Econometrics

Jerzy Mycielski

2010

Omitted and insignificant variables

- Two models:

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{u} \tag{1}$$

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \tag{2}$$

- Two cases:
  - omitted variables: we estimate model (1) but in reality model (2) is valid ($\boldsymbol{\beta}_2 \neq \mathbf{0}$)
  - insignificant variables: we estimate model (2) but model (1) is valid ($\boldsymbol{\beta}_2 = \mathbf{0}$).
- Omitted variables problem has much more serious consequences that insignificant variables problem.

Omitted variables

Example

The researcher wants to verify the effectivity of some drug. He divided randomly the sample of patients into the treated group which was given the drug and the control group which was given placebo. Then the researcher evaluated the change of health of the treated and untreated patients according. It is known however, that the measure of health, which was used, is influenced by some additional characteristics of patient such as age. Is possible find an unbiased estimate of the effect of the drug if we omit these additional characteristics?

Omitted variables

Answer: Yes, if the sample was really randomly divided into treated and untreated groups. In such a case there is no correlations between characteristics omitted in the regression and the participation dummy $Corr_{X_1 X_2} = 0$.

Omitted variables

Example

Correlation between the logarithm of wage and interviewer number

Regression results

logNETPAY	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
INTV	.0016346	.0000989	16.53	0.000	.0014408	.0018284
_cons	5.557534	.0042232	1315.95	0.000	5.549256	5.565812

Omitted variables

Regression with voivodships dummy and dummy for the city size
Part of the regression table

logNETPAY	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
INTV	-.0002166	.0001482	-1.46	0.144	-.0005071	.0000738
_IVOI1_3	-.1495124	.0428622	-3.49	0.000	-.2335268	-.0654981
...						
_IVOI1_97	-.1219227	.0275238	-4.43	0.000	-.1758722	-.0679731
_ITOWN2_1	-.0789742	.019422	-4.07	0.000	-.1170433	-.040905
...						
_ITOWN2_9	-.2471119	.0166571	-14.84	0.000	-.2797616	-.2144623
_cons	5.90414	.0154814	381.37	0.000	5.873795	5.934485

Omitted variables

- Variable related to interviewer number is now insignificant!
- Explanation: correlation between the voivodship number and city size (omitted variables) and the interviewer number.
- Regression of interviewer number on voivodship and city size dummies gives:

```
R-squared = 0.5861
```

Omitted variables

- The simplest case one omitted variable, one included variable

$$\mathrm{E}\left(\widetilde{\boldsymbol{\beta}}_1\right) - \beta_1 = \beta_2 \frac{s_{x_2}}{s_{x_1}} \rho_{x_1 x_2}$$

  - omitted variable x_2 positively correlated with x_1, coefficient β_2 positive - coefficient β_1 overestimated
  - omitted variable x_2 positively correlated with x_1, coefficient β_2 negative - coefficient β_1 underestimated
  - omitted variable x_2 negatively correlated with x_1, coefficient β_2 positive - coefficient β_1 underestimated
  - omitted variable x_2 negatively correlated with x_1, coefficient β_2 negative - coefficient β_1 overestimated
- These results are also often used in the context of multiple regression (although they are not exactly valid in this case), when the omitted variable is correlated with *one* variable included in the model

Omitted variables

direction of the bias

Example

Simple linear model was build in which the number of children born in some area was explained by the number of storks living in the area. It was found that there is a significant relationship between these two variables. Does it imply that storks bring babies?

Omitted variables

Answer: In Poland the birth rate is higher in the countryside than in the urban areas ($\beta_2 > 0$). It is also the case that most storks are living on the countryside ($\rho_{x_1 x_2} > 0$). Important variable related to whether the area in question is an urban area was omitted in the model. Positive estimate of the parameter for the variable number of storks is probably related to the omitted variable bias of the estimator ($\mathrm{E}\left(b_1\right) = \beta_1 + \beta_2 \frac{s_{x_2}}{s_{x_1}} \rho_{x_1 x_2} > 0$ even if $\beta_1 = 0$).

Omitted variables

direction of the bias

Example

Experience and age

- Dependence of log wage on experience

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exper	.0113283	.0006278	18.04	0.000	.0100975	.012559
_cons	7.36974	.0133627	551.52	0.000	7.343544	7.395935

Omitted variables

direction of the bias

- Dependence of log wage on age and experience

lplaca	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exper	.0058233	.0014101	4.13	0.000	.003059	.0085877
age	.0064003	.0014685	4.36	0.000	.0035214	.0092791
_cons	7.214572	.0380217	189.75	0.000	7.140037	7.289107

- Estimate of the coefficient for experience is much lower

Insignificant variables

- Insignificant variable problem: we estimate model (2) but $\boldsymbol{\beta}_2 = \mathbf{0}$.
- We already know that for valid restrictions $\mathbf{H}\boldsymbol{\beta} = \mathbf{h}$, restricted estimator is unbiased and has smaller variance that unrestricted estimator estimator.
- We conclude that if the restriction $\boldsymbol{\beta}_2 = \mathbf{0}$ is valid (model 1 is true) but we will not use this restriction in estimation (we will estimate model 2), then the estimator will be unbiased but inefficient.

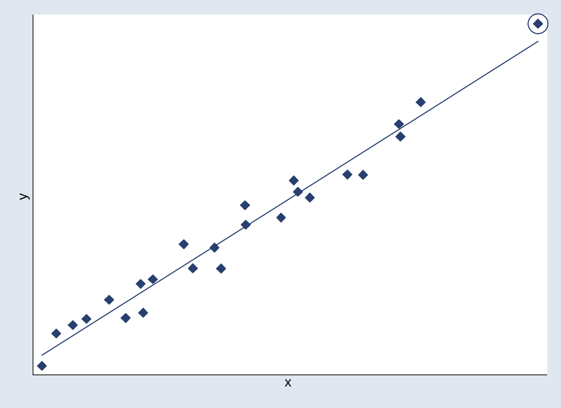
Corollary

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

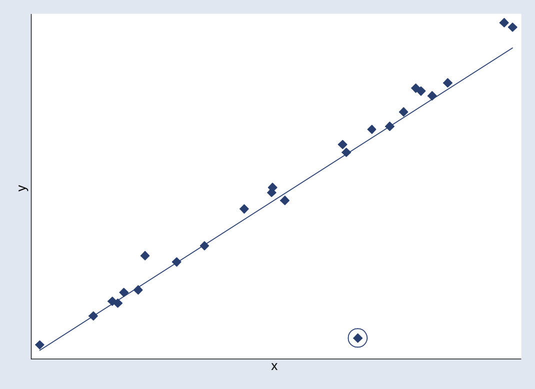
Unusual observations and outliers

- We can have two cases:
    - observations which is unusual in the context of other observations
    - outlier (erroneous observation)

Unusual observation



Outlier



Differences between unusual observations and outliers

- Unusual observations is correct, outlier is erroneous
- Influence of unusual observations and outliers on the regression results is completely opposite:
  - Unusual observation has positive impact on:
    - precision of the estimate of $\boldsymbol{\beta}$
    - fit of the model
  - Outlier has negative impact on
    - precision of the estimate of $\boldsymbol{\beta}$ $\boldsymbol{\beta}$
    - fit of the model

Differences between unusual observations and outliers

Example

We need to compare the profitability of two contracts: A and B. We have data consisting of 10 observations on internal rate of return (IRR) for each of the contacts:
A: $\{10, 8, 8, 9, 11, 10, 8, 9, 11, 10\}$
B: $\{16, 15, 18, 17, 16, -80, 17, 16, 16, 17\}$.
Notice one unusual observations for contract B (it is related to the firm which bankrupted). Should we take into account this observation? Define the dummy variable B which take the value of 1 for contracts from group B.

Differences between unusual observations and outliers

Regression results with one observation omitted:

IRR	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_IB_1	7.155556	.4808912	14.88	0.000	6.140964	8.170147
_cons	9.4	.330972	28.40	0.000	8.70171	10.09829

Regression results with all observations included:

IRR	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_IB_1	-3.5	10.66526	-0.33	0.747	-25.90688	18.90688
_cons	9.4	7.541478	1.25	0.229	-6.444057	25.24406

Detection of unusual observations and outliers

leverage

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

$$h_i = \delta_i' \mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}' \delta_i = \delta_i' \mathbf{P}_X \delta_i = \left(\mathbf{P}_X\right)_{ii}$$
$$= \mathbf{x}_i \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{x}_i'$$

where $\delta_i = [0, \ldots, 0, 1, 0 \ldots, 0]'$ and $\mathbf{P}_X = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$.

- Properties of leverage:
  - for each model
  $$0 \leq h_i \leq 1$$
  - for a model with constant
  $$\frac{1}{n} \leq h_i \leq 1$$
  - observation can be considered unusual if $h_i > \frac{2k}{n}$
- Notice that h_i detect $\mathbf{x}_i$ unusual in the context of other $\mathbf{x}$'s, it does not measure how well $\mathbf{x}_i$ fits the model

Detection of unusual observations and outliers

standardized residuals

- Variance of the vector of residuals is equal to:

$$\begin{aligned} \mathrm{Var}\left(\mathbf{e}\right) &= \mathrm{Var}\left(\mathbf{M}_X \boldsymbol{\varepsilon}\right) = \mathbf{M}_X \left(\mathbf{I}\sigma^2\right) \mathbf{M}_X \\ &= \sigma^2 \mathbf{M}_X \end{aligned}$$

- The variance of residual e_i is equal to

$$\begin{aligned} \mathrm{Var}\left(e_i\right) &= \mathrm{Var}\left(\delta_i' \mathbf{e}\right) = \sigma^2 \delta_i' \mathbf{M}_X \delta_i \\ &= \sigma^2 \left(1 - \delta_i' \mathbf{P}_X \delta_i\right) = \sigma^2 \left(1 - h_i\right) \end{aligned}$$

- Standardized residual is then given by

$$\begin{aligned} \widehat{e}_i &= \frac{e_i}{\sqrt{\mathrm{Var}\left(e_i\right)}} = \frac{e_i}{\sigma\sqrt{1 - h_i}} \\ &\approx \frac{e_i}{s\sqrt{1 - h_i}} \end{aligned}$$

- The impact of the observation on the regression results is especially large if e_i and h_i are both large

- The measure of the impact of one observation on regression fit is called Cook distance.
- It is based on difference between $\widehat{\mathbf{y}}$ obtained from full sample and $\widehat{\mathbf{y}}_{(i)}$ obtained from sample with i-th observation omitted:

$$CD_i = \frac{\left(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_{(i)}\right)' \left(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_{(i)}\right)}{Ks^2} = \frac{\widehat{e}_i^2}{K} \frac{h_i}{1 - h_i}$$

- The observations with $CD_i > 0.5$ and especially these with $CD_i > 1$ should be verified.

Example

Dependence of spending for accommodation on income

Regression results (4111 observations)

lq	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
linc	.4087146	.0139339	29.33	0.000	.3813966	.4360326
_cons	2.768599	.106037	26.11	0.000	2.560709	2.976488

Detection of unusual observations and outliers

Number $\widehat{e} > 2$ is equal to 217 which is about 5% of the sample

Ordered table for leverages 5

q	inc	r2st	lev	cook
375.9	16	3.582841	.0140365	.0911117
414.84	23	3.4911	.0120339	.0740249
400	47	2.904768	.0085492	.036313
132.35	78.9	.5826743	.0064039	.0010943
370.68	118	2.103206	.0049578	.0110109

Detection of unusual observations and outliers

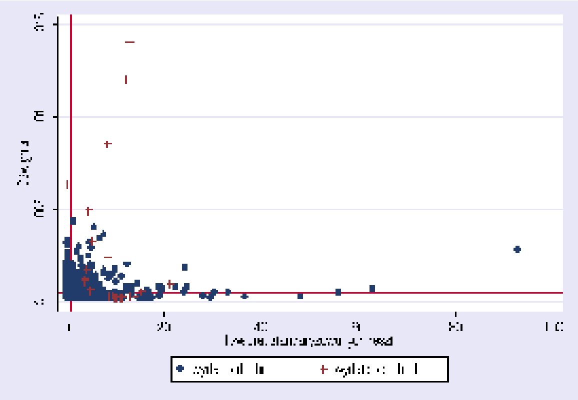
Ordered table for Cook distances

q	inc	r2st	lev	cook
3.67	16150	-9.631348	.0028882	.1314109
375.9	16	3.582841	.0140365	.0911117
414.84	23	3.4911	.0120339	.0740249
400	47	2.904768	.0085492	.036313
2.72	780	-7.928539	.0007519	.0233001

For all observations $q > inc$, this is unusual!

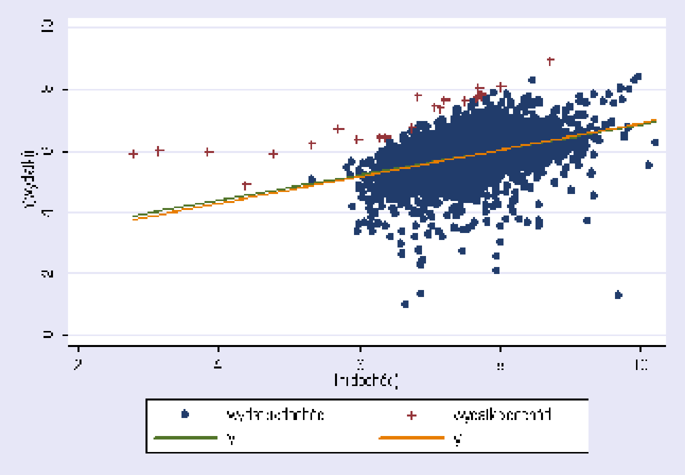
Detection of unusual observations and outliers

Standardized squares of residuals and leverages



Detection of unusual observations and outliers

Regression results for original sample and sample with omitted observations for which $q > inc$



Multicollinearity

- Multicollinearity - strong correlation of explanatory variables
- Difficult to identify (separate) the influences of variables

- x_1 and x_2 are growing "in most cases" together

Example

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

Perfect multicollinearity

- perfect multicollinearity - columns of matrix **X** linearly dependent
- The identification of the influence of explanatory variables on dependent variable impossible

Example

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.

Perfect multicollinearity

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 nonexperimantal data
- We can have a problem if the multicollinearity is strong

Example

dependence of wage on experience

Regression results

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exper_p1	.0821159	.0153068	5.36	0.000	.0521095	.1121223
exper_p2	-.006285	.0021507	-2.92	0.003	-.0105011	-.002069
exper_p3	.0002075	.0001237	1.68	0.093	-.0000349	.00045
exper_p4	-2.70e-06	3.09e-06	-0.87	0.382	-8.76e-06	3.35e-06
exper_p5	1.13e-08	2.77e-08	0.41	0.684	-4.31e-08	6.57e-08
_cons	7.18452	.033636	213.60	0.000	7.118583	7.250458

- Joint test for significance of $exper^5$ and $exper^4$

$F(2, 6503) = 8.35$ [0.0002]

VIF table

Variable	VIF	1/VIF
exper_p3	81085.22	0.000012
exper_p4	72099.95	0.000014
exper_p2	17923.53	0.000056
exper_p5	8874.86	0.000113
exper_p1	600.63	0.001665
Mean VIF	36116.84	

Regression without variable $exper^5$

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exper_p1	.0771847	.0093503	8.25	0.000	.058855	.0955145
exper_p2	-.0054865	.0008796	-6.24	0.000	-.0072108	-.0037621
exper_p3	.0001588	.0000308	5.16	0.000	.0000985	.0002191
exper_p4	-1.45e-06	3.57e-07	-4.07	0.000	-2.15e-06	-7.53e-07
_cons	7.191273	.0292561	245.80	0.000	7.133921	7.248624