INTERMEDIATE QUANTITATIVE METHODS

Measurement Error and Missing Data

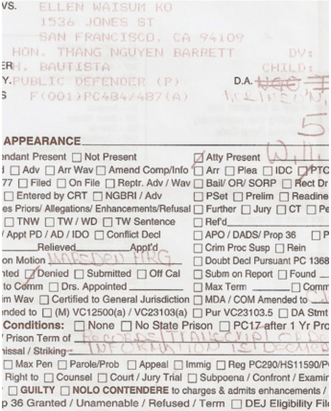
Per Engzell (Nuffield College)

Hilary term 2018

Sociology Department, University of Oxford

MEASUREMENT ERROR

How did we get from this ...


VS. ELLEN WAISUM KO
1536 JONES ST
SAN FRANCISCO, CA 94109
HON. THANG NGUYEN BARRETT DV:
ER H. BAUTISTA CHILD:
Y. PUBLIC DEFENDER (P) D.A.
S F(001)PC484/487(A)
APPEARANCE
endant Present ☐ Not Present ☑ Atty Present
☐ Adv ☐ Arr Wav ☐ Amend Comp/Info ☐ Arr ☐ Plea ☐ IDC ☐ PTC
77 ☐ Filed ☐ On File ☐ Reptr. Adv / Wav ☐ Bail/ OR/ SORP ☐ Rect Dr
☐ Entered by CRT ☐ NGBRI / Adv ☐ PSet ☐ Prelim ☐ Readine
es Priors/ Allegations/ Enhancements/Refusal ☐ Further ☐ Jury ☐ CT ☐ Pe
☐ TNW ☐ TW / WD ☐ TW Sentence ☐ Ref'd
Appt PD / AD / IDO ☐ Conflict Decl ☐ APO / DADS/ Prop 36
Relieved Appt'd ☐ Crim Proc Susp ☐ Rein
on Motion ☐ Doubt Decl Pursuant PC 1368
nted ☑ Denied ☐ Submitted ☐ Off Cal ☐ Subm on Report ☐ Found
to Comm ☐ Drs. Appointed ☐ Max Term ☐ Comm
im Wav ☐ Certified to General Jurisdiction ☐ MDA / COM Amended to
nded to ☐ (M) VC12500(a) / VC23103(a) ☐ Pur VC23103.5 ☐ DA Stmt
Conditions: ☐ None ☐ No State Prison ☐ PC17 after 1 Yr Pro
Prison Term of
nissal / Striking
☐ Max Pen ☐ Parole/Prob ☐ Appeal ☐ Immig ☐ Reg PC290/HS11590/P
Right to ☐ Counsel ☐ Court / Jury Trial ☐ Subpoena / Confront / Examin
☐ GUILTY ☐ NOLO CONTENDERE to charges & admits enhancements /
p 36 Granted / Unamenable / Refused / Term ☐ DEJ Eligibility Fil

... to this?

	id	v3	v4	v5
1	2630	22	3	1930
2	8590	23	2	1745
3	7523	23	4	1700
4	8114	19	1	1730
5	9036	24	6	1000
6	2270	23	6	1010
7	8290	16	4	1000
8	1738	20	2	1600
9	7498	17	4	1130
10	11240	20	4	1845
11	3995	0	0	9900
12	1538	19	5	1715
13	5960	17	1	1300
14	11556	22	3	1755
15	7090	21	2	1800
16	10896	24	4	1930
17	3467	24	3	1800
18	7615	0	0	9900
19	4111	20	1	1700

KNOW YOUR DATA

Your variables do not equal the concepts you want to measure!

Comprehension

↓

Retrieval

↓

Judgement

↓

Reporting

To provide data, every respondent has to

- interpret the question (ideally in the same way as each other and the researcher)
- gather the information from memory
- make a judgement about how it corresponds to given alternatives
- formulate and communicate an answer

In each of these steps, mistakes and random variability are inevitable. Additional errors arise from vague response options, coarseness of categories, coding errors etc.

Adapted from: Robert M. Groves, et al. Survey Methodology (Wiley, 2011).

Often, the x and y in our data are imperfect representations of the concepts we are interested in

For example, we want to know about someone's overall life satisfaction but what we observe is how they ticked a box on a 1–5 scale on a particular day

Even for seemingly straightforward concepts like income, survey responses lack accuracy and response categories may be limited (e.g., 10 different income brackets)

An important type of error is proxy error. Say we want to measure wealth but only know about home ownership (yes/no). This is certainly part of wealth but not all of it!

In formal terms, we are unable to observe the true variable x^* but instead we observe $x = x^* + w$, where w is measurement error

How does this affect the conclusions we are able to draw?

Recap of: Variance, Standard Deviation, Covariance, Correlation

Variance of x

- Written as Var(x) or σ_x^2
- $E\{(x - \mu_x)^2\}$
- The expected squared distance between x and its mean
- Ranges between 0 and ∞

Recap of: Variance, Standard Deviation, Covariance, Correlation

Variance of x

- Written as Var(x) or σ_x^2
- $E\{(x - \mu_x)^2\}$
- The expected squared distance between x and its mean
- Ranges between 0 and ∞

VARIANCE AND COVARIANCE

Recap of: Variance, Standard Deviation, Covariance, Correlation

Variance of x

- Written as Var(x) or σ_x^2
- $E\{(x - \mu_x)^2\}$
- The expected squared distance between x and its mean
- Ranges between 0 and ∞

Recap of: Variance, Standard Deviation, Covariance, Correlation

Variance of x

- Written as Var(x) or σ_x^2
- $E\{(x - \mu_x)^2\}$
- The expected squared distance between x and its mean
- Ranges between 0 and ∞

Recap of: Variance, Standard Deviation, Covariance, Correlation

Variance of x

- Written as Var(x) or σ_x^2
- $E\{(x - \mu_x)^2\}$
- The expected squared distance between x and its mean
- Ranges between 0 and ∞

Recap of: Variance, Standard Deviation, Covariance, Correlation

Variance of x

- Written as Var(x) or σ_x^2
- $E\{(x - \mu_x)^2\}$
- The expected squared distance between x and its mean
- Ranges between 0 and ∞

Recap of: Variance, Standard Deviation, Covariance, Correlation

Variance of x

- Written as Var(x) or σ_x^2
- $E\{(x - \mu_x)^2\}$
- The expected squared distance between x and its mean
- Ranges between 0 and ∞

Covariance of (x, y)

- Written as $=$ Cov(x, y) or σ_{xy}^2
- $E\{(x - \mu_x)(y - \mu_y)\}$
- The expected product of the distance between x and its mean and the distance between y and its mean
- Ranges between $-\infty$ and $+\infty$

Recap of: Variance, Standard Deviation, Covariance, Correlation

Variance of x

- Written as Var(x) or σ_x^2
- $E\{(x-\mu_x)^2\}$
- The expected squared distance between x and its mean
- Ranges between 0 and ∞

Covariance of (x, y)

- Written as $=$ Cov(x, y) or σ_{xy}^2
- $E\{(x-\mu_x)(y-\mu_y)\}$
- The expected product of the distance between x and its mean and the distance between y and its mean
- Ranges between $-\infty$ and $+\infty$

VARIANCE AND COVARIANCE

Recap of: Variance, Standard Deviation, Covariance, Correlation

Variance of x

- Written as Var(x) or σ_x^2
- $E\{(x - \mu_x)^2\}$
- The expected squared distance between x and its mean
- Ranges between 0 and ∞

Covariance of (x, y)

- Written as $=$ Cov(x, y) or σ_{xy}^2
- $E\{(x - \mu_x)(y - \mu_y)\}$
- The expected product of the distance between x and its mean and the distance between y and its mean
- Ranges between $-\infty$ and $+\infty$

Recap of: Variance, Standard Deviation, Covariance, Correlation

Variance of x

- Written as Var(x) or σ_x^2
- $E\{(x - \mu_x)^2\}$
- The expected squared distance between x and its mean
- Ranges between 0 and ∞

Covariance of (x, y)

- Written as $=$ Cov(x, y) or σ_{xy}^2
- $E\{(x - \mu_x)(y - \mu_y)\}$
- The expected product of the distance between x and its mean and the distance between y and its mean
- Ranges between $-\infty$ and $+\infty$

Recap of: Variance, Standard Deviation, Covariance, Correlation

Variance of x

- Written as Var(x) or σ_x^2
- $E\{(x-\mu_x)^2\}$
- The expected squared distance between x and its mean
- Ranges between 0 and ∞

Covariance of (x, y)

- Written as $=$ Cov(x, y) or σ_{xy}^2
- $E\{(x-\mu_x)(y-\mu_y)\}$
- The expected product of the distance between x and its mean and the distance between y and its mean
- Ranges between $-\infty$ and $+\infty$

Recap of: Variance, Standard Deviation, Covariance, Correlation

Variance of x

- Written as Var(x) or σ_x^2
- $E\{(x - \mu_x)^2\} = E\{(x - \mu_x)(x - \mu_x)\}$
- The expected squared distance between x and its mean
- Ranges between 0 and ∞

Covariance of (x, y)

- Written as $=$ Cov(x, y) or σ_{xy}^2
- $E\{(x - \mu_x)(y - \mu_y)\}$
- The expected product of the distance between x and its mean and the distance between y and its mean
- Ranges between $-\infty$ and $+\infty$

Recap of: Variance, Standard Deviation, Covariance, Correlation

Standard deviation of x

- Written as SD(x) or σ_x
- Simply the square root of the variance: $\sqrt{E\{(x-\mu_x)^2\}}$
- Also ranges between 0 and ∞

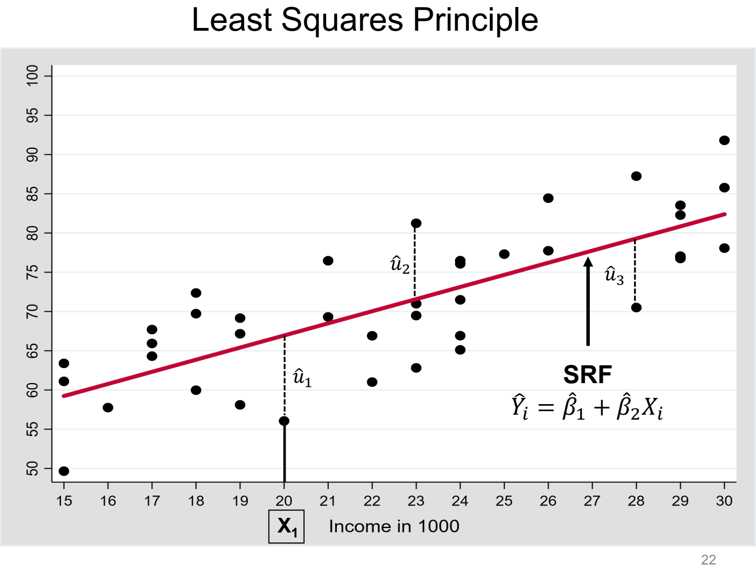
Correlation of (x, y)

- Written as $\mathrm{corr}(x, y)$ or r_{xy}
- The covariance of (x, y) divided by the product of their standard deviations:

$$\frac{E\{(x-\mu_x)(y-\mu_y)\}}{\sigma_x \sigma_y}$$

- Ranges between -1 and $+1$

Remember this?



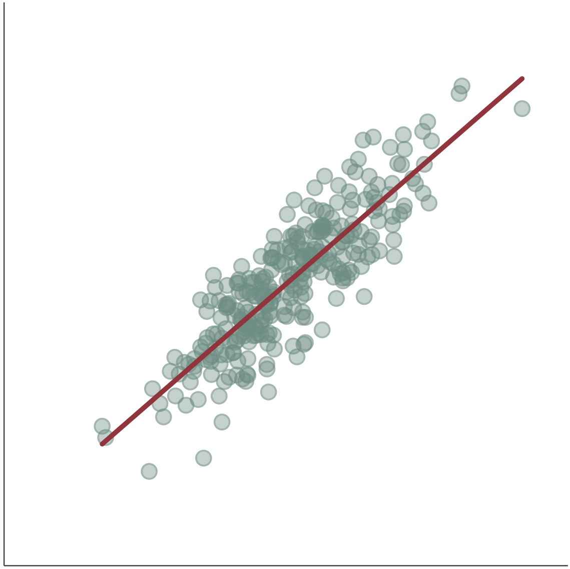
It is a remarkable fact about the world that the slope that minimises the squared deviations is described by:

$$\frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}$$

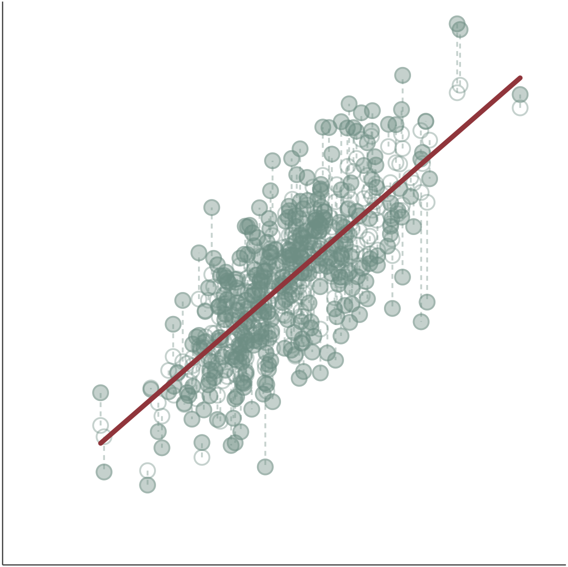
This extends to multivariate regression (but the notation becomes a bit more complex)

Keep this in mind, it will be useful later

- True relationship: $y^* = \beta_1 + \beta_2 x + u$

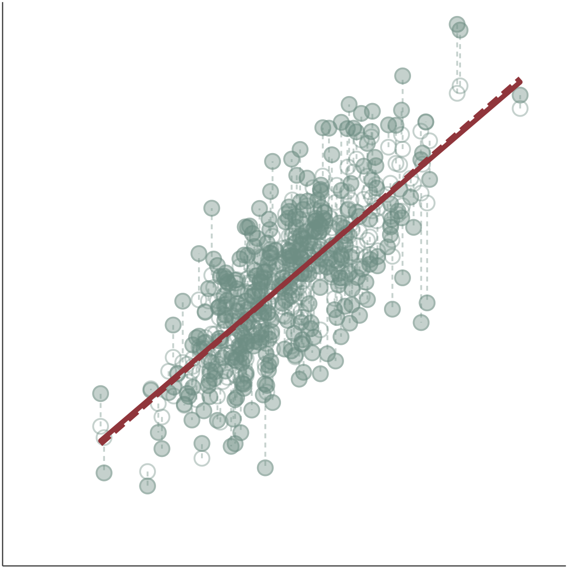


- True relationship: $y^* = \beta_1 + \beta_2 x + u$
- Adding error in y: $y = y^* + w$



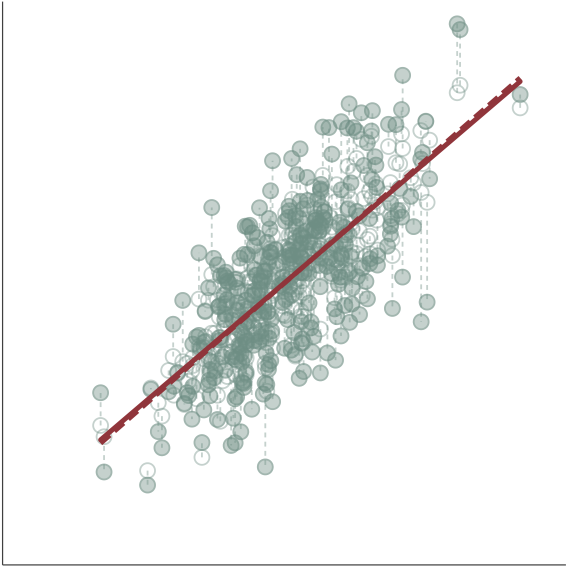
RANDOM NOISE IN Y

- True relationship: $y^* = \beta_1 + \beta_2 x + u$
- Adding error in y: $y = y^* + w$
- Estimated: $y = \beta_1 + \beta_2 x + u$



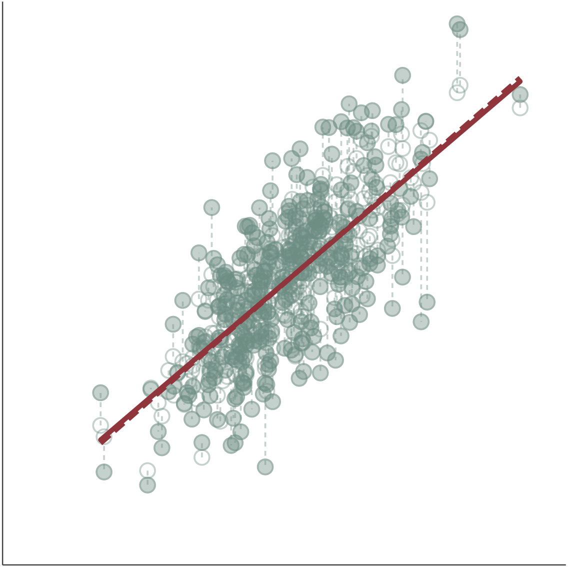
RANDOM NOISE IN Y

- True relationship: $y^* = \beta_1 + \beta_2 x + u$
- Adding error in y: $y = y^* + w$
- Estimated: $y = \beta_1 + \beta_2 x + u$
- No difference here!

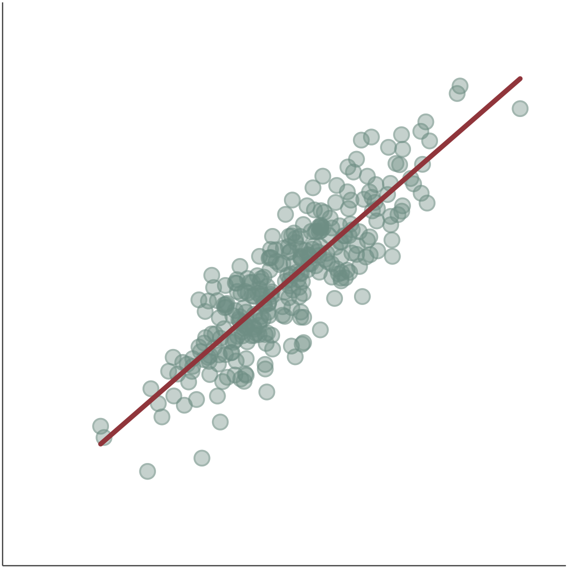


RANDOM NOISE IN Y

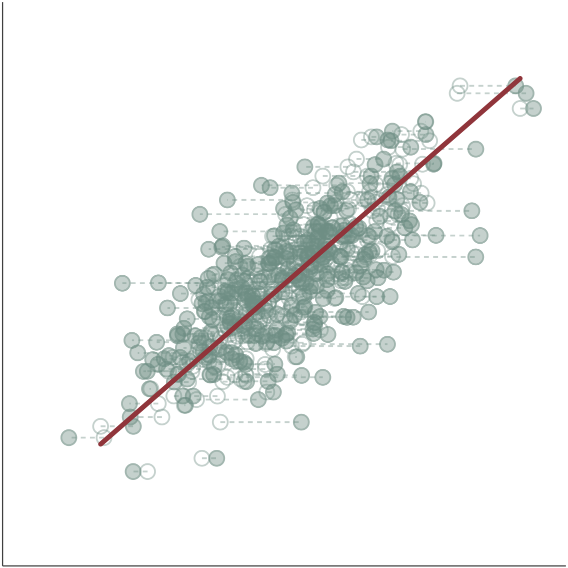
- True relationship: $y^* = \beta_1 + \beta_2 x + u$
- Adding error in y: $y = y^* + w$
- Estimated: $y = \beta_1 + \beta_2 x + u$
- No difference here!
- But: increased sampling variance



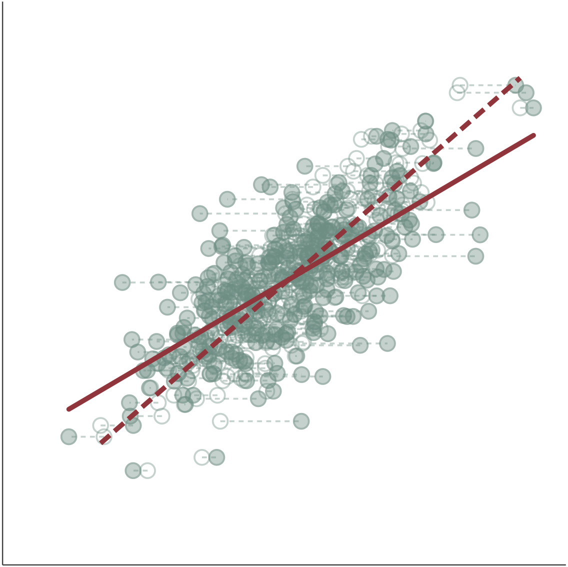
- True relationship:
  $y = \beta_1 + \beta_2 x^* + u$



- True relationship: $y = \beta_1 + \beta_2 x^* + u$
- Adding error in x: $x = x^* + w$

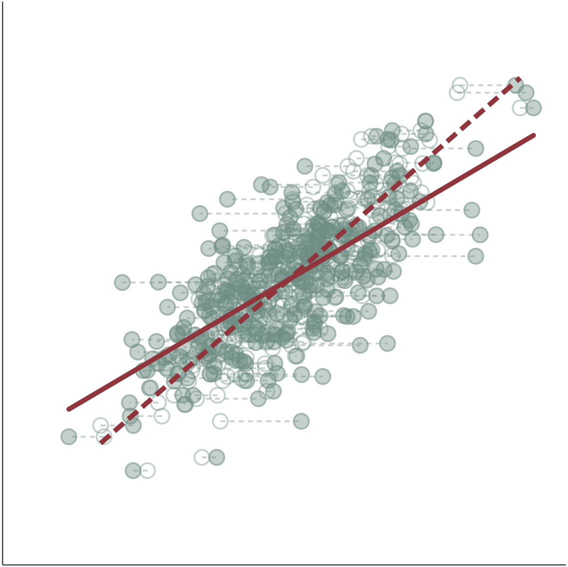


- True relationship: $y = \beta_1 + \beta_2 x^* + u$
- Adding error in x: $x = x^* + w$
- Estimated: $y = \beta_1 + \beta_2 x + u$



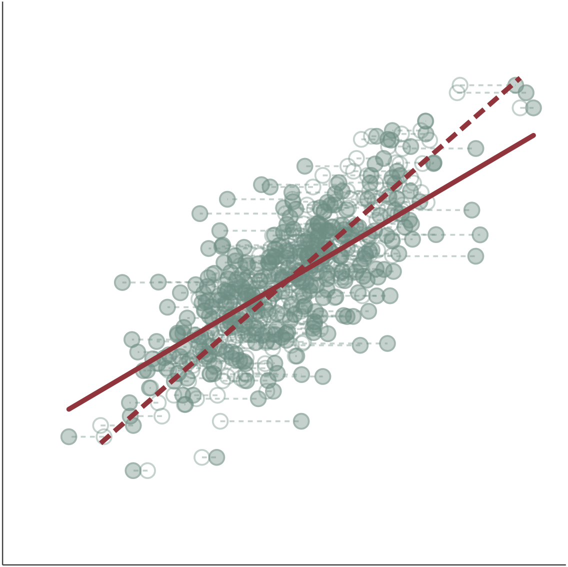
RANDOM NOISE IN X

- True relationship: $y = \beta_1 + \beta_2 x^* + u$
- Adding error in x: $x = x^* + w$
- Estimated: $y = \beta_1 + \beta_2 x + u$
- Look, β_2 is biased downward!



RANDOM NOISE IN X

- True relationship: $y = \beta_1 + \beta_2 x^* + u$
- Adding error in x: $x = x^* + w$
- Estimated: $y = \beta_1 + \beta_2 x + u$
- Look, β_2 is biased downward!
- Also: increased sampling variance



We want to estimate $y_i = \beta_1 + \beta_2 x_i^* + u_i$.

Unable to observe x_i^* we instead observe $x_i = x_i^* + w_i$ where w_i is the measurement error.

Naively substituting x_i for x_i^* we are led to estimate $y_i = \beta_1 + \beta_2 x_i + (u_i - \beta_2 w_i)$.

Given assumptions that x_i^*, w_i, u_i are joint normal with mean and covariance matrices $(\mu_x, 0, 0)$ and $\mathrm{diag}(\sigma^2_{x^*}, \sigma^2_w, \sigma^2_u)$,

$$\text{plim}\ \widehat{\beta}_2 = \beta_2 \left[1 - \frac{\text{Var}(w)}{\text{Var}(x)} \right]$$

The bracketed expression is bounded between 0 and 1 and the result is one of attenuation of the true gradient.

What is going on? We know that:

$$\beta_2 = \frac{\text{Cov}(x, y)}{\text{Var}(x)}$$

Adding random noise to y increases Var(y) but affects neither Cov(x, y) nor Var(x) so β_2 remains the same

Adding random noise to x increases Var(x) while Cov(x, y) remains the same and leads to a lower β_2

In fact, the bias is exactly proportional to the amount of variance in the observed x that is due to measurement error:

$$\text{plim}\ \widehat{\beta}_2 = \beta_2 \frac{\text{Var}(x^*)}{\text{Var}(x)} = \beta_2 \frac{\text{Var}(x^*)}{\text{Var}(x^*) + \text{Var}(w)} = \beta_2 \left[1 - \frac{\text{Var}(w)}{\text{Var}(x)}\right]$$

Example: How strong is the relationship between parents' income and their children's income in adulthood?

$$\ln(y_{\text{child}}) = \beta_1 + \beta_2 \ln(y_{\text{parent}}) + u$$

β_2, called the "intergenerational elasticity", is around 0.50 in the UK and US but for a long time people believed it was closer to 0.15–0.20

- "almost all earnings advantages and disadvantages of ancestors are wiped out in three generations." (Becker and Tomes 1986: 32)

Why? Two types of errors

- Reporting errors, especially when individuals were asked to remember their parents' income
- Using single-year measures of income that vary from year to year, rather than income averaged over many years

What we have seen so far:

```
clear
drawnorm x u w, n(5000)

gen y = x + u
reg y x

replace x = x + w
reg y x
```

Now, let's move to the multivariate case

```
clear

#delimit;
matrix C =
  [  1, .5, .5 \
    .5,  1, .5 \
    .5, .5,  1 ];
#delimit cr

drawnorm x1 x2 x3, means(0 0 0) cov(C) n(5000)
drawnorm u

gen y = x1 + u

reg y x1 x2 x3
```

Source	SS	df	MS
Model	5068.24898	3	1689.41633
Residual	5009.98048	4,996	1.00279833
Total	10078.2295	4,999	2.0160491

Number of obs	=	5,000
F(3, 4996)	=	1684.70
Prob > F	=	0.0000
R-squared	=	0.5029
Adj R-squared	=	0.5026
Root MSE	=	1.0014

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	1.004857	.0172947	58.10	0.000	.9709519	1.038762
x2	.0030808	.0172442	0.18	0.858	-.0307255	.0368871
x3	-.0071945	.0173546	-0.41	0.678	-.0412171	.0268281
_cons	-.0162269	.0141622	-1.15	0.252	-.0439911	.0115373

```
drawnorm w
replace x1 = x1 + w

reg y x1 x2 x3
```

Source	SS	df	MS
Model	3002.10501	3	1000.70167
Residual	7076.12444	4,996	1.41635797
Total	10078.2295	4,999	2.0160491

Number of obs	=	5,000
F(3, 4996)	=	706.53
Prob > F	=	0.0000
R-squared	=	0.2979
Adj R-squared	=	0.2975
Root MSE	=	1.1901

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	.4016158	.0131599	30.52	0.000	.3758167	.4274149
x2	.2012384	.0198348	10.15	0.000	.1623535	.2401234
x3	.2066106	.01986	10.40	0.000	.1676764	.2455449
_cons	-.0238348	.016833	-1.42	0.157	-.0568349	.0091653

Which scenario are you most likely to encounter in the wild?

Which scenario are you most likely to encounter in the wild?

Answer: you should expect the second

Which scenario are you most likely to encounter in the wild?

Answer: you should expect the second

Whenever a variable is measured with (classical) error and there are other correlated variables in the model

- Attenuation for the mismeasured variable becomes worse
- Estimates for correlated covariates will be subject to an opposite, upward bias

If someone claims to have "controlled for" something, you should usually be wary. Often there will be residual confounding (but some variables are nearly error free, like gender).

Example: suppose we want to know if grandparents influence their grandchildren's education

Should we run the following regression?

$$\text{edu}_{\text{child}} = \beta_1 + \beta_2 \text{edu}_{\text{parent}} + \beta_3 \text{edu}_{\text{grandparent}} + u$$

A lot of people do, but it will not tell us much: β_3 is almost certainly positive even if grandparents are of no consequence at all

Why? Parents influence their children in a lot of ways that are not in the model

- Other parent's education (i.e. mother if $\text{edu}_{\text{parent}}$ is father)
- Social class, income, culture, interests, etc
- Genetic inheritance

```
. sysuse census
(1980 Census data by state)

. keep pop death marriage divorce

. codebook, compact

Variable    Obs Unique      Mean     Min       Max  Label
-----------------------------------------------------------------------------
pop          50     50   4518149  401851  2.37e+07  Population
death        50     50  39474.26    1604    186428  Number of deaths
marriage     50     50   47701.4    4437    210864  Number of marriages
divorce      50     50  23679.44    2142    133541  Number of divorces
-----------------------------------------------------------------------------

.
. corr death marriage divorce
(obs=50)

             |    death marriage  divorce
-------------+---------------------------
       death |   1.0000
    marriage |   0.8921   1.0000
     divorce |   0.9003   0.9349   1.0000
```

```
. factor death marriage divorce
(obs=50)

Factor analysis/correlation                    Number of obs    =       50
    Method: principal factors                  Retained factors =        1
    Rotation: (unrotated)                      Number of params =        3
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	2.69024	2.72280	1.0293	1.0293
Factor2	-0.03256	0.01145	-0.0125	1.0168
Factor3	-0.04401	.	-0.0168	1.0000

```
LR test: independent vs. saturated:  chi2(3)  =  185.33 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances
```

Variable	Factor1	Uniqueness
death	0.9238	0.1466
marriage	0.9555	0.0870
divorce	0.9612	0.0762

```
. predict factpop
(regression scoring assumed)

Scoring coefficients (method = regression)
```

Variable	Factor1
death	**0.21062**
marriage	**0.36780**
divorce	**0.42768**

```
. pca death marriage divorce, comp(1)

Principal components/correlation                  Number of obs    =         50
                                                  Number of comp.  =          1
                                                  Trace            =          3
    Rotation: (unrotated = principal)             Rho              =     0.9394
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	2.81832	2.70133	0.9394	0.9394
Comp2	.116989	.0522963	0.0390	0.9784
Comp3	.0646923	.	0.0216	1.0000

Principal components (eigenvectors)

Variable	Comp1	Unexplained
death	0.5718	.07843
marriage	0.5792	.05442
divorce	0.5809	.04883

```
. predict pcapop
(score assumed)

Scoring coefficients
    sum of squares(column-loading) = 1
```

Variable	Comp1
death	**0.5718**
marriage	**0.5792**
divorce	**0.5809**

```
. corr pop fact pca death marriage divorce
(obs=50)
```

	pop	factpop	pcapop	death	marriage	divorce
pop	1.0000					
factpop	0.9679	1.0000				
pcapop	0.9778	0.9986	1.0000			
death	0.9876	0.9443	0.9600	1.0000		
marriage	0.9179	0.9767	0.9724	0.8921	1.0000	
divorce	0.9383	0.9825	0.9753	0.9003	0.9349	1.0000

Factor analysis or principal components are ways to extract the common information in several measures, that can then be stored and entered into a regular regression

However, entering predicted FA or PCA scores into a model ignores remaining uncertainty and leads to underestimation

It is common convention to report this uncertainty as Cronbach's alpha (to access this in Stata, type `alpha x1 x2 ...`):

$$\alpha = \frac{K}{K-1}\left[1 - \frac{\sum \sigma_k^2}{\sigma_{\text{total}}^2}\right]$$

Ranges from 0 to 1, and >.8 often thought "acceptable" (K is number of items, $\sum \sigma_k^2$ is sum of their variances, σ_{total}^2 variance of total)

A more sophisticated approach is to not only minimise error, but try to correct for it

With classical measurement error the bias follows from the signal-to-total-variance ratio, so we need knowledge about that ratio

$$\text{plim } \widehat{\beta}_2 = \beta_2 \frac{\text{Var}(x^*)}{\text{Var}(x)}$$

Such knowledge can come in the form of

a) repeated observations
b) auxiliary data (or ”guesstimate”)

With repeated observations:
Structural equation modeling (SEM), instrumental variables (IV)

With auxiliary data: Errors-in-variables (EIV) regression

Structural equation modeling (SEM) is a way of thinking about models, of drawing them, and an estimation method – all at once

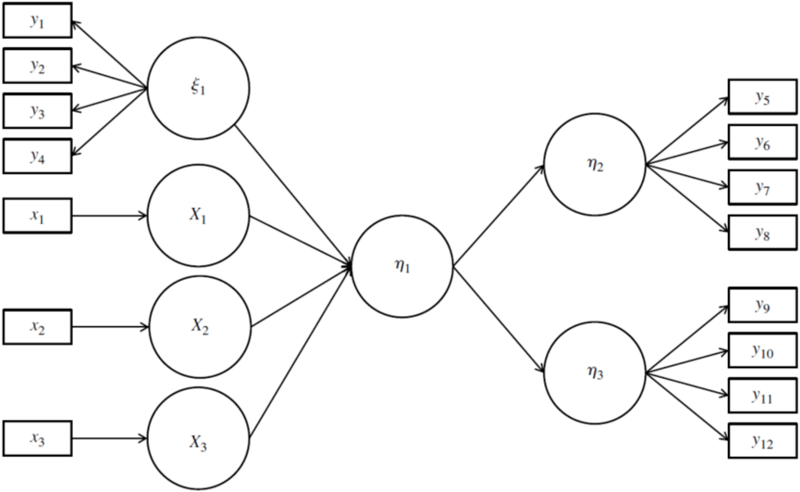
We distinguish between latent/unobserved variables (or constructs, drawn as "bubbles") and manifest/observed variables (or indicators, drawn as boxes)

Syntax differs a bit from the usual Stata convention:

```
sem (x1<-X) (x2<-X) (x3<-X) (y<-X)
```

Lowercase variables (y, x1, x2, ...) are observed variables. Uppercase variables (here: X) are unobserved. Arrow (<-) denotes causality or "is a function of".

Structural equation modeling is very flexible and easy to use, but not very common in sociology


y_1
y_2
y_3
y_4
ξ_1
x_1
X_1
x_2
X_2
x_3
X_3
η_1
η_2
η_3
y_5
y_6
y_7
y_8
y_9
y_{10}
y_{11}
y_{12}

A separate Stata command is `gsem` where ”g” stands for ”Generalized” SEM.

`gsem` can do some things that `sem` can't, like estimating multilevel and nonlinear models with latent variables.

But `sem` is superior to `gsem` in other respects, including survey weighting and clustering, missing data models, goodness-of-fit statistics, etc

Basic syntax is similar between the two

A separate Stata command is `gsem` where "g" stands for "Generalized" SEM.

`gsem` can do some things that `sem` can't, like estimating multilevel and nonlinear models with latent variables.*

But `sem` is superior to `gsem` in other respects, including survey weighting and clustering, missing data models, goodness-of-fit statistics, etc

Basic syntax is similar between the two

* Many of these things can also be done with an independently developed Stata command `gllamm`: generalized linear latent and mixed models (to install this, type `ssc install gllamm`)

Instrumental variables

```
. ivregress 2sls y (x1 = x2 x3)

Instrumental variables (2SLS) regression          Number of obs   =      5,000
                                                  Wald chi2(1)    =     821.10
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =          .
                                                  Root MSE        =     1.4296
```

y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
x1	1.021057	.035633	28.65	0.000	.951218	1.090897
_cons	-.0359843	.0202257	-1.78	0.075	-.075626	.0036573

```
Instrumented:  x1
Instruments:   x2 x3
```

Errors-in-variables (EIV) regression

```
. eivreg y x1 x2 x3, reliab(x1 .50)

                  assumed                  Errors-in-variables regression
    variable    reliability
 ---------------------------               Number of obs     =      5,000
          x1       0.5000                  F(   3,      4996) =     978.29
           *       1.0000                  Prob > F          =     0.0000
                                           R-squared         =     0.4929
                                           Root MSE          =    1.01139
```

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	1.000062	.0278484	35.91	0.000	.9454671	1.054657
x2	-.0211536	.019338	-1.09	0.274	-.0590647	.0167575
x3	.038338	.018338	2.09	0.037	.0023875	.0742884
_cons	-.0355175	.0143139	-2.48	0.013	-.063579	-.0074559

Classical error in x is bad: bias toward the null

Classical error in y: not nearly as bad

With more than one x in the model and correlated variables, measurement error will contaminate even the variables that are accurately measured

Ignorance is not an option: assuming that you observe everything without error is also an assumption

Models such as SEM can be used to correct for error in special cases and given that assumptions are satisified, but often we have to live with uncertainty

If someone claims to have ”controlled for” something, be sceptical and examine the measures critically. Usually there is likely to be residual confounding

Questions?