

1. (30 points) Using data from the Survey of Consumer Finances, I examined the determinants of credit card debt (ccbal), which is a person's unpaid credit card balance. ccbal is either zero or positive for all respondents. The results of an OLS regression and a tobit model are reported below :

COEFFICIENT	OLS	Tobit
Age	-9.242*** (1.53)	-79.05*** (3.14)
Annual household income (in \$1000s)	-0.0782*** (0.023)	-4.56*** (0.23)
Married	445.0*** (41.4)	897.3*** (81.0)
Constant	1902*** (70.1)	465.8*** (139)
Sigma	--	8913*** (37.2)
Observations	73,296	73,296

Standard errors are in parentheses; *** p<0.01, ** p<0.05, * p<0.1); sigma represents the estimated standard error of the residual for the Tobit model.

- Compare the size of the coefficients in the OLS and Tobit models. Are differences between the coefficients consistent with what you would expect? Why or why not? Be sure to justify your answer by explaining why the coefficients would differ between the models.
- Use the Tobit estimates to predict the following for a 50 year old married male with \$100,000 of annual income.
 - the probability of a nonzero credit card balance
 - the probability of a credit card balance greater than \$10,000.
 - the expected credit card balance
 - the expected credit card balance, given that he has a non-zero credit card balance
 - the effect of an additional \$1,000 of income on the expected credit card balance.
- The above sample included data from 8 different years between 1983 and 2004. Suppose that you wanted to test the hypothesis that all the Tobit coefficients (not just the intercepts) were identical in all 8 years. Explain how you could test this hypothesis with details on how you would construct the test statistic, its distribution including degrees of freedom, and under what conditions the test statistic would lead you to reject the null hypothesis.

2. (30 points) A recent study¹ examines the effect of birth weight on a number of outcomes measured shortly after birth (e.g. 1 year mortality) and during adulthood. Because of concerns about the ability to adequately control for family background characteristics, the authors choose a “fixed effects” approach that relies upon the use of twins. A description of their model is below:

II. CONCEPTUAL FRAMEWORK

Following ACL, let

$$(1) \quad y_{ijk} = \alpha + \beta bw_{ijk} + x_{jk}'\gamma + f_{jk} + \varepsilon_{ijk}$$

where subscript i refers to the child, j refers to the mother, and k refers to birth. y_{ijk} is then the outcome of child i born to mother j in birth k , bw_{ijk} is birth weight, x_{jk} is a vector of mother- and birth-specific variables (for example, mother’s education, the year of birth), f_{jk} refers to unobservables that are mother- and birth-specific (for example, the quality of prenatal care, genetic factors), and ε_{ijk} is an idiosyncratic error term assumed independent of all other terms in the equation.

The authors estimate the model using data on a sample of twins. They estimate an OLS model and a model with fixed effects for each mother. (For example, if the sample has 500 pairs of twins, there would be 1000 observations and 500 fixed effects.)

Coefficients on the birth weight variable [actually, it’s included as $\ln(\text{birthweight})$] are presented in the table below with and without fixed effects included. The footnotes to the table describe the other controls that were included in the models.

¹ Black, S.; Devereux, P; and Salvanes, K. “From the Cradle to the Labor Market? The Effect of Birth Weight on Adult Outcomes.” *Quarterly Journal of Economics*, February 2007: 409-439.

TABLE IV
 REGRESSION RESULTS CONSTANT SAMPLE: COEFFICIENT ON LN (BIRTH WEIGHT)

Dependent variable	Male same-sex twins 1978-1986		Female same-sex twins 1967-1977	
	OLS	FE	OLS	FE
1-year mortality	-299.24** (31.74)	-33.20 (22.92)	-390.66** (27.62)	5.84 (23.57)
<i>N</i>	2760		3804	
Five-minute APGAR score	.90** (.15)	.37* (.21)	—	—
Height (males only)	7.55** (.89)	8.56** (.85)	—	—
BMI (males only)	.20 (.40)	2.24** (.54)	—	—
Underweight	-.05 (.03)	-.05 (.05)	—	—
Overweight	.03 (.04)	.17** (.07)	—	—
IQ (males only)	.29 (.20)	1.05** (.28)	—	—
<i>N</i>	1894			
High school completion	—	—	.03 (.04)	.11* (.06)
<i>N</i>			3466	
Full-time work	—	—	.19** (.03)	-.03 (.08)
<i>N</i>			3574	
ln(earnings) FT	—	—	.17** (.07)	.14 (.10)
<i>N</i>			1732	
ln(birth weight of first child)			.19** (.03)	.13** (.06)
<i>N</i>			1722	

Standard errors are given in parentheses. The control variables we use in the OLS estimation are year- and month-of-birth dummies, indicators for mother's education (one for each year), indicators for birth order, indicators for mother's year of birth, and an indicator for the sex of the child. Twin fixed effects regressions include indicators for sex and birth order of the twin (either first born or second born twin). Both cross-sectional and fixed effects regressions for height, BMI, and IQ also include indicator variables for the year the boy was tested by the military. High school completion indicates whether or not the individual has completed at least twelve years of schooling and is restricted to those twenty-one and older. The IQ measure is generated from a composite score from three speeded IQ tests—arithmetic, word similarities, and figures—and is reported in stanine (Standard Nine) units. Earnings are measured as total pension-qualifying earnings reported in the tax registry. These are not topcoded and include labor earnings, taxable sick benefits, unemployment benefits, parental leave payments, and pensions. We restrict attention to individuals aged at least twenty-five. Working full-time indicates whether individuals are full-time, full-year workers. To identify this group, we use the fact that our dataset identifies individuals who are employed and working full time (30+ hours per week) at one particular point in the year (in the second quarter in the years 1986-1995, and in the fourth quarter thereafter). We label these individuals as full-time workers. For ln(birth weight) of child, the sample consists of women born between 1967 and 1977 whose first births occurred by 2004. If the first birth is a twin birth, the woman is discarded from the sample.

** Denotes statistically significant at the 5 percent level.

* Denotes statistically significant at the 10 percent level.

- a. Suppose that a male baby's birthweight increases from 500 to 600 grams. What is the estimated effect on 1 year mortality using the FE model? (Note that birthweight is measured in $\log(\text{grams})$ and that the mortality rate is in deaths per 1,000.)
- b. At the bottom of the above table is a list of controls included in the OLS and FE models. Notice that information such as whether the mother smoked, drank alcohol, or used crack cocaine during pregnancy is not included in either regression. How could the absence of such controls help explain the difference between the OLS and FE estimates on 1 year mortality rates? Describe the underlying econometric rationale for the difference between the estimate as well as any underlying assumptions that you are making about the effects of smoking/drinking on outcomes.
- c. Notice in the footnote that the OLS model includes different controls than the FE model. For example, mother's education is included as a control in the OLS model, but not in the FE model. Why would the authors include the variable in the OLS model, but exclude it from the FE model? Provide the econometric rationale for your answer.
- d. Suppose that someone recommends that the authors estimate a random effects (RE) model instead of the fixed model.
 - i. What advantage does a RE model have over a FE model?
 - ii. What disadvantage does a RE model have over a FE model?
 - iii. What diagnostics could the authors use to determine whether the FE or RE model is "preferred"? Provide details on the construction of any relevant test statistics.

3. (25 points) When I teach the principles of Economics, I use “supplemental instructors” (SI) who hold sessions twice weekly to assist students. Students are not required to attend SI sessions. At each session, the SI asks students to sign in and attendance records are tracked. The Office of Learning Assistance (OLA) uses data on attendance to examine the effect of SI on student outcomes. While OLA uses only simple statistics, it might be summarized by the simple regression below:

$$G_i = a_0 + a_1SI_i + e_i$$

Where G_i represents the course grade received by person i (e.g. $A=4.0$) and SI is a dummy variable indicating whether the person ever attended an SI session.

- a. Given that the SI program is voluntary, why should one be suspect of any estimated value of a_1 as an indicator of the program’s success? Do you think that a_1 will over- or under-state the true effect of SI sessions on student performance? Why?
- b. Suppose that you have data on the student’s ACT score and GPA at Miami in all other courses. Do you expect the addition of such controls would increase or decrease the estimated success of the SI program? Explain.
- c. Suppose that there are “unobservables” (e.g. work ethic) that influence whether a student attends SI sessions and you wish to use a treatment effects model to correct for any potential bias caused by the lack of control for such variables.
 - i. Explain how you would estimate a treatment effects model in this setting and be careful to describe any additional variables that would be necessary to properly estimate the model. (i.e. what differentiates variables used in the two stages of the estimation process?)
 - ii. Explain how you could use the estimates of the treatment effects model to determine whether there is “positive” or “negative” sample selection and whether this would cause the OLS estimates that control for observables (GPA, etc.) to be an over- or under-estimate of the true effect of SI on the grade.
 - iii. Explain how you could use your treatment effects estimate to estimate the difference in grades earned by two students who are identical in terms of observed characteristics, but differ in terms of whether they choose to attend SI sessions.

4. (15 points) I used data from the Survey of Consumer Finances from 1983-2004 to examine the distribution of mortgage rates paid over time. The results of quantile regressions (10th, 50th, and 90th) are presented below. Both age and income are measured as deviations from the sample means. The mortgage rate is measured in basis points (e.g. 750 implies a mortgage rate of 7.50 percent).

COEFFICIENT	10 th	50 th	90 th
Age (deviation from mean)	-0.227** (0.091)	-0.837*** (0.058)	-0.915*** (0.15)
Income (\$1000s) – deviation from mean	-0.0166***	-0.0148***	-0.0184***
Black	(1.20) 18.13*** (4.02)	(1.30) 46.37*** (2.61)	(4.50) 202.2*** (5.98)
Year dummies (1983 omitted)			
1986	-1.680 (9.96)	-149.8*** (6.50)	-246.0*** (14.9)
1989	-47.61*** (7.94)	-215.3*** (5.17)	-396.0*** (11.8)
1992	-175.5*** (7.71)	-327.1*** (5.03)	-488.3*** (11.5)
1995	-248.3*** (7.52)	-450.8*** (4.91)	-628.3*** (11.2)
1998	-235.4*** (7.60)	-464.2*** (4.96)	-641.6*** (11.4)
2001	-261.3*** (7.57)	-496.8*** (4.94)	-684.0*** (11.3)
2004	-467.2*** (7.44)	-653.1*** (4.87)	-865.6*** (11.2)
Constant	897.1*** (7.27)	1200*** (4.75)	1528*** (10.9)
Observations	21589	21589	21589
R-squared	.	.	.

Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

- a. In 1983, what is the 90-10 gap (i.e. difference between the 90th and 10th percentile) in mortgage rates for a person with average income and age for blacks? whites?
- b. The “sub-prime” crisis that is currently wreaking havoc on financial markets in the U.S. was the result of increasingly lax lending standards whereby banks were giving mortgages to an increasingly risky population who had a high chance of default. Presumably, banks were willing to extend loans to this high risk population only if there was a commensurate increase in returns to offset the added risk. Based on the quantile regression results above, do you see any evidence of such behavior in the regression results? Explain.