Diagnostics

Based on a book by Julian J. Faraway

University of Iceland

Data

Under the life-cycle savings hypothesis as developed by Franco Modigliani, the savings ratio (aggregate personal saving divided by disposable income) is explained by per-capita disposable income, the percentage rate of change in per-capita disposable income, and two demographic variables: the percentage of population less than 15 years old and the percentage of the population over 75 years old. The data are averaged over the decade 1960-1970 to remove the business cycle or other short-term fluctuations.

```
library(faraway) # you need to install the package first
data(savings)
```

The dataframe contains the following columns:

`sr`	savings rate - personal saving divided by disposable income
`pop15`	percent population under age of 15
`pop75`	percent population over age of 75
`dpi`	per-capita disposable income in dollars
`ddpi`	percent growth rate of dpi

Diagnostics

- Regression model building is often an iterative and interactive process.
- The first model we try may prove to be inadequate.
- Regression diagnostics are used to detect problems with the model and suggest improvement.
- This is a hands-on process.

Residuals and leverage

Recall that $\hat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$ where $\boldsymbol{H}$ is the hat matrix.

Now,

$$\begin{aligned} \hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = & \quad (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} \\ = & \quad (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon} \\ = & \quad (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon} \end{aligned}$$

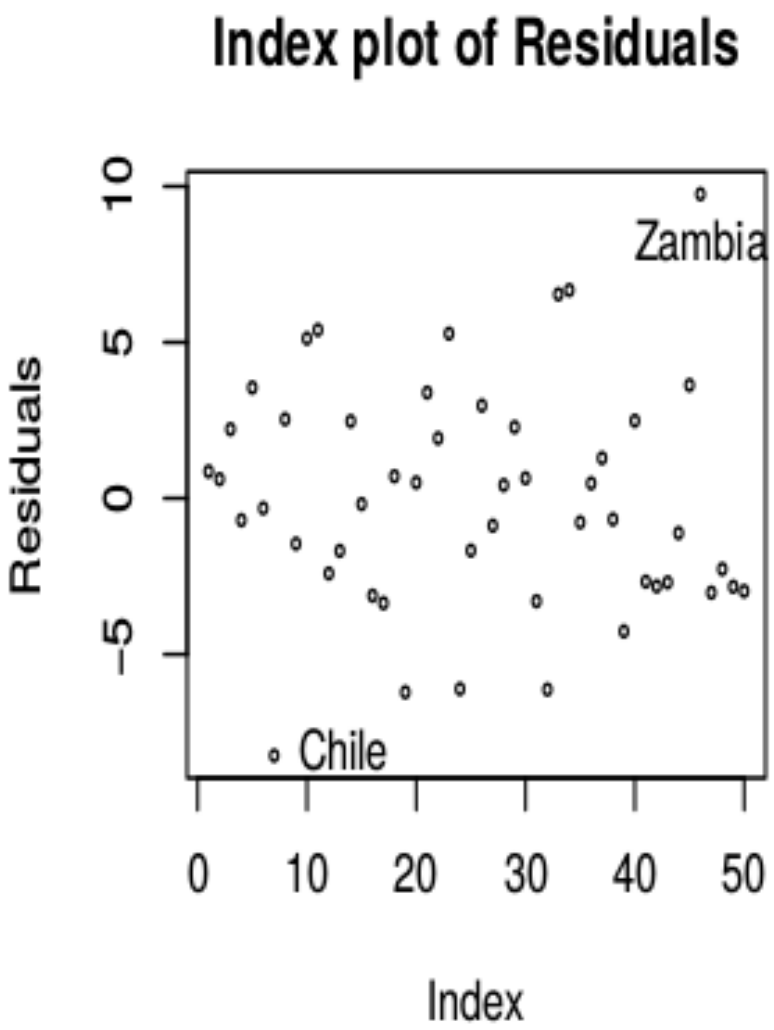
and $\text{var}[\hat{\boldsymbol{\varepsilon}}] = (\boldsymbol{I} - \boldsymbol{H})\sigma^2$ assuming $\text{var}[\epsilon] = \sigma^2\boldsymbol{I}$.

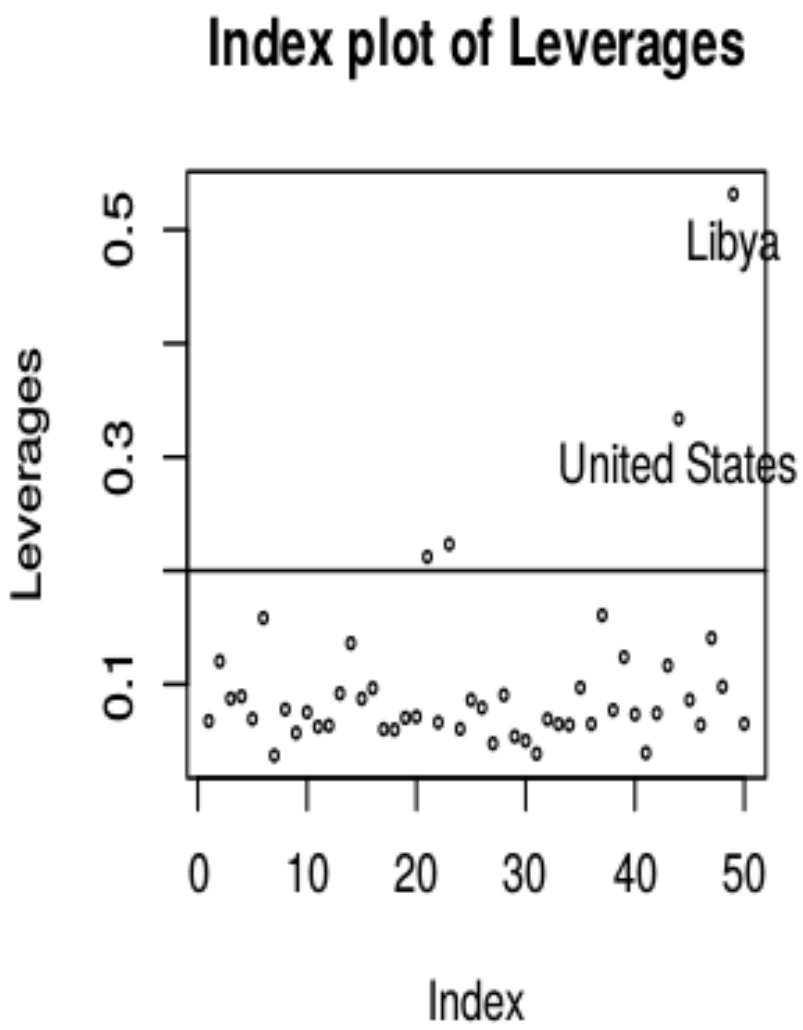
$h_i = H_{ii}$ are called *leverages* and are useful diagnostics.

Leverages

- Leverage is a measure of how far away the independent variable values of an observation are from those of the other observations.
- Large values of h_i are due to extreme values in X.
- High-leverage points are those observations, if any, made at extreme or outlying values of the independent variables such that the lack of neighboring observations means that the fitted regression model will pass close to that particular observation.
- An average value for h_i is p/n and a "rule of thumb" is that leverages of more than $2p/n$ should be looked at more closely.

Outliers


Index plot of Residuals
Residuals
10
5
0
−5
Zambia
Chile
0
10
20
30
40
50
Index


Index plot of Leverages
Leverages
0.5
0.3
0.1
Libya
United States
0
10
20
30
40
50
Index

Studentized residual

We saw earlier that

$$\text{var}[\hat{\varepsilon}_i] = (1 - h_i)\sigma^2$$

This suggests the use of

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

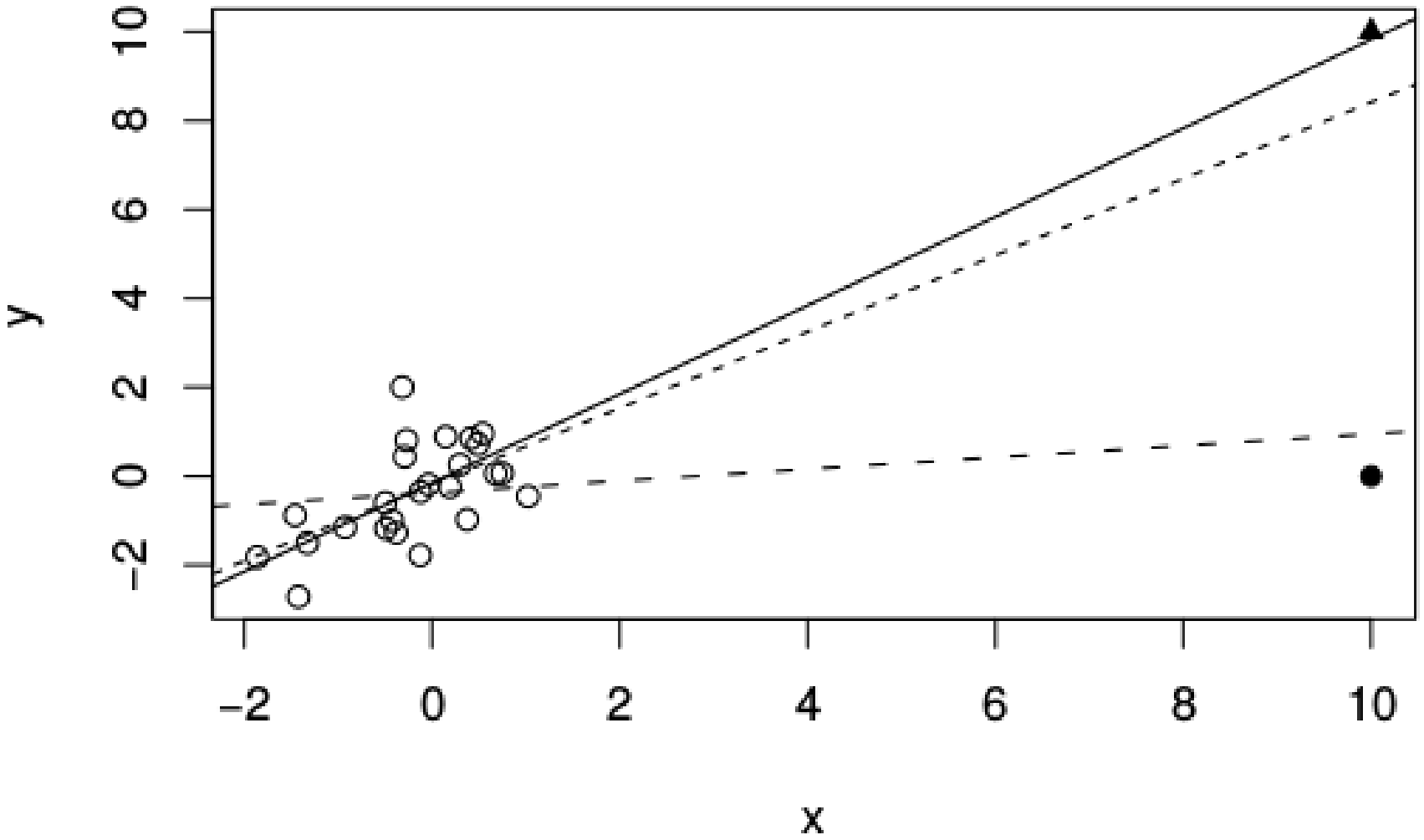
which are called *studentized residuals*

- If the model assumptions are correct $\text{var}[r_i] = 1$
- Studentized residuals are sometimes preferred in residual plots as they have been standardized to have equal variance.

An outlier test

- An outlier is a point that does not fit the current model.
- We need to be aware of such exceptions.
- An outlier test is useful because it enables us to distinguish between truly unusual points and residuals which are large but not exceptional.

Outliers



An outlier test

We exclude point i and recompute the estimates to get $\hat{\boldsymbol{\beta}}_{(i)}$ and $\hat{\sigma}^2_{(i)}$ where (i) denotes that the ith observation has been excluded.

Hence

$$\hat{y}_{(i)} = \boldsymbol{x}_{\boldsymbol{i}}^T \hat{\beta}_{(i)}$$

If $\hat{y}_{(i)} - y_i$ is large, i is an oultlier.

How large is large?

Jackknife residuals

Let us define the *jackknife* (or *externally studentized* or *leave one out studentized* or *crossvalidated*) residuals as

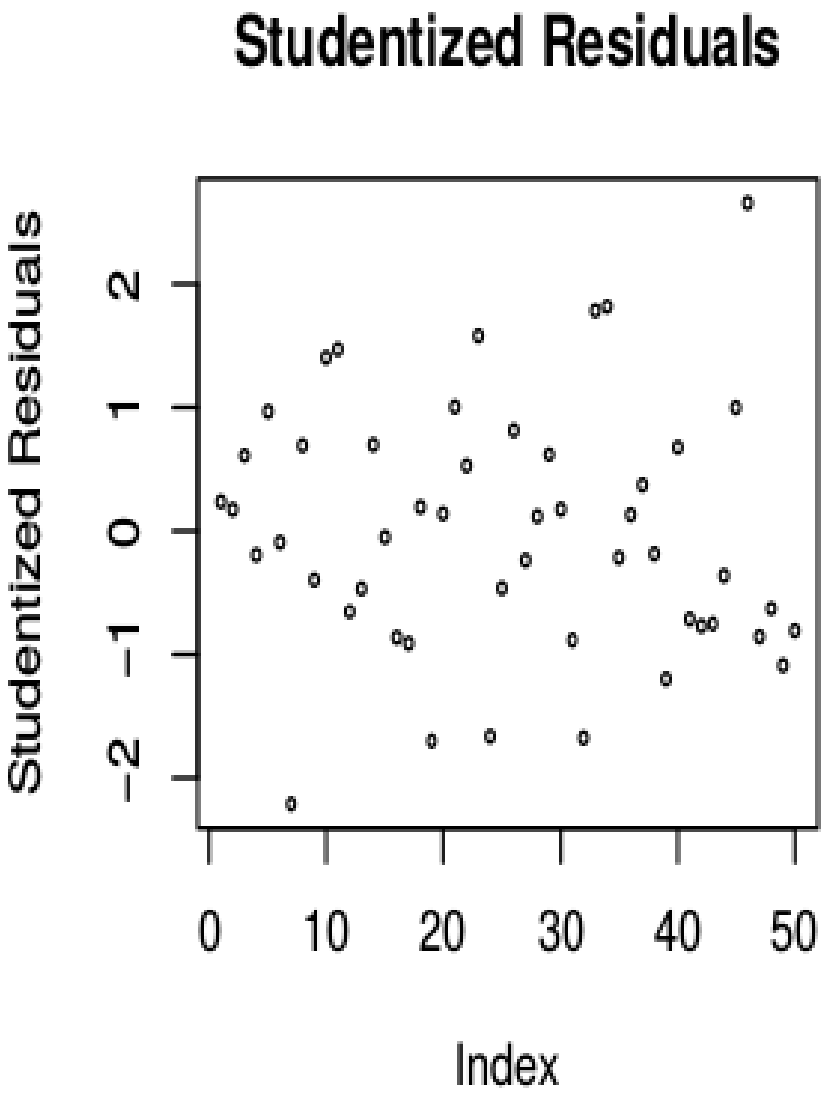
$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_i}} = r_i\sqrt{\frac{n-p-1}{n-p-r_i^2}}$$

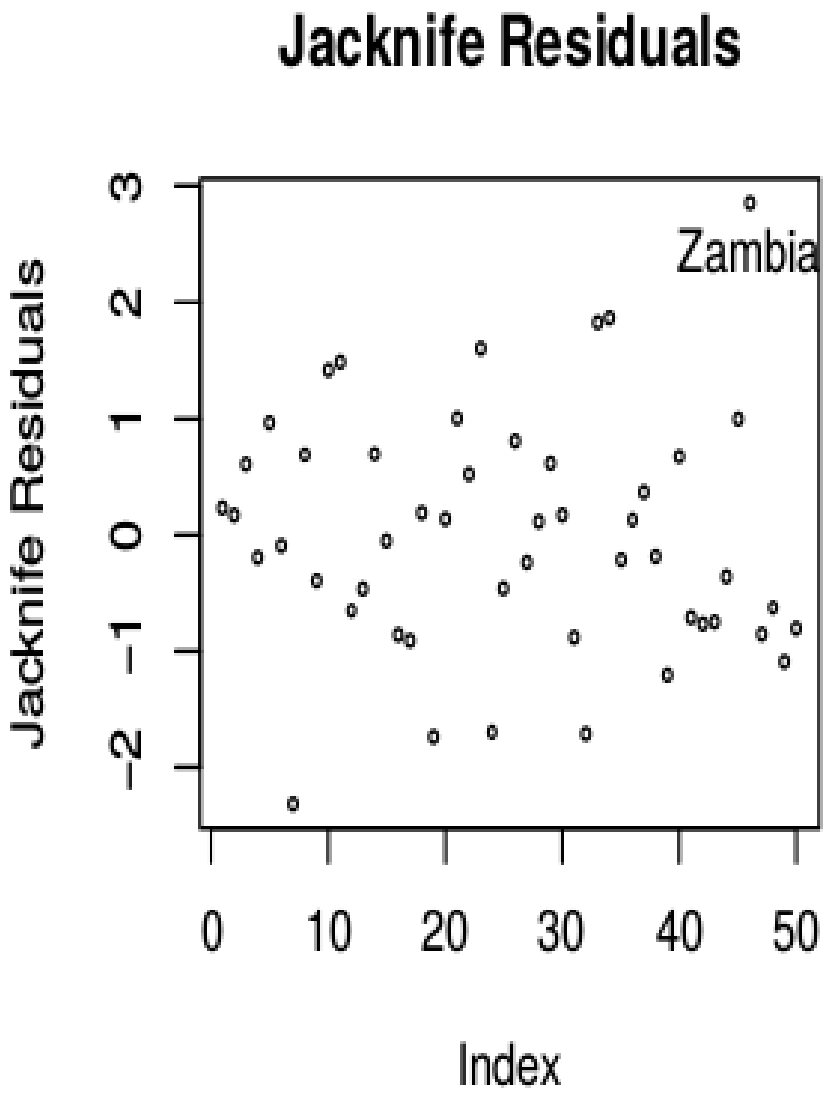
If the model is correct and $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\boldsymbol{I})$, t_i follows a t-distribution with $(n-p-1)$ degrees of freddom.

Jackknife residuals

- Since $t_i \sim t_{(n-p-1)}$ we can calculate a p-value to test whether case i is an outlier.
- However, we are likely to want to test all cases (residuals) so we must adjust the level of the test accordingly.
- We can use Bonferroni correction (conservative method): adjust the α level as α/n.

Outliers


Studentized Residuals
Studentized Residuals
−2
−1
0
1
2
0
10
20
30
40
50
Index


Jacknife Residuals
Jacknife Residuals
Zambia
−2
−1
0
1
2
3
0
10
20
30
40
50
Index

What should be done about outliers

- Check for a data entry error first - these are relatively common.
- Examine the physical context - why did it happen? Sometimes, the discovery of an outlier may be of singular interest. Some scientific discoveries spring from noticing unexpected aberrations.
- Exclude the point from the analysis but try reincluding it later if the model is changed. The exclusion of one or more points may make the difference between getting a statistical significant result or having some unpublishable research.
- To avoid any suggestion of dishonesty, always report the existence of outliers even if you do not include them in your final model.

Influential observations

- An influential point is one whose removal from the dataset would cause a large change in the fit.
- An influential point may or may not be an outlier and may or may not have large leverage but it will tend to have at least one of those two properties.
- Some measures of influence, where the subscripted i indicates the fit without case i are
  - Change in the coefficients $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}$
  - Change in the fit $\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_{(i)}$
- These are hard to judge in the sense that the scale varies between datasets.

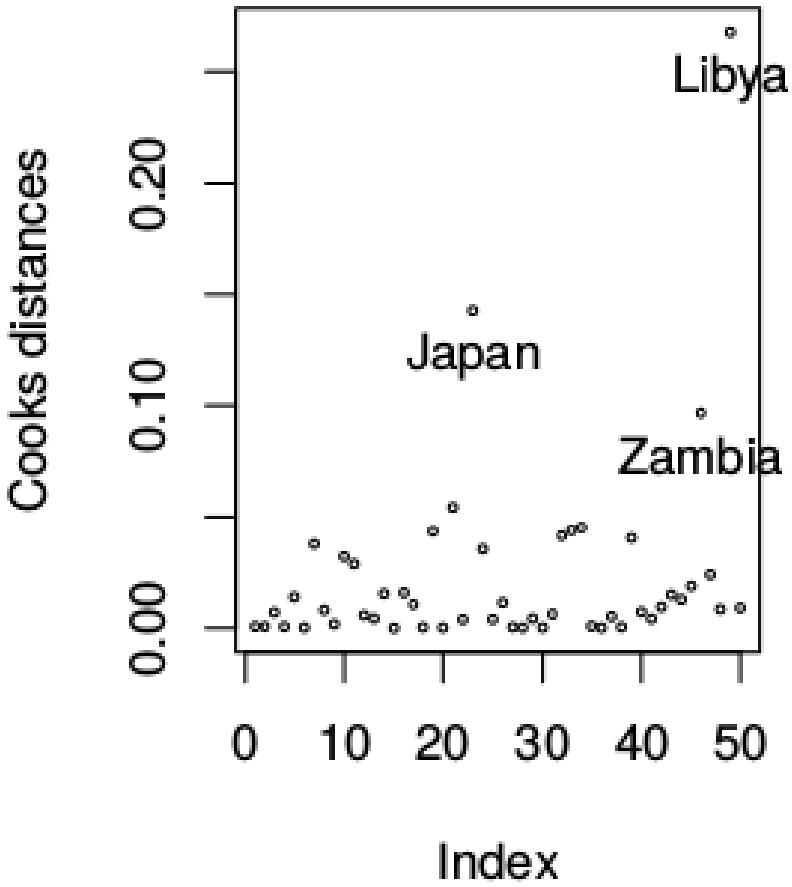
Cook's distance

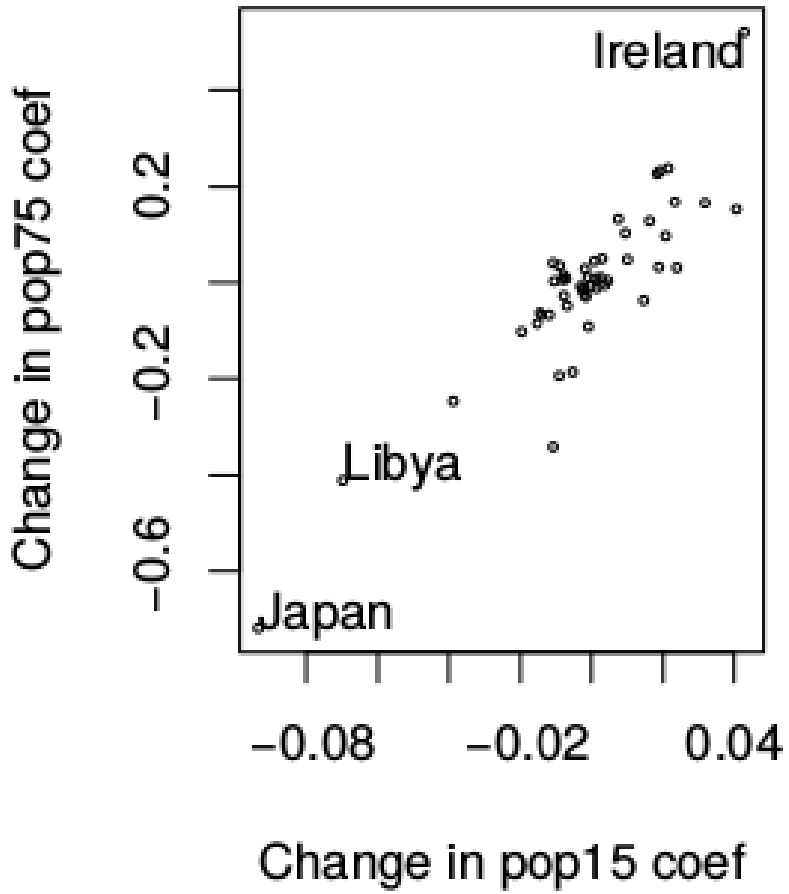
A popular alternative to the measures above is the Cook's distance (Cook statistics) defined as:

$$D_i = \frac{1}{p} r_i^2 \frac{h_i}{1 - h_i}$$

An index plot of D_i can be used to identify influential points.

Cook's distance



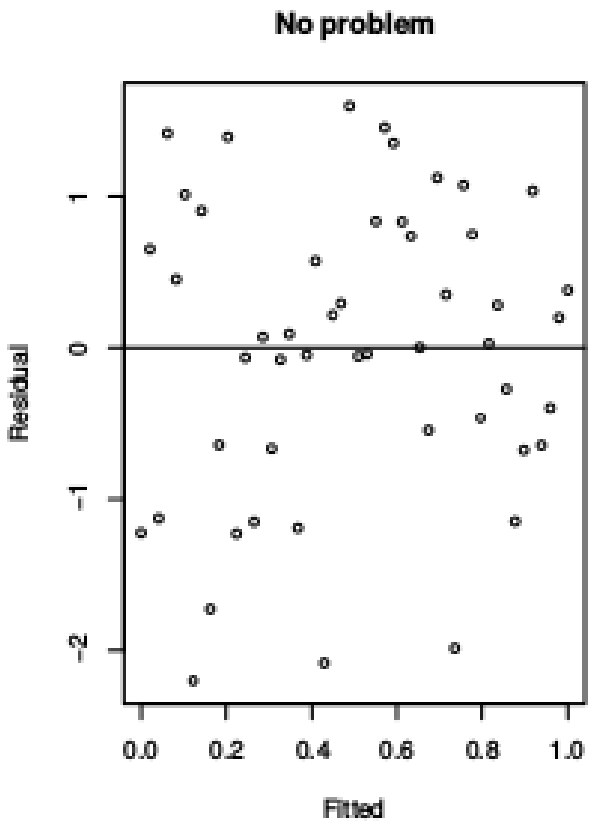


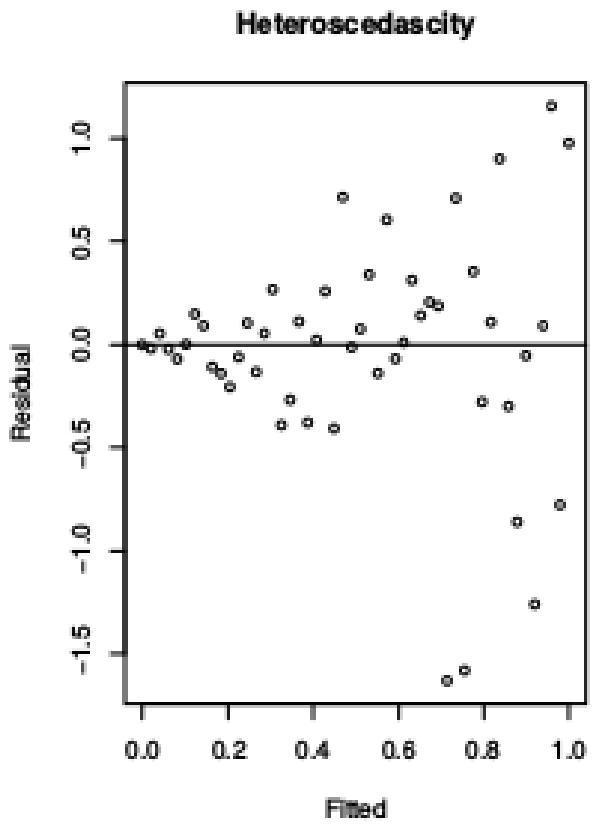
Residual plots

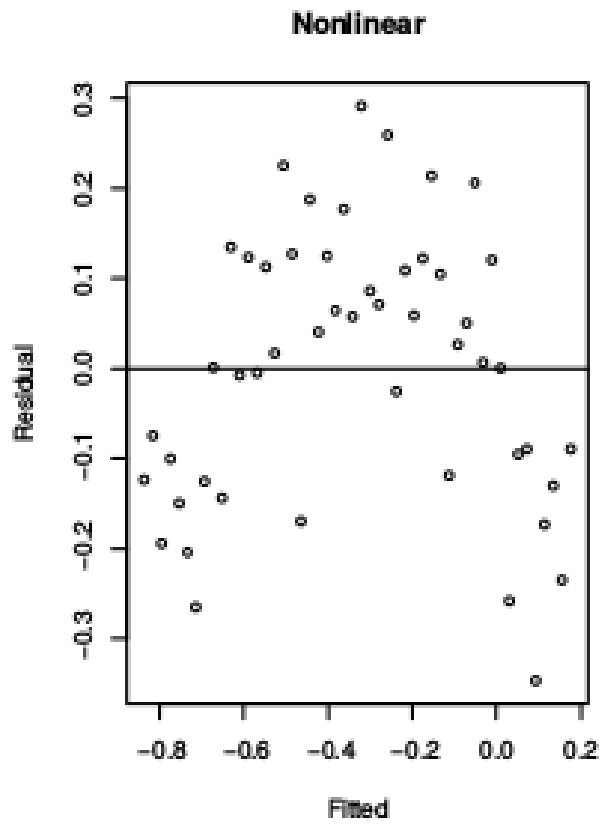
Outliers and influential points indicate cases that are in some way individually unusual but we also need to check the assumptions of the model.

- A plot of $\hat{\varepsilon}$ against $\hat{y}$ is the most importand diagnostic plot.
- If all is well, you should see constant variance in the vertical ($\hat{\varepsilon}$) direction and the scatter should be symmetric vertically about 0.
- Things to look for are heteroscedascity (non-constant variance) and nonlinearity (which indicates some change in the model is necessar).

Residual plots







Residual plots

- You should also plot $\hat{\varepsilon}$ against x_i (for predictors that are both in and out of the model).
- Look for the same things except in the case of plots against predictors not in the model, look for any relationship which might indicate that this predictor should be included.

Non-constant variance

There are two approaches to dealing with non-constant variance.

- Weighted least squares is appropriate when the form of the no-constant variance is either known exactly or there is some known parametric form.
- Alternatively, one can transform y to $h(y)$ where h is chosen so that var$[h(y)]$ is constant.

Non-linearity

In order to check if the systematic part ($\mathrm{E}[X] = \boldsymbol{X}\boldsymbol{\beta}$) of the model is correct we can look at

- Plots of $\hat{\varepsilon}$ against $\hat{y}$ and x_i
- Plots of y against each x_i

But what about the effects of other x on the y vs. x_i plot?

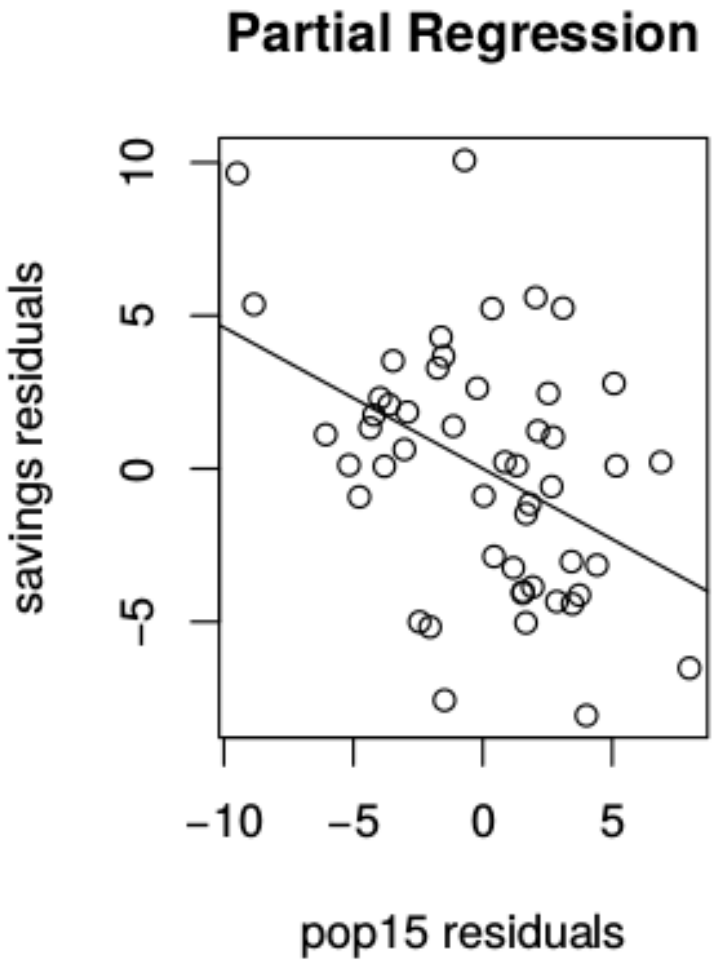
Partial regression or *Added variable* plots can help isolate the effect of x_i on y.

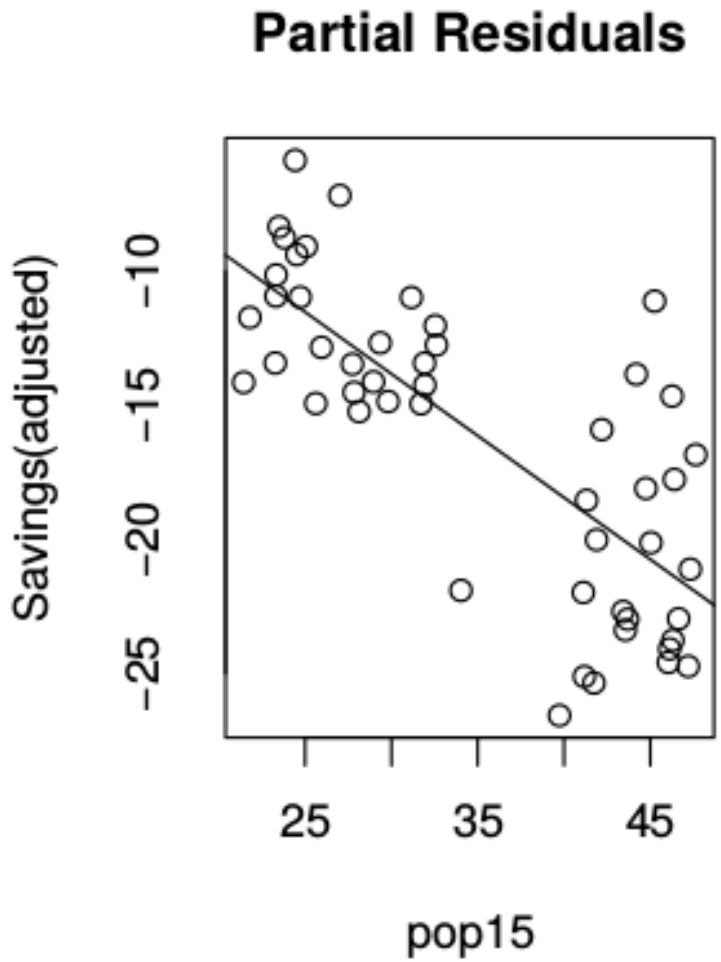
Partial Regression plot

1. Regress y on all x except x_i, get residuals $\hat{\delta}$
   - this represents y with the other X-effect taken out.
2. Regress x_i on all x except x_i, get residuals $\hat{\gamma}$
   - this represents x_i with the other X-effect taken out.
3. Plot $\hat{\delta}$ against $\hat{\gamma}$.

The slope of a line fitted to the plot is $\hat{\beta}_i$. Look for non-linearity and outliers and/or influential points.

Residual plots

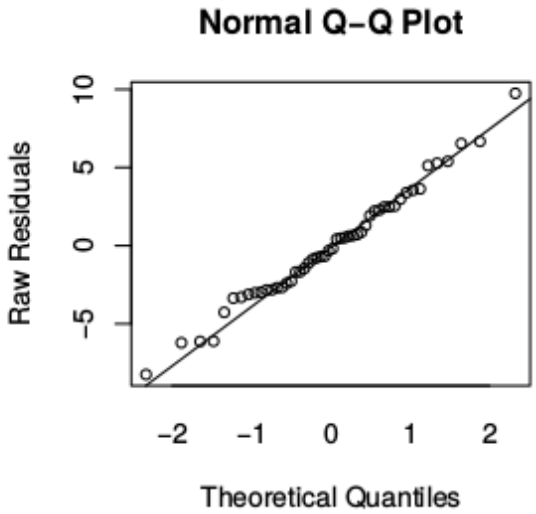


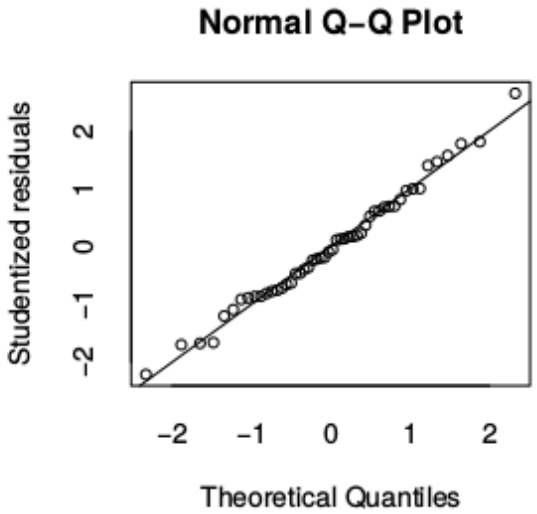


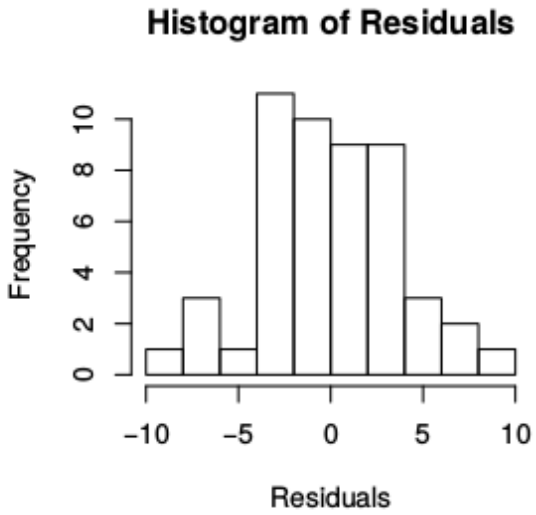
Assessing Normality

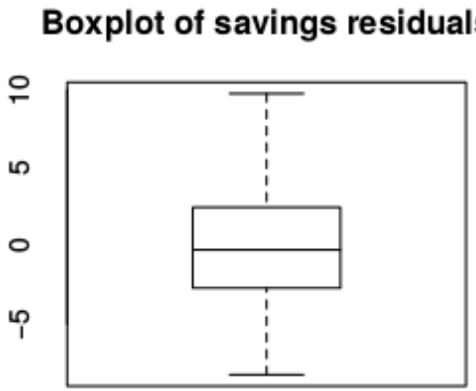
- The test and confidence intervals we use are based on the assumption of normal errors.
- The residuals can be assessed for normality using a *Q-Q plot*.
- The steps are:
  1. Sort the residuals: $\hat{\varepsilon}_{[1]} \leq ...\hat{\varepsilon}_{[n]}$
  2. Compute $u_i = \Phi^{-1}\left(\frac{i}{n+1}\right)$
  3. Plot $\hat{\varepsilon}_{[i]}$ against u_i. If the residuals are normally distributed an approximately straight-line relationship will be observed.

Q-Q plot


Normal Q–Q Plot
Raw Residuals
10
5
0
−5
−2
−1
0
1
2
Theoretical Quantiles


Normal Q–Q Plot
Studentized residuals
2
1
0
−1
−2
−2
−1
0
1
2
Theoretical Quantiles


Histogram of Residuals
Frequency
0
2
4
6
8
10
−10
−5
0
5
10
Residuals


Boxplot of savings residuals
10
5
0
−5

What to do in cases of non-normality?

- A transformation of the response may solve the problem - this is often true for skewed errors.
- Other changes in the model may help.
- Accept non-normality and base the inference on the assumption of another distribution or use resampling methods such as the bootstrap or permutation tests. You don't want to do this unless absolutely necessary. Alternatively use robust methods which give less weight to outlying points. This is appropriate for long tailed distributions.
- For short-tailed distributions, the consequences of non-normality are not serious and can reasonably be ignored.

Correlated errors

We assume that the errors are uncorrelated but for temporally or spatially related data this may well be untrue. For this type of data, it is wise to check the uncorrelated assumption.

- Plot $\hat{\varepsilon}$ against time.
- Use formal tests like the Durbin-Watson or the run test.

If you do have correlated errors, you can use GLS (Chapter 5). This does require that you know Σ or more usually that you can estimate it. In the latter case, an iterative fitting procedure will be necessary as in IRWLS. Such problems are common in Econometrics.

Diagnostics in R

```
fit <- lm(sr ~ pop15 + pop75 + dpi + ddpi, savings,na.action="na.exclude")
summary(fit)
```

```
##
## Call:
## lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings,
##     na.action = "na.exclude")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2422 -2.6857 -0.2488  2.4280  9.7509
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
## pop15       -0.4611931  0.1446422  -3.189 0.002603 **
## pop75       -1.6914977  1.0835989  -1.561 0.125530
## dpi         -0.0003369  0.0009311  -0.362 0.719173
## ddpi         0.4096949  0.1961971   2.088 0.042471 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.803 on 45 degrees of freedom
## Multiple R-squared:  0.3385, Adjusted R-squared:  0.2797
## F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

Diagnostics in R

The `fortify()` method from the `ggplot2` package gives us: $\hat{\varepsilon}$ (`.resid`), h_i (`.hat`), r_i (`.stdresid`) and D_i (`.cooks`):

```
library(ggplot2) # you need to install first if you have not done that already
diag<-fortify(fit)
head(diag)
```

```
##              sr pop15 pop75     dpi ddpi       .hat   .sigma      .cooksd
## Australia 11.43 29.35  2.87 2329.68 2.87 0.06771343 3.843274 0.0008035888
## Austria   12.07 23.32  4.41 1507.99 3.93 0.12038393 3.844361 0.0008175997
## Belgium   13.17 23.80  4.43 2108.47 3.82 0.08748248 3.829661 0.0071546738
## Bolivia    5.75 41.89  1.67  189.13 0.22 0.08947114 3.844055 0.0007278744
## Brazil    12.88 42.19  0.83  728.47 4.56 0.06955944 3.805340 0.0140273514
## Canada     8.79 31.72  2.85 2982.88 2.43 0.15840239 3.845285 0.0003106199
##              .fitted     .resid   .stdresid
## Australia 10.566420  0.8635798  0.23520105
## Austria   11.453614  0.6163860  0.17282943
## Belgium   10.951042  2.2189579  0.61085760
## Bolivia    6.448319 -0.6983191 -0.19245030
## Brazil     9.327191  3.5528094  0.96858807
## Canada     9.106892 -0.3168924 -0.09083873
```

Diagnostics in R

The `rstudent()` method gives us the Jackknife residuals (t_i):

```
head(rstudent(fit))
```

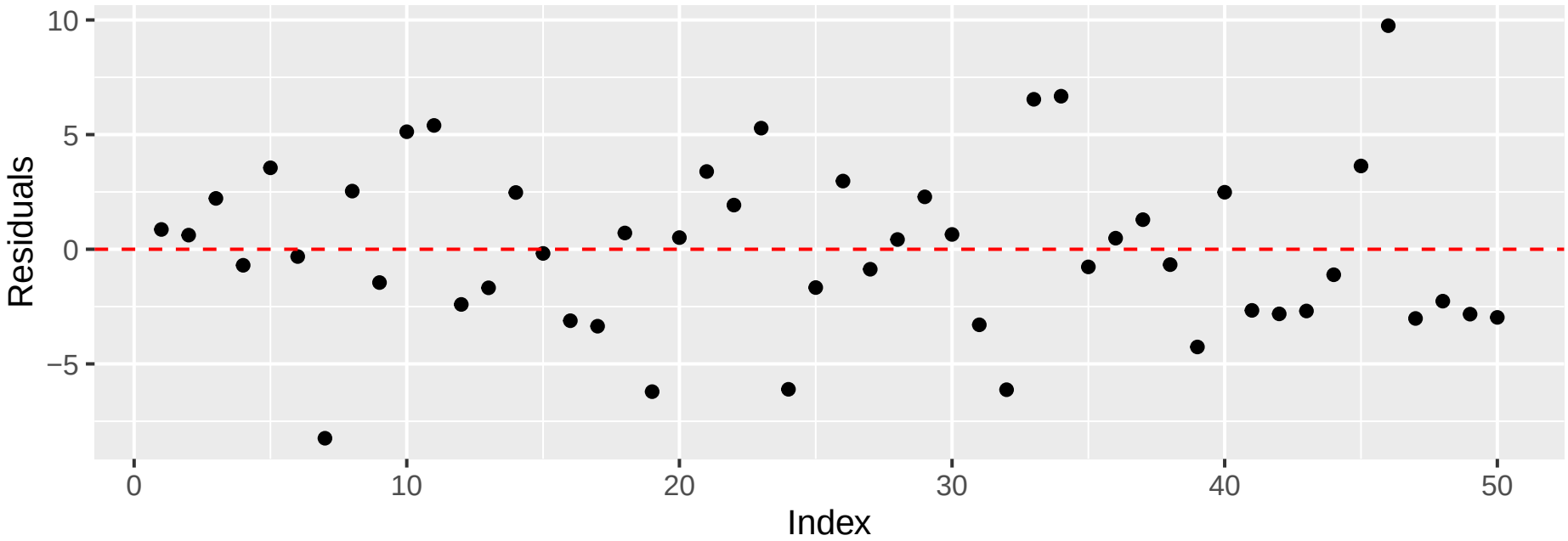
```
##    Australia      Austria      Belgium      Bolivia       Brazil       Canada
##  0.23271611  0.17095506  0.60655220 -0.19037831  0.96790816 -0.08983197
```

Add jackknife residuals to diag dataframe

```
diag$.jack<-rstudent(fit) # add jackknife residuals to diag dataframe
```

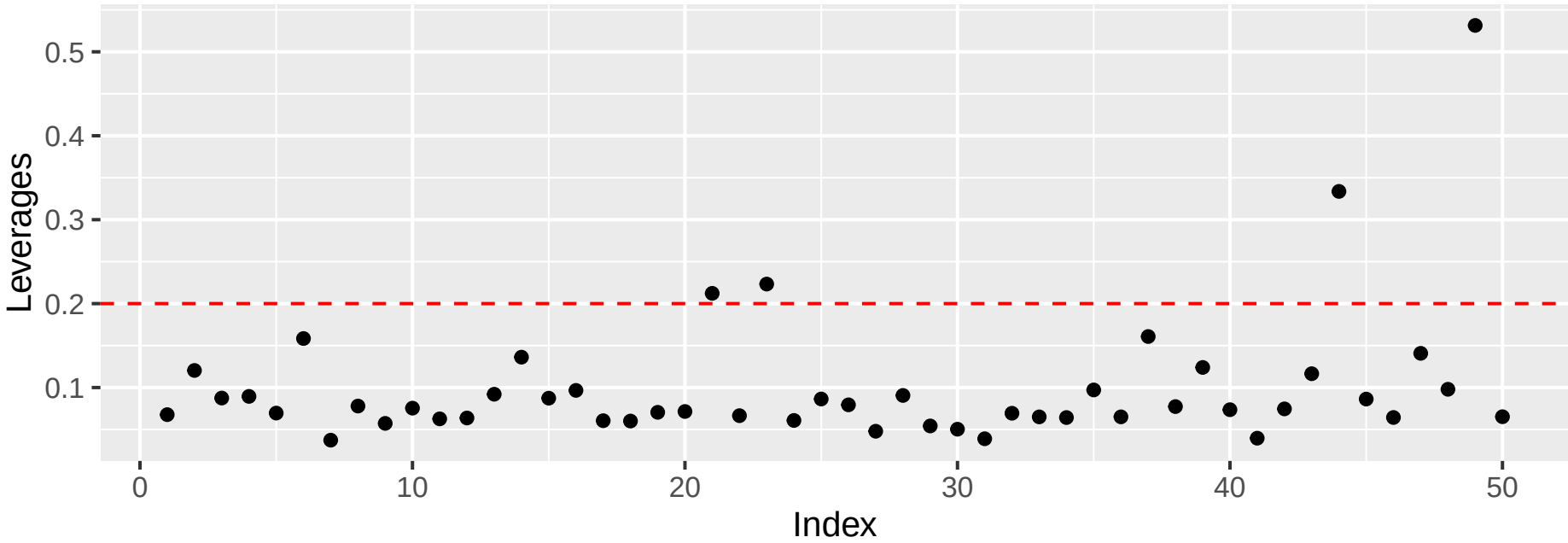
Index plot of residuals

```
p1<-ggplot(diag, aes(x=seq(1:length(.resid)),y=.resid))+geom_point()
    p1<-p1+geom_hline(yintercept=0, col="red", linetype="dashed")
    p1<-p1+xlab("Index")+ylab("Residuals")
    p1
```



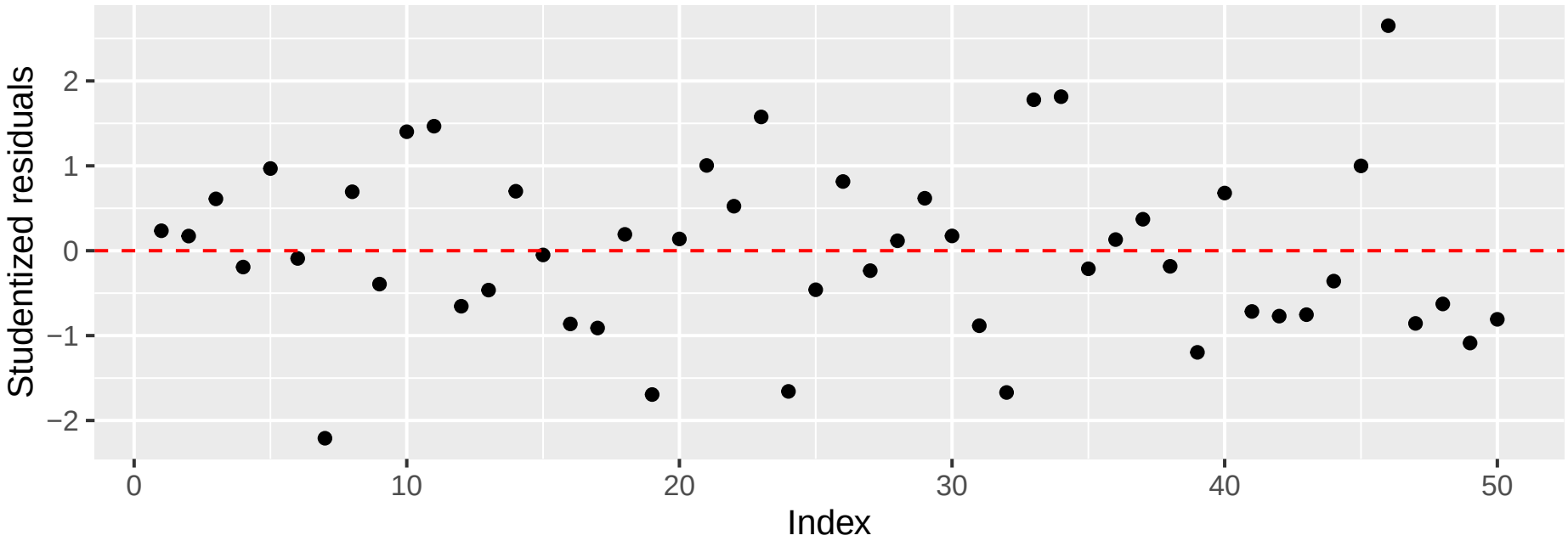
Index plot of leverages

```
p<-length(coef(fit))
n<-length(fitted(fit))
p2<-ggplot(diag, aes(x=seq(1:length(.hat)),y=.hat))+geom_point()
    p2<-p2+geom_hline(yintercept=2*p/n, col="red", linetype="dashed")
    p2<-p2+xlab("Index")+ylab("Leverages")
    p2
```



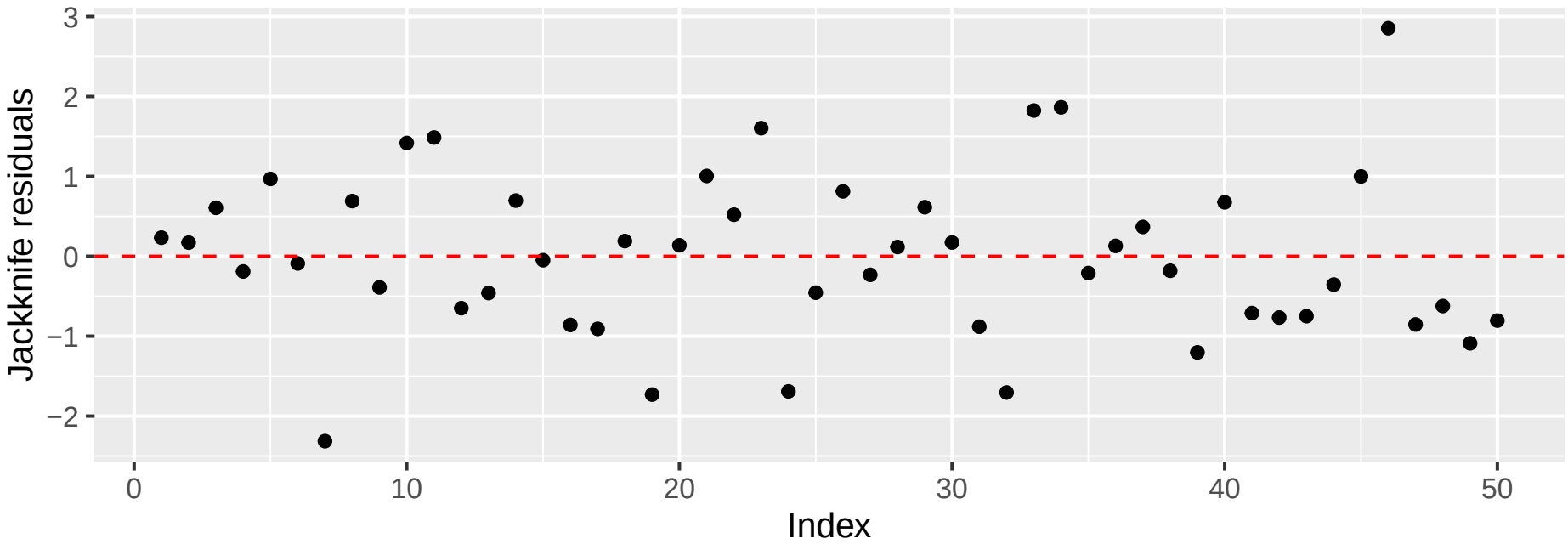
Index plot of studentized residuals

```
p3<-ggplot(diag, aes(x=seq(1:length(.stdresid)),y=.stdresid))+geom_point()
    p3<-p3+geom_hline(yintercept=0, col="red", linetype="dashed")
    p3<-p3+xlab("Index")+ylab("Studentized residuals")
    p3
```



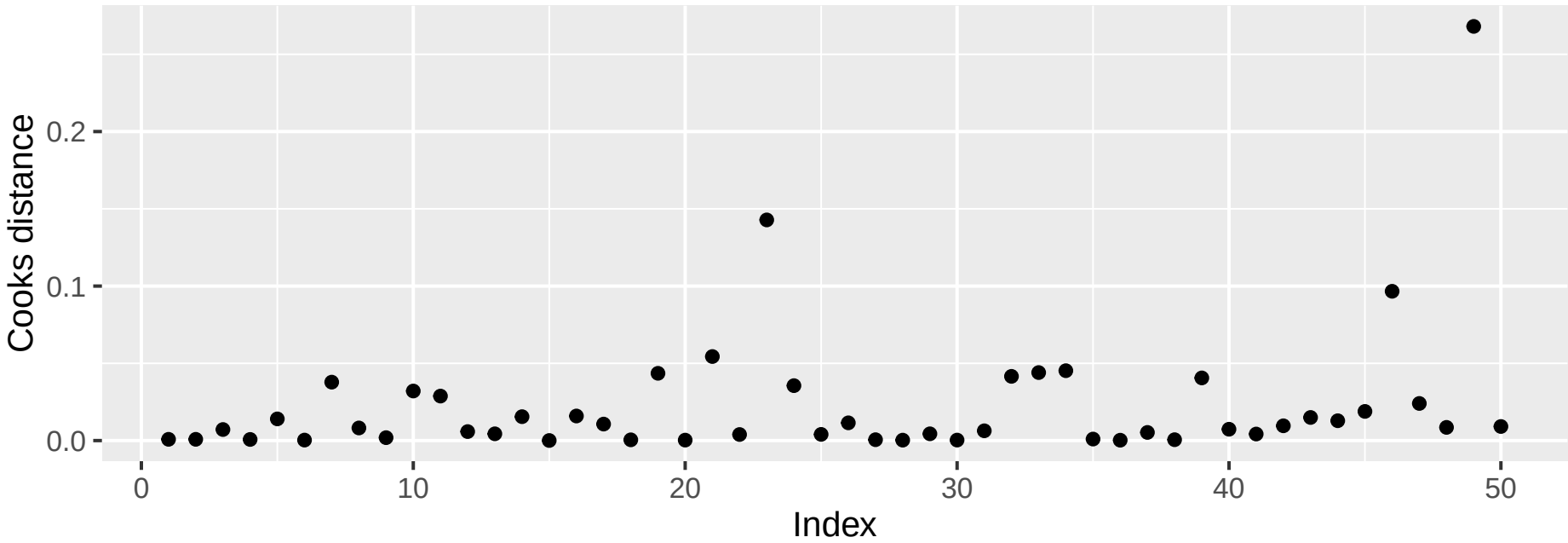
Index plot of jackknife residuals

```
p4<-ggplot(diag, aes(x=seq(1:length(.jack)),y=.jack))+geom_point()
    p4<-p4+geom_hline(yintercept=0, col="red", linetype="dashed")
    p4<-p4+xlab("Index")+ylab("Jackknife residuals")
    p4
```



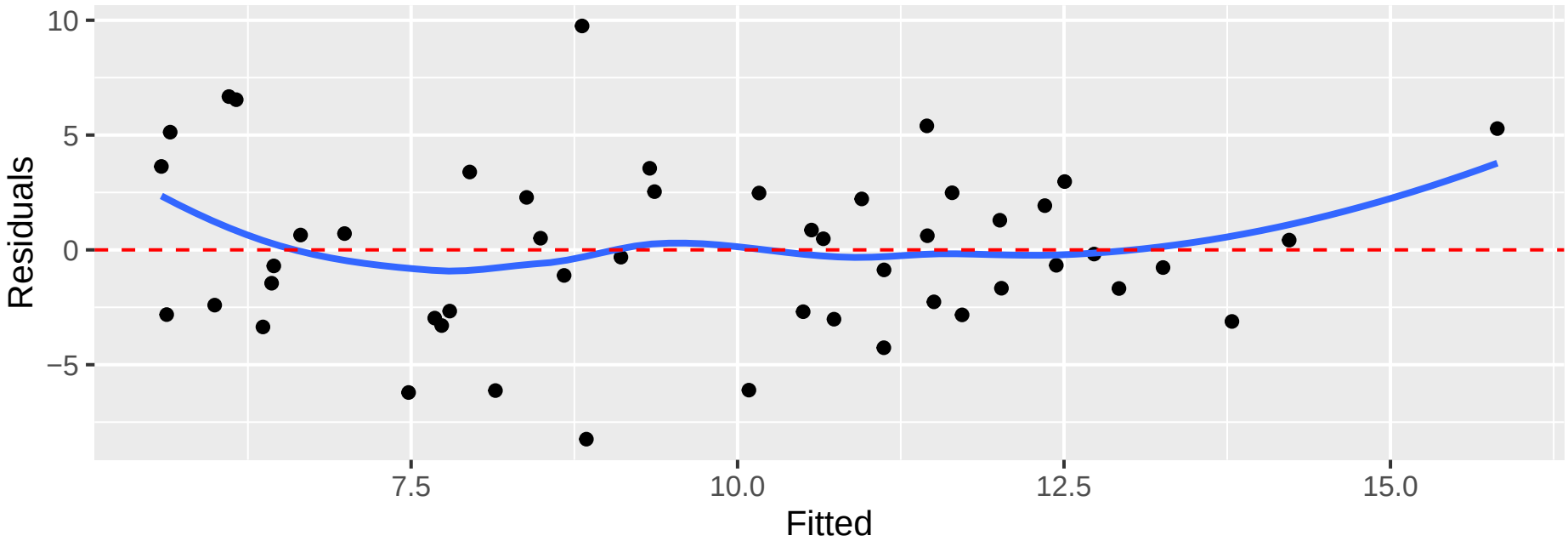
Cooks distance

```
p5<-ggplot(diag, aes(x=seq(1:length(.cooksd)),y=.cooksd))+geom_point()
    p5<-p5+xlab("Index")+ylab("Cooks distance")
    p5
```



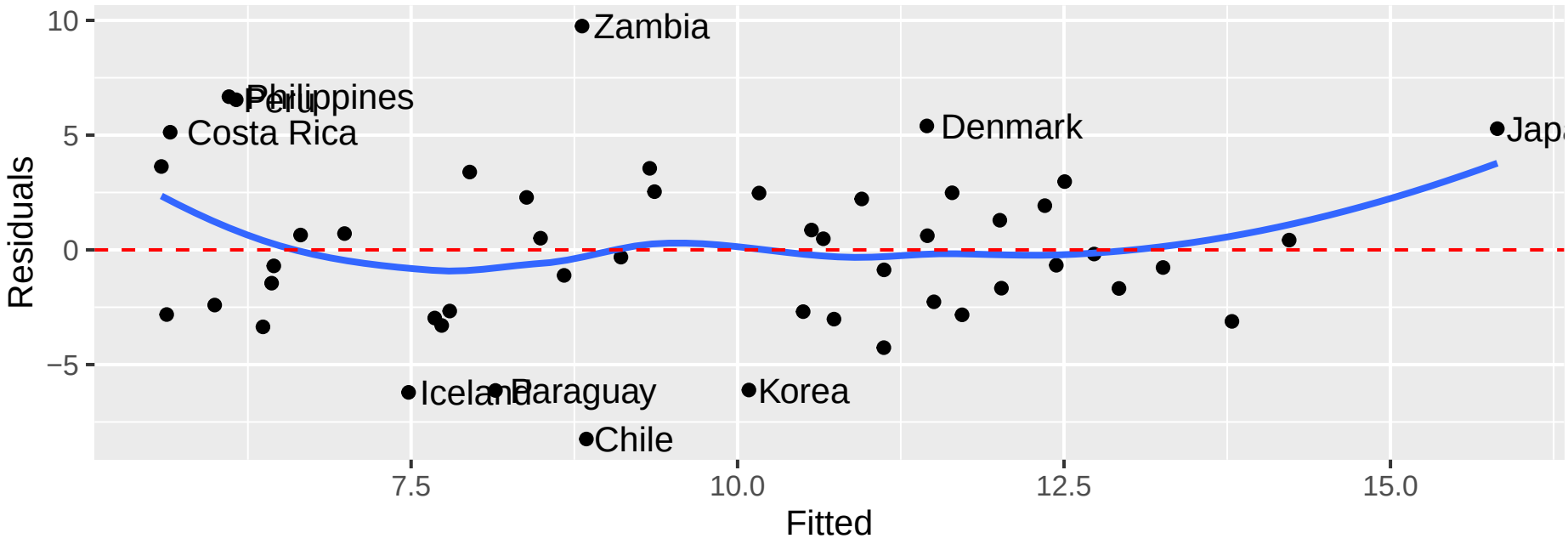
Residual plots

```
p6<-ggplot(diag, aes(x=.fitted,y=.resid))+geom_point()
    p6<-p6+stat_smooth(method="loess",se=F)+
      geom_hline(yintercept=0, col="red", linetype="dashed")
    p6<-p6+xlab("Fitted")+ylab("Residuals")
    p6
```



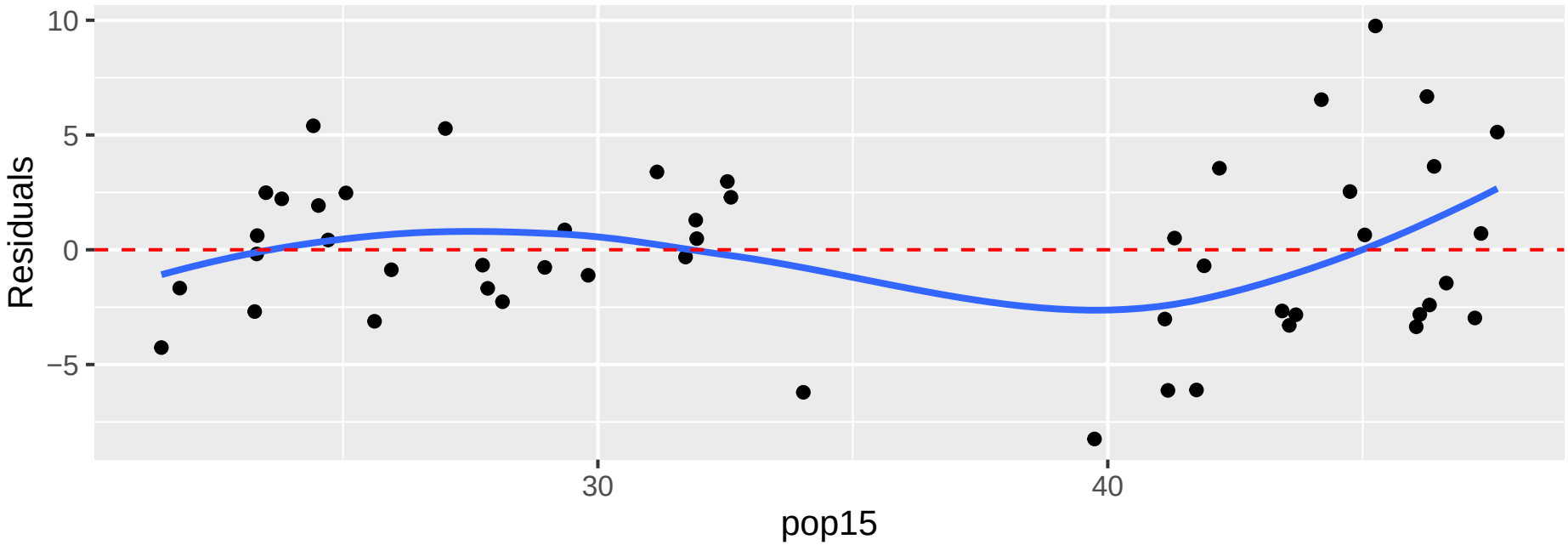
Locate point on the graph

```
p6 + geom_text(aes(label=ifelse(abs(.resid)>5,row.names(savings),"")),
               hjust=-0.1)
```



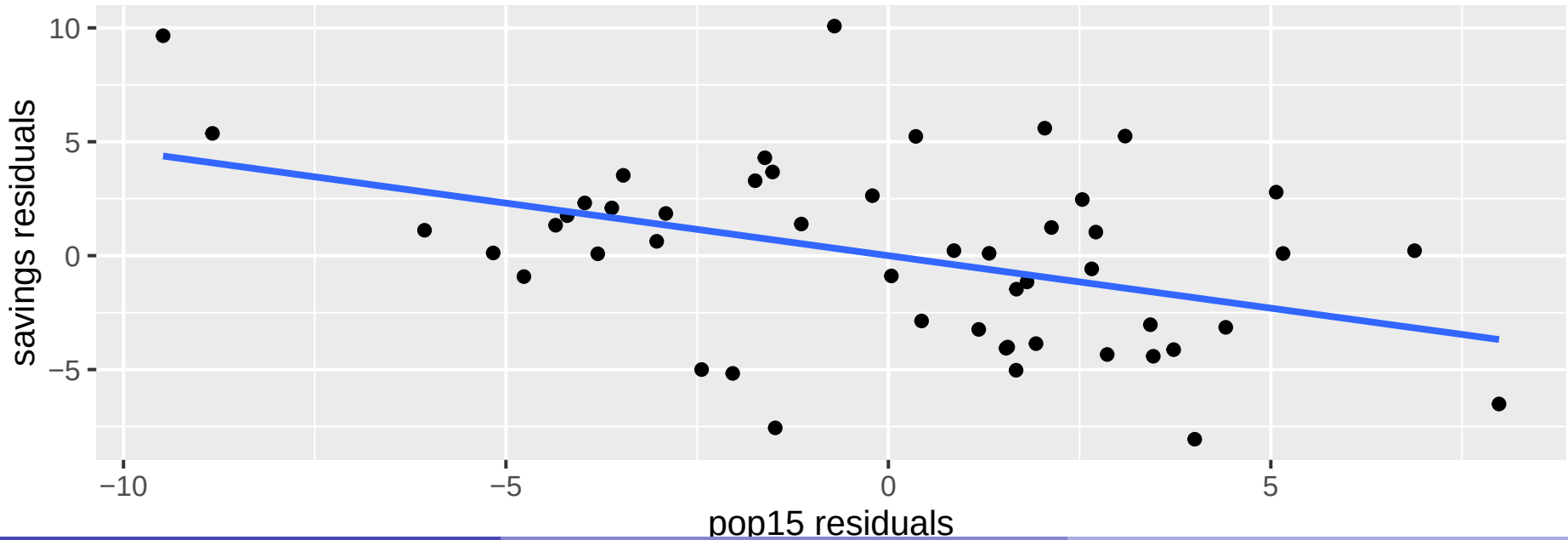
Residual plots

```
p7<-ggplot(diag, aes(x=pop15,y=.resid))+geom_point()
    p7<-p7+stat_smooth(method="loess",se=F)+
      geom_hline(yintercept=0, col="red", linetype="dashed")
    p7<-p7+xlab("pop15")+ylab("Residuals")
    p7
```



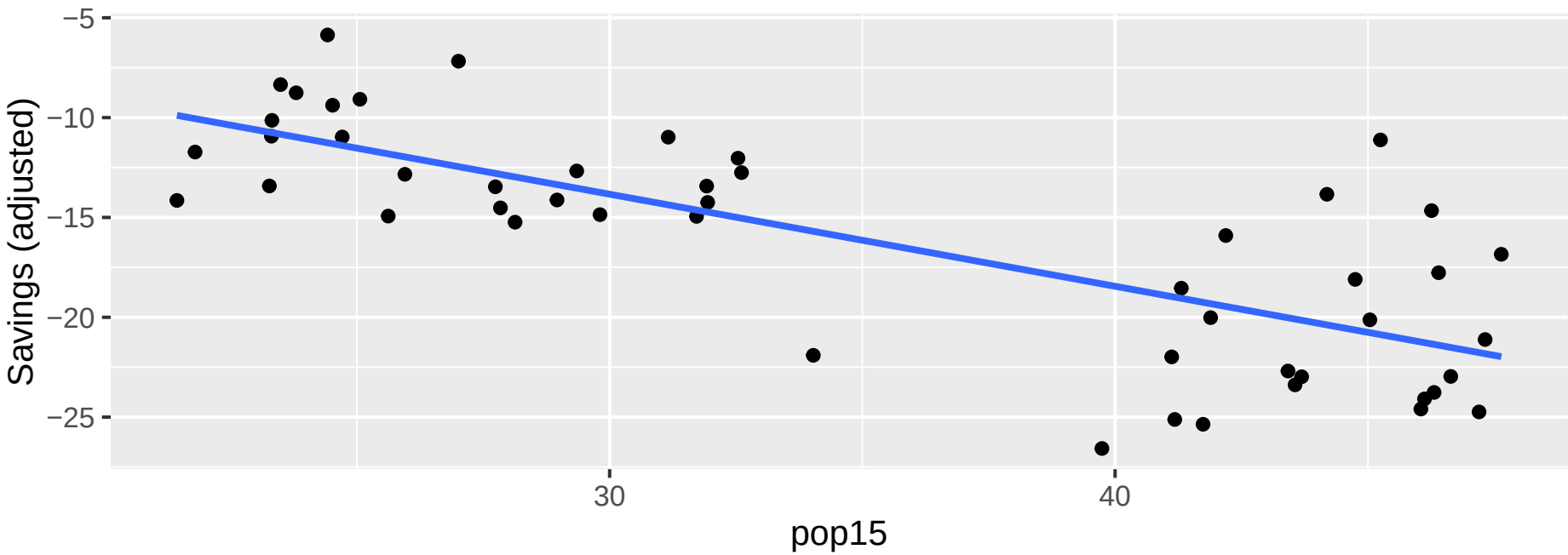
Partial regression plot

```
d<-lm(sr~ pop75 + dpi + ddpi,savings)$res
m<-lm(pop15~pop75 + dpi + ddpi,savings)$res
pr<-data.frame(d=d,m=m)

p8<-ggplot(pr, aes(x=m,y=d))+geom_point()
    p8<-p8+stat_smooth(method="lm",se=F)
    p8<-p8+xlab("pop15 residuals")+ylab("savings residuals")
    p8
```



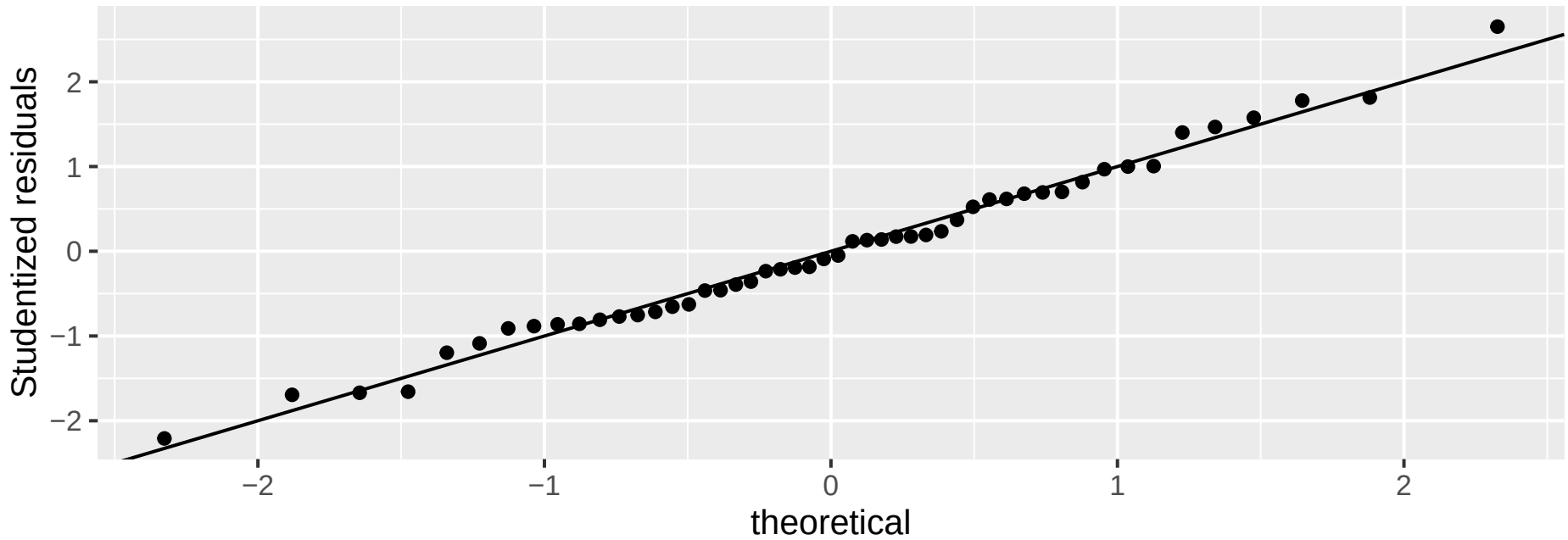
Partial residuals

```
savings$partPop15<-fit$res+fit$coef['pop15']*savings$pop15

p9<-ggplot(savings, aes(x=pop15,y=partPop15))+geom_point()
    p9<-p9+stat_smooth(method="lm",se=F)
    p9<-p9+xlab("pop15")+ylab("Savings (adjusted)")
    p9
```



Q-Q plot of studentized residuals

```
p10<-ggplot(diag, aes(sample = .stdresid)) + stat_qq()
    p10 <- p10 + geom_abline(slope=1)
    p10 <- p10 + ylab("Studentized residuals")
    p10
```



Histogram of residuals

```
p11<-ggplot(diag, aes(.resid)) + geom_histogram(binwidth=2)
    p11 <- p11 + xlab("Residuals")
    p11
```