Jiaqi report

Li Sun

October 12, 2015

Load packages

load necessary packages

Load data

Data comes from Dr. Jiaqi Li in csv file.

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:
3. 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)
4. do not fit ..... 5. fit very well

```
##
##  A  B  N  W
## 40 55  3 14
```

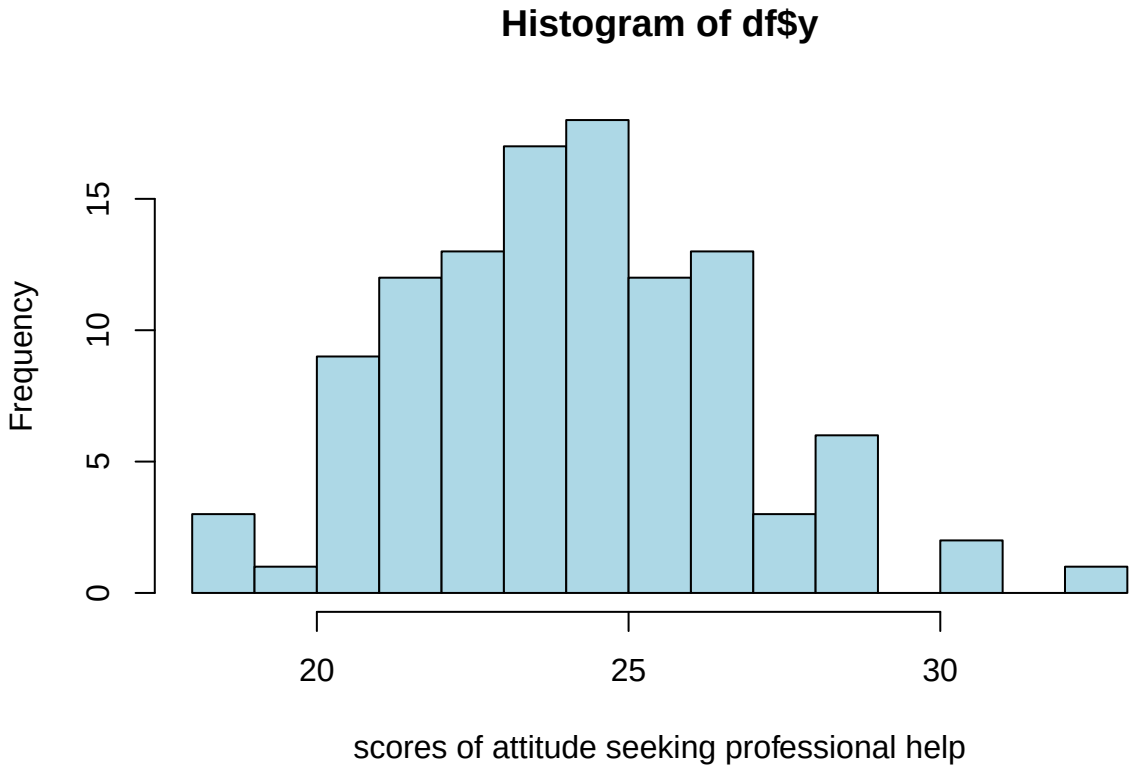
```
##       SLcom4
## SLval4  A  B  N  W
##      A 19 12  0  2
##      B 15 27  1  8
##      N  2  3  1  1
##      W  4 13  1  3
```

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

5. 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.
6. 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.
8. I consider myself as an Asian-American, although deep down, I view myself as an American first.
9. 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:
6. strongly disgree
7. disgree
8. agree
9. 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.

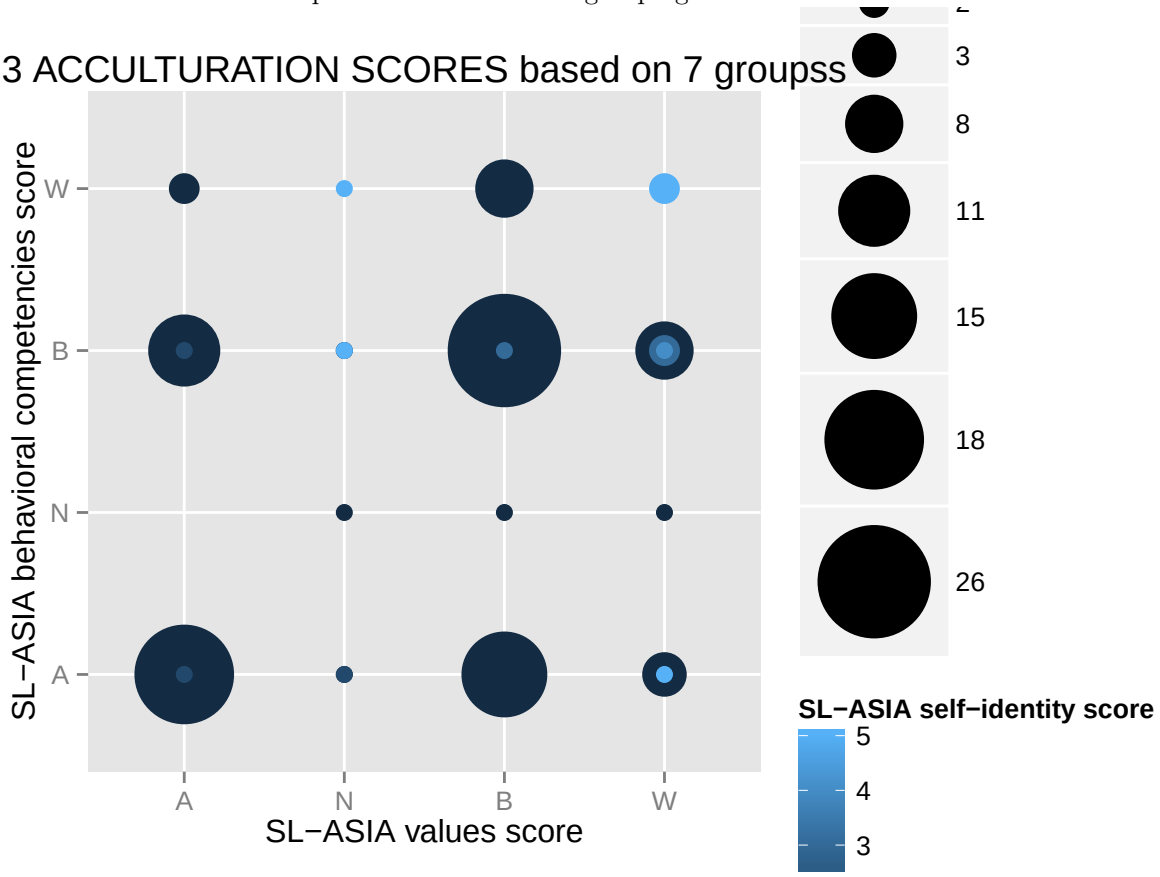



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   18.00   23.00   24.50   24.56   26.00   33.00      32
```

Exploratory data analysis

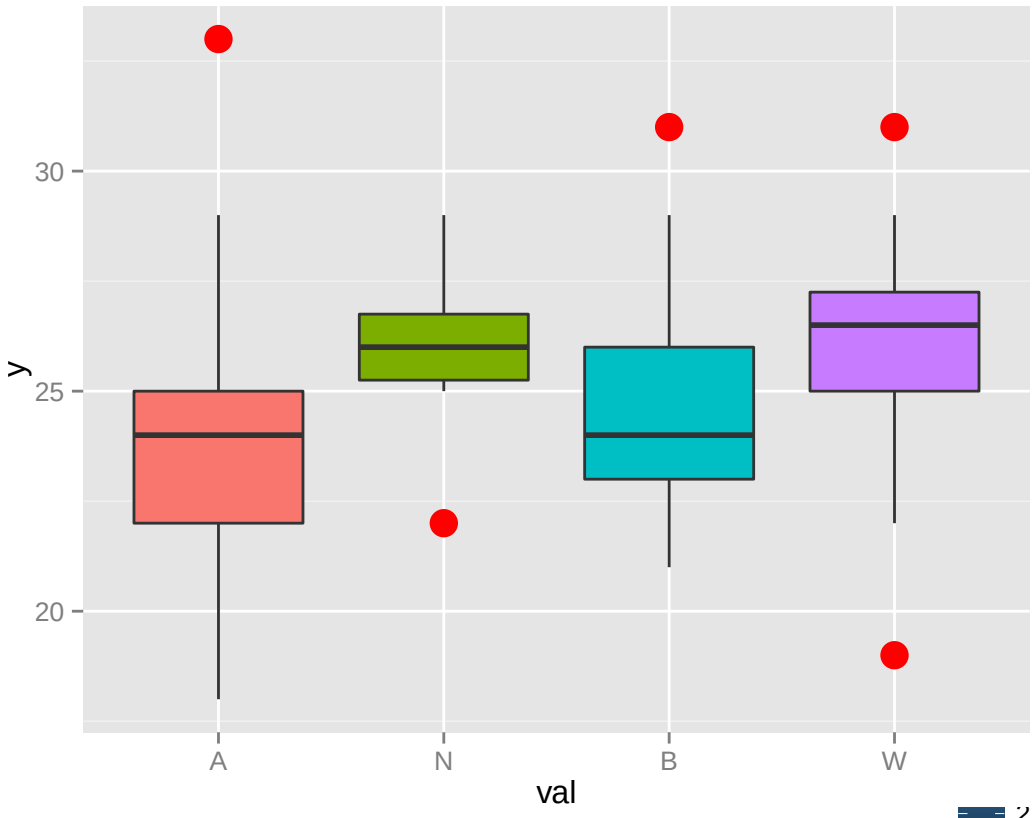
Visualize the relationship among the 3 scores from "Suinn-Lew Asian Self Identity Acculturation"

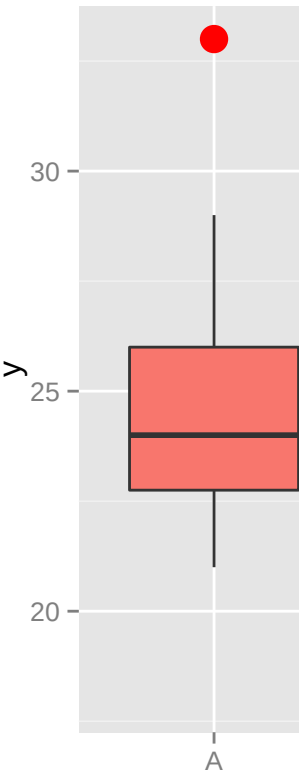
1. Visualize the relationship between different grouping methods of individuals acculturation

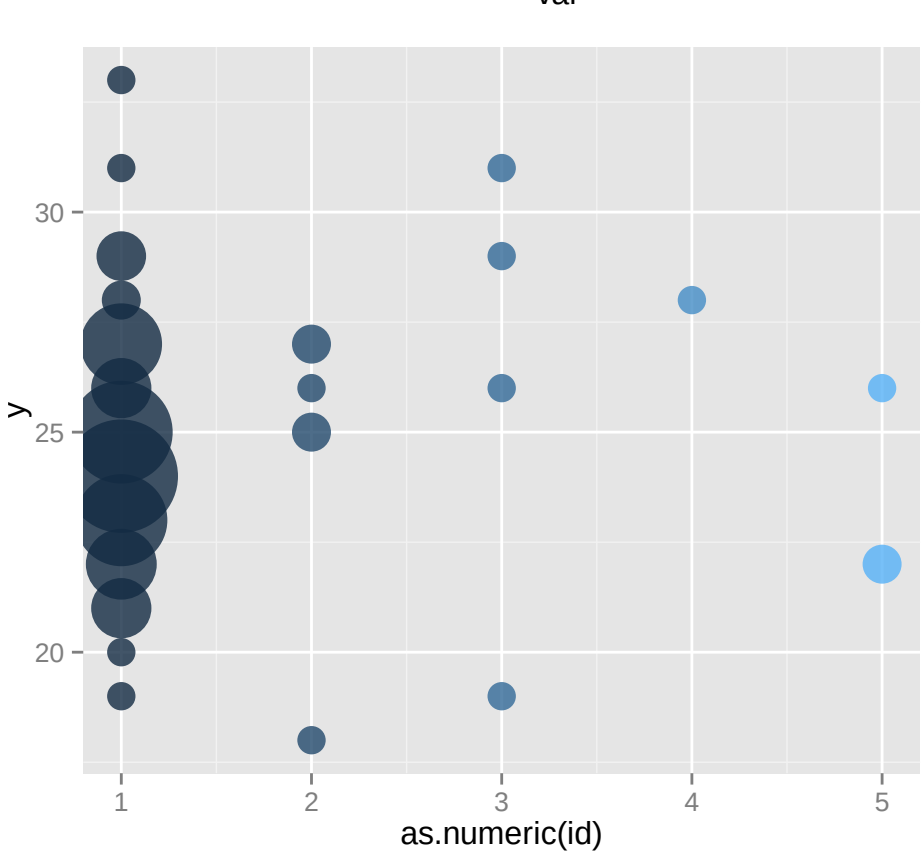


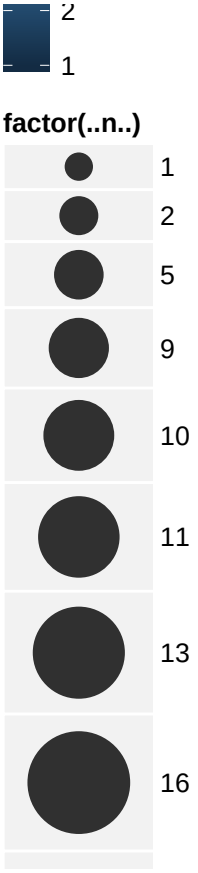
We found that the most asian students identify themselves as asian no matter how good they fit in western life.

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









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

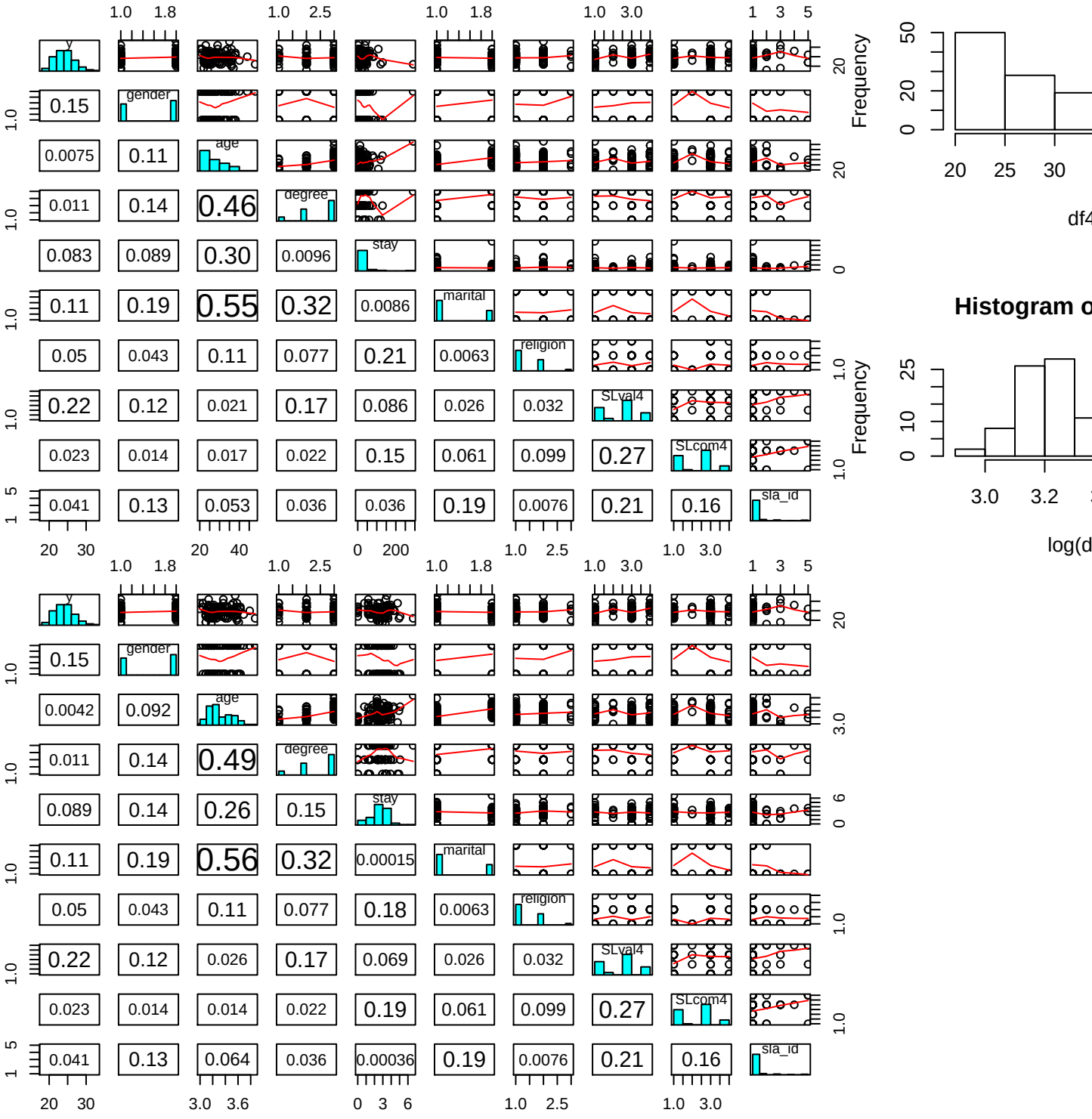
1. Missing values For convenience, we just remove all rows with NAs. We end up having a data with 110 observations and 10 variable.
2. Convert all variable type to ones ready for analysis

```
##         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



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.

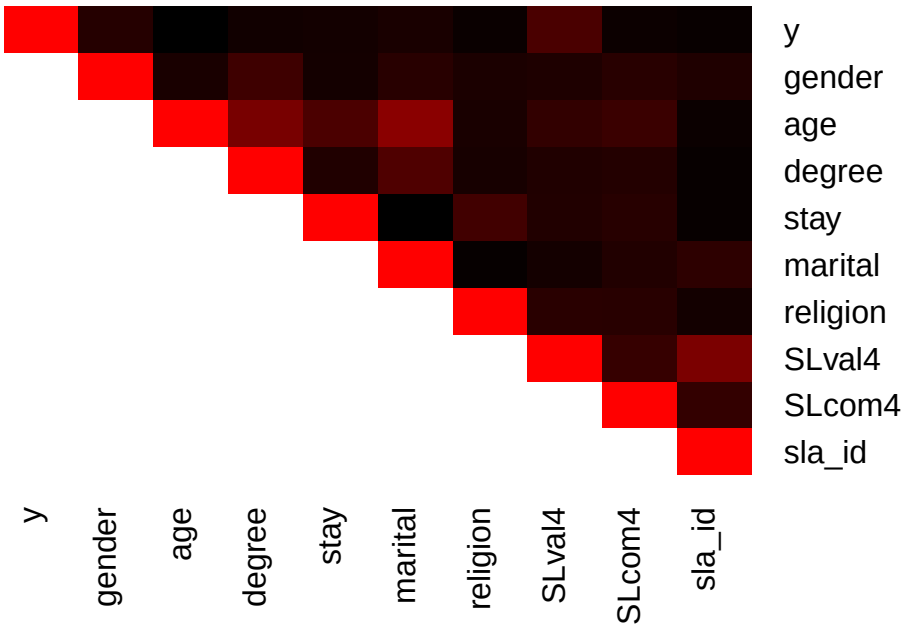
for numeric vs numeric: Pearson's correlation is used, absolute value of this r is categorized as followed, Effect size r Small 0.10 Medium 0.30 Large 0.50[1][2]

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

```
## Warning in chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(...): Chi-squared approximation may be incorrect
```

Visualizing the association



Fit regression

1. Model without interaction
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.

```
##               full_without_int     gender        age    degree      stay
## adj_r_squared       0.04526732 0.03235948 0.03727448 0.1752944 0.1116613
##                     marital   religion    SLval4    SLcom4    sla_id
## adj_r_squared -0.02280438 0.05953353 0.1842698 0.0479008 0.1754162
```

We found, when we include interactions of the following 4 variables to all others, we got dramatically boosted the adjusted r squared. degree stay SLval4 sla_id

The following model is from data without SLcom4 First we try all combinations of the 4 variables interacting with all others from the 4 listed above which gave us best improved adjusted r squared. And all full models are backward selected.

All models we have fmstep: full model without interactions all following are with interactions fmintstep: with all possible interaction fmintstep2: with four variables interacting with all others: degree, stay, SLval4, sla_id fmintstep3: with three variables interacting with all others: degree, stay, sla_id fmintstep4: with three variables interacting with all others: degree, stay, sla_id, and without variable SLcom4 mstaystep: with stay interacting with all others mdgstep: with degree interacting with all others midstep: with sla_id interacting with all others mvalstep: with SLval4 interacting with all others m_staystep: without stay interacting with all others m_dgstep: without degree interacting with all others m_idstep: without sla_id interacting with all others m_valstep: without SLval4 interacting with all others m1step: with degree and stay interacting with all others m2step: with degree and sla_id interacting with all others m3step: with degree and SLval4 interacting with all others m4step: with stay and SLval4 interacting with all others m5step: with stay and sla_id interacting with all others m6step: with SLval4 and sla_id interacting with all others sl1step: with sla_id interacting with degree, s tay interacting with SLval4 degree interacting with age SLval4 interacting with marital stay interacting with sla_id age interacting with stay age interacting with sla_id sl2step: forward select from sl1step sl3: remove alias variable from sl2step sl4: remove triple interacting variable

Compare all models

```
##            df      AIC  bic$BIC   adj.rsqr
## fmstep      7 525.1810 544.0844 0.08469829
## fmintstep  76 421.4242 626.6608 0.69797560
## fmint2step 59 485.2859 644.6142 0.50514579
## fmint3step 50 496.3409 631.3650 0.45063110
## fmint4step 38 502.3099 604.9281 0.39717341
## mstaystep  16 521.6788 564.8865 0.17589868
## mdgstep    31 528.3673 612.0822 0.20826986
## midstep    23 526.5377 588.6487 0.18128841
## mvalstep   11 521.1985 550.9038 0.14633165
## m_staystep 40 500.7715 608.7907 0.41062667
## m_dgstep   47 519.2820 646.2046 0.31879631
## m_idstep   43 498.8679 614.9886 0.42728872
## m_valstep  38 502.3099 604.9281 0.39717341
```

```
## m1step     28 519.6000 595.2134 0.25584509
## m2step     36 516.1957 613.4130 0.30964939
## m3step     37 503.5485 603.4663 0.38755088
## m4step     16 521.6788 564.8865 0.17589868
## m5step     26 514.4556 584.6681 0.28087477
## m6step     34 516.2884 608.1047 0.30209164
```

So far we have tried different variables interacting with all others, which might bring many irrelevant interactions to compromise our model in BIC values. So what about smaller number of interactions? According to all models above, I picked following several interactions sla_id and degree stay and SLval4 degree and age SLval4 and marital stay and sla_id age and stay age and sla_id All of them are maintained in many of above models after backward selection.

This new model is better than ones with similar number of variables. Try forward selection to see if we can get any more significant interactions

sl2step looks the best so far, we will adopt this model and further improve it. there is a question with this model that the design matrix is not full rank. The variable degreeDoc interacting marital status is linear combination of other variables. So lets remove one variable to make it full rank.

```
## Model :
## y ~ gender + age + degree + stay + marital + SLval4 + sla_id +
##     degree:sla_id + stay:SLval4 + age:degree + stay:sla_id +
##     age:sla_id + gender:SLval4 + age:marital + degree:marital +
##     marital:SLval4 + degree:stay + age:degree:sla_id
##
## Complete :
##                                (Intercept) genderwoman age degreemaster
## degreedoctoral:maritalmarried  0           0           0   0
##                                degreedoctoral stay maritalmarried SLval4N
## degreedoctoral:maritalmarried  0              0    1              0
##                                SLval4B SLval4W sla_id degreemaster:sla_id
## degreedoctoral:maritalmarried  0       0       0      0
##                                degreedoctoral:sla_id stay:SLval4N
## degreedoctoral:maritalmarried  0                     0
##                                stay:SLval4B stay:SLval4W age:degreemaster
## degreedoctoral:maritalmarried  0            0            0
##                                age:degreedoctoral stay:sla_id age:sla_id
## degreedoctoral:maritalmarried  0                  0           0
##                                genderwoman:SLval4N genderwoman:SLval4B
## degreedoctoral:maritalmarried  0                   0
##                                genderwoman:SLval4W age:maritalmarried
## degreedoctoral:maritalmarried  0                   0
##                                degreemaster:maritalmarried
## degreedoctoral:maritalmarried -1
##                                maritalmarried:SLval4N
## degreedoctoral:maritalmarried  0
##                                maritalmarried:SLval4B
## degreedoctoral:maritalmarried  0
##                                maritalmarried:SLval4W degreemaster:stay
## degreedoctoral:maritalmarried  0                     0
##                                degreedoctoral:stay age:degreemaster:sla_id
## degreedoctoral:maritalmarried  0                   0
##                                age:degreedoctoral:sla_id
## degreedoctoral:maritalmarried  0
```

so basically the original variable marital is the same as degreeDocmarital + degreeMasmarital, lets remove marital variable

remove 3 variable interaction

Compare all models, model sl3 stands out

	df	AIC	bic$BIC	adj.rsqr
fmstep	7	525.1810	544.0844	0.0846983
fmintstep	76	421.4242	626.6608	0.6979756
fmint2step	59	485.2859	644.6142	0.5051458
fmint3step	50	496.3409	631.3650	0.4506311
fmint4step	38	502.3099	604.9281	0.3971734
mstaystep	16	521.6788	564.8865	0.1758987
mdgstep	31	528.3673	612.0822	0.2082699
midstep	23	526.5377	588.6487	0.1812884
mvalstep	11	521.1985	550.9038	0.1463317
m_staystep	40	500.7715	608.7907	0.4106267
m_dgstep	47	519.2820	646.2046	0.3187963
m_idstep	43	498.8679	614.9886	0.4272887
m_valstep	38	502.3099	604.9281	0.3971734
m1step	28	519.6000	595.2134	0.2558451
m2step	36	516.1957	613.4130	0.3096494
m3step	37	503.5485	603.4663	0.3875509
m4step	16	521.6788	564.8865	0.1758987
m5step	26	514.4556	584.6681	0.2808748
m6step	34	516.2884	608.1047	0.3020916
sl1step	21	520.3380	577.0481	0.2153308
sl2step	33	489.3449	578.4608	0.4508203
sl3	33	489.3449	578.4608	0.4508203
sl4	31	507.5143	591.2292	0.3449916
The best mode	l is	based on th	e criterion	above is sl3.

Results

```
## % latex table generated in R 3.2.3 by xtable 1.8-0 package
## % Tue Jan 12 21:23:25 2016
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrrr}
##   \hline
##  & Estimate & Std. Error & t value & Pr($>$$|$t$|$) \\
##   \hline
## (Intercept) & -6.1776 & 20.3671 & -0.30 & 0.7625 \\
##   age & 1.6816 & 0.9655 & 1.74 & 0.0855 \\
##   degreemaster & -35.5314 & 26.4493 & -1.34 & 0.1830 \\
##   degreedoctoral & 17.7827 & 22.0046 & 0.81 & 0.4215 \\
##   stay & -0.0016 & 0.0341 & -0.05 & 0.9618 \\
##   SLval4N & -10.4843 & 8.2838 & -1.27 & 0.2094 \\
##   SLval4B & -0.6235 & 0.9216 & -0.68 & 0.5007 \\
##   SLval4W & 5.2821 & 1.9589 & 2.70 & 0.0086 \\
##   sla\_id & 42.4749 & 18.9847 & 2.24 & 0.0281 \\
```

```
##   degreemaster:sla\_id & 2.8897 & 24.6748 & 0.12 & 0.9071 \\
##   degreedoctoral:sla\_id & -36.7527 & 20.3543 & -1.81 & 0.0748 \\
##   stay:SLval4N & 0.3131 & 0.2254 & 1.39 & 0.1687 \\
##   stay:SLval4B & 0.0416 & 0.0200 & 2.08 & 0.0407 \\
##   stay:SLval4W & -0.0711 & 0.0320 & -2.23 & 0.0289 \\
##   age:degreemaster & 1.0724 & 1.1956 & 0.90 & 0.3725 \\
##   age:degreedoctoral & -1.2507 & 0.9896 & -1.26 & 0.2101 \\
##   stay:sla\_id & -0.0068 & 0.0293 & -0.23 & 0.8177 \\
##   age:sla\_id & -2.1531 & 0.9169 & -2.35 & 0.0214 \\
##   SLval4A:genderwoman & 2.0006 & 0.8739 & 2.29 & 0.0248 \\
##   SLval4N:genderwoman & 1.3230 & 2.6944 & 0.49 & 0.6248 \\
##   SLval4B:genderwoman & 0.8902 & 0.6432 & 1.38 & 0.1703 \\
##   SLval4W:genderwoman & -2.5128 & 1.3541 & -1.86 & 0.0673 \\
##   age:maritalmarried & -0.4383 & 0.1200 & -3.65 & 0.0005 \\
##   degreemaster:maritalmarried & 7.1636 & 3.3824 & 2.12 & 0.0374 \\
##   degreedoctoral:maritalmarried & 10.8242 & 3.6482 & 2.97 & 0.0040 \\
##   SLval4N:maritalmarried & 11.9039 & 5.7015 & 2.09 & 0.0401 \\
##   SLval4B:maritalmarried & 0.7678 & 1.0218 & 0.75 & 0.4547 \\
##   SLval4W:maritalmarried & 3.3279 & 1.3989 & 2.38 & 0.0198 \\
##   degreemaster:stay & -0.0407 & 0.0257 & -1.58 & 0.1179 \\
##   degreedoctoral:stay & 0.0098 & 0.0209 & 0.47 & 0.6401 \\
##   age:degreemaster:sla\_id & 0.2414 & 1.1274 & 0.21 & 0.8310 \\
##   age:degreedoctoral:sla\_id & 1.9686 & 0.9326 & 2.11 & 0.0380 \\
##    \hline
## \end{tabular}
## \end{table}
```

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.