Problem Set 2

Applied Stats/Quant Methods 1

Due: October 14, 2024

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 Monday October 14, 2024. 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`).

To calculate the Chi-squared statistic by hand, we first have to calculate the expected values for each respective observation. This can be found by dividing the row sum by the grand total and multiplying this by the column total.

```r
# initialise a matrix for the observed data
obs <- matrix(0, 2,3)

# add in the observed data
UC <- c(14,6,7)
LC <- c(7,7,1)
obs[1,] <- UC
obs[2,] <- LC
obs

# initialise new matrix for the expected values
exp <- matrix(0, 2,3)

# calculate the expected value for each cell in the first row and assign to the new empty matrix, then another loop for the second row
for (i in 1:ncol(obs)) {
  exp[1,i] <- (sum(obs[1,])/sum(obs))*(sum(obs[,i]))
}
for (i in 1:ncol(obs)) {
  exp[2,i] <- (sum(obs[2,])/sum(obs))*(sum(obs[,i]))
}
print(exp)
```

```
     [,1]     [,2]     [,3]
[1,] 13.5 8.357143 5.142857
[2,]  7.5 4.642857 2.857143
```

From here, the Chi-squared statistic is given by taking the average of each squared difference between the observed and expected values divided by the expected values.

```r
# we can calculate the chi-squared statistic by matrix subtraction and multiplication
chi2_stat <- sum(((obs - exp)^2)/(exp))
```

```
[1] 3.791168
```

(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 calculate the associated p-value for our test-statistic, we consult the Chi-squared PDF with the appropirate degrees of freedom

```r
# by consulting the Chi-squared PDF, we can find the associated p-value
pchisq(chi2_stat, df = (nrow(exp)-1)*(ncol(exp)), lower.tail = FALSE)
```

```
[1] 0.2849151
```

We can conclude that even with a non-stringent alpha of 0.1, we cannot reject the null hypothesis that the groups are independent.

[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.

To calculate the standardised/adjusted residuals we convert the difference between the expected and observed to units of standard error. For the Chi-squared distribution, the standard error is the square root of the product of each respective expected value by 1 minus the proportions for both the row and column totals

```r
# we already have enough information to calculate the numerator (observed minus expected values), which we can store as a difference matrix
diff <- (obs - exp) ## this is where the error is occurring, can't take one matrix away from the other
diff

# next, I will compute the standard error for each expected value
SE_mat <- matrix(0, 2,3)
gtot <-sum(obs)

# this loop attempts compute the row total and column total for each cell of the observation matrix, compute the standard error and output it to a new matrix
for (i in 1:nrow(SE_mat)) {
  for (j in 1:ncol(SE_mat)) {
    row_tot <- sum(obs[i,])
    col_tot <- sum(obs[,j])
    SE_mat[i,j] <- sqrt(exp[i,j] * (1-(row_tot/gtot)) * (1-(col_tot/gtot)))
  }
}
print(SE_mat)

# now finish the computation
residuals <- diff/SE_mat
round(residuals, 2)
```

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.32	-1.64	1.52
Lower class	-0.32	1.64	-1.52

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

These results are incorrect. The first indication of this is that the matrix is symmetric. Having spent the weekend trouble-shooting this, I know that the issue is with the loop before the residuals are calculated and not with the final simple matrix subtraction and division. Here is the standard error matrix after the loop has run:

```
            [,1]     [,2]      [,3]
[1,] 1.552648 1.43557 1.219377
[2,] 1.552648 1.43557 1.219377
```

This suggests the issue may be with the row indexing, either overwriting the results of the first row over the second or simply computing the first row twice. I have written i-j loops in MATLAB but it has been some time. I broke the process down into two separate loops, like what I used for question 1a. I noticed that some values were correct by comparing these results to calculations of the standard error done by hand, and therefore suspect that standard error function may be the source of the error or a second error.

```
for (i in 1:ncol(SE_mat)) {
  SE_mat[1,i] <- sqrt(
    exp[1,i] * (1-(sum(obs[1,])/gtot)) * (1-(sum(obs[,i])/gtot)))
}
for (i in 1:ncol(SE_mat)) {
  SE_mat[2,i] <- sqrt(
    exp[2,i] * (1-(sum(obs[2,])/gtot)) * (1-(sum(obs[,i])/gtot)))
}
print(SE_mat)
```

I also tried defining a specific standard error function and then using the apply function, but this proved more of an issue. I suspect loops were not necessary, and the entire procedure could be done with apply. I could also have just installed a package that calculated the residuals, but that would have been no fun! ChatGPT was useless and I should simply have started earlier and discussed the issue during office hours. Any suggestions, either for spotting the error in the single loops, the nested loop, or for using the apply function, would be very much appreciated!

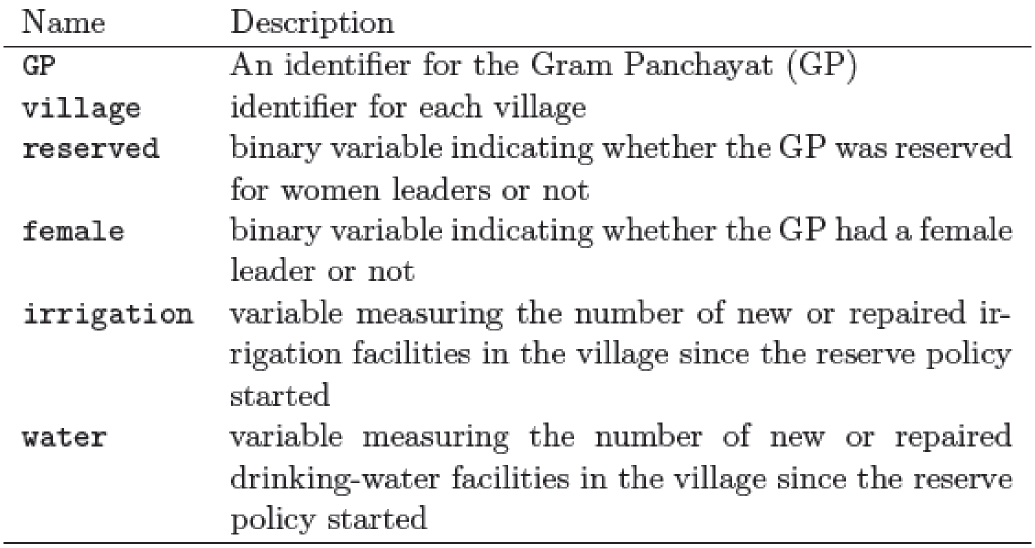
Lastly, in general the residuals help us interpret the distance between the observed and expected values in terms of standard error units rather than raw values.

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.

The null hypothesis here is that there is no correlation between male and female representatives (input variable) and the number of repairs to/new water facilities (response variable). The alternative hypothesis is that there does exist a correlation. As a two-sided test we are not testing whether there is an increase in response variable. This can be summarised as *H0: B-hat = 0 and Ha: B-hat != 0*

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

```r
lastlinelastline
# First inspect whether one third of GPs are women as stated
sum(df_raw$female)/nrow(df_raw)

# Note that here we have a binary input variable, 0 or 1, and a choice of two response variables: irrigation and water.
# We can run two regressions, one for each response variable, or combine their values.
# I will start with the water response variable:

png(file="regression.png")
plot(water~female,
      data = df_raw,
      xlab="Male = 0, Female = 1",
      ylab="New/repair incidences")
lm(water~female, data = df_raw)
abline(lm(water~female, data = df_raw))
dev.off()
cor.test(df_raw$female, df_raw$water)


# to look at the effect on irrigation alone
plot(irrigation~female, data = df_raw)
lm(irrigation~female, data = df_raw)
abline(lm(irrigation~female, data = df_raw))
cor.test(df_raw$female, df_raw$irrigation)


# to look at the combined effect
Y <- rowSums(df_raw[,5:6]) # combine the last two columns into a new vector
df_simple <- df_raw # create a new matrix
df_simple[,5] <- Y # overwrite what's going to be last column
df_simple <- df_simple[,-6] # remove the last column
names(df_simple)[5] = "response" # rename the last column

plot(response~female, data = df_simple)
```

```
lm(response~female, data = df_simple)
abline(lm(response~female, data = df_simple))
cor.test(df_simple$female, df_simple$response)

summary(lm(response~female, data = df_simple))
```

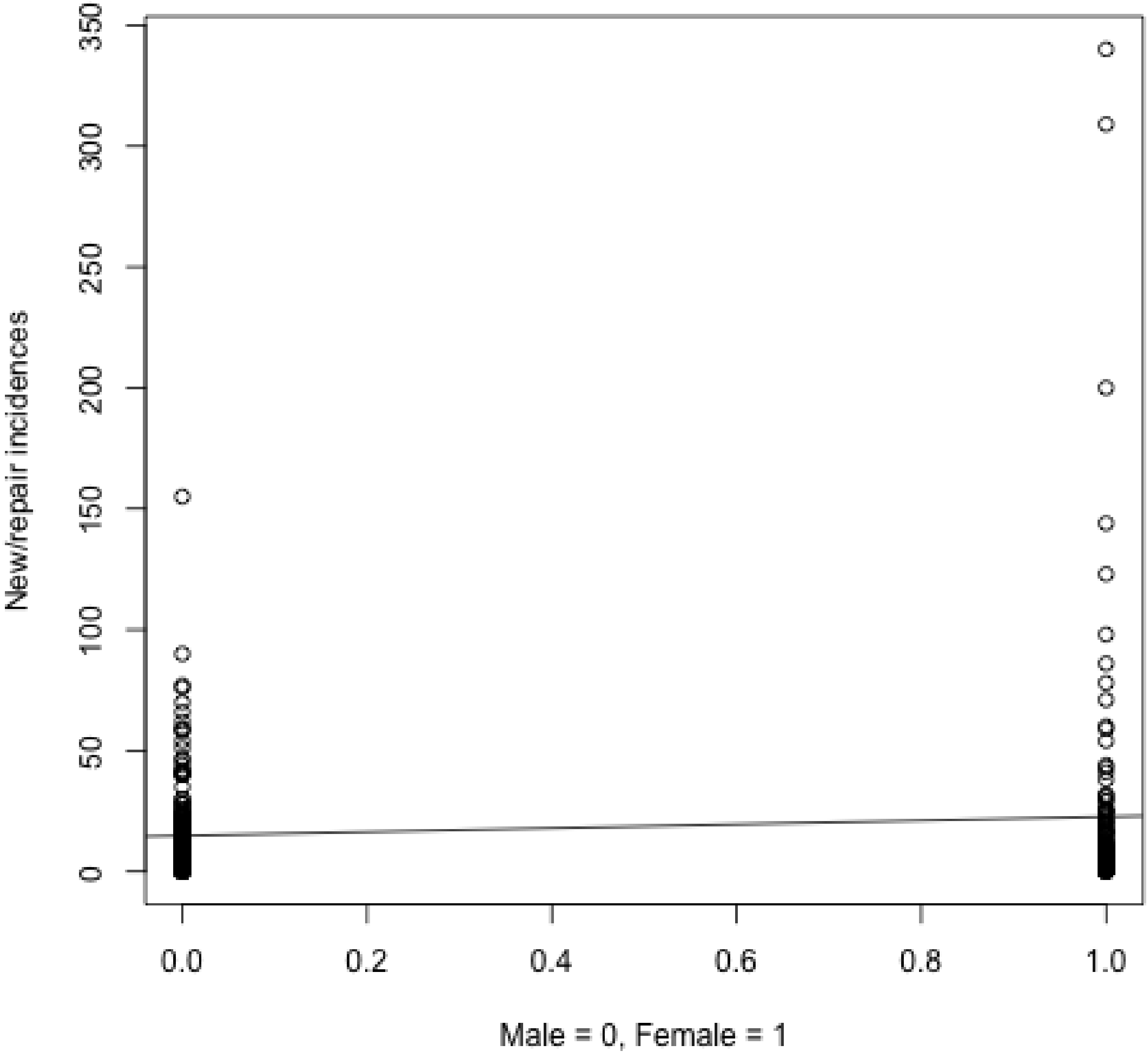
The results of the linear regression run on the merged response variables, and the correlation test, are as follows

```
Coefficients:
(Intercept)       female
14.813        7.864


 Pearson’s product-moment correlation

data:  df_raw$female and df_raw$water
t = 2.0491, df = 320, p-value = 0.04126
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.004563936 0.220363186
sample estimates:
cor
0.1138057
```

Figure 2: Basic plot of binary input variable regression



(c) Interpret the coefficient estimate for reservation policy.

The value we get for the correlation coefficient is 0.11, with a p-value of 0.04. Thus, if we were conducting a hypothesis test with an alpha of 0.05, we would just about be able to reject the null hypothesis that, specifically, there is no statistically significant effect of the reservation policicity on number of new or repaired water facilities in the sample provided. The alternative hypothesis, that there is an effect of the reservation policy, could be accepted at this alpha threshold. The spread around this p-value, as measured by the 95 percent confidence interval is quite large however, which we could interpreted as noisy data and would lend caution to making strong claims about just how effective the policy has been.

The data can also be aggregated to look at the merged response variables, namely changes to both irrigation and water facilities, which dampens the correlation and suggests specificity of the policy to water-drinking facilities rather than a general effect on water-facilities including drinking water and irrigation.