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?

Aboriginal men have `notwhite=1` and `aborig=1` . The former is statistically significant with a big t-value, but the latter is statistically insignificant. A good answer is *either*: yes, they earn less because notwhite is big and significant; *or*, uncertain because we don't know the covariance of the two relevant parameters. The conditional expectation is `-.1239527 + .0332937 = -0.09` but `-.1239527` is also acceptable if they say they're not including the Aborig coefficient because it is insignificant.

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?

the constant gives the conditional expectation of log-earnings for a person that has all other variables equal to zero: a white man with the lowest age and education who lives alone. The t-test tests the hypothesis that this log-earnings is zero, corresponding to annual earnings of $1. These guys are poor, but not that poor. The hypothesis being tested is not very interesting because it tests something so obviously untrue, yielding a gigantic t-value.

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

a household with 1 member has `not_alone=0` and has `CFSIZE=0` ; a household with 2 members has `not_alone=1` and has `CFSIZE=1` , yielding a difference of `| -.1038007 + .0472362 =0.056`.

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

R-squared is low when the variance of epsilon is high. The variance of epsilon is high when there is a lot of unexplained variation. This does not imply that the coefficients are estimated imprecisely or that t-values will be small. T-values depend on the parameter estimate and the estimated std error. The latter shrinks with the variance of X and the size of N and grows with the variance of epsilon. This sample has a big N, so the std errors are small.

e. Why is `_IAGEGRP_9` omitted?

You have to omit one of the categories from each vector of dummy variables to avoid them being collinear in each other. The one omitted is arbitrary. (Stata happens to pick the lowest value.)

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

This average is zero from the first order condition of OLS regression. X'e=0, and X contains a constant, so that 1'e=0.

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

This average is zero from the first order condition of OLS regression. X'e=0, and this implies that there is no correlation between X and e for any X.

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?

   i. If unobserved quality affects earnings *and* is correlated with their regressors, then it causes omitted variable bias through that correlation. If it is uncorrelated with the regressors, it does not cause bias.

   ii. In the former case, I'd have to interpret the coefficients on observed regressors as carrying the load of both their direct effect and an indirect effect through their correlation with the unobserved quality.

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?

i. Is field of study correlated with earnings? Is it correlated with any regressors of interest---ethnic origin? If yes and yes, then excluding it induces bias.
ii. Should it be included? argue either way on the basis of wanting that induced bias or not.

c. Does it matter that they drop all observations for which income from wages and salaries is zero?
i. Is this sample selection random? If minorities are more, or less, likely to be non-workers, then, the nonrandom sampling would induce bias.
ii. If minorities are out of the labour market *because* their opportunities are worse, then excluding this information would change your conclusions about disparity.

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?
i. demonstrate you know what a conditional median is.
ii. how might it solve the problem (could include missings as zeroes).
iii. (it might not work, too---what if workers don't work, e.g., because they are too rich?)

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?
i. are population or education levels correlated with the disturbance term? are they correlated with the law? well, more educated people get married later, and more educated people may want laws that favour people who marry later. In this case, you'd have a correlated missing regressor and induced bias. alternatively, you might argue that these correlations are about zero.

b. Why didn't they use the *random effects* FGLS estimator?
i. random effects for states would require that state effects are uncorrelated with the law. however, they are not. states with low age at first marriage are more likely to be fault. thus, although the RE model would be lower variance in its estimated parameters, it would also have induced bias.

c. It could be that time affects every country differently. Why didn't they interact time dummies with country dummies?
i. this question makes no sense given that it is states. two possible answers: nonsense, there are no countries to dummy out; or, if they interacted time dummies with *state* dummies, they'd have more regressors than observations and couldn't run the regression.

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?
i. the quantile regression would have to be at the person-level, rather than the state-year level. It could have been `xi: qreg age_at_first_marriage i.state nf1`
they don't have to state the code, but rather what the regression would be.

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 $E[\varepsilon] = 0_N$, $E\left[(\varepsilon_i)^2\right] = \dfrac{\sigma^2}{(X_i)^2}$ and $E\left[\varepsilon_i\varepsilon_j\right] = 0$ 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?

$$V(\hat{\beta}) = (X'X)^{-1} X'\Omega X (X'X)^{-1}$$

i. $$\Omega = \sigma^2 diag\left(\frac{1}{X_i^2}\right)$$

ii. Since X is only 1 column:

$$V(\hat{\beta}) = \left(\sum_{i=1}^{N} X_i^2\right)^{-1} \sum_{i=1}^{N} X_i \, \Omega_i X_i \left(\sum_{i=1}^{N} X_i^2\right)^{-1}$$

$$= \left(\sum_{i=1}^{N} X_i^2\right)^{-1} \sigma^2 \left(\sum_{i=1}^{N} X_i^2\right)^{-1} = \sigma^2 \left(\sum_{i=1}^{N} X_i^2\right)^{-2}$$

iii. This is equal to the OLS homoskedastic variance times $\left(\sum_{i=1}^{N} X_i^2\right)^{-1}$. Since X is between 1 and 2, its sum of squares is between N and 4N, both of which are bigger than 1, so that term is smaller than 1. So, the actual variance of the estimated coefficient is smaller than the OLS homoskedastic variance.

iv. Nuance (for extra credit): the estimate of σ^2 would differ between the two estimated variances.

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*?

i. $\hat{\beta}_{GLS} = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}Y$

ii. Since the variance matrix is diagonal, we could use WLS, with weights equal to the minus-one-half matrix of Omega: $w_i = X_i$.

iii. So, we could regress $X_i Y_i$ on X_i^2.

iv. This gives observations where *X=2* twice the weight of 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?

i. unbiased if all those dummy variables are uncorrelated with freight prices *or* uncorrelated with trade volumes.

ii. efficient if those dummy variables would all have coefficients of zero. not efficient otherwise.

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

i. Random effects model is efficient if the above model is unbiased.

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.

i. All of these are feasible, and unbiased. this does *not* have the problem of running out of data, because for each decade/country, there are up to 10 observations (10 years of data). so, this is feasible, but it uses a lot of degrees of freedom. you sacrifice efficiency if their true coefficients are uncorrelated with trade volume or with freight prices, because in that case you could have used the random effects model.