Activity 8 - Gioconda Prada

…

The main goal of the activity is to make sure package `regclass` is installed so that you are set up for the rest of the course. This package contains the commands written specifically for association and regression analysis and contains many datasets.

Once `regclass` is installed, load it up and make it available by running

```
library(regclass)
#If not installed, run the line below. (Of course, remove # sign)
#install.package("regclass")
```

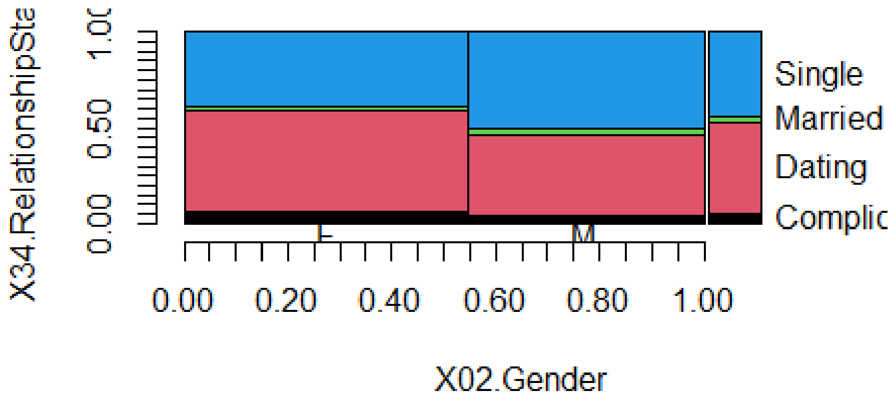
Question 1: After the `regclass` library has been loaded, load in the `SURVEY11` dataset from package `regclass` by running

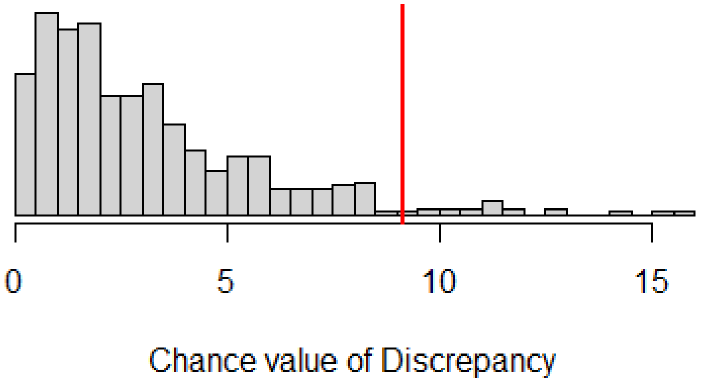
```
data(SURVEY11)
```

This contains information about students from 2011. Let us study a potential association between someone's relationship status `X34.RelationshipStatus` (the *y* variable) and someone's gender `X02.Gender`. If an association existed, this would imply that men and women view definitions of relationship statuses somewhat differently.

Use `associate`, adding the argument `seed=298` so that the set of 500 (default value) permutation datasets generated are the same for everyone.

```
#Your associate command using a seed of 298
associate(X34.RelationshipStatus~X02.Gender,data=SURVEY11,permutations=500,
seed = 298)
```





```
## Association between X02.Gender (categorical) and  X34.RelationshipStatus
(categorical):
## 
##  using 628 complete cases
## Contingency table:
##        y
## x       Complicated Dating Married Single Total
##   F              23    178       8    134   343
##   M              13    120       9    143   285
##   Total          36    298      17    277   628
## 
##  Table of Expected Counts:
##   Complicated Dating Married Single
## F        19.7  162.8     9.3  151.3
## M        16.3  135.2     7.7  125.7
## 
## Conditional distributions of y (X34.RelationshipStatus) for each level of
x (X02.Gender):
## If there is no association, these should look similar to each other and
##  similar to the marginal distribution of y
##          Complicated    Dating    Married    Single
## F         0.06705539 0.5189504 0.02332362 0.3906706
## M         0.04561404 0.4210526 0.03157895 0.5017544
## Marginal  0.05732484 0.4745223 0.02707006 0.4410828
## 
## Permutation procedure:
##   Discrepancy Estimated p-value
##      9.138875             0.034
## With 500 permutations, we are 95% confident that:
##  the p-value is between 0.02 and 0.054
## If 0.05 is in this range, change permutations= to a larger number
```

- Does the mosaic plot suggest that an association exists? Why or why not?

Response: Yes, there are noticeable differences in patterns.

- Let’s estimate *p*-value via theoretical approach for the “discrepancy”” between the conditional distributions of relationship status between genders?

```
ta<- 1-pchisq(9.138875,df=1*3)
print(ta)

## [1] 0.02750033
```

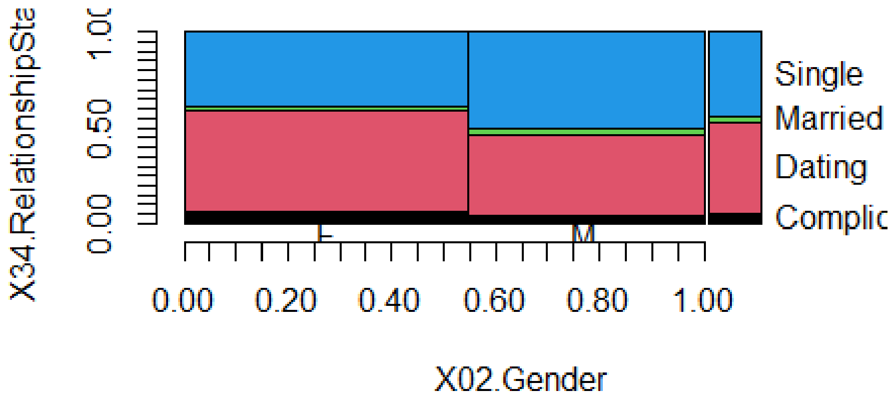
*Response:*0.02750033

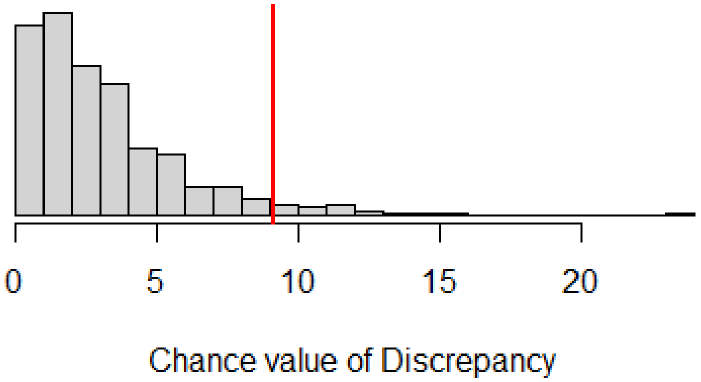
- Let’s estimate *p*-value via permutation approach for the “discrepancy”” between the conditional distributions of relationship status between genders? Why is the test inconclusive?

*Response:*it is inconclusive because the p-value is between 0.02 and 0.054 meaning it is slightly higher than 0.05

- Add the argument `permutations=1500` to make 1500 permutation datasets instead of the default 500 (have the seed still be 298) with seed number equals to 298. The test will now be conclusive. Is the association statistically significant?

```
#Your associate command using 1500 permutations and a seed of 298
associate(X34.RelationshipStatus~X02.Gender,data=SURVEY11,permutations=1500,
seed = 298)
```





```
## Association between X02.Gender (categorical) and  X34.RelationshipStatus
(categorical):
##
##  using 628 complete cases
## Contingency table:
##        y
## x       Complicated Dating Married Single Total
##   F              23    178       8    134   343
##   M              13    120       9    143   285
##   Total          36    298      17    277   628
##
##  Table of Expected Counts:
##   Complicated Dating Married Single
## F        19.7  162.8     9.3  151.3
## M        16.3  135.2     7.7  125.7
##
## Conditional distributions of y (X34.RelationshipStatus) for each level of
x (X02.Gender):
## If there is no association, these should look similar to each other and
##  similar to the marginal distribution of y
##          Complicated    Dating    Married    Single
## F         0.06705539 0.5189504 0.02332362 0.3906706
## M         0.04561404 0.4210526 0.03157895 0.5017544
## Marginal  0.05732484 0.4745223 0.02707006 0.4410828
##
## Permutation procedure:
##   Discrepancy Estimated p-value
##      9.138875             0.034
## With 1500 permutations, we are 95% confident that:
##  the p-value is between 0.025 and 0.044
## If 0.05 is in this range, change permutations= to a larger number
```

*Comment:*there exists a significant association, test result is conclusive

---

Activity 9

Question 1: Load up `SURVEY10` using `data(SURVEY10)`. Making student profiles is valuable from a marketing perspective so that advertisments can be targeted to the right set of people. To make a profile, associations must be analyzed.

a. Run `cor_matrix` on the first 7 columns of the data frame to obtain a correlation matrix (there are so many variables doing `cor_matrix` on the entirety of the data frame results in too many values to look at).

```
data(SURVEY10)
cor_matrix(SURVEY10[,1:7])
```

```
##                     Height Weight DesiredWeight    GPA TxtPerDay
## Height             1.000  0.601         0.774 -0.138    -0.085
## Weight             0.601  1.000         0.871 -0.215    -0.030
## DesiredWeight      0.774  0.871         1.000 -0.212    -0.073
## GPA               -0.138 -0.215        -0.212  1.000     0.072
## TxtPerDay         -0.085 -0.030        -0.073  0.072     1.000
## MinPerDayFaceBook -0.082  0.022        -0.064 -0.027     0.109
##                   MinPerDayFaceBook
## Height                       -0.082
## Weight                        0.022
## DesiredWeight                -0.064
## GPA                          -0.027
## TxtPerDay                     0.109
## MinPerDayFaceBook             1.000
```

b. Now run `all_correlations` on the entirety of the dataframe to get the pairwise correlations between every quantitative variable. Add the argument `sorted="strength"` to have this list sorted by strength (from most negative to most positive). What pair of quantities has the strongest positive, strongest negative, and strongest overall relationship?

```
#all_correlations(SURVEY10,sorted="strength")
head(all_correlations(SURVEY10,sorted="magnitude"))
```

```
##                      var1              var2 correlation          pval
## 1                  Weight     DesiredWeight   0.8713716 9.805054e-218
## 2                  Height     DesiredWeight   0.7741494 1.481493e-140
## 3                  Height            Weight   0.6010430  7.200545e-70
## 4           DesiredWeight  NumBodyPiercings  -0.5880114  3.123079e-66
## 5                 Height  NumBodyPiercings  -0.5818746  1.415521e-64
## 6 PercMoreAttractiveThan OwnAttractiveness   0.5591912  9.503662e-59
```

Response:the strongest positive= Weight and DesiredWeight strongest negative= DesiredWeight and NumBodyPiercings strongest overall relationship= Weight and DesiredWeight

c. Which quantity has the strongest correlation with `GPA`. Run `all_correlations` adding the argument `interest="GPA"` to find out.

```
all_correlations(SURVEY10,interest='GPA',sorted="magnitude")[1,]
```

```
##     var1 var2 correlation        pval
## 1 Weight  GPA  -0.2146987 9.85755e-09
```

Response:Weight

d. The analysis you have done so far has assumed that Pearson’s correlation coefficient provides a suitable measure of the strength of the relationship. This may not be the case.

- Rerun the `all_correlations` command from part b, but add the argument `type="spearman"` so that Spearmans’ rank correlations are considered. You can ignore the `There were 50 or more warnings...` message.

```
#all_correlations with argument type="spearman"
head(all_correlations(SURVEY10,type="spearman",sorted="strength"))
```

```
##               var1                 var2 correlation          pval
## 1    DesiredWeight     NumBodyPiercings  -0.7025526 4.516960e-105
## 2           Height     NumBodyPiercings  -0.6762389  1.392226e-94
## 3           Weight     NumBodyPiercings  -0.5726575  3.745198e-62
## 4 NumBodyPiercings   WeeklyHrsVideoGame  -0.4016186  1.780525e-28
## 5 NumBodyPiercings PercMoreAthleticThan  -0.3072680  9.483049e-17
## 6    DesiredWeight                  GPA  -0.2749433  1.370355e-13
```

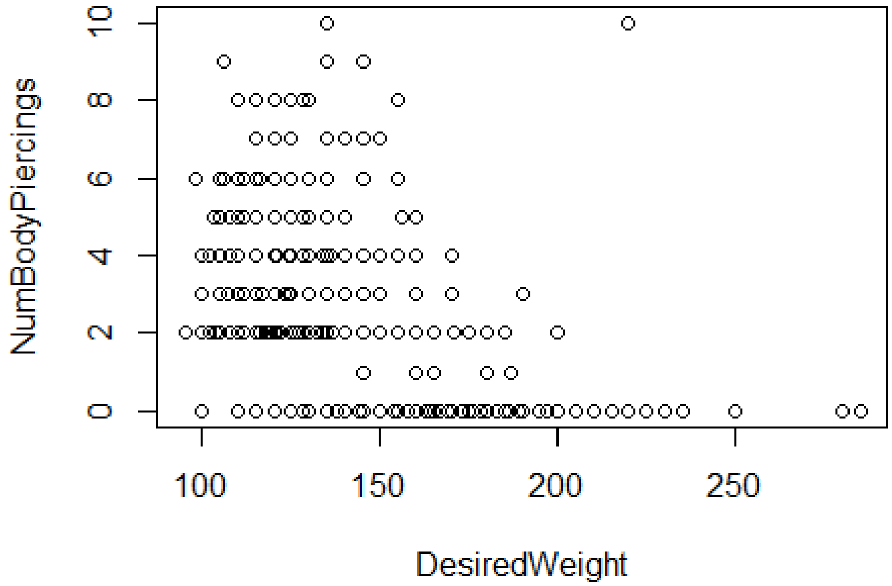
```
#all_correlations(SURVEY10,type="spearman",sorted="magnitude")
```

- What pair of variables have the strong negative association?

Response:DesiredWeight and NumBodyPiercings

- Make a scatterplot of their relationship (same sort of syntax as what you use for associate: `plot(y~x,data=DATA)` and determine if you should be using Pearson or Spearman to gauge the strength.

```
plot(NumBodyPiercings~DesiredWeight, data=SURVEY10)
```



Response:Spearman

- The association is highly statistically significant, but there is obviously not a cause-effect physical law underlying the relationship. Why do you think we see such a strong relationship (i.e., what might be a lurking variable)?

Response: Age

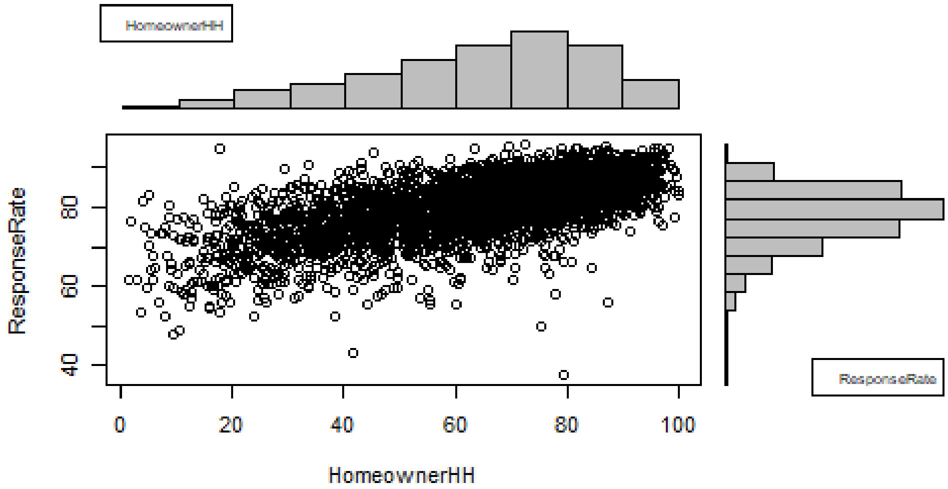
Question 3: Load `EX2.CENSUS` using `data(EX2.CENSUS)`.

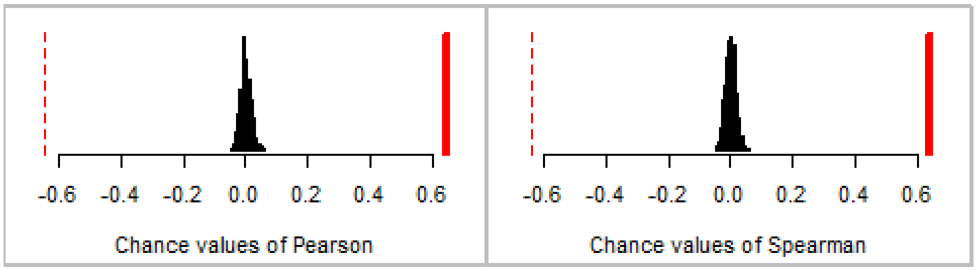
```
data(EX2.CENSUS)
```

b. Is the association between `ResponseRate` and `HomeownerHH` statistically significant?

```
associate(ResponseRate~HomeownerHH, data=EX2.CENSUS)

## Association between HomeownerHH (numerical) and  ResponseRate (numerical)
##  using 3534 complete cases
```





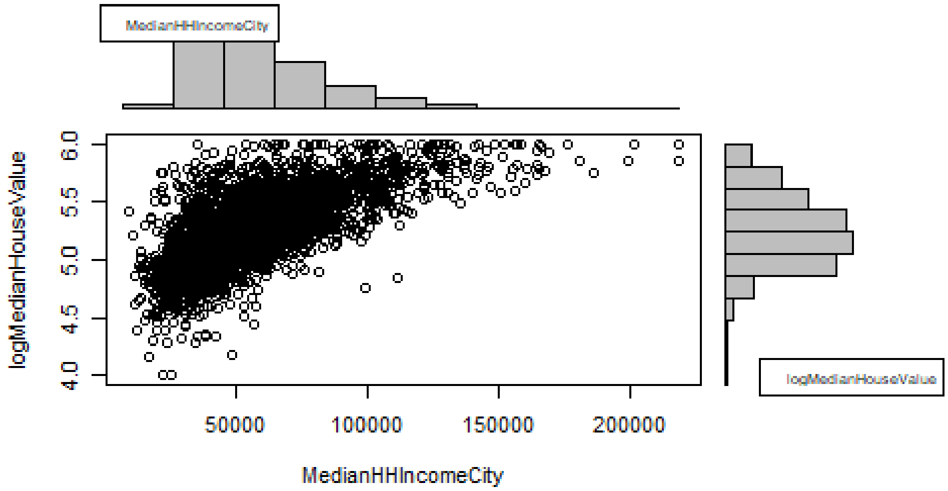
```
## Permutation procedure:
##                                  Value Estimated p-value
## Pearson's r                  0.6454007                 0
## Spearman's rank correlation 0.6421986                 0
## With 500 permutations, we are 95% confident that:
##  the p-value of Pearson's correlation (r) is between 0 and 0.007
##  the p-value of Spearman's rank correlation is between 0 and 0.007
## Note:  If 0.05 is in this range, increase the permutations= argument.
##
##
##
## Advice: If stream of points is well described by an ellipse, use Pearson's
r.
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation
## or try logs, e.g. associate( log10(y)~log10(x) )
```

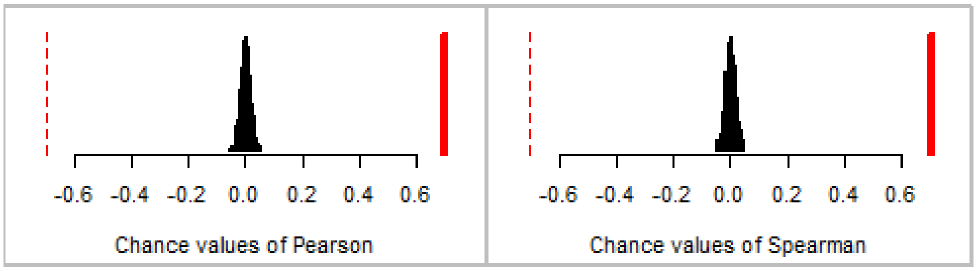
*Response:*yes

c. Describe the direction and form of the relationship between `logMedianHouseValue` and `MedianHHIncomeCity` (the x-variable). Does it have unusual features? Which measure should be used to gauge the strength?

```
associate(logMedianHouseValue~MedianHHIncomeCity, data=EX2.CENSUS)

## Association between MedianHHIncomeCity (numerical) and
logMedianHouseValue (numerical)
##  using 3534 complete cases
```





```
## Permutation procedure:
##                                  Value Estimated p-value
## Pearson's r                  0.7009088                 0
## Spearman's rank correlation 0.7060014                 0
## With 500 permutations, we are 95% confident that:
##  the p-value of Pearson's correlation (r) is between 0 and 0.007
##  the p-value of Spearman's rank correlation is between 0 and 0.007
## Note:  If 0.05 is in this range, increase the permutations= argument.
##
##
##
## Advice: If stream of points is well described by an ellipse, use Pearson's
r.
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation
## or try logs, e.g. associate( log10(y)~log10(x) )
```

*Response:*Positive direction, form is non-linear, it has some heteroscedasticity, spearman