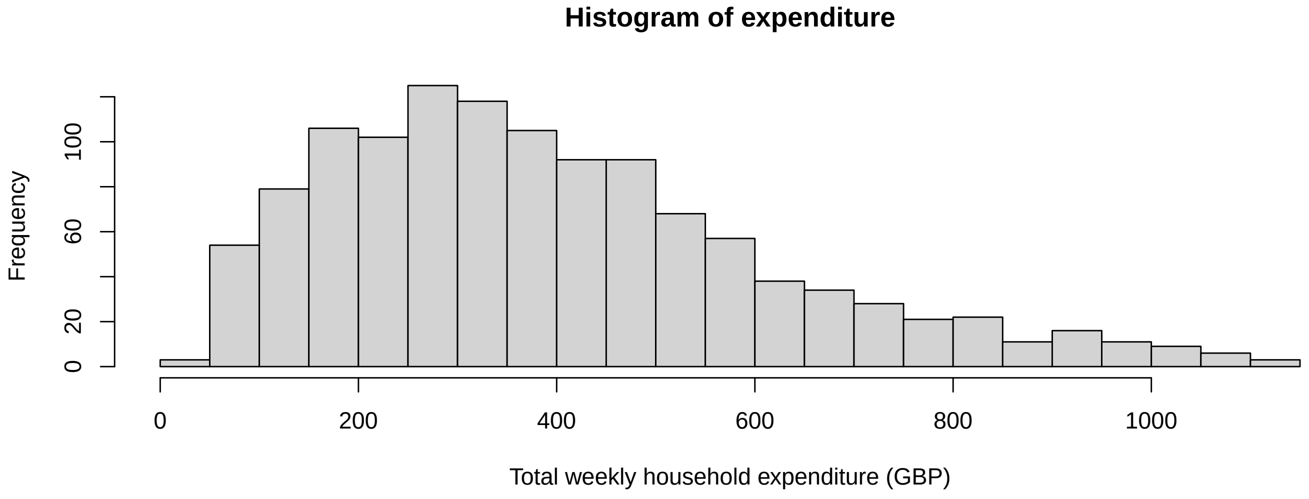
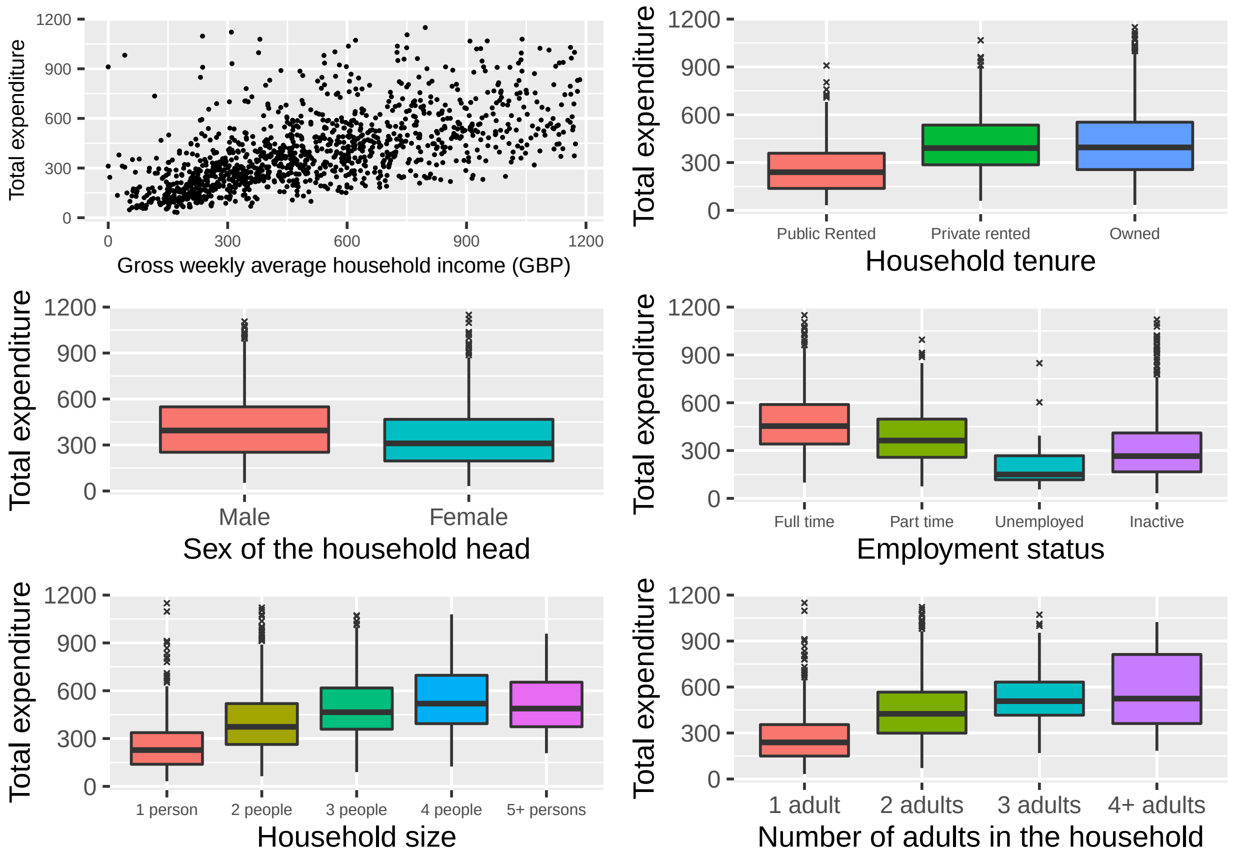
General linear Models Coursework

Student ID: 34273638

Task 1:

Part 1: Assessing distribution of expenditure and relationship between it and covariates





To assess the distribution of expenditure I produced the above histogram, it shows that expenditure has a positively skewed distribution (and thus the normality assumption used in linear regression may be invalid). Using a scatter plot to investigate the relationship between household income and expenditure , it shows there is a clear positive association. Then to investigate the other categorical variables I used boxplots to display differences in the expenditure across the different levels (categories) of each variable. The differences shown in the plots are listed below:

- Different household tenures categories have different weekly expenditure amounts with public rented households having a noticeably lower expenditure compared to the other 2 household levels.
- Households with a female head have a slightly lower weekly expenditure compared to male household heads
- Employment status appears to be associated with expenditure with unemployed households having a dramatically lower expenditure compared to households with full or part time employment status and inactive employment status households having a lower expenditure compared to households with full or part time employment status.
- Household size generally appears to be positively correlated with expenditure, with larger households having a higher weekly expenditure (however 5+ person households break this trend with the expenditure being slightly decreasing, however this could be due to other variables or a small sample size of such households).
- Number of adults in the household also appears positively correlated with expenditure with households containing more adults having a higher expenditure.

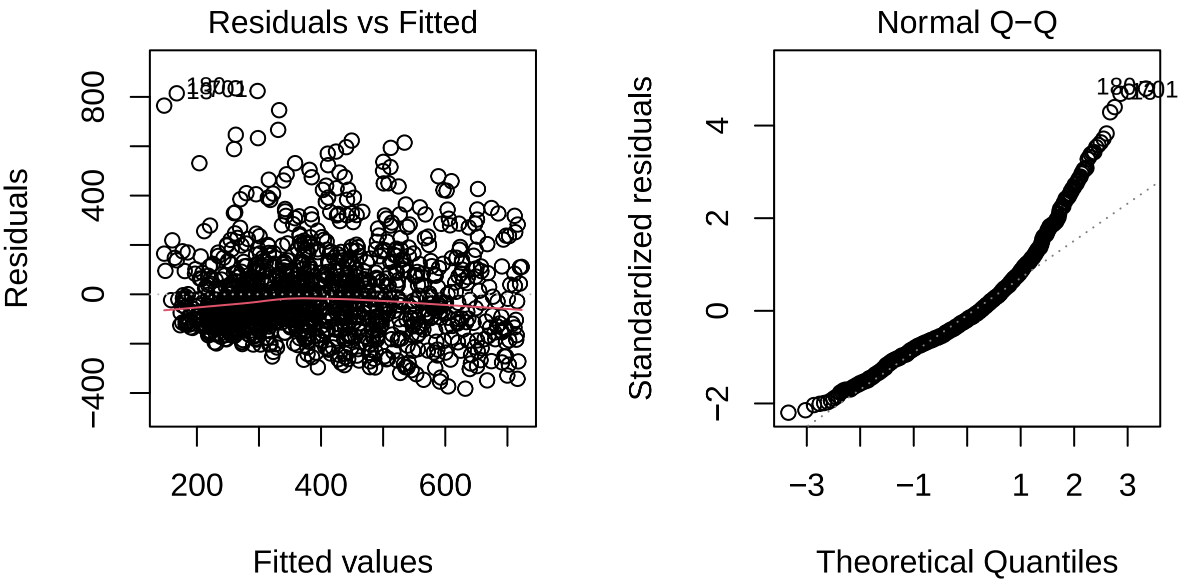
I note that these differences and possible associations may not be due to the variable mentioned but due to a third variable (a confounder) that affects both the categorical variable and expenditure (the response).

Part 2: Regressing expenditure on income

Fitting the model $y_i = \beta_0 + \beta_1 x_i + \epsilon$ where y is the weekly household expenditure and the explanatory variable x is weekly income and $i \in \{1, ..., 1200\}$ is the ith household in the dataset produces the below table (these are the same for all linear models mentioned in parts 2 to 6).It shows the estimated values of the intercept (β_0) and the estimated 'effect' of a 1 pound increase of income on expenditure (β_1) and their associated standard errors.

Table 1: Estimated coefficients and SE for part 2 model

term	**estimate**	**std.error**
(Intercept)	147.224	10.386
income	0.486	0.018



Viewing the above diagnostic plots for this model. The fitted vs residual plot shows the residuals are not equally spread around 0 with a greater spread of residuals above zero, meaning the assumption of linearity may not hold. Furthermore, the QQ normal plot has a positive U shape with the tails diverging heavily away from the straight line. This implies the normality assumption is invalid.

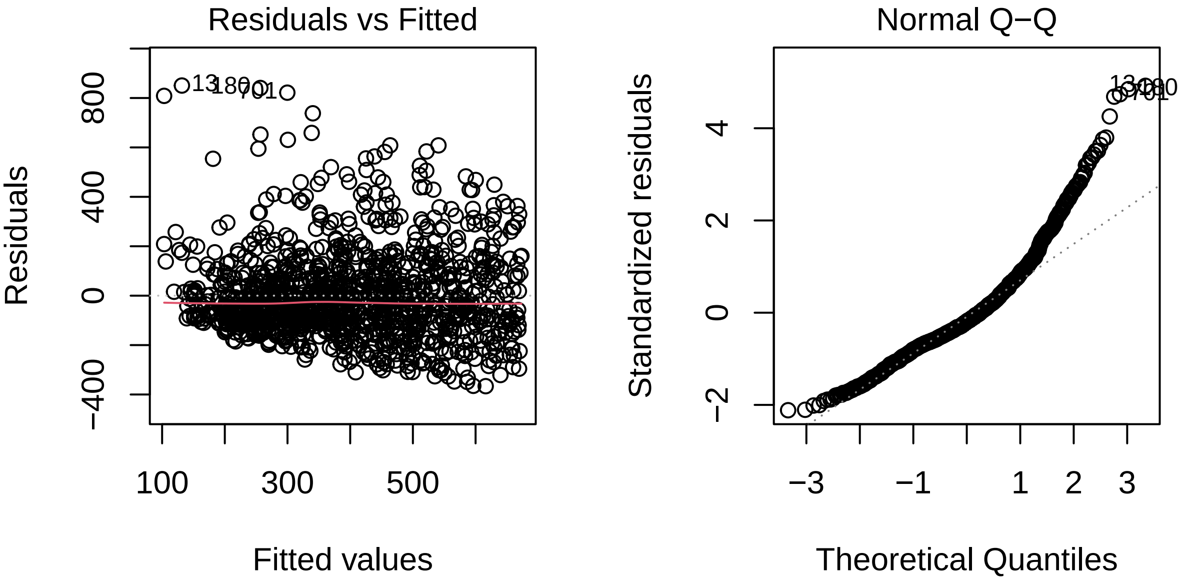
Part 3: Regressing expenditure on income and income squared

Fitting a new linear model $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon$ with an added squared term of income gives the below estimated for the coefficients and their standard errors.

Table 2: Estimated coefficients and SE for part 3 model

term	**estimate**	**std.error**
(Intercept)	103.175	18.051
income	0.690	0.071
I(income^2)	0.000	0.000

Looking at the produced diagnostic plots below for this model, in the fitted vs residual plot the residual values increase as the fitted values increase thus the homoscedasticity assumption may not be valid. The tails in the normal QQ plot also still deviate heavily away from the straight line meaning the normality assumption is invalid.



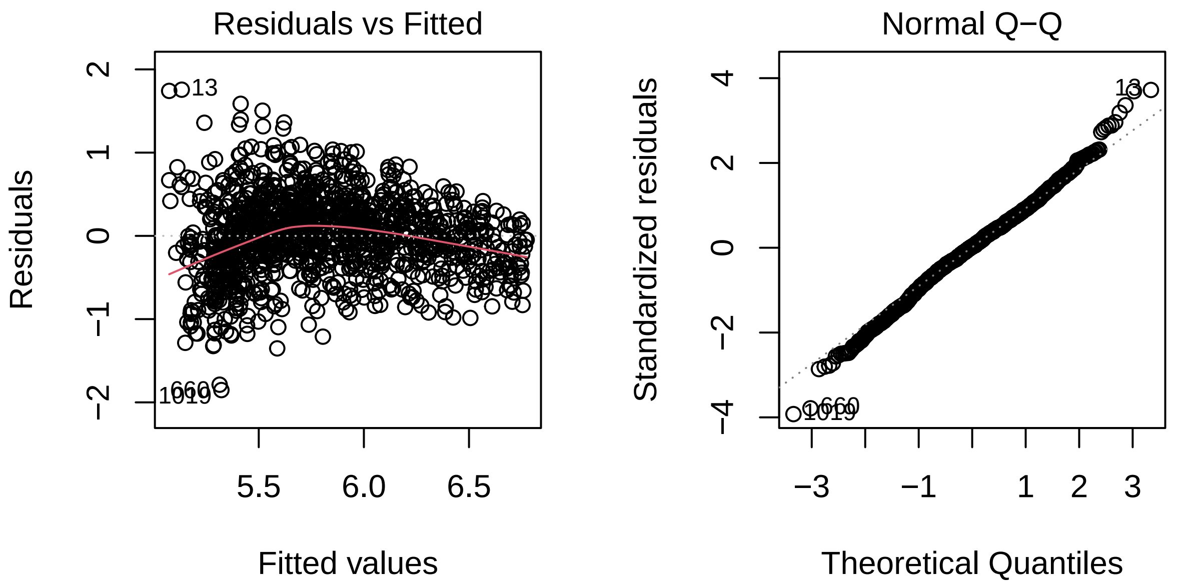
Part 4: Regressing the natural logarithm of expenditure on income

Fitting the same model as in task 2 but with the response of expenditure logged (a log-linear model) $log(y_i) = \beta_0 + \beta_1 x_i + \epsilon$ gives the below estimated coefficients and standard errors.

Table 3: Estimated coefficients and SE for part 4 model

term	**estimate**	**std.error**
(Intercept)	5.074	0.028
income	0.001	0.000

Looking at the produced diagnostic plots above for this model, the data appears to be heteroscedastic as in the fitted vs residual plot the variance in the residuals decreases as the fitted values increases. Furthermore, the linear assumption also appears violated as the residuals appear to exhibit a slight negative quadratic shape. The QQ plot shows the residuals stay close to the straight line thus normality assumption is valid.

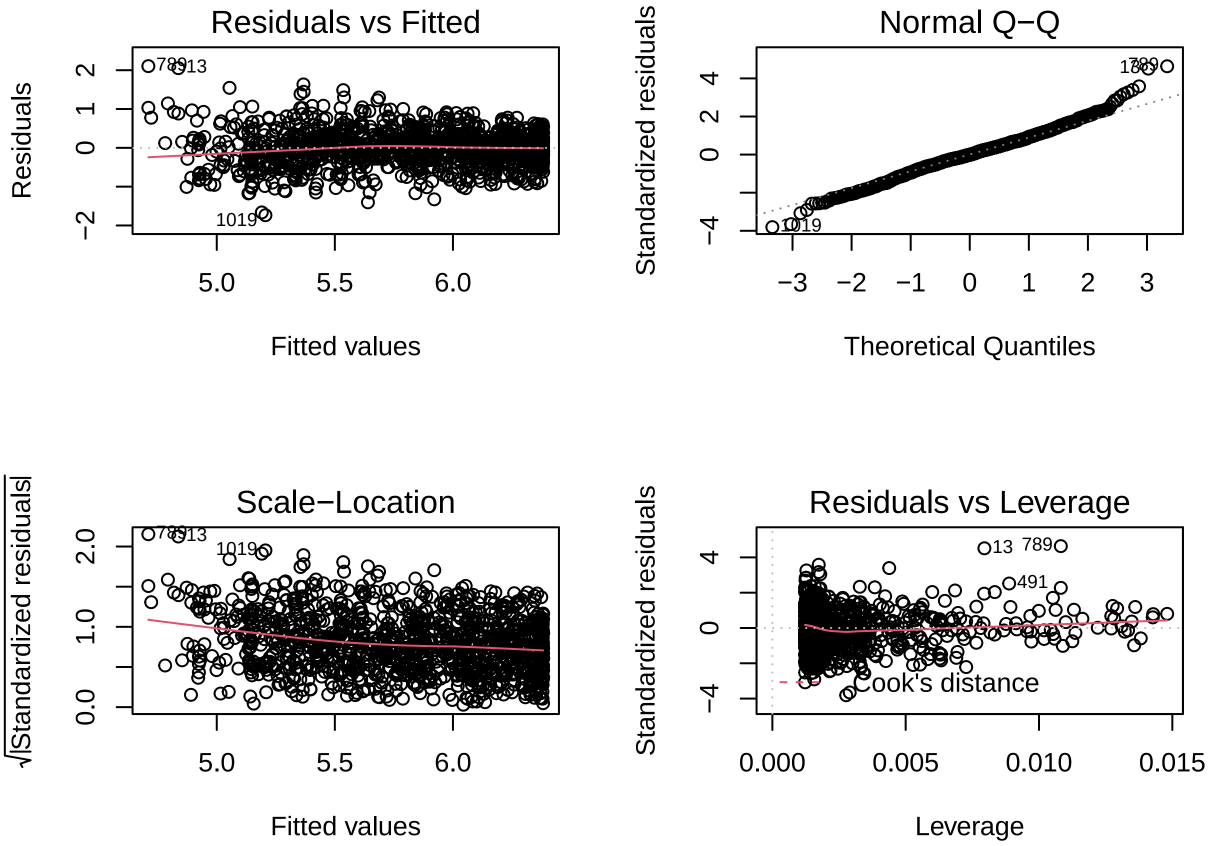


Part 5: Regressing the natural logarithm of expenditure on income and income squared

Fitting the model $log(y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon$ gives the estimated coefficients and standard errors in the table below.

Table 4: Estimated coefficients and SE for part 5 model

term	estimate	std.error
(Intercept)	4.710	0.047
income	0.003	0.000
I(income^2)	0.000	0.000



Producing the diagnostic plots below, the fitted vs residuals plot shows a fairly even scattering of points around 0 and a linear trend, however the variance of the residuals appears to decrease slightly as the fitted values increase. But, looking at the scale-location plot the residuals appear randomly spread around the red line and the red line through them stays fairly horizontal thus homoskedasticity assumptions appears valid. The QQ plot shows the points stay close to the straight line meaning the normality assumption is valid. The residual vs Leverage plot shows no influential points.

Part 6: Comparing models

Table 5: Comparing the AIC and adjusted R squared of models from part 2-5

Model	AIC	Adj.R.squared
Part 2	15792.802	0.384
Part 3	15785.940	0.388
Part 4	1610.637	0.425
Part 5	1528.062	0.464

The model from part 5, $log(y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon$ has the lowest AIC value of **1528.062** and the highest adjusted R squared value of **0.4635055**. Furthermore, from the diagnostic plots the normality assumption is valid and the fitted vs residual plot displays an even scattering of points around 0. Therefore the 4th model is my preferred model to describe the relationship between expenditure and income.

Table 6: Estimated coefficients and SE for part 5 model

term	estimate	std.error
(Intercept)	4.710	0.047
income	0.003	0.000
I(income^2)	0.000	0.000

The log-linear model means that the covarites (income and income squared) have a multiplicative effect on the response (expenditure). The squared term of income included in the model, means that the change in the mean of expenditure now depends on the value of income. Income has a positive association with expenditure with an increase in income causing a percent increase in expenditure. With the initial value of income affecting the magnitude of the percent increase on expenditure. Smaller initial values of income cause a greater percent increase on expenditure (due to the squared income term being negative). In particular, using a realistic increase in weekly income of £20 for a household earning the mean income of £512 gives an increase of $(\frac{exp(532\beta_1+532^2\beta_2)}{exp(512\beta_1+512^2\beta_2)} - 1)$x100% = **3.25%**.

Part 7: Finding a suitable regression model for expenditure

To start I investigated the multicolinearity of the variables in the dataset.

```
##                   GVIF Df GVIF^(1/(2*Df))
## income      21.191489  1        4.603421
## I(income^2) 17.959744  1        4.237894
## hh.size      6.971226  4        1.274717
## hh.adults    7.166821  3        1.388527
## sex.hh       1.222627  1        1.105725
## lab.force    1.837640  3        1.106734
## house.ten    1.283905  2        1.064469
```

None of the variables had a GVIF greater than 5, thus multicolinearity is not an issue with the explanatory variables. However, considering the variables it seems likely that household size is related to number of household adults.

```
##
##                1 adult    2 adults    3 adults   4+ adults
##   1 person   0.831223629 0.000000000 0.000000000 0.000000000
##   2 people   0.101265823 0.680129241 0.000000000 0.000000000
##   3 people   0.046413502 0.130856220 0.602409639 0.000000000
##   4 people   0.018987342 0.127625202 0.240963855 0.750000000
##   5+ persons 0.002109705 0.061389338 0.156626506 0.250000000
```

From the proportional table above there appears to be a relationship between the two variables. To investigate this further I created linear models including each variable separately and both together added to the model from part 5. In the model with only the number of adults variable added using an F-test the variable was significant at $\alpha = 0.05$ with a p-value of **2.271e-11**. However, when adding both number of adults and household size only household size was significant with a p-value of **7.953e-15** and number of adults was no longer significant with a p-value of **0.7249** . Likely the household size was is confounding the effect of number of adults. Thus, I will include the size of a household in for the final model but not the number of adults.

Considering the possible interactions, I decided to only include 2 way interactions as higher order interactions make interpretation of the model much more complicated and there is no clear 3 way interactions that make sense.

First I investigated interactions between income and the factor variables. I added each separately to the model from part 5 and the variable part of the interaction and used an F-test to see if the interaction was significant at the 5% level. income:labour force, income: household size and income: house tenure type were all significant, however income:sex was not. Therefore I will consider the 3 significant interactions for my final model.

Next considering factor variable interactions. There are 6 possible factor interactions, of which I added each to the chosen model separately (plus the variables which were interacting) only sex:household size and labour:tenure were significant at the 5% level. Therefore, these are the only 2 I will consider for the final model.

To choose my final model I will start with a 'full' model that includes all variables that I want to consider. Then using AIC and adjusted R^2 as my model selection criteria I will remove explanatory variables (those that cause the AIC to become lower or increase the adjusted R^2 when removed). After this as the model is for exploratory analysis and not prediction a simpler model that is easier to interpret/ understand is preferable (Occam's razor), therefore if the model can simplified to only include the important variables and this does not significantly affect the AIC and R^2 this further reduced model will be my final model.

My 'full' model regresses the natural logarithm of expenditure on income, income squared, household tenure type, household size, sex of household head, labour force and interactions between income with tenure type, labour force and household size and also interactions between sex of household head and household size and labour force and tenure type. This can be fitted in R using the belwo code:

```
lm.log.full<-lm(log(expenditure)~income*lab.force+I(income^2)+income*house.ten+
                  income*hh.size+sex.hh*hh.size+lab.force:house.ten,data=expend.df)
```

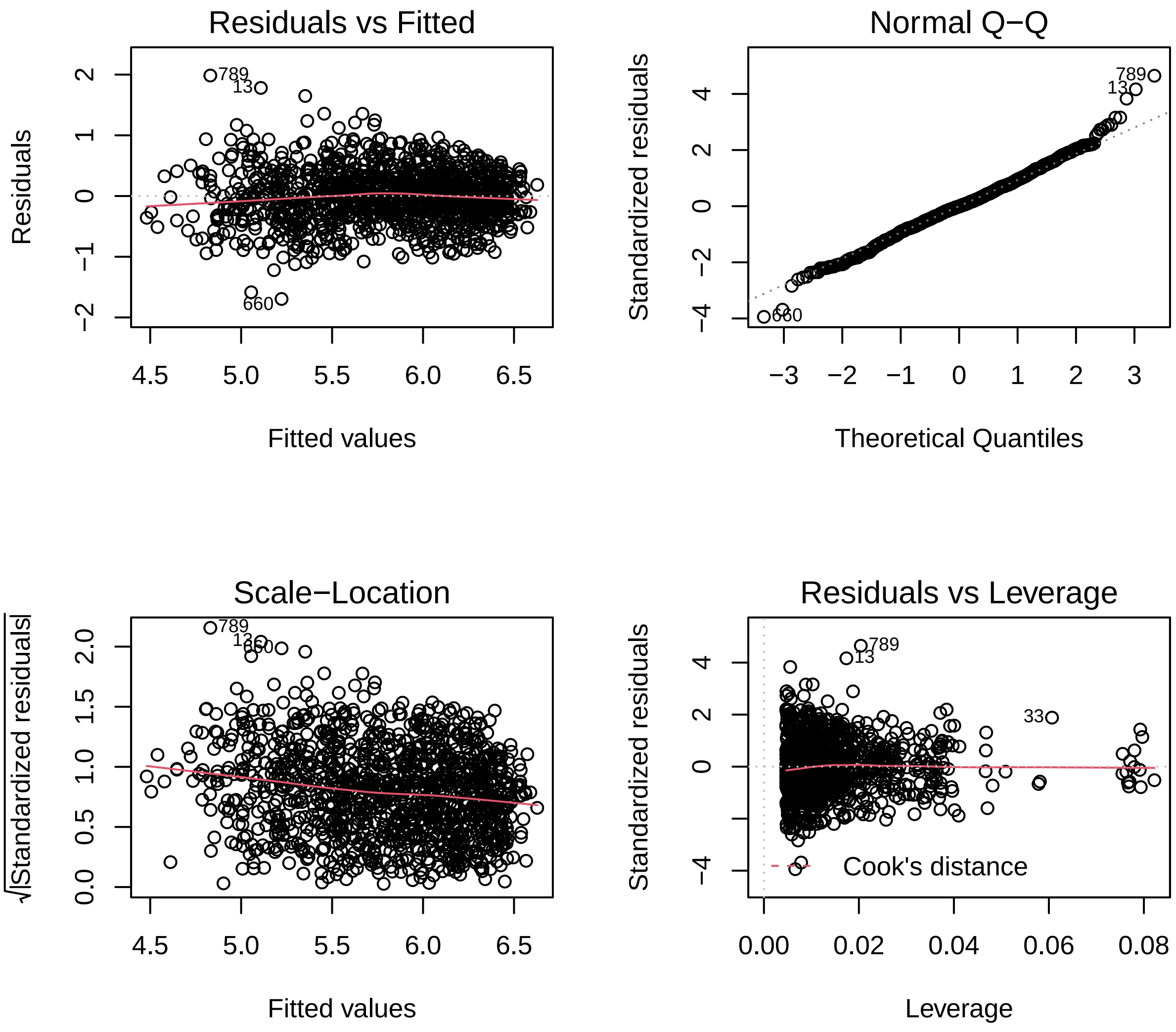
Then using the iterative method of deleting the variable that will reduce the AIC by the largest amount until no decrease in AIC is available. Implementing this algorithm I removed 1 variable, the interaction between labour force and household tenure type as this reduced the AIC from **1391.914** to **1386.878**. All variables left in the updated model are also statistically significant except for income:house tenure.

I then checked how important (the magnitude of impact on the response) the variables left in the model are. To find their impact on expenditure I exponentiated the estimated coefficients and note that due to log-linear model the effect is multiplicative.

term	estimate
income:lab.forcePart time	1.000
income:lab.forceUnemployed	1.001
income:lab.forceInactive	1.000
income:house.tenPrivate rented	1.000
income:house.tenOwned	1.000
income:hh.size2 people	1.000
income:hh.size3 people	1.000
income:hh.size4 people	1.000
income:hh.size5+ persons	1.000

Above is a table of the exponentatied estimated coefficients of all the variables which were approximately 1. These are all the possible interaction levels (excluding the dummy level) of the income interactions. Thus all income interactions are not meaningful/important as their exponentiated value is approximately 1. Therefore, their effect on expenditure is negligible so I remove them from the model as a simpler model is preferred. After removing these 3 variables, removing any more does not further decrease the AIC. Furthermore the difference in AIC between the simpler and more complicated model with the simple models AIC being **1406.732** compared to the more complicated model AIC of **1386.878**. Thus this simplified model is my final model.

Producing diagnostic plots for this final model (seen below) shows no issues with the model fit. The residuals display a random scattering of points indicating the linearity assumption has been met and the residuals variance does not appear to increase/decrease except for a few values but I don't think this is enough evidence to say that the homoscedasticity assumption is invalid. This is further backed up by a straight nearly horizontal red line in the scale-location plot with an even random scattering of residuals around it. The points follow a straight line in the QQ plot with the tails deviating slightly but not extreme enough to indicate the normality assumption is invalid. The cooks distance plot shows no influential points are present in the data.



Part 8: Inference of final model

Let y denote the expenditure of the household, x the income of the household, α^G the sex of the household head, α^S the size of the household, α^T the tenure type, α^L the labour force and α^{SG} the interaction between sex of household head and size of household. Then the final model is:

$$log(y_{ijkwz}) = \beta_0 + \beta_1 x_{ijkwz} + \beta_2 x^2_{ijkwz} + \alpha^G_j + \alpha^S_k + \alpha^T_w + \alpha^L_z + \alpha^{SG}_{jk} + \epsilon_{ijkwz}$$

Where $\alpha^G_M = \alpha^S_1 = \alpha^T_{Priv} = \alpha^L_{FT} = \alpha^{SG}_{Mk} = \alpha^{SG}_{j1} = 0$, $j \in \{M, F\}, k \in \{1, 2, 3, 4, 5+\}, w \in \{Pub, Priv, Own\}, z \in \{FT, PT, U, I\}$ and $i \in \{1, ..., n_{jkwz}\}$ with n_{jkwz} number households with the j^{th} sex of household head, k^{th} number of persons in the household, w^{th} tenure type and z^{th} labour force type.

Due to the model being a log-linear model the explanatory variables have a multiplicative effect on the weekly expenditure. Exponentiating the estimated coefficients to find their effect on expenditure gives the below [*Table 7*]. I note that all values have been rounded to 3 decimal places thus the zero p-values are due to p-values being too small be displayed and similarly the effect of income squared is too small to be shown. My following inference will be based on this table and I use a significance level of 5% for all results.

Table 7: Summary of exponentiated variables affecting expeniture

term	**estimate**	**Confidence.Interval.95.**	**p.value**
(Intercept)	99.520	(87.152 , 113.644)	0.000
income	1.002	(1.002 , 1.003)	0.000
lab.forcePart time	1.012	(0.926 , 1.105)	0.799
lab.forceUnemployed	0.777	(0.659 , 0.915)	0.003
lab.forceInactive	0.950	(0.889 , 1.014)	0.122
I(income^2)	1.000	(1 , 1)	0.000
house.tenPrivate rented	1.326	(1.22 , 1.441)	0.000
house.tenOwned	1.195	(1.115 , 1.281)	0.000
hh.size2 people	1.334	(1.224 , 1.455)	0.000
hh.size3 people	1.530	(1.362 , 1.719)	0.000
hh.size4 people	1.480	(1.313 , 1.668)	0.000
hh.size5+ persons	1.565	(1.349 , 1.816)	0.000
sex.hhFemale	1.153	(1.056 , 1.259)	0.002
hh.size2 people:sex.hhFemale	0.906	(0.803 , 1.023)	0.113
hh.size3 people:sex.hhFemale	0.750	(0.634 , 0.887)	0.001
hh.size4 people:sex.hhFemale	0.968	(0.802 , 1.167)	0.730
hh.size5+ persons:sex.hhFemale	0.858	(0.651 , 1.13)	0.277

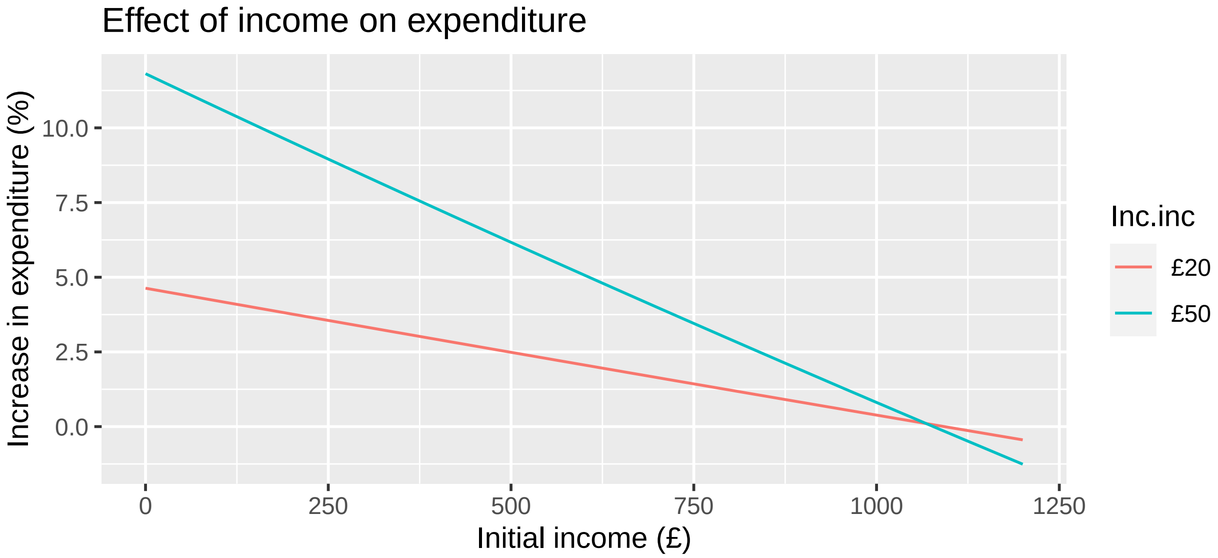
First, interpreting the intercept value, it is for a male headed household of 1 person in full time employment in a publicly rented property who has an income of £0. Clearly this does not make sense (having a job which pays no income).

The difference between full time work and being unemployed was highly statistically significant. With unemployed households compared to full time employment households having a 95% CI of [0.659, 0.915] which corresponds to a **8.5%** to **34.1%** decrease in expenditure with all other variables held constant.

However the difference between full time work and being inactive or in part time work was not significant (the 95% CI included 1) therefore with all other variables constant there is no evidence that being in one of these 3 factors compared to another of influences the expenditure of the household.

The difference between publicly rented properties expenditure and either privately rented or owned properties expenditure were statistically significant. With the associated 95% CI of [1.22, 1.441] and [1.115 , 1.281] which corresponds to a **22%** to **44.1%** and a **11.5% 28.1%** increase in expenditure respectively with all other variables held constant.

Due to income having a squared term the effect of an increase in income on expenditure depends on the initial value of income. To demonstrate this I produced the plot below showing how the percent increase in expenditure varies with the initial income for several different increases in income. For all but the most extreme weekly incomes (over £1000) income has a positive effect on expenditure.



The final 2 variables, sex of household head and household size have an interaction between them. However, only the household size of 3 level has a statistically significant interaction with sex of household head. Therefore as the other levels

The sex of the household head is highly statistically significant with being female head increasing the expenditure by (*5.6%,25.9%*) 95% CI compared to a male head with all other variables held constant. Household size is extremely statistically significant with an increase in expenditure of (*22.4%,45.5*), (*31.3%,66.8%*), (*34.9%,81.6%*) 95% CI's for households of size 2,4,5+ when compared to a household of size 1 with all other variables held constant. For the effect on expenditure between a household of size 1 and size 3 because of the significant interaction with sex if the household head is male the CI is (*36.2%,71.9%*) and (*-15.7%,+52.3%*) if female is it is with all other variables held constant. Note that the for the female head CI 0% is included therefore there is no statistically significant difference in expenditure between a household of size 1 with male head and a household of size 3 with a female head with all other variables held constant.

Overall an increase in income, higher household size, being a female head and living in a owned or privately rented increase the % expenditure when compared to a single male headed household living in a publicly rented house (unless a female head with size 3 household). Whereas being unemployed compared to in full time work decreases the % of expenditure.

All the variables mentioned above were statistically significant in their association with expenditure. However, just because they are significant, it does not mean they influence expenditure. There could be confounding or simple chance causing these results. However, some areas of confidence are:

- income, logically if you make more money you have more available to spend
- household size, larger households are more expensive thus require more spending