ECON4135: Solution to Written Paper 2

26th October 2006

General comments

Some general comments regarding the problem set, and your answers:

- Some of you misunderstood some of the problem. Some problems were ambiguous. I've generally been fairly tolerant and employed considerable good will when correcting. Read through this solution set to see what was expected.
- You should always include output from Stata, to show what you have done (some more detailed comments in footnote 2).
- Remember, you are economists, not statisticians! So, while it of course is crucial to do the estimation and calculations correctly, don't stop there. Try to give your results a (brief) economic interpretation. This is of course particularly important when you are explicitly asked to comment or interpret the results.
- Also, proofread your writing. Try to avoid meaningless sentences, and generally try to be concise.

Problem 1

We have a regression equation given as:

$$\ln Y_i = \alpha + \beta_1 S_i + \beta_2 E_i + \beta_3 E_i^2 + u_i \tag{1}$$

In order to estimate it, we make the following assumptions about the error term, u_i:

$$E(u_i|S_i, E_i) = 0 \tag{2}$$
$$(\ln Y_i, S_i, E_i) \quad are \quad i.i.d.\ vectors \tag{3}$$
$$var(u_i|S_i, E_i) = \sigma^2 \tag{4}$$

Assumption (2) is essential, it ensures that $cov(u_i, S_i) = cov(u_i, E_i) = 0$. This is required for the OLS-estimators to be unbiased and consistent, that is the estimators are on average correct, and as the number of observations increases the probability that the estimators will be very different from the true values becomes small.[1]

Assumption (3) assures that the error terms are independent across observations.

[1] Note that as long as we include a constant term in the regression, $E(u_i) = 0$ is not restrictive. The critical part of assumption (2) is that the covariates does not contain any information about the error term.

The last assumption is of *homoskedasticity* (i.e., equal error term variance across all observations). This assumption is *not* required for the OLS-estimators to be unbiased or consistent, but if it is not satisfied the estimated standard errors of the OLS-estimators will be misleading, and there will exist other unbiased estimators with smaller standard errors. Some students have rather assumed heteroskedasticity, i.e. the error term variance varies over individuals ($var(u_i|S_i, E_i) = \sigma_i^2$), and use robust standard errors (`reg ln_y s e e_2, robust`). This is perfectly ok, and it may even be argued to be prudent.

Using Stata to estimate the model, we get the results shown below.[2]

```
. reg ln_y s e e_2

      Source |       SS       df       MS              Number of obs =    5859
-------------+------------------------------           F(  3,  5855) =  403.06
       Model |  99.5404561     3   33.180152           Prob > F      =  0.0000
    Residual |  481.985139  5855  .082320263           R-squared     =  0.1712
-------------+------------------------------           Adj R-squared =  0.1707
       Total |  581.525595  5858   .09927033           Root MSE      =  .28692

------------------------------------------------------------------------------
       ln_y_ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |   .0599545   .0017712    33.85   0.000     .0564823    .0634267
           e |   .0293536   .0031084     9.44   0.000     .0232601    .0354471
         e_2 |  -.0005593      .0001    -5.60   0.000    -.0007552   -.0003633
       _cons |   7.340757   .0255112   287.75   0.000     7.290746    7.390768
------------------------------------------------------------------------------
```

From the computer output we see that the conditional expectation is given as[3]

$$E(\ln Y_i|S_i, E_i) = 7.34 + .060S_i + .029E_i - .00056E_i^2 \tag{5}$$

We see that the estimated marginal return to schooling ($\hat{\beta}_1$) is equal to 0.060, with a 95% confidence interval of [0.056,0.063]. From the confidence interval we see that β_1 is fairly precisely estimated, and significantly different from 0. If schooling increase by one year, we expect income to increase by 6%.

Problem 2

In Problem 1 we found that $\beta_2 > 0$, and $\beta_3 < 0$. For small values of E_i the effect of another year will be close to β_3, but as E_i increases β_3 will become more important. Thus, the return

[2] When reporting results, you should always include a (part of a) log, specifying both the command you used, and the output from the program. You can do this either by inserting the output into the document, or by appending a log. When estimating large models (as in Problem 4) you may exclude the irrelevant coefficients. The appended or inserted log will be a lot easier to read if you use a fixed width font (for example `courier`) or format the regression output as a table.

[3] When reporting regression results, try to use a meaningful level of precision. Report at least the two first non-zero digits, if you report $\beta_3 = 0.001$, this could mean anything from 0.0005 to 0.0015, which differ by a factor of 3. However, it is seldom relevant to report more than than three to four digits either, the twelfth digit will typically neither be precisely estimated nor interesting. However, when doing calculations you should include a few extra decimal places to avoid error due to lacking numerical precision.

to experience is positive (at least for small values of E_i), but decreasing.

Using for example the command `sum e` I find the average of E_i, $\bar{E}$, to be 15.1:

```
. sum e

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           e |      5859     15.0978     5.95621          0         29
```

Thus, the marginal return to education, evaluated at the mean, can be found as $\beta_1 + 2\beta_2\bar{E} = 0.029 - 2 \cdot 0.00056 \cdot 15.1$. Evaluating this in Stata (as shown here, the expression is evaluated at machine precision, `_b[·]` refers to the estimated coefficients and `r(mean)` refers to a result saved by the previous command, `sum e`):

```
. di _b[e]+2*_b[e_2]*r(mean)
.01246646
```

Thus, at the sample mean experience (15.1 yrs), the estimated marginal return to experience is 1.2% of income per extra year. While still positive, this is less than half of $\hat{\beta}_2$.

For the marginal return to be zero for all values of E_i, we must have $H_0 : \beta_2 = \beta_3 = 0$.[4] We can test this hypothesis using the `test` command, which performs a F-test:

```
. test e e_2

 ( 1)  e = 0
 ( 2)  e_2 = 0

       F(  2,  5855) =  175.29
            Prob > F =    0.0000
```

Since the probability is less than 0.05 (and also less than any other meaningful level of significance), we reject H_0, and accept the alternative $H_1 : \beta_2 \neq 0 or \beta_3 \neq 0$. Looking back at the regression output from Problem 1, this should not be surprising. There we found that both β_2 and β_3 are significantly different from 0. Thus, it was not really necessary to perform the test above.

There will be a linear relationship between $\ln Y_i$ and E_i if $\beta_2 \neq 0$ and $\beta_3 = 0$. From the regression output we see that the hypothesis $H_0 : \beta_3 = 0$ is rejected by a t-test, and thus, we conclude that the relationship between $\ln Y_i$ and E_i is non-linear.

Problem 3

There are several way to check for multicollinearity. A straightforward way is to try to run the regression, and see what happens. If there is multicollinearity, Stata will drop some variable. If not, the regression will proceed as usual, with output reported for all variables.

Alternatively we may generate a variable equal to the sum of all variables suspected to be part of the problem, and inspect this.[5] Below I show how to do this for the regional dummies:

[4] Quite a few misunderstood this question. This is what was expected.

[5] Yet another way is to use the commands `_rmcoll` and `_rmdcoll`.

```
. egen regsum = rsum(ostf-finmark)

. tab regsum

     regsum |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,165       19.88       19.88
          1 |      4,694       80.12      100.00
------------+-----------------------------------
      Total |      5,859      100.00
```

From the log, we see that there are both values of 0 and 1. Thus, we do not have perfect multicollinearity. In the case of the region dummies, the reason for this is that the there is no dummy for Oslo, i.e. Oslo is the reference county. Also for the other dummy sets there is no multicollinearity. For sectors the reference is non-service sectors (agriculture, manufacturing, energy etc.), and for education the reference is educations within science/technology.[6]

Problem 4

Below I show parts of the regression output (omitting most of the dummy variables):

```
. reg ln_y s- finmark

      Source |       SS       df       MS              Number of obs =    5859
-------------+------------------------------           F( 32,  5826) =   78.42
       Model |  175.065341    32   5.4707919           Prob > F      =  0.0000
    Residual |  406.460255  5826  .069766607           R-squared     =  0.3010
-------------+------------------------------           Adj R-squared =  0.2972
       Total |  581.525595  5858   .09927033           Root MSE      =  .26413

------------------------------------------------------------------------------
       ln_y_ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |    .077245   .0019866    38.88   0.000     .0733505    .0811395
           e |    .032566   .0028982    11.24   0.000     .0268845    .0382475
         e_2 |  -.0006226   .0000931    -6.68   0.000    -.0008052     -.00044
      public |  -.2045894    .011631   -17.59   0.000    -.2273906   -.1817882
       servi |   -.023135   .0081435    -2.84   0.005    -.0390994   -.0071707
         :
     finmark |  -.1825931   .0313749    -5.82   0.000    -.2440997   -.1210866
       _cons |   7.345807   .0266337   275.81   0.000     7.293595     7.39802
------------------------------------------------------------------------------
```

We see that the estimated $\hat{\beta}_1$ from the full model is higher than the one we found in Problem 1, and the confidence intervals are not even overlapping.

[6] Finding the reference sector and education is not so easy without specific knowledge. You were expected however to show that multicollinearity is not problem here.

The change in the estimated return to schooling indicates that we may have problems with an omitted variable bias. This means that some variable(s) correlate both with schooling and log earnings. When omitting this variable, its effect on earnings will be attributed to schooling, to the extent that there is a correlation between schooling and the variable in question. Possible candidates for variables that cause a bias when omitted are all the dummy variables. Below I give the correlation pattern between schooling and the sector dummies.

```
. corr s public serv (obs=5859)

             |        s   public     serv
-------------+---------------------------
           s |   1.0000
      public |   0.3641   1.0000
        serv |   0.0945   0.2052   1.0000
```

We see that there is a fairly strong positive correlation between the public sector dummy and schooling. Furthermore, from the regression output we see that income in public sector (sadly, at least to us employed there...) is lower than in other sectors. Thus, omitting the public sector dummy may have led us to understate the significance of schooling. Intuitively, this is because many individuals with much schooling work for low wages in the public sector.

Problem 5

Comparing models (1) and (2), we see that the predictive power of model (2), as measured by R^2, is much larger. This model explains just over 30% of the variation in $\ln Y_i$, as opposed to just over 17% for model (1). Looking at at the adjusted R^2 ($\bar{R}^2$), we see that this increase is more than what we would expect from the simple fact that the number of covariates increases. Furthermore, the models yield (most likely significantly) different results for the main coefficients of interest (β_1), and it is natural to interpret the difference as an omitted variable bias in model (1).

However, another question is what we really want to estimate. In some cases, the possibility of getting well paid private sector jobs may be counted as part of the return of a education. Taking such a view, we may feel that we correct too much if we control for sector.

To sum up, it is not clear which model is the more relevant or better. But in most cases it probably would be reasonable to view model (2) as the complete one, and model (1) as a simpler and possibly biased version.

Below I have included regression output from running OLS on model (2) for females.

```
. reg ln_y s- finmark

      Source |       SS       df       MS              Number of obs =    3247
-------------+------------------------------           F( 32,  3214) =   43.20
       Model |  79.5451549    32  2.48578609           Prob > F      =  0.0000
    Residual |  184.938739  3214  .057541611           R-squared     =  0.3008
-------------+------------------------------           Adj R-squared =  0.2938
       Total |  264.483894  3246  .081479943           Root MSE      =  .23988
```

```
------------------------------------------------------------------------------
       ln_y_ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |   .0704718   .0025202    27.96   0.000     .0655304    .0754132
           e |   .0114862   .0032517     3.53   0.000     .0051106    .0178618
         e_2 |  -.0000584   .0001068    -0.55   0.585    -.0002677     .000151
      public |  -.1775465   .0134934   -13.16   0.000     -.204003   -.1510899
       servi |  -.0522459   .0128918    -4.05   0.000    -.0775229   -.0269689
           :
     finmark |  -.1485397   .0299443    -4.96   0.000    -.2072515   -.0898278
       _cons |   7.345333   .0341969   214.80   0.000     7.278283    7.412383
------------------------------------------------------------------------------
```

Comparing the results with those found in Problem 4, we see that all the β's are smaller than the corresponding ones we found for males. We get a very similar R^2-value, just over 30%. The constant terms are very similar, reflecting the same general income level for males and females. As this is in stark contrast to both public beliefs and the conclusions of most other studies, it is questionable if the sampling technique used in obtaining the data sets is one that makes the data sets representative of the entire population.

Looking at the estimated return to schooling ($\hat{\beta}_1$), we see that $\hat{\beta}_1^{female} < \hat{\beta}_1^{males}$. However the difference is not necessarily significant. If we define the difference in the return to schooling as

$$\delta = \beta_1^{male} - \beta_1^{female} \tag{6}$$

Then the variance of the difference will be given as

$$var(\delta) = var(\beta_1^{male}) + var(\beta_1^{female}) - 2cov(\beta_1^{male}, \beta_1^{female}) \tag{7}$$

Assuming the covariance is 0, the variance of the difference becomes simply the sum of the variances. Thus, testing $H_0 : \delta = 0$ we have the test statistic

$$t_\delta = \frac{\hat{\delta}}{se\left(\hat{var(\delta)}\right)} = \frac{\hat{\delta}}{\sqrt{\hat{var(\delta)}}} = \frac{\hat{\beta}_1^{male} - \hat{\beta}_1^{female}}{\sqrt{var(\hat{\beta}_1^{male}) + var(\hat{\beta}_1^{female})}} \sim t_{N-K} \tag{8}$$

I.e. the estimator t_δ is t-distributed, with $N-K$ degrees of freedom, where N is the combined number of observations and K is the combined number of estimated parameters. From the regression outputs we get the following numbers:

$$t_\delta = \frac{0.0772 - 0.0705}{\sqrt{0.00199^2 + 0.00252^2}} = 2.09 \sim t_{N-K} \tag{9}$$

Since $N-K$ is "large", we can approximate the t-distribution with a standard normal, so the critical value of a two-sided test at the 5%-significance level will be $t_c = 1.96$. As $t_\delta > t_c$ we reject $H_0 : \delta = 0$ and accept the alternative, $H_1 : \delta \neq 0$. Thus, we conclude that the difference is significant, and that the return to schooling is greater for men.[7]

[7] Here we assumed that there was zero covariance between β_1^{male} and β_1^{female}. Alternatively, we could have combined the data for males and females (Stata command `append`), and estimated the difference in return as a interaction coefficient (the coefficient associated with the variable S_i times a dummy variable equal to one if

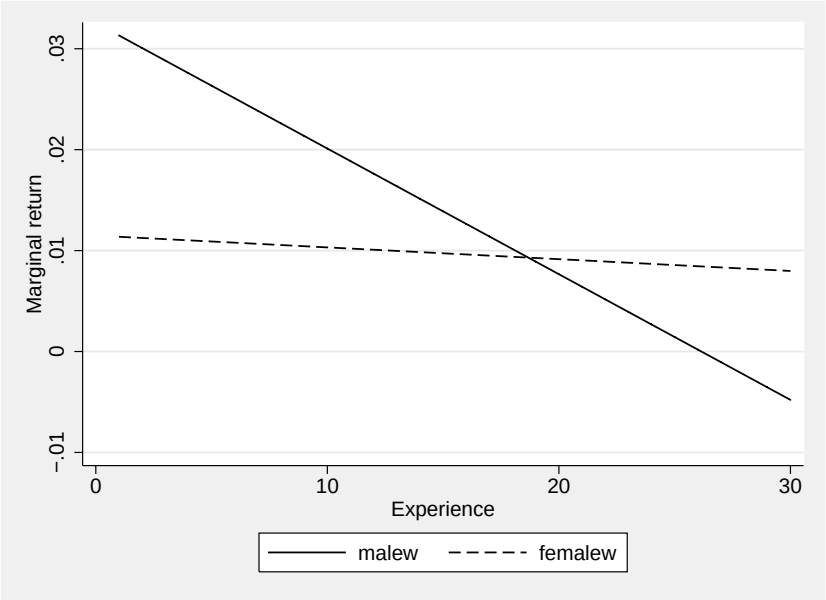

Figure 1: The marginal return to experience.

Problem 6

The marginal return to experience is given as $\beta_2 + 2\beta_3 e_i$. For the males the estimate of this will be $0.33 - 2 \cdot 0.00062E_i$, while for the females it is $0.11 - 2 \cdot 0.000058E_i$. Below I depict this in a figure, letting experience vary from 1 to 30 (remember from Problem 2 that we found experience to vary between 1 and 29 for the males). I also include the code used to generate the graph.[8]

```
. drop _all

. set obs 30 obs was 0, now 30

. egen e = fill(1 2)
```

the individual is female). Then the test comes out as a simple t-test which is included in the default regression output. If we have interaction terms for all variables the results from the regression will be very similar to those obtained by running separate regressions (identical if we allow the error terms to be heteroskedastic with gender, i.e. have different variance for males and females), but we also get the covariances between the coefficients, which may influence whether we reject the hypothesis of the coefficients being equal. In this particular case, this alternative approach yields a t-statistic of 2.05 for δ, which still is significant. Some of you appended the data sets and ran the regression without gender interaction on other variables than schooling. This may be ok, but this means that you assume all the dummies to have similar effects for men and women. If this is not the case, you may get an omitted variable bias. Those who ran a regression including only schooling and the schooling gender interaction can be pretty certain to get OVB, and invalid results from the test, this approach is not recommended.

[8]Note that I could have generated the `male`-variable as `g male=0.33-2*0.00062e`, and similarly for the females. Instead, I use the saved results from a regression of both males' and females' earnings on the covariates, with interaction terms on all the variables for the females (see footnote to Problem 5). If you don't want to mess with the data in memory you can use the command `twoway function`.

```
. g males   = _b[e]+2*_b[e_2]*e

. g females = (_b[e]+_b[kve])+2*(_b[e_2]+_b[kve_2])*e

. line males females e, xti(Experience) scheme(s2mono)
saving(margret,replace) (file margret.gph saved)
```

We see from the figure that the males have a marginal return to experience early in their working lives (i.e. when their experience is low) of over 3%. This is high compared to the females, who start out with a marginal return to experience of just over 1%. The males' marginal return then drop rapidly, and is under 1.5% as their experience tends towards 30 years. However, it is still higher than the initial level of the females' marginal return to experience, which is only barely reduced over their working life.

This difference is remarkable, and it is not clear what causes it. One possibility is that there is a common return to actual working experience. We measure potential working experience however. As females generally have a looser connection to the labor market (due to maternity leave, part time work etc.), the average amount of actual (market) working experience may be lower for any given level of potential experience than for men.

Problem 7

In order to answer this problem we make some assumptions:[9]

- We only care for life cycle income, i.e. schooling does not have any intrinsic value, nor do we have preferences over other results of schooling, such as type of job we expect to get. Thus, maximizing utility amounts to maximize life cycle income, which is calculated as the discounted sum of all income.
- We assume a discount rate of 3%.
- We get no income while studying. There also are no direct costs associated with studying, only foregone earnings (and foregone experience).
- We work from the time we finish schooling until we are 67 years old. Then we die, or (for our purposes equivalently) retire and get a pension which is independent of income prior to retirement.
- Schooling is discrete, i.e. we must have an integer number of years of schooling.
- Income is, to our knowledge, best described by the regression in this exercise. Thus we have no indication that our own error term deviates from zero, or that our return to schooling and experience is different from the average.

Given these assumptions, we get the results presented in the Stata-log below. We note that for both males and females 11 years of schooling after the first 7 yields the highest life cycle

[9] These assumptions are not realistic, but the do serve to make the problem tractable. An interesting exercise is to try to think trough them, think of alternative assumptions, and how changing the assumptions would affect the results. There may also be other assumptions implicit, that I have failed to notice.

income. This is sensitive to our assumptions, however. Try for example to run Stata-code for different values of the discount rate.

Note that as this problem is very open, there were no particular requirements for the answers. The question was meant as an opportunity for you to place the regression results in a context, and to show some economic intuition.

An example of a short answer to this problem could have been: "I assumed that I'm rational. Thus, my optimal level of schooling will be whatever I choose."

```
. foreach g in males females {
  2.   use earningsdata_‘g’ ,clear
  3.   qui reg ln_y s-finmark
  4.   keep in 1/54
  5.   egen age = fill(14 15)
  6.   foreach var of varlist public-finmark {
  7.     qui replace ‘var’ = 0
  8.   }
  9.   forvalues s = 0/11 {
 10.     qui replace s   = ‘s’
 11.     qui replace e   = age - 14 -‘s’
 12.     qui replace e_2 = e^2
 13.     predictnl w‘s’ = exp(xb())
 14.     qui replace w‘s’ = 0 if e<0
 15.     qui replace w‘s’ = w‘s’*0.97^_n
 16.   }
 17.   save lifecycle2_‘g’,replace
 18.   collapse (sum) w*
 19.   xpose ,clear varname
 20.   g s = real(substr(_varname, 2,.))
 21.   drop _varname
 22.   order s
 23.   rename v1 ‘g’
 24.   sort s
 25. }
(5805 observations deleted) file lifecycle2_males.dta saved (3193
observations deleted) file lifecycle2_females.dta saved

. merge s using lifecycle_males

. drop _merge

. format %8.0f males females

. l

     +----------------------+
     |  s   females   males |
     |----------------------|
```

```
 1. |  0     49207   53172 |
 2. |  1     50730   55412 |
 3. |  2     52283   57724 |
 4. |  3     53864   60107 |
 5. |  4     55473   62562 |
    |----------------------|
 6. |  5     57108   65086 |
 7. |  6     58768   67679 |
 8. |  7     60451   70338 |
 9. |  8     62154   73063 |
10. |  9     63877   75848 |
    |----------------------|
11. | 10     65616   78693 |
12. | 11     67368   81591 |
    +----------------------+
```

.