The <u>Z</u>ero <u>I</u>nformation <u>L</u>imit <u>C</u>ondition <u>H</u>ypothesis

Or,

Why Do Some Good Models Go Bad?

Presentation October 13, 2008

Charles R. Nelson

Department of Economics
University of Washington

Based on Nelson and Startz (2007)

To motivate, consider nonlinear regression – say the Phillips Curve

$$y_i = \gamma \bullet (x_i + \beta \bullet z_i) + \varepsilon_i$$

Asymptotic variance of $\hat{\beta}$ is proportional to γ^{-2}.

Thus γ controls the amount of **information** in the data: $I = V^{-1} \propto \gamma^2$.

But γ is unknown in practice, so standard error and for $\hat{\beta}$ based on $\hat{\gamma}$.

This distinction matters for inference: $t = (\hat{\beta} - \beta_0) \bullet \sqrt{\hat{I}}$

Example: sample size n=100; x, and z and ε uncorrelated $N(0,1)$, and $\gamma = .01$.

- MC median standard error is 2 vs. asymptotic 'true' value 10.
- Reflects upward bias in $\hat{\gamma}^2$ that is relatively large when γ is small.
- Too-small standard errors would suggest too-big t-stats, but……
- Actual size of the t-test at the nominal .05 level is only .01!

These distortions occur in a wide class of models that satisfy the *Zero-Information-Limit-Condition* of Nelson & Startz (2006).

Examples: IV, ARMA, GARCH, Unobserved Components.

We would like to understand:

Why is test size too small in some cases, too large in others?

How can we do valid inference in ZILC models?

The Zero-Information-Limit Condition (ZILC).

Parameter of interest is β, but there is also, an 'identifying parameter', γ.

Information measure for $\hat{\beta}$ is the inverse of the variance:

$$I_{\hat{\beta}}(\beta, \gamma, \boldsymbol{\sigma}, \mathbf{W}) = [V_{\hat{\beta}}(\beta, \gamma, \boldsymbol{\sigma}, \mathbf{W})]^{-1}.$$

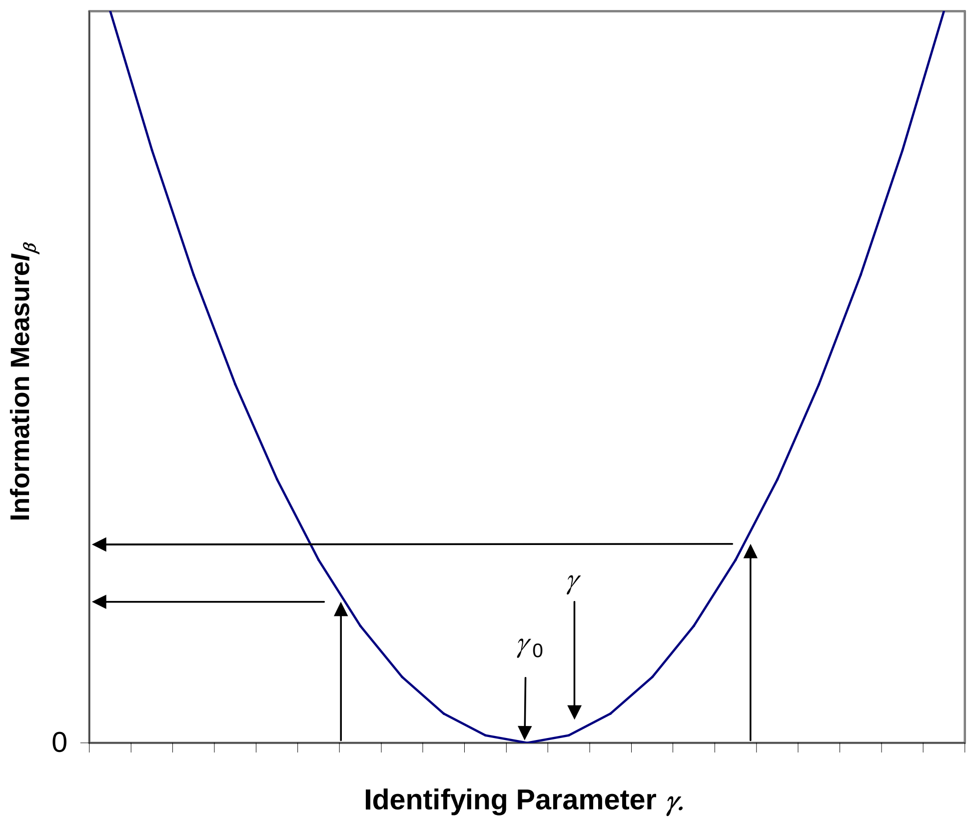
ZILC holds when there is a value of γ, say γ_0, such that

$$\lim_{\gamma \to \gamma_0} I_{\hat{\beta}}(\beta, \gamma, \boldsymbol{\sigma}, \mathbf{W}) = 0.$$

Noting zero and first order derivatives are zero at *ZILC* point:

$$I_{\hat{\beta}}(\beta, \gamma, \boldsymbol{\sigma}, \mathbf{W}) \cong (\gamma - \gamma_0)^2 \bullet I''(\beta, \gamma_0, \boldsymbol{\sigma}, \mathbf{W})/2.$$

Figure 2.1: The *ZILC* point and upward bias in estimated information.



Implication: upward bias in estimating $(\gamma\text{-}\gamma_0)^2$.

$$E(\hat{\gamma}-\gamma_0)^2 = (\gamma-\gamma_0)^2 + V_{\hat{\gamma}}$$

Suggests, though does not prove, that

$$E[I_{\hat{\beta}}(\hat{\beta},\hat{\gamma},\hat{\boldsymbol{\sigma}},\mathbf{W})] > I_{\hat{\beta}}(\beta,\gamma,\boldsymbol{\sigma},\mathbf{W})$$

It is *relative* bias in $\hat{\gamma}$ and $\hat{I}$ that matters since

$$t_{\hat{\beta}} = (\hat{\beta}-\beta_0)\sqrt{\hat{I}_{\hat{\beta}}}\ ,$$

Asymptotic theory is still valid in spite of *ZILC*!

Hypothesis: in models where *ZILC* holds,

estimated information is too large, standard errors too small, and

bias will be larger the closer γ is to the *ZILC* point.

We refer to this conjecture as the ZILC Hypothesis, or

ZILCH

Instrumental Variables.

$$y = \beta x + \varepsilon$$

$$x = \gamma z + v$$

$$V\begin{bmatrix} \varepsilon \\ v \end{bmatrix} = \begin{bmatrix} \sigma_{\varepsilon}^{2} & \rho\sigma_{\varepsilon}\sigma_{v} \\ \rho\sigma_{\varepsilon}\sigma_{v} & \sigma_{v}^{2} \end{bmatrix}$$

Under normality *IV* is also ML, and the inverse of the asymptotic variance is:

$$I_{\hat{\beta}}(\beta, \gamma, \sigma_{\varepsilon}, Z) = \gamma^{2}\left[\frac{T \bullet m_{zz}}{\sigma_{\varepsilon}^{2}}\right].$$

Clearly *ZILC* holds in the *IV* model, with γ_0 being 0.

Strong simultaneity - big ρ - has received most attention

in the ‘weak instrument’ literature:

There t-test rejects the true null hypothesis far too often.

That is the intuitive result: SE too small, $\hat{\beta} / SE$ too big.

But look at what happens when ρ is zero:

* No simultaneity, so least squares is optimal,

but one does *IV* anyway.

* Of interest in practice; isolates the role of the weak instrument.

* *IV* estimator is median unbiased, though its distribution has fat tails.

Estimated information is: $\hat{I} = I_{\hat{\beta}_{IV}}(\hat{\beta}_{IV}, \hat{\gamma}, \hat{\sigma}_{\varepsilon}, \mathbf{Z}) = \hat{\gamma}^2 \left[\dfrac{T \bullet m_{zz}}{s^2_{(y - \hat{\beta}_{IV} x)}} \right]$

Monte Carlo experiment:

$\beta = 0;$

$\gamma = .01;$

$\sigma_{\varepsilon} = \sigma_{\nu} = 1;$

$m_{zz} = 1;$

$\rho = 0;$

Implying that: $I_{\hat{\beta}_{IV}} = .01.$

EViews™ using the standard routines and output. n=100.

MC Sample 1,000 runs.

Median of $\hat{I}$ = 0.21 is much larger than theoretical value of 0.01!

In testing the null hypothesis $\beta = 0$ we use:

(1) true *I*, unknown to a real investigator, and

(2) estimated EViews standard error).

* At a nominal 5% level rejection frequency using true *I* is .03.

* Using t-statistic from EViews, rejection frequency is zero!

How can we overestimate *I* but get t-statistics that are too small?

Normalizing m_{zz}=1:

$$t^2_{\hat{\beta}_{IV}} = (\hat{\beta}_{IV} - \beta)^2 \bullet \hat{I} = (\frac{m_{z\varepsilon}}{\hat{\gamma}})^2 \bullet \left[\frac{T \bullet \hat{\gamma}^2}{s^2_{(y-\hat{\beta}_{IV}x)}} \right] = \left[\frac{\sqrt{T} \bullet m_{z\varepsilon}}{s_{(y-\hat{\beta}_{IV}x)}} \right]^2 .$$

Note $\hat{\gamma}$ is in both the numerator and denominator of.

Numerator has variance σ_ε^2 over an estimate of σ_ε^2.

$\hat{\beta}_{IV}$ is median unbiased but has a large dispersion,

So s^2 tends to be too large; MC median of s^2 is 1.44.

Net effect: t-statistics are too small and frequency of rejection is zilch!

Note ratio representation of $\hat{\beta}$ ***:***

Asymptotic variance of $\hat{\beta} = \hat{\lambda} / \hat{\gamma}$ by 'delta method':

$$V_{\hat{\beta}}(\lambda, \gamma, \sigma, W) = \frac{1}{\gamma^2} \bullet [V_{\hat{\lambda} - \beta\hat{\gamma}}(\beta, \sigma, W)]$$

Thus:

ZILC holds for <u>any ratio of regression coefficients</u>.

Pause while we change to another file!

What do *ZILC* Models have in common?

A common linear representation.

Consider single equation models of the form

$$y_i = f(\beta, \gamma, x_i) + \varepsilon_i ; i = 1,..,N \quad (2.7)$$

ε is a random error and parameters are estimated by G-N.

The linear approximation to the model takes the form:

$$y_i = f(\beta_*, \gamma_*, x_i) + (\beta - \beta_*) \bullet f_{\beta,i} + (\gamma - \gamma_*) \bullet f_{\gamma,i} + e_i$$

$$y_{*,i} = \beta \bullet f_{\beta,i} + \gamma \bullet f_{\gamma,i} + e_i \tag{2.8}$$

Gauss-Newton iterates on the parameter values.

Functional form of the model:

The asymptotic variance for G-N implies the first equality.

ZILC provides the second.

$$I_{\hat{\beta}} = \sigma^{-2} \bullet \left(\sum f_{\beta}^{2} - \frac{(\sum f_{\beta} f_{\gamma})^{2}}{\sum f_{\gamma}^{2}}\right) = (\gamma - \gamma_{0})^{2} \bullet \frac{I_{\hat{\beta}}^{"}}{2} \tag{2.9}$$

It follows from the second equality that:

The first derivative of $f(\beta, \gamma, x_i)$ with respect to β

is proportional to (γ-γ_0).

The linear approximation to a *ZILC* model is then:

$$y_{*,i} = [\beta \bullet (\gamma - \gamma_0)] \bullet g(\beta, x_i) + \gamma \bullet f_\gamma + e_i \tag{2.12}$$

Implication:

In models of the form (2.7) for which *ZILC* holds, the G-N estimate of β can be approximated as the ratio of least squares linear regression coefficients.

To a linear approx the functional form is:

$$\begin{aligned} f(\beta,\gamma,x_i) &= (\gamma-\gamma_0)\bullet G(\beta,x_i) \\ G(\beta,x_i) &= \int g(\beta,x_i)\,d\beta \end{aligned} \qquad (2.13)$$

Appears to restrict the class of models severely.

Actually encompasses **seemingly unrelated** models.

Ratio representation of $\hat{\beta}$ implies spurious inference:.

Asymptotic variance of the estimator $\hat{\beta} = \hat{\lambda} / \hat{\gamma}$:

$$V_{\hat{\beta}}(\lambda, \gamma, \sigma, W) = \frac{1}{\gamma^2} \bullet [V_{\hat{\lambda}-\beta\hat{\gamma}}(\beta, \sigma, W)] \qquad (2.14)$$

Thus:

ZILC holds for any ratio of regression coefficients.

Now the square of the t-statistic is:

$$t_{\hat{\beta}}^2 = \frac{(\hat{\beta}-\beta_0)^2}{\hat{V}_{\hat{\beta}}} = \left[\left(\frac{\hat{\lambda}-\beta_0\hat{\gamma}}{\hat{\gamma}} \right)^2 \Bigg/ \left(\frac{\hat{V}_{[\hat{\lambda}-\beta\hat{\gamma}]}}{\hat{\gamma}^2} \right) \right] = \frac{(\hat{\lambda}-\beta_0\hat{\gamma})^2}{\hat{V}_{[\hat{\lambda}-\beta\hat{\gamma}]}}$$

The notation $\hat{V}_{\hat{\beta}}$ means evaluated at parameter estimates.

In practice V is evaluated using $\beta = \hat{\beta}$.

Standard errors are from estimation algorithm.

Evaluating variance:

$$\hat{V}_{(\hat{\lambda}-\beta\hat{\gamma})} = V_{(\hat{\lambda}-\beta\hat{\gamma})}(\hat{\lambda},\hat{\gamma}) = \hat{V}_{\hat{\lambda}} - 2 \bullet \hat{\beta} \bullet \hat{C}_{\hat{\lambda},\hat{\gamma}} + \hat{\beta}^2 \bullet \hat{V}_{\hat{\gamma}} \qquad (2.16)$$

Consider case where the covariance is zero:

$$\hat{V}_{(\hat{\lambda}-\beta\hat{\gamma})} \approx \hat{V}_{\hat{\lambda}} + \hat{\beta}^2 \bullet \hat{V}_{\hat{\gamma}} = [\hat{V}_{\hat{\lambda}} + \beta^2 \bullet \hat{V}_{\hat{\gamma}}] + (\hat{\beta}^2 - \beta^2) \bullet \hat{V}_{\hat{\gamma}}$$

First term is a proper standard error, augmented by second.

For null hypothesis β=0 we would have:

$$t_{\hat{\beta}}^2 \approx \frac{\hat{\lambda}^2}{\hat{V}_{\hat{\lambda}} + \hat{\beta}^2 \bullet \hat{V}_{\hat{\gamma}}} \leq t_{\hat{\lambda}}^2 .$$

Additional term in denominator is always greater than zero, indeed will be large if β is poorly estimated, so t-statistic **smaller** than the t-statistic for λ.

The size of the t-test for $\boldsymbol{\beta}$ will therefore be too small, though the standard error of $\hat{\beta}$ is too small!

Strong correlation case is important in practice;

t-stats are too small.

The ARMA (1,1) model with near cancellation.

The ARMA(1,1) model may be written:

$$(1-\phi L)y_t = (1-\theta L)\varepsilon_t;\, t=1,..,T;\, \varepsilon_t \sim \text{i.i.d.}\, N(0,\sigma_\varepsilon^2),\ \ |\phi|\langle 1; |\theta|\langle 1.$$

The model is identified only if $\phi \neq \theta$, with asymptotic covariance:

$$V_{\hat{\phi},\hat{\theta}}(\phi,\theta) = T^{-1}\frac{(1-\phi\theta)}{(\theta-\phi)^2}\begin{bmatrix}(1-\phi^2)(1-\phi\theta) & (1-\phi^2)(1-\theta^2)\\ (1-\phi^2)(1-\theta^2) & (1-\theta^2)(1-\phi\theta)\end{bmatrix},$$

What happens when the difference between ϕ and θ is small?

Monte Carlo: $\phi = .01$ and $\theta = 0$ with T=1,000,

t-test null $\theta = 0$ had an actual size .46!

*Surprising in light of the constraint on estimates to fall into -1,1 interval!

*Not limited to sample sizes that economists usually think of as small.

Implies tendency to over-estimate MA order.

Plays key role in model specification.

Example: Unobserved Components Model implies ARMA(p,p).

Re-parameterize the model in terms of θ and difference $\gamma=\phi-\theta$:

$$(1-(\theta+\gamma)L)y_t = (1-\theta L)\varepsilon_t .$$

The covariance matrix $V_{\hat{\theta},\hat{\delta}}(\theta,\delta)$ is:

$$I_{\hat{\theta}}(\theta,\gamma) = \frac{\gamma^2 T}{(1-\theta^2+\theta\gamma)^2(1-\theta^2)}.$$

ZILC holds for the ARMA(1,1), γ with $\gamma_0 = 0$.

$I_{\hat{\theta}}(\hat{\theta},\hat{\gamma})$ depends primarily on the estimate of γ^2 - insensitive to θ.

The bias in estimating γ^2 is variance of $\hat{\gamma}$:

$$V_{\hat{\gamma}}(\theta,\gamma) = T^{-1}[1-(\theta^2-\theta\gamma)^2] \cong T^{-1}(1-\theta^4),$$

Note that *ZILC* does not hold for γ, which is well identified.

Upward bias will be large if γ^2 is small relative to T^{-1}.

For example above, T = 1,000, $\theta = 0$, and $\gamma = .01$:

$$E(\hat{\gamma}^2) = \gamma^2 + V(\hat{\gamma}) = \gamma^2 + T^{-1} = .0001 + .001 = .0011$$

Sampling Experiments.

Monte Carlo DGP is AR(1).

Fit both AR(1), which is correctly specified and well identified, and ARMA(1,1) which is also correctly specified but weakly identified.

True θ is zero, so $\gamma(=\phi)$ is both identifying parameter and AR coefficient.

Series length T is 1,000, and the number of replications is 1,000.

Estimation done within EViews™, standard errors from its non-linear estimation.

Estimation of ϕ in AR(1) model.

Asymptotic standard error of $\hat{\phi}$ is roughly $T^{-0.5}$ or .032.

Corresponds closely to standard deviation in MC sample,

and to the median standard error in the MC sample.

Median I in MC sample is close to the asymptotic value of roughly 1,000.

The t-test has nearly correct size and considerable power.

Thus, for AR(1), asymptotic theory is very accurate in a series of length 1,000.

Estimation of θ in ARMA(1,1).

Identification of both ϕ and θ in the ARMA(1,1) model is conditional on the absolute difference between them.

Actual standard deviation cannot be larger than one because coefficients are bounded within (-1,1).

But downward bias in standard error so strong that median se is well below the actual standard deviation.

Correspondingly, $\hat{I}_{\hat{\theta}}$ is substantially overestimated.

Upward bias is much greater than can be attributed to $\delta^2 T$ alone.

For ϕ=.2, asymptotic theory works well

.

Size of t-test is greatly in excess .05.

Size of t-test using asymptotic formula also excessive.

Do standard errors under null hypotheses help?

It rejects far too infrequently!

Table 3.1.1:
Inference for θ in ARMA(1,1)
DGP is AR (1); T=1,000

	True value of $\phi(=\gamma)$:			
	0.01	0.05	0.10	0.20
Std Dev. of $\hat{\theta}$:				
Asymptotic (true)	3.162	0.632	0.316	0.158
In MC sample	0.659	0.556	0.381	0.180
MC median Std Error	0.359	0.333	0.256	0.152
Information measure $I_{\hat{\theta}}$:				
Asymptotic (true)	0.10	2.50	10.00	40.00
MC median of est. *I*.	7.77	9.01	15.29	43.41
$E\left[\hat{\gamma}^2 T\right]$	1.1	3.5	11	41
Tests of null hypothesis θ=0 at nominal .05 level, rejections :				
EViews t-test.	0.457	0.358	0.245	0.118
Asy SE using $\hat{\phi}, \hat{\theta}$.	0.371	0.341	0.229	0.116
Likelihood ratio test.	0.178	0.142	0.103	0.073

	ARMA(1,1) selected over AR(1), frequency:			
AIC	0.380	0.337	0.251	0.180
SIC	0.040	0.026	0.019	0.014

Identifying role of γ suggests a conditional test:

If it exceeds a critical value designed to give correct size in the experiment for $\phi = .01$,
then the customary t-test is used,
otherwise the null $\theta = 0$ is accepted.
Unfortunately, the size is somewhat excessive.
Size of the likelihood ratio test is also too large,
though less excessive than t-test.

What does work well??
Schwarz Information Criterion (SIC) selects the ARMA(1,1) model over correct AR(1) infrequently.
Outperforms AIC; see Lutkepohl (1991).

Inference for $\gamma(=\phi)$ in the ARMA(1,1).

Identifying parameter γ is well identified,
asymptotic standard error $T^{0.5}$,

MC standard deviation and MC median of standard errors are close!

Better size results are obtained by using the asymptotic standard deviation, $T^{0.5}$.

Power against null γ =0 is comparable to t-test in the AR(1) model.

Asymptotic theory works well for the identifying parameter.

Non-Linear Regression

Consider non-linear regression models of the form

$$y_i = \gamma \bullet f(\beta, w_i) + \varepsilon_i; \quad i = 1,.., N. \tag{3.2.1}$$

where errors ε are i.i.d. normal with standard deviation σ_ε.

$$I_{\hat{\beta}}(\beta, \gamma, \sigma_\varepsilon, W) = \gamma^2 \left[\frac{T \bullet m_{11} \bullet (1 - r_{01}^2)}{\sigma_\varepsilon^2} \right], \tag{3.2.2}$$

where m_{11} denotes sample first moment of the first derivative of f;
r_{01} denotes correlation between the zero and first derivatives.

ZILC holds since I goes to zero as γ approaches zero.

Case where *f* is linear in exogenous variables *x* and *z*:

$$y_i = \gamma(x_i + \beta z_i) + \varepsilon_i . \qquad (3.2.3)$$

Linearization of *f(.)* gives (3.2.3),

with *x* and *z* being the 0 and 1 order derivatives respectively.

Example Staiger, Stock, and Watson (1997) Phillips Curve model

$$\Delta\pi_t = \gamma(u_{t-1} - \bar{u}) + \varepsilon_t$$

π is the inflation rate, u is the unemployment rate, and $\bar{u}$ is NAIRU.

The reduced form for (3.2.3) is the linear regression:

$$y_i = \gamma x_i + \lambda z_i + \varepsilon_i . \tag{3.2.4}$$

Since β is exactly identified:

$$\hat{\beta} = \hat{\lambda} / \hat{\gamma} . \tag{3.2.5}$$

The square of the t-ratio for the null hypothesis $\beta = \beta_0$ may be written:

$$t_{\hat{\beta}}^2 = (\frac{\hat{\lambda}}{\hat{\gamma}} - \beta_0)^2 \bullet \left[\hat{\gamma}^2 \frac{N \bullet m_{11}(1 - r_{01}^2)}{s^2} \right] \quad (3.2.3)$$

As in the IV model, $\hat{\gamma}$ appears in both the denominator and the numerator of the t-ratio, so sampling variation will tend to cancel.

Simulated uncorrelated regressors x and z with unit variances,

β=0 and σ_ε=1, obtained the following output from the EViews™.

Sampling Distributions for Non-Linear Regression

		Information I_β		Standard Error of $\hat{\beta}$		Frequency
γ	N	Asymptotic	Median	Asymptotic	Median	$\lvert t_\beta \rvert > 1.96$
1	**100**	100	108	0.10	0.10	0.049
0.1	100	1	0.88	1.0	1.1	0.001
0.1	1,000	10	9.3	0.32	0.33	0.012
0.1	**10,000**	100	99	0.10	0.10	0.045
0.01	100	0.01	0.20	10	2.2	0.001
0.01	100,000	10	9.1	0.32	0.33	0.013
0.01	**1,000,000**	100	98	0.10	0.10	0.053

Where γ is 0.01, size is not correct at N = 100,000!

Need N = 1,000,000! Why?

$$t_{\hat{\beta}}^2 = \left(\frac{\hat{\lambda}}{\hat{\gamma}}\right)^2 \bullet \frac{\hat{\gamma}^2 \bullet N}{s^2(1+\hat{\beta}^2)} = \hat{\lambda}^2 \bullet \frac{N}{s^2} \bullet \frac{1}{(1+\hat{\beta}^2)} = t_{\hat{\lambda}}^2 \bullet \frac{1}{(1+\hat{\beta}^2)}$$

This is a proper t-statistic – that for classical regression coefficient λ - times a quantity *always* less than one, reducing its dispersion regardless of sample size.

Stochastic Regressors with Near Multicollinearity

DGP: Regressors, x_1 and x_2, are jointly distributed with:

$$Cov(x_{1,i}, x_{2,i}) = \begin{bmatrix} \sigma_1^2 & \gamma\sigma_1\sigma_2 \\ \gamma\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

Error term, ε, independent of regressors with standard deviation σ_ε.

The ith observation on the dependent variable, y_i, is generated by

$$y_i = \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i,$$

Coefficients are to be estimated by least squares.

To check whether *ZILC* applies in this model we note that

$$I(\beta, \gamma, \sigma_{\varepsilon}, W) = \left(\sigma_{\varepsilon}^{2}\left[X'X\right]^{-1}\right)^{-1} = I(\sigma_{\varepsilon}, X),$$

where X denotes the N by 2 matrix of observations drawn from the distribution of the regressors.

Model is not identified if γ=1, in which case I is identically zero,

But *ZILC* does not hold because I is a function of the actual realization of X, not of the parameters of its distribution which includes γ.

In particular, I will depend on the sample correlation between the two regressors in the realization of X, and that will be zero with probability zero.

In the case of Normal errors the t-statistic will have be t-distributed and will reflect appropriately the lack of information arising from correlation between the regressors.

Summary and Conclusions

Weak identification implies spurious inference in some models, but not all.

The Zero-Information-Limit-Condition (*ZILC*) distinguishes models where spurious inference occurs.

Then information or precision overestimated,
standard errors are too small.

ZILC models share a common linear approximation in which parameter of interest is the ratio of regression coefficients.

This common linear representation links seemingly unrelated models such as ARMA and IV.

Whether the t-test is under- or over-sized depends on the correlation between coefficients.

Counterintuitive result is undersized t-tests in the face of underestimated standard errors.

END.