SOCY2200 Statistics **Instructor: Natasha Sarkisian**

Assignment 8. Regression and Chi-Square: Answers

Use hypothesis testing to answer the questions below. ***For each problem, make sure to state your null and research hypotheses in words as well as using formal notation. After finishing the test, state your formal conclusion with regard to the null hypothesis as well as your substantive answer to the question.*** Please make sure to include your Stata output (log and graphs) with this assignment.

1. We expect that negative emotions decrease with age. We use a nationally representative sample of 40 U.S. adults and ask people to record every time they notice having each of the negative emotions on the list we provide. Then we regress the number of times per day that people record negative emotions on their age (in years). We end up obtaining slope b = -0.438, with a standard error of .145.
   a. Write a sentence summarizing how exactly changes in one of these variables are linked to changes in the other in the sample (i.e., interpret the size and direction of the effect in the sample).
   b. Using 95% confidence level, can we conclude that negative emotions decrease with age? Make sure to state your null and research hypotheses in words as well as using formal notation. After finishing the test, state your formal conclusion with regard to the null hypothesis as well as your substantive answer.
   c. Evaluate the probability of Type I and Type II error after doing the test.

   a. As age increases by one year, the number of negative emotions one reports per day decreases by .438 of a time.

   b. Hypothesis testing:
   1. State hypotheses:
   H0: $\beta = 0$ In words: Among adult Americans, age has no effect on the number of times per day someone experiences negative emotions.
   H1: $\beta < 0$ In words: Among adult Americans, age reduces the number of times per day someone experiences negative emotions.
   2. Select alpha: 0.05
   3. Test statistic: Student's t
   4. $t = b/s_b = -0.438/0.145 = -3.02$
   5. Use the table to find critical value: Table B2 (df=n-2 = 40-2=38, alpha = .05, one-tailed) → 1.684
   6. Compare computed & critical value: 3.02 > 1.684
   7. State your decision: We reject H0 in favor of H1.
   8. Conclusion: Based on the sample of 40 U.S. adults, we are 95% sure that, among adult Americans, older age is associated with a lower number of times per day someone experiences negative emotions. (This relationship is statistically significant at .05 level.)

   c. Probability of Type II error is 0 because we rejected the null; probability of Type I error is less than .05.

2. Use Stata on gss2012.dta dataset and focus on variables *educ* and *hrs1* to evaluate whether higher levels of education lead to longer hours spent at work among employed Americans. Please use 99% confidence level in this assessment. Make sure to state your null and research hypotheses in words as well as using formal notation. After finishing the test, state your formal conclusion with regard to the null hypothesis as well as your substantive answer. In addition, do the following:

a. Write a sentence summarizing how exactly changes in one of these variables are linked to changes in the other in the sample (i.e., interpret the size and direction of the effect in the sample).
b. Interpret the intercept coefficient in words.
c. If the effect exists in the population, describe the slope in words using the 99% confidence interval.
d. Evaluate the probability of Type I and Type II error after doing the test.
e. Discuss the practical significance of this effect.
f. State what proportion of variance in your dependent variable is explained by your independent variable.
g. Construct a scatterplot with a lowess curve and a regression line for these two variables. Using this graph, discuss whether the relationship is or is not linear and explain why you see it that way.

Hypothesis testing:

1. H0: $\beta = 0$ In words: Level of education has no effect on hours spent at work among U.S. employees.
   H1: $\beta > 0$ In words: Higher levels of education lead to longer hours spent at work among U.S. employees.
2. We will use alpha of .01
3. Student's t
4. Find in Stata:

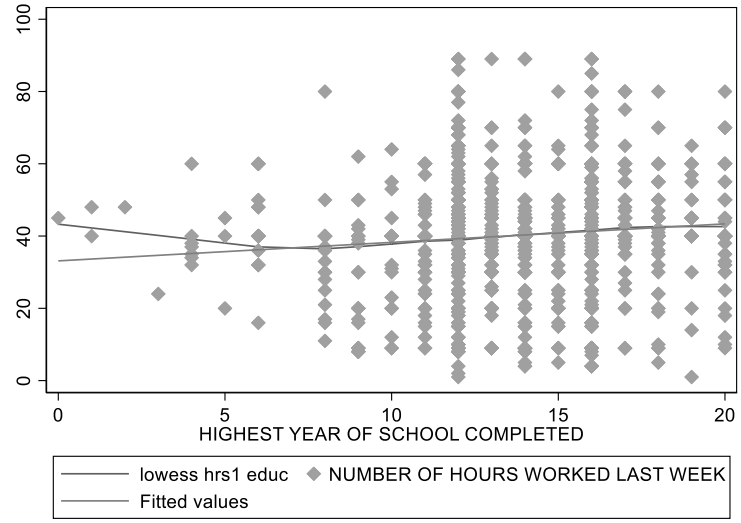
```
. reg hrs1 educ, level(99)
      Source |       SS           df       MS      Number of obs   =     1,165
-------------+----------------------------------   F(1, 1163)      =     12.10
       Model |  2898.15427         1  2898.15427   Prob > F        =    0.0005
    Residual |  278443.497     1,163  239.418312   R-squared       =    0.0103
-------------+----------------------------------   Adj R-squared   =    0.0095
       Total |  281341.652     1,164   241.70245   Root MSE        =    15.473

------------------------------------------------------------------------------
        hrs1 |      Coef.   Std. Err.      t    P>|t|     [99% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5126765   .1473538     3.48   0.001     .1324944    .8928586
       _cons |   33.12847   2.103158    15.75   0.000     27.70219    38.55476
------------------------------------------------------------------------------
```

5. Divide p value by two because it is a one-tailed test, we get 0.0005. We can write the result as b = .513, t=3.48, p <.001.
6. P value of 0.0005 is below our alpha cutoff of .01
7. We reject H0 in favor of H1.
8. Conclusion: Based on data from the GSS 2012 national sample of adults, we are 99.9% sure that, among U.S. adults, higher levels of education are associated with longer hours spent at work. (This relationship is statistically significant at .001 level.)

a. As education increases by one year, hours of work increase by approximately half an hour (.513 of an hour).
b. Based on our regression model, we predict that on average, someone with zero years of education works 33 hours a week for pay.

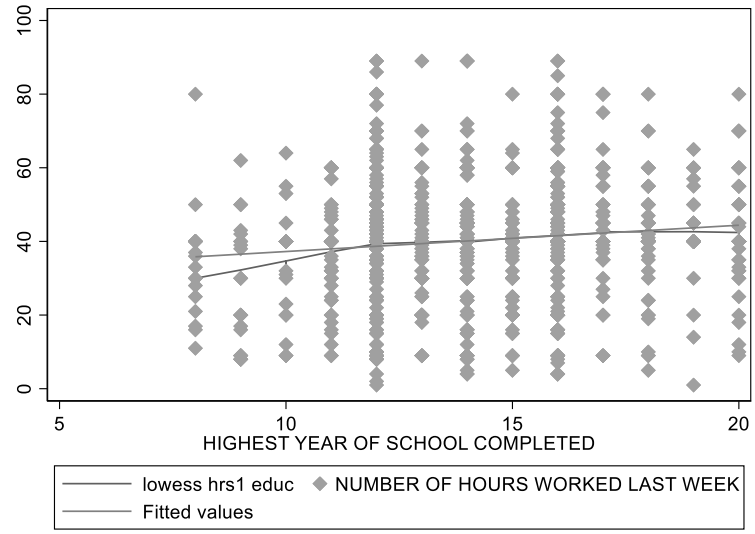
c. We are 99% sure that among employed Americans, as education increases by one year, hours of work increase by between .13 and .89 of an hour (i.e., somewhere between 8 and 53 minutes).
d. Probability of Type I error is .0005 (or 0.05% chances); probability of Type II error is zero because we rejected the null.
e. When one considers that four extra years of education (equivalent of a college degree) translates into between half an hour and 3.5 hours of extra work per week, the size of this effect is quite substantial, so we can consider that it is practically significant, even though it explains only 1 % of variance in hours of work.
f. R-squared shows that only 1% of variance in hours of work is explained by people's level of education.
g. Lowess curve shows that the relationship is rather linear starting at 8 years of schooling; however, the line between 0 and 8 years is going in the opposite direction, so overall, we can say it's not linear. Note, however, that there are few people with less than 8 years of education, so they could be considered outliers.

```
. graph twoway (lowess hrs1 educ) (scatter hrs1 educ) (lfit hrs1 educ)
```



In fact even if I look at values of educ above 6, I get a more linear picture:

```
lowess hrs1 educ if educ>6
```



3. You are interested in evaluating whether gender shapes opinion on death penalty. The survey asked a random sample of 130 American women and 70 American men whether they support or oppose death penalty. The results are as follows: Strongly Support: 30 women, 30 men; Support: 30 women, 20 men; Oppose: 50 women, 15 men; Strongly Oppose: 20 women, 5 men.

a. Construct a contingency table of these data with column percentages by hand.
b. Describe the pattern that you are seeing in the contingency table for the sample.
c. Do a chi-square test of independence on the data using 95% confidence level in order to determine whether there is a relationship between gender and opinion on death penalty.
d. If you conclude that the overall relationship exists in the population, conduct a post-hoc assessment to describe the likely pattern of differences in the population.
e. Evaluate the probability of Type I and Type II error after doing the test.

a. Table:

	C1 Women	**C2 Men**	**Row Totals**
R1 Strongly Support	30 23.08%	30 42.86%	60 30.00%
R2 Support	30 23.08%	20 28.57%	50 25.00%
R3 Oppose	50 38.46%	15 21.43%	65 32.50%
R4 Strongly Oppose	20 15.38%	5 7.14%	25 12.50%
Column Totals	130 100%	70 100%	200 100%

b. Men seem to be more likely than women to strongly support death penalty, and women are more likely than men to oppose or strongly oppose it.

c. Chi-square test:

1. H0: $O_i=E_i$

H1: $O_i \neq E_i$

H0: Among Americans, one's opinion on death penalty does not depend on gender (the two variables are independent).

H1: Among Americans, one's opinion on death penalty depends on gender.

2. Alpha=.05
3. Test statistic: χ^2
4. Calculate:

Cell	$E= (R_t \times C_t) /T$	O-E	$(O-E)^2$	$(O-E)^2/E$
C_1R_1	60*(130/200) = 39	30 – 39 = -9	81	81/39 = 2.08
C_2R_1	(60*70)/200 = 21	30 – 21 = 9	81	81/21 = 3.86
C_1R_2	(50*130)/200 = 32.5	30 – 32.5 = -2.5	6.25	6.25/32.5 = .19
C_2R_2	(50*70)/200= 17.5	20 – 17.5 = 2.5	6.25	6.25/17.5 = 0.36
C_1R_3	(65*130)/200 = 42.25	50 – 42.25 = 7.75	60.0625	60.0625/42.25 = 1.42
C_2R_3	(65*70)/200 = 22.75	15 – 22.75 = -7.75	60.0625	60.0625/22.75 = 2.64
C_1R_4	(25*130)/200 = 16.25	20 – 16.25 = 3.75	14.0625	14.0625/16.25 = 0.87
C_2R_4	(25*70)/200 = 8.75	5 – 8.75 = -3.75	14.0625	14.0625/8.75 = 1.61
Σ		0		χ^2 = 13.03

5. Now we find the critical value: Calculate degrees of freedom df = (4-1)x(2-1) = 3. Assuming a significance level of .05, the critical value of chi-square (from table B5, 3 degrees of freedom) is 7.82.
6. Compare computed value and critical value: 13.03>7.82
7. Decision: The obtained chi-square of 13.03 exceeds the critical value of 7.82, so reject the null hypothesis.
8. Conclusion: Based on the data from 200 Americans, we are 95% certain that there is a relationship between gender and opinion on death penalty in the broader population of Americans.

d. Large residuals (>1.96) correspond to cells C1R1, C2R1, and C2R3. In the population, men are more likely than women to strongly support death penalty (43% vs 23%); men are also less likely than women to oppose it (21% vs 38%).

e. Probability of Type I error is less than .05; probability of Type II error is zero because we rejected the null.

4. Use Stata on gss2012.dta dataset and focus on variables *happy* and *degree* to evaluate whether education has effects on general happiness among Americans.
   a. Construct a contingency table of these data with column percentages.
   b. Describe the pattern that you are seeing in the contingency table for the sample.
   c. Use Stata to do a chi-square test of independence on the data using 99% confidence level in order to determine whether, in the U.S. population, there is a relationship between education and general happiness.
   d. If you conclude that the overall relationship exists in the population, conduct a post-hoc assessment to describe the likely pattern of differences in the population.
   e. Evaluate the probability of Type I and Type II error after doing the test.

      a. See below in Stata

      b. Looking at percentage distribution, it appears that those with a bachelor's degree are especially likely to report being very happy (35% vs 28-31% in other groups), while those with less than high school education are much more likely than any other group to report being not too happy. (25% vs 9-15% in other groups).

      c. Testing:

1. H0: $O_i = E_i$ -- Education is not related to general happiness among Americans.
   H1: $O_i \neq E_i$ -- Education has an effect on general happiness among Americans.
2. Select alpha: We will use alpha of .01
3. Chi-square
4. Calculate in Stata:

```
. tab happy degree, col chi
+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+
      GENERAL |                   RS HIGHEST DEGREE
    HAPPINESS | LT HIGH S  HIGH SCHO  JUNIOR CO   bachelor   graduate |     Total
--------------+---------------------------------------------------------+----------
```

```
   VERY HAPPY |         82        277         47        123         64 |        593
              |      28.98      28.41      31.13      35.14      31.22 |      30.19
--------------+-----------------------------------------------------------+----------
 PRETTY HAPPY |        131        564         82        196        121 |      1,094
              |      46.29      57.85      54.30      56.00      59.02 |      55.70
--------------+-----------------------------------------------------------+----------
NOT TOO HAPPY |         70        134         22         31         20 |        277
              |      24.73      13.74      14.57       8.86       9.76 |      14.10
--------------+-----------------------------------------------------------+----------
        Total |        283        975        151        350        205 |      1,964
              |     100.00     100.00     100.00     100.00     100.00 |     100.00

          Pearson chi2(8) =  42.2607   Pr = 0.000
```

5. $\chi^2 = 42.261$, $p<.001$
6. P-value is below alpha .01
7. We reject the null hypothesis in favor of our research hypothesis.
8. Based on GSS 2012 national sample of U.S. adults, we are 99.9% confident that among Americans, there is a relationship between one's education level and the general degree of happiness one experiences. (This relationship is significant on .001 level.)

d. Posthoc assessment:

```
. tabchi happy degree, adj
          observed frequency
          expected frequency
          adjusted residual
-------------------------------------------------------------------------------------------
GENERAL       |                              RS HIGHEST DEGREE
HAPPINESS     | LT HIGH SCHOOL     HIGH SCHOOL  JUNIOR COLLEGE        bachelor        graduate
--------------+----------------------------------------------------------------------------
   VERY HAPPY |             82             277              47             123              64
              |         85.448         294.386          45.592         105.677          61.897
              |         -0.483          -1.709           0.260           2.225           0.338
              |
 PRETTY HAPPY |            131             564              82             196             121
              |        157.638         543.101          84.111         194.959         114.190
              |         -3.446           1.899          -0.360           0.124           1.012
              |
NOT TOO HAPPY |             70             134              22              31              20
              |         39.914         137.513          21.297          49.364          28.913
              |          5.554          -0.455           0.171          -3.111          -1.890
-------------------------------------------------------------------------------------------
          Pearson chi2(8) =  42.2607   Pr = 0.000
 likelihood-ratio chi2(8) =  39.5845   Pr = 0.000
```

We will use the cutoff for alpha=.01, which is 2.576. Cells C1R2 C1R3 and C4R3 exceed that cutoff. Based on these residuals, we can conclude that in the population, those with a bachelor's degree are especially unlikely to report being not too happy, while those with less than high school education are particularly likely to report being not too happy, and less likely than others to report that they are pretty happy.

e. Probability of Type I error is less than .001 (based on p-value), and probability of Type II error is zero because we rejected the null.