Problem Set 2

Applied Stats/Quant Methods 1

Due: October 15, 2023

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).

```
#creating sums and adding to the table:
df2$Total <- c(sum(df2[1,1:3]), sum(df2[2,1:3]))
df2[nrow(df2) + 1,] <- c(sum(df2$Not_Stopped), sum(df2$Bribe_requested),
    sum(df2$Stopped_or_given_warning), sum(df2$Total))
rownames(df2) <- c("upper_class", "lower_class", "Total")

## manual calc ((row total/grand total)*column total),
# order by rows
f1e <- df2[1,4] / df2[3,4] * df2[3,1] #13.5
f2e <-df2[2,4] / df2[3,4] * df2[3,1] #7.5
f3e <-df2[1,4] / df2[3,4] * df2[3,2] # 8.36
f4e <-df2[2,4] / df2[3,4] * df2[3,2] # 4.64
f5e <- df2[1,4] / df2[3,4] * df2[3,3] # 5.14
f6e <-df2[2,4] / df2[3,4] * df2[3,3] # 2.86

### calculating the statistic:
chisq <- sum((df2[1,1] - f1e)^2/f1e + (df2[2,1] - f2e)^2/f2e + (df2[1,2]
    - f3e)^2/f3e +(df2[2,2] - f4e)^2/f4e + (df2[1,3] - f5e)^2/f5e + (df2
    [2,3] - f6e)^2/f6e)
print(chisq) #3.791
```

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

```
df_df <- (nrow(df2) - 1)*(ncol(df2)-1)
print(df_df) # 6 degrees of freedom
p <- pchisq(3.79, df=6, lower.tail=F)
print(p) # 0.705
```

As the p-value equals 0.706, at the $\alpha = 0.1$ we cannot reject the null hypothesis that the class of the drivers was related to solicitation in bribes.

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

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

```r
z11 <- (df2[1,1] - f1e) / sqrt(f1e*(1-(df2[1,4]/df2[3,4]))*(1-(df2[3,1]/df2[3,4])))
z21 <- (df2[2,1] - f2e) / sqrt(f2e*(1-(df2[2,4]/df2[3,4]))*(1-(df2[3,1]/df2[3,4])))
z12 <- (df2[1,2] - f3e) / sqrt(f3e*(1-(df2[1,4]/df2[3,4]))*(1-(df2[3,2]/df2[3,4])))
z22 <- (df2[2,2] - f4e) / sqrt(f4e*(1-(df2[2,4]/df2[3,4]))*(1-(df2[3,2]/df2[3,4])))
z13 <- (df2[1,3] - f5e) / sqrt(f5e*(1-(df2[1,4]/df2[3,4]))*(1-(df2[3,3]/df2[3,4])))
z23 <- (df2[2,3] - f6e) / sqrt(f6e*(1-(df2[2,4]/df2[3,4]))*(1-(df2[3,3]/df2[3,4])))

z11 <- (df2[1,1] - f1e) / sqrt(f1e)
z21 <- (df2[2,1] - f2e) / sqrt(f2e)
z12 <- (df2[1,2] - f3e) / sqrt(f3e)
z22 <- (df2[2,2] - f4e) / sqrt(f4e)
z13 <- (df2[1,3] - f5e) / sqrt(f5e)
z23 <- (df2[2,3] - f6e) / sqrt(f6e)

df_res <- data.frame(Not_Stopped <- c(z11, z21),
                     Bribe_requested <- c(z12, z22),
                     Stopped_or_given_warning <- c(z13, z23))

colnames(df_res) <- c("Not_Stopped", "Bribe_requested", "Stopped_or_given_warning")
rownames(df_res) <- c("upper_class", "lower_class")
library(xtable)
setwd("/Users/vv/Downloads/StatsI_Fall2023_oct4/problemSets/PS02/template")
print(xtable(df_res, type = "latex"), file = "residuals.tex")
```

	Not_Stopped	Bribe_requested	Stopped_or_given_warning
upper_class	0.14	-0.82	0.82
lower_class	-0.18	1.09	-1.10

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

The standardized residuals show how far away is each observed value from "expectation". Worst predictions are where residuals are closer to |1.96| – whether lower class was requested bribe or stopped/given warning. However, they are still far from letting us to assume that there is a linkage between these two variables.

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.

H_0: $\beta_1 = 0$, in other words, there is no statistically significant relationship between the reservation policy and the number of new or repaired drinking water facilities in the villages.
H_a: $\beta_1 \neq 0$, or there is a statistically significant relationship between the reservation policy and the number of new or repaired drinking water facilities in the villages.

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

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

library(stargazer)
stargazer(model, title = "Reservation policies and improved drinking water facilities",
          out = "PS02_regression_Babaian.tex")
```

Table 1: Reservation policies and improved drinking water facilities

	Dependent variable:
	water
reserved	9.252**
	(3.948)
Constant	14.738***
	(2.286)
Observations	322
R^2	0.017
Adjusted R^2	0.014
Residual Std. Error	33.446 (df = 320)
F Statistic	5.493** (df = 1; 320)
Note:	*p<0.1; **p<0.05; ***p<0.01

(c) Interpret the coefficient estimate for reservation policy.

At the 99% confidence level, we reject H_0: for the villages where GP was reserved for women leaders, the number of new or repaired drinking-water facilities, on average increased by 9.25.