QUEEN MARY UNIVERSITY OF LONDON

MTH5120 Statistical Modelling I

Solutions Exercise Sheet 4

Solution for Question 1

The students could decide to use one of the following three models:

- add a cubic term in the model;
- add a square root;
- do the reciprocal.

1. We start with **Cubic Term** in the model.

   After loading the data by using read.csv as explained in the Practical, we initially estimate a linear regression model with cubic term. Comments:

   - Looking at the R^2, we see a value of 97.32%, which is a slightly higher value with respect to the quadratic regression (see Practical).
   - However, looking at the significance of the estimated parameters of the regressors, we see that the intercept, the parameter related to x_i and to x_i^2 are statistically significant at least at 5%, but the parameter related to the cube x_i^3 is no more statistically significant at 5% since it has p-value equal to 0.05541.
   - Following the principle of parsimony (see lecture notes), a simpler model is to be preferred and in this case one should choose the model with quadratic component only.

   Now, we move to the standardized residuals, and check the assumptions of linearity and constant variance. Figure 1.1 shows the plot of the standardized residuals vs x (left) and standardized residuals vs fitted values (right) for the linearity and constant variance respectively.

   From Figure 1.1, we have no problem of linearity neither of not constant variance. Next we move to the normality assumption and we show the QQ plot and the Shapiro-Wilk test:

   Figure 1.2 shows some problems in the right tails, but the Shapiro-Wilk test showed no problem with normality (p-value equal to 0.2686).

   Next we compute the leverage and the Cook's distance to check the presence of influential observations. Figure 1.3 shows the results for the Leverage (left) and for the Cook's distance (right). For the leverage, we should check with respect to the threshold ($2 \times 4/n = 8/36 = 0.22$) and ($3 \times 4/n = 12/36 = 0.33$). Thus, we have five

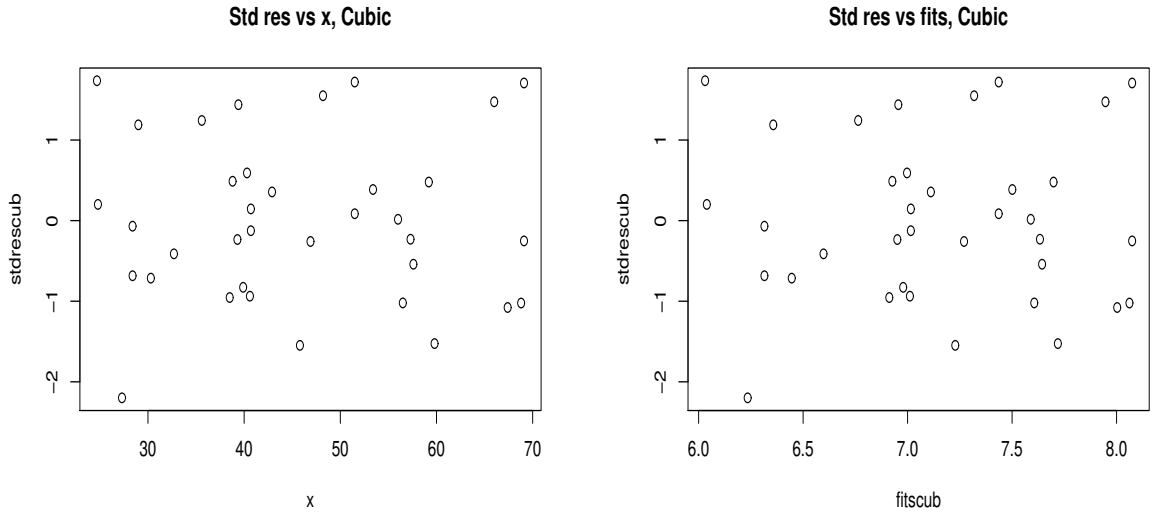

Figure 1.1: Plot of the standardized residuals versus x (left) and vs fitted values (right).

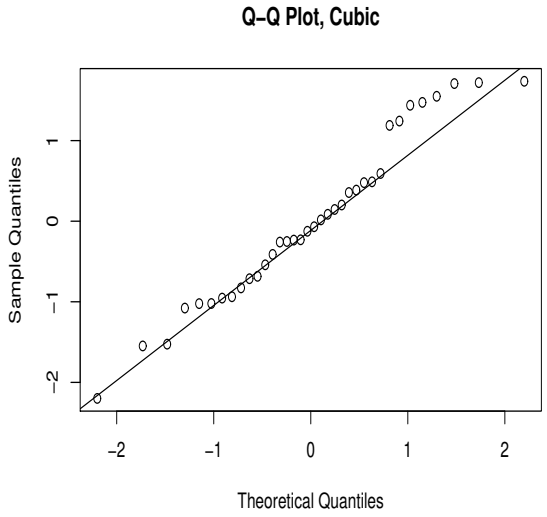

Figure 1.2: Plot of QQ plot.

values greater than the high leverage value and no-one bigger than the very high leverage value. Thus we should keep attention at observation $1, 2, 34, 35$ and 36, which has unusual values of x. For the Cook's distance, there is no value bigger than the 50th percentile of the F distribution with 4 and 32 degrees of freedom.

In conclusion, we check presence of outliers in the y from the standardized residuals and check if any values is greater than 3. In this case no values are bigger than 3, thus no problem of outliers.

2. We move to **Square root transformation of** y.

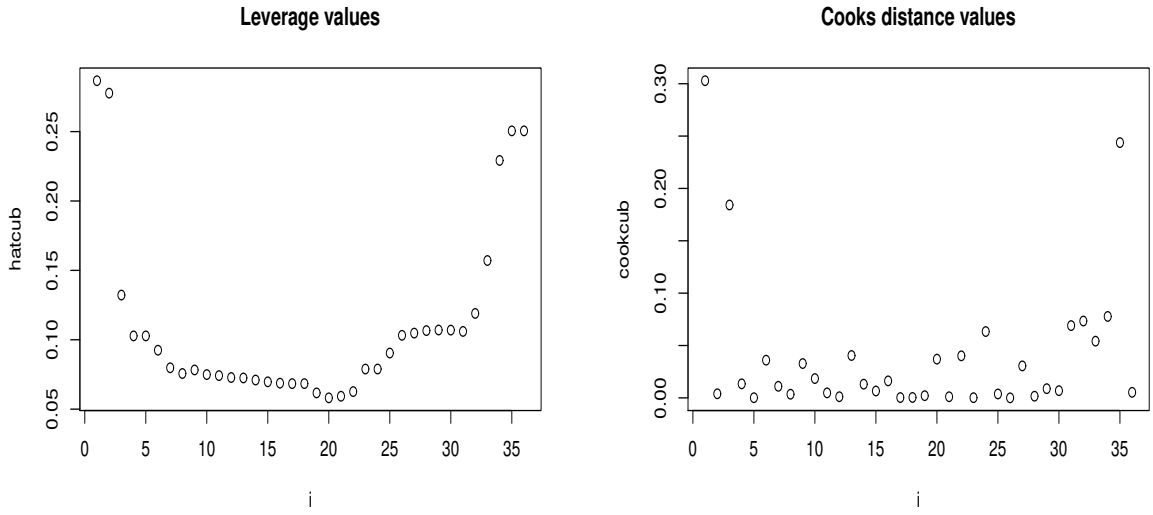

Figure 1.3: Plot of the Leverage valuels (left) and of Cook's distances (right).

After loading the data by using read.csv as explained in the Practical, we initially estimate a linear regression model with Square root transformation of y. Comments:

- Looking at the R^2, we see a value of 96.54%, which is a slightly smaller value with respect to the quadratic regression (see Practical).
- However, looking at the significance of the estimated parameters of the regressors, we see that the intercept is no more statistically significant at 5% since it has p-value equal to 0.398, while the parameter related to x_i is statistically significant at least at 5%.
- Due to problem with the funnel shape, it will be preferred the model with log response and quadratic regressor over the present model.

Now, we move to the standardized residuals, and check the assumptions of linearity and constant variance.

Figure 1.4 shows the plot of the standardized residuals vs x (left) and standardized residuals vs fitted values (right) for the linearity and constant variance respectively.

From Figure 1.4, we have no problem of linearity, while this model has a mild funnel shape in the standardized residuals versus fitted values, suggesting that the variance is not constant. Next we move to the normality assumption and we show the QQ plot and the Shapiro-Wilk test:

Figure 1.5 shows some problems in the right tails, but the Shapiro-Wilk test showed no problem with normality (p-value 0.6396).

Next we compute the leverage and the Cook's distance to check the presence of influential observations. Figure 1.6 shows the results for the Leverage (left) and for the Cook's distance (right). For the leverage, we should check with respect to the threshold

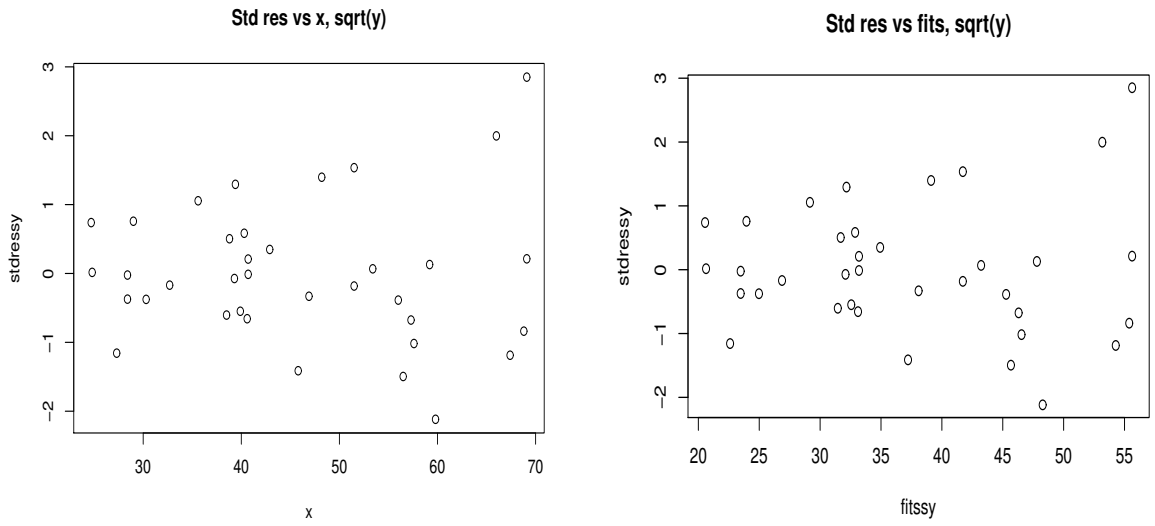

Figure 1.4: Plot of the standardized residuals versus x (left) and vs fitted values (right).

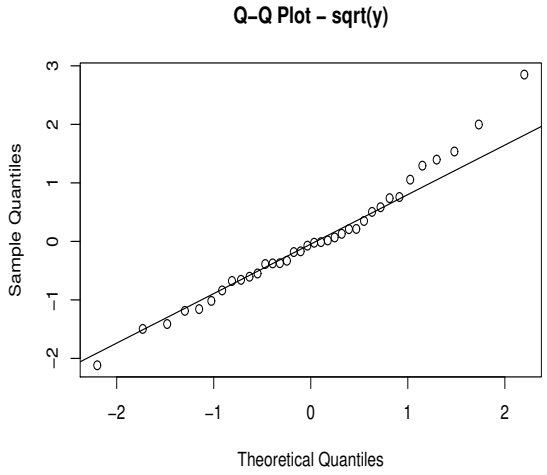

Figure 1.5: Plot of QQ plot.

$(2 \times 2/n = 4/36 = 0.111)$ and $(3 \times 2/n = 6/36 = 0.16)$. Thus, we have two values greater than the high leverage value and no-one bigger than the very high leverage value. Thus we should keep attention at observation 35 and 36 and check observation 34, which has unusual values of x. For the Cook's distance, there is no value bigger than the 50th percentile of the F distribution with 2 and 34 degrees of freedom.

In conclusion, we check presence of outliers in the y from the standardized residuals and check if any values is greater than 3. In this case no values are bigger than 3, thus no problem of outliers (although one should keep attention at observation 35).

3. We move to **reciprocal transformation of** y.

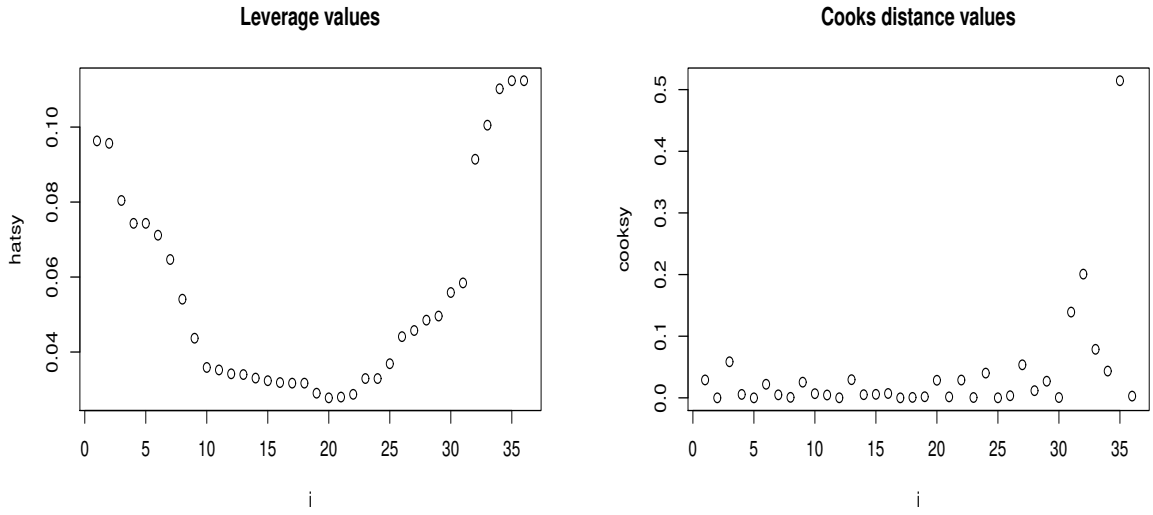

Figure 1.6: Plot of the Leverage valuels (left) and of Cook's distances (right).

After loading the data by using read.csv as explained in the Practical, we initially estimate a linear regression model with reciprocal transformation of y. Comments:

- Looking at the R^2, we see a value of 80.23%, which is a smaller value with respect to the quadratic regression (see Practical).
- However, looking at the significance of the estimated parameters of the regressors, we see that the intercept and the parameter related to x_i are statistically significant at least at 5%.
- Due to problem with the funnel shape and normality assumption, it will be preferred the model with log response and quadratic regressor over the present model.

Now, we move to the standardized residuals, and check the assumptions of linearity and constant variance. Figure 1.7 shows the plot of the standardized residuals vs x (left) and standardized residuals vs fitted values (right) for the linearity and constant variance respectively. From Figure 1.7, we have problems of linearity and of not constant variance, thus we need to add a quadratic term in the model. Next we move to the normality assumption and we show the QQ plot and the Shapiro-Wilk test.

Figure 1.8 shows some problems in the right and left tails, moreover looking at the Shapiro-Wilk test we have problems with normality, since we reject the null hypothesis of normality (p-value 0.01443).

Next we compute the leverage and the Cook's distance to check the presence of influential observations. Figure 1.9 shows the results for the Leverage (left) and for the Cook's distance (right). For the leverage, we should check with respect to the threshold $(2 \times 2/n = 4/36 = 0.111)$ and $(3 \times 2/n = 6/36 = 0.16)$. Thus, we have two values greater than the high leverage value and no-one bigger than the very high leverage value. Thus we should keep attention at observation 35 and 36 and check observation

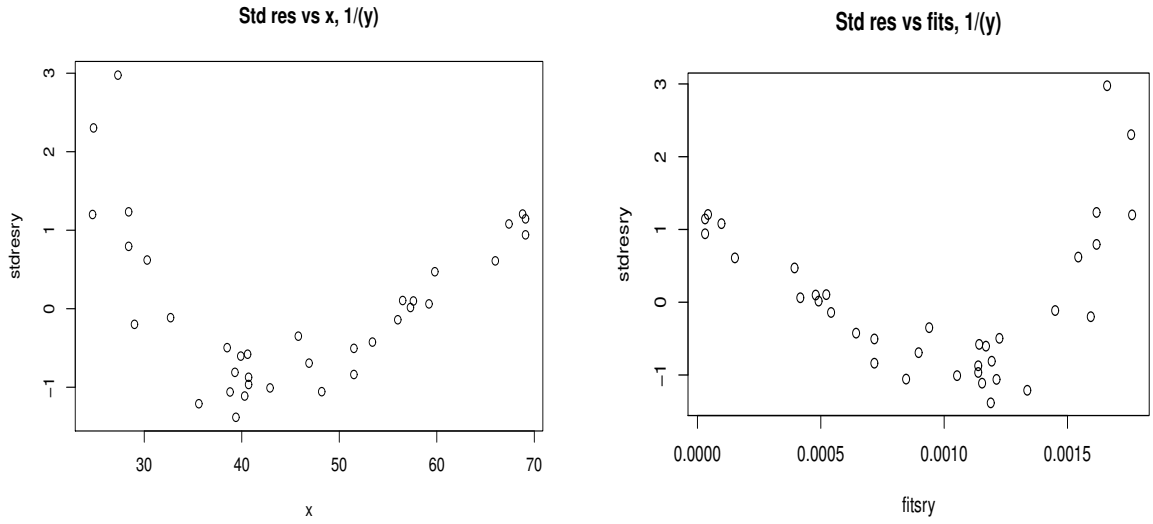

Figure 1.7: Plot of the standardized residuals versus x (left) and vs fitted values (right).

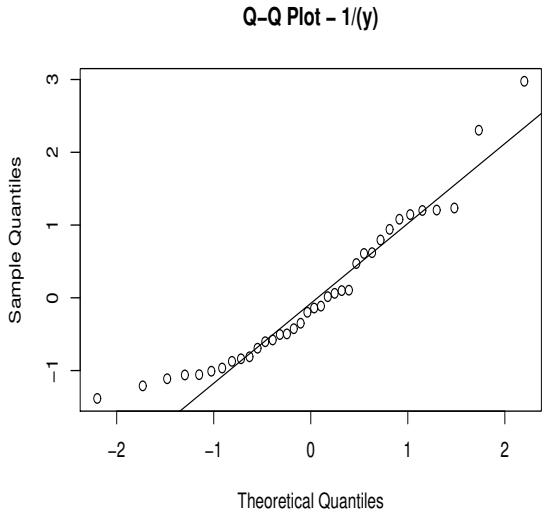

Figure 1.8: Plot of QQ plot.

34, which has unusual values of x. For the Cook's distance, there is no value bigger than the 50th percentile of the F distribution with 2 and 34 degrees of freedom.

In conclusion, we check presence of outliers in the y from the standardized residuals and check if any values is greater than 3. In this case no values are bigger than 3, but the observation 3 could be an outlier.

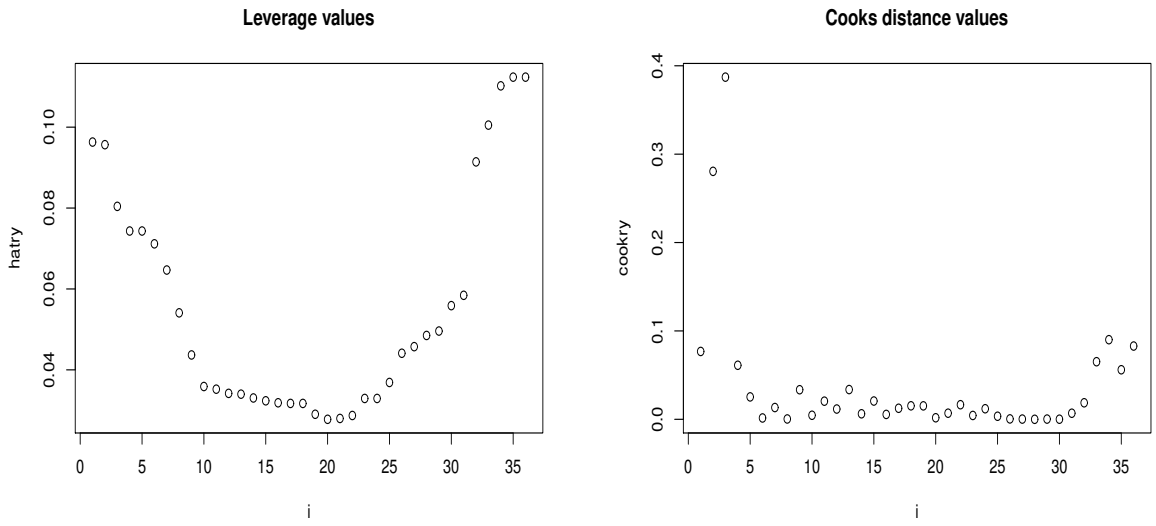

Figure 1.9: Plot of the Leverage valuels (left) and of Cook's distances (right).

Solution of Question 2

For each of the following models, say if it is a linear model or not. If it is not a linear model say if it linearisable. If it is give the linearised model.

(a) This is not a linear model and is not linearisable, since we will have the logarithm of the error term

(b) This is not a linear model but it can be linearizable. The new linear model is

$$log(Y_i - 3) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

since the parameters are linear and also the error are linear

(c) Both models are linear models, because the transformation by mean of square root or of cosine is not on the parameter but on the regressors.