Assignment #3: Module 5 Problem Set

Charles Titus

Grand Canyon University: RES – 866

October 2, 2019

**Directions:**

**Module 5 Problem Set**

**10.15 Learning Activity**

**1. Suppose you are interested in understanding how an employees demographic characteristics, beginning salary, and time at the bank and in the work force are related to current salary. Start by producing scatterplots of salbeg, sex, time, age, edlevel, and work with salnow. Add a fit line to each plot. Check on the variable labels for time and work so you understand what these variables are measuring.**

Below is the scatter plot and it includes a fitted line. The independent variable for this is Salbeg and the dependent variable is Salnow.













**2. Describe the relationships based on the scatterplots. Do they all appear to be linear? Are any relationships negative? What is the strongest relationship?**

By looking at the data one can see that two of the above have a linear relationship with the current salary. One that shows that there is a linear relationship with current salary is beginning salary. The other one that has a linear relationship with current salary is educational level. It is important to note that the scatter plot between sex and current salary is not actually considered a linear relationship because sex is considered a categorical variable. The rest of the scatter plots do not show a linear relationship because they seem to be flat, thus meaning that they are not linear in nature. I did not observe any sort of negative relationship on the scatterplots above. The one that has the strongest relationship can be seen between beginning salary and current salary, and this appears on the fitted line.

**3. Now produce correlations with all these variables. Which correlations with salnow are significant? What is the largest correlation in absolute value with salnow? Did this match what you thought based on the scatterplots?**

There seems to be significant correlations with beginning salary, age of employees, education level, and work experience. This is based upon each being at the .05 level of significance with the p-value of all of them being less than .05. I am kind of confused about the sex correlation because it appears to be significant, but from my understanding this is not true based upon sex being a categorical item.

The strongest correlation can be seen between beginning salary and the current salary because the correlation coefficient is .880 which is what was seen on the scatterplots above for the other question of this assignment.

SPSS:

|  |
| --- |
| **Correlations** |
|  | Beginning salary | Sex of employee | Age of employee | Job seniority | Current salary | Work experience | Current salary |
| Beginning salary | Pearson Correlation | 1 | -.457\*\* | -.011 | -.020 | .880\*\* | .045 | .880\*\* |
| Sig. (2-tailed) |  | .000 | .811 | .668 | .000 | .327 | .000 |
| N | 474 | 474 | 474 | 474 | 474 | 474 | 474 |
| Sex of employee | Pearson Correlation | -.457\*\* | 1 | .052 | -.066 | -.450\*\* | -.165\*\* | -.450\*\* |
| Sig. (2-tailed) | .000 |  | .259 | .148 | .000 | .000 | .000 |
| N | 474 | 474 | 474 | 474 | 474 | 474 | 474 |
| Age of employee | Pearson Correlation | -.011 | .052 | 1 | .052 | -.146\*\* | .804\*\* | -.146\*\* |
| Sig. (2-tailed) | .811 | .259 |  | .262 | .001 | .000 | .001 |
| N | 474 | 474 | 474 | 474 | 474 | 474 | 474 |
| Job seniority | Pearson Correlation | -.020 | -.066 | .052 | 1 | .084 | .003 | .084 |
| Sig. (2-tailed) | .668 | .148 | .262 |  | .067 | .949 | .067 |
| N | 474 | 474 | 474 | 474 | 474 | 474 | 474 |
| Current salary | Pearson Correlation | .880\*\* | -.450\*\* | -.146\*\* | .084 | 1 | -.097\* | 1.000\*\* |
| Sig. (2-tailed) | .000 | .000 | .001 | .067 |  | .034 | .000 |
| N | 474 | 474 | 474 | 474 | 474 | 474 | 474 |
| Work experience | Pearson Correlation | .045 | -.165\*\* | .804\*\* | .003 | -.097\* | 1 | -.097\* |
| Sig. (2-tailed) | .327 | .000 | .000 | .949 | .034 |  | .034 |
| N | 474 | 474 | 474 | 474 | 474 | 474 | 474 |
|  Current salary | Pearson Correlation | .880\*\* | -.450\*\* | -.146\*\* | .084 | 1.000\*\* | -.097\* | 1 |
| Sig. (2-tailed) | .000 | .000 | .001 | .067 | .000 | .034 |  |
| N | 474 | 474 | 474 | 474 | 474 | 474 | 474 |
| \*\*. Correlation is significant at the 0.01 level (2-tailed). |
| \*. Correlation is significant at the 0.05 level (2-tailed). |

**11.16 Learning Activity**

**1. Run a linear regression to predict total family income (income06) with highest year of education (educ). First, do a scatterplot of these two variables and superimpose a fit line. Does the relationship seem linear? How would you characterize the relationship?**

Please note…that I really struggled with section 11.16. I could not get the line on the first graph and then struggled with the other questions as well.

By looking at the scatterplot and the data of the variables it shows that it is linear and it looks to have a positive correlation.



**2. Now run the linear regression. What is the Adjusted R square value? Is the regression significant? What is the B coefficient for educ? Interpret it.**

Based upon the information below the Adjusted R square value is .131. The P-value of the R-Highest degree of coefficient is 0.000 <.05 significance level. This then means that the slop is indeed significant.

The B coefficient from the table below does show that there is an increase of the year in education and the income06 increases at .719 for a level.

|  |
| --- |
| **Model Summary** |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | .363a | .132 | .131 | 5.652 |
| a. Predictors: (Constant), HIGHEST YEAR OF SCHOOL COMPLETED |

|  |
| --- |
| **ANOVAa** |
| Model | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 9386.634 | 1 | 9386.634 | 293.810 | .000b |
| Residual | 61947.138 | 1939 | 31.948 |  |  |
| Total | 71333.772 | 1940 |  |  |  |
| a. Dependent Variable: TOTAL FAMILY INCOME 2006 |
| b. Predictors: (Constant), HIGHEST YEAR OF SCHOOL COMPLETED |

|  |
| --- |
| **Coefficientsa** |
| Model | Unstandardized Coefficients | Standardized Coefficients | t | Sig. |
| B | Std. Error | Beta |
| 1 | (Constant) | 7.948 | .580 |  | 13.695 | .000 |
| HIGHEST YEAR OF SCHOOL COMPLETED | .719 | .042 | .363 | 17.141 | .000 |
| a. Dependent Variable: TOTAL FAMILY INCOME 2006 |

**3. Next add the variables born (born in the U.S. or overseas), age, sex, and number of brothers and sisters (sibs). Check the coding on born so you can interpret its coefficient. First, do a scatterplot of age and sibs with income06. Superimpose a fit line. Does the relationship seem linear? How would you characterize the relationship? Why not do scatterplots of income06 with sex and born?**

Below is the scatter plot of siblings with family income 06. This scatterplot shows that there is a negative relationship between the siblings and total family income. This then means that if there are more brothers and sisters then it will have less family income total.



Below is the scatter plot of income 06 with age. On the scatter plot below on the fit line there is a small positive correlation between the variables and as a result there is a weak relationship in place.

The reason why you would not do scatterplots of income06 with sex and born is because sex and born are both considered categorical variables. As a result of them being considered categorical variables there is no point to do this.



**5. Which variables are significant predictors? What is the effect of each on income06? Which variable is the strongest predictor? The weakest?**

The variables that would be considered significant predictors are age, siblings, and education. The reason is because they are scale variables that can be looked in regards to a linear regression to determine the relationship with income06. The variable that has the strongest predictor is education and the one that seems to be the weakest is siblings.