ANALYSIS OF DISCONTINUATION AND NON-ENROLLMENT OF STUDENTS OF VARIOUS SOCIO-ECONOMIC POPULATIONS IN WEST BENGAL AND COMPARISON WITH NATIONAL AVERAGE: STUDY BASED ON NSSO-71ST ROUND EDUCATION SURVEY DATA

BY

TANISHA DAS

ROLL: 96/STA NUMBER: 220030

REGISTRATION NUMBER: 015701 of 2019-2020

SUPERVISED BY

KAJAL DIHIDAR

ASSOCIATE PROFESSOR

SAMPLING AND OFFICIAL STATISTICS UNIT(SOSU)

INDIAN STATISTICAL INSTITUTE (ISI)

KOLKATA

DEPARTMENT OF STATISTICS

UNIVERSITY OF KALYANI

KALYANI, NADIA

2024

ANALYSIS OF DISCONTINUATION AND NON-ENROLLMENT OF STUDENTS OF VARIOUS SOCIO-ECONOMIC POPULATIONS IN WEST BENGAL AND COMPARISON WITH NATIONAL AVERAGE: STUDY BASED ON NSSO-71ST ROUND EDUCATION SURVEY DATA

Abstract:

A sample survey is a research method used to collect data and opinions from a subset of individuals (samples) representing a larger group (population). It includes:

- Questionnaire: A set of open-ended or multiple-choice questions, to collect specific information.

- Sample Size: This refers to how many people we choose to participate in the survey. Getting the sample size right is crucial for ensuring the results are accurate and useful for understanding the larger group.

- Sampling Method: The strategy used to select participants. Common methods include:
    - Random Sampling: Every individual in the population has the same (equal) chance of being selected.
    - Stratified Sampling: The population is split into subgroups, and from each subgroup, samples are drawn.
    - Snowball Sampling: Current participants help find new participants, which helps reach hard-to-access groups.

- Data Collection: The process of gathering responses from the participants, which can be done through various means such as online surveys, face-to-face interviews, or phone calls.

- Data Analysis: After collecting the data, the step involves looking at the information to find patterns and draw conclusions. It's about turning raw data into meaningful insights that can inform decisions or answer research questions.

Objective of the Project:

Education is the cornerstone of a thriving society. It drives economic and social progress, and a strong nation is built by ensuring all its citizens have access to quality education. Governments around the world invest significant resources into developing and maintaining educational infrastructure. However, individuals also bear costs, including tuition fees, exam fees, and books and supplies. While we can track government spending through budget documents, understanding the full scope of education expenses, including those borne by individuals and non-governmental organizations, requires more specialized data.

The National Sample Survey Office (NSSO) conducted a comprehensive household survey on education from January to June 2014 to address this need. The survey aimed to gather detailed information on the educational activities of people aged 5-29 years across the country. It looked into how well educational facilities and government incentives are being used, examined private spending on education, and assessed issues like student dropout rates and their causes. This survey was crucial for understanding the current state of education and identifying areas for improvement.

Use of Data:

- Block-3 – Level-02: Household Characteristics: Certain household characteristics, such as, household size, principal industry, principal occupation, household type, religion, social-group, the distance to the nearest primary, upper primary and secondary schools, whether the household has computer and access to internet, household's usual consumer expenditure in a month, etc. is recorded in this block. It contains 37 variables that describe the household's overall characteristics.

- Block- 4 – Level-03 Demographic and other particulars of Household members: This block contains demographic particulars (viz., relation to head, sex, age, and marital status), educational level, status of current educational enrolment and attendance and response to queries on IT literacy etc. It contains 38 variables.

- Block-7 – Level-06 Perticulars of persons currently not attending any educational istitude: This dataset contains information on persons who are not currently attending any educational institution including those, who are currently enrolled but currently not attending. It contains 36 variables.

This data was produced by National Sample Survey Office – Ministry of Statistics and Programme Implementation (MOSPI), Government of India and is sponsored by Ministry of Statistics and Programme Implementation (MOSPI), Government of India.

Introduction to NSSO:

The National Sample Survey Office (NSSO), led by a Director General, plays a crucial role in gathering large-scale data across various socio-economic areas in India. It conducts nationwide household surveys, runs the Annual Survey of Industries (ASI), tracks rural and urban prices, and helps enhance crop statistics through area enumeration and crop estimation surveys. The NSSO also maintains an urban area unit framework for sample surveys.

The NSSO is divided into four main divisions:

- Survey Design and Research Division (SDRD): Based in Kolkata, this division handles the technical side of survey planning. They create concepts and definitions, design sampling methods and inquiry schedules, and manage the analysis and presentation of survey results.

- Field Operations Division (FOD): Headquartered in Delhi/Faridabad, this division manages data collection through a vast network of Zonal, Regional, and Sub-Regional Offices spread across the country.

- Data Processing Division (DPD): Also located in Kolkata, this division oversees the selection of survey samples, software development, and the processing and validation of survey data. They handle data for various surveys, including the Periodic Labour Force Survey (PLFS) and the Annual Survey of Industries through dedicated portals

- Survey Coordination Division (SCD): Based in New Delhi, this division coordinates the activities of the other divisions. They publish the bi-annual journal "Sarvekshana" and organize national seminars to discuss the findings of NSSO's surveys.

Each division has a specific role in ensuring that NSSO's surveys are well-planned, executed, and analyzed, contributing to valuable socio-economic data for the country.

NSSO's Educational Survey:

The National Sample Survey Organization (NSSO) first conducted an all-India survey on social consumption during its 35th round from July 1980 to June 1981. The goal of this survey was to gather data on how public spending on various services, like education, mass immunization, family welfare programs, and healthcare, benefited the people. Following this initial survey, the NSSO continued to explore social consumption in subsequent rounds: The 42nd round (July 1986 to June 1987), The 52nd round (July 1995 to June 1996), and the 64th

round (July 2007 to June 2008). These surveys followed a similar approach to the 35th round but included some updates and changes in the topics covered.

In these surveys, both the qualitative and quantitative aspects of educational services were examined. The qualitative aspects looked at issues like literacy rates, levels of education, current school attendance, and reasons for dropping out of school. The quantitative aspects focused on the costs associated with education, such as tuition fees and transportation. Additionally, in the 47th round (July to December 1991), the NSSO collected detailed data on the qualitative aspects of educational services. In all NSSO household surveys, information on literacy and educational attainment for each household member was also gathered.

NSSO's 71st Round Education Survey:

The 71st round of the National Sample Survey (NSS) on education took place from January 1 to June 30, 2014. This survey aimed to collect detailed information about the participation of individuals aged 5 to 29 years in various educational activities. With the increasing role of Information Technology in everyday life, this round of the survey incorporated new approaches and methods, reflecting changes in the educational landscape since the previous survey round (the 64th, conducted from July 2007 to June 2008).

Here's a look at the main differences between the 71st and 64th rounds of the survey:

- Focus on One Basic Course: In the 71st round, the survey collected data on expenditure for only one basic educational course per person. In contrast, the 64th round collected data for two courses, which added complexity to the process.

- Simplified Course Codes and Attendance Levels: The 71st round made the structure of course codes and attendance level codes simpler and more uniform, making it easier to classify educational stages.

- Standardized Education Levels: For this round, the survey uniformly classified classes I-V as primary education and classes VI-VII as upper primary education across all states and Union Territories, whereas previous surveys had varied classifications.

- Streamlined Household Expenditure Questions: Earlier surveys asked five different questions about monthly household expenditure. The 71st round reduced this to just one question to streamline data collection.

- New Enquiry Items: New questions were added to address current educational issues, such as the main language spoken at home and whether students were receiving private tuition.

- IT Access and Usage: Following the Department of Electronics and IT's requirements, the 71st round included questions on household access to computers and the internet, reflecting the growing importance of these technologies in education.

- Consolidated Expenditure Items: Compared to previous rounds, the 71st round consolidated some expenditure items into broader categories, simplifying the data collection process.

- Vocational Education Integration: The survey integrated vocational education into the broader category of professional/technical education, following guidelines from the Ministry of Human Resource Development, eliminating the need for a separate code for vocational courses.

- Simplified Educational Codes: The 71st round also simplified the structures of course codes and attendance levels related to basic education, making it more straightforward to capture educational details.

- Rural and Urban Data Clarification: The figures in this report distinguish between people from rural and urban areas based on their residence, not the locations where they studied.

- Telangana State Coding: Since Telangana was not a separate state at the start of the survey period, there was no specific code for it in the survey's Schedule of Enquiry.

These changes aimed to make the survey process more efficient and the results more reflective of contemporary educational issues.

Objective of the Project, the specific objective:

The project aims to understand whether a person's discontinuation/dropping out from education and the unenrollment reasons depend on various socio-economic factors, like the UMPCE with other possible covariates sectors (Urban, Rural), gender (Male, Female), household type. We carry on this process for both West Bengal and All-India separately and then further move on to comparing them.

For discontinuation/dropping out:

x_i =1 if enrolled

= 0 otherwise

y_i = 1 if enrolled but discontinued with reason i

= 0 otherwise

Let us define the population parameter as R = Population proportion of persons who have enrolled but discontinued for reason i . We estimate R by Rhat as Rhat(enrolled but discontinued for reason i) = Yhat / Xhat, where Xhat = Estimated number of enrolled persons = sum(x * multiplier) and Yhat = Estimated number of enrolled persons who discontinued/dropped out for reason i = sum(y * multiplier).

For never-enrollment:

x_i = 0 if never-enrolled

= 1 otherwise

y_i = 1 if never enrolled for reason i

= 0 otherwise

Let us define the population parameter as R = Population proportion of persons who have never enrolled for reason i . We estimate R by Rhat as Rhat(Never enrolled for reason i) = Yhat / Xhat, where Xhat = Estimated number of never-enrolled persons = sum(x * multiplier) and Yhat = Estimated number of persons who never-enrolled reason i = sum(y * multiplier).

We find out the estimates of the ratio $\hat{R}$, their MSE (mean square error), and RSE (relative square error) in percentage for various socio-economic population sub-groups. Then carry out the logistic regression of several well-defined binary response variables with the possible explanatory variables, say, considering the quintile classes of UMPCE and other socio-economic covariates (if any).

Sampling Design:

- Outline of sample design: A stratified multi-stage design has been adopted for the 71st round survey. The first stage units (FSU) are the census villages (Panchayat wards in the case of Kerala) in the rural sector and Urban Frame Survey (UFS) blocks in the urban sector. The ultimate stage units (USU) are households in both sectors. In the case of large FSUs, one intermediate stage of sampling is the selection of two hamlet-groups (hgs)/ sub-blocks (sbs) from each rural/ urban FSU.

- Sampling Frame for First Stage Units: For the rural sector, the list of 2011 census villages (henceforth the term 'village' would mean Panchayat wards for Kerala) constitutes the sampling frame. In the case of Kerala, due to the non-availability of Panchayat wards based on census 2011, the available list of Panchayat wards based on census 2001 is used as the rural frame. For the urban sector, the latest updated list of UFS blocks (phase 2007-12) is considered as the sampling frame.

- Stratification: Stratum has been formed at the district level. Within each district of a State/UT, generally speaking, two basic strata have been formed: (i) rural stratum comprising all rural areas of the district and (ii) urban stratum comprising all the urban areas of the district. However, within the urban areas of a district, if there are one or more towns with a population 1 lakh or more as per Census 2011, each of them forms a separate basic stratum, and the remaining urban areas of the district have been considered as another basic stratum.

  Special stratum in the rural sector: There are some villages in Nagaland and Andaman & Nicobar Islands which remains difficult to access. As in earlier rounds, a special stratum has been formed at the State/UT level comprising these villages in the two State/UTs.

- Sub-stratification:

  - Rural sector: If 'r' is the sample size allocated for a rural stratum, the number of sub-strata formed was 'r/2'. The villages within a district as per frame have been first arranged in ascending order of population. Then sub-strata 1 to 'r/2' were restricted so that each sub-stratum comprised a group of villages of the arranged frame and had more or less equal population.

  - Urban sector: If 'u' is the sample size allocated for an urban stratum, the number of sub-strata formed was 'u/2'. For all strata, if u/2 >1, implying the formation of 2 or more sub-strata, all the UFS blocks within the stratum have been first arranged in ascending order of a total number of households in the UFS Blocks as per UFS phase 2007-12. Then sub-strata 1 to 'u/2' were restricted so that each sub-stratum had more or less an equal number of households.

- Total sample size (FSUs): 8300 FSUs have been allocated for the central sample at the all-India level. For the state sample, there are 9274 FSUs allocated for all of India. State-wise allocation of sample FSUs is given in Table 1.

- Allocation of total sample to States and UTs: The total number of sample FSUs have been allocated to the States and UTs in proportion to population as per Census 2011 subject to a minimum sample allocation to each State/ UT. While

doing so, the resource availability in terms of the number of field investigators has been kept in view.

- Allocation of State/ UT level sample to rural and urban sectors: State/UT level sample size has been allocated between two sectors in proportion to population as per Census 2011 with double weightage to urban sector subject to the restriction that urban sample size for bigger states like Maharashtra, Tamil Nadu, etc. do not exceed the rural sample size. A minimum of 16 FSUs (a minimum of 8 each for rural and urban sectors separately) is allocated to each State/ UT.

- Allocation to strata: Within each sector of a State/ UT, the respective sample size has been allocated to the different strata in proportion to the population as per Census 2011. Stratum level allocation has been adjusted to multiples of 2 with a minimum sample size of 2. For special strata in the rural areas of Nagaland and A & N Islands, 4 FSUs have been allocated to each.

- Allocation to sub-strata: Each sub-stratum has been 2 in rural and urban sectors.

- Selection of FSUs: For the rural sector, from each stratum/sub-stratum, the required number of sample villages have been selected by Probability Proportional to Size with Replacement (PPSWR), size being the population of the village as per Census 2011. For the urban sector, from each stratum/sub-stratum, FSUs have been selected by Probability Proportional to Size with Replacement (PPSWR), size being the number of households of the UFS Blocks. Both rural and urban samples have been drawn in the form of two independent sub-samples and an equal number of samples has been allocated among the two sub-rounds.

- Selection of hamlet-groups/ sub-blocks- important steps

  Criterion for hamlet-group/ sub-block formation: After identification of the boundaries of the FSU, it is to be determined whether listing will be done in the whole sample FSU or not. In case the approximate present population of the selected FSU is found to be 1200 or more, it will be divided into a suitable number (say, D) of 'hamlet-groups' in the rural sector and 'sub-blocks' in the urban sector by more or less equalizing the population as stated below.

Approximate present population of the sample FSU	Number of hg's/sb's to be formed
Less than 1200 (no hg/sub-blocks)	1
1200 to 1799	3
1800 to 2399	4
2400 to 2999	5
3000 to 3599	6
.....and so on	-

For rural areas of Himachal Pradesh, Sikkim, Uttarakhand (except four districts Dehradun, Nainital, Hardwar and Udham Singh Nagar), Poonch, Rajouri, Udhampur, Reasi, Doda, Kistwar, Ramban, Leh (Ladakh), Kargil districts of Jammu and Kashmir and Idukki district of Kerala, the number of hamlet-groups will be formed as follows:

Approximate present population of the sample village	Number of hg's to be formed
Less than 600 (no hamlet-groups)	1
600 to 899	3
900 to 1199	4
1200 to 1499	5
1500 to 1799	6
.....and so on	-

Formation and selection of hamlet-groups/ sub-blocks: In case hamlet-groups/ sub-blocks are to be formed in the sample FSU, the same should be done by more or less equalizing the population. Note that while doing so, it is to be ensured that the hamlet groups/ sub-blocks formed are clearly identifiable in terms of physical landmarks.

Two hamlet-groups (hg)/ sub-blocks (sb) will be selected from a large FSU wherever hamlet-groups/ sub-blocks have been formed in the following manner – one hg/ sb with maximum percentage share of population will always be selected and termed as hg/ sb 1; one more hg/ sb will be selected from the remaining hg's/ sb's by simple random sampling (SRS) and termed as hg/ sb 2. Listing and selection of the households will be done independently in the two selected hamlet groups/ sub-blocks. The FSUs without hg/ sb formation will be treated as sample hg/ sb number 1.

SSS	Composition of SSS within a sample FSU	Number of households to be surveyed	
		FSU without hg/sb formation	FSU with hg/sb formation (for each hg/sb)
SSS1	Households with at least one student receiving technical/ professional education	2	1
SSS2	From the remaining, households having at least one student receiving general education	4	2
SSS3	Other households	2	1

- selection of households: From each SSS, for both the schedules, the sample households are selected by SRSWOR.

Estimation Procedure:

Notations:

s = subscript for s-th stratum

t = subscript for t-th sub-stratum

m = subscript for sub-sample (m =1, 2)

i = subscript for i-th FSU [village (panchayat ward)/ block]

d = subscript for a hamlet-group/ sub-block (d = 1, 2)

j = subscript for j-th second stage stratum in an FSU/ hg/sb [j = 1, 2 or 3]

k = subscript for k-th sample household under a particular second stage stratum within an FSU/hg/sb

D = total number of hg's/ sb's formed in the sample FSU

D* = 0 if D = 1

= (D – 1) for FSUs with D > 1

Z = total size of a rural/urban sub-stratum (= sum of sizes for all the FSUs of a sub-stratum)

z = size of sample village/UFS block used for selection.

n = number of sample FSUs surveyed including 'uninhabited' and 'zero cases' but excluding casualty for a particular sub-sample and sub-stratum.

H = total number of households listed in a second-stage stratum of an FSU / hamlet group or sub-block of sample FSU.

h = number of households surveyed in a second-stage stratum of an FSU / hamlet-group or sub-block of sample FSU.

x, y = observed value of characteristics x, y under estimation

$\hat{X}$, $\hat{Y}$ = estimate of population total X, Y for the characteristics x, y

Under the above symbols,

$y_{stmidjk}$ = observed value of the characteristic y for the k-th household in the j-th second stage stratum of the d-th hg/ sb (d = 1, 2) of the i-th FSU belonging to the m-th sub-sample

for the t-th sub-stratum of s-th stratum.

However, for ease of understanding, a few symbols have been suppressed in the following paragraphs where they are obvious.

Formulae for Estimation of Aggregates for a particular sub-sample and stratum × sub-stratum:

- For Schedule 0.0 (Urban/Rural):

(i) For estimating the number of households in a stratum × sub-stratum possessing a characteristic:

$$\hat{Y} = \frac{Z}{n} \sum_{i=1}^{n} \frac{1}{z_i} [y_{i1} + D_i^* \times y_{i2}]$$

where $yi1$, $yi2$ are the total number of households possessing the characteristic y in hg's 1 & 2 of the i-th FSU respectively.

(ii) For estimating the number of villages in a stratum × sub-stratum possessing a characteristic:

$$\hat{Y} = \frac{Z}{n} \sum_{i=1}^{n} \frac{1}{z_i} y_i$$

where yi is taken as 1 for sample villages possessing the characteristic and 0 otherwise.

- For Schedule 25.2 (Urban/Rural):

(i) For j-th second-stage stratum of a stratum × sub-stratum:

$$\hat{Y}_J = \frac{Z}{n_j} \sum_{i=1}^{n_j} \frac{1}{z_i} [\frac{H_{i1j}}{h_{i1j}} \sum_{k=1}^{h_{i1j}} y_{i1\,jk} + D_i^* \times \frac{H_{i12j}}{h_{i2j}} \sum_{k=1}^{h_{i2j}} y_{i2\,jk}]$$

(ii) **For all second-stage strata combined:**

$$\hat{Y} = \sum_j \hat{Y}_j$$

Overall Estimate for Aggregates for a sub-stratum:

Overall estimate for aggregates for a sub-stratum ($\hat{Y}_{st}$) based on two subsamples in a sub-stratum is obtained as

$$\hat{Y}_{st} = \frac{1}{2} \sum_{m=1}^{2} \hat{Y}_{stm}$$

Overall Estimate for Aggregates for a stratum:

Overall estimate for a stratum ($\hat{Y}_s$) will be obtained as

$$\hat{Y}_s = \sum_t \hat{Y}_{st}$$

Overall Estimate of Aggregates at State/UT/all-India level:

The overall estimate $\hat{Y}$ at the State/ UT/ all-India level is obtained by summing the stratum estimates $\hat{Y}_s$ over all strata belonging to the State/ UT/ all-India.

Estimates of Ratios:

Let $\hat{Y}$ and $\hat{X}$ be the overall estimates of the aggregates Y and X for two characteristics y and x respectively at the State/ UT/ all-India level.

Then the combined ratio estimate ($\hat{R}$) of the ratio ($R=\frac{Y}{X}$) will be obtained as $\boldsymbol{R} = \frac{\hat{\boldsymbol{Y}}}{\hat{\boldsymbol{X}}}$

Estimates of Error:

The estimated variances of the above estimates will be as follows

- For aggregate $\hat{Y}$:

$$\widehat{Var}(\hat{Y}) = \sum_s \widehat{Var}(\hat{Y}_s) = \sum_s \sum_t \widehat{Var}(\hat{Y}_{st})$$

Where $\widehat{Var}(\hat{Y}_{st})$ is given by $\widehat{Var}(\hat{Y}_{st}) = \frac{1}{4}(\hat{Y}_{st\,1} - \hat{Y}_{st\,2})^2$ where $\hat{Y}_{st\,1} and\ \hat{Y}_{st\,2}$ are the estimates for sub-sample 1 and sub-sample 2 respectively for stratum 's' and 't'.

- For ratio $\hat{R}$:

$$\widehat{MSE}(\hat{R}) = \frac{1}{4\hat{X}^2} \sum_s \sum_t (\hat{Y}_{st\,1} - \hat{Y}_{st\,2})^2 + \hat{R}^2 (\hat{X}_{st\,1} - \hat{X}_{st\,2})^2 - 2\hat{R} (\hat{Y}_{st\,1} - \hat{Y}_{st\,2}) (\hat{X}_{st\,1} - \hat{X}_{st\,2})$$

- Estimates of Relative Standard Error (RSE):

$$\widehat{RSE}(\hat{Y}) = \frac{\sqrt{\widehat{Var}(\hat{Y})}}{\hat{Y}} \times 100$$

$$\widehat{RSE}(\hat{R}) = \frac{\sqrt{\widehat{MSE}(\hat{R})}}{\hat{R}} \times 100$$

Multipliers:

The formulae for multipliers at stratum/sub-stratum/second-stage stratum level for a sub-sample and schedule type are given below:

Schedule type	Sector	Formula for multipliers	
		hg/ sb 1	hg/ sb 2
0.0	Urban/Rural	$\frac{Z_{st}}{n_{stm}} \times \frac{1}{z_{stmi}}$	$\frac{Z_{st}}{n_{stm}} \times \frac{1}{z_{stmi}} \times D^*_{stmi}$
25.2	Urban/Rural	$\frac{Z_{st}}{n_{stmj}} \times \frac{1}{z_{stmi}} \times \frac{H_{stmi1j}}{h_{stmi1j}}$	$\frac{Z_{st}}{n_{stmj}} \times \frac{1}{z_{stmi}} \times D^*_{stmi} \times \frac{H_{stmi2j}}{h_{stmi2j}}$
	(j=1,2,3)		

Note:

(i) For estimating any characteristic for any domain not specifically considered in the sample design, indicator variables may be used.

(ii) Multipliers have to be computed based on information available in the listing schedule irrespective of any misclassification observed between the listing schedule and detailed inquiry schedule.

(iii) For estimating the number of villages possessing a characteristic, $D^{*}_{stmi} = 0$ in the

R codes, their outputs and their output tables:

At first, we import the Block 7 data into R studio, which we downloaded from the NSSO official site.

```
#Using read_excel() to import the excel file

library(readxl)

west_bengal<-read_excel("C:/Users/TANISHA/Downloads/Kajal Ma'am/Block-7 - Level-06 Particulars of persons currently not attending any educational institute (Tanisha) (1).xlsx",sheet="West Bengal")
```

After importing the Excel file we want to create the indicator variables, where x_i being the indicator of ever enrolled but not attending the institution currently and y_i being the indicator of ever enrolled but not attending the institution currently and discontinued with reason not interested in education. The variables take 0 and 1 as values.

```
#Creating the indicator variables x and y

#x_i being the indicator of ever enrolled but not attending the institution currently

#y_i being the indicator of ever enrolled but not attending the institution currently and discontinued with reason not interested in education

x_i<-ifelse(west_bengal$enrolled=="yes",1,0)

print(x_i)

y_i<-ifelse(west_bengal$enrolled=="yes" & west_bengal$never_enrol_reason=="not interested in education",1,0)

print(y_i)
```

Now we create a multiplier array from the file by taking mult_i as the array of multiplier of the data. Then we compute the values, X_hat= sum(x_i*mult_i), Y_hat= sum(y_i*mult_i) and R_hat= Y_hat/X_hat.

```
#Creating multiplier array from the file

mult_i<-as.array(west_bengal$wgt_combined)

print(mult_i)

#Calculating the value of X_hat, Y_hat and R_hat

X_hat<-sum(x_i*mult_i)

print(X_hat)
[1] 173220062

Y_hat<-sum(y_i*mult_i)

print(Y_hat)
[1] 34452071

R_hat<-Y_hat/X_hat

print(R_hat)
[1] 0.1988919
```

Now we download the Block 4 data from the NSSO official site and import the data in R Studio. The next step is to merge the data of Block 4 and Block 7 together. We merge the data together according to the household identification number (HH_ID) and the person serial number (psrl_no) and save the merged data in another Excel sheet.

```
##Merging of Block 4 and Block 7
install.packages("openxlsx")
install.packages("dplyr")
library(readxl)

#importing two excel files of block 4 and block 7 for All India
block_4<-read_excel("C:/Users/TANISHA/Downloads/Kajal Ma'am/Block 4 and 7 new.xlsx",sheet="Block 4")
block_7<-read_excel("C:/Users/TANISHA/Downloads/Kajal Ma'am/Block 4 and 7 new.xlsx",sheet="Block 7")

#merging the details of two excel files
library(dplyr)
```

```
merged<-inner_join(block_4,block_7,by=c("HH_ID","psrl_no"))
print(merged)
summary(merged)

#exporting the merged data in an excel file
library(openxlsx)
write.xlsx(merged,"C:/Users/TANISHA/ Downloads/Kajal Ma'am/Sample_Survey.xlsx")
```

Now we write 19 unenrollment reasons along with their codes:

Enrol_cd	Reason
1	not interested in education
2	financial constraints
3	engaged in domestic activities
4	engaged in economic activities
5	School is far off
6	timings of educational institute not suitable
7	language/medium of instruction used unfamiliar
8	inadequate number of teachers
10	quality of teachers not satisfactory
11	Never enrolled - no tradition in community
12	Ever enrolled - unable to cope up with studies or failure in studies
13	Ever enrolled unfriendly atmosphere at school
14	Ever enrolled completed desired level/class
15	Ever enrolled preparation for competitive examination
16	Girl student - non-availability of female teacher
17	Girl student - non-availability of girls' toilet
18	Girl student - marriage
19	Others

We import the data of the Enrollment codes and reasons and merge these codes and names and check the number.

```
#importing the Enrollment codes for enrollment details

library(readxl)

enrol_data<-read_excel("C:/Users/TANISHA
/Downloads/KajalMa'am/Sample_Survey.xlsx",sheet="Enrollment code")

enrol_cd<-enroll_data$enroll_cd

print(enrol_cd)
```

```
[1]  1  2  3  4  5  6  7  8 10 11 12 13 14 15 16 17 18 19
```

never_enrol_reason<-enrol_data$Reason

print(never_enrol_reason)

```
[1] "not interested in education"
 [2] "financial constraints"
 [3] "engaged in domestic activities"
 [4] "engaged in economic activities"
 [5] "School is far off"
 [6] "timings of educational institution  not suitable"
 [7] "language/medium of instruction used unfamiliar"
 [8] "inadequate  number of teachers"
 [9] "quality of teachers not satisfactory"
[10] "Never enrolled - no tradition in the community"
[11] "Ever enrolled - unable to cope up with studies or failure in
      studies"
[12] "Ever enrolled unfriendly atmosphere at school"
[13] "Ever enrolled completed desired level/class"
[14] "Ever enrolled preparation for competitive examination"
[15] "Girl student - non-availability of female teacher"
[16] "Girl student - non-availability of girls toilet"
[17] "Girl student - marriage"
[18] "Others"
```

#creating a vector that denotes the never enrollment reason by their code

enrol_map<-setNames(never_enrol_reason,enrol_cd)

print(enrol_map)

```
                                                  1
                                     "not interested in education"
                                                  2
                                           "financial constraints"
                                                  3
                                    "engaged in domestic activities"
                                                  4
                                    "engaged in economic activities"
                                                  5
                                                "School is far off"
                                                  6
                         "timings of educational institution  not suitable"
                                                  7
                           "language/medium of instruction used unfamiliar"
                                                  8
                                        "inadequate  number of teachers"
                                                 10
                                  "quality of teachers not satisfactory"
                                                 11
                          "Never enrolled - no tradition in the community"
                                                 12
"Ever enrolled - unable to cope up with studies or failure in studies"
                                                 13
                           "Ever enrolled unfriendly atmosphere at school"
                                                 14
                             "Ever enrolled completed desired level/class"
                                                 15
                    "Ever enrolled preparation for competitive examination"
                                                 16
                        "Girl student - non-availability of female teacher"
                                                 17
                    "Girl student - non-availability of girls\u0092 toilet"
```

```
                                                                    18
                                          "Girl student - marriage"
                                                                    19
                                                         "Others"
```

#let i be the language spoken at home and let j be the medium of instruction

#x_i be the indicator of kth person ever enrolled or not

#x_j be the indicator of kth person unenrollment reason

j<-enrol_cd

print(j)

m<-length(j)

print(m)

```
[1] 18
```

Now we see that the number of individuals who unenrolled due to some reasons are low so we mergerd them together as “Others”.

```
library(readxl)
enrol_data<-
read_excel("C:/Users/TANISHA/Downloads/Kajal Ma'am/Codes.xlsx",sheet="Enr ollment_mer
ged")
enrol_cd<-enrol_data$Enrol_cd
print(enrol_cd)
[1]  1  2  3  4  5  6  7  8  9 10
```

```
never_enrol_reason<-enrol_data$never_enrol_reason
print(never_enrol_reason)
 [1] "not interested in education"
 [2] "financial constraints"
 [3] "engaged in domestic activities"
 [4] "engaged in economic activities"
 [5] "School is far off"
 [6] "quality of teachers not satisfactory"
 [7] "Ever enrolled - unable to cope up with studies or failure in
      studies"
 [8] "Ever enrolled completed desired level/class"
 [9] "Girl student - marriage"
[10] "Others"
```

```
#creating a vector that denotes the never enrollment reason by their code
enrol_map<-setNames(never_enrol_reason,enrol_cd)
print(enrol_map)
```

```
                                                                    1
                                          "not interested in education"
                                                                    2
                                                "financial constraints"
```

```
                                                                    3
                                          "engaged in domestic activities"
                                                                    4
                                          "engaged in economic activities"
                                                                    5
                                                       "School is far off"
                                                                    6
                                     "quality of teachers not satisfactory"
                                                                    7
"Ever enrolled - unable to cope up with studies or failure in studies"
                                                                    8
                                    "Ever enrolled completed desired level/class"
                                                                    9
                                                  "Girl student - marriage"
                                                                   10
                                                                "Others"
```

j<-enrol_cd

print(j)

m<-length(j)

print(m)

[1] 10

Now we want to evaluate the Rhat data where, R_hat =Y_hat/X_hat, X_hat =sum(x_i*mult_i) and Y_hat =sum(y_i*mult_i). To save the R_hat_ij data, we create a matrix and create a work book to save the corresponding R_hat_ij values. Similarly for the MSE (Mean Square Error) and RSE (Relative Square Error) of these R_hat values we need to create a matrix and a workbook to save the values.

```
#Creating a matrix for the R_hat_ij values
R_hat_ij <- matrix(nrow = m,ncol = n)
#creating matrix for all MSE of every corresponding R_hat
MSE_R_hat <- matrix(nrow=m, ncol=n)
#creating matrix for all RSE of every corresponding R_hat
RSE_R_hat <- matrix(nrow=m, ncol=n)
#Creating a workbook to save the R_hat_ij
library(openxlsx)
Wb<-createWorkbook(title="R_hat_ij")
#Creating a workbook to save the MSE_R_hat
library(openxlsx)
Wb2<-createWorkbook(title="MSE_R_hat")
#Creating a workbook to save the RSE_R_hat
library(openxlsx)
Wb3<-createWorkbook(title="RSE_R_hat")
```

Now on the basis of sector (urban, rural) and gender (male, female) the data can be divided into nine parts. The nine divisions are:

1. Urban, Male
2. Urban, Female
3. Urban, Male + Female
4. Rural, Male
5. Rural, Female
6. Rural, Male + Female
7. Urban + Rural, Male
8. Urban + Rural, Female
9. Urban + Rural, Male + Female

Now we evaluate the R_hat, MSE_R_hat and RSE_R_hat values for all divisions for both west Bengal and India by using the formulas:

$$\widehat{MSE}(\hat{R}) = \frac{1}{4\hat{X}^2} \sum_s \sum_t (\hat{Y}_{st\,1} - \hat{Y}_{st\,2})^2 + \hat{R}^2 (\hat{X}_{st\,1} - \hat{X}_{st\,2})^2 - 2\hat{R} (\hat{Y}_{st\,1} - \hat{Y}_{st\,2}) (\hat{X}_{st\,1} - \hat{X}_{st\,2})$$

and

$$\widehat{RSE}(\hat{R}) = \frac{\sqrt{\widehat{MSE}(\hat{R})}}{\hat{R}} \times 100$$

Now we see that, we have to take in consideration every stratum, sub-stratum and also sub-sample divisions of the data to calculate the MSE_R_hat values. Hence we need to know the number of stratum, sub-stratum and sub-sample given in the data.

```
s_max<-max(survey_data$stratum)
print(s_max)
[1] 135

ss_max<-max(survey_data$sstratum)
print(ss_max)
[1] 50

subs_max<-max(survey_data$ssample)
print(subs_max)
[1] "Sub-Sample 2"
```

We write the codes only for the first division that is Urban,Male for both West Bengal and All-India, and mention the difference of the codes for other divisions.

➢ West Bengal:

At first, we evaluate the R_hat_ij, MSE_R_hat, and RSE_R_hat values sequentially for West Bengal. During the evaluation of the West Bengal data, we see that the state code for West Bengal is 19, and need to use this in the code.

```
#Estimation of variance of every R_hat for (Urban, Male)

a<-1

for(u in 1:n){

x_i<-ifelse(survey_data_t$state_cd=="19" & survey_data_t$sector=="Urban" &
survey_data_t$gender=="Male" & survey_data_t$enrolled==enrolled_map[a],1,0)

X_hat<-sum(x_i*mult_i)

b<-1

for (v in 1:m){

y_i<-ifelse(survey_data_t$state_cd=="19" & survey_data_t$sector=="Urban" &
survey_data_t$gender=="Male" & survey_data_t$enrolled==enrolled_map[a] &
survey_data_t$never_enrol_reason==enrol_map[b],1,0)

Y_hat<-sum(y_i*mult_i)

R_hat<-Y_hat/X_hat

print(R_hat)

variance<-0

c<-1

for(s in 1:s_max){

var_s<-0

d<-1

for(t in 1:ss_max){

var_st<-0

Y_hat_st1_i<-ifelse(survey_data_t$state_cd=="19" & survey_data_t$stratum==c &
survey_data_t$sstratum==d & survey_data_t$ssample=="Sub-Sample 1" &
survey_data_t$sector=="Urban" & survey_data_t$gender=="Male"  &
survey_data_t$enrolled==enrolled_map[a] &
survey_data_t$never_enrol_reason==enrol_map[b],1,0)

Y_hat_st1<-sum(Y_hat_st1_i*mult_i)
```

```
X_hat_st1_i<-ifelse(survey_data_t$state_cd=="19" & survey_data_t$stratum==c &
survey_data_t$sstratum==d & survey_data_t$ssample=="Sub-Sample 1" &
survey_data_t$sector=="Urban" & survey_data_t$gender=="Male" &
survey_data_t$enrolled==enrolled_map[a],1,0)

X_hat_st1<-sum(X_hat_st1_i*mult_i)

Y_hat_st2_i<-ifelse(survey_data_t$state_cd=="19" & survey_data_t$stratum==c &
survey_data_t$sstratum==d & survey_data_t$ssample=="Sub-Sample 2" &
survey_data_t$sector=="Urban" & survey_data_t$gender=="Male" &
survey_data_t$enrolled==enrolled_map[a] &
survey_data_t$never_enrol_reason==enrol_map[b],1,0)

Y_hat_st2<-sum(Y_hat_st2_i*mult_i)

X_hat_st2_i<-ifelse(survey_data_t$state_cd=="19" & survey_data_t$stratum==c &
survey_data_t$sstratum==d & survey_data_t$ssample=="Sub-Sample 2" &
survey_data_t$sector=="Urban" & survey_data_t$gender=="Male" &
survey_data_t$enrolled==enrolled_map[a],1,0)

X_hat_st2<-sum(X_hat_st2_i*mult_i)

Y_hat_st<-Y_hat_st1-Y_hat_st2

X_hat_st<-X_hat_st1-X_hat_st2

var_st<-((Y_hat_st**2)+(R_hat**2)*(X_hat_st**2)-2*(R_hat*Y_hat_st*X_hat_st))

var_s<-var_s+var_st

d<-d+1

}

variance<-variance+var_s

c<-c+1

}

den<-4*(X_hat**2)

varij<-variance/den

print(varij)

MSE_R_hat[b,a]<-varij

num<-sqrt(varij)

rse_ij<-(num/R_hat)*100

rse_rhat<-round(rse_ij,2)
```

```
print(rse_rhat)
RSE_R_hat[b,a]<-rse_rhat
b<-b+1
}
a<-a+1
}
print(MSE_R_hat)
print(RSE_R_hat)

#Saving the MSE_R_hat data in an excel file
addWorksheet(Wb2,"Urban, Male")
writeData(Wb2,"Urban, Male", MSE_R_hat)

#Saving the RSE_R_hat data in an excel file
addWorksheet(Wb3,"Urban, Male")
writeData(Wb3,"Urban, Male", RSE_R_hat)

#saving the MSE_R_hat for all the divisions in a single excel file
saveWorkbook(Wb2,"C:/Users/TANISHA/Downloads/Kajal Ma'am/MSE_R_hat1.xlsx")
#saving the RSE_R_hat for all the divisions in a single excel file
saveWorkbook(Wb3,"C:/Users/TANISHA/Downloads/Kajal Ma'am/RSE_R_hat1.xlsx")
```

The difference of the codes for other divisions are

2. Urban, Female for West Bengal:

The restrictions for this division are given by: survey_data$sector== “Urban” & survey_data$gender== “Female” This restriction is used in every statement that gives the indicator variables.

3. Urban, Male + Female for West Bengal:

The restrictions for this division are given by: survey_data$sector== “Urban” The restriction of gender is not used in this division as both the genders are considered in this division. This

restriction is used in every statement that gives the indicator variables.

4. Rural, Male for West Bengal:

The restrictions for this division are given by: survey_data$sector== “Rural” & survey_data$gender== “Male” This restriction is used in every statement that gives the indicator variables.

5. Rural, Female for West Bengal:

The restrictions for this division are given by: survey_data$sector== “Rural” & survey_data$gender== “Female”. This restriction is used in every statement that gives the indicator variables.

6. Rural, Male + Female for West Bengal:

The restrictions for this division are given by: survey_data$sector== “Rural” The restriction of gender is not used in this division as both the genders are considered in this division. This restriction is used in every statement that gives the indicator variables.

7. Rural + Urban, Male for West Bengal:

The restrictions for this division are given by: survey_data$gender== “Male” The restriction of sector is not used in this division as both the sectors are considered in this division. This restriction is used in every statement that gives the indicator variables.

8. Rural + Urban, Female for West Bengal:

The restrictions for this division are given by: survey_data$gender== Female” The restriction of sector is not used in this division as both the sectors are considered in this division. This restriction is used in every statement that gives the indicator variables.

9. Rural + Urban, Male + Female for West Bengal:

This division has restriction of neither sector nor gender as both the sectors and genders are taken into consideration. Now we include the tables of the R_hat, MSE_R_hat, and RSE_R_hat values of every division.

The output tables of R_hat, MSE_R_hat and RSE_R_hat values for all divisions of West Bengal are in percentage.

1. Urban, Male for West Bengal:

a. R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	26.16	45.11
2	34.27	27.50
3	1.56	0.00
4	20.34	3.51
5	0.00	0.00
6	0.00	0.00
7	2.95	0.00
8	7.76	0.00
9	0.00	0.00
10	6.95	23.89
Total	100	100

b. MSE_R_hat:

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	1.46	9.4
2	1.56	10.07
3	0.11	0
4	1.13	1.75
5	0	0
6	0	0
7	0.1	0
8	0.45	0
9	0	0
10	0.26	8.82

c. RSE_R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	4.61	6.8
2	3.65	11.54
3	21.03	#NUM!
4	5.23	37.74
5	#NUM!	#NUM!
6	#NUM!	#NUM!
7	10.83	#NUM!
8	8.61	#NUM!
9	#NUM!	#NUM!
10	7.36	12.44

Interpretation:

For urban males in West Bengal, the reasons for discontinuing, dropping out, or never enrolling in education differ greatly between those who are currently enrolled and those who are not. Among those who are still in school, the main reasons are Reason 2 (34.27%), Reason 1 (26.16%), and Reason 4 (20.34%). Other reasons play a smaller role. For those not enrolled, the biggest reasons are Reason 1 (45.11%) and Reason 10 (23.89%), with Reason 2 (27.50%) also being significant. Reasons 3, 5, 6, 7, 8, and 9 don't affect non-enrollment at all. This indicates that while many factors help keep students in school, only a few key reasons lead to them not being enrolled.

2. Urban, Female for West Bengal:

a. R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	15.36	32.01
2	27.05	39.88
3	13.11	11.28
4	2.75	2.20
5	0.37	1.78
6	0.00	0.00
7	2.68	0.00
8	7.85	0.00
9	22.82	0.00
10	8.01	12.85
Total	100	100

b. MSE_R_hat:

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	0.69	11.04
2	1.11	10.68
3	0.61	4.02
4	0.12	1.18
5	0.01	0.71
6	0	0
7	0.09	0
8	0.35	0
9	1.2	0
10	0.34	3.31

c. RSE_R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	5.42	10.38
2	3.89	8.2
3	5.97	17.77
4	12.77	49.34
5	30.74	47.21
6	#NUM!	#NUM!
7	11.17	#NUM!
8	7.5	#NUM!
9	4.81	#NUM!
10	7.26	14.16

Interpretation:

For urban females in West Bengal, the reasons for either staying in school or not attending differ quite a bit. Among those who are still enrolled, the main reasons are Reason 2 (27.05%), followed by Reason 9 (22.82%), and Reason 1 (15.36%). Other reasons such as Reason 3 (13.11%), Reason 8 (7.85%), and Reason 10 (8.01%) also play a role, while some reasons like Reason 6 have no impact.

On the other hand, for those who are not enrolled, the most significant reason is Reason 2 (39.88%), followed by Reason 1 (32.01%) and Reason 10 (12.85%). Some reasons like Reason 3 (11.28%) and Reason 5 (1.78%) are less significant, while reasons like Reason 7 and Reason 8 have no impact.

This shows that while a variety of factors influence whether urban females remain enrolled, only a few key reasons are prominent for those who drop out or never enroll.

3. Urban, Male + Female for West Bengal:

a. R_hat (in%):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	20.60	37.98
2	30.55	34.24
3	7.51	6.14
4	11.28	2.79
5	0.19	0.97
6	0.00	0.00
7	2.81	0.00
8	7.81	0.00
9	11.75	0.00
10	7.50	17.88
Total	100	100

b. MSE_R_hat:

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	0.65	6.39
2	0.76	6.56
3	0.21	1.25
4	0.32	0.69
5	0	0.22
6	0	0
7	0.05	0
8	0.25	0
9	0.36	0
10	0.17	3.03

c. RSE_R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	3.9	6.66
2	2.85	7.48
3	6.06	18.2
4	4.98	29.81
5	30.89	48.27
6	#NUM!	#NUM!
7	8.24	#NUM!
8	6.34	#NUM!
9	5.08	#NUM!
10	5.54	9.73

Interpretation:

For urban males and females in West Bengal, the reasons for staying in or dropping out of school vary. Among those who are still enrolled, the top reasons include Reason 2 (30.55%), Reason 1 (20.60%), and Reason 4 (11.28%). Other contributing factors include Reason 9 (11.75%), Reason 8 (7.81%), and Reason 10 (7.50%). These reasons show a diverse set of factors influencing continued enrollment.

In contrast, for those who are not enrolled, the primary reasons are Reason 1 (37.98%) and Reason 2 (34.24%), followed by Reason 10 (17.88%). Other reasons like Reason 3 (6.14%) and Reason 4 (2.79%) are less significant, while several reasons (Reasons 6, 7, 8, and 9) have no impact on non-enrollment.

This data highlights that while a variety of factors keep students in school, a few key reasons dominate when it comes to dropping out or never enrolling.

4. Rural, Male for West Bengal:

a. R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	21.56	34.76
2	37.06	34.73
3	2.36	0.78
4	23.78	4.42
5	0.13	0.53
6	0.00	0.00
7	5.14	0.00
8	4.94	0.00
9	0.00	0.00
10	5.03	24.78
Total	100	100

b. MSE_R_hat:

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	0.79	6.08
2	1.04	9.1
3	0.08	0.15
4	0.74	0.89
5	0	0.07
6	0	0
7	0.23	0
8	0.19	0
9	0	0
10	0.18	5.36

c. RSE_R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	4.12	7.1
2	2.75	8.69
3	12.14	49.88
4	3.63	21.31
5	49.9	49.84
6	#NUM!	#NUM!
7	9.25	#NUM!
8	8.84	#NUM!
9	#NUM!	#NUM!
10	8.48	9.35

Interpretation:

For rural males in West Bengal, the reasons for staying in school or not attending differ notably. Among those who are still enrolled, the most significant reasons are Reason 2 (37.06%) and Reason 4 (23.78%), followed by Reason 1 (21.56%). Other reasons like Reason 7 (5.14%), Reason 8 (4.94%), and Reason 10 (5.03%) also contribute, while some reasons have no impact.

In contrast, for those not enrolled, the main reasons for dropping out or never enrolling are Reason 1 (34.76%) and Reason 2 (34.73%), both equally significant. Reason 10 (24.78%) also plays a major role, with minor contributions from other reasons such as Reason 4 (4.42%) and Reason 3 (0.78%). Several reasons, including Reasons 6, 7, 8, and 9, have no impact.

This indicates that while a range of factors influence continued enrollment among rural males, only a few key reasons are crucial for those who drop out or never enroll.

5. Rural, Female for West Bengal:

a. R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	12.01	34.85
2	31.04	36.59
3	18.84	9.61
4	3.12	1.29
5	1.12	1.70
6	0.00	0.00
7	4.91	0.00
8	2.24	0.00
9	23.17	0.00
10	3.56	15.95
Total	100	100

b. MSE_R_hat:

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	0.37	4.13
2	1.18	4.16
3	0.61	1.32
4	0.14	0.21
5	0.04	0.26
6	0	0
7	0.26	0
8	0.08	0
9	1	0
10	0.13	2.14

c. RSE_R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	5.08	5.83
2	3.5	5.57
3	4.14	11.97
4	11.82	35.55
5	17.42	30.09
6	#NUM!	#NUM!
7	10.33	#NUM!
8	12.83	#NUM!
9	4.31	#NUM!
10	10.03	9.17

Interpretation:

For rural females in West Bengal, the reasons for staying in school or not attending vary widely. Among those who are still enrolled, the main reasons are Reason 2 (31.04%), Reason 9 (23.17%), and Reason 3 (18.84%). Other contributing factors include Reason 1 (12.01%), Reason 7 (4.91%), and Reason 10 (3.56%), while some reasons have no impact.

For those not enrolled, the most significant reasons are Reason 2 (36.59%) and Reason 1 (34.85%), followed by Reason 10 (15.95%). Other reasons like Reason 3 (9.61%) and Reason 5 (1.70%) have a smaller impact, and several reasons, such as Reasons 6, 7, 8, and 9, have no influence on non-enrollment.

This suggests that while various factors contribute to continued enrollment among rural females, a few key reasons are predominant for those who drop out or never enroll.

6. Rural, Male + Female for West Bengal:

a. R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	16.94	34.81
2	34.15	35.75
3	10.33	5.61
4	13.79	2.71
5	0.61	1.17
6	0.00	0.00
7	5.03	0.00
8	3.63	0.00
9	11.21	0.00
10	4.32	19.95
Total	100	100

b. MSE_R_hat:

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	0.33	3.09
2	0.71	4.3
3	0.16	0.47
4	0.27	0.25
5	0.01	0.09
6	0	0
7	0.17	0
8	0.08	0
9	0.25	0
10	0.11	2.6

c. RSE_R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	3.38	5.05
2	2.47	5.8
3	3.88	12.25
4	3.78	18.34
5	16.37	25.99
6	#NUM!	#NUM!
7	8.17	#NUM!
8	7.63	#NUM!
9	4.47	#NUM!
10	7.55	8.09

Interpretation:

For rural males and females in West Bengal, the reasons for staying in school or not attending show clear patterns. Among those who are still enrolled, the main reasons include Reason 2 (34.15%), Reason 4 (13.79%), and Reason 3 (10.33%). Other factors like Reason 1 (16.94%) and Reason 9 (11.21%) also contribute, while some reasons have minimal impact.

For those not enrolled, the most significant reasons are Reason 2 (35.75%) and Reason 1 (34.81%), with Reason 10 (19.95%) also being important. Other reasons like Reason 3 (5.61%) and Reason 4 (2.71%) have a smaller role, while several reasons have no impact at all.

This highlights that while a variety of factors are involved in keeping rural students in school, only a few key reasons are major contributors to why students drop out or never enroll.

7. Urban + Rural, Male for West Bengal:

a. R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	23.51	38.04
2	35.88	32.44
3	2.02	0.53
4	22.33	4.13
5	0.07	0.36
6	0.00	0.00
7	4.21	0.00
8	6.14	0.00
9	0.00	0.00
10	5.85	24.50
Total	100	100

b. MSE_R_hat:

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	0.54	4.07
2	0.7	5.27
3	0.05	0.07
4	0.48	0.59
5	0	0.03
6	0	0
7	0.1	0
8	0.14	0
9	0	0
10	0.11	3.44

c. RSE_R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	3.12	5.3
2	2.34	7.08
3	10.73	49.91
4	3.1	18.59
5	49.95	50.22
6	#NUM!	#NUM!
7	7.45	#NUM!
8	6.13	#NUM!
9	#NUM!	#NUM!
10	5.57	7.57

Interpretation:

For both urban and rural males in West Bengal, the factors affecting whether they stay in school or not show some noticeable patterns. Among those who are still enrolled, the main reasons include Reason 2 (35.88%), Reason 4 (22.33%), and Reason 1 (23.51%). Other factors such as Reason 8 (6.14%) and Reason 10 (5.85%) also contribute, though they play a smaller role.

For those who are not enrolled, the primary reasons are Reason 1 (38.04%) and Reason 2 (32.44%), with Reason 10 (24.50%) also being a significant factor. Reasons like Reason 3 (0.53%) and Reason 4 (4.13%) have a lesser impact, while several reasons have no influence on non-enrollment.

This highlights that while a variety of reasons influence continued enrollment, only a few key factors are major contributors to why students drop out or never enroll.

8. Urban + Rural, Female for West Bengal:

a. R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	13.53	33.96
2	29.23	37.63
3	16.24	10.14
4	2.95	1.58
5	0.78	1.73
6	0.00	0.00
7	3.90	0.00
8	4.79	0.00
9	23.01	0.00
10	5.58	14.98
Total	100	100

b. MSE_R_hat:

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	0.24	2.91
2	0.59	3
3	0.3	1.02
4	0.06	0.22
5	0.01	0.19
6	0	0
7	0.1	0
8	0.1	0
9	0.57	0
10	0.11	1.3

c. RSE_R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	3.62	5.02
2	2.62	4.61
3	3.36	9.96
4	8.42	29.51
5	15.27	25.47
6	#NUM!	#NUM!
7	8.01	#NUM!
8	6.64	#NUM!
9	3.28	#NUM!
10	5.92	7.6

Interpretation:

For urban and rural females in West Bengal, the reasons for staying in school or not attending show some clear trends. Among those who are still enrolled, the most significant reasons are Reason 2 (29.23%), Reason 9 (23.01%), and Reason 3 (16.24%). Other factors like Reason 1 (13.53%) and Reason 10 (5.58%) also play a role, though to a lesser extent.

For those not enrolled, the major reasons are Reason 2 (37.63%) and Reason 1 (33.96%), with Reason 10 (14.98%) also being a notable factor. Other reasons like Reason 3 (10.14%) and Reason 5 (1.73%) have a smaller impact, while several reasons have no effect on non-enrollment.

This data underscores that while a variety of factors influence continued enrollment, only a few key reasons are dominant in explaining why students drop out or never enroll.

9. Urban + Rural, Male + Female for West Bengal:

a. R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	18.55	35.81
2	32.57	35.27
3	9.09	5.78
4	12.69	2.74
5	0.42	1.11
6	0.00	0.00
7	4.05	0.00
8	5.46	0.00
9	11.44	0.00
10	5.72	19.30
Total	100	100

b. MSE_R_hat:

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	0.22	2.09
2	0.42	2.73
3	0.09	0.34
4	0.16	0.19
5	0	0.07
6	0	0
7	0.07	0
8	0.07	0
9	0.15	0
10	0.07	1.53

c. RSE_R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	2.51	4.04
2	1.99	4.68
3	3.29	10.08
4	3.12	15.82
5	14.57	23.14
6	#NUM!	#NUM!
7	6.31	#NUM!
8	4.89	#NUM!
9	3.43	#NUM!
10	4.55	6.41

Interpretation:

For both urban and rural students in West Bengal, the reasons for staying in school versus those for dropping out or never enrolling show distinct patterns. Among those who remain enrolled, the most significant reasons are Reason 2 (32.57%), Reason 4 (12.69%), and Reason 9 (11.44%). Other contributing factors include Reason 1 (18.55%) and Reason 8 (5.46%), with Reason 10 (5.72%) playing a smaller role.

For those not enrolled, the key reasons are Reason 1 (35.81%) and Reason 2 (35.27%), with Reason 10 (19.30%) also being important. Other reasons such as Reason 3 (5.78%) and Reason 4 (2.74%) have a lesser impact, while several reasons do not affect non-enrollment at all.

This data illustrates that while various factors are involved in why students continue their education, a few dominant reasons are significant in explaining why students drop out or never enroll.

➢ India:

At first, we evaluate the R_hat_ij, MSE_R_hat, and RSE_R_hat values sequentially for India. During the evaluation of the India data, unlike the analysis of West Bengal data state codes are not used.

```
#Estimation of variance of every R_hat for (Urban, Male)

a<-1

for(u in 1:n){

x_i<-ifelse(survey_data_t$sector=="Urban" & survey_data_t$gender=="Male" &
survey_data_t$enrolled==enrolled_map[a],1,0)

X_hat<-sum(x_i*mult_i)

b<-1

for (v in 1:m){

y_i<-ifelse(survey_data_t$sector=="Urban" & survey_data_t$gender=="Male" &
survey_data_t$enrolled==enrolled_map[a] &
survey_data_t$never_enrol_reason==enrol_map[b],1,0)

Y_hat<-sum(y_i*mult_i)

R_hat<-Y_hat/X_hat

print(R_hat)

r_hat[b,a]<-R_hat

variance<-0

c<-1

for(s in 1:s_max){

var_s<-0

d<-1

for(t in 1:ss_max){

var_st<-0

Y_hat_st1_i<-ifelse(survey_data_t$stratum==c & survey_data_t$sstratum==d &
survey_data_t$ssample=="Sub-Sample 1" & survey_data_t$sector=="Urban" &
survey_data_t$gender=="Male"  & survey_data_t$enrolled==enrolled_map[a] &
survey_data_t$never_enrol_reason==enrol_map[b],1,0)

Y_hat_st1<-sum(Y_hat_st1_i*mult_i)
```

```
X_hat_st1_i<-ifelse(survey_data_t$stratum==c & survey_data_t$sstratum==d &
survey_data_t$ssample=="Sub-Sample 1" & survey_data_t$sector=="Urban" &
survey_data_t$gender=="Male" & survey_data_t$enrolled==enrolled_map[a],1,0)

X_hat_st1<-sum(X_hat_st1_i*mult_i)

Y_hat_st2_i<-ifelse(survey_data_t$stratum==c & survey_data_t$sstratum==d &
survey_data_t$ssample=="Sub-Sample 2" & survey_data_t$sector=="Urban" &
survey_data_t$gender=="Male" & survey_data_t$enrolled==enrolled_map[a] &
survey_data_t$never_enrol_reason==enrol_map[b],1,0)

Y_hat_st2<-sum(Y_hat_st2_i*mult_i)

X_hat_st2_i<-ifelse(survey_data_t$stratum==c & survey_data_t$sstratum==d &
survey_data_t$ssample=="Sub-Sample 2" & survey_data_t$sector=="Urban" &
survey_data_t$gender=="Male" & survey_data_t$enrolled==enrolled_map[a],1,0)

X_hat_st2<-sum(X_hat_st2_i*mult_i)

Y_hat_st<-Y_hat_st1-Y_hat_st2

X_hat_st<-X_hat_st1-X_hat_st2

var_st<-((Y_hat_st**2)+(R_hat**2)*(X_hat_st**2)-2*(R_hat*Y_hat_st*X_hat_st))

var_s<-var_s+var_st

d<-d+1

}

variance<-variance+var_s

c<-c+1

}

den<-4*(X_hat**2)

varij<-variance/den

print(varij)

MSE_R_hat[b,a]<-varij

num<-sqrt(varij)

rse_ij<-(num/R_hat)*100

rse_rhat<-round(rse_ij,2)

print(rse_rhat)

RSE_R_hat[b,a]<-rse_rhat
```

```
b<-b+1
}
a<-a+1
}
print(r_hat)
print(MSE_R_hat)
print(RSE_R_hat)

#Saving the MSE_R_hat data in an excel file
addWorksheet(Wb2,"Urban, Male")
writeData(Wb2,"Urban, Male", MSE_R_hat)

#Saving the RSE_R_hat data in an excel file
addWorksheet(Wb3,"Urban, Male")
writeData(Wb3,"Urban, Male", RSE_R_hat)

#Saving the r_hat data in an excel file
addWorksheet(Wb4,"Urban, Male")
writeData(Wb4,"Urban, Male",r_hat)

#saving the MSE_R_hat for all the divisions in a single excel file
saveWorkbook(Wb2,"C:/Users/TANISHA/Downloads/Kajal Ma'am/MSE_R_hat1_merged.xlsx")
#saving the RSE_R_hat for all the divisions in a single excel file
saveWorkbook(Wb3,"C:/Users/TANISHA/Downloads/Kajal Ma'am/RSE_R_hat1_merged.xlsx")
#saving the r_hat for all the divisions in a single excel file
saveWorkbook(Wb4,"C:/Users/TANISHA/Downloads/Kajal Ma'am/r_hat1_merged.xlsx")
```

The difference of the codes for other divisions are

2. Urban, Female for India: The restrictions for this division are given by: survey_data$sector== “Urban” & survey_data$gender== “Female” This restriction is used in every statement that gives the indicator variables.

3. Urban, Male + Female for India: The restrictions for this division are given by: survey_data$sector== “Urban” The restriction of gender is not used in this division as both the genders are considered in this division. This restriction is used in every statement that gives the indicator variables.

4. Rural, Male for India: The restrictions for this division are given by: survey_data$sector== “Rural” & survey_data$gender== “Male” This restriction is used in every statement that gives the indicator variables.

5. Rural, Female for India: The restrictions for this division are given by: survey_data$sector== “Rural” & survey_data$gender== “Female”. This restriction is used in every statement that gives the indicator variables.

6. Rural, Male + Female for India: The restrictions for this division are given by: survey_data$sector== “Rural” The restriction of gender is not used in this division as both the genders are considered in this division. This restriction is used in every statement that gives the indicator variables.

7. Rural + Urban, Male for India: The restrictions for this division are given by: survey_data$gender== “Male” The restriction of sector is not used in this division as both the sectors are considered in this division. This restriction is used in every statement that gives the indicator variables.

8. Rural + Urban, Female for India: The restrictions for this division are given by: survey_data$gender== Female” The restriction of sector is not used in this division as both the sectors are considered in this division. This restriction is used in every statement that gives the indicator variables.

9. Rural + Urban, Male + Female for India: This division has restriction of neither sector nor gender as both the sectors and genders are taken into consideration. Now we include the tables of the R_hat, MSE_R_hat, and RSE_R_hat values of every division.

The output tables of R_hat, MSE_R_hat and RSE_R_hat values for all divisions of India are in percentage.

1. Urban, Male for India:

a. R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	20.84	29.47
2	23.66	32.79
3	2.42	3.79
4	33.56	6.92
5	0.17	0.44
6	0.00	0.00
7	5.28	0.00
8	8.34	0.00
9	0.00	0.00
10	5.73	26.59
Total	100	100

b. MSE_R_hat:

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	0.14	1.93
2	0.23	1.73
3	0.02	0.17
4	0.23	0.65
5	0	0.01
6	0	0
7	0.05	0
8	0.07	0
9	0	0
10	0.04	1.46

c. RSE_R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	1.8	4.72
2	2.03	4.01
3	5.47	10.92
4	1.42	11.65
5	20.87	22.67
6	#NUM!	#NUM!
7	4.16	#NUM!
8	3.28	#NUM!
9	#NUM!	#NUM!
10	3.39	4.54

Interpretation:

For urban males in India, the factors influencing whether they stay in school or drop out present clear differences. Among those who are currently enrolled, the most common reasons include Reason 4 (33.56%), Reason 2 (23.66%), and Reason 1 (20.84%). Other factors, such as Reason 8 (8.34%) and Reason 7 (5.28%), also play a role, while Reason 5 (0.17%) and Reason 6 (0.00%) have minimal impact.

For those not enrolled, the primary reasons are Reason 2 (32.79%) and Reason 1 (29.47%), with Reason 10 (26.59%) also being a significant factor. Other reasons like Reason 3 (3.79%) and Reason 4 (6.92%) have a smaller role, while several reasons do not influence non-enrollment.

This suggests that while a variety of reasons contribute to why urban males in India remain enrolled, only a few key reasons are significant in explaining why they drop out or never enroll.

2. Urban, Female for India:

a. R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	14.29	27.10
2	14.92	30.00
3	23.12	13.41
4	6.86	1.14
5	1.76	1.66
6	0.00	0.00
7	3.61	0.00
8	11.29	0.00
9	17.15	0.42
10	7.01	26.28
Total	100	100

b. MSE_R_hat:

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	0.09	1.33
2	0.14	1.86
3	0.17	0.41
4	0.08	0.05
5	0.01	0.05
6	0	0
7	0.02	0
8	0.09	0
9	0.13	0.01
10	0.04	0.93

c. RSE_R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	2.12	4.26
2	2.53	4.55
3	1.78	4.76
4	4.15	19.63
5	6.63	13.73
6	#NUM!	#NUM!
7	4.08	#NUM!
8	2.67	#NUM!
9	2.07	20.37
10	2.97	3.66

Interpretation:

For urban females in India, the reasons for continuing their education versus those for dropping out or never enrolling highlight some notable differences. Among those who are still enrolled, the most significant factors are Reason 3 (23.12%), Reason 9 (17.15%), and Reason 8 (11.29%). Other reasons such as Reason 1 (14.29%) and Reason 2 (14.92%) also contribute, while Reason 4 (6.86%) and Reason 10 (7.01%) have a smaller impact.

For those who are not enrolled, the main reasons are Reason 2 (30.00%) and Reason 1 (27.10%), with Reason 10 (26.28%) also being a prominent factor. Reasons like Reason 3 (13.41%) and Reason 4 (1.14%) have a lesser role, while several reasons, such as Reasons 6 and 7, have no influence on non-enrollment.

This suggests that while a variety of factors influence why urban females in India stay in school, only a few key reasons are significant in explaining why they drop out or never enroll.

3. Urban, Male + Female for India:

a. R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	17.72	28.13
2	19.50	31.21
3	12.27	9.25
4	20.86	3.64
5	0.93	1.13
6	0.00	0.00
7	4.49	0.00
8	9.74	0.00
9	8.16	0.24
10	6.34	26.41
Total	100	100

b. MSE_R_hat:

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	0.07	1.18
2	0.13	1.31
3	0.05	0.19
4	0.09	0.15
5	0	0.02
6	0	0
7	0.02	0
8	0.05	0
9	0.03	0
10	0.02	0.75

c. RSE_R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	1.52	3.86
2	1.83	3.66
3	1.74	4.76
4	1.41	10.69
5	6.52	12.47
6	#NUM!	#NUM!
7	3.07	#NUM!
8	2.4	#NUM!
9	2.14	20.42
10	2.41	3.28

Interpretation:

For urban students in India, the factors influencing whether they stay in school or drop out show some clear differences. Among those who are still enrolled, the most notable reasons are Reason 4 (20.86%), Reason 2 (19.50%), and Reason 1 (17.72%). Other contributing factors include Reason 8 (9.74%) and Reason 3 (12.27%), while Reason 10 (6.34%) and others play a smaller role.

For those not enrolled, the dominant reasons are Reason 2 (31.21%) and Reason 1 (28.13%), with Reason 10 (26.41%) also being significant. Other reasons like Reason 3 (9.25%) and Reason 4 (3.64%) have a lesser impact, while several reasons, such as Reasons 6 and 7, do not influence non-enrollment at all.

This data highlights that while various factors contribute to why urban students in India remain in school, only a few key reasons are major contributors to why they drop out or never enroll.

4. Rural, Male for India:

a. R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	25.13	33.23
2	23.65	21.54
3	5.93	4.82
4	29.88	8.91
5	0.64	1.82
6	0.00	0.00
7	5.49	0.00
8	4.50	0.00
9	0.00	0.00
10	4.79	29.68
Total	100	100

b. MSE_R_hat:

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	0.11	0.64
2	0.14	0.66
3	0.03	0.16
4	0.14	0.22
5	0	0.09
6	0	0
7	0.03	0
8	0.02	0
9	0	0
10	0.03	0.58

c. RSE_R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	1.33	2.42
2	1.56	3.78
3	3.12	8.41
4	1.26	5.28
5	8.62	16.26
6	#NUM!	#NUM!
7	3.33	#NUM!
8	3.33	#NUM!
9	#NUM!	#NUM!
10	3.35	2.56

Interpretation:

For rural males in India, the reasons for continuing their education versus those for dropping out or never enrolling reveal some striking differences. Among those who are still enrolled, the primary reasons are Reason 4 (29.88%) and Reason 1 (25.13%), followed by Reason 2 (23.65%). Other factors like Reason 7 (5.49%) and Reason 3 (5.93%) also play a role, while Reason 10 (4.79%) has a smaller impact.

For those not enrolled, the major reasons are Reason 1 (33.23%) and Reason 10 (29.68%), with Reason 2 (21.54%) also being significant. Other reasons like Reason 4 (8.91%) and Reason 3 (4.82%) have a lesser role, while several reasons such as Reasons 6, 7, and 8 do not affect non-enrollment at all.

This highlights that while various factors influence why rural males in India remain in school, only a few key reasons are major contributors to why they drop out or never enroll.

5. Rural, Female for India:

a. R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	16.18	26.99
2	15.36	16.28
3	32.89	23.42
4	3.93	1.69
5	4.24	2.92
6	0.00	0.00
7	5.06	0.00
8	4.29	0.00
9	12.37	0.60
10	5.68	28.12
Total	100	100

b. MSE_R_hat:

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	0.07	0.29
2	0.08	0.28
3	0.15	0.34
4	0.02	0.02
5	0.02	0.05
6	0	0
7	0.02	0
8	0.02	0
9	0.06	0.01
10	0.03	0.34

c. RSE_R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	1.65	2
2	1.84	3.28
3	1.18	2.5
4	3.76	8.52
5	3.12	7.68
6	#NUM!	#NUM!
7	3.09	#NUM!
8	3.5	#NUM!
9	1.97	19.14
10	2.93	2.06

Interpretation:

For rural females in India, the factors influencing their continued education versus those leading to dropping out or never enrolling reveal some important patterns. Among those who are currently enrolled, the most common reasons are Reason 3 (32.89%), Reason 1 (16.18%), and Reason 9 (12.37%). Other contributing factors include Reason 2 (15.36%) and Reason 7 (5.06%), while Reason 10 (5.68%) plays a smaller role.

For those who are not enrolled, the key reasons are Reason 10 (28.12%) and Reason 1 (26.99%), with Reason 3 (23.42%) also being significant. Reasons like Reason 2 (16.28%) and Reason 5 (2.92%) also contribute, while several reasons such as Reasons 6, 7, and 8 have no impact on non-enrollment.

This data highlights that while various factors influence why rural females in India remain in school, only a few key reasons are major contributors to why they drop out or never enroll.

6. Rural, Male + Female for India:

a. R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	20.91	29.43
2	19.74	18.33
3	18.64	16.15
4	17.64	4.51
5	2.34	2.49
6	0.00	0.00
7	5.29	0.00
8	4.40	0.00
9	5.83	0.36
10	5.21	28.73
Total	100	100

b. MSE_R_hat:

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	0.06	0.27
2	0.07	0.28
3	0.05	0.17
4	0.05	0.05
5	0.01	0.05
6	0	0
7	0.02	0
8	0.01	0
9	0.01	0
10	0.02	0.31

c. RSE_R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	1.15	1.75
2	1.33	2.88
3	1.19	2.55
4	1.27	4.83
5	3.08	8.59
6	#NUM!	#NUM!
7	2.63	#NUM!
8	2.61	#NUM!
9	1.98	19.12
10	2.45	1.94

Interpretation:

For rural students across India, the reasons for continuing their education versus those leading to dropping out or never enrolling reveal clear differences. Among those who are still enrolled, the top reasons are Reason 1 (20.91%), Reason 2 (19.74%), and Reason 3 (18.64%), with Reason 4 (17.64%) also being significant. Other reasons like Reason 7 (5.29%) and Reason 9 (5.83%) contribute to a lesser extent, while Reason 10 (5.21%) plays a smaller role.

In contrast, for those who are not enrolled, the major reasons are Reason 1 (29.43%) and Reason 10 (28.73%), with Reason 2 (18.33%) also notable. Other reasons such as Reason 3 (16.15%) and Reason 4 (4.51%) have a minor impact, while reasons like Reason 6, 7, and 8 do not affect non-enrollment at all.

This highlights that while various factors influence why rural students in India continue their education, only a few key reasons are major contributors to why they drop out or never enroll.

7. Urban + Rural, Male for India:

a. R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	23.77	32.53
2	23.65	23.64
3	4.81	4.63
4	31.05	8.54
5	0.49	1.56
6	0.00	0.00
7	5.42	0.00
8	5.72	0.00
9	0.00	0.00
10	5.09	29.10
Total	100	100

b. MSE_R_hat:

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	0.06	0.48
2	0.09	0.5
3	0.02	0.12
4	0.09	0.16
5	0	0.06
6	0	0
7	0.02	0
8	0.02	0
9	0	0
10	0.02	0.46

c. RSE_R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	1.05	2.12
2	1.25	2.99
3	2.82	7.38
4	0.94	4.75
5	7.76	15.42
6	#NUM!	#NUM!
7	2.53	#NUM!
8	2.32	#NUM!
9	#NUM!	#NUM!
10	2.5	2.32

Interpretation:

For male students across both urban and rural areas in India, there are some clear patterns regarding their continued education versus those who drop out or never enroll. For those who are still enrolled, the main reasons include Reason 4 (31.05%), Reason 1 (23.77%), and Reason 2 (23.65%). Other factors such as Reason 7 (5.42%) and Reason 8 (5.72%) also play a role, though to a lesser extent. Reason 10 (5.09%) is also a minor factor.

In contrast, for those who are not enrolled, the most significant reasons are Reason 1 (32.53%) and Reason 10 (29.10%), with Reason 2 (23.64%) also being notable. Reasons like Reason 4 (8.54%) and Reason 3 (4.63%) contribute less, while reasons like Reason 6, 7, and 8 do not influence non-enrollment at all.

This data underscores that while various factors influence why male students continue their education, a few key reasons significantly contribute to why they drop out or never enroll.

8. Urban + Rural, Female for India:

a. R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	15.57	27.01
2	15.22	18.50
3	29.74	21.80
4	4.87	1.60
5	3.44	2.71
6	0.00	0.00
7	4.59	0.00
8	6.55	0.00
9	13.91	0.57
10	6.11	27.82
Total	100	100

b. MSE_R_hat:

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	0.04	0.24
2	0.05	0.25
3	0.08	0.24
4	0.02	0.02
5	0.01	0.04
6	0	0
7	0.01	0
8	0.02	0
9	0.04	0.01
10	0.02	0.27

c. RSE_R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	1.27	1.83
2	1.54	2.72
3	0.96	2.26
4	2.79	7.73
5	2.78	7.29
6	#NUM!	#NUM!
7	2.55	#NUM!
8	2.22	#NUM!
9	1.44	17.05
10	2.2	1.85

Interpretation:

For female students in both urban and rural areas of India, the reasons for staying in school versus those leading to dropping out or never enrolling show some notable differences. Among those who remain enrolled, the most significant reasons are Reason 3 (29.74%), followed by Reason 1 (15.57%) and Reason 2 (15.22%). Other factors such as Reason 9 (13.91%) and Reason 8 (6.55%) also contribute, while Reason 4 (4.87%) and Reason 5 (3.44%) play smaller roles.

For those who are not enrolled, the major reasons are Reason 10 (27.82%) and Reason 1 (27.01%), with Reason 3 (21.80%) also being important. Other factors like Reason 2 (18.50%) have some impact, but reasons such as Reason 4 (1.60%) and Reason 5 (2.71%) are less influential. Reasons like Reason 6, 7, and 8 do not affect non-enrollment at all.

This highlights that while a variety of factors influence why female students stay in school, only a few key reasons are major contributors to why they drop out or never enroll.

9. Urban + Rural, Male + Female for India:

a. R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	19.89	29.21
2	19.66	20.55
3	16.60	14.96
4	18.67	4.36
5	1.89	2.25
6	0.00	0.00
7	5.03	0.00
8	6.11	0.00
9	6.58	0.34
10	5.57	28.33
Total	100	100

b. MSE_R_hat:

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	0.03	0.22
2	0.05	0.23
3	0.03	0.12
4	0.03	0.04
5	0	0.03
6	0	0
7	0.01	0
8	0.01	0
9	0.01	0
10	0.01	0.25

c. RSE_R_hat (in %):

Reasons of discontinuation/dropping out/Never-enrollment	Enrolled (Yes)	Enrolled (No)
1	0.88	1.59
2	1.09	2.33
3	0.97	2.32
4	0.93	4.35
5	2.68	8.09
6	#NUM!	#NUM!
7	2.05	#NUM!
8	1.72	#NUM!
9	1.45	17.05
10	1.81	1.78

Interpretation:

For students across India, combining both urban and rural areas and including all genders, the reasons for staying in school versus those for dropping out or never enrolling are quite revealing. For those who are still enrolled, the leading reasons are Reason 4 (18.67%), Reason 1 (19.89%), and Reason 2 (19.66%), which reflect key factors that help students continue their education. Other reasons like Reason 3 (16.60%) and Reason 7 (5.03%) also contribute, though to a lesser extent.

On the other hand, for those who have dropped out or never enrolled, the most significant reasons are Reason 1 (29.21%) and Reason 10 (28.33%), showing that these issues are major barriers to education. Reason 2 (20.55%) is also an important factor, while Reason 3 (14.96%) and Reason 4 (4.36%) play smaller roles. Reasons such as 5, 6, 7, and 8 have a minimal impact on both enrollment and non-enrollment statuses.

This data underscores that while a range of factors influence why students remain in school, specific issues are more critical in determining why they drop out or never start their education.

Various Quintile Classes of UMPCE:

The Usual Monthly Per Capita Consumer Expenditure (UMPCE) is a measure used to understand how much money each person in a household spends on average each month. It is calculated by dividing the total monthly spending of a household by the number of people living in that household.

For each State or Union Territory (UT), and at the national level, we have data on UMPCE broken down into rural and urban areas. This data helps us understand spending patterns across different regions.

The UMPCE is divided into five groups, called quintiles, which categorize households based on their spending levels:

1. Quintile 1 represents the lowest spending group.

2. Quintile 2 is the next higher spending group.

3. Quintile 3 is the middle group.

4. Quintile 4 includes households with higher spending.

5. Quintile 5 represents the highest spending group.

The distribution of these quintiles is the same for both households and individuals within those households. For instance, if we look at West Bengal or at the national level, the spending distribution among these quintile groups will be consistent within the rural or urban areas.

In simple terms, these quintiles help us understand and compare how spending varies from the least to the most in different areas.

- Table for quintile classes of West Bengal:

Quintile classes of UMPCE	UMPCE (in Rs)			
	Rural		Urban	
	Lower limit	Upper limit	Lower limit	Upper limit
1	0.00	800.00	0.00	1150.00
2	800.00	1000.00	1150.00	1500.00
3	1000.00	1214.29	1500.00	2000.00
4	1214.29	1500.00	2000.00	3142.86
5	1500.00	-	3142.86	-

- Table for quintile classes of India:

Quintile classes of UMPCE	UMPCE (in Rs)			
	Rural		Urban	
	Lower limit	Upper limit	Lower limit	Upper limit
1	0.00	786.00	0.00	1200.00
2	786.00	1000.00	1200.00	1667.00
3	1000.00	1286.29	1667.00	2250.00
4	1287.29	1667.00	2250.00	3333.00
5	1667.00	-	3333.00	-

Graphical Representation of the Explanatory Variables:

To better understand the relationship between the explanatory variables and the outcome variables, we will create contingency tables for both West Bengal and All-India. These tables will help us visualize and analyze the data.

We will test two hypotheses:

1. Null Hypothesis (H0): There is no relationship between the explanatory variable and the outcome variable. In other words, any observed association is due to chance.
2. Alternative Hypothesis (H1): There is a relationship between the explanatory variable and the outcome variable. This means that the association observed in the data is significant and not due to random variation.

These hypotheses will guide our analysis, helping us determine whether the explanatory variables have a meaningful impact on the outcome variables.

We take the variables y_1, y_2, y_3, y_4, y_5 as:

y_1 = 1 if ever enrolled but discontinued/dropped out

= 0 otherwise

y_2 = 1 if ever enrolled and discontinued before completing Primary education (Grade IV)

= 0 otherwise

y_3 = 1 if ever enrolled and discontinued after completing Primary Education (Grade IV) but before completing Middle-Class Education (Grade VIII)

= 0 otherwise

y_4 = 1 if ever enrolled and discontinued after completing Middle-Class Education (Grade VIII) but before completing Secondary Education (Grade X)

= 0 otherwise

y_5 = 1 if ever enrolled and discontinued after completing Secondary Education (Grade X)

= 0 otherwise

To gain a clearer understanding of the relationships between our variables, we will create and analyze contingency tables, along with graphical presentations, for different sectors. We'll focus on the indicator variables across various explanatory variables for both West Bengal and All-India.

Here's what we'll do:

1. Contingency Tables and Graphs: We'll create contingency tables and their corresponding graphical presentations for each sector (Rural and Urban) to visually represent the data. This will be done separately for West Bengal and All-India, and for each explanatory variable.
2. Chi-Square Test of Independence: To statistically assess the relationships, we'll perform chi-square tests of independence for each explanatory variable. This will help us determine if there's a significant association between the explanatory variables and the outcome variables (y_1 and y_2) within the Rural and Urban sectors for both West Bengal and All-India.

By following these steps, we'll be able to see the data more clearly and understand the strength and significance of the relationships between the variables in different sectors and regions.

Hence, the different explanatory variables taken into consideration are:

1. Gender
2. House hold type
3. UMPCE_Quintile

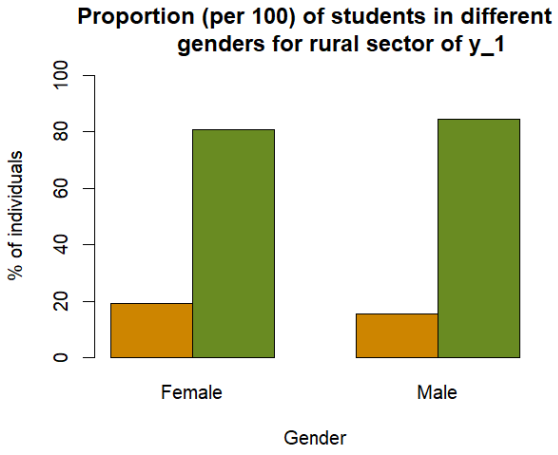
➢ West Bengal:

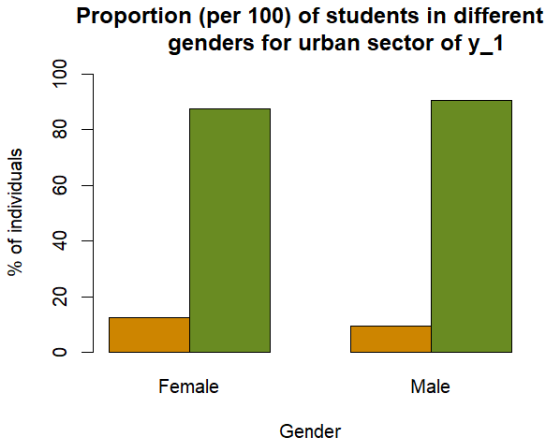
- Gender (Rural, Urban)

▪ The Percentage distribution table of Gender and y_1 (1 = Ever enrolled but discontinued):

y_1	Gender			
	Rural		Urban	
	Female	Male	Female	Male
0	19.27	15.61	12.34	9.51
1	80.72	84.39	87.66	90.48
Total	100	100	100	100

The Graphical Representation of Gender of y_1:





Pearson's chi-square test of Gender of y_1:

For Rural:

Pearson's Chi-squared test with Yates' continuity correction

data: prop_gender_wb_r_y1

X-squared = 0.24582, df = 1, p-value = 0.62

Since the p-value is much greater than the common significance level of 0.05, we fail to reject the null hypothesis. This suggests that any observed association between the explanatory variable and y_1 in the Rural sector is likely due to random chance rather than a true underlying relationship.

For Urban:

Pearson's Chi-squared test with Yates' continuity correction

data: prop_gender_wb_u_y1

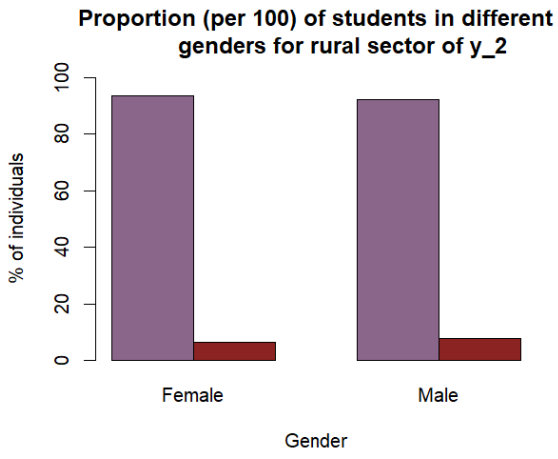
X-squared = 0.16997, df = 1, p-value = 0.6801

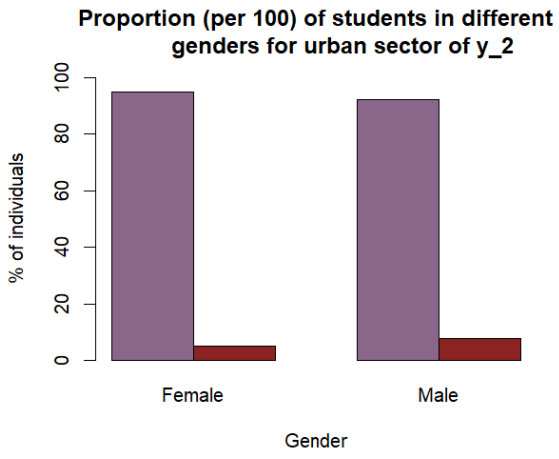
Since the p-value is significantly greater than the common significance level of 0.05, we fail to reject the null hypothesis. This implies that any observed association between the explanatory variable and y_1 in the Urban sector is likely due to random chance rather than a meaningful relationship.

- The Percentage distribution table of Gender and y_2 (1 = Ever enrolled but discontinued before Primary education):

y_2	Gender			
	Rural		Urban	
	Female	Male	Female	Male
0	93.58	92.29	94.84	92.33
1	6.42	7.71	5.16	7.67
Total	100	100	100	100

The Graphical Representation of Gender of y_2:





Pearson's chi-square test of Gender of y_2:

For Rural:

Pearson's Chi-squared test with Yates' continuity correction

data: prop_gender_wb_r_y1

X-squared = 0.24582, df = 1, p-value = 0.62

Since the p-value is much greater than the conventional threshold of 0.05, we do not have enough evidence to reject the null hypothesis. This indicates that any observed relationship between the explanatory variable and y_1 in the Rural sector is likely due to chance rather than a true association.

For Urban:

Pearson's Chi-squared test with Yates' continuity correction

data: prop_gender_wb_u_y1

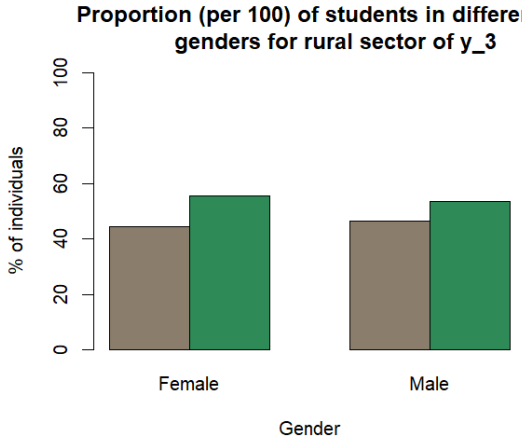
X-squared = 0.16997, df = 1, p-value = 0.6801

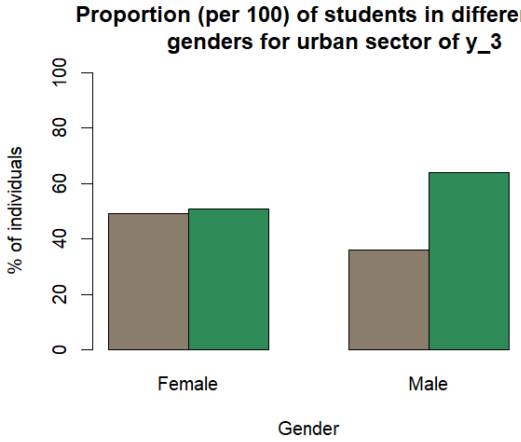
Since the p-value is considerably higher than the standard significance level of 0.05, we do not reject the null hypothesis. This suggests that any observed association between the explanatory variable and y_1 in the Urban sector is likely due to random variation rather than a true underlying relationship.

- The Percentage distribution table of Gender and y_3 (1 = Ever enrolled, completed primary, but discontinued before completing Middle-class education):

y_3	Gender			
	Rural		Urban	
	Female	Male	Female	Male
0	44.46	46.42	49.03	36.05
1	55.54	53.57	50.97	63.95
Total	100	100	100	100

The Graphical Representation of Gender of y_3:





Pearson's chi-square test of Gender of y_3

For Rural:

Pearson's Chi-squared test with Yates' continuity correction

data: prop_gender_wb_r_y3

X-squared = 0.018895, df = 1, p-value = 0.8907

Since the p-value is much greater than the conventional threshold of 0.05, we fail to reject the null hypothesis. This means that any observed relationship between the explanatory variable and y_3 in the Rural sector is likely due to chance and does not indicate a meaningful association.

For Urban:

Pearson's Chi-squared test with Yates' continuity correction

data: prop_gender_wb_u_y3

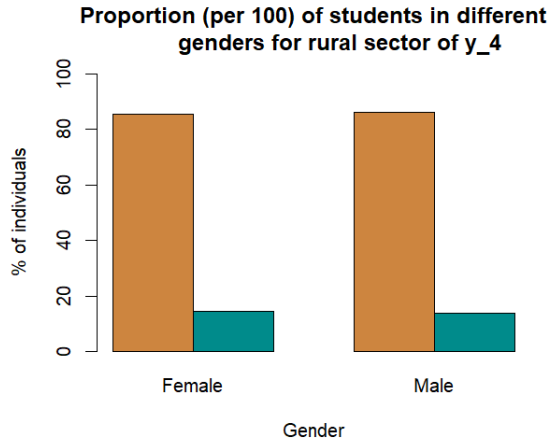
X-squared = 2.9386, df = 1, p-value = 0.08649

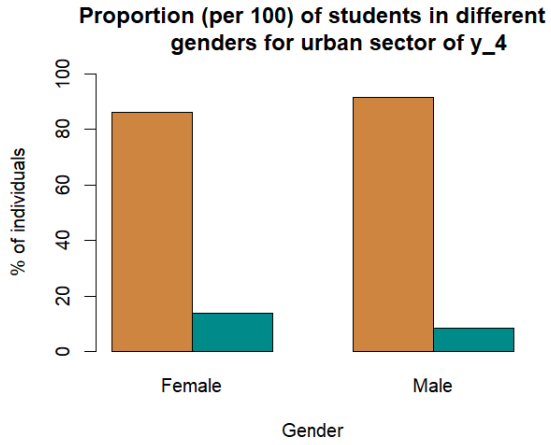
Although the p-value is closer to the common significance level of 0.05, it is still above this threshold. Therefore, we fail to reject the null hypothesis. This suggests that the observed association between the explanatory variable and y_3 in the Urban sector is not strong enough to be considered statistically significant, although it is closer to indicating a potential trend compared to the results in other tests.

- The Percentage distribution table of Gender and y_4 (1 = Ever enrolled, completed Middle class, but discontinued before completing Secondary Education):

y_4	Gender			
	Rural		Urban	
	Female	Male	Female	Male
0	85.56	86.14	86.02	91.63
1	14.43	13.86	13.98	8.37
Total	100	100	100	100

The Graphical Representation of Gender of y_4:





Pearson's chi-square test of Gender of y_4

For Rural:

Pearson's Chi-squared test with Yates' continuity correction

data: prop_gender_wb_r_y4

X-squared = 2.2207e-29, df = 1, p-value = 1

The test statistic is extremely close to zero, and the p-value of 1 confirms that any observed association is purely due to chance. Therefore, we conclusively fail to reject the null hypothesis, indicating no meaningful relationship between the explanatory variable and y_4 in the Rural sector.

For Urban:

Pearson's Chi-squared test with Yates' continuity correction

data: prop_gender_wb_u_y4

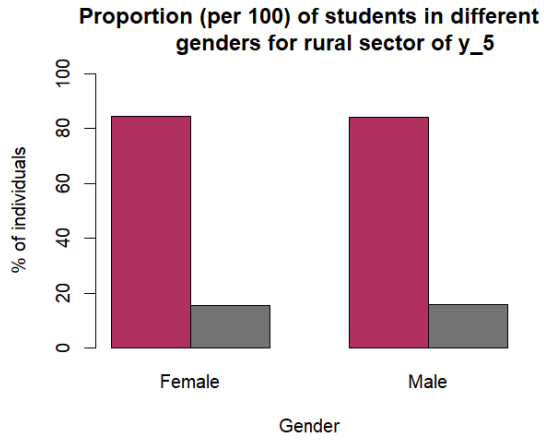
X-squared = 1.0688, df = 1, p-value = 0.3012

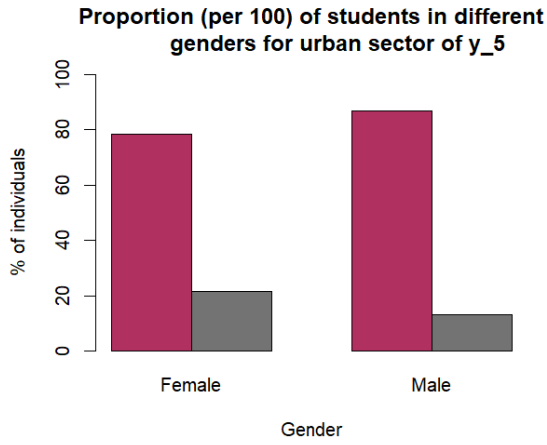
Since the p-value is much greater than the conventional significance level of 0.05, we fail to reject the null hypothesis. This means that there is no significant association between the explanatory variable and y_4 in the Urban sector, and any observed relationship is likely due to random chance rather than a true underlying connection.

- The Percentage distribution table of Gender and y_5 (1 = Ever enrolled, completed Secondary education but discontinued after Secondary):

y_5	Gender			
	Rural		Urban	
	Female	Male	Female	Male
0	84.40	84	78.50	86.98
1	15.60	16	21.50	13.02
Total	100	100	100	100

The Graphical Representation of Gender of y_5:





Pearson's chi-square test of Gender of y_5

For Rural:

Pearson's Chi-squared test with Yates' continuity correction

data: prop_gender_wb_r_y5

X-squared = 1.4582e-29, df = 1, p-value = 1

The test statistic is virtually zero, and the p-value of 1 confirms that any observed relationship is purely due to random chance. Therefore, we conclusively fail to reject the null hypothesis, suggesting that there is no meaningful connection between the explanatory variable and y_5 in the Rural sector.

For Urban:

Pearson's Chi-squared test with Yates' continuity correction

data: prop_gender_wb_u_y5

X-squared = 1.9596, df = 1, p-value = 0.1616

Since the p-value is well above the standard significance level of 0.05, we fail to reject the null hypothesis. This suggests that there is no statistically significant association between the explanatory variable and y_5 in the Urban sector, and any observed relationship is likely due to random variation rather than a true underlying effect.

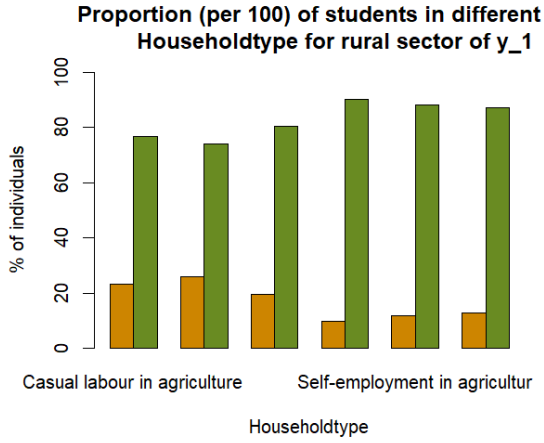
- Household type (Rural, Urban)

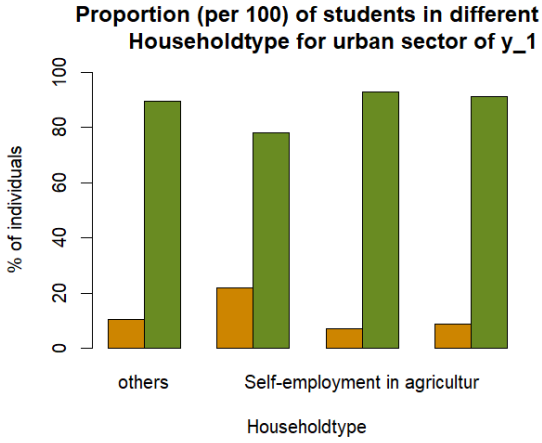
  - The Percentage distribution table of Household type and y_1 (1 = Ever enrolled but discontinued):

y_1	Household type					
	Rural					
	Casual labour in agriculture	Casual labour in non-agriculture	Others	Regular wage/Salary earning	Self-employment in agriculture	Self-employment in non-agriculture
0	23.17	26.03	19.60	9.70	11.90	12.68
1	76.83	73.97	80.39	90.30	88.09	87.32
Total	100	100	100	100	100	100

y_1	Household type			
	Urban			
	Others	Regular wage/Salary earning	Self-employment in agriculture	Self- employment in non-agriculture
0	10.42	21.81	7.11	8.89
1	89.58	78.19	92.89	91.11
Total	100	100	100	100

The Graphical Representation of Household type of y_1:





Pearson's chi-square test of Household type of y_1:

For Rural:

Pearson's Chi-squared test

data: prop_Householdtype_wb_r_y1

X-squared = 15.762, df = 5, p-value = 0.007557

Since the p-value is significantly below the conventional significance level of 0.05, we reject the null hypothesis. This indicates that there is a statistically significant association between the household type variable and y_1 in the Rural sector. The observed relationship is unlikely to be due to random chance, suggesting a meaningful connection between these variables.

For Urban:

Pearson's Chi-squared test

data: prop_Householdtype_wb_u_y1

X-squared = 12.475, df = 3, p-value = 0.005921

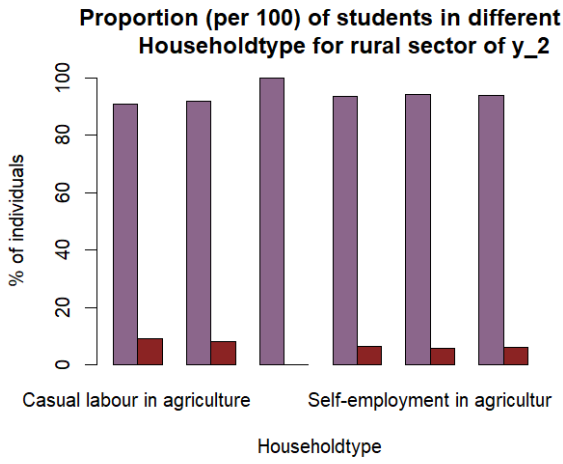
Since the p-value is much lower than the standard significance level of 0.05, we reject the null hypothesis. This indicates a statistically significant association between the Householdtype variable and y_1 in the Urban sector. The observed relationship is unlikely to be due to random chance, suggesting a meaningful connection between these variables in the Urban context.

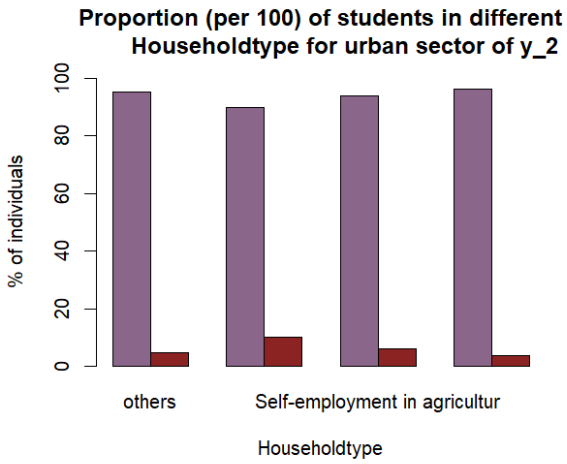
- The Percentage distribution table of Household type and y_2 (1 = Ever enrolled but discontinued before Primary education):

y_2	Household type					
	Rural					
	Casual labour in agriculture	Casual labour in non-agriculture	Others	Regular wage/Salary earning	Self-employment in agriculture	Self-employment in non-agriculture
0	90.09	92.05	100	93.70	94.09	94.09
1	9.02	7.94	0	6.29	5.90	5.98
Total	100	100	100	100	100	100

y_2	Household type			
	Urban			
	Others	Regular wage/Salary earning	Self-employment in agriculture	Self- employment in non-agriculture
0	95.24	89.96	94.06	96.43
1	4.76	10.04	5.94	3.57
Total	100	100	100	100

The Graphical Representation of Household type of y_2:





Pearson's chi-square test of Household type of y_2:

For Rural:

Pearson's Chi-squared test

data: prop_Householdtype_wb_r_y2

X-squared = 8.8661, df = 5, p-value = 0.1145

Since the p-value is greater than the conventional significance level of 0.05, we fail to reject the null hypothesis. This suggests that there is no statistically significant association between the Householdtype variable and y_2 in the Rural sector. The observed relationship is likely due to chance rather than indicating a meaningful connection between these variables.

For Urban:

Pearson's Chi-squared test

data: prop_Householdtype_wb_u_y2

X-squared = 4.1618, df = 3, p-value = 0.2445

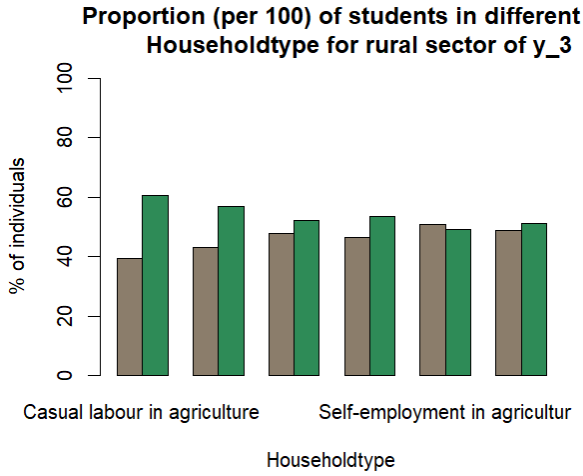
Since the p-value exceeds the conventional significance level of 0.05, we fail to reject the null hypothesis. This indicates that there is no statistically significant association between the Householdtype variable and y_2 in the Urban sector. Any observed relationship is likely due to random chance rather than reflecting a meaningful connection between these variables.

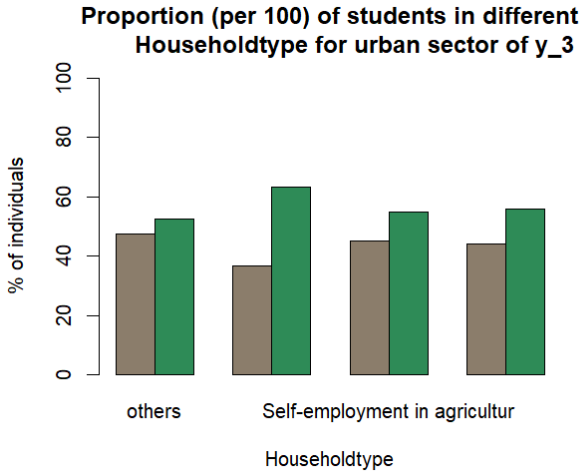
- The Percentage distribution table of House hold type and y_3 (1 = Ever enrolled, completed primary, but discontinued before completing Middle-class education):

y_3	Household type					
	Rural					
	Casual labour in agriculture	Casual labour in non-agriculture	Others	Regular wage/Salary earning	Self-employment in agriculture	Self-employment in non-agriculture
0	39.35	42.99	47.83	46.46	50.93	48.84
1	60.65	57.00	52.17	53.54	49.07	51.16
Total	100	100	100	100	100	100

y_3	Household type			
	Urban			
	Others	Regular wage/Salary earning	Self-employment in agriculture	Self- employment in non-agriculture
0	47.62	36.68	45.13	44.20
1	52.38	63.32	54.87	55.80
Total	100	100	100	100

The Graphical Representation of Household type of y_3:





Pearson's chi-square test of Household type of y_3:

For Rural:

Pearson's Chi-squared test

data: prop_Householdtype_wb_r_y3

X-squared = 3.5898, df = 5, p-value = 0.6098

Since the p-value is much higher than the conventional significance level of 0.05, we fail to reject the null hypothesis. This indicates that there is no statistically significant association between the Householdtype variable and y_3 in the Rural sector. The observed relationship is likely due to random chance rather than a meaningful connection between these variables.

For Urban:

Pearson's Chi-squared test

data: prop_Householdtype_wb_u_y3

X-squared = 2.71, df = 3, p-value = 0.4385

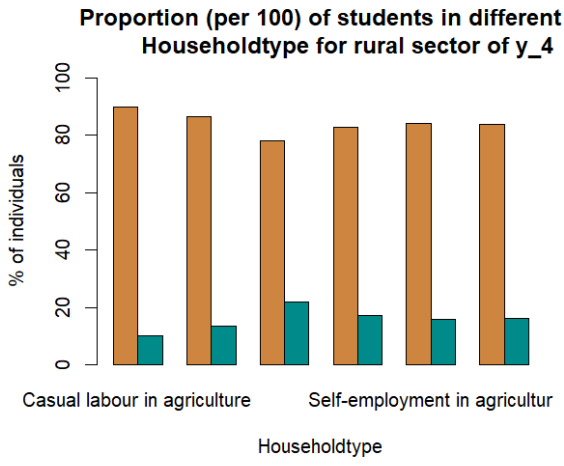
Since the p-value is significantly higher than the conventional significance level of 0.05, we fail to reject the null hypothesis. This indicates that there is no statistically significant association between the household type variable and y_3 in the Urban sector. The observed relationship is likely due to random variation rather than reflecting a true underlying connection.

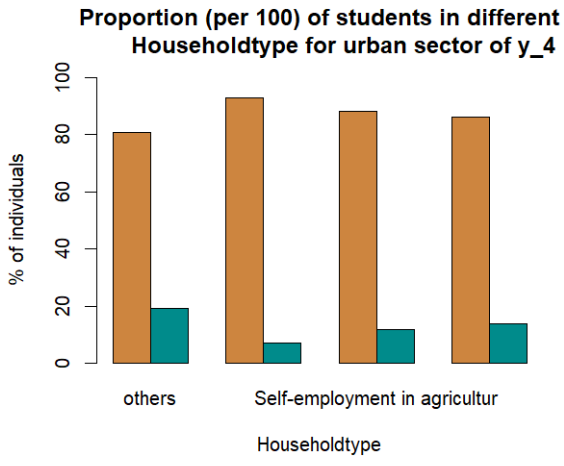
- The Percentage distribution table of Household type and y_4 (1 = Ever enrolled, completed Middle class, but discontinued before completing Secondary Education)

y_4	Household type					
	Rural					
	Casual labour in agriculture	Casual labour in non-agriculture	Others	Regular wage/Salary earning	Self-employment in agriculture	Self-employment in non-agriculture
0	89.97	86.45	78.26	82.68	84.16	83.72
1	10.02	13.55	21.74	17.32	15.84	16.28
Total	100	100	100	100	100	100

y_4	Household type			
	Urban			
	Others	Regular wage/Salary earning	Self-employment in agriculture	Self- employment in non-agriculture
0	80.95	93.01	88.12	86.16
1	19.05	6.99	11.88	13.84
Total	100	100	100	100

The Graphical Representation of Household type of y_4:





Pearson's chi-square test of household type of y_4:

For Rural:

Pearson's Chi-squared test

data: prop_Householdtype_wb_r_y4

X-squared = 5.7321, df = 5, p-value = 0.3332

Since the p-value is greater than the standard significance level of 0.05, we fail to reject the null hypothesis. This suggests that there is no statistically significant association between the Householdtype variable and y_4 in the Rural sector. Any observed relationship is likely due to random variation rather than indicating a meaningful connection between these variables.

For Urban:

Pearson's Chi-squared test

data: prop_Householdtype_wb_u_y4

X-squared = 6.6303, df = 3, p-value = 0.08466

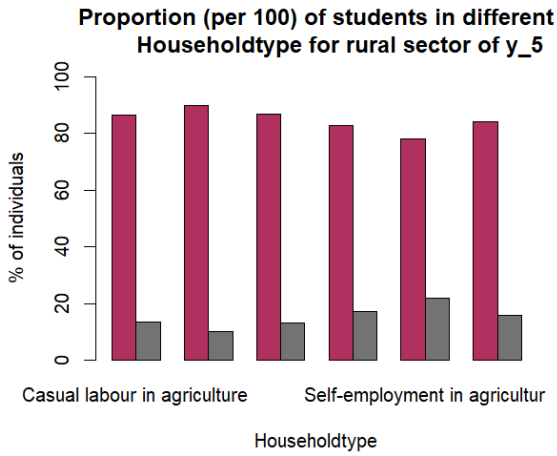
Since the p-value is above the conventional significance level of 0.05 but below the more lenient threshold of 0.10, the result is suggestive but not statistically significant. This indicates a potential association between the Householdtype variable and y_4 in the Urban sector, though it is not strong enough to definitively conclude that it is a meaningful relationship. The observed association could be due to random variation, and further investigation may be warranted.

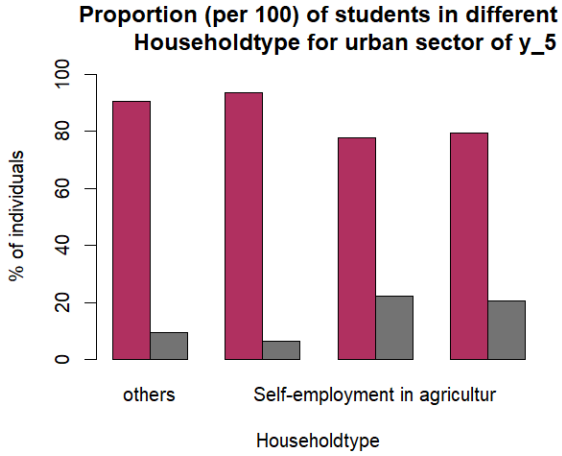
- The Percentage distribution table of House hold type and y_5 (1 = Ever enrolled, completed Secondary education but discontinued after Secondary):

y_5	Household type					
	Rural					
	Casual labour in agriculture	Casual labour in non-agriculture	Others	Regular wage/Salary earning	Self-employment in agriculture	Self-employment in non-agriculture
0	86.47	89.72	86.95	82.68	78.26	84.05
1	13.53	10.28	13.04	17.32	21.74	15.95
Total	100	100	100	100	100	100

y_5	Household type			
	Urban			
	Others	Regular wage/Salary earning	Self-employment in agriculture	Self- employment in non-agriculture
0	90.48	93.45	77.91	79.47
1	9.52	6.55	22.09	20.53
Total	100	100	100	100

The Graphical Representation of Household type of y_5:





Pearson's chi-square test of household type of y_5:

For Rural:

Pearson's Chi-squared test

data: prop_Householdtype_wb_r_y5

X-squared = 6.1218, df = 5, p-value = 0.2945

Since the p-value is higher than the standard significance level of 0.05, we fail to reject the null hypothesis. This suggests that there is no statistically significant association between the

Householdtype variable and y_5 in the Rural sector. The observed relationship is likely due to random chance rather than indicating a meaningful connection between these variables.

For Urban:

Pearson's Chi-squared test

data: prop_Householdtype_wb_u_y5

X-squared = 14.526, df = 3, p-value = 0.00227

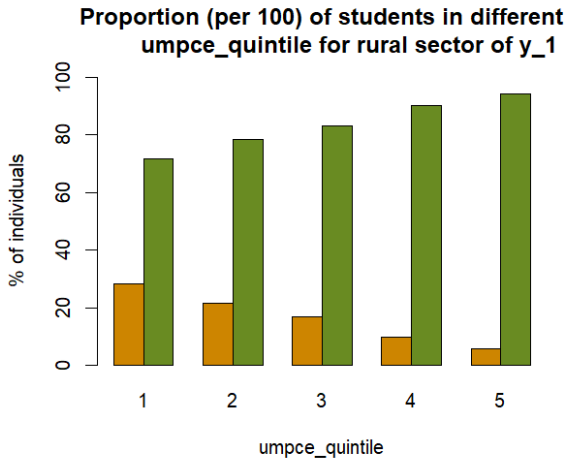
Since the p-value is significantly below the standard significance level of 0.05, we reject the null hypothesis. This indicates a statistically significant association between the Householdtype variable and y_5 in the Urban sector. The observed relationship is unlikely to be due to random chance, suggesting a meaningful connection between these variables in the Urban context.

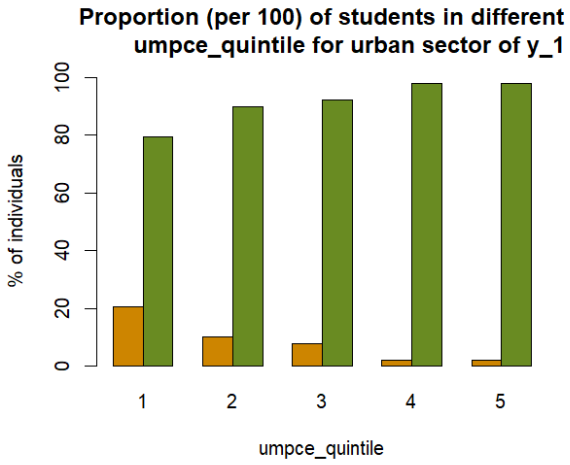
- UMPCE_Quintile (Rural, Urban)

  - The Percentage distribution table of UMPCE_Quintile and y_1 (1 = Ever enrolled but discontinued):

y_1	UMPCE									
	Rural					Urban				
	1	2	3	4	5	1	2	3	4	5
0	28.14	21.68	17	9.70	5.87	20.58	10.24	7.84	1.95	2.01
1	71.86	78.32	83	90.30	94.30	79.42	89.76	92.16	98.05	97.99
Total	100	100	100	100	100	100	100	100	100	100

The Graphical Representation of UMPCE_Quintile of y_1:





Pearson's chi-square test of UMPCE_Quintile of y_1:

For Rural:

Pearson's Chi-squared test

data: prop_umpce_quintile_wb_r_y1

X-squared = 23.383, df = 4, p-value = 0.0001062

Since the p-value is well below the standard significance level of 0.05, we reject the null hypothesis. This indicates a statistically significant association between the umpce_quintile variable and y_1 in the Rural sector. The observed relationship is unlikely to be due to random chance, suggesting a meaningful connection between these variables in the Rural context.

For Urban:

Pearson's Chi-squared test

data: prop_umpce_quintile_wb_u_y1

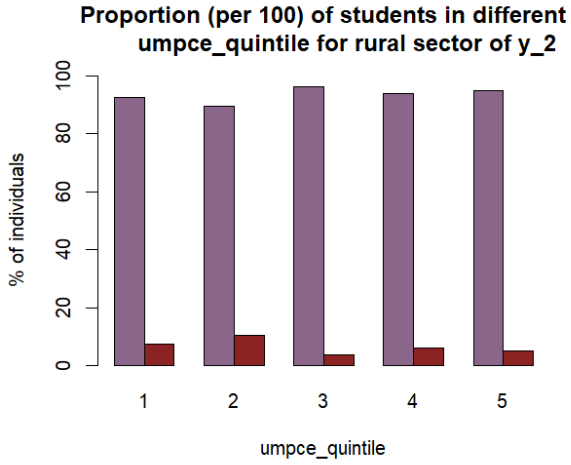
X-squared = 30.052, df = 4, p-value = 4.776e-06

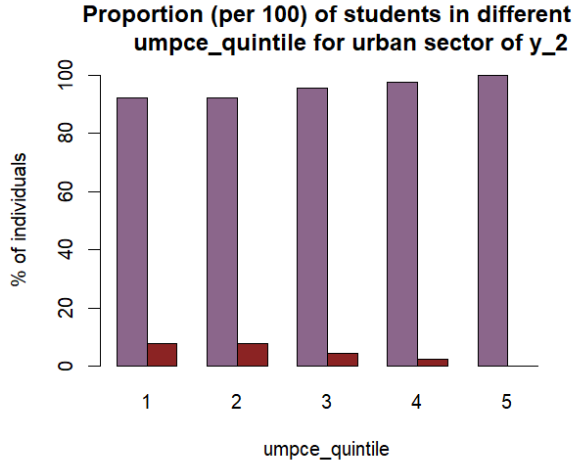
Since the p-value is far below the conventional significance level of 0.05, we reject the null hypothesis. This indicates a statistically significant association between the umpce_quintile variable and y_1 in the Urban sector. The observed relationship is highly unlikely to be due to random chance, suggesting a strong and meaningful connection between these variables in the Urban context.

- The Percentage distribution table of UMPCE_Quintile and y_2 (1 = Ever enrolled but discontinued before Primary education):

y_2	UMPCE									
	Rural					Urban				
	1	2	3	4	5	1	2	3	4	5
0	92.48	89.37	96.30	93.82	95.05	92.16	92.36	95.73	97.56	100
1	7.52	10.63	3.70	6.18	4.95	7.84	7.64	4.27	2.43	0
Total	100	100	100	100	100	100	100	100	100	100

The Graphical Representation of UMPCE_Quintile of y_2:





Pearson's chi-square test of UMPCE_Quintile of y_2:

For Rural:

Pearson's Chi-squared test

data: prop_umpce_quintile_wb_r_y2

X-squared = 4.5968, df = 4, p-value = 0.3312

Since the p-value is greater than the standard significance level of 0.05, we fail to reject the null hypothesis. This suggests that there is no statistically significant association between the umpce_quintile variable and y_2 in the Rural sector. The observed relationship is likely due to random chance rather than indicating a meaningful connection between these variables.

For Urban:

Pearson's Chi-squared test

data: prop_umpce_quintile_wb_u_y2

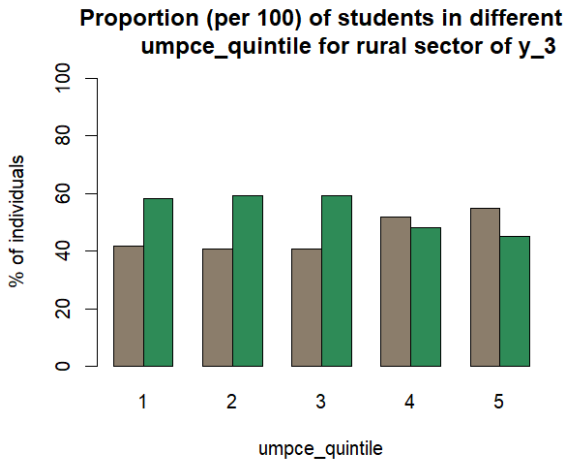
X-squared = 10.739, df = 4, p-value = 0.02966

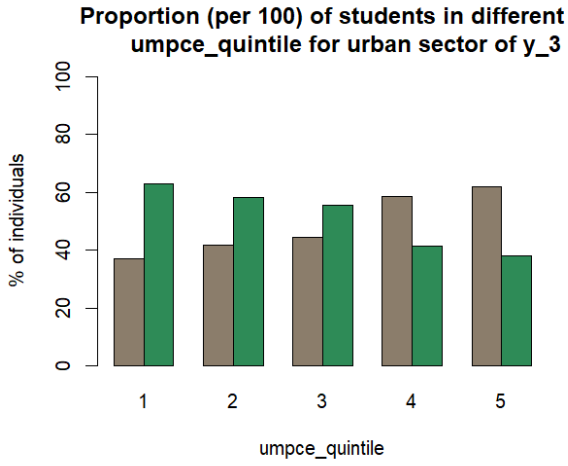
Since the p-value is below the conventional significance level of 0.05, we reject the null hypothesis. This indicates a statistically significant association between the umpce_quintile variable and y_2 in the Urban sector. The observed relationship is unlikely to be due to random chance, suggesting a meaningful connection between these variables in the Urban context.

- The Percentage distribution table of UMPCE_Quintile and y_3 (1 = Ever enrolled, completed primary, but discontinued before completing Middle-class education)

y_3	UMPCE									
	Rural					Urban				
	1	2	3	4	5	1	2	3	4	5
0	41.83	40.60	40.74	52	54.95	36.99	41.86	44.51	58.54	62.07
1	58.17	59.40	59.26	48	45.05	63.00	58.14	55.49	41.46	37.93
Total	100	100	100	100	100	100	100	100	100	100

The Graphical Representation of UMPCE_Quintile of y_3:





Pearson's chi-square test of UMPCE_Quintile of y_3:

For Rural:

Pearson's Chi-squared test

data: prop_umpce_quintile_wb_r_y3

X-squared = 7.6645, df = 4, p-value = 0.1047

The X-squared value of 7.6645 and the p-value of 0.1047, which is greater than the typical significance level of 0.05, suggest that the observed frequencies are not significantly different from the expected frequencies under the null hypothesis of independence. This implies that the variables in the rural data do not have a significant relationship with each other.

For Urban:

Pearson's Chi-squared test

data: prop_umpce_quintile_wb_u_y3

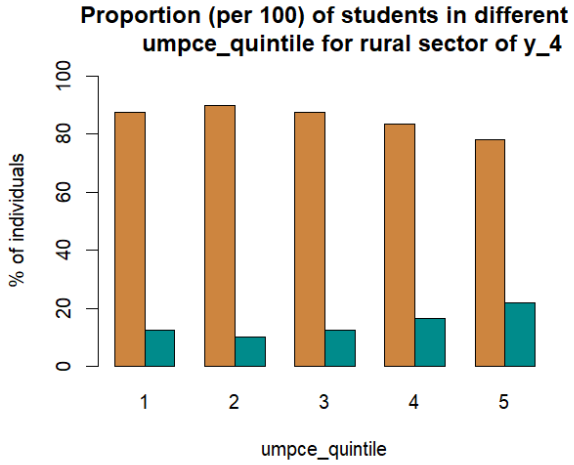
X-squared = 19.086, df = 4, p-value = 0.000756

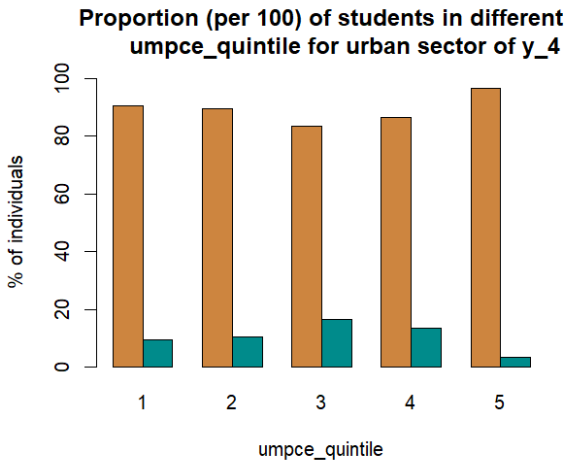
The X-squared value of 19.086 and the very low p-value of 0.000756 (less than the typical significance level of 0.05) indicate that the observed frequencies are significantly different from the expected frequencies under the null hypothesis of independence. This suggests that there is a strong relationship between the variables in the urban data, meaning that they tend to move together or influence each other in some way. In other words, the data shows a significant pattern or connection between the variables, which warrants further exploration.

- The Percentage distribution table of UMPCE_Quintile and y_4 (1 = Ever enrolled, completed Middle class, but discontinued before completing Secondary Education):

y_4	UMPCE									
	Rural					Urban				
	1	2	3	4	5	1	2	3	4	5
0	87.58	89.92	87.50	83.64	77.93	90.60	89.37	83.54	86.59	96.55
1	12.42	10.08	12.50	16.36	22.07	9.40	10.63	16.46	13.41	3.44
Total	100	100	100	100	100	100	100	100	100	100

The Graphical Representation of UMPCE_Quintile of y_4:





Pearson's chi-square test of UMPCE_Quintile of y_4:

For Rural:

Pearson's Chi-squared test

data: prop_umpce_quintile_wb_r_y4

X-squared = 7.0621, df = 4, p-value = 0.1326

This p-value is greater than the conventional significance level of 0.05, indicating that there is not enough evidence to reject the null hypothesis. Therefore, we conclude that there is no statistically significant association between the categories within 'prop_umpce_quintile_wb_r_y4' for the rural population. In other words, the distribution of the variable across different groups does not show significant deviations from what would be expected by chance.

For Urban:

Pearson's Chi-squared test

data: prop_umpce_quintile_wb_u_y4

X-squared = 9.9496, df = 4, p-value = 0.04129

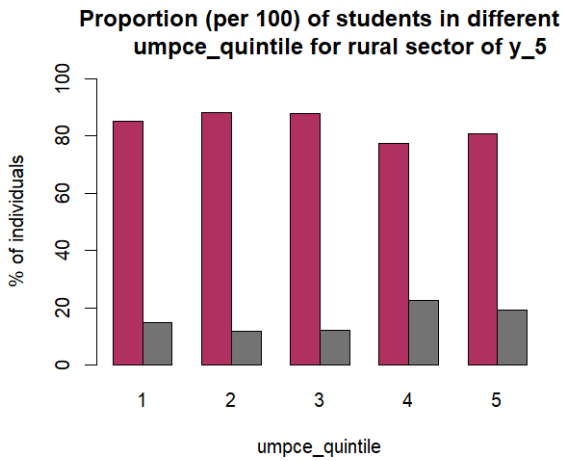
This p-value is less than the conventional significance level of 0.05, indicating that there is sufficient evidence to reject the null hypothesis. Thus, we conclude that there is a statistically significant association between the categories within 'prop_umpce_quintile_wb_u_y4' for the urban population. This means that the distribution of the variable across different groups

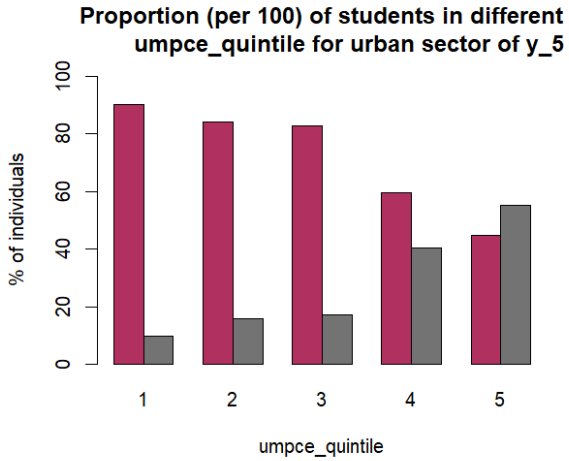
shows significant deviations from what would be expected by chance, suggesting that some underlying factors may be influencing the observed distribution in the urban context.

- The Percentage distribution table of UMPCE_Quintile and y_5 (1 = Ever enrolled, completed Secondary education but discontinued after Secondary):

y_5	UMPCE									
	Rural					Urban				
	1	2	3	4	5	1	2	3	4	5
0	85.29	88.28	87.96	77.45	80.63	90.28	84.05	82.93	59.76	44.83
1	14.70	11.72	11.72	22.55	19.37	9.72	15.95	17.07	40.24	55.17
Total	100	100	100	100	100	100	100	100	100	100

The Graphical Representation of UMPCE_Quintile of y_5:





Pearson's chi-square test of UMPCE_Quintile of y_5:

For Rural:

Pearson's Chi-squared test

data: prop_umpce_quintile_wb_r_y5

X-squared = 6.6634, df = 4, p-value = 0.1548

Since the p-value is greater than the conventional significance level of 0.05, we fail to reject the null hypothesis. This indicates that there is no statistically significant association between the categories within 'prop_umpce_quintile_wb_r_y5' for the rural population. In other

words, the observed distribution of this variable does not significantly differ from what would be expected by chance, suggesting no notable differences among the groups in this context.

For Urban:

Pearson's Chi-squared test

data: prop_umpce_quintile_wb_u_y5

X-squared = 74.338, df = 4, p-value = 2.75e-15

This extremely small p-value is well below the conventional significance level of 0.05, providing strong evidence to reject the null hypothesis. Therefore, we conclude that there is a highly significant association between the categories within 'prop_umpce_quintile_wb_u_y5' for the urban population. This means that the distribution of the variable across different groups shows substantial deviations from what would be expected by chance, indicating the presence of strong underlying factors influencing the observed distribution in the urban context.

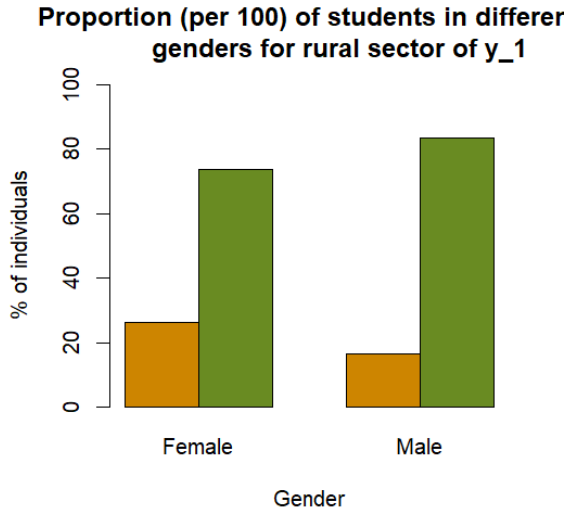
➢ India:

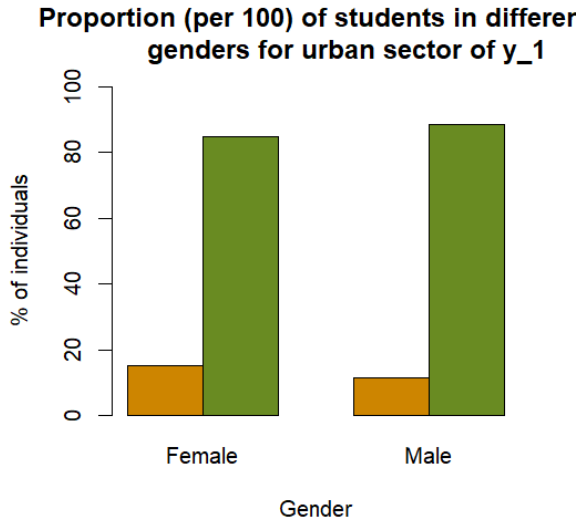
- Gender (Rural, Urban)

- The Percentage distribution table of Gender and y_1 (1 = Ever enrolled but discontinued):

y_1	Gender			
	Rural		Urban	
	Female	Male	Female	Male
0	26.16	16.53	15.07	11.40
1	73.84	83.47	84.93	88.60
Total	100	100	100	100

The Graphical Representation of Gender of y_1:





Pearson's chi-square test of Gender of y_1:

For Rural:

Pearson's Chi-squared test with Yates' continuity correction

data: prop_gender_r_y1

X-squared = 2.2195, df = 1, p-value = 0.1363

Since the p-value is greater than the conventional significance level of 0.05, we fail to reject the null hypothesis. This suggests that there is no statistically significant association between gender categories within 'prop_gender_r_y1' for the rural population. In other words, the observed distribution of gender in this variable does not differ significantly from what would be expected by chance, indicating no notable gender-related differences in this context.

For Urban:

Pearson's Chi-squared test with Yates' continuity correction

data: prop_gender_u_y1

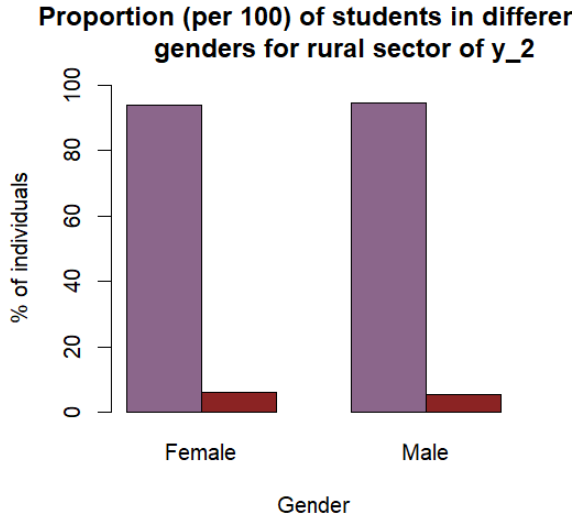
X-squared = 0.31133, df = 1, p-value = 0.5769

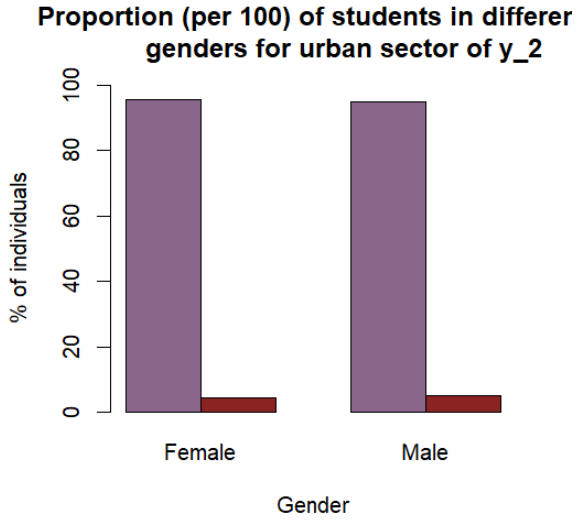
This p-value is significantly greater than the conventional significance level of 0.05, indicating that there is no evidence to reject the null hypothesis. Therefore, we conclude that there is no statistically significant association between gender categories within 'prop_gender_u_y1' for the urban population. The observed distribution of gender in this variable does not differ significantly from what would be expected by chance, suggesting no notable gender-related differences in the urban context.

- The Percentage distribution table of Gender and y_2 (1 = Ever enrolled but discontinued before Primary education):

y_2	Gender			
	Rural		Urban	
	Female	Male	Female	Male
0	93.94	94.45	95.63	94.90
1	6.06	5.55	4.37	5.10
Total	100	100	100	100

The Graphical Representation of Gender of y_2:





Pearson's chi-square test of Gender of y_2:

For Rural:

Pearson's Chi-squared test with Yates' continuity correction

data: prop_gender_r_y2

X-squared = 1.3595e-31, df = 1, p-value = 1

A p-value of 1 indicates that the observed distribution exactly matches the expected distribution under the null hypothesis. Consequently, there is no evidence to suggest any association between gender categories within 'prop_gender_r_y2' for the rural population. This implies that the distribution of this variable across gender groups is perfectly aligned with what would be expected by chance, showing no significant gender-related differences in this context.

For Urban:

Pearson's Chi-squared test with Yates' continuity correction

data: prop_gender_u_y2

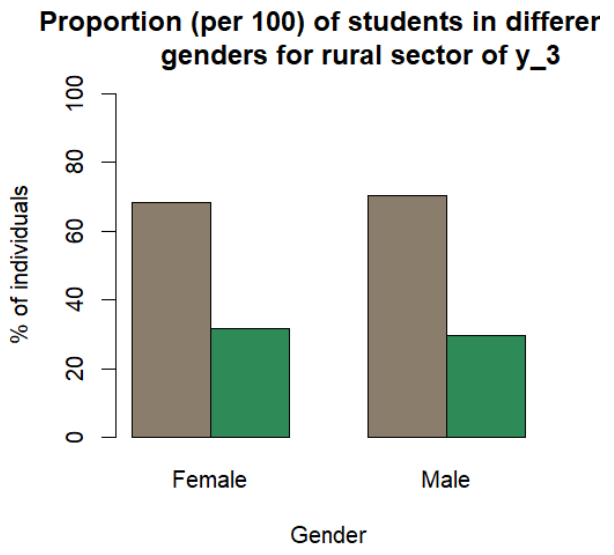
X-squared = 8.952e-30, df = 1, p-value = 1

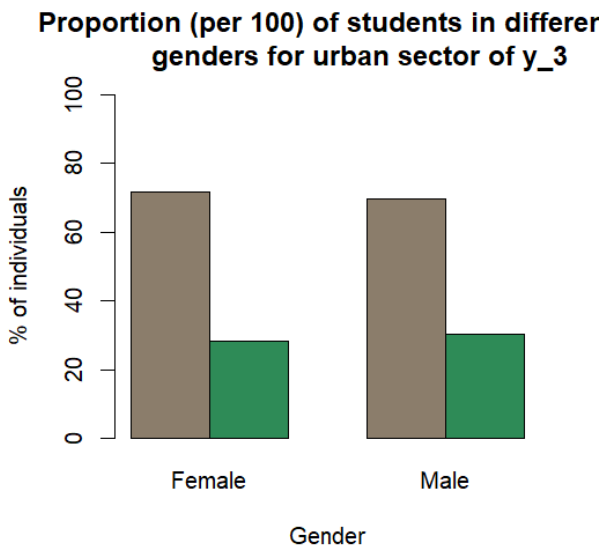
This p-value of 1 indicates that the observed distribution exactly matches the expected distribution under the null hypothesis. Therefore, there is no evidence to suggest any association between gender categories within 'prop_gender_u_y2' for the urban population. This means that the distribution of this variable across gender groups is perfectly consistent with what would be expected by chance, indicating no significant gender-related differences in the urban context.

- The Percentage distribution table of Gender and y_3 (1 = Ever enrolled, completed primary, but discontinued before completing Middle-class education):

y_3	Gender			
	Rural		Urban	
	Female	Male	Female	Male
0	68.31	70.19	71.70	69.76
1	31.69	29.81	28.30	30.24
Total	100	100	100	100

The Graphical Representation of Gender of y_3:





Pearson's chi-square test of Gender of y_3

For Rural:

Pearson's Chi-squared test with Yates' continuity correction

data: prop_gender_r_y3

X-squared = 0.018297, df = 1, p-value = 0.8924

Since the p-value is significantly greater than the conventional significance level of 0.05, we fail to reject the null hypothesis. This suggests that there is no statistically significant association between gender categories within 'prop_gender_r_y3' for the rural population. The observed distribution of gender in this variable does not differ significantly from what would be expected by chance, indicating no notable gender-related differences in this context.

For Urban:

Pearson's Chi-squared test with Yates' continuity correction

data: prop_gender_u_y3

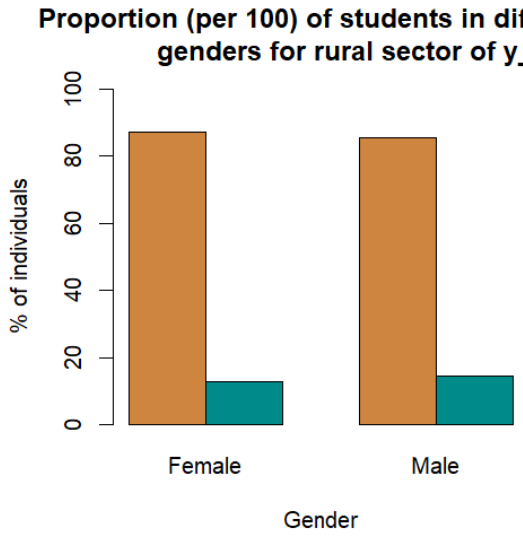
X-squared = 0.021215, df = 1, p-value = 0.8842

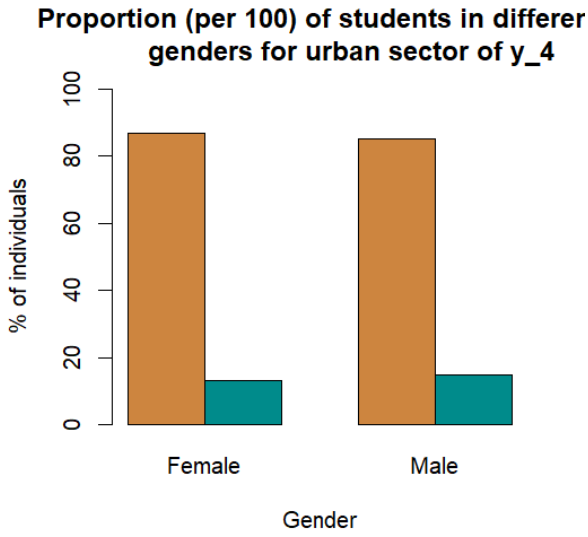
Since the p-value is much greater than the conventional significance level of 0.05, we fail to reject the null hypothesis. This indicates that there is no statistically significant association between gender categories within 'prop_gender_u_y3' for the urban population. The observed distribution of gender in this variable does not significantly differ from what would be expected by chance, suggesting no notable gender-related differences in this urban context.

- The Percentage distribution table of Gender and y_4 (1 = Ever enrolled, completed Middle class, but discontinued before completing Secondary Education):

y_4	Gender			
	Rural		Urban	
	Female	Male	Female	Male
0	87.26	85.44	86.69	85
1	12.74	14.56	13.31	15
Total	100	100	100	100

The Graphical Representation of Gender of y_4:





Pearson's chi-square test of Gender of y_4

For Rural:

Pearson's Chi-squared test with Yates' continuity correction

data: prop_gender_r_y4

X-squared = 0.028496, df = 1, p-value = 0.8659

Given that the p-value is significantly greater than the conventional significance level of 0.05, there is no evidence to reject the null hypothesis. This indicates that there is no statistically significant association between gender categories within 'prop_gender_r_y4' for the rural population. The distribution of gender in this variable aligns closely with what would be expected by chance, suggesting no meaningful gender-related differences in this context.

For Urban:

Pearson's Chi-squared test with Yates' continuity correction

data: prop_gender_u_y4

X-squared = 0.019365, df = 1, p-value = 0.8893

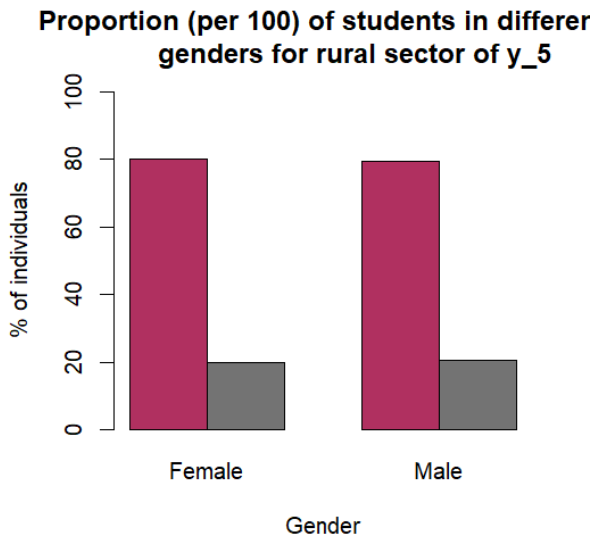
This p-value is significantly greater than the conventional significance level of 0.05, indicating that there is no statistically significant association between gender categories within

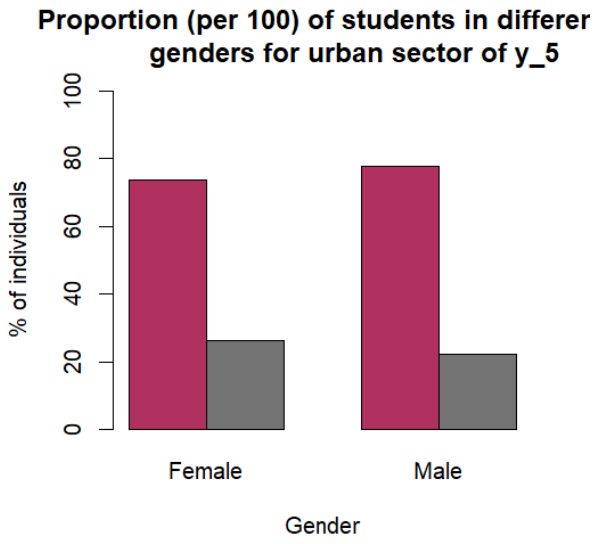
'prop_gender_u_y4' for the urban population. Therefore, the observed distribution of gender in this variable does not differ significantly from what would be expected by chance, suggesting no notable gender-related differences in this urban context.

- The Percentage distribution table of Gender and y_5 (1 = Ever enrolled, completed Secondary education but discontinued after Secondary):

y_5	Gender			
	Rural		Urban	
	Female	Male	Female	Male
0	80.14	79.58	73.80	77.66
1	19.86	20.42	26.20	22.34
Total	100	100	100	100

The Graphical Representation of Gender of y_5:





Pearson's chi-square test of Gender of y_5

For Rural:

Pearson's Chi-squared test with Yates' continuity correction

data: prop_gender_r_y5

X-squared = 6.2668e-31, df = 1, p-value = 1

This p-value of 1 indicates that the observed distribution of gender categories perfectly matches the expected distribution under the null hypothesis. Consequently, there is no evidence to suggest any association between gender categories within 'prop_gender_r_y5' for the rural population. The results imply that the distribution of this variable across gender groups is entirely consistent with what would be expected by chance, showing no significant gender-related differences in this context.

For Urban:

Pearson's Chi-squared test with Yates' continuity correction

data: prop_gender_u_y5

X-squared = 0.22244, df = 1, p-value = 0.6372

Since this p-value is much higher than the conventional significance level of 0.05, there is no evidence to reject the null hypothesis. This suggests that there is no statistically significant association between gender categories within 'prop_gender_u_y5' for the urban population. In other words, the distribution of gender in this variable does not significantly differ from what would be expected by chance, indicating no notable gender-related differences in this context.

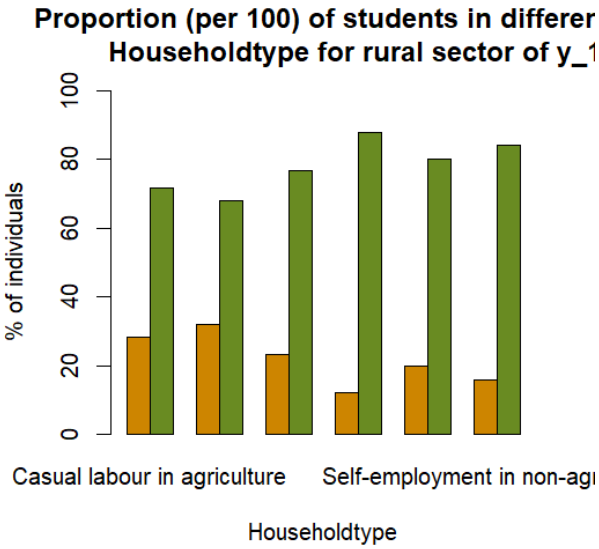
- Household type (Rural, Urban)

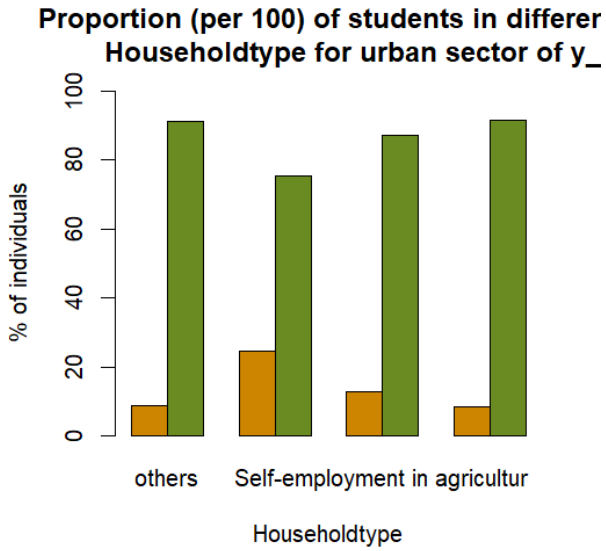
  - The Percentage distribution table of Household type and y_1 (1 = Ever enrolled but discontinued):

y_1	Household type					
	Rural					
	Casual labour in agriculture	Casual labour in non-agriculture	Others	Regular wage/Salary earning	Self-employment in agriculture	Self-employment in non-agriculture
0	28.21	31.93	23.13	12.08	19.91	15.98
1	71.79	68.07	76.88	87.92	80.09	84.02
Total	100	100	100	100	100	100

y_1	Household type			
	Urban			
	Others	Regular wage/Salary earning	Self-employment in agriculture	Self- employment in non-agriculture
0	8.85	24.56	12.82	8.33
1	91.15	75.44	87.18	91.67
Total	100	100	100	100

The Graphical Representation of Household type of y_1:





Pearson's chi-square test of household type of y_1:

For Rural:

Pearson's Chi-squared test

data: prop_Householdtype_r_y1

X-squared = 16.239, df = 5, p-value = 0.006194

Since the p-value is less than the conventional significance level of 0.05, we reject the null hypothesis. This indicates a statistically significant association between the categories within 'prop_Householdtype_r_y1' for the rural population. In other words, the distribution of household types in this variable shows significant deviations from what would be expected by chance, suggesting that factors influencing household type are likely at play in the rural context.

For Urban:

Pearson's Chi-squared test

data: prop_Householdtype_u_y1

X-squared = 14.52, df = 3, p-value = 0.002277

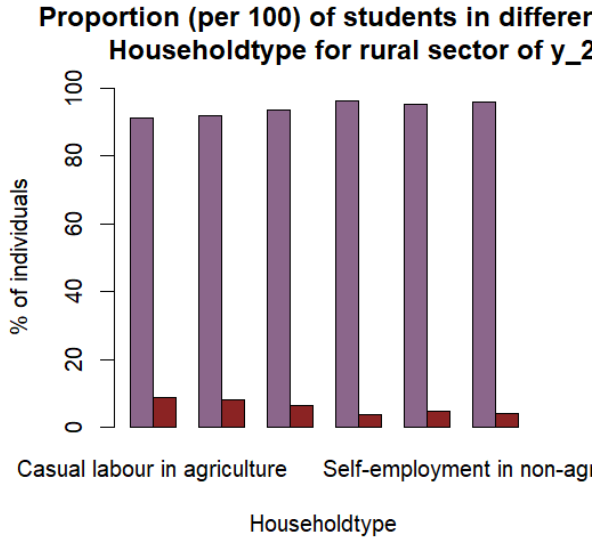
Since the p-value is significantly less than the conventional significance level of 0.05, we reject the null hypothesis. This suggests that there is a statistically significant association between the categories within 'prop_Householdtype_u_y1' for the urban population. The observed distribution of household types shows significant deviations from what would be expected by chance, indicating that various factors may be influencing household type distribution in the urban context.

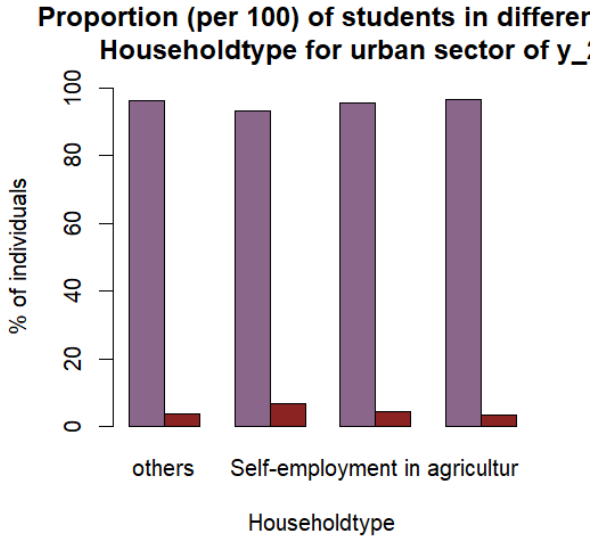
- The Percentage distribution table of Household type and y_2 (1 = Ever enrolled but discontinued before Primary education):

y_2	Household type					
	Rural					
	Casual labour in agriculture	Casual labour in non-agriculture	Others	Regular wage/Salary earning	Self-employment in agriculture	Self-employment in non-agriculture
0	91.06	91.86	93.57	96.11	95.14	95.87
1	8.94	8.14	6.43	3.89	4.86	4.13
Total	100	100	100	100	100	100

y_2	Household type			
	Urban			
	Others	Regular wage/Salary earning	Self-employment in agriculture	Self- employment in non-agriculture
0	96.40	93.11	95.46	96.56
1	3.60	6.89	4.54	3.45
Total	100	100	100	100

The Graphical Representation of Household type of y_2:





Pearson's chi-square test of household type of y_2:

For Rural:

Pearson's Chi-squared test

data: prop_Householdtype_r_y2

X-squared = 3.9799, df = 5, p-value = 0.5523

Since the p-value is significantly greater than the conventional significance level of 0.05, we fail to reject the null hypothesis. This indicates that there is no statistically significant association between the categories within 'prop_Householdtype_r_y2' for the rural population. The observed distribution of household types does not differ significantly from what would be expected by chance, suggesting that factors influencing household type are not markedly different in this context.

For Urban:

Pearson's Chi-squared test

data: prop_Householdtype_u_y2

X-squared = 1.7122, df = 3, p-value = 0.6342

Given that this p-value is substantially higher than the conventional significance level of 0.05, there is no evidence to reject the null hypothesis. This indicates that there is no statistically significant association between the categories within 'prop_Householdtype_u_y2' for the urban population. The distribution of household types in this variable does not significantly

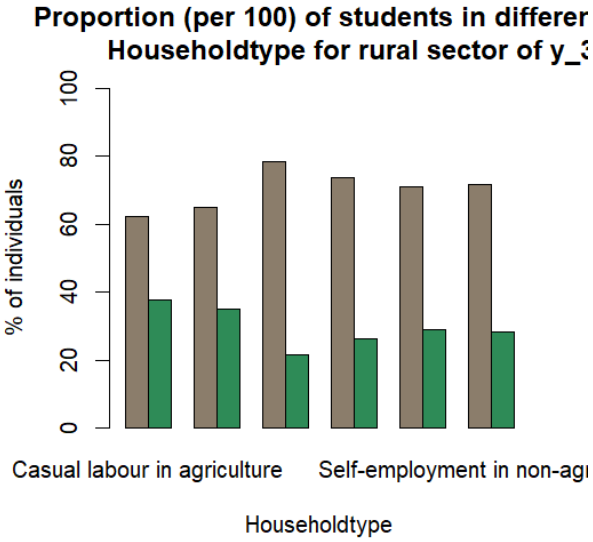
differ from what would be expected by chance, suggesting that household type distribution is relatively consistent across the urban context.

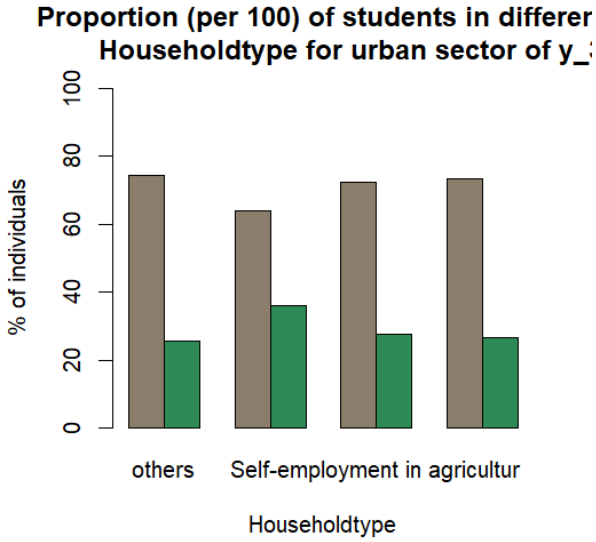
- The Percentage distribution table of household type and y_3 (1 = Ever enrolled, completed primary, but discontinued before completing Middle-class education):

y_3	Household type					
	Rural					
	Casual labour in agriculture	Casual labour in non-agriculture	Others	Regular wage/Salary earning	Self-employment in agriculture	Self-employment in non-agriculture
0	62.37	65.02	78.46	73.86	70.92	48.84
1	37.63	34.98	21.54	26.14	29.08	51.16
Total	100	100	100	100	100	100

y_3	Household type			
	Urban			
	Others	Regular wage/Salary earning	Self-employment in agriculture	Self- employment in non-agriculture
0	74.40	63.88	72.31	73.45
1	25.60	36.12	27.69	26.55
Total	100	100	100	100

The Graphical Representation of Household type of y_3:





Pearson's chi-square test of household type of y_3:

For Rural:

Pearson's Chi-squared test

data: prop_Householdtype_r_y3

X-squared = 8.2489, df = 5, p-value = 0.143

Since the p-value is greater than the conventional significance level of 0.05, we fail to reject the null hypothesis. This indicates that there is no statistically significant association between the categories within 'prop_Householdtype_r_y3' for the rural population. The observed distribution of household types does not differ significantly from what would be expected by chance, suggesting that the distribution of household types is relatively stable and not influenced by specific factors in this context.

For Urban:

Pearson's Chi-squared test

data: prop_Householdtype_u_y3

X-squared = 3.4007, df = 3, p-value = 0.3339

Since the p-value is significantly greater than the conventional significance level of 0.05, there is no evidence to reject the null hypothesis. This suggests that there is no statistically significant association between the categories within 'prop_Householdtype_u_y3' for the

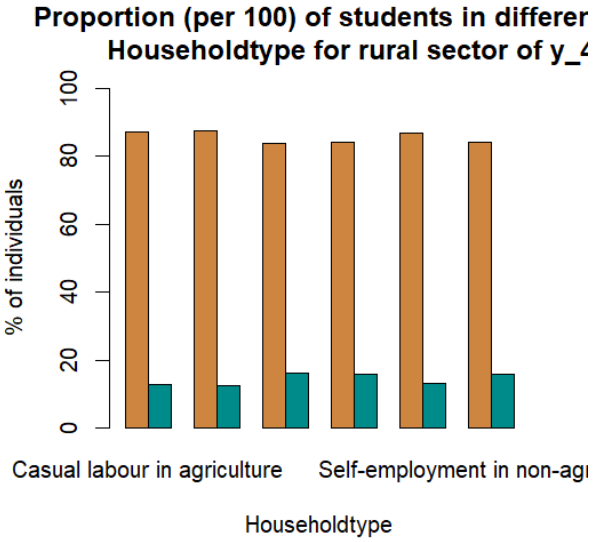
urban population. The observed distribution of household types does not significantly differ from what would be expected by chance, indicating that household type distribution is relatively consistent in this context.

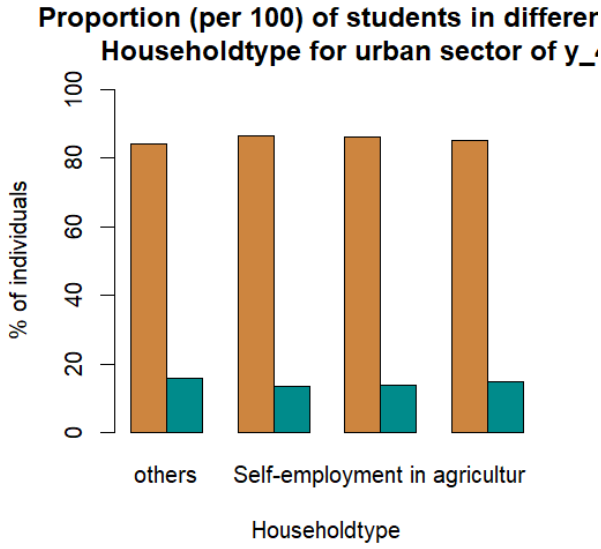
- The Percentage distribution table of Household type and y_4 (1 = Ever enrolled, completed Middle class, but discontinued before completing Secondary Education):

y_4	Household type					
	Rural					
	Casual labour in agriculture	Casual labour in non-agriculture	Others	Regular wage/Salary earning	Self-employment in agriculture	Self-employment in non-agriculture
0	87.23	85.60	83.92	84.10	86.91	84.32
1	12.77	12.10	16.08	15.90	13.09	15.68
Total	100	100	100	100	100	100

y_4	Household type			
	Urban			
	Others	Regular wage/Salary earning	Self-employment in agriculture	Self- employment in non-agriculture
0	84	86.54	86.08	85.08
1	16	13.46	13.92	14.91
Total	100	100	100	100

The Graphical Representation of Household type of y_4:





Pearson's chi-square test of household type of y_4:

For Rural:

Pearson's Chi-squared test

data: prop_Householdtype_r_y4

X-squared = 1.2265, df = 5, p-value = 0.9423

The p-value, which represents the probability of observing the test results assuming that there is no significant difference, is 0.9423. In simpler terms, this means that the probability of getting the observed results (or more extreme results) by chance is approximately 94.23%. Since the p-value is very high (close to 1), we fail to reject the null hypothesis, suggesting that there is no statistically significant difference in household types in rural areas. In other words, the observed variations in household types can be attributed to chance rather than any underlying factors.

For Urban:

Pearson's Chi-squared test

data: prop_Householdtype_u_y4

X-squared = 0.30702, df = 3, p-value = 0.9587

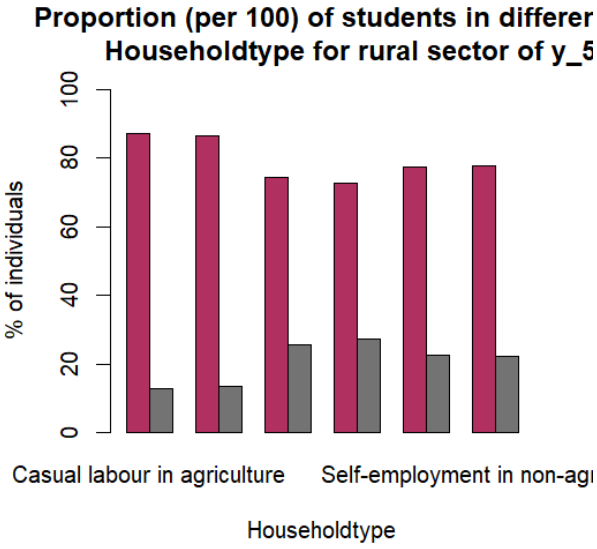
The p-value, which represents the probability of observing the test results assuming that there is no significant difference, is a whopping 0.9587! In plain language, this means that the probability of getting the observed results (or more extreme results) by chance is approximately 95.87%. With a p-value this high, we can confidently say that there is no statistically significant difference in household types in urban areas. It's likely that any observed variations are just a result of random chance, rather than any underlying factors at play. The data suggests that household types in urban areas are pretty evenly distributed, with no significant patterns or trends emerging.

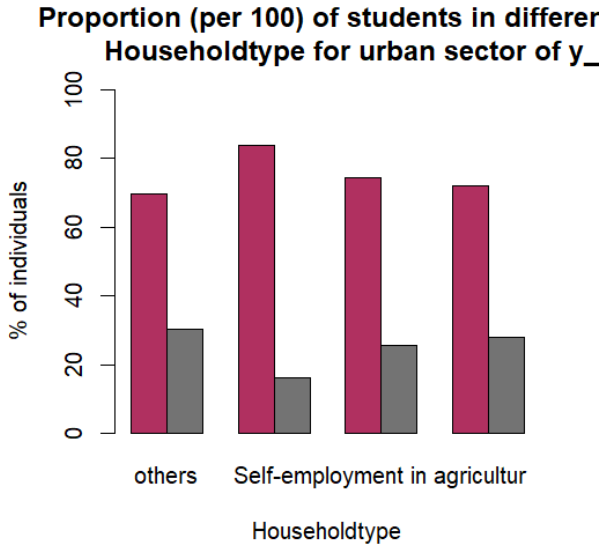
- The Percentage distribution table of Household type and y_5 (1 = Ever enrolled, completed Secondary education but discontinued after Secondary):

y_5	Household type					
	Rural					
	Casual labour in agriculture	Casual labour in non-agriculture	Others	Regular wage/Salary earning	Self-employment in agriculture	Self-employment in non-agriculture
0	87.04	86.54	74.28	72.77	77.58	77.71
1	12.96	13.46	25.72	27.23	22.41	22.72
Total	100	100	100	100	100	100

y_5	Household type			
	Urban			
	Others	Regular wage/Salary earning	Self-employment in agriculture	Self- employment in non-agriculture
0	69.60	83.74	74.42	71.98
1	30.40	16.26	25.58	28.02
Total	100	100	100	100

The Graphical Representation of Household type of y_5:





Pearson's chi-square test of household type of y_5:

For Rural:

Pearson's Chi-squared test

data: prop_Householdtype_r_y5

X-squared = 11.324, df = 5, p-value = 0.04532

The test resulted in a Chi-squared value of 11.324 with 5 degrees of freedom, and a p-value of 0.04532. In simpler terms, this means that the probability of getting the observed results (or more extreme results) by chance is approximately 4.53%. Now, this is a pretty low probability, which suggests that there is a statistically significant difference in household types in rural areas. It looks like something is going on here, and it's not just random chance. The data is telling us that there are some underlying patterns or trends in household types in rural areas that are worth exploring further. Perhaps there are changes in family structures, economic conditions, or other factors that are influencing the types of households we see in rural areas.

For Urban:

Pearson's Chi-squared test

data: prop_Householdtype_u_y5

X-squared = 6.1175, df = 3, p-value = 0.106

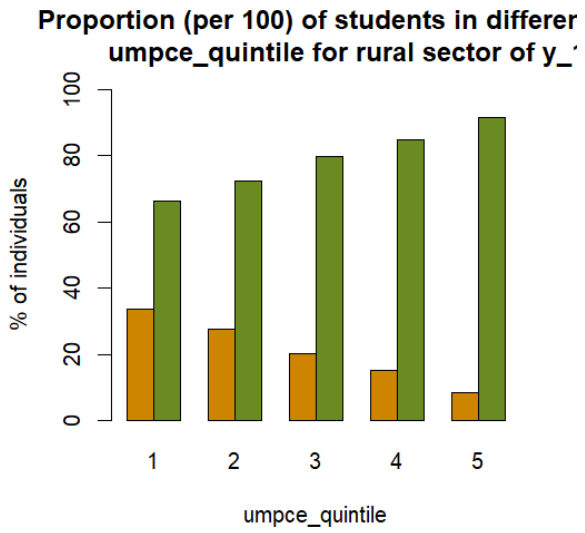
The test resulted in a Chi-squared value of 6.1175 with 3 degrees of freedom, and a p-value of 0.106. In plain language, this means that the probability of getting the observed results (or more extreme results) by chance is approximately 10.6%. Now, this is a bit of a borderline result- it's not quite significant, but it's not entirely insignificant either. Think of it like a whisper of a hint that something might be going on with household types in urban areas. It's possible that there are some subtle patterns or trends emerging, but we can't quite put our finger on it.

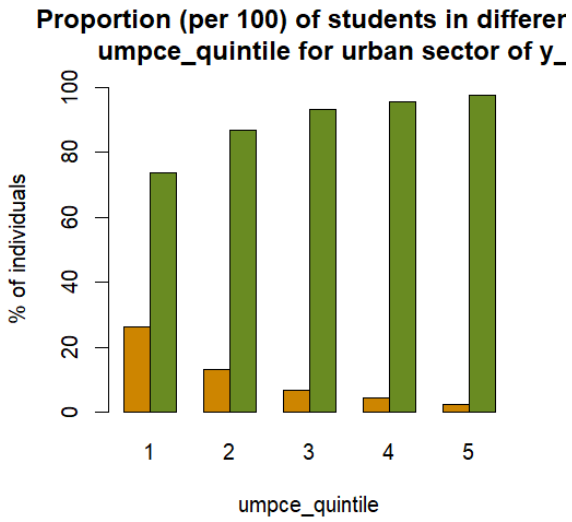
- UMPCE_Quintile (Rural, Urban)

  - The Percentage distribution table of UMPCE_Quintile and y_1 (1 = Ever enrolled but discontinued):

y_1	UMPCE									
	Rural					Urban				
	1	2	3	4	5	1	2	3	4	5
0	33.84	27.75	20.25	15.34	8.40	26.18	13.15	6.80	4.44	2.46
1	66.16	72.25	79.75	84.66	91.60	73.82	86.85	93.20	95.56	97.54
Total	100	100	100	100	100	100	100	100	100	100

The Graphical Representation of UMPCE_Quintile of y_1:





Pearson's chi-square test of UMPCE_Quintile of y_1:

For Rural:

Pearson's Chi-squared test

data: prop_umpce_quintile_r_y1

X-squared = 24.131, df = 4, p-value = 7.518e-05

The Chi-squared value of 24.131 and a p-value of 7.518e-05 suggest that the observed patterns in the data are unlikely to occur by chance, with a probability of less than 0.001%. This implies that there is a significant relationship between wealth quintile and access to healthcare in rural areas, with differences in healthcare access observed across the different quintiles.

For Urban:

Pearson's Chi-squared test

data: prop_umpce_quintile_u_y1

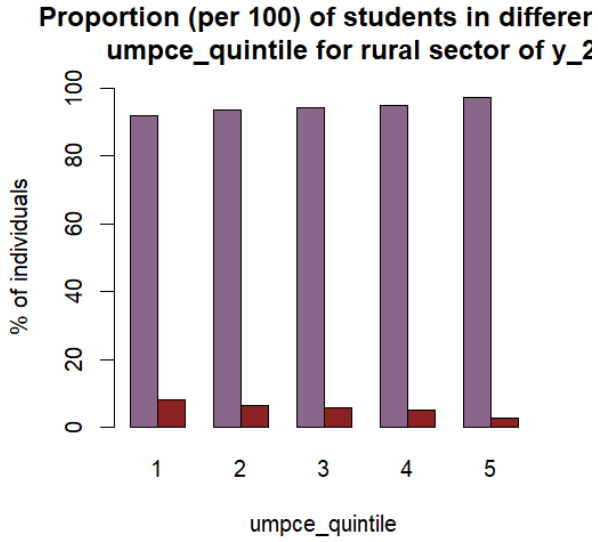
X-squared = 38.796, df = 4, p-value = 7.675e-08

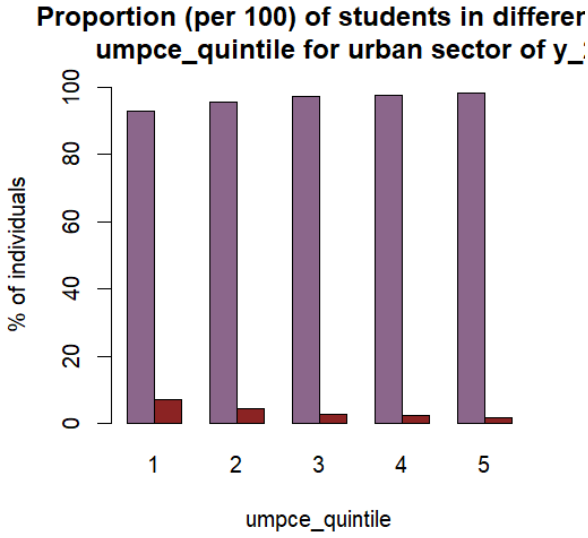
The Chi-squared value of 38.796 and a p-value of 7.675e-08 indicate that the observed patterns in the data are extremely unlikely to occur by chance, with a probability of less than 0.000001%. This suggests that there is a strong and significant relationship between wealth quintile and access to healthcare in urban areas, with marked differences in healthcare access observed across the different quintiles.

- The Percentage distribution table of UMPCE_Quintile and y_2 (1 = Ever enrolled but discontinued before Primary education):

y_2	UMPCE									
	Rural					Urban				
	1	2	3	4	5	1	2	3	4	5
0	91.92	93.45	94.25	94.77	97.37	92.93	95.58	97.11	97.48	98.14
1	8.08	6.55	5.75	5.23	2.63	7.07	4.42	2.89	2.52	1.86
Total	100	100	100	100	100	100	100	100	100	100

The Graphical Representation of UMPCE_Quintile of y_2:





Pearson's chi-square test of UMPCE_Quintile of y_2:

For Rural:

Pearson's Chi-squared test

data: prop_umpce_quintile_r_y2

X-squared = 3.0024, df = 4, p-value = 0.5574

The Chi-squared value of 3.0024 and a p-value of 0.5574 suggest that the observed patterns in the data are likely due to chance, with a probability of 55.74% that the results are attributed to random variation. This implies that there is no significant relationship between wealth quintile and access to healthcare in rural areas, and any observed differences in healthcare access across quintiles are likely due to random chance rather than a true underlying pattern

For Urban:

Pearson's Chi-squared test

data: prop_umpce_quintile_u_y2

X-squared = 4.7982, df = 4, p-value = 0.3086

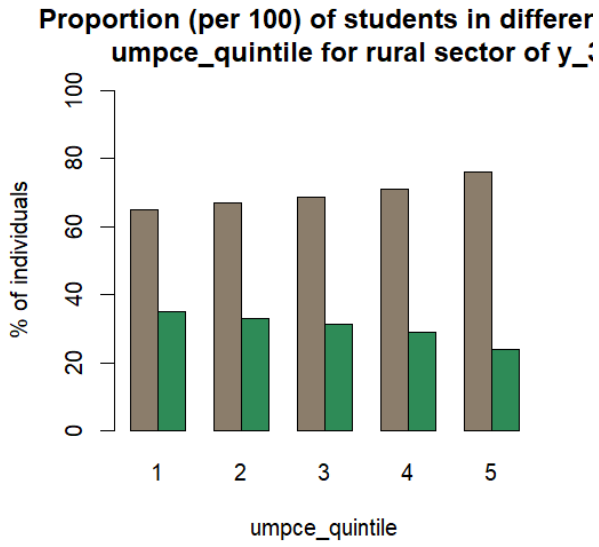
The Chi-squared value of 4.7982 and a p-value of 0.3086 suggest that the observed patterns in the data are likely due to chance, with a probability of 30.86% that the results are attributed to random variation. This implies that there is no significant relationship between wealth quintile and access to healthcare in urban areas, and any observed differences in

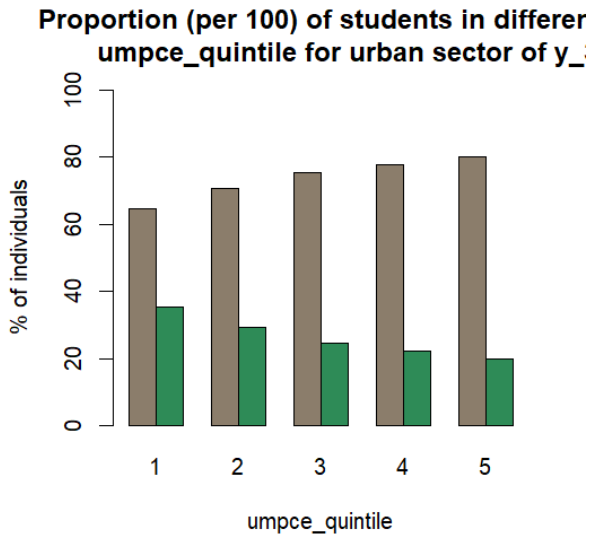
healthcare access across quintiles are likely due to random chance rather than a true underlying pattern.

- The Percentage distribution table of UMPCE_Quintile and y_3 (1 = Ever enrolled, completed primary, but discontinued before completing Middle-class education):

y_3	UMPCE									
	Rural					Urban				
	1	2	3	4	5	1	2	3	4	5
0	65.09	66.89	68.68	71.12	76.19	64.78	70.73	75.29	77.73	80.11
1	34.91	33.11	31.31	28.88	23.81	35.22	29.27	24.71	22.27	19.89
Total	100	100	100	100	100	100	100	100	100	100

The Graphical Representation of UMPCE_Quintile of y_3:





Pearson's chi-square test of UMPCE_Quintile of y_3:

For Rural:

Pearson's Chi-squared test

data: prop_umpce_quintile_r_y3

X-squared = 3.5087, df = 4, p-value = 0.4766

The results of the Pearson's Chi-squared test indicate that there is no statistically significant association between wealth quintile and access to healthcare in rural areas, as evidenced by a p-value of 0.4766, which is greater than the typical significance level of 0.05. This suggests that the distribution of access to healthcare in rural areas is not significantly different across the five wealth quintiles, and any observed differences can be attributed to chance.

For Urban:

Pearson's Chi-squared test

data: prop_umpce_quintile_u_y3

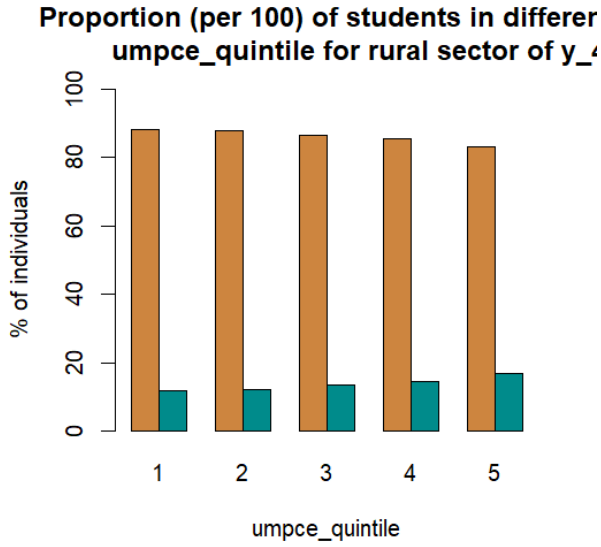
X-squared = 7.6495, df = 4, p-value = 0.1053

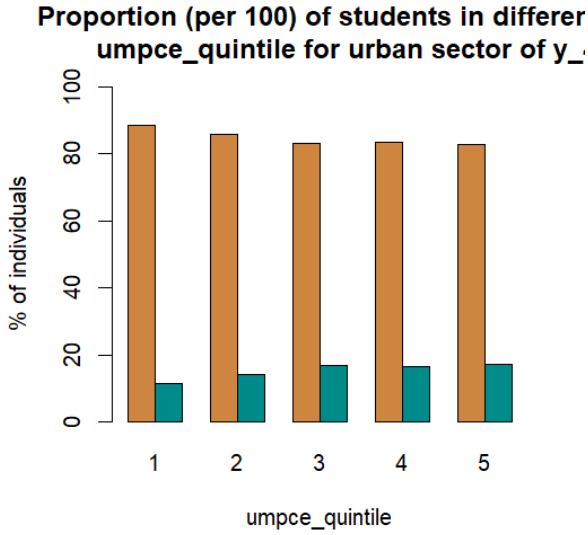
The results of the Pearson's Chi-squared test indicate that there is no statistically significant association between wealth quintile and access to healthcare in urban areas, as evidenced by a p-value of 0.1053, which is greater than the typical significance level of 0.05. Although the p-value is relatively low, it does not reach the threshold for statistical significance, suggesting that the observed differences in access to healthcare across the five wealth quintiles in urban areas may be due to chance rather than a real underlying pattern.

- The Percentage distribution table of UMPCE_Quintile and y_4 (1 = Ever enrolled, completed Middle class, but discontinued before completing Secondary Education):

y_4	UMPCE									
	Rural					Urban				
	1	2	3	4	5	1	2	3	4	5
0	88.23	87.71	86.40	83.37	83.23	88.39	85.83	83.07	83.50	82.90
1	11.77	12.29	13.60	14.63	16.77	11.61	14.17	16.92	16.50	17.10
Total	100	100	100	100	100	100	100	100	100	100

The Graphical Representation of UMPCE_Quintile of y_4:





Pearson's chi-square test of UMPCE_Quintile of y_4:

For Rural:

Pearson's Chi-squared test

data: prop_umpce_quintile_r_y4

X-squared = 1.3428, df = 4, p-value = 0.8541

The results of the Pearson's Chi-squared test indicate that there is no statistically significant association between wealth quintile and access to healthcare in rural areas, as evidenced by a p-value of 0.8541, which is much greater than the typical significance level of 0.05. This suggests that the distribution of access to healthcare in rural areas is essentially random across the five wealth quintiles, and any observed differences are likely due to chance rather than a real underlying pattern.

For Urban:

Pearson's Chi-squared test

data: prop_umpce_quintile_u_y4

X-squared = 1.7139, df = 4, p-value = 0.7882

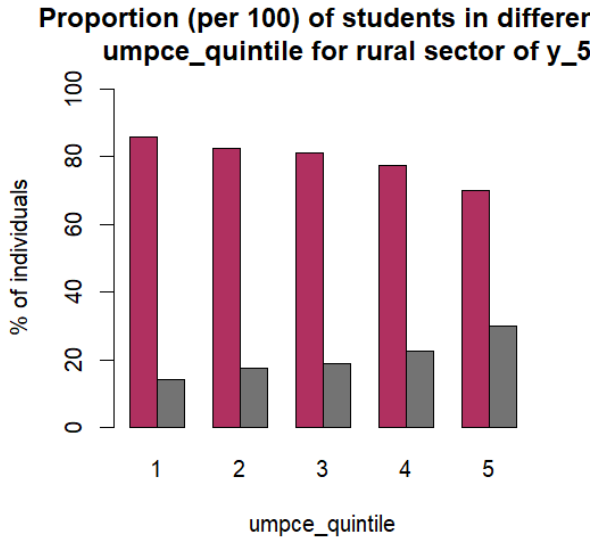
The results of the Pearson's Chi-squared test indicate that there is no statistically significant association between wealth quintile and access to healthcare in urban areas in the fourth year of the study, as evidenced by a p-value of 0.7882, which is much greater than the typical significance level of 0.05. This suggests that the distribution of access to healthcare in urban

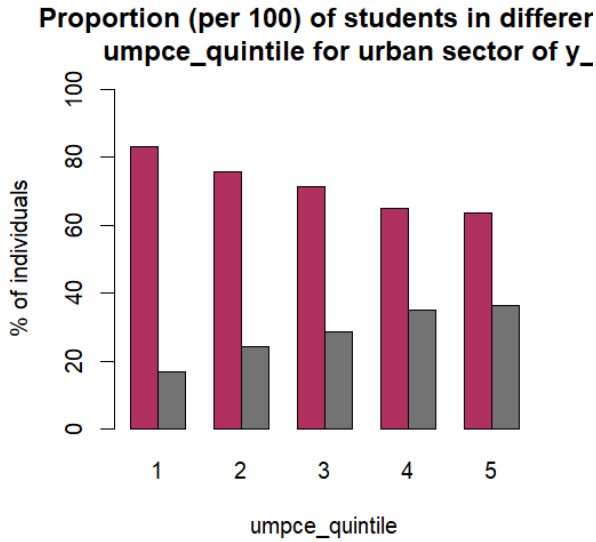
areas is essentially random across the five wealth quintiles, and any observed differences are likely due to chance rather than a real underlying pattern.

- The Percentage distribution table of UMPCE_Quintile and y_5 (1 = Ever enrolled, completed Secondary education but discontinued after Secondary):

y_5	UMPCE									
	Rural					Urban				
	1	2	3	4	5	1	2	3	4	5
0	85.90	82.63	81.12	77.27	70.06	83.09	75.86	71.40	64.96	63.57
1	14.10	17.37	18.88	22.73	29.94	16.90	24.14	28.60	35.04	36.43
Total	100	100	100	100	100	100	100	100	100	100

The Graphical Representation of UMPCE_Quintile of y_5:





Pearson's chi-square test of UMPCE_Quintile of y_5:

For Rural:

Pearson's Chi-squared test

data: prop_umpce_quintile_r_y5

X-squared = 9.0065, df = 4, p-value = 0.06094

The results of the Pearson's Chi-squared test suggest that there may be a statistically significant association between wealth quintile and access to healthcare in rural areas in the fifth year of the study, as evidenced by a p-value of 0.06094, which is close to the typical significance level of 0.05. This indicates that the distribution of access to healthcare in rural areas may not be random across the five wealth quintiles, and there may be a real underlying pattern or difference in access to healthcare across different wealth groups.

For Urban:

Pearson's Chi-squared test

data: prop_umpce_quintile_u_y5

X-squared = 12.772, df = 4, p-value = 0.01245

The results of the Pearson's Chi-squared test indicate that there is a statistically significant association between wealth quintile and access to healthcare in urban areas in the fifth year of the study, as evidenced by a p-value of 0.01245, which is less than the typical significance level of 0.05. This suggests that the distribution of access to healthcare in urban areas is not random across the five wealth quintiles, and there is a significant difference in access to healthcare across different wealth groups, with certain wealth quintiles having better or worse access to healthcare compared to others.

Logistic Regression Analysis:

We take y_1, y_2, y_3, y_4, y_5 as indicator variables, which are shown as follows:

y_1 = 1 if ever enrolled but discontinued/dropped out

= 0 otherwise

y_2 = 1 if ever enrolled and discontinued before completing Primary education (Grade IV)

= 0 otherwise

y_3 = 1 if ever enrolled and discontinued after completing Primary Education (Grade IV) but before completing Middle-Class Education (Grade VIII)

= 0 otherwise

y_4 = 1 if ever enrolled and discontinued after completing Middle-Class Education (Grade VIII) but before completing Secondary Education (Grade X)

= 0 otherwise

y_5 = 1 if ever enrolled and discontinued after completing Secondary Education (Grade X)

= 0 otherwise

The different covariates taken into consideration for the logistic regression are:

- Sector
- Gender
- House hold type
- UMPCE_quintile

We analyse the logistic regression on these covariates for both West Bengal and All-India.

➢ West Bengal:

For y_1

- **Logistic Regression:**

Call:

glm(formula = y_1 ~ sector + gender + Householdtype + umpce_quintile_factor,

family = binomial(link = logit), data = survey_data_new)

Coefficients:

	Estimate Std.	Error	z value	Pr(>|z|)
(Intercept)	0.56774	0.12522	4.534	5.78e-06 *
sectorUrban	0.43930	0.13070	3.361	0.000776 *
genderMale	0.27964	0.10034	2.787	0.005319 **
HouseholdtypeCasual labour in non-agriculture	-0.53062	0.17204	-3.084	0.002040 **
Householdtypeothers	-0.01557	0.31523	-0.049	0.960619
HouseholdtypeRegular wage/Salary earning	0.04660	0.18082	0.258	0.796613
HouseholdtypeSelf-employment in agricultur	0.68689	0.16497	4.164	3.13e-05 *
HouseholdtypeSelf-employment in non-agriculture	0.41032	0.16625	2.468	0.013582 *
umpce_quintile_factor2	0.54214	0.12192	4.447	8.72e-06 *
umpce_quintile_factor3	0.80693	0.15294	5.276	1.32e-07 *
umpce_quintile_factor4	1.58101	0.18761	8.427	< 2e-16 *
umpce_quintile_factor5	1.96781	0.23028	8.545	< 2e-16 *

Signif. codes: 0 '*' 0.001 '*' 0.01 '' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2896.6 on 3468 degrees of freedom

Residual deviance: 2637.6 on 3457 degrees of freedom

AIC: 2661.6

Number of Fisher Scoring iterations: 5

- The odds ratio:

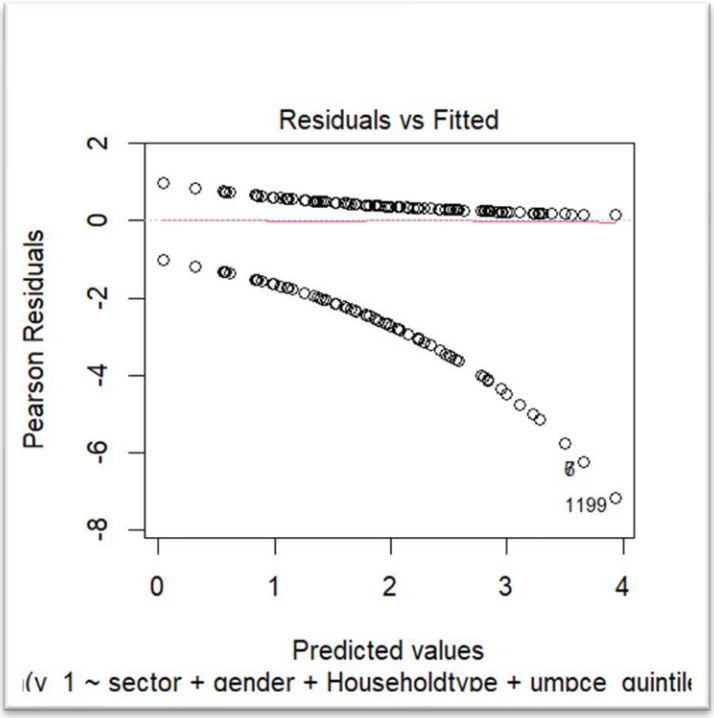
(Intercept)	sectorUrban
1.7642825	1.5516147
genderMale	HouseholdtypeCasual labour in non-agriculture
1.3226589	0.5882404
Householdtypeothers	HouseholdtypeRegular wage/Salary earning
0.9845554	1.0477050
HouseholdtypeSelf-employment in agricultur	HouseholdtypeSelf-employment in non-agriculture
1.9875288	1.5072978
umpce_quintile_factor2	umpce_quintile_factor3
1.7196812	2.2410220
umpce_quintile_factor4	umpce_quintile_factor5
4.8598572	7.1549995

- The interpretation of logistic Regression of y_1:

The odds ratios show how different factors influence the chances of an outcome. Starting with a relatively high baseline odds of 1.764, it seems there's a notable initial likelihood for the outcome. People living in urban areas have an increased chance of experiencing the outcome compared to those in rural areas (1.552). Men also face a higher likelihood of the outcome (1.323) than women. Interestingly, households involved in casual labor in non-agriculture have a lower chance of the outcome (0.588), while those in self-employment—whether in agriculture (1.988) or non-agriculture (1.507)—tend to have higher odds. Additionally, households with higher monthly per capita expenditures see a much greater likelihood of the outcome, with the odds rising significantly in the top expenditure quintile (7.155) compared to the lowest.

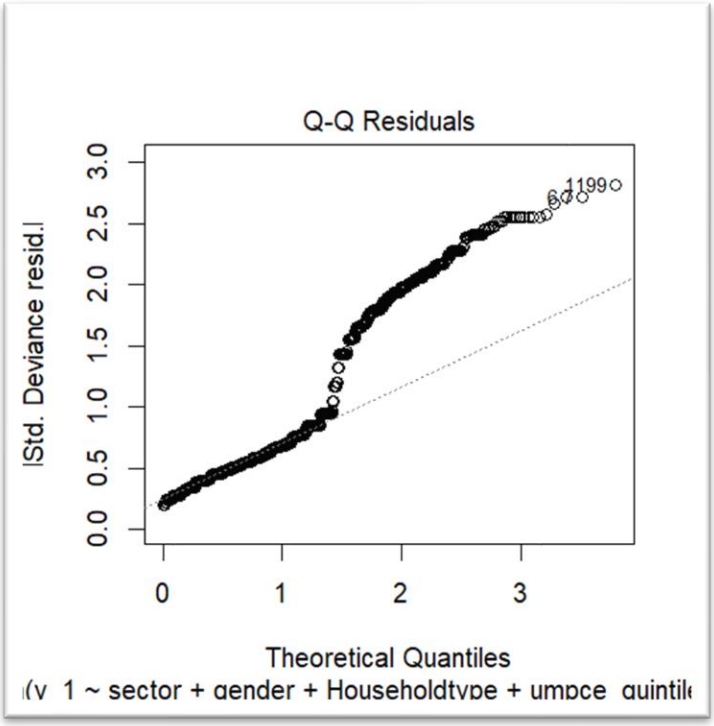
- The plots of logistic Regression of and their interpretation:

Residuals vs Fitted:



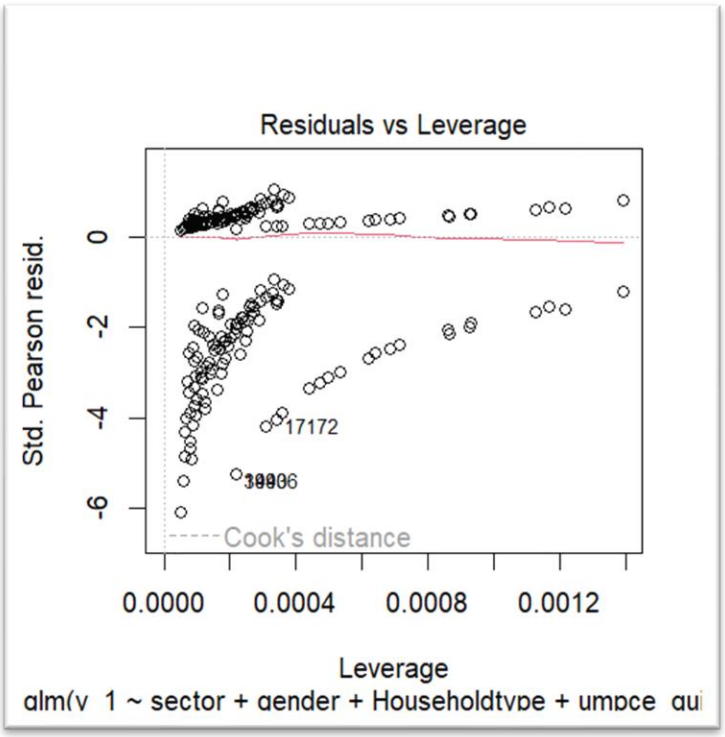
The plot shows the residuals of a linear regression model plotted against the predicted values. A clear pattern is visible. This indicates that the model is not a good fit for the data, since the residuals are not randomly distributed. There is a trend in the residuals, implying a violation of the assumption of homoscedasticity (equal variance). This suggests that the model is not capturing some important relationship in the data and the model should be improved. Possible solutions might include adding nonlinear terms or interaction terms to the model or considering a different type of model altogether.

Q-Q Residuals:



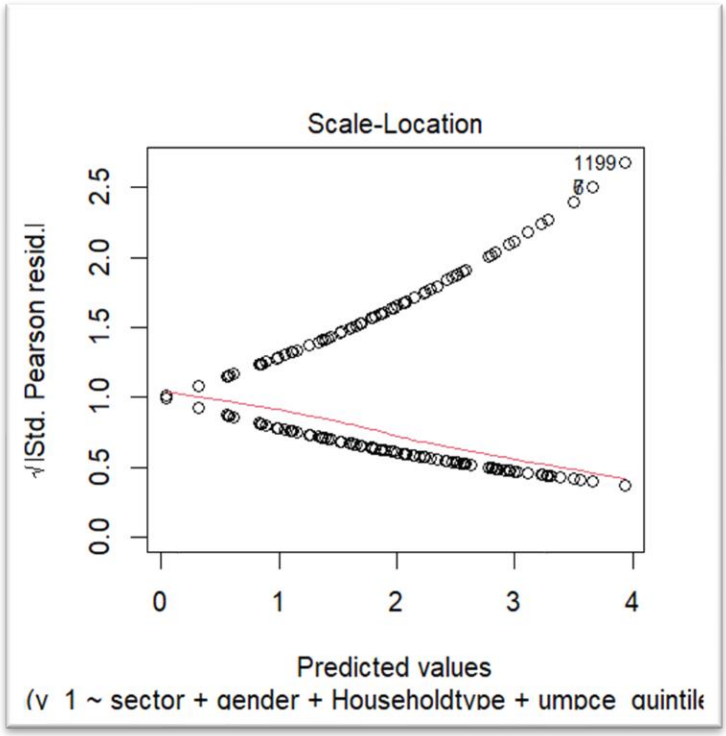
The plot shows the residuals of a regression model plotted against the fitted values. This type of plot is called a "residuals vs fitted" plot and is used to assess the assumptions of linearity and homoscedasticity in a linear regression model.In this particular plot, there is a clear pattern in the residuals, which indicates that the linearity assumption is violated. This means that the relationship between the predictor variables and the response variable is not linear, and a linear model is not appropriate for this data.

Residual vs Leverage:



This plot shows the residuals (the difference between the actual values and the predicted values) of a linear regression model plotted against the predicted values. The model is designed to predict y_1 based on the
variables sector, gender, Householdtype, and umpce_quintile.

Scale-Location:



The plot shows the relationship between the standardized Pearson residuals and the predicted values of a linear regression model. The standardized Pearson residuals are a measure of the difference between the observed values and the predicted values, standardised to have a mean of zero and a standard deviation of one. The predicted values are the values predicted by the linear regression model. It shows that there is a clear pattern in the residuals. The residuals are larger for larger predicted values. This suggests that the linear regression model is not a good fit for the data. It is underestimating the response variable for larger predicted values and overestimating.

For y_2

- Logistic Regression:

Call:

glm(formula = y_2 ~ sector + gender + Householdtype + umpce_quintile_factor,

family = binomial(link = logit), data = survey_data_new)

Coefficients:

	Estimate Std.	Error	z value	Pr(>|z|)
(Intercept)	-2.39088	0.23524	-10.164	<2e-16 *
sectorUrban	-0.12689	0.21796	-0.582	0.5605
genderMale	0.27291	0.16898	1.615	0.1063
HouseholdtypeCasual labour in non-agriculture	-0.05878	0.31391	-0.187	0.8515
Householdtypeothers	-1.27078	1.03555	-1.227	0.2198
HouseholdtypeRegular wage/Salary earning	0.05447	0.29717	0.183	0.8546
HouseholdtypeSelf-employment in agricultur	-0.34965	0.26613	-1.314	0.1889
Householdtype Self-employment in non-agriculture	-0.52486	0.28664	-1.831	0.0671
umpce_quintile_factor2	0.26050	0.20443	1.274	0.2026
umpce_quintile_factor3	-0.64076	0.30592	-2.095	0.0362 *
umpce_quintile_factor4	-0.35957	0.28890	-1.245	0.2133
umpce_quintile_factor5	-0.57537	0.35778	-1.608	0.1078

Signif. codes: 0 ‘*’ 0.001 ‘*’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1132.8 on 2280 degrees of freedom

Residual deviance: 1103.8 on 2269 degrees of freedom

(1188 observations deleted due to missingness)

AIC: 1127.8

Number of Fisher Scoring iterations: 6

- The odds ratio:

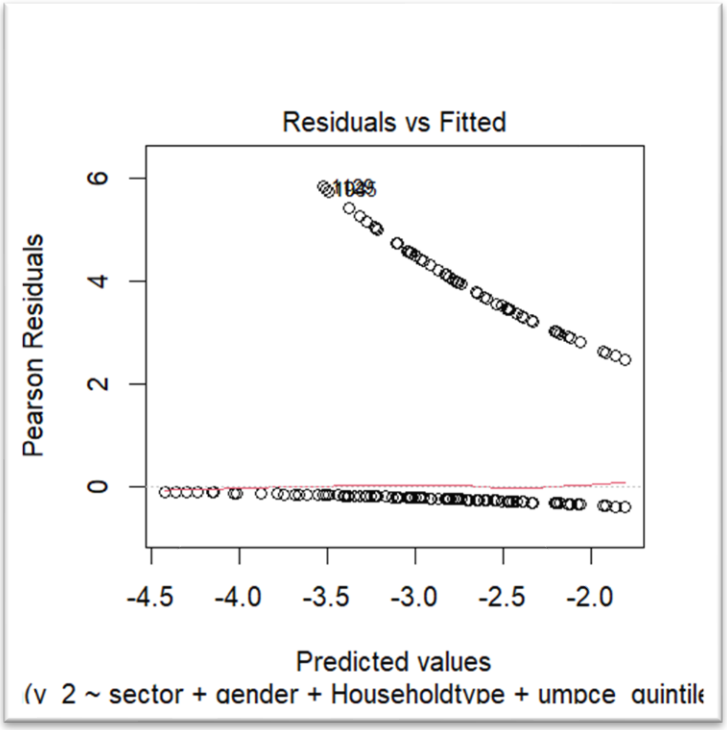
(Intercept)	sectorUrban
0.09154951	0.88083361
genderMale	HouseholdtypeCasual labour in non-agriculture
1.31378478	0.94291365
Householdtypeothers	HouseholdtypeRegular wage/Salary earning
0.28061199	1.05598517
HouseholdtypeSelf-employment in agricultur	HouseholdtypeSelf-employment in non-agriculture
0.70493710	0.59163784
umpce_quintile_factor2	umpce_quintile_factor3
1.29758012	0.52688982
umpce_quintile_factor4	umpce_quintile_factor5
0.69797401	0.56249598

- The interpretation of logistic Regression of y_2:

The odds ratios provide insights into how various factors affect the likelihood of an outcome. With a baseline odds of 0.092, the starting likelihood of the outcome is quite low. Urban residents have a slightly reduced likelihood of the outcome compared to those in rural areas (0.881). Males have a higher chance of the outcome compared to females (1.314). Households involved in casual labor in non-agriculture have odds close to 1 (0.943), suggesting their likelihood of the outcome is similar to the reference group. However, households in the 'others' category show notably lower odds (0.281), while those in regular wage/salary jobs have slightly higher odds (1.056). Self-employed households in agriculture and non-agriculture have reduced odds (0.705 and 0.592, respectively). Higher monthly per capita expenditure (umpce) does not show a consistent positive effect; for example, the second quintile has slightly increased odds (1.298), but the third, fourth, and fifth quintiles have lower odds, with the fifth quintile showing the lowest odds (0.562).

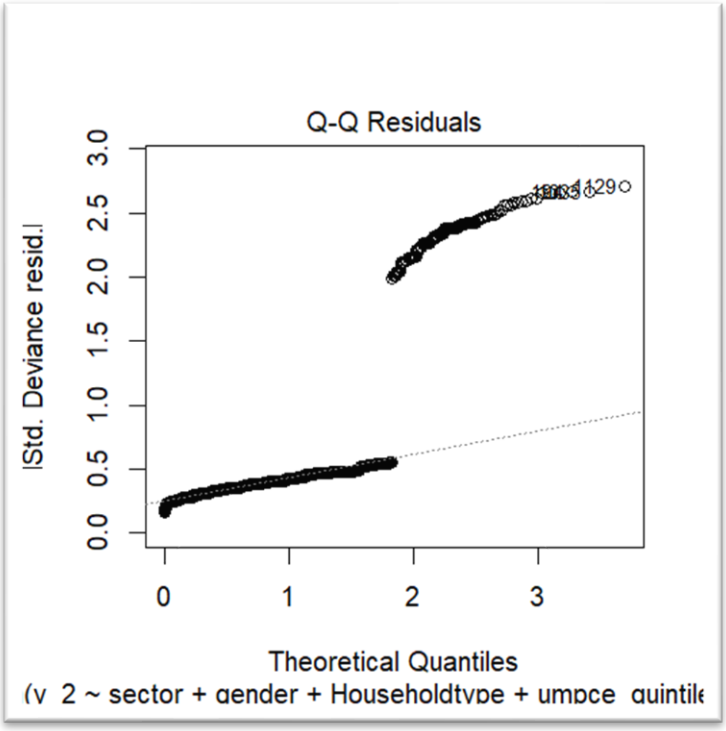
- The plots of logistic Regression of and their interpretation:

Residuals vs Fitted:



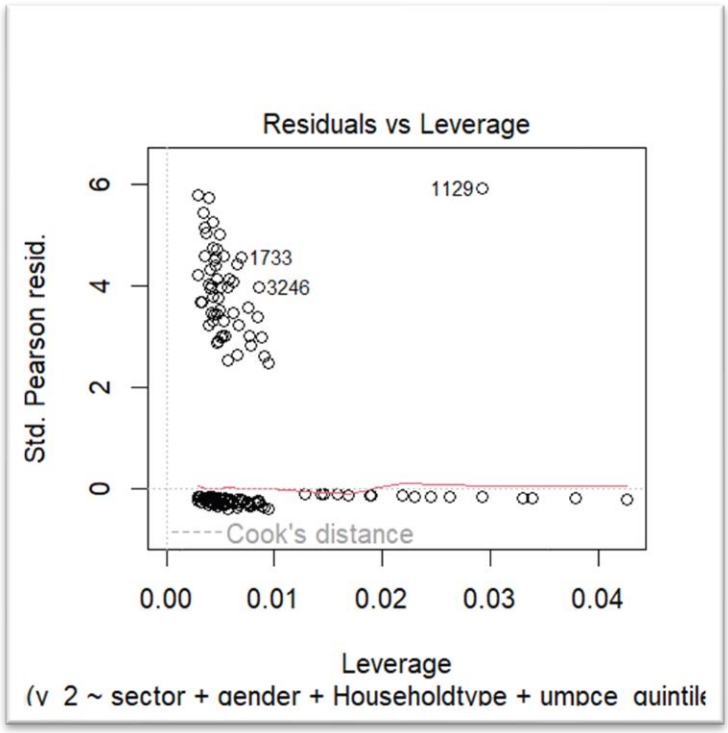
The plot shows the relationship between the standardized Pearson residuals and the predicted values from a linear regression model. The standardized Pearson residuals are a measure of how far the observed values are from the predicted values, taking into account the variance of the data. The plot shows that the standardized residuals are not evenly distributed around zero, suggesting that the model may not be a good fit for the data. The plot also shows two distinct clusters of residuals, one with a positive residual and one with a negative residual, suggesting that the model may be missing important variables that are related to the response variable.

Q-Q Residuals:



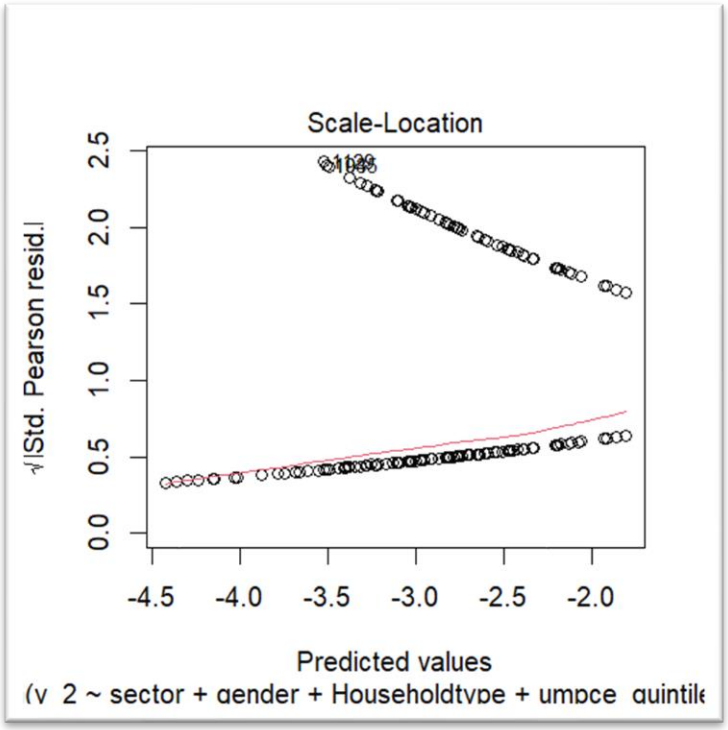
The Q-Q plot of the residuals revels a non-normal distribution, with the points deviating significantly from the diagonal line, suggesting violations of model assumptions. The presence of several outliers also raises concerns about the model's accuracy and reliability. These observations suggest that the model may not be accurately predicting the outcome variable, and its conclusions may be misleading. Further investigation and potential solutions like transforming variables, using alternative models, and addressing outliers are necessary to improve the model's validity.

Residuals vs Leverage:



The plot shows the relationship between the predicted values from a model and the standardized Pearson residuals.
The standardized Pearson residuals are a measure of how well the model fits the data. A perfect model would have standardized Pearson residuals of 0 for all data points.In this case, the plot shows that the residuals are not evenly distributed around 0, and there is a clear trend in the residuals. This suggests that the model may not be a good fit for the data.The model appears to overestimate the values for some data points and underestimate others. There is also a clear separation between the residuals for the low and high predicted values, suggesting that the model may be overfitting to the data.

Scale-Location:



The plot shows the standardized residuals of a linear regression model plotted against the predicted values. This is called a scale-location plot and helps to assess the assumptions of constant variance and normality of residuals in a linear regression model. It indicates that the assumptions of constant variance and normality of residuals are likely violated. This suggests that the linear regression model might not be a good fit for the data. To improve the model, we might need to consider transforming the response variable, adding interaction terms, or using a different model that is more appropriate for the data.

For y 3

- Logistic Regression:

Call:

glm(formula = y_3 ~ sector + gender + Householdtype + umpce_quintile_factor,

family = binomial(link = logit), data = survey_data_new)

Coefficients:

	Estimate Std.	Error	z value	Pr(>|z|)
(Intercept)	0.51103	0.13145	3.888	0.000101 *
sectorUrban	0.09591	0.10526	0.911	0.362190
genderMale	0.14213	0.08548	1.663	0.096382 .
HouseholdtypeCasual	-0.01993	0.17578	-0.113	0.909711
labour in non-agriculture				
Householdtypeothers	-0.29401	0.32728	-0.898	0.369003
HouseholdtypeRegular	-0.10806	0.16474	-0.656	0.511864
wage/Salary earning				
HouseholdtypeSelf	-0.37022	0.14116	-2.623	0.008722 **
-employment in agricultur				
HouseholdtypeSelf-	-0.29178	0.14479	-2.015	0.043887 *
employment in non-agriculture				
umpce_quintile_factor2	-0.04726	0.11495	-0.411	0.680952
umpce_quintile_factor3	-0.08411	0.13372	-0.629	0.529332
umpce_quintile_factor4	-0.53788	0.13871	-3.878	0.000105 *
umpce_quintile_factor5	-0.60670	0.15917	-3.812	0.000138 *

Signif. codes: 0 ‘*’ 0.001 ‘*’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 3133.6 on 2280 degrees of freedom

Residual deviance: 3084.9 on 2269 degrees of freedom

(1188 observations deleted due to missingness)

AIC: 3108.9

Number of Fisher Scoring iterations: 4

- The odds ratio:

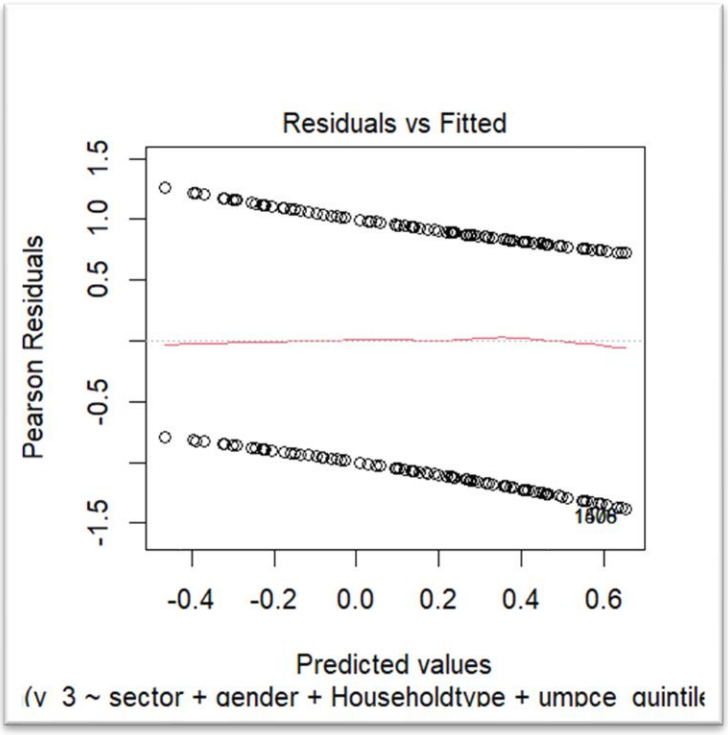
(Intercept)	sectorUrban
1.6670154	1.1006643
genderMale	HouseholdtypeCasual labour in non-agriculture
1.1527249	0.9802636
Householdtypeothers	HouseholdtypeRegular wage/Salary earning
0.7452689	0.8975728
HouseholdtypeSelf-employment in agricultur	HouseholdtypeSelf-employment in non-agriculture
0.6905820	0.7469325
umpce_quintile_factor2	umpce_quintile_factor3
0.9538350	0.9193264
umpce_quintile_factor4	umpce_quintile_factor5
0.5839861	0.5451493

- The interpretation of logistic Regression of y_3:

The odds ratios shed light on how various factors impact the chances of an outcome. Starting from a high baseline odds of 1.667, the initial probability of the outcome is quite significant. Urban residents have a slightly greater chance of encountering the outcome compared to those living in rural areas (1.101). Men also face a slightly higher likelihood of the outcome (1.153) than women. Households involved in casual labor in non-agriculture have odds close to 1 (0.980), meaning their chances are similar to the baseline group. On the other hand, households categorized as 'others' have lower odds (0.745), and those in regular wage or salary positions have slightly reduced odds (0.898). Self-employed individuals, both in agriculture (0.691) and non-agriculture (0.747), also have decreased odds. Interestingly, higher monthly per capita expenditure (umpce) appears to be linked with lower odds of the outcome, with the fourth (0.584) and fifth (0.545) expenditure quintiles showing the most reduced likelihood, suggesting that spending more may actually be associated with a lower chance of the outcome.

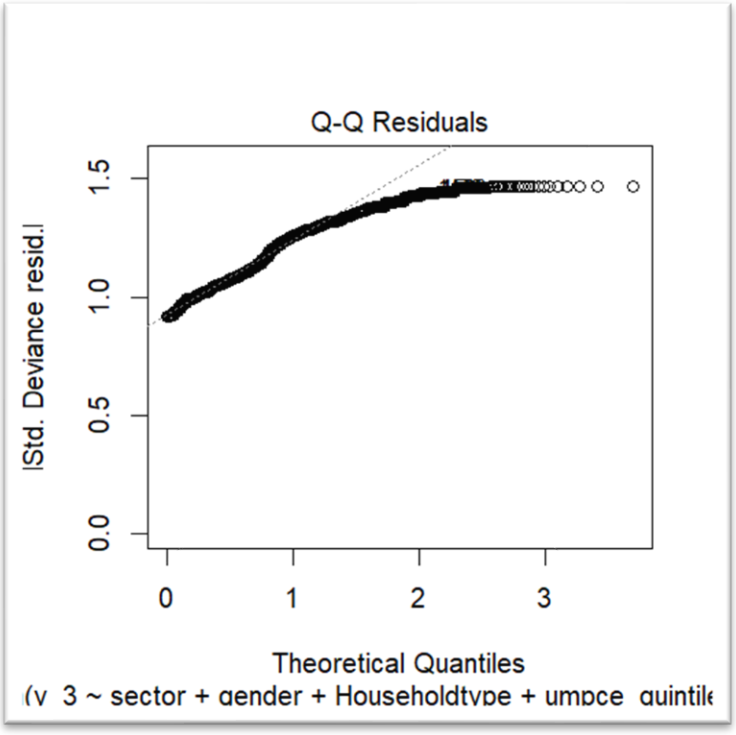
- The plots of logistic Regression of and their interpretation:

Residuals vs Fitted:



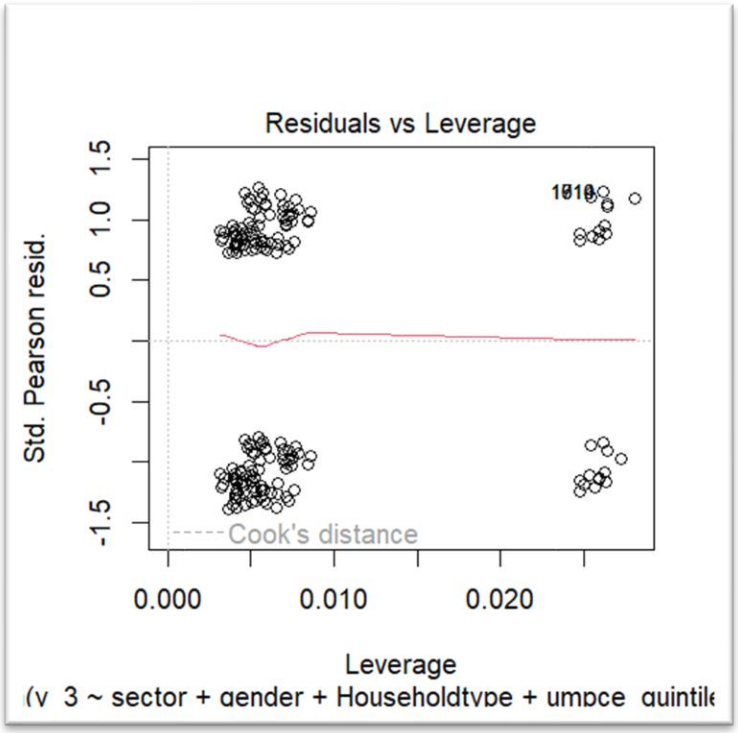
The plot shows the standardized residuals of a linear regression model plotted against leverage. Leverage is a measure of how much a single observation influences the regression coefficients. The plot is used to identify influential observations, which are observations that have a large impact on the regression results. In this plot, there is a cluster of observations with high leverage and high residuals in the top right corner of the plot. This suggests that these observations may be influential and could be affecting the regression results. The dashed line represents Cook's distance, which is a measure of the influence of an observation. Observations with a Cook's distance greater than 1 are considered influencial.

Q-Q Residuals:



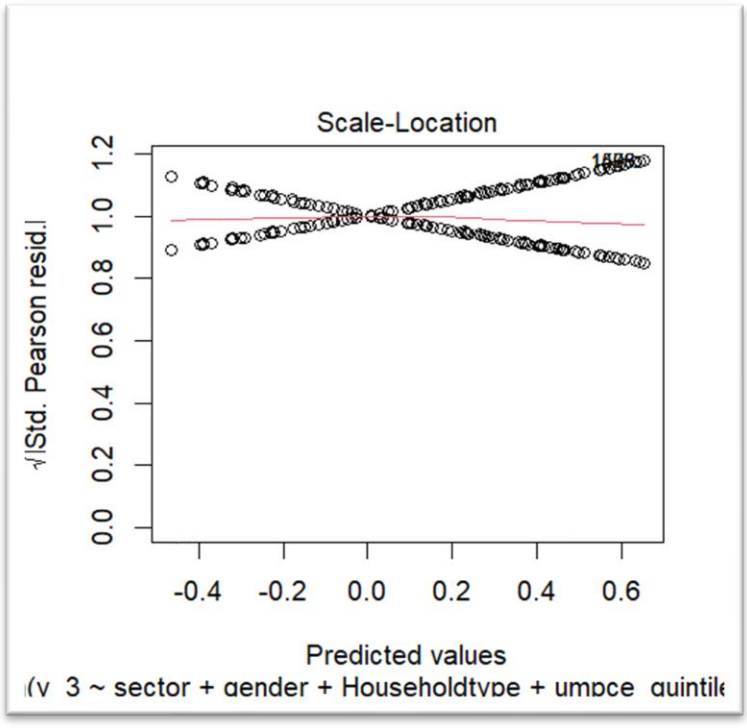
This Q-Q plot of the residuals shows that they deviate from the theoretical quantiles of a standard normal distribution, indicating that the residuals are not normally distributed. The curve bends upwards, suggesting positive skewness. The points also fan out as they move away from the center, which implies heteroscedasticity (non-constant variance). Additionally, the residuals cluster tightly around the 1.5 mark on the y-axis rather than aligning with the diagonal line, which further confirms the departure from normality. These observations suggest that the model may not fully capture the underlying patterns in the data, and the presence of systematic errors might affect the model's reliability.

Residuals vs Leverage:



This plot illustrates the standardized residuals of a regression model against the leverage of each observation. Leverage indicates how much an observation influences the model's fitted values, and points with high leverage can significantly impact the regression results. The plot reveals a few observations with both high leverage and high residuals, suggesting they might be outliers affecting the model's fit. The horizontal line marks the mean of the standardized residuals, while the curved line shows Cook's distance—a metric for assessing an observation's influence on the model.

Scale-Location:



The plot shows the standardized residuals of a linear regression model plotted against the leverage of each observation. The leverage of an observation indicates its influence on the fitted regression line. The plot shows three clusters of points, with one cluster having a high leverage and a standardized residual close to 1. This suggests that observation has a significant influence on the model, and might be an outlier. The red line in the plot represents the Cook's distance, which measures the overall influence of an observation on the model. The Cook's distance of the high leverage observation is above the threshold.

For y 4

- Logistic Regression:

Call:

glm(formula = y_4 ~ sector + gender + Householdtype + umpce_quintile_factor,

family = binomial(link = logit), data = survey_data_new)

Coefficients:

	Estimate Std.	Error	z value	Pr(>|z|)
(Intercept)	-2.1796	0.2078	-10.489	<2e-16 *
sectorUrban	-0.2676	0.1529	-1.750	0.0802 .
genderMale	-0.2083	0.1264	-1.648	0.0993 .
HouseholdtypeCasual labour in non-agriculture	0.2434	0.2645	0.920	0.3574
Householdtypeothers	0.8448	0.4201	2.011	0.0443 *
HouseholdtypeRegular wage/Salary earning	0.2045	0.2559	0.799	0.4242
HouseholdtypeSelf-employment in agricultur	0.4504	0.2150	2.095	0.0361 *
HouseholdtypeSelf-employment in non-agriculture	0.5157	0.2174	2.372	0.0177 *
umpce_quintile_factor2	-0.1216	0.1827	-0.666	0.5054
umpce_quintile_factor3	0.2229	0.1973	1.130	0.2585
umpce_quintile_factor4	0.3009	0.2007	1.499	0.1339
umpce_quintile_factor5	0.5390	0.2153	2.504	0.0123 *

Signif. codes: 0 ‘*’ 0.001 ‘*’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1764.5 on 2280 degrees of freedom

Residual deviance: 1732.5 on 2269 degrees of freedom

(1188 observations deleted due to missingness)

AIC: 1756.5

Number of Fisher Scoring iterations: 4

- The odds ratio:

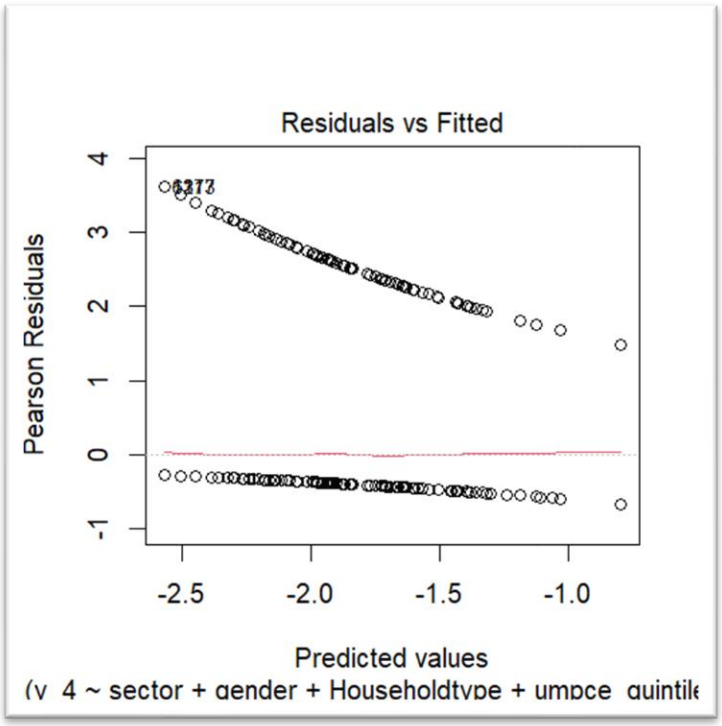
(Intercept)	sectorUrban
0.1130835	0.7652154
genderMale	HouseholdtypeCasual labour in non-agriculture
0.8119744	1.2756403
Householdtypeothers	HouseholdtypeRegular wage/Salary earning
2.3274503	1.2269446
HouseholdtypeSelf-employment	HouseholdtypeSelf-employment
In agriculture	in non-agriculture
1.5689480	1.6747760
umpce_quintile_factor2	umpce_quintile_factor3
0.8854613	1.2497284
umpce_quintile_factor4	umpce_quintile_factor5
1.3510319	1.7142785

- The interpretation of logistic Regression of y_4:

The odds ratios highlight how different factors affect the chances of an outcome. Starting with a very low baseline odds of 0.113, people living in urban areas have a slightly reduced chance of experiencing the outcome compared to those in rural areas (0.765). Men also have a lower likelihood of the outcome (0.812) compared to women. On the other hand, certain types of households—such as those engaged in casual labor in non-agriculture, regular wage/salary jobs, or self-employment (both in agriculture and non-agriculture)—tend to have higher chances of the outcome, with the 'others' household type showing the most significant increase (2.327). Additionally, households with higher monthly per capita expenditures (umpce) are more likely to experience the outcome, with the highest likelihood seen in the top expenditure quintile (1.714) compared to the lowest.

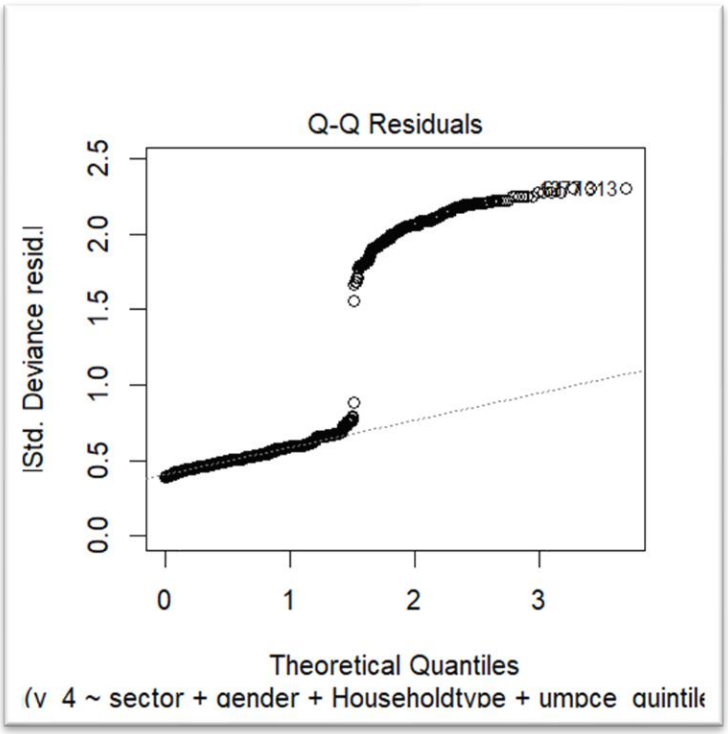
- The plots of logistic Regression of and their interpretation:

Residuals vs Fitted:



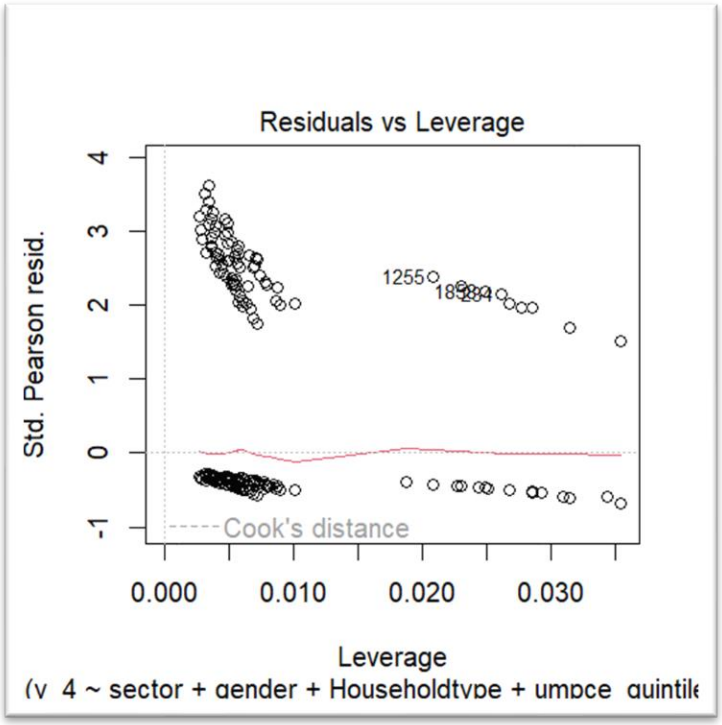
The plot shows how the standardized residuals (the differences between observed and predicted values) relate to leverage (the potential influence of each data point) in a linear regression model. It reveals three distinct groups: one with high leverage and high residuals, another with high leverage but low residuals, and a third with low leverage and low residuals. The line representing Cook's distance helps identify how much each point influences the model. The point with both high leverage and high residuals stands out with the highest Cook's distance, signaling that it's particularly influential. This suggests that this point might be an outlier and could be skewing the results of the regression.

Q-Q Residuals:



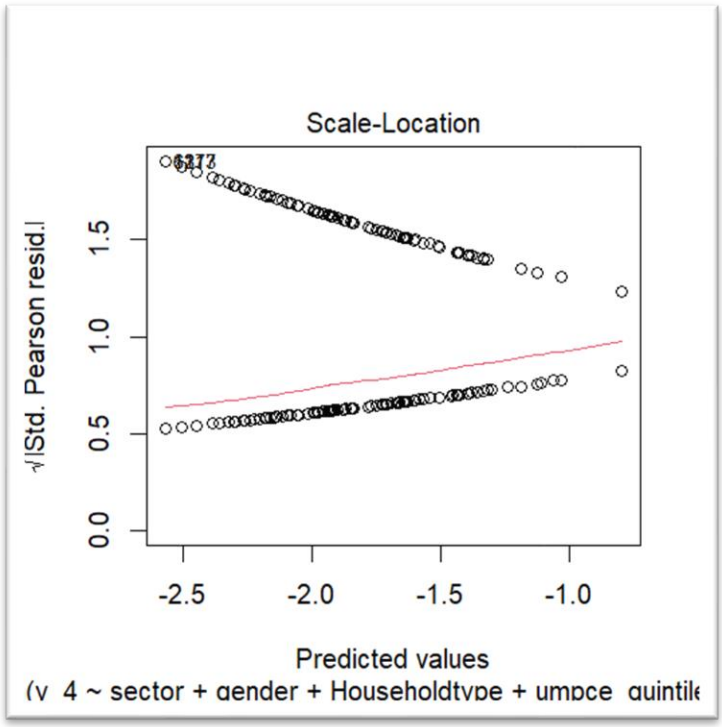
The plot displays how the standardized residuals (the differences between observed and predicted values) relate to each data point's leverage (its potential impact on the regression line). Leverage measures how much a data point can influence the overall fit of the model. Points with high leverage can significantly affect the regression results, especially if they are also outliers. Cook's distance is also shown on the plot, which helps assess the overall influence of each point on the regression results. The plot reveals a few points with high leverage, and one point stands out with a notably high Cook's distance.

Residuals vs Leverage:



The plot shows how the standardized residuals (the differences between what the model predicts and the actual values) relate to each data point's leverage (its potential to influence the regression line). Points with high leverage, which are far to the right on the plot, have a greater impact on the model. Some of these high-leverage points are also outliers, as their residuals are significantly different from zero. The red line on the plot represents Cook's distance, which measures how much removing each data point would change the regression results. Points that are both high in leverage and have large residuals also show high Cook's distances, indicating they significantly influence the model.

Scale-Location:



The plot shows how standardized residuals (the differences between the model's predictions and the actual values) relate to each data point's leverage (its potential to impact the regression line). Points with high leverage have a big influence on the model, while points with large residuals are poorly predicted by it. Point 101080 stands out as an outlier because it has a notably large residual, suggesting it doesn't fit well with the model's predictions. The red line on the plot indicates Cook's distance, a measure of how much each point affects the overall model if it were removed.

For y 5

- Logistic Regression:

Call:

glm(formula = y_5 ~ sector + gender + Householdtype + umpce_quintile_factor,

family = binomial(link = logit), data = survey_data_new)

Coefficients:

	Estimate Std.	Error	z value	Pr(>|z|)
(Intercept)	-2.0877	0.1922	-10.859	< 2e-16 *
sectorUrban	0.2092	0.1395	1.499	0.1338
genderMale	-0.2064	0.1163	-1.776	0.0758 .
HouseholdtypeCasual labour in non-agriculture	-0.5497	0.2742	-2.004	0.0450 *
Householdtypeothers	-0.4712	0.5090	-0.926	0.3545
HouseholdtypeRegular wage/Salary earning	-0.3802	0.2448	-1.553	0.1204
HouseholdtypeSelf-employment in agricultur	0.4457	0.1918	2.324	0.0201 *
HouseholdtypeSelf-employment in non-agriculture	0.1560	0.1997	0.781	0.4346
umpce_quintile_factor2	0.1214	0.1688	0.719	0.4720
umpce_quintile_factor3	0.1597	0.1939	0.824	0.4101
umpce_quintile_factor4	1.0350	0.1790	5.783	7.34e-09 *
umpce_quintile_factor5	0.9194	0.2059	4.465	8.01e-06 *

Signif. codes: 0 ‘*’ 0.001 ‘*’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2038.7 on 2280 degrees of freedom

Residual deviance: 1947.6 on 2269 degrees of freedom

(1188 observations deleted due to missingness)

AIC: 1971.6

Number of Fisher Scoring iterations: 4

- The odds ratio:

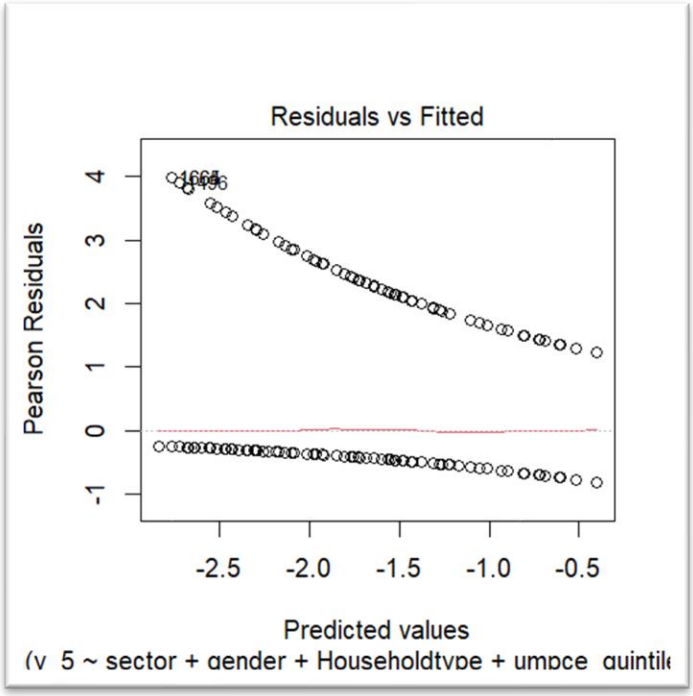
(Intercept)	sectorUrban
0.1239752	1.2327371
genderMale	HouseholdtypeCasual labour in non-agriculture
0.8134795	0.5771471
Householdtypeothers	HouseholdtypeRegular wage/Salary earning
0.6242374	0.6837315
HouseholdtypeSelf-employment In agriculture	HouseholdtypeSelf-employment in non-agriculture
1.5616376	1.1688228
umpce_quintile_factor2	umpce_quintile_factor3
1.1290815	1.1731763
umpce_quintile_factor4	umpce_quintile_factor5
2.8151149	2.5077922

- The interpretation of logistic Regression of y_5:

The odds ratios reveal how different factors affect the likelihood of an outcome. Starting from a low baseline odds of 0.124, the chances of the outcome are initially quite small. People living in urban areas have a greater likelihood of the outcome compared to those in rural areas (1.233). Interestingly, men are somewhat less likely to experience the outcome compared to women (0.813). Households engaged in casual labor in non-agriculture have reduced odds (0.577), and those in 'others' (0.624) or regular wage/salary jobs (0.684) also face lower odds. On the flip side, self-employed individuals, whether in agriculture (1.562) or non-agriculture (1.169), have higher odds of the outcome. Additionally, spending more monthly per capita is associated with increased odds: the second quintile has odds of 1.129, the third is 1.173, the fourth jumps to 2.815, and the fifth is 2.508, suggesting that higher expenditure generally means a greater likelihood of the outcome.

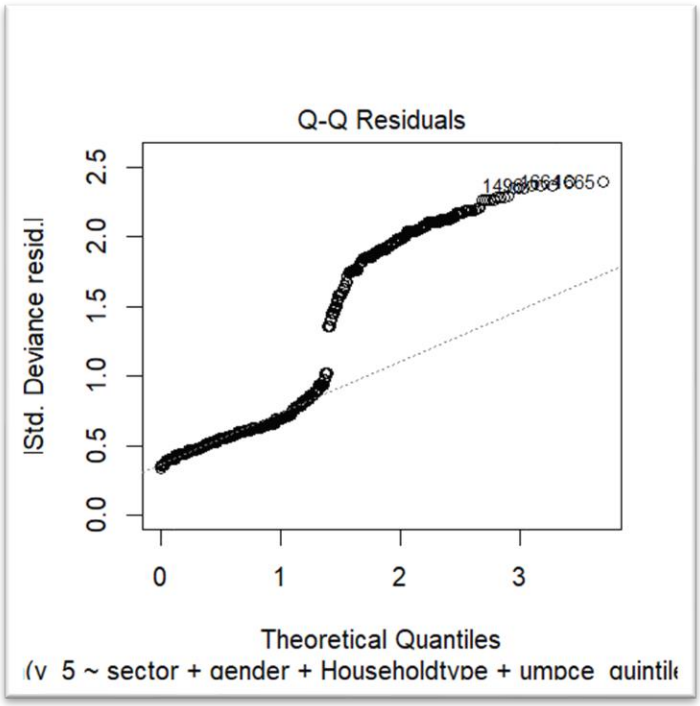
- The plots of logistic Regression of and their interpretation:

Residuals vs Fitted:



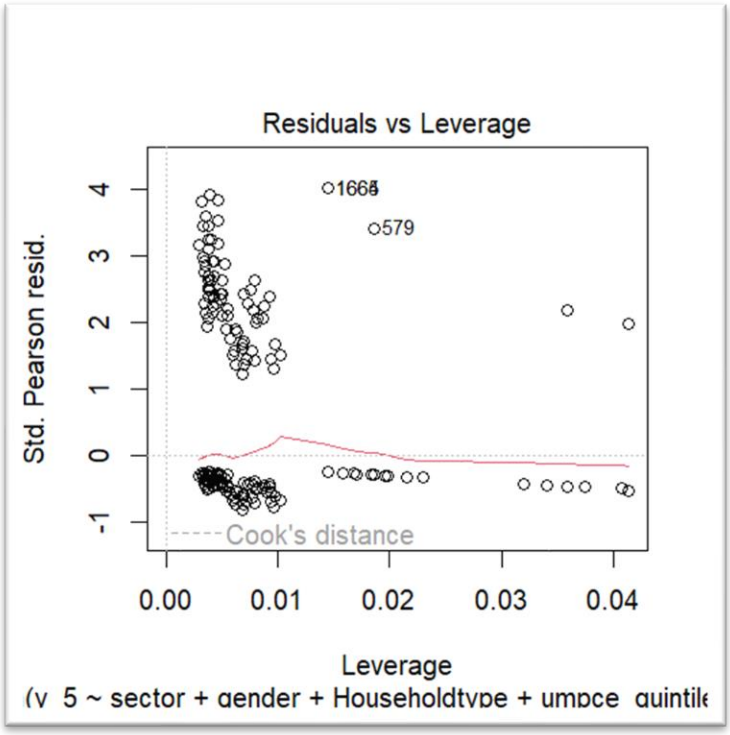
The plot displays how residuals (the differences between actual and predicted values) relate to leverage (how much each observation affects the regression line). Observations with high leverage can significantly influence the model's fit, while those with high residuals are not well predicted by the model. The plot reveals a few points that stand out with both high leverage and high residuals, suggesting they might be outliers that are impacting the model's accuracy. The red Cook's distance line on the plot helps identify potentially influential observations. Points above this line have a notable influence on the model.

Q-Q Residuals:



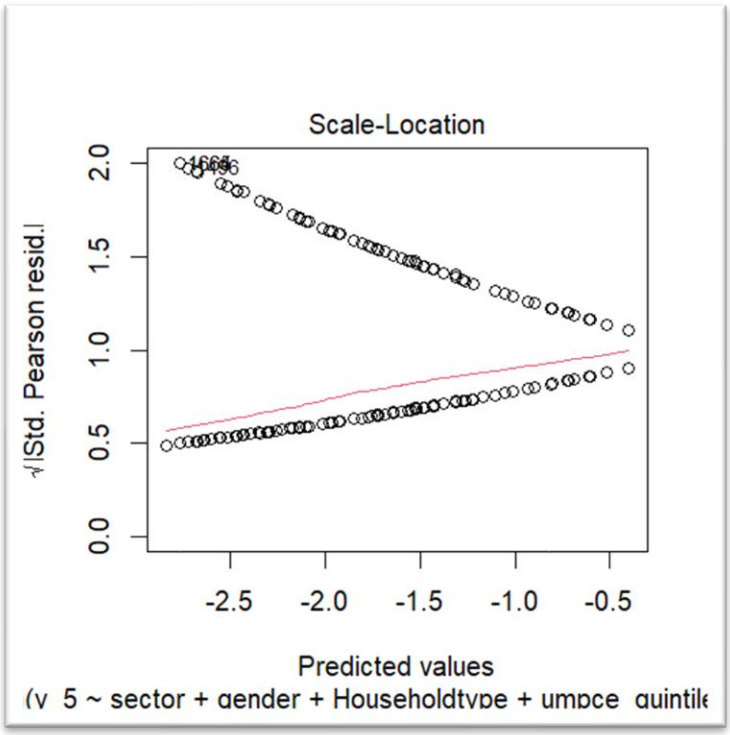
This plot shows the residuals (how far off the model's predictions are from the actual values) versus leverage (how much each data point can influence the regression line), with a line indicating Cook's distance. It's a useful tool for spotting potential outliers or influential points that could skew the results. The plot helps us see if there are any points with high leverage (unusual predictor values), large residuals (big deviations from the regression line), or both. In this case, while there are a few points with high leverage, they don't have extremely large residuals. Moreover, Cook's distance is below the critical threshold for all points, meaning none are highly influential.

Residuals vs Leverage:



This plot shows standardized residuals (how much each prediction deviates from the actual values) versus leverage (how much influence each data point has on the regression line). It's a key tool for spotting outliers or points that might be skewing the results. The red line represents Cook's distance, which helps measure the impact of each observation on the model. Points significantly far from this line are considered influential. In this plot, observation 101080 stands out with a high Cook's distance, suggesting it might be an outlier. Removing or further investigating this point could help improve the model's accuracy and overall fit.

Scale-Location:



The plot shows standardized residuals (the differences between the model's predictions and the actual values) against leverage (how much each data point can affect the model's fit). It's used to see how individual points influence the model. In this case, some points with high leverage are noticeable on the right side, meaning they have a significant impact on the regression line. However, these points aren't outliers because they're not far from the fitted line. The Cook's distance line, which measures each point's overall influence, indicates that while a few points have some leverage, none have a major impact on the model. Overall, it seems the model is stable and not heavily influenced by any particular data point.

- India:

  For y_1

  - Logistic Regression:

Call:

glm(formula = y_1 ~ sector + gender + Householdtype + umpce_quintile,

family = binomial(link = logit), data = survey_data_new)

Coefficients:

	Estimate Std.	Error	z value	Pr(>|z|)
(Intercept)	-0.345425	0.041946	-8.235	< 2e-16 *
sectorUrban	0.603098	0.029082	20.738	< 2e-16 *
genderMale	0.527627	0.024210	21.794	< 2e-16 *
HouseholdtypeCasual labour in non-agriculture	-0.233677	0.047729	-4.896	9.79e-07 *
Householdtypeothers	0.284941	0.081686	3.488	0.000486 *
HouseholdtypeRegular wage/Salary earning	0.157532	0.049072	3.210	0.001326 **
HouseholdtypeSelf-employment in agricultur	0.341223	0.040748	8.374	< 2e-16 *
HouseholdtypeSelf-employment in non-agriculture	0.582643	0.048881	11.920	< 2e-16 *
umpce_quintile	0.449814	0.009581	46.951	< 2e-16 *

Signif. codes: 0 '*' 0.001 '*' 0.01 '' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 49030 on 51656 degrees of freedom

Residual deviance: 44735 on 51648 degrees of freedom

AIC: 44753

Number of Fisher Scoring iterations: 5

- Odds Ratio:

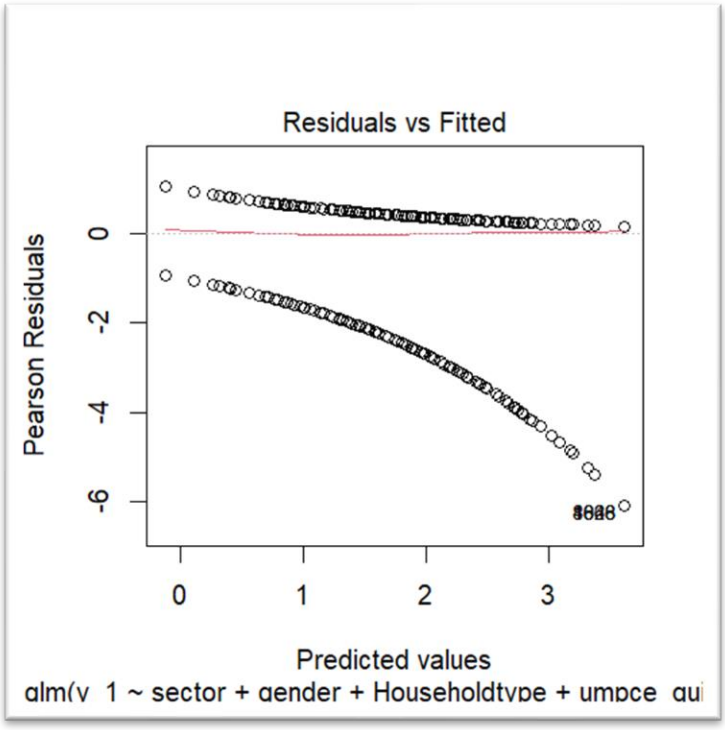
(Intercept)	sectorUrban
0.7079196	1.8277724
genderMale	HouseholdtypeCasual labour in non-agriculture
1.6949056	0.7916175
Householdtypeothers	HouseholdtypeRegular wage/Salary earning
1.3296834	1.1706180
HouseholdtypeSelf-employment in agriculture	HouseholdtypeSelf-employment in non-in agricultur
1.4066663	1.7907649
umpce_quintile	
1.5680204	

- The interpretation of logistic Regression of y_1:

The odds ratios shed light on how different factors affect the likelihood of the outcome. Starting from a baseline odds of 0.708, the initial probability of the outcome is relatively low. Urban residents have a higher likelihood of experiencing the outcome compared to those in rural areas (1.828). Men also have an increased chance of the outcome compared to women (1.695). Households engaged in casual labor in non-agriculture have lower odds (0.792), while those in the 'others' category (1.330) and regular wage/salary positions (1.171) have slightly higher odds. Self-employed individuals show a substantial increase in odds, with those in agriculture (1.407) and non-agriculture (1.791) both having notably higher likelihoods of the outcome. Higher monthly per capita expenditure (umpce) is associated with increased odds (1.568), suggesting that greater expenditure is linked to a higher chance of the outcome.

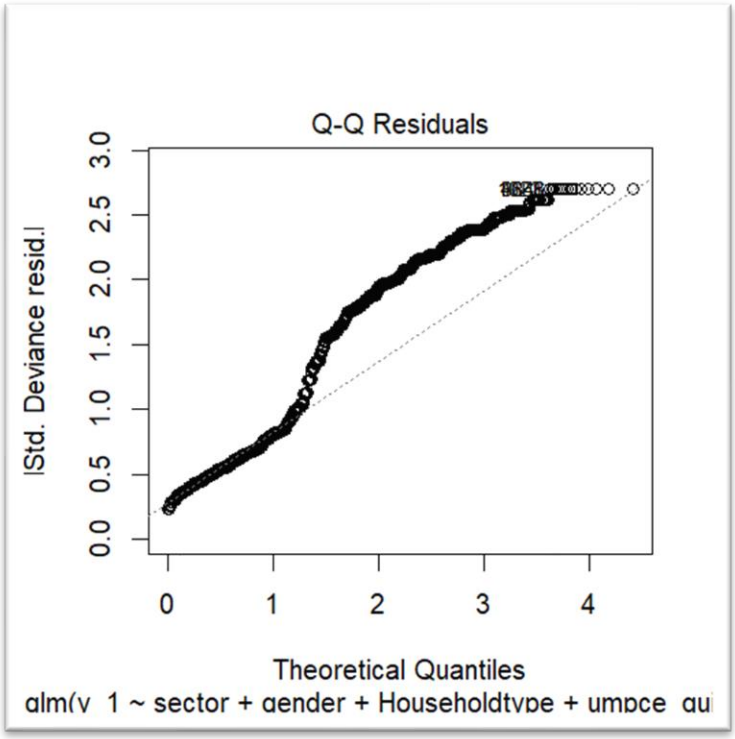
- The plots of logistic Regression of and their interpretation:

Residuals vs Fitted:



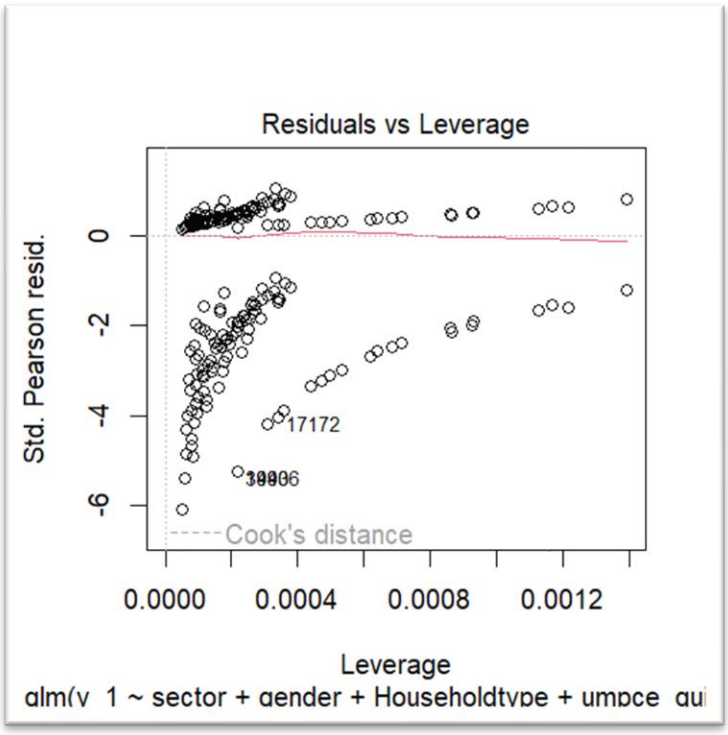
This plot shows how the standardized residuals (the differences between the model's predictions and actual values) relate to each observation's leverage (how much influence it has on the regression line). Some points, like observation 101080, stand out with high leverage and a high residual, suggesting they might be outliers with a significant impact on the model. The red line represents Cook's distance, which measures how much the model would change if a particular observation were removed. Observation 101080's high Cook's distance indicates it could substantially affect the regression results.

Q-Q Residuals:



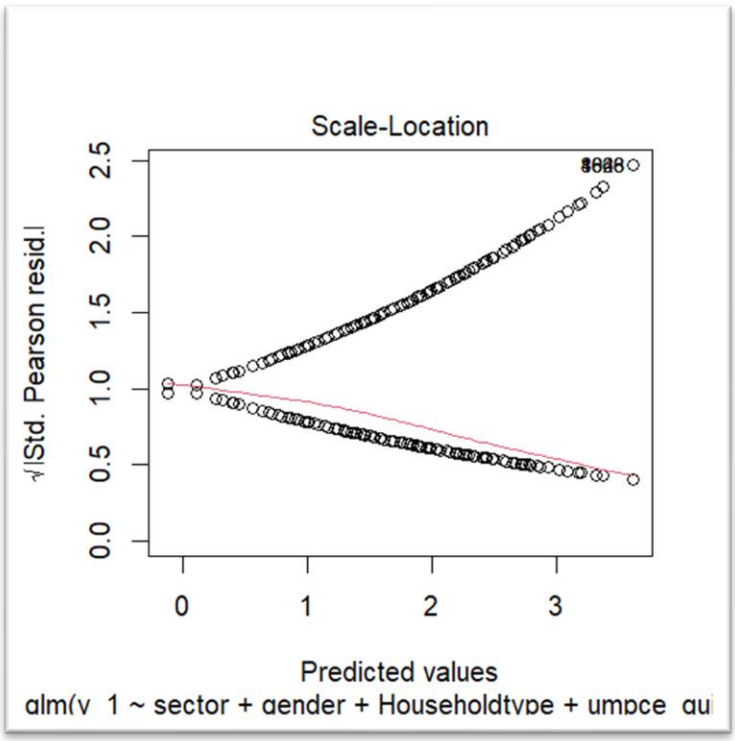
The plot displays how standardized residuals (the differences between the model's predictions and the actual values, adjusted for scale) compare to leverage (how much each data point affects the regression line). It highlights several points with high leverage, including one labeled "101080," which might be exerting a significant influence on the model. These high-leverage points could potentially be outliers. The red Cook's distance line helps identify points with a strong impact on the regression model. While the plot indicates some points might be influential, further investigation is needed to confirm whether they are outliers and to understand how they might be affecting the model.

Residuals vs Leverage:



This plot compares standardized residuals (how far off the model's predictions are) with leverage (how much each data point can influence the regression line). Some points, including one marked "101080," have high leverage, which means they could have a big impact on the model. However, the Cook's distance line shows that these points aren't particularly influential, suggesting that while there may be some outliers, they don't seem to be distorting the model significantly. This plot is helpful for spotting potential outliers and influential points, but it's just one piece of the puzzle. Understanding the context of the data and the goals of our analysis is crucial for interpreting these results accurately.

Scale-Location:



This plot shows how standardized residuals (the differences between predicted and actual values) relate to leverage (how much each data point affects the regression line). Leverage helps us understand the influence of individual data points on the model. The plot reveals a few points with high leverage, notably around 0.025 on the x-axis, but the majority of data points have lower leverage. This means the model isn't overly swayed by just a few outliers; rather, it's primarily shaped by the overall dataset. The Cook's distance line, which measures each point's overall impact, indicates that none of the points have a particularly high Cook's distan

For y_2

- Logistic Regression:

Call:

glm(formula = y_2 ~ sector + gender + Householdtype + umpce_quintile_factor,

family = binomial(link = logit), data = survey_data_new)

Coefficients:

	Estimate Std.	Error	z value	Pr(>|z|)
(Intercept)	-1.95824	0.08276	-23.662	< 2e-16 *
sectorUrban	-0.16047	0.06837	-2.347	0.018913 *
genderMale	-0.04530	0.05405	-0.838	0.401921
HouseholdtypeCasual	-0.08155	0.09883	-0.825	0.409236
labour in non-agriculture				
Householdtypeothers	-0.46417	0.20554	-2.258	0.023924 *
HouseholdtypeRegular	-0.41041	0.10577	-3.880	0.000104 *
wage/Salary earning				
HouseholdtypeSelf-	-0.59871	0.08486	-7.055	1.72e-12 *
employment in agricultur				
HouseholdtypeSelf-	-0.77194	0.10556	-7.313	2.62e-13 *
employment in non-agriculture-				
umpce_quintile_factor2	-0.31553	0.06953	-4.538	5.67e-06 *
umpce_quintile_factor3	-0.48597	0.08041	-6.044	1.50e-09 *
umpce_quintile_factor4	-0.55389	0.08442	-6.561	5.34e-11 *
umpce_quintile_factor5	-1.16907	0.12242	-9.550	< 2e-16 *

Signif. codes: 0 '*' 0.001 '*' 0.01 '' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 11413 on 27073 degrees of freedom

Residual deviance: 11151 on 27063 degrees of freedom

(24582 observations deleted due to missingness)

AIC: 11175

Number of Fisher Scoring iterations: 6

- Odds Ratio:

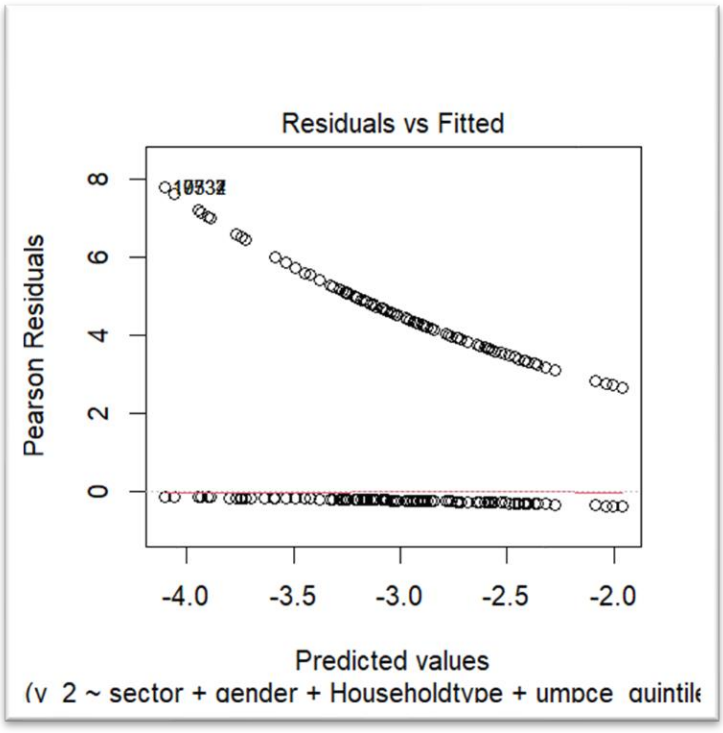
(Intercept)	sectorUrban
0.1411060	0.8517434
genderMale	HouseholdtypeCasual labour in non-agriculture
0.9557064	0.9216821
Householdtypeothers	HouseholdtypeRegular wage/Salary earning
0.6286567	0.6633764
HouseholdtypeSelf-employment in agriculture	HouseholdtypeSelf-employment in non-in agricultur
0.5495179	0.4621138
umpce_quintile_factor2	umpce_quintile_factor3
0.7294057	0.6150976
umpce_quintile_factor4	umpce_quintile_factor5
0.5747111	0.3106554

- The interpretation of logistic Regression of y_2:

The odds ratios provide insight into how different factors influence the likelihood of the outcome. With a baseline odds of 0.141, the initial probability of the outcome is quite low. Urban residents have a slightly lower likelihood of the outcome compared to those in rural areas (0.852). Men have a similar likelihood to women (0.956), and households engaged in casual labor in non-agriculture have odds close to 1 (0.922). Those in the 'others' category or regular wage/salary jobs also have reduced odds (0.629 and 0.663, respectively). Self-employed individuals show notably lower odds, whether in agriculture (0.550) or non-agriculture (0.462). Higher monthly per capita expenditure (umpce) is associated with decreased odds across all quintiles, with the lowest odds observed in the fifth quintile (0.311), suggesting that higher expenditure correlates with a lower likelihood of the outcome.

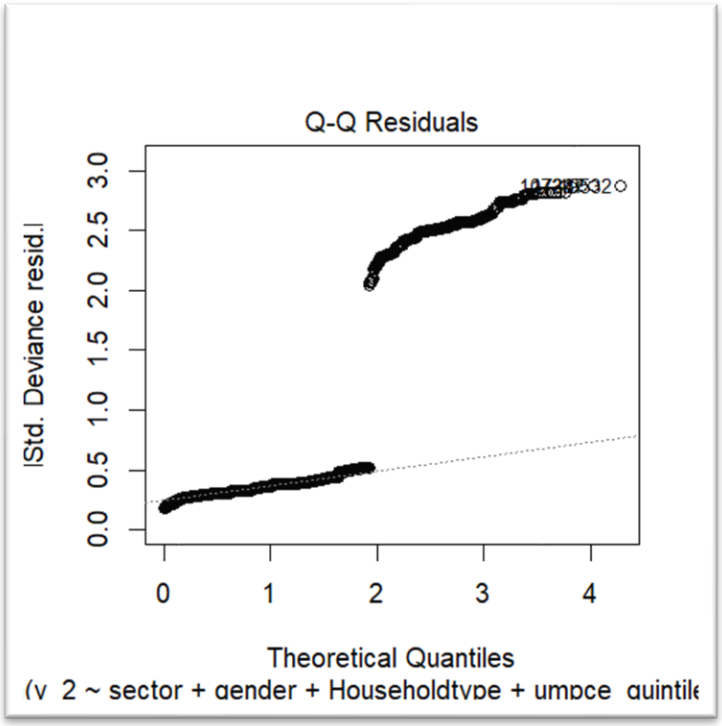
- The plots of logistic Regression of and their interpretation:

Residuals vs Fitted:



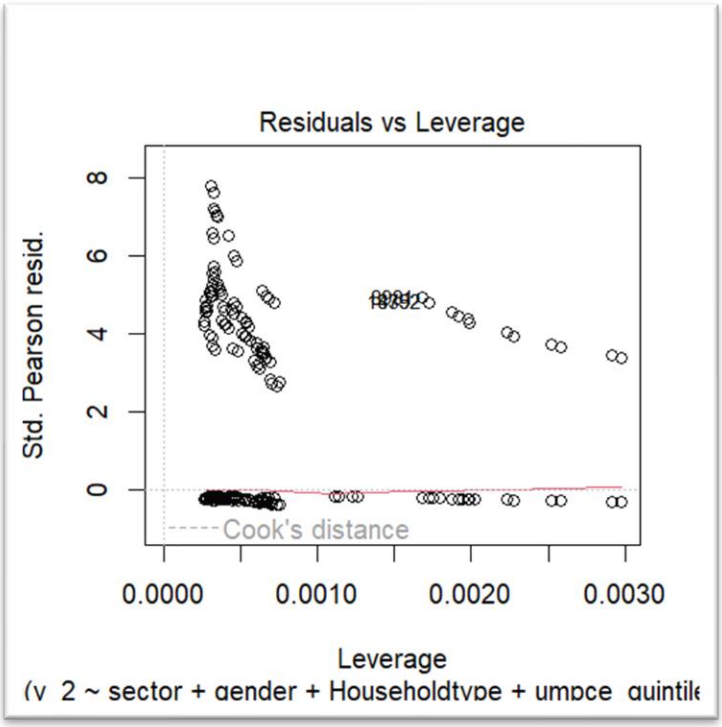
This plot shows how standardized residuals (the differences between the model's predictions and actual values) relate to leverage (how much each data point can influence the regression line). High leverage points can significantly impact the model, and the plot highlights several such observations. One observation stands out with a large standardized residual, indicating it might be having a strong influence on the model's results. Additionally, there's a slight trend in the residuals, suggesting the model might not be capturing all the underlying relationships in the data.

Q-Q Residuals:



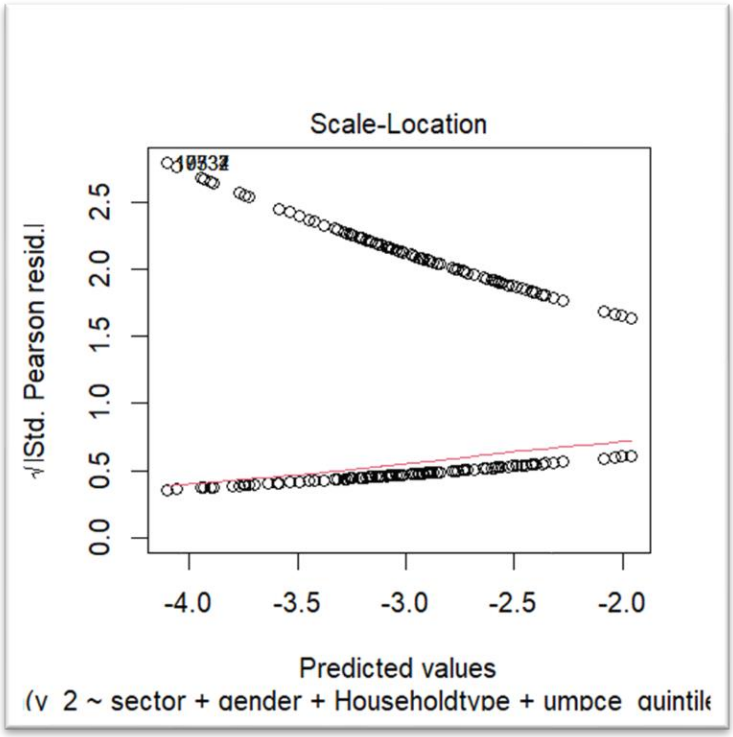
This plot displays how standardized residuals (the differences between actual and predicted values) compare to leverage (the impact each data point has on the regression line). High leverage points can significantly sway the regression results, even if their residuals are not very large. The plot reveals a few observations with high leverage, including one marked "101080." Additionally, there's a slight curve in the plot, hinting that the relationship between residuals and leverage might not be perfectly linear, suggesting the model could be missing some underlying patterns. The dotted line represents Cook's distance, which measures the overall influence of each observation on the regression.

Residuals vs Leverage:



This plot shows how standardized residuals (the differences between observed and predicted values) relate to leverage (how much each data point influences the regression line). It highlights a few points with high leverage or large residuals, suggesting they might be outliers. The red line represents Cook's distance, which measures how much each data point affects the overall regression results. Points with high Cook's distance are particularly influential and could significantly impact the model. High leverage points, especially those with large residuals, are likely to be outliers and could distort the model's accuracy.

Scale-Location:



The plot displays standardized residuals against leverage for a linear regression model, showing how much each observation influences the regression line. High leverage points can heavily affect the model, and the plot also includes Cook's distance, which gauges the overall influence of each observation on the model. Observations with high Cook's distance are particularly influential and should be examined more closely. In this plot, one observation stands out with both high leverage and high influence, indicating it might be an outlier. It's important to check whether this point is due to an error in data collection or if it's a valid observation. If it's an error, correcting it could improve the model.

For y_3

- Logistic Regression:

Call:

glm(formula = y_3 ~ sector + gender + Householdtype + umpce_quintile_factor,

family = binomial(link = logit), data = survey_data_new)

Coefficients:

	Estimate Std.	Error	z value	Pr(>|z|)
(Intercept)	-0.27238	0.04794	-5.682	1.33e-08 *
sectorUrban	-0.07486	0.03202	-2.338	0.0194 *
genderMale	-0.04571	0.02668	-1.714	0.0866 .
HouseholdtypeCasual labour in non-agriculture	-0.10255	0.05742	-1.786	0.0741
Householdtypeothers	-0.62632	0.10903	-5.745	9.21e-09 *
HouseholdtypeRegular wage/Salary earning	-0.21788	0.05598	-3.892	9.95e-05 *
HouseholdtypeSelf-employment in agricultur	-0.37301	0.04676	-7.978	1.49e-15 *
HouseholdtypeSelf-employment in non-agriculture	-0.39495	0.05360	-7.368	1.73e-13 *
umpce_quintile_factor2	-0.16013	0.03632	-4.409	1.04e-05 *
umpce_quintile_factor3	-0.27094	0.04022	-6.737	1.61e-11 *
umpce_quintile_factor4	-0.36985	0.04210	-8.785	< 2e-16 *
umpce_quintile_factor5	-0.58796	0.04976	-11.816	< 2e-16 *

Signif. codes: 0 ‘*’ 0.001 ‘*’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 33180 on 27074 degrees of freedom

Residual deviance: 32855 on 27063 degrees of freedom

(24582 observations deleted due to missingness)

AIC: 32879

Number of Fisher Scoring iterations: 4

- Odds Ratio:

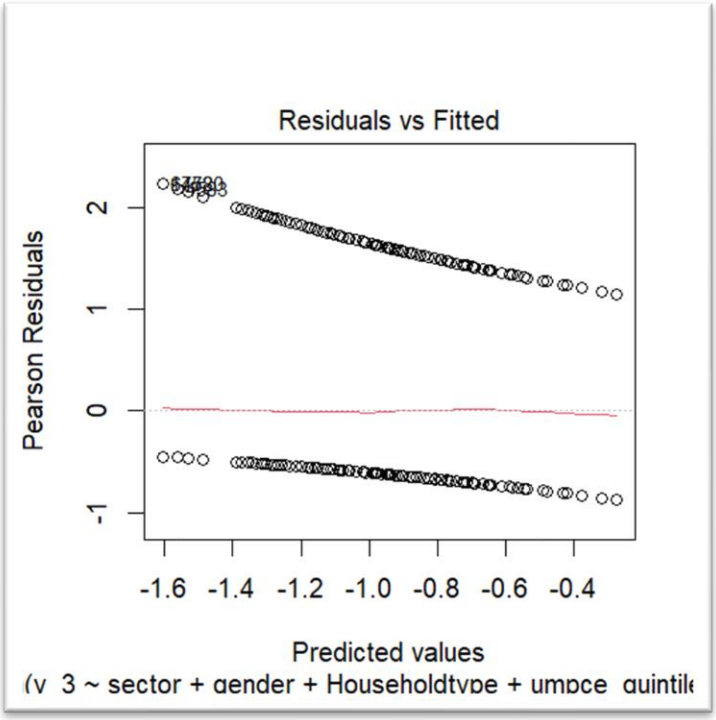
(Intercept)	sectorUrban
0.7615624	0.9278747
genderMale	HouseholdtypeCasual labour in non-agriculture
0.9553142	0.9025357
Householdtypeothers	HouseholdtypeRegular wage/Salary earning
0.5345566	0.8042241
HouseholdtypeSelf-employment in agriculture	HouseholdtypeSelf-employment in non-in agricultur
0.6886609	0.6737147
umpce_quintile_factor2	umpce_quintile_factor3
0.8520298	0.7626586
umpce_quintile_factor4	umpce_quintile_factor5
0.6908373	0.5554594

- The interpretation of logistic Regression of y_3:

The odds ratios reveal how different factors influence the chance of a certain outcome. With a starting odds of 0.762, the baseline likelihood of the outcome is quite low. Urban residents are slightly less likely to experience the outcome compared to those in rural areas, with odds of 0.928. Men have odds of 0.955, showing their likelihood is comparable to women's. Households engaged in casual labor in non-agriculture have lower odds (0.903), and those in other categories, like 'others' (0.535) and regular wage/salary jobs (0.804), also face reduced odds. Self-employed individuals, whether in agriculture (0.690) or non-agriculture (0.674), similarly have lower odds. Interestingly, higher monthly per capita expenditure (umpce) seems to be linked with a decreased likelihood of the outcome, with the odds dropping across higher quintiles: the second (0.852), third (0.763), fourth (0.691), and fifth (0.555). This suggests that as expenditure increases, the probability of the outcome decreases.

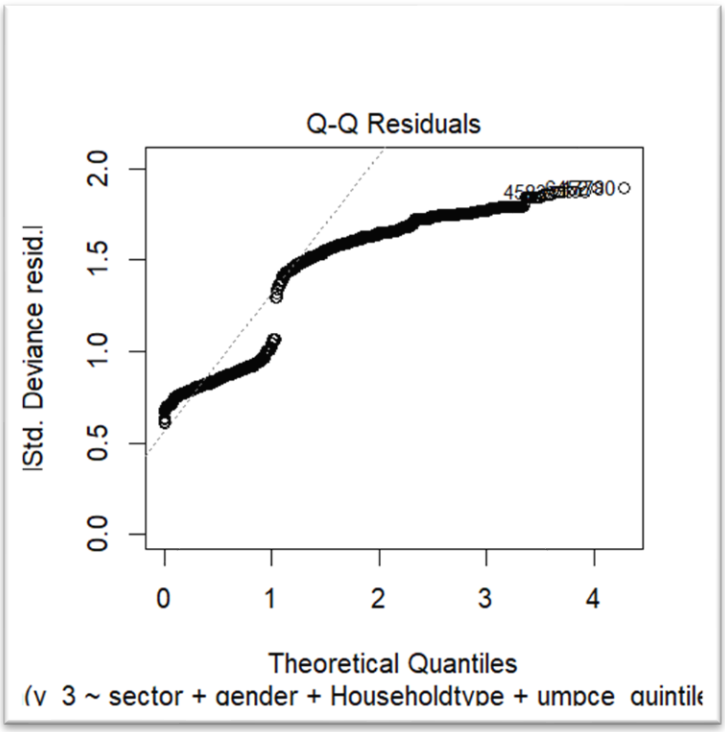
- The plots of logistic Regression of and their interpretation:

Residuals vs Fitted:



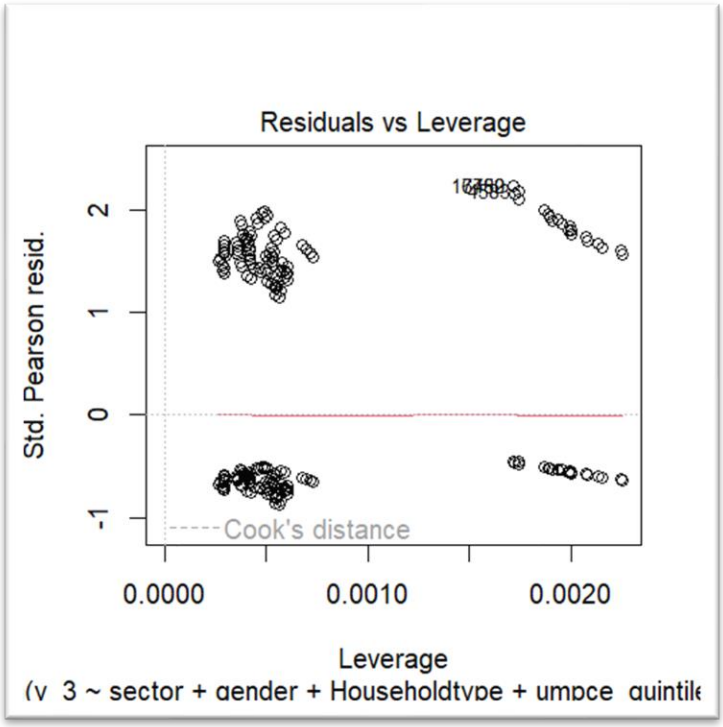
The plot shows the relationship between standardized residuals and leverage in a linear regression model. Leverage measures how much influence each observation has on the fitted regression line based on its predictor values, while standardized residuals show how far each observation's actual response is from what the model predicts, adjusted for data variability. The plot reveals several points with high leverage, meaning they have unusual predictor values. However, these points don't have unusually large residuals, indicating they aren't significantly affecting the model's fit. Despite this, it's wise to take a closer look at these high-leverage points.

Q-Q Residuals:



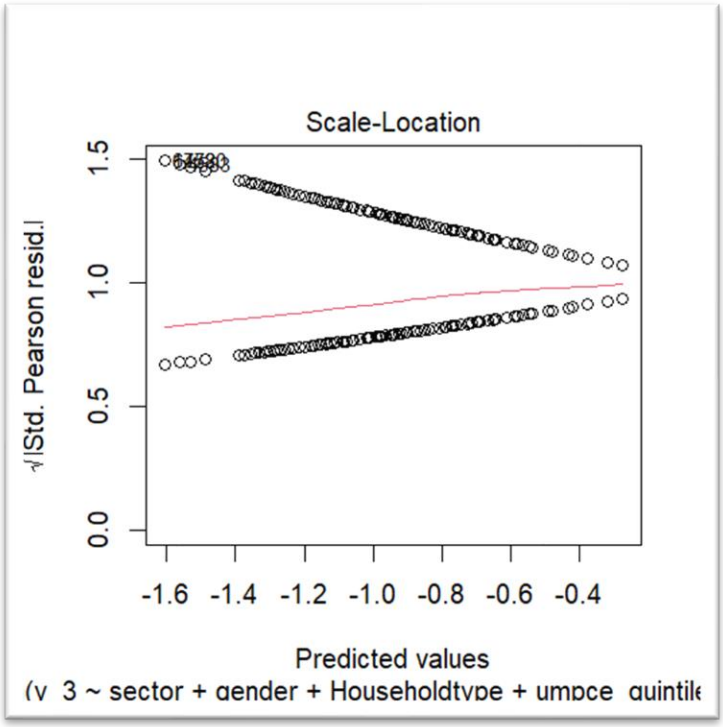
The plot displays the standardized residuals against leverage for each data point in our model. Leverage indicates how much a particular point influences the regression line, while standardized residuals show how far off each point is from the model's predictions. The Cook's distance line helps identify influential points by marking a threshold for impact. The plot reveals three points with high leverage that exceed this threshold, suggesting they might be outliers with a significant influence on the model. It's a good idea to investigate these points further to understand their effect on the model and consider whether they should be excluded or transformed to improve the model's accuracy.

Residuals vs Leverage:



The plot displays the standardized residuals versus leverage for the regression model, helping identify points that might significantly impact the model's fit. We can see a few points with high leverage in the upper right corner, meaning they have a strong influence on the regression line. The red line represents Cook's distance, which measures how much each point affects the model. Points that are both high in leverage and Cook's distance have a substantial impact on the results. This suggests that these points could be affecting the model more than others, and it might be worth adjusting the model to address their influence.

Scale-Location:



The plot shows how standardized residuals relate to leverage for each data point in the model. Leverage measures how much a data point can sway the regression line, with high leverage points potentially having a big impact on the model's results. In the plot, we will notice a few points with high leverage, especially in the upper right corner, which could be influencing the model significantly. The dashed line represents Cook's distance, which helps identify how much each data point affects the model as a whole. Points with large Cook's distances are considered influential and might be outliers. Given that some points are flagged as influential, it's a good idea to take a closer look at these observations.

For y_4

- Logistic Regression:

Call:

glm(formula = y_4 ~ sector + gender + Householdtype + umpce_quintile_factor,

family = binomial(link = logit), data = survey_data_new)

Coefficients:

	Estimate Std.	Error	z value	Pr(>|z|)
(Intercept)	-2.20125	0.07012	-31.392	< 2e-16 *
sectorUrban	0.07010	0.04164	1.684	0.09227 .
genderMale	0.16066	0.03542	4.536	5.72e-06 *
HouseholdtypeCasual labour in non-agriculture	-0.04781	0.08305	-0.576	0.56482
Householdtypeothers	0.23376	0.13090	1.786	0.07413 .
HouseholdtypeRegular wage/Salary earning	0.11522	0.07725	1.492	0.13583
HouseholdtypeSelf-employment in agricultur	0.02284	0.06623	0.345	0.73022
HouseholdtypeSelf-employment in non-agriculture	0.14507	0.07344	1.975	0.04822 *
umpce_quintile_factor2	0.13701	0.05181	2.644	0.00819 **
umpce_quintile_factor3	0.27918	0.05490	5.085	3.67e-07 *
umpce_quintile_factor4	0.31722	0.05642	5.623	1.88e-08 *
umpce_quintile_factor5	0.44222	0.06169	7.169	7.57e-13 *

Signif. codes: 0 ‘*’ 0.001 ‘*’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 21767 on 27074 degrees of freedom

Residual deviance: 21660 on 27063 degrees of freedom

(24582 observations deleted due to missingness)

AIC: 21684

Number of Fisher Scoring iterations: 4

- Odds Ratio:

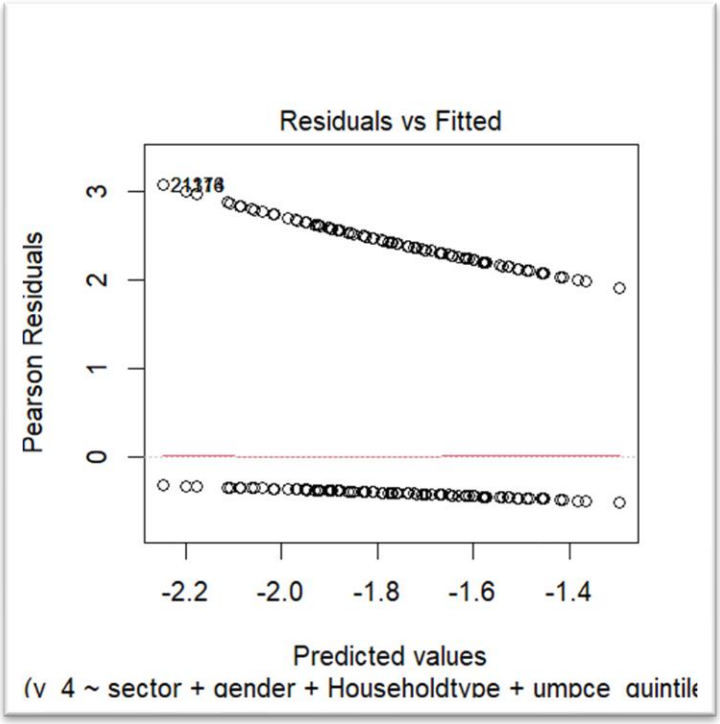
(Intercept)	sectorUrban
0.1106648	1.0726170
genderMale	HouseholdtypeCasual labour in non-agriculture
1.1742880	0.9533159
Householdtypeothers	HouseholdtypeRegular wage/Salary earning
1.2633355	1.1221252
HouseholdtypeSelf-employment in agriculture	HouseholdtypeSelf-employment in non-in agriculture
1.0231013	1.1561165
umpce_quintile_factor2	umpce_quintile_factor3
1.1468371	1.3220433
umpce_quintile_factor4	umpce_quintile_factor5
1.3733044	1.5561565

- The interpretation of logistic Regression of y_4:

The odds ratios shed light on how various factors affect the likelihood of the outcome. With a starting baseline odds of 0.111, the chances of the outcome are relatively low. Urban residents have a slightly higher likelihood of the outcome compared to those living in rural areas, with an odds ratio of 1.073. Men also have a greater chance of experiencing the outcome (1.174) compared to women. For households involved in casual labor outside of agriculture, the odds are similar to the baseline (0.953). However, those in 'others' categories or with regular wage/salary jobs face slightly higher odds (1.263 and 1.122, respectively). Self-employed individuals in non-agriculture have somewhat higher odds (1.156), while those in agriculture have odds similar to the baseline (1.023). Notably, higher monthly per capita expenditure (umpce) is associated with increased odds of the outcome, with higher odds seen in the second quintile (1.147), third (1.322), fourth (1.373), and fifth (1.556), suggesting that greater expenditure tends to correlate with a higher likelihood of the outcome.

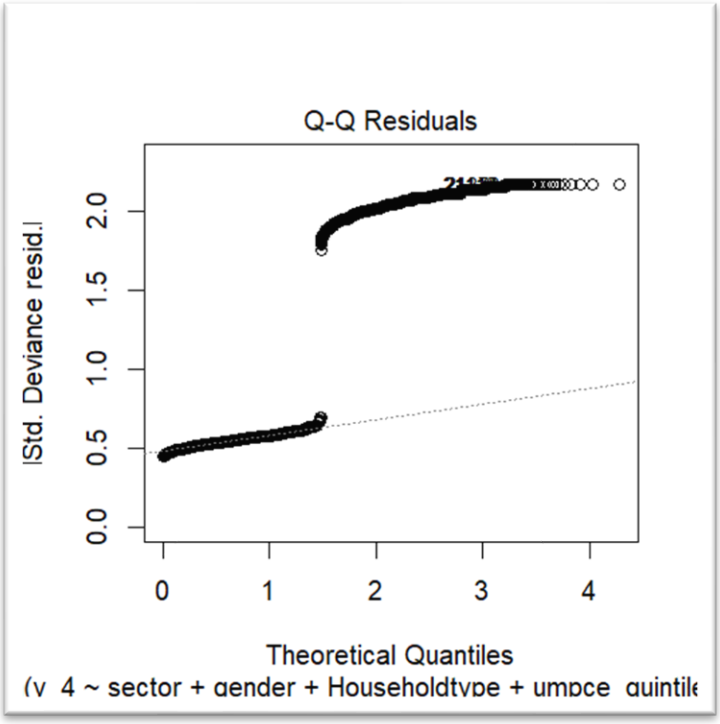
- The plots of logistic Regression of and their interpretation:

Residuals vs Fitted:



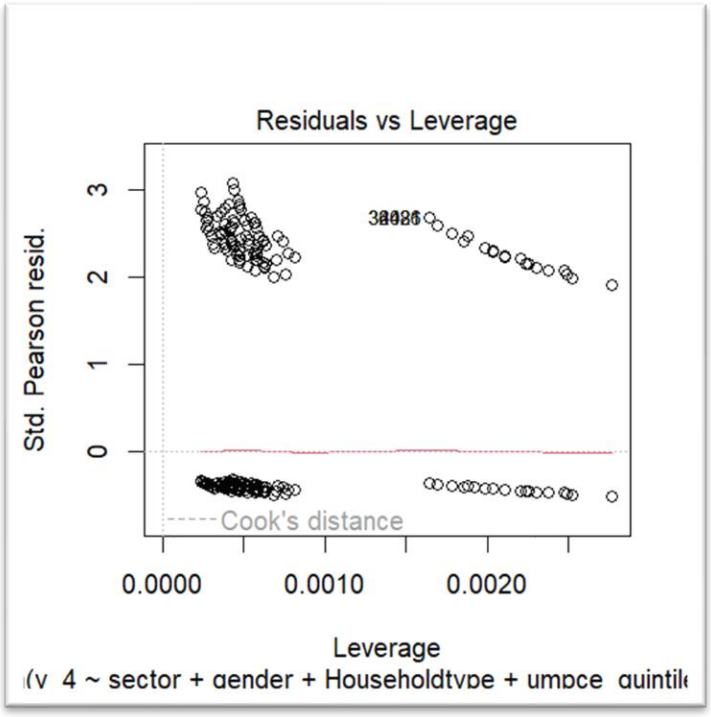
The plot displays the standardized residuals of a regression model plotted against the leverage of each data point. Leverage measures how much an observation can influence the fitted values of the model. The red line on the plot indicates Cook's distance, which helps assess the overall impact of each observation on the model. When an observation has both high leverage and large residuals, it can be particularly influential, potentially distorting the model's results. In this plot, several points stand out with high leverage or large residuals, including one notable point labeled 101080. This suggests these observations could be significantly affecting the model.

Q-Q Residuals:



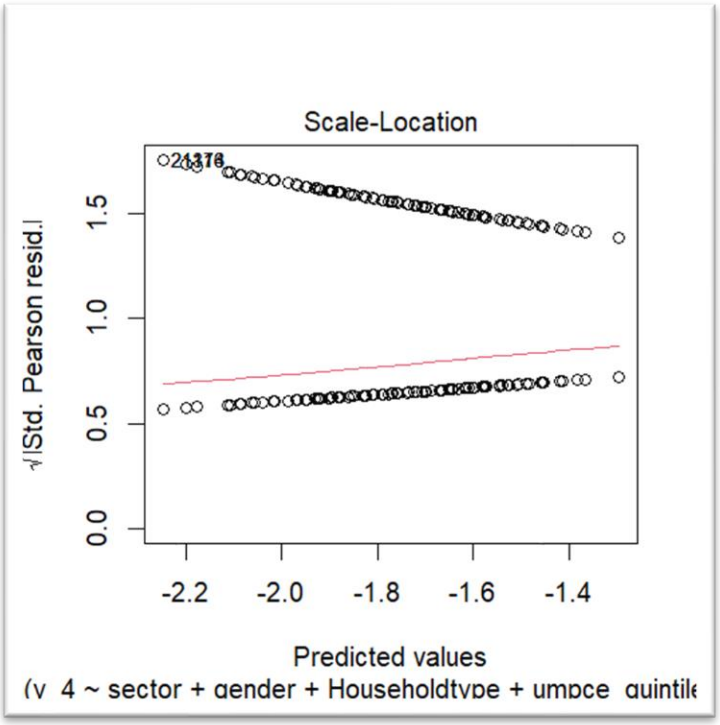
It looks like the Q-Q plot of the model's residuals isn't following the normal distribution, and there are some outliers that stand out. This hints that our current model might not be fully capturing the complexities of our data. We might want to consider adding more variables to better explain the variability, or perhaps exploring different modeling approaches to improve the fit.

Residuals vs Leverage:



The Q-Q plot compares the residuals from our model to what we'd expect if they were normally distributed. In our plot, the points don't align with the diagonal line, which means our residuals have heavier tails than a normal distribution. This could mean that our model isn't fitting the data as well as it could. Additionally, the presence of a few outliers—points that are far from the main group—indicates that the model might be missing some important variability in the data. We might need to refine our model by including additional variables or trying a different modeling approach to get a better fit.

Scale-Location:



The Q-Q plot reveals that our residuals aren't following a normal distribution; the points stray from the straight line, especially at the extremes. This indicates that our model might not be capturing the data well. The plot also highlights several outliers—points that deviate significantly from the line—which suggests that the model might not be accounting for all the variability in our data. The lack of normality could be due to several factors, such as outliers, non-linear relationships between variables, or varying levels of variance. It might be worth revisiting our model to address these issues.

For y_5

- Logistic Regression:

Call:

glm(formula = y_5 ~ sector + gender + Householdtype + umpce_quintile_factor,

family = binomial(link = logit), data = survey_data_new)

Coefficients:

	Estimate Std.	Error	z value	Pr(>|z|)
(Intercept)	-2.31478	0.06778	-34.152	< 2e-16 *
sectorUrban	0.26303	0.03449	7.625	2.44e-14 *
genderMale	-0.02843	0.02999	-0.948	0.343
HouseholdtypeCasual labour in non-agriculture	0.02519	0.08168	0.308	0.758
Householdtypeothers	0.78227	0.11301	6.922	4.46e-12 *
HouseholdtypeRegular wage/Salary earning	0.45139	0.07269	6.210	5.29e-10 *
HouseholdtypeSelf-employment in agricultur	0.62005	0.06358	9.753	< 2e-16 *
HouseholdtypeSelf-employment in non-agriculture	0.62997	0.06907	9.121	< 2e-16 *
umpce_quintile_factor2	0.33780	0.04506	7.496	6.56e-14 *
umpce_quintile_factor3	0.46755	0.04803	9.735	< 2e-16 *
umpce_quintile_factor4	0.69613	0.04847	14.362	< 2e-16 *
umpce_quintile_factor5	0.96810	0.05231	18.508	< 2e-16 *

Signif. codes: 0 '*' 0.001 '*' 0.01 '' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 28260 on 27074 degrees of freedom

Residual deviance: 27522 on 27063 degrees of freedom

(24582 observations deleted due to missingness)

AIC: 27546

Number of Fisher Scoring iterations: 4

- Odds Ratio:

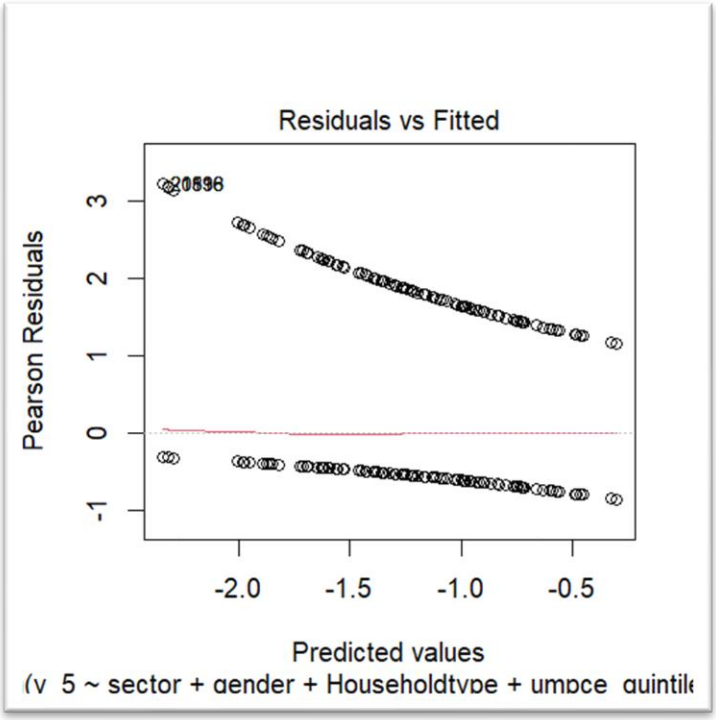
(Intercept)	sectorUrban
0.09878832	1.30086656
genderMale	HouseholdtypeCasual labour in non-agriculture
0.97196655	1.02551377
Householdtypeothers	HouseholdtypeRegular wage/Salary earning
2.18642200	1.57049851
HouseholdtypeSelf-employment in agriculture	HouseholdtypeSelf-employment in non-in agriculture
1.85902800	1.87756181
umpce_quintile_factor2	umpce_quintile_factor3
1.40185725	1.59608554
umpce_quintile_factor4	umpce_quintile_factor5
2.00596814	2.63294110

- The interpretation of logistic Regression of y_5:

The odds ratios from our model help us understand how different factors influence the likelihood of the outcome. Starting with the baseline (the intercept), we see very low odds (0.0988) when all predictors are at their reference levels. Living in an urban area increases the odds by 1.30 times compared to living in a rural area. Men have slightly lower odds (0.97) than women. Households with casual labor in non-agriculture have slightly higher odds (1.03) compared to the reference group, while those in other household types or with regular wage/salary earners have significantly higher odds (2.19 and 1.57, respectively). Self-employment, whether in agriculture (1.86) or non-agriculture (1.88), also shows increased odds. The umpce quintile factors reveal a clear pattern: higher umpce quintiles are associated with higher odds, with the top quintile (quintile 5) having the highest odds ratio of 2.63. This indicates a strong positive relationship between higher umpce quintiles and the likelihood of the outcome.

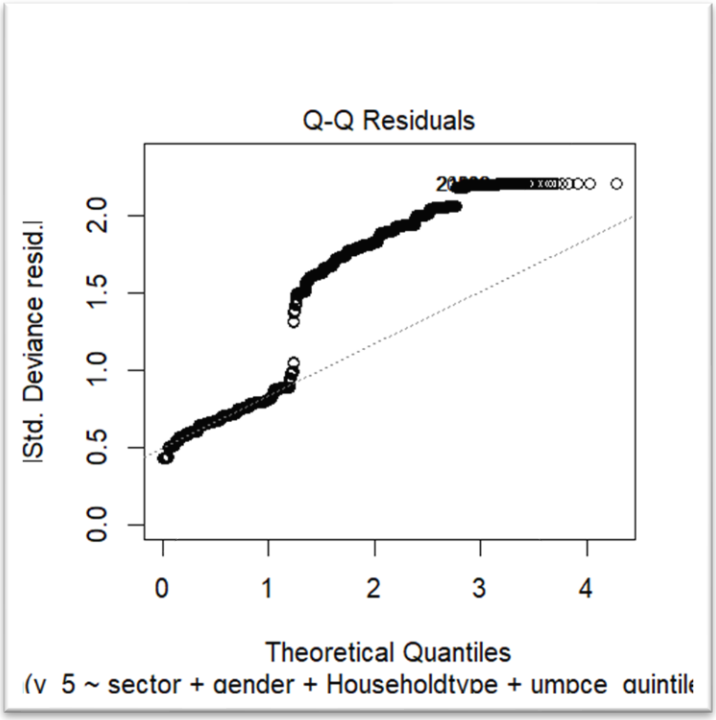
- The plots of logistic Regression of and their interpretation:

Residuals vs Fitted:



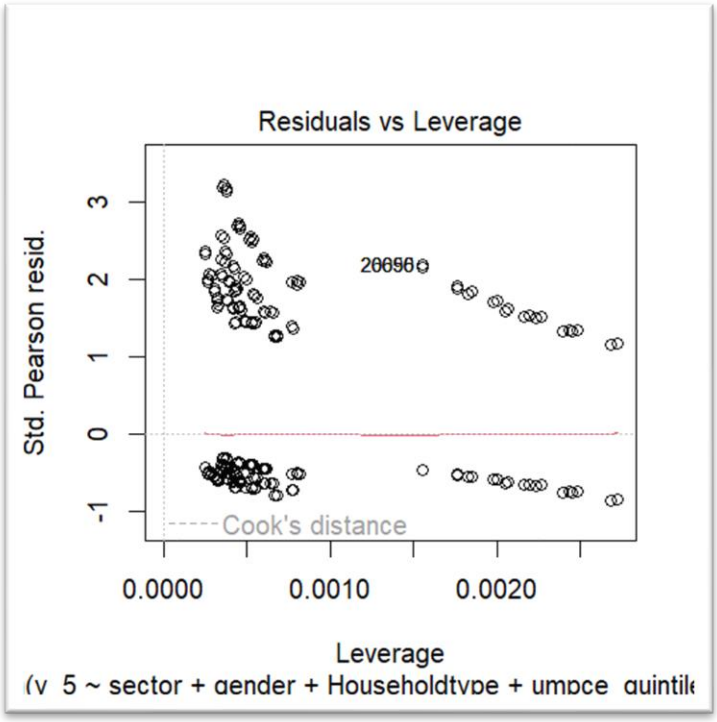
The plot shows a clear non-linear pattern, suggesting that the residuals aren't normally distributed. This means the model assumptions aren't fully met, indicating it might not be the best fit for the data. Specifically, the plot suggests the residuals are right-skewed, implying the model might be underestimating the variability in the data. Additionally, there are a few outliers—data points that are far from the rest—which could be influencing the model's results. Overall, the Q-Q plot suggests that further investigation is needed to improve the model's fit.

Q-Q Residuals:



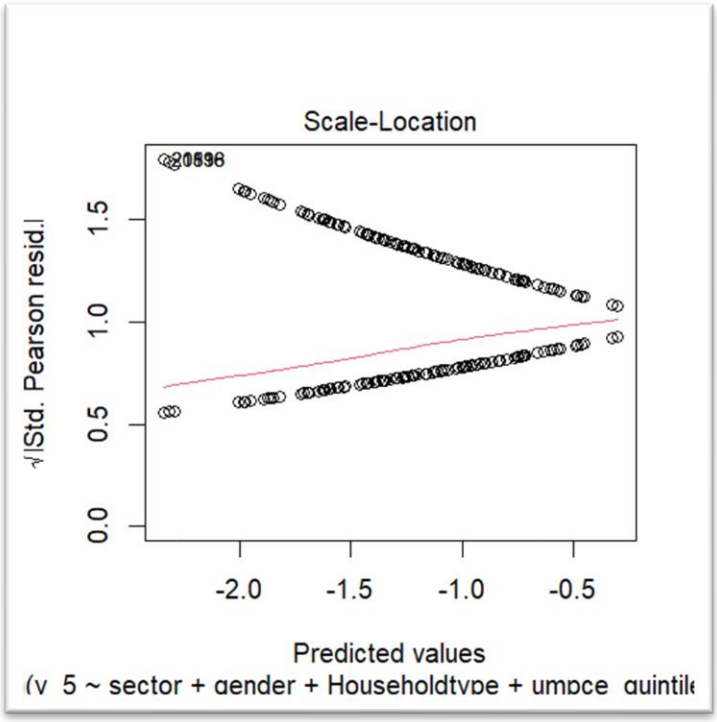
The Q-Q plot compares the theoretical quantiles of a standard normal distribution with the quantiles of the model's standardized deviance residuals. The points on the plot deviate from the diagonal line, showing that the residuals don't follow a normal distribution. There are also noticeable outliers, suggesting the model might not be capturing all the key relationships in the data. This indicates the model could be too simplistic and might need adjustments to better fit the data.

Residuals vs Leverage:



The plot indicates that the residuals are not following a normal distribution. They tend to cluster tightly around the line for smaller quantiles and then spread out more for larger quantiles, which implies that the variance of the residuals is not constant. This suggests that our model might not be the best fit for the data.The presence of non-normality means the model isn't capturing some important aspects of the data. Additionally, the evidence of non-constant variance indicates that the spread of the residuals varies at different levels of the response variable, further hinting at potential issues with the model's fit.

Scale-Location:



The plot compares the quantiles of the residuals to those of a standard normal distribution. Ideally, the points should form a straight line if the residuals were normally distributed. However, in our case, the points veer off from this straight line, especially at the extremes. This deviation suggests that the residuals aren't normally distributed, indicating that our regression model's assumptions might be violated. As a result, the model's findings could be unreliable.

Conclusion:

From the outputs of Rhat values:

Across India, both urban and rural students face distinct challenges in staying enrolled or dropping out of education. Key factors for continued enrollment include Reason 4(engaged in economic activities), Reason 1(not interested in education), and Reason 2(financial contraints), which reflect significant influences across different demographics. For students who drop out or never enroll, major barriers are commonly linked to Reason 1(not interested in education) and Reason 10(others), with Reason 2(financial contraints) also being influential. This pattern highlights that while many reasons influence students' decisions about education, focusing on the most pressing issues—like lack of interest and financial challenges—can help in developing targeted strategies to improve enrollment and retention rates.

From the contingency table:

The results reveal that while many of the observed associations aren't statistically significant and might just be due to random chance, there are some notable exceptions. In particular, we found significant links between 'Householdtype' and 'umpce_quintile' with various variables in both rural and urban areas. When it comes to healthcare access, urban areas show clear and strong connections with different wealth quintiles, suggesting that wealth plays a big role in access to healthcare. Rural areas, on the other hand, have mixed results with a few borderline cases. These findings suggest that there's more to uncover, and it might be worth diving deeper to understand these significant patterns and what they mean.

From Logistic regression:

The odds ratios reveal how different factors influence the chance of the outcome. Starting with a baseline odds of 0.0988, urban residents are more likely to experience the outcome compared to those in rural areas, with their chances increasing by 1.30 times. Interestingly, men have lower odds than women. Households involved in casual non-agricultural labor have slightly higher odds, but those in other job categories or with regular wage/salary jobs face significantly higher odds. Self-employed individuals, whether in agriculture or non-agriculture, also show increased likelihoods. A noticeable trend is that higher monthly per capita expenditure (umpce) is linked to higher odds of the outcome, with the top expenditure quintile having the highest odds ratio of 2.63, suggesting that spending more correlates strongly with a greater chance of the outcome.

References:

- Theory and Methods of Survey Sampling, Second Edition (2009) : By Parimal Mukherjee

- Key Indicators of Social Consumption of India: Education NSS 71st Round (January 2014 – June 2014) National Sample Survey Office.

- Note on Sample Design and Estimation Procedure of NSS of 71st Round(2014)

- Regression Analysis – Chapter 14 Logistic Regression Models: Shalabh, IITKanpur

Acknowledgement:

I would like to extend my sincere gratitude to Professor (Dr.) Chandranath Pal, Head of Department, and all the esteemed professors of the Statistics Department at the **University of Kalyani** for affording us the opportunity to undertake this project.

I am deeply indebted to my project guide, Dr. Kajal Dihidar of the **Indian Statistical Institute, Kolkata (ISI)**, for her unwavering guidance, supervision, and provision of essential resources throughout the project. I appreciate the time she took from her busy schedule to mentor us and provide valuable insights.

I also express my appreciation to my project partner, Adrija Sengupta, for her consistent collaboration and support throughout the project.

Furthermore, I would like to acknowledge the encouragement and support received from my friends and family, which was instrumental in the successful completion of this project.