DSC5211C QUANTITATIVE RISK MANAGEMENT

SESSION 3

Fernando Oliveira
bizfmdo@nus.edu.sg

Regression Analysis and Stationarity in Time Series

Objectives

- Analysis of Regression in Time Series
- The Problem of Autocorrelation
- Durbin-Watson and Autocorrelation
- The Problem of Heteroscedasticity
- Regression with Trending Series
- Testing for Stationarity: Dickey-Fuller Tests

Readings: Enders, Ch. 4
Franses, Ch. 4.2, 4.3
Hanke, Ch. 6, 7, 8.

Simple Linear Regression

- Simple regression estimates and tests a relationship between two variables in a simple linear model

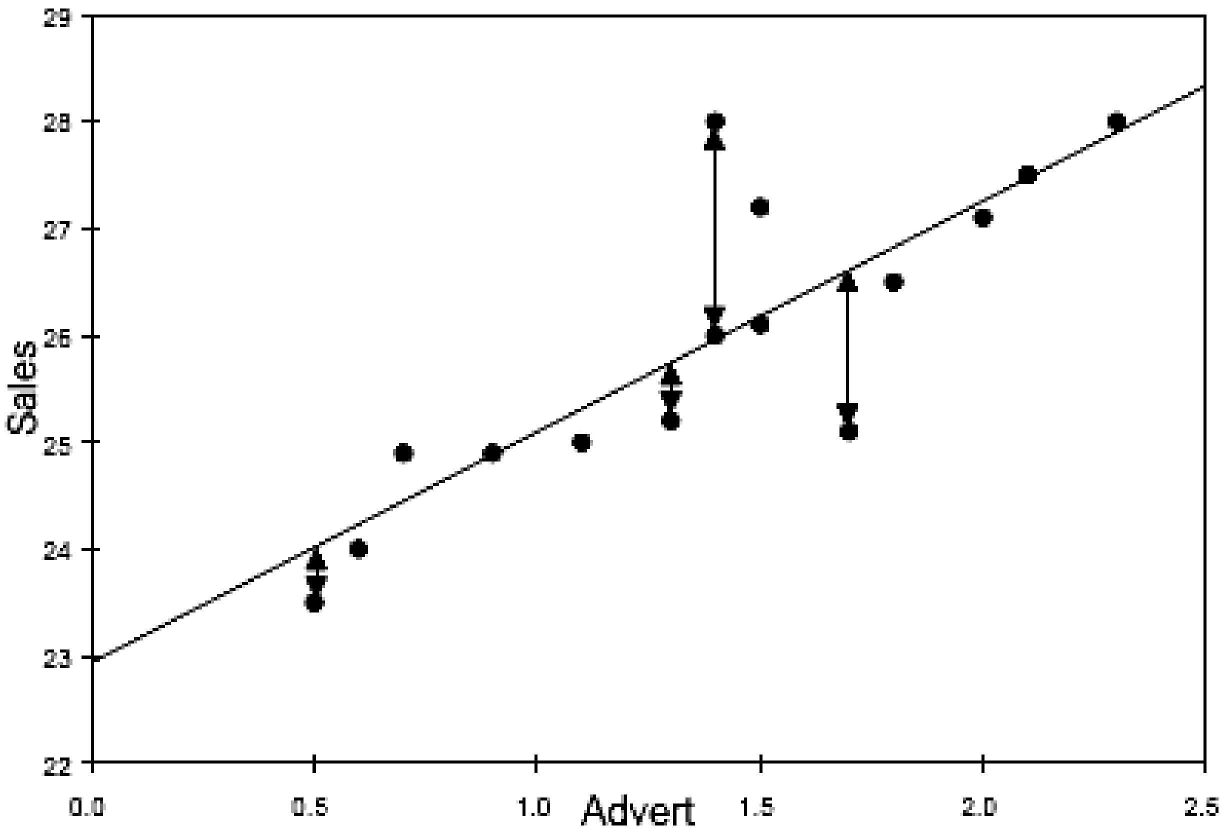
$$y = a + bx + u$$

- Regression model is simply an equation of a line

$$Sales = a + \underbrace{b \cdot Advertising}_{\text{systematic component}} + \underbrace{error}_{\text{random component}}$$

- What are the sales when advertising is zero?
- How much increase in sales from each pound spent on advertising?

Residuals: Graph



Determining the Best Line

- Sales= a+ b ·Advertising + error
- How do we determine the constants ***a*** and ***b***? •

- We choose a line so that residual scatter is minimised:
  1. Average error is zero (the mean of the residuals is zero)
  2. Average squared error is as small as possible. “Least squares” is traditionally used.

- Find the line that minimises the sum of squared residuals (equivalent to minimising residual standard deviation).

$$\min \sum_{i=1}^{n} e_i^2$$

Finding the Best Line

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.83
R Square	0.70
Adjusted R Square	0.67
Standard Error	0.81
Observations	15

	Coeffts	*Std Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	22.94	0.59	39.13	0.000000	21.67	24.21
Advert	2.16	0.39	5.47	0.000108	1.31	3.01

$$Sales = 22.94 + 2.16 Advertising + error$$

Model Building Methodology

- *Graphs and correlation analysis*
  - plot to get a feel for relationship & spot problems
  - correlations measure strength of relationship

- *Model estimation*

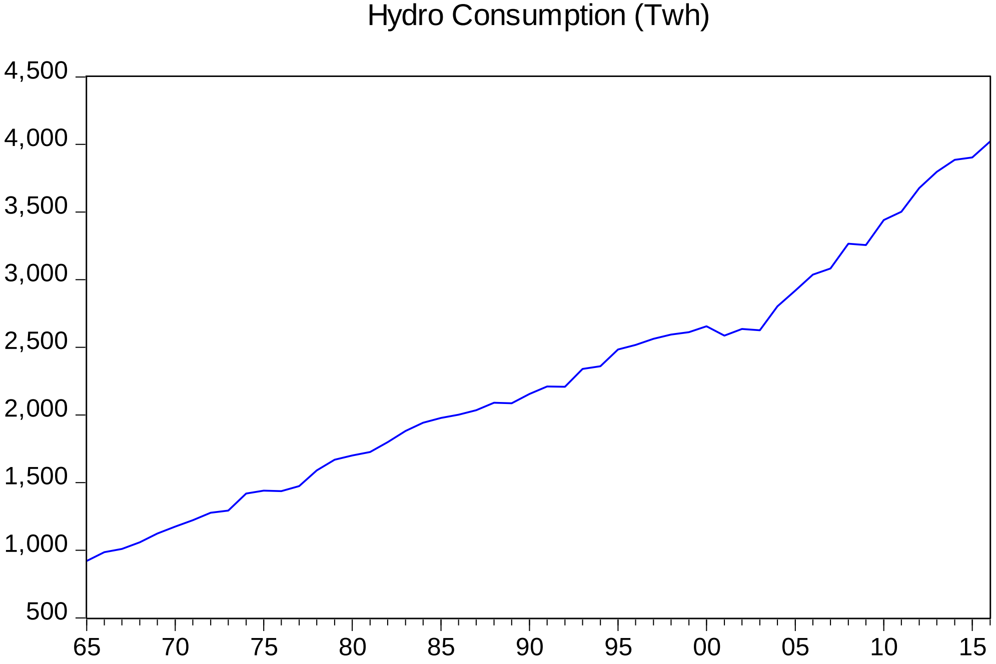
- *Diagnostic evaluation*

  - evaluate validity:
    1. Testing for significance of relationship
    2. Checking if has anything been left out of model

  - evaluate usefulness
    1. How close does model fit data

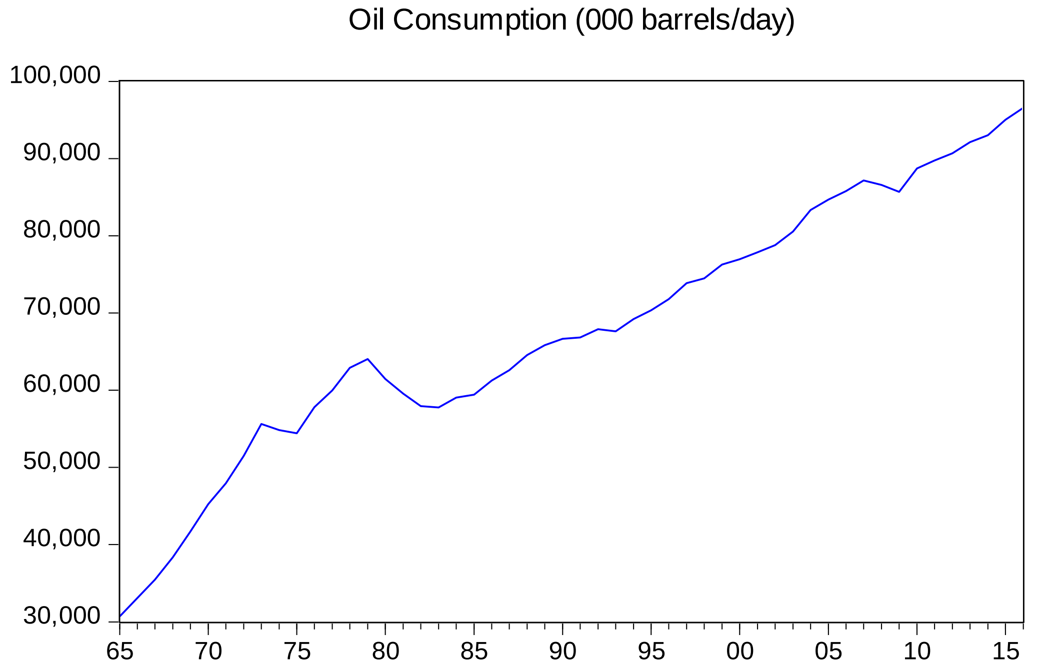
Example: Hydroelectric Consumption

- Source: BP.com
- Frequency: yearly
- Units: TWh
- Date Range: 1965-2016



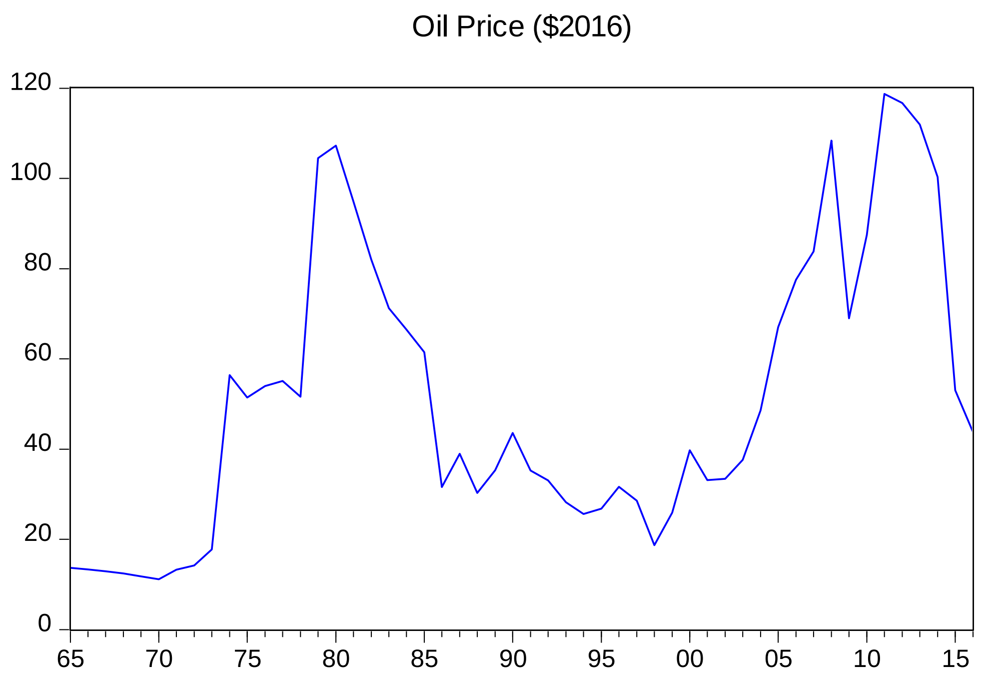
Example: Crude Oil Consumption

- Source: BP.com
- Frequency: yearly
- Units: Thousand Barrels Daily
- Date Range: 1965-2016



Example: Oil Price

- Source: BP.com
- Frequency: yearly
- Units: $/bbl (2016 base)
- Date Range: 1965-2016



Multiple Regression

- Multiple regression estimates and tests a relationship between two variables

$$Y_t = a_0 + a_1 X_t + a_2 W_t + \ldots + e_t$$

- Regression model is simply an equation of a line

$$Hydro_t = 173.7 + 1.09 Gas_t - 0.06 Nuc_t + e_t$$

- Comment on the parameters.

EViews Output

Dependent Variable: Hydro consumption

Sample: 1965 2016
Included observations: 52

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	173.7243	38.43427	4.520036	0.0000
GAS	1.094352	0.038018	28.78477	0.0000
NUCLEAR	-0.063017	0.030536	-2.063697	0.0444

R-squared	0.989253	Mean dependent var	2259.313
Adjusted R-squared	0.988815	S.D. dependent var	857.0553
S.E. of regression	90.64218	Akaike info criterion	11.90768
Sum squared resid	402584.3	Schwarz criterion	12.02025
Log likelihood	-306.5996	Hannan-Quinn criter.	11.95084
F-statistic	2255.302	Durbin-Watson stat	0.390188
Prob(F-statistic)	0.000000		

Residual Assumptions

- Quality of parameter estimates & validity of sig. tests rely upon residuals ~ N(0,σ)

- Residuals must be
  - normally distributed
  - independent (no autocorrelation)
  - same variance (no heteroscedasticity)

- Histogram and plots of residuals enables inspection of assumptions

- Residuals should be simple randomness remaining after deterministic part of variation in *the dependent variable* has been modelled

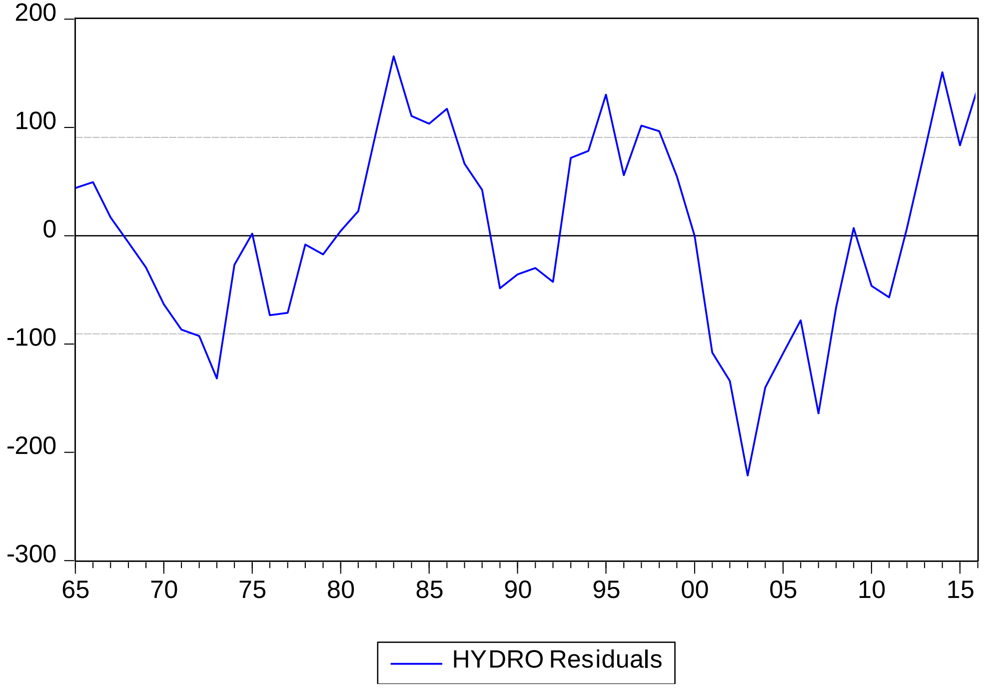
- Any systematic component in the errors should be in the model

Multi-collinearity

- Multi-collinearity means high degree of correlation between two or more explanatory variables
- Result is unreliable coefficient estimates, reflected in big uncertainty on coefficients, i.e. large standard errors, and thus small t-stats
- Prediction may be OK but description will not be
- Conclusion: Check correlations between explanatory variables!

Residuals: Hydroelectric Consumption

- The plot shows that there is a pattern in the residuals: the volatility is increasing over time.



- Heteroscedasticity: the variance of the residuals is not constant

Heteroscedasticity

- The variability in some time series tends to increase with the level of the series.

- The variability increases if the variable is growing at a constant rate than at a constant amount over time.

- The variance of the error term, ε_t, is not constant.

- The standard error of the estimates underestimates the current standard deviation of the error term. These limits are too narrow for using in confidence intervals.

Heteroscedasticity: Solving the Problem

- Sometimes this problem can be solved by simply transforming the data.

- Possible transformations:
  - Transform data: convert current dollars into constant dollars.
  - Use the log linear model

$$Log(Hydro_t) = 4.56 + 1.13*Log(Gas_t) – 0.5\ Log(oil_t) + 0.03\ Log(Oil\_price_t) + e_t$$

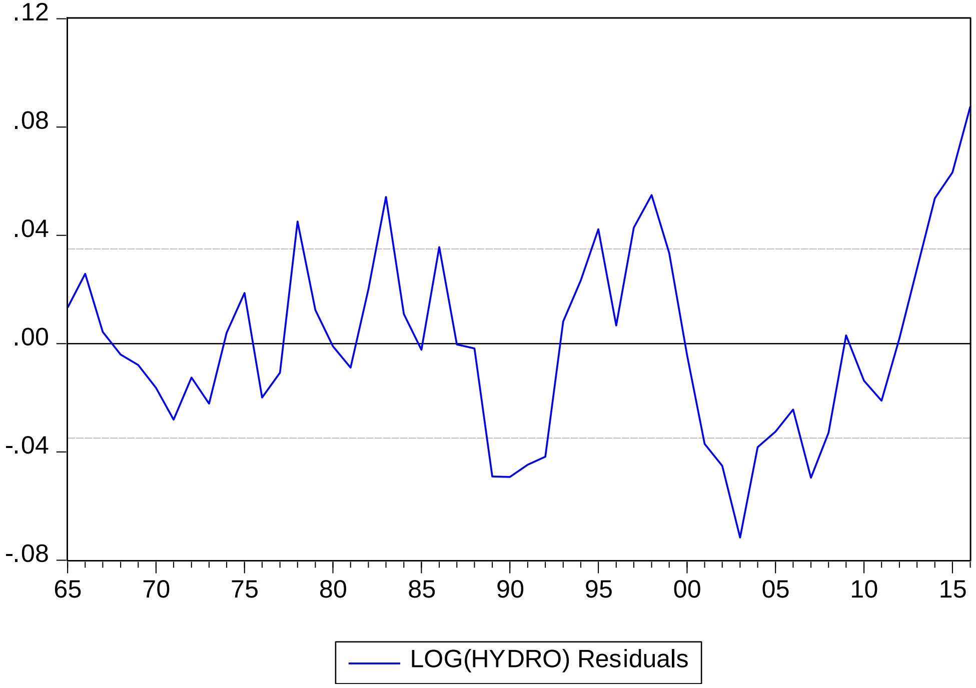
Dependent Variable: LOG(HYDRO)
Method: Least Squares

Sample: 1965 2016
Included observations: 52

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.557423	0.669876	6.803382	0.0000
LOG(GAS)	1.134748	0.055226	20.54723	0.0000
LOG(OIL)	-0.500301	0.097508	-5.130857	0.0000
LOG(OIL_PRICE)	0.032628	0.009695	3.365325	0.0015

R-squared	0.992872	Mean dependent var	7.647408
Adjusted R-squared	0.992427	S.D. dependent var	0.401853
S.E. of regression	0.034971	Akaike info criterion	-3.794820
Sum squared resid	0.058701	Schwarz criterion	-3.644725
Log likelihood	102.6653	Hannan-Quinn criter.	-3.737277
F-statistic	2228.806	Durbin-Watson stat	0.562954
Prob(F-statistic)	0.000000		

Heteroscedasticity: Log linear Model



- The volatility is more homogeneous but at the same time there seems to be large fluctuations. How to interpret this result?

Residual Autocorrelation

- Autocorrelation exists when successive observations over time are related to one another.

- Autocorrelation can occur because the effect of an independent variable on the dependent variable is distributed over time.

- In a regression model autocorrelation is handled by correcting the model:
    - Change the functional form of the model
    - Change the independent variables
    - Changing the form of the error term (Analysed later in the course)

Durbin-Watson Statistic

- *DW* statistic evaluates autocorrelation for residuals placed in same order as the observations. This is generally only interesting if this is time order.

- $DW \approx 2(1\text{-}r)$ where r is autocorrelation
  $0 < DW < 4$
  $DW = 2$ no autocorrelation
  $DW > 2$ negative autocorrelation
  $DW < 2$ positive autocorrelation
  1.5 < Rule of thumb < 2.5: no autocorrelation

- *DW* statistic used instead of *r* because strict tests exist to examine whether *DW* is sig. diff. from 2
- EViews gives *DW*, whilst Excel does not

Residual Autocorrelation: Durbin-Watson Statistic

- *DW is a test for AR(1) serial correlation*.
- *DW* is based on the residuals of the Ordinary Least Squares
- *DW* assumes normally distributed errors
- *DW* is not valid in the presence of heteroscedasticity
- *DW* is not valid in the presence of lagged dependent variables

Autocorrelation: Hydroelectric Consumption

- We have positive autocorrelation, as the DW is 0.56.

- We need to correct the model.

Regression with Trending Series

- Trending variables are likely to have high correlation and lead to a regression with high R^2. This is true regardless of whether they are related.

- Autocorrelation in residuals suggests that the regression is spurious (nonsense)

- Using *differenced* variables (changes) eliminates the trends and thus avoids spurious regression

    - $\Delta Y_t = Y_t - Y_{t-1}$
    - $D\log(Y_t) \sim (Y_t - Y_{t-1})/Y_t$

- Convention is to difference all trending variables used in a regression

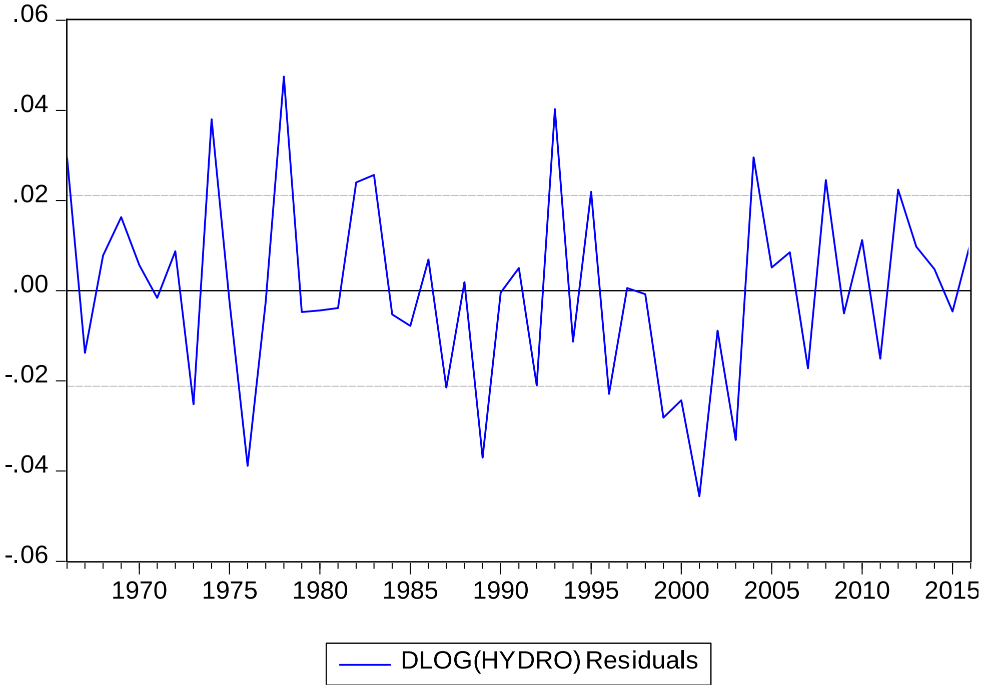
Hydroelectric Consumption – New Model

Dependent Variable: DLOG(HYDRO)
Method: Least Squares

Sample (adjusted): 1966 2016
Included observations: 51 after adjustments

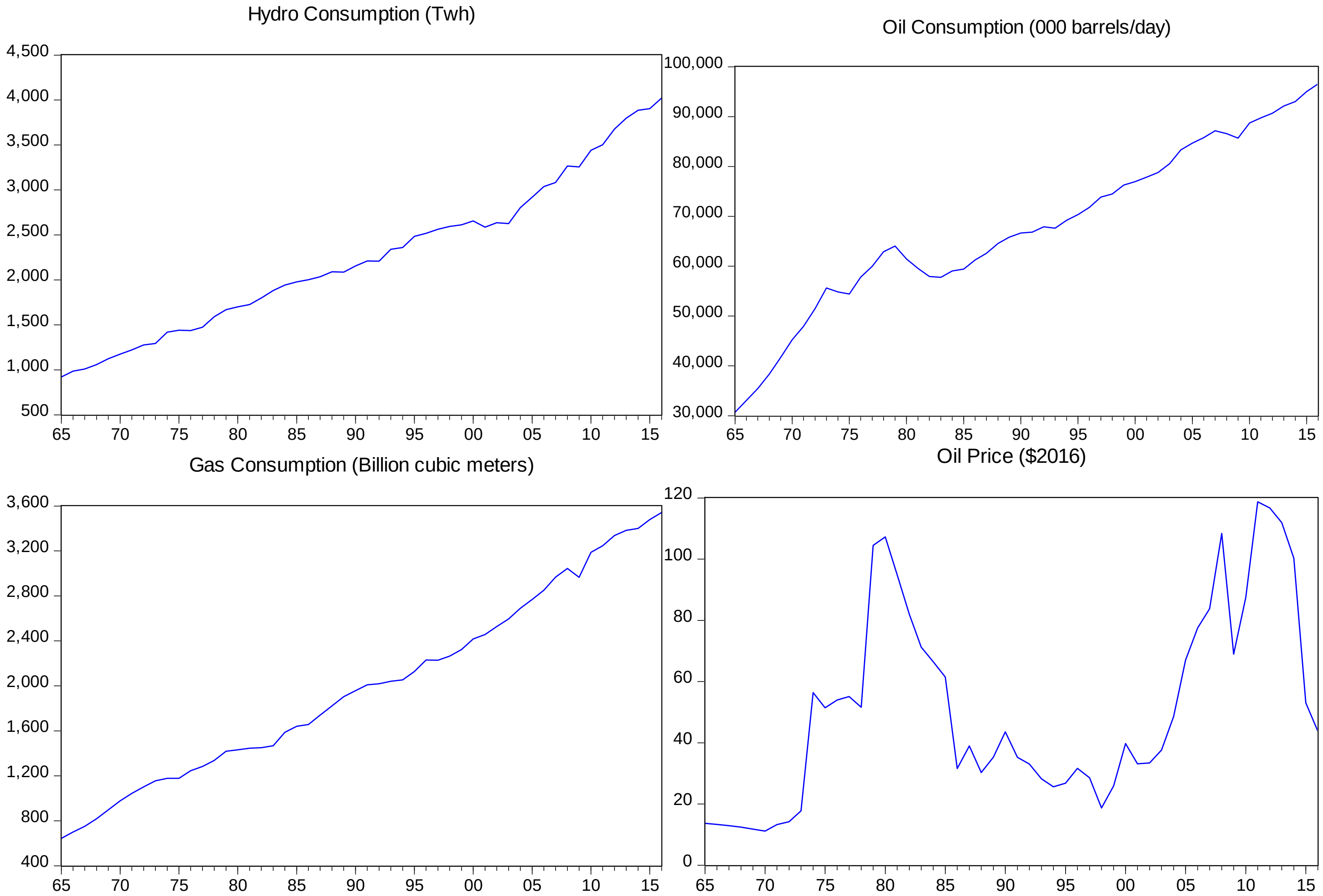
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.019977	0.004904	4.073214	0.0002
DLOG(GAS)	0.249770	0.118389	2.109741	0.0401
DLOG(OIL_PRICE)	0.026221	0.010532	2.489751	0.0163

R-squared	0.228059	Mean dependent var	0.028931
Adjusted R-squared	0.195895	S.D. dependent var	0.023602
S.E. of regression	0.021164	Akaike info criterion	-4.815964
Sum squared resid	0.021501	Schwarz criterion	-4.702328
Log likelihood	125.8071	Hannan-Quinn criter.	-4.772540
F-statistic	7.090456	Durbin-Watson stat	2.240440
Prob(F-statistic)	0.002005		



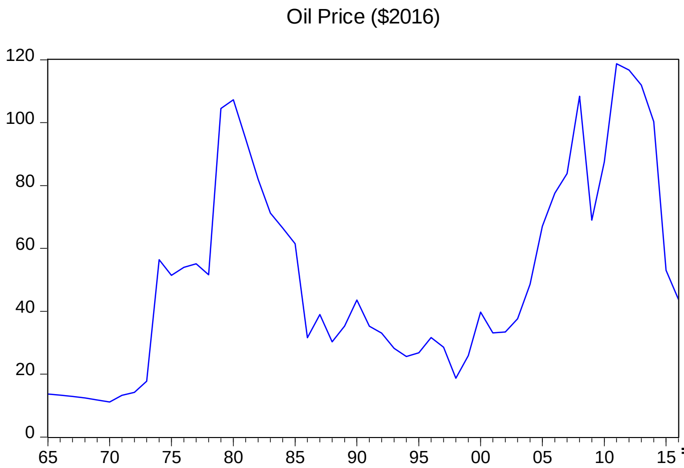
- Is the autocorrelation problem solved?
- How good is the model?
- What is the interpretation of the results?

Nonstationary Series



Testing for Stationarity: Correlogram

- Autocorrelation plot is known as the correlogram. It is very useful for analysing time series.



Sample: 1965 2016
Included observations: 52

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.860	0.860	40.694	0.000
		2	0.709	-0.113	68.964	0.000
		3	0.576	-0.022	87.979	0.000
		4	0.428	-0.148	98.672	0.000
		5	0.324	0.083	104.95	0.000
		6	0.182	-0.262	106.96	0.000
		7	-0.004	-0.257	106.97	0.000
		8	-0.142	-0.004	108.25	0.000
		9	-0.268	-0.102	112.95	0.000
		10	-0.326	0.120	120.05	0.000
		11	-0.363	-0.105	129.08	0.000
		12	-0.408	-0.022	140.75	0.000
		13	-0.422	-0.027	153.59	0.000
		14	-0.428	-0.065	167.15	0.000
		15	-0.374	0.141	177.79	0.000
		16	-0.313	-0.155	185.44	0.000
		17	-0.268	0.008	191.19	0.000
		18	-0.231	-0.151	195.58	0.000
		19	-0.198	0.025	198.93	0.000
		20	-0.130	0.081	200.42	0.000
		21	-0.054	-0.071	200.68	0.000
		22	-0.023	-0.078	200.73	0.000
		23	0.040	0.133	200.89	0.000
		24	0.086	0.054	201.63	0.000

Testing for Stationarity: Correlogram

- Simple test for stationarity is to use autocorrelation function but this is sometimes ambiguous

- Stochastic trend can be identified by testing a_1 =1 in

$$Y_t = a_0 + a_1 Y_{t-1} + \varepsilon_t$$

This is known as testing for a *unit root*

Testing for Stationarity Using Dickey-Fuller Tests

- Natural to run regression and test $a_1 = 1$ versus $a_1 < 1$

$$Y_t = a_0 + a_1 Y_{t-1} + \varepsilon_t$$

- Dickey and Fuller (1979, 1981) showed that if $a_1 = 1$, its regression estimate is biased downwards

  Conventional estimates and t-stats will thus tend to incorrectly reject $a_1 = 1$

- Dickey and Fuller empirically derived critical values for testing ***a_1=1*** which should be used instead of usual *t*-distribution values

Dickey-Fuller Tests

- Instead of testing $\boldsymbol{a_1 = 1}$ against $\boldsymbol{a_1} < 1$ in

$$Y_t = a_0 + a_1 Y_{t-1} + \varepsilon_t$$

  a more convenient test is to run the regression

$$Y_t - Y_{t-1} = a_0 + \gamma Y_{t-1} + \varepsilon_t$$

  where $\gamma = (a_1 - 1)$ and test $\gamma = 0$ against $\gamma < 0$

- Note that $a_1 = 1$ and $\gamma = 0$ imply random walk type behaviour, i.e. nonstationarity

Dickey-Fuller Tests

- Basic tests are:

$$\Delta Y_t = \gamma Y_{t-1} + \varepsilon_t$$

$$\Delta Y_t = a_0 + \gamma Y_{t-1} + \varepsilon_t$$

$$\Delta Y_t = a_0 + a_2 t + \gamma Y_{t-1} + \varepsilon_t$$

- Augmented Dickey-Fuller tests control for higher-order autoregression in new dependent variable ΔY_t:

$$\Delta Y_t = a_0 + a_2 t + \gamma Y_{t-1} + \sum_{i=1}^{m} \beta_i \Delta Y_{t-i} + \varepsilon_t$$

Use m < (no. of observations)$^{1/3}$

- Procedure is to test $\gamma = 0$ after first getting a model with good t-stats for all other parameters. Once good t-stats have been established for all other parameters, focus on: H_0: $\gamma \geq 0$ H_1: $\gamma < 0$.

Dickey-Fuller Tables

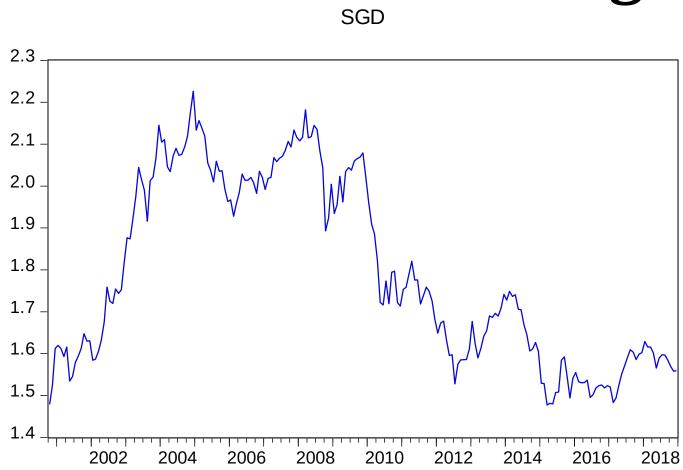
- Dickey-Fuller tables provide critical values for γ

Table A **Empirical Cumulative Distribution of τ**

Probability of a Smaller Value **Sample Size**	**0.01**	**0.025**	**0.05**	**0.10**	**0.90**	**0.95**	**0.975**	**0.99**
No Constant or Time ($a_0 = a_2 = 0$)				τ				
25	−2.66	−2.26	−1.95	−1.60	0.92	1.33	1.70	2.16
50	−2.62	−2.25	−1.95	−1.61	0.91	1.31	1.66	2.08
100	−2.60	−2.24	−1.95	−1.61	0.90	1.29	1.64	2.03
250	−2.58	−2.23	−1.95	−1.62	0.89	1.29	1.63	2.01
300	−2.58	−2.23	−1.95	−1.62	0.89	1.28	1.62	2.00
∞	−2.58	−2.23	−1.95	−1.62	0.89	1.28	1.62	2.00
Constant ($a_2 = 0$)				τ_μ				
25	−3.75	−3.33	−3.00	−2.62	−0.37	0.00	0.34	0.72
50	−3.58	−3.22	−2.93	−2.60	−0.40	−0.03	0.29	0.66
100	−3.51	−3.17	−2.89	−2.58	−0.42	−0.05	0.26	0.63
250	−3.46	−3.14	−2.88	−2.57	−0.42	−0.06	0.24	0.62
500	−3.44	−3.13	−2.87	−2.57	−0.43	−0.07	−0.24	0.61
∞	−3.43	−3.12	−2.86	−2.57	−0.44	−0.07	0.23	0.60
Constant + time				τ_τ				
25	−4.38	−3.95	−3.60	−3.24	−1.14	−0.80	−0.50	−0.15
50	−4.15	−3.80	−3.50	−3.18	−1.19	−0.87	−0.58	−0.24
100	−4.04	−3.73	−3.45	−3.15	−1.22	−0.90	−0.62	−0.28
250	−3.99	−3.69	−3.43	−3.13	−1.23	−0.92	−0.64	−0.31
500	−3.98	−3.68	−3.42	−3.13	−1.24	−0.93	−0.65	−0.32
∞	−3.96	−3.66	−3.41	−3.12	−1.25	−0.94	−0.66	−0.33

Source: This table was constructed by David A. Dickey using Monte Carlo methods. Standard errors of the estimates vary, but most are less than 0.20. The table is reproduced from Wayne Fuller. *Introduction to Statistical Time Series.* (New York: John Wiley). 1976.

Sing Dollar/Euro Exchange Rate



Sample: 2000M10 2018M12
Included observations: 219

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.978	0.978	212.28	0.000
		2	0.956	0.003	416.25	0.000
		3	0.938	0.072	613.56	0.000
		4	0.919	-0.046	803.59	0.000
		5	0.901	0.027	987.01	0.000
		6	0.884	0.022	1164.5	0.000
		7	0.864	-0.077	1335.0	0.000
		8	0.842	-0.059	1497.6	0.000
		9	0.819	-0.038	1652.3	0.000
		10	0.797	-0.015	1799.3	0.000
		11	0.776	0.020	1939.3	0.000
		12	0.755	0.001	2072.6	0.000
		13	0.736	0.027	2200.0	0.000
		14	0.720	0.053	2322.4	0.000
		15	0.704	0.028	2440.1	0.000
		16	0.687	-0.032	2552.8	0.000
		17	0.670	-0.025	2660.3	0.000
		18	0.651	-0.038	2762.4	0.000
		19	0.631	-0.054	2858.9	0.000
		20	0.609	-0.094	2949.0	0.000
		21	0.587	-0.015	3033.1	0.000
		22	0.563	-0.067	3111.0	0.000
		23	0.538	-0.039	3182.4	0.000
		24	0.515	0.025	3248.3	0.000
		25	0.493	0.016	3308.9	0.000
		26	0.468	-0.052	3363.7	0.000
		27	0.447	0.101	3414.2	0.000
		28	0.426	-0.025	3460.2	0.000
		29	0.404	-0.010	3501.9	0.000
		30	0.386	0.042	3540.0	0.000
		31	0.368	-0.006	3574.8	0.000
		32	0.354	0.088	3607.3	0.000
		33	0.341	-0.017	3637.6	0.000
		34	0.327	-0.021	3665.6	0.000
		35	0.311	-0.073	3691.0	0.000
		36	0.298	0.064	3714.5	0.000

Sing Dollar/Euro Exchange Rate Unit Root Test

Null Hypothesis: SGD has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=14)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-1.458925	0.5527
Test critical values:	1% level	-3.460313	
	5% level	-2.874617	
	10% level	-2.573817	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(SGD)
Method: Least Squares

Sample (adjusted): 2000M11 2018M12
Included observations: 218 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
SGD(-1)	-0.017101	0.011722	-1.458925	0.1460
C	0.031024	0.021167	1.465632	0.1442

R-squared	0.009758	Mean dependent var	0.000369
Adjusted R-squared	0.005173	S.D. dependent var	0.037910
S.E. of regression	0.037812	Akaike info criterion	-3.703241
Sum squared resid	0.308828	Schwarz criterion	-3.672191
Log likelihood	405.6533	Hannan-Quinn criter.	-3.690700
F-statistic	2.128463	Durbin-Watson stat	1.931425
Prob(F-statistic)	0.146038		

Sing Dollar/Euro Exchange Rate Unit Root Test

Null Hypothesis: D(SGD) has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=14)

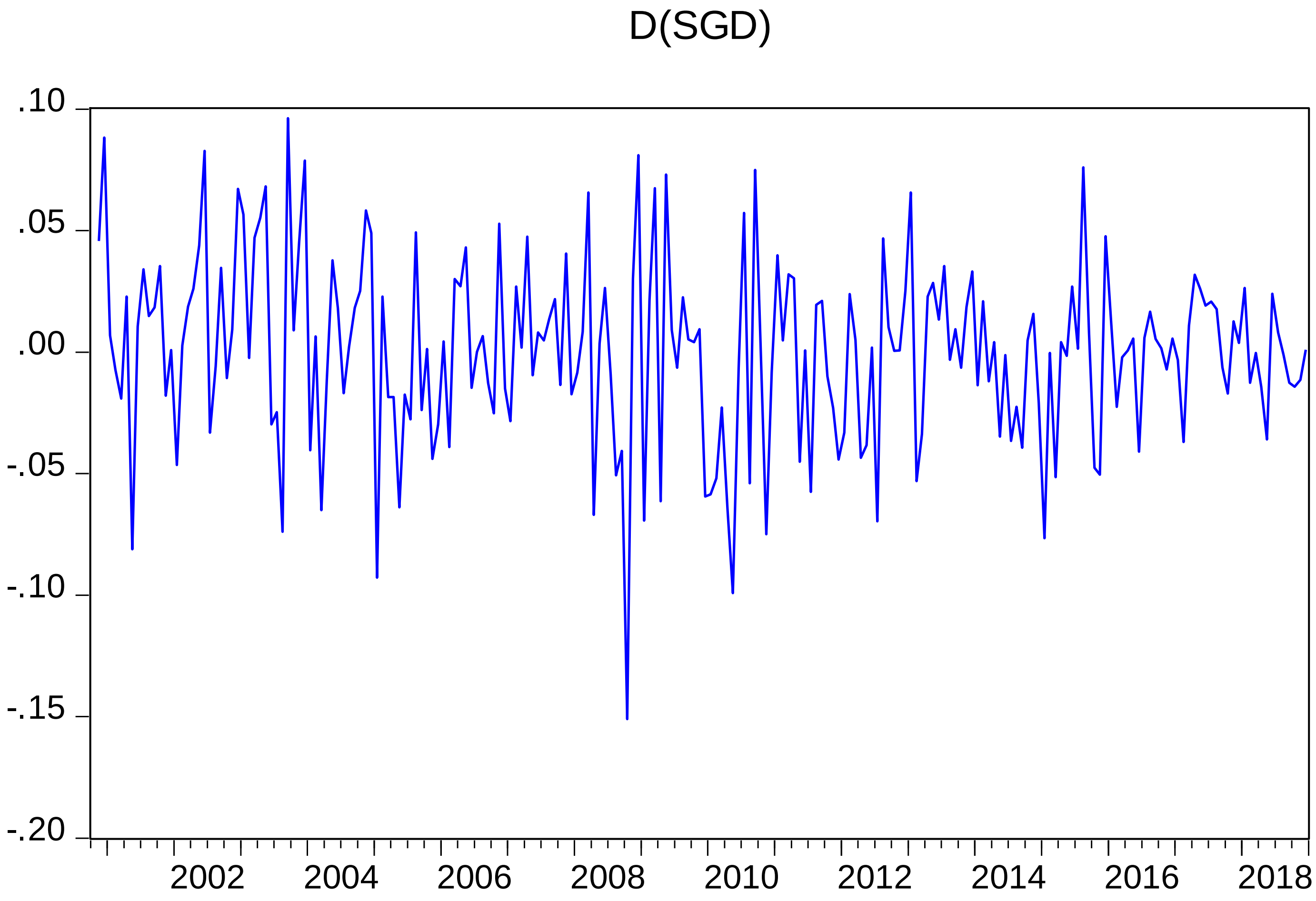
		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-14.36463	0.0000
Test critical values:	1% level	-3.460453	
	5% level	-2.874679	
	10% level	-2.573850	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(SGD,2)
Method: Least Squares
Date: 01/21/19 Time: 12:35
Sample (adjusted): 2000M12 2018M12
Included observations: 217 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
D(SGD(-1))	-0.976132	0.067954	-14.36463	0.0000
C	0.000151	0.002576	0.058680	0.9533

R-squared	0.489726	Mean dependent var	-0.000206
Adjusted R-squared	0.487353	S.D. dependent var	0.053002
S.E. of regression	0.037949	Akaike info criterion	-3.695968
Sum squared resid	0.309629	Schwarz criterion	-3.664817
Log likelihood	403.0126	Hannan-Quinn criter.	-3.683385
F-statistic	206.3425	Durbin-Watson stat	1.999945
Prob(F-statistic)	0.000000		


D(SGD)
.10
.05
.00
-.05
-.10
-.15
-.20
2002
2004
2006
2008
2010
2012
2014
2016
2018

Correlogram D(Sgd/Euro)

Sample: 2000M10 2018M12
Included observations: 218

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.024	0.024	0.1259	0.723
		2	-0.025	-0.025	0.2599	0.878
		3	0.074	0.075	1.4842	0.686
		4	-0.043	-0.048	1.9064	0.753
		5	-0.073	-0.067	3.0959	0.685
		6	0.128	0.125	6.7784	0.342
		7	-0.027	-0.033	6.9463	0.434
		8	0.055	0.072	7.6295	0.470
		9	0.046	0.017	8.1202	0.522
		10	-0.036	-0.027	8.4272	0.587
		11	0.001	0.013	8.4276	0.675
		12	-0.029	-0.053	8.6278	0.734
		13	-0.104	-0.079	11.147	0.598
		14	-0.044	-0.056	11.594	0.639
		15	0.019	0.014	11.676	0.703
		16	0.027	0.042	11.847	0.754
		17	0.034	0.021	12.126	0.792
		18	0.065	0.064	13.154	0.782
		19	0.121	0.138	16.712	0.609
		20	0.033	0.045	16.981	0.654
		21	0.002	0.015	16.982	0.712
		22	0.030	0.025	17.207	0.752
		23	-0.002	-0.002	17.208	0.799
		24	-0.035	-0.042	17.506	0.826
		25	0.099	0.062	19.956	0.749
		26	-0.082	-0.120	21.620	0.709
		27	0.077	0.075	23.111	0.679
		28	0.026	-0.012	23.278	0.719
		29	-0.079	-0.052	24.879	0.685
		30	0.024	0.053	25.032	0.723
		31	-0.033	-0.057	25.312	0.754
		32	-0.060	0.038	26.225	0.754
		33	0.002	-0.024	26.226	0.793
		34	0.012	0.014	26.264	0.826
		35	-0.008	0.014	26.280	0.856
		36	0.031	-0.016	26.539	0.875

Summary

- We have analysed the use of multiple regression for time series analysis.
- *DW* statistic evaluates autocorrelation for residuals
- Correlogram displays autocorrelation function for different lags. This is very important for analysing times series.
- Using *differenced* variables (changes) eliminates the trends and thus avoids spurious regression.
- For some series, the trend is *stochastic*. Classic example is the random walk $Y_t = Y_{t-1} + \varepsilon_t$
- We have analysed the Dickey-Fuller Tests.

EXTRA Readings

Nerlove, M. and F. Diebold, 'Unit Roots in Economic Time Series: A Selective Survey', in T. Bewley (ed.), *Advances in Econometrics*, vol. 8, New York: JAI Press, 1990.