Data Analytics: A Demographic & Socioeconomic Analysis of American Cigarette Smoking

Mitchell Reavis, Ah-Lian Kor, Sanela Lazarevski
Leeds Beckett University, Leeds LS6 3QS
Mitch.Reavis@outlook.com; {A.Kor, S.Lazarevski}@leedsbeckett.ac.uk

Abstract:
This study attempts to model smoking behavior in the United States using Current Population Survey data from 2010 and 2011. An array of demographic and socioeconomic variables is used in an effort to explain smoking behavior from roughly 139,000 individuals. Two regression techniques are employed to analyze the data. These methods found that individuals with children are more likely to smoke than individuals without children; females are less likely to smoke than males; Hispanics, blacks, and Asians are all less likely to smoke than whites; divorcees and widows are more likely to smoke than single individuals; married individuals are less likely to smoke than singles; retired individuals are less likely to smoke than working ones; unemployed individuals are more likely to smoke than working ones; and as education level increases after high school graduation, smoking rates decrease. Finally, it is recommended that encouraging American children to pursue higher education may be the most effective way to minimize cigarette smoking.

1. Introduction

Smoking cigarettes significantly increase an individual's risk of cancer, specifically lung cancer (US Department of Health and Human Services, 2004). This report attempts to use American Current Population Survey to statistically model smoking behavior for demographic and socioeconomic identifiers in the hope of providing meaningful recommendations to policy makers and anti-smoking campaigners to moderate cigarette smoking in America. First, this report provides a general review of literature surrounding business intelligence and data analytics. These findings are then applied to public health literature to examine how business intelligence and data analytics have been leveraged in the industry. Then, literature surrounding smoking behavior based on socioeconomic and demographic identifiers is reviewed, and hypotheses are formed. The data collection process will then be examined, followed by a brief review of a social science data lifecycle. Two methods of regression analysis are then conducted, presented, and explained. Finally, recommendations for policy makers and anti-smoking campaigns are presented based on findings.

2. Review of Literature

2.1 Business Intelligence & Public Health

Business intelligence (BI) is defined as an innovation that leverages data and analytics to assist business planning and decision making (Elbashir et al, 2011). In this sense, "BI is both a process and a product" (Jourdan et al., 2008, p. 121). Fink et al. (2017) adds that, according to information technology executives, business intelligence systems are among the most encouraging and valuable technologies in the present and future. The growth in popularity of BI and supporting systems has largely been spurred by the increase in the amount of data (i.e. 'Big Data') available to firms (Trieu, 2017).

The rising availability of Big Data has allowed BI to venture into different sectors that have rather short histories of data leverage, such as the global public health industry (Davidson, 2015, Thayer et al, 2013). The United States' public health sector, for example, began investing information technology systems and electronic health records in the early 2000's to make health care more efficient and affordable (Steward, 2005; Lyke, 2009). In fact, studies conclude that almost two-thirds of health care companies increased BI spending in 2015; one firm even planned to spend approximately $2 million on BI and analytics in 2015 (Eddy, 2015). By efficiently maintaining electronic patient records of diagnosis and outcome, further creating a large national database, physicians have been able to utilize historical data to mitigate potentially dangerous mistakes and misdiagnoses (Steward, 2005; Kohli & Tan, 2016).

The wide-spread use of data to inform decision making has also been utilized by health care industry regulators and policy makers (Kohli & Tan, 2016). Many studies focus on vice deterrence: discouraging consumption of harmful substances like drugs and alcohol. For example, Hollingworth et al. (2006) conducted a study that analyzed data from the National Longitudinal Survey of Youth and concluded that high alcohol tax-rates, and television advertising bans on alcohol content, reduced drinking among youth and premature mortality among adults. Similarly, Brennan et al. (2016) applied data analytics to aid the UK's Medical Research Council to further develop macro-level alcohol-related health policies to deter binge drinking with high tax.

2.2 Demographics, Socioeconomics, & Smoking Behavior

2.2.1 Gender

Research has used analytics to explore the relationships between broad demographics identifiers and general behavioral tendencies like substance abuse. Chun & Mobley (2010) found a statistically significant increase in the likelihood of substance abuse, risky sexual behavior, poor academic marks, and aggression among American youth males compared to youth females. Additionally, Syamlal et al. (2014) found a statistically significant difference in smoking rates among working men compared to their female counterparts. These findings support the hypothesis: *men will have higher rates of smoking than women when controlling for other demographic and socioeconomic factors*.

2.2.2 Race

Racial differences in smoking habit have also been explored empirically. Fagan et al. (2007) analyzed data from the 2003 American Current Population Survey (CPS) and found that non-Hispanic white Americans were less likely to attempt to quit smoking than their non-Hispanic black counterparts. Further, Soulakova et al. (2017) used 2011 CPS data and observed that both Hispanics and non-Hispanic blacks were more likely to attempt to quit smoking than their non-Hispanic white counterparts. These studies support the research hypothesis: *white civilians will have the highest predicted probability of being smokers when controlling for other demographic and socioeconomic factors.*

2.2.3 Marital Status

Gender differences in risky behavior make it reasonable to expect differences in smoking habits among individuals with differing marital statuses (Chun & Mobley, 2010). Pennanen (2014) found that married individuals were least likely to smoke out of any other marital status (divorced, single, or widowed). Cho et al. (2007) similarly found that married individuals were least likely to smoke, followed by single, widowed, and divorced individuals, in that order. These studies support the research hypothesis: *married individuals will have the lowest smoking rates when controlling for other demographic and socioeconomic factors*.

2.2.4 Children

Research has also briefly explored the effect of children on parent smoking behavior. Millis et al. (2011) found that knowledge of the harmful effects of secondhand smoke limits smoking in households with at least one child. Yet, Halterman et al. (2010) found contradicting results in that the presence of children increases emotional and financial stress among parents, which is linked to higher smoking rates. Acknowledging conflicting results, it can be hypothesized: *the presence of children on smoking behavior will be either minimal or statistically insignificant*.

2.2.5 Employment

From a theoretical view, studying the effect of employment status on smoking rates can be complicated. There is the economic argument that cigarettes are so-called normal goods – goods for which demand and disposable income are positively correlated (Ruhm, 2000). Yet, there is also the psychological argument stating that unemployed, discouraged individuals cope with stress with vices like tobacco (Harris & Edlund, 2005). Despite this double argument, the majority of studies indicate that joblessness is highly correlated with substance use, including cigarettes (Henkel, 2011). Therefore, there is support for the hypothesis: *unemployed individuals are most likely to smoke cigarettes.*

2.2.6 Education

Education has been found to significantly decrease smoking rates. Pennanen et al. (2014) found that less educated people had higher rates of physical nicotine addiction (indicated by biomarkers in blood samples) and general smoking rates. Likewise, De Walque (2007) found that more educated individuals are more likely to be non-smokers and also more likely to be successful at quitting if they did once smoke. These studies support the hypothesis: *years of schooling and smoking prevalence will be negatively correlated.*

3 Methodology

3.1 Data Collection & Variables

The data for this study was collected from the 2010 and 2011 United States Current Population Survey (CPS), which has been systematically accumulated in the Integrated Public Use Microdata Series (IPUMS) by Flood et al. (2015). This sub-section will present all variables used in this study, as well as the applied selection and recoding processes used to arrive at the final data set.

3.1.1 SMOKER

The CPS collected smoking data with the Tobacco Use Supplement (TUS). The data used for this report is comprised of the three most recent surveys available including this supplement. These surveys were conducted January 2011, August 2010, and May 2010. Upon collection, the raw data included 188,119 responses. The *SMOKER* variable was derived from the TSMKER variable in the CPS TUS. All observations for which the TSMKER variable had a label of 'Indeterminate' or 'Not in Universe' were dropped from this study, as these values designate blank or ambiguous responses. Dropping these observations reduced the sample to 139,012 responses. The remaining responses were used to derive the *SMOKER* variable used in this report. Never-smokers and former-smokers were labeled as non-smokers for this report. Former-smokers are considered non-smokers because this report only studies current smoking behavior. The two groups of non-smokers were assigned the value (0) for the derived *SMOKER* variable. Non-daily smokers and everyday smokers were assigned the value (1) for the *SMOKER* variable. It is advantageous to have the *SMOKER* variable as a dichotomous variable (i.e. all responses are either 1 or 0) because it is the dependent variable in this study. These advantages will be discussed in more detail in the Data Analysis section.

3.1.2 CHILD

The *CHILD* variable in this report was derived from the NCHILD variable used in CPS. The NCHILD variable is an integer indicating the number of children present in the household. Observations indicating 1+ children in the household were assigned a (1) for *CHILD*. All responses indicating zero children in the household were assigned (0).

3.1.3 FEMALE

The *FEMALE* variable in this report was derived from the SEX variable used by CPS. CPS assigned males (1) and females (2) in the SEX variable. Although the SEX variable used by CPS is already a binary variable, it had to be re-coded into a 1 or 0 dichotomous variable for proper analysis. All of those who indicated they were female were assigned a value of (1) for this report's *FEMALE* variable; all males were assigned a value of (0). There was no need for a MALE variable in the report because a *FEMALE* value of (0) implies that one is a male.

3.1.4 BLACK // ASIAN // HISPANIC // (WHITE)

This report selected cases in which individuals indicated only a single race, for the three most populous responses (White, Black, Asian) in order to properly control for race demographics. The race variable was derived from a combination of the RACE and HISPAN variables used in CPS. If respondent indicated Hispanic origin, they were assigned a value of (1) for the *HISPANIC* variable used in this study, and a (0) for the *BLACK* and *ASIAN* variables. If a respondent indicated they were black, they were assigned a value of (1) for the BLACK variable used in this study, and a (0) for the *HISPANIC* and *ASIAN* variables. If a respondent indicated they were Asian, they were assigned a value of (1) for the *ASIAN* variable used in this study, and a (0) for the *HISPANIC* and *BLACK* variables. Lastly, a value of (0) for *HISPANIC*, *BLACK*, and *ASIAN* variables implies that the individual is white.

3.1.5 MARRIED // DIVORCED // WIDOWED // (SINGLE)

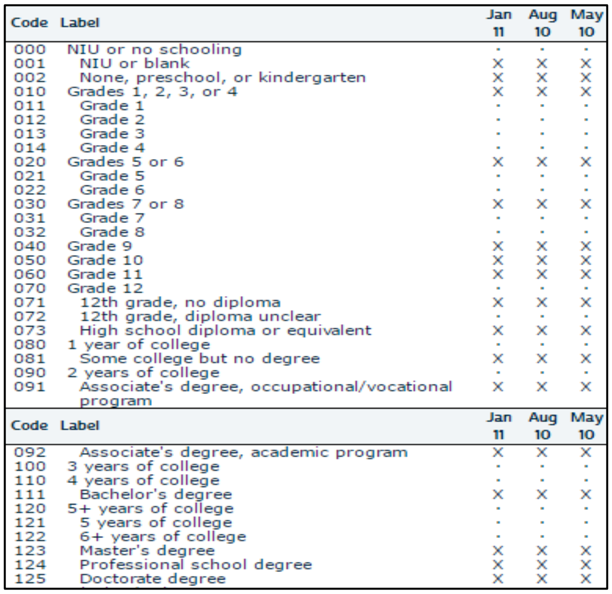
The *MARRIED*, *DIVORCED*, and *WIDOWED* variables in this report were derived from the CPS's MARST variable. Individuals indicating being married and living with their spouse were assigned a value of (1) for *MARRIED* and a (0) for *DIVORCED* and *WIDOWED*. The same logic was applied to those indicating they were divorced or widowed. Cases for which the respondents indicated being married but not living with their spouse, and those who indicated being separated, were not selected for the study due to ambiguity. Lastly, observations that have a value of (0) for *MARRIED*, *DIVORCED*, and *WIDOWED* were single and never married, so no SINGLE variable was created.

3.1.6 FOREIGN

The *FOREIGN* variable in this report was derived from the CPS's NATIVITY variable. Cases of unknown nativity were not selected for this study. Individuals who indicated being born outside of the US were assigned a value of (1) for *FOREIGN*. Individuals who indicated being born in the US were assigned a value of (0) for *FOREIGN*, regardless of the nativity of their parents. The dichotomy of this variable diminished the need to create a NATIVE variable.

3.1.7 UNEMPLOYED // RETIRED // (WORKING)

The *UNEMPLOYED* and *RETIRED* variables in this report were derived from the CPS's EMPSTAT variable. Unemployed, experienced workers were assigned a value of (1) for *EMPLOYED* and a (0) for *RETIRED*. Unemployed, new workers (e.g. recent university graduates) were not included due to minimal responses. Those not in the labor force due to retirement were assigned a value of (1) for *RETIRED* and a (0) for *UNEMPLOYED*. Those employed and working are represented by having values of (0) for both *UMEPLOYED* and *RETIRED*.



Code	Label	Jan 11	Aug 10	May 10
000	NIU or no schooling	·	·	·
001	NIU or blank	X	X	X
002	None, preschool, or kindergarten	X	X	X
010	Grades 1, 2, 3, or 4	X	X	X
011	Grade 1	·	·	·
012	Grade 2	·	·	·
013	Grade 3	·	·	·
014	Grade 4	·	·	·
020	Grades 5 or 6	X	X	X
021	Grade 5	·	·	·
022	Grade 6	·	·	·
030	Grades 7 or 8	X	X	X
031	Grade 7	·	·	·
032	Grade 8	·	·	·
040	Grade 9	X	X	X
050	Grade 10	X	X	X
060	Grade 11	X	X	X
070	Grade 12	·	·	·
071	12th grade, no diploma	X	X	X
072	12th grade, diploma unclear	·	·	·
073	High school diploma or equivalent	X	X	X
080	1 year of college	·	·	·
081	Some college but no degree	X	X	X
090	2 years of college	·	·	·
091	Associate's degree, occupational/vocational program	X	X	X
092	Associate's degree, academic program	X	X	X
100	3 years of college	·	·	·
110	4 years of college	·	·	·
111	Bachelor's degree	X	X	X
120	5+ years of college	·	·	·
121	5 years of college	·	·	·
122	6+ years of college	·	·	·
123	Master's degree	X	X	X
124	Professional school degree	X	X	X
125	Doctorate degree	X	X	X

Figure 1: Possible Responses for the Education variable

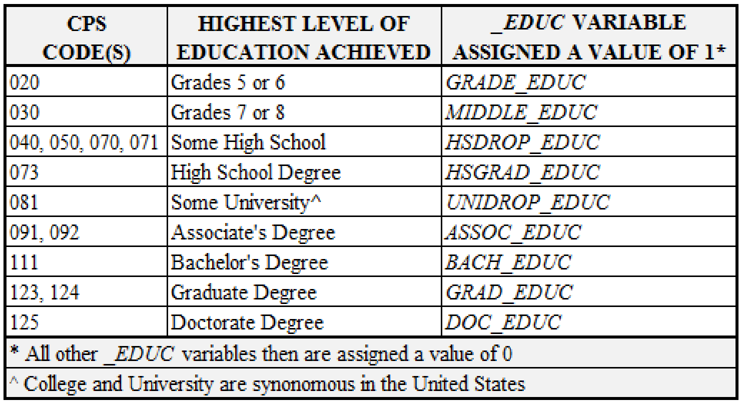
3.1.8 EDUCATION

The *_EDUC* variables in this study required the most extensive recoding process. They were derived from the EDUC variable used in CPS. The range of possible responses are shown in Figure 1 (Flood et al, 2011).

The responses indicate the highest level of education attained. Responses with an (X) indicate at least one respondent identified with that label as the highest level of education achieved. Cases of individuals with (Grade 4 or less) as their highest level of education were not included due to the minimal responses. The remaining observations were coded into nine variables (see Figure 2).

3.1.9 AGE

The *AGE* variable in this study was derived from the AGE variable used in CPS. Respondents aged (18-64) were selected for this survey, as one can purchase tobacco in the US at age 18 (Sullivan, 1990). In order to examine how smoking behavior fluctuates with age increase for this age range, 18 has been subtracted from each CPS AGE. Thus, the new *AGE* variable represents how many years over the age of 18 a respondent is.



CPS CODE(S)	HIGHEST LEVEL OF EDUCATION ACHIEVED	*_EDUC* VARIABLE ASSIGNED A VALUE OF 1*
020	Grades 5 or 6	*GRADE_EDUC*
030	Grades 7 or 8	*MIDDLE_EDUC*
040, 050, 070, 071	Some High School	*HSDROP_EDUC*
073	High School Degree	*HSGRAD_EDUC*
081	Some University^	*UNIDROP_EDUC*
091, 092	Associate's Degree	*ASSOC_EDUC*
111	Bachelor's Degree	*BACH_EDUC*
123, 124	Graduate Degree	*GRAD_EDUC*
125	Doctorate Degree	*DOC_EDUC*
* All other *_EDUC* variables then are assigned a value of 0		
^ College and University are synonomous in the United States		

Figure 2: 9 variables for EDUCATION

3.2 Data Lifecycle

Ball (2012) provides a model for the data lifecycle specifically pertaining to data analysis in social sciences. Given the prevalence of demographic and socioeconomic variables used in this report, it becomes appropriate to adopt this model to study the data lifecycle of CPS. The lifecycle of data is depicted in Figure 3 (Ball, 2012, p. 7).

The data lifecycle of CPS data begins at the Study Concept. Here, the actual CPS survey is designed and data aggregation is determined. The next stage is Data Collection. Here, the physical CPS survey is distributed to a sample of the American population. The data is then coded for computer readability during the Data Processing stage. This is seen in the CPS's use of codes for its education variable (as shown in the *_EDUC* table shown previously). After coding, Data Archiving occurs and the data is included with previous data. In the case of CPS, this would be adding the survey results to the results of prior surveys, maintaining the same coding scheme. During the Data Distribution stage, public access to data is given. This was seen by the IPUMS's ability to collect the CPS data. The next two stages are included in this report. The CPS data was found on the IPUMS site (i.e. Data Discovery phase) and the Data Analysis phase is what will follow in the next sections. Repurposing the Data would be done by the CPS data collectors in the future if they wished to combine or restructure the data to add to its usefulness.

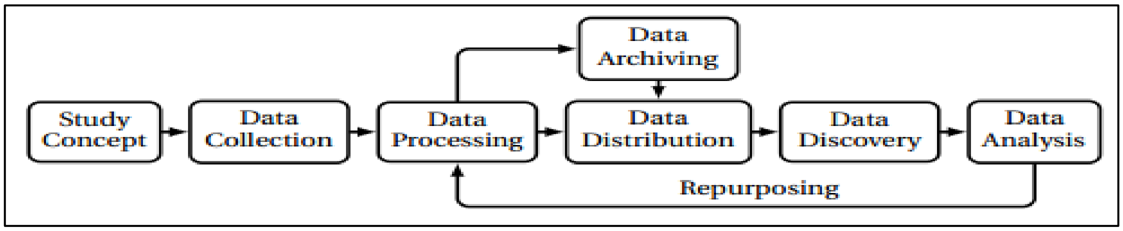


Figure 3: Data Lifecycle

3.3 Data Analysis Techniques

This study will report both descriptive and inferential statistics. Descriptive statistics describe and summarize data (Santucci, 2016). This can be done be measuring central tendencies, measuring variance from those central tendencies, and presenting histograms (Santucci, 2016). Inferential statistics involve employing more complex statistical methods in order to learn about the data, which usually involves a level of confidence (Santucci, 2016).

One of the inferential techniques used in this study was simple linear regression. In simple linear regression, the coefficients of the independent variables included in the regression indicate the increase in the dependent variable for a 1-unit increase in the independent variable. Because this study has a binary variable, *SMOKER*, as the dependent variable, the independent variable coefficients will reflect the increase or decrease in predicted probability of an individual for a 1-unit increase in the independent variable (Long & Freese, 2006). Because many of the independent variables used in the study are also binary variables, the coefficients show the change in predicted probability of smoking when one is assigned a value of 1 (i.e. a 1-unit increase) for that variable, holding all else constant.

The other inferential technique used in this study was binary logistic regression. This method is suitable for regression analysis for which the dependent variable is a binary variable. Instead of changes in predicted probability as given by the simple linear regression, the binary logistic regression will provide a set of relative odds of being a smoker versus the base case (Tranmer & Elliot, 2008).

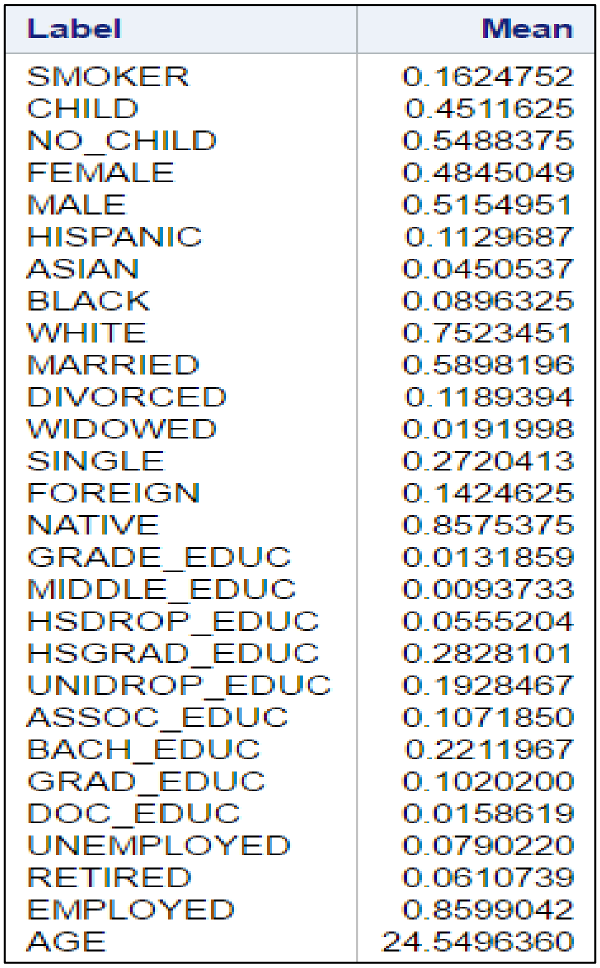
The base case is comprised of all the ‘left out’ variables for the binary variables, or the 0-case for continuous variables (such as *AGE*). In this report, the base case is an 18-year-old (*AGE*=0) male that is white, native-born, childless, single, and employed. The *HS_GRAD* variable will also be dropped out of regressions to avoid an error. Thus, the base case male referenced above will have a highest achieved education level as a high school degree. The base case will be referenced many times in the next section.

4 Findings

4.1 Summary Statistics

The mean (average) for each variable is shown in Figure 4. The base case variables have been added purely for this descriptive section and will be removed for regression analysis. The variation from the central tendencies are irrelevant for binary variables, and thus are not included in the table. For every binary variable, the mean is simply the percentage of observations that have a value of 1 expressed in decimal form. For example, a *SMOKER* mean of roughly 0.162 implies that approximately 16.2% of the sample in this report are classified as current smokers. By the same logic, a mean of 0.451 for *CHILD* implies that 45.1% of the sample have at least one child. Recall that an age of 0 implies one is 18. Therefore, a mean *AGE* of approximately 24.5 implies that the average age of individuals in the sample is roughly 42.5 years old.

The table shows a few points of interest: there are more men than women in the sample by a slim margin; whites represent over 75% of the sample; just under 59% of the respondents are married; over 85% of the sample was born in the US; the overwhelming majority are employed; and a high school degree (which is the base case for education) is the most common level of educational attainment. The next section will use regression methods to discover more about the data set.



Label	Mean
SMOKER	0.1624752
CHILD	0.4511625
NO_CHILD	0.5488375
FEMALE	0.4845049
MALE	0.5154951
HISPANIC	0.1129687
ASIAN	0.0450537
BLACK	0.0896325
WHITE	0.7523451
MARRIED	0.5898196
DIVORCED	0.1189394
WIDOWED	0.0191998
SINGLE	0.2720413
FOREIGN	0.1424625
NATIVE	0.8575375
GRADE_EDUC	0.0131859
MIDDLE_EDUC	0.0093733
HSDROP_EDUC	0.0555204
HSGRAD_EDUC	0.2828101
UNIDROP_EDUC	0.1928467
ASSOC_EDUC	0.1071850
BACH_EDUC	0.2211967
GRAD_EDUC	0.1020200
DOC_EDUC	0.0158619
UNEMPLOYED	0.0790220
RETIRED	0.0610739
EMPLOYED	0.8599042
AGE	24.5496360

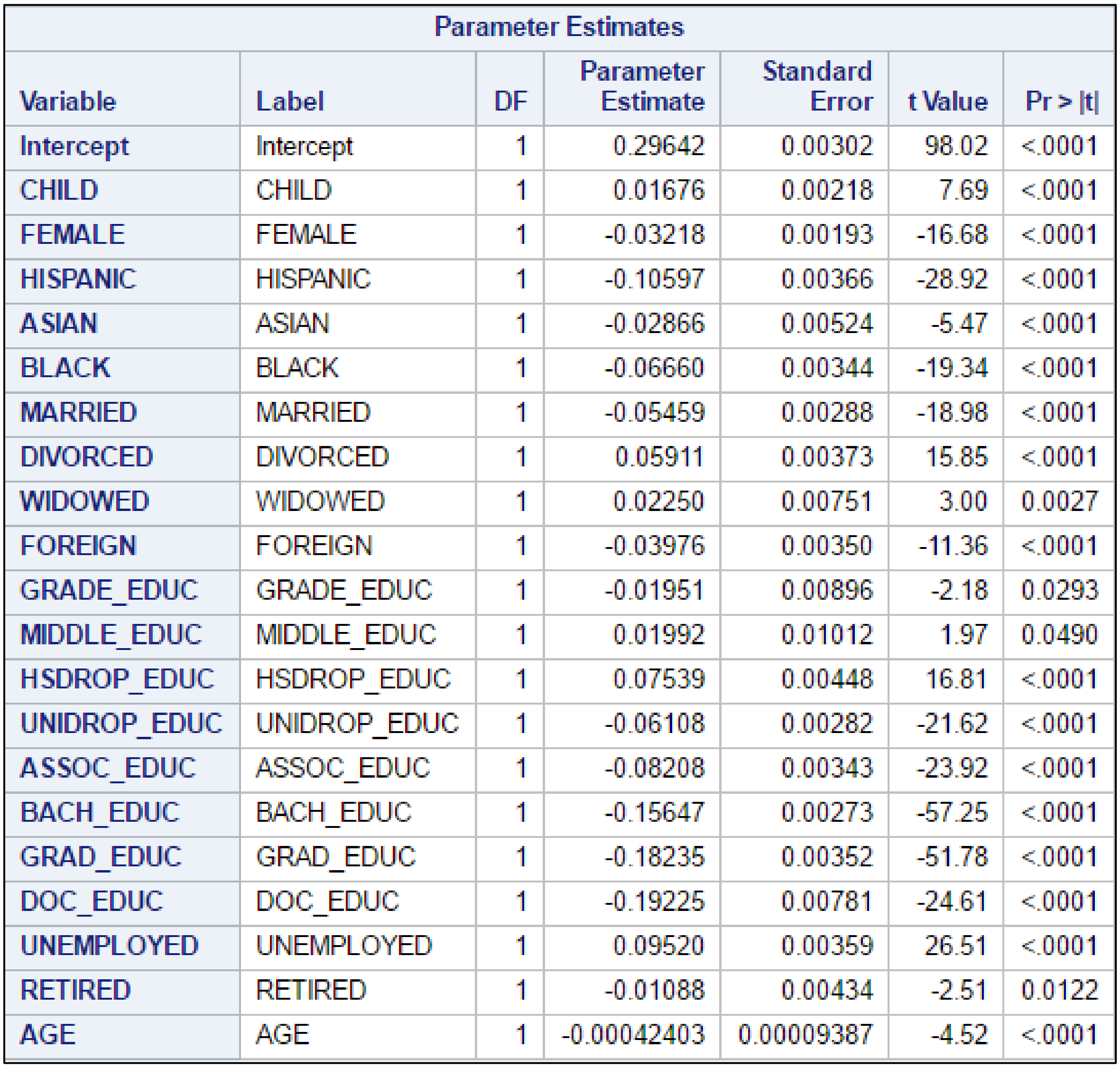
Figure 4: Mean of each variable

4.2 Inferential Statistics

As previously mentioned, two types of regression techniques will be used in this report. The following sub-sections will discuss the results of these statistical methods.

4.2.1 Simple Linear Regression

The results of simple linear regression are shown in Figure 5. Recall the base case is an 18-year-old (AGE=0) male that is white, native-born, childless, single, employed, and has a high school degree as his highest level of education. The intercept is the predicted probability of being a current smoker for this base case. Thus, the base case has a predicted probability of just under 0.30 of being a current smoker. This can be compared to the sample *SMOKER* mean of just 0.16. The binary independent variable parameter estimates represent the change in predicted probability when assigned a (1) in that category versus the base case, when all else is held constant. Therefore, negative coefficients indicate that having a value of (1) assigned for that category decreases the predicted probability of being a current smoker. For example, a *FEMALE* coefficient of approximately -0.032 indicates that being a female decreases the predicted probability of being a current smoker by 3.2 percentage points when compared to a male of the same race, age, education, etc. By the same logic, the probability of being a current smoker for high school drop outs is approximately 7.5 percentage points higher than their high school graduate counterparts. Likewise, Hispanics were found to have the lowest predicted probability of being a current smoker out of any race.



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.29642	0.00302	98.02	<.0001
CHILD	CHILD	1	0.01676	0.00218	7.69	<.0001
FEMALE	FEMALE	1	-0.03218	0.00193	-16.68	<.0001
HISPANIC	HISPANIC	1	-0.10597	0.00366	-28.92	<.0001
ASIAN	ASIAN	1	-0.02866	0.00524	-5.47	<.0001
BLACK	BLACK	1	-0.06660	0.00344	-19.34	<.0001
MARRIED	MARRIED	1	-0.05459	0.00288	-18.98	<.0001
DIVORCED	DIVORCED	1	0.05911	0.00373	15.85	<.0001
WIDOWED	WIDOWED	1	0.02250	0.00751	3.00	0.0027
FOREIGN	FOREIGN	1	-0.03976	0.00350	-11.36	<.0001
GRADE_EDUC	GRADE_EDUC	1	-0.01951	0.00896	-2.18	0.0293
MIDDLE_EDUC	MIDDLE_EDUC	1	0.01992	0.01012	1.97	0.0490
HSDROP_EDUC	HSDROP_EDUC	1	0.07539	0.00448	16.81	<.0001
UNIDROP_EDUC	UNIDROP_EDUC	1	-0.06108	0.00282	-21.62	<.0001
ASSOC_EDUC	ASSOC_EDUC	1	-0.08208	0.00343	-23.92	<.0001
BACH_EDUC	BACH_EDUC	1	-0.15647	0.00273	-57.25	<.0001
GRAD_EDUC	GRAD_EDUC	1	-0.18235	0.00352	-51.78	<.0001
DOC_EDUC	DOC_EDUC	1	-0.19225	0.00781	-24.61	<.0001
UNEMPLOYED	UNEMPLOYED	1	0.09520	0.00359	26.51	<.0001
RETIRED	RETIRED	1	-0.01088	0.00434	-2.51	0.0122
AGE	AGE	1	-0.00042403	0.00009387	-4.52	<.0001

Figure 5: Simple Linear Regression

Perhaps one of the most interesting results is that the predicted probability of smoker for those who have finished high school decreases with each additional level of education. For the *AGE* variable, the coefficient indicates the change in predicted probability of being a current smoker with a 1-year increase in age. A negative coefficient for this variable indicates that smoking rates decrease with age. Interestingly, every single variable used in the simple linear regression are statistically significant variables at a 95% confidence level, as measured by the p-values in the right-most column in the regression table above.

4.2.2 Binary Logistic Regression

Though a different technique, the binary logistic regression should confirm the findings found in the simple linear regression. A graphical result of this regression is shown in Figure 6.

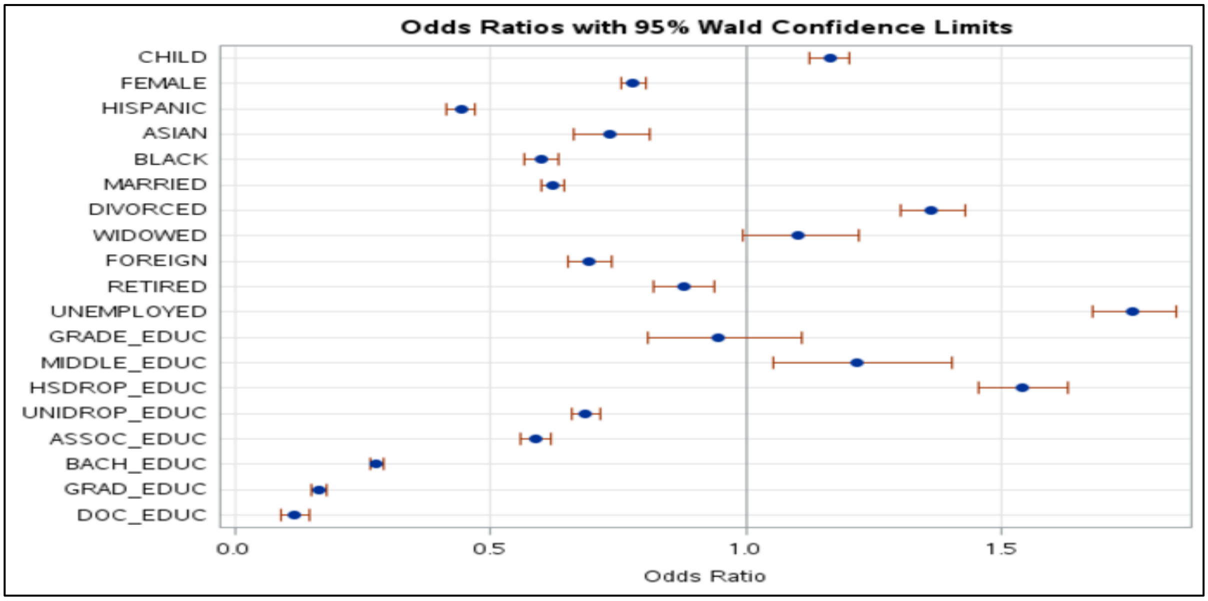


Figure 6: Binary Logistic Regression

The odds ratio line at 1.0 indicates identical odds of being a smoker versus the base case for each category. Therefore, variables which have an odds ratio that are placed to the left of the 1.0 odds ratio line have lower odds of being a current smoker versus the base case. The farther to the right of the 1.0 line, the more likely an individual identifying with that category is to be a smoker; the farther to the left, the less likely. The graph depicts some additional interesting findings: those with children are more likely to smoke than those without; both divorced and widowed individuals are more likely to smoke than single ones; high school drop outs have the highest odds of being a current smoker out of any education level; and the effect of education on smoking rates after high school graduation explained in the simple linear regression can be seen on this graph by the southwest-moving odds estimates after HSDROP_EDUC for remaining education levels - doctoral degree holders are over four times less likely to be smokers than high school drop outs.

5 Recommendations

There are a number of conclusions that policy makers can draw from these regression results, if their goal is the reduce smoking rates in the US. Given the increased odds of being a current smoker when at least one child is in the house hold, policy makers may wish to educate parents on the harmful effects of secondhand smoke on children, as well as the potential example the parents are setting. Additionally, with the knowledge that, based on this report's sample, whites are the most likely to smoke than any other race, anti-smoking campaigns may wish to be launched in areas that are primarily white. Friends and family of divorced and widowed individuals may wish to encourage the avoidance of tobacco use in loved ones stressed with such great change. Unemployment offices may also wish to launch anti-smoking campaigns.

Perhaps the largest area for intervention is in education. While the large majority of the variables used in this study are out of the control of the respondent (e.g. race, gender, etc.), the individual, to a certain extent, can control the amount of education they attain. Anti-smoking agencies may wish to team-up with anti-high school dropout campaigns if they want to most effectively minimize future smoking behavior.

6 Conclusion

This study attempted to model smoking behavior in the US using Current Population Survey data (CPS) from 2010 and 2011. An array of demographic and socioeconomic variables is used to explain smoking behavior of the surveyed individuals. The study found that those with children are more likely to smoke compared to those without children; females are less likely to smoke than males; Hispanics, blacks, and Asians are all less likely to smoke than whites; divorcees and widows are more likely to smoke than single individuals; married individuals are less likely to smoke than singles; retired individuals are less likely to smoke than working ones; unemployed individuals are more likely to smoke than working ones; and as education level increases after high school graduation, smoking rates decrease. Further studies may wish to use more current data once it becomes available and more advanced statistical techniques. Researchers may also wish to add additional explanatory variables, like residence or income.

7 References

Ball, A., 2012. Review of data management lifecycle models.

Brennan, A, Meier, P, Purshouse, R, Rafia, R, Meng, Y, & Hill-Macmanus, D 2016, 'Developing policy analytics for public health strategy and decisions-the Sheffield alcohol policy model framework', Annals Of Operations Research, 1, p. 149, Expanded Academic ASAP, EBSCOhost, viewed 2 May 2017.

Cho, H, Khang, Y, Jun, H, & Kawachi, I 2008, 'Marital status and smoking in Korea: the influence of gender and age', *Social Science & Medicine*, 3, p. 609, Expanded Academic ASAP, EBSCOhost, viewed 3 May 2017.

Chun, H, & Mobley, M 2010, 'Gender and grade-level comparisons in the structure of problem behaviors among adolescents', *Journal Of Adolescence*, 33, pp. 197-207, ScienceDirect, EBSCOhost, viewed 3 May 2017.

Davidson, AJ 2015, 'Creating Value: Unifying Silos into Public Health Business Intelligence', *Frontiers In Public Health Services & Systems Research*, 4, 2, p. 1, Publisher Provided Full Text Searching File, EBSCOhost, viewed 2 May 2017.

De Walque, D 2007, 'Does education affect smoking behaviors? *Journal Of Health Economics*, 26, pp. 877-895, ScienceDirect, EBSCOhost, viewed 3 May 2017.

Eddy, N 2015 ‘Analytics Investment among Health Care Organizations Grows’ http://www.eweek.com/it-management/analytics-investment-among-healthcare-organizations-grows.html; Accessed 2 May, 2017.

Elbashir, M, Collier, P, & Sutton, S 2011, 'The Role of Organizational Absorptive Capacity in Strategic Use of Business Intelligence to Support Integrated Management Control Systems', *Accounting Review*, 86, 1, pp. 155-184, Business Source Premier, EBSCOhost, viewed 2 May 2017.

Fagan, P., Augustson, E., Backinger, C.L., O’Connell, M.E., Vollinger Jr, R.E., Kaufman, A. and Gibson, J.T., 2007. Quit attempts and intention to quit cigarette smoking among young adults in the United States. *American Journal of Public Health*, 97(8), pp.1412-1420.

Fink, L, Yogev, N, & Even, A 2017, 'Business intelligence and organizational learning: An empirical investigation of value creation processes', *Information & Management*, 54, pp. 38-56, ScienceDirect, EBSCOhost, viewed 2 May 2017.

Flood, S, King, M, Ruggles, S, & Warren, JR, *Integrated Public Use Microdata Series, Current Population Survey: Version 4.0.* [dataset]. Minneapolis: University of Minnesota, 2015. http://doi.org/10.18128/D030.V4.0.

Halterman, J.S., Conn, K.M., Hernandez, T. and Tanski, S.E., 2010. ‘Parent knowledge, attitudes, and household practices regarding SHS exposure: a case-control study of urban children with and without asthma,’ *Clinical Pediatrics*, 49(8), pp.782-789.

Harris, K.M. and Edlund, M.J., 2005. Self-medication of mental health problems: New evidence from a national survey. *Health Services Research*, 40(1), pp.117-134.

Henkel, D., 2011, ‘Unemployment and substance use: a review of the literature (1990-2010)’. *Current drug abuse reviews*, 4(1), pp.4-27.

Hollingworth, W., Ebel, B.E., McCarty, C.A., Garrison, M.M., Christakis, D.A. and Rivara, F.P., 2006. Prevention of deaths from harmful drinking in the United States: the potential effects of tax increases and advertising bans on young drinkers. *Journal of Studies on Alcohol*, 67(2), pp.300-308.

Jourdan, Z, Rainer, R, & Marshall, T 2008, 'Business Intelligence: An Analysis of the Literature', *Information Systems Management*, 25, 2, pp. 121-131, Business Source Premier, EBSCOhost, viewed 2 May 2017.

Kohli, R, & Tan, S 2016, 'ELECTRONIC HEALTH RECORDS: HOW CAN IS RESEARCHERS CONTRIBUTE TO TRANSFORMING HEALTHCARE?', *MIS Quarterly*, 40, 3, pp. 553-574, Business Source Premier, EBSCOhost, viewed 2 May 2017.

Long, J.S. & Freese, J., 2006. ‘Regression models for categorical dependent variables using Stata’, *Stata press*.

Lyke, R 2009. ‘Healthcare Reform: An Introduction’ CRS Report for Congress. *Congressional Research Service*, 7-5700, R40517 Available from: <http://fpc.state.gov/documents/organization/126771.pdf> [Accessed 2 May 2017].

Mills, A.L., White, M.M., Pierce, J.P. and Messer, K., 2011. Home smoking bans among US households with children and smokers: opportunities for intervention. *American Journal of Preventive Medicine*, 41(6), pp.559-565.

OECD, 2006, ‘Young Drivers: The Road to Safety’, Organization for Economic Co-Operation and Development, Available from: <http://www.oecd.org/itf/37556934.pdf> [Accessed 3 May 2017].

Pennanen, M., Broms, U., Korhonen, T., Haukkala, A., Partonen, T., Tuulio-Henriksson, A., Laatikainen, T., Patja, K. and Kaprio, J., 2014. ‘Smoking, nicotine dependence and nicotine intake by socio-economic status and marital status’. *Addictive Behaviors*, 39(7), pp.1145-1151.

Ruhm, C.J., 2000. ‘Are recessions good for your health?’ *The Quarterly Journal of Economics*, 115(2), pp.617-650.

Santucci, AC 2016, 'Data description in research', *Salem Press Encyclopedia of Health*, EBSCOhost, viewed 4 May 2017.

Schou, L, Storvoll, E, & Moan, I 2014, 'Alcohol-related sickness absence among young employees: Gender differences and the prevention paradox', *European Journal Of Public Health*, 24, 3, pp. 480-485, CINAHL Complete, EBSCOhost, viewed 3 May 2017.

Soulakova, J, Li, J, & Crockett, L 2017, 'Race/ethnicity and intention to quit cigarette smoking', *Preventive Medicine Reports*, *Vol 5, Iss C*, Pp 160-165 (2017), C, p. 160, Directory of Open Access Journals, EBSCOhost, viewed 3 May 2017.

Steward, M 2005, 'Electronic Medical Records', *Journal Of Legal Medicine*, 26, 4, pp. 491-506, Academic Search Complete, EBSCOhost, viewed 2 May 2017.

Sullivan, LW, 1986, ‘SMOKING AND HEALTH: A National Status Report 2nd Edition – A Report to Congress’ (PDF). nlm.nih.gov. U.S. Department of Health and Human Services. Retrieved 4 May 2017. Available from < https://profiles.nlm.nih.gov/ps/access/NNBBVP.pdf>

Syamlal, G, Mazurek, J, & Dube, S 2014, 'Brief Report: Gender Differences in Smoking Among U.S. Working Adults', *American Journal Of Preventive Medicine*, 47, pp. 467-475, ScienceDirect, EBSCOhost, viewed 3 May 2017.

Thayer, C, Bruno, J, and Remorenko, MB 2013, “Using Data Analytics to Identify Revenue at Risk,” *Healthcare Financial Management* (67:9), pp. 72-78, 80.

Tranmer, M. & Elliot, M., 2008, Binary Logistic Regression. Cathie Marsh Centre for Census and Survey Research. Available from: <http://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/working-papers/2008/2008-20-binary-logistic-regression.pdf> [Accessed 4 May 2017].

Trieu, V 2017, 'Getting value from Business Intelligence systems: A review and research agenda', *Decision Support Systems*, 93, pp. 111-124, ScienceDirect, EBSCOhost, viewed 2 May 2017.

US Department of Health and Human Services, 2004. The health consequences of smoking: a report of the Surgeon General.