# Econometrics

**Natalia Nehrebecka**

Part 4

# Agenda

- Diagnostic tests

# Classical Linear Regression Model Assumptions

1. **Linear in Parameters**
   - *The model in the population can be written as*

$$y_i = \beta_1 + \beta_2 X_{2i} + \cdots \beta_K X_{Ki} + \varepsilon_i$$

     - *where* $\beta_1, \ldots, \beta_K$ *are unknown parameters of interest and* $\varepsilon_i$ *is an unobservable random error of disturbance term*

2. **Non-random Explanatory Variables**
   - *Explanatory variables* $X_{2i}, \ldots X_{Ki}$ *are non-random for* $i = 1, 2, \ldots, N$
3. **Zero Mean of Error Term**
   - *The error term* $\varepsilon_i$ *has an expected value of zero* $E(\varepsilon_i)=0$
4. **No autocorrelation**
   - $Cov(\varepsilon_i, \varepsilon_j) = 0 \quad for\ i \neq j$
5. **Homoscedasticity**
   - $Var(\varepsilon_i) = \sigma^2$ for $i = 1,2, \ldots, N$

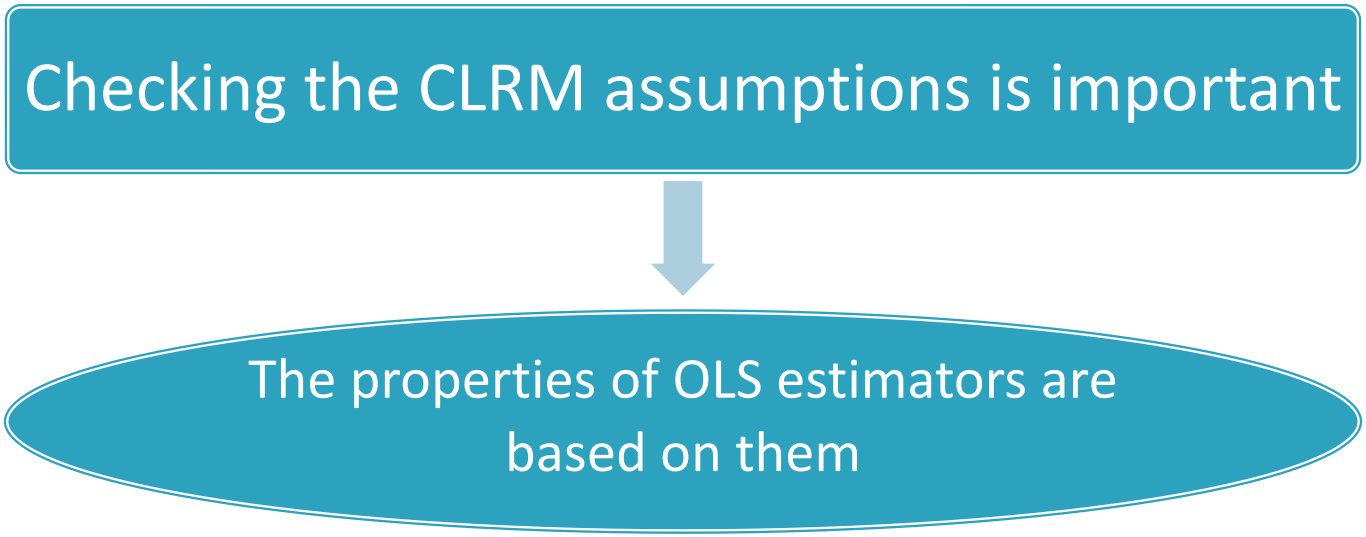
# Classical Linear Regression Model Assumptions

- Additional assumption:

$$\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathrm{I})$$

# Diagnostic tests

- Diagnostic tests are used to verify the CLRM assumptions




- Tests are used after estimating the model

# Regression specification error test (RESET)

- $H_0$: $y_i = X_i\beta + \varepsilon_i$
- $H_1$: $y_i = f(X_i\beta) + \varepsilon_i$

- The idea of RESET is to include squares and possibly higher order fitted values in the regression

$$y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_K X_{Ki} + \boxed{\alpha_1 \hat{y}_i^2} + \boxed{\alpha_1 \hat{y}_i^3} + \varepsilon_i$$

**Test for the exclusion of these terms**

# Ex 1.

```
. des

Contains data from C:\Users\U138301\Downloads\Nowy folder (4)\hprice1.dta
  obs:            88
 vars:            10                          17 Mar 2002 12:21
 size:         2,816
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
price           float   %9.0g                 house price, $1000s
assess          float   %9.0g                 assessed value, $1000s
bdrms           byte    %9.0g                 number of bdrms
lotsize         float   %9.0g                 size of lot in square feet
sqrft           int     %9.0g                 size of house in square feet
colonial        byte    %9.0g                 =1 if home is colonial style
lprice          float   %9.0g                 log(price)
lassess         float   %9.0g                 log(assess
llotsize        float   %9.0g                 log(lotsize)
lsqrft          float   %9.0g                 log(sqrft)
-------------------------------------------------------------------------------
```

# Ex 1.

```
. reg price lotsize sqrft bdrms

      Source |       SS           df       MS      Number of obs   =        88
-------------+----------------------------------   F(3, 84)        =     57.46
       Model |  617130.701         3  205710.234   Prob > F        =    0.0000
    Residual |  300723.805        84   3580.0453   R-squared       =    0.6724
-------------+----------------------------------   Adj R-squared   =    0.6607
       Total |  917854.506        87  10550.0518   Root MSE        =    59.833

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lotsize |   .0020677   .0006421     3.22   0.002     .0007908    .0033446
       sqrft |   .1227782   .0132374     9.28   0.000     .0964541    .1491022
       bdrms |   13.85252   9.010145     1.54   0.128    -4.065141    31.77018
       _cons |  -21.77031   29.47504    -0.74   0.462    -80.38466    36.84405
------------------------------------------------------------------------------
```

1. Check if the function form is correct.

# Ex 1.

```
. ovtest

Ramsey RESET test using powers of the fitted values of price
       Ho:  model has no omitted variables
                  F(3, 81) =      4.26
                  Prob > F =      0.0076
```

Evidence for misspecification

# Regression specification error test (RESET)

- **CLRM assumptions:**

  **(a) Which are not satisfied?**

  **(b) What are the consequences for statistical deduction?**

  **(c) How can we solve this problem?**

# Regression specification error test (RESET)

- **CLRM assumptions:**

  **(a) Which are not satisfied?**

1. **Linear in Parameters**
   - *The model in the population can be written as*

$$y_i = \beta_1 + \beta_2 X_{2i} + \cdots \beta_K X_{Ki} + \varepsilon_i$$

     - *where* $\beta_1, \ldots, \beta_K$ *are unknown parameters of interest and* $\varepsilon_i$ *is an unobservable random error of disturbance term*

# Regression specification error test (RESET)

- **CLRM assumptions:**

  **(b) What are the consequences for statistical deduction?**

1. undermines the economic interpretation of the model (interpretation of estimated parameters)

2. it is impossible to prove the properties of the OLS

# Regression specification error test (RESET)

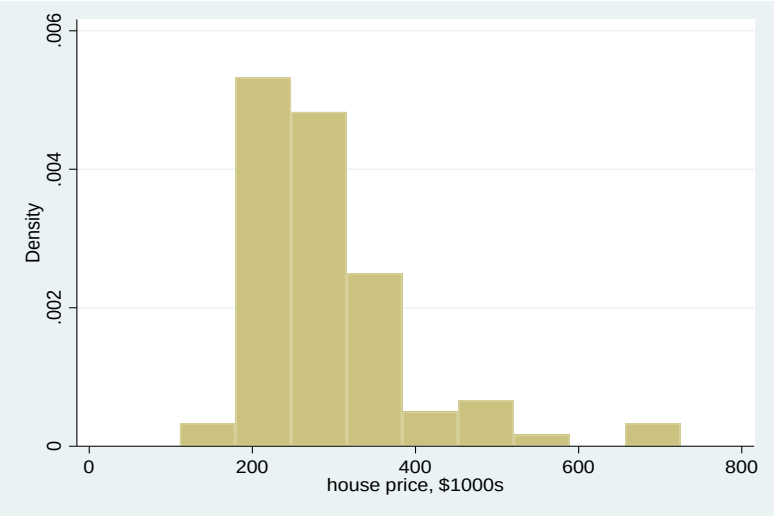
- **CLRM assumptions:**

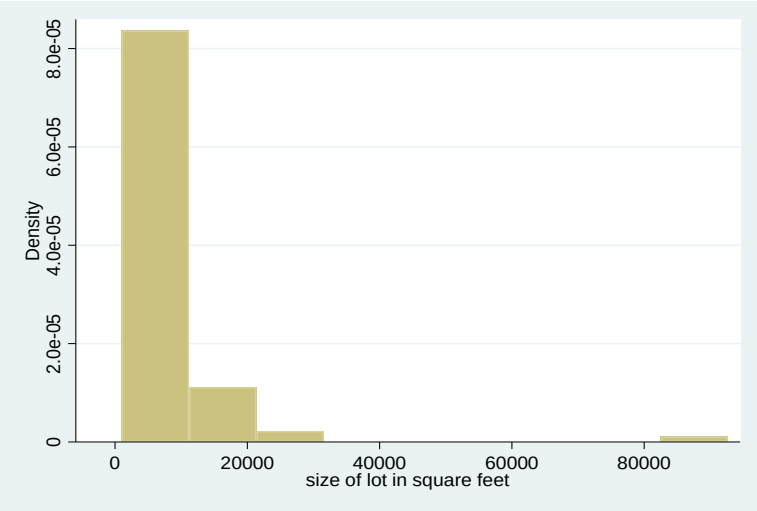
  **(c) How can we solve this problem?**

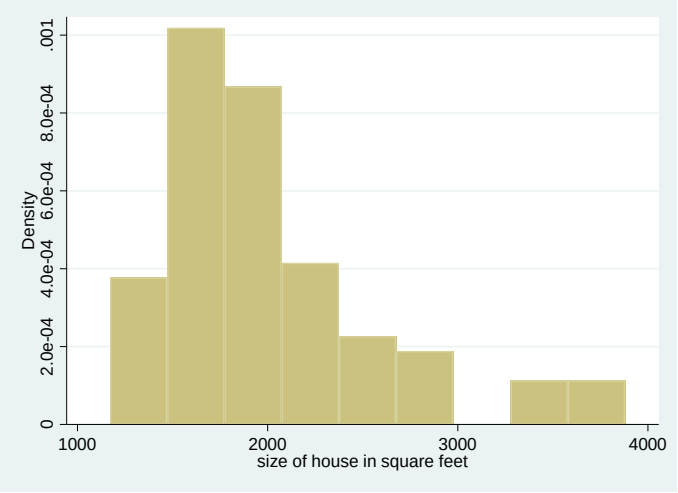
We can try to improve the functional form of the model entering the model:

- interactions between variables,
- make variable transformations,
- use the polynomial model

# Ex 1.


Density
0
.002
.004
.006
0
200
400
600
800
house price, $1000s


Density
0
2.0e-05
4.0e-05
6.0e-05
8.0e-05
0
20000
40000
60000
80000
size of lot in square feet


Density
0
2.0e-04
4.0e-04
6.0e-04
8.0e-04
.001
1000
2000
3000
4000
size of house in square feet

# Ex 1.

```
. reg lprice llotsize lsqrft bdrms

      Source |       SS           df       MS      Number of obs   =        88
-------------+----------------------------------   F(3, 84)        =     50.42
       Model |  5.15504028         3  1.71834676   Prob > F        =    0.0000
    Residual |  2.86256324        84  .034078134   R-squared       =    0.6430
-------------+----------------------------------   Adj R-squared   =    0.6302
       Total |  8.01760352        87  .092156362   Root MSE        =     .1846
------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    llotsize |   .1679667   .0382812     4.39   0.000     .0918404     .244093
      lsqrft |   .7002324   .0928652     7.54   0.000     .5155597    .8849051
       bdrms |   .0369584   .0275313     1.34   0.183    -.0177906    .0917074
       _cons |  -1.297042   .6512836    -1.99   0.050    -2.592191    -.001893
------------------------------------------------------------------------------
. ovtest
Ramsey RESET test using powers of the fitted values of lprice
       Ho:  model has no omitted variables
                 F(3, 81) =      2.45
                 Prob > F =      0.0692
```

# Jarque-Bera Test

- Null and alternative hypothesis:
  - $H_0: \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathrm{I})$
  - $H_1: \boldsymbol{\varepsilon} \nsim N(0, \sigma^2 \mathrm{I})$

# Ex 2.

```
. reg lprice llotsize lsqrft bdrms
      Source |       SS           df       MS      Number of obs   =        88
-------------+----------------------------------   F(3, 84)        =     50.42
       Model |  5.15504028         3  1.71834676   Prob > F        =    0.0000
    Residual |  2.86256324        84  .034078134   R-squared       =    0.6430
-------------+----------------------------------   Adj R-squared   =    0.6302
       Total |  8.01760352        87  .092156362   Root MSE        =     .1846

------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    llotsize |   .1679667   .0382812     4.39   0.000     .0918404     .244093
      lsqrft |   .7002324   .0928652     7.54   0.000     .5155597    .8849051
       bdrms |   .0369584   .0275313     1.34   0.183    -.0177906    .0917074
       _cons |  -1.297042   .6512836    -1.99   0.050    -2.592191    -.001893
------------------------------------------------------------------------------
```

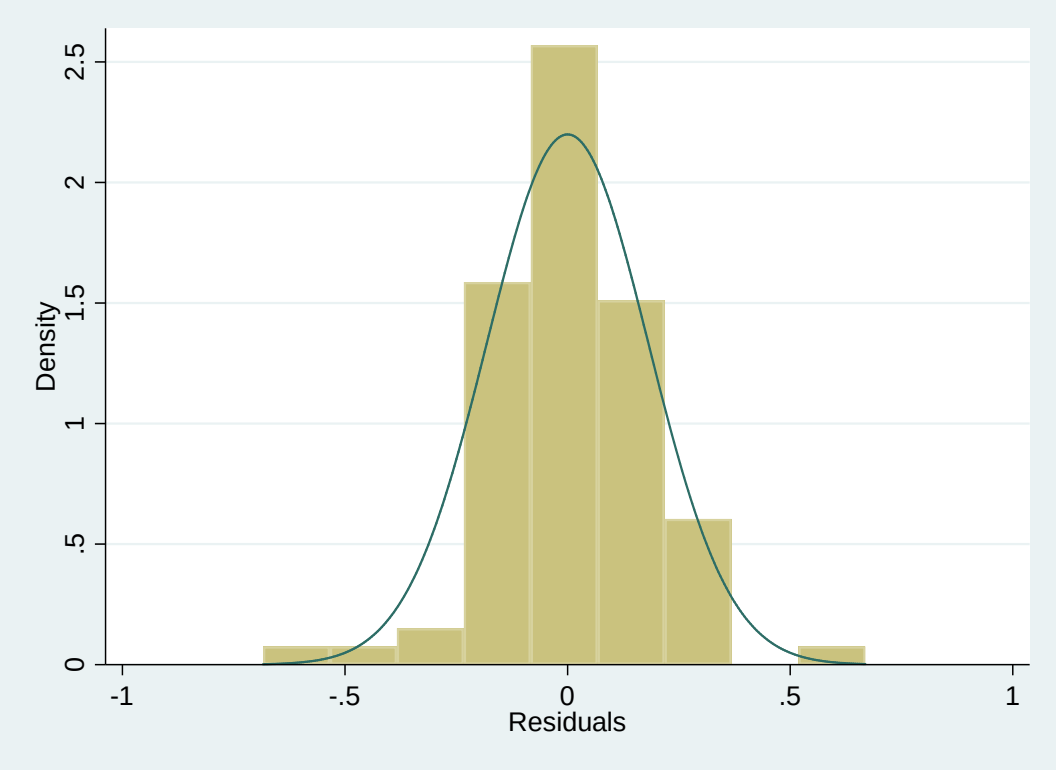
1. Check if residuals have normal distribution.

# Ex 2.

**sktest e**

Skewness/Kurtosis tests for Normality

| Variable | Obs | Pr(Skewness) | Pr(Kurtosis) | joint adj chi2(2) | **Prob>chi2** |
|---|---|---|---|---|---|
| e | 88 | 0.4444 | 0.0007 | 10.25 | **0.0059** |



# Jarque-Bera Test

- **CLRM assumptions:**

  **(a) Which are not satisfied?**

  **(b) What are the consequences for statistical deduction?**

  **(c) How can we solve this problem?**

# Jarque-Bera Test

- **CLRM assumptions:**

  **(a) Which are not satisfied?**

- Interpretation of the test result:
  - H0 rejected $\Longrightarrow$ distribution of the error term is not normal

- Additional assumption of **CRLM**: $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathrm{I})$

# Jarque-Bera Test

- **CLRM assumptions:**

  **(b) What are the consequences for statistical deduction?**

- Consequences of the rejection of H0:
  - assumption of normality of error terms is used when we derive the small sample distribution of estimators and tests statistics.
  - If this assumption is invalid we can only use the asymptotic distributions.

# Jarque-Bera Test

- **CLRM assumptions:**

  **(c) How can we solve this problem?**

- Enlargement of the sample, because for a larger sample the distributions will be closer to known asymptotic distributions

# Chow test

- Null and alternative hypothesis:
  - $\boldsymbol{H_0}: \beta_1 = \beta_2 = \cdots = \beta_m$
  - $\boldsymbol{H_1}$: $H_0$ is not true
- The form of the test statistic:

$$F = \frac{(S \quad - \sum_i \ S_i)/(K(m-1))}{\sum_i \ S_i)/(N - mK)}$$

  - Where:
  - $S_i$ is the residual sum of squares for model estimated on subsample $i$
  - $S$ is the residual sum of squares for model estimated on the full sample

**Critical statistics:**

$$F^* = F(K(m-1); (N - mK))$$

# Ex 3.

```
. reg lwage exper expersq educ  female

      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(4, 521)       =     86.69
       Model |  59.2711314         4  14.8177829   Prob > F        =    0.0000
    Residual |    89.05862       521   .17093785   R-squared       =    0.3996
-------------+----------------------------------   Adj R-squared   =    0.3950
       Total |  148.329751       525   .28253286   Root MSE        =    .41345

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |     .03891   .0048235     8.07   0.000      .029434    .0483859
     expersq |   -.000686   .0001074    -6.39   0.000     -.000897   -.0004751
        educ |   .0841361   .0069568    12.09   0.000     .0704692    .0978029
      female |  -.3371868   .0363214    -9.28   0.000    -.4085411   -.2658324
       _cons |    .390483   .1022096     3.82   0.000     .1896894    .5912767
------------------------------------------------------------------------------
```

1. Check if model parameters are stable.

# Ex 3.

```
. reg lwage exper expersq educ

      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(3, 522)       =     74.67
       Model |  44.5393713         3  14.8464571   Prob > F        =    0.0000
    Residual |   103.79038       522  .198832146   R-squared       =    0.3003
-------------+----------------------------------   Adj R-squared   =    0.2963
       Total |  148.329751       525   .28253286   Root MSE        =    .44591

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0410089   .0051965     7.89   0.000     .0308002    .0512175
     expersq |  -.0007136   .0001158    -6.16   0.000     -.000941   -.0004861
        educ |   .0903658    .007468    12.10   0.000     .0756948    .1050368
       _cons |   .1279975   .1059323     1.21   0.227    -.0801085    .3361035
------------------------------------------------------------------------------
```

# Ex 3.

```
. reg lwage exper expersq educ if female==1

      Source |       SS           df       MS      Number of obs   =       252
-------------+----------------------------------   F(3, 248)       =     24.01
       Model |  11.1496795         3  3.71655983   Prob > F        =    0.0000
    Residual |  38.3839276       248  .154773902   R-squared       =    0.2251
-------------+----------------------------------   Adj R-squared   =    0.2157
       Total |  49.5336071       251  .197345048   Root MSE        =    .39341

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0223715   .0066642     3.36   0.001     .0092459    .0354971
     expersq |  -.0004231   .0001483    -2.85   0.005    -.0007151   -.0001311
        educ |   .0791949   .0103688     7.64   0.000     .0587728    .0996169
       _cons |   .2660838   .1415351     1.88   0.061    -.0126802    .5448479
------------------------------------------------------------------------------
```

# Ex 3.

```
. reg lwage exper expersq educ if female==0

      Source |       SS           df       MS      Number of obs   =       274
-------------+----------------------------------   F(3, 270)       =     58.41
       Model |  30.7328407         3  10.2442802   Prob > F        =    0.0000
    Residual |  47.3513032       270  .175375197   R-squared       =    0.3936
-------------+----------------------------------   Adj R-squared   =    0.3868
       Total |  78.0841439       273  .286022505   Root MSE        =    .41878

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0540175   .0067426     8.01   0.000     .0407427    .0672923
     expersq |  -.0009138   .0001503    -6.08   0.000    -.0012098   -.0006179
        educ |    .090354   .0092666     9.75   0.000       .07211    .108598
       _cons |    .157291   .1364033     1.15   0.250    -.1112583    .4258403
------------------------------------------------------------------------------
```

# Ex 3.

1. Test statictic: $F$

$$F = \frac{[S - (S_1 + S_2)]/(K(m-1)}{(S_1 + S_2)/(N - mK)}$$

2. Critical statistic: $F^*$

$$F^* = F(K(m-1); (N - mK))$$

# Chow test

- **CLRM assumptions:**

  **(a) Which are not satisfied?**

  **(b) What are the consequences for statistical deduction?**

  **(c) How can we solve this problem?**

# Chow test

- **CLRM assumptions:**

  **(a) Which are not satisfied?**

1. **Linear in Parameters**
   - *The model in the population can be written as*

$$y_i = \beta_1 + \beta_2 X_{2i} + \cdots \beta_K X_{Ki} + \varepsilon_i$$

     - *where* $\beta_1, \ldots, \beta_K$ *are unknown parameters of interest and* $\varepsilon_i$ *is an unobservable random error of disturbance term*

# Chow test

- **CLRM assumptions:**

  **(b) What are the consequences for statistical deduction?**

- undermines the economic interpretation of the model (interpretation of estimated parameters)

- it is impossible to prove the properties of the OLS

# Chow test

- **CLRM assumptions:**

  **(c) How can we solve this problem?**

- Interpretation of the test result:
  - H0 rejected ⇒ parameters of the model are not stable
- Consequences of the rejection of H0: model should not be estimated on full sample.
- However, it is often possible to find subsample for which parameters are stable.

# Breusch-Pagan test and White test

- Null and alternative hypothesis:
  - $\boldsymbol{H_0}$: The error term is homoscedastic
  - $\boldsymbol{H_1}$: The error term is heteroskedastic
    - (problem of non-constant error variances is known as HETEROSCEDASTICITY)

# Ex 4.

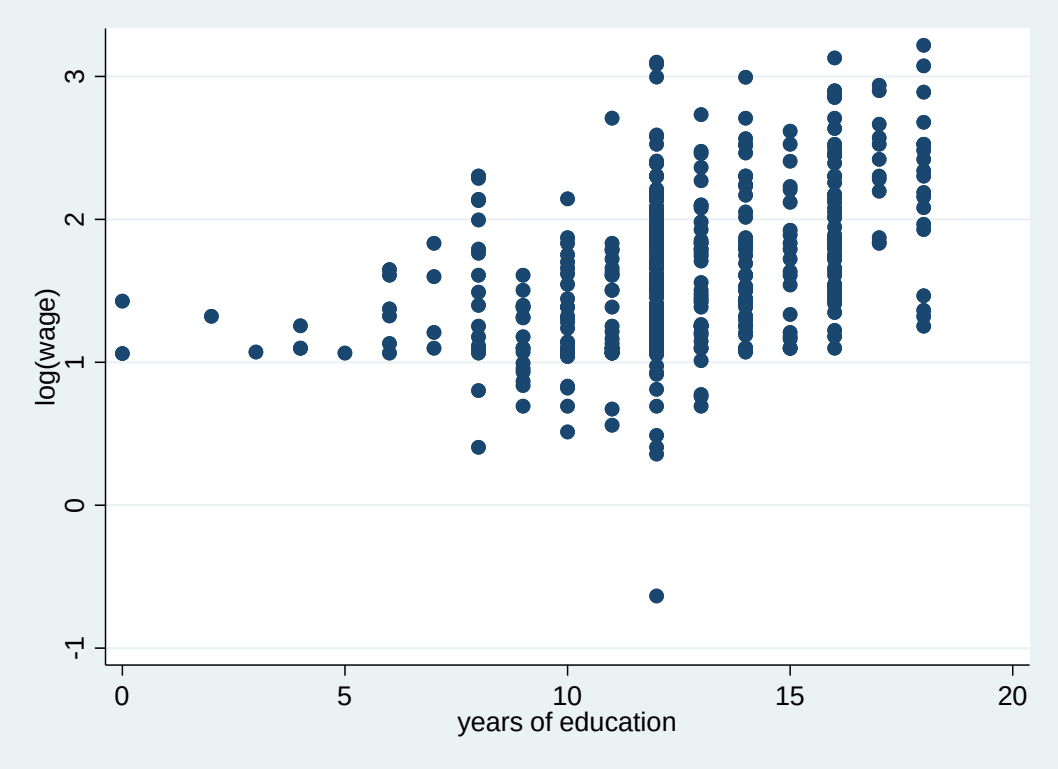
```
. reg lwage educ exper expersq

      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(3, 522)       =     74.67
       Model |  44.5393713         3  14.8464571   Prob > F        =    0.0000
    Residual |   103.79038       522  .198832146   R-squared       =    0.3003
-------------+----------------------------------   Adj R-squared   =    0.2963
       Total |  148.329751       525   .28253286   Root MSE        =    .44591

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0903658    .007468    12.10   0.000     .0756948    .1050368
       exper |   .0410089   .0051965     7.89   0.000     .0308002    .0512175
     expersq |  -.0007136   .0001158    -6.16   0.000     -.000941   -.0004861
       _cons |   .1279975   .1059323     1.21   0.227    -.0801085    .3361035
------------------------------------------------------------------------------
```

**1. Check if there is heteroscedasticity in the model.**

# Ex 4.
## How Does the Heteroskedasticity Look?



# Ex 4.

```
. hettest e, rhs

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: e educ exper expersq

         chi2(4)      =    25.69
         Prob > chi2  =   0.0000

. hettest e, rhs iid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: e educ exper expersq

         chi2(4)      =    20.00
         Prob > chi2  =   0.0005
```

# Ex 4.

```
. imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(8)      =     23.39
         Prob > chi2  =    0.0029

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      23.39      8    0.0029
            Skewness |       0.84      3    0.8406
            Kurtosis |       1.13      1    0.2874
---------------------+-----------------------------
               Total |      25.36     12    0.0132
---------------------------------------------------
```

# Breusch-Pagan test and White test

- **CLRM assumptions:**

  **(a) Which are not satisfied?**

  **(b) What are the consequences for statistical deduction?**

  **(c) How can we solve this problem?**

# Breusch-Pagan test and White test

- **CLRM assumptions:**

  **(a) Which are not satisfied?**

- Interpretation of the test result:
  - $H_0$ rejected ⇒ The error term is heteroskedastic

⇓

Classical Linear Regression Model Assumptions

**5. Homoscedasticity**

- $Var(\varepsilon_i) = \sigma^2$ for $i = 1,2, \dots, N$

# Breusch-Pagan test and White test

- **CLRM assumptions:**

  **(b) What are the consequences for statistical deduction?**

- Consequences of heteroscedasticity for OLS
  - OLS still unbiased and consistent under heteroscedastictiy!
  - Heteroscedasticity invalidates variance formulas for OLS estimators

The usual F-tests and t-tests are not valid under heteroscedasticity

Under heteroscedasticity, OLS is no longer the best linear unbiased estimator (BLUE); there may be more efficient linear estimators

# Breusch-Pagan test and White test

- **CLRM assumptions:**

  **(c) How can we solve this problem?**

Regression with robust standard errors (**White robust estimator**)

# Ex 4.

```
. reg lwage educ exper expersq, robust

Linear regression                               Number of obs     =        526
                                                F(3, 522)         =      71.03
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3003
                                                Root MSE          =     .44591

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0903658   .0077827    11.61   0.000     .0750766     .105655
       exper |   .0410089   .0050237     8.16   0.000     .0311398     .050878
     expersq |  -.0007136   .0001098    -6.50   0.000    -.0009292    -.0004979
       _cons |   .1279975   .1071261     1.19   0.233    -.0824537     .3384487
------------------------------------------------------------------------------
```

# Breusch-Godfrey test

- Null and alternative hypothesis:
  - $\boldsymbol{H_0}$: No autocorrelation
  - $\boldsymbol{H_1}$: Autocorrelation
    - (problem of non-zero error covariances is known as AUTOCORRELATION)

# Ex 5.

```
. tsset year, yearly
        time variable:  year, 1948 to 2003
                delta:  1 year

. reg inf unem

      Source |       SS           df       MS      Number of obs   =        56
-------------+----------------------------------   F(1, 54)        =      3.58
       Model |   31.599858         1   31.599858   Prob > F        =    0.0639
    Residual |  476.815691        54   8.8299202   R-squared       =    0.0622
-------------+----------------------------------   Adj R-squared   =    0.0448
       Total |  508.415549        55  9.24391907   Root MSE        =    2.9715
------------------------------------------------------------------------------
         inf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        unem |   .5023782   .2655624     1.89   0.064    -.0300424    1.034799
       _cons |   1.053566   1.547957     0.68   0.499    -2.049901    4.157033
------------------------------------------------------------------------------
```

1. Check if there is autocorrelation in the model.

# Ex 5.

```
. bgodfrey, lag(1)

Breusch-Godfrey LM test for autocorrelation
------------------------------------------------------------------------------
    lags(p)  |          chi2              df                 Prob > chi2
-------------+----------------------------------------------------------------
       1     |         20.888              1                      0.0000
------------------------------------------------------------------------------
                        H0: no serial correlation
```

# Breusch-Godfrey test

- **CLRM assumptions:**

  **(a) Which are not satisfied?**

  **(b) What are the consequences for statistical deduction?**

  **(c) How can we solve this problem?**

# Breusch-Godfrey test

- **CLRM assumptions:**

  **(a) Which are not satisfied?**

- Interpretation of the test result:
  - $H_0$ rejected ⇒ **Autocorrelation**

⇓

Classical Linear Regression Model Assumptions

4. **No autocorrelation**
   - $Cov(\varepsilon_i, \varepsilon_j) = 0 \quad for\ i \neq j$

# Breusch-Godfrey test

- **CLRM assumptions:**

  **(b) What are the consequences for statistical deduction?**

- Consequences of **autocorrelation** for OLS
  - OLS still unbiased and consistent under **autocorrelation**!
  - **Autocorrelation** invalidates variance formulas for OLS estimators

The usual F-tests and t-tests are not valid under **autocorrelation**

Under **autocorrelation**, OLS is no longer the best linear unbiased estimator (BLUE); there may be more efficient linear estimators

# Breusch-Godfrey test

- **CLRM assumptions:**

  **(c) How can we solve this problem?**

Regression with robust standard errors (**Newey-West robust estimator)**

# Ex 5.

```
. newey inf unem, lag(1)

Regression with Newey-West standard errors      Number of obs     =         56
maximum lag: 1                                  F(  1,        54) =       3.25
                                                Prob > F          =     0.0769

------------------------------------------------------------------------------
             |             Newey-West
         inf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        unem |   .5023782    .278577     1.80   0.077    -.0561351    1.060892
       _cons |   1.053566   1.464589     0.72   0.475    -1.882759     3.98989
------------------------------------------------------------------------------
```

# June, 17 from 1.15 PM to 4.45 PM

1. For what do we use diagnostic tests?
2. What test can be used to verify if the function form of the model is correct? Give $H0$ and $H1$. How it is connected with **CLRM** assumptions?
3. What test can be used to verify if error term is normally distributed? Give $H0$ and $H1$. How it is connected with **CLRM** assumptions? What are the consequences for OLS estimator properties of rejecting the null hypothesis?
4. What tests can be used to test for homoscedasticity in the model? Give $H0$ and $H1$. How it is connected with **CLRM** assumptions? What are the consequences for OLS estimator properties of rejecting the null hypothesis?