Webinar Series in Applied Quantitative Analysis

Date	Topic
	Session One
February 29	Potential Outcomes and Omitted Variable Bias I (Theory)
March 7	Potential Outcomes and Omitted Variable Bias II (Application)
	Session Two
March 21	Difference-in-differences I (Theory)
March 28	Difference-in-differences II (Application)
	Session Three
April 25	Power analysis, clustering and sample size calculations I (Theory)
May 2	Power analysis, clustering and sample size calculations II (Application)
	Session Four
May 23	Fixed-effects I (Theory)
May 30	Fixed-effects II (Application)
	Session Five
June 20	Instrumental variables I (Theory)
June 27	Instrumental Variables II (Application)
	Session Six
July 25	Lagged dependent variables and the Arellano-Bond Estimator I (Theory)
August 1	Lagged dependent variables and the Arellano-Bond Estimator II (Application)

Advancing Evidence.
Improving Lives.

Series Objectives and Target Audience

- Cover methods and issues commonly faced by social scientists when doing empirical work
  - We are open to suggestions for topics, if there is demand, we can continue the series next year

- Target audience is post-graduate students in African universities
  - Economics, Sociology, Agriculture, Population Studies and related fields
  - Anyone with an interest in empirical work and <u>causal inference</u>
  - Must have taken regression and statistics for this to be beneficial to you

- The structure is like a classroom, we will teach theory and then provide examples and demonstrations
  - The first session today, and some of the others, will likely be refreshers

Logistics and other topics

- Participants are muted, chat is disabled, write questions in the Q&A box
  - Selected questions will be answered and posted publicly during the webinar, all Q&A will be posted on the webinar website after the session + video of the session
  - Occasionally we will open the chat to receive responses to questions posed

- You can attend any webinar you want, topics are independent and do not build on each other like a structured course
  - But the first week's topic is an important base to better understand subsequent topics
  - **Registration link again: https://air-org.zoom.us/webinar/register/WN_MJ5UevG-RESBb-iWtEOdmQ**

- We will have a guest speaker for topics 4 and 5 (FE and IV) and maybe others
  - Paul Sirma is handling the Q&A; Jacque Ngao the logistics, Isaac Kiplagat the IT stuff

Selection bias in regression: Linking omitted variable bias with the potential outcomes framework

Sudhanshu (Ashu) Handa

Institute Fellow – AIR
Kenan Eminent Professor of Public Policy, UNC-CH

Modern econometrics is about counterfactual thinking

- Move away from ‘goodness of fit’ and model fitting towards a focus on point estimates and unbiasedness
    - Focus on causal inference

- In this framework the variable of interest can be viewed as a ‘treatment’ which then helps us identify the counterfactual

- Then we ask—is there selection bias? Why or how?

- Use toolkit to resolve the selection bias (experimental designs, fixed-effects, difference-in-differences, etc)

We will link the ‘potential outcomes framework’ (next week) to what we learned in econometrics class—bias in the OLS regression model

Selection can be viewed as an omitted variable which leads to bias

- frailty or the health endowment
- innate ability or special knowledge (master farmers)
- motivation, industriousness

In impact evaluation, selection into program participation is the concern

In general applications, we want to understand whether we have a good counterfactual and if not, find ways to improve it

Housing spending in the U.S. – do renters pay more than non-renters?

Renters pay 8.1% less than non-renters controlling for family size and race

```
. reg log_h renter fam_size white black
```

Source	SS	df	MS
Model	979.660909	4	244.915227
Residual	11640.8373	16,214	.717949754
Total	12620.4982	16,218	.778178458

Number of obs	=	16,219
F(4, 16214)	=	341.13
Prob > F	=	0.0000
R-squared	=	0.0776
Adj R-squared	=	0.0774
Root MSE	=	.84732

log_housing	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
renter	-.0812318	.0141718	-5.73	0.000	-.1090101	-.0534535
fam_size	.1537796	.0045256	33.98	0.000	.144909	.1626502
white	-.1304967	.0260013	-5.02	0.000	-.1814622	-.0795312
black	-.2704562	.0316689	-8.54	0.000	-.3325307	-.2083818
_cons	7.222598	.0288747	250.14	0.000	7.166001	7.279196

Renters pay 9.3% less than non-renters controlling for family size and race <u>and region of residence</u>

Is there selection bias? (Not really—the coefficient of 'renter' hardly changed)
Renters are equally as likely to live in urban and rural areas: No selection into urban/rural

```
. reg log_h renter fam_size white black urban
```

Source	SS	df	MS
Model	1246.70862	5	249.341723
Residual	11373.7896	16,213	.701522828
Total	12620.4982	16,218	.778178458

Number of obs	=	16,219
F(5, 16213)	=	355.43
Prob > F	=	0.0000
R-squared	=	0.0988
Adj R-squared	=	0.0985
Root MSE	=	.83757

log_housing	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
renter	-.0934233	.0140227	-6.66	0.000	-.1209092	-.0659373
fam_size	.1527183	.0044738	34.14	0.000	.1439491	.1614875
white	-.097098	.0257591	-3.77	0.000	-.1475886	-.0466073
black	-.266627	.0313051	-8.52	0.000	-.3279884	-.2052655
urban	.4901121	.0251201	19.51	0.000	.4408739	.5393503
_cons	6.748645	.0374803	180.06	0.000	6.675179	6.82211

Renters pay 11.9% MORE than non-renters controlling for family size and race <u>and region of residence and income</u>

Is there selection bias? Yes—the coefficient of 'renter' completely flipped signs
There is differential selection into income for renters and non-renters. Do renters select into 'low' or 'high' income?

```
. reg log_h renter fam_size white black urban log_pc
```

Source	SS	df	MS
Model	3148.72697	6	524.787828
Residual	9471.77126	16,212	.584244464
Total	12620.4982	16,218	.778178458

Number of obs	=	16,219
F(6, 16212)	=	898.23
Prob > F	=	0.0000
R-squared	=	0.2495
Adj R-squared	=	0.2492
Root MSE	=	.76436

log_housing	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
renter	.1187369	.0133262	8.91	0.000	.092616	.1448578
fam_size	.2094188	.004202	49.84	0.000	.2011824	.2176551
white	-.0578265	.0235176	-2.46	0.014	-.1039236	-.0117294
black	-.1350735	.0286616	-4.71	0.000	-.1912535	-.0788935
urban	.3203993	.0231166	13.86	0.000	.2750882	.3657103
log_pcinc	.3999453	.0070096	57.06	0.000	.3862058	.4136848
_cons	2.658703	.079424	33.47	0.000	2.503023	2.814383

Omitted variable bias formula from our OLS regression class helps us understand selection

-We will show some math for the sake of transparency and completeness, and so you have notes but this is not a math class

-Focus on the intuition, and how to use the bias formula in practice (we will do some examples)

OLS regression refresher (before entering into omitted variable bias)

$For \quad Z_i = \gamma_0 + \gamma_1 P + \sigma_i \quad the\ OLS\ estimator\ of\ \gamma_1\ is\ given\ by$

$\widehat{\gamma_1} = \frac{\sum(P_i - \bar{P})Z_i}{\sum(P_i - \bar{P})^2} = \quad \frac{COV(P,Z)}{VAR(P)}$ (check notes from econometrics class)

- *"Covariance between the outcome and the predictor divided by the variance of the predictor"*

- Remember this formula, will be used later

Omitted variable bias in classical regression

$$Y_i = \beta_0 + \beta_1 P + \beta_2 Z + \mu_i \qquad \textit{True relationship}$$

$$Y_i = \alpha_0 + \alpha_1 P \qquad + \varepsilon_i \qquad \textit{Estimated relationship } (Z \textit{ omitted})$$

From previous slide, recall that OLS estimate of α_1 is given by *"covariance between outcome and predictor divided by variance of the predictor"*

$$\widehat{\alpha_1} = \frac{\sum(P_i - \bar{P})Y_i}{\sum(P_i - \bar{P})^2} = \frac{COV(P, Y)}{VAR(P)}$$

$$\widehat{\alpha_1} = \frac{\sum(P_i - \bar{P})Y_i}{\sum(P_i - \bar{P})^2} = \frac{\sum(P_i - \bar{P})\overbrace{(\beta_0 + \beta_1 P_i + \beta_2 Z_i + \mu_i)}^{\text{Substitute for } Y_i \text{ using true equation}}}{\sum(P_i - \bar{P})^2}$$

$$\widehat{\alpha_1} = \frac{\beta_0 \sum(P_i - \bar{P})}{\sum(P_i - \bar{P})^2} + \frac{\beta_1 \sum(P_i - \bar{P})P_i}{\sum(P_i - \bar{P})^2} + \frac{\beta_2 \sum(P_i - \bar{P})Z_i}{\sum(P_i - \bar{P})^2} + \frac{\sum(P_i - \bar{P})\mu_i}{\sum(P_i - \bar{P})^2}$$

Expand the numerator

Substitute for Y_i using true equation

- $\widehat{\alpha_1} = \frac{\sum(P_i - \bar{P})Y_i}{\sum(P_i - \bar{P})^2} = \frac{\sum(P_i - \bar{P})(\beta_0 + \beta_1 P_i + \beta_2 Z_i + \mu_i)}{\sum(P_i - \bar{P})^2}$

Expand the numerator

- $\widehat{\alpha_1} = \frac{\beta_0 \sum(P_i - \bar{P})}{\sum(P_i - \bar{P})^2} + \frac{\beta_1 \sum(P_i - \bar{P})P_i}{\sum(P_i - \bar{P})^2} + \frac{\beta_2 \sum(P_i - \bar{P})Z_i}{\sum(P_i - \bar{P})^2} + \frac{\sum(P_i - \bar{P})\mu_i}{\sum(P_i - \bar{P})^2}$

Substitute for Y_i using true equation

- $\widehat{\alpha_1} = \frac{\sum(P_i - \bar{P})Y_i}{\sum(P_i - \bar{P})^2} = \frac{\sum(P_i - \bar{P})(\beta_0 + \beta_1 P_i + \beta_2 Z_i + \mu_i)}{\sum(P_i - \bar{P})^2}$

Expand the numerator

- $\widehat{\alpha_1} = \frac{\beta_0 \sum(P_i - \bar{P})}{\sum(P_i - \bar{P})^2} + \frac{\beta_1 \sum(P_i - \bar{P})P_i}{\sum(P_i - \bar{P})^2} + \frac{\beta_2 \sum(P_i - \bar{P})Z_i}{\sum(P_i - \bar{P})^2} + \frac{\sum(P_i - \bar{P})\mu_i}{\sum(P_i - \bar{P})^2}$

- $\widehat{\alpha_1} = \dfrac{\sum(P_i - \bar{P})Y_i}{\sum(P_i - \bar{P})^2} = \dfrac{\sum(P_i - \bar{P})\overbrace{(\beta_0 + \beta_1 P_i + \beta_2 Z_i + \mu_i)}^{\text{Substitute for } Y_i \text{ using true equation}}}{\sum(P_i - \bar{P})^2}$

Expand the numerator

- $\widehat{\alpha_1} = \dfrac{\beta_0 \sum(P_i - \bar{P})}{\sum(P_i - \bar{P})^2} + \dfrac{\beta_1 \sum(P_i - \bar{P})P_i}{\sum(P_i - \bar{P})^2} + \dfrac{\beta_2 \sum(P_i - \bar{P})Z_i}{\sum(P_i - \bar{P})^2} + \dfrac{\sum(P_i - \bar{P})\mu_i}{\sum(P_i - \bar{P})^2}$

- $$\widehat{\alpha_1} = \frac{\sum(P_i - \bar{P})Y_i}{\sum(P_i - \bar{P})^2} = \frac{\sum(P_i - \bar{P})\overbrace{(\beta_0 + \beta_1 P_i + \beta_2 Z_i + \mu_i)}^{\text{Substitute for } Y_i \text{ using true equation}}}{\sum(P_i - \bar{P})^2}$$

Expand the numerator

- $$\widehat{\alpha_1} = \frac{\beta_0 \sum(P_i - \bar{P})}{\sum(P_i - \bar{P})^2} + \frac{\beta_1 \sum(P_i - \bar{P})P_i}{\sum(P_i - \bar{P})^2} + \frac{\beta_2 \sum(P_i - \bar{P})Z_i}{\sum(P_i - \bar{P})^2} + \frac{\sum(P_i - \bar{P})\mu_i}{\sum(P_i - \bar{P})^2}$$

$$\widehat{\alpha_1} = \frac{\beta_0 \sum(P_i - \bar{P})}{\sum(P_i - \bar{P})^2} + \frac{\beta_1 \sum(P_i - \bar{P})P_i}{\sum(P_i - \bar{P})^2} + \frac{\beta_2 \sum(P_i - \bar{P})Z_i}{\sum(P_i - \bar{P})^2} + \frac{\sum(P_i - \bar{P})\mu_i}{\sum(P_i - \bar{P})^2}$$

Sum of deviations from mean is simply = 0

$$\widehat{\alpha_1} = \frac{\beta_0 \sum(P_i - \bar{P})}{\sum(P_i - \bar{P})^2} + \frac{\beta_1 \sum(P_i - \bar{P})P_i}{\sum(P_i - \bar{P})^2} + \frac{\beta_2 \sum(P_i - \bar{P})Z_i}{\sum(P_i - \bar{P})^2} + \frac{\sum(P_i - \bar{P})\mu_i}{\sum(P_i - \bar{P})^2}$$

Sum of deviations from mean is simply = 0

Can be shown that this is = $\sum(P_i - \bar{P})^2$, same as denominator

$$\widehat{\alpha_1} = \frac{\beta_0 \sum(P_i - \bar{P})}{\sum(P_i - \bar{P})^2} + \frac{\beta_1 \sum(P_i - \bar{P})P_i}{\sum(P_i - \bar{P})^2} + \frac{\beta_2 \sum(P_i - \bar{P})Z_i}{\sum(P_i - \bar{P})^2} + \frac{\sum(P_i - \bar{P})\mu_i}{\sum(P_i - \bar{P})^2}$$

Sum of deviations from mean is simply = 0

Can be shown that this is = $\sum(P_i - \bar{P})^2$, same as denominator

COV(P_i, μ_i) = 0 by OLS assumption

$$\widehat{\alpha_1} = \frac{\beta_0 \sum(P_i - \bar{P})}{\sum(P_i - \bar{P})^2} + \frac{\beta_1 \sum(P_i - \bar{P})P_i}{\sum(P_i - \bar{P})^2} + \frac{\beta_2 \sum(P_i - \bar{P})Z_i}{\sum(P_i - \bar{P})^2} + \frac{\sum(P_i - \bar{P})\mu_i}{\sum(P_i - \bar{P})^2}$$

Sum of deviations from mean is simply = 0

Can be shown that this is = $\sum(P_i - \bar{P})^2$, same as denominator

COV(P_i, μ_i) = 0 by OLS assumption

$$\widehat{\alpha_1} = \beta_1 + \frac{\beta_2 \sum(P_i - \bar{P})Z_i}{\sum(P_i - \bar{P})^2}$$

When we estimate (2) and the true relationship is (1)

(1) $Y_i = \beta_0 + \beta_1 P + \beta_2 Z + \mu_i$ $\quad$ *True relationship*

(2) $Y_i = \alpha_0 + \alpha_1 P + \varepsilon_i$ $\quad$ *Estimated relationship (Z omitted)*

We discover that the OLS estimate of α_1 is given by

$$\widehat{\alpha_1} = \beta_1 + \frac{\beta_2 \boxed{\sum(P_i - \bar{P})Z_i}}{\boxed{\sum(P_i - \bar{P})^2}}$$

When we estimate (2) and the true relationship is (1)

(1) $Y_i = \beta_0 + \beta_1 P + \beta_2 Z + \mu_i$ *True relationship*

(2) $Y_i = \alpha_0 + \alpha_1 P + \varepsilon_i$ *Estimated relationship (Z omitted)*

We discover that the OLS estimate of α_1 is given by

$$\widehat{\alpha_1} = \beta_1 + \frac{\beta_2 \boxed{\sum(P_i - \bar{P})Z_i}}{\boxed{\sum(P_i - \bar{P})^2}} = \beta_1 + \beta_2 \boxed{\frac{COV(P,Z)}{VAR(P)}}$$

Do you recognize this?

Refresher of ‘OLS regression refresher’

$For \quad Z_i = \gamma_0 + \gamma_1 P + \sigma_i \qquad the\ OLS\ estimator\ of\ \gamma_1\ is\ given\ by$

$$\widehat{\gamma_1} = \frac{\sum(P_i - \bar{P})Z_i}{\sum(P_i - \bar{P})^2} = \frac{COV(P,Z)}{VAR(P)}$$ (see slide 3)

Refresher of ‘OLS regression refresher’

$For \quad Z_i = \gamma_0 + \gamma_1 P + \sigma_i \qquad the\ OLS\ estimator\ of\ \gamma_1\ is\ given\ by$

$$\widehat{\gamma_1} = \frac{\sum(P_i - \bar{P})Z_i}{\sum(P_i - \bar{P})^2} = \frac{COV(P,Z)}{VAR(P)}$$ (see slide 3)

$$\widehat{\alpha_1} = \beta_1 + \beta_2 \frac{COV(P,Z)}{VAR(P)} = \beta_1 + \beta_2 * \boxed{\widehat{\gamma_1}}$$

When we estimate (2) and the true relationship is (1)

$$(1) \quad Y_i = \beta_0 + \beta_1 P + \beta_2 Z + \mu_i \qquad \textit{True relationship}$$

$$(2) \quad Y_i = \alpha_0 + \alpha_1 P \qquad + \varepsilon_i \qquad \textit{Estimated relationship (Z omitted)}$$

Then the estimate of α_1 is given by

$$\boxed{\widehat{\alpha_1} = \beta_1 + \beta_2 * \widehat{\gamma_1}}$$

"The omitted variable bias formula" or "the bias formula"

"The most important formula of your PhD career"

When we estimate (2) and the true relationship is (1)

$$(1) \quad Y_i = \beta_0 + \beta_1 P + \beta_2 Z + \mu_i \qquad \textit{True relationship}$$

$$(2) \quad Y_i = \alpha_0 + \alpha_1 P \qquad + \varepsilon_i \qquad \textit{Estimated relationship (Z omitted)}$$

Then the estimate of α_1 is given by

$$\widehat{\alpha_1} = \underbrace{\beta_1}_{\text{True ‘impact’ of P on Y}} + \underbrace{\boxed{\beta_2 * \widehat{\gamma_1}}}_{\text{Selection bias}}$$

When we estimate (2) and the true relationship is (1)

$$(1)\quad Y_i = \beta_0 + \beta_1 P + \beta_2 Z + \mu_i \qquad \textit{True relationship}$$

$$(2)\quad Y_i = \alpha_0 + \alpha_1 P \qquad\qquad + \varepsilon_i \qquad \textit{Estimated relationship (Z omitted)}$$

The sources of bias:

$$\widehat{\alpha_1} = \beta_1 + \boxed{\beta_2 * \widehat{\gamma_1}}$$

<u>Selection bias has two elements</u>

β_2 = the relationship between the excluded variable (Z) and the outcome (Y)

$\gamma 1$ = the relationship between the excluded variable (Z) and included variable (P)

Back to our example of the determinants of house spending in the U.S.

Renters pay 9.3% less than non-renters controlling for family size and race and region of residence

```
. reg log_h renter fam_size white black urban

      Source |       SS           df       MS      Number of obs   =    16,219
-------------+----------------------------------   F(5, 16213)     =    355.43
       Model |  1246.70862         5  249.341723   Prob > F        =    0.0000
    Residual |  11373.7896    16,213  .701522828   R-squared       =    0.0988
-------------+----------------------------------   Adj R-squared   =    0.0985
       Total |  12620.4982    16,218  .778178458   Root MSE        =    .83757

------------------------------------------------------------------------------
 log_housing | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      renter |  -.0934233   .0140227    -6.66   0.000    -.1209092   -.0659373
    fam_size |   .1527183   .0044738    34.14   0.000     .1439491    .1614875
       white |   -.097098   .0257591    -3.77   0.000    -.1475886   -.0466073
       black |   -.266627   .0313051    -8.52   0.000    -.3279884   -.2052655
       urban |   .4901121   .0251201    19.51   0.000     .4408739    .5393503
       _cons |   6.748645   .0374803   180.06   0.000     6.675179     6.82211
------------------------------------------------------------------------------
```

-After controlling income, coefficient goes from -.093 to +.119 (an increase of 0.212)

-Use the bias formula to understand selection bias in this case

```
. reg log_h renter fam_size white black urban log_pc

      Source |       SS           df       MS      Number of obs   =    16,219
-------------+----------------------------------   F(6, 16212)     =    898.23
       Model |  3148.72697         6  524.787828   Prob > F        =    0.0000
    Residual |  9471.77126    16,212  .584244464   R-squared       =    0.2495
-------------+----------------------------------   Adj R-squared   =    0.2492
       Total |  12620.4982    16,218  .778178458   Root MSE        =    .76436

------------------------------------------------------------------------------
 log_housing | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      renter |   .1187369   .0133262     8.91   0.000      .092616    .1448578
    fam_size |   .2094188    .004202    49.84   0.000     .2011824    .2176551
       white |  -.0578265   .0235176    -2.46   0.014    -.1039236   -.0117294
       black |  -.1350735   .0286616    -4.71   0.000    -.1912535   -.0788935
       urban |   .3203993   .0231166    13.86   0.000     .2750882    .3657103
   log_pcinc |   .3999453   .0070096    57.06   0.000     .3862058    .4136848
       _cons |   2.658703    .079424    33.47   0.000     2.503023    2.814383
------------------------------------------------------------------------------
```

-Bias formula: estimated coefficient = true coefficient + COV(Y,Z)*COV(Z,P)

-Bias formula: -0.093 = .119 + {COV(income, house spending)*COV(income, renter)}

-0.093 = .119 + { (+) * (-) }

-0.093 = .119 + { -.212}

-Large negative selection bias (renters have significantly lower income than non-renters)

```
. reg log_h renter fam_size white black urban log_pc

      Source |       SS           df       MS      Number of obs   =    16,219
-------------+----------------------------------   F(6, 16212)     =    898.23
       Model |  3148.72697         6  524.787828   Prob > F        =    0.0000
    Residual |  9471.77126    16,212  .584244464   R-squared       =    0.2495
-------------+----------------------------------   Adj R-squared   =    0.2492
       Total |  12620.4982    16,218  .778178458   Root MSE        =    .76436

------------------------------------------------------------------------------
 log_housing | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      renter |   .1187369   .0133262     8.91   0.000      .092616    .1448578
    fam_size |   .2094188    .004202    49.84   0.000     .2011824    .2176551
       white |  -.0578265   .0235176    -2.46   0.014    -.1039236   -.0117294
       black |  -.1350735   .0286616    -4.71   0.000    -.1912535   -.0788935
       urban |   .3203993   .0231166    13.86   0.000     .2750882    .3657103
   log_pcinc |   .3999453   .0070096    57.06   0.000     .3862058    .4136848
       _cons |   2.658703    .079424    33.47   0.000     2.503023    2.814383
------------------------------------------------------------------------------
```

```
. reg childlabour attending_n male  pcincome if ML_ZNZ==1

      Source |       SS           df       MS      Number of obs   =    11,617
-------------+----------------------------------   F(3, 11613)     =    549.12
       Model |  276.630299         3  92.2100997   Prob > F        =    0.0000
    Residual |  1950.09278    11,613  .167923257   R-squared       =    0.1242
-------------+----------------------------------   Adj R-squared   =    0.1240
       Total |  2226.72308    11,616   .19169448   Root MSE        =    .40978

------------------------------------------------------------------------------
 childlabour | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
attending_now |  -.4080822   .0101163   -40.34   0.000     -.427912   -.3882525
        male |   .0272711    .007605     3.59   0.000      .012364    .0421783
    pcincome |  -7.73e-09   4.95e-09    -1.56   0.118    -1.74e-08    1.97e-09
       _cons |   .5843696   .0099986    58.45   0.000     .5647706    .6039686
------------------------------------------------------------------------------
```

-TZ Labour Force Survey: Among children age 6-17, current enrollment is associated with 40.8 percentage point lower probability of child labour (linear probability model)

-We forgot to control for age—omitted variable bias

-What is the direction of bias? Is -40.8 too low or too high?

$$\widehat{\alpha_1} = \beta_1 + \beta_2 * \widehat{\gamma_1}$$

-40.8 = true + {COV(Z,Y)*COV(Z,P)}

-40.8 = true + {COV(age, labor)*COV(age, schooling)}

- What is the correlation between age and child labor? negative or positive?
- What is the correlation between age and schooling? negative or positive?
- Is -40.80 too high or too low?

```
. reg childlabour attending_n male age pcincome if ML_ZNZ==1

      Source |       SS           df       MS      Number of obs   =    11,617
-------------+----------------------------------   F(4, 11612)     =    666.34
       Model |  415.692168         4  103.923042   Prob > F        =    0.0000
    Residual |  1811.03091    11,612  .155962014   R-squared       =    0.1867
-------------+----------------------------------   Adj R-squared   =    0.1864
       Total |  2226.72308    11,616   .19169448   Root MSE        =    .39492

------------------------------------------------------------------------------
 childlabour | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
attending_now |  -.3518561   .0099296   -35.44   0.000    -.3713197   -.3323925
        male |    .028782   .0073294     3.93   0.000     .0144152    .0431488
         age |    .031503    .001055    29.86   0.000      .029435     .033571
    pcincome |  -6.16e-09   4.77e-09    -1.29   0.197    -1.55e-08    3.19e-09
       _cons |   .1835022    .016525    11.10   0.000     .1511105     .215894
------------------------------------------------------------------------------
```

-After controlling age, coefficient goes up -35.2
-Bias term is negative: COV(age, labour)*COV(age, schooling) = (+)* (-)
- -40.8 = -35.2 + {(+)*(-)}
- -40.8 = -35.2 + (-5.6)

In the presence of omitted variable, estimated coefficient is

$$\widehat{\alpha_1} = \beta_1 + \beta_2 * \widehat{\gamma_1}$$

	Omitted variable (Z)	COV(Z,P)=γ_1	COV(Z,Y)=β_2	Direction of bias
Y=earnings, P=training				
Y=health, P=insurance				
Y=harvest, P=HY seeds				
Y=weight, P=nutrition program				
Y=revenue, P=obtained loan				
Y=IPV, P=yrs education				

1) Modern applied economics is about <u>counterfactual thinking</u>

2) Identify the counterfactual

3) Is there selection bias?

4) Use toolkit to resolve the selection bias

   Add relevant covariates (what do we mean by relevant?)

   Study design (matching, discontinuity, randomization)

   Econometrics (fixed-effects, DiD)