Problem Set 2 Answers

Kaley Burg

November 19, 2023

Note

Please note that the R file used in this tex file is `PS02_answers_KB_Cleaned_R.R` because that was the simplified version of my script. However, the mistakes, previous tries, and the code I used to come up with my for loops is in another R script called `PS02_R_KB_safety.R` and this is also in my Github.

Instructions

- Please show your work! You may lose points by simply writing in the answer. If the problem requires you to execute commands in `R`, please include the code you used to get your answers. Please also include the `.R` file that contains your code. If you are not sure if work needs to be shown for a particular problem, please ask.
- Your homework should be submitted electronically on GitHub.
- This problem set is due before 23:59 on Sunday October 15, 2023. No late assignments will be accepted.

Question 1: Political Science

The following table was created using the data from a study run in a major Latin American city.[1] As part of the experimental treatment in the study, one employee of the research team was chosen to make illegal left turns across traffic to draw the attention of the police officers on shift. Two employee drivers were upper class, two were lower class drivers, and the identity of the driver was randomly assigned per encounter. The researchers were interested in whether officers were more or less likely to solicit a bribe from drivers depending on their class (officers use phrases like, "We can solve this the easy way" to draw a bribe). The table below shows the resulting data.

[1] Fried, Lagunes, and Venkataramani (2010). "Corruption and Inequality at the Crossroad: A Multimethod Study of Bribery and Discrimination in Latin America. *Latin American Research Review.* 45 (1): 76-97.

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	14	6	7
Lower class	7	7	1

(a) Calculate the χ^2 test statistic by hand/manually (even better if you can do "by hand" in R).

- First I created a matrix with the data that was given, with 2 rows and 3 columns. I put this in a variable named polscidata. Then I added row and column names to correspond with what was provided. This code is given here:

```
polscidata <- matrix(c(14, 7, 6, 7, 7, 1), nrow = 2, ncol = 3)
colnames(polscidata) <- c("Not Stopped", "Bribe requested", " Stopped
    /given warning")
rownames(polscidata) <- c("Upper class", "Lower class")
```

- Then I made a table with the conditional proportions of the table (in other words, a table with the probabilities of two specific values of each of the variables co-occurring together). This was not necessary to the analysis, but I looked at it regardless. The code is as follows:

```
prop.table(polscidata)
```

- Then I made a table with the sums of the rows and columns and saved it in a new variable. This was mostly to check my work in the later steps

```
tablewmarg <- addmargins(polscidata)
```

- Then I made a new matrix to fill with expected values. I also made the column and row names the same as the previous matrix

```
{expected <- matrix(nrow = 2, ncol = 3)
colnames(expected) <- c("Not Stopped", "Bribe requested", " Stopped/
    given warning")
rownames(expected) <- c("Upper class", "Lower class")}
```

- For the actual "by hand" computation in R, I created a for loop that iterated through value of the matrix "polscidata". For each value, the loop saved the row sums, column sums, and total sum and computed a new value corresponding to the formula: (row sum/total sum)*column sum. Lastly, I added this to the empty matrix ("expected") that I made earlier

```
for (i in 1:nrow(polscidata)) {
   rowsum <- sum(polscidata[i, ])
for (j in 1:ncol(polscidata)) {
  colsum <- sum(polscidata [ ,j])
   totalsum <- sum(polscidata)
```

```
    expected[i,j] <- (rowsum/totalsum)*colsum
}
}
```

- Then, for clarity, I decided to change the name of my variable "polscidata" to "observed" as this aligns with the further computation

```
observed <- polscidata
```

- Next, I made a new variable "X2" to represent my test statistic. This variable is created to correspond with 0, allowing me to use it in a for loop. I then made a for loop that iterates through my matrices and sums together the observed values minus the expected values, squares this value, and then divides it by the expected values. $\sum\left(\frac{(f_{\text{observed}}-f_{\text{expected}})^2}{f_{\text{expected}}}\right)$ Then, after each iteration, it adds this value to my X2 variable, such that X2 ends up being the total sum. X2 then holds the value of my test statistic. The code is as follows:

```
X2 <- 0
for (i in 1:length(observed)) {
  value <- sum(((observed[i] - expected[i])^2)/expected[i])
  X2 <- X2 + value
}
```

- The code returns a value of **3.791168**. This is our test statistic

(b) Now calculate the p-value from the test statistic you just created (in R).[2] What do you conclude if $\alpha = 0.1$?

- To find the p-value, I used code from class. First I made a degrees of freedom variable, then made a new object "pvalue" using the pchisq function:

```
df <- (nrow(observed) - 1)*(ncol(observed) - 1)
```

```
pvalue <- pchisq(X2, df,lower.tail=FALSE)
```

- If alpha = 0.1, we fail to reject the null hypothesis because our p value of **0.1502306** is greater than the alpha value of 0.1. What this means is that we fail to reject the null hypothesis that solicitation of bribe and class are independent of each other.

[2]Remember frequency should be > 5 for all cells, but let's calculate the p-value here anyway.

(c) Calculate the standardized residuals for each cell and put them in the table below.

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.3220306	-1.641957	1.523026
Lower class	-0.3220306	1.641957	-1.523026

- For this, we know that the numerator of the formula of standardized residuals is ok so the numerator of the equation is $f_{expected} - f_{observed}$, so I made a new matrix called "unstandresid with these values using this code:

```
unstandresid <- observed - expected
```

- I then changed the name to "numerator" for the sake of the computation

```
numerator <- unstandresid
```

- Next, I made an empty matrix called denominator. Then, I combined the old for loop that I used to get the expected values to fill the matrix with denominators following this formula for the denominator: $\sqrt{f_{\text{expected}}(1 - \text{row prop.})(1 - \text{column prop.})}$ This code is as follows:

```
{denominator <- matrix(0, nrow = nrow(observed), ncol = ncol(observed
    ))
colnames(denominator) <- c("Not Stopped", "Bribe requested", "
    Stopped/given warning")
rownames(denominator) <- c("Upper class", "Lower class")}
```

```
for (i in 1:nrow(observed)) {
  rowsum <- sum(observed[i, ])
  for (j in 1:ncol(observed)) {
    colsum <- sum(observed [ ,j])
    totalsum <- sum(observed)
    denominator[i,j] <- ((1-(rowsum/totalsum))*
      (1-(colsum/totalsum))*
      (rowsum/totalsum)*colsum)^(1/2)
}
}
```

- This then gave me 2 matrices, 1 with the numerators and another with the denominators. This allowed me to put them together into a new matrix by dividing "numerator" by "denominator". I did this in the creation of a new matrix "stdresids"

```
stdresids <- numerator/denominator
```

- I then checked these values with the chisq.test function and I get the same results

```
chitest <- chisq.test(observed)
chitest$stdres
```

(d) How might the standardized residuals help you interpret the results?

- A residual is the difference between an observed and expected cell frequency. The residual is positive when the observed frequency exceeds the expected value under the null hypothesis. The residual is negative when the observed frequency is smaller than the expected value. The standardized residuals describe the pattern of the association among the cells. A large standardized residual provides evidence against independence (the null hypothesis) in that cell. Values below -3 or above +3 are strong evidence against independence in the given cell (Agresti and Finlay (p. 230)).
- In this case, none of the standardized residuals exceed -3 or +3, so we can conclude that there is no overly strong evidence against independence in any given cell. However, we can conclude that the cells corresponding to "bribe requested" and "class" as well as the cells corresponding to "stopped/given warning" and "class" are most significantly than what would be expected under the null hypothesis.
- In other words, bribes were requested less often for those of upper class and more often for those of lower class than would be expected if class and solicitation are said to be independent.
  - This is the most useful for determining the initial research question of whether officers were more or less likely to solicit a bribe from drivers depending on their class
- Furthermore, participants were stopped/given a warning more often for those of upper class and less often for those of lower class than would be expected if class and solicitation are said to be independent.
- Lastly, participants were not stopped more often for those of upper class and less often for those of lower class than would be expected if class and solicitation are said to be independent
- These residuals and their interpretation can help us interpret our results by allowing us to see that bribe solicitation seems to be influenced by class (note: I am saying influenced by as this is an experimental study with randomized controls, meaning that the language surrounding it can go beyond pure association). Particularly, we see the highest deviation from what would be expected under the null hypothesis in the "bribe requested" category, specifically that bribes were requested from those of lower class standing than would be expected under the null hypothesis. From this we can assume that those of police officers are more likely to solicit a bribe from someone who is of a lower class standing.

Question 2: Economics

Chattopadhyay and Duflo were interested in whether women promote different policies than men.[3] Answering this question with observational data is pretty difficult due to potential confounding problems (e.g. the districts that choose female politicians are likely to systematically differ in other aspects too). Hence, they exploit a randomized policy experiment in India, where since the mid-1990s, $\frac{1}{3}$ of village council heads have been randomly reserved for women. A subset of the data from West Bengal can be found at the following link: `https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv`

Each observation in the data set represents a village and there are two villages associated with one GP (i.e. a level of government is called "GP"). Figure 1 below shows the names and descriptions of the variables in the dataset. The authors hypothesize that female politicians are more likely to support policies female voters want. Researchers found that more women complain about the quality of drinking water than men. You need to estimate the effect of the reservation policy on the number of new or repaired drinking water facilities in the villages.

Figure 1: Names and description of variables from Chattopadhyay and Duflo (2004).

Name	Description
`GP`	An identifier for the Gram Panchayat (GP)
`village`	identifier for each village
`reserved`	binary variable indicating whether the GP was reserved for women leaders or not
`female`	binary variable indicating whether the GP had a female leader or not
`irrigation`	variable measuring the number of new or repaired irrigation facilities in the village since the reserve policy started
`water`	variable measuring the number of new or repaired drinking-water facilities in the village since the reserve policy started

[3] Chattopadhyay and Duflo. (2004). "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India. *Econometrica*. 72 (5), 1409-1443.

(a) State a null and alternative (two-tailed) hypothesis.

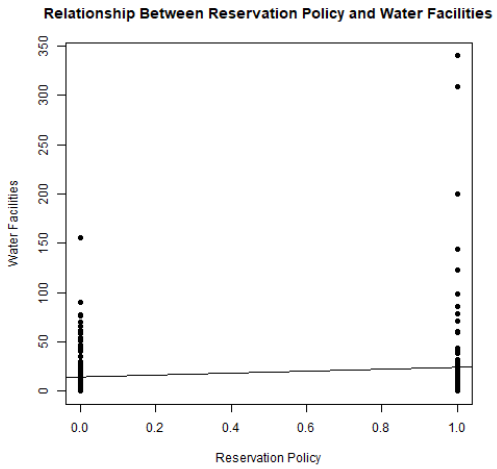
- Our linear model is $Y = \alpha + \beta X$, such that α is the Y-intercept, β is the slope of the line, Y is the outcome vector and X is the vector of the predictor (as seen in the lecture slides)
- As this research question is looking at the effect of the reservation policy on the number of new or repaired drinking water facilities, it makes most sense to look at the coefficient (β_1). The hypotheses for such is (note: the lecture slides referred to β as β_1 and α as β_0 so I will be using that notation in my hypotheses):
  - Null Hypothesis: $\beta_1 = 0$
  - Alternative Hypothesis: $\beta_1 \neq 0$
- The null hypotheses for the intercept is as follows:
  - Null Hypothesis: $\beta_0 = 0$
  - Alternative Hypothesis: $\beta_0 \neq 0$

(b) Run a bivariate regression to test this hypothesis in `R` (include your code!).

- First I created a dataframe with the csv file provided. Then I plotted the reserved variable by the water variable. The question is asking whether the reservation policy (recorded in the reserved variable) has an effect on the number of new or repaired drinking water facilities (recorded in the water variable). I also plotted the regression line using the code for "model" in the next section of code.

```
econdf <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/
    master/PREDICTION/women.csv", header=T)
```

```
png(file = "RegressionPlot.png")
plot(econdf$reserved, econdf$water, pch=16,col = c("black"),
     main="Relationship Between Reservation Policy and Water
    Facilities",
     xlab="Reservation Policy", ylab="Water Facilities")
abline(model)
dev.off()
```



- Next, I fit the linear regression model using the summary function. I did this two ways, as it was stated in tutorial that the latter way is better practice. Then I saved this model into an object named "model"

```
summary(lm(econdf$reserved~econdf$water))
```

```
summary(lm(water ~ reserved, data=econdf))
```

```
model <- summary(lm(water ~ reserved, data=econdf))
```

- I also did this all by hand to show that I was able to (using slides from lecture) and I got the same results. The code is as follows:

```
#bivariate regression by hand too
beta <- sum((econdf$water - mean(econdf$water)) *
              (econdf$reserved - mean(econdf$reserved))) /
              sum((econdf$reserved - mean(econdf$reserved))^2)
beta

alpha <- mean(econdf$water) - beta*mean(econdf$reserved)
alpha


#finding standard deviation estimate to plug in

#saving df1 as the same as model but without summary()
df1 <- lm(water ~ reserved, data = econdf)

sd_estimate <- sqrt(sum(resid(df1)^2)/
                      (dim(econdf)[1] - 2))
sd_estimate
#also can do it this way
sigma(lm(water ~ reserved, data = econdf))


#finding standard errors
```

```r
beta_se <- sd_estimate/sqrt(sum((econdf$reserved -
    mean(econdf$reserved))^2))
beta_se


alpha_se <- sd_estimate * sqrt((1 / dim(econdf)[1]) +
     (mean(econdf$reserved)^2 / sum((econdf$reserved -
      mean(econdf$reserved))^2)))
alpha_se

#finding p value
2*pt((beta - 0)/beta_se, dim(econdf)[1]-2, lower.tail = F)
2*pt((alpha - 0)/alpha_se, dim(econdf)[1]-2, lower.tail = F)

#checking this against model from earlier
model
```

(c) Interpret the coefficient estimate for reservation policy.

- The value of β_0 is **14.738**
- The value of β_1 is **9.252**
- The coefficient estimate (β_1) is stating that an increase of 1 reservation policy is associated with an average increase in new or repaired drinking water facilities in the villages of 9.252.
- However, this is not the best way to word this given that the reservation policy is coded as a binary variable. Rather, the change from no reservation policy to having a reservation policy is associated with an on average increase in new or repaired drinking water facilities in the villages of 9.252
  - The p-value for this value is **0.0197**, so we can reject the null hypothesis at the 0.05 level.
- Regarding β_0, the interpretion is: with no reservation policy, there is an average of 14.738 new or repaired drinking water facilities in the villages.
  - The p-value for this value is **4.22e-10**, so we can reject the null hypothesis at the 0.001 level.
- I think it is also important to note that this is a randomized experimental study, which also changes the interpretation of our results slightly. Under the assumption of causal inference, we would be able to say that **the reservation policy has an effect on the number of new or repaired drinking water facilities in the villages, such that the presence of a reservation policy causes an average increase of 9.25 new or repaired water facilities in the villages.**