Analysis of smoking habits in the UK

Group member's contribution: Questions, Data Description, and Conclusion- Jiaxuan Li; Logistic Regression and Result - Yiran Jia; Logistic Regression Using Best Subset Selection Method - Tianzi Liu; Linear Regression and Result - Jiayin Sha; Hypothesis Test - Xuening Bai

Abstract

The global health problem has always been a topic of great concern to people, and the diseases induced by smoking are gradually increasing. Out of this motivation, we will study the demographic characteristics of smokers and how each factor affects the amount of tobacco they consume. We will use a cross-sectional dataset containing 1691 observations from a smoking survey in the United Kingdom, which includes 12 variables such as the age of the smoker, the number of cigarettes smoked per day, gender, race, nationality, etc. We will use several methods such as Linear Regression, Confidence Interval, Tree regression, and Hypothesis Test to evaluate the relationship between age and smoking frequency, the average age of Caucasian smokers, the average amount of cigarettes smoked per day on weekdays, etc. based on the sample data from the UK.

Questions/Hypotheses

In this research paper, we will explore several questions/hypotheses.

1. **Does the age of a smoker correlate with their smoking habits?**
   We expect that older smokers might smoke more cigarettes due to longer addiction periods, whereas younger smokers might smoke fewer cigarettes due to increased health awareness.
2. **Does the gender of a smoker correlate with their smoking habits?**
   We expect that female smokers consume fewer cigarettes per day compared to male smokers due to differences in smoking habits, social norms, or biological factors between genders.
3. **How does race or ethnicity influence smoking patterns and cigarette consumption?**
   We expect that there might be variations in smoking behaviors across different racial or ethnic groups, influenced by cultural practices, socioeconomic factors, or targeted marketing by tobacco companies.
4. **Is there a relationship between income level and the quantity of tobacco consumption?**
   We expect that individuals with lower income levels might consume more tobacco due to stress-related factors, whereas higher-income individuals might have more resources to quit or reduce smoking.

Data Description

The smoking habits of UK smokers are collected from some surveys and upload on the website OpenIntro (https: //www.openintro.org/data/index.php?data=smoking). This data could be used for analyzing the demographic characteristics of smokers and how each factor affects the amount of tobacco they consume. The data set has 1691 observations on the following 12 variables:

1. **gender**: Gender with levels of Female and Male.
2. **age**
3. **marital_status**: Marital status with levels Divorced, Married, Separated, Single and Widowed.
4. **highest_qualification**: Highest education level with levels A Levels, Degree, GCSE/CSE, GCSE/O Level, Higher/Sub Degree, No Qualification, ONC/BTEC and Other/Sub Degree
5. **nationality**: Nationality with levels British, English, Irish, Scottish, Welsh, Other, Refused and Unknown
6. **ethnicity**: Ethnicity with levels Asian, Black, Chinese, Mixed, White and Refused Unknown.
7. **gross_income**:
   Gross income with levels Under 2,600, 2,600 to 5,200, 5,200 to 10,400, 10,400 to 15,600, 15,600 to 20,800, 20,800 to 28,600, 28,600 to 36,400, Above 36,400, Refused and Unknown.
8. **region**: Region with levels London, Midlands & East Anglia, Scotland, South East, South West, The North and Wales
9. **smoke**: Smoking status with levels No and Yes
10. **amt_weekends**: Number of cigarettes smoked per day on weekends.
11. **amt_weekdays**: Number of cigarettes smoked per day on weekdays.
12. **type**: Type of cigarettes smoked with levels Packets, Hand-Rolled, Both/Mainly Packets and Both/Mainly Hand-Rolled

Methodologies

Logistic Regression

We use the Logistic Model to estimate the impact of individuals' demographic characteristics, such as gender, marital status, income level, etc, on their decision of whether to smoke. The estimation equation is as follows:

$$smoke_i = \frac{exp(\beta_0 + \beta_1 gender_i + \beta_2 age_i + \beta_3 income_i + \beta_4 single_i + \beta_5 married_i + \beta_6 asian_i + \beta_7 black_i + \beta_8 chinese_i + \beta_9 mixed_i + \beta_{10} white_i + \varepsilon_i)}{1 + exp(\beta_0 + \beta_1 gender_i + \beta_2 age_i + \beta_3 income_i + \beta_4 single_i + \beta_5 married_i + \beta_6 asian_i + \beta_7 black_i + \beta_8 chinese_i + \beta_9 mixed_i + \beta_{10} white_i + \varepsilon_i)}$$

where i denotes each individual; $smoke_i$ is a binary variable which equals 1 if the individuals have a smoking habit and 0 when they do not smoke; $gender_i$ equals 1 for males and 0 for females; in the original dataset, each individual belongs to a category of a certain income range, and to better explore the qualitative relationship, we assume that each person $income_i$ is equal to the median of the bracket they belong to; $single_i$, $married_i$, $asian_i$, $black_i$, $chinese_i$, $mixed_i$ and $white_i$ are binary variables, representing the individual's marital status and ethnicity.

Logistic Regression Using Best Subset Selection Method

Considering that using all variables for model fitting may lead to a smaller bias, but large variance when solving realistic problems, we followed up with the best subset selection method to fit the model. Using subset selection to choose a smaller number of predictors from the whole set of predictors can improve prediction accuracy. This is because an appropriately smaller model can reduce the prediction error. The purpose of this section is to find the model that has the highest rate of correctness in determining whether a person is a smoker or not.

Linear Regression

The equation of simple linear regression is:

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

Samples in variable 'age' and samples in variable 'amt_weekdays' are both independent random variables. The Simple Linear Regression model is used to model the association between two quantitative random variables. Specifically, we are going to find the association between Caucasian smokers' age and number of cigarettes smoked per day on weekdays in the UK.

Hypothesis Test

Hypothesis test could check whether our hypothesis meets the condition we state. It needs to state two hypotheses which are null hypothesis H_0 and alternative hypothesis H_a. Then calculate the test statistic from sample data and simulate samples assuming H_0 it is true and calculate the statistic for each sample. Evaluate the evidence H_0 by calculating the p-value. Specifically, we assume that $H_0: \mu = 15$ and $H_a: \mu < 15$ [1].

Result

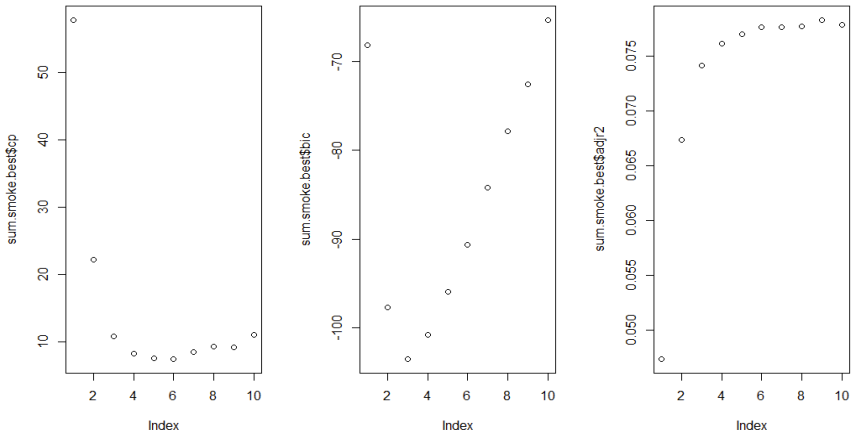
Logistic Regression

As Table 1 (see Appendix) shows, age, income, and marital status have the most significant impact on people's choice of whether to smoke. On the contrary, ethnicity has little effect.

Males are more likely to smoke than females at the 1% significance level. The probability of people smoking is significantly negatively influenced by age and income levels. For marital status, married people have the least probability of smoking, followed by single people. People with 'other' marital status, such as divorced or widowed are more likely to smoke than others.

Logistic Regression Using Best Subset Selection Method

After using the best subset selection method, we used ***cp, bic, and adjr2*** as measures to derive the model.



Performance of models with different numbers of variables

when measuring them using different criteria

Detailed calculations can be found in the Appendix.

(1)Model considering 6 variables which minimizes cp

We choose *gender, age, gross_income, single, married, and asian* as variables to build a model.

```
> coef(smoke.best,6)
 (Intercept)       gender          age gross_income       single      married        asian
   6.475e-01    5.131e-02   -5.579e-03   -4.023e-06   -4.914e-02   -1.394e-01   -1.133e-01
```

(2)Model considering 3 variables which minimizes bic

We choose *age, gross_income, and married* as variables to build a model.

```
> coef(smoke.best,3)
 (Intercept)          age gross_income      married
   6.029e-01   -4.988e-03   -3.296e-06   -1.164e-01
```

(3)Model considering 9 variables which maximizes adjr2

We choose *gender, age, gross_income, single, married, asian, black, chinese, and white* as variables to build a model.

```
coef(smoke.best,9)
(Intercept)        gender           age gross_income        single       married         asian
  7.678e-01     5.281e-02    -5.615e-03    -4.178e-06    -4.855e-02    -1.391e-01    -2.315e-01
       black       chinese         white
 -1.869e-01    -1.952e-01    -1.164e-01
```

After building the above three models, we calculated the error rate of these models and compared it with the error rate of the model constructed by considering only *age* as variable and considering all the variables.

Here are the results:

(a): Error rate when using all 10 variables to build the model

(b): Error rate when considering 6 variables which minimize cp

(c): Error rate when considering 3 variables which minimize bic

(d): Error rate when 9 variables which maximizes adjr2

(e): Error rate when only considering *age* as a variable

```
             smoke
glm.all.pred    0    1
           0 1222  388
           1   48   33
> mean(glm.all.pred==smoke)
[1] 0.7422
> mean(glm.all.pred!=smoke)
[1] 0.2578
```

(a)

```
> table(glm.6.pred,smoke)
           smoke
glm.6.pred    0    1
         0 1228  391
         1   42   30
> mean(glm.6.pred==smoke)
[1] 0.7439
> mean(glm.6.pred!=smoke)
[1] 0.2561
```

(b)

```
> table(glm.3.pred,smoke)
           smoke
glm.3.pred    0    1
         0 1227  392
         1   43   29
> mean(glm.3.pred==smoke)
[1] 0.7428
> mean(glm.3.pred!=smoke)
[1] 0.2572
```

(c)

```
> table(glm.9.pred,smoke)
           smoke
glm.9.pred    0    1
         0 1223  387
         1   47   34
> mean(glm.9.pred==smoke)
[1] 0.7433
> mean(glm.9.pred!=smoke)
[1] 0.2567
```

(d)

```
> table(glm.onlyage.pred,smoke)
                smoke
glm.onlyage.pred    0    1
              0 1270  421
> mean(glm.onlyage.pred==smoke)
[1] 0.751
> mean(glm.onlyage.pred!=smoke)
[1] 0.249
```

(e)

The observation shows that the model constructed by considering all the 10 variables has the lowest accuracy of ***0.7422***. While the accuracy of all three models constructed by utilizing the best subset

selection method is slightly improved. Interestingly, when only one variable *age* is considered to construct the model, the model has the highest accuracy of ***0.751***. This shows that *age* has a significant effect on determining whether a person is a smoker or not.

Linear Regression

From Table 5 we know that for one unit increase in age, the average number of cigarettes smoked per day on weekdays by smokers is expected to increase by 0.3035146 which is the slope of the line.

Table 6 summarizes the age of smokers in the UK, with an average age is 42.71 years. The youngest smoker is 16 years and the oldest smoker is 93 years. Age of middle 50% smokers are between 30 to 54 years.

Table 6 summarizes the number of cigarettes smoked per day from Monday to Friday by Caucasian smokers in the UK. On average, one smoker smokes 13.75 cigarettes per day. Smokers might smoke 0 cigarettes while the highest amount would smoke 55 cigarettes per day. Middle 50 percent of smokers smoke 7 to 20 cigarettes per day on weekdays.

Hypothesis Test

We assume that $H_0: \mu = 15$ and $H_a: \mu < 15$. The null hypothesis $H_0: \mu = 15$ means we predicate that the average frequency of Caucasian smokers should be 15 cigarettes daily on weekdays. In contrast, the alternative hypothesis $H_a: \mu < 15$ predicates that the average number of cigarettes should be less than 15 instead of 15. According to some research, the highest group of average consumption of cigarettes by regular smokers was 10 to 19 in 2020[1]. Additionally, the average number of cigarettes had a falling trend since 1997[3] might be due to the increased sense of health. In this way, we take 15 as an indicator to estimate the average consumption of cigarettes should less than 15.

The p-value we get is 0.00329 (see table 7 in Appendix) so we have weak evidence against the average consumption of cigarettes is 15. In other words, the true average number of cigarettes per day on a weekday is less than 15.

Conclusion

In conclusion, this study has provided significant insights into the demographic characteristics of smokers in the United Kingdom and their patterns of tobacco consumption. Our analyses, encompassing logistic regression, linear regression, and hypothesis testing, have elucidated several key findings.

Age, income, and marital status emerged as significant factors influencing smoking habits. Notably, younger individuals were more likely to smoke, contradicting our hypotheses. Those with lower income levels were more likely to smoke, aligning with our hypothesis that financial stress contributes to higher smoking rates.

Gender also played a crucial role, with males showing a higher propensity to smoke than females. This aligns with our hypothesis that social norms and biological factors might influence smoking habits differently across genders.

Interestingly, ethnicity had a less pronounced impact on smoking habits than anticipated. This finding suggests that while cultural factors may influence smoking initiation, other factors like socioeconomic status and personal circumstances might have a more direct impact on continued tobacco use.

Our study's methodology, particularly the use of best subset selection in logistic regression, enhanced the precision of our models. This approach allowed us to identify the most influential variables, leading to more accurate predictions.

The study's hypothesis testing revealed that the average number of cigarettes smoked per weekday is less than previously estimated, indicating a potential shift in smoking patterns or the effectiveness of public health campaigns.

These findings have significant implications for public health efforts to reduce smoking prevalence. Personalized interventions that address the identified demographic characteristics may be more effective in reducing tobacco consumption. Furthermore, the study emphasizes the importance of continual monitoring of smoking behaviors to understand emerging trends and guide policy decisions.

Appendix

Table 1 Result of Logistic Regression

```
Call:
glm(formula = smoke ~ ., family = binomial, data = smoking_log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4106  -0.7697  -0.5854  -0.3024   2.2804

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.995e+00  6.681e-01   2.986  0.00283 **
gender        3.008e-01  1.264e-01   2.381  0.01729 *
age          -3.283e-02  3.998e-03  -8.214  < 2e-16 ***
gross_income -2.481e-05  5.940e-06  -4.176 2.96e-05 ***
single       -3.777e-01  1.852e-01  -2.039  0.04146 *
married      -8.174e-01  1.518e-01  -5.385 7.22e-08 ***
asian        -1.503e+00  7.255e-01  -2.072  0.03827 *
black        -1.209e+00  7.262e-01  -1.665  0.09585 .
chinese      -1.383e+00  7.880e-01  -1.755  0.07922 .
mixed        -4.085e-01  8.295e-01  -0.493  0.62235
white        -8.270e-01  5.949e-01  -1.390  0.16447
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1898.0  on 1690  degrees of freedom
Residual deviance: 1753.6  on 1680  degrees of freedom
AIC: 1775.6

Number of Fisher Scoring iterations: 4
```

Table 2 Performance of models with different numbers of variables when measuring them using different criteria

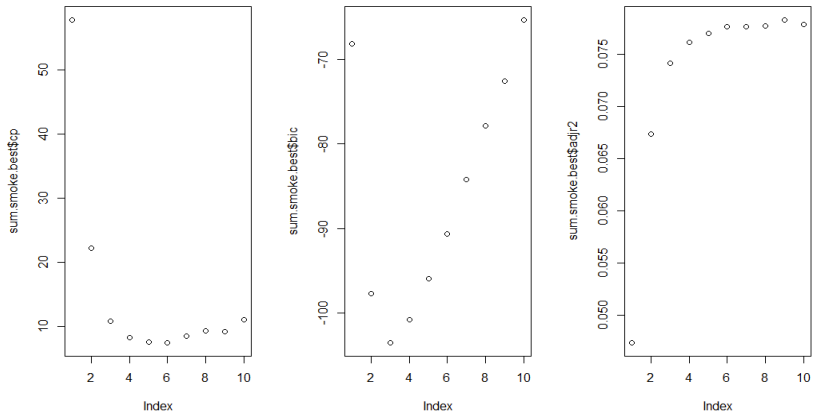


Table 3 Number of variables that maximize model's performance

```
> which.max(sum.smoke.best$adjr2)
he index of maximum value
[1] 9
> which.min(sum.smoke.best$cp)
[1] 6
> which.min(sum.smoke.best$bic)
[1] 3
```

Table 4 Variables chosen when using the best subset selection method considering different numbers of variables

```
Selection Algorithm: exhaustive
          gender age gross_income single married asian black
1  ( 1 )  " "    "*" " "          " "    " "     " "   " "
2  ( 1 )  " "    "*" " "          " "    "*"     " "   " "
3  ( 1 )  " "    "*" "*"          " "    "*"     " "   " "
4  ( 1 )  "*"    "*" "*"          " "    "*"     " "   " "
5  ( 1 )  "*"    "*" "*"          " "    "*"     "*"   " "
6  ( 1 )  "*"    "*" "*"          "*"    "*"     "*"   " "
7  ( 1 )  "*"    "*" "*"          "*"    "*"     "*"   " "
8  ( 1 )  "*"    "*" "*"          " "    "*"     "*"   "*"
9  ( 1 )  "*"    "*" "*"          "*"    "*"     "*"   "*"
10  ( 1 ) "*"    "*" "*"          "*"    "*"     "*"   "*"
          chinese mixed white
1  ( 1 )  " "     " "   " "
2  ( 1 )  " "     " "   " "
3  ( 1 )  " "     " "   " "
4  ( 1 )  " "     " "   " "
5  ( 1 )  " "     " "   " "
6  ( 1 )  " "     " "   " "
7  ( 1 )  "*"     " "   " "
8  ( 1 )  "*"     " "   "*"
9  ( 1 )  "*"     " "   "*"
10  ( 1 ) "*"     "*"   "*"
```

Variables chosen when using the best subset selection method considering different numbers of variables

Table 5 Linear model performance

```
Call:
lm(formula = amt_weekdays ~ age, data = smoking)

Residuals:
    Min      1Q  Median      3Q     Max
-16.579  -6.579  -1.545   5.546  38.645

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.97238    1.27033   7.063 6.82e-12 ***
age          0.11186    0.02782   4.022 6.86e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 6 Summary of age and the number of cigarettes smoked per day from Monday to Friday by Caucasian smokers in the UK

```
> summary(smoking_jy$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.00   30.00   40.00   42.71   54.00   93.00
> summary(smoking_jy$amt_weekdays)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    7.00   12.00   13.75   20.00   55.00
```

Table 7 P-value of Hypothesis Test

```
> p_value
[1] 0.003294
```

Bibliography

1. "Smoking, Drinking and Drug Use among Young People in England, 2021." *NHS Choices*, NHS, 6 Sept. 2022, digital.nhs.uk/data-and-information/publications/statistical/smoking-drinking-and-drug-use-among-young-people-in-england/2021/part-1-smoking-prevalence-and-consumption.
2. *UK smoking data* (no date)*Data Sets*. Available at: https://www.openintro.org/data/index.php?data=smoking.
3. "Health Survey for England 2019 [NS]." *NHS Choices*, NHS, 15 Dec. 2020, digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england/2019.