Predicting Household Income in Later Life: Insights from the SHARE Survey in Sweden

Bin Han

2025-07-11

Table of contents

1 Project description

This study investigates the key predictors of household income among older adults in Sweden using data derived from the SHARE (Survey of Health, Ageing and Retirement in Europe) dataset. The sample includes 420 individuals aged 50 and above, collected during wave 3 and 4 (2008–2011). Using multiple linear regression with a log-transformed outcome variable, we identify the most significant socioeconomic and early-life factors that influence household income. Our final model shows that household income increases with the number of household members, being male, and living with a partner, while it decreases for retired

individuals, those who had fewer books at age 10, and individuals with lower self-rated math or language ability at age 10. **After filtering out extreme income values, the model achieved a mean absolute percentage error (MAPE) of 42.7%, an RMSE of approximately 13,851 SEK, and a MAE of about 10,041 SEK, indicating moderate predictive accuracy.** Although the model is not suitable for precise income prediction, it effectively captures general income trends and highlights the long-term economic influence of early-life conditions and household composition.

2 Load data and R library

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.4
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to be
```

```
library(dplyr)
library(corrplot)
```

```
corrplot 0.95 loaded
```

```
library(car)
```

```
Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:dplyr':

    recode

The following object is masked from 'package:purrr':

    some
```

```r
data<-read.csv("share_selected.csv")
data<-data%>%
  filter(household_income>=3000 & household_income<=75000)
```

3 Exploratory data analysis

3.1 Data overview

```r
glimpse(data)
```

```
Rows: 387
Columns: 11
$ sex                                <chr> "female", "female", "male", "femal~
$ lives_with_partner                 <chr> "yes", "yes", "yes", "yes", "yes",~
$ n_household                        <int> 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2~
$ retired                            <chr> "yes", "yes", "no", "yes", "yes", ~
$ age                                <int> 79, 73, 56, 76, 73, 56, 61, 79, 72~
$ age_partner                        <int> 76, 74, 57, 72, 69, 60, 63, NA, 74~
$ years_of_education                 <int> 12, 11, 11, 12, 8, 7, 15, 5, 11, 2~
$ household_income                   <int> 16525, 40677, 56396, 47009, 17716,~
$ books_age_10                       <chr> ">25", ">25", ">25", ">25", "0-25"~
$ relative_math_ability_at_age_10    <chr> "better", "same/worse", "same/wors~
$ relative_language_ability_at_age_10 <chr> "better", "same/worse", "same/wors~
```

```r
summary(data)
```

```
     sex            lives_with_partner  n_household      retired
 Length:387         Length:387         Min.   :1.000   Length:387
 Class :character   Class :character   1st Qu.:2.000   Class :character
 Mode  :character   Mode  :character   Median :2.000   Mode  :character
                                       Mean   :1.819
                                       3rd Qu.:2.000
                                       Max.   :5.000

      age          age_partner    years_of_education household_income
 Min.   : 50.00   Min.   :43.00   Min.   : 0.00      Min.   : 3297
 1st Qu.: 65.00   1st Qu.:65.00   1st Qu.: 9.00      1st Qu.:16154
 Median : 71.00   Median :68.00   Median :11.00      Median :26191
 Mean   : 70.18   Mean   :68.52   Mean   :11.31      Mean   :28563
 3rd Qu.: 75.00   3rd Qu.:73.00   3rd Qu.:14.00      3rd Qu.:37837
 Max.   :100.00   Max.   :88.00   Max.   :23.00      Max.   :74779
```

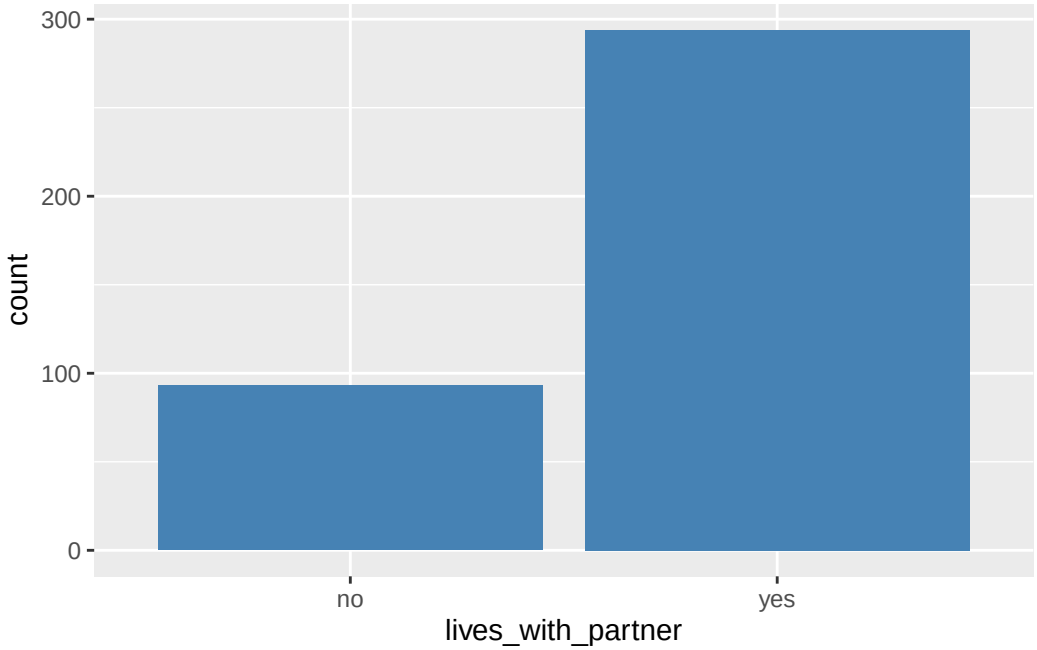
```
                 NA's   :93
 books_age_10       relative_math_ability_at_age_10
 Length:387         Length:387
 Class :character   Class :character
 Mode  :character   Mode  :character
```

```
 relative_language_ability_at_age_10
 Length:387
 Class :character
 Mode  :character
```

```
data|>
  count(sex, name="count")
```

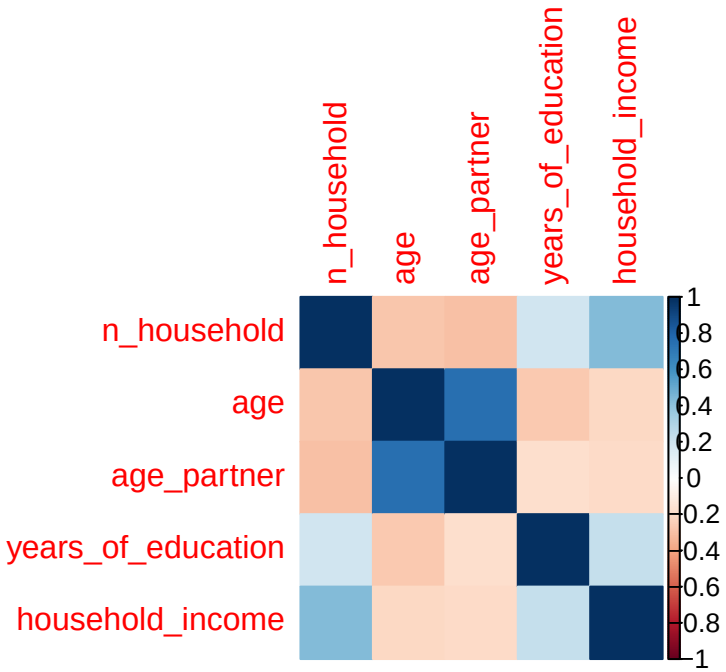
```
     sex count
1 female   216
2   male   171
```

```
data%>%
  ggplot(aes(x=lives_with_partner))+
  geom_bar(fill="steelblue")
```



3.2 Compute correlations

```
data%>%
  select(where(is.numeric))%>%
  cor(use="pairwise.complete.obs")%>%
  corrplot(method="color")
```



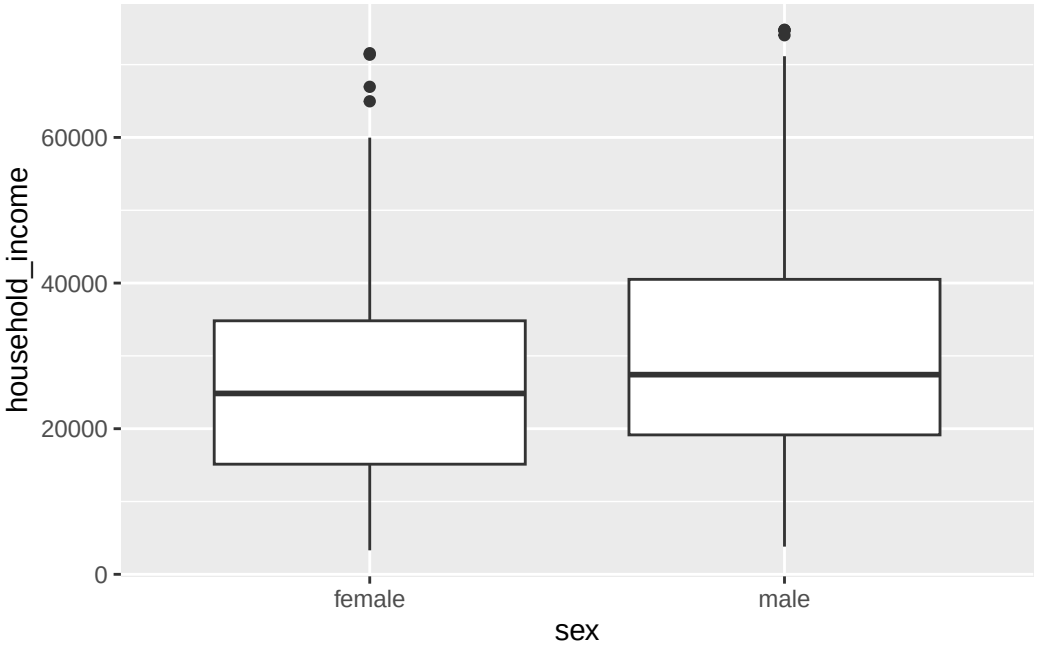
From the correlation plot, we see the household income is proportional to the number of people in the household and the years of education. And the household income is inverse proportional to the age of participant and age of the participant's partner.

3.3 Boxplot and lineplot

Convert the character to the factor first.

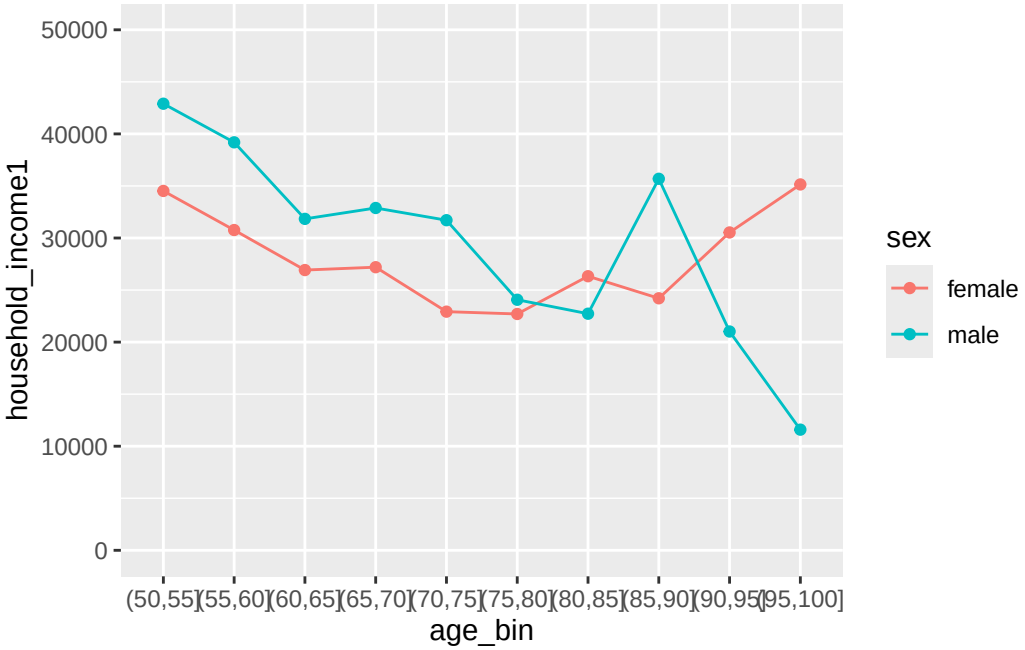
```
data<-data%>%
  mutate(across(where(is.character), as.factor))
```

```
data%>%
  ggplot(aes(x=sex, y=household_income))+
  geom_boxplot()
```




```
data%>%
  mutate(age_bin=cut(age, breaks=10))%>%
  group_by(sex, age_bin)%>%
  summarise(household_income1=mean(household_income))%>%
  ggplot(aes(x=age_bin, y=household_income1, color=sex, group=sex))+
  geom_line()+
  geom_point()+
  ylim(0, 50000)
```

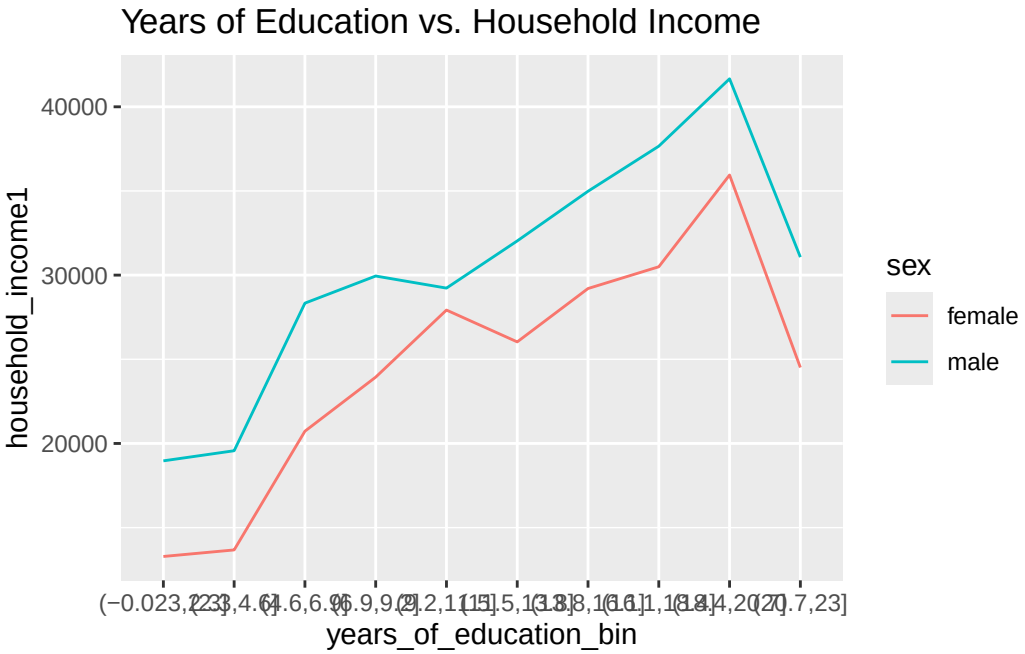
```
`summarise()` has grouped output by 'sex'. You can override using the `.groups`
argument.
```



From the line plot, we can see that the incomes of both males and females decrease between ages 50 and 80. After age 80, female income increases, while male income increases until age 90 and then decreases again until age 100.

```
data%>%
  mutate(years_of_education_bin=cut(years_of_education, breaks=10)) %>%
  group_by(sex, years_of_education_bin) %>%
  summarise(household_income1=mean(household_income))%>%
  ggplot(aes(x=years_of_education_bin, y=household_income1, group=sex, color=sex))+
  geom_line()+
  labs(title="Years of Education vs. Household Income")
```

```
`summarise()` has grouped output by 'sex'. You can override using the `.groups`
argument.
```



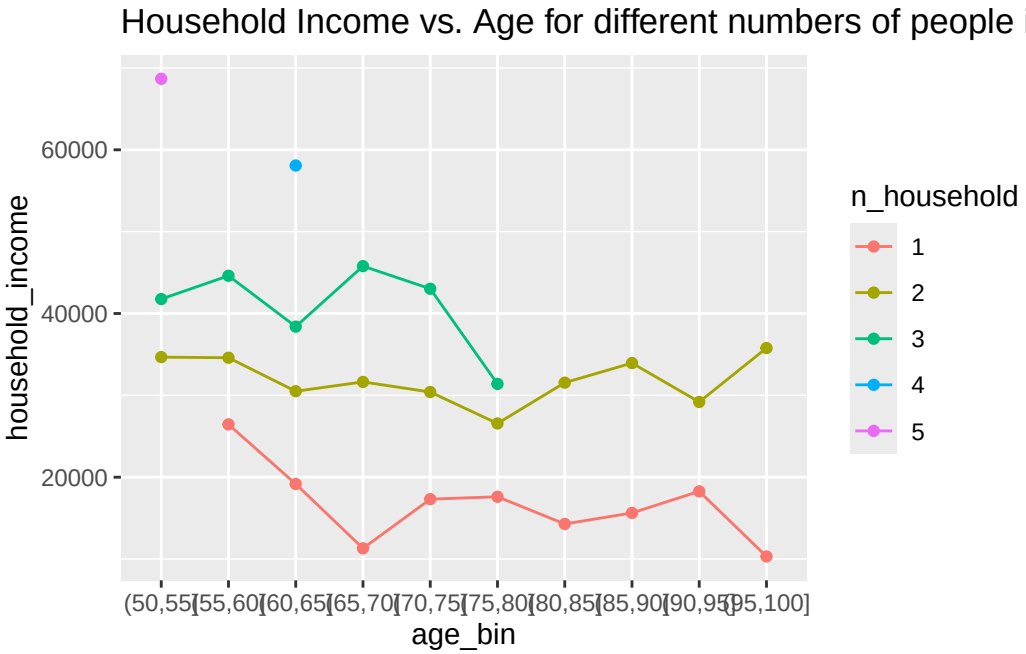
Household income increases with years of education up to 20 years for both males and females. However, from 20 to 23 years of education, household income decreases.

```
data%>%
  mutate(age_bin=cut(age, breaks=10),
         retired=factor(retired, levels=c("no", "yes"), labels=c("Non-Retired", "Retired"))
  group_by(sex, retired, age_bin)%>%
  summarise(household_income1=mean(household_income),.groups = "drop")%>%
  ggplot(aes(x=age_bin, y=household_income1, group=sex, color=sex))+
  geom_line()+
  facet_wrap(vars(retired))+
  labs(title="Household Income vs. Age")
```

For the non-retired 50–75-year-old male, household income decreases. For the non-retired 50–70-year-old female, household income decreases and then increases afterward. Males have higher income than females between the ages of 50 and 70; after age 70, females have more income.

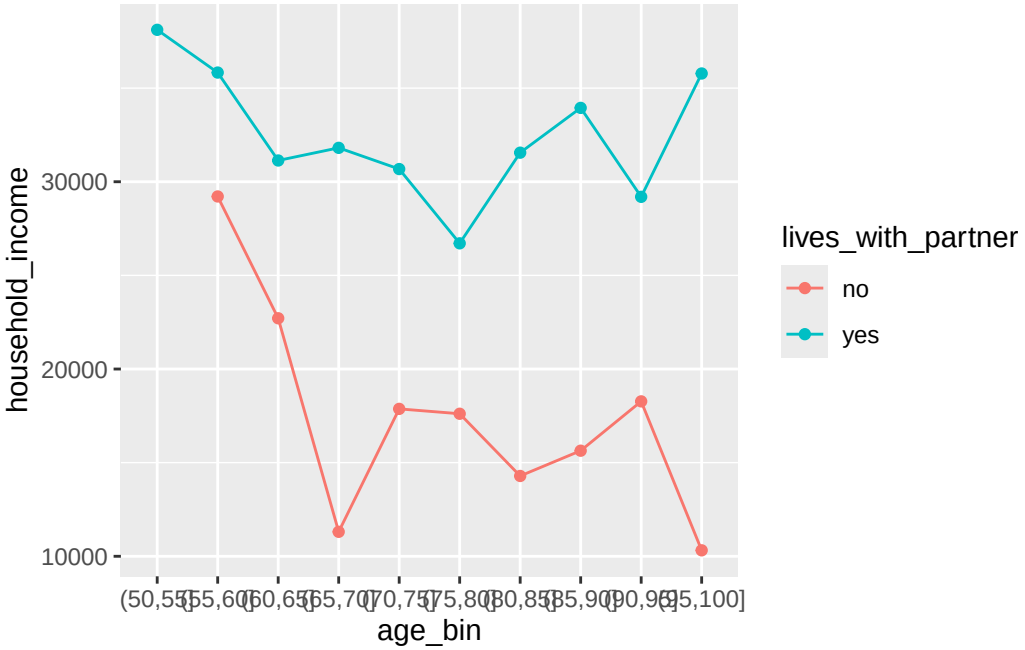
For the retired 60–100-year-old male, household income generally decreases. For retired females, household income increases from age 80 to 100. Males have higher household income than females from age 60 to 90, but females have more income than males from age 90 to 100.

```
data%>%
  mutate(age_bin=cut(age, breaks=10),
         n_household=factor(n_household, levels=c("1", "2", "3", "4", "5"), labels=c("1",
  group_by(age_bin, n_household)%>%
  summarise(household_income=mean(household_income),.groups="drop")%>%
  ggplot(aes(x=age_bin, y=household_income, group=n_household, color=n_household))+
  geom_point()+
  geom_line()+
  labs(title="Household Income vs. Age for different numbers of people in the Household")
```



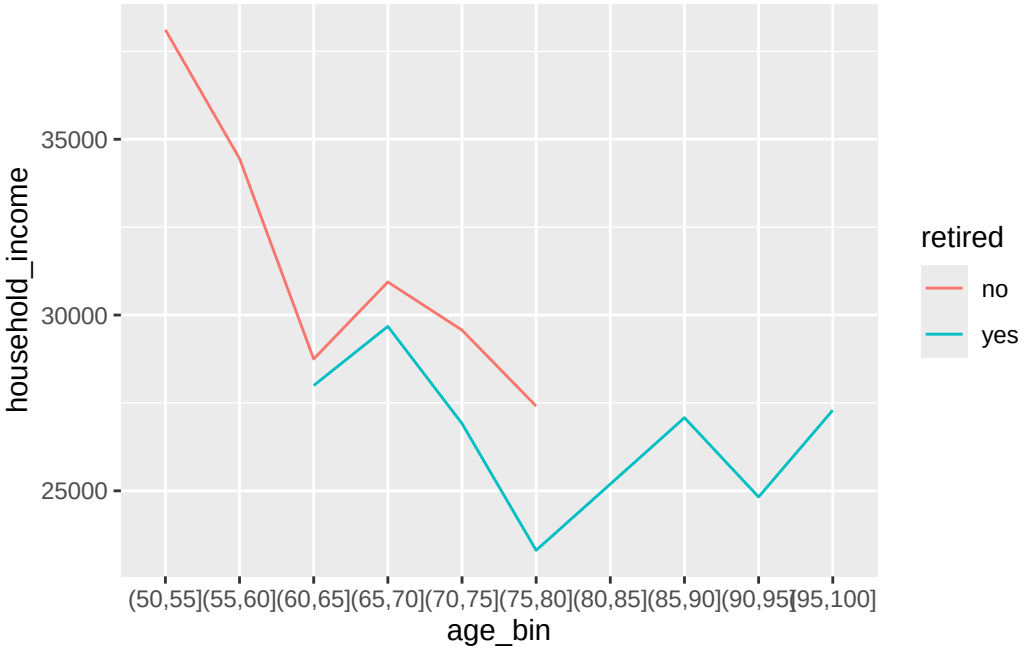
From the line plot, we can see that, generally, households with more members tend to have higher incomes. For different household sizes, household income decreases with increasing age.

```
data%>%
  mutate(age_bin=cut(age, breaks=10))%>%
  group_by(lives_with_partner, age_bin)%>%
  summarise(household_income=mean(household_income), .groups="drop")%>%
  ggplot(aes(x=age_bin, y=household_income, group=lives_with_partner, color=lives_with_par
  geom_point()+
  geom_line()
```



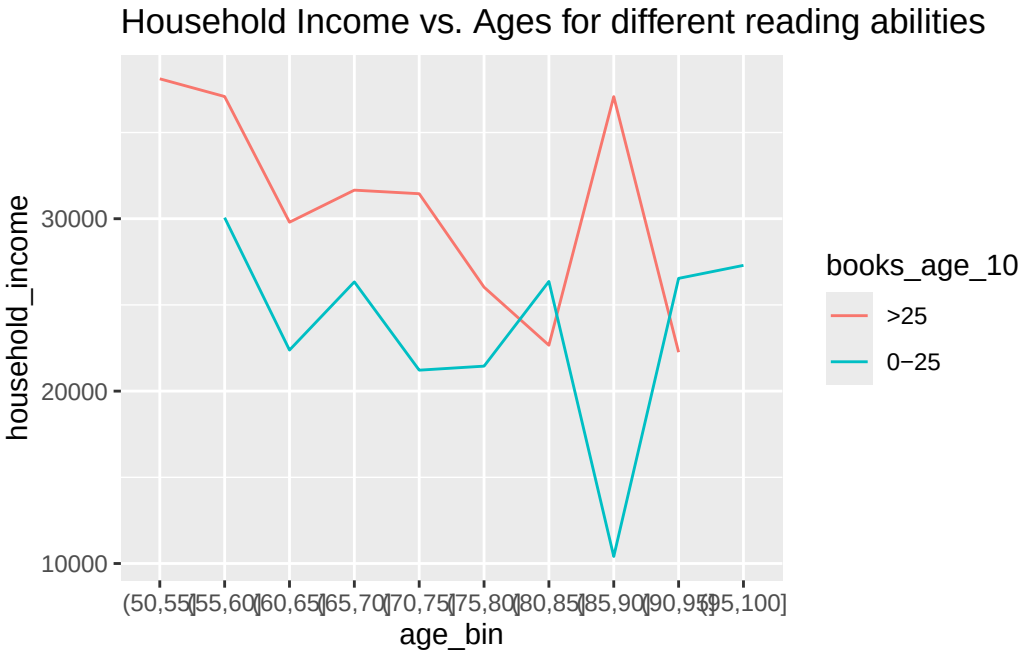
From the line plot, we can see that households with partners tend to have higher incomes than those living alone between ages 55 and 100. Between ages 50 and 55, single-person households have higher incomes than partnered ones.

```
data%>%
  mutate(age_bin=cut(age, breaks=10))%>%
  group_by(retired, age_bin)%>%
  summarise(household_income=mean(household_income), .groups="drop")%>%
  ggplot(aes(x=age_bin, y=household_income, group=retired, color=retired))+
  geom_line()
```



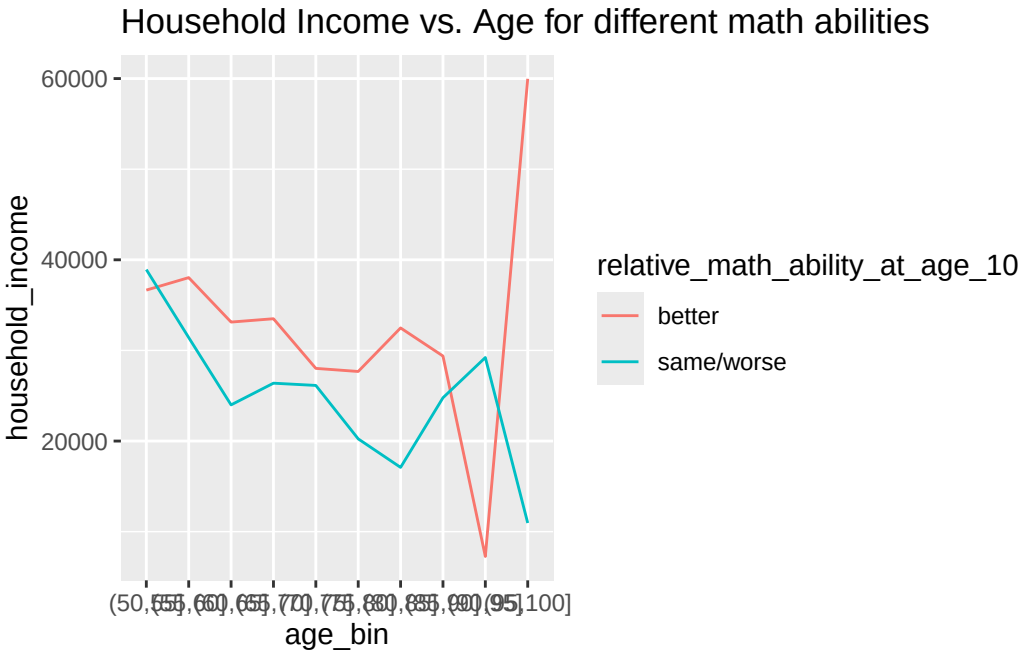
From the line plot, we see the non-retired people between 60-80 years old always have higher income than the retired people. For the non-retired people, the household income decreases with the age increases. For the retired people, the household income increases after age 80.

```
data%>%
  mutate(age_bin=cut(age, breaks=10))%>%
  group_by(age_bin, books_age_10)%>%
  summarise(household_income=mean(household_income), .groups="drop")%>%
  ggplot(aes(x=age_bin, y=household_income, group=books_age_10, color=books_age_10))+
  geom_line()+
  labs(title="Household Income vs. Ages for different reading abilities")
```



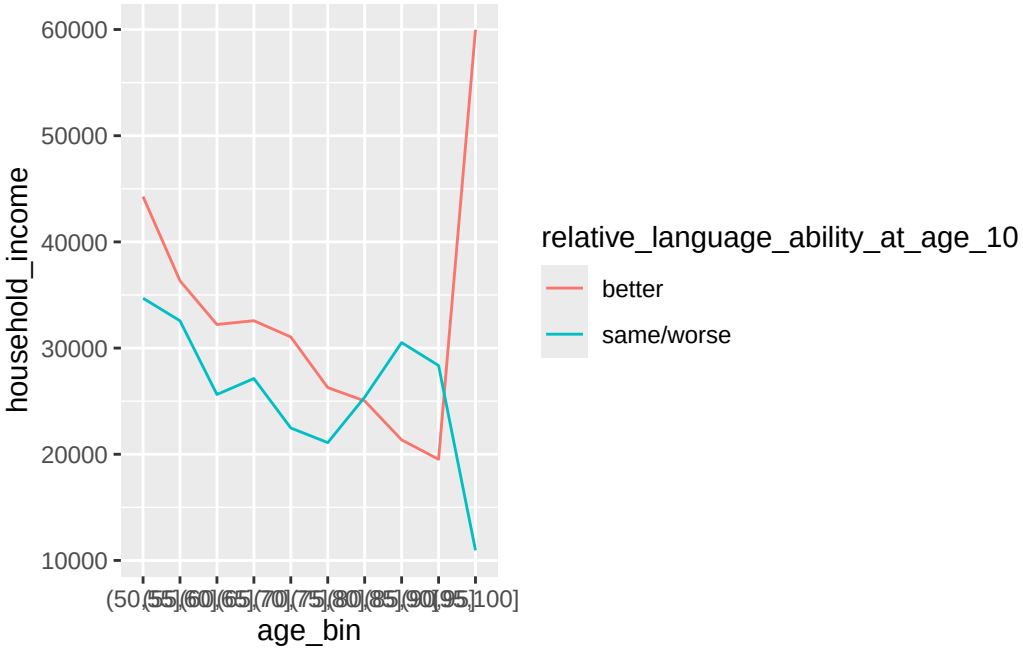
The household income for the participants read over 25 books in bookshelf at age ten is higher than the people who read less than 25 books.

```
data%>%
  mutate(age_bin=cut(age, breaks=10))%>%
  group_by(relative_math_ability_at_age_10, age_bin)%>%
  summarise(household_income=mean(household_income), .groups="drop")%>%
  ggplot(aes(x=age_bin, y=household_income, color=relative_math_ability_at_age_10, group=r
  geom_line()+
  labs(title="Household Income vs. Age for different math abilities")
```



Households with better math ability at age 10 have higher incomes than those with lower math ability across all age periods except [90, 95].

```
data%>%
  mutate(age_bin=cut(age, breaks=10))%>%
  group_by(age_bin, relative_language_ability_at_age_10) %>%
  summarise(household_income=mean(household_income), .groups="drop") %>%
  ggplot(aes(x=age_bin, y=household_income, group=relative_language_ability_at_age_10, col
  geom_line()
```



From the line plot, households with better language ability at age 10 have higher income than those with lower language ability, except between ages 80 and 95.

4 Linear regression

4.1 Initial model

If we put all the variables into our model analysis, there is always level problem. because of the column, lives with partner and age of partner. First we remove the lives with partner.

```
data_clean=data%>%
  select(-c(lives_with_partner))
 model<-lm(log(household_income)~., data=data_clean)
summary(model)
```

```
Call:
lm(formula = log(household_income) ~ ., data = data_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2581 -0.2831 -0.0211  0.3278  1.1013

Coefficients:
```

```
                                            Estimate Std. Error t value
(Intercept)                                10.236685   0.538606  19.006
sexmale                                     0.034368   0.069533   0.494
n_household                                 0.229057   0.114963   1.992
retiredyes                                 -0.131987   0.091011  -1.450
age                                         0.005361   0.006118   0.876
age_partner                                -0.009783   0.008342  -1.173
years_of_education                          0.010880   0.008353   1.303
books_age_100-25                           -0.222220   0.062411  -3.561
relative_math_ability_at_age_10same/worse  -0.062944   0.059457  -1.059
relative_language_ability_at_age_10same/worse -0.182082   0.060856  -2.992
                                           Pr(>|t|)
(Intercept)                                 < 2e-16 ***
sexmale                                    0.621501
n_household                                0.047281 *
retiredyes                                 0.148096
age                                        0.381571
age_partner                                0.241900
years_of_education                         0.193800
books_age_100-25                           0.000434 ***
relative_math_ability_at_age_10same/worse  0.290653
relative_language_ability_at_age_10same/worse 0.003015 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4667 on 284 degrees of freedom
  (93 observations deleted due to missingness)
Multiple R-squared:  0.1927,    Adjusted R-squared:  0.1671
F-statistic: 7.532 on 9 and 284 DF,  p-value: 6.811e-10
```

```r
step_model=step(model, direction=c("both"), trace=1)
```

```
Start:  AIC=-438.25
log(household_income) ~ sex + n_household + retired + age + age_partner +
    years_of_education + books_age_10 + relative_math_ability_at_age_10 +
    relative_language_ability_at_age_10

                                     Df Sum of Sq    RSS     AIC
- sex                                 1   0.05321 61.915 -440.00
- age                                 1   0.16729 62.030 -439.45
- relative_math_ability_at_age_10     1   0.24413 62.106 -439.09
- age_partner                         1   0.29956 62.162 -438.83
- years_of_education                  1   0.36954 62.232 -438.50
<none>                                           61.862 -438.25
- retired                             1   0.45813 62.320 -438.08
```

```
- n_household                           1   0.86472 62.727 -436.17
- relative_language_ability_at_age_10  1   1.95003 63.812 -431.12
- books_age_10                         1   2.76151 64.624 -427.41

Step:  AIC=-440
log(household_income) ~ n_household + retired + age + age_partner +
    years_of_education + books_age_10 + relative_math_ability_at_age_10 +
    relative_language_ability_at_age_10

                                      Df Sum of Sq    RSS     AIC
- relative_math_ability_at_age_10      1   0.29180 62.207 -440.61
- age                                  1   0.31649 62.232 -440.50
- years_of_education                   1   0.34421 62.260 -440.37
- retired                              1   0.41505 62.330 -440.03
<none>                                             61.915 -440.00
- age_partner                          1   0.69348 62.609 -438.72
+ sex                                  1   0.05321 61.862 -438.25
- n_household                          1   0.86276 62.778 -437.93
- relative_language_ability_at_age_10  1   1.92567 63.841 -432.99
- books_age_10                         1   2.90236 64.818 -428.53

Step:  AIC=-440.61
log(household_income) ~ n_household + retired + age + age_partner +
    years_of_education + books_age_10 + relative_language_ability_at_age_10

                                      Df Sum of Sq    RSS     AIC
- age                                  1   0.35556 62.563 -440.94
- retired                              1   0.39006 62.597 -440.78
<none>                                             62.207 -440.61
- years_of_education                   1   0.43508 62.642 -440.56
+ relative_math_ability_at_age_10      1   0.29180 61.915 -440.00
- age_partner                          1   0.73222 62.939 -439.17
+ sex                                  1   0.10089 62.106 -439.09
- n_household                          1   0.90826 63.116 -438.35
- relative_language_ability_at_age_10  1   2.56083 64.768 -430.75
- books_age_10                         1   3.00424 65.211 -428.75

Step:  AIC=-440.94
log(household_income) ~ n_household + retired + age_partner +
    years_of_education + books_age_10 + relative_language_ability_at_age_10

                                      Df Sum of Sq    RSS     AIC
- retired                              1   0.18023 62.743 -442.09
- years_of_education                   1   0.40255 62.965 -441.05
- age_partner                          1   0.40662 62.969 -441.03
<none>                                             62.563 -440.94
```

```
+ age                                    1   0.35556 62.207 -440.61
+ relative_math_ability_at_age_10        1   0.33088 62.232 -440.50
+ sex                                    1   0.29626 62.267 -440.33
- n_household                            1   0.95518 63.518 -438.48
- relative_language_ability_at_age_10    1   2.46087 65.024 -431.60
- books_age_10                           1   2.86047 65.423 -429.79

Step:  AIC=-442.09
log(household_income) ~ n_household + age_partner + years_of_education +
    books_age_10 + relative_language_ability_at_age_10

                                       Df Sum of Sq    RSS     AIC
- years_of_education                    1   0.40859 63.152 -442.18
<none>                                             62.743 -442.09
+ relative_math_ability_at_age_10      1   0.29800 62.445 -441.49
+ retired                              1   0.18023 62.563 -440.94
+ age                                  1   0.14574 62.597 -440.78
+ sex                                  1   0.12708 62.616 -440.69
- n_household                          1   0.94706 63.690 -439.69
- age_partner                          1   1.41929 64.162 -437.52
- relative_language_ability_at_age_10  1   2.57721 65.320 -432.26
- books_age_10                         1   2.82709 65.570 -431.14

Step:  AIC=-442.18
log(household_income) ~ n_household + age_partner + books_age_10 +
    relative_language_ability_at_age_10

                                       Df Sum of Sq    RSS     AIC
<none>                                             63.152 -442.18
+ years_of_education                    1    0.4086 62.743 -442.09
+ relative_math_ability_at_age_10      1    0.3839 62.768 -441.98
+ retired                              1    0.1863 62.965 -441.05
+ age                                  1    0.1248 63.027 -440.77
+ sex                                  1    0.0874 63.064 -440.59
- n_household                          1    1.1283 64.280 -438.98
- age_partner                          1    1.6371 64.789 -436.66
- relative_language_ability_at_age_10  1    3.3782 66.530 -428.86
- books_age_10                         1    3.9383 67.090 -426.40
```

```
summary(step_model)
```

```
Call:
lm(formula = log(household_income) ~ n_household + age_partner +
    books_age_10 + relative_language_ability_at_age_10, data = data_clean)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.3245 -0.2690 -0.0267  0.3161  1.1806

Coefficients:
                                                Estimate Std. Error t value
(Intercept)                                    10.749471   0.434233  24.755
n_household                                     0.259106   0.114025   2.272
age_partner                                    -0.012064   0.004407  -2.737
books_age_100-25                               -0.248684   0.058578  -4.245
relative_language_ability_at_age_10same/worse -0.220954   0.056195  -3.932
                                               Pr(>|t|)
(Intercept)                                     < 2e-16 ***
n_household                                    0.023799 *
age_partner                                    0.006582 **
books_age_100-25                               2.94e-05 ***
relative_language_ability_at_age_10same/worse 0.000106 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4675 on 289 degrees of freedom
  (93 observations deleted due to missingness)
Multiple R-squared:  0.1759,    Adjusted R-squared:  0.1645
F-statistic: 15.42 on 4 and 289 DF,  p-value: 1.921e-11
```

4.2 Model improvement

When the model remove the age of the partner

```
data_clean<-data%>%
  select(-c(age_partner))
model1<-lm(log(household_income)~., data=data_clean)
step_model1<-step(model1, direction="both", trace=1)
```

```
Start:  AIC=-513.83
log(household_income) ~ sex + lives_with_partner + n_household +
    retired + age + years_of_education + books_age_10 + relative_math_ability_at_age_10 +
    relative_language_ability_at_age_10

                                      Df Sum of Sq     RSS     AIC
- age                                  1    0.0049  97.423 -515.81
- years_of_education                   1    0.4671  97.885 -513.98
<none>                                             97.418 -513.83
```

```
- sex                                   1    0.5442  97.963 -513.68
- relative_math_ability_at_age_10       1    0.5443  97.963 -513.68
- retired                               1    0.9950  98.413 -511.90
- lives_with_partner                    1    2.1551  99.574 -507.36
- books_age_10                          1    2.2836  99.702 -506.86
- relative_language_ability_at_age_10   1    2.8898 100.308 -504.52
- n_household                           1    4.1017 101.520 -499.87

Step:  AIC=-515.81
log(household_income) ~ sex + lives_with_partner + n_household +
    retired + years_of_education + books_age_10 + relative_math_ability_at_age_10 +
    relative_language_ability_at_age_10

                                      Df Sum of Sq     RSS     AIC
- years_of_education                   1    0.4624  97.886 -515.98
<none>                                              97.423 -515.81
- relative_math_ability_at_age_10      1    0.5476  97.971 -515.64
- sex                                  1    0.5581  97.981 -515.60
+ age                                  1    0.0049  97.418 -513.83
- retired                              1    1.6794  99.103 -511.20
- lives_with_partner                   1    2.1516  99.575 -509.36
- books_age_10                         1    2.2920  99.715 -508.81
- relative_language_ability_at_age_10  1    2.9147 100.338 -506.40
- n_household                          1    4.1516 101.575 -501.66

Step:  AIC=-515.98
log(household_income) ~ sex + lives_with_partner + n_household +
    retired + books_age_10 + relative_math_ability_at_age_10 +
    relative_language_ability_at_age_10

                                      Df Sum of Sq     RSS     AIC
<none>                                              97.886 -515.98
- sex                                  1    0.5128  98.398 -515.96
+ years_of_education                   1    0.4624  97.423 -515.81
- relative_math_ability_at_age_10      1    0.7199  98.606 -515.14
+ age                                  1    0.0001  97.885 -513.98
- retired                              1    1.9422  99.828 -510.38
- lives_with_partner                   1    2.2664 100.152 -509.12
- relative_language_ability_at_age_10  1    3.4240 101.310 -504.67
- books_age_10                         1    3.5726 101.458 -504.11
- n_household                          1    4.3131 102.199 -501.29
```

```r
summary(step_model1)
```

```
Call:
lm(formula = log(household_income) ~ sex + lives_with_partner +
    n_household + retired + books_age_10 + relative_math_ability_at_age_10 +
    relative_language_ability_at_age_10, data = data_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-1.44687 -0.31555  0.00147  0.30458  1.64562

Coefficients:
                                            Estimate Std. Error t value
(Intercept)                                 9.57481    0.12809  74.753
sexmale                                     0.07689    0.05457   1.409
lives_with_partneryes                       0.28970    0.09780   2.962
n_household                                 0.33886    0.08292   4.087
retiredyes                                 -0.16343    0.05960  -2.742
books_age_100-25                           -0.20596    0.05538  -3.719
relative_math_ability_at_age_10same/worse  -0.09422    0.05643  -1.670
relative_language_ability_at_age_10same/worse -0.20514    0.05634  -3.641
                                           Pr(>|t|)
(Intercept)                                 < 2e-16 ***
sexmale                                    0.159642
lives_with_partneryes                      0.003246 **
n_household                                5.35e-05 ***
retiredyes                                 0.006391 **
books_age_100-25                           0.000230 ***
relative_math_ability_at_age_10same/worse  0.095826 .
relative_language_ability_at_age_10same/worse 0.000309 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5082 on 379 degrees of freedom
Multiple R-squared:  0.3501,    Adjusted R-squared:  0.3381
F-statistic: 29.17 on 7 and 379 DF,  p-value: < 2.2e-16
```

```r
model2<-step_model1
vif(model2)
```

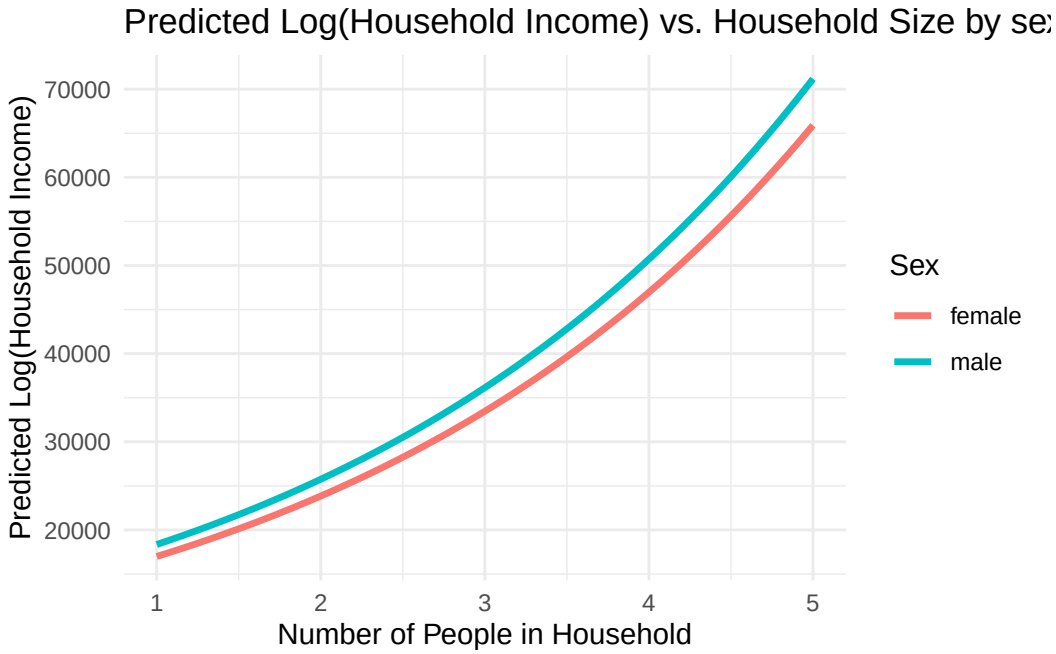
```
                                 sex                  lives_with_partner
                            1.100280                            2.616217
                          n_household                             retired
                            2.697921                            1.058390
                         books_age_10     relative_math_ability_at_age_10
                            1.103304                            1.187072
relative_language_ability_at_age_10
```

```
1.188123
```

Our final model indicates that the log of household income generally increases with the number of household members, being male, and living with a partner. In contrast, household income tends to decrease for retired households, those who had between 0 and 25 books at age 10, and individuals who had lower math or language abilities at age 10.

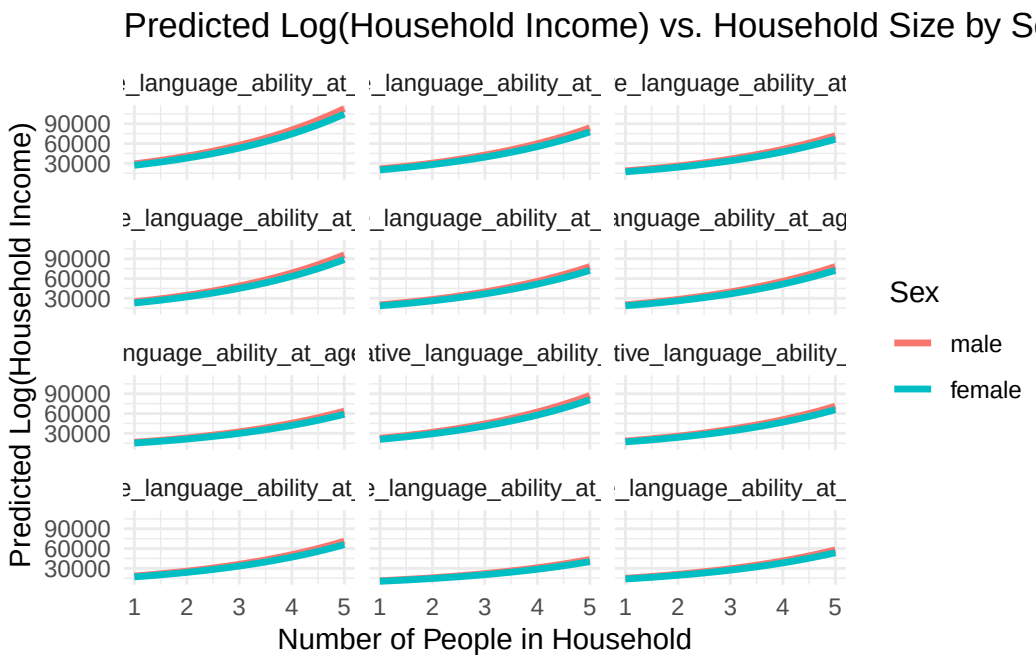
4.3 Prediction Plot

```
data1 <- data %>% select(-age_partner)

new_data <- expand.grid(
  sex=factor(c("male", "female")),
  lives_with_partner=factor("yes", level=c("no", "yes")),
  n_household = seq(min(data1$n_household), max(data1$n_household), length.out = 100),
  books_age_10 = factor(">25", c("0-25", ">25")),
  retired = factor("yes", levels = c("no", "yes")),
  relative_math_ability_at_age_10 = factor("same/worse", levels = c("same/worse", "better"
  relative_language_ability_at_age_10 = factor("same/worse", levels = c("same/worse", "bet
)

# Predict using your final model (assumed as model2)
new_data$predicted_income <- predict(model2, newdata = new_data)

# Plot
ggplot(new_data, aes(x = n_household, y = exp(predicted_income), color = sex)) +
  geom_line(linewidth = 1.2) +
  labs(
    title = "Predicted Log(Household Income) vs. Household Size by sex",
    x = "Number of People in Household",
    y = "Predicted Log(Household Income)",
    color = "Sex"
  ) +
  theme_minimal()
```




```
library(purrr)
data1 <- data %>% select(-age_partner)

topprofiles<-data%>%
  group_by(retired, relative_math_ability_at_age_10, relative_language_ability_at_age_10,
  summarise(count=n(), .groups="drop")%>%
  arrange(desc(count))%>%
  slice_head(n=12)

n_household = seq(min(data1$n_household), max(data1$n_household), length.out = 100)

# Expand to both Male and Female
new_data <- topprofiles %>%
  mutate(profile_id = row_number()) %>%
  group_split(profile_id) %>%
  map_dfr(function(profile) {
    profile_info <- profile %>% select(-count, -profile_id)

    # Add both sexes
    sexes <- c("male", "female")

    map_dfr(sexes, function(sex_val) {
      profile_replicated <- profile_info[rep(1, length(n_household)), ]
      profile_replicated$sex <- factor(sex_val, levels = c("male", "female"))
      profile_replicated$n_household <- n_household
      profile_replicated
```

```
    })
  })

# Predict using your final model (assumed as model2)
new_data$predicted_income <- predict(model2, newdata = new_data, type="response")

# 4. Create facet label
new_data$profile_label <- with(new_data, paste0(
  "Retired: ", retired, ", Relative_math_ability_age_10: ", relative_math_ability_at_age_1
  ", Relative_language_ability_at_age_10:", relative_language_ability_at_age_10,
  ", books_age_10:", books_age_10, ", lives_with_partner:", lives_with_partner
))

# Plot
ggplot(new_data, aes(x = n_household, y = exp(predicted_income), color = sex)) +
  geom_line(linewidth = 1.2) +
  labs(
    title = "Predicted Log(Household Income) vs. Household Size by Sex",
    x = "Number of People in Household",
    y = "Predicted Log(Household Income)",
    color = "Sex"
  ) +
  facet_wrap(~profile_label, ncol=3)+
  theme_minimal()
```

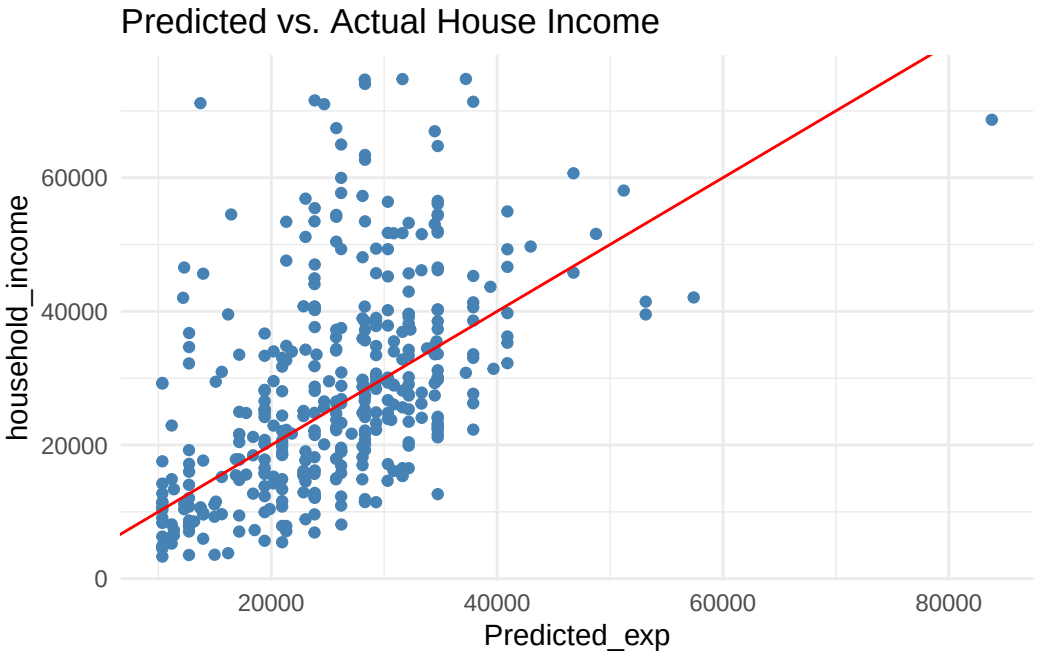


We use the model to predict income based on the 12 most common profiles in the data frame.

We visualize how household income increases with the number of people in the household across different genders.

4.4 Predicted vs. Actual House Income

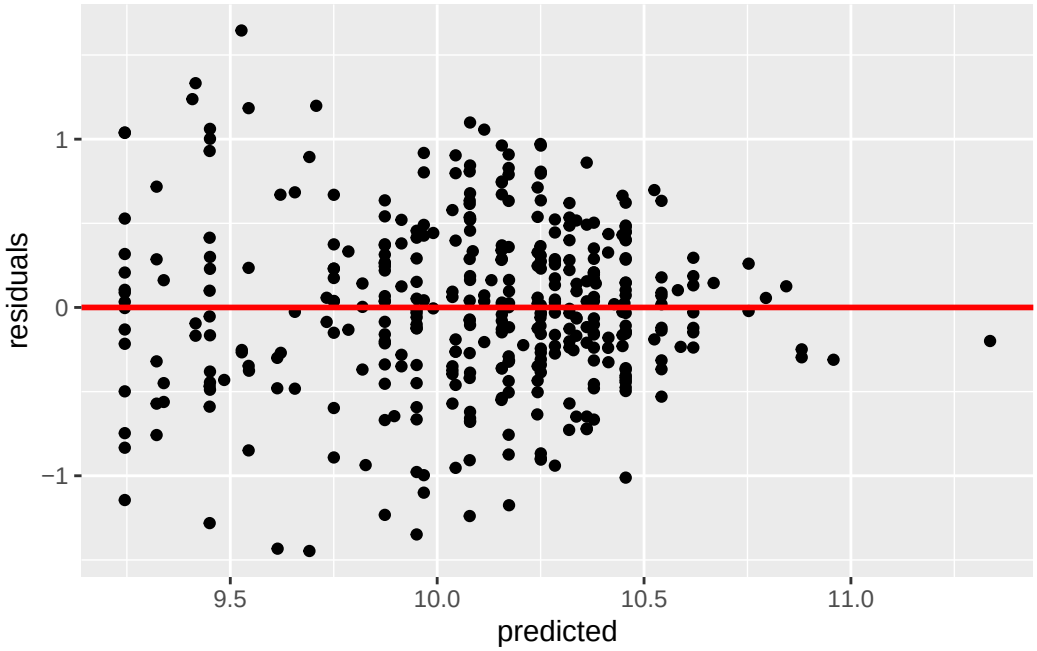
```
data$Predicted=predict(model2, data)
data$Predicted_exp=exp(data$Predicted)

ggplot(data, aes(x=Predicted_exp, y=household_income))+
  geom_point(color="steelblue")+
  geom_abline(slope=1, intercept=0, color="red")+
  labs(title="Predicted vs. Actual House Income")+
  theme_minimal()
```




```
data$predicted=predict(model2)
data$residuals=residuals(model2)

ggplot(data, aes(x=predicted, y=residuals))+
  geom_point()+
  geom_hline(yintercept=0, color="red", size=1, linestyle="solid")
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

```
Warning in geom_hline(yintercept = 0, color = "red", size = 1, linestype =
"solid"): Ignoring unknown parameters: `linestype`
```



4.5 Model Validation

4.5.1 Root Mean Squre Error

```
Rmse<- sqrt(mean((data$Predicted_exp-data$household_income)^2))
Rmse
```

```
[1] 13850.62
```

4.5.2 Mean Absolute Error

```
Mae=mean(abs(data$Predicted_exp-data$household_income))
Mae
```

```
[1] 10040.87
```

4.5.3 Mean Absolute Percentage Error

```
Mape=mean(abs((data$Predicted_exp-data$household_income)/data$household_income))*100
Mape
```

```
[1] 42.68762
```

If we use the original data, the MAPE is 179.1%, that means your model's predictions are off by nearly 1.8 times the actual income on average, which suggests the model is not performing well in terms of accuracy.

If we use the household income is [3000, 75000], our model MAPE is 51%.

A MAPE around 42.7% for log-transformed income is not bad, especially with limited predictors. The model is useful for general trends and comparisons, but not reliable for precise individual prediction.