# Lab 3:  Assessing the Impact of Smoking on Cholesterol

W203: Statistics for Data Science
Date: 2024 April 12
Team Go Bears: Tejasvi Kalakota, Ayushi, Jorge

Data Set Links: https://biolincc.nhlbi.nih.gov/studies/framcohort/
https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset

**Contents**

## 1.1 Introduction

Heart Disease is one of the most debilitating and prevalent ailments plaguing the United States. The CDC reports that heart disease is the leading cause of death across all gender and most ethnic and racial groups in the United States (CDC, 2022). Behavioral Health factors like smoking and alcohol consumption are considered cardiovascular risk factors, but the exact causal relationship between cardiovascular disease between these factors is still not fully understood. In this study we primarily focus on cholesterol. Cholesterol is a thick waxy substance found in tissues throughout the body. Cholesterol serves valuable functions in the body and is essential for numerous cellular processes. However, excessive cholesterol levels in the body can lead to build up of plaque in the arteries leading to an increased risk of heart disease (Goldstein , 2015).  In experimental settings smoking has been documented to alter cholesterol distributions within the body (Rao, 2013). Smoking and high cholesterol levels have a synergistic effect on cardiovascular risk. By modeling cholesterol level among smokers we can help quantify the combined impact of these risk factors on cardiovascular health and identify individuals at higher risk. We will then proceed to develop a second statistical model documenting cholesterol levels

## 1.2 Data Description

### Data Set:

The Framingham Heart Study was a longitudinal cohort study started in 1948 by the National Heart, Lung, and Blood Institute, and it has been pivotal in transforming our understanding of cardiovascular health (Hong, 2023). It consists of decades of data, it's evolved into a multigenerational study examining patterns within cardiovascular diseases. Much of the data from the Framingham Study is publicly available but certain features need to be requested. For our project we use data from this study that has been extracted and joined into a single csv file that is available for public use.

**Variables to operationalize**:

- The dataset contains over 4,000 records with 16 attributes, but we will only use the following attributes:
    - Y Concept: Total Cholesterol
    - X Concept: Smoking Status: 1=patient is a smoker or 0=non-smoker (nominal)
    - X Concept: Sex 1= Male 0 = Female (nominal)
    - X Concept: Age: Patient Age (continuous)
    - X Concept: BMI (continuous)

## 1.3 Methods

In the process of data cleaning, our first step involved addressing missing values, outliers, and inconsistencies within the dataset. Post-cleaning, we partitioned the dataset into training (30%) and testing (70%) subsets to enable effective model validation.

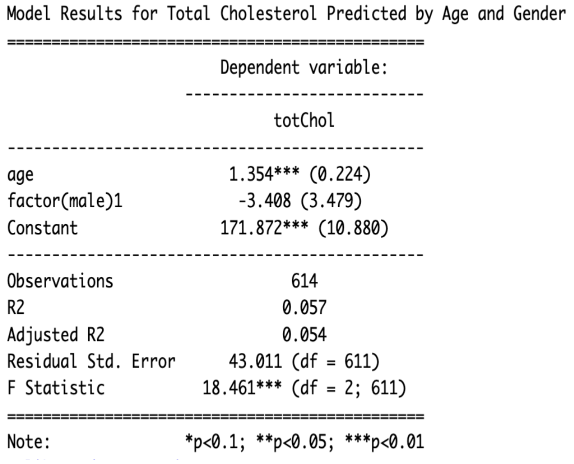
For Model 1, our focus was exclusively on smokers; thus, we isolated records indicating current smokers and subsequently split the data into training and testing subsets. This model utilized sex and age as predictor variables to estimate the total cholesterol levels through linear regression.

Model 2 took a broader approach, incorporating both smokers and non-smokers in the dataset. This model incorporated additional predictor variables, namely sex, age, BMI, and a binary indicator for current smoking status (currentSmoker).

The outcome variable remained the total cholesterol level. In assessing the assumptions of the Classical Linear Model (CLM) for both models, several tests were conducted. Linearity was evaluated via scatterplots of the dependent variable against each predictor. Independence was assumed based on the dataset's nature. Homoscedasticity was examined using residual plots, including a plot of residuals against predicted values. Lastly, the normality of residuals was assessed using Q-Q plots. Additionally, we tested the conditional linear expectation by plotting residuals against predicted values to ensure the robustness and reliability of our models.

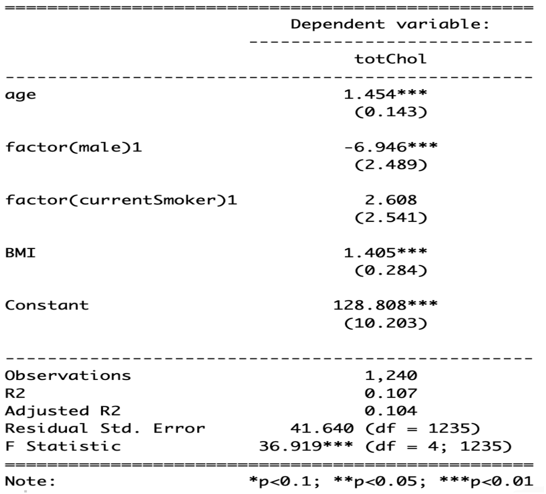
## 1.4 Visuals

## Model 1

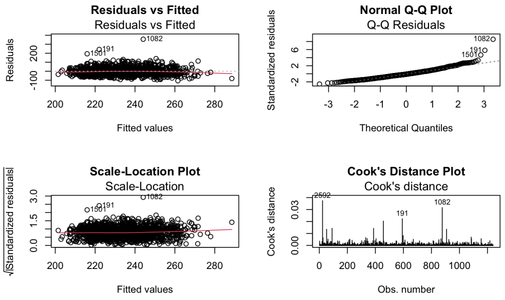


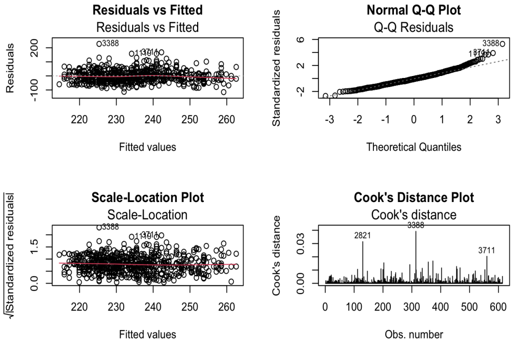
```
Model Results for Total Cholesterol Predicted by Age and Gender
================================================
                          Dependent variable:
                      ---------------------------
                                totChol
------------------------------------------------
age                         1.354*** (0.224)
factor(male)1                -3.408 (3.479)
Constant                   171.872*** (10.880)
------------------------------------------------
Observations                      614
R2                               0.057
Adjusted R2                      0.054
Residual Std. Error       43.011 (df = 611)
F Statistic            18.461*** (df = 2; 611)
================================================
Note:                 *p<0.1; **p<0.05; ***p<0.01
```

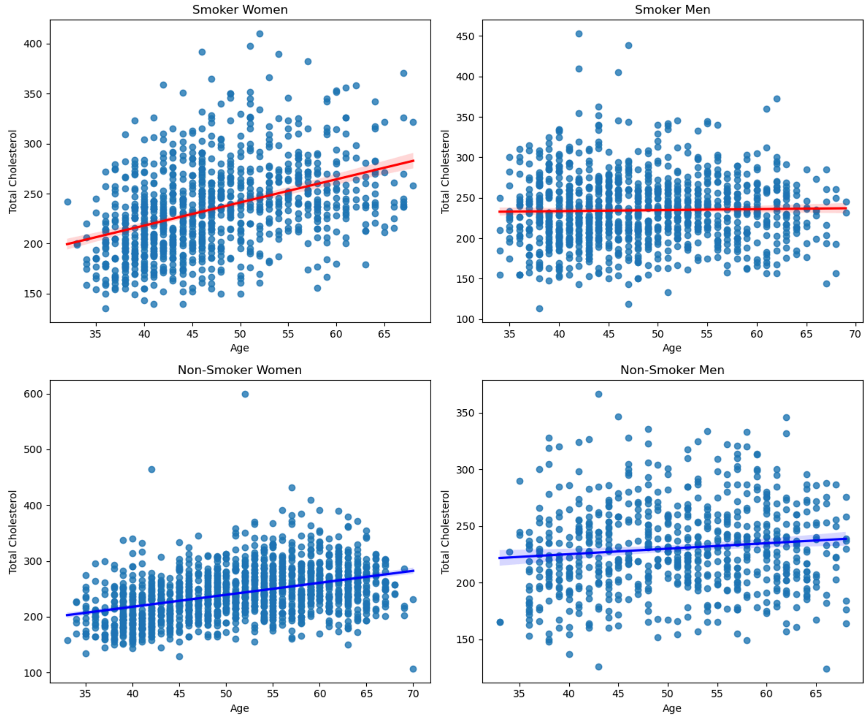
## Model 2



```
============================================================
                                   Dependent variable:
                               ----------------------------
                                         totChol
------------------------------------------------------------
age                                     1.454***
                                        (0.143)

factor(male)1                          -6.946***
                                        (2.489)

factor(currentSmoker)1                   2.608
                                        (2.541)

BMI                                     1.405***
                                        (0.284)

Constant                               128.808***
                                       (10.203)

------------------------------------------------------------
Observations                             1,240
R2                                       0.107
Adjusted R2                              0.104
Residual Std. Error               41.640 (df = 1235)
F Statistic                    36.919*** (df = 4; 1235)
============================================================
Note:                          *p<0.1; **p<0.05; ***p<0.01
```







Description: Females who smoke appeared to have to have increasing cholesterol levels as age progressed, but this was not the case their male counterparts

### 1.5 Results

***Model 1***

In our first model, we explored to see just how age and gender affect cholesterol levels in smokers. The results were quite telling, especially when it came to age. For every additional year in a smoker's life, their cholesterol level is predicted to increase by about 1.354 units on average ( t-statistic == 6.042). Furthermore, our model returned a p-value of less than 0.001, the likelihood that this finding is a fluke is super slim.

Gender, though, didn't seem to play a significant role, which was a bit surprising. Our model suggested that being male doesn't really change cholesterol levels in any noteworthy way. This part of the finding is a little shaky, with a t-value of -0.979 and a p-value of 0.328, indicating that any difference we might be seeing could well be due to chance.

When we look at the overall picture, our model's R-squared value sits at a modest 5.7%, indicating that while age is a factor, it's not the whole story—other unseen factors are likely at play, impacting cholesterol levels such as physical activity, diet, genetics, etc.

The diagnostic plots from R bring these numbers to life. The 'Residuals vs Fitted' plot helps us check that our model doesn't have any weird biases; we want to see those residuals (differences between observed and predicted cholesterol levels) scattered randomly, not forming any patterns.

The 'Normal Q-Q Plot' gives us a heads-up on whether our residuals follow a normal distribution, which is a good sign for our model's validity. The 'Scale-Location Plot' and 'Cook's Distance Plot' help us spot any outliers or influential points that might throw off our results.

Together, these statistics and visual checks from our plots reassure us that while our model isn't capturing everything, it's got a pretty good handle on how age is tied to cholesterol levels in smokers.

***Model 2***

In Model 2, our analysis expanded to include Body Mass Index (BMI) and smoking status alongside age and gender to see if they had any influence on cholesterol levels. Just as with our previous model, age remained a significant predictor of cholesterol levels. This relationship remained the same, showing an increase of approximately 1.354 units in cholesterol per year of age ( t-statistic=6.042, p-value<<<0.001) underscoring the robustness of age as a factor.

BMI also emerged as a noteworthy contributor in this model, indicating a potential correlation between body composition and cholesterol levels. The inclusion of smoking status, however, did not yield a significant impact, suggesting that the relationship between smoking and cholesterol may be influenced by other factors.

The model's R-squared value, while still modest at 5.7%, indicates that while age and BMI offer some explanatory power, they do not fully account for the variance in cholesterol levels. This means that additional variables, possibly lifestyle or genetic factors, might play influential roles in determining cholesterol levels.

The following R plots provide valuable insight into the model's findings. The 'Residuals vs Fitted' plot indicates a random spread of residuals, suggesting an absence of systematic error in the model's predictions.

The 'Normal Q-Q Plot' shows that the residuals closely follow a normal distribution, which is an indicator of model validity. Meanwhile, the 'Scale-Location Plot' and 'Cook's Distance Plot' detect any outliers that could potentially distort the model's predictive capability.

### 1.6 Discussion

Our results emphasize the significant impact of age on cholesterol levels, with each additional year predicting an average increase of approximately 1.354 units. While previous research shows low density lipoprotein cholesterol (LDL) levels typically with age, it is important to note that the current literature represents a more complex picture with respect to age and cholesterol levels (Downer, 2014). This robust relationship between age and cholesterol levels underscores the importance of age-related interventions and monitoring in managing cholesterol among smokers. Additionally, the inclusion of BMI as a notable contributor highlights the potential correlation between body composition and cholesterol levels, aligning with existing research that emphasizes the role of BMI in cardiovascular health (Hussain, 2019). However, the limited influence of gender and smoking status in our models suggests that these factors may be less deterministic than previously thought, encouraging further exploration into other lifestyle, genetic, or environmental variables that may influence cholesterol levels. Overall, our findings contribute to a more nuanced understanding of the multifaceted factors affecting cholesterol levels in smokers, guiding future research and potentially informing targeted interventions for cholesterol management in this population.

### 1.7 Limitations

In our study we did not distinguish between High Density Lipoprotein (HDL) and Low Density Lipoprotein (LDL) cholesterol levels. HDL cholesterol serves a positive function by transporting cholesterol from other tissues back to liver, subsequently reducing your risk of heart disease, while LDL buildup can start to clog one's arteries and make them more vulnerable to cardiovascular ailments (Mayo, 2022). To truly capture cardiovascular risk in the future, we will need to make a distinction between the two types of cholesterol in our modeling strategies.

We also had few patients with BMI values greater than 35, and having this to be the case may misrepresent the average cholesterol levels among this cohort, and more data points are required to properly assess how cholesterol changes with increasing BMI past 35.

Lastly, it's important to recognize that other components like exercise, genetics, and other health risk behaviors can impact cholesterol levels. These are features that can be extracted from the Framingham Data Set, and will be incorporated in future statistical modeling efforts.

### 1.8 Conclusion

In our analysis, age consistently emerged as a significant predictor of cholesterol levels in smokers across both models, with each additional year correlating to an increase of approximately 1.354 units. While BMI also showed influence in Model 2, smoking status did not yield a significant impact in either model. Despite age and BMI offering some explanatory power, other unexplored factors likely contribute to the variance in cholesterol levels among smokers. Further research with additional variables could provide a more comprehensive understanding of these complex relationships.

### 1.9 Bibliography

Londeree, B. R., & Moeschberger, M. L. (1982). Effect of age and other factors on maximal heart rate. *Research Quarterly for Exercise and Sport*, *53*(4), 297–304. https://doi.org/10.1080/02701367.1982.10605252

Linneberg, Allan et al. "Effect of Smoking on Blood Pressure and Resting Heart Rate: A Mendelian Randomization Meta-Analysis in the CARTA Consortium." Circulation. Cardiovascular genetics 8.6 (2015): 832–841. Web.

Hussain, A., Ali, I., Kaleem, W. A., & Yasmeen, F. (2019). Correlation between Body Mass Index and Lipid Profile in patients with Type 2 Diabetes attending a tertiary care hospital in Peshawar. *Pakistan journal of medical sciences*, *35*(3), 591–597. https://doi.org/10.12669/pjms.35.3.7

Zhou, C., Wang, M., Liang, J., He, G., & Chen, N. (2022). Ketogenic Diet Benefits to Weight Loss, Glycemic Control, and Lipid Profiles in Overweight Patients with Type 2 Diabetes Mellitus: A Meta-Analysis of Randomized Controlled Trails. International journal of environmental research and public health, 19(16), 10429. Link

Moser M. (1999). Hypertension treatment and the prevention of coronary heart disease in the elderly. American family physician, 59(5), 1248–1256.

Koliaki, C., Liatis, S., & Kokkinos, A. (2019). Obesity and cardiovascular disease: Revisiting an old relationship. Metabolism, 92, 98-107. doi: 10.1016/j.metabol.2018.10.008

Centers for Disease Control and Prevention. (2021, February 17). About heart disease. Retrieved from CDC Website

*Can we reduce vascular plaque buildup?*. Harvard Health. (2023, August 4). https://www.health.harvard.edu/heart-health/can-we-reduce-vascular-plaque-buildup

Rao Ch, S., & Subash Y, E. (2013). The effect of chronic tobacco smoking and chewing on the lipid profile. *Journal of clinical and diagnostic research : JCDR*, *7*(1), 31–34. https://doi.org/10.7860/JCDR/2012/5086.2663

Goldstein, J. L., & Brown, M. S. (2015). A century of cholesterol and coronaries: from plaques to genes to statins. *Cell*, *161*(1), 161–172. https://doi.org/10.1016/j.cell.2015.01.036

Downer, B., Estus, S., Katsumata, Y., & Fardo, D. W. (2014). Longitudinal trajectories of cholesterol from midlife through late life according to apolipoprotein E allele status. *International journal of environmental research and public health*, *11*(10), 10663–10693. https://doi.org/10.3390/ijerph111010663

Mayo Foundation for Medical Education and Research. (2022, November 3). *HDL cholesterol: How to boost your "good" cholesterol*. Mayo Clinic.https://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/in-depth/hdl-cholesterol/art-20046388#:~:text=This%20is%20why%20LDL%20cholesterol,and%20removed%20from%20your%20body.