Econ 836 Midterm Exam

1. [14 points] Consider the following code and output from a log-wage regression using 2006 Census data on male residents of Toronto. The first line sets the line delimiter to ";".

```
.  use "C:\DATA\2006 Census\pumf2006.dta", clear;
.g insamp=POB==1&AGEGRP>8&AGEGRP<18&COW==4&HDGREE>1&HDGREE<88&WAGES>100&CFSIZE<9&CFSIZE>0&CMA==535&SEX==2&VISMIN<88;
. drop if insamp==0;
(833472 observations deleted)
. generate logwage=log(WAGES);
. generate not_alone=CFSIZE~=1;
. replace CFSIZE=CFSIZE-1;
(11004 real changes made)
. *NOTE for ethnic categories, everything comes from ABOID and VISMIN, and that "|" means "or";
. generate aborig=ABOID<6;
. generate white=VISMIN==13&aborig==0;
. generate chinese=VISMIN==1&aborig==0;
. generate southasian=VISMIN==2&aborig==0;
. generate caribblack=VISMIN==3&aborig==0;
. generate othvismin=VISMIN>4&VISMIN<13&aborig==0;
. generate vm=VISMIN<13&aborig==0;
. generate notwhite=VISMIN<13|aborig==1;
. xi: regress logwage i.AGEGRP i.HDGREE i.MARST not_alone CFSIZE chinese southasian caribblack notwhite
       aborig;
i.AGEGRP          _IAGEGRP_9-17       (naturally coded; _IAGEGRP_9 omitted)
i.HDGREE          _IHDGREE_2-13       (naturally coded; _IHDGREE_2 omitted)
i.MARST           _IMARST_1-5         (naturally coded; _IMARST_1 omitted)

      Source |       SS       df       MS              Number of obs =   11004
-------------+------------------------------           F( 30, 10973) =   97.64
       Model |  2059.31498    30  68.6438326           Prob > F      =  0.0000
    Residual |  7713.98876 10973  .702997245           R-squared     =  0.2107
-------------+------------------------------           Adj R-squared =  0.2086
       Total |  9773.30374 11003  .888239911           Root MSE      =  .83845

------------------------------------------------------------------------------
     logwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 _IAGEGRP_10 |   .3222072   .0283047    11.38   0.000     .2667248    .3776895
 _IAGEGRP_11 |   .4325841   .0305029    14.18   0.000     .3727929    .4923753
 _IAGEGRP_12 |    .518103   .0304299    17.03   0.000      .458455     .577751
 _IAGEGRP_13 |   .5740694    .031979    17.95   0.000     .5113848    .6367541
 _IAGEGRP_14 |   .5695695   .0349644    16.29   0.000      .501033     .638106
 _IAGEGRP_15 |   .5004705   .0408812    12.24   0.000      .420336    .5806049
 _IAGEGRP_16 |   .2319792    .050103     4.63   0.000     .1337683    .3301902
 _IAGEGRP_17 |  -.1181318   .0761101    -1.55   0.121    -.2673213    .0310577
  _IHDGREE_3 |   -.041436   .0399403    -1.04   0.300    -.1197262    .0368542
  _IHDGREE_4 |   .1697118    .039882     4.26   0.000     .0915359    .2478878
  _IHDGREE_5 |   .0435886   .0542511     0.80   0.422    -.0627535    .1499306
  _IHDGREE_6 |   .1232402   .0302268     4.08   0.000     .0639902    .1824901
  _IHDGREE_7 |   .2132195   .0291102     7.32   0.000     .1561582    .2702808
  _IHDGREE_8 |   .2578788   .0407753     6.32   0.000     .1779519    .3378056
  _IHDGREE_9 |   .4763058   .0228869    20.81   0.000     .4314433    .5211682
 _IHDGREE_10 |   .5585993   .0439811    12.70   0.000     .4723885    .6448101
 _IHDGREE_11 |   .3818988      .1296     2.95   0.003     .1278595    .6359382
 _IHDGREE_12 |    .570872   .0326154    17.50   0.000       .50694    .6348041
 _IHDGREE_13 |   .6211139   .0799199     7.77   0.000     .4644565    .7777714
   _IMARST_2 |   .2160764   .0361877     5.97   0.000      .145142    .2870107
   _IMARST_3 |  -.0304484   .0555723    -0.55   0.584     -.13938    .0784833
   _IMARST_4 |  -.2207229   .0370627    -5.96   0.000    -.2933724   -.1480733
   _IMARST_5 |  -.0848602   .1256429    -0.68   0.499    -.3311428    .1614224
   not_alone |  -.1038007   .0294991    -3.52   0.000    -.1616243   -.0459772
      CFSIZE |   .0472362   .0087003     5.43   0.000      .030182    .0642905
     chinese |   .0000648   .0757365     0.00   0.999    -.1483923    .1485219
  southasian |   .0016856   .0790663     0.02   0.983    -.1532985    .1566697
  caribblack |   -.123238    .073679    -1.67   0.094     -.267662     .021186
    notwhite |  -.1239527   .0530695    -2.34   0.020    -.2279785   -.0199269
      aborig |   .0332937   .1001404     0.33   0.740    -.1629995     .229587
       _cons |   10.20313   .0463208   220.27   0.000     10.11233    10.29392
```

a. Do Aboriginal men have lower earnings than white men, conditional on age, education, marital status, and household size? What is the conditional expectation of the difference in log-earnings between Aboriginal and white men?

b. The constant is highly significant, with a t-value of 220. Is this surprising? Why or why not? What is the meaning of the constant term?

c. What is the predicted difference in log-earnings for a household with 1 member versus a household with 2 members?

d. Why is R-squared (equal to V($X\hat{\beta}$)/V(Y)) so low when so many coefficients have big t-values?

e. Why is `_IAGEGRP_9` omitted?

f. What is the average of the residual vector $e = Y - X\hat{\beta}$?

g. Is the residual *e* correlated with household size (the variable `CFSIZE`)?

2. [8 points] Pendakur and Pendakur (1998) estimate models of earnings which control for education, and investigate the differences in earnings across ethnic groups.

a. If there were unobserved quality variation for people with the same reported education level, how would this affect your interpretation of the estimates?

b. Assume that 'field-of-study' is available in the data (it is). Should it be included in the regression? Does excluding it induce bias? Why?

c. Does it matter that they drop all observations for which income from wages and salaries is zero?

d. Suppose these authors wanted to investigate the conditional median of log-earnings rather than the conditional mean. Would this give them a way to deal with zeroes?

2. [8 points] Allen, Pendakur and Suen (2005) estimates a panel model with the standard deviation of the log of age at first marriage on the LHS and no-fault status and state and year dummies on the RHS.

a. They do not include any information about the population of the state in the model. Likewise, there is not information on education levels in the state. Does this matter? Under what conditions does it not matter? Are these plausible conditions?

b. Why didn't they use the *random effects* FGLS estimator?

c. It could be that time affects every country differently. Why didn't they interact time dummies with country dummies?

d. These authors regress median age at first marriage in a state on legal characteristics. Could they have run a quantile regression to address their question? If so, what quantile regression?

4. [4 points] Suppose that: $Y_i = X_i\beta + \varepsilon_i$, for *i=1,…,N*; X is a single column with X a range between 1 and 2; and $E[\varepsilon] = 0_N$, $E\left[(\varepsilon_i)^2\right] = \frac{\sigma^2}{(X_i)^2}$ and $E[\varepsilon_i\varepsilon_j] = 0$ for all *i* not equal to *j*. Here, the variance of the disturbance decreases with *X*, and there are no correlations in disturbances across observations.

a. What is the standard error of the OLS estimate of the (scalar) parameter in this case? Is it larger or smaller than the standard error given in regression output which assumes homoskedasticity? How much bigger or smaller?

b. Derive the GLS estimator for this case, and show how you would implement it. In what way would it treat observations where *X=2* differently from those where *X=1*?

5. [6 points] Jacks and Pendakur (2010) estimate the effect of freight prices on international trade volumes.

a. Suppose they regressed Y on covariates X but not on country dummies or decade dummies. Under what conditions is this estimator unbiased? Under what conditions is it efficient?

b. Is there a better estimator for the case when it is unbiased but not efficient? If so, what? If not, why not?

c. What about regressing Y on covariates X and dummies for each decade (but not each country)? What about regressing Y on X and dummies for each decade and each country and the interaction of these two vectors of dummy variables.