# University College London
# Department of Economics

## G023: Econometric Theory and Methods*
## Answers to Exercise 3

1. *Prepare and examine the data.*

   (a) The survey is designed to be representative of Great Britain. However because of non-response it tends to have an over representation of older people and households without children. Average household size is 1.29 persons per household. The means of the binary variables tell us the proportion of the sample possessing the indicated characteristic, for example, 80% of households have microwave ovens, 94% have freezers. There is a lot of variation in recorded income and in food expenditures. One relatively low income household spent nearly twice on food what they recorded as income. Maybe they had a party during the recording week. Some households spent nothing on food during the recording week. Maybe they were on vacation most of the time, ate entirely from food stocks, or did not eat at home.

   (b) There is a great deal of dispersion in this graph, reflecting the wide range of per capita family incomes and food expenditures. The dispersion in food expenditures per capita seems to increase as we go to higher levels of family income. This looks like heteroskedastic not homoskedastic variation. It is clear that food expenditures per head are on average higher in households with higher income per head, but the slope of the relationship becomes close to zero at higher levels of household income per head.

   (c) Food share is clearly lower on average in households with higher log per capita family income. The relationship looks as if it may be linear over a quite wide range of values, but rather flat at higher income levels and steeper at lower income levels. The variation in the graph appears to be heteroskedastic with *higher* variance in food share at *lower* income levels.

*

2. *Investigate and augment the Working-Leser specification of the food expenditure Engel curve.*

   (a) Here are the OLS estimates

   | Coefficient | Estimate | Est. std. err. |
   |---|---|---|
   | $\beta_1$ | 0.673 | 0.016 |
   | $\beta_2$ | -0.105 | 0.0033 |

   The approximate 95% confidence interval is $[-0.111, -0.099]$. Note that the heteroskedasticity evident in the scatter plot you drew for part (b) of Question 1 suggests that the homoskedasticity assumption underlying the standard error calculation is unlikely to hold. One way too proceed in the light of this would be to specify a model for this heteroskedastic variation, estimate it, then use a GLS estimator and the associated standard errors. Here you are asked to compute heteroskedasticity robust standard errors. They tend to be significantly larger than conventional standard errors leading to larger confidence intervals. For example instead of 0.0033 above there is 0.0053. There is little effect on hypothesis test outcomes at the test sizes used in this question.

   (b) Here are the OLS estimates.

   | Coefficient | Estimate | Est. std. err. |
   |---|---|---|
   | $\gamma_1$ | 0.761 | 0.017 |
   | $\gamma_2$ | -0.115 | 0.0033 |
   | $\gamma_3$ | 0.068 | 0.0045 |

   $\hat{\gamma}_2 + \hat{\gamma}_3 = -0.047$ which is $c'\hat{\gamma}$ where $c' = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix}$. We obtain $s\left(c'(X'X)^{-1}c\right)^{1/2} = 0.0040$ and

   $$\frac{\hat{\gamma}_2 + \hat{\gamma}_3}{s\left(c'(X'X)^{-1}c\right)^{1/2}} = -11.71.$$

   Under the null hypothesis $H_0 : \gamma_2 = -\gamma_3$ and this as a realisation of a random variable with approximately an $N(0,1)$ distribution, so to conduct an approximate size .05 test we compare with $-1.96$ and reject the null hypothesis.

   Comparing the sum of squared residuals obtained with and without the restriction imposed we have $\hat{\varepsilon}'_R\hat{\varepsilon}_R = 27.62$, $\hat{\varepsilon}'_U\hat{\varepsilon}_U = 26.19$ and the test statistic

   $$\frac{(\hat{\varepsilon}'_R\hat{\varepsilon}_R - \hat{\varepsilon}'_U\hat{\varepsilon}_U)}{\hat{\varepsilon}'_U\hat{\varepsilon}_U/n} = 137.12$$

   where $n = 2510$ is the number of observations. We compare this with the 0.95 quantile of a $\chi^2_{(1)}$random variable which is 3.84 ( which is $(1.96)^2$)and reject the null hypothesis. Note that the statistic just obtained is exactly the square of the one obtained earlier.

   (c) The estimated coefficient on $lfincpc^2$ is 0.0334 with an estimated standard error of 0.00349 giving a Wald statistic of 9.55 which exceeds 1.96 so we reject the hypothesis. This positive coefficient indicates a convex relationship as suggested by the scatter plot in part (b) of Question (1). The cubed term has an estimated coefficient of $-0.0099$ with an estimated standard error of 0.00297 giving a Wald statistic of $-3.35$ and again we reject the hypothesis. If we draw the fitted cubic equation we find that the cubic term moderates the convexity slightly but does not remove it. Doing a joint test comparing sums of squared residuals with and without the squared and cubic terms, gives a statistic

$$S = \frac{(27.62439 - 26.53527)}{26.53527/2510} = 103.02$$

which, for an approximate size 0.05 test, we compare with the 0.95 quantile of a $\chi^2_{(2)}$ random variable, namely 5.99. We reject the joint null hypothesis.

(d) There might be regional variations in prices of foods, or regional variations in tastes. But, the coefficients on the region indicators are all very similar. Omitting the intercept, the smallest (region 5) is 0.665 and the largest (region 9) is 0.688, a difference of just 0.023, and all the estimated standard errors are around 0.017. There seems no evidence here pointing to practically significant regional variations in the relationship between food expenditure and household income.

When you try the different styes of estimation you will find that the estimated coefficient on the region indicator for say region 1, when no intercept is included, say $\hat{\delta}_1$, is exactly equal to the estimated intercept when that is included, $\hat{\beta}_0$, plus the estimated coefficient on the region one indicator in the intercept included estimation. If region 1 were the region you excluded to allow an intercept to be estimated then $\hat{\beta}_0 = \hat{\delta}_1$ and another region indicator, say for region $i$, has an estimated coefficient equal to $\hat{\delta}_i - \hat{\delta}_1$. Do the algebra to show that this must happen. The estimated standard errors vary across estimations in accordance with the rules for determining variances of linear functions of estimators. Check this too, using algebra and using the estimated variance of the OLS estimators.

(e) Comparing the sum of squared residuals from 10 region specific estimations with the sum of squared residuals from the estimation in part (a) gives the following test statistic.

$$S = \frac{(27.62439 - 27.33151)}{27.33151/2510} = 26.89$$

which, for an approximate size 0.05 test, we compare with the 0.95 quantile of a $\chi^2_{(18)}$ random variable, namely 28.87. The null hypothesis of equality of slope and intercept coefficients across regions cannot be rejected.

3. *Nutrition.*

(a) $\beta_1$ could represent the average rate of nutrient consumption by consumers not counted in the counts of adults and children. These could be omitted people or maybe pets. However if consumption by such missing consumers is correlated with the number of counted people then we would expect the estimated coefficients on the observed counts to pick up some of the nutrient intake of the uncounted.

Here are the OLS estimates.

| | Energy (kcal/person/day) | | Fat (g/person/day) | |
|---|---|---|---|---|
| Coefficient | Estimate | Est. std. err. | Estimate | Est. std. err. |
| $\beta_1$ | 638 | 136 | 26 | 8.0 |
| $\beta_2$ | 1753 | 84 | 80 | 5.0 |
| $\beta_3$ | 1514 | 100 | 72 | 5.8 |
| $\beta_4$ | 975 | 48 | 39 | 2.8 |

The estimated average intake rates are of plausible orders of magnitude given that this data tells us only about food eaten at home and excludes alcohol and probably under-records confectionery and snack foods. The rather large value for $\hat{\beta}_1$ probably arises because of an element of misspecification. A finer disaggregation of people by ages (you do not have data on this) produces a smaller value for $\beta_1$, of the order of 200 kcal/day.

Conducting size 0.05 tests we do not reject the null of equality of male and female average intake rates for energy or fat. Conducting size 0.10 tests we reject equality for energy but not for fat.

(b) Here you have to multiply the fat intake coefficients by 9 to convert to kcal/person/day before taking ratios of coefficients. The results are shown below.

| | Prop of energy from fat | |
|---|---|---|
| Type | Estimate | Est. std. err. |
| Adult males | 0.41 | 0.013 |
| Adult females | 0.43 | 0.017 |
| Children | 0.36 | 0.014 |

The estimated proportion is very similar for adult males and females, but smaller for children. Only for children is the Government recommendation approached. Actually that recommendation only applies to adults!

To calculate an estimated standard error for the estimated proportion of energy from fat for adult males you can use the delta method, covered in the notes. For this problem you need to take into account the covariance of the estimated coefficients on number of adult males in two separate regression estimations. We have, with $y_E$ denoting energy intakes and $y_F$ denoting fat intakes,

$$\begin{aligned} y_E &= X\beta_E + \varepsilon_E \\ y_F &= X\beta_F + \varepsilon_F \end{aligned}$$

and estimates

$$\begin{aligned} \hat{\beta}_E &= (X'X)^{-1}X'y_E \\ \hat{\beta}_F &= (X'X)^{-1}X'y_F \end{aligned}$$

whose variance matrix is

$$Var\left[\begin{array}{c}\hat{\beta}_E\\ \hat{\beta}_F\end{array}\right]=\left[\begin{array}{cc}\sigma_{EE}(X'X)^{-1} & \sigma_{EF}(X'X)^{-1}\\ \sigma_{EF}(X'X)^{-1} & \sigma_{FF}(X'X)^{-1}\end{array}\right]$$

where $\sigma_{EF}=Cov(\varepsilon_i^E,\varepsilon_i^F|X)$. This can be estimated using the mean of the cross products of the residuals from the two fitted equations. Then it is straightforward to obtain the covariance of the two estimated coefficients.

The formula for the approximate variance of the ratio of two estimators is given at the end of the lecture notes on "Approximate Inference". Applying this gives the entries in the table above.

There is one point to note. To get the proportion of energy from fat for adult males you have to multiply the estimated coefficient on number of adult males in the fat equation by 9 and then divide by the estimated coefficient on number of adult males in the energy equation.. Because of this multiplication, when you apply the formula from the lecture notes you have to multiply the estimated variance of the fat coefficient estimate by $9^2=81$ and the estimated covariance of the fat and energy adult male coefficient estimates by 9. Why?

(c) The nonlinear least squares estimates and the associated estimated standard errors are shown in the table below.

| | Energy (kcal/person/day) | | Fat (g/person/day) | |
|---|---|---|---|---|
| Coefficient | Estimate | Est. std. err. | Estimate | Est. std. err. |
| $\beta_1$ | 812 | 188 | 33 | 10.8 |
| $\beta_2$ | 2209 | 245 | 100 | 14.5 |
| $\beta_3$ | 1870 | 212 | 88 | 12.8 |
| $\beta_4$ | 1122 | 90 | 45 | 4.8 |
| $\alpha$ | -0.044 | 0.019 | -0.042 | 0.026 |

The estimated coefficients on income per head are negative and significantly different from zero using a size 0.05 test. However the coefficients are very small. The negative relationship with income may arise because higher income households eat out more and so spend less on food at home. The variation in the coefficients across adult males, females and children is similar to that found when we fitted a linear model. For example, for energy the ratio of the estimated coefficients on adult males and females is 1.18 in the nonlinear model and 1.16 in the linear model.

Note that the magnitudes of the coefficients are quite different comparing the linear and nonlinear models. This is because in the nonlinear model the coefficients on counts of household members are estimates of intakes when log income per head is zero, i.e. income per head is 1. The average value of log income per head in the sample is 4.84, so to get a better comparison we can multiply the estimated coefficients, e.g., for energy, in the nonlinear model by $\exp(-0.044\times 4.84)=0.81$. This gives values of 1789, 1514 and 908, for respectively adult males, females and children, much closer to the values obtained in the linear model.

The estimated income coefficients in the energy and fat models are very similar, suggesting (why?) that the percentage of energy from fat is insensitive to income. The multiplicative form used here assumes that, as income increases all members of the household experience the same proportionate change in nutrient intakes, which seems reasonable at least as a first order approximation. It might be the case that e.g., males' intakes change faster with income then females', but this would still suggest a multiplicative type model but with interaction terms involving products of counts of household members of each type and functions of income.