STEVE G MWANGI
R LAB 4
02/10/2020

TMATH 410 Computer Assignment 4
Lectures 6-8, Chapter 3
Data
For this assignment we will 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)

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

Submitted

C2. Explain why we would want to express the total values into those per capita, and the population value into population density.

By using the population density and crimes/1000 people, we are creating a standard way to make comparisons for different counties with different populations and land area. The process creates a standardized set of values that can be compared.

C3 (2). For this data set, we will be using per capita personal income, population density, and percentage of population with high school diplomas to predict the per 1000 serious crimes. Explain why we might choose the crime rate as a response variable. Also describe what relationships with any you might expect between per capita serious crimes and the three predictor variables we have named here.

The crime rate makes a better response variable since it can be said to be the effect of some overlying cause in the county, while the other values (per capita personal income, population density, and percentage of population with HS diplomas) are generally defining the Counties. In short, the predictors' causes are hard to specify or pinpoint (e.g. population density defines the county, but cannot be changed by taking land area and increasing/decreasing it), but for Crime rate it can easily be hypothesized to be affected by the other variables, and an effect resulting from them, and not the other way around.

C4 (3). Just for the Western region, create the following scatterplots (publication-quality!). You can produce a graphic with 6 total panels (2 rows, 3 columns) by using the par command: par(mfrow=c(2,3)). Copy the graphs into your document

a) X=per capita personal income (PerCapInc), Y= crime rate (PerThousCrime)
b) X=population density, Y=crime rate
c) X=percentage of population with high school diploma (PercHS), Y=crime rate
d) per capita personal income, population density
e) pre capita personal income, percentage of population with high school diploma
f) population density, percentage of population with high school diploma

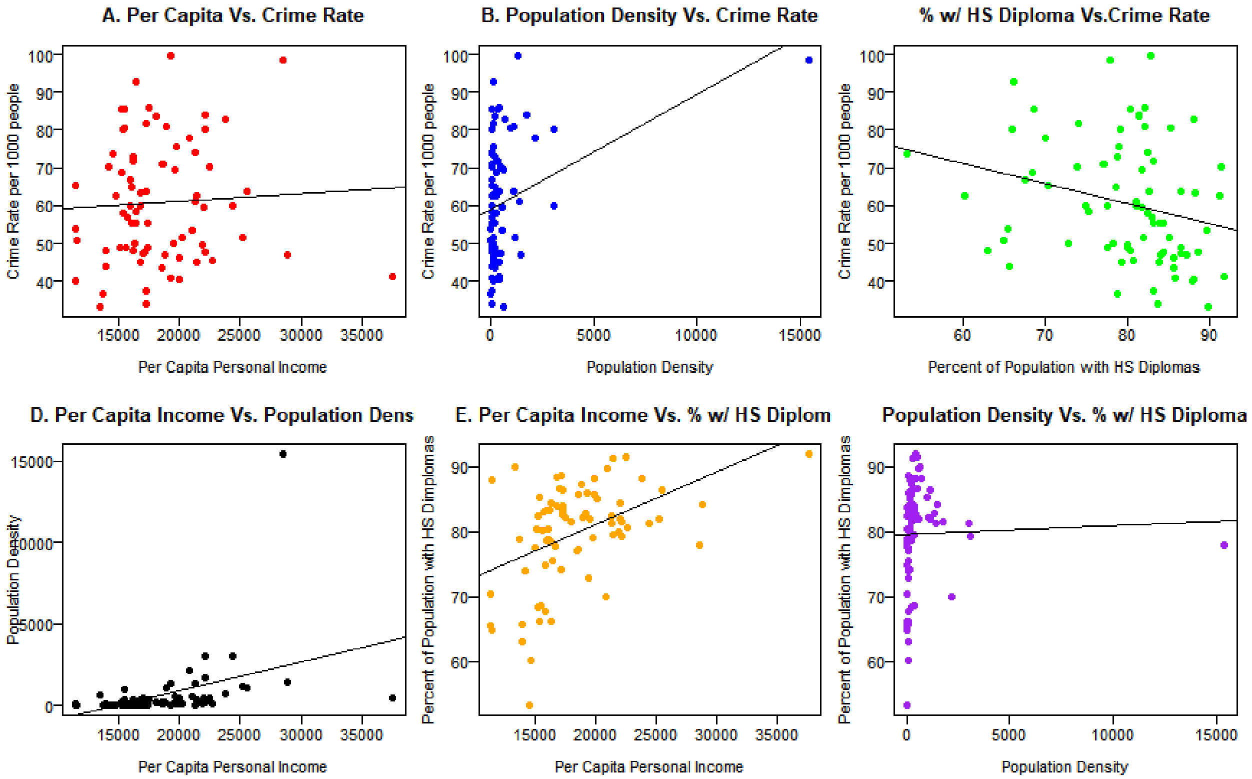

Figure 1 Scatter Plots of the Western Regions with Labelled Axes on all 6, and their corresponding fitted lines

C5. (2) Comment on any patterns you see in the graphs in C4. In particular, is there anything in these plots that you anticipate might complicate your regression analysis?

Plots B, D and F have a point that is off the line and some of the points have high leverage, considering how far they are from where X-bar should be.

Plots A, C, and E have a somewhat reasonable distribution for use with least squares estimate. Additionally, in D, aside from one point that has great influence at about point X =28000, Y = 15000, the other points could be used in the least squares regression.

Plots B and F have too many points that might not necessarily be unique if the effects of some of the high leverage points at 15000 are removed. Might pose a *collinearity problem*.

C6 (3) Just for the Western region estimate the multiple regression model with the number of serious crimes predicted by a linear combination of per-capita income, population density, and percentage of population with high school diplomas. Present your results in a table that follows the format below. $\mathbf{T}_{0.025,\ 73}$ = 1.992997

Coefficient	Estimate	Standard Error	P-value	Lower 95% Confidence Bound	Upper 95% Confidence Bound
Intercept	**1.030e+02**	**1.705e+01**	**5.97e-08**	**69.0194**	**136.9806**
Per-Capita Income	**1.772e-04**	**4.875e-04**	**0.71723**	**-7.9439e-4**	**0.0011**
Population Density	**2.932e-03**	**1.030e-03**	**0.00575**	**8.7921e-4**	**0.0050**
Percent High School	**-5.912e-01**	**2.416e-01**	**0.01680**	**-1.0727**	**-01097**

Table 1Western Region estimated multiple regression values

C7 (3) Provide an interpretation for each regression coefficient estimated in C6. In your interpretation include the estimate for the coefficient and the context of the data and consider whether the coefficient is statistically significant.

Using the P-values and the confidence intervals we see that:

For the intercept, population density and the percent with high school diplomas, the *p-values are less than α* in which case we reject the null hypothesis that the estimates are 0. Only the p-value for the Per-Capita income has a p-value that lets us fail to reject the null hypothesis that the coefficient's true value could be 0. This is further validated by the 95% Confidence interval estimates which show that for all the other values other than that of the Per-Capita Income, their upper and lower bounds are regions that do not include 0, while that of the Per-Capita Income includes the 0 value in its interval; as per the hypothesis testing results. Given the 95% confidence intervals for the Intercept, Population Density and Percent with High School Diplomas excluding the value 0, we see that they are statistically significant; given also that their p-values are quite small, showing that the probability of observing the difference if no difference exists is rather minute, and thus there's higher probability that it's not just randomness affecting the data.

C8. (2) Report the values for MSE, R^2, and the adjusted R^2. Interpret the values and explain why they're different. Remember, the residual standard error given in R is the √MS

MSE = 14.6^2 = 213.16. This is the mean of the squared errors or mean of the residual errors squared. It highlights how close the regression line is close to the data points. In this case, relative to the units of the plot, we see 213.16 means that the points are close.

R^2 = 0.1938. This value signifies the proportion of the variance of the response variables that is predictable based on the explanatory variable; it assesses the quality of the fit. In this case, 19.38% of the total variation in the crime rate can be accounted for by the 3 predictors used in C6 above.

***Adjusted* R^2 = 0.1606. Also used to judge the quality of the fit. This value adjusts the R^2 value above based on the degrees of freedom and accommodates for changes that are based on additional or a smaller number of predictors. It shows how much the quality of the model adjusts given different number of predictors. In this case, this value is less than the given R^2 as should be. But since we don't have another adjusted *Adjusted* R^2 to compare with, it can't tell us much(other than echo what R-squared implies) by itself.**

C9. (2) Create a boxplot and normal quantile plot of the residuals of your multiple linear regression (publication-quality!). Provide the plot.

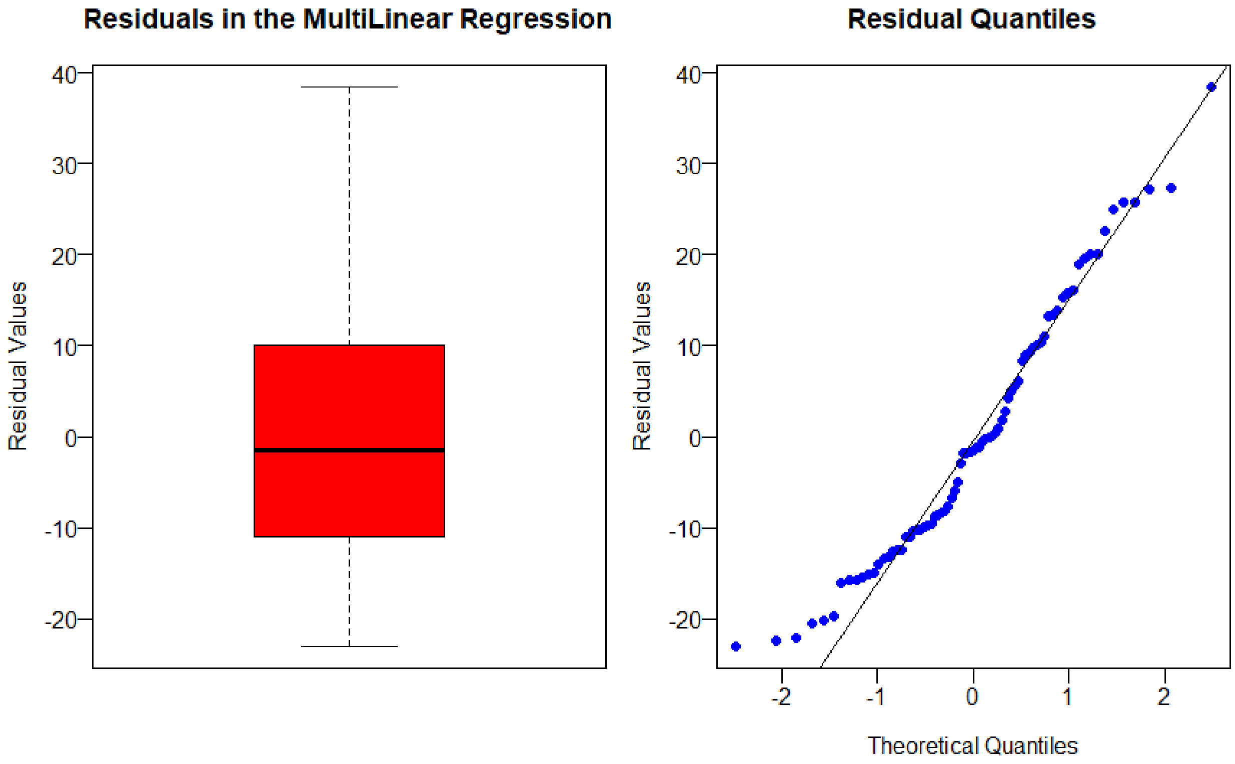

Figure 2 Boxplot and Normal Quantile plot for the residuals in the multiple linear regression

C10. (2) Describe what you see in the graphs in C9. In particular, do you see any patterns that indicate issues with the regression analysis?

Even though the boxplot and the qqplot could be said to be a little right skewed, majority of the points lie on the qqline and the boxplot shows that the distribution of the residuals is about the median value which happens to be at about 0. In short, it seems to good data without overt issues in the regression analysis.

C11 (2). Use the R predict function to find point and interval estimates for each combination of predictor variables. Make sure that you consider whether a confidence or prediction interval is appropriate for this question. Report the values for your point and interval predictions in a publication-quality table.

Fitted Value	**Lower CI**	**Upper CI**
63.2167775035635	55.7840650069002	70.6494900002267
59.0689179036159	55.393792487241	62.7440433199907
95.9270498600945	75.6335522178702	116.220547502319

Figure 3 The Fitted Values and the Confidence Intervals

Fitted Value	**Lower PI**	**Upper PI**
63.2167775035635	33.1780568009892	93.2554982061377
59.0689179036159	29.7331705428396	88.4046652643922
95.9270498600945	60.4459905052193	131.40810921497

Figure 4 The Fitted Values and the Prediction Intervals

Predictors	**Min**	**1st Qu**	**Median**	**Mean**	**3rd Qu.**	**Max**
PerThousCrimes	33.41	48.13	59.66	60.88	71.61	99.7
PopDens	13.26	66.75	186.65	605.44	442.49	15403.38
PercHS	53.20	77.30	81.50	79.69	84.40	91.90
PerCapInc	11379	15701	17268	18323	20786	37541

Figure 5 Summaries of the Observations used in the model

For the 2nd fitted value, the predicted value is not appropriate considering that the observations, the Population Density starts at a minimum of 13.26, while the value used in the prediction is 10, which is less. Overall the Confidence Interval is appropriate, but not the Prediction due to the population density value being outside the bounds of the observed data values.

C12 (3). Write 1-2 paragraphs discussing the model you have estimated for this assignment.

Population Density is the most significant predictor, followed by the Percentage with High School Diplomas then finally we have Per Capita Income. This can be seen in the adjusted R-squared values for the corresponding points when they are removed from the model:

PopDens Removed:
***Multiple R-squared: 0.1044,	Adjusted R-squared: 0.08015***

PerCapInc Removed:
Multiple R-squared: 0.1923, Adjusted R-squared: 0.1705

PercHS Removed:
Multiple R-squared: 0.1276, Adjusted R-squared: 0.1041

The direction and strength of the relationship of the two most significant predictors is positive correlation between population density and crime rate, and negative correlation between crime rate and number of people with high school diplomas. The magnitude is about 0.1938.

When it comes to the Per Capita Income being one with low significance, that was unexpected. Since it is easy to assume it correlates with number of people with High School Diplomas. But this could be due to the discrepancy in wealth distribution, with people who are extremely rich, causing high per capita income, making it an as good as chance predictor. For the population density, it makes sense that it would correlate with crime rate, since the more people in a region, the higher the chance that some of them are not law abiding. Also for the people with highschool diplomas, if less people have diplomas, it means that they are likely to earn less, hence be compelled to seek other ways of earning some extra money, by hook or crook.

When it comes to how useful this model would be in telling us about the different factors affecting crime rate, this model highlights that population density is one of the significant factors to consider, together with the highest education reached by the people. For next steps I would recommend we investigate whether the per capita income is related to the poverty levels in the counties, and whether the poverty levels affects crime rate, and by how much if it does. Additionally, I would also recommend that we look into some of the population density factors that are outliers and have high influence and leverage, especially considering some of the values seen in the scatter plots in C4.