STAT6123 Coursework

Etienne Latif (31045499)

October 2022

1 Task 1

1.1 Question 1

We begin by assessing the distribution of the 'expenditure' variable, representing the total expenditure in GBP of households. As expenditure is a continuous variable, a frequency histogram and boxplot are appropriate visual representations for this analysis.

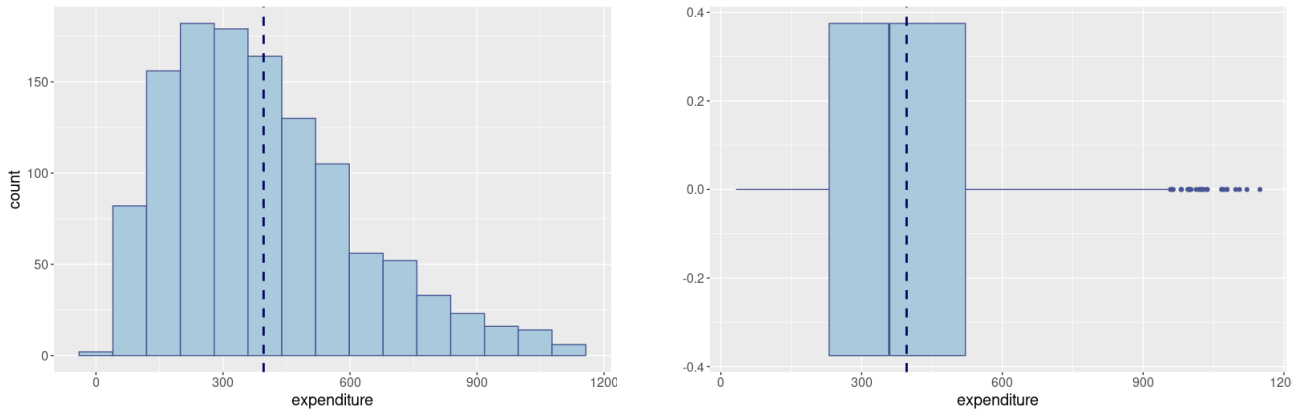

Figure 1: Frequency histogram (left) and boxplot (right) of the 'expenditure' variable. The mean (395.89) value of the variables is shown by the dashed line.

Looking at Figure 1, from the histogram, we see that expenditure is clearly a unimodal distribution, with the peak of the histogram visibly to the left of the mean. The long right tail is clear visual evidence from the histogram that the variable is right-skewed and unlikely to be normally distributed. Further evidence of asymmetry (and hence non-normality) is provided by the boxplot; the right whisker is significantly longer than the left, and the median is closer to the bottom of the box (the 1st-quartile). There is also a number of outliers at the top end of the boxplot (higher than the third quantile plus 1.5 times the interquartile range), indicating there are values that are extremely high above the mean.

Next we briefly explore the relationship between expenditure and income.

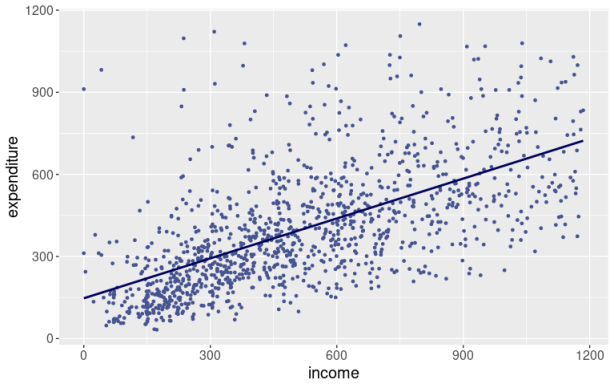

Figure 2: Scatter plot of 'expenditure' against 'income'. The dark blue line is a simple line of best fit, generated by a linear regression of expenditure on income.

It is immediately evident from Figure 2 that there is a large amount of variance in the value of expenditure for any fixed value of income. There appears to be a general positive correlation between the two variables, particularly we can note that the mean and minimum values of expenditure increase as income increases, however the most extreme high values of expenditure appear to be similar across all observed values of income. A simple line of best fit has been included to emphasise this relationship.

We turn our attention now to the relationships between expenditure and the categorical variables 'house.ten', 'sex.hh', 'lab.force', 'hh.size' and 'hh.adults'. The appropriate plot to determine the relationship between a continuous variable and a categorical variables is a boxplot. We generated boxplots for 'expenditure' for each of the variables, as shown below.

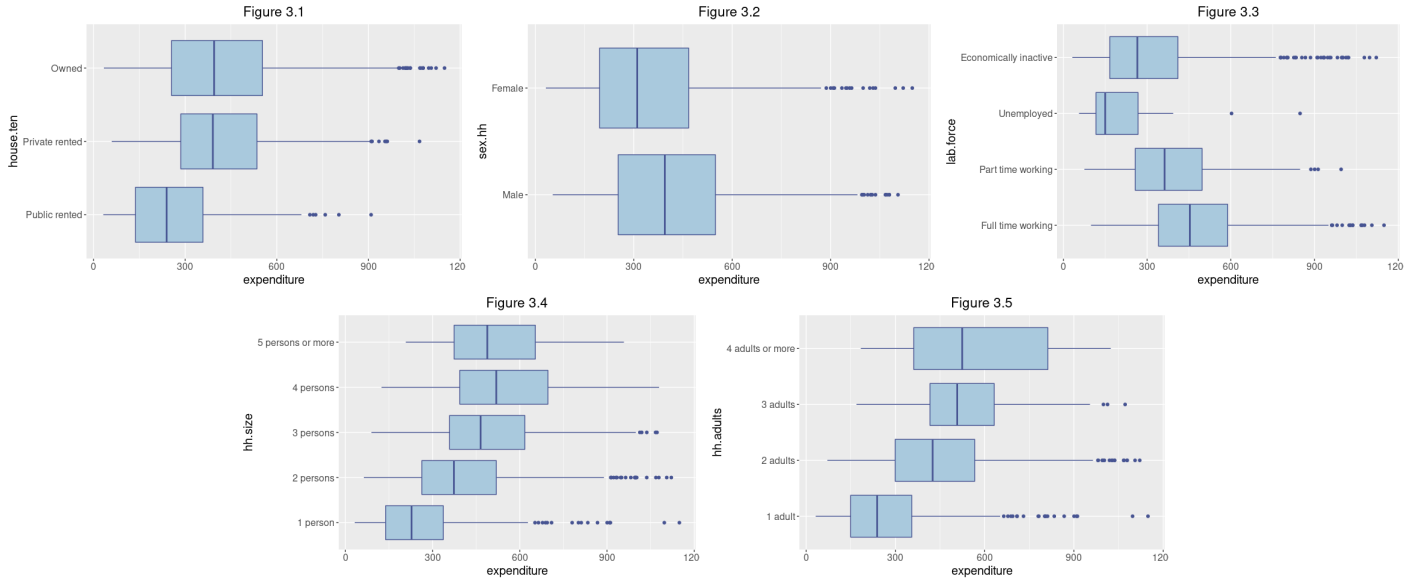

Figure 3: Boxplots of 'expenditure' against 'house.ten' (3.1), 'sex.hh' (3.2), 'lab.force' (3.3), 'hh.size' (3.4) and 'hh.adults' (3.5).

From Figure 3.1 we see there is an increase in median expenditure from 'Public rented' to 'Private rented' and 'Owned'. There is little difference in average expenditure between 'Private rented' and 'Owned', though there are more outliers and a greater interquartile range for 'Owned'. There are outliers at the high end of expenditure for each level. It is notable that we would not expect a clear linear/polynomial relationship between expenditure and household tenure, as the latter is a nominal categorical variable (no natural ordering of the values). Looking at Figure 3.2 we find that the median expenditure for households with male heads is higher than those with females by approximately 70GBP per week. The minimum expenditure for both sexes is very similar, and both also have outliers at the higher end of expenditure.

Figure 3.3, unsurprisingly shows some more clear distinctions in expenditure for different employment levels. As we would expect, there is a noticeable increase in expenditure from unemployed to part time working to full time working, though the difference between part time and full time working is less steep than that from unemployed to part time working. We also note that the median expenditure at unemployed is quite close to the 1st-quartile, implying a strong right skew. Interestingly, the economically inactive category had a median greater than unemployed but lower than part time working, with a significant number of outliers and a very long right tail. The definition of economically inactive used for this data set was not given, but from a review of an article by the Office for National Statistics [1] it appears that the term is used broadly to refer to people of working age who are unlikely to work in the foreseeable future. Speculatively, the significant right skew may be the result of the majority of economically inactive people not being largely wealthy, with outliers being those that are rich enough to not need to work in the foreseeable future. While a comparison of income against employment status can be viewed in the accompanying R code, and appears to provide evidence for this hypothesis, this can not be proved here and is ultimately beyond the scope of this report.

The remaining variables, household size and number of adults in household, may be treated as ordinal variables, increasing as number of people/adults per household increases. From Figure 3.4, we see median expenditure increasing from 1 person to 4 persons, then slightly decreasing from 4 persons to 5+ persons. We again see a pattern of increasing minimum expenditures across the entire group, but find a decreasing number of outliers at the highest values of expenditure as the number of persons increases (and hence the spread of expenditure decreases as number of persons increases). This may be due to larger households being more likely to be families and therefore having similar financial considerations, though again this is beyond the scope of this paper. From Figure 3.5 we see the median expenditure increasing from 1 adult to 3 adults, and very little difference betwen 3 adults and 4 adults or more. We see the same relationship in spread as with household size, except with a significantly larger interquartile range for families with 4 adults or more compared to the other levels of the variables.

1.2 Question 2

We start with a simple linear regression of 'expenditure' on 'income', taking the following parametric form:

$$\text{expenditure} = \alpha + \beta_1 \text{income} \tag{1}$$

The maximum likelihood estimates of these parameters and their standard errors are given in Table 1. A plot of the studentised residuals against fitted values and a QQ-plot are displayed in Figure 4.

Coefficient	Maximum likelihood estimate	Standard error
α	147.224	10.386
β_1	0.486	0.018

Table 1: Table displaying the maximum likelihood estimates and their standard errors for the coefficients in the simple linear regression of 'expenditure' on 'income'.

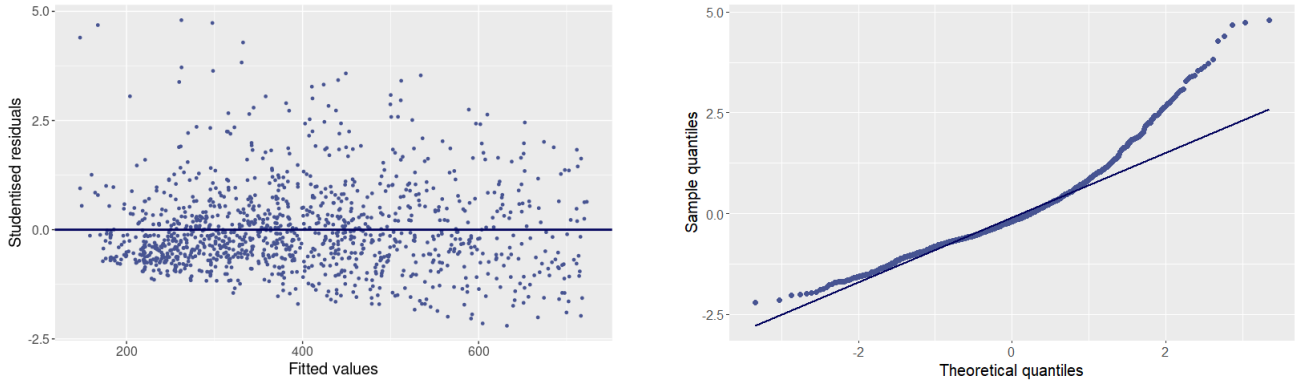

Figure 4: Plot of the studentised residuals against fitted values (left) and a QQ-plot of the quantiles of the studentised residuals against the theoretical quantiles of a standard Gaussian distribution (right) for a linear regression of 'expenditure' on 'income'.

From the plot of studentised residuals against fitted values in Figure 4 we can ascertain that the assumption of homoscedasticity does not appear to hold; there is greater variance at lower fitted values than at higher values. The assumption of linearity is not easy to judge from this plot, but it does appear the model is underpredicting more for lower fitted values, and overpredicting more at higher fitted values. This general trend can be observed across the range of fitted values, but is not entirely clear. The violation of the normality assumption is more evident, as the curvature of the QQ-plot in Figure 4 clearly shows that the distribution has a long right tail and hence is right skewed.

1.3 Question 3

We now add 'income' squared as a predictor to the model given by Equation 1, to obtain the following parametric form:

$$\text{expenditure} = \alpha + \beta_1 \text{income} + \beta_2 \text{income}^2 \tag{2}$$

The maximum likelihood estimates of these parameters and their standard errors are given in Table 2. The residuals against fitted values plot and QQ-plot are given in Figure 5.

Coefficient	Maximum likelihood estimate	Standard error
α	103.200	28.600
β_1	0.690	0.071
β_2	-1.761×10^{-4}	5.910×10^{-5}

Table 2: Table displaying the maximum likelihood estimates and their standard errors for the coefficients in the linear regression of 'expenditure' on 'income' and 'income' squared.

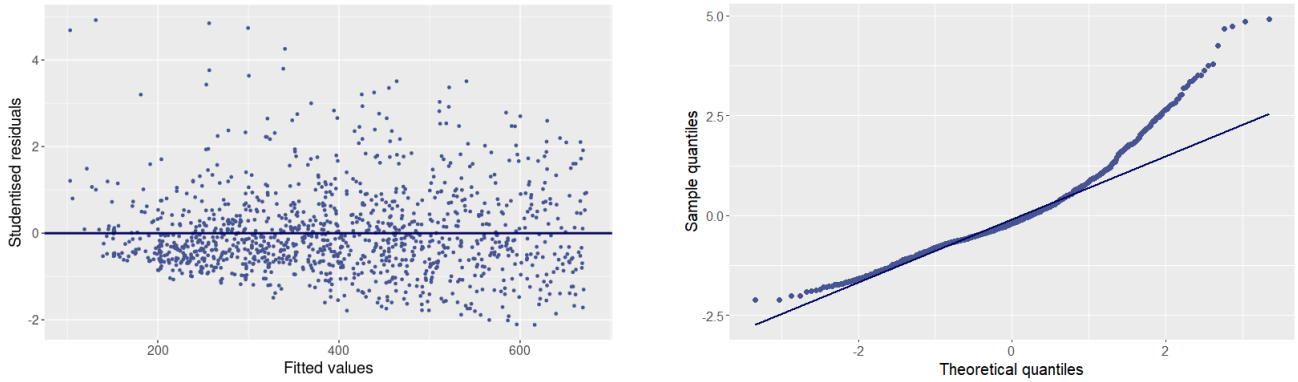

Figure 5: Plot of the studentised residuals against fitted values (left) and a QQ-plot of the quantiles of the studentised residuals against the theoretical quantiles of a standard Gaussian distribution (right) for a linear regression of 'expenditure' on 'income' and 'income' squared.

The assumption violations do not appear to be fixed by the addition of the quadratic term. Both plots in Figure 5 look very similar to those in Figure 4. Intuitively we would not expect the addition of a quadratic term to remedy the violation of the normality and homoscedasticity assumptions, but we also find that there still seems to be a subtle pattern present in the residuals (which we might have expected a quadratic term to fix).

1.4 Question 4

We now apply a transformation to the response, and regress the natural logarithm of 'expenditure' on 'income'. Applying this transformation is a typical method to attempt to fix non-normality and heteroscedasticity issues. The parametric form of this model is given in Equation 3 (where log(.) denotes the natural logarithm).

$$\log(\text{expenditure}) = \alpha + \beta_1 \text{income} \tag{3}$$

The maximum likelihood estimates of these parameters and their standard errors are given in Table 3. The residuals against fitted values plot and QQ-plot are given in Figure 6.

Coefficient	Maximum likelihood estimate	Standard error
α	5.074	0.028
β_1	1.436×10^{-3}	4.821×10^{-5}

Table 3: Table displaying the maximum likelihood estimates and their standard errors for the coefficients in the linear regression of the natural logarithm of 'expenditure' on 'income'.

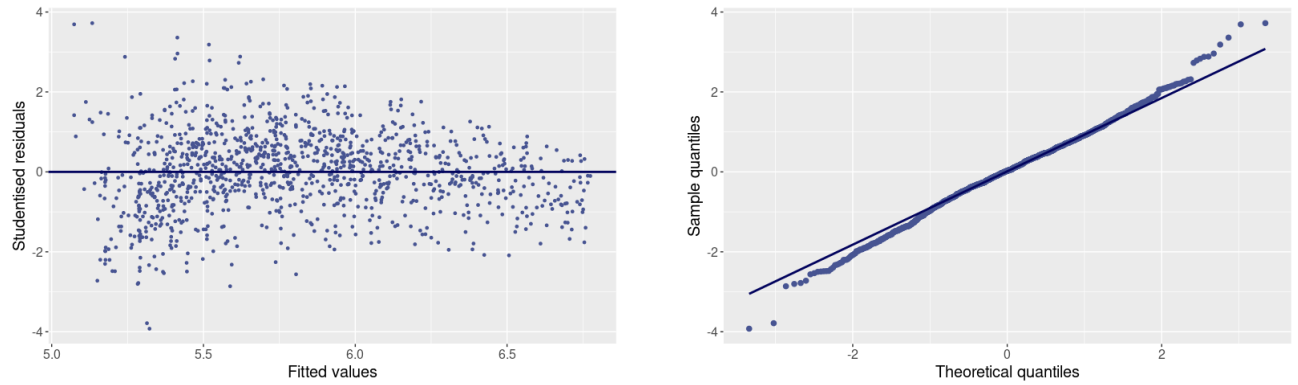

Figure 6: Plot of the studentised residuals against fitted values (left) and a QQ-plot of the quantiles of the studentised residuals against the theoretical quantiles of a standard Gaussian distribution (right) for a linear regression of the natural logarithm of 'expenditure' on 'income'

From Figure 6 it appears that the assumption of linearity is reasonable, though a slight (negative) quadratic relationship seems feasible, particularly in the denser region of points. However, heteroscedasticity is still an issue, as there is noticeably more variance at lower fitted values, creating the characteristic "cone" shape. The assumption of normality seems more reasonable for this model, with fewer outliers at the extremes of the plot, though the slight "S" shape reveals that the distribution has slightly long tails compared to the Gaussian distribution.

1.5 Question 5

We now add the quadratic 'income' term to the model in Equation 3 to obtain the following model:

$$\log(\text{expenditure}) = \alpha + \beta_1 \text{income} + \beta_2 \text{income}^2 \tag{4}$$

The maximum likelihood estimates of these parameters and their standard errors are given in Table 4. The residuals against fitted values plot and QQ-plot are given in Figure 7.

Coefficient	Maximum likelihood estimate	Standard error
α	4.710	0.047
β_1	3.117×10^{-3}	1.858×10^{-4}
β_2	-1.453×10^{-6}	1.554×10^{-7}

Table 4: Table displaying the maximum likelihood estimates and their standard errors for the coefficients in the linear regression of the natural logarithm of 'expenditure' on 'income' and 'income' squared.

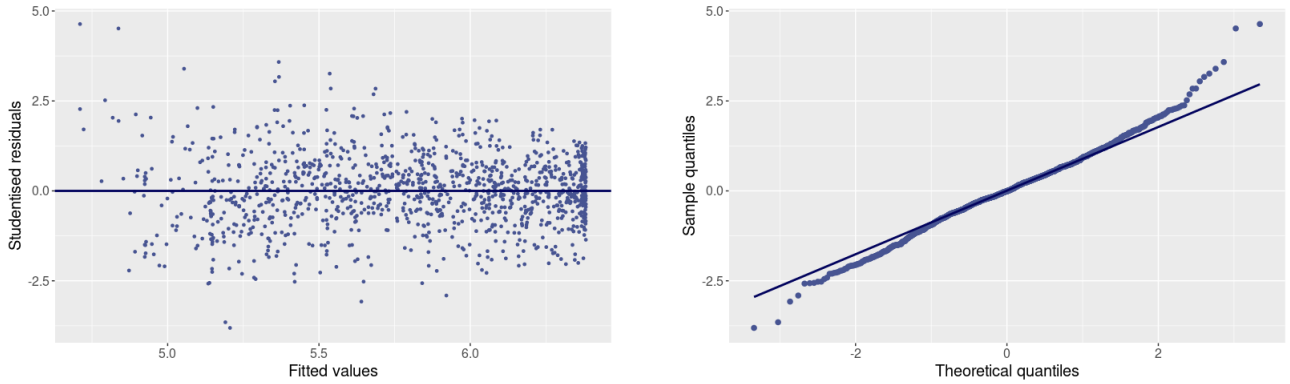

Figure 7: Plot of the studentised residuals against fitted values (left) and a QQ-plot of the quantiles of the studentised residuals against the theoretical quantiles of a standard Gaussian distribution (right) for a linear regression of the natural logarithm of 'expenditure' on 'income' and 'income' squared.

Looking at Figure 7, there is no longer any clear pattern in the residuals, suggesting the relationship between the independent and dependent variables has been well modelled. Homoscedasticity also seems to be a reasonable assumption, with the exception of some outliers for smaller fitted values. The normality assumption again seems fairly reasonable, despite there still being slightly long tails compared to the standard Gaussian distribution.

1.6 Question 6

As the models given by Equations 1 and 2 violate our linear modelling assumptions significantly, we will choose to focus our attention on the models given by Equations 3 and 4, modelling the natural logarithm of the response variable by some covariates. We notice that the models have the same parametric form, but the set of predictors for Model 3 is a subset of those in Model 4, so we can say that Model 3 is nested within Model 4. Firstly, note that there is a small increase in the adjusted R^2 value from Model 3 (0.425) to Model 4 (0.464), suggesting that the more complex model explains a larger proportion of the variance in the dependent variable. Considering a t-test with null hypothesis $\beta_2 = 0$ against alternative hypothesis $\beta_2 \neq 0$, we get an extremely small p-value ($< 2 \times 10^{-16}$), giving us evidence at the standard 5% significance level to reject the null hypothesis and include the quadratic term in the model. As the models vary by only one parameter, it is redundant to compare the two with an F-test, as the result will be the same. Hence we conclude that with the addition of the quadratic term, Model 4 fits the data significantly better than Model 3, and is therefore our preferred model.

Let us denote Y = expenditure and x = income so that our model is $\log Y = 4.71 + 3.117 \times 10^{-3} x - 1.453 \times 10^{-6} x^2$. Taking the exponential function of both sides we get $Y = \exp(4.71 + 3.117 \times 10^{-3} x - 1.453 \times 10^{-6} x^2)$, so that expenditure is a negative quadratic function of income. It is easier to maximise the equation for $\log Y$ (and is equivalent to maximising Y directly), hence using simple calculus we get:

$$\begin{aligned} \frac{d[\log Y]}{dx} &= 3.117 \times 10^{-3} - 2.906 \times 10^{-6} x \\ \frac{d^2[\log Y]}{dx^2} &= -2.906 \times 10^{-6} \end{aligned} \tag{5}$$

Omitting the trivial mathematical details, we have expenditure as a negative quadratic function of income with maximum point $x = 1072.61$. Hence we note that the rate of change of expected expenditure with income is also a function of income. We expect

expenditure to increase as income increases with income up to income = 1072.61 GBP per week and expenditure = 590.91 GBP, but then decrease after this point. We can also find an expression for the percentage change in expenditure for a unit change in income by letting $\log Y_1 = 4.71 + 3.117 \times 10^{-3}x - 1.453 \times 10^{-6}x^2$ and $\log Y_2 = 4.71 + 3.117 \times 10^{-3}(x+1) - 1.453 \times 10^{-6}(x+1)^2$. Clearly $\log Y_2 - \log Y_1 = \log \frac{Y_2}{Y_1} = 3.117 \times 10^{-3} - 1.453 \times 10^{-6}(2x+1) = 3.116 \times 10^{-3} - 2.906 \times 10^{-6}x$ and hence by taking the exponential function and then subtracting one from both sides we get $\frac{Y_2 - Y_1}{Y_1} = \exp(3.116 \times 10^{-3} - 2.906 \times 10^{-6}x) - 1$, or a $\left[\exp(3.116 \times 10^{-3} - 2.906 \times 10^{-6}x) - 1\right] \times 100\%$ increase in expected expenditure for a unit increase in income.

We only have data up to income = 1184.472, and would not be able to extrapolate this relationship beyond this point (as is poor practice regardless) or we would eventually obtain negative expenditure for high values of income due to the negative second derivative, which is invalid. Finally, the intercept $\alpha = 4.71$ is the expected log expenditure, translating to an expected expenditure of 111.05 GBP, given that income is equal to zero.

1.7 Question 7

In this question we attempt to build a suitable regression model for expenditure. Throughout the model building process we shall track the best model we have found so far, with this "best model" being initially set as the model we concluded to be our preferred model in the last question, given by Equation 4. We shall build upon this model by attempting to add a single variable or interaction term to our best model at a time, and using an F-test with a 5% significance level to assess the fit of the more complex model (the anova() command in R makes the F-test convenient, although a t-test is equally valid as the compared models will always differ by a single parameter). For models with an F-test p-value of below 0.05 we will conclude that the model fits the data significantly better enough to justify the trade-off in reduced parsimony, and this model shall become our new best model; otherwise we shall reject the new proposed model in favour of the more parsimonious model that we have so far obtained. The exact results of these F-tests can be viewed in the R script, but are excluded from this report for concision.

It is generally poor practice to include interaction terms of a variable in a model if the main effect of that variable is not included in the model (or more generally we should not include an interaction between terms where each of the lower order interactions containing those terms are not also included in the model). Consequently we will begin by attempting to add the main effects of 'house.ten', 'sex.hh', 'lab.force', 'hh.size', and 'hh.adults'. Using the procedure already described, we find the addition of 'house.ten', 'lab.force' and 'hh.size' to the model to be significant. We choose to not add the other variables to the model and hence will not attempt to add interactions containing these variables.

Next we attempt to include the 2-way interactions between the income variable and each of the three categorical variables that have been added to the model. Interestingly, the order in which these interactions are added to the model appears to affect whether or not the interaction provides a significantly better fit to the data. Specifically when the interaction between income and 'house.ten' is added before the other two interactions the interaction appears to be significant, but when it is added after the other two interactions it does not. This is proven in the R code by considering three F-tests comparing the model with all three interactions against the models that are obtained by dropping one interaction at a time; we find that the model including the income and 'house.ten' interaction is not a significantly better fit. As a result, we choose to include just the interactions between income and 'lab.force' and between income and 'hh.size' to the model.

Considering the addition of the three possible 2-way interactions between the categorical variables of interest, we find none of these interactions to significantly improve the fit of our model, and hence do not include them in our model.

As there are no higher order interactions possible such that every lower order interaction of the variables is included in the model, we conclude that the best model is given in parametric form by the following equation:

$$\begin{aligned}
\log(\text{expenditure}) = {} & \alpha + \beta_1 \times \text{income} + \beta_2 \times \text{income}^2 \\
& + \gamma_{11} \times [\text{house.ten} = \text{Private rented}] + \gamma_{12} \times [\text{house.ten} = \text{Owned}] \\
& + \gamma_{21} \times [\text{lab.force} = \text{Part time working}] + \gamma_{22} \times [\text{lab.force} = \text{Unemployed}] \\
& + \gamma_{23} \times [\text{lab.force} = \text{Economically inactive}] \\
& + \gamma_{31} \times [\text{hh.size} = 2 \text{ persons}] + \gamma_{32} \times [\text{hh.size} = 3 \text{ persons}] + \gamma_{33} \times [\text{hh.size} = 4 \text{ persons}] \\
& + \gamma_{34} \times [\text{hh.size} = 5 \text{ persons or more}] \\
& + \delta_{11} \times \text{income} \times [\text{lab.force} = \text{Part time working}] + \delta_{12} \times \text{income} \times [\text{lab.force} = \text{Unemployed}] \\
& + \delta_{13} \times \text{income} \times [\text{lab.force} = \text{Economically inactive}] \\
& + \delta_{21} \times \text{income} \times [\text{hh.size} = 2 \text{ persons}] + \delta_{22} \times \text{income} \times [\text{hh.size} = 3 \text{ persons}] \\
& + \delta_{23} \times \text{income} \times [\text{hh.size} = 4 \text{ persons}] + \delta_{24} \times \text{income} \times [\text{hh.size} = 5 \text{ persons or more}]
\end{aligned} \tag{6}$$

Note that we have used the notation of [true] = 1, [false] = 0. The maximum likelihood estimates and standard errors for the coefficients are given in the table below:

Coefficient	Maximum likelihood estimate	Standard error
α	4.901	9.738×10^{-2}
β_1	1.644×10^{-3}	2.701×10^{-4}
β_2	-4.335×10^{-7}	1.981×10^{-7}
γ_{11}	2.563×10^{-1}	4.257×10^{-2}
γ_{12}	1.748×10^{-1}	3.506×10^{-2}
γ_{21}	-1.392×10^{-1}	1.023×10^{-1}
γ_{22}	-6.314×10^{-1}	1.535×10^{-1}
γ_{23}	-3.022×10^{-1}	7.921×10^{-2}
γ_{31}	4.145×10^{-1}	6.612×10^{-2}
γ_{32}	3.650×10^{-1}	1.093×10^{-1}
γ_{33}	5.139×10^{-1}	1.342×10^{-1}
γ_{34}	4.803×10^{-1}	1.866×10^{-1}
δ_{11}	2.318×10^{-4}	1.790×10^{-4}
δ_{12}	1.078×10^{-3}	5.440×10^{-4}
δ_{13}	4.697×10^{-4}	1.340×10^{-4}
δ_{21}	-4.641×10^{-4}	1.377×10^{-4}
δ_{22}	-2.487×10^{-4}	1.812×10^{-4}
δ_{23}	-3.736×10^{-4}	2.074×10^{-4}
δ_{24}	-3.215×10^{-4}	2.841×10^{-4}

Table 5: The maximum likelihood estimates and standard errors of the regression coefficients of our preferred model for expenditure.

Note that the model has an adjusted R^2 value of 0.524, indicating that 52.4% of the variance in the data is explained by the explanatory variables in the model. Examining the studentised residuals versus fitted values plot and QQ-plot of the studentised residuals against the standard Gaussian distribution given in Figure 8 we find that our linear modelling assumptions are well satisfied. Our residuals show no clear pattern, implying the relationship between expenditure and the explanatory variables is well modelled. A certain amount of heteroscedasticity is still present, especially due to outliers at lower fitted values, but is an improvement over that which is present in Figure 7. Finally, our normality assumption seems to be reasonable, despite slightly long tails (particularly at the higher quantiles) in comparison to the standard Gaussian distribution.

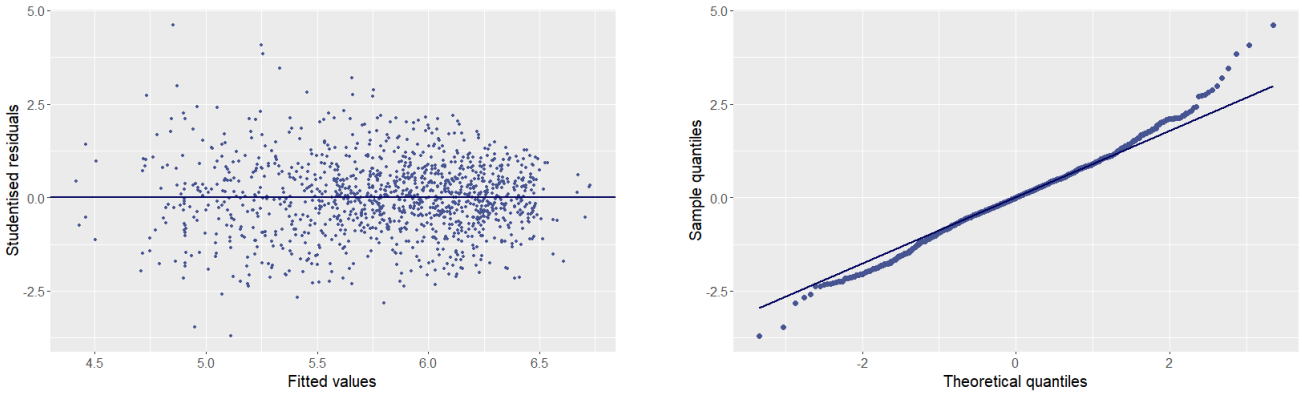

Figure 8: Plot of the studentised residuals against fitted values (left) and a QQ-plot of the quantiles of the studentised residuals against the theoretical quantiles of a standard Gaussian distribution (right) for the linear model given by Equation 6 and Table 5.

1.8 Question 8

We now interpret the parameters of the model specified by Equation 6 and Table 5 and explain the relationship between expenditure and the explanatory variables that is implied by the model.

The simplest relationship to explain is that between expenditure and 'house.ten'. Notice that 'Public rented' was used as the reference category. The values of $\gamma_{11} = 2.563 \times 10^{-1}$ and $\gamma_{12} = 1.748 \times 10^{-1}$ give us how much higher the expected log expenditure is for those with private rented and owned houses respectively in comparison to those public rented houses, when all other variables are held constant. The effects of the different levels of these factors on expected expenditure are multiplicative, with private rented house observations having an expected expenditure $e^{2.563 \times 10^{-1}}$ times greater than public rented houses and owned house observations having an expected expenditure $e^{1.748 \times 10^{-1}}$ times greater than public rented houses (again, when all other variables are held constant). This is shown by taking the exponential function of both sides of Equation 6 to get expenditure $=$ $\exp(\gamma_{11} \times [\text{house.ten} = \text{Private rented}] + \gamma_{12} \times [\text{house.ten} = \text{Owned}]) \exp(C)$ where C is all the other parameters in the model.

It is not common practice to interpret the main effects of categorical variables which are involved in interactions in the model. As such we shall look at the effect of different levels of employment status and household size on the relationship between income

and expenditure. Let us again denote expenditure by Y and income by x. Then, we have an expression for the effect of income on expenditure by:

$$\begin{aligned} \log Y &= \beta_1 x + \beta_2 x^2 + \delta_c x + C \\ &= (\beta_1 + \delta_c)\, x + \beta_2 x^2 \end{aligned} \tag{7}$$

Where C denotes all the effects not involving income and δ_c is the sum of the δ coefficients that are obtained for a given observation of employment status and household size. Then by defining $\log Y_1 = (\beta_1 + \delta_c)\, x + \beta_2 x^2$ and $\log Y_2 = (\beta_1 + \delta_c)\,(x+1) + \beta_2 (x+1)^2$ we get $\log \frac{Y_2}{Y_1} = \beta_1 + \beta_2 + \delta_c + 2\beta_2 x$ and therefore $\frac{Y_2 - Y_1}{Y_1} = \exp(\beta_1 + \beta_2 + \delta_c + 2\beta_2 x) - 1$. In other words for a unit increase in income, we expect a $[\exp(\beta_1 + \beta_2 + \delta_c + 2\beta_2 x) - 1] \times 100\%$ increase in expenditure. For the 20 combinations of employment status and household size, the expected percentage change in expenditure is given below.

Employment status (lab.force)	**Household size** (hh.size)	$\boldsymbol{\delta_c}$ $(\delta_{1i} + \delta_{2j})$	**Percentage change in expenditure for a unit increase in income** ($\times 100\%$)
Full time working	1 person	0	$e^{1.644\times10^{-3} - 8.67\times10^{-7}x} - 1$
Full time working	2 persons	-4.641×10^{-4}	$e^{1.179\times10^{-3} - 8.67\times10^{-7}x} - 1$
Full time working	3 persons	-2.487×10^{-4}	$e^{1.395\times10^{-3} - 8.67\times10^{-7}x} - 1$
Full time working	4 persons	-3.736×10^{-4}	$e^{1.270\times10^{-3} - 8.67\times10^{-7}x} - 1$
Full time working	5 persons	-3.215×10^{-4}	$e^{1.322\times10^{-3} - 8.67\times10^{-7}x} - 1$
Part time working	1 person	2.318×10^{-4}	$e^{1.875\times10^{-3} - 8.67\times10^{-7}x} - 1$
Part time working	2 persons	-2.323×10^{-4}	$e^{1.411\times10^{-3} - 8.67\times10^{-7}x} - 1$
Part time working	3 persons	-1.69×10^{-5}	$e^{1.627\times10^{-3} - 8.67\times10^{-7}x} - 1$
Part time working	4 persons	-1.418×10^{-4}	$e^{1.502\times10^{-3} - 8.67\times10^{-7}x} - 1$
Part time working	5 persons	-8.97×10^{-5}	$e^{1.554\times10^{-3} - 8.67\times10^{-7}x} - 1$
Unemployed	1 person	1.078×10^{-3}	$e^{2.722\times10^{-3} - 8.67\times10^{-7}x} - 1$
Unemployed	2 persons	6.139×10^{-4}	$e^{2.257\times10^{-3} - 8.67\times10^{-7}x} - 1$
Unemployed	3 persons	8.293×10^{-4}	$e^{2.473\times10^{-3} - 8.67\times10^{-7}x} - 1$
Unemployed	4 persons	7.044×10^{-4}	$e^{2.348\times10^{-3} - 8.67\times10^{-7}x} - 1$
Unemployed	5 persons	7.565×10^{-4}	$e^{2.400\times10^{-3} - 8.67\times10^{-7}x} - 1$
Economically inactive	1 person	4.697×10^{-4}	$e^{2.113\times10^{-3} - 8.67\times10^{-7}x} - 1$
Economically inactive	2 persons	5.6×10^{-6}	$e^{1.649\times10^{-3} - 8.67\times10^{-7}x} - 1$
Economically inactive	3 persons	2.21×10^{-4}	$e^{1.865\times10^{-3} - 8.67\times10^{-7}x} - 1$
Economically inactive	4 persons	9.61×10^{-5}	$e^{1.740\times10^{-3} - 8.67\times10^{-7}x} - 1$
Economically inactive	5 persons	1.482×10^{-4}	$e^{1.792\times10^{-3} - 8.67\times10^{-7}x} - 1$

Table 6: The percentage change in expected expenditure per unit increase in income for each possible level of the factors that interact with income.

We may also note that the intercept $\alpha = 4.901$ represents the expected value of log expenditure when income is zero and all categorical variables take their reference level. In such cases, the expected value of expenditure is $\exp(\alpha) = 134.424$ GBP.

2 Task 2

2.1 Question 1

We are working with a generalised linear model with exponential distribution response variable Y_i ($i \in \{1, ..., 1000\}$). Note that this distribution is a member of the exponential family of distributions:

$$\begin{aligned} f(y_i; \theta_i) &= \theta_i \exp(-\theta_i y_i) \\ &= \exp(y_i(-\theta_i) + \log \theta_i + 0) \end{aligned} \tag{8}$$

As the probability density function is in canonical form (i.e. $a(y_i) = y_i$, using the notation of the lecture notes) we can easily calculate the mean and variance of the response using the formulae from the lecture notes and the expressions for $b'(\theta_i), b''(\theta_i), c'(\theta_i), c''(\theta_i)$ given in the assignment sheet:

$$\begin{aligned} E[a(Y_i)] \equiv E[Y_i] &= -\frac{c'(\theta_i)}{b'(\theta_i)} = -\frac{1/\theta_i}{-1} = \frac{1}{\theta_i} = \mu_i \\ Var[a(Y_i)] \equiv Var[Y_i] &= \frac{b''(\theta_i)c'(\theta_i) - c''(\theta_i)b'(\theta_i)}{b'(\theta_i)^3} = \frac{-\left(-1/\theta_i^2\right)(-1)}{(-1)^3} = \frac{1}{\theta_i^2} = \mu_i^2 \end{aligned} \tag{9}$$

The mean response is modelled as a function of a single explanatory variable x_i by a non-canonical link function and systematic component as given: $\log \mu_i = \beta_0 + \beta_1 x_i \equiv \eta_i$. The systematic component can be expressed as $\boldsymbol{x}_i^T \boldsymbol{\beta}$ where $\boldsymbol{x}_i = [1, x_i]^T$, and $\boldsymbol{\beta} = [\beta_0, \beta_1]^T$.

We begin by finding the score vector $\boldsymbol{u}(\boldsymbol{\beta}) = [u_0(\boldsymbol{\beta}), u_1(\boldsymbol{\beta})]^T \equiv [u_0, u_1]^T$. Using the lecture notes, we have:

$$\begin{aligned} u_j &= \sum_{i=1}^{1000} \left(\frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \right) \\ &= \sum_{i=1}^{1000} \left(\frac{(y_i - \mu_i) x_{ji}}{Var[Y_i]} \frac{\partial \mu_i}{\partial \eta_i} \right) \quad \forall j \in \{0, 1\} \end{aligned} \tag{10}$$

We have already obtained an expression for $Var[Y_i]$, and the derivative term is obtained simply:

$$\frac{\partial \mu_i}{\partial \eta_i} = \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^{-1} = \left(\frac{\partial}{\partial \mu_i} \log \mu_i \right)^{-1} = \mu_i \tag{11}$$

Hence we can simplify Equation 10 to obtain:

$$\begin{aligned} u_0 &= \sum_{i=1}^{1000} \frac{(y_i - \mu_i)}{\mu_i^2} \mu_i = \sum_{i=1}^{1000} \frac{y_i - \mu_i}{\mu_i} \\ u_1 &= \sum_{i=1}^{1000} \frac{(y_i - \mu_i) x_i}{\mu_i^2} \mu_i = \sum_{i=1}^{1000} \frac{(y_i - \mu_i) x_i}{\mu_i} \end{aligned} \tag{12}$$

Which can be expressed as a vector as below:

$$\boldsymbol{u}(\boldsymbol{\beta}) = [u_1, u_2]^T = \left[\sum_{i=1}^{1000} \frac{y_i - \mu_i}{\mu_i}, \sum_{i=1}^{1000} \frac{(y_i - \mu_i) x_i}{\mu_i} \right]^T = \boldsymbol{X}^T \boldsymbol{z} \tag{13}$$

where $\boldsymbol{z} = [z_1, ..., z_n]^T, z_i = \frac{y_i - \mu_i}{\mu_i}$ (note this is not the same as the $\boldsymbol{z}$ defined in the IWLS section of the lectures) and $\boldsymbol{X} = \left[\boldsymbol{x}_1^T, ..., \boldsymbol{x}_n^T\right]^T$ (the design matrix). As $\boldsymbol{z}$ depends on the unknown parameters $\boldsymbol{\beta}$ we can not explicitly calculate the values of $\boldsymbol{u}$ without an estimate of $\boldsymbol{\beta}$. Now turning our attention to the information matrix, we can use the following derivation described in the lecture notes:

$$\begin{aligned} I_{jk} &= E[U_j U_k] \\ &= E\left[\left(\sum_{i=1}^{1000} \frac{(Y_i - \mu_i) x_{ji}}{Var[Y_i]} \frac{\partial \mu_i}{\partial \eta_i} \right) \left(\sum_{l=1}^{1000} \frac{(Y_l - \mu_l) x_{kl}}{Var[Y_l]} \frac{\partial \mu_l}{\partial \eta_l} \right) \right] \\ &= \sum_{i=1}^{1000} \sum_{l=1}^{1000} \frac{x_{ji} x_{kl}}{Var[Y_i] Var[Y_l]} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \mu_l}{\partial \eta_l} E[(Y_i - \mu_i)(Y_l - \mu_l)] \\ &= \sum_{i=1}^{1000} \frac{x_{ji} x_{ki}}{Var[Y_i]} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \\ &= \sum_{i=1}^{1000} x_{ji} x_{ki} = \boldsymbol{X}^T \boldsymbol{X} \end{aligned} \tag{14}$$

Where the final line arises from noticing that $Var[Y_i] = \mu_i^2 = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$. As the Fisher information matrix is small (2x2) and depends on only known quantities we can explicitly express the entire matrix:

$$\boldsymbol{I} = \boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} 1000 & \sum_{i=1}^{1000} x_i \\ \sum_{i=1}^{1000} x_i & \sum_{i=1}^{1000} x_i^2 \end{bmatrix} = \begin{bmatrix} 1000 & 607356.6 \\ 607356.6 & 390065229.4 \end{bmatrix} \tag{15}$$

2.2 Question 2

We can derive the Fisher scoring algorithm to find the maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ using the expression in the lecture notes:

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(m+1)} &= \hat{\boldsymbol{\beta}}^{(m)} + \left[\boldsymbol{I}^{(m)}\right]^{-1} \boldsymbol{u}^{(m)} \\ &= \hat{\boldsymbol{\beta}}^{(m)} + \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \boldsymbol{X}^T\boldsymbol{z}^{(m)} \end{aligned} \tag{16}$$

Note that the Fisher information matrix does not depend on the current estimate $\hat{\boldsymbol{\beta}}$. Implementing the algorithm in R yields the estimates $\hat{\boldsymbol{\beta}}_0 = -0.12115$ and $\hat{\boldsymbol{\beta}}_1 = 0.0052629$ (both to 5 significant figures).

2.3 Question 3

Following the lecture notes, for large samples (as we have) the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by $\boldsymbol{I}^{-1}$, which in this case is:

$$\boldsymbol{I}^{-1} = \begin{bmatrix} 1.8414 \times 10^{-2} & -2.8672 \times 10^{-5} \\ -2.8672 \times 10^{-5} & 4.7207 \times 10^{-8} \end{bmatrix} \tag{17}$$

with each value rounded to 5 significant figures.

2.4 Question 4

The t-statistic for β_1 can be calculated by:

$$\begin{aligned} t &= \frac{\hat{\boldsymbol{\beta}}_1}{\sqrt{\hat{Var}\left[\hat{\boldsymbol{\beta}}_1\right]}} \\ &= \frac{\hat{\boldsymbol{\beta}}_1}{\left[\boldsymbol{I}^{-1}\right]_{11}^{1/2}} \\ &= 24.223 \end{aligned} \tag{18}$$

Using R to find the probability that a t-distribution with 997 degrees of freedom takes a value at least as extreme as the calculated test statistic returns zero, implying a probability negligible to zero. For a hypothesis test of $\mathrm{H}_0 : \beta_1 = 0$ vs $\mathrm{H}_1 : \beta_1 \neq 0$, we have significant evidence to reject H_0, hence we conclude that the variable is significant at the standard 5% significance level.

References

[1] O. for National Statistics, "Economic inactivity," 2021. [Online]. Available: https://www.ethnicity-facts-figures.service.gov.uk/work-pay-and-benefits/unemployment-and-economic-inactivity/economic-inactivity/latestdownload-the-data