QMSSGR5015-Lab4-Yue_Ma

Yue Ma

10/24/2018

```
library(plyr)
setwd("/Users/mayue/Desktop/qmss/data analysis/lab4")
w = read.csv("WVS.csv")
w = rename(w, c("V70" = "creation"))
table(w$creation)
```

```
##
##     1     2     3     4     5     6
## 15197 18770 16939 10653  7606  2521
```

```
w = rename(w, c("V71" = "money"))
table(w$money)
```

```
##
##     1     2     3     4     5     6
##  6200 10638 13013 13232 18574 10358
```

```
w = rename(w, c("V242" = "age"))
w = rename(w, c("V2" = "country"))
```

The variable "V70" is related to respondents' attitude on creation of their life. The question asks the respondents if it is important to this person to think up new ideas and be creative; to do things one's own way. 1 is "Very much like me", and 6 is "Not at all like me". The variable "V71" is related to respondents' attitude on money of their life. The question asks the respondents if it is important to this person to be rich; to have a lot of money and expensive things. 1 is "Very much like me", and 6 is "Not at all like me". The tables above show the summary results of the answers.

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

```
lm1 = lm(creation ~ age + as.factor(country), data = w, subset = (country == 276 | country == 392))
summary(lm1)
```

```
##
## Call:
## lm(formula = creation ~ age + as.factor(country), data = w, subset = (country ==
##     276 | country == 392))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8734 -0.8268  0.1650  1.0005  3.3048
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.448422   0.061922  39.540  < 2e-16 ***
## age                      0.008225   0.001124   7.315 3.06e-13 ***
## as.factor(country)392 0.775238   0.038025  20.388  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.229 on 4182 degrees of freedom
##   (304 observations deleted due to missingness)
## Multiple R-squared:  0.1022, Adjusted R-squared:  0.1018
## F-statistic: 238.1 on 2 and 4182 DF,  p-value: < 2.2e-16
```

```
lm2 = lm(money ~ age + as.factor(country), data = w, subset = (country == 276 | country == 392))
summary(lm2)
```

```
##
## Call:
## lm(formula = money ~ age + as.factor(country), data = w, subset = (country ==
##     276 | country == 392))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0496 -0.7351  0.0834  0.8739  2.5024
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.246237   0.056222   57.74   <2e-16 ***
## age                   0.013967   0.001018   13.72   <2e-16 ***
## as.factor(country)392 0.979379   0.034448   28.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.126 on 4292 degrees of freedom
##   (194 observations deleted due to missingness)
## Multiple R-squared:  0.1925, Adjusted R-squared:  0.1921
## F-statistic: 511.5 on 2 and 4292 DF,  p-value: < 2.2e-16
```

```
w = rename(w, c("V10" = "happy"))

lm3 = lm(money ~ age + as.factor(country) + happy, data = w, subset = (country == 276 | country == 392))
summary(lm3)
```

```
##
## Call:
## lm(formula = money ~ age + as.factor(country) + happy, data = w,
##     subset = (country == 276 | country == 392))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0154 -0.7332  0.0800  0.8713  2.5435
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.159548   0.074272  42.541   <2e-16 ***
## age                   0.014125   0.001031  13.705   <2e-16 ***
## as.factor(country)392 0.979804   0.034927  28.053   <2e-16 ***
## happy                 0.042675   0.026882   1.587    0.112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.127 on 4212 degrees of freedom
##   (273 observations deleted due to missingness)
## Multiple R-squared:  0.1934, Adjusted R-squared:  0.1929
## F-statistic: 336.7 on 3 and 4212 DF,  p-value: < 2.2e-16
```

Firstly, I want to take a look at the question that if one's age and country (Germany vs. Japan) affect his or her attitude on creation of life. From the results, we can see that the p-value is < 2.2e-16, so it is statitically significally. And the estimate of "age" is 0.008225. It indicates that for every age increasing, a person believes it is 0.008 less important to think up new ideas and be creative; to do things one's own way. The estimate of "country" is 0.775238. It indicates that Japanese respondents will think creation 0.78 unit less important than German respondents. The results fit my expectation, because people are always more creative during their young ages and become more mature with age increasing. Also, due to different culture, western education system pay more attention to creation development on average.

Then, I want to take a look at the question that if one's age and country (Germany vs. Japan) affect his or her attitude on money of life. From the results, we can see that the p-value is < 2.2e-16, so it is statitically significally. And the estimate of "age" is 0.013967. It indicates that for every age increasing, a person believes it is 0.001 less important to be rich; to have a lot of money and expensive things. The estimate of "country" is 0.979379. It indicates that Japanese respondents will think money 0.98 unit less important than German respondents. The results fit my expectation as well, because people get more life experience with age increasing and realize that there are many other things that are more important than money in life. In this question, respondents from Japan and Germany have similar answers.

Perhaps people who think money less important are just more happy and optimistic, so I add variable "V10" of happiness level as well. "V10" is the question that taking things all together, would you say you are: where 1 is "Very happy", and 4 is "Not at all happy". From the results, we can see that p-value of variable "happy" is 0.112, which is more than 0.05, so it is not statistically significant. As a result, the happiness and optimism level does not affect people's attitude on money so much.

2. Add an interaction term to that model that you think might moderate the original relationship between X1 and X2. Explain why you think an interaction might be present and in what direction it would work. Explain your results. Did it work out? Yes? No?

```r
lm4 = lm(money ~ age*as.factor(country) + happy, data = w, subset = (country == 276 | country == 392))
summary(lm4)
```

```
##
## Call:
## lm(formula = money ~ age * as.factor(country) + happy, data = w,
##     subset = (country == 276 | country == 392))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9641 -0.8008  0.0757  0.9208  2.7476
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.845291   0.086411  32.927  < 2e-16 ***
## age                          0.021010   0.001420  14.795  < 2e-16 ***
## as.factor(country)392        1.698685   0.108348  15.678  < 2e-16 ***
## happy                        0.028943   0.026802   1.080     0.28
## age:as.factor(country)392 -0.014380   0.002053  -7.005 2.87e-12 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.12 on 4211 degrees of freedom
##   (273 observations deleted due to missingness)
## Multiple R-squared:  0.2027, Adjusted R-squared:  0.202
## F-statistic: 267.7 on 4 and 4211 DF,  p-value: < 2.2e-16
```

```
anova(lm3, lm4)
```

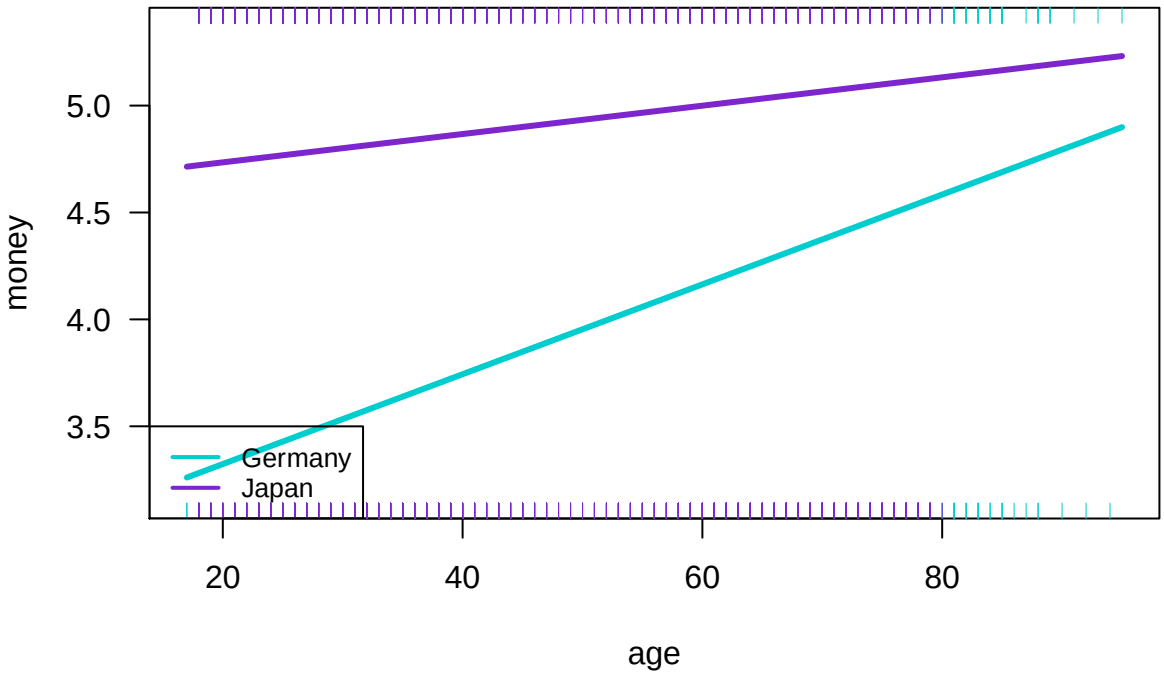
```
## Analysis of Variance Table
##
## Model 1: money ~ age + as.factor(country) + happy
## Model 2: money ~ age * as.factor(country) + happy
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   4212 5348.0
## 2   4211 5286.4  1    61.594 49.063 2.872e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lm5 = lm(money ~ age + I(age^2) + as.factor(country) + happy, data = w, subset = (country == 276 | coun
summary(lm5)
```

```
##
## Call:
## lm(formula = money ~ age + I(age^2) + as.factor(country) + happy,
##     data = w, subset = (country == 276 | country == 392))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0503 -0.7584  0.0255  0.8524  2.6796
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.737e+00  1.497e-01  18.278  < 2e-16 ***
## age                     3.359e-02  6.082e-03   5.522 3.55e-08 ***
## I(age^2)               -1.953e-04  6.016e-05  -3.247  0.00118 **
## as.factor(country)392   9.696e-01  3.503e-02  27.679  < 2e-16 ***
## happy                   4.196e-02  2.685e-02   1.563  0.11823
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.126 on 4211 degrees of freedom
##   (273 observations deleted due to missingness)
## Multiple R-squared:  0.1954, Adjusted R-squared:  0.1947
## F-statistic: 255.7 on 4 and 4211 DF,  p-value: < 2.2e-16
```

```
w2 <- subset(w, country == 276 | country == 392) ## subset the data to use in visreg below ##

library(visreg)
visreg(lm(money ~ age*as.factor(country) + happy, data = w2),
       xvar = "age", by = "country", overlay = T, partial = F, band = F, legend = F,
       line = list(col = c("cyan3", "purple3")))
legend("bottomleft", c("Germany", "Japan"), lwd = 2, col = c("cyan3", "purple3"), cex = 0.8)
```



Firstly, I add an interaction term to that model that I think might moderate the original relationship to see if age and country may interact to predict the attitude on money in life. From the results, we can see that net of other factors, Japanese respondents gain back -0.014380 when they get age increasing. When the respondent is from Germany(country == 276), they will think 0.021 less important on money as the age increases. So when it gains back -0.014380, Japanese respendents ultimately will think 0.00662 (0.021-0.014380) less important on money as the age increases.

Then I see if adding the interaction improves my model. From the results, we can see that the probability is 2.872e-12 and less than 0.001, which indicates that it is statistically significant. As a result, I think that it is better to add the interaction, thus improving my model.

Next, I include a quadratic on age. From the results, the estimate of age square is -1.953e-04. It indicates that net of other factors, with age increases, at first, people think money less and less important, but then at a certain point, for each age point squared, people think money more and more important. The results did not reach my expectation. Meantime, it reduces the adjuste R-squared, so I do not think that adding the quadratic on age can improve my model.

Finally, I display the results of the fitted model with interaction term but without quadratic term. The results showed in the plot reaches my expectation and fits the analysis above. Net of other factors, people think money on life less important as the age increases and age affect more on German than on Japanese.

3. Give me an update on your independent project. What do you plan to investigate? What are your hypotheses? What data are you using? How can we help?

My project question is "What are the effects of pay gap between the executive and employees of listed companies on corporate performance?" I successfully searched a dataset related to Chinese listed companies in CSMAR database and it has the R-readable format. The CSMAR is China Stock Market Financial Statements Database, it contains a list of all financial indicators of listed companies in China I need, including EBIT, ROTA, Total Net Assets, etc. On the Office for National Statistics (http://www.ons.gov.uk), I also got some related datasets, however, it needs more clean manipulation. I will continue to search related dataset of listed companies in the United States through Bloomberg. Then, I will narrow my research topic to a specific region.

My hypotheses are as below: Hypothesis 1: There is an "inverted U-shaped" curve relationship between the pay gap between executives and general staff of listed companies in China and corporation performance. Hypothesis 2: The "inverted U-shaped" curve relationship exists in different countries and regions with different inflection points.