STAT 8020 R Lab 6: Multiple Linear Regression II

Whitney

September 08, 2020

Contents

Species diversity on the Galapagos Islands

First Step: Load the data

```
#installinstall.packages("faraway")
library(faraway)
data(gala)

gala
```

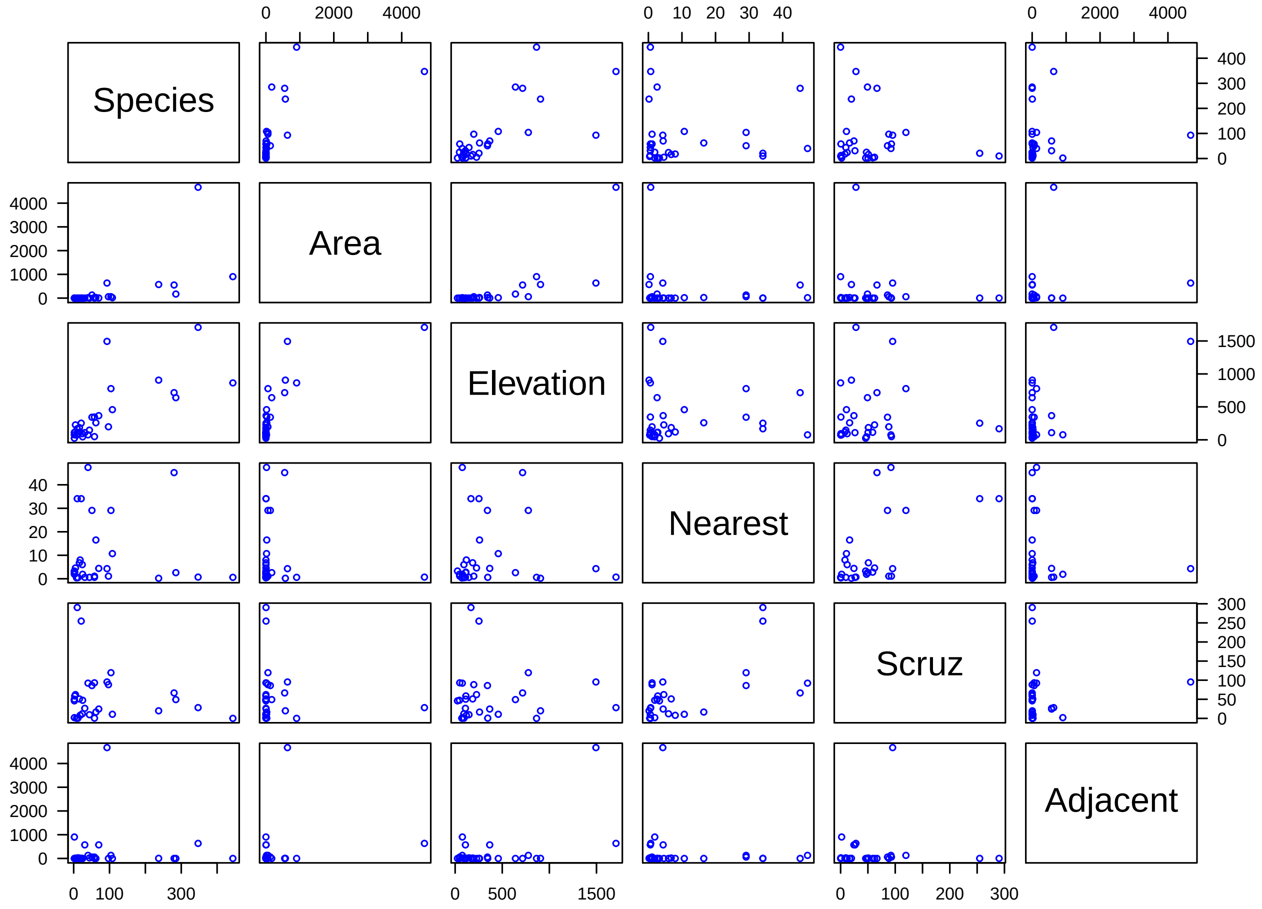
```
##              Species Endemics    Area Elevation Nearest Scruz Adjacent
## Baltra            58       23   25.09       346     0.6   0.6     1.84
## Bartolome         31       21    1.24       109     0.6  26.3   572.33
## Caldwell           3        3    0.21       114     2.8  58.7     0.78
## Champion          25        9    0.10        46     1.9  47.4     0.18
## Coamano            2        1    0.05        77     1.9   1.9   903.82
## Daphne.Major      18       11    0.34       119     8.0   8.0     1.84
## Daphne.Minor      24        0    0.08        93     6.0  12.0     0.34
## Darwin            10        7    2.33       168    34.1 290.2     2.85
## Eden               8        4    0.03        71     0.4   0.4    17.95
## Enderby            2        2    0.18       112     2.6  50.2     0.10
## Espanola          97       26   58.27       198     1.1  88.3     0.57
## Fernandina        93       35  634.49      1494     4.3  95.3  4669.32
## Gardner1          58       17    0.57        49     1.1  93.1    58.27
## Gardner2           5        4    0.78       227     4.6  62.2     0.21
## Genovesa          40       19   17.35        76    47.4  92.2   129.49
## Isabela          347       89 4669.32      1707     0.7  28.1   634.49
## Marchena          51       23  129.49       343    29.1  85.9    59.56
```

```
## Onslow             2        2    0.01        25     3.3  45.9     0.10
## Pinta            104       37   59.56       777    29.1 119.6   129.49
## Pinzon           108       33   17.95       458    10.7  10.7     0.03
## Las.Plazas        12        9    0.23        94     0.5   0.6    25.09
## Rabida            70       30    4.89       367     4.4  24.4   572.33
## SanCristobal     280       65  551.62       716    45.2  66.6     0.57
## SanSalvador      237       81  572.33       906     0.2  19.8     4.89
## SantaCruz        444       95  903.82       864     0.6   0.0     0.52
## SantaFe           62       28   24.08       259    16.5  16.5     0.52
## SantaMaria       285       73  170.92       640     2.6  49.2     0.10
## Seymour           44       16    1.84       147     0.6   9.6    25.09
## Tortuga           16        8    1.24       186     6.8  50.9    17.95
## Wolf              21       12    2.85       253    34.1 254.7     2.33
```

```
#Out the data in csv
#write.csv(gala, file = "gala.csv")
```

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
```

Model 1: Fitting a simple linear regression

Here we use *Elevation* as the predictor as it has the highest correlation with *Species*

```r
M1 <- lm(Species ~ Elevation, data = galaNew)
summary(M1)
```

```
##
## Call:
## lm(formula = Species ~ Elevation, data = galaNew)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -218.319  -30.721  -14.690    4.634  259.180
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.33511   19.20529   0.590     0.56
## Elevation    0.20079    0.03465   5.795 3.18e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78.66 on 28 degrees of freedom
## Multiple R-squared:  0.5454, Adjusted R-squared:  0.5291
## F-statistic: 33.59 on 1 and 28 DF,  p-value: 3.177e-06
```

Model 2: Adding *Area*

```r
M2 <- lm(Species ~ Elevation + Area, data = galaNew)
summary(M2)
```

```
##
## Call:
## lm(formula = Species ~ Elevation + Area, data = galaNew)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -192.619  -33.534  -19.199    7.541  261.514
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.10519   20.94211   0.817  0.42120
## Elevation    0.17174    0.05317   3.230  0.00325 **
## Area         0.01880    0.02594   0.725  0.47478
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 79.34 on 27 degrees of freedom
## Multiple R-squared:  0.554,  Adjusted R-squared:  0.521
## F-statistic: 16.77 on 2 and 27 DF,  p-value: 1.843e-05
```

Model 3: Adding *Adjacent*

```
M3 <- lm(Species ~ Elevation + Area + Adjacent, data = galaNew)
summary(M3)
```

```
##
## Call:
## lm(formula = Species ~ Elevation + Area + Adjacent, data = galaNew)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -124.064  -34.283   -8.733   27.972  195.973
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.71893   16.90706  -0.338  0.73789
## Elevation    0.31498    0.05211   6.044  2.2e-06 ***
## Area        -0.02031    0.02181  -0.931  0.36034
## Adjacent    -0.07528    0.01698  -4.434  0.00015 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61.01 on 26 degrees of freedom
## Multiple R-squared:  0.746,  Adjusted R-squared:  0.7167
## F-statistic: 25.46 on 3 and 26 DF,  p-value: 6.683e-08
```

Full Model

```
M4 <- lm(Species ~ ., data = galaNew)
summary(M4)
```

```
##
## Call:
## lm(formula = Species ~ ., data = galaNew)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -111.679  -34.898   -7.862   33.460  182.584
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.068221  19.154198   0.369 0.715351
## Area        -0.023938   0.022422  -1.068 0.296318
## Elevation    0.319465   0.053663   5.953 3.82e-06 ***
## Nearest      0.009144   1.054136   0.009 0.993151
## Scruz       -0.240524   0.215402  -1.117 0.275208
## Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.98 on 24 degrees of freedom
## Multiple R-squared:  0.7658, Adjusted R-squared:  0.7171
## F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```

Parameter Estimation

```r
X <- model.matrix(M4)
y <- gala$Species
# regression parameters
beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y
#beta_hat_faster <- solve(crossprod(X), crossprod(X, y))
# fitted values
y_hat <- X %*% solve(t(X) %*% X) %*% t(X) %*% y
```

ANOVA

```r
anova(M4)
```

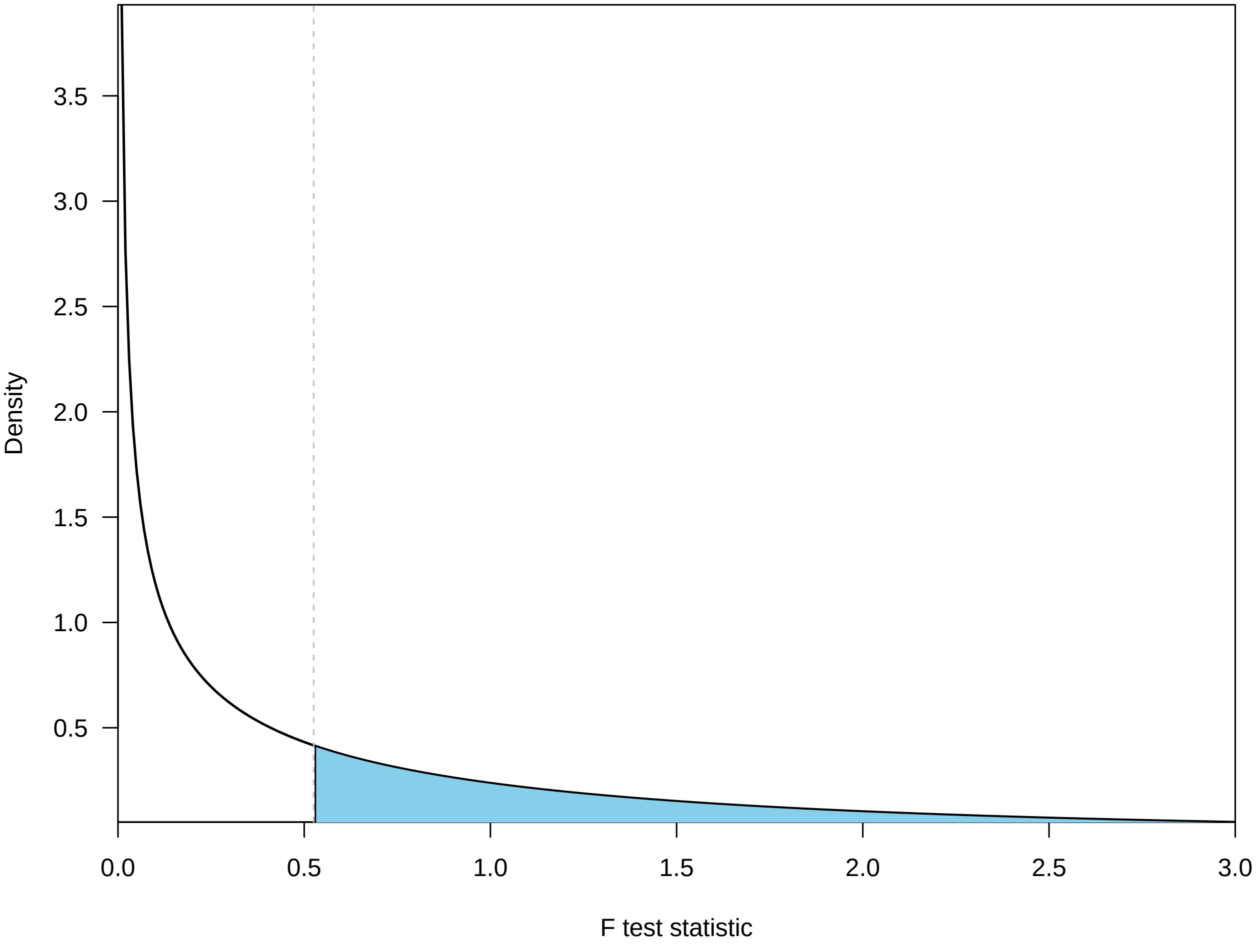
```
## Analysis of Variance Table
##
## Response: Species
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Area       1 145470  145470 39.1262 1.826e-06 ***
## Elevation  1  65664   65664 17.6613 0.0003155 ***
## Nearest    1     29      29  0.0079 0.9300674
## Scruz      1  14280   14280  3.8408 0.0617324 .
## Adjacent   1  66406   66406 17.8609 0.0002971 ***
## Residuals 24  89231    3718
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

General Linear Test

```r
anova(M1, M2)
```

```
## Analysis of Variance Table
##
## Model 1: Species ~ Elevation
## Model 2: Species ~ Elevation + Area
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 173254
## 2     27 169947  1      3307 0.5254 0.4748
```

```
par(las = 1, mar = c(4.1, 4.1, 1.1, 1.1))
xg <- seq(0, 3, 0.01)
yg <- df(xg, 1, 27)
plot(xg, yg, type = "l", xaxs = "i", yaxs = "i", lwd = 1.6,
     xlab = "F test statistic", ylab = "Density")
abline(v = 0.5254, lty = 2, col = "gray")
polygon(c(xg[xg > 0.5254], rev(xg[xg > 0.5254])),
        c(yg[xg > 0.5254], rep(0, length(yg[xg > 0.5254]))),
        col = "skyblue")
```



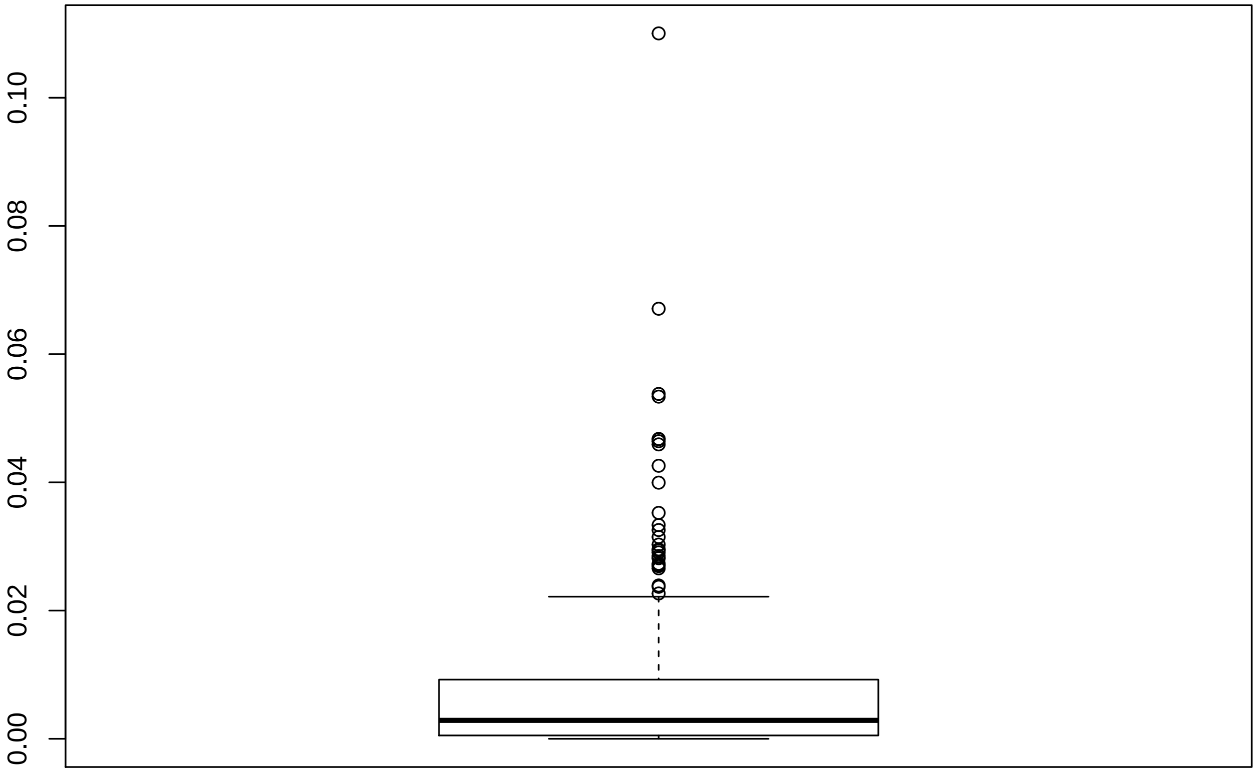
```
Full <- M4
Reduce <- lm(Species ~ Elevation + Adjacent, data = galaNew)

anova(Reduce, Full)
```

```
## Analysis of Variance Table
##
## Model 1: Species ~ Elevation + Adjacent
## Model 2: Species ~ Area + Elevation + Nearest + Scruz + Adjacent
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     27 100003
## 2     24  89231  3     10772 0.9657  0.425
```

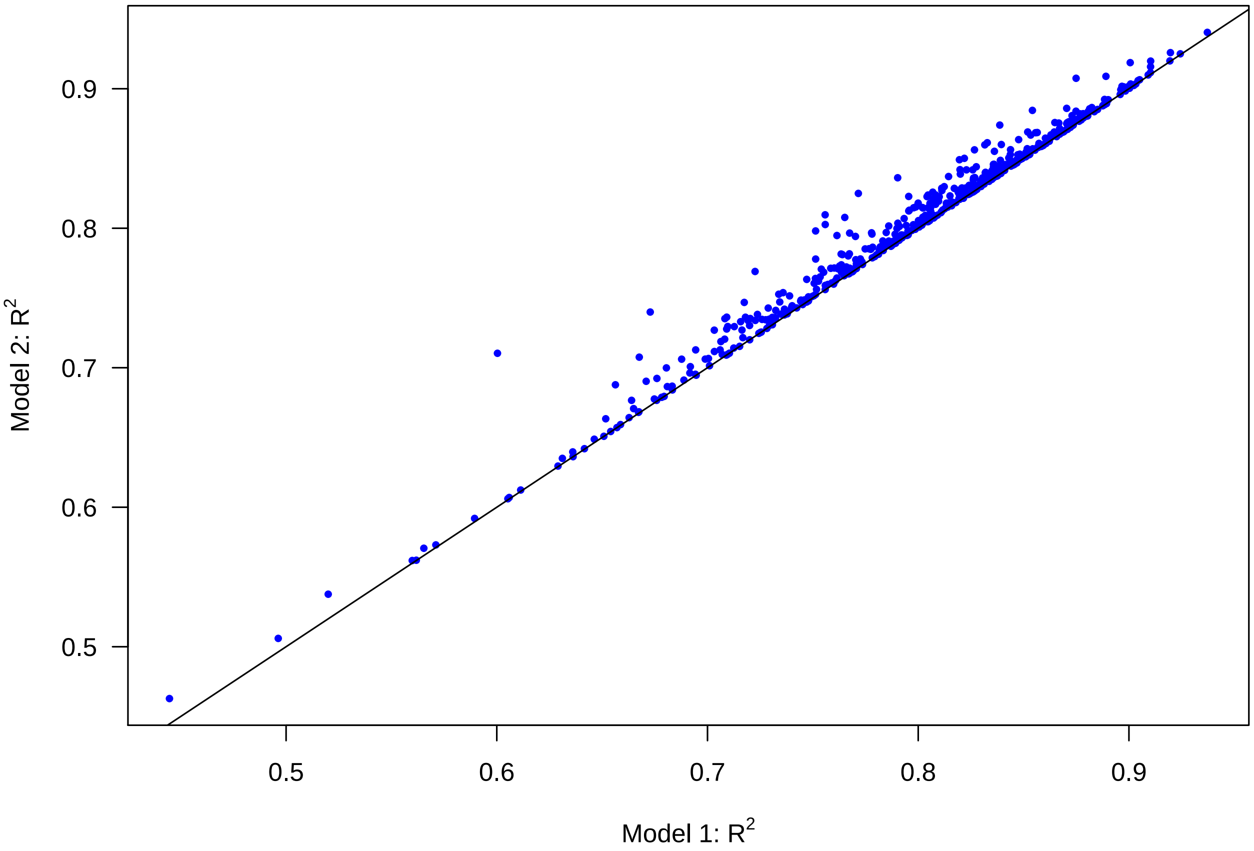
Simulation

R^2 vs. R^2_{adj}

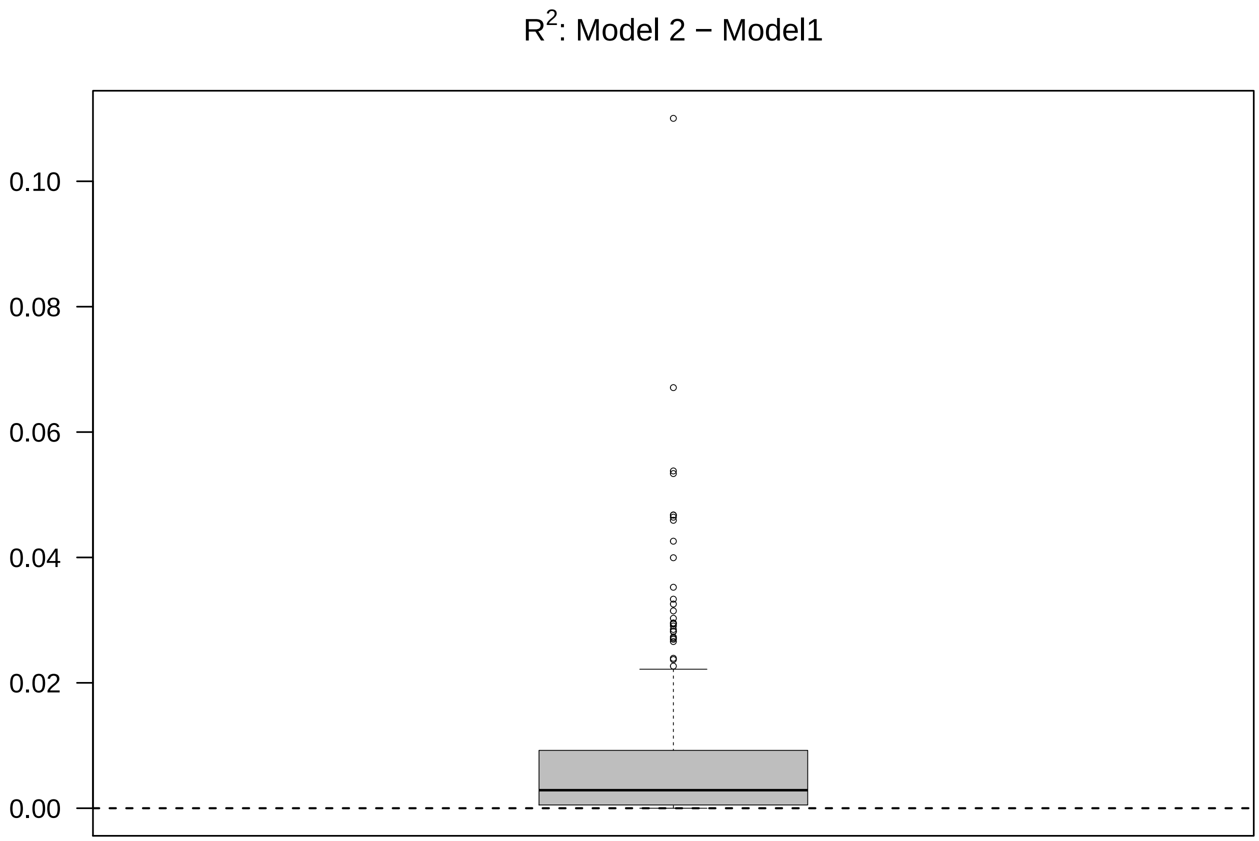
```r
set.seed(123)
N = 500
x1 <- replicate(N, rnorm(30))
x2 <- replicate(N, rnorm(30))
y1 <- apply(x1, 2, function(x) 5 + 2 * x + rnorm(30, 0, 1))
R.sq <- array(dim = c(N, 4))
for (i in 1:N){
  R.sq[i, 1] = summary(lm(y1[, i] ~ x1[, i]))$r.squared
  R.sq[i, 2] = summary(lm(y1[, i] ~ x1[, i]))$adj.r.squared
  R.sq[i, 3] = summary(lm(y1[, i] ~ x1[, i] + x2[, i]))$r.squared
  R.sq[i, 4] = summary(lm(y1[, i] ~ x1[, i] + x2[, i]))$adj.r.squared
}



boxplot(R.sq[, 3] - R.sq[, 1])
```



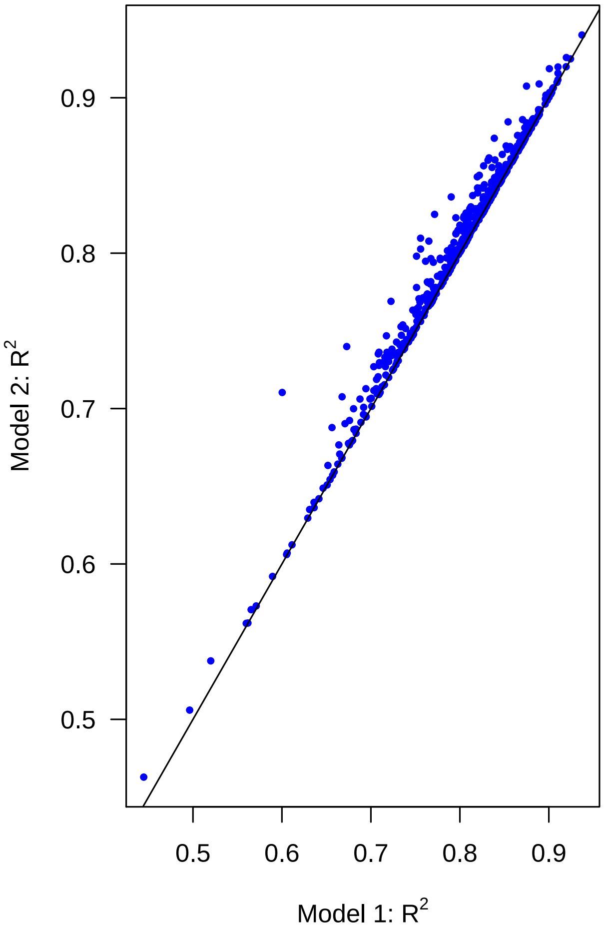
```r
par(las = 1)
plot(R.sq[, 1], R.sq[, 3], pch = 16, cex = 0.65, col = "blue",
     xlab = expression(paste("Model 1: ", R^2)),
     ylab = expression(paste("Model 2: ", R^2)))
abline(0, 1)
```

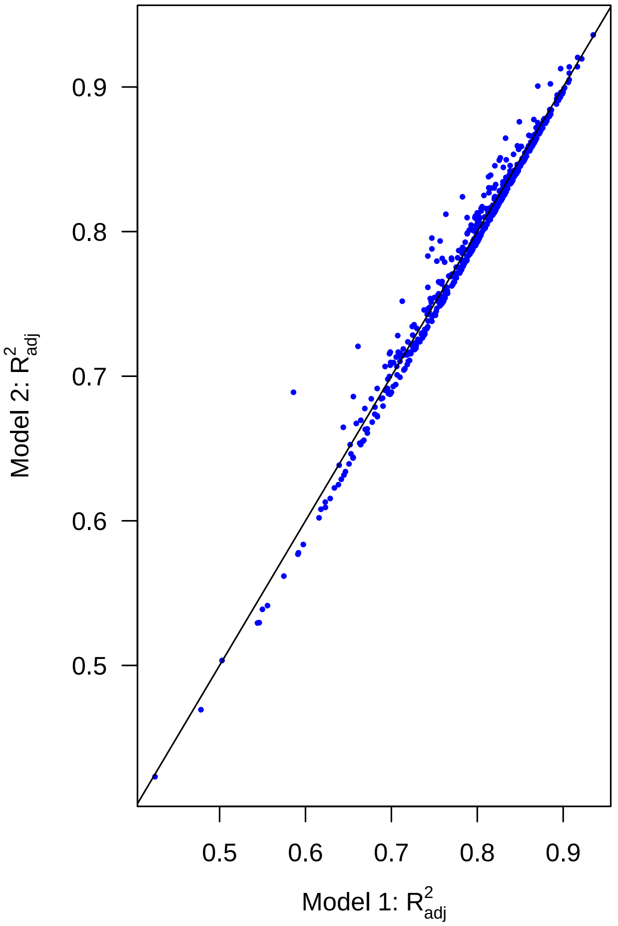


```
boxplot(R.sq[, 3] - R.sq[, 1], las = 1,
        main = expression(paste(R^2, ": Model 2 - Model1")),
        col = "gray", boxwex = 0.5, staplewex = 0.25,
        outwex = 0.25, lwd = 0.5, outcex = 0.5)
abline(h = 0, lwd = 1.25, lty = 2)
```

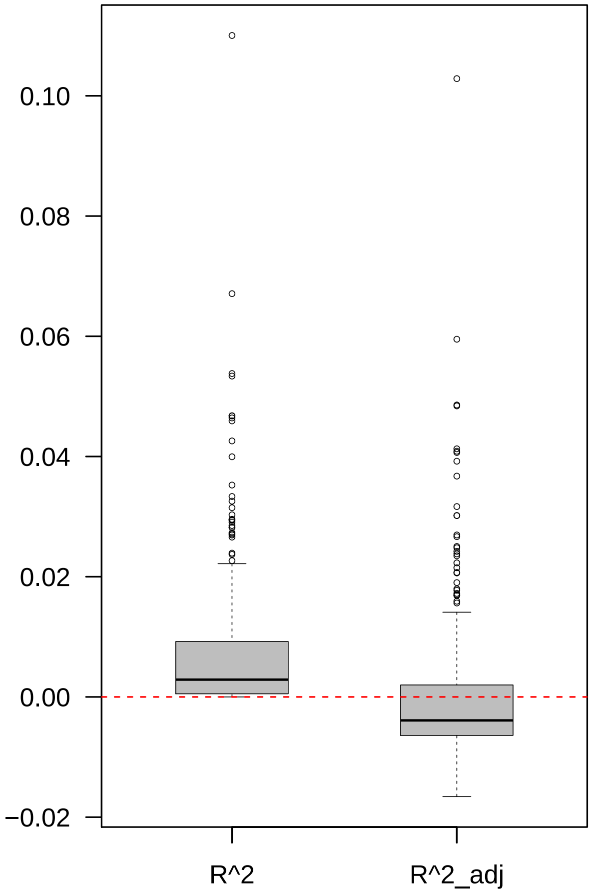


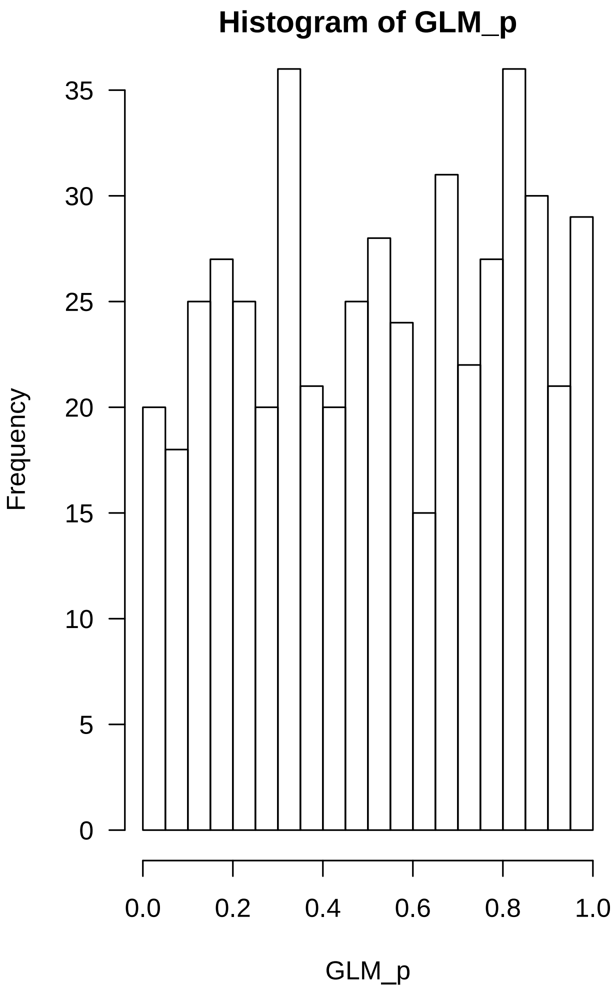
```
par(las = 1, mfrow = c(1, 2), mar = c(5.1, 4.6, 1.1, 1.1))
plot(R.sq[, 1], R.sq[, 3], pch = 16, cex = 0.65, col = "blue",
     xlab = expression(paste("Model 1: ", R^2)),
     ylab = expression(paste("Model 2: ", R^2)))
abline(0, 1)

plot(R.sq[, 2], R.sq[, 4], pch = 16, cex = 0.5, col = "blue",
     xlab = expression(paste("Model 1: ", R[adj]^2)),
     ylab = expression(paste("Model 2: ", R[adj]^2)))
abline(0, 1)
```





```r
boxplot(R.sq[, 3] - R.sq[, 1], R.sq[, 4] - R.sq[, 2], las = 1,
        col = "gray", boxwex = 0.5, staplewex = 0.25,
        outwex = 0.25, lwd = 0.5, outcex = 0.5)
abline(h = 0, lty = 2, col = "red")
axis(1, at = 1:2, labels = c("R^2", "R^2_adj"))

GLM_p <- numeric(500)

for (i in 1:500){
  reduce <- lm(y1[, i] ~ x1[, i])
  full <- lm(y1[, i] ~ x1[, i] + x2[, i])
  GLM_p[i] <- anova(reduce, full)$`Pr(>F)`[2]
}

hist(GLM_p, 30, las = 1)
```





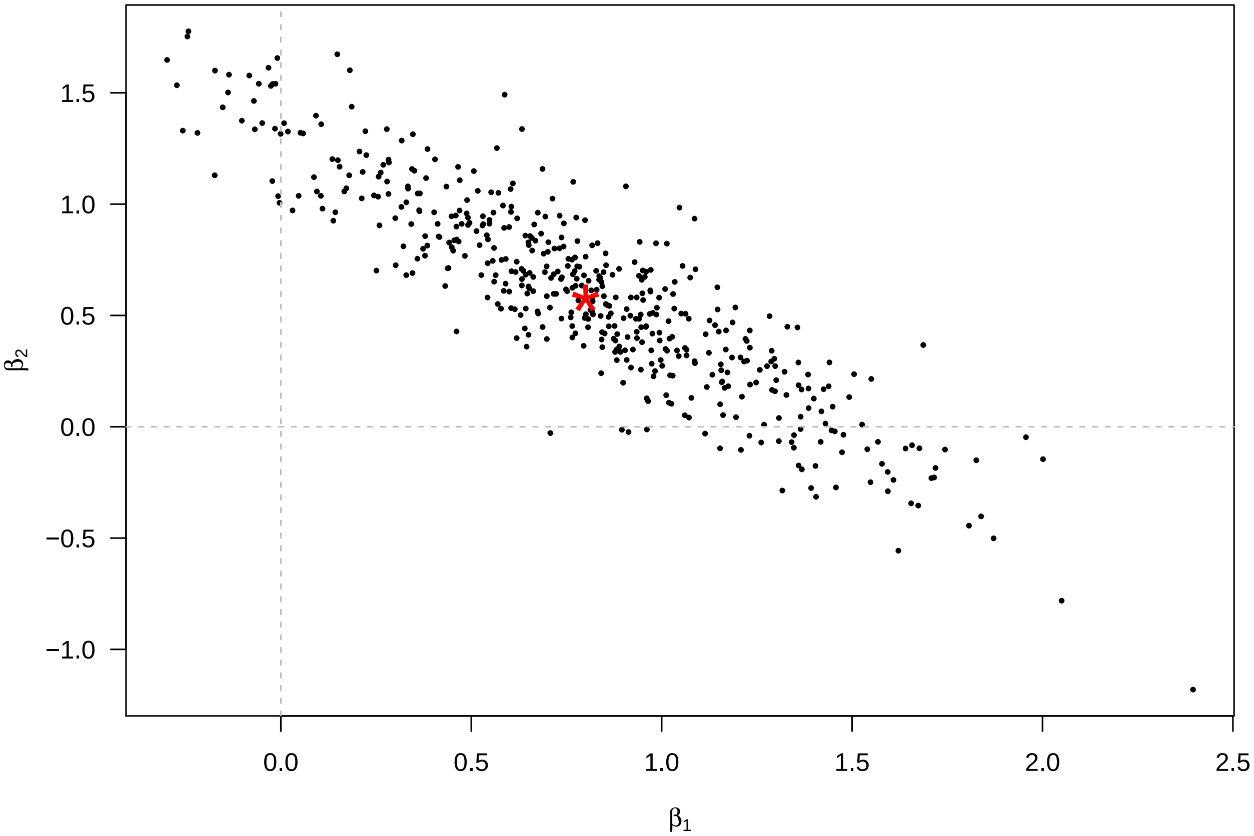
Multicollinearity

```
library(MASS)

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)
```

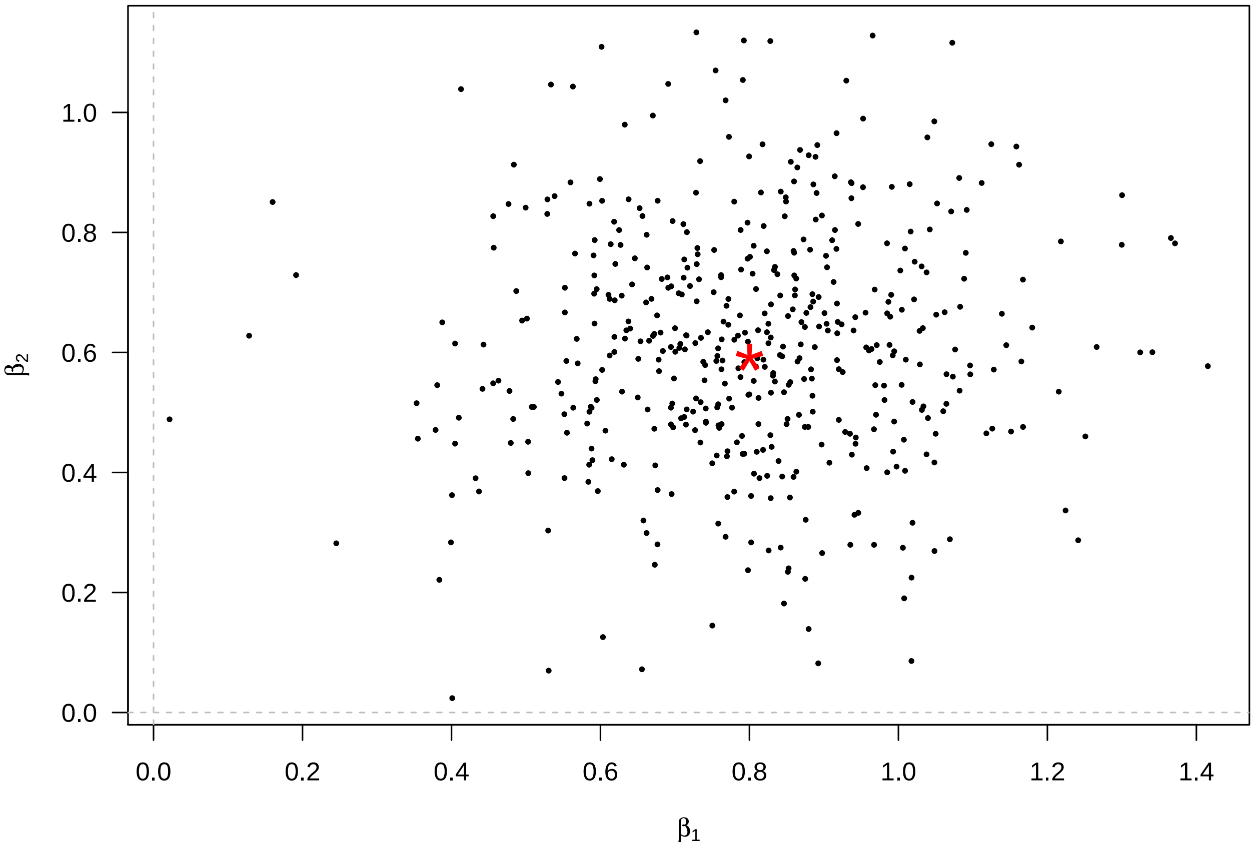
```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
## 0.2892 0.5918 0.6735 0.6610 0.7329 0.8783
```

```
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")
```



```
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")
```

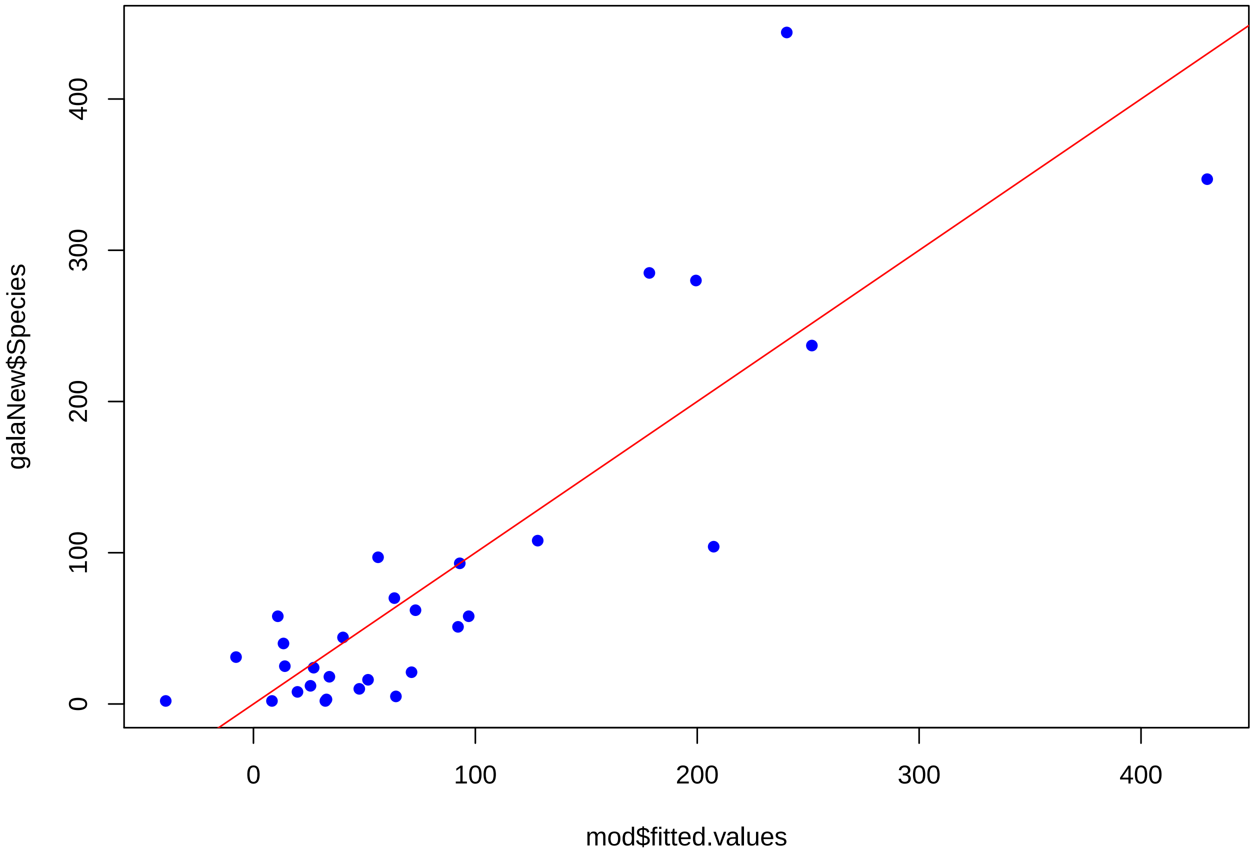


```
R.sq_M2 <- numeric(N)
for (i in 1:N){
  R.sq_M2[i] <- summary(lm(y1[, i] ~ x1[, 1, i] + x1[, 2, i]))$r.squared
}
summary(R.sq_M2)
```

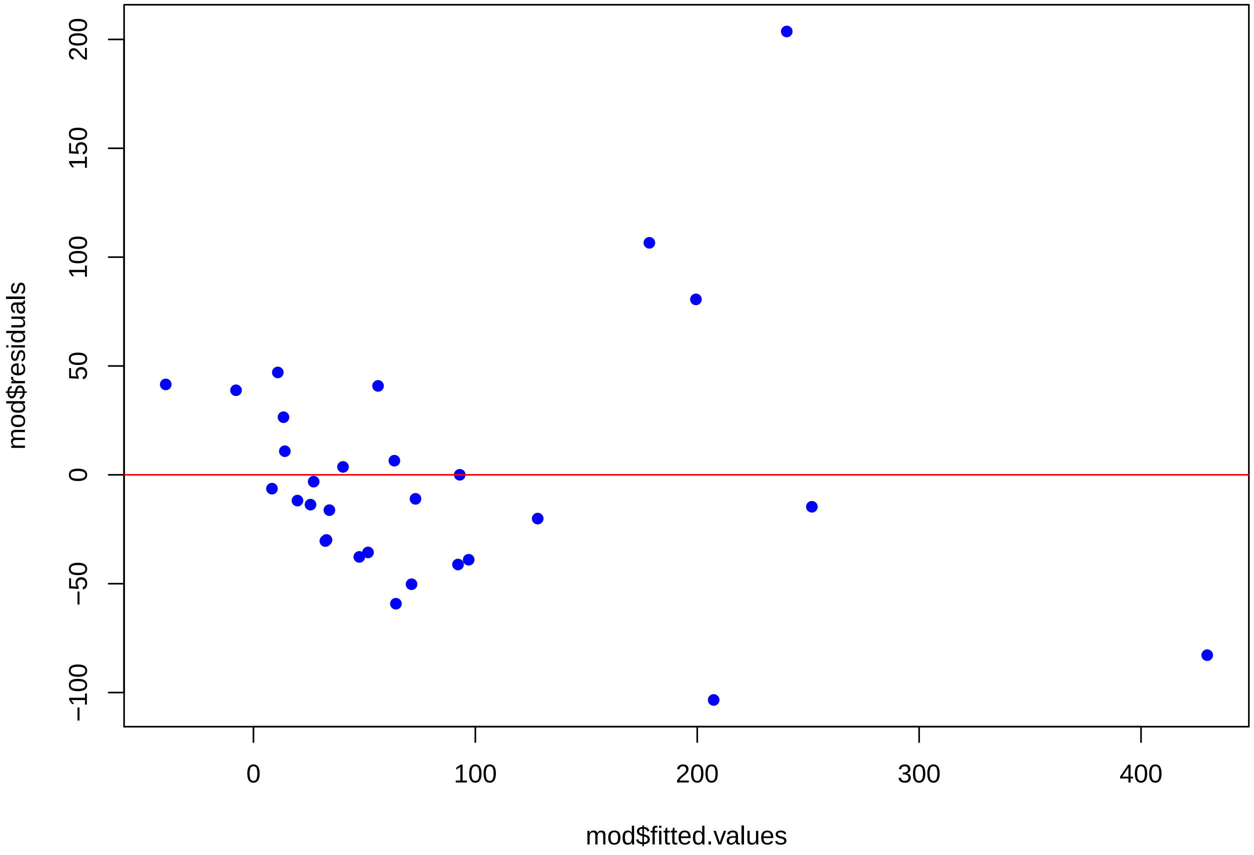
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06369 0.43340 0.52510 0.51762 0.61201 0.82718
```

Model Diagnostics

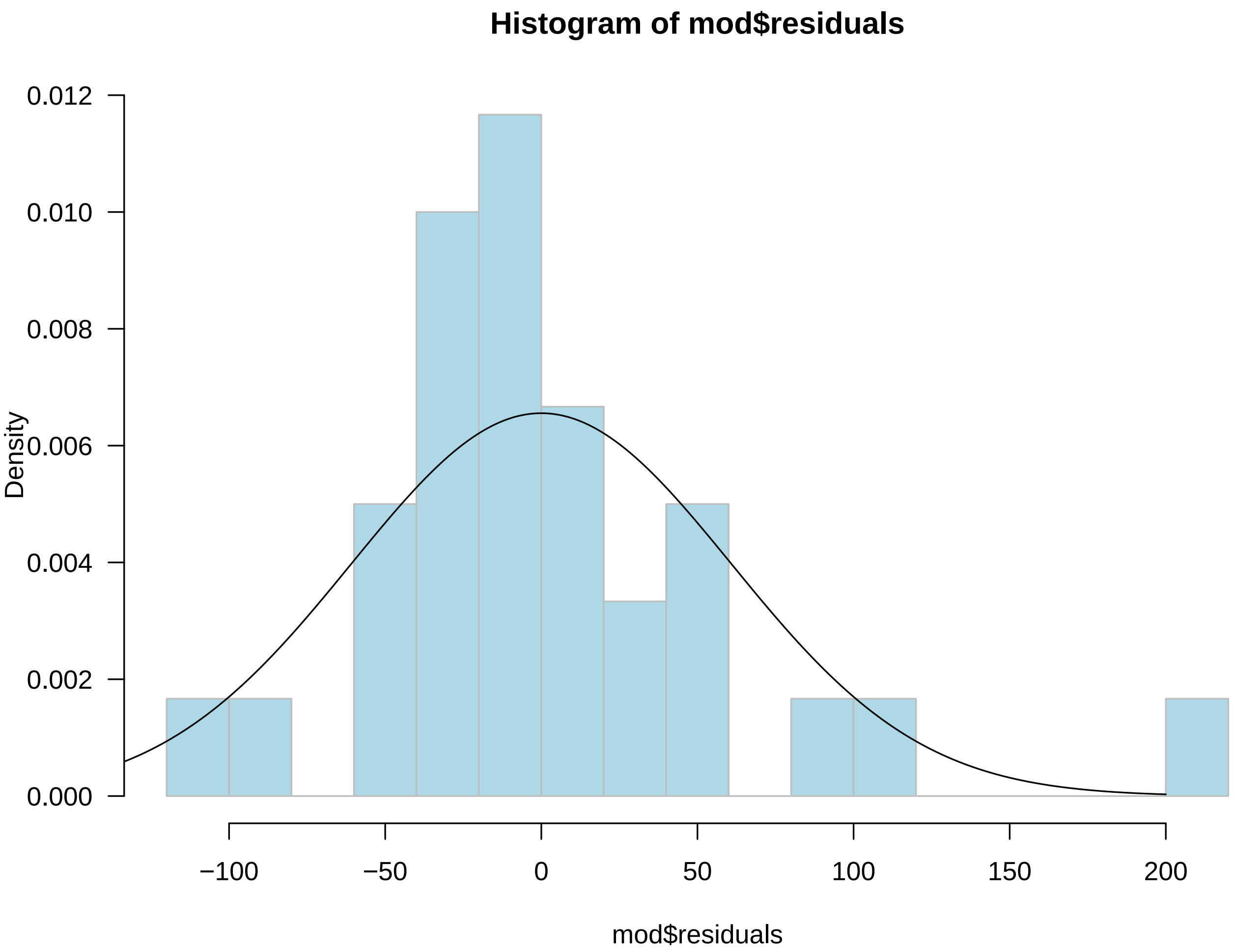
```
mod <- lm(Species ~ Elevation + Adjacent, data = galaNew)
plot(mod$fitted.values, galaNew$Species, pch = 16, col = "blue")
abline(0, 1, col = "red")
```



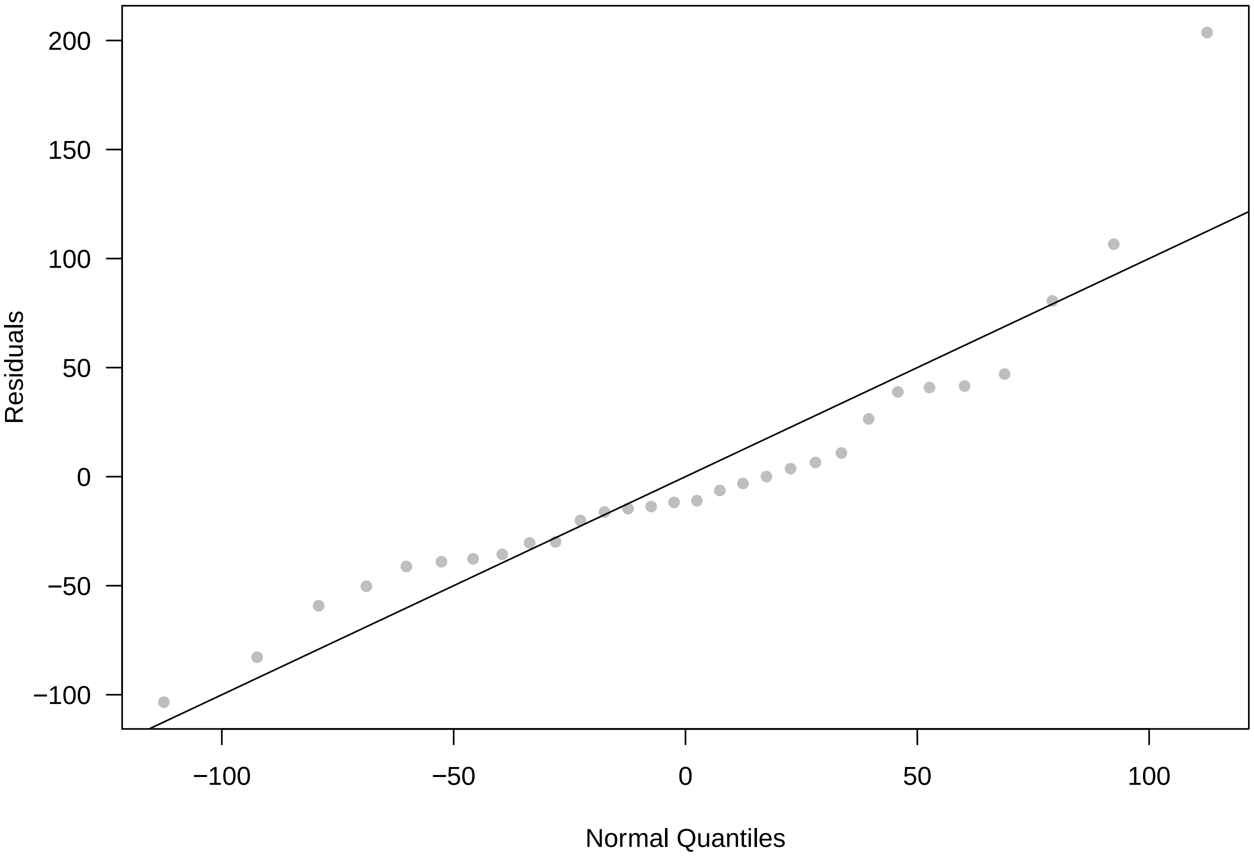
```
plot(mod$fitted.values, mod$residuals, pch = 16, col = "blue")
abline(h = 0, col = "red")
```



```
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)
```



Model Selection

```r
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 dplyr::select() masks MASS::select()
```

```r
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 ~ Area + Elevation + Nearest + Scruz + Adjacent, data = gala, nvmax = 5)
summary(models)
```

```
## Subset selection object
## Call: regsubsets.formula(Species ~ Area + Elevation + Nearest + Scruz +
##     Adjacent, data = gala, 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)
data.frame(
  Adj.R2 = which.max(res.sum$adjr2),
  CP = which.min(res.sum$cp),
  BIC = which.min(res.sum$bic)
)
```

```
##   Adj.R2 CP BIC
## 1      4  2   2
```

```
full <- lm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
           data = gala)
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 = gala)
##
## Coefficients:
## (Intercept)    Elevation     Adjacent
##     1.43287      0.27657     -0.06889
```