CONSUMER EXPENDITURE SURVEY

REGRESSION EXERCISES

c.dougherty@lse.ac.uk
January 2002

TABLE OF CONTENTS

DOWNLOADING THE DATA SET

You will find the Consumer Expenditure Survey data sets in three different formats at the Introduction to Econometrics web site: Stata format for Stata users, EViews format for EViews users, and Excel for everybody else. Nearly all regression packages are able to import Excel files. To download, click on the filename and follow the instructions in the dialogue box. Note that you may wish to download to a directory other than that specified as the default in the dialogue box. If you are using the Stata file, the file will be called CES.DTA and it will be ready for use. If you are using the EViews file, it may download as CES_WF1.BIN. It should have extension WF1, not BIN, so rename it as CES.WF1. To do this, go to My Computer, browse until you find the downloaded file, click on File and then Rename. You will now be able to delete the .bin extension and to replace _WF1 with .WF1.

1. DESCRIPTION OF THE DATA SET

The intention is that the exercises should provide material for practical work for a small group of students working in parallel, each student fitting expenditure functions with a different category of expenditure.

The data set has been derived from the Quarterly Interview Survey of the Consumer Expenditure Survey (henceforward CES) undertaken by the U.S. Department of Labor, Bureau of Labor Statistics. This nationally representative survey has a sample of about 5,000 households, each household being interviewed five times, the first time to gather basic data about the household, and the four other times at quarterly intervals to gather data on expenditures. The households in the present data set entered the quarterly survey at the beginning of 1995 and the data give the total expenditure by category over the 1995 calendar year. The variables are as follows:

Household characteristics

SIZE Number of persons in the household.
SIZEAM Number of adult males (males older than 15) in the household.
SIZEAF Number of adult females (females older than 15) in the household.
SIZEJM Number of junior males (males aged 2 through 15) in the household.
SIZEJF Number of junior females (females aged 2 through 15) in the household.
SIZEIN Number of children aged less than 2 in the household.
REFAGE Age of the reference person in the household (the individual who owns or rents the dwelling).
REFEDUC Education of the reference person, coded as

0 Never went to school
1 Elementary school only(1-8 years)
2 Some high school, but did not graduate
3 High school graduate, no college
4 Some college, but did not graduate
5 College graduate
6 Graduate school

REFRACE Ethnicity of the reference person, coded as

1 White
2 Black
3 American Indian, Aleut, Eskimo
4 Asian or Pacific Islander
5 Other

HHTENURE Household tenure, coded as

1 Owned with mortgage
2 Owned without mortgage
3 Owned, mortgage not reported
4 Rented
5 Occupied without payment of cash rent
6 Student housing

Expenditure variables

EXP	Total household expenditure, including some items not listed as variables below.
FDHO	Food and nonalcoholic beverages consumed at home.
FDAW	Food and nonalcoholic beverages consumed away from home, excluding meals as pay in kind
SHEL	Housing, excluding expenditure on utilities, household operations, and household equipment. In the case of owned dwellings it comprises mortgage interest, property taxes, and the cost of maintenance, repairs, and insurance. In the case of rented dwellings, it consists of rent, including rent as pay in kind. *SHEL* also includes the recurrent costs of vacation houses, expenditure on lodging away from home, and the cost of school housing. Note that this category of expenditure does *not* include purchases of dwellings.
TELE	Telephone services.
DOM	Domestic services, such as condo housekeeping and management, gardening, and babysitting and child day care.
TEXT	Household textiles such as bathroom, bedroom, kitchen and dining room linens, curtains and cushions.
FURN	Furniture.
MAPP	Major household appliances, such as dishwashers, refrigerators, clothes washers, stoves and ovens, air conditioners, floor cleaning machines and sewing machines.
SAPP	Small appliances and miscellaneous housewares.
CLOT	Clothing
FOOT	Footwear
GASO	Gasoline and motor oil
TRIP	Public transportation on out-of-town trips.
LOCT	Local public transportation.
HEAL	Health care, comprising health insurance, medical services, prescription drugs, and medical supplies.
ENT	Entertainment, comprising fees and admissions, televisions, radios, and sound equipment, pets, toys, and playground equipment, and other related equipment and services.
FEES	Membership fees of recreational and health clubs, fees for participant sports, admission fees for movies, theatre, concerts, opera, and sporting events, and fees for recreational instruction.
TOYS	Toys, games, hobbies, playground equipment, and pets, including veterinarian expenses.
READ	Reading.matter, such as newspapers, magazines, and books
EDUC	Education, such as tuition fees, school books, supplies and equipment for elementary school, high school, and college, and other types of school.
TOB	Tobacco products and supplies such as cigarettes, cigars, and pipe tobacco..

All the expenditure variables are measured in current dollars.

2. EXERCISES

You should choose, or be assigned by your instructor, one of the categories of expenditure listed above. In the exercises below you will develop a regression specification for this category, starting with a simple regression model and gradually improving it. For ease of exposition, the exercises refer to a non-existent category called *CAT*. Follow the instructions, replacing *CAT* with the name of your category.

In all of the regressions you should include a constant. Most regression packages automatically assume that a constant is included, unless specifically indicated otherwise. However some, including EViews and its TSP predecessors, require you to specify a constant if you wish to include one.

If you are working with just one category of expenditure, it may be helpful to simplify the data set by deleting the expenditure variables relating to the other categories.

Exercise 1 Simple regression analysis

Is expenditure on your category related to total expenditure?

Regress *CAT* on *EXP*, total household expenditure in 1995. Give an economic interpretation of the regression coefficients and perform *t* tests where appropriate.

Exercise 2 Multiple regression

Is expenditure on your commodity related to household size as well as total household expenditure?

Obviously larger households will tend to spend more, but in principle the household size effect might be picked up to some extent by total household expenditure, which is also larger for larger households. The purpose of this exercise is to see whether size has a separate effect, controlling for total expenditure. Regress *CAT* on *EXP* and *SIZE*. Give an economic interpretation of the regression coefficients and perform appropriate statistical tests.

Exercise 3 Multiple regression

Is expenditure per capita on your commodity related to household size as well as total household expenditure per capita?

This exercise is parallel to the previous one but uses an alternative specification that is in some ways more satisfactory. Generate two new variables *CATPC* = *CAT*/*SIZE* and *EXPPC* = *EXP*/*SIZE* and regress *CATPC* on *EXPPC* and *SIZE*. Give an economic interpretation of the regression coefficients and perform appropriate statistical tests.

Exercise 4 Multiple regression

Is household composition a determinant of expenditure per capita on your commodity, controlling for household expenditure per capita?

It is reasonable to suppose that household composition may affect expenditure on some categories of expenditure since adults may spend more than children on some categories and less on others. The purpose of this exercise is to see whether size has a separate effect, controlling for total expenditure.

Regress *CATPC* on *EXPPC* and *SIZEAM*, *SIZEAF*, *SIZEJM*, *SIZEJF*, and *SIZEIN*. Give an economic interpretation of the regression coefficients and perform appropriate statistical tests. Comparing this regression with that of *CATPC* on *EXPPC* and *SIZE*, is there evidence that this refinement leads to improved explanatory power?

Exercise 5 Nonlinear regression analysis

Is expenditure on your category related to total expenditure and household size? An alternative model specification

Define a new variable *LGCAT* as the (natural) logarithm of expenditure on your category. Define a new variable *LGEXP* as the logarithm of total household expenditure. Also define a new variable *LGSIZE* as the logarithm of household size. Regress *LGCAT* on *LGEXP* and *LGSIZE*. (Just to reassure anyone who has any doubt: you are not really constructing a variable called *LGCAT*, but the equivalent variable for your commodity. For example, if your commodity is clothing, you should construct *LGCLOT* as the natural logarithm of *CLOT*, and you should regress *LGCLOT* on *LGEXP* and *LGSIZE*.)

Exercise 6 Nonlinear regression analysis

Is expenditure on your category per capita related to total expenditure per capita? An alternative model specification

Define a new variable *LGCATPC* as the logarithm of expenditure per capita on your category. Define a new variable *LGEXPPC* as the logarithm of total household expenditure per capita. Regress *LGCATPC* on *LGEXPPC*, provide an interpretation of the coefficients, and perform appropriate statistical tests. Explain how this specification is related to that in the previous exercise.

Exercise 7 Nonlinear regression analysis

Is expenditure on your category per capita related to household size as well as to total expenditure per capita? An alternative model specification

Regress *LGCATPC* on *LGEXPPC* and *LGSIZE*, provide an interpretation of the coefficients, and perform appropriate statistical tests. Explain how this specification is related to that in the two previous exercises.

Exercise 8 A Box-Cox test

Is a logarithmic specification preferable to a linear specification for an expenditure function?

Define *CATPCST* as *CATPC* scaled by its geometric mean *LGCATST* as the logarithm of *CATPCST*. Regress *CATPCST* on *EXPPC* and *SIZE* and regress *LGCATST* on *LGEXPPC* and *LGSIZE*. Use a Box-Cox test to see if there is a significant difference in the goodness of fit of these equations.

Exercise 9 Use of a dummy variable

Does ethnicity have an effect on household expenditure?

The variable *REFRACE* in the data set is coded 1 if the reference individual in the household, usually the head of the household, is white and it is coded greater than 1 for other ethnicities. Define a dummy variable *NONWHITE* that is 0 if *REFRACE* is 1 and 1 if *REFRACE* is greater than 1. Regress *LGCATPC* on *LGEXPPC*, *LGSIZE*, and *NONWHITE*, provide an interpretation of the coefficients, and perform appropriate statistical tests.

Exercise 10 Use of a dummy variable with multiple categories

Does education have an effect on household expenditure?

The variable *REFRACE* in the data set is coded 1 if the reference individual in the household, usually the head of the household, is white and it is coded greater than 1 for other ethnicities. Define a dummy variable *NONWHITE* that is 0 if *REFRACE* is 1 and 1 if *REFRACE* is greater than 1. Regress *LGCATPC* on *LGEXPPC*, *LGSIZE*, and *NONWHITE*, provide an interpretation of the coefficients, and perform appropriate statistical tests.

Exercise 11 *F* test of the explanatory power of a group of dummy variables

Evaluate whether the ethnicity dummies as a group have significant explanatory power for educational attainment by comparing the residual sums of squares in the regressions in Exercises 7 and 10.

Exercise 12 Evaluation of the effect of changing the omitted category in a regression with dummy variables

Repeat Exercise 10 making *EDUCDO* the reference (omitted) category. Evaluate the impact on the interpretation of the coefficients and the statistical tests.

Exercise 13 Chow test

Does ethnicity have an effect on household expenditure?

The variable *REFRACE* in the data set is coded 1 if the reference individual in the household, usually the head of the household, is white and it is coded greater than 1 for other ethnicities. Define a dummy variable *NONWHITE* that is 0 if *REFRACE* is 1 and 1 if *REFRACE* is greater than 1. Regress *LGCATPC* on *LGEXPPC*, *LGSIZE*, and *NONWHITE*, provide an interpretation of the coefficients, and perform appropriate statistical tests.

Exercise 14 Equivalence of a Chow test and an *F* test on a complete set of dummy variables

Does ethnicity have an effect on household expenditure?

The variable *REFRACE* in the data set is coded 1 if the reference individual in the household, usually the head of the household, is white and it is coded greater than 1 for other ethnicities. Define a dummy variable *NONWHITE* that is 0 if *REFRACE* is 1 and 1 if *REFRACE* is greater than 1. Regress

LGCATPC on *LGEXPPC*, *LGSIZE*, and *NONWHITE*, provide an interpretation of the coefficients, and perform appropriate statistical tests.

Exercise 15 Omitted variable bias

Does the omission of total household expenditure or household size give rise to omitted variable bias?

Regress *LGCATPC* (1) on both *LGEXPPC* and *LGSIZE*, (2) on *LGEXPPC* only, and (3) on *LGSIZE* only. Assuming that (1) is the correct specification, analyze the likely direction of the bias in the estimate of the coefficient of *LGEXPPC* in (2) and that of *LGSIZE* in (3). Check whether the regression results are consistent with your analysis.

Exercise 16 *F* test of a restriction

Is expenditure per capita on your category related to total household expenditure per capita?

The model specified in Exercise 6 is a restricted version of that in Exercise 5. Perform an *F* test of the restriction.

Exercise 17 *t* test of a restriction

Is expenditure per capita on your category related to total household expenditure per capita?

The model specified in Exercise 6 is a restricted version of that in Exercise 5. Perform a *t* test of the restriction and check that you obtain the same answer as in the previous exercise. What is the relationship between the two tests?

Exercise 18 Goldfeld-Quandt tests for heteroscedasticity

Is the disturbance term in the expenditure function heteroscedastic?

Sort the data by *EXPPC* (your instructor will have to tell you how to do this), regress *CATPC* on *RXPPC* and *SIZE*, and perform a Goldfeld–Quandt test to test for heteroscedasticity in the *EXPPC* dimension. Repeat using *LGCATPC* as the dependent variable

Exercise 19 The decision to purchase: linear probability model, logit, and probit analysis

What factors affect the decision to make a purchase of your category?

Define a new variable *CATBUY* that is equal to 1 if the household makes any purchase of your category and 0 if it makes no purchase at all. Regress *CATBUY* on *EXPPC*, *SIZE*, *REFAGE*, and *REFEDUC* using (1) the linear probability model, (2) the logit model, and (3) the probit model. Calculate the marginal effects at the mean of *EXPPC*, *SIZE*, *REFAGE*, and *REFEDUC* for the logit and probit models and compare them with the coefficients of the linear probability model.

Exercise 20 tobit and OLS regressions compared

How does the constraining of expenditure affect the regression results?

Most of the categories of expenditure in the data set have large numbers of observations with 0 expenditure. We know that for such commodities OLS will tend to yield biased estimates and that tobit analysis should be preferred. To see how much difference this refinement makes, regress *CATPC* on *EXPPC* and *SIZE* (1) using OLS and the unconstrained observations only, (2) using OLS and all 869 observations, including the 0 observations, and (3) using tobit analysis. If you have been using one of the categories wit very few 0 observations, the three regressions are bound to give very similar results, so choose another category for this exercise. If you have deleted the data on all the other categories, download the data set from the website a second time.

3. INTERPRETATION OF A LOGARITHMIC REGRESSION

Nonlinear relationships of the type

$$Y = \beta_1 X^{\beta_2}$$

are very common in economic theory, for example in demand functions and Cobb–Douglas production functions. If Y is related to X in this way, its elasticity with respect to X is constant and equal to β_2. Differentiating Y with respect to X, we have

$$\frac{dY}{dX} = \beta_2 \beta_1 x^{\beta_2 - 1} = \beta_2 \frac{Y}{X}$$

and hence the elasticity of Y with respect to X, given by the left side of the next equation, is equal to β_2.

$$\frac{\dfrac{dY}{Y}}{\dfrac{dX}{X}} = \beta_2$$

The elasticity is the proportional change in Y per proportional change in X. One way of putting this into concrete terms is to say that, if X changed by 1 percent, Y would change by β_2 per cent.

To fit equations of this type, you linearize the equation by taking the logarithms of both sides:

$$\begin{aligned} \log Y &= \log \beta_1 X^{\beta_2} \\ &= \log \beta_1 + \log X^{\beta_2} \\ &= \log \beta_1 + \beta_2 \log X \end{aligned}$$

So far we have not specified whether we are taking logarithms to base e or to base 10. We shall always use e as the base, so we shall be using what are known as "natural" logarithms. This is standard in econometrics. Purists sometimes write *ln* instead of *log*, but this is unnecessary. Nobody uses logarithms to base 10 any more. They were tabulated in the dreaded log tables that were universally employed for multiplying or dividing large numbers until the early 1970s. With the invention of the pocket calculator, they have become redundant, along with the slide rule. They are not missed.

If we write $Y' = \log Y$, $X' = \log X$, and $\beta_1' = \log \beta_1$, the equation may be re-written

$$Y' = \beta_1' + \beta_2 X'$$

and you can fit it using ordinary regression analysis. All regression packages have built-in facilities for generating new variables like Y' and X' from existing ones. The coefficient of X' will be a direct estimate of the elasticity β_2 and the intercept will be an estimate of $\log \beta_1$. To obtain an estimate of β_1, you calculate $e^{\beta_1'}$.