Work the following problem, based on Hill, Griffiths, and Lim's Problem 10.1, exploring the use of instrumental variables in a regression.

Using state-level data, a research wants to examine the relationship between median rent paid (*RENT*) and median house price (*MDHOUSE* in $1,000). The percentage of the state population living in an urban area (*PCTURBAN*) is used as a control, so our model is

$$RENT_i = \beta_0 + \beta_1 MDHOUSE_i + \beta_2 PCTURBAN_i + u_i,$$

where β_1 is the primary coefficient of interest.

The following table represents a sequence of regressions performed to explore the model:

Equation # Dep. Var. Est. method	(1) RENT OLS	(2) MDHOUSE OLS	(3) MDHOUSE OLS	(4) RENT OLS	(5) RENT 2SLS	(6) UHAT OLS
Constant	125.9 (14.19)	−18.67 (12.00)	7.225 (8.936)	120.7 (12.43)	120.7 (15.71)	−62.85 (26.95)
MDHOUSE	1.521 (0.228)			2.240 (0.268)	2.240 (0.305)	
PCTURBAN	0.525 (0.249)	0.182 (0.115)	0.616 (0.131)	0.0815 (0.244)	0.0815 (0.305)	−0.283 (0.258)
FAMINC		2.731 (0.682)				4.448 (1.532)
REG2		−5.095 (4.122)				−6.768 (9.262)
REG3		−1.778 (4.073)				4.847 (9.151)
REG4		13.41 (4.048)				−18.77 (9.096)
VHAT				−1.589 (0.398)		
n	50	50	50	50	50	50
R^2	0.669	0.691	0.317	0.754	0.599	0.226
SSR	20259.6	3767.6	8322.2	15054.0	24565.7	19019.9

Standard errors are in parentheses below coefficient estimates.

1. What signs do you expect for β_1 and β_2? Why?

2. Column (1) presents the OLS estimates of the equation. Why might we be concerned about endogeneity of *MDHOUSE*? What effect would this have on the OLS estimator for β_1?

3. To deal with this possible endogeneity, there are four instruments available: median family income (*FAMINC*, in $1,000), and the region of the country (*REG2, REG3, REG4*). The OLS

regressions in columns (2) and (3) can be used to construct an F statistic that allows you to test whether the instruments are strong or weak. (Remember that we don't use the standard critical value for this test.) Are the instruments strong?

4. The OLS residuals from the reduced-form regression in column (2) are called *VHAT* and are added to the basic OLS model in column (4) to perform a simple Hausman test. What do we conclude from column (4) about the validity of the OLS regression in column (1)?

5. Column (5) presents 2SLS estimates of the model. Compare the results to the OLS results in column (1). Are there important differences? The coefficient estimates (but not their standard errors) are identical between columns (4) and (5). Why?

6. Column (6) regresses the residuals *UHAT* from column (5) on the exogenous variables of the model. How do we use this regression to test the overidentifying restrictions on the model? What is the result and what does it mean?