QMSS – Data Analysis Final Independent Project

Do Smoking Habits Have Effects on Coronary Diseases?

Wenxin Xi

Introduction

For insurance and consulting companies, it is important to decide the factors that influence the cost and the probability of future risk. When people buy same insurance products, they usually pay different premium because of their personal situations. For life & health insurance products, insurance companies care about their gender, age, smoking habits and other factors to decide their premium. There exists a possibility that through observing and controlling insurant behaviors to lower their future risk instead of adding premium for older people. In this situation, insurance company will encounter less claims and insurant can spend less to keep their guarantee. Insurance companies can make dynamic pricing models by observing insurant behaviors. People need to pay higher premium when they are older since older people usually face higher risk to be sick. But there is a question that age is not the only factor deciding the risk. Healthy habits can help old people to be healthier than the young. Also, insurant can also face less risk if they get more healthy habits like sleeping regularly, quitting smoke and heavy drinking.

My research question is exploring the relationship between smoking habits and coronary diseases. Many severe symptoms happened due to coronary heart disease including chest pain (angina), shortness of breath and heart attack. Health insurance products like critical illness insurance usually cover heart attack and any other diseases due to coronary diseases. The relationship between smoking habits and coronary diseases may help insurance companies to make changes on pricing models. Insurance companies could charge different premium based on insurant' smoking habits.

Overall, the hypothesis is that with smoking frequency/time increasing, the risk of coronary disease increases.

Data Set and Variables

Dataset Information

1. **Name**

National Health Interview Survey 2020, CSV data file

2. **Access**

https://www.cdc.gov/nchs/nhis/2020nhis.htm

NHIS is a cross-sectional household interview survey. The target population for the NHIS is the civilian noninstitutionalized population residing within the 50 states and the District of Columbia at the time of the interview.

The U.S. Census Bureau, under a contractual agreement, is the data collection agent for the National Health Interview Survey. NHIS data are collected continuously throughout the year by Census interviewers. Nationally, about 750 interviewers (also called "Field Representatives" or "FRs") are trained and directed by health survey supervisors in the U.S. Census Bureau Regional Offices to conduct interviews for NHIS.

3. **Pros and Cons**

The NHIS is conducted using computer-assisted personal interviewing. Face-to-face interviews are conducted in respondents' homes, but follow-ups to complete interviews may be conducted over the telephone. In 2019, 34.3% of the Sample Adult interviews were conducted at least partially by telephone. The interviewers are trained meaning survey data will be reliable. The data also contains tobacco use information that is what I need to deal with my research question. But tobacco use information is not as quantitative as I expected. The questionnaire has tobacco use information but lacks detailed options.

Key Independent Variables

1. **SMOKENOW:**

   Table 1: table of "SMOKENOW"

Don't Smoke	Smoke Some Days	Smoke Every Day
8142	877	2736

This variable is asking respondents do they smoke cigarettes **now**. I recode SMKNOW_A from the original dataset by using reverse function since I want this variable in the order of increasing frequency range from 1, meaning "don't smoke", to 3, meaning "smoke every day". I also rename the answers for making it more distinct. Moreover, I exclude the answers "refused", "not ascertained" and "don't know".

2. **SMOKESTATUS:**

   Table 2: table of "SMOKESTATUS"

Never smoker	Past Smoker	Current Smoker
19224	8142	3613

This variable is asking about cigarette history. I recode SMKCIGST_A from the original dataset by using reverse function. The answers range from 1, meaning "never smoker", to 3, meaning "current smoker". I also make combinations of the original answers. I combined "current every day smoker" and "current some day smoker" into "current smoker" since I want to use this variable to compare three categories that are current smoker, past smoker and never smoker. I also exclude the meaningless answers "smoker, current status unknown" and "unknown if ever smoked".

3. **SMOKEAGE:**

   Table 3: table of "SMOKEAGE"

under 18	19-30	31-42	above 42
7602	3585	196	63

This variable is asking the age that respondents started smoking regularly. I recode SMKAGE_A from the original dataset by making categories. Since some ages only have few respondents, I make intervals in order to have more individuals in each interval. I exclude the meaningless answers “refused”, “not ascertained” and “don’t know”.

4. **SMOKEAMOUNT:**

Table 4: table of “SMOKEAMOUT”

under 10	11-20	21-30	above 30
3979	2611	472	660

This variable is asking the number of cigarettes that respondents smoke per day. I recode FORNUMCIG_A from the original dataset by making categories. Since some numbers only have few respondents, I make intervals in order to have more individuals in each interval. I exclude the meaningless answers “refused”, “not ascertained” and “don’t know”.

Key dependent variables

1. **CHD:**

Table 5: table of “CHD”

0	1
29586	1901

This variable shows whether respondents have coronary heart disease or not. I recode CHDEV_A from the original dataset by renaming the variable and making it a dummy variable

(1 = having coronary heart disease; 0 = don't have coronary heart disease). I also exclude meaningless answers "refused", "not ascertained" and "don't know".

2. **ANGINA:**

   Table 6: table of "ANGINA"

0	1
30898	592

This variable shows whether respondents have angina or not. I recode ANGEV_A from the original dataset by renaming the variable and making it a dummy variable (1 = having angina; 0 = don't have angina). I also exclude meaningless answers "refused", "not ascertained" and "don't know".

3. **HA:**

   Table 7: table of "HA"

0	1
30362	1168

This variable shows whether respondents have heart attack or not. I recode MIEV_A from the original dataset by renaming the variable and making it a dummy variable (1 = having heart attack; 0 = don't have heart attack). I also exclude meaningless answers "refused", "not ascertained" and "don't know".

4. **STROKE:**

   Table 8: table of "stroke"

0	1
30478	1060

This variable shows whether respondents have a stroke or not. I recode STREV_A from the original dataset by renaming the variable and making it a dummy variable (1 = having stroke; 0 = don't have stroke). I also exclude meaningless answers "refused", "not ascertained" and "don't know".

Controlled Variables

Since alcohol use, gender and diabetes are also risk factors for coronary disease, I choose them as controlled variables. Men are generally at greater risk of coronary artery disease. However, the risk for women increases after menopause. Diabetes is associated with an increased risk of coronary artery disease. Type 2 diabetes and coronary artery disease share similar risk factors, such as obesity and high blood pressure. Heavy alcohol use can lead to heart muscle damage. It can also worsen other risk factors of coronary artery disease.[1]

1. **ALCOHOL**

   Table 9: table of "ALCOHOL"

abstainer	former drinker	current drinker
3343	5967	21519

This variable is asking about alcohol use history. I recode DRKSTAT_A from the original dataset by making categories. The answers range from 1, meaning "abstainer", to 3, meaning "current drinker". I exclude the answer "drinking status unknown".

2. **FEMALE**

[1] Coronary artery disease, https://www.mayoclinic.org/diseases-conditions/coronary-artery-disease/symptoms-causes/syc-20350613

Table 10: table of "FEMALE"

0	1
14521	17045

This variable means female respondents. I recode SEX_A from the original dataset by making it into a dummy variable (1 = female; 0 = male). I also exclude meaningless answers "refused", "not ascertained" and "don't know".

3. **DIABETES**

Table 10: table of "DIABETES"

0	1
28180	3356

This variable shows whether respondents have diabetes or not. I recode DIBEV_A from the original dataset by renaming the variable and making it a dummy variable (1 = having diabetes; 0 = don't have diabetes). I also exclude meaningless answers "refused", "not ascertained" and "don't know".

Descriptive Statistics

Table 11 shows the descriptive information of variables I have used for my research. The information of dummy variables is on the top side.

The observations are 31568 but they all have missing values. For dummy variables, the probability of getting those diseases is very low that is around 0.04 (mean = 0.04). But the overall observations are sufficient that the number of respondents who have coronary are enough

for me to work on my research. For CHD, skewness equals 3.7, the distribution is highly positively skewed; the kurtosis is 12, showing that the distribution's tails are longer and fatter than normal distribution.

From the summary of categorical variables, as we can see, people who don't smoke take the largest part. For SMOKESTATUS, 62.1% respondents don't smoke and only 11.7% respondents are current smokers.

Table 11: descriptive data of variables

	n	miss	p.miss	mean	sd	median	p25	p75	min	max	skew	kurt
chd	31568	81	0.257	0.06	0.2	0	0	0	0	1	3.7	12
angina	31568	78	0.247	0.02	0.1	0	0	0	0	1	7.1	48
ha	31568	38	0.120	0.04	0.2	0	0	0	0	1	4.9	22
stroke	31568	30	0.095	0.03	0.2	0	0	0	0	1	5.2	25
female	31568	2	0.006	0.54	0.5	1	0	1	0	1	-0.2	-2
diabetes	31568	32	0.101	0.11	0.3	0	0	0	0	1	2.6	5

Summary of categorical variables

strata: Overall

var	n	miss	p.miss	level	freq	percent	cum.percent
smokenow	31568	19813	62.8	Don't Smoke	8142	69.3	69.3
				Smoke Some Days	877	7.5	76.7
				Smoke Every Day	2736	23.3	100.0
smokestatus	31568	589	1.9	Never smoker	19224	62.1	62.1
				Past Smoker	8142	26.3	88.3
				Current Smoker	3613	11.7	100.0
smokeage	31568	20122	63.7	under 18	7602	66.4	66.4
				19-30	3585	31.3	97.7
				31-42	196	1.7	99.4
				above 42	63	0.6	100.0
smokeamount	31568	23846	75.5	under 10	3979	51.5	51.5
				11-20	2611	33.8	85.3
				21-30	472	6.1	91.5
				above 30	660	8.5	100.0
alcohol	31568	739	2.3	abstainer	3343	10.8	10.8
				former drinker	5967	19.4	30.2
				current drinker	21519	69.8	100.0

Models

Multiple Linear Probability Model

Since my dependent variables are dummy variables, at first, I choose multiple linear probability model in my research.

$$P = \alpha+\beta_1X_1+\beta_2X_2+\mu \text{ (linear)}$$

The major advantage of the linear model is its interpretability. In the linear model, if the coefficient is 0.03, that means that a one-unit increase in X is associated with a 3percentage points increase in the probability that Y is 1. Just about everyone has some understanding of what it would mean to increase by 3 percentage points their probability of outcome is 1.

CHD:

Table 12: Linear Probability Result 1

```
Call:
lm(formula = chd ~ as.numeric(smokestatus), data = sub)

Residuals:
     Min       1Q   Median       3Q      Max
-0.09410 -0.07153 -0.04896 -0.04896  0.95104

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.026388   0.003203   8.239   <2e-16 ***
as.numeric(smokestatus) 0.022572   0.001942  11.622   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2372 on 30907 degrees of freedom
  (659 observations deleted due to missingness)
Multiple R-squared:  0.004351,  Adjusted R-squared:  0.004319
F-statistic: 135.1 on 1 and 30907 DF,  p-value: < 2.2e-16
```

Before applying multiple variables into model, I use linear probability model first to find the relationship between SMOKESTATUS and CHD. From table 12, it shows that, each one level

up in SMOKESTATUS, a respondent is about 2.26 percentage points more likely to have coronary disease (CHD == 1). It is statistically significant since p < 0.001. But the adjusted R-sq is quite low which means I can only explain about 0.43 percent of the variation in relation between SMOKESTATUS and CHD.

Table 13: Multiple Linear Probability Result 1

```
Call:
lm(formula = chd ~ as.numeric(smokestatus) + as.numeric(alcohol) +
    diabetes, data = sub)

Residuals:
     Min       1Q   Median       3Q      Max
-0.23309 -0.06220 -0.04563 -0.02906  0.97094

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.056921   0.005815   9.788   <2e-16 ***
as.numeric(smokestatus)  0.021844   0.001936  11.281   <2e-16 ***
as.numeric(alcohol)     -0.016569   0.001993  -8.314   <2e-16 ***
diabetes                 0.127206   0.004358  29.192   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2335 on 30705 degrees of freedom
  (859 observations deleted due to missingness)
Multiple R-squared:  0.03531,   Adjusted R-squared:  0.03522
F-statistic: 374.7 on 3 and 30705 DF,  p-value: < 2.2e-16
```

Then I add two controlled variables ALCOHOL and DIABETES.

Net of ALCOHOL and DIABETES, each one level up in SMOKESTATUS, a respondent is about 2.18 percentage points more likely to have coronary disease (CHD == 1). It is statistically significant since p < 0.001. But the adjusted R-sq is still very low that is 0.03522 which means I can only explain about 3.5 percent of the variation in relation between SMOKESTATUS and CHD.

Table 14

	Dependent variable:	
	chd	
	(1)	(2)
as.numeric(smokestatus)	0.023***	0.022***
	(0.002)	(0.002)
as.numeric(alcohol)		-0.017***
		(0.002)
diabetes		0.127***
		(0.004)
Constant	0.026***	0.057***
	(0.003)	(0.006)
Observations	30,909	30,709
R2	0.004	0.035
Adjusted R2	0.004	0.035
Residual Std. Error	0.237 (df = 30907)	0.234 (df = 30705)
F Statistic	135.079*** (df = 1; 30907)	374.662*** (df = 3; 30705)
Note:		*p<0.1; **p<0.05; ***p<0.01

Additional variables don't have much effect on coefficient of SMOKESTATUS. As we can see, P-value is quite small in multiple linear probability models. However, normality of errors assumption is violated in this model, so p-values are questionable. I will try an alternative model later.

Angina:

Table 15: Multiple Linear Probability Result 2

```
Call:
lm(formula = angina ~ as.numeric(smokestatus) + as.numeric(alcohol) +
    diabetes, data = sub)

Residuals:
     Min       1Q   Median       3Q      Max
-0.07890 -0.02176 -0.01708 -0.00780  0.99220

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.012571   0.003369   3.732 0.000191 ***
as.numeric(smokestatus)  0.009277   0.001121   8.276  < 2e-16 ***
as.numeric(alcohol)     -0.004683   0.001155  -4.056 5.01e-05 ***
diabetes                 0.043181   0.002525  17.104  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1352 on 30708 degrees of freedom
  (856 observations deleted due to missingness)
Multiple R-squared:  0.01295,   Adjusted R-squared:  0.01285
F-statistic: 134.3 on 3 and 30708 DF,  p-value: < 2.2e-16
```

Net of ALCOHOL and DIABETES, each one level up in SMOKESTATUS, a respondent is about 0.93 percentage points more likely to have angina (angina == 1). It is statistically significant since $p < 0.001$. But the adjusted R-sq is quite low which means I can only explain about 1.2 percent of the variation in relation between SMOKESTATUS and ANGINA.

Logistic Regression Model

$$\textbf{Log [P (CHD = 1)/P (CHD=0)] = } \alpha + \beta X1 + \beta X2 + \mu$$

Since multiple linear probability model has some limitations, I choose logistic regression model to compare and contrast. Moreover, log transformation in Logit Model helps to make the dependent variables more “normal” since dependent variables are highly positively skewed. The regression coefficient for each independent variable measures the effect of a one unit change in that variable on the logit (log-odds) of the dependent variable. It is used to understand the relationship between the dependent variable and one or more independent variables by

estimating probabilities using a logistic regression equation. This model can help me predict the likelihood of an event happening or a choice being made.

CHD:

Table 16: Logistic Regression Model Result 1

```
Call:
glm(formula = chd ~ as.numeric(smokestatus) + as.numeric(alcohol) +
    diabetes, family = binomial, data = sub)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9263  -0.3451  -0.2999  -0.2604   2.6081

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.87675    0.09751 -29.501   <2e-16 ***
as.numeric(smokestatus)  0.37328    0.03286  11.358   <2e-16 ***
as.numeric(alcohol)   -0.28792    0.03415  -8.432   <2e-16 ***
diabetes               1.42073    0.05480  25.926   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13964  on 30708  degrees of freedom
Residual deviance: 13141  on 30705  degrees of freedom
  (859 observations deleted due to missingness)
AIC: 13149

Number of Fisher Scoring iterations: 6
```

From table 16, the coefficient for SMOKESTATUS is 0.37328. For each category increase in SMOKESTATUS, a respondent increases the logit by 0.373 of having coronary disease, net of ALCOHOL and DIABETES.

Table 17: Odds Ratios Result 1

```
(Intercept) as.numeric(smokestatus)     as.numeric(alcohol)                diabetes
 0.05631765              1.45248443              0.74982182              4.14015574
```

An odds ratio (OR) is a measure of association between an exposure and an outcome. The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.

- OR=1 Exposure does not affect odds of outcome
- OR>1 Exposure associated with higher odds of outcome
- OR<1 Exposure associated with lower odds of outcome

From table 17, I have odd ratio for three variables.

The odd ratio for SMOKESTATUS is about 1.45. For each one level up in the respondent's smoke status, on average, increases a respondent's odds of having coronary disease by 1.45, net of ALCOHOL and DIABETES. The odds of having coronary disease always go up by 45% when the respondents' smoke status one level increased, net of other variables, because (1.45-1) *100% = 45%. This is a proportionate increase, not an absolute increase.

Angina:

Table 18: Logistic Regression Model Result 2

```
Call:
glm(formula = angina ~ as.numeric(smokestatus) + as.numeric(alcohol) +
    diabetes, family = binomial, data = sub)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5468  -0.1983  -0.1752  -0.1397   3.0446

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -4.33202    0.17198 -25.190  < 2e-16 ***
as.numeric(smokestatus)  0.45621    0.05521   8.263  < 2e-16 ***
as.numeric(alcohol)   -0.24979    0.05982  -4.176 2.97e-05 ***
diabetes               1.38854    0.09156  15.166  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5753.5  on 30711  degrees of freedom
Residual deviance: 5459.2  on 30708  degrees of freedom
  (856 observations deleted due to missingness)
AIC: 5467.2

Number of Fisher Scoring iterations: 7
```

From table 18, the coefficient for SMOKESTATUS is 0.45621. For each category increase in SMOKESTATUS, a respondent increases the logit by 0.456 of having angina, net of ALCOHOL and DIABETES.

Table 19: Odds Ratios Result 2

(Intercept)	as.numeric(smokestatus)	as.numeric(alcohol)	diabetes
0.01314096	1.57807427	0.77896274	4.00897332

The odd ratio for SMOKESTATUS is about 1.58. For each one level up in the respondent's smoke status, on average, increases a respondent's odds of having angina by 1.58, net of ALCOHOL and DIABETES. The odds of having angina always go up by 58% when the respondents' smoke status one level increased, net of other variables, because (1.58-1) *100% = 58%. This is a proportionate increase, not an absolute increase.

HA:

Table 20: Logistic Regression Model Result 3

```
Call:
glm(formula = ha ~ as.numeric(smokeage) + as.numeric(smokeamount) +
    as.numeric(alcohol) + diabetes, family = binomial, data = sub)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0833  -0.3866  -0.2908  -0.2518   2.6334

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.03466    0.27320  -7.448 9.51e-14 ***
as.numeric(smokeage)   0.11642    0.09090   1.281      0.2
as.numeric(smokeamount) 0.29308   0.04599   6.372 1.86e-10 ***
as.numeric(alcohol)   -0.60349    0.07909  -7.631 2.34e-14 ***
diabetes               1.12391    0.10326  10.884  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3710.9  on 7614  degrees of freedom
Residual deviance: 3456.8  on 7610  degrees of freedom
  (23953 observations deleted due to missingness)
AIC: 3466.8

Number of Fisher Scoring iterations: 6
```

I want to find relationship between SMOKEAGE&SMOKEAMOUNT and HA.

The Hypothesis is that the younger the respondents started to smoke, the risk of having a heart attack is higher. From table 20, the coefficient of SMOKEAGE is 0.11642 representing each level up in SMOKEAGE (older group) increase the logit by about 0.12 of the respondents having a heart attack, net of other factors. This is not consistent with my hypothesis. There may exist other variable like smoking years instead of smoking age will be consistent with what I hypothesized.

I hypothesize that the larger amount the respondents smoke, the risk of having a heart attack is higher. From table 20, the coefficient of SMOKEAMOUNT is 0.29308 representing each level up in SMOKEAMOUNT (larger smoke amount) increase the logit by about 0.29 of having a heart attack, net of SMOKEAGE, ALCOHOL and DIABETES. It is consistent with my hypothesis.

Table 21: Odds Ratios Result 3

(Intercept)	as.numeric(smokeage)	as.numeric(smokeamount)	as.numeric(alcohol)	diabetes
0.1307248	1.1234668	1.3405510	0.5469011	3.0768658

The odd ratio for SMOKEAMOUNT is about 1.34. For each one level up in the respondent's smoke amount, on average, increases a respondent's odds of having heart attack by 1.34, net of other factors. The odds of having heart attack always go up by 34% when the respondents' smoke amount one level increased, net of other variables, because (1.34-1) *100% = 34%. This is a proportionate increase, not an absolute increase.

STROKE:

Table 22: Logistic Regression Model Result 4

```
Call:
glm(formula = stroke ~ as.numeric(smokenow) + as.numeric(alcohol) +
    diabetes + female, family = binomial, data = sub)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6600  -0.3408  -0.2640  -0.2589   2.6126

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.80437    0.21314  -8.466  < 2e-16 ***
as.numeric(smokenow)  0.03969    0.05115   0.776    0.438
as.numeric(alcohol)  -0.52607    0.07139  -7.369 1.72e-13 ***
diabetes              0.79798    0.10248   7.786 6.89e-15 ***
female               -0.03646    0.08766  -0.416    0.677
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4491.9  on 11648  degrees of freedom
Residual deviance: 4369.0  on 11644  degrees of freedom
  (19919 observations deleted due to missingness)
AIC: 4379

Number of Fisher Scoring iterations: 6
```

I want to explore the relationship between SMOKENOW and STROKE.

From table 22, it shows that each level up in SMOKENOW (smoke more frequently) increase the logit by about 0.039 of the respondents having a stroke, net of other factors. This is consistent with my hypothesis that smoking more frequently will cause higher risk of having a stroke. But the p-value is extremely high ($p > 0.05$) that it is not statistically significant.

Conclusion

In my research question, because of the limitations of multiple linear probability models, Logit Model is my best choice. First of all, normality of errors assumption is violated in LPM model, so p-values are questionable. Secondly, expected values can be out of the legitimate range of 0-1.

Moreover, homoskedasticity is violated, since the variance of the errors is determined by the percent who are 1s, within each category of X.

The disadvantage of logistic model is that it lacks interpretability but log transformation in the analysis help to make the variable more “normal”. Because the log odds scale is so hard to interpret, it is common to report logistic regression results as odds ratios. Odds ratios are used to compare the relative odds of the occurrence of the outcome of interest, given exposure to the variable of interest. The odds ratio can also be used to determine whether a particular exposure is a risk factor for a particular outcome, and to compare the magnitude of various risk factors for that outcome. Moreover, odds ratios are most commonly in case-control studies that are consistent with my research.

In this case, I will choose Logistic Regression Model.

Results from Logistic Regression Model support my hypothesis that smokers will face higher risk of having coronary diseases. I use different variables related to smoking habits and coronary diseases including specific diseases due to coronary diseases to find the relationship. Most of results support my hypothesis.

I want to explore the daily behaviors and other diseases more detailed in the future. This work is the part of what I want to research. My thesis will cover more various daily behaviors including smoking, drinking and etc. Also, I will also include more kinds of diseases in my research.