Jiaqi report

Li Sun

October 12, 2015

Brief introduction

This is data analysis by multiple linear regression on small scale survey on how the asian or western views affecting the propensity of seeking professional counselling.

Load packages

load necessary packages

Load data

Data comes from Dr. Jiaqi Li in csv file format.

Data cleaning

1. variables needed to be edited: 'age', 'stay'.

- There is one value which is "adult", we change that one to 18, according to the info that he has 89 TOFLE score and stayed in US for 18 month. We change this to 25, which is average age of F1 student who came here after undergraduate (23.5) and master study(26.5). There is also another invalid entry which is 3. We believe it is imput error. and change it to 30
- For stay, there is an entry which is "whole life", we change that to 12 * age.

2. Construct new variable "SL-ASIA values score" ("SLval"), based on the SUINN-LEW ASIA SCALE (question 22~23) coding: 1:Do not believe . . . . . . 5:strongly believe

```
##
##  A  B  N  W
## 33 51  7 21
```

3. Construct new variable "SL-ASIA behavioral competencies score" ("SLcom"), based on the SUINN-LEW ASIA SCALE (question 24~25): 1:do not fit . . . . . . 5:fit very well

```
##
##  A  B  N  W
## 40 55  3 14
##       SLcom4
## SLval4  A  B  N  W
##      A 19 12  0  2
##      B 15 27  1  8
##      N  2  3  1  1
##      W  4 13  1  3
```

4. Construct new variable "SL-ASIA self-identity score" ("sla_id"), based on the SUINN-LEW ASIA SCALE (question 26) coding

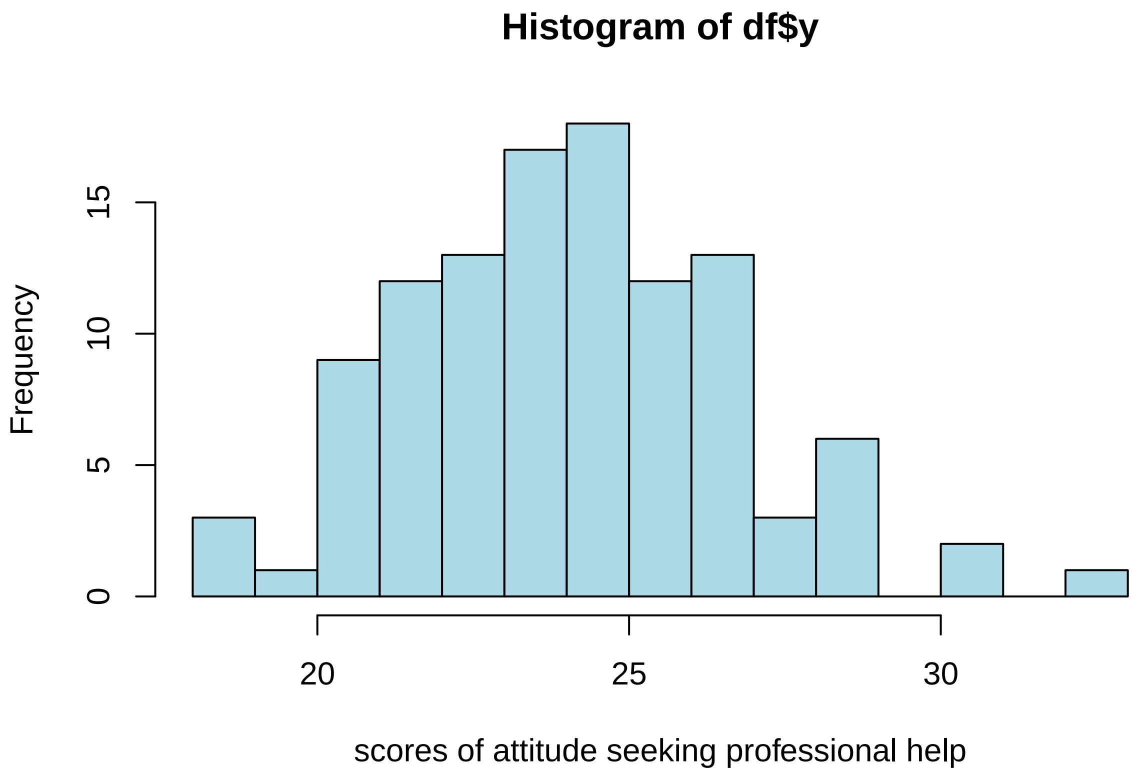
- 1. I consider myself basically an Asian person (e.g., Chinese, Japanese, Korean, Vietnamese, etc.). Even though I live and work in America, I still view myself basically as an Asian person.
- 5. I consider myself basically as an American. Even though I have an Asian background and characteristics, I still view myself basically as an American.
- 7. I consider myself as an Asian-American, although deep down I always know I am an Asian.
- 9. I consider myself as an Asian-American, although deep down, I view myself as an American first.
- 10. I consider myself as an Asian-American. I have both Asian and American characteristics, and I view myself as a blend of both.

We found 0.8571429 responders chose 1 and there is one value "2" which is not included in coding rubric and we will change it to 9 because 9 is missing here. And as suggested we will treat this variable as numeric. From 1 to 5, 1 is very asian and 5 is very american identification.

5. Construct new val indicating individual attitude to counseling. for all original question values:

- 1. strongly disgree
- 2. disgree
- 3. agree
- 4. strongly agree

Calculating based on Whittlesey, V. (2001). Diversity activities for psychology. Boston: Allyn and Bacon, and Fischer, E., and Farina, A. (1995). Attitudes toward seeking psychological professional help: A shortened form and considerations for research. Journal of College Student Development, 36, 368-373.



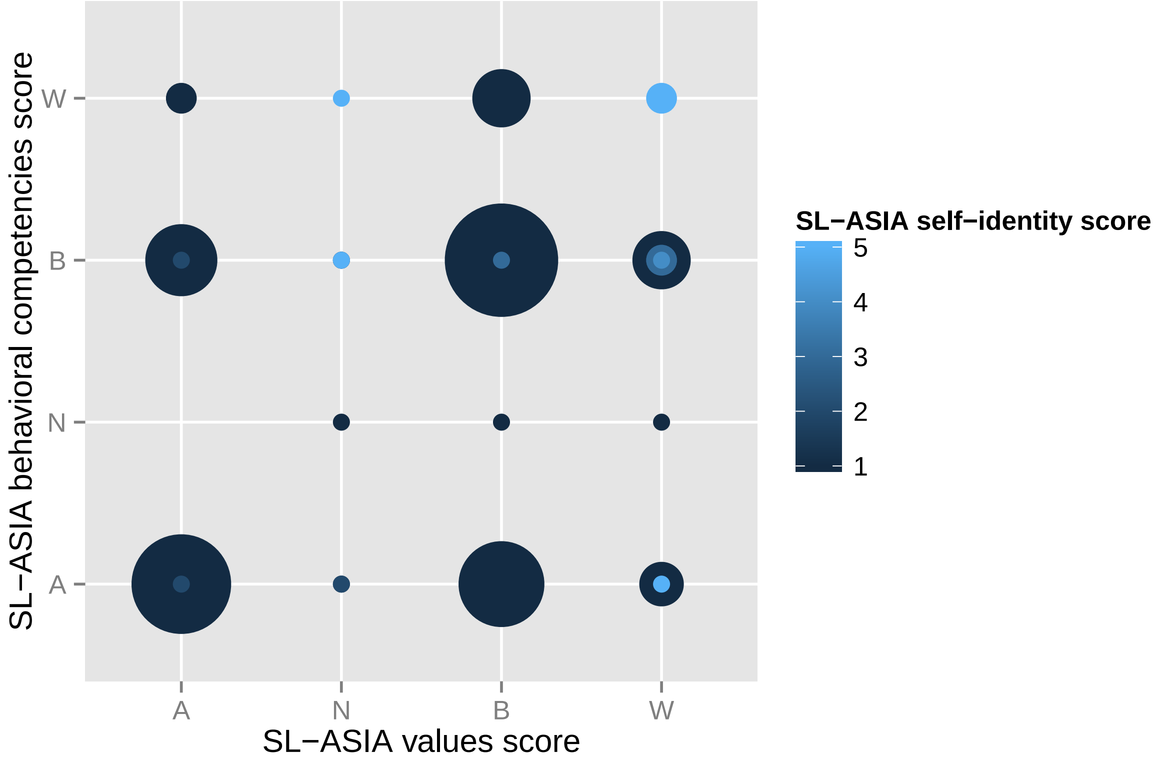
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   18.00   23.00   24.50   24.56   26.00   33.00      32 
```

Exploratory data analysis

Visualize the relationship among the 3 scores from "Suinn-Lew Asian Self Identity Acculturation"

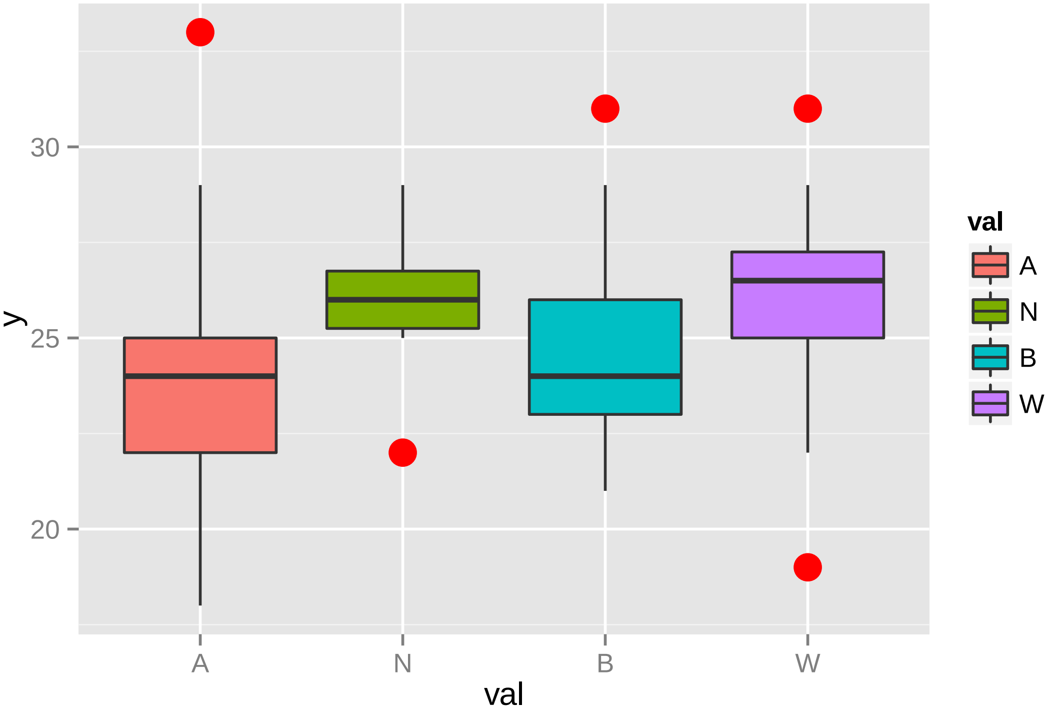
1. Visualize the relationship between different grouping methods of individuals acculturation

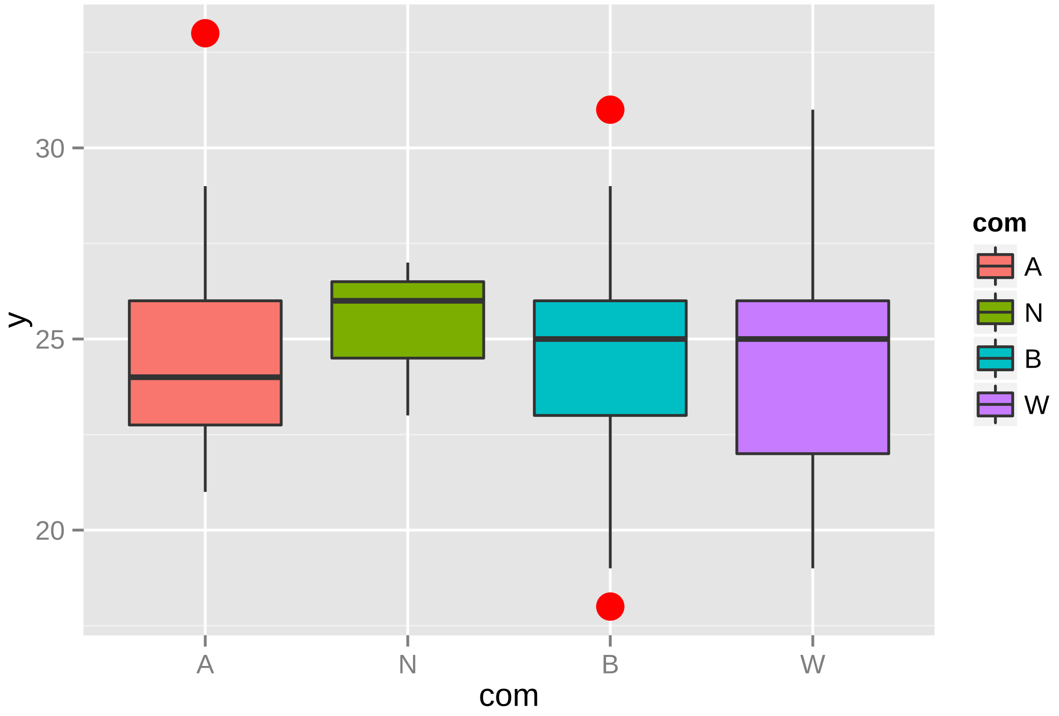
3 ACCULTURATION SCORES based on 7 groupss

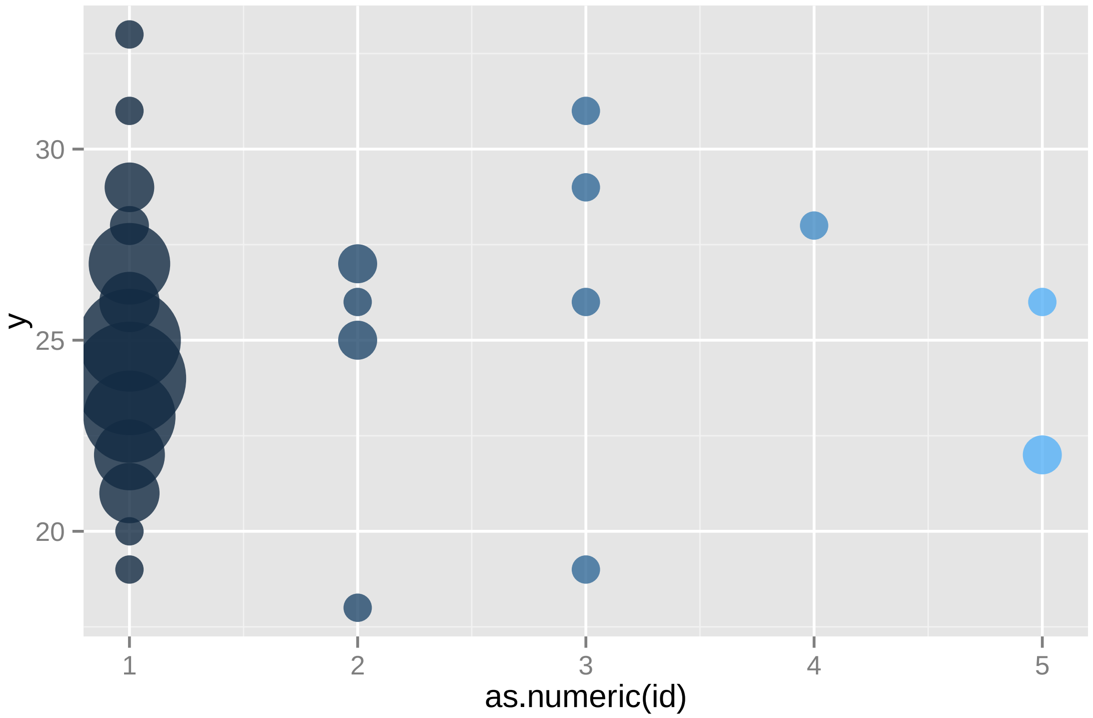


We found that the most asian students identify themselves as asian no matter how good they fit in western life.

2. Visualize the relationship between the 3 scores and attitudes for seeking professional counseling. This need







```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## SLval4         3   68.7  22.905   3.467 0.0188 *
## Residuals   106  700.3   6.607
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 32 observations deleted due to missingness

##                 Df Sum Sq Mean Sq F value Pr(>F)
## SLcom4        3    2.4   0.791   0.109  0.954
## Residuals   106  766.7   7.233
## 32 observations deleted due to missingness
```

```
## 
## Call:
## lm(formula = y ~ sla_id, data = df)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -6.6633 -1.5280 -0.0956  1.4720  8.4720 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  24.3927     0.4730  51.573   <2e-16 ***
## sla_id        0.1353     0.3156   0.429    0.669    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.666 on 108 degrees of freedom
##   (32 observations deleted due to missingness)
## Multiple R-squared:  0.001699,   Adjusted R-squared:  -0.007545 
## F-statistic: 0.1838 on 1 and 108 DF,  p-value: 0.669
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## 
## Fit: aov(formula = y ~ SLval4, data = df)
## 
## $SLval4
##          diff        lwr      upr     p adj
## B-A 0.5347594 -0.9641736 2.033692 0.7881643
## N-A 2.0151515 -0.9625717 4.992875 0.2952344
## W-A 2.1318182  0.2305184 4.033118 0.0214901
## N-B 1.4803922 -1.4153596 4.376144 0.5432242
## W-B 1.5970588 -0.1731060 3.367224 0.0923380
## W-N 0.1166667 -3.0063958 3.239729 0.9996659
```

The only significant result is difference of response between people with asian value and western value

(3D plot here: https://plot.ly/~rikku1983/35/visualization-of-acculturation-and-attitude-for-seeking-professional-counseling/?share_key=0poL7ODE8G2eg9itKxkbs6)

Making data ready for modeling

1. Missing values For convenience, we just remove all rows with NAs. We end up having a data with 110 observations and 10 variable.

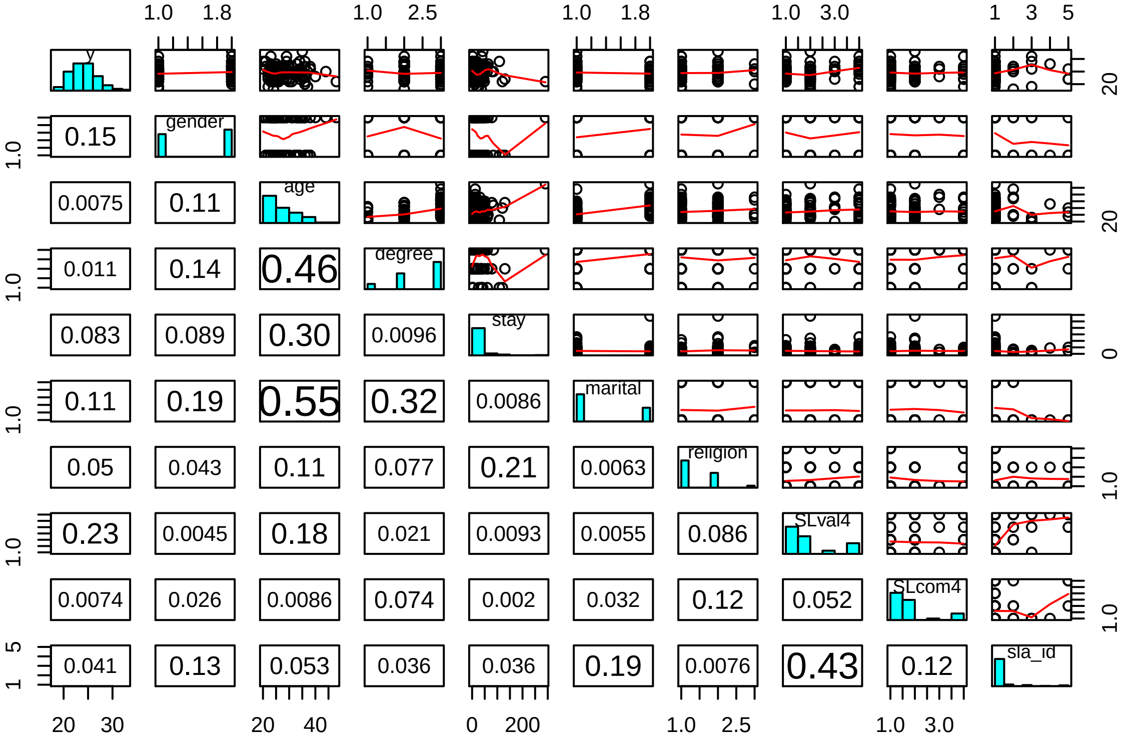
2. Convert all variable type to ones ready for analysis

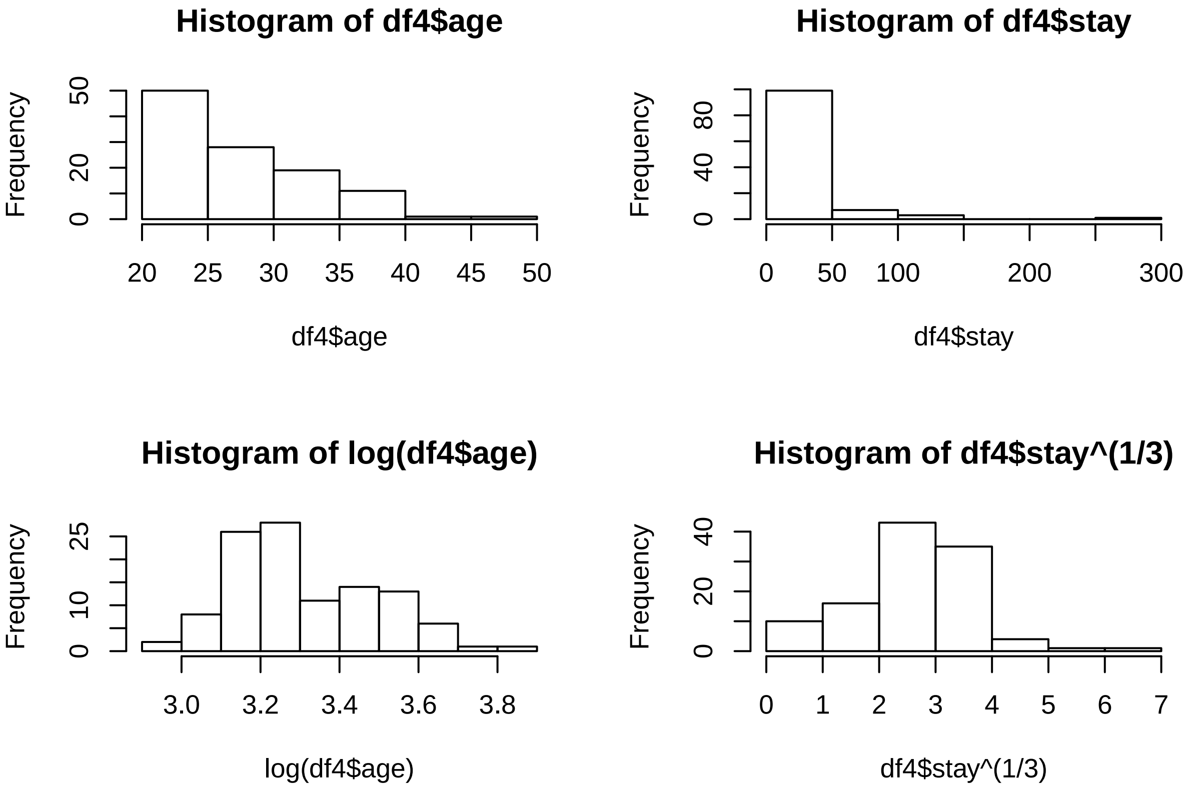
```
##         y    gender       age    degree      stay   marital  religion 
## "numeric"  "factor" "numeric"  "factor" "numeric"  "factor"  "factor" 
##    SLval4    SLcom4    sla_id 
##  "factor"  "factor" "numeric" 
```

Analysis of relationship between each variables

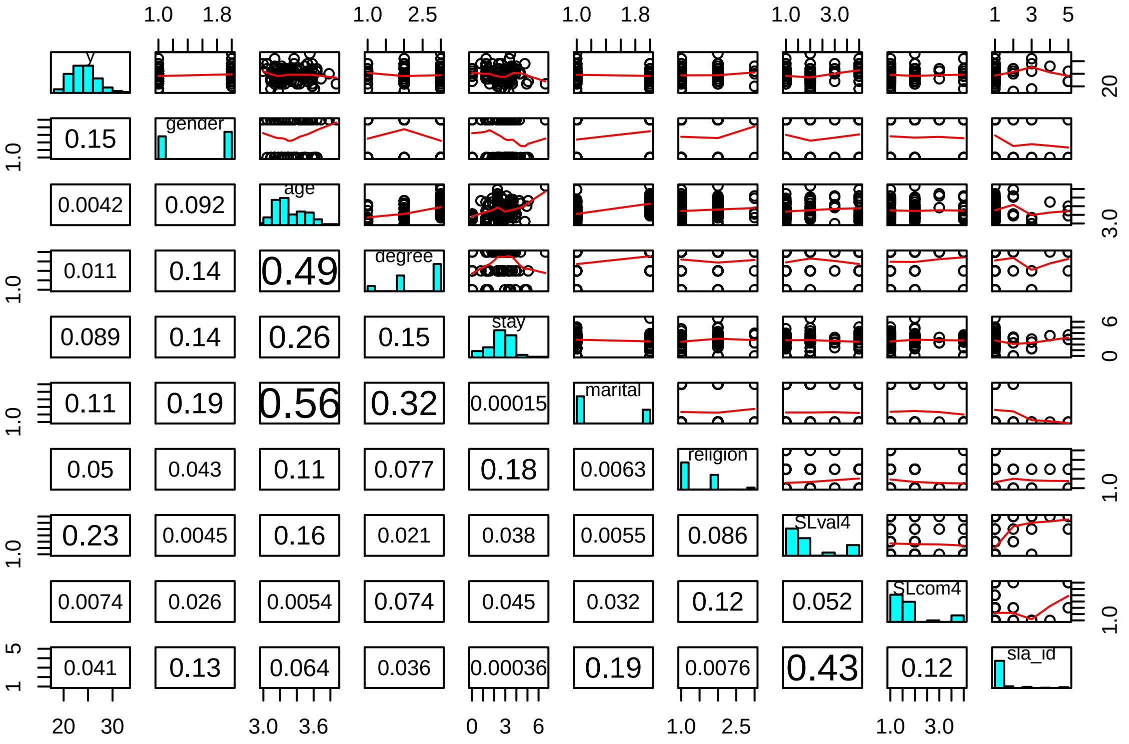
Association between different variables

In this part, we start to look into relationship between different variables in this table by studying there correlations





After transformation of two numeric variables: stay and age



Compare df4 and df5, the correlation between age and response drop from 0.0075 to 0.0042, and the correlation between stay and response increase from 0.083 to 0.089. Basically, not very significant change were observed. So we will live with non-transformed data. Among our predictors, we observe highest correlations between age and marital status (0.55) and, age and degree(0.46).

Due to the correlations showed in this figuer is just pearson correlation which might not be appropriate for categorical data. So we will build the effect size matrix by using more appropriate measure for different type of data.

for numeric vs numeric: Pearson's correlation is used, absolute value of this r is categorized as followed, Effect size r

Small | Medium | Large ———|————–|——- 0.10 | 0.30 | 0.5

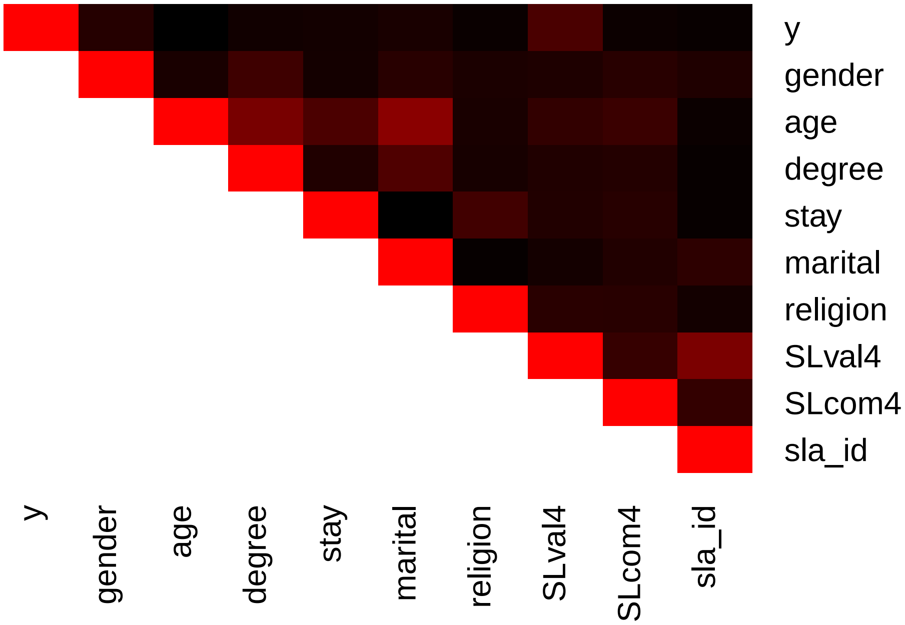
for numeric vs categorical: R square from one-way ANOVA is taken and the square root value is used so that we can compare it to other effect size

for categorical vs categorical: Cramer's V

df*	small	medium	large
1	.10	.30	.50
2	.07	.21	.35
3	.06	.17	.29

Visualizing the association

	y	gender	age	degree	stay	marital	religion	SLval4	SLcom4	sla_
y	1	0.1531631	0.0075212	0.0751640	0.0825411	0.1080973	0.0510631	0.2989148	0.0555590	0.04121
gender	NA	1.0000000	0.1065054	0.2493480	0.0887596	0.1668671	0.1151840	0.1267234	0.1640109	0.13189
age	NA	NA	1.0000000	0.4751252	0.3030065	0.5471982	0.1087090	0.2043054	0.2377670	0.05349
degree	NA	NA	NA	1.0000000	0.1331011	0.3158574	0.0990311	0.1343610	0.1465046	0.03642
stay	NA	NA	NA	NA	1.0000000	0.0086450	0.2622564	0.1373329	0.1629078	0.03631
marital	NA	NA	NA	NA	NA	1.0000000	0.0324457	0.0874728	0.1398424	0.18525
religion	NA	NA	NA	NA	NA	NA	1.0000000	0.1677118	0.1663755	0.08341
SLval4	NA	NA	NA	NA	NA	NA	NA	1.0000000	0.2218145	0.48696
SLcom4	NA	NA	NA	NA	NA	NA	NA	NA	1.0000000	0.20893
sla_id	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.00000



Fit regression

1. Model without interaction

```
##
## Call:
## lm(formula = y ~ gender + marital + SLval4, data = df4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6717 -1.5493 -0.0574  1.3081  8.4456
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     24.1113     0.4732  50.951   <2e-16 ***
## genderwoman      0.8625     0.4998   1.726   0.0874 .
## maritalmarried  -0.7973     0.5245  -1.520   0.1316
## SLval4A         -0.4194     0.5717  -0.734   0.4648
## SLval4N          1.6894     1.1018   1.533   0.1283
## SLval4W          1.5603     0.6707   2.326   0.0219 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.541 on 104 degrees of freedom
## Multiple R-squared:  0.1267, Adjusted R-squared:  0.0847
## F-statistic: 3.017 on 5 and 104 DF,  p-value: 0.01388
```

2. Model with interactions

Model with all possible interactions after backward selection, has much higher R squared but also much more predictors, and the design matrix is not full rank any more. So colinearity and multicolinearity is brought in. Let's try less interaction. We will try just one variable interacting with all others to see if there would be any improvement of r squared.

3. more parsimoneous model

```
##
## Call:
## lm(formula = y ~ degree + stay + marital + SLval4 + sla_id +
##     sla_id:age + degree:sla_id + degree:age + stay:SLval4, data = df4)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.979 -1.266 -0.161  1.303  7.965
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            27.20910    1.61749  16.822  < 2e-16 ***
## degreemaster          -16.28808    4.94447  -3.294  0.00140 **
## degreedoctoral        -12.67931    4.65295  -2.725  0.00768 **
## stay                    0.03273    0.01575   2.078  0.04043 *
## maritalmarried         -1.56708    0.63258  -2.477  0.01504 *
## SLval4A                0.68564    0.73452   0.933  0.35300
## SLval4N                4.51769    2.04907   2.205  0.02994 *
## SLval4W                4.17352    0.97731   4.270  4.7e-05 ***
## sla_id                 3.21046    2.03763   1.576  0.11852
## sla_id:age            -0.25015    0.09402  -2.661  0.00919 **
## degreemaster:sla_id    2.39256    1.13481   2.108  0.03769 *
## degreedoctoral:sla_id  3.98390    1.41284   2.820  0.00587 **
## degreemaster:age       0.52561    0.17489   3.005  0.00341 **
## degreedoctoral:age     0.34156    0.13727   2.488  0.01462 *
## stay:SLval4A          -0.04641    0.01801  -2.576  0.01156 *
## stay:SLval4N          -0.14936    0.10345  -1.444  0.15215
## stay:SLval4W          -0.08663    0.02520  -3.437  0.00088 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.382 on 93 degrees of freedom
## Multiple R-squared:  0.3139, Adjusted R-squared:  0.1958
## F-statistic: 2.659 on 16 and 93 DF,  p-value: 0.001756
```

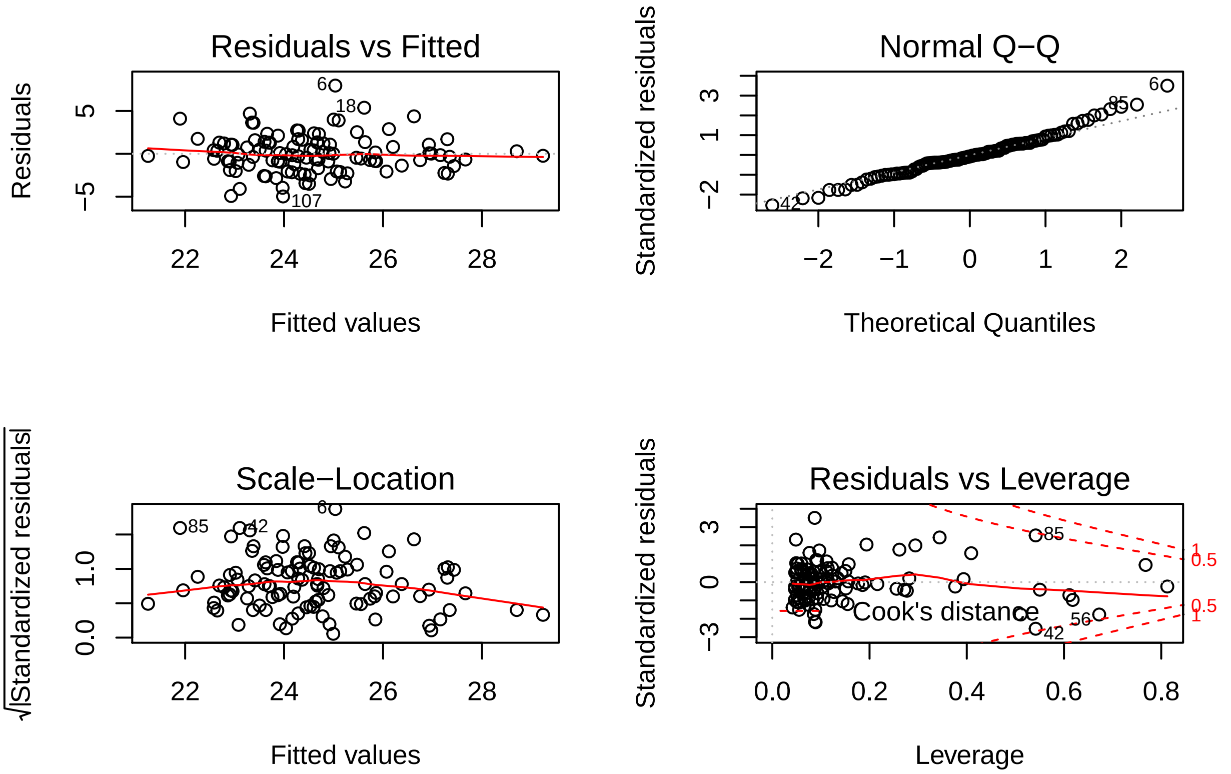
All models we have

Model name	Description
fmstep:	full model without interactions
fmintstep:	with all possible interactions
betint:	with some interactions

Compare all models

	df	AIC	BIC	Ajusted R2
no interaction	6	525.1810	544.0844	0.0846983
all interaction	75	421.4242	626.6608	0.6979756
some interaction	17	520.6470	569.2557	0.1958169
# Final model				

```
##
## Call:
## lm(formula = y ~ degree + stay + marital + SLval4 + sla_id +
##     sla_id:age + degree:sla_id + degree:age + stay:SLval4, data = df4)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.979 -1.266 -0.161  1.303  7.965
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            27.20910    1.61749  16.822  < 2e-16 ***
## degreemaster          -16.28808    4.94447  -3.294  0.00140 **
## degreedoctoral        -12.67931    4.65295  -2.725  0.00768 **
## stay                      0.03273    0.01575   2.078  0.04043 *  
## maritalmarried           -1.56708    0.63258  -2.477  0.01504 *  
## SLval4A                   0.68564    0.73452   0.933  0.35300    
## SLval4N                   4.51769    2.04907   2.205  0.02994 *  
## SLval4W                   4.17352    0.97731   4.270  4.7e-05 ***
## sla_id                    3.21046    2.03763   1.576  0.11852    
## sla_id:age               -0.25015    0.09402  -2.661  0.00919 ** 
## degreemaster:sla_id       2.39256    1.13481   2.108  0.03769 *  
## degreedoctoral:sla_id     3.98390    1.41284   2.820  0.00587 ** 
## degreemaster:age          0.52561    0.17489   3.005  0.00341 ** 
## degreedoctoral:age        0.34156    0.13727   2.488  0.01462 *  
## stay:SLval4A             -0.04641    0.01801  -2.576  0.01156 *  
## stay:SLval4N             -0.14936    0.10345  -1.444  0.15215    
## stay:SLval4W             -0.08663    0.02520  -3.437  0.00088 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.382 on 93 degrees of freedom
## Multiple R-squared:  0.3139, Adjusted R-squared:  0.1958
## F-statistic: 2.659 on 16 and 93 DF,  p-value: 0.001756
```



Final formula y ~ degree + stay + marital + SLval4 + sla_id + sla_id:age + degree:sla_id + degree:age + stay:SLval4

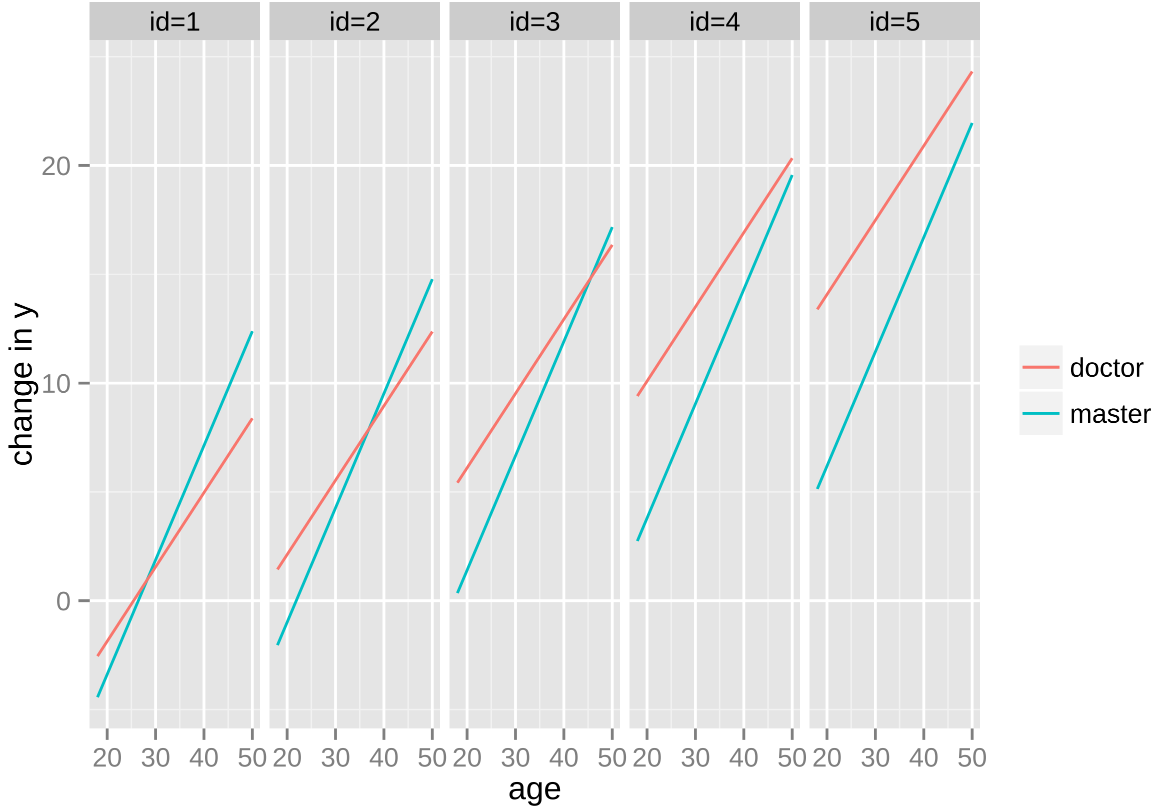
This indicates

1. If a chinese student with age 20, undergraduate degree, just arrive US, not married, got "B" as his value, and got "1" in identification, his altitude for seeking professional counseling is predicted to be 27.2091 + 3.21046 - 0.25015 * 20 = 25.41656

2. Effect of degree Compare degrees assume all other parameters are the same

- 2.1 the difference of y between master degree and undergraduate degree: -16.28808 + 2.39256 * sla_id + 0.52561 * age
- 2.2 the difference of y between doctor degree and undergraduate degree: -12.67931 + 3.98390 * sla_id + 0.34156 * age

Visualize it



2. Effect of stay length(in month) in US

- 3.1 For people with Asian Value (got "A" from question 22 and 23): every one more month people has stayed in US will decrease the y value by 0.01368
- 3.2 For people with Bicultural value (got "B" from question 22 and 23): every one more month people has stayed in US will increase the y value by 0.03273
- 3.3 For people with Western value (got "W" from question 22 and 23): every one more month people has stayed in US will decrease the y value by 0.0539
- 3.4 For people with neither value (got "N" from question 22 and 23): every one more month people has stayed in US will decrease the y value by 0.116628

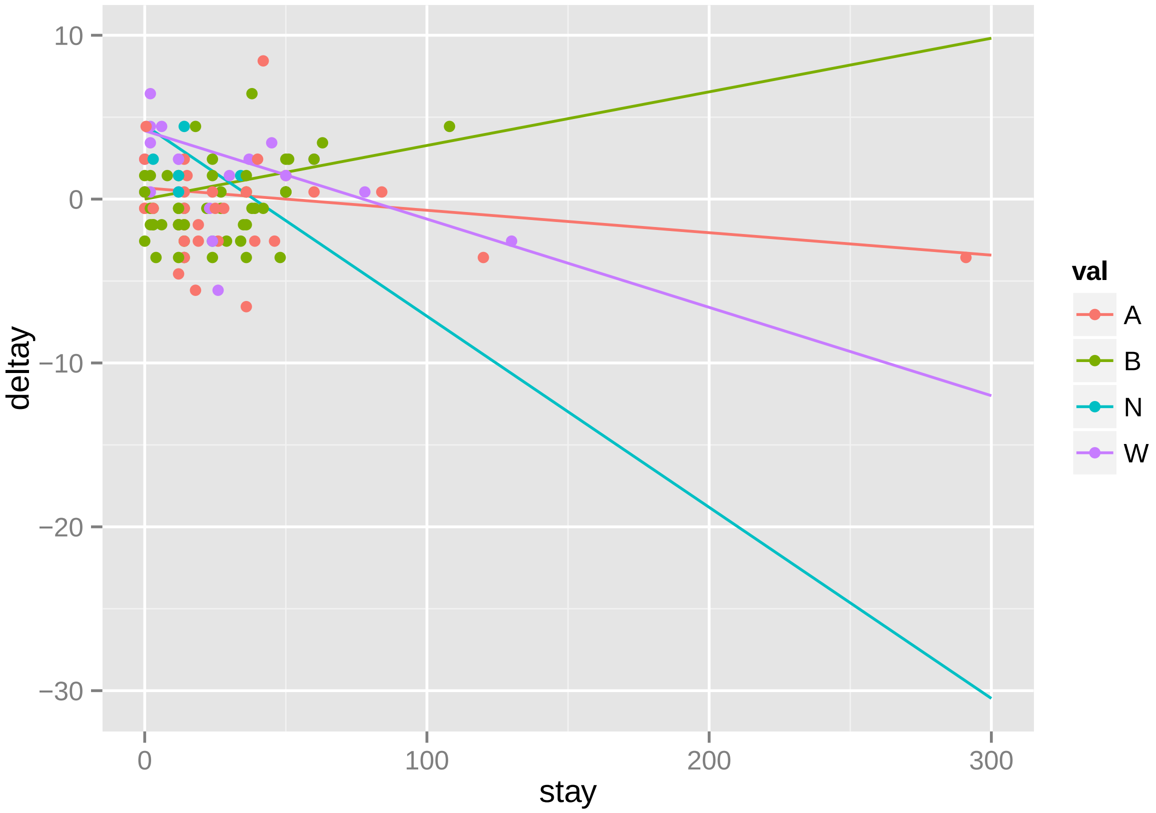
3. married people tend to be less willing for seeking professional counseling, because when all other variable are the same, married people would have 1.56708 less in y value

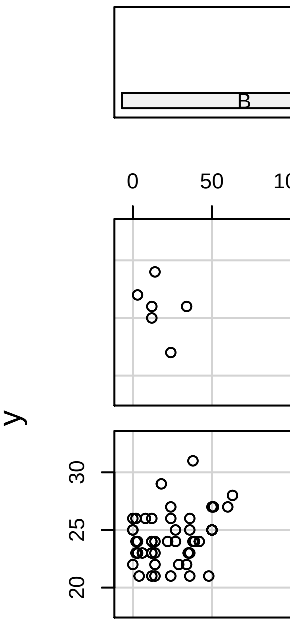
4. Compare different values

- 4.1 the difference of y between W and A: 3.487880 - 0.040218 * stay
- 4.2 the difference of y between B and A: -0.685641 + 0.046410 * stay

- 4.3 the difference of y between N and A: 3.832049 - 0.102949 * stay

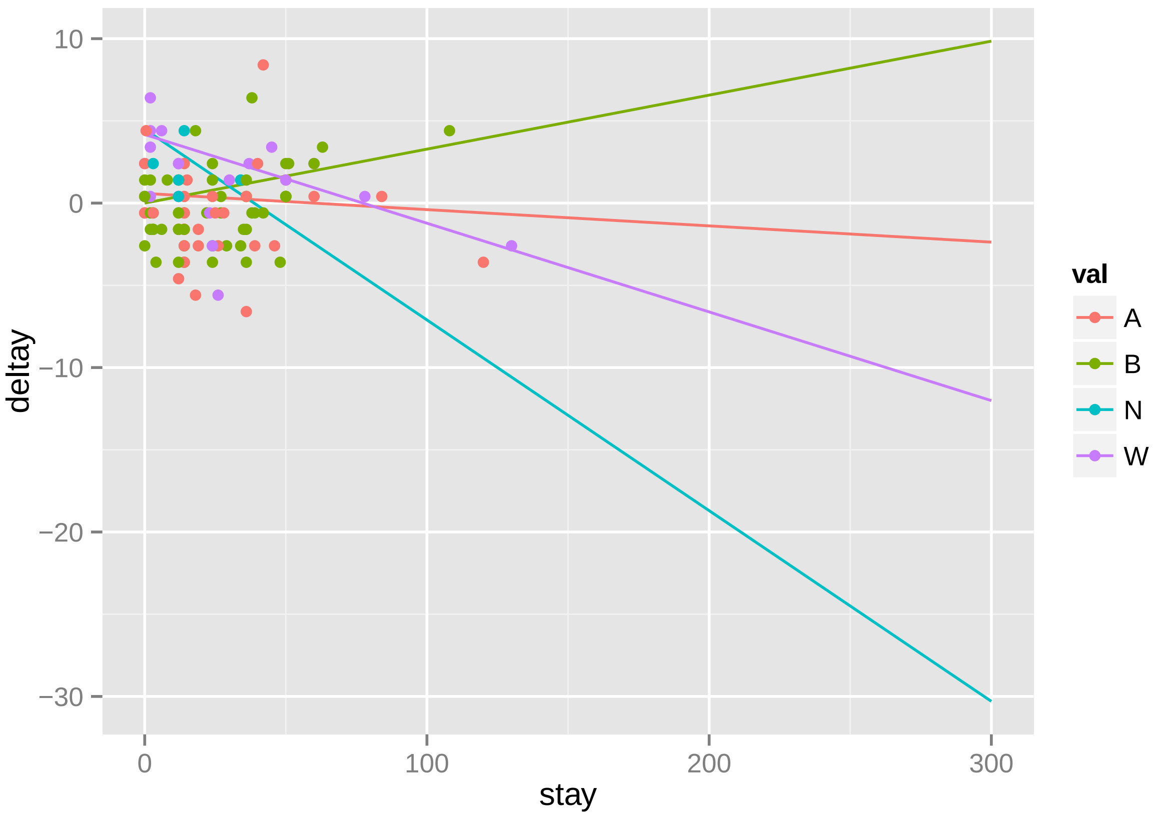
visualize effect of stay and values on changes of y comparing with y given value is “B” and stay=0, and all other variables are the same

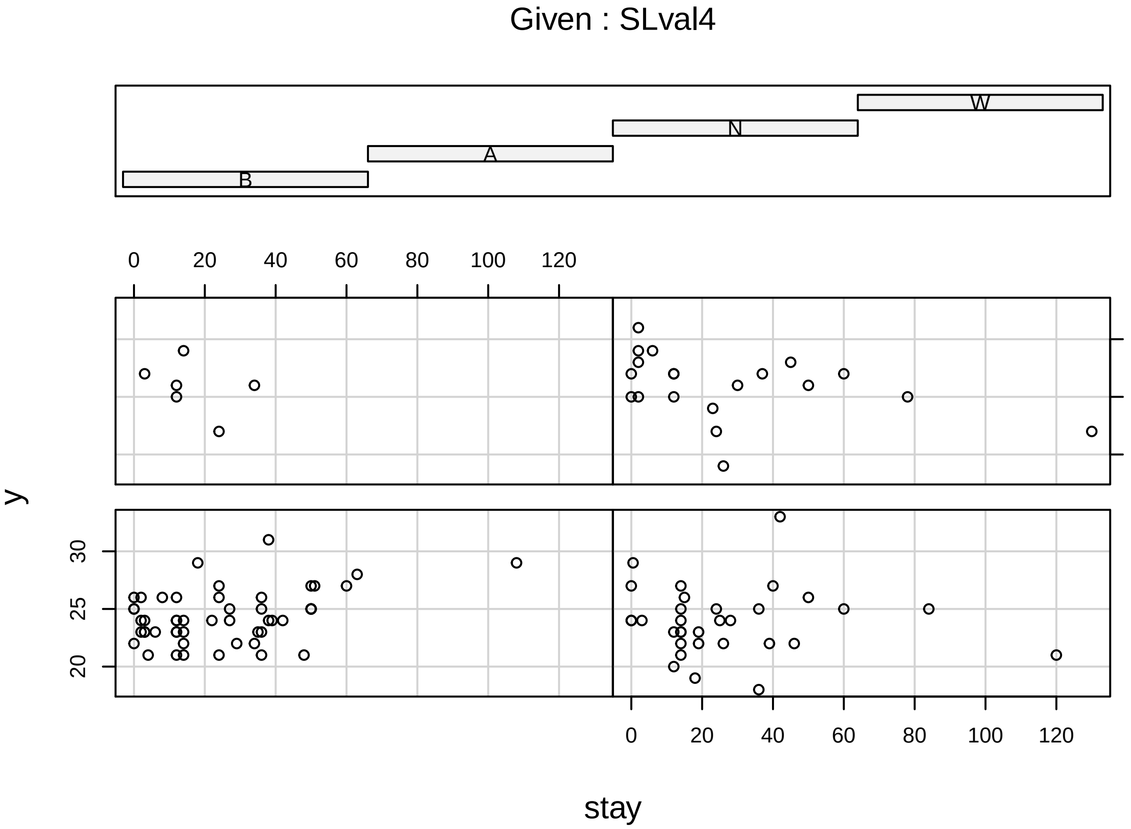




We notice there is a point of value = “A” outling and might have huge leverage. Lets remove it and refit the model

```
##
## Call:
## lm(formula = y ~ degree + stay + marital + SLval4 + sla_id +
##     sla_id:age + degree:sla_id + degree:age + stay:SLval4, data = df6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9847 -1.2720 -0.1824  1.3019  7.8797
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              27.15623    1.64047  16.554  < 2e-16 ***
## degreemaster            -16.35247    4.97687  -3.286 0.001441 **
## degreedoctoral          -12.88270    4.75223  -2.711 0.008007 **
## stay                      0.03283    0.01583   2.073 0.040926 *
## maritalmarried           -1.56237    0.63611  -2.456 0.015921 *
## SLval4A                   0.59180    0.83467   0.709 0.480105
## SLval4N                   4.49134    2.06242   2.178 0.031986 *
## SLval4W                   4.17165    0.98233   4.247 5.18e-05 ***
## sla_id                    3.33550    2.11273   1.579 0.117823
## sla_id:age             -0.25524    0.09683  -2.636 0.009848 **
## degreemaster:sla_id     2.38321    1.14126   2.088 0.039540 *
## degreedoctoral:sla_id   4.01350    1.42535   2.816 0.005953 **
## degreemaster:age        0.53113    0.17727   2.996 0.003513 **
## degreedoctoral:age      0.35012    0.14247   2.458 0.015863 *
## stay:SLval4A           -0.04271    0.02375  -1.798 0.075402 .
## stay:SLval4N           -0.14879    0.10400  -1.431 0.155920
## stay:SLval4W           -0.08677    0.02534  -3.425 0.000921 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.394 on 92 degrees of freedom
## Multiple R-squared:  0.3027, Adjusted R-squared:  0.1814
## F-statistic: 2.496 on 16 and 92 DF,  p-value: 0.003312
```



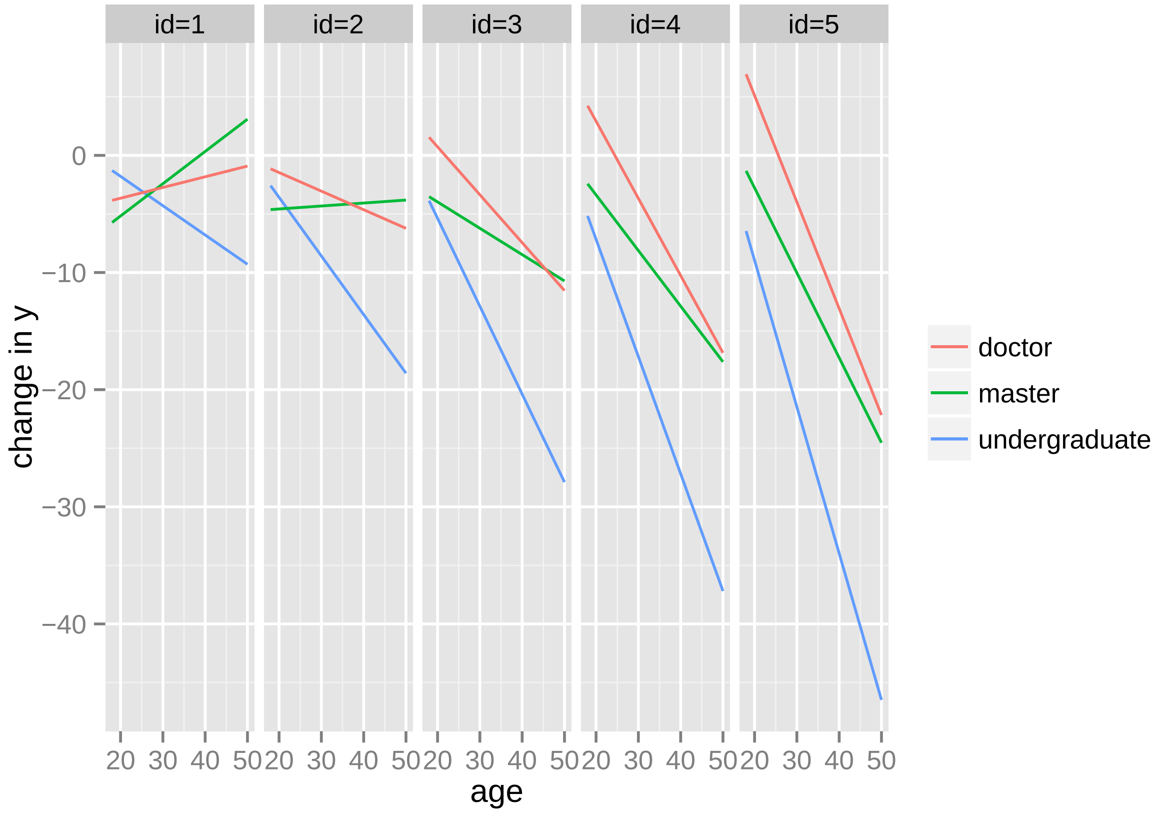


We found the point has negligible effects, so we will maintain it.

5. effect of sla_id

- 5.1 For undergraduate degree 3.210456 - 0.250148 * age
- 5.1 For master degree 3.210456 - 0.250148 * age + 2.392557
- 5.2 For Doctor degree 3.210456 - 0.250148 * age + 3.983903

Lets visulize changes of y due to degree, sla_id and age together with all other variables stay the same



6. effect of age

- 6.1 doctor degree
  - 6.1.1 sla_id = 1: 0.341559-0.250148 = 0.091411 > 0
  - 6.1.2 sla_id = 2-5: 0.341559 - 0.250148 * sla_id < 0
- 6.2 master degree
  - 6.2.1 sla_id=1-2: 0.525614 - 0.250148 * sla_id > 0
  - 6.2.2 sla_id=3-5: 0.525614 - 0.250148 * sla_id < 0

Result:

This model explains 31.39% of variation in y. It contains 17 variables (including dummy variables). two main interaction groups are revealed: 1, stay length and values. 2. degree, age and identifications. This model indicating

*Analyzing first interacting group shows

1. Interestingly, stay length has negative impact on y within groups of people with values of "A", "W" and "N". While stay length has positive impact on y within group of people with values of "B".
2. Western values shows highest positive impact when people stay in US between 5.48653 and 48.12083 months, which constitutes 0.6272727 of all our observations. And western value showed more positive impact on our y than asian value when the stay length is less than 81.24943 months, which consititutes 0.9545455 of our all obervations. This tells us western values has most positive impact, especially more positive than asian values, on our y in most of our samples. Try less complex model.

*The second interacing group involves degree, age and identifications, analyzing it shows:

1. in most cases, age is a negative factor on our y, except in groups of people with master degree and identification of 1 and 2 and group of people with doctoral degree with identification of 1. Which means if you are more asian identified and with advanced degree, the older you are, the more likely you are seeking for professional counseling.
2. More significantly, the more western you identify yourself, the age tend to have more negative impact on y. The higher degree you have, the more likely you are seeking for professional counseling. And interetingly, advanced degree tend to lower the rate of decressing of y due to age increasing.
3. Moreover, within group of people with undergraduate degree and at the same age, the more western they identified themself, the smaller y we get. And this difference increases along the age. in advanced degree, this trend changed. For example, in group of people with doctoral degree, during young age, younger than 30, the more western they identify themselves, the higher y they get. However, when age is bigger, they got less y instead.

*Also we observe married people has less y.

Reference 1. Jacob Cohen (1988). Statistical Power Analysis for the Behavioral Sciences (second ed.). Lawrence Erlbaum Associates. 2. Cohen, J (1992). “A power primer”. Psychological Bulletin 112 (1): 155-159. doi:10.1037/0033-2909.112.1.155. PMID 19565683.