17. **Introduction to Classification Problems in Healthcare Data Analysis**

In the last part of my project, I focused on regression methods, where the response variable was quantitative. In this section, I will explore **classification problems**, which involve response variables with two or more discrete values. Classification is a very common problem in healthcare data analysis and is even more prevalent than regression in many machine learning applications. For example, I might predict whether a patient will survive or succumb to a given disease, or determine if a new diagnostic test will be positive or negative. These kinds of classification problems are critical in healthcare and will be the focus of this part of my project.

**Understanding Categorical Variables**

To start, it's important to understand what **categorical variables** look like. These variables take on discrete values with no inherent ordering. For example, a patient's eye color (brown, blue, green) is a categorical variable with three possible values. Similarly, a classification task might involve predicting whether an email is "spam" or "ham" (legitimate). In healthcare, I might be interested in predicting whether a patient is classified as "high-risk," "medium-risk," or "low-risk" for a certain condition.

**What is a Classifier?**

A classifier is a function that takes a feature vector (like patient characteristics) as input and assigns one of the possible categories as an output. Mathematically, the classifier maps the input features to a set of discrete values, denoted as CCC. For example, in a binary classification problem like determining whether an email is spam or ham, the classifier assigns the email to one of these two categories based on its features.

Although classification problems are often framed this way, in practice, I am usually more interested in estimating the **probability** that an observation (such as a patient) belongs to each category. For example, in healthcare, it might be more useful to estimate the probability that a patient has a high risk of heart disease rather than just classifying them as high-risk or not. This probability estimation helps prioritize interventions—for instance, treating a patient with a 98% risk of a heart attack more urgently than one with a 90% risk.

**Example: Predicting Credit Card Default**

To illustrate classification concepts, I use the **Credit Card Default** dataset. This dataset includes variables such as a cardholder's balance and income. The response variable indicates whether the cardholder defaulted on their credit card payment.

By visualizing the data with a scatter plot of balance against income, and color-coding the response variable (defaulted or not defaulted), I observe that balance appears to be a more significant predictor of default status than income. There is a clearer separation between those who defaulted and those who did not based on their balances, whereas income does not show much separation between the groups.

Box plots of these two variables confirm this observation. The plot shows a more distinct difference in the distributions of balance between those who defaulted and those who did not, while income distributions overlap significantly.

**Can We Use Linear Regression for Classification?**

An interesting question is whether I can use linear regression for classification problems. For example, I could encode the response variable as 0 (no default) or 1 (default) and perform a linear regression of the response on the predictors (balance and income). If the predicted value is greater than 0.5, I classify the observation as a default; otherwise, as a non-default.

For a binary classification problem, this approach can work reasonably well and is equivalent to linear discriminant analysis, which I will discuss later. However, there are limitations. Linear regression might produce probabilities outside the range [0, 1], which are not valid probabilities. To address this, I will introduce **logistic regression**, which is more appropriate for classification tasks.

**Linear Regression vs. Logistic Regression**

The plot of the fitted linear regression line illustrates its limitations in classification tasks. The line can extend below 0 or above 1, which does not make sense for probability estimation. For example, predicting a probability of -0.2 or 1.2 for a patient's survival is nonsensical.

In contrast, **logistic regression** provides a better fit, as shown in the plot. It ensures that predicted probabilities remain within the range [0, 1] and adjusts the curve to match the data more accurately, especially where there are clear separations between classes.

**Challenges with More than Two Categories**

When dealing with more than two categories, such as a patient's condition in an emergency room (e.g., stroke, drug overdose, epileptic seizure), linear regression becomes even less appropriate. If I arbitrarily assign numbers like 1, 2, and 3 to these categories, it may imply an ordering or equal intervals between conditions, which is not necessarily true.

For multi-class problems, a better approach is **multi-class logistic regression** or **discriminant analysis**, which I will cover later in the project. These methods handle multiple categories without implying any false order or equal distance between them.

**Conclusion**

This section introduces the concept of classification problems in healthcare data analysis, highlighting their importance and frequency. By understanding the limitations of linear regression and the appropriateness of logistic regression, I can choose the right methods to analyze healthcare data effectively. Moving forward, I will explore various classification techniques, including multi-class logistic regression and discriminant analysis, to handle more complex datasets and provide actionable insights for healthcare decision-making.