DSA 8020 R Lab 4: Model Selection and Model Checking

Whitney

February 16, 2021

Contents

Savings rates in 50 countries **1**

Savings rates in 50 countries

The savings data frame has 50 rows (countries) and 5 columns (variables):

1. `sr`: savings rate - personal saving divided by disposable income *This variable will be used as the response*
2. `pop15`: percent population under age of 15
3. `pop75`: percent population over age of 75
4. `dpi`: per-capita disposable income in dollars
5. `ddpi`: percent growth rate of dpi

The data is averaged over the period 1960-1970.

Data Source: Belsley, D., Kuh. E. and Welsch, R. (1980) *Regression Diagnostics* Wiley.

Load the dataset

Code:

```
data(savings, package = "faraway")
head(savings)
```

```
##              sr pop15 pop75     dpi ddpi
## Australia 11.43 29.35  2.87 2329.68 2.87
## Austria   12.07 23.32  4.41 1507.99 3.93
## Belgium   13.17 23.80  4.43 2108.47 3.82
## Bolivia    5.75 41.89  1.67  189.13 0.22
## Brazil    12.88 42.19  0.83  728.47 4.56
## Canada     8.79 31.72  2.85 2982.88 2.43
```

```
savings
```

```
##                   sr pop15 pop75     dpi  ddpi
## Australia      11.43 29.35  2.87 2329.68  2.87
## Austria        12.07 23.32  4.41 1507.99  3.93
```

```
## Belgium        13.17 23.80  4.43 2108.47  3.82
## Bolivia         5.75 41.89  1.67  189.13  0.22
## Brazil         12.88 42.19  0.83  728.47  4.56
## Canada          8.79 31.72  2.85 2982.88  2.43
## Chile           0.60 39.74  1.34  662.86  2.67
## China          11.90 44.75  0.67  289.52  6.51
## Colombia        4.98 46.64  1.06  276.65  3.08
## Costa Rica     10.78 47.64  1.14  471.24  2.80
## Denmark        16.85 24.42  3.93 2496.53  3.99
## Ecuador         3.59 46.31  1.19  287.77  2.19
## Finland        11.24 27.84  2.37 1681.25  4.32
## France         12.64 25.06  4.70 2213.82  4.52
## Germany        12.55 23.31  3.35 2457.12  3.44
## Greece         10.67 25.62  3.10  870.85  6.28
## Guatamala       3.01 46.05  0.87  289.71  1.48
## Honduras        7.70 47.32  0.58  232.44  3.19
## Iceland         1.27 34.03  3.08 1900.10  1.12
## India           9.00 41.31  0.96   88.94  1.54
## Ireland        11.34 31.16  4.19 1139.95  2.99
## Italy          14.28 24.52  3.48 1390.00  3.54
## Japan          21.10 27.01  1.91 1257.28  8.21
## Korea           3.98 41.74  0.91  207.68  5.81
## Luxembourg     10.35 21.80  3.73 2449.39  1.57
## Malta          15.48 32.54  2.47  601.05  8.12
## Norway         10.25 25.95  3.67 2231.03  3.62
## Netherlands    14.65 24.71  3.25 1740.70  7.66
## New Zealand    10.67 32.61  3.17 1487.52  1.76
## Nicaragua       7.30 45.04  1.21  325.54  2.48
## Panama          4.44 43.56  1.20  568.56  3.61
## Paraguay        2.02 41.18  1.05  220.56  1.03
## Peru           12.70 44.19  1.28  400.06  0.67
## Philippines    12.78 46.26  1.12  152.01  2.00
## Portugal       12.49 28.96  2.85  579.51  7.48
## South Africa   11.14 31.94  2.28  651.11  2.19
## South Rhodesia 13.30 31.92  1.52  250.96  2.00
## Spain          11.77 27.74  2.87  768.79  4.35
## Sweden          6.86 21.44  4.54 3299.49  3.01
## Switzerland    14.13 23.49  3.73 2630.96  2.70
## Turkey          5.13 43.42  1.08  389.66  2.96
## Tunisia         2.81 46.12  1.21  249.87  1.13
## United Kingdom  7.81 23.27  4.46 1813.93  2.01
## United States   7.56 29.81  3.43 4001.89  2.45
## Venezuela       9.22 46.40  0.90  813.39  0.53
## Zambia         18.56 45.25  0.56  138.33  5.14
## Jamaica         7.72 41.12  1.73  380.47 10.23
## Uruguay         9.24 28.13  2.72  766.54  1.88
## Libya           8.89 43.69  2.07  123.58 16.71
## Malaysia        4.71 47.20  0.66  242.69  5.08
```

1. Perform the best subset selection and select the "best" model using R^2_{adj}

Code:

```r
library(tidyverse)
library(caret)
library(leaps)
models <- regsubsets(sr ~ ., data = savings)
(res.sum <- summary(models))
```

```
## Subset selection object
## Call: regsubsets.formula(sr ~ ., data = savings)
## 4 Variables  (and intercept)
##       Forced in Forced out
## pop15     FALSE      FALSE
## pop75     FALSE      FALSE
## dpi       FALSE      FALSE
## ddpi      FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: exhaustive
##          pop15 pop75 dpi ddpi
## 1  ( 1 ) "*"   " "   " " " "
## 2  ( 1 ) "*"   " "   " " "*"
## 3  ( 1 ) "*"   "*"   " " "*"
## 4  ( 1 ) "*"   "*"   "*" "*"
```

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

```
##      Adj.R2       Cp       BIC
## 1 0.1910048 7.906993 -3.805036
## 2 0.2574811 4.446603 -5.232912
## 3 0.2932620 3.130920 -4.865619
## 4 0.2796525 5.000000 -1.098852
```

Answer:

We would select the model that includes `pop15`, `pop75`, and `ddpi` based on R^2_{adj}.

2. Perform a stepwise selection using AIC

Code:

```r
full <- lm(sr~., data = savings)
step(full, direction = "both")
```

```
## Start:  AIC=138.3
## sr ~ pop15 + pop75 + dpi + ddpi
##
##         Df Sum of Sq    RSS    AIC
## - dpi    1     1.893 652.61 136.45
## <none>               650.71 138.30
```

```
## - pop75  1    35.236 685.95 138.94
## - ddpi   1    63.054 713.77 140.93
## - pop15  1   147.012 797.72 146.49
##
## Step:  AIC=136.45
## sr ~ pop15 + pop75 + ddpi
##
##         Df Sum of Sq    RSS    AIC
## <none>               652.61 136.45
## - pop75  1    47.946 700.55 137.99
## + dpi    1     1.893 650.71 138.30
## - ddpi   1    73.562 726.17 139.79
## - pop15  1   145.789 798.40 144.53

##
## Call:
## lm(formula = sr ~ pop15 + pop75 + ddpi, data = savings)
##
## Coefficients:
## (Intercept)        pop15        pop75         ddpi
##     28.1247      -0.4518      -1.8354       0.4278
```

Answer:

We would again select the model that includes `pop15`, `pop75`, and `ddpi`.

3. Perform a general linear F-test (with $\alpha = 0.1$) to choose between the full model (i.e., using the all 4 predictors) and the reduce model that inlcude `pop15`, `pop75`, and `ddpi` as the predictors

Code:

```
reduce <- lm(sr ~ pop15 + pop75 + ddpi, data = savings)
anova(reduce, full)
```

```
## Analysis of Variance Table
##
## Model 1: sr ~ pop15 + pop75 + ddpi
## Model 2: sr ~ pop15 + pop75 + dpi + ddpi
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     46 652.61
## 2     45 650.71  1    1.8932 0.1309 0.7192
```

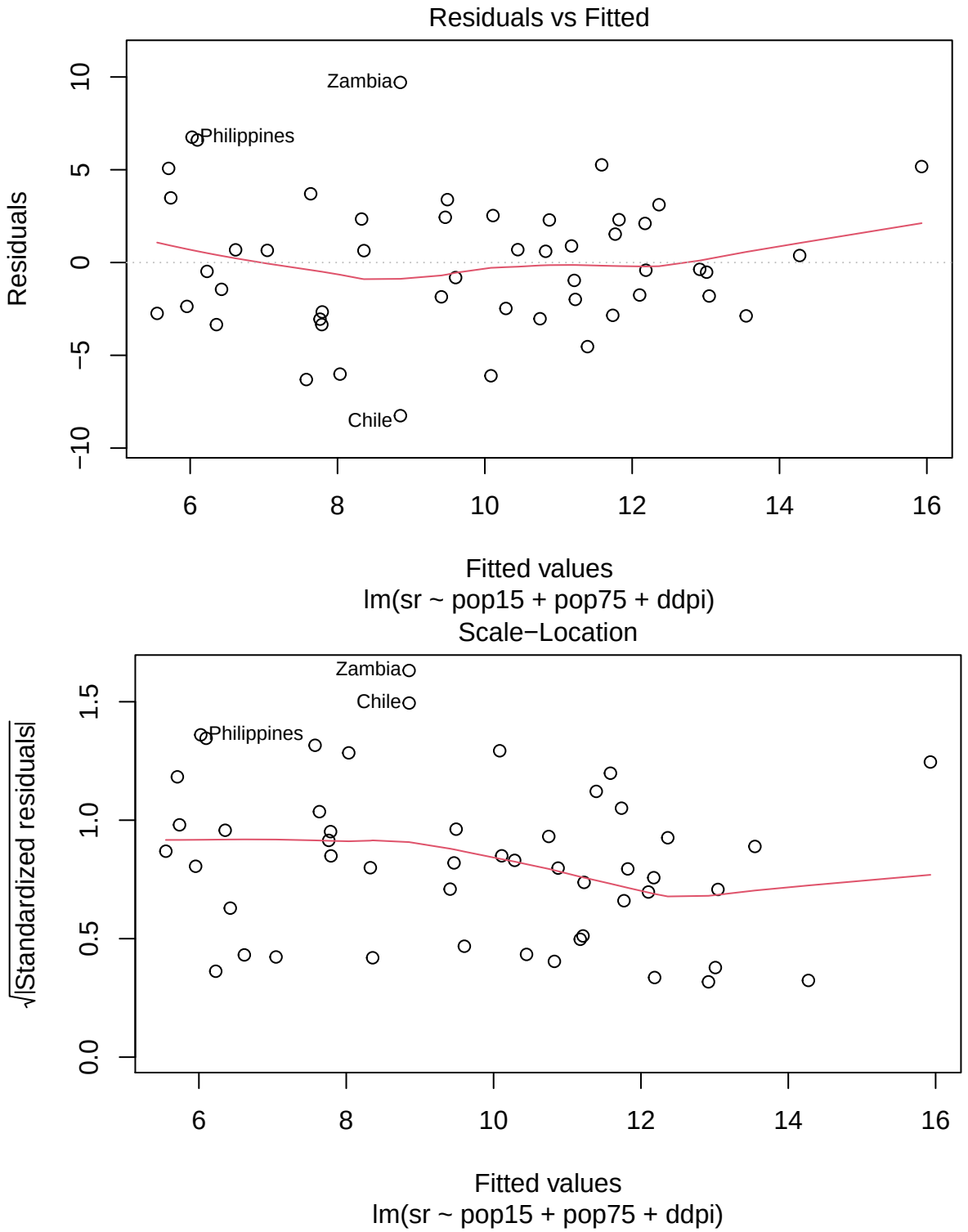
Answer:

The p-value is greater than α, therefore we fail to reject H_0, meaning that we do not have sufficent evidence to include `dpi`.

4. Make a residual plot of the model selected by AIC and comment the model assumptions

Code:

```r
aicModel <- step(full, direction = "both", trace = F)
plot(aicModel, which = c(1, 3))
```



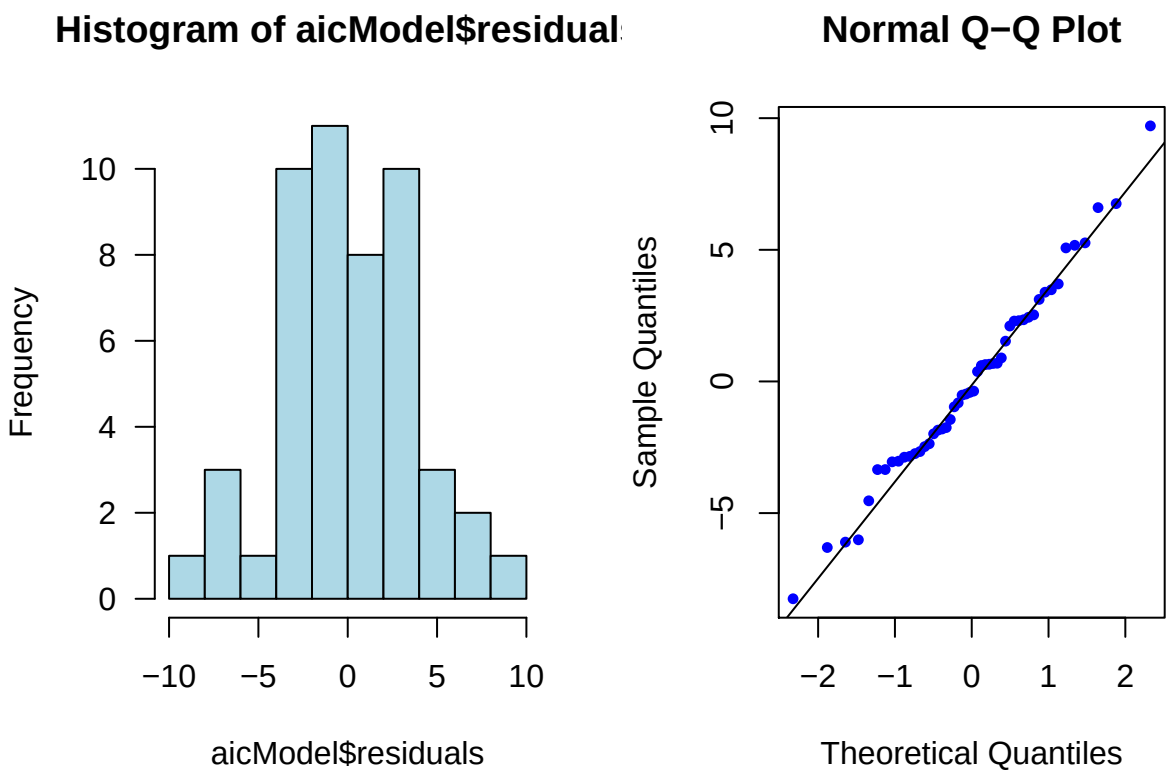
Answer:

The linearity assumption and constant varaince assumption are reasonable based on the residual plots.

5. Use both histogram and qqplot to examine the normality assumption on error

Code:

```
par(mfrow = c(1, 2))
hist(aicModel$residuals, 10, las = 1, col = "lightblue")
qqnorm(aicModel$residuals, pch = 16, cex = 0.75, col = "blue")
qqline(aicModel$residuals)
```



Answer:

The normality assumption on error term is reasonable.

6. Calculate the leverage values to check if there is any high leverage points (i.e., $h > \frac{2p}{n}$)

Code:

```
X <- model.matrix(aicModel)
H <- X %*% solve((t(X) %*% X)) %*% t(X)
lev <- hat(X)
which(lev >= (2 * 4) / 50)
```

```
## [1] 21 23 49
```

```
row.names(savings)[which(lev >= (2 * 4) / 50)]
```

```
## [1] "Ireland" "Japan"   "Libya"
```

```r
## You can also use "hatvalues" to get leverage values: Thanks Lee
hatvalues(aicModel)
```

```
##      Australia        Austria        Belgium        Bolivia         Brazil
##     0.03232688     0.07980303     0.08449336     0.06399229     0.05526517
##         Canada          Chile          China       Colombia     Costa Rica
##     0.02839638     0.03722618     0.07276974     0.05728419     0.07400503
##        Denmark        Ecuador        Finland         France        Germany
##     0.05289186     0.06288264     0.08386715     0.13466283     0.06905292
##         Greece      Guatamala       Honduras        Iceland          India
##     0.06644123     0.06045315     0.05837532     0.06560688     0.06214726
##        Ireland          Italy          Japan          Korea     Luxembourg
##     0.16057383     0.04961285     0.21685941     0.06079056     0.08290531
##          Malta         Norway    Netherlands    New Zealand      Nicaragua
##     0.06894117     0.04345605     0.08935089     0.05247101     0.04963440
##         Panama       Paraguay           Peru    Philippines       Portugal
##     0.03748794     0.06276881     0.06297762     0.06091143     0.06466419
##   South Africa South Rhodesia          Spain         Sweden    Switzerland
##     0.04187055     0.13068766     0.04288517     0.08453760     0.05680013
##         Turkey        Tunisia United Kingdom  United States      Venezuela
##     0.03963730     0.07146547     0.09173390     0.04574418     0.07446401
##         Zambia        Jamaica        Uruguay          Libya       Malaysia
##     0.06331501     0.14070190     0.05781064     0.53130855     0.06168896
```

Answer:

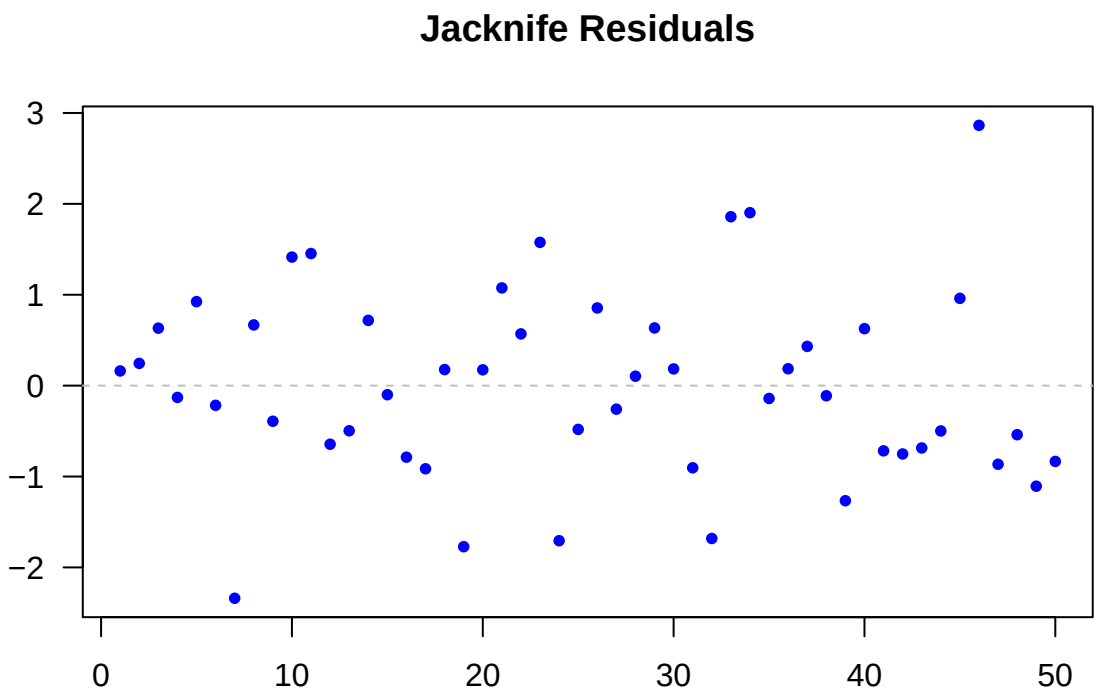
There are three countries with high level points Ireland, Japan, Libya

7. Compute jackknife residuals to identify outlier(s)

Code:

```r
jack <- rstudent(aicModel)

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




```
which.max(jack)
```

```
## Zambia
##     46
```

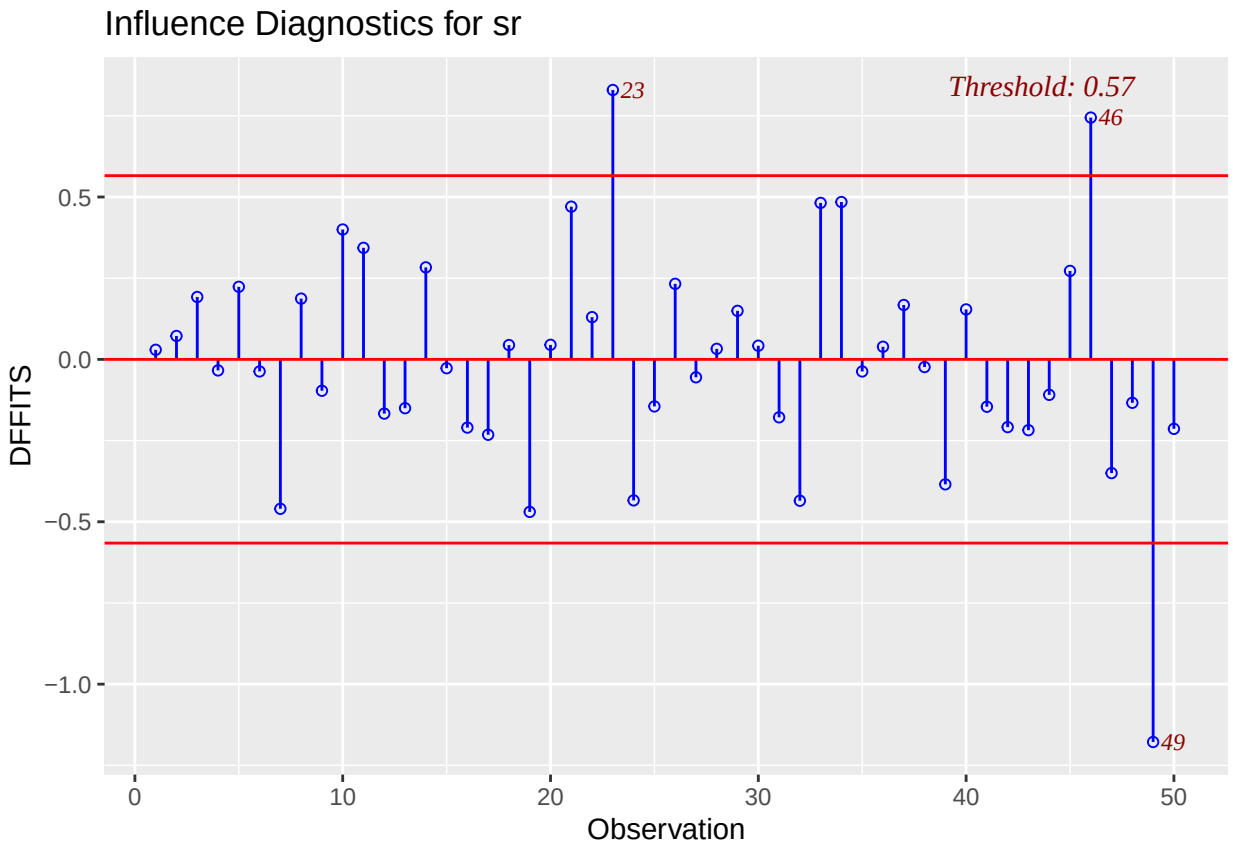
Answer:

There is no obvious outlier

8. Identifying influential observations by computing DFFITS

Code:

```
library(olsrr)
ols_plot_dffits(aicModel)
```




```
row.names(savings)[c(23, 46, 49)]
```

```
## [1] "Japan"  "Zambia" "Libya"
```

Answer:

Japan, Zambia, Libya