

## ECONOMETRICS COMPREHENSIVE EXAM – STUDY GUIDE FOR PART I

This portion of the exam will consist of six multi-part questions that will be similar in format to the questions from the midterm and final exams, but shorter:

- 1) Definitions
- 2) Short answer
- 3) Math (multi-part question involving computations and/or proofs)
- 4) Designing an empirical strategy
- 5) Stata code
- 6) Interpreting results

Below are practice problems for each of these categories, adapted from the homework assignments, review sheets, and exams. This guide has been designed so that if you know everything on it you should do very well on the exam. If you feel you need more practice, working through the other homework and exam questions might be useful.

### 1. Definitions (when appropriate, either a verbal or mathematical definition is fine)

Random variable

Parameter

Statistic

Estimator

Estimate

Unbiased

Efficiency (A concise definition would be “a relative measure of the sampling variance of an estimator; a lower variance indicates higher efficiency.”)

Consistency

Central limit theorem

Econometrics

### 2. Short answer

- Explain the mean squared error criteria and the minimum variance unbiased estimator criteria for choosing an estimator. Which of the two is more widely used, and (generally speaking) why is this the case?
- Briefly describe the difference between cross-sectional, time-series, pooled cross-sectional, and panel data.
- Distinguish between statistical and economic significance, and explain briefly why both are necessary in order to have an important result.
- State (either in words or mathematical notation or both) the assumptions needed to show (in matrix form) that the OLS estimator is unbiased.
- State the assumptions needed to prove the Gauss-Markov Theorem, and then state the Gauss-Markov Theorem. (Hint: the assumptions are the same ones needed to show that  $Var(\hat{\beta}|X) = \sigma^2(X'X)^{-1}$ .)

**2. Short answer (continued)**

- State (either in words or mathematical notation or both) the assumptions needed to show that  $\hat{\beta}|\mathbf{X}$  is distributed as multivariate normal with mean  $\beta$  and variance-covariance matrix  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ .
- If I describe an example of one of the three situations discussed in class where we should use a set of dummy variables, be able to design the appropriate set of dummy variables, taking into consideration that one needs to be excluded to prevent perfect collinearity.
- If I give you a scenario (i.e. we include as regressors both cigarettes smoked per day and cigarettes smoked per week=7\*cigarettes smoked per day), be able to tell if there is perfect collinearity or not, and also explain why.
- Why does over controlling bias our coefficient estimator (assuming we are estimating the total effect instead of the direct effect) but multicollinearity does not?
- What are the consequences of heteroskedasticity on the OLS estimator's unbiasedness, variance, hypothesis testing, fit statistics, and status as the best linear unbiased estimator?
- Discuss the two general options we have when faced with heteroskedasticity and give one benefit to each approach. Why is it generally acceptable with large samples to simply correct the OLS standard errors?
- Why is instrumental variables estimation considered a "high-risk, high-reward" technique? (For the "high-risk" part, you should include a discussion of how 1) it is hard to find an instrument that is both strong and exogenous, and 2) we can't really test whether the instrument is truly exogenous, so the theoretical argument that it is exogenous must be very solid. For the "high-reward" part, your answer should involve stating of the sources of endogeneity that valid IV estimation corrects for as well as any it does not correct for, plus the comment that no other econometric technique corrects for as many of these sources.)

**3. Math**

1. Suppose that X is a random variable with a pdf given by  $f(x) = \frac{1}{9}x^2$  if  $0 < x < 3$ , 0 otherwise.
  - a) Determine an equation for the cumulative distribution function  $F(x)$  and verify that the total probability is equal to 1.
  - b) Find the expected value of X.
  - c) Find the variance of X.
2. Let X be a random variable representing the outcome of rolling a six-sided die where the side with five dots had been replaced with six dots.
  - a) What is the probability density function  $f(x)$ ? Verify that the sum of the probabilities is 1.
  - b) What is the cumulative distribution function  $F(x)$ ?
  - c) Graph the probability distribution function and cumulative distribution function.
  - d) Find the expected value of X.
  - e) Find the variance of X.

3. Suppose that we have the following data for x and y.

#	x	y
1	10	17
2	1	7
3	7	6

- Calculate the sample means of x and y.
- Calculate the sample variances and standard deviations of x and y.
- Calculate the standard error of the means of x and y.
- Calculate the 95% confidence interval for the mean of x.
- Calculate the sample covariance of x and y.
- Calculate the sample correlation between x and y.
- Test whether the mean of x is different from 2.
- Test whether the mean of x is different from the mean of y.

4. Suppose we are studying the relationship between skipping class and college GPA, controlling for ACT score and high school GPA. Defining “skipped” as the number of classes skipped per week, we assume the following population relationship:

$$colGPA = \beta_0 + \beta_1 hsGPA + \beta_2 ACT + \beta_3 skipped + u$$

and we obtain the following estimated equation (standard errors in parentheses):

$$colGPA = 1.39 + .412hsGPA + .015ACT - .083skipped$$

(.33)    (.094)            (.011)            (.026)

where  $n=141$  and  $R^2=.234$ .

- Can we reject  $H_0: \beta_3 = 0$  at the 1% level in favor of  $H_1: \beta_3 \neq 0$ ?
- Can we reject  $H_0: \beta_3 = -1$  at the 5% level in favor of  $H_1: \beta_3 > -1$ ?
- Find the 95% confidence interval for  $\beta_3$ .
- Can we reject  $H_0: \beta_1 = \beta_2$  at the 1% level in favor of  $H_1: \beta_1 > \beta_2$ ?
- Conduct an F-test to test the null hypothesis that hsGPA and ACT are jointly equal to zero. Use a 5% significance level.

5. Suppose we have data on two variables for three individuals. Our data is:

Individual	x	Y
1	4	2
2	2	5
3	1	10

We estimate the following population model using ordinary least squares

$$y = \beta_0 + \beta_1 x + u$$

and obtain  $\hat{\beta}_0 = 11.5$  and  $\hat{\beta}_1 = -2.5$ .

- Find the predicted values of y for each of our three individuals.
- Calculate the residuals for each of our three individuals.
- Calculate the SSE, SSR, SST,  $R^2$ , and adjusted  $R^2$ .
- Calculate the standard error of  $\hat{\beta}_1$ .
- Is x statistically significant at the 5% level?

6. Say we're interested in studying the effect of additional police on crime. We assume police and crime are jointly determined through the following simultaneous equations model:

$$crime = \alpha_1 police + \beta_{10} + u_1$$

$$police = \alpha_2 crime + \beta_{20} + \beta_{21} taxrevenue + u_2$$

where all our variables are per capita and city-level, and we assume tax revenue to be exogenous.

- a) Derive the reduced-form model for police, while showing that the OLS estimator for  $\alpha_1$  is biased if  $\alpha_2 \neq 0$  (assume  $\alpha_1 \alpha_2 \neq 1$ ). What is the total effect of tax revenue on police in terms of the parameters from the structural model? Why is it different from  $\beta_{21}$ ?
- b) Derive the reduced-form model for crime.
- c) Is the parameter  $\alpha_1$  identified? What about  $\alpha_2$ ? Explain.

#### 4. Designing an Empirical Strategy

For the remaining questions, suppose we have a Stata dataset with the following variables:

age	age in years
educ	highest grade completed
black	1 if race is black, 0 otherwise
other	1 if race is neither black nor white, 0 otherwise
married	1 if married, 0 otherwise
hwk	average hours worked per week
bmi	body mass index
female	1 if female, 0 otherwise
fam	family size
inc	household income in dollars
vighwk	hours per week of vigorous exercise
drmo	alcoholic drinks per month
cigday	cigarettes smoked per day
cigtax	state cigarette excise tax (dollars per pack)
charity	amount of charitable contributions
tax	total amount of income taxes paid

1. Suppose we are interested in estimating the effect of work hours on alcohol consumption. The theoretical effect is ambiguous: on one hand, working more leaves less time for leisure activities, so drinking may drop, but on the other hand, working more may increase stress levels, so drinking may rise.

a) In the simple regression model :  $drmo = \beta_0 + \beta_1hwk + u$ ,

do you think that  $\hat{\beta}_1$  is an unbiased estimator of the causal impact of work hours on drinks per month? Why/why not?

b) In the multiple regression model:

$$drmo = \beta_0 + \beta_1hwk + \beta_2age + \beta_3educ + \beta_4black + \beta_5other + \beta_6married + \beta_7female + \beta_8inc + u,$$

explain why  $\hat{\beta}_1$  is more likely to be an unbiased estimator of the causal impact of work hours on drinks per month than it was in part a. Nonetheless, give an example of how omitted variable bias could still be a problem.

c) For the multiple regression model in part b, provide an argument that income should be excluded from the model, even if it is statistically significant and also correlated with hwk.

2. Suppose we estimate the equation

$$charity_i = \beta_0 + \beta_1inc + \beta_2tax + u.$$

a) Do you suspect that inc and tax are perfectly collinear? If not, do you suspect that multicollinearity may be a problem in our estimation?

b) If multicollinearity is a problem, explain why simply omitting one of the independent variables is not an ideal solution if our aim is to estimate the marginal propensity to donate.

c) Explain why reverse causality may be a problem with our analysis.

3. Suppose we estimate the effect of income on body mass index using the model:

$$bmi = \beta_0 + \beta_1 inc + \beta_2 age + \beta_3 hsgrad + \beta_4 colgrad + \beta_5 black + \beta_6 other + \beta_7 married + \beta_8 hwk + \beta_9 female + \beta_{10} fam + u.$$

Argue that, even including all the controls, our OLS estimator of the causal effect of income on bmi could suffer from bias from omitted variables, measurement error in the independent variable, measurement error in the dependent variable, or reverse causality. Which of these would a valid IV estimator eliminate?

4. Suppose we are writing a paper entitled “Smoking: Bad for Your Health and Your Wallet?” in which we test the theory that smoking reduces your income. The idea is that smoking may decrease your productivity by hurting your health, lowering your earnings. Let’s begin by estimating a simple linear regression model with annual household income as the dependent variable and cigday as the only independent variable:

$$inc = \beta_0 + \beta_1 cigday + u.$$

- Do you think  $\hat{\beta}_1$  is likely to be an unbiased estimator for  $\beta_1$ ? To defend your answer, pick one variable from the list that you think “belongs” in the model, meaning that omitting it could lead to bias, and explain your choice.
- Pick one variable from the list that, if added, would likely result in an over controlling problem, and explain why this is the case.
- Let’s add age, educ, black, other, and female as controls and estimate  $inc = \beta_0 + \beta_1 cigday + \beta_2 age + \beta_3 educ + \beta_4 black + \beta_5 other + \beta_6 female + u$ . Give an example of how our estimator  $\hat{\beta}_1$  could still suffer from omitted variable bias, even after adding these controls.
- If we were to add the variables married, hwk, bmi, fam, vighwk, and drmo as controls in our reduced-form model, how would that change the interpretation of our coefficient on cigday?
- In the model given by equation (2), explain why cigday likely suffers from non-classical measurement error. Do we know in which direction this measurement error would bias our estimator  $\hat{\beta}_1$ ?
- Explain why reverse causality/simultaneity may also be a problem in our analysis. (**con’t.**)
- Would a properly-executed instrumental variables approach solve the problem of omitted variable bias? What about measurement error in the independent variable? What about reverse causality/simultaneity? (No explanations necessary. By properly-executed, I mean both of the key assumptions are satisfied.)
- Consider the following structural model:

$$inc = \beta_0 + \beta_1 cigday + \beta_2 age + \beta_3 educ + \beta_4 black + \beta_5 other + \beta_6 female + u$$

$$cigday = \beta_0 + \beta_1 cigtax + \beta_2 age + \beta_3 educ + \beta_4 black + \beta_5 other + \beta_6 female + u.$$

Suppose we estimate this model using two-stage least squares, using cigtax as an instrument for cigday. Formally state the two assumptions needed for cigtax to be a valid instrument for cigday in this context. (Use the actual variable names not labels like x, y, and z.) For both assumptions, discuss whether or not you think they are likely to hold.

1. For our study of the effect of work hours on drinking, write Stata code to perform the following operations.
  - a) Find the sample means of *hwk* and *drmo*.
  - b) Find the correlation between *hwk* and *drmo*, as well as whether or not it is statistically significant.
  - c) Estimate the simple linear regression model
 
$$drmo = \beta_0 + \beta_1 hwk + u$$
  - d) Estimate the multiple linear regression model
 
$$drmo = \beta_0 + \beta_1 hwk + \beta_2 age + \beta_3 educ + \beta_4 black + \beta_5 other + \beta_6 married + \beta_7 female + u$$
  - e) Conduct a t-test of  $H_0: \beta_4 = \beta_5$  against the two-sided alternative.
  - f) Conduct an F-test to see if the control variables “belong” in the model. Formally, test  $H_0: \beta_2 = 0, \beta_3 = 0, \beta_4 = 0, \beta_5 = 0, \beta_6 = 0, \beta_7 = 0, \text{ and } \beta_8 = 0$  against  $H_1: H_0 \text{ is not true}$ .
  - g) Estimate the simple model in part c again, but use a quadratic functional form for *hwk*. Before running the regression, create the necessary variable(s) to do this. Assuming the sample mean for *hwk* is 30, calculate the marginal effect of *hwk* on *drmo* at the mean.
  
2. For our study of the effect of smoking on income, write Stata code to perform the following operations.
  - a) Run the simple linear regression:  $inc = \beta_0 + \beta_1 cigday + u$ .
  - b) Run this regression but use a log-level functional form. Create the necessary variable(s). (Assume that no one in the sample has an income of 0.)
  - c) Returning to a level-level functional form, run the multiple linear regression:
 
$$inc = \beta_0 + \beta_1 cigday + \beta_2 age + \beta_3 educ + \beta_4 black + \beta_5 other + \beta_6 female + u$$
  - d) Execute the White test for heteroskedasticity.
  - e) Execute the Breusch-Pagan test for heteroskedasticity.
  - f) Run the multiple linear regression with heteroskedasticity-robust standard errors.
  - g) Store your estimates under the name “ols.”
  - h) Run the multiple regression in part c, but use a weighted least squares estimator with the heteroskedasticity function  $h(x) = \text{age}$ .
  - i) Create a set of dummy variables to divide the sample into three groups: people who smoke 0 cigarettes per day, people who smoke more than 0 but no more than 15 cigarettes per day, and people who smoke more than 15 cigarettes per day.
  - j) Run the regression from part f (OLS multiple regression with robust standard errors), but replace *cigday* with your new set of dummy variables. (Watch out for perfect collinearity.)
  - k) Ceteris paribus, find the predicted difference between the incomes of people who smoke more than 15 cigarettes per day and people who smoke more than 0 but no more than 15 cigarettes per day.
  - l) Create the interaction term *cigday\*female*.
  - m) Return to the use of a linear functional form for *cigday*. Run the regression from part f, but add the interaction term.
  - n) Ceteris paribus, find the predicted effect of *cigday* on income for women.

2. (continued)

- o) Estimate the structural model in (3) and (4) with two-stage least squares, using cigtax as an instrument for cigday. Use the option that reports the results from both stages.
- p) Store your estimates under the name “iv.”
- q) Conduct a Hausman test comparing your estimates from the regressions in parts o and f.

## 6. Interpreting Results

1. For our study of the effect of work hours on drinking, suppose that if we estimate the simple linear regression model we obtain the following Stata output:

Source	SS	df	MS			
Model	20239.281	1	20239.281	Number of obs =	7489	
Residual	5716797.56	7487	763.563185	F( 1, 7487) =	26.51	
Total	5737036.84	7488	766.164109	Prob > F =	0.0000	
				R-squared =	0.0035	
				Adj R-squared =	0.0034	
				Root MSE =	27.633	

  

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
drmo						
hwk	.0809038	.0157143	5.15	0.000	.0500995	.1117082
_cons	8.92635	.619701	14.40	0.000	7.711562	10.14114

Suppose also that if we estimate the multiple linear regression model (excluding income), we obtain the following Stata output:

Source	SS	df	MS			
Model	266836.037	7	38119.4338	Number of obs =	7488	
Residual	5465529.92	7480	730.685819	F( 7, 7480) =	52.17	
Total	5732365.96	7487	765.642575	Prob > F =	0.0000	
				R-squared =	0.0465	
				Adj R-squared =	0.0457	
				Root MSE =	27.031	

  

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
drmo						
hwk	.0302414	.0161589	1.87	0.061	-.0014345	.0619174
age	-.3381229	.1397713	-2.42	0.016	-.612114	-.0641318
educ	-.2849688	.1265659	-2.25	0.024	-.5330735	-.0368641
black	-3.643506	.7159052	-5.09	0.000	-5.046882	-2.240131
other	-3.100133	1.358236	-2.28	0.022	-5.762659	-.4376079
married	-3.096517	.6613095	-4.68	0.000	-4.392869	-1.800164
female	-10.73453	.6449336	-16.64	0.000	-11.99878	-9.470277
_cons	38.07319	6.498279	5.86	0.000	25.33474	50.81164

- a) In the simple regression, is hwk statistically significant at the 1% level, 5% level, 10% level, or none of these levels? What about in the multiple regression?
- b) Explain in words exactly what the coefficient estimate for hwk in the multiple regression model is telling us.
- c) Does it appear that our estimator for the effect of work hours on drinking is biased if we do not include any control variables?
- d) Finally, suppose that, if we estimate the log-log model, we obtain the following Stata output:

## 6. Interpreting Results (continued)

Source	SS	df	MS			
Model	1461.53184	8	182.691481	Number of obs = 7050		
Residual	14236.0945	7041	2.02188531	F( 8, 7041) = 90.36		
				Prob > F = 0.0000		
				R-squared = 0.0931		
				Adj R-squared = 0.0921		
				Root MSE = 1.4219		
Total	15697.6263	7049	2.22692954			

  

ldrmo	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lhwk	.0884399	.0124888	7.08	0.000	.0639581	.1129218
age	-.0261883	.0075734	-3.46	0.001	-.0410345	-.0113421
educ	.0127175	.0072843	1.75	0.081	-.001562	.0269969
black	-.3670578	.0393518	-9.33	0.000	-.4441993	-.2899164
other	-.2256616	.0736883	-3.06	0.002	-.3701129	-.0812103
married	-.2672116	.0383244	-6.97	0.000	-.342339	-.1920843
female	-.6268929	.0345838	-18.13	0.000	-.6946876	-.5590983
inc	2.49e-06	2.73e-07	9.15	0.000	1.96e-06	3.03e-06
_cons	2.457395	.3538117	6.95	0.000	1.763817	3.150972

What is the approximate “work hour elasticity” of drinks per month? Do you consider the effect of work hours on drinking to be economically significant? Why/why not?

2. Let’s examine the results from our study of the effect of smoking on income.

a) Suppose that, when we estimate the simple regression model with a level-level functional form, we get the following output:

Source	SS	df	MS			
Model	8.1092e+11	1	8.1092e+11	Number of obs = 6733		
Residual	3.6314e+13	6731	5.3950e+09	F( 1, 6731) = 150.31		
				Prob > F = 0.0000		
				R-squared = 0.0218		
				Adj R-squared = 0.0217		
				Root MSE = 73451		
Total	3.7125e+13	6732	5.5147e+09			

  

inc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
cigday	-1199.385	97.82884	-12.26	0.000	-1391.16	-1007.609
_cons	75946.12	1000.387	75.92	0.000	73985.05	77907.2

Interpret the coefficient estimate for cigday. Is cigday significant at the 1% level, 5% level, 10% level, or none of these?

b) When we estimate the simple regression model with a log-level functional form, we get:

Source	SS	df	MS			
Model	200.119238	1	200.119238	Number of obs = 6484		
Residual	6948.8456	6482	1.07202184	F( 1, 6482) = 186.67		
				Prob > F = 0.0000		
				R-squared = 0.0280		
				Adj R-squared = 0.0278		
				Root MSE = 1.0354		
Total	7148.96484	6483	1.10272479			

  

linc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
cigday	-.0193278	.0014146	-13.66	0.000	-.0221009	-.0165547
_cons	10.86325	.0143178	758.72	0.000	10.83518	10.89131

Interpret the coefficient estimate for cigday.



inc	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]
cigday	-697.4414	67.75755	-10.29	0.000	-830.2676 -564.6151
age	960.1898	387.8958	2.48	0.013	199.7911 1720.588
educ	10028.92	435.3543	23.04	0.000	9175.485 10882.35
black	-30974.62	1527.138	-20.28	0.000	-33968.29 -27980.94
other	-17197.19	2924.14	-5.88	0.000	-22929.43 -11464.95
female	-10368.85	1652.438	-6.27	0.000	-13608.16 -7129.552
_cons	-87284.74	17898.67	-4.88	0.000	-122371.8 -52197.67

Has correcting for heteroskedasticity affected our coefficient estimate for cigday? What about the standard error? Explain.

- f) We obtain the following output from the regression with the set of dummy variables for smoking. “Light smoker” is an indicator variable equal to 1 if the individual smokes more than 0 but no more than 15 cigarettes per day, while “heavy smoker” is an indicator variable equal to 1 if the individual smokes more than 15 cigarettes per day.

Linear regression

Number of obs = 6730  
 F( 7, 6722) = 148.75  
 Prob > F = 0.0000  
 R-squared = 0.1895  
 Root MSE = 66894

inc	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]
light smoker	-13650.84	1983.967	-6.88	0.000	-17540.04 -9761.633
heavy smoker	-15877.83	2005.954	-7.92	0.000	-19810.14 -11945.53
age	942.7419	387.7302	2.43	0.015	182.6678 1702.816
educ	9962.178	437.5534	22.77	0.000	9104.435 10819.92
black	-29769.53	1527.064	-19.49	0.000	-32763.06 -26776
other	-16441.16	2912.449	-5.65	0.000	-22150.48 -10731.83
female	-10020.09	1653.056	-6.06	0.000	-13260.6 -6779.576
_cons	-85301.38	17915.29	-4.76	0.000	-120421 -50181.73

Interpret the coefficient estimate for heavy smoker.

- g) Estimates for the model with the interaction term cigday\*female (named cigday\_female):

Linear regression

Number of obs = 6730  
 F( 7, 6722) = 151.95  
 Prob > F = 0.0000  
 R-squared = 0.1890  
 Root MSE = 66914

inc	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]
cigday	-668.9856	89.71486	-7.46	0.000	-844.8552 -493.1161
age	959.7959	387.9463	2.47	0.013	199.2982 1720.294
educ	10027.39	435.678	23.02	0.000	9173.328 10881.46
black	-30978.75	1527.572	-20.28	0.000	-33973.27 -27984.22
other	-17201.69	2924.603	-5.88	0.000	-22934.84 -11468.55
female	-10071.82	1940.197	-5.19	0.000	-13875.23 -6268.423
cigday_female	-65.65291	135.0304	-0.49	0.627	-330.3553 199.0494
_cons	-87388.73	17892.04	-4.88	0.000	-122462.8 -52314.67

Using a 5% significance level, can we conclude that the ceteris paribus effect of smoking on income is different for women and men? Explain.

## 6. Interpreting Results (continued)

Part I/Page 12

h) Below is the output from the first stage of our two-stage least squares estimation, along with output obtained if we exclude “cigtax” from the regression.

```

First-stage regressions
-----
First-stage regression of cigday:
OLS regression with robust standard errors
-----
Total (centered) SS      = 555734.9717
Total (uncentered) SS  = 694192
Residual SS            = 518796.6112
Number of obs          = 6609
F( 6, 6602)           = 81.19
Prob > F                = 0.0000
Centered R2            = 0.0665
Uncentered R2          = 0.2527
Root MSE               = 8.865
-----
cigday |      Coef.   Robust      t   P>|t|   [95% Conf. Interval]
        |           Std. Err.
-----+-----
    age |   .0962126   .0488445    1.97  0.049   .0004616   .1919636
    educ |  -.8617157   .0401174   -21.48  0.000  -.9403587  -.7830726
   black |  -1.78937   .2343978    -7.63  0.000  -2.248866  -1.329875
    other | -3.437686   .3782474   -9.09  0.000  -4.179173  -2.696199
 female |  -.6459796   .2190409    -2.95  0.003  -1.075371  -.2165886
   cigtax | -1.844452   .9154972    -2.01  0.044  -3.639122  -.049781
    _cons |  13.18974   2.271489    5.81  0.000   8.736889  17.64259
-----
Linear regression
Number of obs          = 6730
F( 5, 6724)           = 98.63
Prob > F                = 0.0000
R-squared              = 0.0657
Root MSE               = 8.8447
-----
cigday |      Coef.   Robust      t   P>|t|   [95% Conf. Interval]
        |           Std. Err.
-----+-----
    age |   .0860956   .0481192    1.79  0.074  -.008376   .1805672
    educ |  -.8581609   .0395348   -21.71  0.000  -.9356617  -.78066
   black |   -1.69     .2239404   -7.55  0.000  -2.128994  -1.251006
    other | -3.310528   .3733033   -8.87  0.000  -4.042321  -2.578736
 female |  -.6940665   .2163175    -3.21  0.001  -1.118117  -.2700158
    _cons |  13.18217   2.241167    5.88  0.000   8.788775  17.57557
-----

```

Conduct an F-test to determine if our analysis suffers from a “weak instrument” problem. What do you conclude?

i) Below is the output from the second stage of our two-stage least squares estimation.

```

IV (2SLS) regression with robust standard errors
-----
Total (centered) SS      = 3.63344e+13
Total (uncentered) SS  = 6.88699e+13
Residual SS            = 3.69295e+13
Number of obs          = 6609
F( 6, 6602)           = 144.74
Prob > F                = 0.0000
Centered R2            = -0.0164
Uncentered R2          = 0.4638
Root MSE               = 74751
-----

```

inc	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
cigday	-4476.463	3769.337	-1.19	0.235	-11864.23	2911.301
age	1365.19	549.3612	2.49	0.013	288.4622	2441.918
educ	6675.85	3290.757	2.03	0.042	226.0844	13125.62
black	-37799.14	6535.293	-5.78	0.000	-50608.08	-24990.2
other	-29274.51	13501.35	-2.17	0.030	-55736.66	-2812.355
female	-12513.27	3084.794	-4.06	0.000	-18559.36	-6467.187
_cons	-39791.51	52898.56	-0.75	0.452	-143470.8	63887.76

Instrumented: cigday  
 Included instruments: age educ black other female  
 Excluded instruments: cigtax

Comment on how the coefficient estimate, standard error, and level of significance for cigday are different than in the OLS regression reported in part e.

j) Explain why we cannot implement the overidentification test.

k) Our output from the Hausman test is:

	---- Coefficients ----			
	(b) iv	(B) ols	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
cigday	-4476.463	-697.4414	-3779.022	3768.728
age	1365.19	960.1898	405.0005	389.0174
educ	6675.85	10028.92	-3353.067	3261.832
black	-37799.14	-30974.62	-6824.524	6354.361
other	-29274.51	-17197.19	-12077.31	13180.89
female	-12513.27	-10368.85	-2144.419	2604.881

b = consistent under Ho and Ha; obtained from ivreg2  
 B = inconsistent under Ha, efficient under Ho; obtained from regress

Test: Ho: difference in coefficients not systematic

$$\begin{aligned} \text{chi2}(6) &= (b-B)'[(V_b-V_B)^{-1}](b-B) \\ &= 4.26 \\ \text{Prob}>\text{chi2} &= 0.6411 \end{aligned}$$

Assuming our instrumental variables estimator is consistent, can we conclude that our OLS estimator is inconsistent? Why/why not? Use a 5% significance level.