Econometric Analysis ECMT2160 Exam notes

Charles Christopher Hyland

Semester 2 2017

Abstract

Welcome! Hopefully these notes help you to ace ECMT2160!

Contents

1 Time Series Data **1**
- 1.1 Introduction . . . 1
- 1.2 Gauss-Markov Assumptions . . . 1
  - 1.2.1 Assumption 1: Linearity . . . 1
  - 1.2.2 Assumption 2: No Perfect Collinearity . . . 1
  - 1.2.3 Assumption 3: Strict Exogeneity . . . 2
  - 1.2.4 Assumption 4: Homoskedasticity . . . 2
  - 1.2.5 Assumption 5: No autocorrelation . . . 3
  - 1.2.6 Assumption 6: Normality . . . 4
- 1.3 Classical Linear Model Assumptions . . . 5

2 Static and Distributed Lag Models **6**
- 2.1 Static Models . . . 6
- 2.2 Testing for Serial Correlation . . . 7
  - 2.2.1 Breusche-Godfrey Test for serial correlation . . . 9
- 2.3 Effects of serial correlation . . . 10
- 2.4 Correcting Serial Correlation . . . 11
  - 2.4.1 Cochrane-Orcutt Correcting Serial Correlation (Feasible GLS) . . . 12
- 2.5 Finite Distributed Lag (FDL) Models . . . 13
- 2.6 Infinite Distributed Lag (IDL) Models . . . 16
- 2.7 Geometric Distributed Lag (GDL)/Koyck distributed Lag Models . . . 17
  - 2.7.1 Application of Koyck . . . 18
- 2.8 Conclusion . . . 18

3 Deterministic and Stochastic Trends **19**
- 3.1 Trending Data . . . 19
- 3.2 Stochastic Trends . . . 22

4 Autoregressive Process **27**

5 Vector Autoregression and Error Correction **32**
- 5.1 VAR . . . 33
- 5.2 Cointegration . . . 34
  - 5.2.1 Testing Cointegration . . . 35
  - 5.2.2 Error Correction Model . . . 35
  - 5.2.3 Vector Error Correction Model . . . 35

6 Forecasting **37**
- 6.1 Point Forecast . . . 37

1 Time Series Data

1.1 Introduction

Time series are data collected at fixed intervals and stored chronologically. Therefore, unlike cross-sectional data, it's very important the way in which the data is ordered. From this, we can think of time series data as a **realisation of random variables indexed by time**.

$$\{Y_t \quad \text{t} = 1,..,\text{T}\} = \{Y_1, .., Y_T\}$$

We can refer to this sequence as a **stochastic process**. Intuitively, we can ever only observe the one realisation of a random variable for time series data since we only live once #YOLO.

Time series data can incorporate trends over years whilst time series at monthly/quarterly frequencies can contain seasonality in the data.

1.2 Gauss-Markov Assumptions

OLS on time series can be unbiased but we need quite strict assumptions for this to work. We can use alot of assumptions that were similar for cross sectional OLS.

1.2.1 Assumption 1: Linearity

For a time series process: $\{y_t : t = 1, .., T\}$

$$y_t = \beta' x_t + \epsilon_t$$

where $x_t = (x_{t1},x_{t2},...,x_{tk})$' and β is the associated set of parameters. We assume that y_t is a linear combination of β terms and an error term.

1.2.2 Assumption 2: No Perfect Collinearity

Each $x_t j$ varies somewhat over time and no explanatory variable is an **exact linear function of the others**. This rules out **perfect correlation** between predictors. If we had multicollinearity though, it does not violate the assumption but does cause the variance of the estimators to be high and affect standard inferences.

1.2.3 Assumption 3: Strict Exogeneity

The conditional expectation of the error term is zero: $\mathrm{E}(\epsilon_t|X) = 0 \; \forall$ t

Where X = (x_1,...,x_T)' and is a matrix of explanatory variables. Anything that causes the results in time t to be correlated with any explanatory variables at any time period is a violation of strict exogeneity assumption. We have the case that there is no correlation or relationship between ϵ_t and the explanatory variables for all of time.

We can also have a less restrictive form of the assumption of **weak exogeneity**.

$E(\epsilon_t|x_t, x_{t-1}, x_{t-2}, ...) = 0$ This means that conditional expectation of error term is not zero for all time periods. This is the case of there is no relationship from time t and prior to that. We can extend this to **contemporaneous exogeneity** of $E(\epsilon_t|x_t) = 0$. This imples a lack of correlation between the explanatory variables and the error term for within that time period only. This contemporaneous exogeneity is sufficient for consistency but not for unbiasedness of OLS.

Under the assumptions of *linearity, no perfect collinearity,* and *strict exogeneity*, the OLS estimator is unbiased.

$$E(\hat{\beta}) = \beta \tag{1}$$

We get unbiasedness without restricting the correlation across time in the explanatory vavriables. Therefore, the x_t terms are allowed to be correlated with each other across times. Furthermore, the error terms ϵ_t are also allowed to be correlated across time. We don't know how precise these estimators are though and don't know anything regarding certainty. Therefore, we need other assumptions to be satisifed in order for us to test hypothesis and construct confidence intervals so that we can derive the estimator's distribution.

With **contemporaneous exogeneity**, this is sufficient for OLS estimators to be **consistent**.

1.2.4 Assumption 4: Homoskedasticity

The conditional variance of the error term is 0.

$$Var(\epsilon_t|X) = \sigma^2 \quad t = 1, ..., T \tag{2}$$

Here, we are saying that error terms have constant variance for all of time (strict form).

1.2.5 Assumption 5: No autocorrelation

The error terms are uncorrelated over time now.

$$Cov(\epsilon_t, \epsilon_s|X) = 0 \quad \forall t \neq s \tag{3}$$

We are saying here that there is no covariance between error terms across all time periods.

With assumption 4 and 5, we can now derive the **spherical disturbances** assumption whereby

$$E(\epsilon\epsilon'|X) = \sigma^2 I \tag{4}$$

where I is the T-dimensional identity matrix. Its size is the number of time periods we are looking at. Since it is an identity matrix, that means we don't have any covariance between error terms across time but we do have variance terms for the error terms.

Note that we have 3 different kinds of correlations that may differ across time series regression and need to consider.

1) Correlation between x_{tj} and x_{sj}. This is the correlation between a predictor and itself across time. This is not really an issue, unless there is a perfect linear relationship.

2) Correlation between x_{tj} and ϵ_s. This violates the strict exogeneity assumption and this leads to OLS being biased as a result. However, it doesn't affect weak exogeneity assumption if the error term is in a different period.

3) Correlation between ϵ_t and ϵ_s. This violates no autocorrelation assumption and only affects the **efficiency** of the OLS estimator.

We can have strict exogeneity satisfied but no autocorrelation is violated when ONLY the third (and even first) assumption is not satisfied. We can have a predictor uncorrelated with error terms but we may have error terms correlated with each other.

We can measure the level of **autocorrelation** between 2 time periods by:

$$\rho = \frac{\sigma_{x_t x_s}}{\sigma_{x_t}\sigma_{x_s}}$$

$$\text{t} \neq \text{s}$$

We look at the covariance between time t and s, divided by the standard deviation of t and s respectively. From this, letting s come before t in time, we can assume that s = t -k, where k = 1,2,... We can rewrite the equation to get:

$$\rho_k = \frac{\sigma_{x_t x_{t-k}}}{\sigma_{x_t}\sigma_{x_{t-k}}}$$

$$\text{t} \neq \text{s}$$

We do this because instead of comparing 2 distinct time periods t and s, we can rexpress s as the difference between t and s. So if t =5 and s =2, we can express s as 5-3 (t-k) where k =3. So now we are looking at time period t, and the k^{th} difference.

However, a thing to note is that σ_{x_t} and $\sigma_{x_{t-k}}$ is that they both come from the same stochastic process. From the assumption of **stationarity** whereby the variance is constant throughout the time series, we have

$$\sigma_{x_t} = \sigma_{x_{t-k}} \equiv \sigma_0$$

$$\forall \text{ t}$$

$$\sigma_{x_t x_{t-k}} = \sigma_{x_s x_{s-k}} \equiv \sigma_k$$

$$\forall \text{ t,s}$$

Here, the variance is the same for all time and that the covariance is also the same for all time periods too.

We can sub in σ_k for the covariance and σ_0 for the variance in our autocorrelation formula of $\rho = \frac{\sigma_{x_t x_{t-k}}}{\sigma_{x_t}\sigma_{x_{t-k}}}$.

From this, the autocorrelation of a stationary time series is given from:

$$\rho_k = \frac{\sigma_k}{\sigma_0^2}$$

$$\text{where} \quad \sigma_k = E[(x_t - \bar{x})(x_{t-k} - \bar{x})]$$

$$\text{and} \quad \sigma_0^2 = E[(x_t - \bar{x})^2]$$

Note that we only have 1 $\bar{x}$ since the mean is the same due to stationary data.

Gauss-Markov Theorem: Under assumption 1 to 5, OLS estimators are the best linear unbiased estimators BLUE. Recall that we need *strict exogeneity* for assumption 3 for this to hold.

1.2.6 Assumption 6: Normality

ϵ_t is iid as normal random variables with zero mean and σ^2 variance.

$$\epsilon_t \sim \text{N}(0,\sigma^2)$$

$$\text{t} = 1,2,...,\text{T}$$

The *Gauss Markov Assumptions and normality assumption* gives us the **Classical Linear Model assumptions** (CLM) for time series. We need this so that we can perform exact inferences.

1.3 Classical Linear Model Assumptions

From assumptions 1-6, we can now carry out statistical inference procedures for time series data. t statistics are $\sim t_{T-k-1}$ distributions under the null, where k is the number of exogenous variables in our regression. Furthermore, confidence intervals have the specified confidence levels and F-statistics have exact F-distributions.

From this, we have a set of assumptions (the most important being strict exogeneity) under which OLS is unbiased.

When spherical disturbances assumption holds, OLS is BLUE and usual OLS variance formulas apply. However, serial correlation tends to be a problem even when strict exogeneity holds. Furthermore, we can add normality and this leads to exact inference. However, this tends to be highly unrealistic for many datasets.

2 Static and Distributed Lag Models

First, we define a stationary process is where probability distributions are stable over time. More formally, the joint distirbution of random variables from any set of time periods remain unchanged. So the joint distribution of $\{x_{t_1}, x_{t_2}, ...\}$ is the same as $\{x_{t_1+h}, x_{t_2+h}, ...\}$ for h $\geq$ 1 and therefore the series is **identically distributed**. We know that seasonal and trending data is already non-stationary from this. The idea of stationarity is different to the idea of **weakly dependent** as this idea states that the correlation between $corr(x_t, x_{t+h}) \to 0$ as $h \to \infty$. This means that the process is asymptotically uncorrelated. With this, we can also define the idea of **covariance stationarity** which requires 3 conditions:

1) E(x_t) = c, whereby c is a constant.

2) var(x_t) = k, whereby k is a constant.

3) cov(x_t, x_{t+h}) = ρ_h whereby the covariance only depends on h, and not t.

Note that stationarity is a **property of a process, not time series themselves**. These processes generate time series.

Recall there are 3 assumptions we want with time series data: linearity in parameters, exogeneity, and uncorrelated error terms.

2.1 Static Models

A static model is only for the current time period and relates two or more time series with each other.

$$y_t = \alpha_t = \beta_1 x_{1,t} + ... + \beta_k x_{k,t} + \epsilon_t$$

whereby β_j, j=1,...,k looks at only the **contemporaneous** relationship between $\{x_j\}$ and {y}. Furthermore, $\epsilon_t \sim$ iid(0,σ^2). ϵ_t is referred to as **white noise**. In order words, white noise are where observations aren't autocorrelated with each other, homoskedastic, and indexed by time.

Recall that serial correlation is the correlation of a variable with itself over time. Serial correlation for the error terms does not affect the bias or consistencies of least squares. If $E(\epsilon_t|X) = 0$, then explanatory variables are strictly exogeneous leading to unbiased OLS. If instead, $E(\epsilon_t \mid x_t) = 0$, then explanatory variables are only *contemporaneously exogenous* and therefore OLS is consistent provided that the time series are *weakly dependent*. Weakly dependence refers to the fact that a x_t and x_{t+h} are almost independent as h $\to \infty$. This means as the variabels get further apart in time, the correlation between the variables become smaller and smaller.

However, autocorrelated errors (serial correlation) means that we have issues with statistical inference, even in the case of large samples. Furthermore, measures such as R^2 and $\bar{R^2}$ are invalidated. These goodness of fit measures are useless if the serial correlation is a result of a spurious regression if series {y} and (some of the) {x} have *unit roots*. Just briefly explaining, unit roots means that the effects of any shocks do not disappear over time (which is the opposite of weakly dependent). However, if the data is *weakly dependent* (if data is weakly dependent, then we don't have an unit root since unit roots are highly persistent/strong dependent), then these measures are now reliable!

To elaborate more on *unit roots*, we can think of them as a **stochastic trend** in time series. If we had a time series of:

$$y_t = c + \alpha_1 y_{t-1} + \epsilon_{t-1}$$

the coefficient α_1 is a root. We expect this process to always converge back to the value of c when $\alpha < 1$. If we set c = 0 and $\alpha = 0.5$, if y_{t-1} was 100, then today it's 50, tomororow 25, and so on until it gets to 0. Here, we can see that this series will converge back to c. However, if we had a root that is a **unit**, or in other words, when $\alpha = 1$, we see that the series will never converge back to c. From this, we can see that the time series will never recover back to its expected value and therefore the process is very susceptible to shocks and hard to predict. 3 ways for autocorrelation to occur:

1) Ommited variable bias.

2) Functional misspecfication

3) Measurement error in the independent error

2.2 Testing for Serial Correlation

We can test for serial correlation! We specify simple alternative models that allow the errors to be serially correlated and use the model to test the null that the errors are not serially correlated. From this, we can derive the first-order autocorrelation:

$$\epsilon_t = \rho \epsilon_{t-1} + v_t$$

whereby $v_t \sim$ iid(0,σ_v^2). v_t is a white noise process so that if $\rho = 0$, then that means the error term ϵ_t is just iid. The error term is a function of its own lag. We don't include an intercept due to zero conditional mean. We can think of this as seeing whether does the previous error term have a relationship/effect on the current error term. If it does ($\rho \neq 0$), then there is autocorrelation as the error terms are related (correlated). The null hypothesis of this is that there is no serial correlation such that H_0: $\rho = 0$. Often though, $\rho > 0$, when there is serial correlation but we still use a two-sided alternative. However

in practice, we can't actually observe error terms ϵ_t so we instead use the OLS residual $\hat{\epsilon_t}$. From this, if the explanatory variables are strictly exogenous, we can use a simple t-test. Furthermore, we can actually use the t-test as long as $E(\epsilon_t|x_t, x_{t+1}) = 0$ or that the error in a given time period is uncorrelated with regressors contemporaneously and in the next time period. So here, if we are using just a single lagged residual, we only lose one observations since if we had 100 observations, we can only take 99 lag variables and therefore we can't test it on one observation so we drop that observation.

Steps to test for serial correlation are (under strict exogeneity):

Step 1. Set up a time series model and run the regression:

$$y_t = \alpha + \beta_1 x_{1,t} + ... + \beta_k x_{k,t} + \epsilon_t$$

whereby t=1,...,T

Step 2. Using the residuals from step 1, run the regression

$$\hat{\epsilon_t} = \rho \hat{\epsilon_{t-1}} + v_t$$

whereby t=2,...,T. Doesn't matter if we include an intercept or not! They are asymptotically equivalent. v_t is a white noise and iid. If we don't reject $\rho = 0$, then $\hat{\epsilon}_t = v_t$ so therefore $\hat{\epsilon}_t$ are a white noise process.

Step 3. Compute the **t-statistic** for $\hat{\rho}$ and test whether H_0: $\rho = 0$. Therefore if we reject, there is autocorrelation with the error terms and we cannot do inferences.

The test mentioned has large-sample justifications and tends to work well. Standard errors are wider if small time period and therefore we might not reject even if $\hat{\rho}$ is "large" since there is alot of room for error. Furthermore, standard errors are smaller if large time period so much more likely to reject $\hat{\rho}$ even if its small. This is due to the fact we have not considered statistical vs practical significance. It may be the case that for a large sample size, we just happened to have found some correlation (and therefore since sample size is large, it's easy to reject things). Note that we must assume homoskedasticity for this test, if heteroskedasticity, then we use robust t-statistics (these are smaller standard errors so it is harder to reject).

From this, we can check for even higher-order autocorrelation.

Step 1. Set up a time series model and run the regression:

$$y_t = \alpha + \beta_1 x_{1,t} + ... + \beta_k x_{k,t} + \epsilon_t$$

whereby t=1,...,T

Step 2. Using the residuals from step 1, run the regression

$$\hat{\epsilon_t} = \rho_1 \hat{\epsilon_{t-1}} + \rho_2 \hat{\epsilon_{t-2}} + ... + \rho_q \hat{\epsilon_{t-q}} + v_t$$

whereby t = q+1,...,T. Doesn't matter if we include an intercept or not! They are asymptotically equivalent.

Step 3. We then can use a **F-statistic** to test the joint hypothesis that H_0: $\rho_1 = ... = \rho_q = 0$ in the usual way. Therefore, we can test multiple residuals. If we don't reject null, then $\hat{\epsilon}_t = v_t$ so the residuals are a white noise process.

We now consider the case whereby the regressors are no longer strictly exogenous.

2.2.1 Breusche-Godfrey Test for serial correlation

In the scenario that regressors are not strictly exogenous, we can no longer run the previous test for serial correlation. Instead, we now need to consider these endogenous regressors when testing for serial correlation, which then leads us to the **Breusche-Godfrey test**. Our estimates of ρ in the auxillary regression are going to be biased and now we need to correct it. What we can do is to include all of the explanatory variabels from step 1 to the residual autocorrelation regression in step 2 and then run the F-statistic again. **Note we include the intercept** γ_0. We are attempting to model:

$$\hat{\epsilon_t} = \rho_1 \hat{\epsilon_{t-1}} + \rho_2 \hat{\epsilon_{t-2}} + ... + \rho_q \hat{\epsilon_{t-q}} + \gamma_1 x_{1,t} + ...\gamma_k x_{k,t} + \gamma_0$$

We run a F-statistic for joint hypothesis of $H_0 : \rho_1 = ... = \rho_q = 0$. This is known as the *Breusch-Godfrey test for* q^{th} *order residual autocorrelation.* We have corrected for endogenous variables since if predictors are related to error terms, by including them into the equation for the test, we now consider the effects of the regressors as well. We can also include any number of lagged dependent variables as well to regress on. We can also use robust t-statistics in case of heteroskedasticity. In practice, we don't know how large q is so we normally just try random stuff. We can also use the **Lagrange multipler** to test for it, where LM is:

$$LM = (n - q)R_u^2$$

where n is sample size, q is the order residual autocorrelation we are testing for, and the R_u^2 is the R-squared from regressing the residual on lagged residuals and regressors. Here, $LM \sim \chi_q^2$ distribution and recall that is a upper one sided test (we can also correct for heteroskedasticity with this).

An additional note is that if we want to check for seasonal forms of serial correlation, then for quarterly seasonality, we check for:

$$\hat{\epsilon_t} = \rho_1 \hat{\epsilon}_{t-4}$$

and then do a t-test on $\rho_1 = 0$ or not. If it was yearly, then check on t-12.

2.3 Effects of serial correlation

Recall that OLS parameter is:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{t=1}^{n} x_i \epsilon_i}{SST_x}$$

whereby $d_i = x_i$ since $\bar{x} = 0$ (normally it is $d_i = x_i - \bar{x}$). We simply let $d_i = x_i$ for this question then. Recall that variance of OLS estimator is then:

$$Var(\hat{\beta}_1) = Var(\frac{\sum_{t=1}^{n} x_i \epsilon_i}{SST_x})$$

$$= \frac{1}{SST_x^2} Var(\sum_{t=1}^{n} x_i \epsilon_i)$$

since SST_x is a constant. Now we ignore the fraction in the front (because I'm lazy). Recall we are conditioning on $\mathbf{X}$ so x_i is a constant.

$$= Var(\sum_{t=1}^{n} x_t \epsilon_t | \boldsymbol{X})$$

$$= \sum_{t=1}^{n-1} \sum_{j=1}^{n-t} x_t x_{t+j} Cov(\epsilon_t, \epsilon_{t+j})$$

$$= \sum_{t=1}^{n-1} x_t^2 Var(\epsilon_t) + 2 \sum_{t=1}^{n-1} \sum_{j=1}^{n-t} x_t x_{t+j} Cov(\epsilon_t, \epsilon_{t+j})$$

Recall that $Cov(\epsilon_t, \epsilon_{t+j}) = E(\epsilon_t, \epsilon_{t+j})$.

$$= \sum_{t=1}^{n-1} x_t^2 Var(\epsilon_t) + 2 \sum_{t=1}^{n-1} \sum_{j=1}^{n-t} x_t x_{t+j} E(\epsilon_t, \epsilon_{t+j})$$

Since ϵ_t is a AR(1) serial correlation, it has the form of a random walk. Recall that for a random walk, $E(\epsilon_t, \epsilon_{t+j}) = \rho^j \sigma^2$.

$$= \sum_{t=1}^{n-1} x_t^2 Var(\epsilon_t) + 2 \sum_{t=1}^{n-1} \sum_{j=1}^{n-t} x_t x_{t+j} \rho^j \sigma^2$$

$$= \sum_{t=1}^{n-1} x_t^2 Var(\epsilon_t) + 2\sigma^2 \sum_{t=1}^{n-1} \sum_{j=1}^{n-t} x_t x_{t+j} \rho^j$$

since σ^2 is a constant. Now recall that $SST_x = \sum_{t=1}^{n} x_t^2$.

$$= SST_x Var(\sum_{t=1}^{n} \epsilon_t) + 2\sigma^2 \sum_{t=1}^{n-1} \sum_{j=1}^{n-t} \rho^j x_t x_{t+j}$$

Now we add in our earlier fraction in the front $\frac{1}{SST_x^2}$

$$= \frac{SST_x Var(\epsilon_t)}{SST_x^2} + \frac{2\sigma^2 \sum_{t=1}^{n-1} \sum_{j=1}^{n-t} \rho^j x_t x_{t+j}}{SST_x^2}$$

$$= \frac{Var(\epsilon_t)}{SST_x} + \frac{2\sigma^2 \sum_{t=1}^{n-1} \sum_{j=1}^{n-t} \rho^j x_t x_{t+j}}{SST_x^2}$$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_x} + \frac{2\sigma^2 \sum_{t=1}^{n-1} \sum_{j=1}^{n-t} \rho^j x_t x_{t+j}}{SST_x^2}$$

where $var(\epsilon_t = \sigma^2$ since it is still homoskedastic. The first term is the usual variance of an $\hat{\beta}_1$. The second term is the bias. Note that if $\rho > 1$, then the variance of $\hat{\beta}_1$ for an OLS estimate will be underestimating the true variance since it does not take into account of the second term. From this, usual OLS when assuming no serial correlation (but there actually is), will cause the standard errors to be smaller than what it actually should be, leading to test statistics that are too big and then committing too many type 1 errors (falsely rejecting null hypothesis).

2.4 Correcting Serial Correlation

Let us suppose we know from the *Breusch-Godfrey test* that our model suffers from serial correlation. We know that the efficiency of OLS estimators is affected. Our standard errors will be off meaning that we can no longer carry out inferences. We also know that goodness of fit is off if the data is not weakly dependent either. Terrible news! From that, we can correct the serial correlation.

Assume we have a linear regression with serially correlated errors and strict exogeneity assumption holds:

$$a) \quad y_t = \beta_0 + \beta_1 x_t + \epsilon_t$$

Whereby t=1,...,T and also that:

$$\epsilon_t = \rho\epsilon_{t-1} + v_t$$

such that $v_t \sim \text{iid}(0,\sigma_v^2)$. Here, both ϵ_t and v_t (our white noise) are assumed to be independent error processes. We can say that ϵ_t is an AR(1) process but we assume that $|\rho| < 1$ or else some bad stuff is gonna happen since it'll be unstable. We can do some magical algebraic manipulation. Let us then look at t-1:

$$y_{t-1} = \beta_0 + \beta_1 x_{t-1} + \epsilon_{t-1}$$

We can multiply though by ρ

$$b) \quad \rho y_{t-1} = \rho\beta_0 + \rho\beta_1 x_{t-1} + \rho\epsilon_{t-1}$$

$$a) \quad y_t = \beta_0 + \beta_1 x_t + \epsilon_t$$

We can now subtract b from a.

$$y_t - \rho y_{t-1} = (\beta_0 - \rho\beta_0) + (\beta_1 x_t - \rho\beta_1 x_{t-1}) + (\epsilon_t - \rho\epsilon_{t-1})$$

$$\tilde{y_1} = (1-\rho)\beta_0 + \beta_1\tilde{x_t} + \epsilon - \rho\epsilon_t$$

Since we know that $v_t = \epsilon_t - \rho\epsilon_t$

$$\tilde{y_t} = \tilde{\beta_0} + \beta_1\tilde{x_t} + v_t$$

whereby t=2,...,T. We have taken the *quasi-difference* of the variables. We let that: $\tilde{y_t} = y_t - \rho y_{t-1}$, $\tilde{x_t} = x_t - \rho x_{t-1}$ and $\tilde{\beta_0} = (1-\rho)\beta_0$. Furthermore, recall $\epsilon_t - \rho\epsilon_{t-1} = v_t$ for a white noise process. Now we have corrected for serially correlated errors and now have a white noise process!!! Inference is all good now. In fact, we satisfy all *Gauss-Markov assumptions*. Note that if $\rho = 1$, we have then just differenced it (not quasi), but we assumed that $|\rho| < 1$. However, note that this only works for the observations from 2,...,t, since we cannot difference the first period. Therefore, for estimators to be BLUE, we need to specify that the first time period is simply:

$$y_1 = \beta_0 + \beta_1 x_1 + \epsilon_1$$

and all the v_t's are uncorrelated with ϵ_1 so our issue is fine now and we achieved BLUE estimators.

2.4.1 Cochrane-Orcutt Correcting Serial Correlation (Feasible GLS)

We can use the *Cochrane-Orcutt* procedure to estimate the model parameters.

1) Regress y_t on x_t and obtain the residuals $\hat{\epsilon}$

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t \quad t = 1, ..., T$$

2) Regress $\hat{\epsilon}_t$ on $\hat{\epsilon_{t-1}}$ and obtain $\hat{\rho}$

$$\hat{\epsilon_t} = \rho \hat{\epsilon_{t-1}} + v_t$$

3) From quasi-differenced variables (whereby we multiply the lag y_{t-1} by ρ and then difference y_t and $y_{t-1}\rho$), we regress $\tilde{y_t}$ on $\tilde{x_t}$ to obtain parameter estimates

$$\tilde{y_t} = \tilde{\beta_0} + \beta_1 \tilde{x_t} + \tilde{\epsilon_t}$$

From this, the usual standard errors, t and F statistics are asymptotically valid. An issue is that using $\hat{\rho}$ instead of ρ causes estimators to be biased but it does lead to more efficiency asymptotically (assuming weakly dependent time series) and consistent. Even if $\tilde{\epsilon_t}$ is not normal, we stil have approximately t and F-statistics. Usually, the Cochrane-Orcutt procedure is iterative whereby we repeat steps 2 and 3, to derive new parameter estimates until there is no change in the estimate of ρ from successive iterations. To run this command in stata, we would go: *prais dependentvar independentvar, corc.* Additionally, we can regress y on x and save residuals. Then regress residuals on their lags and save $\hat{\rho}_1$. From this, we go $y - \rho y_{t-1}$ and $x - \rho x_{t-1}$ to generate our new variables. Note that if we included our first observation $y_1 = \beta_0 + \beta_1 x_1 + \epsilon_1$, this is known as the **Prais-Winsten estimation**. Either method should not give us results that differ greatly. More importantly, we may notice that these the Cochrane-Orcutt method may result in coefficents that do not differ too much from OLS (*if the variables are I(0)* BUT the standard errors can be significantly higher to account for the serial correlation (and therefore some variable may now not be statistically significant and also less economically important). **Note that we cannot compare the R^2 of these models with the OLS models since the dependent and independent variables are now different**.

On a final note, we note that sometimes OLS and Cochrane-Orcutt/Prais-Winsten coefficients can vary significantly. This could be because the variables are not related in levels (without differencing) but they are after we quasi-difference them. This leads to an issue since if we are trying to look at static relationships between levels of variables, but they are not I(0), then OLS won't produce consistent estimations. This can arise if the variables have unit roots and doing FGLS (Cochrane-Orcutt) will eliminate these unit roots.

2.5 Finite Distributed Lag (FDL) Models

Instead of looking at things in a contemporaneous measure, we can now add in lag effects to see how things in the past affect today and the future! Let's create a variable Z that

we think can affect y for up to 2 periods in the future. From this, we derive the *finite distributed lag model* as:

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + \epsilon_t$$

which includes two lags. For example, we could model:

$$umbrellassold_t = \alpha_0 + \delta_0 rain_t + \delta_1 rain_{t-1} + \delta_2 rain_{t-2} + \epsilon_t$$

which tells us that umbrellas sold today is affected by amount of rain today, yesterday, and 2 days ago (since if it's been raining for past few days, people will start giving in and buying their own umbrellas instead of stealing it from poor souls at Fisher library).

We can generalise this by if we want a FDL of order q:

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + ... + \delta_q z_{t-q} + \epsilon_t$$

We specify q on what we believe (or statistical test) is a good value of lags to include. δ_0 is the **impact propensity** which tells us the *immediate change* in y when z increases by 1 unit. So if $\delta_t = 0.5$, then a 1 unit increase in z will increase y by 0.5 units ceterus paribus. As an example, suppose we had the FDL(2) model (with no error terms to keep it simple):

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2}$$

Mathematically, imagine we had that t=t-1, $z_t = 20$ units, and then at t=t, we increase it by 1 unit to 21 and then at t=t+1, it drops back down to 20.

$$y_{t-1} = \alpha_0 + \delta_0(20) + \delta_1(20) + \delta_2(20)$$

$$y_t = \alpha_0 + \delta_0(20+1) + \delta_1(20) + \delta_2(20)$$

$$y_{t+1} = \alpha_0 + \delta_0(20) + \delta_1(20+1) + \delta_2(20)$$

$$y_{t+2} = \alpha_0 + \delta_0(20) + \delta_1(20) + \delta_2(20+1)$$

$$y_{t+3} = \alpha_0 + \delta_0(20) + \delta_1(20) + \delta_2(20)$$

Notice that time is increasing by 1 period every equation. We see that the effect of the 1 unit increase propagates throughout the model. So if we take the difference between the first 2 equations (or at time t vs t-1):

$$= \delta_0(1)$$

and we can see that at time t, this is the immediate impact. Furthermore, at time t+1, we difference the first and 3rd equation to get:

$$= \delta_1(1)$$

which tells us the affect of z_{t-1} on y_t or the change in y *one period* after the *temporary* change in z_t. Furthermore, if both z and y are in logarithmic form, the impact propensity is also known as the **short run (instantaneous) elasticity**. This is similar to our static model we had earlier. If we believe there to be no impact propensity, we can just drop z_t from the regression. (or if it is genuinely 0, then $\delta_0 = 0$).

Using the same example as before, we now have that if we increase z_t by 1 unit **permanently**, we have that:

$$y_{t-1} = \alpha_0 + \delta_0(z_t) + \delta_1(z_{t-1}) + \delta_2(z_{t-2})$$

$$y_t = \alpha_0 + \delta_0(z_t + 1) + \delta_1(z_{t-1}) + \delta_2(z_{t-2})$$

$$y_{t+1} = \alpha_0 + \delta_0(z_t + 1) + \delta_1(z_{t-1} + 1) + \delta_2(z_{t-2})$$

$$y_{t+2} = \alpha_0 + \delta_0(z_t + 1) + \delta_1(z_{t-1} + 1) + \delta_2(z_{t-2} + 1)$$

and if we take $t + i \quad i \to \infty$ or into the future, the effect of a permanent increase in z_t is the sum of all the δ. The summation of all δ values is known as the **long run propensity** s.t LRP $= \sum_{t=0}^{t=q} \delta_t$. The LRP tells us that if we permanently change z at a given time, the LRP is the ceteris paraibus change in y after the change in z has passed through all q time periods. So if minimum wage rose by 1 dollar an hour, then substitute the value of 1 for z_t and sum the coefficients. If y and z are both in natural logarithms, then the LRP is known as the **long run elasticity**. Note we can also include dummy variables to account for certain events. Furthermore, note that since z_t is serially correlated with itself through time (which is fine and should be expected!), it will be difficult in getting precise estimates of each δ. From this, even if we run a F-test on $\delta_0 = ... = \delta_q = 0$ and find that it is statistically significant, we still don't know which lag period is significant with y_t (they may be jointly significant but individually insignificant). If $\delta_1 = \delta_2 = ... = \delta_q = 0$, then it will suggest to us to use a static model since lags are insignificant. *Note that LRP can be used for a single lagged regressor in the case of multiple regressors.* If we had:

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2}$$

and we found LRP $= \delta_0 + \delta_1 + \delta_2 = \lambda$, we can then compute a significance test on the LRP and its confidence intervals. We let: $\theta = \delta_0 + \delta_1 + \delta_2$ and re-express as: $\delta_0 = \theta - \delta_1 - \delta_2$. Sub into the above equation to get:

$$y_t = \alpha_0 + (\delta_1 + \delta_2 - \theta)z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2}$$

$$y_t = \alpha_0 + \theta z_t + \delta_1 z_{t-1} - \delta_1 z_t + \delta_2 z_{t-2} - \delta_2 z_t$$

$$y_t = \alpha_0 + \theta z_t + \delta_1(z_{t-1} - z_t) + \delta_2(z_{t-2} - z_t)$$

Regress this final equation to get $\hat{\theta}$ which equals our LRP. Now though, we have a standard error from this regression which we can then use to run a significance test on the long run propensity and compute confidence intervals.

2.6 Infinite Distributed Lag (IDL) Models

IDL models are FDL models but with q = ∞. We model a dependent variable to current and all past values of the independent variable. IDL models are quite unrealistic in the sense that we don't have data that has existed for an infinite time horizon.

$$y_t = \alpha + \delta_0 z_t + \delta_1 z_{t-1} + ... + \epsilon_t$$

Even though its unrealistic, this still gives us good approximations and interesting things to look at! However, we do require the fact that $\delta_j = 0$ as $j \to \infty$. Here, the effect of a z_t variable on y_t decreases the further back in time we go! (Which makes sense since inflation in 1900 shoudn't affect inflation in 2017). Therefore, the effects are strongest for more recent and contemporaneous effects. Note that this doesn't mean that 2016's effect has to be larger than 2015's, just that when it's really far back, it gets smaller.

From all this, we can interpret IDL models quite similar to FDL models. For a temporary change in z, so at time = -1, z = 0, then at time = 0, we have that z increases by 1 unit and one period later, reverts back to its initial level 0 and stays at 0 for the rest of time. Then, for any period $h \geq 0$, we have that:

$$y_h = \alpha + \delta_j + \epsilon_h$$

since all the other terms will have died off. δ_h here is the temporary change in E(y_h) given a one unit temporary change in z at time 0. More formally, we have E(y) = α + δ_h $\forall$ h $\geq$ 0. Like we said earlier, for IDL to make sense, we need that $\delta_h \to 0$ as time progresses or $h \to \infty$ (so the effect dies off). This means that **there is no long run effect from a temporary change in z**. E(y) $\to \alpha$ as h $\to \infty$ since $\delta \to 0$. More generally, δ_h measures the change in the expected value of y after h periods. However though, after sufficient time, this means we expect the average value of y to be a constant. We can still consider the Long run propensity to be LRP = $\delta_0 + ... + \delta_h$ as h $\to \infty$. From this, we can say that the infinite sum is well definied since δ_j must converge to 0 and therefore IDL models can be approximated by $\delta_1 + ... + \delta_p$ for a sufficiently large p. Then that means, for permanent increase in z, we have that: $E(y_h) = \alpha + \delta_0 + ... + \delta_h$ or the Long run propensity plus α. We implictly assume strict exogeneity for this model whereby even all future values of z_t are uncorrelated with the error term. An issue with this model is that it assumes that future values of z_t are constant and never changing which isn't realistic (since if z_t was interest rates, this would definitely be changing in 20 years time compared to today). Therefore, we use a weak exogeneity assumption instead (where the error is uncorrelated with current and past z).

2.7 Geometric Distributed Lag (GDL)/Koyck distributed Lag Models

We need to place some assumptions on the infinite Distributed lag model so that we can actually estimate it (since currently there are an infinite number of parameters). We can use the **geometric/Koyck** distributed lag (GDL) model. From this, the new δ_j depends on two parameters:

$$\delta_j = \gamma\rho^j$$

whereby $|\rho| < 1$, j=0,1,...

We can let γ and ρ be positive or negative, but we need to make sure that $\rho < \gamma$ so that $\delta_j \to 0$ as time period $j \to \infty$. So if γ is 1 and $\rho = 0.5$, then $\delta_1 = 0.5$, $\delta_2 = 0.25$, ... until it converges to 0. This means that something in far away in time doesn't have much of an effect. Furthermore, ρ has to be less than 1 or else the terms will explode (try the above example whereby we set ρ to 1.5 instead of 0.5, you'll see δ_j will keep growing). Therefore from this, we can estimate the IDL model but we make sure we set the coefficients $\delta_j = \gamma\rho^j$.

So for the GDL, we have:

$$y_t = \alpha + \delta_0 z_t + \delta_1 z_{t-1} + ... + \epsilon_t$$

whereby $\delta_j = \gamma\rho^j$. The *impact propensity* is $\delta_0 = \gamma_0$ so the sign of impact propensity is determiend by the sign of γ.

$$y_t = \alpha + \gamma\rho^0 z_t + \gamma\rho^1 z_{t-1} + \gamma\rho^2 z_{t-2} + ... + \epsilon_t$$

then becomes:

$$y_t = \alpha + \gamma z_t + \gamma\rho z_{t-1} + \gamma\rho^2 z_{t-2} + ... + \epsilon_t$$

Note that if $\gamma > 0$ and $\rho > 0$, then all lag coefficients are positive. However, if $\rho < 0$, then the lag coefficients will alternative in sign. We can also compute the long-run propensity by using the sum of geometric series which is:

$$1 + \rho + \rho^2 + ... = \frac{1}{1-\rho}$$

which, when we apply it to y_t to compute the LRP:

$$LRP = \gamma + \gamma\rho + \gamma\rho^2 + ... = \frac{\gamma}{1-\rho}$$

It also follows that the LRP will have the same sign as γ.

2.7.1 Application of Koyck

If we wanted to actually estimate this model, first we have our original equation and its lagged:

$$y_t = \alpha + \gamma z_t + \gamma\rho z_{t-1} + \gamma\rho^2 z_{t-2} + ... + \epsilon_t$$

$$y_{t-1} = \alpha + \gamma z_{t-1} + \gamma\rho z_{t-2} + \gamma\rho^2 z_{t-3} + ... + \epsilon_t$$

We then multiply the second equation by ρ.

$$\rho y_{t-1} = \rho\alpha + \rho\gamma z_{t-1} + \gamma\rho^2 z_{t-2} + \gamma\rho^3 z_{t-3} + ... + \rho\epsilon_t$$

We then subtract this equation from our original equation and lots of terms will cancel out:

$$y_t - \rho y_{t-1} = (1-\rho)\alpha + \gamma z_t + \epsilon_t - \rho\epsilon_{t-1}$$

Rearrange ρy_{t-1} and define new variables:

$$y_t = \alpha_0 + \gamma z_t + \rho y_{t-1} + \epsilon_t - \rho\epsilon_{t-1}$$

Notice that we now have a FDL model of order 1! Hoever, if we run OLS on this, we get inconsistent estimates as y_{t-1} is correlated with ϵ_{t-1}. We can use two-stage least-squares to solve this but that's for another unit :) (Although if you are interested, we can use an instrumental variable x_{t-1} for y_{t-1} and estimate the model with that instead).

2.8 Conclusion

Static models are quite similar to cross-sectional methods since data ordering doesn't matter (since only 1 time period). They are quite good at estimating contemporaneous relations between two or more variables. However, they aren't able to capture lagged effects which is why distributed lag models are useful for this(although we need to be careful with the order of the data now!). However, we can't use any of these models for forecasting since they don't consider the fact that past values of y, can help to forecast y (note that we only used lags of independent variables and not dependent). Furthermore, they require that future realisations of x need to be known in order to forecast those y values (which is impossible to obtain since we would be in the future and already have the y values if we had those future x values...). Conclusively, the models we have explored so far are good in terms of interpreting casual effects but not so much in actual predictive capabilities.

3 Deterministic and Stochastic Trends

3.1 Trending Data

Economic time series tend to change unidirectionally over time. When we see a series grow or shrink, we need to be careful when modelling and interpreting relationships between 2 or more variables. We wish to avoid **spurious** regressions, whereby we think there is a significant correlation when in fact, there is none. A systematic change in a time series that does not appear to be periodic is known as a trend.

Spurious Relationship: Finding a significant relationship between 2 time series variables when in actual fact, there is no relationship between them. One good example could be the number of movies Jennifer Lawrence has appeared in and the size of Australia's GDP. Both are growing from between 2000-2017. Therefore if we ran a regression on this, we would actually find a statistically significant relationship between these 2 variables! However, we know that Lawrence has no effect on Australia's GDP no matter how good her movies are. We in fact can actually attributes this due to **linear trends**, that is, both variables are growing over time, so of course it may appears that there is some correlation going on!

From this, in order to capture true relationships between regressors and dependent variables, we can add a trend variable (t) in order to help control for the case in which any of the variables in the model are linearly trending.

$$y_t = \beta_0 + \beta_1 x_{1t} + \gamma_t + \epsilon_t$$

Where t = 1,2,...,T

Here, this allows us to control for a linear trend that affects y_t which may relate to trends in x_{1t} and x_{2t}. If assumptions of linearity, no perfect collinearity, and strict exogeneity holds, then if we do not include the trend t, this can cause bias in estimating β_1 and β_2. If we had all the assumptions of the classical linear models satisfied, then we can apply test statistics and confidence intervals in the usual way.

However, when dependent variable is trending, we can still have a high R-square which is very misleading since we overestimate the variance in Y and therefore this leads to a high R^2 even with a time trend (since the time trend is capturing alot of the variance). Therefore, we need to be careful in interpreting the goodness of fit when the dependent variable is trending. Here, the r-square captures the variance of the model.

The time trend t represents our ignorance about omitted factors causing y_t to trend up or down. However, we should still be happy with it since we can still fit y to a trend.

If you think about it, adding a time trend to MLR model, has a nice interpretation of detrending y and all explanatory variables. We show what it means to instead **detrend the variables**.

We can detrend variables by regressing each predictor on a time trend and then differencing the actual value from the predicted value. Therefore, first we estimate:

$$y_t = \hat{\alpha_0} + \hat{\alpha_1} t + \hat{\epsilon_t}$$

and we also estimate:

$$x_t = \hat{\alpha_0} + \hat{\alpha_1} t + \hat{\epsilon_t}$$

From this, we can the obtain the detrended series y_t^* by subtracting estimated trend component from original series:

$$y_t^* = y_t - \hat{\alpha_1} t$$

$$y_t^* = \hat{\alpha_0} + \hat{\epsilon_t}$$

and also for predictors

$$\tilde{x_t} = x_t - \hat{\alpha_1} t$$

$$\tilde{x_t} = \hat{\alpha_0} + \hat{\epsilon_t}$$

This gives us the residuals of each variables. From Frisch-Wraugh theorem, the residuals are left over stuff that aren't related to the time trend. From this, we can run the final regression:

$$y_t^* = \theta_0 + \theta_1 \tilde{x_t} + \epsilon_t$$

An easier to think about it would be, let us regress y_t, $x_{t,1}, x_{t,2}..., x_{t,k}$ on a constant and time trend t.

$$y_t = \alpha_0 + \alpha_1 t + \epsilon_t$$

From that, we save save the residuals $\dot{y_t}, \dot{x_t}....$ We then regress this residuals on each other (where we don't need an intercept since the intercept will just be 0).

$$\dot{y_t} = \dot{x}_{t,1} + \dot{x}_{t,2}$$

These also yield identical parameter estimates. From this, the parameter's coefficient and statistical significance will be the same as in the model with a linear trend except the R-square in the detrended model will be significantly lower. Conclusively, OLS coefficient in model with time trend is the same as running a regression on all detrended variables

(whether or not we needed to detrend the explanatory variables as well) and then fit a regression using the detrended series.

We tend to use linear specification to account for linear specification.

$$y_t = \alpha_0 + \alpha_1 t + \epsilon_t$$

where t = 1,...,T and that $\epsilon_t \sim iid(0, \sigma^2)$. From this, the average value of y_t is a linear function of time:

$$E(y_t) = \alpha_0 + \alpha_1 t$$

whilst ϵ_t represents the deviations about the trend. Here, normally linear trends tend to be good enough of an approximation. Furthermore, we can define the change in y_t fromp period t-1 to t as: $\Delta y_t = y_t - y_{t-1}$. Δ is the temporal difference between periods. From this, the linear trend representation:

$$E(\Delta y_t) = E(y_t) - E(y_{t-1}) - [\alpha_0 + \alpha_1 t] - [\alpha_0 + \alpha_1 (t-1)] = \alpha_1$$

and therefore α_1 is the average change in $y_t \forall$ t. So from the original equation: $y_t = \beta_0 + \beta_1 x_{1t} + \gamma t + \epsilon_t$, the γ tells us what is the change in y net of x's are.

Some data are better approximated by exponential trends. Therefore we can use **exponential trends** whereby for strictly positive variables, we can capture an exponential trend by:

$$y_t = e^{\beta_0 + \beta_1 t + \epsilon_t}$$

then we can transform this further by taking logarithms on both sides to get:

$$ln(y_t) = \beta_0 + \beta_1 t + \epsilon_t$$

and now it is the natural logarithm of the variable that follows a linear trend.

We can also define the change in logs as: $\Delta ln(y_t)$ = ln(y_t) - ln(y_{t-1}). Then from this, under exponential trend representations:

$$E[\Delta ln(y_t)] = [\beta_0 + \beta_1 t] - [\beta_0 + \beta_1 (t-1)] = \beta_1$$

The log difference is equivalent to the growth rate. Effectively, we are estimating ln(y_t) = $\beta_0 + \beta_1 t + \epsilon_t$. However, recall that the change in log approximates the grow rate such that:

$$\beta_1 \approx \frac{\Delta y_t}{y_{t-1}}$$

Thus, note that with expontential trends, the rate of growth is constant. From this, this tells us the average growth over time. An important thing we need to consider in forecasting is the seasonality of the data. For annual data, this is not an issue. If 2 variables follow seasonality, this can lead to spurious regressions and therefore, we can include dummy variables (only 3 or else dummy variable trap) to account for this. Everytime we are in a particular season, the dummy variable activates. This is deterministic since we know exactly the trend depending on the quarter.

3.2 Stochastic Trends

The previous models we looked at tended for us to know the deterministic trend for the time series. However, an alternative to represent the trending series is to use a **random walk** process whereby:

$$y_t = y_{t-1} + \epsilon_t$$

whereby t=1,2,.. We can think of random walks as AR(1) models whereby ρ coefficient = 1. We also assume $\epsilon \sim$ white noise and Gaussian. From this, if we back substituted, we get:

$$y_{t-1} = y_{t-2} + \epsilon_t$$

$$y_t = y_{t-2} + \epsilon_t + \epsilon_{t-1}$$

and from all this, we eventually get

$$y_t = y_0 + \epsilon_1 + ... + \epsilon_t$$

$$y_t = \sum_{t=1}^{T} \epsilon_t = t\sigma^2$$

since we had the assumption that $\epsilon_t \sim N(0, 1)$. Furthermore, we can assume that $y_0 = 0$ since it is the beginning of time.

Since $y_0 = 0$, we have the following properties:

$$E(y_t) = E(y_0) = 0$$

$$var(y_t) = \sum_{t=1}^{T} var(\epsilon_t) = T\sigma_\epsilon^2$$

From this, we can then show that for any t $\geq$ 1, s $\geq$ 0

$$Cov(y_t, y_{t+s}) = Var(y_t) = \sigma_\epsilon^2 t$$

This means that the covariance between any 2 periods that isn't the starting period, the covariance between them is simply the number of time periods t times the error term. Since the covariance is a function of time, our series is not **covariance stationary**. Furthermore, it follows that for any t $\geq$ 1, s $\geq$ 0

$$Corr(y_t, y_{t+s}) = \frac{Cov(y_t, y_{t+s})}{\sqrt{var(y_t)}\sqrt{y_{t+s}}} = \sqrt{\frac{t}{t+s}}$$

Therefore, for sufficiently long time series, this implies that the correlation between y_t and y_{t+h} remains high (close to unity) even as s grows. This is known as **highly persistent**

or **strongly dependent** time series which does not satisfy the law of large numbers or the central limit theorem. Persistent means that the correlation will still persist even for a long time horizon. From this, we can see that y_t does not have correlations that die out fast enough as time distance increases. The bigger the time period, the bigger variance we get as a result and furthermore, we now have a sequence of random variables which do not disappear even as time moves on. The effects of shocks do not disappear and therefore violates the assumption of stationarity and covariance stationarity. It is important as to whether is a series highly persistent or not, since if we are looking at for example policy, then policies from 30 years ago can still be having an effect on GDP and stuff today. Random walks are a special case of a **unit root process** since $\rho = 1$, which means it does not converge to a particular value in the long run.

Do not confused trending and highly persistent. Something can be trending but not be highly persistent. However, often the case that highly persistent also contains a clear trend, such as a **random walk with drift**:

$$y_t = \alpha_0 + y_{t-1} + \epsilon_t$$

We see that back substitution gets us:

$$y_t = \alpha_0 t + \epsilon_t + \epsilon_{t-1} + ... + \epsilon_0 + y_0$$

and then taking expectation gets us:

$$E(y_t) = \alpha_0 t$$

so that the expected value of y_t is growing over time if $\alpha_0 > 0$. Furthermore, best predictions of y_{t+h} is $\alpha_0 h + y_t$, which is today's value plus a drift $\alpha_0 h$.

We note that for future forecasts t+h:

$$E(y_{t+h}) = y_t$$

so for any point in the future, our best forecast is the value today.

Highly persistent series such as random walks causes serious problems for regression analysis. If we regress y on x, then we get parameters that are statistically significant even though they are unrelated. Consider

$$y_t = \beta_0 + \beta_1 x_t + \xi_t$$

whereby $\{y_t\}$ and $\{x_t\}$ are the random walk processes such that:

$$y_t = y_{t-1} + \epsilon_t \quad \epsilon \sim iid(0, \sigma_\epsilon^2)$$

$$x_t = x_{t-1} + \nu_t \quad \nu \sim iid(0, \sigma_\nu^2)$$

We also assume that $Cov(\epsilon_t, \nu_t) = 0$, which then implies that $\{y_t\}$ is independent of $\{x_t\}$. Since y_t is a random walk, then we have that:

$$E(y_t) = E[y_0 + \epsilon_1 + \epsilon_2 + ... + \epsilon_t] = 0$$

and same for:

$$E(x_t) = E[x_0 + \epsilon_1 + \epsilon_2 + ... + \epsilon_t] = 0$$

From this, it would have suggested that:

$$\beta_0 = \beta_1 = 0$$

which then implies that:

$$\xi_t = y_t = \epsilon_t + \epsilon_{t-1} + ... + \epsilon_1$$

whereby ξ_t has a mean of zero, variance grows with t, and is *highly persistent.*

Resultantly, the OLS estimator of β_1 which is $\hat{beta}_1$ does not converge to zero, which means that OLS is inconsistent. Practically, as we get more data, $\hat{\beta}_1$ does not get closer to 0 or even converge to a specific value. From this, we will never figure out that y_t is unrelated to x_t. With random walks, the problem arises because $\{x_t\}$ has too much temporal correlations for the law of large numbers to hold. An even more serious issue is that the t-statistic for $H_0 : \beta_1 = 0$ does not have a t-distribution, even in large samples. Furthermore, the t-statistic rejects the null hypothesis too often and this issue gets worse as sampling size grows (we are more likely to make a type 1 error, rejecting true null, with more data).

Regressions with 2 or more independent random walks results in *spurious regression problem* in time series. The two series are independent but they appear to be strongly related since they could just trend in the same manner. From this spurious regression, the R^2 will be quite large.

A stochastic process $\{y_t : t = 1, 2, ...\}$ is **strict stationarity** if for every collection of time indices $t_1 < t_2 < ... < t_m$, the joint distribution of $(y_{t_1}, y_{t_2}, ..., y_{m_1}$ is the same as the joint distribution of $(y_{t_{1+h}}, y_{t_{2+h}}, ..., y_{m_{1_h}} \forall h \geq 1$.

From this, we can define **weakly stationary/covariance-stationary process** as a stochastic process $\{y_t : t = 1, 2, ...\}$ with a finite second moment $E(y_t^2 < \infty)$ whereby $E(y_t) = \mu$, $Var(y_t) = \sigma^2$, and $Cov(y_t, y_{t+h})$ depends only on h and not t (in other words, the covariance is a function of h, not t). If a strict stationary process has a finite second moment (constant variance), then it must be weakly stationary, but the converse is not necessarily true.

From this, a stationary process is **weakly dependent** if $cov(y_t, y_{t+h} = 0 \quad h \to \infty$. In other words, observations are almost independent as h increases. Therefore, the random walk process is nonstationary.

Furthermore, **trending time series** are also nonstationary (whether they are deterministic or stochastic). However, by detrending or differencing, we can transform nonstationary series into stationary series.

Trend-stationary applies for the case when the trend is deterministic and simply estimating the trend and removing it from the data, then the residuals left will be a stationary series. So as long as we include a trend, everything is fine.

Difference stationary: applies for the case when the trend is stochastic. Differencing the data will yield a stationary series.

Weakly dependent processes are said to be **integrated of order zero** or I(0). Therefore, we don't need to do anything to them before regressions since they satisfy standard limit theorems. Unit root processes like random walks are integrated of order one or I(1). This means that first differencing them will lead to a **weakly dependent** series (and often stationary process). Therefore, a time series that is I(1) is said to be a **difference-stationary process**, even though it is misleading as it emphasiss the staionarity after differencing rather than weak dependence. From this, something that is weakly dependent is not necessarily stationary (since mean and variance isn't necessarily constant) and vice versa. Weakly dependence relates to how the function of the covariance of a series should behave asymptotically whilst weakly stationary just requires the covariance to not be a function of time.

To summarise:

Definition 3.1. A time series is **strictly stationary** when the joint distribution of $Y_t, Y_{t-1}, ..., Y_{t-j}$ does not depend on t. The joint density $p(y_t, y_{t-1}, ..., y_{t-k})$ does not depend on t.

The joint distribution does not depend on time means that the values and such aren't changing due to time factor. If time series are stable, it is easy to estimate since the components are not changing over time. This means that the probability distribution function is the same across any index. However, strict stationarity is a very strong assumption so we have a simpler condition that is sufficient for ARMA models so now we use **weakly stationarity/covariance stationary**. A weakly stationary series is one whereby the mean is constant, variance is constant, and its covariance is *not* a function of time. In the literature, stationary refers to weakly stationary unless specified otherwise.

$$E(Y_t) = \mu$$

$$Var(Y_t) = \sigma^2$$

$$Cov(Y_t, Y_{t-k}) = f(k) \neq g(t)$$

Note that strictly stationary implies weakly stationary data. Do not get this confused with **weakly dependent time series** which states that the correlation for:

$$Corr(X_t, X_{t+h}) \to 0 \quad h \to \infty$$

This means that X_t becomes less correlated with values that are further away from the future. Note that weakly stationary data requires the Covariance to be a function of k and not time, whilst weakly dependence specifies the behaviour of this correlation/covariance needs to decrease to 0 as a function of k. A **strongly dependent/highly persistent** time series is the case in which this does not hold and is the opposite to a weakly dependent series. Also note that an unit root is not trend stationary (a shock will leave a trend stationary series to revert back to its mean whilst an unit root won't).

4 Autoregressive Process

Now we are doing actual time series models by attempting using the serial dependence in the response and predictors. Hence, if a variable is correlated with itself over time, we can regress itself over time. This gives us the name **autoregressive model**. An autoregression is a time series regression in which the dependent AND independent variables belong to the same stochastic process.

$$y_t = \alpha + \beta y_{t-1} + \epsilon_t$$
$$\text{where} \quad \epsilon \sim iid(0, \sigma_\epsilon^2)$$

This is an autoregression of order one or AR(1). Here, we make the predictor the previous period of response. Error term is not autocorrelated and iid unlike response variable. We know β tells us if the past variable was a certain amount, this period's response variable would be last period's response scaled by the parameter β. In reality, there are many other factors affecting y_t and related over time which can be reduced into the relationship with the y_t variable. We can model slow adjustments to long run equilibrium from the AR(1) model. The time it takes for shock (error term) to cause it to return to normal, depends on the size of β. This is why it can also be known as the **persistence** term.

We also have a **highly persistent** time series as a result. Suppose we have:

$$y_t = y_{t-1} + \epsilon_t$$

Where $\epsilon_t \sim iid(0, \sigma^2)$.

From this, we can now substitute recursively lagged dependent variables.

$$y_t = \alpha + \beta y_{t-1} + \epsilon_t$$

Sub in

$$y_{t-1} = \alpha + \beta y_{t-2} + \epsilon_{t-1}$$

To now get

$$y_t = \alpha + \beta(\alpha + \beta y_{t-2} + \epsilon_{t-1}) + \epsilon_t$$

Sub in

$$y_{t-2} = \alpha + \beta y_{t-3} + \epsilon_{t-2}$$

To now get

$$y_t = \alpha + \beta\alpha + \beta^2 y_{t-2} + \beta\epsilon_{t-1} + \epsilon$$
$$y_t = \alpha(1 + \beta) + \beta^2(\alpha + \beta y_{t-3} + \epsilon_{t-2}) + \beta\epsilon_{t-1} + \epsilon_t$$

From this, we can derive the equation.

$$y_t = \alpha \sum_{i=0}^{t-1} \beta^i + \beta^t y_0 + \sum_{i=0}^{t-1} \beta^i \epsilon_{t-i}$$

We assume no *unit root* so the coefficient of $|\beta| < 1$, then as t$\to \infty$, then for each of the term:

$$\beta^t y_0 \to 0$$

So the middle term disappears.

Now for the first term:

$$\alpha \sum_{i=0}^{t-1} \beta^i$$

Recall that the sum of infinite geometric series:

$$\sum_{i=0}^{\infty} x^i = \frac{1}{1-x}$$

Which means that:

$$\alpha \sum_{i=0}^{t-1} \beta^i = \alpha \frac{1}{1-\beta} = \frac{\alpha}{1-\beta}$$

First term of expression is the geometric decay weighted by alpha. For the final term, it is the same:

$$\sum_{i=0}^{t-1} \beta^i \epsilon_{t-i} = \sum_{i=0}^{\infty} \beta^i \epsilon_{t-i}$$

From this, we can get this final convergence of:

$$\frac{\alpha}{1-\beta} + \sum_{i=0}^{\infty} \beta^i \epsilon_{t-i}$$

If value of β is less than 1, we can get a geometric summation. We can take expectation of this expression, the first term is constant, so that give us 0. Expectation of β random

variable is 0. We need to make the assumption that the data is weakly stationary (which relies on the fact that the mean is constant). We can then show that:

$$E(y_t) = E[\frac{\alpha}{1-\beta} + \sum_{i=0}^{\infty} \beta^i \epsilon_{t-i}]$$

$$E(y_t) = \frac{\alpha}{1-\beta} + \sum_{i=0}^{\infty} \beta^i E(\epsilon_{t-i})$$

$$E(y_t) = \frac{\alpha}{1-\beta} = \mu$$

Variance of constant is 0 so first part disappears. For the second part, each ϵ is variance constant. Variance of the sum is the sum of the variance. Variance of each one is σ^2 multipled by β^i. We have geometric decay with respect to β together, to get the final term. We denote it as γ_0.

We have that:

$$Var(y_t) = \frac{\sigma_\epsilon^2}{1-\beta^2} = \gamma_0$$

These are both time-variant, so they do not depend on time.

We have an example for if $\beta_1 = 0.5$ for an AR(1) model. Then in t+1

$y_{t+1} = 0.5y_t = 0.5(1) = 0.5$

$y_{t+2} = 0.5y_{t+1} = 0.5(0.5) = 0.25$

And this will eventually diminish to 0 for t+h, as h $\to \infty$

If β is high, then it'll take longer for this to disappear. If β is negative, then the value will be oscilatting between positive and negative.

Autocovariance Suppose we look at first order autocorrelation whereby $\gamma_1 = \text{E}(y_t y_{t+1})$ Do what we did in assignment 2 to derive it. $\beta = \rho$ for AR(1) model. $\rho_2 = \beta^2$ and etc as we increase the lags.

Recall that: $\text{corr}(y_t, y_{t+j}) = \rho^h$ which means that for:

$$y_t = \rho y_{t-1} + \epsilon_t$$

So when $|\rho| < 1$, $\to \text{corr}(y_t, y_{t+j}) \to 0$ as h $\to \infty$. This means that AR(1) is weakly dependent and I(0).

R-squared is worthless since always goes up by adding additional parameters. Adjusted R-square helps to penalise penalty factors but still not enough penalization to help. Therefore,

we need to use information criteia (AIC/BIC). Absolute value of criteria shouldn't be used, we should look at minimum one. We can compare information criteria between each other. The first term is identical for both terms, we only care about the penalty term. AIC is better for relatively small sample. These models may suggest AR(p) different values for p.

We assume $\alpha = 0$ and $\beta = 1$ gives us a random walk. So if they equal 1, then our unconditional mean/variance doesn't hold anymore.

Weakly dependence Process says that the correlation between x_t and x_{t+h} decreases as h $\to \infty$. Therefore, a *stationary* time series process is weakly dependence if x_t and x_{t+h} are **almost independent** as h $\to \infty$. From this, a weakly dependent process is called an **integrated process of order 0 I(0)**. Weakly dependence replaces the assumption of having random variables is valid since in a large sample, since we can say that it is now equally distributed like the population. From this, we can use the time series data right away without having to make any alterations to the data.

We have an unit root iff $\rho = 1$ for an AR(1) model. When $\rho = 1$, we have an unit root series of:

$$y_t = \alpha + y_{t-1} + \epsilon_t$$

Furthermore, if $\alpha = 0$, and $\rho = 1$, then we have that y_t follows a **random walk process**. In both the case of y_t having an unit root or being a random walk, this makes it a I(1). We can take differences of both sides now:

$$y_t - y_{t-1} = \epsilon_t$$

where $\epsilon \sim iid(0, \sigma^2)$ which means ϵ is a weakly dependent process. Therefore, Δy_t is a weak dependent process (but noy y_t).

Testing for unit root see if $\beta = 1$ or not. We have an issue that in non-stationary data, t-statistic is no longer unreliable. So what we do, we subtract by t-1 on both sides and then we test the γ parameter instead. This is equivalent to testing if $\beta = 1$. We need to apply different critical values. We need to use Dickey-Fuller distribution instead now. This is only one sided since we are ever only interested is β is less than 1. If β is greater than 1, then we will have an explosive process whereby shock grows every year. As we increase number of lags in solution, we have more potential solutions. H_0 states that y_t has an unit root when we run a test. Equivalently, we can reformulate that H_0 is I(1) vs the alternative of I(0).

$$y_t = \alpha + \rho y_{t-1} + \epsilon_t$$
$$y_t - y_{t-1} = \alpha + (\rho - 1) y_{t-1} + \epsilon_t$$

Where we can define that $\theta = (\rho - 1)$ to get:

$$\Delta y_t = \alpha + (\theta) y_{t-1} + \epsilon_t$$

and specify that H_0: $\theta = 0$ vs H_a: $\theta < 0$. The null hypothesis allows us to say that $y_t \sim I(1)$ whilst alternative says that $y_t \sim I(0)$. We can no longer use t-distribution anymore, so we must now use **Dickey-Fuller** table.

Regress the new variables the usual way and compare it to the critical values. Null hypothesis is covariance stationary. If that's not the case, then we take first difference to get I(1) and test again. We can also introduced a trend variable which will change the critical values. Tables are slightly different. With this model, we could residuals that are serially correlated, so therefore we can use the Augmented Dickey-Fuller test.

5 Vector Autoregression and Error Correction

If we had the case for AR models:

$$y_t = \alpha + \beta_1 y + \beta_2 y_{t-2} + \beta_3 y_{t-3} + \epsilon_t$$

The unconditional mean would be arrived at via backward substitution. There is a shortcut to reach it by basing it on the assumption that we are dealing with stationary data. Therefore, dependent variable in the long run will have a constant mean. From this, we can come up with:

$$y^* = E(y_t) = E(y_{t-1}) = ... = E(y_{t-3})$$

Where we then arrive at y*:

$$y* = \alpha + \beta_1 y* + \beta_2 y* + \beta_3 y*$$

And then rearrange to get:

$$y* = \frac{\alpha}{1 - \beta_1 - \beta_2 - \beta_3} = \frac{\alpha}{1 - (\beta_1 + \beta + 2 + \beta_3)} = \mu$$

If β is close 1, then the time series is said to be persistent since if there is a shock, we would be seeing its effect for a long time. We would have convergence is β adds up to something that adds up to 1.

For Vector Autoregressive Process (VAR) we would need to use matrices. We assume weakly dependent time series which mean that they are not really persistent whereby any shocks disappears reasonably quickly. Since both of them are weakly dependent, they are integrated of order 0 and therefore not suffer from spurious regression. We can think of dependent variable is a function of independent variable and LAGGED of independent variable, suggesting a dynamic model. What we can do instead (for an AR(1) model in particular) is:

$$y_t = \alpha + \beta Z_t + \epsilon_t$$

and assuming that ϵ_t is first order autocorrelated so iid assumption is violated and now:

$$\epsilon_t = \rho \epsilon_{t-1} + V_t$$

Whereby $V_t \sim$ iid(0,σ). We can then reparameterize the model by first getting the lag of the model:

$$y_{t-1} = \alpha + \beta Z_{t-1} + \epsilon_{t-1}$$

multiply the lag by ρ to get:

$$\rho y_{t-1} = \rho\alpha + \rho\beta Z_{t-1} + \rho\epsilon_{t-1}$$

and then subtract the initial equation by this new equation to get:

$$y_t = \alpha(1-\rho) - \rho y_{t-1} + \beta Z_t - \rho\beta Z_{t-1} + V_t$$

Our new parameters are now:

$$y_t = \tilde{\alpha} + \tilde{\gamma} y_{t-1} + \tilde{\beta}_0 Z_t + \tilde{\beta}_1 Z_{t-1} + V_t$$

If error terms are autocorrelated, we can add lags of dependent variables in order to remove this autocorrelation. However, parameteres are now changed. So if our model should be static, we have the issue is that now the parameters we have constructed are different to the static model. We can assume no immediate effect on z on y, which means that z_t disappears. We can think of a *feedback effect* whereby z can affect y BUT then y can also affect z. We can write 2 equations affecting y and z. We have the same variables on the RHS of both equations but the parameters differ. We have a lagged y and a lagged z.

5.1 VAR

Going off from the last section, **VAR** is like an autoregressive model whereby instead of 1 dependent variable, we now have M **endogenous** dependent variables (since we now have this feedback effect). Each error term in each equation is assumed to be iid *relative* to own future and past realisation BUT may be correlated between equations. NOTE THAT ERROR TERM IN EQUATION 3 DIFFERS TO EQUATION 4. Bivariate since only 2 variables y and z and AR(1) sinceo nly 1 lag. We can have any multivariate AR of order p whereby we have m different equations.

A Π matrix is a matrix of coefficient parameters nxn matrix whilst the x matrix is a matrix of lagged variables. From this we have:

$$x_{1t} = \alpha_1 + \Pi_{11}^{(1)} + \Pi_{12}^{(1)} x_{2t-1} + ... + \Pi_{1n}^{(1)} x_{nt-1} + \Pi_{11}^{(2)} x_{1t-2} + \Pi_{12}^{(2)} x_{2t-2} + ... + \Pi_{1n}^{(2)} x_{nt-2} + \epsilon_{1t}$$

Intercept, its own lag, first lag of 2nd variable, first lag of 3rd variable and so on. Then we have the second lag of itself, then second lag of 2nd variable and so on. The superscript tells us the coefficient for that lag. Error terms are mean and constant covariance matrix. The error terms are a vector whilst the covariance is a matrix.

We can see an example of 1 lag with 2 variables.

VAR is generalises the AR model. Each equation nests an AR model. So if we set the effect from all other variables to 0, then we are left with an AR model. We can have n variables and believe that some variables affect the other variables. Therefore, the specification in reduced form gives us a model that does not impose any restrictions ahead of time.

Structural AR model requires a restriction on the right hand side. We can just use OLS but if variables are different, then we have to use MLE. To identify the system, it means that if we have 2 endogenous variables with 2 equations so direction of causality may be hard to identify so this structural model will allow us to identify the causality via restrictions made.

The number of equations is decided ahead of time of what variables to include. We need to have enough variables to ensure no omitted variable bias occurs. However, for the number of lags to have, we will apply either the AIC or BIC to determine it. We choose the number of lags with lowest information criteria. We take the natural log of the determinant of the variance-covariance matrix. The 2nd component of this is the penalty factor multiplied by number of estimated parameters and divided by the length of the time series. P is the lagged order.

We do a joint test to see if:

$$\Pi_{12}^{(1)} = \Pi_{12}^{(2)} = 0$$

to see if a given variable in a system causes another variable. This is known as the **Granger Causality** test. Rejecting the null means we have Granger causality. Granger causality differs to usual casuality. There could be no reason for one variable to cause another, but we may find that there is granger causality. This means that there is some information in a given variable that helps us to predict a realisation of variable. So Granger causality is predictive by nature and allows for our model to fit better. We assume that we have I(0) variables and locate the optimal lag lengths. Then we estimate a bivariate with order 3 and then we do a joint test. Weakly dependent ensures we don't get spurious regression. I(1) is not weakly dependent and therefore we have to take the difference to ensure they are weakly dependent.

5.2 Cointegration

Now consider if we have I(1) variables and therefore non-stationary. We can represent these 2 by random walk models. Any linear combination of these 2 variables are also an I(1) process. However, there are some cases whereby there is a linear combination of these two I(1) variables could lead to a linear equation of I(0). If this is the case, we can say that the 2 variables are **cointegrated**. Individually they are random walk processes but there is a long run relationship between the 2 variables and therefore wander around together. So there can be a deterministic trend between the 2 variables.

We could have the case that in the short run, they deviate but in the long run, they are cointegrated. This equilibrium means they converge and is denoted by β. So if we assume that $\beta = 1$, which means there is a 1-1 relationship between y and z. So thus, $y_t - z_t$ is I(0).

A generic form of cointegration can be seen as:

$$y_t - \beta Z - \alpha - \gamma t$$

whereby α and γ is set to 0.

5.2.1 Testing Cointegration

$$y_t - \eta_t = \beta Z_t$$

so LHS is I(0), then RHS MUST be I(0). We can test for the stationarity of the RHS variable.

5.2.2 Error Correction Model

γ is the adjustment parameter. The larger the γ (in absolute terms), the faster to the adjustment in the long run equilibrium. General dynamic models is when the response is a function of lags and also for feedback effects.

There is a tradeoff for exogeneity but we now include a feedback effect between the 2 variables. Variables in our new equation, we have an I(1) model and therefore bad for inferences.

5.2.3 Vector Error Correction Model

We now have first difference of variables on the left hand side.

We estimate a static model and collect the residuals. Lag those residual and include into our model of first differences. These parameters tend to be normalised wrt to β's.

To summarise, we like to see relationships between 2 or more variables. Check their order of integration. If I(0), use VAR. If not, then use VAR with differences and whether should we be carrying out a vector error correction model IFF the variables are cointegrated.

For Granger causality testing in the case of I(x) whereby x>0. Recall we can always work with first differences. If variables are I(1), then we can't use our usual test statistics. Turns out, there is a nice fix to the issue by using **Toda-Yamamoto** approach. Suppose we have 2 variables y and z. The are I(x) where x>0. Let us suppose y is I(1) and z is I(2). We would like to test the hypothesis of granger causality (and we can't use usual Var's levels). Once we identify the order of integration for each variable, we set m to the max number of integration of the 2 variables, so m=2 in this case. Then we proceed as if these 2 variabels

are I(0). We set up a VAR model VAR(P) by choosing a lag length based on AIC/BIC. Suppose we found VAR(2) is the best one, with 2 lags.

It is tempting to take difference of variables until I(0) and then find granger causality. However, this is bad. We aren't testing if a variable is affecting another variable, this just tests whether delta of variables affects delta of other variables, which may give us a different result to what we were after.

Error correction term adds more information. We can think of it as an omitted variable. If two variables are cointegrated but no error correction, then we can think of us having an omitted variable. Therefore including an error correction term allows for us to model that in now.

6 Forecasting

Can't evaluate a forecast until tomorrow happens. Information set I_t is all the information we know. E.g. If we are estimating AR(2), then previous realisations of the random variable is my information set. We use a *loss function* to help decide which model to use.

6.1 Point Forecast

The prediction we make is called a **point forecast**. It is our best case of the random variable of interest in the next period.

Forecast Uncertainty: Things we never know since some parts of the future we will never know.

Model Uncertainty: All models are wrong but some are useful. Our model will differ from the true model and therefore we will get error in our forecasts from this. However, we can minimise this. Here uncertainty, is the difference between our forecast and the true model.

Paramter Uncertainty: We estimate parameters with some uncertainty. The difference here is the difference between true model and our estimated parameters.

Any given loss function should satisfy 3 requirements:

1) If forecast error is 0, then loss function should be 0.

2) If forecast error isn't 0, then loss function should be greater than 0

3) If absolute error from 1 period is bigger than absolute error in another period, then associated loss function should also be bigger from that period.

We can optimise and solve by optimising w.r.t to the forecast. y_{t+1} is a random variable.

We can generate pseudo-forecasting environments. Tradeoff for training-test data split. More training data, in-sample error falls but then we can overfit to our training data. We want enough observations out of sample. RMSFE assumes a quadratic loss function. In addition to point forecast, we should also include interval forecasts.

Larger β means that we will have larger variances for multi-step forecasts. Normally $\beta \leq 1$. Forecast interval will initially widen and eventually stabilise at a level that is equivalent to unconditional variance of a given random variable.

For unit root, we go from an AR model into a random walk whereby y_t is a function of y_{t-1}. Our forecast function will be a flat line. A random walk cannot be forecasted

and therefore we just this time's value y_t to represent tomorrow's value. Additionally, the variance increases as the forecast horizon increases and the forecast interval will never stabilise and will always be increasing.

When comparing forecasts between Vector AR and AR models, we are testing for Granger causality and seeing does variable Z help us to also predict variable Y.