DSA 8020 R Lab 4: Model Selection and Model Checking

Meredith Sliger

Contents

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:

```r
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
```

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

Code:

```r
library(tidyverse)
library(caret)
```

```
## Warning: package ’caret’ was built under R version 4.4.2
```

```
library(leaps)
```

```
## Warning: package ’leaps’ was built under R version 4.4.2
```

```
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 ) "*"   "*"   "*" "*"
```

```
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:

The best subset selection helps figure out which predictors are most useful without making the model overly complicated. Looking at adjusted R^2, the best model includes `pop15`, `pop75`, and `ddpi`. This means these three factors do a good job explaining savings rates, while adding `dpi` doesn’t really improve the model much. So, including `dpi` wouldn’t be worth it since it doesn’t add much value.

2. Perform a stepwise selection using AIC

Code:

```
full <- lm(sr ~ ., data = savings)
aicModel <- step(full, direction = "both", trace = F)
summary(aicModel)
```

```
##
## Call:
## lm(formula = sr ~ pop15 + pop75 + ddpi, data = savings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2539 -2.6159 -0.3913  2.3344  9.7070
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.1247     7.1838   3.915 0.000297 ***
## pop15        -0.4518     0.1409  -3.206 0.002452 **
## pop75        -1.8354     0.9984  -1.838 0.072473 .
## ddpi          0.4278     0.1879   2.277 0.027478 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.767 on 46 degrees of freedom
## Multiple R-squared:  0.3365, Adjusted R-squared:  0.2933
## F-statistic: 7.778 on 3 and 46 DF,  p-value: 0.0002646
```

Answer:

Stepwise selection works by adding and removing predictors to find a model that balances accuracy and simplicity. In this case, it ended up with the same model as the best subset method—`pop15`, `pop75`, and `ddpi`. This reinforces the idea that `dpi` isn’t that important. AIC is useful because it avoids adding unnecessary variables that don’t really help the model.

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

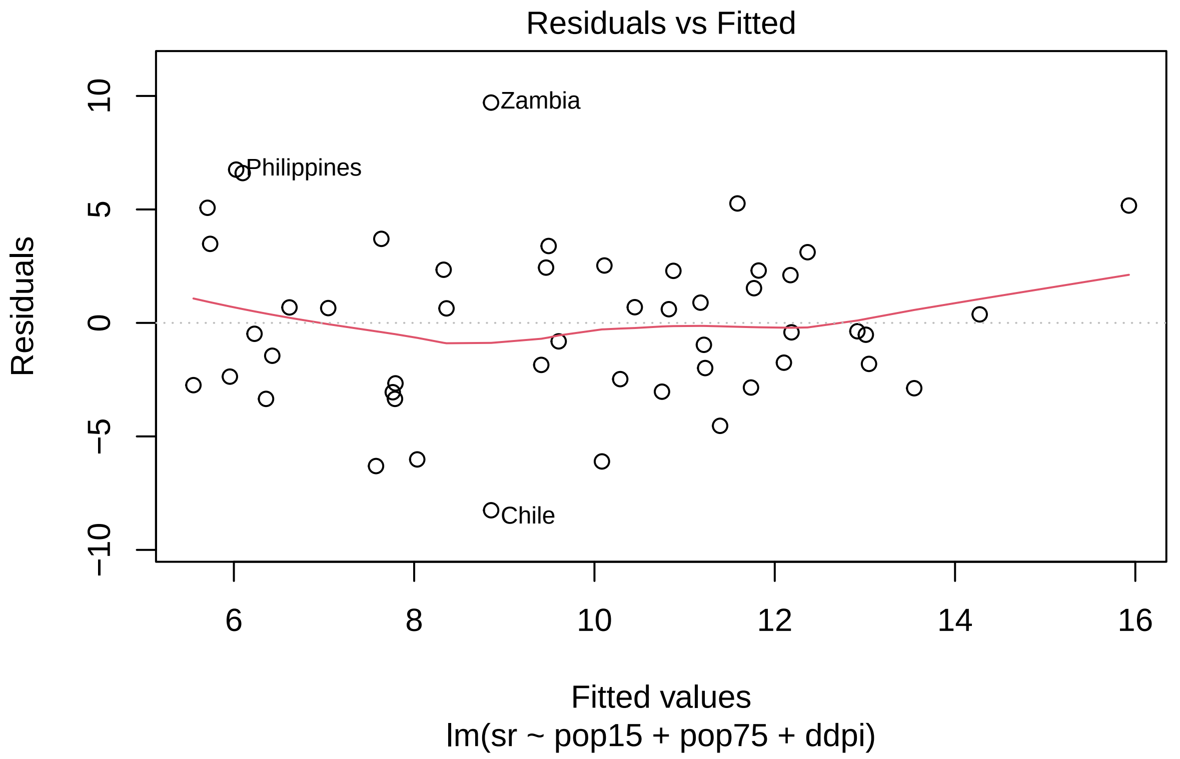
Answer:

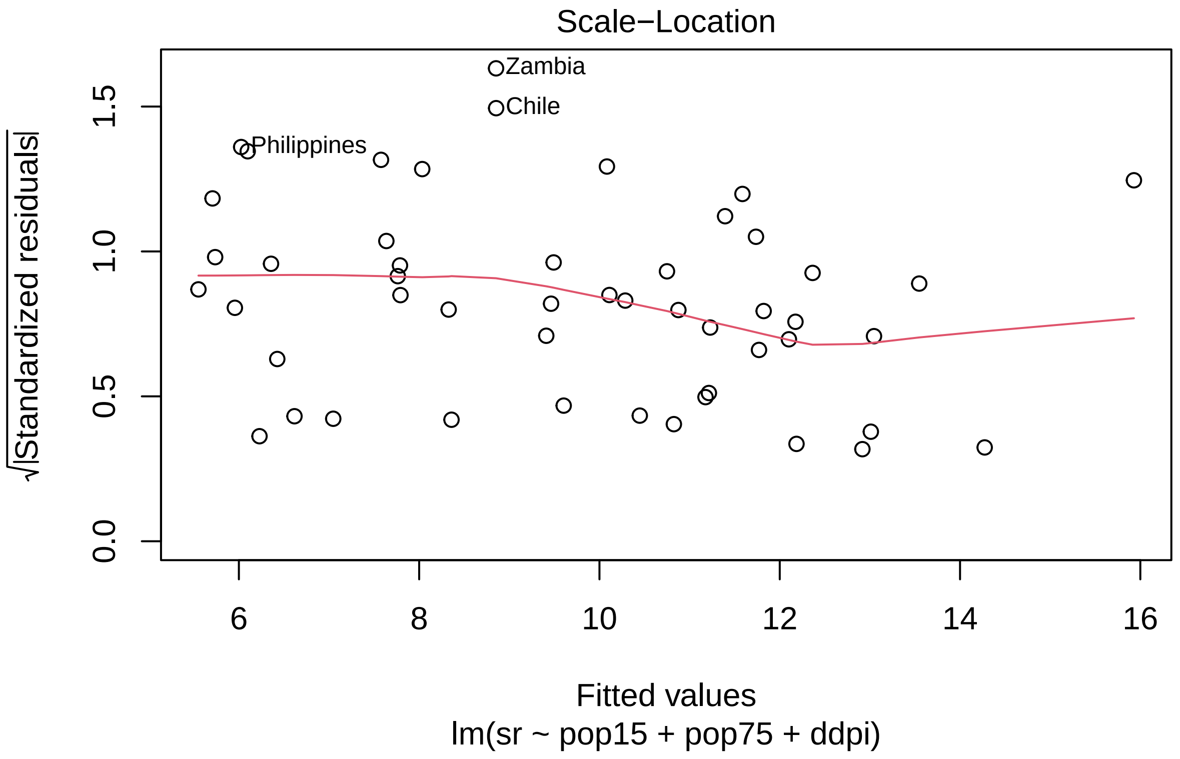
The F-test compares the full model (all four predictors) to the reduced model (without `dpi`) to see if removing `dpi` makes a big difference. The p-value is 0.719, which is way higher than 0.1. That means there’s no strong evidence that `dpi` improves the model. So, the reduced model (with just `pop15`, `pop75`, and `ddpi`) is good enough

4. Make a residual plot of the model selected by *AIC*. Then, comment on the model assumptions

Code:

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





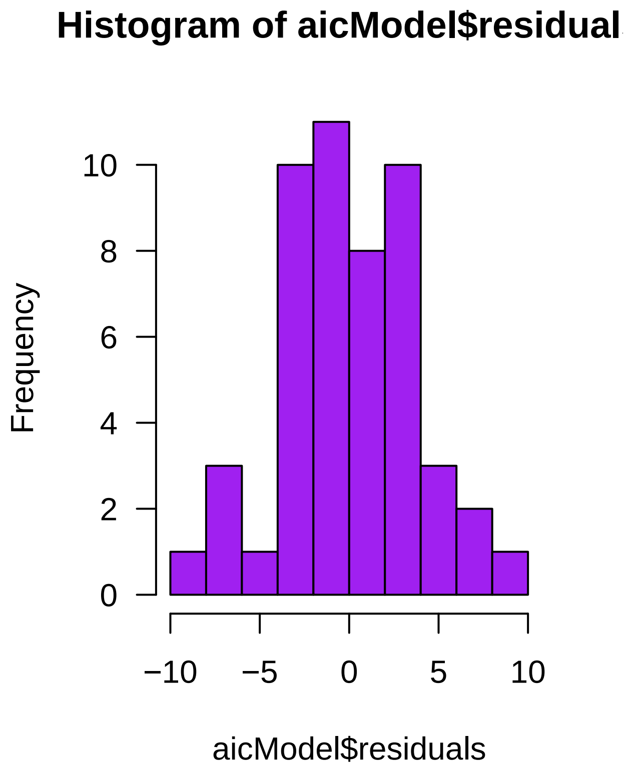
Answer:

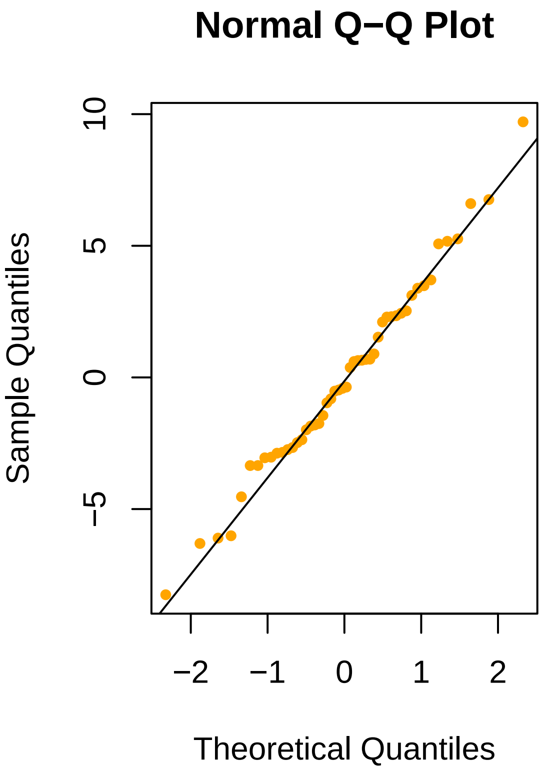
Looking at the residuals vs. fitted values plot, the points seem pretty randomly scattered around zero, which is a good sign. This suggests that the relationship between the predictors and savings rate is fairly linear. The scale-location plot also looks okay—there aren't any obvious patterns, meaning the variance of residuals seems consistent. Overall, there are no major issues with the model's assumptions.

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

Code:

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





Answer:

To check if the residuals follow a normal distribution, I looked at a histogram and a QQ-plot. The histogram is fairly symmetric, which is a good sign. The QQ-plot shows most points falling along the straight line, meaning the residuals are pretty close to normal. There are some slight deviations in the tails, which might suggest a little skewness or a couple of outliers, but nothing too concerning.

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

Code:

```
lev <- hatvalues(aicModel)
p <- length(coef(aicModel)); n <- dim(savings)[1]
which(lev >= (2 * p) / n)
```

```
## Ireland    Japan    Libya
##      21       23       49
```

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

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

Answer:

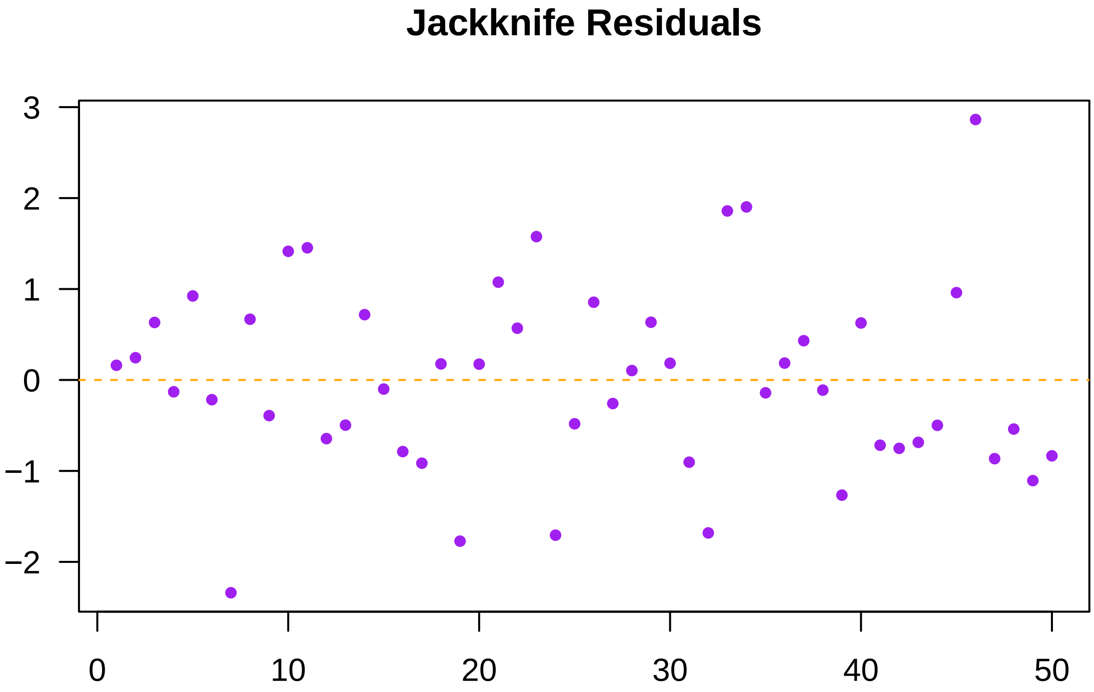
Leverage values tell us how much influence each data point has on the model. A common rule is that if a leverage value is greater than $\frac{2p}{n}$, it might be a high-leverage point. Based on this, Ireland, Japan, and

Libya stand out. These countries may have unusual savings patterns compared to the rest, so they could be affecting the model more than other points.

7. Compute jackknife residuals to identify outlier(s)

Code:

```r
jack <- rstudent(aicModel)
par(las = 1)
plot(jack, pch = 16, cex = 0.8, col = "purple", main = "Jackknife Residuals",
     xlab = "", ylab = "")
abline(h = 0, lty = 2, col = "orange")
```



Answer:

Jackknife residuals help spot outliers by showing how much each data point affects the model when removed. Looking at the plot, most points fall within the normal range, meaning there aren't any extreme outliers. Some points have slightly higher residuals, but they don't seem to be major outliers that would drastically change the model.

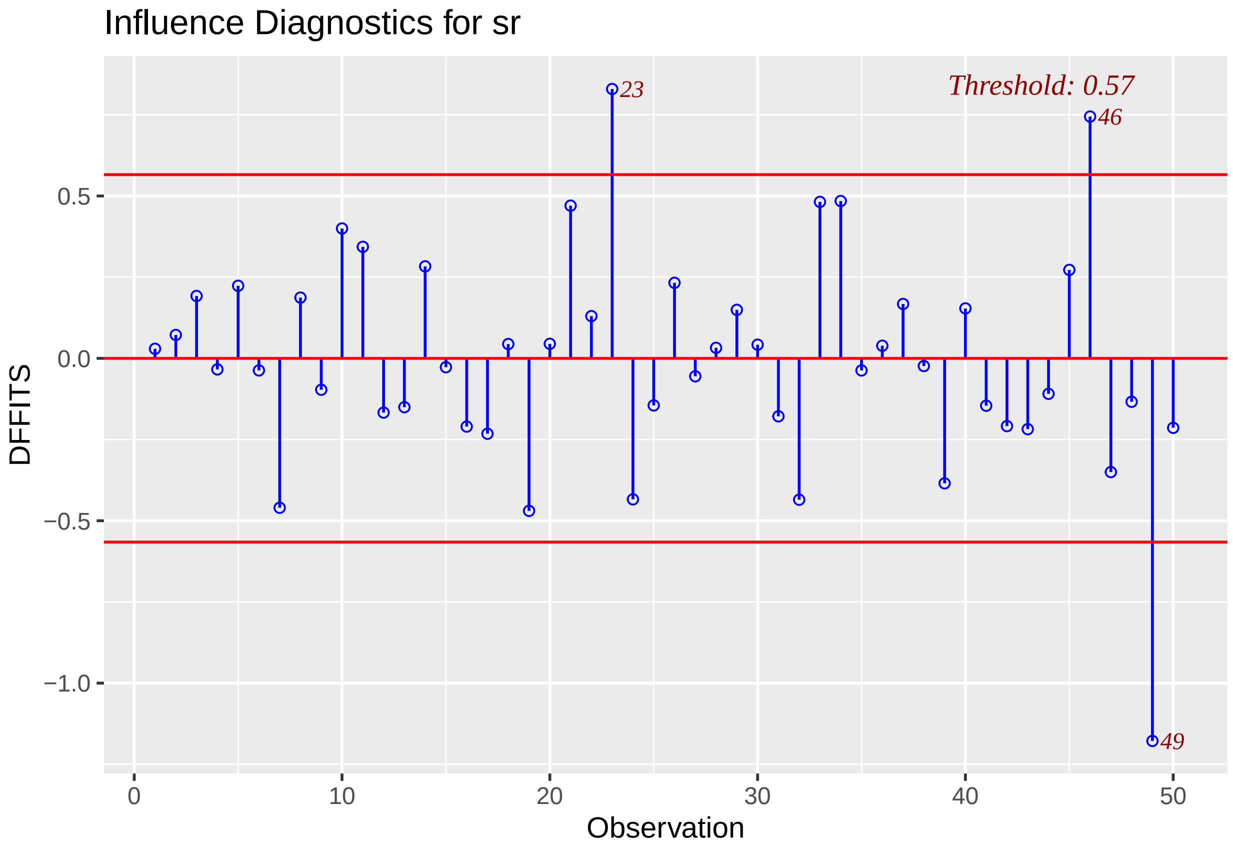
8. Identifying influential observations by computing DFFITS

Code:

```
library(olsrr)
```

```
## Warning: package 'olsrr' was built under R version 4.4.2
```

```
ols_plot_dffits(aicModel)
```



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

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

Answer:

DFFITS measures how much each observation affects the model's predictions. If a point has a high DFFITS value, it means removing it would noticeably change the model. In this case, Japan, Zambia, and Libya are flagged as influential. Their data might be quite different from the other countries, so they could be pulling the model in a certain direction. It might be worth checking if these points should be kept or handled differently.