

ECON 836 Midterm Spring 2010

1. [4 points] Suppose you have a panel of countries as in Islam's growth model. Assume that the true model is $Y_{it} = X_{it}\beta + \theta_i + \varepsilon_{it}$, where $E[X_i' \varepsilon_{it}] = E[\theta_i \varepsilon_{it}] = 0$ and $E[(\theta_i)^2] = \sigma_\theta^2, E[(\varepsilon_{it})^2] = \sigma_\varepsilon^2$.

a. Suppose you run `regress Y X` in Stata. Derive the expectation of the coefficients. Are the estimated coefficients biased?

$$E[\hat{\beta}] = E[(X'X)^{-1} X'Y] = E[(X'X)^{-1} X'X\beta + (X'X)^{-1} X'X(\theta + \varepsilon)]$$

i.
$$= \beta + E[(X'X)^{-1} X'X(\theta + \varepsilon)] = \beta + (X'X)^{-1} X'E[X\theta] + (X'X)^{-1} X'E[X\varepsilon]$$

$$= \beta + (X'X)^{-1} X'E[X\theta]$$

ii. the expectations of the inner products of X and the disturbances are zero by the assumptions give above. But, the expectation of the inner products of X and the fixed effects may not be zero, and represent a bias term.

b. Suppose you run `regress Y X D` in Stata, where D is a set of dummy variables for countries. Now suppose you ran a regression with a left-hand side variable equal to the estimated coefficients on D , and right-hand side variables equal to the average X for each country. Suppose X was significant in this regression. How should you think about a) above?

i. the coefficients on D are estimates of the parameters θ . If X is significant in this regression, then X and θ are probably correlated. If they are correlated, then the bias expression above is not zero, and the estimate in a) is biased.

2. [4 points] Consider the following code and output from a log-wage regression using 2001 Census data on British Columbia residents.

```
#delimit;
generate insamp=pobp<11&agep<65&agep>24&cowp==1&hlosp~=. &
wagesp>0&prov==59;
generate logwage=log(wages);
generate alone=unitssp==1;
recode agep (25/29=1) (30/34=2) (35/39=3) (40/44=4) (45/49=5) (50/54=6) (55/59=7)
(60/64=8) (else=0), gen(agegp);
generate vismin=visminp<5;
generate aborig=abethncp<3;
replace vismin=0 if aborig==1;
generate white=(vismin==0)&(aborig==0);
xi: regress logwage i.agegp i.hlosp i.marsthp i.cmap alone unitssp i.olnp vismin aborig if
( insamp==1&sexp==2);
```

the Stata output is

i.agegp	__Iagegp_0-8	(naturally coded; __Iagegp_0 omitted)
i.hlosp	__Ihlosp_1-14	(naturally coded; __Ihlosp_1 omitted)
i.marsthp	__Imarsthp_1-5	(naturally coded; __Imarsthp_1 omitted)
i.cmap	__Icmap_933-935	(naturally coded; __Icmap_933 omitted)
i.olnp	__Iolnp_1-4	(naturally coded; __Iolnp_1 omitted)

Source		SS	df	MS	Number of obs =	8758
Model		1079.87668	32	33.7461464	F(32, 8725) =	39.91
Residual		7377.23906	8725	.845528832	Prob > F =	0.0000
Total		8457.11574	8757	.96575491	R-squared =	0.1277
					Adj R-squared =	0.1245
					Root MSE =	.91953

logwage		Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
__Iagegp_1		-.4752942	.0460058	-10.33	0.000	-.5654763 -.385112
__Iagegp_2		-.2172347	.0430687	-5.04	0.000	-.3016596 -.1328099

_Iagegp_3		-.1215169	.0429977	-2.83	0.005	-.2058026	-.0372312
_Iagegp_4		-.0357658	.0426757	-0.84	0.402	-.1194202	.0478887
_Iagegp_5		.0141983	.0433798	0.33	0.743	-.0708363	.099233
_Iagegp_6		.007959	.0440889	0.18	0.857	-.0784657	.0943837
_Iagegp_7		(dropped)					
_Iagegp_8		-.3125657	.055457	-5.64	0.000	-.4212746	-.2038568
_Ihlosp_2		.4623549	.2465691	1.88	0.061	-.0209787	.9456885
_Ihlosp_3		.4276838	.2295029	1.86	0.062	-.022196	.8775636
_Ihlosp_4		.6117929	.2295909	2.66	0.008	.1617405	1.061845
_Ihlosp_5		.6476612	.2331167	2.78	0.005	.1906974	1.104625
_Ihlosp_6		.6004473	.2311027	2.60	0.009	.1474316	1.053463
_Ihlosp_7		.6112917	.2296168	2.66	0.008	.1611886	1.061395
_Ihlosp_8		.7055086	.2293639	3.08	0.002	.2559013	1.155116
_Ihlosp_9		.6810592	.2317898	2.94	0.003	.2266965	1.135422
_Ihlosp_10		.7282987	.2303611	3.16	0.002	.2767367	1.179861
_Ihlosp_11		.9076078	.2292791	3.96	0.000	.4581666	1.357049
_Ihlosp_12		.7355061	.2374668	3.10	0.002	.2700152	1.200997
_Ihlosp_13		.9562865	.2323421	4.12	0.000	.5008413	1.411732
_Ihlosp_14		1.060794	.2409619	4.40	0.000	.5884518	1.533136
_Imarsth_2		.3709433	.0462105	8.03	0.000	.2803598	.4615268
_Imarsth_3		.1048837	.0677711	1.55	0.122	-.0279636	.2377311
_Imarsth_4		-.1250447	.0465665	-2.69	0.007	-.216326	-.0337634
_Imarsth_5		-.0425507	.1493504	-0.28	0.776	-.3353126	.2502113
_Icmap_935		-.1541335	.0268411	-5.74	0.000	-.2067484	-.1015185
alone		.2102727	.0384597	5.47	0.000	.1348827	.2856627
unitsp		.0239406	.0092155	2.60	0.009	.0058761	.0420051
_Iolnp_2		-1.468731	.651815	-2.25	0.024	-2.746442	-.1910201
_Iolnp_3		-.0798289	.0347452	-2.30	0.022	-.1479377	-.0117202
_Iolnp_4		-1.053147	.3835622	-2.75	0.006	-1.805019	-.3012743
vismin		-.0912763	.0445894	-2.05	0.041	-.1786821	-.0038705
aborig		-.1962926	.053485	-3.67	0.000	-.3011358	-.0914494
_cons		9.651174	.235676	40.95	0.000	9.189194	10.11315

a Why is `_Iagegp_7` dropped?

You have to drop 1 element of a list of dummies. Stata chose this one for age.

b Why is `white` not a regressor in the regression?

Here, the list of dummies would be `white`, `vismin`, `aborig`.

c How is it that so many coefficients are significant, and yet R-squared is only 12%? Does this suggest a problem in the model?

R-squared measures the size of sigma-squared relative to the variance of Y . Just because sigma-squared is big doesn't have anything to do with how good X is at predicting the mean of Y . It just says that Y has a lot of conditional variance remaining after X is conditioned out.

d The constant is highly significant, with a t-value of 41. Is this surprising? Why or why not?

The constant measures the expectation of the log of earnings for white man aged 55-69, with unreported level of schooling, etc. The t-test tests whether or not this is zero. Zero is the log of \$1. Thus, the t-stat is big because this kind of person has an expected earnings level very far from \$1. Not surprising.

3. [4 points] Suppose that $Y_i = X_i\beta + \varepsilon_i$, where X is a single column with X between 1 and 2,

$E[X_i'\varepsilon_i] = 0$ and $E[(\varepsilon_i)^2] = \sigma^2 X_i$. Here, the variance of the disturbance rises linearly with X .

What is the variance of the OLS estimate of the (scalar) parameter in this case? Is it larger or smaller than the standard regression output? How much bigger or smaller?

a. $V[\hat{\beta}] = (X'X)^{-1} X'E(\varepsilon\varepsilon')X(X'X)^{-1} = (X'X)^{-1} X'DX(X'X)^{-1}$, where D is a diagonal matrix with $E[(\varepsilon_i)^2] = \sigma^2 X_i$ on the main diagonal. Writing this out in summation form:

$$b. V[\hat{\beta}] = \frac{1}{\sum_{i=1}^N x_1^2} \sum_{i=1}^N \sigma^2 x_1^3 \frac{1}{\sum_{i=1}^N x_1^2} = \sigma^2 \frac{1}{\sum_{i=1}^N x_1^2} \frac{\sum_{i=1}^N x_1^3}{\sum_{i=1}^N x_1^2} = \mu_{\text{classical OLS}}[\hat{\beta}] \frac{\sum_{i=1}^N x_1^3}{\sum_{i=1}^N x_1^2}$$

- c. since X is between 1 and 2, its cube is larger than its square, so this true variance is larger than the reported variance in standard regression output.
4. [4 points] Islam estimates a cross-country panel model of per-capita income on a bunch of right-hand side variables. a) Why didn't Islam use the random effects GLS estimator?; b) It could be that time affects every country differently. Why didn't Islam interact time dummies with country dummies?
- random effects model requires that country effects are orthogonal to X . maybe he didn't think this was true.
 - if you interact time dummies with country dummies, you have a term for every data point. plus, you have X . this is more columns of regressors than rows of observations, and so is not identified.
5. [4 points] Why are error terms in regressions mean-zero? Why do we say it doesn't matter if there is measurement error in Y ? In contrast, why might it matter if there is measurement error in X ?
- linear regressions satisfy $X'e=0$, so that error terms are mean-zero by construction of the estimator.
 - if Y is measured with error, then $Y=Z+u$, where Z is the true regressand and u is the measurement error. Then, $Z=XB+e-u$, and the composite error term is still spherical.
 - If $X=Z+u$, and u is correlated with the disturbance term, then this induces endogeneity. Also, there is the problem of pushing the estimated parameter towards zero, but we haven't talked about that yet.
6. [4 points] If you have heteroskedasticity of unknown form, you cannot get a consistent estimator for the weighting matrix, because you need one element for every observation.
- How does the White hetero-robust covariance matrix estimator get around this problem?
 - This estimator uses an estimate of $X'\Omega X$ ($K \times K$) instead of Ω ($N \times N$). Thus, a consistent estimate can be constructed. They should show the covariance matrix of the OLS estimator under heteroskedasticity, and show the role of these matrices.
 - How does it affect the estimated coefficients, compared to the OLS estimates?
 - it only changes the estimated covariance matrix of the OLS estimated coefficient vector--it does not affect the estimated coefficients.
7. [4 points] Provide Stata (or MATLAB) code to create dummies for people who are in the following 5 categories: (1) registered Indians; (2) non-registered Indians who self-identify as a) North American Indian; b) Metis; c) Inuit; (3) people who are neither registered Indians nor who self-identify as Aboriginal.

```
tab ABSRP
```

ABORIGINAL IDENTITY		Freq.	Percent	Cum.
1	Non-Aboriginal population	429,180	97.41	97.41
2	Single North American Indian	7,105	1.61	99.02
3	Single Métis	3,400	0.77	99.80
4	Single Inuit	498	0.11	99.91
5	Multiple Aboriginal responses	73	0.02	99.93
6	Aboriginal responses not included elsew	329	0.07	100.00
Total		440,585	100.00	

```
tab REGINP
```

REGISTERED OR TREATY INDIAN INDICATOR		Freq.	Percent	Cum.
1	Registered under the Indian Act	6,498	1.47	1.47
2	Not registered under the Indian Act	434,087	98.53	100.00
Total		440,585	100.00	

One could use something like

```
gen regind=REGINP==1
gen NAI=(ABSRP==2)&(REGINP>1)
gen Metis=(ABSRP==3)&(REGINP>1)
gen Inuit=(ABSRP==4)&(REGINP>1)
gen nonAbor=(REGINP==2)&(ABSRP==1)
drop if ABSRP>4
```

this exhausts the possibilities. lots of other codings would work, too.