STAT 8020 R Lab 7: Multiple Linear Regression III

Whitney

September 09, 2020

Contents

Multicollinearity

Simulation

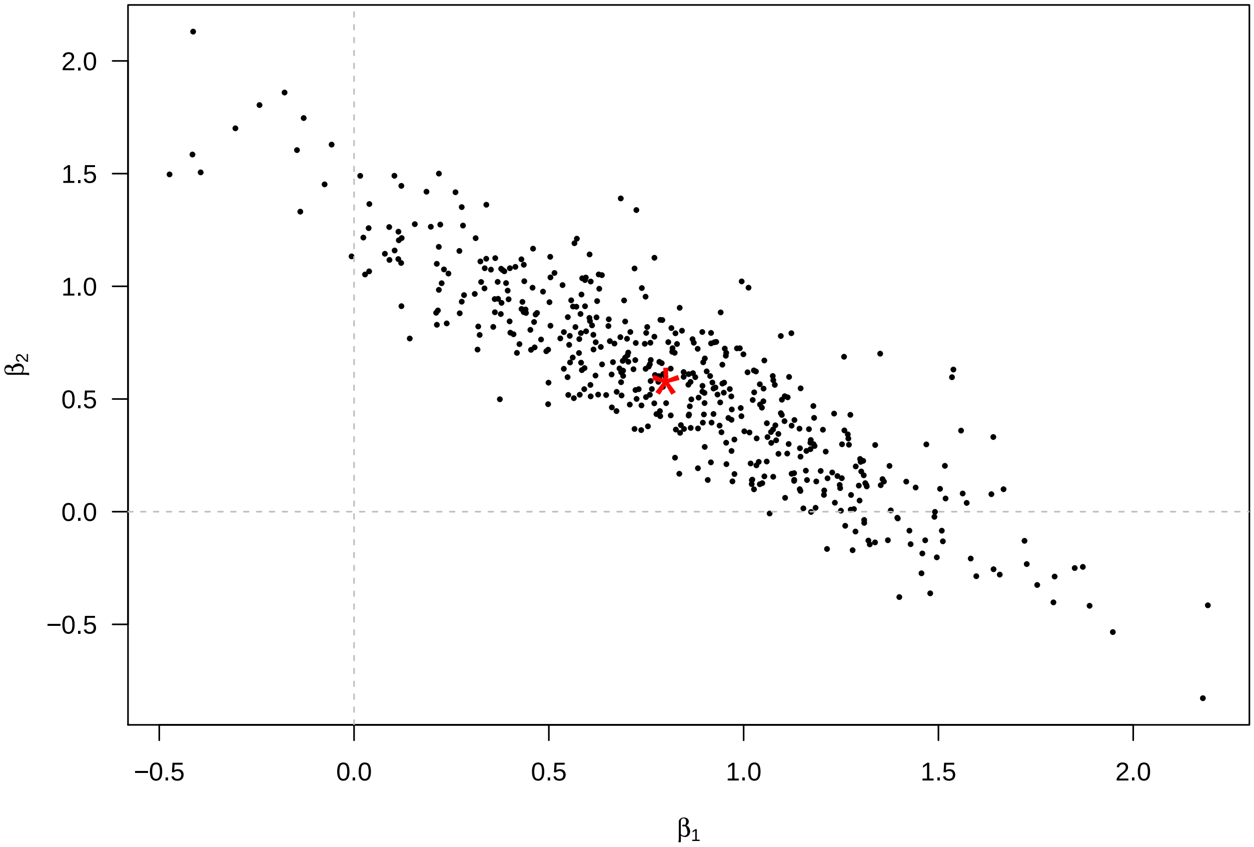
```
library(MASS)
N = 500
x <- replicate(N, mvrnorm(n = 30, c(0, 0), matrix(c(1, 0.9, 0.9, 1), 2)))
y <- array(dim = c(30, N))
for (i in 1:N){
  y[, i] = 4 + 0.8 * x[, 1, i] + 0.6 * x[, 2, i] + rnorm(30)
}
beta <- array(dim = c(3, N))
for (i in 1:N){
  beta[, i] <- lm(y[, i] ~ x[, 1, i] + x[, 2, i])$coefficients
}

R.sq_M1 <- numeric(N)
for (i in 1:N){
  R.sq_M1[i] <- summary(lm(y[, i] ~ x[, 1, i] + x[, 2, i]))$r.squared
}

summary(R.sq_M1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2470  0.6088  0.6779  0.6677  0.7422  0.8917
```

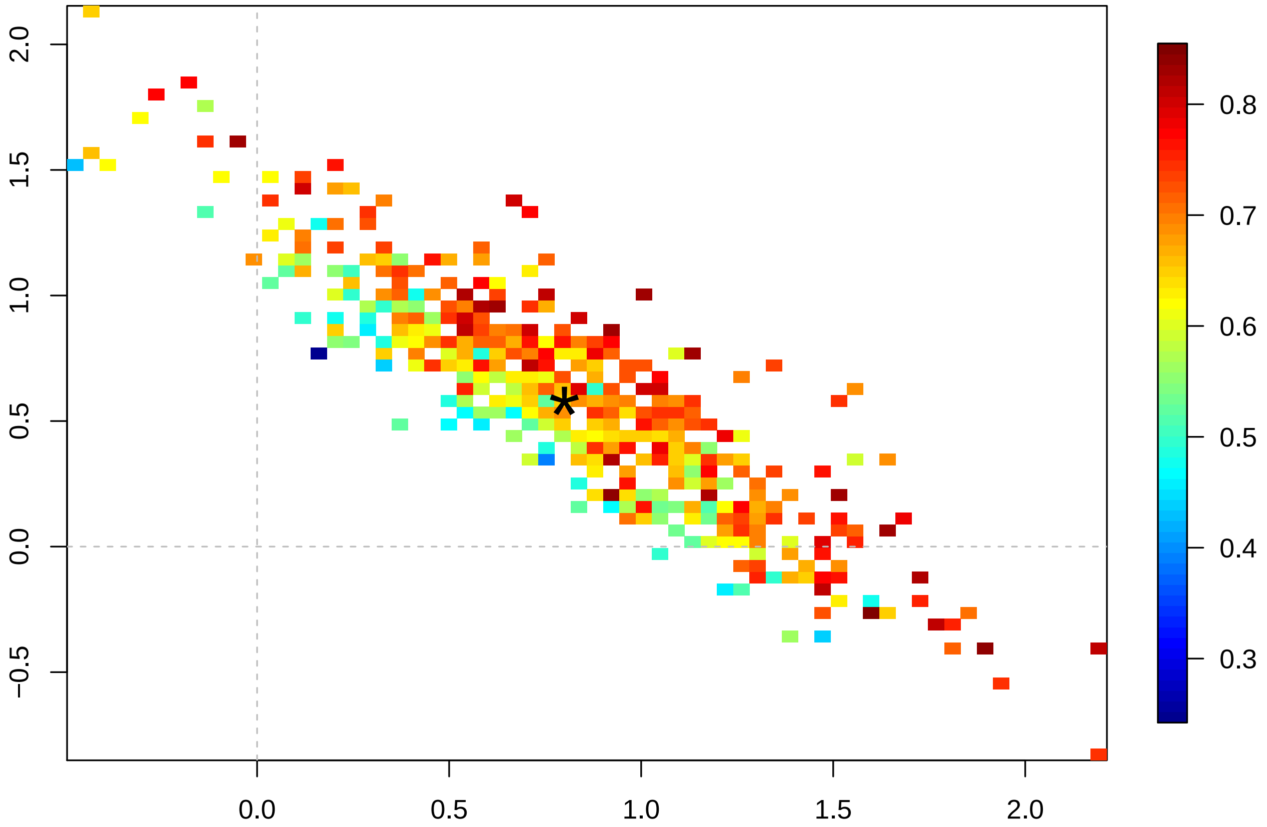
```
plot(beta[2,], beta[3,], pch = 16, cex = 0.5,
     xlab = expression(beta[1]),
     ylab = expression(beta[2]), las = 1)
points(0.8, 0.6, pch = "*", cex = 3, col = "red")
abline(h = 0, lty = 2, col = "gray")
abline(v = 0, lty = 2, col = "gray")

library(fields)
```

```
## Loading required package: spam
```

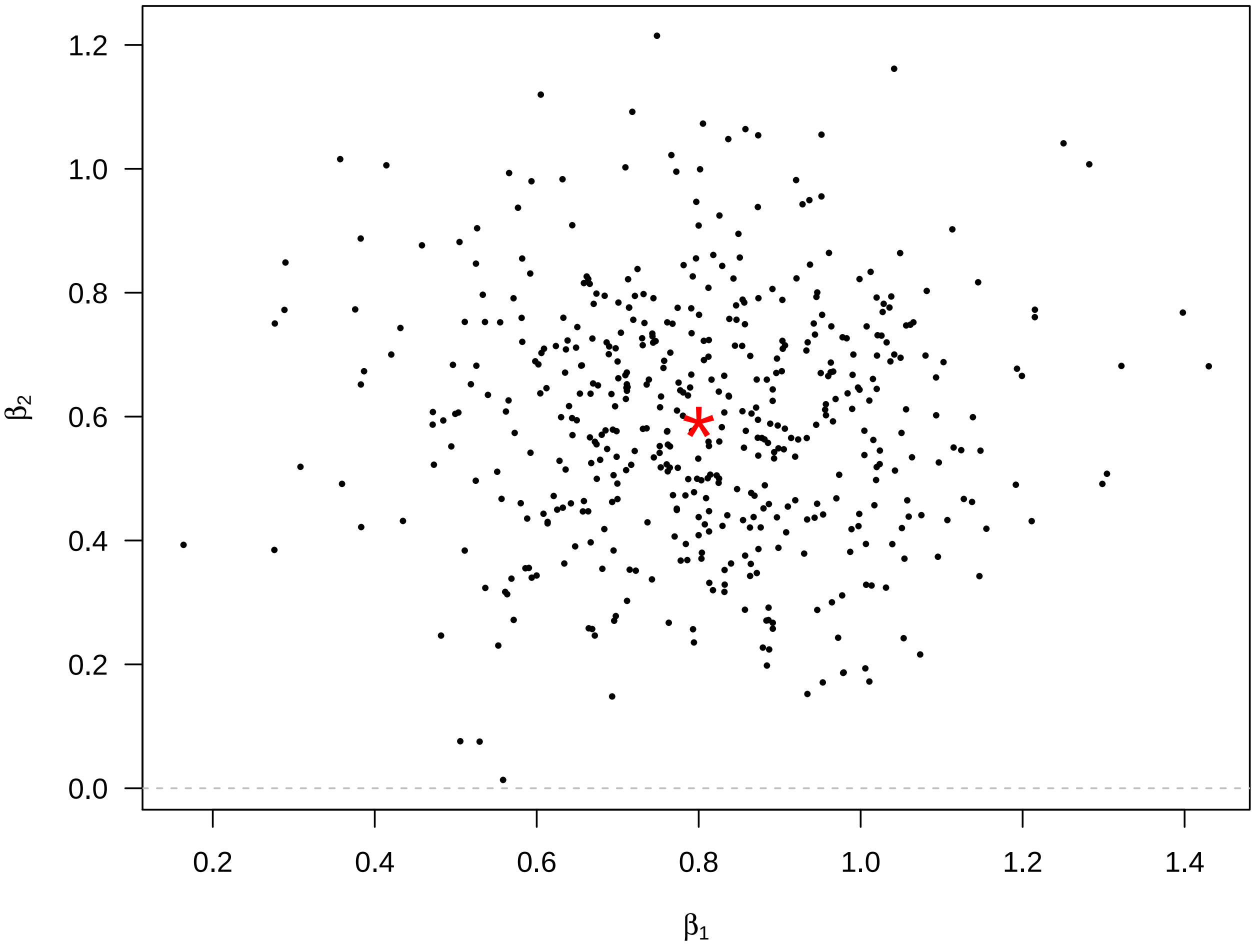
```
## Loading required package: dotCall64
## Loading required package: grid

## Spam version 2.4-0 (2019-11-01) is loaded.
## Type 'help( Spam)' or 'demo( spam)' for a short introduction
## and overview of this package.
## Help for individual functions is also obtained by adding the
## suffix '.spam' to the function name, e.g. 'help( chol.spam)'.

##
## Attaching package: 'spam'

## The following objects are masked from 'package:base':
##
##     backsolve, forwardsolve

## Loading required package: maps

## See https://github.com/NCAR/Fields for
##  an extensive vignette, other supplements and source code
```



```
quilt.plot(beta[2,], beta[3, ], R.sq_M1)
points(0.8, 0.6, pch = "*", cex = 3)
abline(h = 0, lty = 2, col = "gray")
abline(v = 0, lty = 2, col = "gray")
```



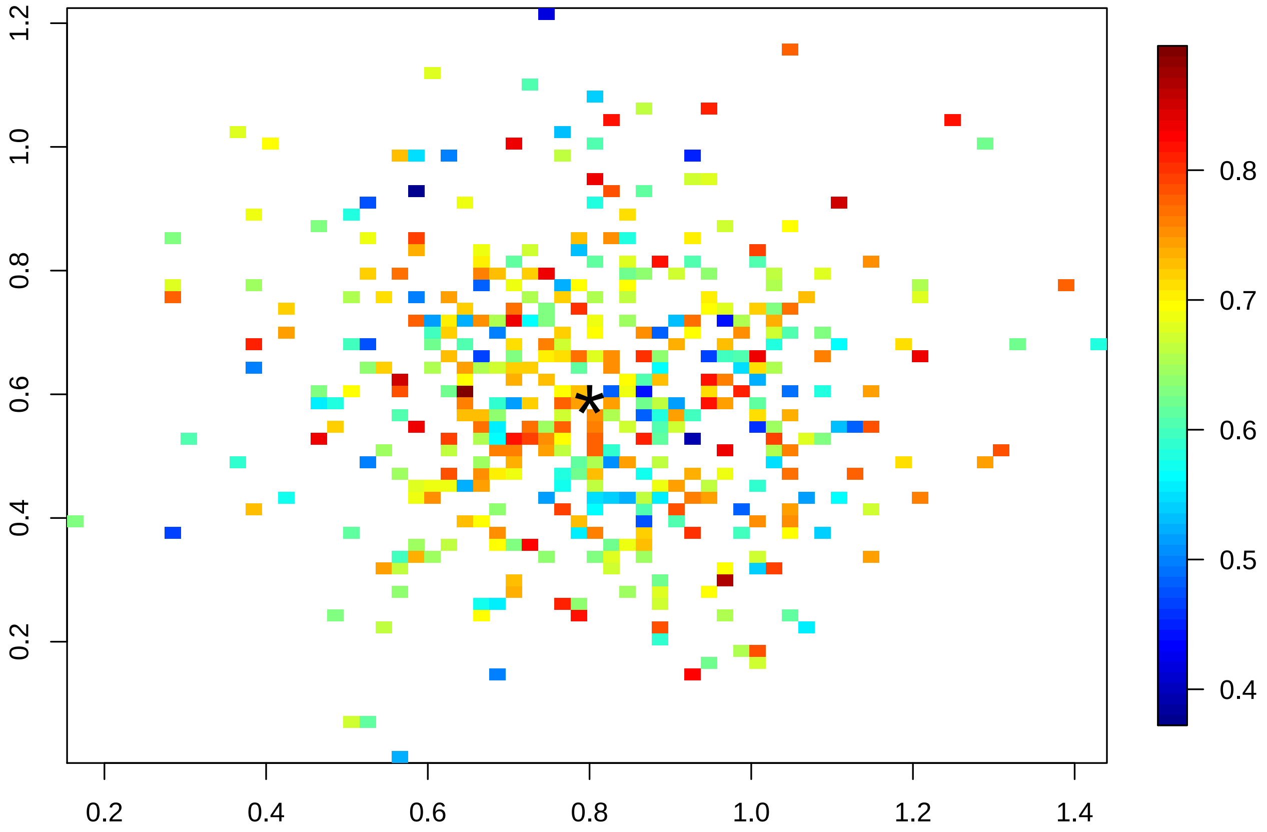
```
x1 <- replicate(N, mvrnorm(n = 30, c(0, 0), matrix(c(1, 0, 0, 1), 2)))
y1 <- array(dim = c(30, N))
for (i in 1:N){
  y1[, i] = 4 + 0.8 * x1[, 1, i] + 0.6 * x1[, 2, i] + rnorm(30)
}
beta1 <- array(dim = c(3, N))
for (i in 1:N){
  beta1[, i] <- lm(y1[, i] ~ x1[, 1, i] + x1[, 2, i])$coefficients
}

plot(beta1[2,], beta1[3,], pch = 16, cex = 0.5,
     xlab = expression(beta[1]),
     ylab = expression(beta[2]), las = 1)
points(0.8, 0.6, pch = "*", cex = 3, col = "red")
abline(h = 0, lty = 2, col = "gray")
abline(v = 0, lty = 2, col = "gray")
```



```
R.sq_M2 <- numeric(N)
for (i in 1:N){
  R.sq_M2[i] <- summary(lm(y1[, i] ~ x1[, 1, i] + x1[, 2, i]))$r.squared
}
summary(R.sq_M2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1186  0.4459  0.5303  0.5238  0.6144  0.8210
```

```
library(fields)
quilt.plot(beta1[2,], beta1[3, ], R.sq_M1)
points(0.8, 0.6, pch = "*", cex = 3)
abline(h = 0, lty = 2, col = "gray")
abline(v = 0, lty = 2, col = "gray")
```



Species diversity on the Galapagos Islands

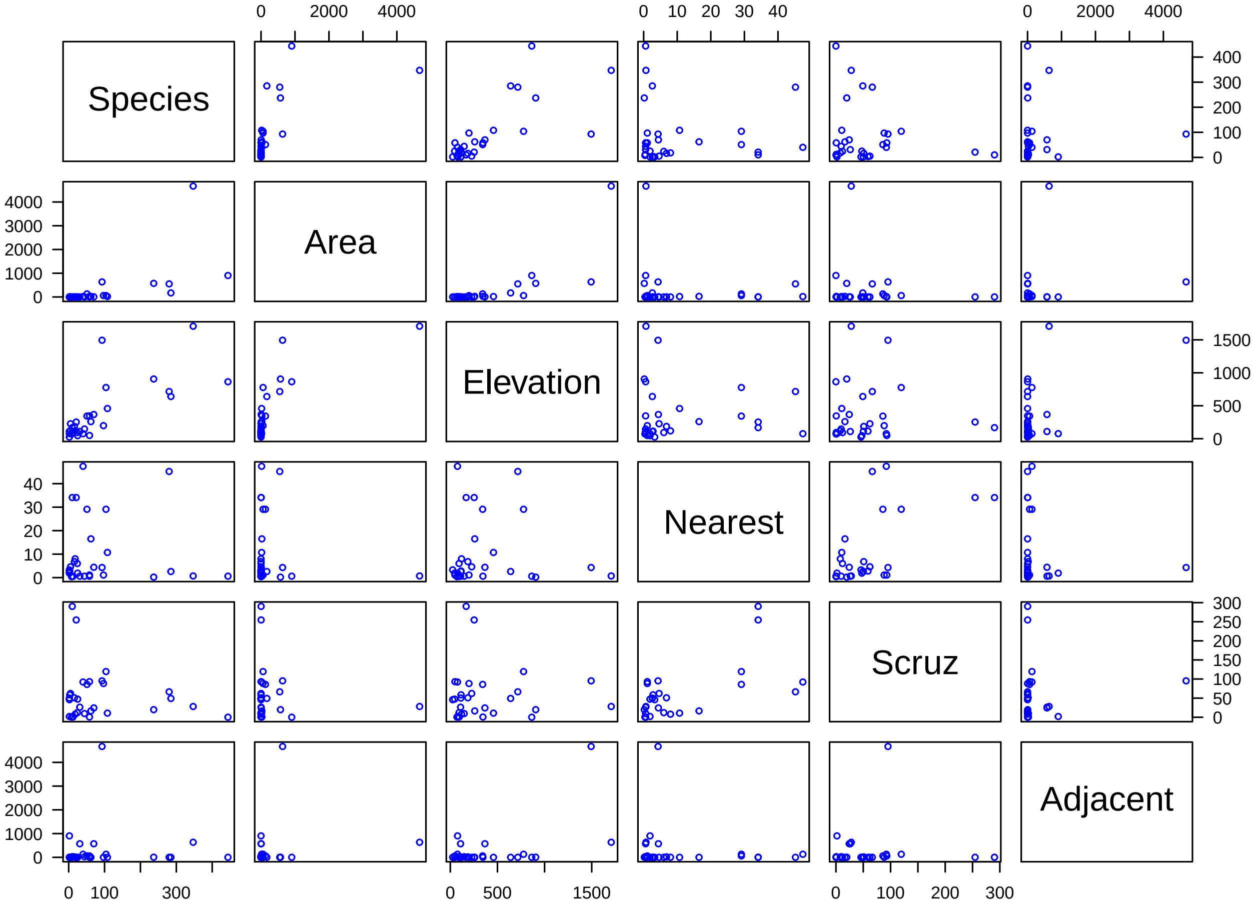
```
library(faraway)
```

```
##
## Attaching package: 'faraway'
## The following object is masked from 'package:maps':
##
##     ozone
```

```
data(gala)
```

Plot the pairwise scatterplots

```
galaNew <- gala[, -2]
plot(galaNew, cex = 0.75, col = "blue", las = 1)
```



Correlation matrix

```r
cor(galaNew)
```

```
##                Species       Area   Elevation     Nearest       Scruz
## Species     1.00000000  0.6178431  0.73848666 -0.01409407 -0.17114244
## Area        0.61784307  1.0000000  0.75373492 -0.11110320 -0.10078493
## Elevation   0.73848666  0.7537349  1.00000000 -0.01107698 -0.01543829
## Nearest    -0.01409407 -0.1111032 -0.01107698  1.00000000  0.61541036
## Scruz      -0.17114244 -0.1007849 -0.01543829  0.61541036  1.00000000
## Adjacent    0.02616635  0.1800376  0.53645782 -0.11624788  0.05166066
##               Adjacent
## Species     0.02616635
## Area        0.18003759
## Elevation   0.53645782
## Nearest    -0.11624788
## Scruz       0.05166066
## Adjacent    1.00000000
```

Variance inflation factor

```r
m <- lm(Species ~ ., data = galaNew)
vif(m)
```

```
##      Area Elevation   Nearest     Scruz  Adjacent
##  2.928145  3.992545  1.766099  1.675031  1.826403
```

```
## Check
r.sq_ele <- summary(lm(Elevation ~ Area + Nearest + Scruz + Adjacent, data = galaNew))$r.square

vif <- 1 / (1 - r.sq_ele)
```

Model Selection

```
library(tidyverse)
```

```
## -- Attaching packages ----------------- tidyverse 1.3.0 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts -------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::map()    masks maps::map()
## x dplyr::select() masks MASS::select()
```

```
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'lattice'

## The following object is masked from 'package:faraway':
##
##     melanoma

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(leaps)
models <- regsubsets(Species ~ ., data = galaNew, nvmax = 5)
summary(models)
```

```
## Subset selection object
## Call: regsubsets.formula(Species ~ ., data = galaNew, nvmax = 5)
## 5 Variables  (and intercept)
##           Forced in Forced out
## Area          FALSE      FALSE
## Elevation     FALSE      FALSE
## Nearest       FALSE      FALSE
## Scruz         FALSE      FALSE
## Adjacent      FALSE      FALSE
## 1 subsets of each size up to 5
## Selection Algorithm: exhaustive
##          Area Elevation Nearest Scruz Adjacent
## 1  ( 1 ) " "  "*"       " "     " "   " "
## 2  ( 1 ) " "  "*"       " "     " "   "*"
## 3  ( 1 ) " "  "*"       " "     "*"   "*"
## 4  ( 1 ) "*"  "*"       " "     "*"   "*"
## 5  ( 1 ) "*"  "*"       "*"     "*"   "*"
```

```
res.sum <- summary(models)

criteria <- data.frame(
  Adj.R2 = res.sum$adjr2,
  Cp = res.sum$cp,
  BIC = res.sum$bic)

criteria
```

```
##      Adj.R2        Cp       BIC
## 1 0.5291255 20.599003 -16.84525
## 2 0.7181425  2.897184 -29.93078
## 3 0.7258462  3.193068 -28.49317
## 4 0.7283816  4.000075 -26.54733
## 5 0.7170651  6.000000 -23.14622
```

```
full <- lm(Species ~ ., data = galaNew)
step(full)
```

```
## Start:  AIC=251.93
## Species ~ Area + Elevation + Nearest + Scruz + Adjacent
##
##             Df Sum of Sq    RSS    AIC
## - Nearest    1         0  89232 249.93
## - Area       1      4238  93469 251.33
## - Scruz      1      4636  93867 251.45
## <none>                    89231 251.93
## - Adjacent   1     66406 155638 266.62
## - Elevation  1    131767 220998 277.14
##
## Step:  AIC=249.93
## Species ~ Area + Elevation + Scruz + Adjacent
##
##             Df Sum of Sq    RSS    AIC
## - Area       1      4436  93667 249.39
## <none>                    89232 249.93
## - Scruz      1      7544  96776 250.37
## - Adjacent   1     72312 161544 265.74
## - Elevation  1    139445 228677 276.17
##
## Step:  AIC=249.39
## Species ~ Elevation + Scruz + Adjacent
##
##             Df Sum of Sq    RSS    AIC
## - Scruz      1      6336 100003 249.35
## <none>                    93667 249.39
## - Adjacent   1     69860 163527 264.11
## - Elevation  1    275784 369451 288.56
##
## Step:  AIC=249.35
## Species ~ Elevation + Adjacent
```

```
##
##             Df Sum of Sq    RSS    AIC
## <none>                   100003 249.35
## - Adjacent   1     73251 173254 263.84
## - Elevation  1    280817 380820 287.47
```

```
##
## Call:
## lm(formula = Species ~ Elevation + Adjacent, data = galaNew)
##
## Coefficients:
## (Intercept)    Elevation     Adjacent
##     1.43287      0.27657     -0.06889
```

```r
step(full, direction = "backward")
```

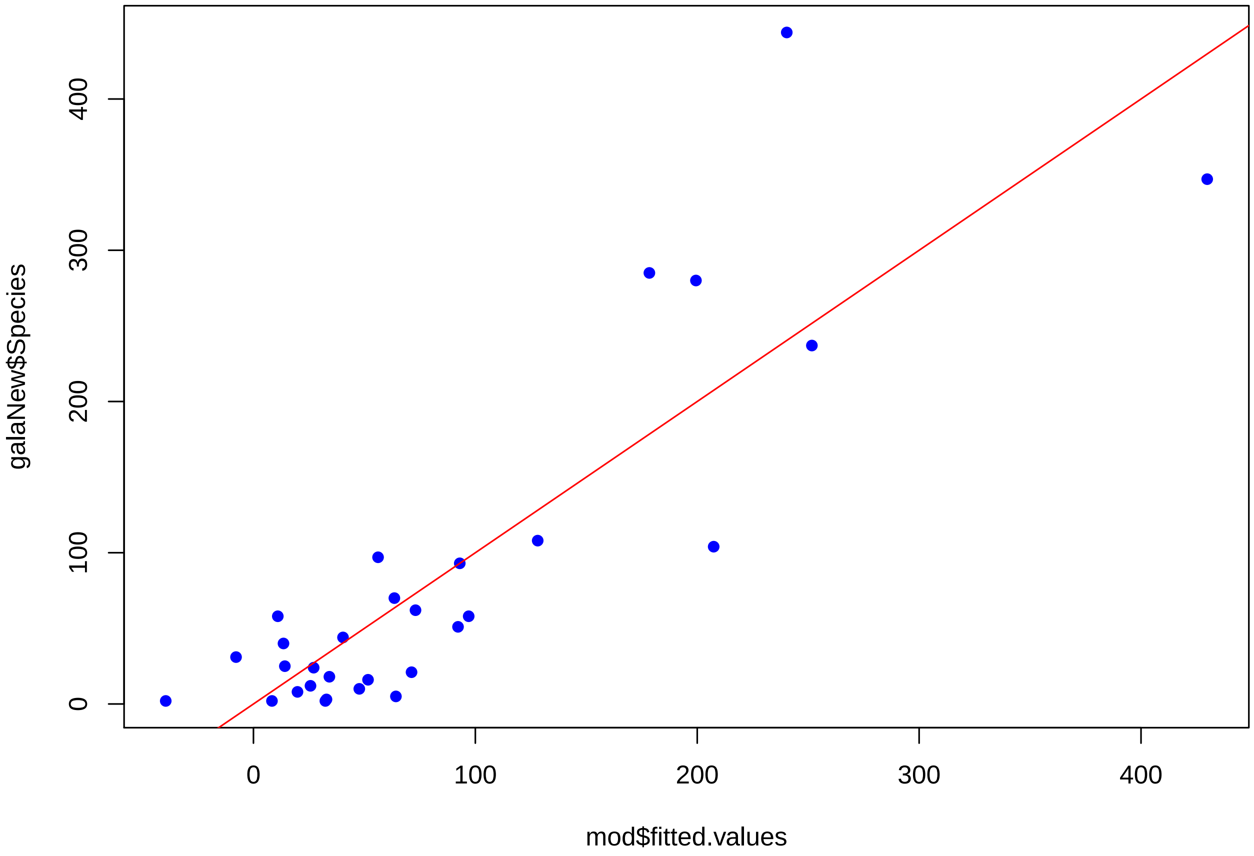
```
## Start:  AIC=251.93
## Species ~ Area + Elevation + Nearest + Scruz + Adjacent
##
##             Df Sum of Sq    RSS    AIC
## - Nearest    1         0  89232 249.93
## - Area       1      4238  93469 251.33
## - Scruz      1      4636  93867 251.45
## <none>                    89231 251.93
## - Adjacent   1     66406 155638 266.62
## - Elevation  1    131767 220998 277.14
##
## Step:  AIC=249.93
## Species ~ Area + Elevation + Scruz + Adjacent
##
##             Df Sum of Sq    RSS    AIC
## - Area       1      4436  93667 249.39
## <none>                    89232 249.93
## - Scruz      1      7544  96776 250.37
## - Adjacent   1     72312 161544 265.74
## - Elevation  1    139445 228677 276.17
##
## Step:  AIC=249.39
## Species ~ Elevation + Scruz + Adjacent
##
##             Df Sum of Sq    RSS    AIC
## - Scruz      1      6336 100003 249.35
## <none>                    93667 249.39
## - Adjacent   1     69860 163527 264.11
## - Elevation  1    275784 369451 288.56
##
## Step:  AIC=249.35
## Species ~ Elevation + Adjacent
##
##             Df Sum of Sq    RSS    AIC
## <none>                   100003 249.35
## - Adjacent   1     73251 173254 263.84
## - Elevation  1    280817 380820 287.47
```

```
##
## Call:
```

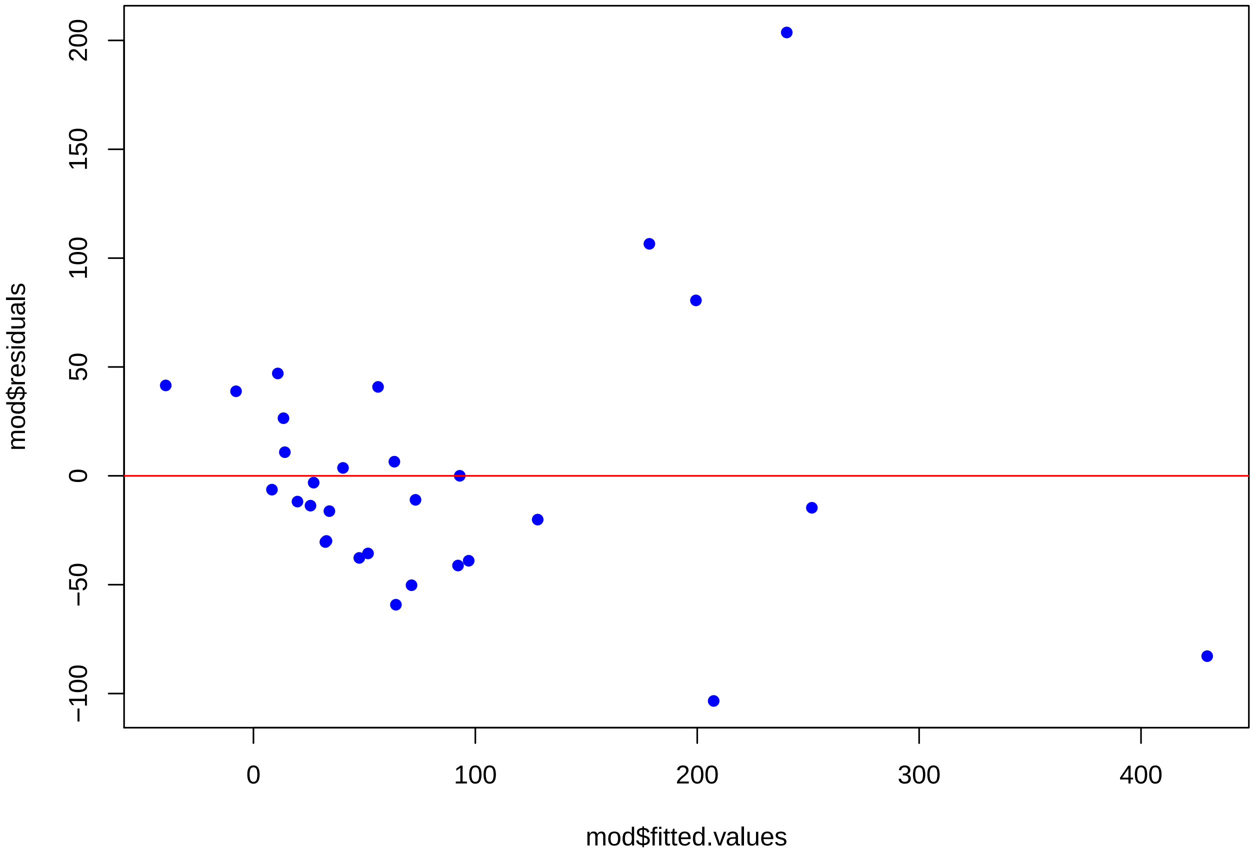
```
## lm(formula = Species ~ Elevation + Adjacent, data = galaNew)
##
## Coefficients:
## (Intercept)    Elevation     Adjacent
##     1.43287      0.27657     -0.06889
```

Model Diagnostics

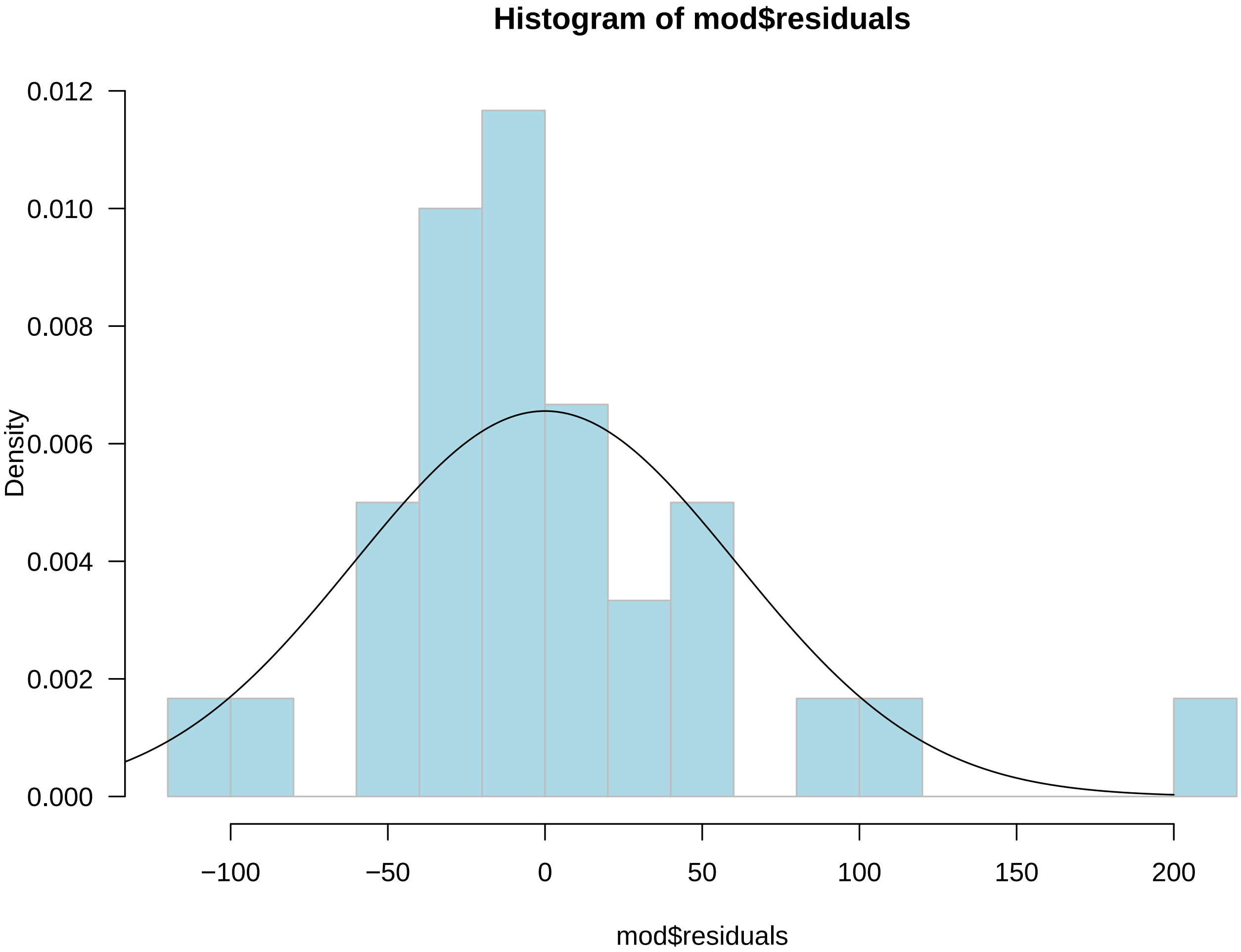
```
mod <- lm(Species ~ Elevation + Adjacent, data = galaNew)
plot(mod$fitted.values, galaNew$Species, pch = 16, col = "blue")
abline(0, 1, col = "red")
```



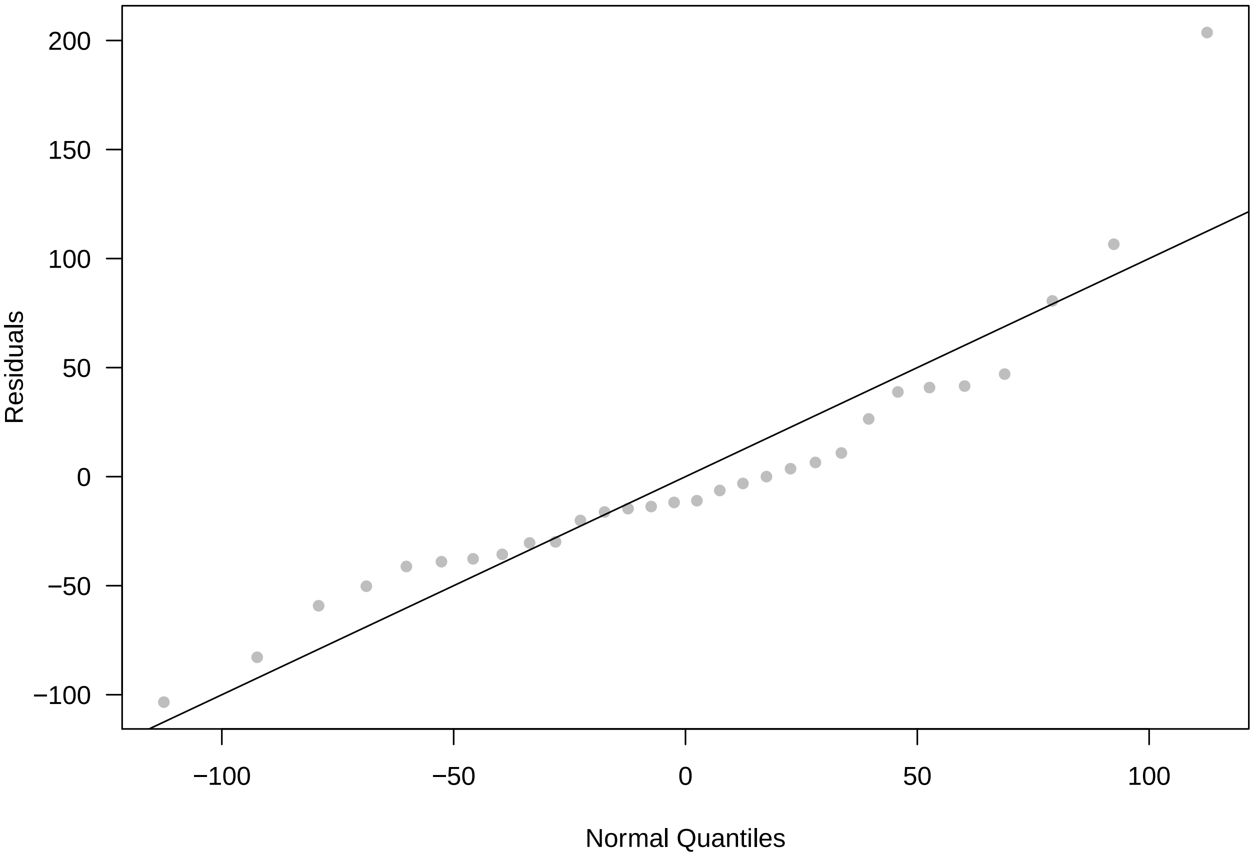
```
plot(mod$fitted.values, mod$residuals, pch = 16, col = "blue")
abline(h = 0, col = "red")
```



```
par(las = 1)
hist(mod$residuals, 12, prob = T,
     col = "lightblue", border = "gray")
xg <- seq(-200, 200, 1)
yg <- dnorm(xg, 0, 60.86)
lines(xg, yg)
```



```
plot(qnorm(1:30 / 31, 0, 60.86), sort(mod$residuals), pch = 16,
     col = "gray", xlab = "Normal Quantiles", ylab = "Residuals")
abline(0, 1)
```



200
150
100
50
0
−50
−100
Residuals
−100
−50
0
50
100
Normal Quantiles