Chapter 7

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 improvements. This is a hands-on process.

7.1 Residuals and Leverage

We start with some basic diagnostic quantities - the residuals and the leverages. Recall that $\hat{y} = X(X^TX)^{-1}X^Ty = Hy$ where H is the hat-matrix. Now

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

So var $\hat{\varepsilon}$ = var $(I-H)\varepsilon = (I-H)\sigma^2$ assuming that var $\varepsilon = \sigma^2 I$. We see that although the errors may have equal variance and be uncorrelated the residuals do not.

$h_i = H_{ii}$ are called *leverages* and are useful diagnostics. We see that var $\hat{\varepsilon}_i = \sigma^2(1-h_i)$ so that a large leverage for h_i will make var $\hat{\varepsilon}_i$ small — in other words the fit will be "forced" to be close to y_i. The h_i depends only on X — knowledge of y is required for a full interpretation. Some facts:

$$\sum_i h_i = p \qquad h_i \geq 1/n \quad \forall i$$

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. Large values of h_i are due to extreme values in X. h_i corresponds to a Mahalanobis distance defined by X which is $(x-\bar{x})^T\hat{\Sigma}^{-1}(x-\bar{x})$ where $\hat{\Sigma}$ is the estimated covariance of X.

Also notice that var $\hat{y}$ = var $(Hy) = H\sigma^2$ so vâr $\hat{y}_i = h_i\hat{\sigma}^2$

We'll use the savings dataset as an example here. First fit the model and make an index plot of the residuals:

```
> data(savings)
> g <- lm(sr ~ pop15 + pop75 + dpi + ddpi, savings)
> plot(g$res,ylab="Residuals",main="Index plot of residuals")
```

The plot is shown in the first panel of Figure 7.1

We can find which countries correspond to the largest and smallest residuals:

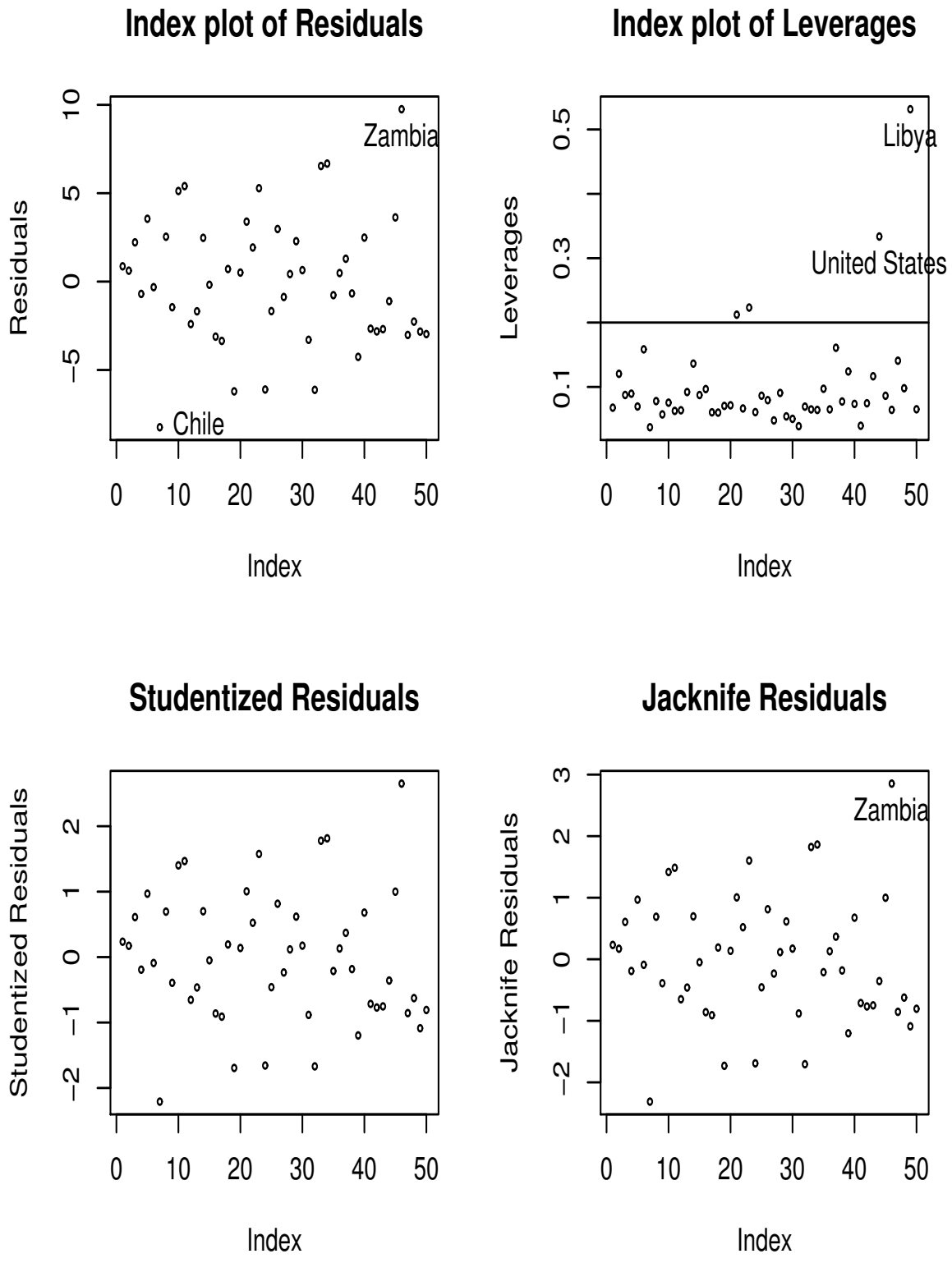

Figure 7.1: Residuals and leverages for the savings data

```
> sort(g$res)[c(1,50)]
  Chile  Zambia
-8.2422  9.7509
```

or by using the `identify()` function. We first make up a character vector of the country names using `row.names()` which gets the row names from the data frame.

```
> countries <- row.names(savings)
> identify(1:50,g$res,countries)
```

Click on the left mouse button next to the points you are interested in to identify them. When you are done, click on the middle (if not available, the right) mouse button. I have identified Chile and Zambia on the plot.

Now look at the leverage: We first extract the X-matrix here using `model.matrix()` and then compute and plot the leverages (also called "hat" values)

```
> x <- model.matrix(g)
> lev <- hat(x)
> plot(lev,ylab="Leverages",main="Index plot of Leverages")
> abline(h=2*5/50)
> sum(lev)
[1] 5
```

Notice that the sum of the leverages is equal to p which is 5 for this data. Which countries have large leverage? We have marked a horizontal line at $2p/n$ to indicate our "rule of thumb". We can see which countries exceed this rather arbitrary cut-off:

```
> names(lev) <- countries
> lev[lev > 0.2]
      Ireland         Japan United States         Libya
      0.21224       0.22331       0.33369       0.53146
```

The command `names()` assigns the country names to the elements of the vector `lev` making it easier to identify them. Alternatively, we can do it interactively like this

```
identify(1:50,lev,countries)
```

I have identified Libya and the United States as the points with the highest leverage.

7.2 Studentized Residuals

As we have seen var $\hat{\varepsilon}_i = \sigma^2(1-h_i)$ this suggests the use of

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

which are called (internally) studentized residuals. If the model assumptions are correct var $r_i = 1$ and $corr(r_i, r_j)$ tends to be small. Studentized residuals are sometimes preferred in residual plots as they have been standardized to have equal variance.

Note that studentization can only correct for the natural non-constant variance in residuals when the errors have constant variance. If there is some underlying heteroscedascity in the errors, studentization cannot correct for it.

We now get the studentized residuals for the savings data:

```
> gs <- summary(g)
> gs$sig
[1] 3.8027
> stud <- g$res/(gs$sig*sqrt(1-lev))
> plot(stud,ylab="Studentized Residuals",main="Studentized Residuals")
```

Notice the range on the axis. Which residuals are large? In this case, there is not much difference between the studentized and raw residuals apart from the scale. Only when there is unusually large leverage will the differences be noticeable.

7.3 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 may effect the fit — see Figure 7.2. The two additional points marked points both have high leverage because they are far from the rest of the data. ▲ is not an outlier. • does not have a large residual if it is included in the fit. Only when we compute the fit without that point do we get a large residual.

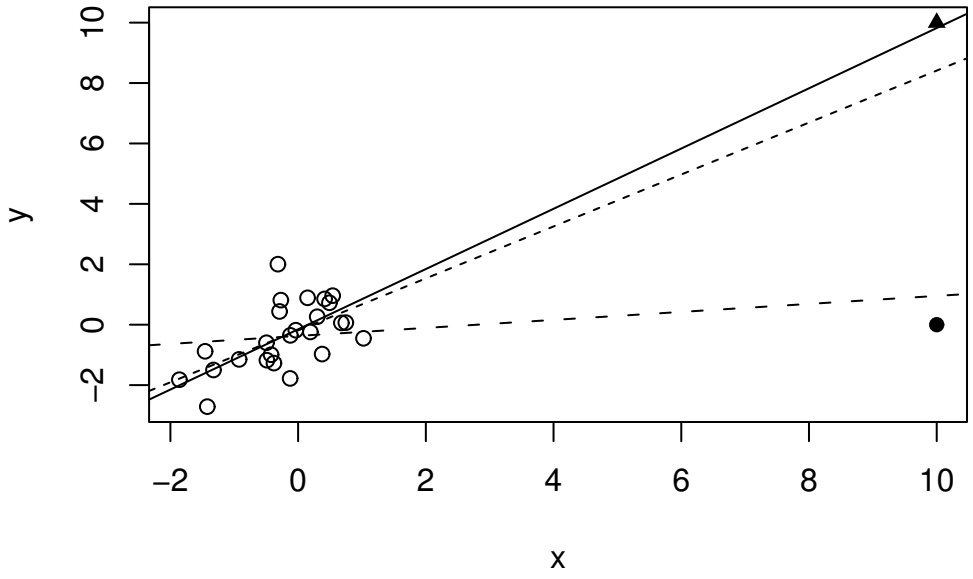

Figure 7.2: Outliers can conceal themselves. The solid line is the fit including the ▲ point but not the • point. The dotted line is the fit without either additional point and the dashed line is the fit with the • point but not the ▲ point.

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

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

If $\hat{y}_{(i)} - y_i$ is large then case i is an outlier. Just looking at $\hat{\varepsilon}_i$ misses those nasty points which pull the regression line so close to them that they conceal their true status. How large is large?

$$\text{var}\,(\hat{y}_{(i)} - y_i) = \sigma^2(1 + x_i^T (X_{(i)}^T X_{(i)}) x_i)$$

and so

$$\hat{\text{var}}\,(\hat{y}_{(i)} - y_i) = \hat{\sigma}_{(i)}^2(1 + x_i^T (X_{(i)}^T X_{(i)}) x_i)$$

Define the jackknife (or externally studentized or crossvalidated) residuals as

$$t_i = \frac{y_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)}(1 + x_i^T (X_{(i)}^T X_{(i)}) x_i)^{1/2}}$$

which are distributed t_{n-p-1} if the model is correct and $\varepsilon \sim N(0, \sigma^2 I)$. Fortunately there is an easy way to compute t_i:

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

which avoids doing n regressions.

Since $t_i \sim t_{n-p-1}$ and we can calculate a p-value to test whether case i is an outlier. However, we are likely to want to test all cases so we must adjust the level of the test accordingly. Even though it might seems that we only test one or two large t_i's, by identifying them as large we are implicitly testing all cases. Suppose we want a level α test. Now P(all tests accept) = 1 - P(At least one rejects) $\geq 1 - \sum_i P$(Test i rejects) $= 1 - n\alpha$. So this suggests that if an overall level α test is required then a level α/n should be used in each of the tests. This method is called the Bonferroni correction and is used in contexts other than outliers as well. It's biggest drawback is that it is conservative — it finds fewer outliers than the nominal level of confidence would dictate. The larger that n is, the more conservative it gets.

Now get the jacknife residuals for the savings data:

```
> jack <- rstudent(g)
> plot(jack,ylab="Jacknife Residuals",main="Jacknife Residuals")
> jack[abs(jack)==max(abs(jack))]
Zambia
2.8536
```

The largest residual of 2.85 is pretty big for a standard normal scale but is it an outlier? Compute the Bonferroni critical value:

```
> qt(.05/(50*2),44)
[1] -3.5258
```

What do you conclude?

Notes

1. Two or more outliers next to each other can hide each other.

2. An outlier in one model may not be an outlier in another when the variables have been changed or transformed. You will usually need to reinvestigate the question of outliers when you change the model.

3. The error distribution may not be normal and so larger residuals may be expected. For example, day-to-day changes in stock indices seem mostly normal but large changes occur not infrequently.

4. Individual outliers are usually much less of a problem in larger datasets. A single point won't have the leverage to affect the fit very much. It's still worth identifying outliers if these type of points are worth knowing about in the particular application. For large datasets, we need only worry about clusters of outliers. Such clusters are less likely to occur by chance and more likely to represent actual structure. Finding these cluster is not always easy.

What should be done about outliers?

1. Check for a data entry error first. These are relatively common. Unfortunately, the original source of the data may have been lost.

2. 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. Another example of the importance of outliers is in the statistical analysis of credit card transactions. Outliers in this case may represent fraudulent use.

3. 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. This can lead to difficult decision about what exclusions are reasonable. To avoid any suggestion of dishonesty, always report the existence of outliers even if you do not include them in your final model.

   It's dangerous to exclude outliers in an automatic manner. NASA launched the Nimbus 7 satellite to record atmospheric information. After several years of operation in 1985, the British Antartic Survey observed a large decrease in atmospheric ozone over the Antartic. On further examination of the NASA data, it was found that the data processing program automatically discarded observations that were extremely low and assumed to be mistakes. Thus the discovery of the Antartic ozone hole was delayed several years. Perhaps, if this had been known earlier, the CFC phaseout would have been agreed earlier and the damage could have been limited.

Here is an example of a dataset with multiple outliers. Data are available on the log of the surface temperature and the log of the light intensity of 47 stars in the star cluster CYG OB1, which is in the direction of Cygnus.

Read in and plot the data:

```
> data(star)
> plot(star$temp,star$light,xlab="log(Temperature)",
  ylab="log(Light Intensity)")
```

What do you think relationship is between temperature and light intensity? Now fit a linear regression and add the fitted line to the plot

```
> ga <- lm(light ~ temp, data=star)
> abline(ga)
```

The plot is shown in Figure 7.3 with the regression line in solid type.

Is this what you expected? Are there any outliers in the data? The outlier test does not reveal any.

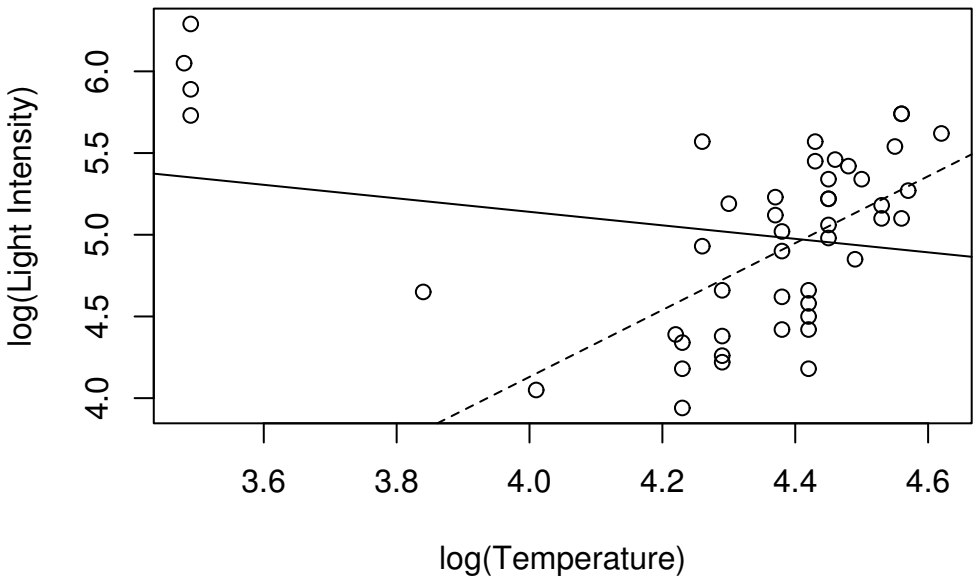

Figure 7.3: Regression line including four leftmost points is solid and excluding these points is dotted

```
> range(rstudent(ga))
[1] -2.0494  1.9058
```

We need not bother to actually compute the critical value since these values are clearly not large enough. The four stars on the upper left of the plot are giants. See what happens if these are excluded

```
> ga <- lm(light ~ temp, data=star, subset=(temp>3.6))
> abline(ga$coef,lty=2)
```

This illustrates the problem of multiple outliers. We can visualize the problems here, but for higher dimensional data this is much more difficult.

7.4 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. In Figure 7.2, the ▲ point is not an influential point but the • point is.

Here are some measures of influence, where the subscripted (i) indicates the fit without case i.

1. Change in the coefficients $\hat{\beta} - \hat{\beta}_{(i)}$
2. Change in the fit $X^T(\hat{\beta} - \hat{\beta}_{(i)}) = \hat{y} - \hat{y}_{(i)}$

These are hard to judge in the sense that the scale varies between datasets. A popular alternative are the Cook Statistics:

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})^T (X^T X)(\hat{\beta} - \hat{\beta}_{(i)})}{p\hat{\sigma}^2}$$

$$= \frac{(\hat{y}-\hat{y}_{(i)})^T(\hat{y}-\hat{y}_{(i)})}{p\hat{\sigma}^2}$$
$$= \frac{1}{p}r_i^2\frac{h_i}{1-h_i}$$

The first term, r_i^2, is the residual effect and the second is the leverage. The combination of the two leads to influence. An index plot of D_i can be used to identify influential points.

Continuing with our study of the savings data:

```
> cook <- cooks.distance(g)
> plot(cook,ylab="Cooks distances")
> identify(1:50,cook,countries)
```

The Cook statistics may be seen in Figure 7.4. I have identified the largest three values.

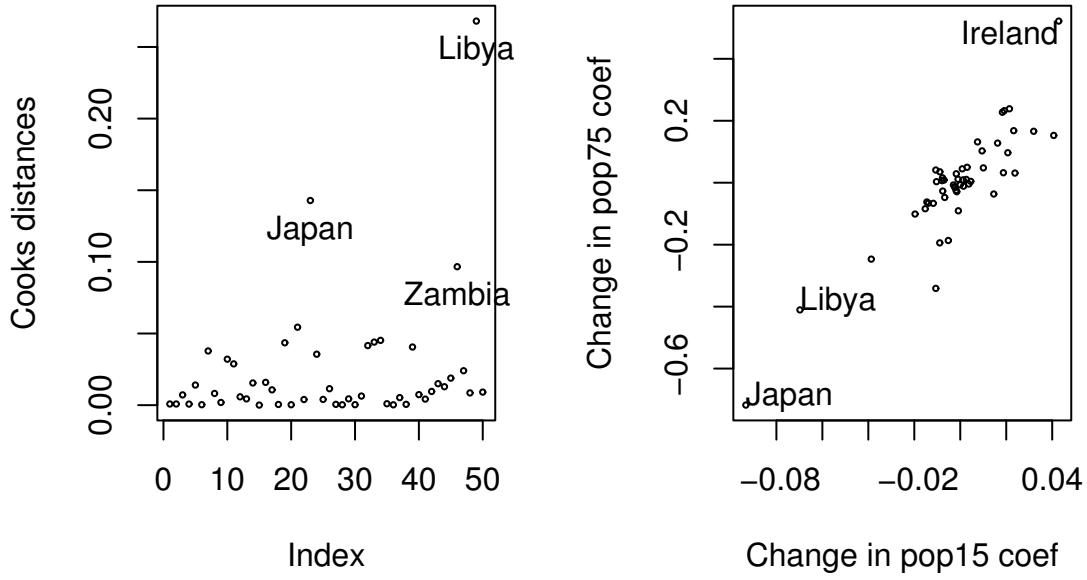

Figure 7.4: Cook Statistics and $\hat{\beta}-\hat{\beta}_{(i)}$'s for the savings data

Which ones are large? We now exclude the largest one and see how the fit changes:

```
> gl <- lm(sr ~ pop15+pop75+dpi+ddpi,savings,subset=(cook < max(cook)))
> summary(gl)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.524046   8.224026    2.98   0.0047
pop15       -0.391440   0.157909   -2.48   0.0171
pop75       -1.280867   1.145182   -1.12   0.2694
dpi         -0.000319   0.000929   -0.34   0.7331
ddpi         0.610279   0.268778    2.27   0.0281

Residual standard error: 3.79 on 44 degrees of freedom
Multiple R-Squared: 0.355,      Adjusted R-squared: 0.297
F-statistic: 6.07 on 4 and 44 degrees of freedom,       p-value: 0.000562
```

Compared to the full data fit:

```
> summary(g)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.566087   7.354516    3.88  0.00033
pop15       -0.461193   0.144642   -3.19  0.00260
pop75       -1.691498   1.083599   -1.56  0.12553
dpi         -0.000337   0.000931   -0.36  0.71917
ddpi         0.409695   0.196197    2.09  0.04247

Residual standard error: 3.8 on 45 degrees of freedom
Multiple R-Squared: 0.338,      Adjusted R-squared: 0.28
F-statistic: 5.76 on 4 and 45 degrees of freedom,       p-value: 0.00079
```

What changed? The coefficient for `ddpi` changed by about 50%. We don't like our estimates to be so sensitive to the presence of just one country. It would be rather tedious to do this for each country but there's a quicker way:

```
> ginf <- lm.influence(g)
> plot(ginf$coef[,2],ginf$coef[,3],xlab="Change in pop15 coef",
              ylab="Change in pop75 coef")
> identify(ginf$coef[,2],ginf$coef[,3],countries)
```

We just plotted the change in the second and third parameter estimates when a case is left out as seen in the second panel of Figure 7.4. Try this for the other estimates - which countries stick out? Consider Japan:

```
> gj <- lm(sr ~ pop15+pop75+dpi+ddpi,savings,
           subset=(countries != "Japan"))
> summary(gj)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.940171   7.783997    3.08   0.0036
pop15       -0.367901   0.153630   -2.39   0.0210
pop75       -0.973674   1.155450   -0.84   0.4040
dpi         -0.000471   0.000919   -0.51   0.6112
ddpi         0.334749   0.198446    1.69   0.0987

Residual standard error: 3.74 on 44 degrees of freedom
Multiple R-Squared: 0.277,      Adjusted R-squared: 0.211
F-statistic: 4.21 on 4 and 44 degrees of freedom,       p-value: 0.00565
```

Compare to the full data fit - what qualitative changes do you observe? Notice that the `ddpi` term is no longer significant and that the the R^2 value has decreased a lot.

7.5 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. Plot $\hat{\varepsilon}$ against $\hat{y}$. This is the most important diagnostic plot that

you can make. 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 necessary). In Figure 7.5, these three cases are illustrated.

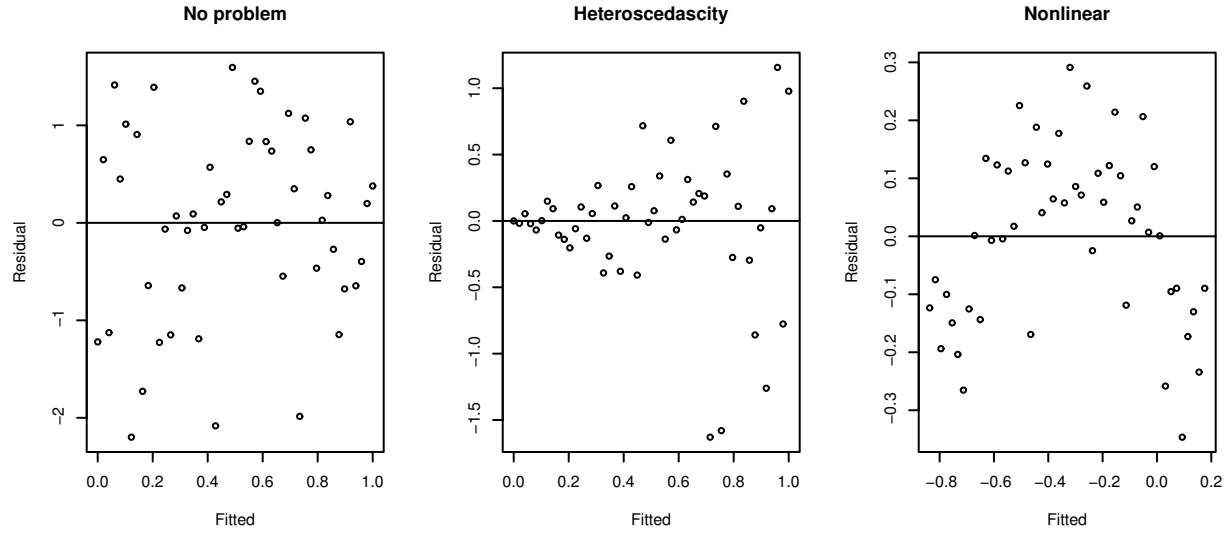

Figure 7.5: Residuals vs Fitted plots - the first suggests no change to the current model while the second shows non-constant variance and the third indicates some nonlinearity which should prompt some change in the structural form of the model

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.

We illustrate this using the savings dataset as an example again:

```
> g <- lm(sr ~ pop15+pop75+dpi+ddpi,savings)
```

First the residuals vs. fitted plot and the abs(residuals) vs. fitted plot.

```
> plot(g$fit,g$res,xlab="Fitted",ylab="Residuals")
> abline(h=0)
> plot(g$fit,abs(g$res),xlab="Fitted",ylab="|Residuals|")
```

The plots may be seen in the first two panels of Figure 7.5. What do you see? The latter plot is designed to check for non-constant variance only. It folds over the bottom half of the first plot to increase the resolution for detecting non-constant variance. The first plot is still needed because non-linearity must be checked.

A quick way to check non-constant variance is this regression:

```
> summary(lm(abs(g$res) ~ g$fit))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.840      1.186    4.08  0.00017
g$fit         -0.203      0.119   -1.72  0.09250
```

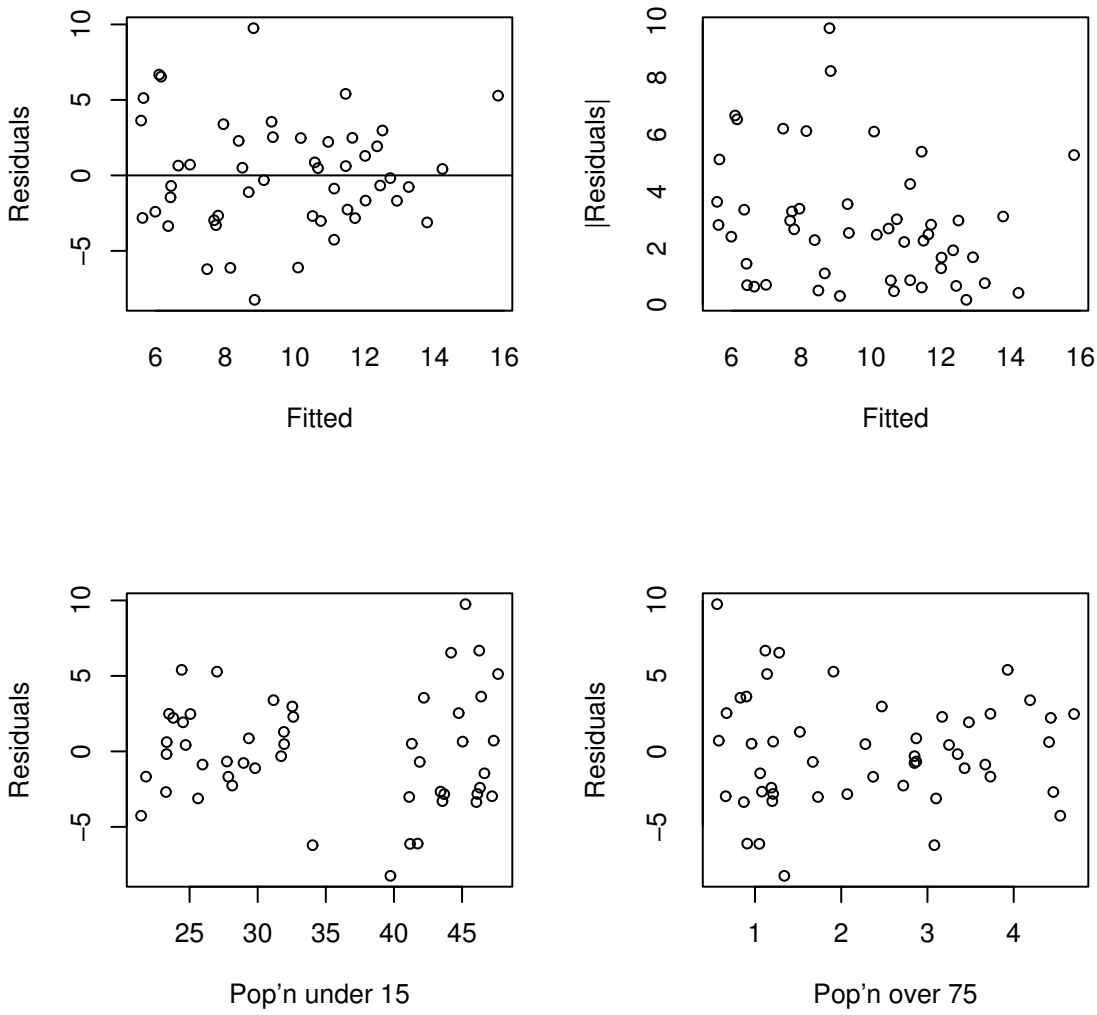

Figure 7.6: Residual plots for the savings data

```
Residual standard error: 2.16 on 48 degrees of freedom
Multiple R-Squared: 0.0578,     Adjusted R-squared: 0.0382
F-statistic: 2.95 on 1 and 48 degrees of freedom,       p-value: 0.0925
```

This test is not quite right as some weighting should be used and the degrees of freedom should be adjusted but there doesn't seem to be a clear problem with non-constant variance.

It's often had to judge residual plots without prior experience so let's show how to generate some of the artificial variety. The following four `for()` loops show

1. Constant Variance
2. Strong non-constant variance
3. Mild non-constant variance
4. Non-linearity

```
> par(mfrow=c(3,3))
> for(i in 1:9) plot(1:50,rnorm(50))
> for(i in 1:9) plot(1:50,(1:50)*rnorm(50))
> for(i in 1:9) plot(1:50,sqrt((1:50))*rnorm(50))
> for(i in 1:9) plot(1:50,cos((1:50)*pi/25)+rnorm(50))
> par(mfrow=c(1,1))
```

In this case we know the truth - do you think you would be able to come to the right conclusions for real data? Repeat to get an idea of the usual amount of variation. I recommend the artificial generation of plots as a way to "calibrate" diagnostic plots. It's often hard to judge whether an apparent feature is real or just random variation. Repeated generation of plots under a model where there is no violation of the assumption that the diagnostic plot is designed to check is helpful in making this judgement.

Now look at some residuals against predictor plots:

```
> plot(savings$pop15,g$res,xlab="Population under 15",ylab="Residuals")
> plot(savings$pop75,g$res,xlab="Population over 75",ylab="Residuals")
```

The plots may be seen in the second two panels of Figure 7.5. Can you see the two groups? Let's compare and test the variances. Given two independent samples from normal distributions, we can test for equal variance using the test statistic of the ratio of the two variance. The null distribution is a F with degrees of freedom given by the two samples.

```
> var(g$res[savings$pop15 > 35])/var(g$res[savings$pop15 <35])
[1] 2.7851
> table(savings$pop15 > 35)
FALSE  TRUE
   27    23
> 1-pf(2.7851,22,26)
[1] 0.0067875
```

A significant difference is seen

7.6 Non-Constant Variance

There are two approaches to dealing with non-constant variance. Weighted least squares is appropriate when the form of the non-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. To see how to choose $h()$ consider this

$$\begin{aligned} h(y) &= h(Ey) + (y - Ey)h'(Ey) + \ldots \\ \text{var } h(y) &= h'(Ey)^2 \text{var } y + \ldots \end{aligned}$$

We ignore the higher order terms. For var $h(y)$ to be constant we need

$$h'(Ey) \propto (\text{var } y)^{-1/2}$$

which suggests

$$h(y) = \int \frac{dy}{\sqrt{\text{var } y}} == \int \frac{dy}{\texttt{SD}y}$$

For example if var $y = \text{var}\,\varepsilon \propto (Ey)^2$ then $h(y) = \log y$ is suggested while if var $\varepsilon \propto (Ey)$ then $h(y) = \sqrt{y}$.

Graphically one tends to see *SDy* rather then var y. Sometimes $y_i \leq 0$ for some i in which case the transformations may choke. You can try $\log(y+\delta)$ for some small δ but this makes interpretation difficult.

Consider the residual-fitted plot for the Galapagos data:

```
> gg <- lm(Species ~ Area + Elevation + Scruz + Nearest + Adjacent, gala)
> plot(gg$fit,gg$res,xlab="Fitted",ylab="Residuals",
                 main="Untransformed Response")
```

We can see non-constant variance in the first plot of Figure 7.7.

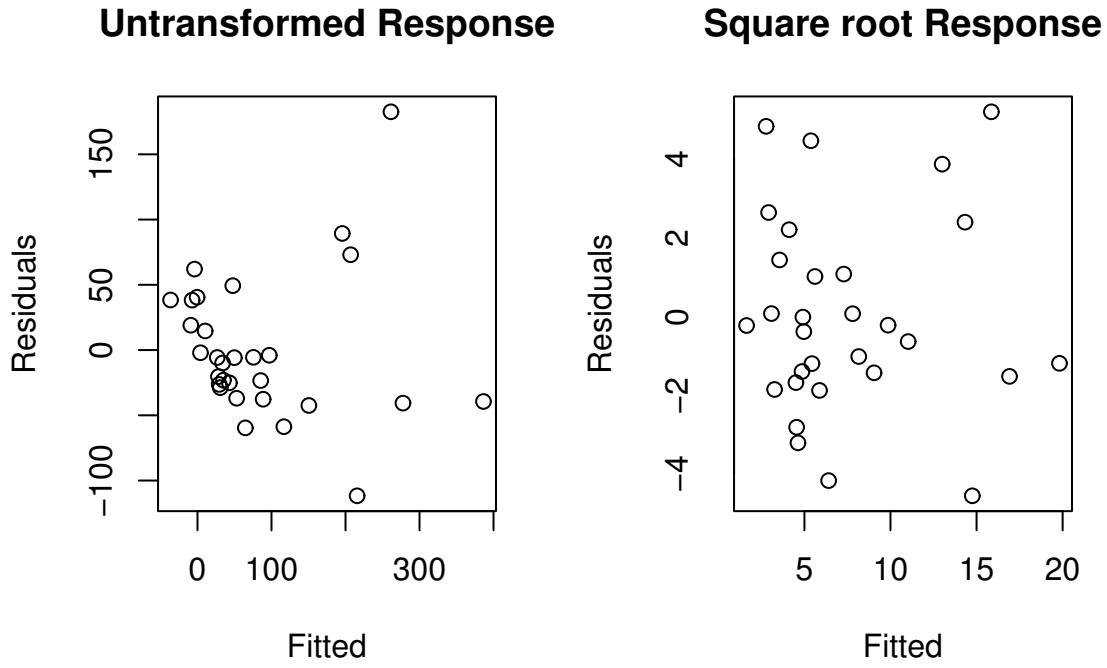

Figure 7.7: Residual-Fitted plots for the Galapagos data before and after transformation

We guess that a square root transformation will give us constant variance:

```
> gs <- lm(sqrt(Species) ~ Area+ Elevation+ Scruz+ Nearest+ Adjacent, gala)
> plot(gs$fit,gs$res,xlab="Fitted",ylab="Residuals",
                 main="Square root Response")
```

We see in the second plot of Figure 7.7 that the variance is now constant. Our guess at a variance stabilizing transformation worked out here, but had it not, we could always have tried something else. The square root transformation is often appropriate for count response data. The poisson distribution is a good model for counts and that distribution has the property that the mean is equal to the variance thus suggesting the square root transformation. It might be even better to go with a poisson regression rather than the normal-based regression we are using here.

There are more formal tests for non-constant variance — for example one could start by regressing $|\hat{\varepsilon}|$ on y or x_i but there is a problem in specifying the alternative hypothesis for such a test. A formal test may be good at detecting a particular kind of non-constant variance but have no power to detect another. Residual plots are more versatile because unanticipated problems may be spotted.

A formal diagnostic test may have reassuring aura of exactitude about it, but one needs to understand that any such test may be powerless to detect problems of an unsuspected nature. Graphical techniques are usually more effective at revealing structure that you may not have suspected. Of course, sometimes the interpretation of the plot may be ambiguous but at least one can be sure that nothing is seriously wrong with the assumptions. For this reason, I usually prefer a graphical approach to diagnostics.

7.7 Non-Linearity

Lack of fit tests can be used when there is replication which doesn't happen too often, but even if you do have it, the tests don't tell you how to improve the model. How do we check if the systematic part ($Ey = X\beta$) of the model is correct? We can look at

1. Plots of $\hat{\varepsilon}$ against $\hat{y}$ and x_i
2. Plots of y against each x_i.

but what about the effect 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.

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 $\hat{\beta}_i$ which adds some insight into the meaning of regression coefficients. Look for non-linearity and outliers and/or influential points.

Partial Residual plots are a competitor to added variable plots. These plot $\hat{\varepsilon} + \hat{\beta}_i x_i$ against x_i. To see where this comes from, look at the response with the predicted effect of the other X removed:

$$y - \sum_{j \neq i} x_j \hat{\beta}_j = \hat{y} + \hat{\varepsilon} - \sum_{j \neq i} x_j \hat{\beta}_j = x_i \hat{\beta}_i + \hat{\varepsilon}$$

Again the slope on the plot will be $\hat{\beta}_i$ and the interpretation is the same. Partial residual plots are reckoned to be better for non-linearity detection while added variable plots are better for outlier/influential detection.

We illustrate using the savings dataset as an example again: First we construct a partial regression (added variable) plot for `pop15`:

```
> d <- lm(sr ~ pop75 + dpi + ddpi,savings)$res
> m <- lm(pop15 ~ pop75 + dpi + ddpi,savings)$res
> plot(m,d,xlab="pop15 residuals",ylab="Saving residuals",
                    main="Partial Regression")
```

Compare the slope on the plot to the original regression and show the line on the plot (see Figure 7.7).

```
> lm(d ~ m)$coef
(Intercept)           m
 5.4259e-17 -4.6119e-01
> g$coef
(Intercept)       pop15       pop75         dpi        ddpi
 28.5660865  -0.4611931  -1.6914977  -0.0003369   0.4096949
> abline(0,g$coef['pop15'])
```

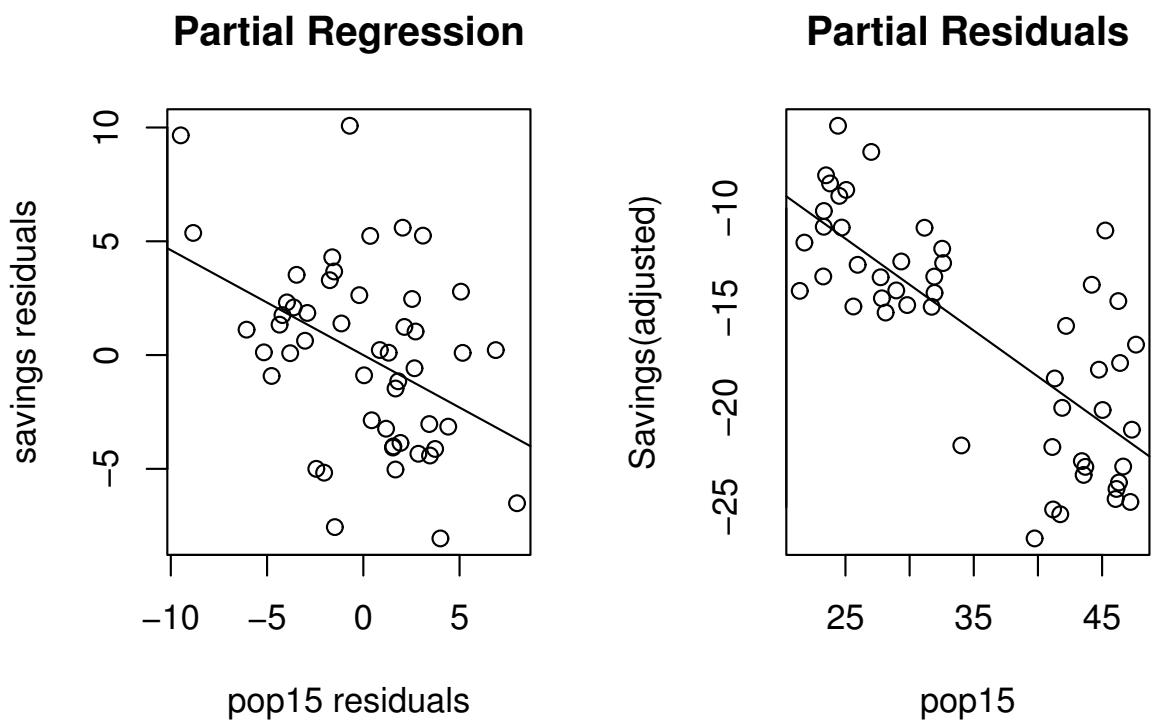

Figure 7.8: Partial regression and residual plots for the savings data

Notice how the slope in the plot and the slope for `pop15` in the regression fit are the same.

The partial regression plot also provides some intuition about the meaning of regression coefficients. We are looking at the marginal relationship between the response and the predictor after the effect of the other predictors has been removed. Multiple regression is difficult because we cannot visualize the full relationship because of the high dimensionality. The partial regression plot allows us to focus on the relationship between one predictor and the response, much as in simple regression.

A partial residual plot is easier to do:

```
> plot(savings$pop15,g$res+g$coef['pop15']*savings$pop15,xlab="pop'n under 15",
  ylab="Saving(adjusted)",main="Partial Residual")
> abline(0,g$coef['pop15'])
```

Or more directly:

```
> prplot(g,1)
```

Might there be a different relationship in the two groups?

```
> g1 <- lm(sr ~ pop15+pop75+dpi+ddpi,savings,subset=(pop15 > 35))
> g2 <- lm(sr ~ pop15+pop75+dpi+ddpi,savings,subset=(pop15 < 35))
> summary(g1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.433969  21.155028   -0.12     0.91
pop15        0.273854   0.439191    0.62     0.54
```

```
pop75       -3.548477   3.033281   -1.17     0.26
dpi          0.000421   0.005000    0.08     0.93
ddpi         0.395474   0.290101    1.36     0.19

Residual standard error: 4.45 on 18 degrees of freedom
Multiple R-Squared: 0.156,      Adjusted R-squared: -0.0319
F-statistic: 0.83 on 4 and 18 degrees of freedom,       p-value: 0.523

> summary(g2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.961795   8.083750    2.96   0.0072
pop15       -0.385898   0.195369   -1.98   0.0609
pop75       -1.327742   0.926063   -1.43   0.1657
dpi         -0.000459   0.000724   -0.63   0.5326
ddpi         0.884394   0.295341    2.99   0.0067

Residual standard error: 2.77 on 22 degrees of freedom
Multiple R-Squared: 0.507,      Adjusted R-squared: 0.418
F-statistic: 5.66 on 4 and 22 degrees of freedom,       p-value: 0.00273
```

Can you see the difference? The graphical analysis has shown a relationship in the data that a purely numerical analysis might easily have missed.

Higher dimensional plots can also be useful for detecting structure that cannot be seen in two dimensions. These are interactive in nature so you need to try them to see how they work. Two ideas are

1. Spinning - 3D plots where color, point size and rotation are used to give illusion of a third dimension.
2. Brushing - Two or more plots are linked so that point which are *brushed* in one plot are highlighted in another.

These tools look good but it's not clear whether they actually are useful in practice. Certainly there are communication difficulties as these plots cannot be easily printed. Many statistical packages allow for this kind of investigation. XGobi is a useful free UNIX-based tool for exploring higher dimensional data that has now been made extended to Windows also as Ggobi. See `www.ggobi.org`

```
> library(xgobi)
> xgobi(savings)
```

or

```
> library(Rggobi)
> ggobi(savings)
```

Most of the functionality can be discovered by experimentation and the online help.

7.8 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: $\hat{\varepsilon}_{[1]} \leq \ldots \hat{\varepsilon}_{[n]}$
2. Compute $u_i = \Phi^{-1}(\frac{i}{n+1})$
3. Plot $\hat{\varepsilon}_{[i]}$ against u_i. If the residuals are normally distributed an approximately straight-line relationship will be observed.

Let's try it out on the same old data:

```
> qqnorm(g$res,ylab="Raw Residuals")
> qqline(g$res)
```

See the first plot of Figure 7.8 - `qqline()` adds a line joining the first and third quartiles - it's useful as a guide. We can plot the (externally) studentized residuals:

```
> qqnorm(rstudent(g),ylab="Studentized residuals")
> abline(0,1)
```

See the second plot of the figure. Because these residuals have been normalized, they should lie along a 45 degree line.

Histograms and boxplots are not as sensitive for checking normality:

```
> hist(g$res,10)
> boxplot(g$res,main="Boxplot of savings residuals")
```

We can get an idea of the variation to be expected in QQ-plots in the following experiment. I generate data from different distributions:

1. Normal
2. Lognormal - an example of a skewed distribution
3. Cauchy - an example of a long-tailed (platykurtic) distribution
4. Uniform - an example of a short-tailed (leptokurtic) distribution

Here's how to generate 9 replicates at a time from each of these test cases:

```
> oldpar <- par()
> par(mfrow=c(3,3))
> for(i in 1:9) qqnorm(rnorm(50))
> for(i in 1:9) qqnorm(exp(rnorm(50)))
> for(i in 1:9) qqnorm(rcauchy(50))
> for(i in 1:9) qqnorm(runif(50))
> par(oldpar)
```

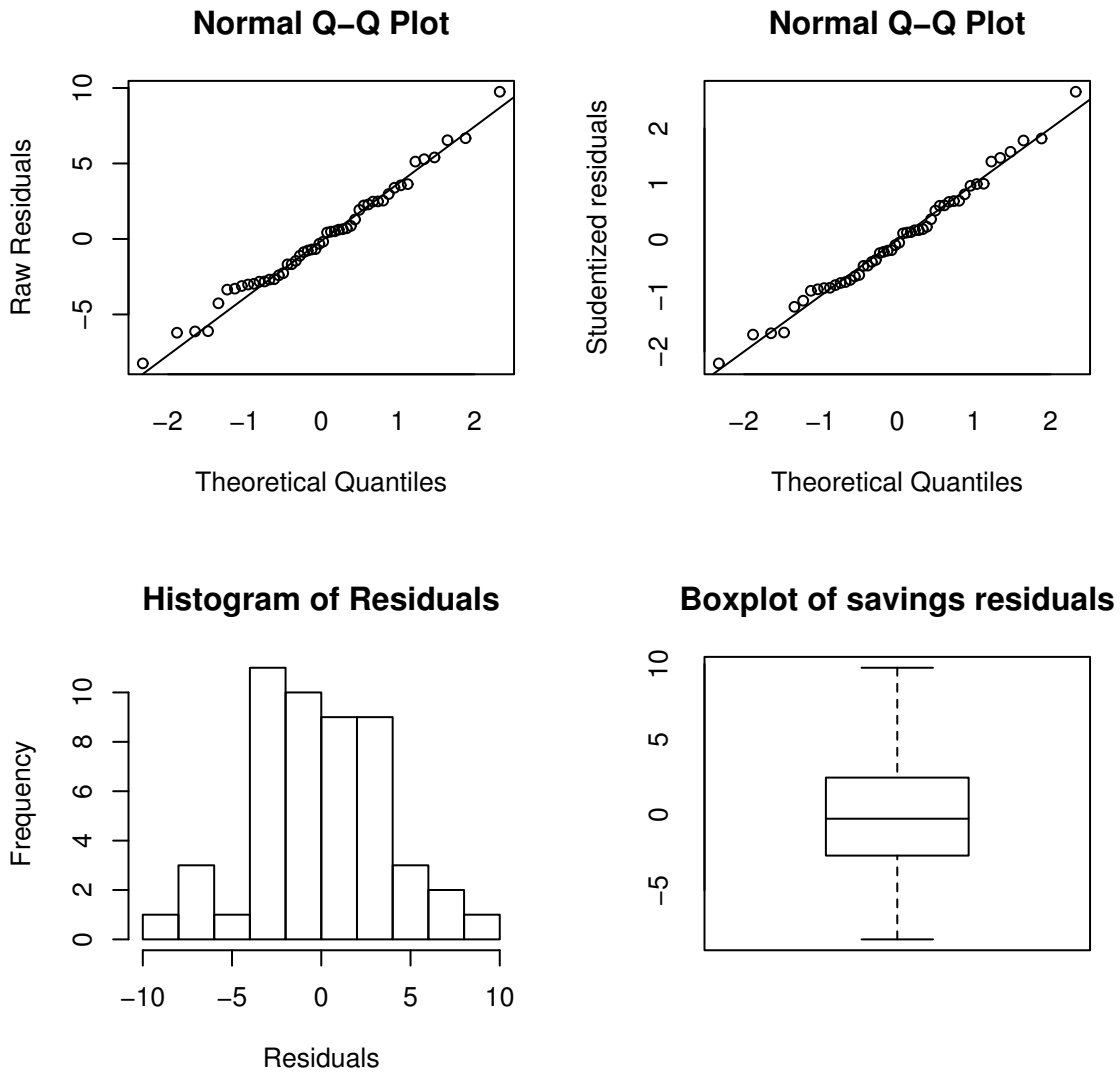

Figure 7.9: Normality checks for the savings data

We save the original settings for the graphics layout in `oldpar` and restore it after we are done. This is a useful trick when you want to experiment with changing these settings.

In Figure 7.8, you can see examples of all four cases:

It's not always easy to diagnose the problem in QQ plots.

The consequences of non-normality are

1. that the least squares estimates may not be optimal - they will still be BLUE but other *robust* estimators may be more effective.

2. that the tests and confidence intervals are invalid. However, it has been shown that only really long-tailed distributions cause a problem. Mild non-normality can safely be ignored and the larger the sample size the less troublesome the non-normality.

What to do?

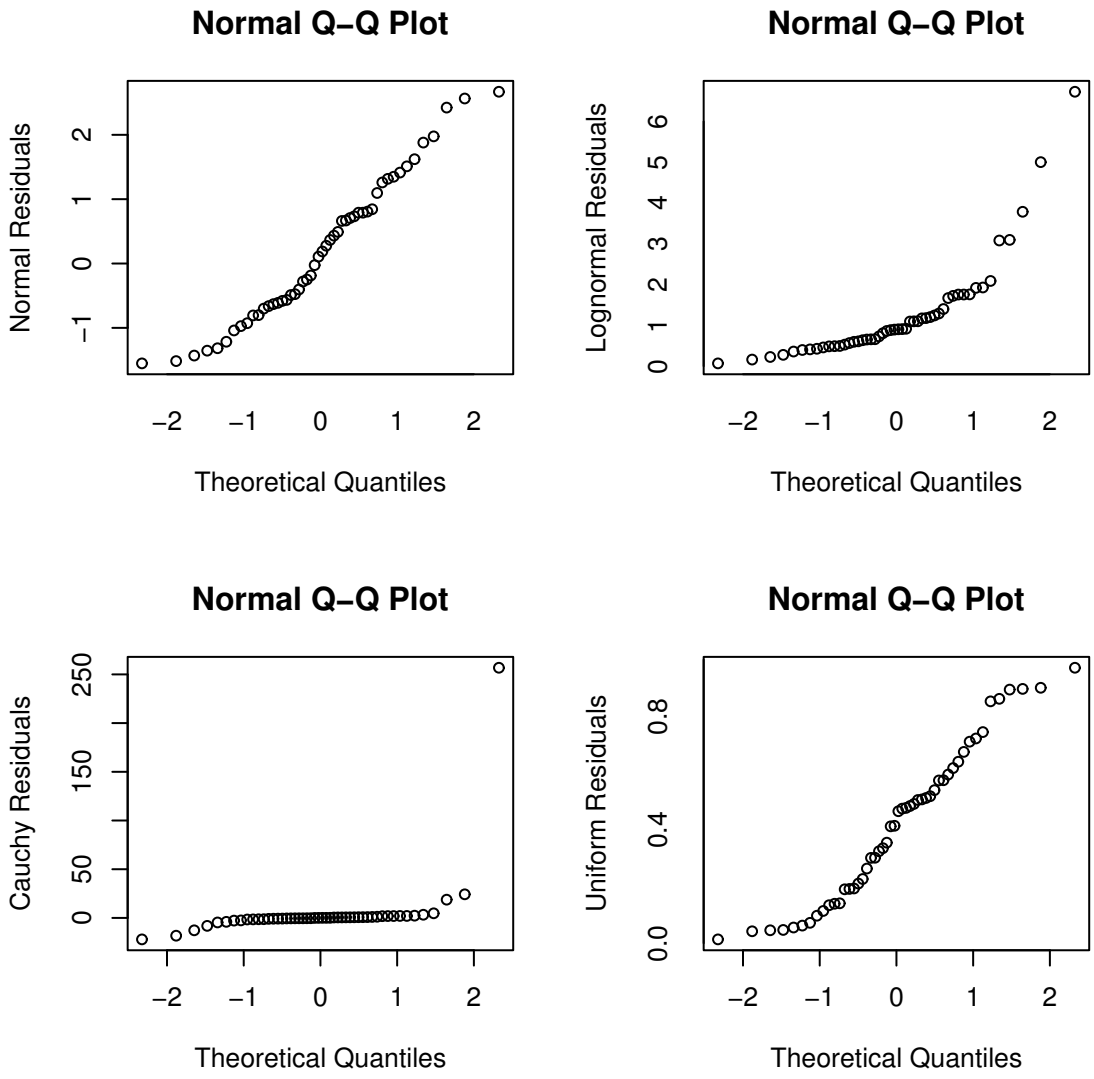

Figure 7.10: QQ plots of simulated data

1. A transformation of the response may solve the problem - this is often true for skewed errors.
2. Other changes in the model may help.
3. 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.
4. For short-tailed distributions, the consequences of non-normality are not serious and can reasonably be ignored.

There are formal tests for normality such as the Kolmogorov-Smirnov test but these are not as flexible as the Q-Q plot. The p-value is not very helpful as an indicator of what action to take. After all, with a large dataset, even mild deviations from non-normality may be detected, but there would be little reason to

abandon least squares because the effects of non-normality are mitigated by large sample sizes. For smaller sample sizes, formal tests lack power.

7.9 Half-normal plots

Half-normal plots are designed for the assessment of positive data. They could be used for $|\hat{\varepsilon}|$ but are more typically useful for diagnostic quantities like the leverages or the Cook Statistics. The idea is to plot the data against the positive normal quantiles

The steps are:

1. Sort the data: $x_{[1]} \leq \ldots x_{[n]}$
2. Compute $u_i = \Phi^{-1}(\frac{n+i}{2n+1})$
3. Plot $x_{[i]}$ against u_i.

We are usually not looking for a straight line relationship since we do not necessarily expect a positive normal distribution for quantities like the leverages. (If the X is multivariate normal, the leverages will have a χ^2_p distribution but there is usually no good reason to assume multivariate normality for the X.) We are looking for outliers which will be apparent as points that diverge substantially from the rest of the data.

We demonstrate the half-normal plot on the leverages and Cook statistics for the savings data:

```
> halfnorm(lm.influence(g)$hat,labs=countries,ylab="Leverages")
> halfnorm(cooks.distance(g),labs=countries,ylab="Cook Statistics")
```

The plots are chosen in Figure 7.11 — I have plotted the country name instead of just a dot for the largest two cases respectively to aid identification. The `halfnorm()` function comes from the book library.

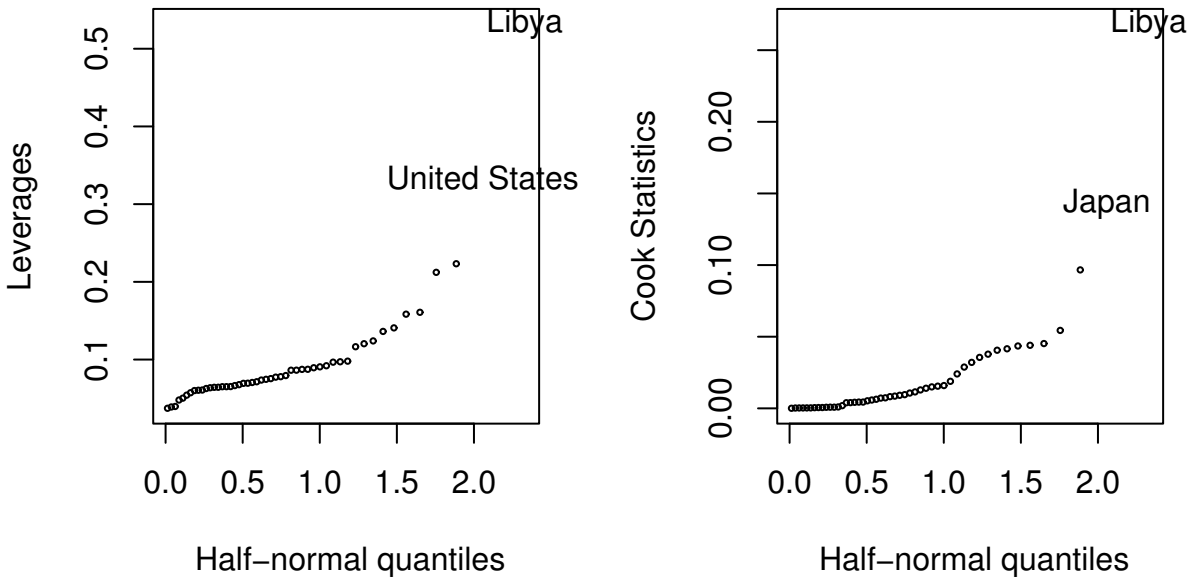

Figure 7.11: Half-normal plots for the leverages and Cook statistics

Libya shows up clearly as unusual in both plots

7.10 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.

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

If you do have correlated errors, you can use GLS. 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.

For the example, we use some taken from an environmental study that measured the four variables ozone, solar radiation, temperature, and wind speed for 153 consecutive days in New York.

```
> data(airquality)
> airquality
    Ozone Solar.R Wind Temp Month Day
1      41     190  7.4   67     5   1
2      36     118  8.0   72     5   2
3      12     149 12.6   74     5   3
4      18     313 11.5   62     5   4
5      NA      NA 14.3   56     5   5
etc..
```

We notice that there are some missing values. Take a look at the data: (plot not shown)

```
> pairs(airquality,panel=panel.smooth)
```

We fit a standard linear model and check the residual-fitted plot in Figure 7.10.

```
> g <- lm(Ozone ~ Solar.R + Wind + Temp,airquality)
> summary(g)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -64.3421    23.0547   -2.79   0.0062
Solar.R       0.0598     0.0232    2.58   0.0112
Wind         -3.3336     0.6544   -5.09  1.5e-06
Temp          1.6521     0.2535    6.52  2.4e-09

Residual standard error: 21.2 on 107 degrees of freedom
Multiple R-Squared: 0.606,      Adjusted R-squared: 0.595
F-statistic: 54.8 on 3 and 107 degrees of freedom,      p-value:    0
> plot(g$fit,g$res,xlab="Fitted",ylab="Residuals",
                  main="Untransformed Response")
```

Notice how there are only 107 degrees corresponding to the 111 complete observations. The default behavior in R when performing a regression with missing values is to exclude any case that contains a missing value. We see some non-constant variance and nonlinearity and so we try transforming the response:

```
> gl <- lm(log(Ozone) ~ Solar.R + Wind + Temp,airquality)
> plot(gl$fit,gl$res,xlab="Fitted",ylab="Residuals",main="Logged Response")
```

Suppose we are now otherwise satisfied with this model and want to check for serial correlation. The missing values in the data were not used in the construction of the model but this also breaks up the sequential pattern in the data. I get round this by reintroducing missing values into the residuals corresponding to the omitted cases.

```
> res <- rep(NA,153)
> res[as.numeric(row.names(na.omit(airquality)))] <- gl$res
```

First make an index plot of the residuals — see Figure 7.10.

```
> plot(res,ylab="Residuals",main="Index plot of residuals")
```

Is there any evidence of serial correlation? Now plot successive residuals:

```
> plot(res[-153],res[-1],xlab=expression(hat(epsilon)[i]),
              ylab=expression(hat(epsilon)[i+1]))
```

Do you see any problem? Let's check

```
> summary(lm(res[-1] ~ -1+res[-153]))
Coefficients:
          Estimate Std. Error t value Pr(>|t|)
res[-153]    0.110      0.105    1.05      0.3

Residual standard error: 0.508 on 91 degrees of freedom
Multiple R-Squared: 0.0119,     Adjusted R-squared: 0.00107
F-statistic:  1.1 on 1 and 91 degrees of freedom,       p-value: 0.297
```

We omitted the intercept term because the residuals have mean zero. We see that there is no significant correlation.

You can plot more than just successive pairs if you suspect a more complex dependence. For spatial data, more complex checks are required.

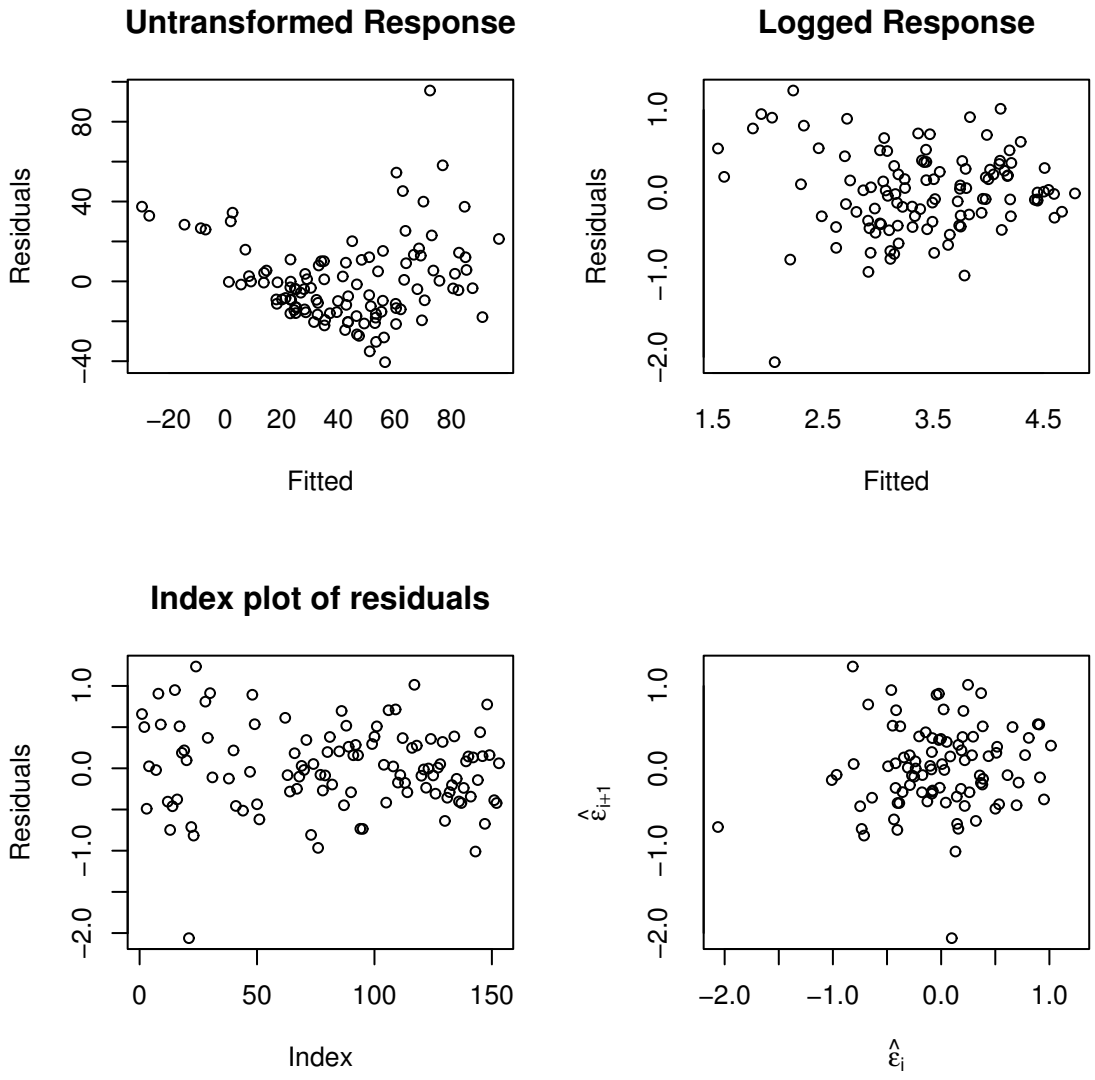

Figure 7.12: Checking for correlated errors - Index plot and scatterplot of successive residuals