Lecture 23

The "Modern" Approach to Building a Multiple Time Series Regression

You should refer to Exercise 9 for a discussion of Spurious Regressions and unbalanced Regression equations. The bottom line of that exercise is that

(i) it is very dangerous (an inappropriate) to run regressions on "slow turning" data either with or without trend and

(ii) unless you regress an I(0) dependent variable on I(0) explanatory variables your regression analysis will almost certainly <u>not</u> uncover significant relationships between the dependent variable and explanatory variables when, in fact, significant relationships exist and therefore

(iii) One has to be very careful in running multiple regressions on time series data!

A time series X_t is $I(0)$ if it <u>does not</u> have to be differenced in order to make it <u>stationary</u> and weakly dependent. A time series $\{X_t : t=1,2,\cdots\}$ is <u>stationary</u> when

(i) The unconditional mean of X_t, $E(X_t)$, is constant

(ii) The unconditional variance of X_t, $Var(X_t)$ is constant

(iii) for any $t, h \geq 1$, $Cov(X_t, X_{t+h})$ depends only on h and not t.

(p.361 Wooldridge)

Furthermore, a stationary time series process $\{X_t : 1,2,\cdots\}$ is <u>weakly dependent</u> if X_t and X_{t+h} are "almost independent" as $h \to \infty$.

On the other hand, if the time series X_t needs to be differenced as in $\Delta X_t = X_t - X_{t-1}$ in order to make it stationary and weakly dependent then we refer to the time series as being I(1).

As Exercise 9 notes, we can have spurious regressions if we regression a dependent variable that is I(1) on one or more explanatory variables that are I(1). Moreover, if we regress an I(1) dependent variable on an I(0) explanatory variable or vice versa we have an unbalanced equation that most likely will miss a significant relationship between variables unless the regression is run in balanced form.

For some examples of I(0) and I(1) variables see the SAS program Learn Unit Root.sas.

In the next lecture (24) we will discuss a test called the Augmented Dickey-Fuller (ADF) test that will allow use to distinguish when to take a difference in the data to make the series I(0) and when to leave it alone (see

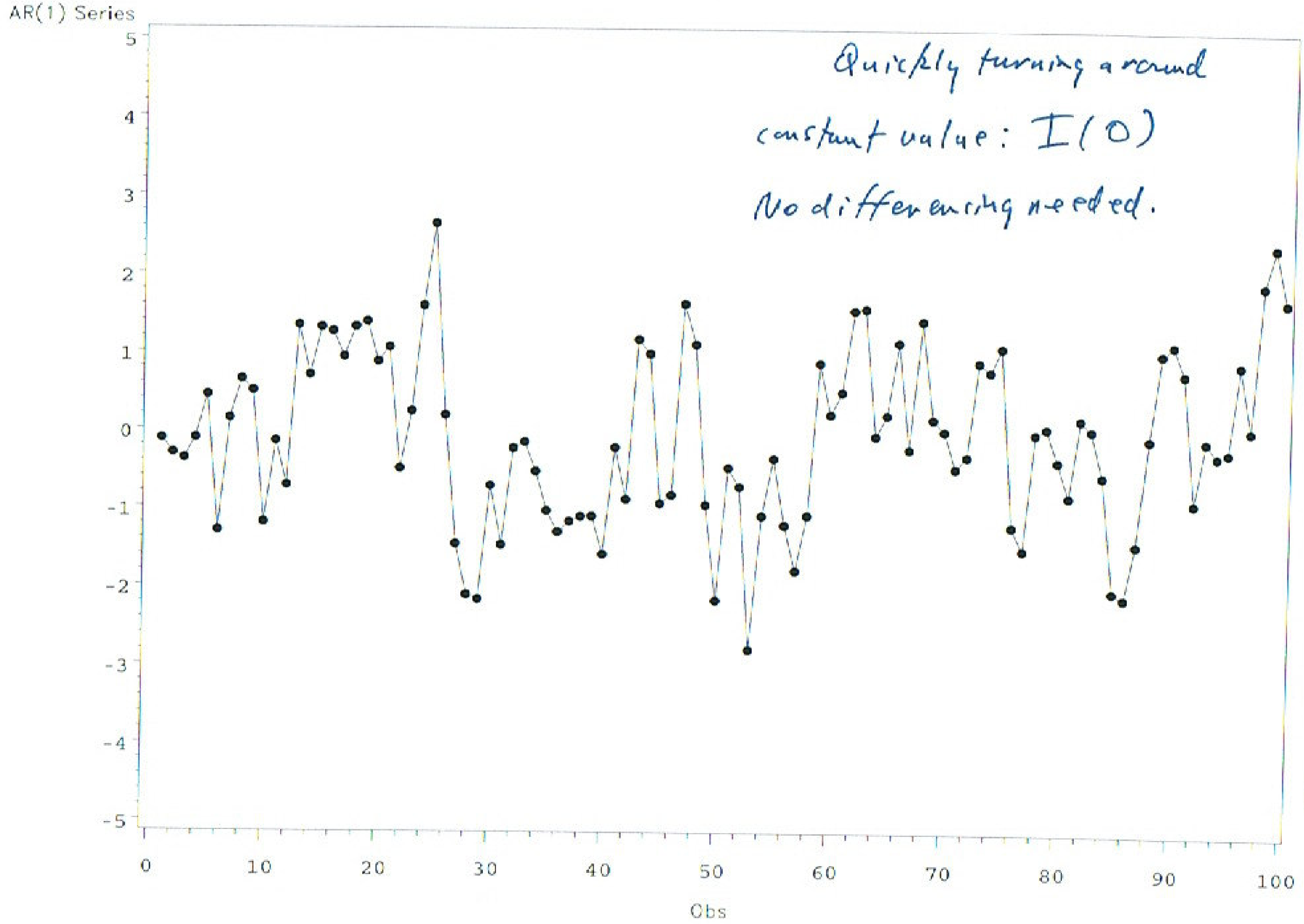
Figures 1 vs. 2 from LearnUnitRoot.sas) on the one hand or when to difference a series or to detrend it (see Figures 3 and 4).

Figures 1 – 4 from LearnUnitRoot.sas are reproduced on the following 4 pages.

Figure 1

Figures from
LearnUnitRoot.sas

Monte Carlo AR(1) data with phi(1) = 0.5
X=Time Y=AR(1) Series



5

Figure 2

Monte Carlo Random Walk Without Drift Data
X=Time Y=RW Series

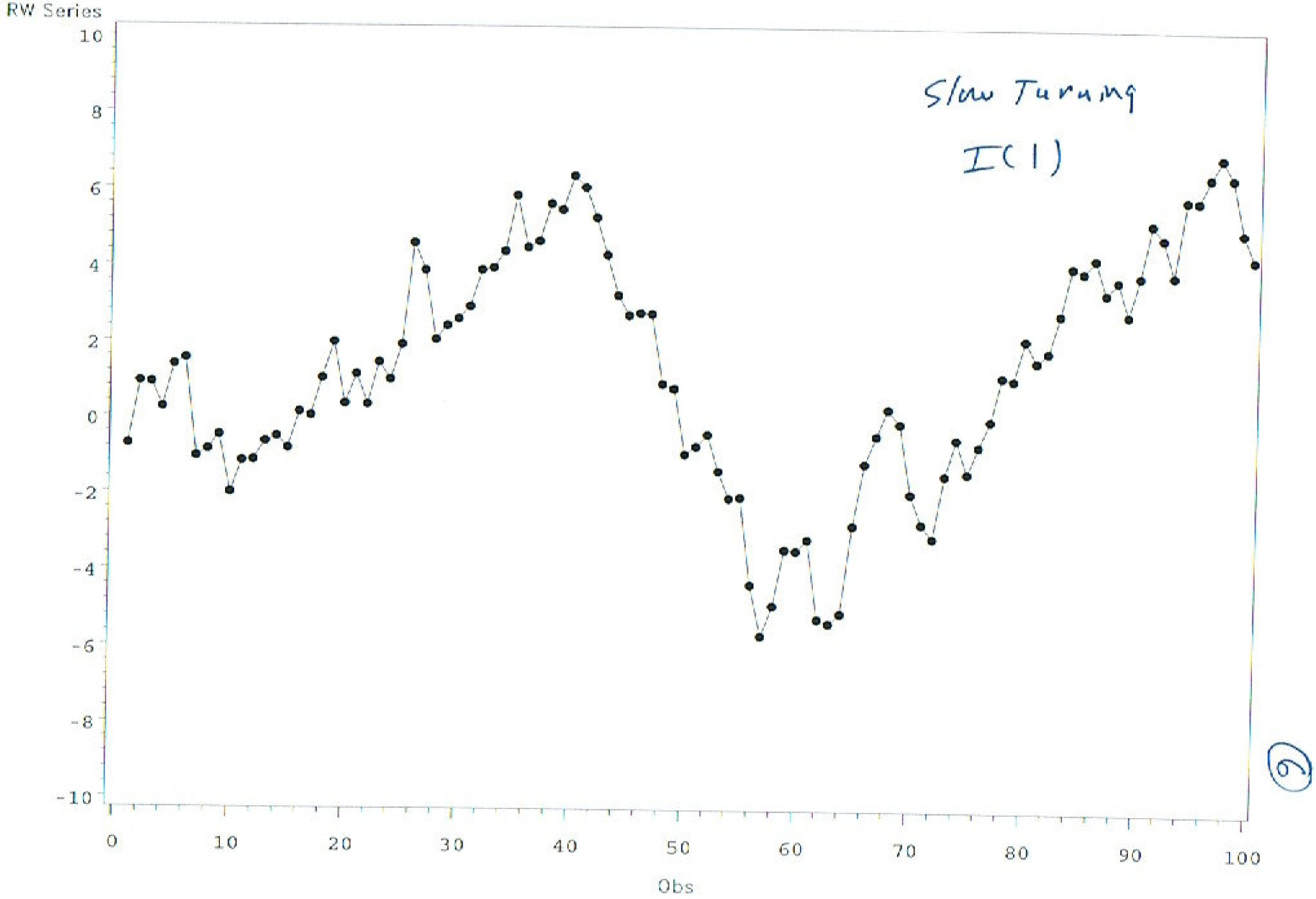


Figure 3

Monte Carlo Random Walk With Drift Data
X=Time Y=RW Series

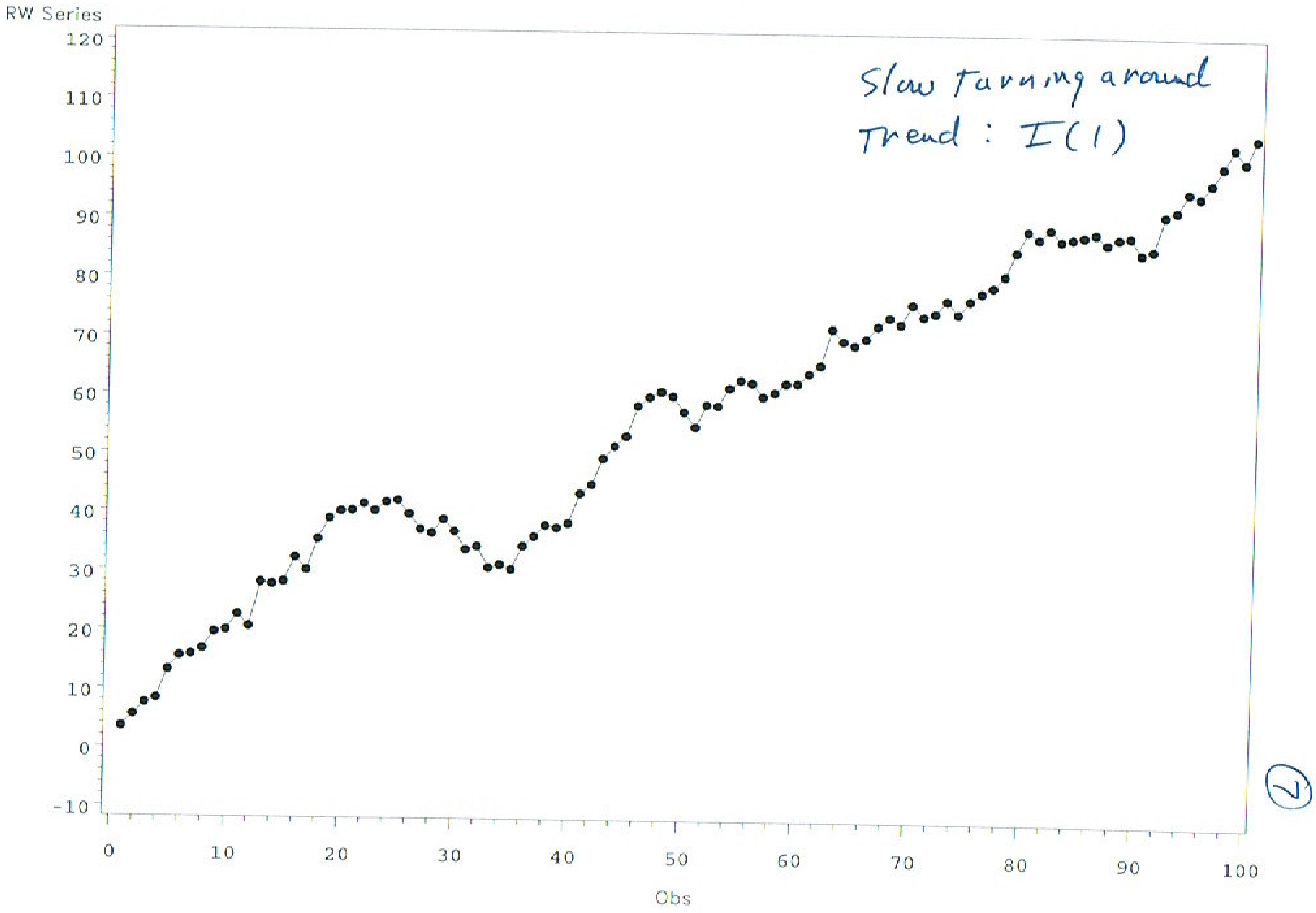
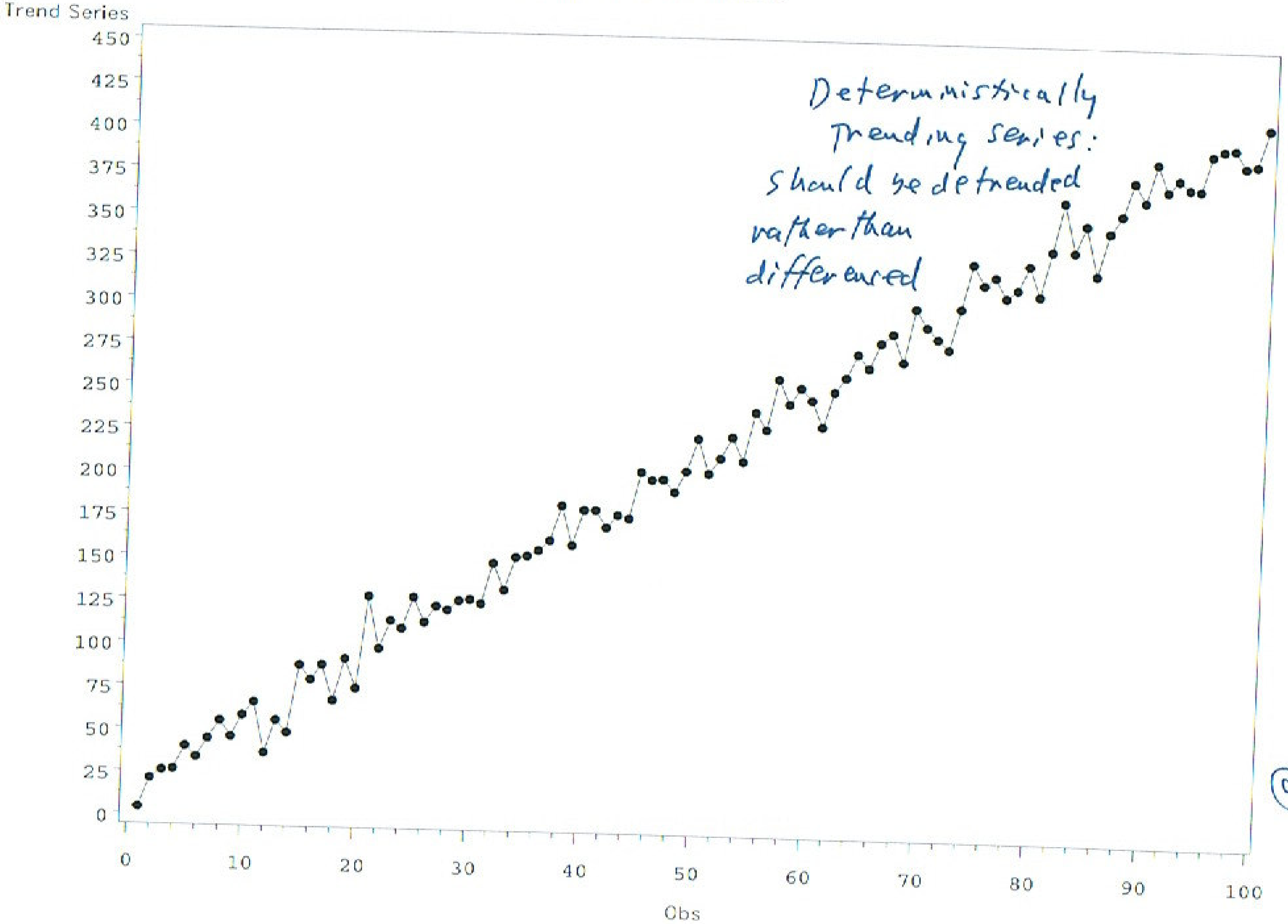


Figure 4

Deterministic Trend Data
X=Time Y=Trend Series



Now let us consider the following strategy for building a time series regression involving an explanatory variable. (The "Modern" Approach).

steps:

1) Transform dependent variable and all explanatory variables to stationary and weakly dependent form: Usually take difference of data or difference of natural logs of data.

2) If data needs to be differenced to make it stationary and weakly dependent, it is called I(1) data. If the data is stationary and weakly dependent as is, then it is called I(0) data. Before running your regression make sure that your regression is balanced in the sense that the dependent variable (maybe after differencing) is I(0) while

all of the explanatory variables are (possibly after differencing) are $I(0)$ as well.

3) Let $y_t \sim I(0)$ and $X_t \sim I(0)$.

First build an autoregressive model for y_t, namely,

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + u_t. \quad (1)$$

Determine the number of autoregressive terms to retain by reducing the lag p until the last lag term is statistically equivalent. Alternatively one could determine the order of p by minimizing a goodness-of-fit criterion like the Akaike Information Criterion (AIC) or the Schwartz Bayesian Criterion (SBC).

4) Now add "distributed lags" $X_t, X_{t-1}, \cdots, X_{t-r}$ to (1) above resulting in

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \beta_0 X_t + \beta_1 X_{t-1} + \cdots + \beta_r X_{t-r} + u_t. \qquad (2)$$

Hopefully we are able to choose the length of the distributed lag in (2) by the number of distributed lag terms that are statistically significant or that minimize the AIC or SBC goodness-of-fit criteria. Note that the first few terms of the distributed lag (β_0, β_1, etc.) may not be statistically significant and therefore should be dropped if there is a substantial delay in the effect that X_t has on y_t.

5) Make sure that your eq. (2) is "dynamically complete" (Wooldridge, p.380) in that the residuals u_t of eq. (2) are serially

uncorrelated at all lags (i.e. the residuals are "white noise"). To determine if (2) is dynamically complete and that the residuals of (2) are white noise we can use the Box-Pierce Q-statistic

$$Q = T \sum_{j=1}^{m} r_j^2 \qquad (3)$$

to test the null hypothesis that the residuals are white noise vs. the alternative hypothesis that they are not. Here r_j denotes the correlation between residuals j-periods apart. Under the null hypothesis that the residuals of (2) are white noise and thus that the equation is dynamically complete the Q-statistic of (3) is distributed as a χ^2_m random variable asymptotically. If the null hypothesis is rejected then we either have to add more autoregressive terms and/or more distributed lag terms until the residuals are white noise and thus that the model is dynamically complete. If none of this helps, then we probably need to go out and hunt for <u>another</u> explanatory variable with distributed that maybe will help make our model dynamically complete.

6) Even if the residuals of our model are uncorrelated they still might be heteroskedastic. If this is the case we can use heteroskedastic (White's) standard errors to properly adjust the t-statistics of our model <u>or</u> we can model the heteroskedasticity using the ARCH or GARCH specification (pp. 416-417 in Wooldridge) for which Robert Engle won the 2003 Nobel Prize in Economics.

To see the above modeling strategy in action see the EVIEWS program fertil3.wf1 and Example 10.4 (p. 338 in Wooldridge), Example 11.6 (p. 378) and Example 11.8 (p. 382).

(Compare and contrast the spurious regression result of Example 10.4 with the final regression model chosen in fertil3.wf1. The final model is balanced, dynamically complete (i.e. has white noise residuals), and all terms are statistically significant.

See the following pages.

Eq. 1

15

Dependent Variable: GFR
Method: Least Squares
Date: 11/23/03 Time: 15:37
Sample: 1 72
Included observations: 72

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	98.68176	3.208129	30.75991	0.0000
PE	0.082540	0.029646	2.784166	0.0069
WW2	-24.23840	7.458253	-3.249876	0.0018
PILL	-31.59403	4.081068	-7.741610	0.0000

R-squared	0.473415	Mean dependent var	95.63194
Adjusted R-squared	0.450184	S.D. dependent var	19.80464
S.E. of regression	14.68506	Akaike info criterion	8.265492
Sum squared resid	14664.27	Schwarz criterion	8.391973
Log likelihood	-293.5577	F-statistic	20.37801
Durbin-Watson stat	0.176873	Prob(F-statistic)	0.000000

This is spurious regression because GFR is I(1) and PE is I(1).

Eq. 2

Dependent Variable: CGFR
Method: Least Squares
Date: 11/23/03 Time: 15:38
Sample(adjusted): 2 72
Included observations: 71 after adjusting endpoints

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.784780	0.502040	-1.563182	0.1226
CPE	-0.042678	0.028367	-1.504469	0.1370

R-squared	0.031761	Mean dependent var	-0.835211
Adjusted R-squared	0.017729	S.D. dependent var	4.258743
S.E. of regression	4.220823	Akaike info criterion	5.745702
Sum squared resid	1229.259	Schwarz criterion	5.809439
Log likelihood	-201.9724	F-statistic	2.263427
Durbin-Watson stat	1.355472	Prob(F-statistic)	0.137025

Association between $CGFR = GFR_t - GFR_{t-1}$ and $CPE = PE_t - PE_{t-1}$ is much less than the correlation between GFR and PE in Eq. 1.

Eq. 3

(17)

Dependent Variable: CGFR
Method: Least Squares
Date: 11/26/04 Time: 14:23
Sample(adjusted): 6 72
Included observations: 67 after adjusting endpoints

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.403923	0.514947	-0.784398	0.4358
CGFR_1	0.352397	0.124798	2.823743	0.0064
CGFR_2	-0.242193	0.130045	-1.862375	0.0673
CGFR_3	0.209921	0.129949	1.615404	0.1113
CGFR_4	0.173825	0.124382	1.397509	0.1672

R-squared	0.212547	Mean dependent var	-0.829851
Adjusted R-squared	0.161744	S.D. dependent var	4.366679
S.E. of regression	3.997971	Akaike info criterion	5.681146
Sum squared resid	990.9938	Schwarz criterion	5.845676
Log likelihood	-185.3184	F-statistic	4.183727
Durbin-Watson stat	2.066959	Prob(F-statistic)	0.004604

starting to build the
Autoregressive core part of
time series regression

Eq. 4

Dependent Variable: CGFR
Method: Least Squares
Date: 11/26/04 Time: 14:23
Sample(adjusted): 4 72
Included observations: 69 after adjusting endpoints

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.721171	0.509080	-1.416617	0.1613
CGFR_1	0.346188	0.120606	2.870413	0.0055
CGFR_2	-0.195280	0.120264	-1.623761	0.1092

R-squared	0.119219	Mean dependent var	-0.863768
Adjusted R-squared	0.092529	S.D. dependent var	4.307073
S.E. of regression	4.102973	Akaike info criterion	5.703805
Sum squared resid	1111.070	Schwarz criterion	5.800941
Log likelihood	-193.7813	F-statistic	4.466759
Durbin-Watson stat	1.891651	Prob(F-statistic)	0.015158

using 5% level of significance the order of the Autoregressive core is one (with only CGFR-1).

Eq. 5

Dependent Variable: CGFR
Method: Least Squares
Date: 11/26/04 Time: 14:25
Sample(adjusted): 6 72
Included observations: 67 after adjusting endpoints

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.589532	0.477586	-1.234401	0.2219
CGFR_1	0.332987	0.122426	2.719917	0.0085
CPE	-0.055431	0.026898	-2.060780	0.0437
CPE_1	0.008585	0.027402	0.313282	0.7552
CPE_2	0.099322	0.026825	3.702543	0.0005
CPE_3	0.009671	0.029546	0.327325	0.7446
CPE_4	-0.044016	0.027148	-1.621289	0.1102

R-squared	0.345489	Mean dependent var	-0.829851
Adjusted R-squared	0.280038	S.D. dependent var	4.366679
S.E. of regression	3.705152	Akaike info criterion	5.555933
Sum squared resid	823.6892	Schwarz criterion	5.786274
Log likelihood	-179.1238	F-statistic	5.278587
Durbin-Watson stat	1.908406	Prob(F-statistic)	0.000201

Now we add on the distributed lag part of the equation (CPE, CPE_1, ..., CPE_4).

Eq. 6

Dependent Variable: CGFR
Method: Least Squares
Date: 11/26/04 Time: 14:26
Sample(adjusted): 4 72
Included observations: 69 after adjusting endpoints

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.702159	0.453799	-1.547292	0.1267
CGFR_1	0.300242	0.105903	2.835056	0.0061
CPE	-0.045472	0.025642	-1.773367	0.0809
CPE_1	0.002064	0.026778	0.077080	0.9388
CPE_2	0.105135	0.025590	4.108366	0.0001

R-squared	0.318113	Mean dependent var	-0.863768
Adjusted R-squared	0.275495	S.D. dependent var	4.307073
S.E. of regression	3.666090	Akaike info criterion	5.505832
Sum squared resid	860.1737	Schwarz criterion	5.667724
Log likelihood	-184.9512	F-statistic	7.464282
Durbin-Watson stat	1.941419	Prob(F-statistic)	0.000053

Building the distributed lag part of model. Looking for significant lagged terms at, say, the 5% level of significance.

Eq. 7

Dependent Variable: CGFR
Method: Least Squares
Date: 11/26/04 Time: 14:27
Sample(adjusted): 4 72
Included observations: 69 after adjusting endpoints

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.702281	0.450313	-1.559541	0.1237
CGFR_1	0.298518	0.102720	2.906144	0.0050
CPE	-0.044947	0.024532	-1.832195	0.0715
CPE_2	0.105629	0.024583	4.296814	0.0001

R-squared	0.318049	Mean dependent var	-0.863768
Adjusted R-squared	0.286575	S.D. dependent var	4.307073
S.E. of regression	3.637949	Akaike info criterion	5.476940
Sum squared resid	860.2535	Schwarz criterion	5.606453
Log likelihood	-184.9544	F-statistic	10.10493
Durbin-Watson stat	1.938759	Prob(F-statistic)	0.000015

We need to drop the contemporaneous distributed lag term CPE because (i) it is of the wrong sign (we expect the effect to be positive) and (ii) it is not statistically significant at the 5% level.

Eq. 8

Dependent Variable: CGFR
Method: Least Squares
Date: 11/26/04 Time: 14:27
Sample(adjusted): 4 72
Included observations: 69 after adjusting endpoints

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.771480	0.456668	-1.689367	0.0959
CGFR_1	0.284249	0.104237	2.726952	0.0082
CPE_2	0.106962	0.025007	4.277254	0.0001

R-squared	0.282830	Mean dependent var	-0.863768
Adjusted R-squared	0.261097	S.D. dependent var	4.307073
S.E. of regression	3.702336	Akaike info criterion	5.498310
Sum squared resid	904.6815	Schwarz criterion	5.595445
Log likelihood	-186.6917	F-statistic	13.01418
Durbin-Watson stat	1.924444	Prob(F-statistic)	0.000017

The Final Equation!

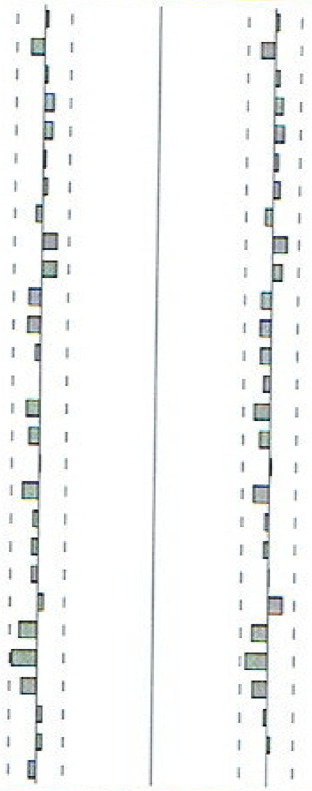
Interpretation:
There is a tendency for the change in the fertility rate to be negative (i.e. decline over time in GFE) with the declines being positively related (i.e. above average decline followed by above average decline; below average decline followed by below average decline) while the change in personal exemption has a two period delayed effect (takes time to recognize change in law and to have a child.)

Residual Analysis of Eq. 8

Correlogram of Residuals

Date: 11/26/04 Time: 14:40
Sample: 4 72
Included observations: 69

Autocorrelation | Partial Correlation



	AC	PAC	Q-Stat	Prob
1	0.034	0.034	0.0825	0.774
2	-0.122	-0.124	1.1788	0.555
3	0.038	0.048	1.2862	0.732
4	0.098	0.081	2.0071	0.734
5	0.090	0.095	2.6234	0.758
6	0.029	0.043	2.6878	0.847
7	0.047	0.061	2.8601	0.898
8	-0.053	-0.066	3.0884	0.929
9	0.136	0.138	4.5912	0.868
10	0.134	0.096	6.0791	0.809
11	-0.108	-0.098	7.0665	0.794
12	-0.120	-0.106	8.3013	0.761
13	-0.043	-0.094	8.4619	0.812
14	0.005	-0.058	8.4643	0.864
15	-0.117	-0.140	9.7057	0.838
16	-0.097	-0.097	10.584	0.834
17	0.015	0.024	10.604	0.876
18	-0.147	-0.147	12.678	0.810
19	-0.047	-0.036	12.893	0.844
20	-0.055	-0.048	13.196	0.869
21	-0.057	0.016	13.529	0.889
22	0.052	0.138	13.808	0.908
23	-0.161	-0.145	16.568	0.830
24	-0.226	-0.201	22.145	0.571
25	-0.137	-0.136	24.238	0.506
26	0.057	-0.026	24.607	0.541
27	0.051	0.041	24.907	0.580
28	-0.063	0.006	25.383	0.607

Showing the residuals of Eq. 8 are white noise. All Q-statistics have probability values above 0.05 no matter what lag (m) we choose.

Eq. 9 WW2 and Pill are no longer significant!

Dependent Variable: CGFR
Method: Least Squares
Date: 11/26/04 Time: 14:28
Sample(adjusted): 4 72
Included observations: 69 after adjusting endpoints

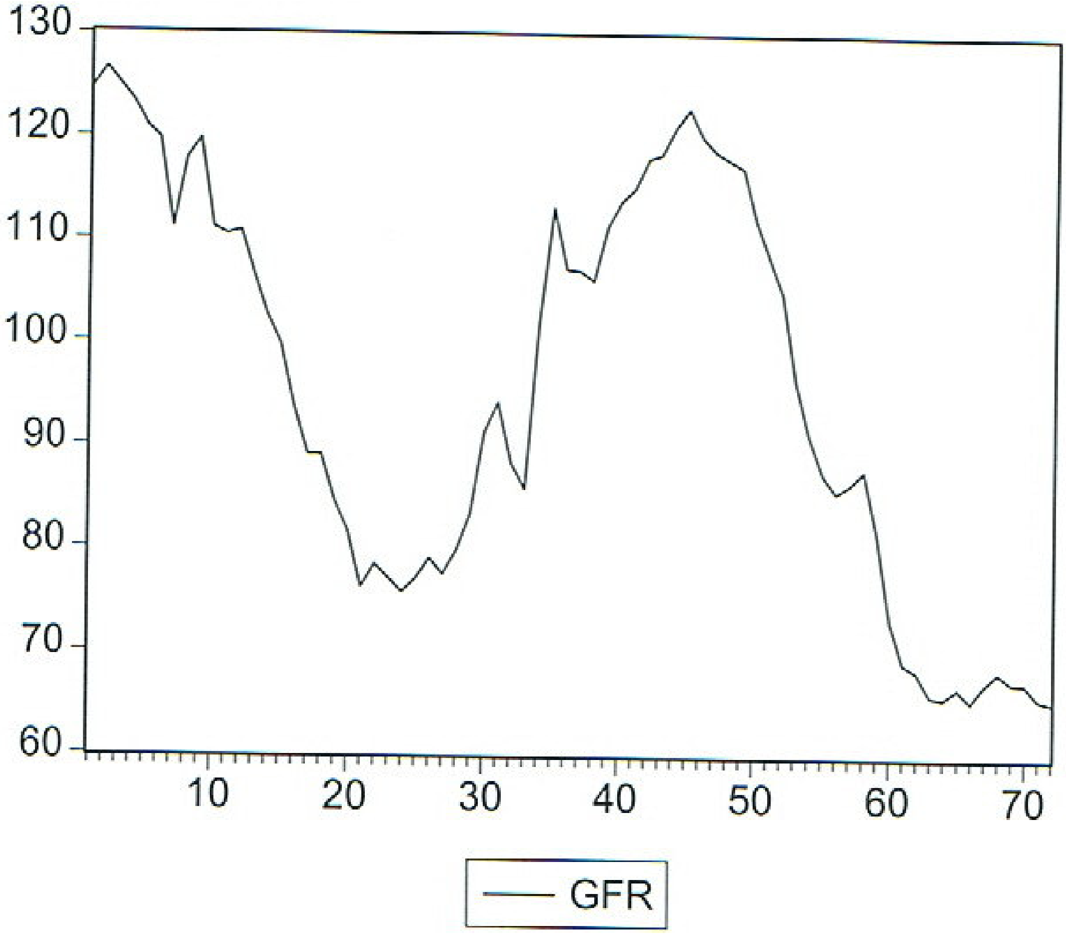
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.441621	0.578834	-0.762949	0.4483
CGFR_1	0.281678	0.109783	2.565766	0.0126
CPE_2	0.110304	0.026804	4.115184	0.0001
WW2	-1.370988	1.879007	-0.729634	0.4683
PILL	-0.745572	1.009490	-0.738563	0.4629

R-squared	0.293560	Mean dependent var	-0.863768
Adjusted R-squared	0.249408	S.D. dependent var	4.307073
S.E. of regression	3.731507	Akaike info criterion	5.541206
Sum squared resid	891.1455	Schwarz criterion	5.703098
Log likelihood	-186.1716	F-statistic	6.648780
Durbin-Watson stat	1.931026	Prob(F-statistic)	0.000154

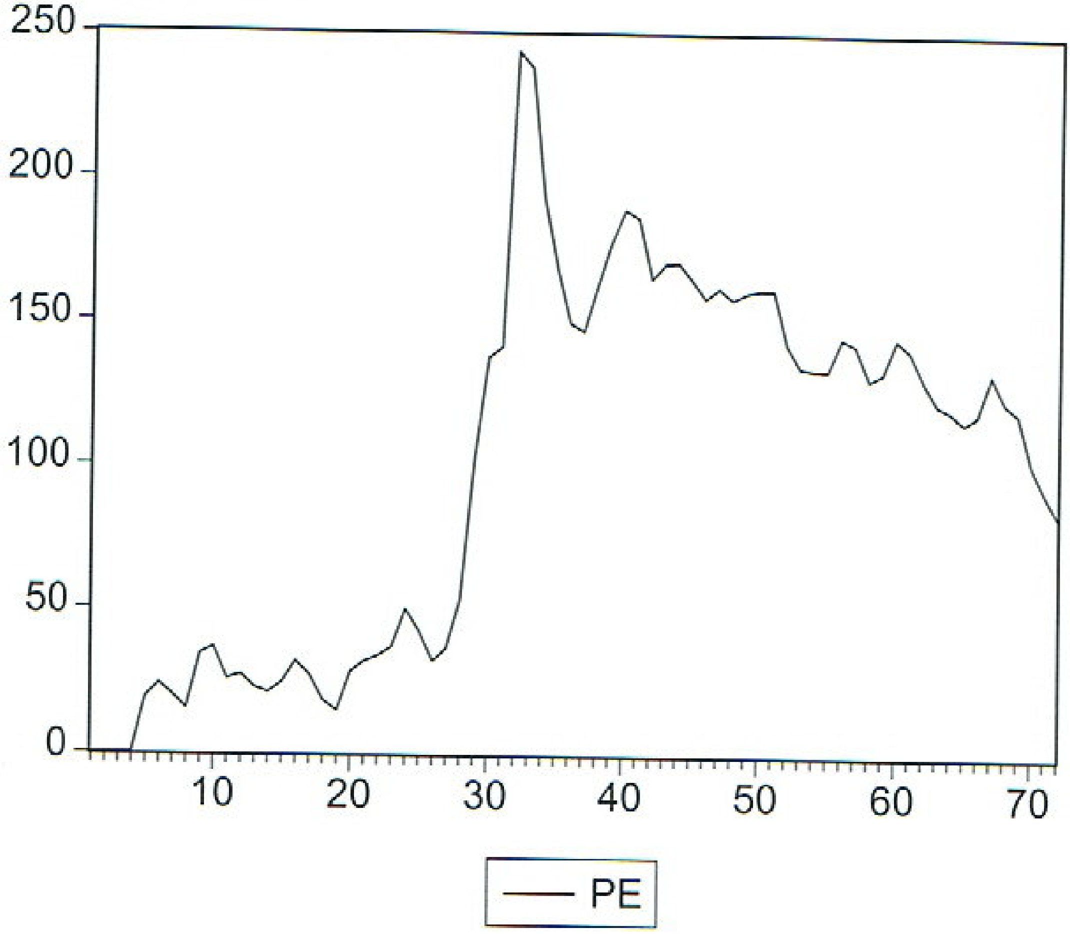
When we try to add the pulse dummy WW2 and the step dummy Pill to our equation we find them to be statistically insignificant unlike what we thought was the case in the spurious regression (Eq. 1). A joint test of the significance of WW2 and Pill, not reported here, indicates that they are jointly insignificant as well.

25

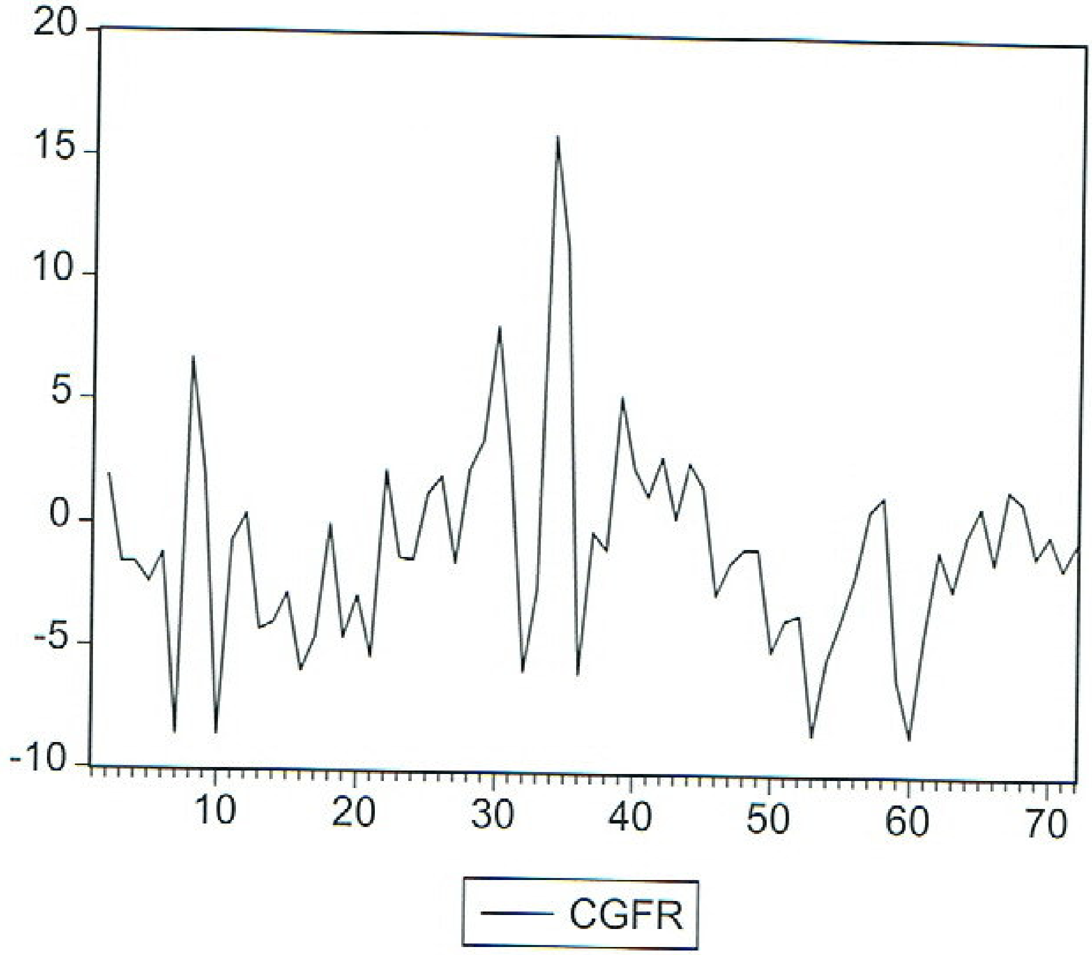
Case 3 ADF test
should be used



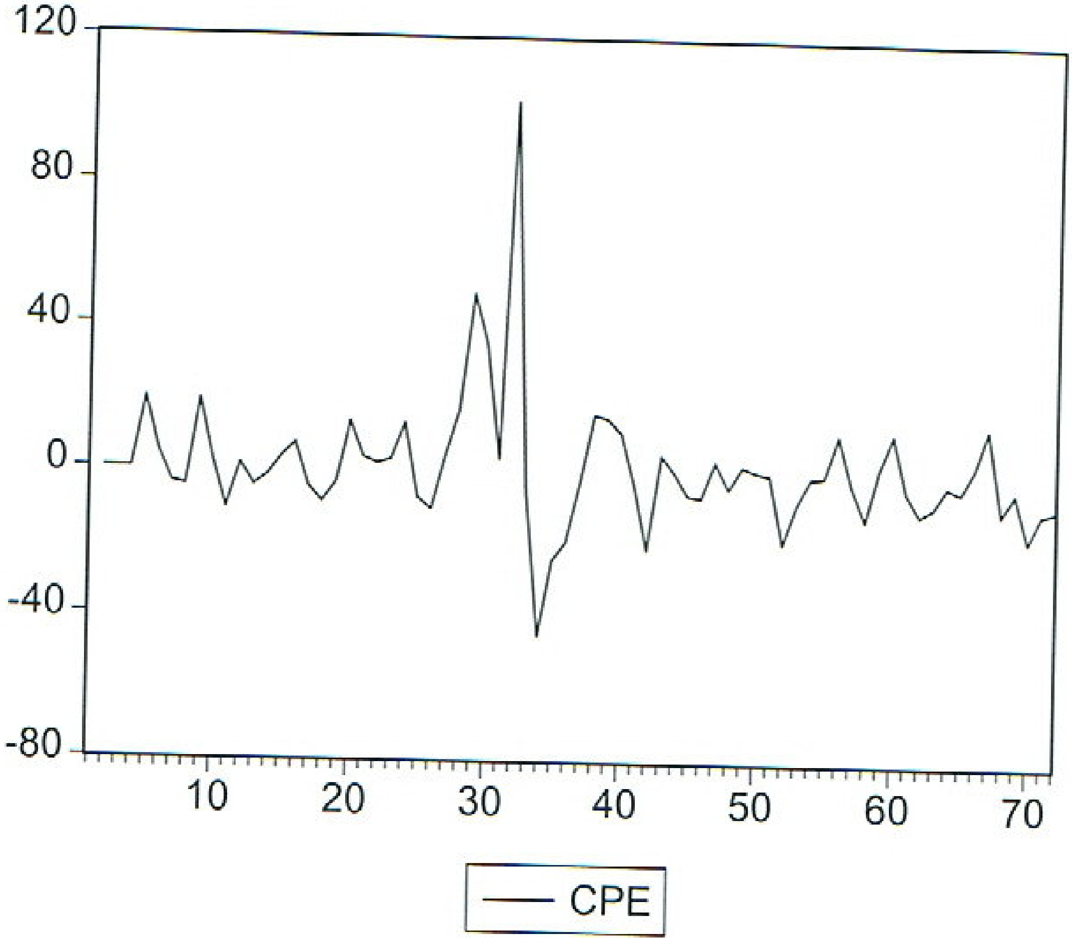
Case 2 ADF test
should be used



After Differencing GFR it is now stationary and weakly dependent

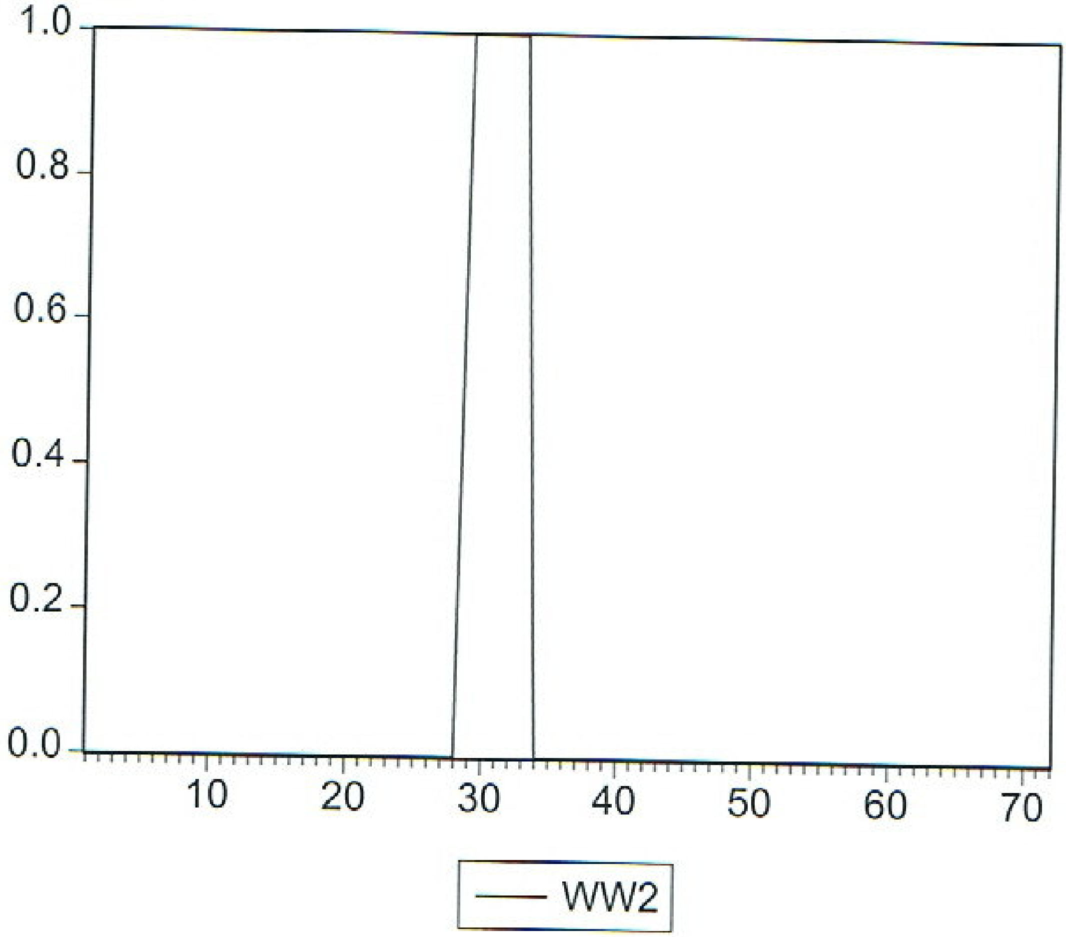


After Differencing PE
it is now stationary and
weakly dependent
(at least approximately)



Pulse

World War II/dummy



"Pill" step dummy