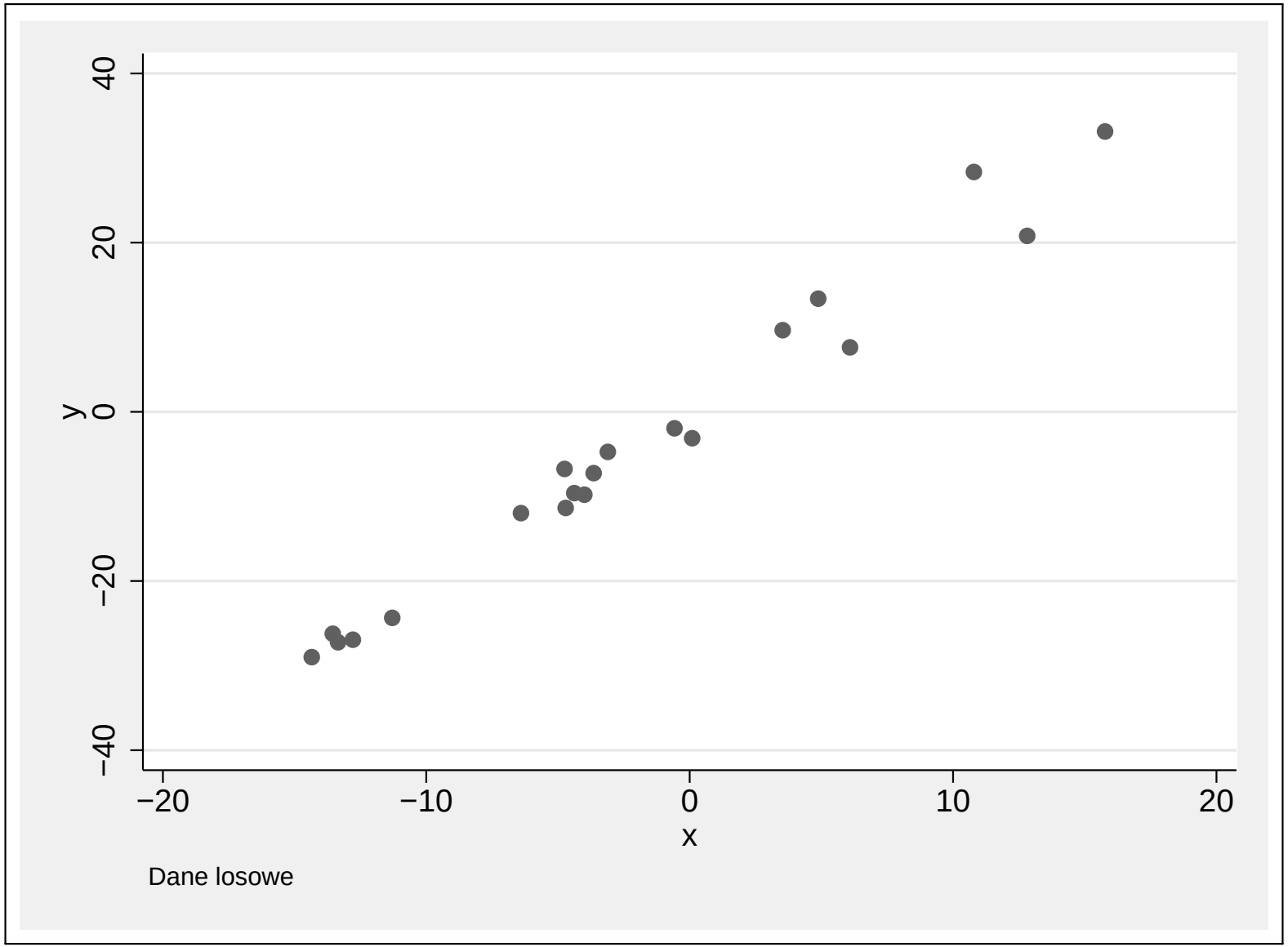
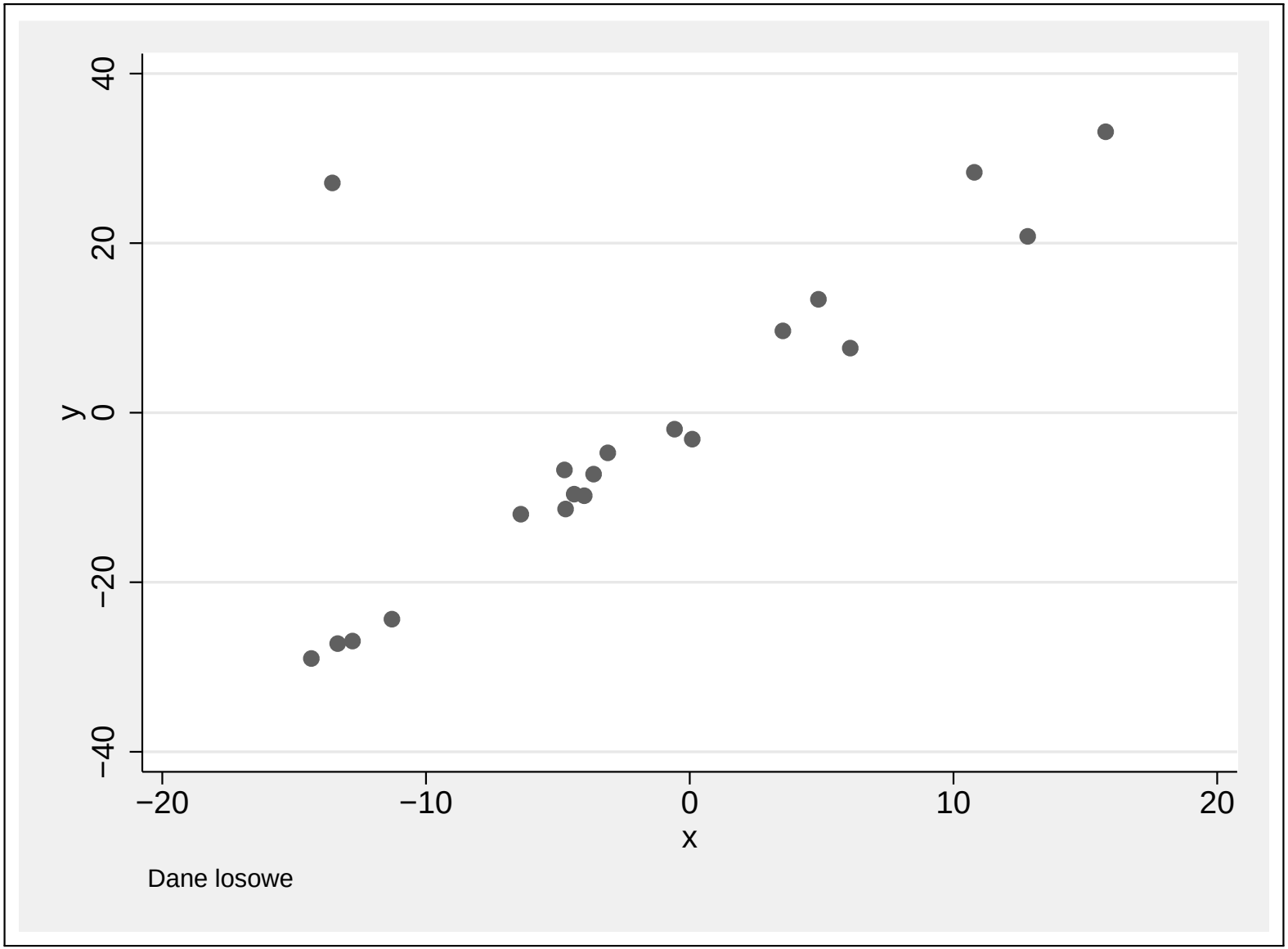
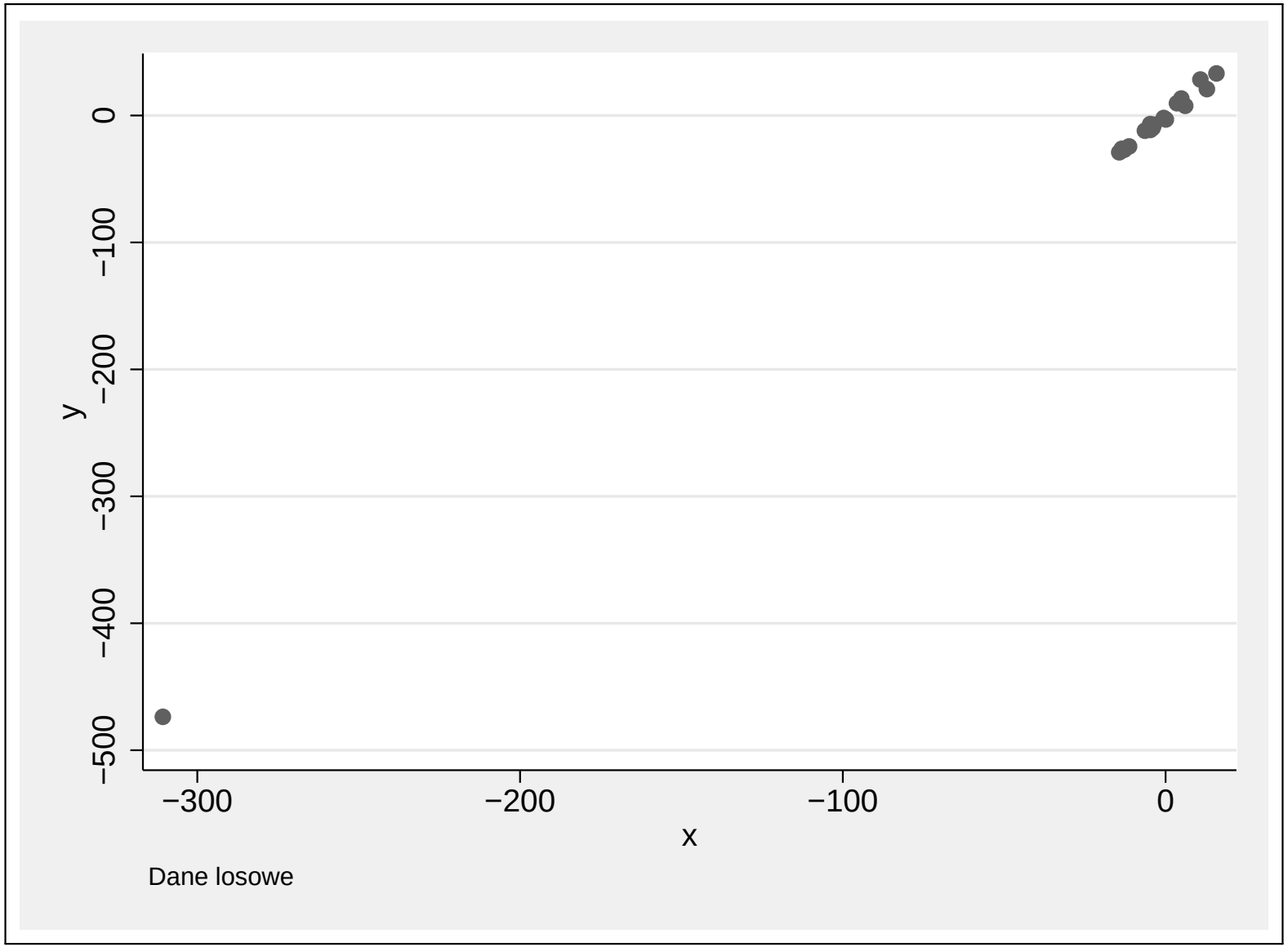
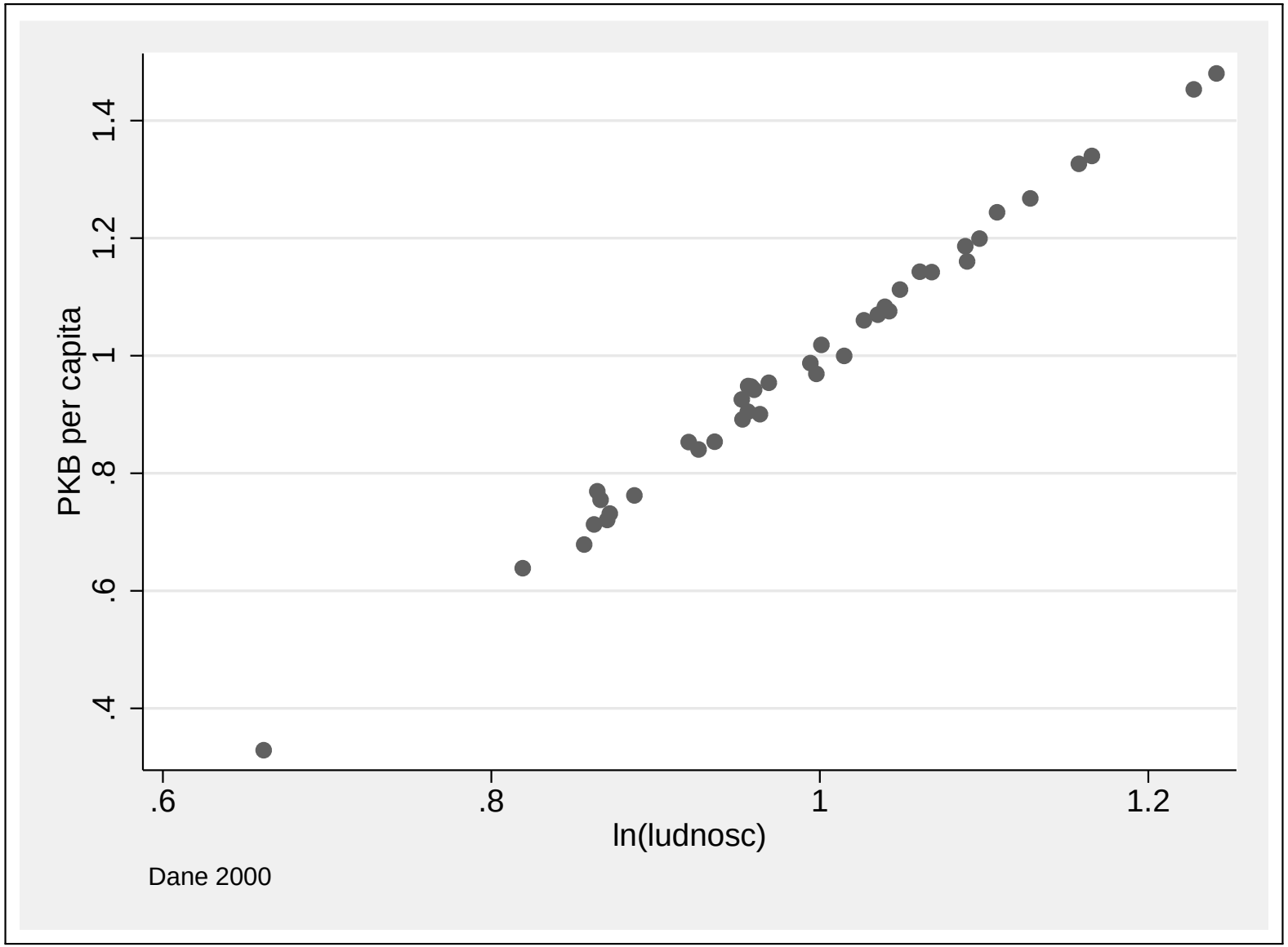
Problemy ze zbiorem danych empirycznych

- Obserwacje nietypowe
  - Naturalna nietypowość
  - Błędy w kodowaniu

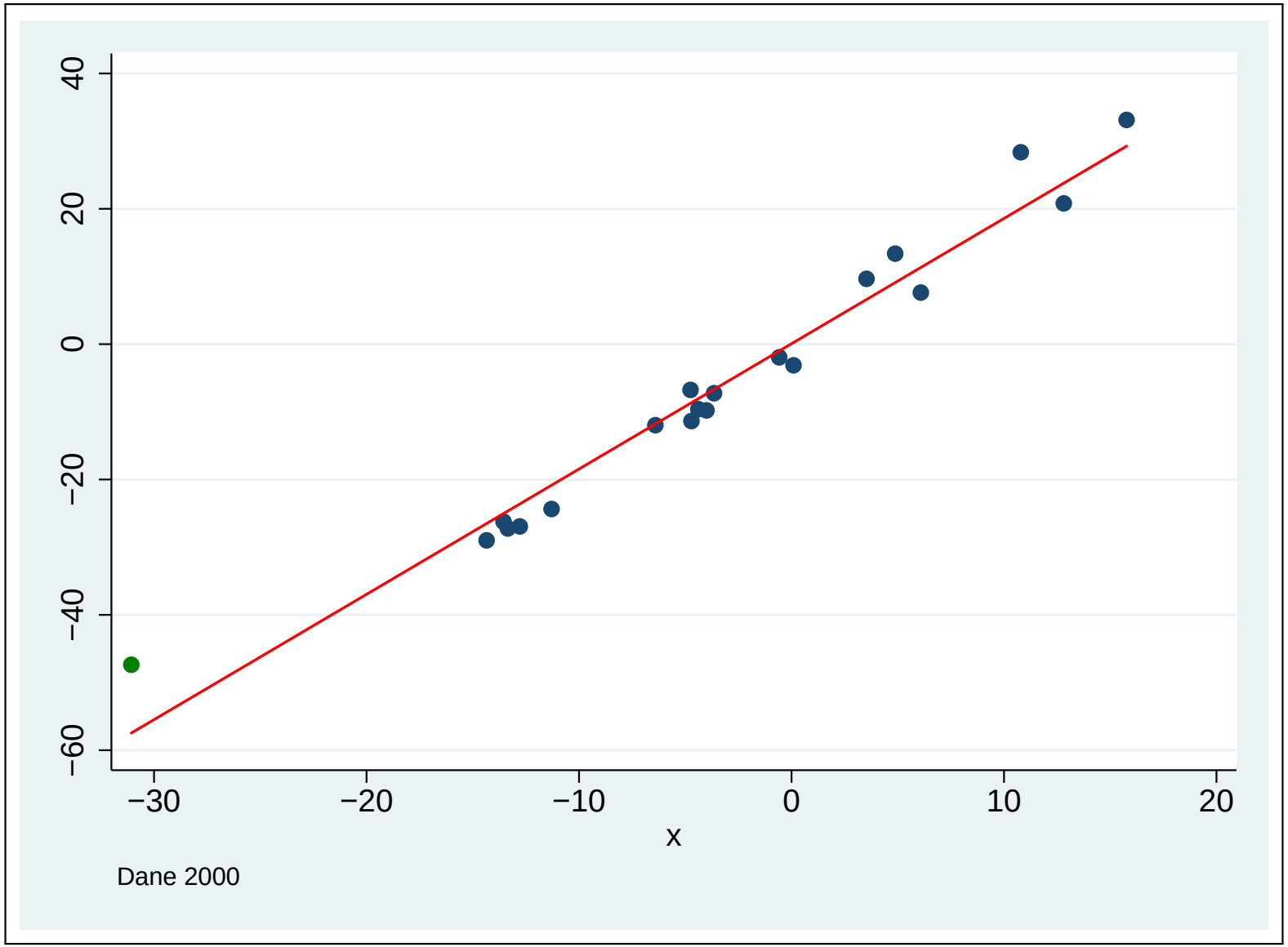

40
20
0
−20
−40
y
−20
−10
0
10
20
x
Dane losowe

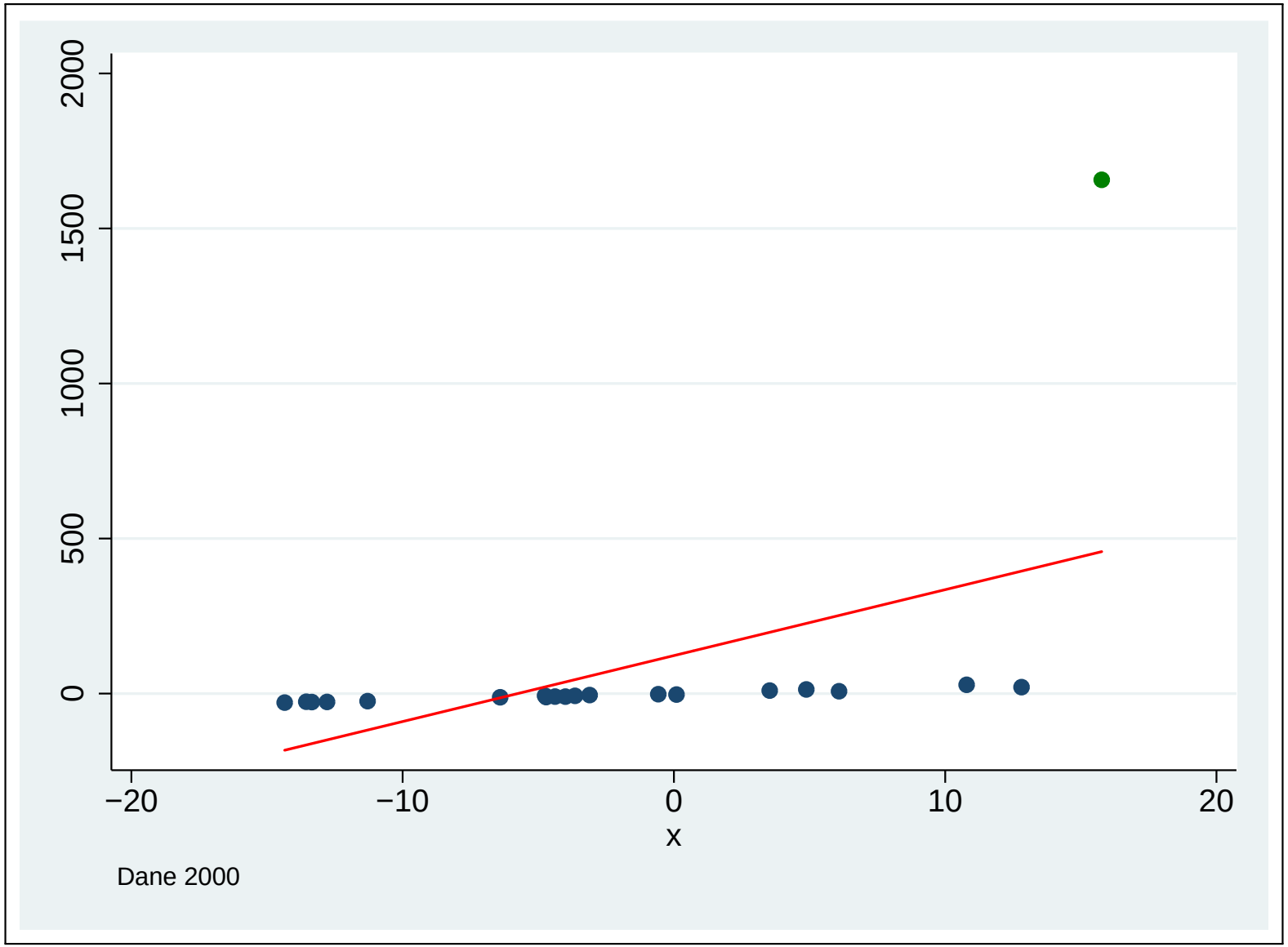

40
20
0
−20
−40
y
−20
−10
0
10
20
x
Dane losowe


0
−100
−200
−300
−400
−500
y
−300
−200
−100
0
x
Dane losowe


PKB per capita
.4
.6
.8
1
1.2
1.4
.6
.8
1
1.2
ln(ludnosc)
Dane 2000

- Obserwacje nietypowe
  - Naturalna nietypowość
  - Błędy w kodowaniu
- Obserwacje błędne


40
20
0
−20
−40
−60
−30
−20
−10
0
10
20
x
Dane 2000


2000
1500
1000
500
0
−20
−10
0
10
20
x
Dane 2000

Przekształcona macierz obserwacji

- Macierz rzutu

$$\hat{y} = \sum_i P_{i,i} y_i$$

- Statystyka dźwigni

$$h_i = \delta_i' X (X'X)^{-1} X' \delta_i = \delta_i' P \delta_i = P_{ii}$$

Przekształcone reszty

- Standaryzowana reszta

$$\hat{e}_i = \frac{e_i}{se(e_i)} = \frac{e_i}{\sigma^2(1-h_i)} \sim N(0,1)$$

- Studentyzowana reszta

$$\tilde{e}_i = \frac{e_i}{se(e_i)} = \frac{e_i}{S^2(1-h_i)} \sim t_{N-k}$$

Statystyki

- DFITS (Welsh i Kun 1977)

$$DFITS_i = \frac{\tilde{e}_i}{S_i\sqrt{1-h_i}}\sqrt{\frac{h_i}{1-h_i}} = r_i\sqrt{\frac{h_i}{1-h_i}}$$

- Odległość Cooka (Cook 1977)

$$CD_i = \frac{1}{k}\frac{e_i^2}{S^2}\frac{h_i}{(1-h_i)^2} \quad \sim F(2, n-2)$$

- Miara Hadi'ego (Hadi 1992)

$$H_i = \frac{h_i}{1-h_i} + \frac{k}{h_i}\frac{\hat{e_i}^2}{1-\hat{e_i}^2}$$

```
      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  1,    18) =    6.32
       Model |  684499.002     1  684499.002            Prob > F      =  0.0216
    Residual |  1948549.82    18  108252.768            R-squared     =  0.2600
-------------+------------------------------           Adj R-squared =  0.2189
       Total |  2633048.82    19  138581.517            Root MSE      =  329.02

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   21.26678    8.45736     2.51   0.022     3.498525    39.03503
       _cons |   122.4676   75.77701     1.62   0.123    -36.73403    281.6691
------------------------------------------------------------------------------
```

	rstd	l	df	cd
1.	-.3500357	.0516287	-.0796416	.0033351
2.	-.1215037	.0533006	-.0280295	.0004156
3.	-.1473605	.0522694	-.0336522	.0005988
4.	-.1628845	.0514844	-.0369066	.00072
5.	4.242484	.2621113	285.7606	3.196726
6.	-1.073506	.1605906	-.4716647	.1102365
7.	-.3986501	.0533206	-.0923531	.0044755
8.	.4379718	.1329183	.1675418	.0147024
9.	.0055467	.0619816	.0013856	1.02e-06
10.	.2998164	.1052582	.1001867	.0052874
11.	.3980449	.1248007	.146722	.0112965
12.	-.7804019	.094788	-.2496792	.0318867
13.	-1.269868	.1978922	-.6424296	.1989222
14.	.4560831	.1359385	.1768297	.0163627
15.	-.0880952	.054482	-.0205554	.0002236
16.	-.1053486	.0543353	-.0245484	.0003188
17.	-.5926796	.0713023	-.1611768	.0134846
18.	-.6755151	.0826231	-.1995612	.0205491
19.	.5061014	.1483649	.2067648	.0223112
20.	-.1907013	.0506093	-.0428325	.0009693

Wariancję każdego z estymatorów pojedynczych parametrów modelu można zapisać jako:

$$var(b_j) = \frac{\sigma^2}{(1 - r_{j\cdot}^2)\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2} = \frac{\sigma^2}{(1 - r_{j\cdot}^2)S_{jj}^2}$$

W takim przypadku są spełnione założenia KMRL, ale występują następujące problemy:

1. niewielkie zmiany w zbiorze danych powodują duże zmiany w otrzymywanych oszacowaniach parametrów.
2. współczynniki równania regresji mają duże błędy standardowe, oraz mogą być nieistotne statystycznie, nawet gdy łącznie są istotne, a współczynnik R^2 modelu jest wysoki
3. współczynniki równania regresji mają „złe", czyli niezgodne z teorią znaki, albo są zbyt małe lub zbyt duże.

Statystykę służącą do pomiaru siły korelacji jest mnożnik nadwyżkowej wariancji *ang. Variance Inflation Factor (VIF)*. Jest to prosty test oparty na statystyce R^2.

$$VIF = \frac{1}{1 - r_{j.}^2}$$

```
      Source |       SS       df       MS              Number of obs =   16162
-------------+------------------------------           F(  5, 16156) =  469.93
       Model |  324.309105     5   64.861821            Prob > F      =  0.0000
    Residual |  2229.92615 16156  .138024644            R-squared     =  0.1270
-------------+------------------------------           Adj R-squared =  0.1267
       Total |  2554.23525 16161  .158049332            Root MSE      =  .37152

------------------------------------------------------------------------------
    lzarobki |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     kobieta |  -.2298019   .0058916   -39.01   0.000    -.2413501   -.2182538
        wiek |   .1840604   .0488994     3.76   0.000     .0882121    .2799087
       wiek2 |  -.0051581   .0019593    -2.63   0.008    -.0089986   -.0013177
       wiek3 |   .0000623   .0000337     1.85   0.065    -3.80e-06    .0001284
       wiek4 |  -2.66e-07   2.11e-07    -1.26   0.207    -6.80e-07    1.47e-07
       _cons |    3.62026   .4410593     8.21   0.000     2.755735    4.484785
------------------------------------------------------------------------------
```

Variable	VIF	1/VIF
wiek3	290954.38	0.000003
wiek2	253156.31	0.000004
wiek4	39067.63	0.000026
wiek	26062.69	0.000038
kobieta	1.01	0.989069
Mean VIF	121848.40	

```
      Source |       SS       df       MS              Number of obs =   13986
-------------+------------------------------           F(  3, 13982) =  676.37
       Model |  265.657711     3  88.5525704            Prob > F      =  0.0000
    Residual |  1830.56521 13982  .130922987            R-squared     =  0.1267
-------------+------------------------------           Adj R-squared =  0.1265
       Total |  2096.22292 13985  .149890806            Root MSE      =  .36183

------------------------------------------------------------------------------
    lzarobki |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     kobieta |  -.2204251   .0062728   -35.14   0.000    -.2327207   -.2081295
        wiek |   .0192319   .0013819    13.92   0.000     .0165232    .0219405
        staz |  -.0104626   .0013581    -7.70   0.000    -.0131247   -.0078006
       _cons |   5.450074   .0285721   190.75   0.000     5.394069    5.506079
------------------------------------------------------------------------------
```

Variable	VIF	1/VIF
staz	18.94	0.052802
wiek	18.89	0.052935
kobieta	1.04	0.965235
Mean VIF	12.96	