# Ch 6: Violations of regression assumptions – autocorrelation

# Outline of the lecture

1. The nature of autocorrelation
2. Causes of autocorrelation
3. Graphical tests for autocorrelation
4. Statistical tests for autocorrelation
5. Remedial measures

   **- Fitting autocorrelated errors (most practical approach)**

# 1.1 Autocorrelation

- Auto= "self"
- Autocorrelation means "correlated with one self"
- The classical multiple linear regression model assumes that

$$\text{Cor}(u_i, u_j) = 0 \ (i \neq j)$$

- However, if we have autocorrelation

$$\text{Cor}(u_i, u_j) \neq 0 \text{ for some } i \neq j$$

- **If the modelling assumptions break down we may not be able to trust computer-generated statistical output**

# 1.2 Autocorrelation

- **Occurs whenever you have correlation between observations that have been in ordered in time or space**
- Observations or measurements taken close together typically take similar values
- Autocorrelation is typically associated with time series data
- **This is especially relevant for datasets in accounting, finance and economics etc if the data have been recorded over time – e.g. dailly, weekly, monthly, quarterly, yearly prices etc**

# 1.3 Nature of autocorrelation

• The classical multiple linear regression model assumes no autocorrelation between the disturbances $u_i$

$$\text{Cor}(u_i, u_j) = 0 \ (i \neq j)$$

• This model assumes that errors are not influenced by errors corresponding to other observations
• E.g. data on real gold price and real stock price
• Naive application of the classical linear regression model would assume that errors pertaining to observations for say 2000 and 2001 would be independent
• **Clearly this is a big assumption**
• **Gives an indication of how autocorrelation can be caused by economic and financial data being collected over time**

# 1.4 Nature of autocorrelation

• If the assumptions underpinning the classical linear regression model are wrong then you cannot take computer-generated statistical output at face value

• Ignoring autocorrelation may

- Under-estimate $\sigma^2$
- Over-estimate $R^2$
- Usual $t$-tests and $F$-tests are no longer valid and may lead to misleading conclusions about statistical significance
- Generally may lead to an over-confidence on the part of the analyst

# 2.1 Causes of autocorrelation

1. Inertia
2. Non-stationarity
3. Model mis-specification
4. Cobweb phenomenon
5. Data limitation and manipulation

# 2.2 Causes of autocorrelation – inertia

- Observations collected together over time
- Inertia
  - Economic time series tend to exhibit cyclical behaviour – touches on really deep themes
  - Examples include GNP, price indices, production figures, employment statistics etc
  - Since these series tend to be quite slow moving
  - Effect of inertia is that successive observations are highly correlated
- **This is an extremely common phenomenon in financial and economic time series**

# 2.3 Causes of autocorrelation – non-stationarity

- Observations collected together over time
- Non-stationarity
  - If $X$ and $Y$ are non-stationary it is possible that the error term will also be non-stationary
  - Error term will then exhibit autocorrelation
  - Common problem in finance and economics
- **Observations recorded over time may lead to autocorrelation**

# 2.4 Causes of autocorrelation – model specification errors

• A model specification error occurs if important variables that should be included in the model are excluded or if the model has the wrong function form. In either case the set-up of the model is incorrect or "mis-specified"

• **Specification errors mean that it is important to consider financial/economic theory and/or other relevant academic literature – mathematics is only part of the story**

• Specification errors mean that residuals from an incorrectly fitted model may exhibit a systematic pattern, rather than a purely random pattern, and so may be autocorrelated

• A commonly encountered example is a failure to account for the lagged effects of economic variables

# 2.5 Causes of autocorrelation – Cobweb Phenomenon

- Commonly encountered with agricultural commodities
- Economic agents like farmers etc commonly base decisions on the prevailing price from last year when deciding how many goods to supply to the market
- E.g. the amount of crops farmers supply to the market at time $t$ might have the form

$$\text{Supply}_t = \beta_1 + \beta_2 P_{t-1} + u_t \tag{1}$$

- The disturbances $u_t$ in equation (1) are therefore unlikely to be completely random and patternless because they represent the actions of intelligent economic agents (e.g. farmers)
- **Net result may be autocorrelated error terms**
- Reflects economic themes of wider importance – i.e. financial and economic variables always represent the collective investment decisions of many different people and so will often be inherently more complicated than the mathematical and statistical tools used to describe them!

# 2.6 Causes of autocorrelation – data manipulation

- For example quarterly data may smooth out some of the wild fluctuations typically seen in monthly sales figures
- For example low frequency census and economic survey data may be interpolated
- Such data transformations may be inevitable and unavoidable, in social sciences unlike physical sciences data quality may be variable, but may induce systematic patterns and autocorrelation into the disturbances $u_i$ of a regression model
- **No magic solution – but this issue remains an important consideration throughout**

# 3.1 Detection of autocorrelation

1. Graphical methods
   - Simpler but may be more robust and more informative in applied project work
2. Statistical tests
   - Runs test
   - Durbin-Watson test

• **In applied project work graphical and statistical methods may serve as a useful cross-check of each other!**

# 3.2 Graphical tests for autocorrelation

1. Time series plot of residuals
   - Plot residuals over time
   - Check to see if any evidence of a systematic pattern exists
2. Auto-correlation plot
   - Natural to think that $u_t$ and $u_{t-1}$ may be correlated
   - Plot $\hat{u}_t$ against $\hat{u}_{t-1}$

# 3.3 Reading the data into R

- Data is in the file `L3eg1data.txt` available on Canvas
- Save to your USB stick then read in the data using the `read.table` command

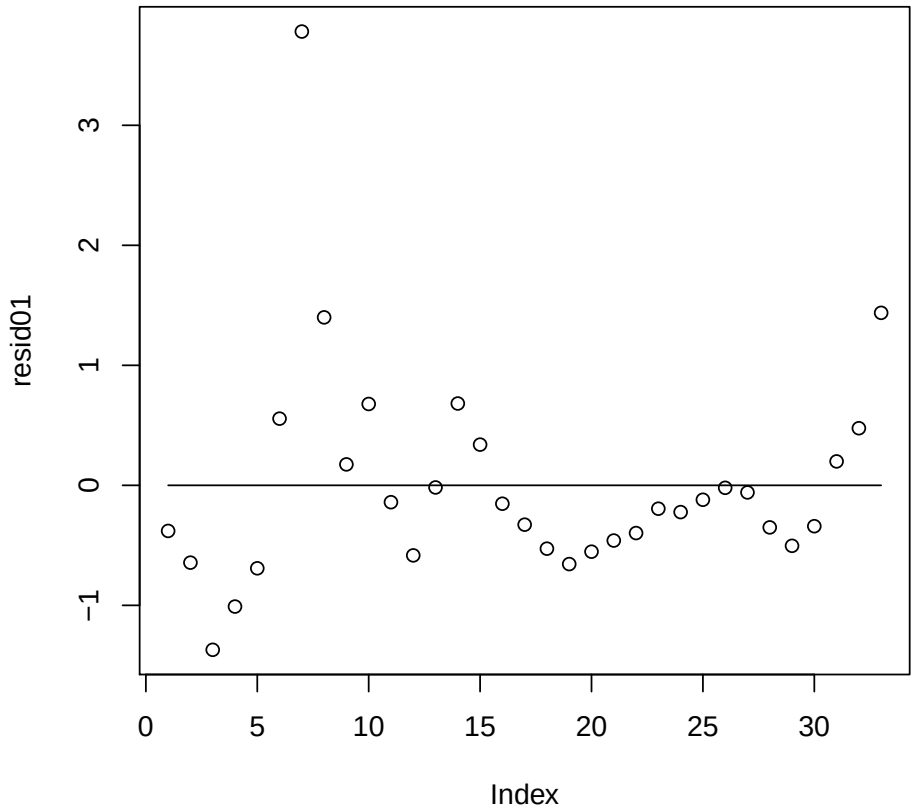
```
data1<-read.table(‘‘E:L3eg1data.txt")
```

- The dataset contains two columns containing the real gold price (left column) and the real stock price (right column)
- In R assign variables linking the real gold price to the first column and the real stock price to the second column. (Note that this has to match the name `data1` given in the above sequence of commands)

```
realgoldprice<-data1[,1]
realstockprice<-data1[,2]
```

# 3.4 Graphical tests for autocorrelation

- Time series plot suggests some evidence for autocorrelation
- Thing to look for here is successive runs of residuals either side of the line



# 3.5 Graphical tests for autocorrelation in R

1. Fit the regression model
2. Time series plot of the residuals produced

1. Fit the regression model

```
a.lm<-lm(realgoldprice~realstockprice)
```

2. Time series plot of the residuals produced

```
resid01<-a.lm$resid
plot(resid01)
length(resid01)
33
lines(seq(1:33), rep(0, 33))
```

# 3.6 Autocorrelation plot of residuals in R

• Autocorrelation plot of residuals

```
length(resid01)
```

33

**Lag 0**

- Lose one observation here – actually the first observation

```
resid12<-resid01[2:33]
```

**Lag 1**

- Lose one observation here – actually the last observation

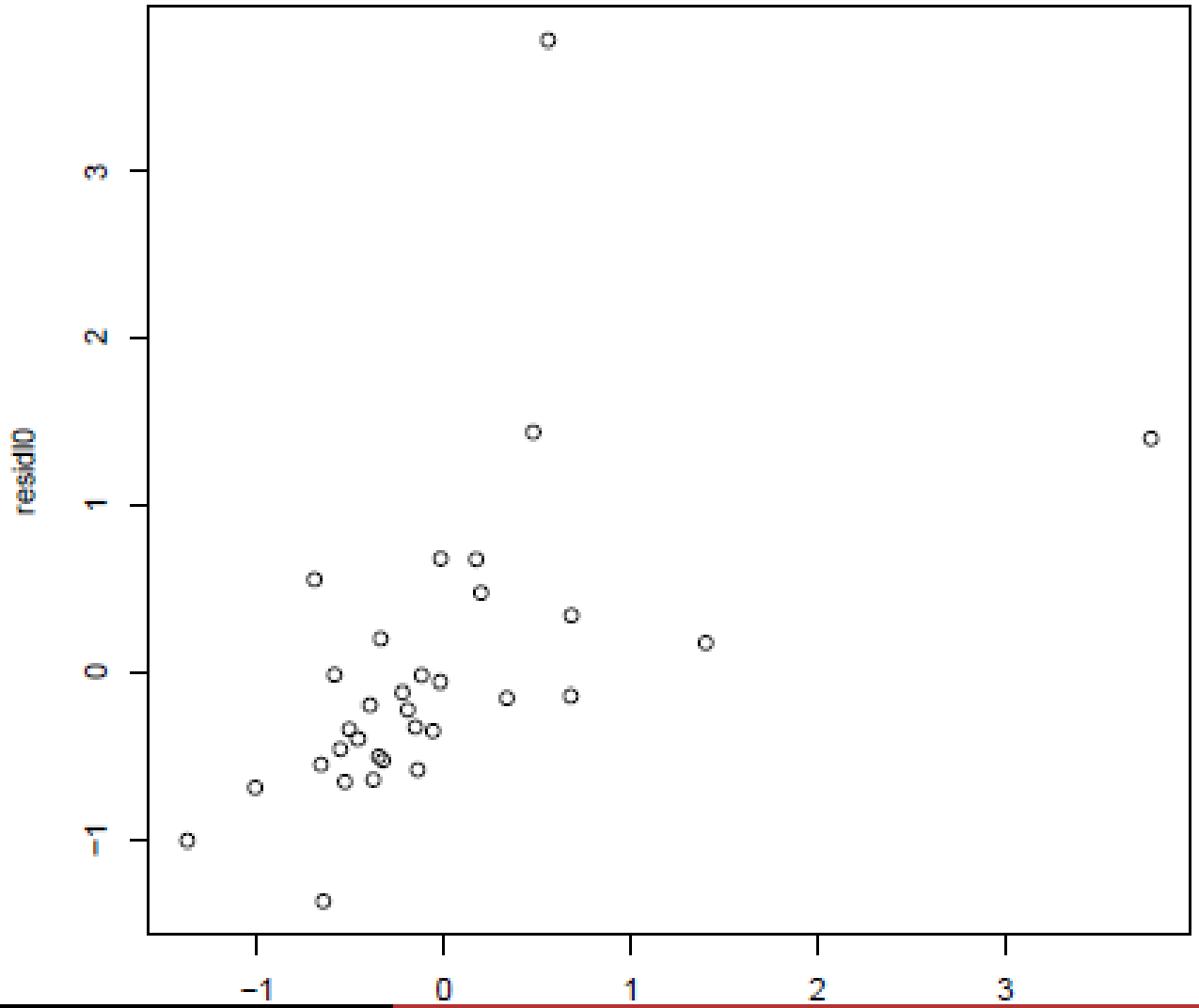
```
resid11<-resid01[1:32]
```

• **Then just have to use the basic `plot` command**

```
plot(resid11, resid12)
```

# 3.7 Autocorrelation plot

- Autocorrelation plot suggests positive autocorrelation



# 4.1 Statistical tests for autocorrelation

1. Runs test
   - Under the classical multiple linear regression model residuals are equally likely to be positive or negative
2. Durbin-Watson test
   - Test to see if residuals are AR(1)
3. Breusch-Godfrey Lagrange Multiplier test
   - More general test for autocorrelation but outside the scope of this course

# 4.2 Runs test

- Conceptually simple but not very powerful
- Define

$$\begin{aligned} R &= \text{Number of runs} \\ E[R] &= \frac{2N_1N_2}{N} + 1 \\ \sigma_R^2 &= \frac{2N_1N_2(2N_1N_2 - N)}{N^2(N-1)} \end{aligned}$$

- $N$=Total number of observations, $N_1$=number of positive residuals, $N_2$=number of negative residuals
- Perform a $t$-test using the bottom row of your $t$-tables (in case this sounds incredibly complicated this is just an approximation based on using the normal distribution)

$$t = \frac{R - E[R]}{\sigma_R}$$

# 4.3 Runs test – example

- In the gold market example the residuals have the following signs (-,-,-,-,-)(+,+,+,+,+)(-,-,-)(+,+)(-,-,-,-,-,-,-,-,-,-,-,-,-,-,-)(+,+,+)
- This gives us R=6 runs or 6 different subsequences of residuals that each share the same sign
- Form the $t$-statistic ($N_1 = 10$ positive residuals, $N_2 = 23$ negative residuals, $N = 33$)

$$|t| = \frac{R - E[R]}{\sigma_R} = \frac{|6 - 14.9394|}{\sqrt{5.6365}} = 3.765 \text{ (3 d.p.)}$$

- From tables $t_\infty(0.025) = 1.96$
- So there is evidence from the runs test ($p < 0.05$) that there is autocorrelation in the residuals

# 4.4 Running a runs test in R

1. Produce a character vector in R that determines if the residuals are positive or negative using the command `factor`
   ```
   residsign<-1*(resid01>0)
   residsign<-factor(residsign)
   ```
2. Load the `tseries` package in R (see below)
3. Apply the function `runs.test`
   ```
   runs.test(residsign)
   ```

• This presents evidence `Standard Normal = -3.7653`, `p-value = 0.0001663<0.05` that the residuals are autocorrelated

# 4.5 Background: using optional packages in R

- Want to run a runs test in R
- To do this you have to download the R package `tseries` – which stands for time series
- In R to upload packages use

`Packages` $\longrightarrow$ `load packages` $\longrightarrow$ `tseries` $\longrightarrow$ `OK`

# 4.6 Installing non-standard packages from CRAN

- In R to see what packages are available for loading use `Packages`$\longrightarrow$`Load packages`
- If the package is not there you might have to download the required package from CRAN
- To do this use `Packages`$\longrightarrow$`Install package(s) ...`$\longrightarrow$`Choose a CRAN mirror`
- Better to choose the UK (London or Bristol) or wherever you are in the world
- You should then be able to see a long list of packages that are available for download
- **This online community and long list of written packages is really the best thing about R**

# 4.7 Potential problems with the university computer network

- On my work computer standard packages load fine
- On your personal computer should be able to update the list of packages via the R repository CRAN
- This maintenance of packages and an active online R community is probably the best thing about R and the best reason to use it
- However, it may not be possible to update packages directly on a PC on the university network. This may be different for different institutions and indeed for different rooms on campus
- **The work-around this is to save the package code to a USB stick and then use the option**

`Packages`$\longrightarrow$`Install package(s) from local files ...`

- **Best thing might be to bring your laptop into class and see if we can get R working ...**
- **On my laptop I found loading R packages fine but fiddlier than it should have been ...**

# 4.8 Durbin-Watson test

• **What proportion of the residual sum of squares can be explained by the correlation between successive residuals**

- (Residual sum of squares means information that cannot be explained by the model)

• Durbin-Watson test is important historically

• **My experience suggests that the Durbin-Watson test is impractical for large problems**

# 4.9 Durbin-Watson test

- Most popular econometric appraoch
- Durbin-Watson $d$-statistic is defined as the the sum of squared differences in successive residuals relative to the residual sum of squares

$$
\begin{aligned}
d &= \frac{\text{Squared distance between successive residuals}}{\text{Residual Sum of Squares}} \\
&= \frac{\sum_{t=2}^{n} \left(\hat{u}_t - \hat{u}_{t-1}\right)^2}{\sum_{t=2}^{n} \hat{u}_t^2}
\end{aligned}
$$

# 4.10 Assumptions underpinning the Durbin-Watson test – Gujarati and Porter, Ch. 12

1. The regression model includes the intercept term
2. The $X$ variables are non-stochastic
3. The disturbances are auto-correlated of order one:

$$u_t = \rho u_{t-1} + \epsilon_t$$

4. The error term $u_t$ is normally distributed
5. The regression model does not include lagged values of the dependent variable, such as $Y_{t-1}$ etc, on the right hand side
6. There are no missing observations in the data

# 4.11 Interpretation of the Durbin-Watson statistic

| Null hypothesis | Decision | If |
|---|---|---|
| No positive autocorrelation | Reject | $0 < d < d_L$ |
| No positive autocorrelation | No decision | $d_L \leq d \leq d_U$ |
| No negative autocorrelation | Reject | $4 - d_L < d < 4$ |
| No negative autocorrelation | No decision | $4 - d_u \leq d \leq 4 - d_L$ |
| No autocorrelation (positive or negative) | Do not reject | $d_U < d < 4 - d_U$ |

# 4.12 Durbin-Watson test in R

- Need to load the R package `lmtest` which stands for linear model (regression tests)
- This works by applying the function `dwtest` to a named linear regression model
- For this lecture example

```
dwtest(a.lm)
DW = 0.90424, p-value = 0.0001215
alternative hypothesis:  true autocorrelation is
greater than 0
```

# 5.1 Regression with autocorrelated residuals

- Standard regression theory and programmes assume that the regression residuals are uncorrelated
- Slides 3.6 and Slide 3.10 present graphical evidence that residuals are correlated
- Slides 4.4 and 4.12 present formal statistical evidence that residuals are correlated
- **It is important here that the graphical and statistical evidence serve as a cross-check of each other**
- Also "know" from the wider context that since gold and stock prices are collected over time some autocorrelation is likely to be present
- This is a practical problem with a practical solution being to fit a regression model with autocorrelated errors

# 5.2 Regression with autocorrelated errors in R

- The basic R command that is needed here is `arima` which can be used here to combine regression with an element of time series
- The way this works is as follows

`arima(y-variable, xreg=X-variables, order=c(`$p, d, q$`)`

- In the `xreg` part specify the regression formula to be used in the model
- In the `order` section specify the ARIMA($p, d, q$) component of the model. For this module we would only really need `order=c(1, 0, 0)` to specify an autoregressive model of order 1
- For our lecture example use

```
arima(realgoldprice, xreg=realstockprice, order=c(1, 0, 0))
```

# 5.3 Analysis of regression with autocorrelated errors in R

• The R commands on Slide 5.2 produce the following (have to compute the $t$-statistics yourself!)

```
Coefficients:
ar1 intercept realstockprice
0.5578 3.9406 -0.0487
s.e.  0.1545 0.5675 0.0247
```

• `length(goldprice)`=33 so this tells us we have 33-3 estimated parameters =30 residual degrees of freedom

• Calculate the $t$-statistics as

```
coeff<-c(0.5578, 3.9406, -0.0487)
ese<-c(0.1545, 0.5675, 0.0247)
t<-abs(coeff)/ese
t
3.610356 6.943789 1.971660
2*(1-pt(t, 30))
1.100235e-03 1.032974e-07 5.793185e-02
```

# 5.4 Interpretation of the results

• From the $t$-statistics evidence of autocorrelation in the residuals ($p = 0.001100235$)

- Autocorrelation reflects the fact that it is harder to establish a genuine link between real gold prices and real stock prices than might be first thought

• Weak evidence of a relationship between the real stock price and the real gold price ($p = 0.05793185$)

- On this course continue to investigate whether or not $p < 0.05$ or $p > 0.05$

- Being in this corridor of uncertainty $0.05 < p < 0.1$ reflects that the interpretation may be more nuanced in harder examples

- **Issues over nuanced interpretations have been important in MSc dissertations in the past**