19. **Multivariate Logistic Regression for Health Risk Factors in the United States**

Now, I'm going to explore how to handle situations with multiple variables using a multivariate logistic regression model. The earlier models only considered single variables, but in real-world scenarios, I want to consider all relevant factors together. So, when I have a collection of variables, it's essential to build a model that accounts for each of them.

The transformation of the probability remains the same, but now I'm dealing with a more general linear model that includes an intercept and a coefficient for each of the variables. When I invert that transformation, it again gives a probability between 0 and 1, which is exactly what I need. Using tools like R, I can include variables like balance, income, and student status in my model. The result will provide me with three coefficients, their standard errors, z statistics, and p-values.

Interestingly, when I include multiple variables, I notice that the coefficient for student status changes from positive in the single-variable model to negative in the multivariate model. This shift isn't an error; it illustrates the complexity of interpreting coefficients in a multiple regression model due to correlations between variables.

To visualize this, I can look at the relationship between credit card balance and default rates, distinguishing between students and non-students. Students tend to have higher balances, which explains why they may initially appear to have higher default rates. However, when controlling for balance, it turns out that, for each level of balance, students actually have a lower default rate than non-students. This is something I can only tease out using multivariate logistic regression, which considers these correlations.

**Example: Heart Disease Risk Factors in the United States**

Moving to another example, I want to focus on heart disease risk factors in the United States. This involves examining a dataset from a study conducted among American males who experienced a heart attack (myocardial infarction) and a control group who did not. The study included 160 men who had suffered a heart attack and 302 men who had not, all of whom were white males aged between 15 and 64 from a particular region in the U.S.

The study measured seven different predictors or risk factors, which I can visualize in a scatter plot matrix. This matrix is a useful way to explore the relationships between variables, with each plot showing a pairwise relationship. I can code the heart disease status into the plot: red points represent those who had a heart attack, and blue points represent those who did not. For instance, in the plot of tobacco usage against systolic blood pressure, I notice that individuals with high tobacco use and high blood pressure tend to be red points, indicating a higher likelihood of heart attack.

One risk factor that stands out is family history, a categorical variable that is crucial to consider. The data shows that individuals with a family history of heart disease have a higher risk, as evidenced by the greater number of red points in that category.

My goal here isn't necessarily to predict the probability of heart disease but to understand the role of different risk factors. For this, I use a generalized linear model (GLM) for logistic regression. In my analysis, I find that certain factors like tobacco usage, low-density lipoprotein (LDL, a type of cholesterol), family history, and age are significant. However, others like obesity and alcohol usage are not, which may initially seem surprising.

This is likely due to correlations among variables. For example, age and tobacco use are correlated, as are alcohol use and LDL cholesterol. The presence of significant variables like LDL may reduce the need for others, such as alcohol, in the model. These correlations must be carefully considered when interpreting the results of a multivariate logistic regression.

**Conclusion**

Through this project, I learned that multivariate logistic regression is a powerful tool for understanding the relationships between multiple predictors and a binary outcome. In health studies, where various factors interact and correlate, this method provides a clearer picture of which factors genuinely matter, allowing for better-targeted interventions and a deeper understanding of the underlying risks.