Problem Set 2

Samanta Nedzinskaite

October 16, 2022

Question 1 (40 points): 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.

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

[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.

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

```
tab <- matrix(c(14, 6, 7, 7, 7, 1), byrow = TRUE, nrow = 2, ncol = 3)
tab

#Step 1: finding expected frequencies
expected_frequencies <- data.frame()

for (i in 1:2){
  expected_frequencies[i,1] <- (sum(tab[i,]) * sum(tab[,1])) / sum(tab)
  expected_frequencies[i,2] <- (sum(tab[i,]) * sum(tab[,2])) / sum(tab)
  expected_frequencies[i,3] <- (sum(tab[i,]) * sum(tab[,3])) / sum(tab)
}
expected_frequencies

rownames = c("Upper Class", "Lower Class")
colnames = c("Not Stopped", "Bribe requested", "Stopped/given warning")
names(expected_frequencies)[1:3] <- colnames
rownames(expected_frequencies)[1:2] <- rownames
expected_frequencies
```

Result:

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

```
#Step 2: finding test statistic

test_statistic <- sum(((tab - expected_frequencies)^2/expected_
    frequencies))
test_statistic
```

[1] 3.791168

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

```
df <- ((nrow(tab) -1)*(ncol(tab) -1))
alpha <- 0.1
pvalue <- pchisq(test_statistic , df = 2, lower.tail=FALSE)
pvalue
pvalue < alpha
```

[1] 0.1502306

[1] FALSE

We fail to reject the null hypothesis because our p-value is not less than our $\alpha = 0.1$. This suggests that we do not have enough evidence from our sample to claim that officers were more or less likely to solicit a bribe from drivers depending on their class.

[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.

```
stand_residual <- data.frame()

for (i in 1:2){
  stand_residual[i,1] <- (tab[i,1] - expected_frequencies[i, 1])/
    sqrt(expected_frequencies[i,1]*(1-sum(tab[i,])/sum(tab))*(1-sum(tab
    [,1])/sum(tab)))
  stand_residual[i,2] <- (tab[i,2] - expected_frequencies[i, 2])/
    sqrt(expected_frequencies[i,2]*(1-sum(tab[i,])/sum(tab))*(1-sum(tab
    [,2])/sum(tab)))
  stand_residual[i,3] <- (tab[i,3] - expected_frequencies[i, 3])/
    sqrt(expected_frequencies[i,3]*(1-sum(tab[i,])/sum(tab))*(1-sum(tab
    [,3])/sum(tab)))
}

stand_residual
```

	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?

The standardizied residual measures the significance of the difference between the observed and the expected values in a regression model. It is useful in interpreting results because it can help to identify outliers in the model. Typically, a standardized residual with a value above 3 would be considered as an outlier in the model. Outliers are a problem for establishing statistical confidence, because the presence of outliers increase the variability in our data.

Question 2 (40 points): 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

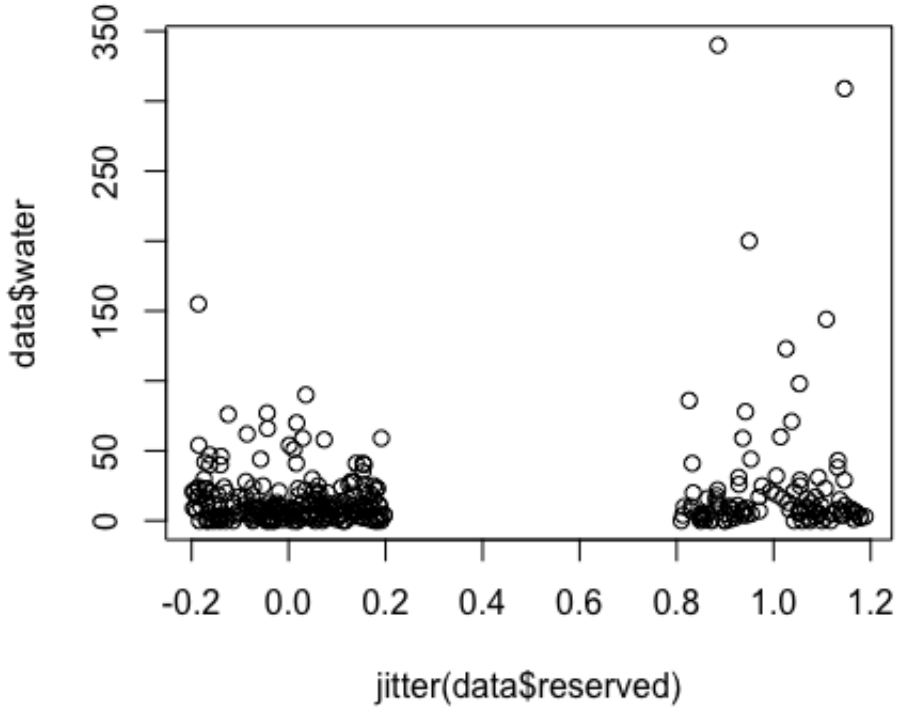
[3] Chattopadhyay and Duflo. (2004). "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India. *Econometrica*. 72 (5), 1409-1443.

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

Null hypothesis: there is no relationship between the reservation policy and the number of new or repaired drinking water facilities in the villages.

Alternative hypothesis: in the villages where the reservation policy is active, the number of new or repaired drinking water facilities either increased or decreased.

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




```r
data$D <- data$reserved == 1
dummy_model <- lm(water ~ D, data = data)
summary(dummy_model)
```

```
Call:
lm(formula = water ~ D, data = data)

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 ***
DTRUE          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
```

(c) Interpret the coefficient estimate for reservation policy.

From the summary of our regression model above, we can see that the expected number of mean water facilities available in districts where reservation policy is in place (coded as zero, and in our dummy model as TRUE) is 14.7+9.2 = **23.9**. Meanwhile, in villages where the reservation policy is not in place (coded as 0, and in our dummy model as FALSE), the mean number of expected available water facilities is **14.7.**

```
                2.5 %   97.5 %
(Intercept) 10.240240 19.23640
DTRUE        1.485608 17.01924
```

If we run a ninety-fife percent confidence interval for the true difference in means between the two groups in our model, we can reject the null hypothesis that there is no difference in group means, since 0 lies outside [1.485608, 17.01924]