**A Study of Causal Effects that Drives Students to use Cigarettes**

By Parker Barouch

December 11, 2017

Applied Econometrics; ECON 5670

**I Introduction:**

The question of this paper is to answer what are the causes that bring children and/or teenagers to smoke. In particular, what are the exact causes or stressors that bring school children/teens to smoke? Smoking has been tied to many health risks and can cause economic hardship (a more thorough explanation of the harm it can do will be in the next section) so this is a very important question regarding every child in the US. With the answering of this question schools will know which factors hinder or encourage their students from smoking. A school armed with this knowledge can readjust these factors in order to give the smallest probability of their students from taking up smoking or find those that are most at risk and curb the habit from ever starting. The information gained from this study is not just limited to the school level, but can also be used by parents or used by the federal government to enact more effective policy to curb the number of children smoking in schools.

**II Non-Econometric Background:**

This section will explain the dangers of smoking and general information about cigarettes. If the reader is already well versed in this subject they may skip this section. The first commercial cigarette was made in 1865 by Washington Duke on his farm in Raleigh, North Carolina. A cigarette, or cigaret, is made from finely cut tobacco leaves rolled in thin paper for smoking. The cigarette is ignited on one end, causing the cigarette and the tobacco that is left to smolder. This allows the smoke from the cigarette to enter you mouths and lungs where it is then inhaled out. Modern cigarettes are commonly white with a filter placed on one end (the yellow portion) as this can be seen in picture 1.

[](https://www.google.com/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&cad=rja&uact=8&ved=0ahUKEwiry8mg5ZvXAhWriFQKHR87BbgQjRwIBw&url=http://www.istockphoto.com/photos/cigarette&psig=AOvVaw1NZdLnlrV6F8MLV64OJ1h-&ust=1509571213687072)Picture 1

In the cigarette, tobacco contains many chemicals. In particular, cigarettes contain a chemical called nicotine that is a very addictive psychoactive chemical. This chemical makes it very hard for consumers of cigarettes to quit smoking. In effect, the smoker of these cigarettes also inhales a host of other chemicals (about 4,000) of which, over 50 are carcinogenic. About half of cigarette smokers die from tobacco-related disease and on average died 10 years younger[[1]](#footnote-1). The continued use of cigarettes also have many side effect including, but not limited to, shortness of breath, loss of sense of smell or taste, grey appearance, damage to reproduction organs, and heart disease. I hope the reader is understanding there are many downsides to cigarettes and I have not even gone into the expenses of cigarettes which can take up a large portion of an avid smoker budget.

**III Data Description:**

The data that will be used in this analysis comes from The Center of Disease Control and Prevention (CDC) National Youth Tobacco Survey (NYTS). The survey is a national survey conducted in the United States for middle and high school youth’s. The survey ask a variety of questions including tobacco-related beliefs, attitudes, behavior, and exposure to pro- and anti-tobacco influences. The NYTS serves as a baseline for comparing progress towards meeting *Healthy People* *2020* goals for reducing tobacco use among youths. The survey data was collected in 2009 for this analysis and was conducted by choosing 222 randomly sampled schools of which 205 participated. There were 24,266 randomly selected students, of which 22,679 participated, giving the survey an 84.8 percent response rate. The survey may not be representative of all youth in the United States since it was given in schools and does not consider the youth who have dropped out of school and who are more likely to be smoking[[2]](#footnote-2).

The Dependent variable I will be using for this analysis is Cig\_ind. Cig\_ind (standing for cigarette indicator) is defined as one if the surveyed person has once in their life smoked a whole cigarette, otherwise if they have not smoked or only had a couple puffs of a cigarette the variable will be zero. The data is cross-sectional meaning we only have one time-period (2009) and we will only see that entity once in the data. The dependent variable in this analysis has a unique feature to it. Such that, the variable may only take on a value of zero or one. This variable comes up a lot in econometrics studies so we have given it a few special name such as a binary or dummy variable. The next step in the analysis is to check the frequency of the binary variable. Ideally, we would like there to be an even split such that 50% of the observed dependent variables are zero and the other 50% are ones.

|  |  |  |
| --- | --- | --- |
| Table 1: Frequency Table of Cig\_ind | | |
| Cig\_ind | Frequency | Percent |
| 0 | 17579 | 77.51% |
| 1 | 5100 | 22.49% |

The information in Table one tells us that 77.51% of the dependent variable is made up of zero’s and 22.49% are made up of one’s. There is a rule of thumb in econometrics for binary dependent variables that at least 20% of the dependent variable are zeros or ones. The reason for this rule is so there is sufficient variation for the model to measure. You may think of this as a sort of potency like measuring Kool-Aid mix to put in your drink to enjoy on a hot summer day. If you do not put enough Kool-Aid mix in your drink the water will drown out the taste of the mix. The same thing can happen in the model that there is not enough zeros or ones the model will be drowned out by the other variable and will not provide good predictions.

Another interesting item to look at is the summary statistics of the independent variables conditioned on whether that student has smoked or not smoked. Though, before I show this it would be helpful to know what the variables are. Therefore, I will show in the table below a data dictionary which provides all of the definition of each variable you will see in this paper.

| Table 2: Data Dictionary (Continued) | | |
| --- | --- | --- |
| ID | Variable | Definition |
| 1 | Agei | The current age of student "i" |
| 2 | Malei | If student "i" is a male returns 1 else returns a 0 |
| 3 | Hispanici | If student "i" is Hispanic return 1 else returns a 0 |
| 4 | AIndiani | If student "i" is Native American or Alaskan Native return 1 else returns a 0 |
| 5 | Asiani | If student "i" is Asian return 1 else returns a 0 |
| 6 | Blacki | If student "i" is Black return 1 else returns a 0 |
| 7 | Hawaiiani | If student "i" is Hawaiian return 1 else returns a 0 |
| 8 | Whitei | If student "i" is White return 1 else returns a 0 |
| 9 | Missi | The number of days student "i" has missed of school in the past 30 days |
| 10 | Cig\_indi | If student "i" has smoke a whole cigarette once in their life returns 1 else returns 0 |
| 11 | Otobi | If student "i" has used chewing tobacco, snuff, or dip once in their life returns 1 else returns 0 |
| 12 | Age\_otobi | The age of student "i" when they first tried chewing tobacco, snuff, or dip |
| 13 | Use\_otobi | The number of times student "i" used chewing tobacco, snuff, or dip in the past 30 days |
| 14 | Cigari | If student "i" has smoked a cigar, cigarillos, or little cigar returns 1 else returns 0 |
| 15 | Day\_cigari | The number of times student "i" has smoke a cigar, cigarillos, or little cigar in the past 30 days |
| 16 | Pipei | If student "i" has smoked tobacco in a pipe returns 1 else returns 0 |
| 17 | Day\_pipei | The number of times student "i" has smoke tobacco in a pipe in the past 30 days |
| 18 | Roomi | The number of times student "i" has been the same room with someone else who was smoking in the past seven days |
| 19 | Cari | The number of times student "i" has been in the car with someone else who was smoking in the past seven days |
| 20 | Harmfuli | If student "i" believes smoking is harmful returns 1 else returns 0 |
| 21 | Pharmfuli | If student "i" believes smoking is probably harmful returns 1 else returns 0 |
| 22 | Pnharmfuli | If student "i" believes smoking is probably not harmful returns 1 else returns 0 |
| 23 | Nharmfuli | If student "i" believes smoking is probably not harmful returns 1 else returns 0 |
| 24 | Live\_smokei | If student "i" lives with someone in their household that smokes return 1 else return 0 |
| 25 | Live\_otobi | If student "i" lives with someone in their household that uses chewing tobacco, snuff, or dip return 1 else return 0 |
| 26 | Friends\_ti | The number of student "i" four closest friends that smoke |
| 27 | Friends\_unsure\_ti | If student "i" does not know if their four closest friends smoke returns 1 else returns 0 |
| 28 | Friends\_otobi | The number of student "i" four closest friends that use chewing tobacco, snuff, or dip |
| 29 | Friends\_unsure\_oti | If student "i" does not know if their four closest friends use chewing tobacco, snuff, or dip returns 1 else returns 0 |
| 30 | Friends\_usei | Friends\_ti + Friends\_otobi |
| 31 | Rule\_noi | If smoking in student "i" house is not allowed anywhere inside of the home return 1 else return 0 |
| 32 | Rule\_si | If smoking in student "i" house is allowed inside some places of the home return 1 else return 0 |
| 33 | Rule\_ai | If smoking in student "i" house is allowed inside anywhere in home return 1 else return 0 |
| 34 | Rule\_nai | If smoking in student "i" house is allowed anywhere in the home return 1 else return 0 |
| 35 | Rule\_Mi | If student "i" did not fill anything out for the four above variables return 1 else return 0 |
| 36 | Dy\_smoke\_yi | If student "i" says they will definitely smoke a cigarette anytime during the next year returns 1 else returns 0 |
| 37 | Py\_smoke\_yi | If student "i" says they will probably smoke a cigarette anytime during the next year returns 1 else returns 0 |
| 38 | Pn\_smoke\_yi | If student "i" says they will probably not smoke a cigarette anytime during the next year returns 1 else returns 0 |
| 39 | Dn\_smoke\_yi | If student "i" says they will definitely not smoke a cigarette anytime during the next year returns 1 else returns 0 |
| 40 | Na\_smoke\_yi | If student "i" does not answer the four above variables returns 1 else returns 0 |
| 41 | Dy\_smoke\_bfi | If student "i" was offered a cigarette by one of their best friends and would definitely take it return 1 else return 0 |
| 42 | Py\_smoke\_bfi | If student "i" was offered a cigarette by one of their best friends and would probably take it return 1 else return 0 |
| 43 | Pn\_smoke\_bfi | If student "i" was offered a cigarette by one of their best friends and would probably not take it return 1 else return 0 |
| 44 | Dn\_smoke\_bfi | If student "i" was offered a cigarette by one of their best friends and would definitely not take it return 1 else return 0 |
| 45 | Na\_smoke\_bfi | If student "i" left the above four variables blank return 1 else return 0 |
| 46 | Par\_warni | If student "i" was told not to smoke by their parent (or guardian) return 1 else return 0 |
| 47 | Smoke\_cooli | If student "i" believes smoking makes young people look cool or fit in return 1 else return 0 |
| 48 | Smoke\_cool\_mi | If student "i" left the above variable blank return 1 else return 0 |
| 49 | Say\_no\_mi | If student "i" left this question blank or said they were unsure if they practiced saying no to tobacco use in school return 1 else return 0 |
| 50 | Say\_noi | If student "i" practice ways of saying no to tobacco use in class return 1 else return 0 |
| 51 | Taught\_smoke\_mi | If student "i" was taught of the danger of tobacco use in school missing or not sure return 1 else return 0 |
| 52 | Taught\_smoke\_badi | If student "i" was taught of the danger of tobacco use in school return 1 else return 0 |
|  |  |  |

Now that we clearly know what every variable stands for I will present the summary statistics unconditioned and conditioned on the dependent variable.

| Table 3: Summary statistics of independent variables, where Cig\_ind = 0 (Continued) | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Number of observations | Mean | Standard Deviation | Minimum | Maximum |
| Age | 17,579 | 14.24 | 2.07 | 9 | 21 |
| Male | 17,579 | 0.48 | 0.50 | 0 | 1 |
| Hispanic | 17,579 | 0.27 | 0.44 | 0 | 1 |
| AIndian | 17,579 | 0.06 | 0.24 | 0 | 1 |
| Asian | 17,579 | 0.06 | 0.24 | 0 | 1 |
| Black | 17,579 | 0.21 | 0.41 | 0 | 1 |
| Hawaiian | 17,579 | 0.03 | 0.17 | 0 | 1 |
| White | 17,579 | 0.54 | 0.50 | 0 | 1 |
| Miss | 17,579 | 1.22 | 2.09 | 0 | 11 |
| Otob | 17,579 | 0.04 | 0.20 | 0 | 1 |
| Age\_otob | 17,579 | 0.49 | 2.52 | 0 | 17 |
| Use\_otob | 17,579 | 0.76 | 4.62 | 0 | 30 |
| Cigar | 17,579 | 0.08 | 0.27 | 0 | 1 |
| Day\_cigar | 17,579 | 0.11 | 1.34 | 0 | 30 |
| Pipe | 17,579 | 0.95 | 0.22 | 0 | 1 |
| Day\_pipe | 17,579 | 0.06 | 1.16 | 0 | 30 |
| Room | 17,579 | 0.65 | 1.67 | 0 | 7 |
| Car | 17,579 | 0.49 | 1.11 | 0 | 5 |
| Harmful | 17,579 | 0.71 | 0.45 | 0 | 1 |
| Pharmful | 17,579 | 0.19 | 0.39 | 0 | 1 |
| Pnharmful | 17,579 | 0.04 | 0.19 | 0 | 1 |
| Nharmful | 17,579 | 0.04 | 0.19 | 0 | 1 |
| Live\_smoke | 17,579 | 0.31 | 0.46 | 0 | 1 |
| Live\_otob | 17,579 | 0.08 | 0.27 | 0 | 1 |
| Friends\_t | 17,579 | 0.37 | 0.87 | 0 | 4 |
| Friends\_unsure\_t | 17,579 | 0.13 | 0.34 | 0 | 1 |
| Friends\_otob | 17,579 | 0.16 | 0.62 | 0 | 4 |
| Friends\_unsure\_ot | 17,579 | 0.11 | 0.31 | 0 | 1 |
| Friends\_use | 17,579 | 0.53 | 1.24 | 0 | 8 |
| Rule\_no | 17,579 | 0.73 | 0.44 | 0 | 1 |
| Rule\_s | 17,579 | 0.08 | 0.26 | 0 | 1 |
| Rule\_a | 17,579 | 0.03 | 0.17 | 0 | 1 |
| Rule\_na | 17,579 | 0.12 | 0.33 | 0 | 1 |
| Rule\_M | 17,579 | 0.03 | 0.18 | 0 | 1 |
| Dy\_smoke\_y | 17,579 | 0.01 | 0.11 | 0 | 1 |
| Py\_smoke\_y | 17,579 | 0.04 | 0.19 | 0 | 1 |
| Pn\_smoke\_y | 17,579 | 0.14 | 0.35 | 0 | 1 |
| Dn\_smoke\_y | 17,579 | 0.77 | 0.42 | 0 | 1 |
| Na\_smoke\_y | 17,579 | 0.03 | 0.18 | 0 | 1 |
| Dy\_smoke\_bf | 17,579 | 0.01 | 0.10 | 0 | 1 |
| Py\_smoke\_bf | 17,579 | 0.04 | 0.19 | 0 | 1 |
| Pn\_smoke\_bf | 17,579 | 0.15 | 0.36 | 0 | 1 |
| Dn\_smoke\_bf | 17,579 | 0.78 | 0.41 | 0 | 1 |
| Na\_smoke\_bf | 17,579 | 0.02 | 0.14 | 0 | 1 |
| Par\_warn | 17,579 | 0.57 | 0.49 | 0 | 1 |
| Smoke\_cool | 17,579 | 0.91 | 0.28 | 0 | 1 |
| Smoke\_cool\_m | 17,579 | 0.02 | 0.15 | 0 | 1 |
| Say\_no\_m | 17,579 | 0.19 | 0.39 | 0 | 1 |
| Say\_no | 17,579 | 0.31 | 0.46 | 0 | 1 |
| Taught\_smoke\_m | 17,579 | 0.15 | 0.36 | 0 | 1 |
| Taught\_smoke\_bad | 17,579 | 0.37 | 0.48 | 0 | 1 |
|  |  |  |  |  |  |

| Table 4: Summary statistics of independent variables, where Cig\_ind = 1 (Cont.) | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | Number of observations | Mean | Standard Deviation | Minimum | | Maximum | |
| Age | 5,100 | 15.79 | 1.82 | 9 | | 21 | |
| Male | 5,100 | 0.54 | 0.50 | 0 | | 1 | |
| Hispanic | 5,100 | 0.30 | 0.46 | 0 | | 1 | |
| AIndian | 5,100 | 0.07 | 0.26 | 0 | | 1 | |
| Asian | 5,100 | 0.04 | 0.19 | 0 | | 1 | |
| Black | 5,100 | 0.17 | 0.37 | 0 | | 1 | |
| Hawaiian | 5,100 | 0.03 | 0.17 | 0 | | 1 | |
| White | 5,100 | 0.61 | 0.49 | 0 | | 1 | |
| Miss | 5,100 | 1.96 | 2.65 | 0 | | 11 | |
| Otob | 5,100 | 0.30 | 0.46 | 0 | | 1 | |
| Age\_otob | 5,100 | 3.76 | 6.08 | 0 | | 17 | |
| Use\_otob | 5,100 | 3.13 | 8.61 | 0 | | 30 | |
| Cigar | 5,100 | 0.62 | 0.49 | 0 | | 1 | |
| Day\_cigar | 5,100 | 1.72 | 5.38 | 0 | | 30 | |
| Pipe | 5,100 | 0.70 | 0.46 | 0 | | 1 | |
| Day\_pipe | 5,100 | 0.94 | 4.51 | 0 | | 30 | |
| Room | 5,100 | 2.40 | 2.74 | 0 | | 7 | |
| Car | 5,100 | 0.95 | 1.40 | 0 | | 5 | |
| Harmful | 5,100 | 0.48 | 0.50 | 0 | | 1 | |
| Pharmful | 5,100 | 0.33 | 0.47 | 0 | | 1 | |
| Pnharmful | 5,100 | 0.10 | 0.30 | 0 | | 1 | |
| Nharmful | 5,100 | 0.07 | 0.25 | 0 | | 1 | |
| Live\_smoke | 5,100 | 0.51 | 0.50 | 0 | | 1 | |
| Live\_otob | 5,100 | 0.17 | 0.38 | 0 | | 1 | |
| Friends\_t | 5,100 | 1.70 | 1.56 | 0 | | 4 | |
| Friends\_unsure\_t | 5,100 | 0.16 | 0.37 | 0 | | 1 | |
| Friends\_otob | 5,100 | 0.65 | 1.19 | 0 | | 4 | |
| Friends\_unsure\_ot | 5,100 | 0.11 | 0.32 | 0 | | 1 | |
| Friends\_use | 5,100 | 2.35 | 2.24 | 0 | | 8 | |
| Rule\_no | 5,100 | 0.61 | 0.49 | 0 | | 1 | |
| Rule\_s | 5,100 | 0.13 | 0.33 | 0 | | 1 | |
| Rule\_a | 5,100 | 0.08 | 0.27 | 0 | | 1 | |
| Rule\_na | 5,100 | 0.15 | 0.36 | 0 | | 1 | |
| Rule\_M | 5,100 | 0.03 | 0.16 | 0 | | 1 | |
| Dy\_smoke\_y | 5,100 | 0.28 | 0.45 | 0 | | 1 | |
| Py\_smoke\_y | 5,100 | 0.32 | 0.47 | 0 | | 1 | |
| Pn\_smoke\_y | 5,100 | 0.18 | 0.38 | 0 | | 1 | |
| Dn\_smoke\_y | 5,100 | 0.18 | 0.38 | 0 | | 1 | |
| Na\_smoke\_y | 5,100 | 0.03 | 0.18 | 0 | | 1 | |
| Dy\_smoke\_bf | 5,100 | 0.24 | 0.43 | 0 | | 1 | |
| Py\_smoke\_bf | 5,100 | 0.33 | 0.47 | 0 | | 1 | |
| Pn\_smoke\_bf | 5,100 | 0.22 | 0.41 | 0 | | 1 | |
| Dn\_smoke\_bf | 5,100 | 0.19 | 0.39 | 0 | | 1 | |
| Na\_smoke\_bf | 5,100 | 0.02 | 0.15 | 0 | | 1 | |
| Par\_warn | 5,100 | 0.57 | 0.49 | 0 | | 1 | |
| Smoke\_cool | 5,100 | 0.79 | 0.41 | 0 | | 1 | |
| Smoke\_cool\_m | 5,100 | 0.03 | 0.17 | 0 | | 1 | |
| Say\_no\_m | 5,100 | 0.17 | 0.38 | 0 | | 1 | |
| Say\_no | 5,100 | 0.22 | 0.41 | 0 | | 1 | |
| Taught\_smoke\_m | 5,100 | 0.15 | 0.36 | 0 | | 1 | |
| Taught\_smoke\_bad | 5,100 | 0.49 | 0.50 | 0 | | 1 | |
|  |  |  |  | |  | |  | |

Now that I have shown the summary statistics conditioned on the dependent variable and all of the statistics make sense. The next important part is to predict the signs that we would expect by running an analysis.

| Table 5: Expected Sign and Explanation (Continued) | | | |
| --- | --- | --- | --- |
| ID | Variable | Expected Sign | Explanation |
| 1 | Age | -/+ | This will probably have a quadratic shape where when they are about to hit critical years for smoking a lot of students will begin smoking, but it will decrease as they get older. |
| 2 | Male | ? | Research has shown that male and female students at this age are smoking at equal rates so I cannot give an educated guess at this time. |
| 3 | Hispanic | ? | Not enough information is known to give an educated guess on what sign this variable will be. |
| 4 | AIndian | ? | Not enough information is known to give an educated guess on what sign this variable will be. |
| 5 | Asian | ? | Not enough information is known to give an educated guess on what sign this variable will be. |
| 6 | Black | ? | Not enough information is known to give an educated guess on what sign this variable will be. |
| 7 | Hawaiian | ? | Not enough information is known to give an educated guess on what sign this variable will be. |
| 8 | White | ? | Not enough information is known to give an educated guess on what sign this variable will be. |
| 9 | Miss | + | The more a student misses of class will send a signal that they care less about school and maybe is a "rebel" which means they are more likely to do illegal things like smoking. |
| 10 | Otob | + | If a student uses other tobacco products they will probably will be smoking too. |
| 11 | Age\_otob | + | If the student has used other tobacco products that student is probably more inclined to smoke. |
| 12 | Use\_otob | + | The more a student uses other tobacco products the more likely that student is to be smoking too. |
| 13 | Cigar | + | The more a student smokes a cigar another tobacco product they probably will smoke too. |
| 14 | Day\_cigar | + | The more a student smoke cigar probably indicates they are more likely to be smoking too. |
| 15 | Pipe | + | The more a student smokes a pipe another tobacco product they probably will smoke too. |
| 16 | Day\_pipe | + | The more a student smoke pipe probably indicates they are more likely to be smoking too. |
| 17 | Room | + | The more times a student is in the room with someone else smoking the more desensitized to it they will become and the more likely they are to be smoking too. |
| 18 | Car | + | The more times a student is in the car with someone else smoking the more desensitized to it they will become and the more likely they are to be smoking too. |
| 19 | Harmful | - | If the student believes smoking to be harmful the less likely they are to be smoking. |
| 20 | Pharmful | - | If the student believes smoking to be probably harmful the less likely they are to be smoking. |
| 21 | Pnharmful | + | If the student believes smoking to be probably not harmful the more likely they are to be smoking. |
| 22 | Nharmful | + | If the student believes smoking to be not harmful the more likely they are to be smoking. |
| 23 | Live\_smoke | + | If the student live with someone that is smoking the more desensitized to it they will become and the more likely they are of smoking too. |
| 24 | Live\_otob | + | If the student lives with someone who is using other tobacco products will be more desensitized to it and will be more likely to be smoking. |
| 25 | Friends\_t | + | The more friends of the student that is smoking the more likely that student is smoking too. |
| 26 | Friends\_unsure\_t | + | If the student is unsure if their friends are smoking they probably are not very close to them meaning they are more of a loaner so might take the habit of smoking to compensate for loneliness. |
| 27 | Friends\_otob | + | The more friends of the student that are using other tobacco products the more likely that student is smoking too. |
| 28 | Friends\_unsure\_ot | + | If the student is unsure if their friends are using other tobacco products they probably not very close to them meaning they are more of a loaner so might take the habit of smoking to compensate for loneliness. |
| 29 | Friend\_unsure | + | If more of your friends use other tobacco or tobacco products the more likely you are to. |
| 30 | Rule\_no | - | If there is a strict rule in their house the parent(s) are stricter and probably do not allow their children to smoke. |
| 31 | Rule\_s | + | If there is some smoking in the house that is allowed the student has a role model that probably smokes and therefore is more inclined to smoke. |
| 32 | Rule\_a | + | If there is smoking in the house that is allowed the student has a role model that probably smokes and therefore is more inclined to smoke. |
| 33 | Rule\_na | - | If there is no set rule for smoking in the house no one probably smokes so the student would not either. |
| 34 | Rule\_M | - | If the parents have not bothered to set a rule there probably is a mutual understanding of not smoking in the house so the student would not smoke either. |
| 35 | Dy\_smoke\_y | + | If the student believes they will definitely be smoking in a year they probably already smoking. |
| 36 | Py\_smoke\_y | + | If the student believes they will probably be smoking in a year they probably already smoking. |
| 37 | Pn\_smoke\_y | - | If the student believes they will probably not be smoking in a year they probably not already smoking. |
| 38 | Dn\_smoke\_y | - | If the student believes they will definitely not be smoking in a year they probably not already smoking. |
| 39 | Na\_smoke\_y | - | If the student skipped this problem they may be planning on smoking relatively soon and don’t want to get in trouble so they probably are not smoking now but plan to do so in the relative future. |
| 40 | Dy\_smoke\_bf | + | If the student would smoke a cigarette if given by friend they probably already smoking. |
| 41 | Py\_smoke\_bf | + | If the student would probably smoke a cigarette if given by friend they probably already smoking. |
| 42 | Pn\_smoke\_bf | - | If the student would probably not smoke a cigarette if given by friend they probably are not smoking. |
| 43 | Dn\_smoke\_bf | - | If the student would not smoke a cigarette if given by friend they probably not smoking. |
| 44 | Na\_smoke\_bf | + | If the student left this question blank they probably don’t want to get any friend in trouble who are smoking which may mean they are probably smoking too. |
| 45 | Par\_warn | - | If the parents have warned the student not to smoke they probably have learned of the dangers of smoking and will therefore be less inclined to smoke. |
| 46 | Smoke\_cool | + | If the student sees other peers smoking and think that is cool they probably are more inclined to be smoking as well. |
| 47 | Smoke\_cool\_m | - | If the student left this question blank they probably have not given this subject any thought and probably are not smoking themselves. |
| 48 | Say\_no\_m | + | If the student leaves this blank r does not know if they have been trained to say no to smoking they probably do not know the dangers of smoking and therefore are more inclined to smoke. |
| 49 | Say\_no | - | If the student was taught to say no to smoking they probably will be more easily able to not smoke if inclined by friends. |
| 50 | Taught\_smoke\_m | + | If the student does not remember or does not know if they have been taught how to say no to smoking they can be more easily persuaded by friend to smoke and therefore more inclined to smoke themselves. |
| 51 | Taught\_smoke\_bad | - | If the student was taught that smoking is bad they are more inclined to not be smoking. |

This process is done not just as an exercise, but as a sort of error checker. If I were to present a model of revenue for a luxurious steak house and said an increase in consumer income decreased the steak house’s revenue. This would be a large red flag of the validity of the analysis. This prediction is not logical and need to be reexamined. The two main reason for this is that there is something wrong with the logic that we are using and/or that there is a problem with the model. If there is a problem with the logic or model this will be shown through illogical coefficients for the model and further examination is needed.

| Table 6: Summary statistics of independent variables (Cont.) | | | | | |
| --- | --- | --- | --- | --- | --- |
| Variable | Number of observations | Mean | Standard Deviation | Minimum | Maximum |
| Age | 22,679 | 14.59 | 2.12 | 9 | 21 |
| Male | 22,679 | 0.50 | 0.50 | 0 | 1 |
| Hispanic | 22,679 | 0.28 | 0.45 | 0 | 1 |
| AIndian | 22,679 | 0.06 | 0.25 | 0 | 1 |
| Asian | 22,679 | 0.06 | 0.23 | 0 | 1 |
| Black | 22,679 | 0.20 | 0.40 | 0 | 1 |
| Hawaiian | 22,679 | 0.03 | 0.17 | 0 | 1 |
| White | 22,679 | 0.56 | 0.50 | 0 | 1 |
| Miss | 22,679 | 1.38 | 2.25 | 0 | 11 |
| Cig\_ind | 22,679 | 0.22 | 0.42 | 0 | 1 |
| Otob | 22,679 | 0.10 | 0.30 | 0 | 1 |
| Age\_otob | 22,679 | 1.22 | 3.89 | 0 | 17 |
| Use\_otob | 22,679 | 1.30 | 5.85 | 0 | 30 |
| Cigar | 22,679 | 0.20 | 0.40 | 0 | 1 |
| Day\_cigar | 22,679 | 0.47 | 2.89 | 0 | 30 |
| Pipe | 22,679 | 0.89 | 0.31 | 0 | 1 |
| Day\_pipe | 22,679 | 0.26 | 2.40 | 0 | 30 |
| Room | 22,679 | 1.05 | 2.09 | 0 | 7 |
| Car | 22,679 | 0.59 | 1.19 | 0 | 5 |
| Harmful | 22,679 | 0.66 | 0.47 | 0 | 1 |
| Pharmful | 22,679 | 0.22 | 0.42 | 0 | 1 |
| Pnharmful | 22,679 | 0.05 | 0.22 | 0 | 1 |
| Nharmful | 22,679 | 0.04 | 0.20 | 0 | 1 |
| Live\_smoke | 22,679 | 0.35 | 0.48 | 0 | 1 |
| Live\_otob | 22,679 | 0.10 | 0.30 | 0 | 1 |
| Friends\_t | 22,679 | 0.67 | 1.20 | 0 | 4 |
| Friends\_unsure\_t | 22,679 | 0.14 | 0.35 | 0 | 1 |
| Friends\_otob | 22,679 | 0.27 | 0.81 | 0 | 4 |
| Friends\_unsure\_ot | 22,679 | 0.11 | 0.31 | 0 | 1 |
| Friends\_use | 22,679 | 0.94 | 1.70 | 0 | 1 |
| Rule\_no | 22,679 | 0.71 | 0.46 | 0 | 1 |
| Rule\_s | 22,679 | 0.09 | 0.28 | 0 | 1 |
| Rule\_a | 22,679 | 0.04 | 0.20 | 0 | 1 |
| Rule\_na | 22,679 | 0.13 | 0.34 | 0 | 1 |
| Rule\_M | 22,679 | 0.03 | 0.18 | 0 | 1 |
| Dy\_smoke\_y | 22,679 | 0.07 | 0.26 | 0 | 1 |
| Py\_smoke\_y | 22,679 | 0.10 | 0.30 | 0 | 1 |
| Pn\_smoke\_y | 22,679 | 0.15 | 0.36 | 0 | 1 |
| Dn\_smoke\_y | 22,679 | 0.64 | 0.48 | 0 | 1 |
| Na\_smoke\_y | 22,679 | 0.03 | 0.18 | 0 | 1 |
| Dy\_smoke\_bf | 22,679 | 0.06 | 0.24 | 0 | 1 |
| Py\_smoke\_bf | 22,679 | 0.10 | 0.30 | 0 | 1 |
| Pn\_smoke\_bf | 22,679 | 0.16 | 0.37 | 0 | 1 |
| Dn\_smoke\_bf | 22,679 | 0.65 | 0.48 | 0 | 1 |
| Na\_smoke\_bf | 22,679 | 0.02 | 0.14 | 0 | 1 |
| Par\_warn | 22,679 | 0.57 | 0.49 | 0 | 1 |
| Smoke\_cool | 22,679 | 0.88 | 0.32 | 0 | 1 |
| Smoke\_cool\_m | 22,679 | 0.03 | 0.16 | 0 | 1 |
| Say\_no\_m | 22,679 | 0.19 | 0.39 | 0 | 1 |
| say\_no | 22,679 | 0.29 | 0.45 | 0 | 1 |
| taught\_smoke\_m | 22,679 | 0.15 | 0.36 | 0 | 1 |
| taught\_smoke\_bad | 22,679 | 0.39 | 0.49 | 0 | 1 |

The next portion of this analysis I will be exploring the summary statistics above. The average age we see in this survey is 14.5 years old with the youngest one being nine years old and the oldest being 21. This has some conflicting information for age, generally sixth graders (lowest grade in the data) are around eleven years old and the youngest person in the survey is nine. This can be explained by having a bright student whom may have skipped some grades. Next, most people graduate high school at seventeen or eighteen while the oldest person in the sample is 21. This can be explained by a student being held back for a couple of years such that when they graduate they will be older than eighteen In the sample 50% of the population is male while the other half is female. This survey does not account for a non-binary person (someone who does not see themselves as masculine or feminine) in the survey the surveyed must identify as one or the other.

The next variable is Hispanic, where 28% of those sampled identified as such. The next variable is actually part of a larger group where it makes more sense to talk about the group than one at a time. The variables Aindian (American Indian or Alaskan Native), Asian, Black, Hawaiian, and white total up to about 91% of the surveyed meaning the other nine percent did not identify as any of these. The most common race was White, with 56% identifying as such, the second largest is Black with 20% identifying as such and the smallest group was Hawaiian or Alaskan Native with only three percent. The next variable miss shows that the average student missed 1.4 days of school a month which seems a little high and is probably this high because of the few students who miss a lot of school and are bringing up the average. Some of these students have missed zero days while the greatest amount is eleven days, which could be due to a moderate/severe illness. The Cig\_ind average tells us 22% of students have smoked a whole cigarette in their life. Use of dip, snuff, or chewing tobacco have been used by about ten percent of the student sample. The age of when students have used the other tobacco is 1.2 which does not make much sense. This is currently measured by the age at which they have used other tobacco products if they have not used the products a zero is entered giving us this strange average. The average student also uses the other tobacco products 1.3 days out of every month with the maximum a person using it is every day.

There is about 20% of the student sample that have smoked a cigar before. While the average student smoke one cigar every two months and some students who smoke it every day. There is also 89% of the student sample who has smoke a pipe once in their life. The average student also smoke one of these pipes once every four months and some students who smoke a pipe every day. The average student also spends one day a week in a room with someone else who smokes and some students every day are in a room with someone who smokes. The average student also spends one day every two weeks in the car with someone who smoke and some students who experience the most of it deal with smoking in the car five times a week.

The next variables are better understood if treated as a group harmful, pharmful, pnharmful, and nharmful where 88% of the student believes that cigarettes are probably or definitely harmful. There is about 35% of students who live with someone who smokes and only 10% of students live with someone who uses other tobacco products. The average students has 0.7 friends out of four who do actually smoke while 15% of the student sample are unsure if their friends smoke. Some student friends group have all four of their friends smoke. The students also have about 0.27 friends out of four that use other tobacco products while some have all four friends that use other tobacco products. There is also about 11% of student that do not know if their friend smoke. Next, I will look at rule\_no, rule\_s, rule\_a, rule\_na, and rule\_m where 71% of the students have strict rules of no smoking in the house and 23% of the students are allowed to smoke anywhere or in some places in the home.

The next group of variables are dy\_smoke\_y, py\_smoke\_y, pn\_smoke\_y, dn\_smoke\_y, and na\_smoke\_y where 17% of the students said they would probably or definitely smoke while three percent of the students left this blank. The next group of variables are dy\_smoke\_bf, py\_smoke\_bf, pn\_smoke\_bf, dn\_smoke\_bf, and na\_smoke\_bf where 16% of the students would take the cigarette if their best friend offered it to them while two percent left this question blank. A majority of parents (57%) have warned their children of smoking. A large majority of students (about 88%) thought smoking looks cool or helps young people fit in while three percent of students left this question blank. There was only 29% of students who have been trained on how to say no to smoking and 19% of student left this question blank or did not remember being trained on this subject. Surprisingly only 39% of the students have been taught in school that smoking is bad while 15% left this question blank or did not remember being taught this.

One thing I would also like to check before the end of this section is something called the dummy rule and largest-smallest rule (or smallest-largest rule). If this rule is violated it may be impossible to create a model without some intervention. To explain this it will be easier for me to show you.

| Table 7: Dummy rule and Largest/Smallest rule (and vice-versa) (Cont.) | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | where cig\_ind = 0 | | where cig\_ind = 1 | |  |
| Variable | Min0 | Max0 | Min1 | Max1 | Problem indicator |
| Age | 9 | 21 | 9 | 21 | 0 |
| Male | 0 | 1 | 0 | 1 | 0 |
| Hispanic | 0 | 1 | 0 | 1 | 0 |
| AIndian | 0 | 1 | 0 | 1 | 0 |
| Asian | 0 | 1 | 0 | 1 | 0 |
| Black | 0 | 1 | 0 | 1 | 0 |
| Hawaiian | 0 | 1 | 0 | 1 | 0 |
| White | 0 | 1 | 0 | 1 | 0 |
| miss | 0 | 11 | 0 | 11 | 0 |
| otob | 0 | 1 | 0 | 1 | 0 |
| age\_otob | 0 | 17 | 0 | 17 | 0 |
| use\_otob | 0 | 30 | 0 | 30 | 0 |
| cigar | 0 | 1 | 0 | 1 | 0 |
| day\_cigar | 0 | 30 | 0 | 30 | 0 |
| Pipe | 0 | 1 | 0 | 1 | 0 |
| day\_pipe | 0 | 30 | 0 | 30 | 0 |
| room | 0 | 7 | 0 | 7 | 0 |
| car | 0 | 5 | 0 | 5 | 0 |
| harmful | 0 | 1 | 0 | 1 | 0 |
| pharmful | 0 | 1 | 0 | 1 | 0 |
| pnharmful | 0 | 1 | 0 | 1 | 0 |
| nharmful | 0 | 1 | 0 | 1 | 0 |
| live\_smoke | 0 | 1 | 0 | 1 | 0 |
| live\_otob | 0 | 1 | 0 | 1 | 0 |
| friends\_t | 0 | 4 | 0 | 4 | 0 |
| friends\_unsure\_t | 0 | 1 | 0 | 1 | 0 |
| friends\_otob | 0 | 4 | 0 | 4 | 0 |
| friends\_unsure\_ot | 0 | 1 | 0 | 1 | 0 |
| rule\_no | 0 | 1 | 0 | 1 | 0 |
| rule\_s | 0 | 1 | 0 | 1 | 0 |
| rule\_a | 0 | 1 | 0 | 1 | 0 |
| rule\_na | 0 | 1 | 0 | 1 | 0 |
| rule\_M | 0 | 1 | 0 | 1 | 0 |
| dy\_smoke\_y | 0 | 1 | 0 | 1 | 0 |
| py\_smoke\_y | 0 | 1 | 0 | 1 | 0 |
| pn\_smoke\_y | 0 | 1 | 0 | 1 | 0 |
| dn\_smoke\_y | 0 | 1 | 0 | 1 | 0 |
| na\_smoke\_y | 0 | 1 | 0 | 1 | 0 |
| dy\_smoke\_bf | 0 | 1 | 0 | 1 | 0 |
| py\_smoke\_bf | 0 | 1 | 0 | 1 | 0 |
| pn\_smoke\_bf | 0 | 1 | 0 | 1 | 0 |
| dn\_smoke\_bf | 0 | 1 | 0 | 1 | 0 |
| na\_smoke\_bf | 0 | 1 | 0 | 1 | 0 |
| par\_warn | 0 | 1 | 0 | 1 | 0 |
| smoke\_cool | 0 | 1 | 0 | 1 | 0 |
| smoke\_cool\_m | 0 | 1 | 0 | 1 | 0 |
| say\_no\_m | 0 | 1 | 0 | 1 | 0 |
| say\_no | 0 | 1 | 0 | 1 | 0 |
| taught\_smoke\_m | 0 | 1 | 0 | 1 | 0 |
| taught\_smoke\_bad | 0 | 1 | 0 | 1 | 0 |
| friends\_use | 0 | 8 | 0 | 8 | 0 |
| agesq | 81 | 441 | 81 | 441 | 0 |

So for the dummy rule what I am checking is whether the min and max both vary. So for friends\_use where cig\_ind equals zero friends\_use varies between zero and one, where cig\_ind equals one the friends\_use also varies from zero to one. This is good, it passes the dummy rule, and if it did not pass we may have to get rid of the variable because we would not be able to compute a model. All of the dummy variables pass this rule so we may proceed. The next thing to look at is largest-smallest rule (and smallest-largest) this is used for continues variables. So looking at agesq we see that they overlap with each other (the min max for cig\_ind = 0 is from 81 to 441 and min max for cig\_ind = 1 is from 81 to 441 i.e. they both occupy the same space). All continues variables pass this rule so we may proceed.

The data for this analysis comes from the Center of Disease control (CDC) National Youth Tobacco Survey (NYTS) and can be found at <https://www.cdc.gov/tobacco/data_statistics/surveys/nyts/index.htm> while the code book which defines all of the questions asked to the students can be found at the same site. The survey has multiple years but for this analysis we will only be looking at the year 2009.

**IV Theoretical Model and Estimation Method**

Due to this variable special nature it excludes us from using Ordinary Least Squared (OLS). If we were to estimate OLS using this type of variable we would be running a model called linear probability model (LPM). This model is “ok” if we would like to get a general idea of what is happening, but when it comes to predicting and what the causal relationship is the LPM starts to fall apart. The first drawbacks of the LPM is heteroscedasticity, heteroscedasticity refers to the situation where the error has different variances given different values for the explanatory variable. For example, let look at the relationship between the number of hours playing certain carnival games and how many tickets won. The first game we will look at is a skill based game where I will make the assumption that your skill level will remain the same no matter how long you are playing.

In this game we see that the variance of the error (how far away the dot is from the trend line) is relatively the same given different levels of hours playing the skill based carnival game. This is what we would like to see in the model, but we may see something else called heteroscedasticity.

The error term for the luck based game is a bit more erratic depending on the hours played in the game. The reason behind this is that you may be on a winning a streak as in the more you play the more you win, but some people may have the opposite a losing streak. This is why we see a larger disparity the more you play. The difference between graph 1 and graph 2 is hopefully obvious at this point. That in graph 2 the variance changes based on the number of hours played while in graph 1 the variance stays the same no matter how many hours you have played. One last note graph 2 is only one example of the many different forms of heteroscedasticity. If the variance changes given different x value then we can say there is heteroscedasticity.

Now that I have explained what heteroscedasticity is you might be thinking why should we care? If there is heteroscedasticity in the model the causal relationship (betas) will be biased meaning we will not get the true causal relationship. The OLS estimators may also not be efficient meaning they have not achieved the smallest variance. This is basically saying we need to control for the heteroscedasticity in OLS or use a different model and in this case a different model is more appropriate.

The second drawback of LPM is the predictions are not bounded between zero and one. To explain this I will illustrate it through an example, if we were trying to measure the probability of a person making over $70,000 given a certain amount of experience. If we had data on their experience and if they do make over $70,000 (one if the person does, otherwise zero) we can graph it and do a regression like in the graph below.

As we can see in Graph 3 there is a clear positive relationship such that the more experience you have the more likely you are to be making over $70,000. There are two problems with this graph the first being if you have less than about two years of experience you have a negative probability of making over $70,000 and if you have over ~37 years of experience the probability of making over $70,000 goes above 100%. This makes no sense, it is impossible to have a probability over 100% or below zero percent.

The third drawback of the LPM is having constant partial effects, going back to the example a one unit increase in experience had a fixed amount increase in the probability that you would make over $70,000. This also does not make sense, usually in the beginning years of working experience will have a larger effect on your salary than the later part (i.e. not a constant partial effect).

The correct model we should be using is either a Probit or a Logit model. They both are used in similar circumstances, but the probit model is generally more accurate so we will be using this model for the analysis. Before I get into how we estimate this model I need to create a new variable called a latent variable. A latent variable is a variable that is not directly observed but is inferred through another variable that is observed. In this model the latent variable will be Cig\_indi\* which is not a binary variable but a continuous one. This latent variable is actually in the dependent variable Cig\_indi because Cig\_indi is defined as such.

(1.)

Cig\_indi\* that is contained within Cig\_indi is an unobservable index of the desire to smoke. For example, those students whose desire to smoke (Cig\_indi\* is greater than zero) will smoke because they have a desire to do so. Otherwise, they see smoking as bad or have no desire to smoke (Cig\_indi\* will be less than or equal to zero) and will choose not to smoke.

Using this latent variable it allows for more variation and different observations that appeared the same with Cig\_indi now can be differentiated. For example, two people who choose to smoke one being a chain smoker who smoke a pack a day and the other being a social smoker who maybe smokes a pack every month. With the original Cig\_indi they would both be given a one and the model would not see a difference between the two. With the latent variable Cig\_indi\* the model will be able to differentiate between the two. Where the chain smoker will be given a greater value than the social smoker. With this latent variable the hidden information is uncovered and we will produce a better more sound model.

To estimate this model we will be using something called maximum likehood estimation (MLE). MLE is used in non-linear regressions because these non-linear regressions (Probit model) has no closed form or in other words you cannot solve for them like you can for OLS using a simple equation. The first step of MLE is to create the probability density function (PDF) of the dependent variable. The PDF in this specific case is used to find the probability of the dependent variable falling within a certain range of values. For example, we may use a PDF to find the probability of a student passing a test.

(2.)

Once the PDF is found we will create the “likehood function” this is done by multiplying all of the marginal density functions at each observation. That is multiply the PDF at the first observation (marginal density function), and multiply at the PDF at the second observation (also a marginal density function), and so on until the nth observation.

(3.)

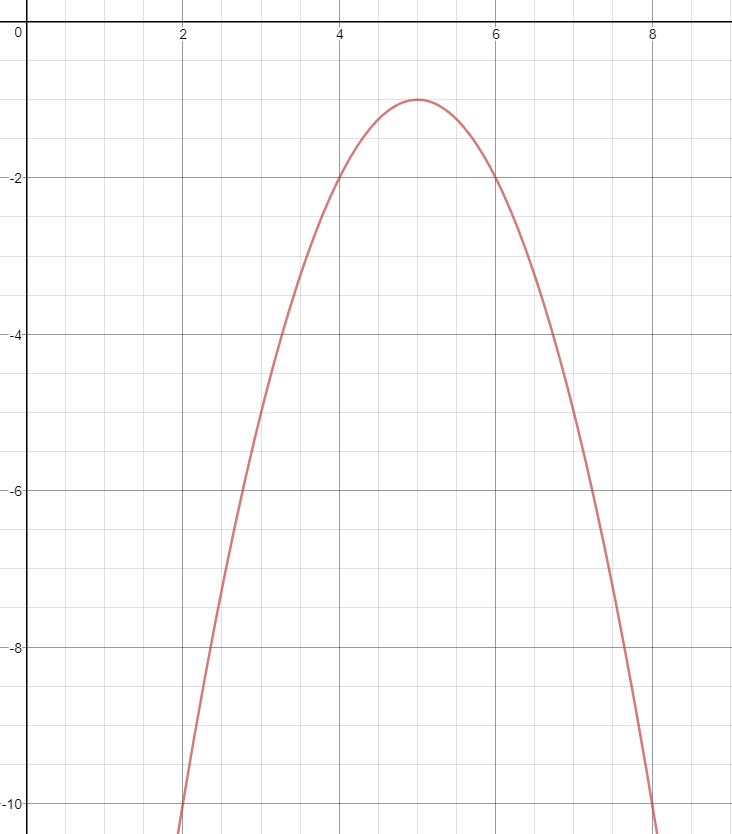


Now we would like to find the maximum of this equation (3.) though this can be very challenging because all of the marginal density functions are multiplied together. To combat this I can take the natural log of both sides which is considered a monotonic transformation. This monotonic transformation allows us to keep the same order of the variables or in other words the maximum of the function and the monotonic transformation of that function will be in the same place. This allows us to solve this problem in a reasonable time frame.

(4.)

Now if I was doing this by hand I would take a partial derivative with respect to beta j find the largest value and call it a day. Unfortunately, computers and calculus problem do not mix well with each other. The computer will use a more iterative process to estimate the maximum. For example, let the graph below be a likehood function we are trying to maximize.

Graph 4



The computer will first pick a random number let us say it is three. The next step is it will choose to either step left or right by one unit. If it chooses to step left by one unit it will find that it went lower from where it originally started and will now check the other way where it finds it increases. The computer will continue to step right by one unit as long as it increasing. So the computer will look at four see it has increased, then step to five and see it has increased once again, then take another step to six where it will see it has decreased. The computer then knows there is a local maximum to the left of six so it will cut the steps in half (step will be a 0.5 instead of one) and continue the other way. The computer will step to 5.5 see it has increased so the computer will again step to the left to five seeing it increased yet again it will once more step to the left to 4.5 seeing it has decreased it will cut the steps in half and continue the other way doing this process for a set amount of iterations. The computer will then give an approximated MLE if the computer criterion was satisfied and was able to find a maximum.

This next part of this section will cover all of the equations that will be used in this analysis. The first thing I will cover in this section is p-values. Whenever any statistician talks about p-values they are accompanied by at least two hypothesis. In this paper the p-values that are used later will have these same hypothesis.

(5.)

(This is called the alternative hypothesis)

Where is the coefficient under consideration

The p-value is the probability that the null hypothesis is correct or in other words the probability that the regressors has no causal relationship with the dependent variable. For example, let us regress income on (the number of years of education). If we were unsure if beta two had a causal relationship with income we can use hypothesis testing. First we would assume that

(null hypothesis) and find the probability of this being true. If the p-value = 0.01 (the probability of the null hypothesis being true is one percent) and we had a significance level (alpha) of 0.05 we would reject the null and say that beta two is statistically significant or in other words there is probably a causal relationship between education and income.

The marginal effect is used to gage how big or small the effect the parameter has on the dependent variable. With an OLS regression we can just take the beta as the marginal effect but with Probit because it is non-linear we cannot do this.

(6.)

This is a very useful item to have because the betas in the regression have no direct interpretation due to the nature of the probit model. There is also another useful metric we can gain from this. The average marginal effect which tells us the marginal effect for the average student in our dataset. This is easily calculated by adding all of the marginal effects for each of the observation and dividing by the number of observations.

Average Partial Effect = (7.)

(E.g. age, income, or etc.)

To calculate the probability of the event happening (i.e. the student has smoked a cigarette) I will use this formula.

(8.)

To calculate the probability of the event not hapeening (i.e. the student has not smoked a cigarette) I will use this formula.

(9.)

Endogeneity means a regressors is correlated with the error term. For example, if we were trying to regress wage on education and from this regression we were missing the regressor ability, we could say there is endogeneity if the covariance of the error term and education does not equal zero or in other words they relate with each other. If there is in fact endogeneity in the model this means that are estimators are wrong or in plain English the model has the causation wrong.

(10.)

Akaike Information Criterion (AIC) is an estimator that measures the relative quality of a statistical model, in particular, by measuring the model’s total error. AIC does not have a test for the model (no hypothesis testing) instead it gives you a number which only gives information if and only if it is compared to another AIC from another model with the smaller AIC being the preferred model. The formula for AIC is as follows

(11.)

.

To find a “better” model we can use AIC as a measurement by seeing the one with the smaller AIC (the model total error is smaller relative to the other model).

Percent correct predictions is calculated by finding the predicted probability for both outcomes and whichever outcome is more likely (the predictions with a probability greater than 50%) we say that value is the predicted one (for example, if the probability of being zero is 75% while the probability of being one is 25% we would make the prediction that the value will be zero). We will calculate this for each observation and make the prediction, once this is completed we will see the accuracy of the models prediction. This is calculated by creating two subgroups one where the outcome did not happen (zero’s) and another group where it did happen (one’s). Then in the did not happen group we will count the number of times we predicted the event did not happen and divide this by then total amount of times the event actually did not happen, this is repeated for the other group as well. This will give us the percentage of correctly predicted predictions.

**V Empirical Analysis and Model Selection**

There are many approaches that are available to find a “good” regression model. For this paper I will use the specific to general model approach. To use this approach I first create a regression model with all the variables that will, in theory, explain the dependent variable. I then start to add new variables that may also explain the dependent variable. Though, before I add any new variables I will first look at the p-values for each regressor to check whether they are significant and if they are significant I will check the sign of the regressor to see if it conforms to the logic that is in table five. If the regressors p-value is not significant I will reexamine the theory of should it be in the model. If I find my theory wanting the regressor will be dropped otherwise if the theory still holds water after further review, the regressor will be kept in the model. If the regressor is significant and has the wrong sign I will reexamine my theory to make sure it is correct, if I find a flaw in my theory that explains the different sign the regressor will be kept as is or if the regressor is non-linear I will create the appropriate non-linear regressor. Otherwise if my theory is sound and the regressor is linear then the regressor with a significant p-value and wrong sign will be dropped from the model because if left in it will cause the model to be unsound and give incorrect predictions. This will be redone each time any regressor is added to the model. The next criterion that will be checked once I create a new model is AIC I will make sure this value decreases with each regressor added so that my model is improved. Once I reach the point where I am unable to add new regressors that improves my model I will start to look for interaction terms. Interaction terms are usually created by multiplying or dividing different variables together.

The first thing to creating an interaction term is asking does theory or logic dictate this be in the model and if so we will add this to the model. For example, if we were asked to estimate the number of patrons that go to the mall we can use the number of square feet of the mall and the sum of total revenue brought in by each store. If I were to use these regressors independently these regression may give odd results for some malls. In general if a mall is small we would expect not many people entering the mall and vice-versa for a large mall. Though this may not always be true just because a mall is large it could be falling on hard times and unable to induce customer to shop while a small mall who may be booming and people are crowded in every inch of the mall. To combat this problem we can create an interaction term by dividing the sum of total revenue of each store by the total amount of square feet in the mall. Now by seeing the revenue by square feet we will be able to capture these booming small malls and the bust large malls.

The last criterion I will look at to choose my final model is the percent correctly predicted. This will give me a general idea in actual terms how well my model performs. Ideally we would like to be able to correctly predict each event with 80% accuracy, but as long as our predictions are better than the previous method (and not better just due to random chance) then the model we are using is fine. This last process is used when I am left with only a couple of different models with similar AIC and statistically significant regressors.

Once I have gone through this process I will end up with my final model. I will present a few different forms of this model the first will be the population form.

(12.)

The population form is unknown but we may estimate it by using MLE. This will approximate the population form of the model and will allow us to make predictions. The main difference between these written equations is the little hat that is above the betas and Cig\_ind. This signifies the parameters are the estimated values of the population form.

(13.)

Now that the model is specified I am able to use software to estimate the beta hats. I will present this two ways the first will be similar to the equation form that is seen above and the other will be in a table format. I will present more information in the table because in general it is easier to read.

(14.)

(P-values for these regressors will be reported in table 8)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 8: Probit estimates, Dependent Variable: Cig\_ind | | | | |
| ID # | Parameters | Estimate | t-Value | P-Value\* |
| 1 | Intercept | -5.9129 | -9.12 | <0.0001 |
| 2 | Age | 0.4900 | 5.68 | <0.0001 |
| 3 | Agesq | -0.0107 | -3.77 | 0.0002 |
| 4 | Hispanic | 0.1552 | 4.66 | <0.0001 |
| 5 | Black | 0.1247 | 3.21 | 0.0013 |
| 6 | White | 0.1487 | 4.36 | <0.0001 |
| 7 | Miss | 0.0190 | 3.59 | 0.0003 |
| 8 | Cigar | 0.9679 | 33.85 | <0.0001 |
| 9 | Room | 0.0773 | 12.05 | <0.0001 |
| 10 | Harmful | -0.2196 | -5.64 | <0.0001 |
| 11 | Pharmful | -0.0864 | -2.05 | 0.0408 |
| 12 | Live\_smoke | 0.1997 | 6.76 | <0.0001 |
| 13 | Friends\_use | 0.0977 | 13.13 | <0.0001 |
| 14 | Rule\_no | 0.1359 | 3.83 | 0.0001 |
| 15 | Rule\_s | 0.1696 | 3.44 | 0.0006 |
| 16 | Rule\_a | 0.2397 | 3.76 | 0.0002 |
| 17 | Dy\_smoke\_y | 0.6374 | 10.79 | <0.0001 |
| 18 | Dn\_smoke\_y | -0.5035 | -14.47 | <0.0001 |
| 19 | Dy\_smoke\_bf | 0.4067 | 6.43 | <0.0001 |
| 20 | Dn\_smoke\_bf | -0.6470 | -18.59 | <0.0001 |
| 21 | Smoke\_cool | 0.0557 | 1.52 | 0.1287 |
| \*Ho: beta\_j = 0, | | | | |
| Ha: beta\_j != 0 (read as beta j not equal to zero) | | | | |

Before we analyze the coefficients of the parameter we first have to analyze the p-values. In this case I will present three tables with parameters that are significant at the 99%, 95%, and less than 90% (insignificant) level.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 9: Regressors significant at the 99% level | | | | |
| ID # | Parameter | Estimate | t-value | P-value\* |
| 1 | Intercept | -5.9129 | -9.12 | <0.0001 |
| 2 | Age | 0.4900 | 5.6800 | <0.0001 |
| 3 | agesq | -0.0107 | -3.7700 | 0.0002 |
| 4 | Hispanic | 0.1552 | 4.6600 | <0.0001 |
| 5 | Black | 0.1247 | 3.2100 | 0.0013 |
| 6 | White | 0.1487 | 4.3600 | <0.0001 |
| 7 | miss | 0.0190 | 3.5900 | 0.0003 |
| 8 | cigar | 0.9679 | 33.8500 | <0.0001 |
| 9 | room | 0.0773 | 12.0500 | <0.0001 |
| 10 | harmful | -0.2196 | -5.6400 | <0.0001 |
| 12 | live\_smoke | 0.1997 | 6.7600 | <0.0001 |
| 13 | friends\_use | 0.0977 | 13.1300 | <0.0001 |
| 14 | rule\_no | 0.1359 | 3.8300 | 0.0001 |
| 15 | rule\_s | 0.1696 | 3.4400 | 0.0006 |
| 16 | rule\_a | 0.2397 | 3.7600 | 0.0002 |
| 17 | dy\_smoke\_y | 0.6374 | 10.7900 | <0.0001 |
| 18 | dn\_smoke\_y | -0.5035 | -14.4700 | <0.0001 |
| 19 | dy\_smoke\_bf | 0.4067 | 6.4300 | <0.0001 |
| 20 | dn\_smoke\_bf | -0.6470 | -18.5900 | <0.0001 |
| \*Ho: beta\_j = 0, | | | | |
| Ha: beta\_j != 0 (read as beta j not equal to zero) | | | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 10: Regressors significant at the 95% level | | | | |
| ID # | Parameter | Estimate | t-value | P-value\* |
| 11 | pharmful | -0.0864 | -2.05 | 0.0408 |
| \*Ho: beta\_j = 0, | | | | |
| Ha: beta\_j != 0 (read as beta j not equal to zero) | | | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 11: Regressors that are insignificant | | | | |
| ID # | Parameter | Estimate | t-value | P-value\* |
| 21 | smoke\_cool | 0.0557 | 1.52 | 0.1287 |
| \*Ho: beta\_j = 0, | | | | |
| Ha: beta\_j != 0 (read as beta j not equal to zero) | | | | |

As you have probably noticed the variable Smoke\_cool is insignificant though it is still in the model. The variable was kept for a reason being this explains an attitude to a student smoking. If a student believes smoking is cool they are more like to smoke.

The next thing I need to check is does the regressors conform to the logic that is stated in Table 12.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 12: Regressors coefficient and their expected signs | | | | |
| ID # | Variable | Coefficient | Expected Sign | Sign Conforms to Logic |
| 1 | Intercept | -5.9129 | . | . |
| 2 | Age | 0.4900 | -/+ | TRUE |
| 3 | Agesq | -0.0107 |  |  |
| 4 | Hispanic | 0.1552 | ? | . |
| 5 | Black | 0.1247 | ? | . |
| 6 | White | 0.1487 | ? | . |
| 7 | Miss | 0.0190 | + | TRUE |
| 8 | Cigar | 0.9679 | + | TRUE |
| 9 | Room | 0.0773 | + | TRUE |
| 10 | Harmful | -0.2196 | - | TRUE |
| 11 | Pharmful | -0.0864 | - | TRUE |
| 12 | Live\_smoke | 0.1997 | + | TRUE |
| 13 | Friends\_use | 0.0977 | + | TRUE |
| 14 | Rule\_no | 0.1359 | - | FALSE |
| 15 | Rule\_s | 0.1696 | + | TRUE |
| 16 | Rule\_a | 0.2397 | + | TRUE |
| 17 | Dy\_smoke\_y | 0.6374 | + | TRUE |
| 18 | Dn\_smoke\_y | -0.5035 | - | TRUE |
| 19 | Dy\_smoke\_bf | 0.4067 | + | TRUE |
| 20 | Dn\_smoke\_bf | -0.6470 | - | TRUE |
| 21 | Smoke\_cool | 0.0557 | + | TRUE |

All of the regressors but one conformed to our logic so we will need to delve into why I was incorrect. Rule\_no or no smoking allowed into the house, I thought would have a negative sign because they have strict parents. Meaning they would have less of a chance to smoke. Though, this make actually be a double edge sword meaning because the parents are strict the students may be rebellious they may do it in spite of their parents or the parents may have already caught the student smoking so that is why the rule was put in place. Hence, the coefficient carries a positive sign or is more likely to smoke. With that cleared up there is no evidence that the model is unsound so we may proceed.

The next thing I want to look at is the total impact for each regressor. The Probit model coefficients are not directly transferable into the probability of a student smoking. This is due to a major problem that would come up if this was true. For example, I will assume there is constant partial effects (please note this is untrue, I am just saying this for arguments sake) meaning the betas are the change in probability. If we wanted to estimate the probability of someone having a pool in their backyard by regressing if you have a pool on the count of the number of neighbors who have pools. If we found the beta to be 0.75 for count of neighbor pools this would mean there is 150% chance (if both of your neighbors have pools) of you having a pool in your back yard, which is impossible. Therefore, regressors have non-constant effects and betas are not directly transferable into probability. In order to get the effect for the individual variable we have to look at marginal effects. In order to see the effect of an individual regressor I will randomly choose four students (two who have and have not smoked a cigarette) and look at their marginal effects.

| Table 13: Variables and their Marginal Effects, with respect to students (Cont.) | | | | |
| --- | --- | --- | --- | --- |
| Variable and Marginal Effect | Student #4,364 | Student #4,583 | Student #13,176 | Student #19,132 |
| cig\_ind | 0 | 1 | 1 | 0 |
| Age | 16 | 15 | 14 | 16 |
| agesq | 256 | 225 | 196 | 256 |
| Hispanic | 1 | 0 | 0 | 0 |
| Black | 0 | 1 | 0 | 0 |
| White | 0 | 0 | 1 | 1 |
| miss | 1 | 1 | 1 | 0 |
| cigar | 1 | 0 | 1 | 0 |
| room | 0 | 0 | 1 | 7 |
| harmful | 1 | 0 | 1 | 1 |
| pharmful | 0 | 0 | 0 | 0 |
| live\_smoke | 0 | 0 | 1 | 1 |
| friends\_use | 2 | 0 | 3 | 3 |
| rule\_no | 1 | 0 | 0 | 0 |
| rule\_s | 0 | 0 | 1 | 0 |
| rule\_a | 0 | 0 | 0 | 0 |
| dy\_smoke\_y | 0 | 0 | 0 | 0 |
| dn\_smoke\_y | 1 | 1 | 0 | 0 |
| dy\_smoke\_bf | 0 | 0 | 0 | 0 |
| dn\_smoke\_bf | 1 | 1 | 0 | 0 |
| na\_smoke\_bf | 0 | 0 | 0 | 0 |
| smoke\_cool | 1 | 1 | 0 | 0 |
| Marginal effect of Age on the probability of cig\_ind | 15.67% | 3.02% | 17.28% | 19.36% |
| Marginal effect of agesq on the probability of cig\_ind | -0.34% | -0.07% | -0.38% | -0.42% |
| Marginal effect of Hispanic on the probability of cig\_ind | 4.96% | 0.96% | 5.47% | 6.13% |
| Marginal effect of Black on the probability of cig\_ind | 3.99% | 0.77% | 4.40% | 4.93% |
| Marginal effect of White on the probability of cig\_ind | 4.76% | 0.92% | 5.25% | 5.88% |
| Marginal effect of miss on the probability of cig\_ind | 0.61% | 0.12% | 0.67% | 0.75% |
| Marginal effect of cigar on the probability of cig\_ind | 30.95% | 5.97% | 34.14% | 38.24% |
| Marginal effect of room on the probability of cig\_ind | 2.47% | 0.48% | 2.73% | 3.05% |
| Marginal effect of harmful on the probability of cig\_ind | -7.02% | -1.36% | -7.75% | -8.68% |
| Marginal effect of pharmful on the probability of cig\_ind | -2.76% | -0.53% | -3.05% | -3.41% |
| Marginal effect of live\_smoke on the probability of cig\_ind | 6.39% | 1.23% | 7.04% | 7.89% |
| Marginal effect of friends\_use on the probability of cig\_ind | 3.12% | 0.60% | 3.45% | 3.86% |
| Marginal effect of rule\_no on the probability of cig\_ind | 4.35% | 0.84% | 4.79% | 5.37% |
| Marginal effect of rule\_s on the probability of cig\_ind | 5.42% | 1.05% | 5.98% | 6.70% |
| Marginal effect of rule\_a on the probability of cig\_ind | 7.67% | 1.48% | 8.46% | 9.47% |
| Marginal effect of dy\_smoke\_y on the probability of cig\_ind | 20.39% | 3.94% | 22.48% | 25.19% |
| Marginal effect of dn\_smoke\_y on the probability of cig\_ind | -16.10% | -3.11% | -17.76% | -19.90% |
| Marginal effect of dy\_smoke\_bf on the probability of cig\_ind | 13.01% | 2.51% | 14.35% | 16.07% |
| Marginal effect of dn\_smoke\_bf on the probability of cig\_ind | -20.69% | -3.99% | -22.82% | -25.56% |
| Marginal effect of smoke\_cool on the probability of cig\_ind | 1.78% | 0.34% | 1.96% | 2.20% |

I will now go into detail over student #4,364, please note the other students marginal effect can be interpreted similarly.

For a student who has the same traits as student #4,364, each one unit increase in Age increases the probability of smoking by 15.67%, ceteris paribus.

For a student who has the same traits as student #4,364, each one unit increase in agesq increases the probability of smoking by -0.34%, ceteris paribus.

For a student who has the same traits as student #4,364, if the student was not Hispanic that would decrease the probability of smoking by 4.96%, ceteris paribus.

For a student who has the same traits as student #4,364, if the student had been reports yes to Black the probability of smoking for that student would increases by 3.99%, ceteris paribus.

For a student who has the same traits as student #4,364, if the student had been reported yes to White the probability of smoking would increase by 4.76%, ceteris paribus.

For a student who has the same traits as student #4,364, each one unit increase in miss increases the probability of smoking by 0.61%, ceteris paribus.

For a student who has the same traits as student #4,364, if the student had not smoked a cigar there would be a decrease in the probability of smoking by 30.95%, ceteris paribus.

For a student who has the same traits as student #4,364, each one unit increase in room increases the probability of smoking by 2.47%, ceteris paribus.

For a student who has the same traits as student #4,364, if the student does not think cigarettes are harmful the probability of smoking would increase by 7.02%, ceteris paribus.

For a student who has the same traits as student #4,364, if the student had reported yes to pharmful the probability of smoking would increase by -2.76%, ceteris paribus.

For a student who has the same traits as student #4,364, if the student had reported yes to live\_smoke the probability of smoking would increase by 6.39%, ceteris paribus.

For a student who has the same traits as student #4,364, each one unit increase in friends\_use increases the probability of smoking by 3.12%, ceteris paribus.

For a student who has the same traits as student #4,364, if the student did not have the rule\_no the probability of smoking would decrease by 4.35%, ceteris paribus.

For a student who has the same traits as student #4,364, if the student had reports yes to rule\_s the probability of smoking would increase by 5.42%, ceteris paribus.

For a student who has the same traits as student #4,364, if the student had reports yes to rule\_a the probability of smoking would increase by 7.67%, ceteris paribus.

For a student who has the same traits as student #4,364, if the student had reports yes to dy\_smoke\_y the probability of smoking would increase by 20.39%, ceteris paribus.

For a student who has the same traits as student #4,364, if the student did not say yes to dn\_smoke\_y the probability of smoking would increase by 16.1%, ceteris paribus.

For a student who has the same traits as student #4,364, if the student had reports yes to dy\_smoke\_bf the probability of smoking would increase by 13.01%, ceteris paribus.

For a student who has the same traits as student #4,364, if the student did not say yes to dn\_smoke\_bf the probability of smoking would increase by 20.69%, ceteris paribus.

For a student who has the same traits as student #4,364, if the student did not say yes to smoke\_cool the probability of smoking would decrease by 1.78%, ceteris paribus.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 14: Variables and their predicted probabilities, with respect to students | | | | |
| Variable and the Predicted Probability | Student #4,364 | Student #4,583 | Student #13,176 | Student #19,132 |
| cig\_ind | 0 | 1 | 1 | 0 |
| Age | 16 | 15 | 14 | 16 |
| agesq | 256 | 225 | 196 | 256 |
| Hispanic | 1 | 0 | 0 | 0 |
| Black | 0 | 1 | 0 | 0 |
| White | 0 | 0 | 1 | 1 |
| miss | 1 | 1 | 1 | 0 |
| cigar | 1 | 0 | 1 | 0 |
| room | 0 | 0 | 1 | 7 |
| harmful | 1 | 0 | 1 | 1 |
| pharmful | 0 | 0 | 0 | 0 |
| live\_smoke | 0 | 0 | 1 | 1 |
| friends\_use | 2 | 0 | 3 | 3 |
| rule\_no | 1 | 0 | 0 | 0 |
| rule\_s | 0 | 0 | 1 | 0 |
| rule\_a | 0 | 0 | 0 | 0 |
| dy\_smoke\_y | 0 | 0 | 0 | 0 |
| dn\_smoke\_y | 1 | 1 | 0 | 0 |
| dy\_smoke\_bf | 0 | 0 | 0 | 0 |
| dn\_smoke\_bf | 1 | 1 | 0 | 0 |
| na\_smoke\_bf | 0 | 0 | 0 | 0 |
| smoke\_cool | 1 | 1 | 0 | 0 |
| Probability of cig\_ind taking on the value of 0 | 74.70% | 97.33% | 30.98% | 44.48% |
| Probability of cig\_ind taking on the value of 1 | 25.30% | 2.67% | 69.02% | 55.52% |

The interpretation for the probability of smoking a cigarette is very similar for all of these students so I will only present one, but again this can be easily applied to the others. For a student who is similar to student #4,364 there is a 74.70% chance that they are smoking, ceteris paribus. Note this person was not smoking so the model made a correct prediction.

**VI Addressing Heteroscedasticity:**

Before I get into the main point of how to determine if there is heteroscedasticity and how to fix it, you should know what it is. Heteroscedasticity is a non-constant variance of the population error term. If the reader would like a more through explanation and example of heteroscedasticity there is a comprehensive explanation of what heteroscedasticity for the dependent variable (though it is directly transferable to the explanatory variables) in section IV if you would like to refresh your memory. The problem with heteroscedasticity is that causes our hypotheses test to be biased so if there is heteroscedasticity and we do not control for it our p-values may be non-sense.

The first thing we must check is if we have heteroscedasticity, to do this we will employ a test called the Breusch-Pagan test. The first thing I do will be create the regression that we will be using. Then I will calculate the error term, square it, and then run a regression with each regressor. To illustrate this I will use the previous example with the skilled and luck based carnival games. First I will look at the skilled based game (homoscedasticity) I will reproduce the original graph and linear regression and then the error squared and linear regression.

Here the linear regression has a slope that is no different from zero because the variance of the error does not change with the hours spent at the carnival. Therefore, I can say there is homoscedasticity. Now I will look at a heteroscedasticity example.

Here the linear regression has a slope that is different from zero because the variance of the error does change with the hours spent at the carnival. Therefore, I can say there is heteroscedasticity.

I can also show this in more general terms first I will specifying the regression.

(15.)

Then I will create a new variable errors\_squared for each observation.

(16.)

Then run an auxiliary regression or in other word a simple regression where errors\_squared is my dependent variable and one explanatory variable.

(17.)

Then I will compute the p-value for eta hat to see if it is different from zero (Ho: eta hat = 0, Ha: eta hat does not equal 0). If it is not different I can say there is not heteroscedasticity otherwise there is heteroscedasticity. Equation 17 is calculated for each regressor so if I have five explanatory variables I will run the equation five times for each regressor.

I will calculate the Breusch-Pagan test for all of the variables, but I will only present the estimate for eta one because this is the only estimate that we care about.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 15: Breusch-Pagan test and estimates/p-value for eta one | | | |
| Parameter | Estimate | p-Value\* | Heteroscedasticity indicator |
| Age | -0.0022 | 0.8579 | 0 |
| agesq | 0.0001 | 0.8949 | 0 |
| Hispanic | 0.1910 | 0.0035 | 1 |
| Black | 0.3536 | <0.0001 | 1 |
| White | -0.2739 | <0.0001 | 1 |
| miss | 0.0890 | <0.0001 | 1 |
| cigar | 0.4696 | <0.0001 | 1 |
| room | 0.0650 | <0.0001 | 1 |
| harmful | -0.1279 | 0.0044 | 1 |
| pharmful | -0.0878 | 0.0921 | 0 |
| live\_smoke | 0.1373 | 0.0177 | 1 |
| friends\_use | 0.0669 | 0.0003 | 1 |
| rule\_no | -0.2358 | <0.0001 | 1 |
| rule\_s | 0.2203 | 0.0272 | 1 |
| rule\_a | 0.1441 | 0.2606 | 0 |
| dy\_smoke\_y | 0.2957 | 0.0817 | 0 |
| dn\_smoke\_y | 0.0668 | 0.3811 | 0 |
| dy\_smoke\_bf | 0.4788 | 0.024 | 1 |
| dn\_smoke\_bf | -0.0334 | 0.6243 | 0 |
| smoke\_cool | -0.3875 | <0.0001 | 1 |
| \*Ho: beta\_j = 0, | | | |
| Ha: beta\_j != 0 (read as beta j not equal to zero) | | | |

The table tells us that there are many variables that have heteroscedasticity so we must correct for it. There are two methods I may use to correct for heteroscedasticity generalized least squares (GLS) or robust standard errors. GLS has a few more restrictions and is usually very hard or impossible to estimate if there is more than one regressor that has heteroscedasticity. This was true in my original model. Therefore, I will use robust standard errors which will give me a new regression with the corrected p-values.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 16: Probit estimates with robust standard errors,  Dependent Variable: Cig\_ind | | | | |
| ID # | Parameters | Estimate | t-Value | P-Value\* |
| 1 | Intercept | -5.9129 | -7.38 | <0.0001 |
| 2 | Age | 0.4900 | 4.62 | <0.0001 |
| 3 | Agesq | -0.0107 | -3.08 | 0.0020 |
| 4 | Hispanic | 0.1552 | 4.56 | <0.0001 |
| 5 | Black | 0.1246 | 3.06 | 0.0020 |
| 6 | White | 0.1487 | 4.29 | <0.0001 |
| 7 | Miss | 0.0190 | 3.24 | 0.0010 |
| 8 | Cigar | 0.9679 | 33.20 | <0.0001 |
| 9 | Room | 0.0773 | 11.65 | <0.0001 |
| 10 | Harmful | -0.2196 | -5.37 | <0.0001 |
| 11 | Pharmful | -0.0864 | -1.98 | 0.0480 |
| 12 | Live\_smoke | 0.1997 | 6.68 | <0.0001 |
| 13 | Friends\_use | 0.0977 | 12.47 | <0.0001 |
| 14 | Rule\_no | 0.1359 | 3.76 | <0.0001 |
| 15 | Rule\_s | 0.1696 | 3.29 | 0.0010 |
| 16 | Rule\_a | 0.2397 | 3.63 | <0.0001 |
| 17 | Dy\_smoke\_y | 0.6374 | 10.84 | <0.0001 |
| 18 | Dn\_smoke\_y | -0.5035 | -14.70 | <0.0001 |
| 19 | Dy\_smoke\_bf | 0.4067 | 6.35 | <0.0001 |
| 20 | Dn\_smoke\_bf | -0.6470 | -18.96 | <0.0001 |
| 21 | Smoke\_cool | 0.0557 | 1.44 | 0.1490 |
| \*Ho: beta\_j = 0, | | | | |
| Ha: beta\_j != 0 (read as beta j not equal to zero) | | | | |

There was no changes with this correction all variables that were significant before stayed significant and all regressors that were insignificant stayed insignificant.

**VII Conclusion:**

In order to make suggestions on what can be done to discourage students from smoking, I will first look at the effects of the regressors. The variables that increased the probability of students smoking were if a student has smoked a cigar, their current age, do they live with somebody that smokes, do their friends smoke or use other tobacco products, and do they miss school. The variables that had negative effects were their beliefs on if smoking was harmful.

The biggest contributor to explain whether a student was going to smoke was if they have smoked a cigar before and the biggest negative factor was their belief if smoking was harmful. In order to stop or hinder students from smoking we should create a more strict policy for smoking cigars. Overall smoking cigars is the most commonly used tobacco product after cigarettes[[3]](#footnote-3). Cigars may be acting like a gateway drug that once tried cigarettes seem less bad and students will be more likely to smoke cigarettes. Another policy is to more specifically target parents or people who live with students by making sure they know the harmful effects and that it may cause their children or children around them to smoke too. Lastly, provide a class that gives the students more understanding of the harmful effect of smoking both physically and economically.

**VIII Shortcomings:**

There are probably a few more regressors that should have been added including family income, whether they live in an apartment, condo, or house, how many anti-tobacco advertisements they see weekly, count of siblings, count of any non-family members that live with them, do they have a job, and GPA or some sort of test for their IQ.

The model probably suffers from some endogeneity with some people having more addictive personality that are therefore more inclined to smoke cigarettes and cigars. A more appropriate study would be to follow these same students for a couple years and have them retake the survey every year. Then I would be able to more control for the addictive personality and be able to provide a more accurate model.

**IX Acknowledgements**

I would like to gratefully thank Jeanette Barouch for proofreading this paper and provided valuable comments and suggestions for improvement.

**X References:**

Corey, C. G., Dube, S. R., Ambrose, B. K., King, B. A., Apelberg, B. J., & Husten, C. G. (2014, August). Cigar Smoking Among U.S. Students: Reported Use After Adding Brands to Survey Items. Retrieved December 10, 2017, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4521384/>

Doll, R., Peto, R., Boreham, J., & Sutherland, I. (2004, June 24). Mortality in relation to smoking: 50 years' observations on male British doctors. Retrieved December 10, 2017, from <http://www.bmj.com/content/328/7455/1519>

National Youth Tobacco Survey. (n.d.). Retrieved December 10, 2017, from <https://www.healthypeople.gov/2020/data-source/national-youth-tobacco-survey>

**XI Appendix:**

%let Path = D:\Applied\;

%include "&path.Macros.sas";

libname pb "D:\Applied\Research Project";

**%macro** ifm(var, replace\_with); /\*replaces missing values with user specified numbers\*/

if &var = **.** then

&var = &replace\_with;

**%mend** ifm;

**proc** **import** datafile = "&Path.Research Project\NYTS 2009 Dataset.xlsx"

out = Survey

dbms = xlsx;

**run**;

**data** clean;

set Survey;

Age = qn1 + **8**;

Grade = qn3 + **5**;

if qn2 = **2** then

Male = **1**;

else

Male = **0**;

if qn4 = **1** then

Hispanic = **1**;

else

Hispanic = **0**;

if qn5a = **1** then

AIndian = **1**;

else

AIndian = **0**;

if qn5b = **1** then

Asian = **1**;

else

Asian = **0**;

if qn5c = **1** then

Black = **1**;

else

Black = **0**;

if qn5d = **1** then

Hawaiian = **1**;

else

Hawaiian = **0**;

if qn5e = **1** then

White = **1**;

else

White = **0**;

if qn7 = **1** then

miss = **0**;

else if qn7 = **2** then

miss = **1**;

else if qn7 = **3** then

miss = **2**;

else if qn7 = **4** then

miss = **6**;

else

miss = **11**;

if qn10 > **2** then

cig\_ind = **1**;

else

cig\_ind = **0**;

if qn38 = **1** then

otob = **1**;

else

otob = **0**;

if qn39 = **1** then

age\_otob = **0**;

else

age\_otob = qn39 + **6**;

if qn40 = **1** then

use\_otob = **0**;

else if qn40 = **2** then

use\_otob = **1**;

else if qn40 = **3** then

use\_otob = **3**;

else if qn40 = **4** then

use\_otob = **6**;

else if qn40 = **5** then

use\_otob = **10**;

else if qn40 = **6** then

use\_otob = **20**;

else

use\_otob = **30**;

if qn43 = **1** then

cigar = **0**;

else if qn43 > **1** then

cigar = **1**;

if qn44 = **1** then

day\_cigar = **0**;

else if qn44 = **2** then

day\_cigar = **1**;

else if qn44 = **3** then

day\_cigar = **3**;

else if qn44 = **4** then

day\_cigar = **6**;

else if qn44 = **5** then

day\_cigar = **10**;

else if qn44 = **6** then

day\_cigar = **20**;

else if qn44 = **7** then

day\_cigar = **30**;

if qn45 = **1** then

Pipe = **0**;

else if qn45 > **1** then

Pipe = **1**;

if qn46 = **1** then

day\_pipe = **0**;

else if qn46 = **2** then

day\_pipe = **1**;

else if qn46 = **3** then

day\_pipe = **3**;

else if qn46 = **4** then

day\_pipe = **6**;

else if qn46 = **5** then

day\_pipe = **10**;

else if qn46 = **6** then

day\_pipe = **20**;

else if qn46 = **7** then

day\_pipe = **30**;

if qn51 = **1** then

room = **0**;

else if qn51 = **2** then

room = **1**;

else if qn51 = **3** then

room = **3**;

else if qn51 = **4** then

room = **5**;

else if qn51 = **5** then

room = **7**;

if qn52 = **1** then

car = **0**;

else if qn52 = **2** then

car = **1**;

else if qn52 = **3** then

car = **3**;

else if qn52 = **4** then

car = **5**;

else if qn52 = **5** then

car = **7**;

if qn52 = **1** then

harmful = **1**;

else

harmful = **0**;

if qn52 = **2** then

pharmful = **1**;

else

pharmful = **0**;

if qn52 = **3** then

pnharmful = **1**;

else

pnharmful = **0**;

if qn52 = **4** then

nharmful = **1**;

else

nharmful = **0**;

if qn53 = **1** then

live\_smoke = **1**;

else

live\_smoke = **0**;

if qn54 = **1** then

live\_otob = **1**;

else

live\_otob = **0**;

if qn55 < **6** then

friends\_t = qn55 - **1**;

if qn55 = **6** or qn55 = **.** then

friends\_unsure\_t = **1**;

else

friends\_unsure\_t = **0**;

if qn56 < **6** then

friends\_otob = qn56 - **1**;

if qn56 = **6** or qn56 = **.** then

friends\_unsure\_ot = **1**;

else

friends\_unsure\_ot = **0**;

rule\_no = **0**;

rule\_s = **0**;

rule\_a = **0**;

rule\_na = **0**;

rule\_M = **0**;

if qn57 = **.** then

rule\_M = **1**;

else if qn57 = **1** then

rule\_no = **1**;

else if qn57 = **2** then

rule\_s = **1**;

else if qn57 = **3** then

rule\_a = **1**;

else if qn57 = **4** then

rule\_na = **1**;

dy\_smoke\_y = **0**;

py\_smoke\_y = **0**;

pn\_smoke\_y = **0**;

dn\_smoke\_y = **0**;

na\_smoke\_y = **0**;

if qn58 = **1** then

dy\_smoke\_y = **1**;

else if qn58 = **2** then

py\_smoke\_y = **1**;

else if qn58 = **3** then

pn\_smoke\_y = **1**;

else if qn58 = **4** then

dn\_smoke\_y = **1**;

else

na\_smoke\_y = **1**;

dy\_smoke\_bf = **0**;

py\_smoke\_bf = **0**;

pn\_smoke\_bf = **0**;

dn\_smoke\_bf = **0**;

na\_smoke\_bf = **0**;

if qn60 = **1** then

dy\_smoke\_bf = **1**;

else if qn60 = **2** then

py\_smoke\_bf = **1**;

else if qn60 = **3** then

pn\_smoke\_bf = **1**;

else if qn60 = **4** then

dn\_smoke\_bf = **1**;

else

na\_smoke\_bf = **1**;

if qn62 < **4** and qn62 ~= **.** then

par\_warn = **1**;

else

par\_warn = **0**;

if qn65 > **2** then

smoke\_cool = **1**;

else if qn65 ~= **.** then

smoke\_cool = **0**;

if smoke\_cool = **.** then

smoke\_cool\_m = **1**;

else

smoke\_cool\_m = **0**;

say\_no\_m = **0**;

if qn78 = **1** then

say\_no = **1**;

else if qn78 = **2** then

say\_no = **0**;

else if qn78 = **.** or qn78 = **3** then

say\_no\_m = **1**;

taught\_smoke\_m = **0**;

if qn81 = **1** then

taught\_smoke\_bad = **0**;

else if qn81 = **2** then

taught\_smoke\_bad = **1**;

else if qn81 = **3** or qn81 = **.** then

taught\_smoke\_m = **1**;

**run**;

/\*proc contents data = clean;\*/

/\*run;\*/

**data** clean1 (drop = qn:);

set clean;

**run**;

**proc** **means** n nmiss mean stddev min max data = clean1;

**run**;

**data** clean2;

set clean1;

%***ifm***(Age, **15**);

%***ifm***(grade, **9**);

%***ifm***(age\_otob, **0**)

%***ifm***(cigar, **0**);

%***ifm***(day\_cigar, **0**);

%***ifm***(pipe, **0**);

%***ifm***(day\_pipe, **0**);

%***ifm***(room, **0**);

%***ifm***(car, **0**);

%***ifm***(friends\_t, **0**);

%***ifm***(friends\_otob, **0**);

%***ifm***(smoke\_cool, **0**);

%***ifm***(say\_no, **0**);

%***ifm***(taught\_smoke\_bad, **0**);

**run**;

/\*proc means nmiss;\*/

/\*run;\*/

**proc** **means** data = clean2;

**run**;

**data** add\_var;

set clean2;

friends\_use = friends\_t + friends\_otob;

if age >= **18** then

legal = **1**;

else

legal = **0**;

agesq = age\*\***2**;

age\_lt18 = age \* (age < **18**);

age\_gte18 = age \* (age >= **18**);

**run**;

%let Indep =

Age

agesq

/\* Grade\*/

/\* Male\*/

Hispanic

/\* AIndian\*/

/\* Asian\*/

Black

White

miss

/\* otob\*/

/\* age\_otob\*/

/\* use\_otob\*/

cigar

room

/\* car\*/

harmful

pharmful

/\* pnharmful\*/

/\* nharmful\*/

live\_smoke

/\* friends\_t\*/

/\* friends\_otob\*/

friends\_use

/\* friends\_unsure\_t\*/

/\* friends\_unsure\_ot\*/

rule\_no

rule\_s

rule\_a

/\* rule\_na\*/

/\* rule\_M\*/

dy\_smoke\_y

/\* py\_smoke\_y\*/

/\* pn\_smoke\_y\*/

dn\_smoke\_y

/\* na\_smoke\_y\*/

dy\_smoke\_bf

/\* py\_smoke\_bf\*/

/\* pn\_smoke\_bf\*/

dn\_smoke\_bf

/\* na\_smoke\_bf\*/

smoke\_cool

/\* say\_no\_m\*//\*\*/

;

**proc** **qlim** data = add\_var;

nloptions maxiter = **500**;

model cig\_ind = &indep / discrete;

output out = logit\_reg proball;

**run**;

/\*proc export data=add\_var\*/

/\* outfile='D:\Applied\Research Project\data.csv'\*/

/\* dbms=csv\*/

/\* replace;\*/

/\*run;\*/

/\*proc logistic descending data = add\_var;\*/

/\* model cig\_ind = &indep / lackfit ctable pprob = 0.5;\*/

/\* output out = logit\_r p=p;\*/

/\*run;\*/

**data** correct1;

set logit\_r;

if p >= **0.5** then

predict = **1**;

else

predict = **0**;

if cig\_ind = predict then

correct = **1**;

else

correct = **0**;

**run**;

**proc** **means** mean data = correct;

class cig\_ind;

var correct;

**run**;

**data** correct;

set logit\_reg;

if Prob1\_cig\_ind < Prob2\_cig\_ind then

predict = **1**;

else

predict = **0**;

if cig\_ind = predict then

correct = **1**;

else

correct = **0**;

**run**;

**proc** **means** mean data = correct;

class cig\_ind;

var correct;

**run**;

**proc** **means** data = clean2;

var cig\_ind;

**run**;

/\*proc corr;\*/

/\*run;\*/

**proc** **freq** data = clean2;

tables cig\_ind;

**run**;

**proc** **means** data = clean2;

var cig\_ind;

**run**;

**proc** **means** data = clean2;

**run**;

**proc** **means** data = clean2;

where cig\_ind = **0**;

**run**;

**proc** **means** data = clean2;

where cig\_ind = **1**;

**run**;

/\*proc means data = clean2;\*/

/\* where grade = 13;\*/

/\*run;\*/

%***ls***(clean2, cig\_ind);

**proc** **qlim** data = add\_var;

nloptions maxiter = **500**;

model cig\_ind = &indep / discrete;

hetero cig\_ind ~ Age / link = linear;

**run**;

**proc** **qlim** data = add\_var;

nloptions maxiter = **500**;

model cig\_ind = &indep / discrete;

hetero cig\_ind ~ agesq / link = linear;

**run**;

**proc** **qlim** data = add\_var;

nloptions maxiter = **500**;

model cig\_ind = &indep / discrete;

hetero cig\_ind ~ Hispanic / link = linear;

**run**;

**proc** **qlim** data = add\_var;

nloptions maxiter = **500**;

model cig\_ind = &indep / discrete;

hetero cig\_ind ~ Black / link = linear;

**run**;

**proc** **qlim** data = add\_var;

nloptions maxiter = **500**;

model cig\_ind = &indep / discrete;

hetero cig\_ind ~ White / link = linear;

**run**;

**proc** **qlim** data = add\_var;

nloptions maxiter = **500**;

model cig\_ind = &indep / discrete;

hetero cig\_ind ~ miss / link = linear;

**run**;

**proc** **qlim** data = add\_var;

nloptions maxiter = **500**;

model cig\_ind = &indep / discrete;

hetero cig\_ind ~ cigar / link = linear;

**run**;

**proc** **qlim** data = add\_var;

nloptions maxiter = **500**;

model cig\_ind = &indep / discrete;

hetero cig\_ind ~ room / link = linear;

**run**;

**proc** **qlim** data = add\_var;

nloptions maxiter = **500**;

model cig\_ind = &indep / discrete;

hetero cig\_ind ~ harmful / link = linear;

**run**;

**proc** **qlim** data = add\_var;

nloptions maxiter = **500**;

model cig\_ind = &indep / discrete;

hetero cig\_ind ~ pharmful / link = linear;

**run**;

**proc** **qlim** data = add\_var;

nloptions maxiter = **500**;

model cig\_ind = &indep / discrete;

hetero cig\_ind ~ live\_smoke / link = linear;

**run**;

**proc** **qlim** data = add\_var;

nloptions maxiter = **500**;

model cig\_ind = &indep / discrete;

hetero cig\_ind ~ friends\_use / link = linear;

**run**;

**proc** **qlim** data = add\_var;

nloptions maxiter = **500**;

model cig\_ind = &indep / discrete;

hetero cig\_ind ~ rule\_no / link = linear;

**run**;

**proc** **qlim** data = add\_var;

nloptions maxiter = **500**;

model cig\_ind = &indep / discrete;

hetero cig\_ind ~ rule\_s / link = linear;

**run**;

**proc** **qlim** data = add\_var;

nloptions maxiter = **500**;

model cig\_ind = &indep / discrete;

hetero cig\_ind ~ rule\_a / link = linear;

**run**;

**proc** **qlim** data = add\_var;

nloptions maxiter = **500**;

model cig\_ind = &indep / discrete;

hetero cig\_ind ~ dy\_smoke\_y / link = linear;

**run**;

**proc** **qlim** data = add\_var;

nloptions maxiter = **500**;

model cig\_ind = &indep / discrete;

hetero cig\_ind ~ dn\_smoke\_y / link = linear;

**run**;

**proc** **qlim** data = add\_var;

nloptions maxiter = **500**;

model cig\_ind = &indep / discrete;

hetero cig\_ind ~ dy\_smoke\_bf / link = linear;

**run**;

**proc** **qlim** data = add\_var;

nloptions maxiter = **500**;

model cig\_ind = &indep / discrete;

hetero cig\_ind ~ dn\_smoke\_bf / link = linear;

**run**;

**proc** **qlim** data = add\_var;

nloptions maxiter = **500**;

model cig\_ind = &indep / discrete;

hetero cig\_ind ~ smoke\_cool / link = linear;

**run**;

/\*random nuber: 19,132, 4,583, 13,176, and 4,364\*/

**proc** **qlim** data = add\_var;

nloptions maxiter = **500**;

model cig\_ind = &indep / discrete;

output out = logit\_reg\_show Marginal proball;

**run**;

**data** marg\_prob;

set logit\_reg\_show;

if \_n\_ = **19132** or \_n\_ = **4583** or \_n\_ = **13176** or \_n\_ = **4364**;

**run**;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

/\*This runs when the include path is ran when the code directly below is ran in the above code\*/

/\*%include "&path.Macros.sas";\*/

/\*proc import datafile = "C:\Users\pwb0029\Desktop\test.csv"\*/

/\* out = test\*/

/\* dbms = csv\*/

/\* replace;\*/

/\*run;\*/

/\*Floor Cap Miss Macro\*/

%Macro FCM(Variable, Dataset, Floor, Cap, Miss);

**Data** &Dataset.clean; /\*this has a . inbetween so that I may append clean to the old dataset name\*/

set &Dataset;

if &Variable = **.** then /\*this need to be checked before floor because . is treated as negative inffinity\*/

&Variable = &miss;

else if &Variable < &Floor then

&Variable = &Floor;

else if &Variable > &Cap then

&Variable = &Cap;

**run**;

**%Mend** FCM;

/\*Example\*/

/\*%FCM(Number, test, 20, 80, 50)\*/

/\*Replaces one value with another\*/

**%Macro** Cln(Variable, replace, replacewith, Dataset);

data &Dataset.clean;

set &Dataset;

if &Variable = &replace then

&Variable = &replacewith;

Run;

**%Mend** Cln;

/\*Example\*/

/\*%Cln(Age\_\_5\_from, '--', 0, Merged3)\*/

/\*Large\_Small Macro idetifies which variables have this problem\*/

**%Macro** LS(Dataset, DepVar); /\*DepVar = Dependent Variable\*/

proc means data = &Dataset NOPRINT;

class &DepVar;

output out = &Dataset.\_all;

run;

data &Dataset.\_all (Drop = \_Type\_);

set &Dataset.\_all;

if (\_STAT\_ = "MIN" or \_STAT\_ = "MAX") and &DepVar ~= **.**;

run;

proc transpose data = &Dataset.\_all

out = &Dataset.\_all;

run;

data &Dataset.\_all; /\*Identifies which variables have the large-small problem and vice-versa\*/

set &Dataset.\_all (rename = (COL1 = MIN COL2 = MAX COL3 = MIN1 COL4 = MAX1));

Large\_small\_I\_ind = **0**;

Large\_small\_O\_ind = **0**;

if MAX < MAX1 then

if MAX < MIN1 then

Large\_small\_I\_ind = **1**;

else

if MAX1 < MIN then

Large\_small\_O\_ind = **1**;

run;

Data &Dataset.\_problem; /\*Creates a new datset with the problemed variables\*/

set &Dataset.\_all;

if Large\_small\_I\_ind = **1** or Large\_small\_O\_ind = **1**;

run;

**%Mend** LS;

1. Mortality in relation to smoking: 50 years' observations on male British doctors [↑](#footnote-ref-1)
2. National Youth Tobacco Survey [↑](#footnote-ref-2)
3. Cigar Smoking Among U.S. Students [↑](#footnote-ref-3)