Statistics for Data Analytics – CA2

Aditya Arora

X18132120

Contents

[Introduction: 2](#_Toc5515652)

[Objective: 2](#_Toc5515653)

[Data Pre-Processing and Preliminary Analysis: 2](#_Toc5515654)

[Data Sources: 2](#_Toc5515655)

[Preparing Data for Software: 3](#_Toc5515656)

[Assumptions: 3](#_Toc5515657)

[Exploratory Data Analysis: 3](#_Toc5515658)

[Linear Model Creation and Interpretation: 6](#_Toc5515659)

[Initial Model Creation: 6](#_Toc5515660)

[Variable Selection 7](#_Toc5515661)

[Testing Assumptions 9](#_Toc5515662)

[Logistic Regression 11](#_Toc5515663)

[Logistic Regression: Preliminary Analysis 12](#_Toc5515664)

[Logistic Regression Model Creation 12](#_Toc5515665)

[Logistic Regression Model Evaluation 13](#_Toc5515666)

[Testing Predictive Power and Goodness of Fit 14](#_Toc5515667)

[Interpretation 15](#_Toc5515668)

[Findings 15](#_Toc5515669)

[References 15](#_Toc5515670)

[Bibliography 16](#_Toc5515671)

# Introduction:

United Nations Development Programme (UNDP) calculates the Gender Inequality Index (GII), which aims to measure disparities between males and females in a region / country, and highlight areas in need of critical policy intervention, according to according to the GII - United Nations Development Report, available on “<http://data.un.org/DocumentData.aspx?q=gender&id=391>**”** (Accessed: 04-April-2019)

There has been a global effort to increase female representation in parliaments. It is hoped that greater representation of women in legislature would help highlight gender inequality issues more effectively, and lead to reduction in gender disparities.

According to *Katica Roy, “How Fewer Women in Politics Impacts Policy and the Economy”, 07-Mar-2018*, Women in elected office focus on health (including women’s health), among other issues.

Similarly, we would expect higher education levels and labour force participation rates of women to positively impact gender equality.

# Objective:

1. Understand and quantify the effect of ratio of women in parliament, percentage of women with at least secondary education and participation of women in labour force (dependent variables) on adolescent birth rates (independent variable)
2. Estimate the power of GII in predicting if a country belongs to the Low HDI group.

# Data Pre-Processing and Preliminary Analysis:

## Data Sources:

The data used is the Gender Inequality Index data from the United Nations Development Programme’s Human Development Report 2018:

“<http://data.un.org/DocumentData.aspx?q=gender&id=391>”

Columns used from dataset:

**Country (country)**: with 191 countries

**Adolescent birth rate (ABR):** Number of births to women ages 15-19 per 1,000 women ages 15-19.

**Population with at least some secondary education (SEC):** Percentage of the population ages 25 and older that has reached (but not necessarily completed) a secondary level of education.

**Share of seats in parliament (PARL):** Proportion of seats held by women in the national parliament expressed as percentage of total seats. For countries with a bicameral legislative system, the share of seats is calculated based on both houses.

**Labour force participation rate (LABF):** Proportion of the working-age population (ages 15 and older) that engages in the labour market, either by working or actively looking for work, expressed as a percentage of the working-age population.

Separate datasets from data.un.org are used for historical values of ratio of women in politics and ratio of girls in education.

<http://data.un.org/_Docs/SYB/CSV/SYB61_T05_Seats%20held%20by%20Women%20in%20Parliament.csv>

http://data.un.org/\_Docs/SYB/CSV/SYB61\_T05\_Seats%20held%20by%20Women%20in%20Parliament.csv

## Preparing Data for Software:

The data files have been pre-processed to prepare the data for analysis on software R.

The UNDP GII data includes the columns HDI rank and GII score, which are calculated based on the independent variables. These present redundant information in the context of our model, and have been removed using MS Excel, along with other additional columns.

In the historical data on % women in politics and ratio of girls in education, the average of these values by country were calculated using the MS Excel Pivot function, and added to the data file.

## Assumptions:

Linear Regression makes no assumptions regarding the distribution of the variables.

The method makes several assumptions regarding the model like:

1. Linear relationship between outcome and predictor variables
2. No multicollinearity between independent variables
3. Homoscedasticity and Independence of Errors
4. There are no influential data points

The model needs to be tested against these assumptions.

## Exploratory Data Analysis:

The GII data is imported to R:

#Read UNDP GII Data

data = readxl::read\_excel("Stats\\GII Cleaned.xlsx", na = "..")

The following new columns are added:

PARL\_AV: Average of share of seats represented by women for all the years of data available

PARL\_2005: Share of seats of women in parliament as of 2005

PRIM: Ratio of girls to boys in Primary Education

1. Data can be visually inspected for any extreme values using histogram

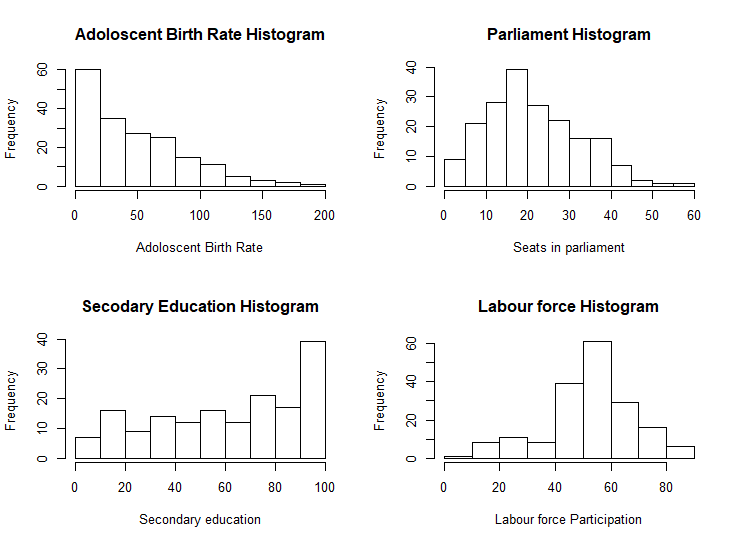
#Histogram of all continuous variables

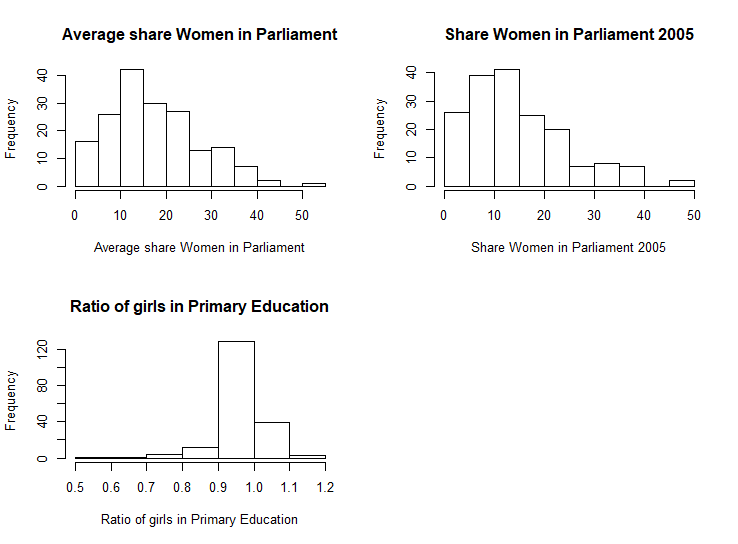
hist(data$`Adolescent birth rate`, xlab = "Adoloscent Birth Rate", main = "Adoloscent Birth Rate Histogram")

hist(data$`Share of seats in parliament`,xlab = "Seats in parliament", main = "Parliament Histogram")

hist(data$`Population with at least some secondary education`,xlab = "Secondary education", main = "Secodary Education Histogram")

hist(data$`Labour force participation rate `,xlab = "Labour force Participation", main = "Labour force Histogram")





1. Correlation matrix created to check multicollinearity:

ABR PARL SEC LABF PARL\_AV PARL\_2005 PRIM

ABR 1.00000000 -0.07108620 -0.67360297 0.23745327 -0.09828329 -0.15605036 -0.39971310

PARL -0.07108620 1.00000000 0.11246755 0.18050415 0.94268643 0.76539662 0.07996179

SEC -0.67360297 0.11246755 1.00000000 -0.07663005 0.14653707 0.20585191 0.36523524

LABF 0.23745327 0.18050415 -0.07663005 1.00000000 0.19685090 0.27343600 0.04950340

PARL\_AV -0.09828329 0.94268643 0.14653707 0.19685090 1.00000000 0.88581343 0.03750461

PARL\_2005 -0.15605036 0.76539662 0.20585191 0.27343600 0.88581343 1.00000000 0.09492719

PRIM -0.39971310 0.07996179 0.36523524 0.04950340 0.03750461 0.09492719 1.00000000

There is a high collinearity between the variables PARL, PARL\_AV, and PARL\_20015, which is expected.

Of the three, the average share of women in the parliament over the years, is most highly correlated with Adolescent Birth Rate.

The PARL and PARL\_2005 variables are removed.

ABR SEC LABF PARL\_AV PRIM

ABR 1.00000000 -0.67360297 0.23745327 -0.09828329 -0.39971310

SEC -0.67360297 1.00000000 -0.07663005 0.14653707 0.36523524

LABF 0.23745327 -0.07663005 1.00000000 0.19685090 0.04950340

PARL\_AV -0.09828329 0.14653707 0.19685090 1.00000000 0.03750461

PRIM -0.39971310 0.36523524 0.04950340 0.03750461 1.00000000

No multicollinearity among independent variables.

1. Relationship between dependent and independent variables can be visually checked for linearity using scatter plots

#Scatterplots to inspect linearity

plot(gii\_data$ABR, gii\_data$PARL, xlab = "Adolescent Birth Rate", ylab = "Current Share of Parliament Seat")

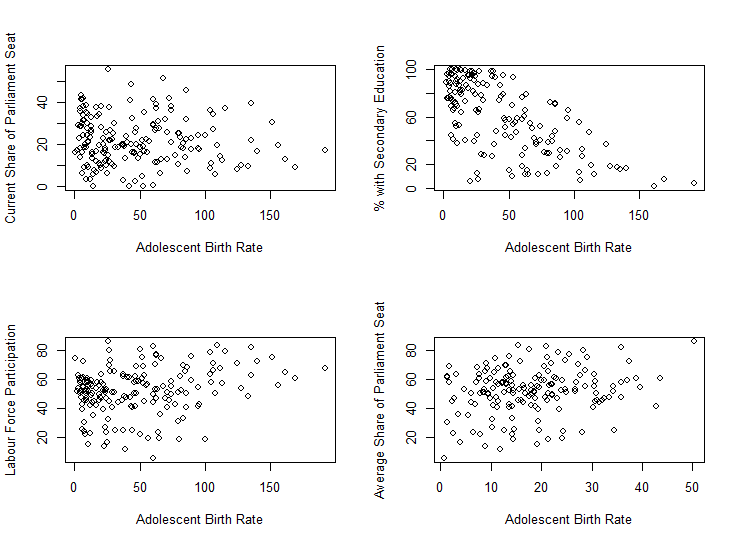
plot(gii\_data$ABR, gii\_data$SEC, xlab = "Adolescent Birth Rate", ylab = "% with Secondary Education")

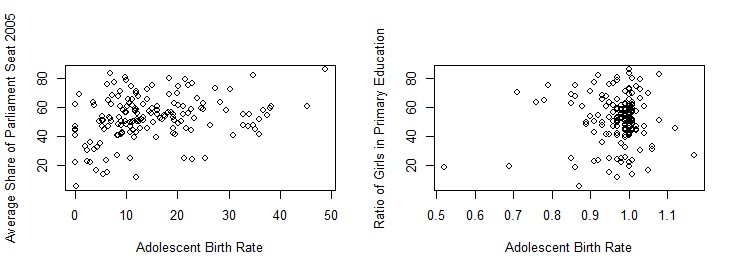
plot(gii\_data$ABR, gii\_data$LABF, xlab = "Adolescent Birth Rate", ylab = "Labour Force Participation")

plot(gii\_data$PARL\_AV, gii\_data$LABF, xlab = "Adolescent Birth Rate", ylab = "Average Share of Parliament Seat")

plot(gii\_data$PARL\_2005, gii\_data$LABF, xlab = "Adolescent Birth Rate", ylab = "Average Share of Parliament Seat 2005")

plot(gii\_data$PRIM, gii\_data$LABF, xlab = "Adolescent Birth Rate", ylab = "Ratio of Girls in Primary Education")





It is observed that the relationship between Adolescent Birth Rate and Ratio of Girls in Primary Education may not be linear.

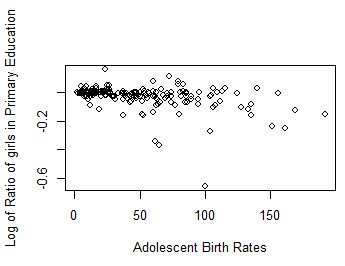
The variable in a ratio not bound by 0 and 1, but cannot be negative.

A log transformation can be applied to attempt to create a linear relationship.

#Log tranformaion on PRIM

gii\_data$PRIM\_LOG = log(gii\_data$PRIM)

plot(gii\_data$ABR, gii\_data$PRIM\_LOG)



This needs to be further examined during model creation.

# Linear Model Creation and Interpretation:

As per above discussion, PARL and PARL\_2005 are not used in the model.

## Initial Model Creation:

The lm() function in R is used to create a multiple variable linear regression model:

#create model with all variables

model\_ABR = lm(ABR ~ SEC + LABF + PARL\_AV + PRIM\_LOG, data = gii\_data)

summary(model\_ABR)

Call:

lm(formula = ABR ~ SEC + LABF + PARL\_AV + PRIM\_LOG, data = gii\_data)

Residuals:

Min 1Q Median 3Q Max

-67.859 -15.814 -3.139 14.681 80.126

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 61.63158 9.89116 6.231 4.76e-09 \*\*\*

SEC -0.73790 0.08199 -9.000 1.16e-15 \*\*\*

LABF 0.57763 0.14709 3.927 0.000133 \*\*\*

PARL\_AV -0.23558 0.23346 -1.009 0.314636

PRIM\_LOG -113.24396 39.59564 -2.860 0.004862 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.96 on 145 degrees of freedom

(41 observations deleted due to missingness)

Multiple R-squared: 0.5171, Adjusted R-squared: 0.5038

F-statistic: 38.82 on 4 and 145 DF, p-value: < 2.2e-16

## Variable Selection

It is observed that PARL\_AV, Average Share of Women in Parliament is not a significant variable in the model.

The variable is removed and the model is re-evaluated:

Call:

lm(formula = ABR ~ SEC + LABF + PRIM\_LOG, data = gii\_data)

Residuals:

Min 1Q Median 3Q Max

-72.040 -16.676 -2.303 16.582 79.689

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 62.29200 9.57732 6.504 1.02e-09 \*\*\*

SEC -0.76629 0.07924 -9.670 < 2e-16 \*\*\*

LABF 0.52522 0.13930 3.770 0.000231 \*\*\*

PRIM\_LOG -118.22394 39.23618 -3.013 0.003021 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.15 on 155 degrees of freedom

(32 observations deleted due to missingness)

Multiple R-squared: 0.5187, Adjusted R-squared: 0.5094

F-statistic: 55.68 on 3 and 155 DF, p-value: < 2.2e-16

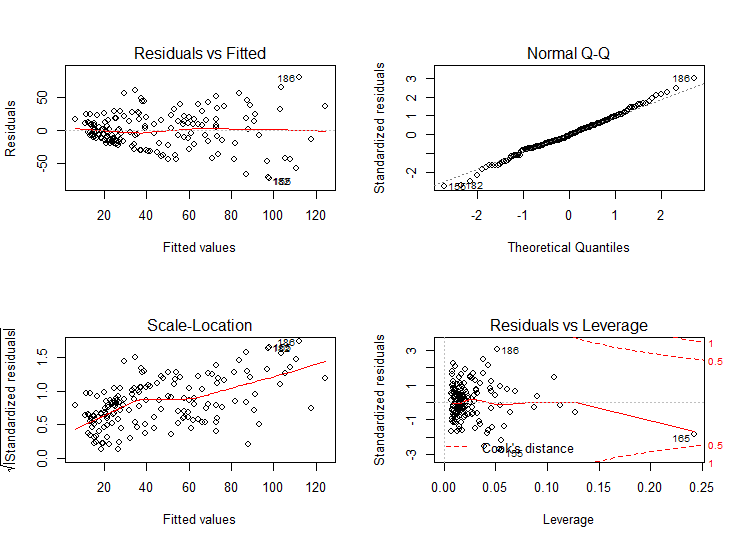
The R-squared value of the model has slightly increased after removing the insignificant variable.

Visual summary of the updated model using R:

#Visual summary of model

par(mfrow = c(2,2))

plot(model\_ABR)



From the first plot, Residuals Vs Fitted values, it is observed that the error variance of the model is not randomly distributed for different fitted values.

This violates the assumption of Independence of Errors of the model.

The observation indicates a possible non-linear relationship between the dependent and at least one of the dependent variables, which is in line with the observation from the Preliminary Analysis.

The variable PRIM\_LOG is removed and the model is re-evaluated:

Call:

lm(formula = ABR ~ SEC + LABF, data = gii\_data)

Residuals:

Min 1Q Median 3Q Max

-77.793 -17.208 0.357 14.635 90.634

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 73.09258 9.19480 7.949 3.35e-13 \*\*\*

SEC -0.86607 0.07491 -11.561 < 2e-16 \*\*\*

LABF 0.47354 0.14225 3.329 0.00108 \*\*

---

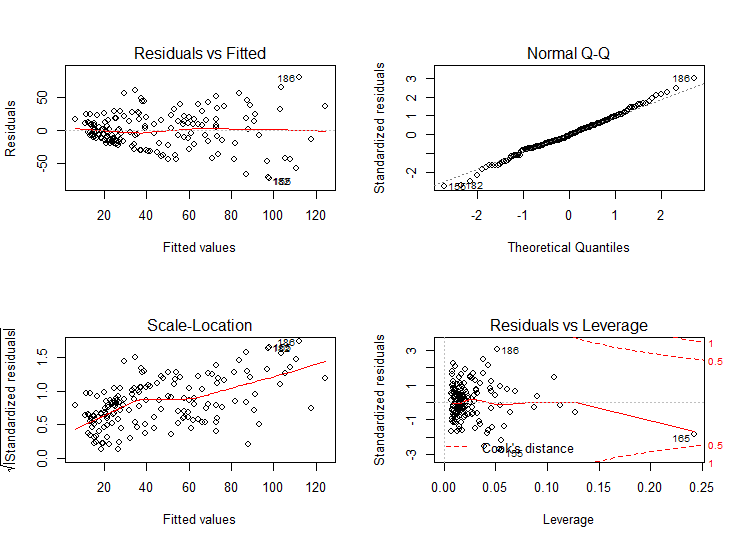
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.89 on 158 degrees of freedom

(30 observations deleted due to missingness)

Multiple R-squared: 0.4895, Adjusted R-squared: 0.4831

F-statistic: 75.76 on 2 and 158 DF, p-value: < 2.2e-16



Although the model R-squared has slightly reduced, the Residual vs Fitted value plot now looks much better.

Also, the Standardised Residuals are generally less than 2, and seems fairly normally distributed.

The standard residuals are also within the Cook’s distance, indicating there are no highly influencing observations.

We can perform further tests on the model to verify that assumptions are met,

## Testing Assumptions

1. Linear Relationship between independent and dependent variables:

This was verified during preliminary analysis

1. Multicollinearity:

Using the correlation matrix, it was observed that no two dependent variables are highly correlated with each other.

A Variation Influence Factor (VIF) test can be conducted to further confirm absence of multicollinearity.

#VIF Test for multicollinearity

vif(model\_ABR)

SEC LABF

1.005907 1.005907

VIF values for both dependent variables are well below 10, and the average value is close to 1. We can confirm absence of multicollinearity, as per Pg 1033, Section 8.7.5, A. Field Discovering Statistics with SPSS.

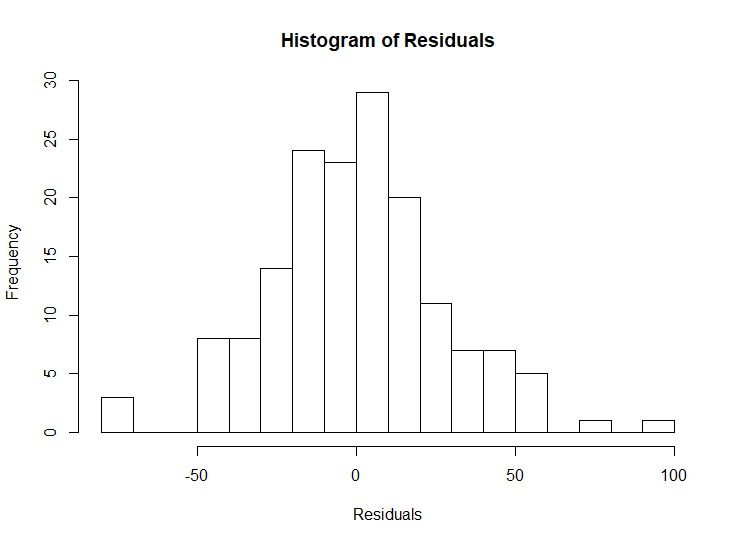
1. Multivariate normality:

The model residuals should be normally distributed.

This can be visually inspected using the Q-Q plot of the standardized residuals, or with a histogram of residual values:

#Histogram of Residuals

hist(model\_ABR$residuals, breaks = 20, xlab = "Residuals", main = "Histogram of Residuals")



The residuals are approximately normally distributed.

Normality can be confirmed by applying the Shapiro Wilk test of normality on the residual values.

#Test residuals normality

shapiro.test(model\_ABR$residuals)

Shapiro-Wilk normality test

data: model\_ABR$residuals

W = 0.9883, p-value = 0.2001

Since the p value is greater than 0.05, the null hypothesis for the Shapiro Wilk test is rejected.

We can confirm that the residuals are normally distributed.

1. Independence of Errors:

This assumption requires that the errors for any two observations are uncorrelated. Violation of this assumption can lead to unreliable confidence intervals for our model, as mentioned in Pg. 949, A. Field., Discovering Statistics using IBM SPSS Statistics 4th edition.

The independence of errors can be tested using the Durbin Watson Test, which tests for serial correlation between the errors.

#Test independence of errors

durbinWatsonTest(model\_ABR)

lag Autocorrelation D-W Statistic p-value

1 0.03719336 1.85744 0.358

Alternative hypothesis: rho != 0

The D-W statistic is close to 2, hence independence of errors is tested.

1. Homoscedasticity

From the plot of Residuals vs Fitted values, it is observed that the error variance slightly with increase in the fitted values, this indicates presence of heteroscedasticity.

We can confirm the same with a Non-Constant Variance (NVC) test.

#Test homoscedasticity

ncvTest(model\_ABR)

Non-constant Variance Score Test

Variance formula: ~ fitted.values

Chisquare = 52.73816 Df = 1 p = 3.811141e-13

The p value is well below the critical value, and presence of heteroscedasticity is confirmed.

Violation of this assumption means that the model confidence intervals are not reliable, and the model cannot be used to produce reliable predictions.

However, the model can still be used for calculating the mean response of the independent variable with changes in the dependent variables.

The Beta values of the regression equations are still valid, however not optimal. Pg. 951, A. Field., Discovering Statistics using IBM SPSS Statistics 4th edition

# Logistic Regression

We want to measure the impact of gender disparity on the overall HDI status of a country.

While the % of women with secondary education is included in the GII calculation, ratio of girls to boys in education is not used.

The combination of these metrics will be helpful in differentiating between countries with generally low levels of education, from countries where there is a lesser ratio of girls, pointing to a potential systemic disadvantage to girls.

The latest available data for ratio of girls in primary, secondary and tertiary education is used from the data.un.org files, and these variables are coded as PRIMR, SECR and TERR.

# Logistic Regression: Preliminary Analysis

A correlation matrix is created in R to check for multicollinearity in dependent variables.

Cor(glm\_vars, use = “pairwise.complete.obs”)

GII SEC PRIMR SECR TERR

GII 1.0000000 -0.8030225 -0.3972269 -0.4179567 -0.3827530

SEC -0.8030225 1.0000000 0.3652352 0.3924499 0.3933266

PRIMR -0.3972269 0.3652352 1.0000000 0.6095399 0.2448537

SECR -0.4179567 0.3924499 0.6095399 1.0000000 0.4689239

TERR -0.3827530 0.3933266 0.2448537 0.4689239 1.0000000

Although some of these variables have a substantial correlation, they are well below the generally acceptable limit of around 0.8, Pg. 1013, A. Field., Discovering Statistics using IBM SPSS Statistics 4th edition

# Logistic Regression Model Creation

A binomial model is created using R with all variables:

glm\_model = glm(HDI ~ GII + PRIMR + SECR + TERR, data = glm\_data, family = "binomial")

Call:

glm(formula = HDI ~ GII + PRIMR + SECR + TERR, family = "binomial",

data = glm\_data)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.31711 0.00770 0.02821 0.17281 1.89781

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 10.267 6.883 1.492 0.13580

GII -19.018 6.108 -3.114 0.00185 \*\*

PRIMR 2.260 5.626 0.402 0.68787

SECR -4.572 3.225 -1.418 0.15629

TERR 3.115 1.550 2.009 0.04455 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 130.843 on 146 degrees of freedom

Residual deviance: 49.215 on 142 degrees of freedom

(44 observations deleted due to missingness)

AIC: 59.215

Number of Fisher Scoring iterations: 8

Although it may seem from the first output that TERR is a significant variable, systematically removing other variables from the model reveals that GII is the only significant predictor in our model.

Call:

glm(formula = HDIL ~ GII, family = "binomial", data = glm\_data)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.10745 -0.28459 -0.05766 -0.00801 2.79206

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -11.963 2.376 -5.036 4.75e-07 \*\*\*

GII 21.207 4.404 4.815 1.47e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 141.658 on 158 degrees of freedom

Residual deviance: 63.761 on 157 degrees of freedom

AIC: 67.761

Number of Fisher Scoring iterations: 7

# Logistic Regression Model Evaluation

The performance of the model can be visually examined by creating a Receiver Operator Characteristic (ROC) Curve, which helps visualize the True Positive and False Positive Rates of the model for different Threshold values.

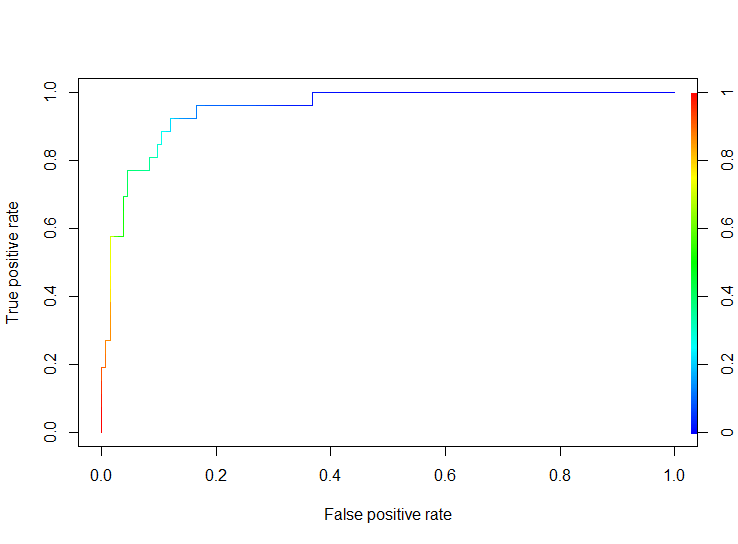
#Operating Characteristics of model

glm\_pred = predict(glm\_model, type = "response")

ROCPred = prediction(glm\_pred, glm\_data$HDIL)

ROCPerf = performance(ROCPred, "tpr", "fpr")

plot(ROCPerf, colorize = T)



The Accuracy statistics of the model using optimal Threshold value can be checked in R.

> #Classification Table

> classification\_table(glm\_model, glm\_model$model[,1])

Actual

Predict 0 1

0 126 6

1 7 20

Specificity: 0.7407407

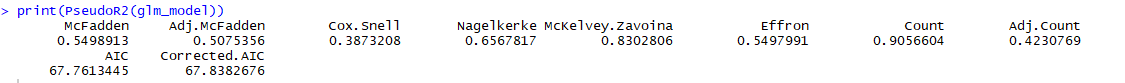
Sensitivity: 0.9545455

Total Accuracy: 0.918239

The model achieves an accuracy of 0.918 against a baseline accuracy of 0.83.

## Testing Predictive Power and Goodness of Fit

The Predictive Power of the model can be estimated using several tests that calculate Pseudo R-Squared value of the model:



The Goodness of Fit of the Model can be estimated with the Hosmer and Lemeshow Test:

> generalhoslem::logitgof(glm\_data$HDIL, fitted(glm\_model))

Hosmer and Lemeshow test (binary model)

data: glm\_data$HDIL, fitted(glm\_model)

X-squared = 3.0555, df = 8, p-value = 0.9308

The null hypothesis for the test is rejected, and we can conclude that the model is significant.

## Interpretation

An increase in the Gender Inequality Index score by 0.01, increases the odds of a country belonging to the Low HDI group by 1.2 times, or 20%.

# Findings

It is found that the share of women in parliament is not a significant variable in predicting the Adolescent Birth Rate in a country.

The Percentage of women with secondary education, and Labour Force participation are both significant, with coefficients -0.86 and 0.47 respectively.

The positive coefficient for Labour Force participation maybe viewed as surprising, as this indicates that counties with higher labour force participation on women, also have higher adolescent birth rate.

Although the ratio of girls to boys in Primary education is a significant variable, it cannot be used to estimate a linear model, due to lack of linearity with Adolescent Birth Rate, the dependent variable.

It may be possible to use further Non-Linear transformation techniques, for ex Box-Cox transformation, on this variable, to see if it can improve the model.

The linear model to estimate the Adolescent Birth Rate is heteroscedastic. It is desirable that more missing variables are identified and introduced to attempt to improve the model.

It is found that the Gender Inequality Index is able to predict whether a country belongs to the Low HDI group, with a 91% accuracy.

An increase in the GII score by 0,01 increases the likelihood of a country being Low HDI by 1.2 times, or 20%.

# References

1. FIELD, A. P., MILES, J., & FIELD, Z. (2012). Discovering statistics using IBM SPSS Statistics. London, Sage.
2. Ray, Katika, How Fewer Women in Politics Impacts Policy and the Economy, Pipeline Equity (  
   <https://www.pipelineequity.com/katicas-voice/women-politics-impacts-policy-economy/>)
3. UNDP, United Nations, Human Development Report

# Bibliography

1. Caffo, Brian (2015). Regression Models for Data Science in R. Lean Publishing.
2. <http://people.duke.edu/~rnau/testing.htm>
3. <https://www.statisticssolutions.com/assumptions-of-multiple-linear-regression/>
4. FIELD, A. P., MILES, J., & FIELD, Z. (2012). Discovering statistics using R. London, Sage.
5. <http://www.sthda.com/english/articles/36-classification-methods-essentials/151-logistic-regression-essentials-in-r/#simple-logistic-regression>