Faculty Name: Prof. José Manuel Magallanes, PhD

Student Name: Samikshya Pandey

Class Name: PUBPOL 542: Computational Thinking For Governance Analytics

File details: Conducting Regression Analysis

File Source: Merged data found in https://github.com/PUBPOL-542-Computational-Thinking/Merge/raw/main/UpdatedTJHemaSamik.csv. The following documents provides instructions on conducting a OLS regression analysis.

First, we need to import the excel files that consists of the data we need to analyze. To do so, we provide the location for the excel files and tell R to read the csv/excel file using the code read.csv.

The data name final data is created.

```
mergecsv = "https://github.com/PUBPOL-542-Computational-Thinking/Merge/raw/main/UpdatedTJHemaSamik.csv"

finaldata = read.csv(mergecsv)


row.names(finaldata)= NULL
```

Once we have imported the data set, we need to ensure the data structure is fit for analysis. To learn more about the data, we call the functions str to learn details of data frame.

```
### veryfying data structure

str(finaldata, width = 50, strict.width = 'cut')
```

```
## 'data.frame':    126 obs. of  6 variables:
##  $ Country     : chr  "Albania" "Algeria" "Ango"..
##  $ lessthan5_50: num  33.8 28.6 89.3 12.2 0.7 0...
##  $ Continent   : chr  "Europe" "Africa" "Africa"..
##  $ FPI         : num  67629 996868 67381 885029 ..
##  $ FDI         : num  0.199 0.145 0.157 0.322 0...
##  $ FIEI        : num  0.607 0.749 0.574 0.607 0...
```

Now to conduct a regression analysis, the first process is to develop hypothesis against which the anlysis will be conducted. For the purpose of this anlysis, we have developed 3 set of hypothesis

Hypothesis 1 FDI decreases as percentage of population earning less than $5.50 increases

Hypothesis 2 FPI decreases as percentage of population earning less than $5.50 increases

Hypothesis 3 Percentage of population earning less than $5.50 decreases as FPI and FIEI advances

```
## hypothesis 1 : FDI decreases as percentage of population earning less than $5.50 increases

hypo1 = formula(FDI~ lessthan5_50)

#hypothesis 2:  FPI decreases as percentage of population earning less than $5.50 increases

```

```
hypo2 = formula(FPI~ lessthan5_50)

#hypothesis 3: Percentage of population earning less than $5.50 decreases as FPI and FIEI advances

hypo3 =formula(lessthan5_50~ FPI*FIEI )
```

After explaining the hypothesis, we need to get the results. Since the dependent variables are not a binary outcome, we can use OLS regression for analysis.

The regression analysis required uses the code glm. This code fits the generalized linear models.We can observe results below:

```
### Getting results

Result1 = glm(hypo1,
              data = finaldata,
              family = 'gaussian')

Result2 = glm(hypo2,
              data = finaldata,
              family = 'gaussian')

Result3 = glm(hypo3,
              data = finaldata,
              family = 'gaussian')
```

Reading results: We call the functions summary to obtain results for each of our hypotheses.

Interpreting from the summary of result, we can learn that:

For the first hypothesis: Can we observe a decrease in FDI when the poverty rate (percentage of people earning less than 5.50 falls) Interpretation for hypothesis 1. We can observe an indirect relationship between poverty and FDI like we had initially hypothesize.

Similar reading can be done for hypothesis 2 and hypothesis 3.

```
### Seeing results

# For the first hypothesis: Can we observe a decrease in FDI when the poverty rate (percentage of peopl

summary(Result1)
```

```
##
## Call:
## glm(formula = hypo1, family = "gaussian", data = finaldata)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.33148  -0.10082  -0.00274   0.07349   0.45416
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.5097954  0.0217208   23.47   <2e-16 ***
## lessthan5_50 -0.0045418  0.0004019  -11.30   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.02599656)
##
##     Null deviance: 6.5442  on 125  degrees of freedom
## Residual deviance: 3.2236  on 124  degrees of freedom
## AIC: -98.317
##
## Number of Fisher Scoring iterations: 2
```

```
### Interpretation for hypothesis 1. We can observe an indirect relationship between poverty and FDI li
```

```
### Results for hypothesis 2

summary(Result2)
```

```
##
## Call:
## glm(formula = hypo2, family = "gaussian", data = finaldata)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2055528  -1189794   -476872    114913  14283255
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2163918     298266   7.255 3.83e-11 ***
## lessthan5_50   -24318       5518  -4.407 2.24e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 4.901977e+12)
##
##     Null deviance: 7.0305e+14  on 125  degrees of freedom
## Residual deviance: 6.0785e+14  on 124  degrees of freedom
## AIC: 4043.4
##
## Number of Fisher Scoring iterations: 2
```

```
# The results also show an indirect relationship
```

```
## results for hypothesis 3

summary(Result3)
```

```
##
## Call:
## glm(formula = hypo3, family = "gaussian", data = finaldata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -67.421  -26.629   -5.349   25.517   68.242
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.047e+02  1.489e+01   7.029 1.29e-10 ***
## FPI         -5.681e-06  8.152e-06  -0.697 0.487165
## FIEI        -9.518e+01  2.405e+01  -3.958 0.000128 ***
## FPI:FIEI     2.418e-06  1.162e-05   0.208 0.835550
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 998.722)
##
##     Null deviance: 160982  on 125  degrees of freedom
## Residual deviance: 121844  on 122  degrees of freedom
## AIC: 1233.7
##
## Number of Fisher Scoring iterations: 2
```

Now after obtaining the regression values, we need to test for the better model. The first way to do this by using a chi-square distribution test.

```
### Searching for a better model

anova(Result1, Result2, test = "Chisq")
```

```
## Warning in anova.glmlist(c(list(object), dotargs), dispersion = dispersion, :
## models with response ’"FPI"’ removed because response differs from model 1
```

```
## Analysis of Deviance Table
##
## Model: gaussian, link: identity
##
## Response: FDI
##
## Terms added sequentially (first to last)
##
##
##              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                           125     6.5442
## lessthan5_50  1   3.3207       124     3.2236 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
anova(Result2, Result3, test = "Chisq")
```

```
## Warning in anova.glmlist(c(list(object), dotargs), dispersion = dispersion, :
## models with response ’"lessthan5_50"’ removed because response differs from
## model 1
```

```
## Analysis of Deviance Table
##
```

```
## Model: gaussian, link: identity
##
## Response: FPI
##
## Terms added sequentially (first to last)
##
##
##              Df   Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                             125 7.0305e+14
## lessthan5_50  1 9.5202e+13       124 6.0785e+14 1.048e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Now,since Result 2 is better and Result 1 and Result 3 is better than result 2, we can conclude that Hypothesis 3, i.e Result 3 fits the model best.

To confirm the results, we can also test the residual value for all three hypothesis. No test the residual, we call the package rsq for library.

```
## Checking the Rsquare to understand the residual value for all 3 hypothesis
library(rsq)
rsq(Result1, adj = T)
```

```
## [1] 0.5034451
```

```
rsq(Result2, adj = T)
```

```
## [1] 0.1284405
```

```
rsq(Result3, adj = T)
```

```
## [1] 0.2245085
```

We can now proceed to visualize the data. For this, we need to call two package from library, dotwhisker and ggplot.

Using the code dwplot, we plot the predicted value of Result 3 against 2 standard deviation.

```
# summary plots to visualize the data

library(dotwhisker)
```

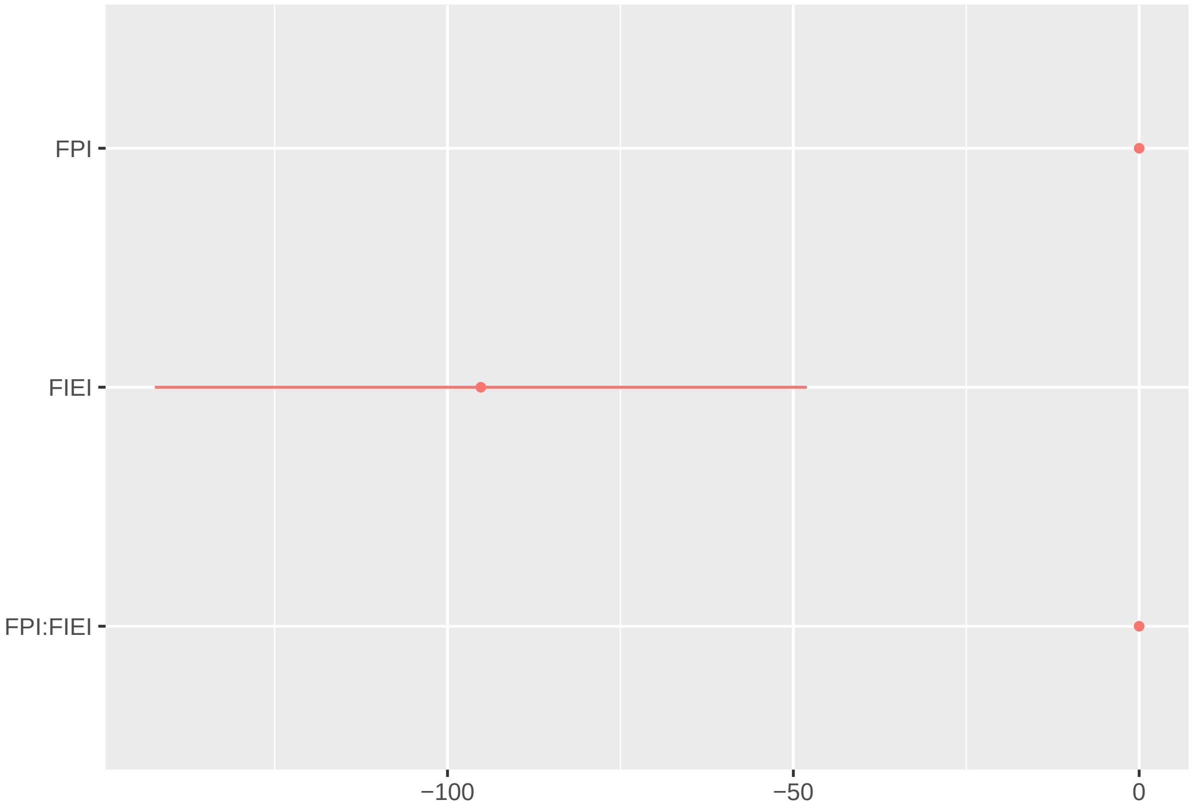
```
## Warning: package ’dotwhisker’ was built under R version 4.0.4
```

```
## Loading required package: ggplot2
```

```
## Warning: package ’ggplot2’ was built under R version 4.0.4
```

```
## Warning in checkMatrixPackageVersion(): Package version inconsistency detected.
## TMB was built with Matrix version 1.3.2
## Current Matrix version is 1.2.18
## Please re-install ’TMB’ from source using install.packages(’TMB’, type = ’source’) or ask CRAN for a
## Registered S3 method overwritten by 'broom.mixed':
##   method      from
##   tidy.gamlss broom
```

```
library(ggplot2)

dwplot(Result3,by_2sd = F)
```



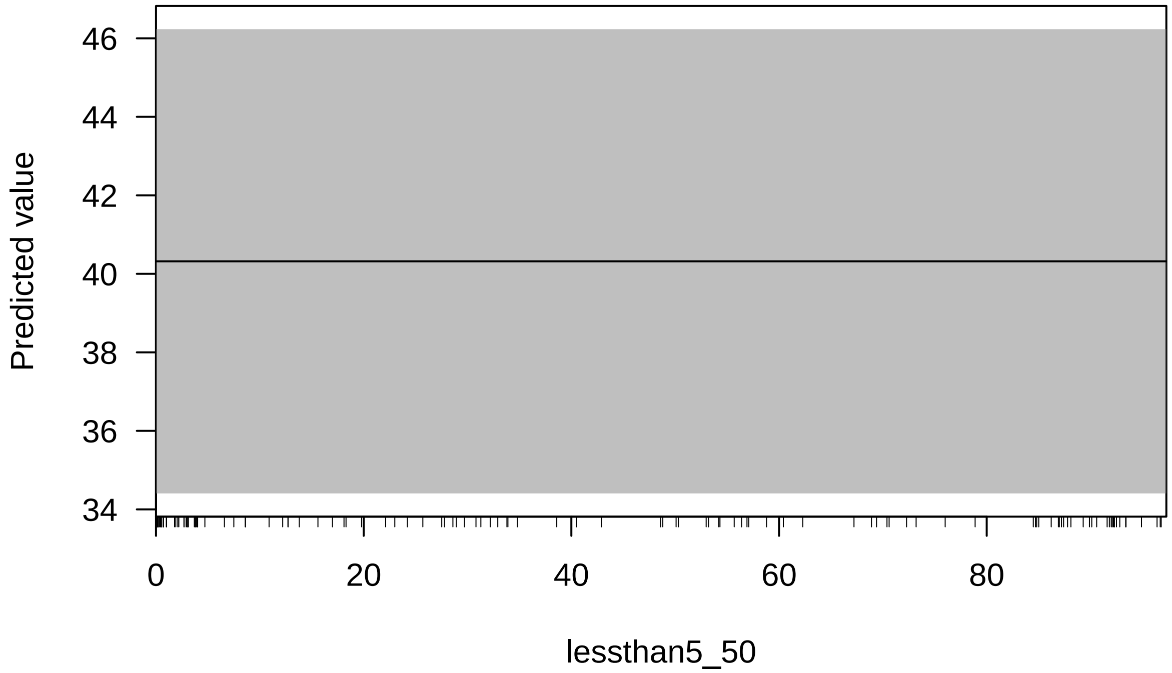
To capture the result of predicted value againsts given value, we can also use margins. For this, we must call margins from library.

The two graph uses data value collected for Result 3 and result 1.

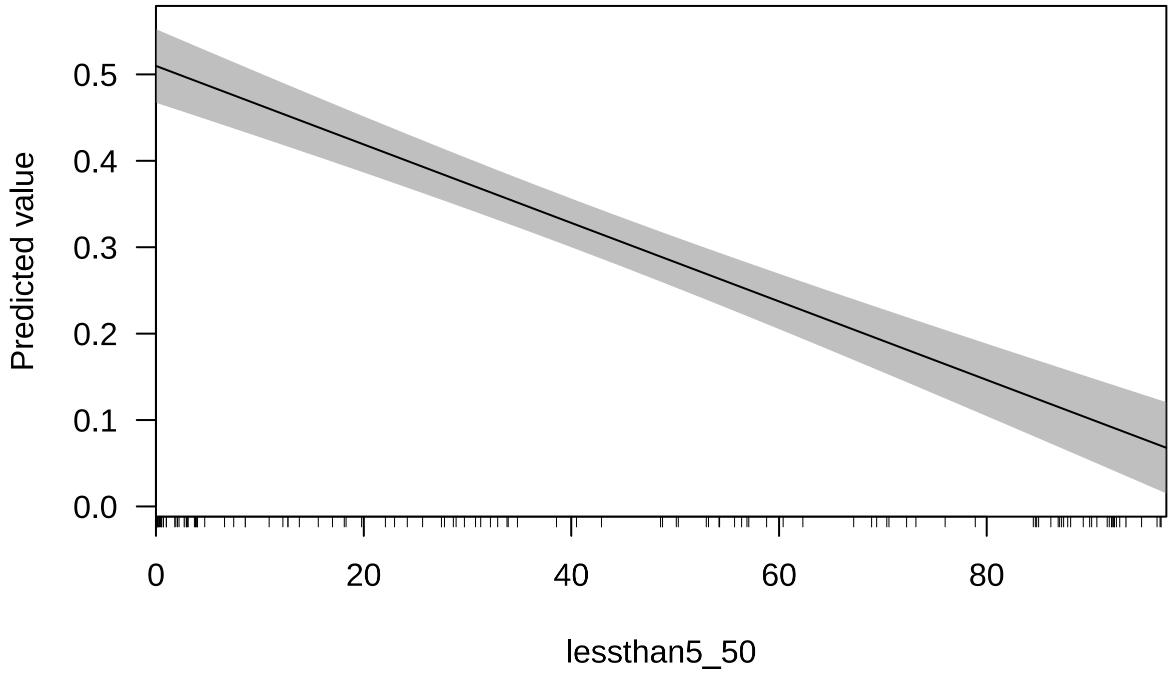
```
# Using the margins library

library(margins)
```

```
## Warning: package 'margins' was built under R version 4.0.4
```

```
cplot(Result3,'lessthan5_50')
```

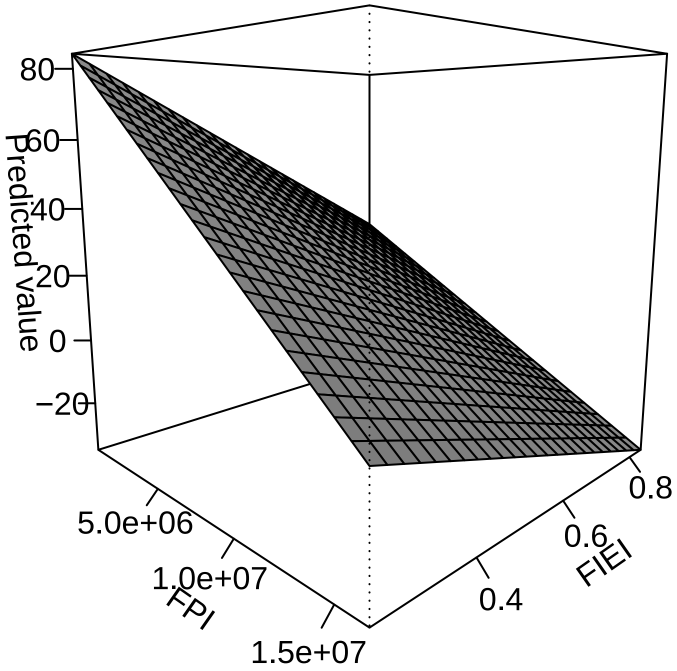


```
cplot(Result1,'lessthan5_50')
```



Another way to visaluze the result is looking for interactions between our values. To do this, we can call for perspe code, as can be oberved below.

```
## Looking into the interactions
persp(Result3)
```



The above codes above, therefore, shows a tutorial in conducting regression analysis in R.

Thank you!