Problem Set 2

Applied Stats/Quant Methods 1

Due: 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

(a) The χ^2 test statistic is calculated as follows:

Read in the data as a matrix.

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

Calculate the expected values, then calculate the difference between the observed and expected values for each sub-category. Calculate the contribution to the χ^2 statistic.

expected number in class * number of outcomes / total number

difference observed - expected

[1] Fried, Lagunes, and Venkataramani (2010). “Corruption and Inequality at the Crossroad: A Multi-method Study of Bribery and Discrimination in Latin America. *Latin American Research Review.* 45 (1): 76-97.

contribution difference^2/expected)

For example, for the sub-category 'Upper Class' and 'Not Stopped':

	Upper Class, Not Stopped
observed	14
expected	13.5 = (27 * 21 / 42)
difference	0.5 = (14 - 13.5)
chi sq contribution	0.0185 = $(0.5\ )^2$ / 13.5

```
ncols <- length(observed[1,])
nrows <- length(observed[,1])

# get totals
row_tots <- vector("double", nrows)
col_tots <- vector("double", ncols)

totals <- sum(observed)  # total number of observations

# calculate row and column totals, e.g, total for NotStopped, UpperClass,
    etc
for (i in 1:nrows) {row_tots[i] <- sum(observed[i, ])}
for (i in 1:ncols) {col_tots[i] <- sum(observed[, i])}

#get expected = row total * column total / total observations
expected <- observed

for (i in 1:nrows) {
  for (j in 1:ncols) {
    expected[i,j] <- row_tots[i] * col_tots[j] / totals
  }
}

# calculate difference between observed and expected
o_e <- observed
o_e <- (o_e - expected)^2 / expected

#calculate chi-squared value & degrees of freedom
chi_sq_val <- sum(o_e)
df = (nrows-1) * (ncols-1)
```

(b) The p-value from the test statistic is calculated as follows. If $\alpha = 0.1$, we cannot reject the null hypothesis that both subgroups are from the same population (i.e. the difference in experience is not statistically significant).

```
p_value <- pchisq(chi_sq_val, df=df, lower.tail=FALSE)
```

```
  The p-value is 15.02%, alpha is 10%
   We cannot reject the null hypothesis that the two sets are from the
same population
   1 observed cell(s) with less than 5 values
```

The observed and expected values are shown in Figure 1

The results of the builtin `R` chisq.test function are as follows:

```
Pearson's Chi-squared test

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

(c) The standardized residuals are set out in Table 1:

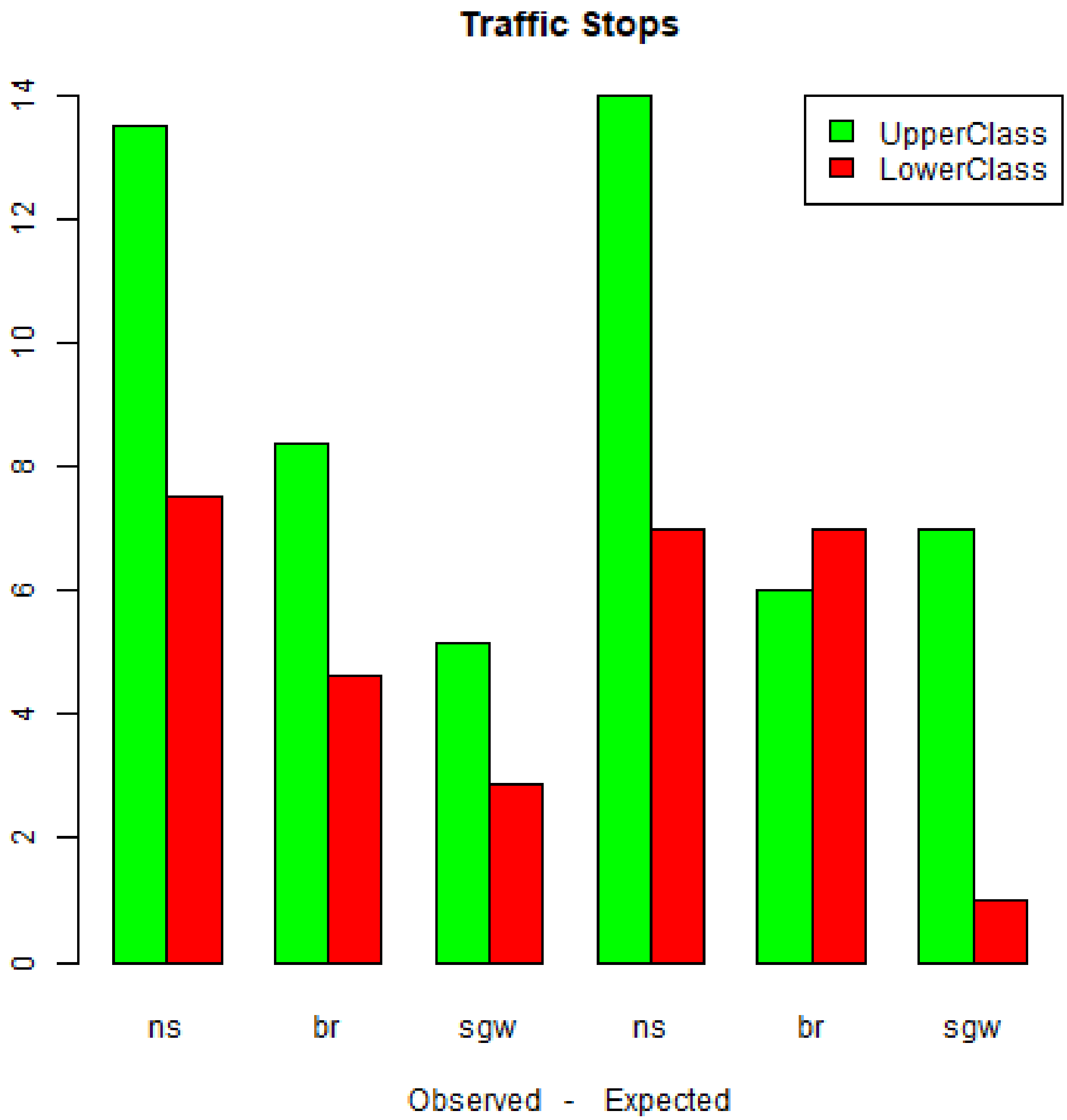


Figure 1: Observed vs Expected values for traffic stop. ns = Not Stopped; br = Bribe Requested; sgw = Stopped Given Warning

Table 1: Standardised Residuals

	NotStopped	BribeRequested	StoppedGivenWarning
UpperClass	0.322	-1.642	1.523
LowerClass	-0.322	1.642	-1.523

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

The biggest contribution to the residuals was from the 'Bribe Requested' variable - fewer upper class individuals were expected to hand over bribes. The difference between the two groups appears to be a combination of fewer upper class drivers being expected to hand over bribes and more of them being given a warning instead the opposite outcome occuring for lower class drivers.

We are not rejecting the null hypothesis, so we are concluding that there may not be any significant relationship between class and the outcomes experienced during traffic stops. The combined effect from the diffent experiences of the two groups was not enough to convince us that class predicts whether or not a driver is asked for a bribe.

Question 2 (40 points): Economics

Chattopadhyay and Duflo were interested in whether women promote different policies than men.[2] 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 2 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 2: 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

[2] 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 The reservation policy has no effect on the number of new or repaired drinking water facilities in the villages.

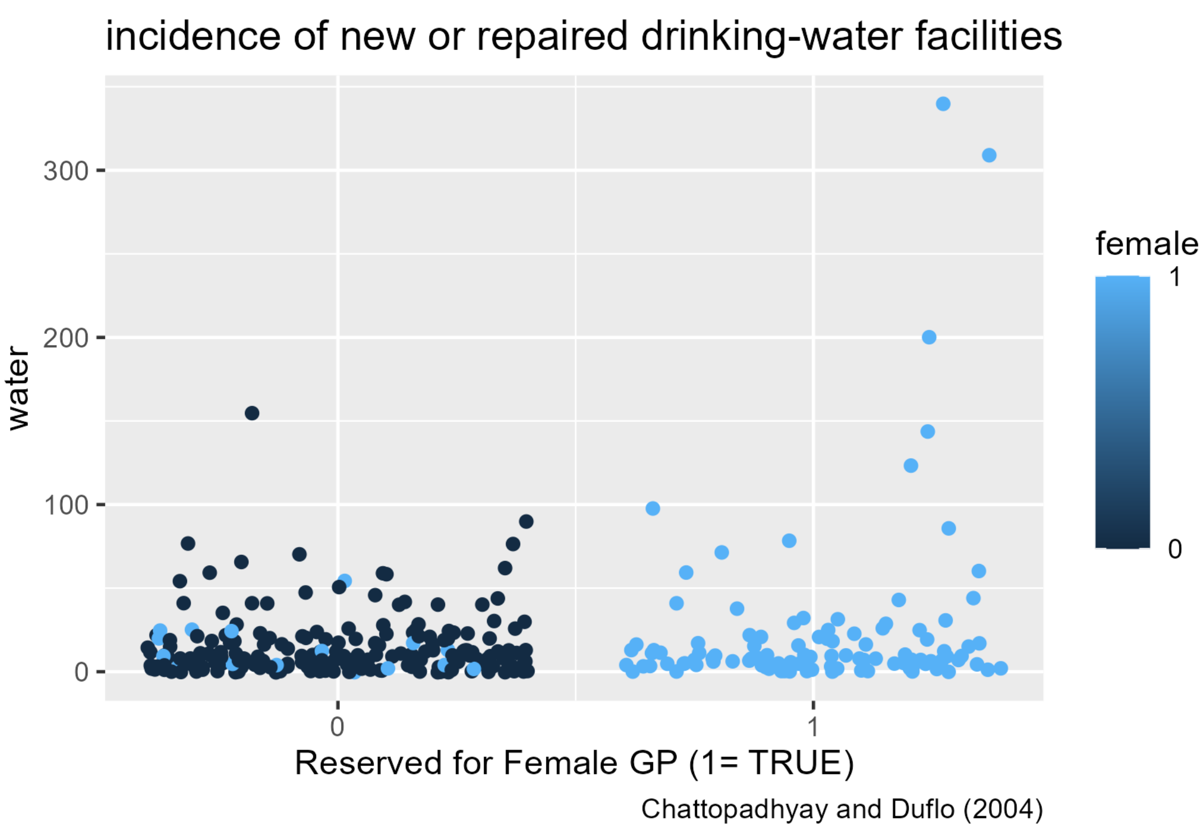
Alternate The reservation policy does have an effect on the number of new or repaired drinking water facilities in the villages.

(b) Bivariate regression to test this hypothesis:.

Import the data.

```
policy <- read_csv("https://raw.githubusercontent.com/kosukeimai/qss/
    master/PREDICTION/women.csv")
```

Figure 3: Drinking water projects, grouped by *reserved* $= [1, 0]$



The relationship between the

x = reserved binary variable indicating whether the GP was reserved for women leaders or not

y = water numeric variable denoting the number of new or repaired drinking water facilities in the villages

was modelled using the Pearson model for linear regression. The code is as follows:

```
#\item [(b)] Run a bivariate regression to test this hypothesis
water <- policy$water                # ie y = response var
reserved <- policy$reserved          # ie x = explanatory var

mean_water <- mean(water)
mean_reserved <- mean(reserved)

n <- length(water)

# calculate sum of squares for reserved and water
ssxx <- sum((reserved - mean_reserved)^2)
ssyy <- sum((water - mean_water)^2)
ssxy <- sum((reserved - mean_reserved)*(water - mean_water) )
# calculate covariance
covxy <- ssxy / n
# check result
cov(x = reserved, y = water, method = "pearson")
#calculate correlation coefficient
corxy <- ssxy / sqrt(ssxx * ssyy)

#calculate estimates for coefficients
beta1 <- ssxy / ssxx
beta0 <- mean_water- mean_reserved*beta1

# calculate standard error values
sse <- sum((water-(beta0 + beta1*reserved))^2)
se <- sqrt(sse / (n-2))

# calculate standard errors for coefficients
s_beta1 <- se * sqrt(1/ssxx)
s_beta0 <- se * sqrt((1/n + mean_reserved ^2 / ssxx))

# calculate the t-test statistics for coefficients
t_beta1 <- beta1 / s_beta1
t_beta0 <- beta0 / s_beta0

# calculate r^2 and p values
r2 <- 1 - (sse / ssyy)
p_beta1 <- 2*pt(t_beta1, df=n-2, lower.tail = FALSE)
p_beta0 <- 2*pt(t_beta0, df=n-2, lower.tail = FALSE)
```

This gives the following results:

Table 2: coefficients for linear regression model water - reserved

	estimate	Std Error	t value	pr(>|t|)
intercept	14.738	2.286	6.446	$4.216474e-10$
reserved	9.252	3.948	2.344	$1.970398e-02$

Table 3: results for linear regression model water - reserved

residual error	degrees of freedom	$R\hat{}2$	covariance	correlation
33.4457	320	0.0169	2.0624	0.1299

The estimate for β_0 is 14.738; the estimate for β_1 is 9.252, where $y = \beta_0 + \beta_1 * x$; the response variable (y) is the incidence of investment in drinking water projects; the explanatory variable (x) is 1 if the GP position is reserved for a woman, 0 otherwise. The pvalue is 0.0197, so at a confidence level of 5%, we reject the null hypothesis that the two variables are independent. The R^2 value suggests that our model accounts for less than 2% of the variance in our water values.

(c) Interpret the coefficient estimate for reservation policy.

We expect that where the GP position is not reserved for a female, the average number of drinking water projects will be 14.738 and that this will increase by 9.252 if the position is reserved.

Caveats

Code in Appendix.

Outliers

On inspection, it is clear that the data, and the model, are significantly affected by outliers (see Figure 4 and Table 4).

Figure 4: Boxplot of number of drinking water projects, grouped by reserved

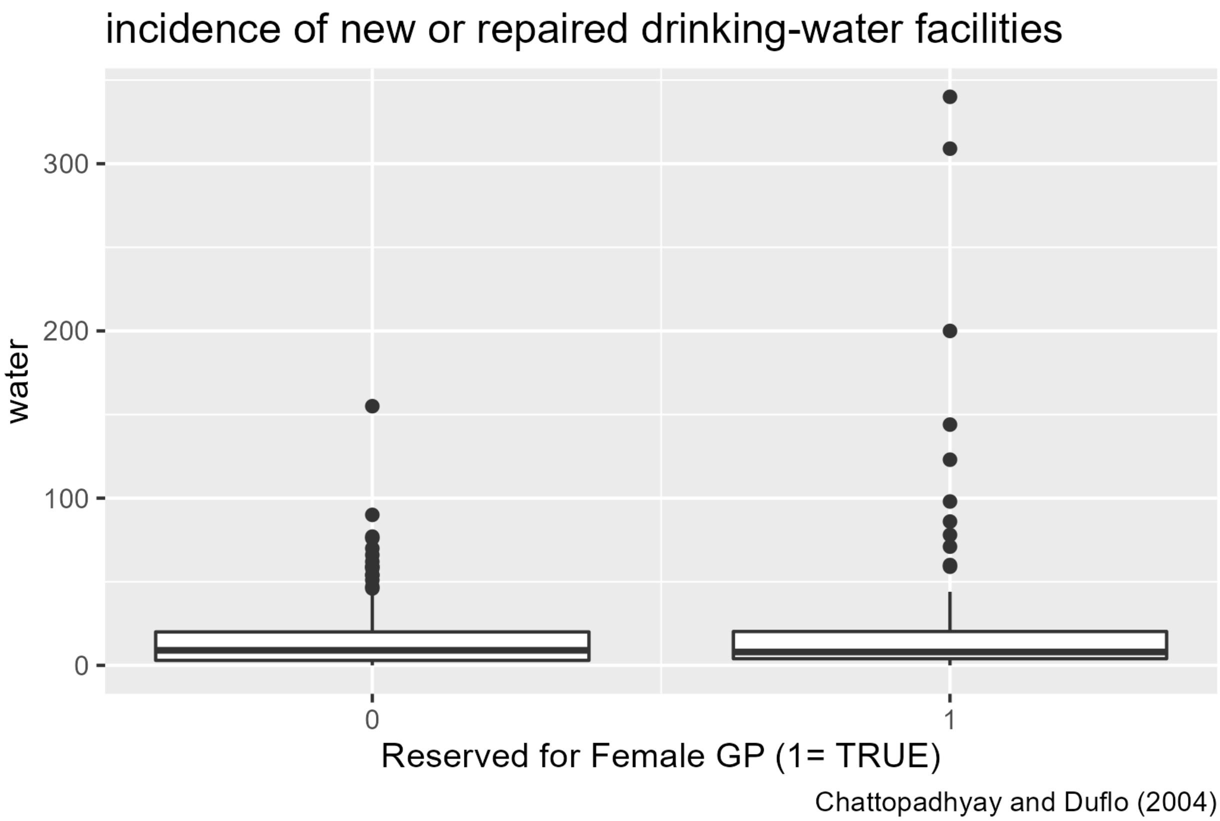


Table 4: Outliers in water incidence

reserved	mean_water	count_water	q3	iqr	outlier_limit
0	68.267	15	c('75%' = 20)	17	c('75%' = 45.5)
1	142.545	11	c('75%' = 20.25)	16.25	c('75%' = 44.625)

The data was modelled with outliers excluded and the results were as in Table 5

The estimate for β_0 is 10.7035; the estimate for β_1 is -0.1571 (p-value = 0.9015). Using this data, we cannot reject the hypothesis that water projects and reserved status are independent. The expected number of drinking water projects decreases by 0.1571 if the village is reserved for a female GP.

However, we have no data to support the idea that the outliers are bad data. We are more likely to conclude that the data is heavily skewed.

Table 5: Pearson Linear Regression - Water Reserved - excluding outliers

	water
reserved	−0.157
	(1.268)
Constant	10.704***
	(0.726)
N	296
R^2	0.0001
Adjusted R^2	−0.003
Residual Std. Error	10.243 (df = 294)
F Statistic	0.015 (df = 1; 294)

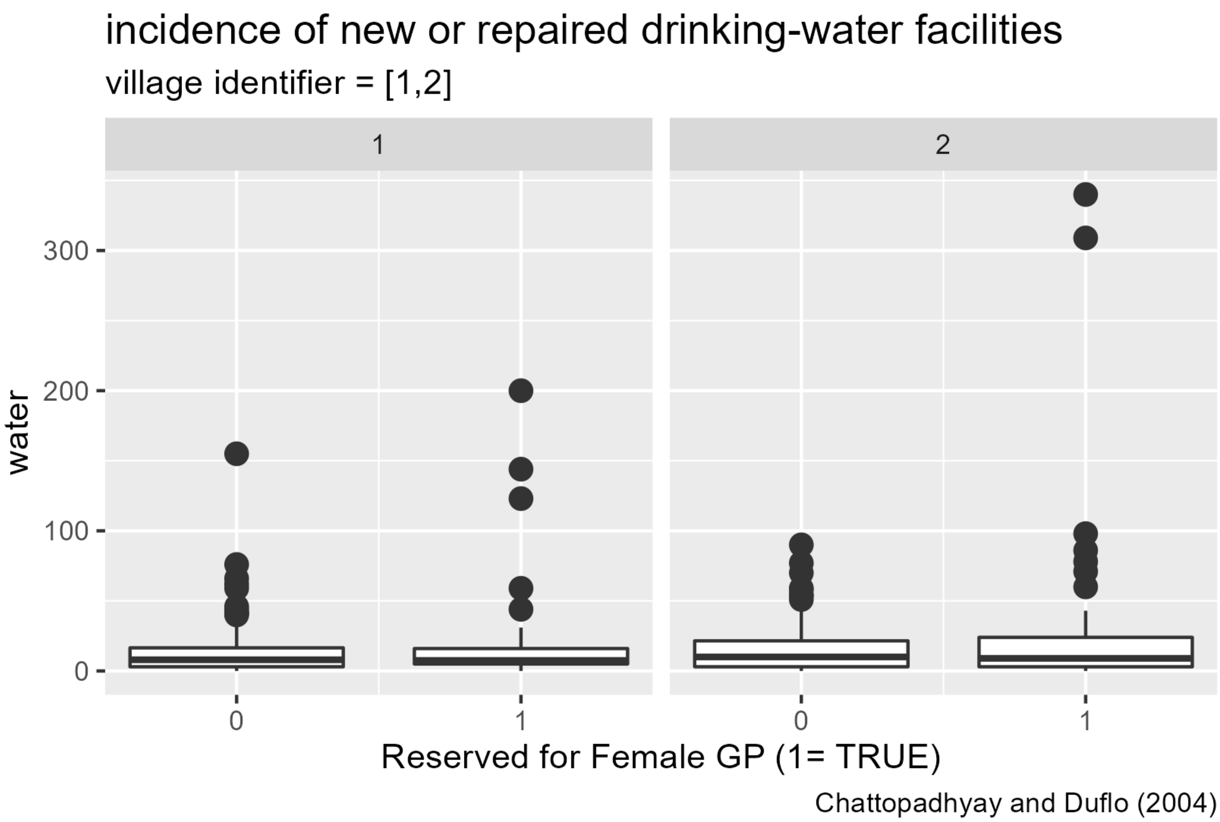
*p < .1; **p < .05; ***p < .01

Villages

The assumption in using a linear regression model is that each village is a separate case and each case is independent. However, in this study each GP is associated with two villages, so there is a risk that the values for each village are not independent.

As seen in Figure 5, the profile for the two sets of data has some differences, mainly the extra high values of the outliers in the $village == 2$ dataset.

Figure 5: Drinking water projects, grouped by Village



A χ^2 test was run on binned values, and this did not reject the hypothesis that the two samples were from the same population.

```
	Pearson's Chi-squared test

data:  one_counts and two_counts
X-squared = 12, df = 9, p-value = 0.2133
```

The linear model with the two villages combined (so our units are now GPs, not villages), gives the same expected values, but with lower confidence as we now have fewer data points.

When the two sets of villages are considered separately the estimate for β_1 is 5.130 for *village* == 1 (p-value = 0.2506) and 13.374 for *village* == 2 (pvalue = 0.04172).

This suggests that splitting or combining our data by village does not add greatly to our information about whether *reserved* is a predictor for *water*.

Table 6: Pearson Linear Regression - Water Reserved - Village = 1

	water
reserved	5.130
	(4.450)
Constant	13.907***
	(2.577)
N	161
R^2	0.008
Adjusted R^2	0.002
Residual Std. Error	26.656 (df = 159)
F Statistic	1.329 (df = 1; 159)

*p < .1; **p < .05; ***p < .01

Table 7: Pearson Linear Regression - Water Reserved - Village = 2

	water
reserved	13.374**
	(6.515)
Constant	15.570***
	(3.773)
N	161
R^2	0.026
Adjusted R^2	0.020
Residual Std. Error	39.028 (df = 159)
F Statistic	4.215** (df = 1; 159)

*p < .1; **p < .05; ***p < .01

Appendix - Code

```r
##################################################################
# Imelda Finn, 22334657
# POP77003 - Stats I
# clear global .envir, load libraries, set wd
##################################################################

# remove objects
rm(list=ls())

# detach all libraries
detachAllPackages <- function() {
  basic.packages <- c("package:stats", "package:graphics", "package:grDevices", "package:utils", "package:datasets", "package:methods", "package:base")
  package.list <- search()[ifelse(unlist(gregexpr("package:", search()))==1, TRUE, FALSE)]
  package.list <- setdiff(package.list, basic.packages)
  if (length(package.list)>0)  for (package in package.list) detach(package, character.only=TRUE)
}
detachAllPackages()

# load libraries
pkgTest <- function(pkg){
  new.pkg <- pkg[!(pkg %in% installed.packages()[,  "Package"])]
  if (length(new.pkg))
    install.packages(new.pkg,  dependencies = TRUE)
  sapply(pkg,  require,  character.only = TRUE)
}

# load necessary packages
lapply(c("ggplot2", "stargazer", "tidyverse", "stringr"),  pkgTest)

# function to save output to a file that you can read in later to your docs
output_stargazer <- function(outputFile, appendVal=TRUE, ...) {
  output <- capture.output(stargazer(...))
  cat(paste(output, collapse = "\n"), "\n", file=outputFile, append=appendVal)
}


# set working directory to current parent folder
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

######################
# Problem 1
######################

#Question 1 (40 points): Political Science

#The following table was created using the data from a study run in a major
# Latin American city.
```

```r
# 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 \\
observed  <- matrix( c (14, 6, 7, 7, 7, 1), nrow = 2, byrow = TRUE)

# create data structure with named dimensions
cols <- c("NotStopped", "BribeRequested",  "StoppedGivenWarning")
rows <- c("UpperClass", "LowerClass")

#\item [(a)]
#Calculate the $\chi^2$ test statistic by hand/manually\\

###----------------------------------------  0 start listing of code from here
ncols <- length(observed[1,])
nrows <- length(observed[,1])

# get totals
row_tots <- vector("double", nrows)
col_tots <- vector("double", ncols)

totals <- sum(observed)  # total number of observations

# calculate row and column totals, e.g, total for NotStopped, UpperClass, etc
for (i in 1:nrows) {row_tots[i] <- sum(observed[i, ])}
for (i in 1:ncols) {col_tots[i] <- sum(observed[, i])}

#get expected = row total * column total / total observations
expected <- observed

for (i in 1:nrows) {
  for (j in 1:ncols) {
    expected[i,j] <- row_tots[i] * col_tots[j] / totals
  }
}

```

```
# calculate difference between observed and expected
o_e <- observed
o_e <- (o_e - expected)^2 / expected

#calculate chi-squared value & degrees of freedom
chi_sq_val <- sum(o_e)
df = (nrows-1) * (ncols-1)

cat(str_glue("The chi_squared statistic is {round(chi_sq_val,3)}"))
cat(str_glue("The chi_squared degrees of freedom is {df}"))

# plot of observed and expected values
png("graphics/obs_exp.png")
barplot(cbind(expected, observed ), legend.text = rows,
        names.arg = c("ns", "br", "sgw", "ns",  "br", "sgw"),
        args.legend = list(x = "topright"),
        main = "Traffic Stops", beside = TRUE, col = c("green", "red"),
        xlab = "Observed   -    Expected")
dev.off()


#\item [(b)]
#Now calculate the p-value from the test statistic you just created R
# .\footnote{Remember frequency should be $>$ 5 for all cells, but let's calculate
# the p-value here anyway.}  What do you conclude if $\alpha = 0.1$?\\

p_value <- pchisq(chi_sq_val, df=df, lower.tail=FALSE)
alpha <- 0.1

# p > alpha, can't reject null
if (p_value > alpha ) txt <- "cannot " else txt <- ""

# should have min of 5 values in each observed cell
cells_under <- length(observed[observed<5])

cat(str_glue("The p-value is {round(p_value*100,2)}%, alpha is {alpha*100}%.")
)
cat(str_glue("We {txt}reject the null hypothesis that the two sets are from the\n same population."))
cat(str_glue("note: {cells_under} observed cell(s) with less than 5 values."))

# \item [(c)] Calculate the standardized residuals for each cell and put them in the table below.

z <- observed
for (i in 1:nrows) {
  row_prop<- (1 - (row_tots [i] / totals))
  for (j in 1:ncols) {
    col_prop<- (1-  (col_tots[j] / totals))
    z[i,j] <- (observed[i,j] - expected[i,j])  /sqrt (expected[i,j]* row_prop
```

```r
    * col_prop)
  }
}

z_df <- data.frame(round(z,3), row.names = rows)
names(z_df) <- cols

print(z_df)

# output results for Zij values to .tex file
output_stargazer(z_df, outputFile="std_residuals.tex", type = "latex",
                 appendVal=FALSE,
                 title="Standardised Residuals",
                 summary = FALSE,
                 style = "apsr",
                 table.placement = "htb",
                 label = "tab:StandardisedResiduals",
                 rownames = TRUE
                 )


# check result
chisq.test(observed)
# Pearson's Chi-squared test

#data:  observed
#X-squared = 3.7912, df = 2, p-value = 0.1502

# \item [(d)] How might the standardized residuals help you interpret the results?

#  fewer upper class individuals asked for bribes and more given warnings;
#  the contribution from lower class drivers expected to give bribes is nearly
#  equivalent to the contribution from upper class drivers getting warnings

# ########################################################################################
# Problem 2
######################

#Question 2 (40 points): Economics
#Chattopadhyay and Duflo were interested in whether 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, 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:
#    \url{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~\ref{fig:women_desc} 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.
# Names and description of variables from Chattopadhyay and Duflo (2004)
# 1 `GP`  Identifier for the Gram Panchayat   
# 2 `village`  identifier for each village
# 3 `reserved`  binary variable indicating whether the GP was reserved for women leaders or not
# 4 `female`  binary variable indicating whether the GP had a female leader or not
# 5 `irrigation` variable measuring the number of new or repaired irrigation facilities in the village since the reserve policy started
# 6 `water`  variable measuring the number of new or repaired drinking-water facilities in the village since the reserve policy started

#\item [(a)] State a null and alternative (two-tailed) hypothesis.
# null: no diff in incidence of new or repaired drinking-water facilities
#  in the village since the reserve policy started
#   ie 'water' is independent of 'reserved'
# alternate: the incidence of new or repaired drinking-water facilities is
#  correlated to the reservation policy


policy <- read_csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv")
#write.csv(policy,"Data/policy.csv")
policy<-read_csv("Data/policy.csv")

summary(policy)

plot(policy$water)
boxplot(policy$water)
# lots of outliers, distribution is skewed right (mean > median)

plot(policy$water, policy$irrigation)
```

```
pairs(policy[4:7])

sum(policy$reserved)   # 108 of 322 villages have reserved GP (54 GPs)
sum(policy$female)     # 124 of 322 villages have female GP (62 GPs)

#\item [(b)] Run a bivariate regression to test this hypothesis
water <- policy$water                # ie y = response var
reserved <- policy$reserved          # ie x = explanatory var

mean_water <- mean(water)
mean_reserved <- mean(reserved)

n <- length(water)

# calculate sum of squares for reserved and water
ssxx <- sum((reserved - mean_reserved)^2)
ssyy <- sum((water - mean_water)^2)
ssxy <- sum((reserved - mean_reserved)*(water - mean_water) )
# calculate covariance
covxy <- ssxy / n
# check result
cov(x = reserved, y = water, method = "pearson")
#calculate correlation coefficient
corxy <- ssxy / sqrt(ssxx * ssyy)

#calculate estimates for coefficients
beta1 <- ssxy / ssxx
beta0 <- mean_water- mean_reserved*beta1

# calculate standard error values
sse <- sum((water-(beta0 + beta1*reserved))^2)
se <- sqrt(sse / (n-2))

# calculate standard errors for coefficients
s_beta1 <- se * sqrt(1/ssxx)
s_beta0 <- se * sqrt((1/n + mean_reserved ^2 / ssxx))

# calculate the t-test statistics for coefficients
t_beta1 <- beta1 / s_beta1
t_beta0 <- beta0 / s_beta0

# calculate r^2 and p values
r2 <- 1 - (sse / ssyy)
p_beta1 <- 2*pt(t_beta1, df=n-2, lower.tail = FALSE)
p_beta0 <- 2*pt(t_beta0, df=n-2, lower.tail = FALSE)

# output results as two tables
cols <- c("estimate ", "Std Error",  "t value", "pr(>|t|)")
rows <- c("intercept", "reserved")

```

```r
beta_vals <- data.frame(matrix(c(round(beta0, 3), round(s_beta0, 3),
                    round(t_beta0, 3), p_beta0,
                    round(beta1, 3), round(s_beta1, 3), round(t_beta1, 3), p_
    beta1),
                        nrow = 2, byrow = TRUE), row.names = rows)

names(beta_vals) <- cols

print(beta_vals)

# output results for beta values to .tex file
output_stargazer(beta_vals, outputFile="policy_model.tex", type = "latex",
                  appendVal=FALSE,
                  title="coefficients for linear regression model water -
    reserved ",
                  summary = FALSE,
                  style = "apsr",
                  table.placement = "htbp!",
                  label = "tab:coefficients",
                  rownames = TRUE
)

result_vals <- tibble(`residual error`= round(se, 4), `degrees of freedom`= n
    -2,
                        `R^2`= round(r2, 4), `covariance`= round(covxy,4),
                        `correlation`= round(corxy,4))



output_stargazer(result_vals, outputFile="policy_model.tex", type = "latex",
                  appendVal=TRUE,
                  title="results for linear regression model water - reserved "
    ,
                  summary = FALSE,
                  style = "apsr",
                  label = "tab:results",
                  rownames = FALSE
)

result_cols <- tibble(round(se, 4), n-2, round(r2, 4),
                       round(covxy,4), round(corxy,4))
names(result_cols) <- c("residual error", "degrees of freedom", "R^2",
                         "covariance","correlation")

#check r^2
r<-cov(reserved, water) / (sd(reserved)* sd(water))

#check correlation coefficient
cor(policy$water, policy$reserved)  # .1299
# water increases with increase in reserved (ie reserved = TRUE), not strong

```

```

output_stargazer(water_model, outputFile="water_model.tex", type = "latex",
                 appendVal=FALSE,
                 title="Pearson Linear Regression - Water ~ Reserved",
                 style = "apsr",
                 table.placement = "htbp!",
                 label = "model:water_reserved"
)



water_model <- lm(water ~ reserved , data = policy)
summary(water_model)
#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



output_stargazer(water_model, outputFile="water_lm.tex", type = "latex",
                 appendVal=FALSE,
                 title="Pearson Linear Regression - Water ~ Reserved",
                 style = "apsr",
                 table.placement = "htbp!",
                 label = "model:water_reserved"
)

#-----------------------------------------




p<- ggplot(policy, aes(reserved, water, colour=female, group_by(female)))
p + geom_jitter() +
  scale_x_continuous(breaks = seq(0, 1, by = 1)) +
  scale_color_continuous(breaks = seq(0, 1, by = 1)) +
  labs(title ="incidence of new or repaired drinking-water facilities",
       x = "Reserved for Female GP (1= TRUE)",
       caption = "Chattopadhyay and Duflo (2004)",
```

```r
      alt = "Boxplot of incidence of new or repaired drinking-water facilities, by reserved [1,0]",
  )
ggsave("graphics/resrvd_water.png")

##==========================================================================
# consider outliers

p<- ggplot(policy, aes(reserved, water, group_by(reserved)))
p + geom_boxplot( aes(group=reserved)) +
  scale_x_continuous(breaks = seq(0, 1, by = 1)) +
  labs(title ="incidence of new or repaired drinking-water facilities",
     x = "Reserved for Female GP (1= TRUE)",
     caption = "Chattopadhyay and Duflo (2004)",
     alt = "Boxplot of incidence of new or repaired drinking-water facilities, by reserved [1,0]",
)
ggsave("graphics/resrvd_water_boxplot.png")

outliers_tbl <- policy %>%
  group_by(reserved) %>%
  mutate(iqr = IQR(water), q3 = quantile(water, .75), outlier_limit = q3 + iqr * 1.5 ) %>%
  filter(water > outlier_limit ) %>%
  mutate(mean_water = round(mean(water),3), count_water = n()) %>%
  select(reserved, mean_water, count_water, q3, iqr, outlier_limit) %>%
  unique()

#reserved mean_water count_water    q3   iqr outlier_limit
#<dbl>      <dbl>       <int> <dbl> <dbl>         <dbl>
#  1          0         68.3           15   20     17               45.5
#  2          1        143.            11   20.2   16.2             44.6


output_stargazer(outliers_tbl, outputFile="water_outliers.tex", type = "latex",
                  appendVal=FALSE,
                  title="Outliers in water incidence",
                  summary = FALSE,
                  style = "apsr",
                  digits= 3,
                  table.placement = "htbp!",
                  label = "tab:wateroutliers",
                  rownames = FALSE
)


# there are fewer outliers in the reserved=1 cohort, but their average
# value is significantly higher

no_outlier_water <- policy %>%
```

```
  group_by(reserved) %>%
  mutate(outlier_limit = quantile(water, .75) + IQR(water) * 1.5) %>%
  ungroup() %>%
  filter(water <= outlier_limit)

outlier_model <-lm(water ~ reserved , data = no_outlier_water)

output_stargazer(outlier_model, outputFile="outlier_model.tex", type = "latex",
                 appendVal=FALSE,
                 title="Pearson Linear Regression - Water ~ Reserved - excluding outliers",
                 style = "apsr",
                 table.placement = "htbp!",
                 label = "tab:noOutliers"
)

summary(outlier_model)


# coefficient for beta0 goes to -0.1571 - with no significance
# (p-value is 0.9015, df= 294)
# same result if exclude sample outliers (ie not by reserved)
##=================================================================
# assumption is that each village is a separate case and each case is independent
# but, each GP relates to 2 villages - need to check for impact of combining villages

# inspect data
p<- ggplot(policy, aes(reserved, water, group_by(reserved)))
p + geom_boxplot( outlier.size = 3, aes(group=reserved)) +
  scale_x_continuous(breaks = seq(0, 1, by = 1)) +
  labs(title ="incidence of new or repaired drinking-water facilities",
       subtitle = "village identifier = [1,2]",
       x = "Reserved for Female GP (1= TRUE)",
       caption = "Chattopadhyay and Duflo (2004)",
       alt = "Boxplot of incidence of new or repaired drinking-water facilities, by reserved [1,0]",
  ) +
  facet_wrap(policy$village)
ggsave("graphics/village_water_boxplot.png")

reserved_water_tab <- policy %>%
  group_by(reserved) %>%
  summarise(n = n(), sum_water = sum(water)) %>%
  mutate(prop_reserved = round(n / sum(n), 4), sum_water, prop_water_reserved =
           round(sum_water / sum(sum_water), 4)) %>% # mutate after our summarise to find the proportion
  arrange(desc(prop_reserved))
```

```
str(reserved_water_tab)
reserved_water_tab

sum(policy$water)

# see if villages are from same population
one_village_policy <- policy %>%
  group_by(GP) %>%
  filter(village ==1)

two_village_policy <- policy %>%
  group_by(GP) %>%
  filter(village ==2)


hist(one_village_policy$water)$counts
#[1] 130  16   8   3   0   0   1   2   0   1
hist(two_village_policy$water)$counts
#[1] 146  13   0   0   0   0   2
hist(one_village_policy$water)$breaks
#[1]   0  20  40  60  80 100 120 140 160 180 200

# coerce counts of water variable into suitably sized bins
one_counts <- hist(one_village_policy$water, breaks = c(0, 20, 40, 60, 350))$counts
two_counts <- hist(two_village_policy$water, breaks = c(0, 20, 40, 60, 350))$counts
# run chisq test - null: both from same population

chi_village <- chisq.test(one_counts, two_counts)


villagetab <- matrix(c(one_counts, two_counts), nrow = 2, byrow = TRUE)
chi_village

output_stargazer(tibble(villagetab), outputFile="village_bins.tex", type = "latex",
                 appendVal=FALSE,
                 title="Binned data for village dataset comparison",
                 summary = FALSE,
                 style = "apsr",
                 table.placement = "htbp!",
                 label = "tab:villageBins",
                 rownames = TRUE
)


#   Pearson's Chi-squared test

#data:  one_counts and two_counts
```

```
#X-squared = 12, df = 9, p-value = 0.2133


#tibble(`village1` = one_counts,`village2`= two_counts )

# run regression model on each set of villages
one_model <- lm(water ~ reserved, data = one_village_policy)
two_model <- lm(water ~ reserved, data = two_village_policy)

summary(one_model)
summary(two_model)

output_stargazer(one_model, outputFile="village_model.tex", type = "latex",
                 appendVal=FALSE,
                 title="Pearson Linear Regression - Water ~ Reserved - Village = 1",
                 style = "apsr",
                 table.placement = "htb",
                 label = "tab:village1"
)

output_stargazer(two_model, outputFile="village_model.tex", type = "latex",
                 appendVal=TRUE,
                 title="Pearson Linear Regression - Water ~ Reserved - Village = 2",
                 style = "apsr",
                 table.placement = "htb",
                 label = "tab:village2"
)


# or combine the villages

combined_village_policy <- policy %>%
  group_by(GP) %>%
  mutate (sum_water = sum(water), sum_irrigation = sum(irrigation)) %>%
  select(GP, reserved, female, sum_water, sum_irrigation) %>%
  unique()

# run model - scaled by 1/2 to get equivalent values to 1 village coefficients
cvp <- lm(sum_water/2 ~ reserved, data = combined_village_policy)

summary(cvp)

output_stargazer(cvp, outputFile="villages_combined.tex", type = "latex",
                 appendVal=FALSE,
                 title="Pearson Linear Regression - Water ~ Reserved - Villages combined",
                 style = "apsr",
                 table.placement = "htb",
                 label = "tab:combinedVillages"
```

```
)


# refs

# Foundations of Statistics for Data Scientists; with R and Python
# https://en.wikipedia.org/wiki/Least_squares
#
```