Course Name :Basic Statistics using GUI-R (RKWard)
Module : Regression Continued
Week 6 Lecture : 1

Harsh Pradhan, Assistant Professor,
Institute of Management Studies, BHU
https://bhu.ac.in/Site/FacultyProfile/1_5?FA000562

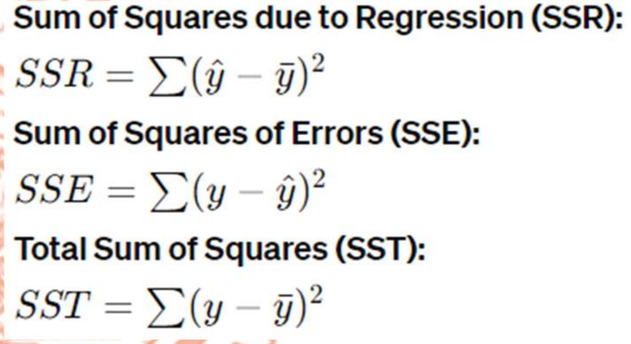
b = SSxy / SSxx
a = ȳ - b * x̄ = 2.96

$SST=\sum(y-\bar{y})2$

SST = SSyy

R² = 1 - (SSE / SST)

DF_regression = 1 (since there is one independent variable)
DF_total = n - 1



Sum of Squares due to Regression (SSR):

$$SSR = \sum(\hat{y} - \bar{y})^2$$

Sum of Squares of Errors (SSE):

$$SSE = \sum(y - \hat{y})^2$$

Total Sum of Squares (SST):

$$SST = \sum(y - \bar{y})^2$$

R Script for Regression

```
library(stats)
library(car)

library(tidyverse);library(performance)
performance::check_model(Model)



summary(Model)
anova(Model)
plot(Model),
#Calculated values
predict(Model)
#Generate residuals
predict(Model)
#SS error
deviance(Model)
#Coeeff
coefficients(Model)
#confidence interval
```

R Script for Regression

```
library(stats)
library(car)


confint(Model)
#vif if there are two or more independent variable
car::vif(Model)
#normality of residuals
qqplot(Model)
#insignificant p-value absence of autocorrel in residuals. Errors are dependent
durbinWatsonTest(Model)
# homoscedasticity, sign p value means constant variance
ncvTest(Model)
#plots res  vs fit, qq plot , std residual vs fitted, residual vs leverage
par(mfrow=c(2,2)); plot(Model)
```

Course Name :Basic Statistics using GUI-R (RKWard)

Module : Chi Square

Week 6 Lecture : 2

Harsh Pradhan, Assistant Professor,
Institute of Management Studies, BHU
https://bhu.ac.in/Site/FacultyProfile/1_5?FA000562

Chi Square

test of independence

Goodness of Fit

Goodness of Fit

- Is a dice fair?
- 1 -9 2-7 3-6 4-4 5-5 6 -5

1,2,3,4,5,6- 6.. Ideal….

	1	2	3	4	5	6
expected	1/6	1/6	1/6	1/6	1/6	1/6
observed	9/36	7/36	1/6	4/36	5/36	5/36

Chi square =summation (obs-expect) square/sum expected,
and degree of freedom, then we compare tobserved and tcalc , at specific p value

Course Name :Basic Statistics using GUI-R (RKWard)

Module : Chi Square Continued

Week 6 Lecture : 3

Harsh Pradhan, Assistant Professor,
Institute of Management Studies, BHU
https://bhu.ac.in/Site/FacultyProfile/1_5?FA000562

Goodness of Fit

- Is a dice fair?
- 1 -9 2-7 3-6 4-4 5-5 6 -5

1,2,3,4,5,6- 6.. Ideal….

	1	2	3	4	5	6
expected	1/6	1/6	1/6	1/6	1/6	1/6
observed	9/36	7/36	1/6	4/36	5/36	5/36

Chi square =summation (obs-expect) square/sum expected,
and degree of freedom, then we compare tobserved and tcalc , at specific p value

-
  ```
  tb= table(x,y)
  chisq.test(tb)
  ```

Chi-square Test of Independence

Chi-square test is used to analyze whether there exists any association between two or more categorical variables. Chi-square test tests the hypothesis whether two or more samples drawn from the same population have similar characteristics or not.

```
tb=table(survey$Smoke,survey$Exer) #contingency table
```

```
chisq.test(tb)
```

If p>0.05, they are independent

```
vcd::assocstats(table(chi$gender,chi[["laptop"]]))
```

Chi-square Test of Association

- If we fail to accept the null hypothesis of independence between variables using chi-square test of independence, we might be interested to know measure of association between variables in order to gauge the strength of the relationships present.

- The type of treatment in the **Treatment** variable are Placebo and Treatment, while treatment outcomes in **Improved** variable are none, some, and marked.

Non Parametric Test

- Nonparametric tests are useful when distributional assumptions of parametric tests such as linearity, normality, and equality of variance are violated.
- There is a whole class of nonparametric tests that are available for analyzing data which are ordinal in nature.
- Some data are ordinal by their definition like ranking of brands, while in other cases, data like ratio or interval are required to be converted to ranks because they do not fulfil assumptions of parametric tests.

Nonparametric test	Parametric Alternative
1-sample sign test	One-sample Z-test, One sample t-test
1-sample Wilcoxon Signed Rank test	One sample Z-test, One sample t-test
Friedman test	Two-way ANOVA
Kruskal-Wallis test	One-way ANOVA
Mann-Whitney test	Independent samples t-test
Mood’s Median test	One-way ANOVA
Spearman Rank Correlation	Correlation Coefficient

```
wilcox.test(my.csv.data$CSE_1,mu=3.5)
t.test(my.csv.data$CSE_1,mu=3.5)
```

a.k.a. **Log Odds**

or **Logit**

Intercept

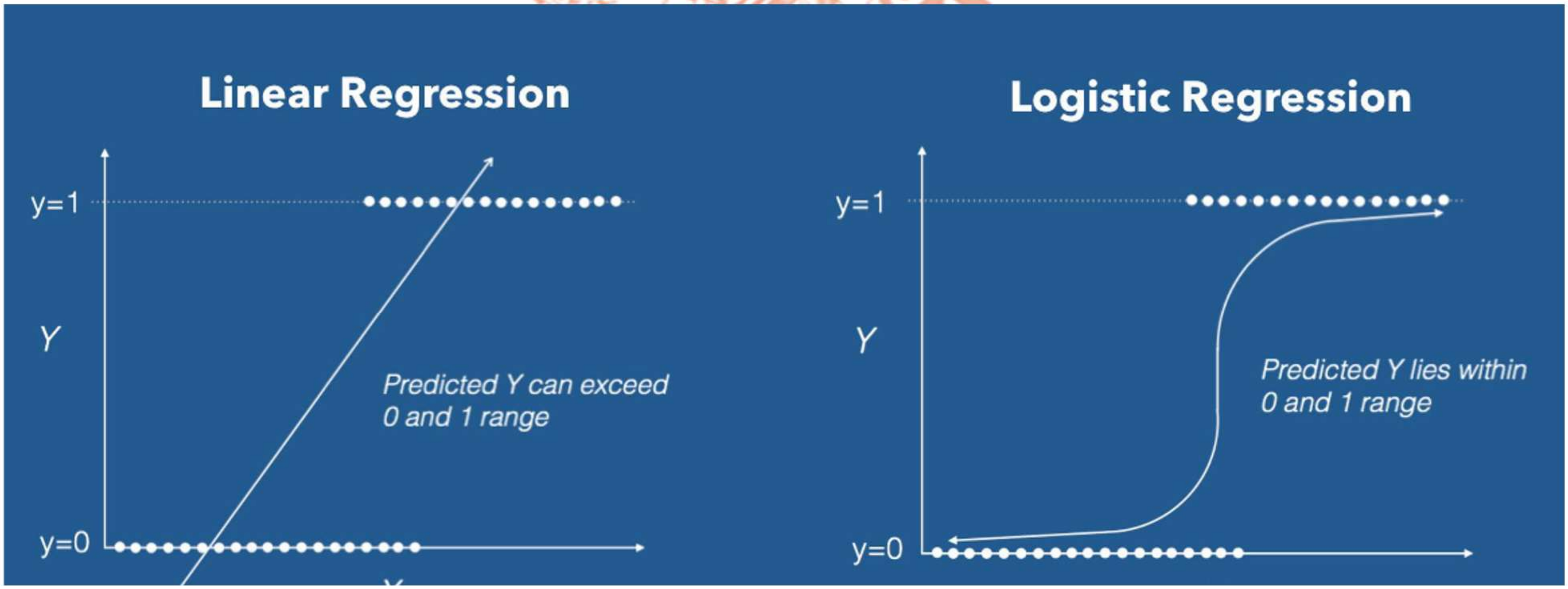
$$\boxed{\log\left(\frac{P}{1-P}\right)} = \beta_0 + \beta_1 X$$

P/(1-P)… odds ratio

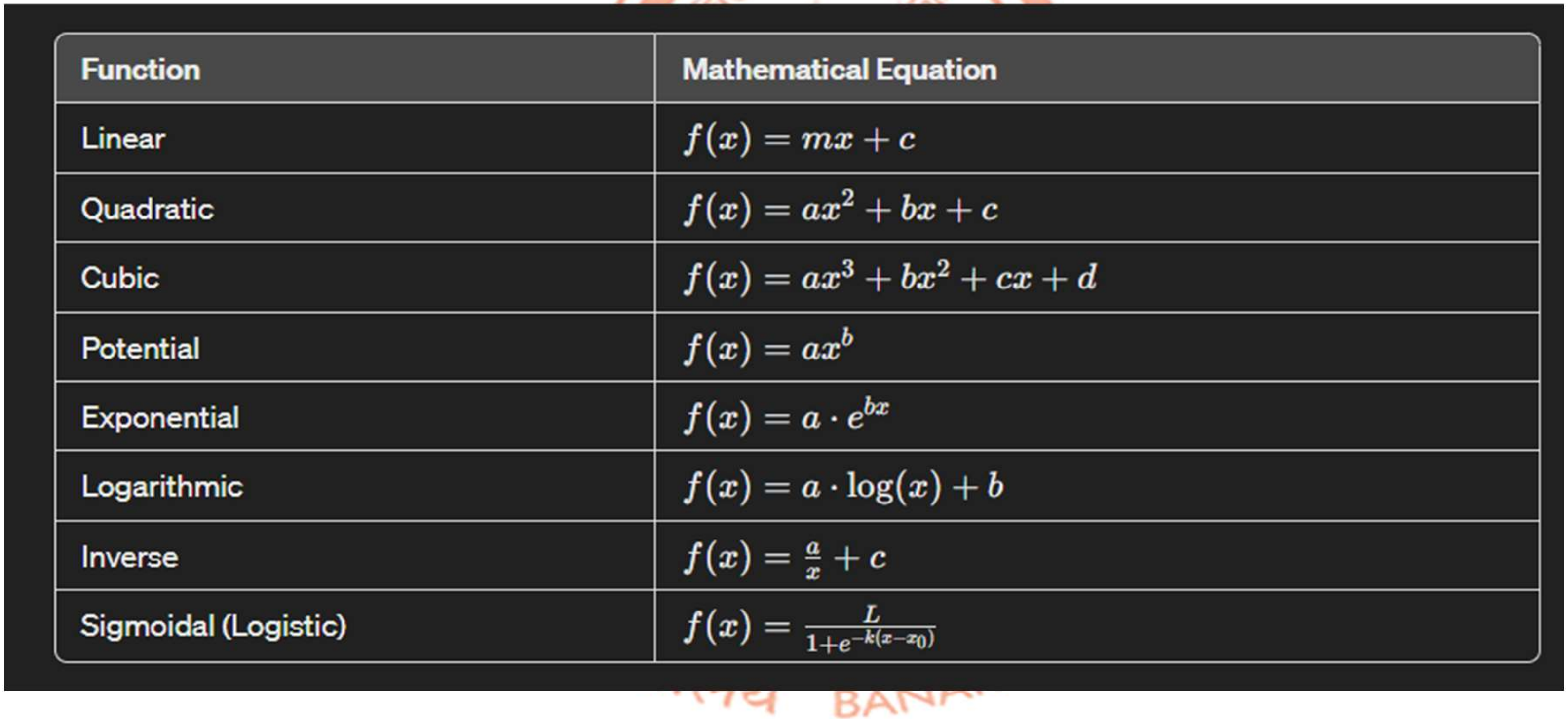
Log [P/(1-P)] is a straight line with Beta-0 as intercept , beta-1 as slope

Log [p/(1-p)]= b +ax, p /(1-p)=exp(b).exp(ax), 1/(2p-1)= [1 +exp(a+bx)]/[exp(a+bx)-1]

p = 1/(1+exp(-a-bx)) ; p= 1/(1+ exp(-y))



Regression Function



Function	Mathematical Equation
Linear	$f(x) = mx + c$
Quadratic	$f(x) = ax^2 + bx + c$
Cubic	$f(x) = ax^3 + bx^2 + cx + d$
Potential	$f(x) = ax^b$
Exponential	$f(x) = a \cdot e^{bx}$
Logarithmic	$f(x) = a \cdot \log(x) + b$
Inverse	$f(x) = \frac{a}{x} + c$
Sigmoidal (Logistic)	$f(x) = \frac{L}{1+e^{-k(x-x_0)}}$

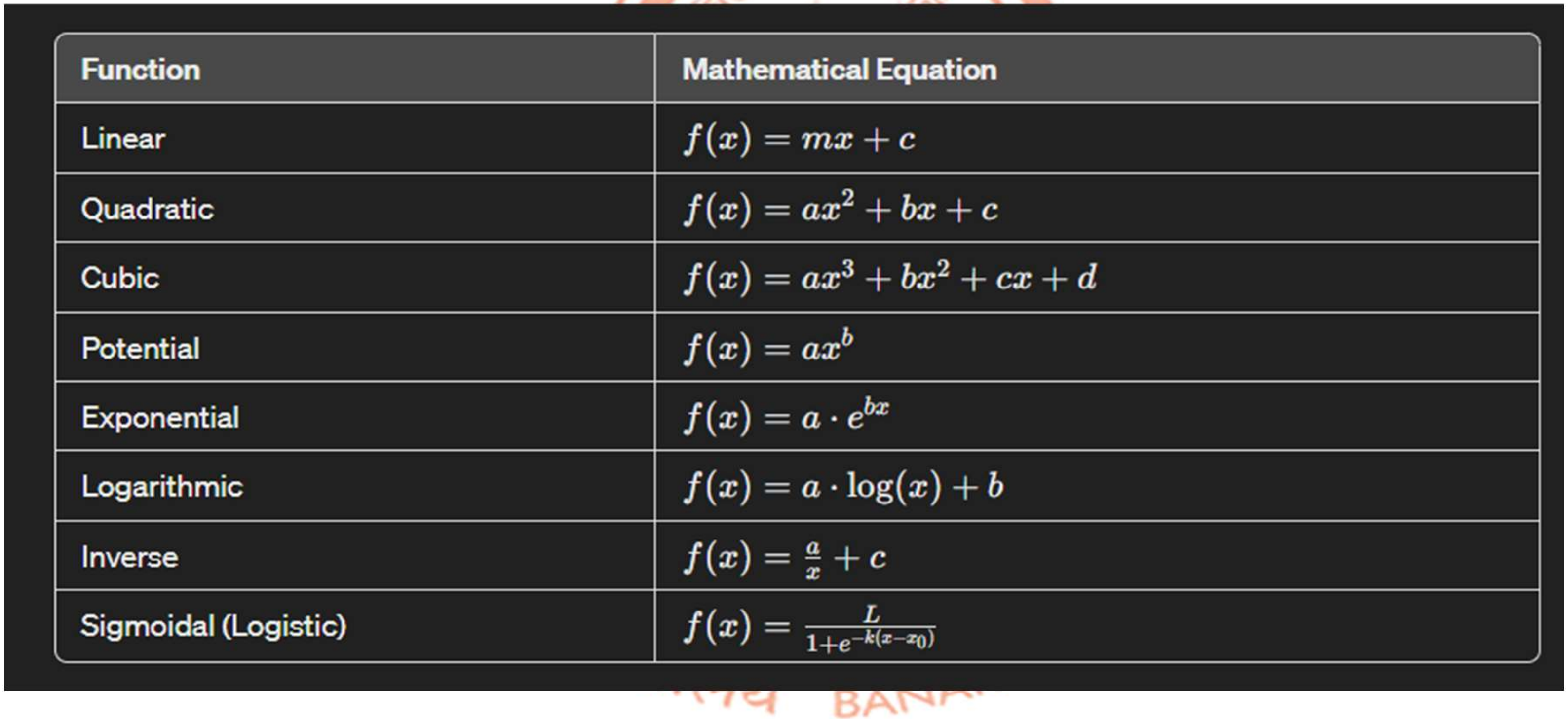
Course Name :Basic Statistics using GUI-R (RKWard)

Module : Non-Linear Function

Week 6 Lecture : 4

Harsh Pradhan, Assistant Professor,
Institute of Management Studies, BHU
https://bhu.ac.in/Site/FacultyProfile/1_5?FA000562

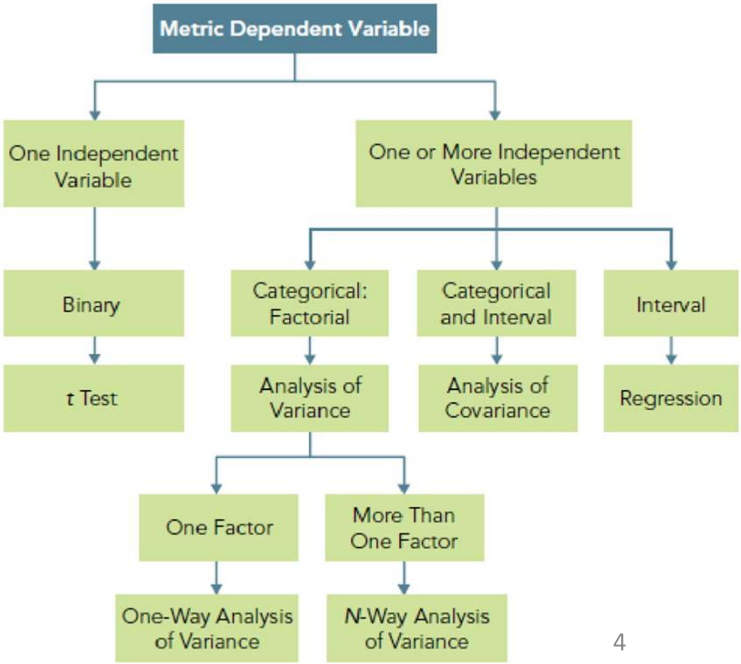
Regression Function



Function	Mathematical Equation
Linear	$f(x) = mx + c$
Quadratic	$f(x) = ax^2 + bx + c$
Cubic	$f(x) = ax^3 + bx^2 + cx + d$
Potential	$f(x) = ax^b$
Exponential	$f(x) = a \cdot e^{bx}$
Logarithmic	$f(x) = a \cdot \log(x) + b$
Inverse	$f(x) = \frac{a}{x} + c$
Sigmoidal (Logistic)	$f(x) = \frac{L}{1+e^{-k(x-x_0)}}$

Logistic Regression

	Ind-metric	Indep-non metric
Dep=metric	Regression	Hypo-testing/ ANOVA
Dep=Not metric	Logistic	Chisq



a.k.a. **Log Odds**

or **Logit**

Intercept

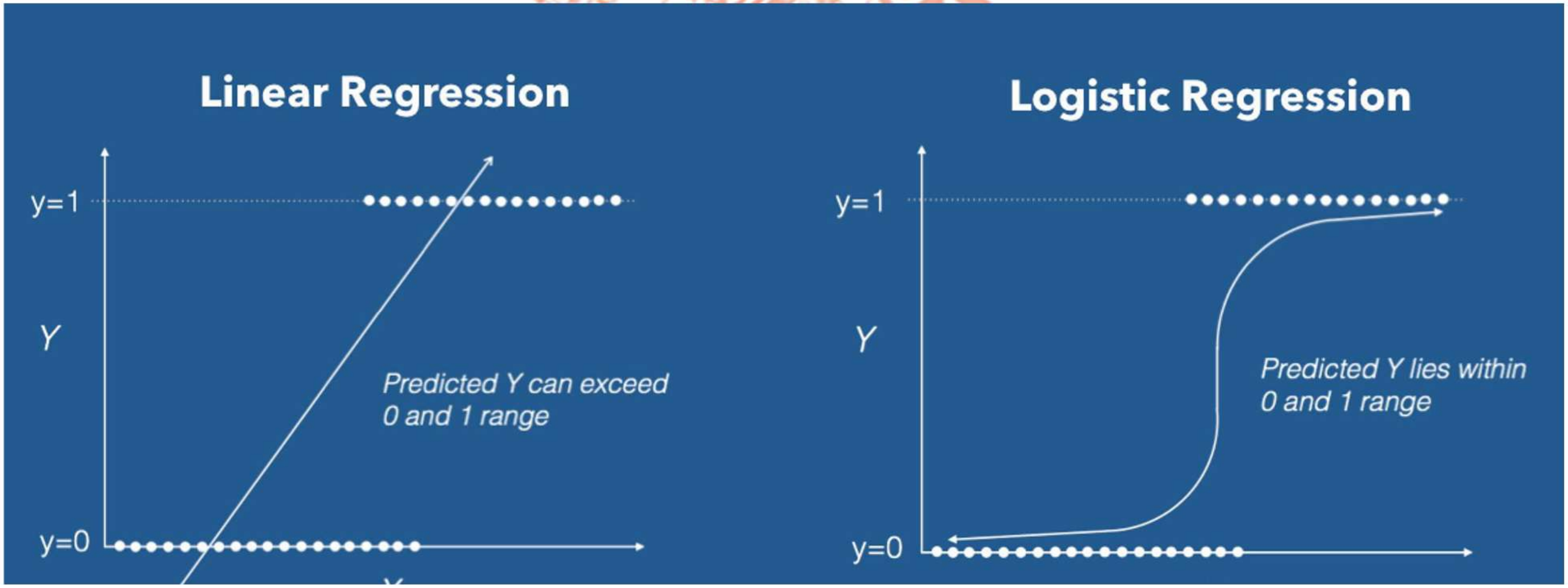
$$\boxed{\log\left(\frac{P}{1-P}\right)} = \beta_0 + \beta_1 X$$

P/(1-P)… odds ratio

Log [P/(1-P)] is a straight line with Beta-0 as intercept , beta-1 as slope

Log [p/(1-p)]= b +ax, p /(1-p)=exp(b).exp(ax), 1/(2p-1)= [1 +exp(a+bx)]/[exp(a+bx)-1]

p = 1/(1+exp(-a-bx))   ;   p= 1/(1+ exp(-y))



- y=seq(0,1,by=.05) ; x=log(y/(1-y))
- plot(y=y,x=x);lines(y=y,x=x)
- log<-glm(data=my.csv.data,Age2cat ~ CSE_3, family=binomial())
- plot(y=log$linear.predictors,x=my.csv.data$CSE_3)

- Summary(log)
- coef(log)
- Y has to be 0 or 1
- Plotting Curve

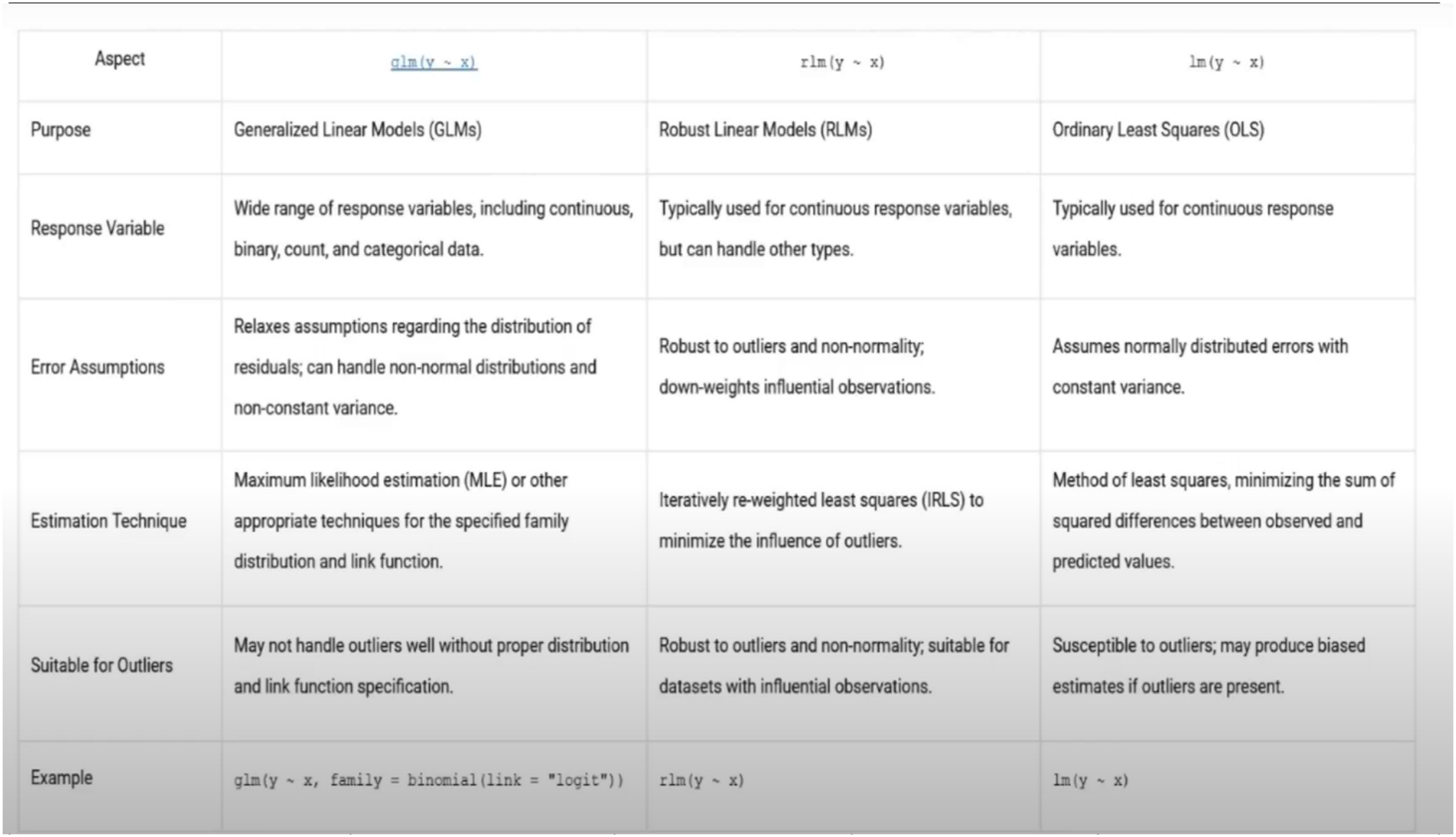
Detailed Write Up on Logistic

Course Name :Basic Statistics using GUI-R (RKWard)

Module : Distributions

Week 6 Lecture : 5

Harsh Pradhan, Assistant Professor,
Institute of Management Studies, BHU
https://bhu.ac.in/Site/FacultyProfile/1_5?FA000562



Aspect	`glm(y ~ x)`	`rlm(y ~ x)`	`lm(y ~ x)`
Purpose	Generalized Linear Models (GLMs)	Robust Linear Models (RLMs)	Ordinary Least Squares (OLS)
Response Variable	Wide range of response variables, including continuous, binary, count, and categorical data.	Typically used for continuous response variables, but can handle other types.	Typically used for continuous response variables.
Error Assumptions	Relaxes assumptions regarding the distribution of residuals; can handle non-normal distributions and non-constant variance.	Robust to outliers and non-normality; down-weights influential observations.	Assumes normally distributed errors with constant variance.
Estimation Technique	Maximum likelihood estimation (MLE) or other appropriate techniques for the specified family distribution and link function.	Iteratively re-weighted least squares (IRLS) to minimize the influence of outliers.	Method of least squares, minimizing the sum of squared differences between observed and predicted values.
Suitable for Outliers	May not handle outliers well without proper distribution and link function specification.	Robust to outliers and non-normality; suitable for datasets with influential observations.	Susceptible to outliers; may produce biased estimates if outliers are present.
Example	`glm(y ~ x, family = binomial(link = "logit"))`	`rlm(y ~ x)`	`lm(y ~ x)`

Aspect	Generalized Linear Model (GLM)	Ordinary Least Squares (OLS)
Scope of Application	Wide range of response variables including continuous, binary, count, and categorical data.	Specifically designed for continuous response variables.
Link Function	Allows for the specification of a link function to model the relationship between the linear predictor and the mean of the response variable.	Does not incorporate a link function; assumes a linear relationship between predictors and response variable.
Assumptions	Relaxes assumptions regarding the distribution of residuals, allowing for different error distributions and non-constant variance.	Assumes normally distributed errors with constant variance (homoscedasticity) and independence of observations.
Estimation Technique	Estimates parameters using maximum likelihood estimation (MLE) or other techniques suited to the specified distribution and link function.	Estimates parameters using the method of least squares, minimizing the sum of squared differences between observed and predicted values.
Applications	Well-suited for modeling a wide range of responses, including binary outcomes (logistic regression), count data (Poisson regression), and categorical outcomes (multinomial regression).	Commonly used for linear regression when modeling continuous outcomes.
Example	`glm(y ~ x, family = binomial(link = "logit"))`	`lm(y ~ x)`

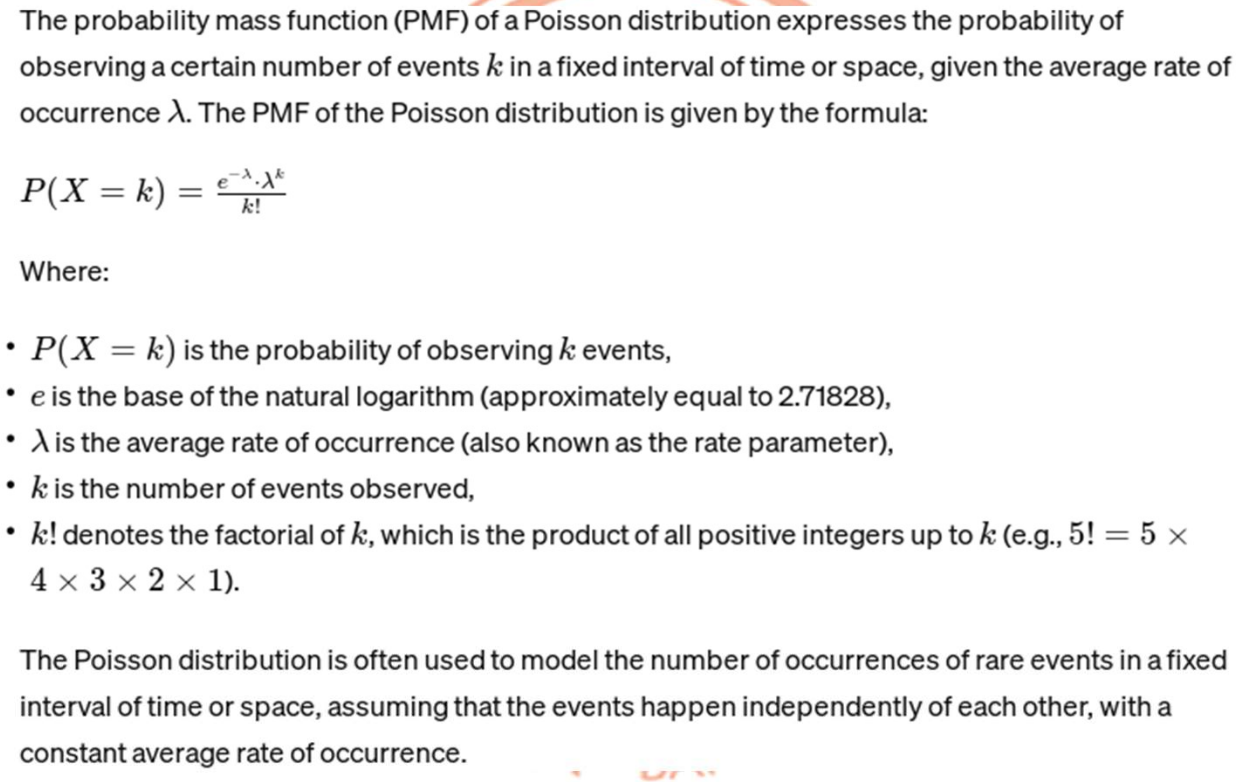
BLUE in Regression

In the context of regression analysis, "BLUE" stands for "Best Linear Unbiased Estimators." It refers to a set of estimators that have several desirable properties:

Best: BLUE estimators are the best among all unbiased linear estimators. This means they have the smallest possible variance compared to other unbiased linear estimators.
Linear: BLUE estimators are linear functions of the observed data. This property ensures that they can be easily computed and interpreted.
Unbiased: BLUE estimators have an expected value (or mean) that equals the true population parameter being estimated. This property ensures that, on average, the estimator does not systematically overestimate or underestimate the true parameter.

Achieving the "best" property (i.e., having the smallest variance) often involves mathematical techniques such as minimizing the mean squared error or maximizing likelihood. In ordinary least squares (OLS) regression, the coefficients obtained are BLUE estimators under the assumptions of the classical linear regression model, assuming that the Gauss-Markov assumptions hold.

In summary, BLUE estimators are highly desirable in regression analysis because they provide estimates of the parameters that are both efficient (minimum variance) and unbiased (accurate).



The probability mass function (PMF) of a Poisson distribution expresses the probability of observing a certain number of events k in a fixed interval of time or space, given the average rate of occurrence λ. The PMF of the Poisson distribution is given by the formula:

$$P(X=k)=\frac{e^{-\lambda}\cdot\lambda^{k}}{k!}$$

Where:

- $P(X=k)$ is the probability of observing k events,
- e is the base of the natural logarithm (approximately equal to 2.71828),
- λ is the average rate of occurrence (also known as the rate parameter),
- k is the number of events observed,
- $k!$ denotes the factorial of k, which is the product of all positive integers up to k (e.g., $5!=5\times4\times3\times2\times1$).

The Poisson distribution is often used to model the number of occurrences of rare events in a fixed interval of time or space, assuming that the events happen independently of each other, with a constant average rate of occurrence.

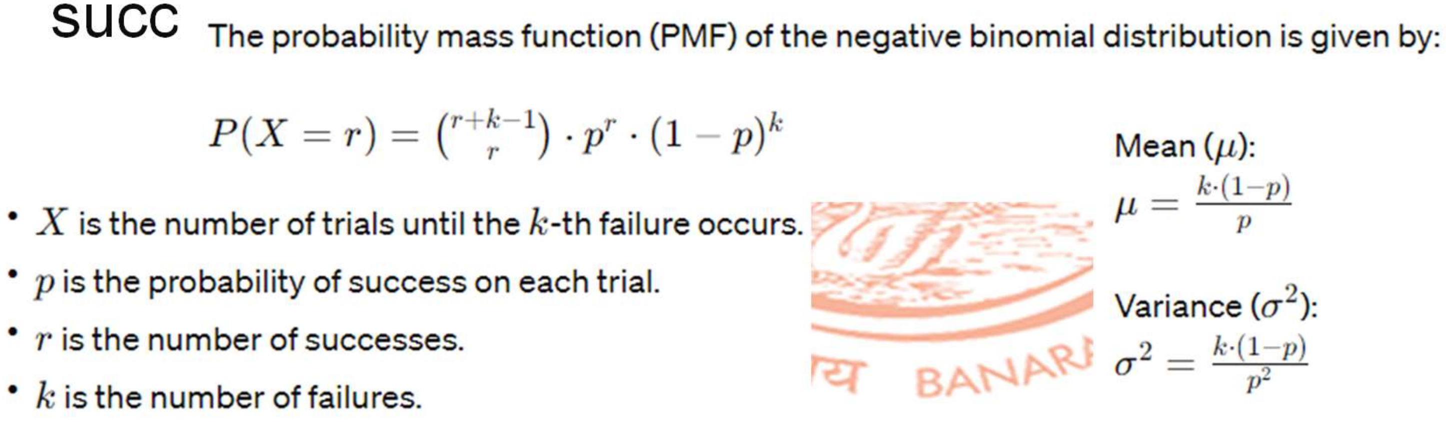
```
# Set the rate parameter (lambda)
lambda <- 2

# Generate random Poisson values
random_values <- rpois(10, lambda)
print(random_values)

# Calculate probabilities for specific values
prob_0 <- dpois(0, lambda)
prob_5 <- dpois(5, lambda)
```

Negative Binomial

- Describes the number of successes in a sequence of independent and identically distributed Bernoulli trials (binary outcomes with a fixed probability of success) before a specified number of failures (or non-successes) occur. In other words, it gives the probability of observing a certain number of succ



The probability mass function (PMF) of the negative binomial distribution is given by:

$$P(X = r) = \binom{r+k-1}{r} \cdot p^r \cdot (1-p)^k$$

- X is the number of trials until the k-th failure occurs.
- p is the probability of success on each trial.
- r is the number of successes.
- k is the number of failures.

Mean (μ):

$$\mu = \frac{k\cdot(1-p)}{p}$$

Variance (σ^2):

$$\sigma^2 = \frac{k\cdot(1-p)}{p^2}$$