Data Simulation Project

Uma Ramasubramanian

May 7, 2017

```
set.seed(101)
generateDataset<-function(N){
  data.frame(matrix(0, nrow = N, ncol = 11))
  age<-0
  canc<-0
  cvd<-0
  diab<-0
  hbp<-0
  education<-0
  smoke<-0
  functional.limitation<-0
  physical.activity<-0
  bmi<-0
  sitting<-0
  data.frame(age,canc,cvd,diab,hbp,education,smoke,functional.limitation,physical.activity,bmi,sitting)

  # Age was normally distributed given the mean and standard deviation
  age<-rnorm(N,55.6,5.4)

  # Education was a factor variable with different levels of education
  # 0= none, 1=School Certificate, 2= High School, 3= Trade 4= Certificate/diploma 5= University degree
  education<-factor(sample(0:5, N, replace=TRUE, prob=c(0.07, 0.128, 0.104, 0.172,0.21,0.317)),ordered=T
                    levels=c('0','1','2','3','4','5'))
  # Smoke as binary variable of did or did not smoke
  smoke<-rbinom(N,1,0.48)

  # Functional limitation as factor with 4 levels
  # 0= no limitation , 1 = mild limitation, 2= mod limitation, 3 = severe limitation

  functional.limitation<-factor(sample(0:3, N, replace=TRUE, prob=c(0.445,0.205, 0.169, 0.181)),ordered=
                                levels= c('0','1','2','3'))
  no<-which(functional.limitation==0)
  length.no<-length(no)
  mild<-which(functional.limitation==1)
  length.mild<-length(mild)
  mod<-which(functional.limitation==2)
  length.mod<-length(mod)
  severe<-which(functional.limitation==3)
  length.severe<-length(severe)

  # Sitting is a factor variable with 4 levels
  # Level 1 = 0 to < 4 hours of sitting
  # Level 2 = 4 to < 6 hours of sitting
  # Level 3 = 6 to < 8 hours of sitting
  # Level 4 = >= 8 hours of sitting

  # This variable is correalted with the amount if functional limitation as people with more limitation
```

```r
# for longer hours. This variable was simulated using the functional limitation indices and randomly
# them to different levels based on domain knowledge

sitting <- factor( rep("level4",N), ordered=T,levels =c("level1", "level2", "level3", "level4"))
mod1<-sample(mod,round(length(mod)*0.35),replace = FALSE)
no1<-sample(no,round(length.no*0.30),replace = FALSE)
mild1<-sample(mild,round(length.mild*0.20),replace = FALSE)
sitting[c(no1,mild1,mod1)]<-"level1"


no2<-sample(subset(no, !(no %in% no1)),round(length.no*0.30),replace = FALSE)
mild2<-sample(subset(mild, !(mild %in% mild1)),round(length.mild*0.25),replace = FALSE)
mod2<-sample(subset(mod,!(mod %in% mod1)),round(length.mod*0.10),replace = FALSE)
sitting[c(no2,mild2,mod2)]<-"level2"

mod3<-sample(subset(mod, !(mod %in% c(mod1,mod2)),round(length.mod*0.75),replace = FALSE))
no3<-sample(subset(no, !(no %in% c(no1,no2))),round(length.no*0.35),replace = FALSE)
sitting[c(mod3,no3)]<- "level3"

# The chronic conditions diabetes, canc, cvd and hbp wa simulated using logistic function to take int
# of age and level of sitting

logistic <- function(t) 1 / (1 + exp(-t))
canc <- runif(length(age))< .05*logistic((age-50)/10) + .065*logistic((as.numeric(sitting)-3)/2)
cvd<- runif(length(age))< .06*logistic((age-45)/2) + .07*logistic((as.numeric(sitting)-3)/2)
diab <- runif(length(age))< .07*logistic((age-50)/10) + .086*logistic((as.numeric(sitting)-3)/2)
hbp <- runif(length(age))< .2*logistic((age-45)/2) +.2*logistic((as.numeric(sitting)-3)/2)

# Physical activity is a factor variable with 4 levels. This level is again dependent os the function
# on the minutes people spend with physical activity
# Sed (zero mins)
# Low Active(0-149 mins)
#Sufficiently active(150-299 mins)
#Highly active(300- 539 mins)
# Very highly active( 540 + mins)

physical.activity <- factor( rep("veryhigh",N), ordered=T,levels =c("sed", "low", "suff", "high", "ve
low.mod1<-sample(mod,round(length(mod)*0.8),replace = FALSE)
low.no1<-sample(no,round(length.no*0.03),replace = FALSE)
low.mild1<-sample(mild,round(length.mild*0.05),replace = FALSE)
physical.activity[c(low.no1,low.mild1,low.mod1)]<-"low"

suff.no2<-sample(subset(no, !(no %in% low.no1)),round(length.no*0.28),replace = FALSE)
suff.mild2<-sample(subset(mild, !(mild %in% low.mild1)),round(length.mild*0.45),replace = FALSE)
suff.mod2<-sample(subset(mod,!(mod %in% low.mod1)),round(length.mod*0.05),replace = FALSE)
physical.activity[c(suff.no2,suff.mild2,suff.mod2)]<-"suff"

high.no3<-sample(subset(no, !(no %in% c(no1,no2)),round(length.no*0.1),replace = FALSE))
high.mild3<-sample(subset(mild, !(mild %in% c(low.mild1,suff.mild2)),round(length.mild*0.1),replace =
physical.activity[c(high.no3,high.mild3)]<-"high"

severe.sed<-sample(severe,length.severe,replace = FALSE)
severe.no4<-sample(subset(no, !(no %in% c(low.no1,suff.no2,high.no3)),round(length.no*0.1),replace = I
```

```
  physical.activity[c(severe.sed,severe.no4)]<-"sed"

  # BMI was factor variable with 4 levels of underweight, normal weight, overweight and obese.This
  # variable is also correalted with the level of physical activity. Lower level of physical activity,
  bmi<- factor( rep("normalweight",N), ordered=T,levels =c("underweight","normalweight", "overweight","

  sed<-which(physical.activity=="sed")
  length.sed<-length(sed)
  low<-which(physical.activity=="low")
  length.low<-length(low)
  suff<-which(physical.activity=="suff")
  length.suff<-length(suff)
  high<-which(physical.activity=="high")
  length.high<-length(high)
  veryhigh<-which(physical.activity=="veryhigh")
  length.veryhigh<-length(veryhigh)

  high1<-sample(high,round(length.high*0.01),replace = FALSE)
  veryhigh1<-sample(veryhigh,round(length.veryhigh*0.015),replace = FALSE)
  bmi[c(high1,veryhigh1)]<-"underweight"

  high2<-sample(subset(high, !(high %in% high1)),round(length.high*0.05),replace = FALSE)
  veryhigh2<-sample(subset(veryhigh, !(veryhigh %in% veryhigh1)),round(length.veryhigh*0.05),replace =
  low2<-sample(low,round(length.low*0.6),replace = FALSE)
  sed2<-sample(sed,round(length.sed*0.7),replace = FALSE)
  suff2<-sample(suff,round(length.suff*0.2),replace = FALSE)
  bmi[c(high2,veryhigh2,low2,sed2,suff2)]<-"obese"

  low3<-sample(subset(low, !(low %in% low2),round(length.low*0.1),replace = FALSE))
  sed3<-sample(subset(sed, !(sed %in% sed2),round(length.sed*0.1),replace = FALSE))
  suff3<-sample(subset(suff, !(suff %in% c(suff2)),round(length.suff*0.10),replace = FALSE))
  bmi[c(sed3,suff3,low3)]<-"overweight"

  # In all the above variables, some noise was added as some individuals may over report or underreport
  # Their hours of sitting, physical activity
  data.frame(age,canc,cvd,diab,hbp,education,smoke,functional.limitation,physical.activity,bmi,sitting)
}
```

```
data<-generateDataset(50000)
```

```
# Add new column counting the number of chronic diseases reported
data$chronic.disease<-rowSums(data=='TRUE')
```

```
summary(data)
```

```
##       age           canc            cvd             diab
##  Min.   :33.23   Mode :logical   Mode :logical   Mode :logical
##  1st Qu.:51.99   FALSE:46849     FALSE:45610     FALSE:45858
##  Median :55.58   TRUE :3151      TRUE :4390      TRUE :4142
##  Mean   :55.62   NA's :0         NA's :0         NA's :0
##  3rd Qu.:59.27
##  Max.   :79.74
##     hbp          education     smoke        functional.limitation
##  Mode :logical   0: 3440   Min.   :0.0000   0:22226
##  FALSE:36053     1: 6573   1st Qu.:0.0000   1:10293
```

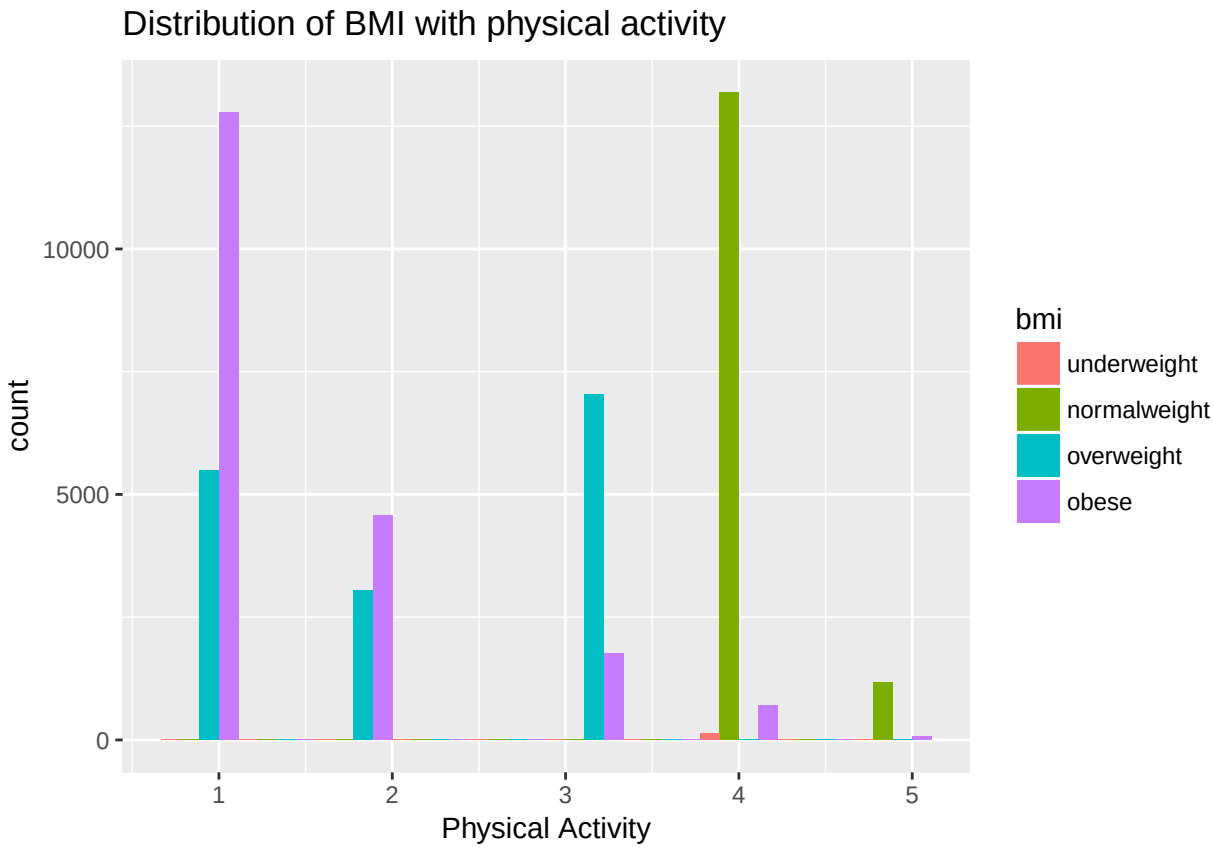
```
##  TRUE :13947     2: 5310   Median :0.0000   2: 8401
##  NA's :0         3: 8568   Mean   :0.4791   3: 9080
##                  4:10492   3rd Qu.:1.0000
##                  5:15617   Max.   :1.0000
##  physical.activity          bmi          sitting      chronic.disease
##  sed     :18271    underweight :  159   level1:11667   Min.   :0.0000
##  low     : 7626    normalweight:14372   level2:10081   1st Qu.:0.0000
##  suff    : 8807    overweight  :15577   level3:12400   Median :0.0000
##  high    :14036    obese       :19892   level4:15852   Mean   :0.5126
##  veryhigh: 1260                                       3rd Qu.:1.0000
##                                                       Max.   :4.0000
```

The data simulated looks close to the researchers data. Some additional noise was added to encounter for patient reporting the values of functional limitation, sitting time and physical activity level in error.

```
library(ggplot2)
```

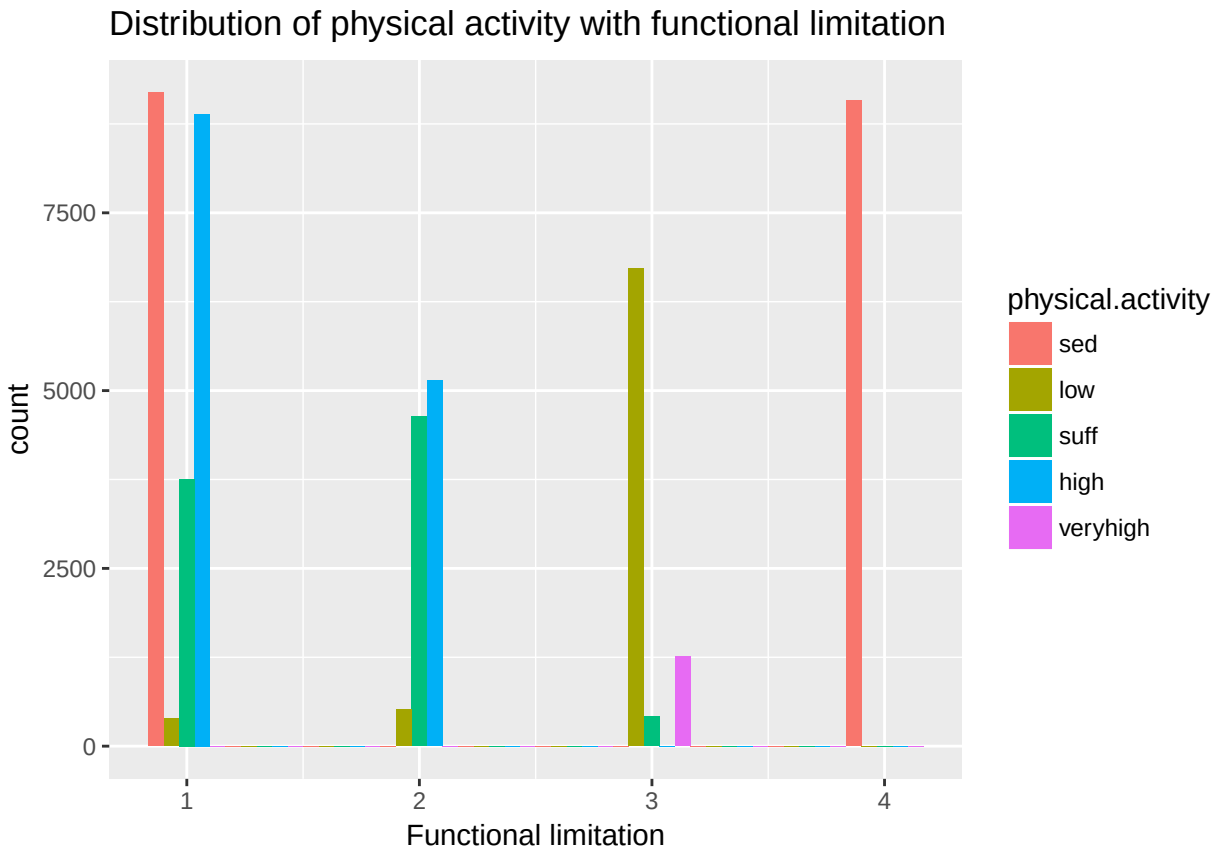
```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

```
p<-ggplot(data, aes(x = as.numeric(physical.activity))) + geom_histogram(aes(fill=bmi),position = 'dodg
p+labs(title="Distribution of BMI with physical activity",x="Physical Activity")
```



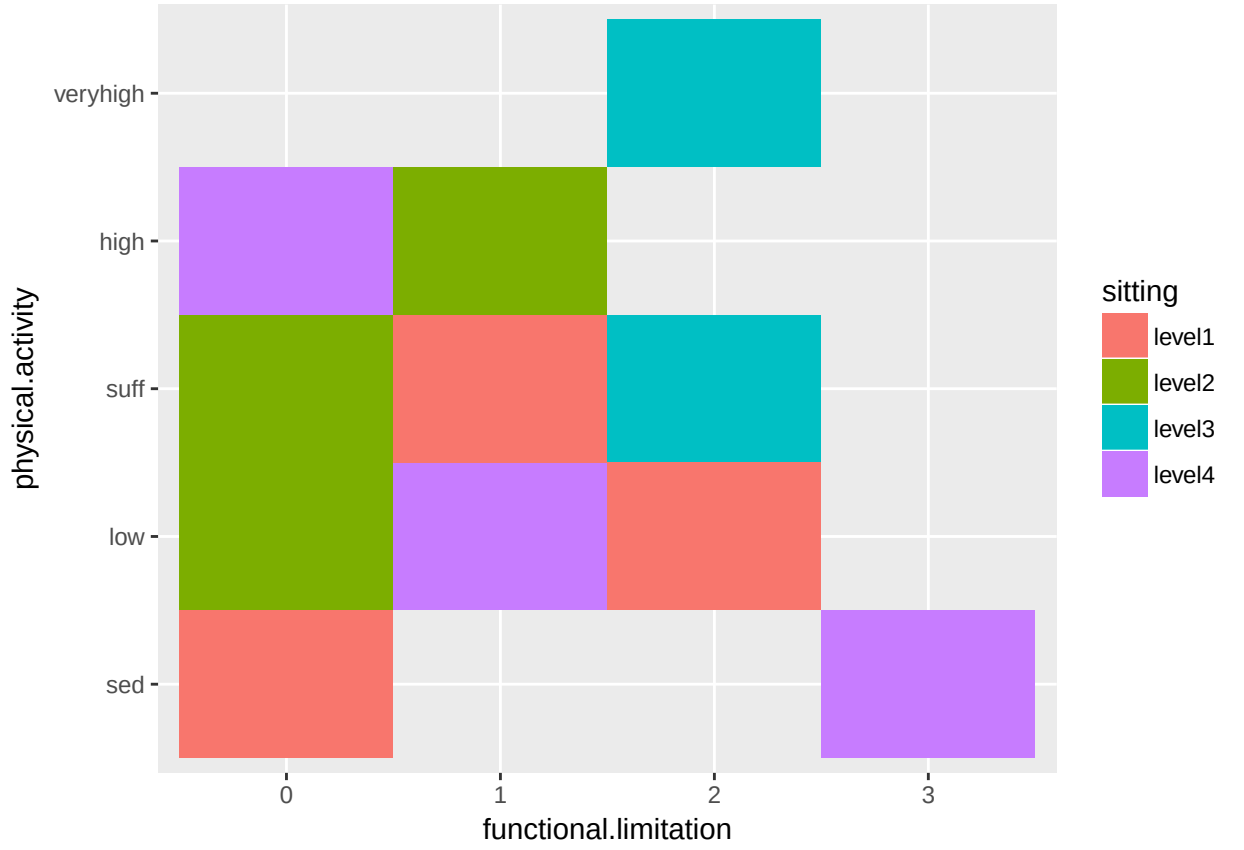
Most of population, who is low in physical activity tend to be overweight or obese. There is distibution of overweight in other levels as well to encounter for error in reporting of the level og physical activity, weight and other factors like genetic make up, muscle mass contributing to higher BMI

```
p<-ggplot(data, aes(x = as.numeric(functional.limitation))) + geom_histogram(aes(fill=physical.activity
p+labs(title="Distribution of physical activity with functional limitation ",x="Functional limitation")
```



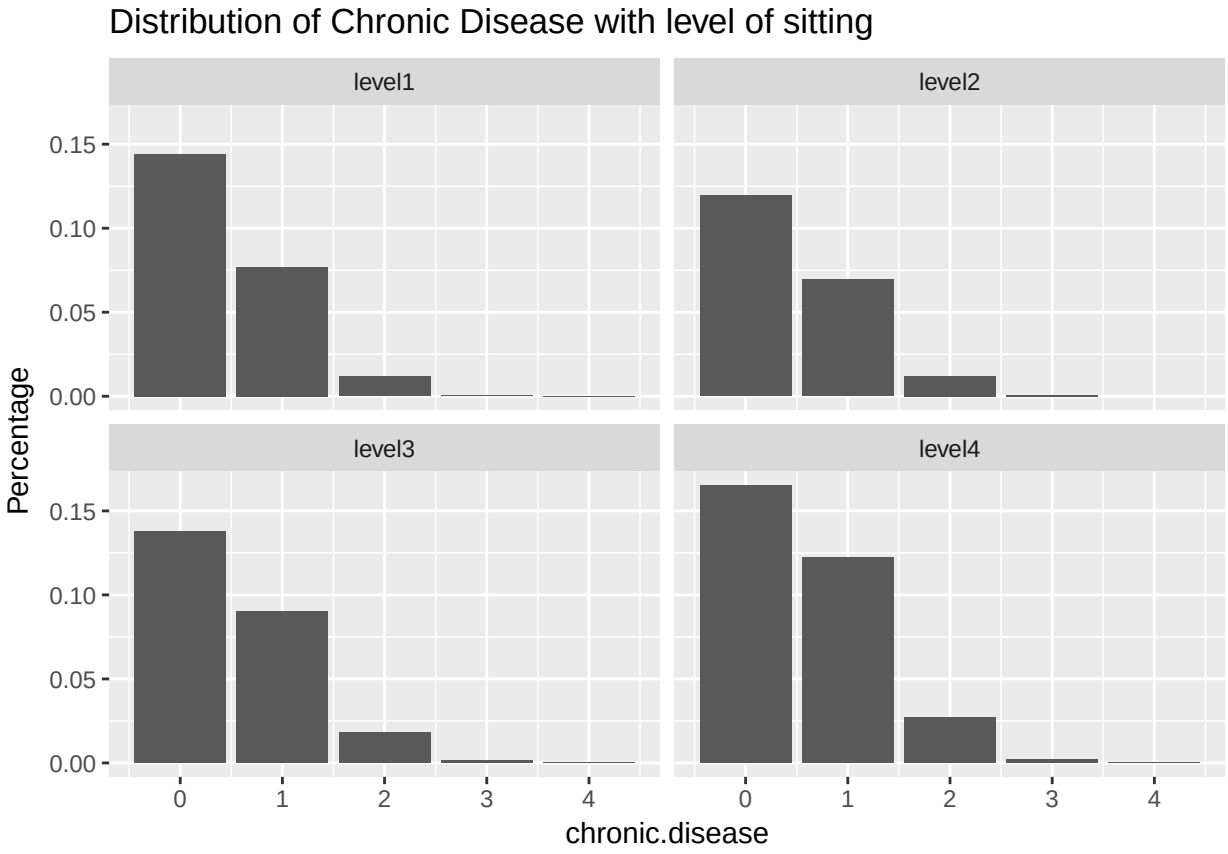
Higher level of functional limitation is assoicated with more sedentary lifestyle.However, there is a large proportion of the population that has no functional limitation and is sedentary.

```
ggplot(data = data, aes(x = functional.limitation, y = physical.activity)) +
  geom_tile(aes(fill = sitting))
```



There is a high percentage of sedentary individuals in all levels of functional limitations. Most of those individuals have higher level of sitting as well.

```
g<-ggplot(data,aes(chronic.disease)) + stat_count(aes(y = ((..count..)/sum(..count..))))
p<-g+ facet_wrap(~sitting)
p + labs (title= 'Distribution of Chronic Disease with level of sitting',y= 'Percentage')
```



As the level of sitting increases, the probability of the person having chronic diseases increases. People with higher level of sitting have higher probability of getting 2 or more conditions

```
data$sitting<-factor(data$sitting, ordered=FALSE)
model1<-glm(formula = canc~sitting, family = binomial(),data = data)
summary(model1)
```

```
##
## Call:
## glm(formula = canc ~ sitting, family = binomial(), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3918  -0.3918  -0.3690  -0.3294   2.4254
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.88703    0.04140 -69.739  < 2e-16 ***
## sittinglevel2  0.03079    0.06036   0.510     0.61
## sittinglevel3  0.23406    0.05501   4.255 2.09e-05 ***
## sittinglevel4  0.35833    0.05134   6.980 2.96e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 23520  on 49999  degrees of freedom
```

```
## Residual deviance: 23453  on 49996  degrees of freedom
## AIC: 23461
##
## Number of Fisher Scoring iterations: 5
```

The model on the whole is significant and level 4 of sitting is slightly significant to increase the odds of developing cancer by a factor of 0.3 and level 3 by a factor of 0.2

```
model2<-glm(formula = diab~sitting, family = binomial(),data = data)
summary(model2)
```

```
##
## Call:
## glm(formula = diab ~ sitting, family = binomial(), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4521  -0.4521  -0.4275  -0.3715   2.3271
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.63878    0.03711 -71.106  < 2e-16 ***
## sittinglevel2  0.10472    0.05323   1.967   0.0492 *
## sittinglevel3  0.29221    0.04888   5.979 2.25e-09 ***
## sittinglevel4  0.40947    0.04579   8.943  < 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: 28565  on 49999  degrees of freedom
## Residual deviance: 28467  on 49996  degrees of freedom
## AIC: 28475
##
## Number of Fisher Scoring iterations: 5
```

Diabetes is highly correlated with all levels of sitting. On observing the coefficents we can see that the odds of developing diabetes increases as the level of sitting increases. The coefficents of higher level of sitting is almost 3 times of level 2 sitting

```
model3<-glm(formula = cvd~sitting , family = binomial(),data = data)
summary(model3)
```

```
##
## Call:
## glm(formula = cvd ~ sitting, family = binomial(), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4509  -0.4509  -0.4337  -0.4007   2.2637
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.48185    0.03470 -71.526  < 2e-16 ***
## sittinglevel2  0.09032    0.04996   1.808 0.070616 .
## sittinglevel3  0.16525    0.04681   3.530 0.000415 ***
```

```
## sittinglevel4  0.24676    0.04389   5.622 1.89e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 29742  on 49999  degrees of freedom
## Residual deviance: 29707  on 49996  degrees of freedom
## AIC: 29715
##
## Number of Fisher Scoring iterations: 5
```

Higher level of sitting is associated with higher probability of developing cardiovascular disease. The coefficent of level 4 sitting is almost 0.2 ie sitting for greater than 8 hours increases the log odd of developing cardiovascular disease by a factor of 0.24, whereas level 3 sitting(6 to $<$ 8 hrs of sitting) increases th odds y a factor of 0.16

```
model4<-glm(formula = hbp~sitting, family = binomial(),data = data)
summary(model4)
```

```
##
## Call:
## glm(formula = hbp ~ sitting, family = binomial(), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8644  -0.8204  -0.7805   1.5268   1.6863
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.14568    0.02164 -52.947  < 2e-16 ***
## sittinglevel2  0.11304    0.03131   3.610 0.000306 ***
## sittinglevel3  0.22945    0.02938   7.809 5.77e-15 ***
## sittinglevel4  0.35377    0.02761  12.814  < 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: 59195  on 49999  degrees of freedom
## Residual deviance: 59011  on 49996  degrees of freedom
## AIC: 59019
##
## Number of Fisher Scoring iterations: 4
```

Sitting for greater than 8 hours increases the odds of developing high blood pressure by a factor of 0.35, level 3 sitting increases the odds by 0.23, level 2 sitting increases the odds by 0.11

```
data$chronic.disease<-as.factor(ifelse(data$chronic.disease==0,0,1))
model5<- glm(chronic.disease~sitting,family = binomial(),data=data)
summary(model5)
```

```
##
## Call:
## glm(formula = chronic.disease ~ sitting, family = binomial(),
##     data = data)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1411  -1.0828  -0.9822   1.2751   1.3860
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.47809    0.01905 -25.100  < 2e-16 ***
## sittinglevel2  0.10254    0.02782   3.686 0.000227 ***
## sittinglevel3  0.25132    0.02626   9.570  < 2e-16 ***
## sittinglevel4  0.39199    0.02481  15.799  < 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: 68417  on 49999  degrees of freedom
## Residual deviance: 68130  on 49996  degrees of freedom
## AIC: 68138
##
## Number of Fisher Scoring iterations: 4
```

The above model, clearly shows that sitting hours is correlated with the development of chronic diseases. Higher the hours of sitting high the odds of developing chronic diseases

```
data$functional.limitation<-factor(data$functional.limitation, ordered=FALSE)
data$physical.activity<-factor(data$physical.activity, ordered=FALSE)

model6<- glm(chronic.disease~sitting + age + physical.activity + functional.limitation + smoke,family =
summary(model6)
```

```
##
## Call:
## glm(formula = chronic.disease ~ sitting + age + physical.activity +
##     functional.limitation + smoke, family = binomial(), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3467  -1.0725  -0.9667   1.2633   1.5242
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.553100   0.098249 -15.808  < 2e-16 ***
## sittinglevel2               0.105153   0.028355   3.708 0.000209 ***
## sittinglevel3               0.232682   0.034157   6.812 9.62e-12 ***
## sittinglevel4               0.385137   0.037719  10.211  < 2e-16 ***
## age                         0.019219   0.001694  11.344  < 2e-16 ***
## physical.activitylow        0.077897   0.063378   1.229 0.219039
## physical.activitysuff      -0.019251   0.035597  -0.541 0.588643
## physical.activityhigh       0.025641   0.038606   0.664 0.506577
## physical.activityveryhigh   0.019588   0.086914   0.225 0.821690
## functional.limitation1      0.001241   0.034509   0.036 0.971322
## functional.limitation2     -0.039755   0.061409  -0.647 0.517380
## functional.limitation3      0.015298   0.045940   0.333 0.739129
## smoke                      -0.001896   0.018144  -0.104 0.916783
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 68417  on 49999  degrees of freedom
## Residual deviance: 67995  on 49987  degrees of freedom
## AIC: 68021
##
## Number of Fisher Scoring iterations: 4
```

Based on the above summary, we can see that the higher level of sitting and age are the significant variables increasing the odds of developing chronic disease. Though age was added to the model to predict chronic diseases, it did not decrease the impact of sitting level on the development of chronic diseases.Physical activity and functional limitations did not conrtribute to the development of chronic diseases.Even smoking which is found to be high risk for chronic diseases like canc, cardiovascular diseases was not as significant as sitting.

Conclusion: Higher level of sitting hours increases the odds of developing chronic conditions. The number of chronic conditions reported also increases as the amount of sitting hours increases. It is important for individuals with higher level of sitting to increase the level of physical activity, to decrease the chances of developing chronic conditions. Regular breaks every hour of sitting is found the decrease the risk significantly.

Limitations: As the study was cross-sectional in nature, we cannot establish whether the volume of sitting time led to the development of these chronic diseases, or whether the presence of these chronic diseases influenced participants' sitting time. The potential for misclassification of the variables used in this analysis must be acknowledged. It is possible that some participants may have incorrectly reported (or failed to report) having a chronic disease, while others may have under- or over-reported their daily sitting time. While these potential misclassifications may have impacted upon the strength of the observed associations, even after adjusting for a range of covariates, sitting time was still strongly and significantly associated with development of chronic diseases. In addition, only data on overall time spent in physical activity in the previous week were included in this study. It is also possible that a proportion of the moderate intensity activity reported in the 45 and Up Study baseline questionnaire was actually light intensity, which may have led to an overestimation of moderate intensity physical activity. The third potential limitation is that the sitting time variables did not delineate specific domains of sitting time, such as office work, driving, other passive travel, and sitting during leisure time.