MARCO ANGRISANI
Exercise 3, G023 Econometric Theory and Methods

In this handout you can find STATA commands to solve Exercise 3, the output given by the program and brief comments to the main results.

3.1 and 3.2 - Food Expenditure Engel Curve

3.1.a

Read the dataset in STATA

```
. use "D:\Teaching\MC3\EX3\food1.dta", clear
```

Descriptive statistics

```
. sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
number	2510	1255.5	724.7189	1	2510
region	2510	5.42988	3.065978	0	9
mic	2510	.8007968	.3994807	0	1
frez	2510	.9394422	.2385649	0	1
memhh	2510	2.497211	1.294357	1	10
earners	2510	1.148606	.9961204	0	6
finc	2510	349.4622	252.1596	32	2950
agehoh	2510	50.87331	16.85788	19	96
adltm	2510	.8749004	.5698408	0	4
adltf	2510	1.003586	.4845018	0	4
child	2510	.6187251	1.014074	0	7
totfdexp	2510	45.44273	31.84617	0	604.96
benefit	2510	.1055777	.3073576	0	1
fat	2510	192.1234	158.2647	0	2837.797
energy	2510	4296.439	2903.637	.806262	30759.88
qfj	2510	133.7128	296.9524	0	5811.258
fincpc	2510	154.6415	111.2148	16	1538
lnfincpc	2510	4.839871	.6364217	2.772589	7.338238
fj	2510	.3609562	.4803736	0	1
x1	2510	4.839871	.6364217	2.772589	7.338238
x2	2510	23.82922	6.164547	7.687248	53.84974
x3	2510	119.2341	45.9939	21.31358	395.1622
fdpc	2510	20.03978	13.5727	0	302.48
fdshare	2510	.1664394	.1243159	0	1.912
lfinc	2510	5.620613	.7081545	3.465736	7.989561
lmemhh	2510	.7807421	.5280412	0	2.302585

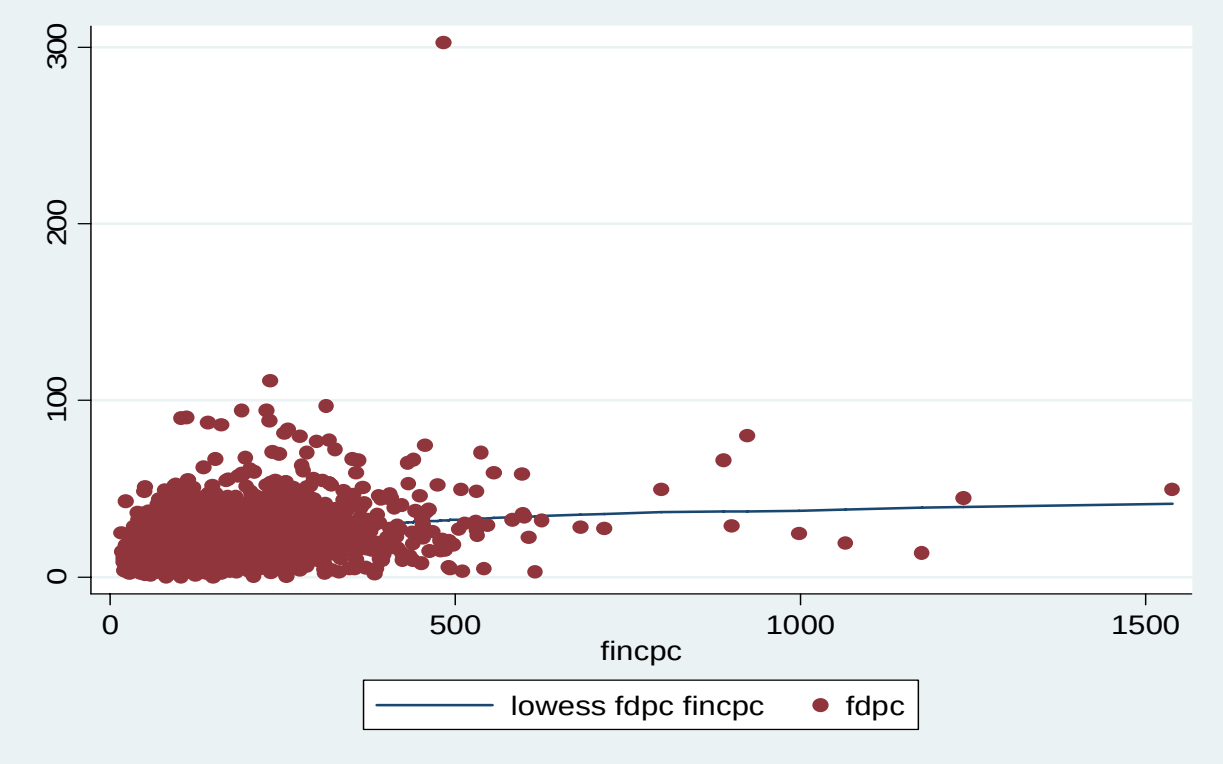
Comments:

1. The mean of memhh is 2.49, this is the average household size.
2. On average only one person earns an income within the household (mean of earners).
3. The average age of the head of the household is roughly 51.
4. Many households in the sample do not have children.
5. Big variation in income and food expenditure.

3.1.b

Graphics and Nonparametric regressions

```
. twoway (lowess fdpc fincpc, bwidth(0.2)) (scatter fdpc fincpc)
```



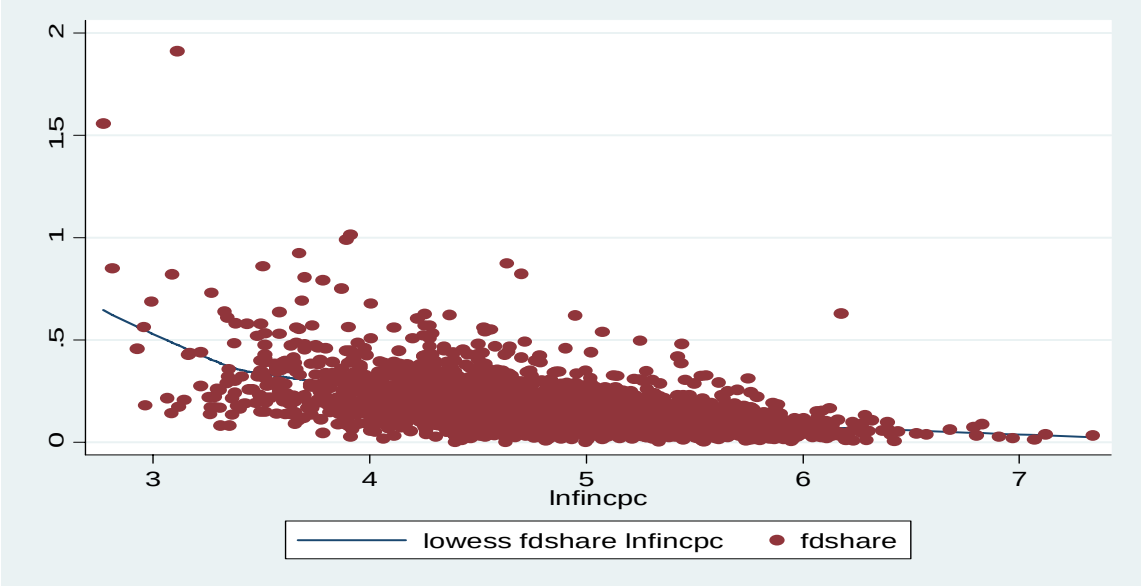
The command *lowess* (STATA 9 or 10) carries out a locally weighted regression. Roughly speaking, this nonparametric method is based on fitting a linear model to observations in a neighbourhood of a point. The width of such a neighbourhood is specified using the option *bwidth* (in this exercise we choose a width of 0.2). Notice that the greater the bandwidth, the greater the smoothing of the regression function. This procedure of fitting a linear model using observations in a neighbourhood of a point is repeated for each point. For example, if you have 1000 observations, this nonparametric estimator requires running 1000 locally weighted regressions.

Comments:

1. Food expenditure per capita increases with income per capita.
2. The conditional mean function (estimated slope of the relationship) becomes flatter.
3. Also the dispersion increases with income levels (hetero rather than homoskedasticy).

3.1.c

```
. twoway (lowess fdpc fincpc, bwidth(0.2)) (scatter fdpc fincpc)
```



Comments:

1. Inverse relationship between food share and (log)income: households with a lower income level tend, on average, to spend a higher fraction of their income in food.
2. The dispersion is relatively higher when income is relatively lower (heteroskedasticity).

3.2.a

OLS estimation of the parameters of the model: $fdshare = \beta_1 + \beta_2 \log fincpc + \varepsilon$ **(1)**

```
. reg fdshare lnfincpc

      Source |       SS       df       MS              Number of obs =    2510
-------------+------------------------------           F(  1,  2508) = 1012.37
       Model |  11.1507845     1  11.1507845            Prob > F      =  0.0000
    Residual |  27.6243907  2508   .01101451           R-squared     =  0.2876
-------------+------------------------------           Adj R-squared =  0.2873
       Total |  38.7751751  2509  .015454434           Root MSE      =  .10495

------------------------------------------------------------------------------
     fdshare |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lnfincpc |  -.1047508   .0032922   -31.82   0.000    -.1112065   -.0982951
       _cons |   .6734199    .016071    41.90   0.000     .6419062    .7049336
```

Using

$$\Pr\left[-1.96 \le \frac{-0.1047508-\beta_2}{0.0032922} \ge 1.96\right] \approx 0.95$$

we obtain the approximate 95% confidence interval for β_2: $\left[-0.111203, -0.098298\right]$.

We reject the hypothesis $H_0: \beta_2 = 0$ using both a test with size 0.05 and a test with size 0.01.

OLS estimation (robust standard errors) of equation (1)

```
. reg  fdshare lnfincpc, robust

Regression with robust standard errors                Number of obs =    2510
                                                      F(  1,  2508) =  391.26
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.2876
                                                      Root MSE      =  .10495

------------------------------------------------------------------------------
             |               Robust
     fdshare |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lnfincpc |  -.1047508   .0052957   -19.78   0.000    -.1151353   -.0943664
       _cons |   .6734199   .0269366    25.00   0.000     .6205997    .7262402
------------------------------------------------------------------------------
```

Comments:

1. Using robust standard errors we take into account the presence of heteroskedasticity (something that seemed quite evident in the first part of the exercise).
2. Robust standard errors tend to be larger than conventional standard errors.
3. Confidence intervals are wider than before.
4. We still reject the null hypothesis $H_0: \beta_2 = 0$ using both a test with size 0.05 and a test with size 0.01.

3.2.b

Extended model: $fdshare = \gamma_1 + \gamma_2 \log finc + \gamma_3 \log memhh + \varepsilon$ **(2)**

```
. reg fdshare lfinc lmemhh

      Source |       SS       df       MS              Number of obs =    2510
-------------+------------------------------           F(  2,  2507) =  602.24
       Model |  12.5836195     2  6.29180976           Prob > F      =  0.0000
    Residual |  26.1915556  2507   .01044737           R-squared     =  0.3245
-------------+------------------------------           Adj R-squared =  0.3240
       Total |  38.7751751  2509  .015454434           Root MSE      =  .10221

------------------------------------------------------------------------------
     fdshare |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lfinc |   -.115337   .0033313   -34.62   0.000    -.1218694   -.1088046
      lmemhh |   .0683164   .0044676    15.29   0.000     .0595558     .077077
       _cons |   .7613665   .0173601    43.86   0.000     .7273249    .7954081
------------------------------------------------------------------------------
```

```
. vce

             |    lfinc   lmemhh    _cons
-------------+---------------------------
       lfinc |  .000011
      lmemhh | -7.5e-06   .00002
       _cons | -.000057  .000026  .000301
```

We use the statistics

$$T = \frac{\hat{\gamma}_2 + \hat{\gamma}_3}{s\left[c'\left(X'X\right)^{-1}c\right]^{1/2}} \overset{a}{\sim} N(0,1)$$

to test $H_0 : \gamma_2 + \gamma_3 = 0$ (where $c' = \begin{pmatrix} 0 & 1 & 1 \end{pmatrix}$). We obtain $T = -11.71$ and we reject H_0.

Alternatively, let *RSSr* be the residuals sum of squares $\hat{\varepsilon}'\hat{\varepsilon}$ of the restricted model in equation (1) and *RSSu* the residuals sum of squares $\hat{\varepsilon}'\hat{\varepsilon}$ of the unrestricted model in equation (2). We can perform the test comparing the quality of the fit of the two estimated models.
More precisely, we can use the statistics:

$$F = \frac{\left(RSSr - RSSu\right)}{RSSu / n} \overset{a}{\sim} \chi^2_{(1)}$$

The residuals sum of squares is provided by STATA in the standard regression output: it is the entrance *Residuals-SS* of the table on the top-left of the output. Therefore, from the two tables above, we get:
RSSr = 27.6243907, *RSSu* = 26.1915556, *n* = 2510, so we have F = 137.12 and we reject H_0.

3.2.c

Extended model with squared and cubed terms.

```
. gen lnfincpcs=lnfincpc^2

. reg fdshare lnfincpc lnfincpcs

      Source |       SS       df       MS              Number of obs =    2510
-------------+------------------------------           F(  2,  2507) =  570.01
       Model |  12.1207744     2  6.06038719           Prob > F      =  0.0000
    Residual |  26.6544008  2507  .010631991           R-squared     =  0.3126
-------------+------------------------------           Adj R-squared =  0.3120
       Total |  38.7751751  2509  .015454434           Root MSE      =  .10311

------------------------------------------------------------------------------
     fdshare |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lnfincpc |  -.4265726   .0338479   -12.60   0.000    -.4929453      -.3602
   lnfincpcs |   .0333773   .0034944     9.55   0.000     .0265251    .0402296
       _cons |   1.435641   .0813474    17.65   0.000     1.276126    1.595156
------------------------------------------------------------------------------

. gen lnfincpcc=lnfincpc^3

. reg fdshare lnfincpc lnfincpcs lnfincpcc

      Source |       SS       df       MS              Number of obs =    2510
-------------+------------------------------           F(  3,  2506) =  385.31
       Model |  12.2399055     3  4.07996851           Prob > F      =  0.0000
    Residual |  26.5352696  2506  .010588695           R-squared     =  0.3157
-------------+------------------------------           Adj R-squared =  0.3148
       Total |  38.7751751  2509  .015454434           Root MSE      =   .1029

------------------------------------------------------------------------------
     fdshare |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lnfincpc |  -1.123473   .2104962    -5.34   0.000    -1.536237   -.7107086
   lnfincpcs |   .1792202     .04362     4.11   0.000     .0936852    .2647552
   lnfincpcc |   -.009992   .0029789    -3.35   0.001    -.0158335   -.0041506
       _cons |   2.524618    .334655     7.54   0.000      1.86839    3.180847
------------------------------------------------------------------------------
```

Comments:

1. We reject the hypothesis that the coefficient on lnfincpcs is zero at 0.05 and 0.01.
2. We reject the hypothesis that the coefficient on lnfincpcc is zero at 0.05 and 0.01.
3. Using the same procedure as before, we can compare the sum of the squared residuals from the unrestricted model (this last one including both a squared and a cubed terms) with the one from the restricted model (the original model in equation (1)) to test the hypothesis that both coefficients are zero.

   We obtain $F = \frac{27.62439 - 26.535269}{27.62439/2510} = 98.96$.

   Confronting this value with the 0.95 quantile of a $\chi^2_{(2)}$, we reject the null hypothesis that the coefficients on the squared and cubed terms are zero.

3.2.d

The model that accounts for such regional variations can be written as:

$$fdshare = \delta_i D_i + \beta_2 \log fincpc + \varepsilon \qquad i = 0,1,2,3,4,5,6,7,8,9 \qquad \textbf{(3)}$$

$$fdshare = \beta_1 + \lambda_i D_i + \beta_2 \log fincpc + \varepsilon \qquad i = 0,1,2,3,4,5,6,7,8 \qquad \textbf{(4)}$$

In equation (3) we include 10 binary variables that capture regional effects on food expenditure and exclude the intercept. If we included the intercept, then there would be perfect multicollinearity among the regressors, since the vector of 1 that allows the model to have an intercept could be expressed as a linear combination of the 10 binary variables (in that case we have a singular data matrix). For this reason, in equation (4) we keep the intercept but we exclude one of the 10 regional dummies (say D_9).

To create binary variables use:

```
. gen d#=(region==#)     (#=0,1,2,3,4,5,6,7,8,9)
```

Or, using the interaction expansion command:

```
. xi, noomit: gen i.region
```

(if the option noomit is not specified, STATA automatically omits the first group, e.g. region 0).

Estimation of equation (3):

```
   . reg fdshare d0 d1 d2 d3 d4 d5 d6 d7 d8 d9 lnfincpc, noconst
```

or

```
   . xi, noomit: reg fdshare i.region lnfincpc, noconst
```

```
      Source |       SS       df       MS              Number of obs =    2510
-------------+------------------------------           F( 11,  2499) =  668.01
       Model |  80.8210925    11  7.34737204           Prob > F      =  0.0000
    Residual |  27.4863255  2499   .01099893           R-squared     =  0.7462
-------------+------------------------------           Adj R-squared =  0.7451
       Total |  108.307418  2510  .043150366           Root MSE      =  .10488

------------------------------------------------------------------------------
     fdshare |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          d0 |    .677663   .0178974    37.86   0.000     .6425677    .7127583
          d1 |   .6838962   .0172704    39.60   0.000     .6500304    .7177619
          d2 |   .6764677   .0176854    38.25   0.000     .6417882    .7111472
          d3 |   .6700149   .0175757    38.12   0.000     .6355505    .7044793
          d4 |   .6708468   .0171497    39.12   0.000     .6372178    .7044759
          d5 |   .6681124   .0177415    37.66   0.000     .6333228    .7029019
          d6 |    .678117    .017392    38.99   0.000     .6440127    .7122213
          d7 |   .6773539   .0176976    38.27   0.000     .6426505    .7120573
          d8 |     .66591   .0191852    34.71   0.000     .6282895    .7035305
          d9 |   .6878544   .0170216    40.41   0.000     .6544765    .7212323
    lnfincpc |  -.1058203   .0033358   -31.72   0.000    -.1123615   -.0992791
------------------------------------------------------------------------------
```

Comments:

1. Each dummy is a highly significant.
2. The estimated values of δ_i's are very similar and very close to the estimated value of β_1 in equation (1), this suggests that the regional effects on food expenditure are negligible.

Estimation of equation (4):

```
    . reg fdshare d0 d1 d2 d3 d4 d5 d6 d7 d8 lnfincpc,
Or
    . char region[omit] 9
    . xi: reg fdshare i.region lnfincpc, noconst
```

```
      Source |       SS       df       MS              Number of obs =    2510
-------------+------------------------------           F( 10,  2499) =  102.64
       Model |  11.2888496    10  1.12888496           Prob > F      =  0.0000
    Residual |  27.4863255  2499   .01099893           R-squared     =  0.2911
-------------+------------------------------           Adj R-squared =  0.2883
       Total |  38.7751751  2509  .015454434           Root MSE      =  .10488

------------------------------------------------------------------------------
     fdshare |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          d0 |  -.0101914   .0095251    -1.07   0.285    -.0288694    .0084865
          d1 |  -.0039582   .0078096    -0.51   0.612    -.0192722    .0113557
          d2 |  -.0113867   .0090152    -1.26   0.207    -.0290648    .0062914
          d3 |  -.0178395     .00828    -2.15   0.031    -.0340759   -.0016031
          d4 |  -.0170076   .0075319    -2.26   0.024     -.031777   -.0022381
          d5 |   -.019742   .0087527    -2.26   0.024    -.0369054   -.0025787
          d6 |  -.0097374   .0084484    -1.15   0.249    -.0263039    .0068291
          d7 |  -.0105005   .0075444    -1.39   0.164    -.0252945    .0042935
          d8 |  -.0219444   .0107725    -2.04   0.042    -.0430684   -.0008205
    lnfincpc |  -.1058203   .0033358   -31.72   0.000    -.1123615   -.0992791
       _cons |   .6878544   .0170216    40.41   0.000     .6544765    .7212323
```

Comments:

1. In equation (4), $\lambda_1, \lambda_2, ..., \lambda_8$ represent differential intercepts by reference to a common intercept β_1.
2. Given that we exclude D_9, we have $\beta_1 = \delta_9$, and for all the others $\delta_i = \beta_1 + \lambda_i \quad i = 0,1,...,8$.
3. Given that $\hat{\lambda}_i = \hat{\delta}_i - \hat{\beta}_1$, $i = 0,1,...,8$, are quite small and in general not statistically different from zero, there is not significant evidence of regional variations.
4. It does not matter which of the 10 indicators we drop in equation (4), since we measure differential intercepts by reference to a different common intercept. For example if we drop D_8 instead of D_9, we get:

```
. reg fdshare d0 d1 d2 d3 d4 d5 d6 d7 d9 lnfincpc
```

```
      Source |       SS       df       MS              Number of obs =    2510
-------------+------------------------------           F( 10,  2499) =  102.64
       Model |  11.2888496    10  1.12888496           Prob > F      =  0.0000
    Residual |  27.4863255  2499   .01099893           R-squared     =  0.2911
-------------+------------------------------           Adj R-squared =  0.2883
       Total |  38.7751751  2509  .015454434           Root MSE      =  .10488

------------------------------------------------------------------------------
     fdshare |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          d0 |    .011753   .0132736     0.89   0.376    -.0142754    .0377814
          d1 |   .0179862    .012106     1.49   0.137    -.0057527     .041725
          d2 |   .0105577   .0129136     0.82   0.414    -.0147647    .0358802
          d3 |   .0041049   .0124164     0.33   0.741    -.0202425    .0284524
          d4 |   .0049369   .0119288     0.41   0.679    -.0184546    .0283283
          d5 |   .0022024   .0127353     0.17   0.863    -.0227705    .0271752
          d6 |    .012207   .0125242     0.97   0.330    -.0123519    .0367659
          d7 |   .0114439   .0119476     0.96   0.338    -.0119843    .0348721
          d9 |   .0219444   .0107725     2.04   0.042     .0008205    .0430684
    lnfincpc |  -.1058203   .0033358   -31.72   0.000    -.1123615   -.0992791
       _cons |     .66591   .0191852    34.71   0.000     .6282895    .7035305
```

Comments:

1. Now $\hat{\beta}_1 = \hat{\delta}_8$, while $\hat{\lambda}_i = \hat{\delta}_i - \hat{\beta}_1 \quad i = 0,1,2,3,4,5,6,7,9$.
2. The *t*-statistics for the hypothesis that a particular binary indicator's coefficient is zero alters as we go between different specification of equation (4), because we change the matrix $\mathbf{X}$ of the regressors and then the matrix $(\mathbf{X}'\mathbf{X})^{-1}$ that we use to compute the *t*-statistics.

3.2.e

The extended model is

$$fdshare = \beta_1 + \lambda_i D_i + \beta_2 \log fincpc + \vartheta_i D_i \log fincpc + \varepsilon \qquad i = 0,1,2,3,4,5,6,7,8 \qquad \textbf{(5)}$$

To generate interaction variables use:
`. gen interac0=lnfincpc*d0` (repeat until interac9)

Estimation of equation (5):

```
. reg  fdshare d0-d8  lnfincpc interac0-interac8
```

```
      Source |       SS       df       MS              Number of obs =    2510
-------------+------------------------------           F( 19,  2490) =   54.87
       Model |  11.4436615    19  .602297975            Prob > F      =  0.0000
    Residual |  27.3315136  2490  .010976511            R-squared     =  0.2951
-------------+------------------------------           Adj R-squared =  0.2898
       Total |  38.7751751  2509  .015454434            Root MSE      =  .10477

------------------------------------------------------------------------------
     fdshare |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          d0 |  -.0163379   .0809919    -0.20   0.840    -.1751563    .1424804
          d1 |   .0594394    .060556     0.98   0.326    -.0593058    .1781847
          d2 |  -.0485005   .0719958    -0.67   0.501    -.1896783    .0926773
          d3 |  -.1126385   .0629785    -1.79   0.074    -.2361342    .0108572
          d4 |   .0249847   .0573772     0.44   0.663    -.0875273    .1374967
          d5 |  -.0631704   .0645564    -0.98   0.328    -.1897601    .0634193
          d6 |   .0276501   .0627531     0.44   0.660    -.0954036    .1507038
          d7 |  -.1468581   .0616333    -2.38   0.017     -.267716   -.0260003
          d8 |   -.142784   .0851855    -1.68   0.094    -.3098258    .0242578
    lnfincpc |  -.1094457   .0058883   -18.59   0.000    -.1209922   -.0978991
    interac0 |   .0010994   .0170043     0.06   0.948    -.0322446    .0344434
    interac1 |  -.0134538   .0124799    -1.08   0.281    -.0379259    .0110182
    interac2 |   .0076775   .0150259     0.51   0.609    -.0217871    .0371421
    interac3 |   .0196418   .0129242     1.52   0.129    -.0057015    .0449852
    interac4 |  -.0089606   .0118065    -0.76   0.448    -.0321121     .014191
    interac5 |    .008949   .0132898     0.67   0.501    -.0171112    .0350091
    interac6 |  -.0081347   .0130533    -0.62   0.533     -.033731    .0174617
    interac7 |   .0275896   .0123698     2.23   0.026     .0033335    .0518457
    interac8 |   .0246111   .0172177     1.43   0.153    -.0091513    .0583735
       _cons |   .7058653     .02951    23.92   0.000     .6479988    .7637319
------------------------------------------------------------------------------
```

Test the hypothesis of NO regional variation: $H_0 : \lambda_i = \vartheta_i = 0 \quad i = 0,1,2,3,4,5,6,7,8$

We test this hypothesis comparing the sum of the squared residuals of the unrestricted (eq. (5)) and the restricted model (eq. (1)). We obtain $F = 26.89$ and we compare it with the 0.95 quantile of the $\chi^2_{(18)}$ distribution (28.87). Therefore we cannot reject the null hypothesis using a test with size 0.05.

3.3 - Model for Nutrition

3.3.a

Model for nutrition:

$$N = \beta_1 + \beta_2 adultm + \beta_3 adultf + \beta_4 child + \varepsilon \qquad \textbf{(6)}$$

Estimation of equation (6) when *N* is energy:

```
. reg  energy adltm adltf child

      Source |       SS       df       MS              Number of obs =    2510
-------------+------------------------------           F(  3,  2506) =  379.63
       Model |  6.6098e       3     2.2033e            Prob > F      =  0.0000
    Residual |  1.4544e    2506     5803624.86         R-squared     =  0.3125
-------------+------------------------------           Adj R-squared =  0.3116
       Total |  2.1154e    2509     8431108.29         Root MSE      =  2409.1

------------------------------------------------------------------------------
      energy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       adltm |   1753.748   84.47855    20.76   0.000     1588.093    1919.403
       adltf |   1514.622   99.55914    15.21   0.000     1319.396    1709.849
       child |   975.4449   47.56874    20.51   0.000     882.1668    1068.723
       _cons |   638.4984   136.1014     4.69   0.000     371.6156    905.3812
------------------------------------------------------------------------------
```

We are going to work with the residuals from this regression, so let us store them in a variable that we will keep in memory. Type the command:

```
. predict res_energy, res
```

now the OLS residuals from the regression above are saved in the variable *res_energy*.

```
. vce
             |    adltm    adltf    child    _cons
-------------+------------------------------------
       adltm |  7136.63
       adltf |  259.662  9912.02
       child |  -128.41 -336.399  2262.79
       _cons | -6424.98  -9966.6 -950.091  18523.6
```

We are going to use this matrix later on, so let us save it matrix as follows:

```
. matrix var_energy=e(V)
```

To view the stored matrix type:

```
. matrix list var_energy
```

and the variance/covariance matrix of the OLS estimator for the “energy” regression will be shown again.

Estimation of equation (6) when N is fat:

```
. reg fat adltm adltf child

      Source |       SS       df       MS              Number of obs =    2510
-------------+------------------------------           F(  3,  2506) =  210.32
       Model |  12640380.1     3  4213460.04           Prob > F      =  0.0000
    Residual |  50204316.5  2506  20033.6459           R-squared     =  0.2011
-------------+------------------------------           Adj R-squared =  0.2002
       Total |  62844696.7  2509  25047.7069           Root MSE      =  141.54

------------------------------------------------------------------------------
         fat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       adltm |   80.37673   4.963372    16.19   0.000       70.644    90.10946
       adltf |   71.65502   5.849402    12.25   0.000     60.18486    83.12517
       child |    38.5584   2.794808    13.80   0.000     33.07803    44.03877
       _cons |   26.03282   7.996373     3.26   0.001     10.35264    41.71299
------------------------------------------------------------------------------
```

We are going to work with the residuals from this regression, so let us store them in a variable that we will keep in memory. Type the command:

```
. predict res_fat, res
```

```
. vce

             |    adltm    adltf    child    _cons
-------------+------------------------------------
       adltm |  24.6351
       adltf |  .896333  34.2155
       child | -.443262 -1.16122  7.81095
       _cons | -22.1785 -34.4039 -3.27964   63.942
```

In order to save this matrix run the command:

```
. matrix var_fat=e(V)
```

To view the stored matrix type:

```
. matrix list var_fat
```

and the variance/covariance matrix of the OLS estimator for the “fat” regression will be shown again.

In order to test the null hypothesis $H_0 : \beta_2 - \beta_3 = 0$ for the two models we use the statistics

$$T = \frac{\hat{\beta}_2 - \hat{\beta}_3}{s\left[c'\left(X'X\right)^{-1}c\right]^{1/2}} \overset{a}{\sim} N(0,1)$$

with $c' = \begin{pmatrix} 0 & 1 & -1 & 0 \end{pmatrix}$.

We obtain:

	T	0.975 quantile	0.95 quantile
N=energy	1.8596	1.96	1.64
N=fat	1.1546	1.96	1.64

Comments:

1. When N is energy, we don't reject the null using a 5% size, but we do using a 10% size.
2. When N is fat, we do not reject the null at both sizes.
3. The intercept can be interpreted as the average nutrient consumption by members not included in adults and children (misreport) or by pets.

3.3.b

Recall from the exercise that fat converts to energy at 9Kcal/g, therefore, before taking ratios of the estimated coefficients, we have to multiply by 9 the estimated coefficients from the fat regression, in order to express the estimated coefficients from the 2 regressions in the same unit. If we do that, we get:

	N=energy	N=fat (times 9)	Ratio: Fat/Energy
$\hat{\beta}_2$	1753	720	0.41
$\hat{\beta}_3$	1514	648	0.43
$\hat{\beta}_4$	975	351	0.36

The ratios in the last column represent the estimated proportion of energy from fat for each of the three types of person. In order to obtain the standard errors of the ratios, we use the "delta method".

Recall that if the approximate distribution of a consistent estimator $\hat{\theta}$ of θ is given by

$$\sqrt{n}\left(\hat{\theta}-\theta\right)\xrightarrow{d}N\left(0,\Omega\right)$$

then the approximate distribution of $h\left(\hat{\theta}\right)$, with $h(\bullet)$ representing a function of the vector of parameters, is

$$\sqrt{n}\left(h\left(\hat{\theta}\right)-h\left(\theta\right)\right)\xrightarrow{d}N\left(0,\Gamma\left(\theta\right)\Omega\Gamma'\left(\theta\right)\right),$$

where $\Gamma=\dfrac{\partial h(\theta)}{\partial\theta'}$.

In our case the vector θ is the vector of parameters from the two regressions (note that the coefficients appear in the same order as in STATA), one using energy as dependent variable and one using fat, namely:

$$\theta'=\left(\beta_{2E}\quad\beta_{3E}\quad\beta_{4E}\quad\beta_{1E}\quad\beta_{2F}\quad\beta_{3F}\quad\beta_{4F}\quad\beta_{1F}\right)=\left(\beta_{E}\quad\beta_{F}\right)$$

The first step is now obtaining an estimate of the variance/covariance matrix Ω.

We know that the variances of the OLS estimators for the energy and fat regressions are

$$Var\left(\hat{\beta}_E\,|X\right)=\sigma_{EE}\left(X'X\right)^{-1}\quad\text{and}\quad Var\left(\hat{\beta}_F\,|X\right)=\sigma_{FF}\left(X'X\right)^{-1}$$

Therefore

$$\Omega=Var\begin{pmatrix}\hat{\beta}_E\\ \hat{\beta}_F\end{pmatrix}|X\Bigg)=\begin{bmatrix}\sigma_{EE}\left(X'X\right)^{-1} & \sigma_{EF}\left(X'X\right)^{-1}\\ \sigma_{EF}\left(X'X\right)^{-1} & \sigma_{FF}\left(X'X\right)^{-1}\end{bmatrix}$$

where $\sigma_{EE}=Var(\varepsilon_E\,|X)$, $\sigma_{FF}=Var(\varepsilon_F\,|X)$ and $\sigma_{EF}=Cov(\varepsilon_E,\varepsilon_F\,|X)$.

Notice that estimates of $\sigma_{EE}\left(X'X\right)^{-1}$ and $\sigma_{FF}\left(X'X\right)^{-1}$ are represented by the two matrices reported above using the command *vce* after the regressions for energy and fat respectively and now stored in memory as *var_energy* and *var_fat*.

Instead, $\sigma_{EF}\left(X'X\right)^{-1}$ is unknown. More precisely, we do not know $\sigma_{EF}=Cov(\varepsilon_E,\varepsilon_F\,|X)$. However, σ_{EF} can be estimated - via analogue principle - by the mean of the cross products of the residuals from the two regressions. In order to do that in STATA, we follow these steps:

```
. gen cross_res=res_energy*res_fat
. egen av_cross_res=mean(cross_res)
. disp av_cross_res
297307.28
```

So we get $\hat{\sigma}_{EF}=297307.28$.

We can recover $(X'X)^{-1}$ from either $Var(\hat{\beta}_E|X)=\sigma_{EE}(X'X)^{-1}$ or $Var(\hat{\beta}_F|X)=\sigma_{FF}(X'X)^{-1}$.
For example, after the regression using energy as the dependent variable, we have the estimated $Var(\hat{\beta}_E|X)=\sigma_{EE}(X'X)^{-1}$ (stored as *var_energy*) and an estimate of $\sigma_{EE}=Var(\varepsilon_E|X)$, namely $\hat{\sigma}_{EE}=5803624.86$. We can easily recover $(X'X)^{-1}$ dividing the matrix *var_energy* by $\hat{\sigma}_{EE}=5803624.86$:

```
. matrix inv_xprimex=var_energy/5803624.86
. matrix list inv_xprimex
```

$$(X'X)^{-1}=\begin{bmatrix} 0.0012 & 0.0000 & -0.0000 & -0.0011 \\ 0.0000 & 0.0017 & -0.0001 & -0.0017 \\ -0.0000 & -0.0001 & 0.0004 & -0.0002 \\ -0.0011 & -0.0017 & -0.0002 & 0.0032 \end{bmatrix}$$

(the same result is obtained if we divide *var_fat* by $\hat{\sigma}_{FF}=20033.6459$).

To obtain $\hat{\sigma}_{EF}(X'X)^{-1}$, type:

```
. matrix var_energy_fat=inv_xprimex*297307.28
. matrix list var_energy_fat
```

$$\hat{\sigma}_{EF}(X'X)^{-1}=\begin{bmatrix} 365.5947 & 13.3019 & -6.5782 & -329.1378 \\ 13.3019 & 507.7716 & -172330 & -510.5676 \\ -6.5782 & -17.2330 & 115.9176 & -48.6712 \\ -329.1378 & -510.5676 & -48.6712 & 948.9247 \end{bmatrix}$$

An estimate of Ω is obtained using:

```
. matrix omega=[var_energy,var_energy_fat\var_energy_fat,var_fat]
. matrix list omega
```

$$\hat{\Omega}=\begin{bmatrix} 7136.63 & & & & 365.5947 & & & \\ 259.662 & 9912.02 & & & 13.3019 & 507.7716 & & \\ -128.41 & -336.399 & 2262.79 & & -6.5782 & -17.233 & 115.9176 & \\ -6424.98 & -9966.6 & -950.091 & 18523.6 & -329.1378 & -510.5676 & -48.6712 & 948.9247 \\ 365.5947 & & & & 24.6351 & & & \\ 13.3019 & 507.7716 & & & 0.89633 & 34.2155 & & \\ -6.5782 & -17.233 & 115.9176 & & -0.443262 & -1.16122 & 7.81095 & \\ -329.1378 & -510.5676 & -48.6712 & 948.9247 & -22.1785 & -34.4039 & -3.2764 & 63.942 \end{bmatrix}$$

The last term that we need to know in order to compute the standard errors of our ratios is the matrix Γ. Recalling that

$$\theta' = \begin{pmatrix} \beta_{2E} & \beta_{3E} & \beta_{4E} & \beta_{1E} & \beta_{2F} & \beta_{3F} & \beta_{4F} & \beta_{1F} \end{pmatrix}$$

and given that

$$h(\theta) = \begin{pmatrix} 9\beta_{2F}/\beta_{2E} & 9\beta_{3F}/\beta_{3E} & 9\beta_{4F}/\beta_{4E} \end{pmatrix}',$$

we have:

$$\Gamma(\theta) = \begin{bmatrix} \frac{\partial(9\beta_{2F}/\beta_{2E})}{\partial\beta_{2E}} & \frac{\partial(9\beta_{2F}/\beta_{2E})}{\partial\beta_{3E}} & \cdots & \frac{\partial(9\beta_{2F}/\beta_{2E})}{\partial\beta_{1F}} \\ \frac{\partial(9\beta_{3F}/\beta_{3E})}{\partial\beta_{2E}} & \frac{\partial(9\beta_{3F}/\beta_{3E})}{\partial\beta_{3E}} & \cdots & \frac{\partial(9\beta_{3F}/\beta_{3E})}{\partial\beta_{1F}} \\ \frac{\partial(9\beta_{4F}/\beta_{4E})}{\partial\beta_{2E}} & \frac{\partial(9\beta_{4F}/\beta_{4E})}{\partial\beta_{3E}} & \cdots & \frac{\partial(9\beta_{4F}/\beta_{4E})}{\partial\beta_{1F}} \end{bmatrix}$$

$$= \begin{bmatrix} -9\beta_{2F}/\beta_{2E}^2 & 0 & 0 & 0 & 9/\beta_{2E} & 0 & 0 & 0 \\ 0 & -9\beta_{3F}/\beta_{3E}^2 & 0 & 0 & 0 & 9/\beta_{3E} & 0 & 0 \\ 0 & 0 & -9\beta_{4F}/\beta_{4E}^2 & 0 & 0 & 0 & 9/\beta_{4E} & 0 \end{bmatrix}$$

Evaluating this matrix in $\hat{\theta} = \begin{pmatrix} \hat{\beta}_E & \hat{\beta}_F \end{pmatrix}$ (estimated coefficients from the two previous regressions), we get:

$$\Gamma(\hat{\theta}) = \begin{bmatrix} -0.000235 & 0 & 0 & 0 & 0.00513 & 0 & 0 & 0 \\ 0 & -0.000281 & 0 & 0 & 0 & 0.00594 & 0 & 0 \\ 0 & 0 & -0.000365 & 0 & 0 & 0 & 0.009225 & 0 \end{bmatrix}$$

To define this matrix in STATA, use the command:

```
. matrix gamma=[-0.000235,0,0,0,0.00513,0,0,0\
                0,-0.000281,0,0,0,0.00594,0,0\0,0,-0.000365,0,0,0,0.009225,0]

. matrix list gamma
```

Finally, the estimated variance/covariance matrix for the ratios is

$$\Gamma\left(\hat{\theta}\right)\hat{\Omega}\Gamma'\left(\hat{\theta}\right)=\begin{bmatrix} 0.000161 & 0.00000 & -0.00000 \\ 0.00000 & 0.000295 & -0.00002 \\ -0.00000 & -0.00002 & 0.000186 \end{bmatrix}$$

To compute this matrix in STATA, type:

```
. matrix var_ratios=gamma*omega*gamma'

. matrix list var_ratios
```

Taking the square root of the elements on the mail diagonal, we obtain the estimated standard errors for the estimated ratios:

	Estimated Ratios	Estimated Standard Errors
Adult males	0.41	0.013
Adult females	0.43	0.017
Children	0.36	0.014

Comments:

1. The estimated proportion of energy from fat is very similar for males and females.
2. The estimated proportion for children is lower.
3. The recommendation is that this proportion should not exceed 35% except for young people: it seems that adults do not follow closely this recommendation.

3.3.c (optional)
See “official” solution if you are interested in.