ME104 Linear Regression Analysis: Problem Set 6

Problems with Errors

1. Today we use again the dataset `parking.dta`. The data concern diplomats from 146 countries stationed at the United Nations in New York City.

`violations`	the number of parking tickets which were issued to diplomatic vehicles from a country and which were not paid (annual average number for the period 11/199711/2002).
`corruption`	a measure of the level of corruption in a country for 1998, with higher levels indicating higher levels of corruption. This reflects both social norms and level of legal enforcement, but only the contribution of the norms may be transferred to a diplomats environment in New York.
`logdipl`	logarithm of the number of diplomats at the countrys UN mission.
`loggdp`	the logarithm of the countrys GDP per capita in 1998 (in year-2000 US dollars).

   (a) After generating a new variable using the command `gen logviol=ln(1+violations)`, fit a linear model for `logviol given corruption logdipl and loggdp`.

   (b) Graphically inspect homoscedasticity of residuals using the `rvfplot` command. What do you conclude?

   (c) Use the Breusch-Pagan test to test the null hypothesis that the variance of the residuals is homogenous using the command `estat hettest`. Interpret.

   (d) The Breusch-Pagan test does not work well for non-linear forms of heteroskedasticity and when the errors are not normally distributed. For such cases a special case of the Breusch-Pagan can be used. This test is the White's test estimated using the stata command `estat imtest, white` or just `imtest, white`. Run the test and interpret the result. For more tests see `help regress postestimation`.

   (e) After predicting the residuals of the fitted regression model using the stata command `predict residuals, resid` use the commands `qnorm` and `pnorm` to check the normality of the residuals. What do you conclude?

   (f) A formal test for normality of the residuals is the Smirnov-Kolmogorov test that you can perform using the command `sktest resid`. Interpret.

   (g) Perform two model specification tests using the command `linktest` and `ovtest`. Interpret the results of these tests.

2. Additional work: Structural equations
   The data file used is `hiedatashortnew.dta`. The following variables are included:

`age`	current age in years.
`education`	years of schol completed.
`income`	family income for year preceding enrollment.
`sex`	participant's sex.
`urban`	=1 if lived in urban area.
`nearsch`	=1 if participant lived near college ten years before enrollment.

   The goal is to estimate the percentage effect on income of getting an extra year of education, controlling for age, sex and urban. It is commonly thought that education is correlated with the error term in the income equation (unobserved ability). This would result in OLS over-estimating the effect of education on the log income. It is hard to find instruments since they need to be uncorrelated with the error term, yet help to predict years of schooling. In this example, some information on how far a participant lived from college 10 years earlier is used as instrument.

   (a) This `ivreg` command computes the 2SLS estimates. The dependent variable is `logincome`. The regressors that are assumed exogenous are `age sex` and `urban`. The regressors assumed endogenous is `education`. The instrumental variable is `nearsch`. The key assumption is that distance from college is not correlated with the error in the income equation, but do help to explain years of schooling.

```
1 . ivreg logincome age sex urban (educ = nearsch)

  Instrumental variables (2SLS) regression

        Source |       SS       df       MS              Number of obs =    1679
  -------------+------------------------------           F(  4,  1674) =    9.44
         Model | -2955.82362     4 -738.955904           Prob > F      =  0.0000
      Residual |  4058.24546  1674  2.42428044           R-squared     =       .
  -------------+------------------------------           Adj R-squared =       .
         Total |  1102.42184  1678  .656985603           Root MSE      =   1.557

  ------------------------------------------------------------------------------
     logincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
  -------------+----------------------------------------------------------------
     education |   .6114749   .2428244     2.52   0.012     .1352034    1.087746
           age |   .0360272   .0104272     3.46   0.001     .0155754    .0564789
           sex |   -.708351   .1675626    -4.23   0.000    -1.037005   -.3796968
         urban |   .0104216   .1239284     0.08   0.933    -.2326492    .2534925
         _cons |   -6.61945   3.334928    -1.98   0.047    -13.16052   -.0783822
  ------------------------------------------------------------------------------
  Instrumented:  education
  Instruments:   age sex urban nearsch
  ------------------------------------------------------------------------------
```

(b) We also estimate the same equation by OLS in order to compute the Hausman test statistic.

```
1 . reg logincome age education sex urban

      Source |       SS       df       MS              Number of obs =    1679
-------------+------------------------------           F(  4,  1674) =   63.56
       Model |  145.364797     4  36.3411994           Prob > F      =  0.0000
    Residual |  957.057044  1674  .571718664           R-squared     =  0.1319
-------------+------------------------------           Adj R-squared =  0.1298
       Total |  1102.42184  1678  .656985603           Root MSE      =  .75612

------------------------------------------------------------------------------
   logincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0143356   .0016154     8.87   0.000     .0111673     .017504
   education |   .0794863   .0072232    11.00   0.000     .0653188    .0936537
         sex |  -.4482515   .0575321    -7.79   0.000     -.561094    -.335409
       urban |  -.0843647   .0564106    -1.50   0.135    -.1950074     .026278
       _cons |   .6814894   .1168803     5.83   0.000     .4522424    .9107363
------------------------------------------------------------------------------
```

(c) The command `hausman` computes the Hausman test statistic. The null hypothesis is that the OLS estimator is consistent. If accepted, we probably would prefer to use OLS instead of 2SLS. Discuss the results of the hausman test.

```
1 . hausman ivreg .,constant sigmamore df(1)

  Note: the rank of the differenced variance matrix (1) does not equal the number of coefficients being
        tested (5); be sure this is what you expect, or there may be problems computing the test.
        Examine the output of your estimators for anything unexpected and possibly consider scaling your
        variables so that the coefficients are on a similar scale.

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     ivreg          .          Difference          S.E.
-------------+----------------------------------------------------------------
   education |    .6114749     .0794863        .5319886        .1176998
         age |    .0360272     .0143356        .0216915        .0047991
         sex |    -.708351    -.4482515       -.2600995        .0575457
       urban |    .0104216    -.0843647        .0947863         .020971
       _cons |    -6.61945     .6814894        -7.30094        1.615296
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from ivreg
            B = inconsistent under Ha, efficient under Ho; obtained from regress

    Test:  Ho:  difference in coefficients not systematic

                  chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       20.43
                Prob>chi2 =      0.0000
                (V_b-V_B is not positive definite)
```