Statistical Analysis Empirical Exercise

Min Jiang; F2

1.1 Find the sample size.

```
v1hh <- read.csv("C:/studydata/Statistical Analysis/R/v1hh.csv")##insert data
View(v1hh)
##1.1 Find  the sample  size
n <- length(v1hh$mpid_hh
```

n	6163L

We get the sample size is 6163.

1.2 Find the mean and standard deviation for the following variables:
templecompany_hh
lendmoney_hh
keroricego_hh

```
> summary(v1hh)
       X              mpid_hh          pid_hh
 Min.   :   1    Min.   :1002    Min.   : 1001
 1st Qu.:1542    1st Qu.:1048    1st Qu.: 1048
 Median :3082    Median :1089    Median : 1089
 Mean   :3082    Mean   :1091    Mean   : 1677
 3rd Qu.:4622    3rd Qu.:1140    3rd Qu.: 1140
 Max.   :6163    Max.   :1174    Max.   :77777
 keroricego_hh     templecompany_hh   keroricecome_hh
 Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
 1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000
 Median :0.00000   Median :0.00000   Median :0.0000
 Mean   :0.04056   Mean   :0.01136   Mean   :0.0404
 3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000
 Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
  lendmoney_hh     borrowmoney_hh     samecaste_hh
 Min.   :0.00000   Min.   :0.00000   Min.   :0.000
 1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000
 Median :0.00000   Median :0.00000   Median :1.000
 Mean   :0.04024   Mean   :0.04511   Mean   :0.656
 3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.000
 Max.   :1.00000   Max.   :1.00000   Max.   :1.000
 sameoccupation_hh
 Min.   :0.0000
 1st Qu.:0.0000
 Median :1.0000
 Mean   :0.5694
 3rd Qu.:1.0000
 Max.   :1.0000
> sd(v1hh$templecompany_hh)
[1] 0.105976
> sd(v1hh$lendmoney_hh)
[1] 0.1965379
> sd(v1hh$keroricego_hh)
[1] 0.1972954
```

So templecompany_hh’s mean is 0.01136, standard deviation is 0.105976.
lendmoney_hh’s mean is 0.04024, standard deviation is 0.1965379.

keroricego_hh’s mean is 0.0404, standard deviation is 0.1972954.

1.3 Find the correlation coefficient between keroricego_hh and samecaste_hh

```
> cor(v1hh$keroricego_hh, v1hh$samecaste_hh)
[1] -0.09176947
```

The correlation coefficient between keroricego_hh and samecaste_hh is -0.09176947.

1.4 How many pairs (in total) have the relation “lendmoney_hh” in this village?

```
> sum(v1hh$lendmoney_hh)
[1] 248
```

There are 248 pairs (in total) have the relation “lendmoney_hh” in this village.
We can also calculate it through 1.1 and 1.2 questions. We can get the mean of lendmoney_hh is 0.0404 and the sample size is 6163. So there are 0.0404*6163=248 pairs (in total) have the relation “lendmoney_hh” in this village

1.5 Is the average number relation “lendmoney_hh” greater than 0. 035 at 5% significance level? Run the t-test below and report/interpret what you find.

```
> t.test(v1hh$lendmoney_hh, mu = 0.035, alternative = c("greater"))

        One Sample t-test

data:  v1hh$lendmoney_hh
t = 2.0931, df = 6162, p-value = 0.01819
alternative hypothesis: true mean is greater than 0.035
95 percent confidence interval:
 0.03612161        Inf
sample estimates:
 mean of x
0.04024014
```

We can get p-value is 0.01819<0.05, so we reject H0 and get the conclusion that the average number relation “lendmoney_hh” is greater than 0. 035 at 5% significance level.
Or 95 percent confidence interval= 0.03612161< mean of x=0.04024014, so we reject H0 and get the same conclusion that the average number relation “lendmoney_hh” is greater than 0. 035 at 5% significance level.

1.6. Construct a confidence level for the population mean for lendmoney_hh, at confidence level 95%.

```
meanlendmoney <- mean(v1hh$lendmoney_hh)
n <- length(v1hh$lendmoney_hh)
s <- sd(v1hh$lendmoney_hh)
se <- s/sqrt(n)
c <- qt(0.975,n-2)
CI <- c(meanlendmoney-c*se, meanlendmoney+c*se)
```

CI	num [1:2] 0.0353 0.0451

So we get a confidence level for the population mean for lendmoney_hh, at confidence level 95% is (0.0353, 0.0451).

2. Simple linear regressions

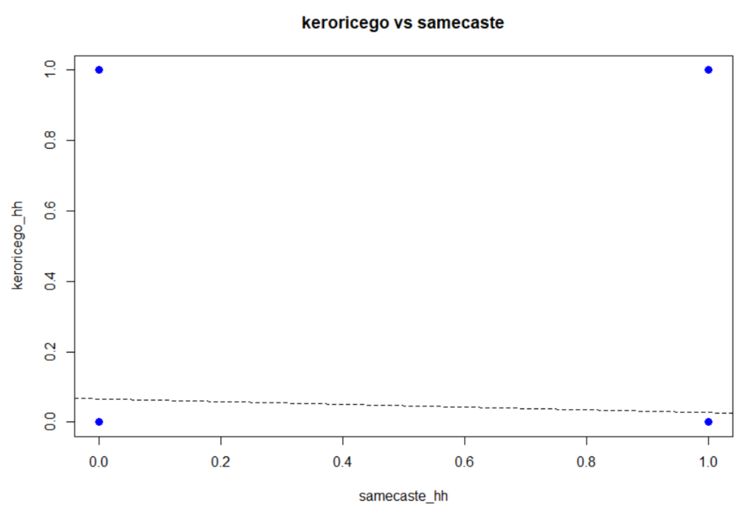
```
lm1 <- lm(keroricego_hh ~ samecaste_hh, data = v1hh)
summary(lm1)
plot(keroricego_hh ~ samecaste_hh, data = v1hh, pch = 16, cex = 1.3,col = "blue", main = "keroricego vs samecaste", xlab = "samecaste_hh", ylab = "keroricego_hh")
abline(lm1,lty=2)
```

```
Call:
lm(formula = keroricego_hh ~ samecaste_hh, data = v1hh)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.06557 -0.06557 -0.02745 -0.02745  0.97255 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.065566   0.004267  15.365  < 2e-16 ***
samecaste_hh -0.038111   0.005269  -7.234 5.27e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1965 on 6161 degrees of freedom
Multiple R-squared:  0.008422,	Adjusted R-squared:  0.008261 
F-statistic: 52.33 on 1 and 6161 DF,  p-value: 5.271e-13
```



2.1 Find the intercept and slope. Find the p-values for both. How do you interpret the slope (together with its p-value)?
Compare your answer here to your answer to question 1.3, what can you find?
Plot the regression line, what can you find?

Intercept is 0.065566, its P-value is < 2e-16 ***. Intercept is significantly different from 0 at 0.1% level. Slop is -0.038111, its P-value is 5.27e-13 ***. Slop is significantly differed from 0 at 0.1% level.

On average, the likelihood of having the relationship of borrowing/lending rice is 6.6% if the two hh's do not belong to the same caste. Belonging to same caste decreases that likelihood by 3.8%, to 6.6- 3.8= 2.8%. The effect of same caste is very significant (extremely small p-value)

The correlation coefficient, denoted by r, tells us how closely data in a scatterplot falls along a straight line. The closer that the absolute value of r is to 1, the closer that the data is described by a linear equation. In this condition, the correlation coefficient is -0.09176947, which shows the data cannot be described by a linear equation suitably. We can also find Multiple R-squared is 0.008422, which also shows the data cannot be described by a linear equation suitably. But in question 2.1 shows "keroricego _hh" and "samecaste_hh" are related, so we had better find other function to fit the data.
The correlation coefficient of keroricego _hh and samecaste_hh is so low that it seems to show there are not relationship between. But in question 2.1, the P-value of the slop show they are related at a 0.1% significance level.

Plot the regression line, we can see keroricego _hh and samecaste_hh are negative correlated. The line very close to the samecastes_hh line, which means whether the two hh's belong to the same caste, the first hh does not tend to lend rice to the second hh.

2.2 Repeat what you did in 2.1, with "templecompany_hh" being the new dependent variable (instead of "keroricego_hh").

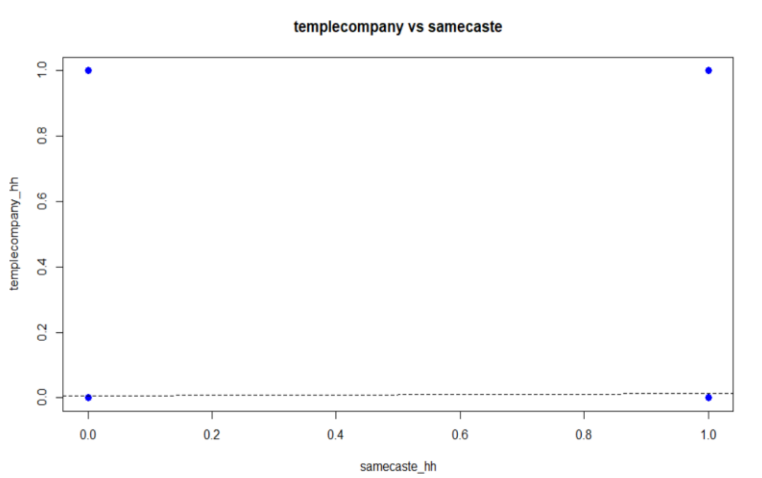
```
lm2 <- lm(templecompany_hh ~ samecaste_hh, data = v1hh)
summary(lm2)
plot(templecompany_hh  ~ samecaste_hh, data = v1hh,pch = 16, cex = 1.3,col = "blue", main = "templecompany vs samecaste", xlab = "samecaste_hh", ylab = "templecompany_hh")
abline(lm2,lty=2)
```

```
Call:
lm(formula = templecompany_hh ~ samecaste_hh, data = v1hh)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.01410 -0.01410 -0.01410 -0.00613  0.99387 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept)  0.006132   0.002300   2.666  0.00770 **
samecaste_hh 0.007966   0.002840   2.805  0.00505 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1059 on 6161 degrees of freedom
Multiple R-squared:  0.001275,  Adjusted R-squared:  0.001113 
F-statistic: 7.867 on 1 and 6161 DF,  p-value: 0.005049
```



Intercept is 0.006132, its P-value is 0.00770 **. Intercept is significantly different from 0 at 1% level, but not at 0.1% level.
Slop is 0.007966, its P-value is 0.00505 **. Slop is significantly differed from 0 at 1% level, but not at 0.1% level.
On average, the likelihood of having the relationship of going to temple together is 0.61% if the two hh's do not belong to the same caste. Belonging to same caste increases that likelihood by 0.80%, to 0.61+ 0.80= 1.41%. The effect of same caste is very significant (extremely small p-value)

```
> cor(v1hh$templecompany_hh, v1hh$samecaste_hh)
[1] 0.03571211
```

In this condition, the correlation coefficient is 0.03571211, which shows the data cannot be described by a linear equation suitably. We can also find Multiple R-squared is 0.001275, which also shows the data cannot be described by a linear equation suitably. But in question 2.2 shows "templecompany_hh" and "samecaste_hh" are related, so we had better find other function to fit the data.
The correlation coefficient of templecompany_hh and samecaste_hh is so low that it seems to show there are not relationship between. But in question 2.2, the P-value of the slop show they are related at a 1% significance level.

Plot the regression line, we can see templecompany_hh and samecaste_hh are positive correlated. The line very close to the samecastes_hh line, which means whether the two hh's belong to the same caste, the two hh's don't tend to go to temple together.

2.3 Is there any difference(s) between what you found in 2.1 and 2.2? (e.g., in terms of slope, p-value, etc.)
If you found any differences, can you tell some story/intuition to justify those?

1)

Belonging to same caste decreases the likelihood of having the relationship of borrowing/lending rice by 3.8%. Probably because it two households are in the same social class, then they may have the same economic difficulties. They may tend to borrow rice from higher caste.

Belonging to same caste increases the likelihood of having the relationship of going to temple by 0.80%. Probably because if two household are in the same caste, then they may have the common topics and the same affordable vehicles to go to the temple.

2)

The P-value of slop in 2.1 is 5.27e-13 ***. The P-value of slop in 2.2 is 0.00505**.5.27e-13 is much smaller than 0.00505. It means Belonging to same caste is more likely to affect the relationship of borrowing/lending rice than the relationship of going to temple. Probably because the problem of survival is the most important problem. People can participate in other recreational and religious activities only when they have enough to eat.

3. Multiple linear regressions

3.1 Report the following information:
Coefficients of the two explanatory variables,
Corresponding p-values,
(Adjusted) R-squared,

```
> lm3 <- lm(keroricego_hh ~ templecompany_hh + lendmoney_hh, data = v1hh)
> summary(lm3)

Call:
lm(formula = keroricego_hh ~ templecompany_hh + lendmoney_hh,
    data = v1hh)

Residuals:
     Min       1Q   Median       3Q      Max
-0.64132 -0.01911 -0.01911 -0.01911  0.98089

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.019112   0.002221   8.605  < 2e-16 ***
templecompany_hh 0.124130   0.020533   6.045 1.58e-09 ***
lendmoney_hh     0.498074   0.011072  44.987  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1702 on 6160 degrees of freedom
Multiple R-squared:  0.2564,    Adjusted R-squared:  0.2561
F-statistic:  1062 on 2 and 6160 DF,  p-value: < 2.2e-16
```

Slop of templecompany_hh: 0.124130; Corresponding p-values: 1.58e-09 ***;
Slop of lendmoney_hh: 0.498074; Corresponding p-values: < 2e-16 ***;
(Adjusted) R-squared: 0.2561

3.2 Interpret your result based on coefficients and p-values you found in 3.1.
Intercept is 0.019112, which means on average, the likelihood of having the relationship of borrowing/lending rice is 1.9% if the two hh's neither go to temple together nor having the relationship of borrowing/lending money.

Slop of templecompany_hh is 0.124130; Corresponding p-values is 1.58e-09 ***; Slop is significantly differed from 0 at 0.1% level. Going to temple together increases the likelihood of having the relationship of borrowing/lending rice by 12.4%, to 1.9+12.4=14.3%. (Still don't have the relationship of borrowing/lending money)

Slop of lendmoney_hh is 0.498074; Corresponding p-values is < 2e-16 ***; Slop is significantly differed from 0 at 0.1% level. having the relationship of borrowing/lending money increases the likelihood of having the relationship of borrowing/lending rice by 49.8%; to 1.9+49.8=51.7%. (Still don't go the temple together)

3.3 For a pair of households such that "templecompany_hh = 1" and "lendmoney_hh = 1", what is the predicted/fitted value (probability) that they also have the relation "keroricego_hh_fit"?
(a) First calculate the fitted value based on what you answered in 3.1.
(b) Verify your answer using the following code in R:
keroricego_hh_fit <- predict(lm3)
which(v1hh$templecompany_hh == 1 & v1hh$lendmoney_hh == 1)
keroricego_hh_fit[XXX]

(a) When templecompany_hh = 1 and lendmoney_hh = 1, we plug them into the formula:

keroricego_hh= 0.019+0.124 templecompan
y_hh+ 0.498 lendmoney_hh

Then we get: keroricego_hh= 0.019+0.124+0.498=0.641
So the predicted/fitted value (probability) that they also have the relation "keroricego_hh_fit" is 64.1%.

(b)

```
> keroricego_hh_fit <- predict(lm3)
> which(v1hh$templecompany_hh == 1 & v1hh$lendmoney_hh == 1)
 [1]  245 2003 3543 4144 4165 4185 4743 4759 5093 5170 5173 5413 5418
[14] 5512
```

```
> keroricego_hh_fit[245]
      245
0.6413163
> keroricego_hh_fit[2003]
     2003
0.6413163
> keroricego_hh_fit[4185]
     4185
0.6413163
> keroricego_hh_fit[5512]
     5512
0.6413163
```

We get when mpid_hh=245 2003 3543 4144 4165 4185 4743 4759 5093 5170 5173 5413 5418 5512, "templecompany_hh = 1" and "lendmoney_hh = 1",the probability of each one is 64.1%, which fits the answer in 3.2 question.