Regression and Linear Models (37252)

Lecture 3 - Simple Linear Regression II

Lecturer: Joanna Wang
Notes adopted from Scott Alexander

School of Mathematical and Physical Sciences, UTS

Autumn 2024

Acknowledgement

These notes are based, in part, on earlier versions prepared by Dr Ed Lidums and Prof. James Brown.

3. Lecture outline

Topics:

- revisiting the R example from Lecture 2
  - T-tests
- taking stock
- model fit
  - ANOVA and F-test
  - ANOVA and R^2
  - leverage and influence
- checking model assumptions
  - residual analysis
- continuing the R example

3. R example Lecture 2

Consider again the example from last week's lecture.

Recall we looked to model the relationship between the logarithm of per-capita GNI and life expectancy, using the country-by-country data contained in the R data set newdata.csv.

We began by proposing an underlying, theoretical reality

$$LifeExp|LogGNI = \beta_0 + \beta_1 \times LogGNI + \epsilon$$

and looked to build the model

$$\widehat{LifeExp}|LogGNI = \hat{\beta}_0 + \hat{\beta}_1 \times LogGNI$$

as an approximation of

$$\mathbb{E}[LifeExp|LogGNI] = \beta_0 + \beta_1 \times LogGNI.$$

3. R example Lecture 2

The **Coefficients** table returned by R

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     17.2798     2.9577   5.842 2.28e-08 ***
newdata$LogGNI  13.4659     0.7408  18.177  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
```

allowed us to extract the parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ and define the simple linear regression equation for our model as

$$\widehat{LifeExp}|LogGNI = 17.280 + 13.466 \times LogGNI.$$

3. R example Lecture 2 – T-tests

The T-statistics returned by R

$$t^*_{\hat{\beta}_0} = \frac{\hat{\beta}_0 - \beta^*_0}{S_{\hat{\beta}_0}} = \frac{\hat{\beta}_0 - 0}{S_{\hat{\beta}_0}} = \frac{17.280}{2.958} = 5.842$$

and

$$t^*_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - \beta^*_1}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - 0}{S_{\hat{\beta}_1}} = \frac{13.466}{0.741} = 18.177$$

are associated with the single-sample, two-tail T-tests on the 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.

3. R example Lecture 2 – T-tests

The critical value associated with a two-tail T-test with $n - 2 = 185$ degrees of freedom is

$$t_{0.975} = 1.97287,$$

allowing each null hypothesis H_0 of the preceding tests to be rejected in favour of the alternatives H_A as

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

Of course, we can see this directly from the p-values or the 95% CIs (using R command **confint(mod1)**)

$$11.445 \leq \beta_0 \leq 23.115$$

$$12.004 \leq \beta_1 \leq 14.927$$

of which zero is not a member.

Note. The constants $t^*_{\hat{\beta}_0}$ and $t^*_{\hat{\beta}_1}$ are, respectively, the values that the RVs $T_{\hat{\beta}_0}$ and $T_{\hat{\beta}_1}$ have taken for this particular set of data (see Week 2 Lecture Notes).

3. R example Lecture 2 – T-tests

We can use the information in the R Coefficients table to test other hypotheses on the true values of the parameters.

For instance, to test the hypothesis that $\beta_0 > 15$ define the hypotheses

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

The statistic for this upper-tail single-sample T-test is calculated as

$$t^*_{\hat{\beta}_0} = \frac{\hat{\beta}_0 - \beta^*_0}{S_{\hat{\beta}_0}} = \frac{17.280 - 15}{2.958} \approx 0.771$$

which has an associated p-value of $p \approx 0.221$ (R command **1-pt(0.771, 185)**).

As $p > \alpha = 0.05$ (the significance level we have selected for this example) we cannot reject H_0.

3. Taking stock

So far we have looked at fitting a simple linear model and assessing it in terms of CIs on the estimated parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ and CIs of its predictions of $\mathbb{E}[Y|x]$ and $Y|x$.

We saw that the parameter CIs were returned by R and that the prediction CIs could be calculated after running the model.

But from last week's lab you will have noticed that R outputs much more than just the Coefficients table.

In this week's lecture we will look deeper into the problem of linear regression and begin to answer some of the questions posed at the end of Week 2 Lecture Notes.

3. Model fit – ANOVA and F-test

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

Consider the **total sum of squares** or **total variation** about the mean

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

With a little algebra this can be decomposed as

$$SST = SSR + SSE, \tag{2}$$

where the total sum of squares due to the regression is given by

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 \tag{3}$$

and the sum of squared errors

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

3. Model fit – ANOVA and F-test

Under the assumption that the residuals $\hat{e}_i \sim N(0, \sigma^2)$ and independent, if

$$\beta_1 = 0$$

then the RV

$$F = \frac{SSR}{S^2} = \frac{SSR}{\frac{SSE}{n-2}} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y}_i)^2}{\frac{\sum_{i=1}^{n}\hat{\epsilon}_i^2}{n-2}} \tag{4}$$

follows an $F(1, n-2)$ distribution.

Note: there is a more general description of this RV F in terms of a quotient of chi-squared RVs, but we'll leave this until required.

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

3. 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 $\mathrm{T}(n-2)$ RV has the same distribution as an $\mathrm{F}(1, n-2)$ RV.

Recall the $\mathrm{T}(n-2)$ RV used for the T-test on β_1, which with hypothesised value $\beta_1 = 0$ has the form

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

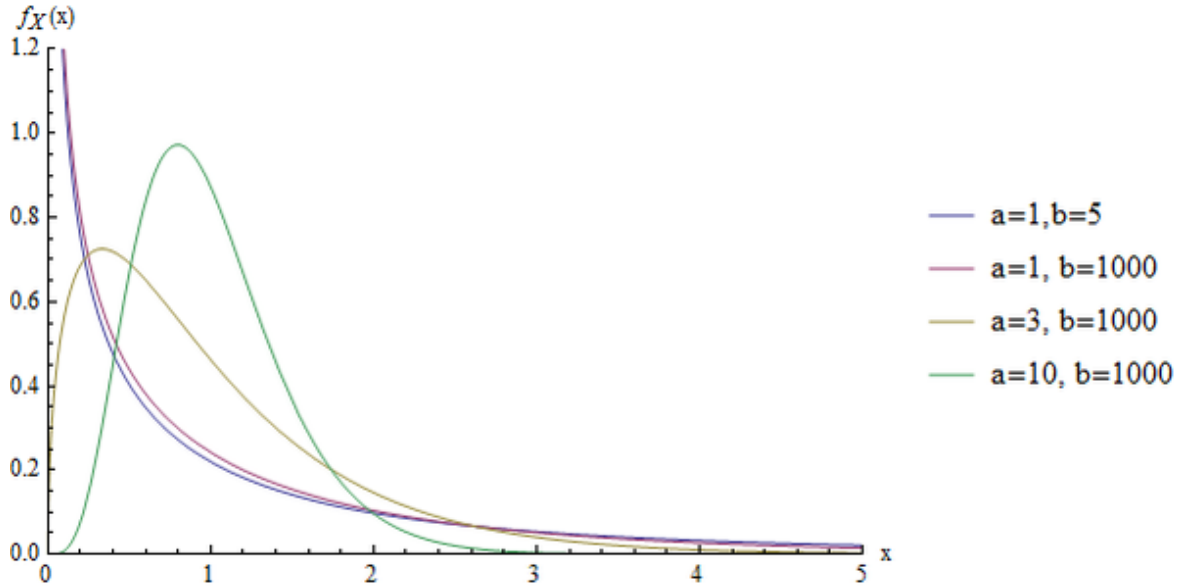
The square of this RV follows the same distribution as F, or more formally

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

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

3. Model fit – ANOVA and F-test

The F-distribution has two parameters, details of which we will go into next week. The associated PDFs for a selection of parameter values are displayed below.



Example: PDF of $X \sim \mathrm{F}(a, b)$ RV

The most extreme events occur in the upper tail, so we use upper tail hypothesis tests.

3. Model fit – ANOVA and F-test

F-test hypotheses

In the context of simple regression, the hypotheses are

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

Test statistic

The test statistic f^* is the value that the RV F, given in (4), takes for the particular model built on the data sample (X_i, Y_i), $i \in \{1, \ldots, n\}$.

Rejection of null hypothesis

H_0 is **rejected** in favour of H_A at **significance level** α if

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

where $f_{1-\alpha}$ is the **quantile** satisfying

$$\text{Prob}\big(F > f_{1-\alpha}\big) = \alpha.$$

3. Model fit – ANOVA and F-test

Rejection of null hypothesis continued

Equivalently, H_0 is rejected if the **p-value**

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

The null hypothesis H_0 is **retained** in any other case.

The argument used here is that if it is so unlikely for an $F(1, n-2)$-distributed RV F to take values at least as large as f^*, then perhaps F cannot be so distributed.

But we know it must be if $\beta_1 = 0$ and so reject this hypothesis in favour of the alternative $\beta_1 \neq 0$.

3. Model fit – ANOVA and R^2

The decomposition of total variation given by (2) can also be used to provide a quantitative measure of model fit.

Recall this decomposition

$$SST = SSR + SSE.$$

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

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

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{5}$$

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

For simple linear regression, R^2 provides a tool for comparing alternative models, looking for the model that, all else being equal, has the higher R^2. For multiple regression a modification is necessary.

3. Model Fit – leverage and influence

The method of least squares, and measures such as R^2, use, as a definition of variation, squared errors.

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

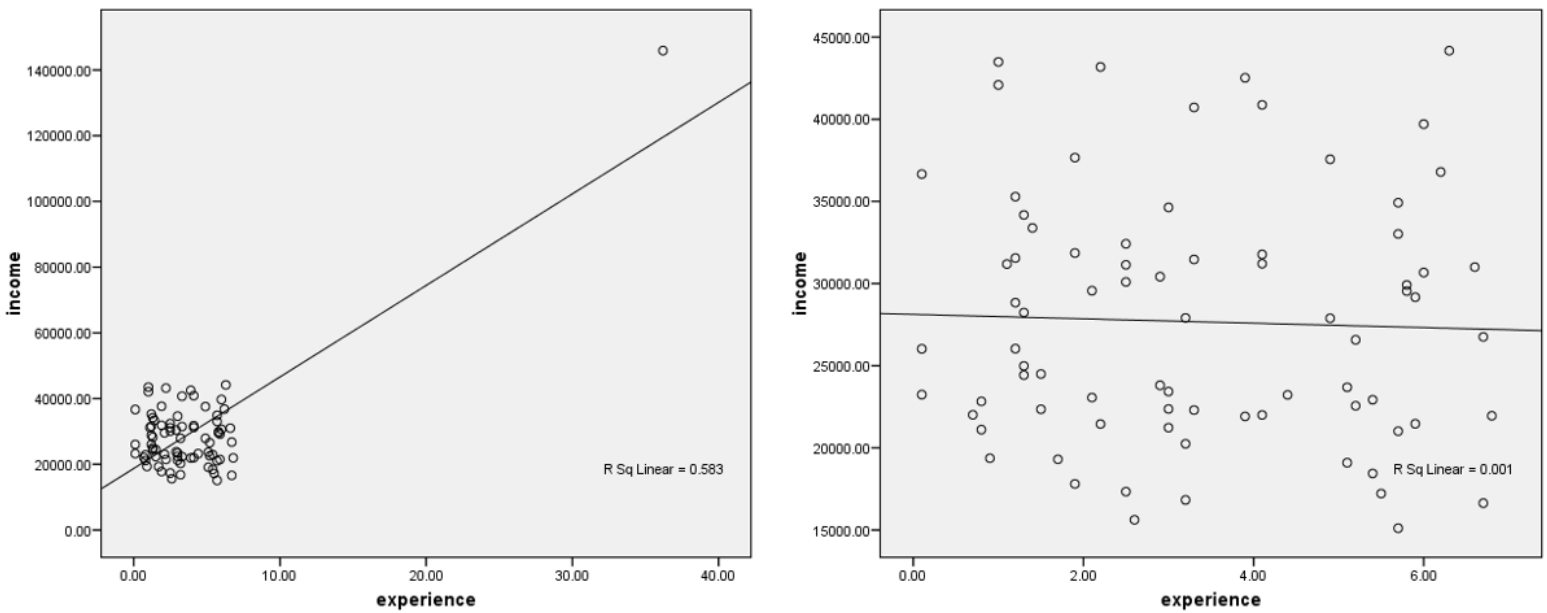
$$\frac{\hat{\epsilon}_i^2}{\hat{\epsilon}_j^2} = 4.$$

One could reasonably expect Y_i to have greater **influence** on model parameters than Y_j through the minimisation of SSE.

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

3. 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.

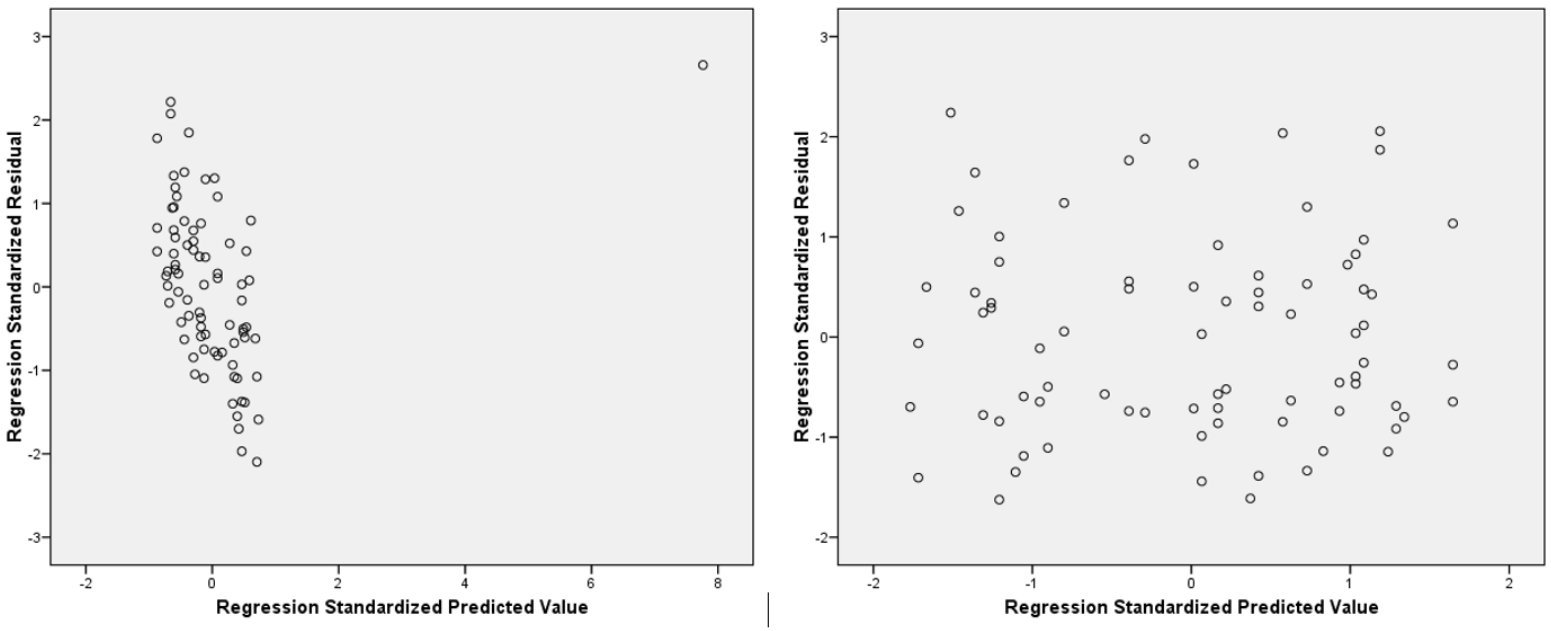


Example: Leverage, influence and model selection

The radical change in the slope of the line and reduction in R^2 says it all.

3. Model Fit – leverage and influence

Another way to look at this example is by comparing before and after scatter plots of **standardised predictions** against **standardised residuals** (we will come back to residual analysis later).



Example: Leverage, influence and model selection

The second plot appears closer to that expected of two standard normal independent RVs (no extreme outlier).

3. Model Fit – leverage and influence

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

$$h_{i,i} = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{S_{XX}} = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}, \tag{6}$$

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

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

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 (10) used in residual analysis.

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

3. Model Fit – leverage and influence

As a rule of thumb, data points with $D_i > \frac{4}{n-m-1}$ (or $D_i > \frac{4}{n}$) are considered potentially influential with $D_i > 1$ a strong indicator.

Another 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 (11).

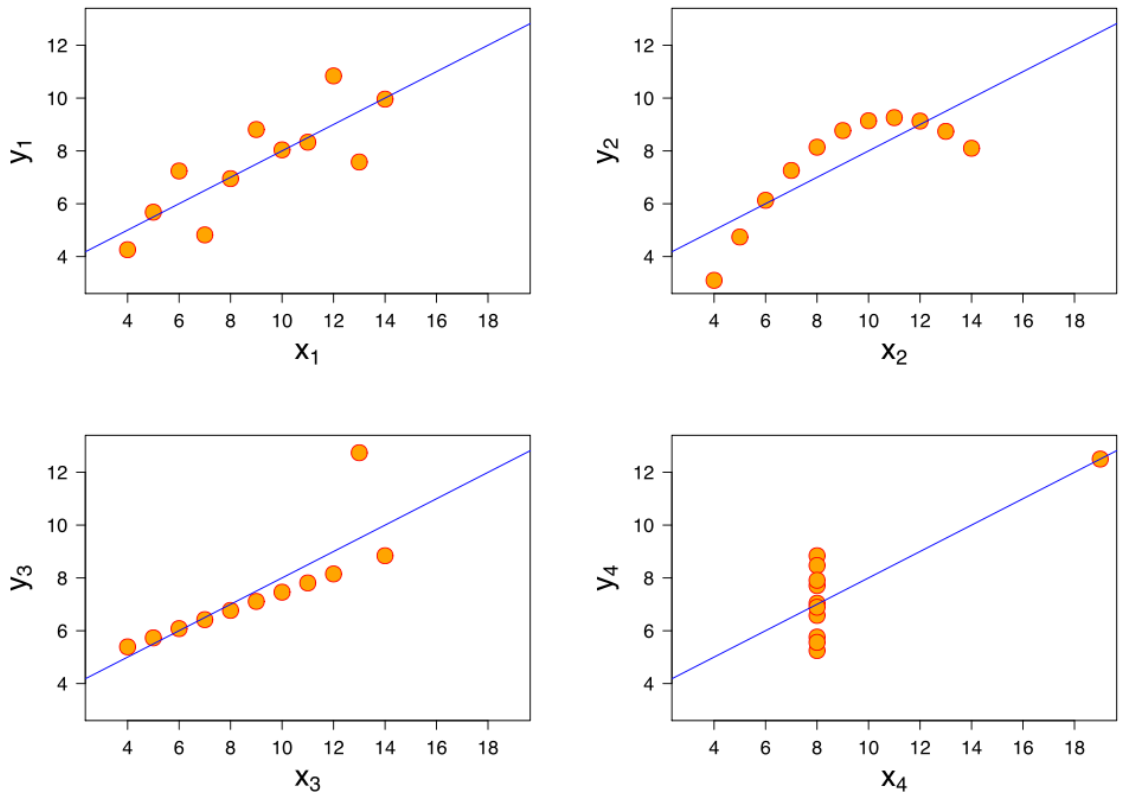
$DFITS_i$ is another measure of the sensitivity of the model to the removal of the i-th observation from the sample data.

As a rule of thumb, data points with $|DFITS_i| > 2\sqrt{\frac{m+1}{n-m-1}}$, m the number of independent variables, should be considered possibly influential.

3. 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

3. 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 minimised 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 an assumption about the properties of the residuals, an assumption that is now embedded in the methods that have been developed.

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

3. Checking model assumptions

Recall the model assumed to describe the population

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

and our model of $\mathbb{E}[Y|x]$, the regression line described by

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

We initially assumed that ϵ was a RV with constant variance σ^2.

However, even this assumption was unnecessary and could have been replaced with a more general description of ϵ as representing the component(s) not captured by $\beta_0 + \beta_1 x$.

The critical assumption we made later was that the error terms

$$\epsilon_i \sim \mathrm{N}(0, \sigma^2) \tag{7}$$

and **independent**.

3. Checking model assumptions

By the properties of the normal distribution this can be re-stated as

$$\epsilon_i \sim \mathrm{N}(0, \sigma^2) \tag{8}$$

with

$$\rho_{\epsilon_i, \epsilon_j} \equiv \mathrm{corr}(\epsilon_i, \epsilon_j) = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases} \tag{9}$$

for each $i, j \in \{1, \ldots, n\}$, where Pearson's correlation coefficient is described in Lecture 1.

N.B. The property used is zero correlation implies independence for normally-distributed RVs.

As estimates of the errors ϵ_i, the residuals $\hat{\epsilon}_i$ should mimic their behaviour – we check the assumptions on ϵ_i via $\hat{\epsilon}_i$.

3. 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 and charts
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.

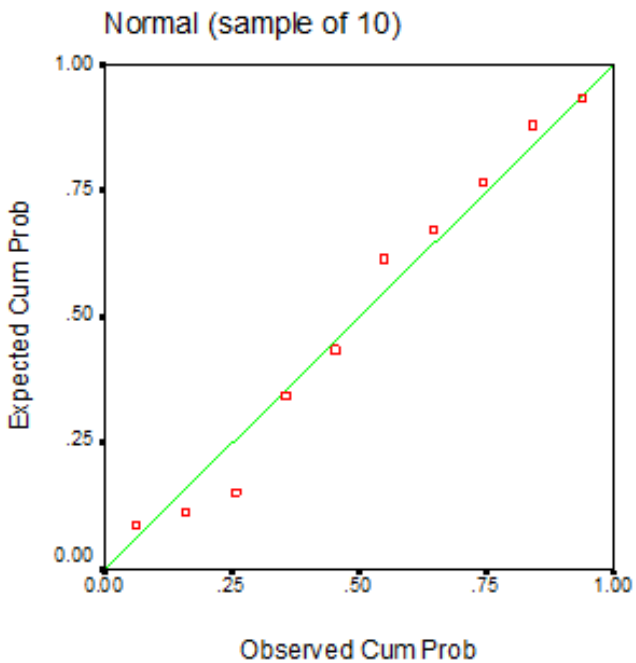
We know the residuals should be normally distributed with zero mean, so when we look at plots of residuals we want to see them distributed symmetrically either side of zero, with more closer to zero than further away etc.

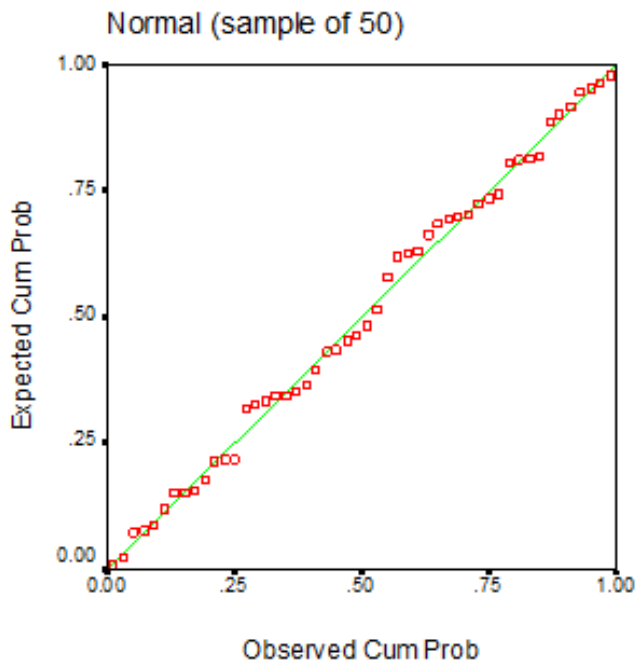
We also know that they should be uncorrelated, so there should be no discernable pattern when plotted against the independent variable or the regression prediction of the dependent variable.

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

3. Checking model assumptions – residual analysis

Below are two scatter plots showing **theoretical cumulative probability** against **observed cumulative probability** for different sized samples of normally distributed (psuedo) RVs generated by algorithm.

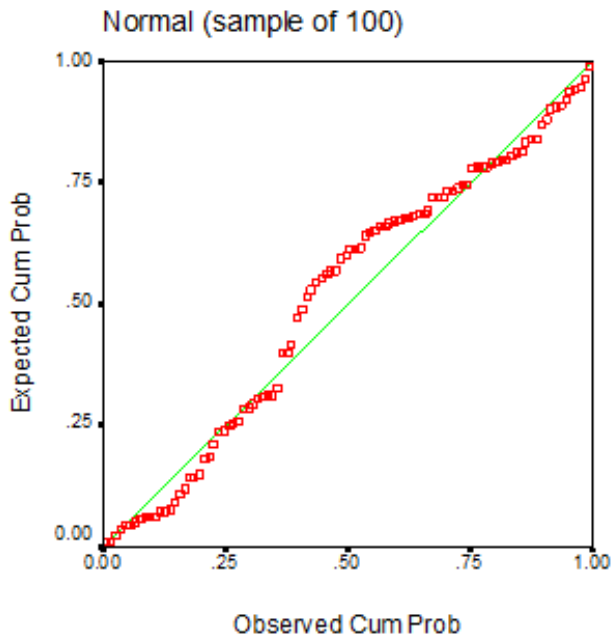


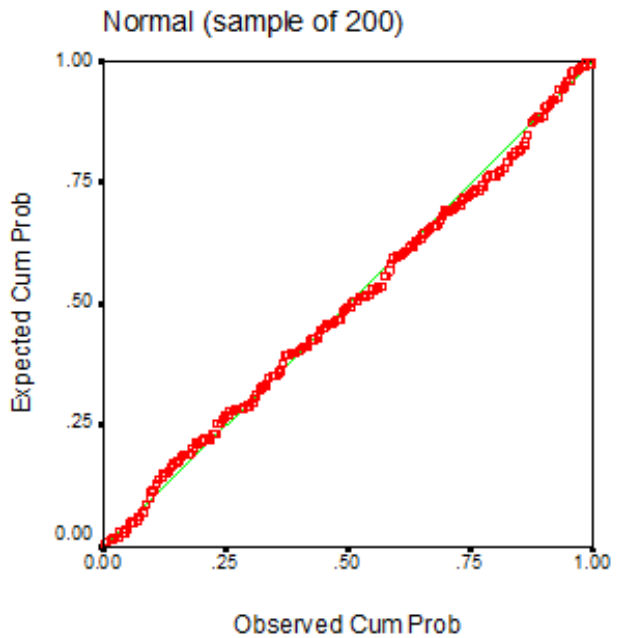


Deviation from theoretical behaviour is indicated by departure from the green straight line.

3. Checking model assumptions – residual analysis

Here is the continuation of this experiment for larger sample sizes.



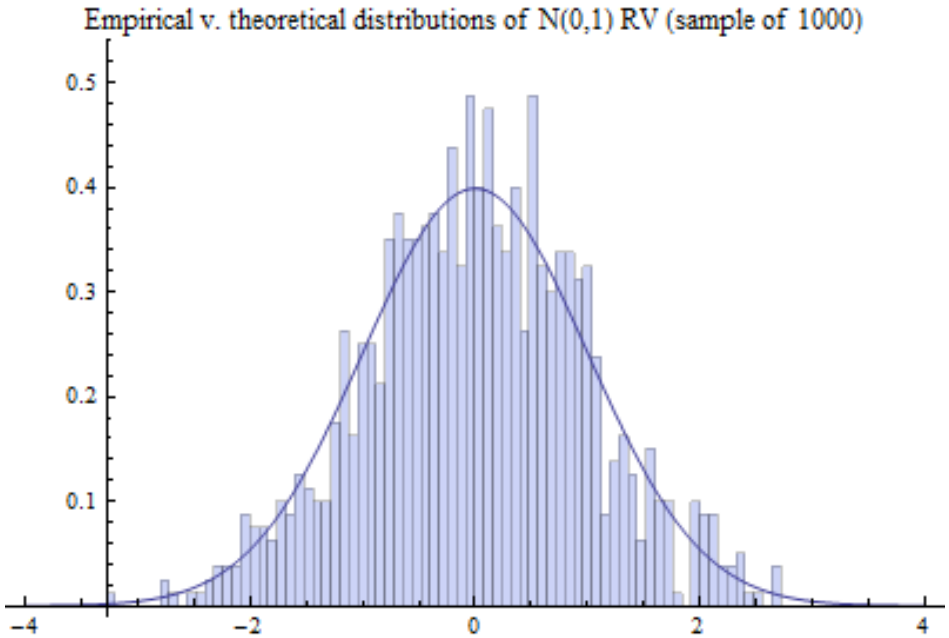


We see observed behaviour approaching theoretical behaviour as the sample size increases.

We can perform similar analyses using R in the context of regression.

3. Checking model assumptions – residual analysis

Another version of this sort of analysis we have seen in previously.

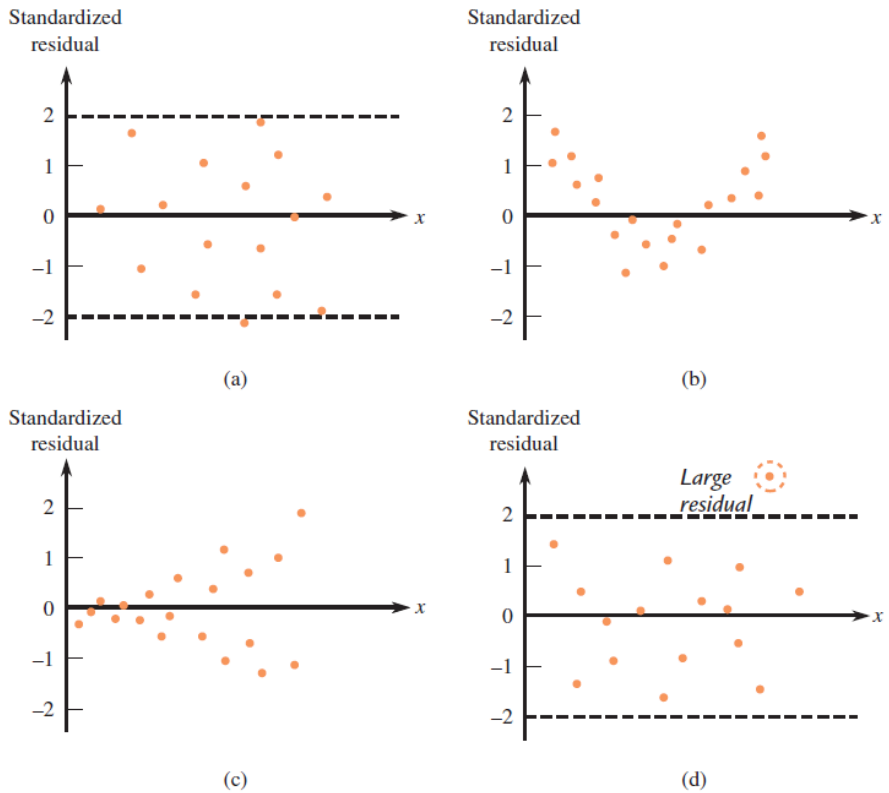


We are looking for the empirical distribution, here displayed as a histogram, to closely match the theoretical PDF.

We apply these sort of tools to regression using R.

3. Checking model assumptions – residual analysis

Below are some typical plots of residuals against dependent variable.



Example of standardised residual plots. Source Peck et al. (2012) page 769

Plot (a) is fine, (b) and (c) show patterns and (d) a large residual.

3. Checking model assumptions – residual analysis

The measure of leverage (6) can be used to define quantities useful for **residual analysis**, such as the **internally-Studentised residual**

$$\hat{t}_i = \frac{\hat{\epsilon}_i}{S\sqrt{1-h_{ii}}} \tag{10}$$

and **externally-Studentised residual**

$$\hat{d}_i = \frac{\hat{\epsilon}_i}{S^{(i)}\sqrt{1-h_{ii}}}, \tag{11}$$

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

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

3. Checking model assumptions – residual analysis

Finally we mention a test for **serial correlation**, the Durbin-Watson test.

This statistic is calculated as

$$dw = \frac{\sum_{i=2}^{n}(\hat{\epsilon}_i - \hat{\epsilon}_{i-1})^2}{\sum_{i=1}^{n}\hat{\epsilon}_i^2}. \tag{12}$$

The statistic satisfies

$$0 \leq dw \leq 4$$

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

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

3. R example continued

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

```
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 variation in the data about the mean.

```
> durbinWatsonTest(mod1)
 lag Autocorrelation D-W Statistic p-value
   1      -0.1620183       2.323897   0.026
 Alternative hypothesis: rho != 0
```

We also see a Durbin-Watson statistic of $dw = 2.324$ which, although not excessive, does point to weak negative serial correlation in the residuals $\hat{\epsilon}_i$.

3. R example continued

Below is the ANOVA table.

```
> summary(aov(mod1))
             Df Sum Sq Mean Sq F value Pr(>F)
LogGNI        1   9301    9301   330.4 <2e-16 ***
Residuals   185   5208      28
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
```

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-statistic of 330.4 and associated p-value is extremely small, we are happy to reject the null hypothesis that the true value of β_1 is zero.

3. R example continued

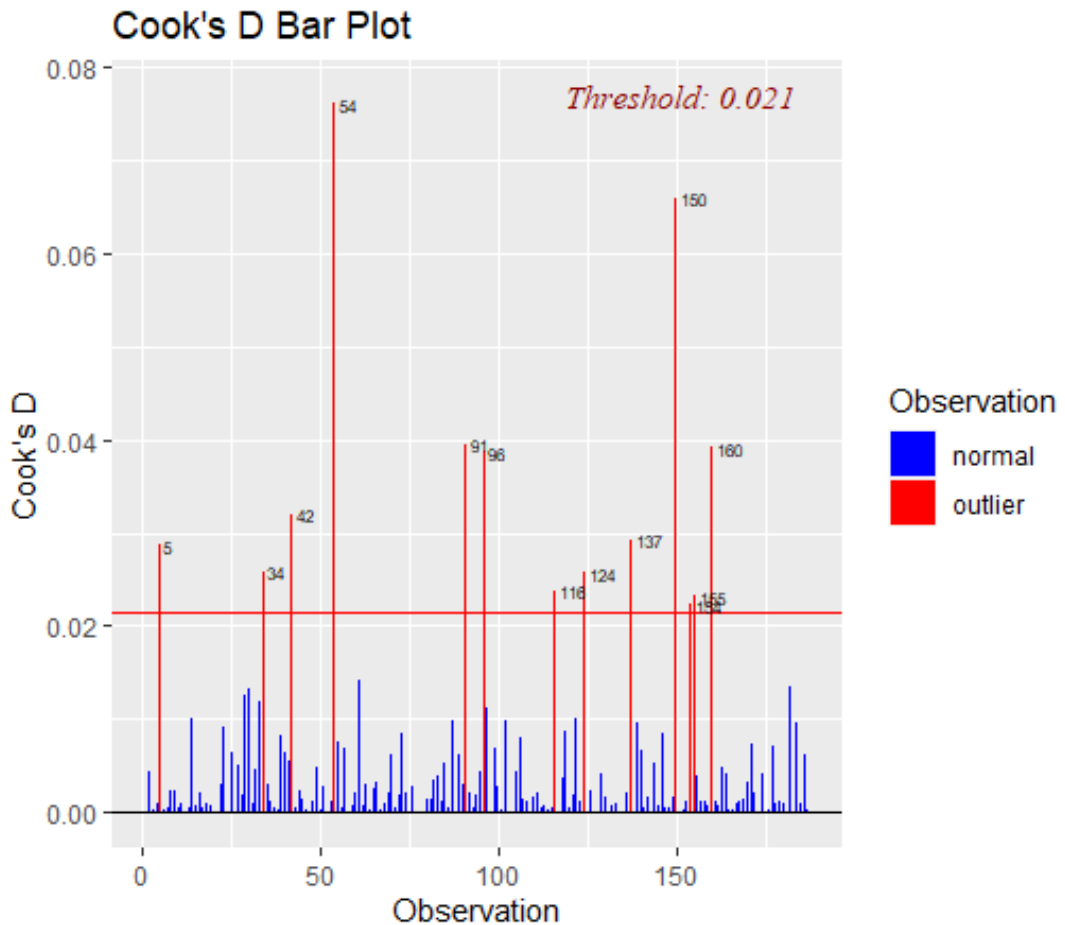
```
> summary(predict(mod1))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  52.93   65.49   71.37   70.58   75.78   85.63
> summary(mod1$resid)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -22.647  -2.267   1.020   0.000   3.354   8.938
> round(summary(rstandard(mod1)) ,3)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -4.286  -0.430   0.194   0.000   0.635   1.690
> cooksD<-cooks.distance(mod1)
> round(summary(cooksD), 3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   0.001   0.005   0.005   0.076
```

The maximum Cook's distance is 0.076. The critical value for our case is $\frac{4}{n-m-1} = \frac{4}{187-2} \approx 0.0216$ indicating at least one point which may have some degree of excessive influence.

3. R example continued

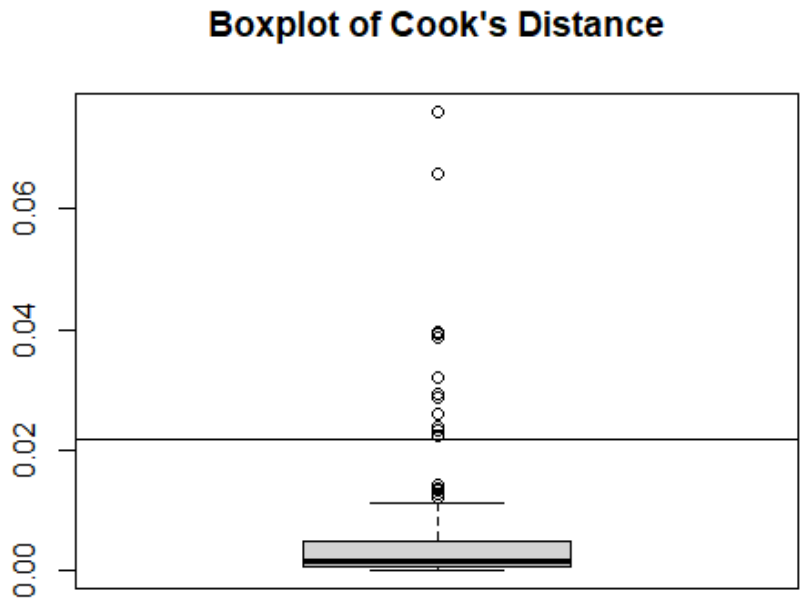
We can investigate this further with plots of Cook's distance D_i.

```
> library('olsrr')
> ols_plot_cooksd_bar(mod1) # this function uses 4/n as the threshold!
```



3. R example continued

We can also look at the boxplot.



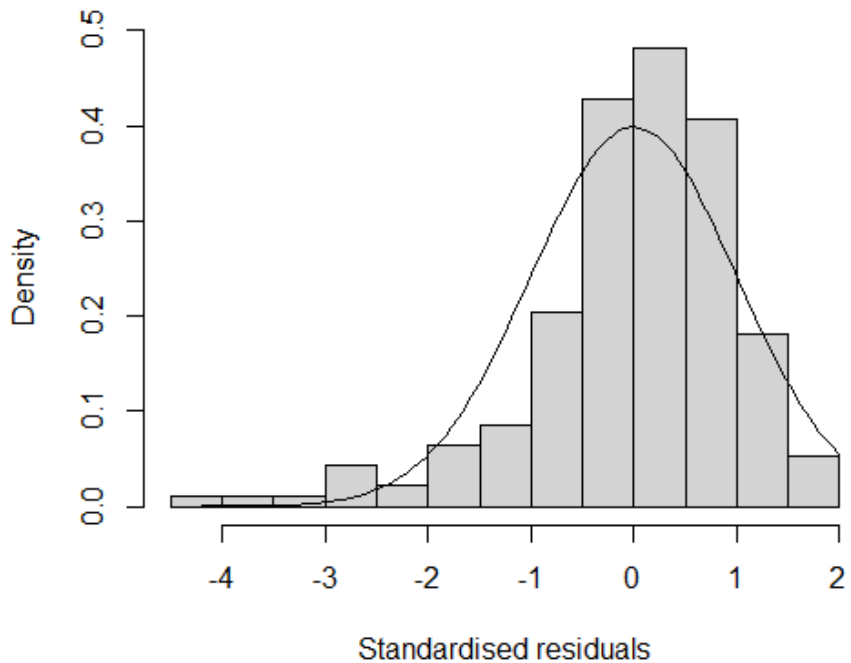
The box plots identify plenty of outliers, not all of which are in excess of the critical values 0.0216 for D_i .

Still, we could exclude the main offenders, re-perform the regression and see if the model improves.

3. R example continued

Now we look at plots of the standardised residuals.

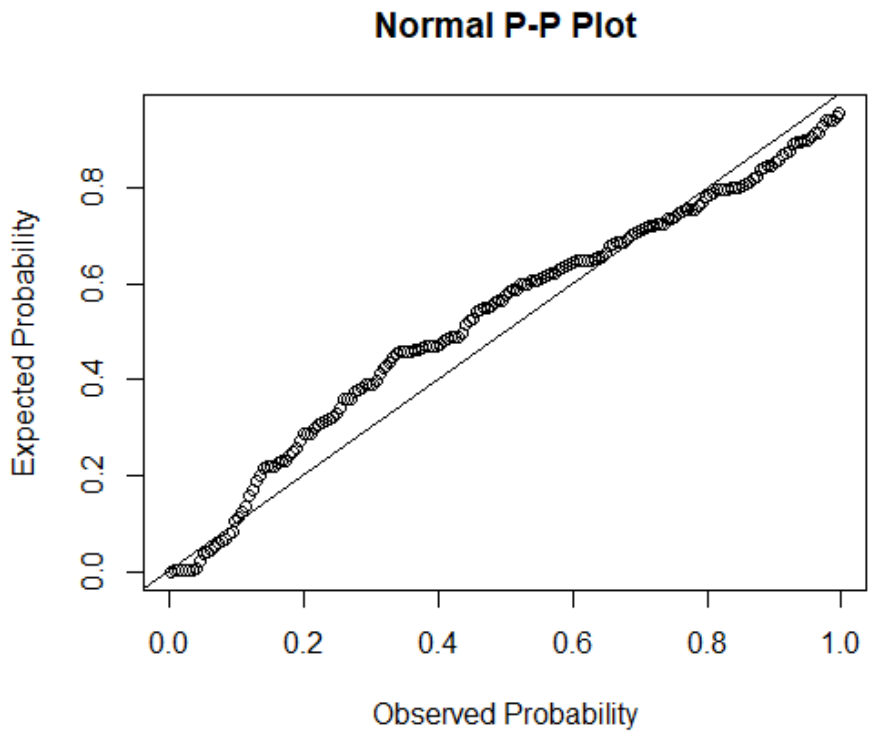
```
> hist(mod1.st.resid, xlab = "Standardised residuals", freq
= F, main = "")
> curve(dnorm, add = T)
```



The histogram is not symmetrical and shows clear negative skewness, a property not possessed by the standard normal distribution N(0,1).

3. R example continued

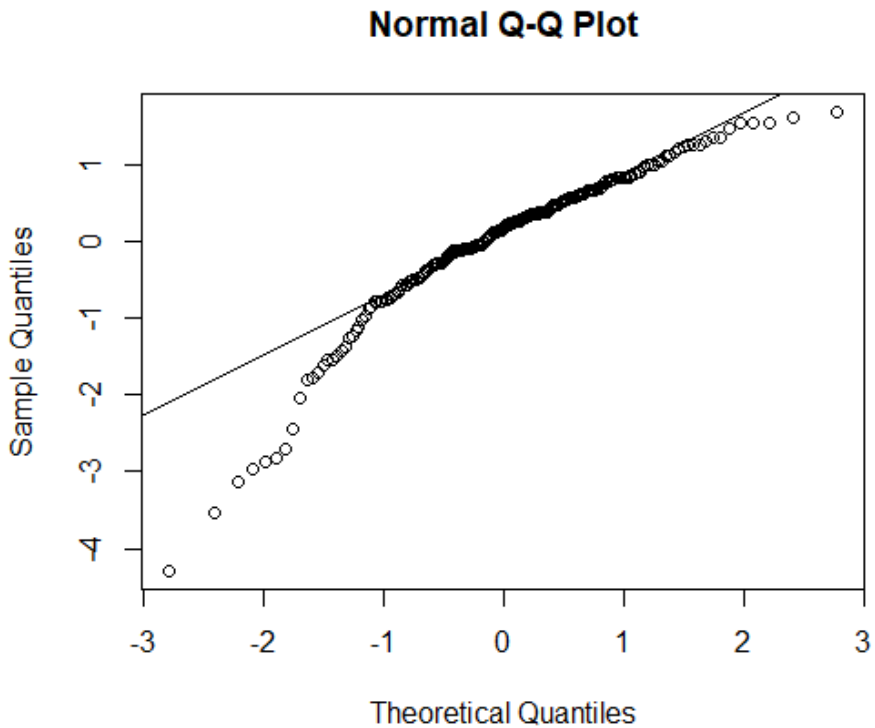
```
> probDist <- pnorm(mod1.st.resid)
> plot(ppoints(length(mod1.st.resid)), sort(probDist), main
= "Normal P-P Plot", xlab = "Observed Probability",
ylab = "Expected Probability")
> abline(0,1)
```



The PP Plot also indicates departure from normality.

3. R example continued

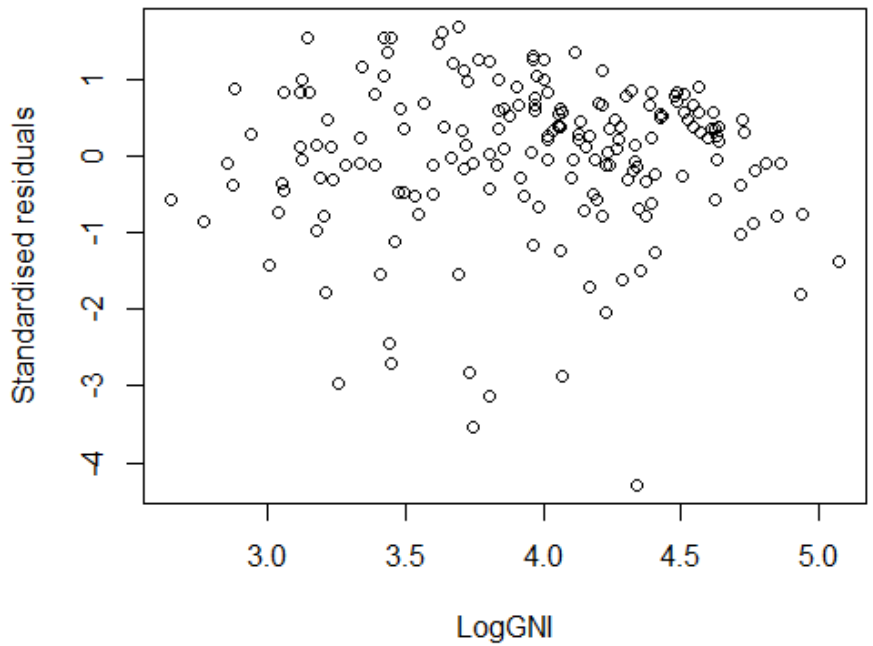
```
> qqnorm(mod1.st.resid)
> qqline(mod1.st.resid)
```



The QQ Plot also indicates departure from normality.

3. R example continued

Finally we look at the standardised residuals, plotted against the dependent variable *LogGNI*.



There is a hint of a pattern indicating possible serial correlation. Large negative observations below -3 is also cause for concern.

All in all, our analysis indicates departure from model assumptions.

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.