RM01 Research Methods

SPSS Exercises

SPSS Exercises Solutions

(Version: November 2021)

By Helen Bao

Department of Land Economy

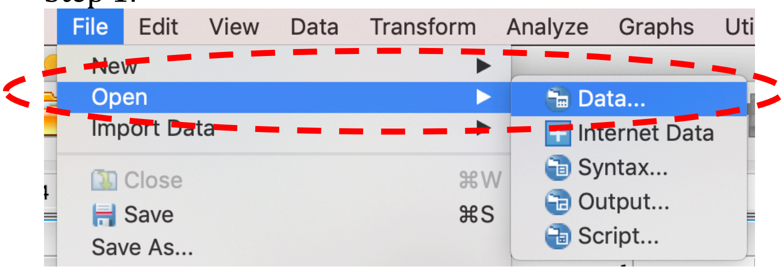
University of Cambridge

SPSS Exercises

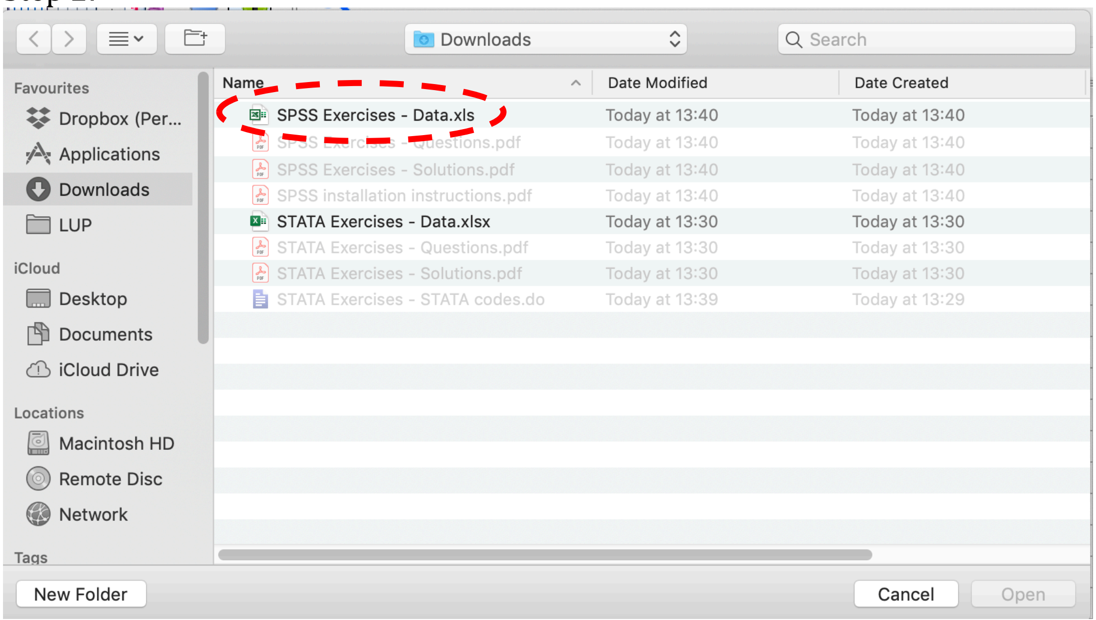
1. Quarterly data on property price index and GDP are given in Worksheet ‘example1’. Variable T is a quarterly time index (e.g., T = 1 for the first quarter). Use SPSS to complete the following tasks.

   a) Open the file in SPSS and view the data

   Step 1:

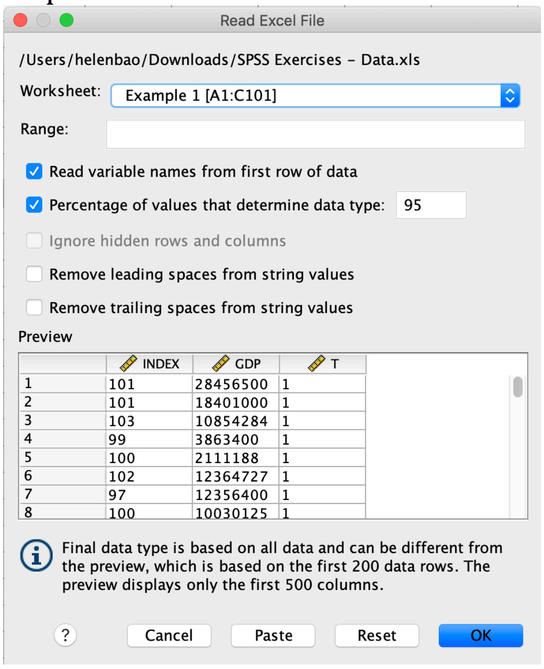


   Step 2:

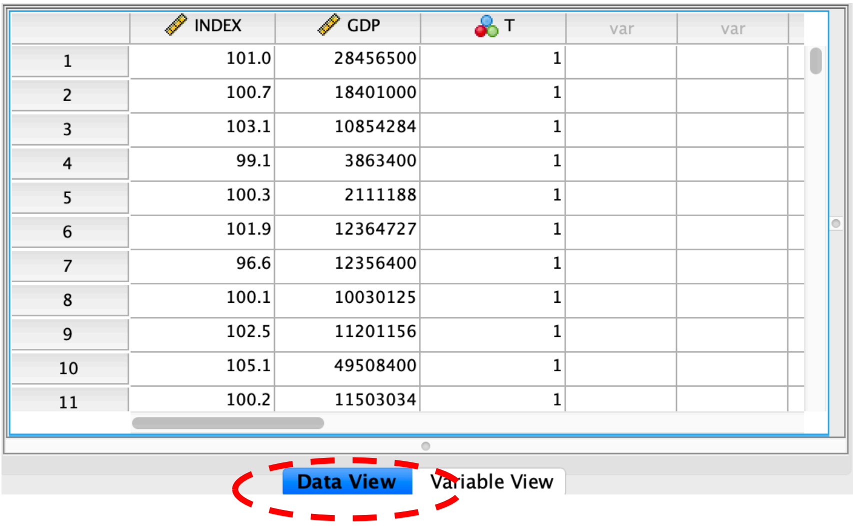


   Note: You may need to change the ‘Files of types’ to be ‘Excel’ in order to make Excel files visible in this window.

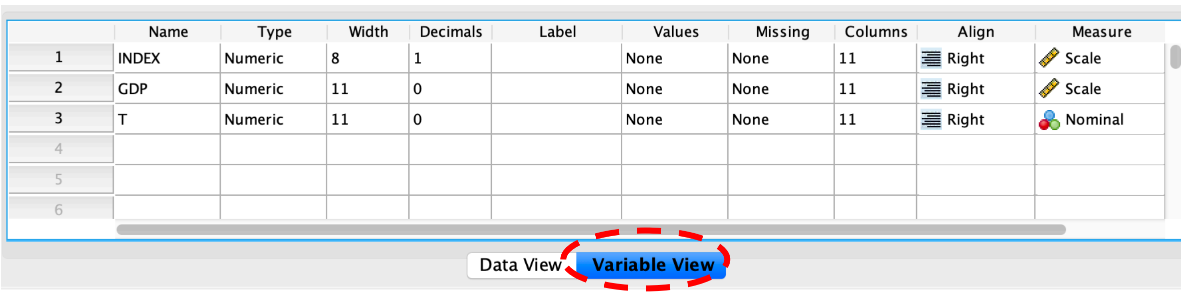
Step 3:



Data view: to view and edit data

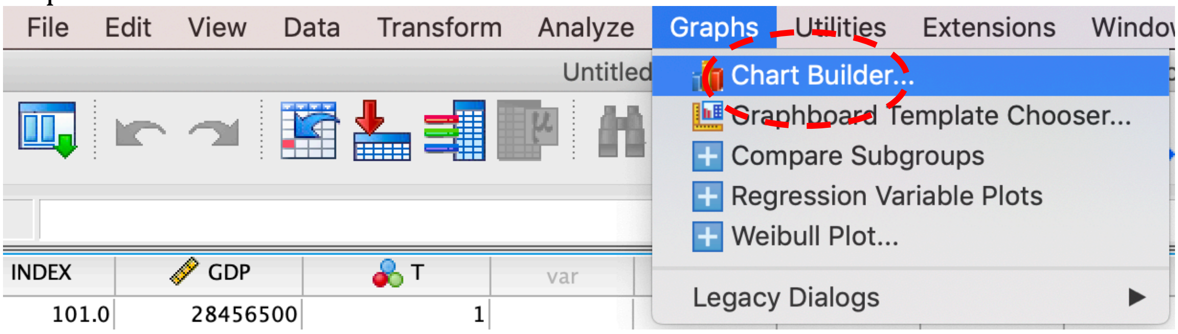


Variable view: to view and edit variable attributes

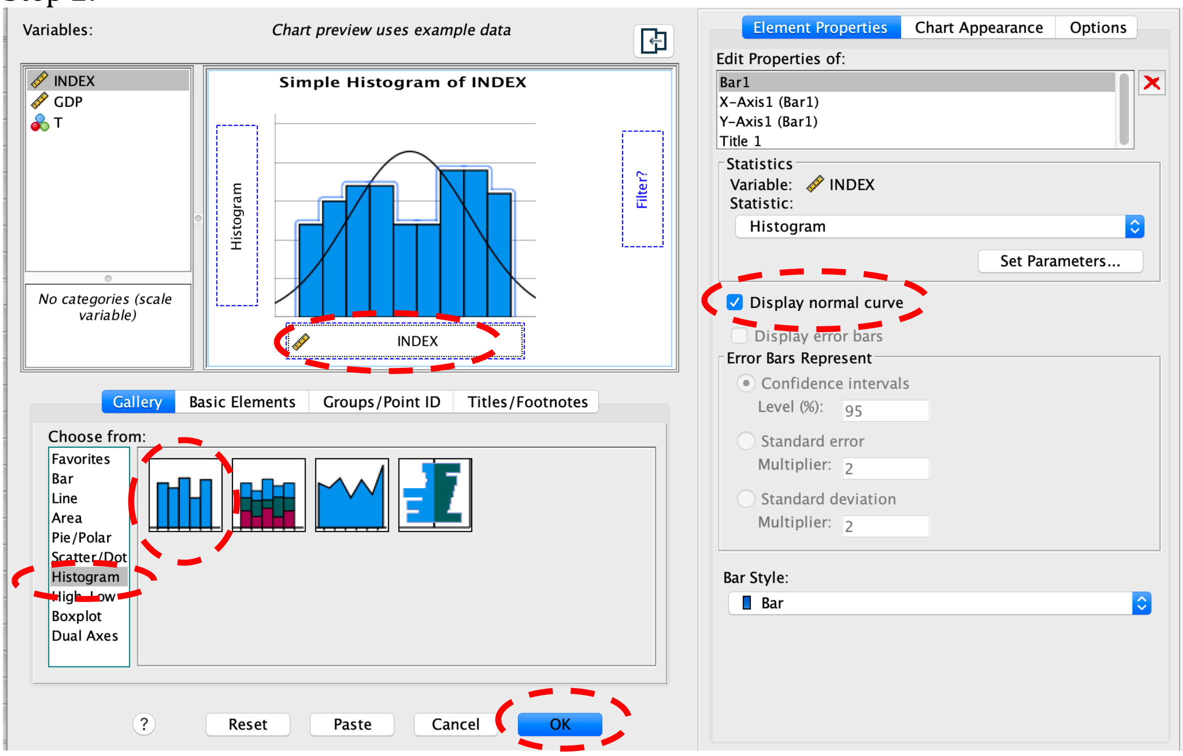


b) Create a histogram of variable INDEX

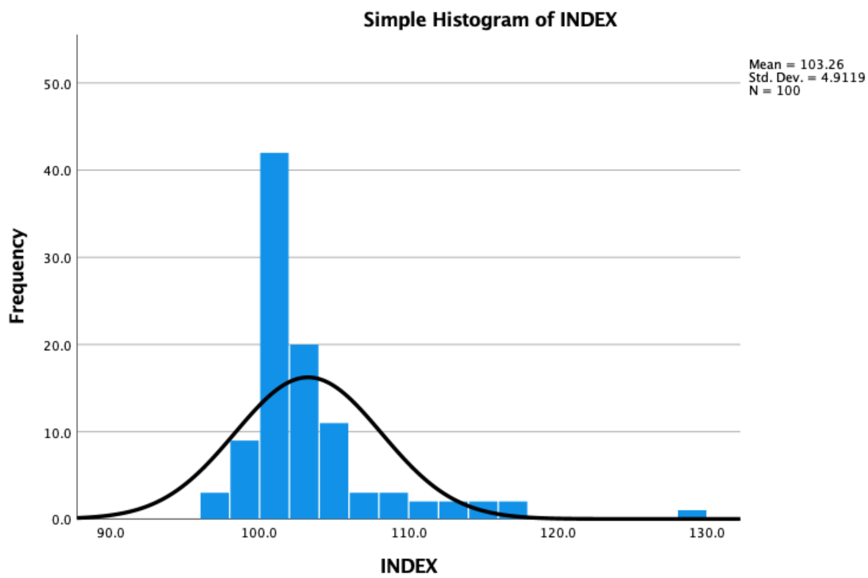
Step 1:



Step 2:

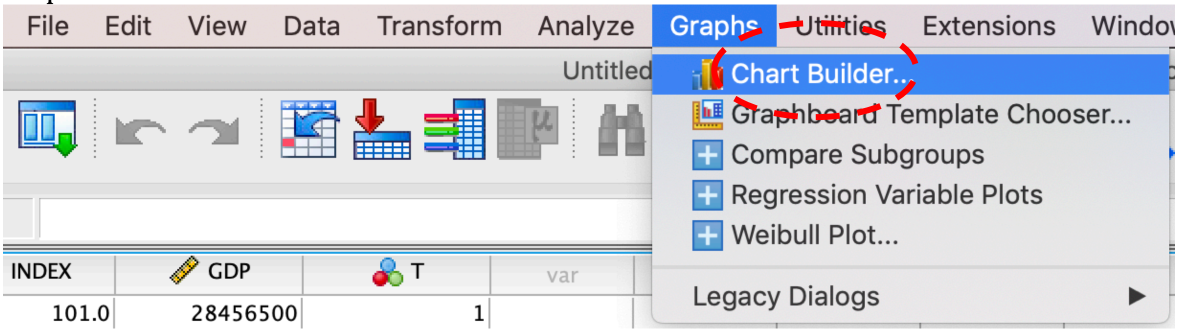


Output:

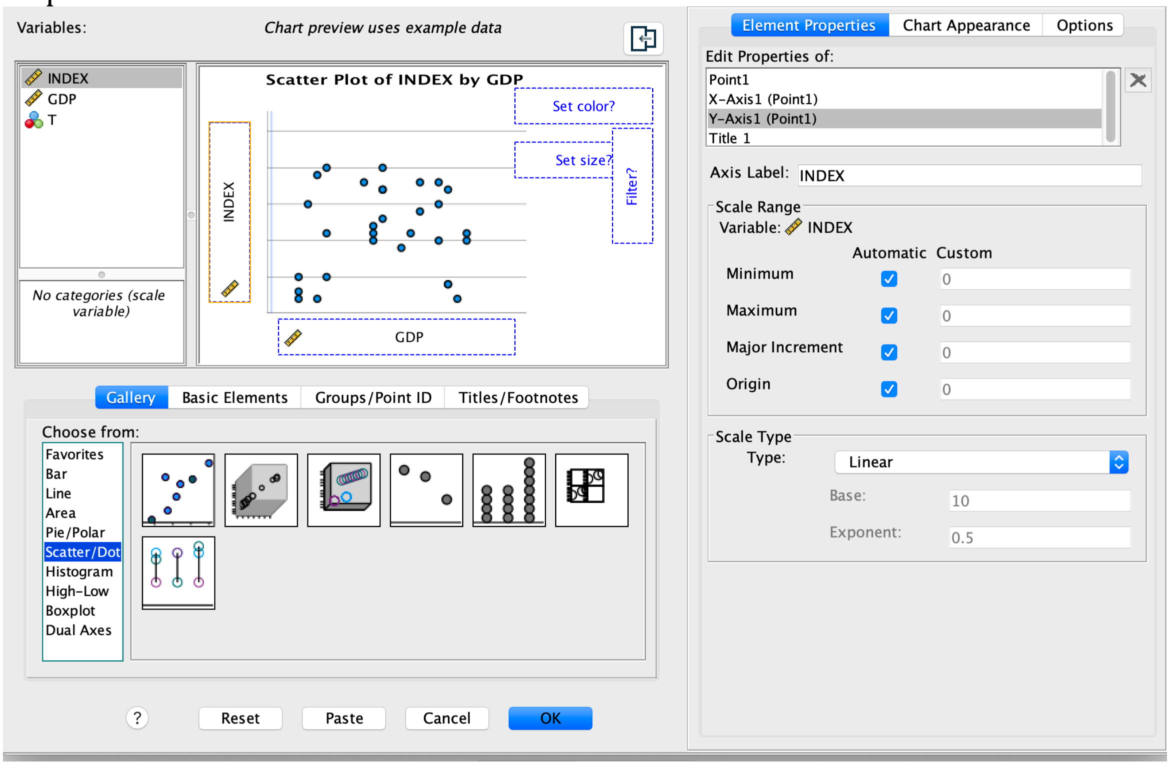


c) Create a scatter plot between INDEX and GDP

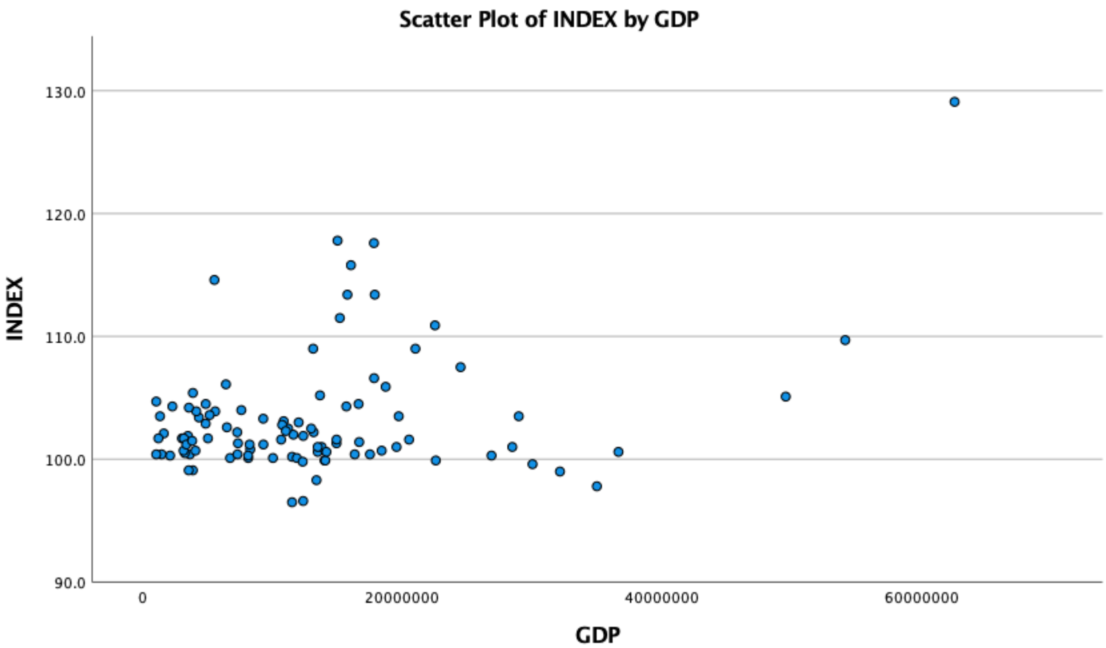
Step 1:



Step 2:

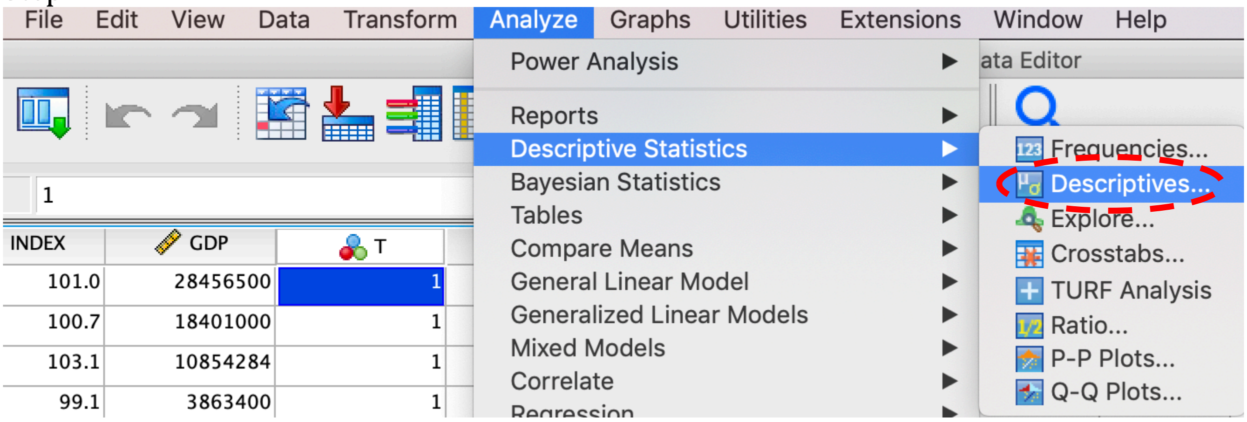


Output:

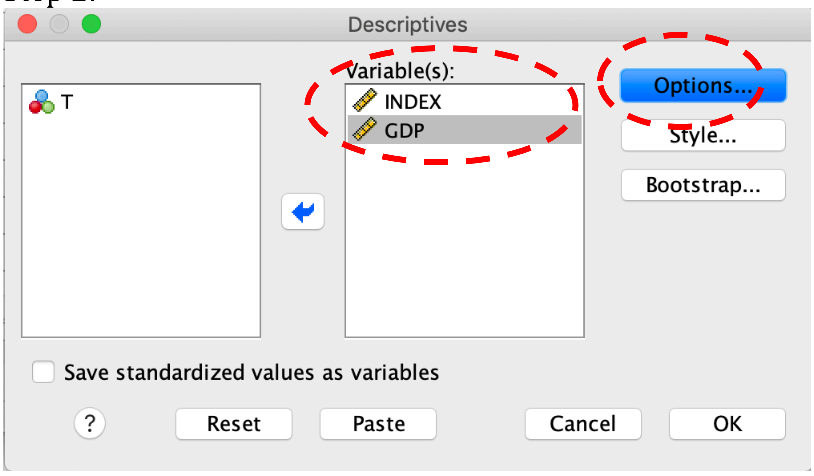


d) Generate descriptive statistics for INDEX and GDP

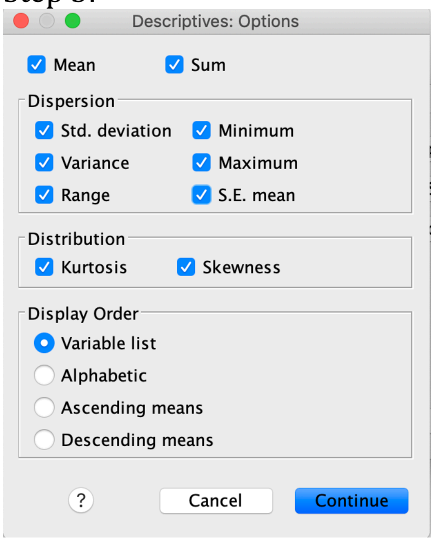
Step 1:



Step 2:



Step 3:



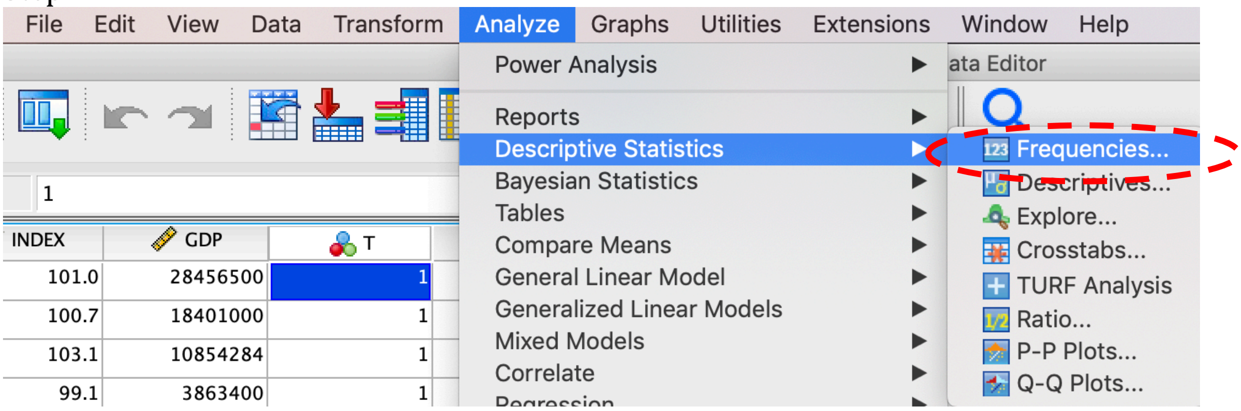
Output:

Descriptive Statistics

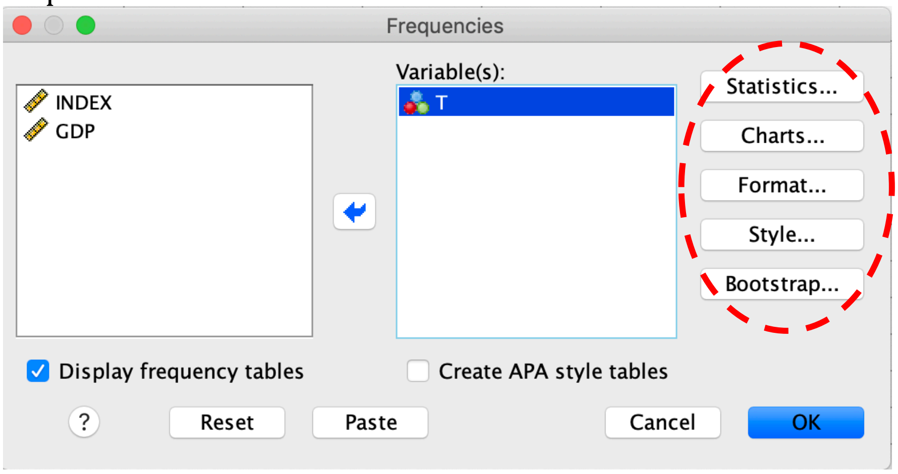
	N	Range	Minimum	Maximum	Sum	Mean		Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
INDEX	100	32.6	96.5	129.1	10326.0	103.260	.4912	4.9119	24.127	2.493	.241	8.277	.478
GDP	100	61463225	1044875	62508100	1.E+9	13067838.9	1090432.76	10904327.6	1.189E+14	2.061	.241	5.885	.478
Valid N (listwise)	100												

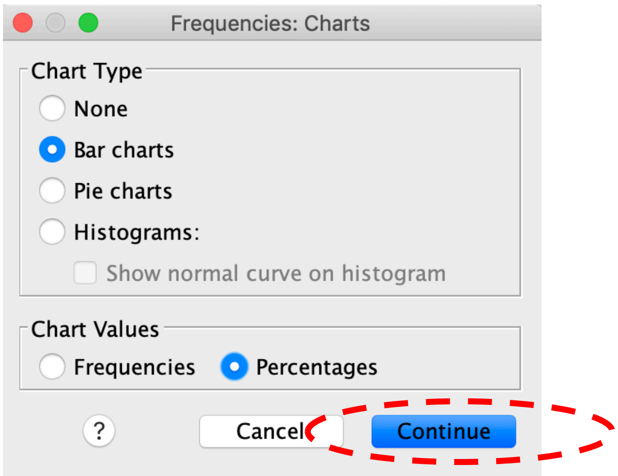
e) Generate frequency statistics for T

Step 1:



Step 2:

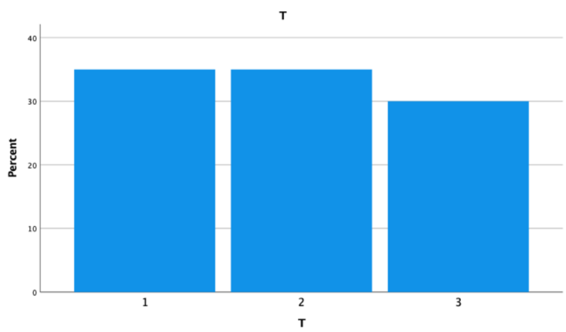




Output:

T

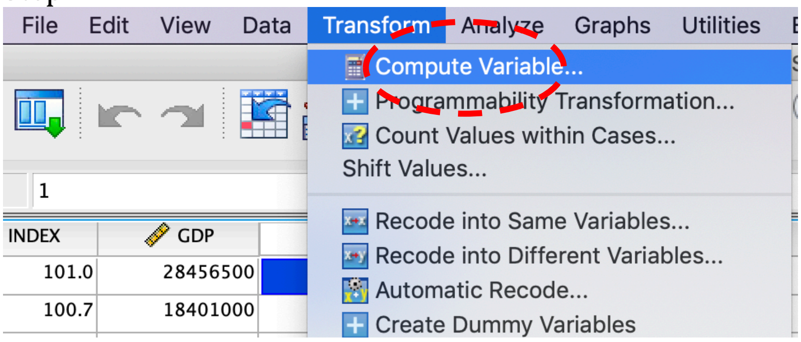
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	35	35.0	35.0	35.0
	2	35	35.0	35.0	70.0
	3	30	30.0	30.0	100.0
	Total	100	100.0	100.0	



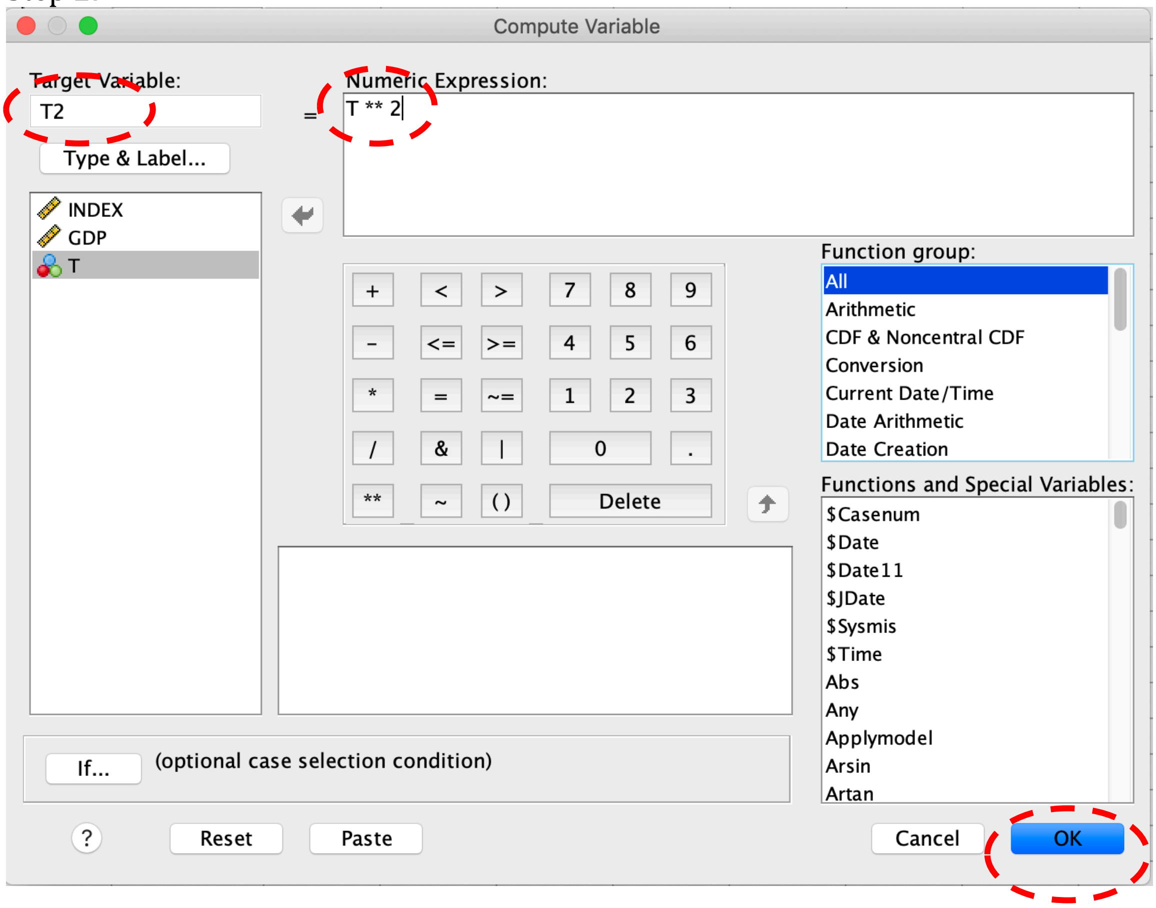
f) Generate two new variables: T2 = T*T, and LNGDP = ln(GDP)

8

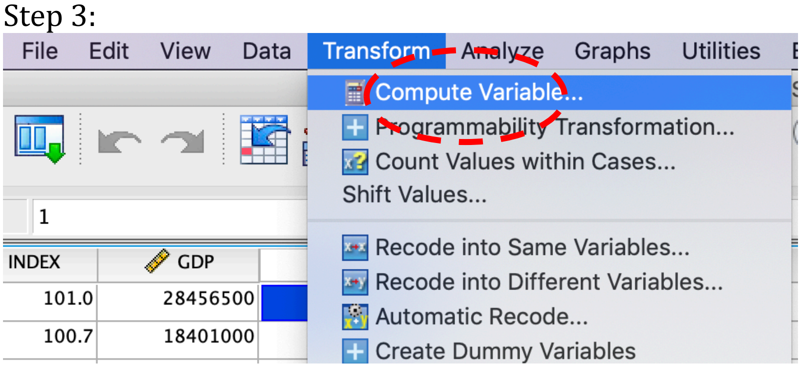
Step 1:



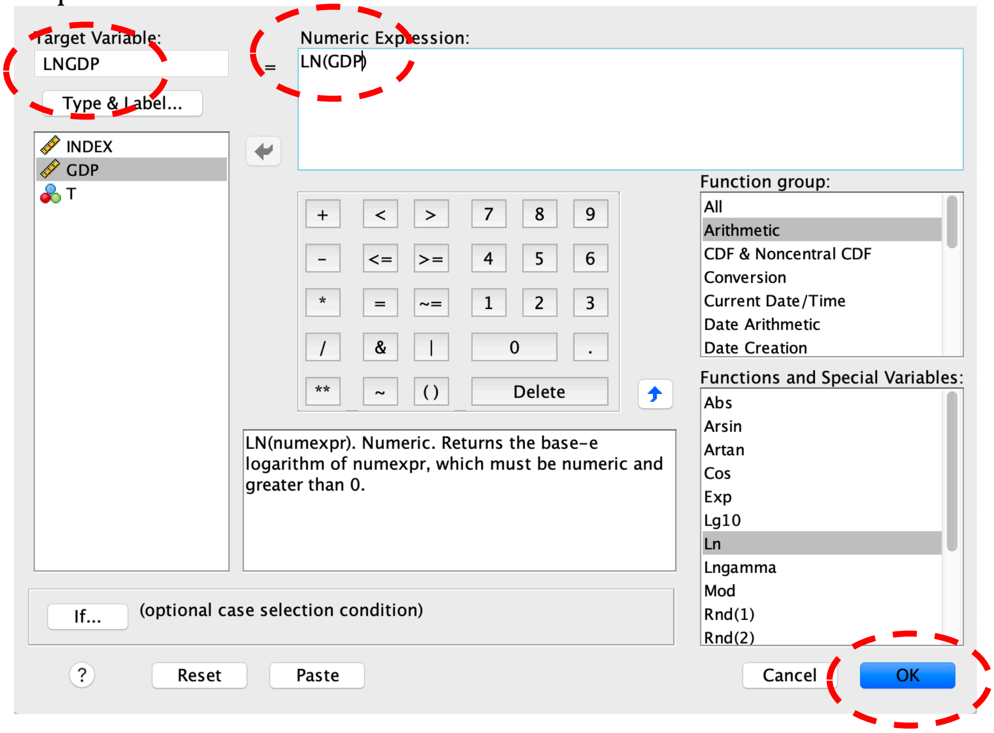
Step 2:



Step 3:

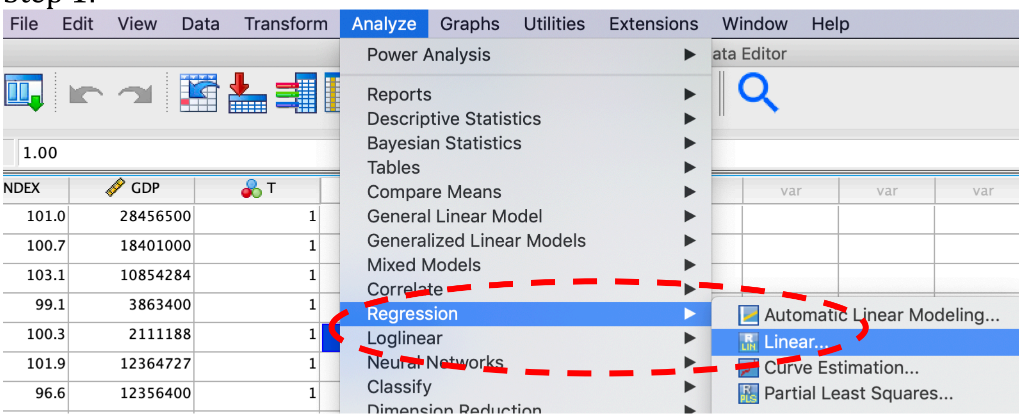


Step 4:

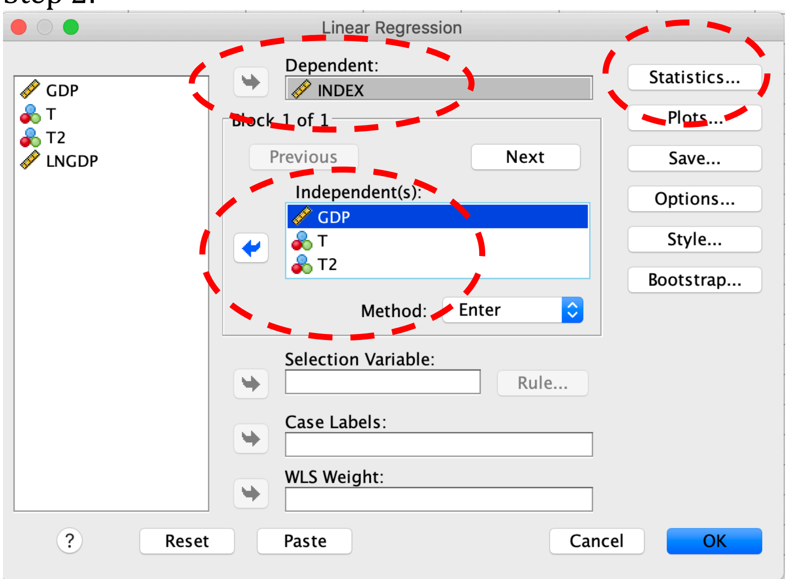


g) Estimate the regression model $INDEX = \beta_0 + \beta_1 GDP + \beta_2 T + \beta_3 T2 + \varepsilon$, where T2 is T squared. Obtain collinearity statistics and autocorrelation test statistics

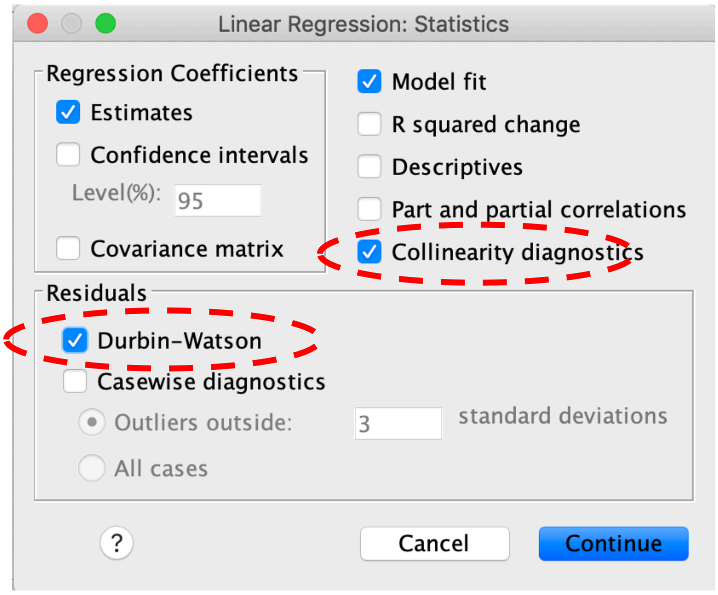
Step 1:



Step 2:



Step 3:



Output:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.439[a]	.192	.167	4.4827	1.596

a. Predictors: (Constant), T2, GDP, T

b. Dependent Variable: INDEX

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	459.489	3	153.163	7.622	.000[b]
	Residual	1929.071	96	20.094		
	Total	2388.560	99			

a. Dependent Variable: INDEX

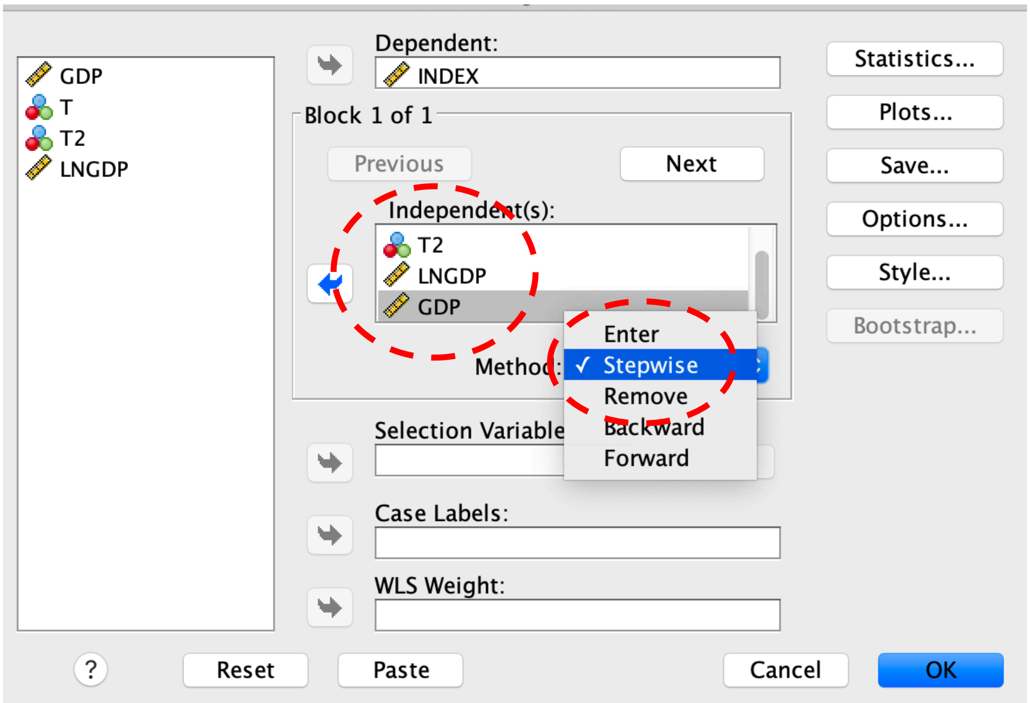
b. Predictors: (Constant), T2, GDP, T

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	99.780	3.356		29.728	.000		
	GDP	1.391E-7	.000	.309	3.301	.001	.961	1.040
	T	-.078	3.780	-.013	-.021	.984	.022	46.049
	T2	.408	.942	.270	.433	.666	.022	46.157

h) Use stepwise selection method to determine the best set of regressors to predict the value of INDEX

Include all variables in the 'independent(s)' list. Choose model selection methods (e.g., backward or forward) from the 'Method..' dropdown menu.



Outputs (Stepwise selection):

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.359[a]	.129	.120	4.6078	
2	.439[b]	.192	.176	4.4595	1.596

a. Predictors: (Constant), GDP

b. Predictors: (Constant), GDP, T2

c. Dependent Variable: INDEX

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	307.798	1	307.798	14.497	.000[b]
	Residual	2080.762	98	21.232		
	Total	2388.560	99			
2	Regression	459.480	2	229.740	11.552	.000[c]
	Residual	1929.080	97	19.887		
	Total	2388.560	99			

a. Dependent Variable: INDEX

b. Predictors: (Constant), GDP

c. Predictors: (Constant), GDP, T2

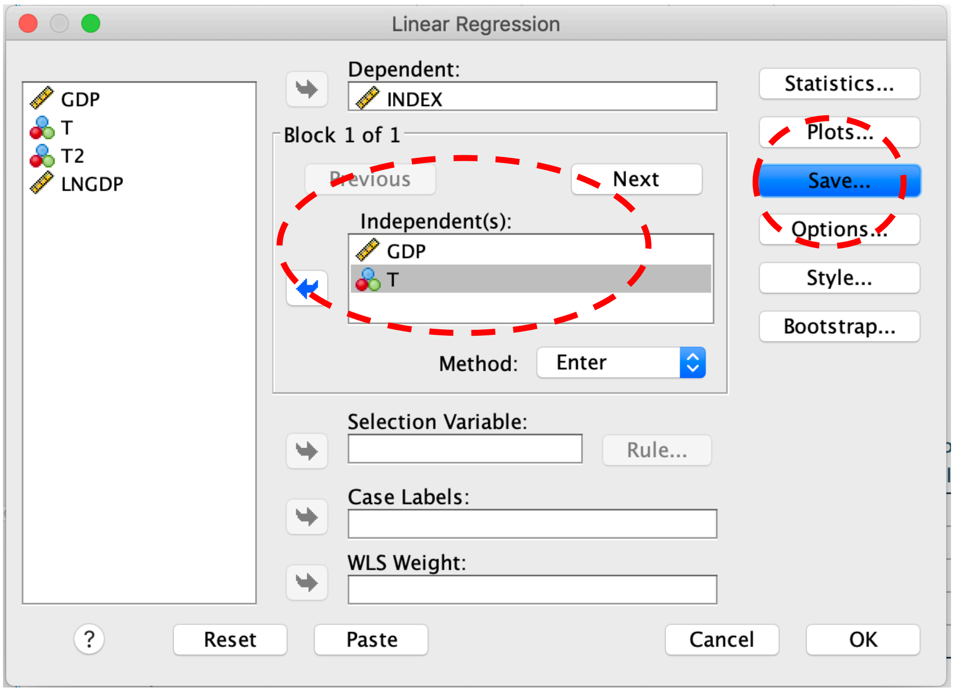
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	101.147	.721		140.220	.000		
	GDP	1.617E-7	.000	.359	3.807	.000	1.000	1.000
2	(Constant)	99.713	.870		114.608	.000		
	GDP	1.391E-7	.000	.309	3.320	.001	.962	1.040
	T2	.388	.141	.257	2.762	.007	.962	1.040

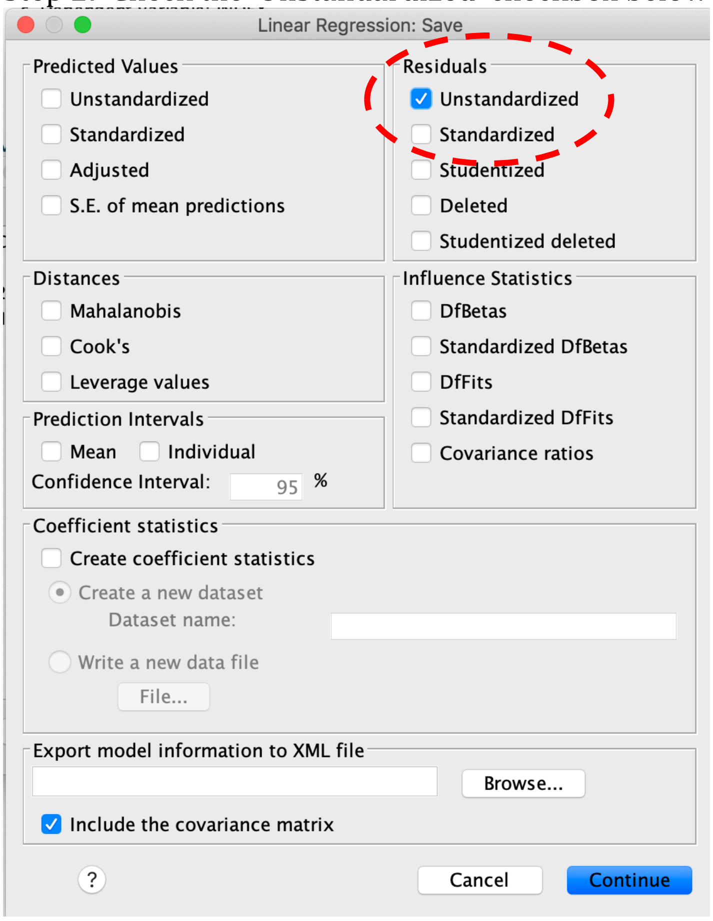
i) Generate a scatter plot between the residuals and GDP for the regression model $INDEX = \beta_0 + \beta_1 GDP + \beta_2 T + \varepsilon$.

12

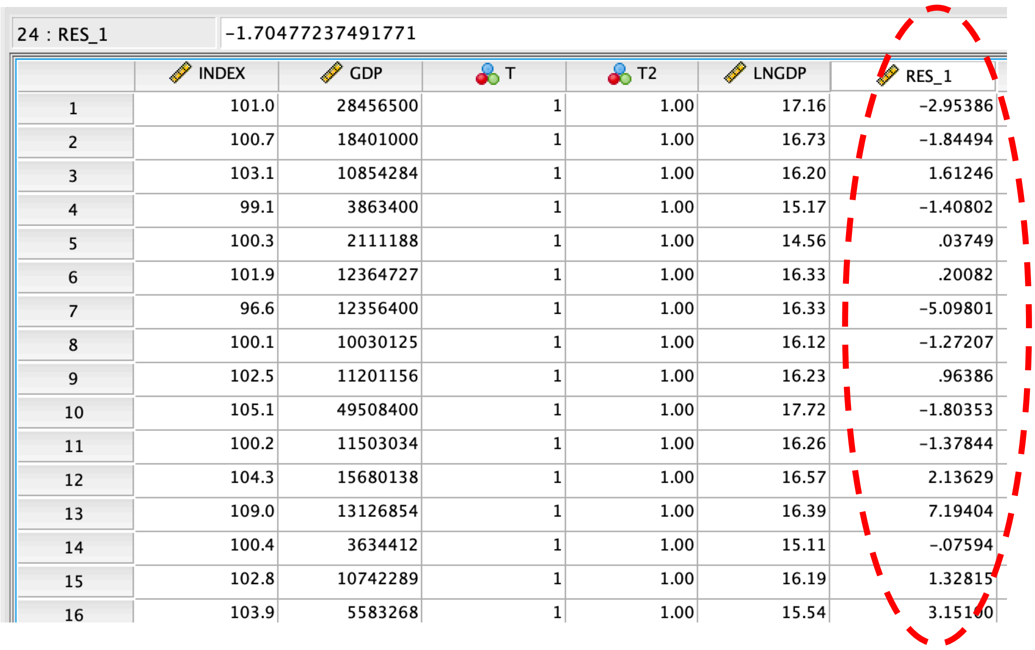
Step 1: Choose GDP and T only as the independent variables, and use 'Enter' method. Click the 'Save' button.



Step 2: Check the 'Unstandardized' checkbox below 'Residuals'. Click 'Continue'.



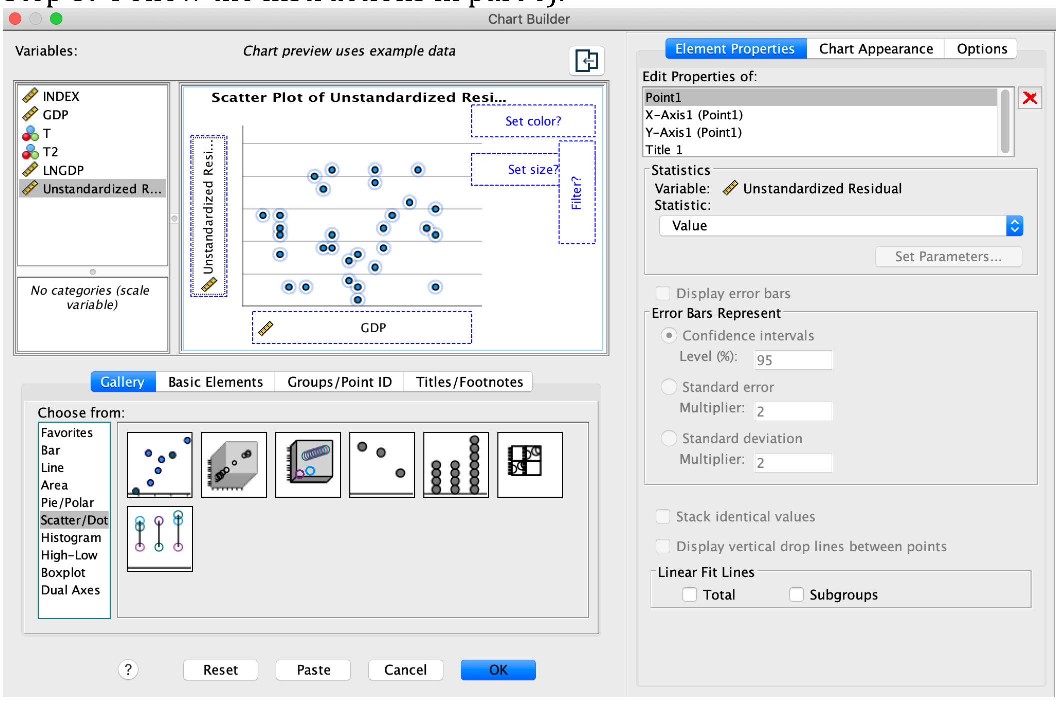
A new variable 'RES_1' is created to store the residuals of your regression model.



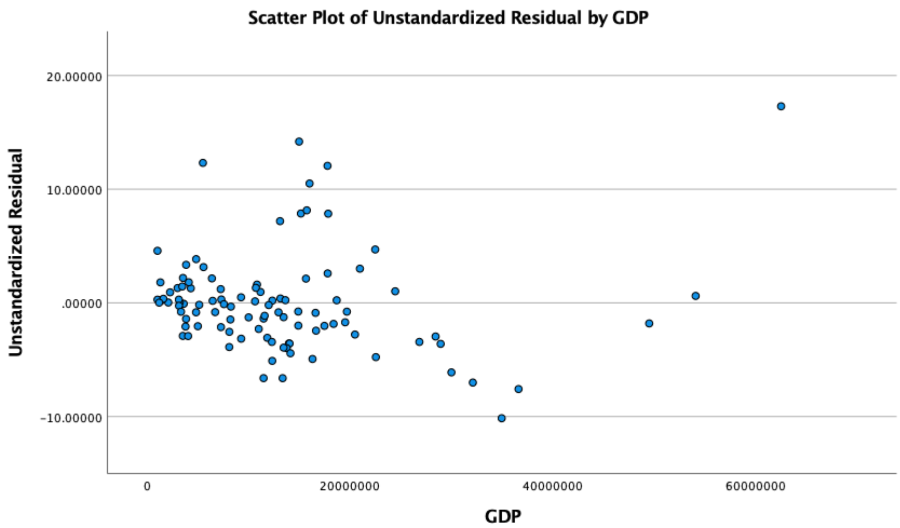
24 : RES_1 -1.70477237491771

	INDEX	GDP	T	T2	LNGDP	RES_1
1	101.0	28456500	1	1.00	17.16	-2.95386
2	100.7	18401000	1	1.00	16.73	-1.84494
3	103.1	10854284	1	1.00	16.20	1.61246
4	99.1	3863400	1	1.00	15.17	-1.40802
5	100.3	2111188	1	1.00	14.56	.03749
6	101.9	12364727	1	1.00	16.33	.20082
7	96.6	12356400	1	1.00	16.33	-5.09801
8	100.1	10030125	1	1.00	16.12	-1.27207
9	102.5	11201156	1	1.00	16.23	.96386
10	105.1	49508400	1	1.00	17.72	-1.80353
11	100.2	11503034	1	1.00	16.26	-1.37844
12	104.3	15680138	1	1.00	16.57	2.13629
13	109.0	13126854	1	1.00	16.39	7.19404
14	100.4	3634412	1	1.00	15.11	-.07594
15	102.8	10742289	1	1.00	16.19	1.32815
16	103.9	5583268	1	1.00	15.54	3.15100

Step 3: Follow the instructions in part c).

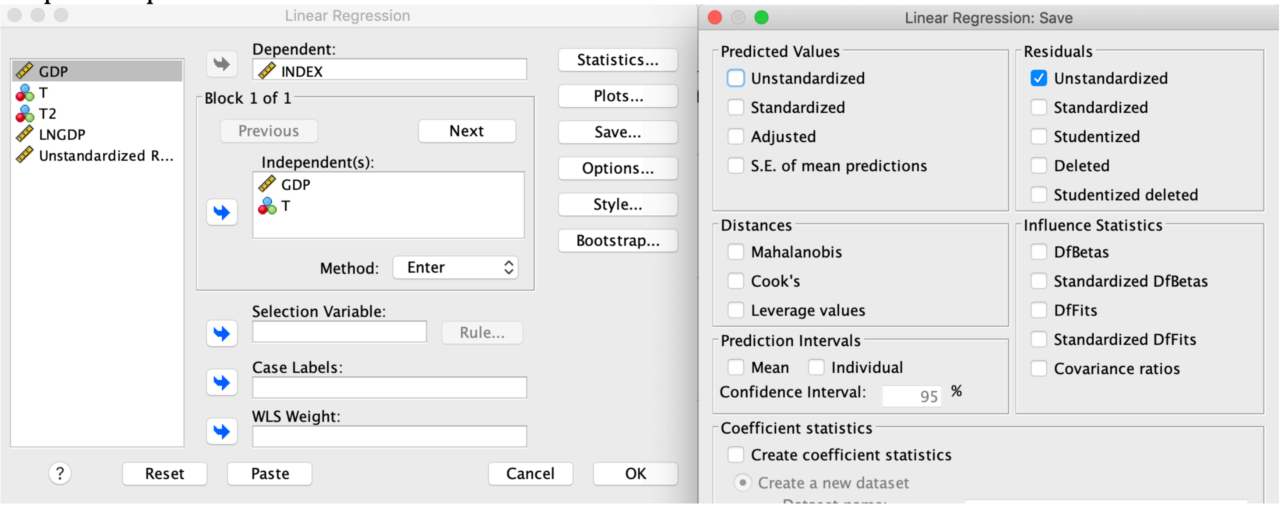


Output:

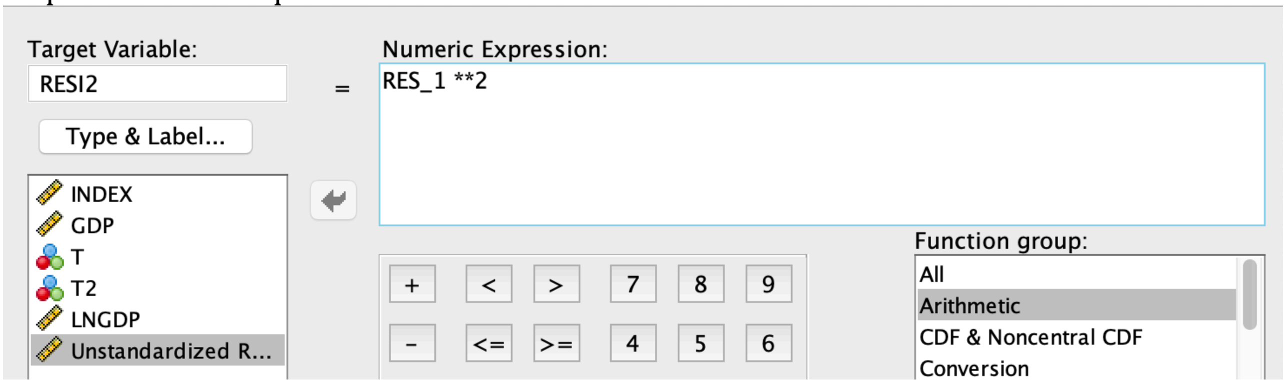


j) Perform a White heteroskedasticity test on the model $INDEX = \beta_0 + \beta_1 GDP + \beta_2 T + \varepsilon$.

Step 1: Request the unstandardized residuals to be saved as a new variable.



Step 2: Create the squared term of the unstandardised residuals and name it as RESI2.

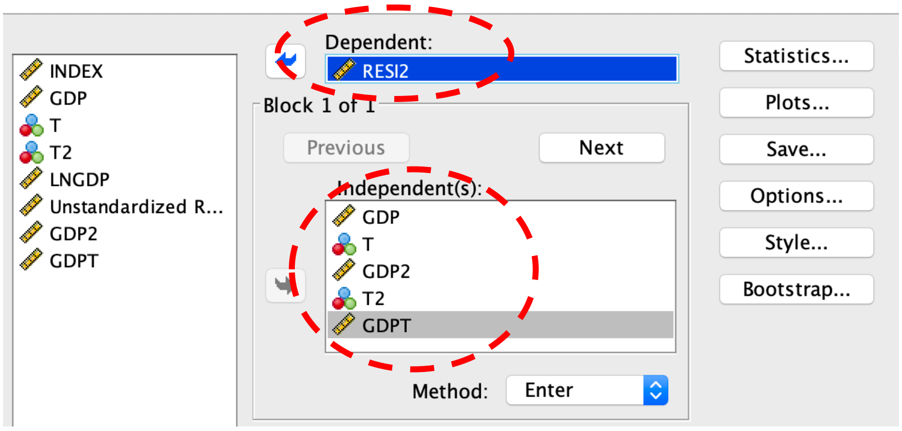


Step 3: Create the squared and cross terms of all independent variables, i.e., GDP2 = GDP*GDP, GDPT = GDP*T. Note that T2 has already been created in Part f).

	INDEX	GDP	T	T2	LNGDP	RES_1	RESI2	GDP2	GDPT
1	101.0	28456500	1	1.00	17.16	-2.95386	8.73	8.10E+14	28456500.00
2	100.7	18401000	1	1.00	16.73	-1.84494	3.40	3.39E+14	18401000.00
3	103.1	10854284	1	1.00	16.20	1.61246	2.60	1.18E+14	10854284.00
4	99.1	3863400	1	1.00	15.17	-1.40802	1.98	1.49E+13	3863400.00
5	100.3	2111188	1	1.00	14.56	.03749	.00	4.46E+12	2111188.00
6	101.9	12364727	1	1.00	16.33	.20082	.04	1.53E+14	12364727.00
7	96.6	12356400	1	1.00	16.33	-5.09801	25.99	1.53E+14	12356400.00
8	100.1	10030125	1	1.00	16.12	-1.27207	1.62	1.01E+14	10030125.00
9	102.5	11201156	1	1.00	16.23	.96386	.93	1.25E+14	11201156.00
10	105.1	49508400	1	1.00	17.72	-1.80353	3.25	2.45E+15	49508400.00
11	100.2	11503034	1	1.00	16.26	-1.37844	1.90	1.32E+14	11503034.00
12	104.3	15680138	1	1.00	16.57	2.13629	4.56	2.46E+14	15680138.00
13	109.0	13126854	1	1.00	16.39	7.19404	51.75	1.72E+14	13126854.00
14	100.4	3634412	1	1.00	15.11	-.07594	.01	1.32E+13	3634412.00

Step 4: Estimate the following regression model

$$RESI2 = \beta_0 + \beta_1 GDP + \beta_2 T + \beta_3 GDP2 + \beta_4 T2 + \beta_5 GDPT + \varepsilon$$



Output:

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	75741.626	5	15148.325	12.320	<.001[b]
	Residual	115584.250	94	1229.620		
	Total	191325.876	99			

a. Dependent Variable: RESID2

b. Predictors: (Constant), GDPT, T, GDP2, GDP, T2

Coefficients[a]

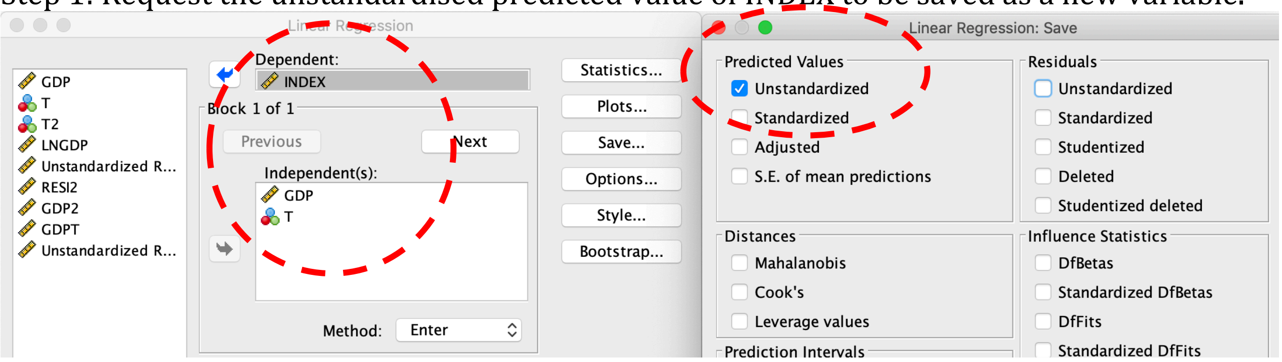
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	12.938	27.508		.470	.639
	GDP	-3.337E-6	.000	-.828	-3.087	.003
	T	9.694	29.608	.178	.327	.744
	T2	-5.427	7.453	-.401	-.728	.468
	GDP2	2.527E-14	.000	.327	1.494	.139
	GDPT	1.767E-6	.000	1.152	4.143	<.001

a. Dependent Variable: RESID2

Because the F test statistic is significant at the 5% level, the model suffers from heteroskedasticity problem. This may be caused by the interaction term between GDP and T, as the coefficient estimate of GDPT is significant.

k) Perform a RESET test on the final model from part h)

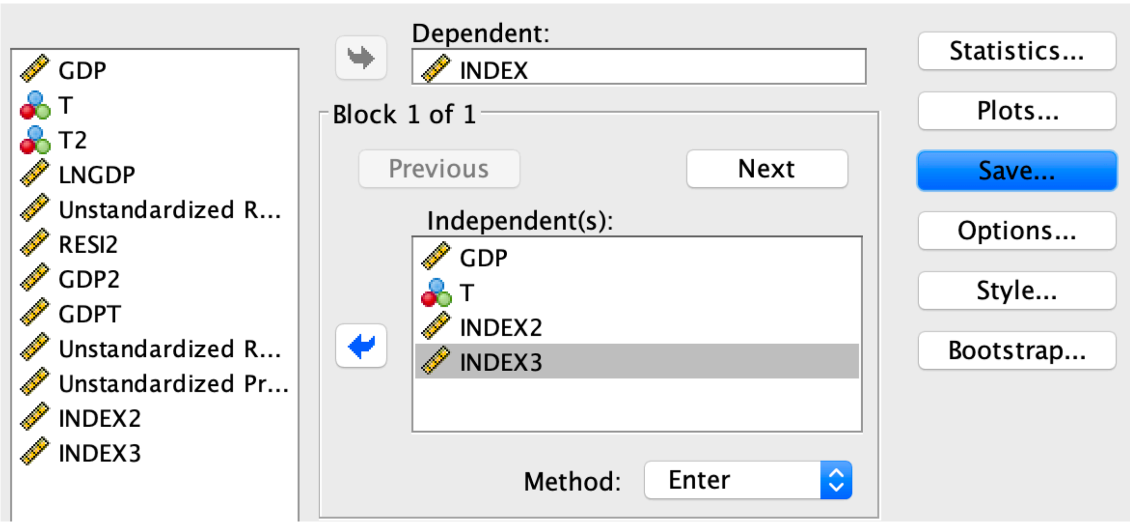
Step 1: Request the unstandardised predicted value of INDEX to be saved as a new variable.



Step 2: Generate the squared and cubed terms for the unstandardised predicted value of INDEX. Name them as INDEX2 and INDEX3 respectively.

PRE_1	INDEX2	INDEX3
103.95386	10806.41	1123367.58
102.54494	10515.47	1078307.83
101.48754	10299.72	1045293.39
100.50802	10101.86	1015318.17
100.26251	10052.57	1007896.01
101.69918	10342.72	1051846.37
101.69801	10342.49	1051810.17
101.37207	10276.30	1041729.33
101.53614	10309.59	1046795.87
106.90353	11428.37	1221732.59
101.57844	10318.18	1048104.62

Step 3: Estimate the full model by including INDEX2 and INDEX3 as new regressors.



Full model output: Note that INDEX2 is excluded from the model due to collinearity. Therefore, the full model is $INDEX = \beta_0 + \beta_1 GDP + \beta_2 T + \beta_3 INDEX3 + \varepsilon$.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	690.092	3	230.031	13.002	.000[b]
	Residual	1698.468	96	17.692		
	Total	2388.560	99			

The reduced model is $INDEX = \beta_0 + \beta_1 GDP + \beta_2 T + \varepsilon$. The output is

ANOVA[a]

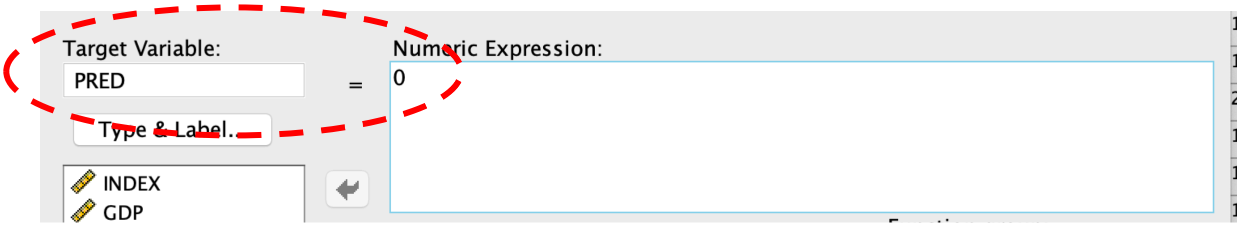
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	455.728	2	227.864	11.435	.000[b]
	Residual	1932.832	97	19.926		
	Total	2388.560	99			

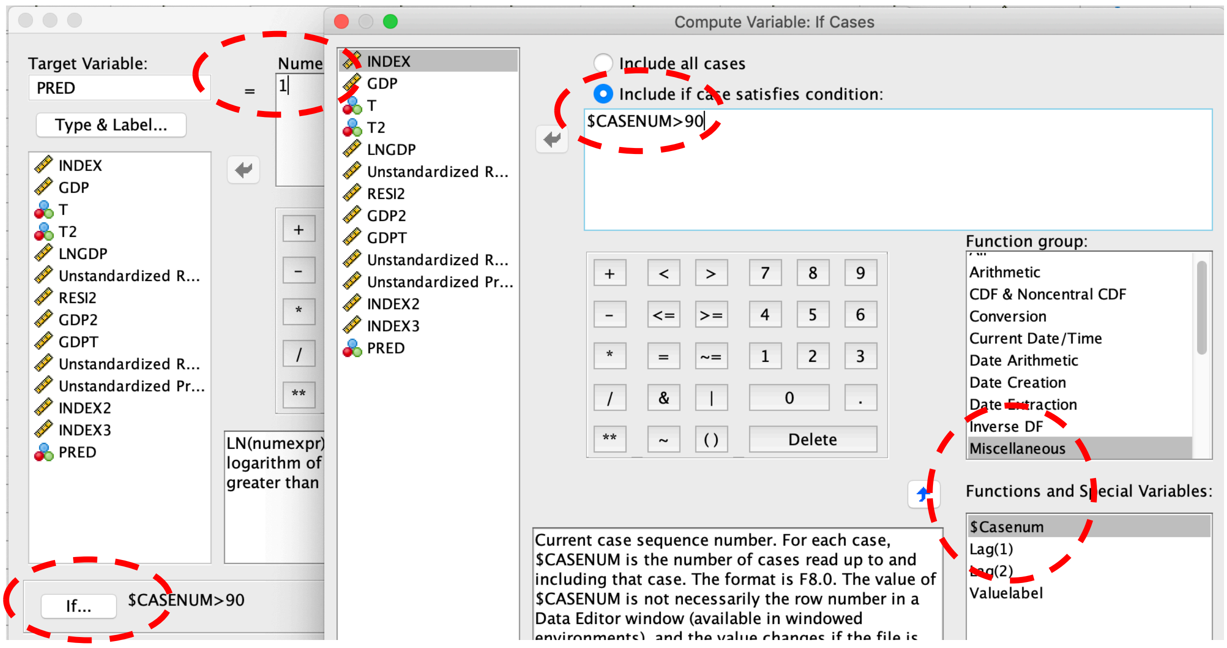
$$F = \frac{(1932.832 - 1698.468)/1}{1698.468/96} = \frac{234.364}{17.692} = 13.2466 > FINV(0.05,1,96) = 3.9402$$

Reject the null hypothesis. The model has misspecification problems.

l) Perform a predictive failure test on the model $INDEX = \beta_0 + \beta_1 GDP + \beta_2 T + \varepsilon$. Reserve the last 10 observations for the test.

Step 1: Create a new variable PRED, which equals 1 for the last 10 observations only, and zero otherwise.

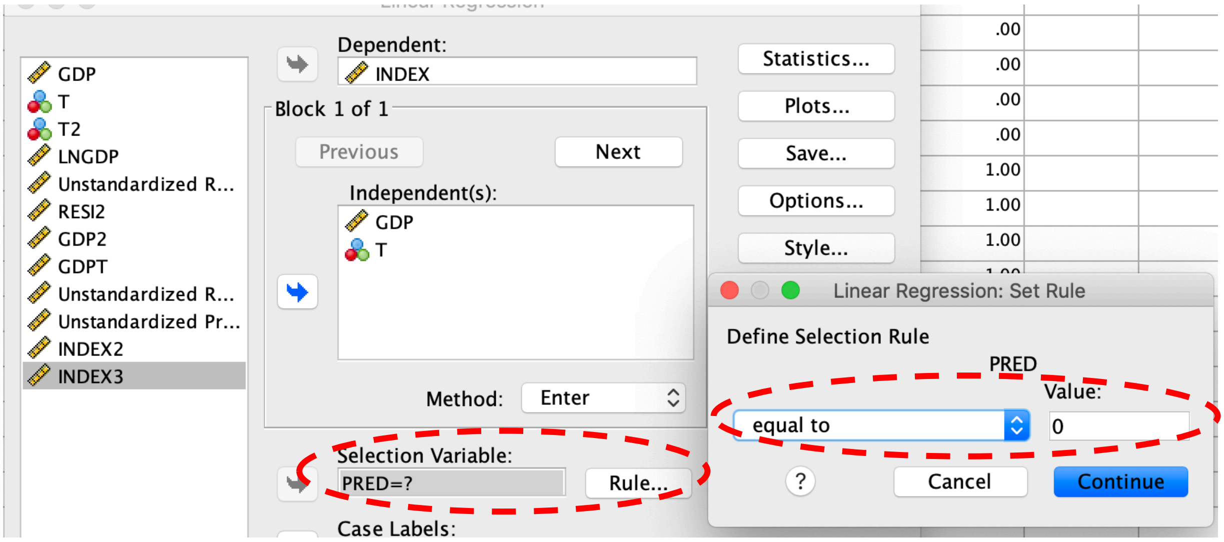




Outputs:

	RES_1	RESI2	GDP2	GDPT	RES_2	PRE_1	INDEX2	INDEX3	PRED
84	-.82475	.68	2.35E+13	1.14E+20	-9.78644	103.72475	10758.82	1115956.17	.00
85	-3.93353	15.47	1.82E+14	2.45E+21	-6.68887	104.93353	11011.05	1155427.89	.00
86	-.10968	.01	5.77E+13	4.38E+20	-14.13426	104.10968	10838.83	1128426.75	.00
87	2.15659	4.65	4.11E+13	2.63E+20	-7.91759	103.94341	10804.23	1123028.82	.00
88	.24174	.06	1.86E+14	2.55E+21	-22.35863	104.95826	11016.24	1156244.99	.00
89	12.06013	145.45	3.17E+14	5.64E+21	116.53470	105.53987	11138.66	1175573.02	.00
90	-2.28969	5.24	1.22E+14	1.34E+21	-13.48424	104.58969	10939.00	1144107.00	.00
91	-.87419	.76	2.76E+14	4.59E+21	-26.18894	105.37419	11103.72	1170045.64	1.00
92	-3.14759	9.91	8.64E+13	8.03E+20	-6.49371	104.34759	10888.42	1136180.22	1.00
93	-10.14487	102.92	1.22E+15	4.28E+22	27.24331	107.94487	11652.09	1257783.79	1.00
94	-3.60212	12.98	8.38E+14	2.43E+22	-41.35644	107.10212	11470.86	1228553.88	1.00
95	-2.04936	4.20	2.53E+13	1.27E+20	-6.50566	103.74936	10763.93	1116750.71	1.00
96	.93409	.87	5.24E+12	1.20E+19	-6.00757	103.36591	10684.51	1104414.23	1.00
97	4.70140	22.10	5.07E+14	1.14E+22	-15.76819	106.19860	11278.14	1197722.91	1.00
98	.23349	.05	3.50E+14	6.55E+21	-30.42572	105.66651	11165.41	1179809.96	1.00
99	-2.07897	4.32	1.45E+13	5.53E+19	-4.71333	103.57897	10728.60	1111257.54	1.00
100	-3.88299	15.08	6.59E+13	5.35E+20	.23754	104.18299	10854.10	1130812.08	1.00

Step 2: Estimate the reduced model ($INDEX = \beta_0 + \beta_1 GDP + \beta_2 T + \varepsilon$) by using the first 90 observations only.



Model 1 Outputs (using the first 90 observations):

ANOVA[a,b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	528.676	2	264.338	13.190	.000[c]
	Residual	1743.508	87	20.040		
	Total	2272.185	89			

Model 2 ($INDEX = \beta_0 + \beta_1 GDP + \beta_2 T + \varepsilon$) outputs (using all 100 observations):

ANOVA[a]

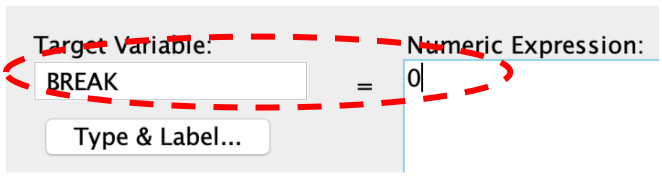
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	455.728	2	227.864	11.435	.000[b]
	Residual	1932.832	97	19.926		
	Total	2388.560	99			

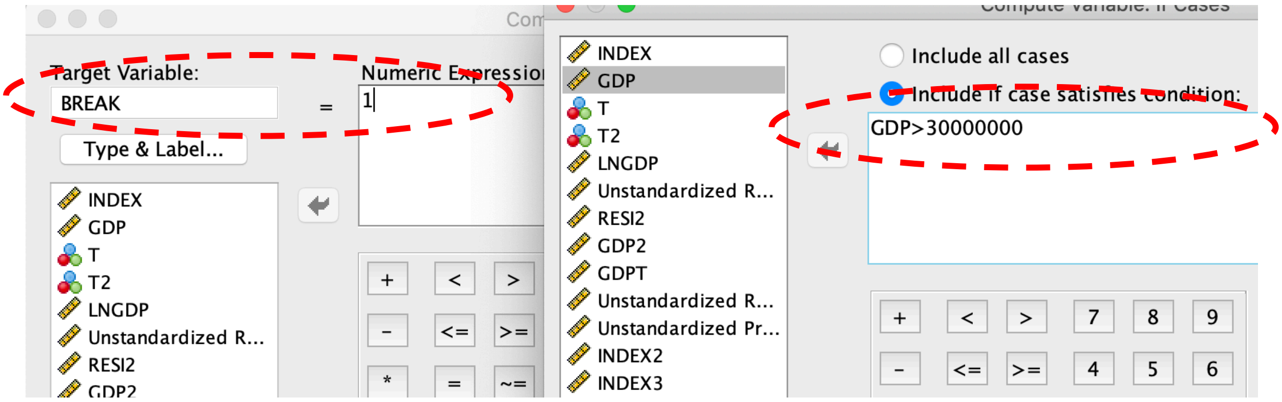
$$F = \frac{(1932.832 - 1743.508)/10}{1743.508/87} = \frac{18.932}{20.040} = 0.9447 < FINV(0.05,10,87) = 1.9413$$

Do not reject the null hypothesis. The model predicts well.

m) Test if there is a structural break at GDP = 30,000,000.

Step 1: Create a dummy variable BREAK, which equals one when GDP > 30,000,000, and zero otherwise.



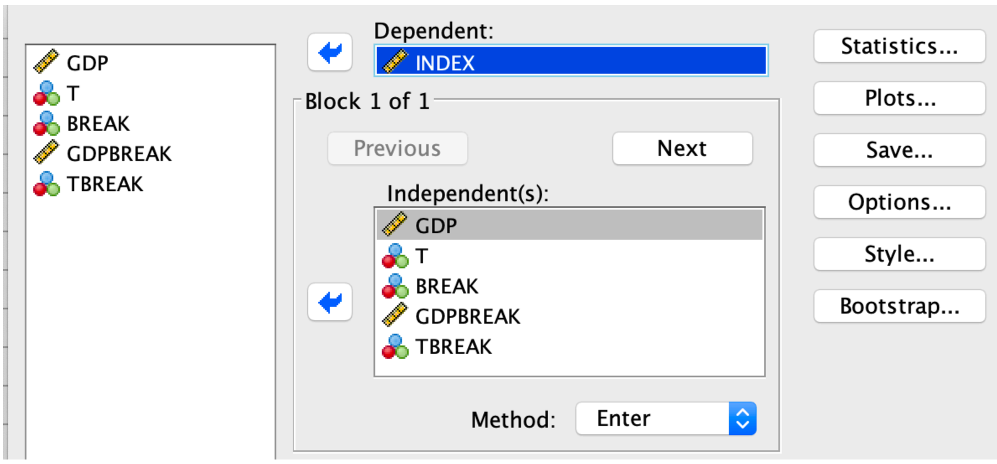


Step 2: Create interaction terms between BREAK and the two regressors GDP and T. Name them as GDPBREAK and TBREAK respectively.

Outputs:

	INDEX	GDP	T	BREAK	GDPBREAK	TBREAK
34	104.7	1048208	1	.00	.00	.00
35	100.7	3150000	1	.00	.00	.00
36	99.0	32127100	2	1.00	32127100.00	2.00
37	101.6	20511600	2	.00	.00	.00
38	100.1	11868132	2	.00	.00	.00
39	103.4	4328880	2	.00	.00	.00
40	101.7	3167000	2	.00	.00	.00
41	99.9	14000238	2	.00	.00	.00
42	99.9	14061000	2	.00	.00	.00
43	96.5	11501802	2	.00	.00	.00
44	99.8	12321276	2	.00	.00	.00
45	109.7	54087600	2	1.00	54087600.00	2.00
46	102.5	12975725	2	.00	.00	.00
47	106.6	17818302	2	.00	.00	.00

Step 3: Estimate the full model $INDEX = \beta_0 + \beta_1 GDP + \beta_2 T + \beta_3 BREAK + \beta_4 GDPBREAK + \beta_5 TBREAK + \varepsilon$.



Full model outputs:

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	909.611	5	181.922	11.563	.000[b]
	Residual	1478.949	94	15.733		
	Total	2388.560	99			

a. Dependent Variable: INDEX

b. Predictors: (Constant), TBREAK, T, GDP, GDPBREAK, BREAK

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	98.877	1.174		84.205	.000
	GDP	1.045E-7	.000	.232	1.672	.098
	T	1.588	.519	.262	3.062	.003
	BREAK	-34.953	7.710	-1.825	-4.534	.000
	GDPBREAK	6.940E-7	.000	1.612	4.825	.000
	TBREAK	1.786	2.204	.224	.810	.420

Reduced model ($INDEX = \beta_0 + \beta_1 GDP + \beta_2 T + \varepsilon$) outputs:

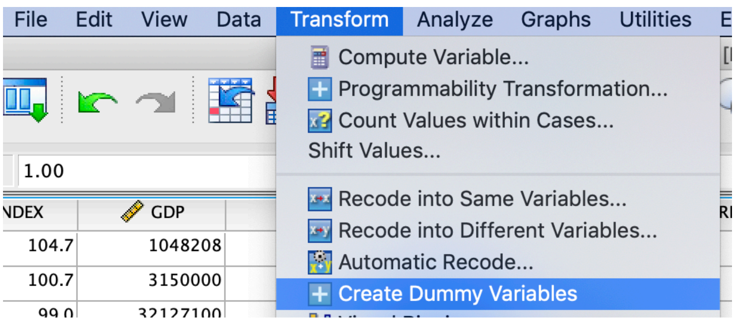
ANOVA[a]

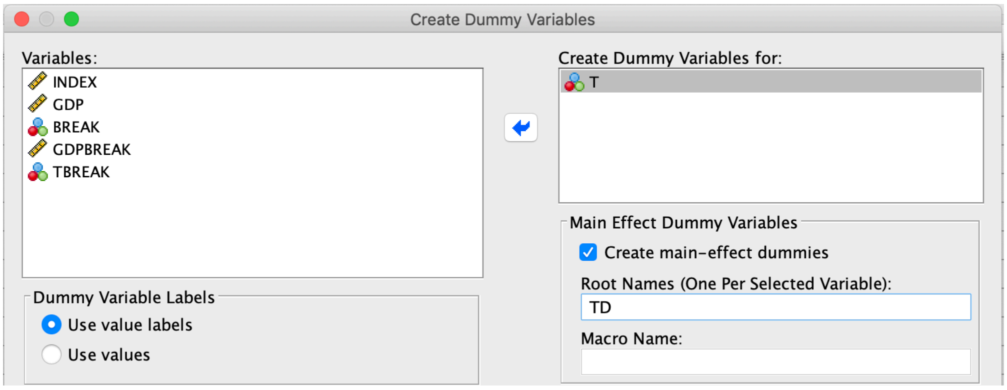
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	455.728	2	227.864	11.435	.000[b]
	Residual	1932.832	97	19.926		
	Total	2388.560	99			

$$F = \frac{(1932.832 - 1478.949)/3}{1478.949/94} = \frac{151.294}{15.733} = 9.616 > FINV(0.05,3,94) = 2.7014.$$

Reject the null hypothesis. There is a structure break at GDP = 30,000,000.

n) Create dummy variables for T

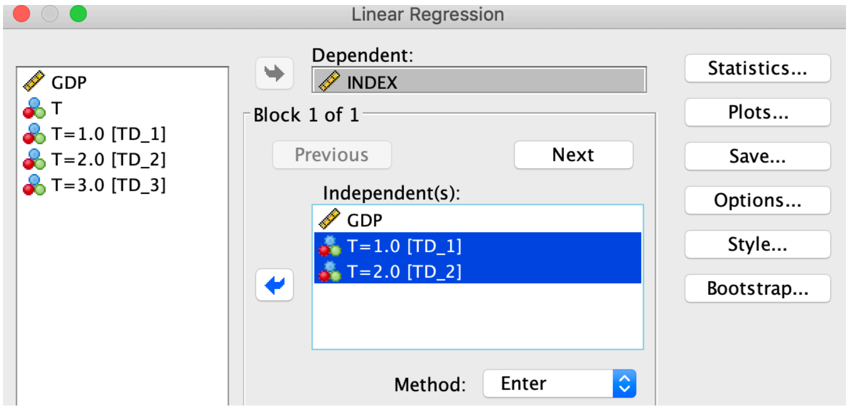




Outputs:

	INDEX	GDP	T	TD_1	TD_2	TD_3
32	101.9	3487465	1	1.00	.00	.00
33	100.4	1044875	1	1.00	.00	.00
34	104.7	1048208	1	1.00	.00	.00
35	100.7	3150000	1	1.00	.00	.00
36	99.0	32127100	2	.00	1.00	.00
37	101.6	20511600	2	.00	1.00	.00
38	100.1	11868132	2	.00	1.00	.00
39	103.4	4328880	2	.00	1.00	.00
40	101.7	3167000	2	.00	1.00	.00
41	99.9	14000238	2	.00	1.00	.00
42	99.9	14061000	2	.00	1.00	.00
43	96.5	11501802	2	.00	1.00	.00

o) Estimate a regression model using the group of dummy variables created in part n)



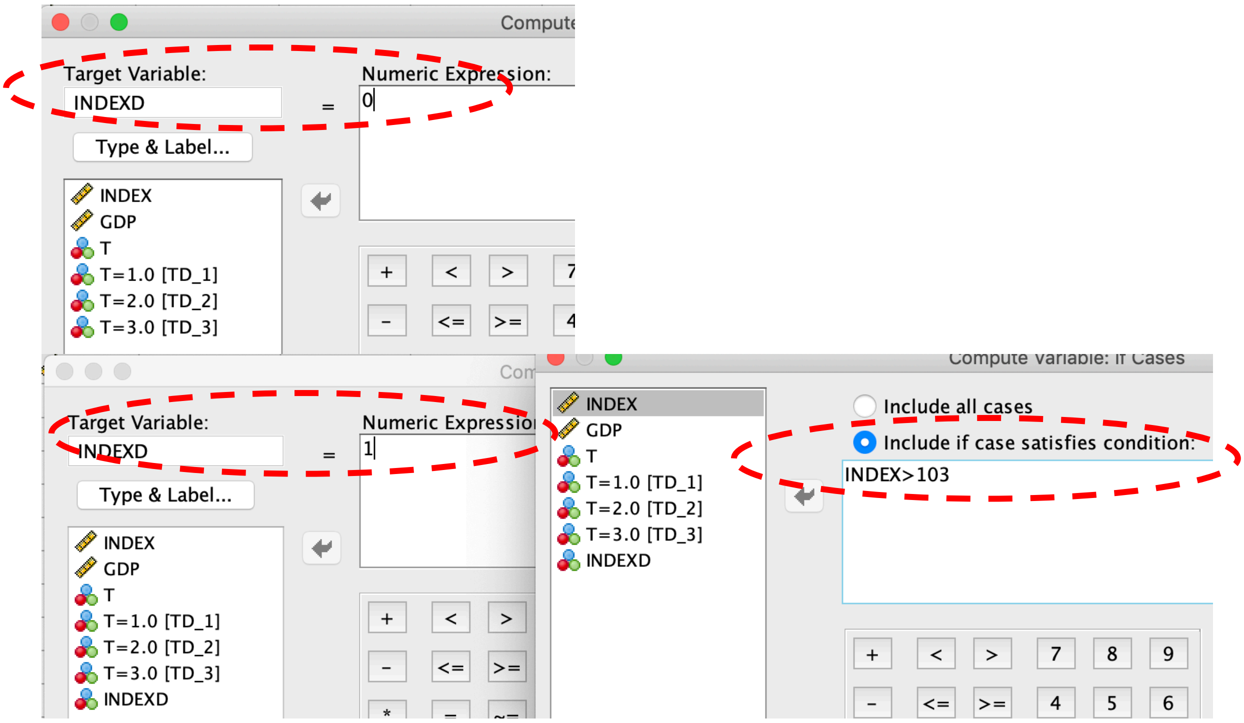
Outputs:

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	103.215	1.066		96.828	.000
	GDP	1.391E-7	.000	.309	3.301	.001
	T=1.0	-3.106	1.136	-.303	-2.733	.007
	T=2.0	-1.960	1.127	-.191	-1.740	.085

a. Dependent Variable: INDEX

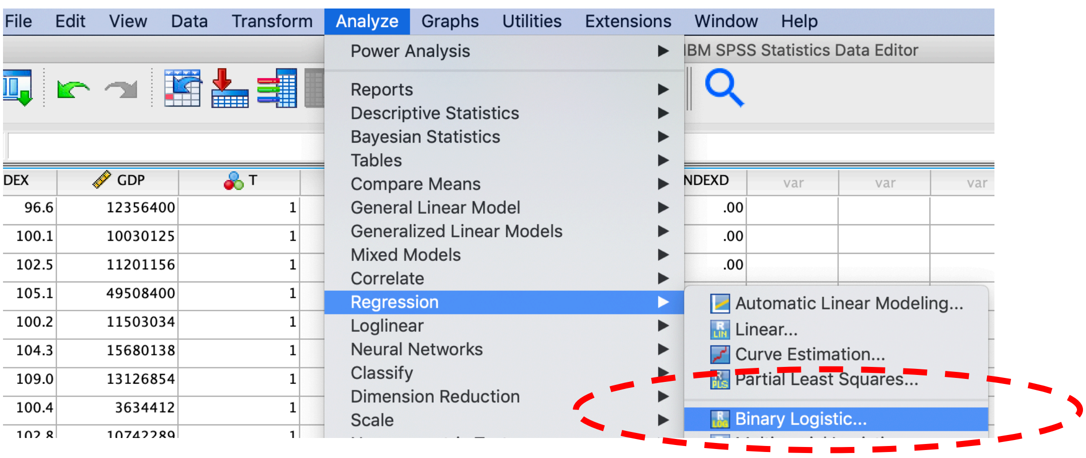
p) Create a dummy variable INDEXD, which equals one when INDEX > 103 and zero otherwise

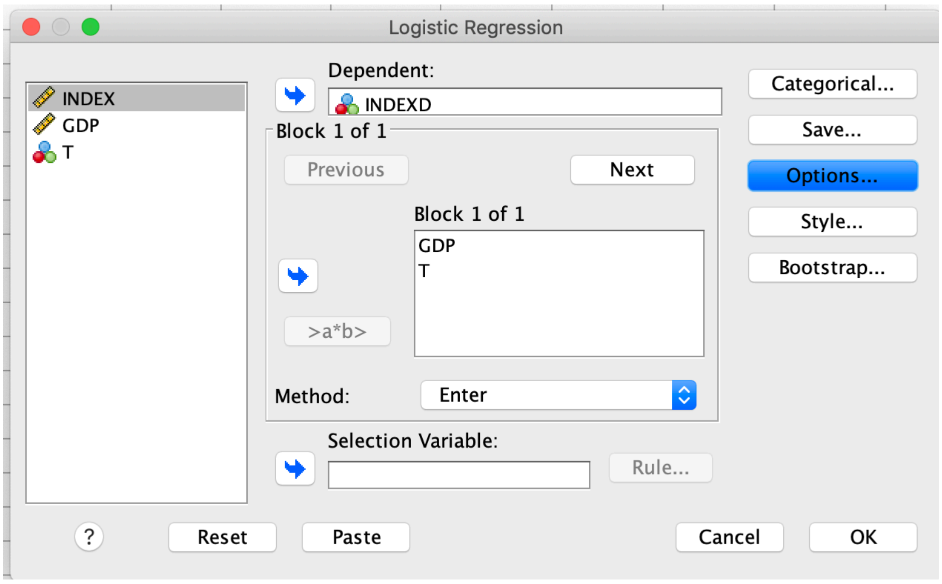


Outputs:

	INDEX	GDP	T	INDEXD
1	101.0	28456500	1	.00
2	100.7	18401000	1	.00
3	103.1	10854284	1	1.00
4	99.1	3863400	1	.00
5	100.3	2111188	1	.00
6	101.9	12364727	1	.00
7	96.6	12356400	1	.00
8	100.1	10030125	1	.00
9	102.5	11201156	1	.00
10	105.1	49508400	1	1.00
11	100.2	11503034	1	.00
12	104.3	15680138	1	1.00
13	109.0	13126854	1	1.00

q) Estimate a logit model by using INDEXD as the dependent variable, and T and GDP as the independent variables





Outputs:

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	119.710[a]	.093	.128

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Classification Table[a]

		Observed	Predicted INDEXD .00	Predicted INDEXD 1.00	Percentage Correct
Step 1	INDEXD	.00	57	8	87.7
		1.00	22	13	37.1
	Overall Percentage				70.0

a. The cut value is .500

Variables in the Equation

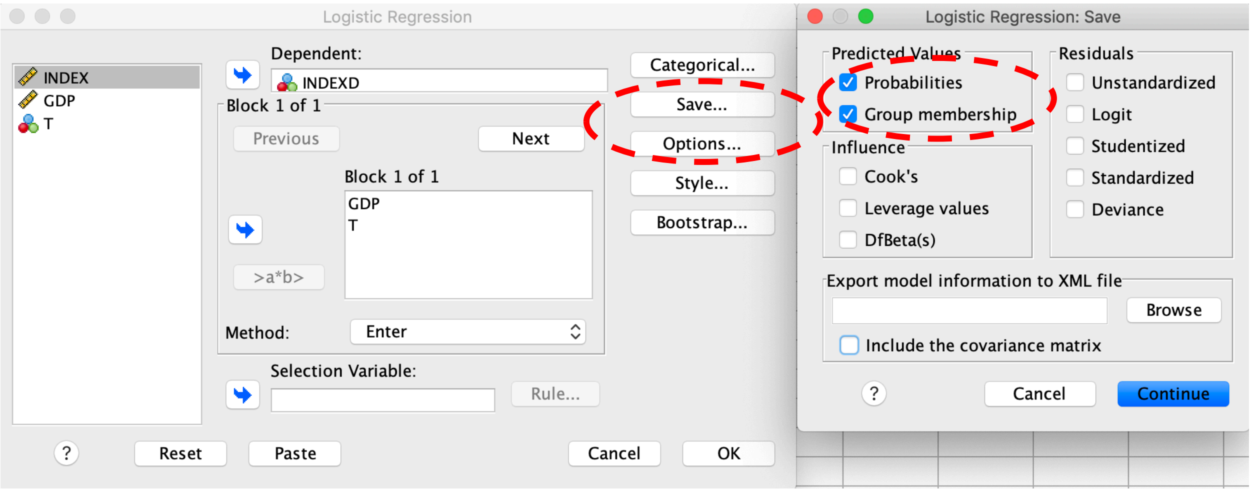
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	GDP	.000	.000	1.721	1	.190	1.000
	T	.704	.283	6.189	1	.013	2.021
	Constant	-2.404	.661	13.223	1	.000	.090

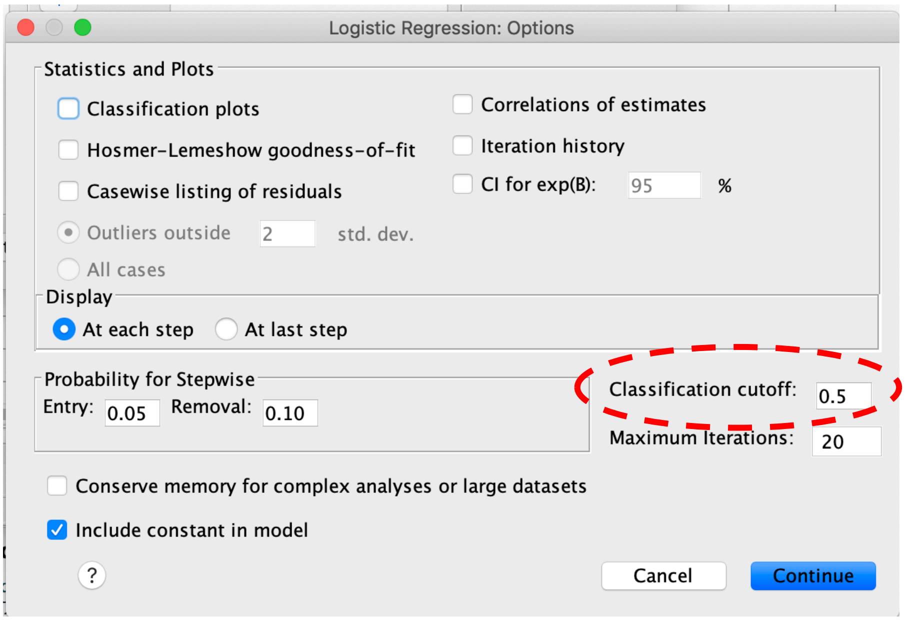
r) Predict the value of INDEXD when T = 1, 2, and 3 and GDP = 30,000000

Step 1: Input the value of T and GDP into the datafile.

	INDEX	GDP	T	INDEXD
97	110.9	22505600	3	1.00
98	105.9	18708046	3	1.00
99	101.5	3809191	3	.00
100	100.3	8120121	3	.00
101	.	30000000	1	.
102	.	30000000	2	.
103	.	30000000	3	.

Step 2: Re-estimate the model, and request the unstandardised predicted value of INDEXD.





Outputs:

	INDEX	GDP	T	INDEXD	PRE_1	PGR_1
96	104.3	2288595	3	1.00	.44239	.00
97	110.9	22505600	3	1.00	.57760	1.00
98	105.9	18708046	3	1.00	.55247	1.00
99	101.5	3809191	3	.00	.45251	.00
100	100.3	8120121	3	.00	.48140	.00
101	.	30000000	1	.	.29063	.00
102	.	30000000	2	.	.45294	.00
103	.	30000000	3	.	.62591	1.00