regression models – HW 2

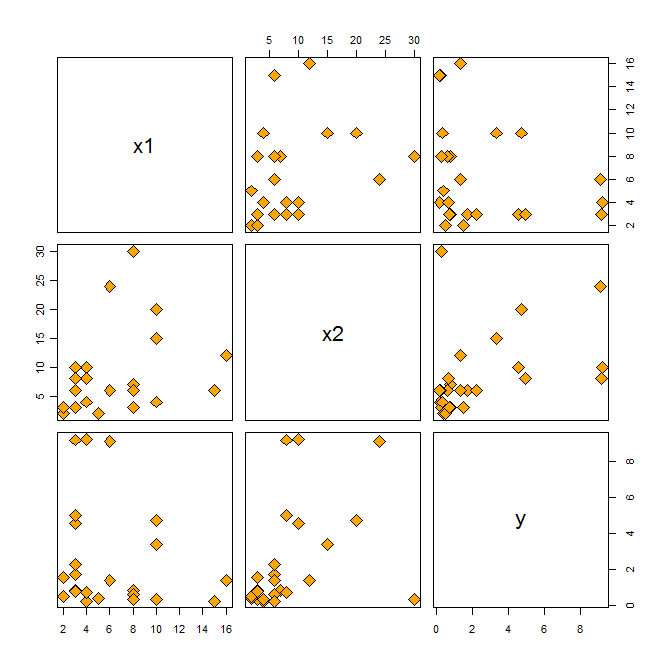
chapter 4-5-6 excercises

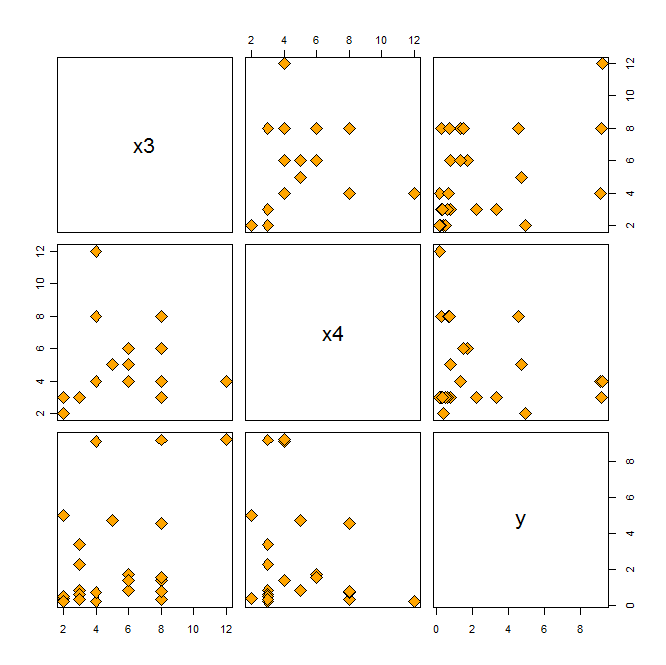
**Thanh Doan – Student ID 0159701**

|  |  |
| --- | --- |
| exercise 4.22 + 6.15 |  |

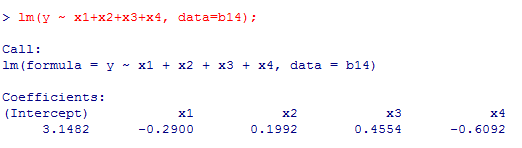
**4.22 a - Fit a multiple regression model** and investigate the adequacy of the model.

Before fitting a regression model, I used to look into the relationship between response variable and various repressors in the form of scatter plot matrix. The relationship between response variable y with x1 does not seem to be a linear relationship. Similarly, the relationship between response variable y with x3 and x4 does not seem to be a linear relationship either.





* **4.22 a.** Fit a multiple regression model where each x1, x2, x3, x4 enter the model linearly



* **4.22 a.** Check the adequacy of this model. First, look at the Normal Probability Plot of the residuals and plot the residuals against fitted values.



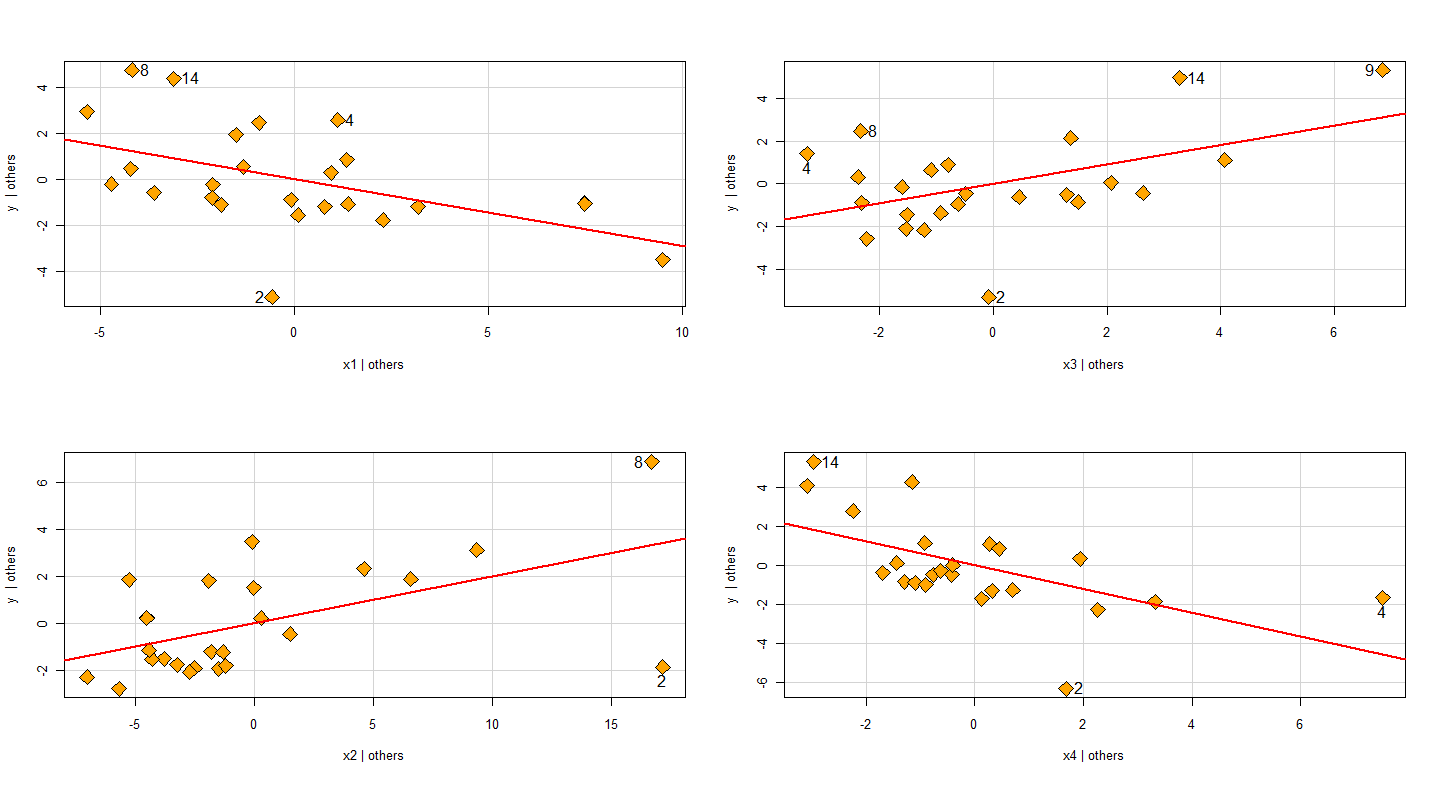


**Comment**

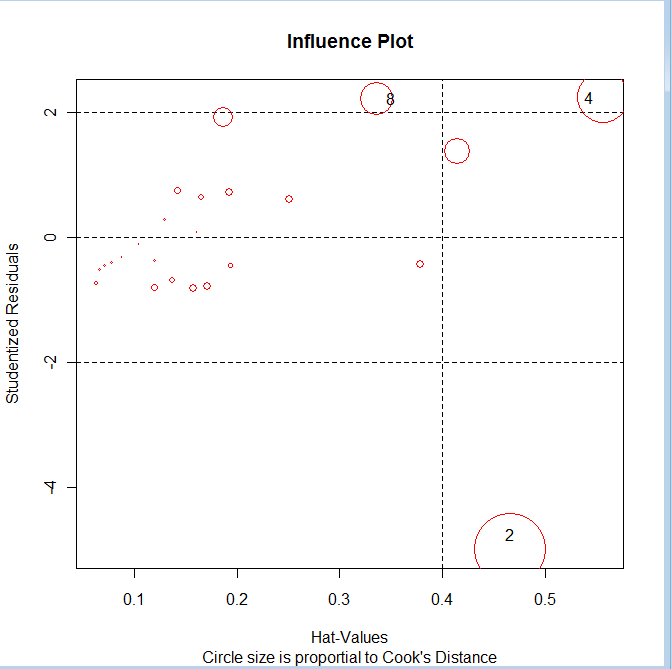
**From the Normal Probability Plot of the residuals and the residuals vs. fitted values plot… it appears that observation 2 is leverage or an influential point. Observation 4 and 8 also deserve further investigation. Also the error does not seem to have constant variance.**

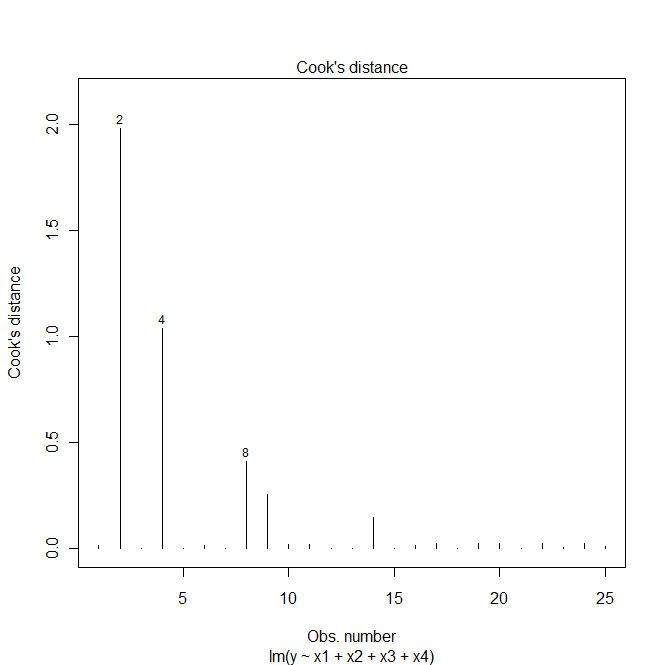
* Check the adequacy of this model. Next, look at partial regression (added-variable) plot, the residuals vs. leverage Plot, the influence plot and the cook’s distance report.

**Added-Variable plots**





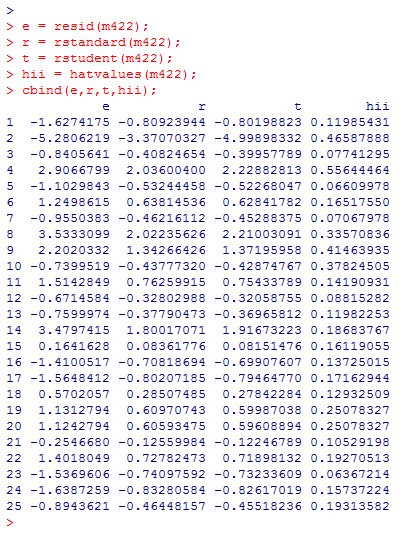


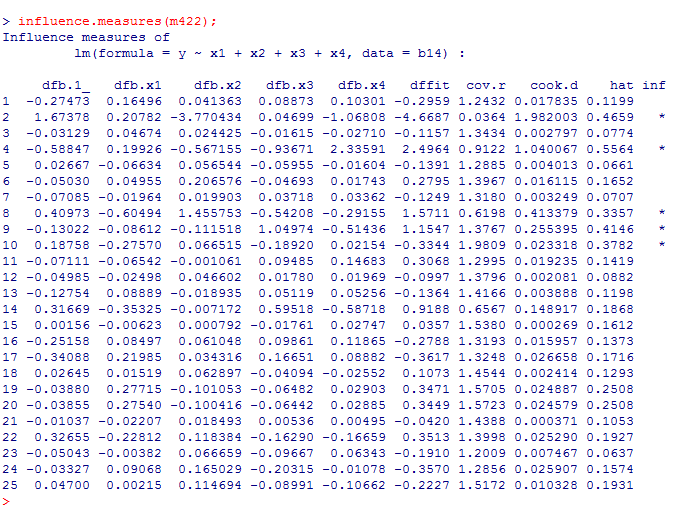


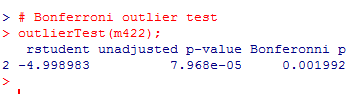
**Comment**

**The Residuals vs. Leverage Plot, the Influence plot, and the Cook’s distance plot all suggest observation 2 and 4 are influential points. Observation 8 also deserves further investigation.**

* Continue check the adequacy of this model. Look into details of various types of residuals, the hat values and various influence measures including dfbetas, dffits and cook’s distance measures. Also conduct Bonferroni outlier test.



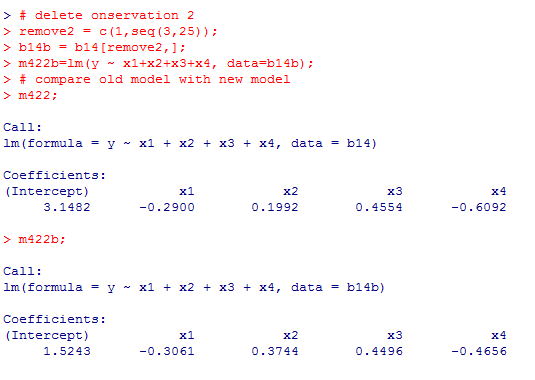




**Comment**

**From all of the above analysis (plots, test), there is strong evidence that observation 2 is a high influence point and a potential outlier. Observation 4 is a moderate influential point. Observation 8, 9, 10 also deserve further investigation.**

**4.22 b – delete observation 2 and refit the model**

****

**Thus, deleting observation 2 changes the intercept and the regression coefficient significantly. The coefficient is changed moderately but are not changed much.**

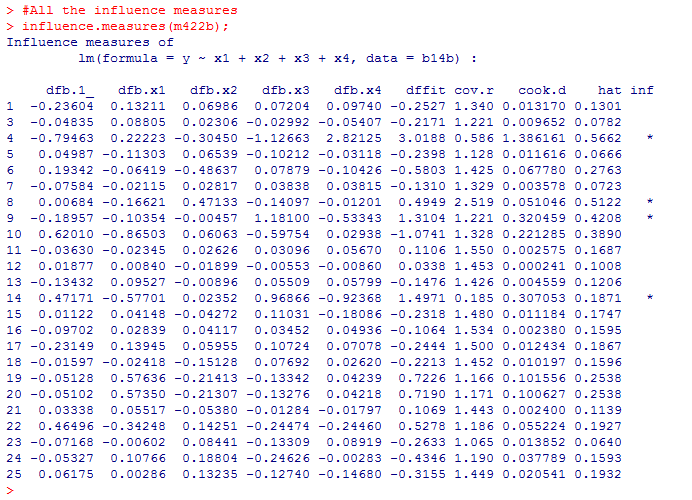
* Check the adequacy of this refitted model without observation 2. Look into below residual analysis and and various influence measures including dfbetas, dffits and cook’s distance measures.

****

****

****

****

****

**Comment:**

**From the Normal Probability Plot of the residuals, the residuals vs. fitted values plot and various diagnostic above… it appears that observation 4 is an influential point. Observation 8, 9, 14 also deserve further investigation. The error does not seem to have constant variance.**

|  |  |
| --- | --- |
| exercise 5.16 |  |

**5.16 a - Fit a multiple regression model**  and plot various residuals vs. predicted response.



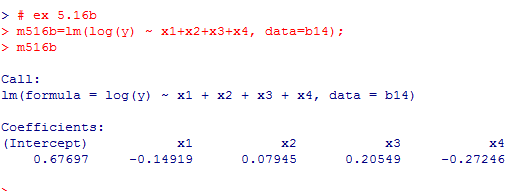




**Comment:**

**The residuals do not have constant variance. Patterns for residual plots are not satisfactory. There is a non-linear pattern to the residuals.**

**5.16 b – Transform the response variable y’ = ln y. Refit the multiple regression and check the adequacy of the model**







**Comment:**

**The residuals still do not have constant variance. Patterns for residual plots are not satisfactory. There is still a non-linear pattern to the residuals and observation 2 is an influence point.**

**5.16 c – Transform both response variable and repressors. Use partial regression, partial residual plots and CERES plots to aid finding suitable transformation.**





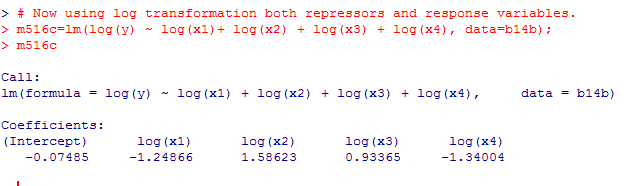






**Comment:**

**From the partial residual and CERES plots, using log transformation both repressors and response variables could linearize the model. The regression equation after perform log transformation on both repressors and response variables are:**

****

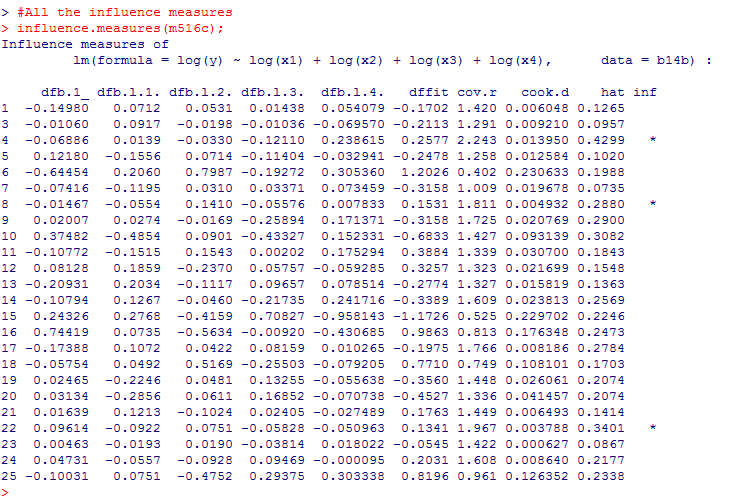
**Now check the adequacy of this model – m516c.**





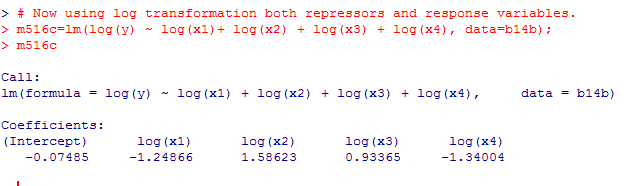






**Comment:**

**The final model – m516c – using log transformation on both response and explanatory variables is adequate in a sense that it fit the data well and produces the residuals that do not violate the assumptions**



**References**

<http://www.stanford.edu/class/stats191/diagnostics.html>

<http://www.statmethods.net/stats/rdiagnostics.html>