# Consequences of measurement error and how to deal with it

Albert Satorra
Universitat Pompeu Fabra
&
BI Norwegian Business School

13.00-14.00, Applications of Psychometrics
The Norwegian Psychometrics Gathering

Oslo, 14th December, 2022

# Overview

# Basics of theory and practice of measurement error

# The simple measurement error model

A universe of units $u_i \in \mathcal{U}$ and a continuous variable $X$ measured with error

$$X(u_i) = x(u_i) + \epsilon(u_i)$$

$X_i \equiv X(u_i); x_i \equiv x(u_i); \epsilon_i \equiv \epsilon(u_i)$
From one observed variable $X$, we get three vars. connected by equality:

$$\text{Observed} = \text{True} + \text{Error}$$

Not yet a model; just a tautology, since

$$\epsilon_i = X_i - x_i \qquad (1)$$

In population notation, we assume

$$X = x + \epsilon$$

$\epsilon$ is random (not determined, as in (1)), of zero mean and uncorrelated with the true value $x$ ($E\epsilon \mid x = 0$); or, a more strong assumption, $\epsilon$ is independent of $x$. We could also assume $\epsilon$ is normally distributed.

# Reliability and validity of a measure X

Assume $\epsilon$ is a centered random variable with variance $\sigma^2_\epsilon$ and distribution free of $x$. In the population $\mathcal{U}$, variation of $X$ is due: to $x$s and to $\epsilon$s.

$$\text{var}(X) = \text{var}(x) + \text{var}(\epsilon) = \sigma^2_x + \sigma^2_\epsilon$$

The *Reliability* of $X$, is the ratio of two variances

$$k_X = \sigma^2_x / \sigma^2_X$$

i.e., the ratio of variation of the true values $x$s (in the population) divided by the variation of the observed $X$. Note that $0 \leq k_K \leq 1$, $\sigma_x = \sqrt{k_X}\sigma_X$ and that $k_x$ is a population ( $\mathcal{U}$ ) dependent measure.

The *Validity* of $X$ is the correlation between the observed $X$ and the true $x$:

$$\text{cor}(X, x) = \text{cov}(x + \epsilon, x)/(\sigma_X \sigma_x)$$
$$= \text{var}(x)/(\sigma_X \sigma_x) = \sigma_x / \sigma_X = k_X^{1/2}$$

1

[1] In psychometrics, validity and reliability are close-related concepts. In survey statistics: validity is associated with $\epsilon$ centered (free of *bias*, equality of expected values of $X$ and $x$); reliability is associated with the *variance* of $\epsilon$.

# $k \neq 1$ do not bias estimation of the mean

Population parameter of interest is

$$\mu \equiv \sum_i x_i / N$$

$N$ is the size of the population. and a variables $X = x + \epsilon$ contaminated with error. Consider a *random sample of n units* from $\mathcal{U}$ We want to assess the estimation of $\mu$ using

$$\bar{X} \equiv \sum_i X_i / n$$

Note

$$\bar{X} = \bar{x} + \bar{\epsilon}$$

Cero mean, finite variance and independence of the$\epsilon_i$s, and the *Law of Large Numbers* ensures $\bar{\epsilon} \to 0$; thus, $\bar{x} \to \mu$ and $\bar{X}$ is consistent for $\mu$ The normality assumption on $\epsilon_i$s is not required, only independence across of the $\epsilon_i$s.

Consistency of $\bar{X}$ as estimator of $\mu$ holds even under measurement error.

## $k \neq 1$ do not bias estimation of a covariance; it bias estimation of the variance

$$X = x + \epsilon_X$$

$$Y = y + \epsilon_Y$$

Reliability of $X$ is $k_X \equiv \sigma_{xx}/\sigma_{XX}$. Clearly estimation of the variance of $X$ (and of $Y$) is affected by measurement error :

$$\sigma_{XX} = \sigma_{xx} + \sigma_{\epsilon_X \epsilon_X} = \sigma_{xx} + (1 - k_X)\sigma_{XX} \neq \sigma_{xx}$$

Estimation of covariance is free of $k$ when the covariance of the errors is zero:

$$\sigma_{XY} = \sigma_{xy} + \sigma_{\epsilon_X \epsilon_Y} = \sigma_{xy}$$

Measurement error does not affect the consistency of $s_{XY} = \sum_i (X_i - \bar{X})(Y_i - \bar{Y})/n$ as estimator of $\sigma_{x,y}$. This result is robust to non-normality.

# $k \neq 1$ do bias estimation of the correlation

From the definition of a correlation

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

and the results $\sigma_X = \sigma_x / \sqrt{k_X}$, $\sigma_Y = \sigma_y / \sqrt{k_Y}$ We obtain

$$\rho_{XY} = \rho_{xy} \times \sqrt{k_X k_Y}$$

Attenuation effect on the correlation when either (or both) $k_X$ and $k_Y$ are strictly smaller than 1 (assuming errors $\epsilon_X$ and $\epsilon_Y$ are uncorrelated).

# $k \neq 1$ do bias estimation of the regression coefficient

The regression coefficient

$$\beta_{X \to Y} \equiv \frac{\sigma_{XY}}{\sigma_X^2} = \frac{\sigma_{xy}}{\sigma_x^2 / k_X} = k_X \times \frac{\sigma_{xy}}{\sigma_x^2} = k_X \times \beta_{x \to y}$$

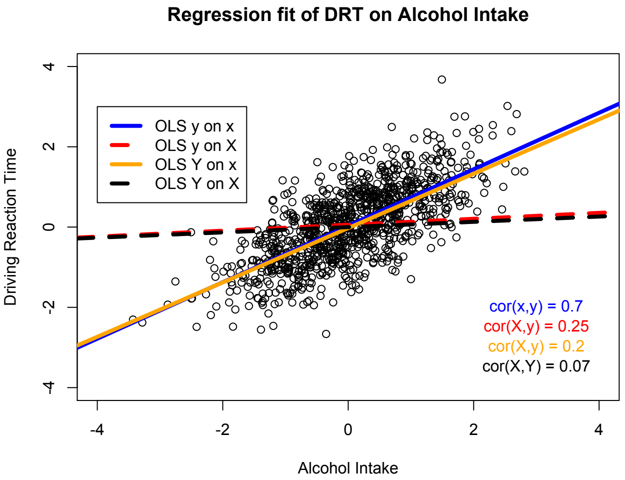
Thus, the (true) regression coefficient $\beta_{x \to y}$ is attenuatetd by the measurement error on $X$. It can be dis-attenuated by

$$\beta_{x \to y} = \beta_{X \to Y} / k_X$$

No attenuation effect when measurment error is on $Y$ .

# Illustration: regression of DRT on AI

Reliabilities: $k_Y = .1; \quad k_X = .1$



# Measurement-error in regression when reliabilities are known

## An example: Fuller's (1987) Farms Operators

# An example: Fuller's (1987) Farms Operators

## *Regression with error-in-variables*

Ex. 3.1.2 of Fuller (1987)

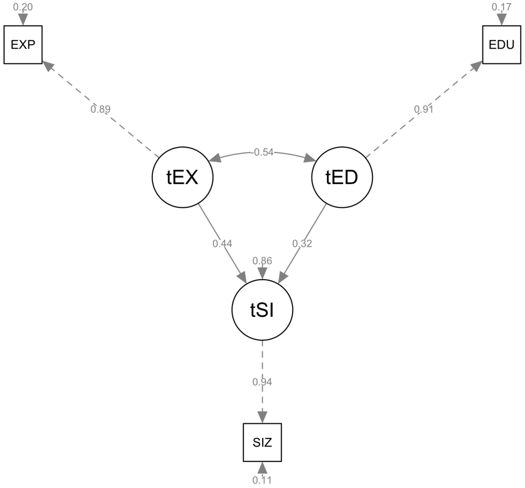
Data from a sample of Iowa farm operators

Y = ln (farm size)
X1 = ln ( # years experience )
X2 = ln (# years education )

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u$$
$$Y = y + e_1$$
$$X_1 = x_1 + e_2$$
$$X_2 = x_2 + e_3$$

(to protect confidentiality,
random error was added to each variable)

# Farms Operators Example: Fitted model (standardized)



# SEM sintaxis of Farm Operators Example

```
Covariances:"SIZE", "EXPERIENCE", "EDUCAT"

       .9148
       .2129 1.006
       .0714 -.449 1.039

reliabilities:  .891,.800,.826
variances of errors: 0.0997132, 0.2012000, 0.1807860
=c(.9148,1.006,1.039)*(1- c(.891,.800,.826))

model1:
latent:  tSIZE=~SIZE;tEXPER=~EXPERIENCE;tEDUC=~EDUCAT
regress: tSIZE  ~   tEXPER + tEDUC
variances of errors:        SIZE        ~~ 0.0997*SIZE
                            EXPERIENCE  ~~ 0.2012*EXPERIENCE
                            EDUCAT      ~~ 0.1807*EDUCAT
```

# Parameter Estimates

| | lhs | op | rhs | est | se | z | pvalue | std.all |
|---|---|---|---|---|---|---|---|---|
| 4 | tSIZE | ~ | tEXPER | 0.440 | 0.112 | 3.923 | 0.000 | 0.437 |
| 5 | tSIZE | ~ | tEDUC | 0.314 | 0.107 | 2.936 | 0.003 | 0.322 |
| 6 | tEXPER | ~~ | tEXPER | 0.799 | 0.107 | 7.494 | 0.000 | 1.000 |
| 7 | tEDUC | ~~ | tEDUC | 0.852 | 0.110 | 7.740 | 0.000 | 1.000 |
| 8 | tEXPER | ~~ | tEDUC | -0.446 | 0.084 | -5.335 | 0.000 | -0.541 |
| 9 | SIZE | ~~ | SIZE | 0.100 | 0.000 | NA | NA | 0.110 |
| 10 | EXPERIENCE | ~~ | EXPERIENCE | 0.201 | 0.000 | NA | NA | 0.201 |
| 11 | EDUCAT | ~~ | EDUCAT | 0.181 | 0.000 | NA | NA | 0.175 |

# Accounting for uncertainty on the estimate of error variances

- Erik Meijer, Edward Oczkowski, Tom Wansbeek (2021) *How measurement error affects inference in linear regression*. **Empirical Economics, 60**, pp. 131–155
- Daniel Oberski and Albert Satorra (2013) *Measurement Error Models With Uncertainty About the Error Variance*, **Structural Equation Modeling, 20**, pp. 409–428

When reliability is not known: bring in multiequation models

# MIMIC model

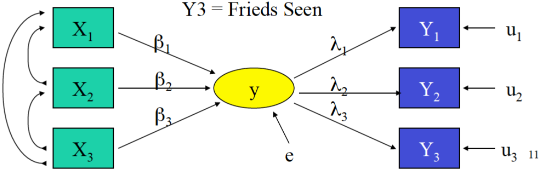
*Mimic model*

Joreskog & Goldberger, JASA (1979)

y =social participation

X1 = Income
X2 = Occupation
X3 = Education

Y1= Church attendance
Y2 = Membership
Y3 = Frieds Seen



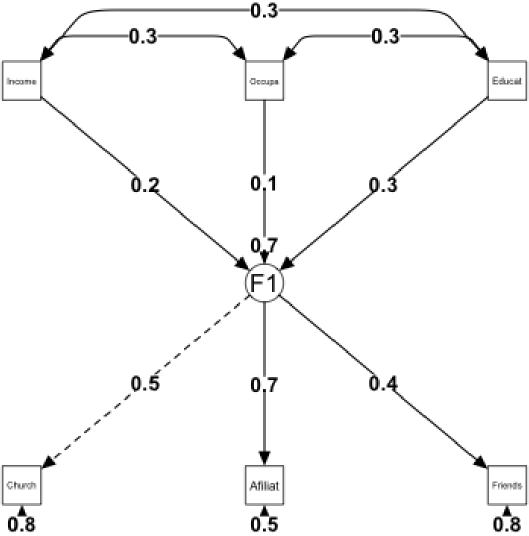
# MIMIC model: ML estimates

ML estimates:

$$
\begin{aligned}
y &= \underset{(.066)}{.269}\, X_1 + \underset{(.065)}{.114}\, X_2 + \underset{(.070)}{.386}\, X_3 + u \\
Y_1 &= \underset{(.046)}{.402}\, y + e_1 \\
Y_2 &= \underset{(.060)}{.634}\, y + e_2 \\
Y_3 &= \underset{(.046)}{.346}\, y + e_3
\end{aligned}
$$

6 overidentifying restrictions. The corresponding chi2 is 12.36 with "P-VALUE" 0.052.

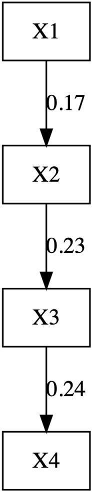
# Path diagram of fitted model



Longitudinal data: accounting for measurement error multiple equations. No information on reliabilities. Testing a Markovian assumption.

# Measurement error in longitudinal research

# Longitudinal data: no account of mesurement error



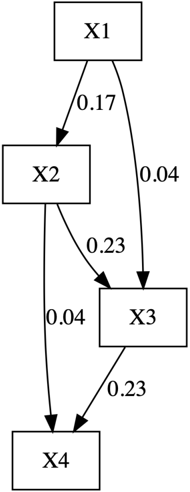

Markovian assumption is rejected. Model fit is chisq =102.351, df =3, p-value = 0.000

# Autorregressive parameter: pop. val is .4

We use a large sample size $n$ to ensure the visibility of bias.

```
Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  X2 ~
    X1                0.174    0.005   34.306    0.000
  X3 ~
    X2                0.231    0.008   28.520    0.000
  X4 ~
    X3                0.243    0.013   18.835    0.000


 Number of observations                        200000

Model Test User Model:
                                               Standard      Robust
  Test Statistic                                550.335     102.351
  Degrees of freedom                                  3           3
  P-value (Chi-square)                            0.000       0.000
  Scaling correction factor                                   5.377
       Satorra-Bentler correction

 modindices (LM tests)
   lhs op rhs      mi    epc
    X4  ~  X2 327.176  0.045
    X2  ~  X4 282.610  0.035
    X2  ~  X3 210.934 -0.203
    X3  ~  X1 210.934  0.039
    X1  ~  X3 210.934  0.028
    X4  ~  X1  35.655  0.016
...
```

# Londitudinal data: no account of mesurement error (model modified to fit)



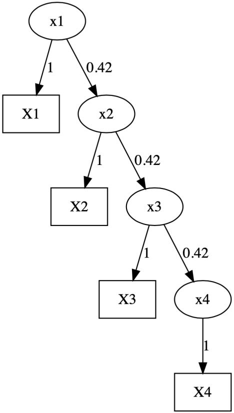

All regression parameters are significant. Deviation from a Markovian

# No account for measurement error, model modified to fit

```
Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  X2 ~
    X1                0.174    0.005   34.306    0.000
  X3 ~
    X2                0.225    0.008   28.296    0.000
  X4 ~
    X3                0.234    0.012   18.717    0.000
    X2                0.045    0.007    6.005    0.000
  X3 ~
    X1                0.039    0.005    7.857    0.000
```

```
Model Test User Model:
                                              Standard      Robust
  Test Statistic                                11.847       3.338
  Degrees of freedom                                 1           1
  P-value (Chi-square)                           0.001       0.068
  Scaling correction factor                                  3.549
       Satorra-Bentler correction
```

# Londitudinal data with account for measurement error (using the simplex model)

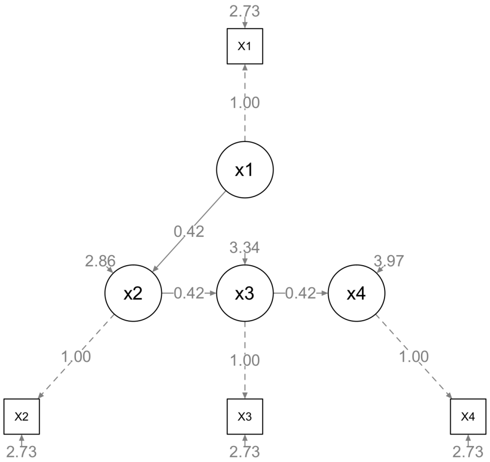



Markovian model (Simplex) is accepted. (SB-scaled) Chi2 = 1.749, df=4, p-value =0.782.

# Autorregressive parameter: pop. val is .4

```
Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  x2 ~
    x1          (a)    0.422    0.027   15.835    0.000
  x3 ~
    x2          (a)    0.422    0.027   15.835    0.000
  x4 ~
    x3          (a)    0.422    0.027   15.835    0.000


Number of observations                           200000

Model Test User Model:
                                              Standard      Robust
  Test Statistic                                15.626       1.749
  Degrees of freedom                                 4           4
  P-value (Chi-square)                           0.004       0.782
  Scaling correction factor                                  8.932
       Satorra-Bentler correction
```

# Londitudinal data: Simplex model (do account for measurement error)




See Section 6.2(Ability scores over time) of Dunn, Everitt and Pickles (1993)

# Measurement error + method effect: MTMM

# Multi-Trait and Multi-Method (MTMM)

- The multitrait-multimethod (MTMM) design (Campbell and Fiske 1959) is probably the best known social science procedure for uncovering systematic measurement error (Bollen and Paxton, 1998)
- Each test or task employed for measurement purposes is a trait method unit a union of a particular trait content with measurement procedures no specific to that content (Campbell and Fiske 1959)

# MTMM: Reliability of $Y_{ij}$

Convergent and discriminant validity of different methods was first assessed in a systematic way by the MTMM design (Campbell and Fiske, 1959), a design where a response or measured variable $Y_{ij}$ associated to trait $i$ ($T_i$) and method $j$ ($M_j$) decomposes as

$$Y_{ij} = t_{ij} T_i + m_{ij} M_j + u_{ij} \quad \text{trait i, method j}$$

with $u_{ij}$ being the random error associated to lack of reliability when measuring $S_{ij} = t_{ij} T_i + m_{ij} M_j$, the part of the response that would be stable across identical repetitions of the measurement process of trait $i$ and method $j$ (Saris and Andrews, 1991, call this the *true score* ). In accordance to this,

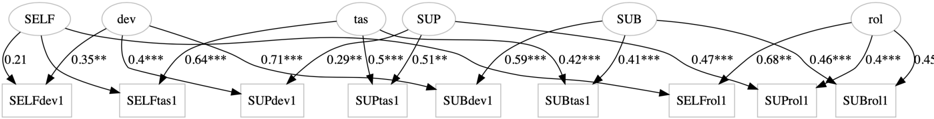
$$k_{ij} \equiv \text{Reliability}(Y_{ij}) = \frac{\text{var}(S_{ij})}{\text{var}(Y_{ij})} = \frac{t_{ij}^2 \text{var}(T_i) + \text{var}(M_j)}{\text{var}(Y_{ij})}$$

We assumed $m_{ij} = 1$ across traits and methods. Its square root, $k_{ij}^{1/2}$, is called the *Validity coefficient*.

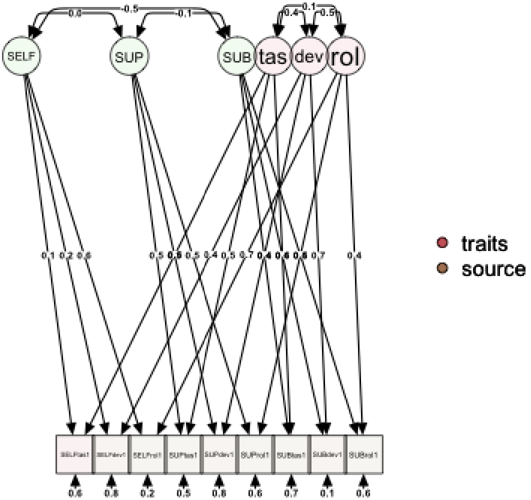
# MTMM with empirical data

```
Here is the variables being used in the empirical example

trait factors are:
tas:The leader monitors whether progress at work is sufficient
dev:The leader is on the lookout for efforts that can be prais
rol:The leader shows respect for others

Assessed by SELF, SUP, SUB.

Number of observations                                 239

Model Test User Model:
                                                    Standard
  Test Statistic                                       8.161
  Degrees of freedom                                      12
  P-value (Chi-square)                                 0.772
```

# Path diagram A: MTMM fit empirical data (std solution)

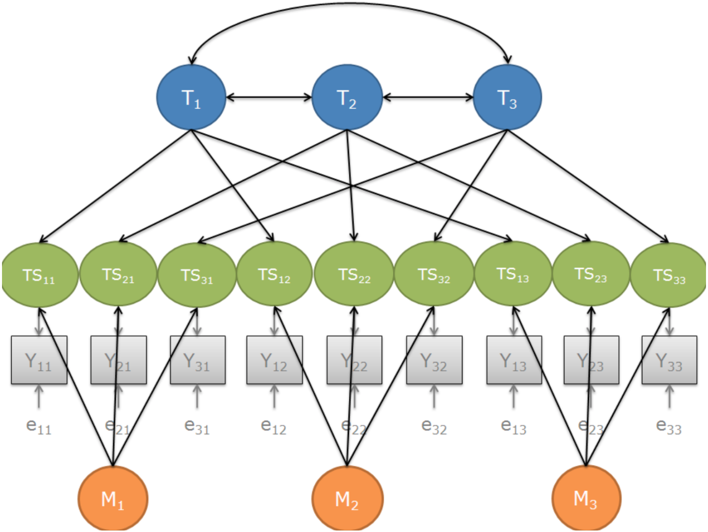


# Path diagram B: MTMM fit empirical data (std solution)

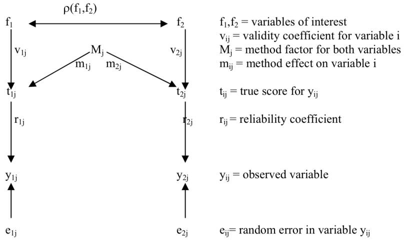


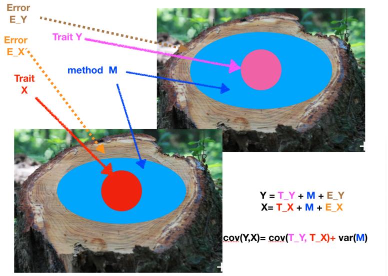
# MTMM: True Score Model

Saris and Andrews (1991)



# Common variation due to method: source of endogeneity (confounding) in the practice of $Y = \alpha + \beta X + \epsilon$




Besides measurement error there is the method effects: a common component of variation due to sharing the same method when measuring $X$ and $Y$.

# Early work on shock-error models and Frisch's (1934) idea of IVs

# Regression with measurement error

"System" of equations:

$$Y_t = y_t + \epsilon_{Yt}; \quad X_t = x_t + \epsilon_{Xt}; \quad y_t = \beta' x_t + d_t$$

In matrix form:

$$\eta_t = B\eta_t + \zeta_t$$

$$z_t = I_q \eta_t + \epsilon_t$$

where: $z_t \equiv (Y_t, X_t')' \, (1 \times q)$; $\eta_t \equiv \begin{pmatrix} y_t \\ x_t \end{pmatrix}$; $B = \begin{pmatrix} 0 & \beta' \\ 0 & I_q \end{pmatrix}$; $\zeta_t \equiv \begin{pmatrix} d_t \\ x_t \end{pmatrix}$; $\epsilon_t \equiv \begin{pmatrix} \epsilon_{Yt} \\ \epsilon_{Xt} \end{pmatrix}$.

More compactly:

$$z = \xi_t + \epsilon_t; \quad \alpha' \xi_t = d_t$$

where

$$\alpha \equiv (1, -\beta')'; \quad \xi_t \equiv (I_{q+1} - B)^{-1} \zeta_t.$$

# Frisch (1934), Reiersøl (1941), Haavelmo (1943)

Ragnar Frisch's (1934) set-up[2]

$$z = \xi_t + \epsilon_t$$

The $\xi_t$ and $\epsilon_t$ are stationary of zero mean with $\mathcal{E}\,\xi_t \epsilon'_{t'} = 0$ for all $t$ and $t'$. Frisch is concerned with the following linear relationships in the systematic part

$$\alpha' \xi_t = 0$$

In our case, $z \equiv (Y_t, X'_t)'$, $\xi_t \equiv (y_t, x'_t)'$, and $\epsilon_t \equiv (\epsilon_Y, \epsilon'_X)'$ are *serially independent measurement error*,

$$\alpha' \xi_t = d_t$$

where $\alpha \equiv (1, -\beta)'$. Let $M_{\xi\xi} \equiv \mathcal{E}\,\xi_t \xi'_t$, and $M_{zz} \equiv \mathcal{E}\,z_t z'_t$,then

$$M_{zz} = M_{\xi\xi} + \Omega$$

where $\Omega \equiv \mathcal{E}\,(\epsilon_t \epsilon'_t)$. This is precisely the simpler set-up of the "Frisch model" (see Aldrich, 1993 for more details on the history of " Frisch model" and the birth of the IV approach).

[2] Original notation is slightly modified

# Ragnar Frisch, 1934: $\beta$ is not identified from second-order moments

Frisch (and Gini) noted that (see Aldrich, 1993 ) in the case of simple regression

$$b_{X \to Y} = \frac{m12}{m22} \leq \beta \leq \frac{m11}{m12} = 1/b_{Y \to X}$$

- Only bounds, not consistent estimator, of $\beta$. $\beta$ is not identified from second moments of the observable variables.
- Jerzy Neyman (1937) conjectured that for a non-normal regressor consistent estimation in the classical errors-in-variables model could be achieved by using third-order moments.
- PAL, M. (1980) provides a consistent estimator of $\beta$ based on third order moments. See below

# Estimation of $\beta$ using higher-order moments

Jennrich and Satorra (2015): Let $X$ and $Y$ be two unobserved random variables with

$$Y = \alpha + \beta X$$

Assume $X$ and $Y$ can only be observed with error as

$$x = X + d \quad \text{and} \quad y = Y + e$$

where $d$ and $e$ have mean zero, are independent among themselves and independent also of $X$ and $Y$. Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a sample from the joint distribution of $(x, y)$. It is shown by PAL (1980, p.352) that if the third central moment of $X$ is nonzero, then

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})^2}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})}$$

is a consistent estimator of $\beta$. We will show how to use IJK methods to show not only this, but also to show that $\hat{\beta}$ is asymptotically normal and to provide a standard error for $\hat{\beta}$. For each data set, we computed $\hat{\beta}$ and the IJK estimate of the asymptotic variance of $\hat{\beta}$. Let $\text{avar}^{IJK}(\hat{\beta})$ be the IJK estimate of the $\text{avar}(\hat{\beta})$. Then

# IV idea: Frisch (1934)

To estimate the $\beta$ in the regression $\alpha'\xi_t = 0$, where: $\alpha \equiv (1, -\beta')'$, $\xi_t \equiv (y_t, x_t')'$, $z_t = \xi_t + \epsilon_t$, $\epsilon_t = (\epsilon_{Yt}, \epsilon_{Xt}')'$, and $z_t$ is the only observable, Frisch assumes: (a) for any $t, t'$, $\mathcal{E}\,\xi_t'\epsilon_{t'} = 0$; and, (b) for $t \neq t'$, $\mathcal{E}\,\epsilon_t'\epsilon_{t'} = 0$, and uses $z_{t-s}$ (the $s$-lagged variable $z_t$, $s$ positive or negative) as *instrument* to identify $\beta$ from the data. The reasoning is:

$$M_{z_t z_{t-s}} \equiv \mathcal{E}\left(z_t z_{t-s}'\right) = \mathcal{E}\left(\xi_t \xi_{t-s}'\right) + \mathcal{E}\left(\epsilon_t \epsilon_{t-s}'\right) = \mathcal{E}\left(\xi_t \xi_{t-s}'\right) \equiv M_{\xi_t \xi_{t-s}}$$

since $\mathcal{E}\,\xi_{t-s}'\epsilon_t = 0$. This implies

$$\alpha' \begin{pmatrix} m_{Y_t Y_{t-s}} & m_{Y_t X_{t-s}} \\ m_{X_t Y_{t-s}} & m_{X_t X_{t-s}} \end{pmatrix} = \alpha' \begin{pmatrix} m_{y_t y_{t-s}} & m_{y_t x_{t-s}} \\ m_{x_t y_{t-s}} & m_{x_t x_{t-s}} \end{pmatrix} = \begin{pmatrix} 0 & 0 \end{pmatrix}$$

since $\alpha' M_{\xi_t \xi_{t-s}} = 0$, $\alpha' M_{z_t, X_{t-s}} = m_{Y_t X_{t-s}} - \beta' m_{X_t X_{t-s}} = 0$, Frisch then deriving the consistent estimator of $\beta$:

$$b_{IV} \equiv m_{Y_t X_{t-s}} m_{X_t X_{t-s}}^{-1}$$

Frisch uses as the IV of $X_t$ the lagged $X_{t-s}$ ($s$ positive or negatives).

# Trygve Haavelmo (1943): shocks in the equations

Equations for consumption $u$ and investment $v$ (or their sum $r$) [thinking structural, causal modeling ]:

$$u_t = \alpha r_t + \beta + x_t$$

$$v_t = \kappa(u_t - u_{t-1}) + y_t$$

$$r_t = u_t + v_t$$

with observable variables $(u_t, r_t)$, $t = 1, 2, \dots N$. Haavelmo refers to the joint probability law of exogeneous random disturbance terms

$$f(x_1, x_2, \dots, x_N, y_1, y_2, \dots, y_N)$$

that determine—up to certain number of unnown parameters (such as the $\kappa$, $\alpha$, $\sigma_x$ or $\sigma_y$ above)— the distribution of the observable variables

$$(u_1, r_1, u_2, r_2, \dots, u_n, r_N)$$

Disturbances (shocks) in equations vs. measurement errors. The footnote in the paper of Koopmans and Hood (1953).

# Errors of observations vs. disturbances, Koopmans and Hood (1953), p. 117:

*That errors of observations are disregarded in this chapter does not imply an a priori judgment that such errors are less important, in their effects on the choice of estimates and on the quality of these estimates, than diturbances in economic behaviour.*[3] *This most be regarded as an empirical question, to be setttled by methods of inference based on models recognizing errors of observation as well as disturbances in behaviour. The emphasis on disturbances in this and other chapters of this volume must be regarded rather as matter of tactics. "Shock-error models" are complicated.*

"Shock-error models" + latent variables cleared up by Karl G. Jöreskog's LISREL approach .

[3]It might be thought that with gradual improvement in the methods of data collection, errors of observation would after a lapse of time be less important than the random elements intrinsic to economic behaviour. However, as Reiersøl as pointed out to one of the authors, as observation improve in accuracy and coverage, it will be possible to introduce more explanatory variables in each equation, thus reducing the variance of "unexplained" disturbances in behavior.

# Sewall Wright's 1934, method of path coefficients

A paper not to miss on the history of the development of structural equation models in economics and social sciences is the *Econometrica* paper Goldberger (1972). It gives an account of the people who first mapped the territory of SEM. It points to the prominent role of Sewall Wright with his method of path analysis. It points also the relevance of the emerging work of Karl G. Jöreskog (in intensive progress at the time, 1971, of Goldberger's Schultz Lecture). Fifty years later, now, it is more than evident the breakthrough contribution of Jöreskog in brushing up the landscape of SEM with his creation, the tool kit LISREL.

# The Jöreskog's SEM approach to measurement error and unobservables

# SEM (LISREL)

Jöreskog, K. G. (1970, ...) develop ML estimation and testing for a general shock-error-latent variable model + producing (with Dag Sörbom) the software LISREL to serve practitioners.

An exact relation $\eta = B\eta$ is contaminated by shocks

$$z = \Lambda\eta + \epsilon$$

$$\eta = B\eta + \zeta$$

with $\Psi := E\epsilon\epsilon'$ and $\Phi := E\zeta\zeta'$. Denote $\xi \equiv \Lambda(I - B)^{-1}\zeta$; then, we can write:

$$z = \Lambda(I - B)^{-1}\zeta + \epsilon = \xi + \epsilon$$

The matrices $B$, $\Lambda$, $\Psi$ and $\Phi$ are functions of $\theta$, the fundamental parameters of the model. The moment structure for the observable vector $z$ is

$$\Sigma_{zz} = \Lambda(I - B)^{-1}\Phi(\Lambda(I - B)^{-1})^T + \Psi = \Sigma_{zz}(\theta)$$

where $\theta$ is the vector of free parameters of the coefficient matrices. [4]

[4] Proprietary software: LISREL, EQS, Mplus, CALIS, `sem` of Stata, AMOS, .... Free software: `sem`, `lavaan` , `OpenMx`, all in `R`, ....

# K. G. Jöreskog's LISREL: the SEM approach

LISREL (SEM):

1. (a unifying) general "shocks-errors-latent variables" variable model. It encompasses regression, simultaneous equations, factor analysis, and combinations of the three.
2. ML estimation and testing of the general model, multiple group, robust se and test statistics (applicable to any subfamily of models)
3. Software for routinary practitioners use (not necessarily statisticians/econometricians) Nowadays: LISREL, EQS, MPlus, sem of Stata, sem and lavaan of free software R, . . . . LISREL was pioneering in the 70s.

A unifying tool for comparative empirical research. As in classical OLS regression, a variety of SEM software producing identical numerical results on a variety of models.

# Endogeneity in regression with errors in variables

# Measurement error and endogeneity in regression

Consider the (SEM) system:

$$Y = \alpha + \beta x + \epsilon; \quad X = x + u \tag{2}$$

where $\epsilon$ and $u$ are centered variables, and $x$, $u$ and $\epsilon$ are mutually uncorrelated. System (2) do imply the regression equation

$$Y = \alpha + \beta X + \delta \tag{3}$$

where $\delta$ has mean zero, but is correlated with the regressor $X$.

Indeed: $Y = \alpha + \beta(X - u) + \epsilon = \alpha + \beta X + \delta; \quad \delta = \epsilon - \beta u$. Thus $\mathrm{cov}(\delta, X) = \mathrm{cov}(\epsilon - \beta u, x + u) = \beta\, \mathrm{var}(u) \neq 0$. The covariance is zero only when $\beta$ or $\mathrm{var}(u)$ (or both) are zero. It is easy to see that the asymptotic limit of the OLS regression coefficient $\sigma_{XY}/\sigma_X^2 = \beta\, k_X$; $k_X = \sigma_x^2/\sigma_X^2$. In (3), due to the nature of $\delta$ generally, $\mathcal{E}(Y \mid X) \neq \alpha + \beta X$; or, equivalently, $\mathcal{E}(\delta \mid X) \neq 0$.

Since $\mathrm{cov}(\delta, X) \neq 0$ it is said the regression (3) suffers of endogeneity of $X$; so OLS will usually be an asymptotically biased estimator of $\beta$. In fact, OLS will consistently estimate $\beta \times k$ where $k$ is the reliability of $X$ as a measure of the true $x$. So, OLS regression attenuates the true regression coefficient, $\beta$, when $k_X \leq 1$.

When $X = x + u$, $X$ is endogeneous in the regression

$$Y = \alpha + \beta X + \delta$$

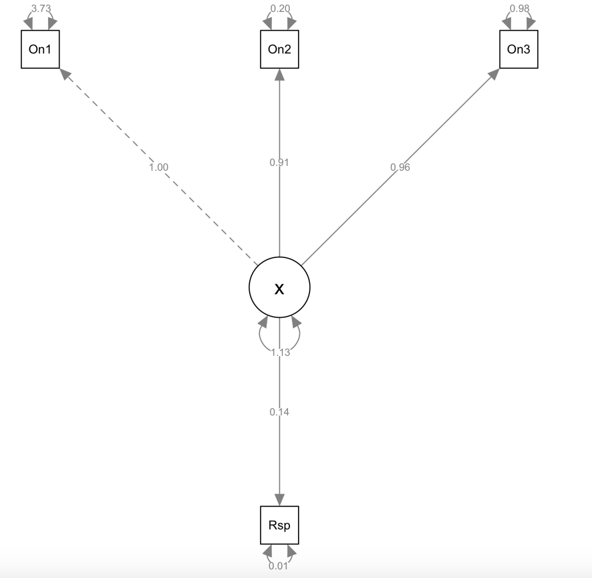
Let $Z$ correlate with $x$ but is uncorrelated with $u$ and the disturbance term $\epsilon$ of (2) (so, it is uncorrelated with the disturbance term $\delta$ of (3). This implies that $Z$ is – in the Frisch-Reiersøl-Haavelmo and Sargan (1958) sense– an instrumental variable (IV) for $X$ in the regression (3). Note that $Z$ could be simply a repeated measure of $x$.

# Regression with multiple indicators of x

We consider data as

```
  head(data)
  Resp Ointake1 Ointake2 Ointake3
1 9.90    11.25    10.12     9.21
2 9.99    10.07    10.64    12.48
3 9.96     5.53    11.08     9.72
4 9.93    12.49     9.34    10.45
5 9.80     8.23     9.12     8.20
6 9.75    11.77     9.10    10.58
....

## SEM
model <- "x =~   Ointake1+ Ointake2 + Ointake3;   Resp ~ x; x~
fit <- sem(model, estimator="MLM",  data = data, auto.var = TF
semPaths(fit, "est",   weighted = FALSE)
```

# SEM fit of a regression with latent factor x



# SEM fit of a regression with latent factor x

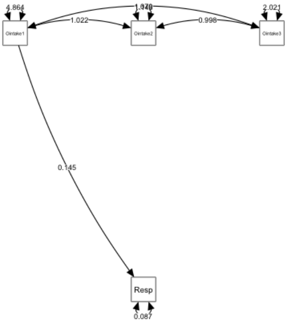
```
parameterestimates(fit, ci=FALSE, standardized=TRUE)[1:5,c(1:7,9)]
```

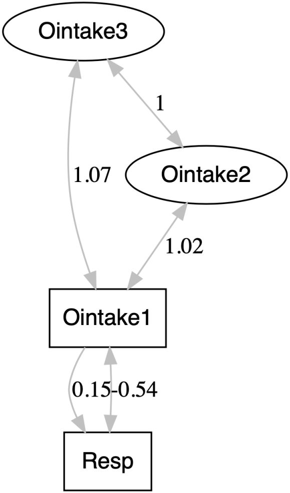
```
   lhs op      rhs   est    se      z pvalue std.all
1    x =~ Ointake1 1.000 0.000     NA     NA   0.482
2    x =~ Ointake2 0.914 0.060 15.146      0   0.907
3    x =~ Ointake3 0.962 0.069 13.912      0   0.719
4 Resp  ~        x 0.144 0.010 15.000      0   0.842
5    x ~~        x 1.131 0.151  7.475      0   1.000
```

```
Model Test User Model:
                                              Standard      Robust
  Test Statistic                                 3.284       3.179
  Degrees of freedom                                 2           2
  P-value (Chi-square)                           0.194       0.204
  Scaling correction factor                                  1.033
       Satorra-Bentler correction
```

# IV -SEM

# IV-SEM Fit of regression with two IVs



# Sintaxis and fit of IV-SEM regression with two IVs

```
## IV_SEM

    model <- "Resp ~  Ointake1;   Resp ~~ Ointake1;
    Ointake1 ~~ Ointake2+Ointake3;
    Ointake2 ~~  Ointake3"
fit <- sem(model, data = data, estimator="MLM",  auto.var = TRUE)

lavaanPlot(name = "plot1", model = fit, coefs = TRUE, stand = FALSE, covs = TRUE,
    digits = 2, stars = FALSE, edge_options = list(color = "grey"), node_options = list(shape = "box",
        fontname = "Helvetica"))

 parameterestimates(fit, ci=FALSE, standardized=TRUE)[1:5,c(1:7,9)]
       lhs op      rhs    est    se       z pvalue std.all
1     Resp  ~ Ointake1  0.145 0.010  14.736      0   1.765
2     Resp ~~ Ointake1 -0.540 0.047 -11.390      0  -0.829
3 Ointake1 ~~ Ointake2  1.022 0.084  12.122      0   0.432
4 Ointake1 ~~ Ointake3  1.070 0.095  11.293      0   0.341
5 Ointake2 ~~ Ointake3  0.998 0.059  16.784      0   0.655


Model Test User Model:
                                              Standard      Robust
  Test Statistic                                 2.339       2.546
  Degrees of freedom                                 1           1
  P-value (Chi-square)                           0.126       0.111
  Scaling correction factor                                  0.919
       Satorra-Bentler correction
```

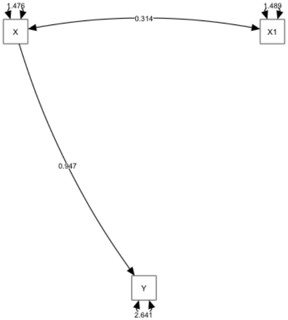
# IV approach using ivpack

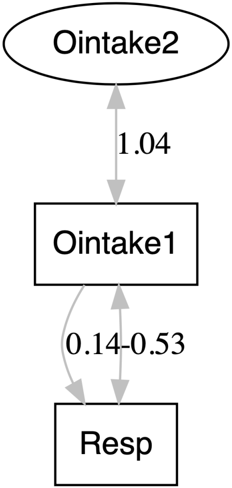
```
library(ivpack)
iv1 <- ivreg(Resp ~ Ointake1 | Ointake2 + Ointake3 , data = data)
summary(iv1, vcov = sandwich, diagnostics = TRUE)


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.550189   0.098566   86.75   <2e-16 ***
Ointake1    0.143836   0.009783   14.70   <2e-16 ***

Diagnostic tests:
                 df1 df2 statistic p-value
Weak instruments   2 859   113.678  <2e-16 ***
Wu-Hausman         1 859   887.597  <2e-16 ***
Sargan             1  NA     2.358   0.125
---
```

# Sargan's test and the test for Weak Instruments

In the case of 2 IV, a chi-square difference test with df=2 is a test of Weak instruments; and Sargan's test (testing if there is a correlation of disturbance term with X) has df =1. In our data, both tests reject the null. Both tests are asymptotically equal to the SEM chi-square (model or difference) tests. Sargan's test is equivalent to the chi-square goodness of fit for the fitted IV-SEM model (in SEM, we robustify this test to non-normality)

# Path diagrams of IV-SEM



# IV-SEM regression with one IV

```
 IV-SEM

    model <- "Resp ~  Ointake1;   Resp ~~ Ointake1;
    Ointake1 ~~ Ointake2 "
fit <- sem(model, data = data, estimator="MLM",  auto.var = TRUE)


  parameterestimates(fit, ci=FALSE, standardized=TRUE)[1:5,c(1:7,9)]
       lhs op      rhs    est    se       z pvalue std.all
1     Resp  ~ Ointake1  0.143 0.010  14.578      0   1.731
2     Resp ~~ Ointake1 -0.526 0.047 -11.134      0  -0.822
3 Ointake1 ~~ Ointake2  1.040 0.085  12.165      0   0.440
4     Resp ~~     Resp  0.084 0.011   7.596      0   2.550
5 Ointake1 ~~ Ointake1  4.864 0.238  20.474      0   1.000

Model Test User Model:
                                              Standard      Robust
  Test Statistic                                 0.000       0.000
  Degrees of freedom                                 0           0
```

# ivpack with one IV

```
library(ivpack)
iv1 <- ivreg(Resp ~ Ointake1 | Ointake2  , data = data)
summary(iv1, vcov = sandwich, diagnostics = TRUE)


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.562503   0.098581   86.86   <2e-16 ***
Ointake1    0.142608   0.009783   14.58   <2e-16 ***

Diagnostic tests:
                 df1 df2 statistic p-value
Weak instruments   1 860     220.4  <2e-16 ***
Wu-Hausman         1 859     841.6  <2e-16 ***
Sargan             0  NA        NA      NA
---
```

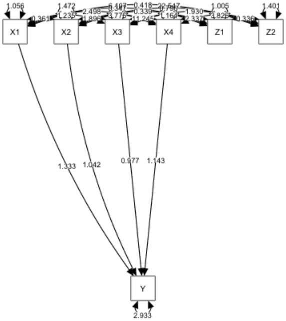
# Estimates in multiple regression, comparison of OLS, ivreg, IV-SEM

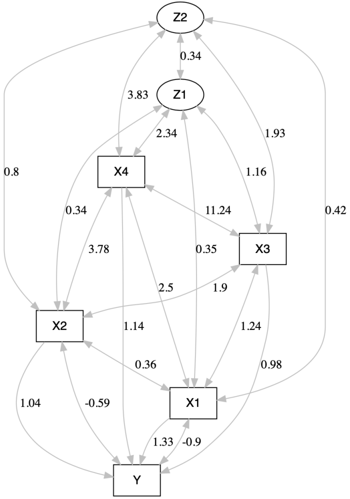
Linear regression model

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

Regressors $X_1$ and $X_2$ suffer from endogeneity due to measurement error. Two indicators $Z_1$ and $Z_2$ of respectively the true value of $X_1$ and $X_2$ serve as IV. Simulated data with $\beta_j = 1, j = 1, 2, 3, 4$ and $n = 2000$

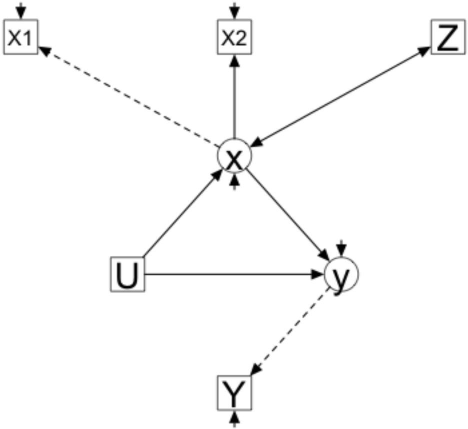
| | *OLS* | | *ivreg* | | *IV-SEM* | | *IV-SEMrob* | |
|---|---|---|---|---|---|---|---|---|
| | Est | se | Est | se | Est | se | Est | se |
| X1 | 0.120 | 0.030 | 1.333 | 0.397 | 1.333 | 0.397 | 1.333 | 0.392 |
| X2 | 0.261 | 0.029 | 1.042 | 0.218 | 1.042 | 0.218 | 1.042 | 0.217 |
| X3 | 0.971 | 0.037 | 0.977 | 0.055 | 0.977 | 0.055 | 0.977 | 0.055 |
| X4 | 1.411 | 0.020 | 1.143 | 0.072 | 1.143 | 0.072 | 1.143 | 0.074 |


X1
X2
X3
X4
Z1
Z2
Y
1.056
1.472
1.005
1.401
1.333
1.042
0.977
1.143
2.933

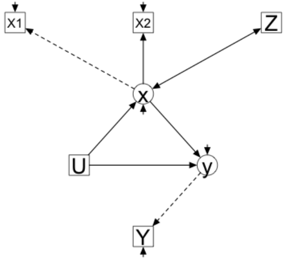

Z2
0.34
Z1
3.83
2.34
1.93
0.8
X4
1.16
0.34
11.24
0.42
X3
3.78
0.35
2.5
1.9
X2
1.14
1.24
0.36
0.98
1.04
-0.59
X1
1.33
-0.9
Y

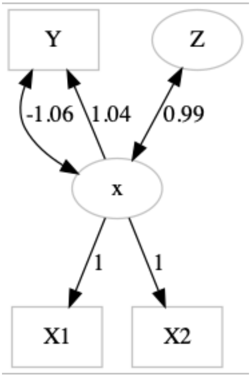
## Sources of endogeneity (shocks vs. errors) and latent variables: the SEM approach

# Sources of endogeneity: shocks vs. errors: e.g., the true value target is $\beta_{x\to y} = 1.0$

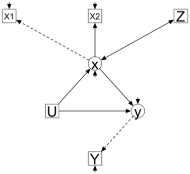


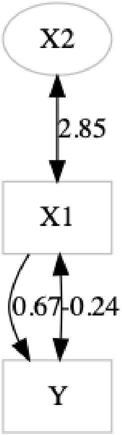
# Estimation of $\beta_{X \to Y}$ using a latent variable model (M0): 1.039(0.031)



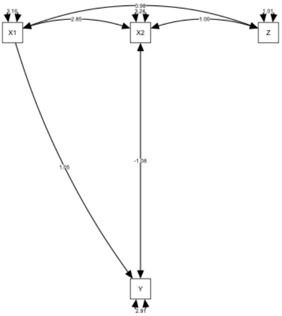



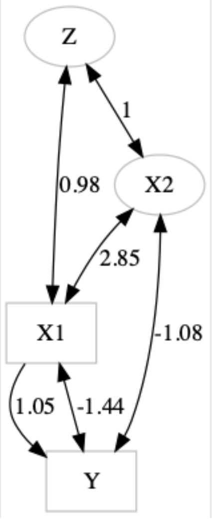
# Estimation of $\beta_{X \to Y}$ with X2 the IV of X1 (M1): 0.673(0.017)



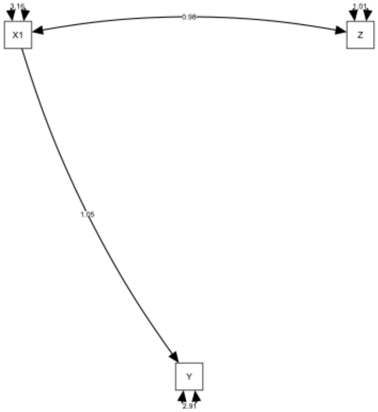
# Estimation of $\beta_{X \to Y}$ with X2 and Z the IVs of X1 (M2): 1.051(0.032)



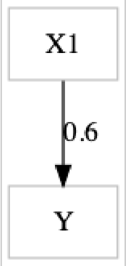

 71 / 75

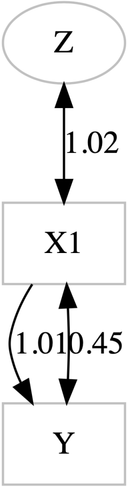
# Estimation of $\beta_{X \to Y}$ with Z the IV of X1 (M3): 1.051(0.032)



# Simple regression vs regression with Z as IV: true $\beta_{X \to Y} = 1$

Left-side graph: Simple regression estimate is: 0.596 (0.015)



# Discussion (cont.)

- Karl G. Jöreskog's LISREL (early 70s): (a) a general model that encompass shocks, errors, and latent variables, (b) ML estimation and testing methods (c) A computer package for practitioners. Now: LISREL, EQS, Mplus, AMOS, sem of Stata, . . . ; free software: sem, lavaan, . . .

# Discussion (cont.)

- Karl G. Jöreskog's LISREL (early 70s): (a) a general model that encompass shocks, errors, and latent variables, (b) ML estimation and testing methods (c) A computer package for practitioners. Now: LISREL, EQS, Mplus, AMOS, sem of Stata, . . . ; free software: sem, lavaan, . . .
- IV-SEM: Instrumental variable (IV) analysis embedded into SEM. The SEM framework delivers IV methods (Sargan's approach and the GMM),when marginalizing the SEM model to a regression equation.

# Discussion (cont.)

- Karl G. Jöreskog's LISREL (early 70s): (a) a general model that encompass shocks, errors, and latent variables, (b) ML estimation and testing methods (c) A computer package for practitioners. Now: LISREL, EQS, Mplus, AMOS, sem of Stata, ...; free software: sem, lavaan, ...
- IV-SEM: Instrumental variable (IV) analysis embedded into SEM. The SEM framework delivers IV methods (Sargan's approach and the GMM),when marginalizing the SEM model to a regression equation.
- Sources of endogeneity. (a) from error: a congeneric measure serves as an IV; (b) from shock, congeneric measure is not a valid IV.

# Discussion (cont.)

- Karl G. Jöreskog's LISREL (early 70s): (a) a general model that encompass shocks, errors, and latent variables, (b) ML estimation and testing methods (c) A computer package for practitioners. Now: LISREL, EQS, Mplus, AMOS, sem of Stata, ...; free software: sem, lavaan, ...
- IV-SEM: Instrumental variable (IV) analysis embedded into SEM. The SEM framework delivers IV methods (Sargan's approach and the GMM),when marginalizing the SEM model to a regression equation.
- Sources of endogeneity. (a) from error: a congeneric measure serves as an IV; (b) from shock, congeneric measure is not a valid IV.
- Impressive Norway contribution to " founding fathers" of psychometrics/econometric methods: ... Ragnar Frisch (1895-1973, Oslo); Olav Reiersøl (1908 – 2001, Oslo); Trygve Haavelmo (Skedsmo, 1911 - Oslo, 1999, Nobel Laureate 1989), Karl B. Jöreskog (affiliation with BI, Oslo).

Thank You!

Albert Satorra (albert.satorra@upf.edu)

This talk here: http://www.econ.upf.edu/ satorra/GatheringOslo2022.pdf