University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

4. Model Diagnostics

Prof. Dr. Miroslav Verbič

miroslav.verbic@ef.uni-lj.si

www.miroslav-verbic.si

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION AMBA ACCREDITED

Ljubljana, October 2025

Motivation

In order to be able to make ***valid* statistical inference** based on an econometric model:

1. The model has to be correctly specified (in terms of the variables and the functional form of the model).

2. The assumptions of the model need to be satisfied.

We focus in this chapter on **the latter**, assuming that the regression model is *already correctly specified*.

Motivation

Key assumptions of the classical linear regression model that cannot be taken for granted, and thus need to be *verified* and (if necessary) its validity needs to be *ensured*:

1. Normality of the disturbances
2. Absence of (perfect) multicollinearity
3. Homoscedasticity
4. Absence of autocorrelation

University of Ljubljana
SEB School of Economics and Business

Motivation

A What the assumption means and what are the key consequences if not fulfilled

B How to verify (test for) the validity of the assumption

C What are the possible solutions in case that the assumption is not fulfilled

EFMD EQUIS ACCREDITED
AACSB ACCREDITED
Association of MBAs ACCREDITED

University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

4.1 Normality of the disturbances

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION OF MBAS ACCREDITED

Meaning of the assumption

NORMAL DISTRIBUTION OF THE STOCHASTIC VARIABLE (DISTURBANCES) *u*

A **What the assumption means and what are the key consequences if not fulfilled**

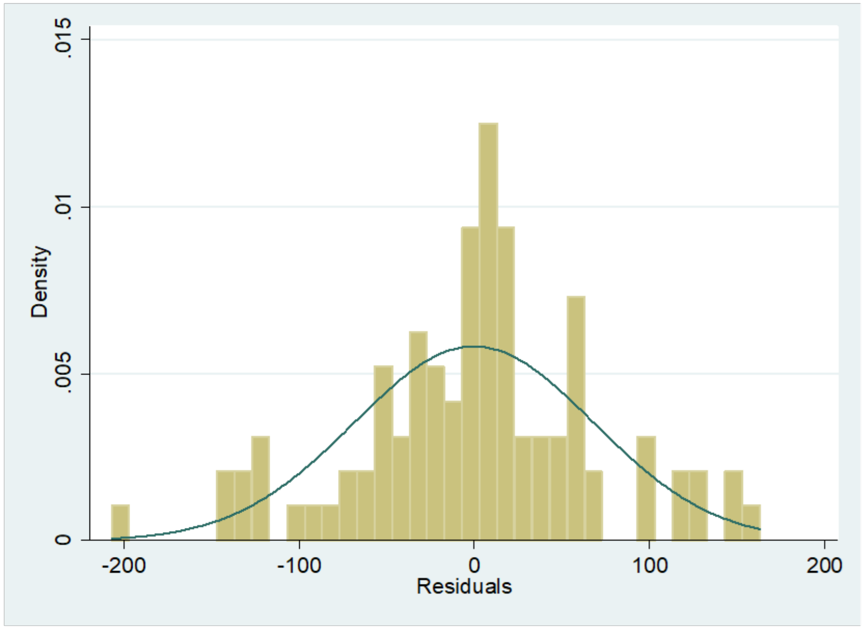
Assumption: $u \sim N\,(0,\ \sigma_u^{\ 2})$ and $y \sim N\,(\mathbf{X}\beta,\ \sigma_u^{\ 2})$

? **The assumption is essential for statistical inference (tests based on t, F and χ^2 distribution are dependent on it)!**

Verifying the validity of the assumption

B How to verify (test for) the validity of the assumption

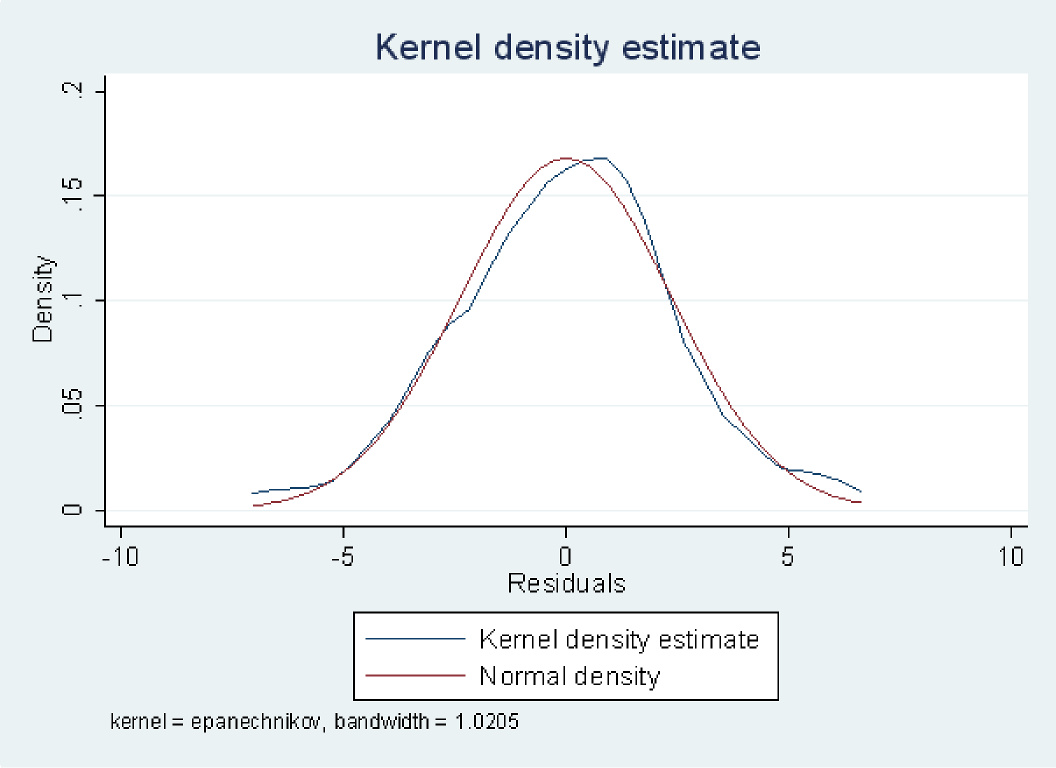
1 Histogram of (standardized) residuals of the sample regression model



Verifying the validity of the assumption

2

Kernel density plot of the residuals of the sample regression model



Verifying the validity of the assumption

3

Jarque – Bera test (1987)

$$JB = n\left[\frac{S^2}{6} + \frac{(K-3)^2}{24}\right] \sim \chi^2_{(2)}$$

$$S^2 = \frac{\left(\frac{1}{n}\sum e^3\right)^2}{\left(\frac{1}{n}\sum e^2\right)^3} \qquad K = \frac{\frac{1}{n}\sum e^4}{\left(\frac{1}{n}\sum e^2\right)^2}$$

H_0: stochastic variable u is normally distributed

H_1: stochastic variable u is **not** normally distributed

Solutions if the assumption is violated

C What are the possible solutions in case that the assumption is not fulfilled

1. Use a robust estimator of regression coefficients, e.g. the method of least absolute deviations.

2. Transformation of variables, e.g. taking logarithms of the dependent variable *y*.

3. Increase sample size, if possible. The validity of this assumption is not crucial for large samples.

University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

4.2 Multicollinearity

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION OF MBAS ACCREDITED

Economic background

The term "multicollinearity" was introduced into econometrics by Ragnar Frisch in 1934, as the *perfect linear dependence among the explanatory variables of the regression model.*

☹ **In general, we do not have controlled experiments in economics. It is thus not possible to control the data generating process, and measure the effects of particular phenomenon in isolation, i.e. independently of the other phenomena.**

☹ **As a consequence, economic phenomena are always related to one another, at least to a certain extent. There always exists a certain amount of *collinearity* among economic variables.**

Meaning of the assumption

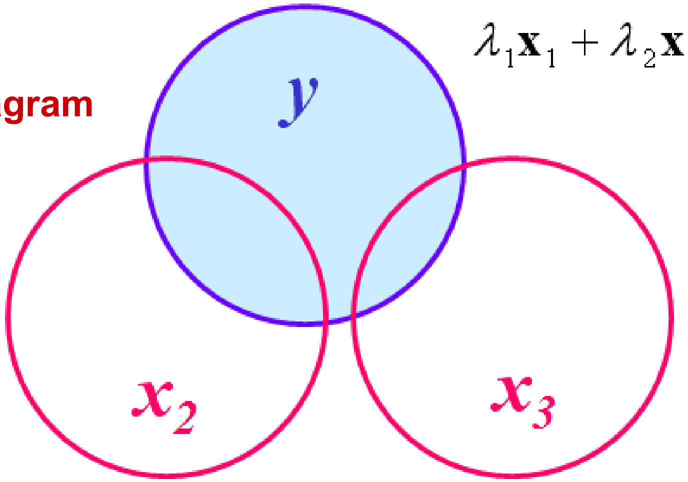
A **What the assumption means and what are the key consequences if not fulfilled**

Assumption:

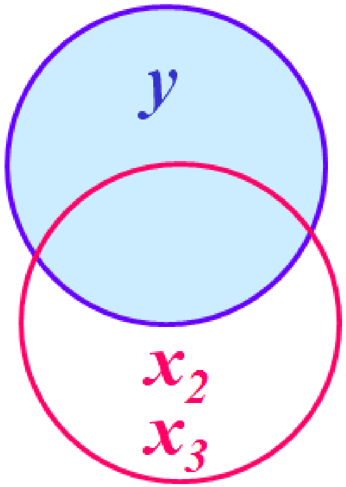
There is no perfect linear dependence among the explanatory variables of the regression model, i.e. no dependence of the type:

$$\lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2 + \ldots + \lambda_k \mathbf{x}_k = 0$$

Venn diagram



No (multi)collinearity



Perfect (multi)collinearity

Meaning of the assumption

I. Perfect multicollinearity

$\mathbf{x}_2 - 3\mathbf{x}_3 = 0$ ⇒ perfect "collinearity": $\mathbf{x}_2 = 3\mathbf{x}_3$

$\mathbf{x}_2 + \mathbf{x}_3 + 2\mathbf{x}_4 = 0$ ⇒ perfect "multicollinearity": $\mathbf{x}_2 = -\mathbf{x}_3 - 2\mathbf{x}_4$

1. Matrix X^TX is singular, i.e. it is not possible to calculate its inverse matrix.
2. $b = (X^TX)^{-1} X^Ty$ is not defined.
3. Multicollinearity relates only to linear dependence among explanatory variables, not non-linear.

Meaning of the assumption

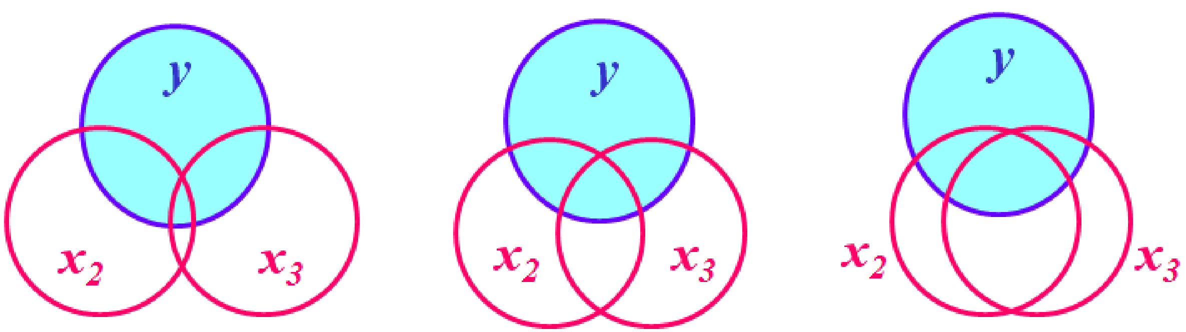
II. Imperfect multicollinearity

$$\lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2 + \ldots + \lambda_k \mathbf{x}_k + \mathbf{v} = \mathbf{0}$$

Each of the explanatory variables can be expressed by the other explanatory variables in the following way:

$$\mathbf{x}_k = -\frac{\lambda_1}{\lambda_k}\mathbf{x}_1 - \frac{\lambda_2}{\lambda_k}\mathbf{x}_2 - \ldots - \frac{\lambda_{k-1}}{\lambda_k}\mathbf{x}_{k-1} - \frac{1}{\lambda_k}\mathbf{v}$$

A particular explanatory variable is not a perfect linear combination of the other explanatory variables.



Weak (multi)collinearity Medium (multi)collinearity Strong (multi)collinearity

Meaning of the assumption

- **The least squares estimator remains *BLUE*, irrespective of (imperfect) multicollinearity.**

- **The variance of regression coefficient estimates increases with increasing multicollinearity.**

- **The variance of regression coefficient estimates is directly reflected on the *t*–statistics. It becomes increasingly difficult to reject H_0.**

- **Regression coefficient estimates and their standard errors become highly sensitive to model specification.**

Meaning of the assumption

The **value of R^2** is not affected substantially due to multicollinearity. Often we have a high value of R^2, but ***statistically insignificant regression coefficient estimates*** at the majority of explanatory variables.

The severity of multicollinearity issues is proportional to the level of multicollinearity among the explanatory variables.

Verifying the validity of the assumption

B How to verify (test for) the validity of the assumption

1 **The sign of one or more regression coefficients is in contradiction to the expectations of economic theory.**

2 **High value of R^2, whereas most of the estimated regression coefficients are statistically insignificant.**

Verifying the validity of the assumption

3

We use "auxilliary regressions":

$$x_{ji} = \beta_1 + \beta_2 x_{2i} + \ldots + \beta_{j-1} x_{j-1} + \beta_{j+1} x_{j+1} + \ldots + \beta_k x_k + v_i$$

$$\Rightarrow R^2_{x_j | \{x_1, \ldots, x_k\} \setminus x_j}$$

and calculate Variance Inflation Factors – *VIF*:

$$VIF_j = \frac{1}{1 - R^2_{x_j | \{x_1, \ldots, x_k\} \setminus x_j}}$$

Solutions if the assumption is violated

C **What are the possible solutions in case that the assumption is not fulfilled**

1. Often "*not doing anything*" is a satisfactory choice.
2. Drop one or more explanatory variables, causing the highest level of multicollinearity in the model.
3. Transformation of variables, e.g. taking first differences or logarithms.
4. Increase sample size, if possible.
5. When sensible and possible, combine time series with cross–sectional data in order to obtain a panel.

4.3 Heteroscedasticity

Meaning of the assumption

HOMOSCEDASTICITY

Origin of the word

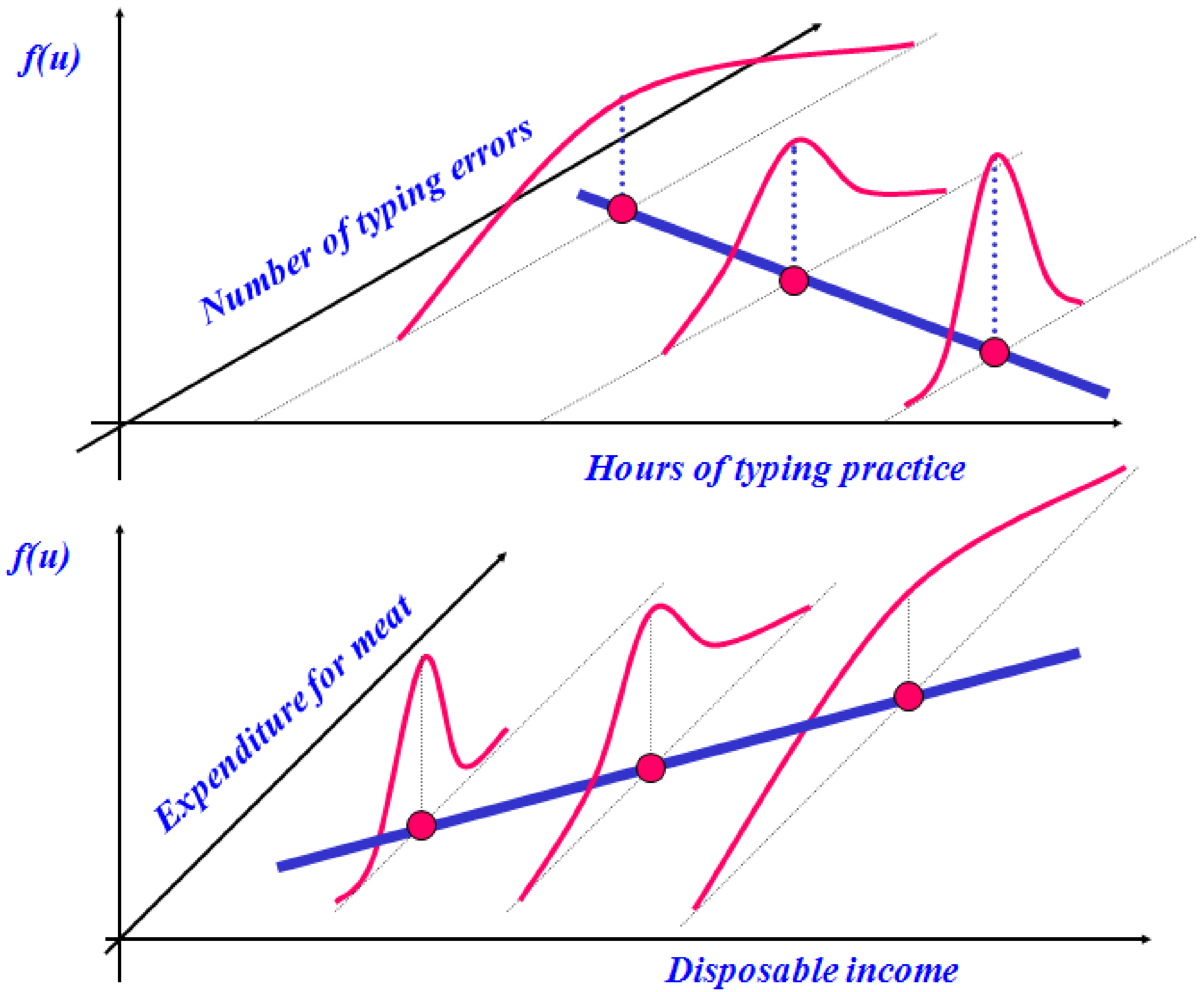
Homo : Skedasticos	Hetero : Skedasticos
Equal : Dispersion	Non-equal : Dispersion
Homoscedasticity	Heteroscedasticity

A What the assumption means and what are the key consequences if not fulfilled

Assumption:

$$Var(u_i|x_i) = E\left[\left(u_i - E(u_i|x_i)\right)^2 \Big| x_i\right] = E\left[u_i^2 \Big| x_i\right] = E(u_i^2) = \sigma^2$$

Meaning of the assumption



Meaning of the assumption

Causes of heteroscedasticity

1. Decision makers learn from their mistakes, i.e. the variability (measured by the variance) decreases with time.
2. Many phenomena increase with time in real terms; e.g. more flexible use of incomes (profits) due to economic growth increases the variability (variance) of stochastic effects.
3. Variability is often a result of poor organization and data collection. One can expect that with time the quality of data increases, and consequently the variance of stochastic effects decreases.
4. With analyses based on cross-sectional data, there are various reasons for heteroscedasticity – we determine them based on the properties of the analysed phenomenon. Heteroscedasticity occurs frequently when the range of values of a variable is very high.
5. Heteroscedasticity is often a consequence of a poor specification of the regression model – “spurious heteroscedasticity”.

Consequences of heteroscedasticity

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

$$E(\mathbf{u}) = \mathbf{0} \quad \text{but} \quad E(\mathbf{u}\mathbf{u}^{\mathrm{T}}) = Var - \text{cov}(\mathbf{u}) = \mathbf{W}$$

1

$$E(\mathbf{b}) = E\left[\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}\right] = E\left[\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\boldsymbol{\beta} + \mathbf{u})\right] =$$

$$= E\left(\boldsymbol{\beta} + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{u}\right) = \boldsymbol{\beta} + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}E(\mathbf{u}) = \boldsymbol{\beta}$$

The estimator of regression coefficients remains unbiased!

Consequences of heteroscedasticity

2

$$Var-\operatorname{cov}(\mathbf{b}) = E\left[(\mathbf{b}-\boldsymbol{\beta})(\mathbf{b}-\boldsymbol{\beta})^{\mathrm{T}}\right] =$$

$$= E\left[\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{u}\mathbf{u}^{\mathrm{T}}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\right] =$$

$$= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}E(\mathbf{u}\mathbf{u}^{\mathrm{T}})\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} =$$

$$= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$$

Homoscedasticity $\Rightarrow$ $Var-\operatorname{cov}(\mathbf{b}) = \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$

The estimator of regression coefficients is not most efficient any more! OLS is not BLUE any more, it is merely LUE!

Consequences of heteroscedasticity

3

The variance estimator of disturbances u is biased.

The variance and covariance estimators of regression coefficients become biased.

Test statistics of regression coefficients are not reliable any more!

Verifying the validity of the assumption

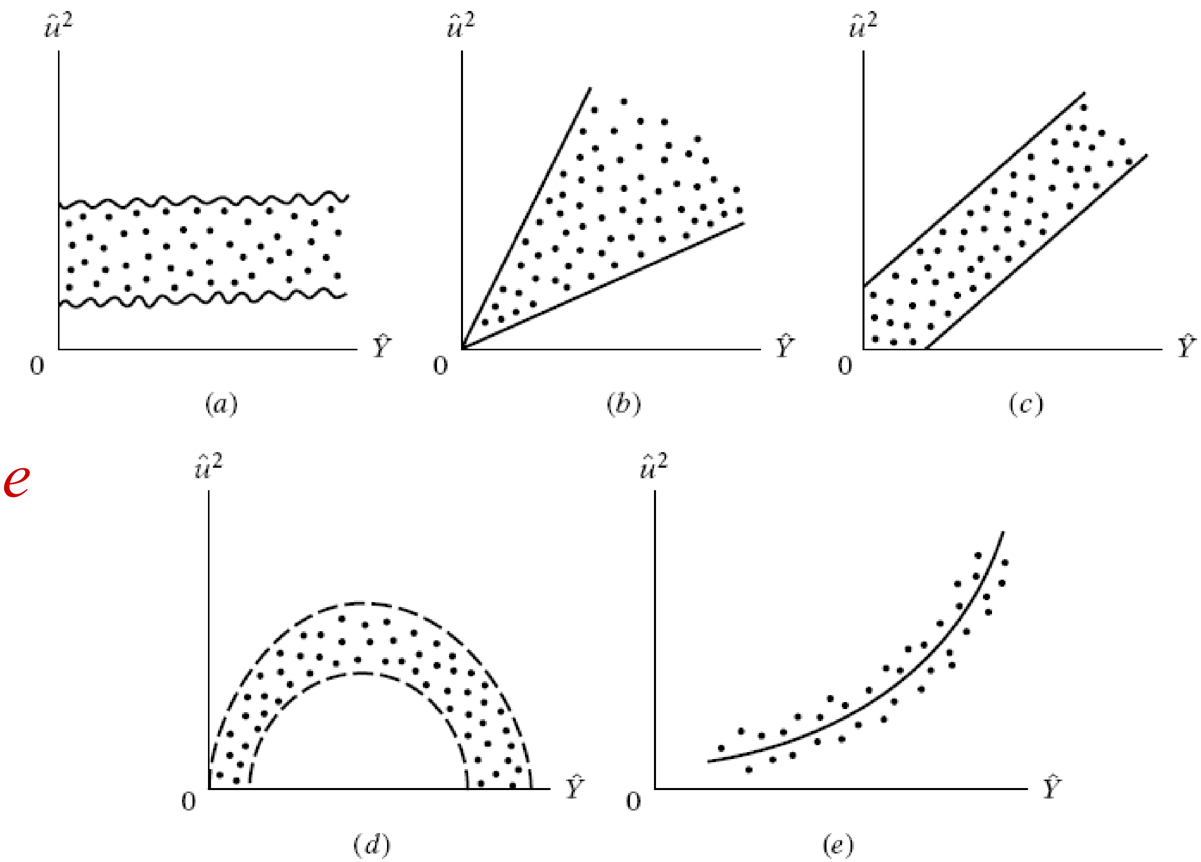
B How to verify (test for) the validity of the assumption

1. Graphic method of detecting heteroscedasticity.
2. Formal statistical tests, the most comprehensive being the White test.

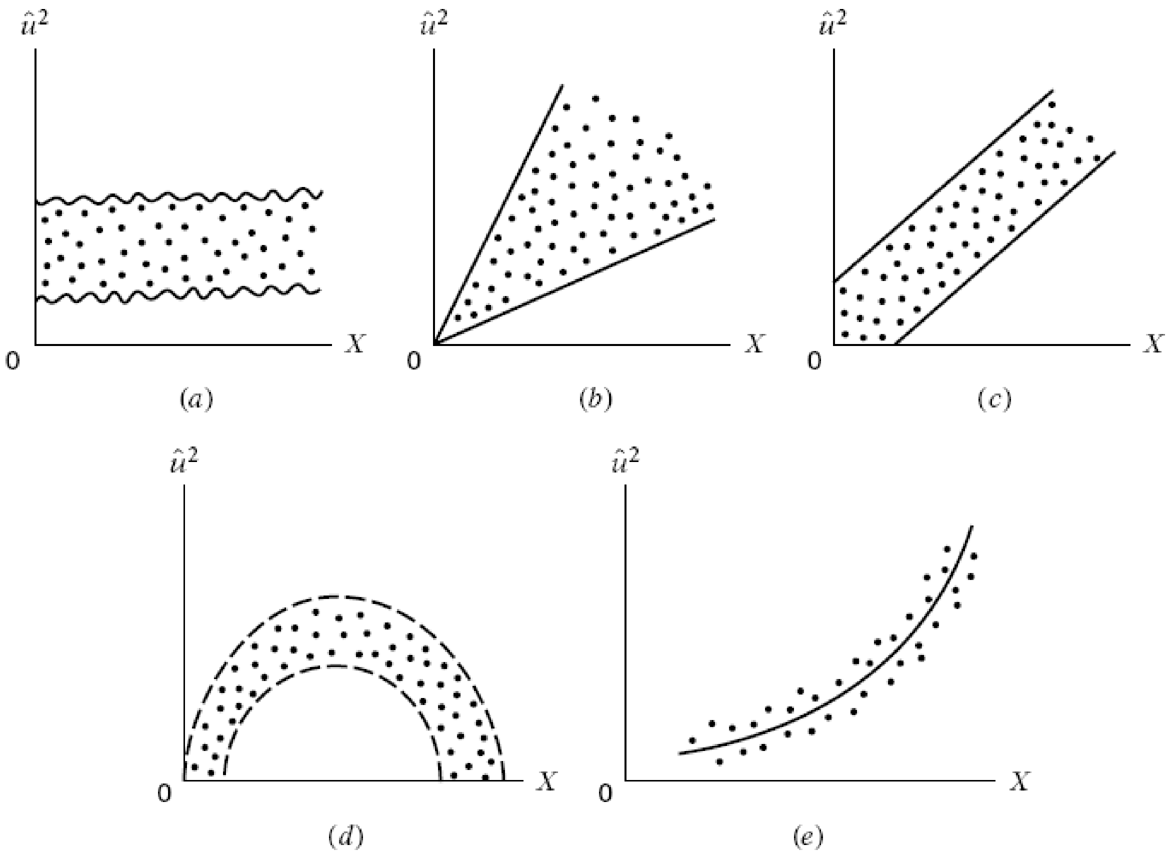
Verifying the validity of the assumption

1. Graphic method of detecting heteroscedasticity

$\hat{u} \equiv e$



Verifying the validity of the assumption



$\hat{u} \equiv e$

Verifying the validity of the assumption

2. White test (1980)

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

a) $$\sigma_i^2 = \alpha_1 + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \alpha_4 x_{2i}^2 + \alpha_5 x_{3i}^2 + \alpha_6 x_{2i} x_{3i}$$

b) **We calculate residuals e_i and estimate the auxilliary regression:**

$$e_i^2 = a_1 + a_2 x_{2i} + a_3 x_{3i} + a_4 x_{2i}^2 + a_5 x_{3i}^2 + a_6 x_{2i} x_{3i} + v_i$$

c) $H_0: \alpha_2 = \alpha_3 = \ldots = \alpha_m = 0$

H_1: At least one α different from 0

d) $$\theta(W) = nR^2 \sim \chi^2_{(m-1)}$$

$\theta \rangle \chi_c^2$ → **reject H_0**

Solutions if the assumption is violated

C What are the possible solutions in case that the assumption is not fulfilled

<table>
<tr><th>Problem</th><th>Heteroscedasticity</th><th>Autocorrelation</th></tr>
<tr><td rowspan="4">Problem management (once detected)</td><td colspan="2">Improvement of the model specification
(eliminates heteroscedasticity and autocorrelation that emerges due to biases)</td></tr>
<tr><td colspan="2">Application of generalized least squares (GLS) estimators:</td></tr>
<tr><td>weighted least squares (WLS) estimator</td><td>generalized difference equation (GDE) estimator:
➢ two-stage procedure
➢ iterative procedure (CORC)</td></tr>
<tr><td colspan="2">If the exact form of the problem is established, this approach eliminates all of the above adverse consequences.</td></tr>
</table>

Solutions if the assumption is violated

Problem	Heteroscedasticity	Autocorrelation
Problem management (once detected)	Robust variance estimators (Huber/White variance estimator)	HAC variance estimators (Newey–West robust variance estimator)
	$u \sim IID$ Estimator loosens the assumption on identical distribution.	$u \sim IID$ Estimator loosens both assumptions (on independence and identical distribution).
	Approach does not affect the regression coefficient estimates.	
	Standard errors regain unbiasedness. Regression coefficient estimator does not necessarily regain efficiency (standard errors are not necessarily the lowest possible).	
	Transformation of variables	AR(I)MAX methodology

Robust variance estimator application

Computer printout of estimation of a money demand regression model (Stata)

```
. regress hm1 ppr rvp rvv czp

      Source |       SS       df       MS              Number of obs =      96
-------------+------------------------------           F(  4,    91) =  527.72
       Model |  11431132.5     4  2857783.12           Prob > F      =  0.0000
    Residual |  492791.936    91    5415.296           R-squared     =  0.9587
-------------+------------------------------           Adj R-squared =  0.9569
       Total |  11923924.4    95  125514.994           Root MSE      =  73.589

------------------------------------------------------------------------------
         hm1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ppr |   1.697766    .513892     3.30   0.001     .6769831     2.71855
         rvp |  -311.6847   45.25178    -6.89   0.000    -401.5718   -221.7976
         rvv |  -11.57513    5.33166    -2.17   0.033    -22.16582    -.98444
         czp |   11.50168   1.472604     7.81   0.000     8.576535    14.42683
       _cons |  -229.2038   125.2134    -1.83   0.070    -477.9248    19.51725
------------------------------------------------------------------------------
```

Robust variance estimator application

```
. whitetst

White's general test statistic :  53.83009  Chi-sq(14)  P-value =  1.4e-06

. regress hm1 ppr rvp rvv czp, robust

Linear regression                                      Number of obs =      96
                                                       F(  4,    91) = 1000.25
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.9587
                                                       Root MSE      =  73.589

------------------------------------------------------------------------------
             |               Robust
         hm1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ppr |   1.697766   .5633882     3.01   0.003     .5786649    2.816868
         rvp |  -311.6847   44.33028    -7.03   0.000    -399.7413   -223.6281
         rvv |  -11.57513   3.532513    -3.28   0.001    -18.59203   -4.558225
         czp |   11.50168    1.32376     8.69   0.000     8.872196    14.13117
       _cons |  -229.2038   58.25138    -3.93   0.000     -344.913   -113.4945
------------------------------------------------------------------------------
```

Robust variance estimator application

Computer printout of estimation of a money demand regression model (R)

```
> mod = lm(hm1 ~ ppr + rvp + rvv + czp, data = money_demand)
> summary(mod)

Call:
lm(formula = hm1 ~ ppr + rvp + rvv + czp, data = money_demand)

Residuals:
Harmonized money aggregate M1
     Min       1Q   Median       3Q      Max
-180.693  -36.611    1.595   38.308  152.114

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -229.2038   125.2134  -1.831  0.07045 .
ppr            1.6978     0.5139   3.304  0.00137 **
rvp         -311.6847    45.2518  -6.888 7.14e-10 ***
rvv          -11.5751     5.3317  -2.171  0.03253 *
czp           11.5017     1.4726   7.810 9.44e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 73.59 on 91 degrees of freedom
Multiple R-squared:  0.9587,    Adjusted R-squared:  0.9569
F-statistic: 527.7 on 4 and 91 DF,  p-value: < 2.2e-16
```

Robust variance estimator application

```
> white_lm(mod, interactions=TRUE)
# A tibble: 1 x 5
  statistic    p.value parameter method       alternative
      <dbl>      <dbl>     <dbl> <chr>        <chr>
1      53.8 0.00000137        14 White's Test greater

> mod_robust = lm_robust(hm1 ~ ppr + rvp + rvv + czp, data = money_demand,
  se_type="HC1")
> summary(mod_robust)

Call:
lm_robust(formula = hm1 ~ ppr + rvp + rvv + czp, data = money_demand,
    se_type = "HC1")

Standard error type:  HC1

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)  CI Lower CI Upper DF
(Intercept) -229.204    58.2514  -3.935 1.626e-04 -344.9130 -113.495 91
ppr            1.698     0.5634   3.013 3.345e-03    0.5787    2.817 91
rvp         -311.685    44.3303  -7.031 3.682e-10 -399.7413 -223.628 91
rvv          -11.575     3.5325  -3.277 1.487e-03  -18.5920   -4.558 91
czp           11.502     1.3238   8.689 1.414e-13    8.8722   14.131 91

Multiple R-squared:  0.9587 ,   Adjusted R-squared:  0.9569
F-statistic:  1000 on 4 and 91 DF,  p-value: < 2.2e-16
```

EFMD EQUIS ACCREDITED
AACSB ACCREDITED
ASSOCIATION OF MBAs ACCREDITED

University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

4.4 Autocorrelation

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION OF MBAS ACCREDITED

Meaning of the assumption

A What the assumption means and what are the key consequences if not fulfilled

Assumption

$$Cov\left(u_i, u_j \middle| x_i, x_j\right) = 0 \quad \Longleftrightarrow \quad Cov\left(u_i, u_j \middle| x_i, x_j\right) \neq 0$$

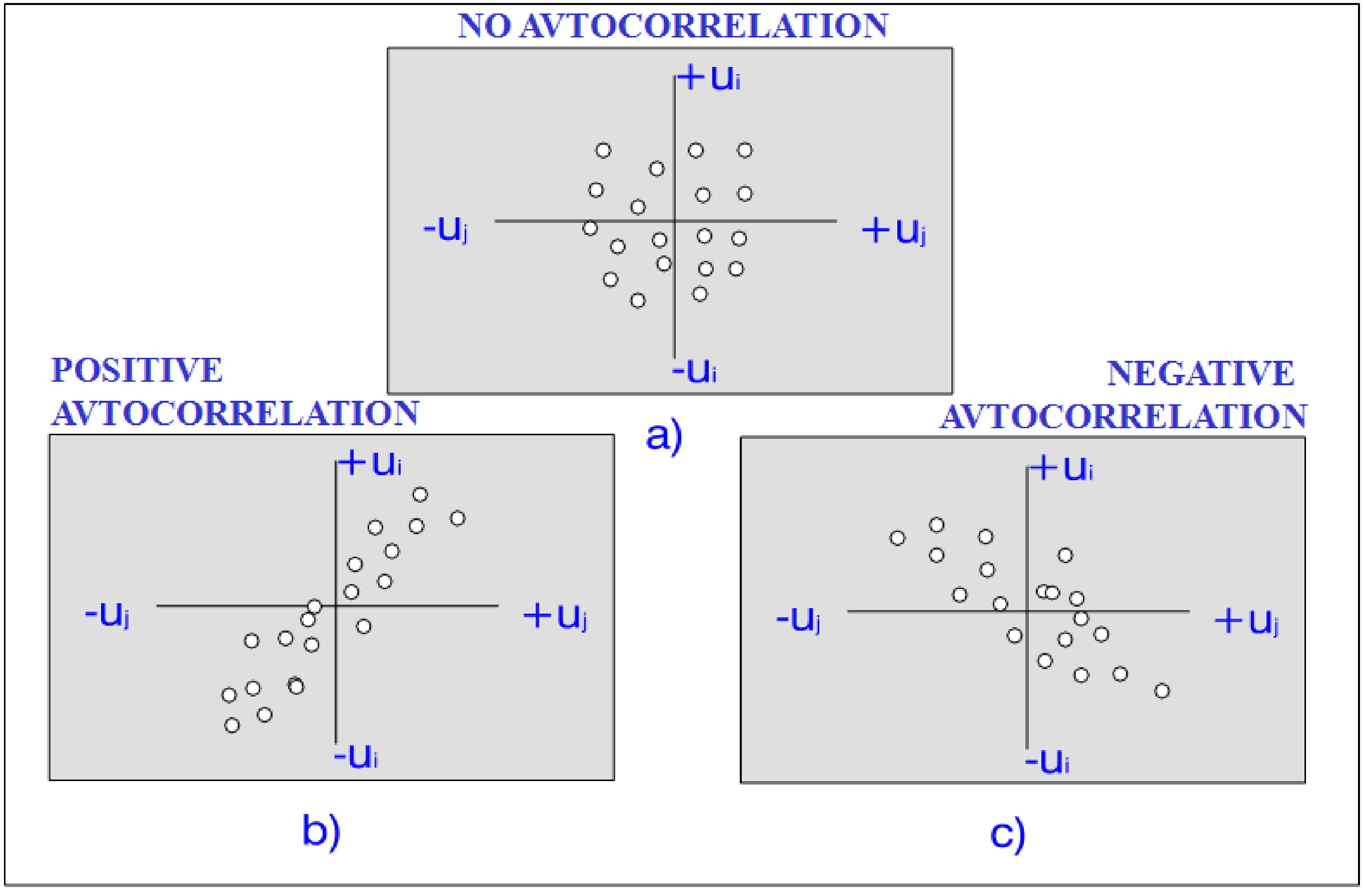
$$i \neq j \qquad\qquad i \neq j$$

No autocorrelation | **Autocorrelation**

We distinguish between *genuine autocorrelation* and *spurious autocorrelation*. The latter is a consequence of a poorly specified regression model.

Meaning of the assumption

Scatter plots of (lagged) stochastic variable *u*:



Meaning of the assumption

Causes of autocorrelation

1. *Genuine*. In most time series there exists an underlying inertia, i.e. its development depends on the phenomenon in the previous time period(s). This is also the main reason for cyclical developments.

2. *Spurious, i.e. model specification errors*.
In empirical research, we often proceed from the most general model, where we can "drop" an important explanatory variable. This is called the excluded-variable specification bias and often causes autocorrelation. Here is an example:

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + u_t$$

Quantity — Price — Income — Price of a substitute or a complement

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + v_t \Rightarrow v_t = \beta_4 x_{4t} + u_t$$

Meaning of the assumption

3 Model specification error due to wrong functional form of the regression model, e.g.:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{2i}^2 + u_i$$

Marginal cost | Product | Product

$$y_i = \beta_1 + \beta_2 x_{2i} + v_i \Rightarrow v_i = \beta_3 x_{2i}^2 + u_i$$

4 Not taking into account lagged variables , i.e. not including lags of (explanatory) variables.

Meaning of the assumption

5 Use of transformations with time series, e.g. averages, sums, trends and various interpolations.

6 Non-stationarity of time series, i.e. first and second moments of a time series not being constant in time.

When the dependent and explanatory variable(s) are non-stationary, it is likely that the stochastic variable is non-stationary as well, and the model exhibits autocorrelation.

Meaning of the assumption

Types (orders) of autocorrelation

1 **First-order autocorrelation (first-order autoregression scheme)**

$$u_t = \rho_1 u_{t-1} + \varepsilon_t \Rightarrow \text{AR(1)}$$

$-1 < \rho_1 < 1$ $\rho = \rho_1$ – coefficient of first-order autocorrelation

2 **Second-order autocorrelation**

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \varepsilon_t \Rightarrow \text{AR(2)}$$

3 ***p*–th order autocorrelation**

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \ldots + \rho_p u_{t-p} + \varepsilon_t \Rightarrow \text{AR}(p)$$

Consequences of autocorrelation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

$$E(\mathbf{u}) = \mathbf{0} \quad \text{but} \quad E\left(\mathbf{u}\mathbf{u}^{\mathrm{T}}\right) = Var - \operatorname{cov}(\mathbf{u}) = \mathbf{W}$$

1

$$E(\mathbf{b}) = E\left[\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}\right] = E\left[\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\boldsymbol{\beta} + \mathbf{u})\right] =$$

$$= E\left(\boldsymbol{\beta} + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{u}\right) = \boldsymbol{\beta} + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}E(\mathbf{u}) = \boldsymbol{\beta}$$

The estimator of regression coefficients remains unbiased!

Consequences of autocorrelation

2

$$Var-\mathrm{cov}(\mathbf{b}) = E\left[(\mathbf{b}-\boldsymbol{\beta})(\mathbf{b}-\boldsymbol{\beta})^{\mathrm{T}}\right] =$$

$$= E\left[\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{u}\mathbf{u}^{\mathrm{T}}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\right] =$$

$$= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}E(\mathbf{u}\mathbf{u}^{\mathrm{T}})\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} =$$

$$= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$$

No autocorrelation $\Longrightarrow$ $Var-\mathrm{cov}(\mathbf{b}) = \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$

The estimator of regression coefficients is not most efficient any more! OLS is not BLUE any more, it is merely LUE!

Consequences of autocorrelation

3

The variance estimator of disturbances u is biased.

The variance and covariance estimators of regression coefficients become biased.

Test statistics of regression coefficients are not reliable any more!

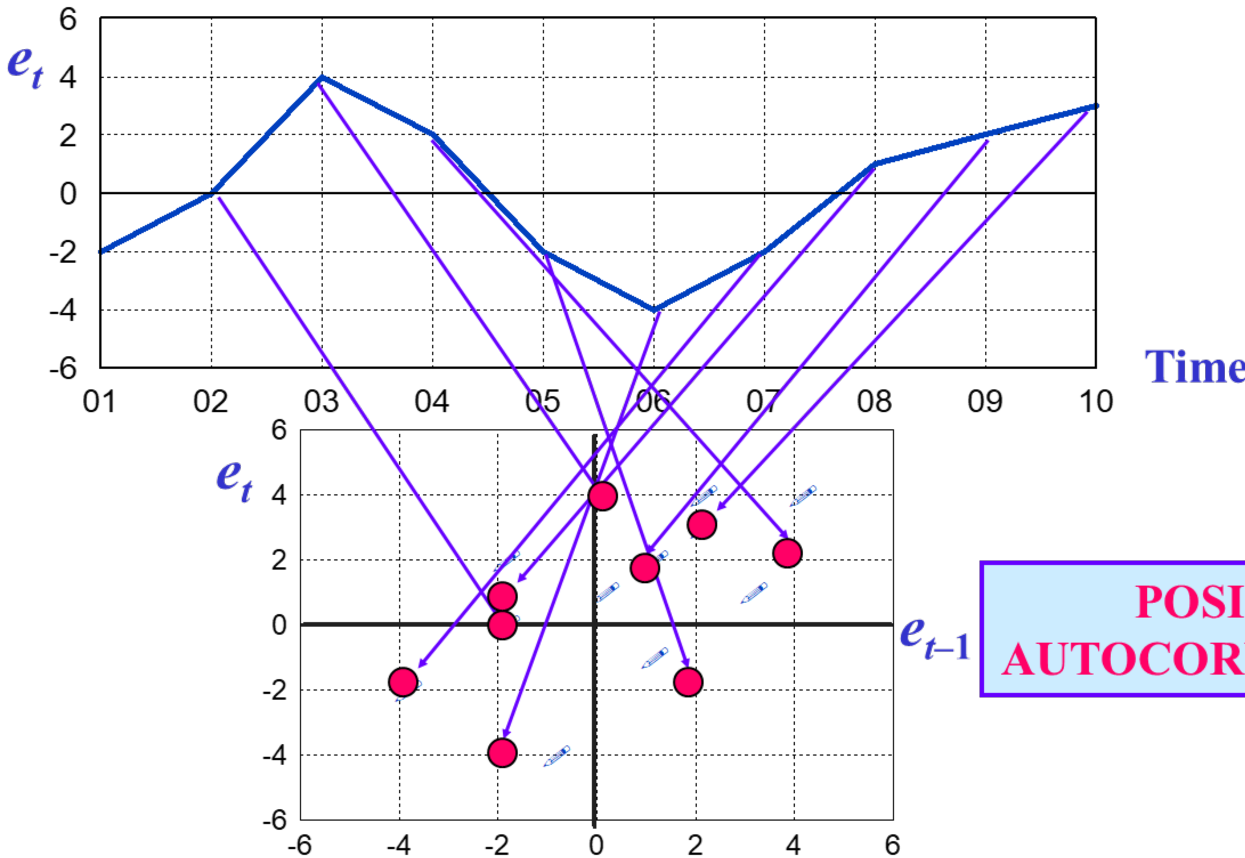
Verifying the validity of the assumption

B **How to verify (test for) the validity of the assumption**

1. Graphic method of detecting autocorrelation.
2. Formal statistical tests, the most comprehensive being the Breusch–Godfrey test.

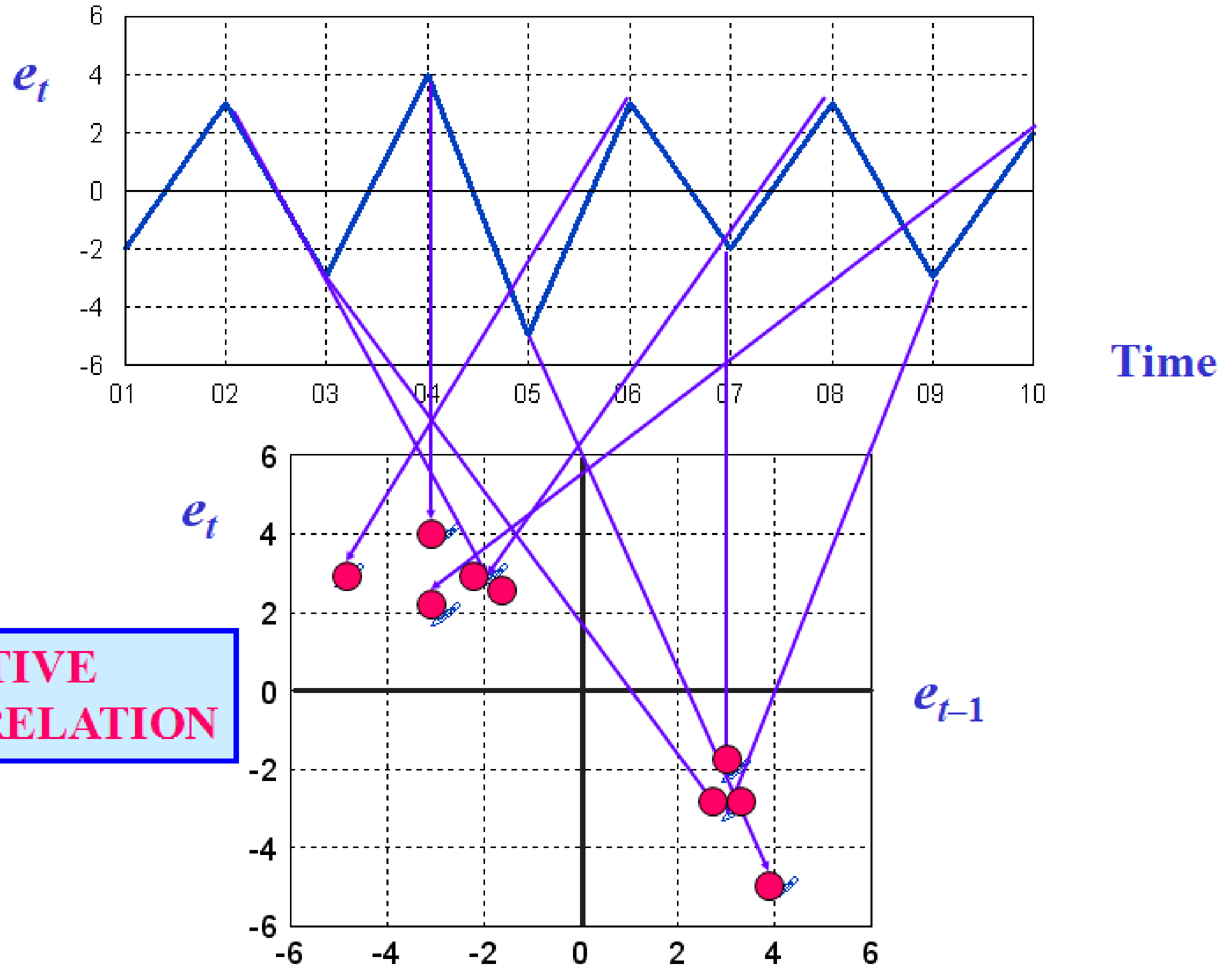
Verifying the validity of the assumption

1. Graphic method of detecting autocorrelation



POSITIVE AUTOCORRELATION

Verifying the validity of the assumption



NEGATIVE AUTOCORRELATION

Verifying the validity of the assumption

2. Breusch–Godfrey test (1978) (Lagrange multiplier test)

1 **Assume any order of autocorrelation, e.g. *p*–th order:**

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots + \rho_p u_{t-p} + \varepsilon_t$$

$H_0: \rho_1 = \rho_2 = \cdots = \rho_p = 0$ ⟹ H_1: At least one ρ_j different from 0

2 **Test statistic**

Estimate auxilliary regression:

$$\hat{e}_t = b_1 + b_2 x_{2t} + \cdots + b_k x_{kt} + \hat{\rho}_1 e_{t-1} + \cdots + \hat{\rho}_p e_{t-p}$$

$$LM = (n-p)R^2 \sim \chi^2_{(p)}$$

- Regression model can include lagged dependent variable as a regressor
- Lag order not known in advance
- Potential problems wih degrees of freedom

Solutions if the assumption is violated

C What are the possible solutions in case that the assumption is not fulfilled

Problem	Heteroscedasticity	Autocorrelation
Problem management (once detected)	Improvement of the model specification (eliminates heteroscedasticity and autocorrelation that emerges due to biases)	
	Application of generalized least squares (GLS) estimators:	
	weighted least squares (WLS) estimator	generalized difference equation (GDE) estimator: ➢ two-stage procedure ➢ iterative procedure (CORC)
	If the exact form of the problem is established, this approach eliminates all of the above adverse consequences.	

Solutions if the assumption is violated

Problem	Heteroscedasticity	Autocorrelation
Problem management (once detected)	Robust variance estimators (Huber/White variance estimator) $u \sim IID$ Estimator loosens the assumption on identical distribution.	HAC variance estimators (Newey–West robust variance estimator) $u \sim IID$ Estimator loosens both assumptions (on independence and identical distribution).
	Approach does not affect the regression coefficient estimates. Standard errors regain unbiasedness. Regression coefficient estimator does not necessarily regain efficiency (standard errors are not necessarily the lowest possible).	
	Transformation of variables	AR(I)MAX methodology

HAC variance estimator application

Computer printout of estimation of a money demand regression model (Stata)

```
. regress hm1 ppr rvp rvv czp

      Source |       SS       df       MS              Number of obs =      96
-------------+------------------------------           F(  4,    91) =  527.72
       Model |  11431132.5     4  2857783.12           Prob > F      =  0.0000
    Residual |  492791.936    91    5415.296           R-squared     =  0.9587
-------------+------------------------------           Adj R-squared =  0.9569
       Total |  11923924.4    95  125514.994           Root MSE      =  73.589

------------------------------------------------------------------------------
         hm1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ppr |   1.697766    .513892     3.30   0.001     .6769831     2.71855
         rvp |  -311.6847   45.25178    -6.89   0.000    -401.5718   -221.7976
         rvv |  -11.57513    5.33166    -2.17   0.033    -22.16582    -.98444
         czp |   11.50168   1.472604     7.81   0.000     8.576535    14.42683
       _cons |  -229.2038   125.2134    -1.83   0.070    -477.9248    19.51725
------------------------------------------------------------------------------
```

HAC variance estimator application

```
. estat bgodfrey, lags(1 2 3 4 5 6 7 8 9 10 11 12)

Breusch-Godfrey LM test for autocorrelation
---------------------------------------------------------------------------
    lags(p)  |          chi2               df                 Prob > chi2
-------------+-------------------------------------------------------------
       1     |         70.771               1                   0.0000
       2     |         70.773               2                   0.0000
       3     |         71.397               3                   0.0000
       4     |         71.398               4                   0.0000
       5     |         71.639               5                   0.0000
       6     |         71.641               6                   0.0000
       7     |         73.403               7                   0.0000
       8     |         73.536               8                   0.0000
       9     |         74.066               9                   0.0000
      10     |         74.112              10                   0.0000
      11     |         74.164              11                   0.0000
      12     |         76.742              12                   0.0000
---------------------------------------------------------------------------
                        H0: no serial correlation
```

HAC variance estimator application

```
. newey hm1 ppr rvp rvv czp, lag(78)

Regression with Newey-West standard errors          Number of obs  =        96
maximum lag: 78                                     F(  4,     91)  =    859.25
                                                    Prob > F        =    0.0000

------------------------------------------------------------------------------
             |             Newey-West
         hm1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ppr |   1.697766   .3108188     5.46   0.000     1.080363     2.31517
         rvp |  -311.6847   64.70783    -4.82   0.000    -440.2189   -183.1505
         rvv |  -11.57513   6.148327    -1.88   0.063    -23.78802     .637769
         czp |   11.50168    .672376    17.11   0.000     10.16609    12.83727
       _cons |  -229.2038   71.30645    -3.21   0.002    -370.8453   -87.56224
------------------------------------------------------------------------------
```

HAC variance estimator application

Computer printout of estimation of a money demand regression model (R)

```
> mod = lm(hm1 ~ ppr + rvp + rvv + czp, data = money_demand)
> summary(mod)

Call:
lm(formula = hm1 ~ ppr + rvp + rvv + czp, data = money_demand)

Residuals:
Harmonized money aggregate M1
     Min       1Q   Median       3Q      Max 
-180.693  -36.611    1.595   38.308  152.114 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -229.2038   125.2134  -1.831  0.07045 .  
ppr            1.6978     0.5139   3.304  0.00137 ** 
rvp         -311.6847    45.2518  -6.888 7.14e-10 ***
rvv          -11.5751     5.3317  -2.171  0.03253 *  
czp           11.5017     1.4726   7.810 9.44e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 73.59 on 91 degrees of freedom
Multiple R-squared:  0.9587,	Adjusted R-squared:  0.9569 
F-statistic: 527.7 on 4 and 91 DF,  p-value: < 2.2e-16
```

HAC variance estimator application

```
> bgtest_tab = as.data.frame(matrix(ncol=4, nrow=12))
> names(bgtest_tab) = c("Order","LM-test","df","p-value")
> for (i in c(1:12)) {
+   a = bgtest(mod,order=i)
+   bgtest_tab[i,1] = i
+   bgtest_tab[i,2] = a$statistic
+   bgtest_tab[i,3] = a$parameter
+   bgtest_tab[i,4] = round(a$p.value,4)
+ }

> bgtest_tab
   Order  LM-test df p-value
1      1 70.77133  1  0.0000
2      2 70.77328  2  0.0000
3      3 71.39680  3  0.0000
4      4 71.39773  4  0.0000
5      5 71.63857  5  0.0000
6      6 71.64141  6  0.0000
7      7 73.40315  7  0.0000
8      8 73.53570  8  0.0000
9      9 74.06588  9  0.0000
10    10 74.11248 10  0.0000
11    11 74.16396 11  0.0000
12    12 76.74155 12  0.0000
```

HAC variance estimator application

```
> coeftest(mod, vcov.=NeweyWest(mod, lag=78, adjust=TRUE, prewhite=FALSE))

t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)    
(Intercept) -229.20375   71.30645 -3.2143   0.00181 ** 
ppr            1.69777    0.31082  5.4622 4.056e-07 ***
rvp         -311.68470   64.70783 -4.8168 5.793e-06 ***
rvv          -11.57513    6.14833 -1.8826   0.06294 .  
czp           11.50168    0.67238 17.1060 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

4. Model Diagnostics

Prof. Dr. Miroslav Verbič

miroslav.verbic@ef.uni-lj.si

www.miroslav-verbic.si

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION AMBA ACCREDITED

Ljubljana, October 2025