British Panel Household Survey (BHPS) Project

Alexander von Paumgartten

Introduction:

In this report we analyze the British Panel Household survey (BHPS) 1991 (Wave 1) that has sample dataset of 3500 records. The BHPS is designed as a research resource for a wide range of social science disciplines and to support interdisciplinary research in many areas. The main objective of the survey is to further understanding of social and economic change at the individual and household level in Britain, to identify, model and forecast such changes, their causes and consequences in relation to a range of socio-economic variables.

The BHPS was designed as an annual survey of each adult (aged 16+) member of a nationally representative sample of more than 5,000 households, making a total of approximately 10,000 individual interviews. The same individuals are re-interviewed in successive waves and, if they split off from original households, all adult members of their new households are also interviewed.

Methods & Analysis:

The objective of this analysis was to determine how the various factors are associated with Annual Household Income of the individual. Initially we had 10 input variables that include demographic as well as socio-economic indicators such as sex, age, political affiliation, no. of cars, weekly hours worked, metropolitan area, labor status, current job, no. of persons in household and housing tenure. Since, our objective is to predict the Annual Household Income which is continuous in nature, **we will be using the machine learning technique of Multiple Linear Regression.**

The columns in the dataset are as below:

```
'data.frame':   3491 obs. of  11 variables:
 $ sex    : Factor w/ 2 levels "1","2": 1 2 2 2 2 1 1 1 1 1 ...
 $ aage   : int  52 54 25 49 39 37 30 20 40 39 ...
 $ AJBHRS : int  40 17 35 35 18 37 42 37 36 50 ...
 $ ancars : int  2 2 3 2 3 1 0 2 3 1 ...
 $ avote  : Factor w/ 4 levels "1","2","3","Others": 2 2 1 1 1 2 2 1 2 2 .
 $ aregion: Factor w/ 4 levels "Others","3","4",..: 1 2 1 1 3 1 2 2 1 2 ..
 $ ajbstat: Factor w/ 4 levels "Others","2","6",..: 2 2 2 2 2 2 2 2 2 2 ..
 $ ajbterm: Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
 $ ahhsize: int  4 4 3 4 4 1 2 2 3 5 ...
 $ atenure: Factor w/ 4 levels "1","2","3","Others": 1 4 2 2 2 3 3 4 1 3 .
 $ afihhyr: num  17226 15035 33254 25116 47035 ...
```

Variable	Description	Type
sex	Sex	Discrete
aage	Age at Date of Interview	Continuous
AJBHRS	No. of hours normally worked per week - Wave 1	Discrete
ancars	Car or van available for private use - Wave 1	Discrete
avote	Political party supported (wave 1)	Discrete
aregion	Region, Metropolitan Area - Wave 1	Discrete
ajbstat	Current labour force status - Wave 1	Discrete
ajbterm	Current job: permanent or temporary - Wave 1	Discrete
ahhsize	Number of Persons in Household - Wave 1	Continuous
atenure	Housing tenure - Wave 1	Discrete

afihhyr	Annual household income (1.9.90-1.9.91) - Wave 1	Discrete

We initiated the data analysis by viewing the summary of the dataset that helped us to understand the data in summarized form using descriptive statistics.

```
      sex             aage           AJBHRS          ancars          avote           aregi
 Min.   :1.000   Min.   :15.00   Min.   : 1.0   Min.   :0.000   Min.   : 1.000   Min.   :
 1st Qu.:1.000   1st Qu.:27.00   1st Qu.:30.0   1st Qu.:1.000   1st Qu.: 1.000   1st Qu.:
 Median :1.000   Median :37.00   Median :37.0   Median :1.000   Median : 2.000   Median :
 Mean   :1.492   Mean   :37.23   Mean   :33.8   Mean   :1.333   Mean   : 2.655   Mean   :
 3rd Qu.:2.000   3rd Qu.:46.00   3rd Qu.:40.0   3rd Qu.:2.000   3rd Qu.: 3.000   3rd Qu.:
 Max.   :2.000   Max.   :79.00   Max.   :97.0   Max.   :3.000   Max.   :11.000   Max.   :
    ajbstat         ajbterm         ahhsize         atenure          afihhyr
 Min.   : 1.000   Min.   :1.000   Min.   :1.000   Min.   :-9.000   Min.   :   259.9
 1st Qu.: 2.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.: 2.000   1st Qu.: 15076.1
 Median : 2.000   Median :1.000   Median :3.000   Median : 2.000   Median : 21408.3
 Mean   : 2.158   Mean   :1.123   Mean   :3.089   Mean   : 2.344   Mean   : 23599.3
 3rd Qu.: 2.000   3rd Qu.:1.000   3rd Qu.:4.000   3rd Qu.: 2.000   3rd Qu.: 29345.0
 Max.   :10.000   Max.   :3.000   Max.   :9.000   Max.   : 8.000   Max.   :195640.3
```

We observed that there are records that have missing values (coded as -9) and also, the variables such as sex, aregion etc. are categorical variables but being considered as continuous variables. As a next step, we remove the records with these missing values and then transform the columns with categorical values into factors. We still observed that there exists multiple categories with very low frequency, and hence, apart from the top 3 categories, we combine all the remaining categories as 'Others' that makes the data manageable and improve its usability.

```
sex           aage           AJBHRS          ancars          avote           aregi
1:1771   Min.   :15.00   Min.   : 1.0   Min.   :0.000   1       :1358   3       :
2:1720   1st Qu.:27.00   1st Qu.:30.0   1st Qu.:1.000   2       :1226   18      :
         Median :37.00   Median :37.0   Median :1.000   3       : 416   4       :
         Mean   :37.24   Mean   :33.8   Mean   :1.333   10      : 251   6       :
         3rd Qu.:46.00   3rd Qu.:40.0   3rd Qu.:2.000   6       :  71   2       :
         Max.   :79.00   Max.   :97.0   Max.   :3.000   4       :  51   8       :
                                                        (Other): 118   (Other):
   ajbstat      ajbterm      ahhsize         atenure          afihhyr
2       :3348   1:3178   Min.   :1.000   2       :2279   Min.   :   259.9
6       :  94   2: 195   1st Qu.:2.000   1       : 455   1st Qu.: 15074.3
9       :  15   3: 118   Median :3.000   3       : 404   Median : 21400.6
1       :   9            Mean   :3.088   7       : 117   Mean   : 23588.2
5       :   8            3rd Qu.:4.000   6       : 110   3rd Qu.: 29310.5
3       :   6            Max.   :9.000   4       :  63   Max.   :195640.3
(Other):  11                             (Other):  63
```

We then start with the data analysis by visualizing box-plots for all the categorical variables and scatterplots for continuous variables in order to identify their relationship with Annual Household Income (afihhyr). The box-plot analysis highlighted the outliers present in the data (Fig.1) within various categorical variables (Fig.1).

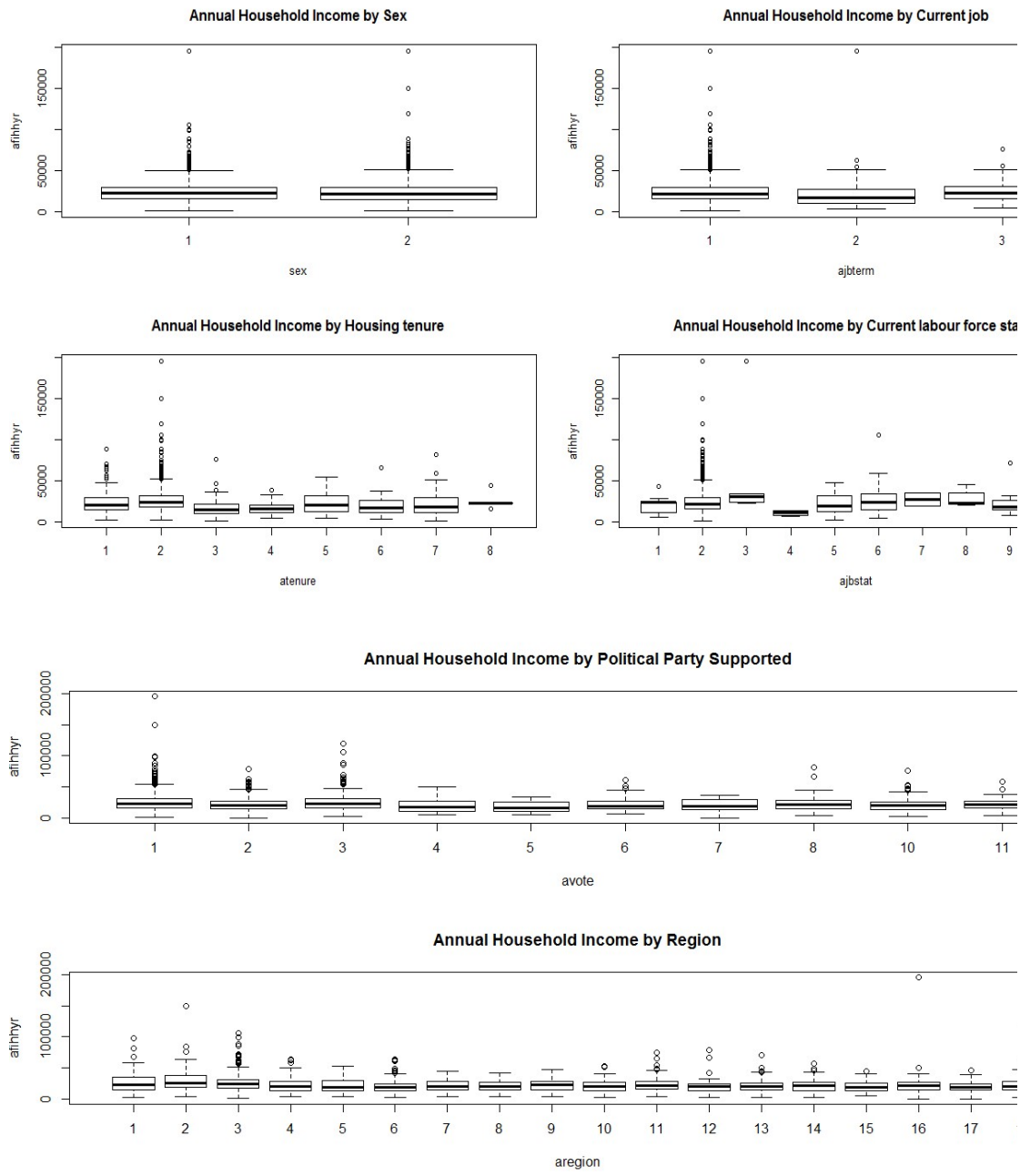

Fig 1: Box plot for all categorical variables

Using scatterplots we identified that afihhyr has strong positive linear relationship with ancars, AJBHRS and ahhsize, while some degree of linear relationship with aage.

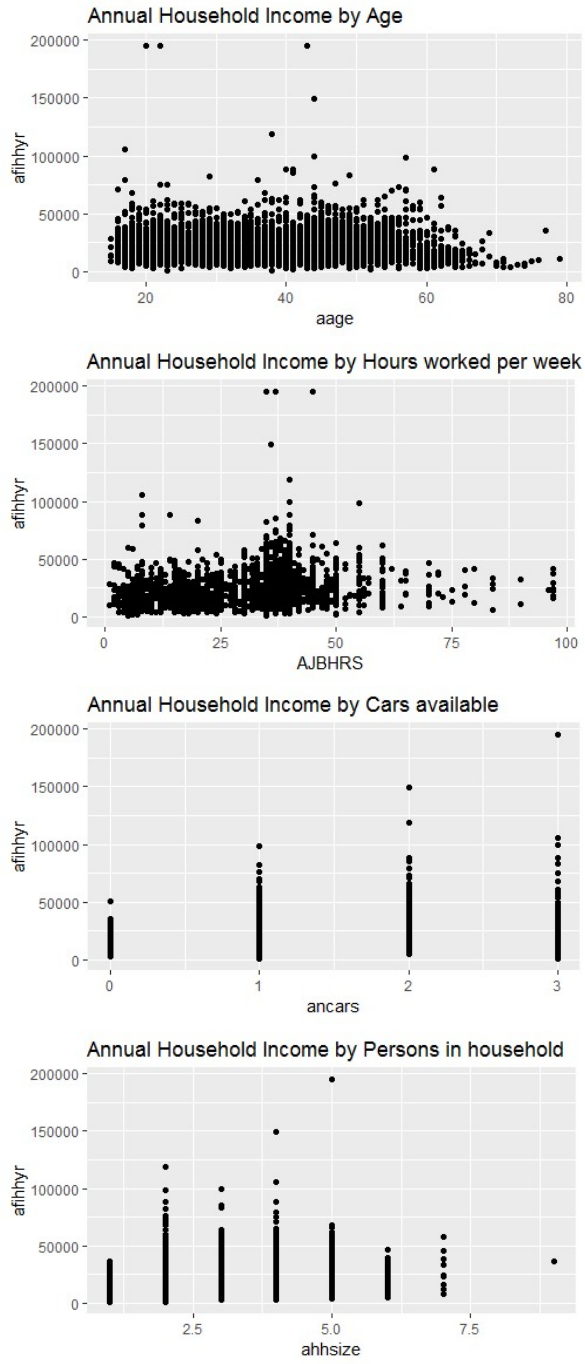

Fig 2: Scatter plot for all continuous variables

To identify the significance of categorical variables, we further performed Analysis-of-Variance (ANOVA) to determine if the classes within the categorical variable have similar afihhyr. We observed that variables avote, aregion and atenure have p-value <0.05 and hence, there is a statistical difference between the classes with respect to afihhyr.

```
> summary(sex_anova)
             Df    Sum Sq   Mean Sq F value Pr(>F)
sex           1 5.452e+08 545203150   2.936 0.0867 .
Residuals  3489 6.480e+11 185714123
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(avote_anova)
             Df    Sum Sq   Mean Sq F value  Pr(>F)
avote         9 1.461e+10 1.624e+09   8.917 2.1e-13 ***
Residuals  3481 6.339e+11 1.821e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(aregion_anova)
             Df    Sum Sq   Mean Sq F value Pr(>F)
aregion      17 2.593e+10 1.525e+09   8.508 <2e-16 ***
Residuals  3473 6.226e+11 1.793e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(ajbstat_anova)
             Df    Sum Sq   Mean Sq F value   Pr(>F)
ajbstat       9 7.509e+09 834329318   4.531 5.96e-06 ***
Residuals  3481 6.410e+11 184140424
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(ajbterm_anova)
             Df    Sum Sq   Mean Sq F value Pr(>F)
ajbterm       2 1.178e+09 589096547   3.174 0.0419 *
Residuals  3488 6.473e+11 185585891
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(atenure_anova)
             Df    Sum Sq   Mean Sq F value Pr(>F)
atenure       7 4.195e+10 5.992e+09   34.41 <2e-16 ***
Residuals  3483 6.066e+11 1.741e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Fig 3: ANOVA test with Annual Household Income for all continuous variables

However, after going through a series of statistical tests and analysis, the final variables that were significant in determining the Average Household Income of the individual are sex, age, political affiliation, no. of cars, weekly hours worked, metropolitan area, labor status, no. of persons in household and housing tenure.

Results and Exploratory Data Analysis:

We then ran the model Model_1 which included all available variables, sex, aage, AJBHRS, ancars, avote, aregion, ajbstat, ajbterm, ahhsize and atenure. Goodness of fit statistics for this model are shown in Table 1 below. Coefficients with asterisks are significant at the 5 % level. All the variables except ajbterm were significant at the 5% level (if we consider ajbterm as a single variable). The R2 and Adj R2 were just above 0.2, which indicate not-so-good fit for model on actual data. They were also within 1% of one another. The F -statistic was highly significant with a value of less than 5%, indicating overall good fit of the model. The Standard Error of the model was very high which indicates a very high standard deviation within the data.

Coefficients	Model_1	Model_1_best	Model_2_log
(Intercept)	*	*	***
sex2	***	***	***
aage	**	**	**
AJBHRS	***	***	***
ancars	***	***	***
avote2	**	**	
avote3			
avoteOthers	***	***	***
aregion3	*	*	**
aregion4	***	***	**
aregion18	**	**	***
ajbstat2	*	*	
ajbstat6			.
ajbstat9	*	.	.
ajbterm2		N/A	.
ajbterm3		N/A	.
ahhsize	***	***	***
atenure2	***	***	***
atenure3	**	**	***
atenureOthers			*
R2/ Adj. R2	0.2426/0.2384	0.2424/0.2387	0.3275/0.3238
F-Statistic Significance	< 2.2e-16	< 2.2e-16	< 2.2e-16
Std. Error	11900	11890	0.4561

Table 1: Goodness of Fit Statistics

As ajbterm was non-significant at the 5% level, it was removed from the regression Model_1. Using the backward elimination method of regression, we then re-ran the model, which formed Model_1_best. Coefficients and goodness of fit statistics for Model_1_best are also shown in Table 1. As can be seen from these statistics the models were comparable in terms fit, with Model_1_best having only significant variables. The Adj R2 also improved by a small amount. This confirm that our initial insight about ajb_term being less significantly related with afihhyr.

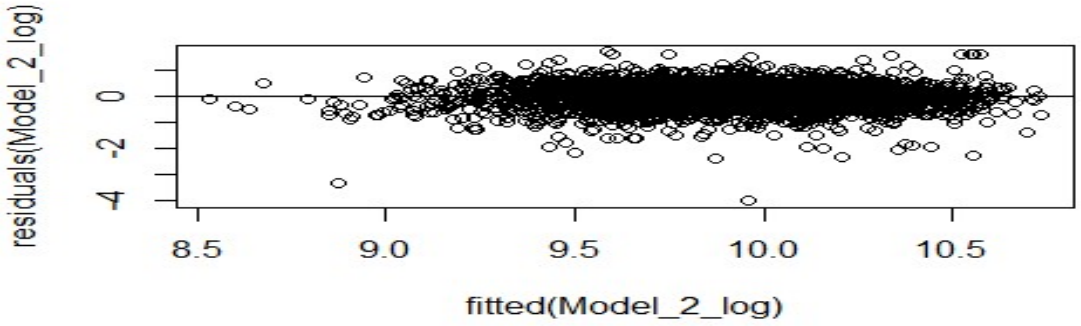

Fig 3: Residual plot for Model_2_log

In Table 2 below are the results for all the relevant regressions we ran.

Coefficients	Model_1	P-values	Model_1_best	P-values	Model_2_log	P-values
(Intercept)	5272.39	0.038076	5262.38	0.036641	8.25535	< 2e-16
sex2	.	0.000417	1606.22	0.000401	0.05609	0.000942
aage	55.55	0.001928	55.61	0.001874	0.06427	0.008378
AJBHRS	176.61	< 2e-16	176.79	< 2e-16	0.23357	< 2e-16
ancars	5886.35	< 2e-16	5884.96	< 2e-16	0.22188	< 2e-16
avote2	-1282.78	0.008585	-1282.38	0.008588	-0.02417	0.196566
avote3	14.11	0.98315	19.62	0.976567	0.01295	0.613244
avoteOthers	-2394.2	0.000258	-2416.47	0.000223	-0.09089	0.000308
aregion3	1279.74	0.015854	1280.25	0.015795	0.06526	0.001341
aregion4	-2716.15	0.000293	-2719.85	0.000287	-0.08203	0.004337
aregion18	2328.34	0.001615	2370.3	0.001293	0.10814	0.000135
ajbstat2	-4721.72	0.021839	-4719.02	0.021662	-0.04462	0.572097
ajbstat6	-1994.85	0.41989	-2056.85	0.404122	0.1663	0.083136
ajbstat9	-7851.96	0.03785	-7232.52	0.051495	-0.27415	0.05904
ajbterm2	-253.52	0.788007	N/A	N/A	-0.06877	0.059539
ajbterm3	988.06	0.392159	N/A	N/A	0.08337	0.06043
ahhsize	1733.34	< 2e-16	1734.27	< 2e-16	0.2993	< 2e-16
atenure2	2745.8	0.000016	2759.24	0.000014	0.13062	6.38E-08
atenure3	-2549.28	0.002995	-2538.44	0.00311	-0.19679	2.12E-09
atenureOthers	-588.51	0.501472	-556.44	0.524666	-0.06774	0.042426

Table 2: Regression results for all variables

We can observe that sex of a person is an important parameter in determining the annual household income (afihhyr) as it is significant in all the models. As the age of the person increases by a year, the afihhyr increases by nearly 54 points, while when using log transformations, with every unit increases in log(aage), the log(afihhyr) increases by 0.064 points. As the AJBHRS increases by a point, the afihhyr increases by nearly 177 points, whereas when using log transformations, with every unit increases in log(AJBHRS), the log(afihhyr) increases by 0.233 points. The number of cars/vans available for private use is a very strong variable as with every single point, the afihhyr increases by 5884 points while with every point increase in log(ancars), the log(afihhyr) increases by 0.221 points. The variable political party supported has a positive impact on the afihhyr when it supports the category 3, and a negative impact when it supports category 2 or any other, however, since the p-value is not less than 0.05, it is susceptible to deviations from the above impact. The region/ metropolitan area of the person i.e. variable aregion has a positive impact on the afihhyr when the person is from category 3 or 18, and a negative impact when is from category 4. The

current workforce status of the person i.e. variable ajbstat has a negative impact on the afihhyr. The current job i.e. variable ajbtern had a low significance in the Model_1 and hence, it was removed in Model_1_best; however, when using the log model i.e. Model_2_log, it came out be a highly significant variable. The number of persons in the household i.e. variable ahhsize has been significant throughout all the models. As the ahhsize increases by a point, the afihhyr increases by nearly 1741 points, whereas when using log transformations, with every unit increases in log(ahhsize), the log(afihhyr) increases by 0.299 points. The housing tenure i.e. variable atenure has a negative impact on the afihhyr and there is a decrease of nearly 11212 points in the afihhyr when the tenure is of category 3.

Although there are certain variables in Model_2_log that have p-value of more than 0.05, they are still significant in predicting the annual household income, as when removing any of them further reduced the r-squared/ adj. r-squared value of the model. Model_2_log was the final model that resulted in the highest adjusted r-squared value and lowest standard error.

The adjusted r-squared value of our first model i.e. Model_1 was 0.2384 when we ran it using all the available variables as Model_1_best, which slightly increased to 0.2387 when we removed the insignificant variable 'ajbterm' from the model, however the standard error was 11920 which is considered to be very high. After introducing log transformations in the data, the adjusted r-squared increased to 0.3238 with a standard error of 0.4561, which is highly significant as compared to the previous models.

The final regression equation is:

log(Average Household Income) =

(8.25535) + (0.05609*sex2) + (0.06427*aage) + (0.23357*AJBHRS) + (0.22188*ancars) + (-0.02417 *avote2) + (0.01295*avote3) + (-0.09089*avoteOthers) + (0.06526*aregion3) + (-0.08203*aregion4) + (0.10814*aregion18) + (-0.04462*ajbstat2) + (0.1663*ajbstat6) + (-0.27415*ajbstat9) + (-0.06877 *ajbterm2) + (0.08337*ajbterm3) + (0.2993*ahhsize) + (0.13062*atenure2) + (-0.19679*atenure3) + (-0.06774*atenureOthers)

The highest value for coefficient of determination i.e. adjusted r-squared that we achieved was 0.3238, which means that the current model is able to explain 32% of the variability within the data to predict the value of annual household income of the person. This sounds reasonable, as we are using very limited variables in the current dataset out of the various other variables that are captured during the survey and might be significant in the prediction. Hence, we it is safe to say that we need more explanatory variables and a larger dataset to better understand the factors that are significant in determining the annual household income.

Summary and Conclusion:

We were given a sample dataset of 3500 records out of the total sample of 5000 households. From this sample, we had various information such as gender, age of the person, average working hours per week, number of car or van for private use, political party supported, region/ metropolitan area, current labour force status, current job, number of persons in Household, housing tenure and annual household income. The objective of this analysis was to understand the factors that are related to the annual household income of a person and how do they impact it.

We initiated this analysis by performing some visual analysis to understand how an individual factor is impacting the annual household income, such as how the annual household income is different for men and women, how does the value of annual household income changes at an average if we reduce or increase the number of cars or number or persons in the household. Having a look at these individual factors visually with respect to the annual household income will help us better understand the relationship between them.

In the next step we ran the regression using all the available factors, and analyzed the results. We could observe that there are certain factors that are not at all related to the annual household income and including them in the regression will do more harm to our analysis than good. Hence, we removed those factors and re-ran our regression analysis until we arrived at a point where all the factors were related to annual household income. This model was good however, there was still some scope of improvements to make it better. As the next step, we tried to transform certain factors of our data by bringing them into a similar scale i.e. same range and re-ran the regression on all available factors. In the results of the regression, we observed that all the factors when transformed to a similar scale were now related to the transformed value of annual household income. We then started to look for the extreme case or less frequent values in our data and remove them in order to further clean our data. Running the regression on this cleaned data further improved the score that is used to measure how the interactions between all the factors is related.

The score that we use to measure the degree of strength for relationship between the available factors and annual household income was 34%. This means that are still some more factors that are related to the annual household income and can help us predict annual household income for households that we haven't surveyed. Also, the size of our dataset is not large enough to give us robust insights. The inference from this data may have some variations depending upon the various other factors that might be captured in our data or factors that are not a part of our data.

Age at the time of survey was one of strong factor to determine the annual household income along with factors such as number of cars owned, number of hours worked per week, average number of persons in the household etc. This is in-line with our intuitive knowledge that as the age of the person increases, the annual household income also increases. Similarly, increase in the number of private cars suggests improvement in the annual household income as well; higher the number of persons in the household better the annual household income of the household and higher the number of hours worked per week suggested higher annual household income. There were certain jobs and areas that displayed a positive impact on the annual household income whereas certain jobs and areas displayed negative impact on the annual household income. Interestingly, how the household had affinity to a political party was also related in determining the annual household income. Using the above knowledge about what are the factors that impact the annual household income and how, we can determine the annual household income for the household basis their given information about these factors.