Answer Key: Problem Set 2

Applied Stats/Quant Methods 1

Jeffrey Ziegler

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 Thursday October 23, 2025. No late assignments will be accepted.*

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 study, confederate made illegal left turns across traffic to draw the attention of the police officers. Two of the confederates were upper class drivers and two were lower class drivers. 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, Brian J, Paul Lagunes, and Atheendar 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 (even better if you can do "by hand" in R).*

$$\text{Expected} = \frac{\sum_{\text{Row}} * \sum_{\text{Column}}}{\sum_{\text{N}}}$$

$$\chi^2 = \sum_{N} \frac{\text{Observed}_i - \text{Expected}_i}{\text{Expected}_i}$$

Let's first try by ourselves:

```
# create matrix to conduct chi-square test
trafficViolations <- matrix(c(14, 6, 7, 7, 7, 1), byrow=T, nrow=2)
rownames(trafficViolations) <- c("Upper class", "Lower class")
colnames(trafficViolations) <- c("Not stopped", "Bribe", "Stopped/warned")
# by hand approach
# create function from chi-square test github.io
byHandChiSquare <- function(table){
  # turn into table
  observedValues <- as.table(table)
  # create sums (row, column, and total)
  grandSum <- sum(observedValues)
  sumRow <- rowSums(observedValues)
  sumCol <- colSums(observedValues)
  # calculate expected values for each observation
  # check "?outer" to see that this takes the outer product
  # of the row and col sum divided by the total sum
  expectedValues <- outer(sumRow, sumCol, "*") / grandSum
  v <- function(r, c, n) c * r * (n - r) * (n - c)/n^3
  V <- outer(sumRow, sumCol, v, grandSum)

  dimnames(expectedValues) <- dimnames(observedValues)
  # create function that calculates each cell residual variance
  # essentially formula on p. 225 in Agresti and Finlay(2009)
  test_statistic <- sum((abs(table - expectedValues))^2 / expectedValues)
  df <- (nrow(observedValues) - 1L) * (ncol(observedValues) - 1L)
  p_value <- pchisq(test_statistic, df, lower.tail = FALSE)
  adjusted_residuals <- (observedValues - expectedValues)/sqrt(
    expectedValues * (1-sumRow/grandSum) * (1-sumCol/grandSum))
  standardized_residuals <- (observedValues - expectedValues)/sqrt(V)
  # return values
  return(list(statistic = test_statistic,
              df = df,
              p.value = p_value,
              observed = observedValues,
              expected = expectedValues,
              adj_res = adjusted_residuals,
              std_res = standardized_residuals))
}
byHandChiSquare(table=trafficViolations)
```

```
$statistic
[1] 3.791168

$df
[1] 2

$p.value
[1] 0.1502306

$observed
            Not_Stopped Bribe_Requested Stopped_Given_Warning
Upper_Class          14               6                     7
Lower_Class           7               7                     1

$expected
            Not_Stopped Bribe_Requested Stopped_Given_Warning
Upper_Class        13.5        8.357143              5.142857
Lower_Class         7.5        4.642857              2.857143

$adj_res
            Not_Stopped Bribe_Requested Stopped_Given_Warning
Upper_Class   0.3220306      -1.5164259             1.6491029
Lower_Class  -0.2740361       1.9295276            -1.5230259

$std_res
            Not_Stopped Bribe_Requested Stopped_Given_Warning
Upper_Class   0.3220306      -1.6419565             1.5230259
Lower_Class  -0.3220306       1.6419565            -1.5230259
```

Now we can check to make sure:

```
# run chi square test with built in function
chisq.test(trafficViolations)
```

```
Pearson's Chi-squared test

data:trafficViolations
X-squared=3.7912, df=2, p-value=0.1502
```

(b) *Now calculate the p-value (in R).*[2] *What do you conclude if* $\alpha = .1$*?*

`pchisq(3.79, df = (2-1)*(3-1), lower.tail = FALSE)` $= 0.1502306$

P-value checks out to our "hand" calculation and the built in function. Cannot reject the null that the two variables of interest are independent.

[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.*

We can do this by hand (see above function), or the standardized residuals are stored in the `chisq.test` object. We're reporting the standardized residuals, `(observed - expected) / sqrt(V)`, where V is the residual cell variance (Agresti, 2007, section 2.4.5 for the case where x is a matrix, $n * p * (1 - p)$ otherwise).

```
# use function to extract standardized residuals
chisq.test(trafficViolations)$stdres
```

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.322	-1.642	1.523
Lower class	-0.322	1.642	-1.523

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

From the frequency table, it is already clear that there is no obvious pattern for a relationship between rows and columns. Further, the standardized residuals turn out to be quite small, which only supports us to be more confident about the lack of the dependency relationship. None of the standardized residuals indicate any of the cells are more or less than we would expect if the two variables were independent. Nevertheless, they do not tell us much, we need the chi-squared test to make conclusions in either case, i.e. whether variables are dependent or not.

Question 2 (20 points): Economics

Chattopadhyay and Duflo were interested in studying the causal effect of having female politicians on policy outcomes.[3] *Do women promote different policies than men? 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 "women.csv" 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 will be asked 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] Raghabendra Chattopadhyay and Esther Duflo. (2004). "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India. Econometrica, Vol. 72, No. 5, pp. 1409-1443.

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

Null: Having reserved seats for female politicians does not change the number drinking water facilities in the villages.

Alternative: The reservation policy has an effect on policy outcomes.

$$H_0 : \beta = 0$$
$$H_A : \beta \neq 0$$

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

After we load our dataset into our working environment, we execute our regression model in which the number of new or repaired water facilities is explained by whether there are reserved seats for female leaders. We then investigate the estimated coefficients of the model using `summary()`.

```
# read in women data from online .csv
women <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/
    master/PREDICTION/women.csv")
# run regression model with water regressed on whether there are reserved
    seats for women
regression_model_problem2 <- lm(water ~ reserved, data=women)
# get summary of model with coefficient estimates
summary(regression_model_problem2)
```

```
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.*

Having reserved seats for female politicians increase the number drinking water facilities in the villages, by 9.2 units. The estimated coefficient is statistically differentiable from zero at the $\alpha = 0.05$ level because the p-value < 0.05 (≈ 0.02).