Assignment 7

Sri Harsha Sudalagunta

2023-04-13

Problem 1.

(a)

Form the given data, there exits a significant relationship between the both variables. The relationship is inversely proportional, with decrease in the one there is an increase in the other.

(b)

Yes, I think its meaningful. But, The given intercept i.e., 4.9754, which mean that even when the resilience is zero, we can have the BDI score as 4.9754 which doesn't make more sense. It may look meaningful technically but not practically.

(c)

The slope of resLow is 9.6538 and the slope of resVeryLow is 15.6453, which mean that the person with resVeryLow can have 6 points high BDI score than the person with resLow resilience.

(d)

Yes based on the given statistic test value i.e., 118.1 and p-value <0.05, there exists an relationship between both of the variables. (e)

i) There will be a high BDI score with the very low when compared to the low resilience which is in line with the previously mentioned study. But, the difference between the groups is same here but difference between categories is different in the previous study.
ii) This imply that difference in the score between the groups is the i.e., 2.76 points. But, the differnce is varied between the groups in the previous study.
iii) Because it treats the resilience categories as numerical, in which the difference between the groups is same. But in the reality case, the difference between the groups is different based on the category. so, this is the reason that i think the model is flawed.

Problem 2.

(a)

I think chi-square test would be appropriate for the particular scenario.

(b)

H0 : caffeinated coffee consumption and risk of depression in women are not related to each other

H1 : caffeinated coffee consumption and risk of depression in women are related to each other

(c)

The proportion of females who suffer from depression = 2607/50739

```
2607/50739
```

```
## [1] 0.05138059
```

The females who is not suffering from depression is 48132/50739

```
48132/50739
```

```
## [1] 0.9486194
```

(d)

E = (total of row * total of column)/total count

```
E = (2607*6617)/50739
E
```

```
## [1] 339.9854
```

Contribution of the cell = (observed-expected)ˆ2/expected

```
C = (373-E)^2/E
C
```

```
## [1] 3.205914
```

(e)

statistic value (q) = 20.93

df = (number of rows -1) * (number of columns-1) = 4

```
test_value = 20.93
df = 4
pchisq(test_value, df, lower.tail = FALSE)
```

```
## [1] 0.0003269507
```

(f)

Based on the p value and test statistic we can assume that there exists a relationship between both of the variables.

From the p-value and test statistic value, there exists a relationship between the risk of depression in women and caffeinated coffee consumption.

(g)

Yes, I think we cannot come to complete conclusion that coffee was being the causation of the depression because there may be confounding factors associated with that. Moreover, from this we can see that there is an association but not the causation.

Problem 3

(a)

Yes, this data can be used . But, based on the sample size we cannot generalize I think.

(b)

Lets calculate do the table

```
c_table <- matrix(c(100, 50, 20, 10, 17, 4), nrow = 2, byrow = TRUE)
rownames(c_table) <- c("Sampled Claims", "No Allowed")
colnames(c_table) <- c("Small", "Medium", "Large")
observed_table = addmargins(c_table)
observed_table
```

```
##                Small Medium Large Sum
## Sampled Claims   100     50    20 170
## No Allowed        10     17     4  31
## Sum              110     67    24 201
```

(c)

E value = (total of row * total of column)/total count

```
E_val = (31*24)/201
E_val
```

```
## [1] 3.701493
```

(d)

As the data counts are more than 10 and the data is in appropriate table format with categories, its use is justified.

(e)

```
test_val = 12.93
d_of_fre = 2
pchisq(12.93 ,2, lower.tail = FALSE)
```

```
## [1] 0.001556991
```

The p-value for the test of no association is 0.001

(f) t)

I think we can calculate the residuals by difference them with the observed and the expected(E) values

```
E <- margin.table(c_table, margin=1) %*% t(margin.table(c_table, margin=2)) / sum(c_table)
E
```

```
##                    Small   Medium     Large
## Sampled Claims 93.03483 56.66667 20.298507
## No Allowed     16.96517 10.33333  3.701493
```

Then subtracting the expected(E) form the residual(r)

```
r <- (c_table - E) / sqrt(E)
r
```

```
##                     Small     Medium       Large
## Sampled Claims  0.7221197 -0.8856149 -0.06625568
## No Allowed     -1.6910359  2.0739034  0.15515535
```

The Positive residuals represents greater than expected frequencies and negatives residuals indicates lower than expected frequency. The residuals are positive for the given data indicating higher than expected frequencies. The residuals have three negative and three positive values.

Problem 4.

a)

we can assume an hypothesis

H0 : Farm environment and risk of childhood asthma are not associated with each other.

H1 : Farm environment and risk of childhood asthma are associated with each other.

```
expo_asthma = 19
expo_noasthma = 966
not_expo_asthma = 11566
not_expo_noasthma = 263757


c_table =  matrix(c(expo_asthma, expo_noasthma, not_expo_asthma, not_expo_noasthma),nrow = 2, byrow = T

c_table
```

```
##                    Asthma_yes Asthma_no
## Exposed farm               19       966
## Not Exposed farm        11566    263757
```

Lets do a chi-square test to see the values

```
test_chi =  chisq.test(c_table)
test_chi
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  c_table
## X-squared = 12.053, df = 1, p-value = 0.0005172
```

The test statistic is 11.41 and p-value is less than 0.05. so, based on this we can see that there is a relationship between both the variables.

(b)

I think yes, the relative risk is an appropriate measure of association for these data. The sample is not on development or diagnosed of the asthma but its on whether they are exposed or not exposed to farm animals. So, i think relative risk is an appropriate measure of association/

(c)

The findings of this study do not provide proof that being around farm animals lowers the chance of having asthma. The results only support a link between early exposure to farm animals and a decreased risk of developing asthma by age 6. This link might be influenced by additional variables or confounding factors associated with it. As it is a retrospective kind of study, the confounding factors may not be addressed.As the farm people work hard, due to which there can be changes in the genetic predisposition of the children form parents, instead of exposure to the farm animals.

Problem 5.

(a)

lets do the odds ratio of the survival = ad/bc

```
a = 14
d= 39
b= 11
c= 26
o = (a*d)/(b*c)
o
```

```
## [1] 1.909091
```

Lets see the relative risk of survival = (a/a+b)/(c/c+d)

```
a = 14
d= 39
b= 26
c= 11
rr = (a/(a+b))/(c/(c+d))
rr
```

```
## [1] 1.590909
```

(b)

The relative risk should always be used to compare the probability of success between two interventions because it is the more comprehensible and natural method. It is preferable in this situation and can be calculated in studies that are prospective.

Problem 6.

(a)

Based on the given information, lets do the chi square test and check the p-value.

```
tea = matrix(c(17, 283, 30, 541), nrow = 2)
rownames(tea) = c("Carcinoma", "No Carcinoma")
colnames(tea) = c("Green Tea", " Doesn't drink Green Tea")
test = chisq.test(tea)
test
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tea
## X-squared = 0.0096778, df = 1, p-value = 0.9216
```

The p value is grater than 0.05 , its insufficient proof to reject the H0 of no association.

Based on this we can assume that they are independent of each other.

(b)

I think finding out the odds ratio would be an appropriate measure.

```
odds = (17/30)/ (283/541)
odds
```

```
## [1] 1.083274
```

The odds ratio is 1.08

The odds ratio is more than 1, so lets calculate the percentage increase

```
percentage_increase = (odds-1)*100
percentage_increase
```

```
## [1] 8.327444
```

There is an 8.3% increase in odds for those who drink green tea regular than the other group.