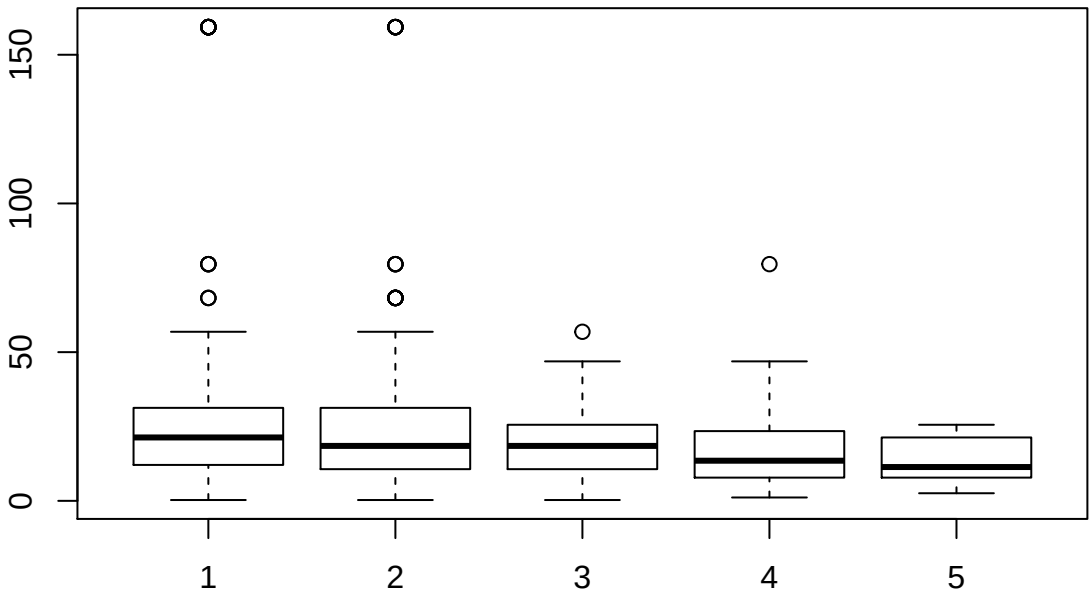
QMSSGR5015-Lab3-Yue_Ma

Yue Ma

2018/10/10

1. Run a simple bivariate regression, and interpret your results. (Did the results fit your expectations? Why? Why not?)

```
setwd("/Users/mayue/Desktop/qmss/data analysis/lab3")
g = read.csv("GSS.2006.csv")
g$realrinc1000s = (g$realrinc)/1000
plot(as.factor(g$rimpskls), g$realrinc1000s)
```




```
lm1 = lm(rimpskls ~ realrinc1000s , data = g, subset = !is.na(big5a2))
summary(lm1)
```

```
##
## Call:
## lm(formula = rimpskls ~ realrinc1000s, data = g, subset = !is.na(big5a2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0849 -0.9215 -0.0152  0.1159  3.0485
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.086384   0.042714  48.846  < 2e-16 ***
## realrinc1000s -0.005272   0.001115  -4.727 2.67e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9178 on 836 degrees of freedom
```

```
##   (678 observations deleted due to missingness)
## Multiple R-squared:  0.02603,    Adjusted R-squared:  0.02487
## F-statistic: 22.35 on 1 and 836 DF,  p-value: 2.671e-06
```

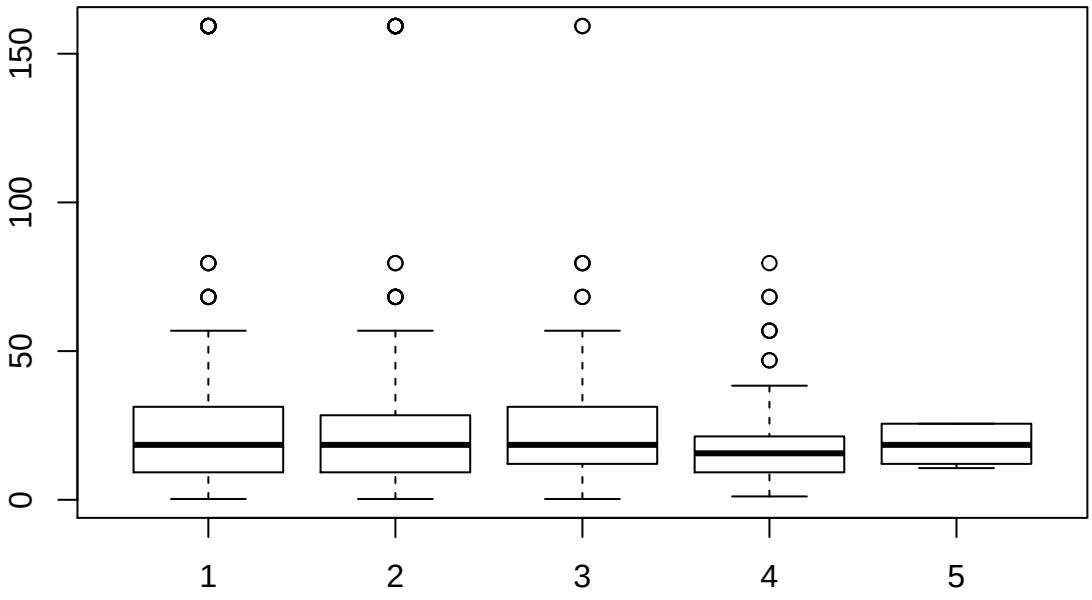
The variable "realrinc" indicates the income of the respondents. I turn income into 1000s of dollars for ease of interpretation. The variable "rimpskls" asked "Do you agree or disagree with the statment that your job gives you a chance to improve your skills?" with 1 strongly agree and 5 strongly disagree.

From the results, we can see that the estimate of "realrinc1000s" is -0.005272. It indicates that for every 1000 dollars more, a person believes it is 0.005 points less to improve their own skills in the job on average. The result fits my expectation, because more income means highr position or more qualified. As the position goes higher, people get less space for improvement. Instead, on some basic levels, especially entry level, people need to learn more to get familiar to the industries and the market, so they may get more chances to improve their skills.

2. Add an additional variable that might mediate or partly "explain" the initial association from that simple regression above – and explain your results. Did it work out? Yes? No?

Perhaps people who make more money are just more outgoing and sociable, so I looked at big5a2, To what extent do you agree or disagree with the following statements? I see myself as someone who is outgoing, sociable with 1 strongly agree and 5 strongly disagree. The lower the score, the more outgoing and sociable people are.

```
plot(as.factor(g$big5a2), g$realrinc1000s)
```



```r
lm2 = lm(rimpskls ~ realrinc1000s + big5a2, data = g)
summary(lm2)
```

```
##
## Call:
## lm(formula = rimpskls ~ realrinc1000s + big5a2, data = g)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4623 -0.8183 -0.0324  0.2026  3.2026
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.774922   0.081587  21.755  < 2e-16 ***
## realrinc1000s -0.004905   0.001106  -4.435 1.04e-05 ***
## big5a2         0.147933   0.033154   4.462 9.23e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9076 on 835 degrees of freedom
##   (3672 observations deleted due to missingness)
## Multiple R-squared:  0.04872,    Adjusted R-squared:  0.04644
## F-statistic: 21.38 on 2 and 835 DF,  p-value: 8.797e-10
```

```r
library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```r
stargazer(lm1, lm2, type = "text")
```

```
##
## ===============================================================
##                                 Dependent variable:
##                     ----------------------------------------
##                                       rimpskls
##                              (1)                    (2)
## ---------------------------------------------------------------
## realrinc1000s             -0.005***              -0.005***
##                            (0.001)                (0.001)
##
## big5a2                                           0.148***
##                                                   (0.033)
##
## Constant                  2.086***               1.775***
##                            (0.043)                (0.082)
##
## ---------------------------------------------------------------
## Observations                838                    838
## R2                         0.026                  0.049
## Adjusted R2                0.025                  0.046
## Residual Std. Error    0.918 (df = 836)       0.908 (df = 835)
```

```
## F Statistic         22.346*** (df = 1; 836) 21.381*** (df = 2; 835)
## ========================================================================
## Note:                                       *p<0.1; **p<0.05; ***p<0.01
```

From the results above, we can get the following conculsions.

Net of how much money someone makes, for each category more outgoing and sociable they feel, they are 0.148 points more to improve their own skills in the job on average. But I can see that the two estimates of variable "realrinc1000s" before and after adding the variable "big5a2" are both -0.005. As a result, this second variable hardly changes the income variable, so it is not mediating the effect really.

3. Run another multiple regression. Tell me how you expect your dependent variable to be affected by the independent variables. Interpret your results.

```
library(plyr)
setwd("/Users/mayue/Desktop/qmss/data analysis/lab3")
d = read.csv("WVS.csv")

d = rename(d, c("V8" = "work"))
d$rwork = 4 - d$work
d$rwork.lab <- ordered(d$rwork, levels = c(1,2,3,4), labels = c("not at all important", "2", "3", "very
table(d$rwork.lab)
```

```
##
## not at all important                   2                   3
##                 5413               18455               45798
##       very important
##                    0
```

```
d = rename(d, c("V239" = "income"))
d$married = ifelse(d$V57 == 1, 1,0)
d$female = ifelse(d$V240 == 2, 1, 0)
d = rename(d, c("V242" = "age"))
lm3 = lm(as.numeric(rwork.lab) ~ income + age + female, d, subset = V2 == 840 & !is.na(married))
lm4 = lm(as.numeric(rwork.lab) ~ income + age + female, d, subset = V2 == 156 & !is.na(married))
summary(lm3)
```

```
##
## Call:
## lm(formula = as.numeric(rwork.lab) ~ income + age + female, data = d,
##     subset = V2 == 840 & !is.na(married))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4103 -0.2937 -0.1724  0.7257  1.0054
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  2.5349991  0.0628883  40.310  < 2e-16 ***
## income      -0.0128126  0.0080156  -1.598   0.1101
## age         -0.0044753  0.0009274  -4.826  1.5e-06 ***
## female      -0.0574764  0.0304584  -1.887   0.0593 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6786 on 1986 degrees of freedom
##   (242 observations deleted due to missingness)
## Multiple R-squared:  0.01509,    Adjusted R-squared:  0.0136
## F-statistic: 10.14 on 3 and 1986 DF,  p-value: 1.247e-06
```

```
summary(lm4)
```

```
##
## Call:
## lm(formula = as.numeric(rwork.lab) ~ income + age + female, data = d,
##     subset = V2 == 156 & !is.na(married))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6219 -0.4151 -0.1022  0.6452  1.1703
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.8298653  0.0668944  42.303  < 2e-16 ***
## income      -0.0006624  0.0085014  -0.078    0.938
## age         -0.0115172  0.0010717 -10.747  < 2e-16 ***
## female      -0.1350550  0.0312928  -4.316 1.67e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.683 on 1904 degrees of freedom
##   (392 observations deleted due to missingness)
## Multiple R-squared:  0.0656, Adjusted R-squared:  0.06413
## F-statistic: 44.56 on 3 and 1904 DF,  p-value: < 2.2e-16
```

The varialbe "V8" is renamed as "work", which indicates how important the work is in your life. The variable "V239" is renamed as "income", which asks where you place yourself in the income distribution in your country. Here is a regression predicting if you think your work is important in your life as a fucntion of income, sex and ages. I also did thsi for the United States and China separately and only if people also answered about the whant job indusry they are in.

From the resuts above, we can draw several conclusions.

For the United States, the estimate of "income" is -0.0128126, which indicates that for each step higher of people's income, they think 0.013 points less importance of their work in their life. THe estimate of "age" is -0.0044753, which indicates that for each higer age people is, they think 0.004 points less importance of their work in their life.

For China, the relationship trend is similar, but the results are different. The estimate of "income" is -0.0006624, much smaller, which indicates that for each step higher of people's income, they think 0.0007 points less importance of their work in their life. The estimate of "age" is -0.0115172, which indicates that for each higer age people is, they think 0.0115 points less importance of their work in their life.

Compared the results in two countries, we can see that the impact of income on people's attitude of work in United States is much larger than in that in China. However, the impact of age is opposite, much larger in China. I assume that this difference is somewhat caused by the diffenrence of labor market regulation and culture in these two countries.

4. Now add another independent variable to that model in Question 3, preferably a set of dummy variables. Tell me why you added that new set of variables and what effect you expected them to have. Did they have an effect? Interpret that new model.

```
lm5 = lm(as.numeric(rwork.lab) ~ income + age + female + married, d, subset = V2 == 840)
summary(lm5)
```

```
##
## Call:
## lm(formula = as.numeric(rwork.lab) ~ income + age + female +
##     married, data = d, subset = V2 == 840)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4239 -0.2958 -0.1653  0.7224  0.9644
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.5353986  0.0628560  40.337  < 2e-16 ***
## income      -0.0100071  0.0081700  -1.225   0.2208
## age         -0.0040762  0.0009545  -4.270 2.04e-05 ***
## female      -0.0590714  0.0304562  -1.940   0.0526 .
## married     -0.0568471  0.0324565  -1.751   0.0800 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6783 on 1985 degrees of freedom
##   (242 observations deleted due to missingness)
## Multiple R-squared:  0.01661,    Adjusted R-squared:  0.01463
## F-statistic: 8.383 on 4 and 1985 DF,  p-value: 1.052e-06
```

```
lm6 = lm(as.numeric(rwork.lab) ~ income + age + female + married, d, subset = V2 == 156)
summary(lm6)
```

```
##
## Call:
## lm(formula = as.numeric(rwork.lab) ~ income + age + female +
##     married, data = d, subset = V2 == 156)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6018 -0.4100 -0.1004  0.6406  1.2063
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.815227   0.069275  40.639  < 2e-16 ***
## income      -0.001165   0.008525  -0.137    0.891
## age         -0.011792   0.001124 -10.493  < 2e-16 ***
## female      -0.134784   0.031297  -4.307 1.74e-05 ***
## married      0.034875   0.042841   0.814    0.416
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.683 on 1903 degrees of freedom
##   (392 observations deleted due to missingness)
## Multiple R-squared:  0.06593,    Adjusted R-squared:  0.06396
## F-statistic: 33.58 on 4 and 1903 DF,  p-value: < 2.2e-16
```

We have added in marital status. As we know, marital status may change people's attitude on the work and make them keep a balance between their work and family in their life. So I want to figure out if marital status can make a difference on the analysis results. I expect that for married people, the impact of income and age on their attitude towards work is less.

From the results, we can see that for the United States, the estimate of "income" is -0.0100071, which indicates that for each step higher of people's income, they think 0.010 points less importance of their work in their life. The estimate of "age" is -0.0040762, which indicates that for each higer age people is, they think 0.004 points less importance of their work in their life.

For China, the relationship trend is similar, but the results are different. The estimate of "income" is -0.001165, much smaller, which indicates that for each step higher of people's income, they think 0.0012 points less importance of their work in their life. The estimate of "age" is -0.011792, which indicates that for each higer age people is, they think 0.0118 points less importance of their work in their life.

5. Now run a partial F test comparing the model in Question 3 to the model in Question 4. Does the F test support the idea of adding those new variables? Why? Why not?

```
anova(lm3, lm5)
```

```
## Analysis of Variance Table
##
## Model 1: as.numeric(rwork.lab) ~ income + age + female
## Model 2: as.numeric(rwork.lab) ~ income + age + female + married
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1   1986 914.58
## 2   1985 913.17  1    1.4112 3.0677 0.08002 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm4, lm6)
```

```
## Analysis of Variance Table
##
## Model 1: as.numeric(rwork.lab) ~ income + age + female
## Model 2: as.numeric(rwork.lab) ~ income + age + female + married
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1904 888.14
## 2   1903 887.83  1   0.30918 0.6627 0.4157
```

The Probability is $0.08002 > 0.05$, and $0.4157 > 0.05$, which indicates that we cannot conclude that a significant difference exists. However, I prefer to say that 0.08 is close to 0.05, so it is at the margin of statistical significance. The results somewhat support the idea of adding the variable of marital status.

The effect in the United States is for my expectation, however, the effect in China is against my expectation. The reason is likely that the traditional views on marriage and family in China gives them more life pressure and make them think more importance of work in their life. Also, the existing social security system in China makes people rely more on their job salary.