Predicting Frequency of Marijuana Usage In the U.S. Population

Abstract: Marijuana, a psychoactive drug commonly consumed in the United States, is becoming easier to obtain--and consequently use--due to its recent legalization in over forty states. As this drug is associated with a wide range of adverse health consequences, we are interested in better understanding whether or not differences in populations would affect marijuana consumption. Specifically, what demographic and background factors can best predict the frequency of marijuana usage in days per year? Through data collected for 12 possible predictor variables from the 2018 National Survey on Drug Use and Health, a multiple linear regression model was fitted using 10 of those variables to predict the [frequency of marijuana use (in number of days per year)]^0.3. The model showed that while the predictor variables improved the fit in comparison to the intercept only model, they could not explain a large portion of variation in the response variable. Future research can consider testing different predictor variables on the frequency of marijuana usage or fitting an alternative model to the data.

1. Background and Introduction

Marijuana, or the dried leaves, flowers, stems, and seeds from plants of genus Cannabis, is a psychoactive drug commonly consumed in the U.S. (1). Marijuana is becoming easier to obtain--and consequently use--due to its recent legalization in over forty states (2). As this drug is associated with a wide range of adverse health consequences, we are interested in better understanding whether or not differences in populations would affect marijuana consumption (3). To answer our research question - **What demographic and background factors can best predict the frequency of marijuana usage?**, we will fit and describe a model to statistically assess the effects of predictors on the response variable.

2. Data and Exploratory Analysis

a. Data and Variables

We used raw data from the 2018 National Survey on Drug Use and Health, an annual survey conducted by the Substance Abuse and Mental Health Services Administration (4). Participants in the survey were chosen through a multistage, stratified experimental design based on the population of certain states. The survey included 56,313 observations from members of noninstitutionalized civilian populations aged 12 years and older within the United States and offers 2,691 variables total that relate to the usage of selected drugs such as marijuana, cocaine, heroin, and alcohol. We used 3 quantitative and 9 categorical predictor variables in our model to predict our response variable, **frequency of marijuana usage in days per year**. **Body mass index** (BMI), the **age when one first used marijuana** (0-81), and the **number of days taken off work in the past year** due to mental health reasons (0-365) were collected for all members of the population. “Bad data,” collected from respondents who skipped or refused to answer survey questions associated with these quantitative variables, was omitted. Unlike the quantitative variables, the categorical variables contained responses from all members of the population. We subsetted each of the categorical variables into smaller, more manageable groups for ease of data analysis and interpretation. We grouped respondent’s **age** into 1 of 8 ranges: “12-15”, “16-19”, “20-23”, “24-29”, “30-34”, “35-49”, “50-64”, and “65+.” For respondent’s **highest level of completed education**, we pooled together data into 1 of 3 groups: “pre-high school”, “high school”, and “college.” Respondent’s **self-reported sex** was divided into either “female” or “male,” just as respondent’s **participation in one or more government assistance programs** was either a “yes” or “no” answer. Respondent’s **race** was categorized into “non-hispanic white,” “non-hispanic black,” “hispanic,” and “other.” **Total yearly family income** for respondents spanned 7 brackets, including “$75,000 or more,” “$50,000 - $74,999,” “$40,000 - $49,999,” “$30,000 - 39,999,” “ $20,000 - 29,999,” “$10,000 - 19,999,” and “less than $10,000.” Respondent’s **employment status** was classified as “full time, “part time,” “unemployment,” “not in the labor force,” and “under the age of 15.” We grouped respondent’s **self-reported health** into “excellent,” “very good,” “good”, and “adequate” categories. **Geographic location** was based on a respondent’s occupancy in a “large metropolitan,” “small metropolitan,” or “non-metropolitan” area. Lastly, after omitting respondents who skipped or refused to answer how many days (0-365) they used marijuana in the past year, we identified 18,621 members of the population. We used their responses to build our model.

b. Exploratory Data Analysis

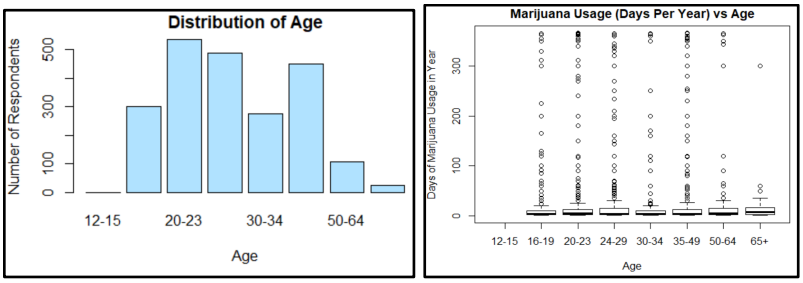
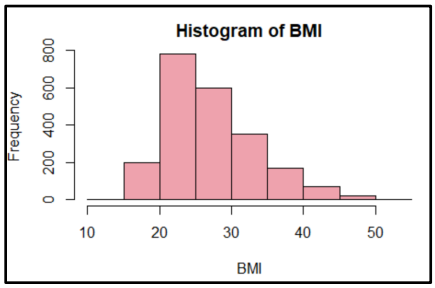
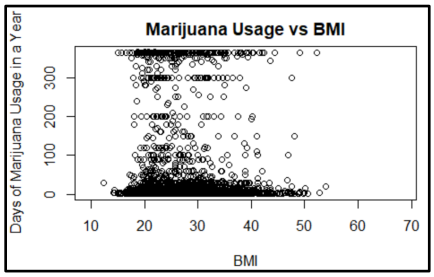


Figure 1. Bar graphs showing the distribution of a few individual variables and boxplots/scatterplots of the predictor vs the response variable

We illustrated two of our predictor variables, age and BMI, and their individual relationship to the response variable **(Figure 1)**. In the bar graph for the age predictor variable, the age range of 35-49 years old is visually disproportionately larger than the rest. From the side-by-side boxplots, the majority of responses for each age range is centered at less than 50 days of marijuana usage per year, yielding outliers that may adversely impact the regression. For the quantitative predictor variable, BMI, the histogram displays a relatively normal distribution, but still slightly right skewed, centered at around a value of 25. The scatter plot of BMI against the response variable illustrates a relationship that does not appear to be linear, but this is an issue we plan to address through model transformations.





3. Model and Results

a. Analytic Methods

As our research question focused on identifying important demographic and background contributing to marijuana consumption, we used a multiple linear regression model to fit our data. A model with all predictors violated nearly all assumptions per model diagnostic plots**.** While boxcox transformation was more efficient and effective than transforming each quantitative predictors and the response variable, we recognized it only could be used for positive values of the response. As a result, we omitted cases where respondents recorded using marijuana for 0 days during the past year to generate a final dataset of 2,185 usable responses. The resulting λ value for this data suggested Y^-0.3 was the best transformation **(Appendix)**. After applying forward, backward, and stepwise selection, which removed variables based on AIC values, we decided to remove two variables, **geographic location** and **participation in one or more government assistance programs,** and use the remaining predictors in the final model.

b. Final Model and Results

Baseline Equation:

$$\frac{1}{MarijuanaUsage^{0.3}} = (0.5161930) + 0.0099482(mjold) + 0.0033616(BMI) - 0.0002916(daysnotworked)$$

Wanting to create a simple model, we did not use interaction terms such that the slopes of each group of a categorical variable are parallel to each other. By contrast, y-intercept can vary based on coefficients from categorical variables, as inferred from t-statistics and associated

p-values. With the final transformed model, all model assumptions are satisfied, and there are no traces of multicollinearity among quantitative variables. As seen from the baseline equation, when all quantitative predictor variables are zero, we would expect the (days of marijuana usage in a year)^-0.3 to be 0.5161930. Slopes associated with the quantitative predictor variables are also revealing. In holding **BMI** and the **number of days taken off work in the past year** constant, we would expect an additional year increase in the **age one first used marijuana** to be associated with, on average, an increase of 0.0099482 (days of marijuana usage in a year)^-0.3.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.162e-01	3.048e-02	16.936	< 2e-16
age.data.mjold	9.948e-03	8.950e-04	11.115	< 2e-16
factor(age.data.health.high)very good	-1.029e-02	1.280e-02	-0.804	0.421354
factor(age.data.health.high)good	-5.872e-02	1.451e-02	-4.046	5.37e-05
factor(age.data.health.high)< good	-7.958e-02	1.877e-02	-4.240	2.32e-05
factor(age.data.grade.col)high	-6.337e-02	1.147e-02	-5.524	3.66e-08
factor(age.data.grade.col)pre high	-7.713e-02	5.797e-02	-1.331	0.183465
age.data.irsex.female	-5.108e-02	9.605e-03	-5.318	1.14e-07
age.data.daysworked	-2.916e-04	8.442e-05	-3.455	0.000561
factor(age.data.age.young)20-23	-6.320e-02	1.651e-02	-3.827	0.000133
factor(age.data.age.young)24-29	-7.654e-02	1.726e-02	-4.435	9.61e-06
factor(age.data.age.young)30-34	-6.422e-02	1.977e-02	-3.249	0.001175
factor(age.data.age.young)35-49	-8.946e-02	1.801e-02	-4.968	7.24e-07
factor(age.data.age.young)50-64	-6.526e-02	2.594e-02	-2.516	0.011936
factor(age.data.age.young)65+	-1.509e-01	4.402e-02	-3.428	0.000618
age.data.BMI2	3.362e-03	7.965e-04	4.220	2.53e-05
factor(age.data.income.high)<10K	-4.587e-03	1.774e-02	-0.259	0.795969
factor(age.data.income.high)19K	-3.992e-02	1.710e-02	-2.334	0.019665
factor(age.data.income.high)29K	-6.104e-02	1.762e-02	-3.465	0.000540
factor(age.data.income.high)39K	3.549e-03	1.737e-02	0.204	0.838068
factor(age.data.income.high)49k	-3.507e-03	1.719e-02	-0.204	0.838413
factor(age.data.income.high)74K	2.158e-03	1.435e-02	0.150	0.880509
factor(age.data.work.full)part time	-6.618e-04	1.263e-02	-0.052	0.958213
factor(age.data.work.full)unemployed	-5.321e-02	2.153e-02	-2.472	0.013515
factor(age.data.work.full)not in force	-2.982e-02	1.402e-02	-2.127	0.033483
factor(age.data.race.white)nonhisp black	-3.304e-02	1.832e-02	-1.804	0.071344
factor(age.data.race.white)nonhisp other	1.829e-02	1.634e-02	1.120	0.262969
factor(age.data.race.white)hisp	2.030e-02	1.419e-02	1.430	0.152727

When compared to α=0.05, the p-value of < 2.2e-16 associated with the F-statistic of 16.21 suggests the predictors improve the model fit in comparison to that of the intercept-only model. However, as the residual standard error of 0.2282 and the adjusted R-squared value of 0.1452 are both small, little variation in the response variable can be explained by our predictors. Admittedly, studies conducted in fields such as psychology often have low R-squared values given the difficulty in properly predicting human behavior (5). Despite this shortcoming, statistical test outcomes still provide useful information. Some of the coefficients that included zero in 95% confidence intervals **(Appendix)** are **employment status** (part time), **self-reported health** (very good), **total yearly family income** (<10K, 39K, 49K, 74K), **race** (all three), and **highest completed education level** (pre-high school). Associated with individual p-values that are larger than α=0.05, these results suggest there may not be enough evidence to conclude a significant difference in the mean number of days per year one uses marijuana exists between these groups and their reference groups.

4. Discussion/Conclusions

Our objective was to determine which demographic and background factors can best predict the frequency of marijuana usage in the U.S. Despite containing 10 predictors from model selection, our final model does not seem to be very effective at predicting our response variable. As this outcome may result from not selecting enough variables to include in our model, we want to explore the effects of adding other predictors. In addition, future work can analyze categorical predictors that were found to be statistically different from their reference groups in greater detail. Recognizing that our categorical variables may have been subsetted too much such as in the case of **total yearly family income**, perhaps pooling together data to create more robust, inclusive groups would result in a more generalizable model. Lastly, since our model cannot be extrapolated to cases where respondents reported not using marijuana in the past year, future work can be undertaken to either attempt an alternative transformation to satisfy model diagnostics or recode the response variable for use in a model such as logistic regression.

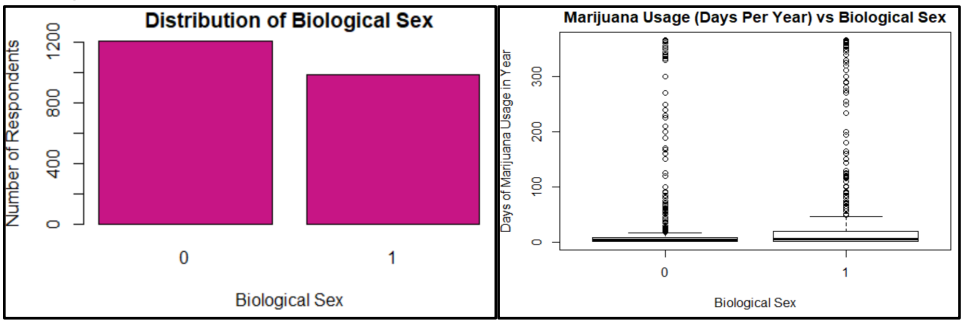
References

1. “Drug Facts: What is Marijuana?.” National Institute on Drug Abuse, 2019.
2. DuPont, RL. “Marijuana Legalization Has Led to More Use and Addiction While Illegal Market Continues to Thrive.” RiverMend Health.
3. Spinola S, Park A, Maisto SA, Chung T. “Motivation Precedes Goal Setting in Prediction of Cannabis Treatment Outcomes in Adolescents.” J Child Adolesc Subst Abuse. 2017;26(2):132–140. doi:10.1080/1067828X.2016.1237917
4. “National Survey on Drug Use and Health 2018 (NSDUH-2018-DS0001,” Substance Abuse and Mental Health Archive, 2018.
5. Onditi, AA. “Relationship between Customer Personality, Service Features and Customer Loyalty in the Banking Sector: A Survey of Banks in Homabay County, Kenya.” International Journal of Business and Social Science. 2013;4(15):132-150.

Appendix

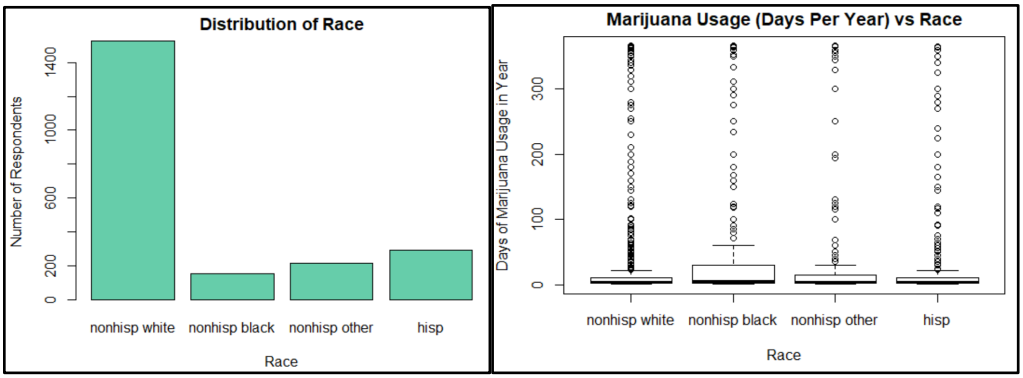
EDA Plots:

Biological Sex:

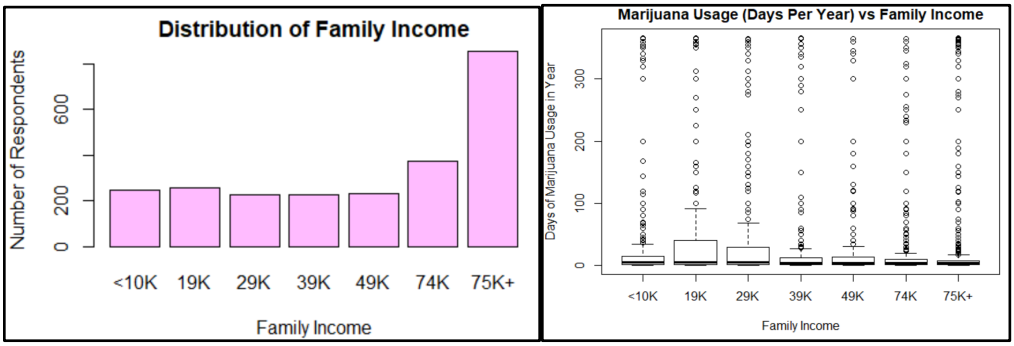


*0=Female, 1=Male

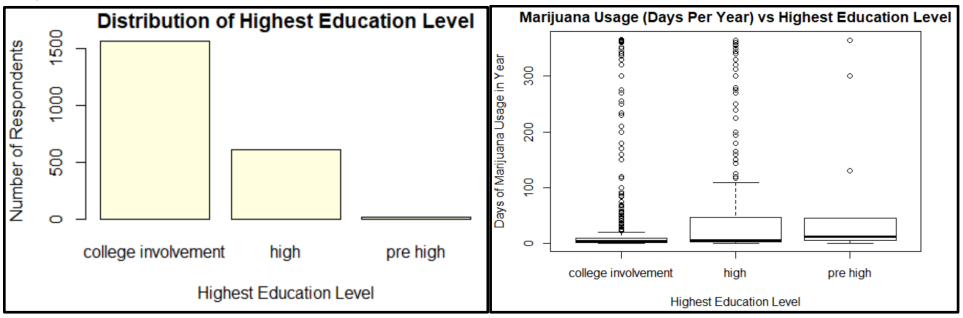
Race:



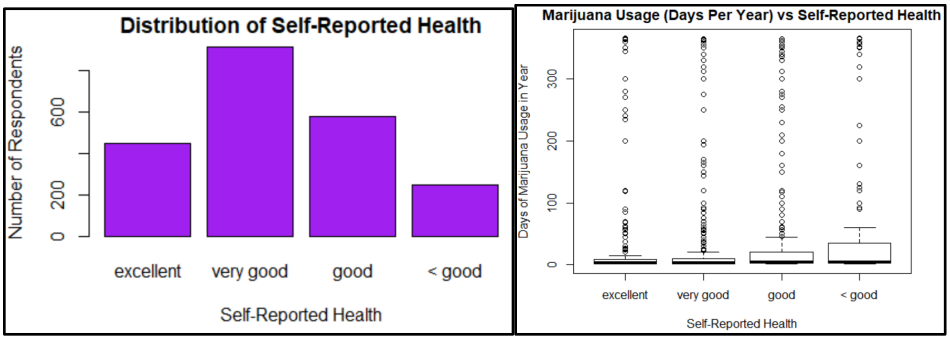
Family Income Level:



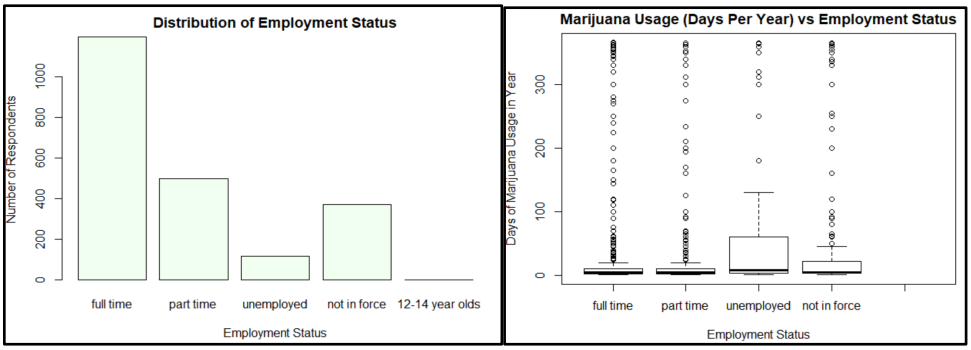
Highest Level of Education:



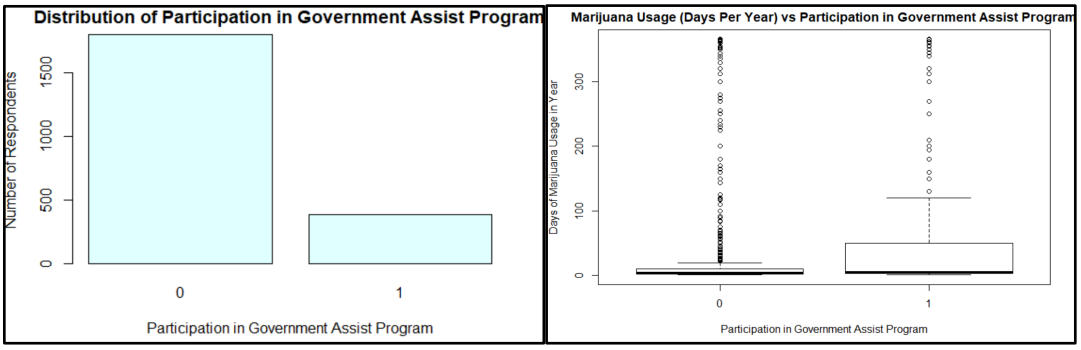
Self-Reported Health:



Employment:

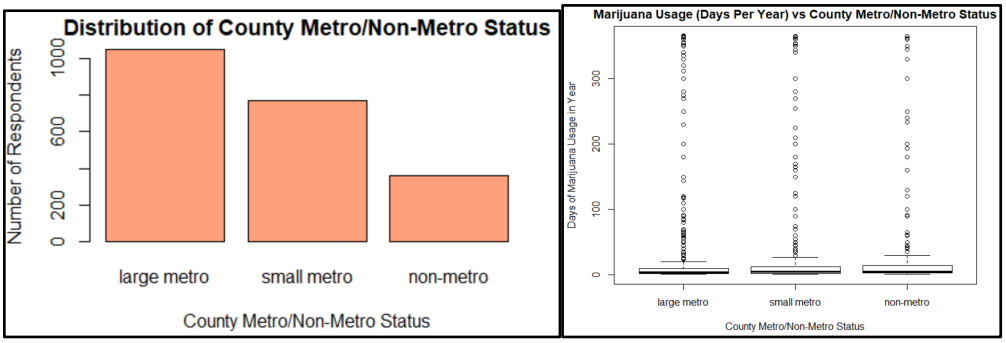


Participation in Government Assist Program:

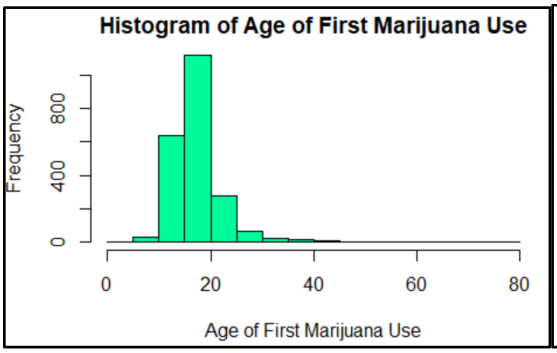


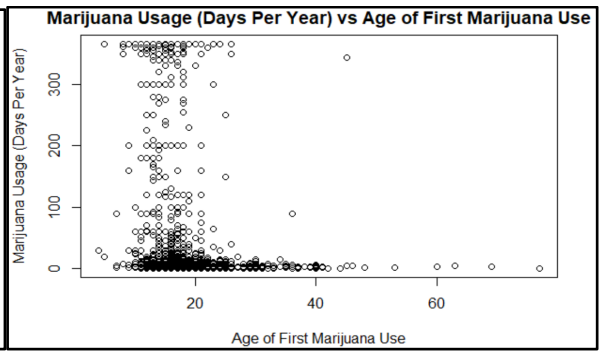
*0=No, 1=Yes

County Metro/Non-Metro Status:

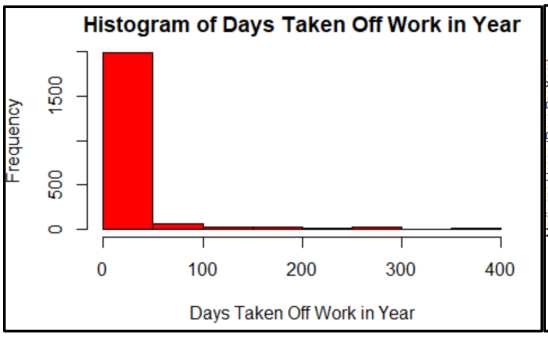


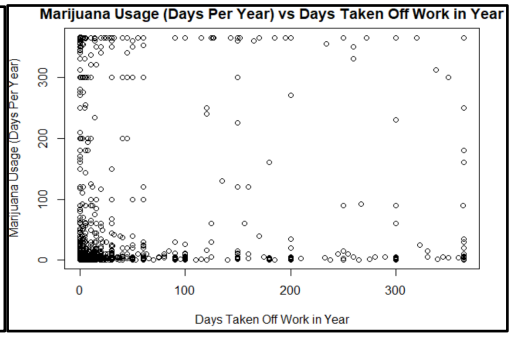
Age of First Marijuana Use:



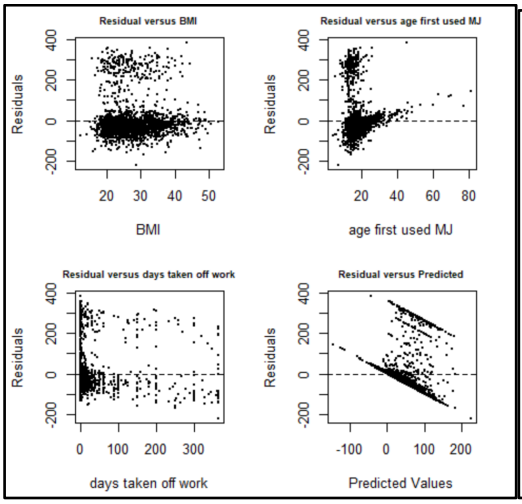


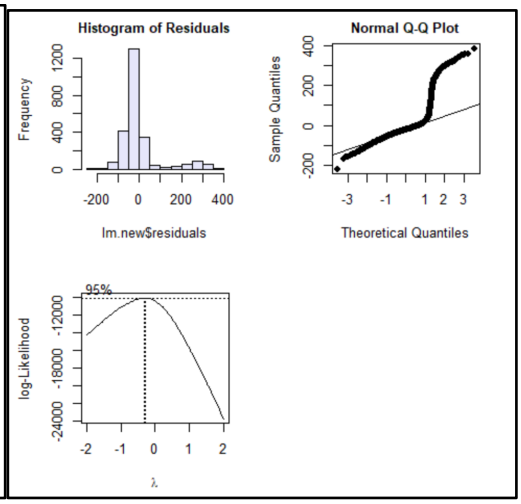
Number of Days Taken Off Work:



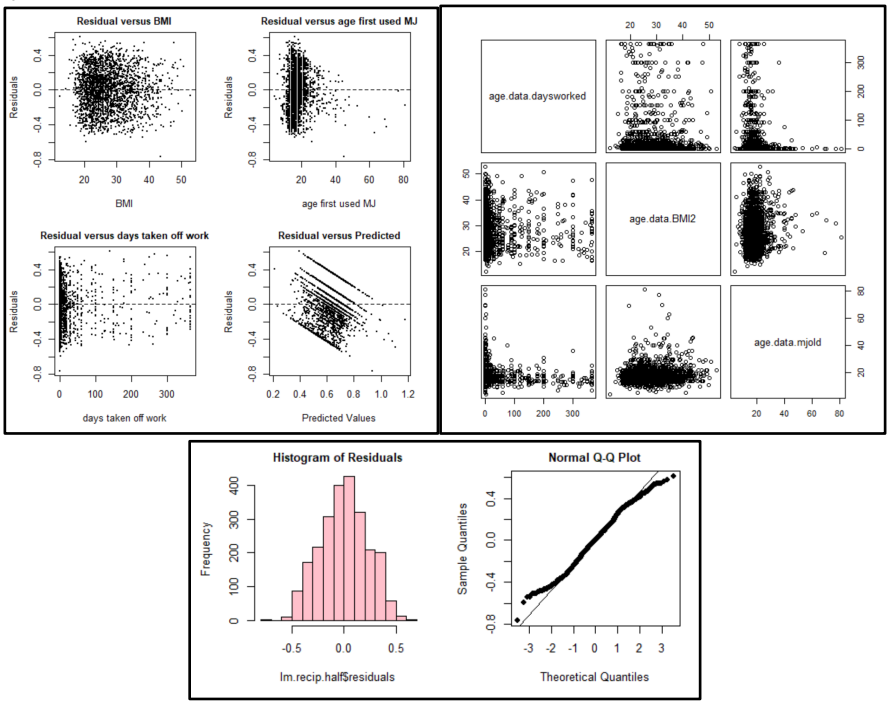


<u>Model Assumption Plots and Boxcox of model with all predictors (N= 2185, before transformation):</u>





Model Assumption Plots and Boxcox of model with all predictors (N= 2185, after transformation utilizing Y^-0.3):



Equations:

Self-Reported Health: *Baseline group is "Excellent"

Very Good: $\frac{1}{MarijuanaUsage^{0.3}} = (0.5161930 - 0.0102916) + 0.0099482(mjold) + 0.0033616(BMI) - 0.0002916(daysnotworked)$

Good: $\frac{1}{MarijuanaUsage^{0.3}} = (0.5161930 - 0.0587189) + 0.0099482(mjold) + 0.0033616(BMI) - 0.0002916(daysnotworked)$

Fair/Bad: $\frac{1}{MarijuanaUsage^{0.3}} = (0.5161930 - 0.0795779) + 0.0099482(mjold) + 0.0033616(BMI) - 0.0002916(daysnotworked)$

Age: *Baseline group is "16-19"

20-23: $\frac{1}{MarijuanaUsage^{0.3}} = (0.5161930 - 0.0631967) + 0.0099482(mjold) + 0.0033616(BMI) - 0.0002916(daysnotworked)$

24-29: $\frac{1}{MarijuanaUsage^{0.3}} = (0.5161930 - 0.0765385) + 0.0099482(mjold) + 0.0033616(BMI) - 0.0002916(daysnotworked)$

30-34: $\frac{1}{MarijuanaUsage^{0.3}} = (0.5161930 - 0.0642212) + 0.0099482(mjold) + 0.0033616(BMI) - 0.0002916(daysnotworked)$

35-49: $\frac{1}{MarijuanaUsage^{0.3}} = (0.5161930 - 0.0894630) + 0.0099482(mjold) + 0.0033616(BMI) - 0.0002916(daysnotworked)$

50-64: $\frac{1}{MarijuanaUsage^{0.3}} = (0.5161930 - 0.0652558) + 0.0099482(mjold) + 0.0033616(BMI) - 0.0002916(daysnotworked)$

65+: $\frac{1}{MarijuanaUsage^{0.3}} = (0.5161930 - 0.1509035) + 0.0099482(mjold) + 0.0033616(BMI) - 0.0002916(daysnotworked)$

Gender: *Baseline group is "Female"

Male: $\frac{1}{MarijuanaUsage^{0.3}} = (0.5161930 - 0.0510794) + 0.0099482(mjold) + 0.0033616(BMI) - 0.0002916(daysnotworked)$

Education: *Baseline group is "College"

High: $\frac{1}{MarijuanaUsage^{0.3}} = (0.5161930 - 0.0633713) + 0.0099482(mjold) + 0.0033616(BMI) - 0.0002916(daysnotworked)$

Pre-High: $\frac{1}{MarijuanaUsage^{0.3}} = (0.5161930 - 0.0771331) + 0.0099482(mjold) + 0.0033616(BMI) - 0.0002916(daysnotworked)$

Race: *Baseline group is "White"

Black: $\frac{1}{MarijuanaUsage^{0.3}} = (0.5161930 - 0.0330426) + 0.0099482(mjold) + 0.0033616(BMI) - 0.0002916(daysnotworked)$

Other: $\frac{1}{MarijuanaUsage^{0.3}} = (0.5161930 + 0.0182915) + 0.0099482(mjold) + 0.0033616(BMI) - 0.0002916(daysnotworked)$

Hispanic: $\frac{1}{MarijuanaUsage^{0.3}} = (0.5161930 + 0.0202958) + 0.0099482(mjold) + 0.0033616(BMI) - 0.0002916(daysnotworked)$

Family Income: *Baseline group is "75K+"

<10K: $\frac{1}{MarijuanaUsage^{0.3}} = (0.5161930 - 0.0045869) + 0.0099482(mjold) + 0.0033616(BMI) - 0.0002916(daysnotworked)$

19K: $\frac{1}{MarijuanaUsage^{0.3}} = (0.5161930 - 0.0399223) + 0.0099482(mjold) + 0.0033616(BMI) - 0.0002916(daysnotworked)$

29K: $\frac{1}{MarijuanaUsage^{0.3}} = (0.5161930 - 0.0610392) + 0.0099482(mjold) + 0.0033616(BMI) - 0.0002916(daysnotworked)$

39K: $\frac{1}{MarijuanaUsage^{0.3}} = (0.5161930 + 0.0035493) + 0.0099482(mjold) + 0.0033616(BMI) - 0.0002916(daysnotworked)$

49K: $\frac{1}{MarijuanaUsage^{0.3}} = (0.5161930 - 0.0035067) + 0.0099482(mjold) + 0.0033616(BMI) - 0.0002916(daysnotworked)$

74K: $\frac{1}{MarijuanaUsage^{0.3}} = (0.5161930 + 0.0021580) + 0.0099482(mjold) + 0.0033616(BMI) - 0.0002916(daysnotworked)$

Employment Status: *Baseline group is "Full Time"

Part-Time: $\frac{1}{MarijuanaUsage^{0.3}} = (0.5161930 - 0.0006618) + 0.0099482(mjold) + 0.0033616(BMI) - 0.0002916(daysnotworked)$

Unemployed: $\frac{1}{MarijuanaUsage^{0.3}} = (0.5161930 - 0.0532128) + 0.0099482(mjold) + 0.0033616(BMI) - 0.0002916(daysnotworked)$

Not in Work Force:
$\frac{1}{MarijuanaUsage^{0.3}} = (0.5161930 - 0.0298215) + 0.0099482(mjold) + 0.0033616(BMI) - 0.0002916(daysnotworked)$

95% Confidence Interval Output for Final Model:

	2.5 %
(Intercept)	0.4564235399
factor(age.data.age.young)20-23	-0.0955754623
factor(age.data.age.young)24-29	-0.1103773517
factor(age.data.age.young)30-34	-0.1029851165
factor(age.data.age.young)35-49	-0.1247756496
factor(age.data.age.young)50-64	-0.1161169309
factor(age.data.age.young)65+	-0.2372198933
age.data.irsex.female	-0.0699134633
factor(age.data.grade.col)high	-0.0858654022
factor(age.data.grade.col)pre high	-0.1908116518
factor(age.data.race.white)nonhisp black	-0.0689583706
factor(age.data.race.white)nonhisp other	-0.0137438664
factor(age.data.race.white)hisp	-0.0075276009
factor(age.data.income.high)<10K	-0.0393697315
factor(age.data.income.high)19K	-0.0734601058
factor(age.data.income.high)29K	-0.0955860164
factor(age.data.income.high)39K	-0.0305034747
factor(age.data.income.high)49k	-0.0372241548
factor(age.data.income.high)74K	-0.0259901920
factor(age.data.work.full)part time	-0.0254272540
factor(age.data.work.full)unemployed	-0.0954290787
factor(age.data.work.full)not in force	-0.0573088789
factor(age.data.health.high)very good	-0.0353861170
factor(age.data.health.high)good	-0.0871763955
factor(age.data.health.high)< good	-0.1163805224
age.data.BMI2	0.0017995841
age.data.mjold	0.0081931112
age.data.daysworked	-0.0004571656

	97.5 %
(Intercept)	0.575962416
factor(age.data.age.young)20-23	-0.030817911
factor(age.data.age.young)24-29	-0.042699617
factor(age.data.age.young)30-34	-0.025457273
factor(age.data.age.young)35-49	-0.054150288
factor(age.data.age.young)50-64	-0.014394690
factor(age.data.age.young)65+	-0.064587200
age.data.irsex.female	-0.032245338
factor(age.data.grade.col)high	-0.040877214
factor(age.data.grade.col)pre high	0.036545509
factor(age.data.race.white)nonhisp black	0.002873210
factor(age.data.race.white)nonhisp other	0.050326890
factor(age.data.race.white)hisp	0.048119168
factor(age.data.income.high)<10K	0.030195938
factor(age.data.income.high)19K	-0.006384438
factor(age.data.income.high)29K	-0.026492308
factor(age.data.income.high)39K	0.037601998
factor(age.data.income.high)49k	0.030210714
factor(age.data.income.high)74K	0.030306238
factor(age.data.work.full)part time	0.024103664
factor(age.data.work.full)unemployed	-0.010996471
factor(age.data.work.full)not in force	-0.002334105
factor(age.data.health.high)very good	0.014802941
factor(age.data.health.high)good	-0.030261345
factor(age.data.health.high)< good	-0.042775331
age.data.BMI2	0.004923543
age.data.mjold	0.011703266
age.data.daysworked	-0.000126093