Diagnnostics and Remedial Measures

AU STAT-615

Emil Hvitfeldt

2021-02-17

Residuals

Diagnostics for predictor variables will be covered in Chapter 10

Diagnostics for the response variable are usually carried out indirectly through an examination of the residuals

$$e_i = Y_i - \hat{Y}_i$$

For the unknown true error $\varepsilon_i = Y_i - E\{Y_i\}$ we know that $E\{\varepsilon_i\} = 0$ and $V\{\varepsilon_i\} = \sigma^2$

Idea

If the fitted model is appropriate for the data at hand, the observed values e_i should reflect the properties assumed for ε_i

Properties of Residuals

Mean

The mean is given by

$$\bar{e} = \frac{\sum e_i}{n} = 0$$

Since this is always true, it provides no information as to whether the true errors ε_i have expected value $E\{\varepsilon_i\} = 0$ 😥

Properties of Residuals

Variance

The variance is given by

$$s^2 = \frac{\sum (e_i - \bar{e})^2}{n-2} = \frac{\sum e_i^2}{n-2} = \frac{SSE}{n-2} = MSE$$

It can be showed that $E\{MSE\} = \sigma^2$ is an unbiased estimator of the variance of the error terms σ^2

So if the model is appropriate, MSE is, an unbiased estimator of the variance of the error terms σ^2

Non-independence

The residuals ε_i are not independent random variables because they involve the fitted values $\hat{Y}_i$ which are based on the same fitted regression function

Departures from model to be studied by residuals

- The regression function is not linear
- The error terms do not have constant variance
- The error terms are not independent
- The model fits all but one or a few outlier observations
- The error terms are not normally distributed
- One or several important predictor variables have been omitted from the model

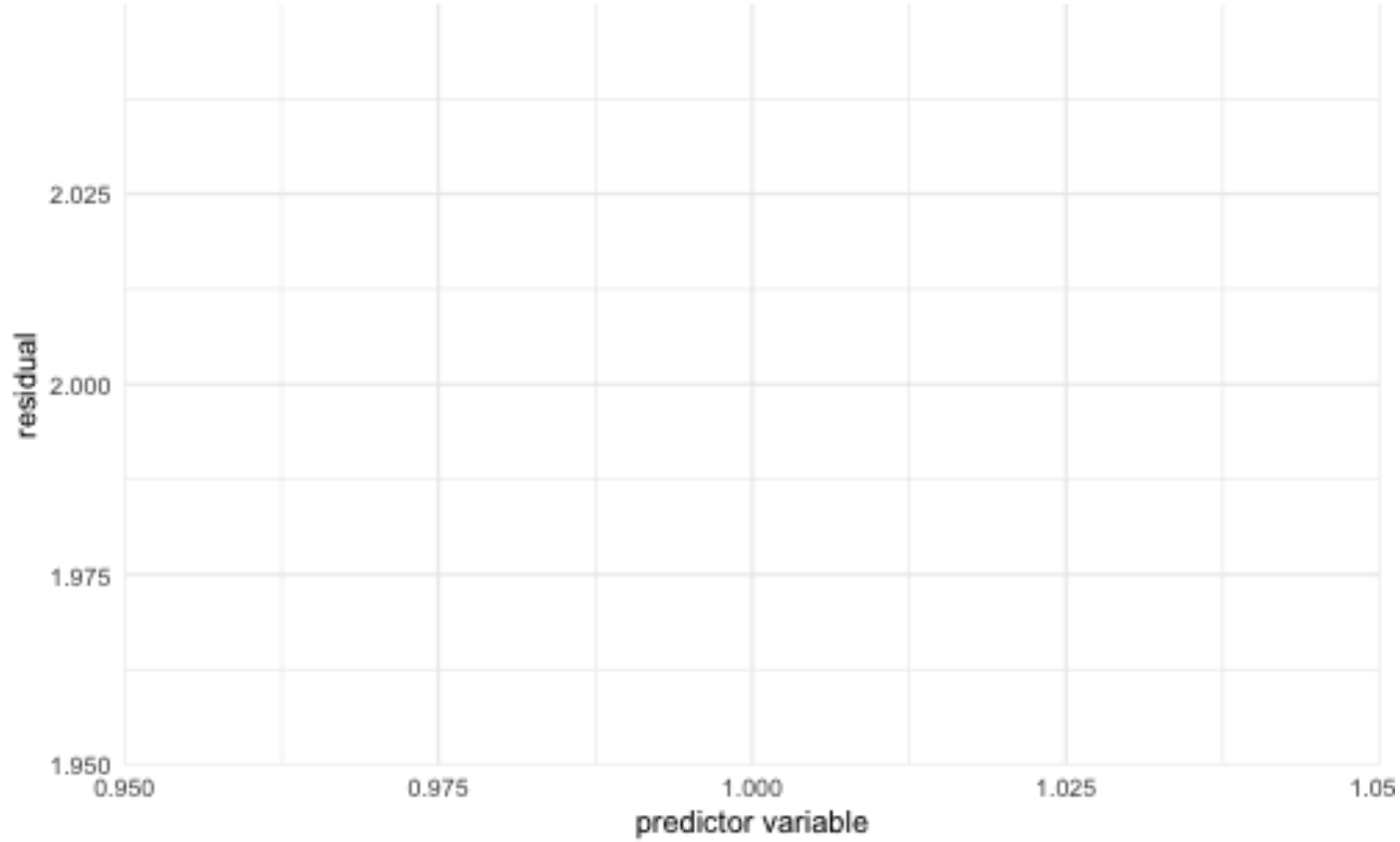
Departures from model to be studied by residuals

- The regression function is not linear
- The error terms do not have constant variance
- The error terms are not independent
- The model fits all but one or a few outlier observations
- The error terms are not normally distributed
- One or several important predictor variables have been omitted from the model

Linearity, Independence, Normality, Equal Variance

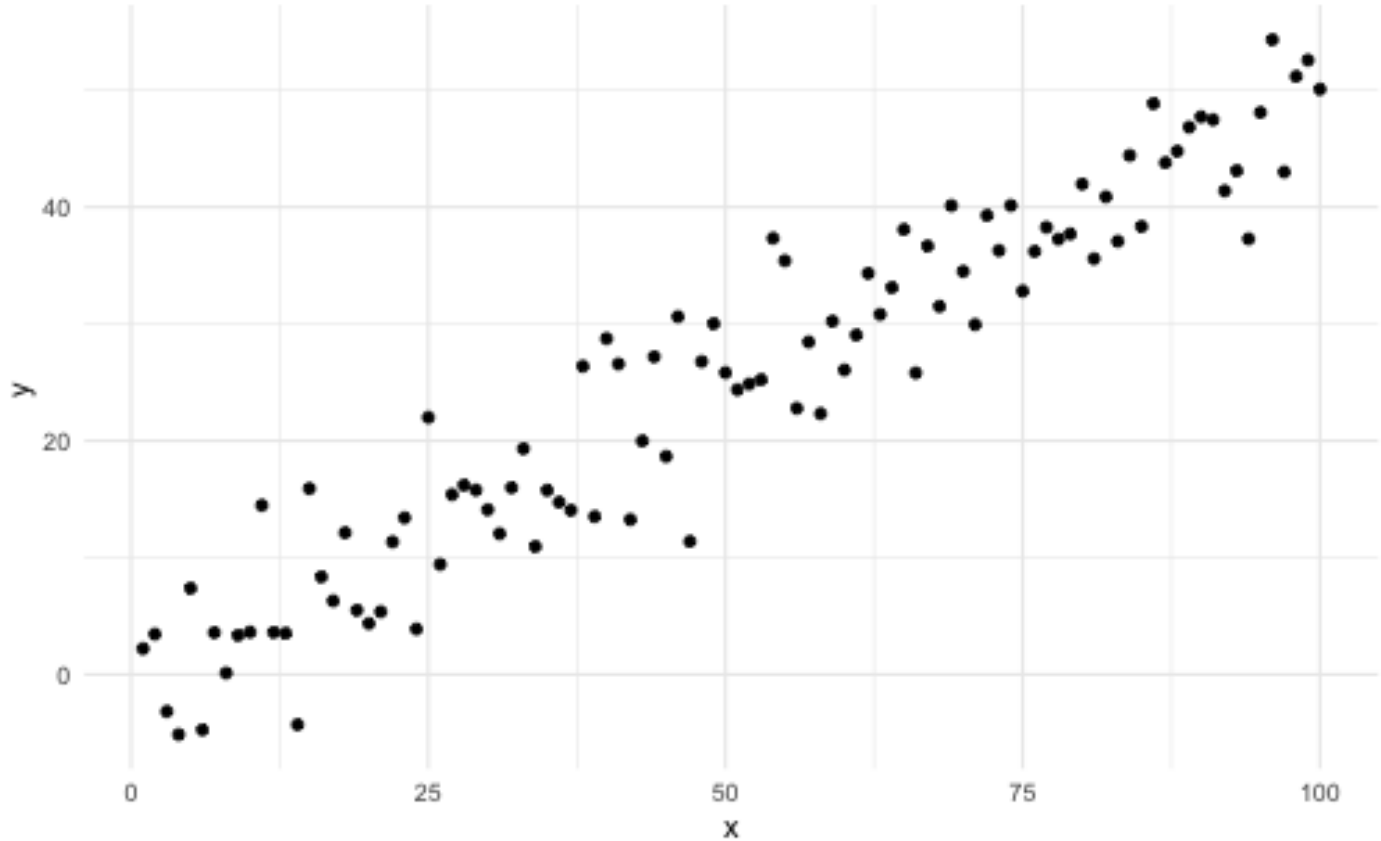
Non-linearity of Regression Function

The residual plot against the predictor variable



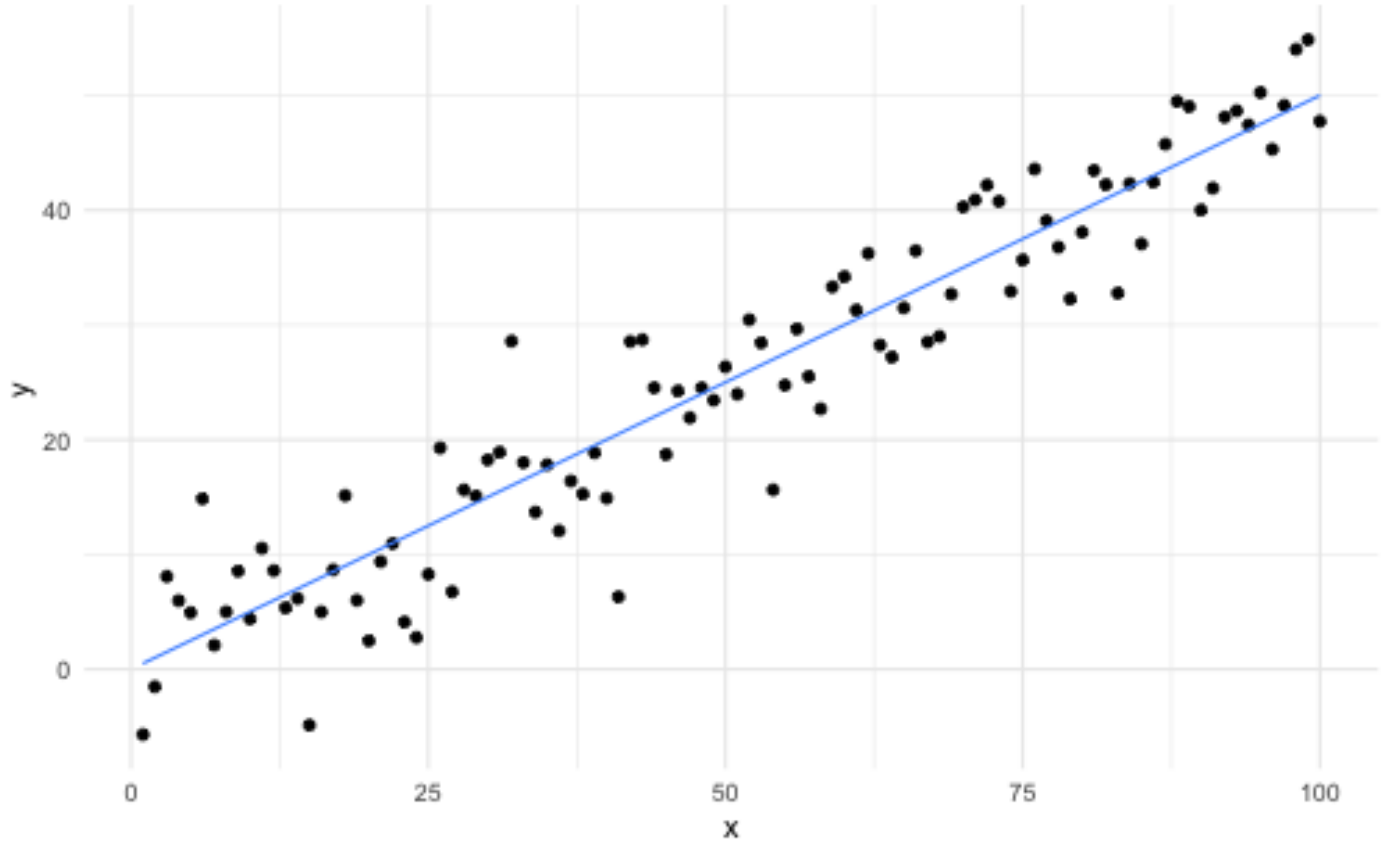
Non-linearity of Regression Function

Consider we have the following data



Non-linearity of Regression Function

the fitted regression line goes here



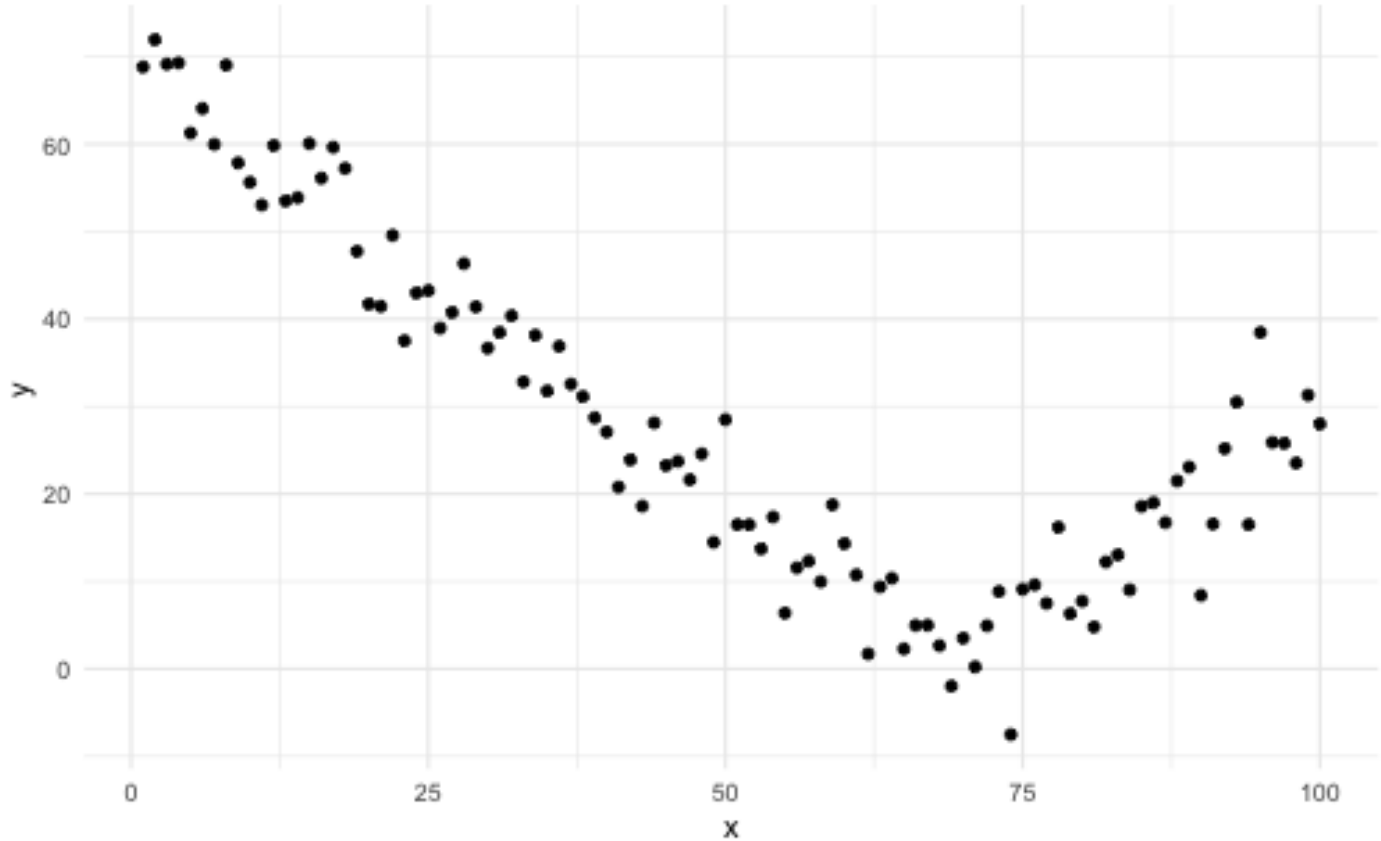
Non-linearity of Regression Function

And the residuals against the predictor variable looks like this

Perfectly linear

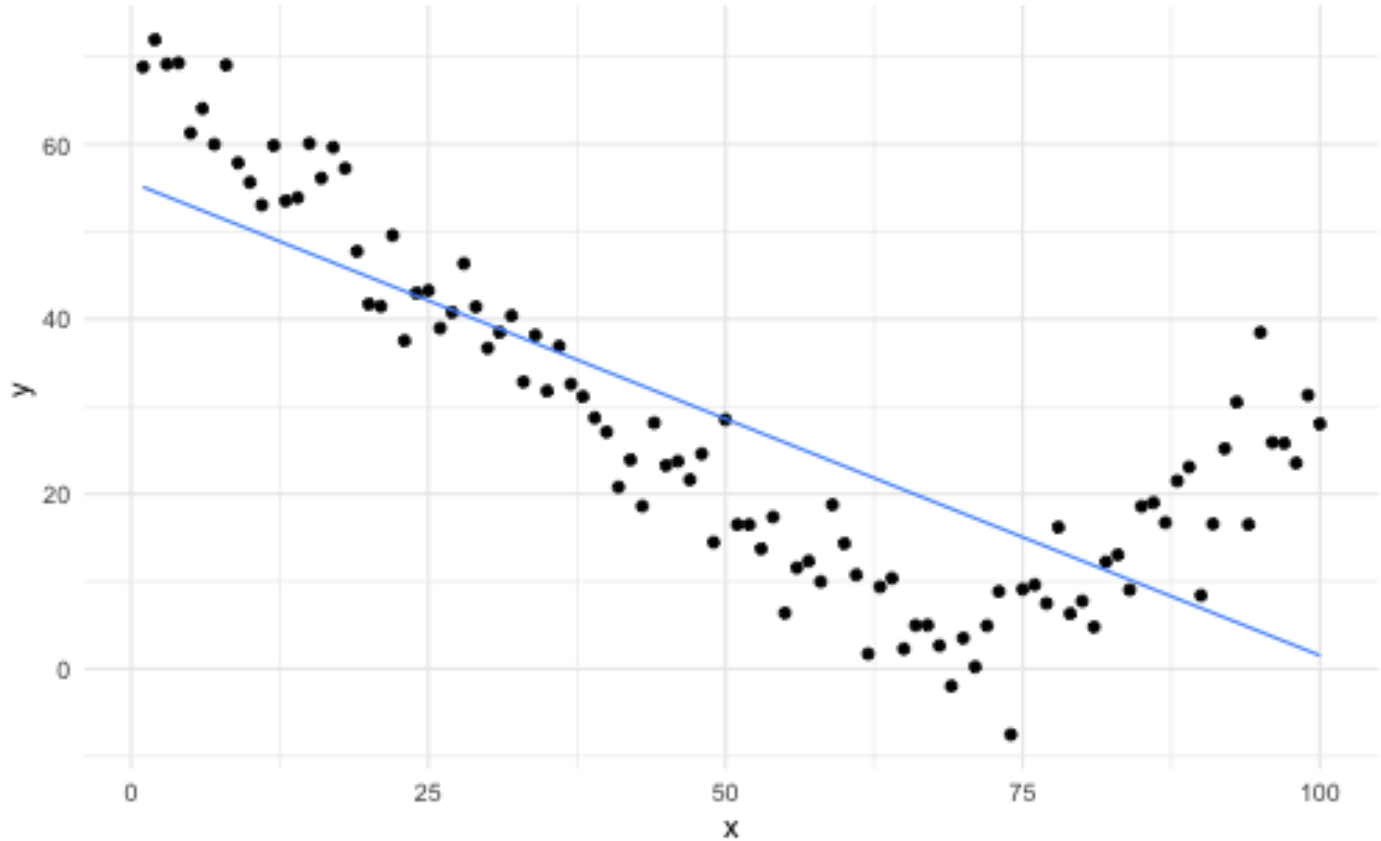
Non-linearity of Regression Function

Consider we have the new following data



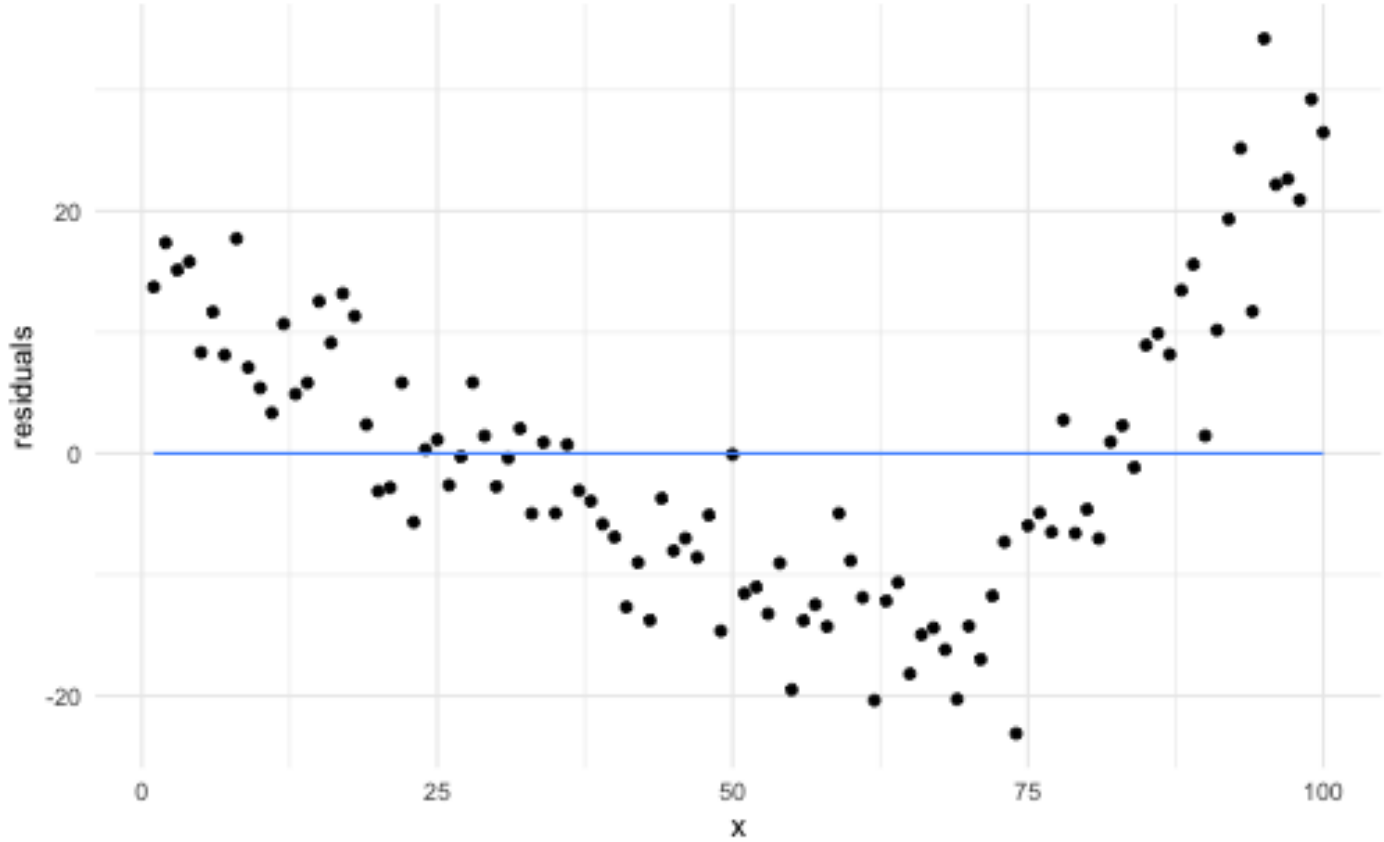
Non-linearity of Regression Function

the fitted line would be



Non-linearity of Regression Function

And the residuals will look like this. Clearly not linear



Non-Constancy of the Error Variance

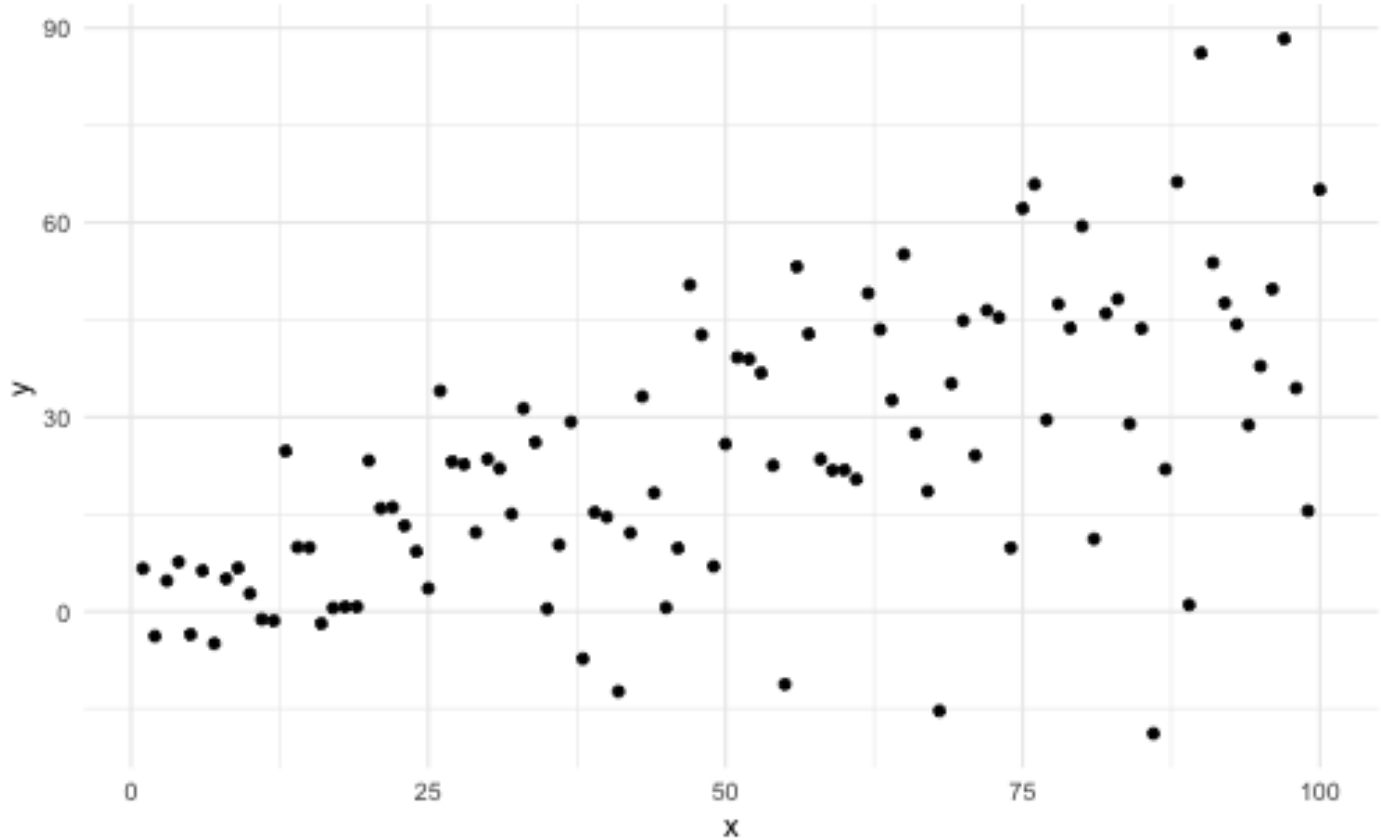
Using residual plots will also enable us to examine whether the variance of the error terms is constant

Non-Constancy of the Error Variance

If we look at this residual plot from earlier. We notice that the spread (variance) of the residuals stay constant throughout the values of X

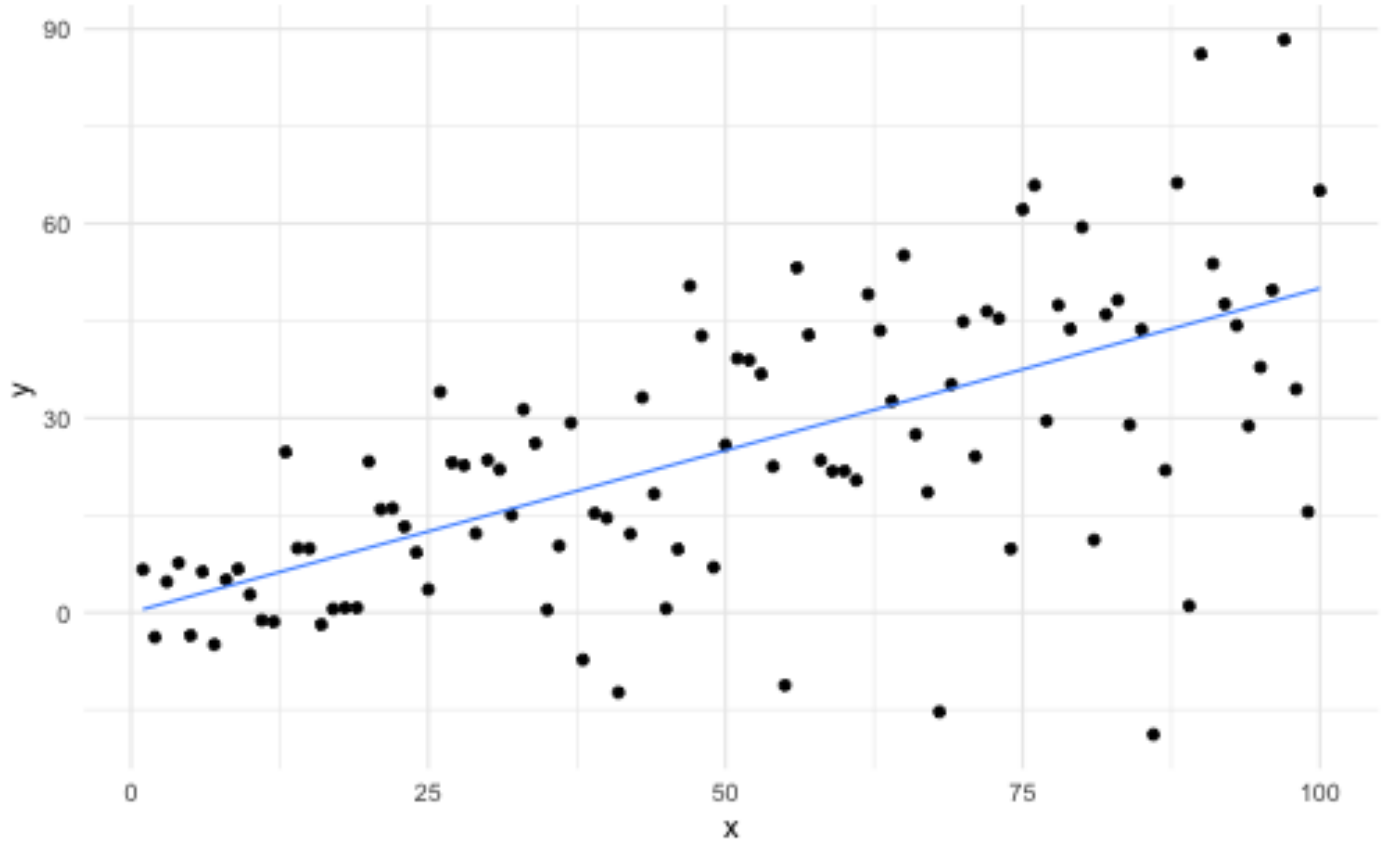
Non-Constancy of the Error Variance

If we take this data right here



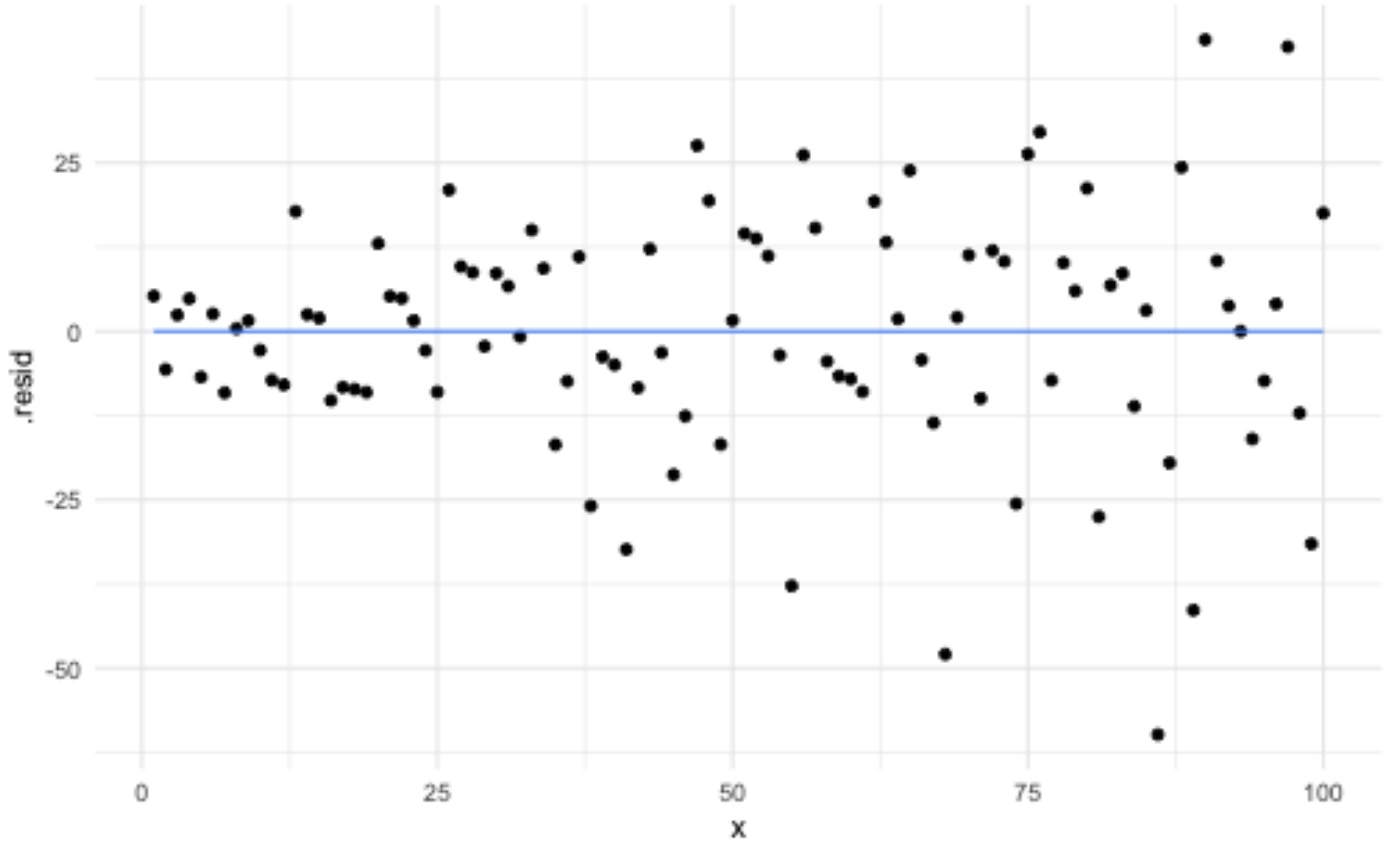
Non-Constancy of the Error Variance

And fit a linear regression line through it



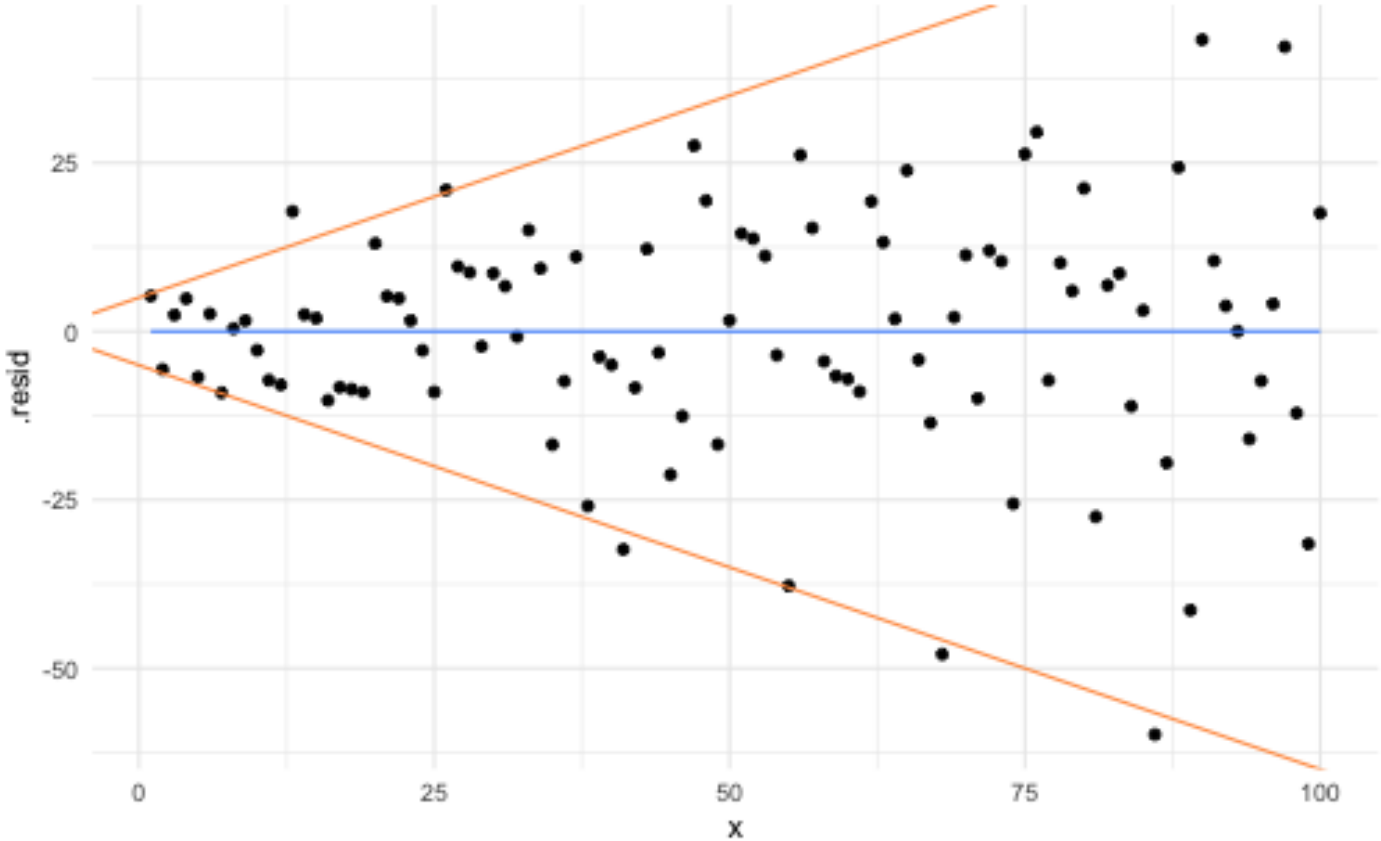
Non-Constancy of the Error Variance

We can take a look at the residuals



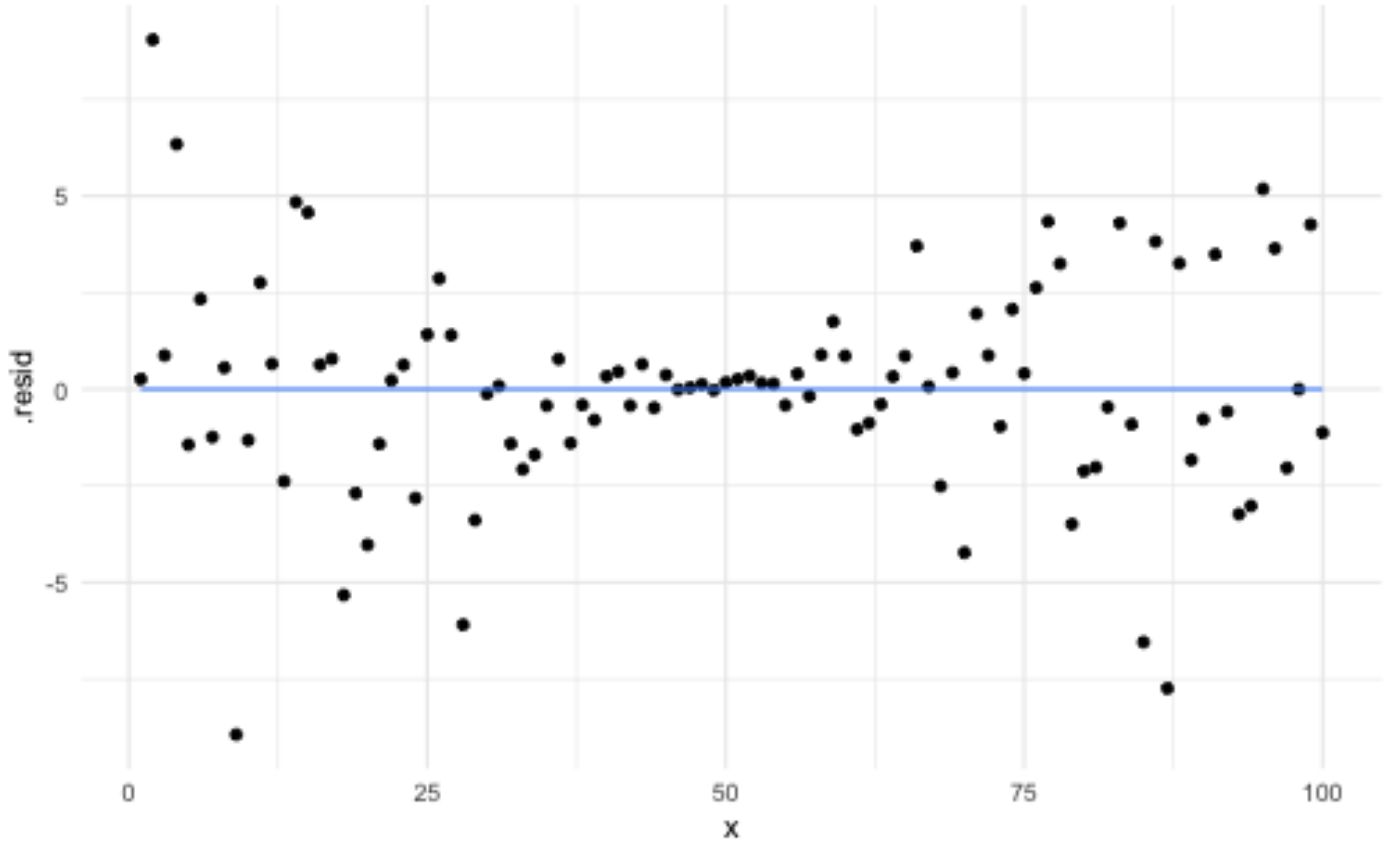
Non-Constancy of the Error Variance

And we notice the spread of the residuals widens as the values get larger



Non-Constancy of the Error Variance

There is no set way the spread of the residuals can be uneven



Presence of outliers

We will discuss this more in Chapter 10

A general strategy is to normalize the error by the square root of the MSE

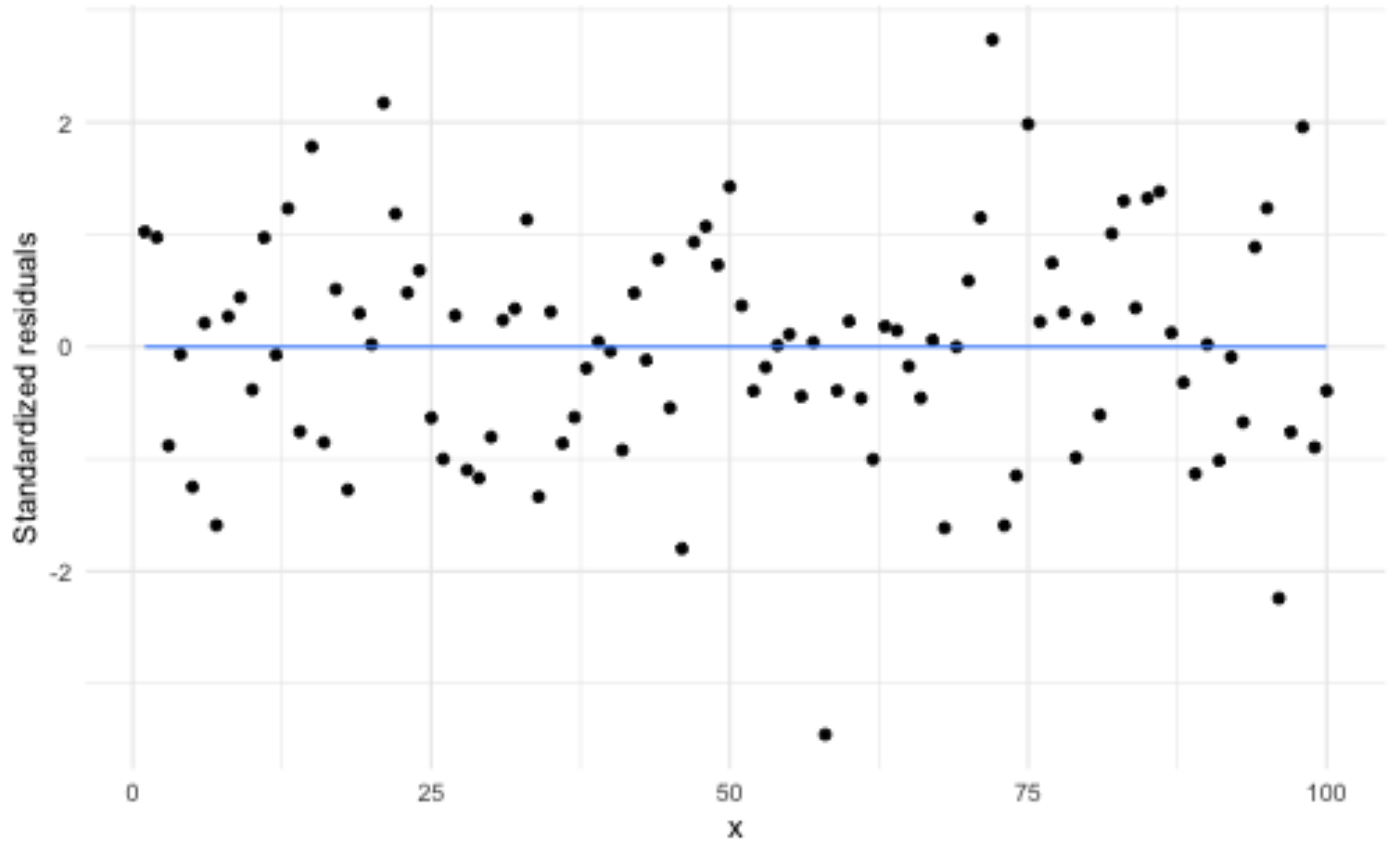
$$\text{Standardized residuals} = \frac{e_i}{\sqrt{MSE}}$$

Presence of outliers

If we plot the standardized residuals against the predictor we see that most of the residuals for this

Presence of outliers

Here it appears that some of the points could be considered outliers



Presence of outliers

Note:

Outliers can affect performance because the sum of the squared deviations minimized

Presence of outliers

You cannot willy-nilly remove outliers

Each outlier removed will by definition improve the fit of the model, but only on the reduced data set

The outliers may carry important information regarding some interactions between predictor variables, or that some important predictors have been excluded from the model

Presence of outliers

A safe rule frequently suggested is to discard an outlier only if there is direct evidence that it represents an error in recording, a miscalculation, etc

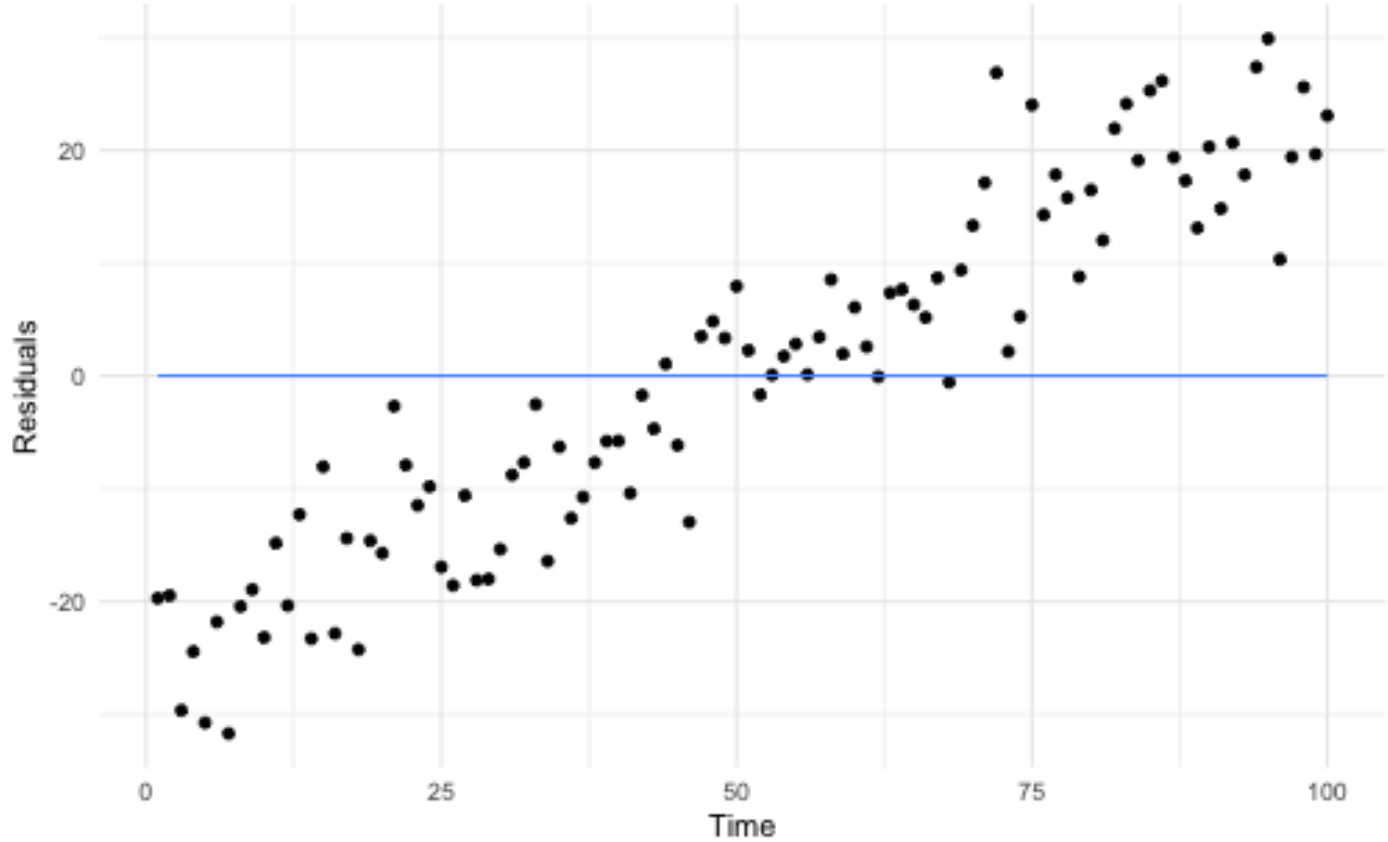
Non-independence of Error Terms

Whenever data are obtained in a time sequence or another type of sequences is it good to prepare a sequence plot of the residuals

The purpose of this is to see whether there is a correlation between the error terms that are near each other in sequence

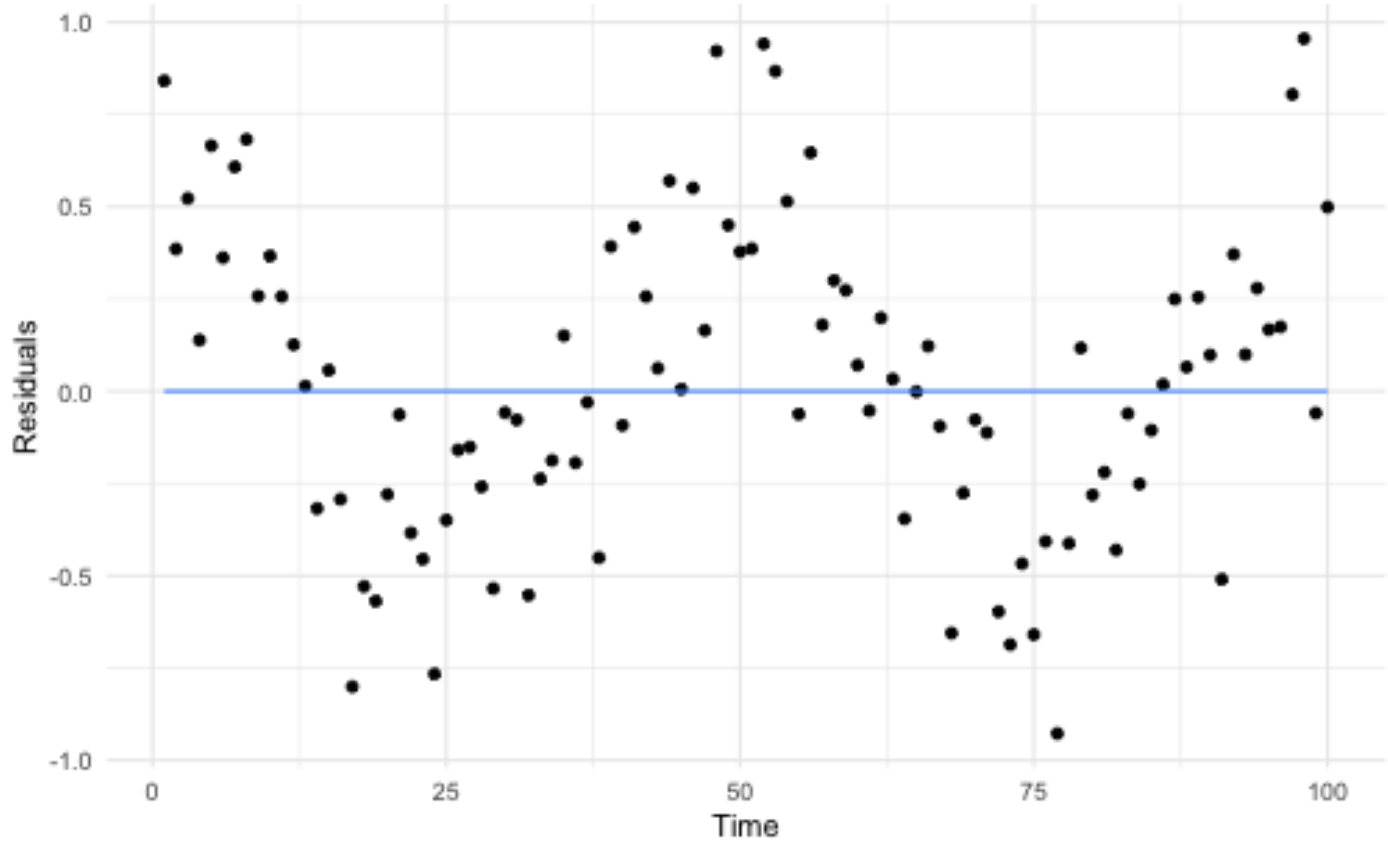
Non-independence of Error Terms

It is important that the errors don't follow a trend along the time axis



Non-independence of Error Terms

For time variables it can be common to see a cyclical non-independence



Non-independence of Error Terms

In general, you can see quite big fluctuations in the data when we talk about time-dependent data

Non-Normality Of Error Terms

The normality of the error terms can be studied by general graphical methods for 1-dimensional data. Histograms are a good first choice

Non-Normality Of Error Terms

Another way is to use Q-Q plots (normal quantile-quantile plots)

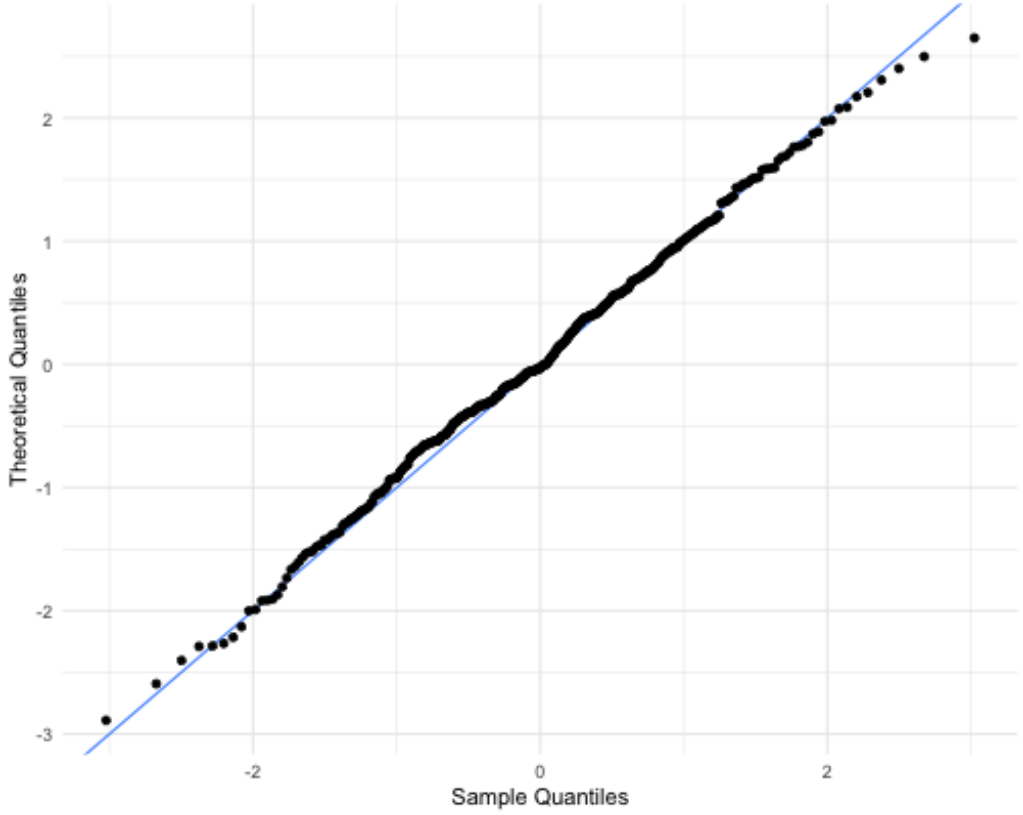
For this type of chart, we plot the sample quantile against the theoretical quantiles

Under the assumption of normal distribution

Non-Normality Of Error Terms

A normal Q-Q plot should look something like this

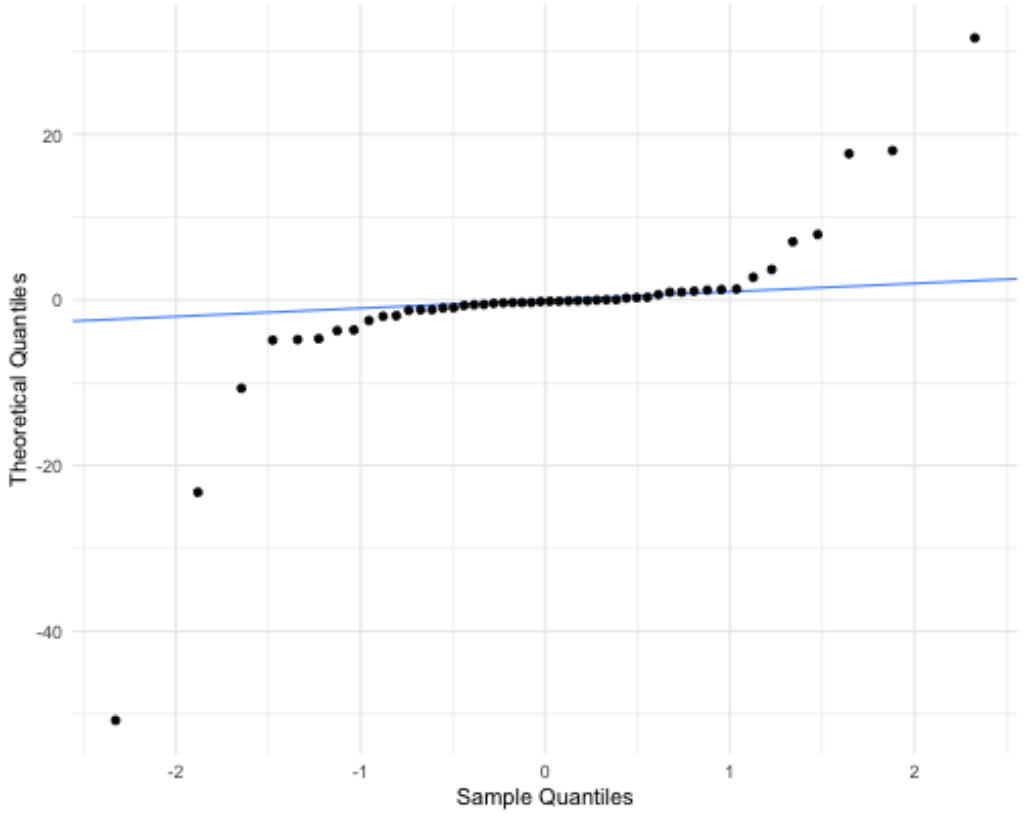
Follow the diagonal fairly close with a couple of deviations at the ends



Non-Normality Of Error Terms

Symptom: ends go under on the left and over on the right

Cause: Symmetrical heavy-tailed



Non-Normality Of Error Terms

Symptom: curved

Cause: skewed right