Department of Decision Science

Faculty of Business

University of Moratuwa

Semester 04

DA2420 – Introduction to Econometrics

Research Question

Does the population of a country increase the labor force of the country?

Group Assignment

Group 10

Group Members:

1. M. W. S. A. U. SILVA - 206121X
2. R. V. N. NETHMINI – 206083F
3. M. N. U. JAYASIRI – 206045R
4. S. M. D. O. A. SAMARAWEERA - 206107K
5. H. C. E. SUDUSINGHE – 206126R
6. A. W. W. M. P. U. BAKMEEWEWA – 206019R

Due Date of Submission

[10/02/2023]

Table of Contents

1. Introduction

1.1. Research Question

The research question: Does the population of a country increase the labour force of the country?

We are interested to look at the relationship between population of the country (X1) and labor force (Y), while controlling for the effects of unemployment (X2) and government spending on education (X2 = control variable). The goal of this study is to see if there is a substantial relationship between population and labor force, and if so, how unemployment and government education spending alter the effect of population on the labor force.

1.2. Conceptual Framework

The concepts of the research question can be operationalized as follows:
Dependent variable (Y): Labour force
Main explanatory variable (X1): Population
Other regressor (X2): Unemployment
Control variable (X3): Government expenditure on education

This forms the conceptual framework of the analysis. By defining these variables, we have specified the concepts that will be measuring and analyzing in the study. The labor force is the dependent variable that is trying to explain or predict based on the values of population and unemployment, while controlling for the effect of government expenditure on education. The population and unemployment are the main explanatory variables that is believed may have an impact on the labor force. The government expenditure on education is included as a control variable to control for any confounding effects on the relationship between population, unemployment, and labor force.

1.3. Hypotheses

Based on the research question, the following hypotheses can be formulated for testing:

1. H0: There is no relationship between population and labor force.
   H1: There is a relationship between population and labor force.

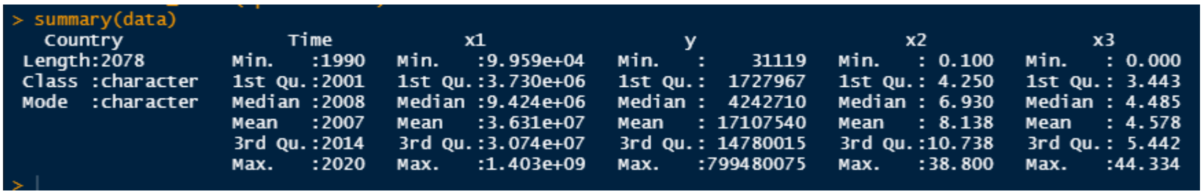
2. H0: The effect of population on labor force is not affected by unemployment.
   H1: The effect of population on labor force is affected by unemployment.

3. H0: The effect of population on labor force is not affected by government expenditure on education.
   H1: The effect of population on labor force is affected by government expenditure on education.

2. Data and methodology

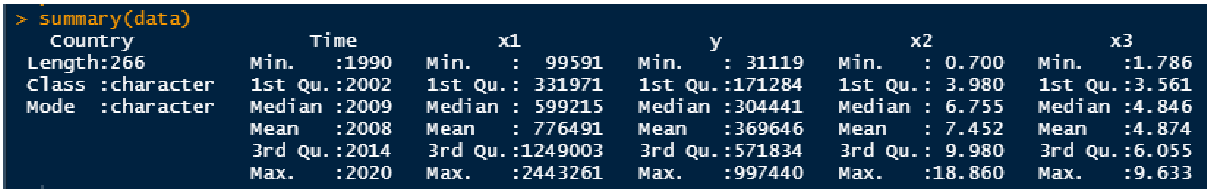
2.1.Description of the dataset

Summary with outliers



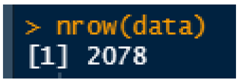
```
> summary(data)
   Country              Time            x1                  y                  x2               x3        
 Length:2078        Min.   :1990   Min.   :9.959e+04   Min.   :    31119   Min.   : 0.100   Min.   : 0.000  
 Class :character   1st Qu.:2001   1st Qu.:3.730e+06   1st Qu.:  1727967   1st Qu.: 4.250   1st Qu.: 3.443  
 Mode  :character   Median :2008   Median :9.424e+06   Median :  4242710   Median : 6.930   Median : 4.485  
                    Mean   :2007   Mean   :3.631e+07   Mean   : 17107540   Mean   : 8.138   Mean   : 4.578  
                    3rd Qu.:2014   3rd Qu.:3.074e+07   3rd Qu.: 14780015   3rd Qu.:10.738   3rd Qu.: 5.442  
                    Max.   :2020   Max.   :1.403e+09   Max.   :799480075   Max.   :38.800   Max.   :44.334  
> 
```

Summary without outliers



```
> summary(data)
   Country              Time            x1                y                x2               x3       
 Length:266         Min.   :1990   Min.   :  99591   Min.   : 31119   Min.   : 0.700   Min.   :1.786  
 Class :character   1st Qu.:2002   1st Qu.: 331971   1st Qu.:171284   1st Qu.: 3.980   1st Qu.:3.561  
 Mode  :character   Median :2009   Median : 599215   Median :304441   Median : 6.755   Median :4.846  
                    Mean   :2008   Mean   : 776491   Mean   :369646   Mean   : 7.452   Mean   :4.874  
                    3rd Qu.:2014   3rd Qu.:1249003   3rd Qu.:571834   3rd Qu.: 9.980   3rd Qu.:6.055  
                    Max.   :2020   Max.   :2443261   Max.   :997440   Max.   :18.860   Max.   :9.633  
```

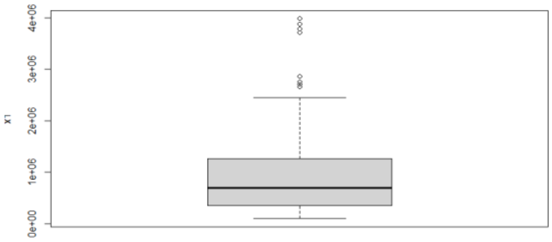
No: of observations after remove N/A values:



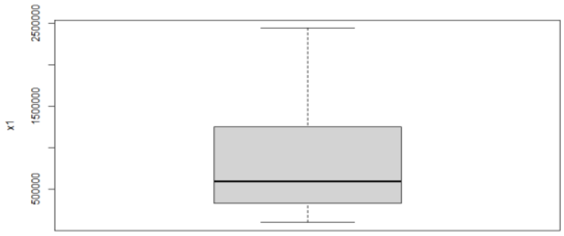
```
> nrow(data)
[1] 2078
```

Identification of outliers - Box plots

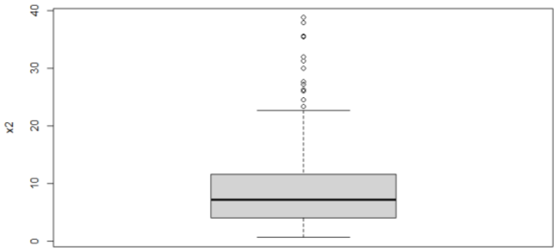
X1 & Y (with outliers)



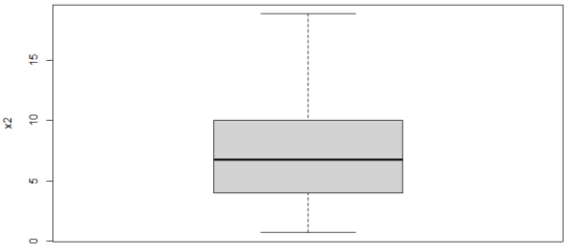
X1 & Y (without outliers)



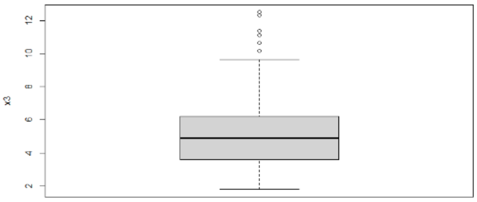
X2 & Y (with outliers)



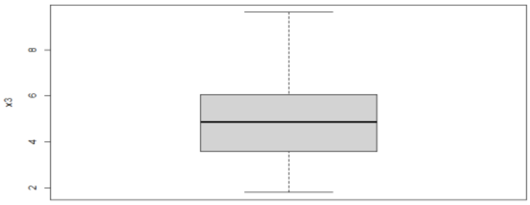
X2 & Y (without outliers)



X3 & Y (with outliers)



X3 & Y (without outliers)



2.2.Regression Model

Based on the research question and the variables identified, a linear regression model can be used for this research question. Since the variables are all in millions, they are converted to logs before running the regression. The functional form of the model would be as follows:

$$\boxed{\begin{array}{c}\text{Labor force (Y)} = \beta0 + \beta1 * \text{Population (X1)} + \beta2 * \text{Unemployment (X2)} + \beta3 * \text{Government} \\ \text{expenditure on education (X3/Control variable)} + \varepsilon\end{array}}$$

$$\boxed{Y = \beta0 + \beta1 * X1 + \beta2 * X2 + \beta3 * X3 + \varepsilon}$$

Where β0, β1, β2, and β3 are the regression coefficients, and ε is the error term. The coefficients represent the change in the dependent variable (Labor force) associated with a unit change in the independent variables, holding all other variables constant. The error term represents the residual or unexplained variation in the dependent variable.

The explanatory variables in the model were selected based on the following reasons:

Population (X1): This variable is the main explanatory variable and represents the size of the population of a country. It is expected that the larger the population, the larger the labor force of the country, so this variable is included in the model to test the relationship between population and labor force.

Unemployment (X2): This variable represents the percentage of the labor force that is unemployed and actively looking for work. Unemployment can be a factor that affects the labor force, so this variable is included in the model to capture its effect on the relationship between population and labor force.

Government expenditure on education (Control variable): This variable represents the government expenditure on education in a given country. Government expenditure on education can have a direct effect on the labor force, so this variable is included as a control variable to control for its effect on the relationship between population and labor force.

These explanatory variables were selected based on their expected effect on the dependent variable and to capture the relationship between population and labor force, while controlling for the effect of unemployment and government expenditure on education.

3. Pooled OLS

3.1.Estimate Regression Model & Interpretation

$$\boxed{\text{Labor force (Y)} = \beta0 + \beta1 * \text{Population (X1)} + \beta2 * \text{Unemployment (X2)} + \beta3 * \text{Government expenditure on education (X3/Control variable)} + \varepsilon}$$

$$\boxed{\ln(Y) = \beta0 + \beta1 * \ln(X1) + \beta2 * \ln(X2) + \beta3 * \ln(X3) + \varepsilon}$$

The dependent variable $\ln(Y)$ represents the natural logarithm of the labor force, which is the number of people who are either employed or actively seeking employment.

The independent variables are $\ln(X1)$, $\ln(X2)$, and $\ln(X3)$, which represent the natural logarithm of the population, unemployment rate, and government expenditure on education, respectively. These variables are included in the model as potential explanatory factors for the variation in the labor force.

The constant term $\beta0$ represents the intercept of the regression line. It represents the value of $\ln(Y)$ when all independent variables equal 0.

The coefficients $\beta1$, $\beta2$, and $\beta3$ represent the change in $\ln(Y)$ associated with a 1% change in $\ln(X1)$, $\ln(X2)$, and $\ln(X3)$, respectively. For example, a positive coefficient for $\beta1$ would indicate that an increase in the population is associated with an increase in the labor force, while a negative coefficient would indicate the opposite. Similarly, positive coefficients for $\beta2$ and $\beta3$ would indicate that an increase in the unemployment rate and government expenditure on education are associated with a decrease in the labor force, while negative coefficients would indicate the opposite. According to the following this regression model covers 88.46% of the variation of Labour force.

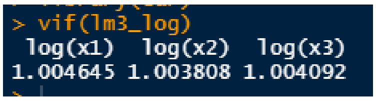
The error term ε represents the unpredictable variation in the labor force that is not explained by the independent variables. It accounts for the random errors and residuals in the model. Overall, this regression model is used to estimate the relationship between the labor force and various socioeconomic factors, and to make predictions about future trends in the labor force based on changes in these factors.

3.2.Diagnostics

Diagnosing and addressing issues such as multicollinearity, heteroscedasticity, model misspecification, and serial correlation are important steps in ensuring that the regression model is accurate and reliable.

3.2.1. Diagnostics for Multicollinearity

Multicollinearity occurs when two or more independent variables in the regression model are highly correlated. To diagnose multicollinearity, you can check the variance inflation factor (VIF) for each independent variable, which measures how much the variance of the estimated regression coefficients increases due to multicollinearity.



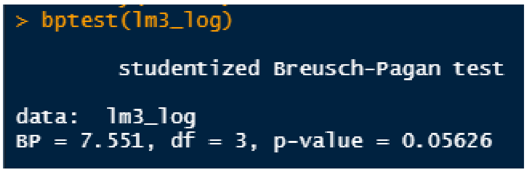
```
> vif(lm3_log)
 log(x1)  log(x2)  log(x3) 
1.004645 1.003808 1.004092 
```

Since VIF factors are lower than 10, there is no multicollinearity in this regression model.

3.2.2. Diagnostics for Heteroscedasticity

The Breusch-Pagan (BP) test is a commonly used statistical test to check for heteroscedasticity in a regression model. Heteroscedasticity occurs when the variance of the errors (residuals) in the model is not constant, but instead depends on the value of the independent variables.

H0: There is no heteroscedasticity in the regression model
H1: There is heteroscedasticity in the regression model



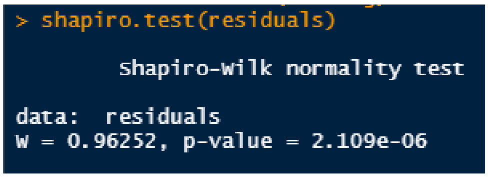
```
> bptest(lm3_log)

	studentized Breusch-Pagan test

data:  lm3_log
BP = 7.551, df = 3, p-value = 0.05626
```

Since P value is greater than significance level (0.05), H0 not rejected which means there is no heteroscedasticity in the regression model. So there’s no need to diagnostic for heteroscedasticity.

3.2.3. Diagnostics for Model Misspecification

Model misspecification occurs when the functional form of the regression model does not fit the underlying relationship between the independent and dependent variables. The Shapiro-Wilk test is one way to check for normality in your data, which is an assumption in many statistical models, including linear regression.

H0: Data is normally distributed.
H1: Data is not normally distributed.



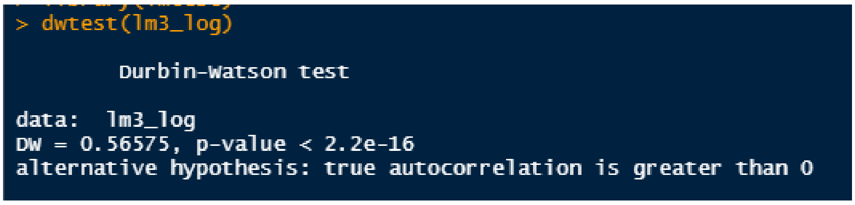
```
> shapiro.test(residuals)

        Shapiro-Wilk normality test

data:  residuals
W = 0.96252, p-value = 2.109e-06
```

Since P value is 0.000002109 and it's less than significance level (0.05), H0 rejected. Which means data is not normally distributed.

In this study, we used a regression model to examine the relationship between labour force and X variable. Our model included three predictor variables: x1, x2, and x3. We conducted several diagnostic tests to assess the validity of our model and ensure that it met the assumptions of the regression analysis. However, the results of the Shapiro-Wilk normality test indicated a deviation from normality in the residuals, suggesting that this model may not be perfectly capturing the relationship between the dependent and independent variables. In this case, given the limitations of our data and the nature of the relationships being examined, we have concluded that the model is appropriate, informative and provide valuable insights into the relationship between labour force and X variable.

3.2.4. Diagnostics for Serial Correlation

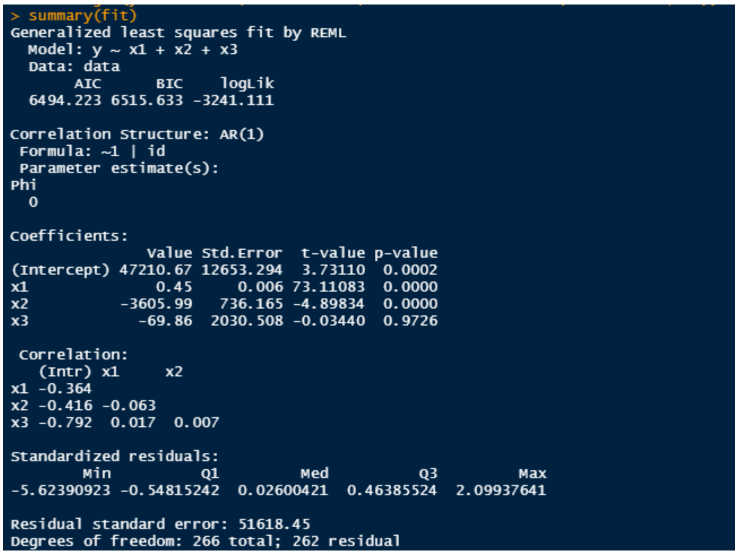
Serial correlation occurs when the error terms are correlated over time. To diagnose serial correlation, we can check the residuals for autocorrelation using the Durbin-Watson test.



```
> dwtest(lm3_log)

        Durbin-Watson test

data:  lm3_log
DW = 0.56575, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

Since P value is greater than significance level (0.05), H0 rejected. There is autocorrelation (positive serial correlation) in this regression model. It means that the residuals are not independent from one another and that the observations are correlated over time. This violates the assumptions of the ordinary least squares (OLS) regression method and can lead to biased and inefficient estimates.

In here we can use Generalized Least Squares (GLS) approach to cure autocorrelation. GLS considers the covariance structure of the residuals to produce more accurate results.



```
> summary(fit)
Generalized least squares fit by REML
  Model: y ~ x1 + x2 + x3
  Data: data
       AIC      BIC    logLik
  6494.223 6515.633 -3241.111

Correlation Structure: AR(1)
 Formula: ~1 | id
 Parameter estimate(s):
Phi
  0

Coefficients:
               Value Std.Error  t-value p-value
(Intercept) 47210.67 12653.294  3.73110  0.0002
x1              0.45     0.006 73.11083  0.0000
x2          -3605.99   736.165 -4.89834  0.0000
x3            -69.86  2030.508 -0.03440  0.9726

 Correlation:
   (Intr) x1     x2
x1 -0.364
x2 -0.416 -0.063
x3 -0.792  0.017  0.007

Standardized residuals:
        Min          Q1         Med          Q3         Max
-5.62390923 -0.54815242  0.02600421  0.46385524  2.09937641

Residual standard error: 51618.45
Degrees of freedom: 266 total; 262 residual
```

The summary of the GLS model fit provides information about the presence of serial correlation in the residuals. The Correlation Structure section shows that an AR(1) structure was used in the GLS model. The estimated parameter for the AR(1) structure, Phi, is 0, which suggests that there is no serial correlation in the residuals.

3.3.Omitted Variables

Omitted variable: Government Expenditure on Education (X3)

In an analysis, it's possible that there are important variables that have been omitted from the analysis that could affect the results. When a significant variable is omitted from the analysis, this can result in biased coefficient estimates. This is known as omitted variable bias. If we consider government expenditure on education (X3) as an omitted variable in this study, it is important to consider its potential impact on the labor force and the main independent variable Population.

Hypothesis 1: Government expenditure on education has a positive effect on the labor force.

This means that as the government invests more in education, the labor force will increase. If this is the case, then omitting X3 (Government expenditure on education) from the analysis would lead to an underestimate of the true effect of the main X variable (X1) on the labor force. This is because X3 and X1 are likely to be positively correlated, meaning that as X1 increases, X3 is also likely to increase. This will cause us to attribute some of the effect of X3 on the labor force to X1, leading to an underestimation of the true effect of X1 on the labor force.

Hypothesis 2: Government expenditure on education has a negative effect on the labor force.

This means that as the government invests more in education, the labor force will decrease. If this is the case, then omitting X3 from the analysis would lead to an overestimate of the true effect of the main X variable (X1) on the labor force.

This is because X3 and X1 are likely to be negatively correlated, meaning that as X1 increases, X3 is likely to decrease. This will cause us to attribute some of the effect of X3 on the labor force to X1, leading to an overestimate of the true effect of X1 on the labor force.

In general, the omitted variable bias occurs when an omitted variable is correlated with both the dependent variable and one or more of the independent variables. The direction and magnitude of the bias will depend on the direction and strength of the correlation between the omitted variable and the independent variables.

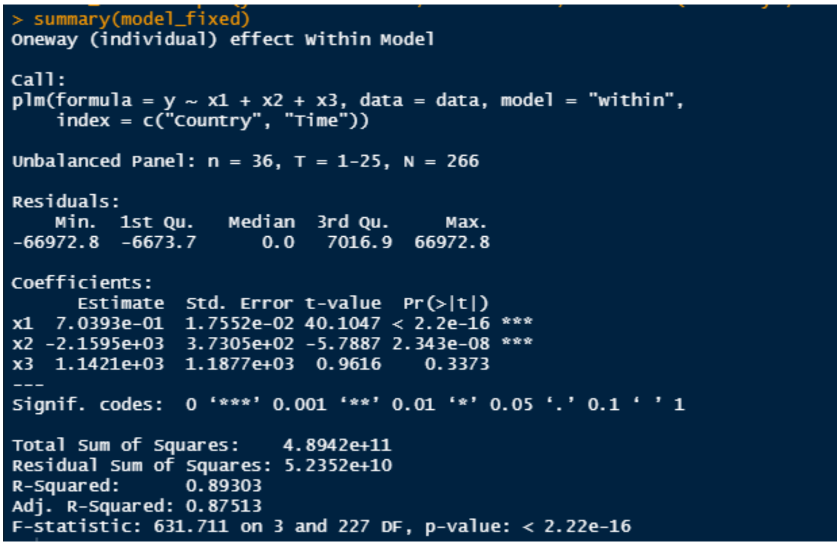
4. Panel Regression

Panel regression is a type of econometric model used to analyze the relationship between a dependent variable and one or more independent variables, where the data is collected for multiple units over multiple time periods. In the context of panel regression, the units could be individuals, firms, countries, etc. and the time periods could be years, quarters, months, etc.

4.1.Panel-regression models

4.1.1. Fixed Effect panel-regression model

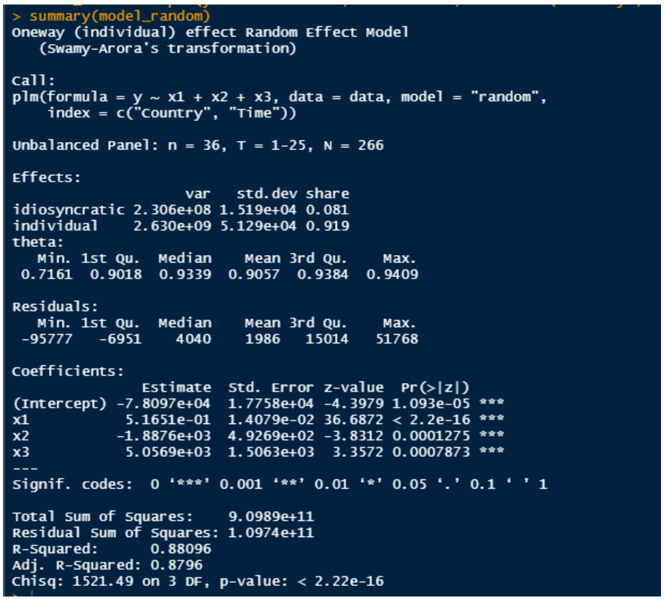
Fixed effects panel regression models control for the unobservable characteristics of each unit by including unit-specific dummy variables in the regression.



```
> summary(model_fixed)
Oneway (individual) effect Within Model

Call:
plm(formula = y ~ x1 + x2 + x3, data = data, model = "within",
    index = c("Country", "Time"))

Unbalanced Panel: n = 36, T = 1-25, N = 266

Residuals:
    Min.  1st Qu.   Median  3rd Qu.     Max.
-66972.8  -6673.7      0.0   7016.9  66972.8

Coefficients:
      Estimate  Std. Error t-value  Pr(>|t|)
x1  7.0393e-01  1.7552e-02 40.1047 < 2.2e-16 ***
x2 -2.1595e+03  3.7305e+02 -5.7887 2.343e-08 ***
x3  1.1421e+03  1.1877e+03  0.9616    0.3373
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    4.8942e+11
Residual Sum of Squares: 5.2352e+10
R-Squared:      0.89303
Adj. R-Squared: 0.87513
F-statistic: 631.711 on 3 and 227 DF, p-value: < 2.22e-16
```

4.1.2. Random Effect panel-regression model

Random effects panel regression models assume that the unit-specific effects are random and independently distributed with mean zero. The random effects model allows for a more flexible specification, as it allows for the unit-specific effects to be correlated with the independent variables.



```
> summary(model_random)
Oneway (individual) effect Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = y ~ x1 + x2 + x3, data = data, model = "random",
    index = c("Country", "Time"))

Unbalanced Panel: n = 36, T = 1-25, N = 266

Effects:
                    var   std.dev share
idiosyncratic 2.306e+08 1.519e+04 0.081
individual    2.630e+09 5.129e+04 0.919
theta:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.7161  0.9018  0.9339  0.9057  0.9384  0.9409

Residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -95777   -6951    4040    1986   15014   51768

Coefficients:
               Estimate  Std. Error z-value  Pr(>|z|)
(Intercept) -7.8097e+04  1.7758e+04 -4.3979 1.093e-05 ***
x1           5.1651e-01  1.4079e-02 36.6872 < 2.2e-16 ***
x2          -1.8876e+03  4.9269e+02 -3.8312 0.0001275 ***
x3           5.0569e+03  1.5063e+03  3.3572 0.0007873 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    9.0989e+11
Residual Sum of Squares: 1.0974e+11
R-Squared:      0.88096
Adj. R-Squared: 0.8796
Chisq: 1521.49 on 3 DF, p-value: < 2.22e-16
```

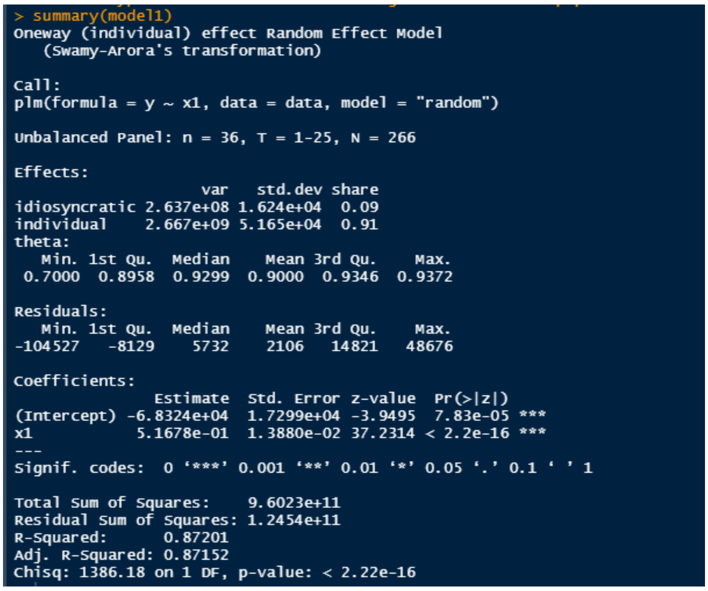
4.2.Identification of the best suited panel-regression model

When considering these two techniques, the adjusted R^2 of random effect model is 87.96% while the adjusted R^2 of fixed effect model is 87.51%. Based on the research question and the effect of population on labor force is different across countries, and that this difference is due to random factors, random effects panel-regression method will be more appropriate since this model allows for the estimation of both within-country and between-country effects.

4.3.Interpretation of Hypothesis

1. Hypothesis 1

H0: There is no relationship between population and labor force.
H1: There is a relationship between population and labor force.

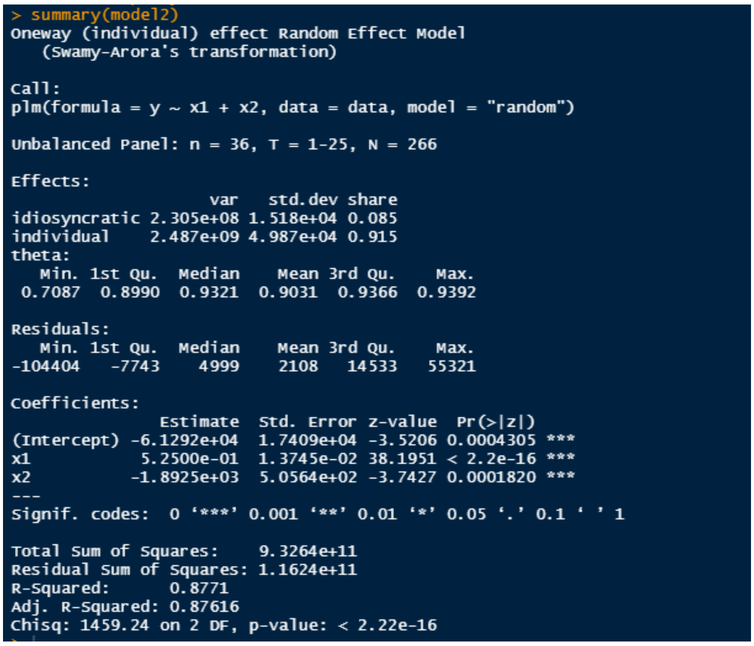


```
> summary(model1)
Oneway (individual) effect Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = y ~ x1, data = data, model = "random")

Unbalanced Panel: n = 36, T = 1-25, N = 266

Effects:
                    var   std.dev share
idiosyncratic 2.637e+08 1.624e+04  0.09
individual    2.667e+09 5.165e+04  0.91
theta:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.7000  0.8958  0.9299  0.9000  0.9346  0.9372

Residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-104527   -8129    5732    2106   14821   48676

Coefficients:
               Estimate  Std. Error z-value  Pr(>|z|)
(Intercept) -6.8324e+04  1.7299e+04 -3.9495  7.83e-05 ***
x1           5.1678e-01  1.3880e-02 37.2314 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    9.6023e+11
Residual Sum of Squares: 1.2454e+11
R-Squared:      0.87201
Adj. R-Squared: 0.87152
Chisq: 1386.18 on 1 DF, p-value: < 2.22e-16
```

Since p value is smaller than the significance level H0 rejected. Therefore, we can conclude that there’s a relationship between population and labor force.

2. Hypothesis 2

H0: The effect of population on labor force is not affected by unemployment. H1: The effect of population on labor force is affected by unemployment.

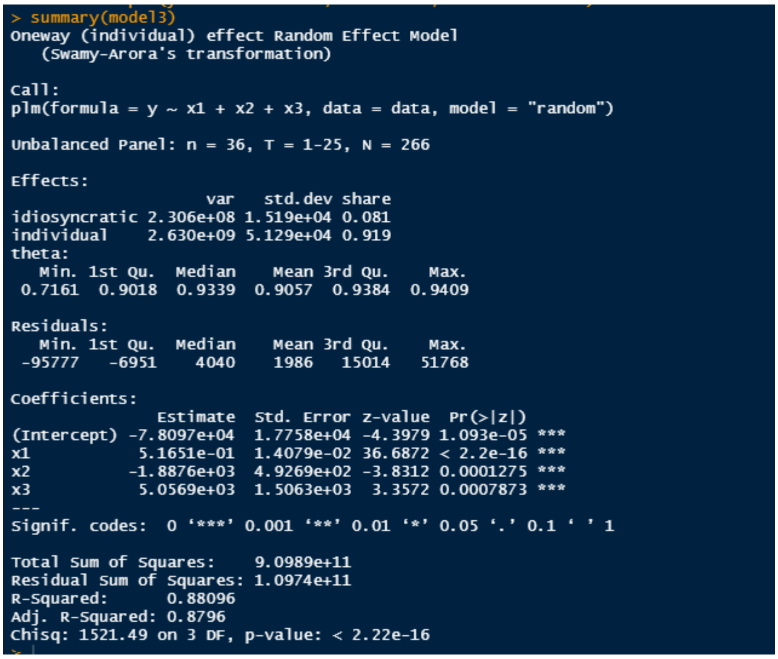


```
> summary(model2)
Oneway (individual) effect Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = y ~ x1 + x2, data = data, model = "random")

Unbalanced Panel: n = 36, T = 1-25, N = 266

Effects:
                    var   std.dev share
idiosyncratic 2.305e+08 1.518e+04 0.085
individual    2.487e+09 4.987e+04 0.915
theta:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.7087  0.8990  0.9321  0.9031  0.9366  0.9392

Residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-104404   -7743    4999    2108   14533   55321

Coefficients:
               Estimate  Std. Error z-value  Pr(>|z|)
(Intercept) -6.1292e+04  1.7409e+04 -3.5206 0.0004305 ***
x1           5.2500e-01  1.3745e-02 38.1951 < 2.2e-16 ***
x2          -1.8925e+03  5.0564e+02 -3.7427 0.0001820 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    9.3264e+11
Residual Sum of Squares: 1.1624e+11
R-Squared:      0.8771
Adj. R-Squared: 0.87616
Chisq: 1459.24 on 2 DF, p-value: < 2.22e-16
```

Since p value is smaller than the significance level H0 rejected. Therefore, we can conclude that the effect of population on labor force is affected by unemployment.

3. Hypothesis 3

H0: The effect of population on labor force is not affected by government expenditure on education.
H1: The effect of population on labor force is affected by government expenditure on education.



```
> summary(model3)
Oneway (individual) effect Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = y ~ x1 + x2 + x3, data = data, model = "random")

Unbalanced Panel: n = 36, T = 1-25, N = 266

Effects:
                   var   std.dev share
idiosyncratic 2.306e+08 1.519e+04 0.081
individual    2.630e+09 5.129e+04 0.919
theta:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.7161  0.9018  0.9339  0.9057  0.9384  0.9409

Residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -95777   -6951    4040    1986   15014   51768

Coefficients:
               Estimate  Std. Error z-value  Pr(>|z|)
(Intercept) -7.8097e+04  1.7758e+04 -4.3979 1.093e-05 ***
x1           5.1651e-01  1.4079e-02 36.6872 < 2.2e-16 ***
x2          -1.8876e+03  4.9269e+02 -3.8312 0.0001275 ***
x3           5.0569e+03  1.5063e+03  3.3572 0.0007873 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    9.0989e+11
Residual Sum of Squares: 1.0974e+11
R-Squared:      0.88096
Adj. R-Squared: 0.8796
Chisq: 1521.49 on 3 DF, p-value: < 2.22e-16
```

Since p value is smaller than the significance level H0 rejected. Therefore, we can conclude that the effect of population on labor force is affected by government expenditure on education.

5. Conclusion

The relationship between population and labor force was studied using the random effect panel regression method. The results of this analysis showed that there was a significant 0.517 between population and labor force, indicating that an increase in population is associated with an increase in labor force.

In addition, the results showed that unemployment had a significant -1890 effect on the relationship between population and labor force. Furthermore, government expenditure on education was found to have a significant 5060 effect on the relationship between population and labor force.

Overall, the random effect panel regression method provides causal estimates of the relationship between population and labor force. Therefore, the results of this study provide evidence for the causal effect of population, unemployment, and government expenditure on education on the labor force.