Problem Set 2

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).

```
observed <- matrix(c(14, 6, 7, 7, 7, 1), nrow = 2, ncol = 3, byrow = TRUE
    )
colnames(observed) <- c("Not Stopped", "Bribe requested", "Stopped/given
    warning")
rownames(observed) <- c("Upper class", "Lower class")
row_marginals <- rowSums(observed)
col_marginals <- colSums(observed)
total_frequency <- sum(observed)
for (i in 1:2) {
  for (j in 1:3) {
    expected[i, j] <- (row_marginals[i] * col_marginals[j]) / total_
    frequency
  }
}
chi_squared <- sum((observed - expected)^2 / expected)
print(chi_squared)
```

[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(observed) - 1) * (ncol(observed) - 1)
p_value <- 1 - pchisq(chi_squared, df)
cat("P-value:", p_value)
```

P-value: 0.1502306
We hypothesize that H0:The likelihood of police soliciting bribes is class-independent

H1:The likelihood of a police officer soliciting a bribe is related to class

If $\alpha = 0.1$, this means that the significance level of the hypothesis test is 0.1. since the resulting p-value has a value of $0.1502306 > 0.1$,

We do not reject H0 that the likelihood of police officers soliciting bribes is unrelated to 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.

```
  for (j in 1:3) {
    standardized_residuals[i, j] <- (observed[i, j] - expected[i,j])/sqrt
   (expected[i,j])
  }
}
print(standardized_residuals)
#d
```

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.1360828	-0.8153742	0.818923
Lower class	-0.1825742	1.0939393	-1.098701

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

(Upper, Not Stopped) is 0.1360828, which means that in the "Upper" category, the observed frequency of "Not Stopped" is slightly higher than the expected frequency, but the difference is not particularly significant.
(Upper, Bribe requested) is -0.8153742, which means that the number of observations of "Bribe requested" in the "Upper" category is significantly lower than the expected frequency, indicating that it occurs less often in this combination.
(Upper, Stopped/given warning) is 0.818923, which means that the observed frequency of "Stopped/given warning" is slightly higher than the expected frequency in the "upper" category, but the difference is not particularly significant.

For the "lower" category, the interpretation of the standardized residuals is similar to that of the upper category.Overall, no data set had a significant effect on the results of the chi-square test.

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.

Null assumption(H0): The reservation policy has no impact on the number of new or rehabilitated drinking water facilities in the village.

Alternative Hypothesis (H1): The reservation policy have an impact on the number of new or rehabilitated drinking water facilities in villages

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

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

Residuals:
Min 1Q Median 3Q Max
-23.991 -14.738 -7.865 2.262 316.009
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

	Estimate Std.	Error	t value	Pr(>|t|)
Intercept	14.738	2.286	6.446	4.22e-10***
reservation	9.252	3.948	2.344	0.0197*

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.

Intercept: The Intercept represents the estimated average number of water facilities when the reservation variable is 0. In this model, the Intercept estimate is 14.738. In this model, the Intercept is estimated to be 14.738, which means that when the reservation variable is 0 (i.e., there is no reservation policy), the average number of water facilities is approximately 14.738.

Coefficient estimate for RESERVATION: In this model, the coefficient estimate for the RESERVATION variable is 9.252. this indicates the effect of the reservation policy on the number of water facilities. The coefficient estimate tells us that each unit increase in reservation is associated with an increase in the number of water facilities of about 9.252 units, holding other factors constant.

t-value and p-value: t-value and p-value are used to measure the significance of the coefficient estimates. In this model, the t-value of the coefficient estimate for the RESERVATION variable is 2.344 and the p-value is 0.0197. This means that the effect of the reservation policy on the number of water resource facilities is statistically significant because the p-value is less than the level of significance typically used (usually 0.05).

Overall, the coefficient estimate has a RESERVATION variable of 9.252, a t-value of 2.344, and a p-value of 0.0197, which implies that the effect of the reservation policy on the number of water facilities is statistically significant. However, the low R-squared value indicates that the model fit is relatively low and the explanatory power of the reservation policy on the number of water facilities is weak.