Econometrics

Natalia Nehrebecka

Part 5

Agenda

- Problems with the data
  - Omitted variables
  - Insignificant variables
    - Omitted variables problem has much more serious consequences that insignificant variables problem.
  - Unusual observations and outliers
  - Multicollinearity

Omitted variables

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon \quad (1)$$

$$\boldsymbol{y = X_1\beta_1 + u} \qquad \textbf{(2)}$$

omitted variables: we estimate model (2) but in reality model (1) is valid ($\beta_2 \neq 0$)

Omitted variables

$$\hat{\beta}_1 = (X_1^{'}X_1)^{-1}X_1^{'}y = (X_1^{'}X_1)^{-1}X_1^{'}(X_1\beta_1 + X_2\beta_2 + \varepsilon) =$$
$$= \beta_1 + (X_1^{'}X_1)^{-1}X_1^{'}X_2\beta_2 + (X_1^{'}X_1)^{-1}X_1^{'}\varepsilon$$

$$E(\hat{\beta}_1) - \beta_1 = (X_1^{'}X_1)^{-1}X_1^{'}X_2\beta_2$$

- If we miss out an important variable it not only means our model is poorly specified it also means that any **estimated parameters are biased**

Omitted variables

- **When we can get the correct parameter estimates even though variables are omitted?**

$$\beta_2 = 0$$

$$X_1^{'} X_2 = 0$$

Ex 1.

```
reg lwage south age educ, robust

Linear regression                               Number of obs     =        935
                                                F(3, 931)         =      54.28
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1505
                                                Root MSE          =     .38878

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       south |  -.1430392    .028152    -5.08   0.000    -.198288   -.0877905
         age |   .0217808   .0041369     5.27   0.000    .0136621    .0298996
        educ |   .0572196   .0059089     9.68   0.000    .0456234    .0688158
       _cons |   5.336631   .1641776    32.51   0.000     5.01443    5.658832
------------------------------------------------------------------------------
```

Ex 1.

```
reg lwage south age, robust

Linear regression                               Number of obs     =        935
                                                F(2, 932)         =      30.99
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0623
                                                Root MSE          =     .40824

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       south |  -.1688684   .0295583    -5.71   0.000     -.226877   -.1108598
         age |    .021169   .0042834     4.94   0.000     .0127629    .0295752
       _cons |   6.136341   .1437233    42.70   0.000     5.854283      6.4184
------------------------------------------------------------------------------
```

Omitted variables

- We estimate the following model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i \quad (1)$$

- Lets assume that the following model is correct:

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

- Then

$$E(\hat{\beta}_1) - \beta_1 = \boldsymbol{\beta_2 \frac{S_{x_2}}{S_{x_1}} \rho_{x_1,x_2}}$$

Ex 2.

EXERCISE 3.1 Two regressions were estimated to explain married woman logarithm of income (logrincome) by the number of years spend on education (educ).

```
      Source |       SS       df       MS              Number of obs =     371
-------------+------------------------------           F(  1,   369) =   45.48
       Model |  10.6867044     1  10.6867044           Prob > F      =  0.0000
    Residual |  86.7135186   369  .234995985           R-squared     =  0.1097
-------------+------------------------------           Adj R-squared =  0.1073
       Total |   97.400223   370  .263243846           Root MSE      =  .48476

------------------------------------------------------------------------------
  logrincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0650681   .0096489     6.74   0.000     .0460944    .0840417
       _cons |   5.549348   .1175245    47.22   0.000     5.318247     5.78045
------------------------------------------------------------------------------
```

1. Interpret the estimated coefficients and R^2.
2. What will probably be the direction of parameter (in front of educ) estimator biase if we don't take into account:
   (a) inteligence of a respondent.
   (b) number of children of a respondent.
   (c) size of a place of living.
   (d) husband's income.

Insignificant variables

$$\boldsymbol{y = X_1\beta_1 + X_2\beta_2 + \varepsilon} \quad \textbf{(1)}$$

$$y = X_1\beta_1 + u \qquad (2)$$

- **Insignificant variables:** we estimate model (1) but model (2) is valid ($\beta_2 = 0$)

Insignificant variables

- In the model with insignificant variables **OLS estimator is inefficient**, that is its variance is higher that the variance of estimator in the model without insignificant variables

Ex 3.

```
. reg lwage educ exper expersq female numdep

      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(5, 520)       =     69.73
       Model |  59.5354775         5  11.9070955   Prob > F        =    0.0000
    Residual |  88.7942739       520  .170758219   R-squared       =    0.4014
-------------+----------------------------------   Adj R-squared   =    0.3956
       Total |  148.329751       525   .28253286   Root MSE        =    .41323

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0813869   .0072958     11.16   0.000      .067054    .0957197
       exper |   .0406966   .0050303      8.09   0.000     .0308144    .0505788
     expersq |  -.0007331   .0001138     -6.44   0.000    -.0009567   -.0005096
      female |  -.3364511   .0363071     -9.27   0.000    -.4077777   -.2651244
      numdep |  -.0194746   .0156521     -1.24   0.214    -.0502238    .0112745
       _cons |   .4369009    .108755      4.02   0.000     .2232478     .650554
------------------------------------------------------------------------------
```

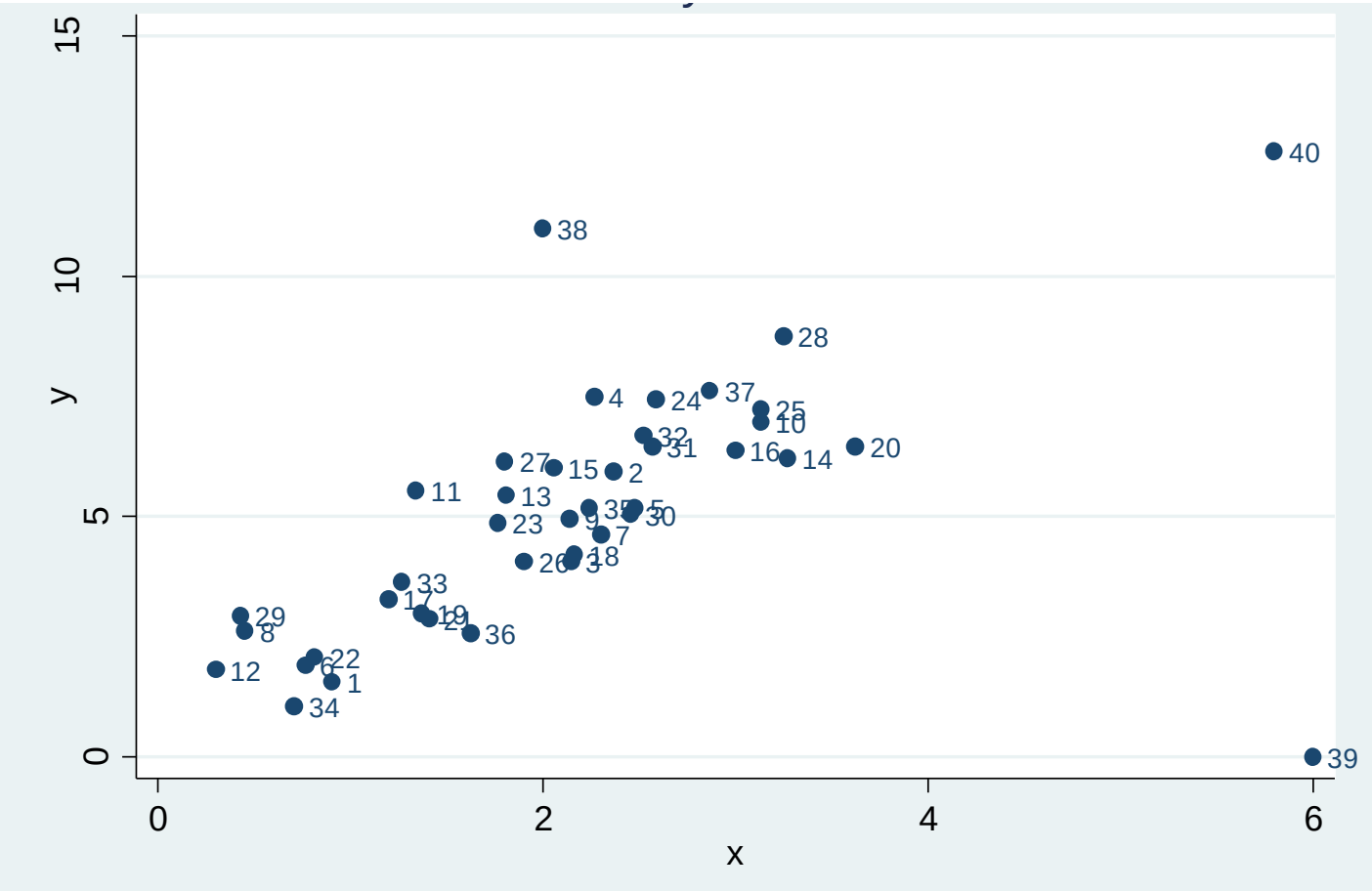
Ex 3.

```
. reg lwage educ exper expersq 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]
-------------+----------------------------------------------------------------
        educ |   .0841361   .0069568    12.09   0.000     .0704692    .0978029
       exper |     .03891   .0048235     8.07   0.000      .029434    .0483859
     expersq |   -.000686   .0001074    -6.39   0.000     -.000897   -.0004751
      female |  -.3371868   .0363214    -9.28   0.000    -.4085411   -.2658324
       _cons |    .390483   .1022096     3.82   0.000     .1896894    .5912767
------------------------------------------------------------------------------
```

Unusual observations and outliers



Unusual observations and outliers

- We can have two cases:

- observations which is unusual in the context of other observations
  - Unusual observations is correct

- **outlier** (erroneous observation)
  - An erroneous observation is an observation that can not be made explained as part of the theoretical economic model which forms the basis of the estimated model

  - Erroneous observations often appear as a result of mistakes at entering observations into the database

Ex 4.

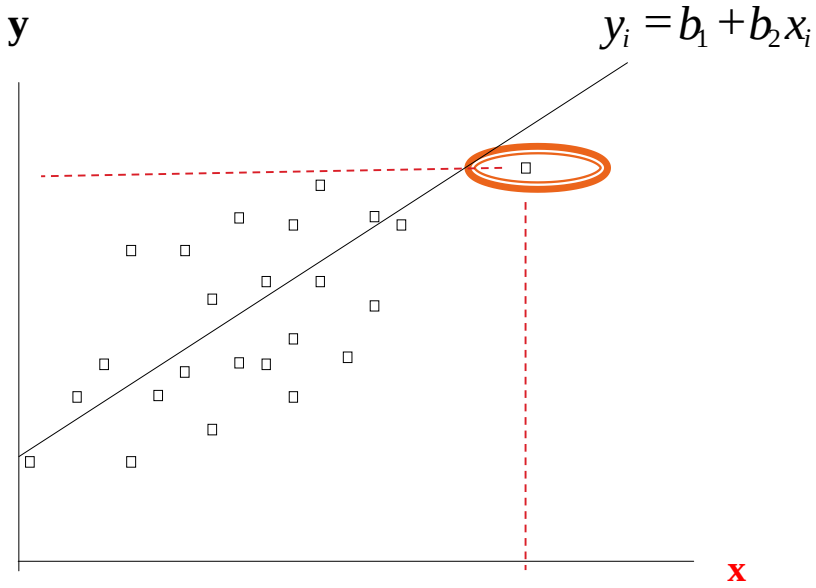
- Using data set obtained from Polish LFS researcher wants to estimate models for the wage equation. Before starting the estimation procedure he calculated descriptive statistics for variable wage. This statistics are reported below.

```
sum wage in 2018

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        wage |       5773    53392.31    32264.34         0      99997


    count if wage==99997
    703
```

Unusual observations and outliers



Detection of unusual observations and outliers

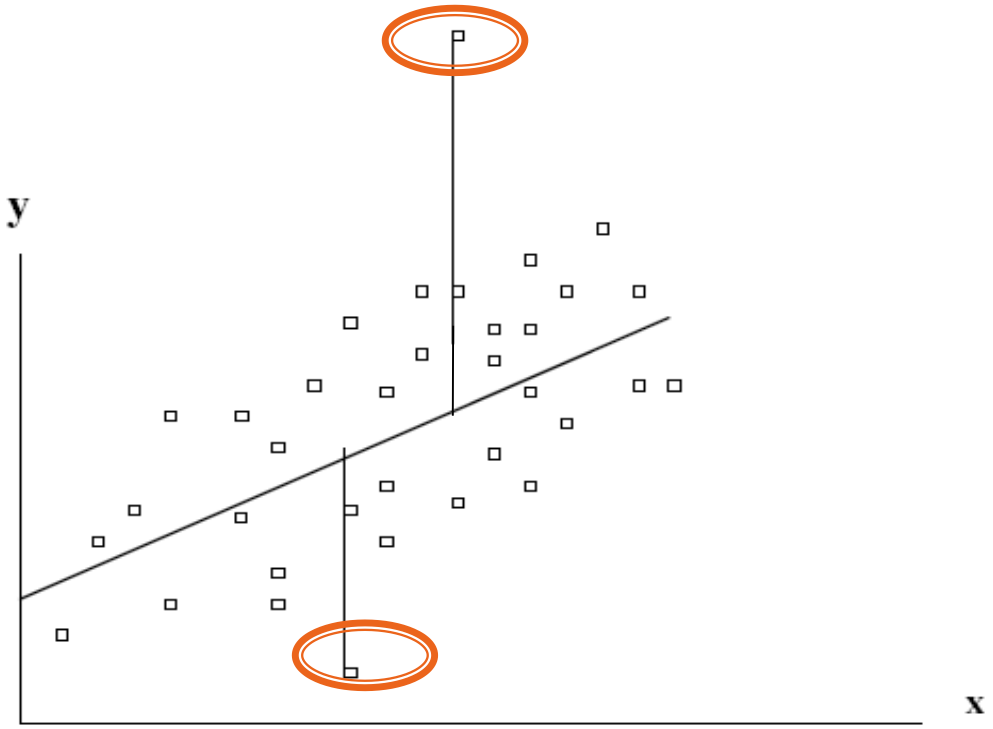
- In order to detect unusual observations we can use **leverage statistics**

$$h_i = x_i(X'X)^{-1}x_i'$$

- observation can be considered unusual if $h_i \geq \frac{2K}{N}$

- It doesn't mean that observation doesn't fit the model, check it we have to analyze **standardized residuals**

Unusual observations and outliers



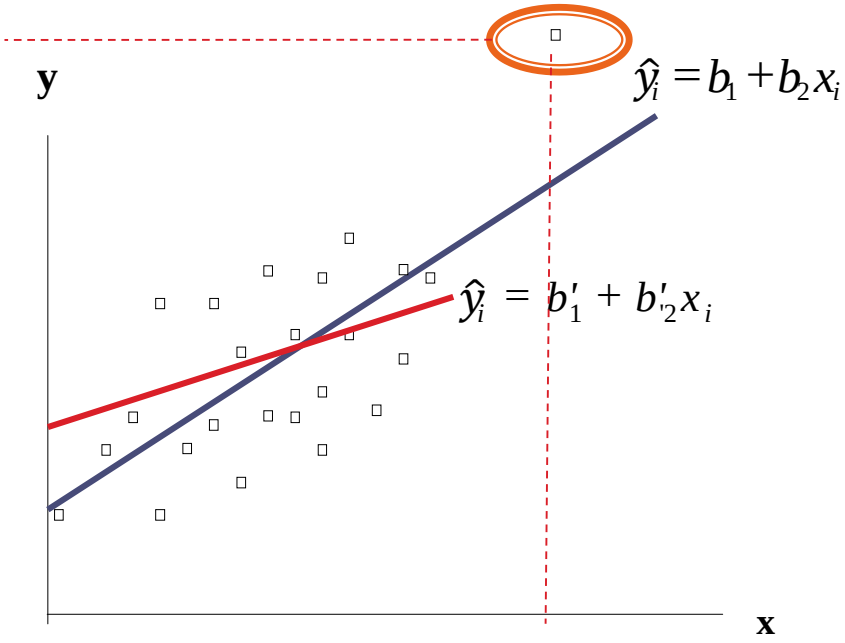
Detection of unusual observations and outliers

- Standardized residuals

$$\hat{e}_i = \frac{e_i}{s\sqrt{1-h_i}} \sim t(N-K)$$

- Observation is not typical and influence estimation if $|\hat{e}_i| > 2$

Unusual observations and outliers



Unusual observations and outliers

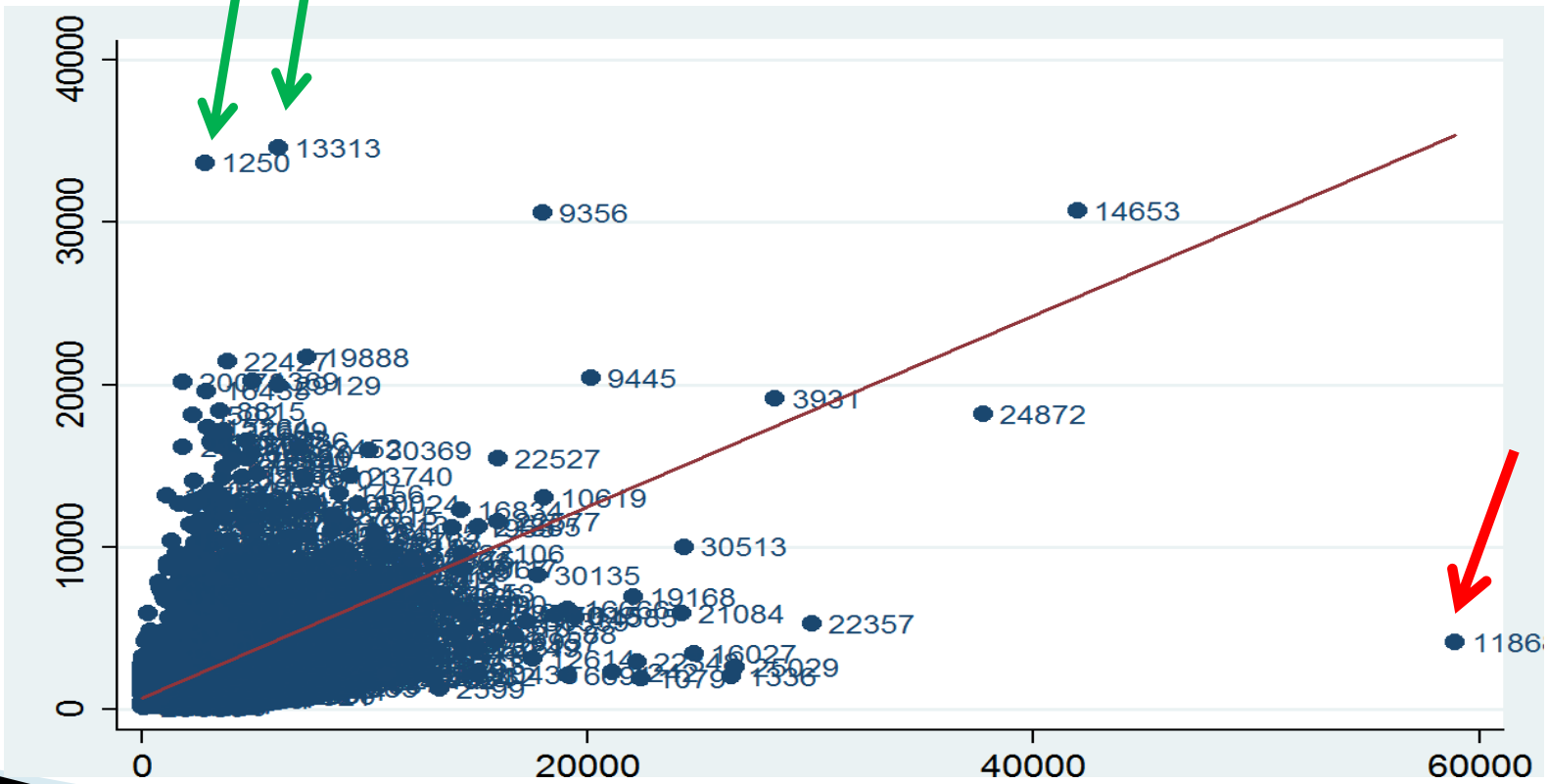
- Influence of unusual observations and outliers on the regression results is completely opposite:

- **Unusual observation has positive impact on:**
  - precision of the estimate of β
  - fit of the model

- **Outlier has negative impact on**
  - precision of the estimate of β
  - fit of the model

Unusual observations and outliers

- The impact of the observation on the regression results is especially large if $\hat{e}_i$ and h_i are both large



Decide on the basis of this graph which observations are unusual and why?

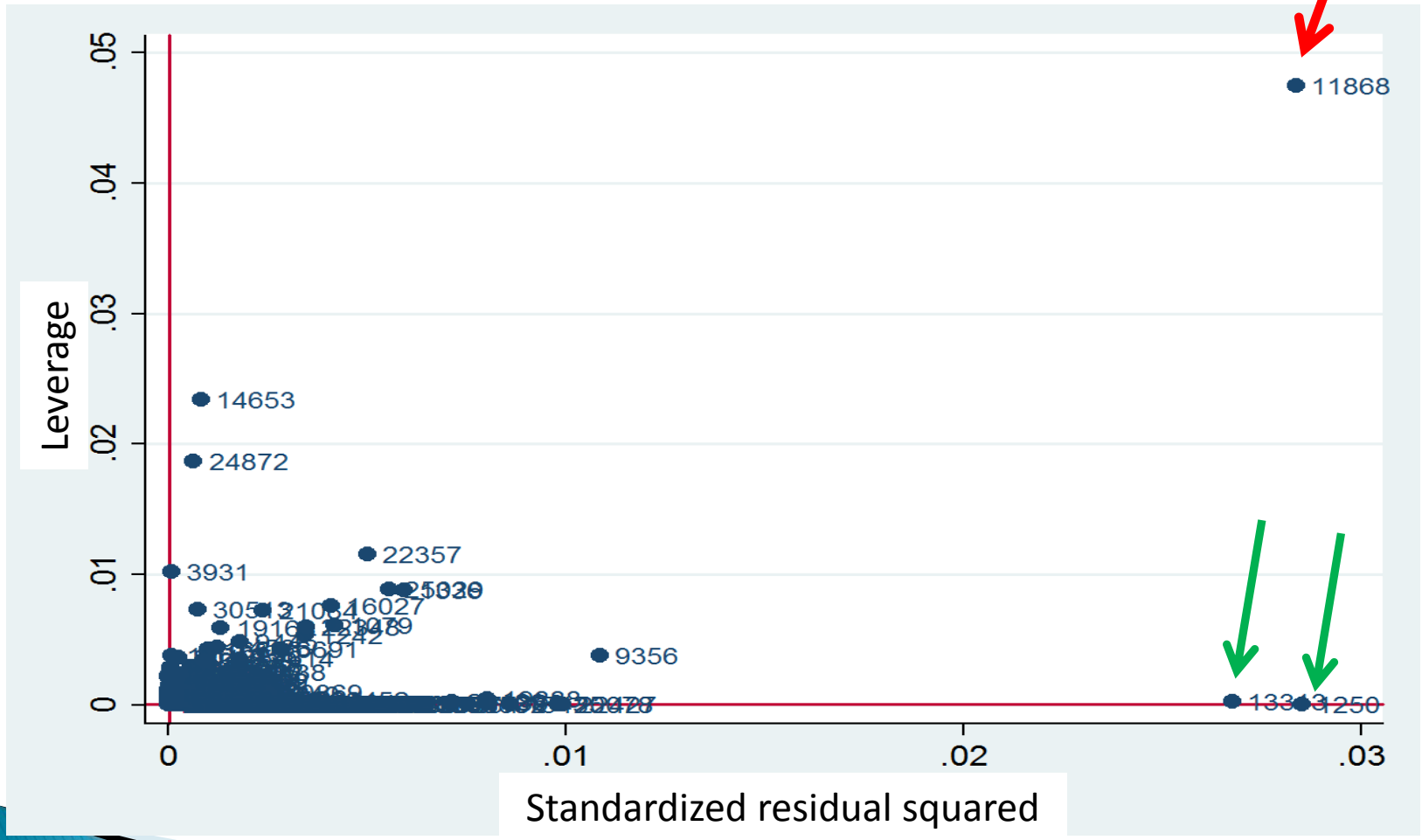
Detection of unusual observations and outliers

- **Cook's Distance**

$$CD_i = \frac{\hat{e}_i^2}{K} \cdot \frac{h_i}{1 - h_i}$$

- The observations with $CD_i > \frac{4}{N}$ should be verified.

Unusual observations and outliers



Decide on the basis of this graph which observations are unusual and why?

Unusual observations and outliers

$$\texttt{expense}_i = \beta_0 + \beta_1 \texttt{wage}_i + \varepsilon_i$$

$$K = 2$$

$$N = 31697$$

Unusual observations and outliers

	numer	wage	expenses	residual_st	leverage	Cook_d
1.	11868	58935	4132	-30.72398	.0474962	23.53513

$$h_i \geq \frac{2K}{N} = \frac{2*2}{31679} \approx 0,00012$$

$$CD_i > \frac{4}{N} = \frac{4}{31679} \approx 0,00012$$

Decide on the basis of this table which observations are unusual and why?

Multicollinearity

- **Collinearity** - strong correlation between explanatory variables
  - Difficult to identify (separate) the influences of variables

  - y is growing with x_1 and x_2
  - which of the variables „causes” the growth of y?

Perfect multicollinearity

- If an independent variable is an exact linear combination of the other independent variables, then we say the model suffers from **perfect collinearity**, and it cannot be estimated by OLS
  - The identification of the influence of explanatory variables on dependent variable impossible

- Model on logarithms:
  - dependent variables: national income Y_t
  - explanatory variables: spending for education E_t , population P_t , spending for education per capita Z_t .

Collinearity!

$$\ln(Z_t) = \ln\left(\frac{E_t}{P_t}\right) = \ln(E_t) - \ln(P_t)$$

Imperfect multicollinearity

- We are talking about **imperfect multicollinearity** if the correlation between exogenous variables are nonzero
- Imperfect multicollinearity is a rule rather than exceptions in non-experimantal data
- We can have a problem if the multicollinearity is strong

reduces the precision of estimates

Ex 5.

Source	SS	df	MS		
Model	216.967865	14	15.4977047	Number of obs	= 1967
Residual	286.280455	1952	.146660069	F(14, 1952)	= 105.67
				Prob > F	= 0.0000
Total	503.24832	1966	.255975748	R-squared	= 0.4311
				Adj R-squared	= 0.4271
				Root MSE	= .38296

ln_wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ_1	.1889953	.0302267	6.25	0.000	.1297153	.2482753
educ_2	.2786084	.0360692	7.72	0.000	.2078703	.3493465
educ_3	.4458425	.0375599	11.87	0.000	.3721808	.5195043
rural	-.0512427	.0179803	-2.85	0.004	-.0865054	-.01598
age	.0358145	.0056632	6.32	0.000	.0247081	.046921
age_2	-.0004565	.0000674	-6.77	0.000	-.0005887	-.0003243
sex	-.1157975	.0501914	-2.31	0.021	-.2142319	-.0173632
w_time	.7589422	.0441721	17.18	0.000	.6723127	.8455717
sexXw_time	-.1165905	.0530097	-2.20	0.028	-.220552	-.012629
status_1	.0928414	.0238605	3.89	0.000	.0460467	.1396362
status_2	-.1109334	.0602626	-1.84	0.066	-.2291193	.0072524
status_3	-.0083049	.0442809	-0.19	0.851	-.0951477	.0785379
exper	.0076781	.0014517	5.29	0.000	.0048311	.0105252
exper_tot	.0026003	.001987	1.31	0.191	-.0012967	.0064972
_cons	5.566119	.1095404	50.81	0.000	5.351291	5.780947

Correlation matrix

	age	exper	exper_tot
age	1.0000		
exper	0.4875	1.0000	
exper_tot	0.9201	0.5743	1.0000

Imperfect multicollinearity

- **Variance Inflation Factor (VIF)** - allows to measure the impact of correlation on the estimation result

$$VIF_k = \frac{1}{1 - R_k^2}$$

Ex 6.

```
Linear regression                               Number of obs =     526
                                                F(  4,   521) =   36.11
                                                Prob > F      =  0.0000
                                                R-squared     =  0.2316
                                                Root MSE      =  .46771

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tenure |   .0057131    .019157     0.30   0.766    -.0319213    .0433475
  tot_tenure |   .0108999   .0189116     0.58   0.565    -.0262524    .0480523
      female |  -.3213711   .0411303    -7.81   0.000    -.4021728   -.2405694
     married |   .1824256    .045194     4.04   0.000     .0936408    .2712104
       _cons |     1.5815   .0425249    37.19   0.000     1.497959    1.665042
------------------------------------------------------------------------------
```

Ex 6.

```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
      tenure |     51.81    0.019302
  tot_tenure |     51.38    0.019465
     married |      1.08    0.924556
      female |      1.06    0.943762
-------------+----------------------
    Mean VIF |     26.33
```

- VIF statistics for this regression were generated in STATA. What is the problem in the model?

Ex 6.

```
. reg lwage tenure female married, robust

Linear regression                               Number of obs     =        526
                                                F(3, 522)         =      46.54
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2312
                                                Root MSE          =      .4674

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tenure |    .016627   .0037012     4.49   0.000     .0093559    .0238982
      female |  -.3202892   .0412866    -7.76   0.000    -.4013975   -.2391808
     married |   .1810318   .0453072     4.00   0.000     .0920249    .2700386
       _cons |   1.581708   .0424218    37.29   0.000     1.498369    1.665046
------------------------------------------------------------------------------
```

Ex 6.

```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
      tenure |      1.09    0.916762
     married |      1.08    0.927801
      female |      1.06    0.945894
-------------+----------------------
    Mean VIF |      1.08
```

Ex 7.

```
. reg lwage exper expersq female married, robust

Linear regression                               Number of obs =     526
                                                F(  4,   521) =   40.76
                                                Prob > F      =  0.0000
                                                R-squared     =  0.2414
                                                Root MSE      =  .46472

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0365473   .0056165     6.51   0.000     .0255136    .0475811
       agesq |  -.0007835   .0001153    -6.79   0.000    -.0010101   -.0005569
      female |  -.3624154   .0401367    -9.03   0.000     -.441265   -.2835658
     married |   .1256961   .0459394     2.74   0.006      .035447    .2159453
       _cons |   1.469439   .0482136    30.48   0.000     1.374722    1.564155
------------------------------------------------------------------------------

.
```

Ex 7.

```
vif

    Variable |       VIF       1/VIF
-------------+----------------------
         age |     15.57    0.064242
       agesq |     14.70    0.068050
     married |      1.28    0.780473
      female |      1.03    0.972158
-------------+----------------------
    Mean VIF |      8.14
```

- VIF statistics for this regression were generated in STATA. What is the problem in the model?

June, 21 from 1.15 PM to 4.45 PM

1. What are the consequences of omitting significant variable in the model?
2. When we can get the correct parameter estimates even though variables are omitted?
3. Why we should remove non-significant variables from the model?
4. Parameters for x_1 and x_2 are positive. Variables are negatively correlated. What is the impact of omitting x_1 on parameter for x_2?
5. What do we mean by unusual observations? When unusual observations can be considered as outliers?
6. When unusual observations has significant impact on estimation?
7. What statistics can be used to find unusual observations?
8. When we say that the variables in the model are perfect collinear? How can you solve this problem?
9. 5. What are the consequences of imperfect collinearity? Using what statistics you can detect imperfect collinearity in model?