Module - 6

ALY 6010 Probability Theory and Statistics

RAHUL AVINASH JADHAV

Northeastern University

College of Professional Studies, Northeastern University, Boston, MA 02115

Contact: jadhav.ra@northeastern.edu

Submitted to Professor: Prof. Roy Wada

Date of submission: 12/18/2021

Introduction

In this assignment we are going to perform operations on the dataset “PSID”. The PSID consist of 14 variables and has 595 observations. The variables in the dataset are id, year, wage2, experience, weeks, occupation, industry, south, smsa, married, gender, union, education, ethnicity.

The operation we are going to perform is regression using lm() function. lm() function in R helps to create simple regression model. We are going to perform regression on main dataset i.e., PSID, and we are also going to create subset from it and perform regression on those subsets and observe the differences.

In addition we are also going to learn how to make dummy variables and what importance does dummy variables have while performing regression.

Analysis

Part 1:

Table1: Regression table on education

Regression table on Education

	Dependent variable:
	Wage
education	0.072***
	(0.006)
Constant	7.056***
	(0.075)
Observations	595
R^2	0.208
Adjusted R^2	0.207
Residual Std. Error	0.390 (df = 593)
F Statistic	156.167*** (df = 1; 593)
Note:	*p<0.1; **p<0.05; ***p<0.01

Regression table on Education

education

We have shown the relationship between wage and education in the table. From the above table you see that with every increase in the education level the wage gets increased by 0.072. The constant value is 7.056. the R square value is 0.21. so, the model fits 21% of our observations.

Part 2:

Table2: Regression table on education level, Gender male, Married

	Dependent variable:
	Wage
education	0.071***
	(0.005)
dummy_Male	0.332***
	(0.066)
dummy_married	0.159***
	(0.053)
Constant	6.640***
	(0.081)
Observations	595
R^2	0.335
Adjusted R^2	0.332
Residual Std. Error	0.358 (df = 591)
F Statistic	99.293*** (df = 3; 591)
Note:	*p<0.1; **p<0.05; ***p<0.01

education Gender Married

From the above table you can see that wage of male is 0.332 more than wage of female. Same as wage of married is 0.159 more than unmarried wage. The r square is 0.33. our regression model is 33% fit for our observations.

Part 3:

Table3: Regression table on 2 subset (Note (1) indicates Male dataset and (2) indicates female dataset)

	Dependent variable:	
	wages	
	(1)	(2)
Education	0.069***	0.102***
	(0.006)	(0.014)
Constant	7.149***	6.244***
	(0.075)	(0.181)
Observations	528	67
R^2	0.217	0.456
Adjusted R^2	0.216	0.448
Residual Std. Error	0.368 (df = 526)	0.284 (df = 65)
F Statistic	145.849*** (df = 1; 526)	54.535*** (df = 1; 65)
Note:		*p<0.1; **p<0.05; ***p<0.01

education

From the above table you can see, wages of male increase with 0.069 with increasing education level and wages of female increases with 0.102 with increase in education level. The r squared for male dataset is 0.21 and for female dataset is 0.448. the regression model for male dataset is 21% fit for our observation and 48% for female dataset.

Summary

- We performed simple regression model on Main dataset, subsets and using dummy variables.
- If we check the dataset, education have major impact on wages. The increase in education level increase the individual wages.

Bibliography

- Kabacoff, R. (2011). *R in action: Data analysis and graphics with R*. Manning.
- M. (n.d.). *Stargazer*. Jakeruss. https://www.jakeruss.com/cheatsheets/stargazer/