AMAT 565—Applied Statistics

How Socio-demographic factors influence health?

Instructor: Karin Reinhold

Group: Avengers

Vineetha Dulla, Harshavardhan Chittaluru

November 30,2023

University At Albany

Section 1: Introduction of Data Set and Purpose of Project

The dataset extracted from the General Social Survey (GSS) via the GSS data explorer encompasses 64,814 rows and 13 columns, comprising categorical variables. These variables include YEAR (ranging from 1972 to 2018), WRKSTAT (representing labor force status), MARITAL (capturing marital status), AGE (categorized into '89 OR OLDER', 'DK', 'NA'), EDUC (indicating the highest year of school completed), SEX (classifying respondents as 'Male' or 'Female'), RACE (divided into 'WHITE', 'BLACK', 'OTHER', 'IAP'), RINCOME (reflecting respondents' income levels, including 'LT $1000', '$1000 to $2999', and so forth up to '$25000 or more', alongside 'Refused' and 'DK'), REGION (categorized by geographic regions), NATDRUG (pertaining to dealing with drug addiction), HEALTH (representing the condition of health from 'EXCELLENT' to 'DK'), HAPPY (measuring general happiness from 'VERY HAPPY' to 'DK'), and SMOKE (indicating whether the respondent smokes, with options 'YES', 'NO', 'DK').

The primary objective of this project involves predicting health conditions and exploring correlations among these variables to discern underlying trends. By analyzing these categorical variables, the aim is to draw connections and insights regarding health conditions while examining the relationships between different factors such as demographics, socio-economic status, lifestyle choices, and reported health outcomes.

Section 2: Data Pre-processing

During the initial data analysis, we identified NA values in each feature and subsequently removed them. This action led to a reduction in the dataset size from 64,814 to 3,314 observations, indicating a depletion rate of 95%. The resultant dataset, named gss_data_wo_na, represents the initial dataset with NA values removed, ensuring a more complete dataset for further analysis.

YEAR	WRKSTAT	MARITAL	AGE	EDUC	SEX	RACE	RINCOME	REGION	NATDRUG
0	21	27	228	177	0	0	27026	0	29636
HEALTH	HAPPY	SMOKE							
17224	4760	48441							

Initially, our dataset contained variables of type <dbl +lbl>, representing double and label data types. To focus solely on the numerical aspect and work with categorical variables containing only double data, we created a modified dataset without the label component using the following code:

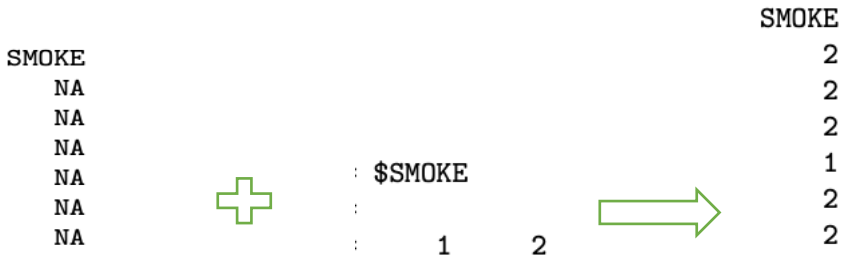
gss_data_ul <- data.frame(sapply(gss_data, haven::zap_labels))

```
WRKSTAT
<dbl+lbl>
1 [WORKING ~
```

transformed to

```
WRKSTAT
       1
       5
```

Considering the dataset's substantial NA values, dropping columns containing NA values could significantly impair our analysis. To address this, a strategy is proposed to replace the NA values within each feature. The resulting dataset, named gss_data_rvalues, will involve substituting NA values in each feature with unique categories specific to that particular feature. This approach allows us to retain valuable information across all variables while mitigating the impact of missing values, enabling a more comprehensive analysis of the dataset.



Utilizing the mice package's methodology, we've employed the "rf" method to impute missing values in our dataset. Configured with m = 5 and maxit = 5, this signifies that during each iteration, the algorithm applies random forest ("rf") to derive the best unique values for imputation. As a result, five distinct datasets are generated, capturing the most suitable imputed values after each iteration. To retrieve the imputed data, the complete() function is employed, extracting the finalized imputed dataset from the iterations.

```
tempdata <- mice(gss_data_ul,m=5,maxit=5,meth="rf",seed=500)
```

```
gss_data_rf_1 <- complete(tempdata,1)
```

After preprocessing steps finally, we got 7 different data sets. They are

gss_data_wo_na – Data set without NA values

gss_data_rf_1
gss_data_rf_2
gss_data_rf_3
gss_data_rf_4

} Data set from mice imputation method

gss_data_rf_5

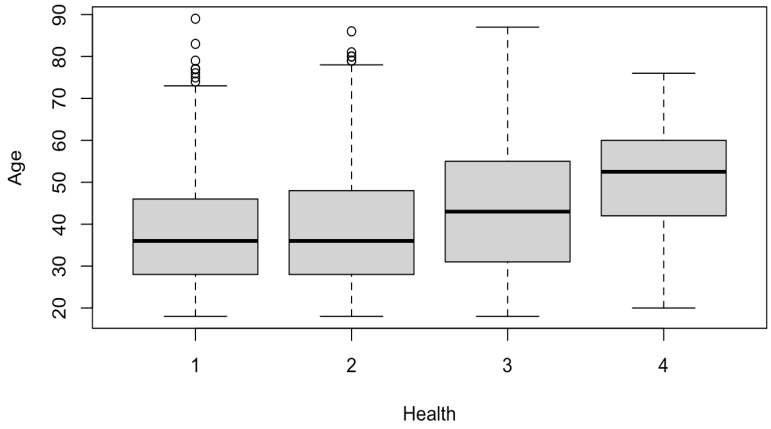
gss_data_rvalues – Data set with random values

Section 3: Exploratory Data Analysis

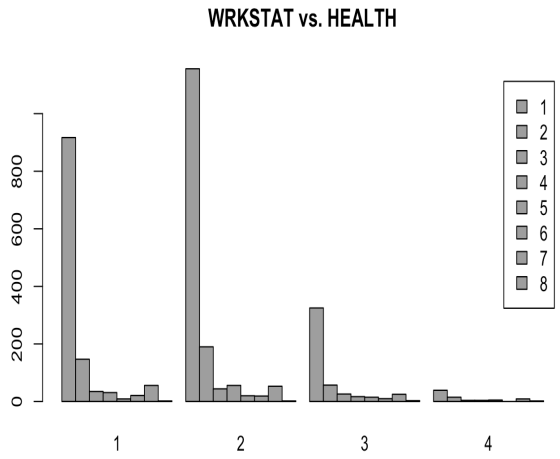
Exploratory data analysis (EDA) is an approach to data analysis to summarize their main characteristics, often with visual methods.

AGE vs HEALTH

In exploring the relationship between the independent variable (AGE) and the response variable (HEALTH) within the "gss_data_wo_na" dataset, boxplots were utilized as visualizations for numeric variables. The boxplots representing the "excellent" and "good" health categories displayed a positive skew, notably leaning towards the lower quartile. Conversely, the boxplot for the "fair" health category exhibited a relatively normal distribution, residing within the inter-quartile range. Finally, the boxplot associated with the "poor" health category showcased a negative skew, positioned towards the upper quartile of the data distribution.

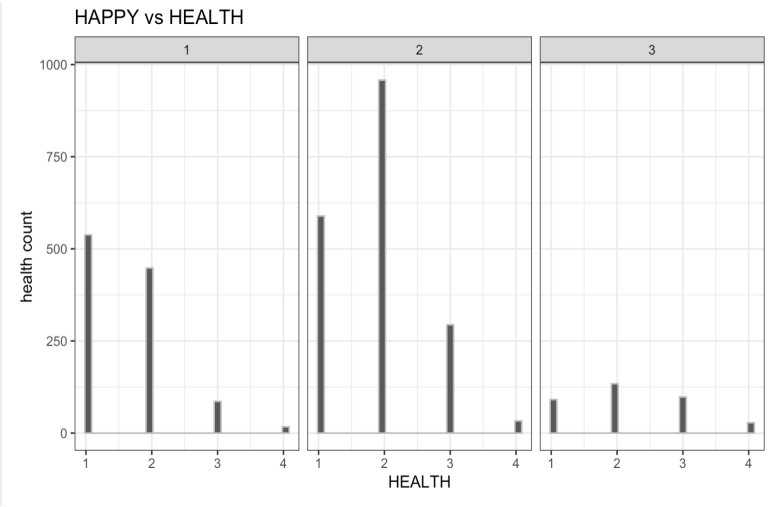


WRKSTAT vs HEALTH



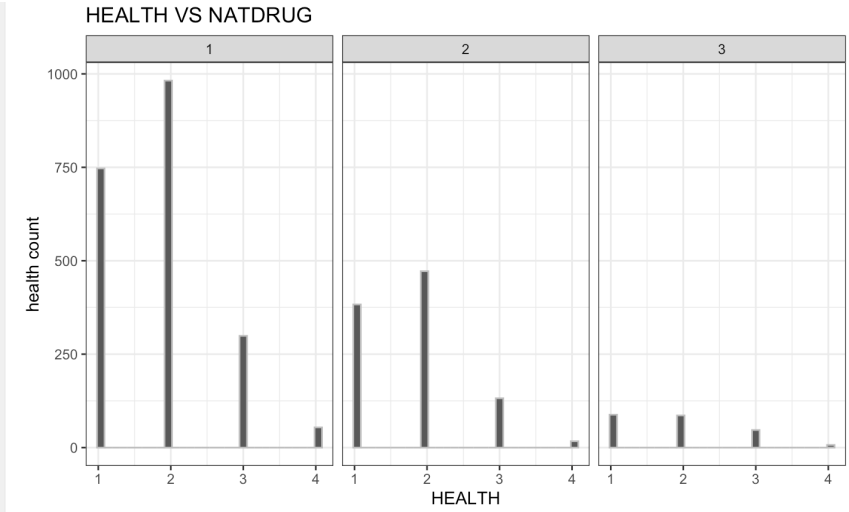
Utilizing ggplot as a visualization method to explore the relationship between "wrkstat" and "health" variables, a notable observation emerged: individuals categorized under "full time" or "part time" employment tend to exhibit an association with "excellent" or "good" health conditions. Conversely, for individuals outside these employment categories, there appears to be a trend indicating a correlation with "poor" health conditions. This suggests a potential relationship where individuals engaged in full-time or part-time work tend to report better health conditions compared to those in other employment categories.

HAPPY vs HEALTH



After utilizing ggplot as a visualization method to explore the relationship between "happy" and "health" variables, an intriguing observation surfaced. Individuals classified under the "very happy" or "pretty happy" categories often displayed a tendency towards reporting "excellent," "good," or "fair" health conditions. Conversely, for individuals categorized as "not too happy," there appears to be a consistent association with "poor" health conditions. This suggests a potential trend where higher levels of happiness correspond to better-reported health conditions, while lower levels of happiness correlate with a higher likelihood of reporting poorer health statuses.

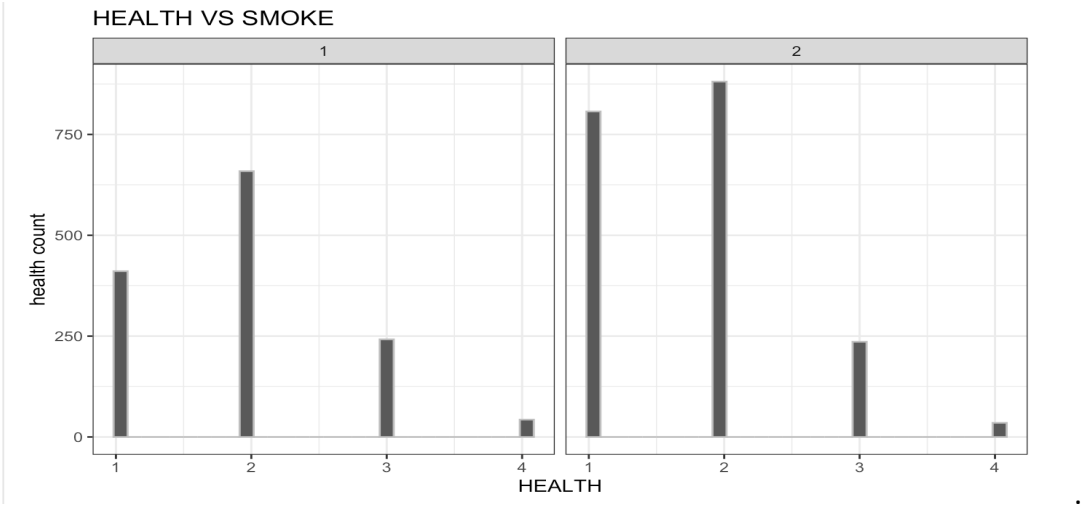
HEALTH vs NATDRUG



Using ggplot as a visualization method to explore the relationship between "health" and "natdrug" variables revealed an interesting trend. Individuals categorized under "too little" or "about right" in terms of drug usage tended to report "excellent," "good," or occasionally "fair" health conditions. Conversely, individuals classified under "too much" drug usage showed a distinct correlation with reporting "poor" health conditions. This suggests a potential association where lower levels of drug usage correspond to a higher likelihood of reporting better health statuses, while higher levels of drug usage correlate with a greater likelihood of reporting poorer health conditions.

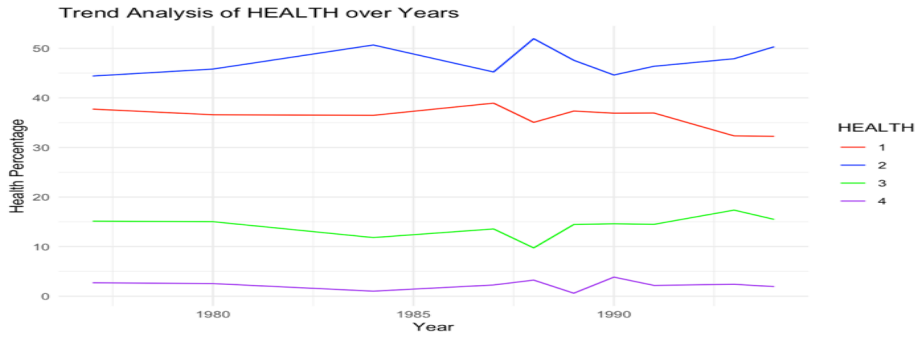
HEALTH vs SMOKE

Ggplot is one of the visualization method for categorical variables to show the relation between health vs smoke. Following the use of ggplot to visualize the relationship between "health" and "smoke" variables, a discernible trend emerged: individuals identified as smokers tended to exhibit health conditions that were notably less favorable, often leaning towards "not that good." Conversely, individuals categorized as non-smokers displayed a wider range of health conditions, spanning from "good" to "fair" and occasionally "excellent." This suggests a potential correlation between smoking status and the reported quality of health, indicating a propensity for poorer health conditions among smokers compared to non-smokers.



YEAR vs HEALTH

Trend analysis is used to show relation between year and health.



Through trend analysis depicting the relationship between "year" and "health," a notable pattern emerged. From 1972 to 1999, the second category denoted by the blue line, representing "good" health conditions, consistently exhibited the highest percentage compared to other health categories. This trend indicates a prolonged prevalence of relatively favorable health conditions categorized as "good" across the specified years within the dataset.

Section 4: Modelling:

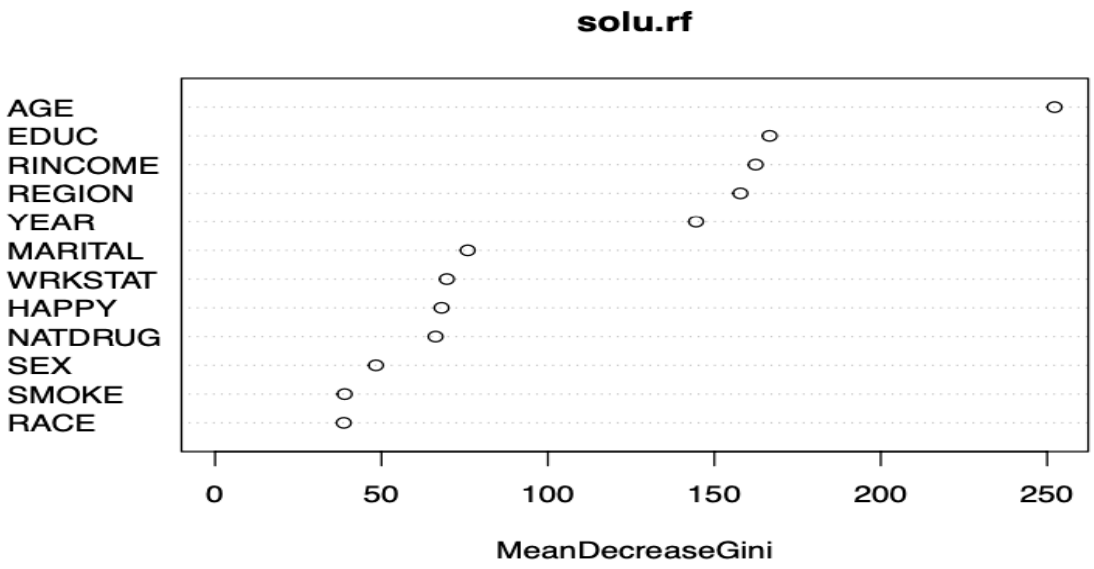
Random Forest:

Random Forest approach is a supervised learning algorithm. It builds the multiple decision trees which are known as forest and glue them together to urge a more accurate and stable prediction. Bagging, also known as bootstrap aggregation, is the ensemble learning method that is commonly used to reduce variance within a noisy dataset. Random forest is an extension of bagging that also randomly selects subsets of features used in each data sample.

To make a prediction at a point x

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_B(x)$$

Importance plot:



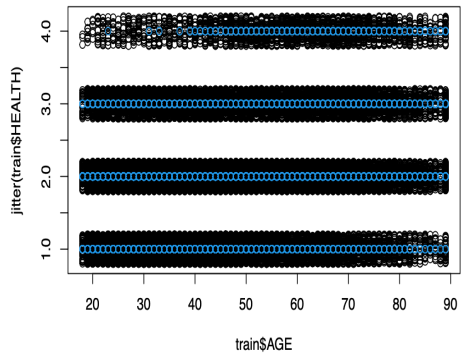
AGE is the important feature out of all the features. Which is why I compare the HEALTH variable with AGE for the plots associated with Random Forest.

Our Model:

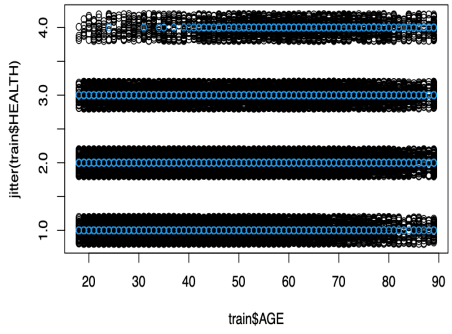
```
model = randomForest(as.factor(HEALTH)~.,data=train)
```

To this model, we train with different datasets that are generated above:

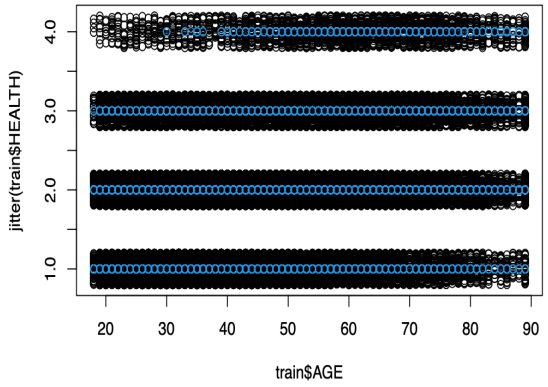
1st Mice Model



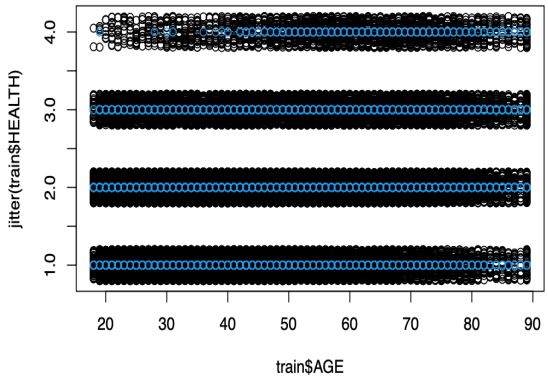
2nd Mice model



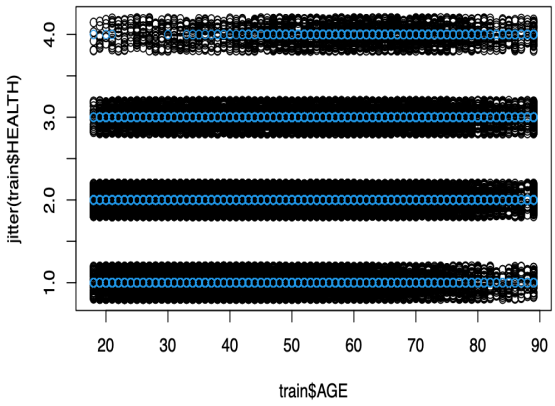
3rd Mice Model



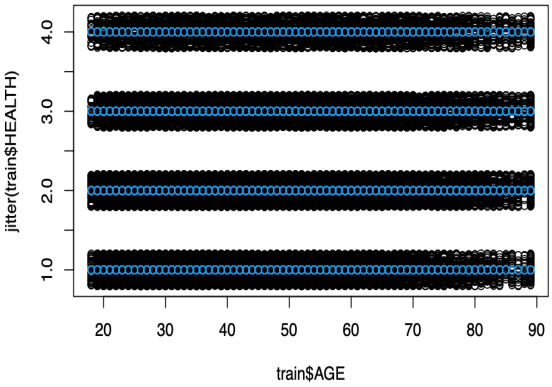
4th Mice Model



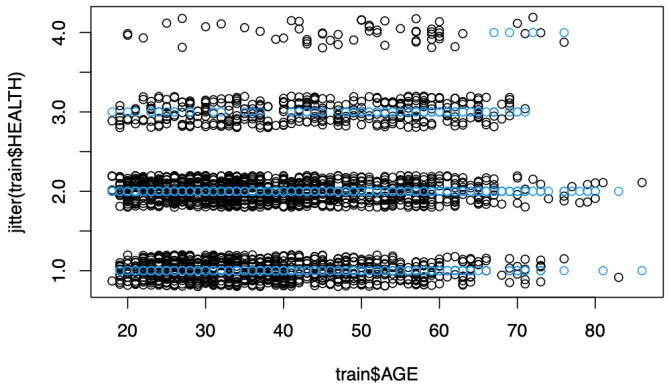
5th Mice Model



Model with Random values



Complete Dataset(Without NA values)



These graphs depict Black circles are original data points, whereas blue points are predicted points from model.

<u>Accuracies of Random Forest:</u>

act	obs 1	2	3	4
1	2259	3370	214	15
2	1776	6358	636	69
3	364	2436	780	128
4	59	537	379	147

Accuracy: 0.4887592

1st Mice model

act	obs 1	2	3	4
1	2230	3356	214	19
2	1738	6421	685	56
3	363	2420	803	124
4	51	534	377	136

Accuracy: 0.4911149

2nd Mice model

act	obs 1	2	3	4
1	2236	3357	212	20
2	1658	6475	671	69
3	348	2401	818	122
4	68	533	369	170

Accuracy: 0.4966969

3rd Mice Model

act	obs 1	2	3	4
1	2193	3325	220	10
2	1760	6510	673	82
3	325	2441	769	135
4	56	510	374	144

Accuracy: 0.4924464

4th Mice Model

act	obs 1	2	3	4
1	2216	3372	217	20
2	1724	6474	637	64
3	344	2446	778	129
4	49	538	362	157

Accuracy: 0.4929073

5th Mice model

act	obs 1	2	3	4
1	1842	3293	379	61
2	1540	5329	782	146
3	519	2333	1014	171
4	287	1111	438	282

Accuracy: 0.4336048

Model with Random values

act	obs 1	2	3	4
1	154	229	8	1
2	126	322	12	0
3	36	80	17	1
4	2	15	4	0

Accuracy: 0.489573

Complete Dataset (Without NA Values)

Based on the accuracy metrics 3rd Mice model gives the best results. But the 3rd mice model does lot of combinations in imputing the data. So, even though, the complete dataset has less accuracy than others, I choose this as the best model because these values are from the original dataset.

Poisson Regression:

In statistics, Poisson regression is a generalized linear model form of regression analysis used to model count data and contingency tables. Poisson regression assumes the response variable Y has a Poisson distribution.

$$P(Y = y) = e^{-\lambda}\frac{\lambda^y}{y!}, \qquad y = 0, 1, 2, \ldots$$

<u>Poisson distribution:</u>

Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event. We assume Mean = Variance = lambda

$$P(Y_i = y_i) = e^{-e^{X_i\beta}} \frac{(e^{X_i\beta})^{y_i}}{y_i!} = e^{-e^{X_i\beta}} \frac{e^{y_i\ X_i\beta}}{y_i!} \qquad \text{where } \lambda_i = e^{X_i\beta}.$$

We estimate p value by Chisquare distribution in Poisson regression, where in linear regression, we use t-distribution. We get Null Deviance and Residual deviance from the poisson model. Larger the deviance represents the model does not fit very well.

e.g., the number of drinks per week; the number of arrests per year

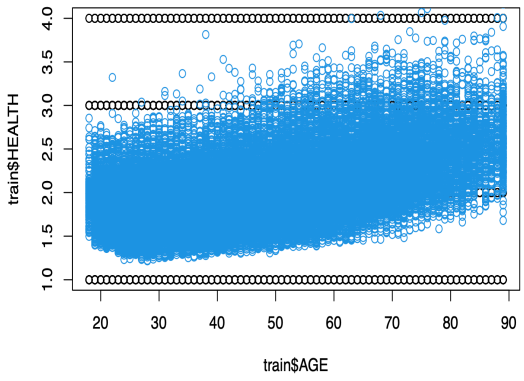
The log likelihood function:

$$\ln L = \sum_{i=1}^{n} (y_i\ \ln \lambda_i - \lambda_i) + \text{constant}$$

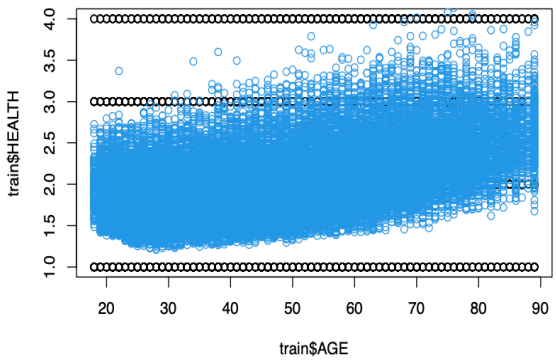
Our Model:

```
model<-glm(HEALTH ~ ., data = train, family = 'poisson')
```

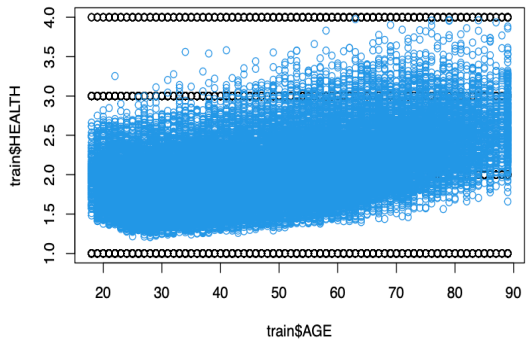
After training the models with our data sets:



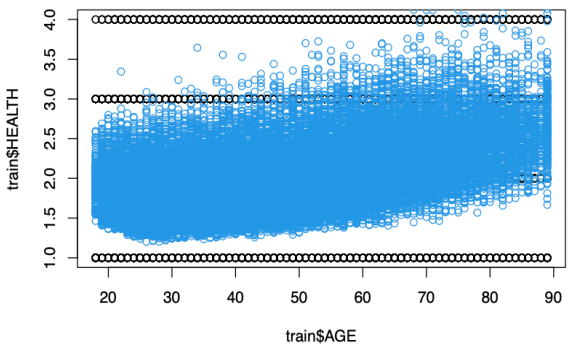
1st Mice Model



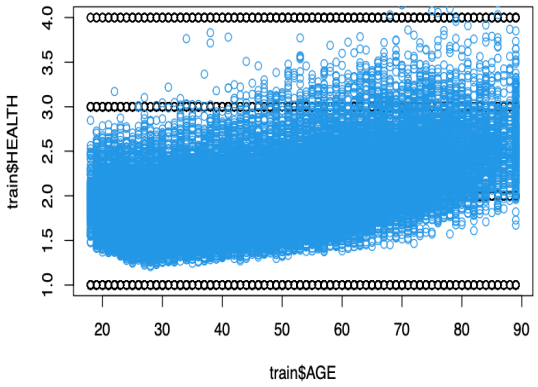
2nd Mice Model



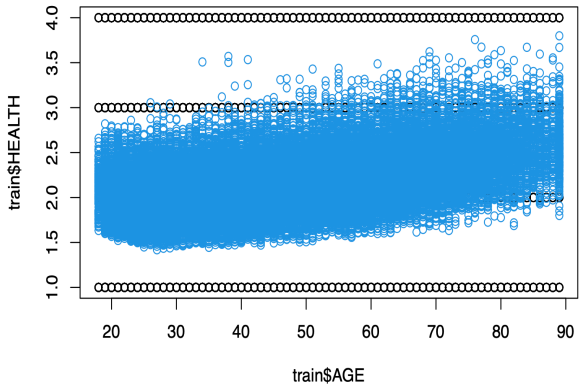
3rd Mice Model



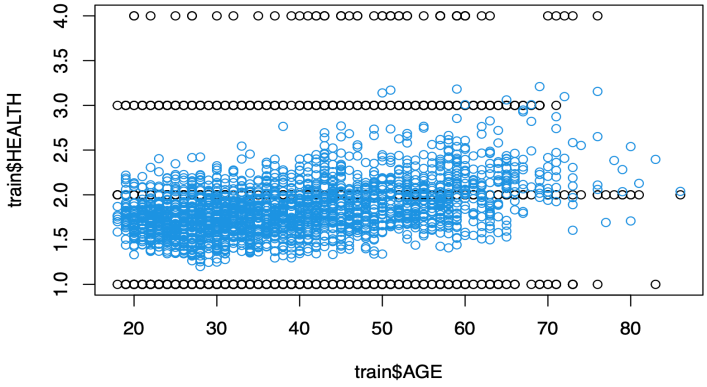
4th Mice Model



5th Mice Model



Model with Random Values



Model with complete data set

These graphs depict black dots are fixed values from dataset, blue circles are predicted by poisson model.

Average Beta Coefficients:

In our case, we took a scenario, instead of comparing all the 5 mice models, we thought of averaging the beta coefficients across the 5 summaries.

Average Variance:

Taking average across the 5 summaries of square of standard errors.

$$\sum_{m}(se)^2$$

Pooled Variance:

It combines the variance estimates within the individual groups. It is a better estimate of the common group variance than either of the individual group variance. It has the assumption of having same number of variables across all groups.

$$Pooled_{va} = Avg_{va} + (1 + \frac{1}{m}) * (\frac{1}{m-1}) * \sum(\beta_i - \beta_{av})^2$$

By using pooled variance all the five mice models, we created a summary object:

	Vars	Estimate	Std_err	Pooled_var	z_value
(intercept)	-3.419096	33.107369	1096.097855	-0.103273	
YEAR	0.002005	0.004394	0.000019	0.456427	
WRKSTAT	0.021638	0.046409	0.002154	0.466253	
MARITAL	0.001820	0.004597	0.000021	0.395857	
AGE	0.004242	0.009257	0.000086	0.458258	
EDUC	-0.023408	0.052498	0.002756	-0.445890	
SEX	-0.011020	0.025433	0.000647	-0.433300	
RACE	0.020349	0.044119	0.001946	0.461230	
RINCOME	-0.004311	0.009544	0.000091	-0.451676	
REGION	0.003028	0.006753	0.000046	0.448409	
NATDRUG	-0.011954	0.027017	0.000730	-0.442443	
HAPPY	0.148327	0.276818	0.076628	0.535829	
SMOKE	-0.063771	0.148812	0.022145	-0.428535	

"Null Deviance: 15899.63 on 45286 degrees of freedom"

"Residual Deviance: 12674.11 on 45274 degrees of freedom"

Pooled of 5 models

```
Call:
glm(formula = HEALTH ~ ., family = "poisson", data = train)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.673325   5.704676  -1.170 0.242082
YEAR         0.003671   0.002891   1.269 0.204275
WRKSTAT      0.002143   0.010576   0.203 0.839421
MARITAL     -0.001896   0.010293  -0.184 0.853824
AGE          0.004668   0.001236   3.777 0.000159 ***
EDUC        -0.025165   0.005750  -4.377 1.20e-05 ***
SEX         -0.004046   0.032949  -0.123 0.902274
RACE         0.022232   0.034279   0.649 0.516623
RINCOME     -0.005749   0.005561  -1.034 0.301193
REGION       0.004793   0.006338   0.756 0.449519
NATDRUG      0.008600   0.025402   0.339 0.734939
HAPPY        0.129316   0.024761   5.223 1.76e-07 ***
SMOKE       -0.068779   0.032199  -2.136 0.032673 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 713.71  on 2306  degrees of freedom
Residual deviance: 615.22  on 2294  degrees of freedom
AIC: 6331.4
```

Complete dataset Model

```
Call:
glm(formula = HEALTH ~ ., family = "poisson", data = train)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.4599134  0.5069415  -4.852 1.22e-06 ***
YEAR         0.0016326  0.0002583   6.320 2.61e-10 ***
WRKSTAT      0.0176051  0.0015075  11.679  < 2e-16 ***
MARITAL      0.0029531  0.0021985   1.343  0.17920
AGE          0.0029416  0.0002045  14.384  < 2e-16 ***
EDUC        -0.0179075  0.0010751 -16.656  < 2e-16 ***
SEX          0.0005255  0.0067738   0.078  0.93817
RACE         0.0182635  0.0060348   3.026  0.00247 **
RINCOME     -0.0005275  0.0009650  -0.547  0.58462
REGION       0.0026228  0.0013159   1.993  0.04625 *
NATDRUG     -0.0015056  0.0041633  -0.362  0.71763
HAPPY        0.0910541  0.0049921  18.240  < 2e-16 ***
SMOKE       -0.1357339  0.0064577 -21.019  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 19162  on 45286  degrees of freedom
Residual deviance: 17098  on 45274  degrees of freedom
```

Model with random values

To compare these three models, we use deviance as metric. We know that lesser the deviance, the better the model. Among these three, Model with complete dataset has very less deviance. I choose Model with complete data set is best.

References:

https://gss.norc.org/get-the-data/spss

https://gss.norc.org/documents/spss/GSS_spss.zip data is taken from this one.