Time series analysis
Day 1.

David Pupovac

What will we do today?

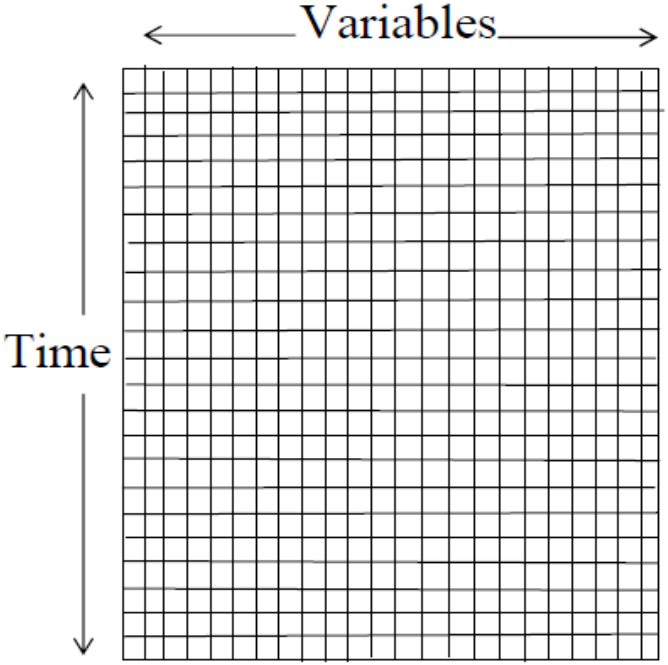
- We will introduce time series data
- We will discuss the issues related to time series from the perspective of linear regression and OLS
- Finally, we will present some methods of dealing with the problem of lack of independence due to the time series structure of data in linear regression setting

Structure of time series

- *Cross-section data* - data collected on different cases, for different variables at the same point in time
- *Time-series data* - data collected on one case, sometimes for only one variable at different points in time.

Time separating successive observations, i.e. sampling interval, should be constant (minor violations are acceptable)

Sampling interval must be short enough for the time series to provide a very close approximation to the original continuous signal



Differences between time-series and cross-sectional analysis

IN TIME SERIES:

- (Typically) fewer cases
- In cross-sectional studies, there is no natural ordering of the observations
- Lack of independence between cases, i.e., “dependent data”
- Greater sensitivity to model specification
- Greater sensitivity to “influential cases”
- Greater premium on theory
    - Theory is what tells us how to correctly specify a model, at least to begin with.
    - Theory also may tell us whether an influential case is an aberration or a critical exemplar.

Differences between time-series and cross-sectional analysis

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

vs.

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} \ldots + \varepsilon_t$$

- However, in contrast to cross-sectional analysis, with time series we can actually assess:

  (a) whether and the extent to which a variable changes and
  (b) whether and the extent to which it persists.

- Time allows for more causal leverage

Before analysis: questions about data

- What is the frequency of measurement (hourly, daily, monthly)
- Are there gaps in the data (missing values – e.g. holidays)
- What is the type of measurement (snapshots, averages - i.e. closing value of Dow Jones or average value)
- Is there adjustment method in measurement (e.g. GDP vs. GDP per capita, seasonal adjustments)
- Are the variables direct measurements or proxies
- Nature of indicators (long term unemployment vs. unemployment; self-reported unemployment vs. payroll unemployment)
- Measurement is a big deal in time series because it can change over time

Stochastic process

- Time series is a realization of a discrete-time stochastic process:

- A stochastic process

$$\{\dots, \boldsymbol{Y_1}, \boldsymbol{Y_2}, \dots, \boldsymbol{Y_t}, \boldsymbol{Y_{t+1}}\} = \{\boldsymbol{Y_t}\}_{\boldsymbol{t=-\infty}}^{\boldsymbol{\infty}}$$

is a sequence of random variables indexed by time t.

- A realization of a stochastic process with T observations is the sequence of observed data

$$\{\boldsymbol{Y_1 = y_1}, \boldsymbol{Y_2 = y_2}, \dots, \boldsymbol{Y_T = y_T}\} = \{\boldsymbol{y_t}\}_{\boldsymbol{t=1}}^{\boldsymbol{T}}$$

The goal of time series modeling is to describe the probabilistic behavior of the underlying stochastic process that is believed to have generated the observed data in a concise way.

In other words goal is detection of this data generating process by inferring from its realization.

Characteristic of the time series stochastic process

- We often describe random sampling from a population as a sequence of independent, and identically distributed (iid) random variables $\boldsymbol{Y_1, Y_2, \ldots}$ such that each $\boldsymbol{Y_i}$ is described by the same probability distribution

- With time series data, we would like to preserve the identical distribution assumption but we **do not want** to impose the restriction that each random variable in the sequence is independent from the other variables

- A characteristic of a time series stochastic process is **ergodicity**.

  – Ergodicity means that two realizations of a time series become ever closer to independence the further they are apart with respect to time

Time series and linear regression

- What does this mean for linear regression equation:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} \ldots + \varepsilon_t$$

…and distribution of y_t specifically (remember, in time series we are often concerned with modeling a single variable).

- It can mean many things and assuming different data generating mechanisms we will come up with different models.

- So, let us review linear regression

Least square method
(simple linear regression)

- Goal: minimize sum of squared errors

$$\mathrm{SSE} = \sum \varepsilon^2 = \sum (y - \hat{y})^2$$

- Define **sum of squares** (e.g. of x) and **sum of products** as:

$$SS_x = \sum (x - \bar{x})^2 \qquad SP_{xy} = \sum (x - \bar{x})(y - \bar{y})$$

*Sum of product **y** × **x** has direct relation to correlation/covariance which are defined as:*

$$cov_{xy} = \frac{SP_{xy}}{n}; \qquad cor_{xy} \frac{SP_{xy}}{\sqrt{SS_x\, SS_y}} = \frac{\sum z_x\, z_y}{n}$$

- Now, you can **obtain estimates** of: $\hat{y} = \beta_0 + \beta_1 x$:

$$\beta_1 = \frac{SP_{xy}}{SS_x} \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

Standard error of the estimate

- Standard error of the estimate gives a measure of the standard distance between regression line and the actual data points, so if:

$$\mathrm{SSE} = \sum (y - \hat{y})^2$$

standard error of the estimate is:

$$\sqrt{\frac{\mathrm{SSE}}{df}} = \sqrt{\frac{SS_y - \beta_1 SP_{xy}}{df}}$$

**df=n-2*

Testing the slope (β_1)

- $H_0: \beta_1 = 0$
- $H_1: \beta_1 \neq 0$

- The test statistic is:

$$se(\beta_1) = \frac{standard\ error\ of\ esitmate}{\sqrt{SS_x}}$$

$$t = \frac{\beta_1 - H_0\ (\beta_1)}{se(\beta_1)}$$

- Using degrees of freedom $(n-1)$, locate corresponding p value on t table. But with sufficiently large sample size (say, $n > 30$) you can just use approximate t critical value scores:

$$\pm 1.645 = 90\%;\ \pm 1.96 = 95\%;\ \pm 2.575 = 99\%$$

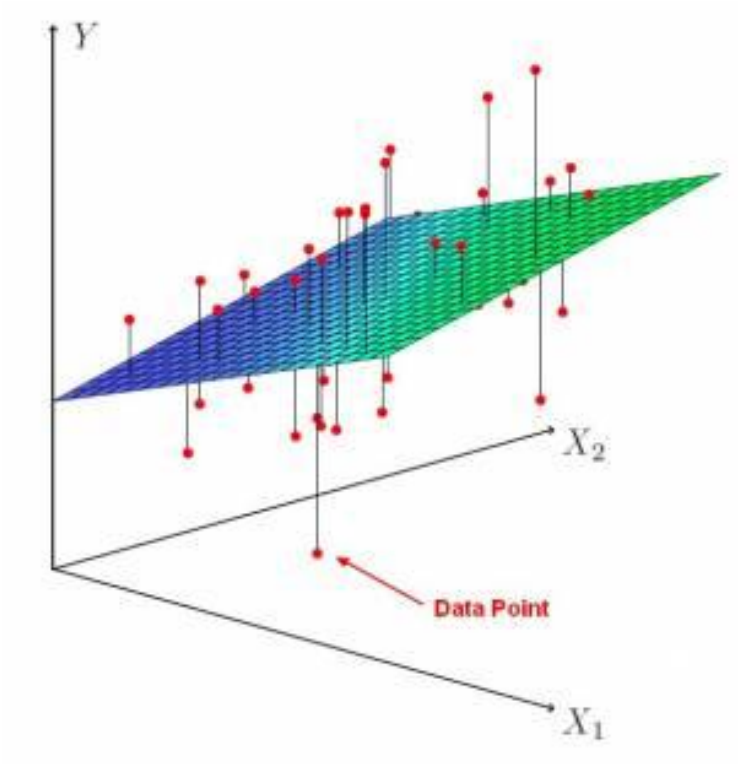
Multiple linear regression

- The multiple linear regression is straightforward extension of simple linear regression and we continue to use OLS in estimation.

- The predictors account for the effect of variable <u>while values of all other variables are kept constant (*ceteris paribus*</u>)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \varepsilon$$

- Some additional aspects:
  - Multicollinearity
  - Overfitting (use adjusted R^2)
  - Use the standardized coefficients to asses the explanatory strength of predictors
  - Extensions:
    - Dummy variables
    - Transformations/polynomials
    - Interaction terms…



OLS matrices

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & X_{11} & X_{21} & \dots & X_{k1} \\ 1 & X_{12} & X_{22} & \dots & X_{k2} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & X_{1n} & X_{2n} & \dots & X_{kn} \end{bmatrix}_{n\times k} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \vdots \\ \beta_k \end{bmatrix}_{k\times 1} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1}$$

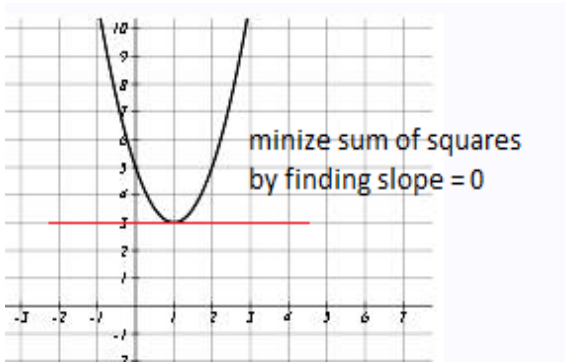
The model has a **systematic component** ($\boldsymbol{X\beta}$)

and a **stochastic component,** $\boldsymbol{\epsilon} = \boldsymbol{y} - \boldsymbol{X\beta}$.

$$\epsilon^T\epsilon = \left(y - X\hat{\beta}\right)^T \times \left(y - X\hat{\beta}\right) = \sum \epsilon^2$$

- To find the $\hat{\beta}$ that minimizes the sum of squared residuals ($\epsilon^T\epsilon$) , we need to take the partial derivative with respect to $\hat{\beta}$ equate the resulting equations to zero, and solve these equations simultaneously to obtain regression coefficients:

$$\frac{\partial \epsilon^T\epsilon}{\partial \hat{\beta}} = -2X^Ty + 2X^TX\hat{\beta} = 0$$



OLS matrices

- From the equation above we get what is called normal equations:

$$(X^T X)\hat{\beta} = X^T y$$

- If the inverse of $X^T X$ exists (i.e. $(X^T X)^{-1}$), then pre-multiplying both sides by inverse gives us:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$ *this minimizes the sum of squared residuals*

  - ***Projection P matrix:*** $X(X^T X)^{-1} X^T$ *also called influence or hat (H) matrix*
  - ***Residual maker:*** $I - X(X^T X)^{-1} X^T$

  - $(X(X^T X)^{-1} X^T)y = \hat{y}$
  - $(I - X(X^T X)^{-1} X^T)y = (I - P)y = \epsilon$

- $Var\_cor(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$ and where $\hat{\sigma}^2 = \frac{\epsilon^T \epsilon}{n-k}$
- Considering $Var\_cor(\hat{\beta}_{OLS})$ the square roots of the variances on the diagonal are **standard errors** – we will get back to this!

- **Applicability of OLS requires is dependent of certain assumptions**

Assumptions (Gauss-Markov): exogeneity - mean of error is zero

$$E[\varepsilon|X] = 0 \; \forall \; i$$

- The errors should have conditional mean zero

$$E[\varepsilon] = 0$$

- The sum of the positive errors should be equal to the sum of the negative errors. Considering residuals, this condition is trivially met because property of OLS is that the sum of residuals is zero

(By leaving the intercept term in the estimated equation you ensure that the residual term is zero-mean)

Assumptions (Gauss-Markov): exogeneity

$$E\left[\varepsilon_i x_{jk}\right] = 0 \; \forall \; i, j$$

No endogenity

- Endogenity causes bias
- Property of an OLS estimator is $X'\epsilon = 0$ – i.e. the observed values of X are uncorrelated with the residuals. For every column x_k of X, $x_k^T \epsilon = 0$. In other words, each regressor has zero sample correlation with the residuals, ϵ.

$$\begin{bmatrix} X_{11} & X_{12} & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & X_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ X_{k1} & X_{k2} & \dots & X_{kn} \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} X_{11} \times e_1 + X_{12} \times e_2 + \dots + X_{1n} \times e_n \\ X_{21} \times e_1 + X_{22} \times e_2 + \dots + X_{2n} \times e_n \\ \vdots \\ \vdots \\ X_{k1} \times e_1 + X_{k2} \times e_2 + \dots + X_{kn} \times e_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}$$

- **Note that this does not mean that in population X is uncorrelated with the disturbances ε; we have to assume this.**

Assumptions (Gauss-Markov): spherical disturbances

- In general, this will mean that uu^T, expected variance–covariance matrix of the disturbances, is $\sigma^2 I$.

- The assumption of $\sigma^2 I$ reflects the assumption of spherical disturbances:

$$\Omega = \sigma^2 I = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

  where $\hat{\sigma}^2 = \frac{\epsilon^T \epsilon}{n-k}$

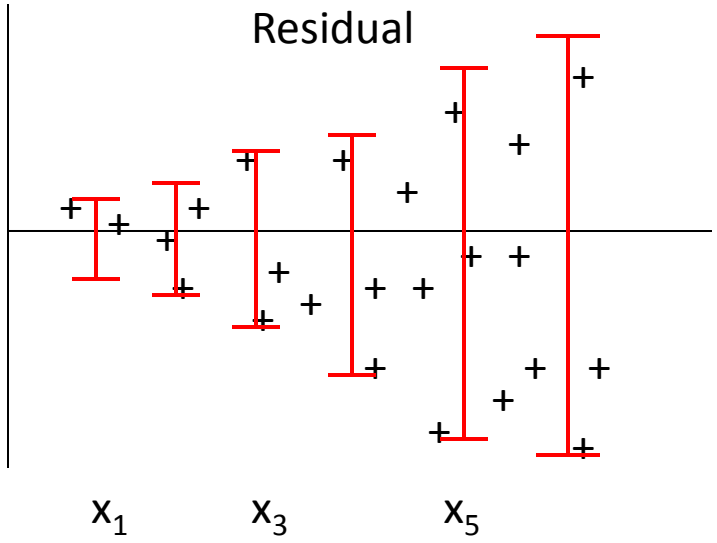
- If error variance-covariance matrix is not $\sigma^2 I$ standard errors are not estimated correctly (efficiently) – the OLS estimates are no longer BLUE

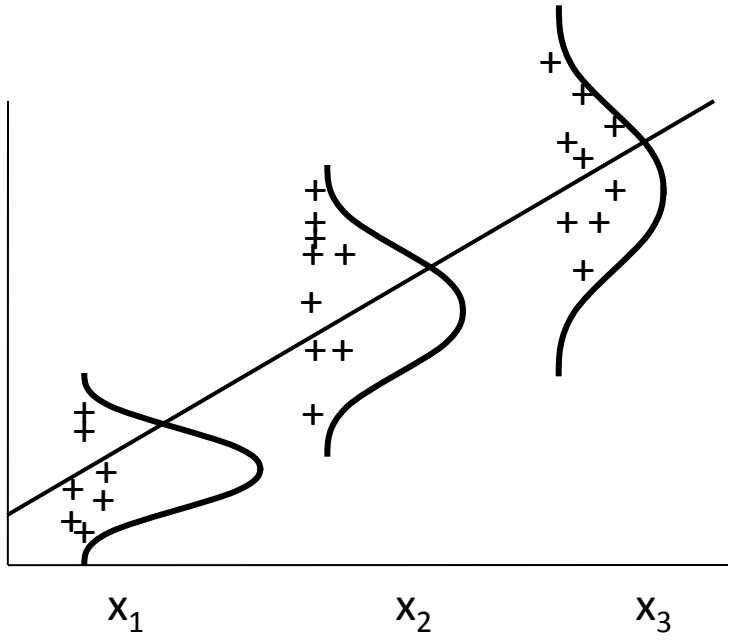
- For our purposes are this means two things:

Assumptions: constant error variance (Gauss-Markov)

1.) No heteroscedasticity

- Spherical disturbances: $E[\varepsilon_i|X] = \sigma^2 \; \forall \; i$

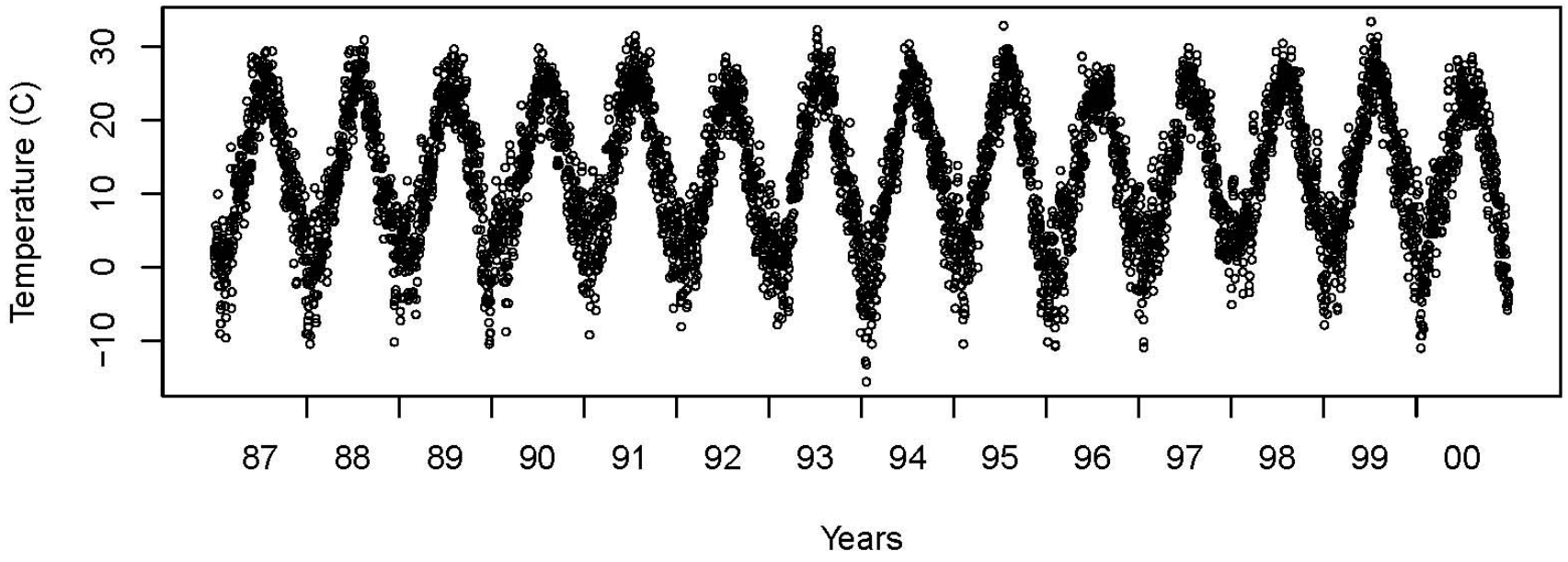




Assumptions: independent errors (Gauss-Markov)

2.) No autocorrelation

- Spherical disturbances: $E[\varepsilon_i \varepsilon_j | X] = 0 \; \forall \; i \neq j$



Assumptions (Gauss-Markov) : no multicollinearity

- $\boldsymbol{X}$ is an $n \times k$ matrix of full rank – the columns of $\boldsymbol{X}$ are linearly independent

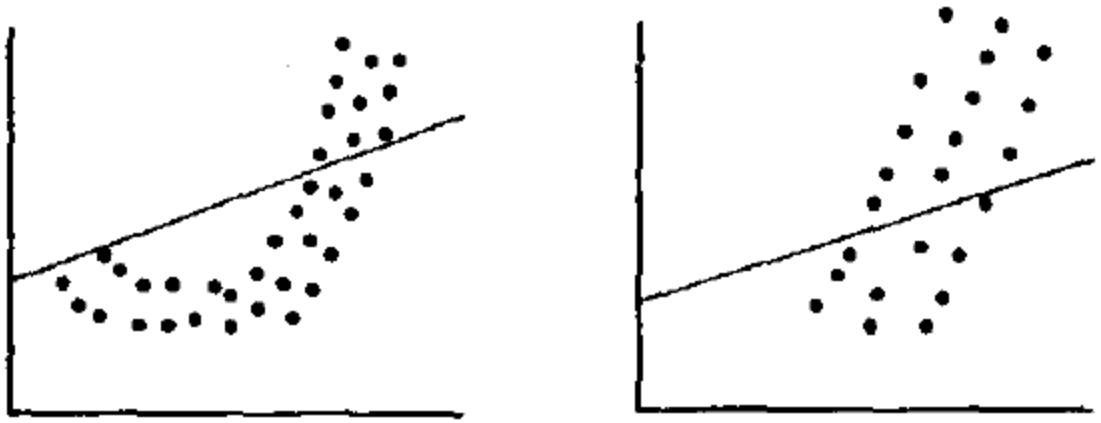
  - **Collinearity** typically arises when two variables that measure the same thing are both included in a multiple regression model.

  - It is possible for collinearity to exist between three or more variables even if no pair of variables has a particularly high correlation. We call this situation **multicollinearity**.

- Detection: VIF, correlation matrices
- Remedy: exclude the unnecessary variable, create a compose indicator, centering

Assumptions: correct model specification

- Linear relation (Gauss-Markov)



- No omitted variables (causes bias)
- No irrelevant variables

From: Lewis-Beck (1980) Applied Regression: An Introduction

Assumptions: normality

$$E[\varepsilon|X] \sim N[0, \sigma^2 I]$$

- For any given X, the distribution of errors is normal
- Also, Y is distributed normally at each value of X
- These assumption is not actually required for the Gauss-Markov Theorem. However, we often assume it to make hypothesis testing easier

Modeling time series

- Let us start with a one of most common approaches to modeling time series in form of $y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} \ldots + \varepsilon_t$

- The goal of the following models is not to model dynamics but to treat it as a nuisance

- In general, these models are recommended when we do not care about dynamic aspects of a time series, when we want to account for remaining autocorrelation in residual, and when the dependence in not large in magnitude.

Basic multivariate analysis

- Autocorrelation (serial correlation): correlation of past values of random variables with its current values

  (correlation in observed values and correlation in errors)

- $E\left[\varepsilon_i \varepsilon_j \middle| X\right] = 0 \; \forall \; i \neq j$ does not hold

- In the face of autocorrelation OLS estimates are not unbiased or inconsistent but they are inefficient. In other words, the standard error is are not correctly estimated.

- Serial correlation may be caused by a specification errors such as:
  - an omitted variable and/or
  - an incorrect functional form

Lagged variables

- Correlation of past values of a variable with its current values is assessed by lags. Let's see what are lags:

$time$	x	x_{t-1}	x_{t-2}
1	0,849468		
2	0,162461	0,849468	
3	0,172593	0,162461	0,849468
4	0,562669	0,172593	0,162461
5	0,639223	0,562669	0,172593
6	0,110537	0,639223	0,562669
7	0,725579	0,110537	0,639223
8	0,302856	0,725579	0,110537
9	0,342341	0,302856	0,725579
10	0,505434	0,342341	0,302856

Defining error autocorrelation

- From $y_t = \beta_0 + \beta_1 x_{1t} + \cdots + \varepsilon_t$ the first-order autocorrelation can be defined as: $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$

  - where: ε_t = the error term of the equation in question
  - ρ = the first-order autocorrelation coefficient
  - u = a classical (not serially correlated) error term

  - Order reflects the number of lags
    - The most commonly assumed kind of serial correlation is first-order serial correlation

- In order to determine mean and variance of ε_t we substitute ε_{t-1} with $\rho\varepsilon_{t-2} + u_{t-1}$:

$$\varepsilon_t = \rho(\rho\varepsilon_{t-2} + u_{t-1}) + u_t = u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \cdots$$ geometrical decay

- Variance $E[\varepsilon_t^2] = \frac{\sigma_u^2}{1-\rho^2}$

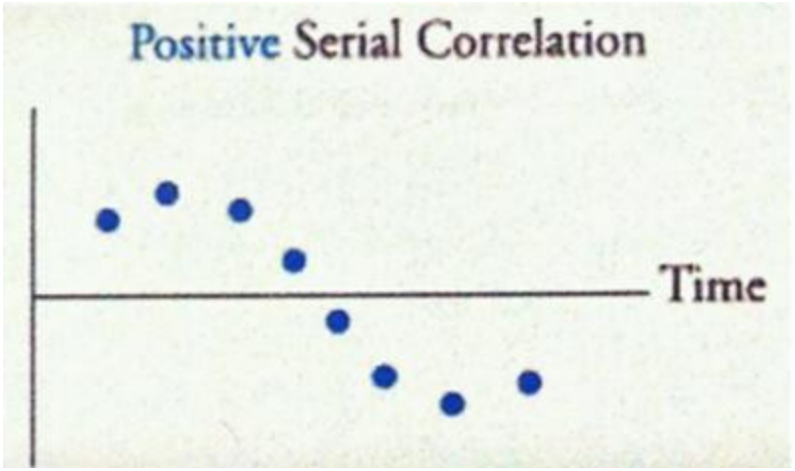
Autocorrelation

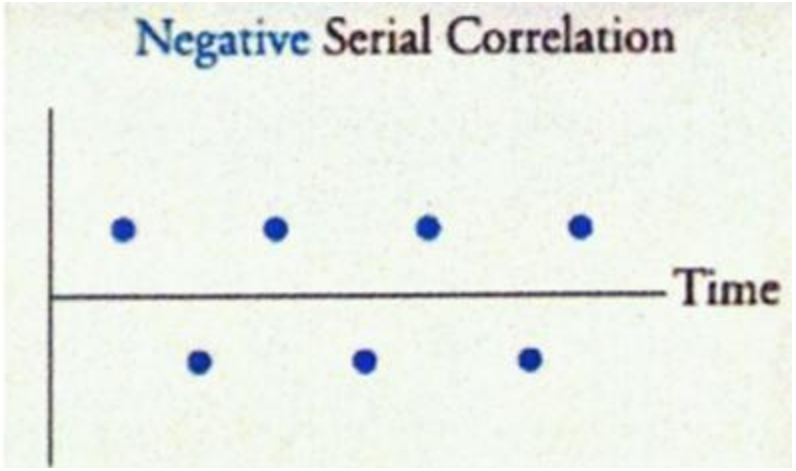
- **The magnitude of ρ indicates the strength of the serial correlation:**
  - If ρ is zero, there is no serial correlation
  - As ρ approaches one in absolute value, the previous observation of the error term becomes more important in determining the current value of ε_t and a high degree of serial correlation exists
- **Autocorrelation ranges:**

$$\mathbf{-1 < \rho < +1}$$

- **Positive**:
  - implies that the error term tends to have the **same sign** from one time period to the next
- **Negative**:
  - implies that the error term has a tendency to **switch signs** from negative to positive and back again in consecutive observations

Positive and negative autocorrelation





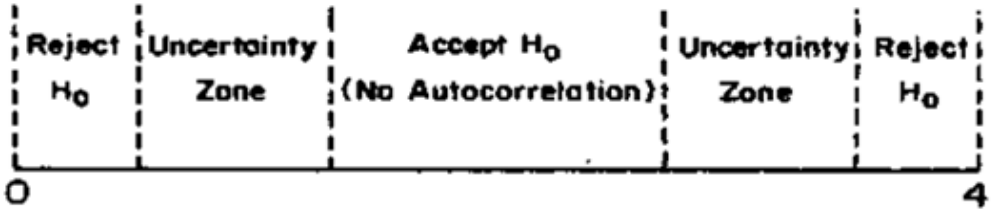
Positive autocorrelation; $\rho > 0$ - very common and consequential, understates standard errors - there will be a tendency to reject the null hypothesis when it should not be rejected.

Negative autocorrelation; $\rho < 0$ - less common and less consequential.

Detection of autocorrelation

- Visualization (plot residuals)
- **Statistical tests:**
- No dependent lags - **Durbin Watson d**: $d = \frac{\sum_t (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_t \varepsilon_t^2}$

The test statistics d runs from 0 to 4. No autocorrelation is about 2.0, let's say 1.4-2.6 (in smaller samples $N = 25$ just below 2.0) – positive autocorrelation if closer to 0, negative autocorrelation if closer to 4. The critical values vary by level of significance, the number of observations, and the number of predictors in the regression equation – so there is Durbin-Watson significance tables.



- Durbin Watson h - Durbin's alternative test, can have dependent lags; it is distributed asymptotically as χ^2 distribution

Detection of autocorrelation

- **Portmanteau test (Ljung-Box)**

$$Q^* = n(n+2)\sum_{k=1}^{h}\frac{r_k^2}{(n-k)}$$

where n is the length of the residual time series, r_k is the kth autocorrelation coefficient of the residuals, and h is the number of lags to test. Large values of Q* indicate that there are significant autocorrelation

- Durbin-Watson and portmanteau test with lagged dependent variables in the regressor matrix are biased in favor of maintaining the null hypothesis of no-autocorrelation

- **Breusch-Godfrey test** – allows for multiple lags

- Breusch-Godfrey is a Lagrange Multiplier (LM) test where we regress residual on X variables and lagged residual(s) and test for significance of latter. For testing we can use either F test or LM statistics and χ^2 distribution

Strategies in dealing with serial correlation in regression models

- There are two common strategies:

  1. Use feasible generalized least squares

  2. Use OLS to estimate the regression coefficients, but use a corrected estimator for the variance-covariance estimator

(there are other additional methods proposed in the literature such as transformation of OLS or instrumental variable estimator)

Generalized least squares

- GLS is an efficient estimator in presence of autocorrelation.
- GLS transforms the model in order to remove heteroskedasticity/autocorrelation from the estimates. It requires the knowledge of the structure of heteroskedasticity or autocorrelation in order to propose the transformed model
- Start with OLS estimation

$$\boldsymbol{Y_t = \beta_0 + \beta_1 X_{1t} + \epsilon_t}$$

where $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$ (due to serial correlation),

- Then lag the new equation by one period and multiply by ρ, obtaining:

$$\boldsymbol{\rho Y_{t-1} = \rho\beta_0 + \rho\beta_1 X_{1t-1} + \rho\epsilon_{t-1}}$$

- Next, subtract the later equation from the former:

$$\boldsymbol{Y_t - \rho Y_{t-1} = \beta_0(1-\rho) + \beta_1(X_{1t} - \rho X_{1t-1}) + u_t}$$

- GLS is only theoretical, we do not typically know the value of $\boldsymbol{\rho}$ - that is why we use **feasible/estimated GLS**

(model cannot contain lagged dependent variable)

Feasible generalized least squares

Still pretty common in many circumstances, particularly where we have small $\boldsymbol{n}$ and dynamics are not of primary interest or concern

- ***Cochrane-Orcutt feasible GLS:***

1. Regress Y on X
2. Generate residuals, i.e., $\epsilon = Y - \beta_0 - \beta_1 X$
3. Regress current residuals on lagged residuals to produce estimate of $\boldsymbol{\rho}$; $\epsilon_t = \rho\epsilon_{t-1} + u$
4. Transform Y and X using $\boldsymbol{\rho}$: $Y_t^* = Y_t - \rho Y_{t-1}$ and $X_t^* = X_t - \rho X_{t-1}$
5. Regress Y_t^* on X_t^*
6. Return to 2 and continue to cycle until convergence (there is little change in $\boldsymbol{\rho}$)

Prais-Winsten approach reintroduces first case with imputation

However , the trend in the economics literature has been not to use GLS to correct for autocorrelation or heteroskedasticity, and instead correct the standard errors directly

Corrected the standard errors

- General procedure:
  - Use OLS to estimate regression coefficients
  - Correct standard errors

- The square roots of the variances on the diagonal of $Var\_cor(\hat{\beta}_{OLS})$ are **standard errors**, where:

$$Var\_cor(\hat{\beta}) = \sigma^2(X^TX)^{-1} \text{ and } \hat{\sigma}^2 = \frac{\epsilon^T\epsilon}{n-k}.$$

- This formula depends on the assumption of spherical disturbances, that is that variance covariance matrix $\Omega = \sigma^2 I$.

$$\Omega = \sigma^2 I = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

- **Heteroscedasticity** - diagonal elements of matrix not are equal
- **Autocorrelation** - off-diagonal elements of matrix are not zero

The variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$

- How do we come to: $Var\_cor(\hat{\beta}) = \sigma^2(X^TX)^{-1}$
- If **<u>population</u>** parameters are β and their respective disturbances are u, than:

$$Var\_cor(\hat{\beta}) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T]$$

- From $\hat{\beta} = (X^TX)^{-1}X^Ty$, by substituting y with $y = X\beta + u$ we get:

$$\hat{\beta} = (X^TX)^{-1}X^T(\boldsymbol{X\beta} + \boldsymbol{u}) = \beta + (X^TX)^{-1}(X^Tu)$$

- By substituting $\hat{\beta}$ with $\beta + (X^TX)^{-1}(X^Tu)$ in $E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T]$ we get:

$$E[(X^TX)^{-1}(X^Tu)((X^TX)^{-1}(X^Tu))^T]$$

- With further simplification we get: $E[(X^TX)^{-1}\boldsymbol{X^Tuu^TX}(X^TX)^{-1}]$
- Now if assume that variance-covariance matrix of disturbances is:

$$E[uu^T] = \boldsymbol{\Omega = \sigma^2 I} \quad \text{we get:}$$

- $Var\_cov(\hat{\beta}) = (X^TX)^{-1}X^T\boldsymbol{\sigma^2 I}X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$

The variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$

- $Var(\hat{\beta}) = \sigma^2(X^TX)^{-1}$ $\quad$ $\boldsymbol{is\ (k \times k) matrix; k - \ includes\ intercept}$

$$Var\_cov(\hat{\beta}) = \begin{bmatrix} var(\hat{\beta}_0) & cov(\hat{\beta}_0, \hat{\beta}_1) & \dots & cov(\hat{\beta}_0, \hat{\beta}_k) \\ cov(\hat{\beta}_1, \hat{\beta}_0) & var(\hat{\beta}_1) & \dots & cov(\hat{\beta}_1, \hat{\beta}_k) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\hat{\beta}_k, \hat{\beta}_0) & cov(\hat{\beta}_k, \hat{\beta}_1) & \dots & var(\hat{\beta}_k) \end{bmatrix}$$

- However, formula $Var\_cor(\hat{\beta}) = \sigma^2(X^TX)^{-1}$ will no longer produce efficient standard errors in the presence of autocorrelation as it cannot be assumed that variance-covariance matrix of disturbances is $\Omega = \sigma^2 I$ i.e.:

$$E[\varepsilon\varepsilon^T|X] = \sigma^2 I$$

- *For heteroscedasticity you can use Eicker-Huber-**White** robust sandwich errors, but this requires independent disturbances (no autocorrelation). Instead we use **Newey-West** standard errors*

White standard errors

- White showed that $X^T \epsilon \epsilon^T X$ is a good estimator of the corresponding expectation term $X^T u u^T X$

- Various heteroskedasticity consistent estimators have been suggested which are constructed by plugging an estimate of type $\widehat{\Omega} = diag(\epsilon_1, \ldots, \epsilon_n)$ into central term (i.e. $X\widehat{\Omega}X_W$) of equation $(X^T X)^{-1} X^T \Omega \, X \, (X^T X)^{-1}$

- Thus, with White standard errors we allowed:

$$\begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_m^2 \end{bmatrix}$$

White and Newey-West standard errors

- If the error terms are not independent, *Ω is not diagonal*. This implies that we have autocorrelation (notice that we still may have heteroskedasticity). Thus, we should allow:

$$\begin{bmatrix} \sigma^2 & \sigma_{1,2} & \cdots & \sigma_{1,m} \\ \sigma_{2,1} & \sigma^2 & \cdots & \sigma_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m,1} & \sigma_{m,2} & \cdots & \sigma^2 \end{bmatrix}$$

- If following White estimator procedure, we allow for off diagonal elements center term in $(X^T X)^{-1} X^T \Omega\, X\, (X^T X)^{-1}$ to be:

$$X^T \Omega\, X = (X^T \hat{u})(\hat{u}^T X) \quad where\ \epsilon = \hat{u}$$

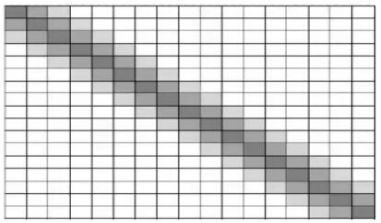
- with $(X^T \hat{u})$ we are calculating if the observed values are related to residuals, **which is always equal to zero.** So our estimates of variances of $\hat{\beta}$ would be zero, so this is useless.

Newey-West standard errors

- Instead we introduce ω weights,

$$\begin{pmatrix} \hat{u}_1^2 & w_1\hat{u}_1\hat{u}_2 & w_2\hat{u}_1\hat{u}_3 & \cdots & \cdots & \cdots \\ w_1\hat{u}_1\hat{u}_2 & \hat{u}_2^2 & w_1\hat{u}_2\hat{u}_3 & w_2\hat{u}_2\hat{u}_4 & \ddots & \vdots \\ w_2\hat{u}_1\hat{u}_3 & w_1\hat{u}_2\hat{u}_3 & \hat{u}_3^2 & w_1\hat{u}_3\hat{u}_4 & \ddots & \vdots \\ \vdots & w_2\hat{u}_2\hat{u}_4 & w_1\hat{u}_3\hat{u}_4 & \hat{u}_4^2 & \ddots & w_2\hat{u}_{n-2}\hat{u}_n \\ & \ddots & \ddots & \ddots & \ddots & w_1\hat{u}_{n-1}\hat{u}_n \\ \cdots & \cdots & \cdots & & & \hat{u}_n^2 \end{pmatrix}$$

- The weights are decreasing with the increased distance in time $(\boldsymbol{\omega_1 > \omega_2 > \omega_3})$ approximately following the schema:



We estimate Newey-West by including White estimator, $X\widehat{\Omega}X_W$, into $X^T\Omega\, X_{NW}$, and where $\left(1 - \frac{l}{L+1}\right)$ are linearly decaying weights

(variations of Newey-West's approach will typically differ in the definition of weights)

- $X^T\Omega\, X_{NW} = X\widehat{\Omega}X_W + \frac{n}{n-k}\sum_{j=1}^{m}\left(1 - \frac{l}{L+1}\right)\sum_{t=l+1}^{n}\epsilon_t\epsilon_{t-l}\,(x_t^T x_{t-l} + x_{t-l}^T x_t)$

Final words

- Neither GLS nor models with Newey-West standard errors can contain lagged dependent variables.

- The assumption of autocorrelation in residuals implies the correlation between lagged dependent variable and residuals.

- This correlation between a regressor and residual violates the assumption of the exogeneity.