Assessing Measurement Errors in the R&D-Innovation-Productivity Relationships

Jacques Mairesse
CREST-ENSAE,
UNU-MERIT and NBER

Stéphane Robin
PRISM - Sorbonne
University of Paris 1

CONCORDi 2017 Annual Conference
Seville, September 27-29, 2017

Motivation

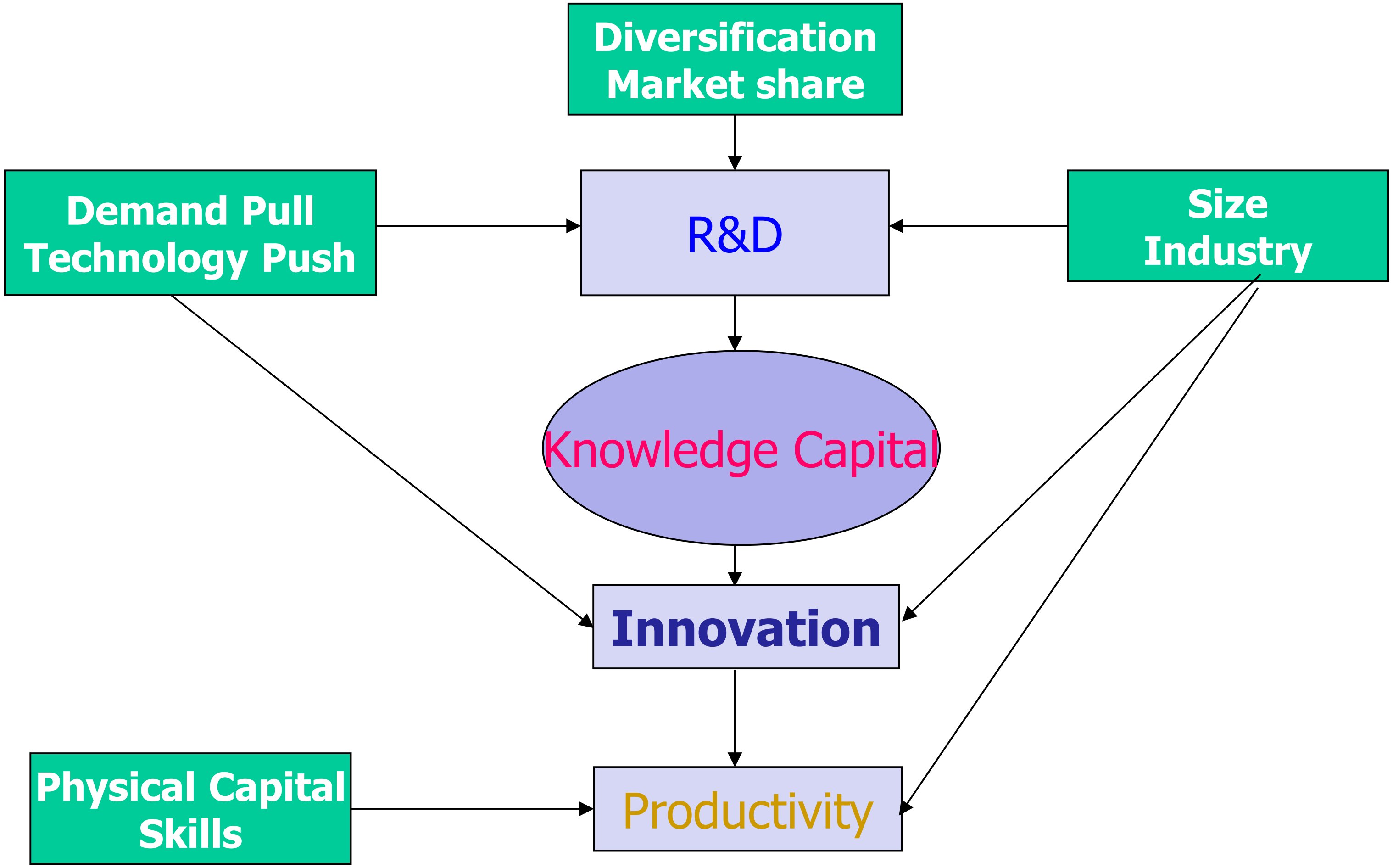
- Mairesse, Mohnen and Kremp (2005) *Ann. Eco & Stats* compare several models derived from CDM to measure the impact of innovation on productivity.

- Their analysis suggests “important measurement errors in the innovation intensity variables, (…) in the innovation binary indicators as well as in the R&D intensity and the R&D doing indicator”.

- Analysis conducted in most CDM-type model using CIS innovation-type data is basically cross sectional.

- Crépon, B., E. Duguet and J. Mairesse (1998), “Research and Development, Innovation and Productivity: An Econometric Analysis at the Firm Level”, *Economics of Innovation and New Technology*, 7(2), 115-158.

The « CDM » model



The « CDM » model

with Bruno Crepon and Emmanuel Duguet

- Brings together the three main fields of investigation in the econometrics of research and innovation
- Proposes a “simple” framework articulating innovative and productive activities
- Takes advantage of the innovation survey information
- Uses estimation methods appropriate to the specification of the model and nature of data

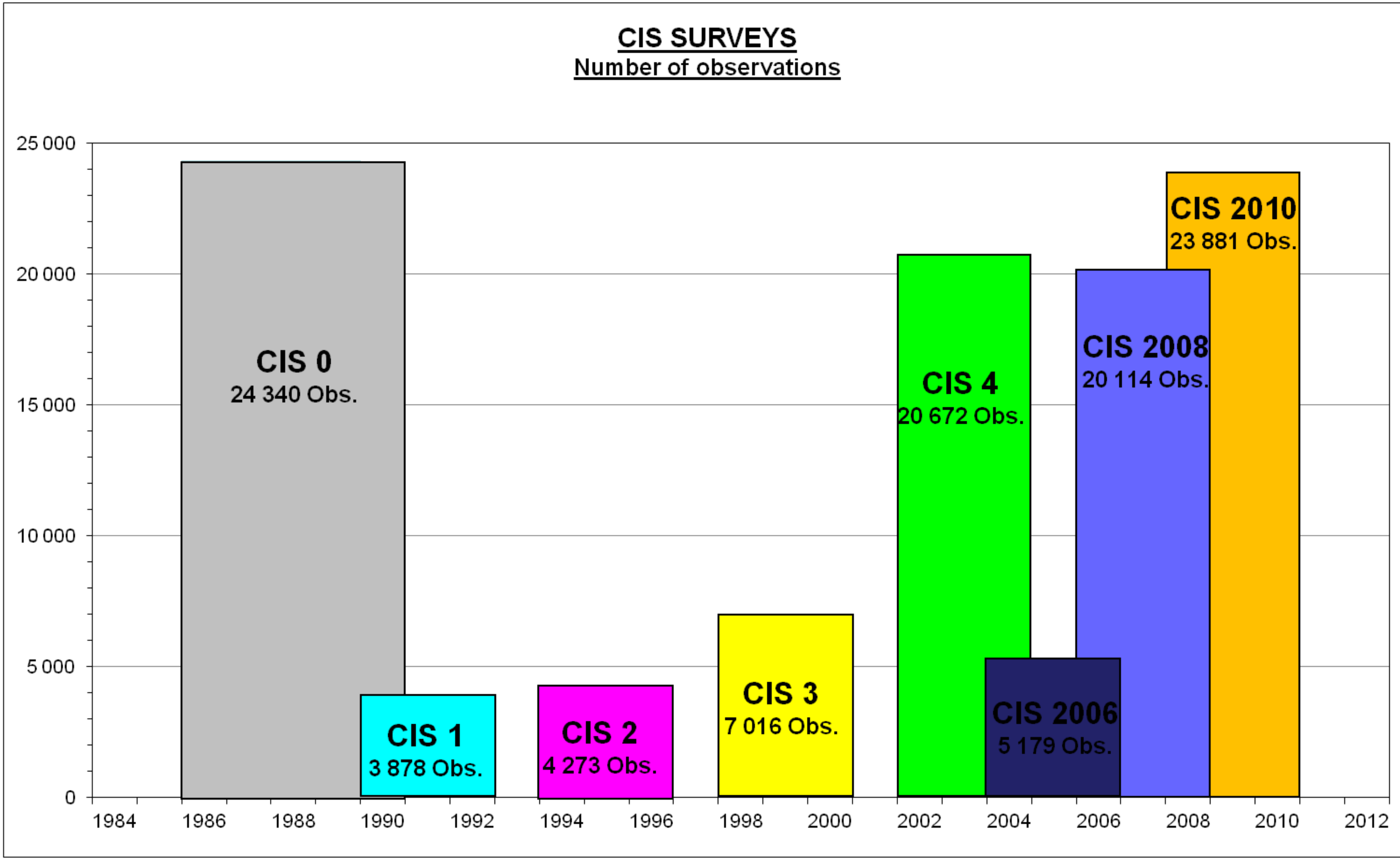
Objective(s) of the present analysis

- Gather several waves of the French CIS, in order to build (balanced and unbalanced) panels, and focus on an unbalanced panel data sample based on CIS2000, CIS2004 and CIS2008.
- Consider three regressions: R&D extended productivity function, and CDM innovation-productivity and R&D-innovation relations and estimate them using four usual panel data OLS estimators: Total and Between in levels and in four year differences
- Assess the importance of measurements errors in R&D and innovation by comparing attenuation biases in both TL/ BL and TD/BD pairs of estimators of the three elasticities of interest, assuming these measurement errors are main source of misspecification and are “white”noise.

Outline of the presentation

- Description of the CIS panel
- Method
- Simple and multiple productivity and innovation regressions
- Illustration of results
- Conclusions so far and research plan

Waves of the French CIS



The “CIS 0 3 4 8” panel

- We merged the prototype CIS 0 (1990) with the 3rd (2000), 4th (2004) and 2008 waves of the French CIS

- Selection rule: keep firms from CIS *t* (*t* = 2000, 2004, 2008) which are observed in CIS 0

```
CIS Wave   |CIS0(90)|CIS3 (2004)|CIS4(2004)|CIS 2008
------------+---------------------------------------
# of firms|  5467  |  2600    |  3021   |  2155
----------+-----------------------------------------
```

- There are:

1059 firms observed in 1990, 2000 *and* 2004

1027 firms observed in 1990, 2004 *and* 2008,

1171 firms observed in 2000, 2004 and 2008,

579 firms observed in 1990, 2000, 2004 *and* 2008.

Mean values of key variables, by CIS wave

```
Variable                 | CIS 0   | CIS 3   | CIS 4   | CIS 2008
-------------------------+-----------------------------------------
# of employees in year t | 328.7   | 505.9   | 362.2   | 392.9
                         |(2267.6) |(2990.8) |(2230.0) |(1621.5)
# of employees in t-2    | NA      | 498.6   | 376.4   | 414.4
                         |         |(2939.6) |(2298.2) |(2330.9)
Labour productivity (t)  | 136.6   | 192.9   | 244.0   | 262.3
                         |(470.4)  |(784.1)  |(982.7)  |(682.9)
Labour productivity (t-2)| NA      | 184.9   | 230.0   | 245.6
                         |         |(671.8)  |(922.6)  |(696.7)
Continuous R&D (1/0)     | NA      | 0.37    | 0.34    | 0.35
                         |         |(0.48)   |(0.47)   |(0.48)
R&D Intensity            | NA      | 2.81    | 5.86    | 3.85
                         |         |(8.00)   |(10.77)  |(15.24)
Product innovator (1/0)  | 0.67    | 0.49    | 0.43    | 0.46
                         |(0.47)   |(0.50)   |(0.50)   |(0.50)
% of innovative sales    | 0.46    | 0.07    | 0.10    | 0.11
 (coded 0/1/2/3 in CIS 0)|(0.77)   |(0.13)   |(0.20)   |(0.21)
-------------------------+-----------------------------------------
```

CIS study panel

- For the analysis conducted in this paper, we focused on the balanced panel constructed by merging the firms that have answered to the three consecutive CIS2000, CIS 2004 and CIS2008 surveys.
- Our sample thus consists of 1171 firms present in 2000, 2004 and 2008 that correspond to 3513 observations in levels and 2342 observations in four-year differences.
- Although not fully representative of the manufacturing sector, it provides a reasonable coverage of medium or large size firms that are more likely to be innovative than smaller firms.
- It is important to stress here that the estimators in differences we will be considering will be in fact four-year-difference estimators

CIS study panel

- For the analysis conducted in this paper, we focused on the balanced panel constructed by merging the firms that have answered to the three consecutive CIS2000, CIS 2004 and CIS2008 surveys.
- Our sample thus consists of 1171 firms present in 2000, 2004 and 2008 that correspond to 3513 observations in levels and 2342 observations in four-year differences.
- Although not fully representative of the manufacturing sector, it provides a reasonable coverage of medium or large size firms that are more likely to be innovative than smaller firms.
- It is important to stress here that the estimators in differences we will be considering will be in fact four-year-difference estimators

Warning

- We do not expect that the magnitude of measurement errors computed for four-year-difference estimators would be much higher than that computed with the level estimators.

- This is contrary to what can be usually observed with an estimator in 'true' first-differences (i.e. with differences between two consecutive years).

- The reason is the fact that the auto-correlations of both R&D and innovation measuring their persistence are much smaller over a period of four years than over one of two.

- NOTE: OLS attenuation biases from classical (or approximately classical) errors-in-variables are strongly exacerbated when one performs regressions in first-differences instead of levels, or, alternatively, when one controls for fixed effects, insofar that the variables affected by measurement errors are often strongly autocorrelated and the measurement errors are non-autocorrelated or weakly so. [Mairesse 1990; Griliches-Mairesse 1998).

Benchmark econometric analysis

- We estimate the following models IN TL/BL and TD/BD:
- (1) RD extended productivity function:

$$\ln(Q/L)_{it} = \beta_0 + \beta_1 \ln(RD/L)_{it} + \beta_2 \ln(C/L)_{it} + \beta_3 \ln(M/L)_{it} + \beta_4 \ln(L)_{it} + \varepsilon_{it}$$

- (2) CDM Innovation-Productivity relation:

$$\ln(Q/L)_{it} = \beta_0 + \beta_1 \mathrm{logit}(SHI)_{it} + \beta_2 \ln(C/L)_{it} + \beta_3 \ln(M/L)_{it} + \beta_4 \ln(L)_{it} + \varepsilon_{it}$$

- (3) CDM RD-Innovation relation:

$$\mathrm{logit}(SHI)_{it} = \beta_0 + \beta_1 \ln(RD/L)_{it} + \beta_4 \ln(L)_{it} + \varepsilon_{it}$$

Classical Errors in Variables (CEV)

- The CEV problem occurs when at least one regressor in an econometric model is affected by measurement issues.

- If this occurs with explanatory variable x, we can write x as:

$$x = x^* + e$$

  - where x^* denotes the **true value** of the variable
  - and e denotes the **measurement error**.

- In the standard linear regression model estimated by OLS, the CEV results in an **attenuation bias** measured by parameter λ.

The attenuation bias

- λ is a synthetic indicator of the importance of measurement errors in an explanatory variable.

- In the standard linear regression model, it is given by :

$$\lambda = \frac{\sigma_e^2}{\sigma_x^2}$$ in a simple regression where the regressor x is affected by a CEV

$$\lambda = \frac{\sigma_e^2}{\sigma_r^2}$$ in a multiple regression where a regressor is affected by a CEV,

where *r* is the error in the regression of “y conditional on all other regressors other than x” on “x conditional on all these other regressors” (Frish-Waugh, see slide 18)

Caveat

- The attenuation bias on the OLS estimator is often presented (using our notation, $x = x^* + e$, where x^* is the true value of x) as:

$$\hat{\beta}_1 \xrightarrow{p} \frac{\sigma^2_{x^*}}{\sigma^2_{x^*} + \sigma^2_e} \beta_1$$

- We slightly rewrite here as:

$$\hat{\beta}_1 \xrightarrow{p} \frac{\sigma^2_x - \sigma^2_e}{\sigma^2_x} \beta_1 = \left(1 - \frac{\sigma^2_e}{\sigma^2_x}\right) \beta_1 = (1 - \lambda)\beta_1$$

with $\sigma^2_x = \sigma^2_{x^*} + \sigma^2_e$ or $\sigma^2_{x^*} = \sigma^2_x - \sigma^2_e$ under the white noise assumption that $corr(x^*, e) = 0$.

Computing λ

- Because the measurement error e is unknown, it is generally impossible to compute λ when estimating $y = \beta x + \varepsilon = \beta(x^* + e) + \varepsilon$ (for instance) on cross-sectional data.

- Because several (related) estimators are available for panel data, it is possible to derive a value of λ from a comparison of a pair of these estimators, if these estimators are differently biased by CEV and assuming no other specification errors .

- See Mairesse (1990) explains how this becomes possible when conducting estimations on panel data . See also:Crépon B. and J. Mairesse (2008) .

Deriving λ from Pooled OLS vs Between

- Let β_1 be the true parameter of interest in a simple regression (i.e., in model (1) or (2) without a control for firm size).

- Let TL denote the "Pooled OLS" estimator in levels, and BL the "between" estimator in levels. Let β_{TL} and β_{BL} denote the estimates of β_1 by TL and BL, respectively.

- Because of the attenuation bias due to the CEV, we can write:

$\beta_{\text{TL}} \to (1-\lambda_{\text{TL}})\beta_1$ and $\beta_{\text{BL}} \to (1-\lambda_{\text{BL}})\beta_1$

which in large samples leads to:

$$\frac{\beta_{TL}}{1-\lambda_{TL}} = \frac{\beta_{BL}}{1-\lambda_{BL}}$$

Simple algebra thus yields:

$$\beta_{TL} - \lambda_{BL}\beta_{TL} = \beta_{BL} - \lambda_{TL}\beta_{BL}$$

$$\Leftrightarrow \beta_{TL} - \frac{\sigma^2_{e\,BL}}{\sigma^2_{x\,BL}}\beta_{TL} = \beta_{BL} - \frac{\sigma^2_{e\,TL}}{\sigma^2_{x\,TL}}\beta_{BL}$$

$$\Leftrightarrow \beta_{TL} - \frac{\sigma^2_{e\,BL}}{\sigma^2_{x\,BL}}\beta_{TL} = \beta_{BL} - \frac{\sigma^2_{e\,TL}}{\sigma^2_{x\,TL}}\beta_{BL}$$

$$\Leftrightarrow \beta_{TL} - \frac{\frac{1}{T}\sigma^2_{e\,TL}}{\sigma^2_{x\,BL}}\beta_{TL} = \beta_{BL} - \frac{\sigma^2_{e\,TL}}{\sigma^2_{x\,TL}}\beta_{BL}$$

using $\sigma^2_{e\,BL} = \frac{1}{T}\sigma^2_{e\,TL}$ (see Mairesse, 1990).

Finally rearranging last equation leads to:

(5)

$$\boxed{\sigma^2_{e\,TL} = \frac{\beta_{TL} - \beta_{BL}}{\dfrac{\beta_{TL}}{T\sigma^2_{x\,BL}} - \dfrac{\beta_{BL}}{\sigma^2_{x\,TL}}}}$$

and :

(6)

$$\boxed{\sigma^2_{e\,BL} = \frac{1}{T}\sigma^2_{e\,TL} = \frac{\beta_{TL} - \beta_{BL}}{\dfrac{\beta_{TL}}{\sigma^2_{x\,BL}} - \dfrac{T\beta_{BL}}{\sigma^2_{x\,TL}}}}$$

where the β's are estimated and all other parameters are constant (T is sample time length and the σ^2's are sample statistics).

from which we derive the formula for λ_{TL} and λ_{BL} :

(5) $$\boxed{\lambda_{TL} = \frac{\sigma^2_{e\,TL}}{\sigma^2_{x\,TL}} = \frac{\beta_{TL} - \beta_{BL}}{\sigma^2_{x\,TL}\left(\frac{\beta_{TL}}{T\sigma^2_{x\,BL}} - \frac{\beta_{BL}}{\sigma^2_{x\,TL}}\right)}}$$

(6) $$\boxed{\lambda_{BL} = \frac{\sigma^2_{e\,BL}}{\sigma^2_{x\,BL}} = \frac{\beta_{TL} - \beta_{BL}}{\sigma^2_{x\,BL}\left(\frac{\beta_{TL}}{\sigma^2_{x\,BL}} - \frac{T\beta_{BL}}{\sigma^2_{x\,TL}}\right)}}$$

where the β's are estimated and all other parameters are constant (T is sample time length and the σ^2's are sample statistics).

Similarly using $\sigma^{2}_{e\;BD} = \frac{2}{T-1}\sigma^{2}_{e\;TD}$ we obtain:

$$\boxed{\sigma^{2}_{e\;TD} = \frac{\beta_{TD}-\beta_{BD}}{\frac{2\beta_{TD}}{(T-1)\sigma^{2}_{x\;BD}} - \frac{\beta_{BD}}{\sigma^{2}_{x\;TD}}}}$$

$$\boxed{\sigma^{2}_{e\;BD} == \frac{2(\beta_{TD}-\beta_{BD})}{\frac{2\beta_{TD}}{\sigma^{2}_{x\;BD}} - \frac{(T-1)\beta_{BD}}{\sigma^{2}_{x\;TD}}} = \frac{\beta_{TD}-\beta_{BD}}{\frac{\beta_{TD}}{\sigma^{2}_{x\;BD}} - \frac{T-1}{2}\frac{\beta_{BD}}{\sigma^{2}_{x\;TD}}}}$$

from which we similarly derive the formula for λ_{TD} and λ_{BD} :

$$\boxed{\lambda_{TD} = \frac{\beta_{TD} - \beta_{BD}}{\sigma^2_{x\,TD}\left(\frac{2\beta_{TD}}{(T-1)\sigma^2_{x\,BD}} - \frac{\beta_{BD}}{\sigma^2_{x\,TD}}\right)}}$$

$$\boxed{\lambda_{BD} = \frac{2(\beta_{TD} - \beta_{BD})}{\sigma^2_{x\,BD}\left(\frac{2\beta_{TD}}{\sigma^2_{x\,BD}} - \frac{(T-1)\beta_{BD}}{\sigma^2_{x\,TD}}\right)} = \frac{\beta_{TD} - \beta_{BD}}{\sigma^2_{x\,BD}\left(\frac{\beta_{TD}}{\sigma^2_{x\,BD}} - \frac{T-1}{2}\frac{\beta_{BD}}{\sigma^2_{x\,TD}}\right)}}$$

where the β's are estimated and all other parameters are constant (*T* is sample time length and the σ^2's are sample statistics).

In a nutshell

- We estimate each of our benchmark models using : (i) “pooled OLS” in levels (TL), (ii) “between” in level (BL), (iii) “first-differences” (TD) and (iv) “between” in differences (BD)
- We compare two pairs of estimators to calculate λ :
  - Pooled OLS in levels and Between in levels: TL/BL
  - First Differences and Between in differences: TD/BD
- Each comparison yields two estimates of σ^2_e and thus two estimated λ.

Multiple regressors and Frisch-Waugh

- We get around this problem using the Frisch-Waugh procedure:
  1. Using OLS, regress the dependent variable *y* on a vector *z* including all regressors except *x*, our (imperfectly measured) variable of interest.
  2. Predict u_1, the residuals of this first regression
  3. Regress *x* on *z* using OLS
  4. Predict u_2, the residuals of this second regression
  5. Regress u_1 on u_2 using panel estimators to get the estimated β's associated with *x* and use these estimates to compute the λ's.

- This allows us to compute λ using the formulas given above in slides (15), (16) and (17).

Example of CDM Innovation-Productivity relation

- $\ln(Q/L)_{it} = \beta_0 + \beta_1 \text{logit}(SHI)_{it} + \beta_2 \ln(C/L)_{it} + \beta_3 \ln(M/L)_{it} + \beta_4 \ln(L)_{it} + \varepsilon_{it}$

- $\text{logit}(SHI)_{it}$ is logit share of the share of innovative sales (i.e. ratio of firm sales in year t for new or substantially improved products introduced in years t-2, t-1 and t in firm total sales)

- We derive four estimated σ^2_e and λ using our four panel estimators and applying the formulas given above.

- To compute standard errors on them we rely on **bootstrap**. Since we may have $\sigma^2_e < 0$ and λ *<0* or *>1,* when we bootstrap the estimations (3000 replications), we use two procedures : (1) keep only the λ *>0 and <1* ; (2) compute σ^2_e as square root of σ^4_e and keeping λ*<1.*

Simple panel regression estimates

$$\ln LP_{it} = \beta_0 + \beta_1 \text{logit (SHI)}_{it} + \varepsilon_{it}$$

	Simple linear regressions			
Variable	**Panel estimator**			
	TL	BL	**TD***	**BD***
% innovative sales (logit transform)	0.04^{***} (0.01)	0.04^{***} (0.02)	0.07^{***} (0.02)	0.06^{***} (0.02)
R²	0.01	0.01	0.04	0.04
Fisher *F*	14.15^{***}	8.03^{***}	21.78^{***}	29.06^{***}
Observations	2112	2112	1398	1398
σ^2_e *(mean value)*	0.36	0.12	0.12	0.12
σ^2_x	3.06	1.72	4.05	1.52
λ: BS method 1 *BS method 2*	0.37** (0.19) 0.36* (0.21)	0.24 (0.16) 0.26 (0.20)	0.08 (0.06) 0.08 (0.06)	0.21 (0.15) 0.20 (0.16)
N: BS method 1 *BS method 2*	1704 2761	1761 2893	1835 2999	1831 2997

- All standard errors are heteroskedasticity-robust whenever possible
- Bootstrapped standard errors for λ . *** Warning: logit (SHI) in levels.**

Multiple panel regression estimates

$$\ln LP_{it} = \beta_0 + \beta_1 \text{logit (SHI)}_{it} + \beta_2 \ln(C/L)_{it} + \beta_3 \ln(M/L)_{it} + \beta_4 \ln(L)_{it} + \varepsilon_{it}$$

	Multiple linear regressions			
Variable	**Panel estimator**			
	TL	BL	**TD***	BD*
% innovative sales (Logit transform)	-0.004 (0.003)	-0.007* (0.003)	0.006**(0.003)	0.01*** (0.002)
Adjusted R^2	0.93	0.93	0.76	0.76
Fisher *F (p-value)*	0.000	0.000	0.000	0.000
Observations	1506	1506	992	992
σ^2_e *(mean value)*	2.13	0.71	1.40	1.40
σ^2_x	2.97	1.75	3.79	1.52
λ	0.71*** (0.15) 0.67 (0.57)	0.40*** (0.11) 0.38 (0.30)	0.35 (0.35) 0.46 (0.28)	0.22 (0.20) 0.49* (0.28)
N: BS method 1 *BS method 2*	2966 2752	2966 2966	783 1329	783 2515

- All standard errors are heteroskedasticity-robust whenever possible
- Bootstrapped standard errors for λ . *** Warning: logit (SHI) in levels.**

Main results

We find significant attenuation biases in all three equations. We thus observe that in the innovation-productivity equation attenuation biases are more important when we use the share of innovative sales than on the R&D productivity equation when we rely on R&D intensity, which is consistent with the conjecture expressed in Mairesse, Mohnen, and Kremp (2005). The evidence that measurement errors in R&D and innovation are unrelated with likely measurement errors in physical capital stock is also noteworthy.

The observation that the measurement errors magnitudes on R&D and corresponding estimated elasticities are different in the R&D-productivity and R&D-innovation equations in both the Level and Difference panel data dimensions is also important. It confirms that it is not enough to take into account errors in variables to correct for estimation biases, and that other source of specification errors have to be considered, with the likely implication of an exacerbation of the measurement errors biases.

Work ahead…

A challenging follow-up to the present exploratory analysis will consist in specifying and estimating a dynamic CDM model that would explicitly relate productivity to innovation and innovation to R&D with lagged and feedback effects. This project will also extend the analysis by Raymond et al. 2015, for France and the Netherlands, which for France is based on a short panel very close to the one used here. It will have to be done on a panel with a longer time dimension, allowing us to scrutinize the three types of specifications errors that are typically taken into account in the CDM framework, selectivity and endogeneity issues as well as errors in variables.

A major objective of the project will be to obtain consistent estimates of the parameters of main interest, but also to characterize these specifications errors and assess their relative importance. A related objective will be to confront the results found in the cross-sectional and longitudinal data dimensions of the data, controlling or not for firm fixed effects or initial conditions.

References

- Mairesse, J. (1990) "Times-series and cross-sectional estimates on panel data: Why are they different and why should they be equal?", in: J. Hartog, G. Ridder, and J. Theeuwes (Eds), *Panel data and labor market studies* (North-Holland, Amsterdam), 81-95.
- Crépon, B., E. Duguet and J. Mairesse (1998), “Research and Development, Innovation and Productivity: An Econometric Analysis at the Firm Level”, *Economics of Innovation and New Technology*, 7(2), 115-158.
- Crépon B. and J. Mairesse (2008) “The Chamberlain Approach to Panel Data: An Overview and Some Simulations”, in Matyas L. and P. Sevestre (Eds) *The Econometrics of Panel Data: A Handbook of the Theory with Applications*, Chap. 5, 113-183.
- Mairesse J., P. Mohnen and E. Kremp (2005) "The Importance of R&D and Innovation for Productivity: A Reexamination in the Light of the French Innovation Survey", *Annals of Economics and Statistics*, 79/80, 486-527.
- Raymond W., J. Mairesse, P. Mohnen and F. Palm (2015) “ Dynamic Models of R&D, Innovation and Productivity: Panel Data Evidence for Dutch and French Manufacturing", *European Economic Review*, 78, 285-306.