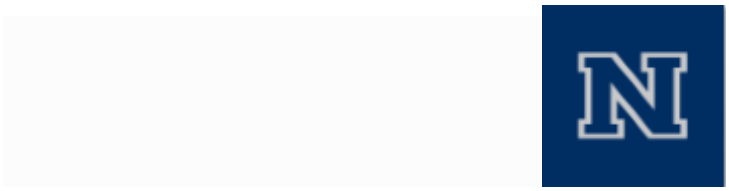
Assignment 4

Biplav Timalsina

March 11, 2018



STAT 757 Applied Regression Analysis

Instructions [20 points]

Modify this file to provide responses to the Ch.4 Exercises in @sheather2009. You can find some helpful code here: http://www.stat. tamu.edu/~sheather/book/docs/rcode/Chapter4.R. Also address the project milestones indicated below. Please email **both** your .Rmd (or roxygen .R) and one of the following either .HTML, .PDF, or .DOCX using the format SURNAME-FIRSTNAME-Assignment4.Rmd and SURNAME-FIRSTNAME-Assignment4.pdf.

Exercise 4.2.3 [60 points total]

3. The Sunday April 15, 2007 issue of the Houston Chronicle included a section devoted to real estate prices in Houston. In particular, data are presented on the 2006 median price per square foot for 1922 subdivisions. The data (HoustonRealEstate.txt) can be found on the book web site. Interest centers on developing a regression model to predict

   Y_i = 2006 median price per square foot

   from

   x_{1i} = %NewHomes (i.e., of the houses that sold in 2006, the percentage that were built in 2005 or 2006)

   x_{2i} = %Foreclosures (i.e., of the houses that sold in 2006, the percentage that were identified as foreclosures)

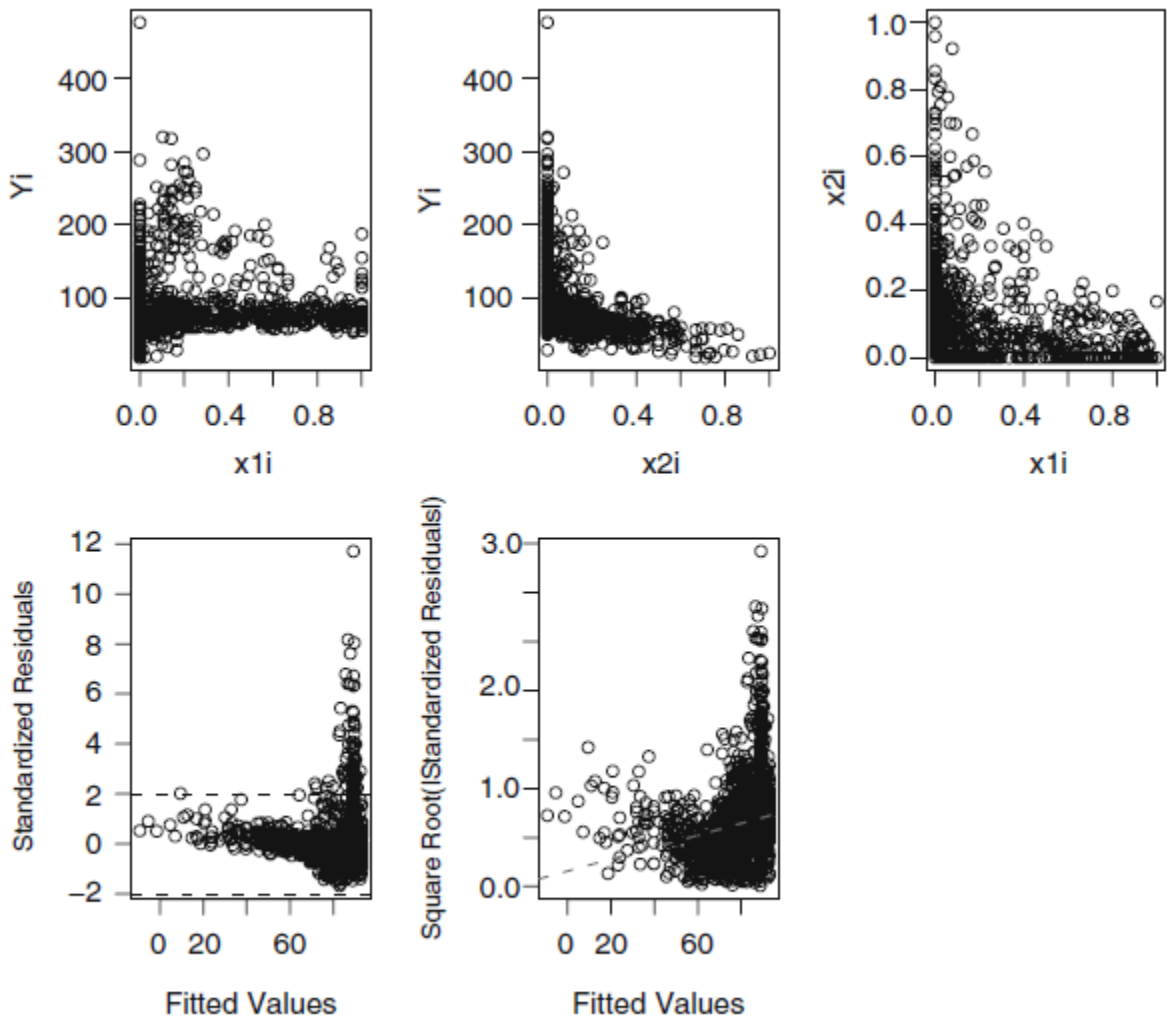


Figure 4.1 Plots associated with model (4.6)

for the $i = 1, \ldots 1922$ subdivisions.
The first model considered was

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e \tag{4.6}$$

Model (4.6) was fit used weighted least squares with weights,

$w_i = n_i$

where
n_i = the number of homes sold in subdivision i in 2006.
Output from model (4.6), in the form of plots, appears in Figure 4.1.

Part a [20 points] Explain it is necessary to use weighted least squares to fit model (4.6) and why wi = ni is the appropriate choice for the weights.

Weighted least squares is used to address the change in variance that different data points in our sample has. From figure 4.1 (a) and (b) we can see that the variance of Y-value (median price of houses) changes for different values of x1i and x2i (which violates the original assumtion that the variance of the error term is uniform). Therefore, we need to address this discrepancy by using weighted least squares to fit the model. The weighted least squares technique is commonly used in the important specialcase when Y_i is the average or the median of n_i observations so that variance of Y_i is inversely proportional to n_i. Because the Y-value is median values in our case, we are using n_i as weights.

Part b [20 points] Explain why (4.6) is not a valid regression model.

We can see from Fig 4.6(d) that the Standardized Residuals graph shows that variance is non-random which indicates the model has not properly addressed the data points. Also, if we observe the scatter plots (a) and (b), we can really not observe a linear (Y varies with x) relationship, rather it looks as it the value of Y remains almost constant for most of X1i and X2i apart from when x1i and X2i is around 0 i.e we see variability in data when x2i and x1i are around 0, and almost no variability in the rest of the domain. Observing closely, we could find exponential relationship (eˆ-x) in these figures. A lot of outliers could be seen in these models from fig (d). This is another indication that our data is not properly represented by our model.

is titled with positive slope which means variance is increasing which violates our assumption that the variance is constant.

Part c [20 points] (c) Describe what steps you would take to obtain a valid regression model (Figure 4.1).

We have understood that 1) there seems to be huge variance in data for x1i and x2i around 0, and little variance in rest part. Thus, we could divide the data points and generate seperate models for different domain of x1i and x2i.

2) We can see negated exponential like relationship (eˆ-x type) from (a) and (b), and can try to model the relationship by using such relationship.

3) On another thought, because Yi doesn't not vary much for X1i and X2i (see how they are almost at 100 for all values of x1i and around 50 for all values of x2i, for most case), we could question the selection of the independent variables as well, and try to explore newer independent variables.

Project milestones [20 points]

1. Review the relevant literature.
2. Identify a gap in knowledge that may be able to be address with your dataset.
3. Update your research question and hypothesis.
4. Draft a preliminary introduction for your written report.

Ans: Done with Dhurba Neupane. Can be seen in his assignment.

References