Varying Coefficient Models Revisited: an Econometric View

Benini, Giacomo; Sperlich, Stefan Andréas; Theler, Raoul

How to cite

BENINI, Giacomo, SPERLICH, Stefan Andréas, THELER, Raoul. Varying Coefficient Models Revisited: an Econometric View. In: Proceedings of the conference ISNPS2014. [s.l.] : [s.n.], 2016. doi: 10.1007/978-3-319-41582-6_5

This publication URL: https://archive-ouverte.unige.ch/unige:83342
Publication DOI: 10.1007/978-3-319-41582-6_5

Varying Coefficient Models Revisited: an Econometric View

Giacomo Benini, Stefan Sperlich and Raoul Theler

Université de Genève, Bd du Pont d'Arve 40, CH - 1211 Genève
Geneva School for Economics and Management

January 9, 2016

Abstract

Disaggregated data are characterized by a high degree of diversity. Nonparametric models are often flexible enough to capture it but they are hardly interpretable. A semiparametric specification that models heterogeneity directly creates the preconditions to identify causal links. Certainly, the presence of endogenous variables can destroy the ability of the model to distinguish correlation from causality. Triangular varying coefficient models that consider the returns as non-random functions, and at the same time exogeneize the problematic regressors are able to add to the flexibility of a semiparametric specification the causal interpretability. Moreover, they make the necessary assumptions much more credible than they typically are in the standard linear models.[1]

1 The Causality Problem in the Presence of Heterogeneous Returns

Disentangling causality from correlation is one of the fundamental problems of data analysis. Every time the experimental methodology – typical in some hard sciences – is not applicable, it becomes almost impossible to separate causality from observed correlations using non-simulated data. The only available alternative is to find a set of non testable assumptions that allow to express the causal links as parameters or as functions, and to subsequently find consistent estimators for the conditional moments or distributions that describe the parameters (or functions) of interest. In particular, consider a response Y to be regressed on an explanatory variable W. The assumption that transforms a simple (cor)relation into a causal effect of W on Y, is often called 'exogeneity'.

Definition 1. *A variable W is weakly exogenous for the parameter of interest ψ, if and only if there exists a re-parametrization λ for the joint density with parameter $\lambda = (\lambda_1, \lambda_2)$ such that*

[1]We thank an anonymous referee and the participants of the ISNPS 2014 meeting in Cadiz for helpful comments and discussion.

1. $f(y, w|\lambda_1, \lambda_2) = f(y|w; \lambda_1) f(w|\lambda_2)$.

2. ψ *depends on* λ_1 *only.*

3. (λ_1, λ_2) *are variation free, i.e.:* $(\lambda_1, \lambda_2) \in (\Lambda_1 \times \Lambda_2)$ *for two given sets* Λ_1, Λ_2.

The factorization presented in Definition 1 implies that the conditional density of Y given W is fully characterized by λ_1, while λ_2 is a so-called nuisance parameter (Engle, Hendry and Richard, 1983). In other words, if the causal impact of W on Y is the objective of interest, then the characterization of the distribution of W is unimportant. This convenient factorization allows to focus exclusively on the relationship between Y and W ignoring all the other associations.

In econometrics, an outcome equation that describes the relationship between Y and W often has a less restrictive moment specification than the one proposed by this definition. Usually, a factorization in the form of

$$E[YW|\lambda_1, \lambda_2] = E[E(Y|W; \lambda_1)W|\lambda_1, \lambda_2] \ , \tag{1}$$

is sufficient to detect the causal impact of W on Y. The problem is that, even for simple economic situations, it is often hard to justify an assumption like (1).

Consider for example the case where an economist wants to study the demand function of soft drinks using the individual consumption of Coca-Cola (X), the individual consumption of Pepsi-Cola (Q) and their respective prices (p_1, p_2) (with $p_1 > p_2$). A typical dataset looks like the one in Figure 1.

Observations	Coca-Cola	Pepsi-Cola
1	0	Q_1^*
2	0	Q_2^*
3	X_3^*	Q_3^*
4	X_4^*	Q_4^*
$\vdots$	$\vdots$	$\vdots$

Figure 1: X^* is the consumption of Coca, while Q^* is the consumption of Pepsi.

From the observation of the data, an econometrician would conjecture that, while the first two cross-section observations (i.e. individuals) may consider Coca-Cola and Pepsi-Cola as perfect substitutes, and therefore, since $p_1 > p_2$, all the income spent on soft drinks goes to Pepsi-Cola, the individuals 3 and 4 prefer to consume a quantity of Coca-Cola X^* different from zero, even though the price of Coca-Cola is higher (see Figure 2).

In other words, since Agent 1 and Agent 3 have different preferences, their optimization process

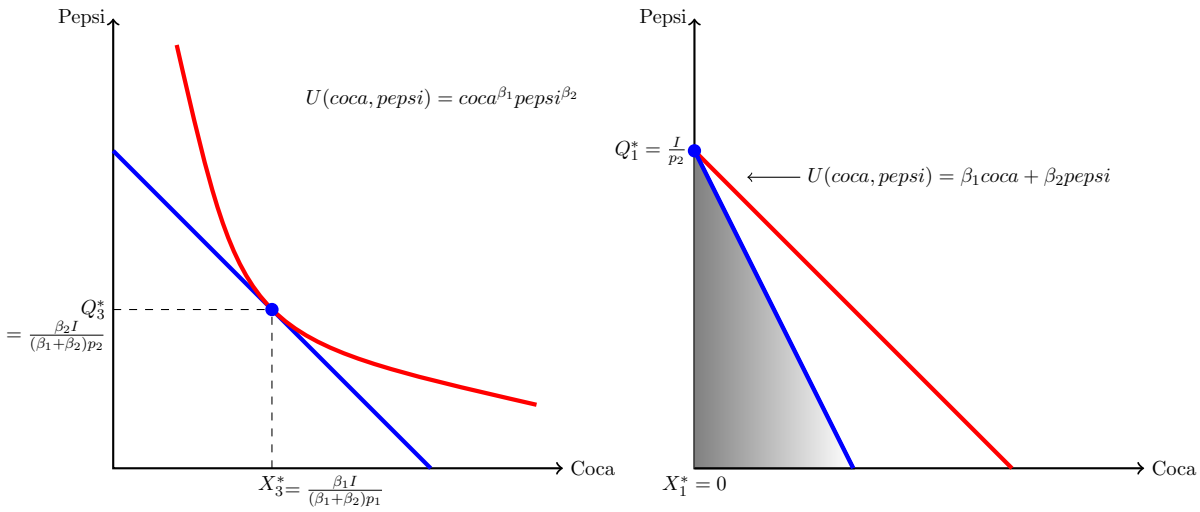

Figure 2: The (individual) demand functions for a given budget constraint $p_1X + p_2Q = I$ varies accordingly to individual preferences. Agent 3 and 4 do not consider Coca-Cola and Pepsi to be equally good (left graph), while Agent 1 and 2 do (right graph).

is different

Agent 1 max process

$$\begin{cases} \max\limits_{X,Q} U(X,Q) = \beta_1 X + \beta_2 Q \\ \text{s.t.} I = p_1 X + p_2 Q \leq I_{ub} \end{cases}$$

Agent 3 max process

$$\begin{cases} \max\limits_{X,Q} U(X,Q) = X^{\beta_1} Q^{\beta_2} \\ \text{s.t.} I = p_1 X + p_2 Q \leq I_{ub} \end{cases}$$

with I_{ub} being the budget constraint. In order to check whether the previous conjecture is true or not, a structural model that enables us to empirically validate the hypothesized choice structure must be specified. If the utility functions are not cardinal, the results of the two maximization processes cannot be compared directly. To the contrary, the study of the expenditure functions allows to monetize the otherwise incommensurable trade-offs between the benefits of the consumptions and their costs. In particular, an expenditure function indicates the minimum amount of money that an individual would need to spend in order to achieve a certain level of utility (given an utility function and a set of prices). If the conjectured choice models are correct, then for those agents that consider Coca-Cola and Pepsi-Cola as perfect substitutes (like Agent 1), the expenditure function should be

$$I(p_1, p_2, \bar{v}(X,Q)) = min\left(p_1 \frac{\beta_1 X + \beta_2 Q}{\beta_1}, p_2 \frac{\beta_1 X + \beta_2 Q}{\beta_2} \right) ,$$

where $\bar{v}(X,Q)$ is the level of utility for the observed consumption levels $(0,Q)$. For individuals that do not consider the two soft drinks as perfect substitutes (like Agent 3), the amount of expenditures for the given level (X,Q) should be

$$I(p_1, p_2, \bar{v}(X,Q)) = \left[p_1 \left(\frac{\beta_1 p_2}{\beta_2 p_1} \right)^{\frac{\beta_2}{\beta_1+\beta_2}} + p_2 \left(\frac{\beta_2 p_1}{\beta_1 p_2} \right)^{\frac{\beta_1}{\beta_1+\beta_2}} \right] X^{\frac{\beta_1}{\beta_1+\beta_2}} Q^{\frac{\beta_2}{\beta_1+\beta_2}} .$$

In this case both Definition 1 and/or assumption (1) are useless, because the required factorization for the vector $W = [X, Q]^T$, given a set of parameters $\lambda_1 = [\beta_1, \beta_2]$, can be true for the perfect substitute case or for the imperfect one, but not for both.

This simple introductory example shows that when micro-data exhibit holes (non participation in the activity of interest), kinks (switching behaviors) and corners (non-consumption or non-participation at specific points in time), then relations like (1) become meaningless (Pudney, 1989). There are at least three solutions to deal with an assumption like (1) in a context where heterogeneity among individuals is a major concern.

A first solution is to aggregate the data and study a much smoother problem (being smoother due to the compensations of the movements in opposite directions) typical for macro-data. Consider, for example, a relation between two variables which at a micro level may be piecewise linear with many nodes. After the aggregation is done, the relationship can probably be approximated by a smooth function that can satisfy equation (1) (Eilers and Marx, 1996). However, if an econometrician is interested in the analysis of individual-level data, in order to describe the economic behavior of individuals or firms, this option does not help.

A second possibility is to accept the heterogeneous nature of the parameters at a micro-level, but to ignore it, and use a parametric (possibly linear) specification with constant coefficients. Let us now abstract from the above example and denote the response by Y and the two explanatory variables X and Q such that

$$Y_i = t(X_i, Q_i) + e_i = \beta_0 + \beta_1 X_i + \beta_2 Q_i + \varepsilon_i \qquad E[\varepsilon_i | X_i, Q_i] = 0 \ . \tag{2}$$

In this case all the heterogeneity is absorbed by the regression disturbance ε. Even if many applied economists recognize the limits of a standard parametric specification that most likely suffers from a functional form misspecification because $t(X_i, Q_i) \neq \beta_0 + \beta_1 X_i + \beta_2 Q_i$, which means that $e \neq \epsilon$, they still use it as an approximation because their least squares estimates (like OLS) converge to the value of β that tries to minimize (we say *try* because its success depends also on other factors like the scedasticity function) the mean-squared prediction error $E[t(X_i, Q_i) - \beta_0 - \beta_1 X_i - \beta_2 Q_i]^2$. It is well known that under homoscedasticity OLS gives the best linear predictor of the non-linear regression function (the mean-squared error (MSE) being the loss function), and this even when there is a functional form misspecification (White, 1980b). However, this property is not useful if the objective of the researcher is to interpret the regression coefficients as a true micro-relationship in the form of $E[Y_i | X_i, Q_i]$, because the standard OLS would typically be inconsistent when estimating the marginal effect of the variables,

$$\hat{\beta}_1^{\,OLS} \to \beta_1 \neq \frac{\partial t(X_i, Q_i)}{\partial X_i} =: \beta_{1i} \qquad \hat{\beta}_2^{\,OLS} \to \beta_2 \neq \frac{\partial t(X_i, Q_i)}{\partial Q_i} =: \beta_{2i}.$$

In particular, if the returns are heterogeneous in the data generating process (DGP), a modeling strategy like (2) might not be able to derive consistent estimates. For example, if $Y_i = \beta_i^T W_i + \varepsilon_i$ with $W = [1, X, Q]^T$, for $\beta_i = [\beta_{0i}, \beta_{1i}, \beta_{2i}]$ is modeled as

$$Y_i = \beta^T W_i + e_i \quad \text{with} \quad \beta = E[\beta_i] \quad \text{and} \quad e_i = W_i^T [\beta_i - \beta] + \varepsilon_i \ ,$$

then the standard OLS estimators would give

$$\begin{aligned}\hat{\beta}^{OLS} &= \left[\sum_{i=1}^{n} W_i W_i^T\right]^{-1} \sum_{i=1}^{n} W_i Y_i = \left[\sum_{i=1}^{n} W_i W_i^T\right]^{-1} \sum_{i=1}^{n} W_i \left[W_i^T E(\beta_i) + e_i\right] \\ &\xrightarrow[n\to\infty]{p} E(\beta_i) + E(W_i W_i)^{-1} E(W_i W_i^T[\beta_i - E(\beta_i)]) + E(W_i W_i^T)^{-1} E(W_i \varepsilon_i) \\ &= E(W_i W_i^T)^{-1} E(W_i W_i^T \beta_i).\end{aligned}$$

From the last equality it follows that $\hat{\beta}^{OLS} \nrightarrow E[\beta_i|W_i]$ unless $E[\beta_i|W_i] = E[\beta_i]$.

A third solution is to conclude that the discreteness and non-linearity typical for micro-data requires to model heterogeneity directly. But how? A first option is to transform the density requirement of Definition 1 into an individual level factorization like

$$f(y_i, w_i|\lambda_{1i}, \lambda_{2i}) = f(y_i|w_i; \lambda_{1i}) f(w_i|\lambda_{2i}) \ , \quad i = 1, ..., n \tag{3}$$

(here w_i needs not to include a 1 for the intercept), where every cross-sectional observation is characterized by a set of individual parameters $(\lambda_{1i}, \lambda_{2i})$. This creates the complication that the parameters (λ_1, λ_2) are no longer variation free, which is not *stricto sensu* a problem because it is possible to transform λ_{1i} into a random coefficient to which it is possible to associate an invariant hyper-parameter θ that characterizes the prior density $f(\lambda_{1i}|w_i, \theta)$. In this specification, the invariance assumption can be reproduced in the form $f(y_i|w_i, g(w_i, \theta))$, where θ is estimated globally by a maximum likelihood or in a neighborhood, e.g. by a kernel-based local likelihood. This Bayesian solution allows to have variation-free hyper-parameters and, at the same time, random coefficients that capture individual heterogeneity due to the randomness of λ_{1i}.

No matter how elegant the solution might look like, it presents many and interdependent problems. The main one is the low degree of robustness of the estimates $\hat{\theta}$. One may use shrinking priors to overcome this, but in order to make sure that the prior decays quickly enough (to produce robust estimates), it is necessary to impose stringent conditions both on the priors' tails and on the decay rates of the tails. This kind of assumptions are very hard to understand in practice and even harder to relate to economic theory.

A less controversial way to directly model heterogeneity is to allow the value of the coefficients to change when and observable variable F, called here 'effect modifier(s)', allows to write equation (2) as

$$Y_i = \beta_i^T W_i + \varepsilon_i \quad \text{with} \quad \beta_i = g(F_i) + \delta_i \ . \tag{4}$$

This is the well-known varying coefficient model (VCM), cf. Hastie and Tibshirani (1993). In this specification Y_i is the dependent variable, W_i is a $d_W \times 1$ vector of explanatory variables, and the coefficient β_i is allowed to vary across i. In particular, it is a function of a $d_F \times 1$ vector of observable variables F_i (which might include also elements of W), while $g(.)$ is a vector of functions of the effect modifier, and δ_i is a stochastic mean-zero disturbance with finite variance. The exogeneity assumption is centered on the idea of correctly estimating the causal impact of W, not the one of F, on Y, therefore it is possible to imagine $\hat{g}$ as the best

nonparametric predictor of β_i for a given F_i. This implicates that the expected value of δ given Q would be equal to zero by construction: $E[\delta_i|F_i] = 0$. The new structure of the model produces a very flexible and yet interpretable semiparametric specification.

The hybrid nature of the VCMs has several advantages. Firstly, it reduces the level of complexity of a pure nonparametric model allowing to interpret the coefficients like in a parametric specification. Secondly, it enables to incorporate insights that come from economic theory into the modeling process. Thirdly, it produces a good trade-off between the loss in fitting ability, which is (hopefully) small compared to the nonparametric specification, and the increased facility of the estimation process, which is almost as easy as in a parametric model.

The empirical potentials of the VCM modeling can be understood re-considering the soft drink example. In this case, depending on whether the agent considers the goods as perfect substitutes or not, the coefficients resulting from the optimal allocations are different. However, in both cases, they are functions of the level of expenditure I, the prices (p_1, p_2) and the quantity consumed (X, Q) by individuals with some characteristics also included in F.

The previous consideration suggests that a VCM, in which the returns are functions of the prices and of the quantities of the goods, allows to keep a linear specification for the expenditure function (or expenditure shares) in the form of

$$Y_i = \beta_{0i} + \beta_{1i}X_i + \beta_{2i}Q_i + \varepsilon_i \tag{5}$$

with $\beta_j = g_j(F_i) + \delta_{ji}$, $j = 0, 1, 2$. In other words, a VCM allows us to transform the structural specification of (2) into a model able to take into account heterogeneity *sive natura*, making an assumption like (1) meaningful and often also plausible. Of course, the presence of numerous effect modifiers makes an equation like (5) hard to compute. To the contrary, a function with few effect modifiers is more easily interpretable and, at the same time, reduces the course of dimensionality of the nonparametric regression $\beta_j = g_j(F_i) + \delta_{ji}$, $j = 0, 1, 2$. Therefore it makes sense to reduce the number of effect modifiers for each j separately (e.g. by canceling those that present a low level of non-linearity with respect to the regressors).

The introduction of a second, more complex, economic example helps explaining the potentials of a VCM, even when the conjectures about the individual-decision making process behind the observed covariates is less easy to deduce than in a simple demand analysis environment. Let's suppose that an applied economist wants to study the impact of education and experiences on wages in a cross-sectional data set. The concerns about the disaggregated nature of the data might induce the researcher to do an a priori analysis that most likely reveals that marginal returns to education vary for different levels of working experience, see e.g. Schultz (2003). Merging the insights that come from the economic theory with the intuitions resulting from the scrutiny of the data we end up with a VCM of the form

$$wage_i = \beta_{0i} + \beta_{1i}educ_i + \varepsilon_i \ , \tag{6}$$

where the intercept and the slope are functions of the level of experience, $\beta_i = g(exp_i) + \delta_i$, with $g(\cdot)$ and δ belonging to $\mathbb{R}^2$. A structural specification like (6) is very appealing because

it corrects the (downward) bias that would emerge using a linear modeling that ignores the interaction between experience and education and therefore systematically underestimates the returns on schooling (Card, 2001). In this new formulation, it is important to discus the role of δ. As indicated above, the nature of δ is not the one of an isotonic deviation from the mean but rather the one of a stochastic disturbance in a nonparametric equation. Therefore the role that δ plays in equation (4) is related to its disturbance-nature.

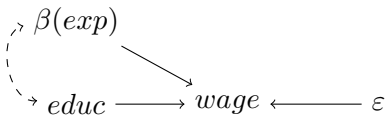

Figure 3: A causal graph of (6) highlighting the causal links among the variables.

Unlike in the soft drinks example, where the choices' structure was easy to *reverse-engineer*,[2] the relationships among the three variables $(wage, educ, exp)$ are more complex. The lack of knowledge about the objectives that individuals have, and the ignorance about the means that they may use to achieve them, does not allow to have an insight in the choice structure based uniquely on the observed covariates. In other words, in the analysis of the wage-education-experience relationship, even assuming that the only objective of (all) individuals is to obtain a high salary, there is no perfect insight about which actions an individual would take in order to achieve this result. For example, in order to have a higher wage, agent i could start working immediately after high school and accumulate experience, which is valued in the labor market. In this case the decision to go to university would be postponed. At the same time, agent j could do the opposite having the same objective function. This means that it is highly probable that the non-linear nature of the discrete data that describe the individual choices can be largely absorbed by $g(exp_i)educ_i$, but there could still be a local deviation form the mean of the level of education, here denoted as $\delta_i educ_i$, due to the uncertainty about the individual decision making process. This is reflected in the causal graph given in Figure 3.

Since each coefficient is not an average given a particular sample realization (but a function), the parameters are allowed to have different degrees of heterogeneity even for the same levels of the effect modifier, reflected in the presence of δ. So, here heterogeneity can explicitly imply deviations from the slopes. In the semiparametric framework, the quantity of local deviation δ is a function of the degree of smoothness of $g(exp_i)$. At the same time, since the primary objective of the research is not to estimate correctly the causal impact of F on Y but rather the one of W, it is sufficient to think of $\hat{g}(\cdot)$ as the best nonparametric predictor of β_i *such that* $E[\delta_i|F_i] = 0$ becomes true by construction. As a result, the average returns to education are equal for all the cross-section observations that have the same level of exp.

[2]Reverse engineering, also called *back engineering*, is the process of extracting knowledge or design information from anything man-made, and re-producing it. In economics, the reverse engineering process consists of extracting the structure of individual preferences from observed outcomes and then reproduce the outcomes using the conjectured informations.

2 Triangular Varying Coefficient Models with Instruments

The previous section highlighted the necessity to model heterogeneity directly in order to make the assumptions of Definition 1 plausible. But still, exogeneity can be violated by the nature of the observed variables irrespectively of the semiparametric characteristics of the VCM. In particular, a regressors could be endogenous in the sense that in the structural equation $Y_i = \beta_i^T W_i + \varepsilon_i$ one has $E[\varepsilon_i|W_i, F_i] \neq 0$. The three usual sources of endogeneity typically mentioned are: the omission of explanatory variables correlated with the included covariates, a measurement error and reversed causality. All the three sources of endogeneity cannot be solved using the varying coefficient approach alone.[3] A popular solution is to introduce some additional variables called instruments.

Definition 2. *A variable Z is called an instrumental variable (IV) for W if*

1. *is partially correlated with the endogenous variable W once the other explanatory variables have been netted out.*
2. *is mean-independent with respect to the stochastic error ε.*

This definition suggests that the addition of a second structural equation to the VCM creates a triangular model able to exogenize W while modeling heterogeneity directly. For simplification, let us set for a moment $dim(W) = dim(Z) = 1$. Keeping a specification like (4), it is sufficient to add a selection equation that relates the endogenous W with the instrument(s) Z, namely

$$W_i = m(Z_i) + \eta_i \qquad\qquad E[\eta_i|Z_i] = 0 \tag{7}$$

and assume a finite variance for η_i. In this formulation the vector of explanatory variables is allowed to contain endogenous components, while Z is a vector of IVs, which may have F as one of its arguments. Furthermore, $m(\cdot)$ is a smooth function, or a vector of smooth functions if $dim(W) > 1$, while ε and η are, respectively, the endogenous error and a stochastic disturbance that has expected value equal to zero and finite variance.

The triangular nature of equations (4) and (7) implies a simple endogeneity mechanism. In order for the error term ε to be correlated with at least one of the explanatory variables W, it must be $cov(\eta, \varepsilon) \neq 0$. To see how the mechanism of the model works in practice, it is useful to consider the simplest possible specification, namely a model that would include only one heterogeneous intercept, one heterogeneous slope and one endogenous explanatory variable. The latter is instrumented by one exogenous variable Z_1, which is correlated with W even if the impact of the (exogenous) effect modifier has been netted out, namely

$$\begin{aligned} Y_i &= \beta_{0i} + \beta_{1i} W_i + \varepsilon_i & E[\varepsilon_i|F_i, W_i] &\neq 0 \\ W_i &= m(F_i, Z_{1i}) + \eta_i & E[\eta_i|F_i, Z_{1i}] &= 0 \ . \end{aligned}$$

In this specification, irrespectively of the relation between the error ε and the two disturbances δ and η, endogeneity comes only through $cov(\varepsilon, \eta)$, see causal graph 4.

[3] However, the most typical, though in economics rarely mentioned, endogeneity problem, i.e. the functional misspecification, can be largely diminished by the VCM.

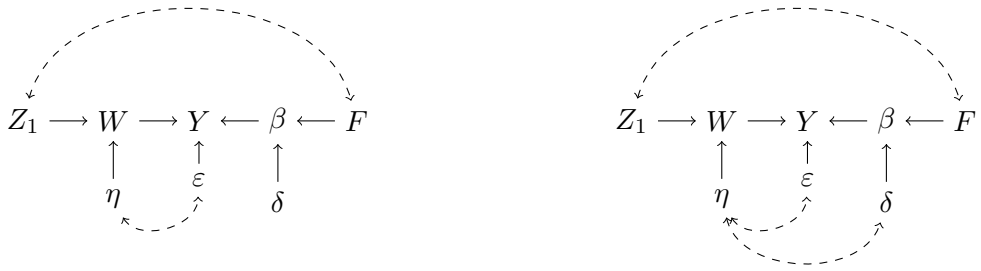

Figure 4: The mechanism of the endogeneity process changes depending on the assumptions about the relationship between the error ε and the stochastic disturbances (η, δ). The left picture is the only possibility in a world of homogeneous coefficients, while the right specification (with $cov(\eta, \delta) \neq 0$) is the situation resulting from introducing a varying coefficient structure. The direct connection between δ and ε is not taken into account because the interest is about the causal link between Y and W for a given level of F.

The considerations about the mechanisms of the endogeneity problem combined with the observation that a VCM is a special case of a semiparametric linear specification, suggest that the model can be identified and later estimated using the *control function* approach (Tesler, 1964). The control function, say $h(\cdot)$, handles the relation between η and ε (irrespectively of the behavior of δ) in the following form

$$\epsilon_i = \delta_i W_i + \varepsilon_i = h(\eta_i) + \vartheta_i \qquad\qquad E[\vartheta_i | \eta_i] = 0 \ . \tag{8}$$

This added to (4) eliminates the endogeneity problem giving unbiased estimates for $g(.)$:

$$E[Y_i | Z_i, \eta_i] = g(F_i) W_i + h(\eta_i) \qquad Z_i = (F_i, Z_{1i}) \ . \tag{9}$$

It is important to notice that the higher complexity of a VCM increases the chance to successfully eliminate the endogeneity problem via the control function approach. Specifically, even if a set of valid instruments $(Z_i)_{i=1}^n$ is available, a linear IV estimator would generally be biased. For example, if the equation $Y_i = \beta_{01} + \beta_{1i} W_i + \varepsilon_i$ (with $\varepsilon_i \perp\!\!\!\perp Z_i$) is modeled using homogeneous coefficients $Y_i = \beta_0 + \beta_1 W_i + e_i$ with $e_i = [\beta_{0i} - \beta_0] + W_i[\beta_{1i} - \beta_1] + \varepsilon_i$ and $\beta_j = E(\beta_{ji})$, $j = 0, 1$, then the instrumentation using Z_i does not produce consistent estimates. Consider for example the case where $dim(Z) = dim(W) \geq 1$. In this setting the estimated returns are

$$\begin{aligned}
\hat{\beta}^{IV} &= \left[\sum_{i=1}^n Z_i W_i^T \right]^{-1} \sum_{i=1}^n Z_i Y_i = \left[\sum_{i=1}^n Z_i W_i^T \right]^{-1} \sum_{i=1}^n Z_i \left[W_i^T E(\beta_i) + e_i \right] \\
&\xrightarrow[n \to \infty]{p} E(\beta_i) + E(Z_i W_i^T)^{-1} E(Z_i W_i^T [\beta_i - E(\beta_i)]) + E(Z_i W_i^T)^{-1} E(Z_i \varepsilon_i) \\
&= E(Z_i W_i^T)^{-1} E(Z_i W_i^T \beta_i) \ .
\end{aligned}$$

The last equality cannot be simplified further unless a new assumption, namely $\beta_i \perp\!\!\!\perp (W_i, Z_i)$, is made - which is clearly in contradiction with the spirit of the model, cf. the causal graphs in Figure 4. Basically, the heterogeneous nature of the returns transforms Z_1 into a 'poor' instrument if the simple linear structure is used.

In order to proceed and correctly estimate the unknown terms in equation (9), it is necessary to impose additional identification conditions. Identification can be obtained imposing a conditional mean independence in the form of

$$E[\epsilon_i|Z_i, \eta_i] = E[\epsilon_i|\eta_i] \qquad \text{CMI}, \tag{10}$$

or a conditional moment restriction

$$E[\epsilon_i|Z_i] = 0 \qquad \text{CMR}. \tag{11}$$

The CMI and the CMR are not equivalent (Kim and Petrin, 2013). The CMI requires Z and η to be additively separable in W, which often is not the case. To the contrary, the CMR can be easily justified by the use of economic primitives that describe the structural specification (Benini and Sperlich, 2016). The use of the CMR, however, requires to include the instrument(s) in the control function, such that the relation between ϵ and η becomes

$$\epsilon_i = h(Z_i, \eta_i) + \vartheta_i \quad E[\vartheta_i|Z_i, \eta_i] = 0 \ . \tag{12}$$

In any case, if the amplitude of the control function increases, a less precise estimate $\hat{g}(.)$ might be produced (multi-functionality). This is the statistical counterpart of the econometric problem called 'weak' instruments, i.e. instruments that are weakly correlated with the endogenous regressors.

The estimation of VCM in its simplest specification has been proposed in different forms. Hastie and Tibshirani (1993) used a smoothing spline based on a penalized least squares minimization, while Fan and Zhang (2008) proposed a kernel weighted polynomials. However, this last method and its surrogates are designed for a single effect modifier for all coefficients, which is a strong limitation in the context we discussed so far.

Estimating an equation like (9) is a more complicated procedure than the one required for a simple VCM. The presence of a control function, which depends upon η, requires the use of specific tools that are designed for additive models. The two most common alternatives are the marginal integration method (Linton and Nielsen, 1995) and the smooth back-fitting (Roca-Pardinas and Sperlich, 2010). The latter method suffers less from the curse of dimensionality and can be applied as part of a 2-steps procedure. The first step consists in the estimation of $m(Z)$ in equation (7) using a standard nonparametric technique. The second step consists in the substitution of the estimated residuals $\hat{\eta}$ into (9), which creates an equation characterized by a finite sample disturbance whose impact can be mitigated asymptotically (Sperlich, 2009). For an exhaustive survey on the VCM estimation techniques see Park, Mammen, Lee and Lee (2013). For a comparison of implementations of these methods in R (R Core Team, 2014), including the control function approach, see Sperlich and Theler (2015).

All the previous considerations are particularly important in the treatment effect literature. For a discrete W with finite support, Imbens and Angrist (1994,1995) named the impact of a treatment (i.e. a change in W) local average treatment effect (LATE). By construction, the LATE can only compute the average of the β_i for the individuals that choose to switch their w

because of an instrument's change. In other words, in the LATE environment, the parameter of interest can only be estimated for people responding to the selection equation and is therefore an instrument (or selection) specific parameter. They imposed a conditional independence assumption in the form of $Y_i(w) \perp\!\!\!\perp Z_i \ \forall w$, as well as the request of independence of the (so called) compliers sub-population to an instrument's change. Reconsidering these assumptions in the presence of heterogeneous returns shows that the LATE is not defined if $cov(\beta, Z) \neq 0$. In the case of a VCM this means that, unless the effect modifier is indeed a constant, the standard independence assumption used to define and identify the LATE is not fulfilled.

The model we outlined above suggests that, if some effect modifiers F_i are observed, they should be used to construct a VCM that makes the LATE conditions more credible. For example, in the case of a binary endogenous treatment W which is instrumented by a binary instrument Z, the varying LATE becomes

$$LATE(q) = \frac{E[Y|F = q, Z = 1] - E[Y|F = q, Z = 0]}{E[W|F = q, Z = 1] - E[W|F = q, Z = 0]} \ .$$

Integrating over q gives the value of the LATE. In this case, the more heterogeneity of returns to W is captured by $g(F)$ the less the LATE will vary over the IVs' choice. In other words, a VCM reduces the typical LATE problem to a minimum because it controls for the correlation between the effect modifier and the instrument. Therefore, the VCM enables to identify a LATE-type parameter that can be estimated nonparametrically regressing Y and W on F and Z. The interesting point here is that the parameter of interest depends on both, the instruments' choice and the values taken by F. An interesting next step would be to find a meaningful model specification that merges the effect modifier and the instrument.

3 An Example

In order to see all the potentials of the triangular VCM specification in practice, it is useful to reconsider the wages-experience-education relationship. Experience and education are crucial variables in the determination of a worker's wage. Yet, labor economists have argued for many years that cognitive and non-cognitive abilities are also critical in order to determine labor market outcomes. A large empirical literature has confirmed the positive connection between cognitive test scores and high wages (Murnane, Willett and Levy, 1995). Unfortunately, many datasets do not provide ability's measures. The lack of information about the skills misleads the researcher to mistake the data generating process (DGP). Even if a VCM modeling strategy is used, if the ability of the individual is not included, such that

$$wage_i = t(educ_i, exp_i, ability_i) + \zeta_i \quad \text{is modeled as} \quad wage_i = g_0(exp_i) + g_1(exp_i)educ_i + \epsilon_i,$$

then the exogeneity assumption $E[\varepsilon_i|educ_i, exp_i] = 0$ does not hold, because of an omitted variable bias. This problem can be solved using an instrument.

There exist at least two classical errors that arise when searching for an IV. The first one is the selection of a variable that is clearly correlated with the endogenous regressor but hardly

independent from the error ε. For example, the level of education (of one) of the parents would be hardly independent from the omitted variable *ability*. A second wrong choice would be the selection of a variable that has the opposite characteristics, namely a variable that is exogenous but that is hardly correlated with the endogenous regressor. For example, the last digit of the person social security number. The choice of good instruments must come from both, a deep knowledge of the origin of the IV and the source of endogeneity.

Take instead the example proposed by Angrist and Krueger (1991). In most American states education legislation requires students to enter school in the calender year when they turn six. Therefore the age at which students start school is a function of the date of birth of the pupils. For example, if the 31st of December is the legal cut-off point, children born in the fourth quarter enter school shortly before turning six, while those born in the first quarter enter school when they are around six years and an half. Furthermore, because compulsory schooling laws require students to remain in school only till they turn 16, these groups of students will be in different grades, or through a given grade to a different degree, when they reach the legal drop out age. The combination of the school start-age policies and the school attendance laws creates a situation where children attend school for different times depending upon their birthdays. Assuming that the day of birth of a person is not correlated with his abilities seems to make the quarter of birth (*qob*) a valid IV. The typical mistake made here is to conclude from no-causality to no-correlation. But firstly, there is clearly the possibility that the IV is correlated with the education of the parents, and secondly, being the youngest could mean to be the smallest and physically weakest in the class resulting in maturity disadvantages. All these facts could change the wage-path invalidating the IV.

Nonetheless, let us consider a VC triangular model with the same instrument proposed by Angrist and Krueger

$$wage_i = g_0(exp_i) + g_1(exp_i)educ_i + h(\eta_i) + \vartheta_i \tag{13}$$

$$educ_i = m(exp_i, qob_i) + \eta_i \ . \tag{14}$$

In this specification they identify the LATE of education on wages for those who do not drop out in spite of the ones that could have thanks to their birth date. Note that, if this is not the parameter of interest, it might have been much better and easier to use a proxy approach instead of an IV one. In order to reverse-engineer the preferences' structure it is necessary to model a situation where a rational individual has to decide, when turning 16, to stay for the rest of the academic year or leave school. In this context, the agent's wage is a function of the years of education *educ*, but also of his unobserved ability, ε. The agent's ability is not observed, but the information set that the student can consult before the decision to stay or not is made includes a signal of his individual ability η, for example his past grades. The cost to stay until the end of the year is a function of an exogenous cost shifter, namely the quoter of birth *qob*, if a student turns 16 in January the cost to stay till the end of the year is higher than if he turns 16 in May, so it makes sense to consider the quoter of birth an argument of the cost function. At the same time, the agent's utility has to be function of the education's choice, the cost-shifters and the unobserved ability, $U(educ, qob, \varepsilon) = p(educ, \varepsilon) - c(educ, qob)$,

where $p(.)$ is the education production function and $c(.)$ is the cost function. The optimal choice problem becomes

$$educ = \underset{\tilde{educ}}{\text{argmax}}\{E[U(\tilde{educ}, qob, \varepsilon)|qob, \eta]\} \ . \tag{15}$$

The specification of the utility function is crucial. The functional form $U(educ, qob, \varepsilon) = p(educ, \varepsilon) - c(educ, qob)$ is not chosen for convenience. The quarter of birth must be part of the cost function, otherwise *qob* would not be valid instruments – but at the same time it cannot be part of the educational production function because otherwise the causal effect of *educ* cannot be excluded from the joint effect of $(educ, qob)$. The costs can depend among the ability's signal, η, if for example a staying-based financial aid is available. This possibility, however, is not taken into account. The decision problem just described is illustrated in Figure 5.

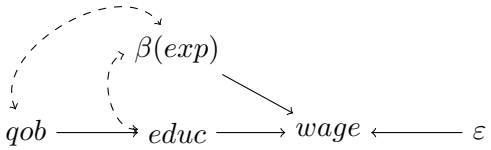

Figure 5: When endogeneity is an issue due to the presence of a model misspecification, the use of VCM is not enough to guarantee causal analysis, and the introduction of IVs becomes necessary to ensure the exogeneity.

In this context the exclusion restriction requires the choice variable *educ* to be separable in (ε, qob). This depends upon the assumptions that the researcher is willing to make about the educational production function $p(\cdot)$ and the cost function $c(\cdot)$.

All the previous considerations show how a model like (13)-(14) is able to: 1. make individual returns heterogeneous, 2. solve the endogeneity problems that are due to the functional form misspecification using the VCM nature of the model, 3. solve the endogeneity problems that are due to the nature of the regressors using IVs, and 4. relate the structural specification to the economic theory providing a rigorous microfoundation of the outcome equation.

References

[1] Angrist, J. D. and Krueger, A. B., 1991, Does Compulsory School Attendance Affect Schooling and Earnings? *The Quarterly Journal of Economics*, 106(4), 979-1014.

[2] Benini, G. and Sperlich, S., 2016, Modeling Heterogeneity by Structural Varying Coefficient Models *Paper presented at the 2015 IAAE Annual Conference.*

[3] Card, D., 2001, Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems. *Econometrica*, 69, 1127-1160.

[4] Eilers, P.and Marx, B., 1996, Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2), 89-121.

[5] Engle, R. F. and Hendry, D. F. and Richard, J-F., 1983, Exogeneity. *Econometrica*, 51(2), 277-304.

[6] Fan, J. and W. Zhang, 2008, Statistical methods with varying coefficient models. *Statistics and Its Interface*, 1(1), 179-195.

[7] Hastie, T. and Tibshirani, R., 1993, Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4), 757-796.

[8] Kim, K. and Petrin, A., 2011, A New Control Function Approach for Non-Parametric Regressions with Endogenous Variables. *NBER Working Paper*, No. 16679.

[9] Murnane, R., Willett, J. and Levy, F., 1995, The Growing Importance of Cognitive Skills in Wage Determination. *Review of Economics and Statistics*, xxxvii(2), 251-266.

[10] Park, B. U., Mammen, E., Lee, Y. K. and Lee, E. R., 2013, Varying Coefficient Regression Models: A Review and New Developments. *International Statistical Review*, 0(0), 1-29.

[11] Linton, O. and Nielsen, J. P., 1995, A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration. *Biometrika*, 82(2), 93-100.

[12] Pudney, S., 1989, *Modelling Individual Choice: The Econometrics of Corners, Kinks, and Holes.* Oxford : Blackwell.

[13] Roca-Pardinas, J. and Sperlich, S. A., 2010, Feasible estimation in generalized structured models. *Statistics and Computing*, 20, 367-379.

[14] R Core Team, 2014, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. *http://www.R-project.org/*

[15] Schultz T. P., 2003, Human Capital, Schooling and Health Returns. *Economics and Human Biology*, 1(2), 207-221.

[16] Sperlich, S., 2009, A Note on Nonparametric Estimation with Predicted Variables. *The Econometrics Journal*, 12, 382-395.

[17] Sperlich, S. and Theler, R., 2015, Modeling Heterogeneity: A Praise for Varying-coefficient Models in Causal Analysis. *Computational Statistics*, 30, 693-718.

[18] Tesler, L., 1964, Iterative Estimation of a Set of Linear Regression Equations. *Journal of the American Statistical Association*, 59, 845-862.

[19] White, H. L., 1980, A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, 48(4), 817-838.