DECEMBER 2, 2020

MULTIPLE REGRESSION ANALYSIS ON POPULATION DATA

I. INTRODUCTION TO MULTIPLE REGRESSION ANALYSIS

Multiple Regression is an advanced version of linear regression. In this regression model, we use two or more independent variables(given as X1, X2,…and so on) to get better predictions or results of the dependent variable(Y) from the regression model.

Multiple Regression Equation can be written as:

$$\mathbf{Y=a+b_1X_1+ b_2X_2+ b_3X_3 +…..+b_kX_k}$$

a is the Intercept (Constant Term),

b_1,……,b_k are slope coefficients for each independent variable.

II. STATEMENT OF OBJECTIVES AND DESCRIPTION OF VARIABLES

This project aims to analyze the factors affecting the annual rate of population change using a multiple regression model. The average annual rate of population change of 140 different countries is predicted depending on the various independent variables like Total Fertility Rate (live births per woman), Life Expectancy at birth(years), Maternal Mortality Ratio, GDP per Capita, Employment to Population Ratio and Migration Rate of the given countries.

The data is sourced from the UNdata website (http://data.un.org/) which consists of population-related data for the years in range 2010-2015. The data is merged from 7 different datasets taken from the above-mentioned website to analyze multiple factors affecting population change rate. Before starting the analysis all the datasets are merged into one complete dataset and the missing values are removed from all the columns for better predictions.

Below is the description of all the variables that are used for analyzing the multiple regression:

Name	Data Type	Unit	Class
Annual Rate Population Change	Numeric	Percentage	Continuous Dependent
Total Fertility Rate	Numeric	Live Births per Woman in %	Continuous Independent
Life Expectancy at Birth	Numeric	Years	Continuous Independent
Maternal Mortality Ratio	Numeric	Deaths per 100,000 live births	Continuous Independent
GDP per capita	Numeric	$	Continuous Independent
Employment To Population Ratio	Numeric		Continuous Independent
Migration Rate	Numeric	per 1,000 population	Continuous Independent

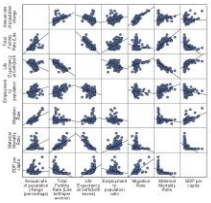
III. ASSUMPTIONS

When you want to use multiple regression to analyze your data, part of the process includes testing to ensure that the details you want to analyze can be analyzed using multiple regression. You need to do this because if your data passes all the assumptions that are needed for multiple regression then only it is reasonable to use multiple regression.

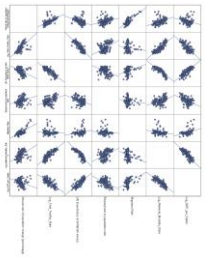
A. Assumption 1:Linearity between Dependent and each of the Independent variables

The very first assumption for multiple regression analysis is that there must be a linear relationship between the dependent variable and each independent variable used in our model. Using the scatter plot matrix in SPSS, we can visualize the linear relationship between the variables at a time.

The linear relationship between the annual rate of population change(dependent variable) and all the independent variables is explained in the scatter plot matrix given below. Total fertility Rate, Life Expectancy at birth, and Migration Rate have a good linear relationship with the Population growth rate looking at the below matrix. Maternal Mortality Ratio and GDP per Capita have weak linearity with the Population growth rate. To improve the linearity the variables are transformed using arithmetic operations in SPSS. This is a multiple trial approach so that we can choose the best transformation that shows linearity between the variables.



After all the trials, Total Fertility Rate, Maternal Mortality Ratio, and GDP per Capita are transformed into their logarithmic values which gave a better linear relationship than the original values. The updated values showed linearity with the dependent variable (annual rate of population change) as seen in the below scatter plot matrix.



B. Assumption 2: Multicollinearity

The variables must not have multicollinearity, which occurs when two or more independent variables are strongly related to each other. This leads to problems in estimating the multiple regression model and cannot easily determine which independent variable is contributing to the variance seen in the dependent variable.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	-3.982	.735		-5.419	.000		
	Log_Total_Fertility_Rate	5.454	.254	.760	21.446	.000	.197	5.084
	Life Expectancy at birth (both sexes)	.037	.006	.224	5.780	.000	.164	6.103
	Log_Maternal_Mortality_Ratio	.333	.083	.163	4.009	.000	.149	6.697
	Migration Rate	.107	.003	.723	38.739	.000	.710	1.409
	Log_GDP_per_Capita	-.060	.098	-.022	-.613	.541	.195	5.120
	Employment to population ratio	.001	.001	.016	.863	.390	.763	1.310

a. Dependent Variable: Annual rate of population change (percentage)

Multicollinearity can be examined in either of the ways 1) To check the correlation matrix for all the variables and if the correlation coefficient value is not in between -0.70 to 0.70 then the predictors are said to be multicollinear. 2) An easier way is to check the VIF values to be less than 10.

In the above table, VIF statistics for all the independent variables is less than 10 hence there is no multicollinearity between the independent variables given to the model.

C. Assumption 3: Independence of Residuals

This assumption is that the residual values must be independent. This means that there should be no relationship between the observations. This assumption can be tested using the Durbin-Watson statistic.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.983[a]	.967	.966	.26668	2.142

a. Predictors: (Constant), Employment to population ratio, Migration Rate, Log_Total_Fertility_Rate, Log_GDP_per_Capita, Life Expectancy at birth (both sexes), Log_Maternal_Mortality_Ratio

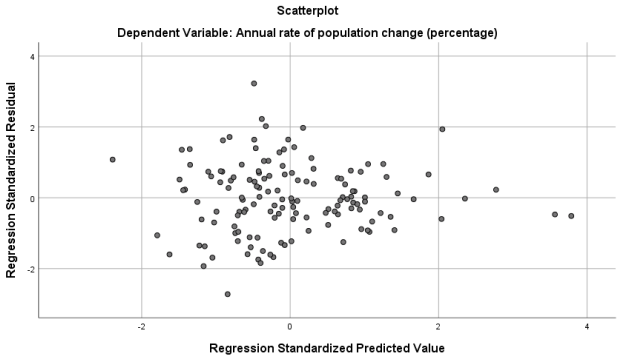
b. Dependent Variable: Annual rate of population change (percentage)

Durbin-Watson statistic is for testing the auto-correlation in between the residual values that ranges from the value 1 to 3 and is considered good when close to 2. The Durbin-Watson statistics for our model has a value of 2.142 which explains that there is very less negative auto-correlation between the residuals.

D. Assumption 4: Homoscedasticity

Homoscedasticity refers to whether these residuals are distributed evenly, or whether they appear to bunch together at some values and scatter far apart at other values.

The below scatterplot appears like noise and does not show any pattern, hence we have homoscedasticity in our defined model.



E. Assumption 5: No significant outliers

Any significant outliers, highly influential points, or high leverage points must not be present in the model. These data points can affect our model and can bias the predicted results. It will make your data less representative overall and might cause differences in the overall predictions. The outliers can be tested in SPSS using Cook's Distance.

Residuals Statistics[a]

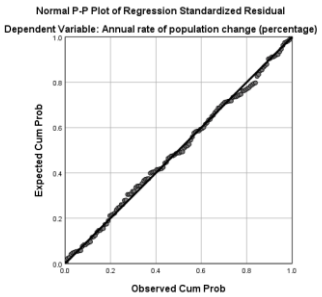
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-1.8363	6.9103	1.5498	1.41550	140
Std. Predicted Value	-2.392	3.787	.000	1.000	140
Standard Error of Predicted Value	.029	.130	.057	.018	140
Adjusted Predicted Value	-1.8581	6.9432	1.5502	1.41624	140
Residual	-.72625	.86076	.00000	.26086	140
Std. Residual	-2.723	3.228	.000	.978	140
Stud. Residual	-2.750	3.314	-.001	1.002	140
Deleted Residual	-.74061	.90716	-.00039	.27381	140
Stud. Deleted Residual	-2.821	3.446	.000	1.011	140
Mahal. Distance	.624	31.885	5.957	5.084	140
Cook's Distance	.000	.085	.007	.013	140
Centered Leverage Value	.004	.229	.043	.037	140

a. Dependent Variable: Annual rate of population change (percentage)

Cook's Distance value must not exceed 1. The points for which the value is exceeding 1 are considered to be outliers. In this data model, the minimum value of Cook's Distance is .000 and the maximum is .085. Hence, our data model does not consist of any outliers or influential data points that need to be removed from the dataset.

F. Assumption 6: Normal P-P Plot of Regression

In this assumption, to see if the residuals/errors are approximately distributed normally represented by a normal P-P plot. The closer the dots plotted on the diagonal line the better the residuals are normally distributed.



In the plot shown above, the residuals are very close to the diagonal line. Hence, it states that the model shows the normal distribution of the residuals plotted against the predicted values.

IV. UNDERSTANDING AND BUILDING MODEL

A. *Descriptive Statistics*

Descriptive statistics enable the data to be summarized and structured so that it can be easily interpreted. Descriptive statistics are useful in explaining simple data attributes, such as summary statistics for variables of size and data calculation. In a large data analysis report, these statistics will assist us to manage the data and present it in a summary table. Descriptive Statistics for the used dataset is shown below.

Descriptive Statistics

	Mean	Std. Deviation	N
Annual rate of population change (percentage)	1.5498	1.43933	140
Log_Total_Fertility_Rate	.4109	.20054	140
Life Expectancy at birth (both sexes)	71.0286	8.62383	140
Log_Maternal_Mortality_ Ratio	1.7842	.70465	140
Migration Rate	1.3458	9.72863	140
Log_GDP_per_Capita	4.0189	.52107	140
Employment to population ratio	118.6650	19.76745	140

B. *Correlation Matrix*

The correlation matrix states the relationships between each independent and dependent variable as well as the relationship between two independent variables. A specific standard is used to look for correlations in between the variables with an absolute value of 0.700. So, the independent variables having relationships with each other higher than the standards are likely to cause a problem(Multicollinearity) in determining the predictions.

In the below correlation table, Total Fertility Rate and Migration Rate are strongly related to the annual rate of population change. Life Expectancy at birth and Maternal Mortality Ratio is moderately correlated with the annual rate change of population. The last two variables Log_GDP_per Capita and Employment to population ratio have a weak correlation with the dependent variable.

Correlations

		Annual rate of population change (percentage)	Log_Total_Fertility_Rate	Life Expectancy at birth(both sexes)	Log_Maternal_Mortality_Ratio	Migration Rate	Log_GDP_per_Capita	Employment to population ratio
Pearson Correlation	Annual rate of population change (percentage)	1.000	.658	-.403	.452	.655	-.295	.265
	Log_Total_Fertility_Rate	.658	1.000	-.853	.848	-.100	-.787	.441
	Life Expectancy at birth (both sexes)	-.403	-.853	1.000	-.889	.263	.823	-.385
	Log_Maternal_Mortality_ Ratio	.452	.848	-.889	1.000	-.251	-.855	.416
	Migration Rate	.655	-.100	.263	-.251	1.000	.397	-.107
	Log_GDP_per_Capita	-.295	-.787	.823	-.855	.397	1.000	-.462
	Employment to population ratio	.265	.441	-.385	.416	-.107	-.462	1.000
Sig. (1-tailed)	Annual rate of population change (percentage)	.	.000	.000	.000	.000	.000	.001
	Log_Total_Fertility_Rate	.000	.	.000	.000	.120	.000	.000
	Life Expectancy at birth (both sexes)	.000	.000	.	.000	.001	.000	.000
	Log_Maternal_Mortality_ Ratio	.000	.000	.000	.	.001	.000	.000
	Migration Rate	.000	.120	.001	.001	.	.000	.103
	Log_GDP_per_Capita	.000	.000	.000	.000	.000	.	.000
	Employment to population ratio	.001	.000	.000	.000	.103	.000	.
N	Annual rate of population change (percentage)	140	140	140	140	140	140	140

C. *Model Building*

a) *Model 1:*

Variables entered to build the first Model are :

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Employment to population ratio, Migration Rate, Log_Total_Fertility_Rate, Log_GDP_per_Capita, Life Expectancy at birth(both sexes), Log_Maternal_Mortality_Ratio[b]	.	Enter

a. Dependent Variable: Annual rate of population change (percentage)

b. All requested variables entered.

After all the assumptions are met, the model is built using all the independent transformed variables(Log_Total_fertility Rate, Life Expectancy at birth (both sexes), Log Maternal Mortality Ratio, Migration Rate of that specific country, Log GDP per capita and Employment to Population Ratio)that have a linear relationship with the dependent variable. The below table shows that the resulting model has 2 non-significant independent variables Log GDP per capita and the Employment to population ratio which does not fulfill the condition of $p<0.05$ having p values of 0.541 and 0.390 respectively.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.983[a]	.967	.966	.26668	2.142

a. Predictors: (Constant), Employment to population ratio, Migration Rate, Log_Total_Fertility_Rate, Log_GDP_per_Capita, Life Expectancy at birth (both sexes), Log_Maternal_Mortality_Ratio

b. Dependent Variable: Annual rate of population change (percentage)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	278.504	6	46.417	652.702	.000[b]
	Residual	9.458	133	.071		
	Total	287.963	139			

a. Dependent Variable: Annual rate of population change (percentage)

b. Predictors: (Constant), Employment to population ratio, Migration Rate, Log_Total_Fertility_Rate, Log_GDP_per_Capita, Life Expectancy at birth(both sexes), Log_Maternal_Mortality_Ratio

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	-3.982	.735		-5.419	.000		
	Log_Total_Fertility_Rate	5.454	.254	.760	21.446	.000	.197	5.084
	Life Expectancy at birth (both sexes)	.037	.006	.224	5.780	.000	.164	6.103
	Log_Maternal_Mortality_ Ratio	.333	.083	.163	4.009	.000	.149	6.697
	Migration Rate	.107	.003	.723	38.739	.000	.710	1.409
	Log_GDP_per_Capita	-.060	.098	-.022	-.613	.541	.195	5.120
	Employment to population ratio	.001	.001	.016	.863	.390	.763	1.310

a. Dependent Variable: Annual rate of population change (percentage)

The output of this model is represented by the regression equation:

Y= (-3.982) + 5.454 (Log_Total_fertility_rate) + 0.037(Life Expectancy at birth) +0.333(Log_Maternal_Mortality_Ratio) + 0.107(Migration Rate) - 0.060(Log_GDP_per_Capita) + 0.001(Employment to population ratio)

Then, the next global hypothesis test is conducted, to check if any of the regression coefficients are other than 0. Using the 0.05 significance level.

$H_0: b_1=b_2=b_3=b_4=b_5=b_6=0$

H_1: Not all the b's are 0.

The ANOVA table interprets the p-value to be 0.000 which is less than the significance level i.e. 0.05, so we reject the null hypothesis and hence determine that at least one of the regression coefficients is not equal to 0.

Next, performing the individual hypothesis where individual regression coefficients are being examined. The coefficients of each independent variable are checked.

$H_0: b_k=0$

$H_1: b_k \neq 0$

The p-values for Log GDP per capita and employment ratio are greater than the significance level(0.05), so the null hypothesis is not rejected for these variables.

To obtain better results from the model, the two non-significant predictors are to be removed from the model one by one (value with the greatest p-value or with smallest t-statistic), and hence we will run Model 2 and 3 with the updated variables.

b) Model 2:

In this model, the independent variable Log GDP per capita is removed as it is the one with the largest p-value of 0.541 in the previous model built and run the analysis again.

Below are the findings for Model 2:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.983[a]	.967	.966	.26605	2.148

a. Predictors: (Constant), Employment to population ratio, Migration Rate, Log_Total_Fertility_Rate, Log_Maternal_Mortality_Ratio, Life Expectancy at birth(both sexes)

b. Dependent Variable: Annual rate of population change (percentage)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	278.478	5	55.696	786.837	.000[b]
	Residual	9.485	134	.071		
	Total	287.963	139			

a. Dependent Variable: Annual rate of population change (percentage)

b. Predictors: (Constant), Employment to population ratio, Migration Rate, Log_Total_Fertility_Rate, Log_Maternal_Mortality_Ratio, Life Expectancy at birth(both sexes)

The global, as well as individual hypothesis on this model, was checked again. It is visible that the R^2 and Adjusted R^2 remain unchanged. Also, the p-value of the Employment to Population ratio is still greater than 0.05 and hence needs to be removed which will be done in the next model.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	-4.254	.584		-7.281	.000		
	Log_Total_Fertility_Rate	5.489	.247	.765	22.196	.000	.207	4.830
	Life Expectancy at birth (both sexes)	.037	.006	.221	5.763	.000	.167	5.996
	Log_Maternal_Mortality_Ratio	.353	.076	.173	4.614	.000	.176	5.698
	Migration Rate	.106	.003	.718	42.285	.000	.853	1.173
	Employment to population ratio	.001	.001	.018	1.008	.315	.795	1.258

a. Dependent Variable: Annual rate of population change (percentage)

c) Model 3:

To satisfy the condition of $p<0.05$, the non-significant variable in the previous model needs to be removed. So, the Employment to population ratio having a p-value of 0.315 is removed and the regression analysis is run again.

After removing this variable, the R^2 and adjusted R^2 values remain unchanged, indicating the removed variables to be completely insignificant.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.983[a]	.967	.966	.26607	2.166

a. Predictors: (Constant), Migration Rate, Log_Total_Fertility_Rate, Log_Maternal_Mortality_Ratio, Life Expectancy at birth(both sexes)

b. Dependent Variable: Annual rate of population change (percentage)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	278.406	4	69.601	983.175	.000[b]
	Residual	9.557	135	.071		
	Total	287.963	139			

a. Dependent Variable: Annual rate of population change (percentage)

b. Predictors: (Constant), Migration Rate, Log_Total_Fertility_Rate, Log_Maternal_Mortality_Ratio, Life Expectancy at birth(both sexes)

The global hypothesis test is conducted, to check if any of the regression coefficients are other than 0. Using the 0.05 significance level.

$H_0: b_1=b_2=b_3=b_4=b_5=b_6=0$

H_1: Not all the b's are 0.

According to the ANOVA table, $p<0.05$. Hence rejecting the null Hypothesis and accepting an alternate Hypothesis for this model and conclude that not all regression coefficients are 0.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	-4.157	.576		-7.213	.000		
	Log_Total_Fertility_Rate	5.537	.243	.771	22.817	.000	.215	4.650
	Life Expectancy at birth (both sexes)	.037	.006	.223	5.825	.000	.167	5.978
	Log_Maternal_Mortality_Ratio	.359	.076	.176	4.720	.000	.177	5.654
	Migration Rate	.106	.003	.717	42.292	.000	.855	1.169

a. Dependent Variable: Annual rate of population change (percentage)

To study the regression coefficients individually, performing the individual hypothesis test.

$H_0: b_k=0$

$H_1: b_k \neq 0$

According to the coefficients table, all the preditors are significant and possess a p-value less than the level of significance(0.05) making this model perfect for regression analysis.

Final Regression equation:

Y= (-4.157) + 5.537 (Log_Total_fertility_rate) + 0.037(Life Expectancy at Birth)+0.359(Log_Maternal_Mortality_Ratio) + 0.106(Migration Rate)

This equation explains that the variable Log_Total_fertility_rate has a strong relationship with the dependent variable(Annual rate of population change) and the other predictors are having moderate and weak relationships.

V. MODEL SUMMARY

As per our analysis, **Model 3** is considered to be the final model to predict the rate of population change.

Below is the summarized analysis of our final model concerning the tables mentioned above in the Model 3 section:

Model Summary:

The Model Summary table is the first table of importance. The R, R^2, adjusted R^2, and the standard error of the estimate are given in this table, which can be used to decide how well a regression model fits the data: R^2 and adjusted R^2 are the coefficients of determination in the model summary. The R^2 value is the proportion of variance in the dependent variable which can be determined by the independent variables. The R^2 value of 96.7% determines that the independent variable has 96.7% of variability with the dependent variable. The adjusted R^2 interprets the strength of the relation between all the independent variables and the annual rate of population change which is about 96.6%. Hence, it is clear that all these independent variables are useful in predicting the value of the annual rate of population change.

ANOVA:

In addition to the explanation provided in Model 3, the value of F statistics is 983.175 which would be compared with the critical value to test the significance of the model. The degree of freedom(df) in the numerator is k i.e.number of independent variables which is 4. The degree of freedom in the denominator is N-(k+1) which computes to 135.
The critical value is between 2.42 and 2.46, which is much smaller than the computed F value by the model.
Hence, rejecting the null Hypothesis of having regression coefficients as 0 as the F value is greater than the critical value.

Coefficients :

Unstandardized B is the coefficient by which each predictor computes the dependent variable if all the remaining predictors are constant in the regression equation.
β_0 =-4.157(constant)
β_1=5.537, an increase in fertility rate by 1% will increase the population change rate by 5.537 %
β_2=0.037, an increase in Life Expectancy at birth by 1 year will increase the population change rate by 0.037%
β_3=0.359, an increase in Maternal Mortality ratio by 1% will increase the population change rate by 0.359%
β_4=0.106, an increase in Migration Rate by 1% will increase the population change rate by 0.106%

Collinearity Diagnostics:

In the Collinearity Diagnostics table, the eigenvalues are used to test the variance among the independent variables. Below is the table of the final Model.

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	Log_Total_Fertility_Rate	Life Expectancy at birth(both sexes)	Log_Maternal_Mortality_Ratio	Migration Rate
1	1	3.780	1.000	.00	.00	.00	.00	.00
	2	.996	1.948	.00	.00	.00	.00	.82
	3	.200	4.346	.00	.07	.00	.02	.05
	4	.022	13.118	.00	.74	.00	.63	.09
	5	.001	64.111	1.00	.18	.99	.35	.03

a. Dependent Variable: Annual rate of population change (percentage)

Casewise Diagnostics:

Casewise Diagnostics provides the actual and predicted values for the dependent variable.

Casewise Diagnostics[a]

Case Number	Std. Residual	Annual rate of population change (percentage)	Predicted Value	Residual
1	-.945	3.30	3.5474	-.25139
2	.215	-.39	-.4511	.05708
3	.308	1.98	1.9010	.08199
4	.559	3.54	3.3953	.14870
5	-1.323	1.04	1.3909	-.35190
6	.452	.33	.2127	.12027
7	-.406	1.54	1.6521	-.10805
8	-.017	.63	.6336	-.00457
9	1.414	1.27	.8899	.37613
10	1.350	2.01	1.6479	.35914

VI. CONCLUSION

After performing multiple regression analysis on the population data for various countries, the fertility rate of a country contributes majorly to the annual change in the rate of population. The Maternal Mortality Ratio in the country, its Migration Rate, and the Life Expectancy of both sexes at the time of birth are also affecting the rate of annual population change in that country to some extent. The final model built is meeting all the assumptions and hence is perfect for predicting annual change in population rate.
According to the dataset used in this project, it concludes that when the Total Fertility Rate, Maternal Mortality Ratio, Migration Rate, and Life Expectancy of the citizens at the time of the birth of a specific country is known then Population Growth Rate can be predicted with 96.7% accuracy.

VII. REFERENCES

[1] W. Aiyin and X. Yanmei, "Multiple Linear Regression Analysis of Real Estate Price," 2018 International Conference on Robots & Intelligent System (ICRIS), Changsha, 2018, pp. 564-568, doi: 10.1109/ICRIS.2018.00145.