Modeling Heterogeneity by Structural Varying Coefficients Models

Giacomo Benini and Stefan Sperlich

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

January 22, 2016

Abstract

High degrees of heterogeneity across economic units makes the estimates of structural parameters uninformative and inaccurate. The present paper develops a flexible triangular semiparametric varying coefficient model able to identify and estimate cross-sectional heterogeneity while increasing the credibility of the instruments used to solve eventual endogeneity problems. In order to make the econometric model theory-consistent all considerations and developments are embedded in the modeling of equilibrium equations rather than in deus-ex-machina remedies.[1]

[1]The second author acknowledges financial support from the Swiss National Science Foundation, project 100018-140295.

1 Introduction

Disentangling causality from correlation is one of the fundamental problems in data analysis. In econometrics and applied economics the need to extrapolate consequential links form observed correlations arises with all its importance in the impact evaluation literature, where the identification and estimation of the total and of the marginal impact of a treatment highlights the value of an experimental methodology.

However, since in applied economics it is impossible to implement a pure experiment, the disentangling is strictly speaking impossible. The best alternative consists of a two-stage process: [1] first find different sets of mostly non-testable assumptions under which the causal returns can be expressed as parameters or as functions which can be defined in terms of moments or distributions; [2] make a second round of assumptions which allows to find consistent estimators of these moments and distributions. Sometimes it is disappointing that applied researchers have different preferences regarding the relative importance of the two stages of the process. Statisticians would normally care less about causality preferring an excellent estimator to a sharp understanding of the consequential links, while econometricians would do the opposite.

However, in order to do *good empirical research* a data analyst has to take into account correct identification and estimation simultaneously because they represent inseparable problems. An optimal solution for identification of causal effects accomplished using an estimate that is not reliable makes the empirical analysis worthless, while an excellent estimator that does not clarify which causal links are implied is useless from a policy perspective. The identification issue is doubtless important for a correct interpretation, but a reasonable empirical study should try to find and use the estimator that, for the given sample and sample size, produces a value that in probability comes closest to the population parameter of interest.[2] Instead of arguing which among the two tasks is more important it is better to find a solution that is not necessarily based on a strict separation between the detection of causality and the good estimation of moments or the distributions that describe the outcome equation. Traditionally, however, this is not the case. In parametric settings the discussion about the detection of the causal links often follows a reverse-engineering process. When an applied economist observes heterogeneous coefficients for cross section observations that have the same dependent and explanatory variables, a natural deduction is that there is an unobserved interaction between the covariates and the stochastic component of the regression and, therefore, causality is in jeopardy. Once endogeneity is detected the search for instrumental variables (IVs) starts. However, since exogeneity is a property of random variables relative to the parameters of interest, if the functional form is misspecified, the whole procedure it is meaningless and the nice properties that parametric specifications have with respect to [2] become irrelevant. On the contrary, in a nonparametric world, where the reverse-engineering

[2] As we cannot calculate probability, this measure is typically replaced by the mean squared error.

process can not be used, the endogeneity problem becomes per se more complicated making the traditional linear IV approaches useless. As a result, if a practitioner is tempted to use nonparametric models in order to avoid functional form misspecification he would be limited by the inability to detect endogeneity, but also by the difficulties to fulfill [2]. In particular, a purely nonparametric IV estimation is hard to interpret, suffers from large variation, and has even slower convergence rates than a standard nonparametric estimator.[3]

Semiparametric models, on the other hand, are able to conciliate the need to investigate causality with estimating assumptions that are required in order to make good quantitative analysis. In particular, every time an endogeneity problem undermines the causal analysis and the solution proposed is to use an instrument, if the model undergoes a functional form misspecification all the hopes to reach a causal identification are lost. Conversely, a semiparametric environment solves, or at least attenuates, the potential functional form misspecification and, at the same time, produces estimates that have a much better trade-off between fitting and estimating than a pure nonparametric specification. Yet, even if today there is no longer a gap between statistics and econometric theory in the use of semiparametrics specifications[4], their use in applied research still quite rare. The intent of this paper is to construct a bridge between the way endogeneity problems are solved in parametric settings and in nonparametric ones combining the best part of both by using a semiparametric specification in the form of a varying coefficient model (VCM).

1.1 Why Varying Coefficient Models?

Varying coefficient models are the best semiparametric specification to explore hidden patterns in the datasets while keeping a simple interpretation of the results. A VCM maintains the linearity assumption of the regressors, which means that endogeneity enters the model like in a parametric framework, but, at the same time, it allows the returns to change, making them non-random functions of possible hidden paths in the dataset. In this context the returns can be interpreted similarly to the traditional way, impossible in a pure nonparametric specification. Since each coefficient is no longer an average given a particular sample realization, but a function, parameters are allowed to have a

[3]Horowitz (2011) says "... nonparametric estimators differ in ways that are important for applied research. Nonparametric estimation is not just a flexible form of parametric estimation" without explicitly saying that they combat endogeneity in various ways and make the validity of IVs more credible. He continues somewhat later "... Some characteristics of nonparametric IV methods may be unattractive ... [as they] can be very imprecise. This is not a defect of the estimators. Rather, it reflects the fact that the data often contain little information about g [the function of interest] when it is identified through IVs...".

[4]The only difference is that power series are very popular in econometrics whereas in statistics they are regarded as a clear step backwards.

different degree of heterogeneity even for the same levels of the effect modifiers. This fact has two important implications. First, unlike any parametric model where cross section heterogeneity comes only through the error term, allowing for deviations from the model hyperplane, here, heterogeneity can come from the deviation of the coefficients reflecting an individual divergence from the mean. From a microeconomic point of view this means that the averaging of individuals' returns made to describe the behavior of a representative agent is wrong. Second, the model can be identified and estimated using the control function approach, which means that, by construction, endogeneity is only the result of the correlation between the stochastic error component in the equation with endogenous explanatory variables and the stochastic disturbance in the selection (or IV) equation. However, since the model considers parameters as non-random functions that are able to adapt with the varying of the effect modifier, the nonparametric part of the model plays a role in eliminating the endogeneity of the model due to functional form misspecification. As a result, the model deals with endogeneity not only by using the control function, but also by modelling heterogeneity directly, de facto funding a solution that does not clearly separate [1] and [2]. Consequently, the choice of the assumptions to use in order to achieve identification is both based on economic theory and linked to the estimation methodology.

1.2 Some standard approaches to nonparametric IV modeling

Since the VCM bonds parametric IV with nonparametric IV modeling it is useful, before presenting the new method, to review some standard approaches to the endogeneity problems in nonparametric settings due to their relative complexity.

Newey and Powell (NP-IV) (1988, 2003) suggested a single equation model identifiable with a conditional moment restriction

$$Y_i = g_0(X_i, Z_{1i}) + \varepsilon_i \quad E[\varepsilon_i|X_i] \neq 0.$$

In this setting Y is an observable scalar random variable, $g_0(.)$ denotes the true (unknown) structural function of interest, X is a scalar explanatory variable, $Z = [Z_1, Z_2]$ is a vector of IVs, and ε is the stochastic error term correlated with X. The conditional moment restriction (CMR) that allows to identify the model takes the form of a mean-independence condition between the stochastic error and the instrument(s)

$$E[\varepsilon_i|Z_i] = 0 \qquad \text{CMR.}$$

The relationship between the structured and the reduced form of the model is a Fredholm integral of the first kind that leads to an ill-posed inverse problem. Therefore, in order to overcome non-continuity, the authors have to restrict the domain of the true function. In particular, they have to impose bounds on the higher-order derivatives that makes the

mapping from reduced to structural form continuous. It should, however, be mentioned that even the authors qualified this method to be less attractive for practitioners.

A different solution was proposed by Newey, Powell and Vella (1999) (NPV-CF). They used a triangular recursive model where the first equation describes the structural relation between the dependent and the endogenous explanatory variable, while the second describes the relation between the endogenous explanatory variable and the IVs

$$\begin{aligned} Y_i &= g_0(X_i, Z_{1i}) + \varepsilon_i \quad E[\varepsilon_i|X_i] \neq 0 \\ X_i &= f_0(Z_{1i}, Z_{2i}) + \eta_i \quad E[\eta_i|Z_i] = 0. \end{aligned}$$

Identification is achieved imposing an exclusion restriction in the form of a conditional mean independence (CMI) assumption

$$E[\varepsilon_i|Z_i, \eta_i] = E[\varepsilon_i|\eta_i] \qquad \text{CMI.}$$

In order to estimate the model, the authors used the control function approach (Telser, 1964), making the stochastic error of the first equation a function of η_i such that

$$\varepsilon_i = \lambda(\eta_i) + \vartheta_i \quad E[\vartheta_i|\eta_i] = 0.$$

Kim and Petrin (2011) (KP-CF) showed that the CMI assumed in the NPV-CF model is hard to justify both in models where endogeneity is the result of demand-supply equilibrium conditions (intrinsically simultaneous models), but also in models where the agent's choice variable is function of a set of exogenous variables and of a signal variable of the stochastic random error (recursive models). Therefore, they replaced this assumption with a conditional moment restriction that can be motivated by economic primitives merging the structure of NPV-CF with the assumptions of the NP-IV.

Darolles, Fan, Florens and Renault (2011) do not specify the endogeneity source and directly start out from

$$Y_i = g_0(X_i) + \varepsilon_i \quad E[\varepsilon_i|X_i] \neq 0.$$

Here, unlike in the NP-IV and in the NPV-CF, $g_0(.)$ is not function of an exogenous variable Z, but only of the endogenous regressor X. The identification and estimation of the model is obtained imposing the condition

$$E[\varepsilon_i - g_0(X_i)|Z_i] = 0.$$

The function $g_0(.)$, in this specification, becomes, like in the NP-IV case, the solution of an ill-posed inverse problem. They analysed identification and overidentification of this model proposing an estimation procedure based on the Tikhonov regularization and presenting the asymptotic properties of the nonparametric IV estimator.

All the previous models did not specify the source of endogeneity, whereas the triangular models (NPV-CF and KP-CF) often insinuate that the source of endogeneity is the result

of a self-selection mechanism. Contrary to the previous specifications, Imbens, Blundell, Newey and Parsson (2007) and Imbens and Newey (2009) studied the triangular recursive model when the multidimensional disturbances are non-separable

$$\begin{aligned} Y_i &= g_0(X_i, \varepsilon_i) \quad E[\varepsilon_i|X_i] \neq 0 \\ X_i &= f_0(Z_i, \eta_i) \quad E[\eta_i|Z_i] = 0, \end{aligned}$$

In this formulation endogeneity is implicit in the functional specification. Since errors are non-additive the unobserved interaction between the stochastic component and the explanatory variables is imposed by construction.

2 The Structural Approach to Varying Coefficient Models

The VCM we propose reaches across the triangular nonparametric IV models, presented in the previous section, and the standard parametric IV models normally used by applied economists. The model is mainly composed by two equations

$$Y_i = \beta_i^T X_i + \varepsilon_i \tag{1}$$

$$X_i = f_0(Z_i) + \eta_i, \quad E[\eta_i|Z_i] = 0 \tag{2}$$

where Y is a $d_y \times 1$ vector of dependent variables, X is a $d_x \times 1$ vector of explanatory variables, which we allow to contain endogenous components, Z is a $d_z \times 1$ vector of IVs and $f(\cdot)$ is a $d_x \times 1$ vector of smooth functions. Furthermore, ε and η are, respectively, the endogenous error and a stochastic disturbance that has expected values equal to zero and finite variance. Irrespectively of its origin, in this triangular setting endogeneity can only be the result of the correlation between the error in equation (1) and the disturbance in equation (2). In other words, since $cov(\varepsilon, \eta) \neq 0$ the expected value of the stochastic error in equation (1) cannot be zero.

The coefficient β_i is allowed to vary over i. In particular, it is a function of a $d_q \times 1$ vector of observable variables Q that we call effect modifiers

$$\beta_i = g_0(Q_i) + \delta_i \text{ such that } E[\delta_i|Q_i] = 0 \tag{3}$$

where $g_0(.)$ is a vector of functions of the effect modifier, and δ is a stochastic mean zero disturbance with finite variance. As our main interest is to correctly estimate the causal impact of X on Y and not the one of Q, it is plausible to choose for $g_0(.)$ the best nonparametric predictor of β_i for a given Q_i. Therefore, we can assume $E[\delta_i|Q_i] = 0$ by construction. It is useful to notice that since per definition an instrument must be partially correlated with the endogenous variable once the other explanatory variables have been netted out, the effect modifier must be an element of the vector Z.

The triangular semiparametric model described by equations (1) to (3) is similar to the NPV-CF model, but with some important differences. Unlike in the NPV-CF the

function $g(X_i, Z_{1i})$ is expressed in a semiparametric form that transforms the model into a VCM with endogenous covariates. Moreover, in order to identify and estimate the model, we do not use the CMI assumption $E[\varepsilon_i|Z_i, \eta_i] = E[\varepsilon_i|\eta_i]$ which has been shown to be implausible in many situations (Kim and Petrin, 2011).

VCM is such an attractive compromise for causality analysis in an environment that presents endogeneity issues for different reasons. To solve any endogeneity problem the first step is to identify its source, namely reversed causality (self-selection), omitted variables or functional form misspecification[5]. Once the source has been investigated a possible solution must be presented (modeling, finding proxies, selection of useful IVs, etc.). The next logical step is to reflect about the potential trade-offs that the new setting presents. IV solutions are characterized by three main issues. First, the set of non-testable assumptions used to achieve identification, normally are very hard to justify (in particular the stochastic independence from potential gains, including the exclusion restriction), allow to identify the *local average treatment effect (LATE)* (Angrist and Imbens, 1994) which is a function of the IV choice, but not the parameters or functions of interest. A second, often underestimated, problem is that IV regressions tend to produce very imprecise estimates. If the goal of the research is to find good point estimates, instead of intervals, an analysis of the mean squared error (MSE) highlights that an unbiased IV point estimate is in probability more distant from the true parameter of interest than a potentially biased direct estimator unless the bias is huge. On top of these two problems there is a more general concern about the interpretability of IV regressions for policy relevant questions (Heckman 2010).

We do not claim that VCM solves all this problems, but that: [1] endogeneity can be strongly reduced if not eliminated. If not, [2] a VCM can weaken the IV assumptions making them more credible, and, at the same time, it [3] reduces the impact of IVs, in the sense that the LATEs vary less over the instruments which is a clear improvement in terms of interpretability.[6] Finally, [4] a good choice of Q can reduce the MSE substantially.

In order to see these four properties in action an intuitive description of an individual choice process can help. Consider the case in which endogeneity is caused by individual i choosing his X_i related to his expected return β_i. The agent might not exactly know his β_i but he knows his Q_i. Endogeneity arises because $\beta_i - E[\beta_i]$ is typically not independent from X_i. If the objective is increasing Y, people with larger β_i are more motivated to rise their X_i. Therefore, any good predictor for β_i will substantially reduce the endogeneity problem and be a better solution than a classic IV approach, at least in practice. Even if we believe that δ_i is correlated with X_i we only need IVs to be independent from β_i

[5] Note that the following discussion would be different if measurement error in X would be the problem.

[6] An alternative way to improve interpretation is to estimate the marginal treatment effect, see Heckman (2010). Certainly, this is still dependent from the chosen instrument, and it requires a sufficiently large sample.

conditional on Q_i and the variation of the LATEs limited by the one of the δ_i, not by the one of the β_i. This means that for different IVs we identify similar parameters.

Substituting equation (3) into (1) shows that the VCM is a special case of a nonparametric model that can be identified and estimated using the control function approach. In the same way the VCM can be identified using a conditional moment restriction

$$E[\epsilon_i|Z_i] = 0 \qquad \text{CMR} \tag{4}$$

motivated by economic primitives, where $\epsilon_i = \delta_i x_i + \varepsilon_i$ is the composite error. Using the CMR assumption the relationship between ϵ and the controls must be a function of η but also of the exogenous variables Z, such that

$$\epsilon_i = h_0(Z_i, \eta_i) + \vartheta_i \qquad E[\vartheta_i|Z_i, \eta_i] = 0 \tag{5}$$

where ϑ is an exogenous disturbance term, and $h(\cdot)$ is typically supposed to be a smooth function, which can be estimated using standard nonparametric techniques. Plugging the control function into equation (1) makes Y a function of (Z, η) and X

$$Y_i = g_0(Q_i)X_i + h_0(Z_i, \eta_i) + \vartheta_i \tag{6}$$

Taking the expectation conditional on the observed variables and imposing the CMR assumption leads to the model that allows to identify the causal impact of X on Y, and that can be estimated with a two-step VCM estimator:

$$E[Y_i|Z_i, \eta_i] = g_0(Q_i)X_i + E[\delta_i|Z_i, \eta_i]X_i + E[\varepsilon_i|Z_i, \eta_i]. \tag{7}$$

Note that the appearance of $\delta_i X_i$ in the error term is not a defect of VCMs, to the contrary, it is an advantage because it makes the error terms explicit. In other words, in all the specifications that do not model heterogeneity directly the whole heterogeneity is absorbed by the regression disturbance ε. In the VCM δ_i is just the left-over of the best prediction of β_i, so it is not hard to argue that either the second factor in (7) is zero, or it is a function of (Z_i, η_i). If we would like to explore further the structure above we could also estimate the additive VCM in the form of

$$E[Y_i|Z_i, \eta_i] = g_0(Q_i)X_i + E[\delta_i|Z_i, \eta_i]X_i + E[\varepsilon_i|Z_i, \eta_i] \tag{8}$$

However, we doubt that this would lead either to a real improvement in estimation or interpretation. Before identifying and estimating $E[Y_i|Z_i, \eta_i] = g_0(Q_i)X_i + h_0(Z_i, \eta_i)$ it is useful to see an economic application where the model shows all its potentials.

Example: Job Training Program

Suppose that a rational individual has to make a choice of either participating to a job training program or not. The agent's wage, Y, is a function of the participation at the

program, $X \in \{0,1\}$, but also of her unobserved ability, ε. The agent's ability is not observed neither by the agent herself nor by the econometricians, but the information set that the worker can consult, before deciding to participate or not to the program, includes a signal of her individual ability η, for example her past working history. The course participation has a cost Z, in particular in this specification there are two cost-shifters, the effect modifier Q, and the instrument Z_2, such that $Z = [Z_1, Z_2]$, with $Q = Z_1$.

While the agents make the choice based on hers individual payoff function $U(X, Z, \varepsilon)$, the econometrician is interested in estimating the production function $m(X, Q, \varepsilon)$. The agent's utility has to be a function of the participation to the program, of the cost-shifters and of the unobserved ability, $U(X, Z, \varepsilon) = m(X, Q, \varepsilon) - c(X, Z_2)$, where $m(.)$ is the production function and $c(.)$ is the cost function. In this framework the optimal choice problem becomes

$$X = \underset{\tilde{X}}{\text{argmax}}\{E[U(\tilde{X}, Z, \varepsilon)|Z, \eta]\}.$$

The specification of the utility function is crucial. The form $U(X, Z, \varepsilon) = m(X, Q, \varepsilon) - c(X, Z_2)$ is not just chosen for convenience. Z_2 must be part of the cost function to be a valid instrument, but, at the same time, it cannot be part of the production function otherwise the causal effect of X cannot be excluded from the joint effect of (X, Z_2). The costs can depend on the ability's signal, η, if for example a merit-based financial aid is available; this possibility is not taken into account for simplicity but it would not change the result. Using the same specification form of equation (1) the production function takes the form,

$$m(X_i, Q_i, \varepsilon_i) = \beta(Q_i)X_i + \varepsilon_i = g(Q_i)X_i + \epsilon_i \qquad \text{with} \qquad \epsilon_i = \delta_i X_i + \varepsilon_i$$

where the effect modifier is exogenous $Q_i = Z_{1i}$, while X is instrumented by Z_2 once the effect of Q_i has been netted out.

Using the previous specification the optimization problem of the agent becomes

$$\begin{aligned} X &= \underset{\tilde{X}}{\text{argmax}}\{E[g(Q)\tilde{X} + \epsilon|Z, \eta] - c(\tilde{X}, Z)\} \\ &= \underset{\tilde{X}}{\text{argmax}}\{E[g(Q)\tilde{X}|Z, \eta] + E[\epsilon|Z, \eta] - c(\tilde{X}, Z)\} \\ &= \underset{\tilde{X}}{\text{argmax}}\{E[g(Q)\tilde{X}|Z, \eta] - c(\tilde{X}, Z)\} \end{aligned}$$

As a result of the individual choice of the optimal X becomes a function of (Z, η). The model can now be expressed in the triangular form

$$\begin{bmatrix} 1 & -g(Q_i) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} Y_i \\ X_i \end{bmatrix} = \begin{bmatrix} 0 \\ f_0(Q_i, Z_{2i}) \end{bmatrix} + \begin{bmatrix} \epsilon_i \\ \eta_i \end{bmatrix}$$

such that the model can be expressed by equations (1)-(2)-(3) recursively in the form of a output-choice equilibrium

$$\begin{aligned} Y_i &= g(Q_i)X_i + \epsilon_i && \text{outcome function} \\ X_i &= f_0(Q_i, Z_{2i}) + \eta_i && \text{choice function} \end{aligned}$$

This new specification has two main advantages: [1] since the model takes into account the possibility that agents participate or not to the program, partly based on their different returns, it is much more realistic that any standard parametric specification. This fact makes the validity of instruments more credible, eases a lot the interpretation of the results and helps to answer relevant policy questions, moreover [2] the model is identified using the CMR. This mean that even if X is not separable from (Z, η), and this depends upon the assumptions that we are willing to make about the functional form of the costs $c(.)$, we can still identify the model. This last finding is particularly useful since virtually every non-linear cost function impedes a division of X from (Z, η), this would not be the case using the standard CMI assumption, $E[\epsilon|Z, \eta] = E[\epsilon|\eta]$.

2.1 Endogeneity and Model Misspecification Error

The most important quality of the hybrid nature of the VCM specification is its ability to deal with one particular kind of endogeneity that is very important in microeconometrics: the functional form misspecification.

In order to see how the model mitigates the functional form misspecification it is necessary to discuss the sources of endogeneity and the mechanism through which it impacts in a model where the coefficients are non-random functions of the effect modifiers. Like in every triangular model, the endogeneity mechanics in a VCM is simple: since by construction $cov(X, \eta) \neq 0$ in order to have $cov(X, \varepsilon) \neq 0$ it must be that $cov(\varepsilon, \eta) \neq 0$, otherwise the regressors and the error would be linearly independent from the error. Considering again the Job Training Program described in the previous. The maximization process leads to an outcome-choice equilibrium in the form

$$\begin{aligned} Y_i &= g(Q_i)X_i + \epsilon_i \\ X_i &= f_0(Q_i, Z_{2i}) + \eta_i \end{aligned}$$

where the effect modifier Q is the first cost-shifter Z_1. This is the simplest possible specification of a VCM where there is one endogenous explanatory variable X, one exogenous effect modifier Q, and one instrument Z_2. The causal links for this outcome-choice equilibrium are portrayed in Figure 1. However, even if the genesis of the correlation between the explanatory variable and the error term is not complicated, its origins can be multiple. In a sense this is not a problem since the different sources of endogeneity can be reassembled under the concept of model misspecification. A researcher can almost always conclude that every time at least one of the assumptions made on the data

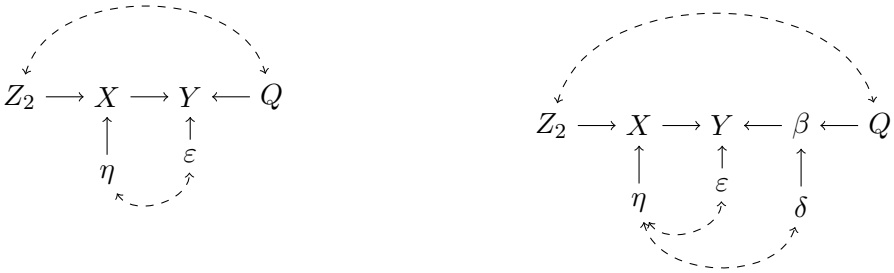

Figure 1: The mechanism of the endogeneity process changes depending upon the assumptions about the relationships between the error ε and the stochastic disturbances (η, δ). The left picture shows a world of homogeneous coefficients where the only source of endogeneity is $cov(\varepsilon, \eta) \neq 0$. The left picture is the result of the introduction of a varying coefficient structure, as a new possibility η can now be correlated with the stochastic mean zero disturbance of (3), $cov(\eta, \delta) \neq 0$. The direct connection between δ and ε is not taken into account because the interest is about the causal link between Y and X for a given level of Q, therefore it is reasonable to consider δ as being exogenous *sive natura*.

generating process (DGP) is not fulfilled, then at least one of the explanatory variables becomes correlated with the error term. Hence, since exogeneity is a property of random variables relative to parameters or functions of interest, it is crucial to understand if the misspecification is the result of the chosen variables, of the parameters or of the function relating them.

For example, if the researcher chooses a linear specification $Y_i = \beta X_i + \gamma Q_i + \epsilon_i$, while the unobserved DGP is an NP-IV model $Y_i = t(X_i, Q_i) + u_i$, then the standard ordinary least squares estimators $(\hat{\beta}^{OLS}, \hat{\gamma}^{OLS})$ cannot measure the (micro-)responses $E[Y_i|X_i, Q_i]$ to a change in (X_i, Q_i), because

$$\hat{\beta}^{OLS} \nrightarrow \frac{\partial t(X_i, Q_i)}{\partial X_i}, \qquad \hat{\gamma}^{OLS} \nrightarrow \frac{\partial t(X_i, Q_i)}{\partial Q_i}.$$

As a result, even thought the correct regressors are chosen, the OLS cannot make individual predictions, but only predictions of aggregate changes (Stoker, 1982). In particular, if the solution of an endogeneity problem is to use the NPV-CF model, and the researcher does not make a functional form mistake in the instrumental equation, the functional form problem becomes:

Data Generating Process	*Parametric (Linear) Misspecification*
$\begin{cases} Y_i = t(X_i, Q_i) + u_i \\ X_i = f(Z_i) + \eta_i \end{cases}$	$\begin{cases} Y_i = \beta X_i + \gamma Q_i + \epsilon_i \\ X_i = f(Z_i) + \eta_i \end{cases}$
$u_i = l(Z_i, \eta_i) + \xi_i$	$\epsilon_i = h(Z_i, \eta_i) + \vartheta_i$

Since the DGP $t(X,Q)$ is mistaken by $\beta X + \gamma Q$ the functional form misspecification is reflected in an indirect misspecification of the control function, which does not link the true error u, but ϵ to (Z,η). This kind of structural misconstruction shows that, even if u is independent from Z, it is not guarantee that the same is true for ϵ, as a result the choice of a *good instrument is irrelevant* if there is a significant functional misspecification error.

If the NPV-CF model is misspecified with the varying coefficient model expressed by equations (1)-(2)-(3)-(5), such that

$$\begin{array}{ll} \textit{Data Generating Process} & \textit{Varing Coefficient Misspecification} \\ \begin{cases} Y_i = t(X_i,Q_i) + u_i \\ X_i = f(Z_i) + \eta_i \end{cases} & \begin{cases} Y_i = \beta_i X_i + \varepsilon_i = g(Q_i)X_i + \epsilon_i \\ X_i = f(Z_i) + \eta_i \end{cases} \\ u_i = l(Z_i,\eta_i) + \xi_i & \epsilon_i = \delta_i X_i + \varepsilon_i = h(Z_i,\eta_i) + \vartheta_i. \end{array}$$

things are different. Like in the linear case the functional misspecification of the error is reflected in an indirect misspecification of the control function, but unlike in the linear model it is easier to find a valid instrument stochastically independent from any functional deviation from the constant return to X. As a result, the smooth function $g(.)$ is able to replicate the results obtainable using the parameters (β,γ) at an aggregate level. However, the VCM has an advantage with respect to the linear model: the nonparametric nature of the varying coefficients helps to assimilate the functional form misspecification error at least on a local level, which means that the functional form misspecification error is not entirely absorbed by the $cov(\varepsilon,\eta)$, but also by the stochastic disturbance of equation (3). In order to see how the functional form misspecification is absorbed by $h(.)$ by construction it is sufficient to compute the expected value of the ϵ

$$\begin{aligned} h(Z_i,\eta_i) &= E[\epsilon_i|Z_i,\eta_i] \\ &= E[Y_i - g(Q_i)X_i|Z_i,\eta_i] \\ &= E[t(X_i,Q_i) + u_i - g(Q_i)X_i|Z_i,\eta_i] \\ &= E[t(X_i,Q_i) - g(Q_i)X_i|Z_i,\eta_i] + E[u_i|Z_i,\eta_i] \\ &= E[t(X_i,Q_i) - g(Q_i)X_i|Z_i,\eta_i] + l(Z_i,\eta_i). \end{aligned}$$

3 Identification and Estimation of Triangular VC Models

3.1 Identification

Equation (8) is characterized by two functions, $g(.)$ and $h(.)$. Considering only the pair of functions that satisfy $E[Y_i|Z_i,\eta_i] = g_0(Q_i)X_i + h_0(Z_i,\eta_i)$ due to the CMR assumption,

then if there exists a pair of functions $\bar{g}(.)$ and $\bar{h}(.)$ such that

$$Pr[\delta(X_i, Q_i) + \kappa(Z_i, \eta_i) = 0] = 1 \tag{9}$$

where $\delta(X_i, Q_i) = (g_0(Q_i) - \bar{g}(Q_i))X_i$ and $\kappa(Z_i, \eta_i) = h_0(Z_i, \eta_i) - \bar{h}(Z_i, \eta_i)$, then $g(.)$ and $h(.)$ are identified.

Theorem 1. *Assume equations (1)-(2)-(3) and (4) are satisfied. If for all $\delta(X_i, Q_i)$ with finite expectation $E[\delta(X_i, Q_i)|Z_i] = 0$ implies $\delta(X_i, Q_i) \xrightarrow{a.s.} 0$ then $g_0(Q_i)$ and $h(Z_i, \eta_i)$ are identified.*

Proof. (by contradiction)
Suppose it is not identified. Then, there must exist functions $\bar{g}(Q_i)$ and $\bar{h}(Z_i, \eta_i)$ such that $\delta(X_i, Q_i) \neq 0$ and $\kappa(Z_i, \eta_i) \neq 0$ but $Pr[\delta(X_i, Q_i) + \kappa(Z_i, \eta_i) = 0] = 1$.
Using the CMR assumption we can write

$$E[\epsilon_i|Z_i] = E[h(Z_i, \eta_i) + \vartheta_i|Z_i] \overset{LIE}{=} E[E[h(Z_i, \eta_i) + \vartheta_i|Z_i, \eta_i]Z_i] = E[h(Z_i, \eta_i)|Z_i] = 0$$

As $\kappa(.)$ is the difference between two possible candidates $h_0(.)$ and $\bar{h}(.)$, it must also be that $E[\kappa(Z_i, \eta_i)|Z_i] = E[h_0(Z_i, \eta_i)|Z_i] - E[\bar{h}(Z_i, \eta_i)|Z_i] = 0$, which means that $E[\delta(X_i, Q_i) - \kappa(Z_i, \eta_i)|Z_i] = E[\delta(X_i, Q_i)|Z_i] = 0$.
Since $E[\delta(X_i, Q_i)|Z_i] = 0$ implies $\delta(X_i, Q_i) \xrightarrow{a.s.} 0$, we have, at the same time, $\delta(X_i, Q_i) = 0$ and $\delta(X_i, Q_i) \neq 0$, which is a contradiction. Given that $g_0(Q_i)$ and $E[Y_i|Z_i, \eta_i]$ are both identified, then also $E[Y_i|Z_i, \eta_i] - g_0(Q_i)X_i$ is identified. □

A sufficient condition for identification is now that the conditional distribution of X_i given Z_i satisfies the completeness condition (Newey and Powell, 2003), which is the nonparametric analog of the rank condition for identification in the linear setting. In this case the completeness condition assuming $E[\delta(X_i, Q_i)] = 0$ implies $\delta(X_i, Q_i) = 0$, i.e. for any $\delta(X_i, Q_i)$ with finite expectations we would have $\delta(X_i, Q_i) \xrightarrow{a.s.} 0$.

3.2 Estimation

There are different methods to compute varying coefficient models (Lee, Mammen and Lee, 2012; Fan and Zang, 1999, 2008). We choose a spline estimator for two reasons. First, spline estimators outperform, both in terms of fitting and prediction quality, all the power series competitors. Second, since spline uses more complex approximations to fit the unknown function modelling the DGP as more data are available, the CMR $E[\epsilon_i|Z_i] = 0$ can be fulfilled by construction readjusting the size of the basis on which the splines are built [7].

[7] Note that the same would not be true for kernel based estimators (Hansen, 2012).

The estimation procedure is implemented using a two-stages spline-based least squared minimization which assumes that all the variables $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ are independent and identically distributed (i.i.d.).

Step 1 The first step requires to compute the controls η_i. The least squared objective function to minimize in order to find the optimal controls is

$$\min \sum_{i=1}^{n} \eta_i^2 = \min \sum_{i=1}^{n} \left\{ X_i - f(Z_i) \right\}^2 = \min_{a_j} \frac{\sum_{i=1}^{n} \left\{ X_i - \sum_{j=1}^{J} a_j B_j(Z_i) \right\}^2}{n} \tag{10}$$

Regressing n data points on the set of B-splines $B_j(\cdot)$, with $j = 1, ..., J$, allows to obtain the parameters $\hat{a}_j$ for $j = 1, ..., J$ that enable to compute the individual residuals of equation (2)

$$\hat{\eta}_i = X_i - \hat{f}(Z_i) = X_i - \sum_{j=1}^{J} \hat{a}_j B_j(Z_i).$$

Step 2 Once the controls are computed it is possible to substitute them and to obtain an output equation that is function of the estimated residuals

$$Y_i = g_0(Q_i) X_i + h_0(Z_i, \hat{\eta}_i) + \vartheta_i. \tag{11}$$

Before proceeding with the second step of the estimation procedure it is necessary to de-mean with respect to Z the conditional basis used to construct the splines that approximate the control function $h(\cdot)$, such that $\bar{h}(Z, \hat{\eta}) = h(Z, \hat{\eta}) - E[h(Z, \hat{\eta})|Z]$. Using the demeaned approximation of $h(\cdot)$, it is possible to proceed with the second minimization

$$\min \sum_{i=1}^{n} \vartheta_i^2 = \min \sum_{i=1}^{n} \left\{ Y_i - g(Q_i) X_i - \bar{h}(Z_i, \hat{\eta}_i) \right\}^2 \tag{12}$$

$$= \min_{b_k, c_l} \frac{\sum_{i=1}^{n} \left\{ Y_i - \sum_{k=1}^{K} b_k B_k(Q_i) X_i - \sum_{l=1}^{L} c_l B_l(Z_i, \hat{\eta}_i) \right\}^2}{n}. \tag{13}$$

Note that the set of B-splines used to approximate $f(Z_i)$, $g(Q_i)$ and $h(Z_i, \hat{\eta}_i)$ do not have to be the same and they can change depending on the degree of smoothness of $f(.)$, $g(.)$ and $h(.)$. This allows the econometrician to play ex-post estimation with the magnitude of δ which regulates the individual local deviation from the mean.

If the number of knots is fixed, the method presented here is simply a two-stages least squared estimator that is not more difficult than a polynomial regression, but that outperforms it and respects the CMR assumption by construction.

3.3 Statistical Properties of the VC Models

The set of parameters obtained combining steps 1 and 2 are minima of the sums (of functions) of data, which means that the estimates are semiparametric m-estimators obtained imposing a moment conditions that depend upon $\hat{g}(.)$ and $\hat{h}(.)$. Therefore, the solution of equation (12) can be thought as a semiparametric m-estimator that solves the following moment condition

$$\frac{\sum_{i=1}^{n} m(R_i, \gamma, \hat{p})}{n} = 0 \tag{14}$$

where, $\gamma = [(a_j)_{j=1}^{J}, (b_k)_{k=1}^{K}, (c_l)_{l=1}^{L}]$ is a $(J + K + L) \times 1$ vector of parameters, $R_i = \{X_i, Y_i, Z_i\}_{i=1}^{n}$ is a collection of i.i.d. random variables with joint distribution F and $\hat{p}$ is a function that depends upon the parameters γ, the variables in R and the estimated residuals $\{\hat{\eta}_i\}_{i=1}^{n}$. The statistical properties of $\hat{\gamma}$, including its asymptotic behavior, have been studied by Newey (1994). Provided that a set of regularity conditions are met (see Appendix A), the derivative of the probability limit of the estimated $\hat{\gamma}$ for a path θ is

$$\frac{\partial plim(\hat{\gamma})}{\partial \theta} = E[d(R)S(R)] \tag{15}$$

where, $d(R)$ is the path-wise derivative and $S(R) = \frac{\partial ln(dF_\theta)}{\partial \theta}$ is the score function of the probability density function dF_θ.

Even though elegant, the previous result is relatively uninteresting in this context since the first step of every semiparametric specification, like VCMs, is to use nonparametric analysis as an exploratory tool to find the correct (varying coefficient) specification. As a result, it is not important to fully characterize the high-order asymptotic distribution of equation (14), to the contrary, what matters it is an accurate description of the finite behavior of the varying coefficient function $\hat{g}(.)$ once an a priori analysis of the data has suggested which variables are effect modifiers. One way to do that it is to implement a (wild-)bootstrap of equation (13) calculating the critical values by re-sampling using the following standard procedure:

1. Estimate $\hat{f}(.)$ and $\hat{\eta}_i = X_i - \hat{f}(Z_i)$ for $i = 1, ..., n$ using equation (10).

2. Using the results from step 1, estimate $\hat{Y}_i$ and its resulting residuals $\hat{\vartheta}_i$ for $i = 1, ..., n$.

3. For each $\hat{Y}_i$ generate a $\hat{Y}_i^{boot} = \hat{g}(Q_i)X_i + \hat{h}(Z_i, \hat{\eta}_i) + \hat{\vartheta}_i\omega_i$ with $\omega_i \overset{iid}{\sim} N(0, 1)$. Repeat the previous procedure B times and define a new sample $R_i^{boot} = \{X_i, Y_i^{boot}, Z_i\}_{i=1}^{n}$ for $boot = 1, ..., B$.

4. For each Y_i^{boot} calculate the function $\hat{g}^{boot}(.)$ and for each $\hat{g}^{boot}(.)$ compute confidence intervals h^{boot} for a preselected bandwidth h.

From the newly obtained B estimates we can calculate standard errors, confidence bands and all the usual measures that characterize a DGP. These, however, are based on the assumption that the prediction of η does not affect the estimation of g_0 in the first order.

To see how the previous procedure works in practice consider the following DGP, where the effect modifier $Q_i = Z_{1i}$ and the instrument Z_{2i} are both drawn from a standard uniform distribution and the disturbance in the IV equation is $\eta \sim N(0, 1)$. In this case, the functional form misspecification problem becomes

Data Generating Process	*Varing Coefficient Misspecification*
$\begin{cases} Y_i = 5 + sin(Z_{1i})X_i + \eta_i^3 + u_i \\ X_i = Z_{1i} + Z_{2i} + \eta_i \end{cases}$	$\begin{cases} Y_i = g(Z_{1i})X_i + h(Z_{1i}, Z_{2i}, \eta_i) + \vartheta_i \\ X_i = f(Z_{1i}, Z_{2i}) + \eta_i \end{cases}$

The resulting simulation is portrayed in Figure (2) and it shows how flexible is the VC Misspecification and how the semiparametric nature of the model is able to mirror the DGP.

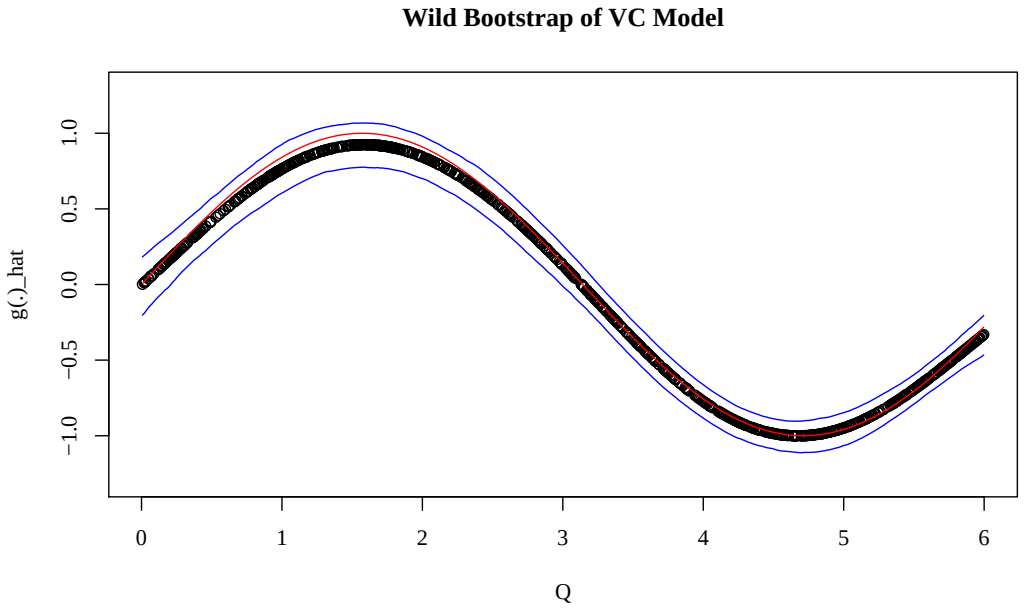

Figure 2: A graphical analysis of the bootstrapped varying coefficient function shows that the VC Misspecification is able to well mimic the DGP (red curve), both with its mean (the dotted black curve) and with the confidence bands for the 5% and the 95% intervals (blue curves).

Presently, for spline estimators of additive VCMs also Bayesian simulation bands are getting more and more popular. A different alternative to the frequentist and the Bayesian methodologies is the use of sub-sampling. This last option has the additional advantage

to repeat both estimation steps for each subsample, i.e. any additional variability of $\hat{g}_0$ derived from the first step is taken into account.

4 Empirical Examples

4.1 Applied Microeconomics

Since the pioneer work of Mincer (1958, 1974), an extensive analysis of the school-experience-earnings relationship has been done. On the basis of both theoretical and empirical arguments the benchmark formula considers the log of earnings as a function of the years of eduction and of the years of potential labor market experience using the following linear specification

$$log(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 exper_i^2 + \varepsilon_i. \tag{16}$$

Equation (16) is the workhorse of research in empirical labor economics (Lemieux, 2003) and it has become the benchmark for every data based analysis due to its ability to fit the data (Sherwing, 1992). However, in recent years, different studies have highlighted that marginal returns to education vary for different levels of working experience (Schultz, 1997). As a consequence, if the wage equation is expressed in a linear form, ignoring the non-linear interaction between experience and education, the returns to education could be systematically underestimated (Card, 2001). A VCM can correct the possible downward bias expressing the returns to education as a function of the different levels of experience such that the Mincer equation becomes

$$log(wage_i) = g_0(exper_i) + g_1(exper_i)educ_i + \delta_i educ_i + \varepsilon_i. \tag{17}$$

The use of a VCM specification has two upsides. Firstly, there is no risk to underestimate the impact of education on earnings due to its non-linear interaction with experience (Cai, Fang, Lin and Su, 2011). Secondly, the VCM allows a form of non-constant returns to experience that goes behind the simple introduction of a squared term which imposes a strict functional specification making (17) similar to the Severance-Lossin and Sperlich (1999) model for the returns to scale.

Irrespectively of the direct modeling of the non-constant returns in equation (17), the exogeneity assumption $E[\varepsilon_i|exper_i, educ_i] = 0$ could be violated due to the omission of crucial unobserved explanatory variables such as the agent's ability. A possible solution is to instrument *educ* using the triangular VCM specification in the form of (1)-(2). We propose as an instrument the father's education. The idea behind this choice is that children that have better educated parents tend to be better educated themselves, independently of individual ability, therefore, the IV equation becomes

$$educ_i = f_0(exper_i, feduc_i) + \eta_i \quad E[\eta_i|exper_i, feduc_i] = 0, \tag{18}$$

where $feduc$ is the education of the father of the householder. Using (17), (18) and a control function like (5) the (equilibrium) outcome equation becomes

$$log(wage_i) = g_0(exper_i) + g_1(exper_i)educ_i + h(feduc_i, exper_i, \eta_i) + \vartheta_i. \tag{19}$$

To construct an empirical counterpart of equation (19) we collect a dataset from the Panel Study in Income Dynamics (PSID) for the 2011 wave. Here, the dependent variable is the logarithm of the hourly wage rate and the two crucial explanatory variables are the number of years worked with the current employer (*exper*) and the completed education level (*educ*). A set of controls is added to check the robustness of the estimates, including the average hours worked last year (h), the level of eduction of the wife ($educwife$), as well as two categorical variables, the race (*race*) and the region of residence (*region*). Dropping the missing values, the dataset has a cross-section dimension of 2649 observations for 8 variables. The standard Mincer linear regression with education instrumented by father's education leads to the following triangular model

$$log(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 exper_i^2 + \varepsilon_i \tag{20}$$

$$educ_i = \gamma_0 + \gamma_1 feduc_i + \gamma_2 exper_i + \gamma_3 exper_i^2 + \eta_i \tag{21}$$

The estimates resulting from a standard two stage least square minimization of (20)-(21) are presented in Table 1.

Table 1: Regression Results

	Dependent variable:
	logwage
educ	0.080^{***}
	(0.014)
exper	0.033^{***}
	(0.003)
I(exper^2)	-0.001^{***}
	(0.0001)
Constant	1.467^{***}
	(0.183)
Observations	2,649
R^2	0.157
Adjusted R^2	0.156
Residual Std. Error	0.507 (df = 2645)
Note:	*p<0.1; **p<0.05; ***p<0.01

The results are in line with standard economic findings. The impact of education and experience are both positive and significant, while the impact of experienced squared

is negative and significant. The estimated coefficients are robust to the introduction of the control variables. A regression that includes also the wife's education, the number of working hours, the race and the region of residence, reduces the constant to $\hat{\beta}_0 = 1.28$, while the magnitude and the levels of significance of the other coefficients remains virtually unchanged ($\hat{\beta}_1 = 0.082$, $\hat{\beta}_2 = 0.025$ and $\hat{\beta}_3 = 0.00026$).

The question now becomes: are equations (20)-(21) a good specification to describe the PSID data?

A first rudimentary graphical analysis shows that a rigid parametric specification like (20)-(21) is unable to capture the non-constant effect of the experience-education relationship on the log of the earnings, suggesting that the results presented in Table 1 are indeed wrong, see Figure (3).

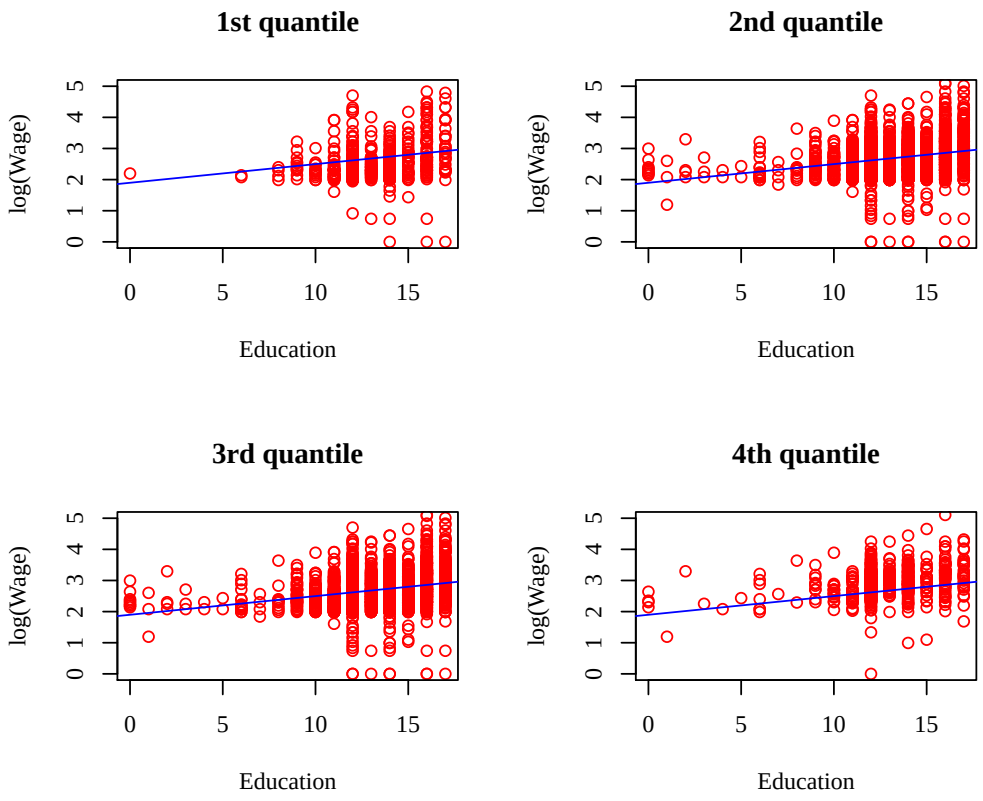

Figure 3: Plotting $log(wage)$ versus $educ$ for low (1st quantile), medium (2nd quantile), high (3rd quantile) and very high (4th quantile) values of $exper$ shows four different regression slopes obtained using a linear (OLS) regression. This fact highlights the necessity to model the impact of education on wages as a function of the level of experience.

Both the intercept and the slope of the four regressions presented in Figure (3) are non-

constant. This empirical finding suggests that a specification like (17) is more appropriate than (16) to model the DGP. Employing the same instrument as in the parametric specification, it is possible to use a two-stages B-Splines like the one presented in equations (10)-(12), leading to the following triangular specification

$$log(wage_i) = g_0(exper_i) + g_1(exper_i)educ_i + \epsilon_i \tag{22}$$

$$educ_i = f_0(exper_i, feduc_i) + \eta_i \tag{23}$$

Model (22)-(23) produces a varying intercept that has a much smaller mean value than $\hat{\beta}_0$ and, to the contrary, a bigger varying slope mean than $\hat{\beta}_1$, see Table 2.

Table 2: Estimated Varying Coefficients

Values	Minimum	1^{ST} Quantile	Median	Mean	3^{RD} Quantile	Maximum
$g_0(exper_i)$	-0.30340	-0.24776	-0.06351	0.00000	0.24664	0.55332
$g_1(exper_i)$	0.07728	0.08855	0.09645	0.09544	0.10272	0.10824

It is interesting to notice that the estimated varying slope $\hat{g}_1(.)$ is bigger than $\hat{\beta}_1$ from its 1^{ST} Quantile onward. The reason is simple, the splines used to fit the function find a $\hat{g}_1(.)$ bigger than $\hat{\beta}_1$ if the years worked with the current employer is lower than 10 and, since 74% of the sample has a level of experience between 0 and 10 years, it is normal that almost all the values of the returns to education are bigger than the parametric estimate, see Figure (4).

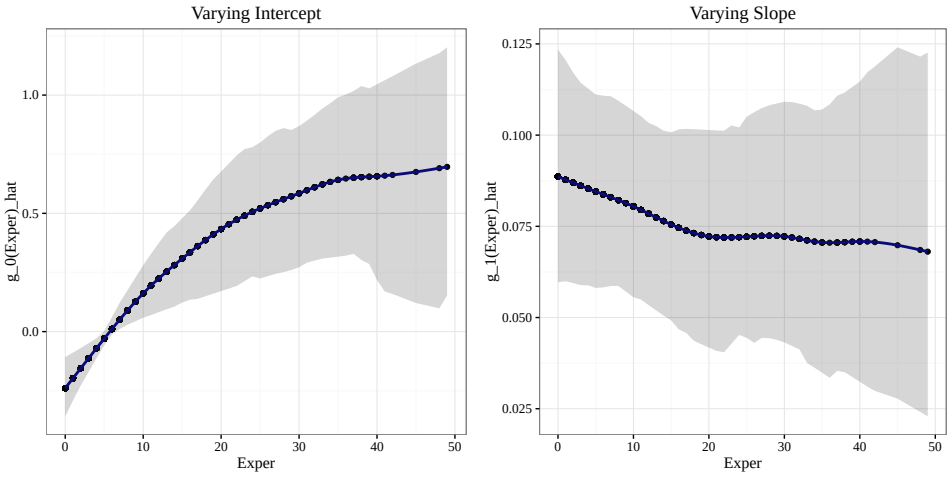

Figure 4: A graphical analysis of the empirical estimates shows that for the largest majority of householders, the one that have less than 10 years of experience, the returns to education are bigger in a VCM than in a standard parametric model.

This means that the estimated VCM confirms the findings of Card and Schultz: the returns to education are non-constant for different levels of experience and the use of a standard linear model underestimates the returns of education on wages for people with low experience.

4.2 Applied Macroeconomics

Solow and Swan (1956) proposed a structural specification of the Neoclassical growth model that relates the total amount of production Y to the capital stock K, the total level of employment L and the available technological shock A using an aggregate production function $F(.)$, in the form

$$Y_i = F(K_i, L_i, A_i). \tag{24}$$

In the last decades many empirical studies have added one or more variables to (24) transforming the growth process into a "theory-of-everything", where any variable is welcome provided it helps to explain the aggregate quantity Y. The problem with this modeling strategy is its inability to explain if the new variable supports the identification of a previously unknown growth driver, or if it changes the causal chains that explain the nature of development. Therefore, instead of adding one or more variables and try to analyses a growth-augmented model, we restudy the Neoclassical equation under the hypothesis that the coefficients, that characterize the production function $F(.)$, depend upon the state of development of an economy. In other words, we consider a model where the growth's drivers are the some one of equation (24), but, the returns are function of the state-of-the-economy. More precisely, we start from the specification proposed by Mankiw, Romer and Weil (1992),

$$lnY_{it} = \rho Y_{i,t-l} + X_{it}^T \beta + \varepsilon_{it} \qquad l \geq 1, \tag{25}$$

for a panel data and transform (25) the βs into functions of the state-of-the-economy. The aim of this strategy is to capture a great deal of information, while keeping a functional form specification that includes only few explanatory variables. The problem now is to find a variable that captures the stage of development of an economy.

Kuznet, in his pioneer work on the relationship between economic growth and income inequality (1955), advanced a simple and yet powerful hypothesis: inequality is a function of the state of development of a nation. More precisely, he postulated that during the first stages of the industrialization process inequality grows, while, when a country is richer, inequality declines. This hypothesis has been confirmed by many empirical studies (Banerjee and Newman, 1993; Aghion and Bolton, 1997; Banerjee and Duflo, 2003). As a result, the measure of inequality can be used as a proxy variable able to capture the state-of-the-economy.

Exploiting the Kuznet curve, we propose a new semiparametric specification that wants to capture the structural impact of capital, labor and technology for different levels of inequality on the aggregate output. In order to do that, we collapse the time dimension of equation (25), in order to get rid of the business-cycle, and analyse an equilibrium equation

$$lnY_i = \rho Y_{i0} + \beta_0 + \beta_{1i} lnK_{i,ph} + \beta_{2i} lnK_{i,h} + \beta_3 lnDep_i + \varepsilon_i, \tag{26}$$

where the two coefficients that multiply the human and the physical capital are function of the amount of inequality in a country[8], identified by the Gini coefficient,

$$\beta_{1i} = g_1(gini_i) + \delta_{1i} \qquad \beta_{2i} = g_2(gini_i) + \delta_{2i}. \tag{27}$$

Unlike equation (25), the cross-sectional specification (27) avoids to produce biased estimate of ρ that tends to be close to one and produce spurious regressions.

To empirically validate the relationship proposed in equations (26)-(27), we use the dataset constructed by Köhler (2014). The panel collects informations about the values of physical and human capital together with the joint depreciation rate of the economy for 81 countries from 1960 to 2007. We average the time dimension of the data, from 1971 to 2007, for all the exploratory variables, with the exception of Y_{i0} and of $gini_i$ for which the 1970 values are used, and get the following specification

$$ln\overline{Y}_i = \rho Y_{i,1970} + \beta_0 + g_1(gini_{i,1970}) ln\overline{K}_{i,ph} + g_2(gini_{i,1970}) ln\overline{K}_{i,h} + \beta_3 ln\overline{Dep}_i + \epsilon_i, \tag{28}$$

with $\overline{K}_{i,ph} = \sum_{t=1971}^{2007} K_{iph,t}/36$ and the same for the other variables and $\epsilon_i = \delta_{1i} lnK_{i,ph} + \delta_{2i} lnK_{i,h} + \varepsilon_i$. This specification ensures that both the autoregressive component and the effect modifier are exogenous[9]. To make sure that also the physical and the human capitals are exogenous we instrument them using as IVs their 1970 values. Once the $\eta = [\eta_{ph}, \eta_h]$ have been computed, they can be introduced in a control function like (5). The (equilibrium) outcome equation leads to the results presented in Table 3 for the parametric part of the model,

[8]Both the joint depreciation rate and the short-run effect are assumed to have constant returns. The idea behind this formulation is that the joint depreciation rate is constant irrespectively of the state of development of the economy and the same is true for the autoregressive part, because the past effect the present in a rich country as much as it does in a poor country.

[9]The depreciation rate is assumed to be exogenous. On the need to assume an exogenous depreciation rate as a maintained hypothesis for a multi state economy, as the one where returns are function of inequality, see Hulten and Wykoff (1980).

Table 3: Regression Results

	Dependent variable:
	lnY
Y_0	0.7737 ***
	(0.0374)
Dep	-0.3485.
	(0.1888)
Constant	1.4596**
	(0.5322)
Observations	81
Adjusted R^2	0.971
Deviance explained	98.1%
Note:	. p<0.1; *p<0.1; **p<0.05; ***p<0.01

and to the one presented in Table 4 for the varying coefficients.

Table 4: Estimated Varying Coefficients

Values	Minimum	1^{ST} Quantile	Median	Mean	3^{RD} Quantile	Maximum
$g_1(gini_{i,1970})$	0.09446	0.12790	0.23330	0.19590	0.24760	0.36140
$g_2(gini_{i,1970})$	0.06113	0.17610	0.22980	0.22580	0.26700	0.37910

It is interesting to make a comparison of these empirical results with the one obtained for the varying Mincer equation. While in equation (22) the varying intercept was only increasing and the varying slope was only decreasing, in the case of the varying Neoclassical growth model the returns, at least for the human capital, are firstly decreasing and later increasing, which makes the use of a varying coefficients even more compelling irrespectively of the density function of the effect modifier, see Figure (5).

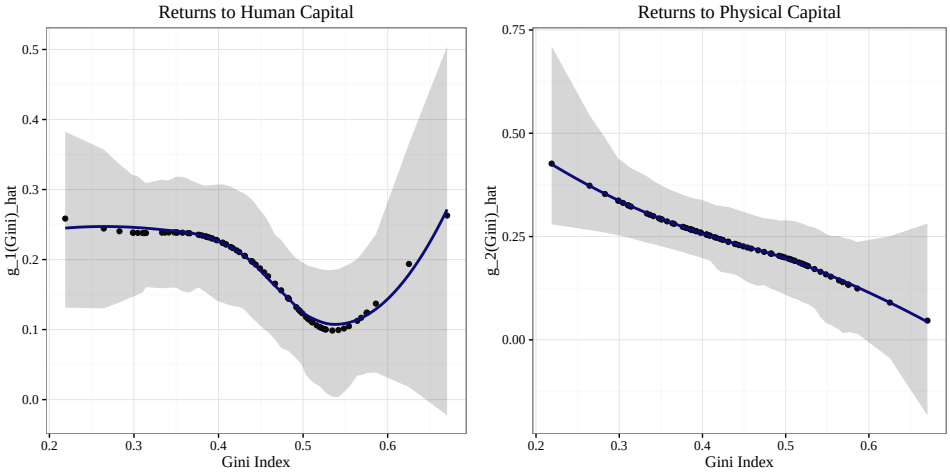

Figure 5: As inequality in a society increases the returns to human capital on per-worker GDP decreases from 0.247 to a minimum of 0.094, when the Gini Index is 0.53, then they start growing again arriving to a maximum of 0.361. To the contrary the returns to physical capital are only decreasing to a maximum of 0.379, when the society is relatively equal to a minimum of 0.061.

5 Conclusions

Recent econometric literature has focused on the possibility for individual returns to be heterogeneous.

In structural models the attempt to micro-found every policy evaluation using a series of (closed form) optimizations has been severely criticized precisely because the representative agent specification does not allow for heterogeneous returns. For example, the pioneering works of Hansen and Singleton (1982, 1983) has been questioned, among others by Summers (1991), because the parameters that characterized the behavior of the representative agent are policy independent for all the sample, implying that every agent has the same response to the policy shock.

The same critique has been advanced in the program evaluation literature. The attempt to reproduce the conclusions of the *ceteris paribus* analysis in a quasi-experimental environment highlights the importance of having an heterogeneous individual treatment effect because the correct computation of the marginal returns of a treatment allows to compute the optimal magnitude of the policy. For example, as notices by Heckman (2010), in order to set the optimal size of a policy it is necessary to compare the returns for the marginal individual and compare it with the marginal cost of the policy, in order to identify the marginal treatment effect.

The VCM represents a good compromise between the idea of having different individual

returns for every cross-section observation or consider the coefficients as fixed.

The semiparametric nature of the VCM is able to capture the non-linear nature of the relationship between the effect modifier and the regressors. At the same time, the presence of the stochastic disturbance in the varying coefficient equation allows for a local level heterogeneity even among agents that have the same level of effect modifier. This last property has the advantage to reduce the level of endogeneity and the distance between the estimated model and the DGP *sive natura*, without using an instrument. In other words, the local level heterogeneity, which does not require to have an interaction between the regressors and the error for the entire hyperplane of the regression, reduces the importance of the instrument in fighting endogeneity. This means that, if, after an explanatory data analysis à la Hastie and Tibshirani (1993), the econometrician suspects that the returns are biased because at least one of the explanatory variables is endogenous, the use of a triangular VCM reduces the level of unexplained heterogeneity at the local level offering a specification that is able to disentangle causal links while producing good estimates of the structural form.

References

[1] Aghion, P., Bolton, P., 1997, A Theory of Trickle-Down Growth and Development. *Review of Economic Studies*, Vol. 64, No. 2, pp 151-172.

[2] Angrist, J., Imbens, G., 1994, Identification and Estimation of Local Average Treatment Effects. *Econometrica*, Vol. 62, No. 2, pp 467-475.

[3] Banerjee, A., Duflo, E., 2003, Inequality and Growth. What Can the Data Say? *Jornal of Economic Growth*, Vol. 8, No. 2, pp 267-299.

[4] Banerjee, A., Newman, F., 1993, Occupational Choice and the Process of Development. *Jornal of Political Economy*, Vol. 101, No. 2, pp 274-298.

[5] Cai, Z., Fang, Y., Lin, M., Su, J., 2011, Semiparametric Varying-coefficient Instrumental Variables Models.

[6] Card, D., 2001, Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems. *Econometrica*, Vol. 69, No. 5, pp 1127-1160.

[7] Darolles, S., Fan, Y., Florens, J.P., Renault, E., 2011, Nonparametric Instrumental Regression. *Econometrica*, 79, No. 5, pp 1541-1565.

[8] Fan, J., Zhang, W., 1999, Statistical Estimation in Varying Coefficient Models. *The Annals of Statistics*, Vol. 27, No. 5, pp 1491-1518.

[9] Fan, J., Zhang, W., 2008, Statistical Methods with Varying Coefficient Models. *Statistics and Its Interface*, Vol. 1(1), pp 179-195.

[10] Hansen, B., 2012, Nonparametric Sieve Regression: Least Squared, Averaging Least Squares, and Cross-Validation.

[11] Hansen, L., Singleton, K., 1982, Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models. *Econometrica*, Vol. 50, No. 5, pp. 1269-1286.

[12] Hansen, L., Singleton, K., 1983, Stochastic Consumption, Risk Aversion, and the Temporal Behavior of Asset Returns. *The Journal of Political Economy*, Vol. 91, No. 2, pp. 249-265.

[13] Hastie, T., Tibshirani, R., 1993, Varying Coefficient Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4), pp. 757-796.

[14] Heckman, J., 2010, Building Bridges Between Structural and Program Evaluation Approaches to Evaluating Policy. *Journal of Economic Literature*, 48(2), pp. 356-398.

[15] Horowitz, J., 2011, Applied nonparametric instrumental variables estimation. *Econometrica*, Vol. 79, No. 2, pp 347-394.

[16] Hulten, C., Wykoff, F., 1980, Economic Depreciation and the Taxation of Structures in United States Manufacturing Industries: an Empirical Analysis. *The Measure of Capital*, University of Chicago Press.

[17] Imbens, G., Newey, W., 2009, Identification and Estimation of Triangular Simultaneous Equations Models without Additivity. *Econometrica*, Vol. 77, No. 5, pp 1481-1512.

[18] Imbens, G., Blundell, R., Newey, W., Persson, T., 2007, Nonadditive Models with Endogenous Regressors. *Advances in Economics and Econometrics*, Boston College Working Papers in Economics, 751.

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

[20] Köhler, M., 2014, Econometric Studies on Flexible Modeling of Developing Countries in Growth Analysis. *PhD Dissertation*, Ch. 4. A Variable-Coefficients Model for Assessing the Returns of Growth Regressions for the Poor and the Rich.

[21] Kuznet, S., 1955, Economic Growth and Income Inequality. *The American Economic Review*, Vol. XLV, No. 1.

[22] Lee, Y.K., Mammen, E., Park, B., 2012, Flexible Generalized Varying Coefficient Regression Models. *The Annals of Statistics*, Vol. 40, No. 3, pp. 1906-1933.

[23] Lemieux, T., 2003, The "Mincer Equation" Thirty Years after Schooling, Experience, and Earnings. *Center for Labor Economics*, University of California, Berkeley, Working Paper N. 62.

[24] Mankiw, G., Romer, D., Weil, N., 1992, A Contribution to the Empirics of Economic Growth. *The Quarterly Journal of Economics*, Vol. 107, No. 2, pp. 407-437.

[25] Mincer, J., 1958, Investment in Human Capital and Personal Income Distribution. *Journal of Political Economy*, Vol. 66, No. 4, pp. 281-302.

[26] Mincer, J., 1974, Schooling, Experience and Earnings. *Columbia University Press*, New York.

[27] Newey, W., 1994, The Asymptotic Variance of Semiparametric Estimators. *Econometrica*, Vol. 62, No. 6, pp. 1349-1382.

[28] Newey, W., Powell, J., 2003, Instrumental Variable Estimation of Nonparametric Models. *Econometrica*, Vol. 71, No. 5, pp. 1565-1578.

[29] Newey, W., Powell, J., Vella, F., 1999, Nonparametric Estimation of Triangular Simultaneous Equations Models. *Econometrica*, Vol. 67, No. 3, pp. 565-603.

[30] Schultz, T., 1997, Human Capital, Schooling and Health, IUSSP, XXIII. General Population Conference, Yale University.

[31] Severance-Lossin, E., Sperlich, S., 1999, Estimation of Derivatives for additive Separable Models. *Statistics: A Journal and Applied Statistics*, Vol. 33, Issue 3, pp. 241-265.

[32] Solow, R., 1956, A Contribution to the Theory of Economic Growth. *Quarterly Journal of Economics*, Vol. 70, No. 1, pp. 65-94.

[33] Sherwing, R., 1992, Distinguished Fellow: Mincering Labor Economics. *Journal of Economic Perspective*, Vol. 6, No. 2, pp. 157-170.

[34] Stoker, T., 1982, The Use of Cross-Section Data to Characterize Macro Functions *Journal of the American Statistical Association*, Vol. 77, No. 378, pp. 369-380.

[35] Summers, L., 1991, The Scientific Illusion in Empirical Macroeconomics. *The Scandinavian Journal of Economics*, Vol. 93, No. 2, pp. 129-148.

[36] Swan, T., 1956, Economic Growth and Capital Accumulation. *Economic Record*, Vol. 32, No. 2, pp. 334-361.

[37] Telser, L., 1964, Iterative Estimation of a Set of Linear Regression Equations. *Journal of the American Association*, Vol. 59, Issue 307, pp. 845-862.

A Appendix

In order to show the statistical properties of the estimated parameters $\hat{\gamma}$ it is helpful to associate them to a family of distributions in the form

$$\hat{\gamma} \to \begin{cases} \Phi = \{F\} & \text{general family of distributions of } r \\ \mu : \Phi \to \mathbb{R}^q & \text{with } q = K + J + L. \text{ If } R_i \text{ has distribution F, then } plim(\hat{\gamma}) = \mu(F) \end{cases}$$

In other words if the parameters are misspecified, they are a function of an unrestricted (except for standard regularity conditions) family of distributions of R, and if they are not misspecified, they become a function of the true distribution F. The limiting behavior of the parameters $\hat{\gamma}$ becomes more comprehensible using the concept of path-wise derivatives. The path-wise derivative of $\mu(F)$ is a $q \times 1$ vector $d(R)$ with $E[d(R)] = 0$ and $E[||d(R)||^2] < \infty$ such that for every path θ the following relation holds

$$\frac{\partial \mu(F_\theta)}{\partial \theta} = E[d(R)S(R)] \tag{29}$$

where $\{F_\theta : F_\theta \in \Phi\}$ is a one-dimensional sub-family of Φ, i.e. a path in Φ which has a probability density function (pdf) dF_θ, and $S(R)$ is the corresponding score function $S(R) = \frac{\partial ln(dF_\theta)}{\partial \theta}$.
The introduction of the concept of path-wise derivatives allows to rethink the family of distribution of $\hat{\gamma}$, the $plim(\hat{\gamma})$ is equal to $\mu(\cdot)$ when the true distribution F_0 is used, i.e. when the path is the one obtained for $\theta = 0$. On this new light it becomes easier to understand what happens to $\hat{\gamma}$ when the path $\theta = 0$ is used.

Consider a general path $\{F_\theta\}$ for a general function $p(\theta) = p(F_\theta)$, while γ is estimated using $\theta = 0$, i.e $\gamma = \gamma_0$. Differentiating the population moment condition for θ gives

$$\begin{aligned} \frac{\partial E_\theta[m(R, \gamma_0, p(\theta))]}{\partial \theta} &= \int \frac{\partial m(R, \gamma_0, p(\theta)) dF_\theta dr}{\partial \theta} \\ &= \int m(R, \gamma_0, p(\theta)) \frac{\partial dF_\theta}{\partial \theta} dr + \int \frac{\partial m(R, \gamma_0, p(\theta))}{\partial \theta} dF_\theta dr \\ &= \int m(R, \gamma_0, p(\theta)) S(r) dr + \int \frac{\partial m(R, \gamma_0, p(\theta))}{\partial \theta} dF_\theta dr \\ &= E[m(R, \gamma_0, p(\theta))S(r)] + E\left[\frac{\partial m(R, \gamma_0, p(\theta))}{\partial \theta}\right] \end{aligned} \tag{30}$$

Assuming that there is an $\alpha(R)$ such that $E[\alpha(R)] = 0$ and $E\left[\frac{\partial m(R, \gamma_0, p(\theta))}{\partial \theta}\right] = E[\alpha(R)S(R)]$ it is possible to re-write equation (30) as

$$\frac{\partial E_\theta[m(R, \gamma_0, p(\theta))]}{\partial \theta} = E[m(R, \gamma_0, p(\theta))S(R)] + E[\alpha(R)S(R)] \tag{31}$$

A simple application of the analytic implicit function transform equation (31) into

$$\begin{aligned}
\frac{\partial \mu(F_\theta)}{\partial \theta} &= -E\Bigg[- \frac{\partial E_\theta[m(R,\gamma_0,p(\theta))]}{\partial \gamma} \mid_{\gamma=\gamma_0} \Bigg]^{-1} \frac{E_\theta[m(R,\gamma_0,p(\theta))]}{\partial \theta} \\
&= -E\Bigg[- \frac{\partial E_\theta[m(R,\gamma_0,p(\theta))]}{\partial \gamma} \mid_{\gamma=\gamma_0} \Bigg]^{-1} \{E[m(R,\gamma_0,p(\theta))S(R)] + E[\alpha(R)S(R)]\} \\
&= E\Bigg\{ \Bigg[\Bigg[\frac{\partial E_\theta[m(R,\gamma_0,p(\theta))]}{\partial \gamma} \mid_{\gamma=\gamma_0} \Bigg]^{-1} [m(R,\gamma_0,p(\theta)) + \alpha(R)] \Bigg] S(R) \Bigg\} \\
&= E[d(R)S(R)] \qquad (32)
\end{aligned}$$

where $d(R) = \left[\frac{\partial E_\theta[m(R,\gamma_0,p(\theta))]}{\partial \gamma} \right]^{-1} [m(R,\gamma_0,p(\theta)) + \alpha(R)]$.
Now it is sufficient to characterize the function $d(R)$ to describe the asymptotic behavior of the estimated parameters, in order to do that it is useful to introduced a theorem

Theorem 2. *(Newey, 1994) Suppose that*

1. *the set of scores for regular paths is linear*
2. *for any $\omega > 0$ any measurable $s(R)$ with $E[s(R)] = 0$ and $E[s(R)]^2 < \infty$ there is a regular path with score $S(R)$ satisfying $E[|s(R) - S(R)|^2] < \omega$*
3. *γ is asymptotically linear and regular*

Then there is $d(R)$ such that $\frac{\partial \mu(F_\theta)}{\partial \theta} = E[d(R)S(R)]$ and $d(R)$ is equal to the influence function $\psi(R)$.

Combining the equation (32) with Theorem 3 it is possible to characterize the asymptotic behavior of the estimated parameters such that

$$\frac{\partial plim(\hat{\gamma})}{\partial \theta} = \frac{\partial \mu(\theta)}{\partial \theta} = E[d(R)S(R)] = E\left[\psi(R) \frac{\partial ln(dF_\theta)}{\partial \theta} \right] \qquad (33)$$