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 in `.pdf` form.*
- *This problem set is due before 8:00 on Friday October 8, 2021. No late assignments will be accepted.*
- *Total available points for this homework is 100.*

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:

```r
# 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:

```r
# 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).

```r
# 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()`.

```r
# 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 ($\approx$0.02).

Question 3 (40 points): Biology

There is a physiological cost of reproduction for fruit flies, such that it reduces the lifespan of female fruit flies. Is there a similar cost to male fruit flies? This dataset contains observations from five groups of 25 male fruit flies. The experiment tests if increased reproduction reduces longevity for male fruit flies. The five groups are: males forced to live alone, males assigned to live with one or eight newly pregnant females (non-receptive females), and males assigned to live with one or eight virgin females (interested females). The name of the data set is ***`fruitfly.csv`****.*[4]

`No`	serial number (1-25) within each group of 25
`type`	Type of experimental assignment 1 = no females 2 = 1 newly pregnant female 3 = 8 newly pregnant females 4 = 1 virgin female 5 = 8 virgin females
`lifespan`	lifespan (days)
`thorax`	length of thorax (mm)
`sleep`	percentage of each day spent sleeping

(a) *Import the data set and obtain summary statistiscs and examine the distribution of the overall lifespan of the fruitflies.*

First, let's load in our data and investigate it by using the `summary()` function, as well as plotting the distribution of lifespan.

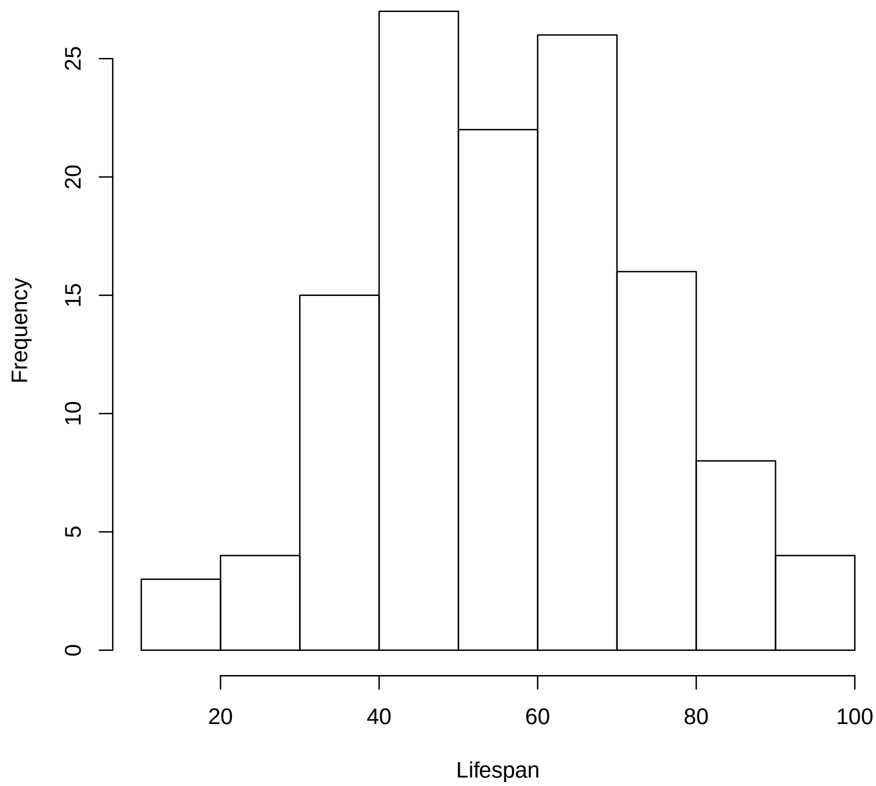
```r
# (a) import the data set
fruitfly <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsI_
    Fall2021/main/datasets/fruitfly.csv")
# summarize the statistics in the data set
summary(fruitfly)
# show the distribution of the overall lifespan of the fruitflies
pdf("plot3_a.pdf")
hist(fruitfly$lifespan, main="", xlab="Lifespan")
dev.off()
```

```
No            type         lifespan          thorax           sleep
Min.   : 1   Min.   :1   Min.   :16.00   Min.   :0.640   Min.   : 1.00
1st Qu.: 7   1st Qu.:2   1st Qu.:46.00   1st Qu.:0.760   1st Qu.:13.00
```

[4] Partridge and Farquhar (1981). "Sexual Activity and the Lifespan of Male Fruitflies". *Nature*. 294, 580-581.

```
Median :13    Median :3    Median :58.00   Median :0.840   Median :20.00
Mean   :13    Mean   :3    Mean   :57.44   Mean   :0.821   Mean   :23.46
3rd Qu.:19    3rd Qu.:4    3rd Qu.:70.00   3rd Qu.:0.880   3rd Qu.:29.00
Max.   :25    Max.   :5    Max.   :97.00   Max.   :0.940   Max.   :83.00
```

Figure 2: Histogram of `lifespan`.

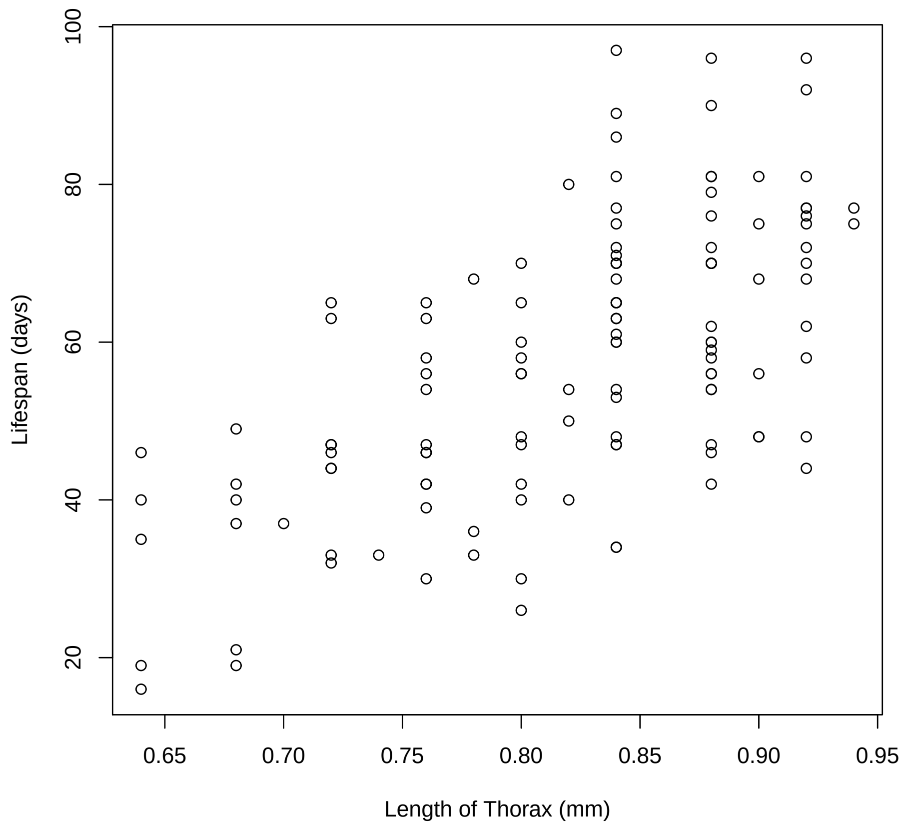


(b) *Plot* ***`lifespan`*** *vs* ***`thorax`****. Does it look like there is a linear relationship? Provide the plot. What is the correlation coefficient between these two variables?*

Let's create a scatter plot of the relationship between `thorax` and `lifespan`. We can see in Figure 3 that there appears to be a positive relationship between the two variables, which is confirmed if we investigate the correlation coefficient (using `cor()`), which is ≈ 0.64.

```
pdf("plot3_b.pdf")
plot(fruitfly$thorax, fruitfly$lifespan,
     xlab = "Length of Thorax (mm)", ylab = "Lifespan (days)")
dev.off()
```

Figure 3: Scatter plot of `thorax` and `lifespan`.




```r
# calculate correlation coefficient between lifespan and thorax
cor(fruitfly$thorax, fruitfly$lifespan)
```

```
[1] 0.6364835
```

(c) *Regress* ***lifespan*** *on* ***thorax****. Interpret the slope of the fitted model.*

```r
# (c) # Run the regression of lifespan on thorax
regression_model_problem3 <- lm(lifespan ~ thorax, data=fruitfly)
# get summary statistics for linear regression model
summary(regression_model_problem3)
```

Using the estimated coefficients and the summary statistics of from the linear regression model, we can calculate the fitted model as $\hat{y} = -61.05 + 144.33x$. The slope of the fitted model is 144.33 and we can interpret the slope as such: when the length of `thorax` increases by 1 mm, the average lifespan a fruitfly increases by 144.33 days.

(d) *Test for a significant linear relationship between* ***`lifespan`*** *and* ***`thorax`****. Provide and interpret your results of your test.*

Table 1: Estimated regression coeffiecient from model executed in part (c).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-61.05	13.00	-4.69	0.00
`thorax`	144.33	15.77	9.15	0.00

For our hypothesis test that the slope $\beta 1 = 0$ in Table 1, the associated test statistic $t = 9.152$ and the $p - value = 1.5e - 10$. The p-value is much less than 0.05 (since level of significance $\alpha = 0.05$), so we can reject the null hypothesis ($\beta 1 = 0$). In other words, there is a statistically differentiable linear relationship from zero between the lifespan of the fruitflies and the length of thorax of the fruitflies.

(e) *Provide the 90% confidence interval for the slope of the fitted model.*

- *Use the formula for typical confidence intervals to find the 90% confidence interval around the point estimate.*

  Let's use the formula to calculate the confidence interval where $\hat{\beta}_1 = 144.33$, $t - score = 1.65$ and $se = 15.77$.

```
# (e) calculate the confidence interval by formula
pointEst <- 144.33
se <- 15.77

# get the t-score
t <- qt(0.95,25*5-2)

# create the upper and lower bounds
lower_CI <- pointEst - t*se
upper_CI <- pointEst + t*se
```

  The resulting interval is [118.19, 170.47] around $\hat{\beta} = 144.33$. The confidence interval does not include zero, which is consistent with the hypothesis test we did in the previous question.

- *Now, try using the function* ***`confint()`*** *in* ***R****.*

  Surprise, surprise, we get the same answer!

```
# now try confint
confint(regression_model_problem3, "thorax", level = 0.9)
```

(f) *Use the* ***predict()*** *function in* ***R*** *to (1) predict an individual fruitfly's lifespan when* ***thorax****=0.8 and (2) the average* ***lifespan*** *of fruitflies when* ***thorax****=0.8 by the fitted model. This requires that you compute prediction and confidence intervals. What are the expected values of lifespan? What are the prediction and confidence intervals around the expected values?*

```
# (f) prediction
# store the two variables in x and y
x <- fruitfly$thorax
y <- fruitfly$lifespan

# predict() function to predict an individual fruitfly lifespan
predict(lm(y~x), newdata = data.frame(x=0.8), interval = "prediction",
    level = 0.90)

# predict() function to predict the average lifespan of fruitflies
predict(lm(y~x), newdata = data.frame(x=0.8), interval = "confidence",
    level = 0.90)
```

First, let's calculate the predicted individual fruitfly lifespan with thorax value of $0.8mm$. The resulting estimated individual value of lifespan for a fruitfly with $0.8mm$ length of thorax is 54.41 days and the corresponding 90% prediction interval is $[31.78, 77.05]$. The resulting estimated average lifespan for fruitflies with $0.8mm$ length of thorax is 54.41 days and the corresponding 90% confidence interval is $[52.33, 56.50]$.

Notice that the prediction interval for an estimated individual lifespan is much wider than the confidence interval for an estimated average lifespan. This is just the same as what we would expect since there is more variability in individual responses than in average responses.

(g) *For a sequence of* ***thorax*** *values, draw a plot with their fitted values for* ***lifespan****, as well as the prediction intervals and confidence intervals.*

```
# (g) create plot of confidence and prediction intervals
# find fitted values and prediction and confidence intervals
# going from .64 to .94 since these are the max and min values of thorax
prep.a <- predict(lm(y~x), newdata = data.frame(x=seq(min(x), max(x),
    0.003)), interval = "confidence")
prep.i <- predict(lm(y~x), newdata = data.frame(x=seq(min(x), max(x),
    0.003)), interval = "prediction")

# open a plot to show the fitted values of lifespan, the prediction and
    confidence intervals
pdf("plot3_g.pdf")
matplot(seq(min(x), max(x), 0.003), cbind(prep.a, prep.i[,-1]), lty = c
    (1,2,2,3,3), type = "l",
        col = c("black", "blue", "green", "red", "purple"),
```

```
        xlab = "Length of Thorax (mm)", ylab = "Fitted Value of Lifespan (days)")

# Add a legend to the plot to show which line represents which value
legend("topleft", legend = c("Fitted Value", "Lower Bound for Confidence Interval",
        "Upper Bound for Confidence Interval", "Lower Bound for Prediction Interval",
        "Upper Bound for Prediction Interval"), lty = c(1,2,2,3,3),
        col = c("black", "blue", "green", "red", "purple"), cex = 0.85)
dev.off()
```

Figure 4: Plot of fitted values, confidence and prediction intervals.