

MalNAF Part 2

Rachael Holmes

13/04/2022

Step 1. Load Libraries

```
library(tidyverse)
library(magrittr) # %>% pipes
library(ggpubr)
library(knitr)
library(ggbiplot) # plotting PCA
library(tidymodels)
library(psych) # pairs.panels
```

Step 2. Import data

Data are for 10 island groups in Malesia selected as they are distinct phytogeographic units. Naturalized plant species and native species data is based on datasets created by RH for this analysis which will be made publicly available when published.

```
mal_rich <- read_csv("Malesia_environmental_data.csv")
```

Step 3 PCA of bioclim

Create uncorrelated climate variables from bioclim data.

```
mal_bioclim <- mal_rich %>%
  as.data.frame() %>% # to drop geometry
  select(Island, bio1_mean, bio2_mean, bio3_mean, bio4_mean, bio5_mean, bio6_mean,
         bio7_mean, bio8_mean, bio9_mean, bio10_mean, bio11_mean, bio12_mean,
         bio13_mean, bio14_mean, bio15_mean, bio16_mean, bio17_mean, bio18_mean,
         bio19_mean) # select bioclim var

mal_bioclim.pca <- mal_bioclim %>%
  select(-Island) %>%
  prcomp(center = TRUE, scale. = TRUE)

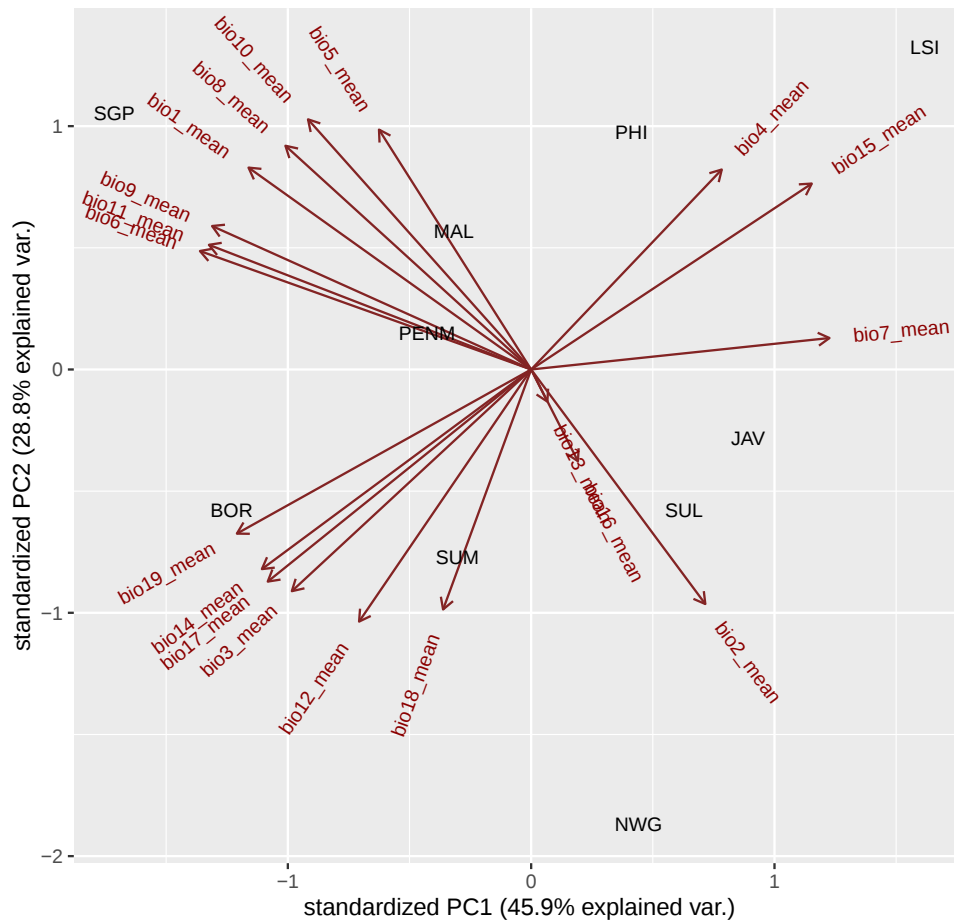
mal_bioclim.pca %>%
  summary()
```

Importance of components:

##	PC1	PC2	PC3	PC4	PC5	PC6	PC7
----	-----	-----	-----	-----	-----	-----	-----

```
## Standard deviation      2.953 2.339 1.8290 0.84703 0.61324 0.5230 0.22276
## Proportion of Variance 0.459 0.288 0.1761 0.03776 0.01979 0.0144 0.00261
## Cumulative Proportion 0.459 0.747 0.9231 0.96087 0.98066 0.9951 0.99767
##                          PC8      PC9      PC10
## Standard deviation      0.17338 0.11931 2.685e-15
## Proportion of Variance 0.00158 0.00075 0.000e+00
## Cumulative Proportion 0.99925 1.00000 1.000e+00
```

```
mal_bioclim.pca %>%
  ggbiplot(labels= mal_bioclim$Island)
```



```
bioclim_vars <- as.data.frame(mal_bioclim.pca$rotation) # extract loadings

bioclim_scores <- as.data.frame(mal_bioclim.pca$x)

bioclim_scores <- bioclim_scores %>%
  add_column(Island = mal_bioclim$Island) # add labels back in

# add scores back into the env var data frame for regression analysis and
# subset only var needed for the analysis

mal_rich_filt <- mal_rich %>%
```

```

left_join(bioclim_scores) %>%
select(Island, Phyto, x_coord, y_coord, area, alt_range,
       perc_urb, perc_agri, pop_dens, GDP_sum, Naturalize,
       PC1, PC2, PC3) %>%
mutate(log10_perc_urb = log10(perc_urb)) %>%
mutate(log10_pop_dens = log10(pop_dens)) %>%
mutate(sqrt_GDP_sum = sqrt(GDP_sum))%>%
mutate(sqrt_Area = sqrt(area))

```

Step 4 Normality

Check out normality of the distributions

```

# First taking a look at dependent variables.

mal_rich_filt %>%
  pull(Naturalize) %>% stats::shapiro.test() # normal

```

```

##
## Shapiro-Wilk normality test
##
## data: .
## W = 0.93401, p-value = 0.4885

```

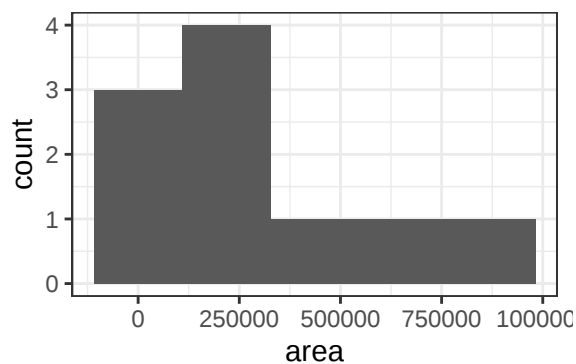
Geographic drivers

Area

```

mal_rich_filt %>%
  ggplot2::ggplot(
    aes(
      x = area
    )
  ) +
  ggplot2::geom_histogram(bins=5) +
  ggplot2::theme_bw()

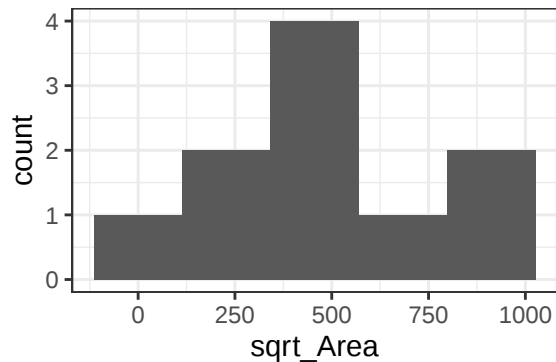
```



```
mal_rich_filt %>%
  pull(area) %>% stats::shapiro.test() # not normal, skewed left with long right tail
```

```
##
## Shapiro-Wilk normality test
##
## data: .
## W = 0.84162, p-value = 0.04614
```

```
mal_rich_filt %>%
  ggplot2::ggplot(
    aes(
      x = sqrt_Area
    )
  ) +
  ggplot2::geom_histogram(bins = 5) +
  ggplot2::theme_bw()
```

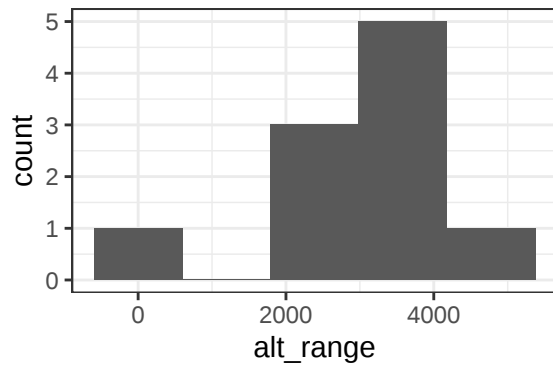


```
mal_rich_filt %>%
  pull(sqrt_Area) %>% stats::shapiro.test() # now normal
```

```
##
## Shapiro-Wilk normality test
##
## data: .
## W = 0.9534, p-value = 0.7088
```

Altitude

```
mal_rich_filt %>%
  ggplot2::ggplot(
    aes(
      x = alt_range
    )
  ) +
  ggplot2::geom_histogram(bins = 5) +
  ggplot2::theme_bw()
```

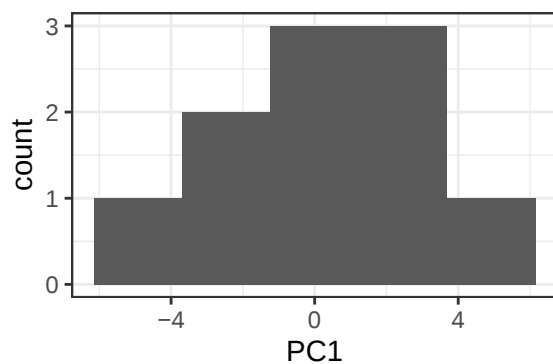


```
mal_rich_filt %>%
  pull(alt_range) %>% stats::shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.89544, p-value = 0.1951
```

Climate

```
mal_rich_filt %>%
  ggplot2::ggplot(
    aes(
      x = PC1
    )
  ) +
  ggplot2::geom_histogram(bins = 5) +
  ggplot2::theme_bw()
```



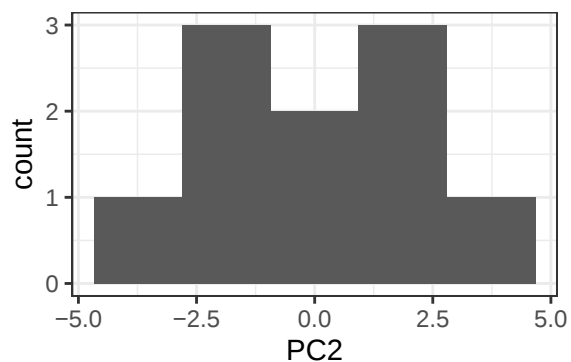
```
mal_rich_filt %>%
  pull(PC1) %>% stats::shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
```

```
## data: .  
## W = 0.97213, p-value = 0.9099
```

```
#pc2
```

```
mal_rich_filt %>%  
  ggplot2::ggplot(  
    aes(  
      x = PC2  
    )  
  ) +  
  ggplot2::geom_histogram(bins = 5) +  
  ggplot2::theme_bw()
```

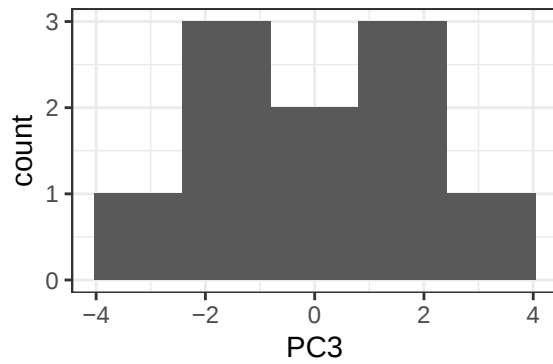


```
mal_rich_filt %>%  
  pull(PC2) %>% stats::shapiro.test()
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: .  
## W = 0.95, p-value = 0.6684
```

```
# pc3
```

```
mal_rich_filt %>%  
  ggplot2::ggplot(  
    aes(  
      x = PC3  
    )  
  ) +  
  ggplot2::geom_histogram(bins = 5) +  
  ggplot2::theme_bw()
```



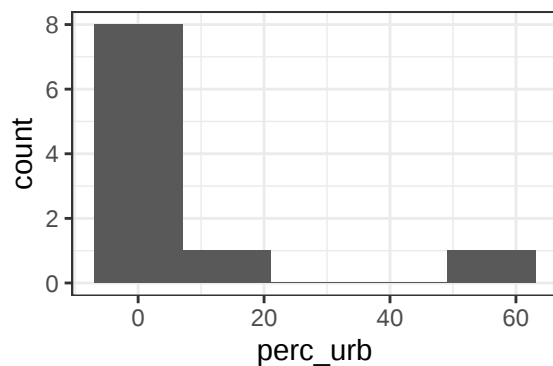
```
mal_rich_filt %>%
  pull(PC3) %>% stats::shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.98217, p-value = 0.9757
```

Anthropogenic

Urban area

```
mal_rich_filt %>%
  ggplot2::ggplot(
    aes(
      x = perc_urb
    )
  ) +
  ggplot2::geom_histogram(bins = 5) +
  ggplot2::theme_bw()
```

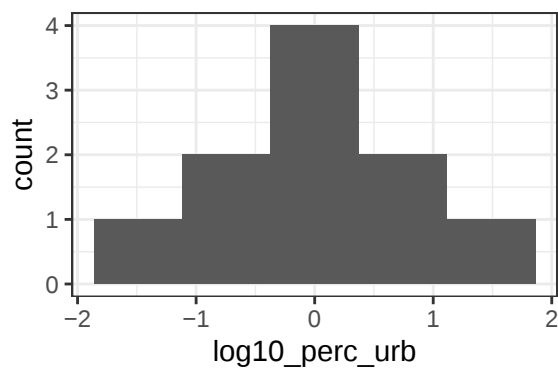


```
mal_rich_filt %>%
  pull(perc_urb) %>% stats::shapiro.test() # not normal
```

```
##
```

```
## Shapiro-Wilk normality test
##
## data: .
## W = 0.46203, p-value = 1.309e-06
```

```
mal_rich_filt %>%
  ggplot2::ggplot(
    aes(
      x = log10_perc_urb
    )
  ) +
  ggplot2::geom_histogram(bins = 5) +
  ggplot2::theme_bw()
```

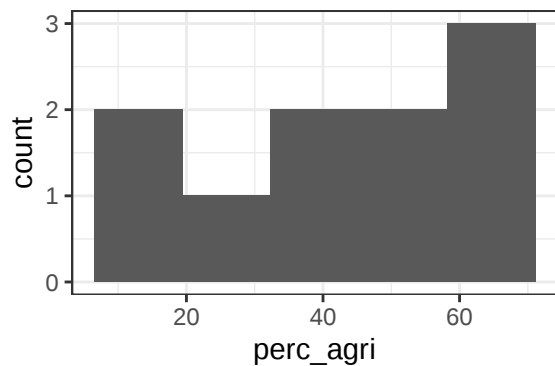


```
mal_rich_filt %>%
  pull(log10_perc_urb) %>% stats::shapiro.test() # now normal
```

```
##
## Shapiro-Wilk normality test
##
## data: .
## W = 0.96372, p-value = 0.8273
```

Agricultural area

```
mal_rich_filt %>%
  ggplot2::ggplot(
    aes(
      x = perc_agri
    )
  ) +
  ggplot2::geom_histogram(bins = 5) +
  ggplot2::theme_bw()
```

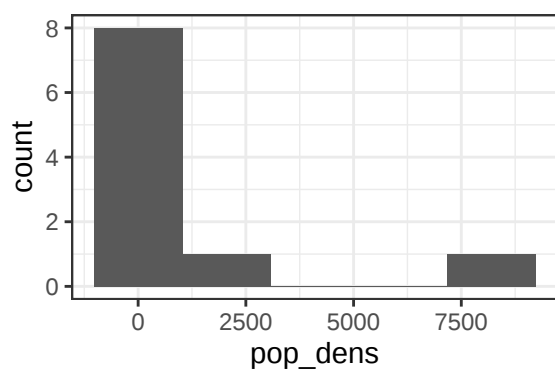


```
mal_rich_filt %>%
  pull(perc_agri) %>% stats::shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.8813, p-value = 0.1351
```

Population density

```
mal_rich_filt %>%
  ggplot2::ggplot(
    aes(
      x = pop_dens
    )
  ) +
  ggplot2::geom_histogram(bins = 5) +
  ggplot2::theme_bw()
```

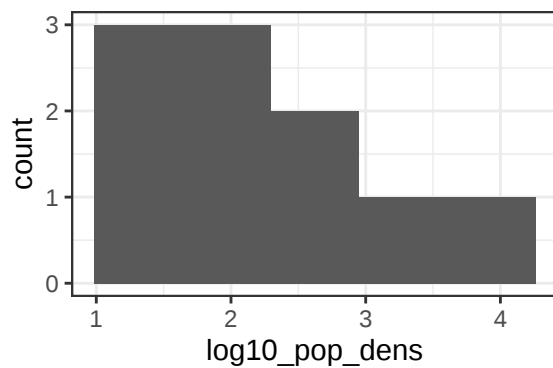


```
mal_rich_filt %>%
  pull(pop_dens) %>% stats::shapiro.test()  # not normal
```

```
##
##  Shapiro-Wilk normality test
##
```

```
## data: .  
## W = 0.45704, p-value = 1.145e-06
```

```
mal_rich_filt %>%  
  ggplot2::ggplot(  
    aes(  
      x = log10_pop_dens  
    )  
  ) +  
  ggplot2::geom_histogram(bins = 5) +  
  ggplot2::theme_bw()
```

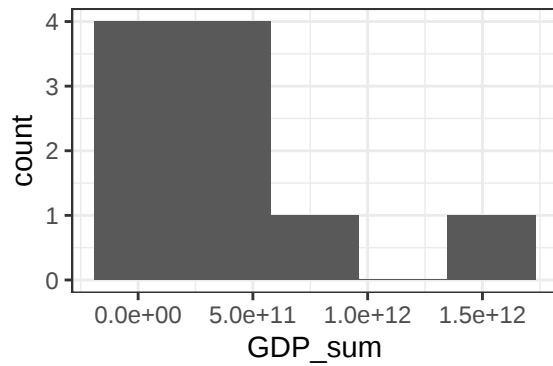


```
mal_rich_filt %>%  
  pull(log10_pop_dens) %>% stats::shapiro.test() # normal
```

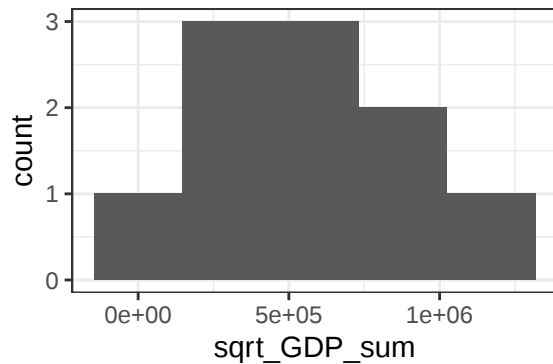
```
##  
## Shapiro-Wilk normality test  
##  
## data: .  
## W = 0.93099, p-value = 0.4577
```

GDP

```
mal_rich_filt %>%  
  ggplot2::ggplot(  
    aes(  
      x = GDP_sum  
    )  
  ) +  
  ggplot2::geom_histogram(bins = 5) +  
  ggplot2::theme_bw()
```



```
mal_rich_filt %>%
  ggplot2::ggplot(
    aes(
      x = sqrt_GDP_sum
    )
  ) +
  ggplot2::geom_histogram(bins = 5) +
  ggplot2::theme_bw()
```



```
mal_rich_filt %>%
  pull(GDP_sum) %>% stats::shapiro.test() # not normal
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.80756, p-value = 0.01791
```

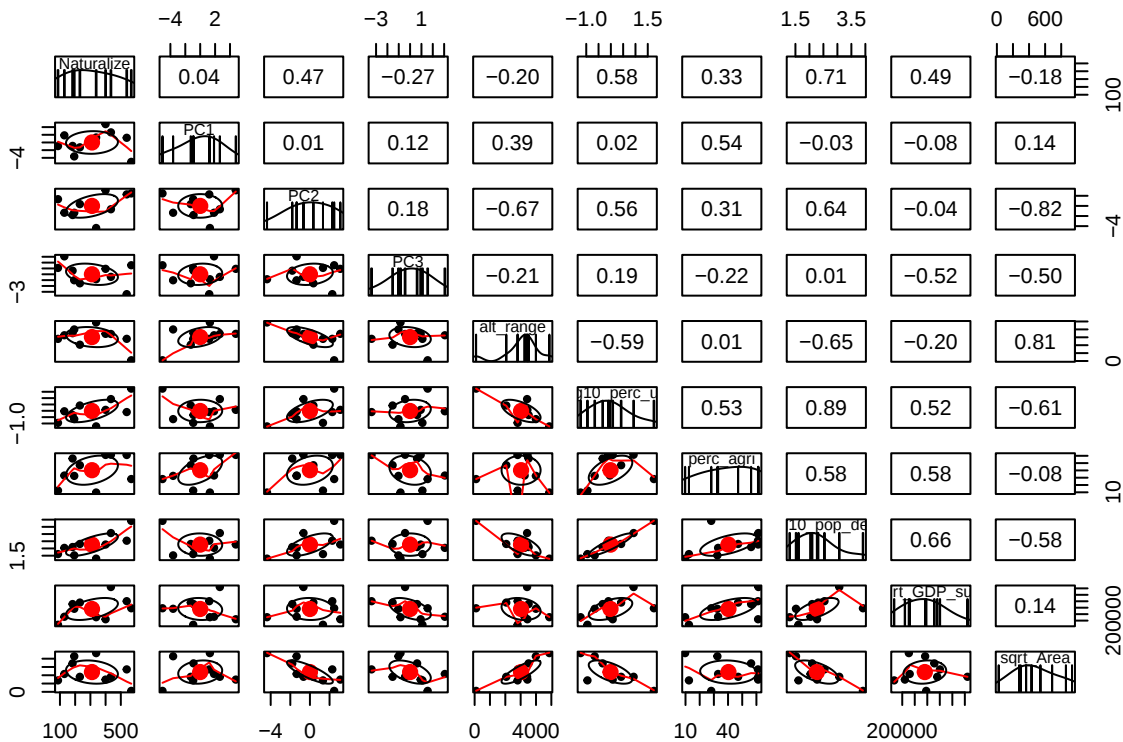
```
mal_rich_filt %>%
  pull(sqrt_GDP_sum) %>% stats::shapiro.test() # now normal (tried log10 but sqrt better)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.95651, p-value = 0.7454
```

Step 5 Correlation tests

Use Spearmans because small dataset even though all variables are normally distributed after transformation

```
mal_rich_filt %>%
  select(
    Naturalize,
    PC1, PC2, PC3, alt_range, log10_perc_urb, perc_agri,
    log10_pop_dens, sqrt_GDP_sum, sqrt_Area
  ) %>%
  pairs.panels(method = "spearman")
```



Population density has the highest correlation with naturalized richness. This is followed by urban area and PC2, but these variables are highly correlated with each other so selecting population density for the regression analysis.

PC3 has strongest negative correlation with richness and has low correlation with the other variables.

Step 6 Multi-Linear Regression

Some more libraries

```
library(lmtest)
library(lm.beta)
library(reshape2)
library(car)
```

Model 1

```
mal_rich_mod1 <- lm(Naturalize ~ log10_pop_dens + PC3, data = mal_rich_filt)

mal_rich_mod1 %>%
  summary()
```

```
##
## Call:
## lm(formula = Naturalize ~ log10_pop_dens + PC3, data = mal_rich_filt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -149.53  -60.94  -23.16   46.81  155.15
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -66.63     109.23  -0.610  0.56112
## log10_pop_dens  166.36      45.84   3.629  0.00841 **
## PC3            -40.57      19.50  -2.080  0.07608 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 106.5 on 7 degrees of freedom
## Multiple R-squared:  0.6975, Adjusted R-squared:  0.6111
## F-statistic: 8.072 on 2 and 7 DF, p-value: 0.01522
```

```
lm.beta(mal_rich_mod1)
```

```
##
## Call:
## lm(formula = Naturalize ~ log10_pop_dens + PC3, data = mal_rich_filt)
##
## Standardized Coefficients::
##      (Intercept) log10_pop_dens      PC3
##      0.0000000      0.7582702     -0.4345800
```

Test the model meets all the assumptions

#1) Normality (Shapiro Wilke)

```
mal_rich_mod1 %>%
  stats::rstandard() %>%
  stats::shapiro.test()
```

```
##
## Shapiro-Wilk normality test
##
## data:  .
## W = 0.95811, p-value = 0.7641
```

```
#Test is not significant so the residuals are normally distributed. The model is robust.
```

```
#2) Homoscedasticity (Breusch-Pagan)
```

```
mal_rich_mod1 %>%  
  lmtest::bptest()
```

```
##  
## studentized Breusch-Pagan test  
##  
## data: .  
## BP = 1.6058, df = 2, p-value = 0.448
```

```
# Not significant so the model is robust.
```

```
# 3) Independance (Durbin-Watson)
```

```
mal_rich_mod1 %>%  
  lmtest::dwtest()
```

```
##  
## Durbin-Watson test  
##  
## data: .  
## DW = 2.6324, p-value = 0.7939  
## alternative hypothesis: true autocorrelation is greater than 0
```

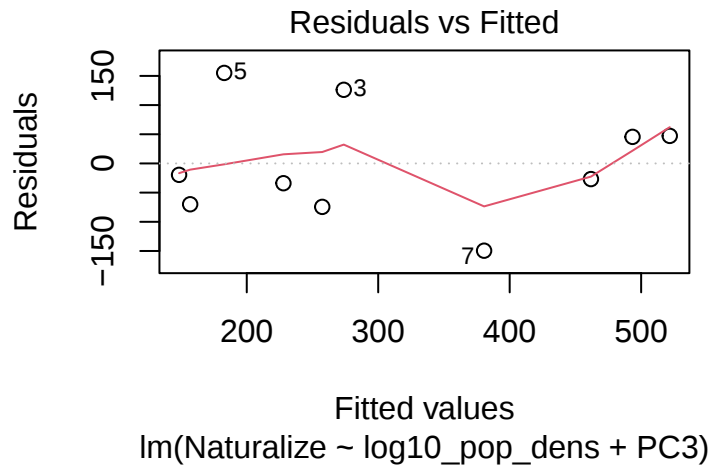
```
# 4) multicollinearity
```

```
mal_rich_mod1 %>%  
  vif()
```

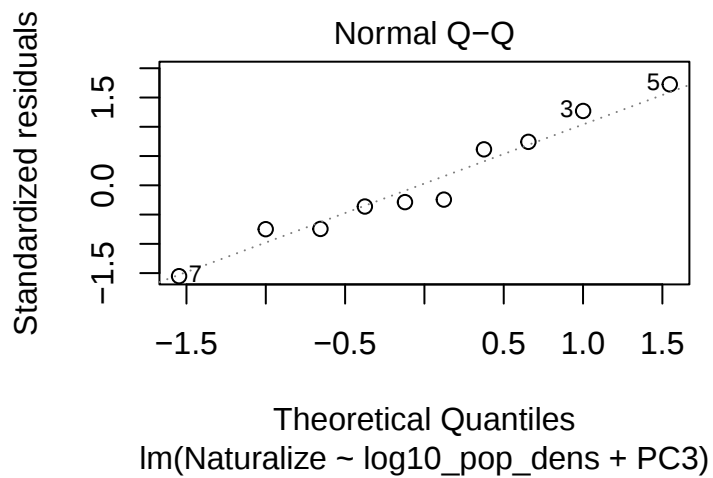
```
## log10_pop_dens          PC3  
##          1.010223          1.010223
```

```
#other checks
```

```
mal_rich_mod1 %>%  
  plot(which = c(1))
```

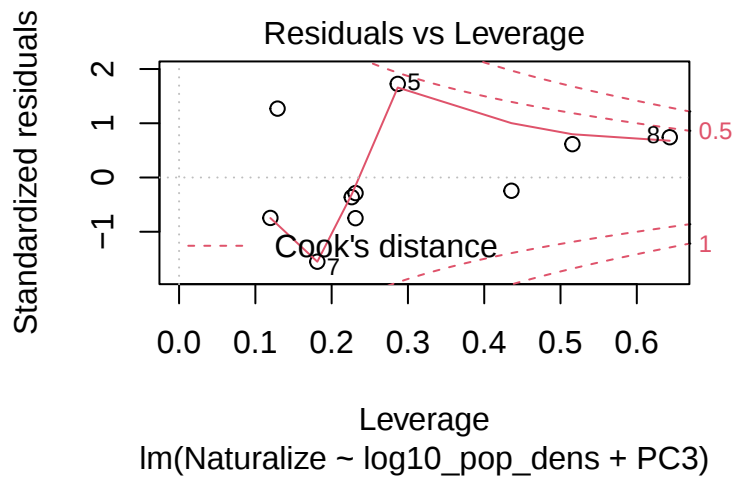


```
mal_rich_mod1 %>%
  plot(which = c(2))
```



```
# cooks distance for outliers

mal_rich_mod1 %>%
  plot(which = c(5))
```



Model is significant and robust meeting all assumptions.

Step 7 Influence of outliers on importance

```
mal_rich2 <- mal_rich_filt %>%
  filter(Island != "SGP") %>%
  filter(Island != "NWG")

mal_rich_filt %$%
  cor.test(Naturalize, perc_agri, method = "spearman")
```

```
##
## Spearman's rank correlation rho
##
## data: Naturalize and perc_agri
## S = 110, p-value = 0.3488
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.3333333
```

```
mal_rich2 %$%
  cor.test(Naturalize, perc_agri, method = "spearman")
```

```
##
## Spearman's rank correlation rho
##
## data: Naturalize and perc_agri
## S = 16, p-value = 0.02178
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.8095238
```