Social Statistics Seminar Series:

Adjustment of Recall Errors in Duration Data Using SIMEX

Jose Pina-Sánchez

- Survey data is plagued by measurement error

Recall errors

- Survey data is plagued by measurement error
- Misunderstanding, self-aquiescence, social desirability, etc

Recall errors

- Survey data is plagued by measurement error
- Misunderstanding, self-aquiescence, social desirability, etc
- In addition, retrospective questions are prone to recall errors

Implications

Classical measurement error

$$X^* = X + U$$

Implications

Classical measurement error

$$X^* = X + U$$

Impact in simple linear regression

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Implications

Classical measurement error

$$X^* = X + U$$

Impact in simple linear regression

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$\beta_1^* = \beta_1 \rho = \beta_1 \frac{S_X^2}{S_X^2 + S_U^2}$$

Implications

Classical measurement error

$$X^* = X + U$$

Impact in simple linear regression

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$\beta_1^* = \beta_1 \rho = \beta_1 \frac{S_X^2}{S_X^2 + S_U^2}$$

Impact in multiple linear regression

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$$

Implications

Classical measurement error

$$X^* = X + U$$

Impact in simple linear regression

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$\beta_1^* = \beta_1 \rho = \beta_1 \frac{S_X^2}{S_X^2 + S_U^2}$$

Impact in multiple linear regression

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$$

$$\beta_1^* = \beta_1 \rho' = \beta_1 \frac{S_{X/Z}^2}{S_{X/Z}^2 + S_U^2}$$

Implications

Classical measurement error

$$X^* = X + U$$

Impact in simple linear regression

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$\beta_1^* = \beta_1 \rho = \beta_1 \frac{S_X^2}{S_X^2 + S_U^2}$$

Impact in multiple linear regression

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$$

$$\beta_1^* = \beta_1 \rho' = \beta_1 \frac{S_{X/Z}^2}{S_{X/Z}^2 + S_U^2}$$

$$\beta_2^* = \beta_2 + \beta_1 (1 - \rho') \Gamma$$

Adjustments are exceptions

- Very little is done among applied statisticians

Adjustments are exceptions

- Very little is done among applied statisticians
- Latent variable estimation cannot always be used

Adjustments are exceptions

- Very little is done among applied statisticians
- Latent variable estimation cannot always be used
- Shortage of data in the form of replicated, validation or instrumental variables

Adjustments are exceptions

- Very little is done among applied statisticians
- Latent variable estimation cannot always be used
- Shortage of data in the form of replicated, validation or instrumental variables
- Complexity of methods

SIMEX

- SIMEX can be an interesting alternative to tackle measurement error in survey data

- SIMEX can be an interesting alternative to tackle measurement error in survey data
- It is relatively easy to implement and, in its standard form, it only requires $\hat{S}_U^2$

SIMEX

- SIMEX can be an interesting alternative to tackle measurement error in survey data
- It is relatively easy to implement and, in its standard form, it only requires $\hat{S}_U^2$
- *"The key idea underlying SIMEX is the fact that the effect of measurement error on an estimator can be determined experimentally via simulation"*

(Carroll et al, 2006, p.98)

SIMEX

- SIMEX can be an interesting alternative to tackle measurement error in survey data
- It is relatively easy to implement and, in its standard form, it only requires $\hat{S}_U^2$
- *"The key idea underlying SIMEX is the fact that the effect of measurement error on an estimator can be determined experimentally via simulation"*

  (Carroll et al, 2006, p.98)
- I proceed to show how that is done for the case of $X^* = X + U$ in $Y = \beta_0 + \beta_1 X^* + \epsilon$

SIMEX

1. Additional variables, $X_k^*(\lambda_k)$, with increasing levels of measurement error are generated by the rule $X_k^*(\lambda_k) = X^* + \sqrt{(\lambda_k)}U$

SIMEX

1. Additional variables, $X^*_k(\lambda_k)$, with increasing levels of measurement error are generated by the rule $X^*_k(\lambda_k) = X^* + \sqrt{(\lambda_k)}U$
2. The outcome model is re-estimated using X^*_k, which produces $\hat{\beta}^*_{1k} = \beta_1 \frac{S^2_X}{S^2_X + (1+\lambda_k)S^2_U}$

SIMEX

1. Additional variables, $X_k^*(\lambda_k)$, with increasing levels of measurement error are generated by the rule $X_k^*(\lambda_k) = X^* + \sqrt{(\lambda_k)}U$

2. The outcome model is re-estimated using X_k^*, which produces $\hat{\beta}_{1k}^* = \beta_1 \frac{S_X^2}{S_X^2+(1+\lambda_k)S_U^2}$

3. Steps 1 and 2 are repeated B in order to obtain $\overline{\hat{\beta}_{1k}^*}$ and reduce Monte Carlo error

SIMEX

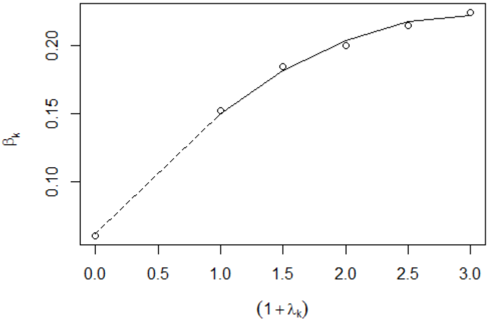
1. Additional variables, $X_k^*(\lambda_k)$, with increasing levels of measurement error are generated by the rule $X_k^*(\lambda_k) = X^* + \sqrt{(\lambda_k)}U$
2. The outcome model is re-estimated using X_k^*, which produces $\hat{\beta}_{1k}^* = \beta_1 \frac{S_X^2}{S_X^2+(1+\lambda_k)S_U^2}$
3. Steps 1 and 2 are repeated B in order to obtain $\overline{\hat{\beta}_{1k}^*}$ and reduce Monte Carlo error
4. $\overline{\hat{\beta}_{1k}^*}$ and λ_k can now be paired and plotted to reproduce the extrapolation function

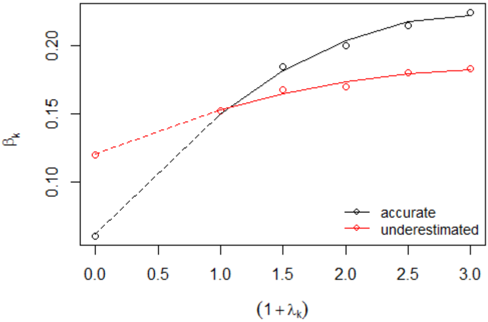
SIMEX

1. Additional variables, $X_k^*(\lambda_k)$, with increasing levels of measurement error are generated by the rule $X_k^*(\lambda_k) = X^* + \sqrt{(\lambda_k)}U$
2. The outcome model is re-estimated using X_k^*, which produces $\hat{\beta}_{1k}^* = \beta_1 \frac{S_X^2}{S_X^2+(1+\lambda_k)S_U^2}$
3. Steps 1 and 2 are repeated B in order to obtain $\overline{\hat{\beta}_{1k}^*}$ and reduce Monte Carlo error
4. $\overline{\hat{\beta}_{1k}^*}$ and λ_k can now be paired and plotted to reproduce the extrapolation function
5. A regression model with $\overline{\hat{\beta}_{1k}^*}$ as the response and λ_k as the explanatory variable is estimated

SIMEX

1. Additional variables, $X_k^*(\lambda_k)$, with increasing levels of measurement error are generated by the rule $X_k^*(\lambda_k) = X^* + \sqrt{(\lambda_k)}U$
2. The outcome model is re-estimated using X_k^*, which produces $\hat{\beta}_{1k}^* = \beta_1 \frac{S_X^2}{S_X^2+(1+\lambda_k)S_U^2}$
3. Steps 1 and 2 are repeated B in order to obtain $\overline{\hat{\beta}_{1k}^*}$ and reduce Monte Carlo error
4. $\overline{\hat{\beta}_{1k}^*}$ and λ_k can now be paired and plotted to reproduce the extrapolation function
5. A regression model with $\overline{\hat{\beta}_{1k}^*}$ as the response and λ_k as the explanatory variable is estimated
6. $\hat{\beta}_{SIMEX}$ can now be calculated by extrapolating to $\lambda_k = -1$


0.10
0.15
0.20
βk
0.0
0.5
1.0
1.5
2.0
2.5
3.0
(1+λk)


0.10
0.15
0.20
β_k
0.0
0.5
1.0
1.5
2.0
2.5
3.0
(1+λ_k)
accurate
underestimated

Recall errors

- The classical measurement error model does not represent recall errors adequately

Recall errors

- The classical measurement error model does not represent recall errors adequately
- Retrospective reports of times to onset/end or durations of an event cannot be negative and will tend to be more accurate the shorter the recall timespan

Recall errors

- The classical measurement error model does not represent recall errors adequately
- Retrospective reports of times to onset/end or durations of an event cannot be negative and will tend to be more accurate the shorter the recall timespan
- e.g. jobseekers' reports on how long they have been unemployed

Recall errors

- The classical measurement error model does not represent recall errors adequately
- Retrospective reports of times to onset/end or durations of an event cannot be negative and will tend to be more accurate the shorter the recall timespan
- e.g. jobseekers' reports on how long they have been unemployed
- Multiplicative models are elegant solutions; $X^* = X \cdot U$

Recall errors

- The classical measurement error model does not represent recall errors adequately
- Retrospective reports of times to onset/end or durations of an event cannot be negative and will tend to be more accurate the shorter the recall timespan
- e.g. jobseekers' reports on how long they have been unemployed
- Multiplicative models are elegant solutions; $X^* = X \cdot U$
- Can also be used to model telescoping effects:

Recall errors

- The classical measurement error model does not represent recall errors adequately
- Retrospective reports of times to onset/end or durations of an event cannot be negative and will tend to be more accurate the shorter the recall timespan
- e.g. jobseekers' reports on how long they have been unemployed
- Multiplicative models are elegant solutions; $X^* = X \cdot U$
- Can also be used to model telescoping effects: for forward telescoping $U \sim N(< 1, S_U^2)$, for backward telescoping $U \sim N(> 1, S_U^2)$

Multiplicative SIMEX

We substitute step 1,

$$X_k^*(\lambda_k) = X^* + \sqrt{(\lambda_k)}U$$

by either:

$$X_k^*(\lambda_k) = exp\left\{log(X^*) + \sqrt{(\lambda_k)}log(U)\right\}$$

(Carroll et al, 2006)

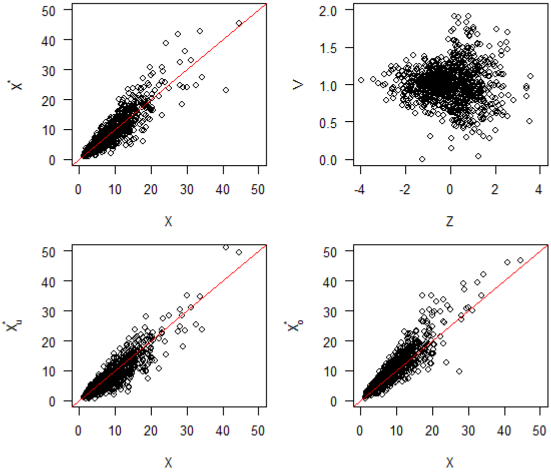
$$X_k^*(\lambda_k) = X^* \cdot U^{\lambda_k}$$

(Biewen et al., 2008)

Simulations

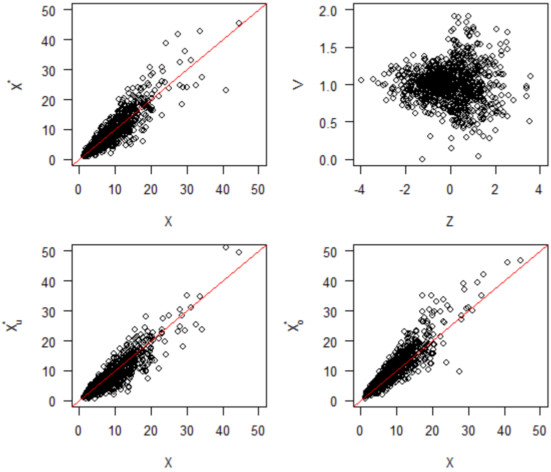
Variable	Mean	SD	Min	Max
Y	.01	.99	-3.01	3.54
Y_{ca}	.5	.5	0	1
Y_{co}	.69	.82	0	4
Z	-.02	1.1	-3.97	3.56
X	8.75	5.23	1.17	44.45

Simulated errors



Simulated errors

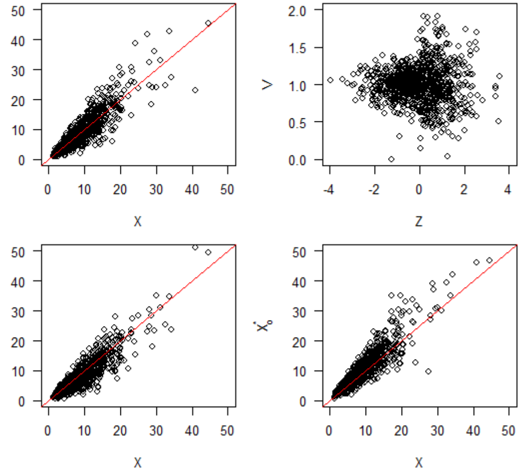
$U \sim N(1, .25)$



Simulated errors

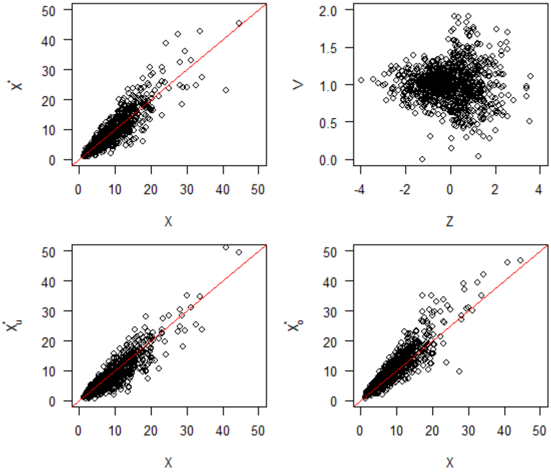
$U \sim N(1, .25)$

$U \sim N(.9, .25)$



Simulated errors

$U \sim N(1, .25)$



$U \sim N(.9, .25)$

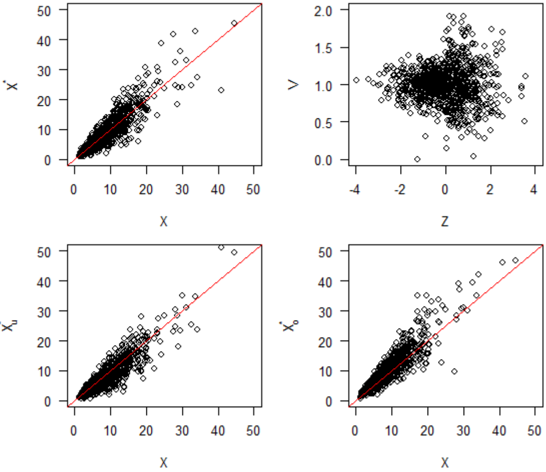
$U \sim N(1.1, .25)$

Simulated errors

$U \sim N(1, .25)$

if $Z < 0$
$U \sim N(1, .30)$

if $Z > 0$
$U \sim N(1, .15)$



$U \sim N(.9, .25)$

$U \sim N(1.1, .25)$

Outcome models

Linear

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$$

Logit

$$\mu = E(Y^{ca})$$

$$log(\frac{\mu}{1-\mu}) = \beta_0 + \beta_1 X + \beta_2 Z$$

Poisson

$$\mu = E(Y^{co})$$

$$log(\mu) = \beta_0 + \beta_1 X + \beta_2 Z$$

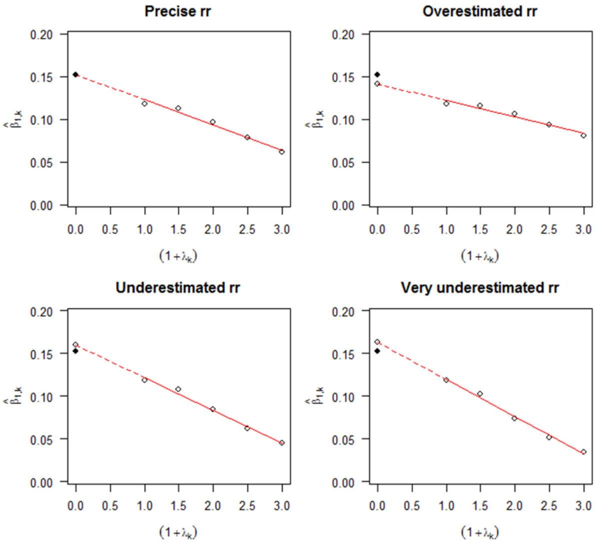
Impact

		Linear				Logit				Poisson			
		Coef	SE	Bias	R.Bias	Coef	SE	Bias	R.Bias	Coef	SE	Bias	R.Bias
True model	β_0	-1.297	.035			-5.997	.388			-1.362	.069		
	β_1	.150	.003			.768	.050			.092	.004		
	β_2	.111	.016			.210	.099			.082	.038		
Naïve: multi.	β_0	-1.013	.038	.284	21.9%	-3.810	.258	2.187	36.5%	-1.198	.065	.284	20.8%
	β_1	.118	.004	-.032	21.2%	.494	.034	-.275	37.5%	.075	.004	-.032	34.6%
	β_2	.156	.019	.045	40.9%	.275	.084	.065	30.9%	.157	.037	.045	55.1%
Naïve: hetero.	β_0	-.998	.039	.372	28.7%	-3.649	.248	2.372	39.5%	-1.178	.065	.372	27.4%
	β_1	.116	.004	-.043	28.8%	.471	.032	-.299	38.9%	.074	.004	-.043	47.0%
	β_2	.169	.020	.067	60.7%	.377	.085	.199	94.8%	.141	.037	.067	81.7%
Naïve: under.	β_0	-.937	.038	.325	25.1%	-3.441	.233	1.962	32.7%	-1.054	.059	.325	23.9%
	β_1	.121	.004	-.024	15.9%	.490	.033	-.183	23.8%	.069	.004	-.024	26.1%
	β_2	.166	.020	.052	46.9%	.321	.084	.124	58.9%	.149	.037	.052	63.2%
Naïve: over.	β_0	-1.028	.038	.245	18.9%	-4.029	.266	1.842	30.7%	-1.110	.061	.245	18.0%
	β_1	.108	.003	-.039	26.1%	.470	.031	-.284	37.0%	.061	.003	-.039	42.7%
	β_2	.150	.019	.053	48.0%	.284	.087	.152	72.4%	.134	.037	.053	64.6%

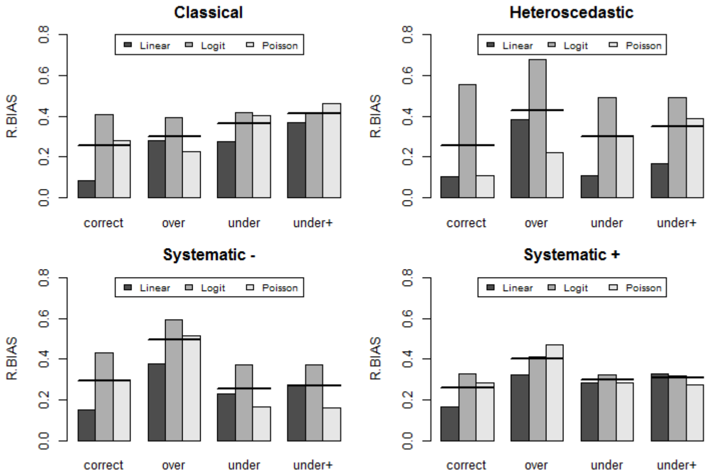
Adjustments

Scenarios	ρ	S_U^2
Correct	.816	.25
Overestimated	.9	.176
Underestimated	.7	.336
Very Underestimated	.6	.412

Results: linear model



Results


Classical
Heteroscedastic
Systematic -
Systematic +
Linear
Logit
Poisson
R.BIAS
0.0
0.2
0.4
0.6
0.8
correct
over
under
under+

Conclusion

- SIMEX is a good candidate to be used as a sensitivity tool when recall errors are suspected:

Conclusion

- SIMEX is a good candidate to be used as a sensitivity tool when recall errors are suspected:
  - Can be easily modified to accommodate memory failures in the form of multiplicative errors

Conclusion

- SIMEX is a good candidate to be used as a sensitivity tool when recall errors are suspected:
  - Can be easily modified to accommodate memory failures in the form of multiplicative errors
  - By shifting the mean of the simulated errors up or down we can account for backward and forward telescoping effects

Conclusion

- SIMEX is a good candidate to be used as a sensitivity tool when recall errors are suspected:
  - Can be easily modified to accommodate memory failures in the form of multiplicative errors
  - By shifting the mean of the simulated errors up or down we can account for backward and forward telescoping effects
  - Easy to implement regardless of the complexity of the outcome model

Conclusion

- SIMEX is a good candidate to be used as a sensitivity tool when recall errors are suspected:
  - Can be easily modified to accommodate memory failures in the form of multiplicative errors
  - By shifting the mean of the simulated errors up or down we can account for backward and forward telescoping effects
  - Easy to implement regardless of the complexity of the outcome model
  - It can be particularly useful when no alternative data is available and all we know stems from previous validation studies

Conclusion

- SIMEX is a good candidate to be used as a sensitivity tool when recall errors are suspected:
  - Can be easily modified to accommodate memory failures in the form of multiplicative errors
  - By shifting the mean of the simulated errors up or down we can account for backward and forward telescoping effects
  - Easy to implement regardless of the complexity of the outcome model
  - It can be particularly useful when no alternative data is available and all we know stems from previous validation studies
  - In all of the 144 estimates studied here SIMEX managed to reduce the bias found in the naive model by at least 22.6%

Conclusion

- SIMEX is a good candidate to be used as a sensitivity tool when recall errors are suspected:
  - Can be easily modified to accommodate memory failures in the form of multiplicative errors
  - By shifting the mean of the simulated errors up or down we can account for backward and forward telescoping effects
  - Easy to implement regardless of the complexity of the outcome model
  - It can be particularly useful when no alternative data is available and all we know stems from previous validation studies
  - In all of the 144 estimates studied here SIMEX managed to reduce the bias found in the naive model by at least 22.6%
  - There are commands to run standard SIMEX in STATA and R, incorporating extensions for multiplicative errors would make it even easier for everyone to consider sensitivity analysis when using retrospective data.