Assignment 7 (Correlation): Answers

1. We are interested to find out if one's hours of sleep are related to level of stress. We draw a random sample of 30 and calculate a sample correlation coefficient r = -.387. On the basis of this correlation coefficient, what can we say about the <u>strength and direction</u> of the relationship between one's hours of sleep and the level of stress in the sample? Using 99% confidence level, can we conclude that there is a relationship between one's hours of sleep and the level of stress in the broader population?

    a. On the basis of this correlation coefficient, what can we say about the <u>strength and direction</u> of the relationship between one's hours of sleep and the level of stress in the sample?

    For the sample, we can say that there is a weak negative relationship between these two variables.

    b. Using 99% confidence level, can we conclude that there is a relationship between one's hours of sleep and the level of stress in the broader population?

    1. State hypotheses:
    H0: $\rho = 0$ In words: There is no relationship between hours of sleep and stress.
    H1: $\rho \neq 0$ → two-tailed test   In words: There is a relationship between hours of sleep and stress.
    2. Select alpha: 0.01
    3. Test statistic: Correlation coefficient itself
    4. The correlation coefficient is r = -.387
    5. Use the table to find critical value: Table B4 (df=n-2=30-2=28, alpha=.01, two-tailed) → 0.4487
    6. Compare computed and critical value: 0.387< 0.4487.
    7. State your decision about H0: Fail to reject H0
    We report: r = -.387, p >.01
    8. Conclusion: Based on the data collected, we cannot conclude with 99% of confidence that there is a relationship between one's hours of sleep and the level of stress in the broader population.

    c. Evaluate the probability of making Type I and Type II errors after conducting this assessment.

    Probability of Type 1 error is 0, probability of Type II error is high because sample size is low.

    d. Calculate coefficients of determination and alienation. What do they tell us about the variance in one's hours of sleep and the level of stress?

    Coefficient of determination (r-squared)=-.387*-.387= .15.
    Coefficient of alienation=1-r^2=1-.15=.85

15% of variance in the level of stress can be explained by the number of hours a person sleeps, while 85% cannot be explained by the number of hours of sleep. (15% of variance is shared between the two variables, and 85% is unique.)

2. Use Stata on gss2012.dta dataset and focus on variables *hrs1* and *tvhours* to evaluate whether, among employed Americans, there is a relationship between the number of hours spent working and the number of hours spent watching TV.

   a. Evaluate this relationship in the sample and describe its strength and direction.

```
. pwcorr hrs1 tvhours, sig

             |     hrs1   tvhours
-------------+------------------
        hrs1 |   1.0000
             |
             |
     tvhours |  -0.1405    1.0000
             |   0.0001
```

   This is a very weak negative relationship; that is, those who work more hours watch fewer hours of TV on average.

   b. Can we conclude (with 99% confidence) that there is a relationship between the two variables in the population?

   1. Null hypothesis: Among employed Americans, there is no relationship between the number of hours spent working and the number of hours spent watching TV
   Research hypothesis: Among employed Americans, there is a relationship between the number of hours spent working and the number of hours spent watching TV.
   H0: $\rho = 0$
   H0: $\rho \neq 0$
   2. Alpha=.01 (since 99% confidence)., two-tailed
   3. Use correlation coefficient as a test statistic (because both variables are interval/ratio).
   4. Use Stata to calculate: Did that above.
   5. The p-value is .0001. We report: r = -.1405, p <.0001.
   6. P value is way below our alpha of .01
   7. We reject the null hypothesis in favor of research hypothesis.
   8. Based on data from a national representative sample, we are 99% confident that among employed Americans, there is a very weak yet statistically significant negative relationship between the number of hours spent working and the number of hours spent watching TV; that is, those who work more hours watch fewer hours of TV on average.

   c. Evaluate the probability of making Type I and Type II errors after conducting this assessment.

   Probability of Type II error is 0, probability of Type I error is based on p-value, so it is smaller than .0001.

   d. Calculate coefficients of determination and alienation. What does each of them tell us?

Coefficient of determination: .1405^2=.02
2% of variance in the number of hours spent working and the number of hours spent watching TV is shared.

Coefficient of alienation: 1-.02=.98
98% of variance in the number of hours spent working and the number of hours spent watching TV is unique, that is, not shared.

e. Use Stata to construct a scatterplot with a lowess curve for *hrs1* and *tvhours* (include hrs1 on the horizontal axis and tvhours on the vertical axis). Using this graph, discuss whether the relationship is or is not linear and explain why you see it that way.

```
. lowess tvhours hrs1
```
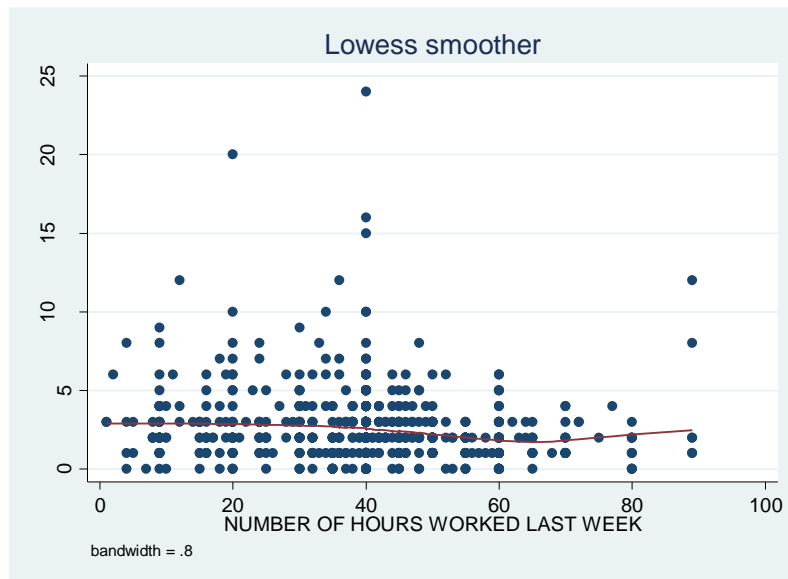

There is a linear decrease for most of the range of the data; there is a trend of an increase in the upper range, but that is mostly because of outliers (people working more than 80 hours).

Side note: if we restrict the range to no higher than 80 hours, we get the following graph:

```
. lowess tvhours hrs1 if hrs1<=80
```

3

Lowess smoother

bandwidth = .8