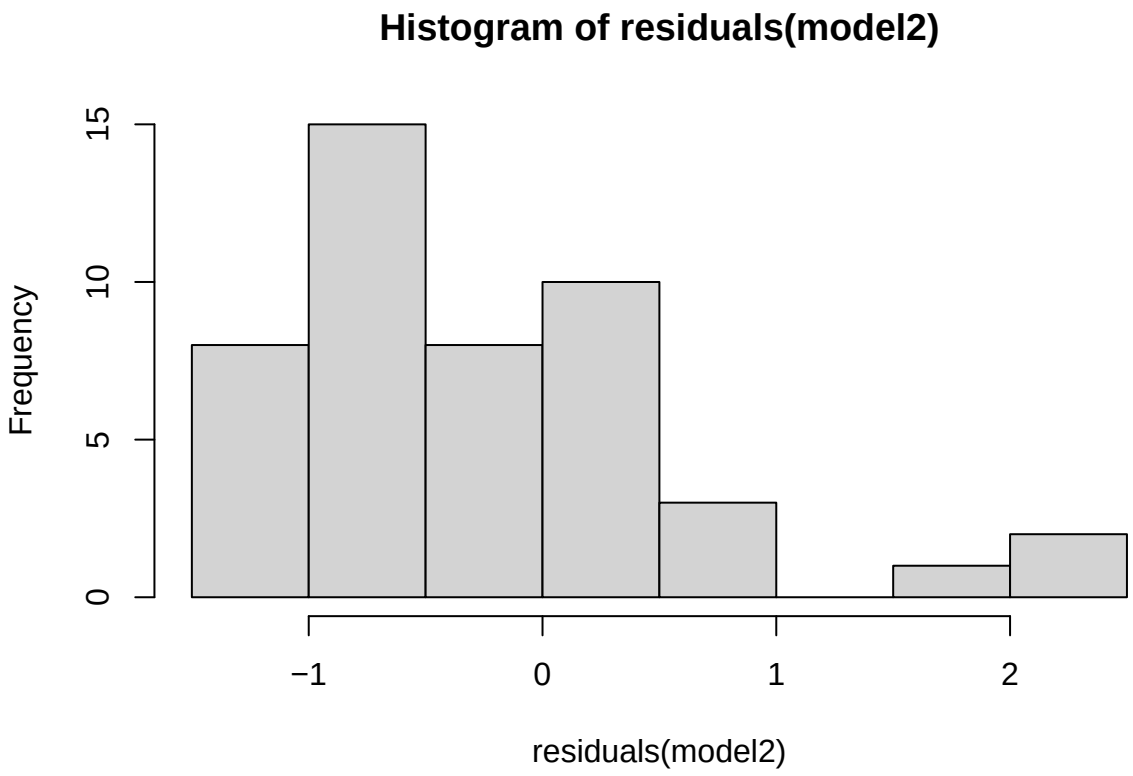
```
##
## Call:
## glm(formula = individualCount ~ shoredistance, family = Gamma(link = log),
##     data = obis_rajidae)
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.679e+00  3.143e-01  14.889  < 2e-16 ***
## shoredistance -1.529e-04  2.025e-05  -7.549 1.57e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 1.337373)
##
##     Null deviance: 120.355  on 46  degrees of freedom
## Residual deviance:  38.442  on 45  degrees of freedom
## AIC: 349.47
##
## Number of Fisher Scoring iterations: 13
```

```
confint(model2, level=0.95)
```

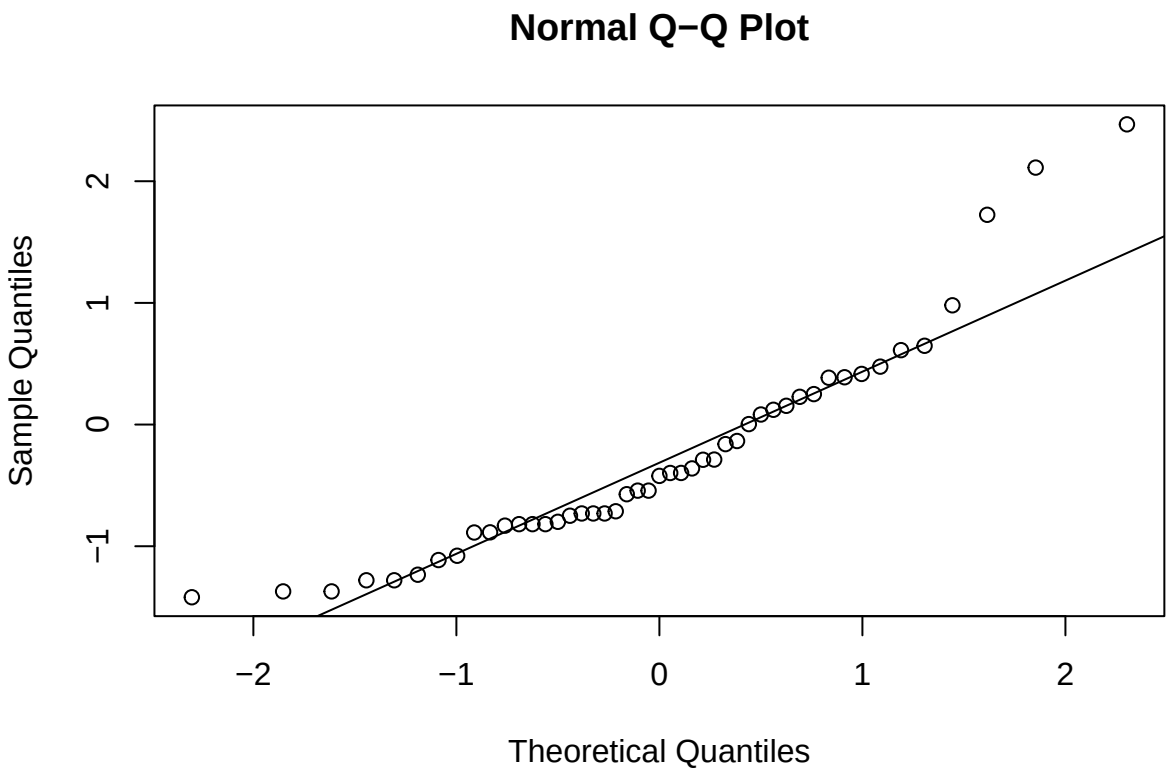
```
## Waiting for profiling to be done...
```

```
##                       2.5 %        97.5 %
## (Intercept)    4.1743529159  5.2616485477
## shoredistance -0.0001855223 -0.0001196208
```

```
hist(residuals(model2))
```



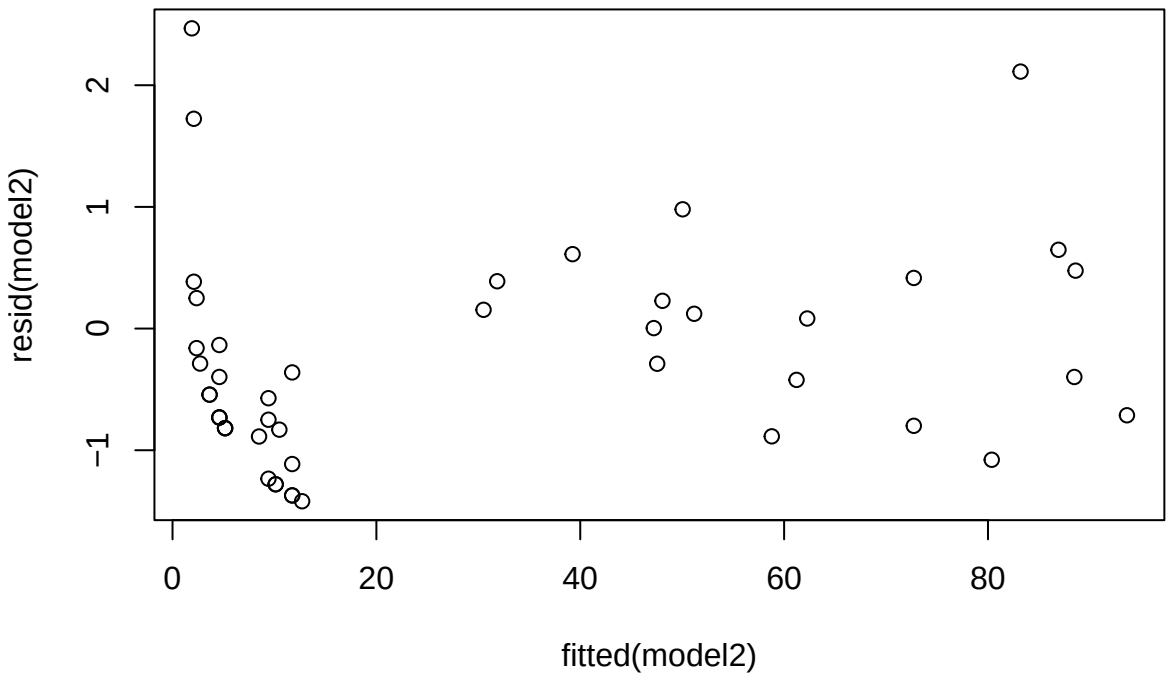

```
qqnorm(residuals(model2))
qqline(residuals(model2))
```




```
shapiro.test(residuals(model2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model2)
## W = 0.89952, p-value = 0.0006952
```

```
plot(fitted(model2), resid(model2))
```



Comparison of normality and homogeneity of the residual values of 'model2' (lognormal gamma distribution) with those of previous model (gaussian distribution; let's call it 'model1'):

1. Normality of residuals

- Shapiro-Wilk test:
  - model1's p-value = 5.437e-10
  - model2's p-value = 0.0006952

Both models do not pass the Shapiro-Wilk test, as both score p-values less than 0.05. However, model2 was able to increase the normality with higher p-value.

2. Homogeneity of residuals

- Plot of fitted values against residuals:
  - model1's plot show some discernible pattern,
  - model2's plot does not show a discernible pattern

Model2 indicates a more homogeneous residuals compared to model1.

Also should be added here that model2 also gives lower p-value for 'shoredistance' in its model summary compared to model1, which indicates a higher probability that it is a better fit to the data

Based on these two (+ additional one) measures, model2 gives a better fit to the data compared to model1.