University of Nottingham
UK | CHINA | MALAYSIA

Introductory Econometrics BUSI2053

Heteroscedasticity and Autocorrelation

Lecture Outline

- The Nature of Heteroscedasticity and how to detect it
- Sources and consequences of Heteroscedasticity
- Remedial Measures
- The Nature and source of Autocorrelationand how to detect it
- Consequences and remedial measures of Autocorrelation

Suggested Reading:

- Chapter 8 & 9: Hill, R.C., Griffiths W.E. and Lim, G.C. Principles of Econometrics, fourth edition, Wiley, 2012 (pp. 298-319) (pp. 335-364)
- Chapter 11 & 12 : Gujarati, D.N. and Porter D.C. Basic econometrics, 5th ed., McGraw-Hill, 2009 (pp. 365-395) (pp. 412-446)
- Chapter 7 & 12: Dougherty, Christopher. Introduction to econometrics. 4th ed. Oxford University Press, 2011 (pp.280-297) (pp.429-456)
- Chapter 16-17: Westhoff, F. An introduction to econometrics: a self-contained approach, MIT Press, 2013

Assumptions of the Simple Linear Regression

SR1:The value of y, for each value of x, is is given by the linear regression: $y_i = \beta_1 + \beta_2 x_i + e_i$

SR2: The expected value of the random error e is: $E(e_i) = 0$

SR3: The variance of the random error e is a constant: $\text{var}(e_i) = \sigma^2 = \text{var}(y_i)$

Note that σ^2 has no subscript i, which means a constant.

SR4: The covariance between any pair of random errors, e_i and e_j is zero (independent): $\text{cov}(e_i, e_j) = \text{cov}(y_i, y_j) = 0$

SR5: The variable x is not random, and must take at least two different values

SR6: Optional $e \sim N(0, \sigma^2)$

Nature of Heteroscedasticity

- In a regression model:

$$y_i = \beta_1 + \beta_2 x_i + e_i$$

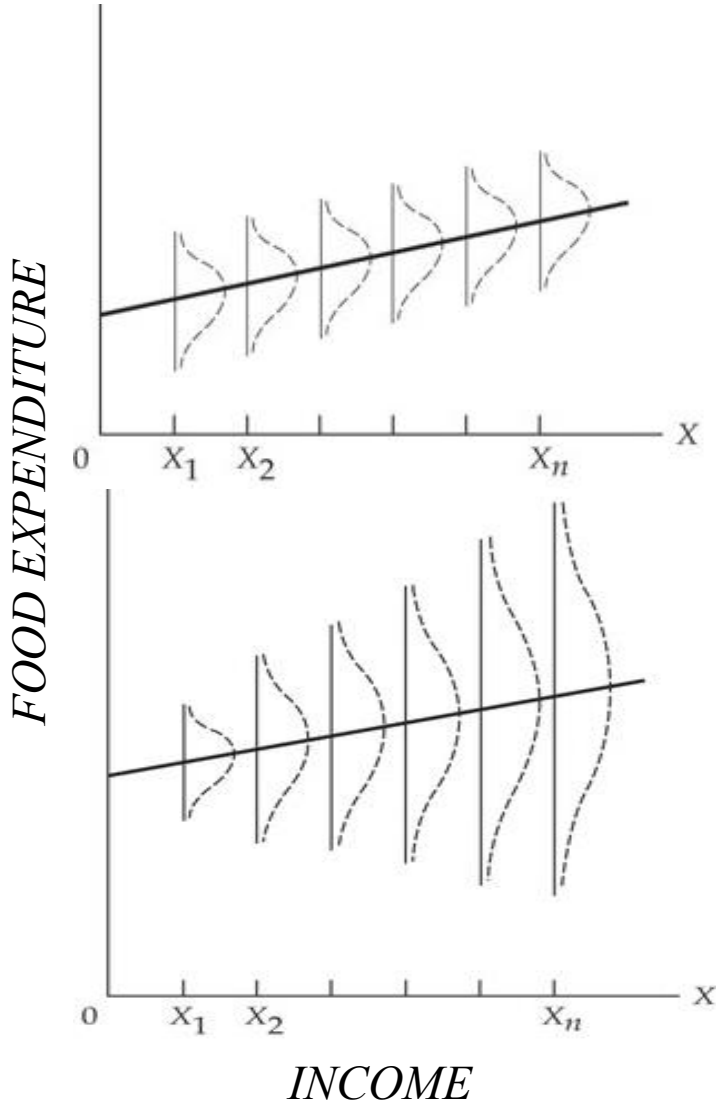
- The probability of getting large positive or negative values for e is higher for larger values of x
  - var(e_i) depend directly on x or var(e_i) increases/decreases as x increases
- When the variances of error terms for all observations are not the same ($\text{var}(e_i) = \sigma_i^2$), we have **heteroscedasticity**
- Assumption 3 of homoscedasticity:

$$\text{var}(y_i) = \text{var}(e_i) = \sigma^2$$

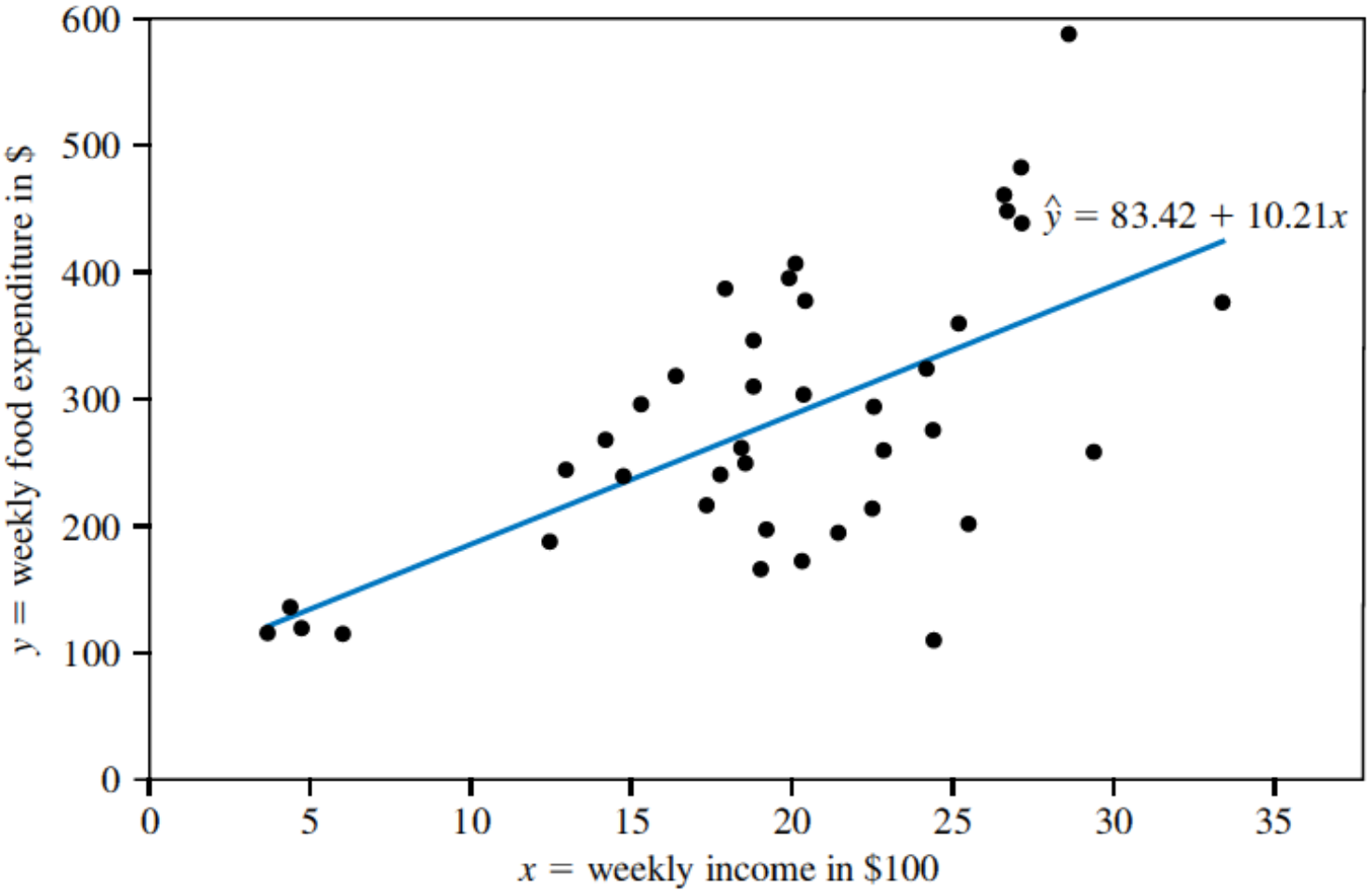
does not hold!

subscript i

without subscript i



Nature of Heteroscedasticity



Least squares estimated food expenditure function and observed data points

Sources of Heteroscedasticity

- Heteroscedasticity is often encountered when using cross-sectional data
- As the size of the economic unit becomes larger, there is more uncertainty associated with the outcomes y

1. Outliers
2. Specification error. For example, omitting an important explanatory variable can look like heteroscedasticity.
3. Error-learning models (as people learn, their errors of behaviour become smaller overtime. For example, number of typing errors)
4. Models including income: as incomes grow, people have more discretionary income. It is similar for profits of companies.

Consequences of Heteroscedasticity

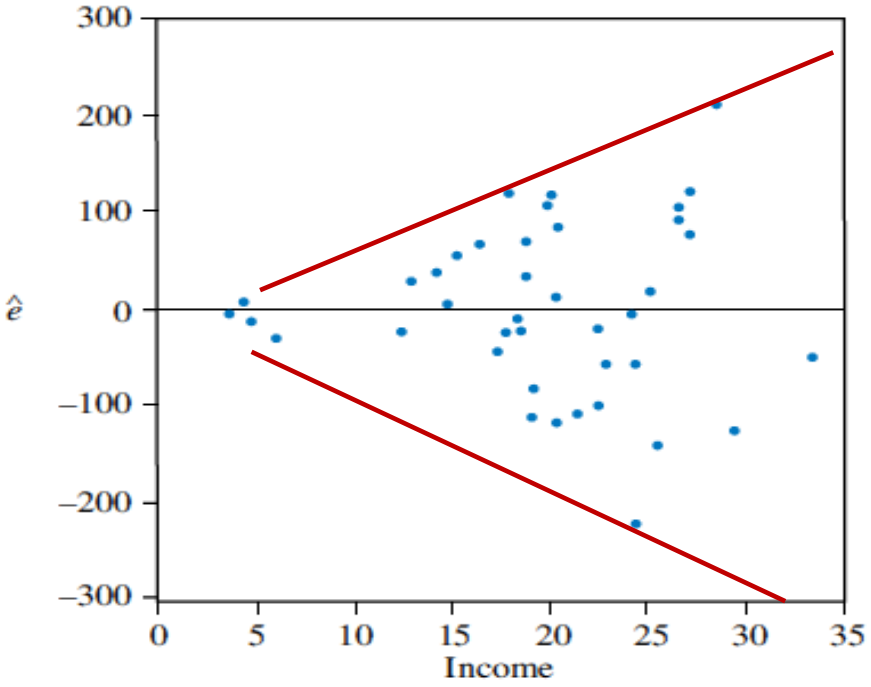
- There are two implications of heteroscedasticity:
  1. The least squares estimator is **still a linear and unbiased estimator**, but it is no longer the best (no longer have minimum variance)
     - There is another unbiased estimator with a smaller variance
  2. The standard errors usually computed for the least squares estimator are incorrect
     - Confidence intervals and hypothesis tests that use these standard errors may be misleading

Detecting Heteroscedasticity

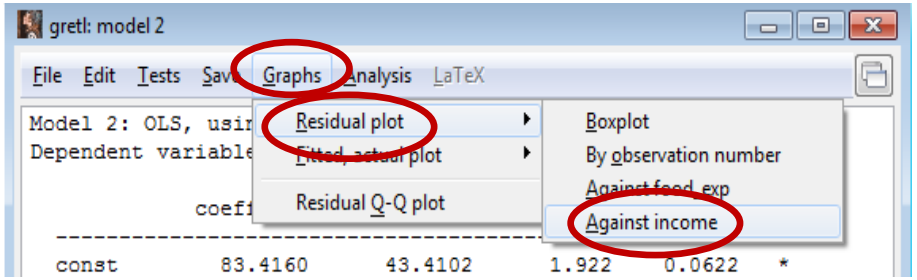
- There are two methods we can use to detect heteroscedasticity
  1. An informal way using residual plots
     - If the errors are homoscedastic, there should be no patterns of any sort in the residuals plot
     - If the errors are heteroscedastic, they may tend to exhibit greater variation in some systematic way

  2. A formal way using statistical tests
     - White's General Heteroscedasticity Test

Detecting Heteroscedasticity (informal way)



Least squares food expenditure residuals plotted against income



In a regression with more than one explanatory variable we can plot the least squares residuals against each explanatory variable, or against, $\hat{y}_i$, to see if they vary in a systematic way

Detecting Heteroscedasticity (formal way)

White's General Heteroscedasticity Test

For demonstration, we will use only 2 independent (numerical) variables

Step 1: Estimate the model with OLS:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$$

Step 2: Save the residuals and estimate the following auxiliary model:

$$\hat{e}_i^2 = \alpha_1 + \alpha_2 x_{1i} + \alpha_3 x_{2i} + \alpha_4 x_{1i}^2 + \alpha_5 x_{2i}^2 + \alpha_6 x_{1i} x_{2i} + v_i$$

Detecting Heteroscedasticity (formal way)

White's General Heteroscedasticity Test (continued)

$$\hat{e}_i^2 = \alpha_1 + \alpha_2 x_{1i} + \alpha_3 x_{2i} + \alpha_4 x_{1i}^2 + \alpha_5 x_{2i}^2 + \alpha_6 x_{1i} x_{2i} + v_i$$

H_0: No heteroscedasticity ($\alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0$)
H_1: There is heteroscedasticity.

Step 3: Obtain R^2 value from the auxiliary regression. Under the **null hypothesis of no heteroscedasticity (which means variance of error is homoscedastic)**,

$$nR^2 \sim \chi^2_{k-1}; \text{ where } n = \text{No of observation}$$

Step 4: If $nR^2 > \chi^2_{k-1}$, reject the null of no heteroscedasticity

k = number of parameters (including the constant) in the auxiliary model

Detecting Heteroscedasticity (formal way)

Testing the Heteroscedasticity for Pizza Example

Regress "pizza" on "income" and "age" and save the squared of residuals.
Using estimated residuals, we run

$$\hat{e}_i^2 = \alpha_1 + \alpha_2 income_i + \alpha_3 age_i + \alpha_4 income_i^2 + \alpha_5\, age_i^2 + \alpha_6\, (income_i \times age_i) + v_i$$

R^2=0.241282; n=40

Test H_0: No heteroscedasticity; H_1: There is heteroscedasticity

$$\chi^2 = nR^2 = 40 \times 0.1888 = 9.651281$$

The 5% critical value is $\chi^2(0.05, 5) = 11.07$

Since $nR^2 = 9.65 < \chi^2_{k-1} = 11.07$; Do Not Reject the null of no heteroscedasticity.
We can conclude that there is no violation of homoscedasticity assumptions in estimating the model.

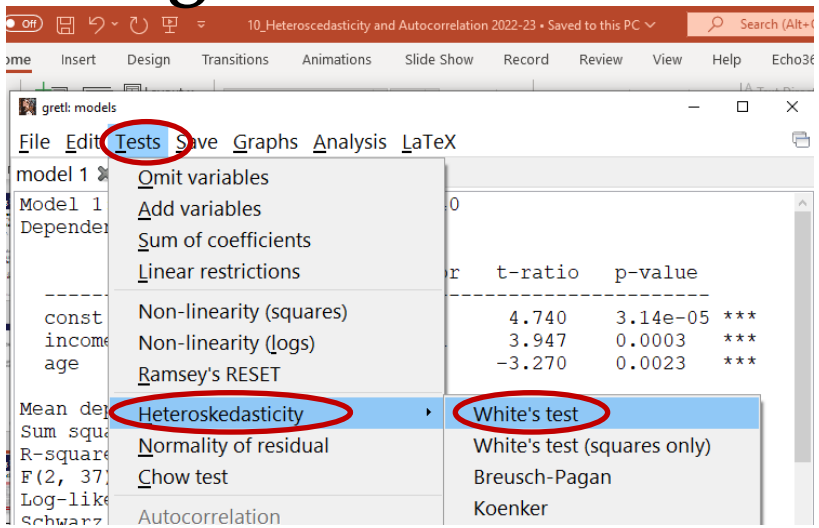
TABLE A.4

χ^2 (Chi-Squared) Distribution: Critical Values of χ^2

	Significance level		
Degrees of freedom	5%	1%	0.1%
1	3.841	6.635	10.828
2	5.991	9.210	13.816
3	7.815	11.345	16.266
4	9.488	13.277	18.467
5	11.070	15.086	20.515
6	12.592	16.812	22.458
7	14.067	18.475	24.322
8	15.507	20.090	26.124
9	16.919	21.666	27.877
10	18.307	23.209	29.588

Detecting Heteroscedasticity (formal way)

White's Test in gretl

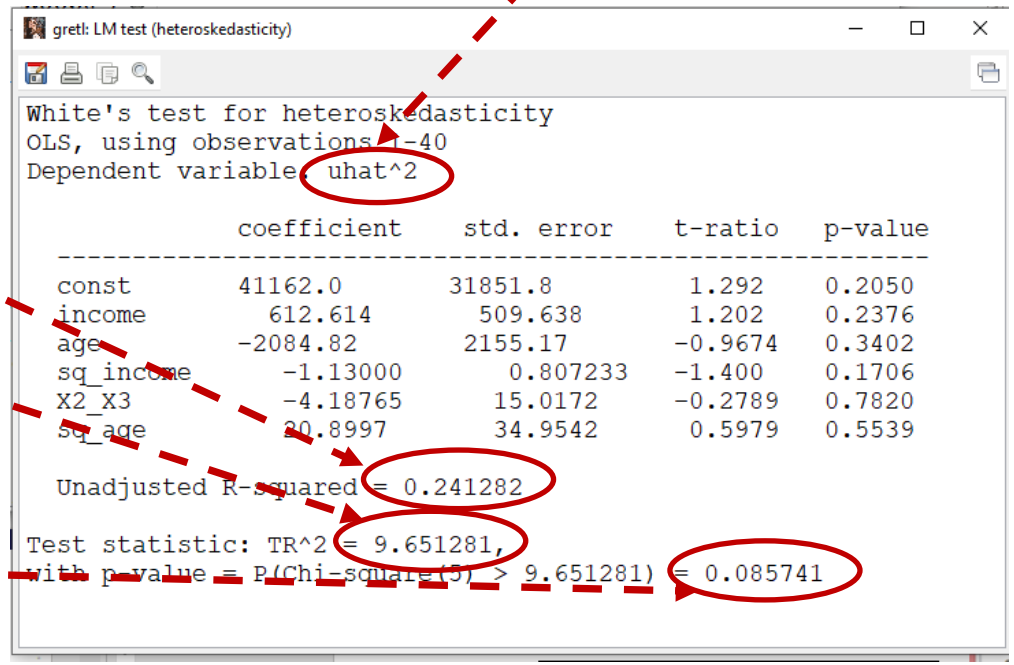


Squared of the estimated residuals

R-squared of the auxiliary model

Chi-square test-statistic=nR^2

P-value of the test statistic



Some remedial measures for Heteroscedasticity

- **Transforming the Dependent Variable**: For example, log transformation can reduce the impact of heteroscedasticity by equalising variance across observations.

- **Robust Standard Errors**: Use heteroscedasticity-consistent standard errors (e.g., White's standard errors). These do not change the estimated coefficients but adjust the standard errors to account for heteroscedasticity, making hypothesis testing reliable.

- **Re-specify the Model**: Check if the model is mis-specified (e.g., omitted variables, incorrect functional form). Add missing variables, square-terms or interaction terms that might explain the heteroscedasticity.

- **Weighted Least Squares (WLS) Method**: Assigning appropriate weights to each observation, ensures that observations with larger variances are given less influence during estimation (See appendix1 that assign $\frac{1}{\sigma}$ as weight).

The Nature of Autocorrelation

Definition

- Autocorrelation means that there is **correlation between members of a series of observations** ordered according to time or space.
- For a time-series data Assumption SR4 can be written as:

$$\text{cov}(y_t, y_s) = \text{cov}(e_t, e_s) = 0 \quad \text{for } t \neq s$$

- Autocorrelation means:

$$\text{cov}(y_t, y_s) = \text{cov}(e_t, e_s) \neq 0 \text{ for } t \neq s$$

YEAR	PRICE	CPI	NYSE
1977	147.98	60.6	53.69
1978	193.44	65.2	53.7
1979	307.62	72.6	58.32
1980	612.51	82.4	68.1
1981	459.61	90.9	74.02
1982	376.01	96.5	68.93
1983	423.83	99.6	92.63
1984	360.29	103.9	92.46

Notes:

1. Time-series observations on a given economic unit, observed over a number of time periods, are likely to be correlated
2. Although autocorrelation is mostly associated with time series data, spatial autocorrelation (i.e., autocorrelation in cross section data) is also a possibility.

Sources of Autocorrelation

1. Inertia: Many economic time-series exhibit business cycles. So, in many time series regression, successive observations are likely to be correlated.
2. Model specification error(s): Omission of an important explanatory variable can result in errors that appear autocorrelated.
3. The cobweb phenomenon as seen for example in agricultural markets can lead to autocorrelation since time t+1's decisions are based on what happened at time t.
4. If one of the explanatory variables should be lagged but is not, then the errors will reflect a systematic pattern.
5. Manipulation, interpolation or extrapolation of data can result in autocorrelation.

Consequences of Autocorrelation

1. The least square estimators are **still linear and unbiased**.

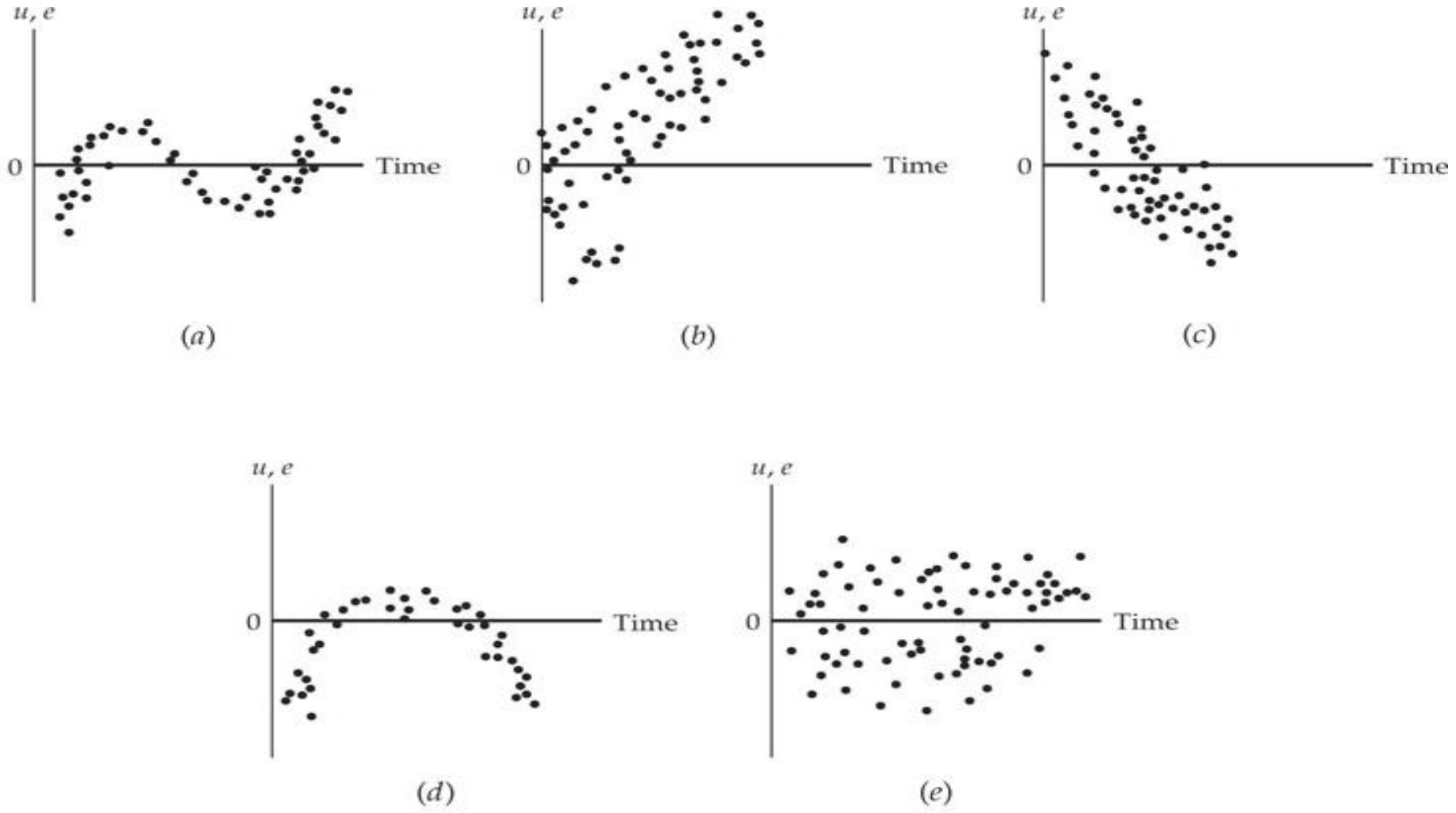
2. But they are not efficient (do not have minimum variance)

3. The estimated variances of coefficients are biased (sometimes seriously underestimate true variances and standard errors).

4. The usual t and F test are not generally reliable.

5. The conventionally computed R^2 may be unreliable

Detecting Autocorrelation (Residual plot)



No Autocorrelation: There should be no systematic (random) pattern in residual plots.

Patterns of Autocorrelation (Essential of Econometrics, 4th ed.; page 314)

Detecting Autocorrelation (Residual plot)

Visual examination of OLS residuals, *e*'s, can show the likely presence of autocorrelation.

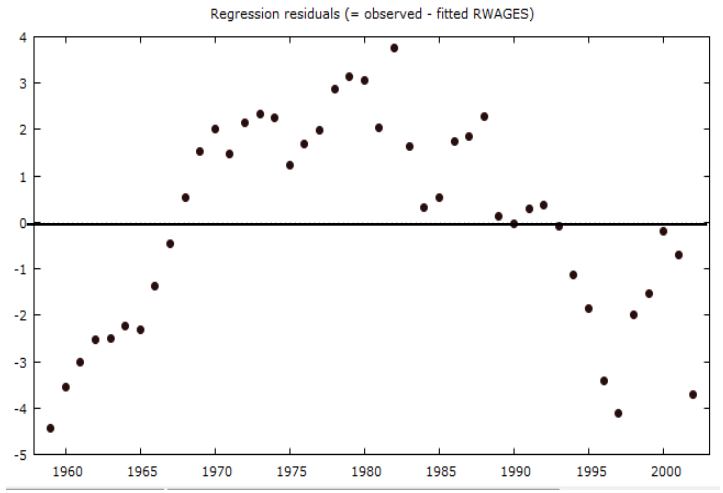
REAL WAGES AND PRODUCTIVITY IN THE U.S. BUSINESS SECTOR, 1959-2002

Data set on Moodle.

OLS, using observations 1959-2002 (T = 44)
Dependent variable: Compensation

	Coeff:	Std. Error	t-ratio	p-value
Const	29.575	1.4605	20.25	<0.0001
Productivity	0.7006	0.0171	40.92	<0.0001

R-squared	0.9755	Adj R-squared	0.9749
F(1, 42)	1674.298	P-value(F)	0.0000
rho	0.8915	D-W	0.2137



Residual plot of the regression

Although the results look good, an examination of time-sequence residual plot indicates that the residuals do not seem to be randomly distributed.

Detecting Autocorrelation (Durbin-Watson test)

Durbin-Watson (DW) statistic: $$\mathrm{DW}(d) = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_i^2}$$

Assumptions:

1. The regression model includes an intercept term.
2. A lagged dependent variable is not one of the explanatory variables
3. There are no missing observations.
4. The disturbances e_t are generated by the following first-order autoregressive, AR(1) mechanism.

$$e_t = \rho e_{t-1} + v_t \text{ where } -1 \leq \rho \leq 1$$

Detecting Autocorrelation (Durbin-Watson test)

Durbin-Watson Test assumption: $e_t = \rho e_{t-1} + v_t$

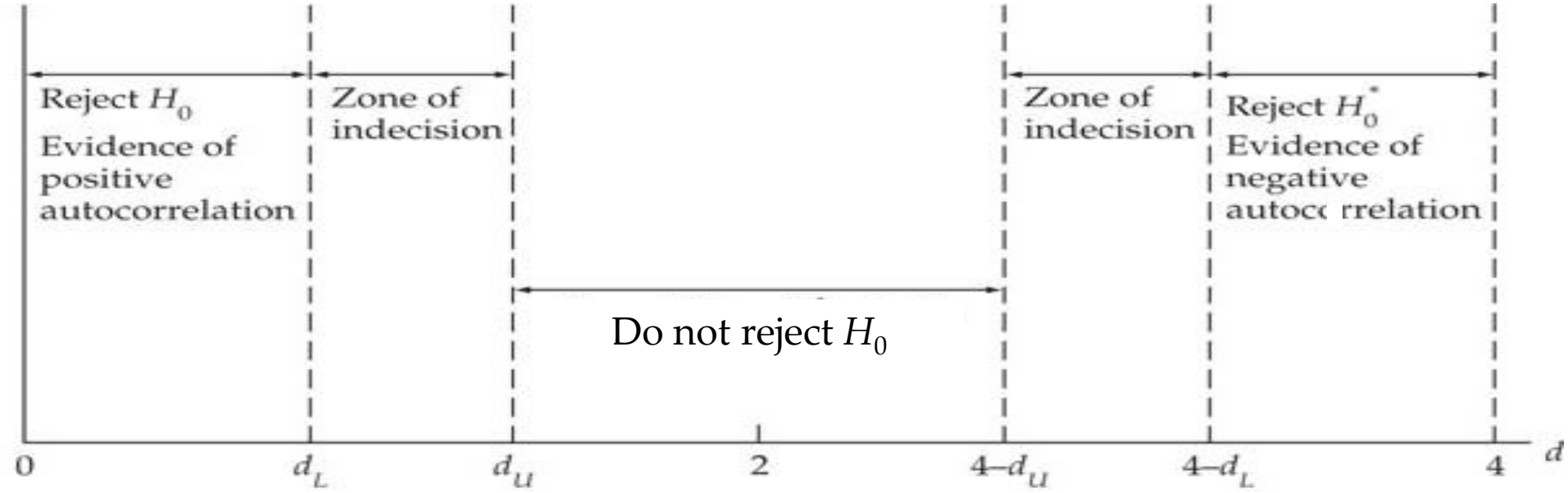
- H_0: No first order autocorrelation ($\rho = 0$)
- H_1: Positive autocorrelation ($\rho > 0$) or negative autocorrelation ($\rho < 0$)

Steps:

1. Run OLS regression and obtain the residuals e_t.
2. Compute DW(d) from the equation (Most econometric softwares now do this routinely).
3. Find out the critical d_L and d_U from the Durban-Watson tables for the given sample size (n) and the given number of explanatory variables.
4. Follow the decision rules given in the table (see slide **23**)

Detecting Autocorrelation (Durbin-Watson test)

Durbin-Watson Test decision making



In our previous example, we have *D-W*=0.2137. From the Durban-Watson tables, d_L=1.475 and d_U=1.566 for n=45 (nearest to 44) and one explanatory variable at 5% significant level. So, there is positive autocorrelation.

k' = no. of independent variables (excluding the constant/intercept)

TABLE D.5A

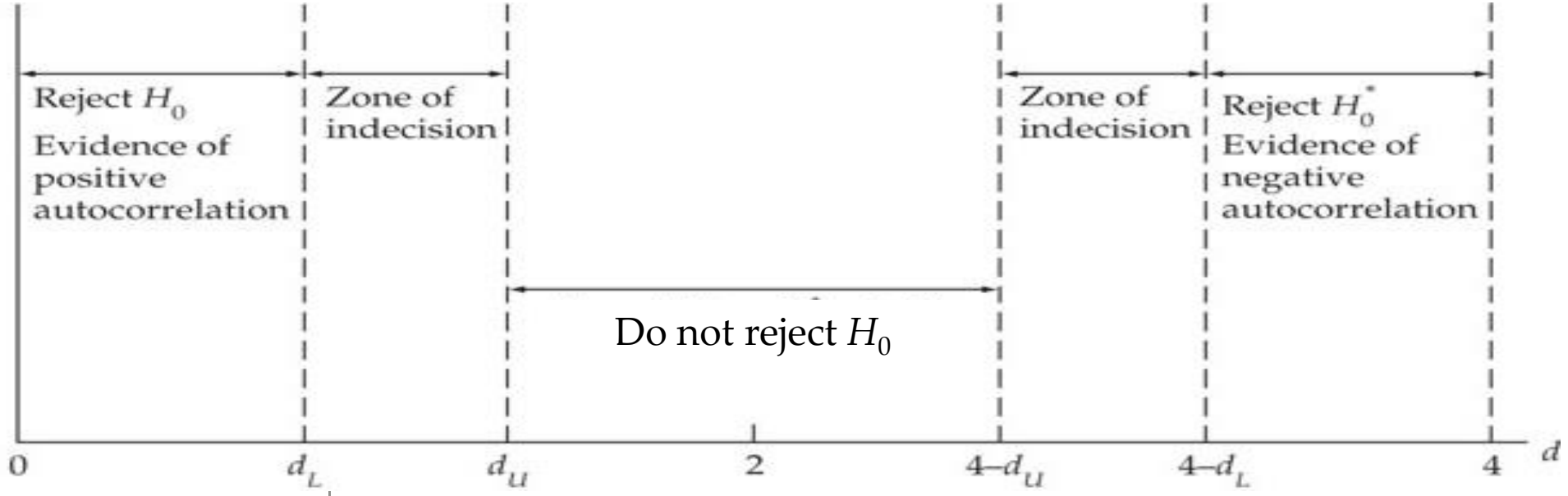
DURBIN–WATSON d STATISTIC: SIGNIFICANCE POINTS OF d_L AND d_U AT 0.05 LEVEL OF SIGNIFICANCE

	$k'=1$		$k'=2$		$k'=3$		$k'=4$		$k'=5$		$k'=6$		$k'=7$		$k'=8$		$k'=9$		$k'=10$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
6	0.610	1.400	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
7	0.700	1.356	0.467	1.896	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
8	0.763	1.332	0.559	1.777	0.368	2.287	—	—	—	—	—	—	—	—	—	—	—	—	—	—
9	0.824	1.320	0.629	1.699	0.455	2.128	0.296	2.588	—	—	—	—	—	—	—	—	—	—	—	—
10	0.879	1.320	0.697	1.641	0.525	2.016	0.376	2.414	0.243	2.822	—	—	—	—	—	—	—	—	—	—
11	0.927	1.324	0.658	1.604	0.595	1.928	0.444	2.283	0.316	2.645	0.203	3.005	—	—	—	—	—	—	—	—
12	0.971	1.331	0.812	1.579	0.658	1.864	0.512	2.177	0.379	2.506	0.268	2.832	0.171	3.149	—	—	—	—	—	—
13	1.010	1.340	0.861	1.562	0.715	1.816	0.574	2.094	0.445	2.390	0.328	2.692	0.230	2.985	0.147	3.266	—	—	—	—
39	1.435	1.540	1.382	1.597	1.328	1.658	1.273	1.722	1.218	1.789	1.161	1.859	1.104	1.932	1.047	2.007	0.990	2.085	0.932	2.164
40	1.442	1.544	1.391	1.600	1.338	1.659	1.285	1.721	1.230	1.786	1.175	1.854	1.120	1.924	1.064	1.997	1.008	2.072	0.952	2.149
45	1.475	1.566	1.430	1.615	1.383	1.666	1.336	1.720	1.287	1.776	1.238	1.835	1.189	1.895	1.139	1.958	1.089	2.022	1.038	2.088
50	1.503	1.585	1.462	1.628	1.421	1.674	1.378	1.721	1.335	1.771	1.291	1.822	1.246	1.875	1.201	1.930	1.156	1.986	1.110	2.044
55	1.528	1.601	1.490	1.641	1.452	1.681	1.414	1.724	1.374	1.768	1.334	1.814	1.294	1.861	1.253	1.909	1.212	1.959	1.170	2.010
60	1.549	1.616	1.514	1.652	1.480	1.689	1.444	1.727	1.408	1.767	1.372	1.808	1.335	1.850	1.298	1.894	1.260	1.939	1.222	1.984
65	1.567	1.629	1.536	1.662	1.503	1.696	1.471	1.731	1.438	1.767	1.404	1.805	1.370	1.843	1.336	1.882	1.301	1.923	1.266	1.964
70	1.583	1.641	1.554	1.672	1.525	1.703	1.494	1.735	1.464	1.768	1.433	1.802	1.401	1.837	1.369	1.873	1.337	1.910	1.305	1.948
75	1.598	1.652	1.571	1.680	1.543	1.709	1.515	1.739	1.487	1.770	1.458	1.801	1.428	1.834	1.399	1.867	1.369	1.901	1.339	1.935
80	1.611	1.662	1.586	1.688	1.560	1.715	1.534	1.743	1.507	1.772	1.480	1.801	1.453	1.831	1.425	1.861	1.397	1.893	1.369	1.925
85	1.624	1.671	1.600	1.696	1.575	1.721	1.550	1.747	1.525	1.774	1.500	1.801	1.474	1.829	1.448	1.857	1.422	1.886	1.396	1.916
90	1.635	1.679	1.612	1.703	1.589	1.726	1.566	1.751	1.542	1.776	1.518	1.801	1.494	1.827	1.469	1.854	1.445	1.881	1.420	1.909
95	1.645	1.687	1.623	1.709	1.602	1.732	1.579	1.755	1.557	1.778	1.535	1.802	1.512	1.827	1.489	1.852	1.465	1.877	1.442	1.903
100	1.654	1.694	1.634	1.715	1.613	1.736	1.592	1.758	1.571	1.780	1.550	1.803	1.528	1.826	1.506	1.850	1.484	1.874	1.462	1.898
150	1.720	1.746	1.706	1.760	1.693	1.774	1.679	1.788	1.665	1.802	1.651	1.817	1.637	1.832	1.622	1.847	1.608	1.862	1.594	1.877
200	1.758	1.778	1.748	1.789	1.738	1.799	1.728	1.810	1.718	1.820	1.707	1.831	1.697	1.841	1.686	1.852	1.675	1.863	1.665	1.874

Detecting Autocorrelation (Homework)

Durbin-Watson Test

n	*K'*	*DW*	*dU*	*dL*	*Conclusion*
40	2	1.23	1.600	1.391	1.23<dL=1.391 [+autocorrelation]
50	5	3.66	1.771	1.335	2.665(4 – dL)<3.66<4 [-autocorrelation]
100	4	2.04	1.758	1.502	dU<2.04<2.242 (4 – dU)[No autocorrelation]
200	6	1.78	1.831	1.707	dL=1.707<1.78<dU=1.831 [Inconclusive]



Some remedial measures for Autocorrelation

Include Lagged Variables: Add a lagged dependent variable (e.g., Y_{t-1}) or other lagged predictors as independent variables. This accounts for relationships over time and helps reduce the autocorrelation in the residuals.

Use Robust Standard Errors (e.g., Newey-West): Even if autocorrelation exists, you can adjust the standard errors to make them robust. This doesn't change the regression coefficients but corrects the errors, ensuring reliable hypothesis testing.

Re-specify the Model: Check if the model is mis-specified (e.g., omitted variables, incorrect functional form).

Generalised Least Squares (GLS): If autocorrelation follows a specific pattern (e.g., AR(1)), GLS can handle it by transforming the data (See appendix 2).

Summary

After this lecture, you should be able to

- Know the definitions and consequences of Heteroscedasticity and Autocorrelation
- Detect Heteroscedasticity and Autocorrelation by using informal/formal methods
- Identify appropriate remedy measures to reduce severity of Heteroscedasticity and Autocorrelation, if not to eliminate completely

University of Nottingham
UK | CHINA | MALAYSIA

Appendix1: WLS for heteroscedasticity

Assume that the error variance is not constant and proportional to σ_i^2:

$$\text{var}(e_i) = \sigma_i^2$$

- By multiplying the original model $y_i = \beta_1 + \beta_2 x_i + e_i$ by the weight $\frac{1}{\sigma_i}$ yields:

$$\left(\frac{y_i}{\sigma_i}\right) = \beta_1\left(\frac{1}{\sigma_i}\right) + \beta_2\left(\frac{x_i}{\sigma_i}\right) + \left(\frac{e_i}{\sigma_i}\right)$$

- **Can be shown that the error term is homoscedastic (using variance rules):**

$$\text{var}\left(\frac{e_i}{\sigma_i}\right) = \frac{1}{\sigma_i^2}\text{var}(e_i) = \frac{1}{\sigma_i^2} \times \sigma_i^2 = 1 \ (\because \text{var}(e_i) = \sigma_i^2)$$

Appendix 2: GLS for Autocorrelation

Let's look at a simple two-variable model:

$$Y_t = \beta_1 + \beta_2 x_t + e_t \qquad (1)$$

and assume that the error terms follow the AR(1) scheme:

$$e_t = \rho e_{t-1} + v_t \qquad (2)$$

Write the regression (1) with a one-period lag as:

$$Y_{t-1} = \beta_1 + \beta_2 x_{t-1} + e_{t-1} \quad (3)$$

Multiplying regression (3) by ρ on both sides:

$$\rho Y_{t-1} = \rho\beta_1 + \rho\beta_2 x_{t-1} + \rho e_{t-1} \; (4)$$

Subtract (4) from (1)

$$(Y_t - \rho Y_{t-1}) = \beta_1(1-\rho) + \beta_2(x_t - \rho x_{t-1}) + (e_t - \rho e_{t-1})$$

$$(Y_t - \rho Y_{t-1}) = \beta_1(1-\rho) + \beta_2(x_t - \rho x_{t-1}) + v_t \quad (5)$$

Note: $e_t = \rho e_{t-1} + v_t$

$v_t = e_t - \rho e_{t-1}$

Appendix 2: GLS for Autocorrelation

$$(Y_t - \rho Y_{t-1}) = \beta_1(1-\rho) + \beta_2(x_t - \rho x_{t-1}) + \upsilon_t \quad (5)$$

We can write equation (5) as:

$$Y_t^* = \beta_1^* + \beta_2 x_t^* + \upsilon_t \qquad (6)$$

where $Y_t^* = (Y_t - \rho Y_{t-1})$ $\quad x_t^* = (x_t - \rho x_{t-1})$

$$\beta_1^* = \beta_1(1-\rho)$$

Now the error term υ_t satisfies the standard OLS assumptions. Apply OLS to the transformed variable Y^* and X^*. The estimators will have the desirable BLUE property.

This transformation is known as Durbin's two-step procedure and the transformed model can be estimated using Generalised Least Square Method.

Appendix 2: GLS for Autocorrelation

How to estimate ρ

1. Assume $\rho = 1$ (First difference method)
2. Estimate ρ from DW statistic (d): $\hat{\rho} \approx 1 - \frac{d}{2}$

In the regression of wages and productivity (Slide no. 22), d=0.2137

$$\hat{\rho} \approx 1 - \frac{0.2137}{2} = 0.8932$$

3. Estimate ρ from OLS residuals, e_t using equation (2).

$$\hat{e}_t = \hat{\rho} e_{t-1}$$

In the regression of wages and productivity (Slide no. 22)

$$\hat{e}_t = 0.8708 e_{t-1} \quad (\text{se} = 0.0719; R^2 = 0.7814)$$