**Assignment**

**1)**

1. How many observations are there?

* 5,801

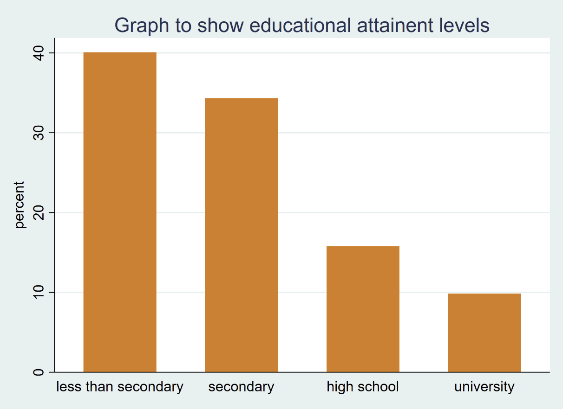
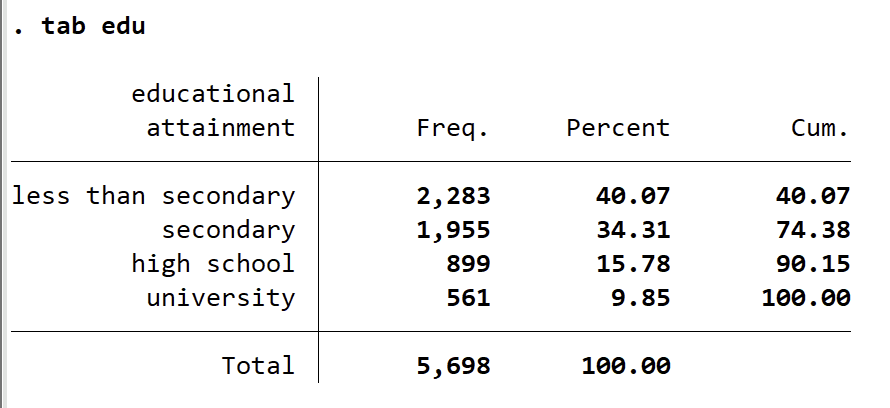
1. How many variables are there?

* 12

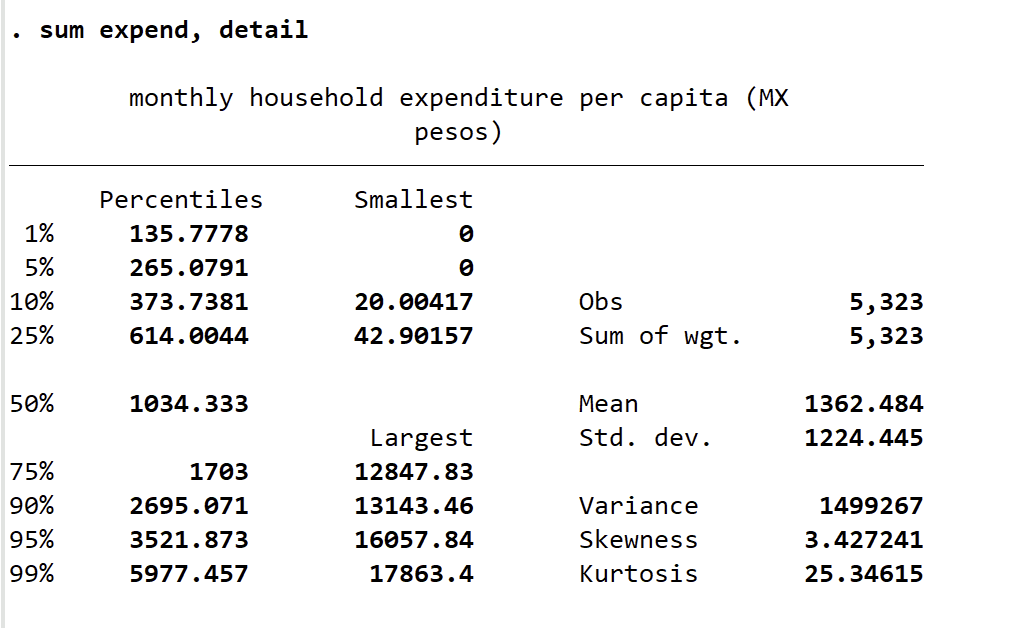
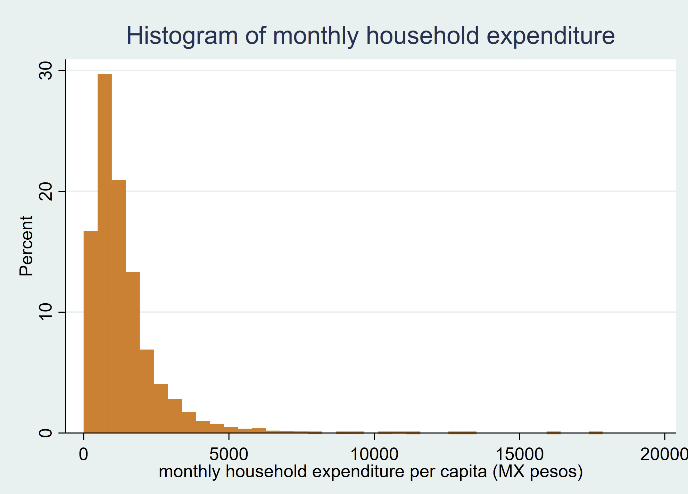
1. What is the population of reference for this specific sample?

* women of reproductive age (aged 15–45) living in Mexico in 2002

**2)**

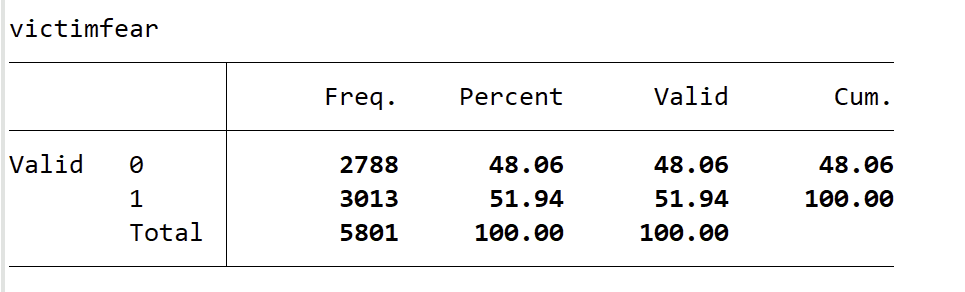


In our sample, out of 5,698 valid cases for educational attainment (which is a categorical and ordinal variable) ‘less than secondary’ is the most frequent level of educational attainment within the sample (40.07%). This is followed by ‘secondary’ levels of educational attainment which also holds a significant amount of the population (34.31%). Following this are ‘high school’ and university levels of education attainment which represent much less of the population respectively (15.78% for ‘high school) (9.85% for ‘university). This also demonstrated by the bar chart. This may be due to how the sample was collected and methods that were used. It should be noted that these variables have been considered with the exclusion of missing values.

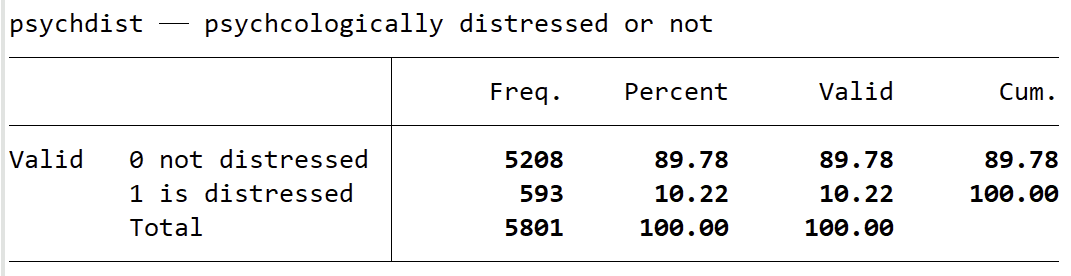


As demonstrated by the summary statistics displayed above, the mean household expenditure is approximately 1362.5 MX$ (MX pesos). The upper quartile (75%) of expenditure is 1703 whilst the lower quartile (25%) is 614.0044 subtracting these from one another gives us an interquartile range of 1,088.9956. The standard deviation of expenditure is 1224.445 which indicates that the spread of the data is dispersed and that data points vary a significant amount from the mean. The kurtosis of the expend variable which is significantly high (normal distribution being 3) thus the data should have a significantly high peak. A skewness value of 3 tells us that the data is positively skewed to the right and that there are more lower data points than high, high expenditure outliers are pulling the data. The histogram validates the conclusions made prior, that the data is positively skewed to the right and that there is a high peak and that there are more observations in bins that are lower value than that of higher.

**3)**

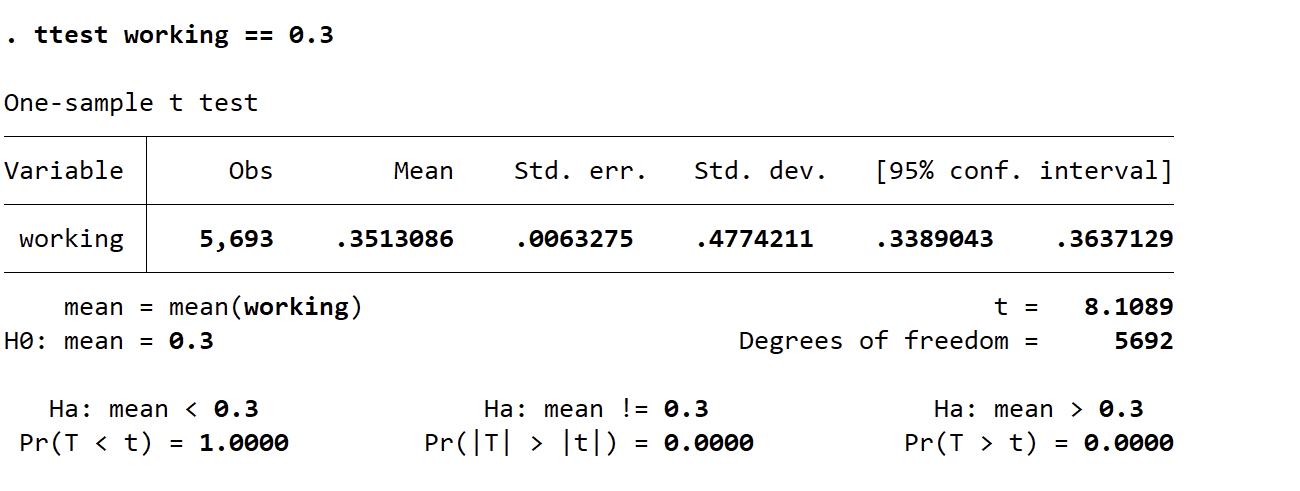


* **victimfear** records whether an individual is scared of being assaulted at day and/or night or neither. category 0 represents ‘fear at neither day nor night’ category 1 represents ‘fear at one/both’. The variable is both categorical and nominal as there are only 2 binary categories. In our sample 48.06% of people have not scared of day/nighttime assault whereas 51.94% are scared of either day/night assualt of both.

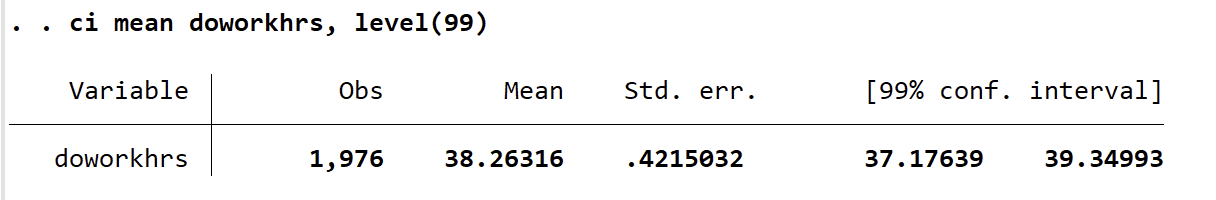


* **psychdist** records either high or low levels of self-reported psychological distress, category 0 represents ‘not distressed’ whilst category 1 represents ‘is distressed’. It is a nominal and categorical variable with binary categories. In our sample 89.78% are not distressed, 10.22% are.
* It is specifically difficult to measure fear and psychological stress firstly due to extraneous variables such as geographic influences, there is no variable/indicator of where in Mexico a respondent lives. Their environment, exposure to crime will influence their fear of assault or psychological stress, thus causing findings to lack external validity. Secondly participants may respond in socially desirable ways, some women may feel weak admitting they fear assault for example and may lie, this affects the internal validity of the study. In relation to the 21 questions asked to form the mhealth variable, the third issue is that these questions are subjective and as the study is self-reported respondents may have different interpretations of questions due to their own experiences, what one individual considers ‘losing interest in things’ or ‘acting pessimistic’ is different from another. This reduces the internal validity of the study. Finally, the study is cross-sectional and only captures the ‘last 4 weeks’, fear and psychological distress change over time and there might have been an unusual event in the last 4 weeks that specifically heightened a respondents fear of assault of psychological distress. There is also overlap between questions, feelings of anxiety, sadness or a lack of appetite may be interlinked and thus questions may be answered incorrectly.

**4)**

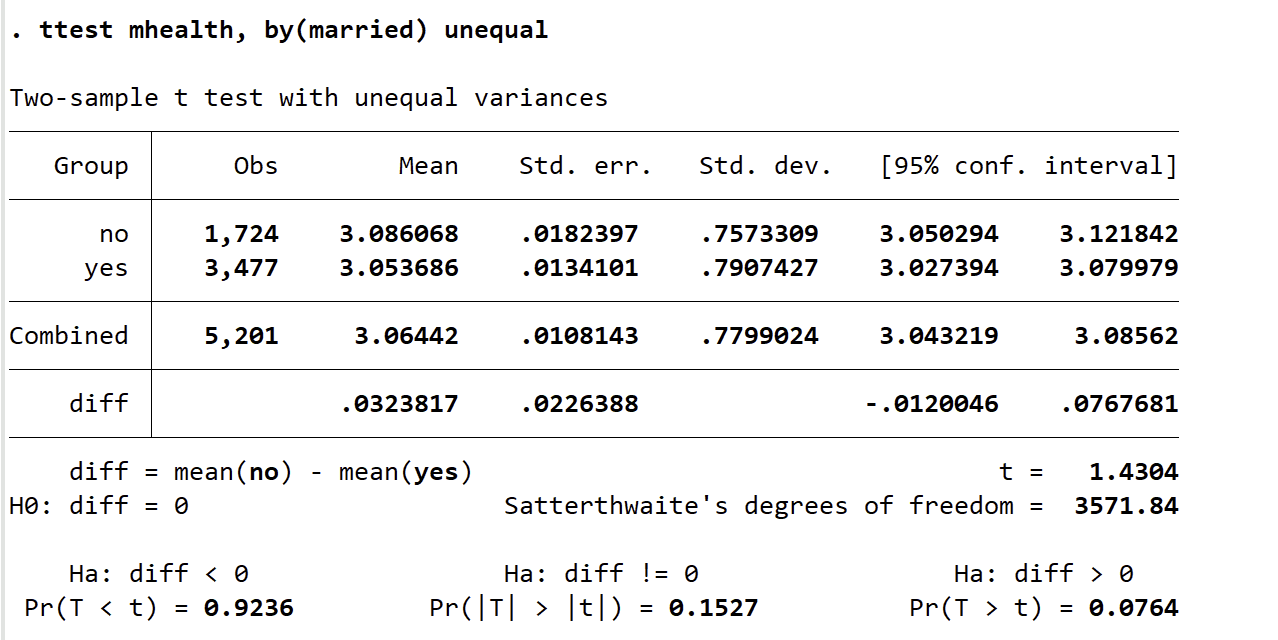
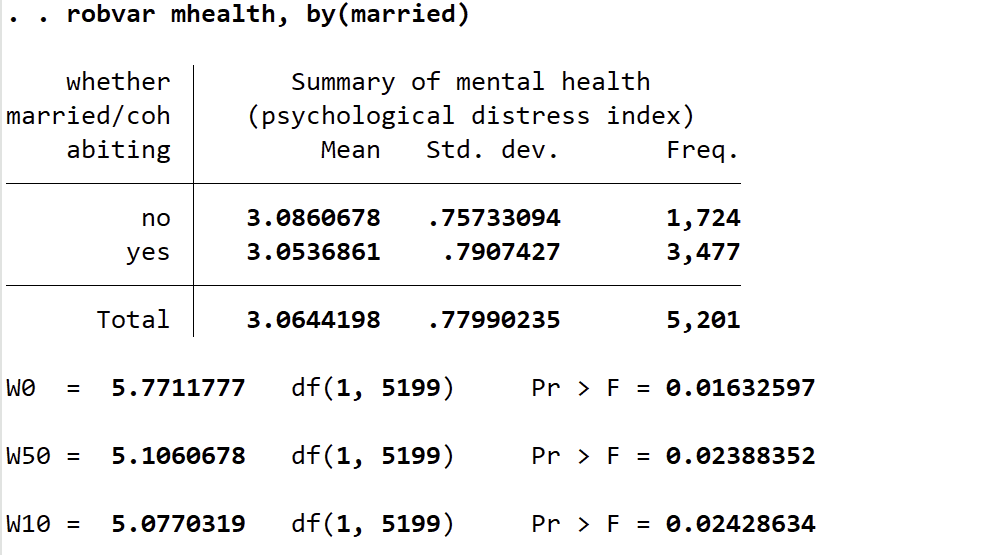
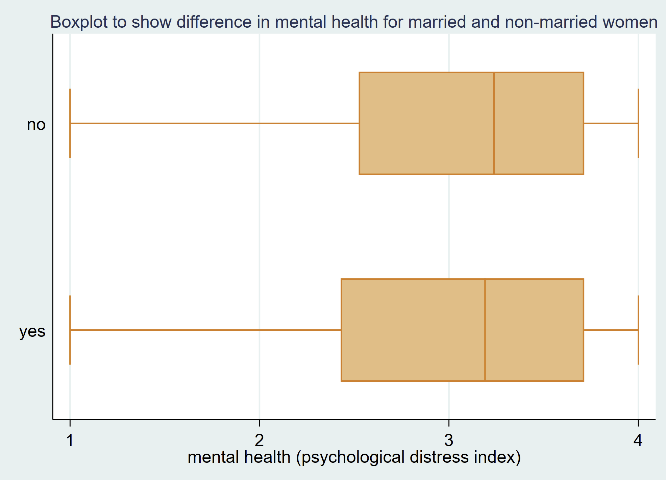


We should reject the null hypothesis that of the women who are of reproductive age living in Mexico in 2002, the proportion of women who work for pay is 0.351 and instead accept the alternate hypothesis that the proportion is not 0.3. This is evident as the 95% confidence interval provides evidence to reject the null as it shows intervals of 0.3389 to 0.3637 which is significantly different from 3 which is out of this range.

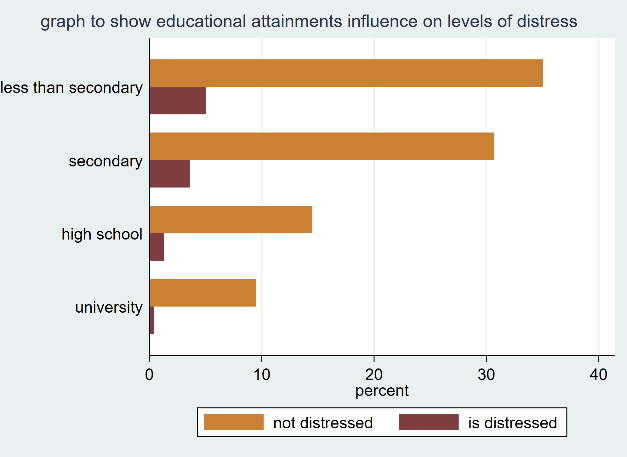
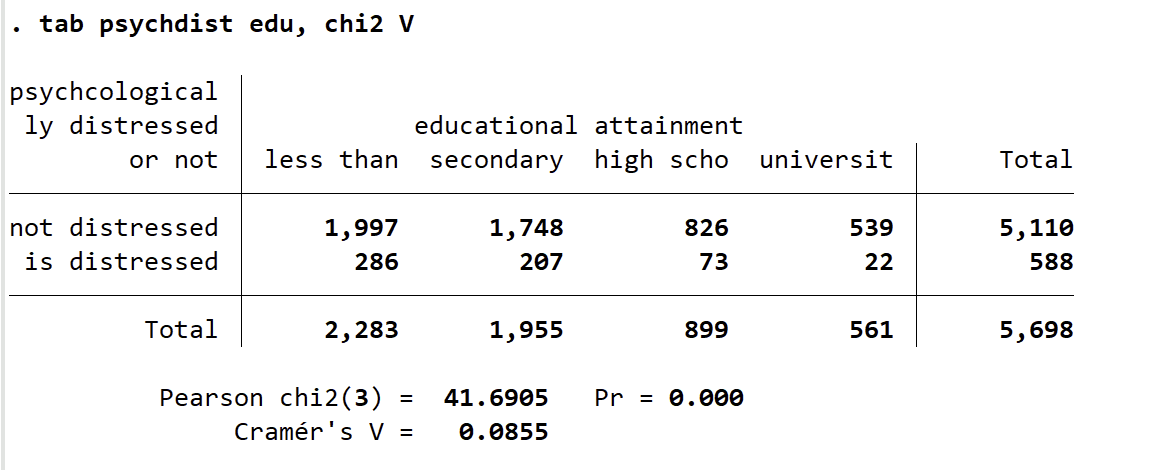
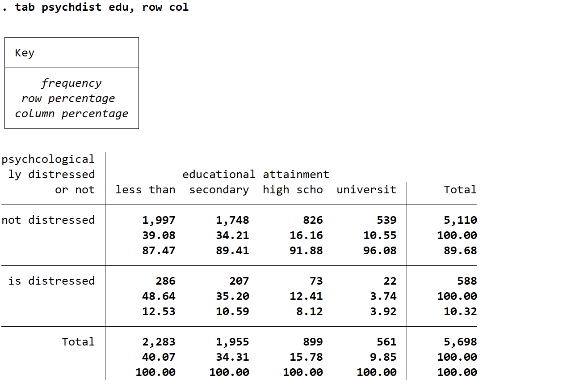
We fail to reject the null hypothesis that of the women who are of reproductive age living in Mexico in 2002, the mean number of working hours for women who work for pay is not significantly different form 40. The table demonstrates that with 99% confidence we can assume that the true population mean is between 37.17639 and 39.34993. This is different from 40 within the population however it is close in proximity, this demonstrates that the sample used is representative of the population of reference (that we are interested in generalising findings to).

* To complete this question, I had to create a new variable **doworkhrs** that specifically only included values above 0 and turned 0 work hours into missing (.)

**5)**



When looking at a ttest, based on a sample of 5,201 women who are 15-45 year old women living in Mexico. We can be 95% confident that the difference in mean mental health is between –0.0120046 and 0.0767681 for women who are married or not-married. We can infer from this and the means of the two categories (3.086 non-married and 3.054 married), that those who are not married tend to have better mental health. We would reject the null hypothesis of equality between in mental health between married and non-married women against as this difference is statistically significant due to a high p-value (0.1587 which is below p<0.05). This is further displayed by the boxplot above which shows that although the IQR is between higher than 2 and around 3.5 for both married and non-married, the IQR for married women starts closer to 2 suggesting a larger proportion of lower values for the married women category. Furthermore, the median for those who are not married is higher than that of those who are married. The IQR and median both validate the inference that women who are married tend to have a lower mental health score. Additionally, the Levene test demonstrates that the differences in variance between the category of married and non-married are different as the Pr > F is 0.01632 which is below our significance level of p<0.05 thus we would reject the null hypothesis that there is no difference in the variation and accept the alternate that there are differences in the variation.

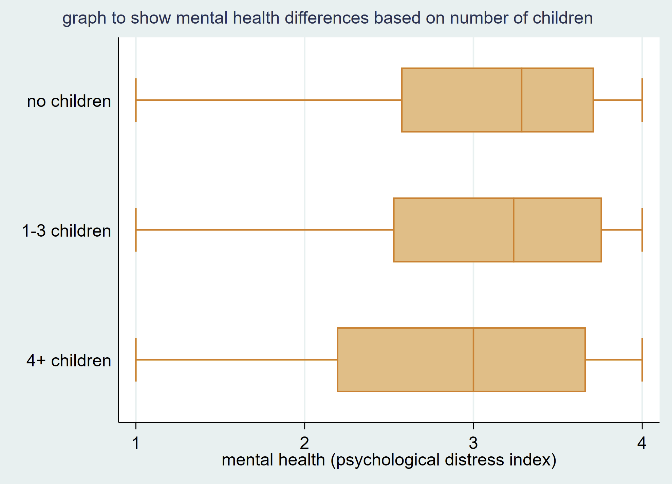
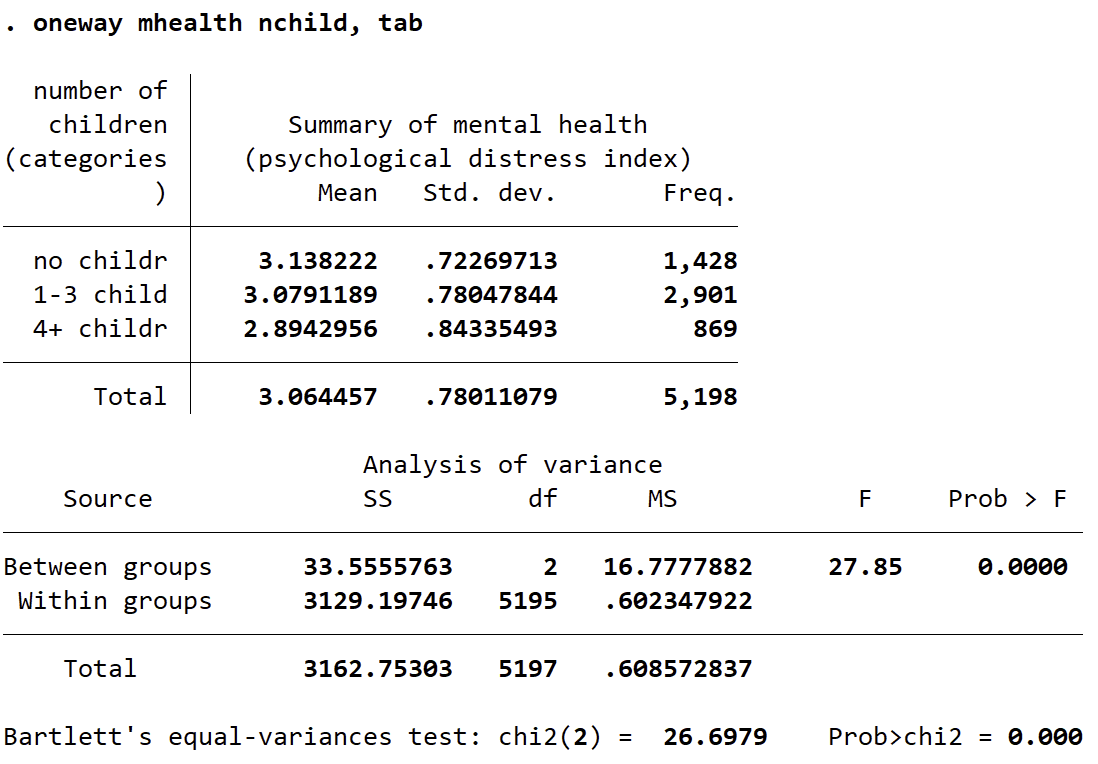


As displayed by the contingency table above, the group that has the highest proportion of distressed women who are 15-45 year old women living in Mexico is those with ‘less than’ educational attainment at a row percentage of 48.64%. The least distressed group is those who have had an educational attainment level of ‘university’ with a row percentage of 3.74%. Furthermore, of those who attended university only 3.92% reported being psychologically distressed. This evidence demonstrates that the higher level of educational attainment an individual achieves, the lower levels of distress they face. This is further demonstrated by the bar graph.

The Chi squared test displays a value of 41.6905 at p<0.0001, this would typically allow us to reject the null hypothesis as it demonstrates the variables have a statistically significant association. To look at the strength of this association we use Cramer’s V but the value provided is 0.085 which indicates no association or weak at best. From this we can infer that there is a statistically significant relationship between the variables, but it is weak.

Iii)

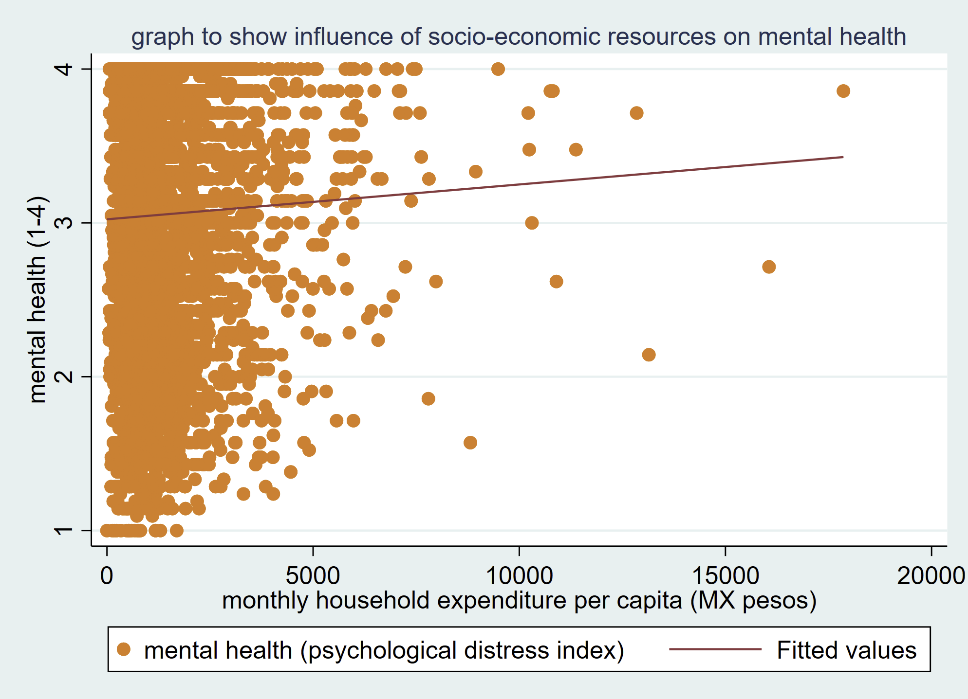
Using the interval-level index for mental health, report on differences in mental health across groups defined by the number of children (*nchild*).

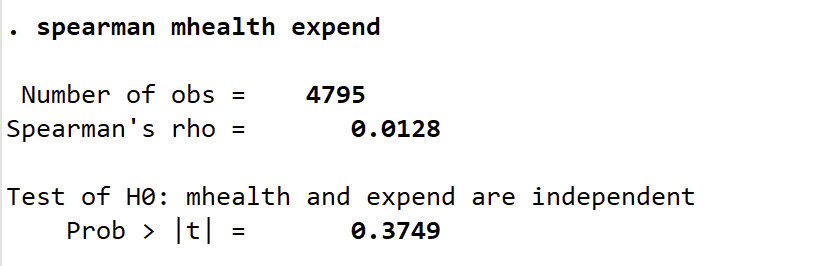
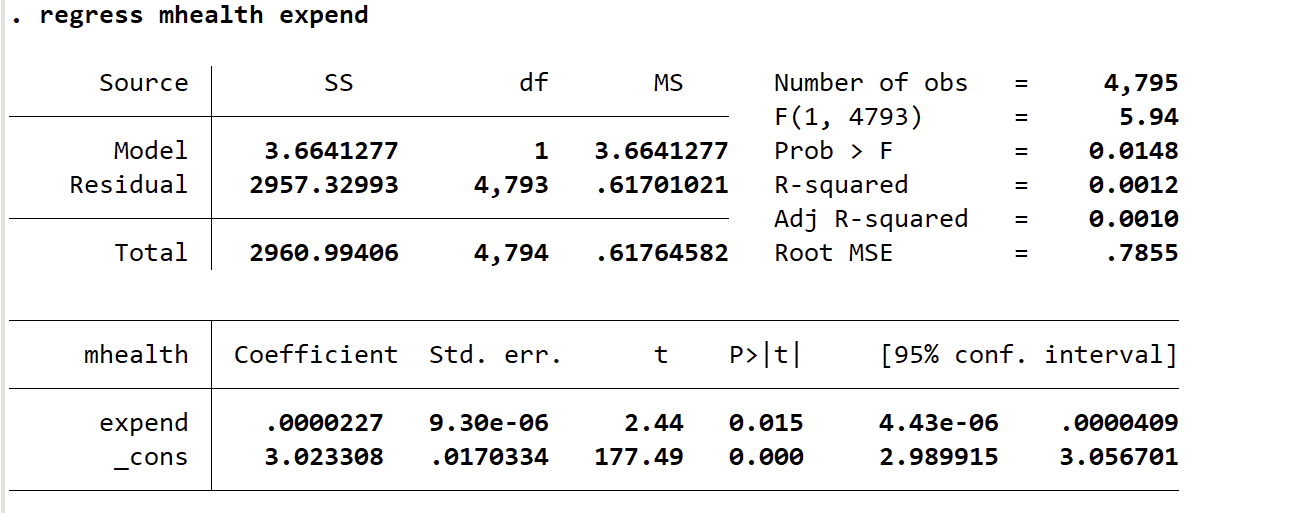


It is evident that there is a correlation between mothers who have more children having better mental health, the mean mental health for those with no children is 3.14 whilst it is 2.89 for those who have 4 or more. This is also extremely statistically significant as the p value is below p>0.0001. As the box plots there is a larger variation for mental health amongst those who have 4 or more children compared to no or 1-3 children.

**6)**

The Independent variable is socio-economic resources as it is hypothesised to have an effect on mental health, the dependent variable is mental health as it is hypothesised to be affected by socio-economic resources.



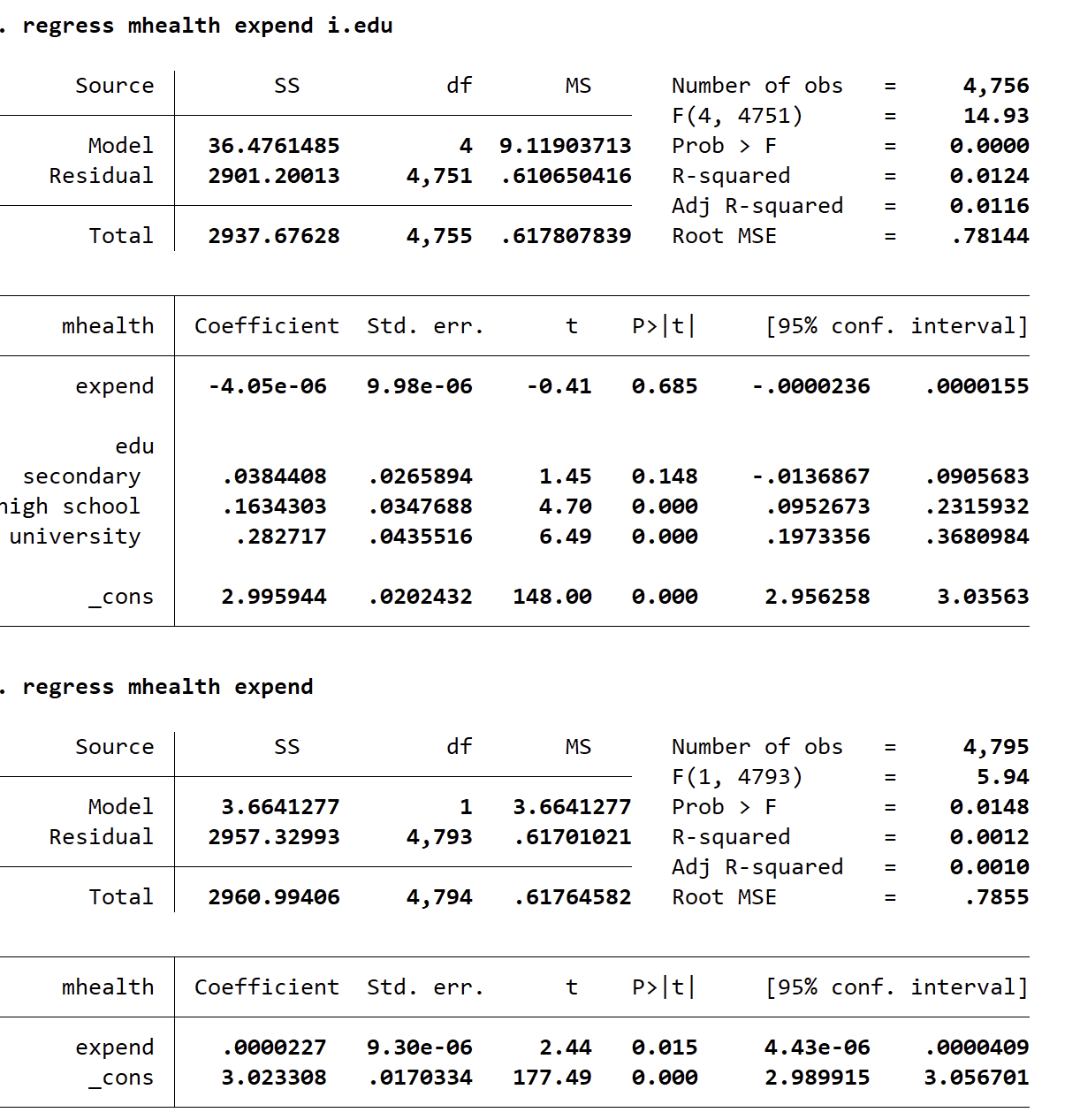


H0- There is an association between household expenditure and mental health

HA- There is no association between household expenditure and mental health

The scatter graph and regression line demonstrate that there is a minimal positive association between socio-economic resources and mental health. Spearman's rho provides further evidence on this as the correlation coefficient is 0.0128, indicating a weak positive association (when –1 is perfect negative and 1 is perfect positive). The p value is below p<0.05, which indicates that there is a statistically significant association between the variables. Additionally, in the regression above the adjusted R-squared value is 0.010 which further suggests that most of the variability in the dependent variable is unexplained by the model. The evidence presented here leads us to fail to reject the null hypothesis that there is a difference as there is a minimal positive association.

Education is a possible confounder between expenditure and mental health as women who have higher levels of education typically will enter higher paying jobs and thus have a higher household expenditure and socioeconomic status.



When education is incorporated into the regression, the adjusted R-squared increases from 0.0010 to 0.0116, this is shows education explains more variation in mental health which was previously unaccounted for in the model. Furthermore, the coefficient for education has changed and become negative this shows that the relationship seen previously between expenditure and mental health is in fact opposite.

education explains a significant amount of the variation in the dependent variable (mental health) that was previously unaccounted for by the model

An issue with estimating a causal relationship is sample selection bias. When researching socio-economic resources and mental health especially with a self-report survey, the respondents who participate may be extremely different from those who refrain from responding as mental health and socioeconomic resources are sensitive topics. This reduces the internal validity of the study as the findings may not be representative of the target population.

**7)**

In this assignment, I have found that around half of the women in our sample fear being assaulted in the night and/or evening. 90% of respondents are not psychologically distressed. Women who are married are typically more psychology distressed than those who are not married. Those who receive higher levels of education are less psychologically distressed. Women with more children are less psychologically stressed. There is a positive association between household expenditure and psychological distress.

H0 = (geographical location has an association psychological distress)

HA = (geographical location has no association with psychological distress)

I believe geographical location will be associated with psychological distress as educational attainment, household expenditure, fear of assault (crime) are all linked to the geographical location someone resides in. The data needed would have to include the ZIP code/states an individual resides in/

**8)**

//BEFORE ANSWERING QUESTIONS

\*exploring data

codebook

browse

list

describe

display

\*checking for missing variables/invalid response coded as -9

codebook mhealth // values are not labelled

codebook expend // values are not labelled

codebook working

codebook work\_hrs

codebook fear\_day

codebook fear\_night

codebook edu

codebook srh

codebook married

codebook age

codebook nchild

\*recoding variables with invalid responses of -9 (missing variables)

replace mhealth =. if mhealth ==-9

replace expend =. if expend ==-9

replace working =. if working ==-9

replace work\_hrs =. if work\_hrs ==-9

replace fear\_day =. if fear\_day ==-9

replace fear\_night =. if fear\_night ==-9

replace edu =. if edu ==-9

replace srh =. if srh ==-9

replace nchild =. if nchild ==-9

//Q1

describe

//Q2

tab edu

graph bar, over(edu)

sum expend, detail

hist expend, percent

//Q3

\*generating new fear of victimisation variables that are sorted into some or no fear

recode fear\_day (1/3 = 1) (4 = 0) ,gen(feardvictim)

recode fear\_night (1/3 = 1) (4 = 0), gen(fearnvictim)

\*creating new variable that is a combination of the recoded fear day and night variables

gen victimfear = fearnvictim | feardvictim

\*labelling

label var victimfear "fear at night and/or day"

label define fear 0 "fear at neither day nor night" 1 "fear at one/both", modify

label values victimfear fear

\*generating psychological distress variable

gen psychdist = (mhealth <2)

\*labelling

label var psychdist "psychcologically distressed or not"

label define dist 0 "not distressed" 1 "is distressed" , modify

label values psychdist dist

\*describing

tab victimfear

tab psychdist

//Q4

//i

\*conducting a one-way t-test to calculate if the proportion of women who work for pay is 0.3

ttest working == 0.3

\*generating new work\_hrs variable that only includes those who are working

recode work\_hrs (0 = .), gen(doworkhrs)

//ii

\*caluclating 99% confidence interval for weekly working hours for women who work for pay

ci mean doworkhrs, level(99)

ttest doworkhrs == 40 if doworkhrs == 1

//Q5

//i

\*Levene test

robvar mhealth, by(married)

\*using a two-sample ttest to test for the differences in mental health between married and non-married women

ttest mhealth, by(married) unequal

graph hbox mhealth, over(married)

//ii

\*chi sqaured test for thre relationship between categorical variables of educational attainment and psychological distress

tab psychdist edu, chi2 V

\*tabulation with row and column percentages

tab psychdist edu, row col

\*bar chart as visualisation

graph hbar, over(psychdist) over(edu) asyvar

//iii

\*ANOVA test for mental health differences in relation to number of children

oneway mhealth nchild, tab

\*visualisation for mental health differences in relation to number of children

hist mhealth, by(nchild)

graph hbox mhealth, over(nchild)

//Q6

//ii

\*association between mental health and socio-economic resources

regress mhealth expend

twoway scatter mhealth expend|| lfit mhealth expend

spearman mhealth expend

//iii

\*testing if education is a confounding variable for household expenditure and mental health with a regre

regress mhealth expend i.edu

regress mhealth expend

**[End]**