Duke DASI project

Stephanie S

Friday, April 10, 2015

1, Introduction:

Is there have a relationship between age(variable name: age)with job satisfaction(variable name: satjob) (Negative or positive relationship between them)? This is an interesting question for the reason that it is helpful for society study, and further exploration. This project conducted a sociological survey used to collect data on demographic characteristics and attitudes of residents of the United States. Why is it interesting to explore? This data are used in numerous newspaper, magazine, and journal articles, by legislators, policy makers, and educators.

2, About data:

The data was collected by given a questionnaires that has to be filled/answered.There are observation : 57061 (cases) and 144 columns.Variables are age and satjob(codebook: satisfaction of job).

variable: age (numerical variable)—-Age of respondent(min=18, max=89, mean=45.7, median=43.0, NAs=202).

variable: satjob(categorical variable)—it is the satisfaction of job(Very Satisfied , Mod. Satisfied, A Little Dissat, Very Dissatisfied, NA's)(see the link for more information http://d396qusza40orc.cloudfront.net /statistics%2Fproject%2Fgss1.html (http://d396qusza40orc.cloudfront.net/statistics%2Fproject%2Fgss1.html)).

```
rm(list=ls())
load(url("http://bit.ly/dasi_gss_data"))
dim(gss)
```

```
## [1] 57061   114
```

The type of study is observational study for the reason it is tooks a questionnaires to be filled. ANOVA and pairwise tests to analysis those relationship between variables. Population of interest is all average age with different satisfactions on jobs(different group of attitude). The finding can generalize to that population for the reason it uses collected data draw from survey. Potential bias shoud be some individuals who are randomly selected but they are not respond to the survey;individuals who are easily accessible are more likely to be included in the sample. There can be a causal link between age and satjob since it uses hypothesis test and compare each group means and find there is at least one group are different,thus age have causal link with satisfaciton of job.

3, Exploratory analysis:

Brief table summary of variable satjob and create a boxplot side by side to visulaize data. Clean data first:

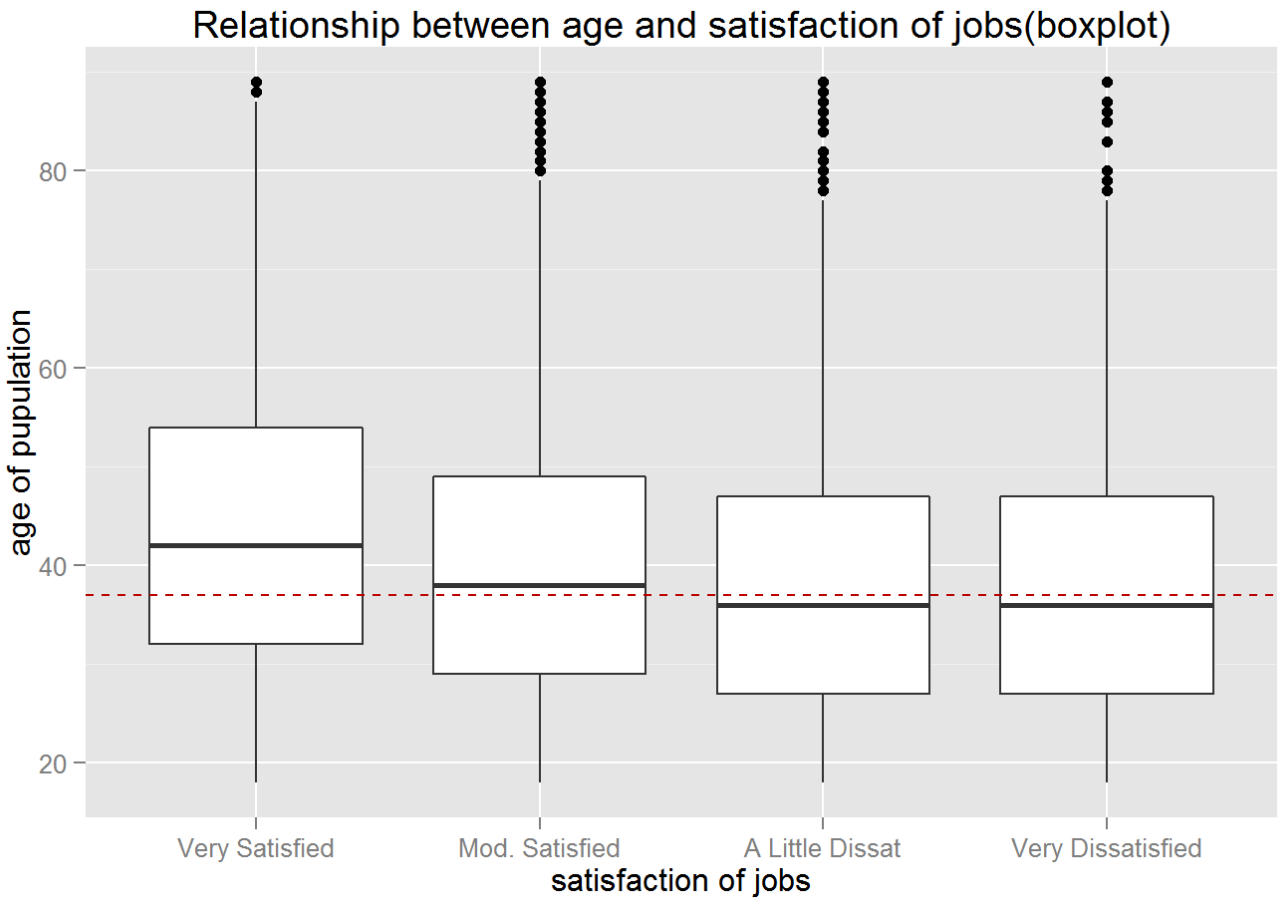
```
library(ggplot2)
target<-gss[,c("caseid","age","satjob")] #target is a data frame that contain only var
 we want
head(target)
```

```
##   caseid age          satjob
## 1      1  23 A Little Dissat
## 2      2  70            <NA>
## 3      3  48  Mod. Satisfied
## 4      4  27  Very Satisfied
## 5      5  61            <NA>
## 6      6  26  Mod. Satisfied
```

```
#remove row contain NAs. Cleaning data:
target_clean<-target[complete.cases(target),]
table(target_clean$satjob)
```

```
##
##    Very Satisfied    Mod. Satisfied   A Little Dissat Very Dissatisfied
##             19654             15693              4099              1707
```

```
g<-ggplot(target_clean,aes(satjob, age)) +
        geom_boxplot( data=target_clean, stat="boxplot", position ="dodge",
                      outlier.shape=16, outlier.size=2) +
        ggtitle("Relationship between age and satisfaction of jobs(boxplot)") +
        labs(x="satisfaction of jobs",y="age of pupulation") +
        geom_hline(aes(yintercept=37), linetype="dashed", colour="#BB0000")
print(g)
```



The finding is that among groups of satjob, the means of ages are less likely to be significant from each other. Next step is to use ANOVA(F statistic) to analysis those data.

4, Inference:

Using ANOVA, we state that null hypothesis is all the means ages of different group of satisfaction on their jobs are equal. So the alternative hypothesis is at least one pair of means are different. Condition for ANOVA is the groups of satjob are independent from each other(no pairing); also distribution of response variable within each group appear approximately normal, thus it fits for ANOVA.

From project requirement: we use method for one numerical (age) and one categorical variable(satjob) (with 4 levels)hypothesis test only

- compare means across several groups;
- no defined parameter of interest, ANOVA and pairwise tests.

```
#clearer to see means of each group:
tapply(target_clean$age, target_clean$satjob, mean)
```

```
##    Very Satisfied    Mod. Satisfied   A Little Dissat Very Dissatisfied
##          43.88354          40.19206          38.12954          37.77622
```

```
tapply(target_clean$age, target_clean$satjob, sd)
```

```
##    Very Satisfied    Mod. Satisfied   A Little Dissat Very Dissatisfied
##          14.69392          13.84934          13.45393          13.26103
```

Oneway ANOVA:

```
s<-lm(age~satjob, data=target_clean)
summary(s)
```

```
##
## Call:
## lm(formula = age ~ satjob, data = target_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.884 -11.192  -1.884   9.808  51.224
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               43.8835     0.1013  433.31   <2e-16 ***
## satjobMod. Satisfied      -3.6915     0.1520  -24.29   <2e-16 ***
## satjobA Little Dissat     -5.7540     0.2438  -23.60   <2e-16 ***
## satjobVery Dissatisfied   -6.1073     0.3583  -17.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.2 on 41149 degrees of freedom
## Multiple R-squared:  0.02444,    Adjusted R-squared:  0.02437
## F-statistic: 343.7 on 3 and 41149 DF,  p-value: < 2.2e-16
```

```
anova(s) #summary(aov.out) same
```

```
## Analysis of Variance Table
##
## Response: age
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## satjob         3  207821   69274  343.65 < 2.2e-16 ***
## Residuals 41149 8294878     202
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
aov.out<-aov(age~satjob, target_clean)
```

Calculate p-value in ANOVA:

```
pf(343.65,3,41149, lower.tail=FALSE)
```

```
## [1] 1.968292e-220
```

```
#Because pvalue is smaller than alpha. We will reject null hypothesis.
```

Pairwise test:

```
pairwise.t.test(target_clean$age, target_clean$satjob, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  target_clean$age and target_clean$satjob
##
##                   Very Satisfied Mod. Satisfied A Little Dissat
## Mod. Satisfied    < 2e-16        -              -
## A Little Dissat   < 2e-16        < 2e-16        -
## Very Dissatisfied < 2e-16        2.5e-11        0.39
##
## P value adjustment method: none
```

5, Conclusion:

(1)It means that we will reject null hypothesis, and the conclusion is there is at least one group of means that are significant different from each other.

(2)We assumes that alpha level is 0.05 for all tests. From pairwise t test, we can see that age means of A little satisfied group has not significant different with group Very Dissatisfied. There are significant difference on means of age(between Mod satisfied and Very satisfied, between A little satisfied and very satisfied, between Very dissatisfied and very satisfied, between Mod satisfied and a little satisfied, between very dissatisfied and mod satisfied).

6, References:

1, Where data from : http://bit.ly/dasi_gss_data (http://bit.ly/dasi_gss_data)

2, Data citation : http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/34802/version/1 (http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/34802/version/1)

7, Appendix:

```
head(target_clean, 50)
```

```
##    caseid age           satjob
## 1       1  23   A Little Dissat
## 3       3  48    Mod. Satisfied
## 4       4  27    Very Satisfied
## 6       6  26    Mod. Satisfied
## 7       7  28    Very Satisfied
## 8       8  27   A Little Dissat
## 9       9  21    Mod. Satisfied
## 10     10  30    Mod. Satisfied
## 12     12  56    Very Satisfied
## 13     13  54    Very Satisfied
## 14     14  49    Mod. Satisfied
## 15     15  41    Very Satisfied
## 16     16  54    Mod. Satisfied
## 19     19  46    Very Satisfied
## 21     21  57    Very Satisfied
## 22     22  58    Very Satisfied
## 23     23  21    Mod. Satisfied
## 26     26  53    Mod. Satisfied
## 27     27  42    Mod. Satisfied
## 28     28  42   A Little Dissat
## 29     29  20    Mod. Satisfied
## 30     30  23    Very Satisfied
## 31     31  26    Mod. Satisfied
## 32     32  25    Very Satisfied
## 35     35  21    Mod. Satisfied
## 36     36  27    Mod. Satisfied
## 39     39  58    Mod. Satisfied
## 40     40  51    Mod. Satisfied
## 42     42  53    Very Satisfied
## 43     43  39    Mod. Satisfied
## 47     47  25    Very Satisfied
## 48     48  49    Very Satisfied
## 49     49  40    Mod. Satisfied
## 50     50  43    Mod. Satisfied
## 51     51  46    Very Satisfied
## 52     52  37    Mod. Satisfied
## 53     53  46    Very Satisfied
## 56     56  35    Mod. Satisfied
## 59     59  37    Very Satisfied
## 60     60  57    Very Satisfied
## 61     61  39    Mod. Satisfied
## 63     63  51    Mod. Satisfied
## 64     64  43    Very Satisfied
## 65     65  39    Very Satisfied
## 67     67  30    Mod. Satisfied
## 68     68  47    Mod. Satisfied
## 69     69  31    Mod. Satisfied
## 71     71  45    Very Satisfied
## 75     75  19 Very Dissatisfied
```

```
## 76     76  23    Mod. Satisfied
```