

# Practical Report 2

## Acceptable residual plots

In this practical we will explore simulated data to illustrate residual diagnostics which are part of the linear modelling process. Residual plots allows us to judge the appropriateness of the assumption of linearity and homoscedasticity.

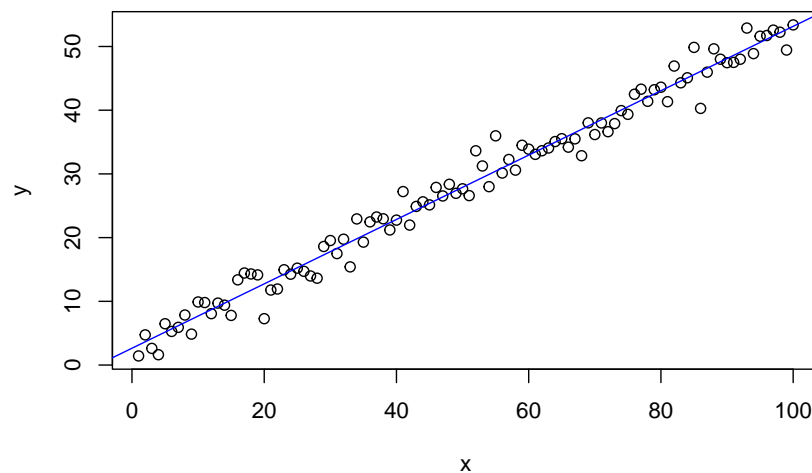
### Section 1, Question 1

Linear model:

$$Y_j = 3 + \frac{1}{2}x_j + \epsilon_j, \quad j = 1, \dots, 100,$$

where  $\epsilon_1, \dots, \epsilon_{100}$  are iid  $N(0, 4)$ . (Remark: `rnorm` takes standard deviation as the argument, see in `?rnorm`.)

### Section 1, Question 2

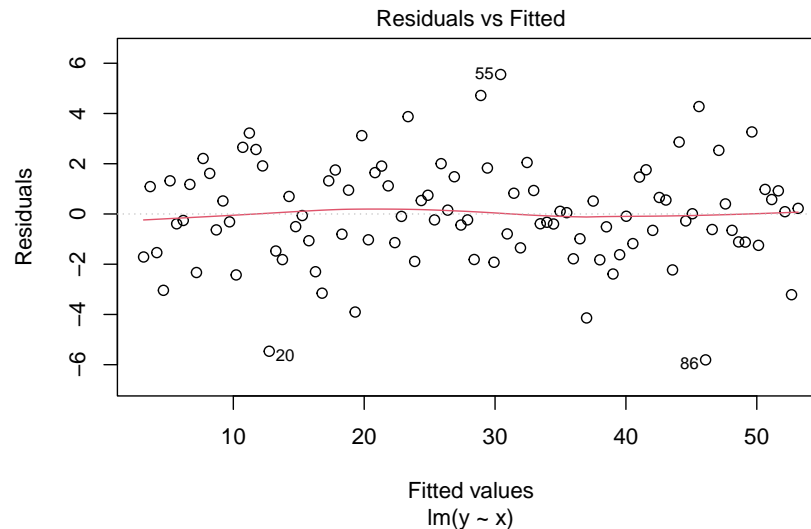


```
## (Intercept)          x
## 2.6376093 0.5052425
```

We can see that the estimated intercept and slope are reasonably close to the true parameter, 3 and 0.5. We can double-check this by the 95% confidence intervals.

```
##          2.5 %    97.5 %
## (Intercept) 1.8305475 3.4446711
## x          0.4913678 0.5191172
```

## Section 1, Question 3



The residuals scatter around the zero horizontal line and the variation of the residuals appears relatively constant. This implies no evidence of violations against the assumptions.

## Section 1 Question 4

Next, we sample Dataset 2 with only 20 data points.

Harder to tell if satisfying the model assumptions. We can see some curvature and uneven variance of the residuals even if the data is sampled from a linear model.

Question to think: how the confidence intervals change?

```
##          2.5 %    97.5 %
## (Intercept) 2.3082453 4.1203836
## x2         0.3753934 0.5266678
```

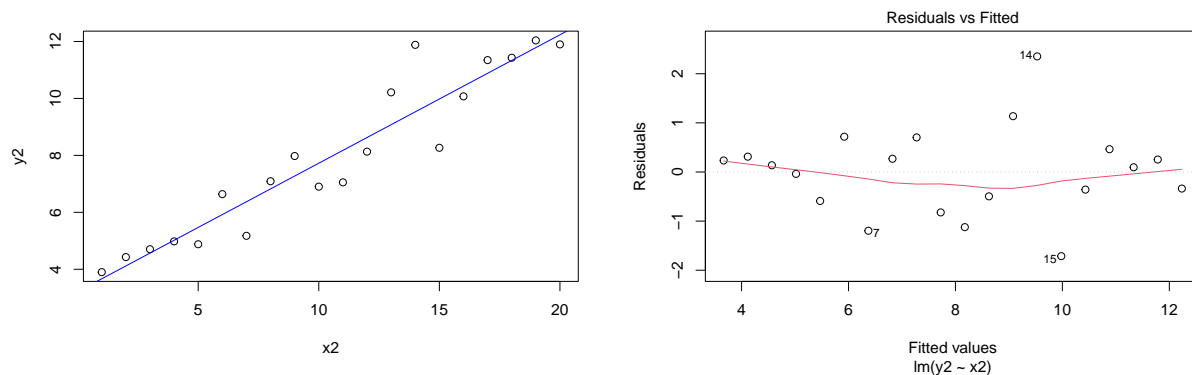


Figure 1: Left: Right:

```
## [1] "model2 intercept CI length: 1.81213825014775"
## [1] "model2 slop CI length: 0.151274378348051"
## [1] "model1 intercept CI length: 1.61412358684011"
## [1] "model1 slop CI length: 0.0277494061464172"
```

## Unacceptable residual plots

The figure below shows the scatterplot of Dataset 3, a dataset sampled from the model

$$Y_j = 10 + \frac{1}{2}x_j + \epsilon_j,$$

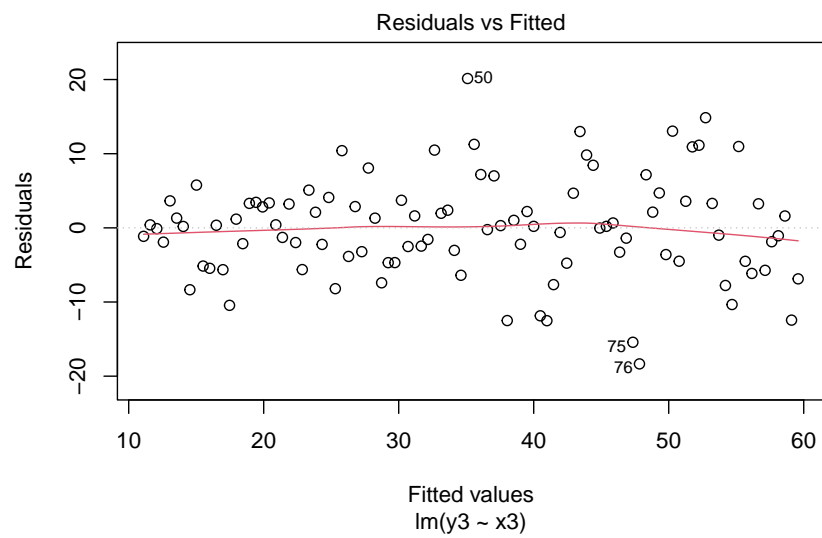
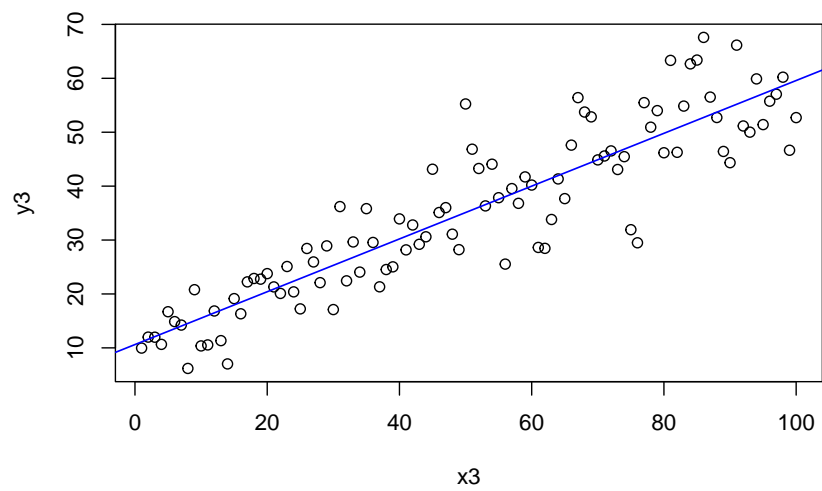
where  $x_j = j$  and  $\epsilon_j \sim N(0, x_j)$  for  $j = 1, \dots, 100$ . Furthermore, the errors  $\epsilon_1, \dots, \epsilon_{100}$  are independent. We fitted a simple linear regression model to the data and added the fitted line to the scatterplot.

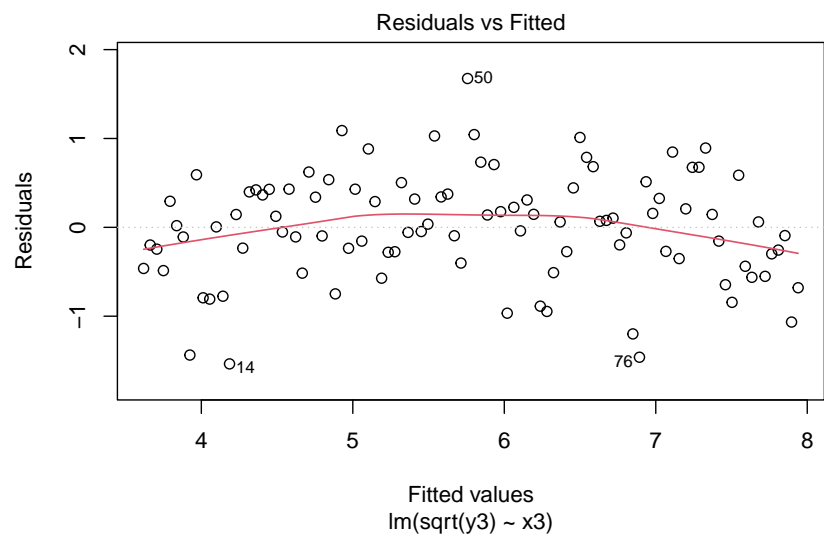
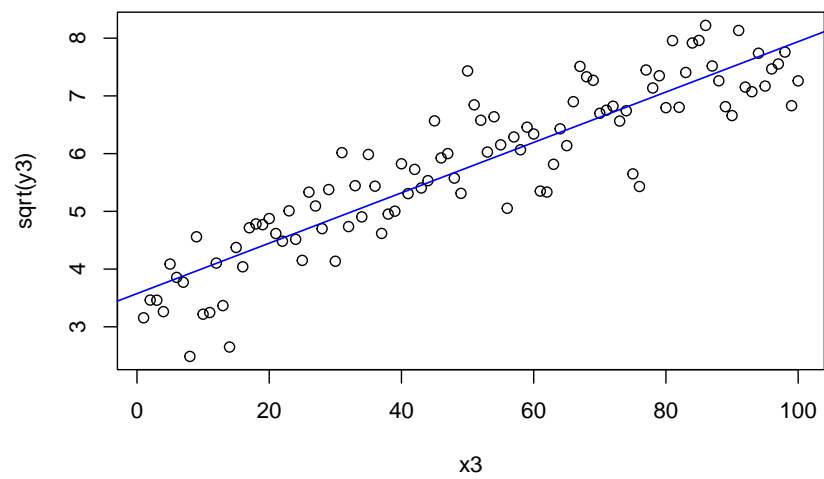
### Section 2 Question 1

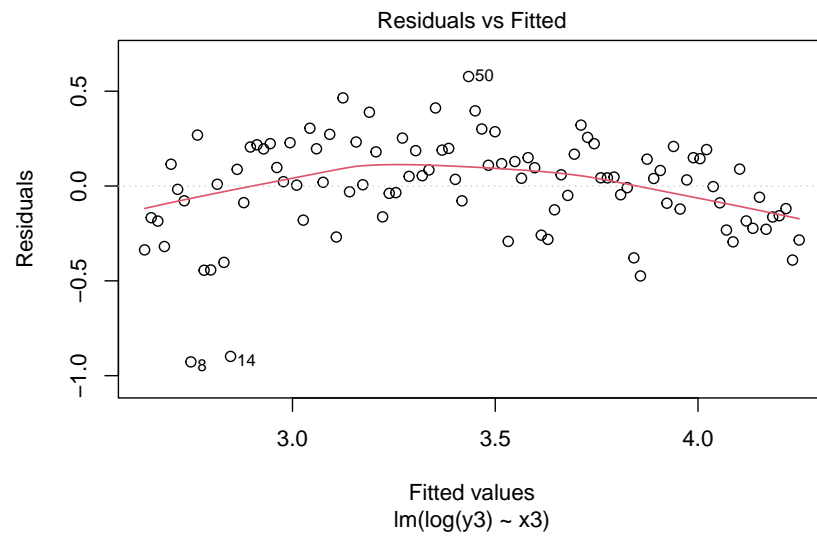
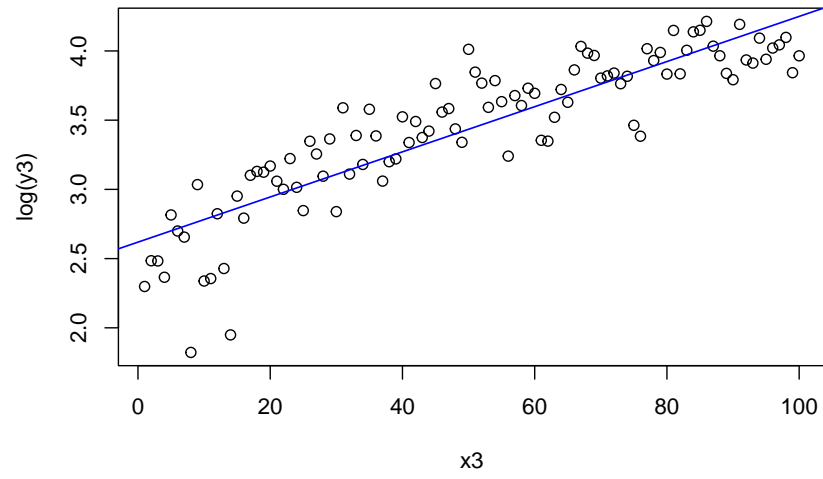
The residuals scatter around the zero horizontal line. **But** the variation of the residuals disperses as the fitted values increasing.

### Section 1, Question 2

**Comments:** Linearity is a more important assumption than homoscedasticity. Thus, if we cannot resolve the heteroscedasticity without the linearity assumption becoming unreasonable, then we choose a model for which the linearity assumption is reasonable, even if it suffers from heteroscedasticity.







### Section 2 Question 3

Our final artificial dataset is sampled from the model

$$Y_j = 100 - 20x_j + x_j^2 + \epsilon_j,$$

where  $\epsilon_1, \dots, \epsilon_{100}$  are iid  $N(0, 400)$ .

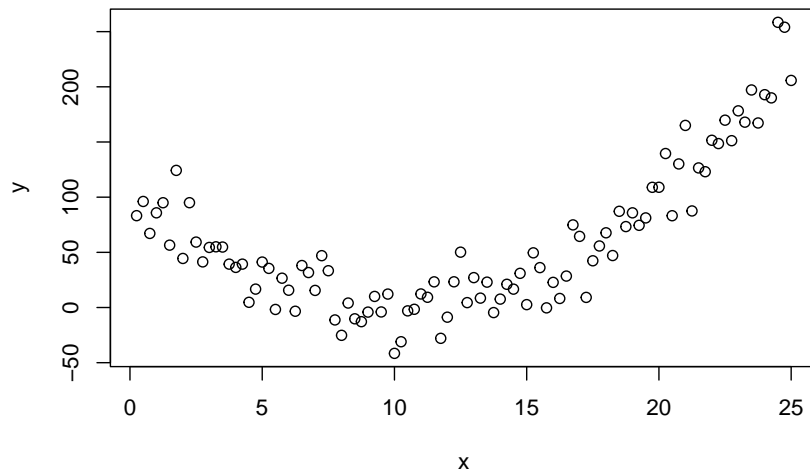
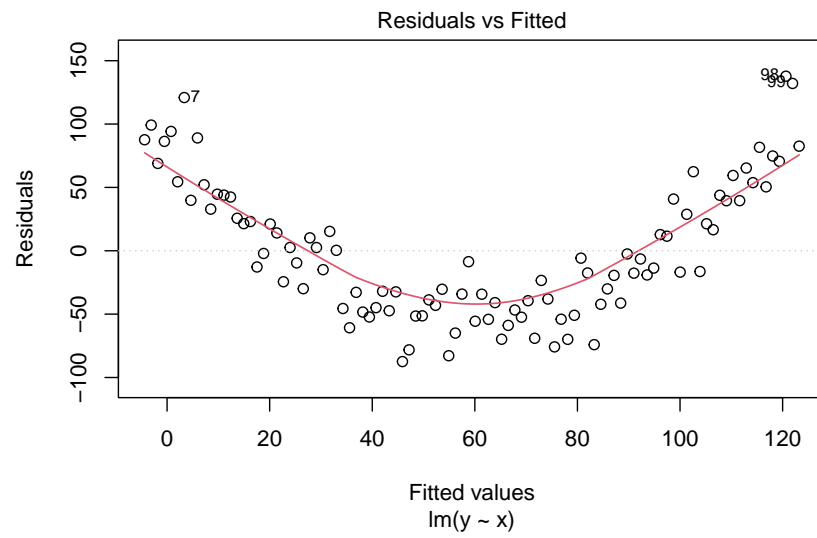
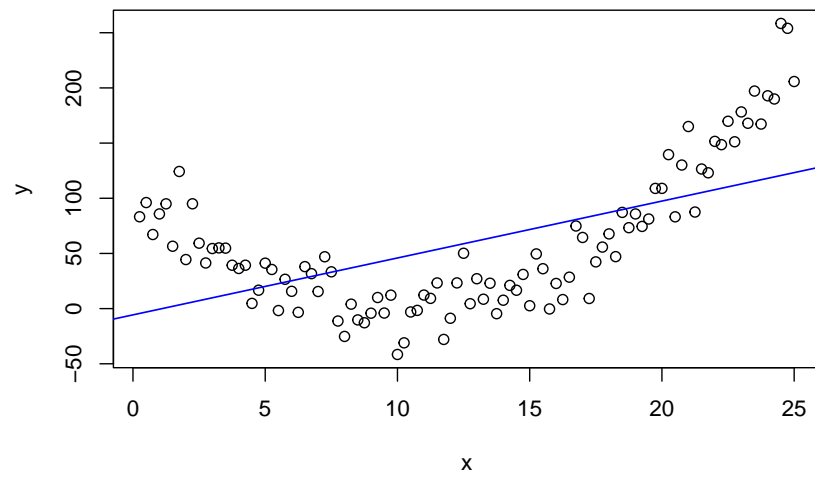


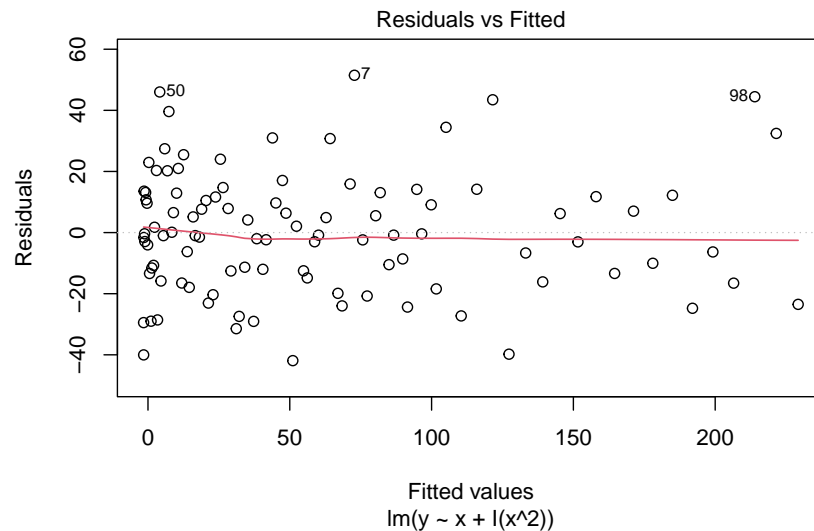
Figure 2: Scatterplot of Dataset 4.

### Section 2, Question 4

$$Y_j = \beta_0 + \beta_1 x_j + \epsilon_j.$$



$$Y_j = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \epsilon_j.$$



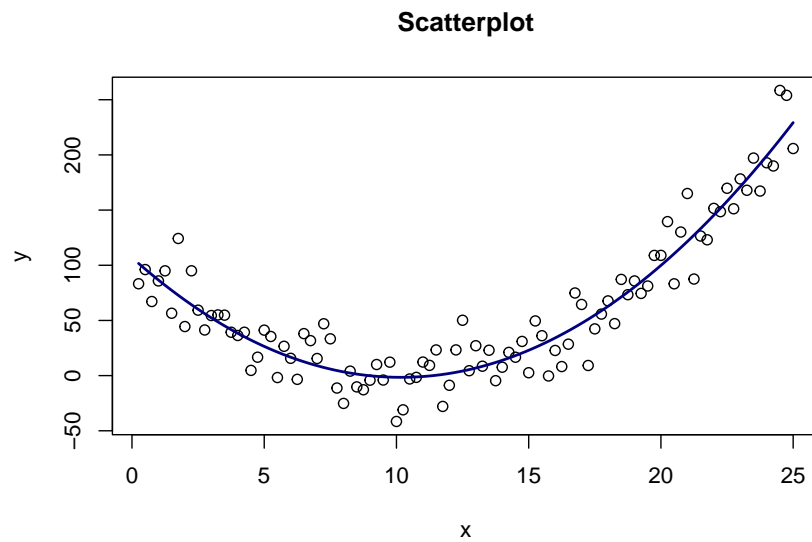
Only the residual plot of the quadratic regression model is a null plot.

## Section 2, Question 5 and 6

We make use of the function `quadraticPlot` to visualise the quadratic regression model. For element  $i$  in the `ax` vector, `predict(m, newdata=list(x=ax))` provides  $\hat{\beta}_0 + \hat{\beta}_1 ax[i] + \hat{\beta}_2 ax[i]^2$ .

```
quadraticPlot <- function(x, y){
  # Produce a scatterplot of the data
  plot(x, y, xlab="x", ylab="y", main="Scatterplot")
  # Fit a quadratic regression model
  m <- lm(y ~ x + I(x^2))
  # Create a vector ax of length 101
  # spanning the range of the explanatory variable
  ax <- seq(min(x), max(x), length.out=101)
  # predict the response for each value in ax
  fitted.curve <- predict(m, newdata=list(x=ax))
  # Fit a curve through the predicted response values and
  # add to this to the plot
  lines(ax, fitted.curve, col="navy", lwd=2)
}
```

This scatterplot shows that the quadratic regression model provides a good fit to the data.



## Section 2, Question 7

Consider

$$Y_j = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \epsilon_j, \quad j = 1, \dots, n.$$

One unit change of  $x_j$  will affect both  $\beta_1$  and  $\beta_2$ . Hence  $\beta_1$  and  $\beta_2$  should not be interpreted individually.

## (Intercept)	x	I(x <sup>2</sup> )
## 106.92425	-21.33701	1.04922