UNIT-V.

Generalised Linear Models:- (GLM)

Introduction:-

The GLM is a unification of both linear and non linear regression models that also allows the incorporation of nonnormal response distributions. In a GLM the response variable distribution must only be a member of the exponential family, which includes the normal, poisson, binomial, exponential and gamma distributions as members. Further more, the normal - error linear model is just a special case of the GLM, so in many ways, the GLM can be thought of as a unifying approach to many aspects of empirical modeling and data analysis.

We begin our presentation of these models by considering the case of logistic regression This is a situation where the response variable has only two possible outcomes, generally called "success" and "failure" and denoted by 0 and 1. Notice that the response is essentially qualitative, since the designation "success" or "failure" is entirely arbitrary. then we consider the situation where the response variable is a count, such as the number of defects in a unit of product or the number of relatively rare events such as the number of Atlantic hurricanes that make landfall on the united States in a year. Finally, we discuss how all these situations are unified by the GLM.

Logistic Regression models:-

Models with a Binary Response variable:-

Consider the situation where the response variable in a regression problem takes on only two possible values, 0 and 1. These could be arbitrary assignments resulting from observing a qualitative response. For Example, the response could be the outcome of a functional electrical test on a semiconductor device ... are either a "success", which means the device works properly, or a "failure", which could be due to a short, an open, or some other functional problem.

Suppose that the model has the form

$$y_i = x_i'\beta + \varepsilon_i \rightarrow (1)$$

where $x_i' = [1, x_{i1}, x_{i2}, \ldots, x_{ik}]$

$\beta' = [\beta_0, \beta_1, \beta_2, \ldots, \beta_k]$,

and the response variable y_i takes on the value either 0 or 1. We will assume that the response variable y_i is a Bernoulli random variable with probability distribution as follows:

y_i	Probability
1	$P(y_i = 1) = \pi_i$
0	$P(y_i = 0) = 1 - \pi_i$

Now since $E(\varepsilon_i) = 0$, the expected value of the response variable is

$$E(y_i) = 1(\pi_i) + 0.(1 - \pi_i)$$

$$= \pi_i$$

This implies that

$$E(y_i) = x_i'\beta = \pi_i$$

This means that the expected response given by the reponse function $E(y_i) = x_i'\beta$ is just the probability that the response variable takes on the value 1.

There are some substantive problems with the regression model in equation (1). First, note that if the response is binary, then the error terms ε_i can only take on two values, namely

$\varepsilon_i = 1 - x_i'\beta$ when $y_i = 1$

$\varepsilon_i = -x_i'\beta$ when $y_i = 0$

Consequently, the errors in this model cannot possibly be normal. Second, the errors variance is not constant, since

$$\sigma^2_{y_i} = E\{y_i - E(y_i)\}^2$$

$$= (1-\pi_i)^2 \pi_i + (0-\pi_i)^2 (1-\pi_i)$$

$$= \pi_i (1-\pi_i)$$

Notice that this last expression is just

$$\sigma_{y_i}^2 = E(y_i)[1-E(y_i)]$$

Since $E(y_i) = x_i' \beta = \pi_i$. This indicates that the variance of the observations (which is the same as the variance of the errors because $\varepsilon_i = y_i - \pi_i$, and π_i is a constant) is a function of the mean. Finally, there is a constraint on the response function, because

$$0 \leq E(y_i) = \pi_i \leq 1$$

This restriction can cause serious problems with the choice of a linear response function, as we have initially assumed in equation ①. It would be possible to fit a model to the data for which the predicted values of the response lie outside the 0, 1 interval.

Generally, when the response variable is binary, there is considerable empirical evidence indicating that the shape of the response function should be nonlinear. A monotonically increasing (or decreasing) S-shaped (or reverse S-shaped) function, such as shown in figure ①, is usually employed. This function is called the logistic response function, and has the form

$$E(y) = \frac{\exp(x'\beta)}{1+\exp(x'\beta)} \longrightarrow ②$$

or equivalently,

$$E(y) = \frac{1}{1+\exp(-x'\beta)} \longrightarrow (3)$$

The logistic response function can be easily linearized. One approach defines the structural portion of the model in terms of a function of the response function mean. Let

$$\eta = x'\beta \longrightarrow (4)$$

be the linear predictor where η is defined by the transformation.

$$\eta = \ln \frac{\pi}{1-\pi} \longrightarrow (5)$$

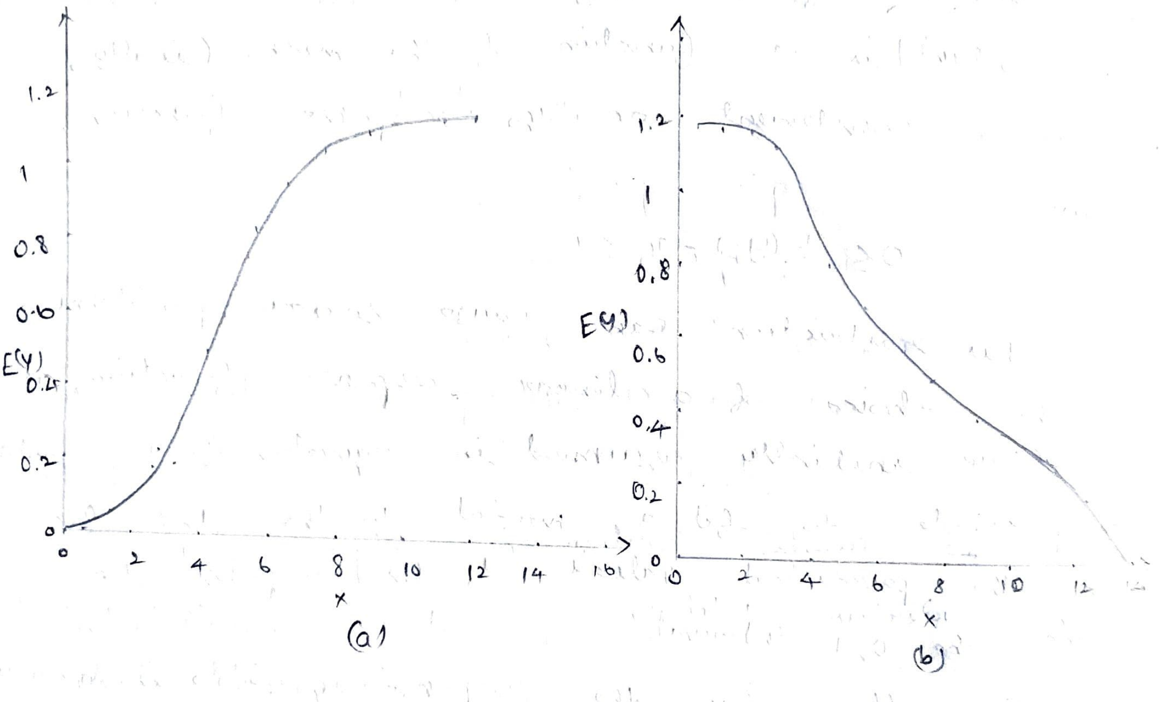


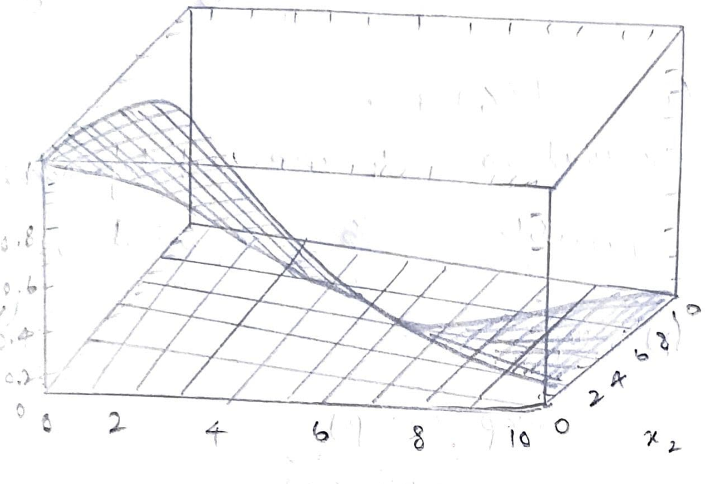
Figure (1) :- Examples of the logistic response function

a). $E(y) = 1/(1+e^{-6.0-1.0x})$,

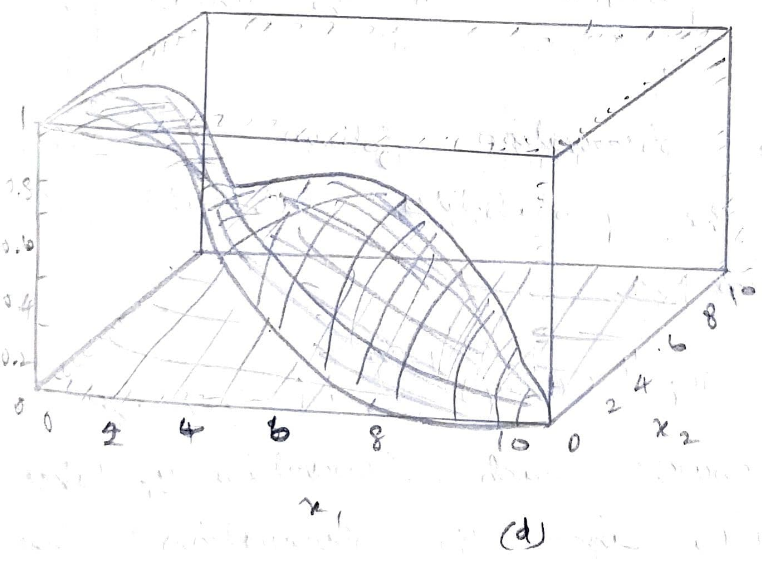
b). $E(y) = 1/(1+e^{-6.0+1.0x})$.

This transformation is often called the logit transformation of the probability π, and the ratio $\pi(1-\pi)$ in the transformation is called the odds. Sometimes the logit transformation is called the log-odds.

There are other functions that have the same shape as the logistic function, and they can also be obtained by transforming π using the cumulative normal distribution. This produces a probit regression model. The probit regression model is less flexible than the logistic regression model and probably not as widely used because it cannot easily incorporate more than



(c)



(d)

Figure 13.1 (Continued)

(c). $E(y) = 1/(1 + e^{-5.0+0.65x_1+0.4x_2})$, and

(d). $E(y) = 1/(1 + e^{-5.0+0.65x_1+0.4x_2+0.15x_1x_2})$.

One predictor variable, Another possible transformation is the complementary loglog transformation of π, given by $\ln[-\ln(1-\pi)]$. This results in a response function that is not symmetric about the value $\pi = 0.5$.

Estimating the parameters in a Logistic Regression Model :-

The general form of the logistic regression model is

$$y_i = E(y_i) + \varepsilon_i$$

where the observations y_i are independent Bernoulli random variables with expected values

$$E(y_i) = \pi_i$$

$$= \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}$$

We will use the method of maximum likelihood to estimate the parameters in the linear predictor $x_i'\beta$.

Each sample observation follows the Bernoulli distribution, so the probability distribution of each sample observation is

$$f_i(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}, \quad i = 1, 2, \ldots, n$$

and of course each observation y_i takes on the value 0 or 1, Since the observations are independent, the likelihood function is just

$$L(y_1, y_2, \ldots, y_n, \beta) = \prod_{i=1}^{n} f_i(y_i)$$

$$= \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

It is more convenient to work with the log-likelihood.

$$\ln L(y_i, y_2, \ldots, y_n, \beta) = \ln \prod_{i=1}^{n} f_i(y_i)$$

$$= \sum_{i=1}^{n} \left[y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right)\right] + \sum_{i=1}^{n} \ln(1-\pi_i)$$

Now since $1-\pi_i = [1+\exp(x_i'\beta)]^{-1}$ and η_i

$\eta_i = \ln[\pi_i/(1-\pi_i)] = x_i'\beta$, the log-likelihood can be written as

$$\ln L(y, \beta) = \sum_{i=1}^{n} y_i x_i'\beta - \sum_{i=1}^{n} \ln[1+\exp(x_i'\beta)]$$

Often in logistic regression models we have repeated observations or trials at each level of the x variables. This happens frequently in designed experiments. Let y_i represent the number of 1's observed for the ith observations and n_i be the number of trials at each observation. Then the log-likelihood becomes

$$\ln L(y, \beta) = \sum_{i=1}^{n} y_i \pi_i + \sum_{i=1}^{n} n_i \ln(1-\pi_i) - \sum_{i=1}^{n} y_i \ln(1-\pi_i)$$

Numerical search methods could be used to compute the maximum likelihood estimates (or MLEs) $\hat{\beta}$. However, it turns out that we can use iteratively reweighted least squares (IRLS) to actually find the MLEs. For details of this procedure, refer to Appendix C.13. There are several excellent computer programs that implement IRLS for the logistic regression model, such as SAS PROC GENMOD and S-PLUS.

Let $\hat{\beta}$ be the final estimate of the model parameters that the above algorithm produces. If the model assumptions are correct, then we can show that asymptotically.

$E(\hat{\beta}) = \beta$ and $Var(\hat{\beta}) = (X' V^{-1} X)^{-1}$

The estimated value of the linear predictor is $\hat{\eta}_i = x_i' \hat{\beta}$, and the fitted value of the logistic regression model is often written as

$$\hat{y}_i = \hat{\pi}_i = \frac{\exp(\hat{\eta}_i)}{1 + \exp(\hat{\eta}_i)}$$

$$= \frac{\exp(x_i' \hat{\beta})}{1 + \exp(x_i' \beta)}$$

$$= \frac{1}{1 + \exp(-x_i' \hat{\beta})}.$$

Poisson Regression :-

We now consider another regression modeling scenario where the response variable of interest is not normally distributed. In this situation the response variable represents a count of some relatively rare event, such as defects in a unit of manufactured product, errors or "bugs" in software, or a count of particulate matter or other pollutants in the environment. The analyst is interested in modeling the relationship between the observed counts and potentially useful regressor or predictor variables. For example, an engineer could be interested in modeling the relationship between the observed number of defects in a unit of product and production conditions when the unit was actually manufactured.

We assume that the response variable y_i is a count, such that the observation $y_i = 0, 1, \ldots$. A reasonable probability model for count data is often the Poisson distribution

$$f(y) = \frac{e^{-\mu} \mu^{y}}{y!}, \quad y = 0, 1, \ldots \qquad \rightarrow (13.27)$$

where the parameter $\mu > 0$. The poisson is another example of a probability distribution where the mean and variance are related. In fact, for the Poisson distribution it is straightforward to show that,

$$E(y) = \mu \quad \text{and} \quad Var(y) = \mu$$

That is, both the mean and variance of the Poisson distribution are equal to the parameter μ.

The Poisson regression model can be written as

$$y_i = E(y_i) + \varepsilon_i, \quad i = 1, 2, \dots, n \rightarrow (13.28)$$

We assume that the expected value of the observed responce can be written as

$$E(y_i) = \mu_i$$

and that there is a function g that relates the mean of the response to a linear predictor, say

$$g(\mu_i) = \eta_i$$
$$= \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$
$$= x_i' \beta \rightarrow (13.29)$$

The function g is usually called the link function. The relationship between the mean and the linear predictor is

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(x_i'\beta) \rightarrow (13.30)$$

There are several link functions that are commonly used with the Poisson distribution. One of these is the identity link.

$$g(\mu_i) = \mu_i = x_i' \beta \rightarrow (13.31)$$

when this link is used, $E(y_i) = \mu_i = x_i'\beta$ since $\mu_i = g^{-1}(x_i'\beta) = x_i'\beta$. An other popular link function for the Poisson distribution is the log link

$$g(\mu_i) = \ln(\mu_i) = x_i'\beta \quad \longrightarrow (13.32)$$

For the log link in Eq (13.32), the relationship between the mean of the response variable and the linear predictor is

$$\mu_i = g^{-1}(x_i'\beta)$$

$$= e^{x_i'\beta} \quad \longrightarrow (13.33)$$

(linear) predictor

The (log link) is particularly attractive for Poisson regression because it ensures that all of the predicted values of the response variable will be nonnegative.

The method of maximum likelihood is used to estimate the parameters in Poisson regression. The development follows closely the approach used for logistic regression. If we have a random sample of n observations on the response y and the predictor x, then the likelihood function is

$$L(y, \beta) = \prod_{i=1}^{n} f_i(y_i)$$

$$= \prod_{i=1}^{n} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \quad \longrightarrow (13.34)$$

$$= \frac{\prod_{i=1}^{n} \mu_i^{y_i} \exp\left(-\sum_{i=1}^{n} \mu_i\right)}{\prod_{i=1}^{n} y_i!}$$

where $\mu_i = g^{-1}(x_i'\beta)$. Once the link function is specified, we maximize the log-likelihood.

$$\ln L(y, \beta) = \sum_{i=1}^{n} y_i \ln(\mu_i) - \sum_{i=1}^{n} \mu_i - \sum_{i=1}^{n} \ln(y_i!) \quad \longrightarrow (13.35)$$

Iteratively reweighted least squares can be used to find the maximum likelihood estimates of the parameters in Poisson regression, following an approach similar to that used for logistic regression. Once the parameter estimates $\hat{\beta}$ are obtained, the fitted Poisson regression model is

$$\hat{y}_i = g^{-1}(x'_i \hat{\beta}) \quad \rightarrow (13.36)$$

For example, if the identity link is used, the prediction equation becomes

$$\hat{y}_i = g^{-1}(x_i' \hat{\beta})$$
$$= x_i' \hat{\beta}$$

and if the log link is specified, then

$$\hat{y}_i = g^{-1}(x_i' \hat{\beta})$$
$$= \exp(x_i' \hat{\beta})$$

Inference on the model and its parameters follows exactly the same approach as used for logistic regression. That is, model deviance is an overall measure of goodness of fit, and tests on subsets of model parameters can be performed using the difference is deviance between the full and reduced models. These are likelihood ratio tests. Wald inference, based on large-sample properties of maximum likelihood estimators, can be used to test hypotheses and construct confidence intervals on individual model parameters.

Interpretation of the parameters in a Logistic Regression Model:-

It is relatively easy to interpret the parameters in a logistic regression model, Consider first the case where the linear predictor has only a single regressor, so that the fitted value of the model at a particular value of x, say x_i, is

$$\hat{\eta}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

The fitted value at $x_i + 1$ is

$$\hat{\eta}(x_i+1) = \hat{\beta}_0 + \hat{\beta}_1(x_i+1)$$

and the difference in the two predicted values is

$$\hat{\eta}(x_i+1) - \hat{\eta}(x_i) = \hat{\beta}_1$$

Now $\hat{\eta}(x_i)$ is just the log-odds when the regressor variable is equal to x_i, and $\hat{\eta}(x_i+1)$ is just the log-odds when the regressor is equal to $x_i + 1$.

× [TABLE 13.2 Output from SAS PROC GENMOD for the Logistic Regression Model in Example 13.1

The GENMOD procedure

Model Information

Description	Value
Data set	WORK.LUNG
Distribution	BINOMIAL
Link Function	LOGIT
Dependent Variable	R
Dependent variable	N
Observation used	8
Number of Events	44
Number of Trials	391

Criteria for Assessing Goodness of Fit

Criterion	DF	Value	Value DF
Deviance	6	6.0508	1.0085
Scaled Deviance	6	6.0508	1.0085
Pearson Chi-square	6	5.0285	0.8381
Scaled Pearson X2	6	5.0285	0.8381
Log Likelihood		-109.6637	

Analysis of Parameter Estimates

Parameter	DF	Estimate	Std. Error	Chi-Square	Pr > Chi
INTERCEPT	1	4.7965	0.5686	71.1606	0.0001
X	1	0.0935	0.0157	36.7084	0.0001
SCALE	0	1.0000	0.0000		

Note: The Scale parameter was held fixed

Last Evaluation of the Negative of the Hessian

Parameter	PRM1	PRM2	Scale
PRM1	32.87874	1153.472	0
PRM2	1153.472	44669.1	0
Scale	0	0	

Therefore, the difference is the two fitted values is

$$\hat{\eta}(x_{i+1}) - \hat{\eta}(x_i) = \ln(\text{odds}_{x_{i+1}}) - \ln(\text{odds}_{x_i})$$

$$= \ln\left(\frac{\text{odds}_{x_{i+1}}}{\text{odds}_{x_i}}\right)$$

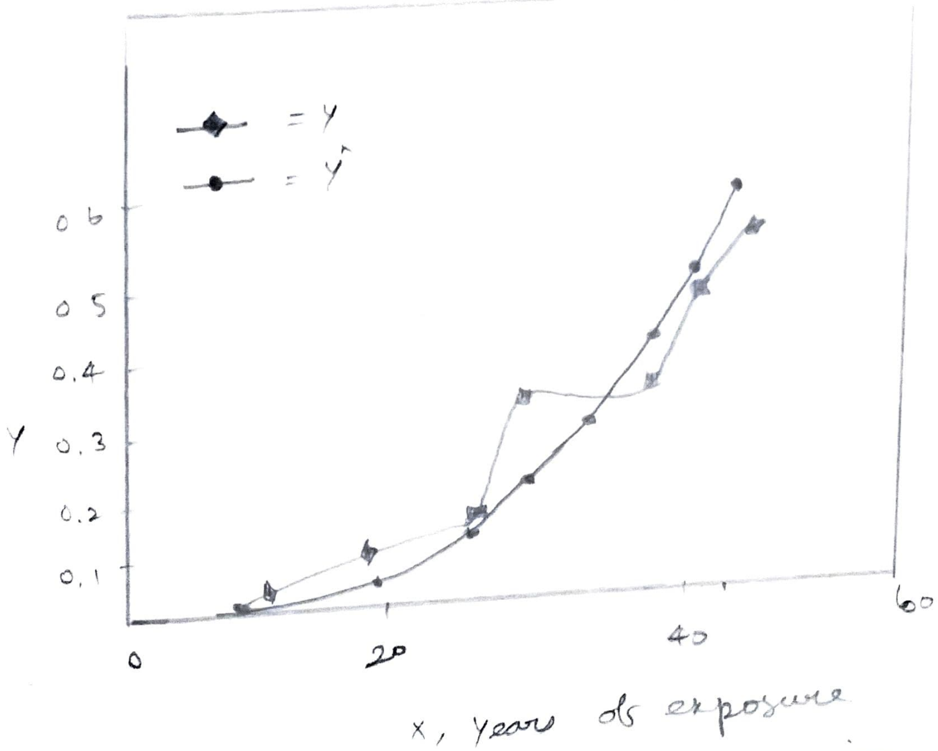


Figure 13.3. The fitted logistic regression model for the [illegible] data from Table 13.1.

If we take antilogs, we obtain the odds ratio

$$\hat{O}_R = \frac{\text{odds}_{x_i+1}}{\text{odds}_{x_i}} = e^{\hat{\beta}_1}$$

The odds ratio can be interpreted as the estimated increase in the probability of success associated with a one-unit change in the value of the predictor variable. In general, the estimated increase in the odds ratio associated with a change of d units in the predictor variable is $\exp(d\hat{\beta}_i)$

The interpretation of the regression coefficients in the multiple logistic regression model is similar to that for the case where the linear predictor contains only one regressor. That is, the quantity $\exp(\hat{\beta}_j)$ is the odds ratio for regressor x_j, assuming that all other predictor variables are constant.