**Prakhar Rathi**

**1810110169**

**Q2. Does the sex of an individual affect educational attainment? Is there any evidence that the educational attainment of males is different from that of females?**

**Answer: (a)** In this question, we wish to find out whether the educational attainment of an individual is affected by their sex, hence, we wish to find out whether there is any relationship between the sex of an individual and the number of years of schooling. We will be using the following variables to model the relationship.

**Variable Definitions**

|  |  |
| --- | --- |
| **Variable Name** | **Variable Description** |
| S | Number of years of Schooling |
| ASVABC | measure of cognitive ability of a person |
| SM | years of schooling of the respondent’s mother |
| SF | years of schooling of the respondent’s father |
| MALE | dummy variable which is 1 for male respondents and 0 for female respondents |

This question has a variable which is qualitative in nature and hence does not have a continuous numerical value. The variable ‘MALE’ is therefore called a dummy variable because it represents categorical data i.e. sex of the individual. Since the sex of an individual can take on 2 values, we have taken one dummy variable to avoid the dummy variable trap.

So, we can represent our population equation as follows:

|  |  |
| --- | --- |
|  | 1 |

where β0, β1, β2, β3 and β4 are the coefficients, values of which we will try to estimate though the regression analysis. From the above equation, based on the value of MALE variable, we can get the equations 2 and 3.

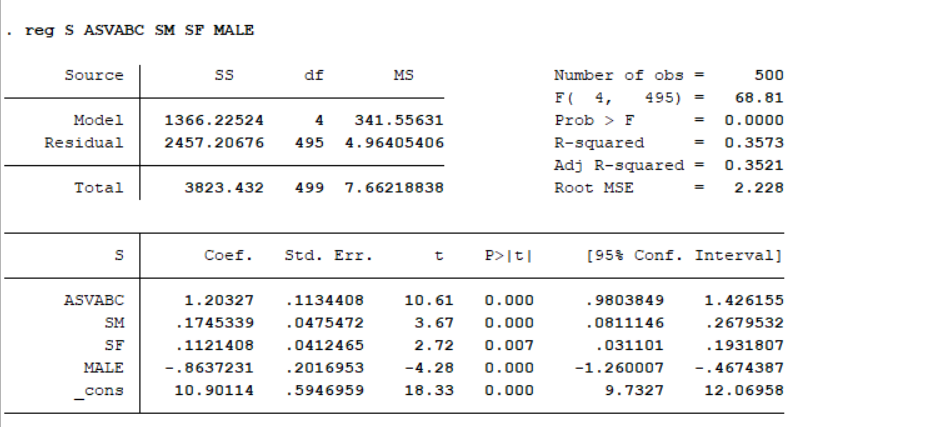
|  |  |
| --- | --- |
|  | 2 |
| Here, equation 2 is for males and equation 3 is for females. The average value of the error term would be zero so we can eliminate that. On subtracting 3 from 2, we get | 3 |

|  |  |
| --- | --- |
|  | 4 |

Here, β4 represents the additional number of years a person might get to attend school for on the basis of their gender. Now, we can write the regression equation as:

|  |  |
| --- | --- |
|  | 5 |

If the regression coefficients that we estimate are statistically significant then the discrepancy in the number of schooling years is also statistically significant. On running the regression analysis on STATA, we get this output.



Based on the above result, we can write the equation as follows.

|  |  |
| --- | --- |
|  | 6 |

This means that, holding SM, SF and the cognitive ability of a person constant, Males would attain 0.86 years less of education as opposed to females. However, we need to ascertain whether the coefficient of the MALE dummy variable is statistically significant or not. We can do this by performing a two-tailed *t*-test on the coefficient of the dummy variable. Our null hypothesis will be that the coefficient of the dummy variable is 0 meaning there is no effect of dummy variable on the years of schooling. Our alternative hypothesis will be that the dummy variable coefficient is not equal to 0 and it does impact the

|  |  |
| --- | --- |
|  | 7 |

We can calculate the t-stat as follows:

|  |  |
| --- | --- |
| The t-statistic value has been calculated as 2.28 and it can be seen that the STATA output is also the same for this null hypothesis. The degrees of freedom for this sample are 495 and the t-crit value at the 1 percent significance level is 2.586. Since, |t-stat| > |t-crit|, we will reject the null hypothesis at this level and accept the alternative hypothesis. Even at the 0.1 percent, the t-crit value is 3.31 which will lead to a rejection even at the 0.1 percent value with an even smaller risk of Type I error. Thus, we can say that the regression coefficient is statistically significant. The R-squared value is low but not very low to cause a lot concern.  The p-value for this regression is less than 0.001 which means that the probability of obtaining the corresponding t-statistic as a matter of chance is lower than 0.1% so we can reject the null hypothesis at that level.  Before I state my conclusion, I want to test the data for multicollinearity as well to make sure that the explanatory variables aren’t highly correlated. Multicollinearity occurs when one explanatory variable is linearly dependent on other variables. If there was perfect collineriaty then STATA would not have given any output but even if there is a high degree of multicollinearity then it’s a problem. The concern is that, as the degree of multicollinearity increases, the regression model estimates of the coefficients become unstable and the standard errors for the coefficients can get wildly inflated.  There are multiple ways to check for multicollinearity and I will be using two of them. The first one is variance inflation factor (vif). VIF is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable. A generally accepted practice is that a variable whose VIF values are greater than 10 may merit further investigation. [1] | 8 |

The second approach is more iterative in nature. I regress each of the Xi variables on the remaining X variables and find the corresponding R2i. If the R2i is high then Xi is likely to be correlated with the rest of the X variables. It is to be noted that I have only performed this analysis on the continuous variables. The results of the test are as follows:-

|  |  |
| --- | --- |
| **Xi** | **R2i** |
| SM | 0.34 |
| SF | 0.35 |
| ASVABC | 0.17 |
| MALE | 0.02 |

These R2 values are not very high which again confirms that our data isn’t very highly correlated. So, we can conclude that the sex of an individual does affect their educational attainment. **Controlling for all other factors, men are likely to attend 0.86 (approx. 11 months) less years of schooling as opposed to women.**

In this case, we will witness an intercept shift where the shift will be equal to the difference between the education level of men and women when only the SEX variable varies while everything else is constant.

|  |  |
| --- | --- |
|  |  |

Since, none of the slope variables are changing and β4 is added to the intercept, hence, only the intercept changes and the curves for men and women remain parallel. There is a parallel upward shift in case of women.

**Using your EAWE data set, define a slope dummy variable MALEASVC as the product of MALE and ASVABC: MALEASVC = MALE\*ASVABC. Regress S on ASVABC, SM, SF, ETHBLACK, ETHHISP, MALE, and MALEASVC, interpret the equation and perform appropriate statistical tests to comment on the use of the interaction variable in this regression, and whether that improves model specification.**

We are continuing our exploration to see whether the sex on an individual affects their ability to attain education. This time some new variables have been introduced.

|  |  |
| --- | --- |
| **Variable name** | **Definition** |
| MALEASVC | Interaction between male and ASVABC. |
| ETHHISP | Dummy variable to signify whether the ethnicity of the individual is Hispanic or not. 1 = Hispanic, 0 = not Hispanic |
| ETHBLACK | Dummy variable to signify whether the ethnicity of the individual is Black or not. 1 = Hispanic, 0 = not black |

This time we have made our estimates even more specific and we have reduced some of the unobserved factors by introducing dummy variables for Ethnicity of people. It is largely accepted that ethnicity of an individual can also affect their ability to attain education and so we are holding it constant for this test. Additionally, we have also introduced another variable called MALEASVC which is the product of MALE and ASVABC. This is a slope dummy variable and represents interaction between a dummy variable (MALE) and a continuous variable (ASVABC).

So, we can express this new model as follows:

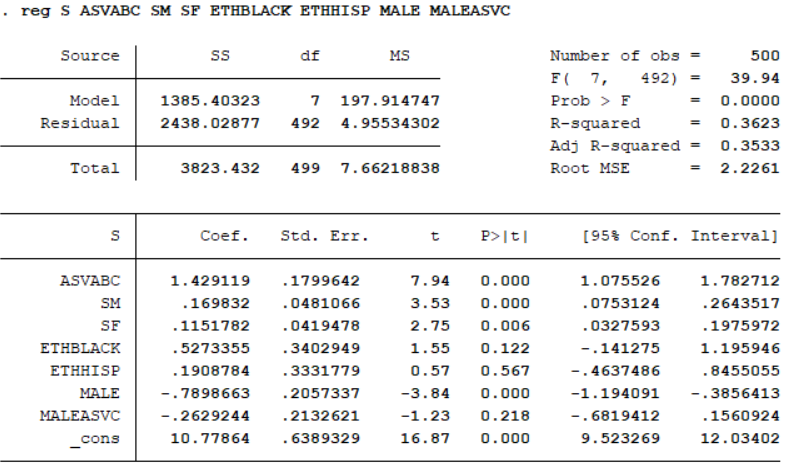
|  |  |  |  |
| --- | --- | --- | --- |
| where MALEASVC can be written as MALE\*ASVABC. We can write the equation for both males and females by replacing the dummy variable as 1 or 0 respectively.   |  |  | | --- | --- | |  |  | |  |
|  |  |

|  |  |
| --- | --- |
|  |  |

The above equation represents two things. First, if men and women had no cognitive ability (ASVABC = 0) then how many additional years of education would men be able to attain if β4 > 0. If it was less than 0 then women would be getting additional years of schooling if both men and women had no cognitive ability. Second, β5 represents the additional effect of increase of cognitive ability by 1 unit on the educational attainment of men. This additional effect would be positive if the coefficient was positive otherwise it would be negative. This is based on the reasoning that the cognitive ability might have different effects on the level of education of men and women. This term which now affects the slope coefficient of ASVABC will allow for different slopes for men and women which was not the case earlier. Now, we will witness a change in the intercept **and the** slope of the curve. However, this is contingent on whether our coefficients are statistically significant or not. This can be tested with our regression analysis and t-tests.

After running the regression on STATA, we get the following results (below) and our equation can be written as:

|  |  |
| --- | --- |
|  |  |



We need to check the significance of coefficients of MALE and MALEASVC.

Our hypotheses are:

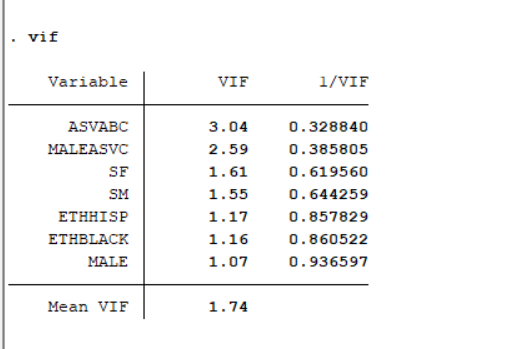
For male the t-statistic value is -3.84. At the 1 percent level for df = 492 is 2.59. Since, |t-stat| > t-crit, we will reject the null hypothesis and accept that there is a relationship between the sex and the level of education. Even at the 0.1 percent level, the t-crit is 3.31 which is lower than the |t-stat| at that level. So, we can reject the null hypothesis at the 0.1 percent level as well. The same can be concluded from our p-value. This is not surprising because we have tested this earlier as well. However, let’s see whether the cognitive ability has different effects for men and women.

Our hypotheses are:

We want to see whether the coefficient of the interaction variable is statistically significant or not. The t-stat value can be calculated as -1.23. At the 5 percent and 492 degrees of freedom, the t-crit value is 1.96. It can be clearly seen that the |t-stat| < t-crit value. So, we cannot reject the hypothesis at the 5 percent level. The p-value in the output clearly shows that the probability of getting our t-statistic value by chance which is matter of concern. So, we cannot accept that β5 ≠ 0.

We can now conclude a couple of things. The effect of ASVABC on the level of education is not different for men and women but there is still a difference between the level of education of men and women which is seen in both the cases. This means there will be a difference of intercepts in those curves. We can say that keeping all other factors same, women attain education for 0.79 years (approx. 10 months) more than men.

I have also performed a VIF test and the results show that all values are way below 10. The ASVABC and MALEASVC do show a certain degree of collineriaty but it is not high enough to raise a lot of concern.



Finally, our **adjusted R2 value does increase** in the second case which means that we are doing a **little better** with this model as compared to the previous case (roughly 0.1% more variance is explained by the new model as compared to the previous one). The model specification is improved. However, if the slope dummy variable is not statistically significant, we must question whether this increase is due to the variable MALEASVC or the ethnicity variables that we introduced. This requires further investigation.

**REFERENCES**

[1] *UCLA. (n.d.). Regression with Stata Chapter 2 – Regression diagnostics. IDRE Stats – Statistical Consulting Web Resources. https://stats.idre.ucla.edu/stata/webbooks/reg/chapter2/stata-webbooksregressionwith-statachapter-2-regression-diagnostics/*