# GENERALIZED LINEAR MODELS

# PROBIT PROBLEM

PRODUCED BY: RISHI DEY CHOWDHURY

# Table of Contents



**About This File**

This file was created for the benefit of all teachers and students wanting to learn about Generalized Linear Models

SECTION 1

# Before Generalized Linear Models

Until now, we have seen linear models which can be depicted by the following box diagram and it modelled some linear system taking some inputs (maybe categorical or continuous) and random error generating a prediction for the continuous response variable.

## 1.1 Why didn't the Basic Linear Models suffice?

In several scenarios, we are interested in modelling a discrete response using all these covariate and factor inputs, but the Linear Models we saw so far are limited to prediction of continuous response variable. Thus, it lead to the Generalized Linear Models.


Figure 1: Basic Linear Model and Generalized Linear Model

2

SECTION

# History of Generalized Linear Models

Before delving into the mathematical formulation, we explore how GLM came into existence. It was originally formulated and created by statistician (or better known as Toxicologist[1]) DJ Finney. He was interested in finding out the strength[2] of various poisons he was working on. DJ Finney was trying to get a numerical value for lethal dose[3] of poisons. As the lethal dose varies from one living creature to another even in the same population, we can consider it to be a random variable. Just like we do for blood pressure, etc. So, what Mr. Finney did was that for each mice Mr. Finney calculated the lethal dose which is simply like a random sample from the population. Let these random variables be

$$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$$

we are not constrained to using Normal distribution but that is what Finney did and hence we will use it for our discussion now. In terms of toxicology, $\hat{\mu}$ is called potency whereas $\frac{1}{\hat{\sigma}^2}$ is called reliability of the poison i.e. reliability gives us an idea of the variation between the lethal dose amount resulting from various observations.

## 2.1 Why the above experiment feels reasonable?

This at a first place looks like a very simple problem of finding $\hat{\mu}$ and $\hat{\sigma}^2$. So, you might think we can ideally carry out this experiment. But what happens when you actually carry out the

[1] A toxicologist works on things like estimating the strength of some poison.

[2] One way to quantify the strength of poison is for example, Potassium Cyanide and Arsenic, how much of each poisonous chemical is enough to kill a person of a population. If a poison requires only a very little amount to kill a person compared to the other then that poison is more lethal/stronger

[3] The minimum dose of the poison that is required to kill any unit of a particular population.

experiment is you start with a mouse and slowly start applying the poison to it, by pushing the syringe down until the moment that the mouse dies. So, we simply record the measurement of the dosage used. This gives my $X_1$ and we carry on the experiment until we have a reasonable sample.

## 2.2 Why it isn't reasonable due to practical issues?

But this simplistic idea simply doesn't work out as death is not a clearly disearnable event, unlike blood pressure, etc. When we start applying the poison to the mouse, it first enters a coma state where the mouse is either living or dead. So, we have to stop the experiment and take the mouse on a different table and do some time consuming experiment to decide if the mouse is living or dead which takes 10 minutes (say). If the mouse is not dead it's internal body system is working at a tremendous rate in removing much of the poison from the blood system to urinary blood. So, we cannot just take the mouse back to the first table and continue with the experiment if the mouse is not dead. This is because the effective amount of poison inside the blood system of the mouse has changed as a result all we get from this experiment is we apply a fixed dose and look at the mouse which is either living or is in comma stage in which case, we have to carry out the experiment and see whether the mouse is actually living or dead. We cannot continue the experiment any further. So, after each experiment we either have the mouse living or dead. Thus, we cannot observe the $X_i s$ directly.

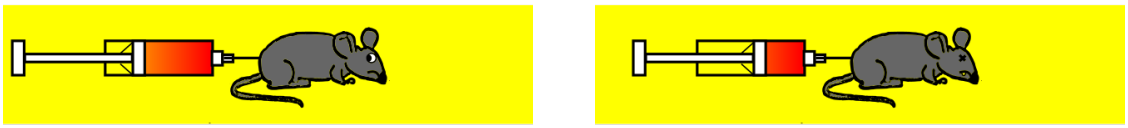

Figure 2: Start and End of Trial

## 2.3 New way to present the problem statement with this changed setup

Mathematically, the situation is like this that suppose the true lethal dose of a medicine is $X$ and the applied dose is $d$. We now observe the state of the mouse which is either dead or alive. All that we can conclude is if the mouse is dead $X$ must be less than or equal to $d$. On the other hand if the mouse is alive then $X$ must be greater than $d$. Thus, we can't observe $X$ directly and $X$ is thus called a latent variable. So, we want to estimate the $\mu$ and $\sigma^2$ but we can not observe the $X_i s$ directly. Thus, DJ Finney introduced a new type of data collection which is commonly known as bio acid.