Analyzing Determinants of Health Expenditure in UK Households: A Multivariate Regression Approach*

yifan tong

April 21, 2024

This study explores the factors influencing health expenditure in UK households, a critical issue within the contexts of public health and economic policy. Utilizing data from the Living Costs and Food Survey UK Data Archive, this research employs multiple linear regression models to analyze the impact of demographic, economic, income, housing, occupational, and geographic variables on health expenditure. The analysis reveals a significant positive association between household size and health expenditure, suggesting that larger households face higher health costs. Additionally, the age of the oldest household member and household income significantly predict health expenses, emphasizing the effects of age-related health needs and economic status on healthcare access and affordability. The study also considers the number of rooms in accommodation and occupational class as predictors, reflecting broader socio-economic impacts on health expenditure. Despite a low R-squared value in the models, indicating that other unmeasured factors might also influence health costs, the findings provide valuable insights for policymakers and healthcare providers aiming to enhance healthcare resource management. Future research could further refine these models by considering non-linear relationships and more complex interactions between the variables

Table of contents

*Code and data are available at: LINK.

1 Introduction

The escalating costs of healthcare are a significant concern globally, influencing both public health outcomes and economic stability. In the UK, the intricacies of healthcare expenditure, driven by diverse demographic, economic, and social factors, present a compelling case for detailed analysis. The total healthcare expenditure in the UK reached £280.7 billion in 2021, accounting for 12.4% of the GDP, with a notable impact from the government's pandemic response leading to increased spending in areas like preventive care and pharmaceuticals.

This study addresses the critical question: "What are the factors influencing health expenditure in UK households?" Despite extensive research in the field, gaps remain in understanding the specific contributions of factors like household size, income, and age, particularly in the context of recent economic fluctuations and health crises. By employing a multivariate regression approach using data from the Living Costs and Food Survey, this research dissects the influence of these variables on healthcare spending.

The findings reveal a significant correlation between household size and health expenditure, with larger households incurring higher costs. Moreover, the age of the oldest household member and overall household income emerged as crucial predictors of expenditure, underscoring the influence of age-related health needs and financial capacity on spending. These insights are pivotal for policymakers and healthcare providers, who require a nuanced understanding of spending drivers to allocate resources effectively.

The paper is structured as follows: Section 2 describes the data source, Section 3 presents the regression models and results, Section 4 discusses the implications of these findings, and

recommendations for future research and policy formulation. By integrating detailed statistical analysis with comprehensive data, this study provides a robust foundation for targeted healthcare funding strategies, aiming to enhance the efficiency of resource management in the UK's public health system.

2 Data

2.1 Data Management

This research was conducted using the R statistical programming environment (R Core Team 2023). Data management and analysis were primarily facilitated by the `tidyverse` package (Wickham et al. 2019), with additional support from `broom` (Robinson, Hayes, and Couch 2022) for converting statistical outputs into tidy formats. Tables were styled and managed using `knitr` (Xie 2023) and `kableExtra` (Zhu 2021), while visualizations were created with `ggplot2` (Wickham 2016). The `corrplot` (Wei and Simko 2021) package was utilized for correlation matrices, and regression diagnostics employed the `car` (Fox and Weisberg 2022) and `lmtest` (Zeileis and Hothorn 2022) packages.

2.2 Source

The data for this study was sourced from the Living Costs and Food Survey hosted by the UK Data Service (UK Data Service 2022). This survey provides comprehensive insights into household expenditures, including detailed information on food, living costs, and various demographic and economic factors. It is a valuable resource for analyzing household behavior and economic conditions across the UK.

2.3 Summary Statistics

Because there are as many as 1913 variables in the original data, I only choose some of the variables that are relevant to the research problem. The final data has 5133 samples and 10 variables. Each row in the dataset represents a unique household and its corresponding values for these variables.

1. P606t: Total Health expenditure for children and adults. This is a numeric value.
2. A049: Household size, represented numerically.
3. A071: Sex of the oldest person in the household. This is a nominal value with three categories:
    - 1 for all male,
    - 2 for all female,
    - 3 for mixed sex.

4. p344p: Gross normal weekly household income, top-coded. This is a numeric value.
5. incanon: Anonymised household income and allowances, a numeric value.
6. p493p: Indicates whether the household is wealthy or not, based on anonymised data. This is a nominal value with two categories:
    - 0 for Not wealthy,
    - 1 for Wealthy.
7. a070p: Age of the oldest person in the household, anonymised and numeric.
8. a114p: Number of rooms in accommodation, anonymised and numeric.
9. A094: NS - SEC 8 Class of household reference person. This is a nominal value with various categories representing different occupational classes.
10. Gorx: Government Office Region modified. This is a nominal value with categories representing different regions.

Here are summary tables (Table 1, Table 2) displaying the statistics of numeric variables and categorical variables from dataset.

Table 1: Summary statistics of numerical variables

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
P606t	5133	7.41	39.00	0.00	0.00	0.28	3.81	1505.68
A049	5133	2.36	1.24	1.00	1.00	2.00	3.00	9.00
p344p	5133	723.07	510.63	0.00	335.65	588.22	981.35	2134.09
incanon	5133	739.03	498.28	0.00	368.86	596.54	983.52	2144.51
a070p	5133	54.96	16.12	17.00	42.00	55.00	68.00	80.00
a114p	5133	5.72	1.72	1.00	5.00	6.00	7.00	10.00

These tables provides a clear overview of the data distribution for each variable, helping in understanding the dataset's characteristics and informing potential analyses.

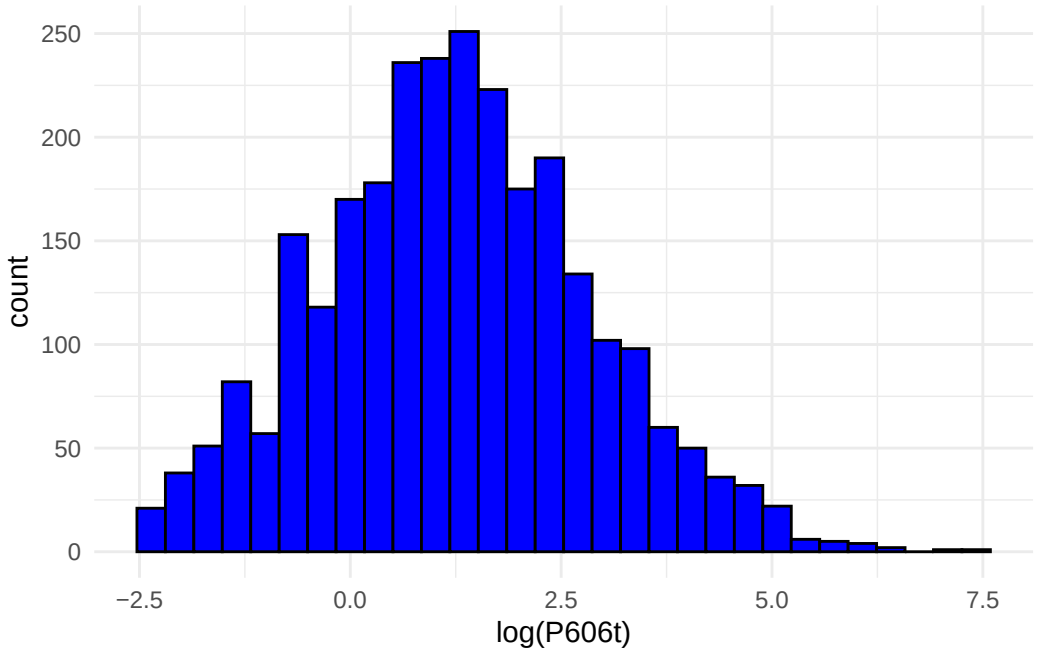
Table 2: Summary statistics of categorical variables

A071	p493p	A094	Gorx
1:2790	0:5121	12 :1610	8 : 740
2:1970	1: 12	3 : 961	2 : 588
3: 373		2 : 445	6 : 498
		7 : 410	5 : 470
		4 : 395	9 : 468
		8 : 351	3 : 459
		(Other): 961	(Other):1910

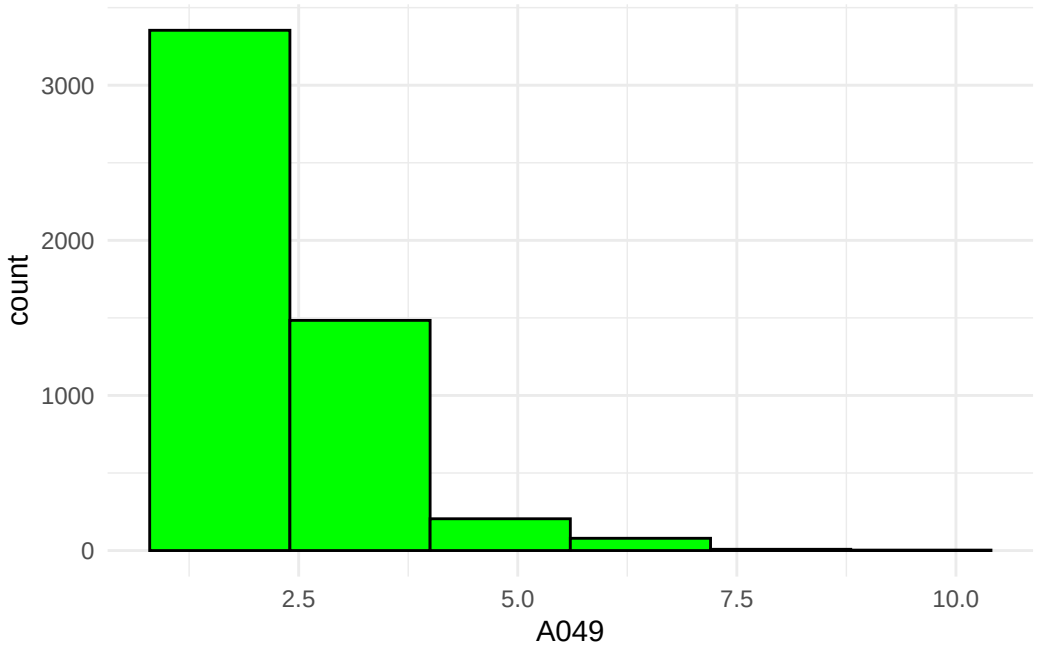
2.4 Data Visualisations

Based on the summary statistics (Table 1, Table 2) and the visualizations (Figure 1, Figure 2, Figure 3, Figure 4) created for each variable in dataset, here are some analyses and insights:

1. Total Health Expenditure (P606t):
    - To make the pictures more meaningful, I use log transformation. The distribution is highly skewed, with most values clustered near zero. This suggests that a majority of households have low health expenditure, but there are a few households with very high health costs.
    - Possible factors influencing this could be differences in health needs, insurance coverage, or access to healthcare services.
2. Household Size (A049):
    - Most households consist of 1 to 4 members.
    - Larger households might have different dynamics in terms of health expenditure, potentially due to more diverse health needs or economies of scale.
3. Weekly Household Income (p344p) and Anonymised Household Income (incanon):
    - Both income distributions are right-skewed, indicating that most households have lower incomes with a smaller proportion having substantially higher incomes.
    - There could be a correlation between income levels and health expenditure, possibly due to better access to healthcare or ability to afford more services in higher-income households.
4. Age of Oldest Person in Household (a070p):
    - The distribution is somewhat uniform but slightly skewed towards older ages.
    - Older age groups might have higher health expenditures due to age-related health issues.
5. Rooms in Accommodation (a114p):
    - Most households have 5 to 7 rooms, which might correlate with household size and indirectly affect health expenditure.
6. Sex of Oldest Person in Household (A071):
    - The gender distribution shows a predominance of either all male or all female as the oldest person in the household.
    - This demographic factor could be explored to see if there's a gender-related pattern in health expenditure.
7. Wealthy Household Indicator (p493p):

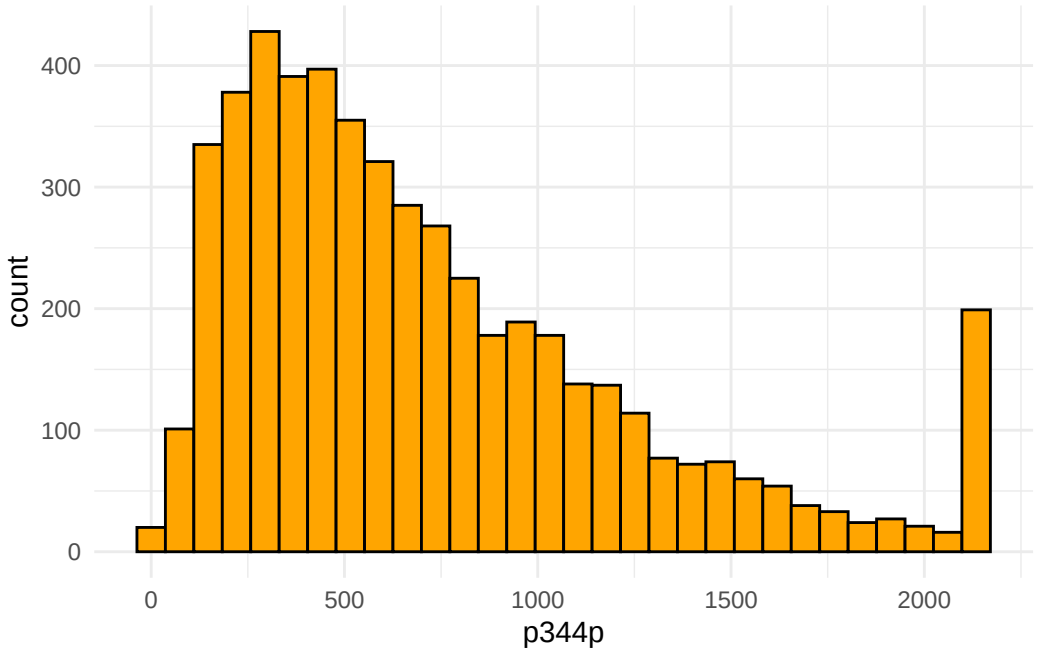


(a) Distribution of Total Health Expenditure log(P606t)

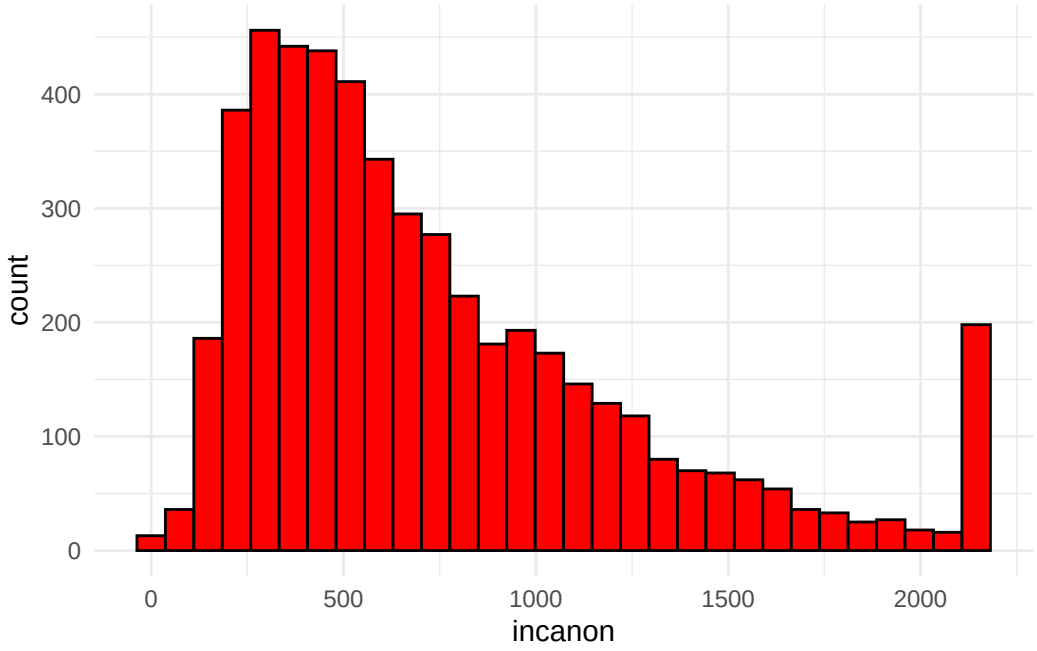


(b) Distribution of Household Size (A049)

Figure 1: Distribution of Total Health Expenditure and Household Size

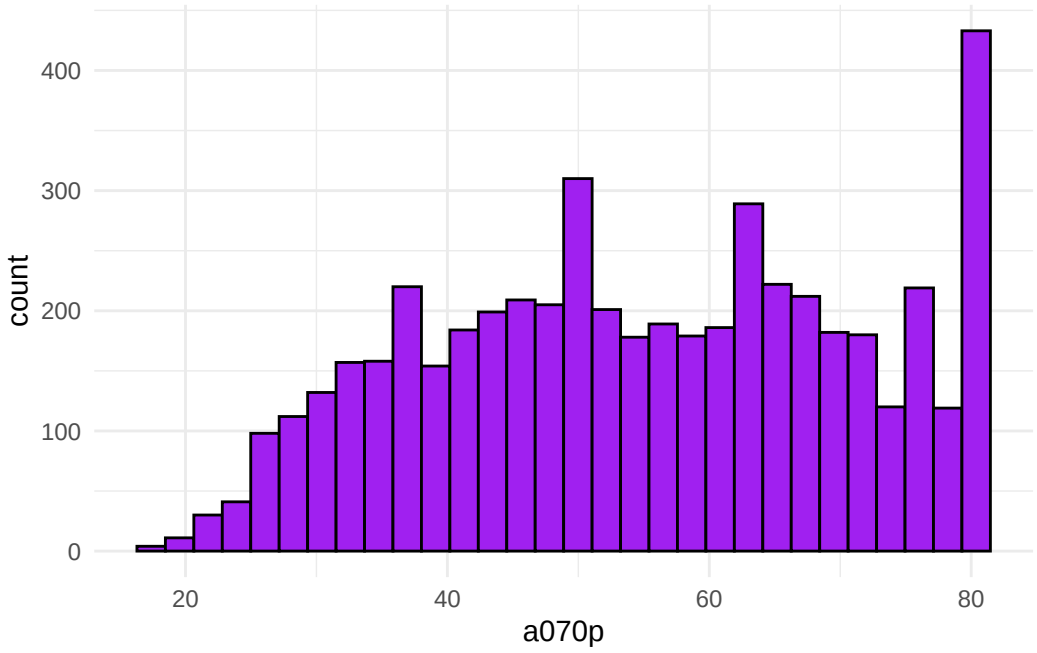


(a) Distribution of Weekly Household Income (p344p)

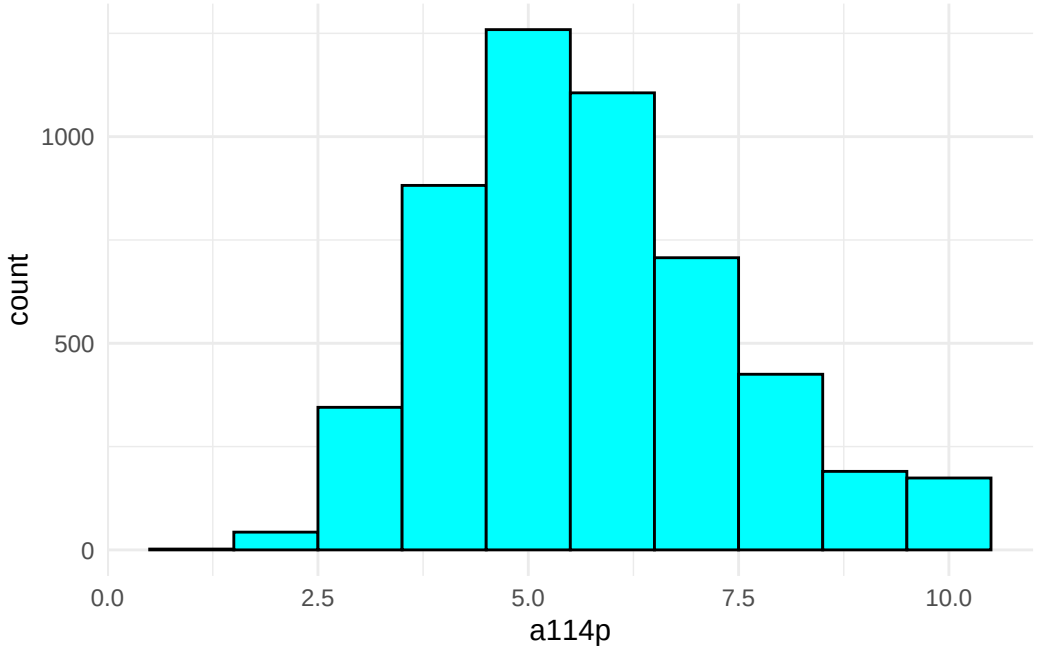


(b) Distribution of Anonymised Household Income (incanon)

Figure 2: Distribution of Weekly Household Income and Anonymised Household Income

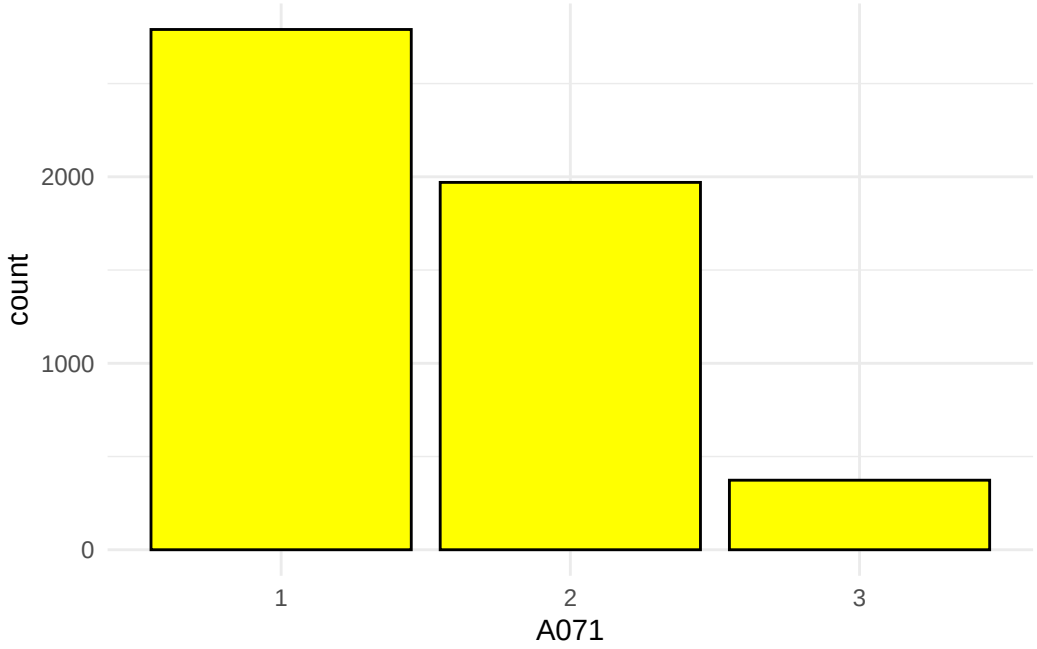


(a) Distribution of Age of Oldest Person in Household (a070p)

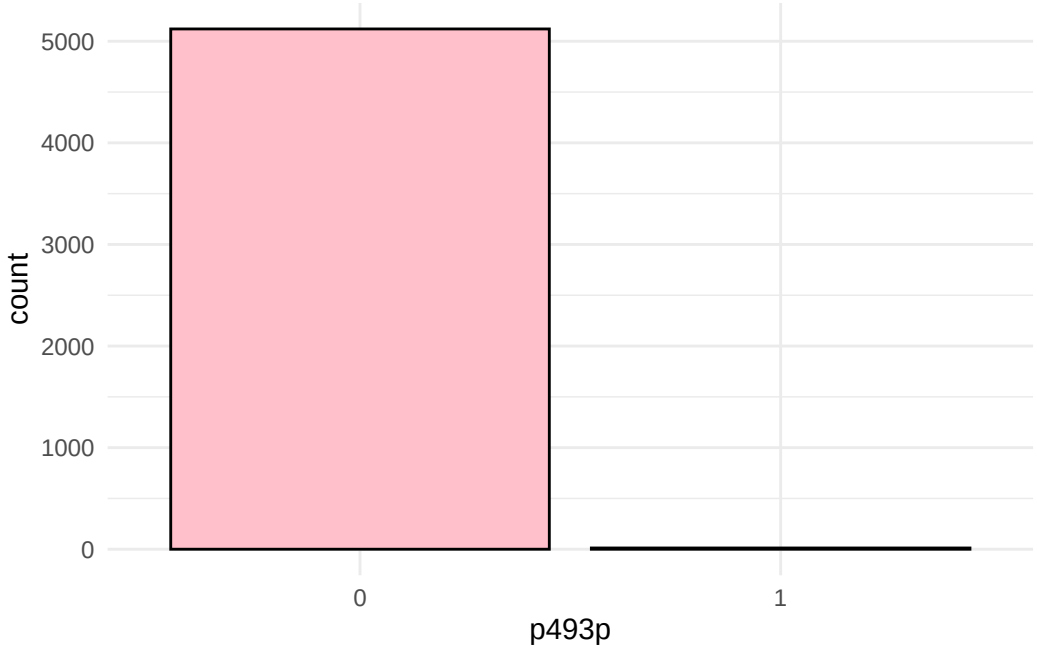


(b) Distribution of Rooms in Accommodation (a114p)

Figure 3: Distribution of Age of Oldest Person in Household and Rooms in Accommodation



(a) Distribution of Sex of Oldest Person in Household (A071)



(b) Distribution of Wealthy Household Indicator (p493p)

Figure 4: Distribution of Sex of Oldest Person in Household and Wealthy Household Indicator

- Most households are not classified as wealthy.
- The distinction between wealthy and non-wealthy households could be significant in terms of healthcare spending, with wealthier households potentially having higher expenditure.

8. Categorical Variables (A071, A094, Gorx):
    - The distribution of categorical variables like the sex of the oldest person, NS-SEC class, and government office region might offer insights when cross-examined with health expenditure. For instance, regional differences could indicate varying healthcare costs or access across regions.

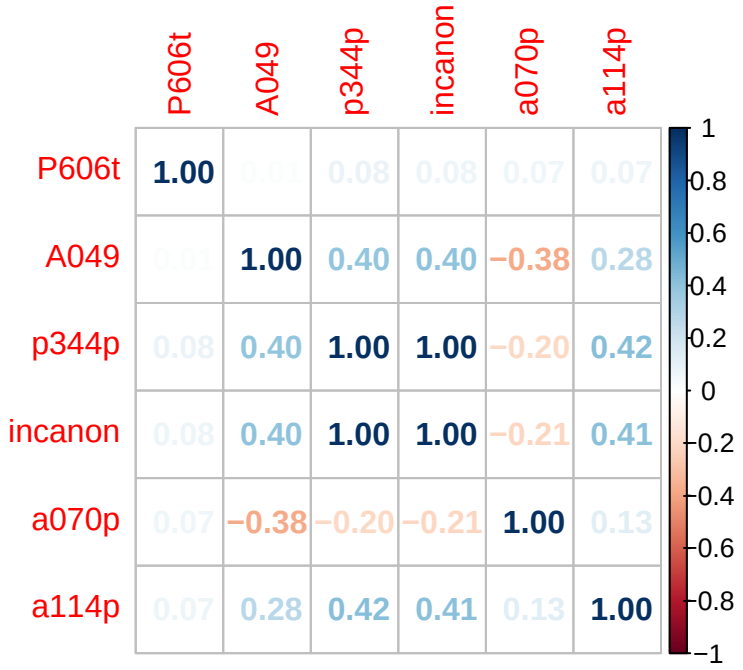


Figure 5: Correlation between numerical variables

From the correlation plot (Figure 5) of the numeric variables in dataset, we can draw several conclusions:

1. Health Expenditure (P606t):
    - There doesn't appear to be a strong correlation between health expenditure and other variables in the dataset. This suggests that health expenditure in this dataset is not strongly predictable by the other measured factors like household size, income, age of the oldest person, or number of rooms in the accommodation.
    - However, correlation does not imply causation, and there might be other unmeasured variables or non-linear relationships that affect health expenditure.

2. Household Size (A049) and Rooms in Accommodation (a114p):
    - There is a moderate positive correlation between household size and the number of rooms in the accommodation. This is intuitive as larger households typically require more rooms.
3. Income Variables (p344p and incanon):
    - There is a very strong correlation between the two income variables (p344p and incanon), indicating that they are likely measuring the same underlying attribute (household income) in slightly different ways.
4. Age of the Oldest Person (a070p):
    - The age of the oldest person in the household does not show a strong correlation with other variables. This suggests that the age factor, at least in isolation, is not a strong determinant of factors like income or health expenditure in this dataset.

3 Regression Models and Results

3.1 Model 1

The regression model built using log-transformed health expenditure (P606t) as the target variable and household size (A049), sex of the oldest person in the household (A071), and age of the oldest person in the household (a070p) as predictors provides the following insights (Table 3):

1. Model Fit:
    - R-squared: 0.037. This indicates that about 3.7% of the variability in the log-transformed health expenditure is explained by the model. It's a relatively low value, suggesting that these predictors alone do not strongly explain the variation in health expenditure.
2. Coefficients:
    - A049 (Household Size): Coefficient = 0.7802, p-value < 0.001. This suggests a significant positive relationship between household size and health expenditure. Larger households tend to have higher health expenditures.
    - a070p (Age of Oldest Person): Coefficient = 0.0406, p-value < 0.001. There's a significant positive relationship between the age of the oldest person in the household and health expenditure. Older age is associated with higher health expenditures.
    - A071 (Sex of Oldest Person):

Table 3: Linear regression model summary of the first model and second model

	Model 1	Model 2
(Intercept)	−7.6685	−8.6719
	(0.3776)	(0.3844)
A049	0.7802	0.5237
	(0.0646)	(0.0680)
A0712	−0.2852	−0.0721
	(0.1564)	(0.1558)
A0713	0.9097	0.6371
	(0.2910)	(0.2887)
a070p	0.0406	0.0444
	(0.0049)	(0.0049)
incanon		0.0018
		(0.0002)
p493p1		−2.3896
		(1.5197)
Num.Obs.	5133	5133
R2	0.037	0.059
R2 Adj.	0.036	0.058
AIC	31 596.9	31 479.0
BIC	31 636.2	31 531.3
Log.Lik.	−15 792.462	−15 731.488
F	48.719	53.787
RMSE	5.25	5.19

- A0712 (All Female): Coefficient = -0.2852, p-value = 0.068. This coefficient is not statistically significant at the 5% level, suggesting that having an all-female oldest person does not significantly differ in health expenditure compared to the base category (all-male).
- A0713 (Mixed Sex): Coefficient = 0.9097, p-value = 0.002. This is significant, indicating that households with a mixed-sex oldest person have higher health expenditures compared to households with an all-male oldest person.

3.2 Model 2

The second regression model, which includes both demographic and economic/income data (since the correlation coefficient between incanon and p344p is 1, in order to avoiding multi-collinearity problem, I only use incanon predictor), yields the following results (Table 3):

1. Household Size (A049):

- Coefficient: 0.5237
- This suggests that an increase in household size is associated with an increase in health expenditure, which aligns with the expectation that larger households might have higher health-related costs.

2. Age of Oldest Person (a070p):

   - Coefficient: 0.0444
   - Indicates that as the age of the oldest person in the household increases, there's a slight increase in health expenditure.

3. Sex of the Oldest Person (A071) Categories:

   - A071_2 (All Female): Coefficient of -0.0721, not statistically significant (p-value: 0.643).
   - A071_3 (Mixed Sex): Coefficient of 0.6371, suggesting households with mixed sex as the oldest person are associated with higher health expenditure compared to the baseline (all male), and this effect is statistically significant (p-value: 0.027).

4. Anonymised Household Income (incanon):

   - Coefficient: 0.0018
   - Indicates a positive association between household income and health expenditure, which might reflect the ability of higher-income households to spend more on health.

5. Wealthy Household Indicator (p493p1):

   - Coefficient: -2.3896
   - This negative coefficient suggests that households classified as wealthy have lower health expenditure, but this result is not statistically significant (p-value: 0.116).

6. Overall Model Performance:

   - The R-squared value is 0.059, indicating that the model explains approximately 5.9% of the variability in the log-transformed health expenditure. While this is an improvement over the first model, it still leaves a large portion of the variance unexplained, suggesting that other unmeasured factors might be influencing health expenditure.

3.3 Model 3

The third regression model, which includes demographic information, economic and income data, housing characteristics, occupational and employment data, and geographic information, provides the following insights (Table 4):

1. Household Size (A049):

   - Coefficient: 0.5041

- Suggests a positive association between household size and health expenditure, indicating larger households tend to have higher health expenditures.

2. Age of Oldest Person (a070p):
   - Coefficient: 0.0327
   - Indicates a slight increase in health expenditure with the age of the oldest person in the household.
3. Sex of the Oldest Person (A071) Categories:
   - A071_2 (All Female): Coefficient of -0.0598, not statistically significant.
   - A071_3 (Mixed Sex): Coefficient of 0.5931, suggesting a slightly higher health expenditure in households with mixed sex compared to all male households.
4. Anonymised Household Income (incanon):
   - Coefficient: 0.0014
   - Reflects a positive correlation between household income and health expenditure.
5. Wealthy Household Indicator (p493p_1):
   - Coefficient: -1.8057
   - Indicates a potential negative association between being classified as wealthy and health expenditure, though not statistically significant.
6. Rooms in Accommodation (a114p):
   - Coefficient: 0.1899
   - Suggests that a higher number of rooms in accommodation is associated with increased health expenditure.
7. Occupational and Employment Data (A094) and Geographic Information (Gorx):
   - Various coefficients for different occupational classes and regions, some of which are statistically significant.
   - These coefficients suggest varying health expenditures based on the occupational class of the household reference person and the household's region.
8. Model Performance:
   - R-squared: 0.077, indicating that about 7.7% of the variability in log-transformed health expenditure is explained by the model. This is an improvement over the previous models but still leaves a substantial amount of variance unexplained.
   - The presence of some non-significant predictors suggests the potential for model refinement.

3.4 Model 4

Based on the analyses of the three regression models and considering the significance of coefficients, the distribution of variables, and potential multicollinearity issues, I recommend the following predictors for final model:

- Household Size (A049): This variable consistently showed a significant positive relationship with health expenditure.
- Age of the Oldest Person in the Household (a070p): Age could be a relevant factor, especially if older individuals have different health needs.
- Anonymised Household Income (incanon): Income is a crucial factor in health expenditure, as it can influence access to healthcare services and the ability to afford healthcare.
- Rooms in Accommodation (a114p): The number of rooms was significant in the previous model and might reflect the household's overall economic status or lifestyle, which could impact health expenditure.
- NS - SEC 8 Class of Household Reference Person (A094): Some occupational classes showed significant coefficients in the previous models, consider including them.
- Government Office Region Modified (Gorx): There are regions demonstrated a significant impact on health expenditure, they might be worth including.

The model coefficients and summary (I only show the first 15 coefficients in Table 4) are:

1. Household Size (A049):
    - Coefficient: 0.5201
    - Indicates that larger households are associated with higher health expenditures.
2. Age of Oldest Person (a070p):
    - Coefficient: 0.0323
    - Suggests a slight increase in health expenditure with the increasing age of the oldest person in the household.
3. Anonymised Household Income (incanon):
    - Coefficient: 0.0015
    - Reflects a positive association between household income and health expenditure.
4. Rooms in Accommodation (a114p):
    - Coefficient: 0.1920
    - Indicates that more rooms in accommodation are associated with increased health expenditure.
5. Occupational Class of Household Reference Person (A094) and Geographic Information (Gorx):

Table 4: Linear regression model summary of the third model and fourth model

	Model 3	Model 4
(Intercept)	−9.3326	−9.3711
	(0.6868)	(0.6775)
A049	0.5041	0.5201
	(0.0707)	(0.0702)
A0712	−0.0598	
	(0.1562)	
A0713	0.5931	
	(0.2872)	
a070p	0.0327	0.0323
	(0.0068)	(0.0067)
incanon	0.0014	0.0015
	(0.0002)	(0.0002)
p493p1	−1.8057	
	(1.5144)	
a114p	0.1899	0.1920
	(0.0502)	(0.0502)
A0942	0.0569	0.0942
	(0.4891)	(0.4889)
A0943	0.2526	0.2685
	(0.4611)	(0.4609)
A0944	0.9728	0.9917
	(0.5112)	(0.5110)
A0945	0.1705	0.1863
	(0.5240)	(0.5237)
A0946	0.7102	0.7436
	(0.5337)	(0.5335)
A0947	−0.2204	−0.2198
	(0.5172)	(0.5172)
A0948	0.0824	0.0990
	(0.5289)	(0.5290)
Num.Obs.	5133	5133
R2	0.077	0.075
R2 Adj.	0.071	0.071
AIC	31 429.5	31 429.9
BIC	31 632.3	31 613.1
Log.Lik.	−15 683.740	−15 686.959
F	14.591	16.017
RMSE	5.14	5.14

- Various coefficients for different occupational classes and regions are included. Some of these coefficients are statistically significant, suggesting that occupation and region can influence health expenditure.
- For example, the negative and significant coefficient for A0949 (Never worked and long term unemployed) suggests lower health expenditure for this group.

6. Model Performance:

   - R-squared: 0.075, which means the model explains approximately 7.5% of the variability in log-transformed health expenditure.

4 Conclusion and Discussion

The analysis of the Living Costs and Food Survey UK Data Archive aimed to understand the factors influencing total health expenditure in households. Using linear regression models, we explored various demographic, economic, income, housing, occupational, and geographical predictors. The final model incorporated key variables such as household size, age of the oldest person, household income, number of rooms, occupational class, and region.

Key findings include a positive association between household size and health expenditure, suggesting that larger households tend to incur higher health costs. This could be due to more diverse health needs or simply more individuals requiring healthcare. The age of the oldest person in the household also showed a positive relationship with health expenditure, reflecting potentially higher healthcare needs in older age. Household income was another significant predictor, underscoring the impact of economic status on healthcare affordability and access.

The number of rooms in accommodation, a proxy for socioeconomic status, was positively related to health expenditure, possibly indicating that more affluent households have greater healthcare spending. Occupational class and geographic variables showed varying impacts, suggesting that socio-economic and regional factors influence health expenditure, though these relationships were not always straightforward or statistically significant.

The model diagnostics revealed some concerns. The presence of heteroskedasticity and potential multicollinearity issues, along with the Ramsey's RESET test results, suggest that the linear model might not fully capture the complexity of the relationships. Additionally, the residuals vs fitted values plot indicated potential non-linearity and the presence of outliers.

In conclusion, while the regression models provided valuable insights into factors affecting health expenditure, they also highlighted the complexity of predicting healthcare costs. The modest R-squared values suggest that other unmeasured factors might play a significant role. Future research could explore more sophisticated models, including non-linear relationships and interaction effects, to better understand the dynamics of health expenditure. Addressing potential outliers and ensuring model assumptions are met would further enhance the robustness of the findings.

Appendix

.1 Model Diagnostics

The analysis of the final regression model using various diagnostic tests provides the following insights.

Table 5: Variance-inflation factors of model

GVIF	Df	GVIF^(1/(	2*Df)	)
A049	1.472784	1	1.213583	
a070p	2.286864	1	1.512238	
incanon	1.904756	1	1.380129	
a114p	1.434300	1	1.197623	
A094	3.117332	11	1.053039	
Gorx	1.110677	11	1.004783	

The VIF values (Table 5) for most predictors are below 10, suggesting that multicollinearity is not a severe concern for these variables.

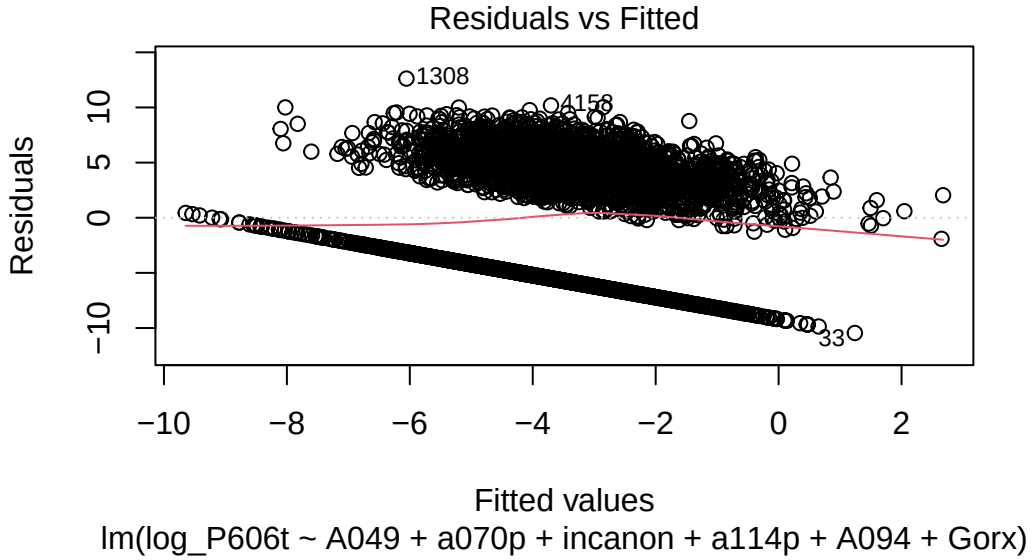


Figure 6: The scatterplot between model residuals and fitted values

The Residuals vs Fitted Values Plot (Figure 6) would typically show whether the residuals have non-linear patterns. Ideally, the residuals should be randomly scattered around zero without any apparent pattern. In this plot, while there is some randomness in the spread of residuals, there appears to be a slight pattern, particularly with residuals not being evenly

distributed across different ranges of fitted values. This suggests the presence of non-linearity in the relationship between predictors and the response variable. The plot shows some signs of non-constant variance (heteroscedasticity), as indicated by the spread of residuals that seems to vary across the range of fitted values.

```
Multiple parameters; naming those columns df1, df2
```

Table 6: Ramsey's RESET test for model

df1	df2	statistic	p.value m	ethod
2	5104	9.154635	0.0001075	RESET test

The RESET test (Table 6) has a low p-value (0.0001057), suggesting that the model may suffer from specification issues. This means the current model might not adequately capture the relationship between predictors and the target variable.

Table 7: Breusch-Pagan test against heteroskedasticity for model

statistic	p.value	parameter m	ethod
144.1207	0	26	studentized Breusch-Pagan test

The Breusch-Pagan test (Table 7) indicates the presence of heteroskedasticity (p-value $<$ 0.0001). This suggests that the variance of residuals is not constant, which can affect the reliability of standard errors and thus the inference drawn from coefficient estimates.

The QQ plot (Figure 7) shows some deviation from the line, particularly in the tails, suggesting that the residuals may not be perfectly normally distributed. While some deviation is common in real-world data, extreme deviations might affect the validity of some assumptions of the linear regression model.

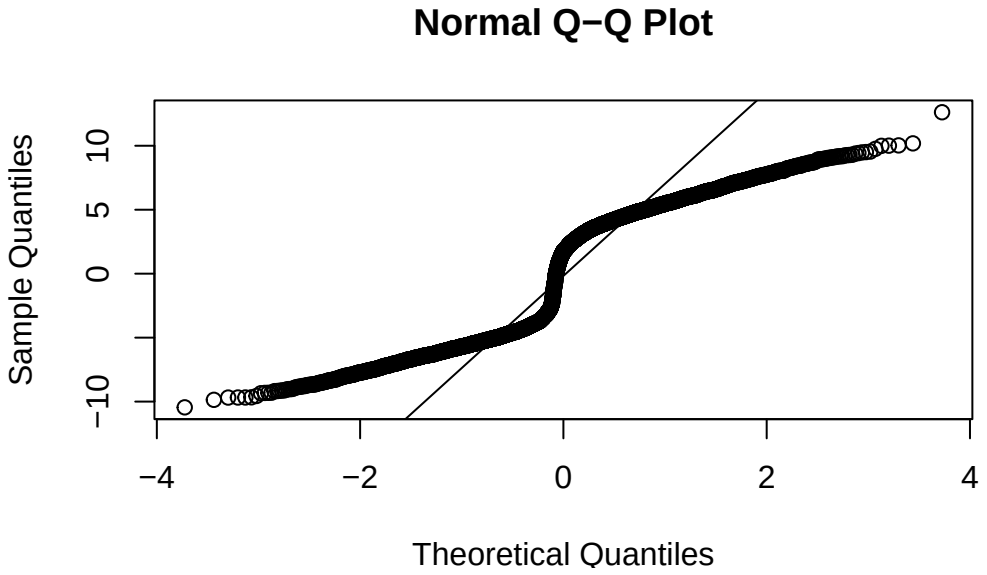


Figure 7: Normal QQ plot of the model residuals

References

Fox, John, and Sanford Weisberg. 2022. *Car: Companion to Applied Regression*. https://CRAN.R-project.org/package=car.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, Alex Hayes, and Simon Couch. 2022. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

UK Data Service. 2022. "Living Costs and Food Survey." Available online at UK Data Service. https://ukdataservice.ac.uk/.

Wei, Taiyun, and Viliam Simko. 2021. *Corrplot: Visualization of a Correlation Matrix*. https://CRAN.R-project.org/package=corrplot.

Wickham, Hadley. 2016. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://CRAN.R-project.org/package=knitr.

Zeileis, Achim, and Torsten Hothorn. 2022. *Lmtest: Testing Linear Regression Models*. https://CRAN.R-project.org/package=lmtest.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.