Topic 5
Violations of Classical Linear Regression

Learning objectives

- Explain the problems of econometrics
  - Assumption violations
    1. Multicollinearity
    2. Heteroscedasticity
    3. Autocorrelation
  - Terminology
  - Sources
  - Detection
  - Consequences
  - Remedies

Assumption Violation: Multicollinearity

- Multicollinearity exists when at least some of the predictor variables are correlated among themselves so that there is a linear relationship between two or more variables.
- Suppose there are three variables: X_1, X_2, X_3, the variables are linearly related if: $X_1 = a_1 X_2 + a_2 X_3$
- Three cases can be distinguished
  - ***Perfect Multicollinearity***
    - Perfect linear relationship among the variables
    - Usually introduced into a problem by accident
  - ***Near-Perfect Multicollinearity***
    - Almost perfect linear relationship
    - Typical of economic data
  - ***No Multicollinearity***
    - No linear relationship at all

Sources of multicollinearity

1. Data collection method e.g. sampling over a limited range of the values taken by the regressors in the population.

2. Constraints on the model or in the population being sampled e.g. in a regression of electricity consumption (X_1), on income (X_2) and house size (X_3); high X_2 implies X_3

3. Model specification e.g. adding polynomial terms to a regression model when the range of the X variable is small.

4. An overdetermined model i.e. when a model has more explanatory variables than the number of observations.

5. In a time series model, there may be regressors that share a common trend i.e. they all increase or decrease over time.

Detection of multicollinearity

1. Independent variable(s) considered critical in explaining the model's dependent variable are not statistically significant according to the tests.
2. High R^2, highly significant F-test, but few or no statistically significant t-tests
3. Parameter estimates drastically change values and become statistically significant when excluding some independent variables from the regression
4. A simple test for multicollinearity is to conduct "artificial" regressions between each independent variable (as the "dependent" variable) and the remaining independent variables
5. Variance Inflation Factors (VIF_j) are calculated as: $VIF_j = \frac{1}{(1 - R_j^2)}$
   - $VIF_j = 2$, for example, means that variance is twice what it would be if X_j, was not affected by multicollinearity
   - A $VIF_j > 10$ is clear evidence that the estimation of B_j is being affected by multicollinearity

Consequences of multicollinearity

1. OLS estimators will have large variance and covariances making precise estimated difficult
2. Because of large variances CIs are wider
3. Because of large variances t ratio are statistically insignificant
4. R^2 can be very high
5. OLS estimators and standard errors can be sensitive to small changes in the data.

Remedies of multicollinearity

1. Do nothing; if it is a data deficiency problem we increase the sample size with additional or new data.
2. Combine cross-sectional and time series data (pooling the data)
3. Drop a variable(s): exclude the independent variables that appear to be causing the problem
4. Transformation of variables e.g. choose log functional form.

Assumption Violation: Heteroskedasticity

- OLS makes the assumption that $V_j(\varepsilon) = \sigma^2$ *for all* $_j$ *i.e.*, the variance of the error term is constant (Homoskedasticity). If the error terms do not have constant variance, they are said to be heteroskedastic.
  - Hetero (different or unequal) is the opposite of Homo (same or equal)
  - Skedastic means spread or scatter
  - Homoskedasticity = equal spread
  - Heteroskedasticity = unequal spread
- Assume that in the two-variable model $Y_i = \beta_1 + \beta_2 X_i + u_i$, Y represents savings and X represents income. Figures 11.1 and 11.2 show that as income increases, savings on the average also increase. But in Figure 11.1 the variance of savings remains the same at all levels of income, whereas in Figure 11.2 it increases with income. It seems that in Figure 11.2 the higher income families on the average save more than the lower-income families, but there is also more variability in their savings.

Homeskedastic disturbances

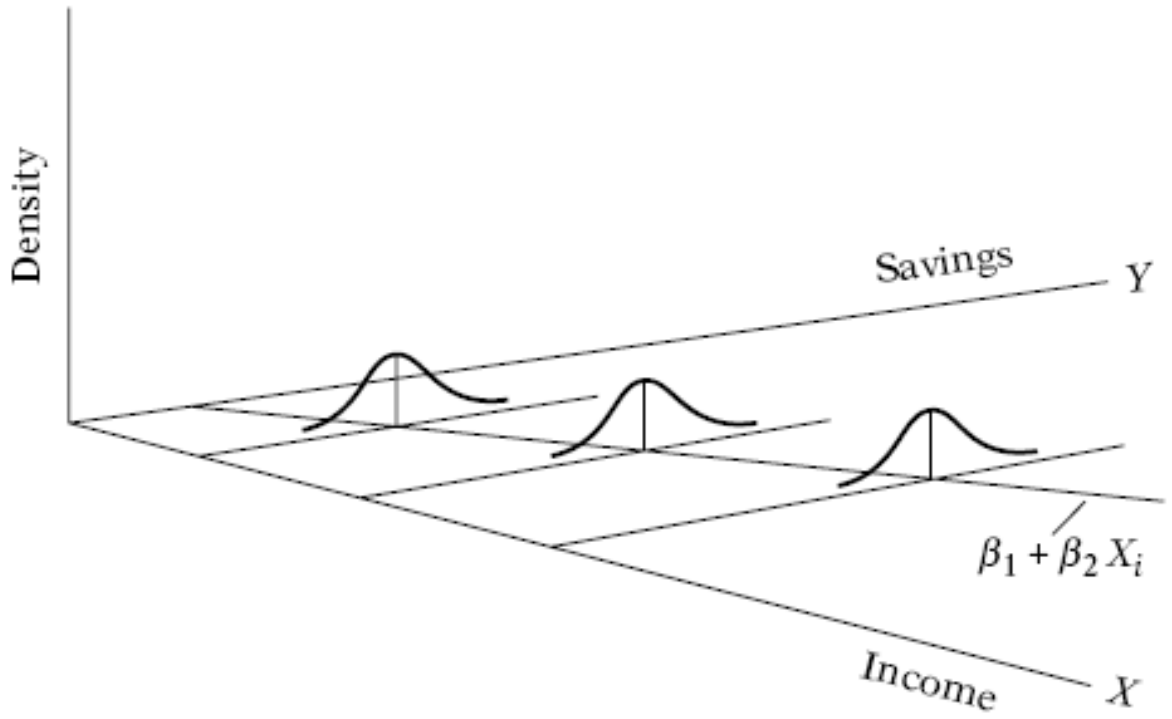


FIGURE 11.1 Homoscedastic disturbances.

Heteroskedastic disturbances

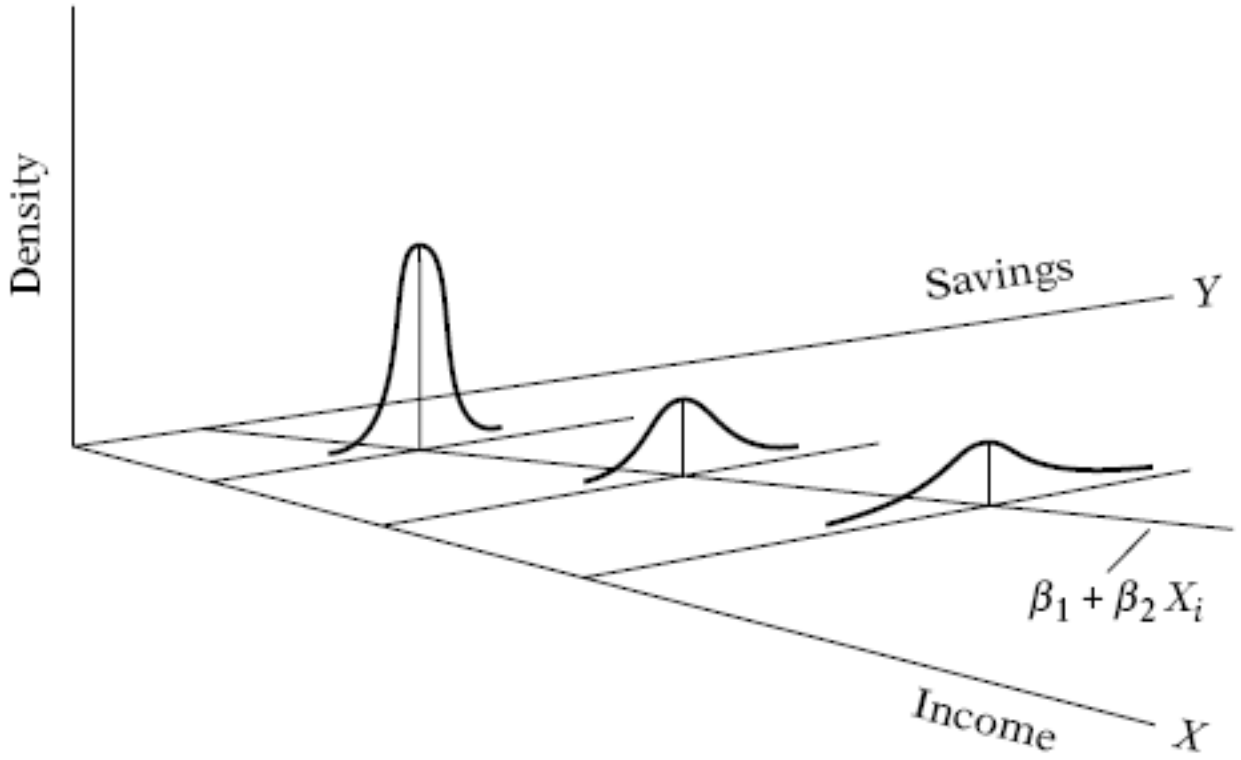


FIGURE 11.2 Heteroscedastic disturbances.

Sources of heteroskedasticity

1. Following the error-learning models, as people learn, their errors of behavior become smaller over time. In this case, σ^2_i is expected to decrease.
2. As incomes grow, people have more discretionary income and hence more scope for choice about the disposition of their income. Hence, σ^2_i is likely to increase with income. Similarly, companies with larger profits are generally expected to show greater variability in their dividend policies than companies with lower profits.
3. As data collecting techniques improve, σ^2_i is likely to decrease. Thus, banks that have sophisticated data processing equipment are likely to commit fewer errors in the monthly or quarterly statements of their customers than banks without such facilities.
4. Heteroskedasticity can also arise as a result of the presence of outliers, (either very small or very large) in relation to the observations in the sample. The inclusion or exclusion of such an observation, especially if the sample size is small, can substantially alter the results of regression analysis.

Sources of heteroskedasticity

5. Another source of heteroskedasticity arises from violating Assumption 9 of CLRM, namely, that the regression model is correctly specified, very often what looks like heteroskedasticity may be due to the fact that some important variables are omitted from the model. But if the omitted variables are included in the model, that impression may disappear.
6. Another source of heteroskedasticity is skewness in the distribution of one or more regressors included in the model. Examples are economic variables such as income, wealth, and education. It is well known that the distribution of income and wealth in most societies is uneven, with the bulk of the income and wealth being owned by a few at the top.
7. Other sources of heteroskedasticity: As David Hendry notes, heteroskedasticity can also arise because of
   - incorrect data transformation (e.g., ratio or first difference transformations)
   - incorrect functional form (e.g., linear versus log–linear models).

Sources of heteroskedasticity

- Note that the problem of heteroscedasticity is likely to be more common in cross-sectional than in time series data. In cross-sectional data, members may be of different sizes, such as small, medium, or large firms or low, medium, or high income. In time series data, on the other hand, the variables tend to be of similar orders of magnitude. Examples are GNP, consumption expenditure, savings.

Detection of heteroskedasticity

Informal methods

1. Nature of the Problem: Very often the nature of the problem under consideration suggests whether heteroskedasticity is likely to be encountered. In cross-sectional data involving heterogeneous units, heteroskedasticity may be the rule rather than the exception. Thus, in a cross-sectional analysis involving the investment expenditure in relation to sales, rate of interest, etc., heteroskedasticity is generally expected if small-, medium-, and large-size firms are sampled together.
2. Graphical inspection of residual (scatter diagrams): If there is no a priori or empirical information about the nature of heteroscedasticity, in practice one can do the regression analysis on the assumption that there is no heteroscedasticity and then do an examination of the residual squared $u\hat{}^2_i$ to see if they exhibit any systematic pattern.

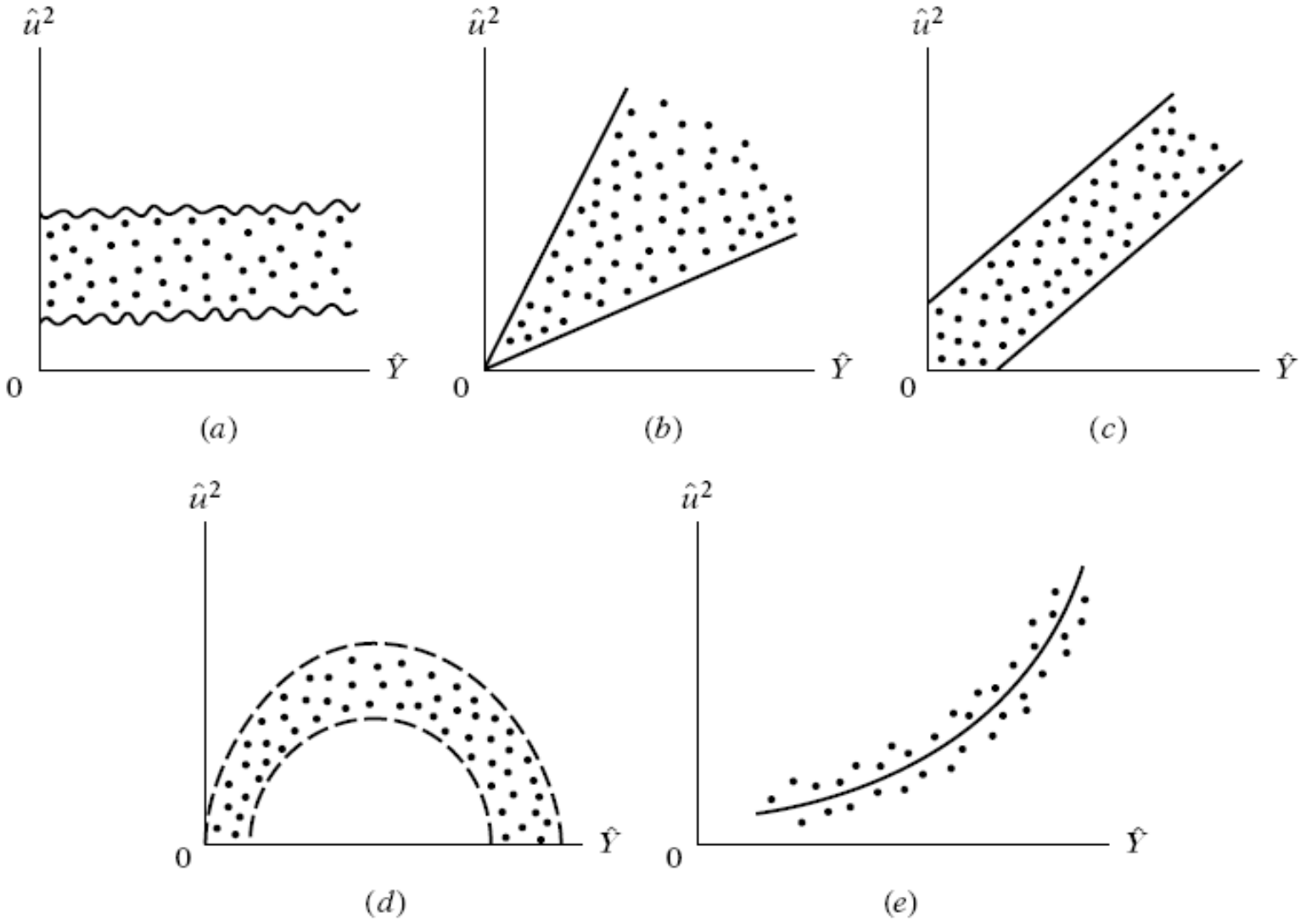


FIGURE 11.8 Hypothetical patterns of estimated squared residuals.

Detection of heteroskedasticity

Formal methods

1. White's Test
   - Estimate the regression using OLS and obtain the residuals.
   - Regress the squared residuals (as the dependent variables) on all the X variables, and all squared values of the X variables. Obtain R^2 from this regression.
   - Under the null hypothesis that there is no heteroscedasticity, it can be shown that sample size (n) times the R^2 obtained from the auxiliary regression asymptotically follows the chi-square distribution with df equal to the number of regressors (excluding the constant term) in the auxiliary regression.)
   - We can again use either the *F or LM* statistic to test the following hypothesis for homoskedasticity, $H0 : \delta_1 = \delta_2 = 0$
   - The Lagrange Multiplier (LM) test statistics, $LM = n * R^2$
   - If the chi-square value obtained exceeds the critical chi-square value at the chosen level of significance, the conclusion is that there is heteroskedasticity. If it does not exceed the critical chi-square value, there is no heteroskedasticity

Consequences of heteroskedasticity

- Assuming all other assumptions are in place, the assumption guaranteeing unbiasedness of OLS is not violated. Consequently OLS is unbiased in this model
- However the assumptions required to prove that OLS is efficient are violated. Hence OLS is not BLUE in this context
- We can devise an efficient estimator by reweighing the data appropriately to take into account heteroskedasticity
- If there is heteroskedasticity in our data and we ignore it then the standard errors of our estimates will be incorrect
- However, if all the other assumptions hold our estimates will still be unbiased.
- Since the standard errors are incorrect inference may be misleading

Remedies for heteroskedasticity

1. Generalized Least Squares / Weighted Least Squares
   - In a perfect world, we would actually know what heteroskedasticity we could expect—and we would then use 'weighted least squares'.
   - WLS essentially transforms the entire equation by dividing through every part of the equation with the square root of whatever it is that one thinks the variance is related to.
   - In other words, if one thinks one's variance of the error terms is related to $X1^2$, then one divides through every element of the equation (intercept, each bx, residual) by X1.
   - In this way, one creates a transformed equation, where the variance of the error term is now constant (because you've "weighted" it appropriately).
2. Where the error variance is unknown, remedies based on assumptions about the error variance are used but they amount to WLS
3. Respecification of the model using a different functional form e.g. log transformation.

Assumption Violation: Autocorrelation

- Autocorrelation also known as serial correlation occurs in time-series studies when the errors associated with a given time period carry over into future time periods.
- For example, if we are predicting the growth of stock dividends, an overestimate in one year is likely to lead to overestimates in succeeding years.
- In situation like this, the assumption of no auto or serial correlation in the error term that underlies the CLRM will be violated.
- Error term is correlated with itself (serial correlation):

  $Cov(e_i, e_j) \neq E(e_i e_j) \neq 0 \quad i \neq j$

Sources of autocorrelation

1. Inertia - Macroeconomics data experience cycles/business cycles.
2. Specification bias- excluded variable case

   Appropriate equation: $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + u_t$

   Estimated equation: $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + v_t$

   Estimating the second equation implies: $v_t = \beta_4 X_{4t} + u_t$

3. Specification bias- incorrect functional form

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{2t}^2 + v_t$$

$$Y_t = \beta_1 + \beta_2 X_{2t} + u_t$$

$$u_t = \beta_3 X_{2t}^2 + v_t$$

Sources of autocorrelation

4. Cobweb Phenomenon

   In agricultural market, the supply reacts to price with a lag of one time period because supply decisions take time to implement. This is known as the cobweb phenomenon.

   Thus, at the beginning of this year's planting of crops, farmers are influenced by the price prevailing last year.

5. Lags

$$Consumption_t = \beta_1 + \beta_2 Consumption_{t-1} + u_t$$

   The above equation is known as autoregression because one of the explanatory variables is the lagged value of the dependent variable.

   If you neglect the lagged the resulting error term will reflect a systematic pattern due to the influence of lagged consumption on current consumption.

Sources of autocorrelation

6. Data manipulation

$$Y_t = \beta_1 + \beta_2 X_t + u_t \qquad Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + u_{t-1}$$

$$\Delta Y_t = \beta_2 \Delta X_t + v_t$$

This equation is known as the first difference form and dynamic regression model. The previous equation is known as the level form.

Note that the error term in the first equation is not autocorrelated but it can be shown that the error term in the first difference form is autocorrelated.

7. Nonstationarity

When dealing with time series data, we should check whether the given time series is stationary.

A time series is stationary if its characteristics (e.g. mean, variance and covariance) are constant over time (time invariant) that is, they do not change over time.

If that is not the case, we have a nonstationary time series.

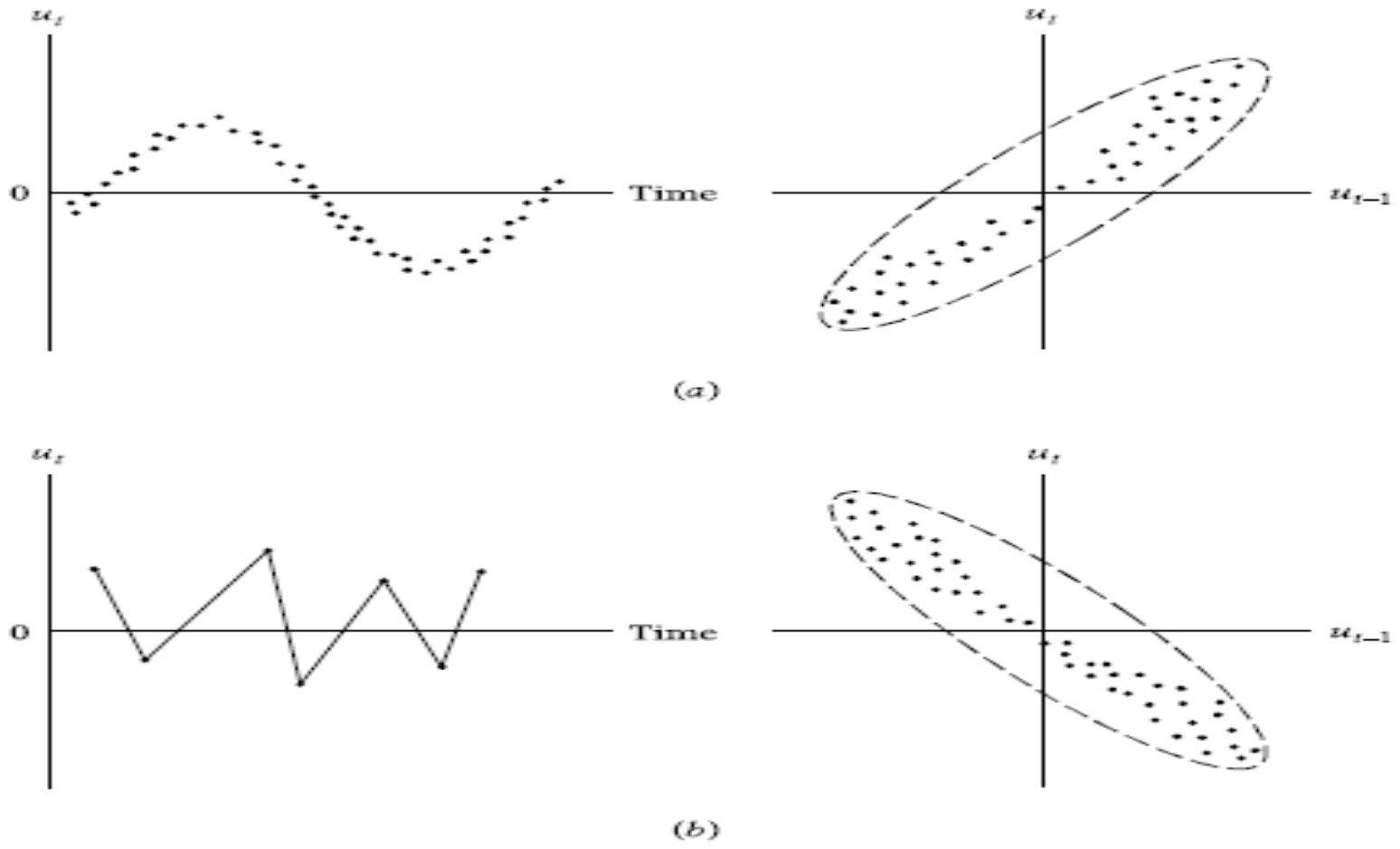
Detecting autocorrelation

Informal methods

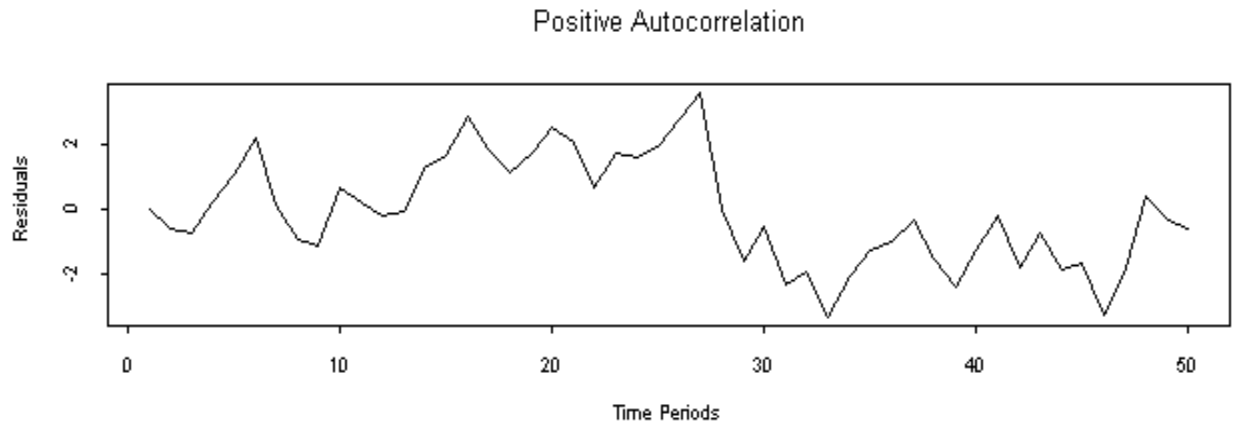
Graphical Method

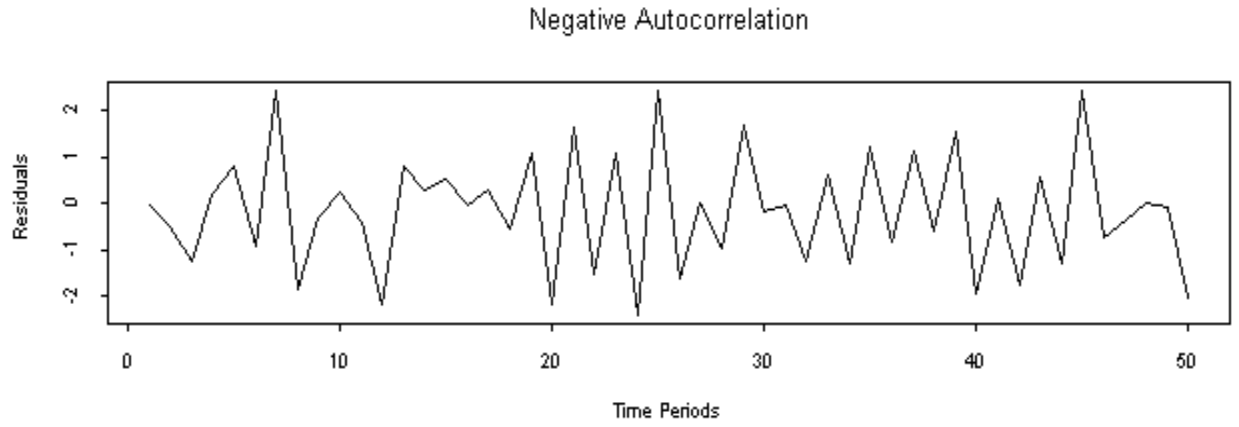
- Plot residuals against time where residuals are estimates of disturbance term; Can highlight violations; Look for nonrandom patterns
- We can also plot ordinary residuals against lagged ordinary residuals
- If positive autocorrelation exists
  - Residuals will follow a sine wave-type
  - Negative residuals tend to be followed by negative residuals while positive residuals tend to be followed by positive residuals
  - Any jaggedness due to random white noise
- If negative autocorrelation exists
  - then negative numbers are followed immediately by positive numbers in almost all cases
  - Any jaggedness due to white noise

Detecting autocorrelation



(a) Positive and (b) negative autocorrelation.





Typical Residual Patterns These two panels show typical residual patterns under positive and negative autocorrelation.

Notice how the positively autocorrelated series is smoother and more sine wave-like than the negatively autocorrelated series.

Detecting autocorrelation

Formal methods

1. The Durbin Watson Test

$$d = \frac{\sum_{t=2}^{t=n} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{t=n} \hat{u}_t^2}$$

H_0 : $\rho = 0$ i.e the u's are not auto correlated

H_a: $\rho \neq 0$ i.e. the u's are auto correlated

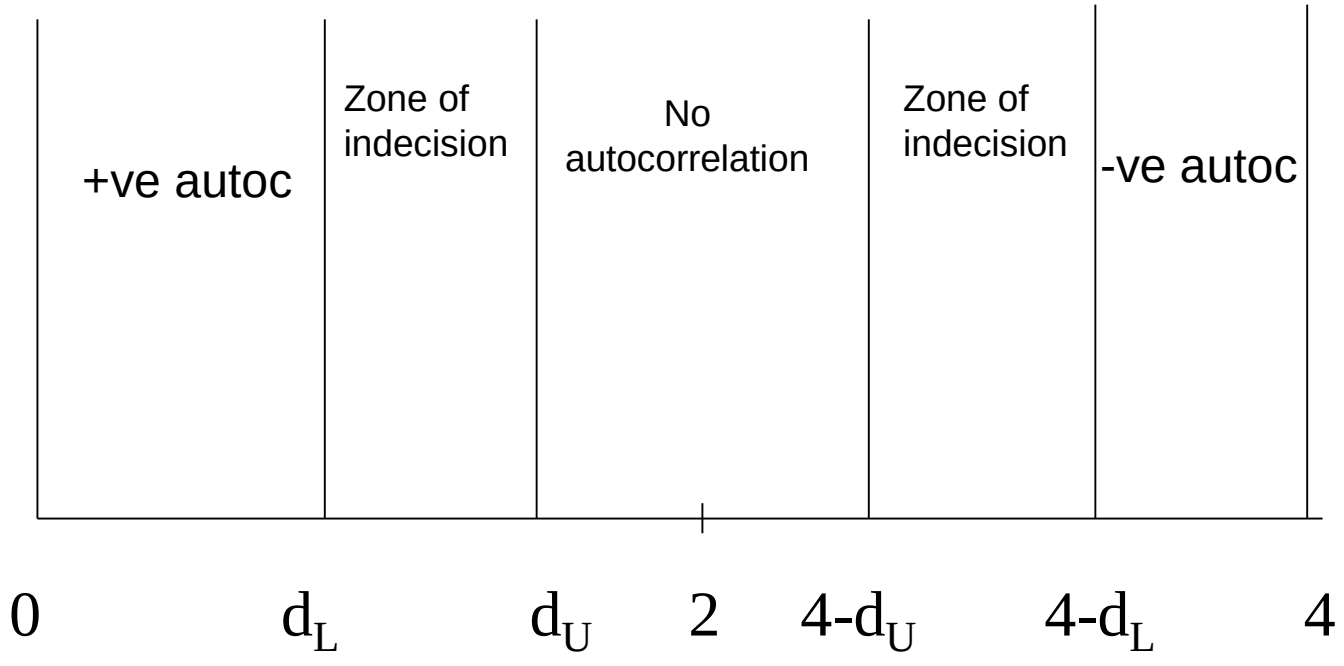
- Durbin-Watson have derived a lower bound d_L and an upper bound d_U such that if the computed d lies outside these critical values, a decision can be made regarding the presence of positive or negative serial correlation.

Detecting autocorrelation

With a large sample $d = 2(1- \rho)$

- But since $-1 \leq \rho \leq 1$, this implies that $0 \leq d \leq 4$ because:
- $\rho = 1$ implies $d = 2(1-1) = 0$ this is the lower bound d_u
  and we have positive autocorrelation
- $\rho = -1$ implies $d = 2(1-(-1)) = 4$ this is the upper bound d_L
  and we have negative autocorrelation
- If the statistic lies near the value 2, there is no serial correlation
- But if the statistic lies in the vicinity of 0, there is positive serial correlation.
- The closer the *d* is to zero, the greater the evidence of positive serial correlation.
- If it lies in the vicinity of 4, there is evidence of negative serial correlation
- If it lies between d_L and d_U / $4 - d_L$ and $4 - d_U$, then we are in the zone of indecision.

The Durbin Watson Test



Consequences of autocorrelation

- The OLS estimator is still unbiased.
- The OLS estimator is inefficient; that is, it is not BLUE.
- The estimated variances and covariances of the OLS estimates are biased and inconsistent.
- If there is positive autocorrelation, and if the value of a right-hand side variable grows over time, then the estimate of the standard error of the coefficient estimate of this variable will be too low and hence the t-statistic too high.
- Hypothesis tests are not valid.

Remedies for autocorrelation

- If autocorrelation is due to omission of an explanatory variable, the solution will require the inclusion of the omitted variable
- If autocorrelation is due to misspecification of the functional form of the model, then we need to correctly specify the model
- Transform data to correct the problem
- We can apply the method of first differences that assumes $\rho = 1$
- Durbin's method to estimate ρ and parameters of the model
- Cochrane-Orcutt method that transforms the original model and applies OLS in an iterative process