Sibling Competition and Marital Sorting in South Asia

Skand Goel

September 16, 2016

1 Introduction

In many developing countries, lack of female empowerment and mobility imply that access to resources within the household, and outside it, are governed sharply by marital norms. In South Asia, these resource constraints interact with the peculiar tradition of arranged marriage, where families play an important role in deciding who marries whom. Vogl (2013) finds that women with younger sisters are likely to marry earlier and have a worse marital match than single daughters or women with younger brothers. This paper extends the analysis in Vogl to study how potential brides (or their families, in this case) substitute match attributes in the presence of sibling competition. In order to do so, I adopt the semiparametric approach proposed by Chiappori, Oreffice and Quintana-Domeque (2012) (henceforth COQ).

Vogl (2013) studies sibling competition and marriage institutions in South Asia (Bangladesh, India, Nepal and Pakistan) using DHS data. He compares marriage age and marital outcomes for women who have younger brothers with those who have younger sisters. In the absence of sex selection by parents, this identifies the effect of next-born sibling's gender. Exploiting this natural experiment within the family, he documents that girls with younger sisters are more likely to leave home early (a good proxy for marriage in virilocal societies such as South Asia). Presence of siblings, especially younger sisters, in the family puts pressure on older sisters to get married earlier. This, in turn, appears to have long term impacts as girls who leave home earlier have lower human capital indicators and are matched with "low-quality" (in terms of

wealth, education and occupation) husbands[1].

The more interesting part of Vogl's argument deals with long run consequences of sibling competition. The effect of a younger sister on marriage age is statistically and economically significant, so are the effects on human capital acquisition. However, the quality of the spousal match does not show a drastic difference. Nonetheless, it is possible that effects of sibling competition on match quality operates in another way. Specifically, it is possible that a girl's family that is too eager to get her married off trades-off attributes of a potential husband in a different way as compared to a family that has until now resisted the pinch of social stigma.

The paper is structured as follows. The next section briefly discusses the COQ methodology and attempts to replicate their findings with in the US data. Unfortunately, it is not possible to clean the data in the same way the authors have done, so the results do not match up. The third section replicates the relevant results from Vogl (2013) and moves on to estimating substitution patterns between husbands' attributes. I conclude with some important issues for future research.

2 Estimating substitution patterns in matching

2.1 COQ methodology

In order to understand differences in tradeoffs in matching made at the time of marriage by , I use the COQ methodology. A matching model is defined by a finite population of men and women with some observable and unobservable characteristics. Each individual is characterized by such a vector. A matching (equilibrium) is a product measure on the space of all characteristics such that the marginals equal initial distributions of those characteristics in the population. Typically, there is also a stability requirement.

COQ assume that individuals' preferences over possible partners can be summarized by a *single index* that depends on all the relevant *observable* attributes of the partner. Formally, suppose $X_i = (X_i^1, \ldots, X_i^k)$ is a vector of k observable attributes of female i. The single index (separability) assumption states that these attributes matter for *any* man j only through some function (or index) $I(X)$. As a result, if there are two females i and i' with different characteristics but $I(X_i) = I(X_{i'})$ and if they have the same vector of unobservables ($\eta_i = \eta_{i'}$), then they are perfect substitutes on the

[1]Long term impacts have only been studied for Nepal because of limited data on long term outcomes in the other three countries.

marriage market. The second assumption of conditional independence states that the distribution of η conditional on $I(X)$ is atomless and indpendent of X. Symmetric conditions hold for any male j with observables $Y_i = (Y_i^1, \ldots, Y_i^p)$, unobservables $\epsilon_i = (\epsilon_i^1, \ldots, \epsilon_i^q)$ and index function $J(.)$. In principle, no restrictions (parametric or otherwise) are imposed on the index I.

Due to separability, any difference between agents with the same index will be on account of unobservables. However, conditional independence implies that given the same index, these agents will have the same distribution of observables. As a result, two individuals with the same value of the index will have the same distribution of characteristics. In particular, separability implies that the distribution, and so any moment, of $Y_i^k|X_j$ will be a function only of $I(X_j)$, for example

$$\mathbb{E}\left[Y_i^k|X_j\right] = \phi_k\left[I(X_j)\right]. \tag{1}$$

Note that the marginal rate of substitution between female characteristics r and t is defined as

$$MRS_j^{r,t} = \frac{\partial I/\partial X_j^t}{\partial I/\partial X_j^r}.$$

From (1) it is clear that this can be recovered from the data by regression as follows

$$\frac{\partial I/\partial X_j^t}{\partial I/\partial X_j^r} = \frac{\partial \mathbb{E}\left[Y_i^k|X_j\right]/\partial X_j^t}{\partial \mathbb{E}\left[Y_i^k|X_j\right]/\partial X_j^r} \tag{2}$$

Estimation. The estimation strategy followed by COQ is based on (2). Furthermore, they assume I and J are linear functions. In this case, $MRS_j^{r,t}$ can just be recovered by regressing any male attribute s on female attributes r and t:

$$Y^s = \beta_0 + \beta_1^s X^t + \beta_2^s X^r$$

(2) further implies the restriction that for any two male attributes s and s'

$$\frac{\beta_1^s}{\beta_2^s} = \frac{\beta_1^{s'}}{\beta_2^{s'}}. \tag{3}$$

COQ estimate this multiple system of equations by Seemingly Unrelated Regression and use (3) as a test for empirical identification of the MRS in this model.

2.2 Replication of COQ results

After cleaning their data, COQ are left with 659 observations. Using the author provided data preparation do-file, I had 778 observations. However, some variables were missing in many of these observations, e.g. log wage of the husband was missing for 255 observations, which possibly accounts for much of the discrepancy seen here.

Table 1 below displays correlations between attributes in the sample. Matching patterns are qualitatively similar to those in the COQ, but magnitudes differ. Most importantly, unlike COQ, I do not find statistically significant relationships between husband's log wage and wife's attributes. Consistent with this, in Table 2, I do not recover the strong substitution between husband's wage and husband's BMI. This is true for both standard and augmented regressions. While the estimated MRS in each of the equation is statistically indistinguishable from zero, the test for linear specification of the index function holds up. The constrained model is not statistically different as per the likelihood-ratio test, which is consistent with empirical identification of MRS.

Looking at the MRS between wife's characteristics in Table 3, there are similar issues as above. One thing that stands out is that the linear specification of the index function is quite rejected. This is bolstered by the results in Table 4, which allows for heterogeneity of MRS by spousal height. Estimated MRS under the proportionality restriction imposed in (3) is small and statistically indistinguishable from zero. This constrained model is also statistically different from the unconstrained model.

Table 1 Sample Correlation

Variables	Wife's BMI	Husband's BMI	Wife's education	Husband's log wage	Husband's education
Wife's BMI	1.000				
Husband's BMI	0.210	1.000			
	(0.000)				
Wife's education	-0.131	-0.097	1.000		
	(0.000)	(0.007)			
Husband's log wage	-0.038	0.096	0.074	1.000	
	(0.387)	(0.027)	(0.090)		
Husband's education	-0.199	-0.002	0.623	0.113	1.000
	(0.000)	(0.963)	(0.000)	(0.009)	

Notes: p-values in parantheses.

Table 2 SUR Regressions of wife's characteristics on husband's characteristics

	Standard Regressions		Augmented Regressions	
	Wife's BMI	Wife's Education	Wife's BMI	Wife's Education
	Unconstrained Model			
Husband's log wage	-0.189	0.127	-0.237	0.068
	(0.149)	(0.134)	(0.120)	(0.105)
Husband's BMI	0.315***	-0.120***	0.313***	-0.129***
	(0.059)	(0.042)	(0.069)	(0.038)
Observations	524		523	
Within columns:	-0.601	-1.051	-0.758*	-0.531
Husband's log wage/	(0.234)	(1.390)	(0.150)	(0.655)
Husband's BMI	$\chi^2(1) = 0.162$		$\chi^2(1) = 0.088$	
	p-value = 0.687		p-value = 0.766	
Across columns:	0.023	0.040	0.031*	0.021
Husband's log wage ×	(0.018)	(0.041)	(0.019)	(0.029)
Husband's BMI	$\chi^2(1) = 0.170$		$\chi^2(1) = 0.087$	
	p-value = 0.680		p-value = 0.768	
	Constrained Model			
Ratio of coefficients	-0.666***		-0.697	
	(0.449)		(0.444)	
Husband's BMI	0.310	-0.123***	0.314***	-0.127***
	(0.055)	(0.035)	(0.053)	(0.032)
LR Test	$\chi^2(1) = 0.432$			
	p-value = 0.511			

Notes: Standard errors in parantheses. *p-value<0.1 **p-value<0.05 ***p-value<0.01

Table 3 SUR Regressions of husband's characteristics on wife's characteristics

	Standard Regressions		Augmented Regressions	
	Husband's BMI	Husband's log wage	Husband's BMI	Husband's log wage
	Unconstrained Model			
Wife's education	-0.107**	0.570***	-0.133***	0.514***
	(0.037)	(0.024)	(0.048)	(0.026)
Wife's BMI	0.176***	-0.094***	0.176***	-0.097***
	(0.022)	(0.022)	(0.032)	(0.017)
Observations	773		773	
Within columns:	-0.609**	-6.084***	-0.755**	-5.322***
Wife's education/	(0.284)	(1.192)	(0.322)	(1.236)
Wife's BMI	$\chi^2(1) = 11.156$		$\chi^2(1) = 18.926$	
	p-value = 0.001		p-value = 0.000	
Across columns:	0.010**	0.100***	0.013**	0.090***
Wife's education ×	(0.004)	(0.019)	(0.006)	(0.015)
Wife's BMI	$\chi^2(1) = 30.888$		$\chi^2(1) = 18.815$	
	p-value = 0.000		p-value = 0.000	
	Constrained Model			
Ratio of coefficients	-4.255***		-3.908***	
	(0.679)		(0.928)	
Wife's BMI	0.046***	-0.129***	0.057***	-0.126***
	(0.011)	(0.019)	(0.018)	(0.021)
LR Test	$\chi^2(1) = 43.636$			
	p-value = 0.000			

Notes: Standard errors in parantheses. *p-value<0.1 **p-value<0.05 ***p-value<0.01

Table 4 SUR Regressions of individual characteristics on spousal attributes allowing for spousal height interaction

	Wife's BMI	Wife's Education	Husband's BMI	Husband's log wage
	Unconstrained Model			
β	-0.396*	0.172	-0.080	0.003
	(0.204)	(0.195)	(0.079)	(0.038)
π	0.461***	-0.137**	0.212***	-0.033**
	(0.087)	(0.066)	(0.045)	(0.014)
ρ	6.211*	-0.493	2.061	-1.196
	(3.512)	(2.948)	(2.274)	(0.898)
θ	0.512*	-0.085	-0.076	0.019
	(0.300)	(0.205)	(0.088)	(0.050)
δ	-0.318**	0.033	-0.040	0.038*
	(0.145)	(0.118)	(0.075)	(0.022)
	Constrained Model			
κ	-0.856		-0.605	
	(0.749)		(1.879)	
ψ	0.454***	-0.132**	0.198***	-0.027*
	(0.090)	(0.052)	(0.069)	(0.014)
λ	7.199**	-0.466	-0.401	0.378
	(3.366)	(0.574)	(2.190)	(1.223)
ξ	-0.331**	0.028	0.023	-0.017
	(0.148)	(0.027)	(0.095)	(0.065)
Observations	519		519	
LR Test	$\chi^2(1) = 9.597$		$\chi^2(1) = 7.304$	
	p-value = 0.002		p-value = 0.007	

Notes: Standard errors in parantheses. *p-value<0.1 **p-value<0.05 ***p-value<0.01

3 Sibling Competition and marriage

3.1 Existing results

I begin by replicating relevant results from Vogl (2013). Vogl uses Demographic and Health Surveys (DHS) from Bangladesh, India, Nepal, and Pakistan. The specific survey years are Bangladesh (1993–94, 1996–97, 1999–2000, 2004, 2007), India (1992–93, 1998–99, 2005–6), Nepal (1996, 2001, 2006), and Pakistan (1990–91, 2006–7). The first set of data come from the DHS fertility history module, where women list all their live births and some attributes of children (including coresidence). The fertility history module does not track these births once they leave the parental house. Long run outcomes are derived from the sibling history module of DHS, which asks respondents to

list all children ever born to their biological mothers. Unfortunately, the only nationally representative sibling history module in South Asia comes from the 2006 Nepal DHS.

Table 5 Coresidence by gender of younger sibling

	(1) First subsequent pregnancy (conditional on ≥1 more pregnancy)	(2) Second subsequent pregnancy (conditional on ≥2 more pregnancy)
younger sister	-0.0375***	-0.0181*
	(0.0103)	(0.0105)
Mean among women w/ a younger brother	0.437	0.407
N	7702	6585

Notes: OLS estimates. Brackets contain standard errors clustered at the PSU level. Only observations with singleton current and subsequent births are included. Each cell reports a coefficient from a separate regression. The dependent variable equals 1 if the woman resides with her mother, 0 otherwise. All regressions include fixed effects for age, mother's region of residence, survey year, and the exact composition of older siblings by birth order and sex. Regressions also control for spacing from the previous birth, maternal and paternal educational attainment, maternal age, and rural residence.
Source. DHS Fertility Histories.

Table 5 presents the main result in Vogl (2013) for Nepal. Column 1 regresses coresidence on the presence of an immediate younger sister vs an immediate younger brother, while column 2 does the same for one after the next youngest. Presence of a younger sister as opposed to a brother makes coresidence less likely. Table 6 presents the more important result in our context. Sibling competition is associated with less educated and less skilled husbands. However, these magnitudes are not large, and the small negative do not appear when we look at a standardized wealth index and spouse age. Therefore, it might be useful to study actual marriage patterns and how they differ among those who face sibling competition.

3.2 Estimating matching patterns

Table 6 Coresidence by gender of younger sibling

	(1) Skilled occupation	(2) Educational attainment	(3) Wealth index	(4) Age
younger sister	-0.0397***	-0.185*	-0.0366	0.0525
	(0.0124)	(0.106)	(0.0327)	(0.150)
Mean among women w/ a younger brother	0.427	4.107	0.00933	42.98
N	6379	6289	3346	5965

Notes: Brackets contain standard errors clustered at the PSU level. Only ever married women between ages 30 and 49 included. Women born in the same year as a sibling, women with two next-youngest siblings born in the same year, and women with no younger siblings are excluded. All regressions control for religion, spacing from the respondent's birth, the year the respondent's mother initiated childbearing, and birth and survey year fixed effects and include fixed effects for the exact composition of older siblings by birth order and sex.

This sibling rivalry is unusual in that it is not captured by the standard budget constraint. Vogl's explanation for this phenomenon is based on a model of marriage search where the marriage market value of a girl decreases with age and there is a cost to the family (possibly psychological or stigma) associated with an unmarried daughter of marriageable age. As age goes up, the supply of feasible matches goes down and the expected cost to the family increases. This generates predictions that are borne out by the data as discussed above. Note that age is a more important factor as per this mechanism so we control for it in our regressions. Moreover, COQ's setup is very general - it can accomodate search as well. We need to assume that matching technology is symmetric and also satisfies a single index type of assumption. These are reasonable assumptions to make in our context.

Table 7 Sample correlation

Variables	wife's age	wife's edu	wife's height	husband's age	husband's edu	husband in skilled occ.	Wealth index
wife's age	1.000						
wife's edu	-0.159	1.000					
	(0.000)						
wife's height	-0.035	0.097	1.000				
	(0.020)	(0.000)					
husband's age	0.729	-0.130	-0.030	1.000			
	(0.000)	(0.000)	(0.055)				
husband's edu	-0.172	0.548	0.112	-0.207	1.000		
	(0.000)	(0.000)	(0.000)	(0.000)			
husband in skilled occ.	-0.120	0.282	0.032	-0.101	0.370	1.000	
	(0.000)	(0.000)	(0.034)	(0.000)	(0.000)		
Wealth index	-0.027	0.570	0.086	-0.010	0.516	0.395	1.000
	(0.117)	(0.000)	(0.000)	(0.562)	(0.000)	(0.000)	

Notes: p-values in parantheses.

Table 7 displays matching patterns in the data. These are indicative of assortative matching - more educated, taller (an indicator of good health in malnourished populations) tend to have younger, more educated, skilled and wealthier husbands. The two male characteristics I consider are age and education. Age is not as much of a decider in a man's marriage as is it is a woman's. Indeed, the above mechanism does not hinge on age of the man at all - in fact, waiting longer is more likely to get you an older husband - so age becomes a quality attribute that does not govern the search process. Education is a good proxy for income and personal acheivement (rather than wealth, which might be bequeathed). The results with wealth were not very different.

In order to understand substitution patterns of women, I run SUR regression in the same manner as Table 1. Every regression controls for own age and has has region-survey year-locality(rural/urban) fixed effects. Table 8 does this for the full sample, while Table 9 does it separately for the two groups defined by younger sibling gender. For the full sample, the simple linear specification that I use is not enough to identify MRS. The restrictions implied by 3 do not hold, either in the Wald test of comparing coefficient ratios or in the likelihood ratio test. Unfortunately, similar results show up in the regressions by sibling gender. While MRSs estimated in the two equations for each sample (defined by sibling gender) is statistically indistinguishable from each other, they are also very noisy estimates. Moreover, the likelihood ratio test rejects the null of the constrained model being nested in the unconstrained model.

Table 8 SUR Regression of wife's characteristics on husband's characteristics - Full sample

	Wife's BMI	Wife's Education
Husband's age	49.545**	0.017
	(19.806)	(0.014)
Husband's wealth	150.232	1.893***
	(275.099)	(0.267)
Observations	3,222	
Within columns:	0.330	0.009
Husband's age/	(0.628)	(0.008)
Husband's wealth	$\chi^2(1) = 0.264$	
	p-value = 0.607	
Across columns:	93.794**	2.567
Husband's age ×	(39.146)	(4.787)
Husband's wealth	$\chi^2(1) = 6.157$	
	p-value = 0.013	
	Constrained Model	
Ratio of coefficients	0.323***	
	(0.099)	
Husband's BMI	150.738***	0.089**
	(42.077)	(0.0363)
LR Test	$\chi^2(1) = 4.585$	
	p-value = 0.032	

Notes: Standard errors in parantheses. Every regression controls for own age and has has region-survey year-locality(rural/urban) fixed effects.

Table 9 SUR Regression of wife's characteristics on husband's characteristics - By sibling gender

	Younger brother		Younger sister	
	Wife's BMI	Wife's Education	Wife's BMI	Wife's Education
	Unconstrained Model			
Husband's age	35.124	-0.003	56.906**	0.029
	(30.396)	(0.021)	(27.734)	(0.021)
Husband's wealth	720.772*	2.330***	-618.368	1.230**
	(320.941)	(0.299)	(518.444)	(0.537)
Observations	1663		1559	
Within columns:	0.049	-0.001	-0.092	0.024
Husband's age/	(0.049)	(0.010)	(0.097)	(0.016)
Husband's wealth	$\chi^2(1) = 1.386$		$\chi^2(1) = 1.299$	
	p-value = 0.239		p-value= 0.254	
Across columns:	81.841	-1.945	69.980	-17.892
Husband's age ×	(70.436)	(15.522)	(49.817)	(15.551)
Husband's wealth	$\chi^2(1) = 2.048$		$\chi^2(1) = 2.940$	
	p-value = 0.152		p-value = 0.086	
	Constrained Model			
Ratio of coefficients	-8.703		-0.311	
	(110.261)		(0.105)***	
Husband's BMI	-4.202	0.000	-192.926	0.003
	(53.202)	(0.003)	(58.856)***	(0.042)
LR Test	$\chi^2(1) = 57.007$		$\chi^2(1) = 66.497$	
	p-value = 0.000		p-value = 0.000	

Notes: Standard errors in parantheses. Every regression controls for own age and has has region-survey year-locality(rural/urban) fixed effects.

Conclusion

This paper reviewed Vogl's findings on sibling competition in South Asia and attempted to use COQ's methodology to further understand marital sorting patterns in the presence of such a phenonmenon. While results from this analysis are inconclusive, this is an important question given the pervasiveness of the issue. Understanding marriage patterns and their consequences in developing countries is becoming an exciting area of research (Banerjee et al 2012; Sautmann 2014; Fulford 2015). This is also a phenomenon in rapid flux as more women join the labor force and holds of the joint family dwindle.

Two important questions that have not yet been studied in this context come to mind. The interaction between marital sorting and fertility choice has not receievd much attention in the literature. It is an important attribute over which people match,

and it has first order consequences for the people involved. Also, macroeconomic implications of household sorting in developing countries still constitutes an incipient literature (Cuberes et al forthcoming). Marriage sorting also governs labor force participation, but it does so in general equilibrium.

These are exciting avenues for research, and this paper is an exploratory attempt in that direction.

References

Banerjee, A., Duflo, E., Ghatak, M. and Lafortune, J., 2013. Marry for what? Caste and mate selection in modern India. American Economic Journal: Microeconomics, 5(2), pp.33-72.

Chiappori, P.A., Oreffice, S. and Quintana-Domeque, C., 2012. Fatter attraction: anthropometric and socioeconomic matching on the marriage market. Journal of Political Economy, 120(4), pp.659-695.

Fulford, S.L., 2015. Marriage migration in India: Vast, varied, and misunderstood. Working Paper 820, Boston College. URL http://fmwww. bc. edu/EC-P/wp820. pdf.

Sautmann, A., 2011. Partner Search and Demographics: The Marriage Squeeze in India. Available at SSRN 1915158.

Teignier, M. and Cuberes, D., 2014. Aggregate Costs of Gender Gaps in the Labor Market: A Quantitative Estimate.

Vogl, T.S., 2013. Marriage Institutions and Sibling Competition: Evidence from South Asia. Quarterly Journal of Economics, 128(3), pp.1017-1072.

Methods appendix

I found it hard to implement non-linear SUR in canned packages in STATA, R or Python. Most packages such as systemfit (in R) or sureg (in STATA) are meant for linear models. So I estimated most models using the "gmm" package in STATA. It can handle non-linear systems of equations. Moreover, estimating SUR as GMM has the added advantage of providing a natural way of doing Likelihood-ratio tests. Nonetheless, it is difficult to incorporate many fixed effects in this setup. Consequently, I demeaned data and bootstrapped standard errors for correct inference. Also, the gmm package is limited in its optimization capabilities. It is very sensitive to initialization values (as is apparent from the estimates) and does not allow for non-derivative based optimization routines. Whenever "gmm" did not work, I used STATA's "nlsur" com-

mand that is meant for non-linear SUR regressions. This command, however, does not easily facilitate likelihood-ratio tests.