Testing the Homozygosity per Island

Grace Saville

16/05/2022

1. Dataset loading

```
data <- read.csv("./data/RStudio/ratsSNPs_clean.csv")

copy <- data # making a copy of the data
head(colnames(copy), n = 20)
```

```
##  [1] "island"              "registration.number" "genus"
##  [4] "species"             "sex"                 "country"
##  [7] "state_province"      "island.1"            "locality"
## [10] "site"                "geo_lat"             "geo_long"
## [13] "collector"           "collecting.date"     "field.number"
## [16] "Populatie"           "X299_CHR1_114679736" "X13_CHR1_116614092"
## [19] "X14_CHR1_124857905"  "X15_CHR1_134869867"
```

```
copy <- copy[,-c(2:16)] # removing unnecessary columns for this analysis
```

2. Replaceing SNP's with symbols for heterozygous and homozygous

```
dim(copy) # 370 rows 283 columns
```

```
## [1] 370 283
```

```
unique(unlist(copy[,17:283])) # checking what SNP combinations are present
```

```
##  [1] "?"   "A:A" "G:G" "A:G" "T:T" "A:T" "C:C" "C:T" "T:G" "C:G" "A:C"
```

```
het <- c("A:G", "A:T", "C:T", "T:G", "C:G", "A:C") # vector heterozygous combinations
hom <- c("G:G", "A:A", "C:C", "T:T") # vector of homozygous combinations

# Replacing specific combos
for (i in 1:nrow(copy)){
  copy[i,][copy[i,] %in% het] <- "E"
```

```
  copy[i,][copy[i,] %in% hom] <- "O"
}

rm(i)

copy[copy == "?"] <- NA # replacing ? with NA's
```

3. Calculating heterozygous and homozygous totals per specimen

```
o.freq <- vector()
e.freq <- vector()
na.freq <- vector()
for (i in 1:370) {
  x <- sum(grepl("O", copy[i,], fixed = TRUE)) # counting row O's
  o.freq <- append(o.freq, x) # adding sum to vector
  x <- sum(grepl("E", copy[i,], fixed = TRUE)) # counting row E's
  e.freq <- append(e.freq, x)
  x <- sum(is.na(copy[i,])) # counting row NA's
  na.freq <- append(na.freq, x)
  rm(x)
}

rm(i)

per.specimen <- data.frame(copy$island, o.freq, e.freq, na.freq) # making a df
# with the freq totals
perc.o <- round(as.vector((
    per.specimen$o.freq / (per.specimen$o.freq + per.specimen$e.freq)
  ) * 100), digits = 3) # percentage without incl missing
per.specimen$perc.o <- perc.o

perc.missing <- round(as.vector((
    per.specimen$na.freq / (
      per.specimen$o.freq + per.specimen$e.freq + per.specimen$na.freq)
    ) * 100), digits = 3) # percentage of missing data in row/specimen

per.specimen$perc.missing <- perc.missing

rm(o.freq, e.freq, na.freq, perc.o, perc.missing)

str(per.specimen)
```

```
## 'data.frame':    370 obs. of  6 variables:
##  $ copy.island : chr  "Borneo_002" "Borneo_003" "Borneo_004" "Borneo_005" ...
##  $ o.freq      : int  105 135 149 95 98 137 153 198 97 152 ...
##  $ e.freq      : int  15 18 34 39 26 17 35 45 28 26 ...
##  $ na.freq     : int  162 129 99 148 158 128 94 39 157 104 ...
##  $ perc.o      : num  87.5 88.2 81.4 70.9 79 ...
##  $ perc.missing: num  57.4 45.7 35.1 52.5 56 ...
```

4. Calculating heterozygous and homozygous totals per island

```
data <- data[order(data$island, decreasing = FALSE),] # ordering df alphabetically
head(names(data), n = 20)
```

```
##  [1] "island"              "registration.number" "genus"
##  [4] "species"             "sex"                 "country"
##  [7] "state_province"      "island.1"            "locality"
## [10] "site"                "geo_lat"             "geo_long"
## [13] "collector"           "collecting.date"     "field.number"
## [16] "Populatie"           "X299_CHR1_114679736" "X13_CHR1_116614092"
## [19] "X14_CHR1_124857905"  "X15_CHR1_134869867"
```

```
data[c(grep("Mainland", data$island.1)),c(1, 8)] # checking which populations fall
```

```
##         island island.1
## 365 Cambodia_1 Mainland
## 363 Laos___001 Mainland
## 364 Laos___002 Mainland
## 369 Thailand01 Mainland
## 370 Thailand02 Mainland
```

```
# in the mainland category
# unique(data$island)
shortpopnames <- as.character(
  c(
    "aotea",
    "borneo",
    "cambodia",
    "grtmercury",
    "halmaher",
    "hatutaa",
    "honuea",
    "kaikura",
    "kamaka",
    "kayangel",
    "laos",
    "late",
    "luzon",
    "malenge",
    "mohotani",
    "motukawa",
    "newbrita",
    "newguine",
    "normanby",
    "rakiura",
    "reiono",
    "rimatuu",
    "slipper",
    "southland",
    "sulawesi",
```

```
    "tahanea",
    "thailand",
    "wake"
  )
) # writing shortened names as is in the data df so I can use the character
# strings with grep()

# splitting the data df into df objects by island:
for (i in 1:length(shortpopnames)) {
  y <- as.vector(grep(
    shortpopnames[i], copy[,1], ignore.case = TRUE, value = FALSE))
  assign(paste(shortpopnames[i]), copy[y,])
}
rm(y, i)

# making a "mainland" df:
mainland <- rbind(cambodia, thailand, laos)
rm(cambodia, thailand, laos)
shortpopnames
```

```
##  [1] "aotea"      "borneo"     "cambodia"   "grtmercury" "halmaher"
##  [6] "hatutaa"    "honuea"     "kaikura"    "kamaka"     "kayangel"
## [11] "laos"       "late"       "luzon"      "malenge"    "mohotani"
## [16] "motukawa"   "newbrita"   "newguine"   "normanby"   "rakiura"
## [21] "reiono"     "rimatuu"    "slipper"    "southland"  "sulawesi"
## [26] "tahanea"    "thailand"   "wake"
```

```
shortpopnames <- shortpopnames[-c(3, 11, 27)] # removing the names now in mainland
shortpopnames <- append(shortpopnames, "mainland")

# double checking the dfs have the correct specimens per island:
length(shortpopnames) #26
```

```
## [1] 26
```

```
for (i in 1:26) {
  print(shortpopnames[i])
  y <- get(shortpopnames[i])
  print(y[,1])
}
```

```
## [1] "aotea"
##  [1] "Aotea__001" "Aotea__002" "Aotea__003" "Aotea__004" "Aotea__005"
##  [6] "Aotea__006" "Aotea__007" "Aotea__008" "Aotea__009" "Aotea__010"
## [1] "borneo"
##  [1] "Borneo_002" "Borneo_003" "Borneo_004" "Borneo_005" "Borneo_006"
##  [6] "Borneo_008" "Borneo_012" "Borneo_013" "Borneo_014" "Borneo_015"
## [11] "Borneo_016" "Borneo_017" "Borneo_019" "Borneo_020" "Borneo_021"
## [16] "Borneo_022" "Borneo_023" "Borneo_024"
## [1] "grtmercury"
## [1] "GrtMercury"
## [1] "halmaher"
```

```
##  [1] "Halmahera1" "Halmahera2" "Halmahera3" "Halmahera6" "Halmahera7"
##  [6] "Halmahera8" "Halmahera9" "Halmaher10" "Halmaher11" "Halmaher19"
## [11] "Halmaher23" "Halmaher24"
## [1] "hatutaa"
##  [1] "Hatutaa_01" "Hatutaa_02" "Hatutaa_03" "Hatutaa_04" "Hatutaa_05"
##  [6] "Hatutaa_06" "Hatutaa_07" "Hatutaa_08" "Hatutaa_09" "Hatutaa_10"
## [11] "Hatutaa_11" "Hatutaa_12" "Hatutaa_13" "Hatutaa_14" "Hatutaa_15"
## [16] "Hatutaa_16" "Hatutaa_17" "Hatutaa_18" "Hatutaa_19" "Hatutaa_20"
## [21] "Hatutaa_21"
## [1] "honuea"
##  [1] "Honuea_001" "Honuea_002" "Honuea_003" "Honuea_004" "Honuea_005"
##  [6] "Honuea_006" "Honuea_007" "Honuea_008" "Honuea_009" "Honuea_010"
## [11] "Honuea_011" "Honuea_012" "Honuea_013" "Honuea_015" "Honuea_016"
## [16] "Honuea_017" "Honuea_018" "Honuea_019" "Honuea_020" "Honuea_021"
## [1] "kaikura"
##  [1] "Kaikura_01" "Kaikura_02" "Kaikura_03" "Kaikura_04" "Kaikura_05"
##  [6] "Kaikura_06" "Kaikura_07" "Kaikura_08" "Kaikura_09" "Kaikura_10"
## [11] "Kaikura_11" "Kaikura_12" "Kaikura_13" "Kaikura_14" "Kaikura_15"
## [16] "Kaikura_16" "Kaikura_17" "Kaikura_18" "Kaikura_19" "Kaikura_20"
## [1] "kamaka"
##  [1] "Kamaka_001" "Kamaka_002" "Kamaka_003" "Kamaka_004" "Kamaka_005"
##  [6] "Kamaka_006" "Kamaka_007" "Kamaka_008" "Kamaka_010" "Kamaka_011"
## [11] "Kamaka_012" "Kamaka_013" "Kamaka_014" "Kamaka_015" "Kamaka_016"
## [16] "Kamaka_017" "Kamaka_018" "Kamaka_019" "Kamaka_020" "Kamaka_021"
## [1] "kayangel"
##  [1] "Kayangel01" "Kayangel02" "Kayangel03" "Kayangel04" "Kayangel05"
##  [6] "Kayangel06" "Kayangel07" "Kayangel08" "Kayangel09" "Kayangel10"
## [11] "Kayangel12" "Kayangel14" "Kayangel16" "Kayangel18" "Kayangel20"
## [1] "late"
##  [1] "Late_Is_01" "Late_Is_02" "Late_Is_03" "Late_Is_04" "Late_Is_05"
##  [6] "Late_Is_06" "Late_Is_07" "Late_Is_08" "Late_Is_09" "Late_Is_10"
## [11] "Late_Is_11" "Late_Is_12" "Late_Is_13" "Late_Is_14" "Late_Is_15"
## [16] "Late_Is_16" "Late_Is_17" "Late_Is_18" "Late_Is_19" "Late_Is_20"
## [21] "Late_Is_21"
## [1] "luzon"
## [1] "Luzon__001"
## [1] "malenge"
##  [1] "Malenge_04" "Malenge_11" "Malenge_14" "Malenge_15" "Malenge_17"
##  [6] "Malenge_18" "Malenge_19" "Malenge_20" "Malenge_22" "Malenge_23"
## [11] "Malenge_24" "Malenge_25"
## [1] "mohotani"
##  [1] "Mohotani01" "Mohotani02" "Mohotani03" "Mohotani04" "Mohotani05"
##  [6] "Mohotani06" "Mohotani07" "Mohotani08" "Mohotani09" "Mohotani10"
## [11] "Mohotani11" "Mohotani12" "Mohotani13" "Mohotani14"
## [1] "motukawa"
##  [1] "Motukawan1" "Motukawan2" "Motukawan3" "Motukawan4" "Motukawan5"
##  [6] "Motukawan6" "Motukawan7" "Motukawan8" "Motukawan9" "Motukawa10"
## [11] "Motukawa11" "Motukawa12" "Motukawa13" "Motukawa14" "Motukawa15"
## [16] "Motukawa16" "Motukawa17" "Motukawa18" "Motukawa19" "Motukawa20"
## [21] "Motukawa21"
## [1] "newbrita"
##  [1] "NewBritai2" "NewBritai3" "NewBritai6" "NewBritai7" "NewBritai8"
##  [6] "NewBrita14" "NewBrita16" "NewBrita23" "NewBrita24" "NewBrita25"
## [1] "newguine"
```

```
## [1] "NewGuinea1" "NewGuinea8" "NewGuine17"
## [1] "normanby"
## [1] "Normanby08"
## [1] "rakiura"
##  [1] "Rakiura_01" "Rakiura_02" "Rakiura_03" "Rakiura_04" "Rakiura_05"
##  [6] "Rakiura_06" "Rakiura_07" "Rakiura_08" "Rakiura_09" "Rakiura_10"
## [11] "Rakiura_11" "Rakiura_12" "Rakiura_13" "Rakiura_14" "Rakiura_15"
## [16] "Rakiura_16" "Rakiura_17" "Rakiura_18" "Rakiura_19" "Rakiura_20"
## [21] "Rakiura_21"
## [1] "reiono"
##  [1] "Reiono_001" "Reiono_002" "Reiono_003" "Reiono_004" "Reiono_005"
##  [6] "Reiono_006" "Reiono_007" "Reiono_008" "Reiono_009" "Reiono_010"
## [11] "Reiono_011" "Reiono_012" "Reiono_013" "Reiono_014" "Reiono_015"
## [16] "Reiono_016" "Reiono_017" "Reiono_018" "Reiono_019" "Reiono_020"
## [21] "Reiono_021"
## [1] "rimatuu"
##  [1] "Rimatuu_01" "Rimatuu_02" "Rimatuu_03" "Rimatuu_04" "Rimatuu_05"
##  [6] "Rimatuu_06" "Rimatuu_07" "Rimatuu_08" "Rimatuu_09" "Rimatuu_10"
## [11] "Rimatuu_11" "Rimatuu_12" "Rimatuu_13" "Rimatuu_14" "Rimatuu_15"
## [16] "Rimatuu_16" "Rimatuu_17" "Rimatuu_18" "Rimatuu_21"
## [1] "slipper"
##  [1] "Slipper_01" "Slipper_02" "Slipper_03" "Slipper_04" "Slipper_05"
##  [6] "Slipper_06" "Slipper_07" "Slipper_08" "Slipper_09" "Slipper_10"
## [11] "Slipper_11" "Slipper_12" "Slipper_13" "Slipper_14" "Slipper_15"
## [16] "Slipper_16" "Slipper_17" "Slipper_18" "Slipper_19" "Slipper_20"
## [21] "Slipper_21"
## [1] "southland"
## [1] "Southland1"
## [1] "sulawesi"
##  [1] "Sulawesi_1" "Sulawesi_2" "Sulawesi_3" "Sulawesi_4" "Sulawesi_5"
##  [6] "Sulawesi_6" "Sulawesi_7" "Sulawesi_8" "Sulawesi_9" "Sulawesi10"
## [11] "Sulawesi11" "Sulawesi12" "Sulawesi13" "Sulawesi14" "Sulawesi15"
## [16] "Sulawesi16" "Sulawesi17" "Sulawesi18" "Sulawesi19" "Sulawesi20"
## [21] "Sulawesi21" "Sulawesi22"
## [1] "tahanea"
##  [1] "Tahanea_01" "Tahanea_02" "Tahanea_03" "Tahanea_04" "Tahanea_05"
##  [6] "Tahanea_06" "Tahanea_07" "Tahanea_08" "Tahanea_09" "Tahanea_10"
## [11] "Tahanea_11" "Tahanea_12" "Tahanea_13" "Tahanea_14" "Tahanea_15"
## [16] "Tahanea_16" "Tahanea_17" "Tahanea_18" "Tahanea_19" "Tahanea_20"
## [1] "wake"
##  [1] "Wake_Is_01" "Wake_Is_02" "Wake_Is_03" "Wake_Is_04" "Wake_Is_05"
##  [6] "Wake_Is_06" "Wake_Is_07" "Wake_Is_08" "Wake_Is_09" "Wake_Is_10"
## [11] "Wake_Is_11" "Wake_Is_12" "Wake_Is_13" "Wake_Is_14" "Wake_Is_15"
## [16] "Wake_Is_16" "Wake_Is_17" "Wake_Is_18" "Wake_Is_19" "Wake_Is_20"
## [1] "mainland"
## [1] "Cambodia_1" "Thailand01" "Thailand02" "Laos___001" "Laos___002"
```

```
rm(y)
```

```
# counting values:
per.island <- data.frame()
for (i in 1:length(shortpopnames)) {
  o <- sum(grepl("O", unlist(get(shortpopnames[i])), fixed = TRUE)) # counting row O's
  e <- sum(grepl("E", unlist(get(shortpopnames[i])), fixed = TRUE)) # counting row E's
```

```
  na <- sum(is.na(get(shortpopnames[i]))) # counting row NA's
  vec <- as.vector(c(shortpopnames[i], nrow(get(shortpopnames[i])), o, e, na))
  per.island <- rbind(per.island, vec)
}
rm(o, e, na, vec, i)

colnames(per.island) <- c("island", "specimens.based.on", "o.freq", "e.freq", "na.freq") # renaming col
str(per.island)
```

```
## 'data.frame':    26 obs. of  5 variables:
##  $ island            : chr  "aotea" "borneo" "grtmercury" "halmaher" ...
##  $ specimens.based.on: chr  "10" "18" "1" "12" ...
##  $ o.freq            : chr  "2505" "2622" "258" "1756" ...
##  $ e.freq            : chr  "232" "464" "24" "296" ...
##  $ na.freq           : chr  "83" "1990" "0" "1332" ...
```

```
per.island$o.freq <- as.integer(per.island$o.freq) # changing the numbers from
# characters to integers
per.island$e.freq <- as.integer(per.island$e.freq)
per.island$na.freq <- as.integer(per.island$na.freq)

# calculating percentages:
perc.o <- round(as.vector((
  per.island$o.freq / (per.island$o.freq + per.island$e.freq)) * 100), digits = 3)
# percentage without incl missing
per.island$perc.o <- perc.o
perc.missing <-
  round(as.vector((
    per.island$na.freq / (per.island$o.freq + per.island$e.freq + per.island$na.freq)
  ) * 100), digits = 3) # percentage of missing data in row/specimen
per.island$perc.missing <- perc.missing

rm(perc.o, perc.missing, het, hom)
rm(list = shortpopnames)
```

5. Saving

```
# getwd()
# write.csv(per.island, "./results/heterozygosity_testing_results_table.csv", row.names = FALSE)
```

6. Considering island size and distance from mainland

```
island.km <- read.csv("./data/Raw_data/island_size_data.csv", header = TRUE)
head(island.km, n = 10)
```

```
##    ID              ISLAND ALTERNATIVE_ISLAND_NAME  area_km2 elevation_m
```

```
## 1  19                Aotea    Great Barrier island    285.00         627
## 2   1               Borneo                           748168.10        4175
## 3  27       Doubtful Sound                    Patea 150437.00        3724
## 4  26 Great Mercury Island                   Ahuahu     18.72         231
## 5   4            Halmahera Jilolo, Gilolo, Jailolo  18039.60        1635
## 6  13              Hatutaa                   Hatutu      7.98         428
## 7  15               Honuea                               0.28           8
## 8  17       Kaikura Island Kaik?ura, Selwyn island      5.64         185
## 9   9               Kamaka                               0.50         166
## 10 12             Kayangel              Ngcheangel      1.70           2
##       Larger_unit Archipelago_Group           Smaller_unit     REGION  X X.1
## 1     New Zealand       New Zealand              Northland  Polynesia NA  NA
## 2            PhBS         Indonesia  Greater Sunda Islands      Sunda NA  NA
## 3     New Zealand       New Zealand Fiordland on Southland  Polynesia NA  NA
## 4     New Zealand       New Zealand              Northland  Polynesia NA  NA
## 5        Moluccas    Maluku Islands                          Wallacea NA  NA
## 6  Remote Oceania         Marquesas                         Polynesia NA  NA
## 7  Remote Oceania   Society Islands         Tetiaroa atoll  Polynesia NA  NA
## 8     New Zealand       New Zealand              Northland  Polynesia NA  NA
## 9  Remote Oceania   Gambier Islands                         Polynesia NA  NA
## 10     Micronesia  Caroline Islands                 Palau Micronesia NA  NA
##    X.2 X.3 X.4 X.5 X.6 X.7 X.8 X.9 X.10 X.11 X.12 X.13 X.14 X.15 X.16 X.17 X.18
## 1   NA  NA  NA  NA  NA  NA  NA  NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## 2   NA  NA  NA  NA  NA  NA  NA  NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## 3   NA  NA  NA  NA  NA  NA  NA  NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## 4   NA  NA  NA  NA  NA  NA  NA  NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## 5   NA  NA  NA  NA  NA  NA  NA  NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## 6   NA  NA  NA  NA  NA  NA  NA  NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## 7   NA  NA  NA  NA  NA  NA  NA  NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## 8   NA  NA  NA  NA  NA  NA  NA  NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## 9   NA  NA  NA  NA  NA  NA  NA  NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## 10  NA  NA  NA  NA  NA  NA  NA  NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
##    X.19 X.20 X.21 X.22 X.23 X.24 X.25 X.26 X.27 X.28 X.29 X.30 X.31 X.32 X.33
## 1    NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## 2    NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## 3    NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## 4    NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## 5    NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## 6    NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## 7    NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## 8    NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## 9    NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## 10   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
##    X.34 X.35 X.36
## 1    NA   NA   NA
## 2    NA   NA   NA
## 3    NA   NA   NA
## 4    NA   NA   NA
## 5    NA   NA   NA
## 6    NA   NA   NA
## 7    NA   NA   NA
## 8    NA   NA   NA
## 9    NA   NA   NA
## 10   NA   NA   NA
```

```
island.km <- island.km[,c(2, 4)]
island.km <- island.km[order(island.km$ISLAND, decreasing = FALSE),] # sorting
```

```
distance.from.ML <- data[,c(1,8,11,12)]
distance.from.ML <- distance.from.ML[!duplicated(distance.from.ML$island.1),]
# keeping only 1 coordinate for each island
distance.from.ML <- distance.from.ML[
  order(distance.from.ML$island.1, decreasing = FALSE),] # sorting alphabetically
row.names(distance.from.ML) <- seq(nrow(distance.from.ML)) # renaming row numbers
# to be sequential
head(distance.from.ML) # checking
```

```
##       island                island.1   geo_lat  geo_long
## 1 Aotea__001 Aotea (Great Barrier I) -36.23000  175.4300
## 2 Borneo_002                  Borneo  -0.51020  117.0912
## 3 Southland1          Doubtful Sound -45.31667  166.9833
## 4 GrtMercury    Great Mercury Island -36.58333  175.9167
## 5 Halmaher10               Halmahera   1.16250  127.8872
## 6 Hatutaa_01                 Hatutaa  -7.92000 -140.5700
```

```
#shortening some of the long island names to match what is already in the island.km df:
distance.from.ML[1,2] <- "Aotea"
distance.from.ML[20,2] <- "Rakiura"
distance.from.ML[22,2] <- "Rimatuu"

#decided to keep Cambodia as the mainland coordinates, will use this as the
# base for distance from mainland
distance.from.ML <- distance.from.ML[,-1] # removing specimen ID column since
# no longer necessary
```

```
km.from.ML <- vector()
for (i in 1:nrow(distance.from.ML)) {
  x <- distGeo(as.vector(distance.from.ML[13,c(3,2)]), as.vector(distance.from.ML[i,c(3,2)]))
  km.from.ML <- append(km.from.ML, x)
}

km.from.ML <- km.from.ML / 1000 # converting from metres to kilometres
distance.from.ML$km.from.ML <- km.from.ML # adding the kms to the distance df

rm(i, x, km.from.ML)
```

7. Merging the dataframes for easier analysis

```
names(island.km)
```

```
## [1] "ISLAND"   "area_km2"
```

```
names(distance.from.ML)
```

```
## [1] "island.1"   "geo_lat"    "geo_long"   "km.from.ML"
```

```
island.km <- merge(
  island.km, distance.from.ML, by.x = "ISLAND", by.y = "island.1", all = TRUE)

rm(distance.from.ML)

island.km$ISLAND
```

```
##  [1] "Aotea"                "Borneo"               "Doubtful Sound"
##  [4] "Great Mercury Island" "Halmahera"            "Hatutaa"
##  [7] "Honuea"               "Kaikura Island"       "Kamaka"
## [10] "Kayangel"             "Late Island"          "Luzon"
## [13] "Mainland"             "Malenge"              "Mohotani"
## [16] "Motukawanui"          "New Britain"          "New Guinea"
## [19] "Normanby Island"      "Rakiura"              "Reiono"
## [22] "Rimatuu"              "Slipper Island"       "Sulawesi"
## [25] "Tahanea"              "Wake Island"
```

```
per.island$island
```

```
##  [1] "aotea"      "borneo"     "grtmercury" "halmaher"   "hatutaa"
##  [6] "honuea"     "kaikura"    "kamaka"     "kayangel"   "late"
## [11] "luzon"      "malenge"    "mohotani"   "motukawa"   "newbrita"
## [16] "newguine"   "normanby"   "rakiura"    "reiono"     "rimatuu"
## [21] "slipper"    "southland"  "sulawesi"   "tahanea"    "wake"
## [26] "mainland"
```

```
per.island$island <- c(
  "Aotea",
  "Borneo",
  "Great Mercury Island",
  "Halmahera",
  "Hatutaa",
  "Honuea",
  "Kaikura Island",
  "Kamaka",
  "Kayangel",
  "Late Island",
  "Luzon",
  "Malenge",
  "Mohotani",
  "Motukawanui",
  "New Britain",
  "New Guinea",
  "Normanby Island",
  "Rakiura",
  "Reiono",
  "Rimatuu",
```

```
  "Slipper Island",
  "Doubtful Sound",
  "Sulawesi",
  "Tahanea",
  "Wake Island",
  "Mainland"
) # editing the names to match those in the other df so I can merge them

per.island <- merge(per.island, island.km, by.x = "island", by.y = "ISLAND", all = TRUE)
```

8. Saving again

```
# getwd()
# write.csv(per.island, "./results/RStudio_Homozygosity/homozygosity_testing_results_table.csv", row.na
```

9. Linear regression and Plots

```
per.island <- read.csv(
  "./results/RStudio_Homozygosity/homozygosity_testing_results_table.csv", header = TRUE)
```

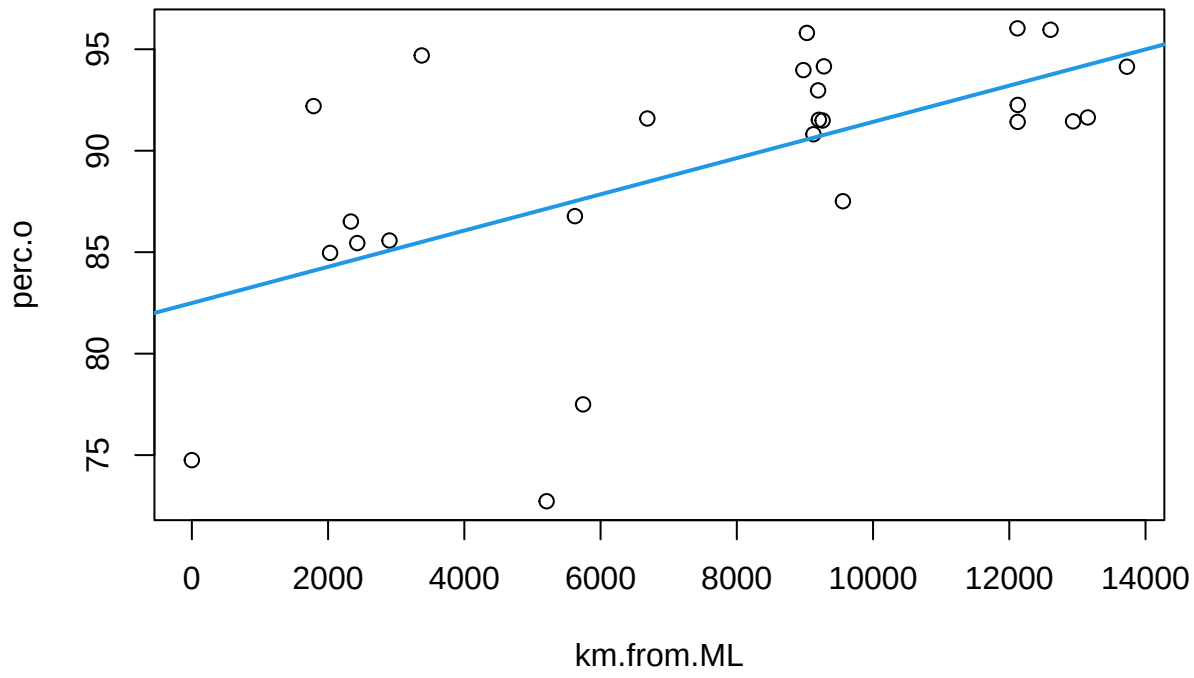
Issues to consider before/during the statistical analyses:

- Some islands may be outliers because they are based on too few specimens (e.g. due to removal because of too many missing SNPs)
- Some islands may have specimens that are closely related to each other if they were sampled from the same site therefore may not accurately represent the population
- Sample size is small therefore diagnostics could be overinterpreted and results might not represent reality well
- Distance from the mainland may not be the best measure because it doesn't indicate the difficultly of reaching the island in all cases, from example Normanby isl. is right next to New Guinea which is very large and likely has a diverse population that can easily move to Normanby and back.

9a. Distance

i. Model Construction and Diagnostics

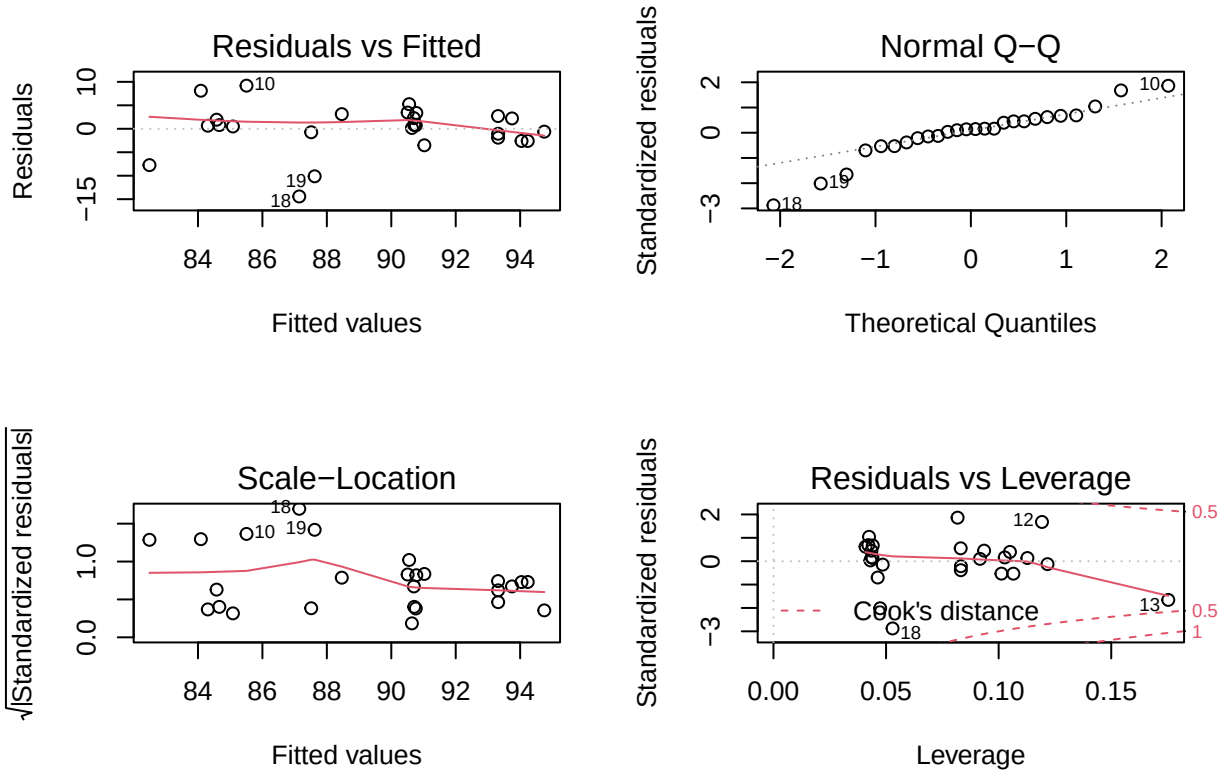
```
testLM <- lm(perc.o ~ km.from.ML, data = per.island) # model

plot(perc.o ~ km.from.ML, data = per.island)
abline(coef = coef(testLM), col = 4, lwd = 2)
```



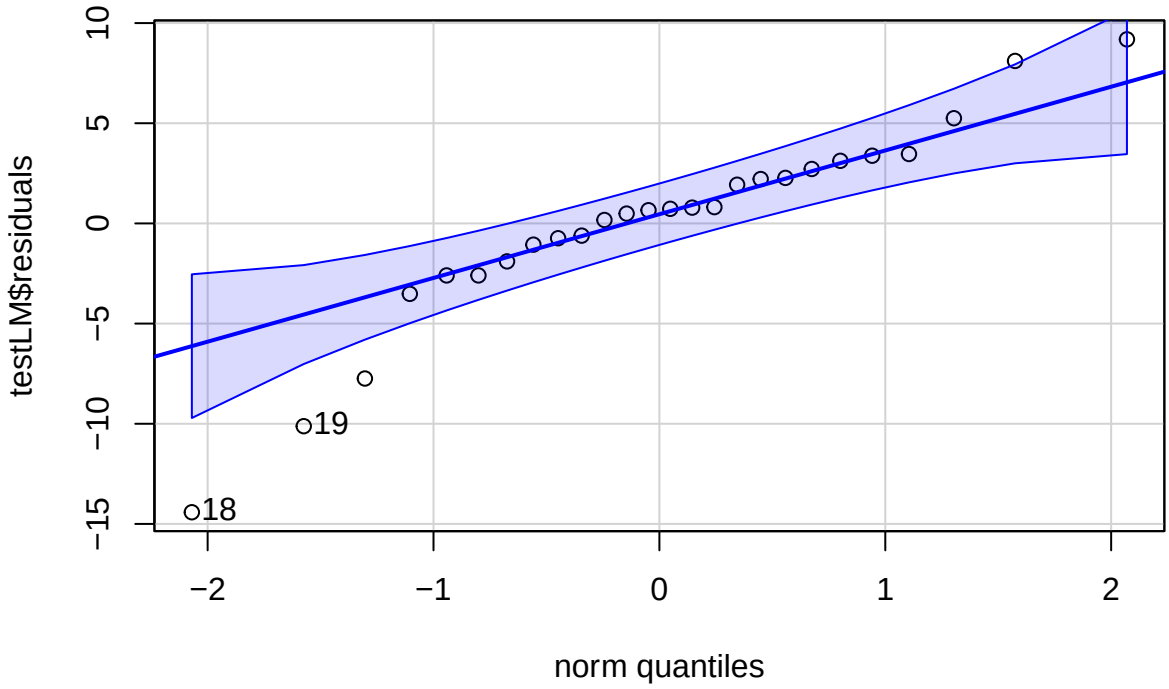

```
summary(testLM) # model results
```

```
##
## Call:
## lm(formula = perc.o ~ km.from.ML, data = per.island)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4165  -1.6867   0.6933   2.6053   9.1871
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.249e+01  2.156e+00  38.262  < 2e-16 ***
## km.from.ML  8.927e-04  2.470e-04   3.614  0.00139 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.149 on 24 degrees of freedom
## Multiple R-squared:  0.3525, Adjusted R-squared:  0.3255
## F-statistic: 13.06 on 1 and 24 DF,  p-value: 0.001387
```

```
par(mfrow = c(2, 2)) # changes the number of plots visible at once
plot(testLM) # diagnostic plots
```




```
par(mfrow = c(1, 1))
qqPlot(testLM$residuals, line = "quartiles") # non-normal dist
```




```
## [1] 18 19
```

```
shapiro.test(testLM$residuals) # indicates non-normality of resid
```

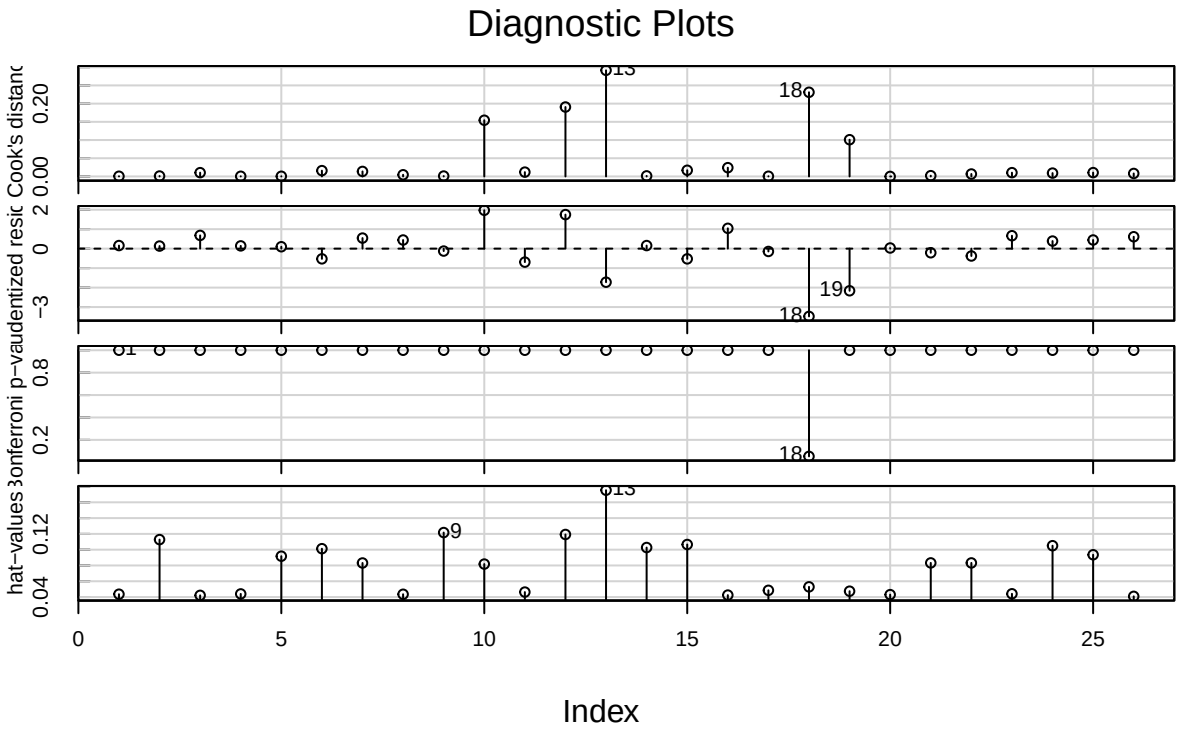
```
##
##  Shapiro-Wilk normality test
##
## data:  testLM$residuals
## W = 0.91825, p-value = 0.04092
```

```
hist(testLM$residuals, breaks = 10) #
```


```
ncvTest(testLM) # homoscedasticity test: H0 of constant variance is rejected
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 5.719721, Df = 1, p = 0.016775
```

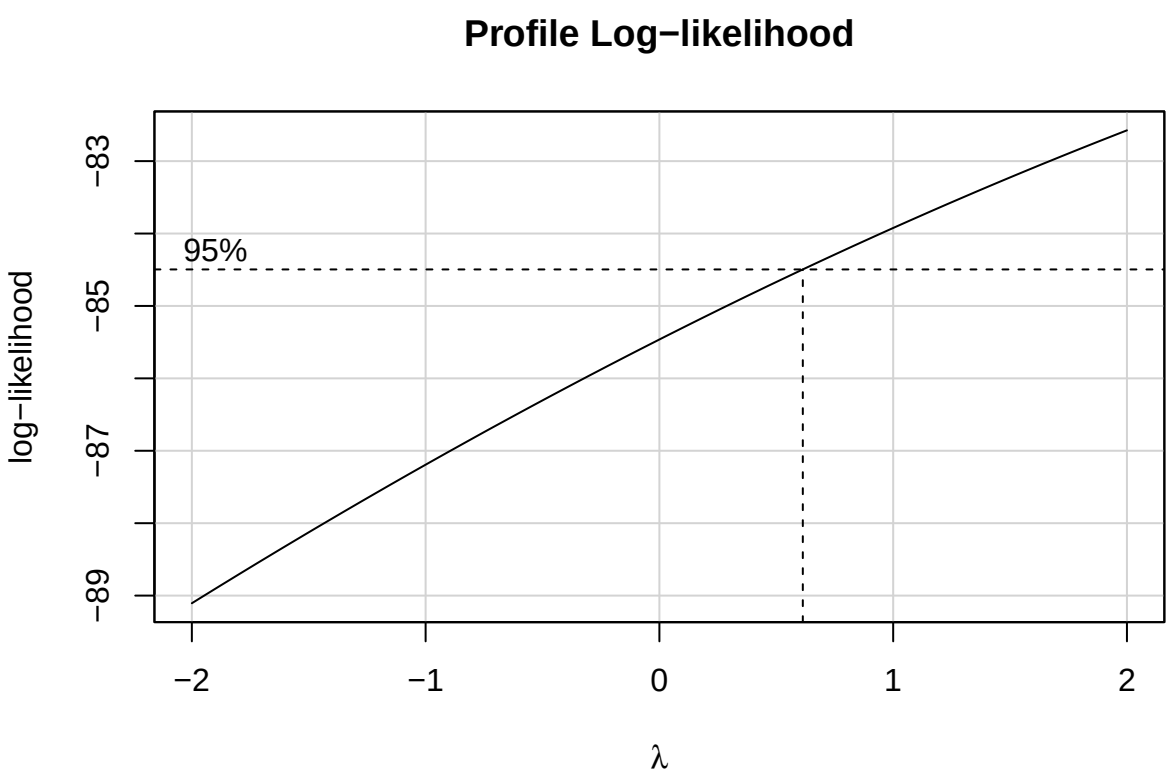
```
influenceIndexPlot(testLM) # outliers
```




```
# Cooks distances: none larger than 0.5,
# Studentised residuals: point 18 less than -3
# Bonferroni p-value: point 18 smaller then 0.05,
# Hat-values: point 13 influential, higher than 1
outlierTest(testLM)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 18 -3.479726          0.0020257     0.052668
```

```
boxCox(testLM) # suggests sqrt transformation
```



Notes on diagnostic plots

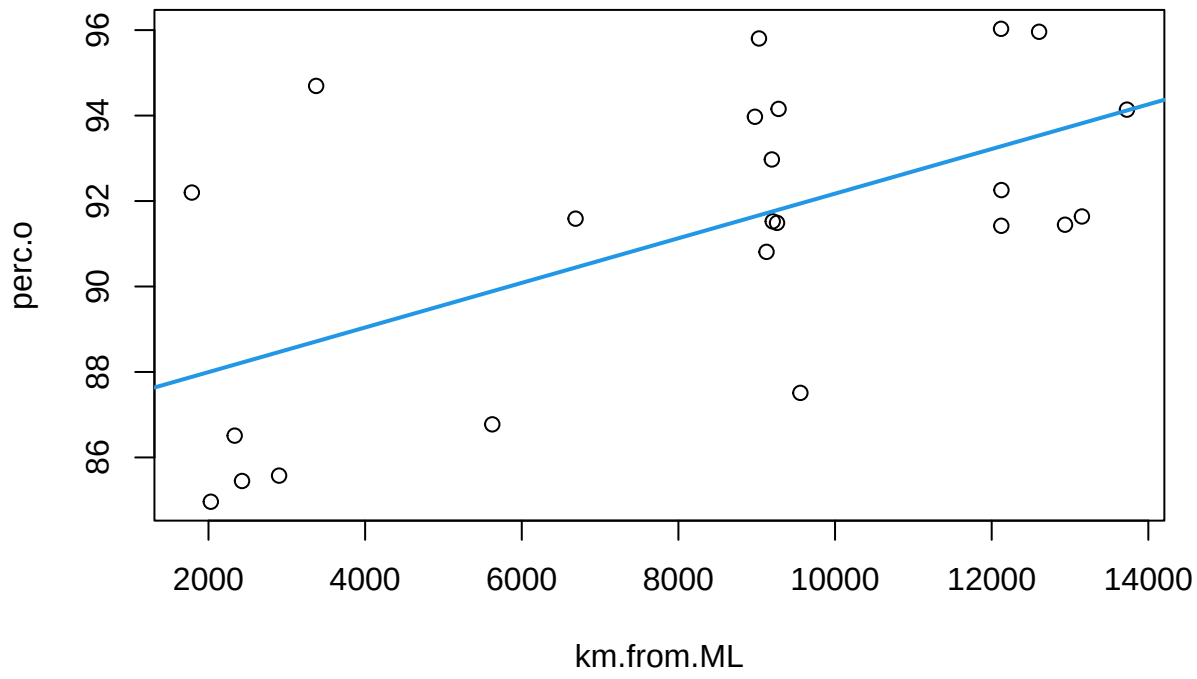
- Residuals vs. Fitted plot shows a line not quite horizontal, but close enough to flat consider the relationship linear. Cone shape indicates heteroscedasticity. Questionable points: 10 (Kayangel), 18 (New Guinea), 19 (Normanby Island).
- Normal QQ plot shows residuals generally in line, indicating normality, however the tails are fat ant bottom tail prominent, indicating fat tails and/or slight left-skewed distribution (towards the right). The Shapiro-Wilk test (only just) rejects H0 that residuals are normal. Questionable points are again: 10 (Kayangel), 18 (New Guinea), 19 (Normanby Island).
- Scale-Location plot is used to indicate constant residual variance with a line that does not trend up or down overall. Here it may be flat or trend down however it is not entirely clear due to the questionable points (10 (Kayangel), 18 (New Guinea), 19 (Normanby Island)) pulling a section upwards. The NCV test indicates non-constant variance.
- Residuals vs Leverage (Cook's distance) plot assesses outliers. None of the points are over 1 (Cook's distance line which would make them statistical outliers) nor are any over 0.5 (which would make them questionable). Point 13 (Mainland) has high leverage, however this makes sense because the distance to the Mainland from the Mainland is 0, which is an unusually low number in this model. Point 18 (New Guinea) is identified as a potential issue in the outlier tests.

ii. Adjusted Model and Diagnostics

I have decided to remove the Mainland point for several reasons; it is an outlier with a high leverage and a value of 0, also not valuable since the Mainland is the reference point, not a new value. Testing removing 18 and 19 (New Guinea and Normanby island) saw an improvement in normality, which makes the model more reliable, however I don't think they are true outliers outside of the dataset, so the results from this

model will come with this caveat. Additionally, New Guinea and Normanby island are based off of 3 and 1 specimen respectively, which mean the points may not represent their populations accurately.

```
z <- per.island[-c(13, 18, 19),] # removing points mentioned
LM <- lm(perc.o ~ km.from.ML, data = z) # model

plot(perc.o ~ km.from.ML, data = z)
abline(coef = coef(LM), col = 4, lwd = 2)
```

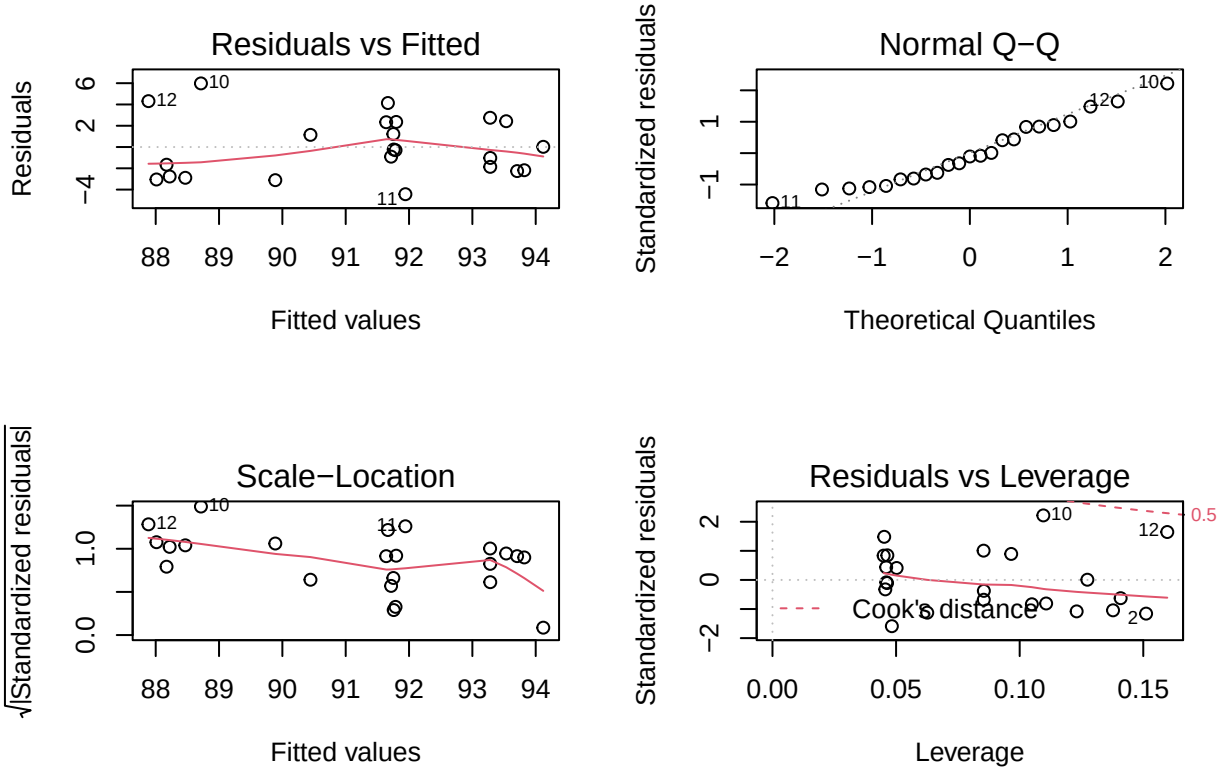



```
summary(LM) # model results
```

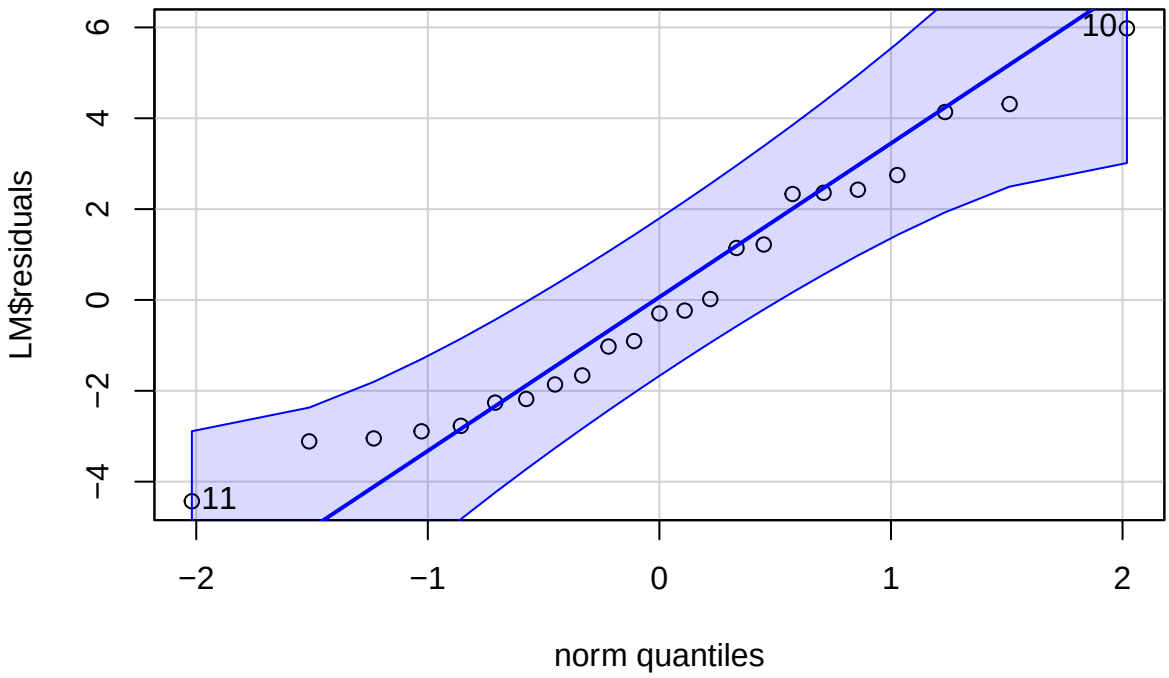
```
##
## Call:
## lm(formula = perc.o ~ km.from.ML, data = z)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4315 -2.2204 -0.2978  2.3458  5.9795
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.695e+01  1.379e+00  63.035  < 2e-16 ***
## km.from.ML  5.221e-04  1.509e-04   3.459  0.00235 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.858 on 21 degrees of freedom
## Multiple R-squared:  0.363,  Adjusted R-squared:  0.3326
## F-statistic: 11.96 on 1 and 21 DF,  p-value: 0.002349
```

```
par(mfrow = c(2, 2))
plot(LM) # diagnostic plots
```




```
par(mfrow = c(1, 1))
qqPlot(LM$residuals, line = "quartiles") # normal residual distribution
```

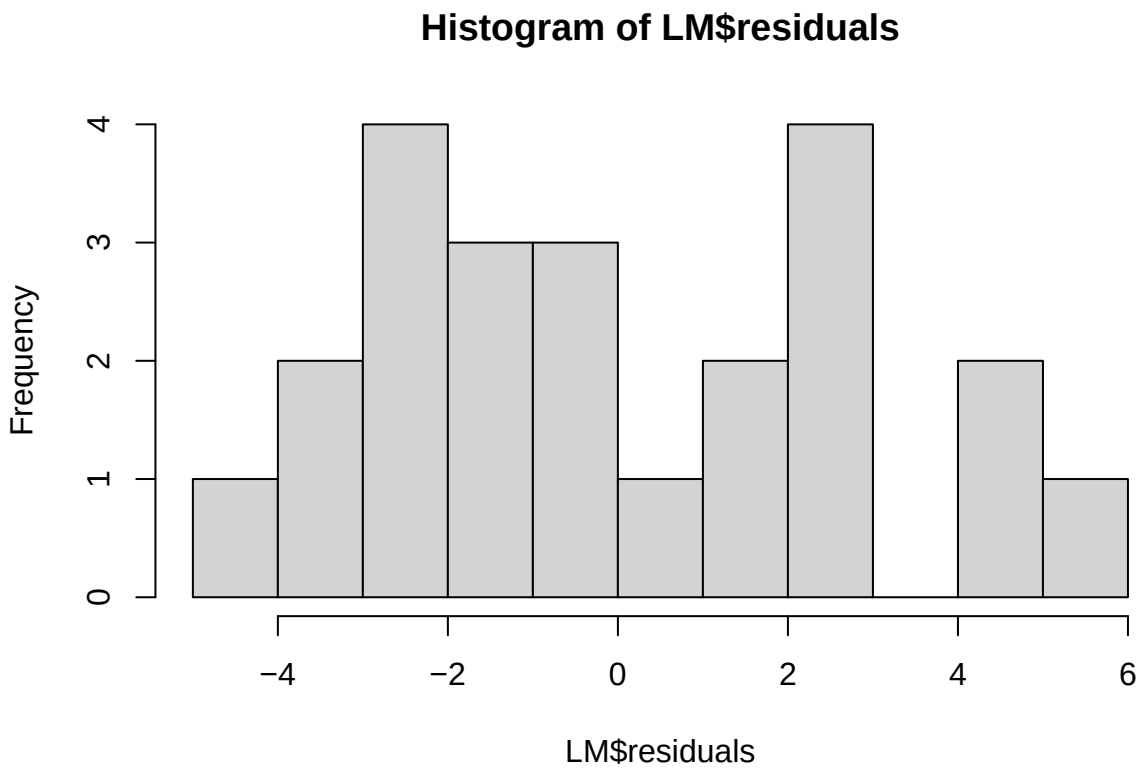



```
## [1] 10 11
```

```
shapiro.test(LM$residuals) # indicates normality of residuals
```

```
##
##  Shapiro-Wilk normality test
##
## data:  LM$residuals
## W = 0.95512, p-value = 0.3723
```

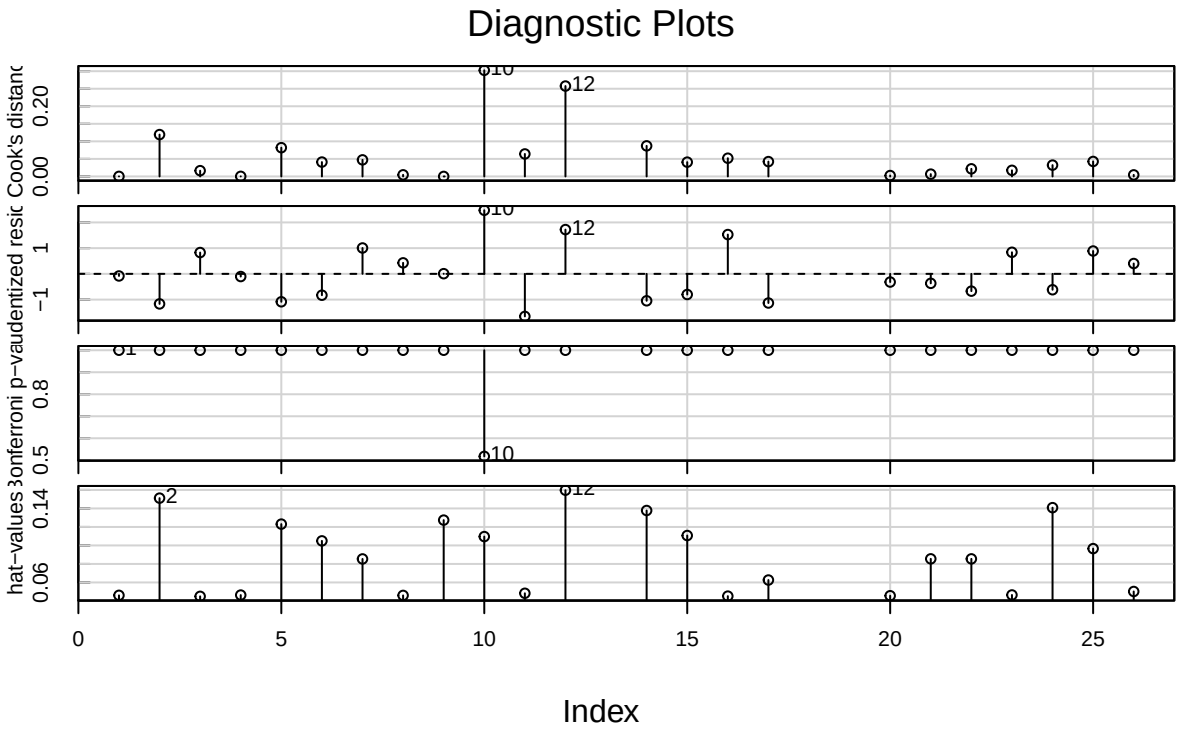
```
hist(LM$residuals, breaks = 10)
```




```
ncvTest(LM) # homoscedasticity test: H0 of constant variance is not rejected
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 2.469013, Df = 1, p = 0.11611
```

```
influenceIndexPlot(LM) # nothing overly concerning
```




```
outlierTest(LM)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 10 2.472633           0.022509       0.5177
```

```
# not significant
```

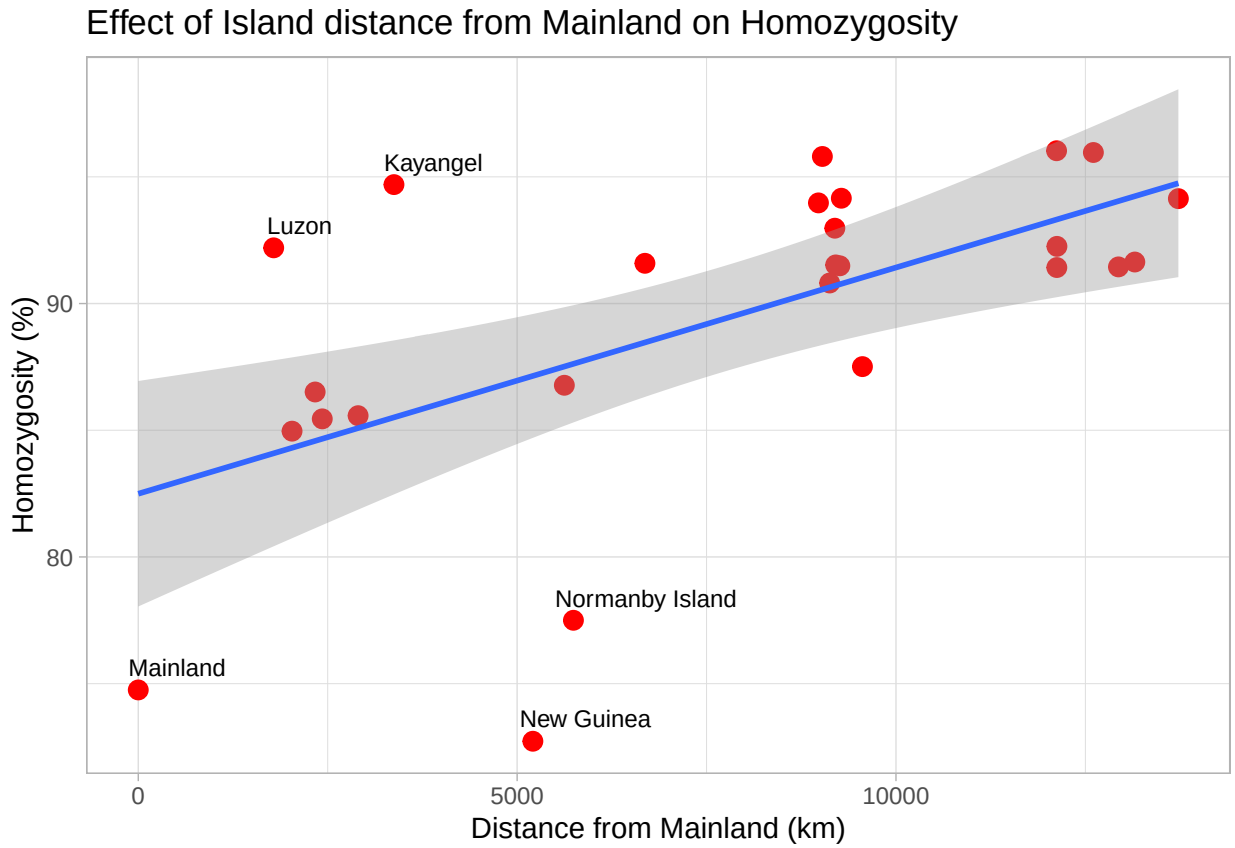
The distribution of residuals is now normal, although there appears to be some level of heteroscedasticity visible in the plots (residuals vs. fitted, scale-location, residualPlot) even though the ncv test does not reject HO of constant variance of the residuals. I am not too concerned with this because the sample size is small and I do not want to over-analyse the diagnostics when my goal is merely to check for a correlation.

- Higher t-value is generally more significant (above 2), related to p-value. High enough here.
- The residual standard error is very low compared to the intercept estimate, which is ideal.
- Multiple R-squared indicates ~36% of the variance is explained by the model
- Significant p-value(s), can reject H0 where x and y are not correlated.

vi. Model Plot

```
ggplot(data = per.island, aes(x = km.from.ML, y = perc.o, label = island)) +
  geom_point(colour = "red", size = 3) +
  geom_smooth(method = 'lm', se = TRUE, level = 0.95) +
  geom_text(
    data = subset(per.island, km.from.ML < 6500 &
                    (perc.o < 80 | perc.o > 90)),
    hjust = 0.1,
    vjust = -0.8,
    size = 3
  ) +
  ggtitle("Effect of Island distance from Mainland on Homozygosity") +
  xlab("Distance from Mainland (km)") + ylab("Homozygosity (%)") +
  theme_light()
```

```
## `geom_smooth()` using formula 'y ~ x'
```

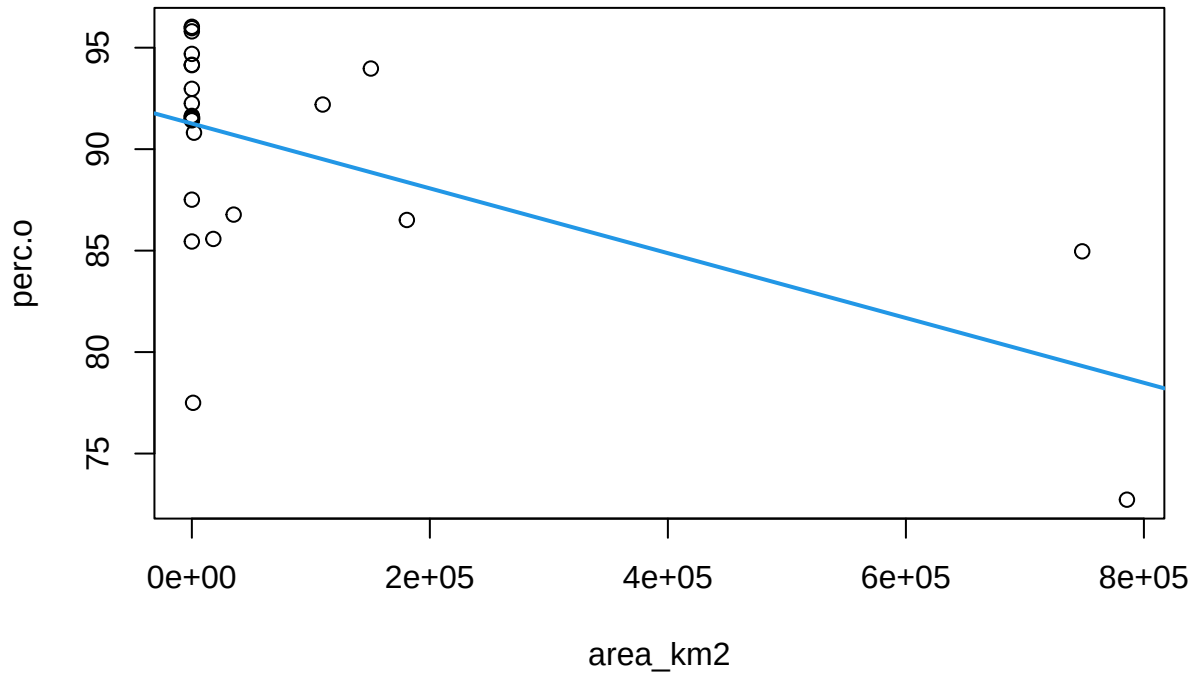


9b. Area

i. Model Construction and Diagnostics

```
testLM2 <- lm(perc.o ~ area_km2, data = per.island) # model
```

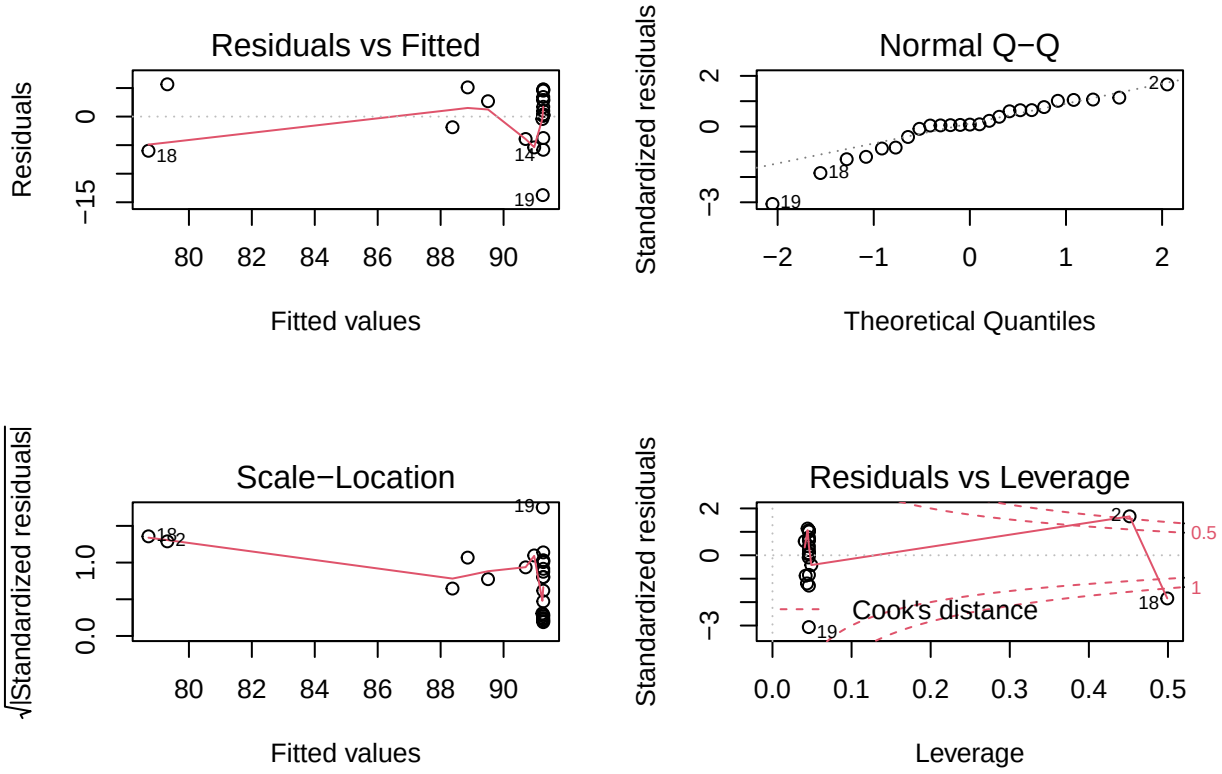
```r
plot(perc.o ~ area_km2, data = per.island) # points clustered in small area_km2
abline(coef = coef(testLM2), col = 4, lwd = 2)
```



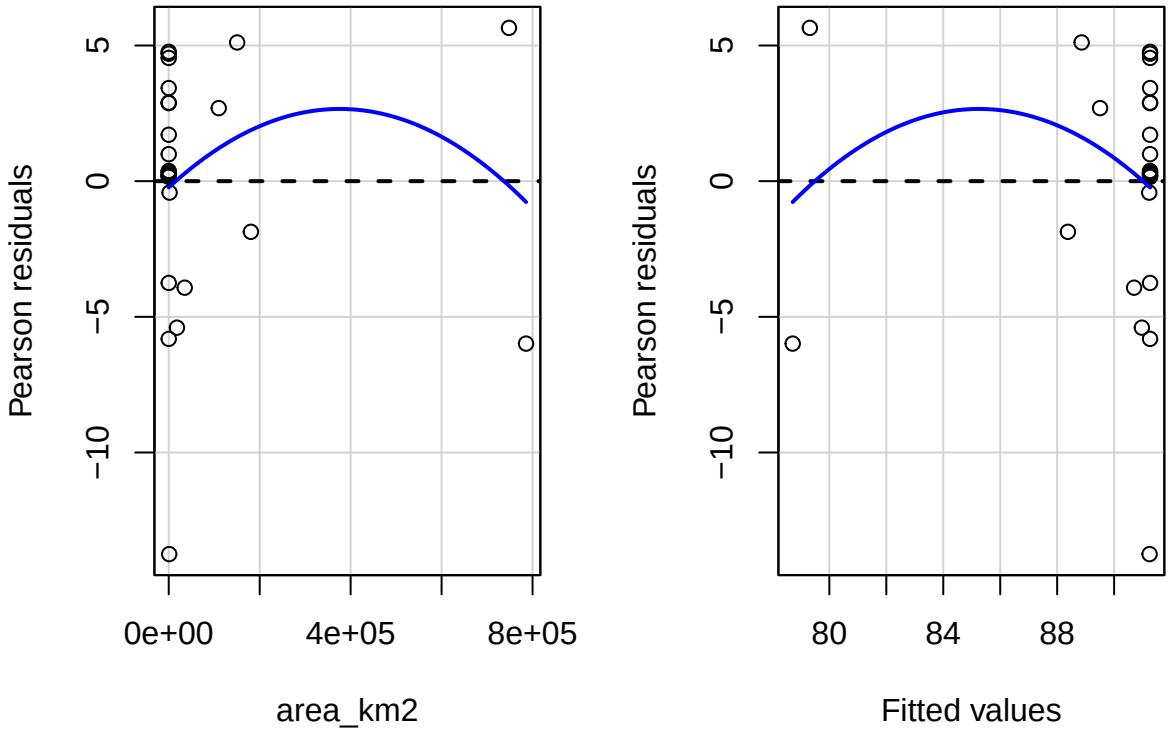

```r
summary(testLM2) # model results
```

```
##
## Call:
## lm(formula = perc.o ~ area_km2, data = per.island)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.7456  -1.8668   0.3259   2.8939   5.6496
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.126e+01  9.854e-01  92.612  < 2e-16 ***
## area_km2    -1.597e-05  4.414e-06  -3.618  0.00144 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.589 on 23 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3627, Adjusted R-squared:  0.335
## F-statistic: 13.09 on 1 and 23 DF,  p-value: 0.001445
```

```
par(mfrow = c(2, 2)) # changes the number of plots visible at once
plot(testLM2) # diagnostic plots
```

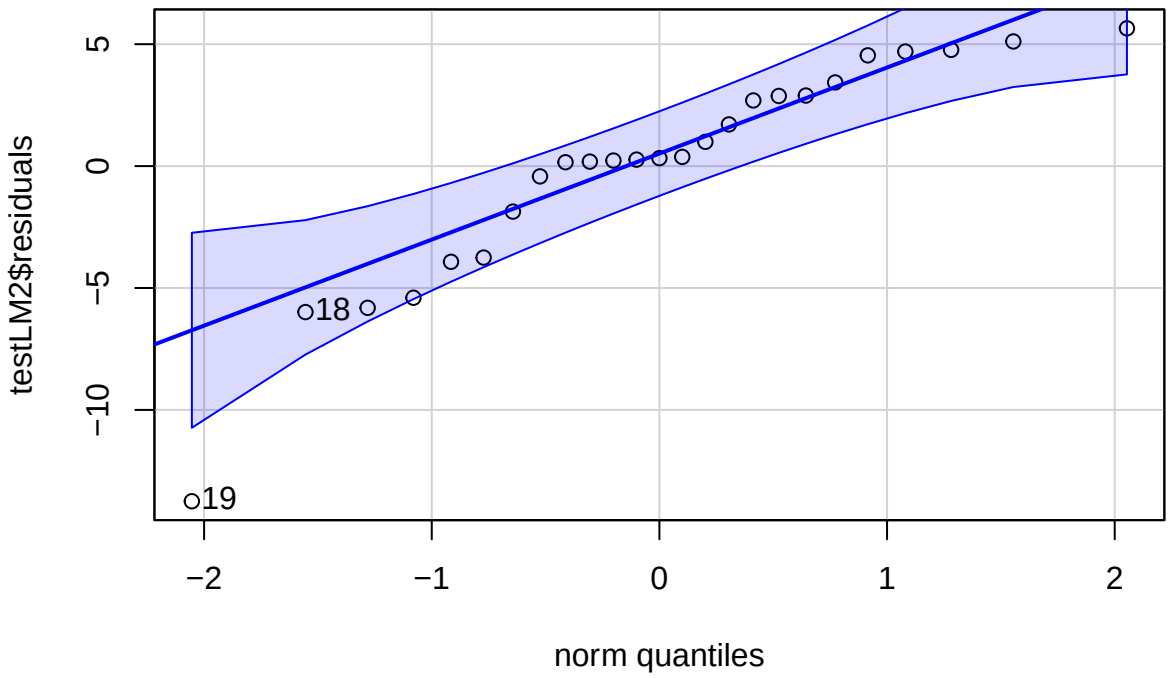



```
par(mfrow = c(1, 1))
residualPlots(testLM2) # Spread of points doesn't look good, there's lots of bunching
```




```
##            Test stat Pr(>|Test stat|)
## area_km2      -0.655           0.5192
## Tukey test    -0.655           0.5124
```

```
qqPlot(testLM2$residuals, line = "quartiles") # appears normal dist, outlier point 19
```




```
## 19 18
## 18 17
```

```
shapiro.test(testLM2$residuals) # indicates non-normality of residuals if alpha is 0.05,
```

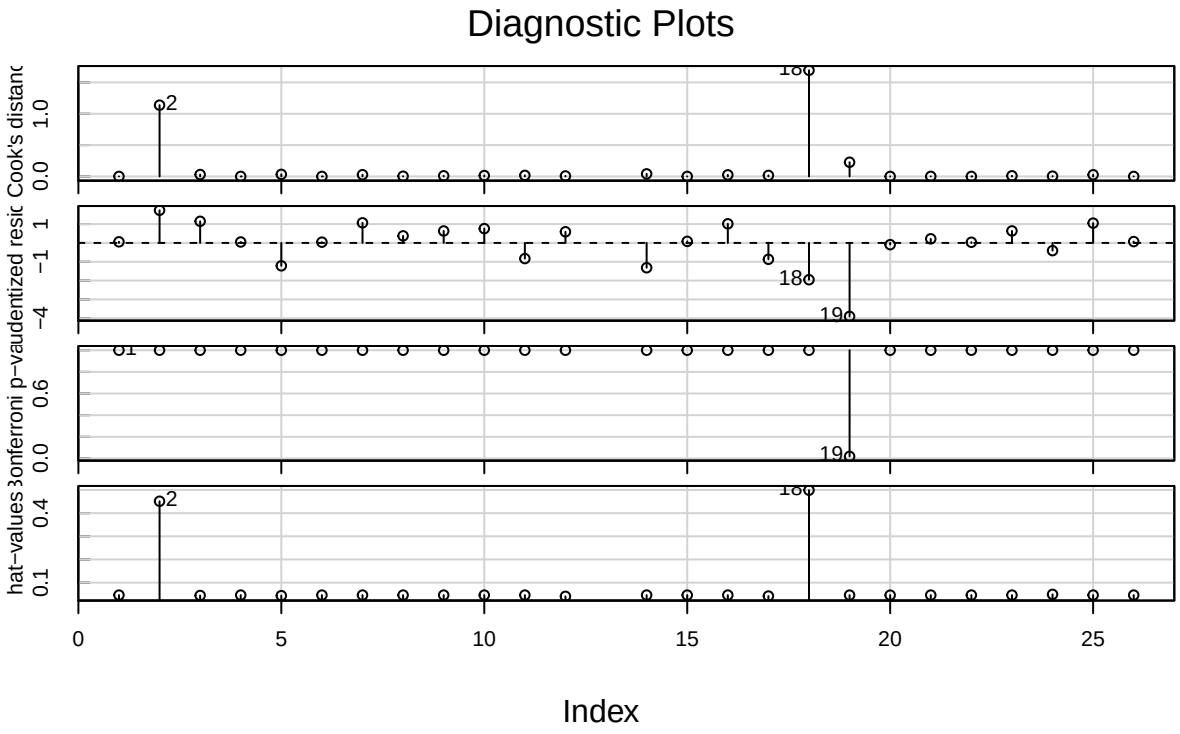
```
##
##  Shapiro-Wilk normality test
##
## data:  testLM2$residuals
## W = 0.89572, p-value = 0.01482
```

```
# I think this significant p-value is caused by point 19
hist(testLM2$residuals, breaks = 10, xlim = c(-15,10))
```


```
ncvTest(testLM2) # homoscedasticity test: H0 of constant variance is not rejected,
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.4593862, Df = 1, p = 0.49791
```

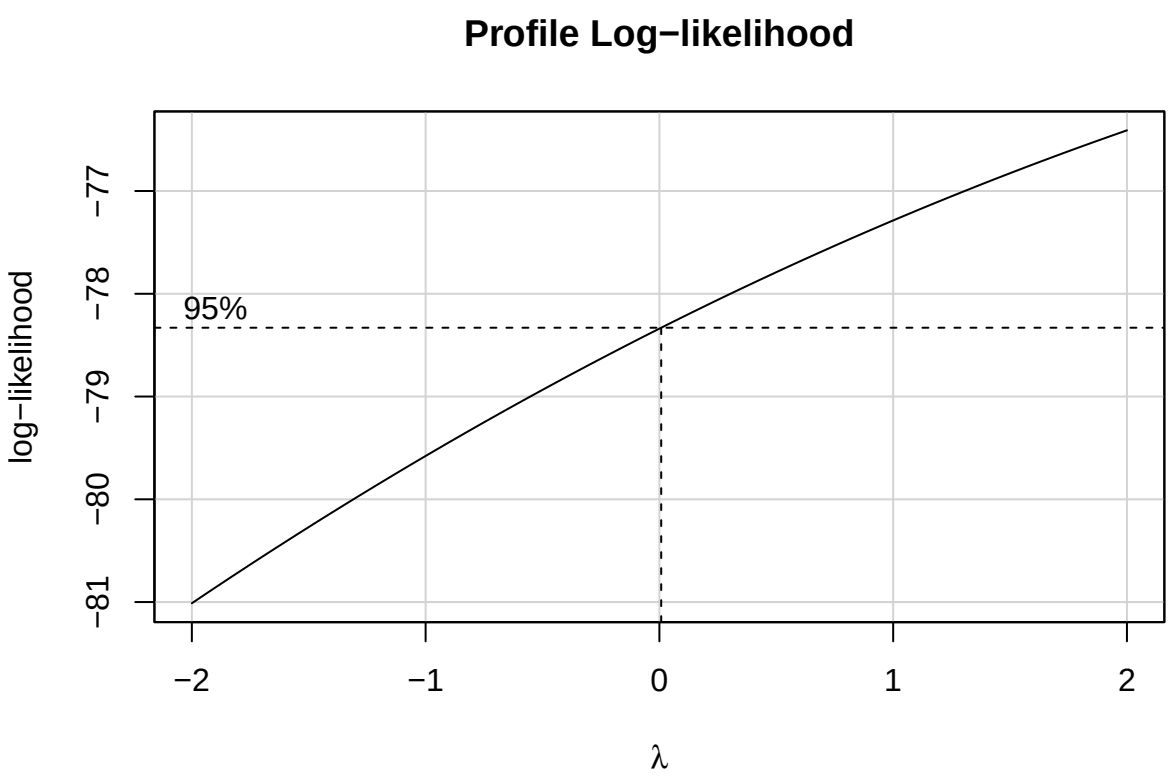
```
# but scale location plot above (plot(testLM2)) still looks peculiar.

influenceIndexPlot(testLM2) # outliers
```




```
# Cooks distances: points 2 and 18 larger than 0.5,
# Studentised residuals: point 19 less than -3
# Bonferroni p-value: point 19 smaller then 0.05,
# Hat-values: points 2 and 18 different, perhaps influential but not higher than 1
outlierTest(testLM2) # 19: rstudent = -3.900653, unadjusted p-value = 0.00076823, Bonferroni p = 0.0192
```

```
##     rstudent unadjusted p-value Bonferroni p
## 19 -3.900653         0.00076823     0.019206
```

```
boxCox(testLM2) # suggests log transformation
```



Notes on diagnostic plots

- Residuals vs. Fitted plot shows a line not horizontal, therefore relationship may not be linear here. Points 18 and 19 are more extreme.
- Normal QQ plot shows residuals generally in line, indicating normality although there is a fat lower tail. The Shapiro-Wilk test indicates non-normality of residuals, although not if the alpha is set to 0.01. Questionable point is once again 18 (New Guinea).
- Scale-Location plot is used to indicate constant residual variance with a line that does not trend up or down overall. Line trends downwards which means non-constant variance, although strangely the ncvTest indicates H0 of constant variance should not be rejected.
- Residuals vs Leverage (Cook's distance) plot assesses outliers. Points 2 (Borneo) and 18 (New Guinea) is over the 0.5 Cook's line and high leverage and therefore is noteworthy. This could be explained by them being the largest island areas in the dataset/model. The outlier tests also bring attention to point 19.
- The Box-Cox plot indicates a transformation to perform and here lambda lines up almost perfectly with 0 == log transformation.

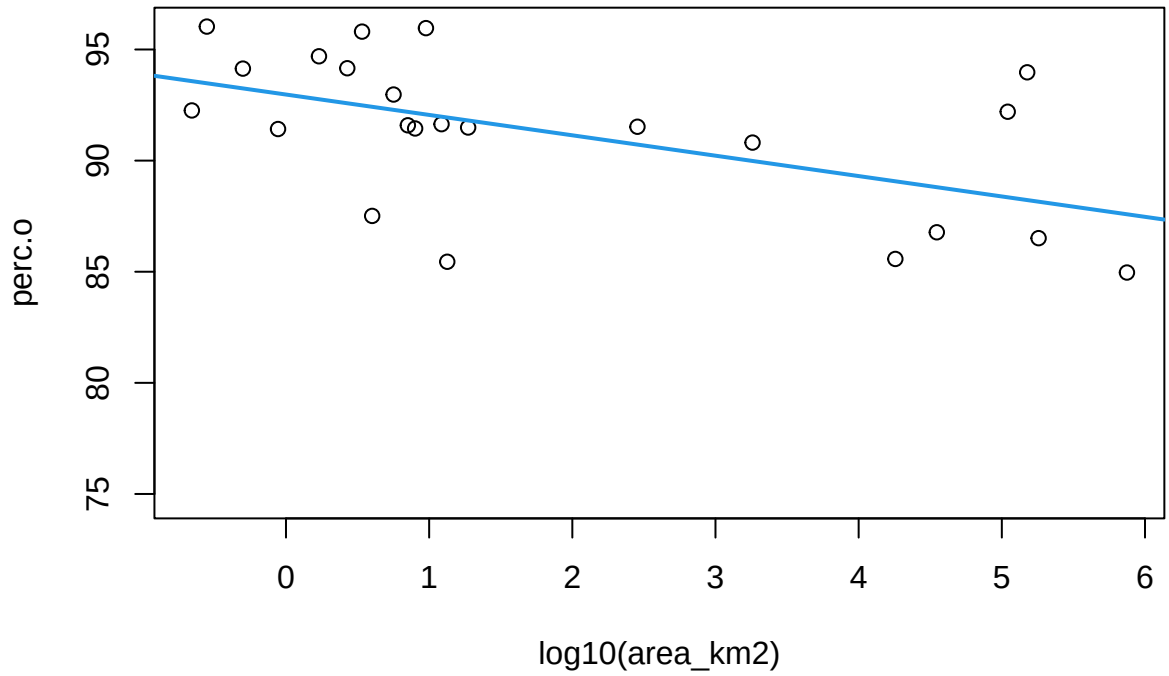
ii. Adjusted Model and Diagnostics

Trying log10 of the area since there's a cluster of small islands. Bigger islands will be "squeezed" more than smaller islands so that ratios/distances between points are preserved, abline is estimated over them (but if "unsqueezed" the abline would become curved). A log10 relationship might make sense because the effect of islands size decreases as the islands get bigger.

I'm hesitant to remove any outliers here, but I'm seeing how it goes removing New Guinea and Normanby Island again. Removing them doesn't mean I will not consider them again, however due to having so few

data points I think they strongly influence the results and it may be easier to see the general picture without them. Additionally, Normanby Island is missing 57% and New Guinea 56% of the SNPs, making them less reliable.

```
z <- per.island[-c(18, 19),] # removing New Guinea and Normanby Island again

LM2 <- lm(perc.o ~ log10(area_km2), data = z)

plot(perc.o ~ log10(area_km2), data = z)
abline(coef = coef(LM2), col = 4, lwd = 2)
```

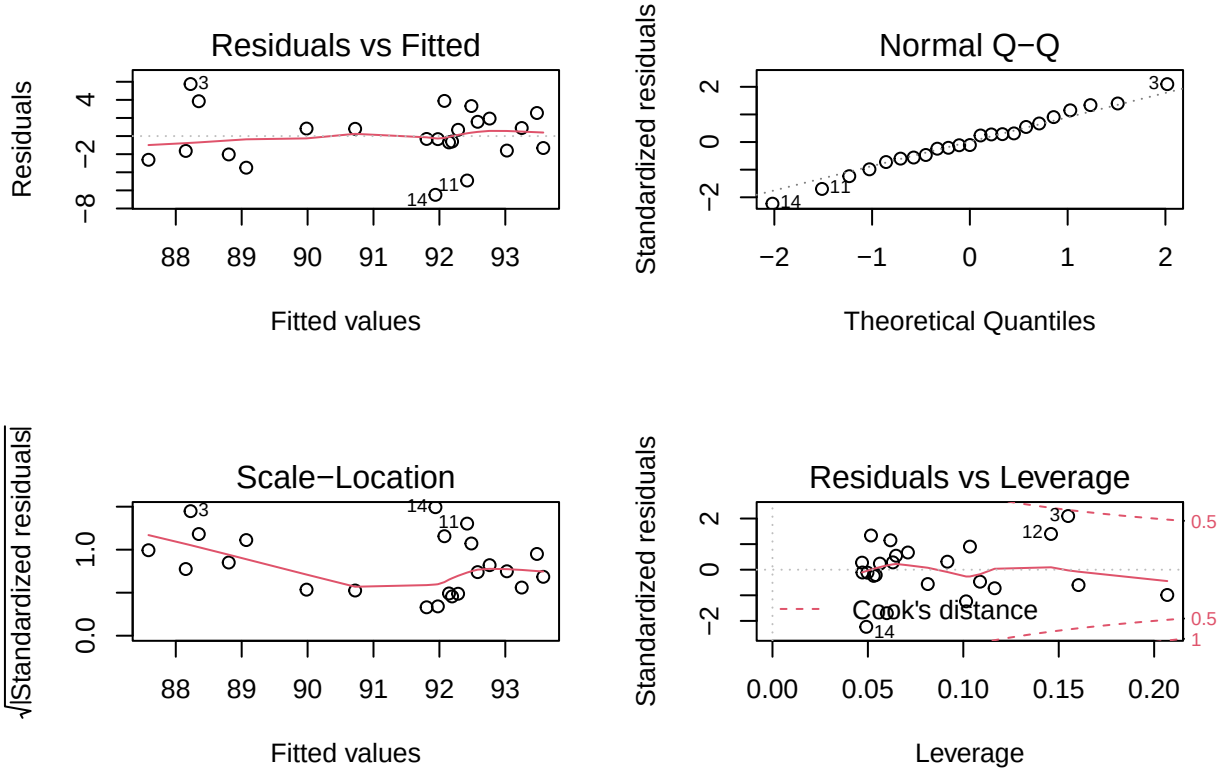



```
summary(LM2)
```

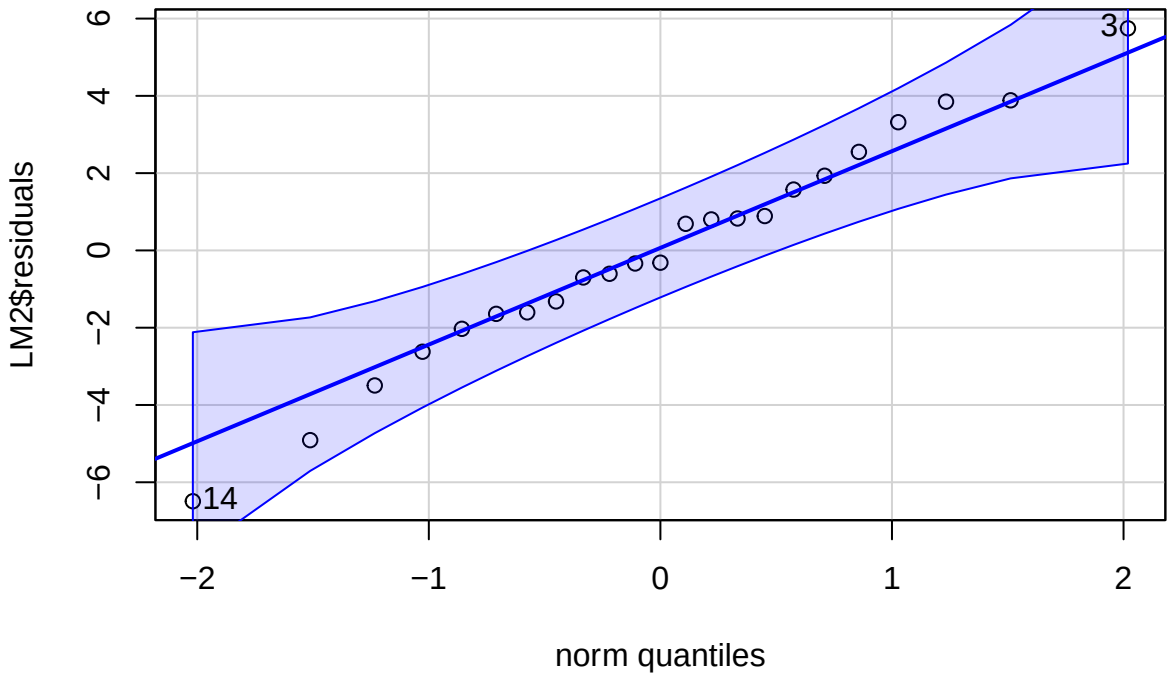
```
##
## Call:
## lm(formula = perc.o ~ log10(area_km2), data = z)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.491 -1.622 -0.317  1.754  5.747
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      92.9729     0.8395 110.750  < 2e-16 ***
## log10(area_km2)  -0.9171     0.3012  -3.045  0.00616 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.982 on 21 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3062, Adjusted R-squared:  0.2732
## F-statistic:  9.27 on 1 and 21 DF,  p-value: 0.006159
```

```
par(mfrow = c(2, 2)) # changing the number of plots visible at once
plot(LM2) # diagnostic plots; relationship now linear and outliers are now
```




```
# less influential

par(mfrow = c(1, 1))
qqPlot(LM2$residuals, line = "quartiles") # normal
```

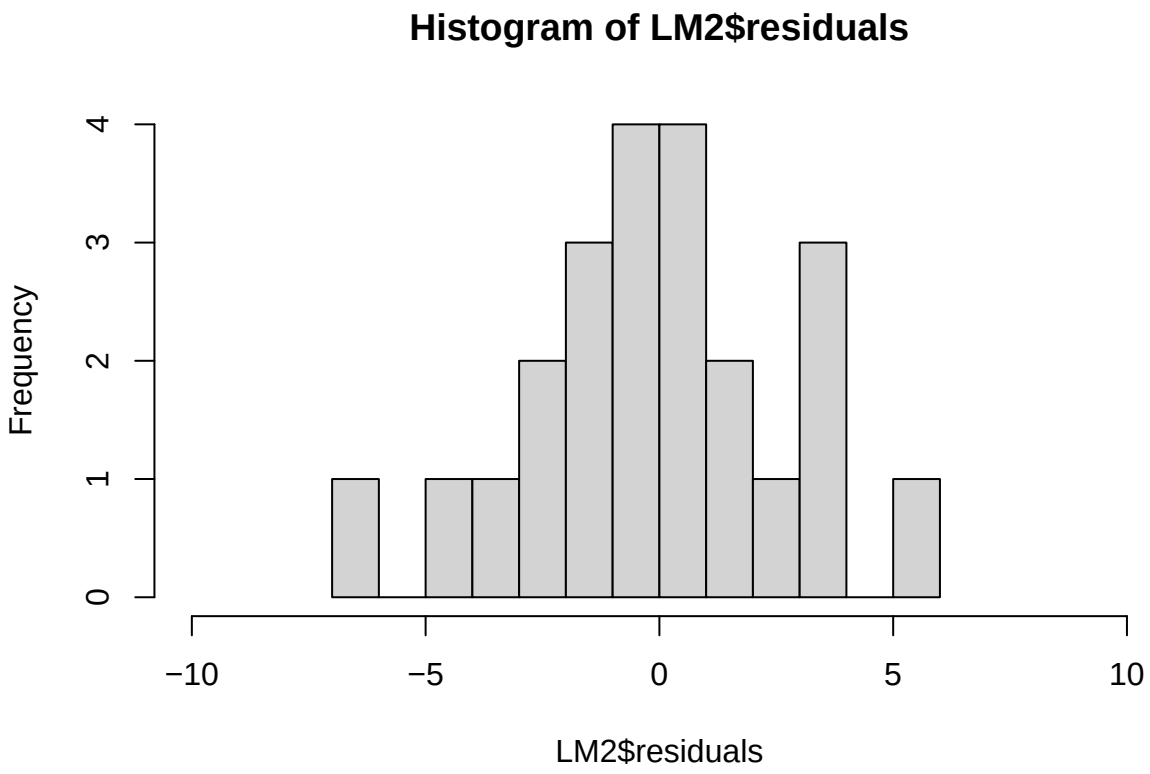



```
## 14  3
## 13  3
```

```
shapiro.test(LM2$residuals) # indicates normality of residuals
```

```
##
##  Shapiro-Wilk normality test
##
## data:  LM2$residuals
## W = 0.98906, p-value = 0.9945
```

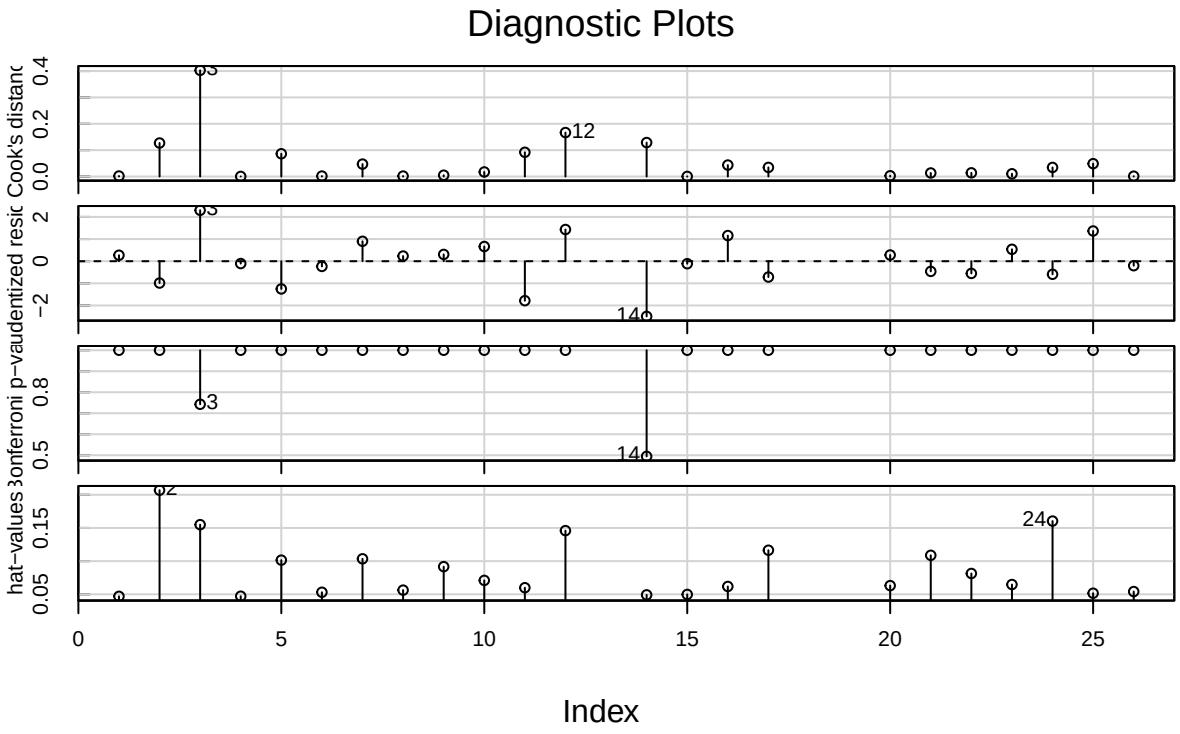
```
hist(LM2$residuals, breaks = 10, xlim = c(-10,10))
```




```
ncvTest(LM2) # homoscedasticity test: H0 of constant variance is not rejected
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.8664813, Df = 1, p = 0.35193
```

```
influenceIndexPlot(LM2) # outliers
```




```
# Cooks distances: none larger than 0.5,
# Studentised residuals: none less than -3 or more than 3
# Bonferroni p-value: none smaller then 0.05,
# Hat-values: points 2 influential, higher than 1
outlierTest(LM2)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 14 -2.493984           0.021508      0.49468
```

- Spread of residuals seems equal enough on both sides
- t-values are above or below 2, indicating the coefficients are more reliable
- Residual standard error much smaller than the intercept estimate, which is good
- R^2 indicates ~30% of the homozygosity points can be explained by the explanatory variable.
- F-statistic is higher than 1 and significant, meaning there is a relationship present.

vi. Model Plot

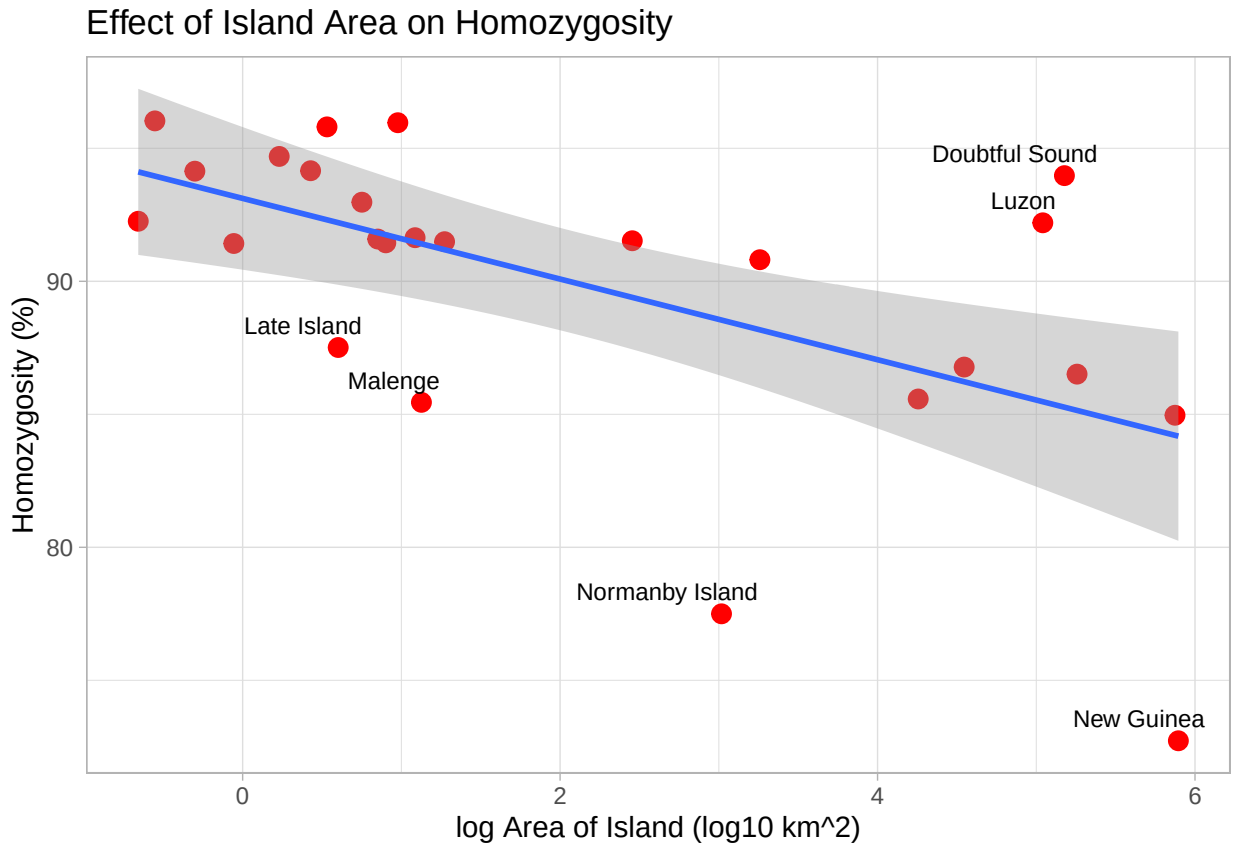
```
ggplot(data = per.island, aes(
  x = log10(area_km2),
  y = perc.o,
  label = island
```

```
)) +
  geom_point(colour = "red", size = 3) +
  geom_smooth(method = 'lm', se = TRUE, level = 0.95) +
  geom_text(
    data = subset(per.island,
                  (log10(area_km2) < 2 & perc.o < 90) |
                    (log10(area_km2) > 2.9 & (perc.o > 92 | perc.o < 80))
                  ),
    hjust = 0.8,
    vjust = -0.8,
    size = 3
  ) +
  ggtitle("Effect of Island Area on Homozygosity") +
  xlab("log Area of Island (log10 km^2)") + ylab("Homozygosity (%)") +
  theme_light()
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



9c. Multiple Regression: both Distance and Area

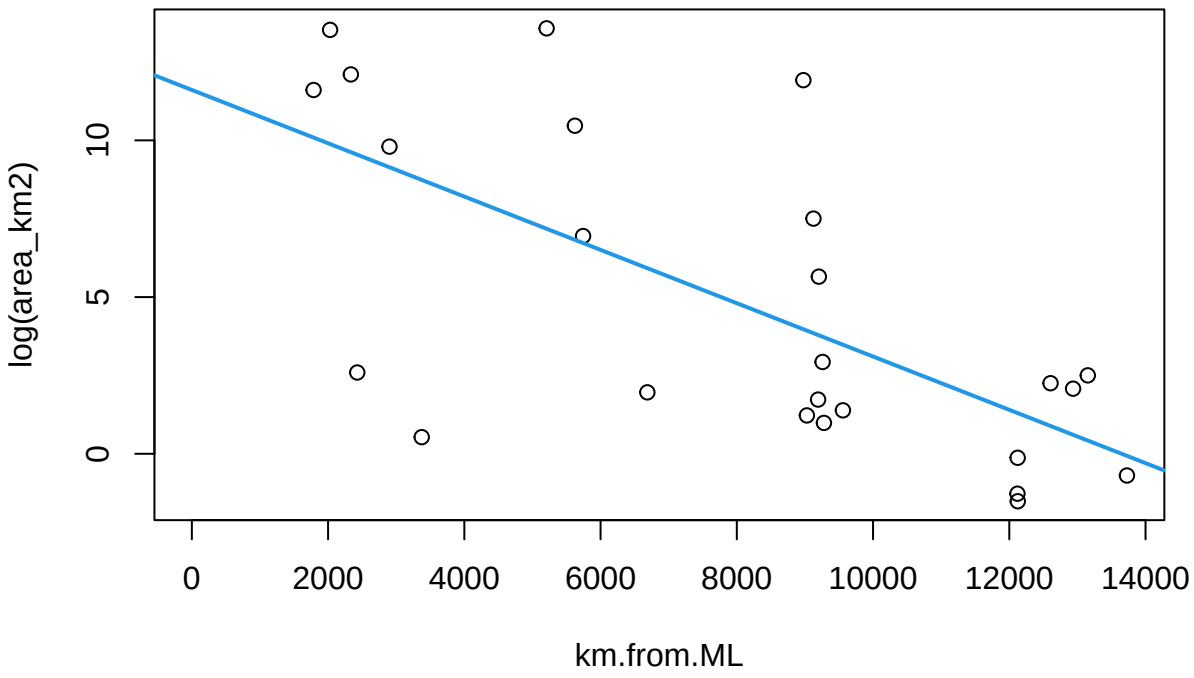
i. Model Construction and Diagnostics

Taking a moment to look at if there's a relationship between area and distance from the mainland:

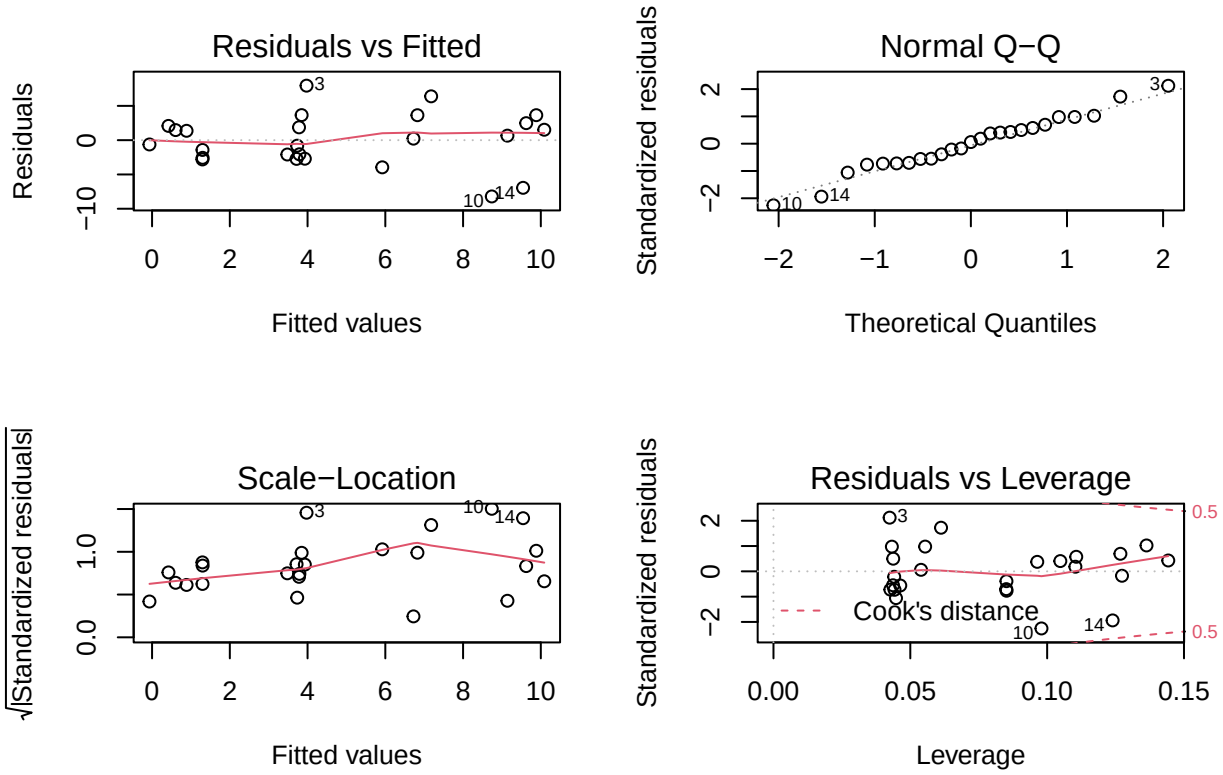
```
par(mfrow = c(1, 1))
adLM <- lm(log(area_km2) ~ km.from.ML, data = per.island)
summary(adLM)
```

```
##
## Call:
## lm(formula = log(area_km2) ~ km.from.ML, data = per.island)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2082 -2.5735  0.2227  2.0792  7.9477
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.6091301  1.7653224   6.576 1.04e-06 ***
## km.from.ML  -0.0008506  0.0001983  -4.289 0.000274 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.828 on 23 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4444, Adjusted R-squared:  0.4203
## F-statistic:  18.4 on 1 and 23 DF,  p-value: 0.0002738
```

```
plot(log(area_km2) ~ km.from.ML, data = per.island)
abline(coef = coef(adLM), col = 4, lwd = 2)
```




```
par(mfrow = c(2, 2))
plot(adLM)
```




```
shapiro.test(adLM$residuals) # p-value = 0.8809, normality
```

```
##
##  Shapiro-Wilk normality test
##
## data:  adLM$residuals
## W = 0.97979, p-value = 0.8809
```

```
cor.test(per.island$km.from.ML, log(per.island$area_km2), method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  per.island$km.from.ML and log(per.island$area_km2)
## t = -4.2892, df = 23, p-value = 0.0002738
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8404009 -0.3686031
## sample estimates:
##       cor
## -0.6666402
```

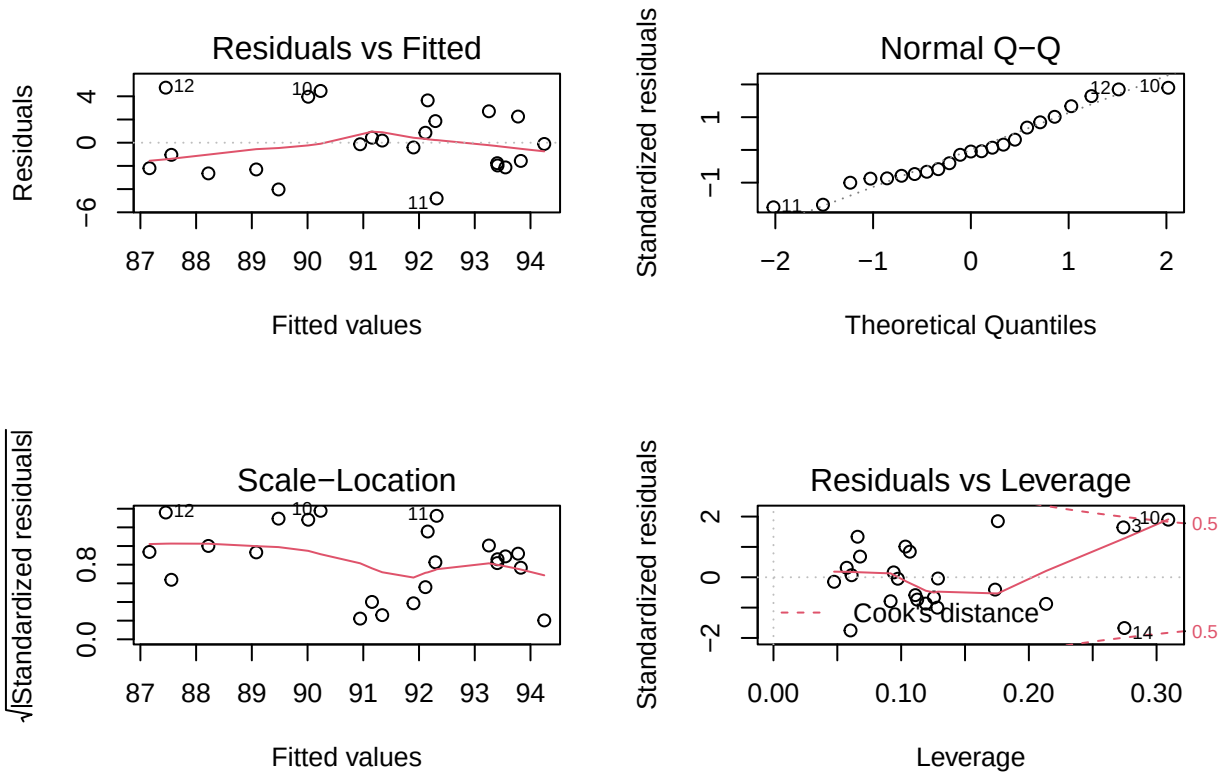
```
#correlated
```

The above plot shows that generally the area of the islands decreases with distance (e.g. Borneo and New Guinea are large, Reiono and Tahanea are small) which I believe is a combination of sampling issues and how the islands are naturally on the Pacific. The correlation between these two variables is an issue with only a basic linear models because I cannot test the affect of each variable compared with each other because their effect is similar (collinearity in this case). I will go through diagnostics to check if the effect is great enough to warrant a different method e.g. GLM.

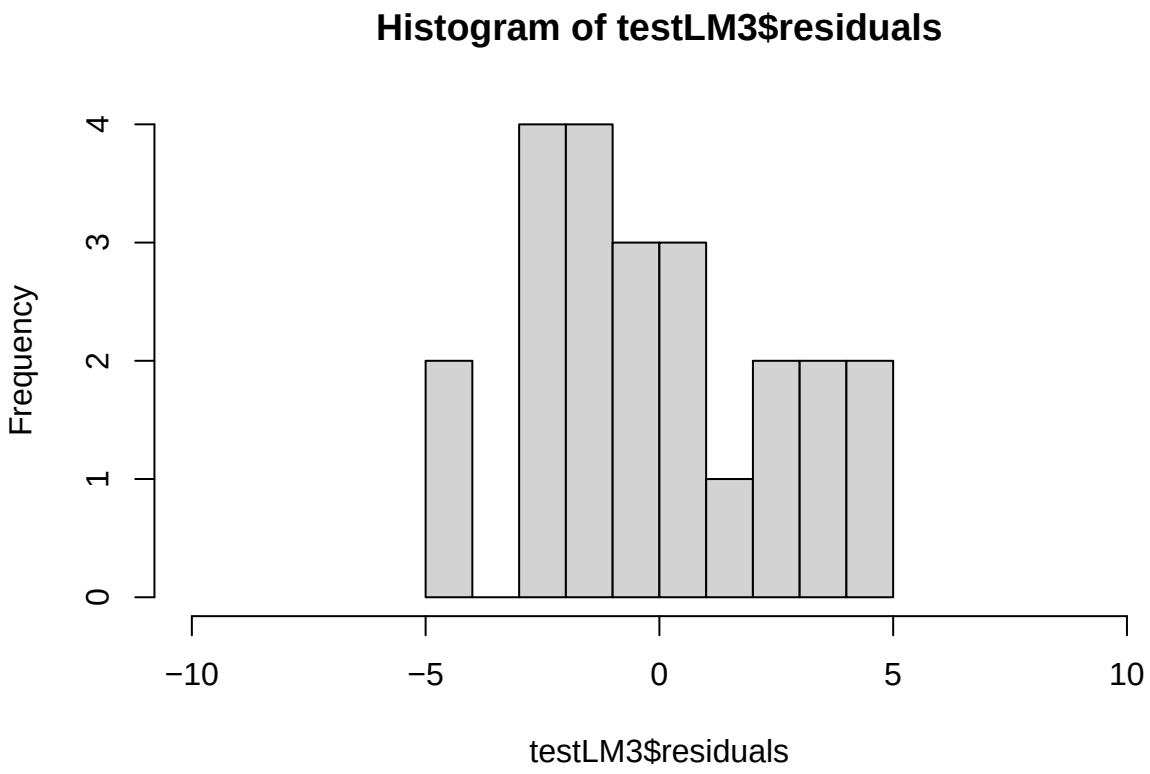
```
z <- per.island[-c(18, 19),] # starting by removing New Guinea and Normanby Island
# again (after brief check of diagnostic plots)

testLM3 <- lm(perc.o ~ km.from.ML + log(area_km2), data = z) # not sure if I
# should keep log transformation
summary(testLM3)
```

```
##
## Call:
## lm(formula = perc.o ~ km.from.ML + log(area_km2), data = z)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8047 -2.0464 -0.1323  2.0577  4.7434
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   89.1097394  2.2543620  39.528   <2e-16 ***
## km.from.ML     0.0003642  0.0001989   1.832   0.0819 .
## log(area_km2) -0.1985810  0.1651672  -1.202   0.2433
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.828 on 20 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4059, Adjusted R-squared:  0.3465
## F-statistic: 6.832 on 2 and 20 DF,  p-value: 0.005478
```

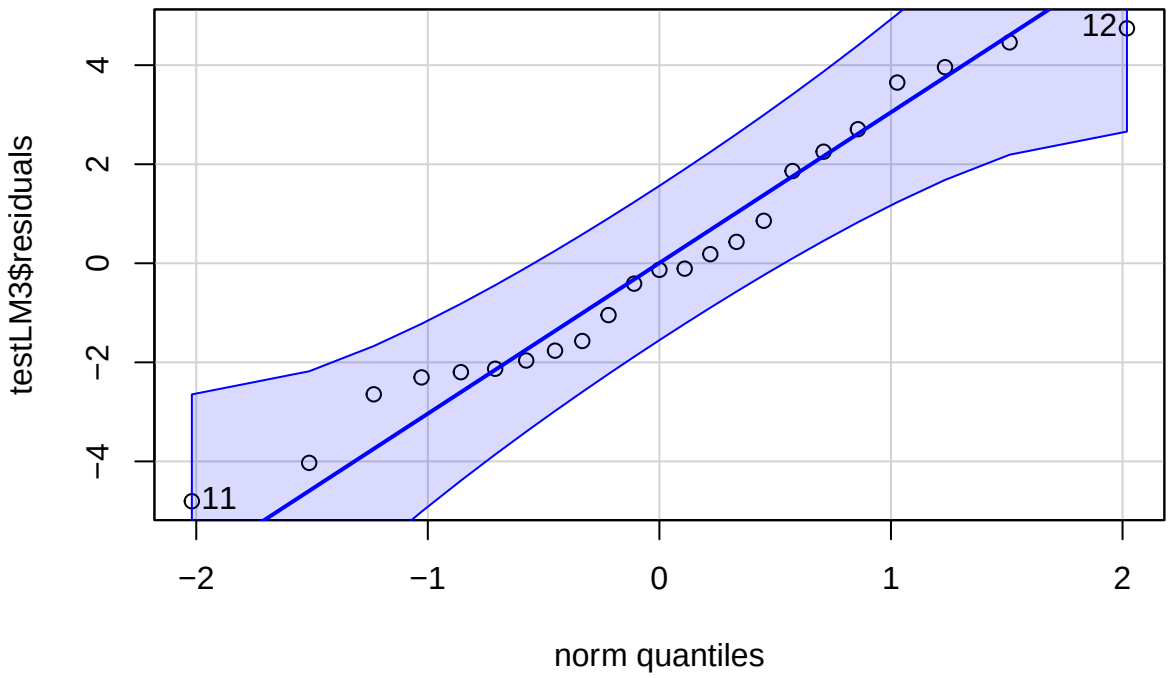
```
par(mfrow = c(2, 2))
plot(testLM3)
```




```
par(mfrow = c(1,1))
hist(testLM3$residuals, breaks = 10, xlim = c(-10,10))
```




```
qqPlot(testLM3$residuals, line = "quartiles") # normal
```




```
## [1] 11 12
```

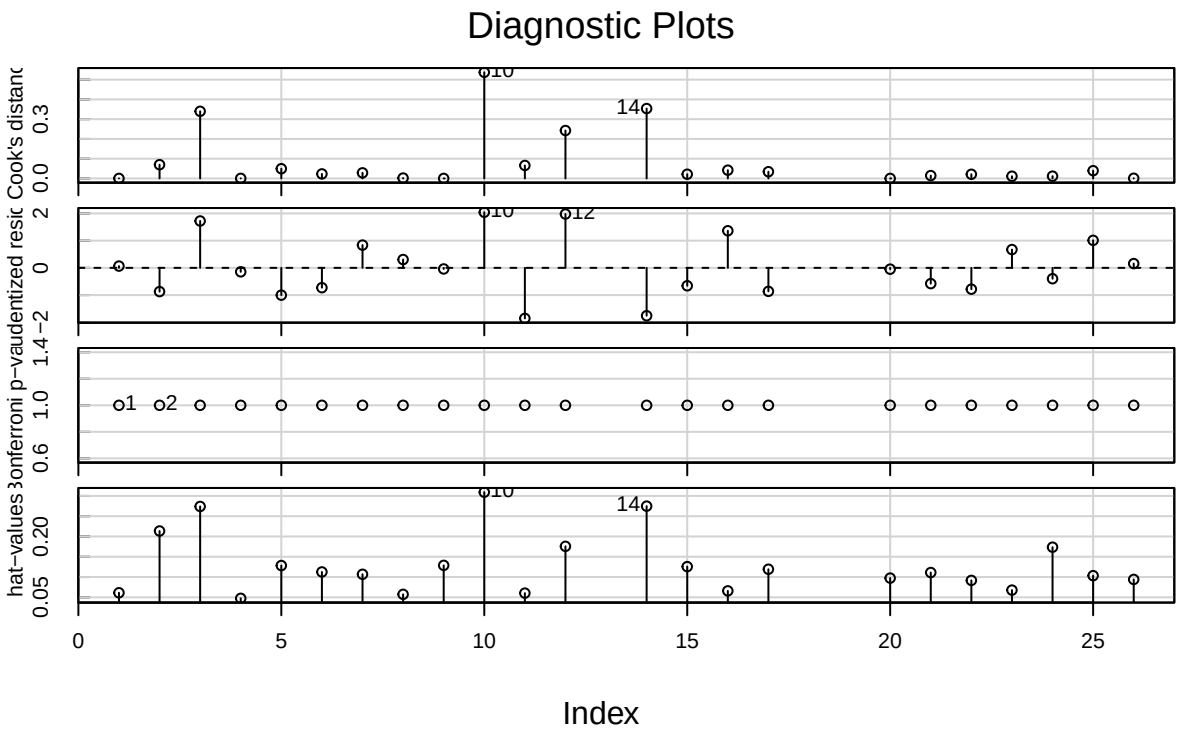
```
shapiro.test(testLM3$residuals) # indicates normality of residuals
```

```
##
##  Shapiro-Wilk normality test
##
## data:  testLM3$residuals
## W = 0.96015, p-value = 0.4665
```

```
ncvTest(testLM3) # homoscedasticity test: H0 of constant variance is not rejected
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.163531, Df = 1, p = 0.28073
```

```
influenceIndexPlot(testLM3) # outliers
```




```
# Cooks distances: no. 2 larger than 0.5,
# Studentised residuals: none less than -3 or more than 3
# Bonferroni p-value: none smaller then 0.05,
# Hat-values: none influential, higher than 1
outlierTest(testLM3)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 10 2.042642           0.055213           NA
```

```
mctest::mctest(testLM3, type = "b") # tests for collinearity, most tests did not detect
```

```
##
## Call:
## omcdiag(mod = mod, Inter = Inter, detr = detr, red = red, conf = conf,
##     theil = theil, cn = cn)
##
##
## Overall Multicollinearity Diagnostics
##
##                        MC Results detection
## Determinant |X'X|:         0.5641         0
## Farrar Chi-Square:        11.7359         1
## Red Indicator:             0.6602         1
```

```
## Sum of Lambda Inverse:     3.5453         0
## Theil's Method:            0.4659         0
## Condition Number:          7.6872         0
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
##
## ==================================
##
## Call:
## imcdiag(mod = mod, method = method, corr = FALSE, vif = vif,
##     tol = tol, conf = conf, cvif = cvif, ind1 = ind1, ind2 = ind2,
##     leamer = leamer, all = all)
##
##
## All Individual Multicollinearity Diagnostics Result
##
##                   VIF    TOL     Wi  Fi Leamer   CVIF Klein   IND1 IND2
## km.from.ML    1.7727 0.5641 16.226 Inf 0.7511 3.1836     1 0.0269    1
## log(area_km2) 1.7727 0.5641 16.226 Inf 0.7511 3.1836     1 0.0269    1
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## km.from.ML , log(area_km2) , coefficient(s) are non-significant may be due to multicollinearity
##
## R-square of y on all x: 0.4059
##
## * use method argument to check which regressors may be the reason of collinearity
## ==================================
```

```
# e.g. condition number = 7.6872 (10-30 indicates presence, issue if >30) and
# VIF = 1.727 (issue if >5)
```

- Residuals appear well distributed (approximately 5 on either side of 0, which the median is close to)
- The t-values are low (between -2 and 2) which is not ideal, estimates may be unreliable
- Here the explanatory variables are no longer significant, which may mean that both contribute to the overall model but the exact effect of each variable is unclear. This matches up with my findings that the two explanatory variables are at least somewhat colinear.
- 23 observations - 3 parameters = 20 degrees of freedom
- The R^2 value indicates that 40% of the variance is explained (35% when considering the extra parameter added)
- The f-test is above 1 and significant (the higher the better when only with a small number of data points), therefore the variance that is explained is significant (?)

Even though the colinearity test indicated colinearity isn't high between the explanatory variables, I believe it is still present (e.g. visible when modelling the variables against each other) and having an effect on the results above. Both the p-values are not significant, the t-values are low and the standard error is high in this mixed model. The best way forward is generally to remove the colinear variable, and/or accept they are intertwined. An additional problem could be omitted variable bias, where another variable has an important effect on the model but has not been included. For example, the presence of another species on an island.

vi. Alternative models (GLM)

```
# distance ~ area
c <- glm(perc.o ~ km.from.ML + area_km2, data = z, family = gaussian(link = "identity"))
summary(c) # km.from.ML significant, AIC 118.34
```

```
##
## Call:
## glm(formula = perc.o ~ km.from.ML + area_km2, family = gaussian(link = "identity"),
##     data = z)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.5625  -1.9098  -0.4521   2.3708   5.3633
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.783e+01  1.632e+00  53.813   <2e-16 ***
## km.from.ML   4.436e-04  1.697e-04   2.613   0.0166 *
## area_km2    -4.328e-06  4.290e-06  -1.009   0.3251
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 8.160259)
##
##     Null deviance: 269.23  on 22  degrees of freedom
## Residual deviance: 163.21  on 20  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 118.34
##
## Number of Fisher Scoring iterations: 2
```

```
# distance ~ log area
c2 <- glm(perc.o ~ km.from.ML + log(area_km2), data = z, family = gaussian(link = "identity"))
summary(c2) # intercept still significant, not km or area. AIC 117.88
```

```
##
## Call:
## glm(formula = perc.o ~ km.from.ML + log(area_km2), family = gaussian(link = "identity"),
##     data = z)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.8047  -2.0464  -0.1323   2.0577   4.7434
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   89.1097394  2.2543620  39.528   <2e-16 ***
## km.from.ML     0.0003642  0.0001989   1.832   0.0819 .
## log(area_km2) -0.1985810  0.1651672  -1.202   0.2433
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for gaussian family taken to be 7.997446)
##
##     Null deviance: 269.23  on 22  degrees of freedom
## Residual deviance: 159.95  on 20  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 117.88
##
## Number of Fisher Scoring iterations: 2
```

```
# (lower preferred when comparing models). Results the same as the linear model above.
```

```
with(summary(c2), 1 - deviance/null.deviance) # R^2 value of 0.4058967. This is not
```

```
## [1] 0.4058967
```

```
# high therefore there may be additional variables to add to the model to explain
# the variance.
```

v. Model Plot

```
ggplot(data = per.island, aes(
  x = km.from.ML,
  y = log10(area_km2) + 1,
  label = island
)) +
  geom_point(mapping = aes(colour = perc.o), size = 3) +
  geom_smooth(method = 'lm', se = TRUE, level = 0.95) +
  scale_color_gradient2(
    low = "navy",
    mid = "darkorchid1",
    high = "red",
    midpoint = 85,
    name = "Homozygosity (%)"
  ) +
  geom_text(
    data = subset(
      per.island,
      (log10(area_km2) < 2 & km.from.ML < 7500) |
        (log10(area_km2) > 4 & km.from.ML > 5000)
    ),
    hjust = 0.3,
    vjust = -0.8,
    size = 3
  ) +
  ggtitle("Island Homozygosity change with Distance and Area increase") +
  xlab("Distance from Mainland (km)") + ylab("Area (log10 km²) + 1") +
  theme_light()
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```