Paper for Presentation next week: Water for Life: The Impact of the Privatization of Water Services on Child Mortality, *Journal of Political Economy*, Galiani, Gertler and Schargrodsky, Vol. 113, pp. 83-120, February 2005

Intro Comments:

The purpose of this lecture is to think about an alternative tool for dealing with selection-type problems: namely propensity score matching. This builds on the treatment effects approach that was developed earlier in the course.

Let's set up the problem slowly:

Say we have a program we want to evaluate (eg. some form of teacher training, workforce program, unemployment re-skilling, on-the-job training, tax incentive scheme, R&D team structure, managerial structure or whatever...).

Say there is a population of people who could be exposed to the program. The program is suspected to affect the realisation of some outcome variable Y_i. What we are interested in is either the average treatment effect or the average treatment effect on the treated.

If people were randomly allocated to the treatment ATE and ATET would be the same thing and we could proceed by just comparing sample means of the outcome variable (or maybe conditional means if we wanted to boost finite sample properties).

That is, given a large enough sample we could average away both observable and unobservable characteristics of the population.

What happens when assignment to treatment is not random?

Here it helps to have a model of how we think people might select into the treatment. Lets imagine we have some sort of model in the back of our head that is some sort of idea that people who have a better chance of doing well in the treatment will select into it. Some of what makes someone do well is observed and some is unobserved. Most empirical applications will have aspect of this in them.

Let selection be indicated by a variable D where $D = 1$ if assigned to the program and $D = 0$ if not, and let Z be observed characteristics of individuals in the sample from the population of interest and let the outcome for an individual be Y_1 if they were to get treatment and Y_0 if they were to not get treatment.

So ATE and ATET are (resp.)

$$E(Y_1 - Y_0)$$

$$E(Y_1 - Y_0 | D = 1)$$

We assume either

a) 'Strict Ignorability'

(Y_0, Y_1) and D are independent conditional on Z

equivalently

$$Pr(D = 1|Y_1, Y_0, Z) = Pr(D = 1|Z)$$

or

$$E(D = 1|Y_1, Y_0, Z) = E(D = 1|Z)$$

b) 'Conditional Mean Independence' (which is weaker than the above)

$$E(Y_0|Z, D = 1) = E(Y_0|Z, D = 0) = E(Y_0|Z)$$

basically we use Strict Ignorability when going after ATE and Conditional Mean Independence when going after ATET

We will also assume that

$$0 < Pr(D = 1) < 1$$

That is there is a positive probability of either participating or not (this can be weakened a bit).

Note that strict ignorability means that we have "selection on observables" loosely speaking. That is, unobserved characteristics *that might be related to the utimate outcome* are not determining who recieves the treatment.

Conditional Mean Independence means that the expected effect of not getting treated only depends on observed stuff. That is (for instance), conditional on observables, people who do better in the absence of treatment are just as likely to get treatment as someone who would do worse. Notice that, the extent that unobservables drive the outcome following treatment and this in turn affects selection, is not important.

To see this last point note that:

$$E(Y_1 - Y_0|D = 1) = E(Y_1|D = 1) - E_{Z|D=1}E_{Y_0}(Y_0|D = 1, Z)$$

Hence

$$E(Y_1 - Y_0|D = 1) = E(Y_1|D = 1) - E_{Z|D=1}E_{Y_0}(Y_0|D = 0, Z)$$

Where the last bit can be estimated by matching the untreated to the treated based on the treated's Z's.

The last thing to bear in mind is that the Z's should not be causily determined by whether treatment is recieved or not. (ie. if you are matching on marital status, then the marital status should not be affected by treatment...)

You should have a sense of how this matching approach is going to work now, so lets get into a little more detail

Econometric Issue: imagine that the Dimension of Z is large. If the elements of Z are discrete then you worry that the grid you are matching on may be either too coarse or that you won't have a strong enough sense of closeness to really make matching sensible.

More worrying is that if the variables in Z are continuous, and $E(Y_0|D = 0, Z)$ is estimated nonparametrically (say using a kernel version of a local regression), then convergence rates will be slow due to what is called a curse of dimensionality problem. This will mean that inference is (potentially) not possible.

Rosenbaum and Rubin prove the following theorem (which is very cool):

$$E(D|Y, Z) = E(D|Z) \Rightarrow E(D|Y, Pr(D = 1|Z))$$
$$= E(D|Pr[D = 1|Z])$$

That is, when Y_0 outcomes are independent of program participation conditional on Z, they are also independent of participation conditional on the probability of participation.

So, when matching on Z is valid, matching on $Pr(D = 1|Z)$ is also valid.

So, provided $Pr(D = 1|Z)$ can be estimated paramterically (ie. with a parametric estimator like a probit) we can avoid this curse of dimensionality issue.

How to justify these assumptions: An Example

It is often useful to write down a simple model to get a sense of how restrictive the assumptions above might be in a particular context. Here is an example.

Say we are evaluating the effect of a training program. Say people apply to the program based on expected benefits.

If treatment occurs, training starts at t=0 and lasts through to period 10. The information that people have when they choose to participate is W (which may or may not be observed by the econometrican).

$$D = 1 \text{ if } E\left(\sum_{j}^{T} = 10\frac{Y_{1j}}{(1+r)^j} - \sum_{k=1}^{T}\frac{Y_{0k}}{(1+r)^k}W\right) > \epsilon + Y_{00}$$

If

$$f(Y_{0k}|\epsilon + Y_{00}, Z) = f(Y_{0k}|Z) \text{ then}$$

$$E(Y_{0k}|Z, D = 1) = E(Y_{0k}|Z, \epsilon + Y_{00} < \gamma W) = E(Y_{0k}|Z)$$

Then matching would be OK.

However, this would put restrictions on the correlation structure of the earnings residuals over time. ie, if $X = W$ and $Y_{00} = Y_{0k}$ then you have a problem becasue knowing that a person selected into the program tells you something about future earnings if treatment is not recieved.

So you have to think about how X and W are related and how the no-treatment earnings evolve over time more deeply. (see Lalonde, Heckman and Smith (1999) for a similar example)

Basic Set-Up of the matching Estimator

Let the object of interest be $\widehat{\alpha}$

$$\widehat{\alpha} = \frac{1}{n_1} \sum_{i \in I_1} \left[Y_{1i} - \widehat{E}(Y_{0i} | D = 1, P_i(Z) \right]$$

where

$$\widehat{E}(Y_{0i} | D = 1, P_i(Z)) = \sum_{j \in \{j | P_j(Z) \in C(P_i(Z))\}} W(i, j) Y_{0j}$$

Typically you'll use a probit or somthing similar for estimating the propensity and then take the weighted average distance between each Y_{1i} and the stuff it is matched to, and then average over all of that.

Except in specific instances, where the bootstrap does not work, standard errors are obtained by bootstrapping over the entire (two-step) proceedure. To work out whether the bootstrap is OK or not you should check Abadie and Imbens (2004), this will be an issue if you are doing some forms of what are called nearest neighbor matching (see below). They have code for fixing this issue on their website

Implementation

There are more or less three types of ways to implement this:

1. Nearest Neigbor Matching

This is the most traditional. Each treated pair is matched with their *nearest neigbor* so that

$$C(P) = \min_j \|P_i - p_j\|$$

this can be implemented with or without replacement. Doing it with replacement seems better to me: without replacement the estimates will depend on the order in which you match.

2. Stratification or Interval Matching

Here the common support of P is partitioned into intervals and treatment effects are computed through simple averaging within intervals. The overall impact is obtained by weighted averages of the treatement effects in each interval where the weights are given by the proportion of D=1 in each interval.

3. Kernel and Local Linear Matching

The idea here is to accomodate many to one matching by weighting different untreated, matched, observations, by how far away they are from the treated observation. This is done using a kernel of some sort.

In the framework above, this in implemented by setting the weighting functions equal to

$$W(i,j) = \frac{G\left(\frac{P_j - P_i}{a}\right)}{\sum_{k \in C(P)} G\left(\frac{P_k - P_i}{a}\right)}$$

where $G\left(\frac{P_j - P_i}{a}\right)$ is a kernel of some sort (see for instance Pagan and Ullah (200?)), and a is a bandwidth

If the kernel is bounded between -1 and 1 then

$$C(P_i) = \{|\frac{P_j - P_i}{a}| \leq 1\}$$

Heckman, Ichimura and Todd (1997) have an extension of this they call local linear matching that has somewhat better (nicer) assymptotic properties, particularly when you might be worried about distributions hitting bounderies and things like that.

When you use Kernel matching you can bootstrap to your hearts content.

Why matching and not regression?

If you were awake you should have been wondering why do all this when you could run a regression that is something like

$Y = \beta_0 + \alpha D + \gamma G(Z) + \epsilon$ (for the ATET assumptions)

(a similar regression implementation exists for ATE - see Wooldridge's text book at 611 - 612)

The short answer is that a lot of the time it is not going to matter whether you use a regression apporach or a matching approach.

The slightly more nuanced answer is that matching focuses on modelling the selection process which regression assumes you can model the outcome. When we understand the selection mechanism much better than the outcome mechanism, then matching is likely to be more convincing. It also invovles slightly fewer assumptions (regression tends to require a little more functional form to be imposed, because you need to take a stand on the G(Z) function above).

When you are worried about selection on unobservables, you might wonder about how to handle this. If treatment is at date zero, and you have data on the treated and untreated at dates -1 and 1, then you might this a differences in differences type approach could be combined with a matching approach.

You would be right.

A differences-in-differences extension of matching exists (see Heckman, Ichimura and Todd (1997) and Heckman, Ichimura Smith and Todd (1998)). Allan will be joining us later to talk through an implementation of this style of estimator in his recent research.

An application of matching:

Dehejia and Wahba (1999), Causal Effects in Non-experimental Studies: Reevaluating the Evaluation of Training Programs, Journal of the American Statistical Association, 94,448, pp. 1053-62

Chandra and Collard-Wexler, Mergers in Two-Sided Markets: An Application to the Canadian Newspaper Industry, JEMS, Forthcoming.

Our application involves studying the effect of a series of mergers in the Canadian newspaper industry. During the period 1995 to 1999, about 75% of Canada's daily newspapers changed ownership. Two newspaper chains in particular, Hollinger and Quebecor, acquired the majority of newspapers that changed hands. Not only did national concentration figures increase significantly, but county-level data indicate that multi-market contact also increased greatly over this period.

Background on the Canadian Mergers

The Canadian newspaper mergers can be traced to three large business acquisitions between 1996 and 2000:

•Through a series of deals in 1995 and 1996, Hollinger Inc. acquired a controlling stake in the Southam group of newspapers (which included 16 daily newspapers) as well as completed the purchase of 25 daily newspapers from the Thomson group and 7 independent dailies.

•On March 1st, 1999, Quebecor Inc. acquired the Sun Media chain of newspapers, including 14 daily papers, in a $983 million deal. Quebecor surpassed a bid by Torstar for purchasing Sun Media, but in turn sold four of its existing dailies to Torstar.

•On July 31st, 2000, Canwest purchased 28 daily newspapers from Hollinger Inc. The $3.5 billion purchase constituted the largest media deal in Canada's history. It allowed Canwest to go from having a zero stake in the Canadian newspaper market to becoming the country's biggest publisher, with 1.8 million daily readers.

Table I.
Aggregate Summary Statistics

Variable	Obs	Mean	SD	Min	Max
Weekday circ.	515	47,206	74,041	1,000	494,719
Saturday circ.	408	68,366	106,508	2,675	739,108
Sunday circ.	139	110,750	112,708	13,693	491,105
Average price ($)	515	0.58	0.15	0.21	1.04
Average pages	491	39.7	26.3	8	140
Weekday ad. rate ($)	511	2.3	3.0	0.4	25.6
Saturday ad. rate	399	2.9	3.7	0.5	26.9
Sunday ad. rate	137	4.0	2.7	1.0	12.5
Evening paper	515	0.52	0.50	0	1
French	515	0.11	0.31	0	1
Ad. rate per 10 K readers	511	0.98	0.86	0.22	7.70

Source: Editor and Publisher Magazine.

Table II.
County Level Summary Statistics

Variable	Obs	Mean	SD	Min	Max
Newspaper-Counties					
Weekday circ.	3,612	4,638	16,020	1	220,930
Saturday circ.	2,007	4,719	19,020	3	305,227
Sunday circ.	2,789	4,233	16,134	0	188,326
Weekly circulation	3,612	31,446	108,994	9	1598,203
Weighted Herfindahl (Group)	3,612	0.61	0.19	0.34	1
Counties					
Total daily circ.	1,053	15,909	38,366	1	324,940
Total weekly circ.	1,053	107,880	262,910	62	2353,779

Source: Audit Bureau of Circulation (ABC) and Statistics Canada.

Data

Our primary data source is Editor & Publisher Magazine, which is the weekly magazine of the newspaper industry.

Supplementary data are obtained from county level circulation figures provided by the Audit Bureau of Circulations (ABC). ABC is an independent, not-for-profit organization that is widely recognized as the leading auditor of periodical information in North America and many other countries. The ABC dataset contains extremely detailed information on the circulation of 73 Canadian newspapers for the years 1995-1999. These 73 newspapers constitute the major selling dailies in Canada.

Table III.
Newspaper Ownership by Group

Ownership	Daily Circulation	National Market Share
1995		
Southam	1285,746	0.26
Thomson	997,425	0.20
Torstar	494,719	0.10
Sun Media	472,054	0.09
Quebecor	421,841	0.08
Trans Canada (JTC)	283,472	0.06
Others	1058,793	0.21
Aggregate national circulation	5014,050	
1999		
Hollinger/Southam	2211,945	0.44
Quebecor/Sun Media	1160,572	0.23
Thomson	536,346	0.11
Torstar	460,654	0.09
Trans Canada (JTC)	257,316	0.05
Others	345,218	0.07
Aggregate national circulation	4972,051	
2002		
Canwest	1575,936	0.33
Quebecor	973,059	0.20
Torstar	671,231	0.14
Trans Canada (JTC)	415,345	0.09
Hollinger	259,523	0.05
Others	918,383	0.19
Aggregate national circulation	4813,477	

Table IV.
Fraction of Counties with Multi-Market Contact

		Hollinger	Quebecor	JTC	Torstar
Hollinger	1995 (90)	–	0.28	0.37	0.49
	1999 (199)	–	0.74	0.90	0.55
Quebecor	1995 (123)	0.38	–	0.97	0.09
	1999 (128)	0.73	–	0.98	0.98

Note: Figures in parentheses refer to the number of counties in which each chain—Hollinger or Quebecor—was present in the corresponding year.

Model

Consider the following Hotelling model which formalizes this intuition. There are two newspapers located at the end points of the line segment on $[0,1]$. Denote the newspaper at 0 by A and at 1 by B. There is a continuum of readers distributed uniformly along this line segment. The utility to a reader located at i from reading newspaper A is given by:

$$u_{iA\epsilon} = \delta(k_A) - p_A - \alpha \cdot i + \epsilon \tag{1}$$

Here α represents the reduction in utility experienced by readers further away from the newspaper, $\delta(k_A)$ is the quality of the newspaper which can depend on the quantity of ads k_A in the newspaper, p_A is the price of newspaper A, and ϵ represents an idiosyncratic taste for newspapers. We assume that ϵ follows a uniform distribution given by:

$$\varepsilon \sim U\left(0, \gamma\right]$$

This allows readers' preferences to vary along two dimensions: their relative taste for newspapers A and B, and their overall taste for newspaper reading. These two taste parameters are orthogonal to each other. The assumption that ε is different from zero is important, since if there is no ε then a newspaper can perfectly screen readers.

The utility from newspaper B is analogously given by:

$$u_{iB\epsilon} = \delta(k_B) - p_B - \alpha \cdot (1 - i) + \epsilon$$

Publishers earn revenue from newspaper sales, as well as from advertising. Advertisers are located at the endpoints 0 and 1 and have a greater valuation of readers located closer to them. Specifically, assume that advertisers receive profits of q for each consumer that buys a product at their store. The probability that a consumer located at i who reads the newspaper will buy the product from an advertiser located at 0 is given by:

$$P^0(i) = \frac{\beta}{q} - \frac{w}{q} \cdot i$$

Thus, readers located further away from the advertiser are less likely to visit the store, and w captures the decrease in the probability of visiting a store if a consumer is located further away from the store. This implies that the advertiser's willingness to pay for a consumer located at i is given by $\beta - \omega \cdot i$. Analogously, the willingness to pay by an advertiser at 1 for the same reader is $\beta - \omega \cdot (1 - i)$.

The revenue of newspaper A from selling to a reader located at i is given by:

$$R_A(i) = p_A + \beta - \omega \cdot i$$

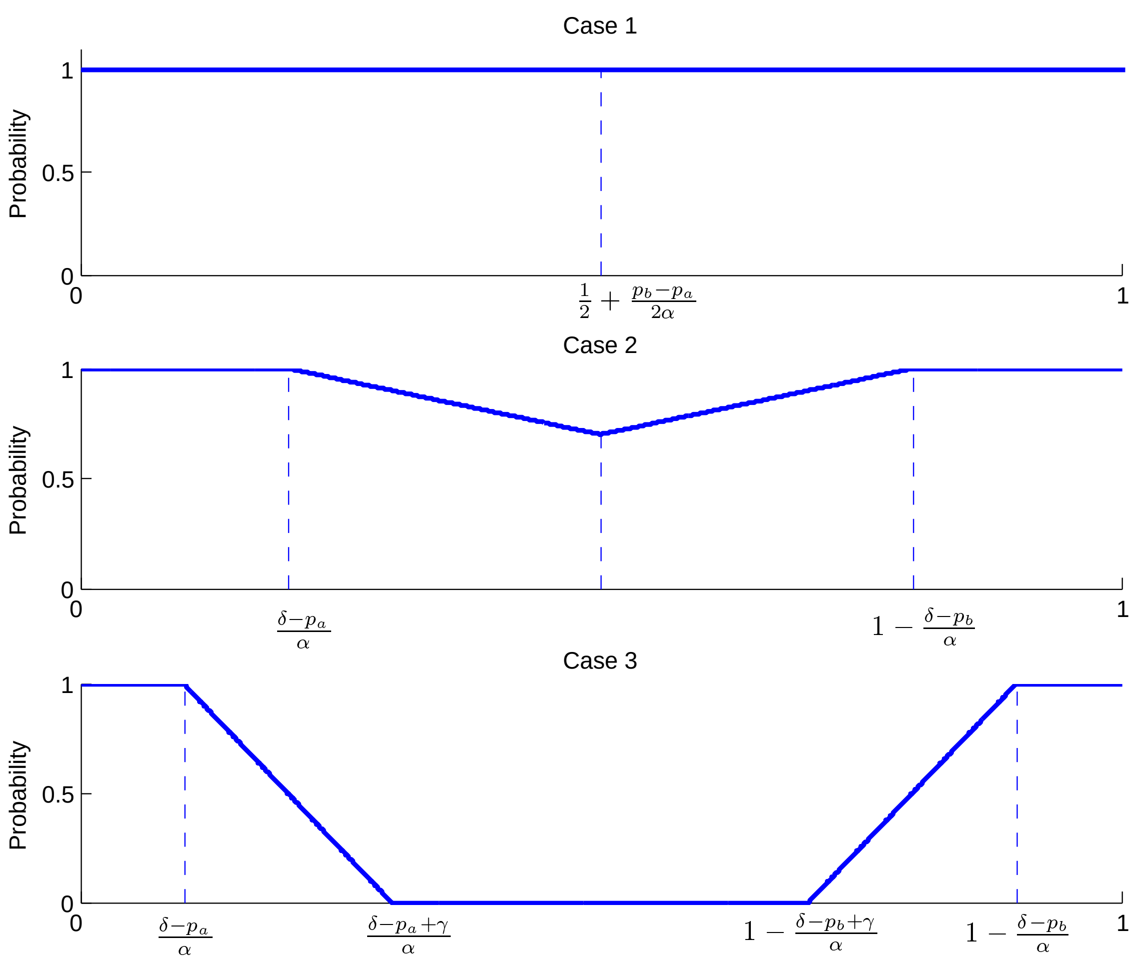
and note that the newspaper can extract all of the advertisers' surplus.

The total profit to newspaper A is given by

$$\Pi_A = \int_0^1 [p_A + \beta - \omega \cdot i] P(i = A) di - C(q_A)$$

where $C(q_A)$ is the newspaper's cost of delivering q_A papers, and $q_A = \int_0^1 P(i = A) di$.

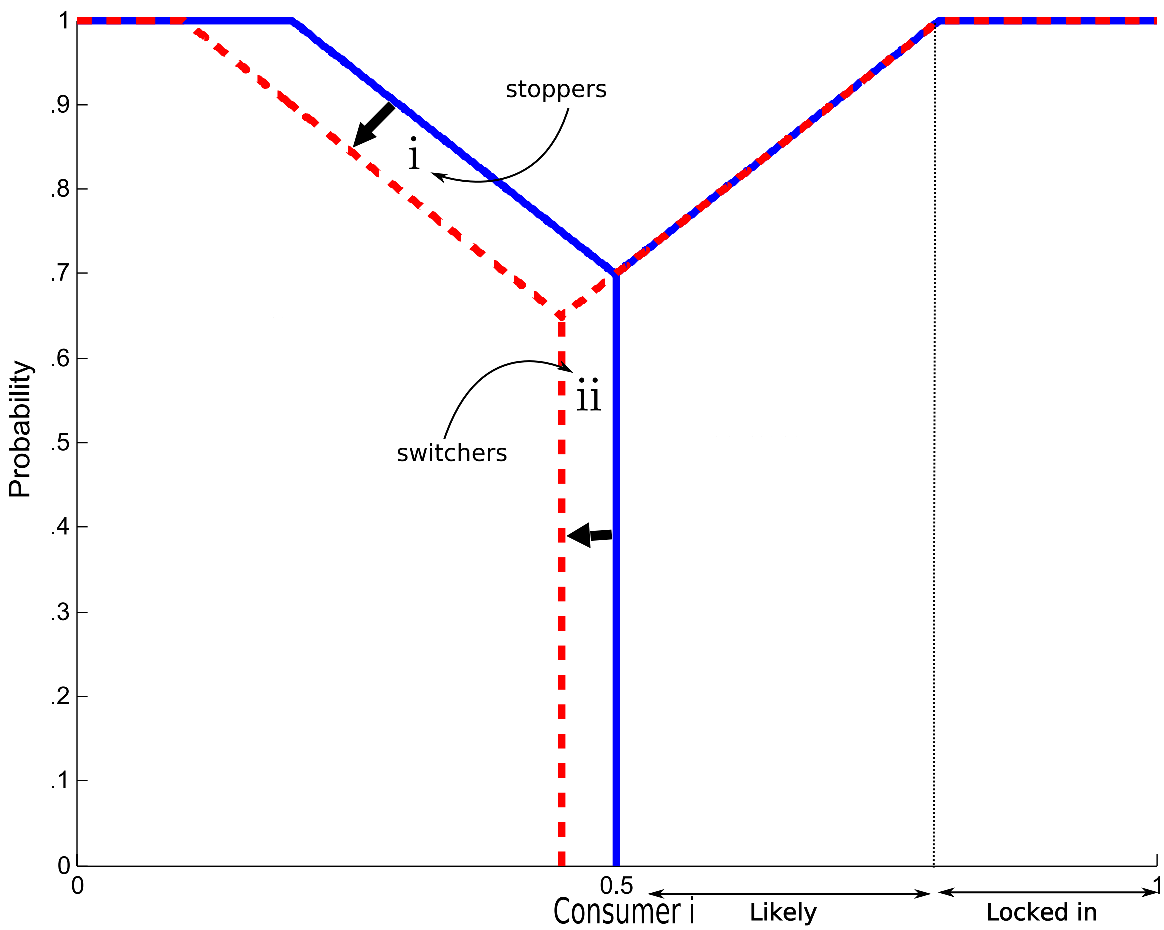
There are three possible cases:



The probability that a reader at i purchases newspaper A is given by:

$$P(i = A) = \begin{cases} 1 & \text{if} \quad i \in \left[0, \frac{\delta - p_A}{\alpha}\right] \\ 1 - \frac{\alpha \cdot i - \delta - p_A}{\gamma} & \text{if} \quad i \in \left[\frac{\delta - p_A}{\alpha}, \frac{1}{2} + \frac{p_B - p_A}{2\alpha}\right] \\ 0 & \text{if} \quad i \in \left[\frac{1}{2} + \frac{p_B - p_A}{2\alpha}, 1\right] \end{cases}$$

$$R_A = \int_0^{\frac{\delta - p_A}{\alpha}} (p_A + \beta - \omega i)\, di + \int_{\frac{\delta - p_A}{\alpha}}^{\frac{1}{2} + \frac{p_A - p_B}{2\alpha}} (p_A + \beta - \omega i) \left[1 - \frac{\alpha i - p_A - \frac{p_A - p_B}{2\alpha}\delta}{\gamma}\right] di$$



When newspapers A and B merge, the price of the newspaper A will now reflect the effect of p_A on profits of newspaper B.

Specifically, the sign of the change in price depends on $\frac{\partial \Pi_B}{\partial p_A}$ given by:

$$\frac{\partial \Pi_B}{\partial p_A} = \frac{1}{2\alpha}\left(p_B + \beta - \omega\left(\frac{1}{2} + \frac{p_A - p_B}{2\alpha}\right) - \frac{\partial C}{\partial q}\right) \quad (2)$$

Thus, the sign of $\frac{\partial \Pi_B}{\partial p_A}$ depends on the the sign of $\left(p_B + \beta - \omega\left(\frac{1}{2} + \frac{p_A - p_B}{2\alpha}\right) - \frac{\partial C}{\partial q}\right)$, the profitability of the consumer who is indifferent between newspaper A and newspaper B. Call the consumer located at $\frac{1}{2}+\frac{p_B-p_A}{2\alpha}$ the *switching consumer*, i.e. the consumer who is indifferent between purchasing newspaper A and newspaper B. In particular, a necessary condition for the switching consumer to yield negative value to the newspaper that they purchase is that the marginal cost of the newspaper, $\frac{\partial C}{\partial q}$, is higher than the price charged to readers, p_A.

We now turn to the effect of a merger on advertising price. Typically, advertising prices are quoted on a per-thousand basis, i.e. it is assumed that total prices are proportional to the number of readers. The price per reader for newspaper A (denoted p_A^{ra}) is given by:

$$p_A^{ra} = \frac{\int_0^1 (\beta - \omega i) P(i = A) di}{\int_0^1 P(i = A) di} \tag{3}$$

which can be rewritten as:

$$p_A^{ra} = \frac{\int_0^{\frac{\delta - p_A}{\alpha}} (\beta - \omega i)\, di + \int_{\frac{\delta - p_A}{\alpha}}^{\frac{1}{2}+\frac{p_A-p_B}{2\alpha}} (\beta - \omega i) \left[1 - \frac{\alpha i - p_A - \frac{p_A - p_B}{2\alpha}\delta}{\gamma}\right] di}{\frac{\delta - p_A}{\alpha} + \int_{\frac{\delta - p_A}{\alpha}}^{\frac{1}{2}+\frac{p_A-p_B}{2\alpha}} \left[1 - \frac{\alpha i - p_A - \frac{p_A - p_B}{2\alpha}\delta}{\gamma}\right] di} \tag{4}$$

The price per reader for advertisers will increase after the merger if the price charged to readers increases. If p_A increases, then p_A^{ra} will increase as well, since the average i of readers of newspaper A goes down. We have already established that the price charged to readers could rise or fall after the merger depending on the profitability of the switching consumer. Thus the change in the price for advertisers is ambiguous as well.

Results

We identify the average treatment effect of the merger using both difference-in-differences and difference-in-differences matching methods. We compare newspapers that changed hands versus those that did not; as well as those in the dominant newspaper chains versus the rest. Because the predictions of the model are ambiguous, i.e. they depend on parameters of the valuation of advertisers and consumers that are difficult to estimate, we use difference-in-difference and matching approaches to evaluate the impact of mergers on prices.

Notice that we are adopting the language of natural-experiments; however in reality we do not believe that the treatment and control groups are randomly chosen representative samples, since firms self-select into these groups. Nevertheless, since these labels have become commonplace in the quasi-experimental literature in economics, we shall continue to use them here. Moreover, it is not clear that a truly natural experiment is useful for a Competition Authority deciding on whether to approve a merger. The collection of mergers that come before the Competition Authority is never exogenous since firms initiate mergers. In addition, mergers which are likely to increase market power will also be more profitable for the merging firms.

We will look at two different merger treatments T_{it}:

A. Newspapers with changed ownership between 1995 and 1999/2002 .

B. Newspapers acquired by Quebecor or Hollinger between 1995 and 1999 and by Canwest between 1999 and 2002.

We study the effect of these treatments on several outcome variables (henceforth denoted y_{it}) of interest to analyzing the effect of mergers.

The standard method for difference-in-differences calculations involves comparing the changes in the means for two groups – the treatment and control groups – before and after the treatment. The outcomes are determined by:

$$y_{it} = \underbrace{\mu_i}_{\text{newspaper fixed effect}} + \underbrace{\delta_t}_{\text{year effect}} + \underbrace{\alpha T_{it}}_{\text{treatment effect}} + u_{it} \tag{5}$$

where α is the effect of mergers on the outcome variable, and we allow for time trends (δ_t) and newspaper fixed effects (μ_i). The difference in difference estimator is just the difference between the change in Δy_{it} for the merged group and unmerged group:

$$\alpha = E(y_{it} - y_{it-1}|T_{it} = 1) - E(y_{it-1} - y_{it}|T_{it} = 0) \tag{6}$$

For the difference in difference estimate of α to be correct, we need to assume that assignment to the merger group is not confounded: $T_{it} \perp (u_{it} - u_{it-1})$.

For instance, if it was the case that Hollinger acquired small newspapers, and the ad rates for small newspaper were falling from 1995 to 2002, this would violate unconfoundedness. We relax this assumption by presenting estimates using the difference-in-differences matching estimators which only requires unconfoundedness conditional on observables, i.e. $T_{it} \perp (u_{it} - u_{it-1}) | X_{it}$. The difference-in-differences matching estimators will yield similar conclusions as the straight difference in differences estimator.

TABLE V.
DIFF-IN-DIFF MATCHING ESTIMATE OF THE EFFECT OF OWNERSHIP CHANGES AND OWNERSHIP BY HOLLINGER OR QUEBECOR USING THE NEAREST NEIGHBOUR MATCHING ESTIMATOR

	Ownership Change				Hollinger-Quebecor		Canwest-Quebecor	
	1995–1999		1995–2002		1995–1999		1995–2002	
Change in Variable	Coef.†	S.E.	Coef.†	S.E.	Coef.†	S.E.	Coef.†	S.E.
Circulation price	−0.01	(0.02)	0.03	(0.07)	0.00	(0.02)	0.01	(0.02)
Ad rate	0.13	(0.15)	−0.21	(0.42)	0.24	(0.18)	0.24	(0.23)
Average pages	−1.86	(1.57)	0.82	(4.40)	2.90	(1.55)	1.38	(2.03)
Rate per 10 K	−0.13*	(0.06)	−0.10	(0.28)	−0.12	(0.08)	−0.01	(0.11)
Log circ.	0.00	(0.02)	−0.10	(0.16)	0.00	(0.02)	0.06	(0.05)
Circulation daily	288	(1,448)	−4,688	(8,335)	1,014	(1,166)	1,945	(2,730)
N	97		92		97		92	

*Significant at the 5% level.
†Matching variables: daily circulation, circulation price, pages, province, ad rate per 10 K, ad rate.

Control Function Techniques

Suppose we have a probit regression

$$y_{it} = X_{it}\beta + \alpha p_{it} + \epsilon_{it} \geq 0$$

But we are worried that price is correlated with the product level unobservable, i.e. $E[\epsilon_{it} p_{it}] > 0$. This will bias our estimates of α in the probit. Note as well that there is no way to do IV, since there isn't an unobservable ϵ that we can back out directly.

Thus one idea is to use a *control function* instead. We can decompose ϵ_{it} into:

$$\epsilon_{it} = \xi_{it} + \eta_{it}$$

where $E[p_{it}\eta_{it}] = 0$, i.e. the uncorrelated error term. So all that we are really worried about is the ξ_{it} component. Now suppose we can build a control for ξ_{it}, something that will pick out this term. One idea would be to use the product's market share s_{jt}, where we might know that $\xi_{it} = f(s_{jt})$, market share is monotonically related to ξ_{it}. Then we can run the following probit, putting in the "control function" into the regression:

$$y_{it} = X_{it}\beta + \alpha p_{it} + f(s_{jt}) + \eta_{it} \geq 0$$

and now we don't have a biased estimate of α.

Control Function Techniques: Productivity

Another place where control functions are commonly used are in production function analysis. Say we have a production function such as:

$$y_{it} = \beta_0 + \beta_l l_{it} + \beta_k k_{it} + \omega_{it} + \epsilon_{it}$$

where ω_{it} is the true productivity and ϵ_{it} is the measurement error. The simultaneity and selection problems imply that $E[l_{it}\omega_{it}] > 0$ and $E[k_{it}\omega_{it}] > 0$. However, the measurement error term is not an issue, hence $E[l_{it}\epsilon_{it}] = 0$ and $E[k_{it}\epsilon_{it}] = 0$.

Suppose investment is strictly increasing in ω_{it}. Then we have:

$$i_{it} = i^*(\omega_{it}, k_{it})$$

but this means that we can invert this function to get

$$\omega_{it} = g(i_{it}, k_{it})$$

And put this into the production function regression:

$$y_{it} = \beta_0 + \beta_l l_{it} + \beta_k k_{it} + g(i_{it}, k_{it}) + \epsilon_{it}$$

Now you can see that we won't be able to get at β_k, but we can estimate the labor coefficient β_l consistently in this setup. If you look at the first stage of Olley and Pakes (1996), this is exactly what they are doing.