DATA 583 Regression Report

Ujjwal Upadhyay, Shveta Sharma, Varshita Kyal

2023-03-24

Background

The dataset undertaken for analysis is regarding the life expectancy of 193 countries for a period of 16 years (2000-2015) along with the factors impacting it. The factors are mainly classified broadly into four categories namely immunization, mortality, economic and social. The purpose of this report is to examine how these 20 explanatory variable contribute to the life expectancy. The link to the dataset : https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

Scientific Hypothesis

1) *First Research Hypothesis* :

Null Hypothesis (H_0) : All predictors influencing life expectancy in "Developed" or "Developing" countries have same effect.

Alternate Hypothesis (H_a) : One or More predictors influencing life expectancy in "Developed" or "Developing" countries have different effect.

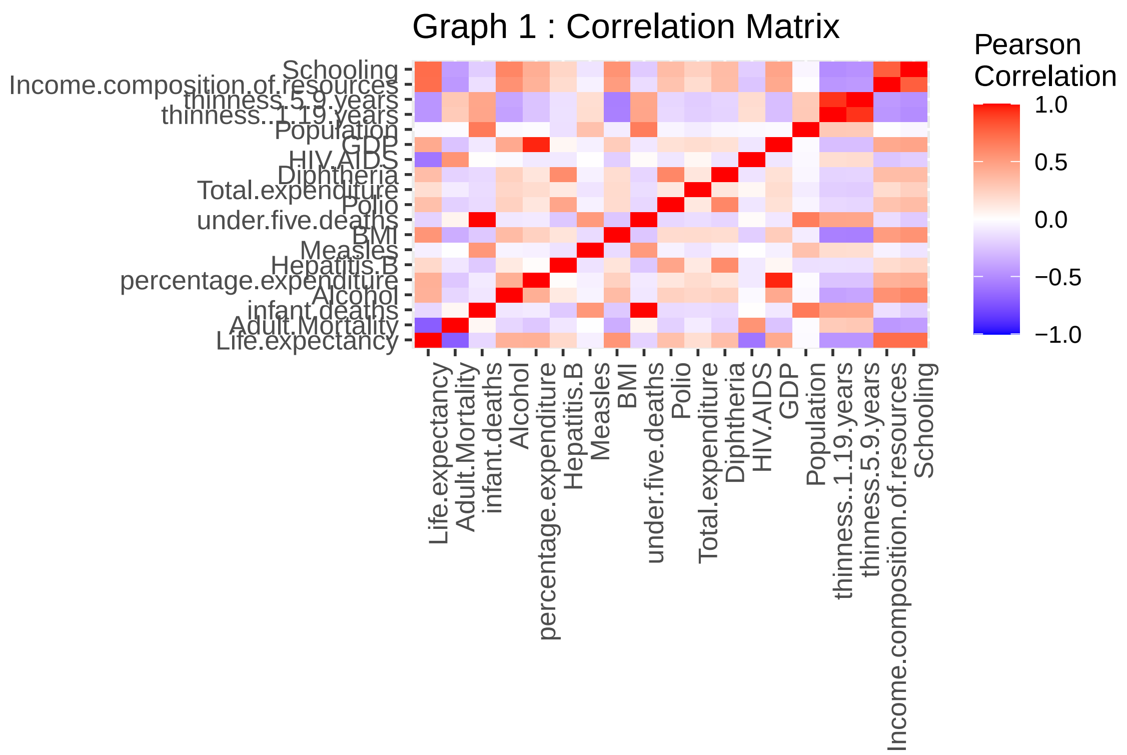
For example, we think expenditure on health as a percentage of Gross Domestic Product(GDP) per capita is a strong and significant factor affecting life expectancy in "Developing" countries but it might not have an equal amount of effect on life expectancy in "Developed" countries. Therefore, throughout the project we will build different models to best fit the data and check whether or not our hypothesis are correct.

2) *Second Research Hypothesis* :

Null Hypothesis (H_0) : All the independent variables such as life expectancy, adult mortality, alcohol consumption, percentage expenditure of GDP on health, hepatitis B, measles, bmi, etc are not significant predictors of whether a country is "Developed" or "Developing".

Alternative Hypothesis (H_a) : At least one of the independent variables such as life expectancy, adult mortality, alcohol consumption, percentage expenditure of GDP on health, hepatitis B, measles, bmi, etc are significant predictors of whether a country is "Developed" or "Developing".

Multicollinearity


Graph 1 : Correlation Matrix

VIF (Variance Inflation factor)

Table 1: VIF values for the predictor variables

Predictor	VIF
Adult.Mortality	1.686903
infant.deaths	180.960466
Alcohol	1.872242
percentage.expenditure	8.033868
Hepatitis.B	1.626886
Measles	1.424736
BMI	1.918680
under.five.deaths	179.985220
Polio	1.900256
Total.expenditure	1.157331
Diphtheria	2.344290
HIV.AIDS	1.456909
GDP	8.690851
Population	1.495219
thinness..1.19.years	7.617167
thinness.5.9.years	7.728310
Income.composition.of.resources	3.301878
Schooling	3.829151

As we observed in Pearson correlation plot in exploratory data analysis, there is high positive correlation between infant deaths and under five deaths. There were some other predictors that demonstrated multi-collinearity. To detect multi-collinearity, we used various statistical methods, such as correlation matrices, and variance inflation factors (VIF). As multicollinearity is detected, we reduced its effects by removing one or more of the highly correlated independent variables. For instance, we removed infant deaths and only kept under five deaths. From the figure above,we can see that VIF for infant deaths and under five deaths is over 10. We kept a cutoff point of VIF=10 to reduce multi-collinearity.

Linearity Assumption

The response variable "Life Expectancy" is nearly normally distributed. We did observe some variables that were skewed but since they represented natural variation, we didn't get rid of the outliers. We expected data to be somewhat skewed for those attributes as there is huge variation and diversification in countries.

Regression Analysis

For the first research hypothesis, we would perform regression analysis with two different families of GLM. We have included interaction term, "Developed" which is a binary variable categorizing the countries in dataset as "Developed" or "Developing". One of the families is gaussian family. As the response variable i.e. "Life Expectancy" is continuous and normally distributed, we opted to use the gaussian (normal) distribution to model it. The other family we used for comparison with gaussian is gamma family as "Life Expectancy" histogram was somewhat positively skewed and gamma family can handle continuous data with a positive skew that cannot be modeled well using the gaussian distribution. Moreover, In GLMs, the gamma distribution is often used with a logarithmic link function to ensure that the predicted values are positive which is acceptable case in terms of life Expectancy.

Variable Selection

Generalized Linear Model(GLM) (family - " gaussian") : One of the models we used is Generalized Linear Model(GLM) (family - " gaussian") with interaction term using a new binary column "Developed", wherein, 1 represents "Developed" and 0 represents "Developing". This binary attribute is tested for interaction with other continuous predictors like adult mortality rate, alcohol consumption, expenditure on health as a percentage of gross domestic product per capita(%), hepatitis B, measles, BMI, under five deaths, polio, general government expenditure on health as a percentage of total government expenditure(%), diphteria, HIV/AIDS, GDP, population, malnutrition(1-19 years), malnutrition(5-9 years), income composition of resources and schooling. We tried both AIC(Alkaline Information Criterion) and BIC (Bayesian Information Criterion) with backward elimination, forward and "both" selection techniques to model to select variables for best model selection. Backward and "both" selection techniques provides similar results with close to 18 predictors values whereas forward selection technique doesn't work well as it fails to provide a parsimonious model. Moreover, In between AIC and BIC, BIC provided a list 15 significant predictors resulting in a much simpler model.

Generalized Linear Model(GLM) (family - "Gamma(link='log')") : GLM with gamma family provided very similar results compared to gaussian family. The technique used for variable selection was similar to other models to keep it consistent. The significant and non-significant predictors in best model were also similar. We would further do a model comparison to determine which model provides better metric scores for test results.

Model Comparison

Table 2: Anova table for GLM (family= gaussian)

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1474	22765.11	NA	NA	NA
1457	22441.04	17	324.0667	0.2244948

The difference in deviance between the two models (full and stepwise) of GLM with gaussian. A large deviance indicates a poor fit. In this case, the deviance is 324. The p-value for a chi-squared test of the null hypothesis that the difference in deviance between the two models is equal to the degrees of freedom. In this case, the p-value is 0.224, which indicates that there is no significant difference in deviance between the full model and stepwise model at the 0.05 level of significance. The high p-value suggests that the difference in deviance between the two models is not statistically significant, meaning that there is not enough evidence to reject the null hypothesis that both models fit the data equally well. Overall, the analysis suggests that there is not enough evidence to prefer full model over stepwise model as a better fit for the data.

Table 3: Anova table for GLM (family= gamma(log='link'))

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
2112	8.050062	NA	NA	NA
2095	7.956233	17	0.0938289	0.0915463

Here, we see that for GLM with family = gamma has deviance of 0.09 and p-value is 0.09, which indicates that there is no significant difference in deviance between the full model and stepwise model at the 0.05 level of significance.The high p-value suggests that the difference in deviance between the two models is not statistically significant, meaning that there is not enough evidence to reject the null hypothesis that both models fit the data equally well. Overall, the analysis suggests that there is not enough evidence to prefer full Model over stepwise model for GLM (family=gamma) as a better fit for the data.

Table 4: Model comparison results

Model	AIC	BIC	Residual_Deviance	Null_Deviance
Full Model - gaussian	8337.489	8517.911	22441.041165	139563.4638
Stepwise Model - gaussian	8324.852	8415.063	22765.107903	139563.4638
Full Model - gamma	12136.130	12328.670	7.956233	45.8037
Stepwise Model - gamma	12127.095	12223.365	8.050062	45.8037

The AIC (Akaike information criterion) and BIC (Bayesian information criterion) are both measures of model fit that balance model complexity and goodness of fit. In general, lower AIC and BIC values indicate better model fit.

In this case, the AIC and BIC values of the stepwise model are lower than those of the full model, indicating that the stepwise model has a better balance of model complexity and goodness of fit. However, the residual deviance of the full models is lower than that of the stepwise models, suggesting that the full model has a better fit to the data.

Overall, while the stepwise model - gaussian has a better balance of complexity and goodness of fit according to the AIC and BIC, the full model - gaussian may still have a better fit to the data according to the residual deviance.

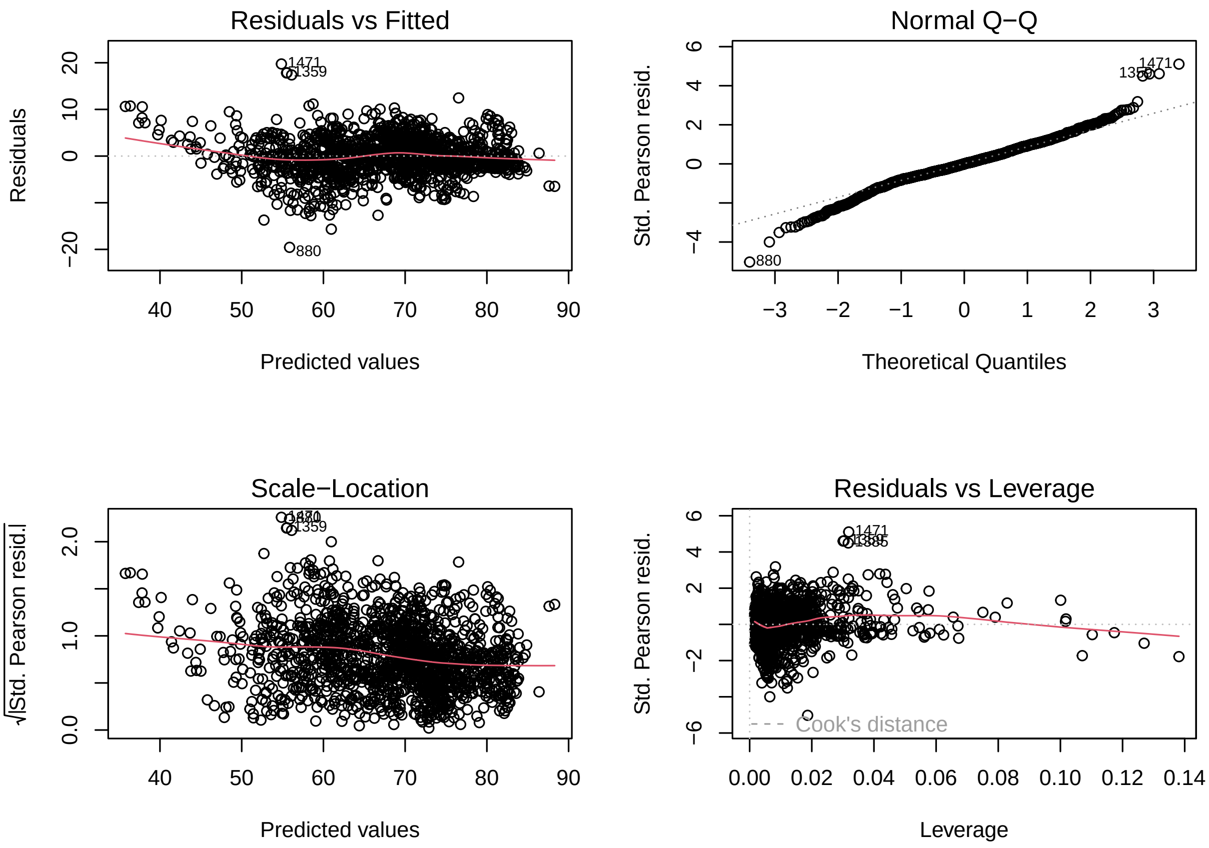
Table 5: Model comparison results

Model	Adj_R_Squared	Test_MSE
Full Model - gaussian	0.8367494	14.57522
Stepwise Model - gaussian	0.8357250	14.40608
Full Model - gamma	0.8236439	4183.64208
Stepwise Model - gamma	0.8230004	4183.55120

As per the given table, we can observe that GLM (family = "Gaussian") stepwise model has adjusted r^2 value of 0.83, hence, it can explain 83% variation in model. The test MSE is also 14.5 which is lower than gamma family GLM full/stepwise model. We do observe that full model and stepwise models are quite similar in terms of test statistics, but, for sake of parsimony, we would go for stepwise model as it is simpler with 95% of predictors in the model being highly significant. But before making that selection, we should have a look at diagnostic plots for stepwise models of both gaussian and gamma families.

Diagnostic Plots

Graph 2: GLM(family = Gaussian) Stepwise Regression Plot



Scatter plot of residuals vs. fitted values: There was no pattern in the plot as the predicted values of the dependent variable and the residuals are scattered. Hence, the linear assumptions are met such as non-linearity, heteroscedasticity, or outliers.

Normal probability plot (QQ plot) of residuals: The residuals are normally distributed, the points in the plot will fall approximately along a straight line.

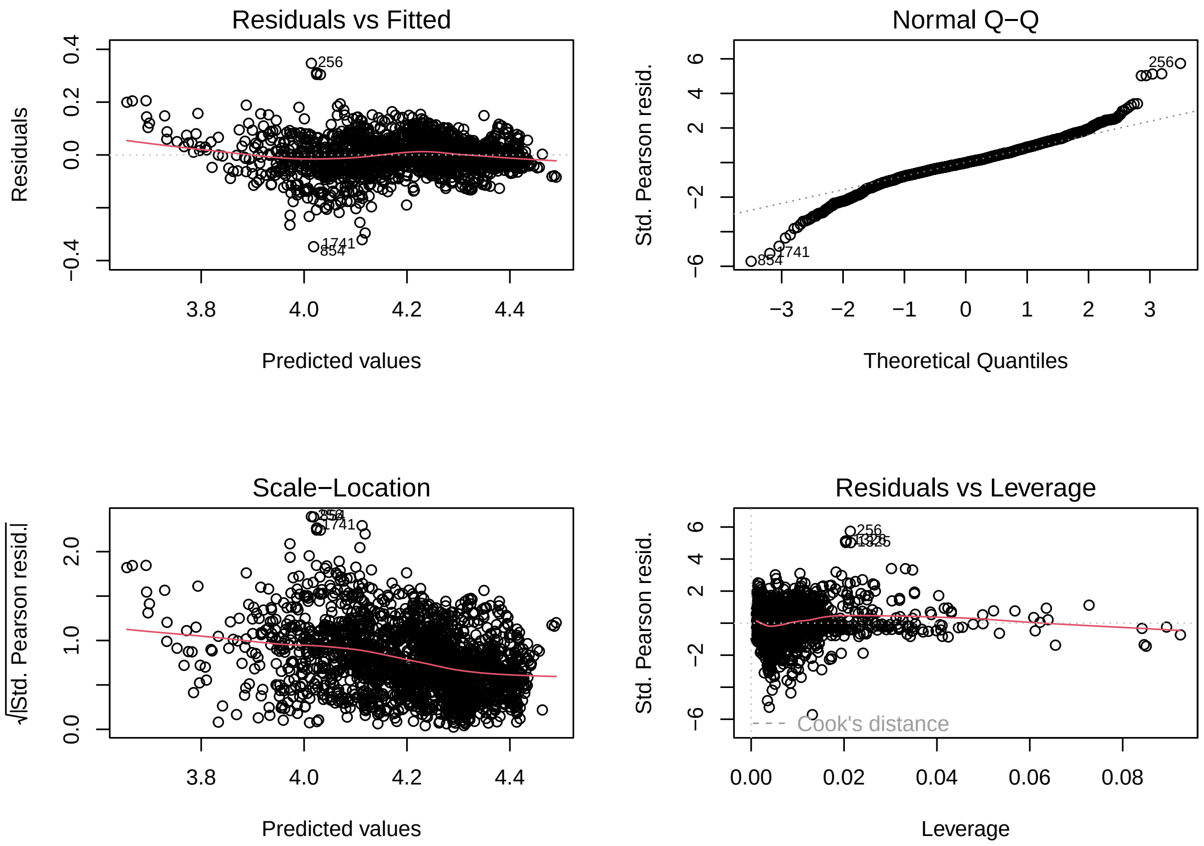
Scale-location plot: This plot shows that the variance is constant, the points in the plot are evenly scattered around a horizontal line.

Residuals vs. leverage plot: This plot shows that there are no High leverage observations or outliers that might influence on the model.

Cook's distance plot: This plot shows there are no observations with a high Cook's distance that may have a large influence on the model.

Hence, there are no major issues with this GLM model and it seems like a good fit on our dataset for analyzing life expectancy.

Graph 3: GLM(family = Gamma (log='link')) Stepwise Regression Plot



The diagnostic plots for gamma family GLM model are very similar to gaussian family. Hence, there is no major issue in this GLM gamma family stepwise model.

Logistic Regression

For second research hypothesis we perform classification using logistic regression. It is a statistical method used to analyze and model the relationship between a dependent variable (also called the response or target variable) and one or more independent variables (also called predictors or features). Unlike linear regression, which models a continuous response variable, logistic regression models the probability of a binary outcome (i.e., whether an event will or will not occur). It is a type of generalized linear model (GLM) that uses a logit link function to transform the probability of the binary outcome to a linear model. The output of a logistic regression model is typically the predicted probability of the binary outcome for a given set of predictor values. The model is trained by maximizing the likelihood of the observed data, and the coefficients of the independent variables represent the log odds ratio of the outcome for a unit change in the predictor or feature variable, keeping all other variables as constant.

Modelling

We have splitted the whole dataset into train and test sets with 70% used for training and remaining 30% as testing. Logistic regression is fitted on the training dataset which showed many predictors such as measles, bmi, polio as insignificant, so we used BIC(Bayesian Information Criterion) in a backward selection procedure for feature selection which gave only five predictors as significant in classifying the country as developed or developing.

Table 6: Confusion Matrix

	Predicted Developed	Predicted Developing
Actual Developed	75	21
Actual Developing	16	526

The above selected model is used to make predictions on the test dataset. The predicted probabilities are converted to predicted classes using a decision threshold of 0.5 which is used to compute the confusion matrix which is a table that compares the actual and predicted labels of a classification model. The above confusion matrix correctly predicted 75 samples as “Developed”, incorrectly predicted 21 samples as “Developing” when they were actually “Developed”, incorrectly predicted 16 samples as “Developed” when they were actually “Developing” and correctly predicted 526 samples as “Developing”.

Model Comparison

Table 7: Anova table for Logistic Regression

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1484	366.3203	NA	NA	NA
1471	351.0440	13	15.27636	0.2904191

The difference in deviance between the two models (full and stepwise) of Logistic Regression. A small deviance indicates a good fit. In this case, the deviance is 15.276. The p-value for a chi-squared test of the null hypothesis that the difference in deviance between the two models is equal to the degrees of freedom. In this case, the p-value is 0.2904, which indicates that there is no significant difference in deviance between the full model and stepwise model at the 0.05 level of significance. The high p-value suggests that the difference in deviance between the two models is not statistically significant, meaning that there is not enough evidence to reject the null hypothesis that both models fit the data equally well. Overall, the analysis suggests that there is not enough evidence to prefer full model over stepwise model as a better fit for the data.

Performance Metrics of Logistic Regression

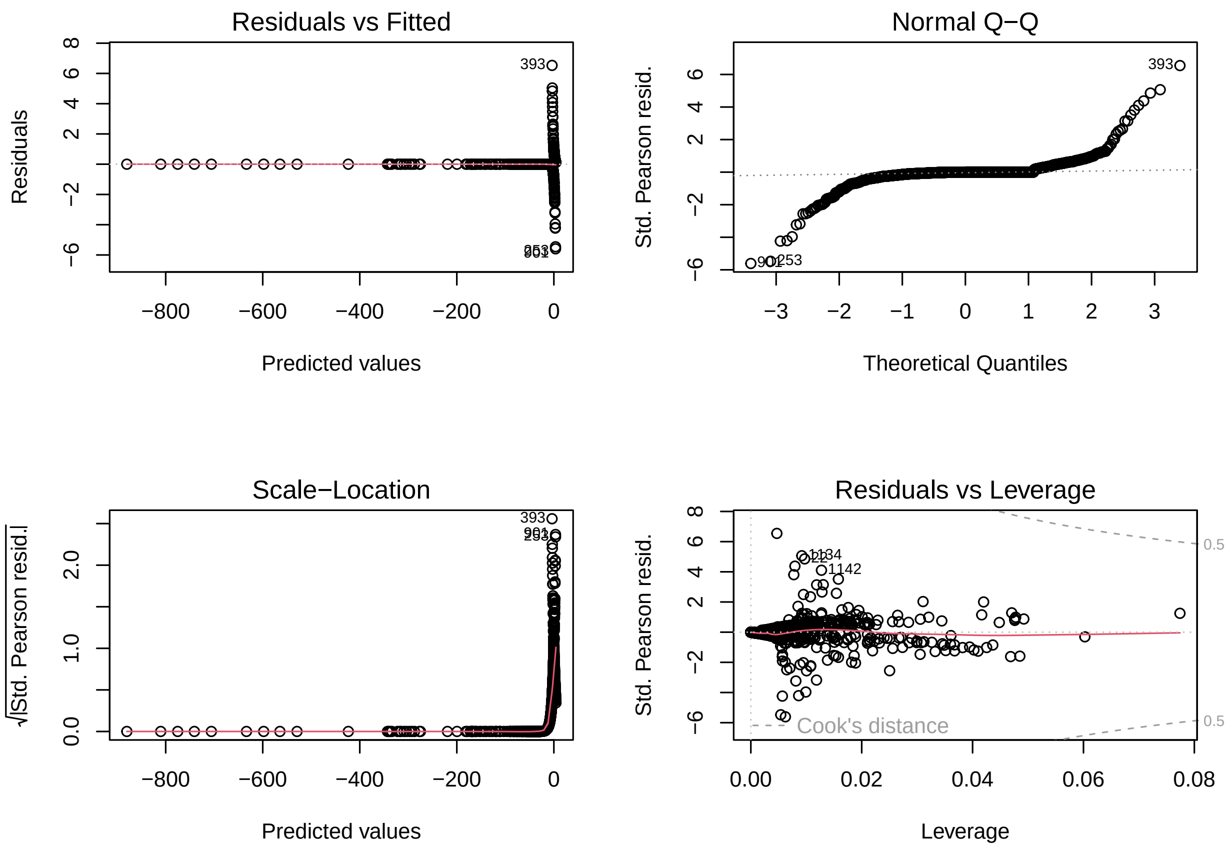
Table 8: Performance metrics for the model

Metric	Value
Accuracy	0.9420
Precision	0.9616
Recall	0.9705
F1 score	0.9660

From the confusion matrix we can compute the performance metrics as accuracy, precision, recall, and F1 score.

Accuracy: It measures the proportion of correct predictions made by the model. The accuracy of the model is 0.942, which means that the model has correctly predicted 94.2% of the cases. **Precision**: It measures the proportion of true positives (correctly predicted positive cases) among all the positive predictions made by the model. The precision of the model is 0.9616 which means that when the model predicts a country as developed correctly 96.16% of the time. **Recall**: The recall of the model is 0.9705, which means that the model has correctly identified 97.05% of the developed countries. **F1 score**: It is the harmonic mean of precision and recall, which provides a balanced measure of both precision and recall. In this case, the F1 score of the model is 0.966, which means that the model's overall performance is good considering both precision and recall.

Graph 4 : Diagnostic Plots - Classification



The diagnostic plot for a logistic regression model suggests:

Residual plot: The residuals are not distributed approximately symmetrically distributed around zero.

Normal Q-Q plot: The plot of the residuals does not follow a straight line, indicating there no normality of the residuals.

Scale-location plot: The plot of the absolute residuals against the fitted values shows some patterns, indicating that the variance of the residuals is not constant across the range of the fitted values.

Cook's distance plot: The plot of Cook's distance against the observation number does not show any observations with large values, indicating the absence of influential outliers.

Therefore, this diagnostic plot for a logistic regression model show some patterns, indicating that the model does not fit the data well and some statistical assumptions are compromised.

Results

Regression Results

As discussed in detail above, a stepwise GLM (with interactions) model with family="gaussian" is the best model in terms of parsimony, AIC, BIC, Test MSE and adjusted R^2.

GLM model,where the response variable is Life expectancy and the statistically significant predictors are adult mortality rate, alcohol, percentage expenditure, BMI, polio, diphtheria, HIV AIDS, income composition of resources, schooling, percentage expenditure:developed, BMI:developed, malnutrition(1-19)years:developed, income composition of resources:developed, and schooling:developed.

A one-unit increase in Adult Mortality is associated with a decrease in Life expectancy of 0.0128 years, holding all other variables constant. Similarly, a one-unit increase in Alcohol is associated with a decrease in Life expectancy of 0.226 years, holding all other variables constant.

Some of the variables have interactions with Developed, which means their effect on life expectancy is different depending on whether the country is developed or not. For example, a one-unit increase in Income composition of resources is associated with an increase in Life expectancy of 8.134 years, holding all other variables constant, but this effect is stronger for developed countries, as the interaction term Income composition of resources:Developed has a coefficient of 33.655. Also, the coefficient for percentage expenditure is 0.00106, which means that for every one-unit increase in percentage expenditure, the outcome variable is expected to increase by 0.00106 units, holding all other variables constant. However, the coefficient for the interaction term "percentage expenditure:Developed" is -0.001066, which means that the effect of percentage expenditure on the outcome variable is different for developed countries compared to non-developed countries. Specifically, the effect is expected to be lower for developed countries compared to non-developed countries.

Similarly, the coefficients for "BMI" and "thinness..1.19.years" indicate their effects on the outcome variable, while the coefficients for the interaction terms "BMI:Developed" and "thinness..1.19.years:Developed" indicate how these effects vary for developed countries compared to non-developed countries.

Overall, this model suggests that the effects of certain predictor variables on the outcome variable may vary depending on whether a country is considered developed or not.The p-values of the coefficients indicate their statistical significance. The smaller the p-value, the stronger the evidence against the null hypothesis of no effect. All the coefficients except for thinness..1.19.years are statistically significant at a 5% significance level.

The final equation of the model is:

$LifeExpectancy = 49.128015 - 0.012866 * AdultMortality - 0.226491 * Alcohol + 0.001060 * percentageexpenditure + 0.066353 * BMI + 0.026517 * Polio + 0.025679 * Diphtheria - 0.493725 * HIV/AIDS - 0.018728 * thinnessamongchildrenunder5years + 8.134580 * HDI + 0.944727 * Schooling - 0.001066 * percentageexpenditure :$

$Developed - 0.066152 * BMI : Developed - 1.485421 * thinnessamongchildrenunder5years : Developed + 33.655204 * HDI : Developed - 1.258324 * Schooling : Developed$

where HDI is the Income Composition of Resources in terms of the Human Development Index.

Classification Results

The logistic regression model is predicting the probability of a country being categorized as "Developed" based on the following predictor variables: adult mortality, alcohol, hepatitis B, under-five deaths, and income composition of resources.

The equation for the logistic regression model is:

$log(oddsofbeingDeveloped) = -18.793671 - 0.007982 * AdultMortality + 0.353858 * Alcohol + 0.022927 * HepatitisB - 0.347856 * underfivedeaths + 18.388387 * Incomecompositionofresources$

The coefficients for each predictor variable indicate the direction and strength of the relationship between the predictor and the probability of a country being categorized as "Developed". For example, as alcohol increases by 1 unit, the log odds of being classified as "Developed" increase by 0.353858 all else being equal. Also, as adult mortality increases by 1 unit, the log odds of being classified as "Developed" decreases by 0.007982 all else being equal.

The p-values for each coefficient indicate the statistical significance of the relationship between the predictor and the outcome. A p-value less than 0.05 is generally considered statistically significant, which means that we can reject the null hypothesis that the predictor has no effect on the outcome.

The deviance residuals measure the goodness of fit of the model. A smaller residual deviance indicates a better fit of the model to the data. In this case, the residual deviance is 366.32, which is smaller than the null deviance of 1204.61, indicating that the model is a good fit for the data.

Overall, the model suggests that adult mortality, alcohol, hepatitis B, under-five deaths, and income composition of resources are all significant predictors of a country being categorized as "Developed".

Conclusion

First Scientific Research Hypothesis

As per the regression results from Stepwise Generalized Linear Model (family = " gaussian") , we reject the null hypothesis in the favor of alternate hypothesis. Hence, we conclude that *One or More predictors influencing life expectancy in "Developed" or "Developing" countries have different effect.* The important variables identified in the GLM interaction model are adult mortality rate, alcohol, percentage expenditure, Bmi, polio, diphtheria, HIV AIDS, income composition of resources, schooling, percentage expenditure:developed, BMI:developed, malnutrition(1-19)years:developed, income composition of resources:developed, and schooling:developed. As discussed above and proven through statistical test metrics, the most appropriate model discovered for first scientific research hypothesis is GLM interaction model(family=" gaussian").A certain statistical assumption around life expectancy data being normally distributed may be violated as we know the data is slightly right skewed but as we have observed in diagnostic plots, it didn't effect the results in a significant way. The current model has a adjusted R^2 value of 0.83 i.e. it explains only 83% variation in data. This is believed to be an good adjusted R^2 and doesn't seem like an overfit model.

Second Scientific Research Hypothesis

In conclusion, we reject the null hypothesis that none of the independent variable is significant in classifying a country into "Developed" or "Developing". From the Logistic regression analysis with BIC method for variable selection is the most approriate model which state that adult mortality, alcohol, hepatitis B, under-five deaths, and income composition of resources are significant predictors for classifying a country into “Developed” or “Developing”.

From the diagnostic plot we observe that the model is overfitting the data and not generalizing well to new data.Hence inspite of 94% classification in training set, we might see a decline when exposed to new data corresponding to year after 2015.

Technical Difficulties and Future Work

For first research hypothesis we also tried fitting GAM(General Additive Model) with family ” gaussian”, we faced performance issue while running the model and technical difficulties during variable selection of the model.The result from the model seemed unrealistic as all the diagnostic plots were showing overfitting of the model.

Future Work

We can include more in-depth analysis of GAM in future scope which can help us to resolve the technical difficulties. Therefore, we suggest not concluding GAM as best fit for our dataset.We would also like to run our regression and classification model on life expectancy data post year 2015 to validate its performance in predicting life expectancy and accuracy of results.