



ECON90024 – FORECASTING IN ECONOMICS & BUSINESS

LECTURE 8: STOCHASTIC TRENDS, UNIT ROOT TESTS AND ARCH

TODAY'S LECTURE

- Deterministic vs. Stochastic Trends
- Unit Root Tests
- Overview of what we have learned so far.
- ARCH Models

DETERMINISTIC VS. STOCHASTIC TRENDS

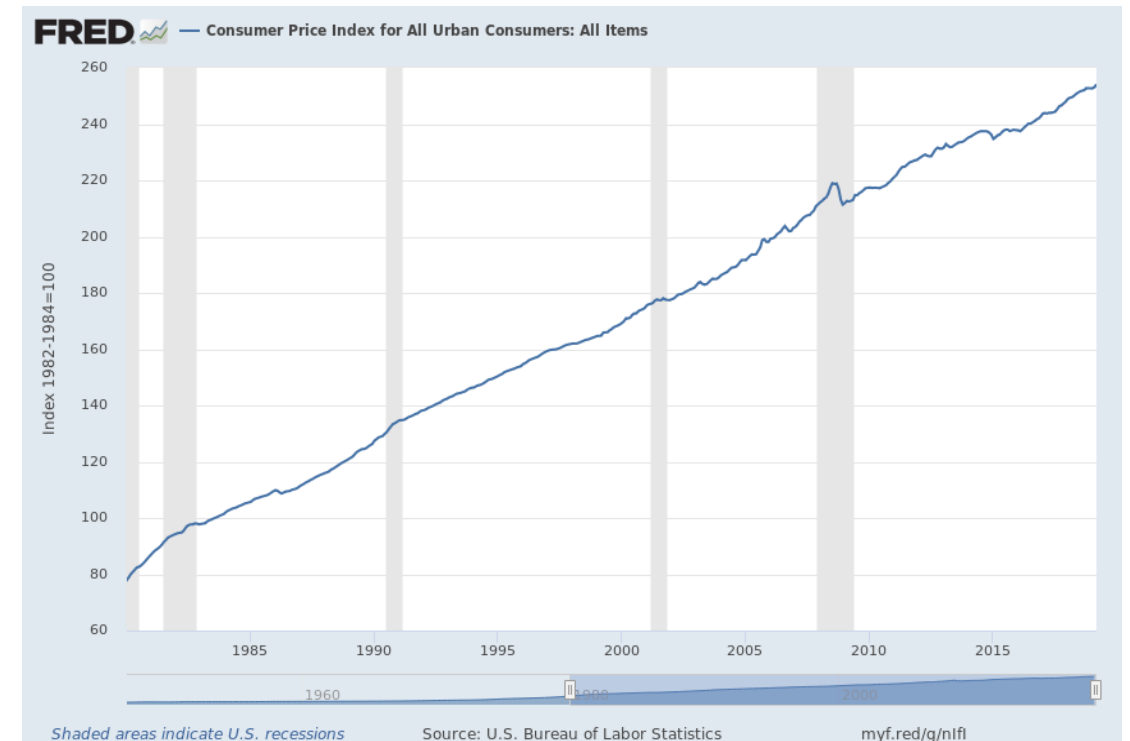
- When faced with a time series that exhibits a clear trend, our approach so far has been to specify a time series model with a deterministic trend:

$$Y_t = \alpha + \delta t + u_t$$

$$u_t = \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \dots$$

$$\varepsilon_t \sim i.i.d.(0, \sigma^2)$$

- We call such a process **trend stationary** because if one subtracts the trend $\alpha + \delta t$, the result is a stationary process.
- Deterministic trends are easy to estimate and understand. However, in specifying a deterministic trend model, we are making a very strong assumption about the data generating process.



DETERMINISTIC VS. STOCHASTIC TRENDS

- An alternative way to characterize a trend in a time series is to treat it as the result of an accumulation of permanent shocks (i.e. **a stochastic trend**). An example of such a process would be a random walk with drift

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

$$\varepsilon_t \sim i.i.d.(0, \sigma^2)$$

- Such a process is often called a **unit root process** because one or more of the solutions to the lag polynomial is unity (i.e. equal to 1). To see this, note that we can rewrite the above model as,

$$(1 - L)Y_t = \delta + \varepsilon_t$$

- Clearly the solution to $(1 - z) = 0$ is $z = 1$.

DETERMINISTIC VS. STOCHASTIC TRENDS

Deterministic Trend	Stochastic Trend
$Y_t = \delta t + \varepsilon_t$	$Y_t = Y_{t-1} + \delta + \varepsilon_t$ $Y_t = \delta + \delta + \dots + \varepsilon_t + \varepsilon_{t-1} + \dots + \varepsilon_0$
$E[Y_t] = \delta t$ $var(Y_t) = \sigma^2$	$E[Y_t] = \delta t$ $var(Y_t) = t\sigma^2$

- Here we are supposing that $Y_0 = 0$ and that $\varepsilon_t \sim i.i.d.(0, \sigma^2)$
- We can see that whether we specify a data generating process as having a deterministic trend or a stochastic trend will have quite different implications for forecasting!

DETERMINISTIC VS. STOCHASTIC TRENDS

- To compare the forecasts, let's first consider the generalized deterministic trend model

$$Y_t = \alpha + \delta t + \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \dots$$

$$\varepsilon_t \sim i.i.d.(0, \sigma^2)$$

- This is a generalized specification because Wold's Representation theorem tells us that any covariance stationary process may be written as an infinite distributed lag of white noise,

$$\Psi(L)\varepsilon_t = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}$$

- Conditioning on Ω_t , it follows that the h -step ahead forecast will be given by

$$E[Y_{t+h} | \Omega_t] = \alpha + \delta(t+h) + \psi_h \varepsilon_t + \psi_{h+1} \varepsilon_{t-1} + \dots$$

DETERMINISTIC VS. STOCHASTIC TRENDS

- The h -step ahead forecast error will then be,

$$Y_{t+h} - E[Y_{t+h} | \Omega_t] = \varepsilon_{t+h} + \psi_1 \varepsilon_{t+h-1} + \cdots + \psi_{h-1} \varepsilon_{t+1}$$

- With the forecast error variance being,

$$\sigma_h^2 = \sigma^2(1 + \psi_1^2 + \cdots + \psi_{h-2}^2 + \psi_{h-1}^2)$$

- As we increase the the forecast horizon and let $h \rightarrow \infty$, the forecast error variance will grow, however it will be bounded since Wold's representation theorem also states that

$$\sum_{i=0}^{\infty} \psi_i^2 < \infty$$

- Therefore, the added uncertainty from forecasting a covariance stationary process farther into the future becomes negligible.

DETERMINISTIC VS. STOCHASTIC TRENDS

- Let's now consider a generalized unit root process,

$$Y_t = Y_{t-1} + \delta + \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \dots$$

$$\varepsilon_t \sim i.i.d.(0, \sigma^2)$$

- The first thing to recognize is that the first difference of a unit root process will be covariance stationary,

$$Y_t - Y_{t-1} = (1 - L)Y_t = \Delta Y_t = \delta + \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \dots$$

- Such a process is also commonly known as an *integrated process of order 1*.
- A process with two unit roots will require a second difference to be taken in order to achieve covariance stationarity and will therefore be an integrated process of order 2.

$$(1 - L)^2 Y_t = \kappa + \psi(L) \varepsilon_t$$

- A general process is denoted as an ARIMA(p, d, q), where p refers to the number of autoregressive lags, d refers to the order of integration, and q refers to the number of moving average lags.

DETERMINISTIC VS. STOCHASTIC TRENDS

- To compute a h -step ahead forecast of a unit root process, we first recognize that a h -step ahead forecast of its change can be computed as,

$$E[\Delta Y_{t+h} | \Omega_t] = \Delta \hat{Y}_{t+h|t} = \delta + \psi_h \varepsilon_t + \psi_{h+1} \varepsilon_{t-1} + \dots$$

- Since the level of the variable at time $t + h$ is simply the sum of changes between t and $t + h$,

$$Y_{t+h} = \Delta Y_{t+h} + \Delta Y_{t+h-1} + \dots + \Delta Y_{t+1} + Y_t$$

- Then it must be case that the h -step ahead forecast be given by

$$E[Y_{t+h} | \Omega_t] = \hat{Y}_{t+h|t} = \Delta \hat{Y}_{t+h|t} + \Delta \hat{Y}_{t+h-1|t} + \dots + \Delta \hat{Y}_{t+1|t} + Y_t$$

DETERMINISTIC VS. STOCHASTIC TRENDS

- Writing it all out, we have that

$$\begin{aligned} E[Y_{t+h}|\Omega_t] &= \hat{Y}_{t+h|t} \\ &= \{\delta + \psi_h \varepsilon_t + \psi_{h+1} \varepsilon_{t-1} + \dots\} + \{\delta + \psi_{h-1} \varepsilon_t + \psi_h \varepsilon_{t-1} + \dots\} + \dots + \{\delta + \psi_1 \varepsilon_t + \psi_2 \varepsilon_{t-1} + \dots\} + Y_t \end{aligned}$$

- Collecting the terms, we obtain,

$$E[Y_{t+h}|\Omega_t] = \hat{Y}_{t+h|t} = h\delta + Y_t + (\psi_h + \psi_{h-1} + \dots + \psi_1)\varepsilon_t + (\psi_{h+1} + \psi_h + \dots + \psi_2)\varepsilon_{t-1} + \dots$$

- Then, the forecast error will be given by,

$$Y_{t+h} - \hat{Y}_{t+h|t} = \{\Delta Y_{t+h} + \Delta Y_{t+h-1} + \dots + \Delta Y_{t+1} + Y_t\} - \{\Delta \hat{Y}_{t+h|t} + \Delta \hat{Y}_{t+h-1|t} + \dots + \Delta \hat{Y}_{t+1|t} + Y_t\}$$

DETERMINISTIC VS. STOCHASTIC TRENDS

- To see what the forecast error looks like in terms of the innovations, we recognize that,

$$\Delta Y_{t+h} - \Delta \hat{Y}_{t+h|t} = \{\varepsilon_{t+h} + \psi_1 \varepsilon_{t+h-1} + \cdots + \psi_{h-1} \varepsilon_{t+1}\}$$

$$\Delta Y_{t+h-1} - \Delta \hat{Y}_{t+h-1|t} = \{\varepsilon_{t+h-1} + \psi_1 \varepsilon_{t+h-2} + \cdots + \psi_{h-2} \varepsilon_{t+1}\}$$

\vdots

$$\Delta Y_{t+1} - \Delta \hat{Y}_{t+1|t} = \varepsilon_{t+1}$$

- Then, putting it all together we obtain,

$$Y_{t+h} - \hat{Y}_{t+h|t} = \varepsilon_{t+h} + (1 + \psi_1) \varepsilon_{t+h-1} + (1 + \psi_1 + \psi_2) \varepsilon_{t+h-2} + \cdots + (1 + \psi_1 + \psi_2 + \cdots + \psi_{h-1}) \varepsilon_{t+1}$$

DETERMINISTIC VS. STOCHASTIC TRENDS

- It follows that the forecast error variance (or the mean squared error MSE) will be given by,

$$E \left[(Y_{t+h} - \hat{Y}_{t+h|t})^2 \right] = \sigma_h^2 = \{1 + (1 + \psi_1)^2 + (1 + \psi_1 + \psi_2)^2 + \cdots + (1 + \psi_1 + \psi_2 + \cdots + \psi_{h-1})^2\} \sigma^2$$

- The forecast error variance again increases with the length of the forecasting horizon h , though in contrast to the trend-stationary case, it does not converge to any fixed value as $h \rightarrow \infty$.
- Instead, it asymptotically approaches a linear function of h with slope

$$(1 + \psi_1 + \psi_2 + \cdots + \psi_{h-1})^2 \sigma^2$$

DETERMINISTIC VS. STOCHASTIC TRENDS

- To see this more clearly, let's consider an ARIMA(0,1,1) model,

$$Y_t = Y_{t-1} + \delta + \varepsilon_t + \theta \varepsilon_{t-1}$$

$$\varepsilon_t \sim i.i.d.(0, \sigma^2)$$

- In this specific case, the h -step ahead forecast error variance will simplify from,

$$E \left[(Y_{t+h} - \hat{Y}_{t+h|t})^2 \right] = \sigma_h^2 = \{1 + (1 + \theta)^2 + (1 + \theta)^2 + \dots + (1 + \theta)^2\} \sigma^2$$

- To

$$E \left[(Y_{t+h} - \hat{Y}_{t+h|t})^2 \right] = \sigma_h^2 = \{1 + (h - 1)(1 + \theta)^2\} \sigma^2$$

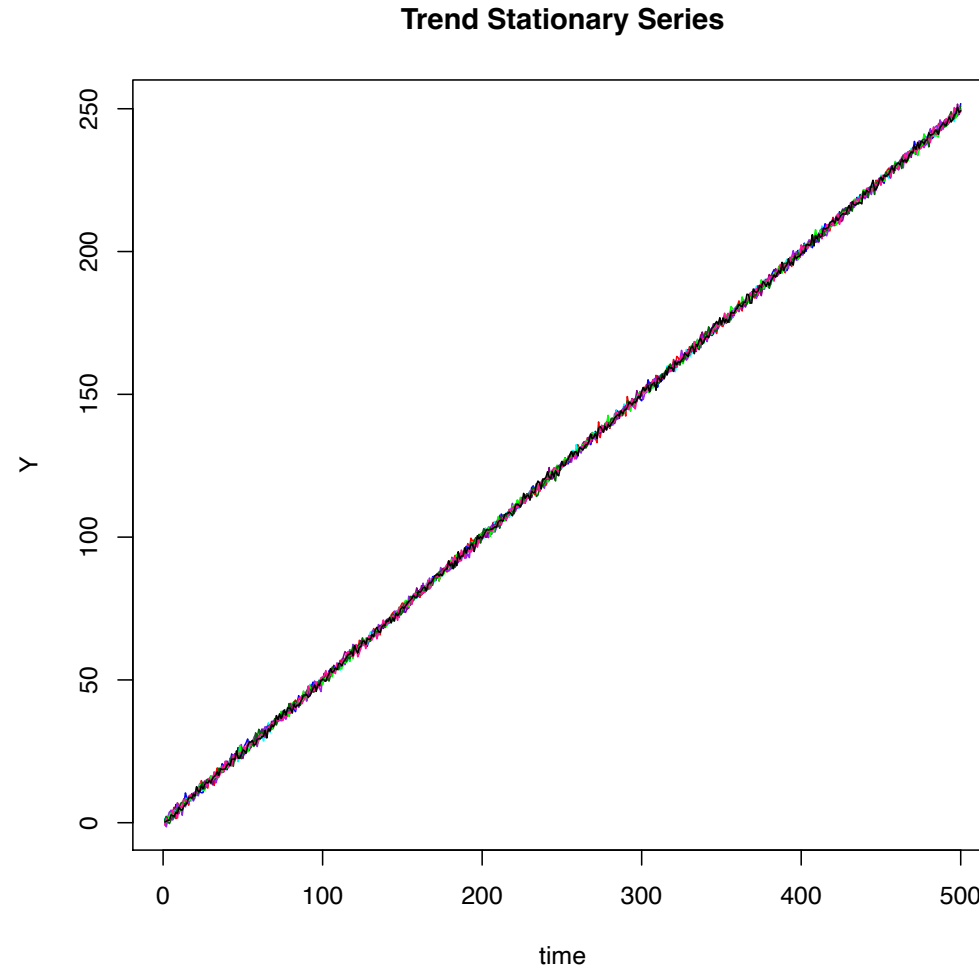
- Therefore, we have shown explicitly that for a stationary process, the forecast error variance reaches a finite bound as the forecast horizon grows large, whereas for a unit root process, the forecast error variance grows linearly with the forecast horizon.

DETERMINISTIC VS. STOCHASTIC TRENDS

- Here are some simulated trend stationary series that are generated from the model,

$$Y_t = 0.5t + \varepsilon_t$$

$$\varepsilon_t \sim i.i.d. N(0,1)$$

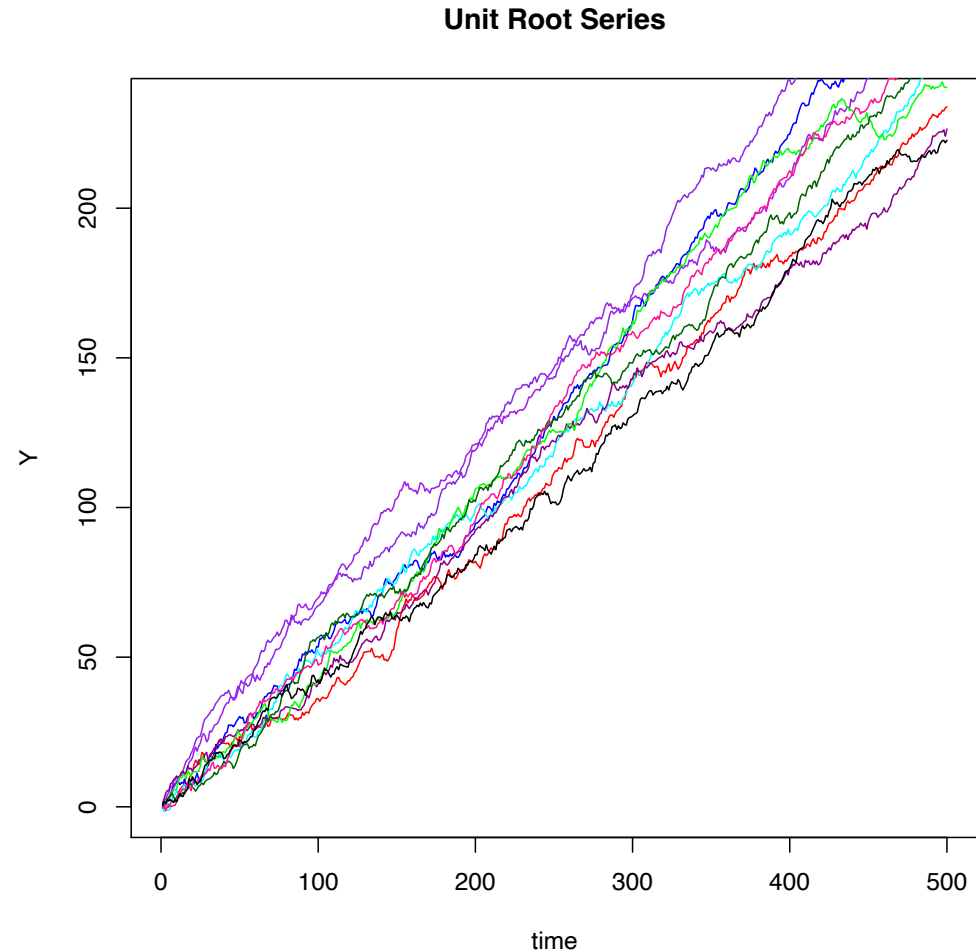


DETERMINISTIC VS. STOCHASTIC TRENDS

- Here are some simulated series from a unit root process that are generated from the model,

$$Y_t = Y_{t-1} + 0.5 + \varepsilon_t$$

$$\varepsilon_t \sim i.i.d. N(0,1)$$

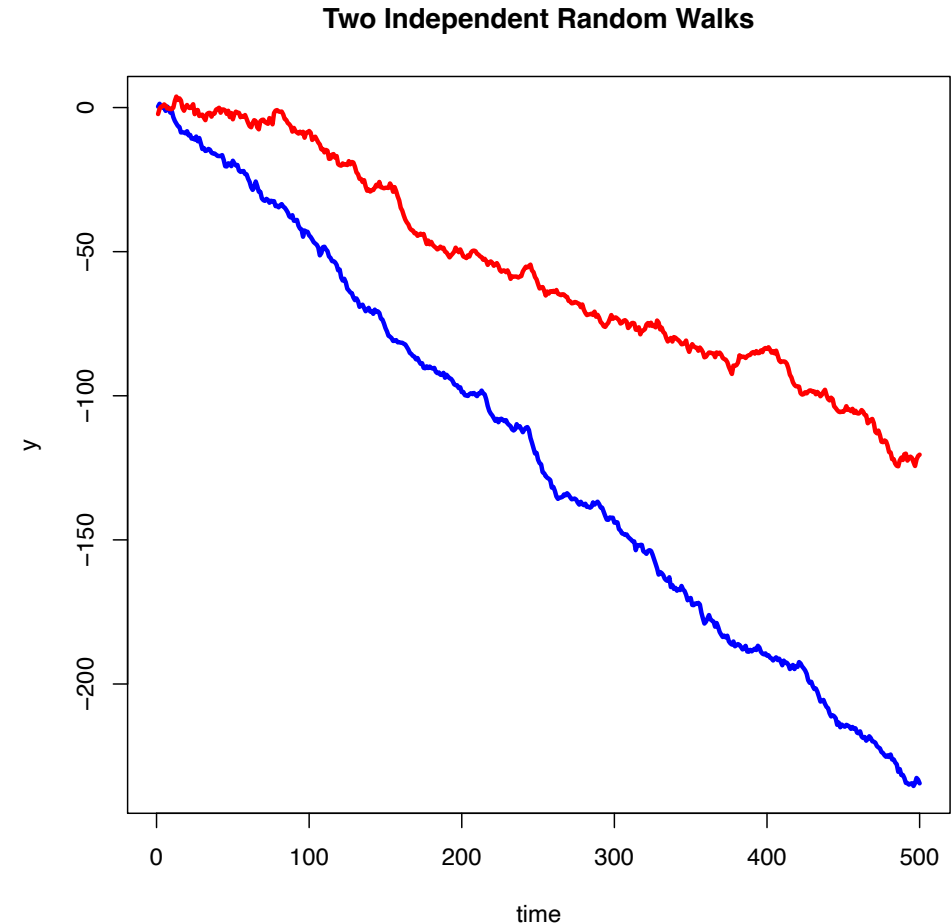


SPURIOUS TIME SERIES REGRESSIONS

- We have to be extremely careful when we specify and estimate time series models that involve multiple variables that have unit roots.
- Two variables that are integrated could exhibit an extremely high correlation without an underlying structural relationship. This often results in what is known as a ***spurious regression***.
- To illustrate, let's consider two independent random walks:

$$Y_t = Y_{t-1} - 0.5 + \varepsilon_t$$

$$X_t = X_{t-1} - 0.2 + u_t$$



SPURIOUS TIME SERIES REGRESSIONS

- When we estimate a regression, we obtain the following results:

```
Call:
lm(formula = y ~ x)

Residuals:
    Min       1Q   Median       3Q      Max
-21.3141  -6.4463  -0.2548   7.1985  20.1954

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.72330    0.73319  -21.45  <2e-16 ***
x             1.83997    0.01082  170.11  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.998 on 498 degrees of freedom
Multiple R-squared:  0.9831, Adjusted R-squared:  0.983
F-statistic: 2.894e+04 on 1 and 498 DF, p-value: < 2.2e-16
```

- These estimates, standard errors and R^2 are telling us that we could use X to forecast Y , but this is clearly not the case!

SPURIOUS TIME SERIES REGRESSIONS

- If we compute the **first differences** of each variable $\Delta Y_t = Y_t - Y_{t-1}$, $\Delta X_t = X_t - X_{t-1}$ and estimate the same regression, we obtain,

```
Call:
lm(formula = ydiff ~ xdiff)

Residuals:
    Min       1Q   Median       3Q      Max
-2.95812 -0.60398  0.01043  0.66872  2.72752

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.46016    0.04354  -10.569  <2e-16 ***
xdiff        0.04379    0.04001   1.094    0.274
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9492 on 497 degrees of freedom
Multiple R-squared:  0.002404,    Adjusted R-squared:  0.0003971
F-statistic: 1.198 on 1 and 497 DF,  p-value: 0.2743
```

- The stochastic trends of these two independent random walks can make it seem as if both series share the same trend, especially in small samples. However, when we take the first differences, the association disappears.

TESTING FOR UNIT ROOTS

- As the previous slides have shown, the presence of unit roots in the data generating process will have deep implications when it comes to forecasting.
- Since it can be difficult to determine the existence of a unit root simply through a visual inspection of the time series, we should always rely on a statistical approach.
- As a first step, let's consider the AR(1) model,

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

- If we subtract Y_{t-1} from both sides of the equation, we obtain,

$$Y_t - Y_{t-1} = \Delta Y_t = (\phi - 1)Y_{t-1} + \varepsilon_t$$

- From this representation, we can see that if Y_t is a random walk, $\phi = 1$ and thus $(\phi - 1) = 0$. Therefore, we could run the following simple linear regression

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

- And test the null hypothesis $H_0: \rho = 0$! This is known as the Dickey-Fuller test for a unit root.

TESTING FOR UNIT ROOTS

- The Dickey-Fuller test statistic takes the same form as the standard t statistic for a test of significance.

$$DF = \frac{\hat{\rho} - \rho}{\hat{\sigma}_{\hat{\rho}}} = \frac{\hat{\rho}}{\hat{\sigma}_{\hat{\rho}}}$$

- Most software packages will perform the Dickey-Fuller test with the option of the following specifications:

1. $\Delta Y_t = \rho Y_{t-1} + \varepsilon_t$
2. $\Delta Y_t = \mu + \rho Y_{t-1} + \varepsilon_t$
3. $\Delta Y_t = \mu + \delta t + \rho Y_{t-1} + \varepsilon_t$

TESTING FOR UNIT ROOTS

- Unit roots can also manifest across higher order autoregressive dynamics. Consider the following AR(3) model,

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \varepsilon_t$$

- Written in terms of the lag operator,

$$(1 - \phi_1 L - \phi_2 L^2 - \phi_3 L^3)Y_t = \varepsilon_t$$

- This process will possess a unit root if $z = 1$ is a solution to the polynomial equation

$$(1 - \phi_1 z - \phi_2 z^2 - \phi_3 z^3) = 0$$

- Thus if we set $z = 1$, this will imply that

$$\phi_1 + \phi_2 + \phi_3 = 1$$

TESTING FOR UNIT ROOTS

- If we take our original model

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \varepsilon_t$$

- And we add and subtract the terms $\phi_2 Y_{t-1}$, $\phi_3 Y_{t-1}$, $\phi_3 Y_{t-2}$ to the right hand side, we will obtain,

$$Y_t = (\phi_1 + \phi_2 + \phi_3)Y_{t-1} - (\phi_2 + \phi_3)(Y_{t-1} - Y_{t-2}) + \phi_3(Y_{t-2} - Y_{t-3}) + \varepsilon_t$$

- Which we can write more compactly as

$$Y_t = \beta_0 Y_{t-1} + \beta_1 (Y_{t-1} - Y_{t-2}) + \beta_2 (Y_{t-2} - Y_{t-3}) + \varepsilon_t$$

- Again, if we subtract Y_{t-1} from both sides, we obtain

$$Y_t - Y_{t-1} = \Delta Y_t = (\beta_0 - 1)Y_{t-1} + \beta_1 \Delta Y_{t-1} + \beta_2 \Delta Y_{t-2} + \varepsilon_t = \rho Y_{t-1} + \beta_1 \Delta Y_{t-1} + \beta_2 \Delta Y_{t-2} + \varepsilon_t$$

- Under the null hypothesis of a unit root, $\rho = \phi_1 + \phi_2 + \phi_3 - 1 = 0$.

TESTING FOR UNIT ROOTS

- In general, an **Augmented Dickey-Fuller** test with k lags will involve the following estimated regression,

$$\Delta Y_t = \rho Y_{t-1} + \sum_{j=2}^k \beta_j \Delta Y_{t-j+1} + \varepsilon_t$$

- Where again, the test statistic is computed as,

$$ADF = \frac{\hat{\rho} - \rho}{\hat{\sigma}_{\hat{\rho}}} = \frac{\hat{\rho}}{\hat{\sigma}_{\hat{\rho}}}$$

- As with the standard Dickey-Fuller test, most statistical packages will give you the option of the following specifications:

1. $\Delta Y_t = \rho Y_{t-1} + \sum_{j=2}^k \beta_j \Delta Y_{t-j+1} + \varepsilon_t$
2. $\Delta Y_t = \mu + \rho Y_{t-1} + \sum_{j=2}^k \beta_j \Delta Y_{t-j+1} + \varepsilon_t$
3. $\Delta Y_t = \mu + \delta t + \rho Y_{t-1} + \sum_{j=2}^k \beta_j \Delta Y_{t-j+1} + \varepsilon_t$

- It must be noted however that the sampling distribution of this test statistics will be nonstandard. The critical values for the test cannot be read off a standard Student t table. Software packages will report p-values that correspond to a special set of Dickey-Fuller distribution.

SELECTING LAG LENGTH FOR ADF TEST

- When performing an Augmented Dickey-Fuller test we have to select an appropriate lag length k .
- If we set k to be too small, then we may fail to account for any higher order dynamics that may exist. This may lead us to incorrectly reject the null-hypothesis of a unit root (i.e. make a Type-I error).
- If we set k to be too large, then this increases the probability of making a Type-II error (i.e. wrongly failing to reject the null hypothesis of a unit root).

SELECTING LAG LENGTH FOR ADF TEST

- To find an appropriate k , we follow the procedure set out by Ng & Perron (1995):

1. Set $k_{max} = \left\lceil 12 \left(\frac{T}{100} \right)^{1/4} \right\rceil$
2. Estimate the ADF test regression with $k = k_{max}$
3. If the absolute value of the t-statistic for testing the significance of the last lagged difference is greater than 1.6 then set $k = k_{max}$ and perform the unit root test.
4. Otherwise, reduce the lag length by one and repeat the process.

A GENERAL APPROACH TO ESTIMATING A TIME SERIES MODEL

1. Plot the time series data that you have collected.
2. Specify and estimate using OLS any deterministic components (i.e. trend and seasonality) that you believe to exist, along with the mean of the series.
3. Generate the residuals from your deterministic regression. This will be the demeaned, detrended and seasonally adjusted series.
4. Perform an ADF test on your demeaned, detrended and seasonally adjusted series to determine the order of integration.
5. If the null-hypothesis of a unit root is not rejected, take the first difference of the data and perform the unit root test again. Once you reject the null-hypothesis of a unit root, proceed to the next step with your differenced data.

A GENERAL APPROACH TO ESTIMATING A TIME SERIES MODEL

6. Generate the sample ACF and PACF of the demeaned, detrended and seasonally adjusted series. These will give you a sense of the dependence structure of the time series.
7. Estimate a range of $\text{ARMA}(p, q)$ and choose your preferred model using the AIC and BIC.
8. Once you have chosen your preferred model, you can use it to compute point and interval forecasts!

WHAT WE HAVE LEARNED SO FAR

- We began by conceptualizing all time series as comprising of a set of fluctuations occurring at different time scales:

$$Y_t = T_t + S_t + C_t + \varepsilon_t$$

- The trend T_t and seasonal S_t components can be modeled in a deterministic way, however, the shape and details of these deterministic components must be determined in a logical way (i.e. not arbitrary!). Moreover, we can estimate deterministic trends and seasons easily using OLS.
- Cyclical fluctuations C_t are stable and can be represented using covariance-stationary time series models.

WHAT WE HAVE LEARNED SO FAR

- A covariance-stationary time series must have the following stochastic properties:
 1. A constant mean: $E[Y_t] = \mu$
 2. A finite variance: $E[(Y_t - \mu)^2] < \infty$
 3. The covariance between any two terms in the series depends only on the relative position of the two terms.

$$E[(Y_t - \mu)(Y_{t-j} - \mu)] = E[(Y_{t+k} - \mu)(Y_{t+k-j} - \mu)] = \gamma_j$$

$$\forall j, k$$

- These three properties ensure a stable dependence structure over time. When this is true, the past can tell us a lot about the future.

WHAT WE HAVE LEARNED SO FAR

- If a time series Y_t is covariance stationary, then Wold's representation hypothesis tell us that it can be represented as an infinite distributed lag of white noise, also known as the ***general linear model***:

$$Y_t = \Psi(L)\varepsilon_t = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}$$

$$\varepsilon_t \sim iid(0, \sigma^2)$$

- Where

$$\psi_0 = 1$$

$$\sum_{i=0}^{\infty} \psi_i^2 < \infty$$

- This is a powerful result as it tells us that even though we can never know the true form of a covariance stationary data generating process, it can always be represented in this way!

WHAT WE HAVE LEARNED SO FAR

- The general linear model has an infinite number of parameters $\{\psi_i\}_{i=1}^{\infty}$ which means that it can never be estimated with a finite data set.
- We must instead approximate the general linear model with an ARMA(p, q) of which $AR(p)$ and $MA(q)$ models are special cases.

$$Y_t = \phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}$$

$$\varepsilon_t \sim iid(0, \sigma^2)$$

- Written in lag operator notation:

$$\Phi(L)Y_t = \Theta(L)\varepsilon_t$$

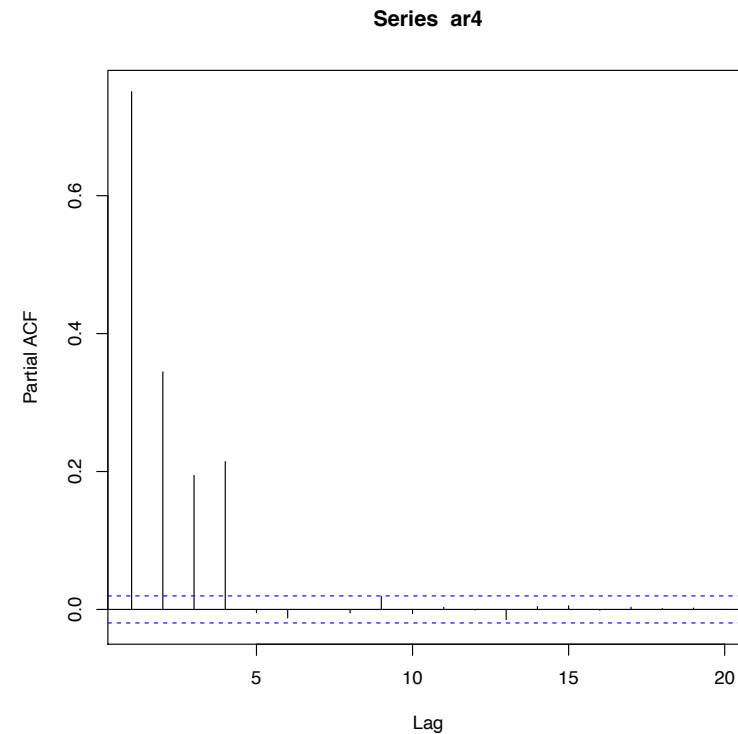
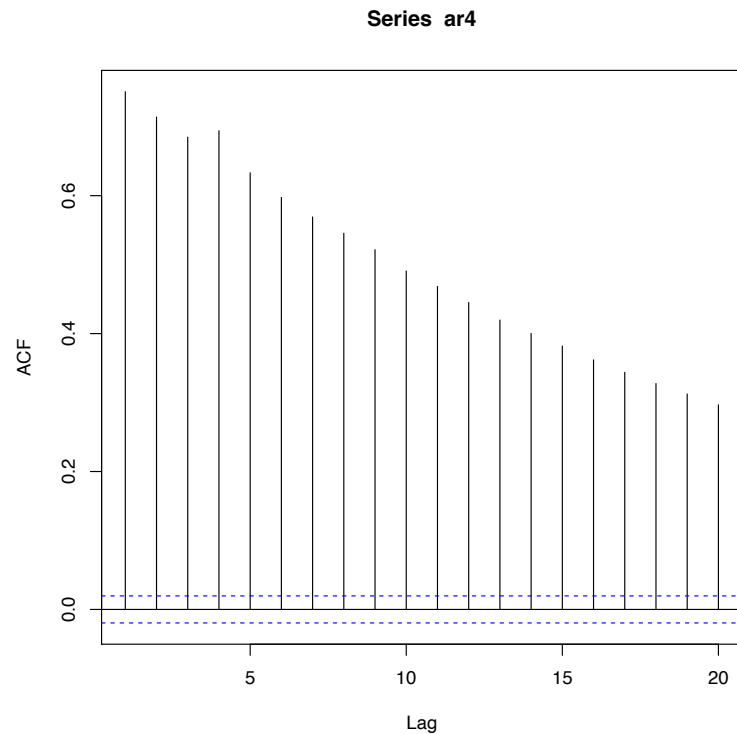
$$\Phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p$$

$$\Theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \dots + \theta_q L^q$$

- The ARMA(p, q) is invertible if all the roots of $\Theta(L)$ lie outside the unit circle.
- The ARMA(p, q) is stationary if all the roots of $\Phi(L)$ lie outside the unit circle.

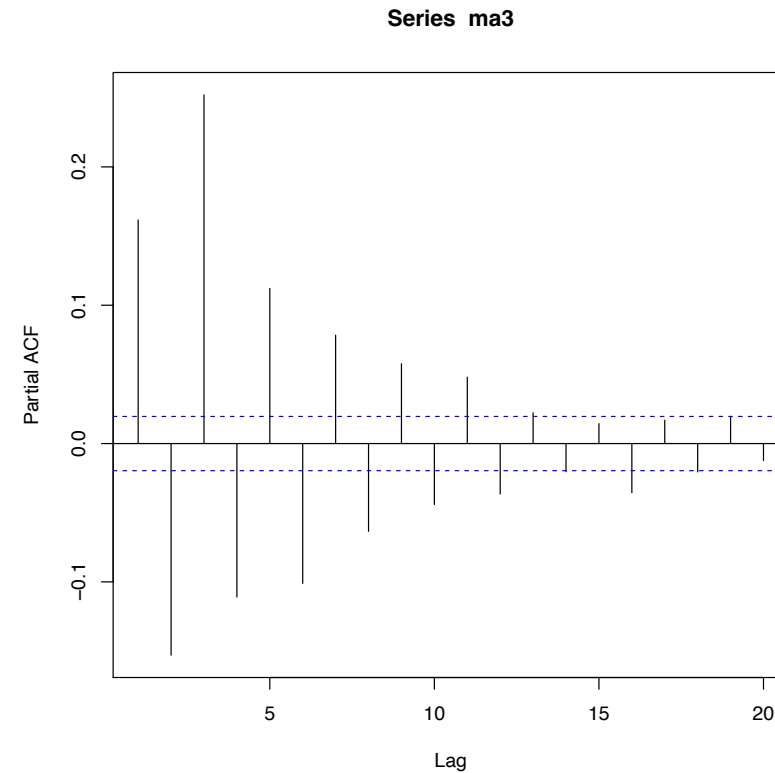
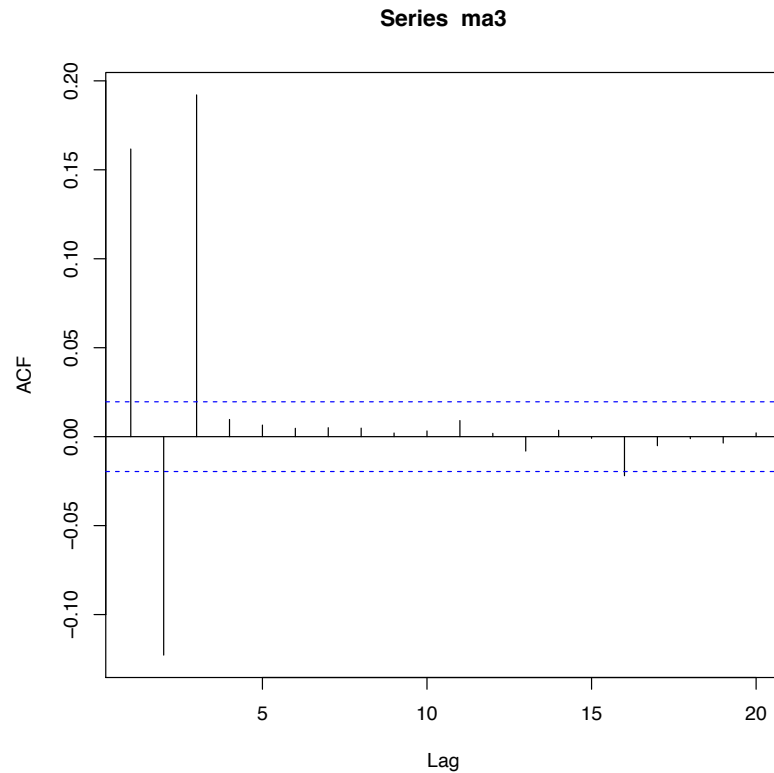
WHAT WE HAVE LEARNED SO FAR

- When thinking about how many AR and MA terms to include in our model, it is useful to think about the manner in which they manifest in the shapes of the autocorrelation and partial autocorrelation functions of the time series.
- A pure autoregressive model will have an autocorrelation function ρ_j that decays gradually to zero as $j \rightarrow \infty$ and a partial autocorrelation function that cuts off at the order of the autoregression (i.e. lag p).



WHAT WE HAVE LEARNED SO FAR

- On the other hand, the a pure moving average model will have an autocorrelation function ρ_j that cuts off at the order of the moving average (i.e. lag q) and a partial autocorrelation that gradually decays to zero as $j \rightarrow \infty$



WHAT WE HAVE LEARNED SO FAR

- The sample ACF and PACF can be very useful for identifying AR and MA orders in clear cut cases. However, in most instances, our time series model will include both AR and MA terms.
- Therefore our primary method of choosing our ARMA specification will be information criteria.
- Let k be the number of parameters to be estimated in the model, then,

$$AIC = 2k - 2\log L(\boldsymbol{\theta})$$

$$BIC = \log(n)k - 2\log L(\boldsymbol{\theta})$$

For a normal likelihood function, the values that maximize the log likelihood will be the same values that minimize the sum of squared errors. Therefore these values should be equivalent up to a scaling constant.

WHAT WE HAVE LEARNED SO FAR

$$Y_t = c + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

$$\varepsilon_t \sim iid N(0, \sigma^2)$$

- Generating the point and interval forecasts then involves the following steps:
 1. The m -step ahead point forecast is given by $E[Y_{T+m} | \Omega_T]$
 2. The m -step ahead forecast error is given by $Y_{T+m} - E[Y_{T+m} | \Omega_T]$
 3. From the structure of the ARMA model, we know that the forecast error will be a linear function of *i. i. d.* innovations ε_t from which we can compute the variance of the forecast error as a function of $\boldsymbol{\theta} = \{c, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q, \sigma^2\}$
 4. Then the 95% interval forecast for the m -step ahead forecast will be defined by

$$E[Y_{T+m} | \Omega_T] \pm 1.96\sigma_m$$

5. We will use the MLE estimates to compute all of these objects.

LIMITS OF ARMA MODELS

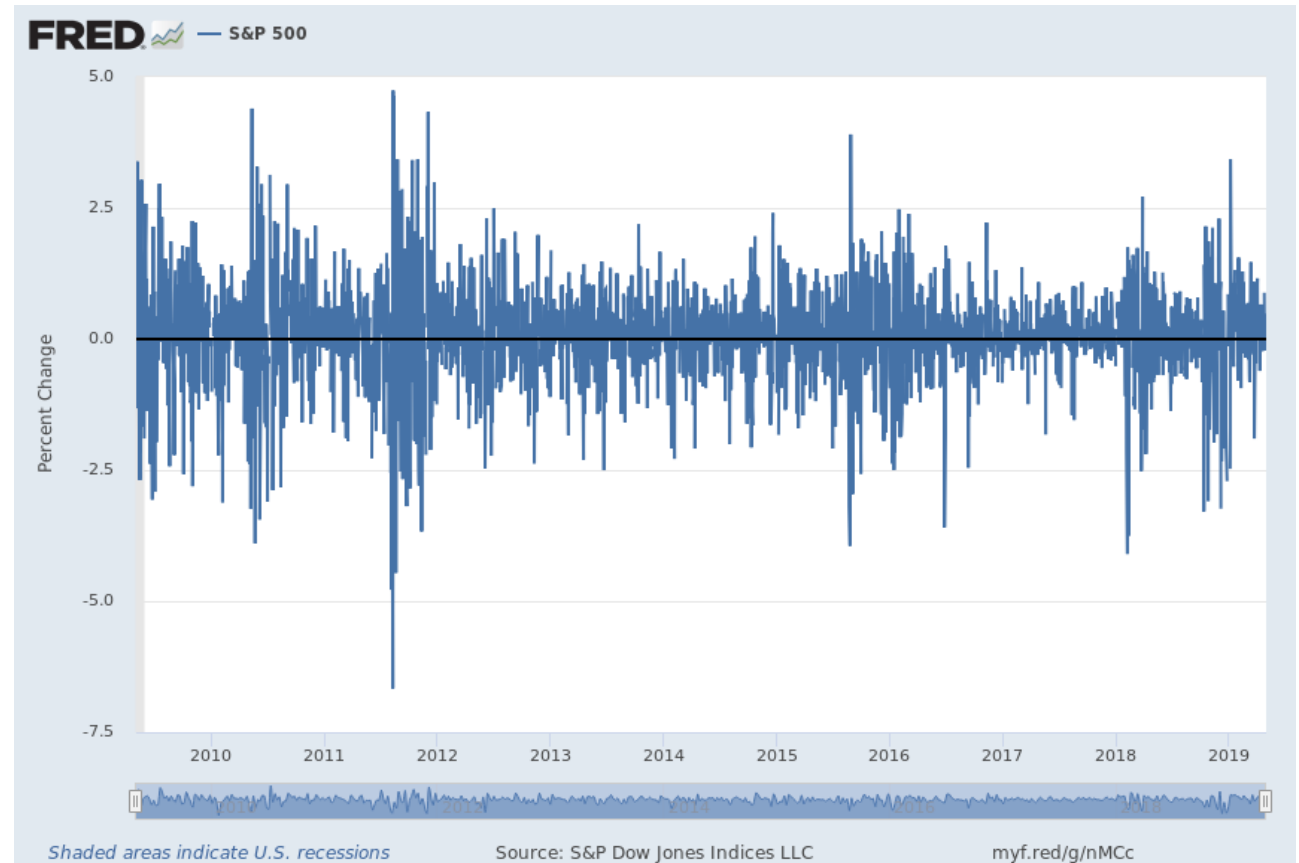
- Recall that ARMA processes have the following conditional and unconditional moments:

AR PROCESS	MA PROCESS
$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + \varepsilon_t$	$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q}$
$E[Y_t] = 0$	$E[Y_t] = 0$
$E[Y_t \Omega_{t-1}] = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p}$	$E[Y_t \Omega_{t-1}] = \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q}$
$var(Y_t) = \gamma_0 = \phi_1 \gamma_1 + \phi_2 \gamma_2 + \dots + \phi_p \gamma_p + \sigma^2$	$var(Y_t) = \gamma_0 = \sigma^2 + \theta_1^2 \sigma^2 + \theta_2^2 \sigma^2 + \dots + \theta_q^2 \sigma^2$
$var(Y_t \Omega_{t-1}) = E[\varepsilon_t^2 \Omega_{t-1}] = \sigma^2$	$var(Y_t \Omega_{t-1}) = E[\varepsilon_t^2 \Omega_{t-1}] = \sigma^2$

- When a process has a constant conditional variance, shocks have no impact on the volatility of a time series.

MOTIVATING VOLATILITY MODELS

- When we look at plots of many financial time series, we observe clear volatility clustering.
- History has shown us again and again that large negative shocks are followed by periods of high volatility.
- One way of characterizing this kind of time series behaviour is to say that there exists time-variation in the conditional variance of the series.



AUTOREGRESSIVE CONDITIONAL HETROSKEDEASTICITY (ARCH) MODEL

- Let Y_t be a stationary ARMA(p, q) model:

$$Y_t = \mu_t + \varepsilon_t$$
$$\mu_t = \alpha + \sum_{i=1}^p \phi_i Y_{t-i} + \sum_{i=1}^q \theta_i \varepsilon_{t-i}$$

- The conditional mean and variance of this model is given by

$$E[Y_t | \Omega_{t-1}] = \mu_t = \alpha + \sum_{i=1}^p \phi_i Y_{t-i} + \sum_{i=1}^q \theta_i \varepsilon_{t-i}$$

$$E[(Y_t - \mu_t)^2 | \Omega_{t-1}] = \sigma_t^2 = \text{var}(\varepsilon_t | \Omega_{t-1})$$

- In addition to specifying an equation for the conditional mean, we are now interested in specifying a model in which we allow the conditional variance σ_t^2 to evolve over time.

TESTING FOR ARCH EFFECTS

- Before we can proceed to specify and estimate an ARCH model, we need to test whether ARCH effects are present. Let e_t be the residuals of the mean equation,

$$e_t = y_t - \hat{\mu}_t$$

- The first thing that we can do is to compute Ljung-Box and Box-Pierce statistics with to the squared residuals $\{e_t^2\}_{t=1}^T$. We can use these computed statistics to test the null hypothesis that the first m lags of the ACF of the squared residuals are jointly zero.
- A second approach would be to estimate the following linear regression

$$\varepsilon_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_m \varepsilon_{t-m}^2 + u_t$$

- And then test the null hypothesis

$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_m = 0$$

- This is known as the Lagrange multiplier test and was devised by Engel (1982). It is simply a test of overall significance (i.e. an F-test) of an AR(m) model.

TESTING FOR ARCH EFFECTS

- To illustrate, let's work with the returns on the S&P 500. Let the level of the index be given by P_t and let the percentage returns be given by

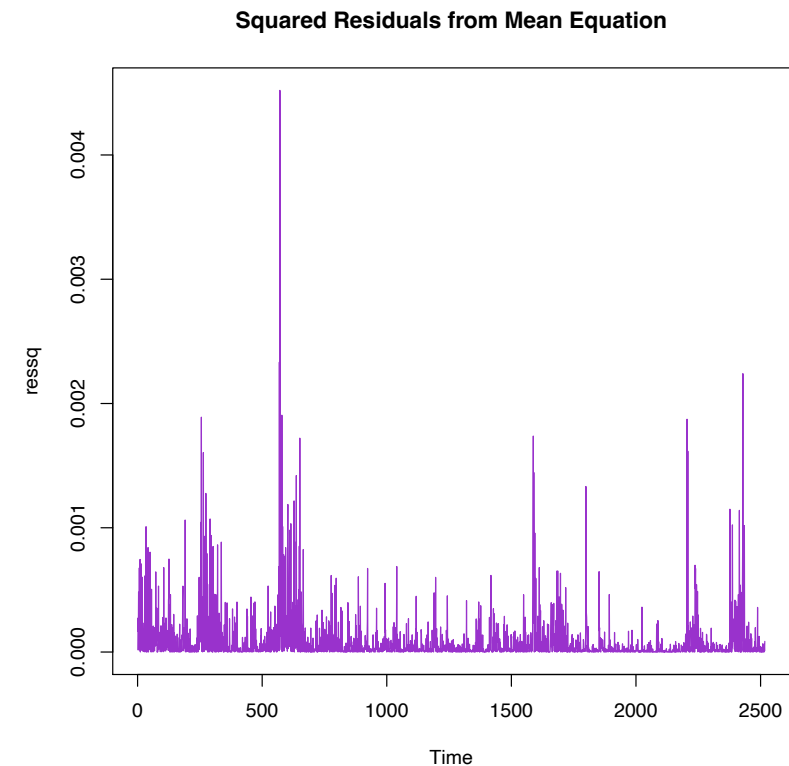
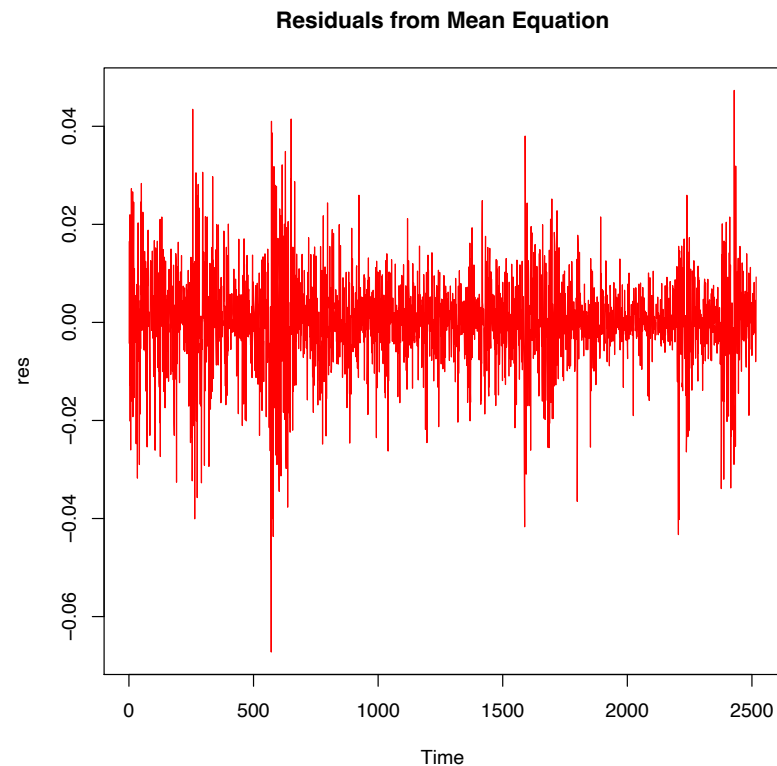
$$R_t = \frac{P_t - P_{t-1}}{P_t} \approx \log(P_t) - \log(P_{t-1})$$

- For a detailed explanation of using a log difference to approximate a percentage change, see (<https://stats.stackexchange.com/questions/244199/why-is-it-that-natural-log-changes-are-percentage-changes-what-is-about-logs-th/244237>)
- For the mean equation, let's specify an ARMA(1,1)

$$R_t = c + \phi R_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}$$

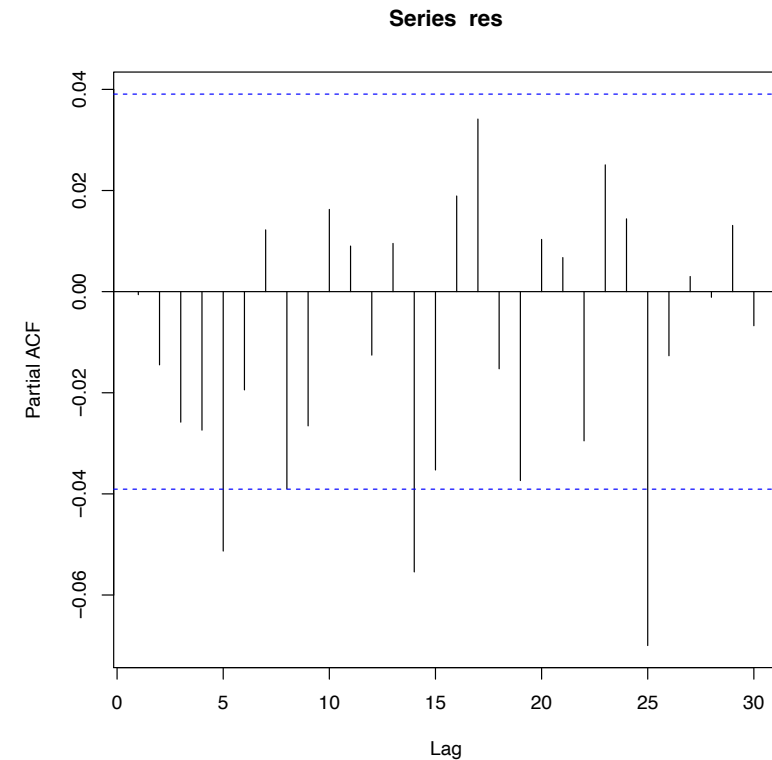
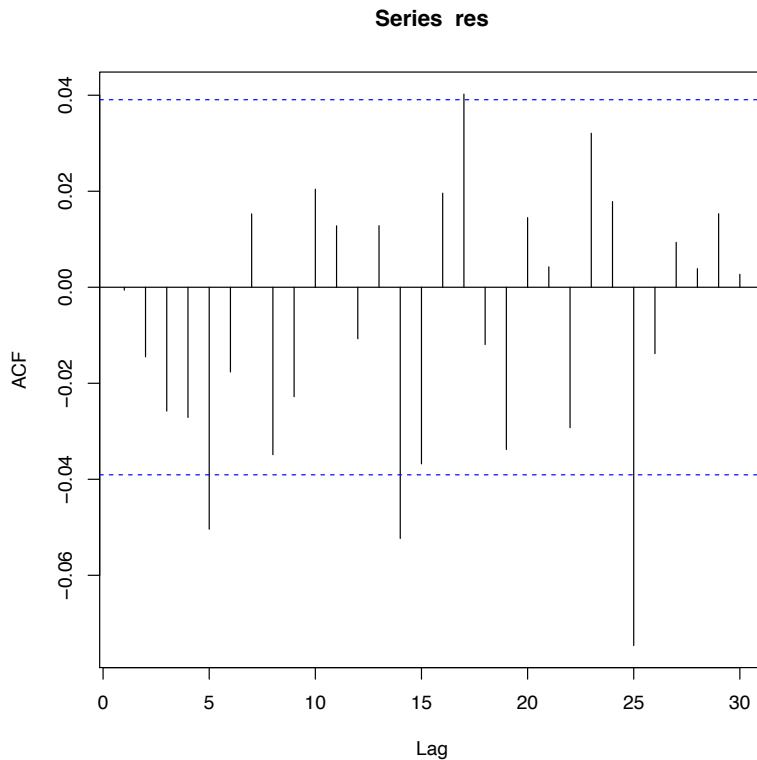
TESTING FOR ARCH EFFECTS

- The residuals and squared residuals obtained from the mean equation estimation are as follows:



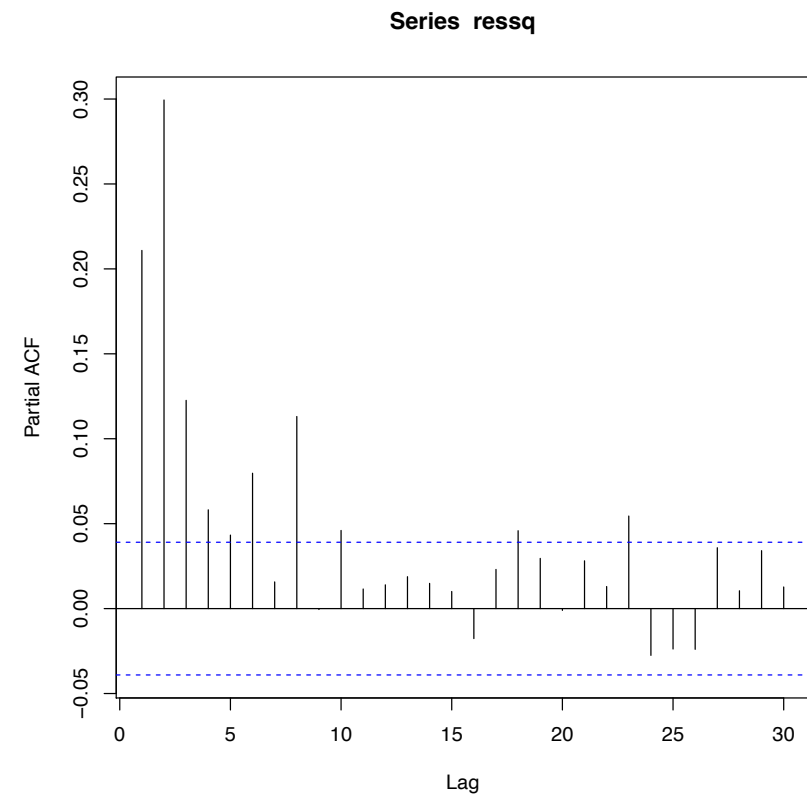
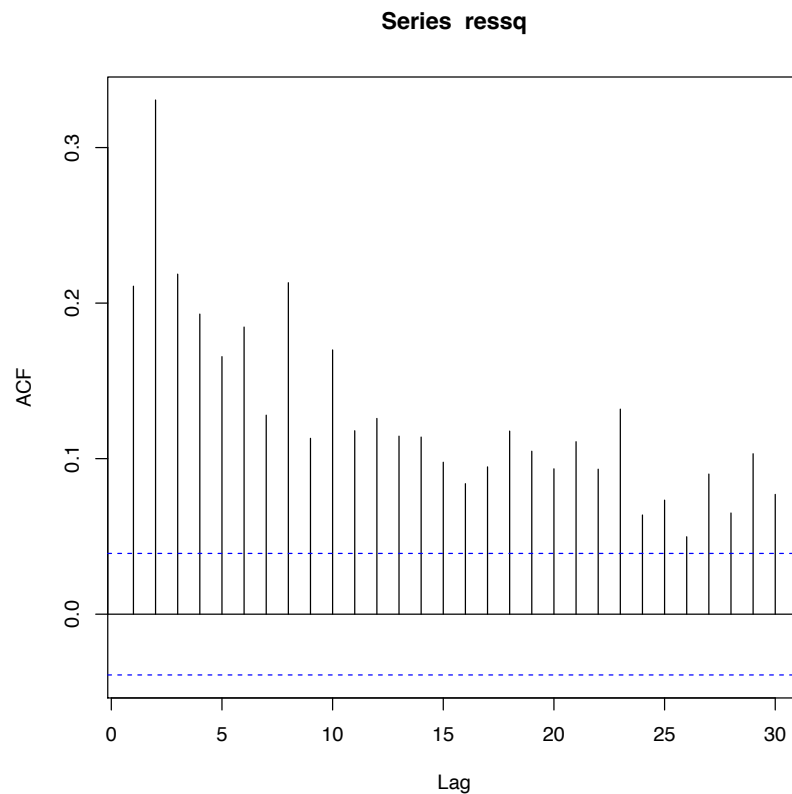
TESTING FOR ARCH EFFECTS

- Looking at the sample ACF and PACF of the residuals, we don't see too much structure,



TESTING FOR ARCH EFFECTS

- We have a much different story when we look at the sample ACF and PACF of the squared residuals:



TESTING FOR ARCH EFFECTS

- The dependence in the squared residuals would also be confirmed by Box tests and the ARCH LM test.

```
> Box.test(ressq, lag = m, type = "Box-Pierce")
```

Box-Pierce test

```
data: ressq  
X-squared = 1790.7, df = 51, p-value < 2.2e-16
```

```
> Box.test(ressq, lag = m, type = "Ljung-Box")
```

Box-Ljung test

```
data: ressq  
X-squared = 1802, df = 51, p-value < 2.2e-16
```

Call:

```
lm(formula = ressq[6:T_r] ~ ressq[5:(T_r - 1)] + ressq[4:(T_r -  
2)] + ressq[3:(T_r - 3)] + ressq[2:(T_r - 4)] + ressq[1:(T_r -  
5)])
```

Residuals:

Min	1Q	Median	3Q	Max
-0.0009014	-0.0000623	-0.0000415	0.0000009	0.0038343

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.966e-05	4.864e-06	8.155	5.46e-16 ***
ressq[5:(T_r - 1)]	1.003e-01	1.995e-02	5.029	5.29e-07 ***
ressq[4:(T_r - 2)]	2.593e-01	2.001e-02	12.956	< 2e-16 ***
ressq[3:(T_r - 3)]	1.048e-01	2.057e-02	5.097	3.71e-07 ***
ressq[2:(T_r - 4)]	5.419e-02	2.001e-02	2.708	0.00682 **
ressq[1:(T_r - 5)]	4.349e-02	1.994e-02	2.181	0.02924 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0002037 on 2506 degrees of freedom
Multiple R-squared: 0.1475, Adjusted R-squared: 0.1458
F-statistic: 86.71 on 5 and 2506 DF, p-value: < 2.2e-16

AUTOREGRESSIVE CONDITIONAL HETROSKEDEASTICITY (ARCH) MODEL

- Having detected ARCH effects, we are now motivated to specify and estimate a model!
- The first model that we will look at is the standard ARCH(m), which is written as

$$Y_t = \mu_t + \varepsilon_t$$

$$\varepsilon_t = \sigma_t v_t$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \cdots + \alpha_m \varepsilon_{t-m}^2$$

$$v_t \sim i.i.d.(0,1), \quad \alpha_i \geq 0 \quad \forall i$$

- Therefore, according to this model, the dependence structure of the error ε_t is driven by the time variation in its conditional variance σ_t^2 .
- Specifically, large, shocks in the past will imply a large conditional variance σ_t for the innovation ε_t .

ARCH(1)

- To better understand the ARCH model, let's derive its stochastic properties. Let's consider an ARCH(1) process,


$$\varepsilon_t = \sigma_t v_t$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2$$

- Given the information set Ω_{t-1} , we can easily derive the conditional mean of ε_t as well as its unconditional mean

$$E[\varepsilon_t | \Omega_{t-1}] = E[\sigma_t v_t | \Omega_{t-1}] = \sigma_t E[v_t | \Omega_{t-1}] = 0$$

Law of iterated
expectations


$$E[\varepsilon_t] = E[E[\varepsilon_t | \Omega_{t-1}]] = 0$$

ARCH(1)

- The conditional variance of ε_t is computed as

$$E[\varepsilon_t^2 | \Omega_{t-1}] = E[\sigma_t^2 v_t^2 | \Omega_{t-1}] = \sigma_t^2 E[v_t^2 | \Omega_{t-1}] = \sigma_t^2$$

- While the unconditional variance of ε_t is given by

$$E[\varepsilon_t^2] = E[E[\sigma_t^2 v_t^2 | \Omega_{t-1}]] = E[\sigma_t^2] = E[\alpha_0 + \alpha_1 \varepsilon_{t-1}^2] = \alpha_0 + \alpha_1 E[\varepsilon_{t-1}^2]$$

- Since we are assuming that ε_t is a covariance stationary process, it must be the case that

$$E[\varepsilon_t^2] = E[\varepsilon_{t-1}^2]$$

- Therefore we can solve for the unconditional variance as

$$E[\varepsilon_t^2] = \frac{\alpha_0}{1 - \alpha_1}$$

Therefore in order for the unconditional variance to be positive, it must be the case that $0 \leq \alpha_1 < 1$

ARCH(1)

- An ARCH model also allows for some interesting behaviour in the tails of the density function of ε_t . To understand this, we must go beyond mean and variance of ε_t and consider its fourth moment.
- In general, the k -th raw moment of a random variable X is given by

$$E[X^k]$$

- If the random variable X has a mean μ_X , its k -th central moment is given by

$$E[(X - \mu)^k]$$

- If the random variable X has a mean μ_X and variance of σ_X^2 , its k -th standardized moment is given by

$$E \left[\left(\frac{X - \mu_X}{\sigma_X} \right)^k \right]$$

ARCH(1)

- It should be clear then that the variance of a random variable is its second central moment

$$\sigma_X^2 = E[(X - \mu_X)^2]$$

- As it turns out, the skewness of a random variable is given by its third standardized moment

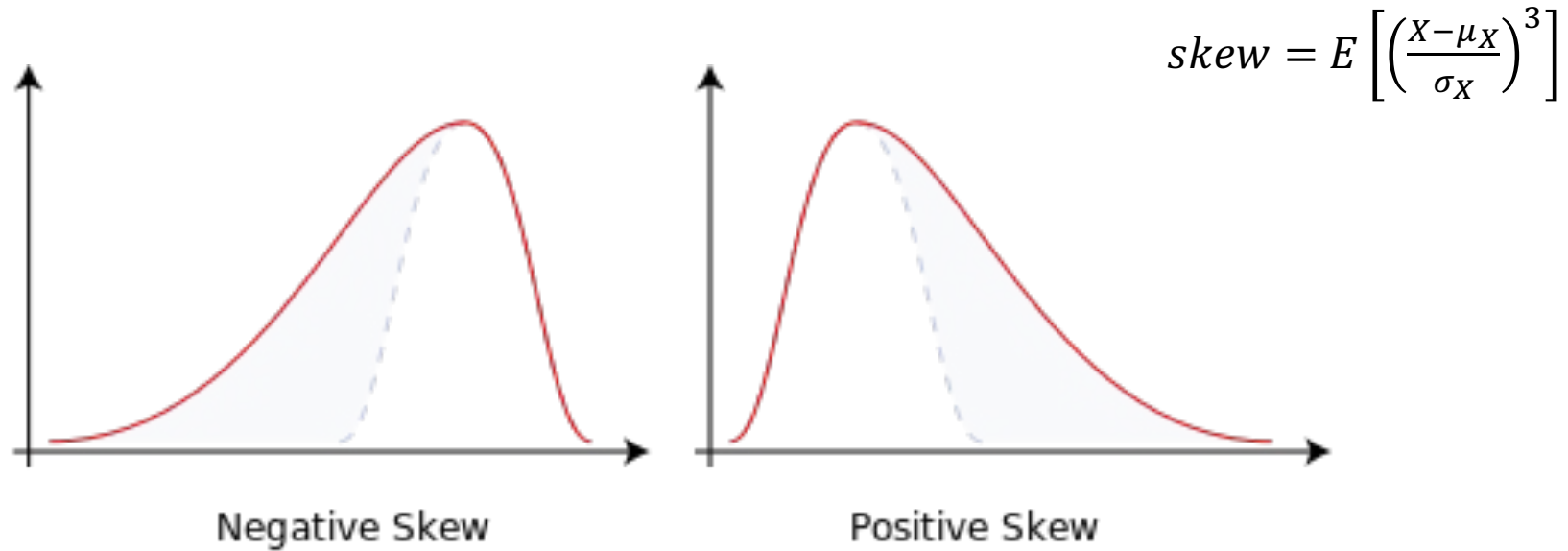
$$skew = E \left[\left(\frac{X - \mu_X}{\sigma_X} \right)^3 \right]$$

- And the kurtosis of a random variable is given by its fourth standardized moment

$$kurt = E \left[\left(\frac{X - \mu_X}{\sigma_X} \right)^4 \right] = \frac{E[(X - \mu_X)^4]}{(\sigma_X^2)^2}$$

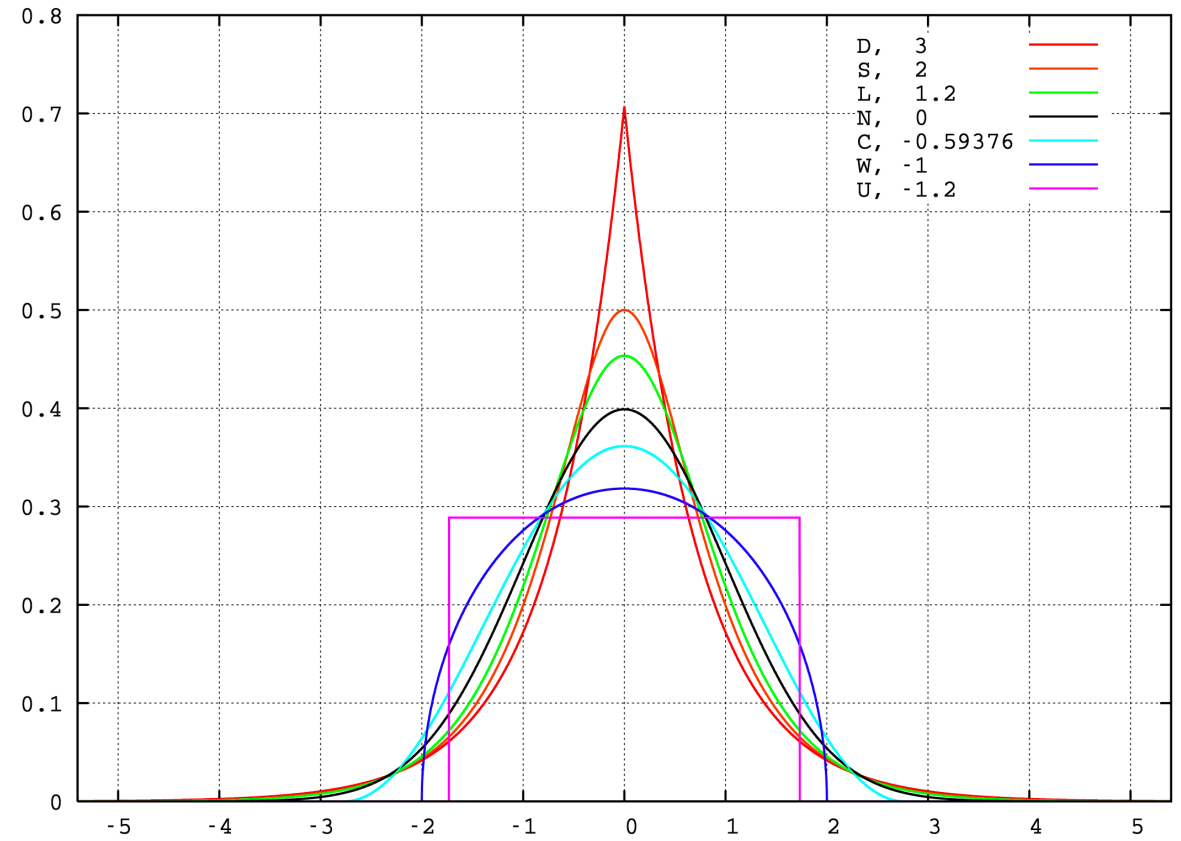
ARCH(1)

- The skewness of a random variable describes the presence of asymmetry in the tails of its probability density (or distribution) function.



ARCH(1)

- The kurtosis of a random variable describes the relative concentration of probability mass in the tails of its density (or distribution) function.
- A density function that exhibits a high degree of kurtosis (also known as leptokurtic) its mass concentrated around its mean and in its tails. Such densities are described as highly peaked and fat tailed.
- A density function that exhibits a low degree of kurtosis (also known as platykurtic) has its probability mass distributed evenly across a moderate range of values.



ARCH(1)

- The fourth central conditional moment of our ARCH(1) process is given by

$$E[\varepsilon_t^4 | \Omega_{t-1}] = E[\sigma_t^4 v_t^4 | \Omega_{t-1}] = \sigma_t^4 E[v_t^4 | \Omega_{t-1}]$$

- Let's make the assumption that v_t is a standard normal, that is, $v_t \sim i.i.d N(0,1)$. A feature of all normal random variables is that they have a kurtosis of 3. Therefore,

$$E[\varepsilon_t^4 | \Omega_{t-1}] = 3\sigma_t^4 = 3(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2)^2$$

- Again, we compute the fourth central unconditional moment using the Law of Iterated Expectations,

$$E[\varepsilon_t^4] = E[E[\varepsilon_t^4 | \Omega_{t-1}]] = 3E[(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2)^2] = 3E[\alpha_0^2 + 2\alpha_0\alpha_1 \varepsilon_{t-1}^2 + \alpha_1^2 \varepsilon_{t-1}^4]$$

ARCH(1)

- Passing the expectations operator through, we obtain,

$$E[\varepsilon_t^4] = 3(\alpha_0^2 + 2\alpha_0\alpha_1E[\varepsilon_{t-1}^2] + \alpha_1^2E[\varepsilon_{t-1}^4])$$

- Since we have already solved for $E[\varepsilon_{t-1}^2]$, we can write

$$E[\varepsilon_t^4] = 3\left(\alpha_0^2 + 2\alpha_0\alpha_1\left(\frac{\alpha_0}{1-\alpha_1}\right) + \alpha_1^2E[\varepsilon_{t-1}^4]\right)$$

- Collecting terms, we obtain

$$E[\varepsilon_t^4] = 3\alpha_0^2\left(1 + 2\left(\frac{\alpha_1}{1-\alpha_1}\right)\right) + 3\alpha_1^2E[\varepsilon_{t-1}^4]$$

ARCH(1)

- If we make the assumption that our process is fourth order stationary, then $E[\varepsilon_t^4] = E[\varepsilon_{t-1}^4]$, so that

$$E[\varepsilon_t^4](1 - 3\alpha_1^2) = 3\alpha_0^2 \left(1 + 2 \left(\frac{\alpha_1}{1 - \alpha_1} \right) \right) = 3\alpha_0^2 + \frac{6\alpha_0^2\alpha_1}{1 - \alpha_1} = \frac{3\alpha_0^2 + 3\alpha_0^2\alpha_1}{1 - \alpha_1}$$

- Simplification yields,

$$E[\varepsilon_t^4] = \frac{3\alpha_0^2(1 + \alpha_1)}{(1 - 3\alpha_1^2)(1 - \alpha_1)}$$

- Therefore in order for the fourth moment of our process to be positive, not only must $0 \leq \alpha_1 < 1$, it must also be the case that $(1 - 3\alpha_1^2) > 0$, that is,

$$0 < \alpha_1^2 < \frac{1}{3}$$

ARCH(1)

- Having computed the unconditional fourth centered moment of the process, we can proceed to compute its unconditional kurtosis,

$$kurt_{\varepsilon} = \frac{E[\varepsilon_t^4]}{(E[\varepsilon_t^2])^2} = \frac{3\alpha_0^2(1 + \alpha_1)}{(1 - 3\alpha_1^2)(1 - \alpha_1)} \times \frac{(1 - \alpha_1)^2}{\alpha_0^2} = \frac{3(1 - \alpha_1^2)}{1 - 3\alpha_1^2} > 3$$

- Where the last inequality is obtained by imposing the restriction that $0 \leq \alpha_1^2 < \frac{1}{3}$
- This result tells us that the tail distribution of ε_t is heavier than that of a normal distribution. That is, the shock of an ARCH(1) process is more likely than a Gaussian white noise series to produce extreme observations.

ESTIMATING ARCH MODELS

- As with ARMA models, the method of maximum likelihood is typically used to estimate ARCH models. Suppose we had an ARCH(m) in which the white noise innovations v_t are assumed to be standard Gaussian,

$$\varepsilon_t = \sigma_t v_t$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \cdots + \alpha_m \varepsilon_{t-m}^2$$

$$v_t \sim i.i.d. N(0,1)$$

- Then the likelihood function associated with this model for a sample size of T would be given by

$$L(\boldsymbol{\theta}; \varepsilon_1, \varepsilon_2, \dots, \varepsilon_T) = f(\varepsilon_1, \dots, \varepsilon_m; \boldsymbol{\theta}) \times \prod_{t=m+1}^T f(\varepsilon_t | \varepsilon_{t-1}; \boldsymbol{\theta}) = f(\varepsilon_1, \dots, \varepsilon_m; \boldsymbol{\theta}) \times \prod_{t=m+1}^T \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left\{-\frac{\varepsilon_t^2}{2\sigma_t^2}\right\}$$

ESTIMATING ARCH MODELS

- As with the case in ARMA models, it is much easier to work with the conditional likelihood in which we condition on the first m values,

$$L(\boldsymbol{\theta}; \varepsilon_{m+1}, \dots, \varepsilon_T | \varepsilon_1, \dots, \varepsilon_m) = \prod_{t=m+1}^T \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left\{-\frac{\varepsilon_t^2}{2\sigma_t^2}\right\}$$

- Note that once we condition on the first m values, σ_t^2 can be evaluated recursively.
- The MLE estimates will be values of $\hat{\alpha}_0, \hat{\alpha}_1, \dots, \hat{\alpha}_m$ that maximize the log likelihood.

ESTIMATING ARCH MODELS

- We can estimate an ARCH model in R using the `garch()` function. From our sample PACF of the squared residuals we see dependence up until the 10th lag, so let's estimate an ARCH(10).
- The residuals generated from a GARCH object are standardized residuals:

$$u_t = \frac{e_t}{\hat{\sigma}_t}$$

- For a properly specified ARCH model, the standardized residuals should be observations from a sequence of *i. i. d*(0,1) variables.

```
Call:
garch(x = res, order = c(0, 10))
```

```
Model:
GARCH(0,10)
```

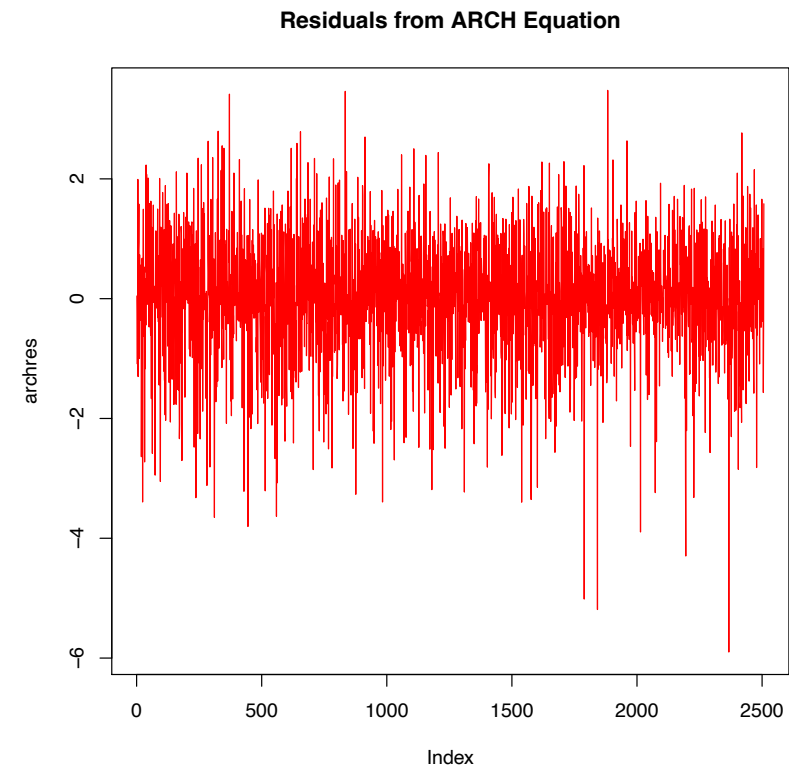
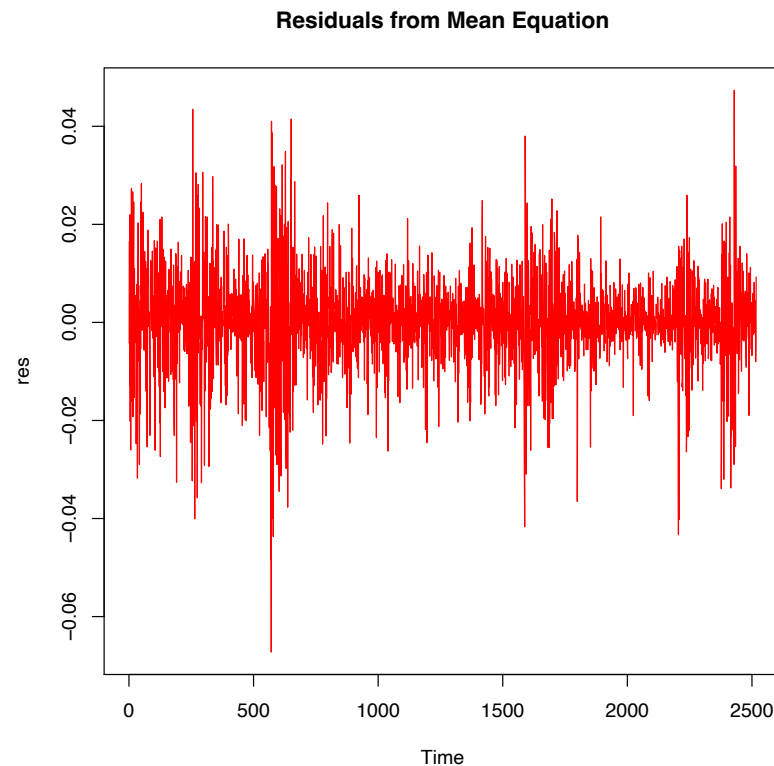
```
Residuals:
      Min       1Q   Median       3Q      Max
-5.89923 -0.50568  0.01735  0.58414  3.47756
```

```
Coefficient(s):
      Estimate Std. Error t value Pr(>|t|)
a0  1.868e-05  1.180e-06  15.826 < 2e-16 ***
a1  1.085e-01  1.484e-02   7.309 2.69e-13 ***
a2  1.532e-01  2.162e-02   7.086 1.38e-12 ***
a3  1.133e-01  2.150e-02   5.271 1.36e-07 ***
a4  9.846e-02  2.038e-02   4.832 1.35e-06 ***
a5  5.701e-02  1.553e-02   3.672 0.000241 ***
a6  6.420e-02  1.922e-02   3.339 0.000839 ***
a7  5.698e-02  1.306e-02   4.362 1.29e-05 ***
a8  4.581e-02  1.624e-02   2.820 0.004803 **
a9  4.274e-02  1.753e-02   2.437 0.014795 *
a10 7.141e-02  1.792e-02   3.985 6.74e-05 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

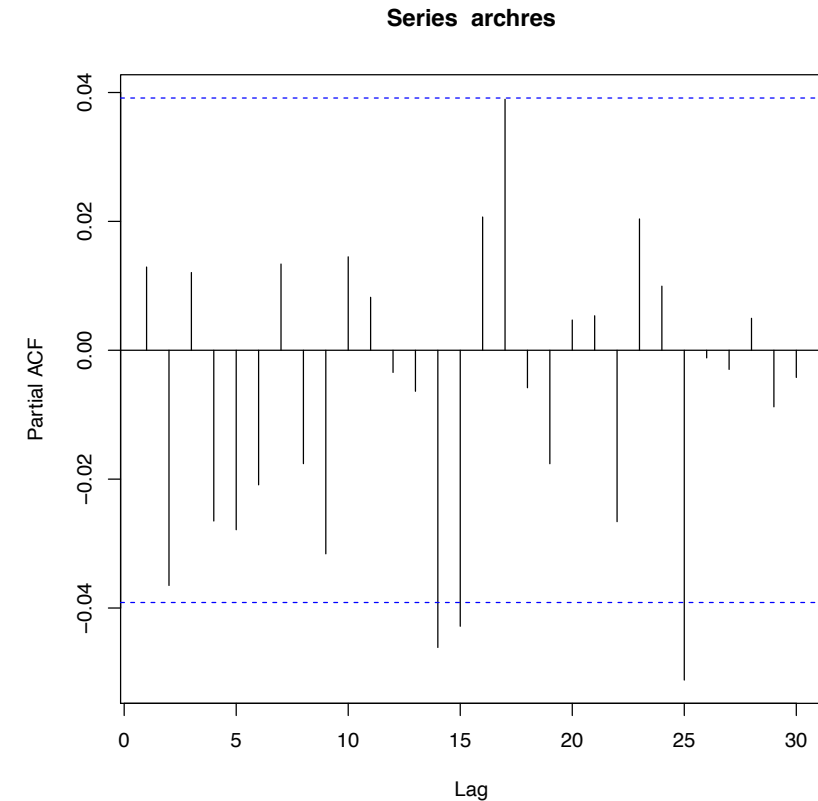
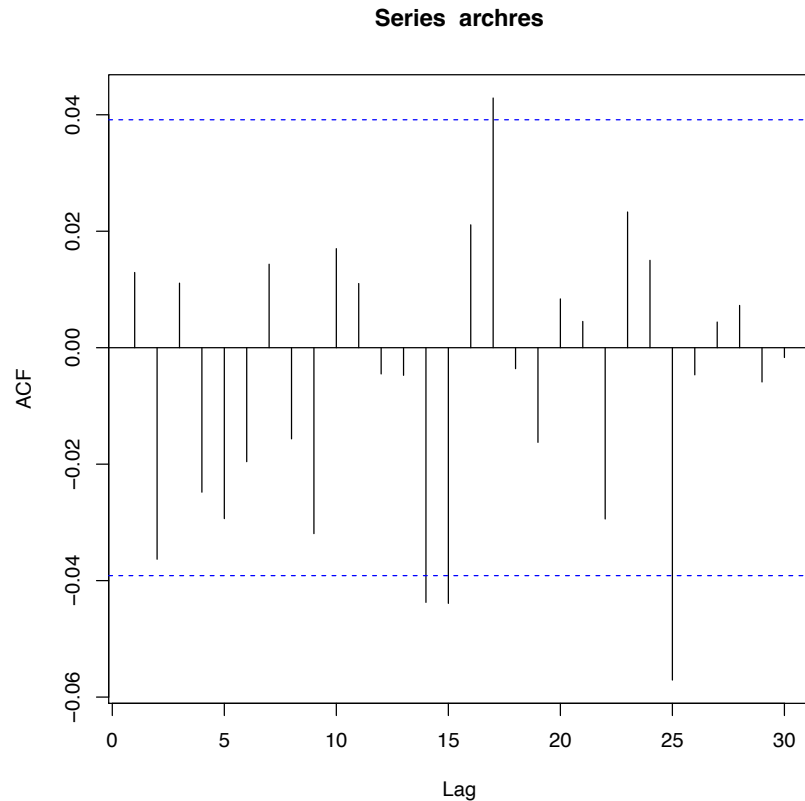
ESTIMATING ARCH MODELS

- Comparing the residuals from the ARCH estimation and the mean equation:



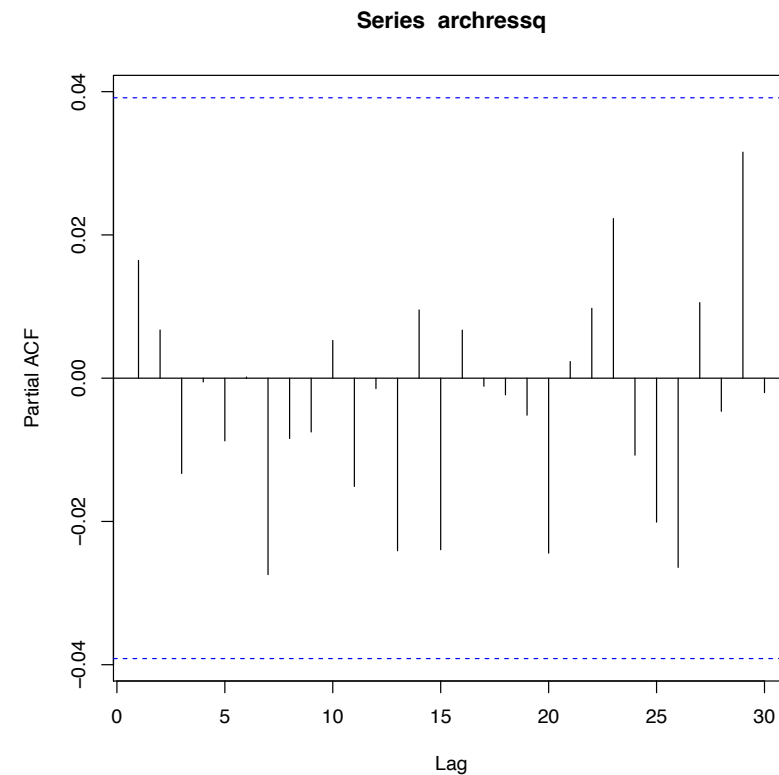
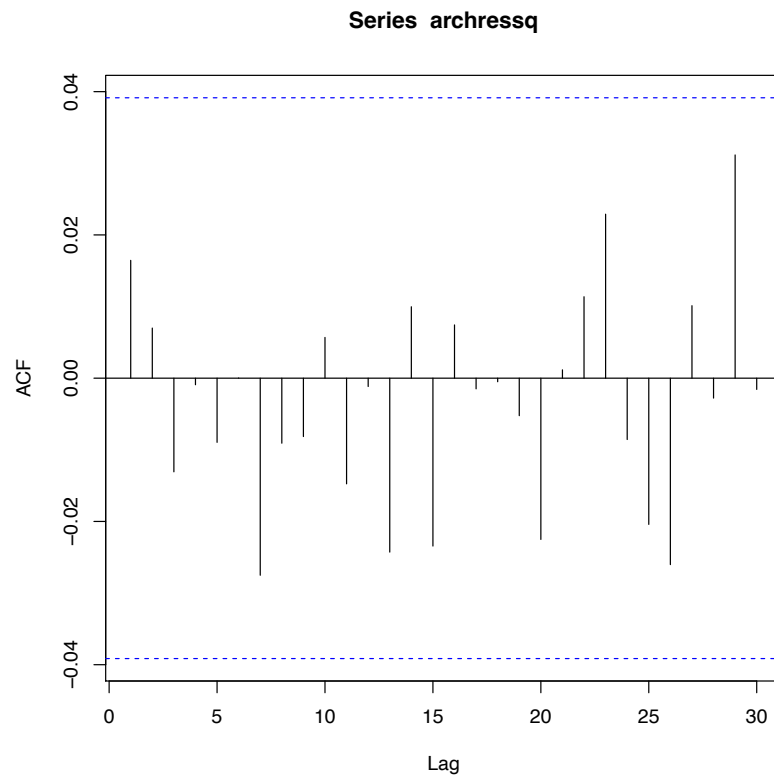
ESTIMATING ARCH MODELS

- Looking at the sample ACF and PACF of the residuals from the ARCH estimation:



ESTIMATING ARCH MODELS

- Looking at the sample ACF and PACF of the squared residuals from the ARCH estimation



ESTIMATING ARCH MODELS

- Performing our Box tests on the squared residuals from our ARCH estimation we can clearly see that the dependence in the conditional variance has been removed!

```
> Box.test(archressq, lag = m, type = "Box-Pierce")
```

Box-Pierce test

```
data: archressq  
X-squared = 28.524, df = 51, p-value = 0.9954
```

```
> Box.test(archressq, lag = m, type = "Ljung-Box")
```

Box-Ljung test

```
data: archressq  
X-squared = 28.865, df = 51, p-value = 0.9947
```

LIMITATIONS OF ARCH MODELS

- While the ARCH model is able to accommodate some important features of financial time series, it nevertheless has the following shortcomings:

1. The model assumes that positive and negative shocks have the same effects on volatility. We can see this directly in the conditional variance equation:

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \cdots + \alpha_m \varepsilon_{t-m}^2$$

2. The ARCH parameters are restricted to lie in a rather small range. For instance in an ARCH(1), it must be the case that, $0 \leq \alpha_1^2 < \frac{1}{3}$. This means that we may have to include a lot of terms in order to account for dependence.