Applied Stats - Problem Set 2

Yana Konshyna

October 14, 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).

Code in R:

```r
# Creating matrix
regressMat <- matrix(c(14, 6, 7, 7, 7, 1), nrow = 2, byrow = T)

# Calculating f expected values ”by hand”
#The first row of the table
fe11 <- 27*21/42
fe12 <- 27*13/42
fe13 <- 27*8/42

#The second row of the table
fe21 <- 15*21/42
fe22 <- 15*13/42
fe23 <- 15*8/42

# Calculate the chi-squared test statistic
chi_squared <- (regressMat[1,1] - fe11)^2/fe11 + (regressMat[1,2] - fe12)^2/fe12 +
  (regressMat[1,3] - fe13)^2/fe13 + (regressMat[2,1] - fe21)^2/fe21 +
  (regressMat[2,2] - fe22)^2/fe22 + (regressMat[2,3] - fe23)^2/fe23
chi_squared
```

Output:

```
> chi_squared
[1] 3.791168
```

Conclusion: When the fo (observed value) and fe (expected value) tend to be close for each cell of the table and chi-squared is relatively small, there isn't enough evidence to reject the H0.

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

Code in R:

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

```r
p_value <- pchisq(chi_squared, df = (nrow(regressMat) - 1)* (ncol(regressMat) - 1), lower.tail = FALSE)
p_value
```

Output:

```
> p_value
[1] 0.1502306
```

Conclusion: The p-value = 0.1502306 bigger than $\alpha = 0.1$. That means there isn’t enough evidence to reject null hypothesis (H0). The variables are independent. Thus, there isn’t strong evidence to suggest that the difference in police interaction(”being stopped”, ”not being stopped”, ”bribe requested”) between ”Upper class” and ”Lower class” drivers is statistically significant.

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

Code in R:

```r
result <- chisq.test(regressMat)
stand_residuals <- result$stdres
table <- as.table(stand_residuals)
colnames(table) <- c("Not Stopped", "Bribe requested", "Stopped/given warning")
rownames(table) <- c("Upper Class", "Lower class")
print(table)
```

Output:

```
> print(table)
            Not Stopped   Bribe requested   Stopped/given warning
Upper Class   0.3220306        -1.6419565               1.5230259
Lower class  -0.3220306         1.6419565              -1.5230259
```

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.3220306	-1.6419565	1.5230259
Lower class	-0.3220306	1.6419565	-1.5230259

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

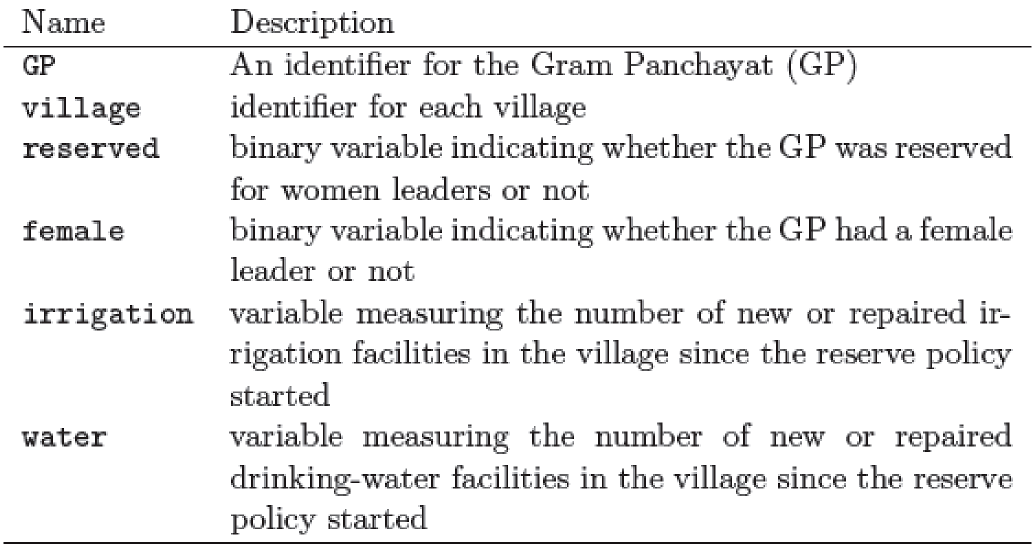
Conclusion: Positive residual results show that the observed values are larger than the predicted values. A negative residual shows that the observed values are smaller than the predicted values. The smaller the absolute value of the residual, the better the prediction is, since the predicted values are closer to the observed values. According to the standardized residuals, "Upper class" drivers are more likely to be stopped and warned, while "Lower class" drivers are more likely to be solicited for bribes.

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.

H0 = There is no effect between the reservation policy in villages where council heads have been randomly reserved for women and the number of new or repaired drinking water facilities in the villages (there is no relationship between independent variable ”reserved” X and dependent variable ”water” Y.) Ha = there is a relationship between X and Y.

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

Code in R:

```r
# Viewing csv file with data and assigning data to vector df
df <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv")
View(df)

# Perform the bivariate regression
model <- lm(water ~ reserved, data = df)

# Summarize the regression results
summary(model)
```

Output:

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

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 ***
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
```

(c) Interpret the coefficient estimate for reservation policy.

Conclusion: The p-value = 0.0197 is less than 0.05, so there isn't enough evidence to not reject H0. Therefore, variables ”water” and ”reserved” are dependent. There is an effect between the reservation policy in villages where council heads have been randomly reserved for women and the number of new or repaired drinking water facilities in the villages. The coefficient estimate 9.252 for ”reserved” variables means that, on average, compared to villages which aren't reserved for women council heads, the villages whose council heads have been randomly reserved for women, will have 9.252 more new or repaired drinking water facilities.