Paper 441-2013

Short-term costs of smoking during pregnancy: Geometric multidimensional approach

Violeta Balinskaite, University of Bologna

ABSTRACT

Smoking during pregnancy imposes a considerable economic burden on society. This phenomenon has been studied fairly extensively in the United States, but little is known about its costs within the European Union. This paper aims to evaluate the neonatal costs of a mother who smokes during pregnancy compared to the alternative of her not smoking. Geometric multidimensional approach that is used for analysis involves the use of conditional multiple correspondence analysis as a tool for investigating the dependence relationship between covariates and the assignment-to-treatment indicator variable within a strategy whose final aim is to find balanced groups.

INTRODUCTION

Smoking in pregnancy is a major public health concern, posing risk to both mother and child. It is well-known cause of many complications of pregnancy, adverse fetal and infant outcomes, and suspected cause of some subtle and long-term outcomes in offspring. During the past decade, an increasing number of studies have been published focusing on the economic costs associated with smoking in pregnancy. The cost impact of smoking during pregnancy on both women and their children is an important measure for health care professionals making decisions about how to allocate health care resources.

The study reported in this paper focus as well on the economic costs associated with smoking in pregnancy. In our work, we deal with observational data, where selection process is complex and usually involving some combination of self-, administrator-, or other third-person-selection. In the last decades, some of the greatest minds in economics and statistics have studied the problem of self-selection with the resulting approaches being widely accepted and used as the best fixes, but these solutions to the bias that arise from self- or administrator-selection are imperfect, and many researches reserve their strongest causal inference for data from experimental rather than observational studies.

In this paper, we apply new geometric multidimensional approach: it is an algorithmic approach presented by Camillo and D'attoma. This analysis involves testing whether the information matrix X and the assignment-to-treatment indicator vector T are unrelated by subgroups, then collecting outcome for this units being compared.

METHODOLOGY

The idea is to summarize the multivariate difference in covariates across treatment groups in terms of between-group inertia and then testing it with a multivariate imbalance test. The method consists of three simple steps:

First Step: We compute global imbalance (GI) measure on the whole sample, and perform hypothesis test to evaluate its significance. The aim of measuring imbalance is to summarize the difference between the multivariate space of the pre-treatment covariates for the treated and the multivariate space of the pre-treatment covariates for the controls. In this step we want to see if the difference between simple comparison of outcome of treatment and control groups is biased by selection. To calculate the measure and test Global Imbalance, we use SAS macro program. Description of GI and SAS® marco program is defined in appendix.

Second Step: If imbalance exists, then we adopt so called *tandem approach*. It is a two-steps approach: first we use Multiple Correspondence Analysis (MCA) to obtain a low-dimensional representation of the X-space; second we apply cluster analysis to identify homogeneous groups on the basis of the low-dimensional MCA coordinates. To choose an appropriate number of clusters, we examine a tree diagram called dendrogram.

Third Step: We measure and test balance within each cluster, and compute local average treatment effect (ATE) within balanced groups (eliminating observations in unbalanced clusters). To a certain extent this step mirrors that of the step in a propensity score analysis where one identifies treatment and comparison group cases that are matched according to their assigned propensity score, and using those T-C pairs (or groups) to estimate a treatment effect. In this case, we use outcomes from resulting treatment and comparison cases that are assigned in each cluster to generate the treatment impact for that subgroup.

DATA

The data came from Emilia-Romagna administration office *Servizio Sistema Informativo Sanita e Politiche Sociali* database. The final dataset arises from a complex work of data manipulation of different types of data: purely administrative data (SDO) and administrative data based on questioner (CedAP). CedAP contains information about parents' socio-demographic characteristics, pregnancy, birth, and infant; SDO contains data of all hospital admissions occurred in public and accredited private hospitals in Emilia-Romagna region and outside. Our final data includes 15,473 observations of births between January and June in 2010. Of these, 2380 are observations where mother indicated that she was smoking during pregnancy and 13,093 are observations were mother indicated that she did not smoke.

We consider 53 categorical pre-treatment covariates and variable SMOKING as treatment indicator. Our outcome variable consists of the cost of birth and the cost of hospitalization during the first six months. Table 1 shows part of selected baseline traits separately for mothers who were smoking ("treatment" group), and those who did not smoke ("control" group). We can see that the treatment and control groups differ from one another, for example, a greater proportion of those who smokes are Italians (86.76% versus 69.36% in control group), are single (45.08% versus 24.33), and are occupied (73.36% versus 63.34%).

Table 1
Part of Selected Baseline Characteristics, by Treatment status

Baseline variable (%)	Chi-Square	p-value	Overall	Treatment	Control	Difference
Mother's age group :	0.3264	0.8494				
< 31			33.45	33.95	33.35	0.6
31-40			60.3	59.87	60.38	-0.51
> 40			6.26	6.18	6.27	-0.09
Mother's nationality:	479.596**					
Italian			72.04	86.76	69.36	17.4
Other EU citizen			4.09	6.47	3.66	2.81
Citizen of LMLIC*			14.21	2.52	16.34	-13.82
Others			9.66	4.24	10.65	-6.41
Mother's marital status:	524.885**					
single			27.52	45.08	24.33	20.75
married			65.11	47.27	68.36	-21.09
separated			1.29	2.77	1.02	1.75
divorced			0.93	1.34	0.86	0.48
widowed			0.14	0.25	0.12	0.13
not stated			5.0	3.28	5.32	-2.04
Mother's activity status:	220.353**					
occupied			64.88	73.36	63.34	10.02
unemployed			4.28	6.97	3.8	3.17
in search for first job			0.11	0.21	0.09	0.12
student			1.27	1.13	1.3	-0.17
housewife			23.77	15.34	25.3	-9.96
other			0.07	0.25	0.04	0.21
not available			5.61	2.73	6.13	-3.4
Previous conceptions (yes)			58.59	53.49	59.52	-6.03

NOTES: Low middle and low income countries according to UN. **p- value <0.0001.*

RESULTS

We started our analysis with computing the GI for this data set. As shown in Table 2, calculated GI value is 0.004 and it falls in the critical region. We can interpret it as the presence of

imbalance in data, and therefore demanding adjustment in order to estimate a treatment effect that is not biased by selection.

Table 2
Balance in the Overall Sample

Treatment	Control	GI	Interval	Balance
2380	13093	0.004	(0, 0.0004)	no
15.38%	84.62%			

The second step of our analysis is to use the cluster analysis on MCA coordinates to find groups of comparable units. MCA was carried out using all 53 pre-treatment covariates and the result of the MCA is a set of new variable (factorial coordinates) that are continuous and orthogonal to one another.

To identify homogeneous groups with new variables, we choose to use a hierarchical clustering method, and the Wald's method as group proximity measure. We most closely examined 10-, 42-, 44-, 69- and 72- cluster solution and have chosen the 42-cluster solution set because it discards fewer numbers of units (around 33.5%) with respect to other solutions. Table 3 presents the results of cluster analysis in terms of balance, including the proportion of treatment and control cases that each cluster includes and local effects. We accepted balance for nine clusters where GI is on the interval limits, and nice of the clusters result in having unbalance. These nine clusters represent about 33.5% of the observations (of which 60% are in clusters 1 and 30, short description in Table 6) of the original sample and we exclude them from our third analytic step.

Table 3
Balance and Effects, by Clusters

Cluster	Treatment (%)	Control (%)	GI	Interval	Balance	p-value	Local effect
1	13.04	86.96	0.008	(0, 0.003)	no	0.20	-
2	17.42	82.58	0.026	(0, 0.032)	yes	0.14	-667.23
3	8.33	91.67	0.031	(0, 0.054)	yes	0.30	-828.76
4	16.4	83.6	0.023	(0, 0.022)	yes	0.77	-301.74
5	16.08	83.92	0.013	(0, 0.01)	yes	<0.05	-1175.12
6	15.3	84.7	0.016	(0, 0.011)	no	<0.05	-
7	18.11	81.89	0.007	(0, 0.005)	yes	<0.05	-737.61
8	12	88	0.038	(0, 0.047)	yes	0.70	-175.61
9	14	86	0.013	(0, 0.014)	yes	0.45	-278.89
10	13.92	86.08	0.019	(0, 0.048)	yes	0.55	-229.69
11	18.14	81.86	0.016	(0, 0.02)	yes	0.97	-16.13
12	17.23	82.77	0.008	(0, 0.012)	yes	0.23	684.93
13	12.37	87.63	0.012	(0, 0.012)	yes	0.47	-109.67
14	12.37	87.63	0.006	(0, 0.005)	yes	0.66	113.25
15	17.34	82.66	0.009	(0, 0.012)	yes	0.66	151.11

16	19.61	80.39	0.011	(0, 0.015)	yes	0.27	-401.62
17	13.82	86.18	0.012	(0, 0.012)	yes	0.73	-81.77
18	20.19	79.81	0.005	(0, 0.005)	yes	0.95	15.8
19	20.18	79.82	0.007	(0, 0.007)	yes	<0.05	-324.85
20	17.51	82.49	0.003	(0, 0.003)	yes	<0.05	-736.65
21	11.54	88.46	0.066	(0, 0.07)	yes	0.22	4095.48
22	11.57	88.43	0.022	(0, 0.031)	yes	0.63	346.48
23	19.64	80.36	0.02	(0, 0.024)	yes	0.58	329.04
24	34.48	65.52	0.06	(0, 0.112)	yes	<0.05	-801.48
25	12.77	87.23	0.028	(0, 0.022)	no	0.59	-
26	8.57	91.43	0.037	(0, 0.058)	yes	<0.05	-971.74
27	10.57	89.43	0.02	(0, 0.018)	yes	0.22	-530.73
28	29.79	70.21	0.046	(0, 0.07)	yes	0.39	1356.83
29	22.32	77.68	0.012	(0, 0.013)	yes	0.89	-76.02
30	8.22	91.78	0.009	(0, 0.004)	no	0.23	-
31	3.85	96.15	0.013	(0, 0.03)	yes	<0.05	-555.31
32	6.05	93.95	0.015	(0, 0.011)	no	0.89	-
33	10.6	89.4	0.024	(0, 0.017)	no	0.50	-
34	18.25	81.75	0.023	(0, 0.029)	yes	0.64	-378.42
35	6.8	93.2	0.024	(0, 0.017)	no	0.44	-
36	5.41	94.59	0.047	(0, 0.031)	no	0.55	-
37	9.85	90.15	0.016	(0, 0.007)	no	0.18	-
38	19.17	80.83	0.019	(0, 0.029)	yes	0.28	-200.4
39	19.53	80.47	0.018	(0, 0.027)	yes	0.22	2094.47
40	29.76	70.24	0.005	(0, 0.005)	yes	0.12	-444.25
41	14.52	85.48	0.022	(0, 0.05)	yes	<0.05	-1193.82
42	15.61	84.39	0.007	(0, 0.008)	yes	0.55	211.57

In the last step, we compare the cost for mothers who were smoking with those who did not separately within each balanced clusters. We then compute the average causal effect (ATE). The results are shown in Table 4. It shows that there is no smoking impact on the cost and it is rather unexpected.

Table 4
Average Causal Effect

Method	Treated	Control	ATE
42-Clusters (Ward Method)	2380	13093	-249.03 (23.94)

We keep this result as our first step of analyzing the smoking impact on the cost using the data set of Emilia-Romagna region. In this analysis our used treatment group can be divided into 3 sub-groups: smoking during pregnancy, stopped smoking in the first months of pregnancy, stopped smoking in the first days of pregnancy. Table 5 presents the naïve estimator of the average causal effect of these three sub-groups. Our next step would be to compute ATE where treatment group consists of mothers who were smoking during pregnancy.

Table 5
Naïve Estimator of ATE

Stopped smoking in the first day	Stopped smoking in the first months	Smoking
-175.69	-431.22	47.91

CONCLUSION AND FUTURE RESEARCH

The main aim of this work has been to measure smoking impact on cost using new algorithmic approach which expunge selection bias from non-experimental data, and therefore facilitate estimation of unbiased treatment effects. Applied method consists of three steps: first identifying whether bias exists, then performing a cluster analysis on MCA coordinates, and in the last step comparing treatment and comparison cases within balanced clusters to estimate the average treatment effect. To identify if bias exist, we computed GI measure and performed hypothesis test to evaluate its significance. We used SAS Macro program created by F. Camillo and I. D'Attoma.

In our application, we excluded 33.5% of the observations because the selection bias has not been eliminated in few clusters. We got rather unexpected results, which show that there is no smoking impact on the cost. We keep this result as our first step of analysis and for our future work we will measure impact for more precisely defined treatment group.

REFERENCES

1. E. K. Adams, V. P. Miller and others, "Neonatal health care costs related to smoking during pregnancy", *Health Econ.*, **11** (2002), 193-206.
2. P. Arabie, L. Hubert, "Cluster Analysis in marketing research", in R.P. Bagozzi (Ed) *Advanced methods of marketing research*, Oxford,UK: Blackwell, 1994.
3. I. D'Attoma, F. Camillo, "A multivariate strategy to measure and test global imbalance in observational studies", *Expert Systems with Applications*, **38** (2011), 3451-3460.
4. I. D'Attoma, F. Camillo, "%GI SAS Macro: A SAS Macro for Measuring and Testing Global Imbalance of Covariates within Subgroups", Journal of Statistical Software, **51**(2012), 1-19.
5. B. Escofier, "Analyse des correspondances multiples conditionelle", in E. Diday (Ed), *Data Analysis and Informatics*, North Holland, Amsterdam: Elsevier Science, 1988.
6. J. Estadella, T. Ajula and S. Thiò-Henestrosa, "Distribution of the inter and intra inertia in conditional MCA", *Computational Statistics*, **20** (2005), 449-463.

7. C. Godfrey, K. E. Pickett, S. Parrott, N. D. Mdege and D. Eapen, "Estimating the costs to the NHS of smoking in pregnancy for pregnant women and infants", Project of the University of York 2011.
8. A. Hackshaw, C. Rodeck and S. Boniface, "Maternal smoking in pregnancy and birth defects: a systematic review based on 173 687 malformed cases and 11.7 million controls", Human Reproduction Update, July 2011.
9. D. P. Miller, K. F. Villa, S. L. Hogue and D. Sivapathasundaram, "Birth and first-year costs for mothers and infants attributable to maternal smoking", *Nicotine & Tobacco Research*, **3** (2001), 25-35.
10. S. L. Morgan, C. Winship, "Counterfactuals and Causal Inference. Methods and Principles for Social Research, Cambridge, 2007.
11. L. R. Peck, F. Camillo and I D'Attoma, "A promising new approach to eliminating selection bias", *The Canadian Journal of Program Evaluation*, **24** (2010), 31-56.
12. L. R. Peck, "Using Cluster Analysis in Program Evaluation", *Evaluation Review*, **29** (2005), 178-196.

ACKNOWLEDGMENTS

I would like to thank to my supervisor Prof. Furio Camillo for encouraging me to submit a paper to the SAS Global Forum 2013.

APPENDIX

Computing Global Imbalance and Its Significance Test

The Global Imbalance Measure

The GI measure is expressed as:

$$GI = \frac{1}{Q}\sum_{t}\sum_{j}\frac{b_{tj}^2}{k_t k_j} - 1$$

where:

Q denotes the number of pre-treatment covariates,
$t \in T$, where T denotes the number of treatment levels,
$j \in J_Q$, where J_Q is the set of all categories of the Q pre-treatment covariates,
b_{tj} is the number of units with category j in the treatment group t,
k_t is the size of group t,
k_j is the number of units with category j.

The GI measure is the result of using the Conditional Multiple Correspondence Analysis framework (Escofier, 1988) to quantify the between group inertia. When the dependence among

categorical baseline covariates (**X**) and the treatment assignment (**T**) is outside the control of researchers, displaying the relationship among them on a factorial space represents a first step for discovering the hidden relationship. In presence of dependence, any descriptive factorial analysis may unfold this link. Usually, the problem of the factorial decomposition of the variance related to the juxtaposition of the **X** matrix and **T** is faced within the MCA framework.
Referring to MCA, the structure of the data matrix eigenvectors and eigenvalues decomposition process, could be strongly influenced by the presence of an external conditional variable. Hence, a conditional analysis could be useful in order to isolate the part of the variability of the **X**-space due to the assignment mechanism.
Referring to Huygens' overall inertia decomposition of total inertia ($\mathbf{I}_{total}$) as within-groups ($\mathbf{I}_{within}$) and between-groups ($\mathbf{I}_{between}$), Conditional MCA consists in a factorial decomposition of the within-group inertia. In turn, conditional MCA could be also considered as an intra analysis since the inertia induced by the conditioning variable (**T**) is not taken into account.

The Imbalance Test

To determine the significance of the detected imbalance, we perform a multivariate imbalance test. The null hypothesis of no dependence among **X** and **T** is specified as:

$$H_0: \mathbf{I}_{within} = \mathbf{I}_{total}$$

On the basis of the asymptotic distribution function of $\mathbf{I}_{between}$ expressed as:

$$I_{between} \sim \frac{\chi^2_{(T-1)(J-1),\alpha}}{nQ}$$

the interval of plausible values for GI is defined as:

$$GI \in \left(0, \frac{\chi^2_{(T-1)(J-1),\alpha}}{nQ}\right)$$

with *n* as the sample size, *Q* as the number of baseline covariates and $\chi^2_{(T-1)(J-1)}$ as the chi-square value with *(T-1)(J-1)* degrees of freedom.
If the measured GI is outside the interval, then the null hypothesis of no dependence among **X** and **T** is rejected and it shows that, given all covariates involved in the selection mechanism, it does not exist dependence between the information matrix **X** and **T**, or if it exist, it is not statistically significant.

A SAS Macro for Measuring and Testing GI of Covariates withing Subgroups

```
%MACRO GI(library=,dsn=,out=,firstclu=,lastclu=,id=,group_var=,
         balance_var=,Q=, treat=,alpha=, multitreat=);
```

where

- `library` is the name of the directory in which information is located.
- `dsn` is the name of the SAS data set to be read. It must contain Q categorical covariates, the treatment indicator variable, the ID variable and the group membership variable. A group could be the result of any classification analysis conducted separately before running *%GI*.
- `out` is the name of the SAS output data set.
- `firstclu` is the number of the last group to analyze. It is a numeric value.
- `lastclu` is the number of the last group to analyze. It is a numeric value.
- `id` is the ID variable.
- `group_var` is the name of the variable that denotes the group membership.
- `balance_var` includes the name(s) of the baseline categorical variable(s) to be balance checked. The name(s) may be listed in any order and separated by blanks. The variable(s) must be numeric. No missing values are allowed.
- `Q` is the number of categorical variables on which simultaneously check imbalance.
- `treat` is the name of the treatment indicator variable. It must be a numeric value.
- `alpha` is the significance level to be used in testing GI.
- `multitreat` denotes the number of treatment levels.

The macro computes for each group GI measure using the SAS/IML language. First it counts treatment and control units for each group, and then creates a disjunctive table for each group. To compute GI measure, this macro program will create the following matrices which will be used within the IML procedure: **Z** that includes the Q baseline categorical covariates in disjunctive form, **L** that includes the t treatment level indicators, **B** that is the Burt table, the result of the inner product of the indicator matrix **Z**, the **Band** matrix that is a contingency table which crosses the categories of the baseline categorical covariates with each treatment level. Before quitting, the *%GI* macro deletes temporary datasets created during the implementation to avoid cluttering and errors in case the macro is invoked again.

Example of Macro Used for Data Preparation

```
options mstored sasmstore=data;
%macro intpat (dat1,dat2,b,var1,var2,var3,var4)/store;
data int1 (rename=(&var1=&b&&var1));
set &dat1;
keep chiavesdo &var1;
run;
proc sort data=int1 nodupkey;by chiavesdo;run;
proc transpose data=&dat1 out=int2(drop = _name_ _label_)
prefix=&b&&var2;
```

```
    by chiavesdo ;
    var &var2;
run;
proc transpose data=&dat1 out=int3(drop = _name_ _label_)
prefix=&b&&var3;
    by chiavesdo ;
    id &var2;
    var &var3;
run;
proc transpose data=&dat1 out=int4(drop = _name_ _label_)
prefix=&b&&var4;
    by chiavesdo ;
    id &var2;
    var &var4;
run;
proc sort data=int1; by chiavesdo;run;
proc sort data=int2; by chiavesdo;run;
proc sort data=int3;by chiavesdo;run;
proc sort data=int4; by chiavesdo;run;
data &dat2;
merge int1 int2 int3 int4;
by chiavesdo;
run;
%mend intpat;
```

Table 6
Part of Selected Baseline Characteristics, Cluster 1 and Cluster 30

	Custer 1		Custer 30	
	Treatment	Control	Treatment	Control
Size	232	1547	108	1314
Mother's age group (%):				
<31	17.65	13.80	15.59	25.72
31-40	48.24	40.07	15.59	17.07
>40	2.35	2.32	0.59	1.02
Father's age group (%):				
<31	2.65	1.85	7.94	8.25
31-40	28.53	17.76	16.76	26.41
>40	9.71	5.01	6.76	8.32
not available	27.35	31.57	0.29	0.84
Mother's marital status (%):	37.94	17.40	11.18	4.18
single	20.29	23.68	17.94	36.47
married	1.18	0.33	1.47	0.29
separated	0.29	0.40	0.29	0.25

divorced	0	0.04	0	0.07
widowed	8.53	14.35	0.88	2.54
not stated				
Mother's activity status	49.12	28.70	4.41	5.38
(%):	1.18	0.36	5.0	3.16
occupied	0	0.07	0	0
unemployed	0.88	0.07	0.29	0.87
in search for first job	0.59	0.36	22.06	34.36
student	0	0	0	0.04
housewife	16.47	26.63	0	0
other				
not available				

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Violeta Balinskaite
Work Phone: 00393895378638
E-mail: violeta.balinskaite2@unibo.it