10.22441/ihasj.2021.v4i3.05

COMPARISON OF IMBALANCED DATA METHODS ON LOGISTICS REGRESSION
(Case Study: Poverty in Indonesia In 2018)

Pardomuan Robinson Sihombing
BPS- Statistics Indonesia, Jalan dr. Sutomo No. 6-8 Jakarta
robinson@bps.go.id

Abstract - Poverty is still one of the main problems in economic development and inequality, unemployment, and economic growth. This study aims to model poverty directly by using a discrete choice model using binomial regression. The data used is imbalanced data, where one of the value categories is relatively small. In this study, the logistic regression method applies several resample techniques. They include undersampling, oversampling, a combination of both, and Cost-Sensitive Learning (CSL). The results obtained that both sampling techniques provide optimal results when viewed from the indicators of accuracy, specificity, sensitivity, and AUC. In addition, the results show that in households in rural areas, the head of the household is female, unmarried, has low education, married at an early/old age, and has a large household size, has a greater chance of being poor than other categories. So that targeted and comprehensive policy is needed so that the poverty rate can continue to be reduced and welfare increases.

Keywords: imbalanced; logistics; poverty; resample

INTRODUCTION

Poverty is still one of the main problems in economic development and inequality, unemployment, and economic growth. One of the pillars of the Sustainable Development Goals (SDGs) in the field of social development is the achievement of the fulfillment of quality fundamental human rights in a fair and equal manner to improve welfare for all people, with the main priority being ending poverty in all its forms everywhere). Indonesia's poverty percentage data itself shows a declining trend from year to year, as shown in Figure 1, where the 2019 poverty rate based on the March 2019 Susenas data was 9.41 percent.

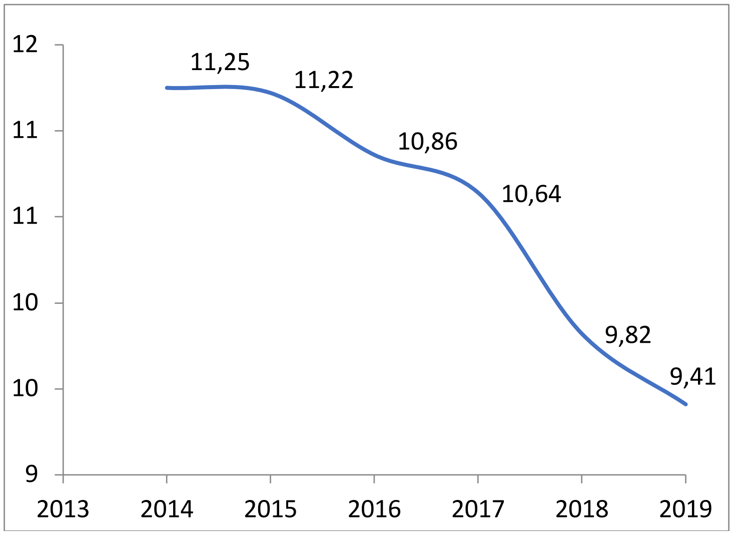

Figure 1 Percentage of Indonesia's Poor 2014-2019

There are two approaches to modeling the factors that influence poverty. The first approach uses a regression approach between consumption expenditure per adult equivalent to several potential explanatory variables called the consumption approach. In the second approach, we can model poverty directly using a discrete choice model. The discrete approach in question is to categorize poverty into two categories based on household consumption expenditure compared to the poverty line of a region (Fissuh & Harris, 2005). According to Han and Kamber (2012), classification is the process of finding a model or function that can describe and distinguish classes of data or concepts. The aim is to predict the unknown class of an object of observation. Several classification methods are commonly used, including Classification and Regression Trees (CART), Naïve Bayes, Random Forest, Rotation Forest, K-Nearest Neighbor (KNN), Artificial Neural Networks (ANN), Support Vector Machine (SVM), discriminant analysis, and logistic regression.

Logistic regression can be a robust classifier by providing classification opportunities and covering multi-class classification problems. However, unfortunately, this method is based on the assumption that the amount of data is evenly distributed between different classes. Whereas in real life, there are events that show that the amount of data is not balanced between different classes, and when the imbalance condition is called imbalanced data (Maalouf & Trafalis, 2011). In particular, for independent binary observations, Czado and Santner (1992) show that incorrectly assuming a logistic link leads to a substantial increase in the bias and mean squared error of the parameter estimates and the predicted probabilities asymptotically and in a finite sample. King and Zeng (2001) tate that when logistic regression is used in imbalanced data, the classification tends to eliminate opportunities from the minority class because the predicted value will tend to be in the majority class so that the accuracy of the resulting classification is not good. Especially the data used is data in enormous amounts (big data).

Therefore, in this study, we will examine and apply the logistic regression method by paying attention to imbalanced data and large data sets using data division with a deterministic method (holdout) and stratified 10-fold cross-validation (CV) on the dataset. For resample method, this research compares four resample methods. They include Undersampling, oversampling, a combination of both, and Cost-Sensitive Learning (CSL) in modeling the classification of poor status in Indonesia.

METHODS

Data Sources and Research Variables

The data sources used in this study came from the March 2018 Susenas data. The research variables used were:

Table 1. Research Variable

Variable	Name	Information
Y	Poor Status	0 Not Poor 1 Poor
X1	Area Type	0 Urban 1 Rural
X2	Marital status	0 Not Married and Divorced 1 Married
X3	Gender of household	0 female 1 Male
X4	Education	0 No School 1 Elementary and Junior School 2 senior school 3 Universities
X5	Business field	0 Primary 1 Secondary 2 Tertiary
X6	Number of household members	
X7	Age	

Logistics Regression

Logistic regression is a regression model that shows the effect of predictor variables, either continuous or categorical, on response variables in the form of categorical data. In binary logistic regression, the response variable consists of two (binary) categories, namely 0 and 1. Where the response variable for each observation follows the Bernoulli distribution and with Logistics Regression model as follows:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \quad (1)$$

Partial parameter testing to determine whether a predictor variable significant effect on the response variable. The hypothesis used in this test:

$H_0 : \beta_j = 0$
$H_1 : \beta_j \neq 0$ where j=1, 2, …, p
with wald test statistics

$$W = \frac{\beta_0}{se\,(\beta_j)} \sim N(0,1) \quad (2)$$

The decision to reject H0 if the test statistic value |W | > Z means that the predictor variable significantly affects the response variable.

Imbalanced Data Method

The method used is widely known as the 'Sampling Method.' In general, this method aims to convert unbalanced data into a balanced distribution using several mechanisms. Modification occurs by resizing the original data set and providing the same balance proportions. The first method is undersampling. This method works with the majority class, i.e., reduces the number of observations from the majority class to create a balanced data set. The second method is oversampling, where this method works with minority classes. The method used is to replicate observations from the minority class to balance the data. The third method is In simple words, instead of replicating and adding observations from the minority class, it overcomes the imbalance by generating artificial data. The third method is the method of both/combine sampling from the oversampling and undersampling methods. The fourth method uses Cost-Sensitive Learning (CSL), where this method considers the cost of misclassification of False Positive and False Negative.

Modeling

At the modeling stage, a model will be formed that can distinguish data classes. To do the At this stage, there are two types of data sets, namely training data and testing data. Training data is a dataset that is used to build a model. In contrast, testing data is used to calculate the model's performance formed by comparing the actual data labels and the data labels from the model classification results. This research uses two methods to form training data and to test research data, namely K-Fold Cross-Validation and deterministic/holdout.

a. K-Fold Cross-Validation (K-Fold CV)

Validation of data testing is carried out on datasets that are not used in model formation. Cross-Validation is a form of resampling that takes several samples from all observations and turns them into training data for the model (Nisbet, Robert, Elder, & Miner, 2009)

b. Deterministic/ Holdout

The division of the data set into training data and testing data can use a deterministic/holdout method, namely by determining the distribution ratio of the two datasets yourself.

For example, the dataset ratio can use 8:2, meaning that 0.8 of the total data for training data and the remaining 0.2 for testing data can produce performances.

Evaluation

Evaluation is carried out to choose which method of dividing the data set, and classification can produce accuracy. Evaluation in this research is by paying attention to the Confusion Matrix. Confusion Matrix is a tool to determine the extent to which classifiers can recognize or predict data classes. Confusion Matrix is a table measuring mxm with m=number of classes (Han, Jiawei, Kamber, & Pei, 2012). The column section is filled by the actual label for each class, while the predicted class label fills the row section.

Table 2. Confusion Matrix

Confusion Matrix		*Actual Class*		*Total*
		Yes	*No*	
Predicted Class	*Yes*	*TP*	*FP*	*P'*
	No	*FN*	*TN*	*N'*
Total		*P*	*N*	

In general, classification accuracy is used to measure accuracy, namely the proportion of the correct frequency classified with the total sample. In addition to seeing the accuracy, we can see the sensitivity. Sensitivity + is the proportion of the class of concern/want that is correctly predicted. Specificity - is the proportion of classes that are not of concern/unwanted to be correctly predicted. If the accuracy rate is high, however sensitivity and specificity are low, then the classification can be said to be not good (Han, Kamber, & Pei, 2012)

$$Akurasi = \frac{TP+TN}{P+N} \tag{3}$$

$$Sensitivity{+}= \frac{TP}{P} \tag{4}$$

$$Specificity{-}= \frac{TN}{N} \tag{5}$$

Another classification performance evaluation measure is the Receiver Operating Characteristic (ROC) curve. ROC curve is an analysis curve that describes the performance of a classification model on two dimensions between sensitivity as the y-axis and (1-specificity) as the x-axis (Fawcett, 2006). The single value can measure classification performance on the ROC curve area Under Curve the ROC (AUC). Table 3 shows the standard classification categories based on the AUC value (Gorunescu F. , 2011).

Table 3. Classification Category Based on AUC

AUC value	Classification Category
0.90 – 1.00	*Excellent classification*
0.80 - 0.90	*Good classification*
0.70 – 0.80	*Fair classification*
0.60 – 0.70	*Poor classification*
0.50 – 0.60	*Failure*

RESULTS AND DISCUSSION

Descriptive Analysis

Before conducting further analysis of the relationship between variables, a descriptive analysis of the research variables was conducted. There are 10.79 percent of Indonesia's population who are still below the poverty line. Differences in facilities and infrastructure in urban and rural areas can affect the level of welfare of a household. As many as 8.50 percent of Indonesia's poor live in rural areas. When viewed from the marital status, it is dominated by residents who have a partner by 9.49 percent. Meanwhile, when viewed from gender to head of household, the poor are dominated by household headed by a woman.

Primary business fields (agriculture and mining) still dominate in Indonesia. This sector also dominates the poor in Indonesia at 7.45 percent. In terms of education, Indonesia is quite good where as much as 9.42 percent of the population has completed tertiary education. The poor population in Indonesia is still dominated by households whose household heads have elementary and junior high school education.

Table 4a. Descriptive Analysis of Research Variables

Poverty Status	Area Type		Marital status		Gender		Total
	rural	Urban	Single and Divorce	Marry	Male	Female	
Not Poor	50.63	38.58	13.37	75.84	79.24	9.97	89.21
Poor	8.50	2.29	1.30	9.49	9.61	1.18	10.79
Total	59.13	40.87	14.67	85.33	88.85	11.15	100.00

Table 5b. Descriptive Analysis of Research Variables

Poverty Status	Business Sector			Highest Education			
	Primary	Secondary	Tertiary	No school	Elementary and Junior School	Senior high school	University
Not Poor	37.70	15.00	36.50	3.95	52.37	23.64	9.25
Poor	7.45	1.39	1.95	1.17	7.97	1.48	0.18
Total	45.16	16.39	38.45	5.12	60.34	25.12	9.42

Model Selection

Results Classification of models on imbalanced data without any treatment using the Logistics Regression model generally results in higher average accuracy and sensitivity values compared to data that has been treated to overcome imbalances with undersampling, oversampling, and combine / both sampling schemes. andCost Sensitive Learningboth with holdout (80:20) and k-fold data sharing techniques. However, on the other hand, the classification results without treatment resulted in lower specificity and pseudo R square than the four treatment models with both data sharing models.

When viewed from the AUC Value, the values are relatively the same with the scheme without treatment or with the four treatment techniques, ranging from 77.10% to 77.49%. The highest AUC value is found in the holdout data division technique (80:20). The highest average accuracy and sensitivity values that imbalanced data can produce without treatment are 89.371 percent and 99.30 percent, respectively, with the holdout data sharing technique, and 89.29 percent and 99.35 percent with the K-fold technique. While the average specificity value produced is only 6.15 percent with the holdout data sharing technique and 6.20 percent with the K-fold technique. If the value is small enough, the type of imbalanced data will have a shallow level of accuracy in classifying the few categories, in this case, the poor households, into the poor category. This misclassification will affect government policies in terms of assistance provided to the poor so that the assistance provided is not on target.

With statistical techniques to overcome imbalanced data, the resulting model increases the pseudo R square value from 17.68 percent to 18.12 percent. On the other hand, the average specificity value increased, ranging from 70.48 percent to 70.96 percent. In addition, there was a decrease in sensitivity to range from 69.12 percent to 69.89 percent. In other words, the handling of imbalanced data makes the average value of specificity and sensitivity more balanced, resulting in a lower average overall accuracy value, ranging from 69.86 percent to 70.31 percent, compared to the Classification with imbalanced data without treatment. On the other hand, the classification error becomes more balanced.

Table 6. Average Logistic Regression Model Classification Performance

Resampling	Share Data	Pseudo R Square	AIC	Accuracy	Specificity	Sensitivity	AUC
No Treatment	holdout	0.1420	122108	0.8937	0.0615	0.9930	0.7748
	aggregate cfold	0.1426	137001	0.8929	0.0620	0.9935	0.7723
Under Sampling	holdout	0.1805	50917	0.7000	0.7064	0.6935	0.7709
	aggregate cfold	0.1795	57340	0.7019	0.7048	0.6989	0.7729
Over Sampling	holdout	0.1810	420348	0.7019	0.7057	0.6981	0.7749
	aggregate cfold	0.1812	472722	0.7015	0.7051	0.6979	0.7740
Both/Combine Sampling	holdout	0.1792	236130	0.7031	0.7096	0.6966	0.7745
	aggregate cfold	0.1796	265514	0.7015	0.7077	0.6952	0.7730
Cost Sensitive Learning	holdout	0.1768	236817	0.7007	0.7083	0.6929	0.7725
	aggregate cfold	0.1772	266284	0.6986	0.7061	0.6912	0.7710

By considering the average of various criteria, namely the pseudo-R square value, specificity value, sensitivity value, and AUC, the best Logistics Regression model is a combined/both sampling scheme with a holdout technique. If it looks at the more detailed classification results for each fold, as presented in the table, the pseudo r square value is 17.92 percent, accuracy is 70.31 percent, sensitivity is 70.96 percent, specificity is 69.66 percent, and AUC is 77.45 percent with the following equation:

$g(x) = \ln\left[\frac{P(Y=1|x)}{P(Y=0|x)}\right]$ = -0.028-0.478Urban-0.269Married-0.081Male-0.748SD_SMP-1.449SMA-2.403PT-0.472Secondary-0.798Tertiary+0.464number_art-0.008age

Table 7. Hypothesis Test Results Table

Variable	Category	Coefficients:	P. Value
	(Intercept)	-0.028	0.410
Area Type	Urban	-0.478	0.000
Marital status	Married	-0.269	0.000
Gender	Male	-0.081	0.000
Education	Elementary and Junior School	-0.748	0.000
	Senior high school	-1,449	0.000
	University	-2.403	0.000
Business field	Secondary	-0.472	0.000
	Tertiary	-0.798	0.000
Number of household members	Number of household members	0.464	0.000
Age	Age	-0.008	0.000

From the results above, it can be seen that all variables and their categories significantly affect a person's poor status with a coefficient of determination of 17.92 percent, which means that variations from a person's poor status can be explained by explanatory variables and their categories of 17.92 percent. From the accuracy value of 70.96 percent, it is greater than the cut-off of 50 percent, which means that the model can accurately predict the inferior status of households by 6 percent. Meanwhile, the ROC curve and the AUC value of 77.45 have shown promising results according to the AUC criteria (Gorunescu F. , 2011)

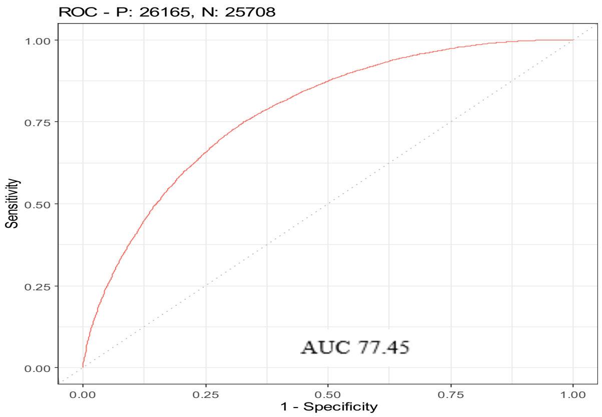

Figure 2. ROC and AUC

Interpretation/ Discussion

The location of the household is often associated with the poverty status of a household. This result can be attributed to differences in access to primary facilities such as education, transportation, health, and the economy. The results of this study indicate the influence of the status of the area of residence on the poverty status of a household with a negative coefficient. This result means that households living in rural areas have a higher chance of becoming poor than households living in urban areas. This result is in line with several previous studies which state that households in rural areas are more vulnerable to being poor due to limited access(Anyanwu, 2014).

Marital status can be linked to responsibility for household expenses. The results showed a significant effect between marital status and poor status with a negative coefficient. This result can be interpreted that someone who has marital status has a lower chance of being poor. This result is due to The income generated becomes cumulative of household members (Anyanwu, 2014).

In general, the head of the household is male. Some household heads can be headed by women if there is a divorce, either divorced or divorced. In some cases, female-headed households tend to have a higher chance of being poor. The results showed a significant effect between the sex of the head of the household with poor status with a negative coefficient. This result means that households that a man heads have a lower chance of being poor. This result is caused because the income generated becomes the cumulative ofBuvinić & Rao Gupta **(1997)** concludes that there are three main reasons why female-headed households are more vulnerable to becoming poor, namely: female-headed households tend to be households with high dependents; women tend to have lower incomes; and the interaction between household structure factors and gender inequality factors in the economy which further exacerbates poverty opportunities.

Education is one of a person's efforts to improve the standard of living. Higher educational attainment will tend to increase the earning potential of individuals, and consequently, increased income will help them out of poverty (Awan, Sarwar, W Muhammad, & M. Waqas, 2011) . This study also shows consistent results where the education of the head of the household has a significant effect with a negative coefficient. This result means that the higher the education of the head of the household, the fewer opportunities for poverty.

Business fields can be grouped into three, namely primary, secondary and tertiary. Primary business fields are directly related to natural processing, such as agriculture, fishery, livestock, and mining. At the same time, the secondary business field concerns the processing of the results of the primary business field, namely industrial activities, electricity, and water. In comparison, the tertiary business field concerns the trade and services business field. The results of this study indicate a significant effect, and the coefficient is negative; this means that families whose heads of household work in the secondary and tertiary sectors have a lower tendency to be poor than those whose main income comes from the primary sector. The results of this study are in line with the results of research conducted by (2015) , which states that the shift from the agricultural sector is effective in alleviating poverty.

Household size indicates the number of people who usually live in the household. The more people living in a household, the more resources are needed to make the household members prosperous. The results of this study indicate that compared to households consisting of only one person, groups of households consisting of 2 or more people have a higher tendency to live in poverty. The results of this study are in line with several studies in developing countries which state that the more the number of household members, the lower the average consumption per capita, which indicates that the household is getting closer to poor status (Anyanwu, 2014).

Age linked to productivity. Someone who is in productive age can work and generate income that can meet the needs of his family. The results showed a significant effect between age and poor status with a negative coefficient. This result means that the more unproductive age, the more inadequate opportunities will increase; this is in line with the results of research conducted by Gonder and Xing (2012) , which shows that after passing the productive age, a person's income will tend to decrease so that the level of vulnerability to being poor will be higher.

CONCLUSIONS AND SUGGESTIONS

From the result above, the conclusions drawn from this research are that the imbalanced data handling model used is both/combine sampling technique with holdout/deterministic data sharing method. Rural areas have a higher probability of being impoverished than urban areas. Married people have a lower probability of being poor than unmarried and divorced people. Households headed by a man have a lower probability of being poor than those headed by a woman. Higher education tends to have a lower chance of being poor compared to lower education Pendidikan. Secondary and tertiary business fields have a lower chance of being poor compared to Primary Business Fields. The larger the size of the household members, the higher the probability of being poor compared to the smaller households. The relationship between poverty and age is U-shaped, where older people have a lower chance of being poor than young and old age.

Future studies can compare the results with other classification techniques such as SVM, random forest, rotation forest, nave Bayes, KNN, Decision Trees, ANN, etc. They are using other social, economic, and cultural variables in exploring the factors that influence poverty. Finally, the subsequent studies can use a new data set.

References

Adhi, E. T. (2009). Pelayanan Sanitasi Buruk Akar Dari Kemiskinan. *Jurnal Analisis Sosial, 14*, 76–88.

Anyanwu, J. C. (2014). Marital Status, Household Size and Poverty in Nigeria: Evidence from the 2009/2010 Survey Data. African Development Review. *26*(1), 118–137. doi:https://doi.org/10.1111/1467-8268.12069

Awan, M., Sarwar, M., W Muhammad, M., & M.Waqas. (2011). Munich Personal RePEc Archive Impact of education on poverty reduction Impact Of Education On Poverty Reduction. *International Journal of Academic Research, 3*.

Buvinić, M., & Gupta, G. R. (1997). Female-headed households and femalemaintained families: Are they worth targeting to reduce poverty in Developing Countries. *Economic Development and Cultural Change, 45*(2), 258–280. doi:https://doi.org/10.1086/452273

Czado, C., & Santner, T. (1992). The effect of link misspecification on binary regression inference. *J. Statist. Plann. Inference 33*, 213–231. MR1190622.

Fahar, F. (2015). Kemiskinan dan Ketenagakerjaan di Kepulauan Riau 2014: Permasalahan Dan Implikasi Kebijakan. *Jurnal Ekonomi Keuangan*.

Fawcett, T. (2006). An Introduction to ROC Analysis. Journal of Pattern Recognition Letters. *An Introduction to ROC Analysis. Journal of Pattern Recognition Letters, 27*, 861-874.

Fissuh, E., & Harris, M. (2005). *Modeling Determinants of Poverty in Eritrea: A New Approach*, 1-35.

Gorunescu, F. (2011). *Data Mining Concept, Models and Techniques.* Verlag Berlin Heidelberg: Springer.

Gounder, R., & Xing, Z. (2012). Impact of education and health on poverty reduction: Monetary and non-monetary evidence from Fiji. *Economic Modeling, 29*(3), 787–794. doi:https://doi.org/10.1016/j.econmod.2012.01.018

Han, Jiawei, Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques 3rd Edition.* Massachusetts: Elsevier Inc.

King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. *Journal of Political Analysis, 9*(2), 137-163.

Maalouf, M., & Trafalis, T. (2011). Rare Events and Imbalanced Datasets: An Overview. *Int. Journal Data Mining, Modelling and Management, 3*(4), 375-385.

Macinnes, T., Tinson, A., Hughes, C., Born, T. B., & Aldridge, H. (2015). Monitoring poverty and social exclusion.

Malgesini, G., Cesarini-Sforza, L., Babović, M., Leemkuil, S., Sverrisdóttir, M., & Mareková, S. (2001). Gender and Poverty in Europe. Development, (Ic). 1-14.

Nisbet, Robert, Elder, J., & Miner, G. (2009). *Handbook of Statistical Analysis and Data.* California: Elsevier Inc.

OECD. (2015). *Pensions at a Glance 2015: OECD and G20 indicators.* Retrieved from https://doi.org/10.1787/pension_glance-2015-en

Rahayu, A. S. (2017). Kehidupan Sosial Ekonomi Single Mother dalam Ranah Domestik dan Publik. *Jurnal Analisa Sosiologi*, 6(1).

Sigle, W. R., & McLanahan, S. (2002). For richer or poorer? Marriage as an anti-poverty strategy in the United States. Population. *57*(3), 509–526. doi:https://doi.org/10.2307/3246637

Sinaga, U., & Siregar, H. (2002). Analisis Determinan Kemiskinan Sebelum dan Sesudah Desentralisasi Fiskal. *Jurnal Sosio Economic of Agriculturure and Agribisnis, 6*(2), 1–17.

Sridhar, K. S. (2015). Is urban poverty more challenging than rural poverty? A review. *Environment and Urbanization Asia, 6*(2), 95–108.

Tilak, J. B. (1999). Education and Poverty in South Asia. Prospect, XXIX(4).

UNDP. (1997). *In Human Development Report.* doi:https://doi.org/10.2307/2524904