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)

On Computer

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
```

C1 (4). Submit the R script you used to conduct your analysis.

Notice in this data set that everything is in total amounts for a given county–total income, total population, total crimes. First we will create new columns that transform the total crimes to crimes per 1000 people and the total population into population density (people/land area). Use the code below to make the new columns.

```
# make a new column for population density byt dividing
# the total population by the land area
cdi.df$PopDens=cdi.df$TotalPop/cdi.df$LandArea
# make a new column for crimes per 1000 people by dividing
# the total serious crimes by the total population and then
# multiplying by 1000
cdi.df$PerThousCrimes=cdi.df$TotalSeriousCrime/cdi.df$TotalPop*1000
```

C2 (2). Explain why we would want to express the total values into those per capita, and the population value into population density.

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.

For this analysis we will focus on the Western region (4)

We can create a new dataframe that is only the Western Region by using the code below.

```
# the qualifier inside the brackets tells R to only keep
# the rows that satisfy the condition indicated
cdi4.df=cdi.df[cdi.df$Region==4,]
```

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

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?

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.

Coefficient	Estimate	Standard error	p-value	Lower 95% confidence bound	Upper 95% confidence bound
Intercept					
Per-capita Income					
Population Density					
Percent High School					

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.

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 $\sqrt{MS}$

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

```
# example residual boxplot, assumes your lm object is called cdi4.lm
boxplot(cdi4.lm$residuals,ylab="Residual values",col="grey")
```

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?

You are interested in the expected (or mean) per thousand serious crime rate in three different counties with the following characteristics:

PopDens	PercHS	PerCapInc
900	75	11000
10	80	19000
10000	72	35000

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**

```
# Here is how you structure your data frame
# for the predict function:
newdata=data.frame(PopDens=c(900,10,10000),PercHS=c(75,80,72),PerCapInc=c(110
00,19000,35000))
# have a look at Computer Assignment 3 for the syntax for predictions
```

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

In your discussion make sure to include the following (organized into a narrative, paragraph structure, not just bullet points that respond to the questions):

Which variables were significant predictors and which were not?

For the significant predictors, what were the direction and strength of the relationship?

Did the set of significant predictors and the direction of the relationships make sense relative to what you expected? Why or why not?

Did the set of non-significant predictors make sense relative to what you expected? Why or why not?

How useful do you think this model would be for predicting crime rate in a county? Explain your reasoning and give justification from the model statistics and graphics.

How useful do you think this model would be for understanding different factors that affect crime rate? Explain your reasoning and give justification from the model statistics and graphics.

what are next steps you would recommend to increase our understanding further?