MSCI609 – First steps in data analysis (for group project)

For the course:

1. You should check your raw data for normality.
2. Use the Kolmolgorov-Smirnoff test (it is a nonparametric test).

H_0: the empirical distribution follows the normal distribution

H_1: the empirical distribution is non normal

Formula from (https://www.easycalculation.com/formulas/kolmogorov-smirnov-formula.html)

$$D = \max_{1 \le i \le N} \left(F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i) \right)$$

Where,

D = Maximum Value of Normal Distribution,

N = Numbeformr of Statistic Data,

F = Kolmogorov Smirnov (KS) Index.

3. Do not log transform data before conducting the K-S test. Leave it as raw data.
4. If K-S rejects normality of the raw data then create a frequency distribution.
5. Are there outliers? In other words, are the raw data skewed?
6. If raw data are skewed, log transform the data (x' = ln(x)) and then re-run K-S test.

In practice:

If your data a non-normal, that is OK. Run your OLS regression with the DV and IVs. You need to learn how to make predicted values from your regression function. And then you need to analyze the residuals. Most statistical software can do this for you easily. But I will show you how to do it by hand.

Here is your model: $y = a + bx + e$ (1)

Where: y is number of automobiles purchased per capita per year

x is income per capita per year

a, b are estimated coefficients

e is error term

You have the following data on y and x:

y	x
0.25	15.00
0.19	14.50
0.29	16.80
0.40	20.00
0.42	21.30
0.44	20.10
0.47	22.36
0.38	20.05
0.41	23.05
0.44	24.93

MSCI609 – First steps in data analysis (for group project)
Now perform OLS using the data above to get:
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.91052
R Square	0.829047
Adjusted R Square	0.807678
Standard Error	0.04085
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.065	0.065	38.797	0.000
Residual	8	0.013	0.002		
Total	9	0.078			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-0.1205	0.0796	-1.5134	0.1686
x	0.0247	0.0040	6.2287	0.0003

Your estimated equation is therefore: $y = -0.1205 + 0.0247x + e$ (2)

Where e is the 'error' term or we can call it the 'residual'. We want to analyze the residuals from this regression to see if they are normally distributed. Statistics tells us OLS estimates are BLUE – best linear unbiased estimates. But also, it is *assumed* that the errors (residuals) are normally distributed. Therefore, we need to test if the residuals are indeed normal. We will use the K-S test. How to get the residuals? We use (2) above, and insert the actual x values into the estimated equation to obtain predicted values of y. Then compute: y_actual – y_predicted = residual and use K-S test on the residuals.

x	y_predicted	y_actual	residual
15.00	0.250155	0.25	0.000
14.50	0.237799	0.19	-0.048
16.80	0.294639	0.29	-0.005
20.00	0.373720	0.40	0.026
21.30	0.405847	0.42	0.014
20.10	0.376191	0.44	0.064
22.36	0.432043	0.47	0.038
20.05	0.374956	0.38	0.005
23.05	0.449095	0.41	-0.039
24.93	0.495555	0.44	-0.056