Chapter 5

Regression analysis of household expenditure and income

5

Background

The purpose of this chapter is to demonstrate a multivariate analysis based on the Living Costs and Food survey (LCF). The aim of the analysis is to identify key characteristics of households affecting both household income and household expenditure, using regression techniques. The analysis uses the complete LCF 2008 sample containing 5,850 responding households across Great Britain and Northern Ireland.

This chapter outlines the techniques used for the quality assurance of the modelling, as well as the methodology used. It then presents the main findings of the analysis. Tables 5.3 and 5.4 summarise the regression analyses and provide more detailed results.

This chapter uses technical language to explain the regression techniques used. Therefore this chapter, unlike the others in Family Spending, may be less suitable for readers without a statistical background.

Explanatory variables for household expenditure and income

A number of potential explanatory variables were identified within the LCF dataset for modelling household expenditure and household income. These are variables that are likely to be associated with income and expenditure and are easy to define. Table 5.1 presents these variables and distinguishes between individual characteristics of the Household Reference Person (HRP) and household characteristics.

Table 5.1

Potential key variables to explain household expenditure and income

Individual characteristics
Gender of HRP
Age of HRP
Economic activity status of HRP
Socio-economic status of HRP
Household characteristics
Number of workers in the household
Household composition
Household tenure
Government Office Region
Urban/rural location of household
Gross normal weekly household income[1]

1 Please note that the gross normal weekly household income was considered as a potential predictor for household expenditure only.

Testing the Standard Assumptions

In order to apply a valid regression model, the analysis relies on certain assumptions being met. Firstly, there must be a linear relationship between the dependent and independent variables and secondly the independent variables must be linearly independent. Thirdly, multicollinearity tests can be used to check that the variables are not highly correlated with one another. In addition, the assumption of homoscedasticity requires the errors to have a constant variance, which can otherwise distort the precision of the β coefficient. Finally, the error distribution should also be normal.

The distributions for both dependent variables were found to be positively skewed; they did not follow a normal distribution. Consequently, these variables needed to be transformed, for which a log-transformation was chosen. Figures 5.1 and 5.2 illustrate the skewed distributions of the raw data. Figures 5.3 and 5.4 present the distribution after log-transformation which shows an approximately normal distribution.

Figure 5.1

Histogram of Total Household Expenditure

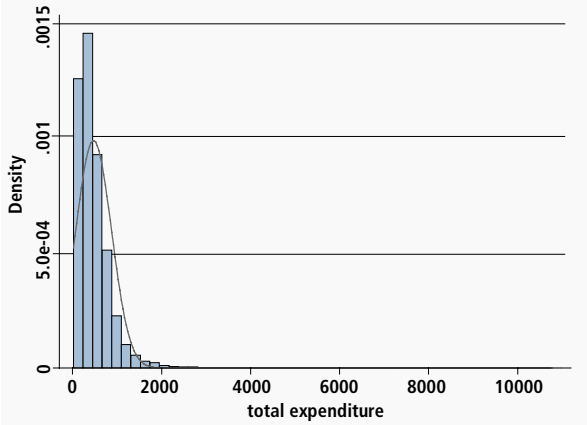


Figure 5.2

Histogram of Gross Normal Household Income

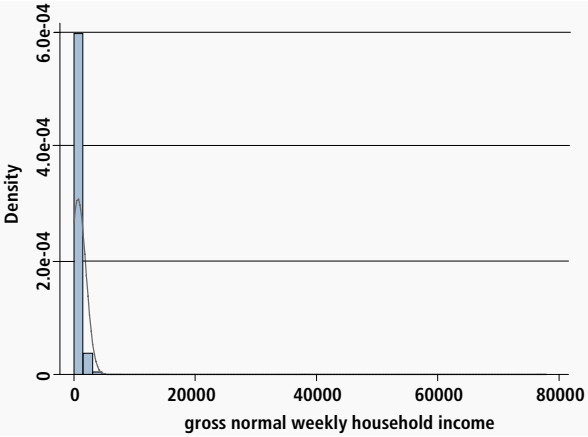


Figure 5.3

Histogram of Log-transformed Total Household Expenditure

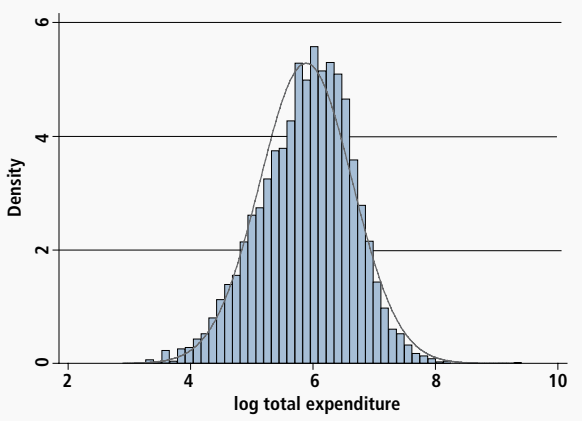
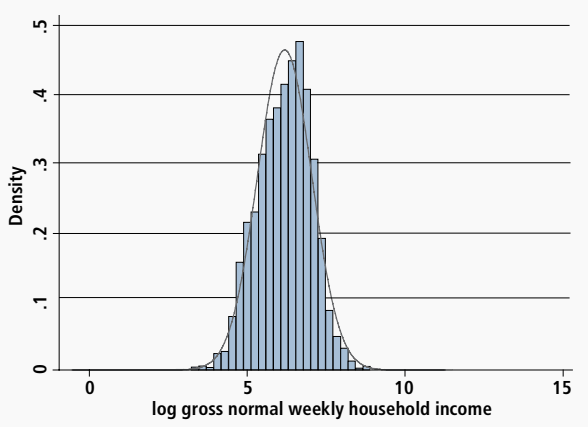


Figure 5.4

Histogram of Log-transformed Gross Normal Household Income



In order to test the linear relationship assumption, plots of the residuals versus predicted values of the model were run. These plots showed that the points were distributed around the diagonal which meant that the second assumption was held.

A multicollinearity test on the LCF dataset revealed that economic status of the HRP and socio-economic status were highly correlated. Also, the number of workers in the household was correlated with household composition and the government office regions were correlated with the urban/rural identifier. Different models were investigated using combinations of these variables and it was found that socio-economic status, household composition and the urban/ rural indicator produced the best fit. Therefore, economic activity status, number of workers and government office regions were excluded from the model.

The LCF sample is likely to include a marginal proportion of households reporting household expenditure and/or income figures that are large enough to be considered outliers. Outliers can have an effect on the assumption of normality and also on the regression slope if the data point is influential. Outliers were detected by using standardised z-scores, which represent the relative position of an individual score compared to the mean and variation of the values in a distribution. In a normally distributed sample, z-scores of cases should not exceed a value of 3.29. The observation of standardised z-scores of the total household expenditure revealed that eight outliers had a z-score higher than six which were dropped for the household expenditure regression model; three cases were dropped for the regression model for household income. These outliers were dropped in order to produce a more robust estimate of the coefficients.

Finally, the assumption of homogeneity of variance was assessed using scatter plots of standardized residuals against standardised predicted values. Additionally, the Breusch-Pegan test for hetereoscedasticity was used to explore whether the estimated variance of residuals in the analysed models were constant. The expected result was a homoscedastic variance of residuals in the sample. However, the observed result of this test revealed that the data for both analysis models were heteroscedastic. Referring first to the analysis for household income, six outliers were excluded to try to solve this data issue. This resulted in an improvement of the test results, as the data appeared to be homoscedastic. Conversely, the heteroscedasticity discovered in the analysis of household expenditure could not be improved by removing outliers from the data model. Therefore care had to be taken in the choice of mode of analysis and interpretation of results.

Multivariate Regression Modelling

Sampling Design

The sampling methodology for the LCF sample differs between Great Britain, conducted by ONS, and Northern Ireland, conducted by the Central Survey Unit of Northern Ireland Social Research Association (NISRA). A representative sample for Great Britain is drawn as a two-stage stratified random sample with clustering from the 'small user' Postal Address File. Postcode sectors are used as the Primary Sampling Units (PSUs), with 18 addresses selected from each PSU to form the monthly interviewer quota. For Northern Ireland a simple random sample of private addresses is drawn from the Valuation and Lands Agency List. (For further information on the LCF sampling methodology, please refer to the LCF Technical Report 2008).

To consider the sampling methodology in the analysis a special multivariate regression model was chosen, which takes the structure of PSUs and geographical strata into account.

Northern Ireland cases were sampled in a different way but included in the same model and therefore Northern Ireland as a whole was considered as one stratum, while each Northern Ireland case represented one PSU. Using the program STATA to analyse the data, the sampling method for the multiple linear regression could be specified through the 'svy' prefix. This method enables the calculation of robust standard errors in the regression model, which removes the bias introduced to the model through the heteroscedastic data. (For further information on this type of regression, please refer to: www.stata.com/help.cgi?=svy))

Statistical Modelling

Multiple linear regression models were chosen as the mode of analysis to identify the effects of individual and household characteristics on household income and expenditure. The dependent variable chosen for the expenditure model was the total household expenditure, which included the total consumption expenditure of the twelve, Classification Of Individual COnsumption by Purpose (COICOP), categories, as well as other expenditure items (e.g. mortgage interest payments, tax payments, holiday spending, cash gifts and charitable donations). For the income model the gross normal weekly household income was chosen as the dependent variable, which was derived from the income of all household members, taking into account not only earnings but also any incomings from self-employment, social security benefits, investments, pensions and annuities, as well as any other sources specified by respondents.

As previously mentioned, the multicollinearity test revealed collinearity between some variables in the original list. Table 5.2 presents the explanatory variables that were included in the final regression models for expenditure and income. The list also indicates the type of variable.

Table 5.2
Regression models used for analysis

Regression model for total household expenditure	Regression model for gross weekly household income
Gender for HRP (categorical)	Age of HRP (continuous)
Socio-economic status of HRP (categorical)	Socio-economic status of HRP (categorical)
Gross weekly household income (continuous)	Household composition (categorical)
Household composition (categorical)	Household tenure (categorical)
Household tenure (categorical)	Urban/rural identifier of household (categorical)
Urban/rural identifier of household (categorical)	

As a result of the test for normal distribution, evidence was found for highly skewed data. In order to conduct analysis based on normally distributed data, the dependent variables for the models were transformed using a natural logarithm. The regression analyses were modelled using the following formula:

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + \beta_2 Z_i + e_i$$

The natural logarithm of household expenditure or household income, ln(Y), was modelled as a function of individual characteristics of the HRP (X_i) and household characteristics (Z_i), and e_i represents the random error term. The model predicting household income included six cases with zero income. Since log-transformation cannot be applied to zero values, and recoding these values to 0.01 increased the homogeneity of variance of residuals, these six cases were excluded from the final model.

Results

To enable interpretation of the results, the regression coefficients need to be back-transformed by using the inverse of the natural logarithm function. It should be noted, when interpreting the results, that the coefficients can be back-transformed in this way however the model becomes multiplicative.

Total Household Expenditure Model

The results show that the explanatory variables in this model accounted for 64 per cent of the variance in total household expenditure ($R^2 = 0.64$). The full regression model is shown in Table 5.3. Examination of individual explanatory variables are summarised below. Unless otherwise stated the results are significant at the 95 per cent level.

- Gender was not significant in the original model and was therefore excluded from the analysis.
- The final analysis shows that the age of the HRP had an effect on the total household expenditure after controlling for all other characteristics in the model. The model shows less than one per cent decrease per unit increase of age.
- The socio-economic status of the HRP had an effect on household expenditure when all other characteristics in the model were kept constant (Table 5.3). In comparison to the reference group (households where the HRP had never worked or was in long-term unemployment), the model shows that households with a HRP employed by a large employer or in a higher management position had 59.5 per cent higher expenditure. This was followed by households with a HRP employed in a high professional occupation with

53.9 per cent higher spending than the reference group. This was closely followed by households with a HRP employed in a lower managerial or professional occupation.

- The household composition also had an effect on total household expenditure within the model (Table 5.3). When controlling for other characteristics, all other household combinations reported significantly higher household expenditure than the reference category of one adult (retired on state pension) households. Households with three or more adults with children had the highest expenditure, closely followed by other large household compositions such as households with three or more adults without children, and households with two adults and three or more children (155.3 per cent, 146.4 per cent and 145.2 per cent respectively).
- Relative to households that rent from local authorities, households owning a property by rental purchase had 64.2 per cent higher spending. This was also the case for households owning a property with a mortgage with 52.0 per cent higher expenditure when all other characteristics in the model were kept constant (Table 5.3).
- The model shows that the expenditure of households located in urban areas was 7.0 per cent lower than spending of rural households, when keeping all other characteristics constant.

Gross Weekly Household Income Model

The analysis shows that 64 per cent of the variance in total household expenditure was explained by the model ($R^2 = 0.64$). The full regression model is shown in Table 5.4. All regression coefficients proved to be significant at the 95 per cent level unless stated otherwise.

- After keeping all other explanatory variables constant, gender proved to be significant, indicating that households with a male HRP had 8.3 per cent higher gross weekly incomes than households with a female HRP. The age coefficient did not result in a significant value and was therefore excluded from the final model.
- The analysis shows that socio-economic status of the HRP had an effect on the gross weekly household income, after controlling for all other characteristics in the model. Results show that compared to the reference group (households where the HRP worked in routine occupations), households with a HRP employed by a large employer or in a higher management position had the largest incomes being 114.4 per cent above the income of reference households. This was followed by households where the HRP was employed in a high professional occupation, indicating a 102.0 per cent higher income. Unsurprisingly, those which were likely to have the lowest household incomes were households where the HRP has never worked or were long-term unemployed.
- Similar to the analysis on household expenditure, the household composition had an effect on gross weekly household income. Compared to the reference category (households with one adult retired on state pension), households with three or more adults without children had the highest income. This was followed by households with three or more adults with children, followed by households with two adults and 3 or more children.
- By observing the tenure type it can be seen that compared to households that rent from local authorities, households that own a property either with a mortgage, by rental purchase or outright, are more likely to have a higher gross weekly income.
- When comparing income in urban and rural households, the analysis showed that the gross weekly household income of urban households was less than 7.0 per cent lower than those of rural households.

Conclusion

The regression models produced for household income and household expenditure differ slightly in terms of the final variables. For the household expenditure model, the age of the HRP coefficient was found to be significant where it was not significant within the income model. Household income was included as an explanatory variable for the expenditure model but as it cannot be used as both a dependent and independent variable was therefore excluded from the income model. The age of the HRP was not found to be significant when modelling household income and was therefore excluded. However, the sex of the HRP coefficient was significant within the household income model but was not for household expenditure. It was therefore excluded from the final expenditure regression model. Apart from those exceptions the variables for both models were the same. This is to be expected because a higher income would generally lead to higher expenditure, so those variables which are significant in the income model would also be likely to have an effect on expenditure.

The section below describes ways in which the model could be improved and also a way to test the model coefficients further.

Further research

Through the test of the homogeneity of variance assumption it was discovered that the data model for household expenditure was biased due to heteroscedasticity of residuals. A possible reason for this could be that the fitted model could not explain cases with higher expenditure. Further analysis is necessary to explore this assumption.

A possible avenue for further investigation may be to explore whether adding an age squared variable to the regression analyses could help to explain more of the variance in the model. Income and expenditure generally increase as the age of the HRP increases before decreasing again. The age squared variable may be more appropriate to model this distribution within the regression analysis.

The inclusion of interaction terms could help improve the fit of the model. The investigation of interaction terms would also reveal how certain individual and household characteristics moderate each other. For example, there may be different linear models for male and female HRP gross income.

To further test the model an investigation could be carried out to identify the extent to which the explanatory variables predict household income and expenditure. One way to do this would be to use the model to predict household income and expenditure. These predicted values could then be compared to the actual values in an alternate dataset.

Table 5.3

Outcome variable: Total household expenditure

Explanatory variables:	Back-transformed coefficient	Significance[1]	Back-transformed 95% confidence interval	
Age of HRP	0.998	0.004	0.997	0.999
Socio-economic status:				
Never worked/long term unemployed		reference		
Large employer/higher management	1.595	0.000	1.398	1.820
High professional occupations	1.539	0.000	1.355	1.748
Lower managerial and professional occupations	1.518	0.000	1.346	1.712
Students	1.476	0.000	1.257	1.732
Small employers and own account workers	1.459	0.000	1.286	1.655
Intermediate occupations	1.449	0.000	1.276	1.646
Lower supervisory and technical occupations	1.388	0.000	1.230	1.568
Semi-routine occupations	1.251	0.000	1.103	1.419
Routine occupations	1.199	0.006	1.054	1.363
Not classified for other reasons	1.139	0.035	1.009	1.285
Gross weekly household income	1.000	0.000	1.000	1.000
Household composition:				
1 adult retired mainly dependent on state pension		reference		
3 or more adults with children	2.553	0.000	2.292	2.844
3 or more adults without children	2.464	0.000	2.235	2.717
2 adults and 3 or more children	2.452	0.000	2.188	2.747
2 adults and 2 children	2.221	0.000	2.006	2.459
2 adults and 1 child	2.129	0.000	1.915	2.366
1 man and 1 woman - other retired household	2.112	0.000	1.929	2.312
1 man and 1 woman - non-retired household	1.997	0.000	1.817	2.196
2 men or 2 women	1.826	0.000	1.620	2.058
1 adult and 2 or more children	1.784	0.000	1.569	2.029
1 man and 1 woman retired mainly dependent on state pension	1.694	0.000	1.524	1.882
1 adult and 1 child	1.560	0.000	1.384	1.758
1 adult - non-retired household	1.323	0.000	1.201	1.457
1 adult - other retired household	1.244	0.000	1.133	1.366
Household tenure:				
Local authority		reference		
Own by rental purchase	1.642	0.000	1.334	2.020
Own with mortgage	1.520	0.000	1.436	1.609
Private rented - unfurnished	1.433	0.000	1.344	1.529
Private rented furnished	1.413	0.000	1.267	1.577
Own outright	1.394	0.000	1.308	1.485
Housing association	1.161	0.000	1.084	1.244
Rentfree	1.128	0.054	0.998	1.274
Urban-rural classification:				
Rural household		reference		
Urban household	0.929	0.000	0.901	0.957
Constant	96.658	0.000	81.871	114.116
R-squared = 0.6388				

1 Significance relates to log tranformed coefficient

5

Table 5.4

Outcome variable: Gross weekly household income

Explanatory variables:	Back-transformed coefficient	Significance[1]	Back-transformed 95% confidence interval	
Sex of HRP:				
Female		reference		
Male	1.083	0.000	1.047	1.120
Socio-economic status:				
Routine occupations		reference		
Large employer/higher management	2.144	0.000	1.963	2.341
High professional occupations	2.020	0.000	1.852	2.202
Lower managerial and professional occupations	1.701	0.000	1.593	1.816
Lower supervisory and technical occupations	1.274	0.000	1.185	1.370
Intermediate occupations	1.229	0.000	1.132	1.333
Small employers and own account workers	1.126	0.007	1.034	1.227
Semi-routine occupations	1.054	0.170	0.978	1.136
Students	0.902	0.279	0.749	1.087
Never worked and long term unemployed	0.708	0.000	0.616	0.814
Not classified for other reasons	0.768	0.000	0.710	0.831
Household composition:				
1 adult retired mainly dependent on state pension		reference		
3 or more adults without children	3.613	0.000	3.284	3.974
3 or more adults with children	3.204	0.000	2.817	3.645
2 adults and 3 or more children	2.594	0.000	2.289	2.940
2 adults and 2 children	2.545	0.000	2.302	2.814
1 man and 1 woman non-retired household	2.510	0.000	2.287	2.754
1 man and 1 woman other retired household	2.509	0.000	2.321	2.712
2 adults and 1 child	2.336	0.000	2.107	2.591
2 men or 2 women	2.251	0.000	1.948	2.601
1 adult and 2 or more children	1.608	0.000	1.435	1.801
1 adult - other retired household	1.525	0.000	1.422	1.635
1 man and 1 woman retired mainly dependent on state pension	1.459	0.000	1.355	1.572
1 adult and 1 child	1.408	0.000	1.262	1.571
1 adult non-retired household	1.323	0.000	1.204	1.453
Household tenure:				
Local authority		reference		
Own with mortgage	1.731	0.000	1.628	1.842
Own by rental purchase	1.662	0.000	1.435	1.923
Own outright	1.529	0.000	1.442	1.622
Housing association	1.182	0.000	1.100	1.271
Rentfree	1.169	0.029	1.016	1.346
Private rented - unfurnished	1.293	0.000	1.205	1.388
Private rented furnished	1.154	0.085	0.980	1.358
Urban-rural classification:				
Rural household		reference		
Urban household	0.935	0.000	0.905	0.966
Constant	141.330	0.000	125.000	159.793
R-squared = 0.6438				

1 Significance relates to log tranformed coefficient