**Problem 4 - (10 points)**

The “calif” SAS dataset lists the populations of Californian cities along with the percentage of the population that are: under 18 (pct\_under\_18), between 18 and 64 (pct\_18\_64) and over 64 (pct\_over).

Run the following code snippet to create new variables ‘log\_population’ and ‘log\_pct\_over’.

data calif\_subset;

set calif ;

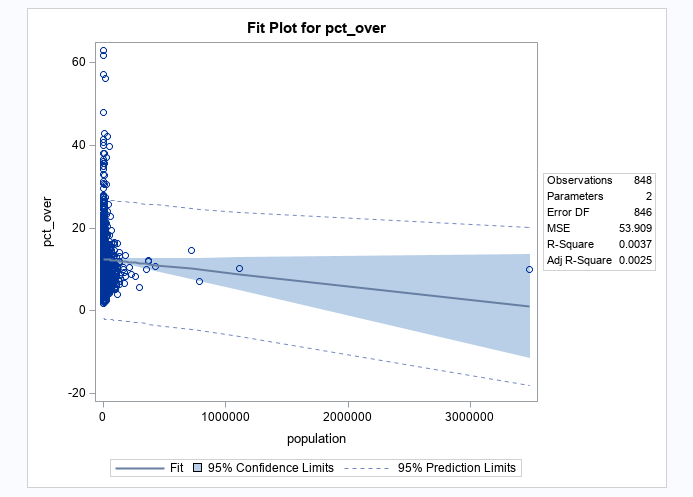
Log\_population=log(population);

Log\_Pct\_over=log(pct\_over);

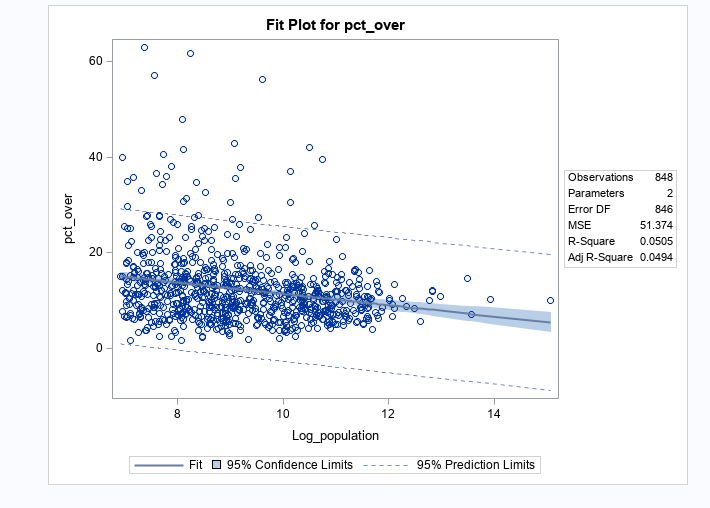
run;

Establish the following three regression models:

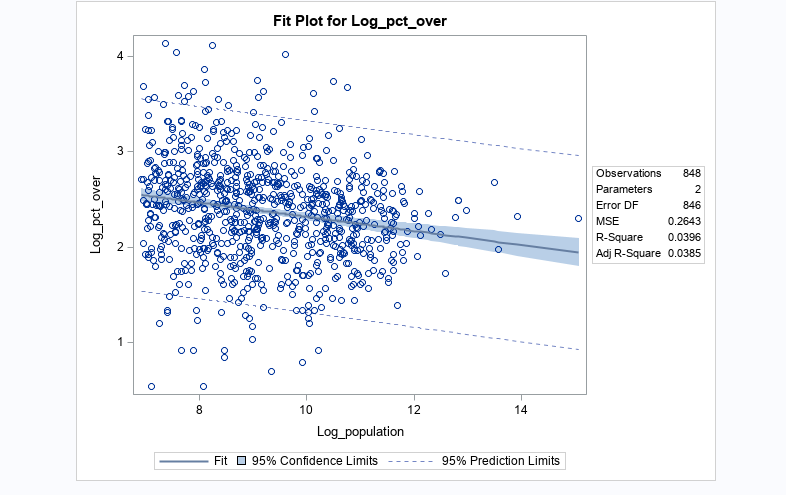
* Pct\_over on Population



* Pct\_over on Log\_Population



* Log\_pct\_over on Log\_population



Which model works the best? Why?

Ans: **The third model is better than the rest since the residuals are better than the second and even thought eh R-Square values is less than the 2nd model. Since we have taken the Log transformation at both the sides the data is normalized which is easy to interpret since there are outliers but we can make the model work relatively well then the rest of the models.**