DSA 8020 R Session 4: Multiple Linear Regression III

Whitney

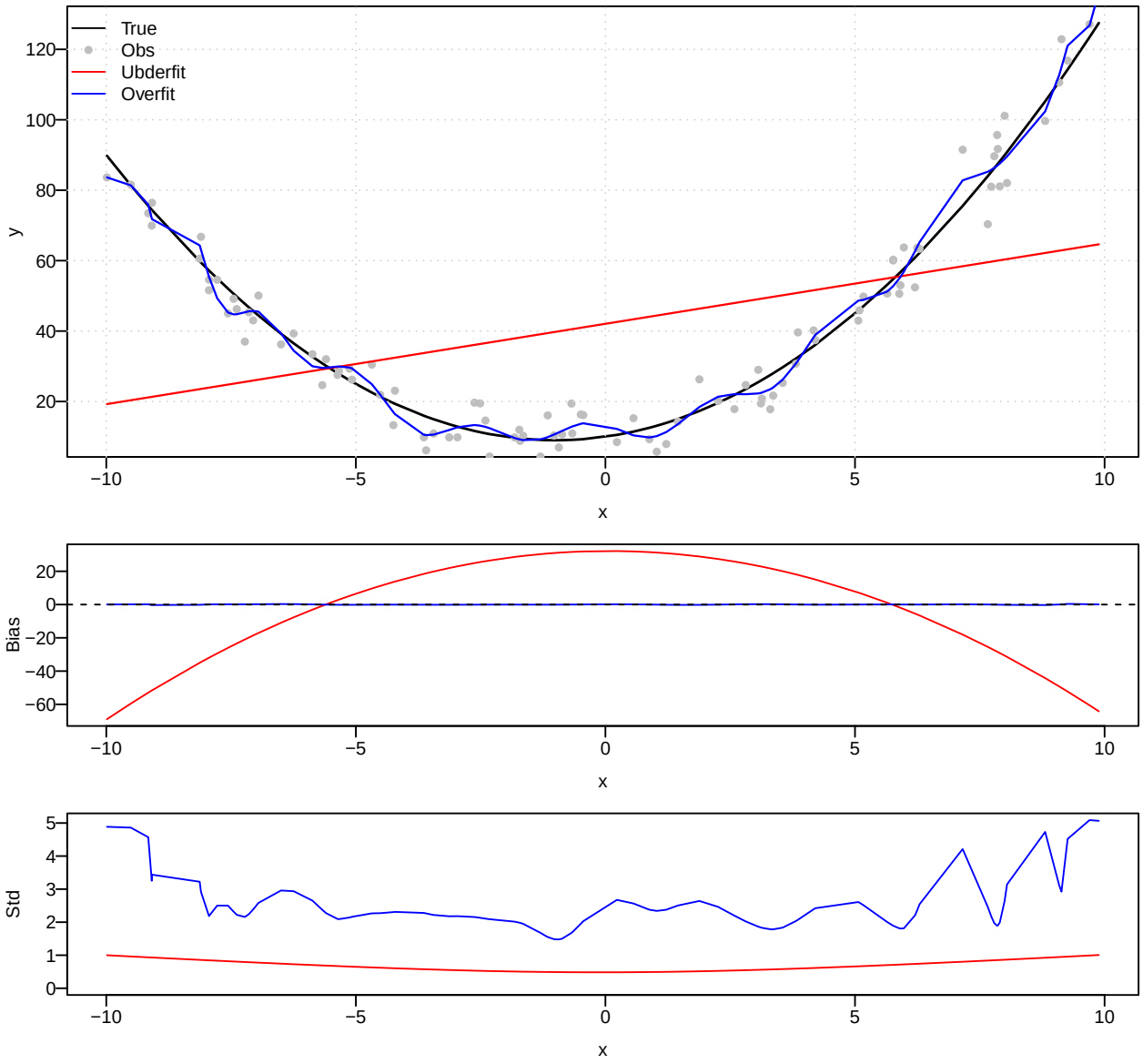
January 29, 2023

Contents

Bias And Variance

```
set.seed(123)
x <- sort(-10 + 20 * runif(100))
z <- 10 + 2 * x + x^2
set.seed(1234)
err <- replicate(500, rnorm(100, 0, 5))
y <- z + err

layout(matrix(c(1,1,2,3), 4, 1, byrow = TRUE))
par(las = 1, mar = c(3.5, 3.5, 0.5, 0.5), mgp = c(2, 0.5, 0))
plot(x, z, type = "l", ylab = "y", lwd = 1.5)
points(x, y[, 5], pch = 16, col = "gray")
grid()
```

```r
lm <- apply(y, 2, function(z) lm(z ~ x))
lines(x, lm[[5]]$fitted.values, col = "red", lwd = 1.2)
qm <- apply(y, 2, function(z) lm(z ~ poly(x, 25)))
lines(x, qm[[5]]$fitted.values, col = "blue", lwd = 1.2)
legend("topleft", legend = c("True", "Obs", "Ubderfit", "Overfit"),
       col = c("black", "gray", "red", "blue"),
       lty = c(1, NA, 1, 1), pch = c(NA, 16, NA, NA),
       bty = "n")

lmpred <- array(unlist(lapply(lm, predict)), dim = c(100, 500))
qmpred <- array(unlist(lapply(qm, predict)), dim = c(100, 500))
bias_lm <- apply(lmpred, 1, mean) - z
bias_qm <- apply(qmpred, 1, mean) - z
plot(x, bias_lm, col = "red", type = "l", ylab = "Bias")
lines(x, bias_qm, col = "blue")
abline(h = 0, lty = 2)
var_lm <- apply(lmpred, 1, var)
var_qm <- apply(qmpred, 1, var)
plot(x, sqrt(var_qm), col = "blue", ylim = c(0, max(sqrt(var_qm))), type = "l",
     ylab = "Std")
lines(x, sqrt(var_lm), col = "red")
```



Model Selection

load the data

```
library(faraway)
data(gala)
galaNew <- gala[, -2]
```

Best Subset Selection

```
#install.packages(c("tidyverse", "caret", "leaps"))
library(tidyverse)
```

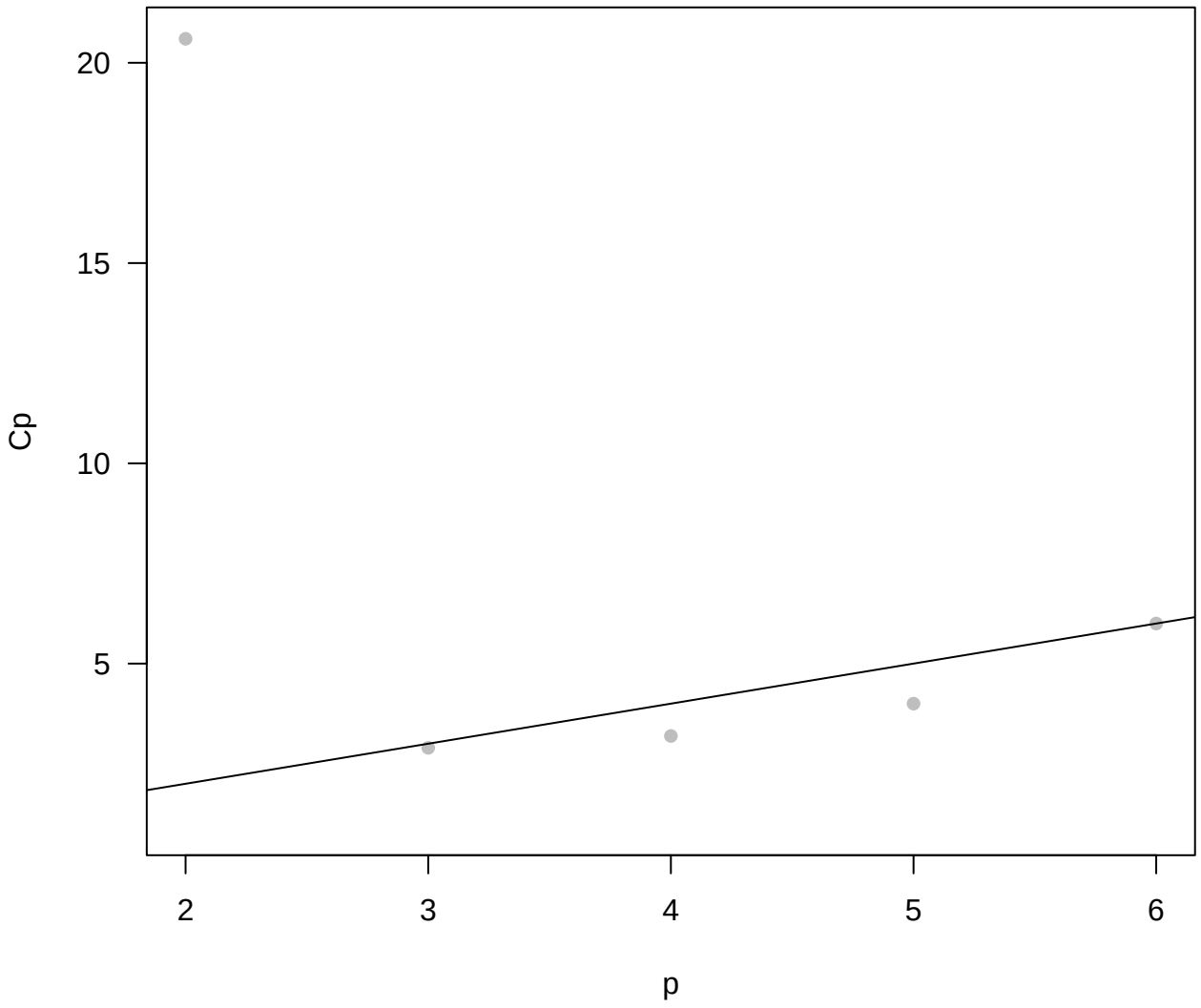
```
library(caret)
library(leaps)
models <- regsubsets(Species ~ ., data = galaNew)
summary(models)
```

```
## Subset selection object
## Call: regsubsets.formula(Species ~ ., data = galaNew)
## 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 ) "*"  "*"       "*"     "*"   "*"
```

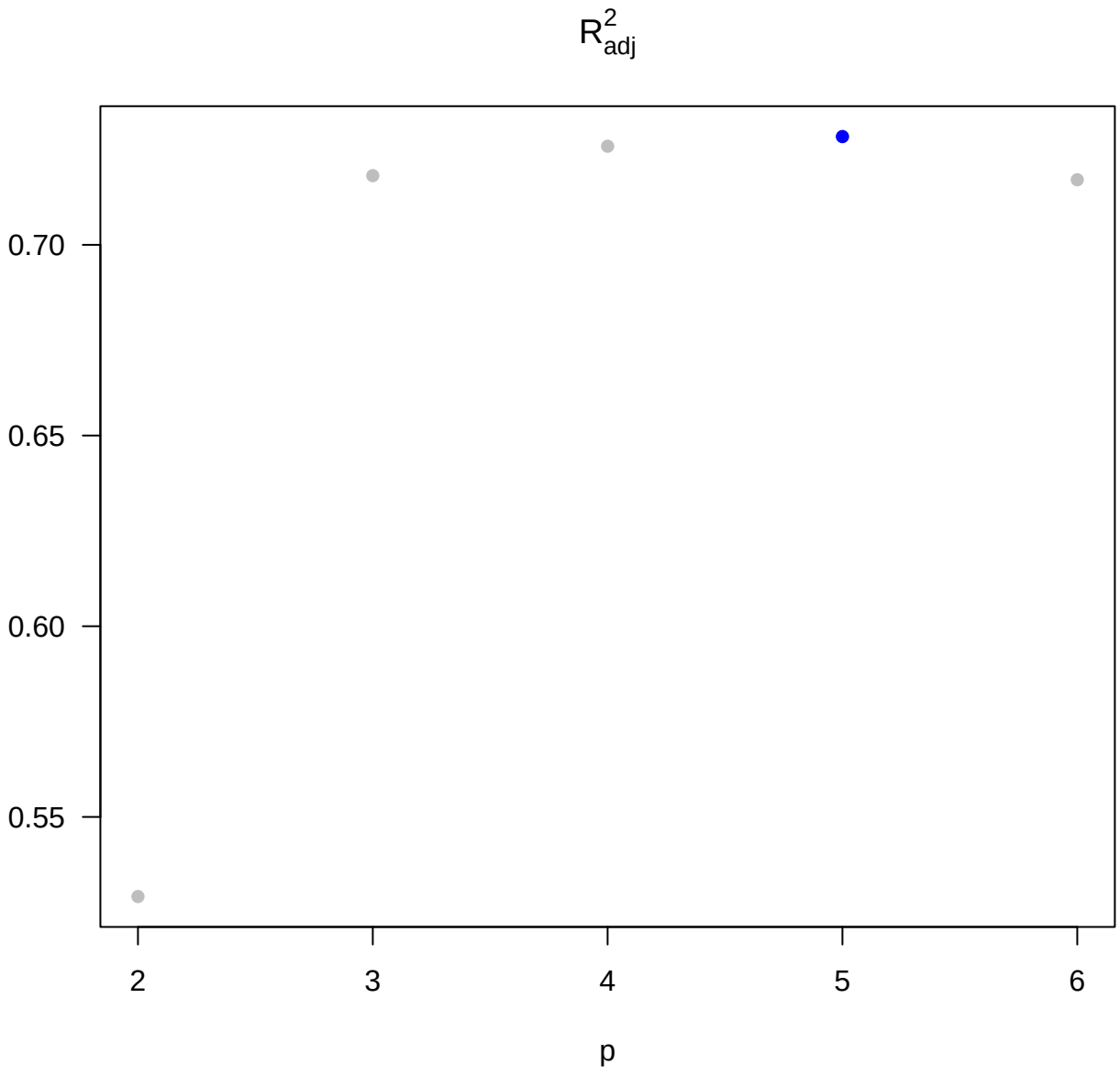
```
res.sum <- summary(models)

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
```

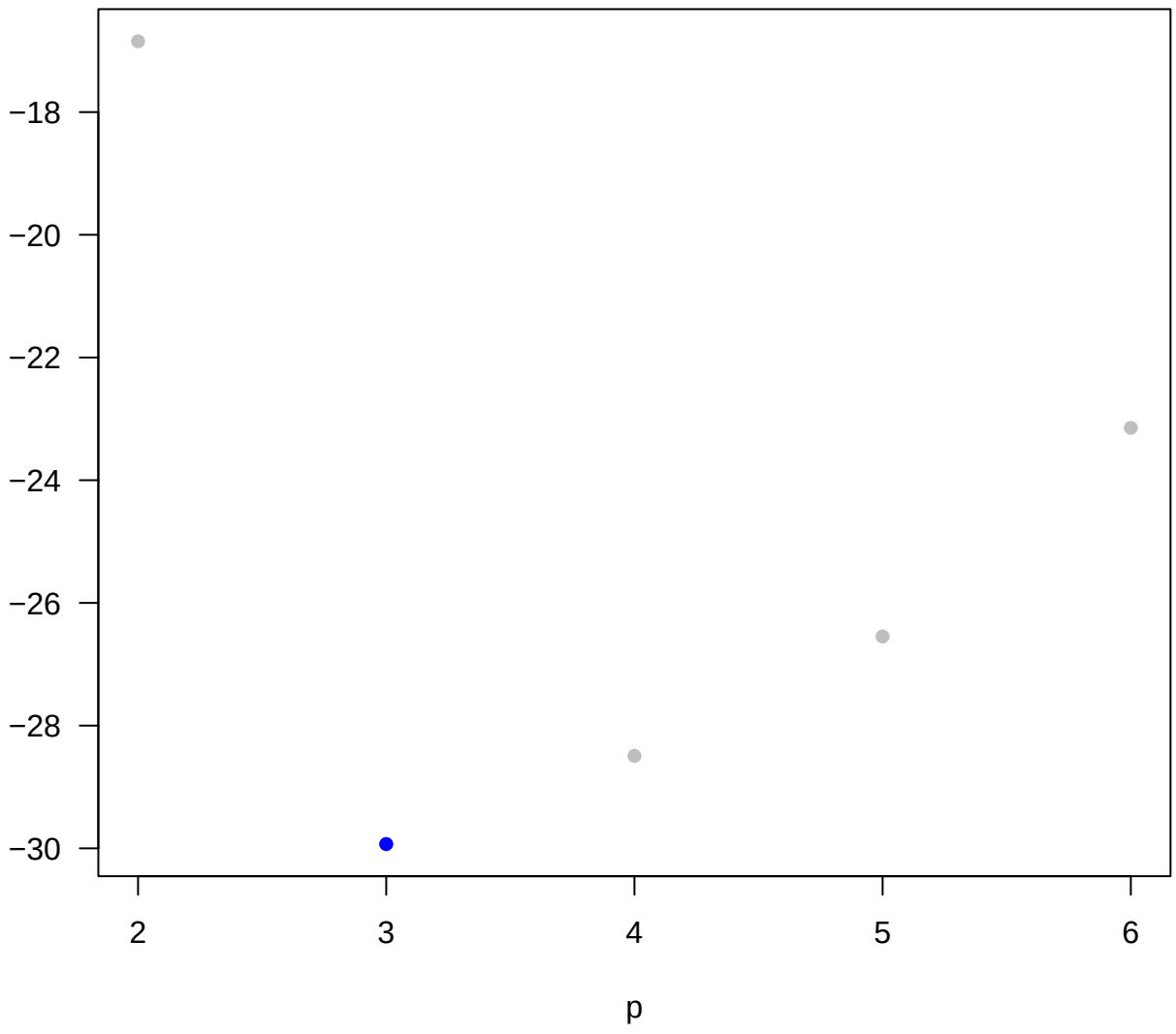
```
plot(2:6, criteria$Cp, las = 1, xlab = "p", ylab = "Cp",
     pch = 16, col = "gray", ylim = c(1, max(criteria$Cp)))
abline(0, 1)
```




```
plot(2:6, criteria$Adj.R2, las = 1, xlab = "p", ylab = "", pch = 16, col = "gray",
     main = expression(R['adj']^2))
points(5, criteria$Adj.R2[4], col = "blue", pch = 16)
```




```
plot(2:6, criteria$BIC, las = 1, xlab = "p", ylab = "", pch = 16, col = "gray", main = "BIC")
points(3, criteria$BIC[2], col = "blue", pch = 16)
```



Backward Selection

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

Stepwise Selection

```
step(full, direction = "both")
```

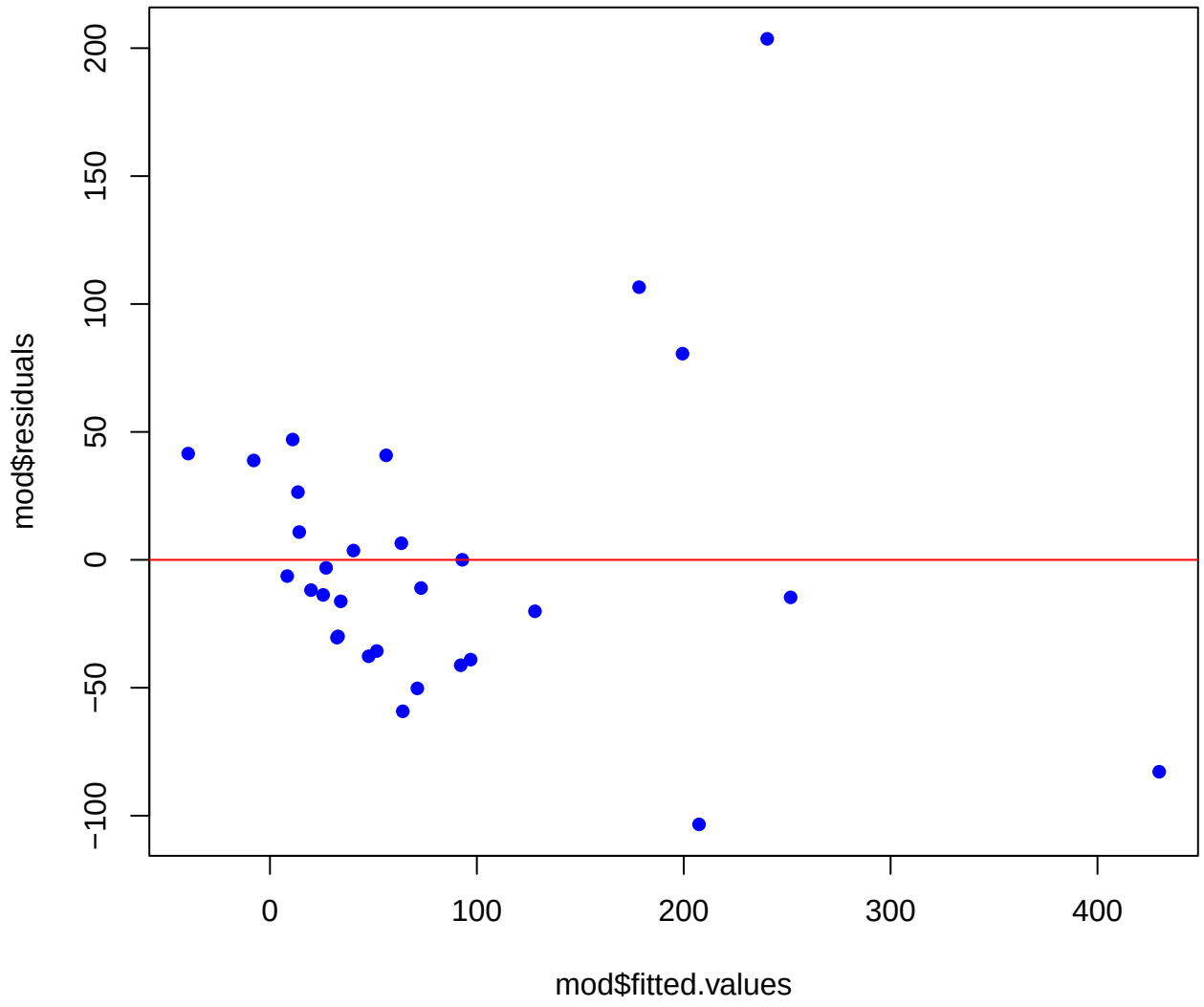
```
## Start:  AIC=251.93
## Species ~ Area + Elevation + Nearest + Scruz + Adjacent
##
##             Df Sum of Sq    RSS    AIC
## - Nearest    1         0  89232 249.93
## - Area       1      4238  93469 251.33
## - Scruz      1      4636  93867 251.45
## <none>                    89231 251.93
## - 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
## + Nearest    1         0  89231 251.93
## - 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
## + Area       1      4436  89232 249.93
## + Nearest    1       198  93469 251.33
## - 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
## + Scruz      1      6336  93667 249.39
## + Area       1      3227  96776 250.37
## + Nearest    1      1550  98453 250.88
## - 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

Residual Plot

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

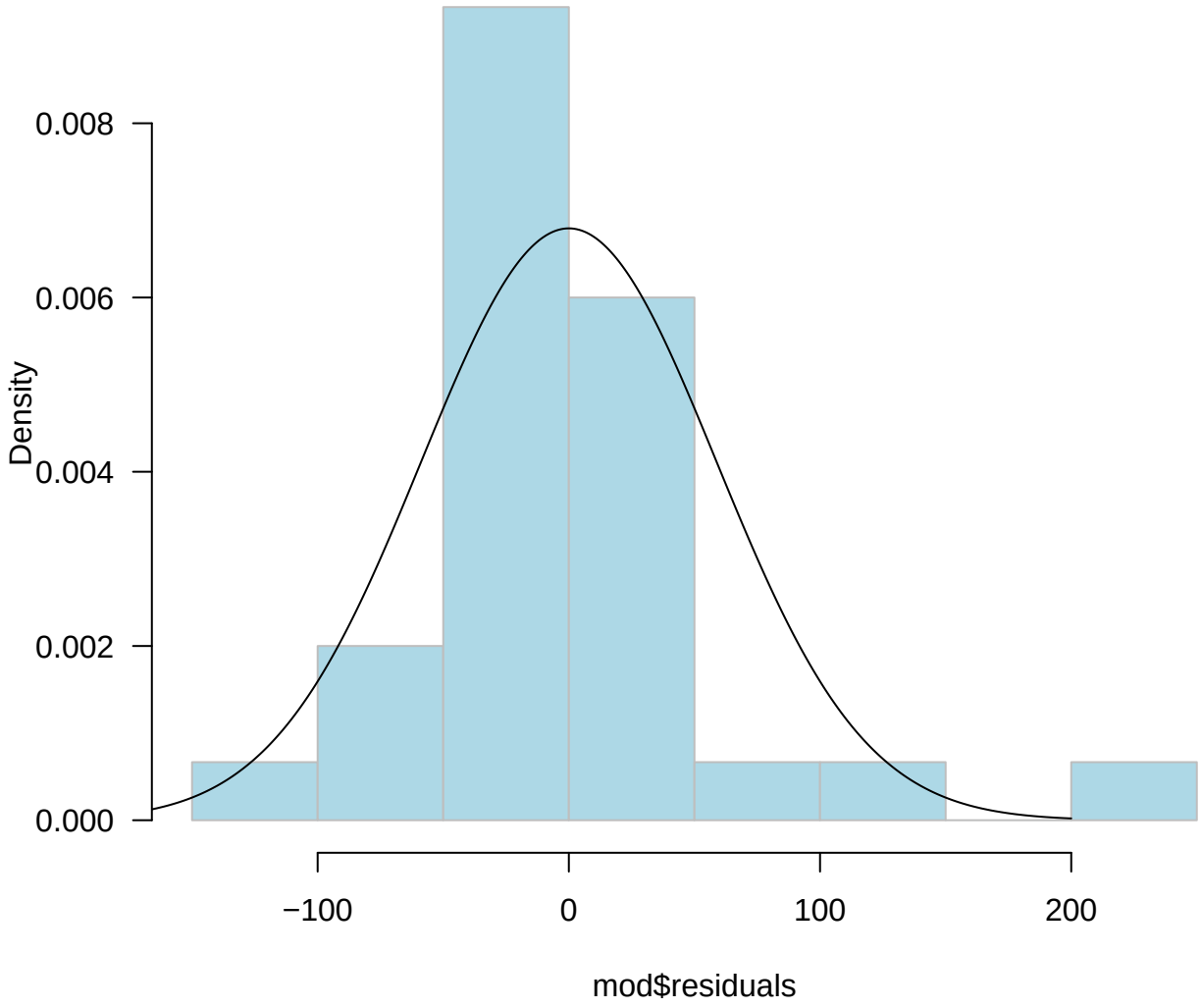


Residual Histogram/QQplot

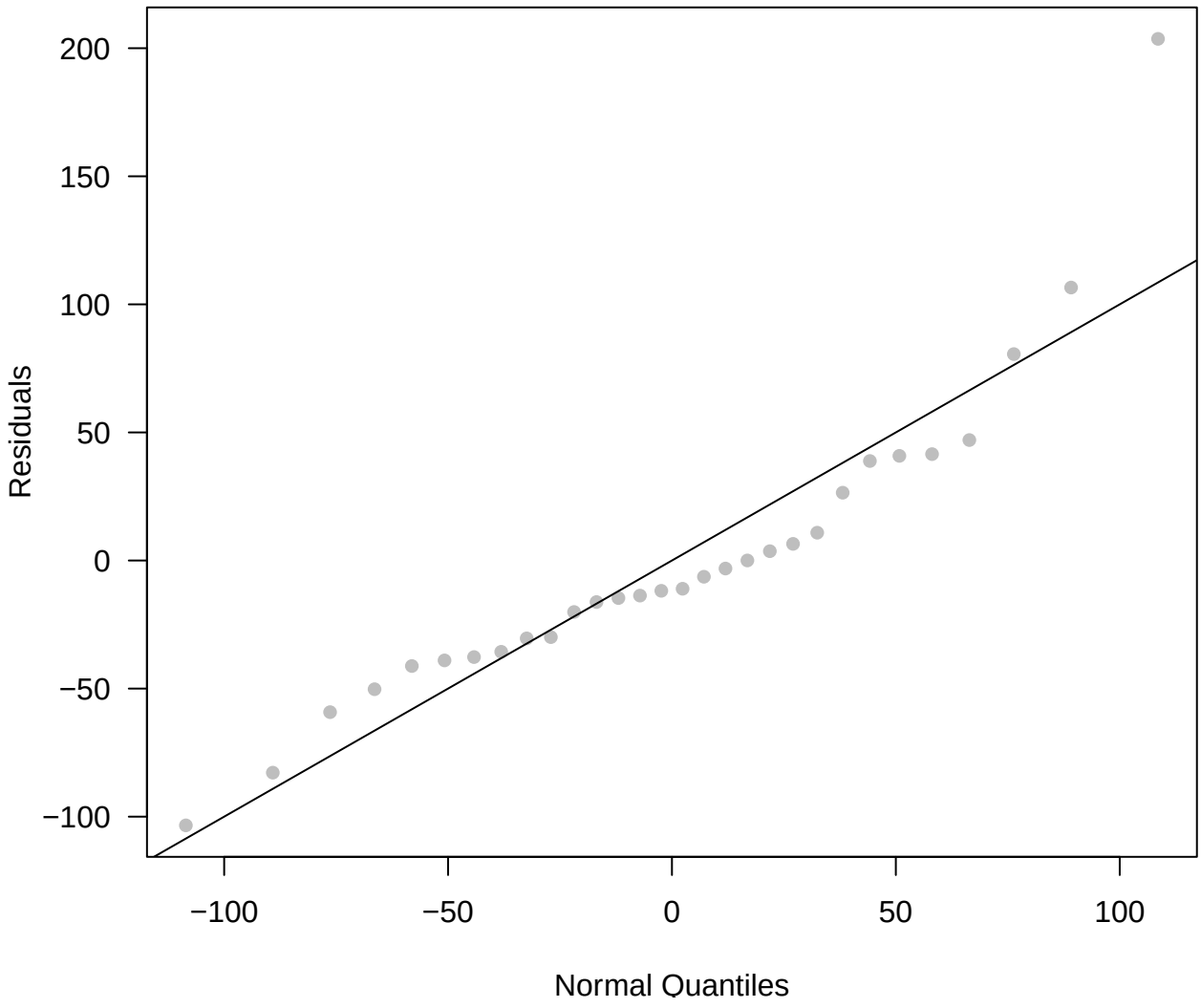
```
(sd <- sd(mod$residuals))
```

```
## [1] 58.72291
```

```
par(las = 1)
hist(mod$residuals, 5, prob = T, col = "lightblue", border = "gray")
xg <- seq(-200, 200, 1)
yg <- dnorm(xg, 0, sd)
lines(xg, yg)
```

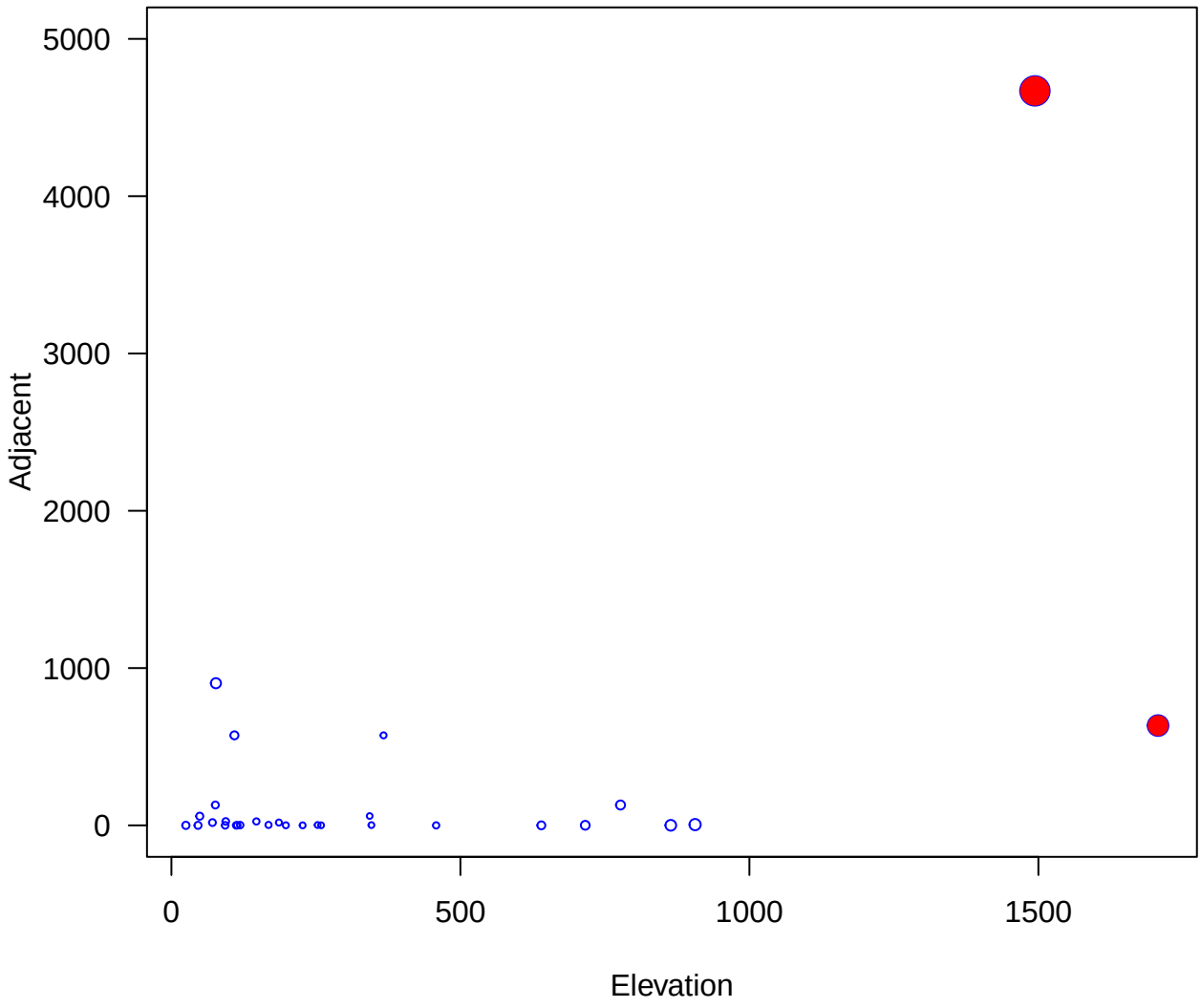



```
plot(qnorm(1:30 / 31, 0, sd), sort(mod$residuals), pch = 16,
     col = "gray", xlab = "Normal Quantiles", ylab = "Residuals")
abline(0, 1)
```



Leverage

```
step_gala <- step(full, trace = F)
X <- model.matrix(step_gala)
H <- X %*% solve((t(X) %*% X)) %*% t(X)
lev <- hat(X)
high_lev <- which(lev >= 2 * 3 / 30)
attach(gala)
par(las = 1)
plot(Elevation, Adjacent, cex = sqrt(5 * lev), col = "blue", ylim = c(0, 5000))
points(Elevation[high_lev], Adjacent[high_lev], col = "red", pch = 16,
       cex = sqrt(5 *lev[high_lev]))
```

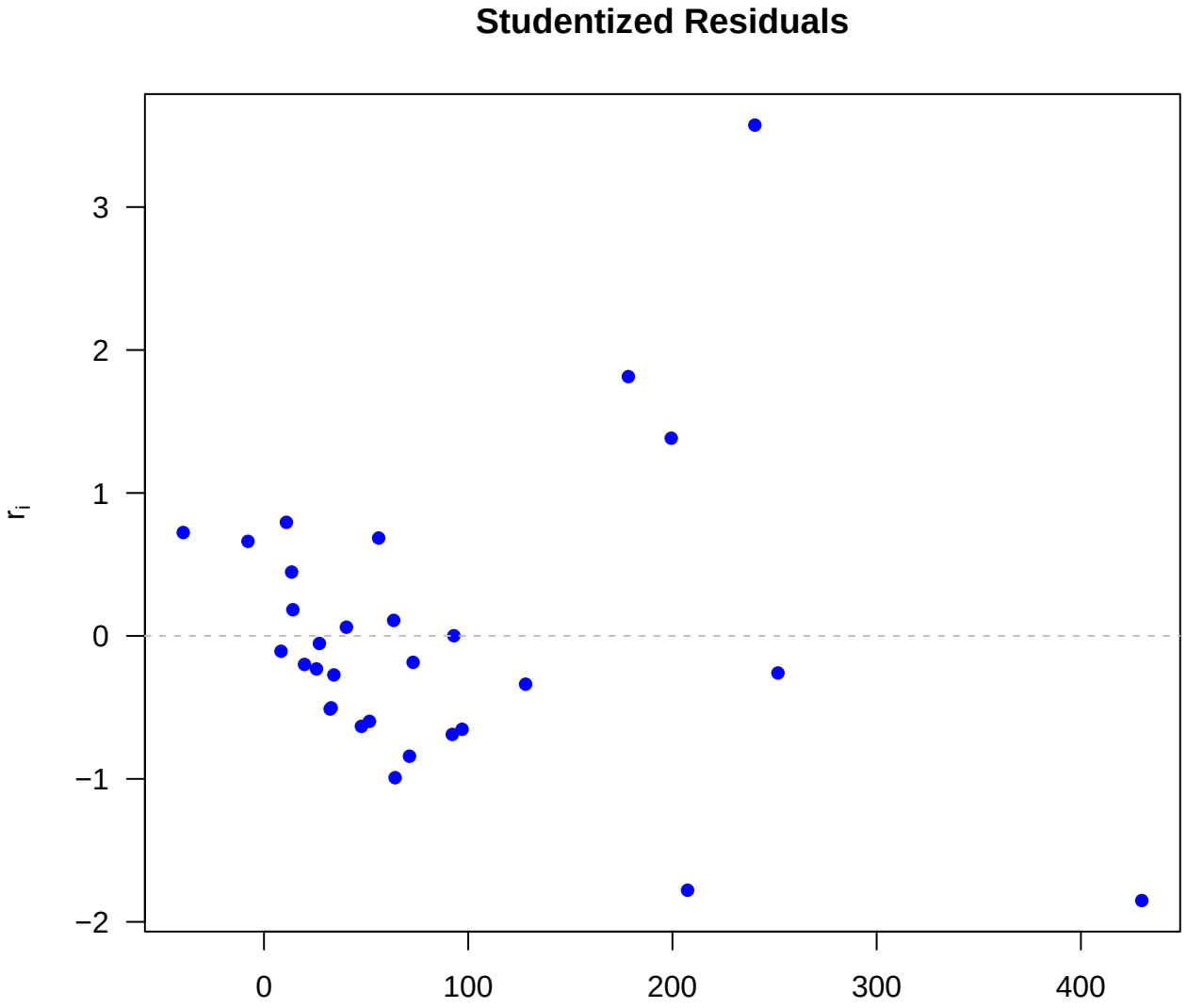


Studentized Residuals

```
gs <- summary(step_gala)
gs$sig
```

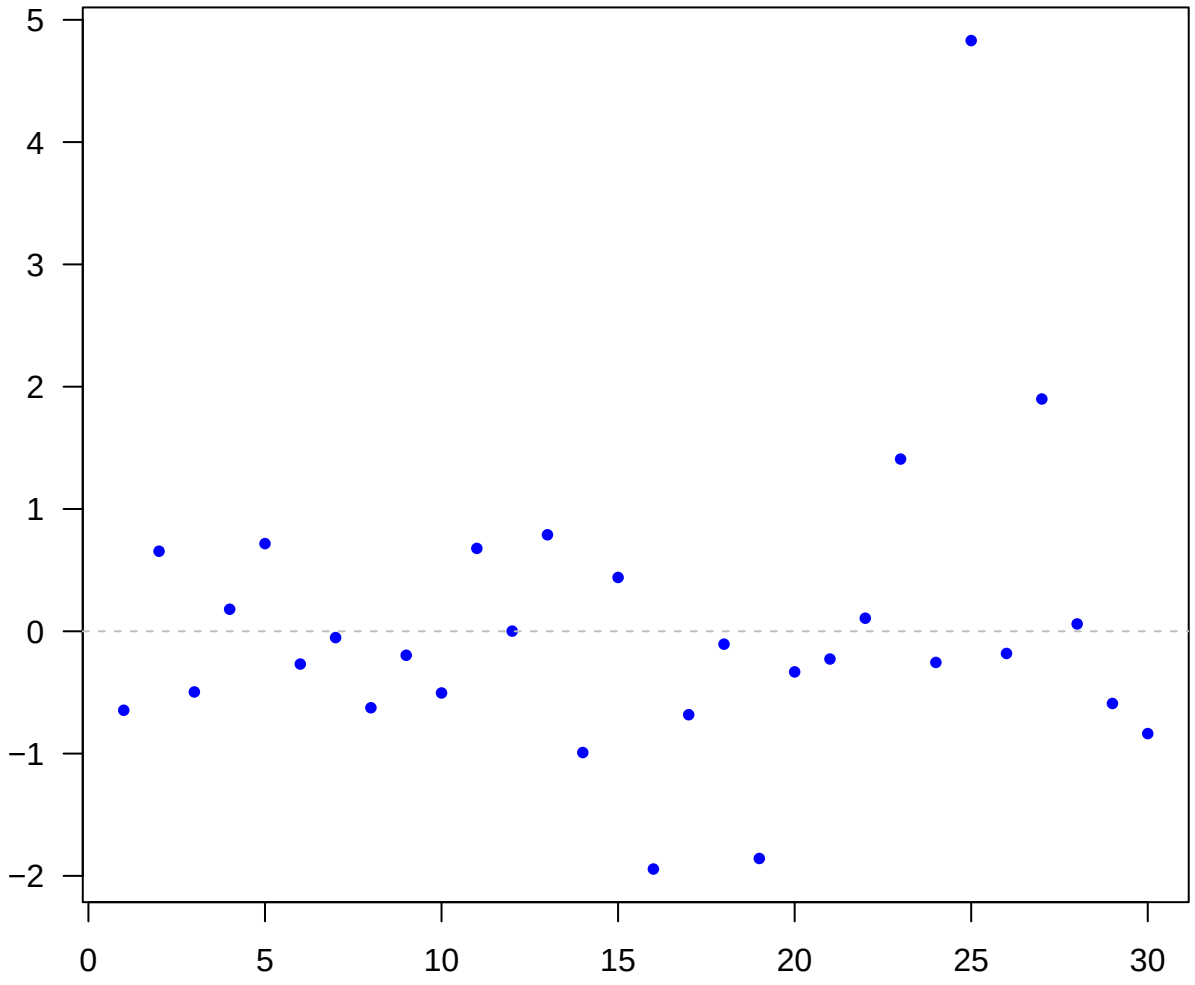
```
## [1] 60.85898
```

```
studRes <- gs$res / (gs$sig * sqrt(1 - lev))

par(las = 1)
plot(step_gala$fitted.values, studRes, pch = 16, col = "blue",
     ylab = expression(r[i]), main = "Studentized Residuals", xlab = "")
abline(h = 0, lty = 2, col = "gray")
```



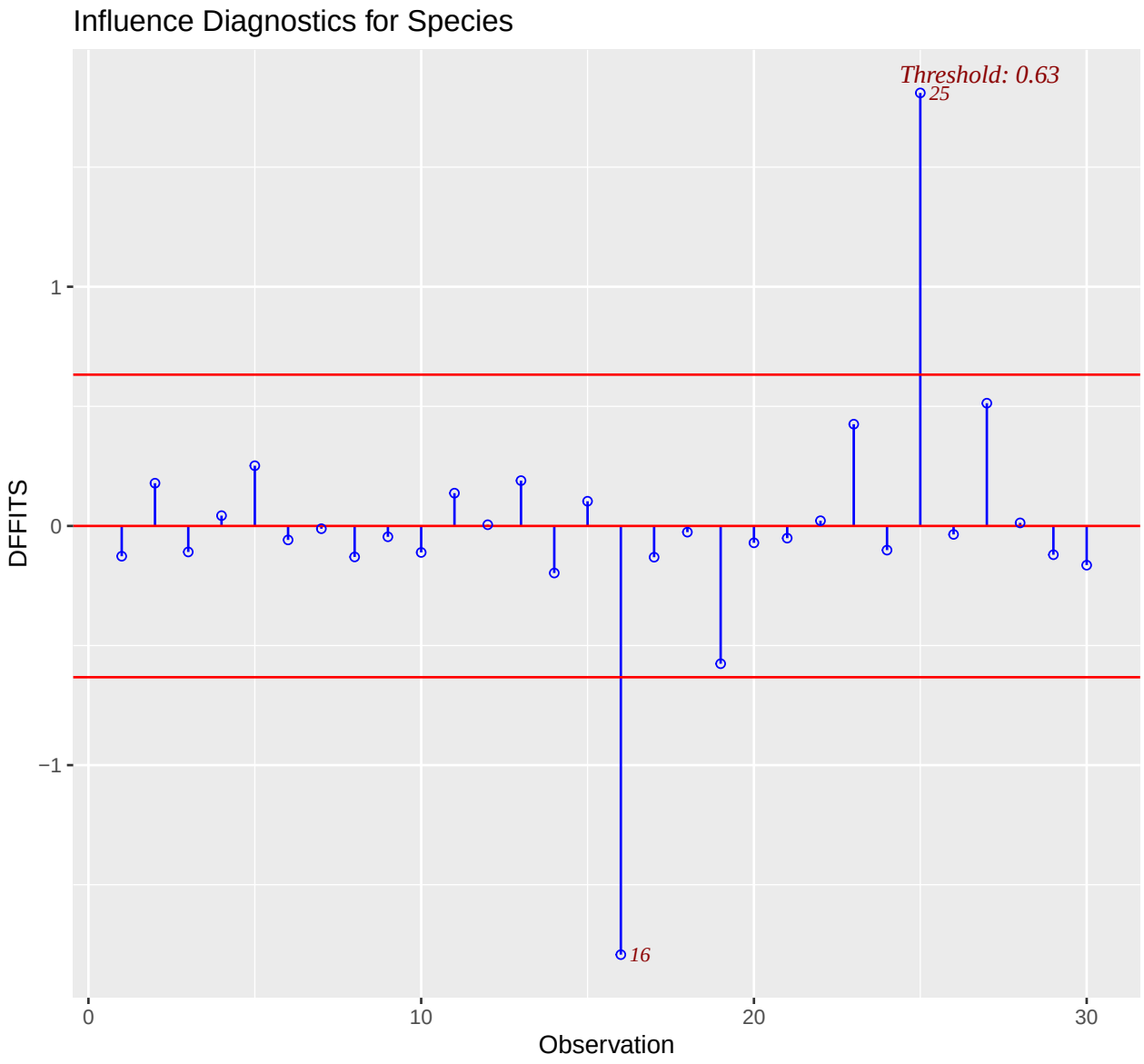
Jackknife Residuals

```
jack <- rstudent(step_gala)

par(las = 1)
plot(jack, pch = 16, cex = 0.8, col = "blue", main =" Jacknife Residuals ",
     xlab = "", ylab = "")
abline(h = 0, lty = 2, col = "gray")
```



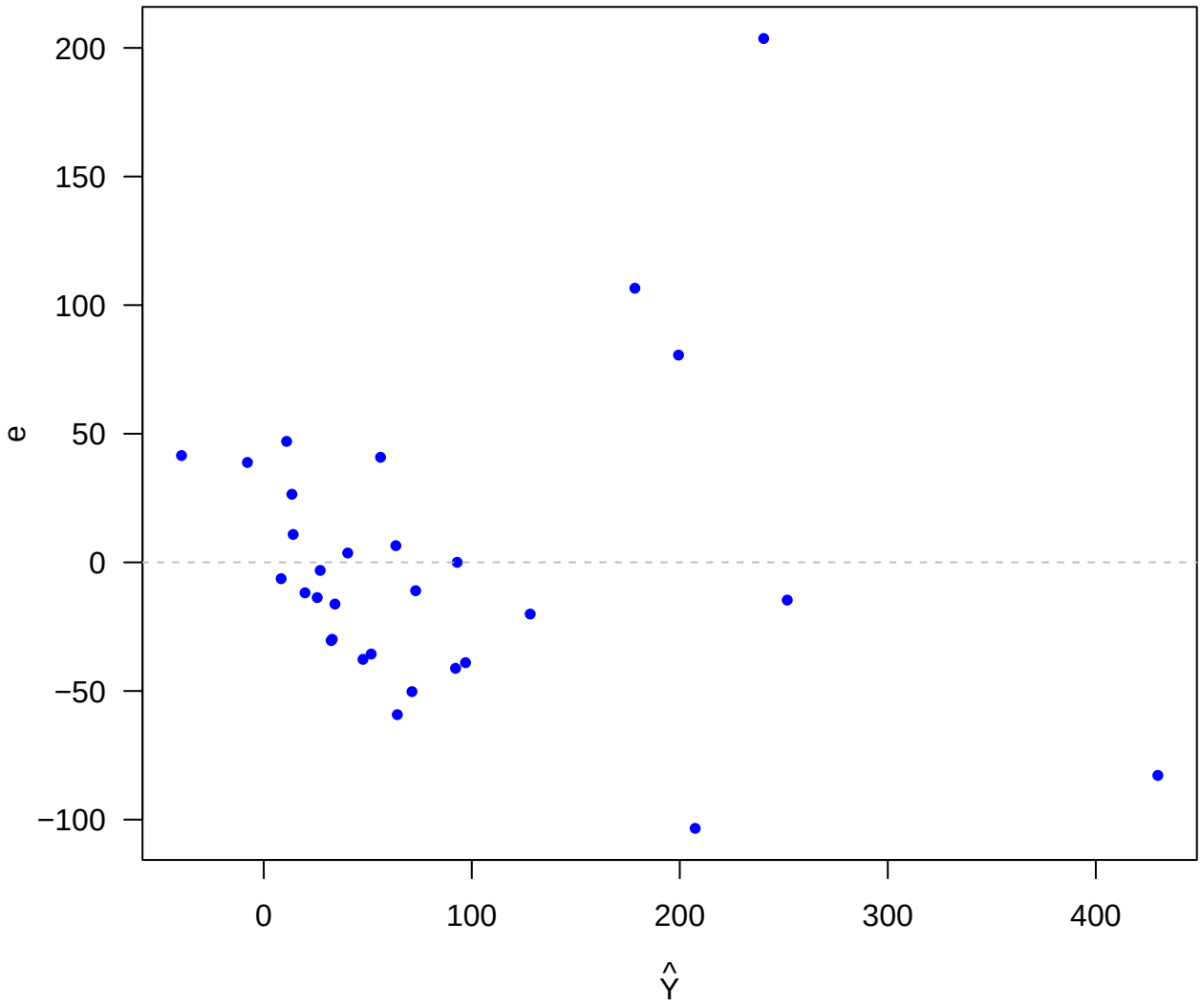
Identifying Influential Observations: DFFITS

```
library(olsrr)
ols_plot_dffits(step_gala)
```

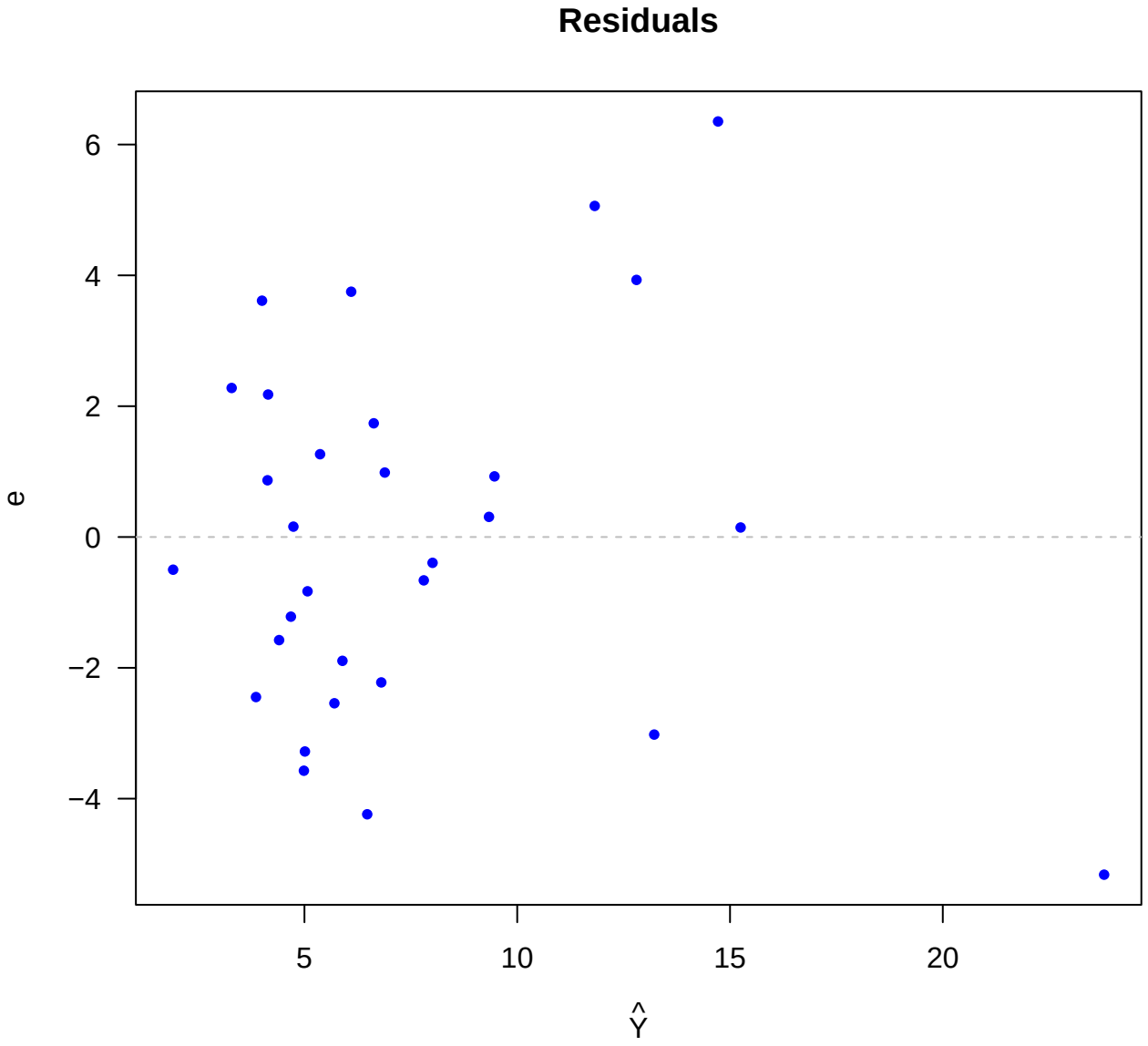


Transformation

```
par(las = 1)
plot(step_gala$fitted.values, step_gala$residuals,
     pch = 16, cex = 0.8, col = "blue", main =" Residuals ",
     xlab = expression(hat(Y)), ylab = expression(e))
abline(h = 0, lty = 2, col = "gray")
```




```
sqrt_fit <- lm(sqrt(Species) ~ Elevation + Adjacent)

par(las = 1)
plot(sqrt_fit$fitted.values, sqrt_fit$residuals,
     pch = 16, cex = 0.8, col = "blue", main =" Residuals ",
     xlab = expression(hat(Y)), ylab = expression(e))
abline(h = 0, lty = 2, col = "gray")
```



Box-Cox Transformation

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:olsrr':
##
##     cement
```

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

```r
boxcox <- boxcox(step_gala, plotit = T,
                 lambda = seq(-0.25, 0.75, by = 0.05))
```