Case Study: Predict Mental Health for Young Americans using GSS Data

Xiao LIANG, Yuxuan LI, Yuying TAN

All group members, including Xiao LIANG, Yuxuan LI, and Yuying Tan, have contributed to this case study equally.

Introduction

Mental disease is distinct to physical illness in nature. On one hand, the outbreak of it has been brought forward dramatically. The prevalence of young Americans aged 18 to 25 with a major depressive episode has climbed from 8.4% to 14.6% in the past decade, the largest increase among all age intervals.[1] One the other hand, early diagnosis would make a great difference. Treatments have been proved effective on improving physical symptoms, emotion management, and social functioning at early stages of mental disorders.[2]

Although medical researchers found a first-degree relative with mental diseases to be the most predictive factor,[3] we would like to flag individuals at high risks by a big-data approach and bring back to light those who might skip primary care due to financial or societal issues. The US government conducts the **General Social Survey (GSS)** every year since 1972. The data collected are readily accessible online, and contain demographic, socio-economic, and life-style information of people who were interviewed in person or by phone. Because these people are randomly sampled from the entire population of US residents, conclusions drawn from this dataset can be generalized at large.[4]

In this case study, we build a **zero-inflated negative binomial (ZINB)** regression model to predict the number of days in a month that one feels mentally unhealthy and at least possibly seeks for help, using the GSS data. The analysis focuses on American residents aged 18 to 24 only, since disease patterns differ remarkably across age groups. We then show by a 5-fold cross-validation that our model makes reasonably reliable predictions and promotes decision-making.

```
# load the original dataset
load("brfss2013.RData")
# select relevant variables and observations
DATA_proj <- brfss2013 %>%
  filter(X_age_g=="Age 18 to 24")%>%
  select(cellfon2, cellfon3, sex, income2, height3, weight2,
         X_state,children,sleptim1,menthlth,smoke100,X_educag,
         X_race,internet,medcost,maxdrnks,fruit1,vegetab1,exeroft1)
# remove the redundant dataset
```

[1] https://www.nimh.nih.gov/health/statistics/major-depression.shtml
[2] https://www.ajmc.com/journals/supplement/2007/2007-11-vol13-n4suppl/nov07-2638ps092-s097?p=2
[3] https://www.bcmj.org/articles/early-detection-depression-young-and-elderly-people
[4] https://gss.norc.org/

```
save(DATA_proj, file="DATA_project.RData")
rm(brfss2013)
```

Data Preparation

The original dataset is massive, including 330 variables and 491 775 observations. Combining the codebook with suggestions given by psychologists, nine numerical and nine categorical variables are picked out in the preliminary selection. All the variables concerning chronic symptoms and physical disturbance are excluded in the first place, on the grounds that they usually affect the elderly only.

NUMERICAL	Description
mentlth	The number of days ones feels mentally unhealthy during the past month
height3	The height of an individual in feet and inches
weight2	The weight of an individual in pounds
children	The number of children one has
sleptim1	The average number of hours one sleeps per day during the past month
maxdrnks	The max number of glasses of alcohol per day during the past month
fruit1	The average grams of raw fruit one takes in per day during the past month
vegetab1	The average grams of vegetables per day during the past month
exeroft1	The average intensity of exercises during the past month

CATEGORICAL	Description
X_state	The state one resides in
X_race	The race one identifies oneself with
sex	The biological gender
X_educag	The highest education level ones has completed
income2	The level of family income per year
cellfon2	Whether one owns a private cellphone
internet	Whether one accesses the internet during the past month
medcost	Whether one has unpaid medical bills
smoke100	Whether one smokes during the past 100 days

We start by transforming variables of interest into more informative formats. For example, dividing 50 states, a federal district, and five territories into seven geographical categories facilitates the comparison of effect. And health-relevant knowledge contained in height and weight can be more succinctly represented by Body Mass Index (BMI).

$$BMI = \frac{weight}{height^2} \times 703$$

```
load("DATA_project.RData")
# Data transformation
DATA <- DATA_proj %>%
  mutate(cellfon2=ifelse(!is.na(cellfon3), 0, cellfon2)) %>%
  mutate(cellphone=factor(ifelse(cellfon2==0,"No","Yes"))) %>%
  mutate(GENDER=factor(ifelse(sex=="Female", "F", "M"))) %>%
  mutate(Class=ifelse(income2 %in% c("Less than $75,000","$75,000 or more"),
                      "upper-middle","middle")) %>%
  mutate(Class=ifelse(income2 %in% c("Less than $10,000","Less than $15,000",
                                     "Less than $20,000"),"lower",Class)) %>%
  mutate(Class=factor(Class)) %>%
  mutate(height3=as.numeric(as.character(height3))) %>%
  mutate(weight2=as.numeric(as.character(weight2))) %>%
  mutate(HEIGHT=floor(height3/100)*12+height3-100*floor(height3/100)) %>%
  mutate(BMI = 703*weight2/HEIGHT^2) %>%
  mutate(DISTRICT=ifelse(X_state%in%c("Maine","Massachusetts","New Hampshire",
                                      "Rhode Island","Vermont","Connecticut"),
                         "New Eng","Others")) %>%
  mutate(DISTRICT=ifelse(X_state%in%c("Illinois","Indiana","Iowa","Michigan",
                                      "Ohio","Wisconsin","Minnesota"),
                         "Lakes",DISTRICT)) %>%
  mutate(DISTRICT=ifelse(X_state%in%c("New York","Pennsylvania","Maryland",
                                      "New Jersey","Delaware",
                                      "District of Columbia"),
                         "Pacific",DISTRICT)) %>%
  mutate(DISTRICT=ifelse(X_state%in%c("Virginia","West Virginia","Kentucky",
                                      "North Carolina","South Carolina",
                          "Georgia","Florida","Alabama","Louisiana",
                          "Mississippi","Tennessee","Arkansas","Missouri"),
                         "South",DISTRICT)) %>%
  mutate(DISTRICT=ifelse(X_state%in%c("Texas","Oklahoma","Kansas","Nebraska",
                                      "South Dakota","North Dakota","Montana",
                                      "Idaho","Wyoming","Utah","Colorado",
                                      "Arizona","New Mexico","Alaska"),
                         "Midwest",DISTRICT)) %>%
  mutate(DISTRICT=ifelse(X_state%in%c("California","Oregon","Washington",
                                      "Nevada","Hawaii"),"West",DISTRICT))
# remove missing data and outliers
DATA <- DATA %>%
  filter(weight2<=330, HEIGHT<=80, sleptim1>1, BMI>10, BMI<40) %>%
  filter(!is.na(menthlth), !is.na(sleptim1), !is.na(children)) %>%
  filter(!is.na(Class),!is.na(GENDER),!is.na(fruit1),!is.na(X_race)) %>%
  filter(!is.na(DISTRICT),!is.na(BMI),!is.na(cellphone),!is.na(medcost)) %>%
  filter(!is.na(smoke100), !is.na(maxdrnks), !is.na(internet)) %>%
  filter(!is.na(exeroft1), !is.na(X_educag), !is.na(vegetab1))
# rearrange the data frame
```

```
DATA <- DATA %>%
  select(menthlth, DISTRICT, X_race, GENDER, X_educag, Class, cellphone,
         internet, medcost, BMI, children, sleptim1, smoke100, maxdrnks,
         fruit1, vegetab1, exeroft1)
save(DATA, file="DATA_cleaned.RData")
```

After removing missing data and obvious outliers (e.g. averagely sleeping one hour per day during the past month), we now work with a dataset of 17 columns and 9 673 rows. It then is split into five subsets of roughly equal size to prepare for the cross-validation later.

```
# k-fold cross validation
k <- 5
set.seed(222)
# split the sample IDs into k subsets
N <- nrow(DATA)
n <- floor(N/k)
ID <- list()
pool <- 1:N
for (i in 1:(k-1))
{
  ID[[i]] <- sample(pool, n, replace=FALSE)
  pool <- setdiff(pool,ID[[i]])
}
ID[[k]] <- pool
rm(pool)
```

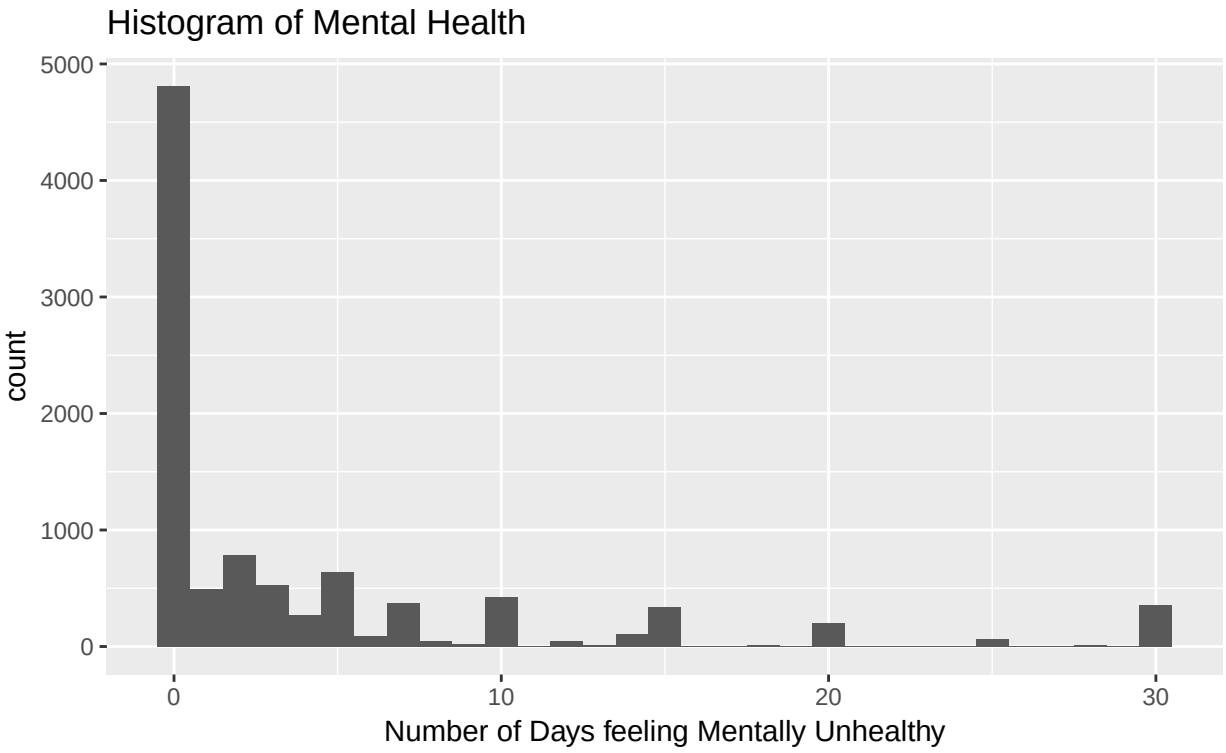
Response Variable and ZINB

Our response variable, `menthlth`, is distributed in a peculiar shape, in that (i) there is a fat tail and (ii) about half of the response are zeros. The ordinary models for count data, such as Poisson regression and binomial regression, are therefore no more suitable. To address the overdispersion and the excess of zeros, we decide to model `menthlth` with the **zero-inflated negative binomial (ZINB)** regression.

```
# compute the proportion of respondents who don't feel mentally unhealthy at all
prop <- round(nrow(DATA %>% filter(menthlth==0)) / N, digits=4)
cat(100*prop,"% of the responses are zeros.")
```

```
## 49.71 % of the responses are zeros.
```

```
# plot the histogram of the number of days feeling mentally unhealthy
ggplot(data=DATA, aes(x=menthlth)) +
  geom_histogram(binwidth=1) +
  xlab("Number of Days feeling Mentally Unhealthy") +
  ggtitle("Histogram of Mental Health")
```



In ZINB, the zero responses are not only contributed by the count process but also a binary process. The latter indicates whether or not an individual is genetically susceptible to mental illness and/or under detrimental circumstances, which we plan to predict by the logistic regression with long-term to permanent factors, such as gender, income, race, and so on. The former only applies when the latter "succeeds" and communicates how many mentally unhealthy days that the intrinsic or extrinsic adversity would elicit in a month, which we intend to predict by the negative binomial regression majorly with short-term to mid-term factors, such as diet, BMI, exercises.

Suppose the j^{th} respondent is exposed to triggering circumstances with probability π_j and feels mentally unhealthy in Y_j days during the past month, then

$$\begin{cases} P(W_j = 1) = \pi_j \\ P(W_j = 0) = 1 - \pi_j \end{cases}$$

where W_j is the binary indicator of exposure, and

$$\begin{aligned} P(Y_j = 0) &= P(W_j = 0) \times P(Y_j = 0|W_j = 0) + P(W_j = 1) \times P(Y_j = 0|W_j = 1) \\ &= (1 - \pi_j) \times 1 \ + \ \pi_j \times p_j^{r_j} \end{aligned}$$

$$\begin{aligned} P(Y_j = k) &= P(W_j = 0) \times P(Y_j = k|W_j = 0) + P(W_j = 1) \times P(Y_j = k|W_j = 1) \\ &= 0 + \pi_j \cdot P(Y_j = k|W_j = 1) \\ &= \pi_j \cdot \binom{k + r_j - 1}{r_j - 1} p_j^{r_j} (1 - p_j)^k \end{aligned}$$

in which $k > 0$ and $Y_j|W_j = 1$ follows the negative binomial distribution $NB(r_j, p_j)$.

$$E(Y_j) = \pi_j \cdot E(Y_j|W_j = 1) = \frac{\pi_j p_j r_j}{1 - p_j}$$

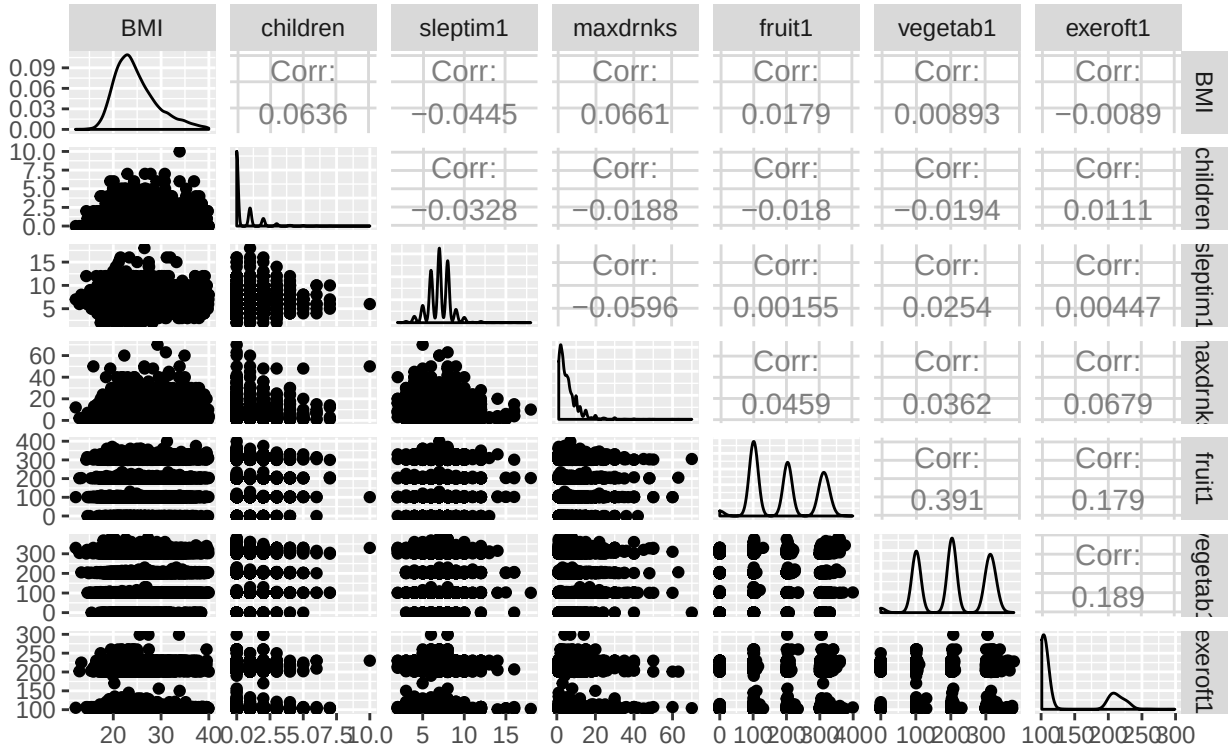
In short, there are three parameters that link the response to the predictors, namely, π, r, and p. Note that the interpretation of r and p is not straightforward. But if we assume that r independent steps are required to conquer the psychological difficulties and one is equally likely to succeed in each attempt at each step, the negative binomial model is not outrageously unreasonable.

Exploratory Data Analysis

The purpose of exploratory data analysis (EDA) is a rough idea of the dependency of the response upon each predictor and the interrelationship among predictors, so that the pool of candidates might be reduced to more managable size.

Numerical Predictors

```
ggpairs(DATA, columns=c(10:12,14:17)) # explore pairwise
```



Other than the plausible correlation between `fruit1` and `vegetab1`, and the possible correlations between `exeroft1` and them, no colinearity among these numerical explanatory variables are found. We thereby construct by **principal component analysis (PCA)** a new variable, `DIET`, to extract 70% of the variances from `fruit1` and `vegetab1`.

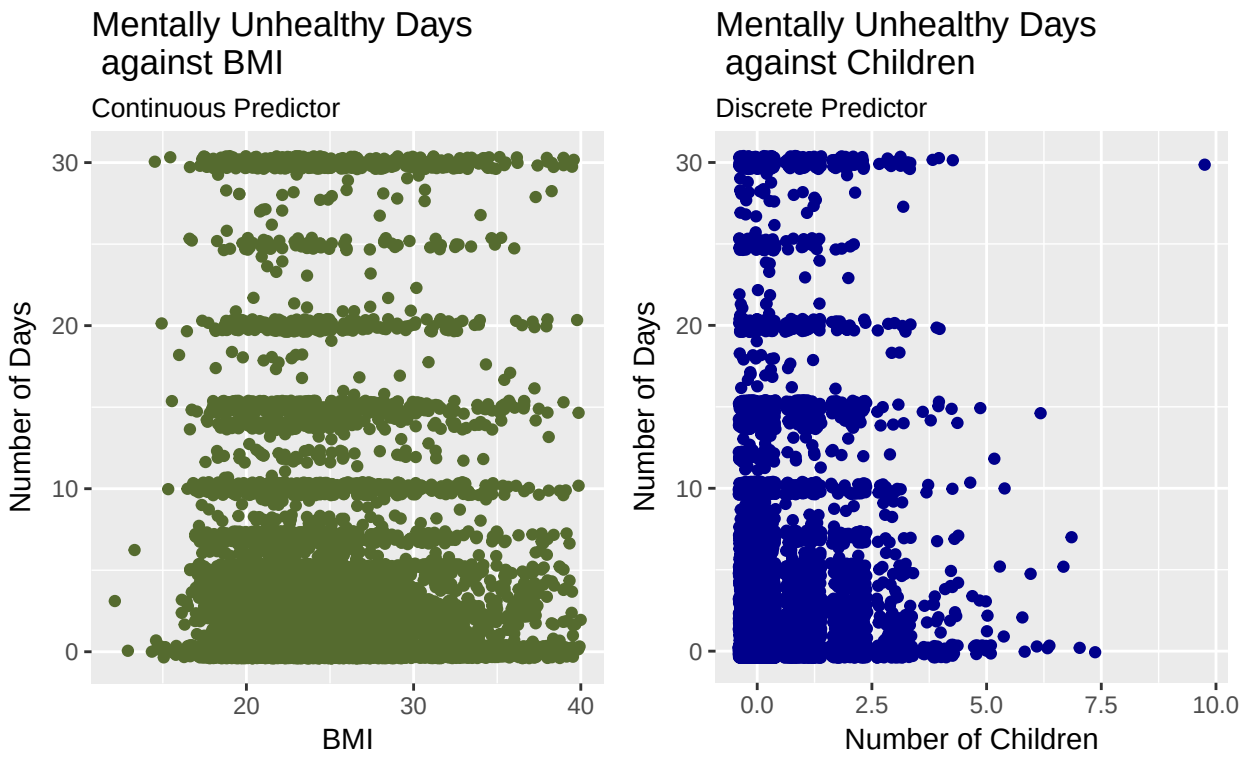
```
# project two variables onto a one-dimension subspace
diet.pca <- prcomp(DATA[,15:16], center = TRUE,scale. = TRUE)
# add the projection to the dataset
```

```
DATA <- DATA %>%
  mutate(DIET=diet.pca$x[,1])
summary(diet.pca)  # check how much variance the projection can explain
```

```
## Importance of components:
##                           PC1    PC2
## Standard deviation     1.1794 0.7804
## Proportion of Variance 0.6955 0.3045
## Cumulative Proportion  0.6955 1.0000
```

It becomes meaningless to compute the correlations between `menthlth` and each predictor as their relationships in ZINB regression are not presumed to be linear any more. Besides, scatterplots hardly display any patterns, regardless of continuous or discrete predictors. This might be resulted by the discreteness and the extreme skewedness of `menthlth`.

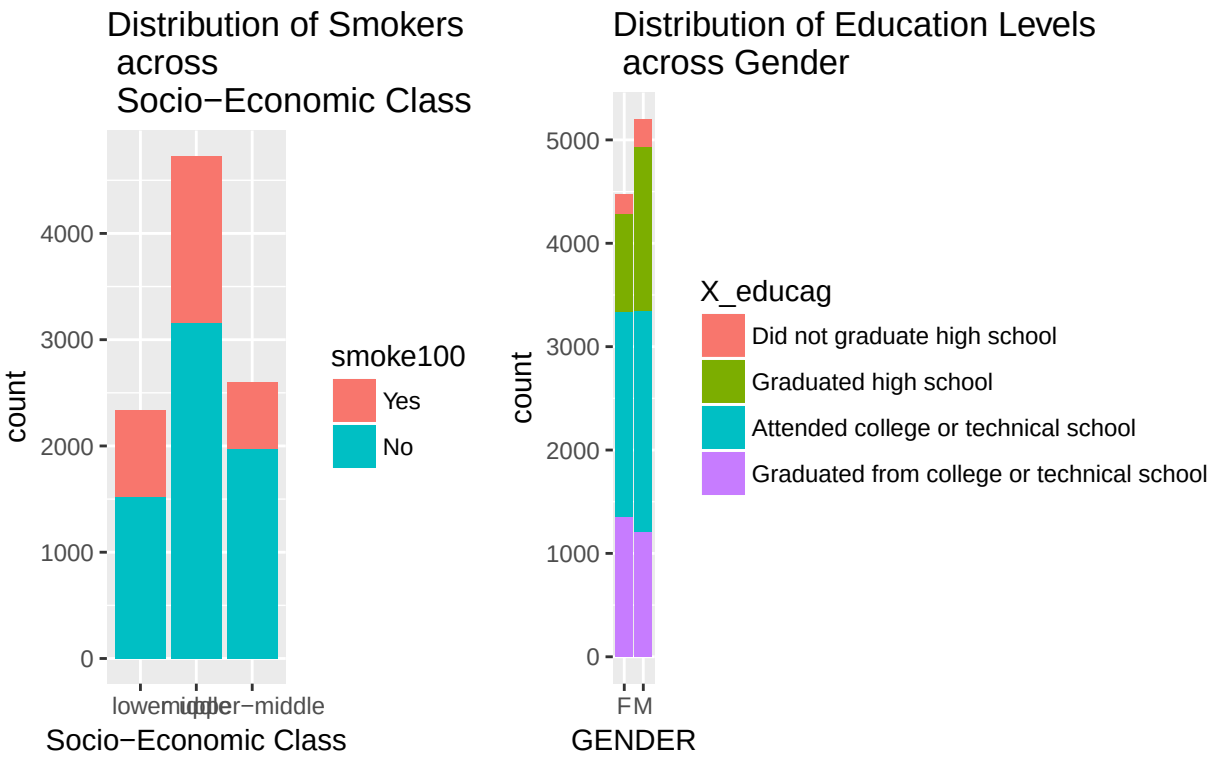
```
g1 <- ggplot(DATA, aes(x=BMI,y=menthlth)) +
        geom_jitter(color="darkolivegreen") + ylab("Number of Days") +
        ggtitle(label=paste("Mentally Unhealthy Days \n against BMI"),
                subtitle="Continuous Predictor")
g2 <- ggplot(DATA, aes(x=children,y=menthlth)) +
        geom_jitter(color="darkblue") +
        ylab("Number of Days") + xlab("Number of Children") +
        ggtitle(label=paste("Mentally Unhealthy Days \n against Children"),
                subtitle="Discrete Predictor")
grid.arrange(g1,g2,nrow=1)
```



Categorical Predictors

Two of the pairwise segmented barplots among categorical predictors suggest statistical dependency between `Class` and `smoke100`, and between `GENDER` and `X_educag`. In specific, the higher the socio-economic class, the fewer people smoke during the past 100 days. And a larger proportion of females have at least obtained some post-secondary education than males have. Given the sample size, $N = 9673$, the differences are rather considerable.

```
# plot smoke100 against socioeconomic class
g3 <- ggplot(DATA, aes(x=Class,fill=smoke100)) +
        geom_bar() +
        xlab("Socio-Economic Class") +
        ggtitle("Distribution of Smokers \n across \n Socio-Economic Class")
# plot education against gender
g4 <- ggplot(DATA, aes(x=GENDER,fill=X_educag)) +
        geom_bar() +
        ggtitle("Distribution of Education Levels \n across Gender")
grid.arrange(grobs=list(g3,g4), widths=c(1.5,2.2),
             layout_matrix=matrix(c(1,2),nrow=1,ncol=2))
```



That said, we decide to retain all of these four predictors and be wary. This is to avoid premature information loss as they describe an individual from distinct perspectives. We do expect at least one to be insignificant when a pair are regressed upon.

Furthermore, `DISRICT` and `X_race` are excluded from the candidates as `menthlth` only fluctuates negligibly across their levels compared to the intrinsic volatility.

```
# summarize the average and standard deviation
# across levels of DISTRICT
kable(DATA %>%
        group_by(DISTRICT) %>%
        summarise(count=n(), avg_mental=mean(menthlth), sd_mental=sd(menthlth)),
      digits=4)
```

DISTRICT	count	avg_mental	sd_mental
Lakes	1661	4.1698	6.9738
Midwest	2803	3.9051	6.9510
New Eng	1017	4.5585	7.2009
Others	185	3.7784	7.8060
Pacific	1112	4.2599	7.2384
South	1819	4.4233	7.5790
West	1076	4.6599	7.5986

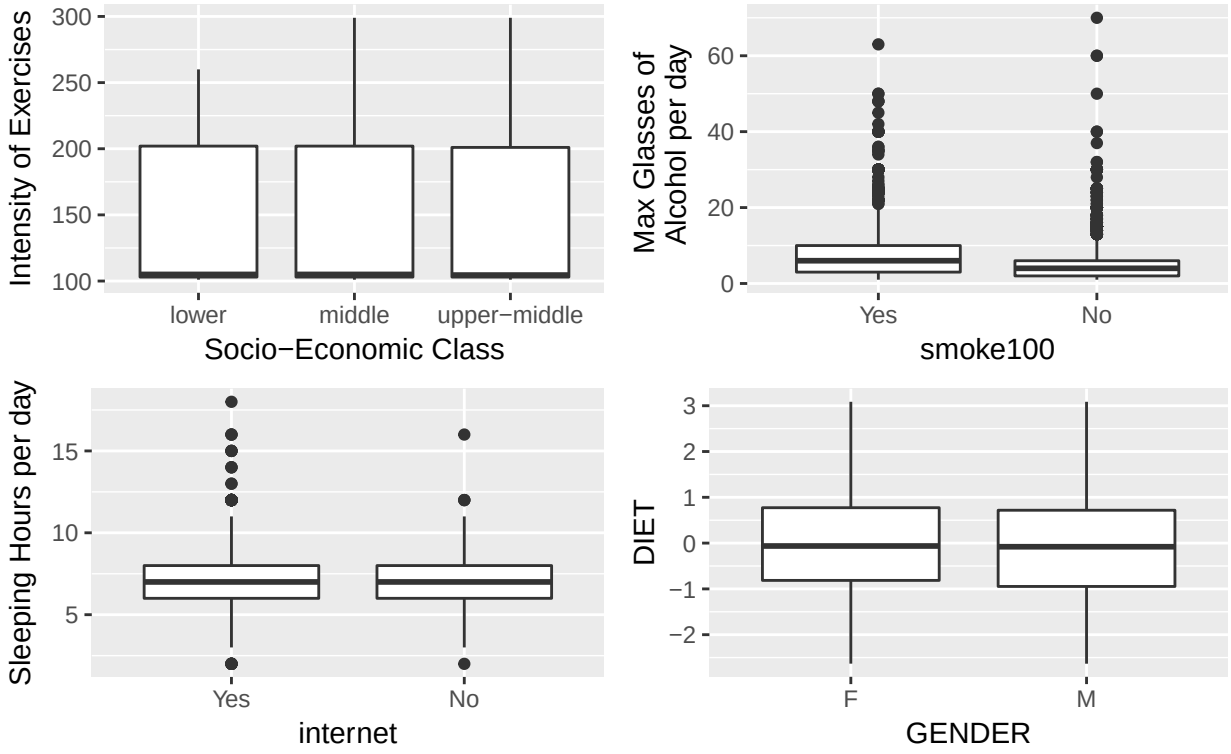
```
# summarize the average and standard deviation
# across levels of race
kable(DATA %>%
        group_by(X_race) %>%
        summarise(count=n(), avg_mental=mean(menthlth), sd_mental=sd(menthlth)),
      digits=4)
```

X_race	count	avg_mental	sd_mental
White only, non-Hispanic	6950	4.1335	7.0444
Black only, non-Hispanic	624	4.5112	7.8960
American Indian or Alaskan Native only, Non-Hispanic	152	5.5987	8.7362
Asian only, non-Hispanic	305	3.2393	6.0283
Native Hawaiian or other Pacific Islander only, Non-Hispanic	72	4.5278	8.4069
Other race only, non-Hispanic	58	6.7069	9.6337
Multiracial, non-Hispanic	363	5.4490	8.2622
Hispanic	1149	4.2898	7.3823

Numerical VS Categorical

We also investigate a few possibilities of numerical explanatory variables depending upon categorical ones. Do smokers drink more alcohol as well? Are people in higher socio-economic class in favor of harder sports? Would female eat healthier? Is the use of internet squeezing out the sleeping time? These questions are brought up on basis of commonsense, yet the barplots provide no marked evidence.

```
g5 <- ggplot(DATA, aes(x=Class, y=exeroft1)) +
  geom_boxplot() +
  xlab("Socio-Economic Class") +
  ylab("Intensity of Exercises")
g6 <- ggplot(DATA, aes(x=internet, y=sleptim1)) +
  geom_boxplot() +
  ylab("Sleeping Hours per day")
g7 <- ggplot(DATA, aes(x=smoke100, y=maxdrnks)) +
  geom_boxplot() +
  ylab("Max Glasses of \n Alcohol per day")
g8 <- ggplot(DATA, aes(x=GENDER, y=DIET)) +
  geom_boxplot()
grid.arrange(grobs=list(g5,g6,g7,g8), widths=c(1,1),
             layout_matrix=matrix(1:4,nrow=2,ncol=2))
```



As a result, we manage to prune the candidates to be six numerical and seven categorical variables.

NUMERICAL	Description
BMI	The body mass index that reflects the relationship between height and weight
children	The number of children one has
sleptim1	The average number of hours one sleeps per day during the past month
maxdrnks	The max number of glasses of alcohol per day during the past month
DIET	The index constructed by PCA from fruit1 and vegetab1
exeroft1	The average intensity of exercises during the past month

CATEGORICAL	Description
GENDER	The biological gender
X_educag	The highest education level ones has completed
Class	The socio-economic class determined by annual family income
cellphone	Whether one owns a private cellphone
internet	Whether one accesses the internet during the past month
medcost	Whether one has unpaid medical bills
smoke100	Whether one smokes during the past 100 days

Backward Model Selection

We start with the full model and drop the most insignificant predictor, if any, until every p-value is sufficiently small. As psychologists recommend, the full model will include the **interaction** between `GENDER` and `children`, `GENDER` and `sleptim1`, `GENDER` and `smoke100`, and `BMI` and `maxdrnks`.

Again, the central idea is to use long-term to permanent factors for the binary process and short-term to mid-term factors for the count process in **ZINB regression**.

```
# randomly pick one subset for testing
# select model with the rest of observations
set.seed(999)
j <- sample(1:k, 1)
TRAIN <- DATA[setdiff(1:N,ID[[j]]),]
# train the full model
m1 <- zeroinfl(menthlth ~ GENDER + children + BMI + sleptim1 + maxdrnks +
                          DIET + exeroft1 + GENDER:children +
                          GENDER:sleptim1 + BMI:maxdrnks
                        | GENDER + X_educag + Class + cellphone +
                          internet + medcost + smoke100 + BMI + GENDER:smoke100,
               data = TRAIN, dist = "negbin", EM = TRUE)
summary(m1)
```

```
##
## Call:
## zeroinfl(formula = menthlth ~ GENDER + children + BMI + sleptim1 + maxdrnks +
##     DIET + exeroft1 + GENDER:children + GENDER:sleptim1 + BMI:maxdrnks |
##     GENDER + X_educag + Class + cellphone + internet + medcost + smoke100 +
##         BMI + GENDER:smoke100, data = TRAIN, dist = "negbin", EM = TRUE)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -0.8012 -0.5521 -0.4693  0.1317  9.6019
##
## Count model coefficients (negbin with log link):
```

```
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.9161208  0.1825880  10.494  < 2e-16 ***
## GENDERM          -0.0296226  0.1651006  -0.179 0.857607
## children          0.0643921  0.0294840   2.184 0.028964 *
## BMI               0.0159308  0.0051086   3.118 0.001818 **
## sleptim1         -0.0840437  0.0159713  -5.262 1.42e-07 ***
## maxdrnks          0.0672903  0.0165125   4.075 4.60e-05 ***
## DIET             -0.0172969  0.0151980  -1.138 0.255078
## exeroft1          0.0014303  0.0003682   3.885 0.000102 ***
## GENDERM:children -0.0053208  0.0434346  -0.123 0.902503
## GENDERM:sleptim1 -0.0353985  0.0229001  -1.546 0.122158
## BMI:maxdrnks     -0.0019054  0.0005836  -3.265 0.001094 **
## Log(theta)       -0.1905160  0.0450034  -4.233 2.30e-05 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                                                      Estimate Std. Error z value
## (Intercept)                                         -2.691338   0.285612  -9.423
## GENDERM                                              0.790621   0.126842   6.233
## X_educagGraduated high school                       -0.044587   0.153003  -0.291
## X_educagAttended college or technical school        -0.112207   0.150956  -0.743
## X_educagGraduated from college or technical school -0.175003   0.156876  -1.116
## Classmiddle                                          0.263860   0.077345   3.411
## Classupper-middle                                    0.233459   0.086918   2.686
## cellphoneYes                                         0.072805   0.072245   1.008
## internetNo                                           0.607087   0.181786   3.340
## medcostNo                                            0.984964   0.113206   8.701
## smoke100No                                           0.477334   0.119349   3.999
## BMI                                                  0.021939   0.006822   3.216
## GENDERM:smoke100No                                  -0.178912   0.142890  -1.252
##                                                     Pr(>|z|)
## (Intercept)                                          < 2e-16 ***
## GENDERM                                             4.57e-10 ***
## X_educagGraduated high school                       0.770737
## X_educagAttended college or technical school        0.457293
## X_educagGraduated from college or technical school 0.264617
## Classmiddle                                         0.000646 ***
## Classupper-middle                                   0.007232 **
## cellphoneYes                                        0.313572
## internetNo                                          0.000839 ***
## medcostNo                                            < 2e-16 ***
## smoke100No                                          6.35e-05 ***
## BMI                                                 0.001300 **
## GENDERM:smoke100No                                  0.210534
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 0.8265
```

```
## Number of iterations in BFGS optimization: 1
## Log-likelihood: -1.705e+04 on 25 Df
```

Not to our surprise, `X_educag` is among the insignificant predictors whereas `GENDER` is extraordinarily significant. Dropping the interaction between `GENDER` and `children` since it is associated with the largest insignificant p-value, we then pass on the rest of candidates to the next step.

```
m2 <- zeroinfl(menthlth ~ GENDER + children + BMI + sleptim1 + maxdrnks +
                          DIET + exeroft1 + GENDER:sleptim1 + BMI:maxdrnks
                        | GENDER + X_educag + Class + cellphone + internet +
                          medcost + smoke100 + BMI + GENDER:smoke100,
               data = TRAIN, dist = "negbin", EM = TRUE)
summary(m2)
```

```
##
## Call:
## zeroinfl(formula = menthlth ~ GENDER + children + BMI + sleptim1 + maxdrnks +
##     DIET + exeroft1 + GENDER:sleptim1 + BMI:maxdrnks | GENDER + X_educag +
##     Class + cellphone + internet + medcost + smoke100 + BMI + GENDER:smoke100,
##     data = TRAIN, dist = "negbin", EM = TRUE)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -0.8010 -0.5521 -0.4692  0.1326  9.5611
##
## Count model coefficients (negbin with log link):
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        1.9167364  0.1825829  10.498  < 2e-16 ***
## GENDERM           -0.0324933  0.1634368  -0.199 0.842409
## children           0.0619331  0.0216565   2.860 0.004239 **
## BMI                0.0159883  0.0050796   3.148 0.001646 **
## sleptim1          -0.0841646  0.0159398  -5.280 1.29e-07 ***
## maxdrnks           0.0674266  0.0164495   4.099 4.15e-05 ***
## DIET              -0.0173619  0.0151891  -1.143 0.253019
## exeroft1           0.0014303  0.0003682   3.885 0.000102 ***
## GENDERM:sleptim1 -0.0353147  0.0228919  -1.543 0.122910
## BMI:maxdrnks      -0.0019115  0.0005806  -3.292 0.000994 ***
## Log(theta)        -0.1906746  0.0449806  -4.239 2.24e-05 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                                                  Estimate Std. Error z value
## (Intercept)                                     -2.692265   0.285558  -9.428
## GENDERM                                          0.791099   0.126820   6.238
## X_educagGraduated high school                   -0.044676   0.153012  -0.292
## X_educagAttended college or technical school    -0.112307   0.150965  -0.744
## X_educagGraduated from college or technical school -0.175068   0.156887  -1.116
```

```
## Classmiddle                                            0.263962   0.077351   3.413
## Classupper-middle                                      0.233587   0.086921   2.687
## cellphoneYes                                           0.072869   0.072250   1.009
## internetNo                                             0.607022   0.181805   3.339
## medcostNo                                              0.985057   0.113223   8.700
## smoke100No                                             0.477755   0.119342   4.003
## BMI                                                    0.021949   0.006822   3.218
## GENDERM:smoke100No                                    -0.179422   0.142864  -1.256
##                                                       Pr(>|z|)
## (Intercept)                                            < 2e-16 ***
## GENDERM                                               4.43e-10 ***
## X_educagGraduated high school                         0.770305
## X_educagAttended college or technical school          0.456921
## X_educagGraduated from college or technical school    0.264470
## Classmiddle                                           0.000644 ***
## Classupper-middle                                     0.007202 **
## cellphoneYes                                          0.313187
## internetNo                                            0.000841 ***
## medcostNo                                              < 2e-16 ***
## smoke100No                                            6.25e-05 ***
## BMI                                                   0.001292 **
## GENDERM:smoke100No                                    0.209153
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 0.8264
## Number of iterations in BFGS optimization: 1
## Log-likelihood: -1.705e+04 on 24 Df
```

By similar argument, we discard `GENDER` in the negative binomial regression and proceed. Note that the interaction between `GENDER` and `sleptim1` would be pointless without `GENDER`, so `GENDER:sleptim1` is also deleted.

```
m3 <- zeroinfl(menthlth ~ children + BMI + sleptim1 + maxdrnks + DIET +
                          exeroft1 + BMI:maxdrnks
                        | GENDER + X_educag + Class + cellphone + internet +
                          medcost + smoke100 + BMI + GENDER:smoke100,
               data = TRAIN, dist = "negbin", EM = TRUE)
summary(m3)
```

```
##
## Call:
## zeroinfl(formula = menthlth ~ children + BMI + sleptim1 + maxdrnks +
##     DIET + exeroft1 + BMI:maxdrnks | GENDER + X_educag + Class + cellphone +
##     internet + medcost + smoke100 + BMI + GENDER:smoke100, data = TRAIN,
##     dist = "negbin", EM = TRUE)
```

```
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -0.7951 -0.5450 -0.4593  0.1301  7.0402
##
## Count model coefficients (negbin with log link):
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.9892093  0.1659313  11.988  < 2e-16 ***
## children      0.0686299  0.0218555   3.140 0.001689 **
## BMI           0.0139650  0.0050954   2.741 0.006131 **
## sleptim1     -0.0985713  0.0115435  -8.539  < 2e-16 ***
## maxdrnks      0.0550472  0.0164679   3.343 0.000830 ***
## DIET         -0.0114259  0.0153103  -0.746 0.455495
## exeroft1      0.0013203  0.0003718   3.552 0.000383 ***
## BMI:maxdrnks -0.0016838  0.0005828  -2.889 0.003863 **
## Log(theta)   -0.2214901  0.0453801  -4.881 1.06e-06 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                                                    Estimate Std. Error z value
## (Intercept)                                       -2.825145   0.295297  -9.567
## GENDERM                                            0.918551   0.133703   6.870
## X_educagGraduated high school                     -0.057179   0.153933  -0.371
## X_educagAttended college or technical school      -0.126494   0.151984  -0.832
## X_educagGraduated from college or technical school -0.193654   0.158046  -1.225
## Classmiddle                                        0.268408   0.078418   3.423
## Classupper-middle                                  0.237942   0.087890   2.707
## cellphoneYes                                       0.074288   0.073006   1.018
## internetNo                                         0.605529   0.184157   3.288
## medcostNo                                          1.019653   0.117057   8.711
## smoke100No                                         0.508039   0.127702   3.978
## BMI                                                0.022428   0.006917   3.243
## GENDERM:smoke100No                                -0.205788   0.149265  -1.379
##                                                   Pr(>|z|)
## (Intercept)                                        < 2e-16 ***
## GENDERM                                           6.42e-12 ***
## X_educagGraduated high school                      0.71030
## X_educagAttended college or technical school       0.40525
## X_educagGraduated from college or technical school 0.22046
## Classmiddle                                        0.00062 ***
## Classupper-middle                                  0.00678 **
## cellphoneYes                                       0.30889
## internetNo                                         0.00101 **
## medcostNo                                          < 2e-16 ***
## smoke100No                                        6.94e-05 ***
## BMI                                                0.00118 **
## GENDERM:smoke100No                                 0.16799
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 0.8013
## Number of iterations in BFGS optimization: 1
## Log-likelihood: -1.708e+04 on 22 Df
```

DIET ought to be excluded at this step. Even though one level of X_educag gives larger p-value, we want to be safe and only look at the smallest p-value associated with a factor.

```
m4 <- zeroinfl(menthlth ~ children + BMI + sleptim1 + maxdrnks + exeroft1 +
                          BMI:maxdrnks
                        | GENDER + X_educag + Class + cellphone + internet +
                          medcost + smoke100 + BMI + GENDER:smoke100,
               data = TRAIN, dist = "negbin", EM = TRUE)
summary(m4)
```

```
##
## Call:
## zeroinfl(formula = menthlth ~ children + BMI + sleptim1 + maxdrnks +
##     exeroft1 + BMI:maxdrnks | GENDER + X_educag + Class + cellphone +
##     internet + medcost + smoke100 + BMI + GENDER:smoke100, data = TRAIN,
##     dist = "negbin", EM = TRUE)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -0.7950 -0.5451 -0.4595  0.1318  6.9920
##
## Count model coefficients (negbin with log link):
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.9781192  0.1651771  11.976  < 2e-16 ***
## children      0.0683209  0.0218437   3.128 0.001762 **
## BMI           0.0140685  0.0050907   2.764 0.005717 **
## sleptim1     -0.0983472  0.0115332  -8.527  < 2e-16 ***
## maxdrnks      0.0549448  0.0164659   3.337 0.000847 ***
## exeroft1      0.0013791  0.0003633   3.796 0.000147 ***
## BMI:maxdrnks -0.0016794  0.0005827  -2.882 0.003951 **
## Log(theta)   -0.2202226  0.0452911  -4.862 1.16e-06 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                                                    Estimate Std. Error z value
## (Intercept)                                        -2.82350    0.29493  -9.573
## GENDERM                                             0.91759    0.13352   6.872
## X_educagGraduated high school                      -0.05654    0.15386  -0.367
## X_educagAttended college or technical school       -0.12574    0.15191  -0.828
## X_educagGraduated from college or technical school -0.19306    0.15797  -1.222
## Classmiddle                                         0.26863    0.07836   3.428
```

```
## Classupper-middle                                          0.23818    0.08782   2.712
## cellphoneYes                                               0.07394    0.07294   1.014
## internetNo                                                 0.60567    0.18405   3.291
## medcostNo                                                  1.01910    0.11689   8.719
## smoke100No                                                 0.50870    0.12753   3.989
## BMI                                                        0.02243    0.00691   3.246
## GENDERM:smoke100No                                        -0.20655    0.14910  -1.385
##                                                           Pr(>|z|)
## (Intercept)                                                < 2e-16 ***
## GENDERM                                                   6.32e-12 ***
## X_educagGraduated high school                             0.713291
## X_educagAttended college or technical school              0.407814
## X_educagGraduated from college or technical school        0.221648
## Classmiddle                                               0.000607 ***
## Classupper-middle                                         0.006686 **
## cellphoneYes                                              0.310691
## internetNo                                                0.000999 ***
## medcostNo                                                  < 2e-16 ***
## smoke100No                                                6.64e-05 ***
## BMI                                                       0.001171 **
## GENDERM:smoke100No                                        0.165943
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 0.8023
## Number of iterations in BFGS optimization: 1
## Log-likelihood: -1.708e+04 on 21 Df
```

All p-values associated with the negative binomial regression component have become sufficiently small. Now we eliminate `cellphone` for the logistic regression and continue the backward selection.

```
m5 <- zeroinfl(menthlth ~ children + BMI + sleptim1 + maxdrnks + exeroft1 +
                          BMI:maxdrnks
                        | GENDER + X_educag + Class + internet + medcost +
                          smoke100 + BMI + GENDER:smoke100,
               data = TRAIN, dist = "negbin", EM = TRUE)
summary(m5)
```

```
##
## Call:
## zeroinfl(formula = menthlth ~ children + BMI + sleptim1 + maxdrnks +
##     exeroft1 + BMI:maxdrnks | GENDER + X_educag + Class + internet +
##     medcost + smoke100 + BMI + GENDER:smoke100, data = TRAIN, dist = "negbin",
##     EM = TRUE)
##
## Pearson residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -0.7956 -0.5455 -0.4599  0.1317  7.0660
##
## Count model coefficients (negbin with log link):
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.9778637  0.1651520  11.976  < 2e-16 ***
## children      0.0689095  0.0218400   3.155 0.001604 **
## BMI           0.0140486  0.0050898   2.760 0.005777 **
## sleptim1     -0.0982219  0.0115307  -8.518  < 2e-16 ***
## maxdrnks      0.0548693  0.0164641   3.333 0.000860 ***
## exeroft1      0.0013792  0.0003632   3.797 0.000147 ***
## BMI:maxdrnks -0.0016774  0.0005827  -2.879 0.003993 **
## Log(theta)   -0.2195366  0.0452452  -4.852 1.22e-06 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                                                   Estimate Std. Error z value
## (Intercept)                                       -2.756598   0.286142  -9.634
## GENDERM                                            0.918357   0.133295   6.890
## X_educagGraduated high school                     -0.051904   0.153557  -0.338
## X_educagAttended college or technical school      -0.118916   0.151525  -0.785
## X_educagGraduated from college or technical school -0.184150   0.157506  -1.169
## Classmiddle                                        0.263586   0.078120   3.374
## Classupper-middle                                  0.226512   0.087009   2.603
## internetNo                                         0.608491   0.183927   3.308
## medcostNo                                          1.015081   0.116653   8.702
## smoke100No                                         0.508113   0.127314   3.991
## BMI                                                0.022195   0.006899   3.217
## GENDERM:smoke100No                                -0.206577   0.148901  -1.387
##                                                   Pr(>|z|)
## (Intercept)                                        < 2e-16 ***
## GENDERM                                           5.59e-12 ***
## X_educagGraduated high school                     0.735355
## X_educagAttended college or technical school      0.432574
## X_educagGraduated from college or technical school 0.242338
## Classmiddle                                       0.000741 ***
## Classupper-middle                                 0.009233 **
## internetNo                                        0.000939 ***
## medcostNo                                          < 2e-16 ***
## smoke100No                                        6.58e-05 ***
## BMI                                               0.001294 **
## GENDERM:smoke100No                                0.165336
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 0.8029
## Number of iterations in BFGS optimization: 1
## Log-likelihood: -1.708e+04 on 20 Df
```

To further trim the model, `X_educag` is crossed out as the smallest p-value associated with it turns out to be greater than any other insignificant p-values.

```
m6 <- zeroinfl(menthlth ~ children + BMI + sleptim1 + maxdrnks + exeroft1 +
                          BMI:maxdrnks
                        | GENDER + Class + internet + medcost + smoke100 +
                          BMI + GENDER:smoke100,
               data = TRAIN, dist = "negbin", EM = TRUE)
summary(m6)
```

```
##
## Call:
## zeroinfl(formula = menthlth ~ children + BMI + sleptim1 + maxdrnks +
##     exeroft1 + BMI:maxdrnks | GENDER + Class + internet + medcost + smoke100 +
##     BMI + GENDER:smoke100, data = TRAIN, dist = "negbin", EM = TRUE)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -0.7978 -0.5440 -0.4611  0.1347  6.9653
##
## Count model coefficients (negbin with log link):
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.9813503  0.1650499  12.005  < 2e-16 ***
## children      0.0670160  0.0217767   3.077 0.002088 **
## BMI           0.0140210  0.0050871   2.756 0.005848 **
## sleptim1     -0.0983639  0.0115264  -8.534  < 2e-16 ***
## maxdrnks      0.0546632  0.0164520   3.323 0.000892 ***
## exeroft1      0.0013765  0.0003631   3.791 0.000150 ***
## BMI:maxdrnks -0.0016700  0.0005822  -2.868 0.004125 **
## Log(theta)   -0.2182204  0.0451649  -4.832 1.35e-06 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.846841   0.258768 -11.002  < 2e-16 ***
## GENDERM             0.927155   0.132919   6.975 3.05e-12 ***
## Classmiddle         0.264555   0.078045   3.390 0.000700 ***
## Classupper-middle   0.216865   0.086708   2.501 0.012381 *
## internetNo          0.641426   0.182599   3.513 0.000443 ***
## medcostNo           0.997406   0.115326   8.649  < 2e-16 ***
## smoke100No          0.484603   0.126259   3.838 0.000124 ***
## BMI                 0.022470   0.006895   3.259 0.001117 **
## GENDERM:smoke100No -0.206193   0.148671  -1.387 0.165470
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 0.8039
```

```
## Number of iterations in BFGS optimization: 1
## Log-likelihood: -1.708e+04 on 17 Df
```

And finally, we drop the interaction between `GENDER` and `smoke100`, and obtain the model in which all explanatory variables are significant.

```
m7 <- zeroinfl(menthlth ~ children + BMI + sleptim1 + maxdrnks + exeroft1 +
                          BMI:maxdrnks
                        | GENDER + Class + internet + medcost + smoke100 + BMI,
               data = TRAIN, dist = "negbin", EM = TRUE)
summary(m7)
```

```
##
## Call:
## zeroinfl(formula = menthlth ~ children + BMI + sleptim1 + maxdrnks +
##     exeroft1 + BMI:maxdrnks | GENDER + Class + internet + medcost + smoke100 +
##     BMI, data = TRAIN, dist = "negbin", EM = TRUE)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -0.7924 -0.5483 -0.4603  0.1327  6.9973
##
## Count model coefficients (negbin with log link):
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.9811733  0.1649541  12.010  < 2e-16 ***
## children      0.0677723  0.0217617   3.114 0.001844 **
## BMI           0.0141044  0.0050850   2.774 0.005542 **
## sleptim1     -0.0984756  0.0115169  -8.551  < 2e-16 ***
## maxdrnks      0.0546012  0.0164476   3.320 0.000901 ***
## exeroft1      0.0013772  0.0003628   3.796 0.000147 ***
## BMI:maxdrnks -0.0016709  0.0005821  -2.870 0.004101 **
## Log(theta)   -0.2150407  0.0449000  -4.789 1.67e-06 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.72053    0.23730 -11.464  < 2e-16 ***
## GENDERM             0.76932    0.06420  11.984  < 2e-16 ***
## Classmiddle         0.26847    0.07781   3.450 0.000560 ***
## Classupper-middle   0.22081    0.08656   2.551 0.010742 *
## internetNo          0.64357    0.18152   3.546 0.000392 ***
## medcostNo           0.99380    0.11437   8.689  < 2e-16 ***
## smoke100No          0.33983    0.06698   5.074  3.9e-07 ***
## BMI                 0.02217    0.00687   3.227 0.001252 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Theta = 0.8065
## Number of iterations in BFGS optimization: 1
## Log-likelihood: -1.708e+04 on 16 Df
```

The final **ZINB** model regresses upon the following predictors for the binary process and the count process respectively.

Binary Process	
GENDER	The biological gender
Class	The socio-economic class determined by annual family income
internet	Whether one accesses the internet during the past month
medcost	Whether one has unpaid medical bills
smoke100	Whether one smokes during the past 100 days
BMI	The body mass index that reflects the relationship between height and weight

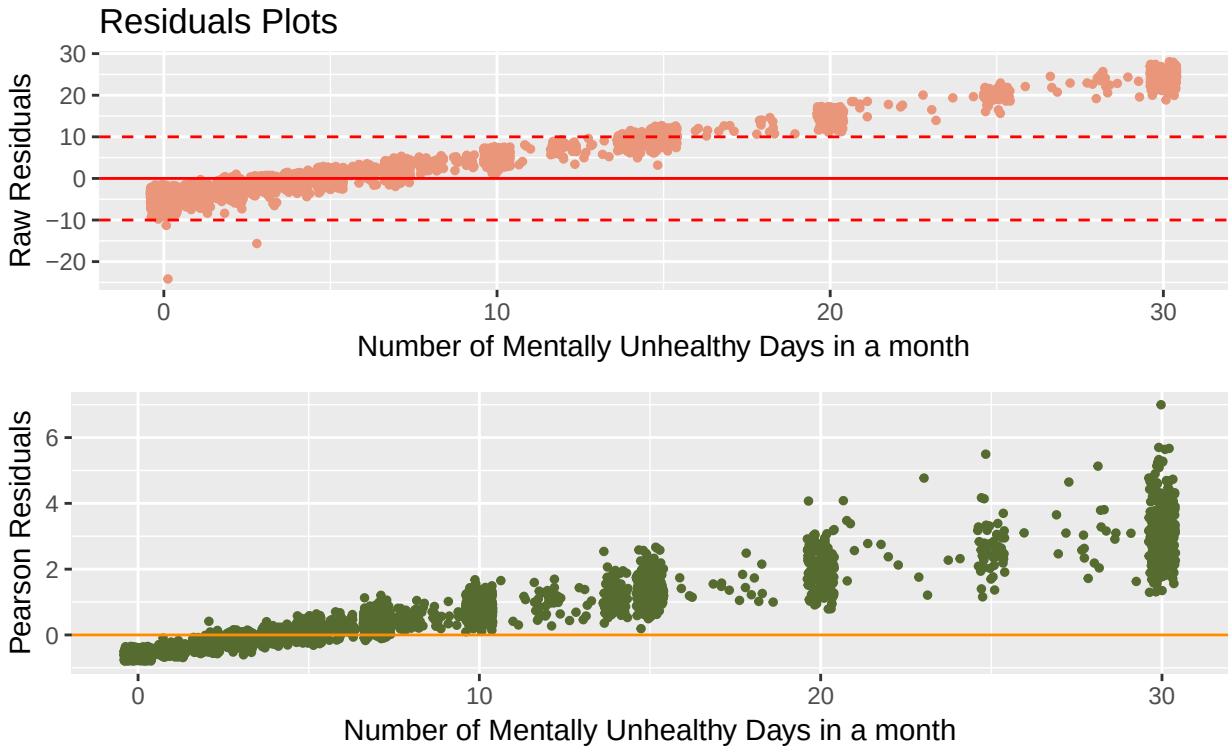
Count Process	
BMI	The body mass index that reflects the relationship between height and weight
children	The number of children one has
sleptim1	The average number of hours one sleeps per day during the past month
maxdrnks	The max number of glasses of alcohol per day during the past month
exeroft1	The average intensity of exercises during the past month
BMI:maxdrnks	The interaction between BMI and maxdrnks

Residuals & Singular Cases

In spite of the increasing underestimation for more severe mental problems, the residuals are mostly concentrated within an acceptable range; that is to say, predictions are only accurate for mild to no conditions.

```
TRAIN <- TRAIN %>%
  mutate(raw_resid = m7$residuals) %>%
  mutate(pearson = resid(m7,type="pearson"))
# plot raw residuals against menthlth
G1 <- ggplot(TRAIN, aes(x=menthlth,y=raw_resid)) +
  geom_jitter(color="darksalmon",size=1) +
  geom_hline(yintercept=c(0,-10,10),color="red",linetype=c(1,2,2)) +
  xlab("Number of Mentally Unhealthy Days in a month") +
  ylab("Raw Residuals") +
  ggtitle("Residuals Plots")
# plot Pearson residuals against menthlth
G2 <- ggplot(TRAIN, aes(x=menthlth,y=pearson)) +
  geom_jitter(color="darkolivegreen", size=1) +
  geom_hline(yintercept=0, color="darkorange") +
```

```
  xlab("Number of Mentally Unhealthy Days in a month") +
  ylab("Pearson Residuals")
grid.arrange(grobs=list(G1,G2),layout_matrix=matrix(1:2,nrow=2,ncol=1))
```



The singular point at the bottom-left of the raw residuals plot catches our attention — a respondent who self-reports to be completely fine is predicted to be mentally compromised for over 20 days in a month. By looking into his case, we learn that he is a lower-class smoker who misses the higher education, uses cellphone and internet, drinks a lot of alcohol, and bears unpaid medical bills, all explaining the curious prediction. Yet the sleeping time contributes exceptionally. Could anybody survive from averagely sleeping two hours per day for a month? Is this record an error? Or more importantly, given all these detrimental factors, is he hiding his mental problems in the interview?

However, the respondent also exercises astonishingly intensely (a hard laborer, perhaps) and maintains perfect BMI. Incorporating the fact that some people are just genetically resistant to psychological damages, we conclude that this case is possible, albeit improbable, and should not be abandoned.

```
# print out the row of the singular case
D <- TRAIN %>%
  mutate(Fitted=m7$fitted.values) %>%
  filter(raw_resid <= -20)
```

menthlth	DISTRICT	X_race	GENDER
0	Midwest	American Indian or Alaskan Native only, Non-Hispanic	M

X_educag	Class	cellphone	internet	medcost	BMI	children
Graduated high school	lower	Yes	Yes	Yes	18.46876	1

sleptim1	smoke100	maxdrnks	fruit1	vegetab1	exeroft1	DIET	Fitted
2	Yes	40	101	301	230	-0.0643763	24.16781

Cross Validation

To evaluate how well the final model fits the training data and generalizes to new data, we repeat the fitting $k = 5$ times, each time treating a different subset as the validation data and merging the other subsets as the training data. By doing so, all observations are used for both training and validation, and each observation is used for validation exactly once. We then assess the model performance by the square root of the mean squared error, both for fitting and for predicting.

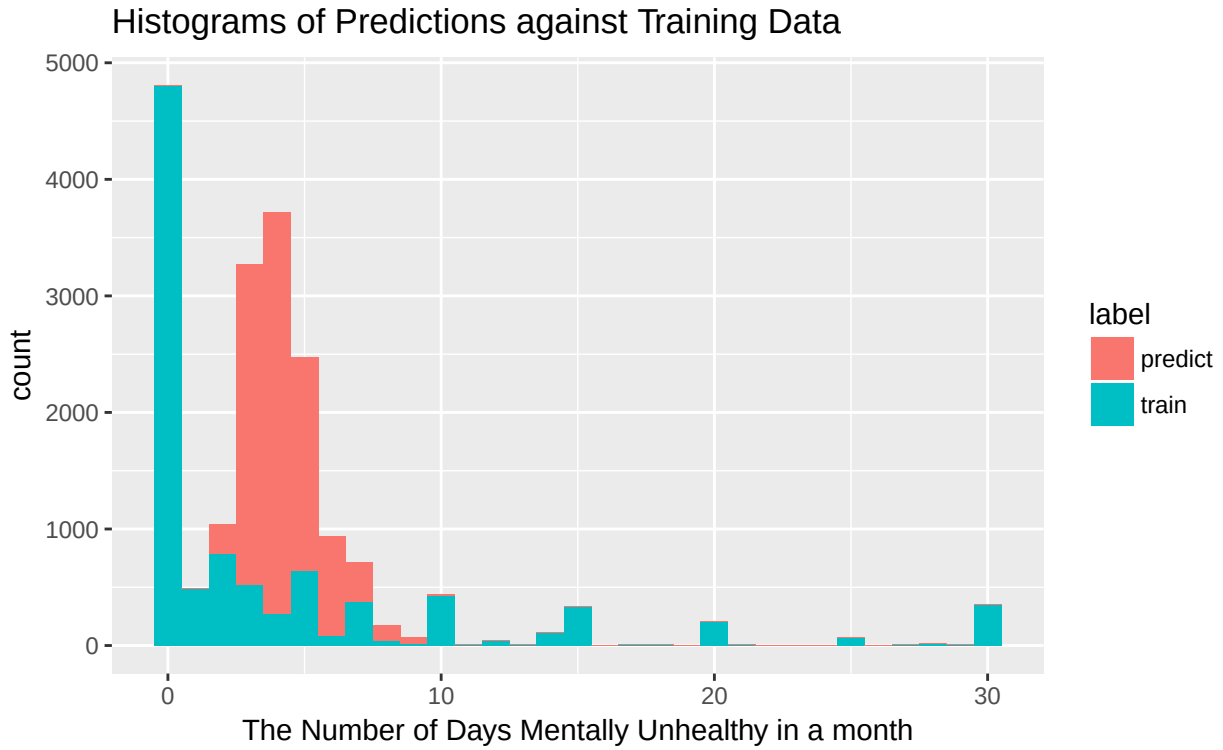
```
# repeat training and validation k times
# each time using the ith subset as validation data
# record sqrt(MSE) in every run
error_fitted <- c()
error_pred <- c()
days_pred <- list()
model <- list()
for (i in 1:k)
{
  TRAIN <- DATA %>% slice(setdiff(1:nrow(DATA),ID[[i]]))
  TEST <- DATA[ID[[i]],]
  model[[i]] <- zeroinfl(menthlth ~ children + BMI + sleptim1 + maxdrnks +
                                    exeroft1 + BMI:maxdrnks
                                  | GENDER + Class + internet + medcost +
                                    smoke100 + BMI,
                         data = TRAIN, dist = "negbin", EM = TRUE)
  error_fitted[[i]] <- sum((model[[i]]$fitted.values-TRAIN$menthlth)^2)
  error_fitted[i] <- sqrt(error_fitted[[i]]/nrow(TRAIN))
  days_pred[[i]] <- predict(model[[i]], TEST)
  error_pred[i] <- sqrt(sum((days_pred[[i]]-TEST$menthlth)^2)/nrow(TRAIN))
}

# print out the errors
ERROR <- data.frame(error_fitted, error_pred)
colnames(ERROR) <- c("Error of Fitting","Error of Predicting")
kable(ERROR, digits=4)
```

Error of Fitting	Error of Predicting
6.9715	3.5418
6.9723	3.5280
6.9672	3.5374
6.9915	3.5068
7.0519	3.3745

Judged by the insubstantial fluctuation in each column above, the predictions made by our final model are rather reliable. Nevertheless, it calls for caution that the validation errors are strikingly smaller than the training errors in every run. It could be simply due to sample sizes; if the model cannot account for certain patterns among the data, adding more samples constantly results in greater unexplained variances.

In fact, as shown in the figure below, the model fails to capture either the fat tail or the excess of zeros in the observations.

```
df1 <- data.frame(DAYS=DATA$menthlth)
df2 <- data.frame(DAYS=predict(m7,DATA))
df1$label <- "train"
df2$label <- "predict"
ggplot(rbind(df1,df2), aes(DAYS,fill=label)) +
  geom_histogram(alpha=1, binwidth=1) +
  xlab("The Number of Days Mentally Unhealthy in a month") +
  ggtitle("Histograms of Predictions against Training Data")
```



The most powerful risk factors (e.g. genes and family history of diseases) have not been gathered by the GSS. Not surprisingly, there is a notable portion of variances that cannot be explained away no matter how good the model is. Given the poor availability of genetic and medical data in general, we propose to assist in identifying vulnerable young Americans with the GSS data to any feasible extent.

In addition, automatic detection always comes down to a judgment — whether one needs to see a doctor or not. Since National Institute of Mental Health advise people who feel mentally compromised for more than five days within a month to seek professional help,[5] we can also employ this threshold to make binary decision about the mental states of respondents.

```
DATA <- DATA %>%
  mutate(DIAGNOSIS=factor(ifelse(menthlth<4,"fine","ill"))) %>%
  mutate(predict = predict(m7,DATA)) %>%
  mutate(PREDICT=factor(ifelse(predict<4, "negative","positive")))
# construct a contingency table
TAB <- table(DATA$PREDICT, DATA$DIAGNOSIS)
kable(TAB)
```

	fine	ill
negative	3770	1058
positive	2840	2005

```
# compute sensitivity
cat("sensitivity:", round(TAB[2,2]/sum(TAB[,2]),digits=4), "\n")
```

```
## sensitivity: 0.6546
```

```
# compute specificity
cat("specificity:", round(TAB[1,1]/sum(TAB[,1]),digits=4), "\n")
```

```
## specificity: 0.5703
```

Both the sensitivity (probability of true positiveness) and the specificity (probability of true negativeness) turn out to be acceptable. In other words, approximately 65% of the respondents who are indeed suffering from mental illness would be correctly detected by our model and advised to see doctors, without harassing too many healthy people and burdening medical workers overwhelmingly.

[5] https://www.nimh.nih.gov/health/education-awareness/index.shtml

Interpretation

We would like to highlight that the **ZINB model** consists of two components, the **logistic regression** for the binary process and the **negative binomial regression** for the count process. And each component involves a different set of explanatory variables — say, X and Z — respectively.

$$E(Y_j) = \pi_j \cdot E(Y_j|W_j = 1) = \frac{\pi_j p_j r_j}{1 - p_j}$$

Logistic Regression

In the binary process, the logit of $\pi_j = P(W_j = 1)$ is linearly modelled by

$$\log(\frac{\pi_j}{1-\pi_j}) = \mathbf{X}_{(j)} \cdot \vec{\beta} + \epsilon_j$$

$$\hat{\pi} = \frac{\exp(\mathbf{X}\hat{\beta})}{1 + \exp(\mathbf{X}\hat{\beta})}$$

Therefore, the **intercept** $\beta_0 \approx -2.7205$ can be interpreted as that when BMI equals to 0 (which is practically impossible) and all factors take the reference level, the probability of an American resident who is aged 18 to 24 being exposed to intrinsic or extrinsic triggering circumstances (i.e. $W = 1$) is $\pi_0 = \exp(\beta_0)/(1 + \exp(\beta_0)) \approx 0.0618$.

The interpretations of the **slopes** concern the **odds** (i.e. $\pi/(1-\pi)$) of such exposure. For instance, everything else held constant, the odds is expected to be multiplied by $\exp(\beta_7) \approx 1.0224$ for an additional unit of BMI on average. And the odds for a female, on average, is expected to be $\exp(\beta_1) \approx 2.1583$ times that for a male, everything else held constant.

Negative Binomial Regression

In the count process, the expected days of feeling mentally unhealthy is logarithmatically linked to the linear combination of $\mathbf{Z}_{(j)}$.[6]

$$\log(\mathbf{E}[Y_j|W_j = 1]) = \mathbf{Z}_{(j)} \cdot \vec{\gamma} + \delta_j$$

$\hat{\gamma}$ has to be interpreted in terms of conditional expectation $E[Y|W = 1]$. Given that an American resident aged 18 to 24 is exposed to triggering circumstances (i.e. $W = 1$), the **intercept** $\gamma_0 \approx 1.9812 \approx \log(7.25)$ expresses the log of the expected number of unhealthy days elicited in a month when other variables all take zeros (again, some cases are practically impossible).

Conditional on $W = 1$, the number of mentally disturbed days in a month is expected to be multiplied by $\exp(\gamma_1) \approx 1.0702$, on average, for one additional child that the young American has, holding other variables constant. The interpretation of the **slopes** is further complicated by the presence of **interaction**. Speciically, conditional on $W = 1$, we expect an additional unit of BMI

[6] https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Negative_Binomial_Regression.pdf

to multiply the number of unhealthy days by $\exp(\gamma_2 + \gamma_6 \cdot \Delta) \approx \exp(0.0141 - 0.0017 \cdot \Delta)$, with Δ being the change in the number of glasses of alcohol that the young American drinks per day at maximum, averagely speaking and everything else held constant. We may also interpret the coefficients pertaining to the max number of drinks in a similar manner.

Conclusion

This case study aims at predicting the number of mentally troubled days in a month, for American residents aged 18 to 24, from ten variables in the GSS data. Professional opinions of psychologists are taken into account and interaction among predictors considered. We opt for the zero-inflated negative binomial regression to address the overdispersion and the excess of zeros found in the response. Using backward selection, we obtain the final model in which every predictor is associated with significant p-values.

A few interesting findings present themselves in the results. For example, under triggering circumstances, an increase in either BMI or the maximum number of drinks per day is expected to boost the occurrence of symptoms when affecting alone; but such rise inclines to be less steep if they work interactively. Can alcohol relieve the anxiety of becoming fat somehow? Is using diverse ways to handle bad mood, eating high-calorie food and getting drunk, healthier than using either one exclusively? Moreover, dropping predictors itself could be revealing. It is widely presumed that the pressure of caring for kids dominantly falls upon mothers, yet the lack of interaction between genders and the number of children tends not to support this prevalent view.

Finally, it is practical to flag young Americans at high risk of mental problems based on the outputs generated by this model. Although more crucial information is not collected by the GSS, and thus the predictions are not pleasingly accurate, this model serves the goal of early intervention and offers a satisfying solution. All codes and data are accessible online.[7]

[7] https://github.com/PawinData/GLM