Regression Diagnostics, Part 2

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \epsilon_i \sim N(0, \sigma^2)$$

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y} \qquad E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta} \qquad Var[\hat{\boldsymbol{\beta}}] = \boldsymbol{\sigma}^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$$

In Part 1, we looked at the Normality and constant variance assumptions. Now, in Part 2, we look at unusual observations.

- **Leverage**
- **Outliers**
- **Influence**

Leverage

Recall,

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

Thus,

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

Now we define,

$$\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$$

Which we will refer to as the hat matrix. The hat matrix is used to project onto the subspace spanned by the columns of $\boldsymbol{X}$. (And is otherwise know as a projection matrix.)

The diagonal elements of this matrix are called the leverages.

$$\boldsymbol{H}_{ii} = h_i$$

Large values of h_i indicate extreme values in $\boldsymbol{X}$, which may influence regression. Note that leverages only depend on $\boldsymbol{X}$.

$$\sum h_i = p$$

Here, p is the number of βs. (Also the trace (and rank) of the hat matrix.)

A common check for a large leverage is to compare to $2 * \frac{p}{n} = 2\bar{p}$, two times the average leverage. A leverage larger than this is considered an observation to be aware of.

For simple linear regression, we have,

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SXX}$$

which suggests that the large leverages occur when x values are far from their mean. (Recall that the regression goes through the point $(\bar{x}, \bar{y})$.)

There are multiple ways to find leverages in R.

```
x <- c(2,6,8,8,12,16,20,20,22,26)
y <- c(58,105,88,118,117,137,157,169,149,202)
X <- cbind(rep(1,10), x)

H <- X %*% solve(t(X)%*%X) %*% t(X)
diag(H)
```

```
##  [1] 0.3535211 0.2126761 0.1633803 0.1633803 0.1070423 0.1070423 0.1633803
##  [8] 0.1633803 0.2126761 0.3535211
```

```
sum(diag(H))
```

```
## [1] 2
```

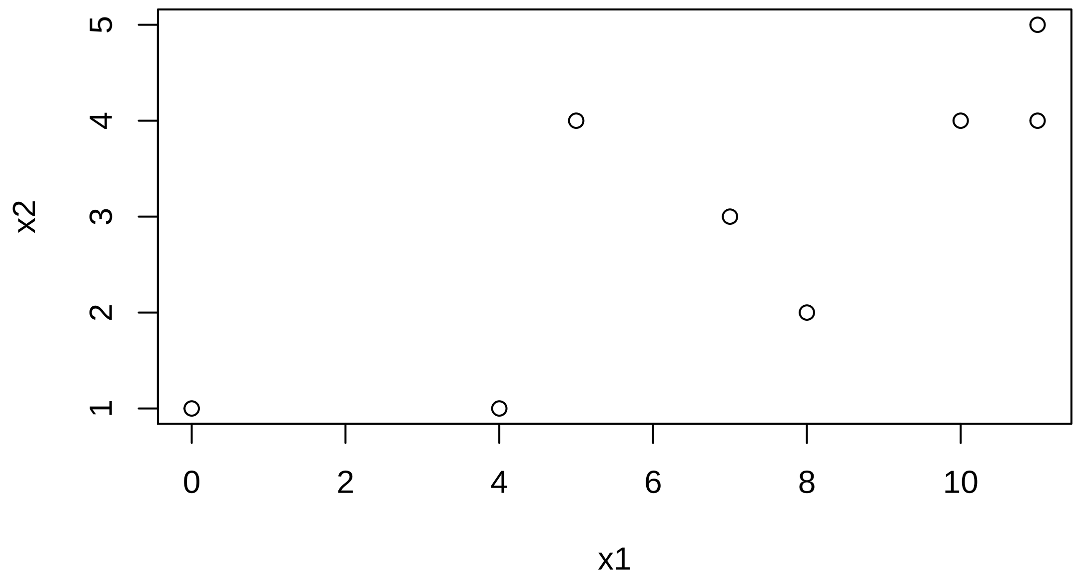
```
fit <- lm(y ~ x)
hatvalues(fit)
```

```
##         1         2         3         4         5         6         7 
## 0.3535211 0.2126761 0.1633803 0.1633803 0.1070423 0.1070423 0.1633803 
##         8         9        10 
## 0.1633803 0.2126761 0.3535211
```

In the next example we will look at how changing the y values for points with different leverages affect regression.

```
x1 <- c( 0,11,11, 7, 4,10, 5, 8)
x2 <- c( 1, 5, 4, 3, 1, 4, 4, 2)
y  <- c(11,15,13,14, 0,19,16, 8)

plot(x1,x2)
```




```
X <- cbind( rep(1,8), x1, x2 )
H <- X %*% solve(t(X)%*%X) %*% t(X)
diag(H)
```

```
## [1] 0.6000 0.3750 0.2875 0.1250 0.4000 0.2125 0.5875 0.4125
```

```
sum(diag(H))
```

```
## [1] 3
```

```
2*mean(diag(H))
```

```
## [1] 0.75
```

```
min(diag(H))
```

```
## [1] 0.125
```

```
max(diag(H))
```

```
## [1] 0.6
```

```
fit <- lm(y ~ x1 + x2)
hatvalues(fit)
```

```
##      1      2      3      4      5      6      7      8
## 0.6000 0.3750 0.2875 0.1250 0.4000 0.2125 0.5875 0.4125
```

```r
lm(y ~ x1 + x2)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Coefficients:
## (Intercept)           x1           x2
##         3.7         -0.7          4.4
```

The first observation has large leverage. Note how changing its y value has a large effect on the regression.

```r
y[1] <- 20
lm(y ~ x1 + x2)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Coefficients:
## (Intercept)           x1           x2
##       8.875       -1.375        4.625
```

The fourth observation has small leverage. (Why is that?) Note that when we change its y value, the regression is barely changed. (The y value for the first observation is first changed back.)

```r
y[1] <- 11
y[4] <- 30
lm(y ~ x1 + x2)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Coefficients:
## (Intercept)           x1           x2
##         5.7         -0.7          4.4
```

```r
mean(x1)
```

```
## [1] 7
```

```r
mean(x2)
```

```
## [1] 3
```

Outliers

Outliers are points which do not fit the model well. They may or may not have a large affect on the model. To identify outliers, we will look for observations with large residuals.

Note,

$$\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$$

Then, under the assumptions of linear regression,

$$Var(e_i) = (1 - h_i)\sigma^2$$

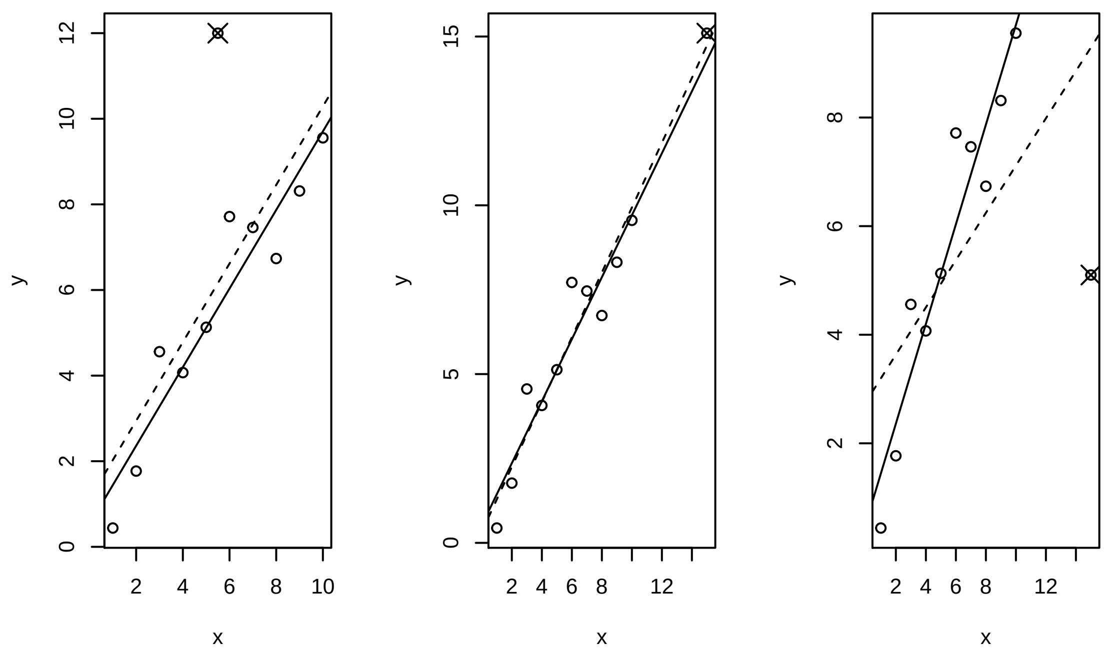
and thus

$$\hat{Var}(e_i) = (1 - h_i)s^2.$$

We can then look at the **standardized residuals** for each observation, $i = 1, 2, \ldots n$,

$$r_i = \frac{e_i}{\sqrt{1 - h_i}}.$$

The following three plots appear in the book. In each plot two regression are fit. First one without the point marked by an X, which can be seen as the solid line. The second regression includes the point with the X, and is seen as a dashed line.



This should convince us that unusual observations can influence the regression, in particular, sometimes unusual observations pull the regression closer to them, as seen in the third plot. For this reason, we will look at **studentized residuals**.

$$t_i = \frac{y_i - \hat{y_{(i)}}}{s_{(i)}(1 + x_i^T(\boldsymbol{X_{(i)}}^T\boldsymbol{X_{(i)}})^{-1}x_i)^{1/2}}$$

Each studentized residuals looks at the residual when we first fit a model without an observation, then predict that observation's y value, and compare it to its actual y value. By doing this, unusual observations have the large residuals we would expect, since they aren't pulling the model close to them. (Since we're omitting them when fitting the model.)

Studentized residuals, don't actually require fitting n regressions, instead there is an easier way to compute them,

$$t_i = \frac{e_i}{s_{(i)}(\sqrt{1-h_i})} = r_i\left(\frac{n-p-1}{n-p-r_i^2}\right)^{1/2} \sim t_{n-p-1}$$

which is a function of the standardized residuals.

Since the studentized residuals follow a t-distribution, we can use this to test for outliers. Call an observation an outlier if

$$t_i > t_{\alpha/2}(n-p-1) \quad \text{OR} \quad t_i < -t_{\alpha/2}(n-p-1).$$

The trouble here is that we are performing n test, so with a large n and a particular α we would expect a number of test to come back significant, despite the observation not being an outlier. We will modify our test using a **Bonferroni correction**, where we now call an observation an outlier if,

$$t_i > t_{\alpha/2n}(n-p-1) \quad \text{OR} \quad t_i < -t_{\alpha/2n}(n-p-1).$$

In the following example, we find observations with large leverage. We also find outliers with and without a Bonferroni correction. We find two outliers without the correction, and none with the correction.

```r
# load the savings data
library(faraway)
data(savings)
?savings

# fit a model with every predictor
mymodel <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data = savings)
summary(mymodel)
```

```
##
## Call:
## lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings)
##
## 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
```

```
# calculate leverages, find the ones we should look at
lev <- hatvalues(mymodel)
lev
```

```
##      Australia        Austria        Belgium        Bolivia         Brazil
##     0.06771343     0.12038393     0.08748248     0.08947114     0.06955944
##         Canada          Chile          China       Colombia     Costa Rica
##     0.15840239     0.03729796     0.07795899     0.05730171     0.07546780
##        Denmark        Ecuador        Finland         France        Germany
##     0.06271782     0.06372651     0.09204246     0.13620478     0.08735739
##         Greece      Guatamala       Honduras        Iceland          India
##     0.09662073     0.06049212     0.06008079     0.07049590     0.07145213
##        Ireland          Italy          Japan          Korea     Luxembourg
##     0.21223634     0.06651170     0.22330989     0.06079915     0.08634787
##          Malta         Norway    Netherlands    New Zealand      Nicaragua
##     0.07940290     0.04793213     0.09061400     0.05421789     0.05035056
##         Panama       Paraguay           Peru    Philippines       Portugal
##     0.03897459     0.06937188     0.06504891     0.06425415     0.09714946
##   South Africa South Rhodesia          Spain         Sweden    Switzerland
##     0.06510405     0.16080923     0.07732854     0.12398898     0.07359423
##         Turkey        Tunisia United Kingdom  United States      Venezuela
##     0.03964224     0.07456729     0.11651375     0.33368800     0.08628365
##         Zambia        Jamaica        Uruguay          Libya       Malaysia
##     0.06433163     0.14076016     0.09794717     0.53145676     0.06523300
```

```
lev_mean <- mean(lev)
sum(lev > 2 * lev_mean)
```

```
## [1] 4
```

```
lev[lev > 2 * lev_mean]
```

```
##       Ireland         Japan United States         Libya
##     0.2122363     0.2233099     0.3336880     0.5314568
```

```
max(lev)
```

```
## [1] 0.5314568
```

```
# calculate the studentized residuals
sresid <- rstudent(mymodel)
sresid
```

```
##      Australia        Austria        Belgium        Bolivia         Brazil
##     0.23271611     0.17095506     0.60655220    -0.19037831     0.96790816
##         Canada          Chile          China       Colombia     Costa Rica
##    -0.08983197    -2.31342946     0.69048169    -0.38946778     1.41731062
##        Denmark        Ecuador        Finland         France        Germany
##     1.48644473    -0.64957871    -0.45986445     0.69640933    -0.04918692
##         Greece      Guatamala       Honduras        Iceland          India
##    -0.85967533    -0.90854545     0.19051919    -1.73119989     0.13729730
##        Ireland          Italy          Japan          Korea     Luxembourg
##     1.00485886     0.52015744     1.60321582    -1.69103214    -0.45560591
##          Malta         Norway    Netherlands    New Zealand      Nicaragua
##     0.81227407    -0.23247367     0.11605663     0.61373189     0.17254242
##         Panama       Paraguay           Peru    Philippines       Portugal
##    -0.88147653    -1.70488128     1.82391409     1.86382587    -0.21040432
##   South Africa South Rhodesia          Spain         Sweden    Switzerland
##     0.12996586     0.36714512    -0.18175853    -1.20293404     0.67532922
##         Turkey        Tunisia United Kingdom  United States      Venezuela
##    -0.71138840    -0.76677907    -0.74959873    -0.35461507     0.99932569
##         Zambia        Jamaica        Uruguay          Libya       Malaysia
##     2.85355834    -0.85376418    -0.62253411    -1.08930326    -0.80489153
```

```r
hist(sresid)

# find the outliers
n <- length(sresid)
p <- length(mymodel$coefficients)
df <- n - p - 1
alpha <- 0.05

# without bonferroni
crit <- qt(1 - 0.05/2,df)
sum(abs(sresid) > crit)
```

```
## [1] 2
```

```r
sresid[abs(sresid) > crit]
```

```
##     Chile    Zambia
## -2.313429  2.853558
```

```r
max(abs(sresid))
```

```
## [1] 2.853558
```

```r
# with bonferroni
crit <- qt(1 - (0.05/2)/n,df)
sum(abs(sresid) > crit)
```

```
## [1] 0
```

```r
sresid[abs(sresid) > crit]
```

```
## named numeric(0)
```

```r
max(abs(sresid))
```

```
## [1] 2.853558
```

```r
# compare with/withou bonferroni, and fdr
pvals <- pt(-abs(sresid),df)*2
padjb <- p.adjust(pvals, method = "bonferroni")
padjf <- p.adjust(pvals, method = "fdr")
cbind(pvals,padjb,padjf)
```

```
##                       pvals     padjb     padjf
## Australia       0.817061004 1.0000000 0.9459747
## Austria         0.865042853 1.0000000 0.9459747
## Belgium         0.547265140 1.0000000 0.8826857
## Bolivia         0.849888351 1.0000000 0.9459747
## Brazil          0.338380698 1.0000000 0.8826857
## Canada          0.928828419 1.0000000 0.9477841
## Chile           0.025433198 1.0000000 0.6358299
## China           0.493517797 1.0000000 0.8826857
## Colombia        0.698808873 1.0000000 0.9459747
## Costa Rica      0.163434746 1.0000000 0.8171737
## Denmark         0.144292553 1.0000000 0.8016253
## Ecuador         0.519341286 1.0000000 0.8826857
## Finland         0.647877878 1.0000000 0.9459747
## France          0.489835308 1.0000000 0.8826857
## Germany         0.960992937 1.0000000 0.9609929
## Greece          0.394627854 1.0000000 0.8826857
## Guatamala       0.368539833 1.0000000 0.8826857
## Honduras        0.849778638 1.0000000 0.9459747
## Iceland         0.090423702 1.0000000 0.6993102
## India           0.891422159 1.0000000 0.9459747
## Ireland         0.320459192 1.0000000 0.8826857
## Italy           0.605561320 1.0000000 0.9459747
## Japan           0.116042382 1.0000000 0.7252649
## Korea           0.097903426 1.0000000 0.6993102
## Luxembourg      0.650913948 1.0000000 0.9459747
## Malta           0.421007458 1.0000000 0.8826857
## Norway          0.817248104 1.0000000 0.9459747
## Netherlands     0.908135746 1.0000000 0.9459747
## New Zealand     0.542552654 1.0000000 0.8826857
## Nicaragua       0.863802347 1.0000000 0.9459747
## Panama          0.382849962 1.0000000 0.8826857
## Paraguay        0.095268739 1.0000000 0.6993102
## Peru            0.074961446 1.0000000 0.6993102
## Philippines     0.069027538 1.0000000 0.6993102
## Portugal        0.834323467 1.0000000 0.9459747
## South Africa    0.897185790 1.0000000 0.9459747
## South Rhodesia  0.715270737 1.0000000 0.9459747
```

```
## Spain          0.856606997 1.0000000 0.9459747
## Sweden         0.235434928 1.0000000 0.8826857
## Switzerland    0.503000424 1.0000000 0.8826857
## Turkey         0.480598229 1.0000000 0.8826857
## Tunisia        0.447307674 1.0000000 0.8826857
## United Kingdom 0.457485731 1.0000000 0.8826857
## United States  0.724572020 1.0000000 0.9459747
## Venezuela      0.323101241 1.0000000 0.8826857
## Zambia         0.006566663 0.3283332 0.3283332
## Jamaica        0.397859978 1.0000000 0.8826857
## Uruguay        0.536803879 1.0000000 0.8826857
## Libya          0.281950432 1.0000000 0.8826857
## Malaysia       0.425210344 1.0000000 0.8826857
```

```
# use a package to find the outliers
library(car)
```


```
outlierTest(mymodel)
```

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##        rstudent unadjusted p-value Bonferonni p
## Zambia 2.853558          0.0065667      0.32833
```

Influence

As we have seen in the plots from the book, some outliers only change the regression a small amount (plot 1) and some outliers have a large effect on the regression. (plot 3) Observations that fall into the later category we will call **influential**.

A common measure of influence is **Cook's Distance**,

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

which is a function of both leverage and standardized residuals. A Cook's Distance is considered large if $D_i > 4/n$, and an observation with a large Cook's Distance is called influential.

```
# calculate cook's distances, find the ones we should look at
cook <- cooks.distance(mymodel)
cook[cook > 4/n]
```

```
##      Japan     Zambia      Libya
## 0.14281625 0.09663275 0.26807042
```

```
# fit a model without the observation that has the largest cook's distance
modified <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data = savings, subset = cook < max(cook))
summary(modified)
```

```
##
## Call:
## lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings,
##     subset = cook < max(cook))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0699 -2.5408 -0.1584  2.0934  9.3732
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.5240460  8.2240263   2.982  0.00465 **
## pop15       -0.3914401  0.1579095  -2.479  0.01708 *
## pop75       -1.2808669  1.1451821  -1.118  0.26943
## dpi         -0.0003189  0.0009293  -0.343  0.73312
## ddpi         0.6102790  0.2687784   2.271  0.02812 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.795 on 44 degrees of freedom
## Multiple R-squared:  0.3554, Adjusted R-squared:  0.2968
## F-statistic: 6.065 on 4 and 44 DF,  p-value: 0.0005617
```

```
# compare to the original model
coef(mymodel)
```

```
##   (Intercept)         pop15         pop75           dpi          ddpi
## 28.5660865407 -0.4611931471 -1.6914976767 -0.0003369019  0.4096949279
```

```r
coef(modified)
```

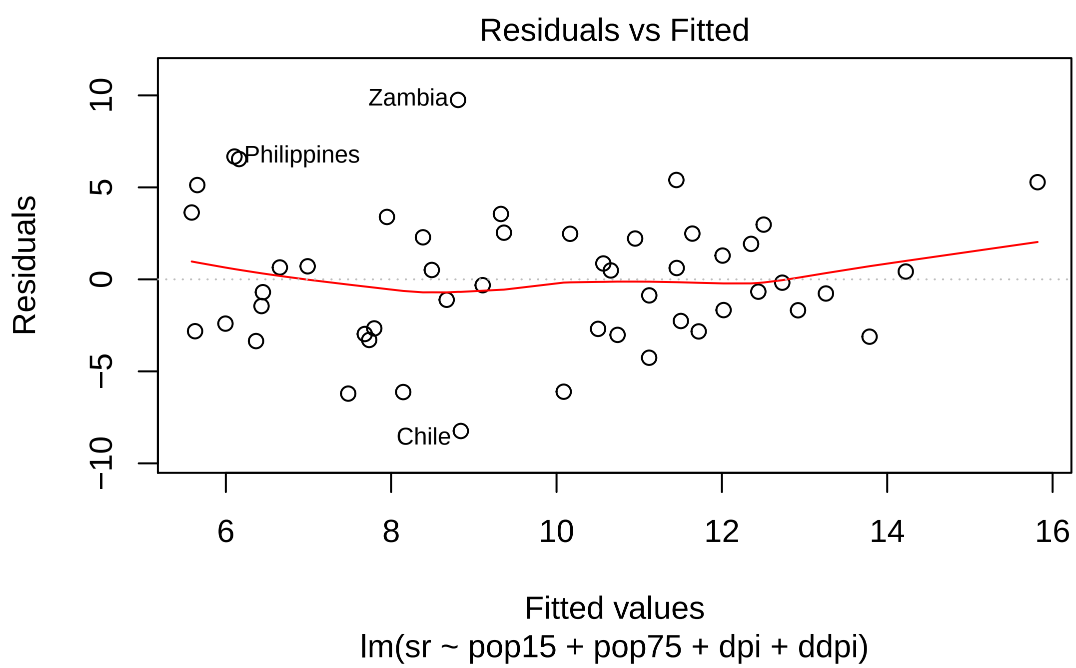
```
##   (Intercept)         pop15         pop75           dpi          ddpi
## 24.5240459788 -0.3914401268 -1.2808669233 -0.0003189001  0.6102790264
```

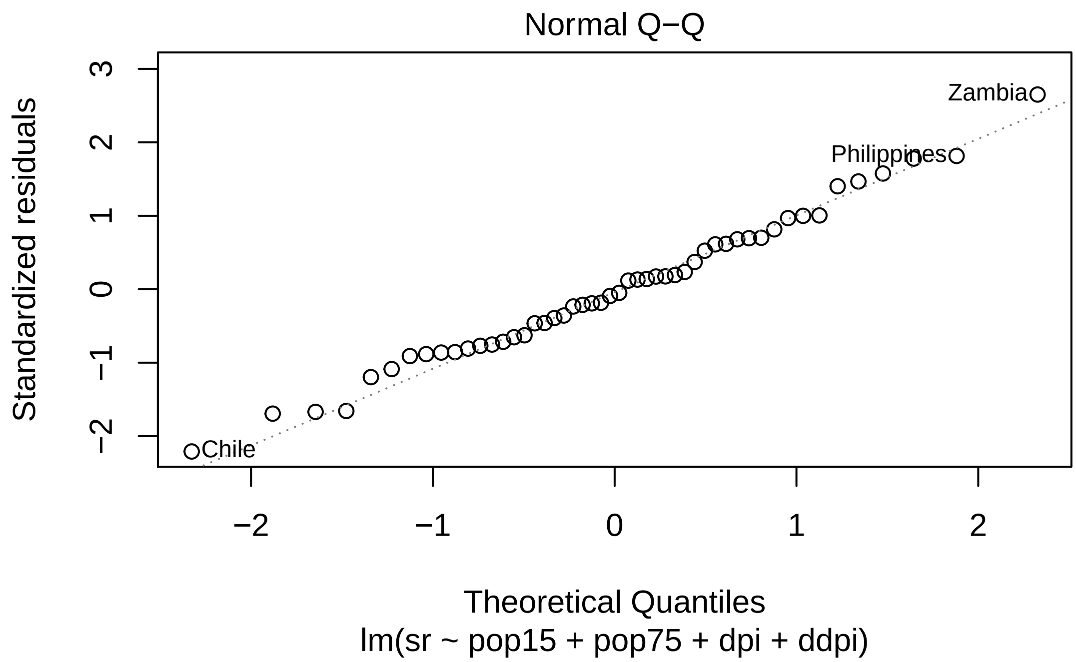
```r
(coef(mymodel) - coef(modified)) / coef(mymodel)
```

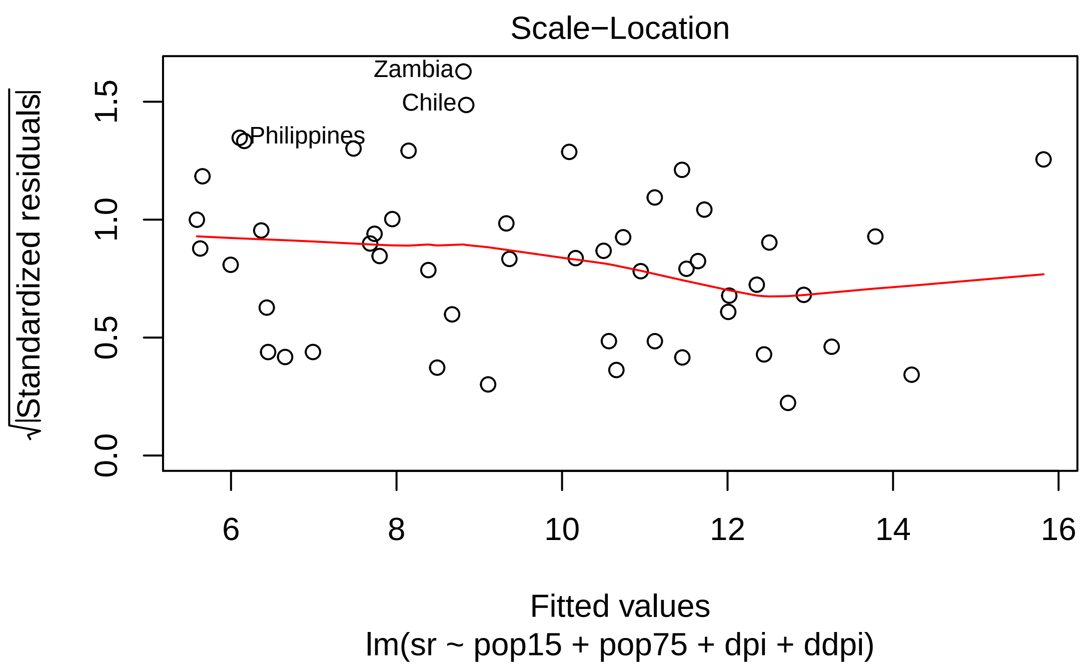
```
## (Intercept)       pop15       pop75         dpi        ddpi
##  0.14149788  0.15124470  0.24276164  0.05343314 -0.48959380
```

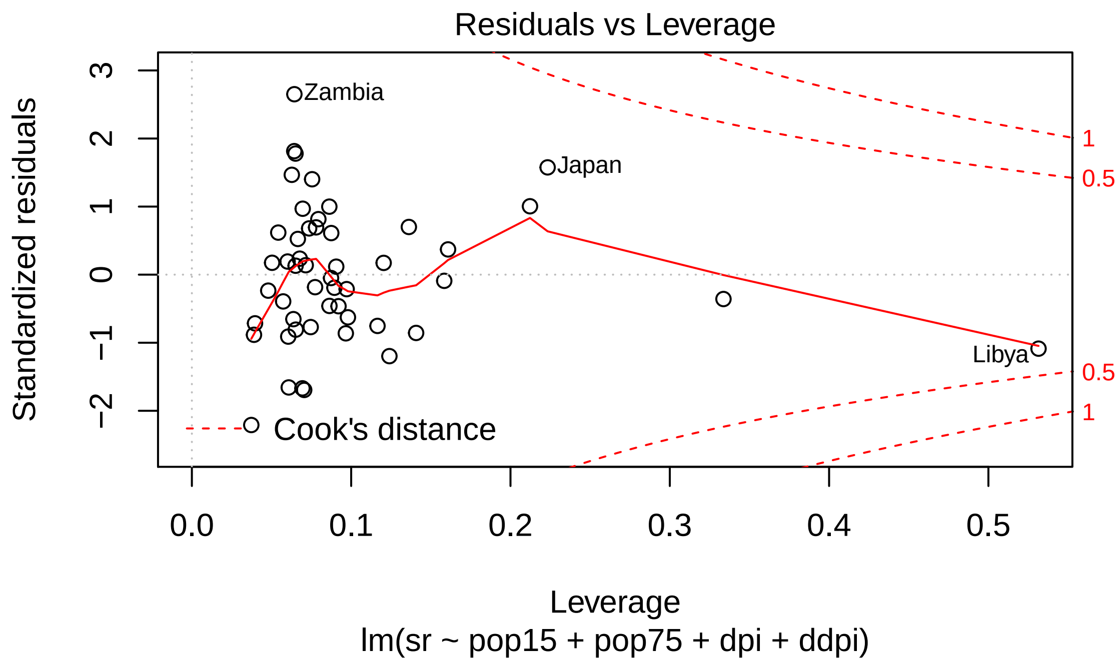
A nice feature of the `lm` function in `R` is the resulting plots when we call `plot` on an object created using `lm`. Doing so outputs a series of plots. The first two are familiar to us. The first is the fitted vs residuals plot. `R` adds a red line which is a moving average, and labels some of the important observations. The second plot is a QQ-plot. The third plot is another fitted vs residuals plot, this time with standardized residuals. The fourth and most interesting plot gives us the Cook's Distance for each point.

```r
#view all diagnostic plots
plot(mymodel)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(sr ~ pop15 + pop75 + dpi + ddpi)
Zambia
Philippines
Chile
3
2
1
0
−1
−2
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
Fitted values
lm(sr ~ pop15 + pop75 + dpi + ddpi)
Zambia
Chile
Philippines
1.5
1.0
0.5
0.0
6
8
10
12
14
16


Residuals vs Leverage
Standardized residuals
Zambia
Japan
Libya
Cook's distance
1
0.5
0.5
1
−2
−1
0
1
2
3
0.0
0.1
0.2
0.3
0.4
0.5
Leverage
lm(sr ~ pop15 + pop75 + dpi + ddpi)