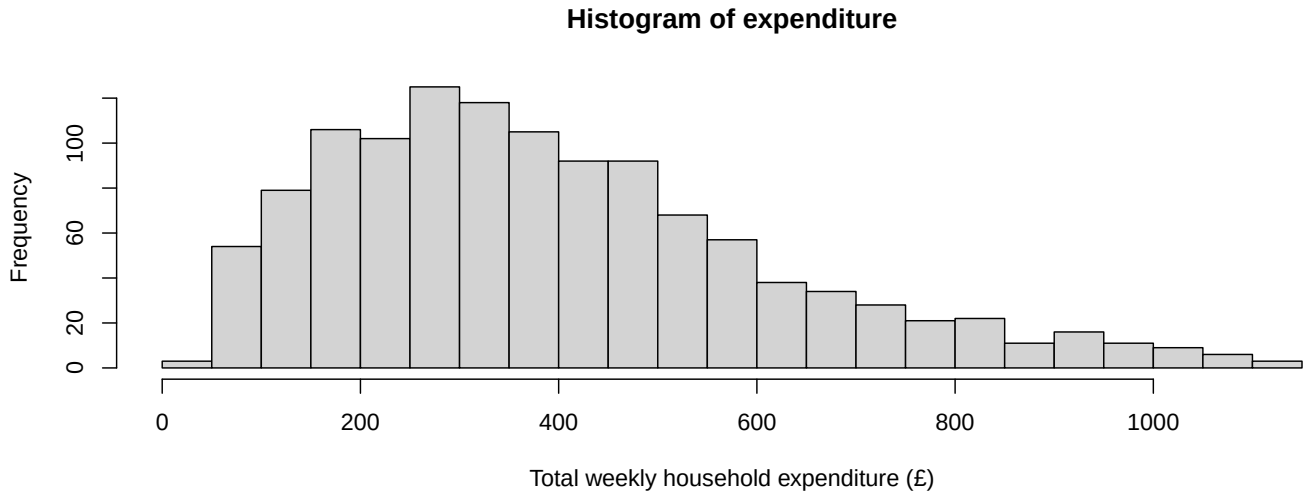
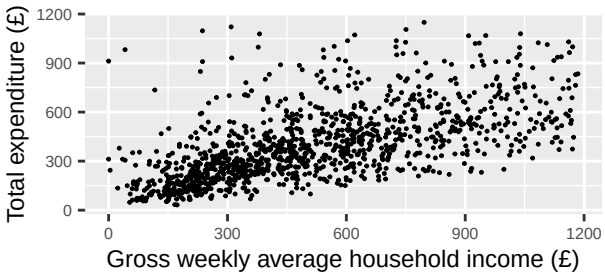
General linear Models (STAT6123) Coursework

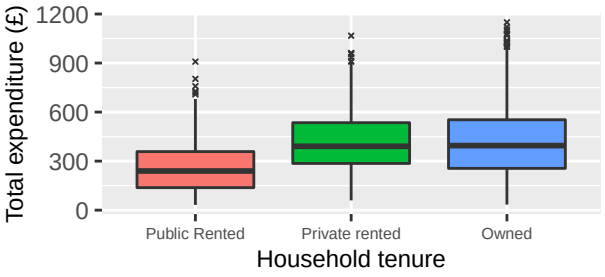
Student ID: 34273638

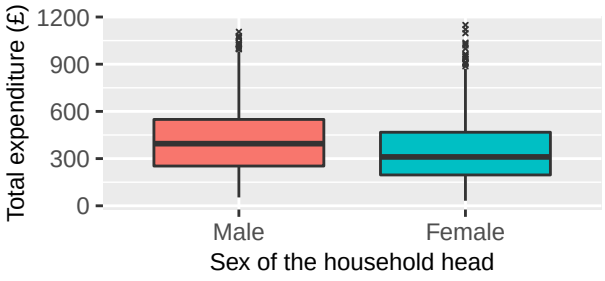
Task 1:

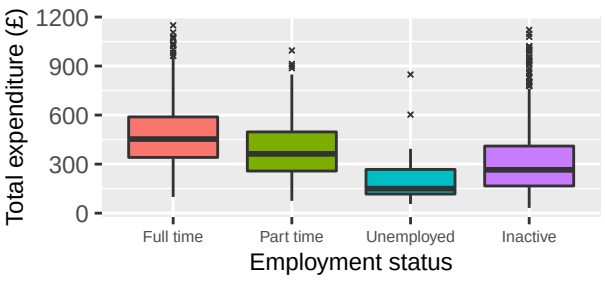
Part 1: Assessing distribution of expenditure and relationship between it and covariates

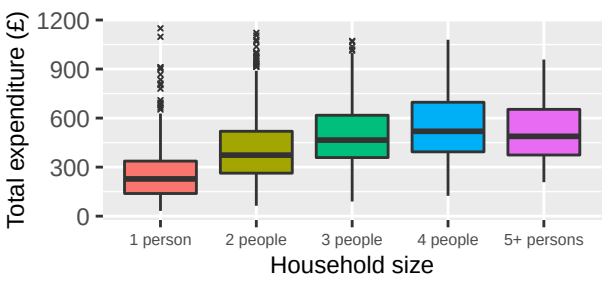


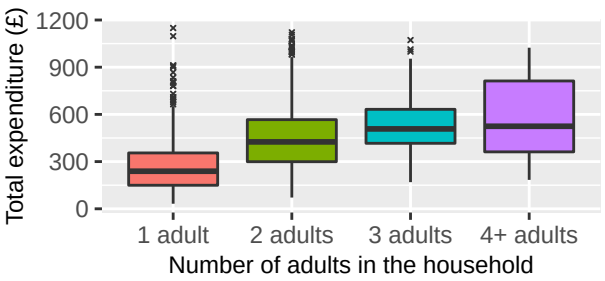












To assess the distribution of expenditure I produced the above histogram, it shows that expenditure has a positively skewed distribution. A larger proportion of households (**56.6%**) spent less than the mean household expenditure of £396. Using a scatter plot to investigate the relationship between household income and expenditure, there is a clear positive association with an increase in income associated with an increase in expenditure. Then to investigate the relationship between the categorical variables and household expenditure I used boxplots to display differences in the expenditure across the different categories of each variable. The differences shown in the plots above are listed below:

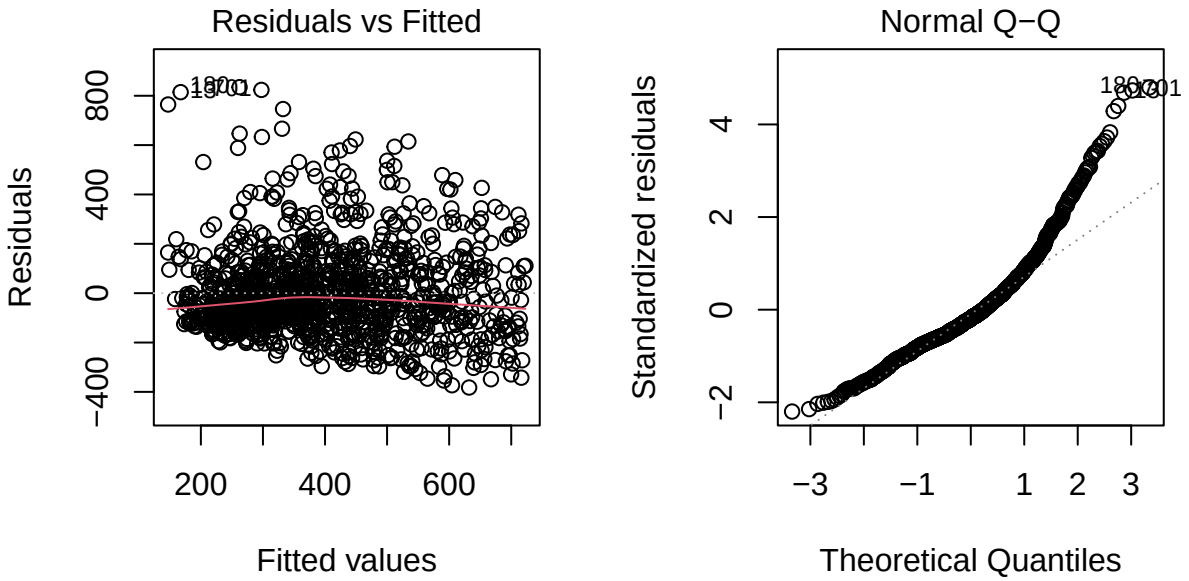
- Different household tenure categories have different weekly expenditure amounts with public rented households having a lower expenditure compared to the other two household tenure categories of owned and privately rented.
- Households with a female household head have a slightly lower weekly expenditure compared to households with a male household head.
- Employment status is associated with expenditure, with unemployed households having a dramatically lower expenditure compared to households with full or part time employment status. Moreover, households with an inactive employment status have a lower expenditure compared to households with full or part time employment status.
- Household size generally appears to have a positive association with expenditure, with larger households having a higher weekly expenditure (however 5+ person households break this trend with the expenditure slightly decreasing, however this could be due to other variables or a small sample size of such households).
- Number of adults in a household has a positive association with expenditure with households containing more adults having a higher expenditure.

Part 2: Regressing expenditure on income

Fitting the model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where the response 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 of the estimated coefficients and their standard errors (SE). I note y and x represent the same variables for all models mentioned in parts 2 to 6.

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 higher density of residuals just below 0 compared with those above 0 (residual distribution is not symmetric about 0). This means the errors are not normally distributed, thus the assumption that the residuals are normally distributed is invalid. Furthermore, the QQ normal plot has a positive U shape with the top tail diverging heavily away from the straight line. This implies the normality assumption is invalid.

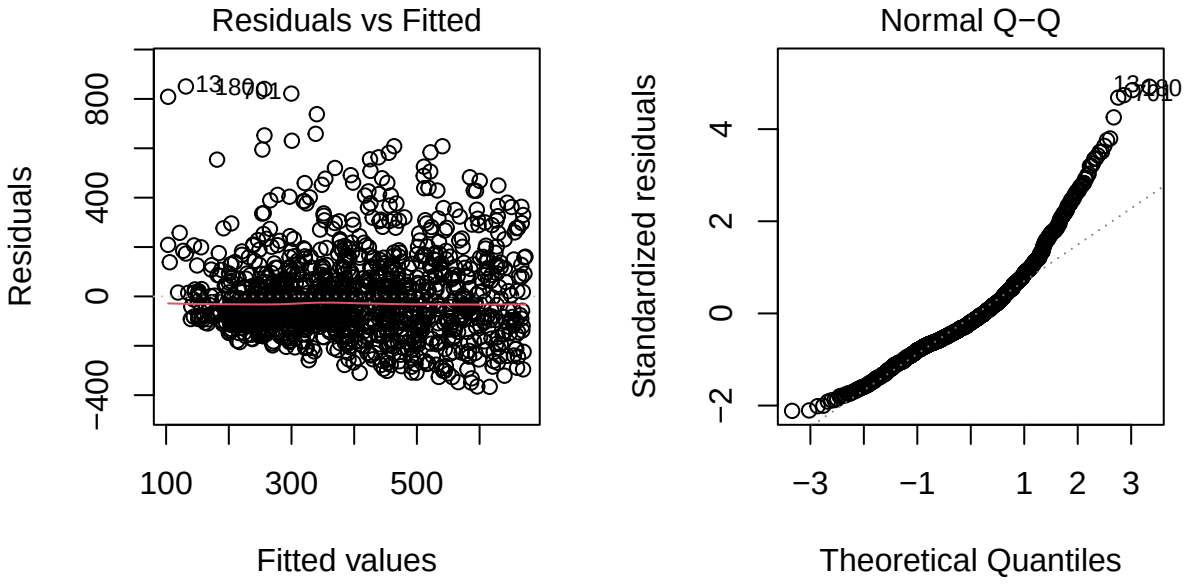
Part 3: Regressing expenditure on income and income squared

Fitting a new linear model $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$ (adding a squared term of income) gives the below estimate for the coefficients and their standard errors.

Table 2: Estimated coefficients and SE for part 3 model

term	estimate	std.error
(Intercept)	103.17533	18.05064
income	0.68975	0.07065
I(income^2)	-0.00018	0.00006

Looking at the produced diagnostic plots below for this model, in the fitted vs residual plot the residual variance increases as the fitted values increase, thus the homoscedasticity assumption is invalid. The tails in the normal QQ plot also still deviate heavily away from the straight line meaning the normality assumption is invalid.



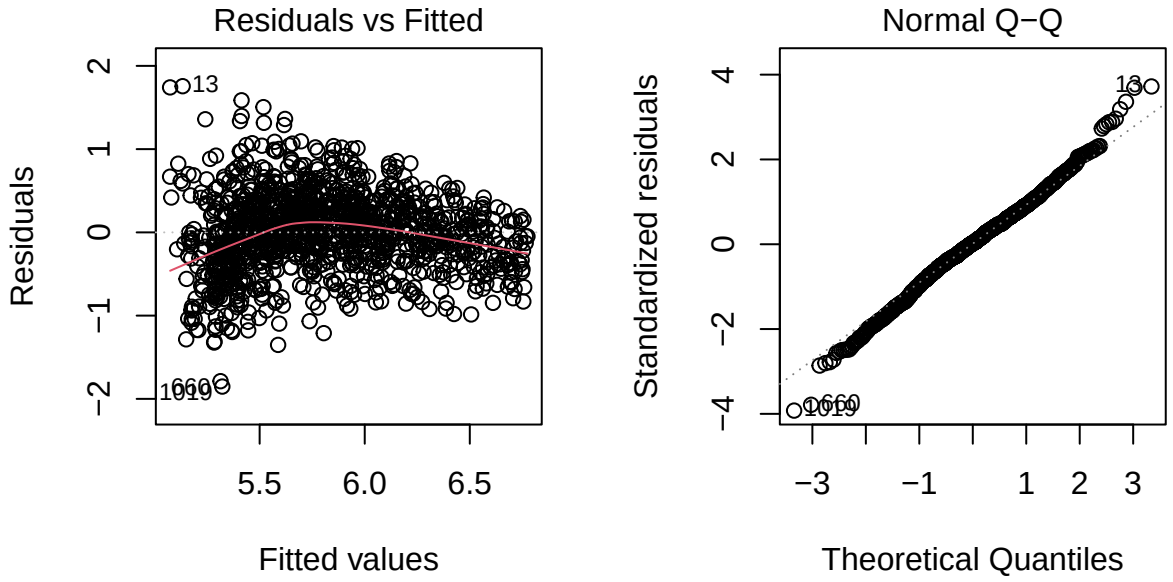
Part 4: Regressing the natural logarithm of expenditure on income

Fitting a log-linear model for expenditure, $log(y_i) = \beta_0 + \beta_1 x_i + \epsilon_i$ gives the below estimated coefficients and standard errors.

Table 3: Estimated coefficients and SE for part 4 model

term	estimate	std.error
(Intercept)	5.07397	0.02819
income	0.00144	0.00005

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

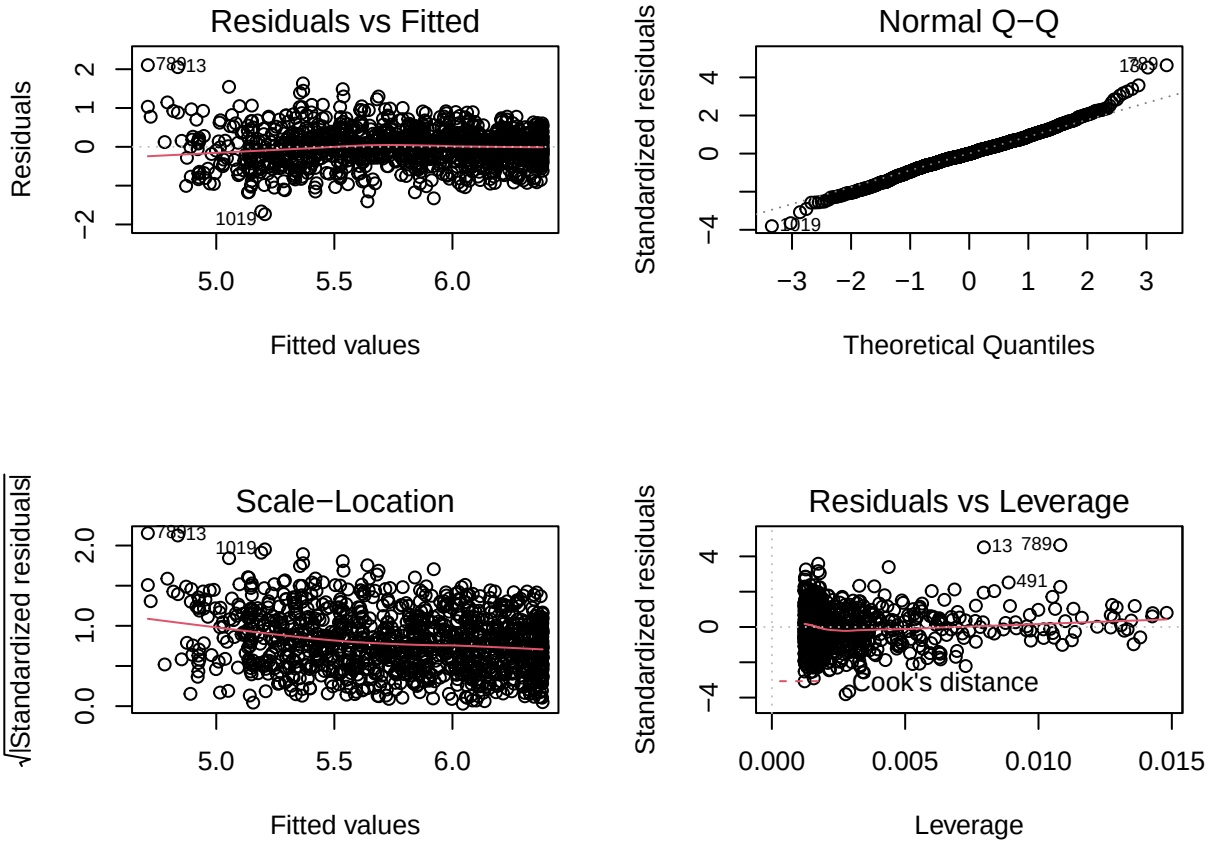


Part 5: Regressing the natural logarithm of expenditure on income and income squared

Fitting the model $log(y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$ 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)	4.7103774	0.0474727
income	0.0031172	0.0001858
I(income^2)	-0.0000015	0.0000002



Producing the diagnostic plots above, the fitted vs residuals plot shows a fairly even scattering of points around 0 and a linear trend, thus the linearity assumption is valid. The scale-location plot shows residuals that are randomly spread around the red line which stays fairly horizontal thus the homoskedasticity assumption is 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 there 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, $log(y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$ has the lowest AIC value of **1528.062** and the highest adjusted R^2 value of **0.464**, thus both model selection criteria select model 5 as the best fitting model of the data. Furthermore, the diagnostic plots find no issues with the model assumptions. Therefore the 4th model from part 5 is my preferred model to describe the relationship between expenditure and income.

The chosen model is a log-linear model which means that the covarites (income and income squared) have a multiplicative effect on the response (expenditure). Both variables were statistically significant at $\alpha = 0.05$. The squared term of income, means that the effect of income on expenditure depends on the initial value of income. Due to the negative value of the squared income regression coefficient the lower the initial value of income the greater the percent increase in expenditure for a given increase in income. Overall, income has a positive association with expenditure in the model with an increase in income causing a percent increase in expenditure (for all but the most extreme initial values of income slightly over £1060). 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%** in household expenditure.

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 the 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 of the number of adults against people in a household there is a clear relationship between the two variables. To investigate this further I created linear models based on the model selected in part 6 which I then added each covariate (household size and number of adults in household) separately and both together. In the model with only the covariate the number of adults 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 the 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** . This is due to the two covariates explaining much of the same variation in the data thus the number of adults becomes insignificant when both are included in the model. Therefore, I will only consider the size of a household for the final model and 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 categorical variables. I added each interaction separately to the model from part 5 and the variables included in the interaction then used an F-test to see if the interaction was significant at the 5% level. Employment status, household size and household tenure type all had significant interactions with income with p-values of **9.417e-05**, **0.002731** and **0.002709** respectively. Therefore, I will consider these 3 significant interactions for my final model.

Next considering the 6 possible 2 way categorical variable interactions. I added each to the chosen model from part 6 separately (plus the variables which were interacting) and found only interactions between sex of the household head and household size, and employment status and household tenure were significant at the 5% level with p-values of **0.017408** and **0.033667**. Therefore, these are the only 2 way categorical interactions 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 my model selection criteria I will remove explanatory variables (those that cause the AIC to become lower 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 be simplified to only include the important variables and this does not significantly affect the AIC this further reduced model will be my final model.

My 'full' model regresses the natural logarithm of expenditure on income, income squared, household tenure, household size, sex of household head, employment status and interactions between income and household tenure, employment status and household size and also interactions between sex of household head and household size and employment status and household tenure. This can be fitted in R using the below 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 I removed 1 variable. This was the interaction between employment status and household tenure type, it reduced the AIC from **1391.914** to **1386.878** when removed. All variables left in the updated model are also statistically significant except for the interaction between income and household tenure.

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

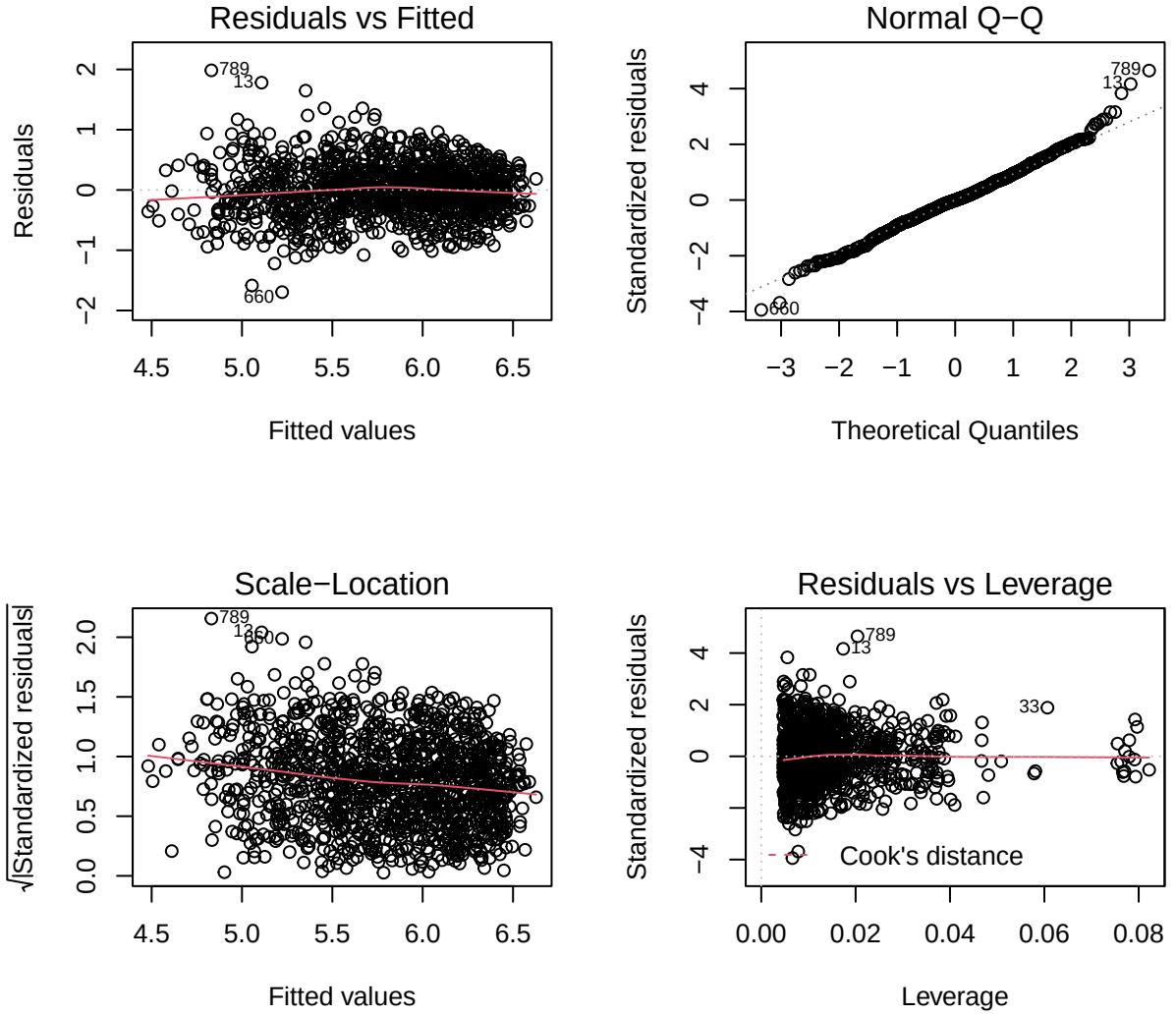
Table 6: Exponentiated coefficents of explantory variables with negliable effect on expenditure

term	estimate
income:lab.forcePart time	1.00031
income:lab.forceUnemployed	1.00089
income:lab.forceInactive	1.00048
income:house.tenPrivate rented	0.99961
income:house.tenOwned	0.99965
income:hh.size2 people	0.99951
income:hh.size3 people	0.99961
income:hh.size4 people	0.99964
income:hh.size5+ persons	0.99967

Above is a table of all the exponentatied regression coefficients which were approximately 1. These were in fact all the possible interaction levels (excluding the dummy level) of the interactions with income. Thus all income interactions are not meaningful/important as their exponentiated value is approximately 1. Therefore, their effect on expenditure is

negligible so I removed 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 is small with the simpler models AIC being **1406.732** compared to the more complicated models AIC of **1386.878**. Thus the simplified model is my final model.

Producing diagnostic plots for this final model (seen below) there is no evidence of issues with the model fit. The residuals display a random scattering of points indicating the linearity assumption has been met. The straight, nearly horizontal red line in the scale-location plot with an even random scattering of residuals around it indicates no issue with homoscedasticity. 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 household tenure type, α^L the employment status 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 $\alpha^G_M = \alpha^S_1 = \alpha^T_{Priv} = \alpha^L_{FT} = \alpha^{SG}_{Mk} = \alpha^{SG}_{j1} = 0$, $j \in \{M, F\}, k \in \{1, 2, 3, 4, 5+\}, w \in \{Pub, Priv, Own\}, z \in \{FT, PT, U, I\}$ (FT is full time, PT is part time, U is unemployed and I is inactive) 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} household tenure type and z^{th} employment staus type.

Due to the model being a log-linear model the explanatory variables have a multiplicative effect on expenditure. Exponentiating the estimated coefficients to find their effect on expenditure gives the below table. I note that all values have been rounded to 4 decimal places thus the zero p-values are due to the p-values being too small to be displayed and similarly for the effect of income squared. The exponentiated estimate of the income squared regression coefficient is 0.99999896 (8 d.p.) with a 95% confidence interval (CI) of [0.99999866, 0.99999927]. My following inference will be based on this table using a significance level of 5%.

Table 7: Summary of multiplicative effect of variables on expenditure

term	**estimate**	**Confidence.Interval.95.**	**p.value**
(Intercept)	99.5204	(87.1525 , 113.6435)	0.0000
income	1.0023	(1.0019 , 1.0027)	0.0000
lab.forcePart time	1.0115	(0.9263 , 1.1046)	0.7988
lab.forceUnemployed	0.7766	(0.6594 , 0.9147)	0.0025
lab.forceInactive	0.9495	(0.8892 , 1.0139)	0.1222
I(income^2)	1.0000	(1 , 1)	0.0000
house.tenPrivate rented	1.3258	(1.2196 , 1.4413)	0.0000
house.tenOwned	1.1953	(1.1154 , 1.2809)	0.0000
hh.size2 people	1.3344	(1.2241 , 1.4547)	0.0000
hh.size3 people	1.5304	(1.3622 , 1.7195)	0.0000
hh.size4 people	1.4803	(1.3135 , 1.6683)	0.0000
hh.size5+ persons	1.5654	(1.349 , 1.8165)	0.0000
sex.hhFemale	1.1529	(1.0561 , 1.2586)	0.0015
hh.size2 people:sex.hhFemale	0.9063	(0.8025 , 1.0234)	0.1129
hh.size3 people:sex.hhFemale	0.7501	(0.6342 , 0.8872)	0.0008
hh.size4 people:sex.hhFemale	0.9675	(0.8021 , 1.167)	0.7298
hh.size5+ persons:sex.hhFemale	0.8580	(0.6512 , 1.1305)	0.2767

The difference between households in full time employment and those unemployed was highly statistically significant. With unemployed households compared to full time employment households having an exponentiated 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. Furthermore, as the CI's of unemployed and part-time employment do not overlap there is a statistically significant increase in expenditure if a household is in part time employment compared to unemployed households with all other variables held constant. However, there was no evidence to support a difference in expenditure between full time employment households and those that were inactive or in part time work (the exponentiated 95% CI's included 1).

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

Sex of household head and household size have an interaction between them and therefore cannot be reported separately. To easily display the effects of both on expenditure simultaneously I created the below 2x2 table which displays the multiplicative effect of sex and household size on expenditure.

Table 8: 95% CI of multiplicative effect on household expenditure of sex of household head and household size compared to a male headed household of 1 person

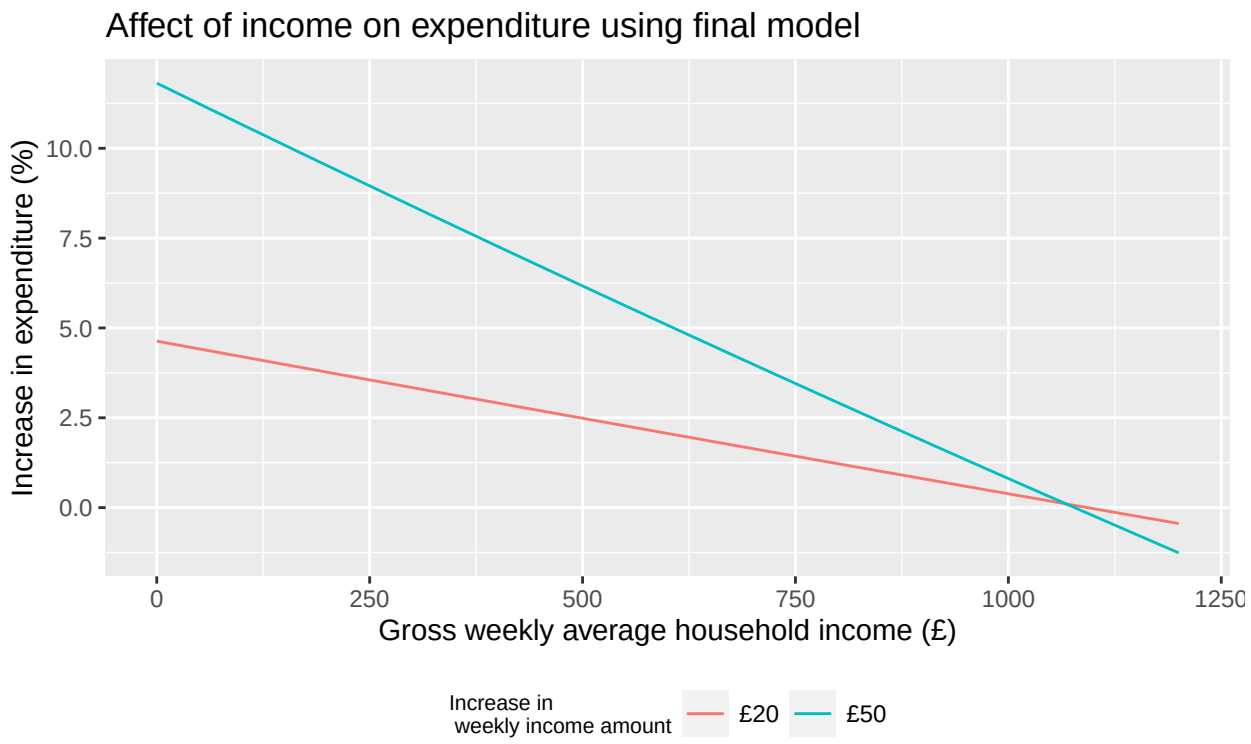
	1 Person	2 Persons	3 Persons	4 Persons	5+ Persons
Male.head	(1, 1)	(1.224, 1.455)	(1.362, 1.719)	(1.313, 1.668)	(1.349, 1.816)
Female.head	(1.056, 1.259)	(1.037, 1.874)	(0.912, 1.92)	(1.113, 2.45)	(0.928, 2.585)

For households of size 1,2 and 4 a female household head corresponds to an increase in expenditure compared to a male household of size 1, however for a household of size 3 or 5+ there is no evidence of a difference in expenditure (the

exponentiated CI included 1) with all other variables held constant. I note the confidence intervals are very large for female headed households of size 2 and above, this means there is alot of variation in the data causing this uncertainty in the results. This could be due to the small sample of households with female heads of size 3 and above (insufficient power) or/and additional confounders affecting female household heads that were not included in the analysis.

Compared to a household of size 1, all larger sized households increased expenditure if the household head was male with an increase of expenditure of **22.4%** to **45.5%**, **36.2%** to **71.9%**, **31.3%** to **66.8%** and **34.9%** to **81.6%** for households of size 2, 3, 4 and 5+ respectively with all other variables held constant. There is no evidence larger households increase expenditure for female headed households as the 95% exponentiated CI's overlap for all household sizes.

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 of a household for several different increases in household income. For all but the most extreme household incomes (slightly over £1000), income has a positive effect on expenditure. With the higher the initial income, the less the percent increase in expenditure for the same increase in income (seen by the negative gradient in the below plot). For an initial income of £512 (the mean income) an increase in income of £20 and £50 increases expenditure by **2.44%** and **6.04%** respectively.



Overall an increase in income (except for households on extremely high incomes), living in an owned or privately rented property compared to a publicly rented property and being in full or part time employment compared to unemployed increases the expenditure of a household. With household sizes larger than 1 increasing income compared to singular households if the household head is male and no proven difference if the household head is female. A female headed household increases expenditure for certain household sizes (1, 2 and 4) and has no proven effect for others (3 and 5+) compared to a male headed household.

It is important to remember that the relationships found in this analysis does not imply causality between the variables and expenditure as they may be affected by confounding, chance or bias from the sample the model was based on. However, some areas of causal confidence are: (1) income, logically if you make more money you have more money available to spend, (2) Employment status, if you are unemployed you are going to be more frugal with purchases. Including additional factors that could potentially be confounding results such as the age of the household head, ethnicity of household and location of household would improve the confidence in the relationships found. Furthermore, a larger sample would enable the effect of the sex of the household head to be better understood for all household sizes.

I note that generalization of these results may not be possible as different countries, cultures and ethnicities may have different attitudes on spending. As there was no information on where/when the original sample was drawn from and the process for the samples selection I cannot recommend what groups/areas these results may be representative of.

Task 2

1. First I note that $f(y;\theta) = \theta e^{-\theta y} = exp(log(\theta) - \theta y)$. Therefore, $\alpha(y) = y, \beta(\theta) = -\theta, c(\theta) = log(\theta)$ and $d(y) = 0$. As $\alpha(y) = y$ the distribution is in canonical form. Thus, we can apply the theory from lectures. The distribution of the $Y_i's$ are all of the same form and independent, thus the joint probability density function of $Y_1, ..., Y_n$ where each $Y_i \sim Exp(\theta_i)$ is:

$$f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta}) = \prod_{i=1}^{n} \theta_i e^{\theta_i y_i}$$

As the link function $g(\mu_i) = log(\mu_i) = \eta_i$ and $\eta_i = \beta_0 + \beta_1 x_i \implies \mu_i = e^{\beta_0 + \beta_1 x_i}$. Using the information provided in the question and lecture notes we can derive the following results:

$$Var(Y_i) = \frac{b''(\theta_i)c'(\theta_i) - b'(\theta_i)c''(\theta_i)}{b'(\theta_i)^3} = \frac{0 - \frac{1}{\theta_i^2}}{-1} = \frac{1}{\theta_i^2} = \mu_i^2, \quad \frac{\partial \mu_i}{\partial \eta_i} = \left(\frac{\partial \eta_i}{\partial \mu_i}\right)^{-1} = \left(\frac{1}{\mu_i}\right)^{-1} = \mu_i$$

$$u(\beta_j) = \sum_{i=1}^{n} \left(\frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}\right) = \sum_{i=1}^{n} \left(\frac{(y_i - \mu_i)x_{ji}}{Var(Y_i)}\frac{\partial \mu_i}{\partial \eta_i}\right) = \sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{\mu_i}x_{ji}\right) = \sum_{i=1}^{n} \left(\frac{y_i}{e^{\beta_0 + \beta_1 x_i}} - 1\right) x_{ji}$$

$$I_{jk} = \sum_{i=1}^{n} \frac{x_{ji}x_{ki}}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 = \sum_{i=1}^{n} x_{ji}x_{ki} \implies \boldsymbol{I} = \boldsymbol{X}^T\boldsymbol{X}, \quad \text{For } \boldsymbol{X} \text{ the design matrix}$$

Thus the updating equation to find the m.l.e. of $\boldsymbol{\beta}$ using the Fisher scoring algorithm is:

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{w}^{(m)}$$

Where, $\boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ . & . \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix}$, $\quad \boldsymbol{\beta}^T = (\beta_0, \beta_1), \quad \boldsymbol{w}^T = (w_1, ..., w_n), \quad w_i^{(m)} = \frac{y_i}{exp(\beta_0^{(m)} + \beta_1^{(m)} x_i)} - 1$

2. Using R to implement the above algorithm and with the initial guess $\boldsymbol{\beta}_0^T = (0, 1)$. The point estimates of the coefficients to 3 d.p. are:

$$\boldsymbol{\beta}^T = (-0.121, 0.005)$$

3. The variance-covariance matrix is defined below to 3 s.f. as:

$$(\boldsymbol{I})^{-1} = (\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix} 1.84e-02 & -2.87e-05 \\ -2.87e-05 & 4.72e-08 \end{bmatrix}$$

4. The t-statistic is derived below for β_1:

$$T_i = \frac{\beta_i}{SE(\beta_i)} \implies T_1 = \frac{0.005}{\sqrt{4.72e-08}} = 24.2$$

Where the standard error of β_i is $\sqrt{[\boldsymbol{I}^{-1}]_{ii}}$ (the square root of ith diagonal element in variance-covariance matrix). Conducting a hypothesis test with $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$. We use the large sample distribution of the regression coefficients (where the regression coefficients are represented as $\hat{\beta}_i$):

$$\frac{\hat{\beta}_i - \beta_i}{SE(\hat{\beta}_i)} \sim N(0,1) \implies \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = T_1 \sim N(0,1) \quad \text{under } H_0$$

Therefore setting the significance of the test to $\alpha = 0.05$

$$z_{0.975} = 1.96 < |T_1| = 24.2$$

Therefore we reject the null that β_1 is not significant at a 5% level of significance and thus the data provides evidence that β_1 is significant.