Problem Set 2

Stefan Keel
Applied Stats/Quant Methods 1

Due: October 15, 2023

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).

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	13.5	8.357143	5.142857
Lower class	7.5	4.642857	2.857143

To calculate the χ^2 test I first created the matrix and then used this code to calculate and fill the new matrix with the results for the expected values for ex (xexp).

```
for (i in 1 : nrow(x)) {
for (j in 1 : ncol(x)) {
xexp[i, j] <- (sum(x[i, ]) * sum(x[, j])) / sum(x)  }
}
```

With these table I then executed the chi-test which yielded in these values:

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.01851852	0.6648352	0.6706349
Lower class	0.03333333	1.1967033	1.2071429

Again an empty table was created and this code was used to fill it with the chi test values:

```
for (i in 1:nrow(chi)) {
for (j in 1:ncol(chi)) {
chi[i, j] <- ((x[i, j] - xexp[i, j])**2) / xexp[i, j]   }
}
```

Summing the cells in the table than produces the chi-squared: $\chi^2 = \mathbf{3.791168}$

This code was used:

```
chi_sqrd <- sum(chi)
print(chi_sqrd)
```

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

The p-value from the test statistic is **0.1502**. Taking that alpha = 0.1. I conclude that we have no sufficient evidence to exclude that the variables are statistically independent. There is evidence that the variables could be dependent because the p-value is bigger than 0.1.

This was calculated in this way:

```
p_chi_sqrd <- pchisq(chi_sqrd, df = (ncol(x) - 1) * (nrow(x) - 1),
lower.tail=FALSE)
print(p_chi_sqrd)

#or the easy way
chi_test <- chisq.test(x)
print(chi_test)
```

[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

To get the standardised residuals I used this code:

```
#getting the standardised residuals from the chi_test
chi_test$stdres

> chi_test$stdres
not stopped bribe requested stopped/given warning
upper class   0.3220306       -1.641957              1.523026
lower class  -0.3220306        1.641957             -1.523026
```

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

”A standardized residual is the raw residual divided by an estimate of the standard deviation of the residuals. It’s a measure of the strength of the difference between observed and expected values. ” (Quote from here: www.isixsigma.com). They help identifying outliers which, in this case, is not the case as all values fall no further than 2 standard deviations. Although they fall further away - it still is within range. The Std. Residual (Not Stopped) of 0.322 is quite good.

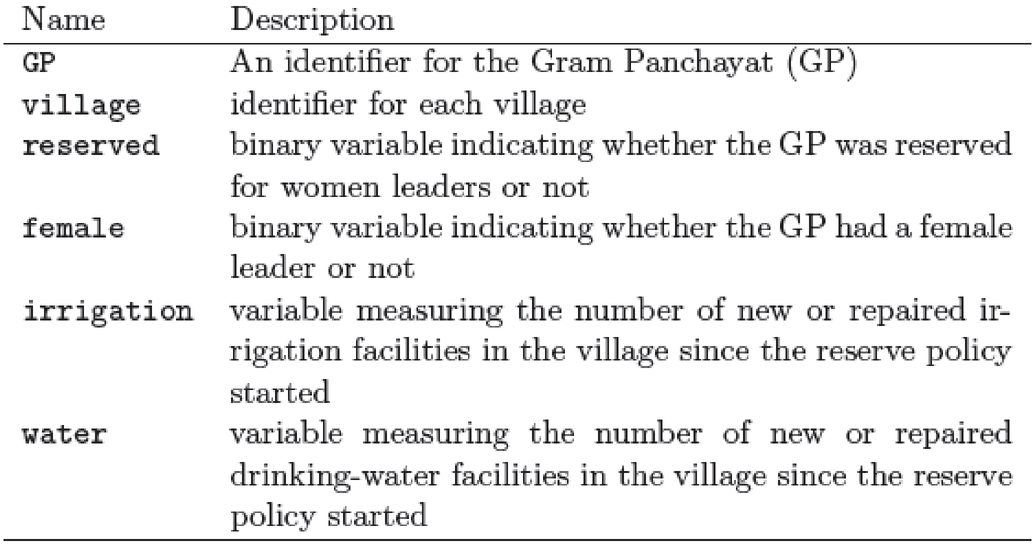
Considering my still developing knowledge and extensive internet research on std. residual interpretation I would like to suggest that ”bribe requested” and ”stopped/warning given” fall further away from our predicted line. This might be an indicator but most likely not statistically significant.

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.

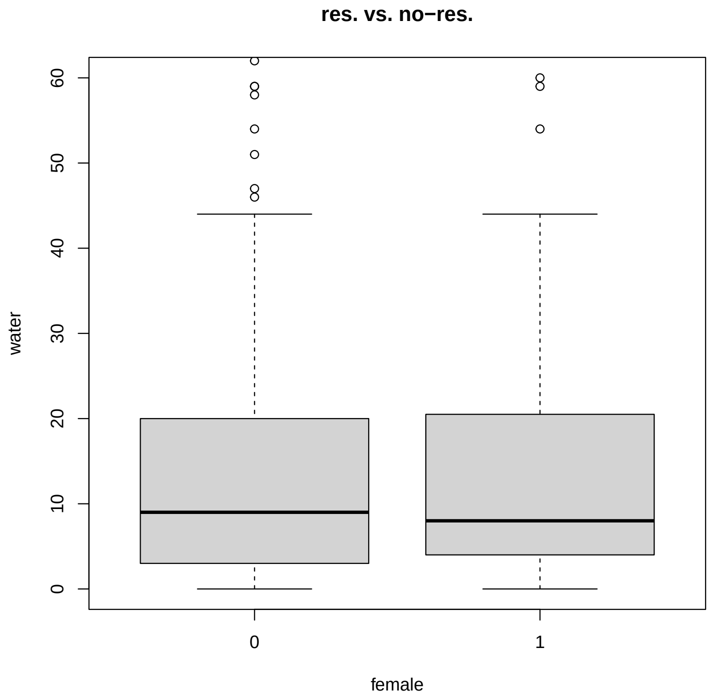
(a) State a null and alternative (two-tailed) hypothesis.

1. HO: The reservation policy for women has no relation to more repaired drinking water facilities.
2. HA: The reservation policy for women has a relation to more repaired drinking water facilities.

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

I started of, after checking the dataframe in R, by boxplotting the water and reservation policy to get a vague understanding of any differences etc.

```
ls(df)  #getting an overview
View(df) #getting an overview
boxplot(water ~ female, data = df, main = "res. vs. no-res.", ylim = c())
#the outliers make the boxplot quite unreadable =  ylim will be set
```



This is the boxplot:

The difference between reserved (1) and unreserved (0) is not strikingly if we look at the mean. Furthermore (for better visualisation) I excluded some outliers by setting the max. scale of Y to 60. Nevertheless, the correlation test had to be run to do the hypothesis test.

```
correlation <- cor(df$water, df$reserved, method = "pearson")
#correlation function
correlation

corrtest <- cor.test(df$water, df$reserved) #correlation test for more info
corrtest

summary(lm(df$water~df$reserved)) #getting the crucial data
```

By running the summary for lm we get these values.

```
Call:lm(formula = df$water ~ df$reserved)
Residuals:
Min      1Q
Median      3Q
Max -23.991 -14.738  -7.865   2.262 316.009
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   14.738      2.286   6.446 4.22e-10 ***
df$reserved    9.252      3.948   2.344   0.0197 *

---Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 33.45 on 320 degrees of freedom
Multiple R-squared:  0.01688,Adjusted R-squared:  0.0138
F-statistic: 5.493 on 1 and 320 DF,  p-value: 0.0197
```

Therefore, we would also reject the H0.

Following values indications were calculated **Correlation: 0.1299079, and the p-value: 0.0197** as well as the confidence intervall with 95percent of: **lower: 0.02090616 higher: 0.23585751**.

As the p-value of 0.02 is lower than 0.05 we have evidence to reject the H0 and to accept HA.

(c) Interpret the coefficient estimate for reservation policy.

The correlation of **0.1299079** is small but it shows a positive correlation between reservation for women and established water facilities. The y-intercept at 0 is 14.738 water installations. By increasing by 1 reservation, additional 9.252 water installations, on average, may be installed. Degrees of freedom indicate that there is more power and therefore the result can be significant (www.sites.utexas.edu)

Concluding: We might have evidence that there is a relation between reservation policy for women and construction of drinking water facilities.