RM01 Research Methods

STATA Exercises

STATA Exercises Solutions

(Version: November 2021)

By Helen Bao

Department of Land Economy

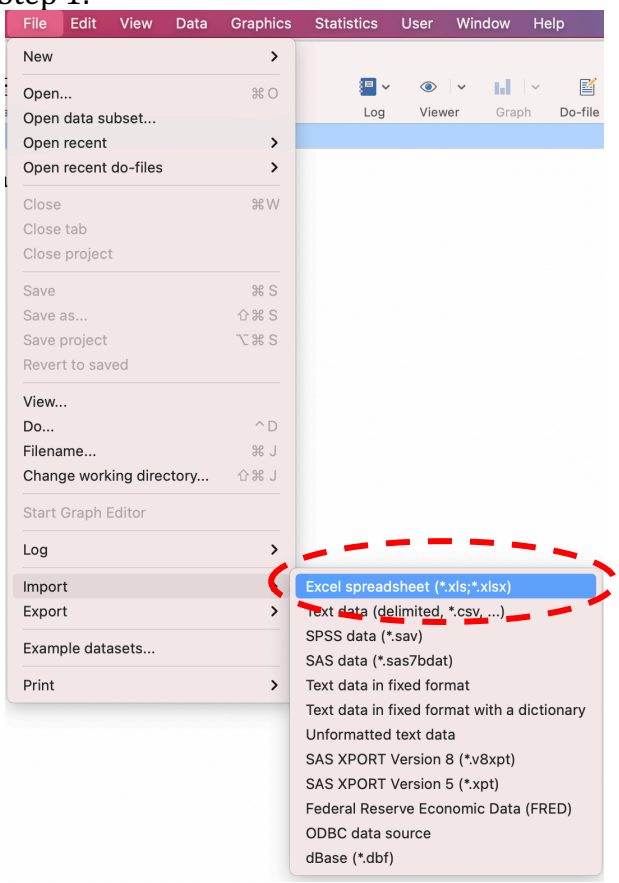
University of Cambridge

STATA Exercises

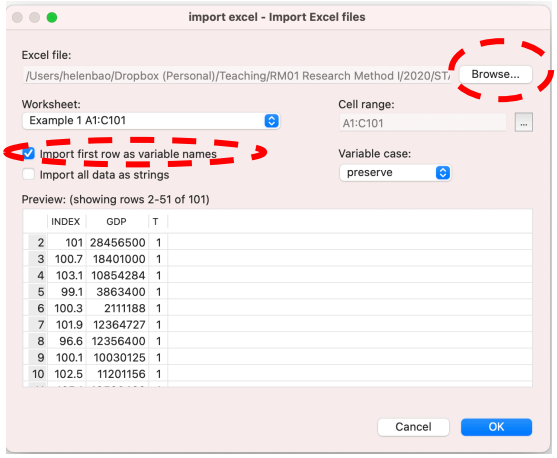
1. Quarterly data on property price index and GDP are given in Worksheet 'example1'. Variable T is a quarterly time index (e.g., T = 1 for the first quarter). Use STATA to complete the following tasks.

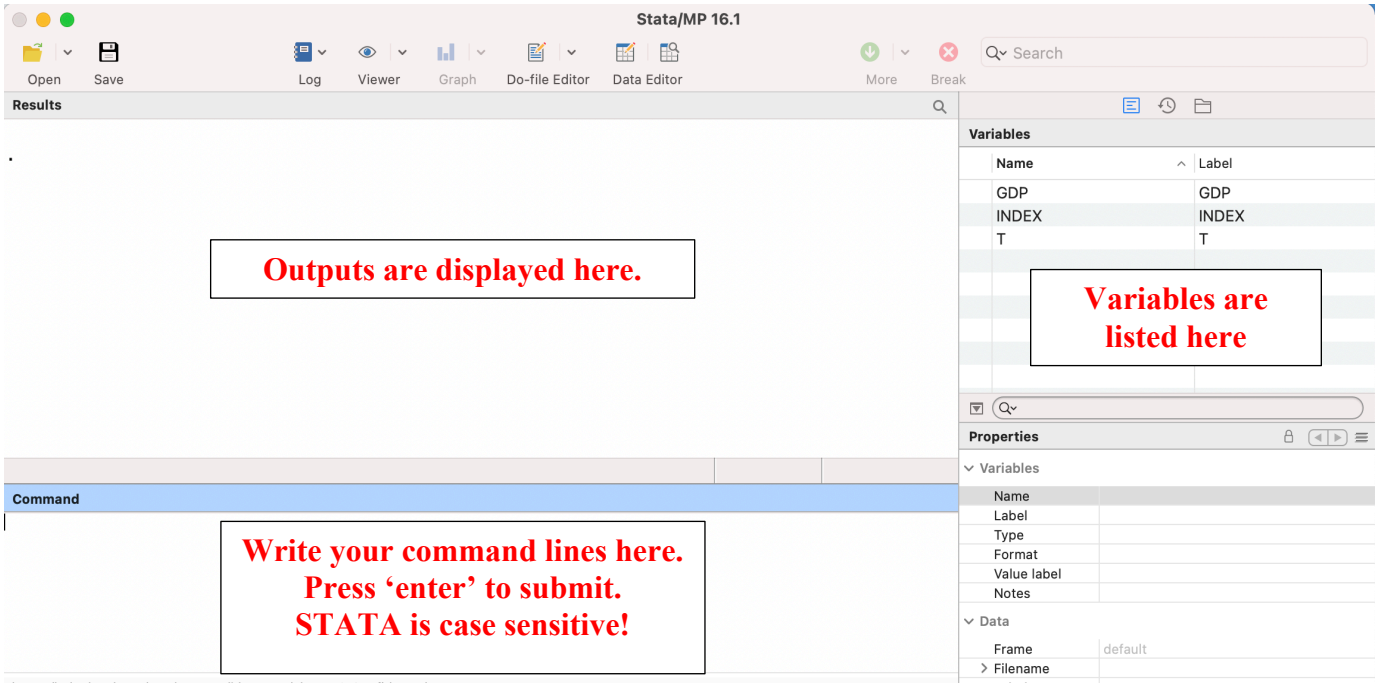
   a) Open the file in STATA and view the data

Step 1:



Step 2: Click 'Browse' to locate the file. Check the 'Import first row as variable names' box.



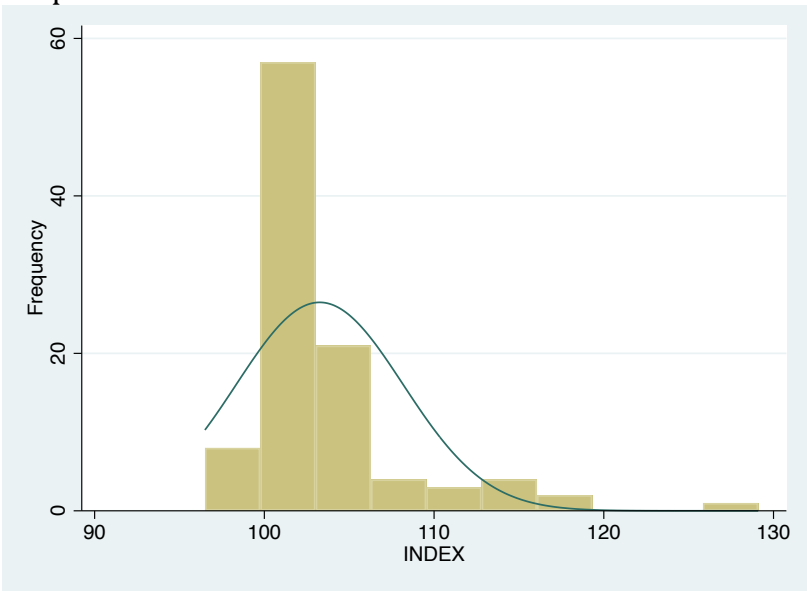


b) Create a histogram of variable INDEX

STATA command:

histogram INDEX, frequency normal

Output:

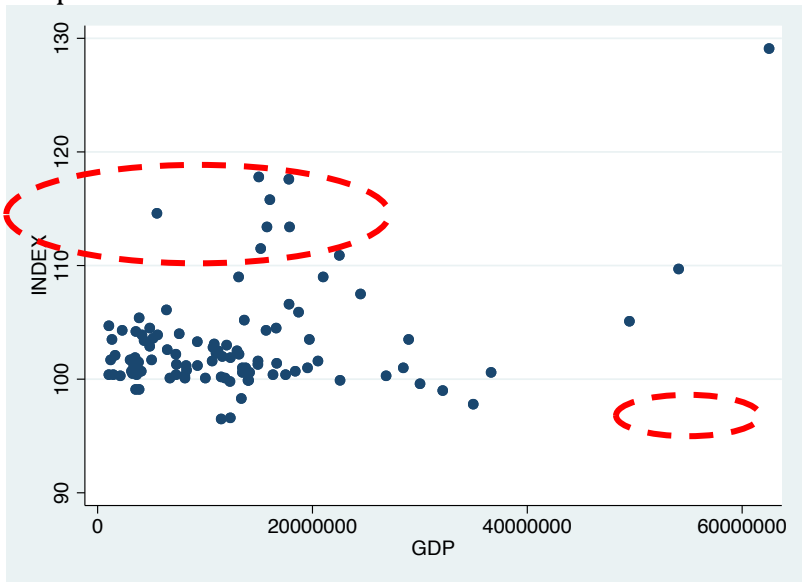


c) Create a scatter plot between INDEX and GDP

STATA command:

twoway (scatter INDEX GDP)

Output:



d) Generate descriptive statistics for INDEX and GDP

STATA command:

summarize INDEX GDP

Output:

Variable	Obs	Mean	Std. Dev.	Min	Max
INDEX	100	103.26	4.911911	96.5	129.1
GDP	100	1.31e+07	1.09e+07	1044875	6.25e+07

e) Generate frequency statistics for T

STATA command:

tabulate T

Output:

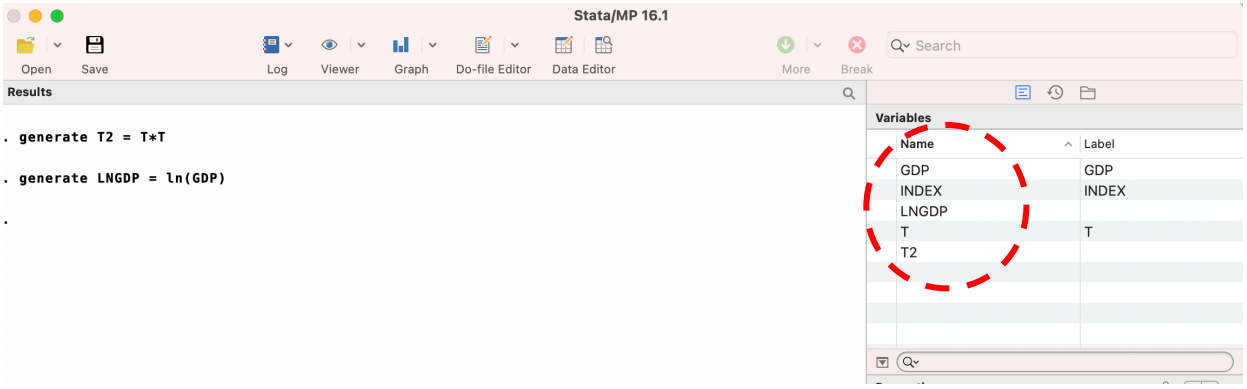
T	Freq.	Percent	Cum.
1	35	35.00	35.00
2	35	35.00	70.00
3	30	30.00	100.00
Total	100	100.00	

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

STATA command:

generate T2 = T*T
generate LNGDP = ln(GDP)

Outputs:



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

STATA commands:

reg INDEX GDP T T2
vif
generate YEAR = _n
tsset YEAR
estat dwatson

Output:

```
. reg INDEX GDP T T2
```

Source	SS	df	MS	Number of obs	=	100
				F(3, 96)	=	7.62
Model	459.488651	3	153.162884	Prob > F	=	0.0001
Residual	1929.07135	96	20.0944932	R-squared	=	0.1924
				Adj R-squared	=	0.1671
Total	2388.56	99	24.1268687	Root MSE	=	4.4827

INDEX	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
GDP	1.39e-07	4.21e-08	3.30	0.001	5.55e-08	2.23e-07
T	-.0776739	3.780335	-0.02	0.984	-7.58158	7.426232
T2	.4076264	.9422173	0.43	0.666	-1.46266	2.277913
_cons	99.77957	3.3564	29.73	0.000	93.11717	106.442

```
. vif
```

Variable	VIF	1/VIF
T2	46.16	0.021665
T	46.05	0.021716
GDP	1.04	0.961352
Mean VIF	31.08	

```
. generate YEAR = _n

. tsset YEAR
        time variable:  YEAR, 1 to 100
                delta:  1 unit

. estat dwatson

Durbin-Watson d-statistic(  4,   100) =   1.59622
```

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

STATA commands:

sw, pe(0.05) pr(0.10): reg INDEX GDP T T2 LNGDP

Outputs (Stepwise selection):

```
. sw, pe(0.05) pr(0.10): reg INDEX GDP T T2 LNGDP
                     begin with full model
p = 0.9644 >= 0.1000  removing T
p = 0.1136 >= 0.1000  removing LNGDP
```

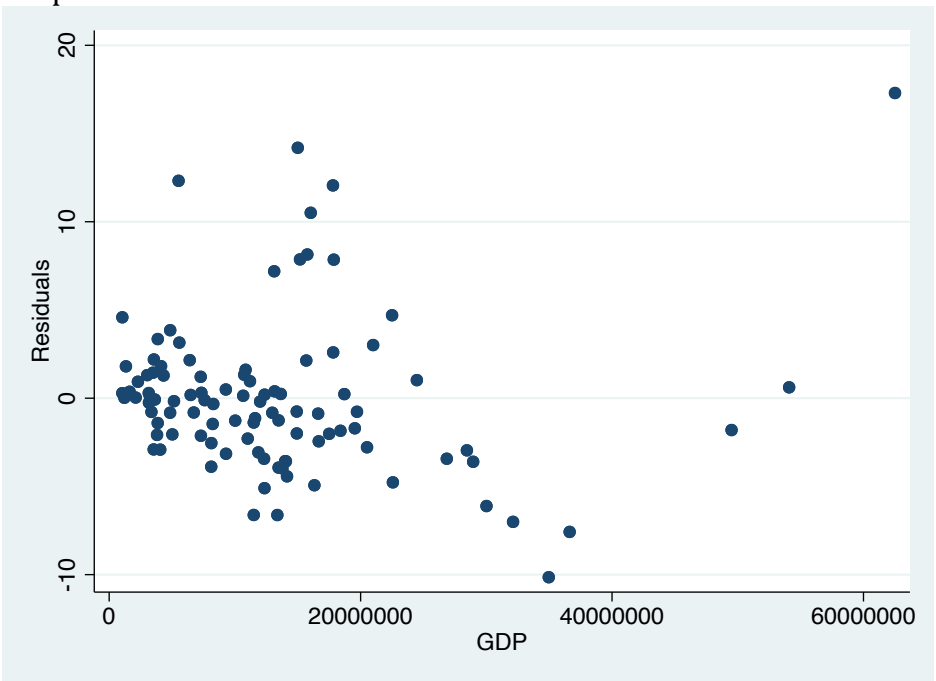
Source	SS	df	MS	Number of obs	=	100
				F(2, 97)	=	11.55
Model	459.480168	2	229.740084	Prob > F	=	0.0000
Residual	1929.07983	97	19.8874209	R-squared	=	0.1924
				Adj R-squared	=	0.1757
Total	2388.56	99	24.1268687	Root MSE	=	4.4595

INDEX	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
GDP	1.39e-07	4.19e-08	3.32	0.001	5.60e-08	2.22e-07
T2	.388486	.1406687	2.76	0.007	.1092976	.6676745
_cons	99.71299	.8700331	114.61	0.000	97.98622	101.4398

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

STATA commands:
reg INDEX GDP T
predict res, residuals
twoway (scatter res GDP)

Output:



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

STATA commands:
estat imtest,white

Output:

```
. estat imtest,white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(5)      =     39.59
         Prob > chi2  =    0.0000

Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	39.59	5	0.0000
Skewness	21.85	2	0.0000
Kurtosis	3.74	1	0.0532
Total	65.17	8	0.0000

Because the Chi square 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)

STATA commands:
estat ovtest

Outputs:

```
Ramsey RESET test using powers of the fitted values of INDEX
       Ho:  model has no omitted variables
                  F(3, 94) =      7.64
                  Prob > F =      0.0001
```

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

STATA commands:
reg INDEX GDP T if _n<91

Model 1 Outputs (using the first 90 observations):

```
. reg INDEX GDP T if _n<91

      Source |       SS           df       MS      Number of obs   =        90
-------------+----------------------------------   F(2, 87)        =     13.19
       Model |  528.676429         2  264.338214   Prob > F        =    0.0000
    Residual |  1743.50846        87  20.0403271   R-squared       =    0.2327
-------------+----------------------------------   Adj R-squared   =    0.2150
       Total |  2272.18489        89  25.5301673   Root MSE        =    4.4766
```

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

STATA commands:
reg INDEX GDP T

Model 2 Outputs (using all 100 observations):

```
. reg INDEX GDP T
```

Source	SS	df	MS		
				Number of obs	= 100
				F(2, 97)	= 11.44
Model	455.727685	2	227.863842	Prob > F	= 0.0000
Residual	1932.83232	97	19.9261063	R-squared	= 0.1908
				Adj R-squared	= 0.1741
Total	2388.56	99	24.1268687	Root MSE	= 4.4639

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

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

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

STATA commands:
generate BREAK=0
replace BREAK=1 if GDP>30000000
generate TBREAK = T*BREAK
generate GDPBREAK=GDP*BREAK
reg INDEX GDP T
reg INDEX GDP T BREAK GDPBREAK TBREAK

Outputs:

New variables

	INDEX	GDP	T	INDEXD	INDEXDHAT	BREAK	TBREAK	GDPBREAK
34	104.7	1048208	1	1	.1581664	0	0	0
35	100.7	3150000	1	0	.1658486	0	0	0
36	99	32127100	2	0	.4671632	1	2	3.21e+07
37	101.6	20511600	2	0	.3907148	0	0	0
38	100.1	11868132	2	0	.3369195	0	0	0
39	103.4	4328880	2	1	.2931648	0	0	0
40	101.7	3167000	2	0	.286724	0	0	0
41	99.9	14000238	2	0	.3498632	0	0	0
42	99.9	14061000	2	0	.3502354	0	0	0
43	96.5	11501802	2	0	.3347194	0	0	0
44	99.8	12321276	2	0	.3396509	0	0	0
45	109.7	54087600	2	1	.6129658	1	2	5.41e+07
46	102.5	12975725	2	0	.3436145	0	0	0
47	106.6	17818302	2	1	.3735935	0	0	0
48	117.8	15003435	2	1	.3560322	0	0	0

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

```
. reg INDEX GDP T
```

Source	SS	df	MS	Number of obs	=	100
				F(2, 97)	=	11.44
Model	455.727685	2	227.863842	Prob > F	=	0.0000
Residual	1932.83232	97	19.9261063	R-squared	=	0.1908
				Adj R-squared	=	0.1741
Total	2388.56	99	24.1268687	Root MSE	=	4.4639

INDEX	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
GDP	1.40e-07	4.19e-08	3.34	0.001	5.70e-08	2.23e-07
T	1.539271	.5649354	2.72	0.008	.4180305	2.660511
_cons	98.42743	1.218401	80.78	0.000	96.00924	100.8456

Full model: $INDEX = \beta_0 + \beta_1 GDP + \beta_2 T + \beta_3 BREAK + \beta_4 GDPBREAK + \beta_5 TBREAK + \varepsilon$

```
. reg INDEX GDP T BREAK GDPBREAK TBREAK
```

Source	SS	df	MS	Number of obs	=	100
				F(5, 94)	=	11.56
Model	909.611055	5	181.922211	Prob > F	=	0.0000
Residual	1478.94895	94	15.7334994	R-squared	=	0.3808
				Adj R-squared	=	0.3479
Total	2388.56	99	24.1268687	Root MSE	=	3.9665

INDEX	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
GDP	1.05e-07	6.25e-08	1.67	0.098	-1.96e-08	2.29e-07
T	1.588375	.5186607	3.06	0.003	.5585618	2.618188
BREAK	-34.95293	7.70962	-4.53	0.000	-50.26056	-19.64529
GDPBREAK	6.94e-07	1.44e-07	4.82	0.000	4.08e-07	9.80e-07
TBREAK	1.786009	2.204446	0.81	0.420	-2.59097	6.162989
_cons	98.87681	1.174234	84.21	0.000	96.54534	101.2083

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

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

Alternatively, use the following command line to obtain the test result directly.

test BREAK GDPBREAK TBREAK

Outputs:

```
. test BREAK GDPBREAK TBREAK

 ( 1)  BREAK = 0
 ( 2)  GDPBREAK = 0
 ( 3)  TBREAK = 0

       F(  3,    94) =    9.62
            Prob > F =    0.0000
```

n) Create dummy variables for T

STATA commands:
tabulate T, generate(TD)

Outputs:

Variables

Name	Label
GDP	GDP
INDEX	INDEX
LNGDP	
T	T
T2	
TD1	T== 1.0000
TD2	T== 2.0000
TD3	T== 3.0000
YEAR	
res	Residuals

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

STATA commands:
reg INDEX GDP i.T

Outputs:

```
. reg INDEX GDP i.T
```

Source	SS	df	MS	Number of obs	=	100
				F(3, 96)	=	7.62
Model	459.488651	3	153.162884	Prob > F	=	0.0001
Residual	1929.07135	96	20.0944932	R-squared	=	0.1924
				Adj R-squared	=	0.1671
Total	2388.56	99	24.1268687	Root MSE	=	4.4827

INDEX	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
GDP	1.39e-07	4.21e-08	3.30	0.001	5.55e-08	2.23e-07
T						
2	1.145205	1.073097	1.07	0.289	-.9848753	3.275286
3	3.105663	1.136365	2.73	0.007	.8499959	5.36133
_cons	100.1095	.8891826	112.59	0.000	98.34451	101.8745

You may use **reg INDEX GDP ib3.T** to omit T=3 as the base category.

```
. reg INDEX GDP ib3.T
```

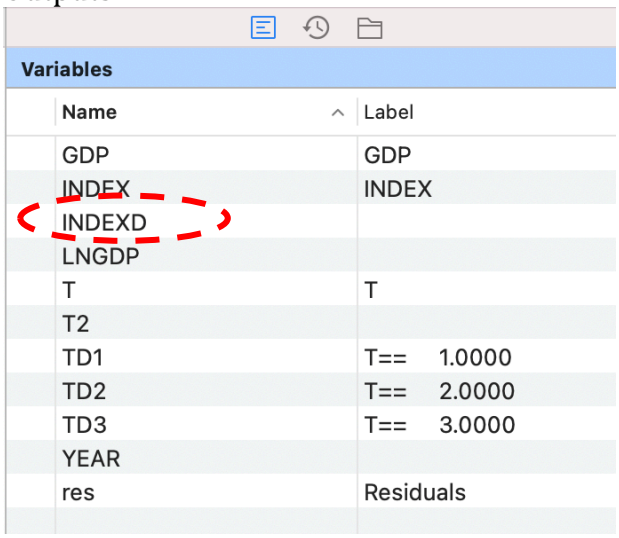
Source	SS	df	MS	Number of obs	=	100
				F(3, 96)	=	7.62
Model	459.488651	3	153.162884	Prob > F	=	0.0001
Residual	1929.07135	96	20.0944932	R-squared	=	0.1924
				Adj R-squared	=	0.1671
Total	2388.56	99	24.1268687	Root MSE	=	4.4827

INDEX	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
GDP	1.39e-07	4.21e-08	3.30	0.001	5.55e-08	2.23e-07
T						
1	-3.105663	1.136365	-2.73	0.007	-5.36133	-.8499959
2	-1.960458	1.126798	-1.74	0.085	-4.197133	.2762176
_cons	103.2152	1.065968	96.83	0.000	101.0993	105.3311

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

STATA commands:
generate INDEXD=0
replace INDEXD=1 if INDEX > 103

Outputs:



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

STATA commands:
logit INDEXD GDP T

Outputs:

```
. logit INDEXD GDP T, nolog

Logistic regression                             Number of obs     =        100
                                                LR chi2(2)        =       9.78
                                                Prob > chi2       =     0.0075
Log likelihood = -59.854919                     Pseudo R2         =     0.0755
```

INDEXD	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
GDP	2.69e-08	2.05e-08	1.31	0.190	-1.33e-08	6.72e-08
T	.7035333	.2827923	2.49	0.013	.1492707	1.257796
_cons	-2.403693	.6610158	-3.64	0.000	-3.69926	-1.108126

r) Predict the value of INDEXD when T = 1, 2, and 3 and GDP = 30,000,000

STATA commands:

```
set obs `=_N+3'
replace GDP=30000000 if missing(INDEX)
forvalues i=1(1)3 {
        replace T=`i' if _n==100+`i'
        }
predict INDEXDHAT
```

Outputs:

	INDEX	GDP	T	INDEXD	INDEXDHAT
91	104.5	16621797	3	1	.5385443
92	101.2	9294858	3	0	.4892992
93	97.8	34968787	3	0	.6566758
94	103.5	28954070	3	1	.6192948
95	101.7	5025271	3	0	.4606373
96	104.3	2288595	3	1	.44239
97	110.9	22505600	3	1	.5775973
98	105.9	18708046	3	1	.5524712
99	101.5	3809191	3	0	.4525129
100	100.3	8120121	3	0	.4813981
101	.	30000000	1	.	.2906258
102	.	30000000	2	.	.4529364
103	.	30000000	3	.	.6259123