Understanding Data and Statistical Design (60117)

Chapter 6

Simple Linear Regression II

Subject Coordinator: Stephen Woodcock
Lecture notes: Scott Alexander

School of Mathematical and Physical Sciences, UTS

Autumn 2024

Chapter outline

Topics:

- R example
  - T-tests
- taking stock
- model fit
  - ANOVA and F-test
  - ANOVA and R^2
  - leverage and influence
- checking model assumptions
  - residual analysis
  - vidual residual analysis
  - statistical residual analysis
- R example continued

R example

Consider again the example from last chapter.

Recall we looked to model the relationship between life expectancy and logarithm of per capita GNI, using the data set life.data.csv.

We began by proposing a population described by

$$LIFE = \beta_0 + \beta_1 \times gniLog + \epsilon$$

and looked to fit the model

$$\widehat{life} = \hat{\beta}_0 + \hat{\beta}_1 \times gniLog$$

to estimate

$$\mathbb{E}[LIFE] = \beta_0 + \beta_1 \times gniLog.$$

R example

R returned the following output.

```
Call:
lm(formula = life ~ gniLog, data = life.data)

Residuals:
    Min      1Q  Median      3Q     Max
-22.647  -2.267   1.020   3.354   8.938

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.2798     2.9577   5.842 2.28e-08 ***
gniLog        5.8482     0.3217  18.177  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.306 on 185 degrees of freedom
Multiple R-squared:  0.6411,    Adjusted R-squared:  0.6391
F-statistic: 330.4 on 1 and 185 DF,  p-value: < 2.2e-16
```

This gave the fitted regression equation

$$\widehat{life} = 17.2798 + 5.8482 \times gniLog.$$

We also computed the 95% CIs of the parameters.

```
                  2.5 %    97.5 %
(Intercept) 11.444555 23.114985
gniLog       5.213424  6.482883
```

R example – T-tests

The **test statistics**

$$t^*_{\hat{\beta}_0} = \frac{\hat{\beta}_0 - \beta^*_0}{s_{\hat{\beta}_0}} = \frac{\hat{\beta}_0 - 0}{s_{\hat{\beta}_0}} = \frac{17.2798}{2.9577} = 5.842$$

and

$$t^*_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - \beta^*_1}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}} = \frac{5.8482}{0.3217} = 18.177$$

were used in **two tail T-tests with hypotheses**

$$H_0: \ \beta_0 = 0$$
$$H_A: \ \beta_0 \neq 0$$

and

$$H_0: \ \beta_1 = 0$$
$$H_A: \ \beta_1 \neq 0$$

respectively.

R example – T-tests

The **rejection quantiles** (significance level $\alpha = 0.05$, $n - 2 = 185$ degrees of freedom) for these two tail tests

$$t_{0.975} = 1.97287$$

allowed each null hypothesis H_0 to be rejected as

$$t^*_{\hat{\beta}_0}, t^*_{\hat{\beta}_1} > t_{0.975}.$$

We could also see this as the **p-values** were below significance level $\alpha = 0.05$.

We could also see this as zero was outside the 95% CIs

$$11.444555 \leq \beta_0 \leq 23.114985,$$

$$5.213424 \leq \beta_1 \leq 6.482883.$$

R example – T-tests

We can use the information supplied by the R summary to test other hypotheses on the true values of the parameters.

For instance, consider the **upper tail T-test**

$$H_0: \ \beta_0 = 15$$
$$H_A: \ \beta_0 > 15.$$

The test statistic is calculated as

$$t^*_{\hat{\beta}_0} = \frac{\hat{\beta}_0 - \beta^*_0}{s_{\hat{\beta}_0}} = \frac{17.2798 - 15}{2.9577} = 0.771$$

which has an associated p-value of $p = 0.221$.

As $p > \alpha = 0.05$ (our usual significance level) we cannot reject H_0.

Other versions of these T-tests can be constructed for both β_0 and β_1.

Taking stock

So far we have looked at fitting a simple linear regression model and analysing it in terms of confidence intervals on the parameters β_0 and β_1 and predictions intervals of $\mathbb{E}[Y]$ and Y.

We saw that the parameter CIs, fitted values and the prediction intervals could be computed easily using R.

In this chapter we look deeper into the problem of linear regression and begin to answer some of the questions posed at the end of the previous chapter.

Model fit – ANOVA and F-test

We now look at another method of testing hypotheses on β_1.

Consider the **sum square total**

$$SST = \sum_{i=1}^{n} (Y_i - \overline{Y})^2.$$

with $\overline{Y}$ the sample mean of the Y_i RVs.

With a little algebra this can be decomposed as

$$SST = SSR + SSE,$$

where the **sum square regression** is given by

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$$

and the minimised **sum square error** by

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} \hat{\epsilon}_i^2.$$

Model fit – ANOVA and F-test

Under the assumptions that the noise terms $\epsilon_i \sim N(0, \sigma)$ and independent and that

$$\beta_1 = 0,$$

the RV

$$F^* = \frac{SSR}{\frac{SSE}{n-2}} = \frac{MSR}{MSE} \sim F(1, n-2), \tag{1}$$

i.e. ***F*-distributed** with 1 numerator degree of freedom and $n-2$ denominator degrees of freedom.

This provides us with an alternative method for testing the null hypothesis that $\beta_1 = 0$.

Model fit – ANOVA and F-test

To be more precise, this provides us with an equivalent method for testing the hypothesis $\beta_1 = 0$, because the square of a $T(n-2)$ RV has the same distribution as an $F(1, n-2)$ RV.

Recall the test statistic for the T-test null hypothesis $\beta_1 = 0$ has the form

$$T_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}.$$

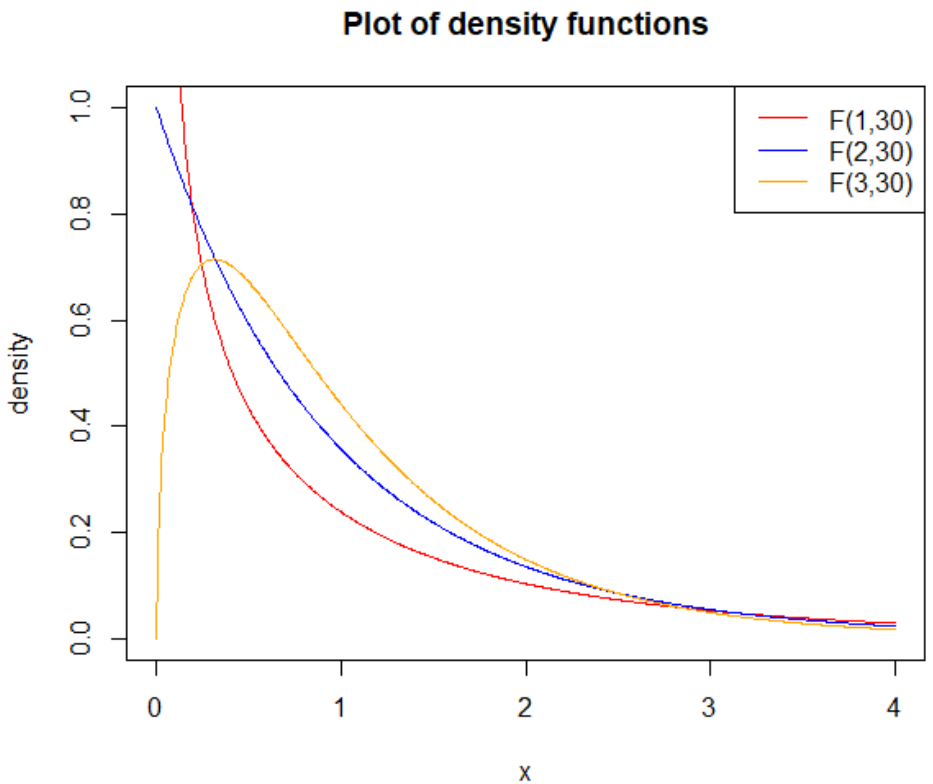
The square of this RV follows the same distribution as F^*, or more formally

$$T^2_{\hat{\beta}_1} \stackrel{d}{=} F^*$$

where $\stackrel{d}{=}$ denotes **equality in distribution**, a weaker form of equality used frequently in probability and statistics.

Model fit – ANOVA and F-test

The F-distribution has two parameters and a selection of PDFs for a variety of parameter choices is displayed below.



The extreme events occur in the upper tail, so the null hypothesis rejection region is in the upper tail.

Model fit – ANOVA and F-test

Hypotheses

In the context of simple linear regression, the hypotheses are

$$H_0: \ \beta_1 = 0$$
$$H_A: \ \beta_1 \neq 0.$$

Test statistic

The test statistic

$$f^* = \frac{msr}{mse}$$

is calculated from the data and is an observation of the RV F^* in (1).

Test decision

H_0 is rejected at significance level α if

$$f^* > f_{1-\alpha},$$

where the quantile $f_{1-\alpha}$ is from $F(1, n-2)$ distribution.

Model fit – ANOVA and F-test

Equivalently, H_0 is rejected if the p-value

$$p = \text{Prob}(F > f^*) < \alpha,$$

where $F \sim F(1, n-2)$.

The null hypothesis H_0 is retained if this is not the case.

Model fit – ANOVA and R^2

The decomposition of total squared variation

$$SST = SSR + SSE$$

can also be used to provide a quantitative measure of model fit.

The larger the proportion of total squared variation explained by the model, the better the fit of the model to the data.

We define this proportion as the **coefficient of determination**

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

which satisfies $0 \leq R^2 \leq 1$.

For simple linear regression, R^2 provides a tool for comparing alternative models on the same data set, looking for the model that, all else being equal, has the higher R^2.

For multiple regression a modification to this statistic is necessary.

Model fit – leverage and influence

The method of least squares involves minimising the sum square error.

Consider two data points, y_i and y_j with residuals satisfying

$$\hat{\epsilon}_i = 2\hat{\epsilon}_j.$$

The ratio of the contribution that these two data points make to *sse* is

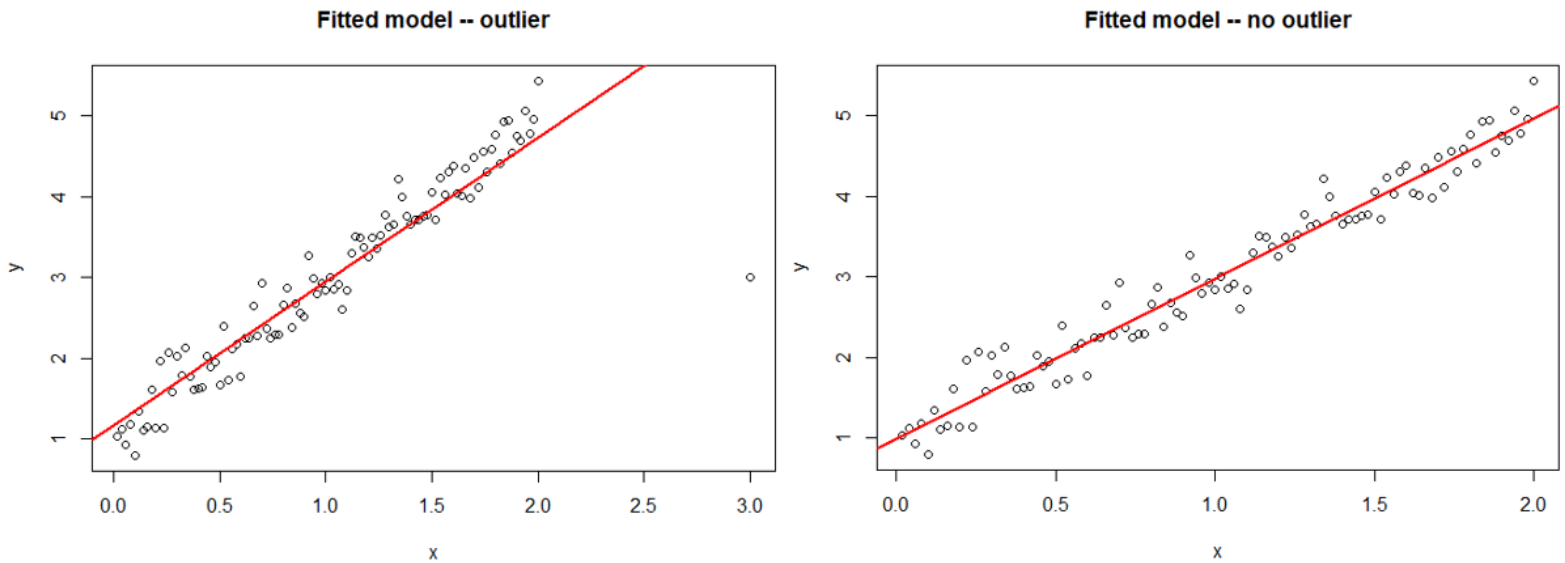
$$\frac{\hat{\epsilon}_i^2}{\hat{\epsilon}_j^2} = 4.$$

Through the minimisation of *sse*, one could reasonably expect y_i to have greater **influence** on the estimation of model parameters than y_j.

This is just a consequence of the model selection process, but sometimes it can lead to unexpected and undesirable effects.

Model fit – leverage and influence

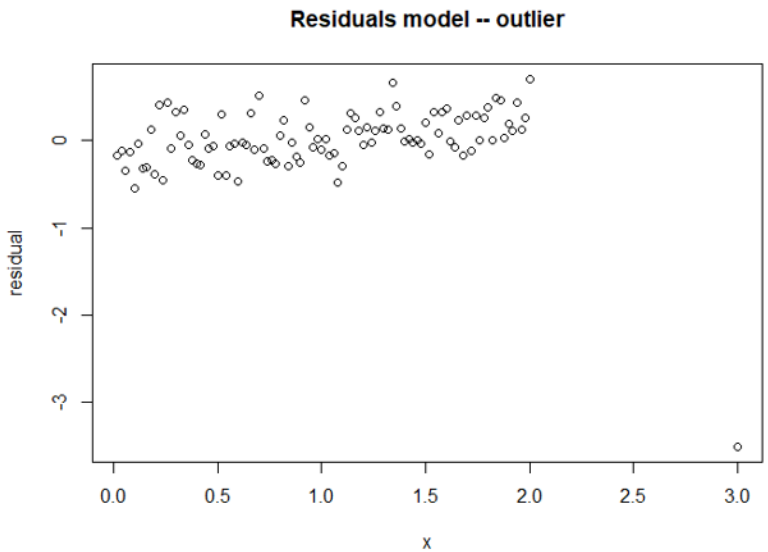
The following example shows least squares models fitted to data sets differing in one component only – in the second model a single **outlier** has been removed.

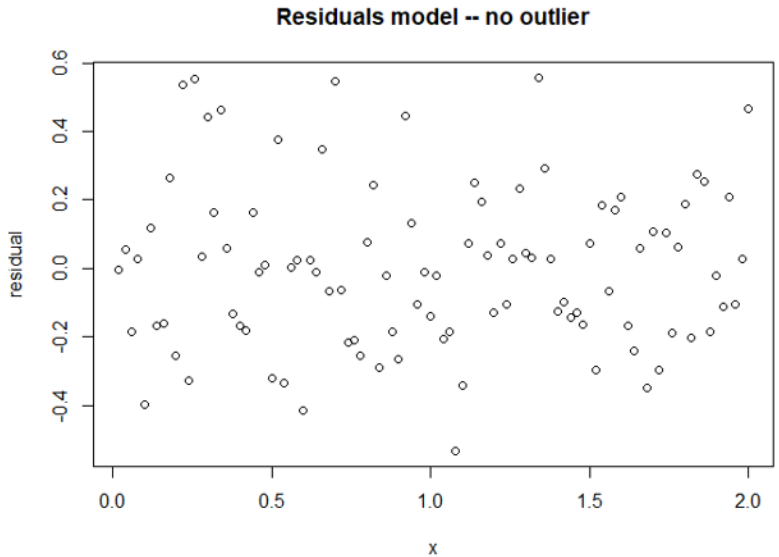


Notice the change in the fitted regression line and increase in R^2 (see R code file)?

Model fit – leverage and influence

Another way to look at this example is by comparing before and after scatter plots of **residuals** against the predictor x.





We will return to residual analysis later.

Model fit – leverage and influence

The relative importance of a point to a model can be quantified in terms of its **leverage**, defined as

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{s_{xx}} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

which we see is a function of the distance between the predictor variable and its sample mean.

We can use leverage to quantify the **influence** a point has on the overall regression model.

Model fit – leverage and influence

To do so we need to first define the **internally Studentised residual**

$$\hat{t}_i = \frac{\hat{\epsilon}_i}{s\sqrt{1 - h_{ii}}} \tag{2}$$

and **externally Studentised reisudal**

$$\hat{d}_i = \frac{\hat{\epsilon}_i}{s^{(i)}\sqrt{1 - h_{ii}}}, \tag{3}$$

where $s^{(i)}$ is the estimate s recalculated after exclusion of observation i.

These quantities weight the residuals $\hat{\epsilon}_i$ according to their leverage – the larger the leverage of a point i the larger $\hat{t}_i$ and $\hat{d}_i$.

The motivation behind use of these modified residuals is that the residuals $\hat{\epsilon}_i$ do not have exactly the same variance/covariance structure as the noise terms ϵ_i (we will return to this next chapter).

Model fit – leverage and influence

COOK'S DISTANCE

One such statistic is **Cook's *D***, which for the **case of *m* independent variables** is defined as

$$D_i = \frac{1}{m} \frac{h_{ii}}{1 - h_{ii}} \hat{t}_i^2$$

with $\hat{t}_i$ the internally-Studentised residual defined in (2).

This statistic can be used to asses the sensitivity of the estimated model parameters to the removal of the i-th observation from the sample data.

As a **rule of thumb**, data points with

$$D_i > \frac{4}{n - m - 1}$$

are considered potentially influential.

When potentially influential points are identified, the model is re-run on the reduced data set excluding these points and compared to the full data set model – significant changes in the parameter estimates confirm the excluded points as influential.

Model fit – leverage and influence

DFITS

A similar statistic is **DFITS**, which for the i-th data point is defined as

$$DFITS_i = \hat{d}_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}}$$

with $\hat{d}_i$ the externally-Studentised residual defined in (3).

$DFITS_i$ is another measure of the sensitivity of the model to the removal of the i-th observation from the sample data.

As a **rule of thumb**, data points with

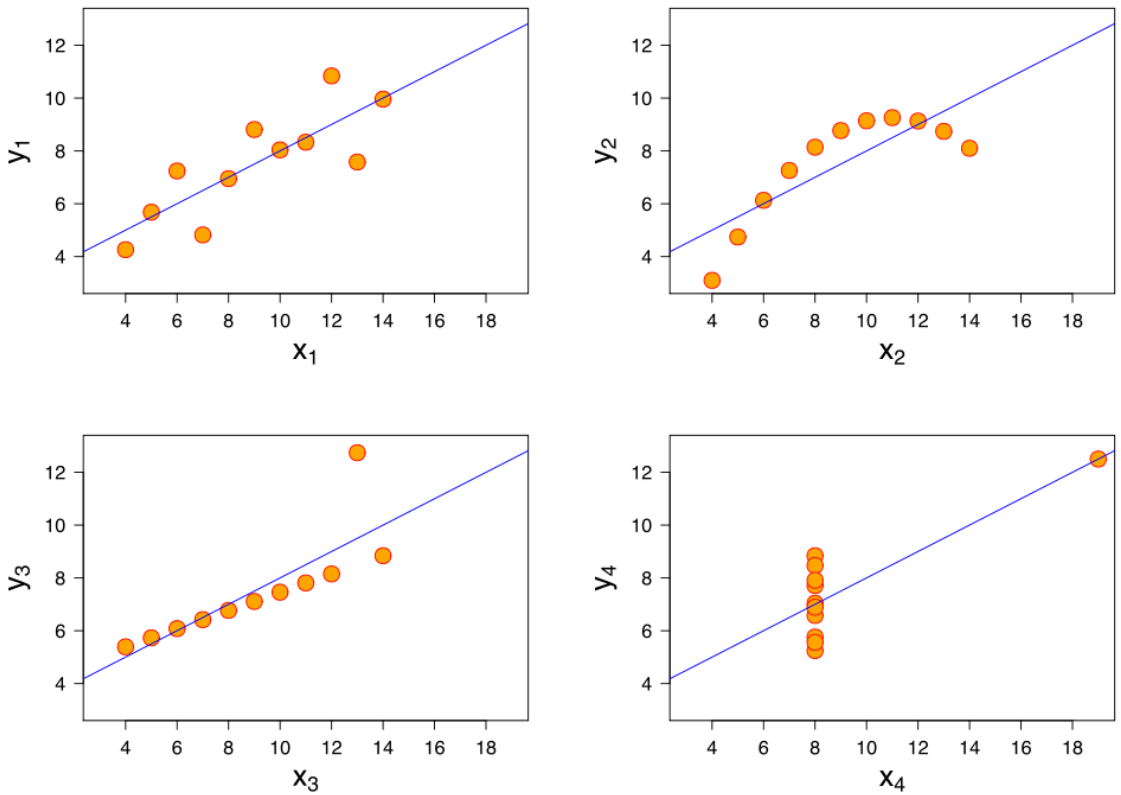
$$|DFITS_i| > 2\sqrt{\frac{m}{n}},$$

m the number of predictors, should be considered potentially influential.

Model fit – leverage and influence

We finish this section with a famous example of Anscombe (1973).

Each regression line is identical, each data set is very different.



Example: Anscombe's Quartet.
Source: https://commons.wikimedia.org/wiki/File:Anscombe.svg

Checking model assumptions

When using the method of least squares to fit a model to data we needed no assumptions.

We used the method of least squares and found the model that minimises *sse*.

But we have gone much further than this and developed tools to place CIs on parameter estimates and model predictions and tools to assess model fit.

Along the way we have relied on assumptions about the properties of the residuals, assumptions that are now embedded in the methods that have been developed.

The **most important** part of building a model is **justifying the assumptions** on which the model relies.

Checking model assumptions

Recall the model assumed to describe the population was

$$Y = \beta_0 + \beta_1 x + \epsilon$$

with our model fitted to sample data

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

The critical assumptions we made were that noise or errors were independent and

$$\epsilon_i \sim N(0, \sigma^2).$$

However, we don't have access to the error terms, but we do have access to the residuals.

As estimates of the noise terms ϵ_i, the residuals $\hat{\epsilon}_i$ should behave similarly – we check the assumptions on ϵ_i via $\hat{\epsilon}_i$.

Checking model assumptions – residual analysis

Verification of model assumptions boils down to analysis of residuals.

This analysis can be performed in two complementary ways:

1. visual inspection via plots
2. numerical inspection via statistical tools.

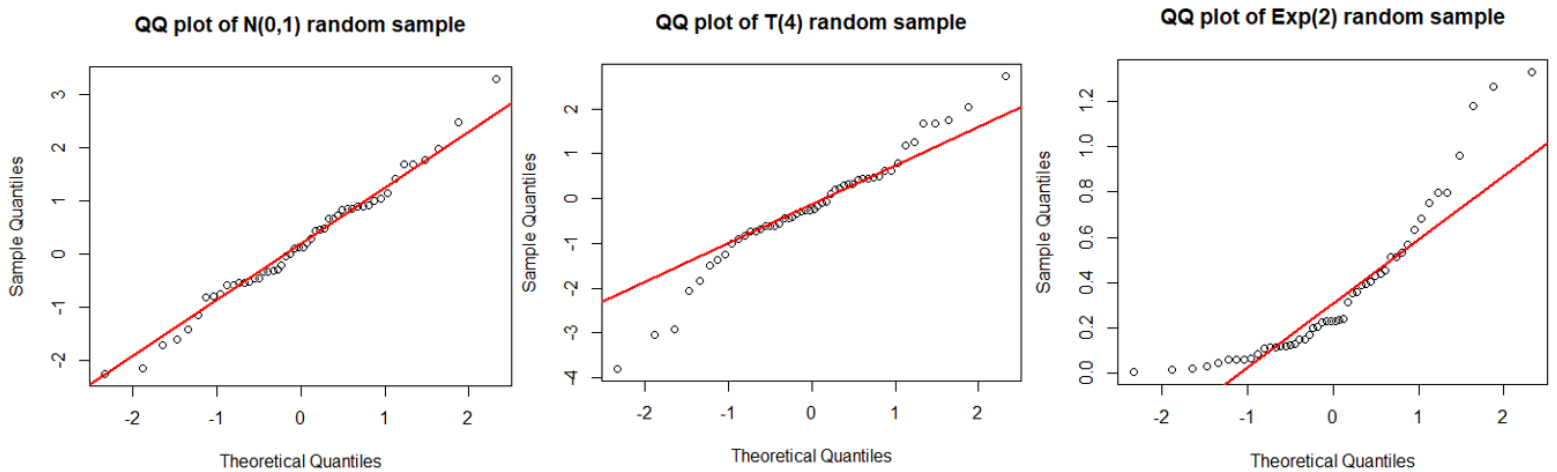
When we look for visual clues, essentially we are looking for some patterns that affirm the assumptions and other patterns that contradict.

We can also use various statistical tools to identify the same sort of behaviour.

Checking model assumptions – visual residual analysis

NORMALITY – QQ PLOTS.

To asses the normality of the residuals visually, we use QQ plots, three examples of which are below. The red line is where the sample quantiles equal the quantiles from a normal distribution.



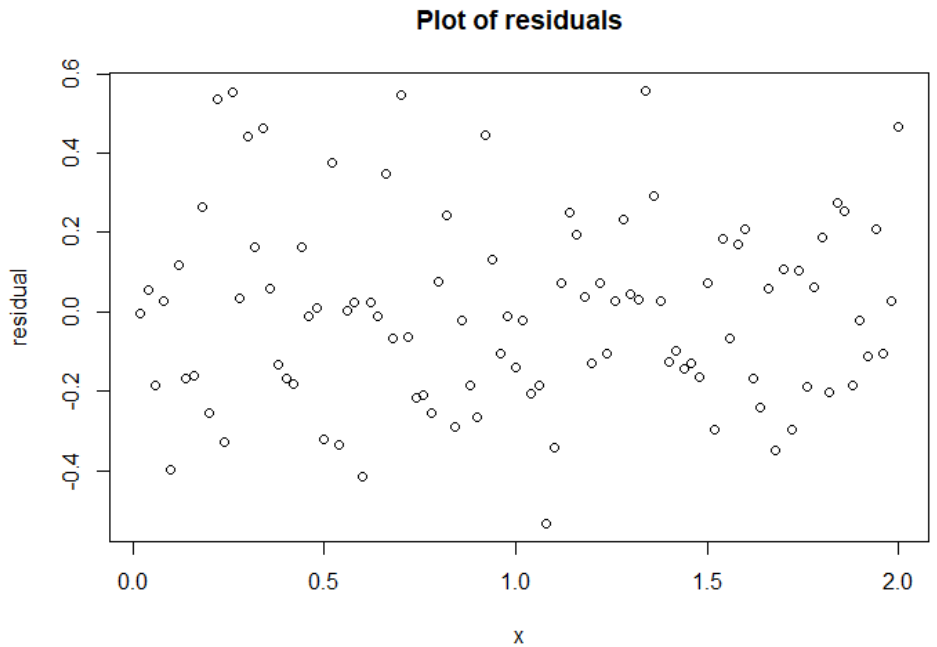
The plot on the left shows a sample from N(0,1) distribution – we see the data closely tracking the red line indicating normally distributed behaviour.

The other two plots show samples from non-normal distributions – we see departures from the red lines indicating non-normal behaviour.

Checking model assumptions – visual residual analysis

INDEPENDENCE – SCATTER PLOTS

To asses the independence of the residuals visually, we use scatter plots, an example of which is below.



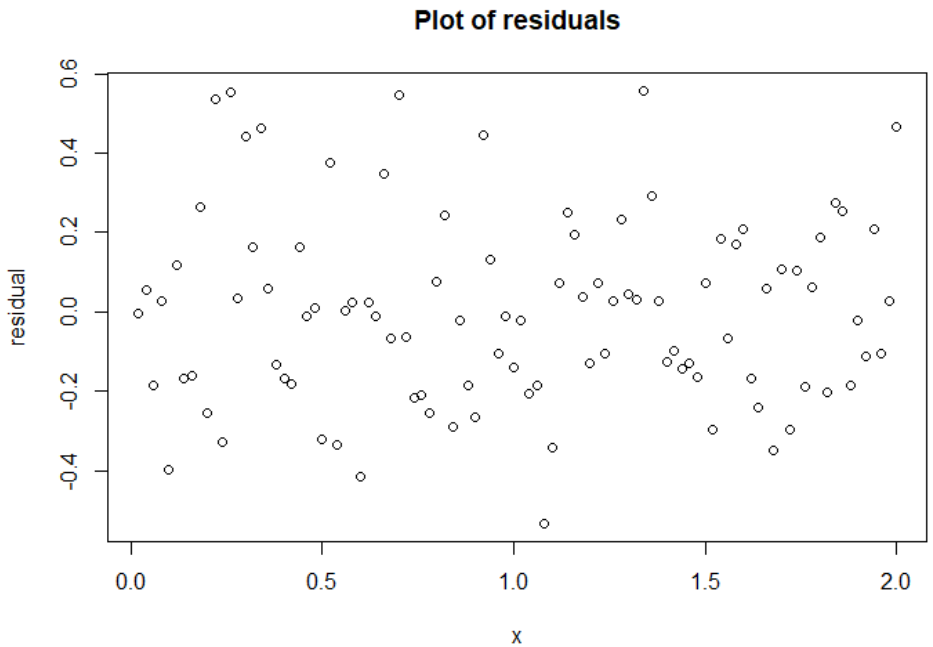
If the residuals are independent we should see no structure or patterns in the plot.

Instead we should just see randomness, which is what we see above, indicating no problem with independence.

Checking model assumptions – visual residual analysis

CONSTANT VARIANCE – SCATTER PLOTS

To asses the variance of the residuals visually, we again use scatter plots.

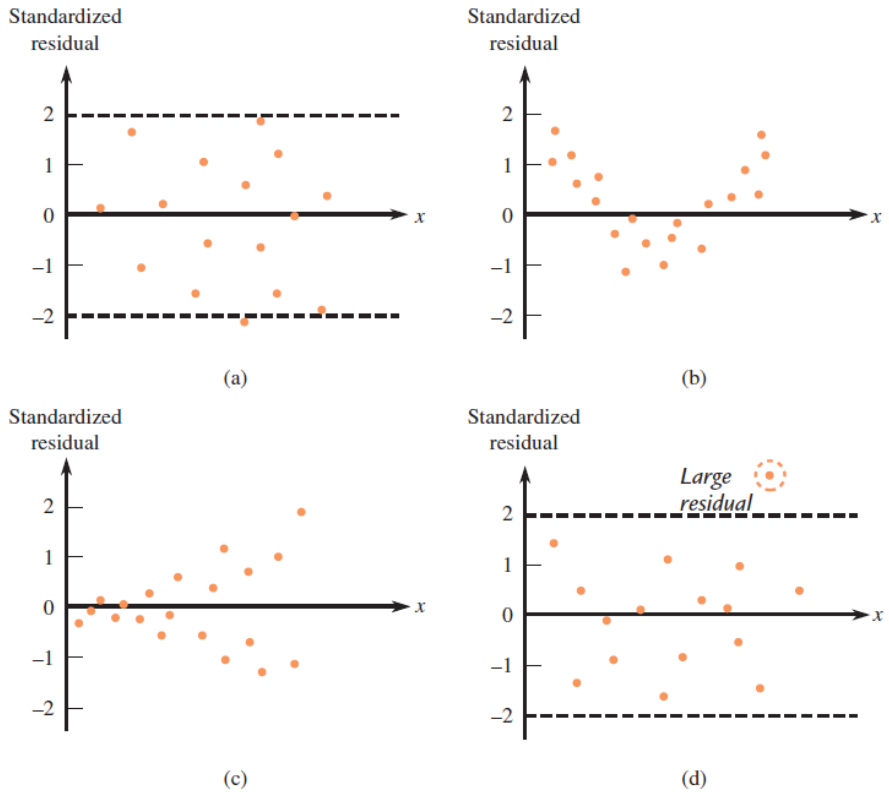


If the residuals are of constant variance, then the vertical range of the points in the plot should be fairly consistent.

That is what we see above, indicating no problem with constant variance.

Checking model assumptions – visual residual analysis

Below are some plots of residuals that are typical in regression analysis.



Source: Peck et al. (2012) page 769

Plot (a) is fine, (b) shows a pattern, (c) shows increasing variance and (d) shows a large residual.

Checking model assumptions – statistical residual analysis

NORMALITY – TEST

There are a variety of normality tests, including the Shapiro-Wilk, Anderson-Darling, Kolmogorov-Smirnov etc.

They all have different test statistics, details of which we won't go into.

Hypotheses

H_0: the residuals $\hat{\epsilon}_i$ are normally distributed

H_A: the residuals $\hat{\epsilon}_i$ are not normally distributed.

Test decision

The null hypothesis is rejected if the p-value $p < \alpha$.

Conclusion

If the null hypothesis is rejected, then there is significant evidence that the residuals are not normally distributed.

Checking model assumptions – statistical residual analysis

INDEPENDENCE – DURBIN-WATSON TEST

The Durbin-Watson test is a test for **autocorrelation** and so can be used to asses the independence assumption.

Hypotheses

H_0: autocorrelation is not present in the residuals

H_A: autocorrelation is present in the residuals.

Test statistic

Without going into details, the statistic is calculated as

$$dw = \frac{\sum_{i=2}^{n}(\hat{\epsilon}_i - \hat{\epsilon}_{i-1})^2}{\sum_{i=1}^{n}\hat{\epsilon}_i^2}.$$

Test decision

The null hypothesis is rejected if the p-value $p < \alpha$.

Conclusion

If the null hypothesis is rejected, then there is significant evidence that the residuals are not independent.

Checking model assumptions – statistical residual analysis

The value of the test statistic not only tells us if autocorrelation is present, but it also tells as what type.

It can be shown that

$$0 \leq dw \leq 4$$

with $dw < 2$ suggesting positive autocorrelation and $dw > 2$ suggesting negative autocorrelation.

Recall that our assumption is that the residuals are uncorrelated, so we are looking for values of close $dw = 2$.

As a **rule of thumb**, if the test statistic takes values

$$dw < 1.5 \quad \text{or} \quad dw > 2.5,$$

then this can be taken as evidence of significant autocorrelation in the residuals (note some statisticians use the values 1 and 3 rather than 1.5 and 2.5).

R example continued

Let's return to our SPSS example and apply what we have learned today.

Below is an extract of the summary information generated by R.

```
Residual standard error: 5.306 on 185 degrees of freedom
Multiple R-squared:  0.6411,    Adjusted R-squared:  0.6391
F-statistic: 330.4 on 1 and 185 DF,  p-value: < 2.2e-16
```

With $R^2 = 0.641$ we see that the model explains around 64% of the squared variation in the response data about the sample mean.

Below is the output from the Durbin-Watson test.

```
        Durbin-Watson test

data:  life.model
DW = 2.3239, p-value = 0.02679
alternative hypothesis: true autocorrelation is not 0
```

We also see a Durbin-Watson statistic of $dw = 2.324$ which, although not excessive, does point to possible negative autocorrelation in the residuals.

Also, p-value of 0.02679 is less that usual significance $\alpha = 0.05$.

R example continued

Below is the ANOVA table.

```
Analysis of Variance Table

Response: life
           Df Sum Sq Mean Sq F value    Pr(>F)
gniLog      1 9301.1  9301.1  330.41 < 2.2e-16 ***
Residuals 185 5207.7    28.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

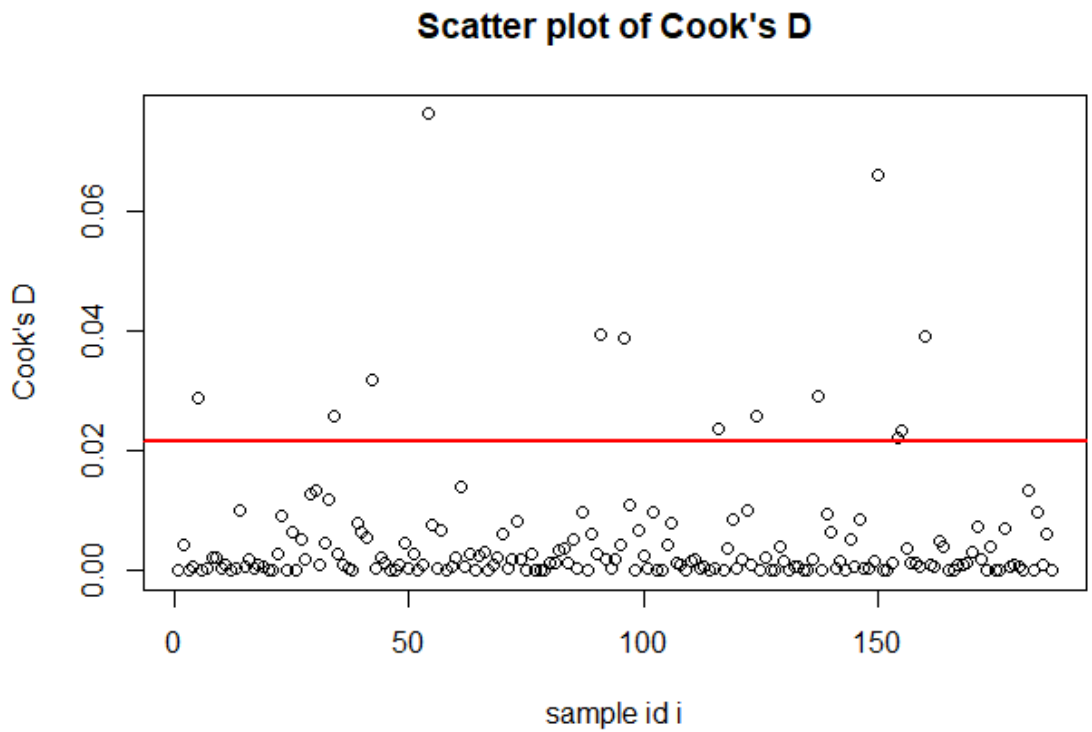
We see the decomposition of total variation *sst* in terms of that due to the model *ssr* and that due to the errors *sse*.

With an F-test statistic of $f^* = 330.41$ and associated p-value reported as $p < 2.2 \times 10^{-16}$ (less than our usual significance level $\alpha = 0.05$), we reject the null hypothesis that $\beta_1 = 0$.

R example continued

Next we look for potentially influential points with a scatter plot of Cook's distance D_i, the red line showing the critical value of Cook's D

$$D_{\text{critical}} = \frac{4}{n-m-1} = \frac{4}{187-1-1} \approx 0.022.$$



The sample points above the red line are potentially influential.

R example continued

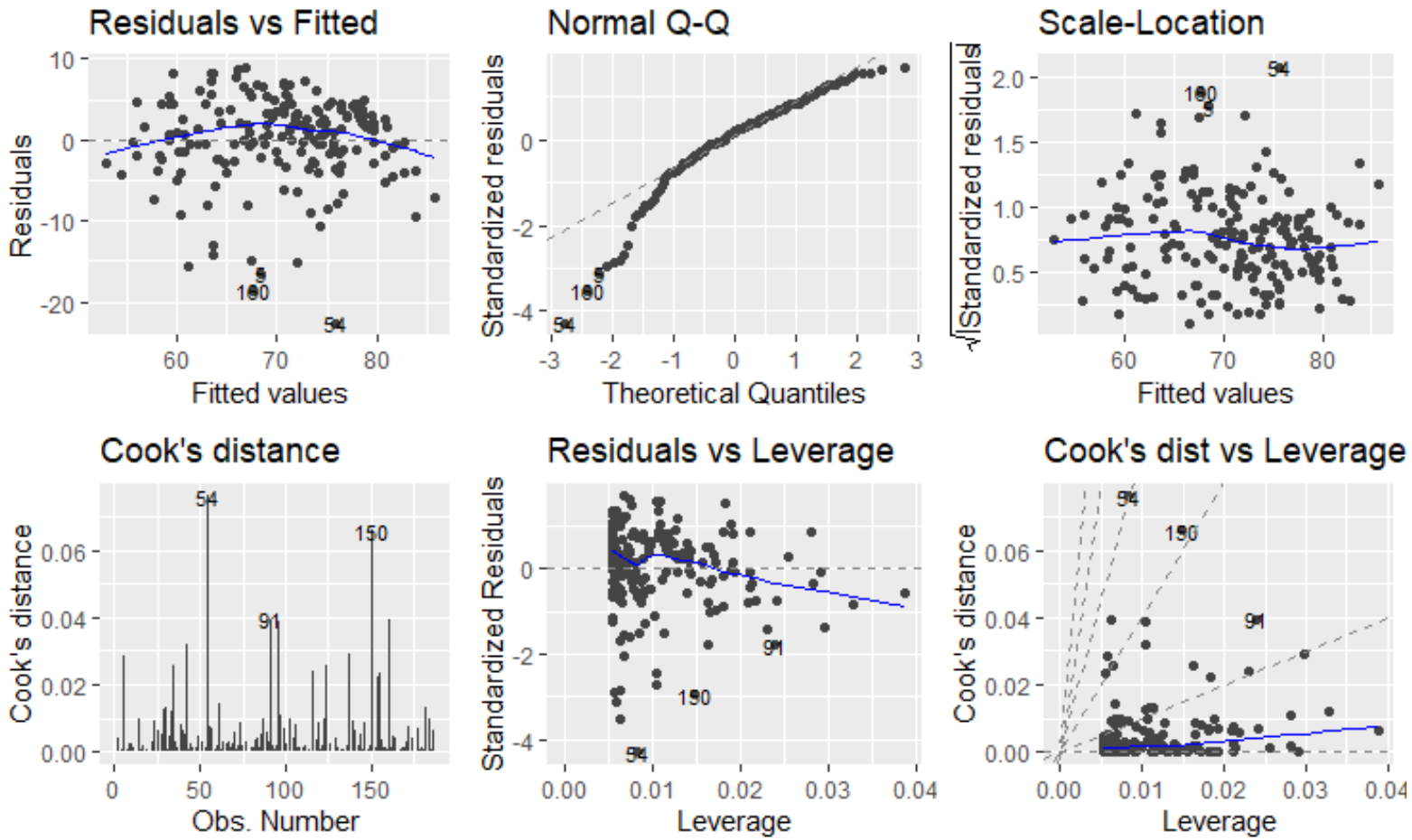
We can have R identify and display the influential points.

	life	gniLog	cooksD
5	51.9	8.751949	0.02873565
34	51.2	7.391415	0.02583807
42	50.7	7.928046	0.03197004
54	53.1	9.997524	0.07612298
91	74.3	11.360007	0.03949446
96	49.4	7.936660	0.03870509
116	50.3	6.918695	0.02377007
124	52.5	8.585412	0.02576283
137	78.4	11.687122	0.02926973
150	45.6	7.503841	0.06595971
154	67.7	7.233455	0.02228774
155	56.9	9.374837	0.02324901
160	49.0	8.619027	0.03923312

We can remove these and create the reduced data set model (see R code file).

R example continued

Now we look at diagnostic plots for the full data set model.



R example continued

Normality. In the "Normal Q-Q" plot, we see significant departure from the line representing normally distributed behaviour – problem with this assumption.

Independence. In the "Residuals vs Fitted" plot, we see signs of negative curvature – problem with this assumption.

Equal variance. In the "Residuals vs Fitted" plot, we see a larger range of residuals in the centre of the plot than on the sides – problem with this assumption.

In this case, all of these assumptions do not appear to have been met, indicating problems with our fitted model.

R example continued

Finally we run a normality test on the residuals for the full data set model.

Define the hypotheses

H_0: the residuals $\hat{\epsilon}_i$ are normally distributed

H_A: the residuals $\hat{\epsilon}_i$ are not normally distributed.

The output from R is below.

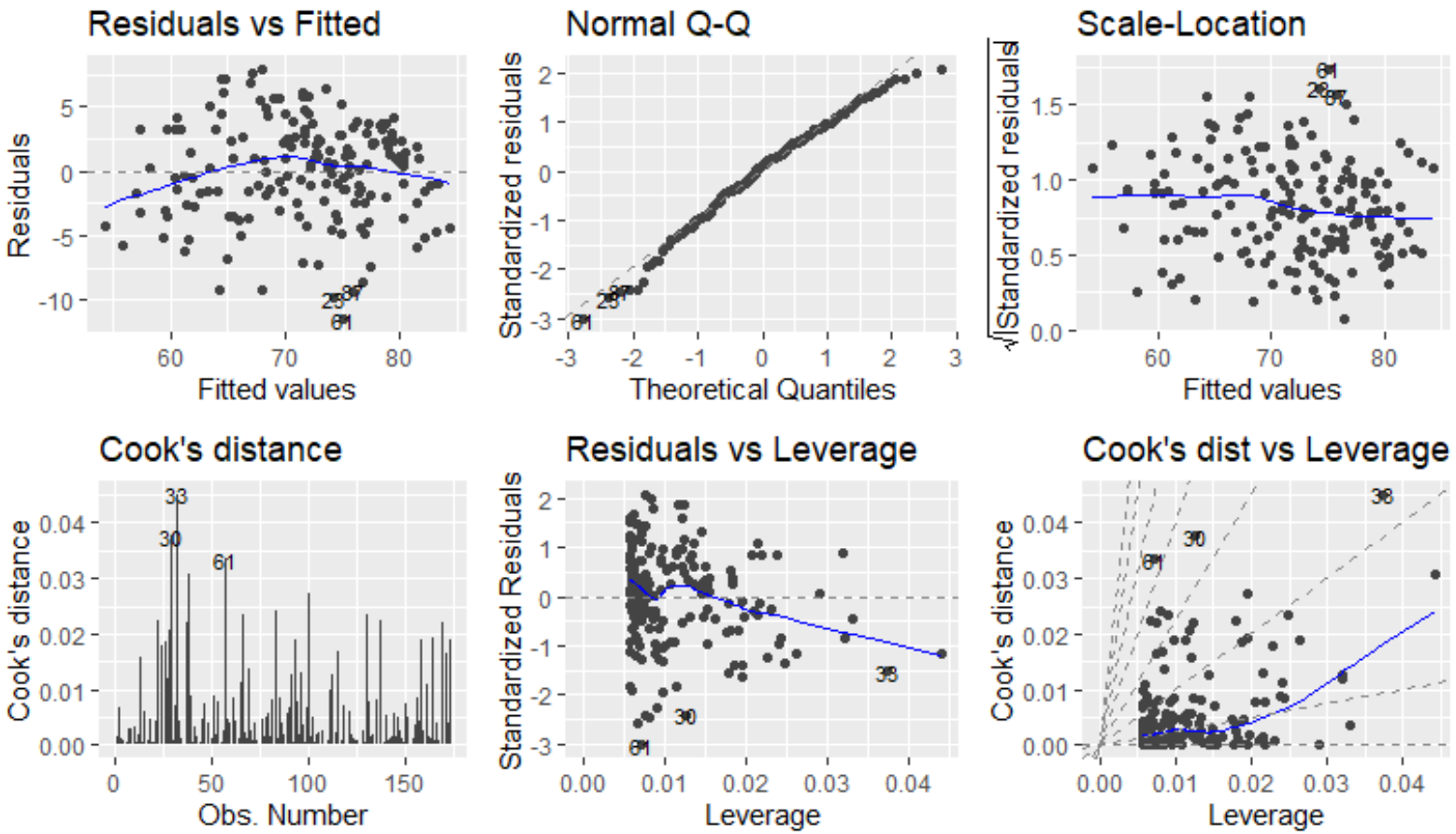
```
        Shapiro-Wilk normality test

data:  life.data$resid
W = 0.90871, p-value = 2.395e-09
```

The p-value is less than our usual significance level of $\alpha = 0.05$, so we reject the null hypothesis and conclude there is strong evidence that the residuals are not normally distributed.

R example continued

Now we look at diagnostic plots for the reduced data set model.



Removing the points with large Cook's D appears to have improved the behaviour of the residuals in terms of the modelling assumptions.

R example continued

Finally we run a normality test on the residuals for the reduced data set model.

Define the hypotheses

H_0: the residuals $\hat{\epsilon}_i$ are normally distributed

H_A: the residuals $\hat{\epsilon}_i$ are not normally distributed.

The output from R is below.

```
        Shapiro-Wilk normality test

data:  life.data.reduced$resid
W = 0.98489, p-value = 0.057
```

The p-value is greater than our usual significance level of $\alpha = 0.05$, so we retain the null hypothesis and conclude there is no strong evidence that the residuals are not normally distributed.

References I

Anscombe, F. J. (1973). Graphs in statistical analysis. *American Statistician*, 27:17–21.

Peck, R., Olsen, C., and Devore, J. (2012). *Introduction to statistics & data analysis*. Brooks/Cole, 4th edition.