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 2022

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

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

University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

4.1 Normality of the disturbances

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION AMBA 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**

this needs to hold for u t-test to be valid

this holds because x are fixed

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

disturbance is distributed normally

?

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

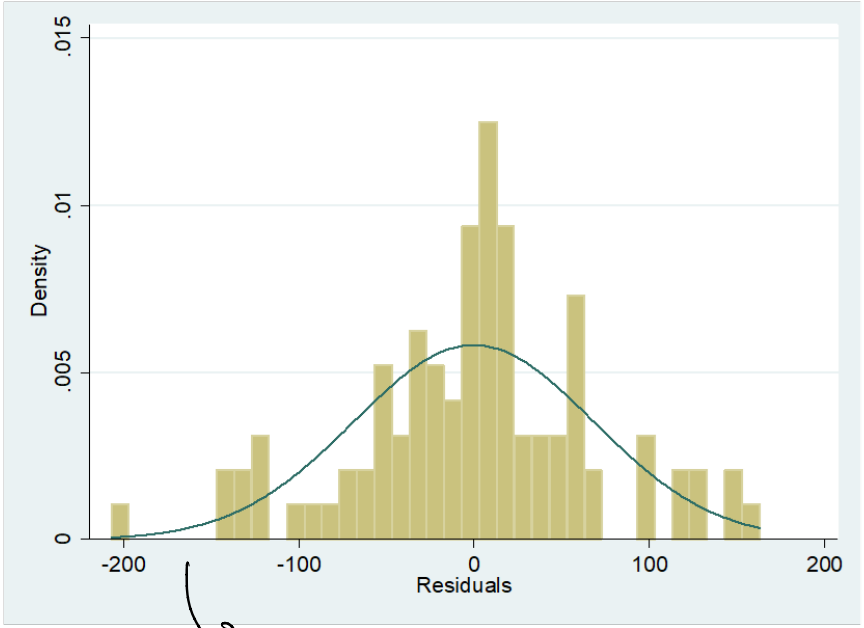
$$X\beta = \beta_1 + \beta_2 X_2 + \ldots + \beta_K X_K$$

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

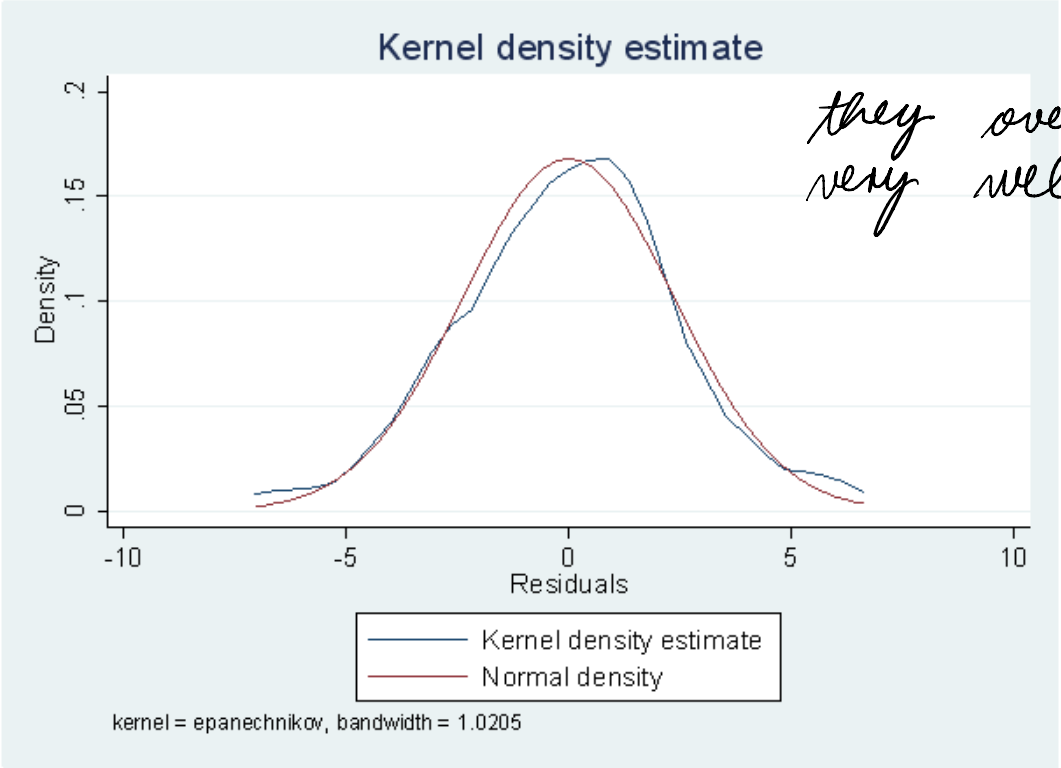


the two functions overlap very well

Verifying the validity of the assumption

2

Kernel density plot of the residuals of the sample regression model



they overlap very well

Central moments of the distribution for the STANDARDIZED NORMAL DISTRIBUTION

Slide 8

$a_1 = \mu = 0$ → first moment is mean

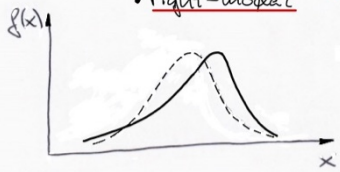
$a_2 = \sigma^2 = 1$

$a_3 = S = 0$ ~ skewness ⇒ measures assimetry
↳ normal distribution $S = 0$

$a_4 = K = 3$ ~ kurtosis
measures tailedness of the distribution, normal distribution $K = 3$

$\vdots$

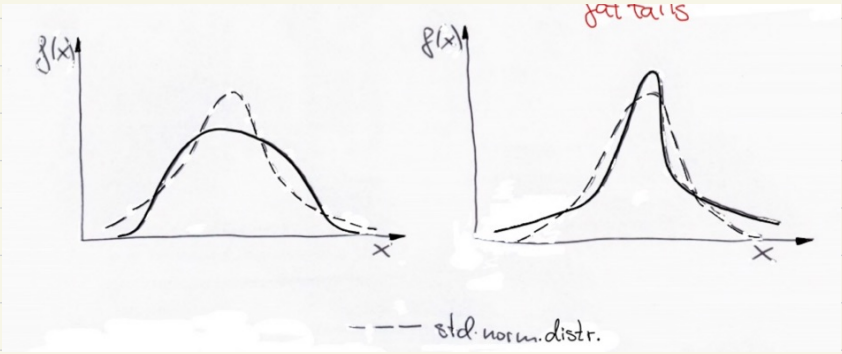
$S < 0$:
- neg. asymmetry
- asymmetry to the left
- right-modal



$S > 0$:
- positive asymmetry
- asymmetry to the right
- left-modal

$K < 3$: platykurtosis, less prob. in the tails

$K > 3$: leptokurtosis, more prob. in the tails, "fat tails" — most often in reality



Verifying the validity of the assumption

This test is most often used in econometrics

Jarque – Bera test (1987)

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

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

In this case we don't want to reject the null hypothesis

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. $\Rightarrow n > 100$

this is important for small samples

↳ In practice this is not very useful

Example 1: We gathered a sample of data for 32 European countries for the year 2003. We have the following variables available (the data are provided in Stata Data file `health.dta`, while the programming code is given in Stata Do file `health-commands-l08.do`):

- life expectancy at birth (*LIFE*; in years);
- health expenditure per capita (*EXP*; in U.S. dollars);
- percentage of smokers among adults (*TOBACCO*);
- consumption of alcohol per capita (*ALCO*; in litres of distilled spirits).

a) Estimate the linear regression model: $LIFE_i = \beta_1 + \beta_2 EXP_i + \beta_3 TOBACCO_i + u_i$ and check validity of the assumption on normality of the disturbances.

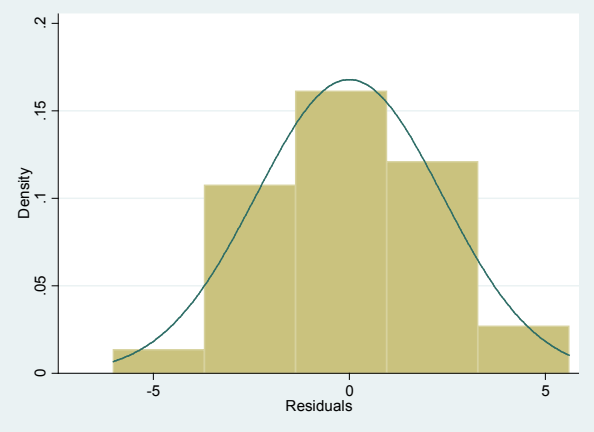
b) Estimate the linear regression model: $LIFE_i = \beta_1 + \beta_2 EXP_i + \beta_3 TOBACCO_i + \beta_4 ALCO_i + u_i$ and check validity of the assumption on (absence of) multicollinearity.

Computer printout of the results in Stata:

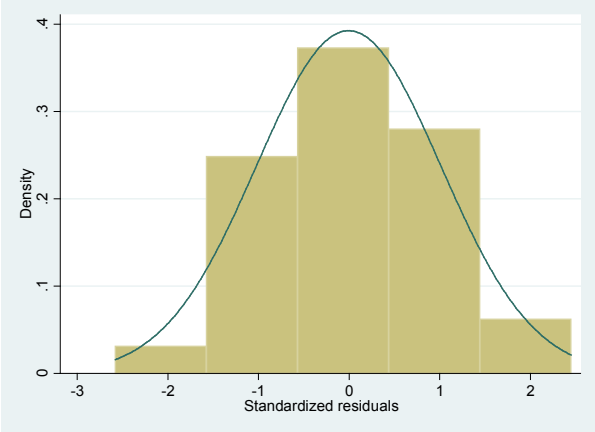
Normality of the disturbances:

small sample

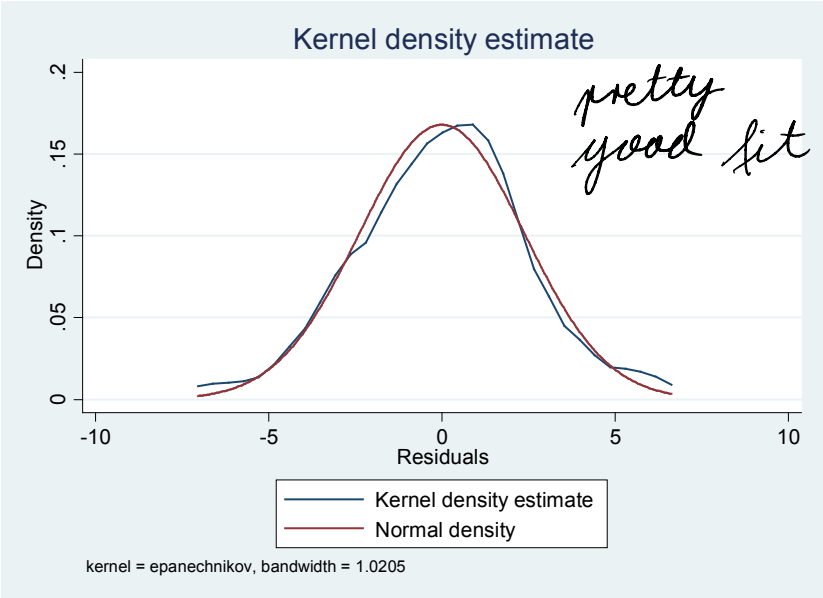
```
. regress life exp tobacco

      Source |       SS       df       MS              Number of obs =      32
-------------+------------------------------           F(  2,    29) =   31.97
       Model |  385.751827     2  192.875914            Prob > F      =  0.0000
    Residual |   174.97295    29  6.03354999            R-squared     =  0.6880
-------------+------------------------------           Adj R-squared =  0.6664
       Total |  560.724777    31   18.087896            Root MSE      =  2.4563

------------------------------------------------------------------------------
        life |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   .0023323   .0003709     6.29   0.000     .0015736    .0030909
     tobacco |  -.2503555   .0889983    -2.81   0.009    -.4323774   -.0683335
       _cons |   79.62409   2.796632    28.47   0.000     73.90433    85.34384
------------------------------------------------------------------------------

. predict elife, resid
. predict elifestd, rstandard

. histogram elife, normal
(bin=5, start=-6.0237689, width=2.3258684)
```




```
. histogram elifestd, normal
(bin=5, start=-2.5803783, width=1.0055467)
```



```
. kdensity elife, normal
(n() set to 32)
```




```
. sum elife, detail
```

```
                        Residuals
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -6.023769      -6.023769
 5%    -3.561575      -3.561575
10%    -2.820123      -2.855744       Obs                  32
25%    -1.700531      -2.820123       Sum of Wgt.          32

50%     .2013203                      Mean           1.28e-08
                       Largest        Std. Dev.      2.375771
75%     1.358591       2.43306
90%      2.43306       2.645627       Variance       5.644289
95%     4.427194       4.427194       Skewness       -.0833723
99%     5.605573       5.605573       Kurtosis       3.410826
```

0
3

due to normal distribution

```
. return list

scalars:
                  r(N) =  32
              r(sum_w) =  32
               r(mean) =  1.28056854010e-08
                r(Var) =  5.644288676812403
                 r(sd) =  2.37577117517921
           r(skewness) =  -.0833723127091076
           r(kurtosis) =  3.410825630496252
                r(sum) =  4.09781932831e-07
                r(min) =  -6.023768901824951
                r(max) =  5.605573177337647
                 r(p1) =  -6.023768901824951
                 r(p5) =  -3.561574935913086
                r(p10) =  -2.820123434066773
                r(p25) =  -1.700530529022217
                r(p50) =  .2013202682137489
                r(p75) =  1.358591318130493
                r(p90) =  2.433059930801392
                r(p95) =  4.427193641662598
                r(p99) =  5.605573177337647
```

```
. scalar obs=r(N)
. scalar s=r(skewness)
. scalar k=r(kurtosis)

. scalar jb=obs*(s^2/6+(k-3)^2/24)
. display jb
.26210863

. display chi2tail(2,jb)
.87717013
```

→ uporabimo formulo na 8. slajdu

p value → we cannot reject Ho

```
. jb6 elife
Jarque-Bera normality test:  .2621 Chi(2)  .8772
Jarque-Bera test for Ho: normality: (elife)
```

izbrisano na vajah

Multicollinearity:

```
. regress life exp tobacco alco

      Source |       SS       df       MS              Number of obs =      32
-------------+------------------------------           F(  3,    28) =   26.30
       Model |  413.850212     3  137.950071            Prob > F      =  0.0000
    Residual |  146.874565    28  5.24552017            R-squared     =  0.7381
-------------+------------------------------           Adj R-squared =  0.7100
       Total |  560.724777    31   18.087896            Root MSE      =  2.2903

------------------------------------------------------------------------------
        life |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   .0018569   .0004023     4.62   0.000     .0010329    .0026809
     tobacco |  -.2238391   .0837702    -2.67   0.012    -.3954346   -.0522436
        alco |  -.6493606   .2805689    -2.31   0.028     -1.22408   -.0746412
       _cons |   81.42053   2.720683    29.93   0.000     75.84746    86.99359
------------------------------------------------------------------------------

. pwcorr exp tobacco alco, sig

             |      exp  tobacco     alco
-------------+---------------------------
         exp |   1.0000
             |
             |
     tobacco |  -0.3017   1.0000
             |   0.0933
             |
        alco |  -0.5510   0.2751   1.0000
             |   0.0011   0.1276

. regress exp tobacco alco

      Source |       SS       df       MS              Number of obs =      32
-------------+------------------------------           F(  2,    29) =    7.08
       Model |    15822811     2  7911405.51            Prob > F      =  0.0031
    Residual |    32415407    29  1117772.65            R-squared     =  0.3280
-------------+------------------------------           Adj R-squared =  0.2817
       Total |    48238218    31  1556071.55            Root MSE      =  1057.2

------------------------------------------------------------------------------
         exp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tobacco |  -38.96759   37.98675    -1.03   0.313    -116.6592    38.72404
        alco |  -356.1345   111.3588    -3.20   0.003    -583.8887   -128.3802
       _cons |   3602.096    1062.97     3.39   0.002     1428.078    5776.114
------------------------------------------------------------------------------
```

Example 1 (PDF, pp. 1, 3-5)

b) Multicollinearity:

- R^2 versus $p(b_j)$. ✓
- Correlation coefficients:

$p(r_{exp,alco}) = 0.0011 < \alpha = 0.05$
$r^2_{exp,alco} = (-0.5510)^2 = 0.3036$

$r^2_{exp,alco} = 0.3036 < R^2 = 0.7381$ ✓

- F-tests based on auxilliary regressions:

M_{AE}: $\widehat{exp_i} = c_1 + c_2 tobacco_i + c_3 alco_i$
$R^2_{exp} = 0.3280$

$H_0: \gamma_j = 0, \forall j = 2,3$
$H_1: \gamma_j \neq 0, \exists j = 2,3$
} always put greek letters in hypothesis

$$F = \frac{R^2_{exp}/(k-2)}{(1-R^2_{exp})/(n-k+1)} = \frac{0.3280/(4-2)}{(1-0.3280)/(32-4+1)} = 7.078$$

(k from M_B → initial model)

$F_c(m_1 = k-2 = 2, m_2 = n-k+1 = 29, \alpha = 0.05) = 3.33$

$F > F_c$, we reject H_0 at $\alpha = 0.05$ and conclude from H_1.

Variable exp_i causes multicollinearity.

$$F = \frac{R^2_{alco}/(k-2)}{(1-R^2_{alco})/(n-k+1)} = \frac{0,3167/(4-2)}{(1-0,3167)/(32-4+1)}$$

$$F = 6,721$$

$$F_c(m_1 = k-2 = 2, m_2 = n-k+1 = 29, \alpha = 0,05) = 3,33$$

$F > F_c \Rightarrow$ We reject the null hypothesis at $\alpha = 0,05$ and conclude from H_1.

$$VIF_{alco} = \frac{1}{1-R^2_{alco}} = \frac{1}{1-0,3167} = 1,46$$

VIF is lower than 10 so we don't have issues with multicollinearity.

$$F = \frac{R^2_{tob}/(k-2)}{(1-R^2_{tob})/(n-k+1)} = \frac{0,1080/(4-2)}{(1-0,1080)/(32-4+1)}$$

$$F = 1,756$$

$$F_c(m_1 = k-2 = 2, m_2 = n-k+1 = 29, \alpha = 0,05) = 3,33$$

We cannot reject the null hypothesis at $\alpha = 0,05$.

$$VIF_{tob} = \frac{1}{1-R^2_{tob}} = \frac{1}{1-0,1080} = 1,12$$

HW: Do the other two auxilliary regressions.

- Calculation of VIF_j and Tol_j:

$$VIF_j = \frac{1}{1-R_j^2}$$

$$VIF_{exp} = \frac{1}{1-R_{exp}^2} = \frac{1}{1-0.3280} = 1.488 < 10 \checkmark$$

we have multicollinearity since it is above 1, but it doesn't cause any issues.

$$Tol_j = 1 - R_j^2 = \frac{1}{VIF_j}$$

$$Tol_{exp} = \frac{1}{1.488} = 0.672 > 0.1 \checkmark$$

Stata: commands ESTAT VIF and COLLIN.

Example 2 (PDF, pp. 5-8) in Stata only.

```
. scalar R2exp=e(r2)
. scalar Fexp=(R2exp/(4-2))/((1-R2exp)/(32-4+1))
. scalar pFexp=Ftail(4-2,32-4+1,Fexp)
. scalar vifexp=1/(1-R2exp)
. scalar toleranceexp=1/vifexp

. display Fexp, pFexp, vifexp, toleranceexp
7.0778306 .00313851 1.4881262 .671986

. regress tobacco exp alco

      Source |       SS       df       MS              Number of obs =      32
-------------+------------------------------           F(  2,    29) =    1.76
       Model |  90.5229799     2    45.26149           Prob > F      =  0.1906
    Residual |  747.496685    29  25.7757477           R-squared     =  0.1080
-------------+------------------------------           Adj R-squared =  0.0465
       Total |  838.019664    31  27.0328924           Root MSE      =   5.077

------------------------------------------------------------------------------
     tobacco |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |  -.0008986    .000876    -1.03   0.313    -.0026902     .000893
        alco |    .458065   .6160993     0.74   0.463    -.8019995    1.718129
       _cons |   28.65308   2.839478    10.09   0.000     22.84569    34.46046
------------------------------------------------------------------------------

. scalar R2tobacco=e(r2)
. scalar Ftobacco=(R2tobacco/(4-2))/((1-R2tobacco)/(32-4+1))
. scalar pFtobacco=Ftail(4-2,32-4+1,Ftobacco)
. scalar viftobacco=1/(1-R2tobacco)
. scalar tolerancetobacco=1/viftobacco

. display Ftobacco, pFtobacco, viftobacco, tolerancetobacco
1.7559719 .19061119 1.1211015 .89197989

. regress alco exp tobacco

      Source |       SS       df       MS              Number of obs =      32
-------------+------------------------------           F(  2,    29) =    6.72
       Model |  30.8785495     2  15.4392748           Prob > F      =  0.0040
    Residual |  66.6360968    29  2.29779644           R-squared     =  0.3167
-------------+------------------------------           Adj R-squared =  0.2695
       Total |  97.5146463    31  3.14563375           Root MSE      =  1.5158

------------------------------------------------------------------------------
        alco |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |  -.0007321   .0002289    -3.20   0.003    -.0012003   -.0002639
     tobacco |   .0408345   .0549226     0.74   0.463    -.0714948    .1531638
       _cons |   2.766477   1.725856     1.60   0.120    -.7632956     6.29625
------------------------------------------------------------------------------

. scalar R2alco=e(r2)
. scalar Falco=(R2alco/(4-2))/((1-R2alco)/(32-4+1))
. scalar pFalco=Ftail(4-2,32-4+1,Falco)
. scalar vifalco=1/(1-R2alco)
. scalar tolerancealco=1/vifalco

. display Falco, pFalco, vifalco, tolerancealco
6.7191656 .00400199 1.4633907 .6833445
```

```
. qui regress life exp tobacco alco
. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
         exp |      1.49    0.671986
        alco |      1.46    0.683345
     tobacco |      1.12    0.891980
-------------+----------------------
    Mean VIF |      1.36

. collin exp tobacco alco, corr rinv
(obs=32)

  Collinearity Diagnostics

                       SQRT                   R-
  Variable      VIF     VIF    Tolerance    Squared
----------------------------------------------------
       exp      1.49    1.22    0.6720      0.3280
   tobacco      1.12    1.06    0.8920      0.1080
      alco      1.46    1.21    0.6833      0.3167
----------------------------------------------------
  Mean VIF      1.36

                         Cond
        Eigenval         Index
---------------------------------
    1     1.7677          1.0000
    2     0.7842          1.5014
    3     0.4481          1.9862
---------------------------------
 Condition Number         1.9862
 Eigenvalues & Cond Index computed from deviation sscp (no intercept)
 Det(correlation matrix)    0.6211

Inverse of correlation matrix

               exp     tobacco        alco
    exp   1.4881262
tobacco   .24169899   1.1211015
   alco   .75351706  -.17517811   1.4633907
```

tolerance

■

Example 2: We analyse production functions for 81 manufacturing companies in the computer manufacturing industry for a given year (the data are provided in Stata Data file `production.dta`, while the programming code is given in Stata Do file `production-commands-l08.do`). We have cross-section data available for the following variables:

- value added as a proxy for the product (Q; in 1,000 monetary units);
- average number of employed workers as a proxy for labour (L);
- sum of tangible and intangible assets as a proxy for capital (K; in 1,000 monetary units).

a) Check validity of the assumption on normality of the disturbances by applying the Jarque–Bera test in the model of linear and log-linear production function.
b) Check validity of the assumption on (absence of) multicollinearity in the model of linear and log-linear production function.

Computer printout of the results in Stata:

Normality of the disturbances:

```
. regress q l k

      Source |       SS       df       MS              Number of obs =      81
-------------+------------------------------           F(  2,    78) =   52.90
       Model |  6.9350e+12     2  3.4675e+12           Prob > F      =  0.0000
    Residual |  5.1130e+12    78  6.5551e+10           R-squared     =  0.5756
-------------+------------------------------           Adj R-squared =  0.5647
       Total |  1.2048e+13    80  1.5060e+11           Root MSE      =  2.6e+05

------------------------------------------------------------------------------
           q |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           l |   9687.383   3640.852     2.66   0.009     2439.003    16935.76
           k |    2.27941   .7553228     3.02   0.003      .775678    3.783142
       _cons |  -11875.29   34865.13    -0.34   0.734    -81286.43    57535.85
------------------------------------------------------------------------------

. predict eq, resid
. sum eq, detail

                        Residuals
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -928124.7      -928124.7
 5%    -205781.2      -693306.2
10%    -149066.2        -520897       Obs                  81
25%     -30861.5      -258838.8       Sum of Wgt.          81

50%    -3095.372                      Mean           -.0010632
                      Largest         Std. Dev.       252809.1
75%     7945.411        350021
90%      60822.5      778072.3       Variance        6.39e+10
95%     166468.2      816731.7       Skewness        1.55663
99%      1310726       1310726       Kurtosis        14.98543

. scalar obs=r(N)
. scalar slin=r(skewness)
. scalar klin=r(kurtosis)

. scalar jblin=obs*(slin^2/6+(klin-3)^2/24)
. display jblin
517.53233

. display chi2tail(2,jblin)
4.16e-113

. jb6 eq
Jarque-Bera normality test:  517.5 Chi(2)  4.e-113
Jarque-Bera test for Ho: normality: (eq)

. gen lq=log(q)
. gen ll=log(l)
. gen lk=log(k)
```

normal distribution
0
3

=) insert in formula

$p = 4.16e\text{-}113 < 0,05$ =) We reject the null hypothesis

this could be missing on the exam

```
. regress lq ll lk

      Source |       SS       df       MS              Number of obs =      81
-------------+------------------------------           F(  2,    78) =  190.75
       Model |  178.261263     2  89.1306313           Prob > F      =  0.0000
    Residual |    36.44752    78  .467275898           R-squared     =  0.8302
-------------+------------------------------           Adj R-squared =  0.8259
       Total |  214.708783    80  2.68385978           Root MSE      =  .68358

------------------------------------------------------------------------------
          lq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ll |   .9645479   .1199229     8.04   0.000     .7257997    1.203296
          lk |   .1885438   .0673358     2.80   0.006     .0544886     .322599
       _cons |   7.546026   .4617465    16.34   0.000      6.62676    8.465293
------------------------------------------------------------------------------

. predict elq, resid
. sum elq, detail

                         Residuals
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -1.225007      -1.225007
 5%    -1.015222      -1.073147
10%    -.8845653      -1.022881       Obs                  81
25%    -.4654464      -1.018599       Sum of Wgt.          81

50%    -.0882534                      Mean           -1.24e-09
                        Largest       Std. Dev.       .674977
75%     .4299134       1.258059
90%     1.020899       1.329013       Variance        .455594
95%     1.214762       1.409848       Skewness        .4283599
99%     1.786791       1.786791       Kurtosis        2.619945

. scalar obs=r(N)
. scalar slog=r(skewness)
. scalar klog=r(kurtosis)

. scalar jblog=obs*(slog^2/6+(klog-3)^2/24)
. display jblog
2.9646371

. display chi2tail(2,jblog)
.22711051

. jb6 elq
Jarque-Bera normality test:  2.965 Chi(2)  .2271
Jarque-Bera test for Ho: normality: (elq)
```

The log function helped us to not reject normality.

$p = .22711051$

↳ we don't reject the null.

Multicollinearity:

```
. collin l k, corr
(obs=81)

  Collinearity Diagnostics

                        SQRT                    R-
  Variable      VIF     VIF    Tolerance    Squared
----------------------------------------------------
         l      3.55    1.88     0.2817      0.7183
         k      3.55    1.88     0.2817      0.7183
----------------------------------------------------
  Mean VIF      3.55
```

neither of these models have problems with collinearity.

Eigenval	Cond Index
1 1.8476	1.0000
2 0.1524	3.4813

Condition Number 3.4813
Eigenvalues & Cond Index computed from deviation sscp (no intercept)
Det(correlation matrix) 0.2817

```
. collin ll lk, corr
(obs=81)
```

Collinearity Diagnostics

Variable	VIF	SQRT VIF	Tolerance	R-Squared
ll	3.45	1.86	0.2896	0.7104
lk	3.45	1.86	0.2896	0.7104
Mean VIF	3.45			

Eigenval	Cond Index
1 1.8428	1.0000
2 0.1572	3.4242

Condition Number 3.4242
Eigenvalues & Cond Index computed from deviation sscp (no intercept)
Det(correlation matrix) 0.2896

■

University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

4.2 Multicollinearity

↳ glej 176. stran (Stata izpis)
za predelavo poglavja

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION AMBA ACCREDITED

Economic background

relationship among the explanatory variables

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

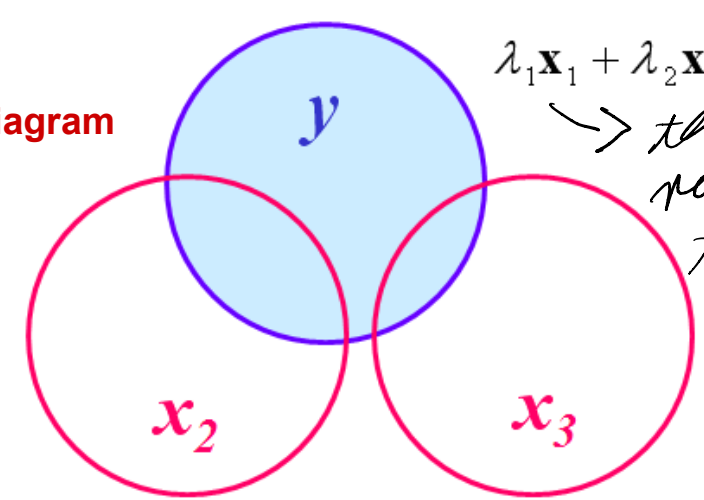
Assumption:

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

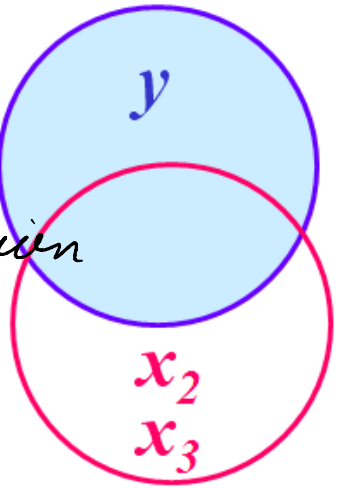
$$\lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2 + \ldots + \lambda_k \mathbf{x}_k = 0$$

→ this is not possible if the assumption holds

Venn diagram



No (multi)collinearity



Perfect (multi)collinearity

X should be correlated to y, while X among themselves should not be correlated.

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 $\mathbf{X}^T\mathbf{X}$ is singular, i.e. it is not possible to calculate its inverse matrix.** ⇒ determinant of X^TX is 0.
2. **$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is not defined.** ↳ it is not possible to calculate b
3. **Multicollinearity relates only to linear dependence among explanatory variables, not non-linear.**

x, x^2 is okay since they don't have a linear relationship.

$\ln x, \ln x^2$ is not okay because they have a linear relationship.

$(\ln x, 2 \ln x)$

Meaning of the assumption

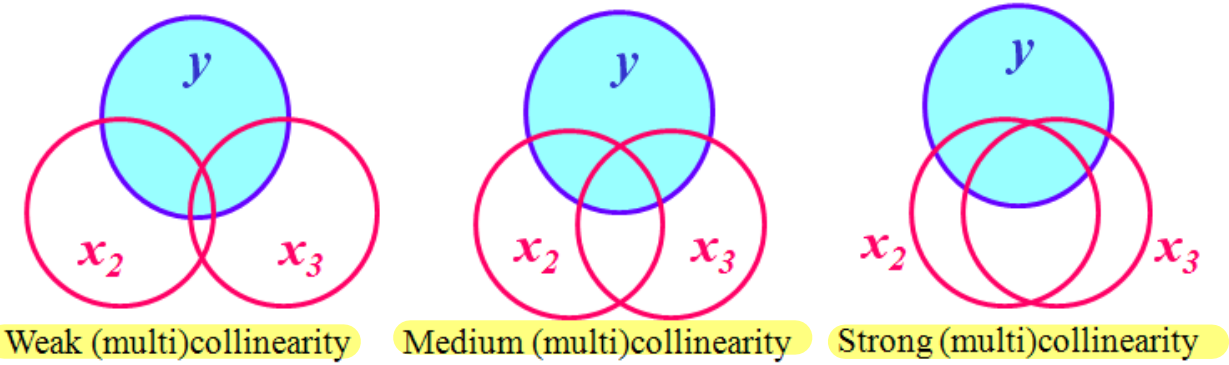
II. Imperfect multicollinearity → most often in practice

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

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.



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.

↳ this is not good because the model needs to be as robust as possible

Multicollinearity

- **Effect of multicollinearity on testing:**

$\downarrow t_j = \frac{b_j - \beta_j}{\sqrt{Var(b_j)}\uparrow}$ vs. t_c Slide 15

↳ t_c is more difficult to reject the null hypothesis when the multicollinearity increases

Meaning of the assumption

High R^2 => high share of y variability can be explained by x.

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

any regression which is not the main regression

3

We use "auxilliary regressions":

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

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

all other x without x_j

and calculate Variance Inflation Factors – *VIF*:

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

Slide 18:

- **F-test based on an auxilliary regression:**

H_0: $\beta_j = 0$, $\forall j = 2, \dots, k$ (no multicollinearity) => null hypothesis
H_1: $\beta_j \neq 0$, $\exists j = 2, \dots, k$ (multicollinearity)
↳ alternate hypothesis

- **Variance inflation factor:**

$1 \leq VIF_j \leq \infty$
↑ no multicollinearity
↑ perfect multicollinearity

$R_j^2 = ?$

Rule of thumb: $VIF_j > 10$, serious multicollinearity issues that need to be addressed. => important

$R_j^2 > 0.9$, whereas ideally it should be 0.

Some experts use tolerance instead of VIF

- **Tolerance:**

$Tol_j = \frac{1}{VIF_j} = 1 - R_j^2$

Rule of thumb: $Tol_j < 0.1$, serious multicollinearity issues that need to be addressed.

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.
   ↳ because the least squares estimator is still blue
2. Drop one or more explanatory variables, causing the highest level of multicollinearity in the model.
   ↳ some variables need to stay in the model because of theory
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.

University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

Homoscedasticity

4.3 Heteroscedasticity

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION AMBA ACCREDITED

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:

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

expected value of $u \Rightarrow 0$

this doesn't change with observations

Homoscedasticity

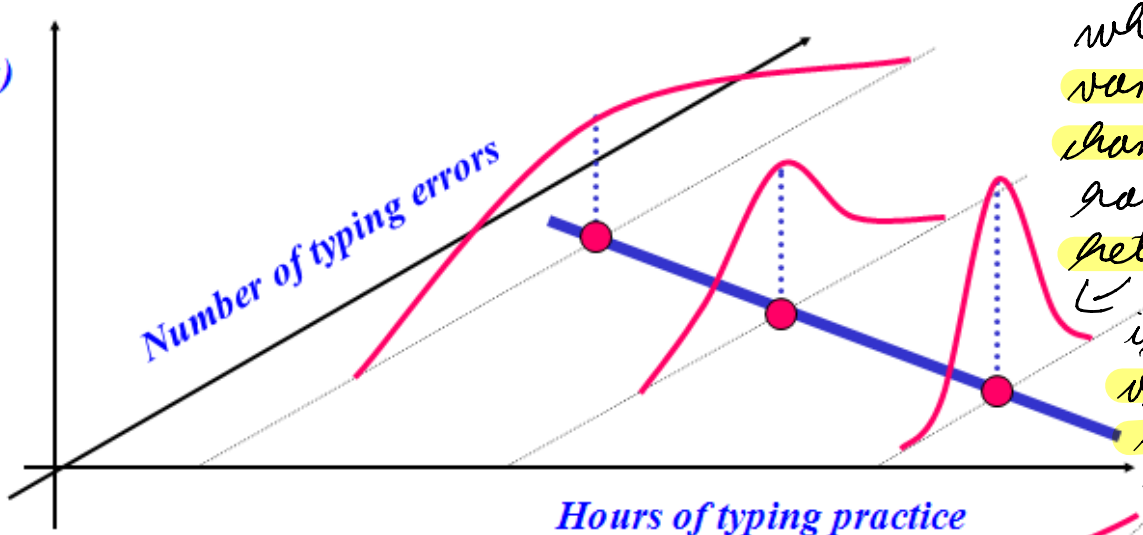
Slide 21.

$Var(u_i | x_i) = \sigma^2$ (homoscedasticity)

Vs.

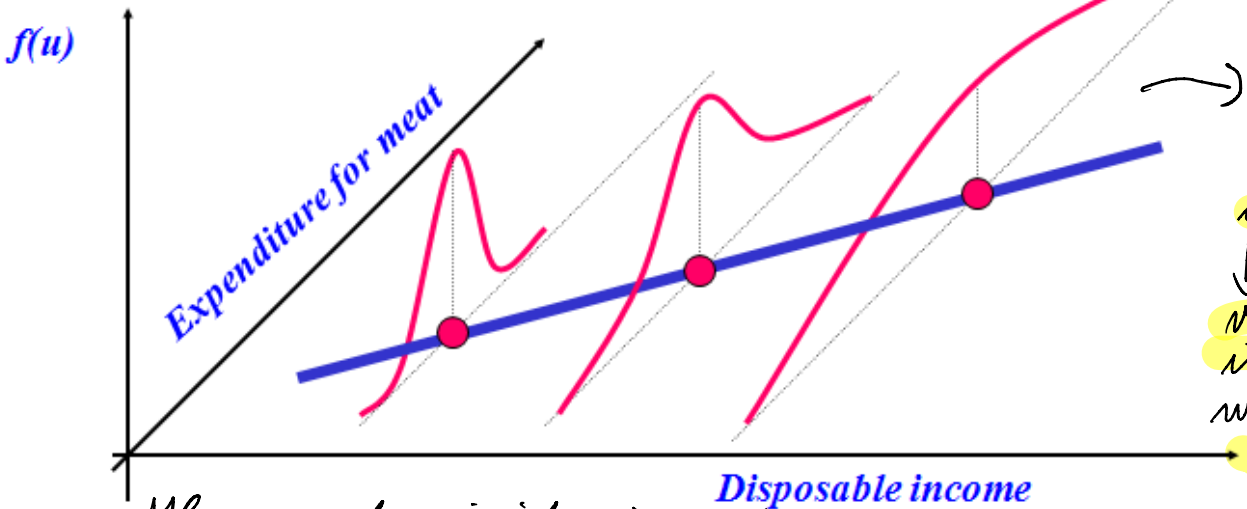
$Var(u_i | x_i) = \sigma_i^2$ (heteroscedasticity)

Meaning of the assumption



when the variability is changing we have heteroscedasticity

if the number of hours typing increases then the variability decreases.



this is also an example of heteroscedasticity

variability increases when income increases

Homoscedasticity is when variability doesn't change at different values.

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

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

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

mathematical expectation of the value

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

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

$\to 0$

$X^T X \beta \cdot (X^T X)^{-1} = \beta$

The estimator of regression coefficients remains unbiased!

Heteroscedasticity

Slides 24-25.

$$\text{Var} - \text{Cov}(u) = W = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix} \neq \sigma^2 \cdot I$$

this isn't equal to σ^2

$$b = (X^T X)^{-1} X^T y = (X^T X)^{-1} X^T (X\beta + u) =$$
$$= \underbrace{(X^T X)^{-1} X^T X}_{I} \beta + (X^T X)^{-1} X^T u =$$
$$= \beta + (X^T X)^{-1} X^T u \Rightarrow b - \beta \text{ is equal to this}$$

$$\left.\begin{aligned} b - \beta &= (X^T X)^{-1} X^T u \\ (b-\beta)^T &= u^T \underbrace{X (X^T X)^{-1}}_{\text{symmetry}} \end{aligned}\right\} \longrightarrow \text{Var} - \text{Cov}(b)$$

as $(ABC)^T = C^T B^T A^T$.

Consequences of heteroscedasticity

2

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

vrednost iz prejšnje strani

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

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

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

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

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

↳ we lose efficiency

Slide 25:

Homoscedasticity, $W = \sigma^2 \cdot I$:

$$\text{Var-Cov}(b) = (X^T X)^{-1} X^T \sigma^2 I \, X (X^T X)^{-1} =$$
$$= \sigma^2 \cdot (X^T X)^{-1} \underbrace{X^T X (X^T X)^{-1}}_{I}$$
$$= \sigma^2 \cdot (X^T X)^{-1}$$

Heteroscedasticity, $W \neq \sigma^2 \cdot I$:

$$\text{Var-Cov}(b) = (X^T X)^{-1} X^T W X (X^T X)^{-1}$$

"sandwich" estimator

↳ because of this we lose efficiency

Consequences of heteroscedasticity

that is an issue

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!

Slides 25-26:

Least squares estimator:

1. Estimator of the regression coefficients β_j:

$$b = (X^T X)^{-1} X^T y$$ ✓ ⇒ this is fine

2. Estimator of the variance of stoch. var. u:

$$s_e^2 = \frac{RSS}{n-k}$$ ✗ ⇒ this becomes biased

3. Estimator of the variance-covariance matrix of regression coefficient estimates:

$$\text{var-cov}(b) = s_e^2 \cdot (X^T X)^{-1}$$ ✗ ⇒ this becomes biased

We also use $e = \hat{u}$.

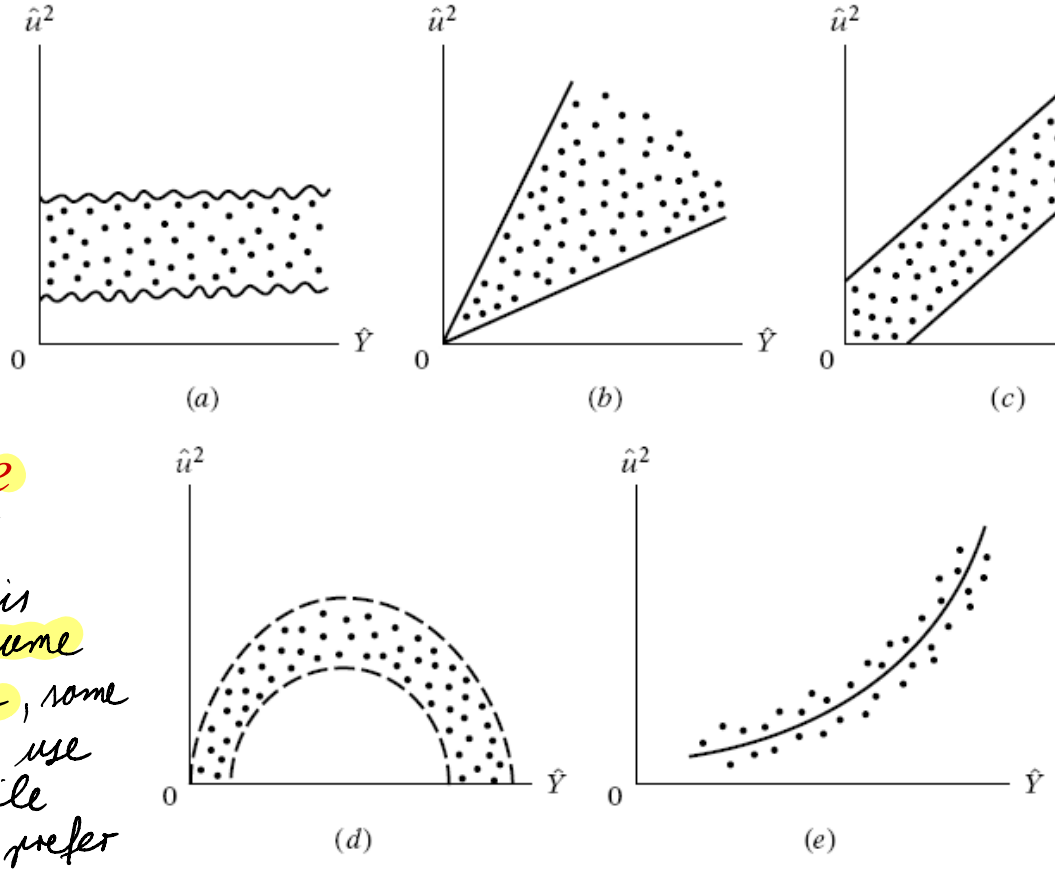
Verifying the validity of the assumption

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

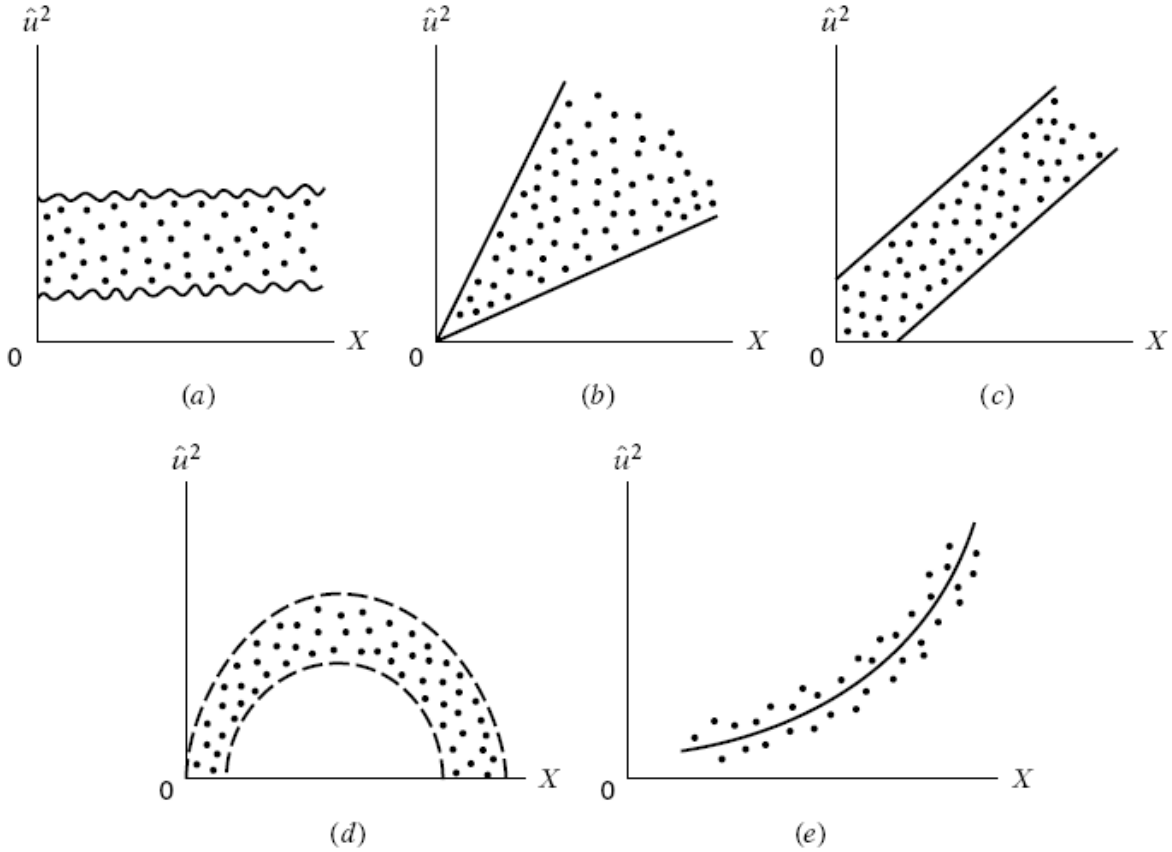


any pattern like this is an indicator of heteroscedasticity ☺ if there is no pattern we have homoscedasticity.

$\hat{u} \equiv e$

this is the same value, some people use e, while others prefer $\hat{u}$.

Verifying the validity of the assumption



$\hat{u} \equiv e$

Verifying the validity of the assumption

White assumed that heteroscedasticity is caused by explanatory variables, their squares and their cross-products.

2. White test (1980)

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

a)

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

if α are 0 we have homoscedasticity

b)

We calculate residuals e_i and estimate the auxilliary regression:

α become a

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

c)

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

H_1: At least one α different from 0

Lagrange multiplier test

d)

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

$\theta > \chi_c^2$ → **reject H_0**

Solutions if the assumption is violated

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

Problem	Heteroscedasticity	Autocorrelation
Problem management (once detected)	Improvement of the model specification ⇒ adding additional variables (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

Slide 32:

Our assumption: $u \sim IID$

independence, identical distribution → homoscedasticity

Problem	Heteroscedasticity	Autocorrelation
Problem management (once detected)	Robust variance estimators (Huber/White variance estimator)	HAC variance estimators (Newey–West robust variance estimator)
	$u \sim IID$	$u \sim IID$
	Estimator loosens the assumption on identical distribution.	Estimator loosens both assumptions (on independence and identical distribution).
	Approach does not affect the regression coefficient estimates. b is left unchanged	
	Standard errors regain unbiasedness. → important	
	Regression coefficient estimator does not necessarily regain efficiency (standard errors are not necessarily the lowest possible). → this is not that important	
	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
------------------------------------------------------------------------------
```

EFMD EQUIS

Slides 33–34:

Example, variables:

- hm1: harmonized money aggregate M1;
- ppr: income of households;
- rvp: interest rate on demand deposits;
- rvv: interest rate on short-term deposits;
- czp: consumer price index.

everything that is greyed out is problematic

Robust variance estimator application

→ White test

```
. whitetst

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

We reject H_0.

```
. 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
```

H_0: homoscedasticity

H_1: heteroscedasticity

hm1	Coef.	Robust 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

b stay the same

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
```

University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

4.4 Autocorrelation

↳ assumption: there is no autocorrelation

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION AMBA ACCREDITED

Autocorrelation

Is about covariance or correlation between stochastic variable u and lagged stochastic variable u, i.e.:

t	u_t	u_{t-1}
1	u_1	–
2	u_2	u_1
3	u_3	u_2
4	u_4	u_3
5	u_5	u_4

Slide 38.

It is a time-series related phenomenon, as it has to do with ordering of observations.

time is linear => year 2022 comes after 2021

Meaning of the assumption

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

Assumption

one lag being correlated with u causes autocorrelation

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

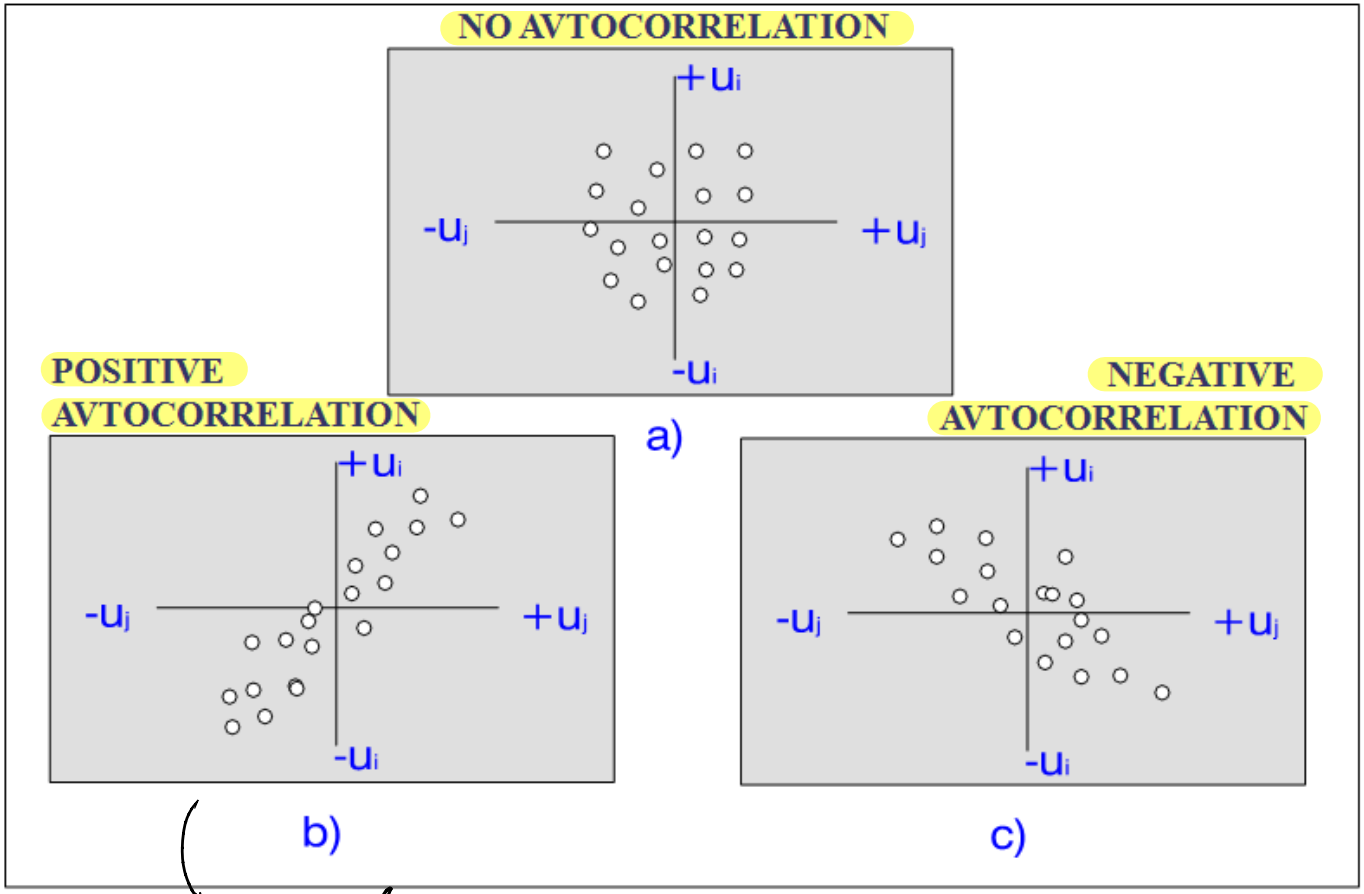
$$i \neq j \qquad\qquad i \neq j$$

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:



most common

Meaning of the assumption

Causes of autocorrelation

Genuine

1. **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. ***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

if we don't put x_4 in the model it exists and goes into the disturbance term - u

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

this causes biasness and autocorrelation

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

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

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.

the mean

the standard deviation and covariance

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)

autoregressive process of order 1

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

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

2 **Second-order autocorrelation**

autoregressive scheme of order 2

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

3 ***p*–th order autocorrelation**

autoregressive scheme of order p

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

Consequences of autocorrelation

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

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

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

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

$= \beta + (X^TX)^{-1} \cdot X^T \cdot 0$ ↳0

$= \beta$

The estimator of regression coefficients remains unbiased!

the b are unaffected by autocorrelation

Consequences of autocorrelation

2

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

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

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

the same as in case of heteroscedasticity

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

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

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

↳ best linear unbiased estimator

Autocorrelation

$Var - Cov(u) = W \neq \sigma^2 \cdot I$ Slides 44-45.

Example, AR(1):

$$W = \frac{1}{1-\rho^2} \cdot \begin{bmatrix} \sigma^2 & \rho^1 & \cdots & \rho^{T-1} \\ \rho^1 & \sigma^2 & \cdots & \rho^{T-2} \\ \rho^2 & \rho^1 & \cdots & \rho^{T-3} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \cdots & \sigma^2 \end{bmatrix}$$

ρ - coefficient of first-order autocorrelation

Homoscedasticity:

$$W = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 & \\ & & \ddots \\ 0 & & \sigma_e^2 \end{bmatrix}$$

Slide 45:

No autocorrelation, $W = \sigma^2 \cdot I$:

$$\begin{aligned} \text{Var-Cov}(b) &= (X^T X)^{-1} X^T \sigma^2 I \, X (X^T X)^{-1} = \\ &= \sigma^2 \cdot (X^T X)^{-1} \underbrace{X^T X (X^T X)^{-1}}_{I} \\ &= \sigma^2 \cdot (X^T X)^{-1} \end{aligned}$$

Autocorrelation, $W \neq \sigma^2 \cdot I$:

$$\text{Var-Cov}(b) = (X^T X)^{-1} X^T W X (X^T X)^{-1}$$

"sandwich" estimator

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!

↳ we cannot do any testing because of that

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.

Slides 45-46:

Least squares estimator:

1. Estimator of the regression coefficients β_j.

   $b = (X^T X)^{-1} X^T y$ ✓ → this is fine

2. Estimator of the variance of stoch. var. u:

   $s_e^2 = \frac{RSS}{n-k}$ ✗ → this becomes biased

3. Estimator of the variance-covariance matrix of regression coefficient estimates:

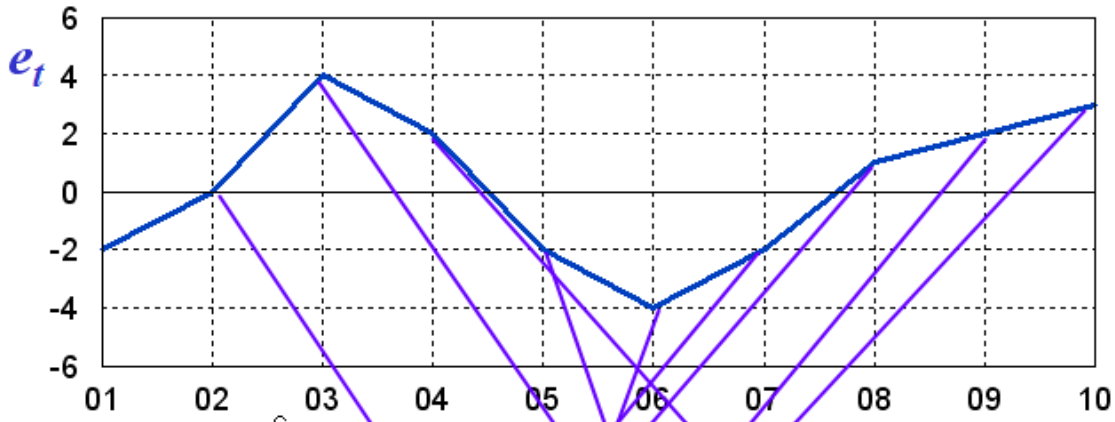
   $\text{var-cov}(b) = s_e^2 \cdot (X^T X)^{-1}$ ✗ → this becomes biased

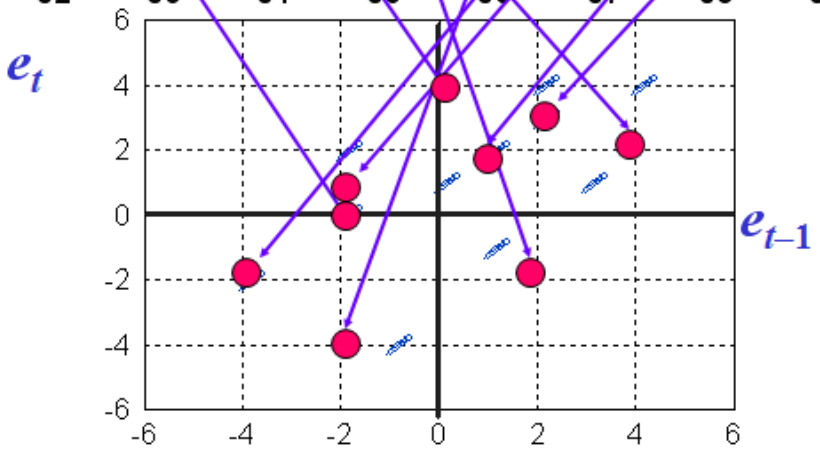
exactly the same story as with heteroscedasticity

↳ we have the same issues

Verifying the validity of the assumption

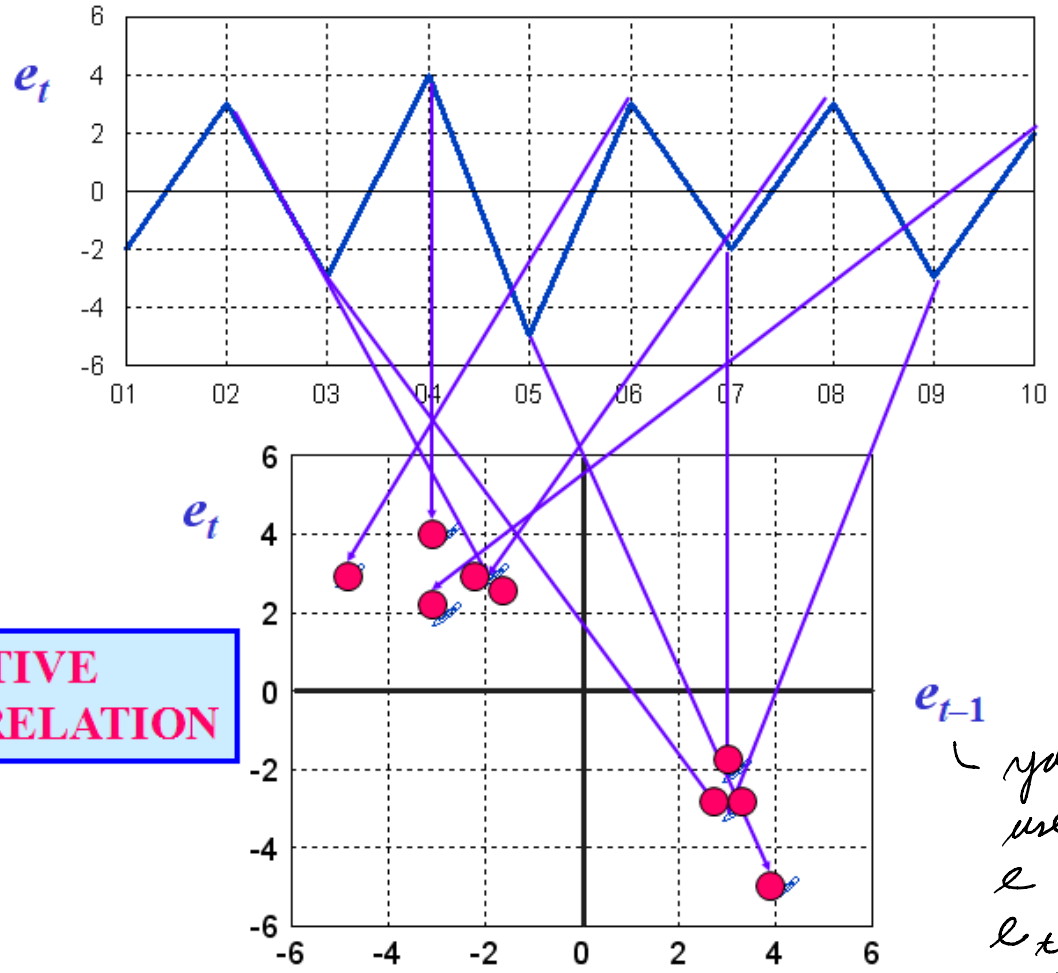
1. Graphic method of detecting autocorrelation





POSITIVE AUTOCORRELATION

Verifying the validity of the assumption



NEGATIVE AUTOCORRELATION

you can use other e but e_{t-1} is the most common

Verifying the validity of the assumption

This is in the handbook

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

↳ auxilleary regression to calculate R^2
then we use R^2 to calculate test statistic

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**

↳ we have autocorrelation

Estimate auxilliary regression:

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

$LM = (n-p)R^2 \sim \chi^2_{(p)}$ — Vilenič stran 35

sample size ⇒ we lose p observations ⇒ see next page

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

t	e_t	e_{t-1}
1	e_1	—
2	e_2	e_1
3	e_3	e_2
4	e_4	e_3

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.	

=) this helps if we have spurious autocorrelation

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)
	$u \sim IID$	$u \sim IID$
	Estimator loosens the assumption on identical distribution.	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

Slide 52:
Our assumption: $u \sim IID$
independence → no autocorrelation
identical distribution
HAC: heteroscedasticity and autocorrelation consistent (estimator of variance).
HAC estimators reduce heteroscedasticity as well as autocorrelation.

HAC variance estimator application

the first couple of lags are most important.

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
------------------------------------------------------------------------------
```

EFMD EQUIS

Slides 53-55:

Example, variables:

- hm1: harmonized money aggregate M1;
- ppr: income of households;
- rvp: interest rate on demand deposits;
- rvv: interest rate on short-term deposits;
- czp: consumer price index.

↳ if we have an issue with autocorrelation everything that is greyed out is problematic.

HAC variance estimator application

Breusch Godfrey test

```
. 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

we reject the null in all 12 lags

this means that we have autocorrelation present

this may be missing on the exam

HAC variance estimator application

HAC variance estimator command

```
. 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
```

standard errors are now unbiased

hm1	Coef.	Newey-West 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 2022