

1. (45 points) Using data from the Health and Retirement Survey for households with at least one person over age 50, I used OLS to estimate a regression model of the determinants of a household's networth (i.e. Assets-liabilities). The control variables included the respondent's age and its square ( $age^2$ ), dummy variables indicating whether the person is married, black, or a smoker; and the number of years of formal education. The regression was estimated with and without the education variable included. Results are in specifications (1) and (2) in the table below.

	Specification (1)		Specification (2)		Specification (3)	
	Coefficient	t-statistic	Coefficient	t-statistic	Coefficient	t-statistic
Intercept	-1431353	-2.16	-1218684	-2.9	-1764956	-2.71
age	51508	2.31	46603	2.15	48595	2.22
age-squared	-385	-2.06	-310	-1.71	-334	-1.83
married	196219	13.22	184654	12.8	186239	12.78
black	-145119	-8.26	-99804	-5.8	-115597	-6.67
Currently smoke? (1=yes, 0=no)	-112357	-7.56	-77059	-5.3	-86921	-5.93
Years of education	--		38563	20.11	27398	16.36
Sample size	6667		6667		6667	

- Based on specification (2), what is the predicted networth for a 60 year old white unmarried smoker with 12 years of education? Provide a brief explanation.
- Adding education to the regression causes the coefficient on the black dummy variable to drop (in absolute terms). Explain what would cause this to happen. **Explain why** these conditions would lead to the observed change in the coefficient on the black dummy.
- Suppose that the black dummy was dropped and replaced by a "non-black" dummy (i.e. a variable that equals one if the person is not black) in specification (2). Describe **any and all** coefficients that would change and, if possible, provide numerical values for any coefficients that change.
- I reestimated specification (2) after adding random noise to years of education. The results are in specification (3). Notice that the coefficient on years of education decreases while that on the black dummy increases (in absolute terms). Explain why you should have expected these two coefficients to change in the direction that they do.
- Based on regression specification (2), other things being the same, at what age is networth maximized? Explain briefly.
- I reestimated specification (2) using  $\log(\text{networth})$  in place of networth. The coefficient on years of education in the new specification is .19. Provide an interpretation of this coefficient.

- g. I estimated a network regression using only the subsample of whites. Coefficient estimates are in the table below along with the sample means for network and all the explanatory variables by race.
- How much of the black-white difference in network can be “explained” by differences in the explanatory variables? Provide a brief description of how you derived your answer.
  - How much of the black-white difference in network is “explained” by racial differences in education? Provide a brief description of how you derived your answer.

Variable	Regression estimates using white sample		Sample mean	
	coefficient	t-statistic	white	black
Network	--	--	349138	138526
age	43467	1.65	57	56
age-squared	-283	-1.29	3292	3247
married	191083	10.82	0.69	0.41
Years of education	42472	18.42	12.1	10.9
Currently smoke? (1=yes, 0=no)	-93927	-5.28	0.28	0.34
Intercept	-1824945	-2.33		

2. (40 points) A recent paper by Martin Schellhorn<sup>1</sup> examines the effect of health insurance deductibles on utilization. The basic regression analysis can be described as follows:

$$visits_i = \beta_0 + deductible_i\beta_1 + X_i\beta_2 + e_i$$

Where  $visits_i$  represents the number of doctor visits person  $i$ ,  $deductible_i$  is the dollar value of the insurance deductible,  $X_i$  is a vector of demographic characteristics that are expected to affect a person’s number of visits. Since a higher deductible means that a person will have to pay a larger share of the costs of visiting the doctor, the expected effect of deductible is negative.

- Since individuals are allowed to choose their health insurance plan and the level of the deductible, Schellhorn argues that the deductible variable is endogenous in the above regression equation. Which direction do you anticipate the coefficient on deductibles would be biased if the regression was estimated without a correction for the endogeneity? Explain the basis for your conclusion.
- How does two-stage least squares eliminate the bias that emerges in an OLS model? Be precise about the source of the endogeneity bias and how 2SLS eliminates the problem.
- If you are going to use two-stage least squares (2SLS), what additional data is required beyond that used in the OLS estimation of the regression above? Explain the necessary properties of this data.

---

<sup>1</sup> The Effect of Variable Health Insurance Deductibles on the Demand for Physician Visits, *Health Economics*, July 2001.

- d. Suppose that the 2SLS estimate of the coefficient on the deductible variable  $\beta_1$  is statistically insignificant (i.e. the null that the coefficient equals zero cannot be rejected at a reasonable level of confidence) but is quantitatively large. You are convinced that deductibles truly matter, but that the 2SLS process is responsible for the lack of significance. What suggestions can you make for "improving" the estimation process? Explain **how and why** these changes would "improve" the process.

3. (15 points) Suppose you have the following model describing the rate of return earned on a pension fund:

$$r_i = \beta_0 + \beta_1 * cstock_i + \beta_2 * otherstock_i + e_i$$

where  $r$  is the rate of return,  $cstock$  is the percentage of the fund's assets invested in the pension sponsor's own company stock, and  $otherstock$  is the percentage of the fund's assets invested in other stocks.

- Given that a pension can invest in company stock, other stock, or bonds, why can't the regression include the percent of assets in bonds as an independent variable? Explain.
- Suppose that you believe that the variance of the residual varies with the percentage of assets in the three categories. If this hypothesis is true, what (if anything) is "wrong" with using OLS to estimate the above regression equation? How would GLS improve the estimates. Explain.
- How could you test the hypothesis proposed in (b)? Explain.