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: the id number of the county (ID), the name of the county (County), the US state the county is in (State), the total land area of the county (LandArea; square miles), the total population (TotalPop), the total number of serious crimes (TotalSeriousCrime), percent of the population age ≥ 25 with at least a high school diploma (PercHS), the percent of the population with at least a bachelor's degree (PercBach), the percent of the population with income below the poverty level (PercPov), percent unemployed (PercUnemp), per capita income (PerCapInc), total personal income (TotInc), and the geographic region in the US (Region; 1=North East,2=North Central,3=South,4=West)

On computer

Note, this is the same dataset we used in assignment 4, so if you have it already no need to download it again. Download the dataset named "CDI_Asgn4.csv" from Canvas and read it into your R session. For this assignment I name the R data frame cdi.df. Verify that your data frame has 13 columns and 440 rows. Below are the first 3 lines of the data set.

```
##   ID      County State LandArea TotalPop TotalSeriousCrime PercHS PercBach
## 1  1 Los_Angeles    CA     4060  8863164            688936   70.0     22.3
## 2  2        Cook    IL      946  5105067            436936   73.4     22.8
## 3  3      Harris    TX     1729  2818199            253526   74.9     25.4
##   PercPov PercUnemp PerCapInc TotalInc Region
## 1    11.6       8.0     20786   184230      4
## 2    11.1       7.2     21729   110928      2
## 3    12.5       5.7     19517    55003      3
```

In Assignment 4 we fit a linear model of per thousand serious crime to a linear combination of per-capita income, population density, and percent of adult population with a high school diploma, for the Western Region (Region==4). For this assignment we will continue working on the Region 4 subset of the data. **Note that from here forward ALL graphs are to be publication-quality!**

Make sure to create a column that measures population density and per thousand serious crime, and then create new data frame that isolates Region 4. Then estimate the linear model that predicts per-thousand serious crime by per-capita income, population density, and percent high school. This time we also want to make sure the rownames in this new data frame are specific to this data frame. Execute the lines of code below in that order!

```
# calculate population density and
cdi.df$PopDens=cdi.df$TotalPop/cdi.df$LandArea
cdi.df$PerThousCrimes=cdi.df$TotalSeriousCrime/cdi.df$TotalPop*1000
cdi4.df=cdi.df[cdi.df$Region==4,]
# and here we rename the rows to be indices pertaining to this
# dataframe only, in sequential order
rownames(cdi4.df)=1:nrow(cdi4.df)
# and now we estimate the linear model
cdi4.lm=lm(PerThousCrimes~PerCapInc+PopDens+PercHS,data=cdi4.df)
```

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

In assignment 4 we created two residual plots, a boxplot and a normal 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?

The R function "influence" calculates important measures we use to evaluate our regression model, including the leverage values ("hat"), the change in coefficient estimate when the ith case is excluded ("coefficients"), the standard deviation of the residuals (externally standardized, "sigma"), and the weighted residuals (which we don't cover in this class). For example, the leverage values of our observations are obtained by (where cdi4.lm is your OLS regression object):

```
cdi4.infl=influence(cdi4.lm)
# first three leverage values
cdi4.infl$hat[1:3]
```

```
##          1          2          3
## 0.05552124 0.01467135 0.05026237
```

C3. (2) Create a vector of the internally studentized residuals (the standardized residuals), using the vector of leverage values that you created above, then produce a boxplot and normal QQ plot of the standardized residuals. Provide the new graphs.

```
# we can create an object that stores the results of the
# summary of the lm object
cdi4.lm.sum=summary(cdi4.lm)
# look at the "attributes" of this new object
attributes(cdi4.lm.sum)
# One of the attributes is "sigma," which is the residual standard error.
# We can use this to calculate our standardized residuals
cdi4.stdRes=cdi4.lm$residuals/(cdi4.lm.sum$sigma*sqrt(1-cdi4.infl$hat))
# The function R standard also calculates the standardized residuals directly
cdi4.stdRes=rstandard(cdi4.lm)
```

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?

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**

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.

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**

```
cdi4.cooks=cooks.distance(cdi4.lm)
# create an index plot of the Cook's distances
par(mfrow=c(1,1),mar=c(3.5,3.5,0.5,0.5),mgp=c(2.5,0.5,0),las=1)
plot(cdi4.cooks,ylab="Cook's distance")
# draw horizontal lines at 1 and 2
abline(h=c(1,2),lwd=2)
```

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

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**.

```
# set up the graphing window
par(mfrow=c(2,2),mar=c(3.5,3.5,1.5,0.5),mgp=c(2.5,0.5,0),las=1)
plot(cdi4.lm)
```

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.

A nice feature of the R residual plot is that it provides the index number (row number) of any potential influential observations. We can then look at those individual observations to help understand why the observation might be influential. Note that the graph flags three points regardless of whether they are problematic. It's up to us to decide if they are!

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**.

```
# This command returns the 5th, 6th, and 20th rows of the cdi4.df object,
# across all columns. For your assignment substitute the rows corresponding
# to the influential observations.
cdi4.df[c(5,6,20),]
```

```
##    ID      County State LandArea TotalPop TotalSeriousCrime PercHS
## 5  12        King    WA     2126  1507319            124959   88.2
## 6  13 Santa_Clara    CA     1291  1497577             77009   82.0
## 20 77      Pierce    WA     1676   586203             41980   83.2
##    PercBach PercPov PercUnemp PerCapInc TotalInc Region   PopDens
## 5      32.8     5.0       4.6     23779    35843      4  708.9929
## 6      32.6     5.0       5.5     25193    37728      4 1160.0132
## 20     17.5     8.7       6.4     16194     9493      4  349.7631
```

```
##    PerThousCrimes
## 5        82.90150
## 6        51.42240
## 20       71.61342
```

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.