word count: 3750

Candidate Number: 26684

Introduction

This report investigates how income, household tenure, number of children, and the economic position of the household reference person affect expenditure patterns across various regions of the UK. I used Bayesian and Hierarchical Bayesian models to estimate these effects and compared the results with those from frequentists linear regression. My aim is to understand whether these variables have a consistent impact on spending in different regions and to determine which modelling approach best captures this relationship. The comparison of different statistical techniques helps us confirm the reliability of my results and provides a clearer picture of spending behaviour across the UK. Such an understanding of household expenditure helps government policymakers in creating more effective economic policies. By understanding how different demographics spend their money, policymakers can tailor fiscal and social policies to boost economic growth, address income inequality, and manage inflation.

My analysis has demonstrated that the Bayesian Hierarchical model, featuring a random intercept, has achieved optimal performance in terms of predictive accuracy and model complexity. The predictive variables incorporated into the model included the household's income, the economic position of the household reference person, and the number of children in the household.

In the initial phase of my analysis, a linear regression model was employed, followed by two Bayesian Regression models, one with Nimble and the other in STAN.

Later, I used Hierarchical Bayesian Models, also performed in both Nimble and STAN, where I chose the individual regions as our hierarchy.

Summary of the data

Our dataset comprises a collection of 5,144 cases from the United Kingdom, spanning from the year 2013 onwards. This dataset has been utilised for the purpose of my modelling and includes six variables.
These variables are as follows: expenditure, which represents the amount in GBP spent by a household and serves as our dependent variable; income, defined as the gross weekly household income; A121r, a categorical variable that denotes the tenure status of the household, distinguishing between publicly rented, privately rented, or owned properties; A093r, a categorical variable that denotes the economic position of the household reference person, distinguishing between full-time working, part-time working, unemployed, or economically inactive; G019r, a categorical variable that denotes the number of children in the household, distinguishing between no child, one child, or two or more children; and Gorx, which is employed as the hierarchical level in my models. The Gorx variable specifies the region within the UK where the household resides, thereby facilitating a more detailed analysis based on geographical segmentation.

Exploratory Data Analysis

Upon conducting an exploratory analysis, it has been observed that the expenditure variable exhibits significant right skewness. The application of a logarithmic transformation has effectively mitigated this skewness. A similar skewness pattern was identified in the income variable, which also improved upon application of the logarithmic transformation. These transformations are expected to improve the accuracy of our model predictions.
Both the expenditure and income variables were centred to minimise the effects of multicollinearity but only in the Bayesian modelling.

Analysis of the bar chart representing household tenure (A121r) reveals a substantial predominance of privately owned households compared to other groups. Regarding variable A093r, which indicates the economic position of the household reference person, it is noted that the largest proportion of individuals is classified as full-time working.

From variable G019r, which denotes the number of children in the household, significant majority of households have no children.

Furthermore, the distribution of cases across different regions varies, with some regions being underrepresented relative to others. This discrepancy may be attributed to the geographical size of the regions in question. The analysis also uncovers a positive linear relationship between income and expenditure, suggesting a direct correlation between these two variables.
Additionally, a comparison of median expenditures across different tenure groups indicates variations, implying a relationship between expenditure and the A121r variable. Similar influences are observed for variables G019r and A093r, pointing to the impact of household tenure status and other demographic factors on expenditure patterns.

Model Predictors

Model 1 predictors:

- Income
- A121r (Household tenure)

Model 2 predictors:

- Income
- A093r (Economic position of household reference person)
- G019r (Number of children in the household)

Model 1 (Frequentists Linear Regression)

The formula for our linear regression model is given by:

$$Y_i = \beta_0 + \beta_1 X + \beta_2 A191r_i + \epsilon_i$$

Y_i is log-transformed expenditure for level i
X is log-transformed income
i index representing the level of A191r (Household tenure)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.550994   0.056485  27.459  < 2e-16 ***
income      0.690510   0.009657  71.505  < 2e-16 ***
A121r2      0.228023   0.022677  10.055  < 2e-16 ***
A121r3      0.126694   0.018704   6.774  1.4e-11 ***
```

From the summary I observed that all predictor variables in the model are statistically significant, as indicated by the p-values less than 0.05.
From the output, we can see that a 1% increase in income will result in a 0.69% increase in expenditure, all else being equal. The model uses publicly rented households as the baseline, and both privately rented households and home owners have expenditures that are higher by 23% and 13%, respectively, relative to publicly rented households.

The Multiple R-squared value is 0.5631, meaning that approximately 56.31% of the variability in the dependent variable is explained by the model. This indicates a reasonably good fit.

Model 2 (Frequentists Linear Regression)

The formula for our linear regression model is given by:

$$Y_{ij} = \beta_0 + \beta_1 X + \beta_2 A093r_i + \beta_3 G019r_j + \epsilon_{ij}$$

Y_{ij} is log-transformed expenditure for levels i and j
X is log-transformed income
i index representing the level of A093r (Economic Position of Household reference person)
j index representing the level of G019r (Number of children in the Household)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.79150    0.07295  24.560  < 2e-16 ***
income       0.67031    0.01093  61.321  < 2e-16 ***
A093r2       0.03131    0.02222   1.409    0.159
A093r3      -0.02146    0.04014  -0.534    0.593
A093r4      -0.08010    0.01684  -4.756 2.03e-06 ***
G019r2       0.09569    0.01997   4.792 1.70e-06 ***
G019r3       0.12788    0.01778   7.191 7.36e-13 ***
```

From the summary we observed that majority of predictor variables in the model are statistically significant, apart from Economic position of Part-time working and Unemployed.

From the output, we can see that a 1% increase in income will result in a 0.67% increase in expenditure, all else being equal.
For Economic position, the model uses Full-time working households as the baseline and for the number of children in the household the baseline are households without children.

The Multiple R-squared value is 0.5657, meaning that approximately 56.57% of the variability in the dependent variable is explained by the model. This indicates a reasonably good fit.

Bayesian Non-hierarchical Modelling

Model 1 (Bayesian Regression with Nimble)

Priors:

for the regression coefficients β

$\beta_d \sim N(0,10^4)$ for $d = 1,\ldots, D$

for the residual variance τ

$\tau \sim \text{Gamma}(2.5, 0.5)$

for standard deviation σ

$$\sigma = \sqrt{1/\tau}$$

Likelihood:

for each observation of expenditure i is

$\text{expenditure}_i \sim N(\mu_i, \tau)$ for $i = 1,\ldots, n$

where μ_i is the linear predictor for the i-th observation:

$$\mu_i = \beta_1 X_{i1} + \beta_2 X_{i2}$$

β_d represents the d-th regression coefficient for D predictors
X_{ij} represents the value of the j-th predictor for the i-th observation

Summary of fixed effects

	mean	sd	2.5%	50%	97.5%	Rhat	n.eff
beta[1]	-1.302670e-06	0.006434687	-0.01274367	-4.081723e-05	0.01252688	1	12000
beta[2]	7.112082e-01	0.008919458	0.69345975	7.112795e-01	0.72878040	1	12000
beta[3]	-1.629466e+00	99.536775582	-197.44030835	-9.617230e-02	191.43378890	1	12000
beta[4]	-2.087243e+00	99.928020021	-197.86626895	-9.119494e-01	193.02659107	1	12000
tau	4.779292e+00	0.094789011	4.59443550	4.778299e+00	4.96598655	1	12000

In the output, beta[1] represents the intercept, beta[2] is the effect of income, beta[3] is the effect of private renting, and beta[4] is the effect of ownership.

We can see that the only significant effect is that of income on household expenditure, since it does not include zero in its 95% credible interval.

The effect of income is positively related to expenditure; hence, as a household's income increases, so does its expenditure. Intuitively, this makes sense, as we earn money to spend on things that we believe make us happier, rather than merely saving indefinitely.

Convergence is confirmed with R-hat values at 1 for all parameters, and the large effective sample sizes (n.eff) indicate robustness in the posterior distributions. The sigma value of approximately 0.457 reflects a reasonable level of precision in the model's predictions relative to the variability in the data.

The reliability of these estimates is further validated by the satisfactory mixing of chains in the trace plots and by the rapid and stable convergence depicted in the autocorrelation function (ACF) plots.

Model 1 (Bayesian Regression with STAN)

Priors:

$\beta \sim N(0, 2.5)$

$\sigma \sim N(0, 5)$

Likelihood:

$y_i \sim N(X_i\beta, \sigma)$ for $i = 1, \ldots, n$

y_i represents the expenditure for observation i
X_i is the vector of predictors for observation i
β is the vector of regression coefficients
σ is the standard deviation of the residuals

Summary of fixed effects

	mean	se_mean	sd	2.5%	97.5%	n_eff	Rhat
sigma	0.4532127	6.003039e-05	0.00452470	0.4446566	0.46216560	5681.162	1.0010526
(Intercept)	-0.1205275	2.592703e-04	0.01649837	-0.1530318	-0.08797923	4049.268	0.9999541
income	0.6903395	1.369978e-04	0.00963543	0.6714355	0.70922553	4946.695	0.9998820
A121r2	0.2277699	3.343075e-04	0.02281235	0.1834031	0.27188372	4656.374	0.9999749
A121r3	0.1265505	2.918259e-04	0.01883730	0.0890143	0.16305178	4166.677	0.9999850

We can see from the output that all coefficients are significant, as none of them include zero in their 95% credible intervals. As income increases, so does expenditure. We can also see that privately rented households have higher expenditure relative to publicly rented ones, and households that are owned also have higher expenditure than public ones.

Interestingly, home ownership seems to lead to lower expenditure than private renting. This would make sense if interest rates are low, and our mortgage payments would be lower than rent.

The Bayesian regression model, as implemented in STAN, shows that all parameters, including sigma, intercept, income, A121r2, and A121r3, have been estimated with a high level of confidence, indicated by the Rhat values very close to 1. The sigma value of approximately 0.453 reflects a reasonable level of precision in the model's predictions relative to the variability in the data.

The effective sample sizes (n.eff) for each parameter are large, suggesting that the posterior estimates are reliable. The mean values of the coefficients indicate the direction and strength of the relationships, and the standard deviations (sd) are relatively small, showing precision in the estimates.

Comparison between software:

Upon comparative analysis of the outputs derived from the two Bayesian models, it is observed that both models have estimated the coefficients with a considerable degree of confidence. This can be confirmed by the diagnostics plots, where we can see good mixing of the two chains in the trace plots and quick drop to 0 and stabilising around this value in the ACF plots.

Notably, there exists a variance in the actual values and significance of specific coefficients, particularly A121r2 and A121r3, across the two models.

Model 2 (Bayesian Regression with Nimble)

Priors:

for the regression coefficients β

$\beta_d \sim N(0,10^4)$ for $d = 1,\ldots, D$

for the residual variance τ

$\tau \sim \text{Gamma}(2.5, 0.5)$

for standard deviation σ

$$\sigma = \sqrt{1/\tau}$$

Likelihood:

for each observation of expenditure i is

$\text{expenditure}_i \sim N(\mu_i, \tau)$ for $i = 1,\ldots, n$

where μ_i is the linear predictor for the i-th observation:

$$\mu_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}$$

β_d represents the d-th regression coefficient for D predictors
X_{ij} represents the value of the j-th predictor for the i-th observation

Summary of fixed effects

	mean	sd	2.5%	50%	97.5%	Rhat	n.eff
beta[1]	-0.009101858	6.735021e-03	-0.02233384	-0.009103974	4.000631e-03	1	9800
beta[2]	0.712462739	8.843838e-03	0.69513075	0.712416713	7.300631e-01	1	10927
beta[3]	0.086645956	2.076738e-02	0.04628643	0.086629704	1.273509e-01	1	9741
beta[4]	-1.042152457	9.938771e+01	-197.40273168	-0.042163464	1.949978e+02	1	11783
beta[5]	-0.995582707	1.000122e+02	-196.84718811	-1.506268863	1.984483e+02	1	12510
beta[6]	0.681388261	9.986029e+01	-194.59458739	0.411960152	1.981163e+02	1	12000
beta[7]	-0.720620528	1.003281e+02	-198.68654689	-0.679774172	1.946945e+02	1	12000
tau	4.794036437	9.548948e-02	4.61058022	4.793879929	4.982901e+00	1	12103

In the output, beta[1] represents the intercept, beta[2] is the effect of income, beta[3] is the effect of part-time working, beta[4] is the effect of unemployed, beta[5] is the effect of economically inactive, beta[6] is the effect of one child and beta[7] is the effect of 2 or more children.

We can see that the only significant effects are that of income and the economic position of household reference person that is part-time working since they both do not include zero in their 95% credible intervals.

The effect of the economic position of household reference person that is part-time working on expenditure is greater relative to reference person that is full-time working. According to the model part-time workers have higher expenditures. This is little counterintuitive given that full-time workers generally earn more due to more hours worked. In some sectors, part-time jobs might offer higher hourly wages compared to similar full-time positions, especially if they require specialised skills or are in high-demand periods. This can increase the disposable income available to part-time workers.

Convergence is confirmed with R-hat values at 1 for all parameters, and the large effective sample sizes (n.eff) indicate robustness in the posterior distributions. The sigma value of approximately 0.457 reflects a reasonable level of precision in the model's predictions relative to the variability in the data.

The reliability of these estimates is further validated by the satisfactory mixing of chains in the trace plots and by the rapid and stable convergence depicted in the autocorrelation function (ACF) plots.

Model 2 (Bayesian Regression with STAN)

Priors:

$\beta \sim N(0, 2.5)$

$\sigma \sim N(0, 5)$

Likelihood:

$y_i \sim N(X_i\beta, \sigma)$ for $i = 1,\ldots, n$

y_i represents the expenditure for observation i
X_i is the vector of predictors for observation i
β is the vector of regression coefficients
σ is the standard deviation of the residuals

Summary of fixed effects

	mean	se_mean	sd	2.5%	97.5%	n_eff	Rhat
sigma	0.451922172	6.106203e-05	0.004568767	0.44300709	0.46076372	5598.291	0.9999824
(Intercept)	-0.006042993	1.775559e-04	0.011935390	-0.02923288	0.01755067	4518.589	1.0000848
income	0.670521034	1.548969e-04	0.010912093	0.64910903	0.69187695	4962.843	0.9999350
A093r2	0.031342383	3.095866e-04	0.022438038	-0.01269188	0.07518869	5252.977	0.9999027
A093r3	-0.021243921	5.534176e-04	0.040262038	-0.09926594	0.05779744	5292.800	0.9999684
A093r4	-0.079762962	2.473679e-04	0.016757523	-0.11276893	-0.04706072	4589.158	1.0001046
G019r2	0.095989763	2.770629e-04	0.019532053	0.05686351	0.13408728	4969.802	1.0002728
G019r3	0.128188596	2.467496e-04	0.017890194	0.09244878	0.16304603	5256.748	1.0002836

From the output, we can see that the model identifies the effects of the number of children on expenditure as significant. Households with one child have higher expenditures relative to those without, and households with two or more children have even higher expenditures compared to households without any children. This is reasonable, as children need to be fed, schooled, and clothed.

The Bayesian regression model, as implemented in STAN, shows that all parameters, including sigma, intercept, income, A121r2, and A121r3, have been estimated with a high level of confidence, indicated by the Rhat values very close to 1.
The effective sample sizes (n.eff) for each parameter are large, suggesting that the posterior estimates are reliable.

Comparison between software:

Upon comparative analysis of the outputs derived from the two Bayesian models, it is observed that both models have estimated the coefficients with a considerable degree of confidence.

Notably, there exists a variance in the actual values and significance of specific coefficients, particularly A093r2, A093r3, A093r4, G019r2 and G019r3, across the two softwares.

Hierarchical Bayesian Modelling

For the upcoming analysis, I intend to implement a Hierarchical Bayesian Regression model that incorporates both random intercepts and slopes, each informed by independent prior distributions.
I will also fit models that incorporate only the random intercept.
This advanced modelling approach will not only include covariates such as A121r, A093r, G019r and Income, but will also integrate the variable 'Gorx' to establish a hierarchical structure, thereby allowing for the accommodation of nested data variations at different levels of the hierarchy.

Model 1 (Hierarchical Regression with STAN)

Random Intercept+Slope

Priors:

Global coefficients and variances

$\beta \sim N(0, 2.5^2)$

$\sigma_{intercept} \sim N(0, 5^2)$

$\sigma_{slope} \sim N(0, 5^2)$

$\sigma_{res} \sim N(0, 5^2)$

Random effects

$\gamma_{intercept}[g] \sim N(0, \sigma_{intercept})$ - Random intercepts for each group g

$\gamma_{slope}[g] \sim N(0, \sigma_{slope})$ - Random slopes for each group g

Likelihood:

expenditure[i] $\sim N(\mu_i, \sigma_{res})$

$\mu_i =$ dot_product$(\beta, X[i])$ + $\gamma_{intercept}$[Gorx[i]] + γ_{slope}[Gorx[i] $\cdot X[i,2]$

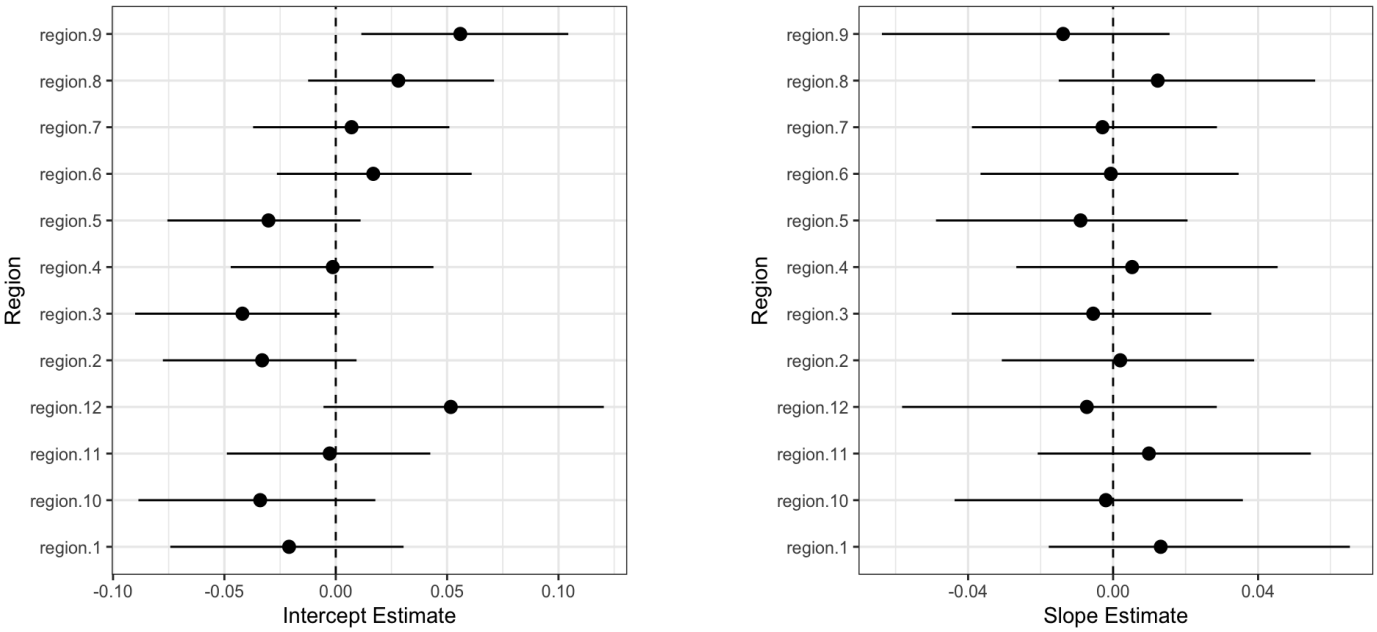
$X[i]$ represents the vector of predictors for observation i, $X[i, 2]$ specifically refers to the second predictor in the vector $X[i]$ and Gorx[i] identifies the group to which observation i belongs.

Summary of fixed effects

	mean	se_mean	sd	2.5%	97.5%	n_eff	Rhat
beta[1]	-1.171466e-01	0.0003919120	0.021264022	-1.585992e-01	-7.516165e-02	2943.8360	1.0008835
beta[2]	6.873528e-01	0.0001788854	0.011877834	6.638822e-01	7.103149e-01	4408.8418	0.9998975
beta[3]	2.242096e-01	0.0003429995	0.022790729	1.797937e-01	2.694917e-01	4414.9863	0.9999679
beta[4]	1.244315e-01	0.0002770095	0.018554004	8.835286e-02	1.600767e-01	4486.2748	1.0005522
sigma_res	4.517089e-01	0.0000605889	0.004410075	4.432890e-01	4.604126e-01	5297.9266	1.0000761

In the output, beta[1] represents the intercept, beta[2] is the effect of income, beta[3] is the effect of private renting, and beta[4] is the effect of ownership.

Summary of random effects



From the summary of the random effects, we can observe variation in both the random intercepts and slopes. These variations are likely influenced by factors not considered in our model, suggesting influences beyond Income and household tenure.

For the random intercept, the highest values correspond to regions 12 and 9, which are Northern Ireland and the South West, respectively. These regions exhibit baselines that are higher relative to the overall baseline.

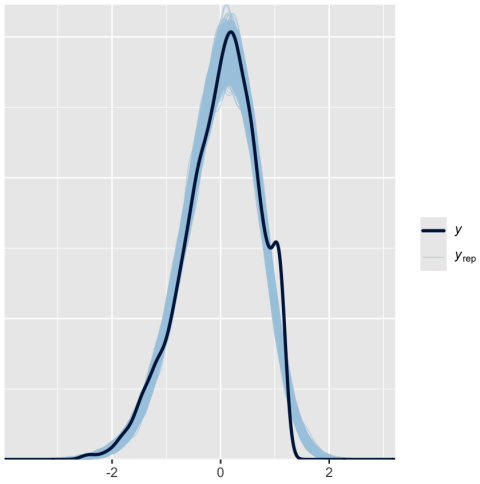
Conversely, the lowest values correspond to regions 10 and 3, which are Wales and Yorkshire and The Humber, respectively. These regions have baselines that are lower relative to the overall baseline.

It is possible that house prices might influence the variability of the random intercepts, particularly since they are significantly higher in the South West than in Yorkshire. It would be reasonable to include average house prices and average rent per household in our future analyses to account for these factors.

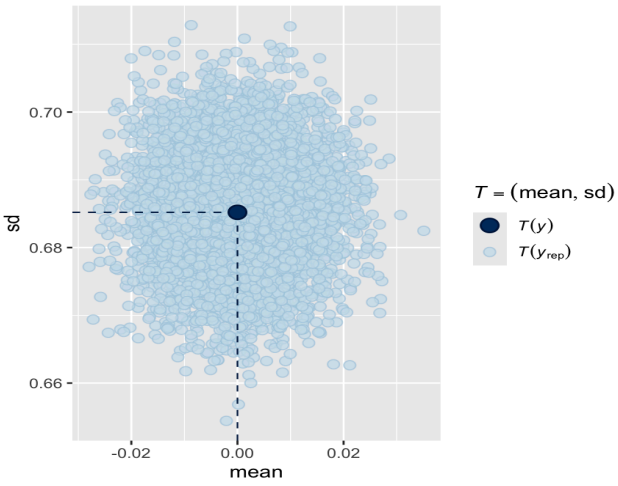
I have also included a random slope effect on the income variable. It is reasonable to assume that the rate of income change varies among different regions. This could be due to more urban areas typically offering more lucrative jobs that provide faster career progression compared to rural areas, where the diversity and lucrativeness of jobs may not be as pronounced.

Variation between random slope estimates appears to be smaller relative to random intercept estimates.

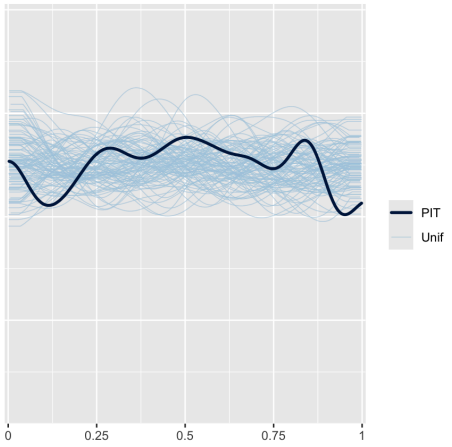
Model diagnostics



From ppc_dense_overlay plot we can see the posterior draws follow the expected distribution reasonably well with minor deviations in the right half of the distribution. Overall the fit looks satisfactory.



The fact that the observed data's statistic (dark blue point) is within the densest area of the simulated data's statistics (light blue points) implies that the model has a good fit to the observed data, further confirming our observation from previous plot.



The PIT plot indicates that the model is not perfectly calibrated. There seem to be areas, particularly around the 0.1 and 0.9 quantiles, where the model might be over-predicting, as indicated by the PIT line being below the Unif line.

Model 1 (Hierarchical Regression with Nimble)

Random Intercept+Slope

Priors:

the global intercept, the coefficient for income, and random effects coefficients for j indices:

$\beta_0, \beta_{Income}, \beta_{A121r[j]} \sim N(0, 4)$

group-specific intercepts and slopes for group g:

interceptGorx[g], slopeGorx[g] ~ $N(0, \tau_{Gorx})$

the precision of the group-specific intercepts and slopes:

τ_{Gorx} ~ Gamma(2.5, 0.5)

$$\sigma_{Gorx} = \sqrt{\frac{1}{\tau_{Gorx}}}$$

the precision of the residuals:

τ_{res} ~ Gamma(2.5, 0.5)

$$\sigma_{res} = \sqrt{\frac{1}{\tau_{res}}}$$

Likelihood:

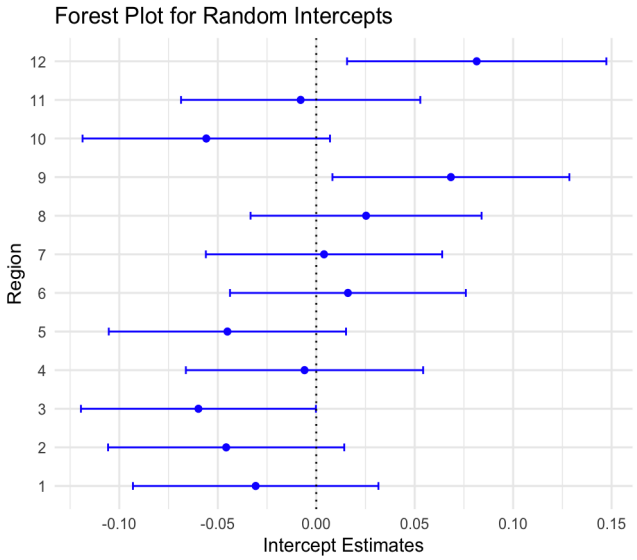
expenditure[i] ~ $N(\mu_i, \tau_{res})$

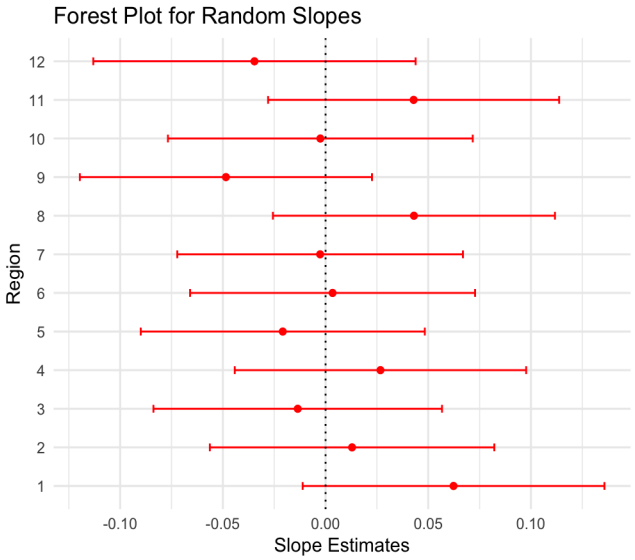
$\mu_i = \beta_0 + \beta_{Income}$ x income[i] + inprod($\beta_{A121r[1:J]}$, A121r[i, 1:J]) + inetrceptGorx[GorxID[i]] +slopeGorx[GorxID[i]] x income[i]

Summary of fixed effects

	mean	sd	2.5%	50%	97.5%	Rhat	n.eff
beta0	-0.008096213	0.90610601	-1.378158	0.076788360	1.0895501	16.75	8
betaA121r[1]	-0.118904736	0.90996045	-1.232198	-0.224060338	1.2008939	17.58	8
betaA121r[2]	0.103754545	0.91025780	-1.010654	-0.001743282	1.4237964	17.66	8
betaA121r[3]	0.002769072	0.90996943	-1.110295	-0.106218596	1.3205158	17.66	8
betaIncome	0.686338474	0.05952629	0.568912	0.686700172	0.8012152	1.00	134
tau_gorx	25.694395298	7.03251084	13.772003	25.055806421	41.0962918	1.00	2559
tau_res	4.901913817	0.09686810	4.711071	4.902889195	5.0914924	1.00	12069

Summary of random effects





Comparison between software:

Upon comparative analysis of the outputs derived from the two Bayesian models, it is observed that model implemented in STAN have estimated the coefficients with a higher degree of confidence.

Notably, there exists a variance in the actual values and significance of some fixed effects coefficients across the two softwares.

The value and spread of random effects appears to be more consistent between the two softwares.

Model 2 (Hierarchical Regression with STAN)

Random Intercept

Priors:

Global coefficients and variances

$\beta \sim N(0, 2.5^2)$

$\sigma_{intercept} \sim N(0, 5^2)$

$\sigma_{res} \sim N(0, 5^2)$

Random effects

$\gamma_{intercept}[g] \sim N(0, \sigma_{intercept})$ - Random intercepts for each group g

Likelihood:

expenditure[i] $\sim N(\mu_i, \sigma_{res})$

$\mu_i =$ dot_product(β, $X[i]$) + $\gamma_{intercept}$[Gorx[i]]

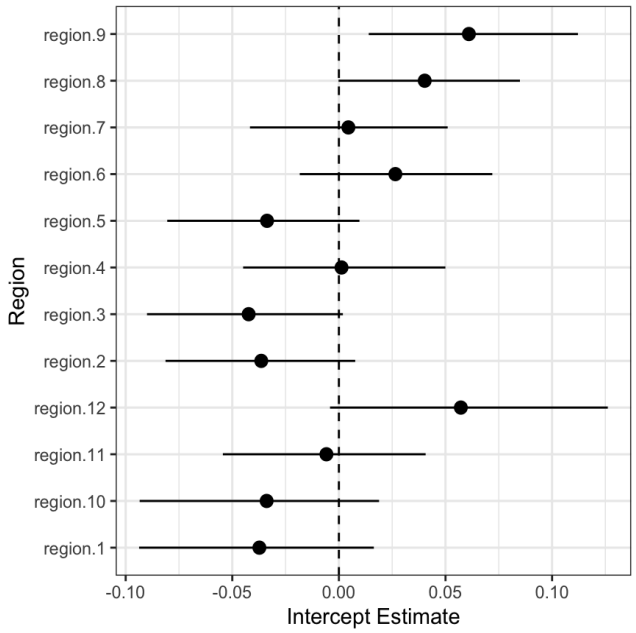
$X[i]$ represents the vector of predictors for observation i and Gorx[i] identifies the group to which observation i belongs.

Summary of fixed effects

	mean	se_mean	sd	2.5%	97.5%	n_eff	Rhat
beta[1]	-6.159764e-03	2.988416e-04	0.018880321	-4.418744e-02	3.126118e-02	3991.505	1.0005193
beta[2]	6.659088e-01	1.448619e-04	0.010750319	6.443139e-01	6.868893e-01	5507.248	0.9997496
beta[3]	3.203298e-02	3.027797e-04	0.021881907	-1.126103e-02	7.505272e-02	5222.962	0.9998124
beta[4]	-2.381569e-02	5.372901e-04	0.039468108	-1.023594e-01	5.377924e-02	5396.037	0.9999475
beta[5]	-8.051099e-02	2.174953e-04	0.016806014	-1.132839e-01	-4.712004e-02	5970.758	0.9998462
beta[6]	1.001046e-01	2.848452e-04	0.020199248	6.047720e-02	1.396422e-01	5028.664	0.9997467
beta[7]	1.281310e-01	2.447870e-04	0.017947839	9.262775e-02	1.627933e-01	5375.854	1.0000536
sigma_res	4.503116e-01	6.244346e-05	0.004467715	4.418267e-01	4.592065e-01	5119.140	1.0006895

In the output, beta[1] represents the intercept, beta[2] is the effect of income, beta[3] is the effect of part-time working, beta[4] is the effect of unemployed, beta[5] is the effect of economically inactive, beta[6] is the effect of one child and beta[7] is the effect of 2 or more children.

Summary of random effects

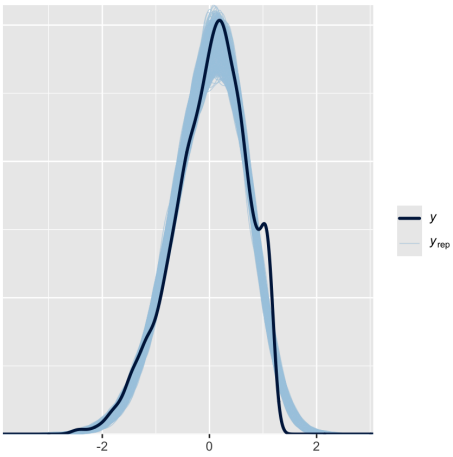


From the summary of the random effects, we can observe variation in the random intercepts. These variations are likely influenced by factors not considered in our model, suggesting influences beyond Income, number of children in the household and economic position of household reference person.

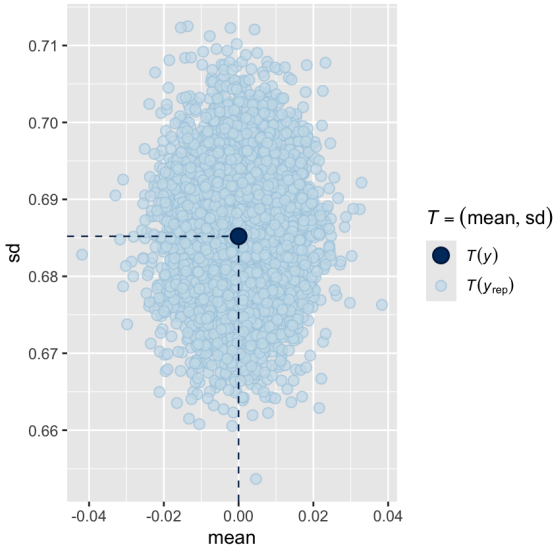
The mean values of these intercepts is very similar to the random intercepts from the model that considered only income and household tenure as predictors. The only meaningful difference is in region 1 (North East) and region 7 (London).

Additionally to house prices, age of household members might be another factor influencing the variability of the random intercepts. Different age groups have different spending priorities and patterns. For example, older adults may spend more on healthcare, while younger adults might spend more on education or entertainment. It would be interesting how different age groups are segment across different regions.

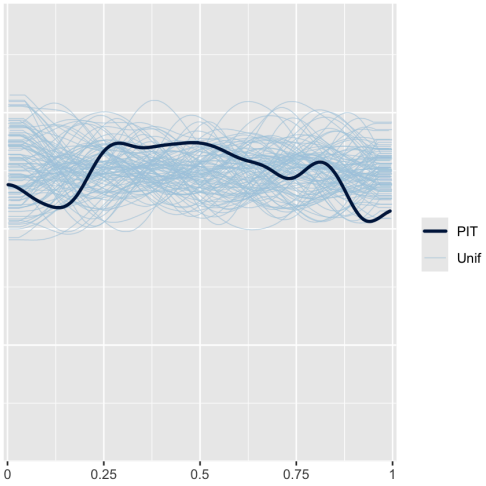
Model diagnostics



From ppc_dense_overlay plot we can see the posterior draws follow the expected distribution reasonably well with minor deviations in the right half of the distribution. Overall the fit looks satisfactory.



The fact that the observed data's statistic (dark blue point) is within the densest area of the simulated data's statistics (light blue points) implies that the model has a good fit to the observed data, further confirming our observation from previous plot



The PIT plot indicates that the model is not perfectly calibrated. There seem to be areas, particularly around the 0.1 and 0.9 quantiles, where the model might be over-predicting, as indicated by the PIT line being below the Unif line.

Model 2 (Hierarchical Regression with Nimble)

Random Intercept

Priors:

the global intercept, the coefficient for income, and random effects coefficients for j indices:

β_0, β_{Income}, $\beta_{A093r[j]}$, $\beta_{G019r[k]}$ ~ $N(0, 4)$

group-specific intercepts and slopes for group g:

interceptGorx[g], slopeGorx[g] ~ $N(0, \tau_{Gorx})$

the precision of the group-specific intercepts and slopes:

τ_{Gorx} ~ Gamma(2.5, 0.5)

$$\sigma_{Gorx} = \sqrt{\frac{1}{\tau_{Gorx}}}$$

the precision of the residuals:

τ_{res} ~ Gamma(2.5, 0.5)

$$\sigma_{res} = \sqrt{\frac{1}{\tau_{res}}}$$

Likelihood:

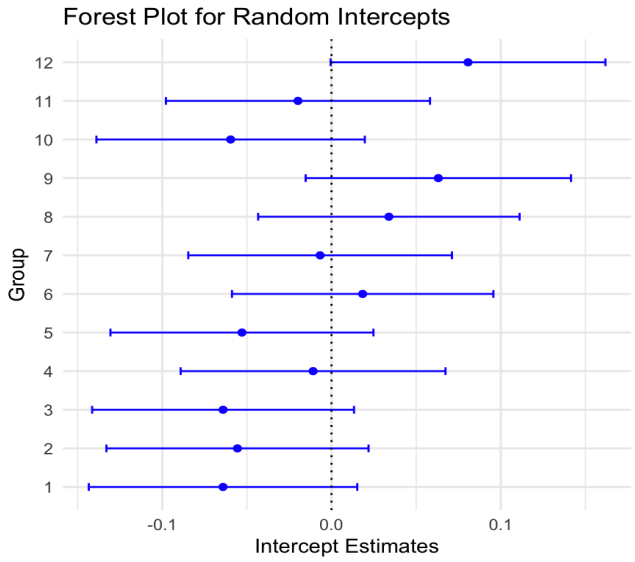
expenditure[i] ~ $N(\mu_i, \tau_{res})$

$\mu_i = \beta_0 + \beta_{Income}$ x income[i] + inprod($\beta_{A093r[1:J]}$, A093r[i, 1:J]) + inprod($\beta_{G019r[1:K]}$, G019r[i, 1:K]) + inetrceptGorx[GorxID[i]]

Summary of fixed effects

	mean	sd	2.5%	50%	97.5%	Rhat	n.eff
beta0	-0.0408142736	0.64287989	-0.9658315	-0.05732771	1.0556777	5.89	5
betaA093r[1]	0.0211991758	0.84184850	-1.3455714	0.03420873	1.1049914	8.40	8
betaA093r[2]	0.0535082075	0.84199539	-1.3123502	0.07750617	1.1390244	8.39	8
betaA093r[3]	-0.0007800802	0.84215879	-1.3607190	0.03652418	1.0865494	8.35	8
betaA093r[4]	-0.0589204062	0.84122063	-1.4238935	-0.03724692	1.0243073	8.40	8
betaG019r[1]	0.0172802510	0.27828632	-0.5486423	0.05493601	0.5139462	4.12	10
betaG019r[2]	0.1184890398	0.27917510	-0.4478167	0.15751100	0.6122575	4.11	10
betaG019r[3]	0.1453691737	0.27892800	-0.4185893	0.18526787	0.6371160	4.10	10
betaIncome	0.6652152959	0.01094917	0.6433624	0.66522252	0.6866878	1.00	5799
tau_gorx	15.6280994111	5.55144838	6.6885796	14.97753753	28.0989087	1.00	3662
tau_res	4.9356310165	0.09815468	4.7440515	4.93569337	5.1296871	1.00	13019

Summary of random effects



Comparison between software:

Upon comparative analysis of the outputs derived from the two Bayesian models, it is observed that model implemented in STAN have estimated the coefficients with a higher degree of confidence.

Notably, there exists a variance in the actual values and significance of some fixed effects coefficients across the two softwares.

The value and spread of random effects appears to be more consistent between the two softwares.

Model comparison

For my analysis, I used the WAIC (Widely Applicable Information Criterion) to compare various models. It is a statistical measure used to compare the fit of various statistical models while accounting for model complexity. WAIC is particularly useful in Bayesian model selection, as it provides a balance between model fit and simplicity by penalising models that are overly complex. The lower the criterion, the better the balance between the two.

Table of WAIC values

	Model 1	Model 2
WITHOUT Random effects	6527.2	6511.8
WITH Random effects	6412.8	6377.5

From the table above, we can see that models with random effects perform better than models without random effects. This indicates that by accounting for hierarchical structures, and therefore variations at different levels of the hierarchy, we can improve the accuracy of our model predictions. The best-performing model is Model 2, which includes Income, A093r, and G019r as our predictors, and a random intercept as our random effect.

Conclusion

My analysis of UK expenditures data, employing Frequentist linear regression, Bayesian regression with Nimble and STAN, and hierarchical Bayesian Regression with random effects, reveals that incorporating regional variability does enhance the accuracy of our models. The models from STAN software have consistently provided more robust coefficient estimates and greater predictive accuracy, as indicated by a smaller standard deviation of the residuals. The analytical evidence suggests that there are significant regional variations in expenditure habits when controlled for income, tenure, number of children and economic position. Furthermore, we can assert that expenditure is positively associated with income and is influenced by the tenure status of households and the number of children in the household.

It is important to note that certain categories within our dataset were underrepresented, which could potentially lead to biased results. This issue is evidenced by the greatly varying counts across regions, and by the unequal representation of subgroup sizes for variables such as tenure and number of children.

For future analyses, it would be prudent to include age groups in our models. As spending habits vary significantly at different stages of life, incorporating this demographic information could enhance our understanding of expenditure patterns across the UK.