STEVE G MWANGI
R LAB 5
02/17/2020

TMATH 410 Computer Assignment 5
Lectures 9-11, Chapter 4-5
Objectives
1. Learn how to perform residual diagnostics
2. Interpret graphical and quantitative methods of residual diagnostics

Data

For this assignment we will continue to investigate predictors of county-level crime totals of serious crime incidents in 1990 (originally obtained by Kutner et al. (2005), Applied Linear Regression Analysis 5th ed, from the Geospatial and statistical Data Center, University of Virginia). These data are from 440 of the most populous US counties.

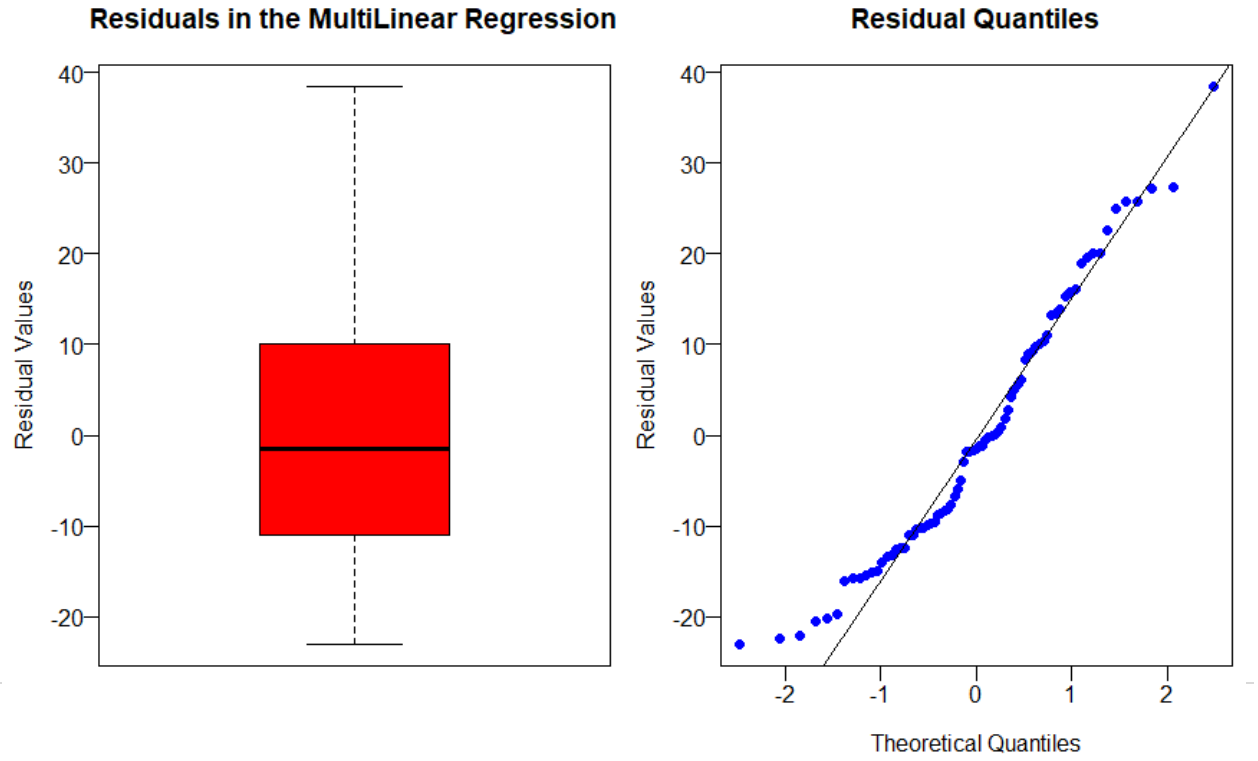
The variables in the data set are:

1. ID number of the county (ID)
2. Name of the county (County)
3. US state the county is in (State)
4. Total land area of the county (LandArea; square miles)
5. Total Population (TotalPop)
6. Total number of serious crimes (TotalSeriousCrime)
7. Percent of the population age ≥ 25 with at least a high school diploma (PercHS)
8. Percent of the population with at least a bachelor's degree (PercBach)
9. Percent of the population with income below the poverty level (PercPov)
10. Percent unemployed (PercUnemp)
11. Per capita income (PerCapInc)
12. Total personal income (TotInc)
13. Geographic region in the US (Region; 1=North East,2=North Central,3=South,4=West)

C1. (4) Submit the R script you used for this assignment.

Submitted

Figure 1 Boxplot and QQ plot of the residuals



C2. (2) Which regression assumption did these two plots (boxplot and a normal QQ plot of the residuals) assess? How did you interpret the plots?

Linearity and Normality Assumption. Interpretation: The residuals on the qq plot followed a pattern that denoted a linear relationship. Additionally, the boxplot denoted that the data was normally distributed, as can be seen by the symmetrical distribution about the mean/median (approx. 0) of the boxplot.

C4. (2) Compare the graphs of the standardized residuals in C3 to the same graphs of the residuals produced in computer assignment 4. What's the same? What's different?

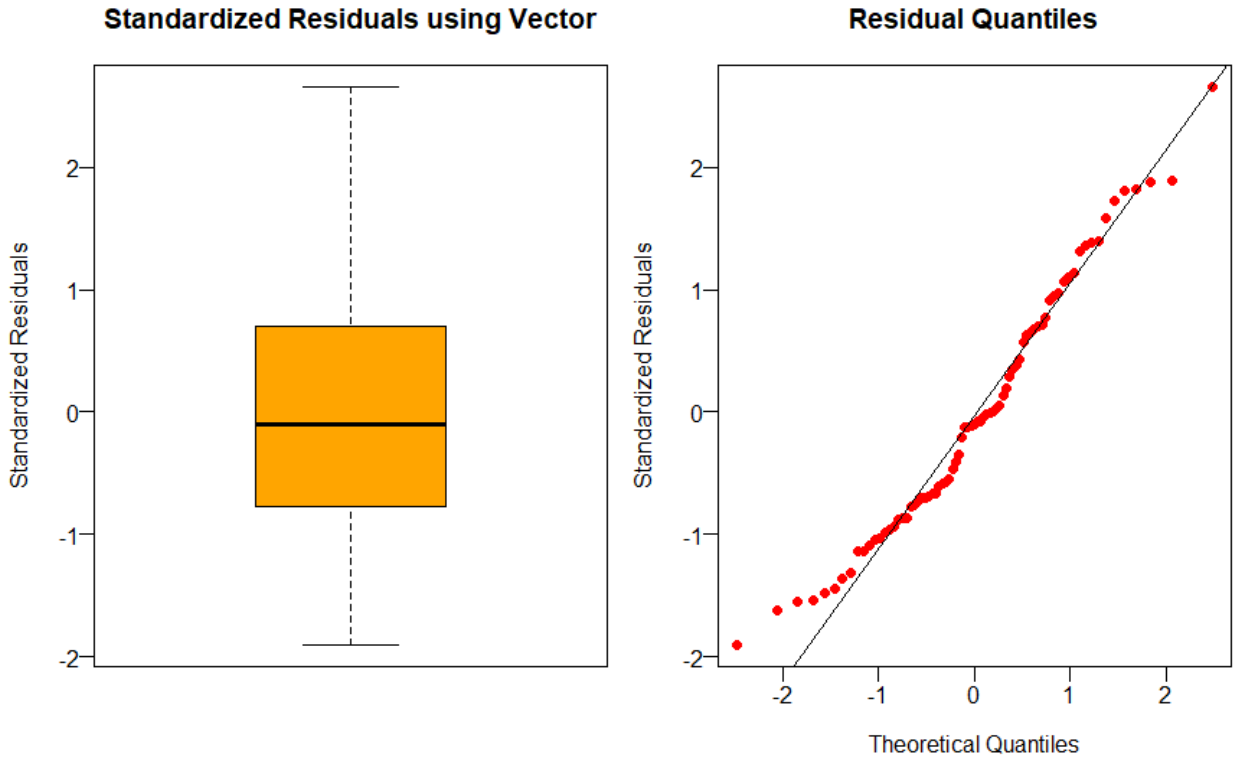


Figure 2 Standardized Residual Boxplot and QQ plot

The plots are similar in appearance to their original plots without the standardized values. The difference lies in the scale for the axes. Whereas these two plots are bounded within 3 units of 0—interval (-2,3), the original plots were bounded by values that were significantly larger with interval of about (-30, 40).

C5. (2) Create scatter plots of the standardized residual values against each of the predictor variables, and against the fitted Y values. This will be a set of 4 plots. You can present the plots in one window by using the par(mfrow=c(2,2)) command. Make sure to create nice labels on both axes! Submit your plot.

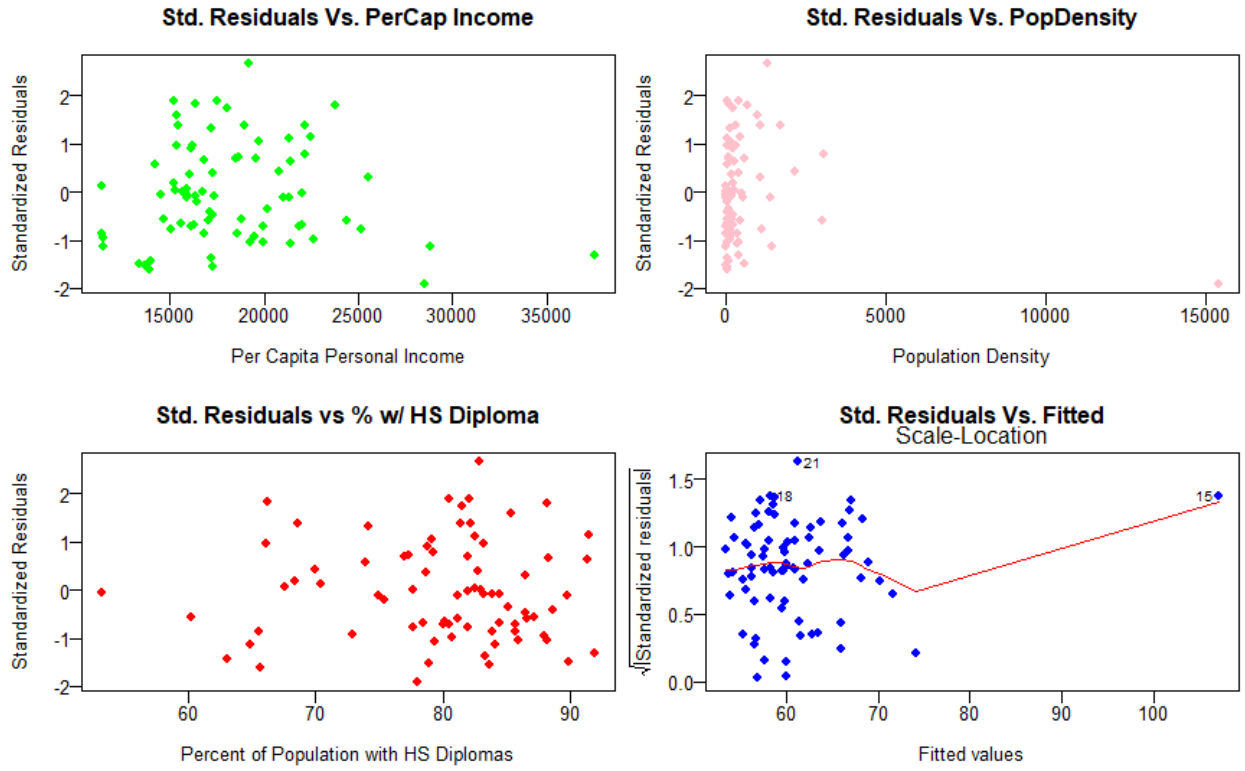


Figure 3 Scatter plots of the Standardized residuals against all predictor variables and then finally against the fitted values

C6 (2) Describe any patterns you see in the graphs in C5. Comment on which regression assumption(s) these plots assesses, and if you notice any potential violations.

These scatterplots assess whether we have a non-constant variance problem or a violation of linearity assumption. Overall, they help us determine if the pattern is random between the residuals and the predictors, which is the case we want. The first three plots could be said to have points evenly distribute about the line y = 0, in which case it indicates that there is no violation of either linearity or constant variance, and the residuals are distributed randomly, with no effects attributable to a possible relation between them and the predictors/X-values. The last plot though shows a violation of constant variance due to the single point that is by itself. Additionally, the Std. Residuals vs Percent with HS diplomas could also be said to somewhat indicate some heteroscedasticity problem. Additionally, the scatter plots point to an outlier that has been consistently distinct from the other points, as is seen in the Std. Residuals Vs. Population Density plot at point (15000, -2).

C7. (2) We can also use R to calculate the Cook's distance for each of our observations using the function cooks. distance. Use the function to create the vector of Cook's distances, and plot those distances. Provide the graph

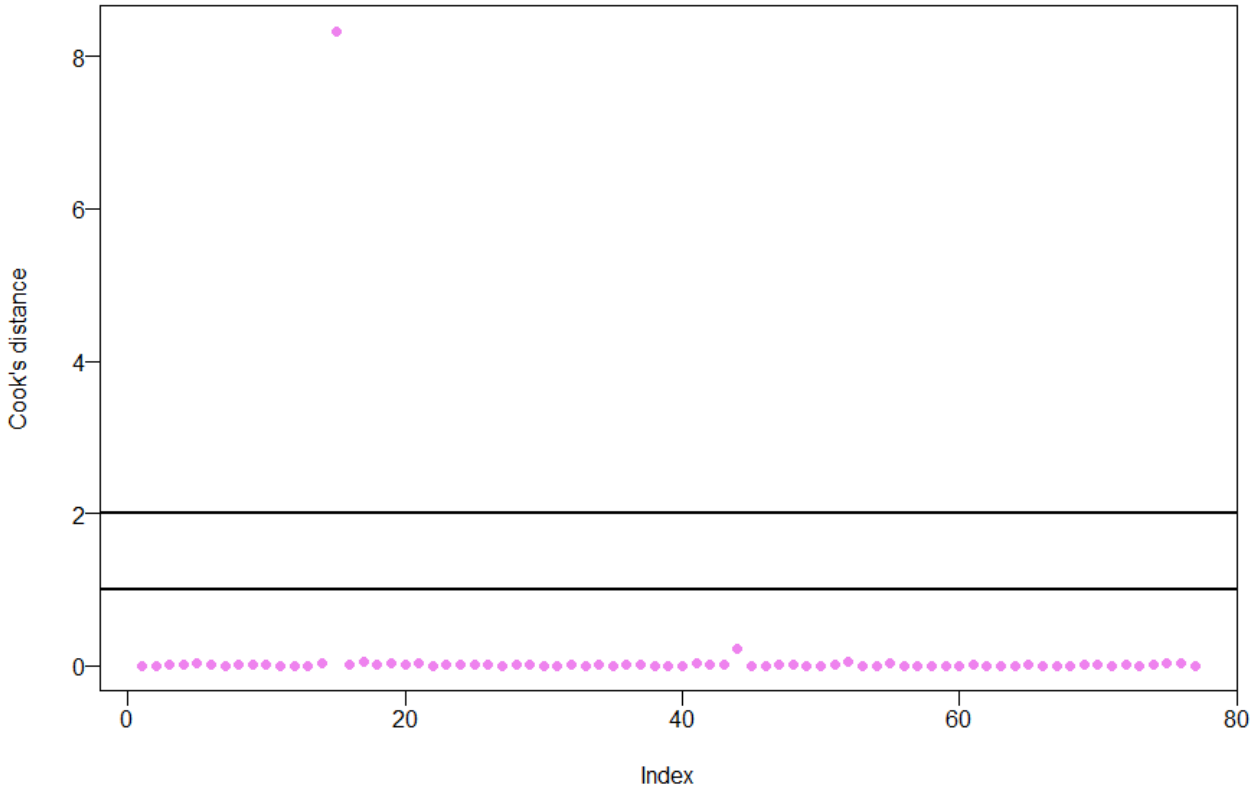


Figure 4 Index plot of the Cook's distance

C8. (2) Comment on any influential observations you see in C7.

The observation at about the index 18 is at Cook's distance of 8. This is significantly greater than all the other points which lie below the cook distance value 1. Hence this single point has great influence on the data we have.

C9. (2) R has a built-in residual diagnostic tool that creates some of our residual plots and highlights any outliers. Use the code below to create the built-in residual diagnostic plots. Submit the plots.

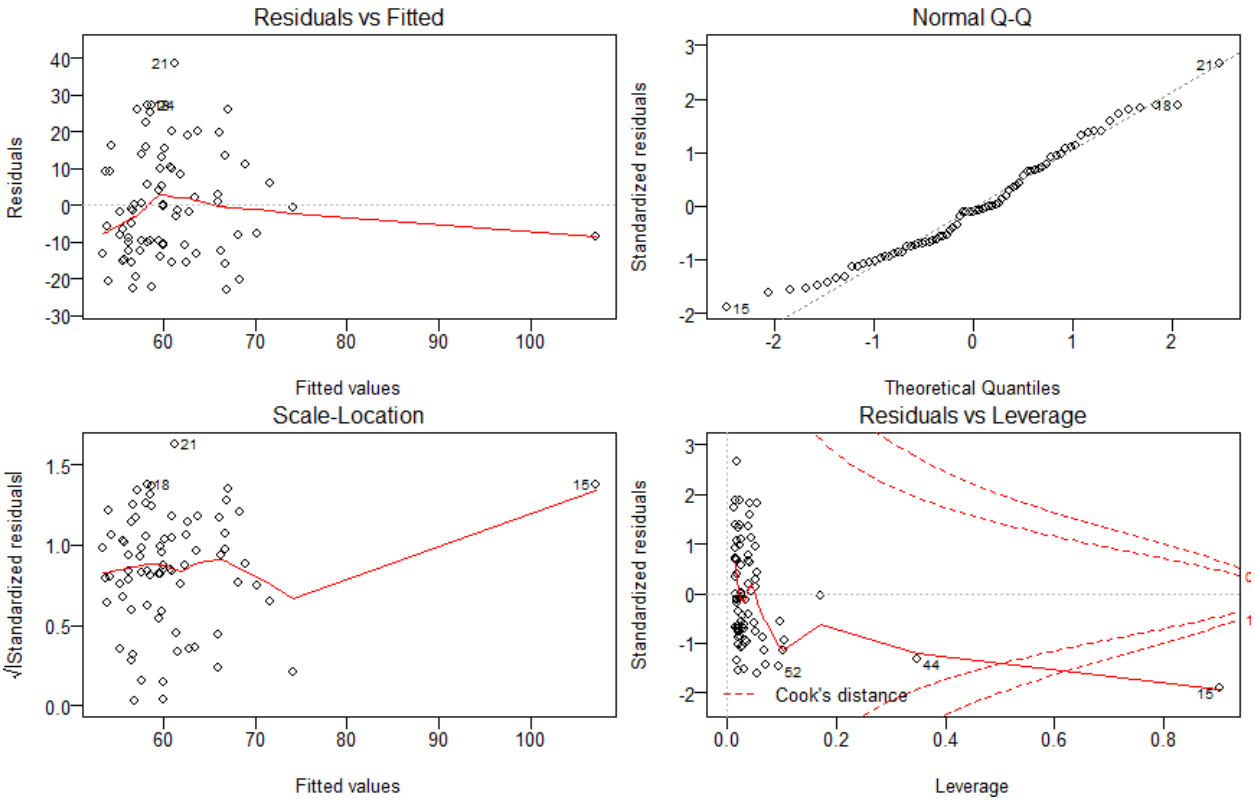


Figure 5 Diagnostic Plots built with R diagnostic tools.

C10. (3) Explain what each plot is composed of in C9. Use the plots to assess the regression assumptions for this model. Note, the $\sqrt{|standardizedresiduals|}$ against fitted values helps to assess constant variances (homoscedasticity). A flat pattern is consistent with constant variance.

Residuals vs Fitted: This plot displays the residual errors against the fitted values. The residual errors are not standardized but the fit line passes at y = 0 and is dotted. The red line helps us see the pattern of the residual points.

Normal Q-Q: The normal QQ-plot allows us to see if the residuals violate the normality assumption. Since most of the points follow the dotted line, normality assumption is not violated.

Scale-Location: This plot helps is to see how the standardized residuals vary over the fitted values. In this case we start off with constant variance or flat pattern, but due to a point labelled 15, the constant variance is violated.

Residuals vs Leverage: The dotted red lines indicate cook's distance relative to the plot. Point 15 again is beyond both the dotted lines, which indicated that is a point of great leverage and possibly great influence on the data, as can be seen by the way the red line bends towards it, showing that the point is skewing the distribution of the data towards it.

C11. (2) Extract the row(s) in the dataframe that R has flagged as influential in the Cook's distance graph. Speculate why those observations might have strong influence, and what we might do to remedy the model. See below for an example. Replace with the id's flagged in the analysis.

Table 1 Extracted values from the flagged points

""	"ID"	"County"	"State"	"LandArea"	"TotalPop"	"TotalSeriousCrime"	"PercHS"
"15"	53	"San_Francisco"	"CA"	47	723959	71234	78
"44"	206	"Marin"	"CA"	520	230096	9460	91.9
"52"	246	"Davis"	"UT"	305	187941	6279	89.9

Table 2 Continued Extracted Values Above

""	"PercBach"	"PercPov"	"PercUnemp"	"PerCapInc"	"TotalInc"	"Region"	"PopDens"	"PerThousCrimes"
"15"	35	9.7	5.6	28532	20656	4	15403.3829787234	98.3950748592116
"44"	44	3	4	37541	8638	4	442.4923076923 08	41.1132744593561
"52"	23.5	5.5	4.5	13394	2517	4	616.2	33.4094210417099

C12. (3) Write a paragraph summarizing what we have learned from the residual diagnostics. Make sure not to just outline what was done, but to explain what we learned from the process about this model in particular.

San Francisco county is a point of high leverage and probably high influence too. This county by itself is significant enough to influence our model, as can be seen with the high Cook's value (>8). The crime rate in the county could be attributable to, or affected by, other factors besides those in this model. As such, this could mean that it is probably best we exclude it from the model we have since it stands out as such an outlier by enhancing heteroscedasticity in the model we have, as can be seen in the standardized residuals vs fitted scatter plots. Additionally, since it has one of the lowest percent of people with high school diplomas, relatively high percent of people who are poor, but still high percent of people with bachelor's degrees compared to other regions and a relatively high per capita income, the relationship between its high population density and the high crime rate could possibly be of greater import, than that between crime rate and the other predictor variables in other counties. As such, since it is a clear outlier that skews both the distribution of the residuals, as can be seen in the residuals vs. fitted, it might be best that altogether that it is not included in the model. Its effects are of high leverage and great influence on our model as can be seen on the Cook's plot, and it might be appropriate that we disdain from using it in the model.