US Youth Smoking Analysis*

Which group of youth should be the target of smoking prevention?

Huakun Shen

27 April 2022

Abstract

An statistical analysis has been conducted to analyze national youth tobacco usage using NYTS survey prodiced by CDC. The use of tobacco is known to be addictive and harmful to human health, thus preventing tobacco consumption during the youth is crucial, as youth is likely the time people first start to consume tobacco. It's important to know who are more likely to smoke in order to target this group of people and apply prevention means. It is found that Males high school students whose race is Hispanic, White and Black (in decreasing order) are more likely to smoke. Curiosity is also an important factor. Keeping students in a healthy environment is important.

1 Introduction

The use of tobacco, cigarette and other products related to smoking is known to be addictive and have a negative effect on humans' health. As many habits are formed during the youth, it may be more effective to intervene during this period. Many factors can affect humans' habits. To effectively influence the youth, we should know which group of people we are targeting (are more likely to smoke). For example, what's the relationship between race, age, sex and smoking. I am also interested in the trend of the percentage of smoking population, so that we can have an idea of whether the smoking prevention means are effective.

After doing the analysis I found that high school males students whose race is Hispanic, white or black are more likely to smoke. Curiosity is also a big factor, lowing students' curiosity could also potentially prevent smoking. However, there is a paradox about curiosity. Teaching students about the risks of smoking could raise their curiosity, not teaching the risk is obviously not improper. So, providing a healthy environment for students and teaching them about the risk of tobacco consumption are important.

2 Data

This analysis is done using R (R Core Team 2020), dplyr(Wickham et al. 2021), tidyverse(Wickham et al. 2019), here(Müller 2020). Graphs and tables are generated using and ggplot2(Wickham 2016), kableExtra(Zhu 2021), and gridExtra(Auguie 2017).

2.1 Data Source

The data used in this report was retrieved from NYTS("National Youth Tobacco Survey (NYTS)" 2022) from CDC. The dataset was designed to provide national data on indicators key to the design, implementation

*Code and data are available at: https://github.com/HuakunShen/NYTS-Analysis

and evaluation of comprehensive tobacco prevention and control programs. The goal of the dataset is to reduce tobacco use among youth. Historical data is available from 1999 to 2021 (some years are skipped). In this report, I used the data of recent years (2015-2020) to see the trend of tobacco use in the youth in the US.

The data collected in different years are different. There are hundreds of columns for each year. So I pick some of the common columns that are representative, they are

- race: Asian, Black, White, Hispanic, AI/AN or NHOPI
- age: age with a offset of -9, i.e. 1 if a person is 9 years old, 2 if 10 years old and so on
- schooltype: middle school or high school
- smoked: whether have been smoked in the past 30 days
- curious: have been curious about smoking a cigarette
  - Definitely yes and probably yes are both considered **TRUE**
  - Definitely no and probably no are both considered **FALSE**

Since the data raw dataset's columns are not labelled in English, I pre-processed the data to make it easier to understand.

The **smoked** column is a combination of multiple survey questions (using **or** operator), the description is as follows. Which means, if either of the following columns are **TRUE**, **smoked** will be TRUE.

Col	Description
ccigt_r	Smoked cigarettes on 1 or more days in the past 30 days
ccigar_r	Smoked cigars, cigarillos, or little cigars on 1 or more days in the past 30 days
cslt_r	Used chewing tobacco, snuff, or dip on 1 or more days in the past 30 days
cpipe_r	Smoked tobacco out of a pipe (not a hookah or waterpipe) on 1 or more days in the past 30 days
cbidis_r	Smoked bidis on 1 or more days in the past 30 days
chookah_r	Smoked tobacco out of a hookah or waterpipe on 1 or more days in the past 30 days
csnus_r	Used snus, such as Camel or Marlboro Snus, on 1 or more days in the past 30 days
cdissolv_r	Used dissolvable tobacco, such as Ariva, Camel orbs, Camel sticks, Camel strips, or Stonewall, on 1 or more
celcigt_r	Used electronic cigarettes, such as Ruyan or NJOY, on 1 or more days in the past 30 days

Basically, all kinds of tobacco use in the past 30 days are considered as **smoked**.

This dataset/survey is chosen because it's published by CDC (Centers for Disease Control and Prevention), which should be the most official and authoritative institution for this type of data in the United States. CDC has two other surveys, they are "Alaska Native Adult Tobacco Survey Guidance Manual" and "American Indian Adult Tobacco Survey" which are for 2 minority groups that can not represent the US. A report from WHO called "WHO Global Tobacco Control Report 2011 - Data" has a 404 not found page. NYTS is the only dataset I found that contains data I need and also historical data for many years.

Figure 1 shows a preview of the dataset (selected properties, 2015) by displaying the distribution of smoking population (not percentage) by various properties. The datasets for other years have exactly the same structure.

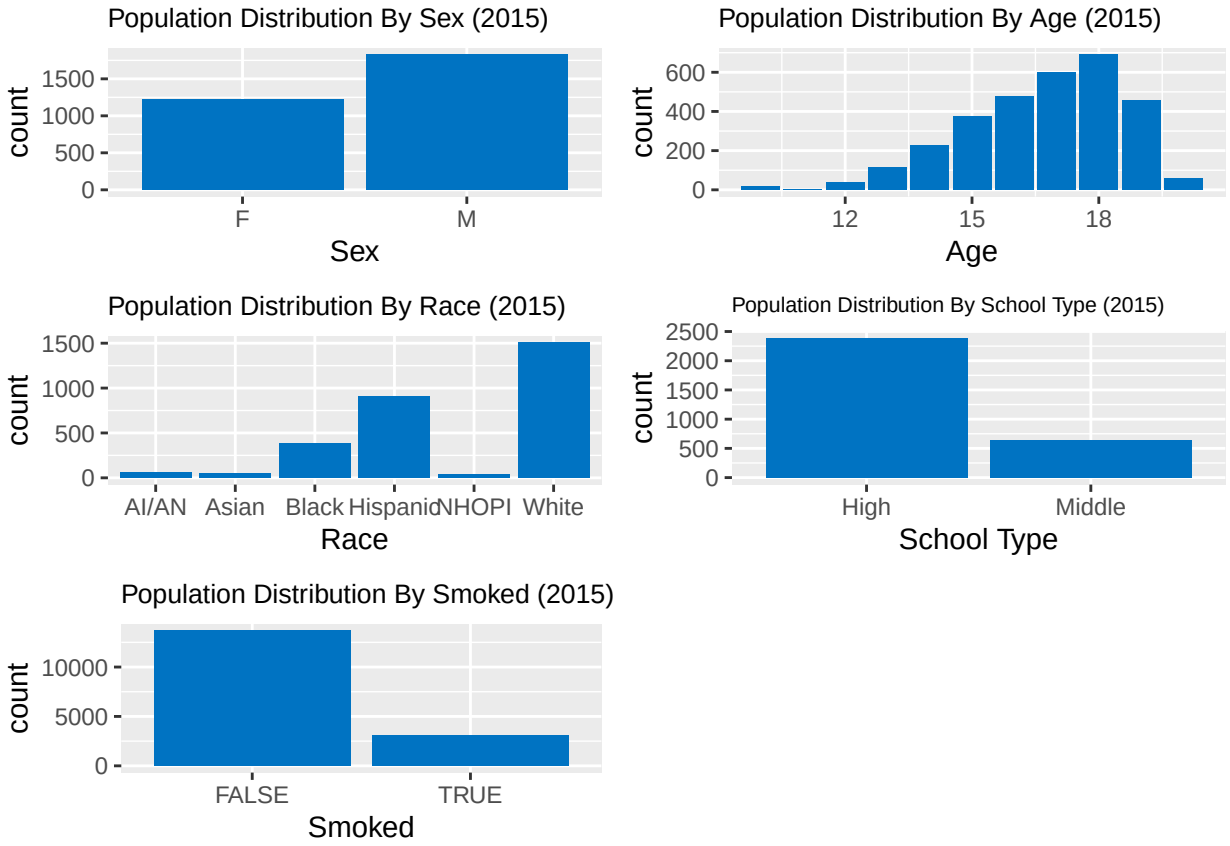

Figure 1: Smoking Population Distribution

3 Model

$$Y \sim Bernoulli(\pi)$$

$$\log \frac{\pi}{1-\pi} = X\beta$$

The response variable Y is **smoked** boolean variable.

The predictor variables include

- Age
- Sex
- Race
- School Type
- curious

The type of model used is GLM (Generalized Linear Model), logistic regression.

Logistic regression is used because the response variable **smoked** has a binary value.

See Model Summary section for description about the intercept/baseline group because this is determined by R (R Core Team 2020).

A model card is included in the appendix, information are obtained from (**penn-GLM?**).

4 Results

4.1 Model Summary

In the following model summaries, the intercept/baseline is the same. The baseline group is Female, Age=9, Asian, high school students who have never been curious about smoking a cigarette.

Since the sample size of AI/AN and NHOPI are too low, these 2 races are removed from the models (so that they don't serve as the baseline group).

Most selected variables are significant except for **schooltype**, which has been significant in 2 of the 5 years.

Overall, Asians, Females, middle school (low-age) students who had less curiosity in smoking are less likely to smoke.

4.1.1 Model 2015

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.9445	0.2472	-24.0439	0.0000
Age	0.2910	0.0209	13.9352	0.0000
SexM	0.5868	0.0496	11.8344	0.0000
raceBlack	0.6654	0.1798	3.7002	0.0002
raceHispanic	0.6443	0.1732	3.7207	0.0002
raceWhite	0.7458	0.1709	4.3643	0.0000
schooltypeMiddle	-0.1269	0.0859	-1.4765	0.1398
curious	2.5200	0.0559	45.0667	0.0000

All variables are significant except for **schooltype**.

4.1.2 Model 2016

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.2677	0.2485	-25.2206	0.0000
Age	0.2788	0.0196	14.2004	0.0000
SexM	0.4597	0.0475	9.6880	0.0000
raceBlack	1.6230	0.1874	8.6588	0.0000
raceHispanic	1.6379	0.1830	8.9524	0.0000
raceWhite	1.6923	0.1811	9.3435	0.0000
schooltypeMiddle	-0.1581	0.0846	-1.8694	0.0616
curious	1.7966	0.0473	37.9585	0.0000

All variables are significant except for **schooltype**.

4.1.3 Model 2017

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.3806	0.2436	-22.0915	0e+00
Age	0.2372	0.0206	11.5317	0e+00
SexM	0.3450	0.0500	6.9060	0e+00
raceBlack	0.9983	0.1777	5.6179	0e+00
raceHispanic	1.0926	0.1730	6.3140	0e+00
raceWhite	1.2451	0.1698	7.3343	0e+00
schooltypeMiddle	-0.3280	0.0907	-3.6168	3e-04
curious	1.7940	0.0500	35.8812	0e+00

All variables are significant.

4.1.4 Model 2018

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.1596	0.2001	-20.7858	0e+00
Age	0.1796	0.0174	10.3019	0e+00
SexM	0.3093	0.0428	7.2214	0e+00
raceBlack	0.5565	0.1517	3.6691	2e-04
raceHispanic	0.7923	0.1420	5.5804	0e+00
raceWhite	1.1153	0.1390	8.0230	0e+00
schooltypeMiddle	-0.8210	0.0768	-10.6922	0e+00
curious	1.7338	0.0430	40.3437	0e+00

All variables are significant.

4.1.5 Model 2019

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.4607	0.1821	-24.4957	0.0000
Age	0.2290	0.0167	13.6933	0.0000
SexM	0.0787	0.0394	1.9959	0.0459
raceBlack	0.9655	0.1293	7.4648	0.0000
raceHispanic	0.9338	0.1220	7.6542	0.0000
raceM	1.0239	0.1938	5.2844	0.0000
raceWhite	1.1413	0.1195	9.5505	0.0000
schooltypeMiddle	-0.3360	0.0701	-4.7930	0.0000
curious	1.6572	0.0395	41.9798	0.0000

All variables are significant except for **schooltype**.

4.1.6 Model 2020

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.9445	0.2472	-24.0439	0.0000
Age	0.2910	0.0209	13.9352	0.0000
SexM	0.5868	0.0496	11.8344	0.0000
raceBlack	0.6654	0.1798	3.7002	0.0002
raceHispanic	0.6443	0.1732	3.7207	0.0002
raceWhite	0.7458	0.1709	4.3643	0.0000
schooltypeMiddle	-0.1269	0.0859	-1.4765	0.1398
curious	2.5200	0.0559	45.0667	0.0000

All variables are significant.

4.2 Smoking Population Trend

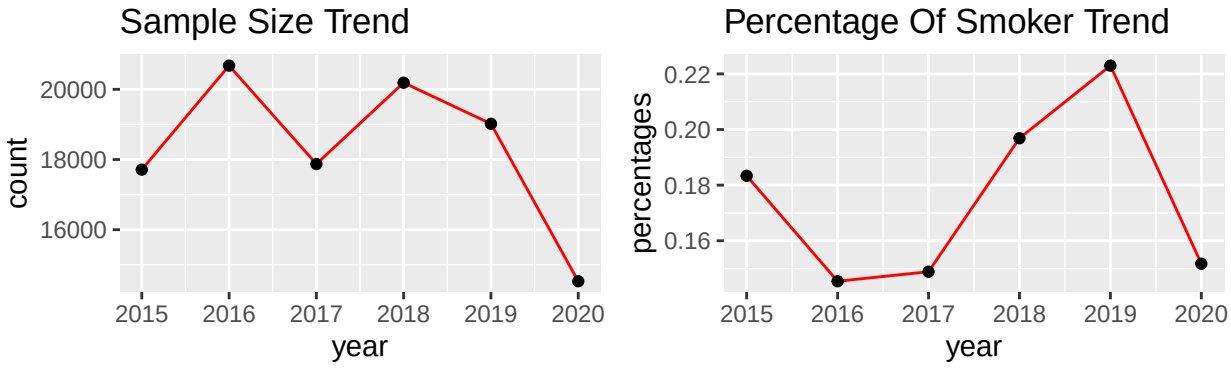

Figure 2: Trend of Sample Size and Smoker Population

From figure **??**fig:trend-plot), we can see a large decrease in smoker percentage from 2015 to 2016, then a large increase in 2018 and 2019. There is a huge decrease in smoking percentage in 2020 with a big decrease in sample size in 2020 as well.

4.3 Density

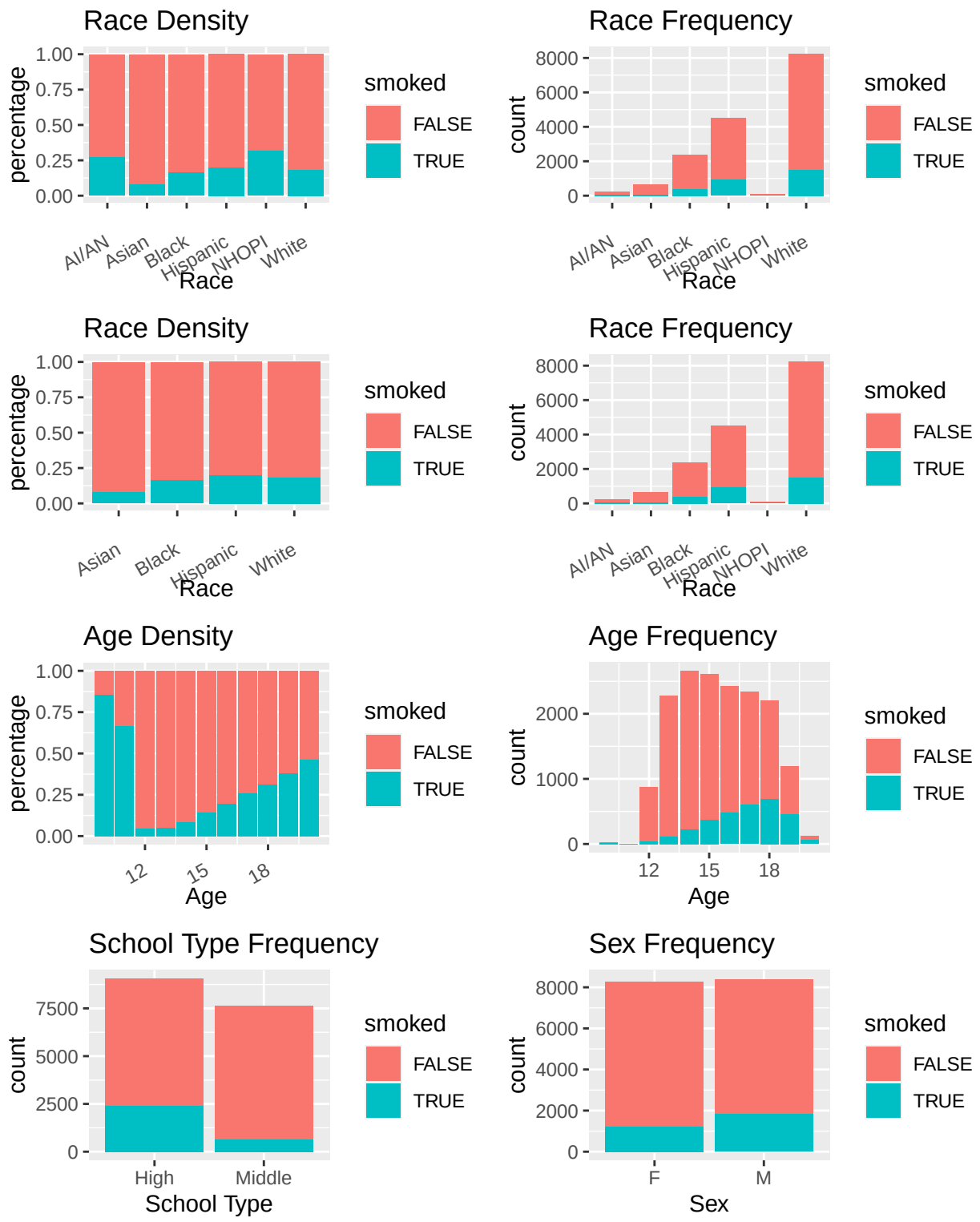

Figure 3: Smoker Density By Group

Table 2: Smoker Percentage By Race

Race	Percentage
Asian	7.68
Black	16.20
Hispanic	20.05
White	18.33

Figure 3 plots the density of smoked population by different groups. The general trend is 1. Higher percentage of males smokes than females 2. As age increases, more adolescents tend to smoke (higher percentage of high school students smoke than middle school) 1. When age is 10-11, the smoking ratio is super high, but this is probably because the sample size is too small. See Age Frequency plot. 3. Higher percentage of AI/AN and NHOPI smoke than other races (but their sample size is very small) 1. See table 2. Hispanic has the highest percentage of smokers, then White, Black and Asian

5 Discussion

5.1 Trend

From figure 2, we can see an overall increasing trend of smoking youth population. The drop of sample size in 2020 could be due to Covid 19, and this could have an impact on the accuracy of collected results. However, even with the large drop in sample size in 2020, the sample size is still pretty large. So, another hypothesis is that smokers has been reduced due to Covid 19 because more people tend to stay home. I believe the overall trend is still an increasing trend with the data of recent years. It may be more accurate if we can look at the historical data of the years before 2015.

5.2 Race

Figure 3 shows that, among the 4 races with the largest populations (Asian, Black, Hispanic and White), Asian has the lowest percentage of smokers. The other 3 groups have relative high smoking percentage.

This hypothesis is also supported by GLM. From Model Summary results, we can see that all variables are significant based on the p-values. Asian is the baseline group, raceBlack, raceWhite and raceHispanic all have positive estimate value.

5.3 Other Variables

In Model Summary section we can see that **SexM** variable is significant and positive for all years. Meaning that more male tends to smoke than female does.

Age is positive, meaning that as age increases, youths are more likely start to smoke.

schooltypeMiddle significant for 2 of the 5 years. It is negative meaning that high school students are more likely to smoke than middle school students are, which is expected as **schooltype** is directly related to age.

curious is also always significant and positive, meaning that students who have been curious in smoking are more likely to smoke later.

5.4 Conclusion

With the evidence from statistical models and plots, we can conclude the target for smoking prevention. The main target are high school male students whose race is hispanic, white or black.

This doesn't mean middle school students should not be focused on. If middle school students start to smoke, it's more likely for them to smoke later or even influence their classmates. This is evidenced by the **curious** variable. As more classmates/friends are smoking, a student should be curious about smoking and are more likely to smoke eventually.

5.5 Limitation

The dataset has a serious limitation when we need to perform an analysis across multiple years. The number of survey questions is increasing every year, more columns are added to the dataset, which is fine. What's troublesome is that many column names are not consistent across years, some questions are even missing in later years. Which could cause serious mistake if not realized, and takes lots of time to find the column names and match them.

Appendix

Model Card - GLM

Model Details

- GLM - Generalized Linear Model
- Developed by John Nelder and R.W.M. Wedderbrun in 1972
- A flexible generalization of ordinary linear regression

Form

$$y_i \sim N(x_i^T \beta, \sigma^2)$$

- y_i (Response Variable) is assumed to follow exponential family distribution with mean $\mu_i = x_i^T \beta$
- x_i contains known covariates
- β contains the coefficients to be estimated
- Fit by least squares and weighted least squares
    - Using SAS's GLM procedure or R's $lm()$ function

Intended Use

GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function. No linear relationship is assumed between the response variable and explanatory variables.

- Binary Logistic Regression (Target is binary)
    - $logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i$
- Poisson Regression (For modelling events whose outcomes are counts)
    - $\log(\lambda_i) = \beta_0 + \beta x_i$
- Advantages over traditional OLS regression
    - No need to transform the response to have a normal distribution
    - Choice of link function is separate from the choice of random component (more flexibility in modeling)
    - Models fitted via MLE, likelihood functions and parameter estimate benefit from asymptotic normal and chi-square distributions

Factors

- Random Component
    - Specifies the probability distribution of the response variable
        - Normal Distribution in ordinary linear regression model
        - Binomial Distribution in binary logistic regression model
- Systematic Component
    - Specifies the explanatory variables $(x_1, x_2, \dots, x_k)$ in the model
- Link Function η or $g(\mu)$
    - Specifies the link between the random and the systematic components
    - Indicates how the expected value of the response variable relates to the linear combination of explanatory variables
    - $\eta = g\big(E(Y_i)\big) = E(Y_i)$ for classical regression
    - $\eta = \log\left(\frac{\pi}{1-\pi}\right) = logit(\pi)$ for logistic regression

Caveats and Recommendations

- The random component (response variable) doesn't not have a separate error term
- Dependent variable Y_i does not need to be normally distributed, typically assumes a distribution from an exponential family (e.g. Binomial, Poisson, Multinomial, Normal, etc.)
- Assumes data Y_i are independently distributed
- Does not assume a linear relationship between the response variable and explanatory variables
- Assume a Linear relationship between the transformed expected response (in terms of the link function) and the explanatory variables (e.g. binary logistic regression $logit(\pi) = \beta_0 + \beta_1 x$))
- Errors are independent (not normally distributed)
- Parameter estimation with MLE

Figure 4: GLM Model Card

References

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

"National Youth Tobacco Survey (NYTS)." 2022. *Centers for Disease Control and Prevention*. Centers for Disease Control; Prevention. https://www.cdc.gov/tobacco/data_statistics/surveys/nyts/index.htm.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.