The Effects of Literacy Rates on Unemployment Rates in India

Prakriti Bansal

Research Question
What is the impact of increasing literacy rates in India on unemployment rates in states and UTs?

Introduction
With increasing unemployment rates, it is becoming necessary to figure out the underlying problems causing this. In this paper, we are going to start by looking at the literacy rate and unemployment rate trends. We will, further, explore more features like corruption, annual growth rate, population, no. of factories and gender ratio.
We are going to be looking at state/UT data for 2011, to explore these trends. [1]

Hypothesis
For the simple regression analysis, the unemployment rates for each state and UT for 2011 was selected as the dependent variable while the literacy rate for each state for 2011 was selected as the independent variable.
Let the trend between unemployment and literacy rate be given as:

$$unemp = \beta_0 + \beta_1(lit) + \varepsilon$$

Let's start by considering that there is no relation between these two variable and therefore, the OLS coefficients are zero.
Both literacy rate and unemployment rate are represented as percentages i.e. the no. of people literate or unemployed per 100 inhabitants.

For multiple regression analysis, in addition to unemployment, five other independent variables were included: annual growth rate, corruption, population, gender ratio and no. of factories.

$$unemp = \beta_0 + \beta_1(lit) + \beta_2(corr) + \beta_3(growth) + \beta_4(pop) + \beta_5(genr) + \beta_6(fac) + \varepsilon$$

We will start by the null hypothesis that, there is no correlation i.e. all the OLS coefficients are zero.

[1] For the datasets used, please refer look through the /Datasets folder. This contains all the files in .csv format.

Data
The table below is the description of data variables used in this report (a separate file of the description can be found in the /Datasets folder).

Table 1: Statistics Summary of Variables

	Annual Growth	Corruption	Factories	Gender Ratio	Literacy	Population	Unemployment rate
count	34.00	34.00	34.00	34.00	34.00	34.00	34.00
mean	1.71	704063.15	6798.56	927.94	77.45	35561.91	3.65
std	0.90	1299129.12	9144.67	80.48	8.38	44701.26	3.94
min	-0.05	0.00	18.00	618.00	61.80	64.00	0.00
25%	1.30	0.00	839.25	896.75	70.78	1588.75	1.33
50%	1.66	33000.00	2931.00	938.50	76.60	21053.00	2.35
75%	2.02	1280875.00	8059.25	972.50	85.05	60944.25	3.78
max	4.51	6741500.00	36996.00	1084.00	94.00	199581.00	17.70

The assumption that there is a linear relationship between the parameters is satisfied as they can be written as linear equation. The assumption that no two variables are collinear is also satisfied as shown by the correlation matrix. The assumption that the sampling is random is satisfied as data from all of the population was used.

Table 2: Correlation matrix

	Annual Growth	Corruption	Factories	Gender Ratio	Literacy	Population
Annual Growth	1.0000	-0.0869	-0.1464	-0.5965	-0.2558	-0.0595
Corruption	-0.0869	1.0000	0.1196	0.1199	-0.1701	0.2429
Factories	-0.1464	0.1196	1.0000	0.2019	-0.0648	0.4383
Gender Ratio	-0.5965	0.1199	0.2019	1.0000	-0.0366	0.1335
Literacy	-0.2558	-0.1701	-0.0648	-0.0366	1.0000	-0.4531
Population	-0.0595	0.2429	0.4383	0.1335	-0.4531	1.0000

Results

The OLS regression results for simple linear regression and multiple linear regression are shown below.

Figure 1: Results for Simple Linear Regression. The dependent variable is the unemployment rate.

OLS Regression Results

Dep. Variable:	2011-12 - Rural+Urban	R-squared:	0.171
Model:	OLS	Adj. R-squared:	0.145
Method:	Least Squares	F-statistic:	6.581
Date:	Fri, 16 Jun 2017	Prob (F-statistic):	0.0152
Time:	18:18:58	Log-Likelihood:	-91.167
No. Observations:	34	AIC:	186.3
Df Residuals:	32	BIC:	189.4
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Bias	-11.3792	5.893	-1.931	0.062	-23.383	0.624
Literacy	0.1941	0.076	2.565	0.015	0.040	0.348

Omnibus:	27.806	Durbin-Watson:	2.021
Prob(Omnibus):	0.000	Jarque-Bera (JB):	55.944
Skew:	1.963	Prob(JB):	7.11e-13
Kurtosis:	7.907	Cond. No.	735.

Figure 2: Results for Multiple Linear Regression. The dependent variable is the unemployment rate.

OLS Regression Results

Dep. Variable:	2011-12 - Rural+Urban	R-squared:	0.547
Model:	OLS	Adj. R-squared:	0.446
Method:	Least Squares	F-statistic:	5.428
Date:	Fri, 16 Jun 2017	Prob (F-statistic):	0.000864
Time:	18:14:27	Log-Likelihood:	-80.895
No. Observations:	34	AIC:	175.8
Df Residuals:	27	BIC:	186.5
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Annual Growth	-3.0661	0.760	-4.034	0.000	-4.626	-1.506
Bias	13.6633	11.576	1.180	0.248	-10.089	37.415
Corruption	-4.73e-07	4.08e-07	-1.158	0.257	-1.31e-06	3.65e-07
Factories	-9.247e-05	6.38e-05	-1.449	0.159	-0.000	3.84e-05
Gender Ratio	-0.0089	0.008	-1.079	0.290	-0.026	0.008
Literacy	0.0621	0.075	0.828	0.415	-0.092	0.216
Population	-1.045e-05	1.47e-05	-0.709	0.484	-4.07e-05	1.98e-05

Omnibus:	8.187	Durbin-Watson:	2.008
Prob(Omnibus):	0.017	Jarque-Bera (JB):	7.447
Skew:	0.769	Prob(JB):	0.0241
Kurtosis:	4.701	Cond. No.	3.37e+07

Simple Regression Model for Unemployment Rates

Resulting from the single regression analysis, the following equation shows the estimated effect of literacy rate on unemployment rate for each Indian state and UT.

$$unemp = -11.38 + 0.19(lit)$$

The regression of literacy on unemployment generates a positive coefficient of 0.19. After applying, the f-test, the p-value comes out to be 0.0152/2 = 0.0076 (about 0.76% which is less than 1%). We divide the p-value by two as this is a two tailed test i.e. the resulting coefficient could have been positive or negative. Therefore, we can reject our null hypothesis and further conclude that the regression is highly significant.
The correlation coefficient of this model (R^2) is about 17.1%. Hence, there must be some correlation between the two variables, although weak.

Multiple Regression Model for Unemployment Rates

Resulting from the multiple regression analysis, the following equation shows the estimated effect of literacy rate, population, annual growth rate, corruption, gender ratio and no. of factories on unemployment for each Indian state and UT.

$$unemp = 13.7 + 0.062(lit) - (4.73 \times 10^{-7})(corr) - 3.07(growth) - (1.05 \times 10^{-5})(pop) - 0.0089(genr) - (9.25 \times 10^{-5})(fac)$$

The multiple linear regression model generates a negative coefficient for the added features. The f-test results in a p-value of 0.000864. Therefore, we reject our null hypothesis and further can conclude that our model is highly significant.
The correlation coefficient of the model is 54.7%. Hence, there is moderate correlation between the independent variables and the dependent variable.

Conclusion
The relationship between literacy and unemployment rate, we started with the null hypothesis. For the simple linear regression the literacy rate had a positive effect on the unemployment rate. This means that as the literacy rate increases, the unemployment rate also increase. Although unexpected, it is possible, since unemployment is the no. of people actively seeking a job, we can deduce that as more people become literate, more people will start looking for a job.

In the multiple regression model, all the other features resulted in a negative effect on the unemployment rate. The multiple regression model resulted in a better correlation coefficient of about 55% and by the f-test, we were able to reject the null hypothesis and prove our model highly significant.

Additional Notes
1. Linear regression plots for all the features can be found in the /Plots folder.
2. The graphical representation of the correlation matrix can be found in the /Plots folder.