The above methods use for analyzing contingency tables. Those methods help us investigate effects of explanatory variables on categorical response variables. Now we will use models as the basis of such analyses. In fact, the methods which we use above, also result from analyzing effects in certain models, but models can handle more complex situations. Such as analyzing simultaneously the effects of several explanatory variables.

A good fitting model has several benefits. The structural form of the model describes the patterns of association and interaction. The sizes of the model parameters determine the strength and importance of the effects. Inferences about parameters evaluate which explanatory variables affect the response variable Y, while controlling effects of possible confounding variables (an outside influence that changes the effect of a dependent and independent variable). Finally, the model's predicted values smooth the data and provide improved estimates of the mean of Y at possible explanatory variable values.

Generalized Linear Models (GLMs)

This broad class of models includes ordinary regression and ANOVA models for continuous responses as well as models for discrete responses.

We will discuss GLM for categorical and other discrete response data.

Components of a GLM:

All GLM have three components:-

1. The random component: identifies the response variable Y and assumes a probability distribution for it.
2. The systematic component: specifies the

explanatory variables for the model.
⇒ The Link function: specifies a function of the expected value (mean) of Y_i which the GLM relates to the explanatory variables through a prediction equation having linear form.

(1) Random Component:

The random component of a GLM identifies the response variable Y and selects a probability distribution for it. Denote the observations on Y by $(Y_1 \cdots Y_n)$. Standard GLMs treat $Y_1 \cdots Y_n$ as independent.

Y_i are binary ⇒ success or failure

Y_i are the number of successes out of a certain fixed number of trials ⇒ Binomial

Y_i are counts ⇒ Poisson or Negative Binomial

Y_i are continous ⇒ Normal. (weight)

(2) Systematic Component:

The systematic component of a GLM specifies the explanatory variables. These enter linearly as predictors on the right-hand side of the model equation. That is the systematic component specifies the variables that are the x_j in the formula:

$$\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

This linear combination of the explanatory variables is called linear predictor.

Some x_j can be based on others in the model $x_3 = x_1 x_2$ (interaction between x_1 & x_2) or $x_3 = x_1^2$ (curvilinear effect of x_1)

GLMs used to emphasize that x-values are treated as fixed rather than as a random variable.

(3) Link function:

The third component of a GLM, the link function, specifies a function $g(\cdot)$ that relates μ (the expected value of Y, the mean of the probability distribution) by $\mu = E(Y)$ to the linear predictor as

$$g(\mu) = \alpha + \beta_1 x_1 + \cdots \beta_k x_k$$

the function $g(\cdot)$, the link function, connects the random and systematic components.

⇒ The simplest link function is $g(u) = u$. This models the mean directly and is called the Identity link. It specifies a linear model for the mean response.

$$u = \alpha + \beta_1 x_1 + \cdots - \beta_k x_k$$

⇒ The link function $g(u) = \log\left(\frac{u}{1-u}\right)$ models the log of an odds. It is appropriate when u is between 0 & 1, such as probability. This is called logit link. A GLM that use logit link is called a logistic regression model.

Each potential probability distribution for Y has one special function of the mean that is called its natural parameter. For normal distribution, it is the mean itself. For binomial, the natural parameter is the logit of the success probability. The link function that uses the natural parameter as $g(u)$ in the GLM is called the canonical link.

Ordinary regression models for continuous responses are special cases of GLMs. The assume a normal dist^n for Y and model its mean directly, using the identity link function.

A GLM generalizes ordinary regression models in two ways: (1) it allows Y to have a dist^n other than the normal. (2) it allows modeling some function of the mean. Both generalizations are important for categorical data.

Historically, early analyses of nonnormal responses often attempted to transform Y so it is approximately normal, with constant variance. Then ordinary regression models using least squares are applicable. In practice it is difficult to do. So GLMs are used because the GLM fitting process uses Maximum likelihood methods for our choice of random component; and we not restricted to normality for that choice. Choice of link function is separate from the choice of random component. It is not chosen to produce normality or stabilize the variance

Now we illustrate the GLM components by introducing the two most important GLMs for discrete response — logistic regression models for binary data and loglinear models for count data.