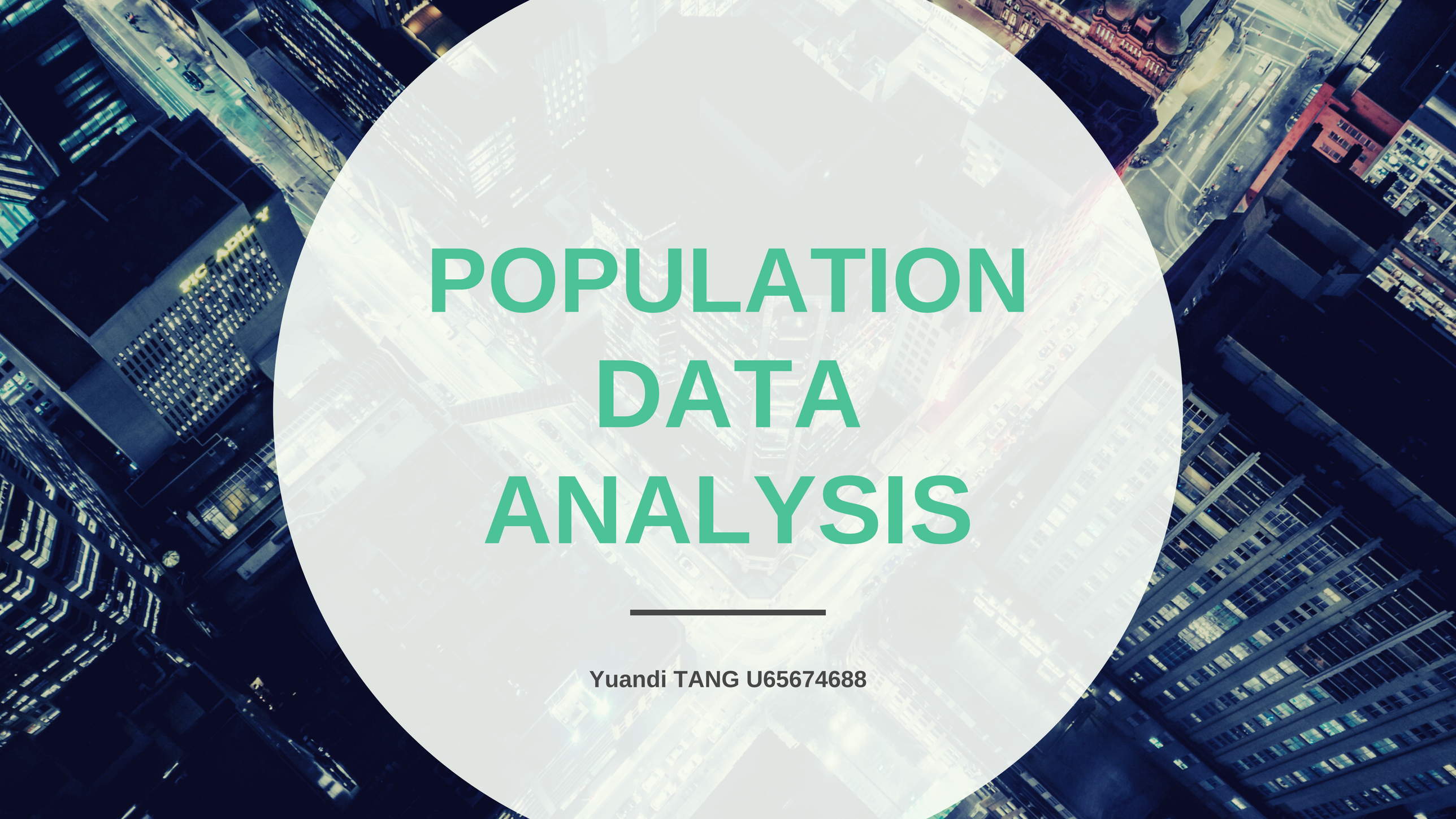


POPULATION DATA ANALYSIS

Yuandi TANG U65674688

QUESTION

HOW MANY PEOPLE ARE THERE IN YOUR COUNTRY

TOPICS

1. RESEARCH SCENARIO

2. RESEARCH QUESTION

3. RESEARCH METHODS

4. CONCLUSION

RESEARCH SCENARIO

The relationship between land area, density, fertility rate, and the population is always very complicated yet interesting. In this research, I take the recent demographic data and use multiple data analysis methods to dive deep into those relationships

RESEARCH QUESTION

1.What is the relationship between population and land area?

2.What is the relationship between population and other factors in the dataset? What is the best model.

3.What is the best model between other variables and the level of median age(>30,1,0)?

DATASET INTRODUCTION

The column information are listed below:

- *Country/Other - Name of countries and dependent territories.*
- *Population (2020) - Population in the year 2020*
- *Yearly Change - Percentage Yearly Change in Population*
- *Net Change - Net Change in Population*
- *Density (P/Km²)- Population density (population per square km)*
- *Land Area (Km²) - Land area of countries / dependent territories.*
- *Migrants (net) - Total number of migrants*
- *Fert. Rate - Fertility rate*

RESEARCH METHODS

- Correlation
- Sample Means
- Simple Linear Regression
- Multiple Linear Regression
- Model Accessment
- Prediction
- Cross validation
- Sample Logistic Regression
- Multiple Logistic Regression
- One sample Proportion Test

CLEANING

make numeric

```
library(car)
library(pROC)
#import data
data <-
read.csv("D:/personal/study/555/555TERMPROJECT/world_population.csv")
#check and cleaning
data <- na.omit(data)
is.integer(data$Population..2020.)
is.integer(data$Land.Area..Km虏.)
is.integer(data$Urban.Pop..)
data$Urban.Pop..<-data$Urban.Pop..[!is.na(data$Urban.Pop..)]
as.integer(data$Urban.Pop..)
is.integer(data$Med..Age)
attach(data)
```

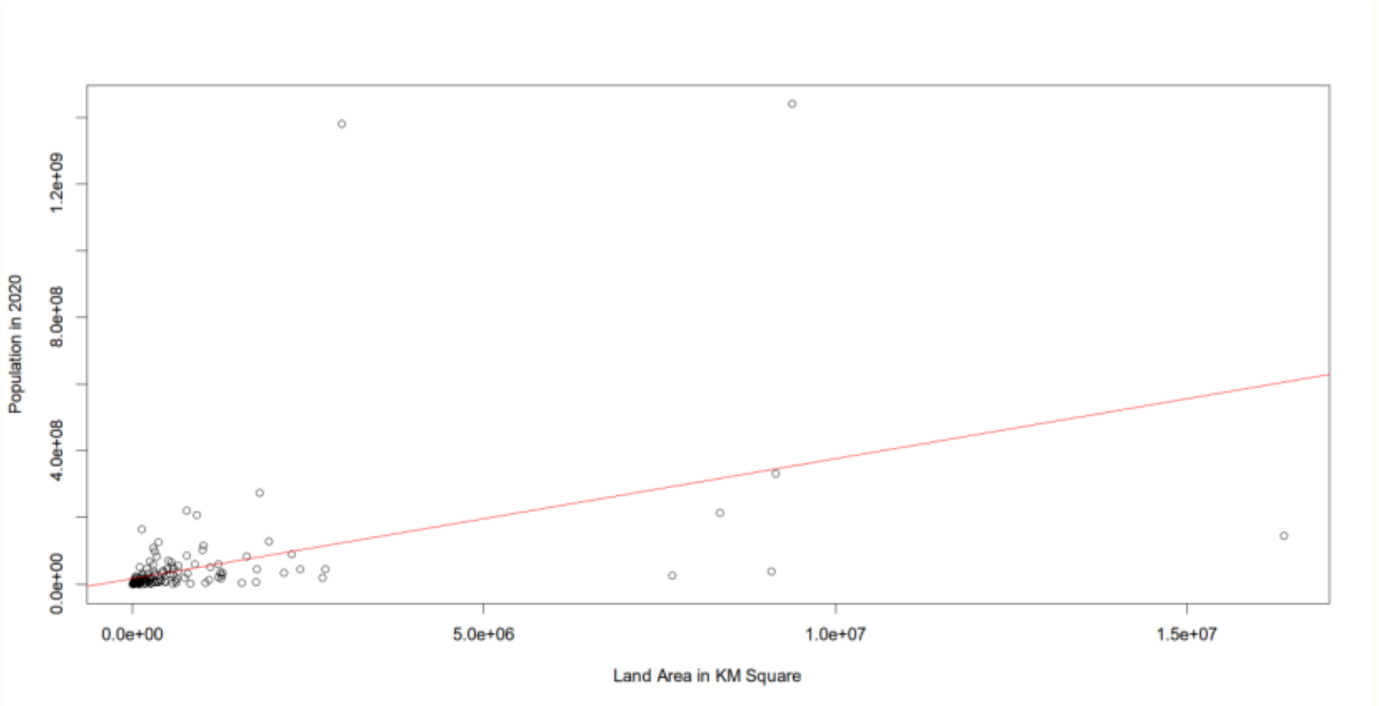
Outlier：

Decide not to remove, coz every country is unique and if delete big countries(CHN IND USA CAN), the model will be so biased from the reality

METHODS

- Correlation
- ANOVA
- ANCOVA
- Simple and Multiple Linear Regression
- Prediction
- Cross-Validation
- Logistics Regression
- Proportion Test

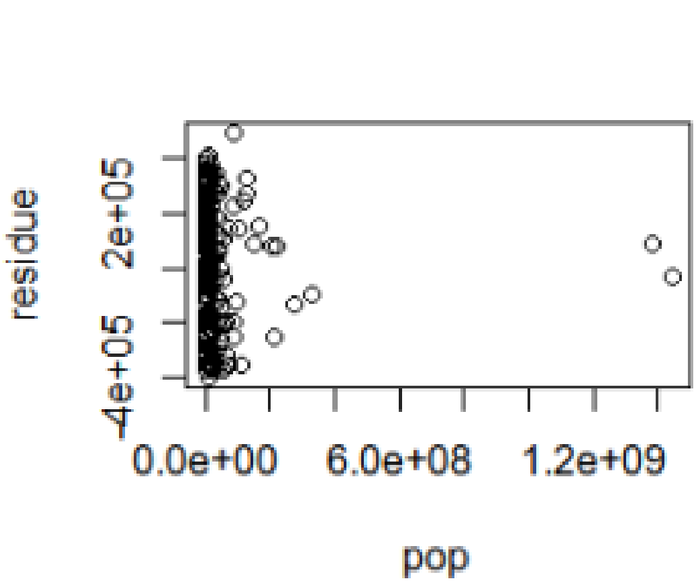
CONCLUSION-SLR

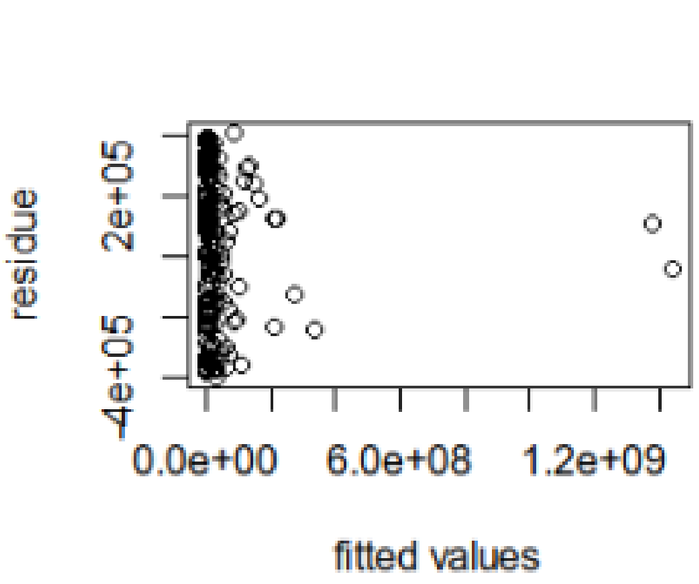


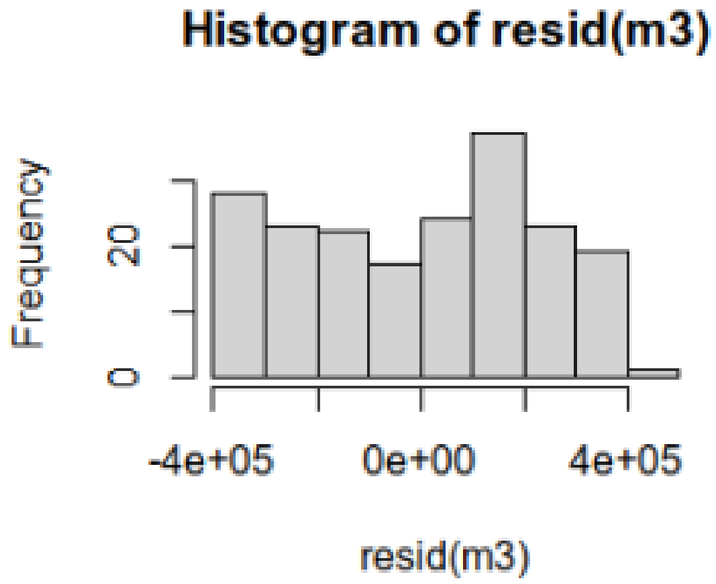
SIMPLE LINEAR REGRESSION

POPULATION~LAND AREA

CONCLUSION-MLR







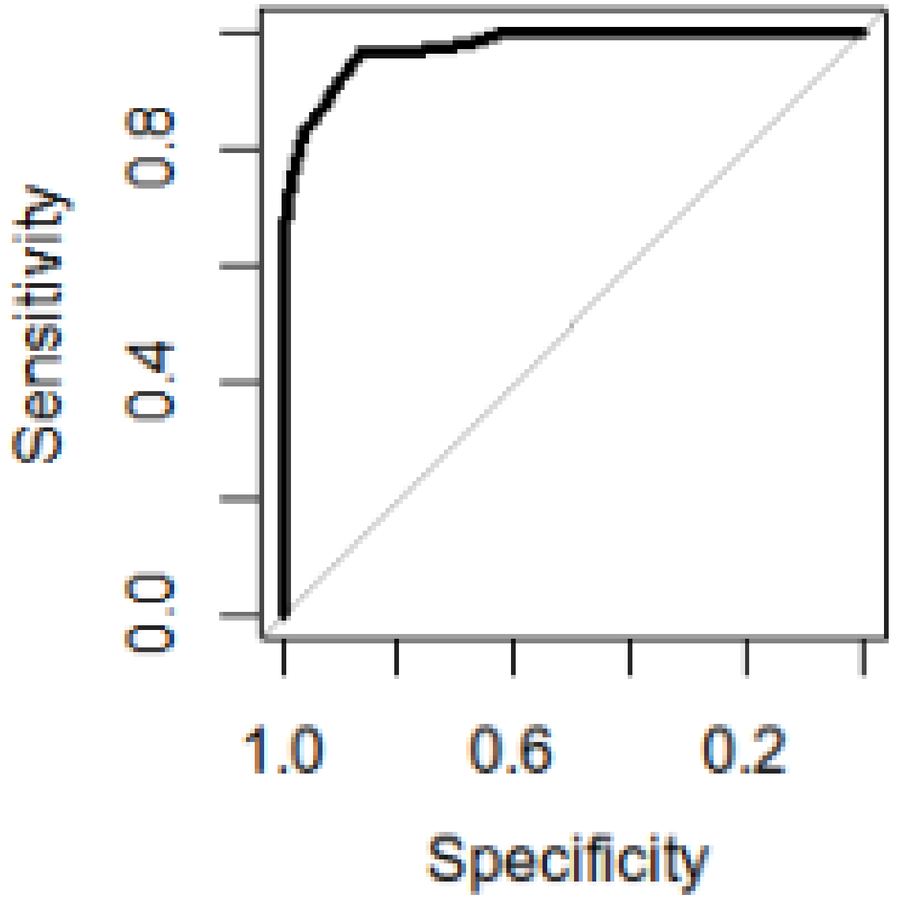
MULTIPLE LINEAR REGRESSION

POPULATION~

WORLD.SHARE+NET.CHANGE+YEARLY.CHANGE+*FERTILITY.RATE.

CONCLUSION-LOG

97.44%

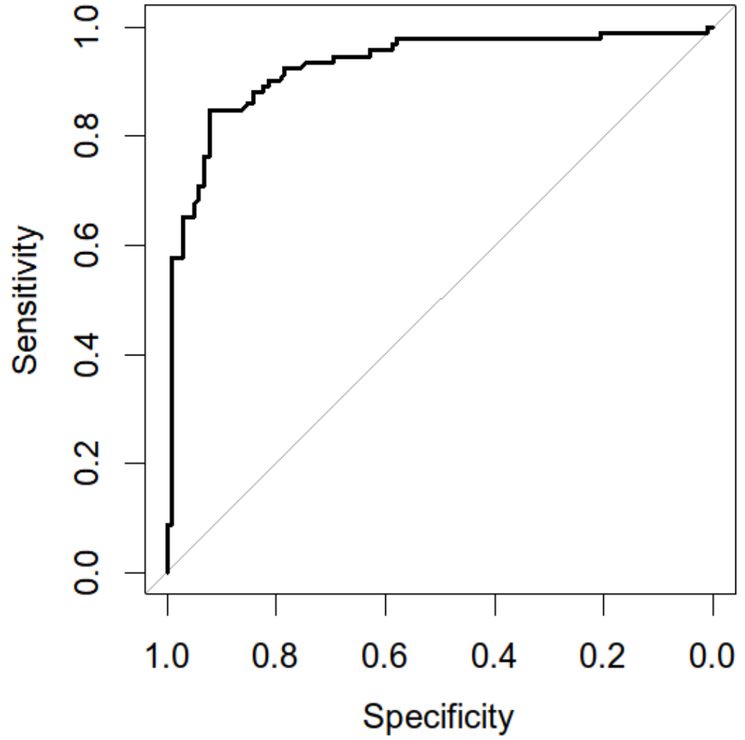


Simple Logistic Regression

Median Age~Urbanpop

CONCLUSION-LOG

92.74%



Multiple Logistic Regression

Median Age~YearlyPopChange+FertilityRate

OUTPUT

- **SLIR (ARS=0.19)** ***-->NOT GOOD,NOT SIGNIFICANT LINEAR***

Population=

3.598e+01***Landarea**+1.604e+07.

- ***MLIR(ARS=0.9 AIC=4794.4)*** ***-->GOOD BUT NOT MEANINGFUL***

Population=

7.791e+07***World.Share**+7.490e-02***Net.Change**-

7.681e+04***Yearly.Change**+5.136e+04***Fertility. Rate**

- **SLOR(AUC=97.44%) -->VERY GOOD, SIGNIFICANT LOGISTIC**

Probability of Level of **median age** = 1/1+e5.7499***urbanpop**-12.8911

- **MLOR(AUC=92.74%) -->VERY GOOD, SIGNIFICANT LOGISTIC**

Probability of Level of **median age**=

1/1+e^(-12.3374-0.52293***fertilityrate**+**yearlychang***0.579)

CONCLUSION

- **THE COUNTRY'S POPULATION IS RELATED BUT NOT HIGHLY RELATED TO THE LAND AREA.**
- **THE COUNTRY'S CURRENT POPULATION IS BETTER RELATED TO YEARLY CHANGE AND FERTILITY RATE, WHICH MAKES SENSE STATISTICALLY AND THEORETICALLY.**
- **A BIGGER URBAN POPULATION AND SMALLER FERTILITY RATE ARE SIGNIFICANT INDICATORS OF MEDIAN AGE(THE AGING EFFECTS) ON POPULATION. THE AGING THEORY PROVED STATISTICALLY.**

LIMITATION

- STATIC DATA, WITHOUT CONSIDERING THE INFLUENCE OF PREVIOUS AND POPULATION CHANGE
- HIGHLY SKEWED BECAUSE IN A REAL SITUATION, CAUSE A DILEMMA

REMOVE OUTLIERS(CHINA/INDIA)

--> WORSE IN INTERPRETATION

KEEP OUTLIERS(CHINA/INDIA)

--> WORSE IN MODEL QUALITY

- LINEAR/LOGISTIC REGRESSION IN THE CLASS WE LEARNED MAY NOT BE A GOOD METHOD TO EXPLAIN THE DATA. FURTHER AND MORE COMPLEX MODELS SHOULD BE APPLIED.

THANKS