Within-Groups Wage Inequality and Schooling: Further Evidence for Portugal

Corrado Andini *

University of Madeira, CEEAplA and IZA

ABSTRACT

This paper provides further evidence on the positive impact of schooling on within-groups wage dispersion in Portugal, using data on male workers from the 2001 wave of the European Community Household Panel. The issue of schooling endogeneity is taken into account by using the newest available instrumental-variable technique for quantile regression, i.e. the control-function estimator due to Lee (forthcoming, 2007). The findings are compared with earlier results based on different techniques, i.e. the instrumental-variable estimator due to Arias, Hallock and Sosa-Escudero (2001) and the standard exogeneity-based estimator due to Koenker and Bassett (1978).

Keywords: Endogeneity, Quantile Regression, Schooling, Wage Inequality.
JEL Classification: I21, J31, C29.

* Please send correspondence to: Corrado Andini, Universidade da Madeira, Campus da Penteada, 9000-390 Funchal, Portugal. Tel.: +351 291 70 50 53, Fax: +351 291 70 50 40, Email: andini@uma.pt, and Website: http://www.uma.pt/andini.

I gratefully acknowledge financial support from the European Commission (EDWIN Project, HPSE-CT-2002-00108). For valuable comments, I would like to thank Santiago Budría, Joop Hartog, Pedro Telhado Pereira, and Robert Wright. The usual disclaimer applies.

I. INTRODUCTION

Many authors agree on the argument that schooling has a positive impact on within-groups wage inequality in Portugal[1]. The evidence is generally based on the standard quantile-regression techniques due to Koenker and Bassett (1978), applied to several types of wage equations. As a matter of example, Hartog, Pereira and Vieira (2001), Machado and Mata (2001) as well as Martins and Pereira (2004) use different specifications of a Mincerian model and find that the return to schooling is increasing along the conditional earnings distribution. Particularly, the return at the ninth decile seems to be significantly higher than the return at the first decile.

Apart from a recent contribution by Andini (forthcoming, 2008), a common feature of the quantile-regression studies on schooling and within-groups wage inequality in Portugal is that no one deals with the endogeneity of schooling which is, typically, a relevant issue in studies focusing on the impact of schooling on the mean of the conditional earnings distribution.

Likewise the ordinary-least-squares estimator, the estimator of Koenker and Bassett (1978) assumes residuals' orthogonality[2]. Therefore, disregarding the endogeneity of schooling may imply inconsistent estimates of its coefficient along the conditional wage distribution, i.e. the estimated impact of schooling on within-groups earnings dispersion may be misleading.

[1] This section partly borrows from Andini (forthcoming, 2008).

[2] Consider the regression model $Y = \beta X + \varepsilon$. The estimator of Koenker and Bassett is consistent at each quantile θ of the conditional distribution of Y if the orthogonality condition holds at each quantile, i.e. $\mathrm{Quant}_{\theta}(\varepsilon_{\theta}|X) = 0 \quad \forall\theta$, while the consistency of the ordinary-least-squares estimator requires that the conditional mean of regression residuals is null, i.e. $E(\varepsilon|X) = 0$.

A further common feature of existing studies is that no one deals with the total impact of schooling on within-groups wage inequality in Portugal, meaning that previously-estimated wage equations do not exclude schooling-dependent covariates, such as industry dummies[3] and labour-market experience[4], from the list of regressors.

To be more precise about the meaning and the relevance of the concept of total return to schooling, let us present the following simple example. First, let us label the logarithm of hourly earnings as $\ln w$, the number of schooling years as s, the so-called potential labour-market experience as $\exp = \text{age} - s - 6$. Second, let us suppose that a wage equation includes potential labour-market experience as control-variable, i.e. $\ln w = \beta_0 + \beta_1 s + \beta_2 \exp + \varepsilon$.

In the latter case, the coefficient of schooling years β_1 does not capture the total effect of schooling on earnings for the simple reason that the total effect is given by $\frac{\partial \ln w}{\partial s} = \beta_1 - \beta_2$.

Although in our simple example one can recover the total effect of schooling by subtracting the estimate of β_2 from the estimate of β_1, it is worth stressing that, when the mathematical law of the schooling-dependent covariate is unknown (i.e. the general case), the recovering exercise becomes impossible. Hence, if the objective of the

[3] Jobs in some industries may require more years of schooling than jobs in other industries.

[4] Pereira and Martins (2004) properly argue that in order "to obtain the full impact of education on wages, one should be careful not to include in the wage equation covariates whose value can depend on education. In the extreme case one should only regress the ln(wage) in education." (p. 526). See also Andini (forthcoming, 2007).

empirical analysis is the total return to schooling, it is important to exclude schooling-dependent covariates from the wage equation.

Following Andini (forthcoming, 2008), this paper tackles a controversial issue. On the one hand, if the potential correlation between residuals and schooling is limited by the insertion in the wage equation of a large set of control-variables, standard quantile-regression techniques are likely to consistently estimate the coefficient of schooling along the conditional wage distribution but unlikely to recover the total returns to schooling due to the likely presence of schooling-dependent covariates among the control-variables. On the other hand, if the set of control-variables excludes schooling-dependent covariates and only includes strictly exogenous regressors, then standard techniques may poorly estimate the total impact of schooling on within-groups wage inequality due to the likely correlation between residuals and schooling. This paper aims at presenting *consistent* estimates the *total* return to schooling along the conditional distribution of Portuguese wages and compares the results with previous findings by Andini (forthcoming, 2008).

II. EMPIRICAL MODEL

Andini (forthcoming, 2008) finds that the difference between the total return at the ninth decile and the total return at the first decile of the conditional earnings distribution in Portugal is likely to range between 4.2% estimated using the standard quantile-regression techniques due to Koenker and Bassett (1978) and 26.2% estimated using the instrumental-variable approach due to Arias, Hallock and Sosa-Escudero (2001). In this paper, we provide further evidence on the total impact of schooling on within-groups wage inequality using the latest available instrumental-variable technique for quantile regression, i.e. the control-function estimator due to Lee (forthcoming, 2007).

To explain the rationale behind our estimation procedure, let us suppose that a variable Y is explained by an exogenous variable X and an endogenous variable S. Further, let us assume that there is an interest in estimating the impact of S (not only on the mean but also) on the shape of the conditional distribution of Y, controlling for X and using quantile-regression techniques. The control-function approach due to Lee (2007, forthcoming) is a two-stage approach similar to the one adopted by Arias, Hallock and Sosa-Escudero (2001, AHS henceforth). The first-stage regression consists of an ordinary-least-squares regression of S on X and instrumental variables Z1, Z2, and so on. The difference between the estimator of AHS and the estimator of Lee is related to the second stage. The method of AHS replaces the actual values of S with first-stage predicted values of S. In contrast, Lee adds a polynomial function of first-stage residuals to the set of the second-stage regressors, formed by the actual values of S and X.

Both approaches aim at providing consistent estimates of the coefficient of the endogenous explanatory variable at several quantiles of the conditional distribution of the dependent variable, using the estimator of Koenker and Bassett (1978, KB henceforth) in the second stage. And, both approaches suffer from the typical loss of efficiency associated with the implementation of instrumental-variable techniques in finite samples, but the method of Lee turns out to be *more efficient* than the method of AHS, in our specific application, at both the first and the ninth decile of the conditional wage distribution. Note that the latter point has special relevance for the issue of recovering a reliable measure of the impact of schooling on within-groups wage inequality.

Based on the empirical specification proposed by Andini (forthcoming, 2008), the estimation procedure of this paper is as follows:

$$(1) \qquad \ln w_i = \upsilon_\theta + \beta_\theta s_i + \sum_j \delta_{\theta j} yb_{ji} + \sum_c \lambda_{\theta c} \hat{\eta}_i^c + \mu_{\theta i}$$

$$s_i = \nu + \sum_j \alpha_j yb_{ji} + \sum_k \phi_k qb_{ki} + \eta_i$$

$$\hat{s}_i = E(s_i | yb_{ji}, qb_{ki})$$

$$\hat{\eta}_i = s_i - \hat{s}_i$$

$$\text{with } \mathrm{Quant}_\theta(\mu_{\theta i} | s_i, yb_{ji}, \hat{\eta}_i) = 0 \text{ for each } \theta .$$

Our notation is as follows: letter i refers to the i-th individual in the sample, $\ln w$ stands for the logarithm of gross hourly earnings, s measures years of schooling, yb is an indicator-variable for year of birth with j going from 1937 to 1984 (year 1936 is the excluded category), qb is an indicator-variable for quarter of birth with k going from 1 to 3 (quarter 4 is the excluded category), θ is a quantile-indicator going from 5 to 95, c represents the order of the control-function.

Data are extracted from the latest available wave of the European Community Household Panel (ECHP), the 2001 wave, and are related to Portuguese male workers. Summary sample statistics are reported in Table 1.

III. RESULTS

First-stage regression results are presented in the Appendix. Specifically, as already discussed by Andini (forthcoming, 2008), the use of a full set of quarters of birth as instrumental variables is not entirely satisfactory because the F-test of excluded instruments does not reject the null (p-value 0.1972; this result seems driven by the third quarter). It is also known, however, that passing the F-test should not be intended as a strict requirement due to the limitations of the test itself, i.e. low power (see Cruz and

Moreira, 2005). This is particularly true when the Sargan test of over-identification is passed (p-value 0.1351), i.e. the model specification is not rejected. In addition, our results are consistent with the findings of Angrist and Krueger (1991) who argue that people born later in the year have slightly more schooling than people born earlier. Note, indeed, that the estimated coefficients for the first, the second and the third quarter of birth are negative and that the excluded category is the fourth quarter of birth. Further, robustness checks highlight that, if the model is just-identified using the fourth quarter of birth only, the estimated coefficient for the last quarter of the year has the expected positive sign and is statistically significant (p-value 0.055). Finally, if only the last two quarters of the year are used as instruments, the estimated coefficients are both positive and the F-test of excluded instruments is passed (p-value 0.096).

It is worth stressing that our estimation results for the wage equation are robust to the identification strategy. As a matter of example, the AHS method provides roughly the same results when one uses the fourth quarter of birth as single instrument for schooling years rather than a full set of indicator-variables for quarters of birth as in the Appendix. Hence, for sake of comparison with our previous estimates of the total return to schooling along the conditional wage distribution, we adopt the same identification strategy as implemented by Andini (forthcoming, 2008).

Figure 1 plots the estimates of the main parameter of interest, i.e. the estimates of β in model (1) which are, in turn, reported in Table 2. Note that the last two columns of Table 2 also report the AHS results and the standard KB results presented by Andini (forthcoming, 2008).

Regarding the choice of the order of the control-function, Lee (2004) does not suggest the existence of an optimal order because Monte-Carlo experiments in finite samples show that "the estimator is not very sensitive to the choice of the order of the polynomial approximations" (p. 15). Indeed, our empirical findings based on a number of different orders of the control-function are consistent with the simulation results. As a matter of example, we report estimates of β based on c going from 1 to 4. The corresponding estimates of the total return to schooling along the conditional wage distribution are labelled as L1, L2, L3, L4.

Depending on the order of the control-function, the total impact of schooling on within-groups wage inequality varies from 5.1% to 7.4% and fits the interval provided by Andini (forthcoming, 2008), who refers to the difference between the return at the ninth decile and the return at the first decile of the conditional earnings distribution. Therefore, one contribution of this paper consists of providing evidence on a smaller interval of impact than previously estimated, i.e. 5.1%-7.4% vs. 4.2%-26.2%.

Further, note that the standard KB techniques disregarding the endogeneity issue suggest a 4.1% gap and therefore underestimate the total impact of schooling on within-groups wage inequality in Portugal, as previously suggested by Andini (forthcoming, 2008), although the magnitude of the downward bias is smaller than one predicted by the AHS method.

Remarkably, the estimator of Lee predicts a pattern of the schooling coefficient, along the conditional wage distribution, that looks very much like the one predicted by the method of AHS, with a peak around the eight decile and a drop around the second decile. Hence, the impact of schooling on within-groups wage inequality is higher when measured as difference between the 25th quantile and the 75th quantile, ranging from 7.6% to 12.2%. Again, this result is not captured by the standard quantile-regression estimator that only highlights a 3.0% gap.

In contrast, the standard exogeneity-based techniques seem to perform relatively fine when considering the conditional average return to schooling. The ordinary-least-squares estimator prospects a 5.7% coefficient, which lies within the interval of 4.7%-7.5% obtained using the control-function approach and the AHS method. The conclusion is that schooling has a positive impact on *between-groups* wage inequality in Portugal. The latter, however, is a well-known result and goes beyond the objective of this paper.

IV. FINAL REMARKS

This paper uses the control-function estimator for quantile regression due to Lee (forthcoming, 2007) and provides further evidence of the total impact of schooling of within-groups wage inequality in Portugal. In our specific application, the estimator of Lee turns out to be more efficient than the estimator of Arias, Hallock and Sosa-Escudero (2001) at two key-deciles of the conditional wage distribution, thus providing more reliable measures of the corresponding total returns to schooling. Specifically, we find that standard exogeneity-based estimator of Koenker and Bassett (1978) underestimates the impact of schooling on earnings dispersion trough its within-groups dimension, although the magnitude of the bias is likely to be less than previously suggested.

The empirical research on the positive association between schooling and within-groups wage dispersion in Portugal is relatively rich in contributions. The finding of a positive association is also supported by several studies using the KB estimator with education levels rather than schooling years. The striking evidence that schooling/education is a strong source of the so-called residual earnings inequality should inspire an effort of *understanding the reasons* behind this stylized fact.

Some authors suggest that one possible explanation of the above-referred fact has to do with educational mismatches in the labour market. However, the existing empirical evidence is not necessarily consistent with this argument (see Budría, 2006). Our perception of the problem is that future research should pay more attention to the issue of school quality. The existing differences in the quality of Portuguese universities and high-schools, which are well-known to everybody living in Portugal, are likely to play an important role in explaining the stylized fact that the wage returns for individuals with the same number of schooling years or the same level of education (and the same observed characteristics) are quite heterogeneous. However, an empirical evaluation of the latter hypothesis needs individual-level data on school quality, which are not currently available.

REFERENCES

Andini, C. (forthcoming, 2008) The Total Impact of Schooling on Within-Groups Wage Inequality in Portugal, Applied Economics Letters, draft available at <http://www.uma.pt/andini>.

Andini, C. (forthcoming, 2007) Returns to Education and Wage Equations: a Dynamic Approach, Applied Economics Letters, draft available at <http://www.uma.pt/andini>.

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

Arias, O., Hallock, K.F. and Sosa-Escudero, W. (2001) Individual Heterogeneity in the Returns to Schooling: Instrumental Variables Quantile Regression Using Twins Data, Empirical Economics, 26(1), 7-40.

Budría, S. (2006) Can Over-Education Account for the Positive Association between Education and Within-Groups Wage Inequality? A Note, MPRA Working Papers, nº 92, Munich Personel RePEc Archive, October.

Hartog, J., Pereira, P.T. and Vieira, J.A.C. (2001) Changing Returns to Education in Portugal during the 1980s and Early 1990s: OLS and Quantile Regression Estimators, Applied Economics, 33(8), 1021-1037.

Koenker, R. and Bassett, G. (1978) Regression Quantiles, Econometrica, 46(1), 33-50.

Lee, S. (2004) Endogeneity in Quantile Regresssion Models: A Control Function Approach, CEMMAP Working Papers, nº CWP08/04, Centre for Microdata Methods and Practice, December.

Lee, S. (forthcoming, 2007) Endogeneity in Quantile Regresssion Models: A Control Function Approach, Journal of Econometrics, draft available at <http://www.sciencedirect.com/science/journal/03044076>.

Machado, J.A.F. and Mata, L. (2001) Earnings Functions in Portugal 1982-1994: Evidence from Quantile Regressions, Empirical Economics, 26(1), 115-134.

Martins, P.S. and Pereira, P.T. (2004) Does Education Reduce Wage Inequality? Quantile Regression Evidence from 16 Countries, Labour Economics, 11(3), 355-371.

Pereira, P.T. and Martins, P.S. (2004) Returns to Education and Wage Equations, Applied Economics, 36(6), 525-531

Table 1. Summary sample statistics

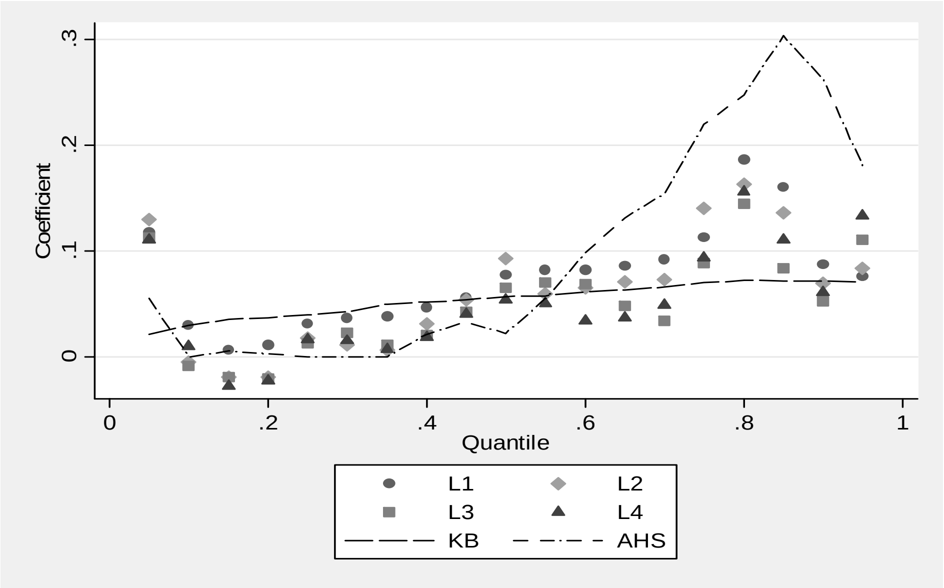
Variable	Obs.	Mean	S.E.	Min	Max
Logarithm of gross hourly wage	1782	6.55	0.47	3.32	8.49
Schooling years	1782	8.80	3.91	3.00	27.0
Year of birth	1782	1965	11.5	1936	1984
Quarter of birth	1782	2.45	1.11	1.00	4.00

Table 2. Conditional Returns to Schooling

Quantile	L1	L2	L3	L4	KB	AHS
5	0.1175	0.1296	0.1124	0.1103	0.0211	0.0550
	(0.0869)	(0.0933)	(0.0859)	(0.0909)	(0.0052)	(0.0828)
10	0.0300	-0.0050	-0.0084	0.0096	0.0298	0.0000
	(0.0631)	(0.0674)	(0.0611)	(0.0728)	(0.0033)	(0.0782)
15	0.0067	-0.0195	-0.0195	-0.0279	0.0352	0.0054
	(0.0452)	(0.0476)	(0.0455)	(0.0551)	(0.0025)	(0.0536)
20	0.0111	-0.0194	-0.0207	-0.0228	0.0368	0.0028
	(0.0444)	(0.0516)	(0.0633)	(0.0533)	(0.0019)	(0.0376)
25	0.0313	0.0177	0.0128	0.0164	0.0396	0.0000
	(0.0335)	(0.0369)	(0.0324)	(0.0494)	(0.0017)	(0.0155)
30	0.0367	0.0115	0.0225	0.0152	0.0427	0.0000
	(0.0378)	(0.0373)	(0.0300)	(0.0421)	(0.0023)	(0.0151)
35	0.0381	0.0066	0.0116	0.0069	0.0497	0.0000
	(0.0311)	(0.0357)	(0.0462)	(0.0400)	(0.0018)	(0.0108)
40	0.0467	0.0314	0.0205	0.0184	0.0518	0.0214
	(0.0304)	(0.0371)	(0.0302)	(0.0310)	(0.0020)	(0.0526)
45	0.0559	0.0540	0.0423	0.0401	0.0537	0.0329
	(0.0335)	(0.0289)	(0.0385)	(0.0391)	(0.0018)	(0.0329)
50	0.0770	0.0925	0.0651	0.0535	0.0570	0.0221
	(0.0668)	(0.0389)	(0.0423)	(0.0501)	(0.0032)	(0.0398)
55	0.0819	0.0594	0.0699	0.0501	0.0577	0.0549
	(0.0445)	(0.0472)	(0.0469)	(0.0563)	(0.0026)	(0.0613)
60	0.0822	0.0651	0.0685	0.0340	0.0614	0.0977
	(0.0509)	(0.0547)	(0.0436)	(0.0517)	(0.0031)	(0.0473)
65	0.0855	0.0711	0.0479	0.0371	0.0631	0.1312
	(0.0440)	(0.0523)	(0.0471)	(0.0522)	(0.0030)	(0.0489)
70	0.0918	0.0727	0.0338	0.0488	0.0658	0.1541
	(0.0542)	(0.0650)	(0.0643)	(0.0647)	(0.0029)	(0.0567)
75	0.1129	0.1401	0.0887	0.0934	0.0698	0.2195
	(0.0675)	(0.0414)	(0.0963)	(0.0632)	(0.0036)	(0.0397)
80	0.1861	0.1626	0.1445	0.1556	0.0723	0.2473
	(0.0551)	(0.0657)	(0.0702)	(0.0589)	(0.0038)	(0.0729)
85	0.1600	0.1360	0.0836	0.1103	0.0716	0.3029
	(0.0642)	(0.0576)	(0.0795)	(0.0797)	(0.0031)	(0.0795)
90	0.0872	0.0695	0.0521	0.0606	0.0717	0.2624
	(0.0813)	(0.0735)	(0.0719)	(0.0680)	(0.0042)	(0.1123)
95	0.0756	0.0834	0.1103	0.1329	0.0706	0.1808
	(0.0990)	(0.1046)	(0.1060)	(0.1105)	(0.0062)	(0.1073)
Mean	0.0751	0.0643	0.0472	0.0479	0.0577	0.0751
	(0.0447)	(0.0447)	(0.0446)	(0.0445)	(0.0029)	(0.0534)
90-10	0.0572	0.0745	0.0605	0.0510	0.0419	0.2624
75-25	0.0816	0.1225	0.0759	0.0770	0.0302	0.2195

Standard errors in parentheses

Figure 1. Conditional Returns to Schooling



Appendix. First-stage regression of schooling years

	Coeff.	Robust S.E.	t	P-value
Quarter of birth				
1	-0.5303	0.2733	-1.94	0.053
2	-0.5162	0.2780	-1.86	0.064
3	-0.3004	0.2675	-1.12	0.262
Year of birth				
1937	-0.1968	0.4193	-0.47	0.639
1938	0.4486	0.8273	0.54	0.588
1939	0.1668	1.1782	0.14	0.887
1940	-0.2851	0.3714	-0.77	0.443
1941	1.2821	1.2007	1.07	0.286
1942	0.0151	0.7509	0.02	0.984
1943	2.6175	1.3599	1.92	0.054
1944	0.4641	1.3862	0.33	0.738
1945	3.5350	1.1231	3.15	0.002
1946	1.7825	0.9022	1.98	0.048
1947	0.1080	0.5157	0.21	0.834
1948	2.4706	1.2484	1.98	0.048
1949	1.3233	0.7644	1.73	0.084
1950	2.5171	0.9047	2.78	0.005
1951	1.4911	0.6433	2.32	0.021
1952	1.2574	0.6183	2.03	0.042
1953	1.2702	0.4896	2.59	0.010
1954	1.5278	0.8376	1.82	0.068
1955	0.5447	0.4065	1.34	0.180
1956	1.6778	0.6129	2.74	0.006
1957	1.6236	0.6541	2.48	0.013
1958	1.4677	0.5944	2.47	0.014
1959	0.6664	0.4798	1.39	0.165
1960	1.1161	0.4322	2.58	0.010
1961	1.6299	0.6509	2.50	0.012
1962	2.3032	0.5585	4.12	0.000
1963	2.2418	0.5484	4.09	0.000
1964	1.4711	0.5304	2.77	0.006
1965	3.2459	0.7877	4.12	0.000
1966	3.1548	0.8257	3.82	0.000
1967	3.8398	0.9662	3.97	0.000
1968	3.0704	0.6384	4.81	0.000
1969	3.1249	0.5733	5.45	0.000
1970	3.7298	0.6285	5.93	0.000
1971	3.8072	0.4555	8.36	0.000
1972	3.9569	0.5066	7.81	0.000
1973	3.1319	0.4842	6.47	0.000
1974	4.3609	0.5063	8.61	0.000
1975	4.4149	0.5147	8.58	0.000
1976	3.6441	0.4565	7.98	0.000
1977	3.9069	0.4371	8.94	0.000
1978	3.5783	0.4024	8.89	0.000
1979	3.8873	0.4110	9.46	0.000
1980	3.9421	0.3377	11.67	0.000
1981	4.0952	0.2941	13.92	0.000
1982	2.7029	0.4154	6.51	0.000
1983	2.6506	0.3425	7.74	0.000
1984	2.6250	0.4438	5.91	0.000
Constant	6.4083	0.2554	25.09	0.000
Sargan test of over-identifying restrictions				0.1351
F-test of excluded instruments				0.1972