(1) Perform exploratory analysis on all variables.

See “Normal distribute plot for all variables”.

(2) Perform simple regression analysis of percent over 64 (pct_over) on the “population” variable (or transformed form of Population).

See “Simple regression for pct_over vs. population” and “Simple regression for pct_over vs. log_population”.

(3) Would you apply any transformation to the “population” variable? What kind of transformation if any?

Log transformation. We can transform the population into log(population).

(4) Is the regression line a good predictor? Why or why not?

For no transformation regression, the regression line is not a good predictor, because the “Pr>F” is 0.0778, which is not smaller than 0.05. And for log transformation regression, the regression line is a good predictor, because the “Pr>F” is <0.0001, and the “Pr>|t|” for intercept and log_population are also <0.0001. These are all highly significant.

Simple regression for pct_over vs. population

The REG Procedure
Model: MODEL1
Dependent Variable: pct_over

Number of Observations Read	848
Number of Observations Used	848

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	168.09475	168.09475	3.12	0.0778
Error	846	45607	53.90946		
Corrected Total	847	45775			

Root MSE	7.34231	R-Square	0.0037
Dependent Mean	12.38113	Adj R-Sq	0.0025
Coeff Var	59.30238		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	12.48476	0.25887	48.23	<.0001
population	1	-0.00000327	0.00000185	-1.77	0.0778

Simple regression for pct_over vs. log_population

The REG Procedure
Model: MODEL1
Dependent Variable: pct_over

Number of Observations Read	848
Number of Observations Used	848

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2313.34102	2313.34102	45.03	<.0001
Error	846	43462	51.37371		
Corrected Total	847	45775			

Root MSE	7.16755	R-Square	0.0505
Dependent Mean	12.38113	Adj R-Sq	0.0494
Coeff Var	57.89087		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	23.26150	1.63999	14.18	<.0001
log_population	1	-1.18193	0.17613	-6.71	<.0001

(5) What are high leverage and influential observations?

Leverage(log transformation):

The highest five leverage points are shown in the following chart, which are 1, 2, 3, 4 and 5.

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0.00117927	381	0.00973781	5
0.00117931	382	0.01227717	4
0.00117934	383	0.01268172	3
0.00117943	384	0.01460307	2
0.00117945	385	0.02190575	1

Influential observation(log transformation):

Cookd and Dffits:

The highest five cookd and dffits points are shown in the following charts. Point 772, 736, 595, 844 and 320 are both in these two graphs, which means these points have the large influential effect on the line.

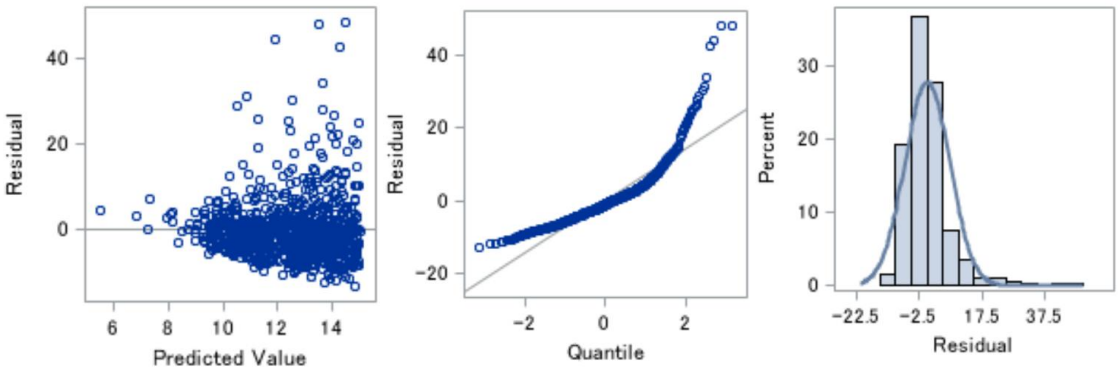
Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1.02508E-11	579	0.0244816	320
8.39650E-11	812	0.0255629	844
3.70796E-10	364	0.0392509	595
4.32523E-10	333	0.0498560	736
4.40982E-09	495	0.0734025	772

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
-1.147E-01	820	0.226321	320
-9.421E-02	796	0.227609	844
-8.494E-02	768	0.287788	595
-8.404E-02	767	0.322436	736
-8.374E-02	715	0.393696	772

(6) Are the residuals normally distributed? Why?

No. The residuals are not normally distributed.

Because of the second graph that these points are not distributed like a line, the residuals are not normally distributed.


Residual
40
20
0
6
8
10
12
14
Predicted Value
Residual
40
20
0
−20
−2
0
2
Quantile
Percent
30
20
10
0
−22.5
−2.5
17.5
37.5
Residual