Applied Biostatistics Assigment 1

Léonard Berney, Mohammad Aquil

Data

The dataset we will be working on contains information on life-cycle savings for the 1960-1970 period in different countries. The data consists of 50 observations on 5 variables:

- sr: personal savings
- pop15: percentage of population under 15
- pop75: percentage of population over 75
- dpi: per-capita disposable income
- ddpi: percentage growth rate of dpi

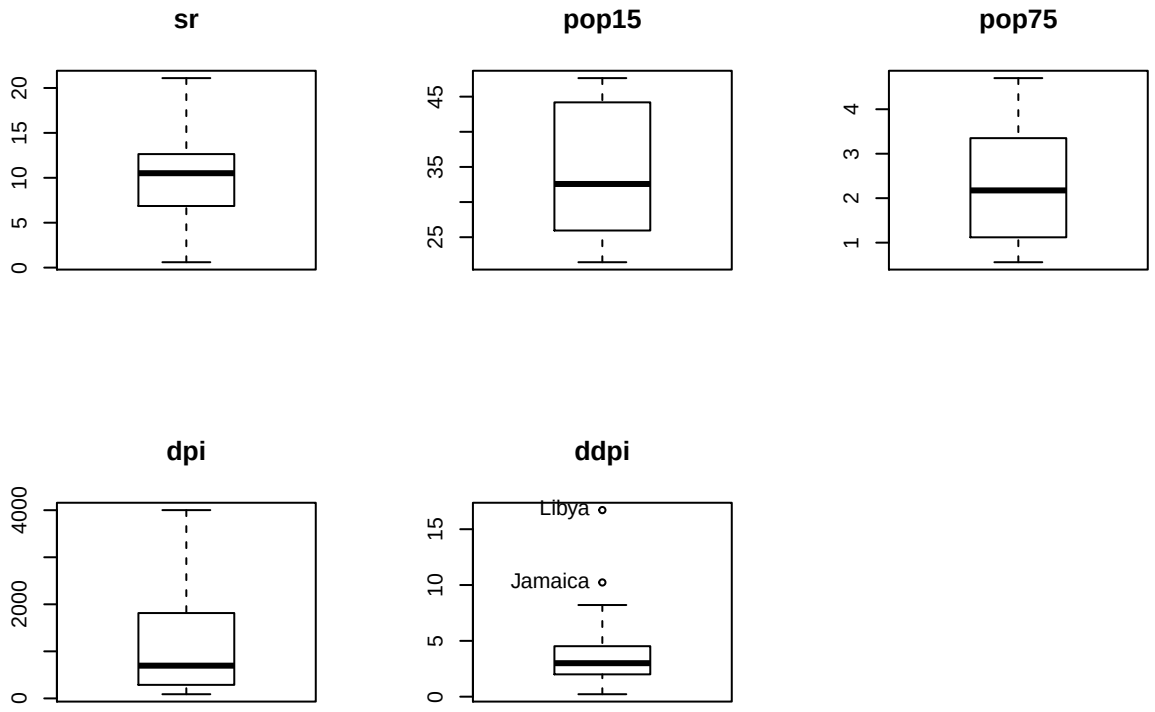

Figure 1: Boxplots of the variables

Linear Model

The objective is to build a linear model that can predict the personal saving ratio of a country. We will start by fitting a model using every variables and then try to prune it as much as possible, without sacrificing too much accuracy.

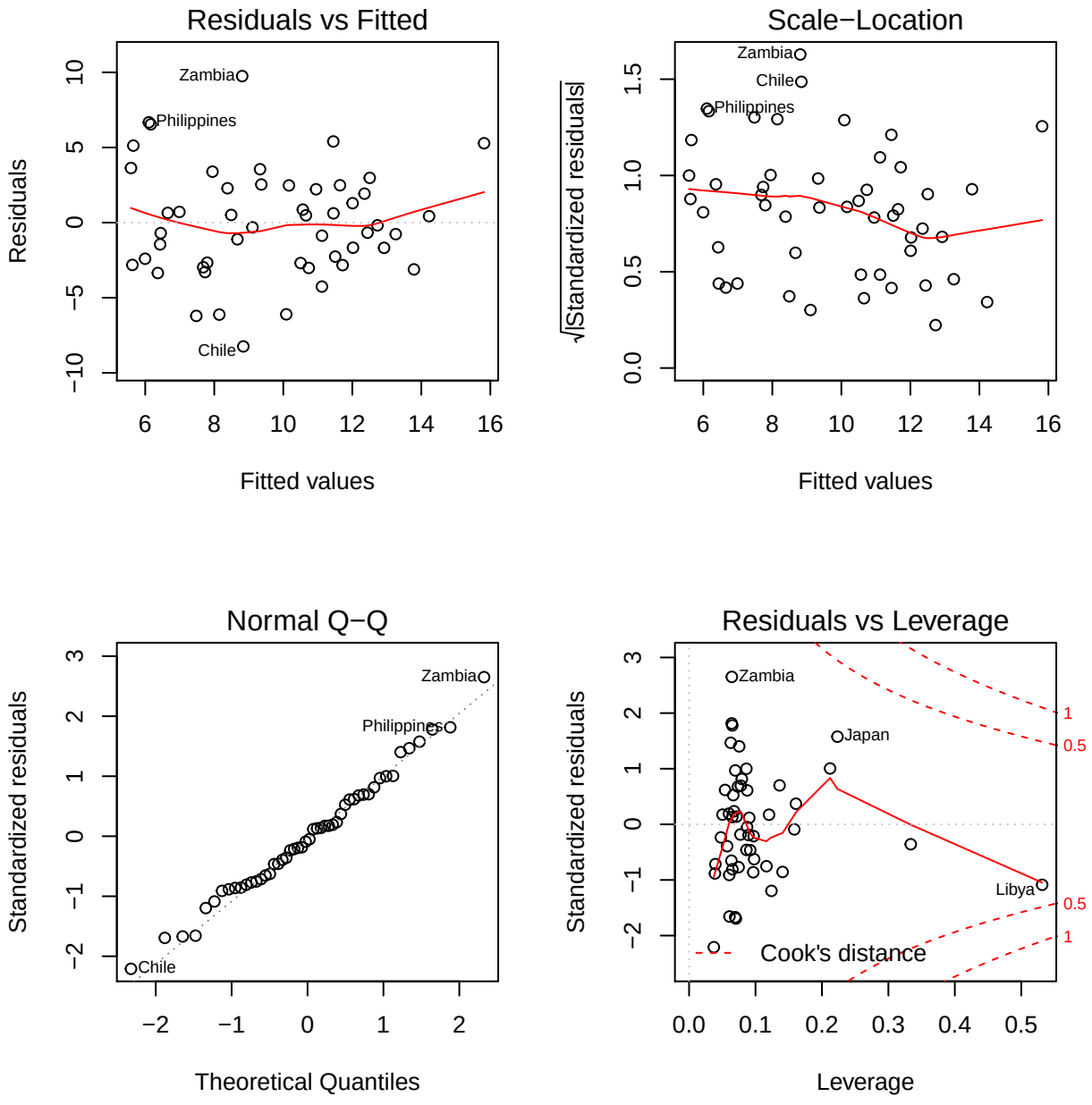

Figure 2: Diagnostic plots for the initial model

From Figure 2 we observe that the residuals are close enough to a normal distribution and are homoscedastic, so no transformation of the data is required before being able to exploit a linear model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	28.5661	7.3545	3.88	0.0003
pop15	-0.4612	0.1446	-3.19	0.0026
pop75	-1.6915	1.0836	-1.56	0.1255
dpi	-0.0003	0.0009	-0.36	0.7192
ddpi	0.4097	0.1962	2.09	0.0425

Table 1: Summary of the full model

Looking at the summary from Table 1, we can see that dpi is most probably useless and when using a confidence level of 95% pop75 is not statically significant either.

We now build a second linear model omitting the dpi variable. By comparing it with the full model using a log likelyhood ratio test, Table 2 shows that we can indeed remove dpi from the model.

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
1	6	-135.10			
2	5	-135.17	-1	0.15	0.7031

Table 2: Log-likelyhood ratio test of the full model (1) and removing dpi (2)

We then remove the variable pop75 from the second model and do another likelyhood ratio test.

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
1	5	-135.17			
2	4	-136.94	-1	3.54	0.0597

Table 3: Log-likelyhood ratio test of the second model (1) and removing pop75 (2)

This time the results are less obvious but at a 95% confidence level we still reject the hypothesis that the second model is better than the third.

Conclusion

We were able to simplify the full model containing 4 variables by removing the 2 least significant ones, leading to the final model:

```
## (Intercept)       pop15        ddpi
##  15.5995758  -0.2163762   0.4428302
```

One thing to note is that the boxplot of ddpi from Figure 1 showed Libya as an outsider but on Figure 2 the residuals vs leverage plot indicated that Libiya was not too influencial on the full model. However, on the same graph for the final model on Figure 3 Libya has a greater leverage compared to the other countries. We also see that other points have had their leverage changed but since Libya was already an outsider, we don't care too much about the changes in the final model.

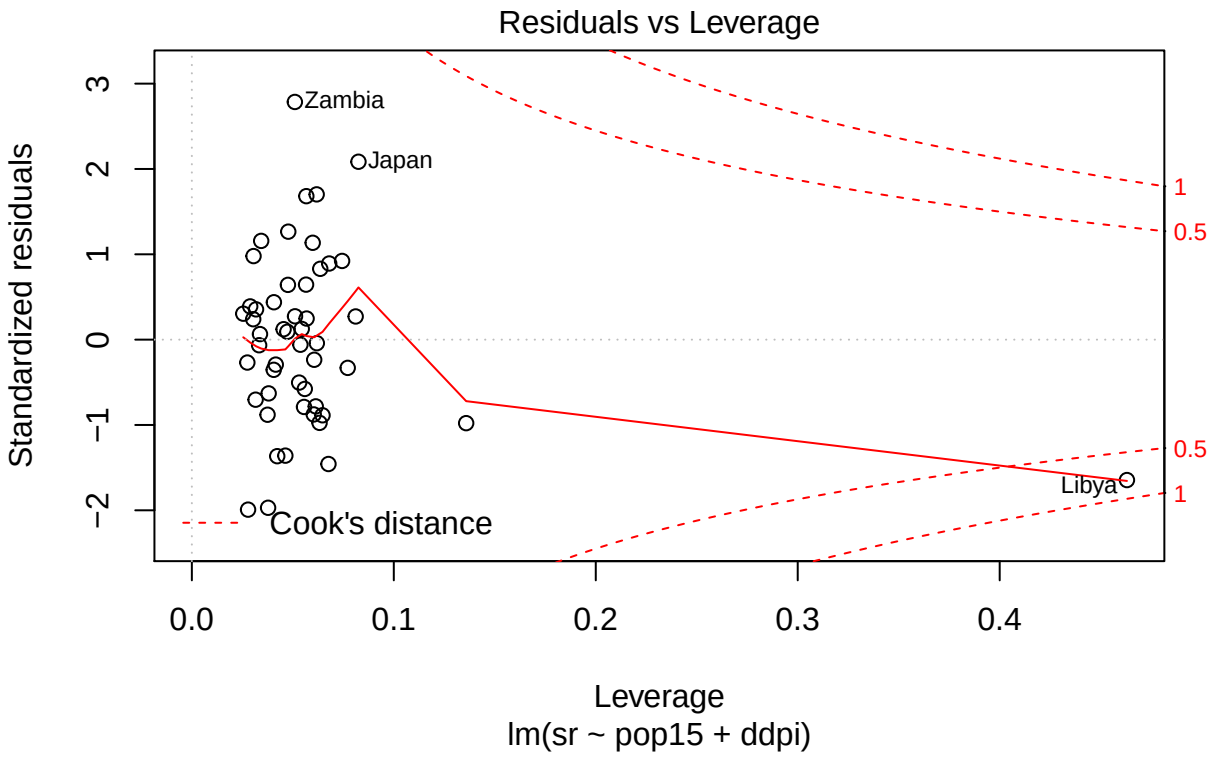

Figure 3: Cook's distance for the final model