Unemployment Percentage Rate Prediction in 2023 by Province in Indonesia Using the Panel Data Regression Method

Catherine Benedicta
Statistics Department
School of Computer Science
Bina Nusantara University

Sekar Azalea
Statistics Department
School of Computer Science
Bina Nusantara University

Diana Petrina
Statistics Department
School of Computer Science
Bina Nusantara University

Margaretha Ohyver
Statistics Department
School of Computer Science
Bina Nusantara University

***Abstract* - Unemployment has always been a serious national issue for all countries. Many factors can cause this issue. Our research aims to determine which factors have a high correlation with the unemployment rate. The Panel Data Regression with Least Square Dummy Variable method is selected to explain the effect of each individual unit of cross section which is unobserved but correctly specifies the model of relation. The estimation of Unemployment Percentage Rate could then be used accordingly as tools for many analyses for the related topic and consideration when making regulations in order to solve the unemployment issue.**

Keywords - Unemployment, Panel Data Regression, Least Square Dummy Variable

I. INTRODUCTION

In this modern era, technologies are rapidly growing from time to time. With this phenomenon, the employment requirements are also getting complex. A population with a large number could also boost unemployment in some regions. According to the Department of Labor, Population Resettlement, and Cooperatives (DISNAKER), Indonesia's population growth rate is not comparable with the job opportunities that exist.[1] Considering the fact that unemployment in Indonesia is growing from time to time, it still becomes a major problem that must be solved, especially for the government. According to Sukirno 2002, this unemployment condition will affect the people's prosperity in certain areas which can lead to another and yet, a bigger problem which is poverty.[2]

According to the Oxford Dictionary, the word 'unemployment' is defined as the state of being unemployed.[3] With this word, it defines the jobless people, actively seeking work, and are available to take a job. And according to BPS 2017, there are categories of unemployment, and one of them is 'open unemployment'.[4] The people who are considered included as open employment are the ones who are not looking for a job and not working, but are considered as a human resource.

A little fact about Indonesia's unemployment rate, based on Trading

Economics' research, the unemployment rate in Indonesia in the first quarter of 2022 has fallen to 5.83%. Although the figure has decreased by about 4% from the first quarter of last year, Indonesia is still ranked 30th out of 49 countries in Asia and 13th out of 24 G20 countries for the country with the lowest unemployment rate.[5]

Of course, after knowing the facts we can conclude that unemployment can bring numerous negative impacts in plenty of aspects. It can lead to an increase in poverty and crime rates, many people are depressed, political instability due to many union demonstrations, and an increase in commercial sex workers due to demands to earn a living.

Then it makes us wonder, what are the factors that can lead to this, especially in Indonesia itself. Generally, the lack of job opportunities will become the number one factor in unemployment. We dig some references related to this topic, and we find the fact of the unemployment factors are :[6]

1. Lack of job fields to accommodate job seekers.
2. Human resources that lack needed skills.
3. Lack of connections and information about work.
4. The lazy culture that still plagues job seekers makes it easy for them to give up looking for opportunities.

We can analyze these points on what is related to recent conditions. The fact that now, the technologies are growing rapidly, logically the job opportunities should also grow. Data analysts, Digital Businesses, and Digital Marketing are examples of this job field. Then, why is unemployment still growing each day? What are the factors that significantly affected recent unemployment? And what actions should be done?

In our hypothesis to overcome the growing unemployment rate, a projection or description is needed in analyzing the factors that have the most significant influence on the percentage level of unemployment.

In this paper, we look forward to finding the best regression model to predict the rate of unemployment in the year 2023 with the data of year 2019, 2020 and 2021. We use the panel data regression method to find the predictive model. Cited from web statistikian.com, the regression data panel is defined as a combination method of cross-section and time-series data, where every unit of a cross-section is measured at a different time[7]. So in other words, panel data is data from the same individuals who are observed over a certain period of time.

In conclusion, we intend to develop a regression model using the regression data panel that allows us to predict the factors that significantly affect the unemployment rate. These predictions might be useful for the government or certain people as a projection of the unemployment topic or as tools for many analyses for the related topic. We hope this research could provide a good and accurate estimation, as well as help to determine which factor has a significant effect based on our model.

II. METHODOLOGY

In this section, we will explain the regression method that we use to find a model that fits our dataset, test the residual assumptions that must be met, and compare the 3 regression effect models as well as an explanation of the dataset that we use as data for analysis and exploration in research.

A. Panel Data Regression Method

The regression data panel is defined as the combination method of cross-section and time-series data, where every unit of a cross-section is measured at a different time.

a. Balanced and Unbalanced

We use **balanced panel** data if the cross-sectional unit has the same amount of time-series data. Otherwise, we use **unbalanced panel** data if the cross-sectional unit does not have the same amount of time-series data.

b. 3 Effect Model (Common, Fixed, and Random)[8]

Common Effect Model. This model is also called Pooled Least Square. Among all of the three effect models, this model is considered the simplest approach. The assumption contained in the regression result in this model is that there is no difference between the value of intercepts and slopes. In other words, the regression coefficient is constant for both individual and time. The OLS method is used for this model with the equation:

$$y_{it} = \alpha + \beta' X_{it} + \varepsilon_{it}$$

For:

$$i = 1, 2, \ldots, N$$
$$t = 1, 2, \ldots, T$$

Where:

N = Number of individuals or cross section.

T = Number of time periods.

From this model NxT can be generated an equation that is equal to T equation of cross section and as much N equation of coherent time or time series.

Fixed Effect Model. This model, which is also called the Least Squares Dummy Variable (LDSV), uses dummy variable technique to estimate the intercept differences between individuals. This model assumes that different intercepts can accommodate the differences between individuals (cross section). The equation of fixed effect model is shown by the following formula:

$$y_{it} = \alpha_i + \beta' X_{it} + \varepsilon_{it}$$

For:

$$i = 1, 2, \ldots, N$$
$$t = 1, 2, \ldots, T$$

Where:

N = Number of individuals or cross section

T = Number of time periods

Random Effect Model or Error Component Model (ECM) or Generalized Least Square (GLS) is a model that will estimate panel data where interference variables may be interconnected between time and between individuals. Unlike the others, this model uses a different method, such as maximum likelihood or general least square. The assumption that may be contained in this model is there may be a difference of intercept for each individual and the intercept is a random variable, if the residuals are interconnected between time and between individuals or cross sections. So in this model there are two residual components. The first is the residual as a whole where the residual is a combination of cross section and time series. The second residual is an individual residual which is a random characteristic of the i-th unit observation and remains at all times. The regression equation of panel data of random effects model is as follows:

$y_{it} = \alpha + \beta' X_{it} + u_i + \varepsilon_{it}$

For:

$i = 1, 2, ..., N$

$t = 1, 2, ..., T$

Where:

N = Number of individuals or cross section

T = Number of time periods

ε_{it} = Residual as a whole where the residual is a combination of cross section and time series.

u_i = Individual residual which is the random characteristic of unit observation of the i-th and remains at all times.

B. Least Square Dummy Variable

In panel data model, the purpose of least square dummy variables is to explain the effect of each individual unit of cross section which is unobserved but correctly specifies the model of relation.

C. Residual Assumption

To obtain an unbiased model, we need to test our regression model with the residual assumption test. This test is one of the terms or conditions that must be met to use a linear regression model. Assumptions that must be fulfilled are :

a. The error must have a normal distribution
b. The error must have a constant variance (*homoscedastic*)
c. The error must be an independent variable (non-autocorrelated)

D. Multicollinearity

Multicollinearity happens when independent variables in the regression model are highly correlated to each other. It makes it hard to interpret the model and also creates an overfitting problem. It is a common assumption that people test before selecting the variables into the regression model.

To detect multicollinearity, we need to do measurement of VIF (Variable Inflation Factors). The formula of VIF,

$VIF_j = \frac{1}{1-R_j^2}, j = 1, 2, ..., p$, with

p = number(s) of predictor variable(s)

R_j^2 = square of multiple correlation coefficient

Note that if the result of VIF_j is higher than 10, it can be concluded that the model has a multicollinearity problem, so we need to treat the problem[9].

E. Chow Test

Chow test is a test to determine the model we are going to choose, whether Common Effect or Fixed Effect[8].

The hypothesis for this test is:

H_0= Common Effect

H_1= Fixed Effect

To run the test, we need to define what are the common and fixed effect methods from our panel data. Then, we run the pooltest(). If the p-value result is less than our alpha, we reject the null hypothesis. And if we reject the null hypothesis, we need to continue to do a hausman test.

F. Hausman Test

Hausman test is a test to determine the model we are going to choose, whether Random Effect or Fixed Effect[8]. The hypothesis for this test is:

H_0= Random Effect

H_1= Fixed Effect

To run the test, we need to define what are the random and fixed effect method from our panel data. Then, we run the phtest(). If the p-value result is less than our alpha, we reject the null hypothesis.

G. Weighted Least Square

The purpose of the Weighted Least Squares (WLS) method is to overcome model regression with non-constant error variance (heteroscedastic)[10]. This purpose will be achieved because WLS has the ability to neutralize the effects of violation of the assumption of heteroscedasticity and can eliminate the unusualness and consistency of the OLS. WLS can be obtained by minimizing

$$\sum_{i=1}^{n} \omega_i (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2,$$

Where ω_i are the weights inversely proportional to the variances of the residuals, x_i are independent variables, y_i are dependent variable, and β_p are the parameters.

H. Dataset

In this paper, our dataset is collected from 7 datasets that are obtained from BPS (Statistics Indonesia). The datasets are :

- “Produk Domestik Regional Bruto” will be used as one of the predictor variables, with the name “pdb”. It defines Indonesia's Gross Regional Domestic Product in units of million rupiahs.
- “Jumlah Penduduk Menurut Provinsi di Indonesia” will be used as one of the predictor variables, with the name “jumlahPenduduk”. It defines the total population in every province in Indonesia.

- “Persentase anak usia 10-17 tahun yang bekerja menurut provinsi” that will be used as one of the predictor variables, with the name “kerja”. It defines the percentage of 10-17 years old kids that are already working in every province in Indonesia.
- “Proporsi Individu Yang Menggunakan Telepon Genggam” that will be used as one of the predictor variables, with the name “HP”. It defines the percentage of people that already have a handphone in every province in Indonesia.
- “Tingkat Penyelesaian Pendidikan Menurut Jenjang Pendidikan dan Provinsi” that will be used as one of the predictor variables, with the name “pendidikan”. It defines the percentage of finished education (We only use High School level) in every province Indonesia has.
- “Upah Rata - Rata Per Jam Pekerja Menurut Provinsi” that will be used as one of the predictor variables, with the name “upah”. It defines the average hourly wages of workers in every province in Indonesia.
- “Tingkat Pengangguran Terbuka Menurut Provinsi” that will be used as our response variable, with the name “pengangguran”. It defines the percentage of open unemployment in every province in Indonesia.

With these datasets, we will combine all of the data to become one-panel data. Then, we estimate the prediction using the panel data.

III. RESULT AND DISCUSSION

Unemployment has always been one of the main problems in many countries, including Indonesia. Our research is meant to predict the unemployment rate in Indonesia for each province and its predictor variable so that the government will have the data to take into consideration while making a decision to decrease the unemployment rate. There are 7 predictor variables that are used in this research and in order to predict the unemployment rate, we chose the Panel Regression.

First, we will check the multicollinearity of the predictor variables using VIF. If any of the values is greater than 10, we need to treat it.

TABLE I. MULTICOLLINEARITY

Variable	VIF
pendidikan	1.990791
jumlahPenduduk	3.887872
HP	1.931854
kerja	1.230272
pdb	4.011150
upah	1.333669

From Table I, we know that our data have no multicollinearity, so we can continue to find the most appropriate effect model using

Chow Test and Hausman Test. Let α, the significant value is equal to 0.05. For both test, if the p-value less than α, then the fixed effect will be selected as the better model.

First, we need to do for the test is to use all of our predictor variables. Table II below states all the p-value results from the test using all of our predictors.

TABLE II. CHOW AND HAUSMAN TEST USING ALL PREDICTOR VARIABLES

Test	P-Value
Chow	4.996e-11
Hausman	1.309e-05

From the table II above, we can conclude that both test results reject the null hypothesis. So, we can conclude that the fixed effect model is appropriate for this case.

The next thing we need to do is to check the significance of all predictor variables to our response variable using the F Test and T test.

TABLE III. SIGNIFICANCE TEST USING ALL PREDICTOR VARIABLES

Test	**P-Value**	
F-test	3.8772e-07	
T-test	**Variable**	**P-Value**
	pendidikan	0.0001878
	jumlahPenduduk	0.3682719
	HP	0.1094832
	kerja	0.4915448
	pdb	0.0017995
	upah	0.5102595

From Table III, the result from our F-Test indicates that there are, at least one predictor variable that has a significant relationship with our response variable. With T-Test, it is confirmed that there are two predictor variables that have a significant relationship with the response variable, which are "pendidikan" and "pdb".

After knowing that "pendidikan" and "pdb" variables have the most significant relationship with the response variable, we do the Chow Test and Hausman Test to remodel but only use the significant predictor variables to get the best model.

In this step we only use the variables that have a significant relationship with our response variable.

TABLE IV. CHOW AND HAUSMAN TEST USING SIGNIFICANT PREDICTOR VARIABLES

Test	P-Value
Chow	2.2e-16
Hausman	4.514e-05

From the table IV above, we can conclude that both test results reject the null hypothesis. So, we can conclude that the fixed effect model is appropriate for this case.

The same thing that we do in this step is to determine whether our two predictor variables have a significant relationship with the response variable

TABLE V. SIGNIFICANCE TEST

USING SIGNIFICANT PREDICTOR VARIABLES

Test	P-Value	
F-test	1.923e-08	
T-test	**Variable**	**P-Value**
	pendidikan	7.35e-09
	pdb	0.0006001

From Table V, the result from our F-test indicates that there are, at least one predictor variable that has a significant relationship with our response variable. With t-test, it is confirmed that indeed those two variables, “pendidikan” and “pdb” have a significant relationship with the predictor variable.

The next step is to test the residual assumptions based on the model that has been obtained. The table VI below states all the p-value results from the residual assumption test. Let α, the significant value is equal to 0.05. For all test, if the p-value greater than α, then the assumption is not violated.

TABLE VI. THE RESIDUAL ASSUMPTION TEST

Test	Result
Kolmogorov-Smirnov	0.3268
Breusch-Pagan	0.01051
Panel Durbin-Watson	0.9993

From table VI above, we can conclude that only the Breusch-Pagan has p-value less than α (0.05). Thus it means that the residual does not have constant variance (heteroscedastic). In this case, what we need to do to treat the violation is to use the weighted least square regression.

Weighted Least Squares (WLS) can be used to overcome heteroscedasticity. To do this method, we need to find the fitted values from our last model, and then we will create the Weighted Least Square Model estimated with significant predictor variables as stated below.

$$\hat{Y} = 0.13732X_{pendidikan} - 0.000011995X_{pdb}$$

Weighted Least Square Summary

R-squared	Adj R-squared	F-statistics
0.41391	0.1031	22.9565

Because our purpose is also to make an estimation model for panel data, we can do it with the Least Square Dummy Variable method to explain the effect of each individual unit of cross section which is unobserved but correctly specifies the model of relation.

LSDV Summary

R-squared	Adj R-squared	F-Statistics
0.9929	0.9887	235.4

By comparing the R-Square value of WLS and LSDV, we know that the R-squared of the LSDV model has a higher value than the WLS’s R-squared. Thus, we will use LSDV to build the model.

Thus, the estimation model will become:

$$\hat{Y} = \lambda_t + \mu_i + 0.01795X_{pendidikan} - 6.09 \times 10^{-6}X_{pdb}$$

Where:

λ_t: Intercept for the time section

μ_i: Intercept for each province.

The intercept value for the time section will be shown in Table VII, while the intercept for the province will be shown in Table VIII.

TABLE VII. INTERCEPT FOR TIME

Year	λ_t
2019	5.450
2020	6.624
2021	6.278

TABLE VIII. INTERCEPT FOR EACH PROVINCE

Province	μ_i
Bali	-1.803
Banten	5.967
Bengkulu	-3.177
D.I. Yogyakarta	-2.715
DKI Jakarta	1.828
Gorontalo	-3.314
Jambi	-1.131
Jawa Barat	15.24
Jawa Tengah	6.901
Jawa Timur	12.31
Kalimantan Barat	-40.58
Kalimantan Selatan	-1.487
Kalimantan Tengah	-1.803
Kalimantan Timur	3.139
Kalimantan Utara	-1.945
Kep. Bangka Belitung	-2.060
Kep. Riau	3.329
Lampung	-0.4892
Maluku	-0.0032
Maluku Utara	-2.107
Nusa Tenggara Barat	-2.914
Nusa Tenggara Timur	-2.563
Papua	-1.690
Papua Barat	-0.273
Riau	2.955
Sulawesi Barat	-3.647
Sulawesi Selatan	1.411
Sulawesi Tengah	-2.317
Sulawesi Tenggara	-2.517
Sulawesi Utara	0.2622
Sumatera Barat	0.465
Sumatera Selatan	0.593
Sumatera Utara	3.862

IV. CONCLUSION

In this paper, we intend to develop a regression model using the Data Panel

Regression that allows us to predict the factors that significantly affect the unemployment rate. The rate of unemployment for a particular year and province could be estimated by developing a regression model by using Least Square Dummy Variable method. And with our research, we know that the factors that have a significant relationship with the rate of unemployment are the level of education and the gross regional domestic product. In conclusion, we hope that the government or certain parties can come up with a solution to help the community in dealing with their education and create a project that can help the national economy in order to lower the unemployment rate in Indonesia.

REFERENCES

[1] Masalah Tenaga Kerja dan Angkatan Kerja di Indonesia | Dinas Tenaga Kerja. https://disnaker.bulelengkab.go.id/informasi/detail/artikel/masalah-tenaga-kerja-dan-angkatan-kerja-di-indonesia-56

[2] Sukirno. Makroekonomi Teori Pengantar Edition: 3 (2002).

[3] Unemployment Definition | Oxford Dictionary. https://www.oxfordlearnersdictionaries.com/definition/english/unemployment

[4] Tenaga Kerja | Badan Pusat Statistik. https://www.bps.go.id/subject/6/tenaga-kerja.html

[5] Unemployment Rate | Trading Economics. https://tradingeconomics.com/country-list/unemployment-rate

[6] Franita, Riska. Analisis Pengangguran Di Indonesia. Universitas Muhammadiyah Tapanuli Selatan (2016).

[7] Penjelasan Metode Analisis Regresi Data Panel | Statistikian. https://www.statistikian.com/2014/11/regresi-data-panel.html

[8] Zulkifar, Rizka. Estimation Model and Selection Method of Panel Data Regression. Universitas Islam Makassar.

[9] Chartterjee, S. & Hadi, A. S. Regression analysis by example. Journal of Applied Statistics vol. 40 (2013).

[10] Hanifah, Nurul. Penerapan Metode Weighted Least Square Untuk Mengatasi Heteroskedastisitas Pada Analisis Regresi Linear. Universitas Pendidikan Indonesia (2016).