Graphical diagnostics of endogeneity*

Xavier de Luna[a] and Per Johansson[b]

[a] Department of Economics, Umeå University,
S-90187 Umeå, Sweden.
E-mail: xavier.deluna@econ.umu.se

[b] IFAU - Office of Labour Market Policy Evaluation,
S-75120 Uppsala, Sweden.
E-mail: per.johansson@ifau.uu.se

Abstract

We show that in sorting cross-sectional data, the endogeneity of a variable may be successfully detected by graphically examining the cumulative sum of the recursive residuals. An interesting case arises with a continuous or ordered (e.g., years of schooling) endogenous variable. Then, a graphical test for misspecification due to endogeneity (e.g., self selection) can be obtained without instrumental variables. Moreover, the sign of the bias implied by this endogeneity becomes deducible through such graphs.

Keywords: CUSUM plot; Recursive residuals; Return to schooling; Self selection.

JEL: C32, C52.

*The first author would like to acknowledge the Wikström Foundation for its financial support.

1 Introduction

In regression models for continuous responses, all sorts of model misspecifications may be diagnosed by an analysis of ordinary residuals, i.e., from an ordinary least square (OLS) estimator. Endogeneity of variables is a notable exception, however; with linear models, this is due to the OLS estimated parameter being consistent for the reduced form parameter (other than the structural parameter).

The purpose of this paper is to show how recursive residuals associated with a specific ordering of the data can be successfully analyzed in order to diagnose the endogeneity of a variable. In particular, we show that a graphical display of the cumulative sum of the recursive residuals obtained by assuming exogeneity is helpful in diagnosing the presence of endogeneity. Moreover, the sign of the bias implied by this endogeneity (e.g., the direction of a self selection bias) is also deducible through such graphs. The use of recursive residuals was earlier advocated by Harvey and Collier (1977) to test for functional misspecification in regression analysis.[1] In particular, they proposed a t-statistic to test that the residuals have expectation zero. This test is directly applicable to test against endogeneity when the data is sorted adequately. Instruments may be needed to obtain such a sorting.

However, an interesting case arises with a continuous or ordered (e.g., years of schooling) variable whose endogeneity is due to selectivity. Sorting the data with respect to this variable and looking at recursive residuals obtained with a model where exogeneity is assumed allow us to diagnose the misspecification. In this case, endogeneity can thus be diagnosed without specifying a model for the alternative, in contrast with the Hausman test (cf. Hausman, 1978) for which certain instruments are needed.

In Section 2, we start by presenting a framework allowing us to introduce special orderings of the data that are useful for diagnosing endogeneity. Section 3 presents the methodology based on the calculation of recursive residuals associated with a relevant ordering of the data. Graphical displays of these residuals as well as the Harvey-Collier test statistic are proposed to diagnose the endogeneity of a variable. Sections 4 to 6 present different areas of application. Thus, Section 4 looks at a text-book example of endogeneity due to simultaneity of two variables. Section 5 considers Garen's

[1] Endogeneity of a variable is often equivalent to a functional misspecification. For instance, if a random coefficient is associated to a continuous endogenous variable (e.g. Garen's (1984) model), the outcome equation is implicitely non-linear in that variable.

(1984) model of selectivity based on a random coefficient. In particular, a real data set concerning returns to schooling is analyzed in detail. Finally, Section 6 discusses the case of endogenous treatment where the propensity score (Rosenbaum and Rubin, 1983) can be used to sort the data to identify self-selection. The paper is concluded with a discussion in Section 7.

2 Sorting scores for endogeneity

We consider an observational study where independent observations are available for a response y, together with a set of exogenous variables $\mathbf{x}$ and a possibly endogenous variable z (denoted the treatment in the sequel). The following linear statistical model is considered

$$y = \mathbf{x}'\boldsymbol{\beta} + \gamma z + \varepsilon,$$

where ε is a zero mean error term. The exogeneity of $\mathbf{x}$ implies that its marginal density, $p(\mathbf{x}; \boldsymbol{\delta})$, $\boldsymbol{\delta} \in D \subseteq \mathbb{R}^d$, can be ignored without loss of information about $\boldsymbol{\beta}$ and γ (see, e.g., Gouriéroux and Monfort, 1995, Chap. 1.5). Similarly, if treatment z is exogenous, its effect can be studied by the sole specification of $p(y|z, \mathbf{x}; \boldsymbol{\beta}, \gamma)$, that is, the density of the error term. However, in a typical observational study, the exogeneity of treatment z must be assessed.

We propose graphical diagnostics of endogeneity by sorting the data with respect to a sorting score. The sorted data should then not be distinguishable from any other random ordering, only under the exogeneity of the variable of interest.

In general, there is no unique sorting score for a given problem, but certain sorting scores will be more useful than others. In order to present results, we must give a minimal description of the alternative hypothesis of endogeneity. Hence, let us consider an unobserved variable u such that

$$E(\varepsilon|\mathbf{x}, z, u) = u. \tag{1}$$

Let

$$m(\mathbf{x}, z; \boldsymbol{\theta}) = \mathbf{x}'\boldsymbol{\beta} + \gamma z,$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}', \gamma)'$. Under H_0 : "z is exogenous", implying within this framework that $E(u|\mathbf{x}, z) = 0$, we have that $m(\mathbf{x}, z; \boldsymbol{\theta}) = E(y|z, \mathbf{x})$, the unbiased

and optimal (with minimum mean squared error of prediction) predictor of y, given $\mathbf{x}$ and z.

Let us define $\boldsymbol{\theta}_c$ (possibly a function of c) as

$$\boldsymbol{\theta}_c = \arg\min_{\boldsymbol{\theta}} E[\{y - m(\mathbf{x},z;\boldsymbol{\theta})\}^2 | z, \mathbf{x}, s < c],$$

where s is a random variable. We call this variable a *sorting score* when it is such that $\boldsymbol{\theta}_c = \boldsymbol{\theta}$, a constant, only under H_0. Such sorting scores will allow us to identify endogeneity by fitting $m(\mathbf{x}, z; \boldsymbol{\theta}_c)$ recursively to data sorted with respect to s and looking for evidence of a varying parameter $\boldsymbol{\theta}_c$. We now define a class of particularly useful sorting scores.

Definition 1 *A monotone sorting score s for $E(y|\mathbf{x}, z)$ is such that $\forall c \in \Omega_s$,*

$$m(\mathbf{x}, z; \boldsymbol{\theta}_c) \lesseqgtr E(y|z, \mathbf{x}) \text{ for all } \mathbf{x}, z \text{ such that } s > c,$$

when H_0 does not hold.

We start by a preliminary result.

Lemma 1 *If $u = \lambda z$, and hence sorting with respect to u or z is identical, then for $s = u$, $\forall c \in \Omega_s$, $m(\mathbf{x}, z; \boldsymbol{\theta}_c) = E(y|z, \mathbf{x})$ for all $\mathbf{x}, z$.*

Proof. We have that $E(y|z, \mathbf{x}) = \mathbf{x}'\boldsymbol{\beta} + (\gamma + \lambda)z$. Moreover, $\boldsymbol{\theta}_c$ is solution of

$$E(y - m(\mathbf{x}, z; \boldsymbol{\theta}_c)|z, \mathbf{x}, s < c) = 0,$$

$\forall c \in \Omega_s$. But here, $E(y - m(\mathbf{x}, z; \boldsymbol{\theta}_c)|z, \mathbf{x}, s < c) = E(y - m(\mathbf{x}, z; \boldsymbol{\theta}_c)|z, \mathbf{x})$ and $\boldsymbol{\theta}_c = (\boldsymbol{\beta}', \gamma + \lambda)'$. The lemma is then proved. ■

Thus, in the situation of the lemma, $s = u$ is not monotone and will not help us identify endogeneity. In fact, λ and γ are not identifiable in such a situation. On the other hand, let u and z be not proportional but linearly dependent as

$$u = \xi_1 z + \xi_2, \tag{2}$$

where ξ_i, $i = 1, 2$ are random variables with $E(\xi_i|\mathbf{x}, z) = \alpha_i$, $\alpha_1 \neq 0$, and $V(\xi_i|\mathbf{x}, z) \geq 0$, for any z. The latter inequality is assumed to be strict for at least one i, $i = 1, 2$. Furthermore, when $V(\xi_1|\mathbf{x}, z) > 0$, z is assumed to be either always non-negative or always non-positive. Then, it can be shown that u is a monotone sorting score:

Proposition 1 *Let u in (1) be such that under endogeneity of z, u and z are linearly dependent as described by (2). Then, u is a monotone sorting score.*

Proof.

We assume endogeneity of z. We can write

$$E(y|\mathbf{x}, z, u) = \boldsymbol{\beta}'\mathbf{x} + \gamma z + u$$

and

$$E(y|\mathbf{x}, z) = \boldsymbol{\beta}'\mathbf{x} + \gamma z + E(u|\mathbf{x},z),$$

where $E(u|\mathbf{x}, z) \neq 0$. Furthermore, for a constant $c \in \Omega_u$ (the sample space of u),

$$E(y|\mathbf{x}, z, u < c) = \boldsymbol{\beta}'\mathbf{x} + \gamma z + E(u|\mathbf{x}, z, u < c).$$

By the linearity assumption (2), we can write $E(u|\mathbf{x}, z, u < c) = \alpha_2^c + \alpha_1^c z$, where $\alpha_i^c = E(\xi_i|\mathbf{x}, z, u < c)$, $i = 1, 2$, are constants although functions of c. Hence,

$$E(y|\mathbf{x}, z, u < c) = \boldsymbol{\beta}'\mathbf{x} + \gamma z + \alpha_2^c + \alpha_1^c z,$$

which is linear and equivalent to $m(x, z; \theta_c)$. Moreover, because $\alpha_i^c < \alpha_i$ when $V(\xi_i|\mathbf{x}, z) > 0$ (this is true for at least one i), $E(u|\mathbf{x}, z, u < c) < E(u|\mathbf{x}, z)$ for any positive z and $E(u|\mathbf{x}, z, u < c) > E(u|\mathbf{x}, z)$ for any negative z. Thereby the monotonicity of u as sorting score is implied. ■

Even if u is unobserved, this result is of practical use because the ordering of u can be retrieved by studying z and its relation to certain instruments, see the example sections below.

A most convenient case arises when z is a monotone sorting score in itself, since no instrumental variables are then required. This situation can arise when $E(u|z)$ is non-linear in z which is, for instance, the case with random coefficient models; see Section 5.

3 Graphical diagnostics

Graphical diagnostics are informal tools for analysis but, at the same time, a very powerful medium for conveying information. A graph may tell more than the value of a test statistic, although both are obviously complementary. Since ordinary residuals are not really appropriate to identify the endogeneity misspecification, we base our analysis on recursive residuals.

3.1 Recursive residuals

Let a set of independent observations $(y_i, \mathbf{x}_i, z_i)$, $i = 1, \ldots, n$, be generated by a model with corresponding density $p(y|z, \mathbf{x}; \boldsymbol{\beta})$. For each $k = q, \ldots, n-1$, a consistent estimate $\widehat{\boldsymbol{\beta}}_k$ of $\boldsymbol{\beta}$, based on $(y_i, \mathbf{x}_i, z_i)$, $i = 1, \ldots, k$, is assumed to be available. Recursive residuals are then obtained by predicting y_j with $E(y_j|z_j, \mathbf{x}_j; \widehat{\boldsymbol{\beta}}_{j-1})$, $j = q+1, \ldots n$. This prediction is an estimate, based on observations $(y_i, \mathbf{x}_i, z_i)$, $i = 1, \ldots, j-1$, of the optimal (mean squared error sense) predictor $E(y_j|z_j, \mathbf{x}_j; \boldsymbol{\beta})$. The recursive residuals are then standardized prediction errors:

$$w_j = \frac{y_j - E(y_j|z_j, \mathbf{x}_j; \widehat{\boldsymbol{\beta}}_{j-1})}{Var(y_j - E(y_j|z_j, \mathbf{x}_j; \widehat{\boldsymbol{\beta}}_{j-1})|z_j, \mathbf{x}_j)}, \qquad j = q+1, \ldots n.$$

Assuming that the involved moments exist and that the model is well specified, these recursive residuals are, at least asymptotically, independent and identically distributed with mean zero and variance one. These properties hold exactly when $\boldsymbol{\beta}$ is known.

Example 1 *The linear Gaussian model, $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_\mathbf{i}$ with ε_i independently and normally distributed with mean zero and variance σ^2, is an important particular case for which recursive residuals were originally studied, e.g., by Brown et al. (1975). For this model, we have, for $j = q+1, \ldots n$,*

$$w_j = \frac{y_j - \mathbf{x}_j'\widehat{\boldsymbol{\beta}}_{j-1}}{\sigma(1 + \mathbf{x}_j'(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}\mathbf{x}_j)^{1/2}},$$

where $\mathbf{X}_{j-1} = (\mathbf{x}_1', \ldots, \mathbf{x}_{j-1}')'$. Assuming that $\mathbf{X}_{j-1}'\mathbf{X}_{j-1}$ are invertible, w_j are homoscedastic, independent, and with standard normal distribution (Brown et al., 1975). No asymptotic argument is needed here.

Kianifard and Swallow (1996) review the application of recursive residuals, see also Dawid (1984) and de Luna and Johansson (2000).

3.2 Cumulative sum and the Harvey-Collier test

A graphical diagnostic tool is obtained by graphically displaying recursive residuals. Their cumulative sum (CUSUM) is most useful. Asymptotically,

the recursive residuals have mean zero for a well-specified model. When misspecification arises, in our case when exogeneity of the treatment does not hold, the recursive residuals will typically have non-zero mean. If we then sort the data with respect to s, a monotone sorting score, the residuals will have positive (or negative) mean throughout the recursion.

The results of Section 2 can be used to obtain such sortings. In the time series context, the aim when inspecting cumulative sums is to detect a change of the parameter values, see e.g. Brown et al. (1975). Most often this is believed to be an abrupt structural change at a unknown point in time. The endogeneity misspecification is instead translated by small but systematic biases in predictions. Thus, a monotone sorting score should be used for these biases to accumulate instead of cancelling each other out, thereby guaranteeing the best visual effect when plotting the cumulative sum of the recursive residuals. Examples illustrate these issues in the next section. The constancy of the bias sign is also relevant for the test presented below to have power.

Harvey and Collier (1977) proposed a simple test based on the sum of the recursive residuals to identify functional misspecification in a regression model. In our context, write

$$\overline{w} = \frac{1}{n-q} \sum_{i=q+1}^{n} w_i$$

the average of the recursive residuals. Then, under H_0, asymptotically (exactly under the normal model), $\overline{w}$ is normally distributed with mean zero and variance $1/(n-q)$, and constitutes a test statistic for H_0. This test is a necessary complement to the CUSUM plot. Note that a simulation study conducted in de Luna and Johansson (2000) showed the good properties of the Harvey-Collier test in comparison with, for example, a classic Hausman test.

4 Application I: consumption and income

Consider the model for y and z:

$$y = \beta x_1 + \gamma z + \varepsilon,$$

where variable z is endogenous, such that

$$z = \alpha x_2 + \nu,$$

with ε and ν correlated[2]. Denote $\mathbf{x} = (x_1, x_2)$. Assuming $E(\varepsilon|\nu)$ to be linear in ν (e.g., bivariate normality), we have

$$E(y|\mathbf{x}, z) = \beta x_1 + \gamma z + \lambda(z - \alpha x_2), \tag{3}$$

where $\lambda = 0$ if and only if ε and ν are uncorrelated. Here, $s = (z - \alpha x_2)$ is chosen as a sorting score. Note that $\lambda s(x_2, z)$ plays the role of the omitted variable u in (1). The assumptions of Proposition 1 are met when, for instance, x_2 and z are jointly normal[3] (so that $E(x_2|z)$ is linear in z), in which case we can say that s is a monotone sorting score. It can be approximated by $(z - \widehat{\alpha} x_2)$, where $\widehat{\alpha}$ is a consistent estimate. Note that $x_2 = x_1$ would lead to the non-identifiability of γ, and the non-applicability of Proposition 1.

Example 2 *We use data on U.S. consumption expenditures (c_t), disposable income (y_t) and government expenditure (g_t), in billions of 1982 dollars, for year t between 1975-1986.*[4] *We assume:*

$$c_t = \gamma_0 + \gamma_1 y_t + \varepsilon_t,$$

where y_t is endogenous such that

$$y_t = \alpha_0 + \alpha_1 g_t + \nu_t. \tag{4}$$

Figure 8 shows how the endogeneity of y_t is revealed by using the residuals from (4) as a sorting score.

[2]Classical examples include: i) y is a quantity of goods and z its price, ii) y is a consumption measure and z disposable income.

[3]This assumption might seem restrictive, but is only needed to ensure that Proposition 1 can be applied. It should be observed, however, that the linearity of $E(s|z)$ is not a necessary condition for monotonicity.

[4]This data is described in Hill, Griffiths and Judge (1997) and obtained from http:\\www.wiley.com.

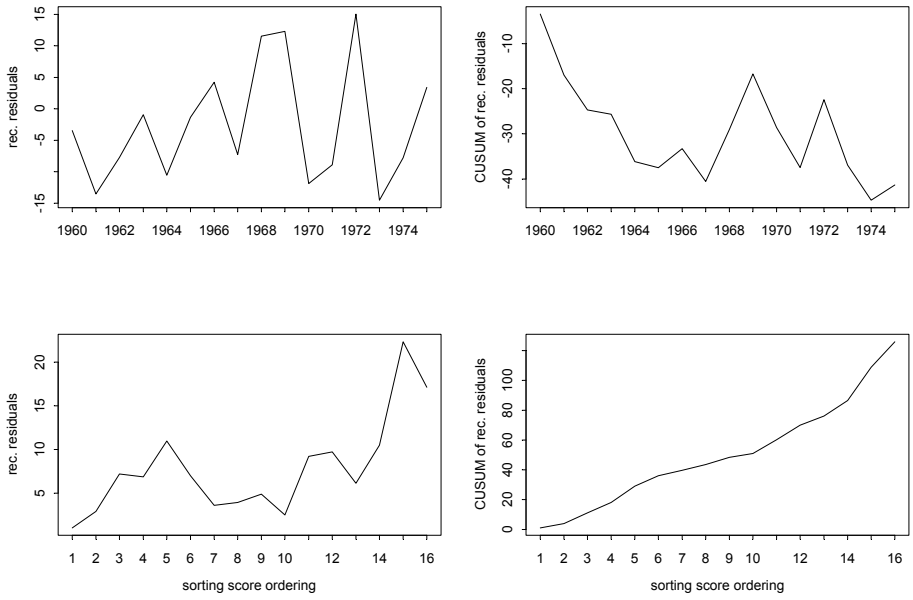


Figure 1: Recursive residuals and their cumulative sum when regressing the consumption expenditures on disposable income (y_t), using time ordering (above panels) $-$HC$= -1.09$; and the ordering using the sorting score: residuals of regressing y_t on the government expenditure variable (below panels) $-$HC$= 5.67$.

5 Application II: return to schooling

5.1 The Garen model

We now consider the selectivity model proposed by Garen (1984, 1988)

$$\begin{aligned} y &= \mathbf{x}'\boldsymbol{\beta} + z\delta + zr + \varepsilon, \\ z &= f(\mathbf{x}^*) + \nu, \end{aligned} \tag{5}$$

where $E(\varepsilon|\mathbf{x}^*, z, r) = 0$ and $\mathbf{x}^*$ contains all the variables in $\mathbf{x}$ and possibly others.[5] For this model,

$$E(y|\mathbf{x}^*, z) = \mathbf{x}'\boldsymbol{\beta} + z\delta + zE(r|\nu). \tag{6}$$

Assuming $E(r|\nu)$ to be linear in ν (e.g., bivariate normality), we have $E(r|\nu) = \lambda(z - f(\mathbf{x}^*))$. Exogeneity of z corresponds to $\lambda = 0$, i.e. uncorrelated r and ν variables. Here, heteroscedasticity is present even if z is exogenous:

$$V(y|\mathbf{x}^*, z) = z^2V(r|\nu) + \sigma^2, \tag{7}$$

where σ^2 is the variance of ε. In this case, neglecting the endogeneity of z leads to a misspecification of the conditional expectation, by assuming it to be linear while $zE(r|\nu)$ is non-linear in $\mathbf{x}^*$ and z.

In this example, not accounting for the endogeneity of z corresponds to omitting the variable[6] $z(z - f(\mathbf{x}^*))$ in (6) which corresponds to u in (1). The non-linearity of $E(y|\mathbf{x}, z)$ may often be hidden by the heteroscedastic noise when examining conventional residuals. On the other hand, recursive residuals can often identify the systematic bias in predictions obtained with the sorting score $s = z(z - f(\mathbf{x}^*))$[7] as illustrated in Example 3. Because f is unknown, an approximate sorting score must be used to estimate this function, yielding $z(z - \widehat{f}(\mathbf{x}^*))$. Notice that within this framework**,** it is possible to proceed without specifying a parametric form for f but instead using a non-parametric estimate.

[5]Garen also considered a pure random effect, i.e. $y = \mathbf{x}'\boldsymbol{\beta} + z\delta + zr + \eta + \varepsilon$, with $E(y|x^*, z) = x'\beta + z\delta + zE(r|\nu) + E(\eta|v)$ and $E(\eta|v) = \rho v$. Here, we omit η for clarity.

[6]The omitted variable is a sorting score since $\lambda = 0$ under exogeneity.

[7]Note that Proposition 1 does not apply here since $E(s|z)$ is not linear in z. However, $E(s|z)$ is quadratic in z and therefore**,** fitting a linear model in z leads to a systematic under-prediction of y and hence**,** s is a monotone sorting score.

Remark 1 *An important property of the diagnostics (CUSUM plots and Harvey-Collier test) is that they have the correct size as soon as* $\lambda = 0$*, even if the random coefficient* r *exists. This is to be contrasted with a classical Hausman type test (typically a conventional test on the residuals of the regression of* z *against instruments, when introduced in the outcome equation) where the null hypothesis is* $r \equiv 0$ *and which therefore has power even against the mere existence of* r*.*

Remark 2 *The Garen model is an econometric translation of a theoretical proposition saying that individuals maximize their present value of future returns. Thus, stating further that individuals know their own coefficient* r*, we should observe a positive correlation between* r *and* z *(endogeneity). If not taken into account, this leads to an underprediction of individual returns to schooling for increasing* z*'s.*[8] *In other words, the theory which has inspired the Garen model also predicts that* z *is a monotone sorting score, leading to recursive residuals with a positive mean.*

Example 3 *The Garen model is considered, and we simulate 100 observations with the following specifications: for* $i = 1, \ldots, 100$*,*

$$\begin{aligned} y_i &= 1 + 2x_{1i} + \gamma_i z_i + \varepsilon_i, \\ \gamma_i &= 1 + r_i, \\ z_i &= x_{1i} - x_{2i} + \nu_i, \end{aligned}$$

with $x_{1i} \sim U(0,1), x_{2i} \sim U(0,1), \varepsilon_i \sim N(0,1)$ *and* r_i *and* ν_i *bivariate normal with expectations zero, variances* 0.36 *and* 1 *respectively, and correlation* -0.5*. Assuming exogeneity* $E(y_i|x_{1i}, z_i) = x_{1i}\beta + z_i\gamma$ *is estimated with OLS. Several types of residual analyses are presented in Figures 2 and 3.*

¿From the residuals plots of Figure 2**,** there seems to be no severe heteroscedasticity. Identifying the misspecification of the conditional mean is not straightforward with these residual plots, although a trained eye may see some structure in the OLS residuals when sorted with respect to the omitted

[8]Using the full sample OLS estimator on (5) would, of course, lead to a positive biased estimate of the mean return to schooling. Here**,** we rather discuss the individual's return to schooling, when sorting with respect to schooling, recursively estimating the model with OLS, and thereafter performing out of sample predictions using this previous OLS estimate.

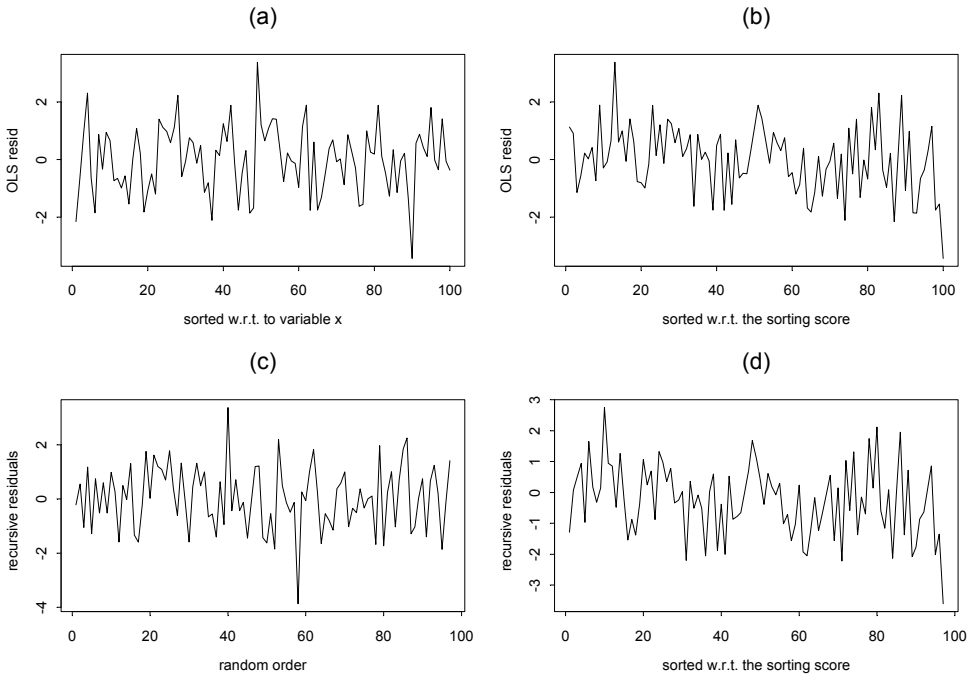


Figure 2: Residuals from the Garen model of Example 7: (a) OLS residuals sorted w.r.t. the variable z; (b) OLS residuals obtained with the ordering of the omitted variable (optimal sorting score); (c) Recursive residuals obtained with a random ordering; (d) Recursive residuals obtained with the optimal sorting score.

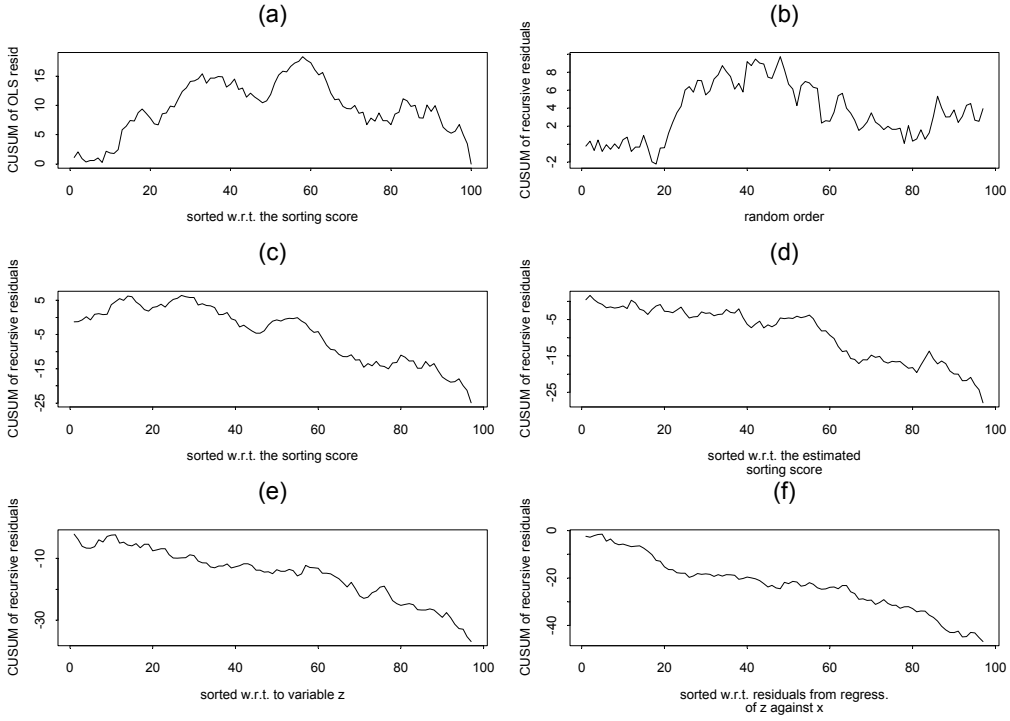


Figure 3: CUSUM plots of various residuals from the Garen model of Example 2: (a) OLS residuals sorted w.r.t. the sorting score $z_i(z_i - f(\mathbf{x}_i))$ $-\mathrm{HC}(r_i \equiv 0) = 0.00$; (b) Recursive residuals obtained with a random ordering $-\mathrm{HC}(r_i \equiv 0) = 0.40$; (c) Recursive residuals obtained with the sorting score $z_i(z_i - f(\mathbf{x}_i))$ $-\mathrm{HC}(r_i \equiv 0) = -2.53$; (d) ditto but with an estimate of the previous sorting score $-\mathrm{HC}(r_i \equiv 0) = -2.83$; (e) ditto but with the sorting score $s(x, z) = z$ $-\mathrm{HC}(r_i \equiv 0) = -3.75$; (f) ditto but with the OLS residuals from the regression of z_i on x_i as the sorting score $-\mathrm{HC}(r_i \equiv 0) = -4.74$.

variable, graph (b), and in the recursive residuals obtained with this same sorting score, graph (d). The CUSUM plots in Figure 3 are more interesting. We note that the recursive residuals obtained with well chosen sorting scores all provide a clear sign of the misspecification of the conditional mean of the model (endogenous treatment) by displaying a systematic departure from zero for the CUSUM trajectory. This neat visual effect is due to the monotonicity property. Note that, as in Section 4, the residuals of the selection equation as well as the endogenous variable itself also seem to provide monotone sorting scores, see the bottom panel of Figure 3. The values of the HC test (for $H_0 : r_i \equiv 0$), given in the caption of the figure, confirm the visual impression.

5.2 U.S. data on return to schooling

The data set[9] analyzed in this section was used by Angrist and Krueger (1991) to study the effects of compulsory school attendance, see also Angrist et al. (1999). It consists of a sample of 329 500 men born in 1930-39 from the 1980 US census. This data set will help us illustrate the kind of insights a graphical display of CUSUM recursive residuals can yield when investigating the exogeneity of a covariate.

The linear model of interest tries to explain the log weekly wage by the number of schooling years, while controlling for an age effect (assumed to be exogenous). Schooling systems differ between states, see Angrist and Krueger (1991, Appendix 2), and for that reason, we perform state-specific analyses. As argued in the previous section, the explanatory variable describing school attendance can be used as sorting score to check its endogeneity which is predicted by the theory. We discard individuals with zero to eleven years of education, in order to avoid effects due to compulsory schooling laws.[10] In particular, the compulsory schooling period should not suffer from a selection bias.

Recursive residuals are computed starting from 13 years of education, 1-12 years cases serving as starting values, together with one individual with 13 years to allow for estimability.[11] In a sense, the question of interest is whether

[9]The data set is available at the following address: http://qed.econ.queensu.ca/jae/

[10]Compulsory schooling laws may in some instances push students to complete a high school degree, see Angrist and Krueger (1991, pp. 1004-1005).

[11]In this application, we have multiple observations for a given number of years of eduction. These are left in their original ordering.

the return to education remains constant after 12 years of education, which most often corresponds to the completion of a high school degree. Recursive residuals are not only useful for diagnosing whether years of schooling are endogenous (selection bias) but also indicate the sign of the selection bias when present, see Remark 2. This is illustrated by the comments below.

Figure 4 displays CUSUM plots of recursive residuals obtained for California, Kansas, New York and Louisiana. We use the Californian case for our main comments: The CUSUM plot indicates that there is actually no selection bias up to year 15 of education (in agreement with Angrist and Krueger's (1991) empirical findings, where they compared OLS and two-stage LS estimates). At year 16 (most often the completion of a University degree) there seems to be a positive selection bias, however. Indeed, although the HC value (1.02) is not significant at this stage, the clear upward trend observed for students with 16 years of education is convincing enough (such a trend is clear in 31 states out of 50; examples include Kansas and New York in Figure 4, while Louisiana is a counter-example). The non-significance of the test is most surely due to the fact that many of the recursive residuals are consistent with the zero mean hypothesis (those corresponding to 13 to 15 years of education). Finally, years 17 to 20 (most often postgraduate studies) do not seem to be rewarded at the same rate as previous years in terms of log wages since there is a strong negative selection bias (over predictions are observed); here, HC is significant (this downward pattern for postgraduate studies is observed in 39 states out of 50). The final HC value is seldom significant unless, as for California and New York, a large number of individuals is available. This is due to the use of a non-monotone sorting score:[12] recursive residuals being not biased (up to year 15) -no selection bias-, then positively biased (year 16) -positive selection bias- and finally downwardly biased (years 17 to 20) -negative selection bias-.

FIGURE 4 CAN BE FOUND AT THE
END OF THIS DOCUMENT

[12]Years of schooling, although not monotone, is an interesting sorting score. The empirical evidence against its monotonicity is actually interesting per se, because theory predicts monotonicity, more precisely positively biased recursive residuals.

Figure 4: Vertical bars indicate years of education; for instance, residuals before the first bar correspond to 13 years, residuals between the first and the second bar are for individuals with 14 years of education and so on, up to 20 years. HC values are: -2.13 for California, 0.73 for Kansas, -5.23 for New York and -1.34 for Louisiana.

6 Application III: self-selection into programs

The standard endogenous treatment model (cf. Heckman, 1978) is such that the choice is described by

$$\begin{aligned} z^* &= \mathbf{x}^{*\prime}\boldsymbol{\alpha} + \varepsilon_1 \\ z &= I(z^* > 0), \end{aligned} \tag{8}$$

and the outcome equation is:

$$y = \mathbf{x}'\boldsymbol{\beta} + z\delta + \varepsilon_2. \tag{9}$$

where z^* is an unobserved latent variable and endogeneity implies that ε_1 and ε_2 are correlated. If ε_1 and ε_2 are bivariate normal and correlated**,** we have that $E(\varepsilon_2|\mathbf{x}^*, z) = \rho\sigma_2^{-1}(\lambda z - \widetilde{\lambda}(1-z))$, where $\lambda = \phi(\mathbf{x}^{*\prime}\boldsymbol{\alpha})/(1-\Phi(\mathbf{x}^{*\prime}\boldsymbol{\alpha}))$ and $\widetilde{\lambda} = \phi(\mathbf{x}^{*\prime}\boldsymbol{\alpha})/\Phi(\mathbf{x}^{*\prime}\boldsymbol{\alpha})$. We assume that $\mathbf{x}^*$ contains at least one variable not included in $\mathbf{x}$.

Assuming joint normality of ε_1 and ε_2, and denoting $\sigma_1^2 = 1$, σ_2^2, ρ, their respective variances and correlation,

$$E(y|\mathbf{x}^*, z) = \mathbf{x}'\boldsymbol{\beta} + \rho\sigma_2^{-1}[\lambda z - \widetilde{\lambda}(1-z)]. \tag{10}$$

The last term in this equation corresponds to the unobserved variable u in (1). In this case**,** the hypotheses of Proposition 1 are fulfilled[13] and u is therefore a monotone sorting score. Here**,** the missing variable is not observed but can be evaluated by using a consistent estimator $\widehat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$.

Note that sorting with respect to $\lambda z - \widetilde{\lambda}(1-z)$ is equivalent to first sorting the sub-sample for which $z = 0$ with respect to $\Phi(x^{*\prime}\alpha) - 1$, or equivalently**,** with respect to $\Pr(z = 1|\mathbf{x}^*) = \Phi(\mathbf{x}^{*\prime}\alpha)$, followed by the sub-sample with

[13]That is**,** u can be rewritten as (2).

$z = 1$, which is also sorted with respect to $\Phi(\mathbf{x}^{*\prime}\alpha)$. $\Pr(z = 1|\mathbf{x}^*)$ is called the propensity score.[14] More generally, i.e. without the need to specify the relation between (8) and (9), the theory often predicts that the propensity score is a monotone sorting score (in that individuals maximize the expected return from the treatment, see e.g. Heckman and Robb 1986) allowing us to test this prediction.

When there are more than two possible exclusive treatments, m say, the outcome can be written as

$$y = \mathbf{x}'\boldsymbol{\beta} + \sum_{k=1}^{m} \delta_k z_k + \varepsilon,$$

where $z_{ki} = 1$ if individual i takes treatment k, $k = 1, \ldots, m$, and zero otherwise. Then,

$$s = \sum_{k=1}^{m} z_k \Pr(z_k = 1|\mathbf{x}^*)$$

is a sorting score under the stringent model assumptions of Lee (1983). More generally, treatments can be compared in pairs, e.g., against the non-treatment class, by using data concerning only two such treatments and then proceeding as in the above binary choice situation.

Finally, when there is a natural order and meaningful numbers can be assigned to treatments, then the situation is similar to a continuous treatment z and, for instance, the Garen (1984) model may be used, as in the case study of Section 5.2. Other models are reviewed in Vella (1998), where a control function is always provided, often in the form of generalized residuals. This control function corresponds to the unobserved variable u and will often provide a useful sorting score.

7 Discussion

In this paper, a graphical analysis of the recursive residuals associated with a sorting of the data has been advocated as a tool for diagnosing endogeneity.

[14] Note that sorting the whole sample with respect to the propensity score does not yield exactly the same sorting. In the latter case, the two sub-samples defined by $z = 0$ and 1 will generally not be fully separated by the sorting, since a non-treated individual may in fact have a similar, and indeed even higher, propensity to be treated than one who is actually treated.

We expect practitioners to find this type of analysis a useful complement to existing tests for exogeneity. A major application area arises when the endogenous variable is continuous or ordered. Indeed, it is then possible to test against endogeneity *without* instrumental variable, by sorting the data with respect to the endogenous variable and looking at the residuals obtained from recursively fitting the outcome equation through the sorted data set. An interesting by-product is that in case of endogeneity, the direction of the bias implied by the endogenous variable is directly available from the CUSUM plot of the recursive residuals**,** as illustrated with the U.S. data on returns to schooling.

When instruments are available**,** our approach is complementary to Hausman-type tests by providing an appealing graphical diagnostic tool. The proposed Harvey-Collier test has, moreover, the advantage of having no power against the presence of a random coefficient in front of an exogenous variable, see Remark 1.

Monotone sorting scores have been emphasized because they ensure the best power when looking at CUSUM plots of recursive residuals. However, as soon as endogeneity implies non-linearity of the conditional expectation, e.g. random coefficient models or non-normality of the error term in the regression equation for the endogenous variable, then any sorting, even a random sorting, may allow the analyst to diagnose endogeneity (by identifying the non-linearity). In this case, even ordinary least squares residual**s** may be sufficient. This is, however, far from certain because this non-linearity is often weak and the residuals heteroscedastic. In this article, we have shown how recursive residuals associated with a monotone sorting score may overcome this difficulty.

8 Reference

Angrist J.D. and Krueger A.B. (1991). "Does Compulsory School Attendance Affect Schooling and Earnings?," *Quarterly Journal of Economics*, **CVI**, 979-1014.

Angrist J.D., Imbens G.W. and Krueger A.B. (1999). "Jackknife Instrumental Variables Estimation," *Journal of Applied Econometrics*, **14**, 57-67.

Brown, R.L., Durbin, J. and Evans, J. M. (1975). "Techniques for Testing

the Constancy of Regression Relationships over Time (with Discussion)," *Journal of the Royal Statistical Society Series B*, **37**, 149-192.

Dawid, A.P. (1984). "Statistical Theory: The Prequential Approach," *Journal of the Royal Statistical Society Series A*, **147**, 278-292.

de Luna, X. and Johansson, P. (2000). "Testing Exogeneity in Cross-section Regression by Sorting Data," *IFAU Working Paper* **2000:2**.

Garen, J. (1984). "The returns to Schooling: A Selectivity Bias Approach with a Continuous Choice variable," *Econometrica*, **52**, 1199-1218.

Garen, J. (1988). "Compensating Wage Differentials and Endogeneity of Job Riskiness," *Review of Economics and Statistics*, **70**, 9-16.

Gouriéroux, C. and Montfort, A. (1995). *Statistics and Econometric Models*, Cambridge University Press. Cambridge.

Harvey, A. and Collier G. (1977). "Testing for Functional Misspecification in Regression Analysis," *Journal of Econometrics*, **6**, 103-119.

Hausman, J.A. (1978). "Specification Test in Econometrics" *Econometrica*, **46**, 1251-1271.

Heckman, J.J. (1978). "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica*, **46**, 931-959.

Heckman J.J. and Robb R. (1986). "Alternative Identifying Assumptions in Econometric Models of Selection Bias," in: H. Wainer, ed., *Drawing infernce from Self selected samples*. Springer-Verlag, Berlin, Germany, **63-107**.

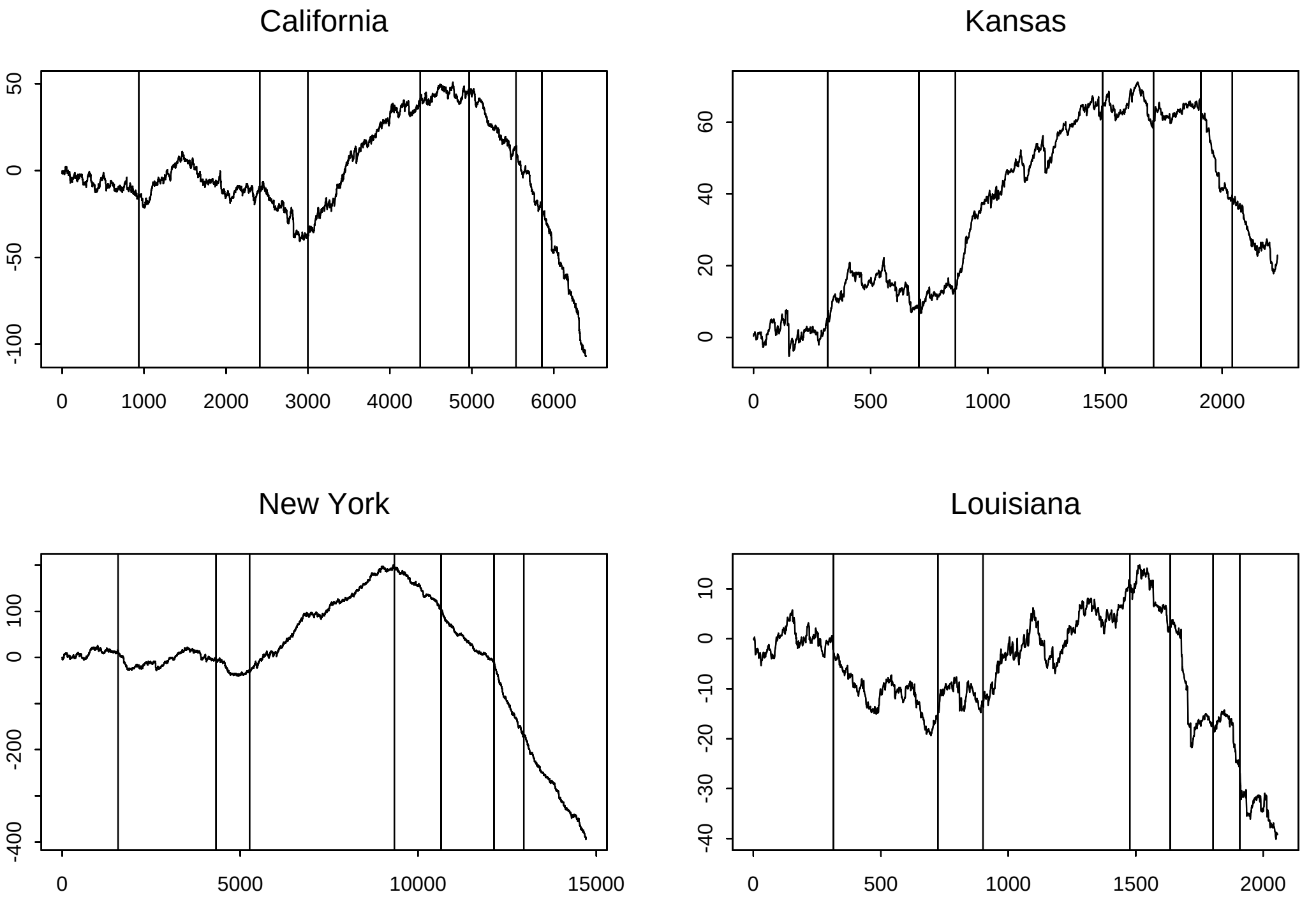
Hill, R.C., Griffiths, W.E. and G.G. Judge (1997). *Undergraduate Econometrics*, John Wiley & Sons, Inc. New York.

Kianifard, F. and Swallow W.H. (1996). "A Review of the Development and Application of Recursive Residuals in Linear Models," *Journal of the American Statistical Association*, **91**, 391-400.

Lee, L.-F., (1983). "Generalized Econometric Models with Selectivity," *Econometrica* **51**, 507-512.

Rosenbaum, P.R. and Rubin, D.B. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effect," *Biometrika*, **70**, 41-55.

Vella, F. (1998). "Estimating Models with Sample Selection Bias: A Survey," *Journal of Human Resources,* **38,** 127-169.


California
50
0
-50
-100
0
1000
2000
3000
4000
5000
6000
Kansas
60
40
20
0
0
500
1000
1500
2000
New York
100
0
-200
-400
0
5000
10000
15000
Louisiana
10
0
-10
-20
-30
-40
0
500
1000
1500
2000