Control Function and Related Methods: Nonlinear Models

Jeff Wooldridge
Michigan State University

Programme Evaluation for Policy Analysis
Institute for Fiscal Studies
June 2012

1. General Approach
2. Nonlinear Models with Additive Errors
3. Models with Intrinsic Nonlinearity
4. “Special Regressor” Methods for Binary Response
1. General Approach

- With models that are nonlinear in parameters, the linear projection approach to CF estimation rarely works (unless the model happens to be linear in the EEVs).

- If $\mathbf{u}_1$ is a vector of “structural” errors, $\mathbf{y}_2$ is the vector of EEVs, and $\mathbf{z}$ is the vector of exogenous variables, we at least have to model $E(\mathbf{u}_1|\mathbf{y}_2, \mathbf{z})$ and often $D(\mathbf{u}_1|\mathbf{y}_2, \mathbf{z})$ (in a parametric context).

- An important simplification is when

$$\mathbf{y}_2 = \mathbf{g}_2(\mathbf{z}, \delta_2) + \mathbf{v}_2$$

(1)

where $\mathbf{v}_2$ is independent of $\mathbf{z}$. Unfortunately, this rules out discrete $\mathbf{y}_2$. 
• With discreteness in $y_2$, difficult to get by without modeling $D(y_2|z)$.

• In many cases – particularly when $y_2$ is continuous – one has a choice between two-step control function estimators and one-step estimators that estimate parameters at the same time. (Typically these have a quasi-LIML flavor.)

• More radical suggestions are to use generalized residuals in nonlinear models as an approximate solution to endogeneity.
2. Nonlinear Models with Additive Errors

• Suppose

\[ y_1 = g_1(y_2, z_1, \gamma_1) + u_1 \]
\[ y_2 = g_2(z, \gamma_2) + v_2 \]

and

\[ E(u_1|v_2, z) = E(u_1|v_2) = v_2 \rho_1 \]

• Assume we have enough relevant elements in \( z_2 \) so identification holds.
• We can base a CF approach on

\[ E(y_1|y_2, z) = g_1(y_2, z, \gamma_1) + v_2 \rho_1 \]  

(2)

• Estimate \( \gamma_2 \) by multivariate nonlinear least squares, or an MLE, to get

\[ \hat{v}_{i2} = y_{i2} - g_2(z_i, \hat{\gamma}_2). \]

• In second step, estimate \( \gamma_1 \) and \( \rho_1 \) by NLS using the mean function in (2).

• Easiest when \( y_2 = z \Gamma_2 + v_2 \) so can use linear estimation in first stage.

• Can allow a vector \( y_1 \), as in Blundell and Robin (1999, Journal of Applied Econometrics): expenditure share system with total expenditure endogenous.
• In principle can have $y_2$ discrete provided we can find $E(u_1|y_2,z)$.
• If $y_2$ is binary, can have nonlinear switching regression – but with additive noise.

• Example: Exponential function:

\[
y_1 = \exp(\eta_1 + \alpha_1 y_2 + z_1 \delta_1 + y_2 z_1 \psi_1) + u_1 + y_2 v_1
\]
\[
y_2 = 1[z_2 \delta_2 + e_2 > 0]
\]

• Usually more natural for the unobservables to be inside the exponential function. And what if $y_1$ is something like a count variable?
• Same issue arises in share equations.
3. Models with Intrinsic Nonlinearity

- Typically three approaches to nonlinear models with EEVs.

  1. Plug in fitted values from a first step estimation in an attempt to mimic 2SLS in linear model. Usually does not produce consistent estimators because the implied form of $E(y_1|z)$ or $D(y_1|z)$ is incorrect.

  2. CF approach: Plug in residuals in an attempt to obtain $E(y_1|y_2, z)$ or $D(y_1|y_2, z)$.

  3. Maximum Likelihood (often limited information): Use models for $D(y_1|y_2, z)$ and $D(y_2|z)$ jointly.

- All strategies are more difficult with nonlinear models when $y_2$ is discrete.
Binary and Fractional Responses

Probit model:

\[ y_1 = 1[\alpha_1 y_2 + z_1 \delta + u_1 \geq 0], \]  

where \( u_1|z \sim Normal(0, 1) \). Analysis goes through if we replace \((z_1, y_2)\) with any known function \( x_1 \equiv g_1(z_1, y_2) \).

- The Rivers-Vuong (1988) approach is to make a homoskedastic-normal assumption on the reduced form for \( y_2 \),

\[ y_2 = z\pi_2 + v_2, \quad v_2|z \sim Normal(0, \tau_2^2). \]  

(4)
• RV approach comes close to requiring

\[(u_1, v_2) \text{ independent of } z.\] (5)

If we also assume

\[(u_1, v_2) \sim \text{Bivariate Normal} \] (6)

with \(\rho_1 = \text{Corr}(u_1, v_2)\), then we can proceed with MLE based on \(f(y_1, y_2|z)\). A CF approach is available, too, based on

\[P(y_1 = 1|y_2, z) = \Phi(\alpha_{\rho_1}y_2 + z_1\delta_{\rho_1} + \theta_{\rho_1}v_2)\] (7)

where each coefficient is multiplied by \((1 - \rho_1^2)^{-1/2}\).
The Rivers-Vuong CF approach is

(i) OLS of $y_{i2}$ on $z_i$, to obtain the residuals, $\hat{v}_{i2}$.

(ii) Probit of $y_{i1}$ on $z_{i1}, y_{i2}, \hat{v}_{i2}$ to estimate the scaled coefficients. A simple $t$ test on $\hat{v}_2$ is valid to test $H_0 : \rho_1 = 0$. 

• Can recover the original coefficients, which appear in the partial effects – see Wooldridge (2010, Chapter 15). Or, obtain average partial effects by differentiating the estimated “average structural function”:

\[
\hat{ASF}(z_1, y_2) = N^{-1} \sum_{i=1}^{N} \Phi(x_i \hat{\beta}_{\rho_1} + \hat{\theta}_{\rho_1} \hat{v}_{i2}), \tag{8}
\]

that is, we average out the reduced form residuals, \( \hat{v}_{i2} \).

• Cover the ASF in more detail later.
The two-step CF approach easily extends to fractional responses: $0 \leq y_1 \leq 1$. Modify the model as

$$E(y_1|y_2, z, q_1) = \Phi(x_1\beta_1 + q_1),$$

(9)

where $x_1$ is a function of $(y_2, z_1)$ and $q_1$ contains unobservables.

- Assume $q_1 = \rho_1 v_2 + e_1$ where $D(e_1|z, v_2) = Normal(0, \sigma^2_{e_1})$.
- Use the same two-step estimator as for probit. (In Stata, glm command in second stage.) In this case, must obtain APEs from the ASF in (8).
• In inference, assume only that the mean is correctly specified. (Use sandwich in Bernoulli quasi-LL.)
• To account for first-stage estimation, the bootstrap is convenient.
• No IV procedures available unless assume that, say, the log-odds transform of $y_1$ is linear in $x_1\beta_1$ and an additive error independent of $z$. 
CF has clear advantages over “plug-in” approach, even in binary response case. Suppose rather than conditioning on \( v_2 \) along with \( z \) (and therefore \( y_2 \)) to obtain \( P(y_1 = 1|z, y_2) \) we use

\[
P(y_1 = 1|z) = \Phi\{[\alpha_1(z\pi_2) + z_1\delta_1]/\omega_1\}
\]

\[
\omega_1^2 = \text{Var}(\alpha_1 v_2 + u_1)
\]

(i) OLS on the reduced form, and get fitted values, \( \hat{y}_{i2} = z_i\hat{\pi}_2 \). (ii) Probit of \( y_{i1} \) on \( \hat{y}_{i2}, z_{i1} \). Harder to estimate APEs and test for endogeneity.
• Danger with plugging in fitted values for $y_2$ is that one might be tempted to plug $\hat{y}_2$ into nonlinear functions, say $y_2^2$ or $y_2z_1$, and use probit in second stage. Does not result in consistent estimation of the scaled parameters or the partial effects.

• Adding the CF $\hat{v}_2$ solves the endogeneity problem regardless of how $y_2$ appears.
Example: Married women’s fraction of hours worked.

```
. use mroz
. gen frachours = hours/8736
. sum frachours

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>frachours</td>
<td>753</td>
<td>.0874729</td>
<td>.0936243</td>
<td>0</td>
<td>.5666209</td>
</tr>
</tbody>
</table>
```
. reg nwifeinc educ exper expersq kidslt6 kidsge6 age huseduc husage

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>20722.898</td>
<td>8</td>
<td>2590.36225</td>
</tr>
<tr>
<td>Residual</td>
<td>81074.2176</td>
<td>744</td>
<td>108.970723</td>
</tr>
<tr>
<td>Total</td>
<td>101797.116</td>
<td>752</td>
<td>135.368505</td>
</tr>
</tbody>
</table>

Number of obs = 753
F( 8, 744) = 23.77
Prob > F = 0.0000
R-squared = 0.2036
Adj R-squared = 0.1950
Root MSE = 10.439

| nwifeinc | Coef. | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|----------|-------|-----------|-------|------|----------------------|
| educ     | 0.6721947 | 0.2138002 | 3.14  | 0.002 | 0.2524713 to 1.091918 |
| exper    | -0.3133239 | 0.1383094 | -2.27 | 0.024 | -0.5848472 to -0.0418007 |
| expersq  | -0.0003769 | 0.0045239 | -0.08 | 0.934 | -0.0092581 to 0.0085043 |
| kidslt6  | 0.9004389 | 0.8265936 | 1.09  | 0.276 | -0.7222947 to 2.523172 |
| kidsge6  | 0.4462001 | 0.3225308 | 1.38  | 0.167 | -0.1869788 to 1.079379 |
| age      | 0.2819309 | 0.1075901 | 2.62  | 0.009 | 0.0707146 to 0.4931472 |
| huseduc  | 1.1882889 | 0.1617589 | 7.35  | 0.000 | 0.8707307 to 1.505847 |
| husage   | 0.0681739 | 0.1047836 | 0.65  | 0.515 | -0.1375328 to 0.2738806 |
| _cons    | -15.46223 | 3.9566    | -3.91 | 0.000 | -23.22965 to -7.694796 |

.predict v2h, resid
. glm frachours educ exper expersq kidslt6 kidsge6 age nwifeinc v2h, fam(bin) link(probit) robust
              note: frachours has noninteger values

Generalized linear models
Optimization : ML
Scale parameter = 1
Deviance = 77.29713199  (1/df) Deviance = 0.103894
Pearson = 83.04923963  (1/df) Pearson = 0.1116253

Variance function: V(u) = u*(1-u/1)  [Binomial]
Link function : g(u) = invnorm(u)  [Probit]

Log pseudolikelihood = -154.3261842  BIC = -4851.007

|            | Coef.    | Robust Std. Err. | z   | P>|z|    | [95% Conf. Interval] |
|-------------|----------|------------------|-----|--------|---------------------|
| frachours   |          |                  |     |        |                     |
| educ        | .0437229 | .0169339         | 2.58| 0.010  | .010533             | .0769128           |
| exper       | .0610646 | .0096466         | 6.33| 0.000  | .0421576            | .0799717           |
| expersq     | -.00096  | .0002691         | -3.57| 0.000  | -.0014875           | -.0004326          |
| kidslt6     | -.4323608| .0782645         | -5.52| 0.000  | -.5857565           | -.2789651          |
| kidsge6     | -.0149373| .0202283         | -0.74| 0.460  | -.0545841           | .0247095           |
| age         | -.0219292| .0043658         | -5.02| 0.000  | -.0213986           | -.0200756          |
| nwifeinc    | -.0131868| .0083704         | -1.58| 0.115  | -.0295925           | .0032189           |
| v2h         | .0102264 | .0085828         | 1.19 | 0.233  | -.0065957           | .0270485           |
| _cons       | -1.169224| .2397377         | -4.88| 0.000  | -1.639102           | -.6993472          |
Fitting exogenous probit model

Probit model with endogenous regressors

Number of obs = 753
Wald chi2(7) = 240.78
Log pseudolikelihood = -3034.3388
Prob > chi2 = 0.0000

| Coef. | Std. Err. | z | P>|z| | [95% Conf. Interval] |
|-------|-----------|---|-------|------------------------|
| nwifeinc | -.0131207 | .0083969 | -1.56 | 0.118 | -.0295782 | .0033369 |
| educ | .0434908 | .0169655 | 2.56 | 0.010 | .0102389 | .0767427 |
| exper | .060721 | .0098379 | 6.17 | 0.000 | .041439 | .0800029 |
| expersq | -.0009547 | .0002655 | -3.60 | 0.000 | -.0014751 | -.0004342 |
| kidslt6 | -.0148507 | .0205881 | -0.72 | 0.471 | -.0552025 | .0255012 |
| kidsge6 | -.0148507 | .0205881 | -0.72 | 0.471 | -.0552025 | .0255012 |
| age | -.0218038 | .0045701 | -4.77 | 0.000 | -.030761 | -.0128465 |
| _cons | -1.162787 | .2390717 | -4.86 | 0.000 | -1.631359 | -.6942151 |

| Coef. | Std. Err. | z | P>|z| | [95% Conf. Interval] |
|-------|-----------|---|-------|------------------------|
| _athrho | .1059984 | .0906159 | 1.17 | 0.242 | -.0716056 | .2836024 |
| _lnsigma | 2.339528 | .0633955 | 36.90 | 0.000 | 2.215275 | 2.463781 |

Instrumented: nwifeinc
Instruments: educ exper expersq kidslt6 kidsge6 age huseduc husage

Wald test of exogeneity (_athrho = 0): chi2(1) = 1.37 Prob > chi2 = 0.2421
• What are the limits to the CF approach? Consider

\[ E(y_1|z, y_2, q_1) = \Phi(\alpha_1 y_2 + z_1 \delta_1 + q_1) \]  

(10)

where \(y_2\) is discrete. Rivers-Vuong approach does not generally work (even if \(y_1\) is binary).

• Neither does plugging in probit fitted values, assuming

\[ P(y_2 = 1|z) = \Phi(z\delta_2) \]  

(11)

In other words, do \textit{not} try to mimic 2SLS as follows: (i) Do probit of \(y_2\) on \(z\) and get the fitted probabilities, \(\hat{\Phi}_2 = \Phi(z\hat{\delta}_2)\). (ii) Do probit of \(y_1\) on \(z_1, \hat{\Phi}_2\), that is, just replace \(y_2\) with \(\hat{\Phi}_2\).
• The only strategy that works under traditional assumptions is maximum likelihood estimation based on $f(y_1|y_2, z)f(y_2|z)$. [Perhaps this is why some, such as Angrist (2001), promote the notion of just using linear probability models estimated by 2SLS.]

• “Bivariate probit” software can be used to estimate the probit model with a binary endogenous variable. Wooldridge (2011) shows that the same quasi-LIML is consistent when $y_1$ is fractional if (10) holds.

• Can also do a full switching regression when $y_1$ is fractional. Use “heckprobit” quasi-LLs.
• A CF approach based on generalize residuals can be justified for “small” amounts of endogeneity. Consider

\[ E(y_1|y_2, z, q_1) = \Phi(x_1 \beta_1 + q_1) \]  \hspace{1cm} (11)

and

\[ y_2 = 1[z \delta_2 + e_2 > 0] \]  \hspace{1cm} (12)

• \((q_1, e_1)\) jointly normal and independent of \(z\).

• Let

\[ \hat{gr}_{i2} \equiv y_{i2} \lambda(z_i \hat{\delta}_2) - (1 - y_{i2}) \lambda(-z_i \hat{\delta}_2) \]  \hspace{1cm} (13)

be the generalized residuals from the probit estimation.
The variable addition test (essentially score test) for the null that $q_1$ and $e_2$ are uncorrelated can be obtained by “probit” of $y_i1$ on the mean function

$$
\Phi(x_i1 \beta_{\rho1} + \eta_{\rho1} \hat{g}r_{i2})
$$

(14)

and use a robust $t$ statistic for $\hat{\eta}_{\rho1}$. (Get scaled estimates of $\beta_1$ and $\eta_1$.)
Wooldridge (2011) suggests that this can approximate the APEs, obtained from the estimated average structural function:

\[
\widehat{ASF}(y_2, z_1) = N^{-1} \sum_{i=1}^{N} \Phi(x_i \hat{\beta}_1 + \hat{\eta}_1 \hat{g}r_{i2})
\] (15)

Simulations suggest this can work pretty well, even if the amount of endogeneity is not “small.”
• If we have two sources of unobservables, add an interaction:

\[
E(y_1|y_2, z) \approx \Phi(x_1 \beta_{\rho_1} + \eta_{\rho_1} r_2 + \omega_{\rho_1} y_2 r_2)
\]  

\[
\hat{ASF}(y_2, z_1) = N^{-1} \sum_{i=1}^{N} \Phi(x_1 \hat{\beta}_{\rho_1} + \hat{\eta}_{\rho_1} \hat{r}_{i_2} + \hat{\omega}_{\rho_1} y_2 \hat{r}_{i_2})
\]  

\[
(16) \hspace{10cm} (17)
\]

• Two df test of null that \(y_2\) is exogenous.
Multinomial Responses

- Recent push by Petrin and Train (2010), among others, to use control function methods where the second step estimation is something simple – such as multinomial logit, or nested logit – rather than being derived from a structural model. So, if we have reduced forms

\[ y_2 = z\Pi_2 + v_2, \quad (18) \]

then we jump directly to convenient models for \( P(y_1 = j|z_1, y_2, v_2) \). The average structural functions are obtained by averaging the response probabilities across \( \hat{v}_{i2} \).
• Can use the same approach when we have a vector of shares, say $y_1$, adding up to unity. (Nam and Wooldridge, 2012.) The multinomial distribution is in the linear exponential family.

• No generally acceptable way to handle discrete $y_2$, except by specifying a full set of distributions.

• Might approximate by adding generalized residuals as control functions to standard models (such as MNL).
Exponential Models

- IV and CF approaches available for exponential models. For $y_1 \geq 0$ (could be a count) write

$$E(y_1|y_2, z, r_1) = \exp(x_1\beta_1 + q_1),$$  \hspace{1cm} (19)

where $q_1$ is the omitted variable independent of $z$. $x_1$ can be any function of $(y_2, z_1)$.

- CF method can be based on

$$E(y_1|y_2, z) = \exp(x_1\beta_1)E[\exp(q_1)|y_2, z].$$  \hspace{1cm} (20)
• For continuous $y_2$, can find $E[\exp(q_1)|y_2, z]$ when $D(y_2|z)$ is homoskedastic normal (Wooldridge, 1997) and when $D(y_2|z)$ follows a probit (Terza, 1998).
• In the probit case,
\[
E(y_1|y_2, z) = \exp(x_1 \beta_1) h(y_2, z\pi_2, \theta_1)
\]
\[
h(y_2, z\pi_2, \theta_1) = \exp(\theta_1^2/2) \{y_2 \Phi(\theta_1 + z\pi_2)/\Phi(z\pi_2)
+ (1 - y_2)[1 - \Phi(\theta_1 + z\pi_2)]/[1 - \Phi(z\pi_2)]\}.
\]
• Can use two-step NLS, where $\hat{\pi}_2$ is obtained from probit. If $y_1$ is count, use a QMLE in the linear exponential family, such as Poisson or geometric.
• Can show the VAT score test is obtained from the mean function

\[ \exp(\mathbf{x}_i \beta_1 + \eta_1 \hat{g}_{i2}) \]  \hspace{1cm} (23)

where

\[ \hat{g}_{i2} = y_{i2} \lambda(z_i \hat{\delta}_2) - (1 - y_{i2}) \lambda(-z_i \hat{\delta}_2) \]

• Convenient to use Poisson QMLE. Computationally very simple. At a minimum might as well test \( H_0 : \eta_1 = 0 \) first.
• As in binary/fractional case, adding the GR to the exponential mean might account for endogeneity, too.

\[
\text{ASF}(y_2, z_1) = N^{-1} \sum_{i=1}^{N} \exp(x_1 \hat{\beta}_1 + \hat{\eta}_1 \hat{gr}_{i2}) \\
= \left[ N^{-1} \sum_{i=1}^{N} \exp(\hat{\eta}_1 \hat{gr}_{i2}) \right] \exp(x_1 \hat{\beta}_1)
\]

• Add \( y_{i2} \hat{gr}_{i2} \) for a switching regression version:

\[
\text{“}E(y_{i1}|y_{i2}, z_i) = \exp(x_{i1} \beta_1 + \eta_1 \hat{gr}_i + \omega_1 y_{i2} \hat{gr}_{i2}) \quad (24)
\]
• IV methods that work for any $y_2$ without distributional assumptions are available [Mullahy (1997)]. If

$$E(y_1|y_2, z, q_1) = \exp(x_1 \beta_1 + q_1)$$

(25)

and $q_1$ is independent of $z$ then

$$E[\exp(-x_1 \beta_1)y_1|z] = E[\exp(q_1)|z] = 1,$$

(26)

where $E[\exp(q_1)] = 1$ is a normalization. The moment conditions are

$$E[\exp(-x_1 \beta_1)y_1 - 1|z] = 0.$$  

(27)

• Requires nonlinear IV methods. How to approximate the optimal instruments?
Quantile Regression

• Suppose

\[ y_1 = \alpha_1 y_2 + z_1 \delta_1 + u_1, \]  

where \( y_2 \) is endogenous and \( z \) is exogenous, with \( z_1 \subset z \).

• Amemiya’s (1982) two-stage LAD estimator is a plug-in estimator.

Reduced form for \( y_2 \),

\[ y_2 = z \pi_2 + \nu_2. \]  

First step applies OLS or LAD to (29), and gets fitted values, \( y_{i2} = z_i \hat{\pi}_2 \). These are inserted for \( y_{i2} \) to give LAD of \( y_{i1} \) on \( z_{i1}, \hat{y}_{i2} \).

2SLAD relies on symmetry of the composite error \( \alpha_1 \nu_2 + u_1 \) given \( z \).
• If $D(u_1, v_2 | z)$ is “centrally symmetric” can use a control function approach, as in Lee (2007). Write

$$u_1 = \rho_1 v_2 + e_1,$$

where $e_1$ given $z$ would have a symmetric distribution. Get LAD residuals $\hat{v}_{i2} = y_{i2} - z_i \hat{\pi}_2$ and do LAD of $y_{i1}$ on $z_{i1}, y_{i2}, \hat{v}_{i2}$. Use $t$ test on $\hat{v}_{i2}$ to test null that $y_2$ is exogenous.

• Interpretation of LAD in context of omitted variables is difficult unless lots of symmetry assumed.

• See Lee (2007) for discussion of general quantiles.
4. “Special Regressor” Methods for Binary Response

• Lewbel (2000) showed how to semi-parametrically estimate parameters in binary response models if a regressor with certain properties is available. Dong and Lewbel (2012) have recently relaxed those conditions somewhat.

• Let $y_1$ be a binary response:

$$y_1 = \begin{cases} 1 & [w_1 + y_2 \alpha_2 + z_1 \delta_1 + u_1 > 0] \\ 0 & \end{cases}$$

$$= \begin{cases} 1 & [w_1 + x_1 \beta_1 + u_1 > 0] \\ 0 & \end{cases}$$

(31)

where $w_1$ is the “special regressor” normalized to have unity coefficient and assumed to be continuously distributed.
• In willingness-to-pay applications, \( w_1 = -cost \), where \( cost \) is the amount that a new project will cost. Then someone prefers the project if

\[
y_1 = 1[\text{wtp} > cost]
\]

• In studies that elicit WTP, \( cost \) is often set completely exogenously: independent of everything else, including \( y_2 \).

• Dong and Lewbel (2012) assume \( w_1 \), like \( z_1 \), is exogenous in (31), and that there are suitable instruments:

\[
E(w_1'u_1) = 0, E(z'u_1) = 0
\]  (32)
• Need usual rank condition if we had linear model without \( w_1 \): \( \text{rank } E(z'x_1) = K_1 \).

• Setup is more general but they also write a linear equation

\[
w_1 = y_2 \pi_1 + z \pi_2 + r_1
\]

\[
E(y_2' r_1) = 0, \ E(z' r_1) = 0
\]

and then require (at a minimum)

\[
E(r_1 u_1) = 0.
\]

• Condition (34), along with previous assumptions, means \( w_1 \) must be excluded from the reduced form for \( y_2 \) (which is a testable restriction).
To see this, multiply (33) by $u_1$, take expectations, impose exogeneity on $w_1$ and $z$, and use (31):

$$E(u_1 w_1) = E(u_1 y_2) \pi_1 + E(u_1 z) \pi_2 + E(u_1 r_1)$$

or

$$0 = E(u_1 y_2) \pi_1$$

(35)

For this equation to hold except by fluke we need $\pi_1 = 0$ (and in the case of a scalar $y_2$ this is the requirement). From (33) this means $y_2$ and $w_1$ are uncorrelated after $z$ has been partialled out. This implies $w_1$ does not appear in the reduced form for $y_2$ once $z$ is included.
• Can easily test the Dong-Lewbel identification assumption on the special regressor. Can hold if $w_1$ depends on $z$ provided that $w_1$ is independent of $y_2$ conditional on $z$.

• In WTP studies, means we can allow $w_1$ to depend on $z$ but not $y_2$.

• Dong and Lewbel application: $y_1$ is decision to migrate, $y_2$ is home ownership dummy. The special regressor is $age$. But does $age$ really have no partial effect on home ownership given the other exogenous variables?
• If the assumptions hold, D-L show that, under regularity conditions (including wide support for \( w_1 \)),

\[
    s_1 = x_1 \beta_1 + e_1
\]

\[
    E(z'e_1) = 0
\]

where

\[
    s_1 = \frac{(y_1 - 1[w_1 \geq 0])}{f(w_1|y_2, z)}
\]

(37)

where \( f(\cdot|y_2, z) \) is the density of \( w_1 \) given \((y_2, z)\).
• Estimate this density by MLE or nonparametrics:

\[
\hat{s}_{i1} = \frac{(y_{i1} - 1[w_{i1} \geq 0])}{\hat{f}(w_{i1}|y_{i2}, z_i)}
\]

• Requirement that \(f(\cdot|y_2, z)\) is continuous means the special regressor must appear additively and nowhere else. So no quadratics or interactions.