Abstract

Survey analysis of spending habit based on the time spend on internet and other demographic factors.

Spending habit survey

***Roli Srivastava***

Table of Contents

[Project Overview 2](#_Toc12450)

[Research Question 2](#_Toc12451)

[Significance of Research 2](#_Toc12452)

[Methodology 2](#_Toc12453)

[Data 3](#_Toc12454)

[Descriptive Analysis 4](#_Toc12455)

[Analysis of internet usage 4](#_Toc12456)

[Goodness of fit 7](#_Toc12457)

[Binomial vs Poisson 7](#_Toc12458)

[Ord Plot 8](#_Toc12459)

[Dist Plot 7](#_Toc12460)

[Analyzing Influence of Internet Usage on Spending Habits 9](#_Toc12461)

[Association Analysis 9](#_Toc12462)

[Chi Square test 9](#_Toc12463)

[CMH Test 9](#_Toc12464)

[AssocPlots 9](#_Toc12465)

[Correspondence Analysis 16](#_Toc12466)

[Regression Modeling 17](#_Toc12467)

[General Linear Regression Model 19](#_Toc12468)

[Log Linear Regression Model 18](#_Toc12469)

[Multicollinearity 20](#_Toc12470)

[Residual Plots 20](#_Toc12471)

[Odds Ratio 21](#_Toc12472)

[Analyzing Influence of Demographics on Spending Habits 24](#_Toc12473)

[Conclusion 26](#_Toc12478)

[Gaps and Future Efforts 33](#_Toc12479)

**References…………………………………………………………………………………………………………………………………………6**

# Project Overview

With the new millennium, the wave of internet has swept the world and especially United States. In the last couple of years, the usage of internet has exponentially gone up. People of all ages now have some sort of electronic device and are connected to the web. Many children under 5 years of age can now be seen with tablets and adults with multiple devices such as laptops, tablets and smart phones. We use internet for many things – playing games, watching movies, shopping, meeting new people and finding jobs among many other. Since we are becoming more and more dependent on the internet, we thought it would be interesting to look at if the usage of internet impacts our lifestyle in any way.

One of the topics that we are going to explore is if our spending habits and internet usage are related. As we spend more time online, we are exposed various advertisements, new technology, fashion trends, celebrity styles, products etc. and this we think will lead to

# Research Question:

*1. Does Internet Usage Influence Spending Habits?*

*2.Does Demographics Influence Spending Habits?*

# Significance of Research

If we can establish a relationship between internet usage and spending habit, then -

Internet heavy advertisement plans may be the best way to go for businesses

It can be used to educate young people about making smart financial decision

It can also be used to educate parents on how their children's spending habit may be forming without their understanding

Establishing this relationship maybe helpful in understanding the trend of where we are headed as technology continues to advance

# Methodology

* Descriptive statistics to analyze the type of data
  + Ord\_plots
  + Distplot
* Relationship studies
  + Chi Square Test
  + CMHtest
  + Assocstats
* Regression
* Tools used: R

# Data

The data set from Kaggle

It includes 1010 responses from young people who were asked to rank their preferences on various topics

The variables in order Strongly disagree 1-2-3-4-5 Strongly agree (integer) : Ranked data

* Internet usage less than an hour day - few hours a day - most of the day – no time at all (categorical). **Internet.usage**

**SPENDING HABITS**

* I enjoy going to large shopping centers. **Shopping.centres**
* I spend a lot of money on partying and socializing. **Entertainment**.**spending**
* I spend a lot of money on my appearance**. Spending.on.looks**
* I spend a lot of money on gadgets. **Spending**.**on.gadgets**
* I will happily pay more money for good, quality or healthy food. **Spending**.**on.healthy.eating**

**DEMOGRAPHICS**

* **Age**: (integer)
* **Height**: (integer)
* **Weight**: (integer)
* **Gender**: Female - Male (categorical)
* **Highest education achieved**: Primary school - Secondary school - Bachelor degree (categorical)
* **I am the only child**: No - Yes (categorical)
* **I spent most of my childhood in a**: City - village (categorical)
* **I lived most of my childhood in a**: house/bungalow - block of flats (categorical)
* str(SurveyData)

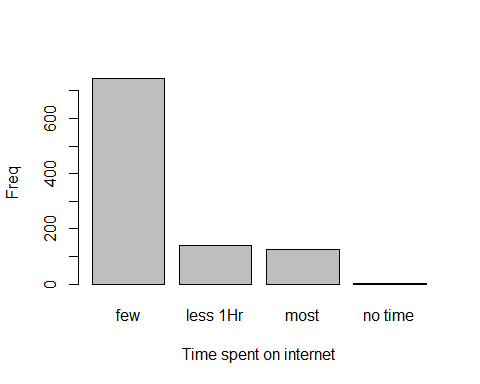
## 'data.frame': 1010 obs. of 16 variables:  
## $ Internet.usage : Factor w/ 4 levels "few","less 1Hr",..: 1 1 1 3 1 1 2 1 1 1 ...  
## $ Finances : int 3 3 2 2 4 2 4 3 2 4 ...  
## $ Shopping.centres : int 4 4 4 4 3 3 3 4 1 4 ...  
## $ Branded.clothing : int 5 1 1 3 4 3 1 4 3 4 ...  
## $ Entertainment.spending : int 3 4 4 3 3 3 3 4 2 2 ...  
## $ Spending.on.looks : int 3 2 3 4 3 1 4 4 1 3 ...  
## $ Spending.on.gadgets : int 1 5 4 4 2 4 1 3 3 2 ...  
## $ Spending.on.healthy.eating: int 3 2 2 1 4 4 5 2 4 4 ...  
## $ Age : int 20 19 20 22 20 20 20 19 18 19 ...  
## $ Height : int 163 163 176 172 170 186 177 184 166 174 ...  
## $ Weight : int 48 58 67 59 59 77 50 90 55 60 ...  
## $ Gender : Factor w/ 3 levels "","female","male": 2 2 2 2 2 3 2 3 2 2 ...  
## $ Education : Factor w/ 7 levels "","college/bachelor degree",..: 2 2 7 2 7 7 7 2 7 7 ...  
## $ Only.child : Factor w/ 3 levels "","no","yes": 2 2 2 3 2 2 2 2 2 2 ...  
## $ Village\_town : Factor w/ 3 levels "","city","village": 3 2 2 2 3 2 3 2 2 2 ...  
## $ building : Factor w/ 3 levels "","block of flats",..: 2 2 2 3 3 2 3 3 3 2 ...

table(SurveyData$Internet.usage)

##   
## few less 1Hr most no time   
## 744 139 124 3

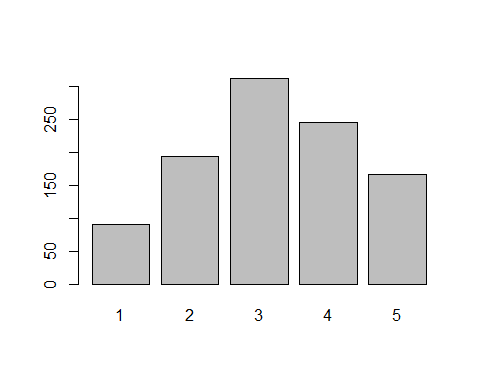
# Descriptive Analysis

Analysis of internet usage : it seems like most of the young people at least spend few hours on internet.



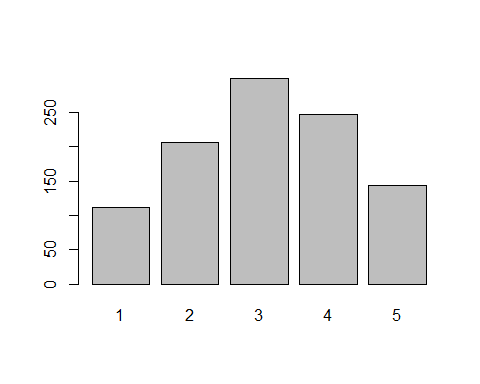
plot(factor(SurveyData$Entertainment.spending))

Analysis of spending on Entertainment : it seems like most of the young people moderately agrees spending on entertainment.



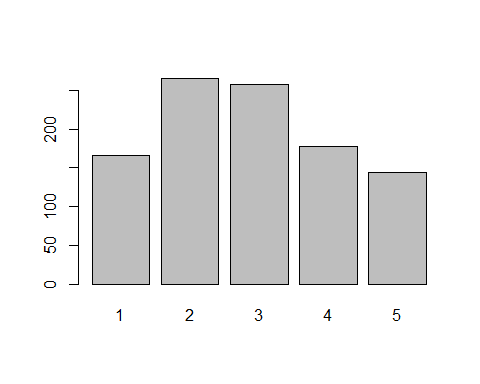
plot(factor(SurveyData$Spending.on.looks))

Analysis of spending on looks : it seems like most of the young people moderately agrees spending on looks.



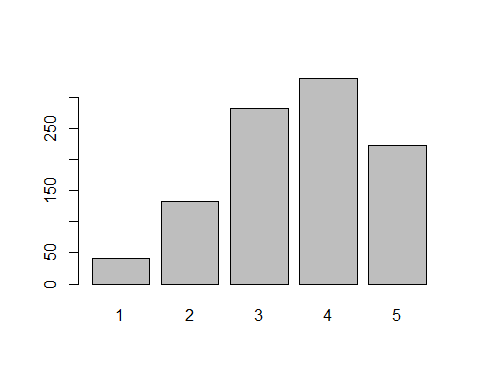
plot(factor(SurveyData$Spending.on.gadgets)):

Analysis of spending on gadgets : it seems like most of the young people moderately agrees spending on gadgets.



plot(factor(SurveyData$Spending.on.healthy.eating)):

Analysis of spending on healthy eating : it seems like most of the young people strongly agrees spending on healthy eating.

SurveyData$entertainment <- (SurveyData$Entertainment.spending)  
SurveyData$looks <- (SurveyData$Spending.on.looks)  
SurveyData$gadgets <- (SurveyData$Spending.on.gadgets)  
SurveyData$eating <- (SurveyData$Spending.on.healthy.eating)  
SurveyData$Internet.usage <-(as.integer(SurveyData$Internet.usage))

## Goodness of fit

summary(goodfit(SurveyData$Internet.usage, type="binomial"))

## Warning in goodfit(SurveyData$Internet.usage, type = "binomial"): size was  
## not given, taken as maximum count

##   
## Goodness-of-fit test for binomial distribution  
##   
## X^2 df P(> X^2)  
## Likelihood Ratio 755.4354 2 9.105325e-165

***## p-value is very high shows significant lack of fit for binomial****.*

summary(goodfit(SurveyData$Internet.usage, type="poisson"))

##   
## Goodness-of-fit test for poisson distribution  
##   
## X^2 df P(> X^2)  
## Likelihood Ratio 976.6713 2 8.289401e-213

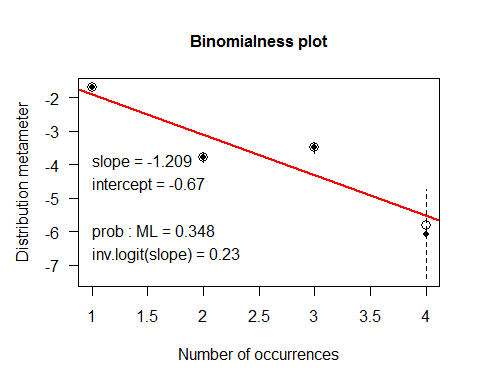
***## p-value is very high shows significant lack of fit for poisson****.*

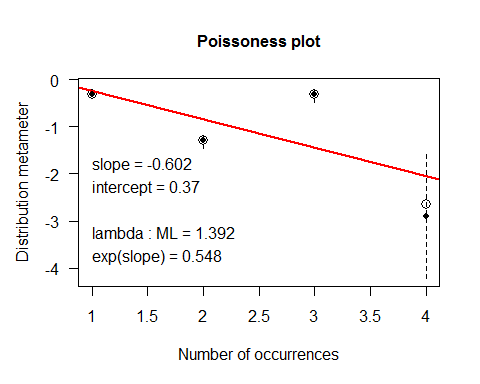
## Binomial vs Poisson

***## neither poisson nor binomial fits the distribution****.*

## Dist Plot

distplot(SurveyData$Internet.usage, type="binomial")

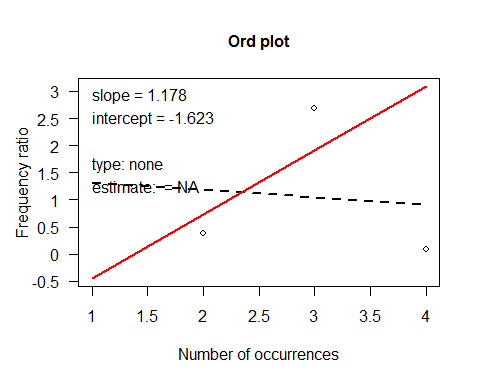
distplot(SurveyData$Internet.usage, type="poisson")



***## It clearly shows in distplot also that internet usage data is not fitting either Binomial or Poisson data.***

## Ord Plot

Ord\_plot(SurveyData$Internet.usage) : positive slope and negative intercept



***## ord plot shows what distribution to use , in our case it is log-series.***

# Analyzing Influence of Internet Usage on Spending Habits

## Association Analysis

### Chi Square test / CMH Test / AssocPlots

Between internet usage and entertainment spending

surveyEnt <- ftable(xtabs(~Internet.usage + entertainment, data = SurveyData ))  
  
chisq.test(surveyEnt)

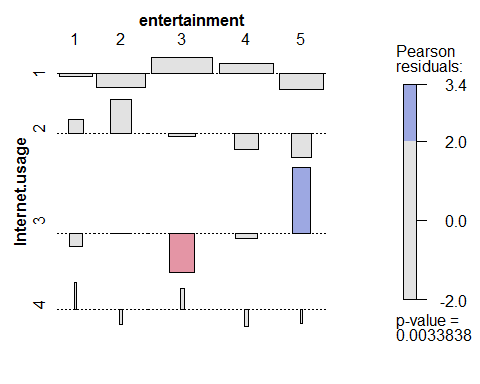
## Warning in chisq.test(surveyEnt): Chi-squared approximation may be  
## incorrect

##   
## Pearson's Chi-squared test  
##   
## data: surveyEnt  
## X-squared = 29.444, df = 12, p-value = 0.003384

# significant relation between internet usage and entertainment spending  
  
assocstats(surveyEnt)

## X^2 df P(> X^2)  
## Likelihood Ratio 28.300 12 0.0049983  
## Pearson 29.444 12 0.0033838  
##   
## Phi-Coefficient : NA   
## Contingency Coeff.: 0.169   
## Cramer's V : 0.099

#By the contingency coefficient, there is moderately strong association between internet usage and entertainment spending  
  
vcd::assoc(surveyEnt, shade=TRUE)



#By the assoc plot, there is p-value is high so there is strong relationship between internet usage and entertainment spending

CMHtest(surveyEnt)

## Cochran-Mantel-Haenszel Statistics   
##   
## AltHypothesis Chisq Df Prob  
## cor Nonzero correlation 1.189 1 0.2755292  
## rmeans Row mean scores differ 13.094 3 0.0044386  
## cmeans Col mean scores differ 11.341 4 0.0229907  
## general General association 29.414 12 0.0034180

*## significant prob value.*

#The results of all four tests lead to the same conclusion that there is a significant association between the internet usage and entertainment spending The test for non-zero correlation, treating both variables as ordinal, has the largest ratio of χ2/df.  
  
surveylooks <- ftable(xtabs(~Internet.usage + looks, data = SurveyData ))  
  
chisq.test(surveylooks) # because both are ranked variable

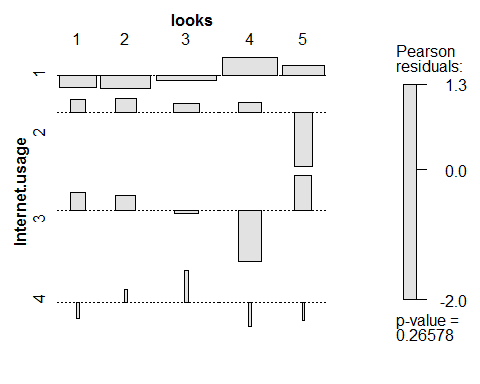
## Warning in chisq.test(surveylooks): Chi-squared approximation may be  
## incorrect

##   
## Pearson's Chi-squared test  
##   
## data: surveylooks  
## X-squared = 14.57, df = 12, p-value = 0.2658

# non-significant relation between internet usage and looks spending  
  
assocstats(surveylooks)

## X^2 df P(> X^2)  
## Likelihood Ratio 16.667 12 0.16255  
## Pearson 14.570 12 0.26578  
##   
## Phi-Coefficient : NA   
## Contingency Coeff.: 0.119   
## Cramer's V : 0.069

#By the contingency coefficient, there is moderately association between internet usage and entertainment spending  
  
vcd::assoc(surveylooks, shade=TRUE)



#By the assoc plot, p-value is non-significant so there not very strong relationship between internet usage and looks spending

CMHtest(surveylooks)

## Cochran-Mantel-Haenszel Statistics   
##   
## AltHypothesis Chisq Df Prob  
## cor Nonzero correlation 2.1603 1 0.14162  
## rmeans Row mean scores differ 3.8559 3 0.27745  
## cmeans Col mean scores differ 4.9725 4 0.29014  
## general General association 14.5559 12 0.26662

*## non-significant prob value.*

#The results of 3 tests lead to the same conclusion that there is a non-significant association between the internet usage and looks spending The test for non-zero correlation, treating both variables as ordinal, has the largest ratio of χ2/df.  
  
  
surveyGadgets <- ftable(xtabs(~Internet.usage + gadgets, data = SurveyData ))  
  
chisq.test(surveyGadgets)

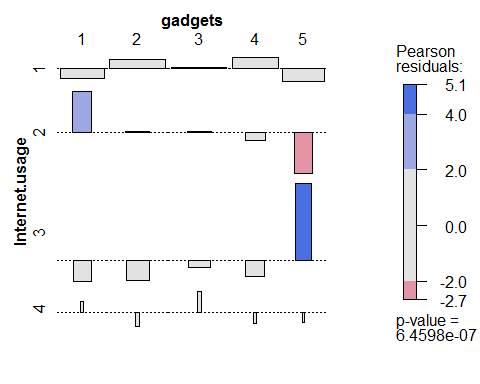
## Warning in chisq.test(surveyGadgets): Chi-squared approximation may be  
## incorrect

##   
## Pearson's Chi-squared test  
##   
## data: surveyGadgets  
## X-squared = 51.899, df = 12, p-value = 6.46e-07

# significant relation between internet usage and gadgets spending  
  
assocstats(surveyGadgets)

## X^2 df P(> X^2)  
## Likelihood Ratio 47.715 12 3.5041e-06  
## Pearson 51.899 12 6.4598e-07  
##   
## Phi-Coefficient : NA   
## Contingency Coeff.: 0.221   
## Cramer's V : 0.131

#By the contingency coefficient, there is moderately strong association between internet usage and gadgets spending  
  
vcd::assoc(surveyGadgets, shade=TRUE)



#By the assoc plot, p-value is high so there is strong relationship between internet usage and gadgets spending

CMHtest(surveyGadgets)

## Cochran-Mantel-Haenszel Statistics   
##   
## AltHypothesis Chisq Df Prob  
## cor Nonzero correlation 4.071 1 4.3626e-02  
## rmeans Row mean scores differ 29.237 3 1.9965e-06  
## cmeans Col mean scores differ 16.389 4 2.5396e-03  
## general General association 51.848 12 6.5965e-07

*## significant prob value.*  
  
#The results of all four tests lead to the same conclusion that there is a significant association between the internet usage and gadgets spending The test for non-zero correlation, treating both variables as ordinal,has the largest ratio of χ2/df.  
  
surveyEating <- ftable(xtabs(~Internet.usage + eating, data = SurveyData ))  
  
chisq.test(surveyEating)

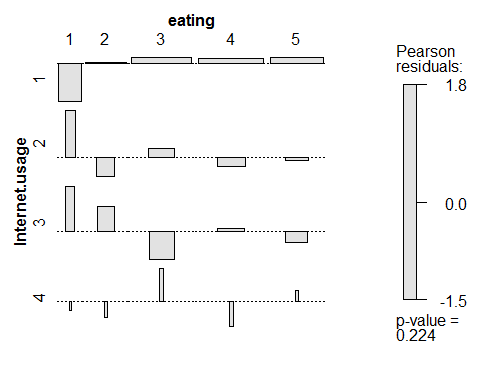
## Warning in chisq.test(surveyEating): Chi-squared approximation may be  
## incorrect

##   
## Pearson's Chi-squared test  
##   
## data: surveyEating  
## X-squared = 15.328, df = 12, p-value = 0.224

# non-significant relation between internet usage and eating spending  
  
assocstats(surveyEating)

## X^2 df P(> X^2)  
## Likelihood Ratio 15.394 12 0.22061  
## Pearson 15.328 12 0.22400  
##   
## Phi-Coefficient : NA   
## Contingency Coeff.: 0.122   
## Cramer's V : 0.071

#By the contingency coefficient, there is moderate association between internet usage and eating spending  
  
vcd::assoc(surveyEating, shade=TRUE)



#By the assoc plot, p-value is non-significant so there is not very strong relationship between internet usage and healthy eating spending

CMHtest(surveyEating)

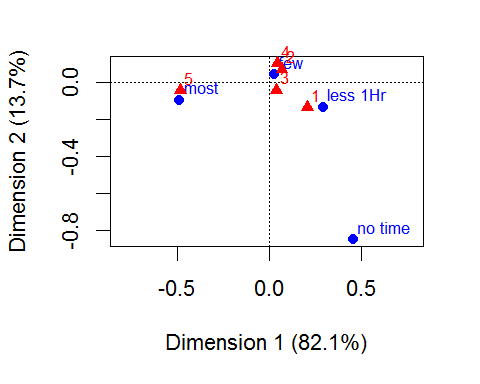
## Cochran-Mantel-Haenszel Statistics   
##   
## AltHypothesis Chisq Df Prob  
## cor Nonzero correlation 2.6102 1 0.106178  
## rmeans Row mean scores differ 2.9105 3 0.405624  
## cmeans Col mean scores differ 7.7845 4 0.099799  
## general General association 15.3125 12 0.224788

#non-significant prob value

#The results of all four tests lead to the same conclusion that there is a non-significant association between the internet usage and eating spending the test for non-zero correlation, treating both variables as ordinal, has the largest ratio of χ2/df.

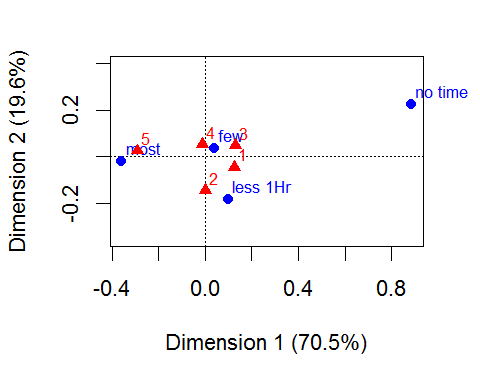
## **Correspondence Analysis**

library(ca)  
gad\_1 <- (xtabs(~SurveyData$Int + gadgets, data = SurveyData ))  
op <- par(cex=1.4,mar=c(5,4,2,2)+.1)  
gadgets.ca <- ca(gad\_1)  
res <- plot(gadgets.ca)



***#For Gadgets it shows that young people who strongly agreed spending on gadget and rated as 5 use to spend most time on internet. Those who rated 1 and 3 on gadget spending spend less than 1 hour or no time on internet. Those who rated 2 and 4 for gadget spending , spend only few hours on internet.***

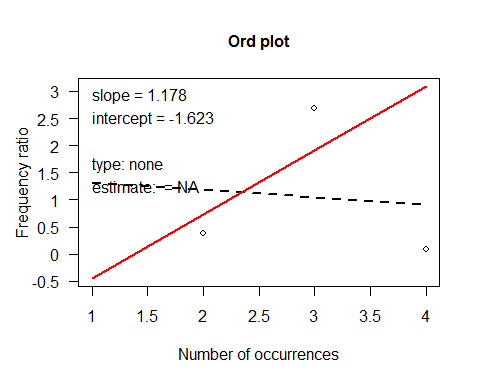
par(op)  
  
gad\_1 <- (xtabs(~SurveyData$Int + entertainment, data = SurveyData ))  
op <- par(cex=1.4,mar=c(5,4,2,2)+.1)  
entertainment.ca <- ca(gad\_1)  
res <- plot(entertainment.ca)



***#For entertainment it shows that young people who strongly agreed spending on entertainment and rated as 5 use to spend most time on internet. Those who rated 1 and 2 on gadget spending spend less than 1 hour on internet. Those who rated 3 and 4 for gadget spending , spend only few hours or no time on internet .***

## **Regression Modeling:**

As we saw in our Ord plot that our internet data was following log-series distribution we will use log linear regression to model our data.



### **Log Linear Regression Model**

library(MASS)  
logModel <- loglm(Internet.usage ~ entertainment + eating + gadgets + looks, data=SurveyData)  
summary(logModel)

## Formula:  
## Internet.usage ~ entertainment + eating + gadgets + looks  
## attr(,"variables")  
## list(Internet.usage, entertainment, eating, gadgets, looks)  
## attr(,"factors")  
## entertainment eating gadgets looks  
## Internet.usage 0 0 0 0  
## entertainment 1 0 0 0  
## eating 0 1 0 0  
## gadgets 0 0 1 0  
## looks 0 0 0 1  
## attr(,"term.labels")  
## [1] "entertainment" "eating" "gadgets" "looks"   
## attr(,"order")  
## [1] 1 1 1 1  
## attr(,"intercept")  
## [1] 1  
## attr(,"response")  
## [1] 1  
## attr(,".Environment")  
## <environment: R\_GlobalEnv>  
## attr(,"predvars")  
## list(Internet.usage, entertainment, eating, gadgets, looks)  
## attr(,"dataClasses")  
## Internet.usage entertainment eating gadgets looks   
## "numeric" "numeric" "numeric" "numeric" "numeric"   
##   
## Statistics:  
## X^2 df P(> X^2)  
## Likelihood Ratio 289.4591 985 1  
## Pearson 340.5478 985 1

## Pearson coefficient is high signifying strong relation between internet usage and spending.

### **General Linear Regression Model**

survey.logistic2 <- glm(Internet.usage ~ gadgets + looks + entertainment + eating,data=SurveyData)  
  
summary(survey.logistic2)

##   
## Call:  
## glm(formula = Internet.usage ~ gadgets + looks + entertainment +   
## eating, data = SurveyData)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.6720 -0.4169 -0.3554 0.5300 2.7517   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.44500 0.09483 15.238 <2e-16 \*\*\*  
## gadgets 0.05028 0.01915 2.625 0.0088 \*\*   
## looks -0.04558 0.02097 -2.174 0.0299 \*   
## entertainment 0.02138 0.02106 1.015 0.3103   
## eating -0.03545 0.02127 -1.666 0.0960 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.4954293)  
##   
## Null deviance: 500.42 on 1001 degrees of freedom  
## Residual deviance: 493.94 on 997 degrees of freedom  
## (8 observations deleted due to missingness)  
## AIC: 2146.8  
##   
## Number of Fisher Scoring iterations: 2

coeftest(survey.logistic2)

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.444996 0.094830 15.2378 < 2.2e-16 \*\*\*  
## gadgets 0.050282 0.019154 2.6251 0.008663 \*\*   
## looks -0.045582 0.020968 -2.1739 0.029711 \*   
## entertainment 0.021379 0.021059 1.0152 0.310009   
## eating -0.035448 0.021273 -1.6664 0.095637 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**## Both gadgets and spending on looks are significant predictor for the internet usage**  **. It shows that for every unit increase in spending on gadgets internet usage increase by 0.05.It shows that for every unit increase in spending on looks internet usage decreases by 0.04 . Rest all entertainment and eating are non-significant.**

### **Multicollinearity**

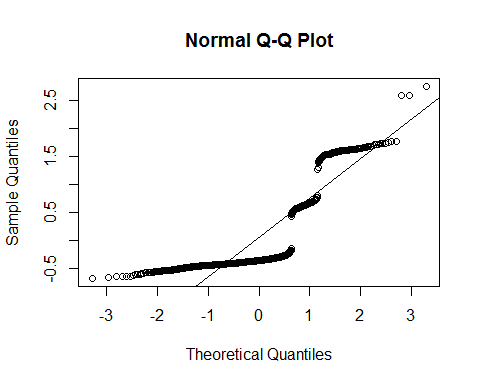
car::vif(survey.logistic2)

## gadgets looks entertainment eating   
## 1.224228 1.292038 1.265751 1.092040

## No multicollinearity is found , all values are below 10.

### **Residual Plots**

qqnorm(residuals(survey.logistic2))  
qqline(residuals(survey.logistic2))



shapiro.test(survey.logistic2$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: survey.logistic2$residuals  
## W = 0.68499, p-value < 2.2e-16

***## W is approaching to 1 so this fine , it is significant as there is more than 1000 observation.***

***## Not very great residual plot but as we have seen that spending itself is not a great predictor of internet, so this was expected. There might be another factor which needs to be analyzed which might be affecting internet usage.***

### **Odds Ratio**

exp(cbind(OddsRatio=coef(survey.logistic2),  
 confint(survey.logistic2)))

## Waiting for profiling to be done...

## OddsRatio 2.5 % 97.5 %  
## (Intercept) 4.2418343 3.5223656 5.108260  
## gadgets 1.0515671 1.0128212 1.091795  
## looks 0.9554415 0.9169728 0.995524  
## entertainment 1.0216089 0.9803014 1.064657  
## eating 0.9651728 0.9257588 1.006265

***## Thus, for a gadget, the odds of spending on gadgets are 1.05 times as large as the odds of internet usage increases.***

**Linear Regression Model:**

using only gadgets and looks as they seems to impact the internet usage.

model2 <- lm(Internet.usage ~ gadgets + looks, data = SurveyData, family = "logistic")

## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :  
## extra argument 'family' will be disregarded

summary(model2)

##   
## Call:  
## lm(formula = Internet.usage ~ gadgets + looks, data = SurveyData,   
## family = "logistic")  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.5830 -0.4041 -0.3571 0.5520 2.6490   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.39191 0.06946 20.038 <2e-16 \*\*\*  
## gadgets 0.04700 0.01837 2.559 0.0106 \*   
## looks -0.04395 0.01957 -2.246 0.0249 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7052 on 1004 degrees of freedom  
## (3 observations deleted due to missingness)  
## Multiple R-squared: 0.008615, Adjusted R-squared: 0.00664   
## F-statistic: 4.362 on 2 and 1004 DF, p-value: 0.01299

***## Both gadgets and looks are significant , R values are low showing that there are other factors also which are affecting internet usage.***

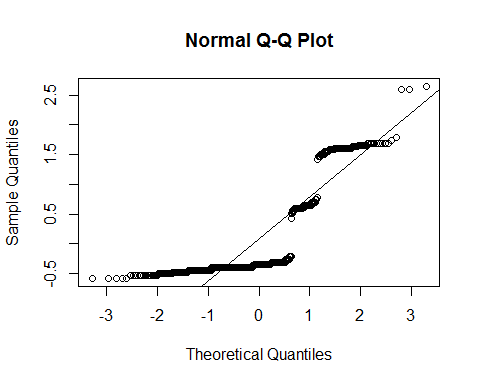
car::vif(model2)

## gadgets looks   
## 1.125926 1.125926

***## so no multicollinearity issue in gadget and looks spending***  
  
  
  
  
modelchi <- model1$deviance-model2$deviance  
cdf <- model2$df.residual-model1$df.residual  
chisqp <- 1 - pchisq(modelchi, cdf)  
***## no difference in both the models***  
  
exp(model2$coefficients)

## (Intercept) gadgets looks   
## 4.0225097 1.0481239 0.9569991

qqnorm(residuals(model2))  
qqline(residuals(model2))



shapiro.test(model2$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: model2$residuals  
## W = 0.66549, p-value < 2.2e-16

***## W is approaching to 1 so this fine , it is significant as there is more than 1000 observation.***

***## Not very great residual plot but as we have seen that spending itself is not a great predictor of internet, so this was expected. There might be another factor which needs to be analyzed which might be affecting internet usage.***

## **MULTINOMIAL Model:**

library("nnet")  
test <- multinom(Internet.usage ~ entertainment + eating + gadgets + looks, data = SurveyData)

## # weights: 24 (15 variable)  
## initial value 1389.066950   
## iter 10 value 771.968397  
## iter 20 value 751.283624  
## iter 30 value 750.986154  
## iter 30 value 750.986148  
## iter 30 value 750.986148  
## final value 750.986148   
## converged

summary(test)

## Call:  
## multinom(formula = Internet.usage ~ entertainment + eating +   
## gadgets + looks, data = SurveyData)  
##   
## Coefficients:  
## (Intercept) entertainment eating gadgets looks  
## 2 -0.769923 -0.07682331 -0.008093706 -0.2039963 -0.03162994  
## 3 -2.137609 0.16058831 -0.180942037 0.3445191 -0.20319944  
## 4 -4.040725 -0.57634082 0.225214064 -0.1952109 -0.06391072  
##   
## Std. Errors:  
## (Intercept) entertainment eating gadgets looks  
## 2 0.3877688 0.08752450 0.08815376 0.08377291 0.08783687  
## 3 0.4297777 0.09504467 0.09409018 0.08612084 0.09319440  
## 4 2.4048415 0.59051741 0.55193747 0.53443241 0.53702349  
##   
## Residual Deviance: 1501.972   
## AIC: 1531.972

#summary  
#1. Model execution output shows some iteration history and includes the final negative log-likelihood 750.986148. This value is multiplied by two as shown in the model summary as the Residual Deviance.  
  
#2. The summary output has a block of coefficients and another block of standard errors. Each block has one row of values corresponding to one model equation. In the block of coefficients, we see that the first row is being compared to Internet. Usage = “less than an hour a day” to our baseline Internet. Usage = “usage” and the second row to Internet. Usage = “most of the day ” to our baseline Internet. Usage = “usage” and the third row to Internet. Usage = “no time at all ” to our baseline Internet. Usage = “usage”  
  
#3. A one-unit increase in entertainment decreases the log odds of being in 'less than an hour a day' vs. usage by 0.076  
  
#4. A one-unit increase in entertainment increases the log odds of being in 'most of the day' vs. usage by 0.16  
  
#4. A one-unit increase in entertainment decreases the log odds of being in 'no time at all' vs. usage by 0.57  
  
#Now we’ll calculate Z score and p-Value for the variables in the model.  
  
z <- summary(test)$coefficients/summary(test)$standard.errors  
z

## (Intercept) entertainment eating gadgets looks  
## 2 -1.985521 -0.8777349 -0.09181352 -2.4351104 -0.3600986  
## 3 -4.973756 1.6896088 -1.92307034 4.0004152 -2.1803825  
## 4 -1.680246 -0.9759929 0.40804271 -0.3652678 -0.1190092

p <- (1 - pnorm(abs(z), 0, 1))\*2  
p

## (Intercept) entertainment eating gadgets looks  
## 2 4.708659e-02 0.38008758 0.92684621 1.488725e-02 0.71877337  
## 3 6.566808e-07 0.09110281 0.05447122 6.323143e-05 0.02922912  
## 4 9.290951e-02 0.32906797 0.68324232 7.149115e-01 0.90526809

***## gadgets look significant***

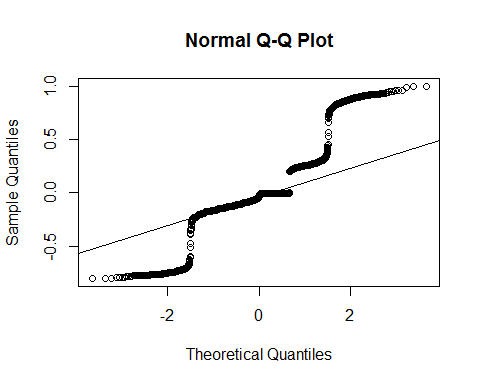
exp(coef(test))

## (Intercept) entertainment eating gadgets looks  
## 2 0.46304873 0.9260535 0.9919390 0.8154654 0.9688651  
## 3 0.11793646 1.1742015 0.8344837 1.4113111 0.8161155  
## 4 0.01758472 0.5619509 1.2525908 0.8226611 0.9380888

head(fitted(test))

## 1 2 3 4  
## 1 0.7377536 0.19637964 0.06279404 0.0030727157  
## 2 0.5894376 0.06685485 0.34316981 0.0005377469  
## 3 0.6791335 0.09151832 0.22864162 0.0007065099  
## 4 0.7022827 0.09981759 0.19692606 0.0009736691  
## 5 0.7576624 0.16313658 0.07594925 0.0032517487  
## 6 0.6870002 0.10478931 0.20594294 0.0022675255

qqnorm(residuals(test))  
qqline(residuals(test))



shapiro.test(test$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: test$residuals  
## W = 0.85599, p-value < 2.2e-16

***## little improved***

***## W is approaching to 1 so this fine , it is significant as there is more than 1000 observation.***

***## Not very great residual plot but as we have seen that spending itself is not a great predictor of internet, so this was expected. There might be another factor which needs to be analyzed which might be affecting internet usage***

# **Analyzing Influence of Demographics on Spending Habits**

**Transforming data and exploratory Analysis**

SurveyData$entertainment <- factor(SurveyData$Entertainment.spending)  
SurveyData$looks <- factor(SurveyData$Spending.on.looks)  
SurveyData$gadgets <- factor(SurveyData$Spending.on.gadgets)  
SurveyData$eating <- factor(SurveyData$Spending.on.healthy.eating)  
  
tabulate(SurveyData$Gender)

## [1] 6 593 411

tabulate(SurveyData$Education)

## [1] 1 212 10 5 81 80 621

tabulate(SurveyData$Only.child)

## [1] 2 754 254

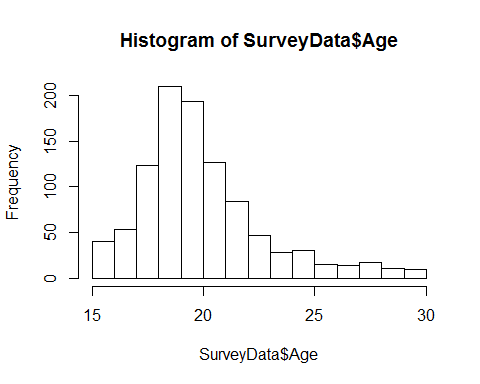
tabulate(SurveyData$Village\_town)

## [1] 4 707 299

tabulate(SurveyData$building)

## [1] 4 595 411

*## all of them are already factored so no need to factor. Looks like balanced design*  
hist.default(SurveyData$Age)



##Now check all possible relationship:   
#age-gender ,age-education , age -only-child ,age-village-town, age-building  
  
model <- aov(SurveyData$Age ~ SurveyData$Gender)   
summary(model)

## Df Sum Sq Mean Sq F value Pr(>F)   
## SurveyData$Gender 2 155 77.39 9.841 5.86e-05 \*\*\*  
## Residuals 1000 7864 7.86   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 7 observations deleted due to missingness

## it seems significant let’s do tukey to see the difference in levels of gender  
TukeyHSD(model)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = SurveyData$Age ~ SurveyData$Gender)  
##   
## $`SurveyData$Gender`  
## diff lwr upr p adj  
## female- -2.0862479 -5.0422915 0.8697958 0.2226353  
## male- -1.3271394 -4.2886533 1.6343746 0.5443273  
## male-female 0.7591085 0.3354595 1.1827576 0.0000840

## not significant so not to worry.  
model <- aov(SurveyData$Age ~ SurveyData$Only.child)   
summary(model)

## Df Sum Sq Mean Sq F value Pr(>F)   
## SurveyData$Only.child 2 82 40.82 5.143 0.006 \*\*  
## Residuals 1000 7937 7.94   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 7 observations deleted due to missingness

## non-significant  
  
model <- aov(SurveyData$Age ~ SurveyData$Village\_town)   
summary(model)

## Df Sum Sq Mean Sq F value Pr(>F)  
## SurveyData$Village\_town 2 2 1.097 0.137 0.872  
## Residuals 1000 8016 8.016   
## 7 observations deleted due to missingness

## non-significant  
model <- aov(SurveyData$Age ~ SurveyData$building)   
summary(model)

## Df Sum Sq Mean Sq F value Pr(>F)  
## SurveyData$building 2 4 1.763 0.22 0.803  
## Residuals 1000 8015 8.015   
## 7 observations deleted due to missingness

## non-significant  
  
## now since there is no multicollinearity, we can proceed to model based on spending - let’s take it one by one first spending on gadgets

**Analyzing the effect of demographics on spending on gadgets**

*model <- polr(gadgets ~ Gender + Education + Age + Only.child + Village\_town + building, data = SurveyData, Hess = TRUE, method = "logistic")  
coeftest(model)*

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value  
## Genderfemale -0.200986 0.835898 -0.2404  
## Gendermale 0.867773 0.837763 1.0358  
## Educationcollege/bachelor degree -0.735528 1.486638 -0.4948  
## Educationcurrently a primary school pupil 0.233064 1.608784 0.1449  
## Educationdoctorate degree -0.749004 1.720296 -0.4354  
## Educationmasters degree -0.540909 1.507517 -0.3588  
## Educationprimary school -0.688133 1.490339 -0.4617  
## Educationsecondary school -0.955850 1.481031 -0.6454  
## Age -0.038769 0.026507 -1.4626  
## Only.childno -0.608239 1.048580 -0.5801  
## Only.childyes -0.393365 1.053191 -0.3735  
## Village\_towncity 0.037087 0.766288 0.0484  
## Village\_townvillage -0.310077 0.775146 -0.4000  
## buildingblock of flats 3.000050 1.148814 2.6114  
## buildinghouse/bungalow 2.989889 1.153348 2.5924  
## Pr(>|t|)   
## Genderfemale 0.810036   
## Gendermale 0.300540   
## Educationcollege/bachelor degree 0.620880   
## Educationcurrently a primary school pupil 0.884843   
## Educationdoctorate degree 0.663373   
## Educationmasters degree 0.719816   
## Educationprimary school 0.644378   
## Educationsecondary school 0.518822   
## Age 0.143898   
## Only.childno 0.562007   
## Only.childyes 0.708858   
## Village\_towncity 0.961409   
## Village\_townvillage 0.689225   
## buildingblock of flats 0.009154 \*\*  
## buildinghouse/bungalow 0.009673 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## looks like only building is impacting , only that is significant p = 0.009673.  
  
model <- polr(gadgets ~ building, data = SurveyData, Hess = TRUE, method = "logistic")  
coeftest(model)

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## buildingblock of flats 2.9287 1.1364 2.5772 0.01010 \*  
## buildinghouse/bungalow 2.7318 1.1371 2.4024 0.01647 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## now since there is no multicollinearity, we can proceed to model based on spending - let’s take it one by one first spending on entertainment

**Analyzing the effect of demographics on spending on entertainment**

model <- polr(entertainment ~ Gender + Education + Age + Only.child + Village\_town + building, data = SurveyData, Hess = TRUE, method = "logistic")  
coeftest(model)

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value  
## Genderfemale -1.459637 0.872681 -1.6726  
## Gendermale -0.770535 0.872658 -0.8830  
## Educationcollege/bachelor degree 1.528275 1.518094 1.0067  
## Educationcurrently a primary school pupil 1.440841 1.608614 0.8957  
## Educationdoctorate degree 1.407090 1.694833 0.8302  
## Educationmasters degree 1.267823 1.538987 0.8238  
## Educationprimary school 1.365240 1.523825 0.8959  
## Educationsecondary school 1.315397 1.512295 0.8698  
## Age -0.038459 0.027086 -1.4199  
## Only.childno -8.133960 0.768471 -10.5846  
## Only.childyes -8.159020 0.770792 -10.5852  
## Village\_towncity -0.354869 0.934427 -0.3798  
## Village\_townvillage -0.618357 0.939763 -0.6580  
## buildingblock of flats 2.230624 1.027778 2.1703  
## buildinghouse/bungalow 2.246391 1.032693 2.1753  
## Pr(>|t|)   
## Genderfemale 0.09473 .   
## Gendermale 0.37747   
## Educationcollege/bachelor degree 0.31432   
## Educationcurrently a primary school pupil 0.37063   
## Educationdoctorate degree 0.40661   
## Educationmasters degree 0.41025   
## Educationprimary school 0.37051   
## Educationsecondary school 0.38462   
## Age 0.15596   
## Only.childno < 2e-16 \*\*\*  
## Only.childyes < 2e-16 \*\*\*  
## Village\_towncity 0.70420   
## Village\_townvillage 0.51070   
## buildingblock of flats 0.03022 \*   
## buildinghouse/bungalow 0.02985 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## looks like spending on entertainment is impacted by Only.child,building  
  
## now since there is no multicollinearity, we can proceed to model based on spending - let’s take it one by one first spending on looks

**Analyzing the effect of demographics on spending on looks**

model <- polr(looks ~ Gender + Education + Age + Only.child + Village\_town + building, data = SurveyData, Hess = TRUE, method = "logistic")  
coeftest(model)

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value  
## Genderfemale -0.077234 0.738731 -0.1045  
## Gendermale -0.555374 0.740316 -0.7502  
## Educationcollege/bachelor degree -0.970989 1.512213 -0.6421  
## Educationcurrently a primary school pupil -1.209251 1.599664 -0.7559  
## Educationdoctorate degree -0.452002 1.709812 -0.2644  
## Educationmasters degree -1.180555 1.534167 -0.7695  
## Educationprimary school -1.359392 1.517616 -0.8957  
## Educationsecondary school -1.149444 1.506514 -0.7630  
## Age -0.040222 0.026800 -1.5008  
## Only.childno -18.329070 0.711431 -25.7636  
## Only.childyes -17.945845 0.712000 -25.2048  
## Village\_towncity -0.017072 0.905686 -0.0188  
## Village\_townvillage -0.446114 0.911501 -0.4894  
## buildingblock of flats 0.687817 0.798576 0.8613  
## buildinghouse/bungalow 0.861123 0.805075 1.0696  
## Pr(>|t|)   
## Genderfemale 0.9168   
## Gendermale 0.4533   
## Educationcollege/bachelor degree 0.5210   
## Educationcurrently a primary school pupil 0.4499   
## Educationdoctorate degree 0.7916   
## Educationmasters degree 0.4418   
## Educationprimary school 0.3706   
## Educationsecondary school 0.4457   
## Age 0.1337   
## Only.childno <2e-16 \*\*\*  
## Only.childyes <2e-16 \*\*\*  
## Village\_towncity 0.9850   
## Village\_townvillage 0.6246   
## buildingblock of flats 0.3893   
## buildinghouse/bungalow 0.2851   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## looks like spending on looks is impacted by Only.child  
  
## now since there is no multicollinearity, we can proceed to model based on spending - let’s take it one by one first spending on eating

**Analyzing the effect of demographics on spending on healthy eating**

model <- polr(eating ~ Gender + Education + Age + Only.child + Village\_town + building, data = SurveyData, Hess = TRUE, method = "logistic")  
coeftest(model)

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value  
## Genderfemale -1.033960 0.813598 -1.2708  
## Gendermale -1.018569 0.815013 -1.2498  
## Educationcollege/bachelor degree -0.632921 1.523072 -0.4156  
## Educationcurrently a primary school pupil -0.476886 1.634616 -0.2917  
## Educationdoctorate degree -0.880127 1.706129 -0.5159  
## Educationmasters degree -0.435274 1.545042 -0.2817  
## Educationprimary school -1.039099 1.528354 -0.6799  
## Educationsecondary school -0.716815 1.517488 -0.4724  
## Age -0.012470 0.027735 -0.4496  
## Only.childno -0.902743 1.442092 -0.6260  
## Only.childyes -0.632125 1.445242 -0.4374  
## Village\_towncity 0.095614 0.955915 0.1000  
## Village\_townvillage 0.030194 0.962079 0.0314  
## buildingblock of flats 0.257051 0.782032 0.3287  
## buildinghouse/bungalow 0.126890 0.788979 0.1608  
## Pr(>|t|)  
## Genderfemale 0.2041  
## Gendermale 0.2117  
## Educationcollege/bachelor degree 0.6778  
## Educationcurrently a primary school pupil 0.7705  
## Educationdoctorate degree 0.6061  
## Educationmasters degree 0.7782  
## Educationprimary school 0.4967  
## Educationsecondary school 0.6368  
## Age 0.6531  
## Only.childno 0.5315  
## Only.childyes 0.6619  
## Village\_towncity 0.9203  
## Village\_townvillage 0.9750  
## buildingblock of flats 0.7425  
## buildinghouse/bungalow 0.8723

## looks like spending on eating is not impacted by any of them

# **Conclusion**

Most of the young people spend few hours a day on internet

the ord plot shows that the data fit the log-series distribution

the chiq-test and CMH relationship tests were significant only for spending on entertainment and gadgets with internet usage

Through our model we saw that gadgets and looks are the two main criteria for spending time on internet where looks were oppositely signified.

For the demographic part – spend on gadgets impacted by person lives in flats or bungalow

spend on entertainment impacted by person lives in flats or bungalow or only child/not

spend on looks impacted by person is only child/not

# **Gaps and Future Efforts**

This data was not sufficiently able to infer internet usage based on just on spending , maybe there are other factors which impacts the internet usage.

Like many people spend time on gaming and that impacts internet usage so need such of more data to get to the root cause.

If internet usage - was collected in continuous variable in time that could have given better picture.

References :

1. book: Discrete Data Analysis with R , Visualization and Modeling techniques for categorical and Count data by Michael Friendly and David Meyer.
2. <https://www.statmethods.net/graphs>
3. <https://stat.utexas.edu/videos/r/statistical-inference>