R-CLASS WORKSHEET

1. Data cleaning

The National Survey of Childhood Health (NSCH) is an annual survey to examines the physical and emotional health of children ages 0-17 years of age. The survey employs a two stage cluster sample, (1) A household is randomly selected then (2) a randomly selected child in the household has their physical and emotional recorded. The 2020 data is available in two tables (a) Screener table with information on the household (b) topical table with information on a randomly selected child. We are interested in creating a database to study childhood depression and its association with certain Childhood Adverse Event (ACE). Variables of interest in the data table are:

Variable Name	Description	Response Code
`ACE1`	Economic hardship?	1 = Never, 2 = Rarely 3 = Somewhat often, 4 = Very often
`ACE3`	Parent/Guardian Divorced?	1=Yes, 2=No
`ACE4`	Death of a parent/Guardian?	1=Yes, 2=No
`ACE5`	Parent/Guardian in jail or prison	1=Yes, 2=No
`ACE6`	Exposure to domestic violence	1=Yes, 2=No
`ACE7`	Victim/witnessed violence in neighborhood	1=Yes, 2=No
`ACE8`	Lived with person with mental illness	1=Yes, 2=No
`ACE9`	Lived with person with drug/alcohol abuse problems	1=Yes, 2=No
`ACE10`	Experienced racism?	1=Yes, 2=No
`K2Q32A`	Has had diagnosis of depression?	1=Yes, 2=No
`SC_AGE_YEARS`	Selected child age in years	

The child identifier is `HHID`

(1) (E) Create a subset of data with adverse childhood events (`HHID, ACE1, ACE3, ..., ACE12, K2Q32A, SC_AGE_YEARS`)

(2) (E) We are interested in children aged 5 to 17 years. Further refine your dataset to include children in this age range

(3) (I) Recode the data `ACE1= 1` if Somewhat often and very often, 2 if otherwise (careful for missing data) using the function `recode()` in package `car`

(4) (I) In SAS, NA is coded as '.'. You are sharing the data with a colleague who codes in SAS, recode all missing from NA to '.'. Has the characteristics of the variables in the data changed?

(5) (A) How many children have one or more missing ACE responses and or depression score? Further refine the dataset to exclude these children

2. Linear Regression Analysis

Perform the following analysis on the `BGSall` dataset:

(1) Estimate the correlation matrix for girls.

(2) For only girls, fit a simple linear regression model regressing `Soma` on `WT9` using the lm (See 1 – 3 below)

(3) Use `coef` to obtain the coefficients

(4) Obtain the confidence intervals, predictions using `confint, predict`

(5) For only girls plot extttSoma against `WT9`, label the axes and provide a title plot

(6) Add the regression line on the plot, using `abline`, try and make th line dashed and blue (hint: `lty =2, col = 'blue'`)