# Assignment 1 Report

| S.No | Name | Roll Number |
|---|---|---|
| 1 | Baddepudi Venkata Naga Sri Sai Vineetha | M20CS054 |
| 2 | Parsa Revanth | M20CS058 |
| 3 | Vinit Ramesh Gore | M20CS064 |
| 4 | Anurag Saraswat | M20CS066 |

# Dataset : Life Expectancy Dataset (WHO)

The dataset related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from the United Nation website. In this project we have considered data from the years 2000-2015 for 193 countries for further analysis.

Dataset consists of immunization factors, mortality factors, economic factors, social factors and other health related factors. On initial visual inspection of the data showed some missing values. The observation indicated that most of the missing data was for population, Hepatitis B and GDP. The missing data were from less known countries like Vanuatu, Tonga, Togo,Cabo Verde etc.

Finding all data for these countries was difficult and hence, it was decided that we exclude these countries from the final model dataset. The final dataset consists of 22 Columns and 2938 rows.

# Data Preprocessing

In Total, there are 22 Variables, 19 of them are Numerical, and 3 of them are Categorical. We will need to remove/mutate some variables:

- The rows containing the missing values are dropped from the dataset.
- The non-numeric variables ‘Country’, ‘Year’ and ‘Status’ as they do not contribute to the prediction of the dependent variable ‘Life Expectancy’.
- Correlation between Life Expectancy and other variables is found. The variables which have correlation greater than -0.2 and less than 0.2 like ‘Population’, ‘Measles’, ‘Infant deaths’, ‘Under five deaths’ and ‘Total expenditure’ are removed.
- The dataset is split into the train and test set in the ratio of 80:20.

# Creating the model

Regression analysis is helpful in modeling the relationship between a response or dependent variable and one or more explanatory or independent variables.

# LINEAR MODELS (LINEAR REGRESSION)

## Simple linear regression:

Simple linear regression describes the existence of linear relationship between the response or dependent variable and explanatory variables. The response variable should follow normal distribution.

Suppose a model has to be developed which describes the relationship between the response variable y in terms of explanatory variable x, the general form of the equation for linear regression would be

$$y = ax + b$$

where a and b are constants or coefficients.

## Multiple linear regression:

Multiple linear regression can be used to describe the linear relationship between response or dependent variable and more than one explanatory variable.

Suppose a model has to be developed which describes the relationship between the response variable y in terms of more than one independent variables say $x_1, x_2, x_3, \ldots, x_n$, the general form of the equation for multiple linear regression would be

$$y = a_1x_1 + a_2x_2 + a_3x_3 + \ldots + a_nx_n + b$$

where $b, a_1, a_2, a_3, \ldots$ are coefficients.

## Linear regression model on dataset:

Using the dataset, the life expectancy of people is predicted. As it is dependent on more than one explanatory variable multiple linear regression is used. Fit the model on the train dataset using lm() function. Using the model, the predictions are made on the test data.

# GENERALISED LINEAR MODEL

## Generalized Linear Models:

Generalized Linear Models are an extension to the standard linear models as the response variables can have error distribution rather than just normal distribution. It has three components

1. Random Component:
    - It specifies the probability distribution of the response variable y
2. Systematic Component:
    - It specifies the linear combination of explanatory variables in the linear predictor
3. Link Function:
    - It specifies the link between random and systematic components
    - It explains the relation between the expected value of the response variable and the predicted values using linear predictor

- This is the function g that is applied to each component of the mean of y that relates it to the linear predictor
- It allows the model to generalize well by linking the regression coefficients to the distribution and they can have non-linear link function

**Class of Generalized Linear Models**

| **Distribution** | **Support of distribution** | **Typical uses** | **Link name** | **Link function,** $X\beta = g(\mu)$ | **Mean function** |
|---|---|---|---|---|---|
| Normal | real: $(-\infty, +\infty)$ | Linear-response data | Identity | $X\beta = \mu$ | $\mu = X\beta$ |
| Exponential | real: $(0, +\infty)$ | Exponential response data, scale parameters | Inverse | $X\beta = \mu^{-1}$ | $\mu = (X\beta)^{-1}$ |
| Gamma | | | | | |
| Inverse Gaussian | real: $(0, +\infty)$ | | Inverse squared | $X\beta = \mu^{-2}$ | $\mu = (X\beta)^{-2}$ |
| Poisson | integer: 0, 1, 2, 3, ... | count of occurrences in fixed amount of time/space | Log | $X\beta = ln(\mu)$ | $\mu = exp(X\beta)$ |
| Bernoulli | integer: {0, 1} | outcome of single yes/no occurrence | Logit | $X\beta = ln(\frac{\mu}{1-}$ | $\mu = \frac{1}{1+exp(-X$ |
| Binomial | integer: 0, 1, 2, 3, … , N | count of # of "yes" occurrences out of N yes/no occurrences | | $X\beta = ln(\frac{\mu}{n-}$ | |
| Categorical | integer: [0, K) | outcome of single K-way occurrence | | $X\beta = ln(\frac{\mu}{1-}$ | |
| | K-vector of integer: [0, 1], where exactly one element in the vector has the value 1 | | | | |
| Multinomial | K-vector of integer: [0, N] | count of occurrences of different types (1 .. K) out of N total K way occurrences | | | |

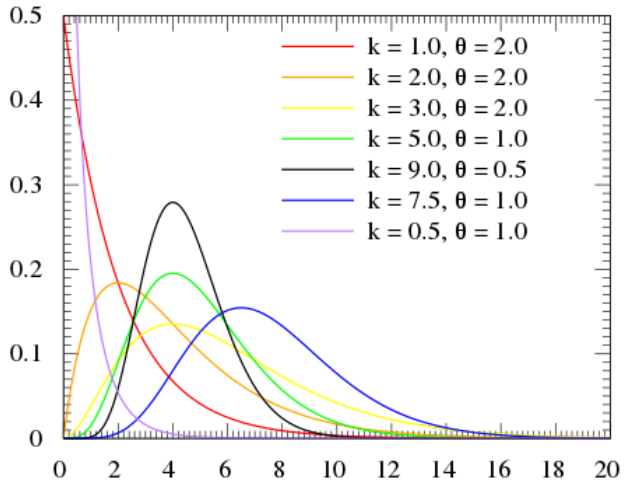
Source: https://en.wikipedia.org/wiki/Generalized_linear_model#Logit_link_function

## Gamma Regression

Gamma regression assumes Gamma distribution as the underlying distribution of the response variable. Thus the relation between the mean and the variance of the response variable is fixed. More precisely, it is given as:

$$\frac{Var(y_i)}{E(y_i)} = \beta_i$$

where $\beta$ is the scale parameter.




A Typical Gamma Distribution, k is the shape parameter, θis the scale parameter.

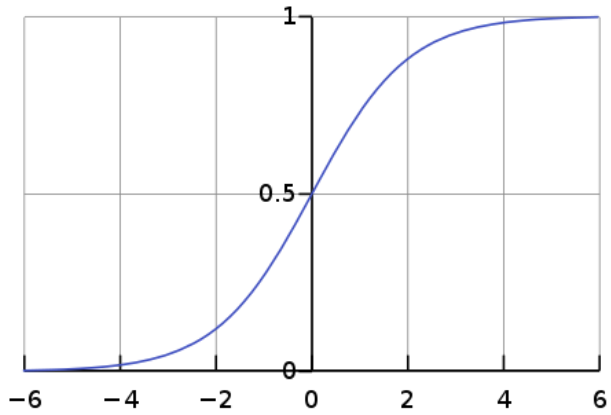
## Logistic Regression:

Logistic regression is used for predicting the probability of the response variables given the set of independent variables. It estimates the probabilities using the underlying logit function.

$$\eta = logit(\boldsymbol{\pi}) = log(\boldsymbol{\pi} / (1 - \boldsymbol{\pi})$$

The general form of the equation for logistic regression is

$$y = 1/(1 + e^{-(b + a_1x_1 + a_2x_2 + a_3x_3 + \ldots)})$$



Source: Wikipedia

# SIGNIFICANCE OF COEFFICIENTS

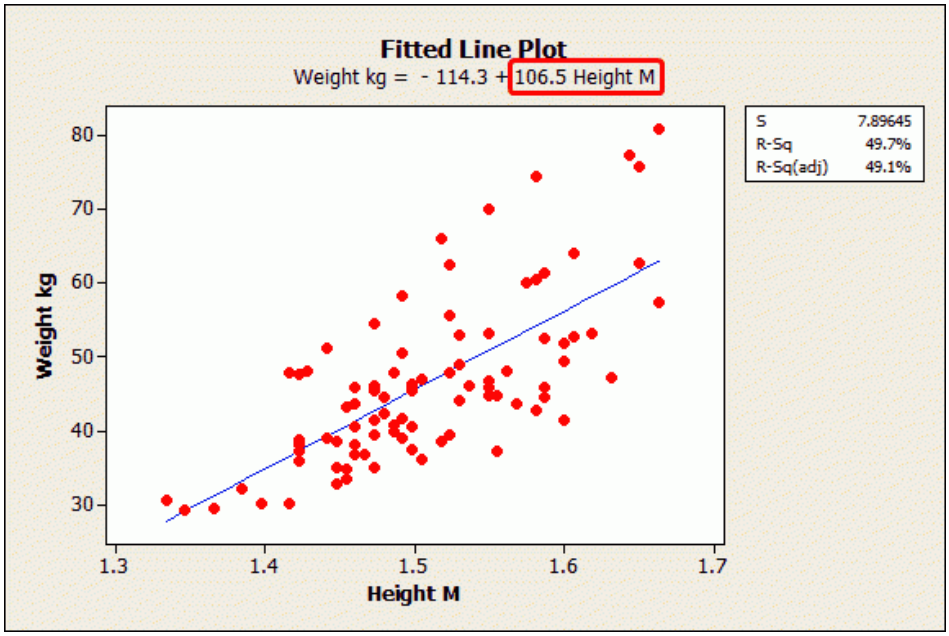
The equation of the line formulated by the multiple Linear Regression (MLR) model that has one dependent variable and *n* independent variables is given as follows:

$$y \sim a_1x_1 + a_2x_2 + a_3x_3 + \ldots + a_nx_n + b$$

Here y is the dependent variable and $x_i$'s are independent, b is the intercept and $a_i$'s are the coefficients. Coefficient $a_i$ determines the mathematical relationship between $x_i$ and y. The sign of coefficient determines whether the variable is positively or negatively correlated with y.

Let us consider an example to understand the significance of coefficients better. Given below is an LR model where "Weight kg" depends upon "Height M".

| **Coefficients:** | | | | | |
|---|---|---|---|---|---|
| | Estimate | Std. Error | t value | Pr(>\|t\|) | Significance |
| (Intercept) | -114.326 | 17.4425 | -6.5544 | 0 | *** |
| Height M | 106.505 | 11.55 | 9.22117 | 0 | *** |




Source: https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/

The coefficient of the "Height M" is 106.5. It means that the variable "Weight kg" would increase with value 106.5 for every unit increase of "Height M".

The coefficients table shows values that help us decide the statistical significance of each variable in the prediction process. A zero significant variable will have no effect on the dependent variable. The different columns shown in the table help us to determine whether to keep the variable in the model or not.

**Estimates**

In Table 1, the estimate column shows the estimated coefficient values. These values can be plugged into the LR formula to get an approximate prediction of "Weight kg".

**Std. Error**

The Std. Error column gives the deviation from these estimates to generate a range of possible values for coefficients. These are the values that are possible to be estimated when samples from the same population with the same sample size are used to fit the LR model.
The smaller these values, the more precise is the model.

**t-value**

The t-value column contains the ratio of estimate and Std. Error. Generally, the confidence interval is 95% i.e. when the t-value lies within the confidence interval, then the estimate is considered to be significant in contributing to the prediction. A larger t-value shows that the estimate is more precise. The t-values close to 0 indicate imprecision of the estimate.

**p-value (Pr(>|t|)**

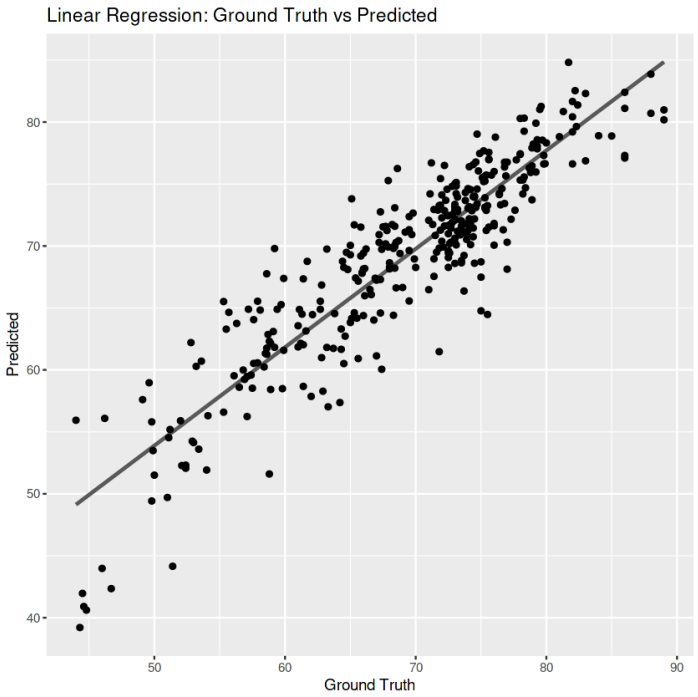
The p-values are used to decide the statistical significance of the independent or response variable. If the p-value is less than the significance level, the estimate is considered to be significant. Generally, the significance level is kept to be 0.05.

**Significance codes**

These are visual cues to quickly determine significant variables. A variable having more stars beside it is more significant. Those having no cue beside them are the least significant.

# RESULTS

## Multiple Regression Model



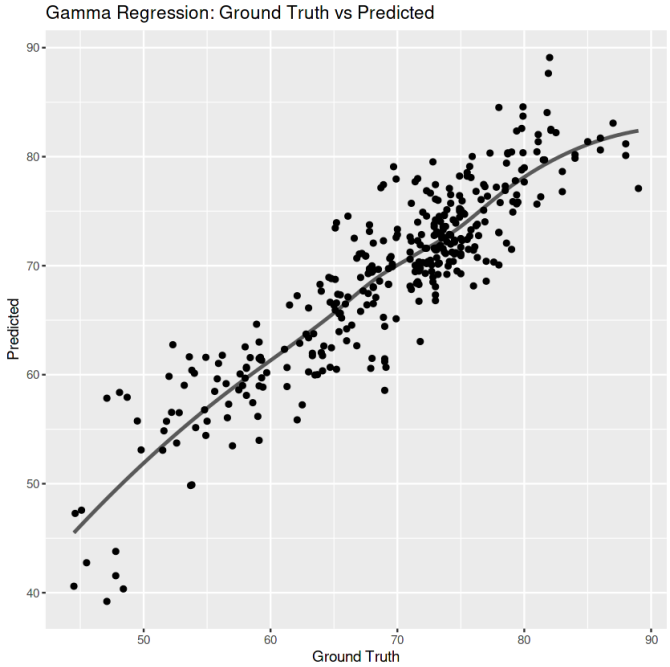

The above figure shows the plot between the predicted and ground truth values of "Life.expectancy". We could see that the regression line learns the best line to fit through this data. Coming to the dataset we used, the Multiple Regression model was at first fitted over all other variables with "Life.expectancy" as the dependent variable. The summary of the MLR model shows the significance of coefficients (Refer Table 1).

**Table 1: Significance of Coefficients**

```
Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     5.235e+01  7.915e-01  66.139  < 2e-16 ***
Adult.Mortality                -1.798e-02  1.056e-03 -17.025  < 2e-16 ***
Alcohol                        -1.123e-01  3.395e-02  -3.309 0.000962 ***
percentage.expenditure          4.192e-04  2.102e-04   1.995 0.046294 *
Hepatitis.B                    -5.181e-03  5.172e-03  -1.002 0.316668
BMI                             3.653e-02  6.695e-03   5.456 5.84e-08 ***
Polio                           1.396e-02  6.099e-03   2.288 0.022284 *
Diphtheria                      2.058e-02  6.567e-03   3.134 0.001763 **
HIV.AIDS                       -4.353e-01  1.996e-02 -21.814  < 2e-16 ***
GDP                             4.632e-06  3.305e-05   0.140 0.888570
thinness..1.19.years           -2.885e-02  5.605e-02  -0.515 0.606847
thinness.5.9.years             -4.813e-02  5.535e-02  -0.870 0.384661
Income.composition.of.resources  9.842e+00  9.353e-01  10.523  < 2e-16 ***
Schooling                       9.353e-01  6.612e-02  14.147  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the coefficient estimates and their p-values, we can conclude that the variables "HIV.AIDS", "Schooling", "Income.composition.of.resources", "Adult.Mortality", "Alcohol" and "BMI" have best significance and are the best contributors to the model. These should be included in the model fitting.

## Gamma Regression Model:



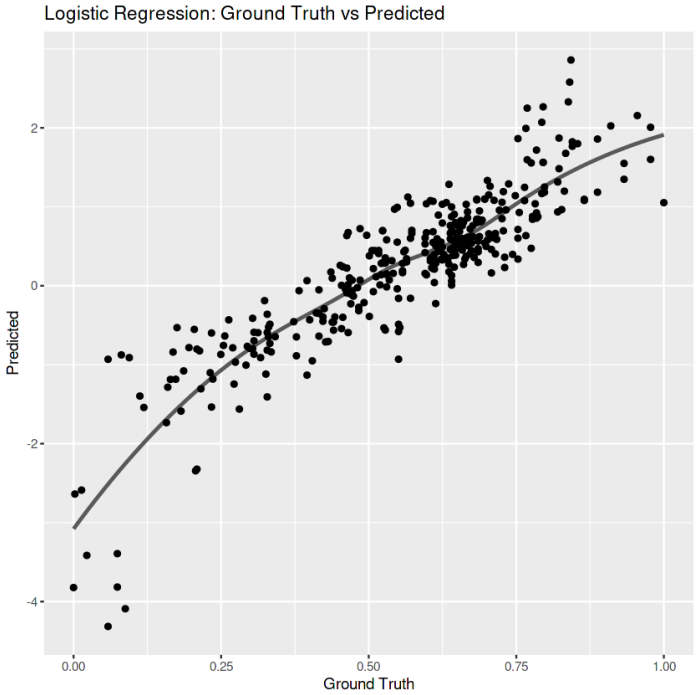

The above figure shows the plot between the predicted and ground truth values of "Life.expectancy". We could see that the gamma regression learns the best curve to fit through this data. The gaussian regression model was at first fitted over all other variables with "Life.expectancy" as the dependent variable. The summary of the MLR model shows the significance of coefficients (Refer Table 2).

**Table 2: Significance of Coefficients**

```
Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     1.785e-02  1.780e-04 100.286  < 2e-16 ***
Adult.Mortality                 4.207e-06  2.497e-07  16.851  < 2e-16 ***
Alcohol                         2.248e-05  6.995e-06   3.214  0.00134 **
percentage.expenditure         -5.557e-08  3.857e-08  -1.441  0.14991
Hepatitis.B                     5.699e-07  1.075e-06   0.530  0.59614
BMI                            -6.430e-06  1.377e-06  -4.669 3.34e-06 ***
Polio                          -2.735e-06  1.362e-06  -2.008  0.04481 *
Diphtheria                     -4.181e-06  1.438e-06  -2.907  0.00371 **
HIV.AIDS                        1.524e-04  5.816e-06  26.211  < 2e-16 ***
GDP                             1.379e-09  6.106e-09   0.226  0.82142
thinness..1.19.years            1.070e-05  1.287e-05   0.831  0.40597
thinness.5.9.years              6.446e-06  1.275e-05   0.506  0.61318
Income.composition.of.resources -2.259e-03  2.132e-04 -10.591  < 2e-16 ***
Schooling                      -1.789e-04  1.419e-05 -12.615  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the coefficient estimates and their p-values, we can conclude that the variables "HIV.AIDS", "Schooling", "Income.composition.of.resources", "Adult.Mortality" and "BMI" have best significance and are the best contributors to the model. These should be included in the model fitting.

## Logistic Regression Model:




The above figure shows the plot between the predicted and ground truth values of "Life.expectancy". We could see that the logistic regression learns the best line to fit through this data. The logistic regression model was at first fitted over all other variables with "Life.expectancy" as the dependent variable. The summary of the MLR model shows the significance of coefficients (Refer Table 3).

**Table 3: Significance of Coefficients**

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -0.86845    0.41748  -2.080 0.037507 *
Adult.Mortality                 -1.28730    0.49449  -2.603 0.009234 **
Alcohol                         -0.14629    0.36781  -0.398 0.690838
percentage.expenditure           1.38372    2.74063   0.505 0.613634
Hepatitis.B                     -0.04085    0.30913  -0.132 0.894858
BMI                              0.21357    0.29938   0.713 0.475605
Polio                            0.07403    0.34418   0.215 0.829708
Diphtheria                       0.18360    0.38040   0.483 0.629335
HIV.AIDS                        -3.41598    0.96989  -3.522 0.000428 ***
GDP                             -0.02421    2.64826  -0.009 0.992706
thinness..1.19.years            -0.05373    0.86242  -0.062 0.950321
thinness.5.9.years              -0.13460    0.88761  -0.152 0.879465
Income.composition.of.resources  0.78142    0.50130   1.559 0.119046
Schooling                        1.52937    0.66055   2.315 0.020597 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the coefficient estimates and their p-values, we can conclude that the variables "HIV.AIDS" have best significance and are the best contributors to the model. These should be included in the model fitting.

# REFERENCES


1. Foundations of Linear and Generalized Linear Models By Alan Agresti
2. Coefficients for Binary Logistic Regression - Minitab Express
3. Understanding Linear Regression Output in R | by Christian Thieme | Towards Data Science
4. How to Interpret P-values and Coefficients in Regression Analysis - Statistics By Jim
5. https://online.stat.psu.edu/stat504/lesson/6/6.1
6. https://cran.r-project.org/web/packages/GlmSimulatoR/vignettes/exploring_links_for_the_gaussian_distribution.html
7. https://en.wikipedia.org/wiki/Generalized_linear_model