Chapters covered : Ch 3, 12

High school graduates and Crime rates

In R, there are many packages; a package bundles R code, data, documentation, and test, is easy to share with others. In this problem we will use a data set named `'crime2005'` available through a R package named `'smss'` (Statistical Methods For The Social Sciences).
First to install the package and obtain the data set, run the following code:

```
install.packages("smss")
library("smss") #runs the package
data(crime2005)
View(crime2005)
```

There are 51 observations (50 states and Washington D.C.) and 8 variables. We are considering a linear regression model between the following variables:

- `HS`: Percent high school graduates in a given state
- `VI2`: violent crime rate (number of violent crimes per 10,000 population in a given state

where `HS` is the explanatory variable and `VI2` is the response variable.
Answer the following questions:

a) Use `summary(lm(crime2005$VI2 ~ crime2005$HS))` command to fit the linear regression model of `HS` and `VI2`.
State the estimate regression equation in the form of $\widehat{VI2} = a + b(HS)$. Interpret both slope and y-intercept.

b) State and interpret the value of r^2 from the summary output from the previous part.

c) Calculate the correlation r between `VI2` and `HS`, and state the strength and the direction of the correlation.

d) Use `plot(y~x)` command to make a scatter plot.

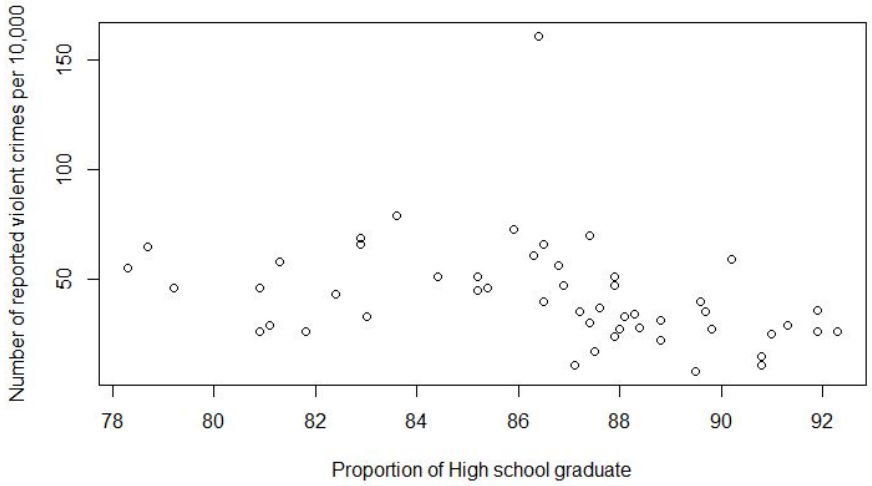
e) Washington DC has the largest violent crime rate in the data set. In the scatter plot, a dot with largest crime rate with 85% HS graduates indicates Washington DC. How would r^2 (the coefficient of determination) change if Washing DC is removed(increase, decrease, or remain the same)? Explain why.

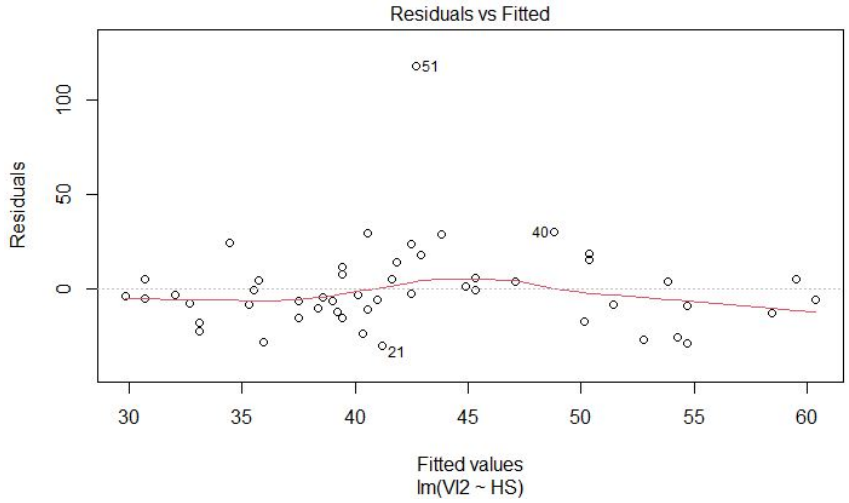
f) Minnesota's percent of high school graduates is 92.3. Find the predicted crime rate of Minnesota. What is the residual of Minnesota?

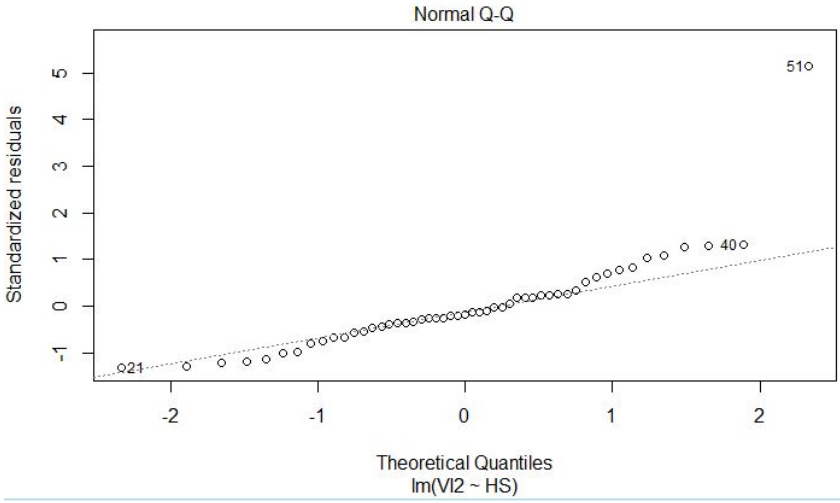
g) Check the linearity, normal error, and constant variance assumptions using the plots below.

h) Construct a 95% confidence interval to estimate β (slope of population regression model). Interpret the result.

i) Conduct a five step hypothesis test on whether the true slope β is different from 0. Use $\alpha = 0.05$. Remember to check assumptions and interpret the conclusion in the context of the problem.


Number of reported violent crimes per 10,000
50
100
150
78
80
82
84
86
88
90
92
Proportion of High school graduate


Residuals vs Fitted
51
40
21
Residuals
0
50
100
30
35
40
45
50
55
60
Fitted values
lm(VI2 ~ HS)


Normal Q-Q
51
40
21
Standardized residuals
-1
0
1
2
3
4
5
-2
-1
0
1
2
Theoretical Quantiles
lm(VI2 ~ HS)