Solution to Series 5

1. **a)** From the three R formulae we can derive the following:

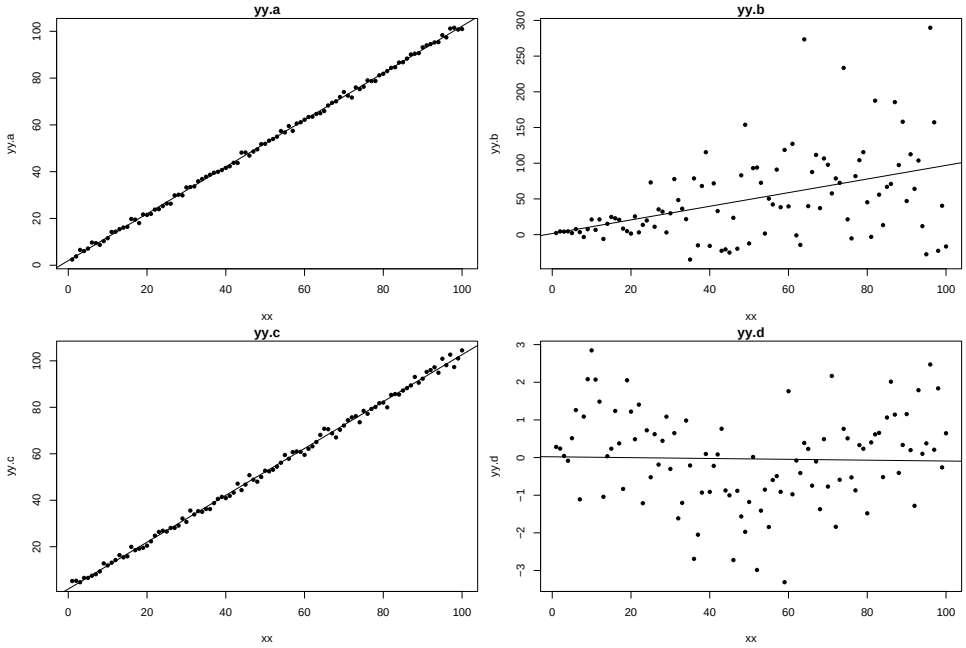
.a Model assumptions valid.

.b Model contains strong non-constant variance.

.c Variance slightly non-constant.

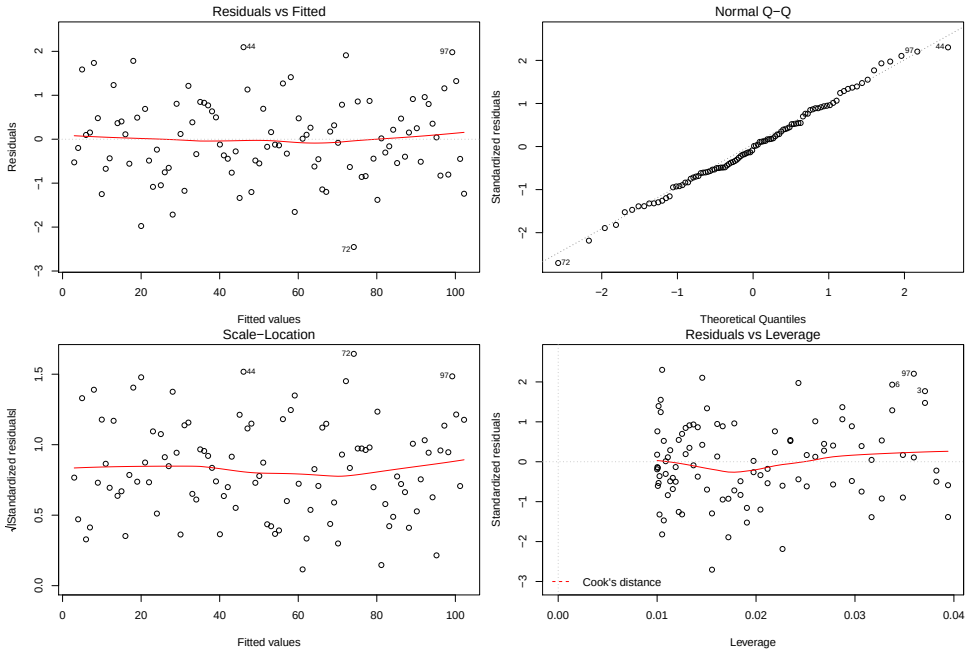
.d Non-linear model.

b)
```
> set.seed(123) #To make data reproducible.
> n <- 100
> xx <- 1:n
> yy.a <- 2+1*xx+rnorm(n)
> yy.b <- 2+1*xx+rnorm(n)*(xx)
> yy.c <- 2+1*xx+rnorm(n)*(1+xx/n)
> yy.d <- cos(xx*pi/(n/2)) + rnorm(n)
> par(mfrow=c(2,2))
> fit.a <- lm(yy.a ~ xx)
> plot(xx, yy.a, main="yy.a", pch=20)
> abline(fit.a)
> fit.b <- lm(yy.b ~ xx)
> plot(xx, yy.b, main="yy.b", pch=20)
> abline(fit.b)
> fit.c <- lm(yy.c ~ xx)
> plot(xx, yy.c, main="yy.c", pch=20)
> abline(fit.c)
> fit.d <- lm(yy.d ~ xx)
> plot(xx, yy.d, main="yy.d", pch=20)
> abline(fit.d)
```



c) Model diagnostics yy.a

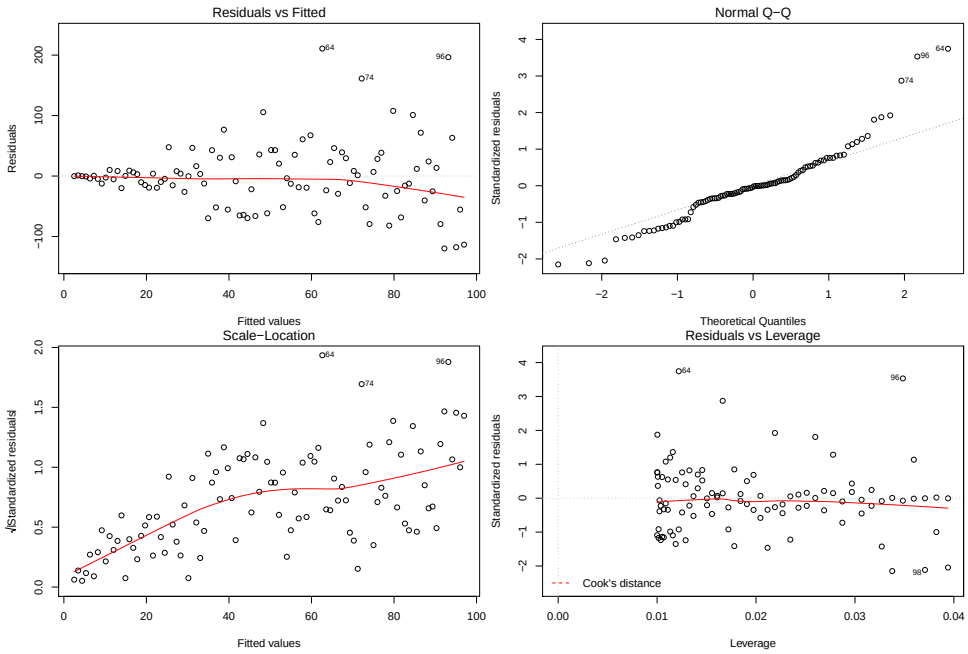
```
> par(mfrow=c(2,2))
> plot(fit.a)
```



`yy.a`: From the Residuals vs Fitted (Tukey-Anscombe) and Scale-Location plots we conclude that the constant variance of the errors assumption is satisfied. Moreover, looking at the Tukey-Anscombe plot, we see that neither the zero-expectation of the errors nor the uncorrelated errors assumptions are violated (the red line seems to be close to the x-axis and we cannot identify a non-random structure in the data). Furthermore, the Q-Q plot does not show strong evidence against the normality assumption.

`Model diagnostics yy.b`

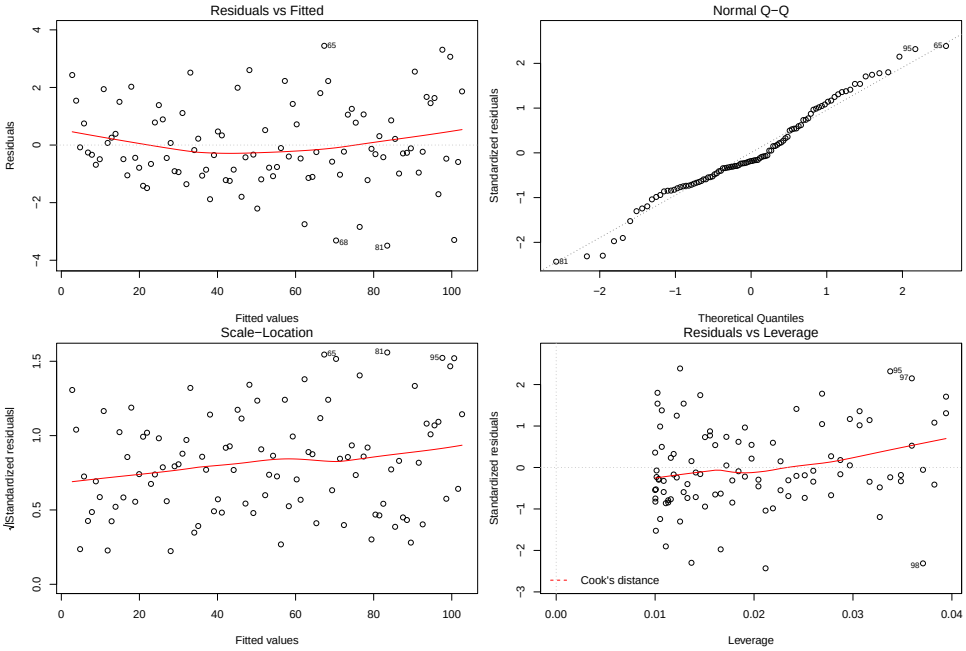
```
> par(mfrow=c(2,2))
> plot(fit.b)
```



`yy.b`: The Tukey-Anscombe and Scale-Location plots show residuals with strong non-constant variance: the residuals are bigger for larger fitted values. From the Tukey-Anscombe plot, we conclude that the zero-expectation and uncorrelated errors assumption are satisfied. The Q-Q plot provide evidence against the normality assumption, which is what we would expect if we look at the model equation.

`Model diagnostics yy.c`

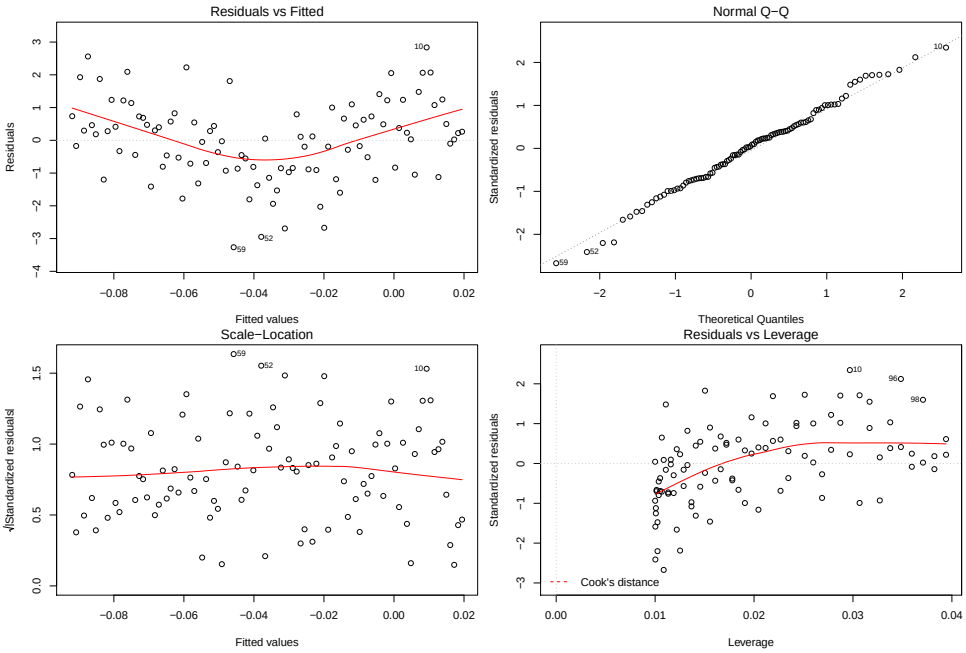
```
> par(mfrow=c(2,2))
> plot(fit.c)
```



`yy.c`: The Tukey-Anscombe plot again show evidence against the non-constant variance assumption. However, it is less accentuated than in the previous example because the residuals have smaller values than in `fit.b`. From this plot, however, we can see that the zero-expectation and uncorrelated errors assumption are satisfied. From the Q-Q plot, we conclude that the normality assumption is slightly violated as we would expect by looking at the model equation.

`Model diagnostics yy.d`

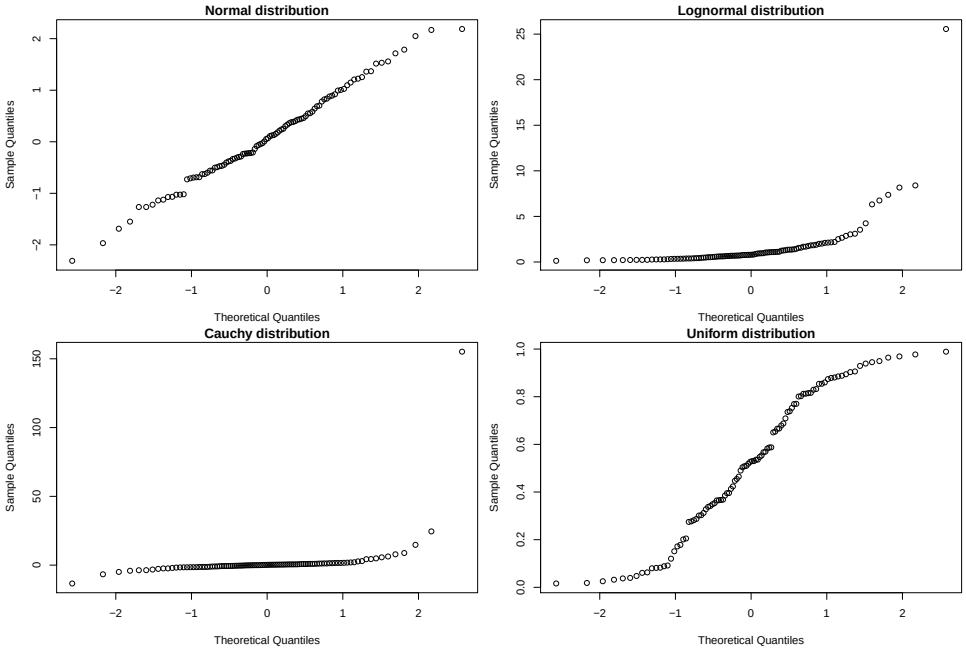
```
> par(mfrow=c(2,2))
> plot(fit.d)
```



`yy.d`: From the Tukey-Anscombe plot we can see that this model is clearly non-linear since it exhibits a U-shaped pattern. Therefore, we can conclude the existence of a non-linear relation between response and predictor. From the Scale-Location, Tukey-Anscombe, and Q-Q plots , we cannot see strong evidence against the assumptions of constant variance, normality and uncorrelated errors.

d) The exercise should be repeated generating new random numbers (remember to change the argument of `set.seed` or just eliminate it). Manipulating the number of observations is also instructive. However, the above described structures are of general nature and will largely remain on the repetitions.

e)
```
> par(mfrow=c(2,2))
> set.seed(123)
> qqnorm(rnorm(n), main=c("Normal distribution"))
> qqnorm(exp(rnorm(n)), main=c("Lognormal distribution"))
> qqnorm(rcauchy(n), main=c("Cauchy distribution"))
> qqnorm(runif(n), main=c("Uniform distribution"))
```



`Normal distribution` The sample quantiles fit nicely to the theoretical quantiles of a normal distribution. Deviations from the diagonal line are to be expected due to randomness.
`Lognormal distribution` The curve is bent upwards. This indicates a positively skewed distribution of the sample points.
`Cauchy distribution` The distribution of the data seems to be fairly symmetric. However, the curve has the shape of an inverted S which indicates that this distribution has heavier tails than those of a Normal distribution.
`Uniform distribution` We have the opposite case of the Cauchy distribution. Here, the curve is S-shaped and we conclude that the distribution of this sample has shorter tails than those of a normal distribution.

f) Repeat the exercise generating new random numbers (remember to change the argument of `set.seed` or eliminate it) and varying the number of observations as well.

2. **a)** In the full, general linear model, K regression lines are fitted. The parameters α and β are intercept and slope for the observations of category 1. The parameters γ_i describe for any category $j = 2, \ldots, K$ how much the intercept is changed with respect to category 1. The parameters δ_j describe how much the slope is changed with respect to category 1.
The regression lines of the full, general model are identical to the ones of the simple regression model for the categories. Tests, however, that compare the slopes and intercepts of the categories are only possible in the full, general model.

b) The null hypothesis is in words: "The slopes of the regression lines are the same for all categories." Mathematically speaking, this means $\delta_2 = \delta_3 = \ldots = \delta_K = 0$. Since the models are nested, we should use an F-test. If the nullhypothesis is true, we have

$$\frac{(SS_{e_0} - SS_e)/(K-1)}{SS_e/(n-2K)} \sim F_{(K-1),(n-2K)}.$$

The quantities SS_{e_0} and SS_e are the sums of the squared residues of the reduced model (i.e. $\delta_j = 0$) and of the full model (i.e. $\delta_j \neq 0$). We can take the expression on the left side as test statistics.

c) The R commands can be found on the question sheet. We obtain the p-value 0.009708.

d) Yes, the bigger model does fit better.

3. **a)** (iv): The sum of the residuals can be non-zero if the model does not contain an intercept term.

b) (i): This situation can occur if the predictors are correlated.