Home assignment # 4
(Suggested solutions)

1. For this exercise, use the data in *fertil.gdt* file. It contains information about a sample of women in Botswana: their education (variable *educ*), their age (variable *age*), and the number of children they have (variable *children*).

   (a) Estimate the model
   $$children = \beta_0 + \beta_1 educ + \varepsilon \quad ,$$
   interpret the coefficient β_1.

   (b) Conduct the Breusch-Pagan test for heteroskedasticity. Does its result justify the use of robust standard errors? If yes, reestimate the model using these robust standard errors and comment on the difference with respect to the model estimated by simple OLS.

   (c) Redefine the model as
   $$children = \beta_0 + \beta_1 educ + \beta_2 age + \beta_3 age^2 + \varepsilon$$
   and estimate it.

      i. Does age have a significant impact on number of children of Botswana women? State the hypothesis, and compute the test statistics by hand. Interpret the results and compare it to the results of the test in Gretl.

      ii. Do you find justification for the inclusion of *age* in quadratic form? [Hint: State the four specification criteria and argue if they are satisfied for the quadratic in age.]

      iii. How does the coefficient β_1 change when compared to part (a)? Does this signal any bias in the model from part (a)? Where does it come from? Explain the sign of this bias.

   (d) Do you think the coefficient β_1 from the model in part (c) may suffer from some omitted variable bias? What variable(s) could be missing in the model and how would the coefficient β_1 change if they were included in the regression? (What is the sign of the potential bias in coefficient β_1?)

   (e) Conduct the RESET test of the model from part (c) and interpret the results.

Solution:

(a) The estimation result is:

```
Model 1: OLS, using observations 1-4361
Dependent variable: children

             coefficient   std. error   t-ratio    p-value
  ------------------------------------------------------------
  const        3.49554     0.0561238     62.28    0.0000    ***
  educ        −0.209650    0.00796014   −26.34    5.41e-142 ***

Mean dependent var   2.267828   S.D. dependent var   2.222032
Sum squared resid    18571.78   S.E. of regression   2.064112
R-squared            0.137287   Adjusted R-squared   0.137089
F(1, 4359)           693.6651   P-value(F)           5.4e-142
Log-likelihood      −9347.408   Akaike criterion     18698.82
Schwarz criterion    18711.58   Hannan-Quinn         18703.32
```

The meaning of the coefficient β_1 is that an additional year of education will reduce the number of children a woman has by 0.2. (Obviously, *children* is not a continuous variable, so this interpretation sounds a little bit odd, but we have to realize we talk here in averages.)

(b) The result of the test is:

```
Breusch-Pagan test for heteroskedasticity
OLS, using observations 1-4361
Dependent variable: scaled uhat^2

             coefficient   std. error   t-ratio    p-value
  ------------------------------------------------------------
  const       1.52659      0.0421310     36.23    1.53e-251 ***
  educ       −0.0899239    0.00597551   −15.05    6.20e-50  ***

  Explained sum of squares = 543.72

Test statistic: LM = 271.859811,
with p-value = P(Chi-square(1) > 271.859811) = 0.000000
```

The p-value equal to zero indicates that we should reject the null hypothesis of homoskedasticity. It means that we have a heteroskedastic error term in the model and we should use robust standard errors.

When we reestimate the model using robust standard errors, we get:

```
Model 2: OLS, using observations 1-4361
Dependent variable: children
Heteroskedasticity-robust standard errors, variant HC0

            coefficient   std. error   t-ratio    p-value
  ------------------------------------------------------------
  const        3.49554     0.0663462     52.69    0.0000    ***
  educ        -0.209650    0.00836023   -25.08    9.34e-130 ***

Mean dependent var   2.267828   S.D. dependent var   2.222032
Sum squared resid    18571.78   S.E. of regression   2.064112
R-squared            0.137287   Adjusted R-squared   0.137089
F(1, 4359)           628.8617   P-value(F)           9.3e-130
Log-likelihood      -9347.408   Akaike criterion     18698.82
Schwarz criterion    18711.58   Hannan-Quinn         18703.32
```

As expected, we have the same coefficients here, but the standard errors are larger, showing that they were underestimated by OLS in part (a).

(c) The estimation result of the redefined model is:

```
Model 3: OLS, using observations 1-4361
Dependent variable: children
Heteroskedasticity-robust standard errors, variant HC0

            coefficient   std. error    t-ratio   p-value
  ------------------------------------------------------------
  const      -4.13831      0.243509     -16.99    9.13e-63 ***
  educ       -0.0905755    0.00604551   -14.98    1.61e-49 ***
  age         0.332449     0.0191983     17.32    5.07e-65 ***
  sq_age     -0.00263082   0.000351804   -7.478   9.07e-14 ***

Mean dependent var   2.267828   S.D. dependent var   2.222032
Sum squared resid    9284.147   S.E. of regression   1.459746
R-squared            0.568724   Adjusted R-squared   0.568427
F(3, 4357)           1923.764   P-value(F)           0.000000
Log-likelihood      -7835.592   Akaike criterion     15679.18
Schwarz criterion    15704.71   Hannan-Quinn         15688.19
```

i. To test for a significant impact of age on number of children of Botswana women, we need to use an F-test:

$$H_0 : \beta_2 = 0 \& \beta_3 = 0 \quad \text{vs} \quad H_A : \beta_2 \neq 0 \vee \beta_3 \neq 0$$

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(n-k)} \sim F_{J,n-k} \quad ,$$

$$F = \frac{(18571.78 - 9284.147)/2}{9284.147/(4361-4)} = 2179 \quad ,$$

$$F_{crit(2,4357;0.95)} = 3.00$$

Therefore, we reject the null hypothesis of insignificant coefficients. The quadratic form of age has a positive impact on number of children. The results of the test in Gretl confirm our conclusion:

```
Restriction set
 1: b[age] = 0
 2: b[sq_age] = 0

Test statistic: Robust F(2, 4357) = 2675.9, with p-value = 0

Restricted estimates:

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const        3.49554     0.0561238     62.28     0.0000    ***
  educ        -0.209650    0.00796014   -26.34     5.41e-142 ***
  age          0.000000    0.000000      NA        NA
  sq_age       0.000000    0.000000      NA        NA

  Standard error of the regression = 2.06411
```

ii. The specification criteria: test of significance of the coefficient, change in R^2, bias in other coefficients, and theory. The significant coefficients for the *age* and age^2 variables, a sharp rise in R^2 and a dramatic change of coefficient for variable *educ* signal that this model performs much better than the one in part (a) and age in quadratic form should indeed be included in the model. Moreover, theory (intuition) also suggests that *age* and age^2 should be included in the model: obviously, very young women have less children than older women, because they have just started their families. The impact of age will be smaller for older women, because after some age, it is biologically impossible to have more children: this justifies the quadratic specification in *age*, reflecting the concave relationship.

iii. The coefficient β_1 decreases significantly in absolute value when compared to part (a), which signals that there was an omitted variable bias given by the omission of the *age* variable. This bias was negative and was given by the fact that *age* is a relevant variable for the model (and its correlation with the dependent variable is positive - see the coefficient in the estimated equation) and that it is correlated with the variable *educ* (this correlation is negative, which we can verify in Gretl and explain by possible development of education in Botswana, leading to higher education of new generations as compared to the old generations). The bias is proportional to the product of these two correlations and this is why we should expect it to be negative.

(d) We can hypothesize that the coefficient β_1 from the model in part (c) may still suffer from some omitted variable bias. A possible variable that could have an impact on the number of children a woman has may be some variable representing socio-economic characteristics of the woman's family. Since we can suppose that women in better social and economic situation may have less children (negative correlation) and more education (positive correlation), we would expect the bias to be negative. Hence, if we had such variable and we included it in the equation, we should expect the coefficient to become even less negative than in part (c) (smaller in absolute value).

(e) The result of the RESET test of the model from part (c) is:

```
Auxiliary regression for RESET specification test
OLS, using observations 1-4361
Dependent variable: children

             coefficient   std. error    t-ratio   p-value
  ---------------------------------------------------------
  const      -2.87905      0.442674     -6.504    8.72e-11 ***
  educ        0.00335782   0.0119351     0.2813   0.7785
  age         0.247641     0.0339369     7.297    3.48e-13 ***
  sq_age     -0.00402865   0.000599533  -6.720    2.06e-11 ***
  yhat^2      0.294301     0.0384241     7.659    2.29e-14 ***
  yhat^3     -0.0210462    0.00537522   -3.915    9.16e-05 ***

Test statistic: F = 41.877395,
with p-value = P(F(2,4355) > 41.8774) = 9.67e-19
```

The very low p-value signals that we should reject the null hypothesis that the model is correctly specified. This signals that there are still some variables which are omitted from our estimation or that our specification has an incorrect functional form.

2. Suppose following investment model was estimated with quarterly data from 1997-2009 (standard errors in parenthesis):

$$\underset{}{I_t =} \underset{(1.10)}{7.70} + \underset{(0.23)}{0.55}\, Y_t + \underset{(0.12)}{0.63}\, Q_{t2} + \underset{(1.03)}{1.55}\, Q_{t3} + \underset{(0.74)}{2.13}\, Q_{t4} \quad ,$$

$$n = 64 \quad , \quad R^2 = 0.72 \quad ,$$

where I_t is the investment in period t, Y_t is the GDP in period t, and dummy variables Q_{ti} are equal to 1 in the i-th quarter and zero otherwise ($i = 2, 3, 4$). Denote the coefficients associated with the dummies δ_2, δ_3 and δ_4.

   (a) What restriction on these parameters would lead to the model:

$$I_t = \beta_0 + \beta_Y Y_t + \delta q_t + \varepsilon_t \quad ,$$

   where $q_t = 0$, 1, 2, 3 in the first, second, third and fourth quarters respectively? Briefly discuss this restriction. [Hint: To find the restrictions, compare the coefficients of the two models (restricted and unrestricted) for each quarter.]

   (b) Test the restriction if the regression R^2 of the restricted model was 0.68.

   (c) Explain how would you test for presence of AR(4) autocorrelation of the error term in this model. Describe all steps that you need to take to conduct the test, the null and alternative hypothesis, and the test statistics.

Solution:

(a) To find the restrictions, we will compare the expected values of I_t in each quarter for both models:

$$\begin{aligned} I_t &= \alpha + \beta_Y Y_t + \delta_2 Q_{t2} + \delta_3 Q_{t3} + \delta_4 Q_{t4} + \eta_t \\ I_t &= \beta_0 + \beta_Y Y_t + \delta q_t + \varepsilon_t \quad . \end{aligned}$$

1^{st} quarter:

$$\begin{aligned} E[I_t] &= \alpha + \beta_Y Y_t \\ E[I_t] &= \beta_0 + \beta_Y Y_t \end{aligned}$$

2^{nd} quarter:

$$\begin{aligned} E[I_t] &= \alpha + \delta_2 + \beta_Y Y_t \\ E[I_t] &= \beta_0 + \delta + \beta_Y Y_t \end{aligned}$$

3^{rd} quarter:

$$\begin{aligned} E[I_t] &= \alpha + \delta_3 + \beta_Y Y_t \\ E[I_t] &= \beta_0 + 2\delta + \beta_Y Y_t \end{aligned}$$

4^{th} quarter:

$$\begin{aligned} E[I_t] &= \alpha + \delta_4 + \beta_Y Y_t \\ E[I_t] &= \beta_0 + 3\delta + \beta_Y Y_t \end{aligned}$$

The comparison gives us the following results:

$$\begin{aligned} &1^{st} \text{ quarter} \Rightarrow \alpha = \beta_0 \\ &2^{nd} \text{ quarter} \Rightarrow \alpha + \delta_2 = \beta_0 + \delta \\ &3^{rd} \text{ quarter} \Rightarrow \alpha + \delta_3 = \beta_0 + 2\delta \\ &4^{th} \text{ quarter} \Rightarrow \alpha + \delta_4 = \beta_0 + 3\delta \end{aligned}$$

which reduces finally to the two restrictions

$$\begin{aligned} \delta_3 &= 2\delta_2 \\ \delta_4 &= 3\delta_2 \quad . \end{aligned}$$

These restrictions assume constant difference between the quarters accross the year (meaning that there is e.g. the same difference between the first an the second quarter as between the second and the third one).

(b) We will test the null hypothesis that these restiction are valid using the standard F-test over restricted and unrestricted models. As given in the setup, $R^2_U = 0.72$ and $R^2_R = 0.68$. Number of restrictions $J = 2$, number of observations $n = 64$, number of parameters of unrestricted model $k = 5$. We construct the F-statistic

$$F = \frac{(R^2_U - R^2_R)/J}{(1 - R^2_U)/(n-k)} = \frac{(0.72 - 0.68)/2}{(1 - 0.72)/(64 - 5)} = 4.2143$$

and when we compare it to the corresponding critical value $F_{2,59} = 3.1531$, we see that we can reject the null hypothesis that the restrictions are valid.

(c) AR(4) autocorrelation of the error term implies that the error term has the following structure:

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \rho_3 \varepsilon_{t-3} + \rho_4 \varepsilon_{t-4} + u_t$$

We can test for AR(4) autocorrelation of the error term using analysis of residuals. Since OLS is consistent even under autocorrelation, the residuals are consistent estimates of the stochastic error term and we can thus use them to test for autocorrelation of the error term. The test proceeds as follows:

i. Estimate the original model by OLS, save the residuals $e_t = I_t - \widehat{I}_t$.

ii. Estimate the model $e_t = \alpha + \rho_1 e_{t-1} + \rho_2 e_{t-2} + \rho_3 e_{t-3} + \rho_4 e_{t-4} + u_t$ by OLS.

iii. Test if $\rho_1 = \rho_2 = \rho_3 = \rho_4 = 0$ using the standard F-test:

$$H_0 : \rho_1 = 0 \& \rho_2 = 0 \& \rho_3 = 0 \& \rho_4 = 0 \text{ vs } H_A : \rho_1 \neq 0 \vee \rho_2 \neq 0 \vee \rho_3 \neq 0 \vee \rho_4 \neq 0$$

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(n-k)} \sim F_{J,n-k} \quad .$$