Biases on variances estimated on large data-sets

François Gardes

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Documents de Travail du
Centre d'Economie de la Sorbonne

CES Working Papers

CES
Centre d'Économie de la Sorbonne
UMR 8174

Biases on variances estimated on large data-sets

François GARDES

2021.22

Maison des Sciences Économiques, 106-112 boulevard de L'Hôpital, 75647 Paris Cedex 13
https://centredeconomiesorbonne.cnrs.fr/
ISSN : 1955-611X

Biases on variances estimated on large data-sets

March 8, 2021

François Gardes, Paris School of Economics, Université Paris 1 Panthéon Sorbonne, Western Catholic University[1]

Abstract

The inverse dependency of the estimated variances over the sample size throws a fundamental question on the validity of the usual statistical methodology, since any hypothesis on the value of a coefficient can be tested negatively by increasing the size of the data-set. I suppose that large data-sets are characterized by a concentration of information on homogenous sub-populations, a spatial autocorrelation of the error terms and the covariates may bias the estimation of variances. Using the corrections of variances under spatial autocorrelation, we obtain variances comparable to an estimation on sub-samples (named efficient sub-samples) the sizes of which are sufficient to contain the information which gives rise to similar estimates to those obtained on the whole population. Moreover, the estimation on efficient data-sets does not necessitate the specification of the spatial autocorrelations which are supposed to bias the estimated variances.

Keywords: dataset, estimated variance, spatial autocorrelation, grouped observations

JEL Classification: C01, C12, C55

Introduction

Edward Leamer (1978) and Deidre McCloskey (1985), among others, remarked that the variances are probably underestimated on large data-sets, which biases all inferences and tests on microdata. Estimates of the coefficient variances in a linear model are generally very low once the data-set is sufficiently large,

[1]Centre d'Economie de la Sorbonne, 106-112 Boulevard de l'Hôpital, 75647, Paris Cedex 13, France. Email : gardes@univ-paris1.fr. Thanks are due for their remarks to Ali Abbas, Julien Boelaert, Francesco Gerotto, Philip Merrigan, Nicholas Sowels for his revision of the text and Christophe Starzec for the preparation of the data.

Documents de travail du Centre d'Economie de la Sorbonne - 2021.22

since increasing the sample size diminishes the standard error by some factor linked to the square root of the size. This occurs for instance for cross-sections of households or for financial data. That important question have recently been accompagnied by critics on the p-value and the usual 95% confidence limit, which does not take into account the statistical type 2 errors provoked by misspecification (Leamer, 1978; McCloskey, 1985; Ziliak and McCloskey, 2009; see also Nuzzo, 2018, for other references).

No plausible explanation of that under-estimation of variances has been proposed yet except the idea that information perhaps does not increase at the same speed as the data size (Leamer discussed by Darnell and Evans, 1990, page 101) and the proposal by Leamer to use Bayesian inference (since the Bayes factor is determined in part by the sample size) and to carry out conventional testing with significance levels which are decreasing functions of sample size. I suppose in this article that the supplementary amount of information contained in large data-sets decreases with the concentration of information on restricted areas (homogenous sub-populations), which creates a spatial autocorrelation of the error terms and the covariates. Angrist and Pischke (2000) note that while "heteroskedasticity rarely leads to dramatic changes in inference, clustering can make all the difference", for instance by an under-estimation of the standard error by a factor 2.65 in the STAR experiment on the score of students in the same class. The estimation on large data-sets provides another evidence of this importance of spatial autocorrelation.

Section 1 presents an example and our basic assumptions, section 2 an alternative method to remove the biases due to spatial autocorrelation and section 3 an application using a French survey containing more than ten thousand households.

1 The under-estimation of variances

1.1 An empirical example

I consider, first the estimation of an Engel curve for four expenditures (under a Working specification where the budget share is regressed on logarithmic income and other covariates) on a cross-section containing 10251 households, then the estimation of the same model over sub-populations of much more limited size taken at random from that survey. Compare for instance in Table 1 the estimation of a linear model, first on the whole data-set, then on five hundred sub-populations which have been taken at random for a certain relative size (1% or one third of the whole survey in column 2 and 3). The estimated income elasticities are in average very similar for these two types of data if the sub-populations contain, for instance, only 1% of the population (105 households), while the estimated variance is much smaller on large data-sets compared to the estimates on sub-populations. For sub-population corresponding in average

to 17% of the sample (an average of the size of the efficient sub-populations[2] for food, housing and transport and other expenditures), the average inflating factor for the standard error compared to the whole sample is for instance 1.49. Everything acts as if small sub-populations are sufficient to recover the values of the parameters while indicating much more dispersed estimates.

Table 1

Estimates of income elasticities (French INSEE survey, 2000)

Data	Whole Population	Sub-Populations (1%)	Efficient Sub-Populations
Food			
Average Income Elasticity:	0.5884	0.5859	0.5883
Empirical s.e.	-	0.0864	0.0128
Estimated s.e.	0.0072	0.0740	0.0134
Number of households	10251	103	2982
Housing			
Average Income Elasticity:	0.8339	0.8400	0.8347
Empirical s.e.	-	0.1051	0.0258
Estimated s.e.	0.0077	0.0796	0.0213
Number of households	10251	103	1350
Transport			
Average Income Elasticity:	0.8143	0.8157	0.8141
Empirical s.e.	-	0.0898	0.0263
Estimated s.e.	0.0088	0.1137	0.0208
Number of households	10251	103	1836
Other expenditures			
Average Income Elasticity:	1.244	1.284	1.253
Empirical s.e.	-	0.1033	0.0345
Estimated s.e.	0.0082	0.0815	0.0303
Number of households	10251	103	759

Note: Empirical s.e. are calculated across 500 sub-populations (with a fixed size) taken at random. Data-set: French INSEE surveys: the Family Expenditure Survey (FES, INSEE BDF 2001) and the Family Time Budget (FTB, INSEE BDT 1999).

[2]The *efficent sub-populations* are defined in section 2.2 as giving rise to a difference in the parameter smaller than 5% for at least 95% of samples having a definite size.

1.2 Basic Hypothesis: Borsuk-Ulam theorem

The first possible explanation could be that the increase in the number of observations creates numerous pairs of close observations, the residual distances of which (squared in the loss function) are much smaller than the average distance among a limited number of pairs of observations which differ largely as concerns both the endogenous variable and the covariates. The examination of a case where the data-set is distributed among cells of equal volume in a two-dimensional space shows that this explanation does not help to explain the different values of the residual variances. An analysis on simulated data[3] concludes to the same result.

Another possible explanation could be based on the Borsuk-Ulam's topological theorem which states that, for any set of observations, there is a tendency to a concentration of these observations on poles acting as barycenters of sub-populations of similar individuals[4]. Economic variables (both households' characteristics and the endogenous variables resulting from their choices) could concentrate households on selected parts of the possible location on the surface corresponding to all observed variables in a data-set [5]. Such a concentration of information in an economic data-set would imply a spatial correlation of the residuals which exists only when the number of observations is large. This spatial autocorrelation in turn will create an under-estimation of the variance-covariance matrix.

Borsuk-Ulam's theorem states that for every continuous mapping g from a n-dimensional sphere into R^n, there exist a point for which $g(x) = g(-x)$. This means that the second point (-x) can be removed from the sphere without loosing any information. In order to apply the Borsuk-Ulam's theorem to a data-set, consider a curve on a surface (y, x) in R^n, with x a set n explanatory variables and y the explained variable. The curve $y = g(x)$ can be considered as immersed in the smooth surface in R^n corresponding to all observation of (y, x) given by the data-set. This surface is thus a hypersurface of dimension (n-1) in the space of dimension n which corresponds to an economic model defined by the equation relating y to x. The empirical counterpart of this theorem is thus that considering the sample of observations (y, x) as immersed into a smooth n-dimensional surface, the point in the sample which is the closest to $(g(-x), x)$ could be removed without changing much the size of the sample. This operation can therefore be repeated till the remaining sub-sample cannot be considered as

[3] Performed by Julien Boelaert.

[4] This theorem has been for the first time used for a similar purpose by the French-american astrophysicist, Gérard de Vaucouleurs, who discovered, in the 50's, an unexpected concentration of galaxies into big galaxy clusters which cannot be explained within the traditional big-bang model: the universe is not as homogenous as was supposed by models based on the big-bang event sending all the matter uniformally in all directions. Tiny initial dissymetries (observed lately as waves in the initial states of the universe) were sufficient to create this concentration by the laws of celestial mechanics. As a consequence, these galaxy clusters exert gravitational effects which accelerate the galaxies.

[5] The concentration of similar individuals taking similar choices may exert an influence on other individuals by changing the costs corresponding to different socio-economic locations, and eventually creates mimetic behavior which increases the concentration.

a good approximation of the embedding smooth surface. A classical application of the theorem is known as the Ham Sandwich Theorem : M_1 to M_K being K bounded measurable subsets of R^n, there exist a n-1 dimensional hyperplane that cuts all of them exactly into half of their measure. Applying this theorem to our statistical sample, the whole population being divided into K homogenous groups (immersed into the subsets M_{K}), a similar reasoning allows to divide by two the number of observations in each group, leaving K sub-populations M'_k composed of similar individuals to the half of each groups $M"_k$ which have been removed (such that $M'_k \cup M"_k = M_k$). That means that the original sample can be finally reduced, by c consecutive operations, to a sub-sample containing 2^c less observations, the reduction finishing as soon as the remaining population cannot be considered as a good representation of the smooth surface (that is when the information contained in this surface is not fully represented by the final sub-population).

Our basic assumption can then be formulated: whenever the sample can be grouped into homogeneous sub-populations, there appears a correlation between the covariates within these sub-populations, since homogeneous populations means probably correlated explanatory variables between similar individuals. Therefore, it can be supposed that within these sub-populations, latent (unobserved) variables (which effects are composed into the residuals) are also correlated, which creates a correlation between the residuals within these sub-populations. This spatial correlation has as a consequence a negative bias in the estimated variances and covariances of the linear model, which depends on the size of the data-set.

2 Estimation of the bias : two methods

The various possible corrections of spatial autocorrelation are presented in the first sub-section. Another method, discussed in the second sub-section, is based on the definition of sub-populations the sizes of which are sufficient to recover all the information contained in the whole population and used to estimate the model. These sub-populations are named *efficient sub-populations*. The estimation on these sub-population gets rid of the biases in the estimation of variances provoked by the spatial autocorrelation due to the repetition of observations.

2.1 Usual Correction methods for the biases issued from a spatial autocorrelation

Bruce Moulton discussed the biases on estimated variances in the linear model when some covariates take a unique value in separate areas of the data-set: for instance the level of unemployment calculated for each state in a sample made over the whole US territory. His articles rely in fact on the formulas obtained by Kloek (1981) and Greenwald (1983). Bruce Greenwald gived a precise analysis of all types of biases as soon as some autocorrelation appears between residuals because of a spatial structure of the data-set (previous researches on the sub-

ject by Engle, Kloeck, Pakes and Pfeffermann were dedicated to the calculus of bounds to the variances for extreme cases of spatial correlation). Greenwald's analysis (equation 7 in Appendix A) shows that the inflating factor increases with the product of the intra-class correlations of the covariate and the residual (which vary according to the level of definition of the clusters) and with the clusters average size and size variance. The product of intra-class-class correlations decreases from 0.026 to 0.021 when the number of clusters decreases from 88 to 30, and to 0.001 for 5 clusters, but these movements are compensated by the increase of the average size and size variances of the clusters. These formulas are discussed in Appendix A.

2.2 Estimation on efficient sub-samples

Whence the econometrician restricts the size of the sub-sample among a basic large population applying the Ham Sandwich theorem, the amount of reduction of the sample can be regulated by a statistical method: the precision of estimates being measured by the empirical variances of the coefficients among k sub-population of equal size (which is very similar to the empirical estimate of the s.e., see Table 1), a test of a difference between the estimates of more than 5% for more than 5% of this set of sub-sample indicates which minimum size is necessary in order to have less than 5% chances to obtain an incorrect estimate (i.e. remote from the value obtained using the full reference population). This minimum size n* can be obtained for instance for the case of the income coefficient in the Engel curve, resolving the following equation between the empirical standard error σ estimated on the whole population of size N, the relative size of the sub-population compared to the whole population and the value of the coefficient estimated on the whole population. Suppose that the correct value of the variance $V(\widehat{\beta})$ is given by an estimation on a sub-sample containing a sufficient amount of information in order to have less than 5% chance that the estimate on the sub-sample differs from the true value estimated on the whole survey by less than 5%. The estimate of the variance on the whole sample will be under-estimated by a coefficient equal to the ratio of the sizes of the sub-sample and the whole sample:

$$V\left(\widehat{\beta}_N\right) = \frac{n^*}{N} V\left(\widehat{\beta}_{n^*}\right) \tag{1}$$

Indeed, the passage from the sub-sample to the whole sample can be considered as a replication as many times as necessary of sub-population containing the same information. The formula defining the minimum size writes:

$$1.96\ \widehat{\sigma}_N \sqrt{\frac{n^*}{N}} = 0.05 \widehat{\beta}_N \tag{2}$$

which implies:

$$n^* = \left[\frac{0.05\widehat{\beta}_N}{1.96\ \widehat{\sigma}_N}\right]^2 N \tag{3}$$

Estimating on sub-samples having this minimum size thus affords a correct average estimate with a variance corresponding to a sufficient amount of information. The final estimate is the value obtained by the estimation on the full population, with a variance which is the average of variances obtained on multiple efficient sub-populations.

Note that the size of the efficient sub-populations differs between commodities, being much greater for food (2982 households) than for the three other aggregates (1856, 1350 and 759 households being necessary for Transport, Housing and All Other expenditures respectively). This size can be interpreted as an indicator of the heterogeneity of the population as concerns the income elasticity of the commodity.[6]

The exactness of the estimation of variances can be checked by the absence of clustering effects measured by usual correction methods discussed in Appendix A.

3 Empirical application

A partition of a French Family Expenditures survey (2000, 10251 households, see a description of the data-set in Appendix B) into cells (defined by age classes, education level and family structure) defines 88 sub-populations containing in average 116 households. The same Working form as in Table 1 is adopted regressing the food budget share over total expenditures, the food price[7], the number of Unit of consumption, age and its square (all these explanatory variables in logarithm) and the proportion of children. The resulting income elasticity is 0.5884 (s.e. 0.0072) for the whole sample. The intra-group correlations of explanatory variables are 0.503 for total expenditure and 0.117 for the food price. The intra-group correlation of the residuals is 0.046. The resulting variance inflation obtained by Greenwald-Moulton's formula (8) for the income coefficient is 2.74, which inflates the standard error of the income elasticity by a factor 3.49 (considering only the intra-group correlation of total expenditures). A regression taking account of the clustering into groups by the Cluster Robust Variance Estimator (equation (6)) gives the same estimate of the income elasticity with a comparable inflating factor of the s.e. (2.99) [8]. A regression on the population aggregated into 88 cells gives a smaller elasticity with a much larger s.e. of

[6] A test of the representatbility of information in efficent sub-populations could rely on the representation in these sub-populations of observations pertaining to each cell of a choosen grouping of the population. Note that a different proportion of each group in an efficient sub-population compared to the whole population may not bias the estimation, at least if the grouping is not endogenous, - i.e. made at random as concerns the residuals of the estimation corresponding to the set of latent variables- since it corresponds to a weighted estimation with weight differing between the two data-sets

[7] Full prices defined in Gardes, 2019.

[8] Note that the inflation factor would be much larger for an estimation on smaller sub-populations of 105 households (1% of the whole sample) which gives also an average income elasticity close to the estimation on the whole population (see Table 1) but with a much larger dispersion.

0.0636 (see Hannan and Burnstein, 1974, for a precise analysis of biases from grouped observations).

These results show that the correction obtained by regressing on efficient sub-populations gives the same inflation factor as the correction of clustering effects by CRVE or the Greenwald-Moulton' formula. Similar results are obtained using the efficient sub-populations method for Housing, Transport and Other expenditures (inflating factors equal to 2.18, 2.10 and 1.68), CRVE (2.76, 2.37 and 3.71) and Greenwald-Moulton's (2.05, 2.89 and 2.00).

Table 2

Various estimates of the income elasticity of food (FES, France, 2000)

Data	Whole sample	Correction of the s.e.	Cluster-Robust Variance Estimator	Grouping into 88 cells	Efficient sub-populations
Elasticity	0.5884	0.5884	0.5884	0.6920	0.5883
s.e.	0.0072	0.0119	0.0216	0.0636	0.0134
Sample size	10251	10251	10251	88	2982
Method	OLS	Greenwald's equation (7)	Equation (6)	OLS	OLS

Note: Estimates of the Opportunity Cost by country, US dollars 2015, Cobb-Douglas specification. data-set: French INSEE surveys: the Family Expenditure Survey (FES, INSEE BDF 2001) and the Family Time Budget (FTB, INSEE BDT 1999).

I conclude that the estimation on efficient sub-populations estimates unbiased coefficients (compared to the estimation on the whole sample) but inflates the estimated variances by the same magnitude than the correction of clustering effects by Greenwald's equation. One can also remark that the correction of clustering effect, for instance by the Greenwald-Moulton's formula, diminishes by 40% when estimating on efficient sub-populations, which indicates that the definition of an efficient size of the data-set eliminates a large proportion of the spatial autocorrelations. On the contrary, grouping the data-set into homogenous cells changes a lot the estimated parameters because of the loss in degrees of freedom, and increases the estimated variances in a proportion much larger than all correction methods applied to individual data. The discussion of the relationship between the number of clusters and the spatial autocorrelation of the residuals must also take into account the change in the autocorrelation of the covariates which could increase together with the autocorrelation of the residuals, both increasing the inflating factors of estimated variances[9].

[9] The product of the intra-correlation of the covariate and the residuals decreases much (from 0.114 to 0.013 for Food expenditure) when the number of clusters is diminished from 88 to 5 (grouping according to the head's age).

Conclusion

Two methods have been proposed to estimate correctly the variances of the parameters of a linear model on a numerous data-set. Both give rise to much greater standard errors, with close inflating factors varying between 2 and 3 for the estimation of Engel curves for four aggregate commodities. The correction by means of clustering suffers from the arbitrary choice of the grouping. Also, clustering is not always possible if discriminant variables are not available to operate an efficient grouping. Besides, some among the formulas used in the literature in the correction of estimated variances apply only in case of one explanatory variable. The correlated changes, across different data-sets, of the spatial correlations of the residuals and of the covariates necessitates an important future study in order to estimate the relation between the structure of the population and the inflating factor of estimated variances. Another interesting problem is whether there exists an aggregate effect of various possible homogeneous groupings introducing independently spatial autocorrelations in the estimation.

On the contrary, the estimation on efficient sub-populations containing enough statistical information is based on a statistical choice which does not suppose an a-priori structure for a grouping of the data-set. Also, the second method is valid whatever the number of explanatory variables. Besides, the estimation on similar data-sets as concerns their size and structure allows to compare estimates performed on independent data-sets, giving rise to a normalization of standard errors given by these estimations. Therefore, in order to compare estimates corresponding to independent data-sets, a convenient method consist in taking at random subsets of each independent data-set corresponding to the same statistical rule defining the efficient sub-population. The second method thus applies to any econometric analysis, by OLS or other statistical models, and also to the estimation of a system of equations.

An interesting question is whether it is possible to split optimally the original sample applying the Borsuk-Ullam theorem: this necessitates the estimation of the hyperplans which split the sample into two similar sub-samples (which have been done at random in our empirical application).

Both methods convey the same information about the bias of the estimated variances, so that these corrected variances or the variances obtained on limited sub-populations should be used on micro-data for any test of significance or specification using the variance-covariance matrix.

Appendix A: Correction of the biases issued from a spatial autocorrelation

Kloek's analysis

Kloek considers the case where all regressors are constant within groups and groups have the same size m. In that case, the ratio between the true GLS

variance and the OLS is:

$$f(\hat{\beta}) = 1 + \rho(m-1) \tag{4}$$

where ρ is the intra-class correlation of the residual. If for example $\rho = 0.05$ and $m = 50$, the factor of under-estimation of standard errors is $\sqrt{f} = 1.86$.

Greenwald's formulas

Greenwald generalizes the analysis of the linear model: $y = X\beta + e$ and his basic formula evaluates the bias between the true variance-covariance matrix and the usual estimate in the OLS estimation of a linear model $\hat{C} = \sigma^2(X'X)^{-1}$:

$$E(\hat{C}) - C = E(\sigma^2)(X^{'}X)^{-1} - (X^{'}X)^{-1}X^{'}VX(X^{'}X)^{-1} \tag{5}$$

with $V = E(ee') = \sigma^2 A$, $\sigma^2 = \frac{1}{T}tr(V)$, T the sample size and A a positive definite $T \times T$ matrix indicating the spatial correlations between the residuals.

The Cluster-Robust Variance Estimator (CRVE Stata option for clustered data) used in Table 2 uses a similar formula with a normalization of the estimate depending on the degrees of liberty[10]. It applies whatever the number of covariates with a large number of clusters $g = 1\ to\ G$ and u_g the vector of residual inside cluster g:

$$V_{clusters}(\beta) = (X^{'}X)^{-1}\left(\frac{G}{G-1}\frac{N-1}{N-k}\sum_g X_g\hat{u_g}\hat{u_g^{'}}X_g^{'}\right)(X^{'}X)^{-1} \tag{6}$$

where k is the number of covariates X, G the number of clusters and N the size of the whole population.

Special cases

Greenwald's case

In the case of the variance component model (observations organized into J groups j with size m_j) with multiple explanatory variables indexed by k, the proportional bias in variances writes:

$$E(\hat{C}_{kk}) = \frac{\sigma^2}{T}\sum\nolimits_j \rho_j\rho_j(k)(m_j - 1) \tag{7}$$

with ρ_j the average correlations within groups of the errors and $\rho_j(k)$ of the explanatory variables. Therefore, if the groups within which the errors are correlated tend to be also relatively homogeneous in terms of the set of explanatory variables, the standard errors of the coefficients tend to be understated[11].

[10] This formula is attributed to Liang and Zegler (1986) by MacKinnon (2016).

[11] Note that in this formula the covariance matrix corresponding to an explanatory variable depends only on the spatial autocorrelation of that variable, and not on its relations with other covariates.

Moulton's formulas

Moulton applies in his empirical studies (1986, 1987, 1990) various simpler forms of these formulas and formulate tests of the biases. These correspond to the formula obtained by Scott and Holt (1982) for a two-stage sampling in the simple linear model and Greenwald formula (7) for clusters of equal size (variance of the clusters size $V(m_j) = 0$:

$$f(\beta) = 1 + [\frac{V(m_j)}{m_j} + m_j - 1]\rho_x\rho_\epsilon \tag{8}$$

with m the average clusters size and ρ_x the intra-correlation of the unique regressor:

$$\rho_x = \frac{\sum_i \sum_{j \neq k} (x_{ij} - \overline{x})(x_{ij} - \overline{x})}{v(x_{ij}) \sum_i m_i(m_i - 1)} \tag{9}$$

where the indices indicates individual j in group i.

Moulton analyses (section 3 of his 1990 paper) the model $y = X\beta + Z\delta + u$ where Z is a set of dummies indicating the groups. He proposes to apply a Breuch-Pagan test for the hypothesis that variances of the dummy coefficients $V(\delta_i) = \sigma_i$ are non-nul and uses the correct variance of the OLS estimate of equation (2) with $V = \sigma^2 I + \sum_i \sigma_i Z_i Z_i'$.

He concludes to an under-estimation of standard errors by 1.3 to 7.

His 1990 article writes the true covariance matrix of the OLS estimates corresponding to equation (2):

$$V(\hat{\beta}_{OLS}) = \sigma^2 (X^{'} X)^{-1} [1 + \rho(N - 1)] \tag{10}$$

where $N = X^{'} Z Z^{'} X (X^{'} X)^{-1}$ which simplifies as in the Kloek case to the formula of the true OLS variance-covariance matrix :

$$C = \sigma^2 (X^{'} X)^{-1} [1 + (m - 1)\rho] \tag{11}$$

with m the common size of all groups and ρ the common value of the correlation between residuals in each group.

These formulas show that the bias in variances increases with this spatial correlation and with the group size, which depend on the total number of observations in the data-set and on the number of groups.

Appendix B: Data-set

The French data-set from INSEE combines at the individual level the monetary and time expenditures into a common, unique goods and services consumption structure by a statistical match of the information contained in two surveys: the Family Expenditure Survey (FES, INSEE BDF 2001) and the Family Time Budget (FTB, INSEE BDT 1999). I define 8 types of activities or time use types compatible with the available data both from FES and BDT: Eating and

cooking time (FTB) and food consumption (FES), cleaning and home maintenance and dwelling expenditures (including imputed rent), clothing maintenance and clothing expenditures, education time and education expenditures, health care time and health expenditures, leisure time and leisure expenditures, transport time and transport expenditures, miscellaneous time use and miscellaneous goods and services. Time uses for all selected activities are regressed on the households' characteristics for all observation units in FTB survey and these estimations serve to predict the time spent on these activities for the corresponding units in the FES survey.

Appendix C: Results for other commodities

Table 3

Estimates of the income elasticity of Housing, Transport and Other expenditures (FES, France, 2000)

Data	Whole sample	Correction of the s.e.	Cluster-Robust Variance Estimator	Efficient sub-populations
Housing	0.8339	0.8339	0.8339	0.8347
s.e.	0.0077	0.0098	0.0158	0.0213
Transport	0.8143	0.8143	0.8143	0.8141
s.e.	0.0088	0.0138	0.0254	0.0208
Other expenditures	1.244	1.244	1.124	1.253
s.e.	0.0082	0.0100	0.0163	0.0303
Sample size	10251	10251	10251	2982
Method	OLS	Greenwald's equation (7)	Equation (6)	OLS

Note: data-set: French INSEE surveys: the Family Expenditure Survey (FES, INSEE BDF 2001) and the Family Time Budget (FTB, INSEE BDT 1999).

References

[1] Angrist, J. D., and Pischke, J-S, 2000, *Almost Harmless Econometrics*, Princeton University Press.

[2] Darnell, A. C., Evans, J. L., 1990, *The Limits of Econometrics*, Edward Elgar.

[3] Gardes, F., 2019, The Estimation of Price Elasticities and the Value of Time in a Domestic Production Framework: an Application Using French Micro-Data, *Annals of Economics and Statistics*, 135, September, 69-100.

[4] Greenwald, B. C., 1983, A General Analysis of the Bias in the Estimated Errors of Least Squares Coefficients, *Journal of Econometrics*, 22, August, 323-338.

[5] Hannan, M.T., Burnstein, L., 1974, Estimation from Grouped Observations, *American Sociological Review*, Vol.39, June, 374-392.

[6] Kloek, T., 1981, OLS Estimation in a Model Where a Microvariable Is Explained by Aggregates and Contemporaneous Disturbances Are Equicorrelated, *Econometrica*, 49, January, 205-207.

[7] Leamer, 1978, *Specification Searches: Ad Hoc Inference with Non Experimental Data*, Wiley.

[8] Liang, Zegler, 1986, Longitudinal Data Analysis Using Generalized Linear Models, *Biometrica*, 73(1), 13-22.

[9] McCloskey, D., 1985, The Loss Function Has Been Mislaid: The Rhetoric of Significance Tests, *American Economic Review*, vol. 75, 2, 201-205.

[10] MacKinnon, J. G., 2015, Wild Cluster Bootstrap Confidence Intervals, *L'Actualité Economique*, vol. 91, 1-2, Mars-Juin.

[11] Moulton, B. R., 1986, Random group Effects and the Precision of Regression Estimates, *Journal of Econometrics*, 32, August, 385-397., 5, April, 275-282.

[12] Moulton, B. R., 1987, Diagnostics for Group Effects in Regression Analysis, *Journal of Economic and Business Statistics*.

[13] Moulton, B. R., 1990, An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units, *Review of Economics and Statistics*, vol. 72, 2, May, 334-338.

[14] Nuzzo, R., 2018, La Malédiction de la Valeur-P, *Pour la Science Hors Série*, 98, Février-Mars, 34-39.

[15] Scott, A.J., Holt, D., 1982, The Effect of Two-Stages Sampling on Ordinary Least Squares Methods, *Journal of the American Association* 77, 848-854.

[16] Ziliak, Stephen, and McCloskey, Deirdre, (2008). *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor, University of Michigan Press, 2009.