June FINAL EXAM: ECONOMETRICS

Answer each question in different folders in two hours

1. The regressions for this problem are based on a sample of 27,326 observations, a survey of health care system usage taken in Germany over 7 years in the 1990s. The four regressions below regress a self-assesed general health measure (SAH) on a series of covariates based on this model

$$SAH = \beta_0 + \beta_1 Age + \beta_2 Age^2 + \beta_3 Married + \beta_4 Female + \beta_5 Hhkids + u$$

Age is measured in years. $Married$ and $Hhkids$ are dummy variables for marital status and whether there are kids in the household, and $Female = 1$ for women, 0 for men. In the first regression, the dummy variable $Female$ is included; in the second, it is omitted. The third regression is the same as the second, for women only; the fourth is the same as the second, but for men only. It is assumed that all the classical assumptions of the regression model are satisfied, and robust to heteroskedasticity standard errors are provided.

   (a) How would you test the hypothesis that all coefficients in the first model except the constant term are equal to zero? Carry out the test and explain under which assumptions your method is valid.

   (b) The coefficient on $Female$ in the first regression is a measure of the average difference in health between men and women with everything else held constant. The underlying null hypothesis is that the health determination mechanism is the same for men and women. The alternative hypothesis is that the health determinations are the same, except there is a constant difference between men and women. Carry out a test of the null hypothesis against the alternative in the context of the first regression. Now, use the results from both the first and the second regressions to carry out the same test. Which method would you prefer?

   (c) Provide the marginal effect of an additional year of age on SAH for someone who has 30 years of age. Using the results given for the first regression, compute a confidence interval for this value ($\widehat{Cov}\left(\hat{\beta}_1, \hat{\beta}_2\right) = -2.8 \times 10^{-7}$).

   (d) The quadratic specification of the model implies (given the results) that the relationship between SAH and age is hill shaped. Find the value of age for which SAH peaks. Explain how would you test the hypothesis that this value is $Age^* = 20$ against the alternative that it is greater than 20?

   (e) The second, third and fourth regressions report the least squares regression results for the model without the $Female$ dummy variable for the pooled sample, the subsample of women and the subsample of men.

   Write the two last models in a single equation and provide the estimated coefficients of all parameters and the SSR when estimating this model by OLS. Using these results, test the hypothesis that the same model applies to both men and women against the alternative hypothesis that the models are different.

First Regression, Pooled, OLS
Dependent variable: SAH
Residuals Sum of squares = 762.3413, R-squared = .1085580

	Coeficiente	Desv. Típica	Estadístico t	Valor p
Constant	-0,09502233	0,02521330	-3,769	0,0002
AGE	0,04424800	0,00391106	11,314	0,0
AGE2	-0,00085563	0,00014667	-5,834	0,0
MARRIED	0,08509255	0,00247111	34,435	0,0
FEMALE	0,00618235	0,00207600	2,978	0,0029
HHKIDS	-0,01748786	0,00214964	-8,135	0,0

Second Regression, Pooled, OLS
Dependent variable: SAH
Residuals Sum of squares = 762.5888, R-squared = .1082687

	Coeficiente	Desv. Típica	Estadístico t	Valor p
Constant	-0,07980746	0,02469379	-3,232	0,0012
AGE	0,04251933	0,00386830	10,992	0,0000
AGE2	-0,00079952	0,00014547	-5,496	0,0000
MARRIED	0,08486639	0,00247030	34,355	0,0000
HHKIDS	-0,01750636	0,00214994	-8,143	0,0000

Third Regression, Female Only, OLS
Dependent variable: SAH
Residuals Sum of squares = 372.6560, R-squared = .1225431

	Coeficiente	Desv. Típica	Estadístico t	Valor p
Constant	-0,20116524	0,03454318	-5,824	0,0000
AGE	0,05933135	0,00557856	10,636	0,0000
AGE2	-0,00147495	0,00021664	-6,808	0,0000
MARRIED	0,11519095	0,00352421	32,686	0,0000
HHKIDS	-0,01321821	0,00310482	-4,257	0,0000

Fourth Regression, Male Only, OLS
Dependent variable: SAH
Residuals Sum of squares = 384.3125, R-squared = .1042340

	Coeficiente	Desv. Típica	Estadístico t	Valor p
Constant	0,01085465	0,03786548	0,287	0,7744
AGE	0,03082032	0,00578582	5,327	0,0000
AGE2	-0,00033408	0,00021216	-1,575	0,1153
MARRIED	0,05491566	0,00346937	15,829	0,0000
HHKIDS	-0,01782500	0,00298198	-5,978	0,0000

2. Consider the following linear regression to explain the return to education in wages for US female workers,

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 smsa + \beta_4 south + u \quad (1)$$

where *wage* is the hourly wage in cents, *educ* are the years of schooling, *exper* is calculated as $age - educ - 6$, *smsa* and *south* are binary variables equal to one if the individual lives in a metropolitan area or in the south, respectively. It is suspected that *educ* is endogenous in this equation, while *exper* is regarded as exogenous. Three potential instrumental variables are available: *motheduc* and $fatheduc$, with information on mother's and father's schooling, respectively, and *nearc*4, which is a binary variable indicating nearness to a 4-year college. The information on the estimation of different models is provided at the end of the question with robust to heteroskedasticity estimates of the standard deviations.

(a) Researcher A estimates model (1) by 2SLS using the three instruments. Evaluate the relevance and exogeneity of the instruments.

(b) Researcher B estimates model (1) by 2SLS using only one instrument, *motheduc*. Evaluate the relevance and exogeneity of the instrument. Explain which strategy is better and if both are valid.

(c) Researcher C believes that given that there are plenty of instruments it is not necessary to control for *smsa* and *south* in the 2SLS fitting and estimate this model instead,

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + v. \quad (2)$$

Explain how you would test for the relevance for *educ* of the three instruments available to estimate (2) and which conclusions you can reach given the available information.

(d) Considering the information provided on the 2SLS estimation of (2), what can you say about the validity of the estimation method in (c)? Explain carefully the tests performed and provide an intuition for your results.

(e) Explain how your 2SLS estimation strategy would change if you consider that *exper* is also endogenous in (2). Why would you suspect that *exper* could be endogenous in (2) even if it is not in (1)?

Model 1: 2SLS, using observations 2–3006 ($n = 2220$)
Dependent Variable: l_wage
With instruments: educ. Instruments: motheduc fatheduc nearc4

	Coeficiente	Desv. Típica	Estadístico t	Valor p
const	4,05743	0,201782	20,1080	0,0000
educ	0,122659	0,0119272	10,2840	0,0000
exper	0,0611716	0,00523885	11,6765	0,0000
smsa	0,136248	0,0199858	6,8172	0,0000
south	−0,133170	0,0184341	−7,2241	0,0000

Model 2: OLS, using observations 2–3006 ($n = 2220$)
Dependent Variable: uhat_ M1 (= residuals of the 2SLS fit of Model 1)

	Coeficiente	Desv. Típica	Estadístico t	Valor p
const	−0,00831427	0,0469708	−0,1770	0,8595
exper	−0,000101583	0,00223589	−0,0454	0,9638
smsa	0,00220166	0,0197041	0,1117	0,9110
south	−0,000374686	0,0181107	−0,0207	0,9835
fatheduc	−0,00318214	0,00313224	−1,0159	0,3098
motheduc	0,00388169	0,00364105	1,0661	0,2865
nearc4	−0,00204680	0,0189841	−0,1078	0,9142

Robust F estad. of joint significance (fatheduc, motheduc, nearc4): 0,461289

Model 3: OLS, using observations 2–3006 ($n = 2220$)
Dependent Variable: educ

	Coeficiente	Desv. Típica	Estadístico t	Valor p
const	13,4685	0,227866	59,1069	0,0000
exper	−0,331590	0,0109005	−30,4196	0,0000
smsa	0,262093	0,0960391	2,7290	0,0064
south	−0,153658	0,0857999	−1,7909	0,0734
fatheduc	0,118582	0,0143585	8,2587	0,0000
motheduc	0,135504	0,0169707	7,9845	0,0000
nearc4	0,207622	0,0927493	2,2385	0,0253

Robust F estad. of joint significance (fatheduc, motheduc, nearc4): 104,402

Model 4: 2SLS, using observations 2–3006 ($n = 2657$)
Dependent Variable: l_wage
With instruments: educ. Instruments: motheduc

	Coeficiente	Desv. Típica	Estadístico t	Valor p
const	3,97667	0,213271	18,6461	0,0000
educ	0,128692	0,0126092	10,2062	0,0000
exper	0,0615520	0,00545155	11,2907	0,0000
smsa	0,131164	0,0182115	7,2023	0,0000
south	−0,148726	0,0180281	−8,2497	0,0000

Model 5: OLS, using observations 2–3006 ($n = 2657$)
Dependent Variable: educ

	Coeficiente	Desv. Típica	Estadístico t	Valor p
const	14,2652	0,195353	73,0229	0,0000
motheduc	0,206877	0,0125450	16,4907	0,0000
exper	−0,359005	0,00970304	−36,9993	0,0000
smsa	0,353921	0,0842524	4,2007	0,0000
south	−0,362810	0,0783040	−4,6334	0,0000

Model 6: 2SLS, using observations 2–3006 ($n = 2220$)
Dependent Variable: l_wage
With instruments: educ. Instruments: motheduc fatheduc nearc4

	Coeficiente	Desv. Típica	Estadístico t	Valor p
const	3,52647	0,200051	17,6278	0,0000
educ	0,158046	0,0116808	13,5304	0,0000
exper	0,0729822	0,00531226	13,7385	0,0000

Model 7: OLS, using observations 2–3006 ($n = 2220$)
Dependent Variable: uhat_ M6 (=residuals of the 2SLS fit of Model 6)

	Coeficiente	Desv. Típica	Estadístico t	Valor p
const	−0,0132662	0,0454804	−0,2917	0,7706
exper	−0,000454856	0,00236147	−0,1926	0,8473
fatheduc	−0,00358098	0,00328008	−1,0917	0,2751
motheduc	0,00135764	0,00387611	0,3503	0,7262
nearc4	0,0562306	0,0192320	2,9238	0,0035

Robust F estad. of joint significance (fatheduc motheduc nearc4): 3,16551

SOME CRITICAL VALUES: $Z_{0.90} = 1.282$, $Z_{0.95} = 1.645$, $Z_{0.975} = 1.96$, $\chi^2_{1,95} = 3.84$, $\chi^2_{1,975} = 5.02$, $\chi^2_{2,95} = 5.99$, $\chi^2_{2,975} = 7.378$, $\chi^2_{3,95} = 7.81$, $\chi^2_{3,975} = 9.3484$, $\chi^2_{4,95} = 9.49$, $\chi^2_{4,975} = 11.1433$, $\chi^2_{5,95} = 11.07$, $\chi^2_{5,975} = 12.8325$, where $\mathbb{P}(Z \leq Z_\alpha) = \alpha$ y $\mathbb{P}\left(\chi^2_m \leq \chi^2_{m,\alpha}\right) = \alpha$, Z is distributed as a normal with mean zero y variance one, and χ^2_m as a chi-square with m degrees of freedom.