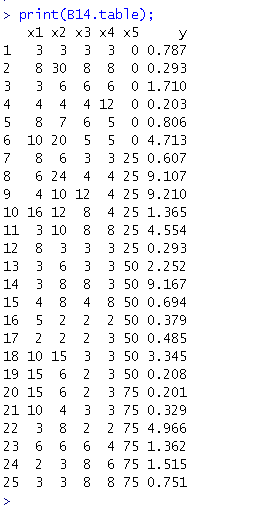
qwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnm

|  |
| --- |
| STAT 5532  Exercise 6.15  5/13/2011  Thanh Doan |

|  |  |
| --- | --- |
| Data set |  |

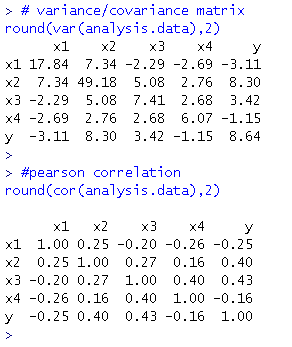
* The dataset contains transient points of an electronic converter.

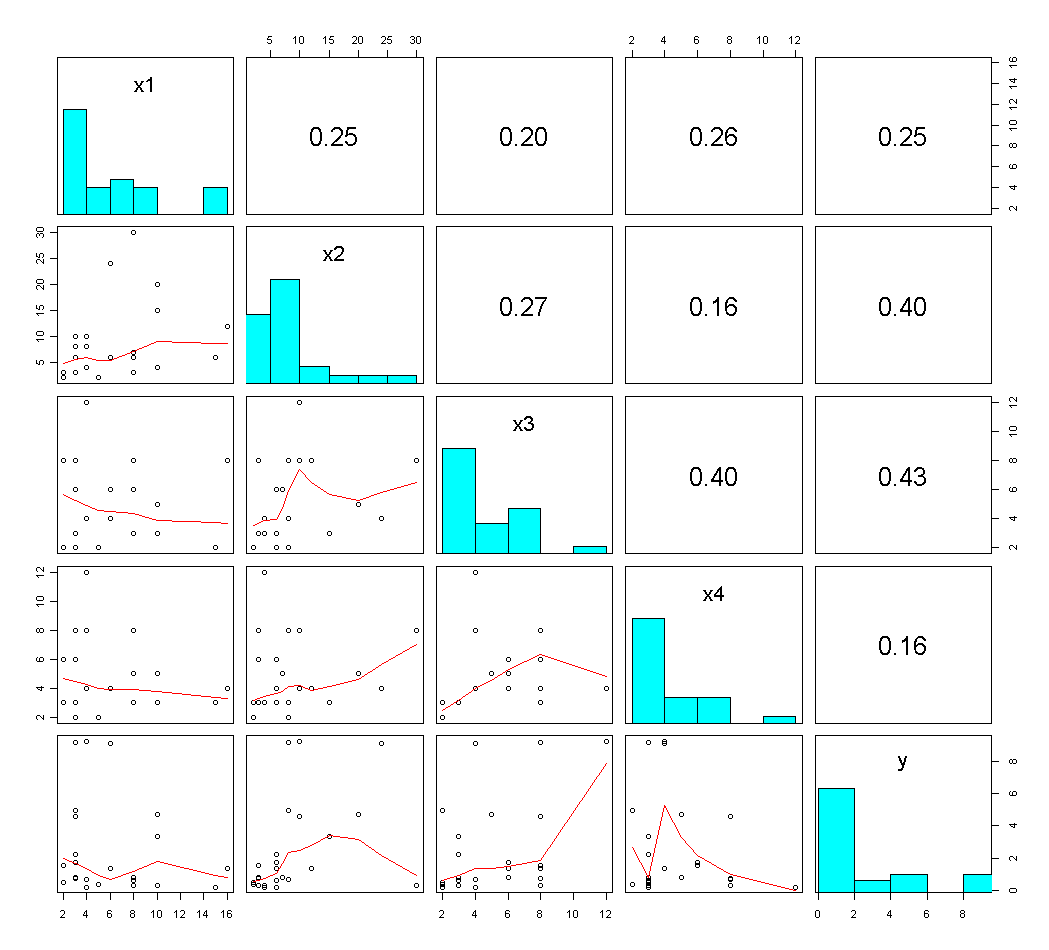


**Objective of the analysis**

* Fit a regression model to all 25 observations but only use x1-x4 as the regressors.
* Investigate this model for influential observations and comment on the findings

1. **Compute the variance of all study variables as well as Pearson correlation to have some understanding about the variability of each variable and the pair-wise correlation between them.**

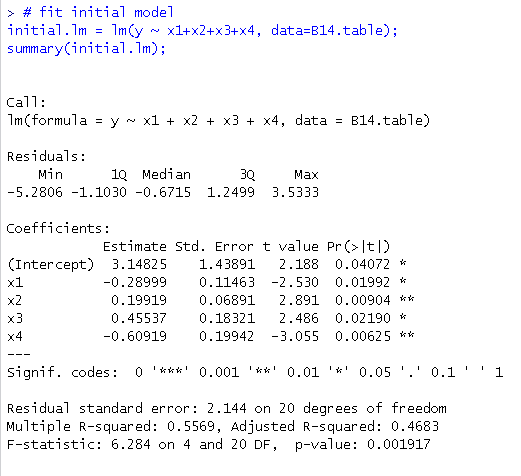




**Comment:**

* Among 4 covariates, x2 and x3 has highest correlation with y.
* Among 4 covariates themselves, indicates the correlation between x3 and x4 is highest.
* The numbers 0.4 and 0.43 are not very high indicating no single regressor dominantly influence the response variable y.

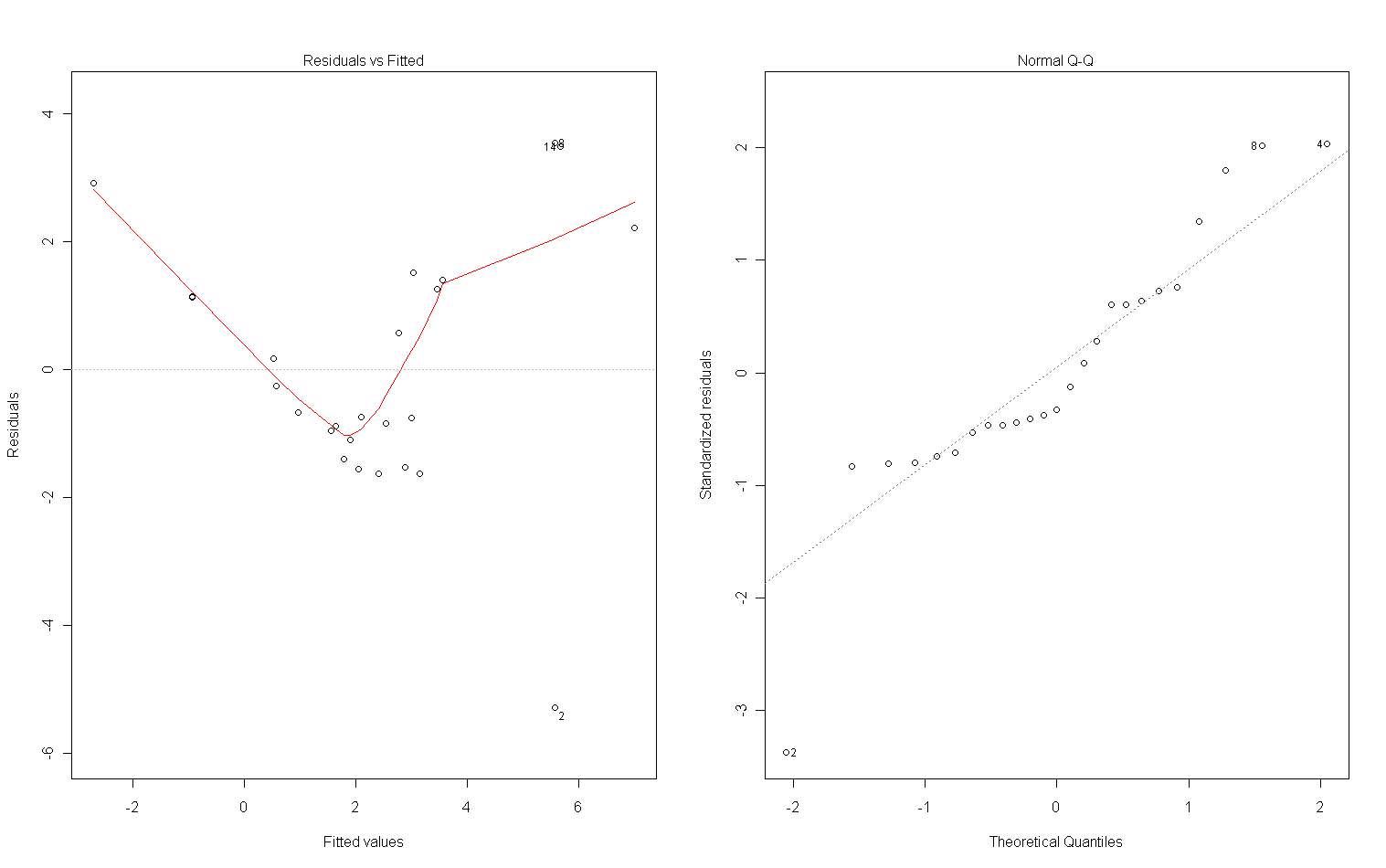
1. **Fit and study an initial (full) multiple regression model using all 4 regressor variables**



**Comments**

* The fitted model is
* The overall F test (F-statistic = 6.284, and its p-value = 0.0019) indicates that the regression is significant. One or more regressors are important.
* The individual t-test and its p-value are all less than 0.5… indicate that all four regressors have effect on the response variable y.
* We will analyze this further because individual coefficient and its corresponding t-test only tell part of the story (and sometimes misleading) due to the nature that their effect is partial and computed as it is given that all other regressors are already in the model.
* The adjusted R2 value is 0.468 is not very high

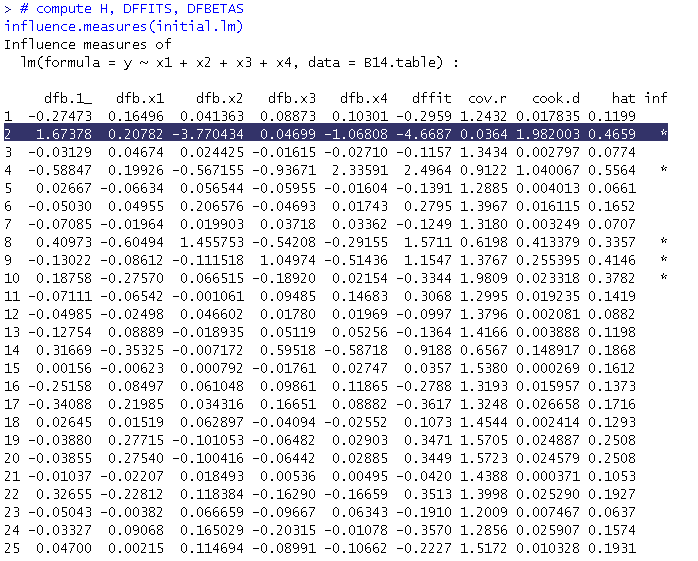
1. **Plot of the residuals versus the predicted values and the normal probability plot of the residuals**



**Comments**

* The normal probability plot of the residuals suggests that… residuals are not necessarily following normality in this case.
* Observation 2 has very high residual value (over 3 std. deviation from the mean zero).
* Observation 4 and 8 also have high residuals (over 2 std. deviation from the mean zero)
* The residuals versus the fitted values plot indicate that… residuals are not evenly split in around zero.
* There is a nonlinear pattern in the residuals versus the fitted values plot.

1. **Compute and analyze the leverage and measures of influence**
2. Compute **H** matrix and investigate hii.
3. Compute Cook’s **D** distance measure and analyze this measure
4. Compute DFFITS and DFBETAS measures of influences.



**Comments**

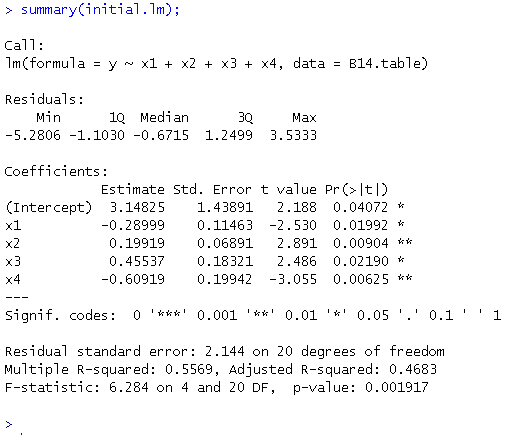
* The average size of hat diagonal is
* Observations 2, 4, and 9 have leverage greater than 2 times the average size of the hat diagonal. Here we use **2p/n** as the cutoff for leverage point
* In supplementary R code, I refit the model without observation 2. Observation 2 has significant impact on both the estimated. Removing Observation 2 also reduce the std. error of all parameter coefficients.
* Observations 2 and 4 have large Cook’s distance measure. Deleting point 2 and 4 would move to the boundary of an approximate 50% or more confidence region of based on the complete dataset.
* Deleting observation 2 and 4 also would move the vector of fitted values significantly.
* DFBETAS0,2 and DFBETAS2,2 and DFBETAS4,2 are large (greater than the cutoff value )
* Similarly observation 4, 8 and 9 also have large DFBETAS measures
* Another diagnostic for investigating the deletion influence of the ith observation on the predicted value is to analyze the DBFITSi
* Observation 2, 4, 8 and 9 have high DBFITSi measure (if we use cutoff point )

**Recommendation**

* From the analysis and findings on step 3 and 4… It is recommended to discuss with the person who collect and/or measure the data.
  + If the owner of the dataset agrees that observation 2 and 4 were recorded wrong then we can remove these influential points and refit the model. Then do further analysis on the new model without these data points.
  + If the person who obtained the data insists that these influential points are valid observations then we need to consider using an estimation method that is not impacted as severely by these influential points as Least Square. *May be we should consider using robust regression to down-weight these observations in proportion to residual magnitude or influence*.
* The next page is the screenshot of the model without two influential points (Observation 2, 4). **Note that:** even after removing points 2 and 4, we might not get a sensible model because here we only consider terms entering model linearly. However this is not the scope of chapter 6 exercise.
* The final model may involve variable transformation (see my work on exercise 5.16), variable selection or using different estimation techniques other than least square (such as using robust regression, nonlinear regression or using GLM modeling). However finding a sensible model using these techniques is not in the scope of exercise 6.15.

**Model 1**

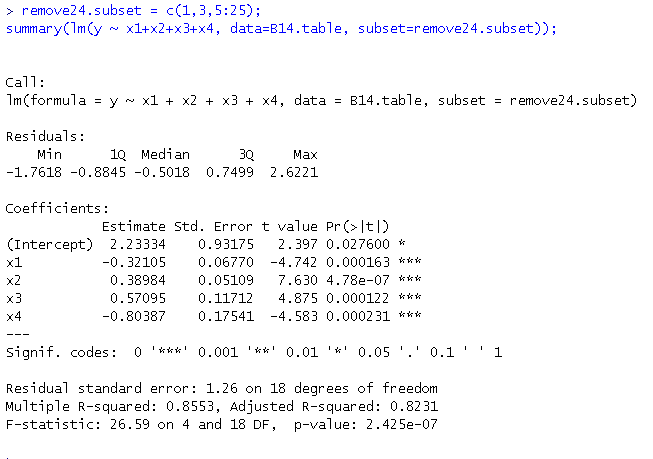
**(Initial model)**



* Model 1 is the initial model and it has larger error in its parameter estimated coefficients
* Equivalently, model 1 has smaller adjusted

**Model 2**

**(No influential points 2 and 4)**

****

* Model 2 has quite different estimated coefficients than model 1
* Model 2 has smaller Std. error for parameter estimation
* Model 2 has larger t statistic and equivalently smaller p-values