Final Project Report (Speed Dating Analysis)

Sangseok Lee

12/9/2019

SUMMARY

1. In a speed dating during four minutes, when male and female gives one point on their partner's attractiveness, sharing values, humor or probablity to decide to meet themselves, it is more likely to increase their possibility to decide to meet their partners. This result, in turn, also seems to affect their matching after their decision since their partner decides to meet themselves based on the same condition.
2. In the same way, when male and female gives one point on their partner's sincerity or ambition, it is likely to decrease their possibiltiy to decide to meet their partners.
3. Giving one point on their partner's intelligence seems to affect a positive effect on female's decision to meet and a negative effect on male's decision to meet.
4. It appears that meeting with the same race does not affect much on people's decision, nor their matching.

Intro

In today's busy world, finding and dating a romantic partner seems more time consuming than ever. As a result, many people have turned to speed dating as a solution that allows one to meet a large number of potential partners during a short amount of time. In this report, I want to explore what speed dating takes to become successful in getting approvals from a potential partner. In addition, since matching is more important than just decision, I am going to explore what factors really affect their matching.

Data background

This data came from an experiment done by Columbia University to figure out "Race preference in dating". The experiment is based on meetings through speed dating eventes where participants engage in four-minute conversations to determine whether or not they are interested in one another.

Purpose of the research

1. Decision model by male and female: I would like to know which variables are the main factors to make people choose.
2. Matching model by male and female : I would like to know which variables are the main factors to make people choose each other.

In order to answer the first question, I use the dataset with 8378 rows and 15 columns and my response variable would be 'dec'variable. In addition, to answer the second question, I only select the people who decide their partner, which brings the dataset with 3375 rows and 23 columns. My response variable for the second model is 'match' variable.

1. The model for the first question

My response variable is 'dec' column and predictors are 11 variables: 'gender','attr','sinc','intel','fun','amb','shar','samerace', 'age_o','prob','like'.

1) Data Cleaning : missing values and categoricla variables

Frist of all, there are 10 NAs in pid(partner ID) column from wave 5, and they can be inferred as 118 because people in wave 5 evaluated everyone except 118 as iid. Second, I dropped 414 rows including missing values that we cannot estimate from 'age_0', 'prob', 'like' columns. 5 no evaluation rows on all six variables(attr,sinc,intel,fun,amb,shar) is also removed because they do not give us any information for the decision. Third, we changed dec(decision), gender, samerace columns as facor variables. Still, we have more than 1,000 missing values(about 12% of the dataset, Appendix 1).

2) Data Cleaning : imputed data

Since some variables such as sinc, intel, amb and shar are relatively hard to estimate with imputed data compared with other variables such as attr and fun, I used two ways for imputation. 1. [Model 1-1] Imputation only for attr, fun (removal for NAs in sinc, intel, amb and shar) : Appendix 2 2. [Model 1-2] Imputation for every missing values in all six variables I am going to compare two models and decide which model is better. From the first way of imputation, I used 'norm' because imputed values keeps observed values rathers than 'pmm' way. In order to avoid multi-collinearity, I made mean-centering for attr, sinc, intel, fun, amb, shar, prob, like. Now I divded this dataset by two : male (3,505 observations) and female(3,403 observations).

3) EDA (Plots in Appendix 3)

When you look at the relationship between decision and other variables, it seems that most variables have positive effect on the response variable in both male and female. However, for 'amb' variable in female, it does not seem to make difference in decision. In addition, for 'age_o' variable in male, male tend to say yes to younger female. When I checked the interaction term between each predictor, 'shared value' and 'inteligence' seems to have an interaction term. Even if you have high score of intelligence, it does not seem to guarantee that you have high score of shared values. However, if you have high score of shared values, it seems to guarantee that you always have high score of shared values. 'shared value' and 'ambition' seems to have an interaction term, which is a relatively weaker relationship than the relationship between 'shared value' and 'intelligence'. In addition, people's decision regarding 'same_race', it does not seem that people prefer their partner as the same race.

4) Model (1-1)

Since people are evaluated at multiple times in their wave, which is equal to the meaning of a group, I assume that there should be a varying intercept according to the wave. Also, I assumed that during four minutes, 'attractive' variable would be the highest indicator that is easy to observe compared to other variables, so I put 'attractive' variable as a varying slope. With these in mind, I analyze this data with a hierarchical logistic regression. When I checked a model with only a random intercept and the other model with a random intercept("wave") and a random slope("attr"), it turned out that the hierarchical model fails to converge, so I choose the model with a random intercept only. For our model, I use "dec" as the response variable, with attr, sinc, fun, amb, intel, shar, samerace, prob, like as the main effect. To build our model, I started with a baseline model with all the main effects. I then manually added an interaction to test whether the new model with the single interaction would converge. Out of the 36 possible interactions for the 9 main effects, I found that interaction term between "intel" and "shar" improved the AIC and BIC of the baseline model as our EDA guided us.

We have two models including a male and a female model.

a) Male model : $wave_i|x_i \sim Bernoulli(\pi_i)$;, i= 1 , . . . , n; j = 1, . . . , 21;

$\log(\frac{\pi_i}{1-\pi_i}) = \beta_0+\gamma_{0j[i]}^{wave}+\beta_{attr}attr_i+\beta_{sinc}sinc_i+\beta_{fun}fun_i+\beta_{intel}intel_i+\beta_{share}share_i+\beta_{samerace}samerace_i+\beta_{prob}prob_i+\beta_{like}like_i$ $\gamma_{0j} \sim N(0, \sigma^2_{wave})$

b) Female Model : $wave_i|x_i \sim Bernoulli(\pi_i)$;, i= 1 , . . . , n; j = 1, . . . , 21;

$\log(\frac{\pi_i}{1-\pi_i}) = \beta_0+\gamma_{0j[i]}^{wave}+\beta_{attr}attr_i+\beta_{sinc}sinc_i+\beta_{fun}fun_i+\beta_{intel}intel_i+\beta_{share}share_i+\beta_{prob}prob_i+\beta_{like}like_i$

	Male			Female		
Coefficient	*Odds Ratios*	*Conf.Int(95%)*	*P-Value*	*Odds Ratios*	*Conf.Int(95%)*	*P-Value*
(Intercept)	0.71	0.56 – 0.90	**0.005**	0.51	0.39 – 0.66	**<0.001**
attr_c	1.78	1.65 – 1.91	**<0.001**	1.40	1.31 – 1.49	**<0.001**
sinc_c	0.76	0.70 – 0.83	**<0.001**	0.84	0.78 – 0.91	**<0.001**
fun_c	1.16	1.08 – 1.25	**<0.001**	1.17	1.09 – 1.26	**<0.001**
amb_c	0.85	0.79 – 0.92	**<0.001**	0.85	0.79 – 0.91	**<0.001**
intel_c	0.91	0.83 – 1.00	**0.047**	1.13	1.03 – 1.24	**0.012**
shar_c	1.08	1.02 – 1.15	**0.012**	1.20	1.13 – 1.28	**<0.001**
samerace: samerace1	0.82	0.69 – 0.99	**0.037**			
prob_c	1.23	1.17 – 1.30	**<0.001**	1.14	1.09 – 1.20	**<0.001**
like_c	1.92	1.74 – 2.12	**<0.001**	1.60	1.47 – 1.75	**<0.001**
intel_c:shar_c	0.93	0.90 – 0.96	**<0.001**	0.97	0.94 – 1.00	0.093
Random Effects						
σ^2	3.29			3.29		
τ_{00}	0.20_{wave}			0.32_{wave}		
ICC	0.06			0.09		
N	21_{wave}			21_{wave}		
Observations	3505			3403		
Marginal R^2 / Conditional R^2	0.580 / 0.604			0.516 / 0.559		

Figure 1: Hierarchical Logistic regression for decision

$\gamma_{0j} \sim N(0, \sigma^2_{wave})$

5) Model Results

a) Male model : The results indicate that attr,fun,amb,prob, like, and the interactions intel:shar were highly significant less than 0.001% significance level. In addition, intel, shar, samerace were also statistically significant at the 0.05% significance level. (varying slope : Appendix 4)

b) Female model : The results indicate that attr, fun, amb, prob, like, shar were highly significant less than 0.001% significance level. In addition, 'intel' was also statistically significant at the 0.05% significance level. Lastly, the interactions intel:shar was less significant with less than 0.01% significance level.(varying slope : Appendix 5)

For example, if the male who met a female without same race during a blind-dating gives one point for 'attracitve' score, it increases 78% in the odds of the male's decision. If a female gives one point for 'attractive' score, it increases 40% in the odds of the female's decision. In addition, the confusion Matrix and ROC curve is in the appendix 6(Male model) and the appendix 7(Female model). The accuracy of the male model is 0.8 with 0.437 threshold and the female model is 0.75 with 0.299 threshold.

6) Model Assessment

I checked the binned residuals plots and there are no patterns or no severe outliers. Appendix 8 illustrates it for the male model and Appendix 9 illustrates it for the female model. When I campared ther 1-1 model with the 1-2 model as illustrated in Appendix 10 and 11, it turns out that the 1-1 model has better AIC and BIC, eventually.

2. The model for the second question

In order to answer the second question, I only select the people who decide their partner, which brings the dataset with 3518 rows and 23 columns. My response variable for this model is 'match' variable.

1) Data Cleaning : missing values and categoricla variables

Frist of all, there are 10 NAs in pid(partner ID) column from wave 5, and they can be inferred as 118 because people in wave 5 evaluated everyone except 118 as iid. Second, I dropped 615 rows including missing values that we cannot estimate from 'age_0', 'prob', 'like', 'attr3_1','shar_o','amb_o','intel_o','sinc_o','prob_o','like_o' columns. Third, we changed match, gender, samerace columns as facor variables. We have only 6 missing values to impute.

2) Data Cleaning : imputed data

I used 'ppm' because imputed values keeps observed values rathers than 'norm' way. In order to avoid multi-collinearity, I made mean-centering for attr3_1, sinc3_1, intel3_1, fun3_1, amb3_1, attr_o, sinc_o, intel_o, fun_o, amb_o,shar_o, prob, like,prob_o, like_o. Now I divded this dataset by two : male (1,289 observations) and female(1,614 observations).

3) EDA (plots in Appendix 13)

When you look at the relationship between match and other variables, it seems that most of self evaluation for themselves does not seem to affect the partner's decision except 'fun' in only male's matching, which means that if a male believes himself as a humours guy, he has high possibility to be matched with his partner. Also, even if you believe that your partner says yes to meet you, it does not seem to affect the partner's decision. In addition, scores of the partners selected by people seems to affect positively on the matching in both male and female. When I checked the interaction term between what people belive(attr3_1, sinc3_1,fun3_1,amb3_1,intel3_1) and the score of their partner(attr_o,sinc_o,fun_o,amb_o,intel_o), it seems that there is no interaction between each pair. In addition, people's matching regarding 'same_race', it does not seem that people do not match according to the same race.

4) Model

In cotrast to the first model using a hierarchical logistic regression, we consider individuals rather than groups.Thus, I use a logistic regression with "match" as the response variable, with attr3_1, sinc3_1, fun3_1, amb3_1, intel3_1, attr_o, sinc_o, fun_o, amb_o, intel_o,shar_o, samerace, prob, like, prob_o, like_o as the main effect. To build our model, I used stepwise selection using AIC.

We have two models including a male and a female model. a) Male model and female model :

$\log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 like(partner)_i + \beta_2 attr(partner)_i + \beta_3 shar(partner)_i + \beta_4 prob_i + \beta_5 amb(partner)_i + \beta_6 fun(partner)_i + \beta_7 prob(partner)_i + \beta_8 sinc(partner)_i + \beta_9 intel(partner)_i + \beta_{10} fun(self)_i + \beta_{11} like_i$

5) Model Results

a) Male model : The results indicate that like_o, attr_o, shar_o, prob, amb_o, fun_o and prob_o were highly significant less than 0.001% significance level. In addition, sinc_o was also statistically significant less than 0.05%. Lastly, fun3_1(self) was less significant with less than 0.1% significance level. For example, if the male who decided to meet a female during a blind-dating obtained one point for 'attracitve' score from the female, it increases 40% in the odds of being matched with the female. In the same way, like_o, shar_o, prob,fun_o, prob_o, intel_o, fun3_1(self) have positive effect on the matching model. On the contrary, amb_o, sinc_o and like have negative effect on the matching model.(if the male who decided to meet a female obtained one point for 'sincere' scroe from the female, it decreases 11% in the odds of being matched with the female)

	Male			Female		
Coefficient	*Odds Ratios*	*Conf.Int(95%)*	*P-Value*	*Odds Ratios*	*Conf.Int(95%)*	*P-Value*
(Intercept)	0.42	0.36 – 0.48	**<0.001**	0.53	0.45 – 0.63	**<0.001**
like_o_c	1.47	1.29 – 1.68	**<0.001**	1.61	1.38 – 1.88	**<0.001**
attr_o_c	1.40	1.28 – 1.54	**<0.001**	1.83	1.63 – 2.06	**<0.001**
shar_o_c	1.20	1.10 – 1.31	**<0.001**	1.08	0.98 – 1.19	0.109
prob_c	1.15	1.07 – 1.25	**<0.001**	1.07	0.98 – 1.17	0.114
amb_o_c	0.78	0.70 – 0.87	**<0.001**	0.82	0.72 – 0.92	**0.001**
fun_o_c	1.22	1.10 – 1.36	**<0.001**	1.22	1.09 – 1.38	**0.001**
prob_o_c	1.15	1.07 – 1.23	**<0.001**	1.14	1.05 – 1.24	**0.002**
sinc_o_c	0.89	0.79 – 0.99	**0.036**	0.75	0.66 – 0.86	**<0.001**
intel_o_c	1.13	0.98 – 1.30	0.102	1.06	0.91 – 1.24	0.444
fun3_1_c	1.07	0.99 – 1.17	0.089	0.92	0.83 – 1.02	0.111
like_c	0.92	0.82 – 1.03	0.134	0.98	0.87 – 1.10	0.701
Observations	1614			1289		
R^2 Tjur	0.360			0.415		

Figure 2: Logistic regression for matching

b) Female model : The results indicate that like_o,attr_o, fun_o, sinc_o and amb_o were highly significant equal or less than 0.001% significance level. In addition, prob_o was also statistically significant less than 0.01% significance level. For example, if a female who decided to meet a male during a blind-dating obtained one point for 'attractive' score, it increases 83% in the odds of being matched with the male. In the same way, like_o, shar_o, prob, fun_o, prob_o, intel_o have positive effect on the matching model. On the contrary, amb_o, sinc_o, fun(self), like have negative effect on the matching model. (If the female who decided to meet a male obtained one point for 'sincere' socre from the male, it decreases 24% in the odds of being matched with the male)

In addition, the confusion Matrix and ROC curve is in the appendix 14(Male model) and the appendix 15(Female model). The accuracy of the male model is 0.76 with 0.325 threshold and the female model is 0.80 with 0.514 threshold.

6) Model Assessment

I checked the binned residuals plots and there are no patterns or no severe outliers. Appendix 16 illustrates it for the male model and Appendix 17 illustrates it for the female model.

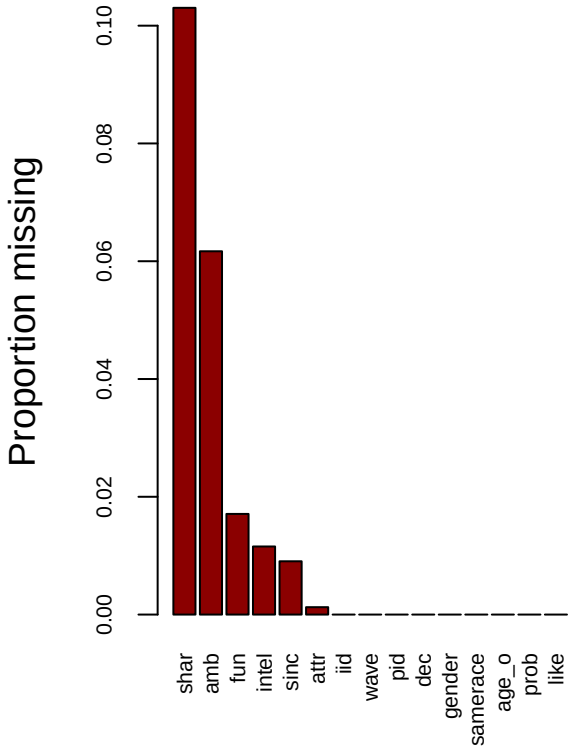
Limitation

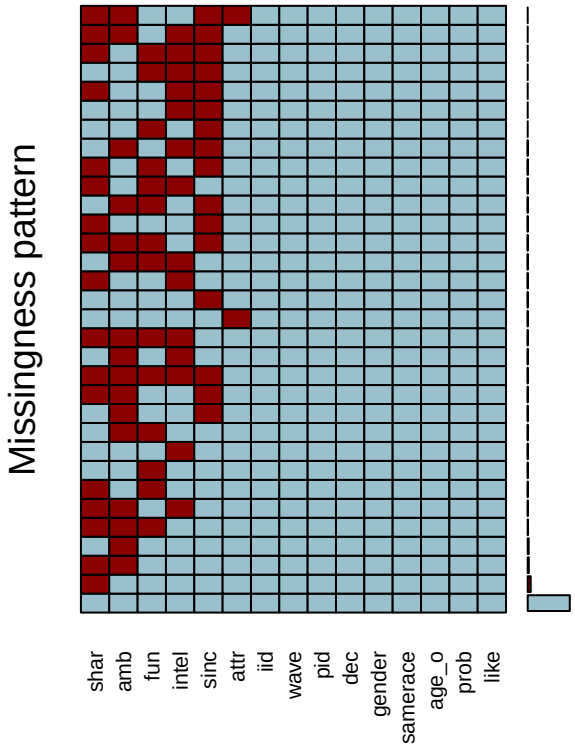
1. It is such a limited time to determine people's characteristic, specially some values that is hard to estimate such as sincerity, ambition or intelligence.
2. There are about 15% of missing values that are hard to impute because they are mostly personal feelings, so that we could not use them.

Appendix 1 (Missing data)

```
## Warning in plot.aggr(res, ...): not enough vertical space to display
```

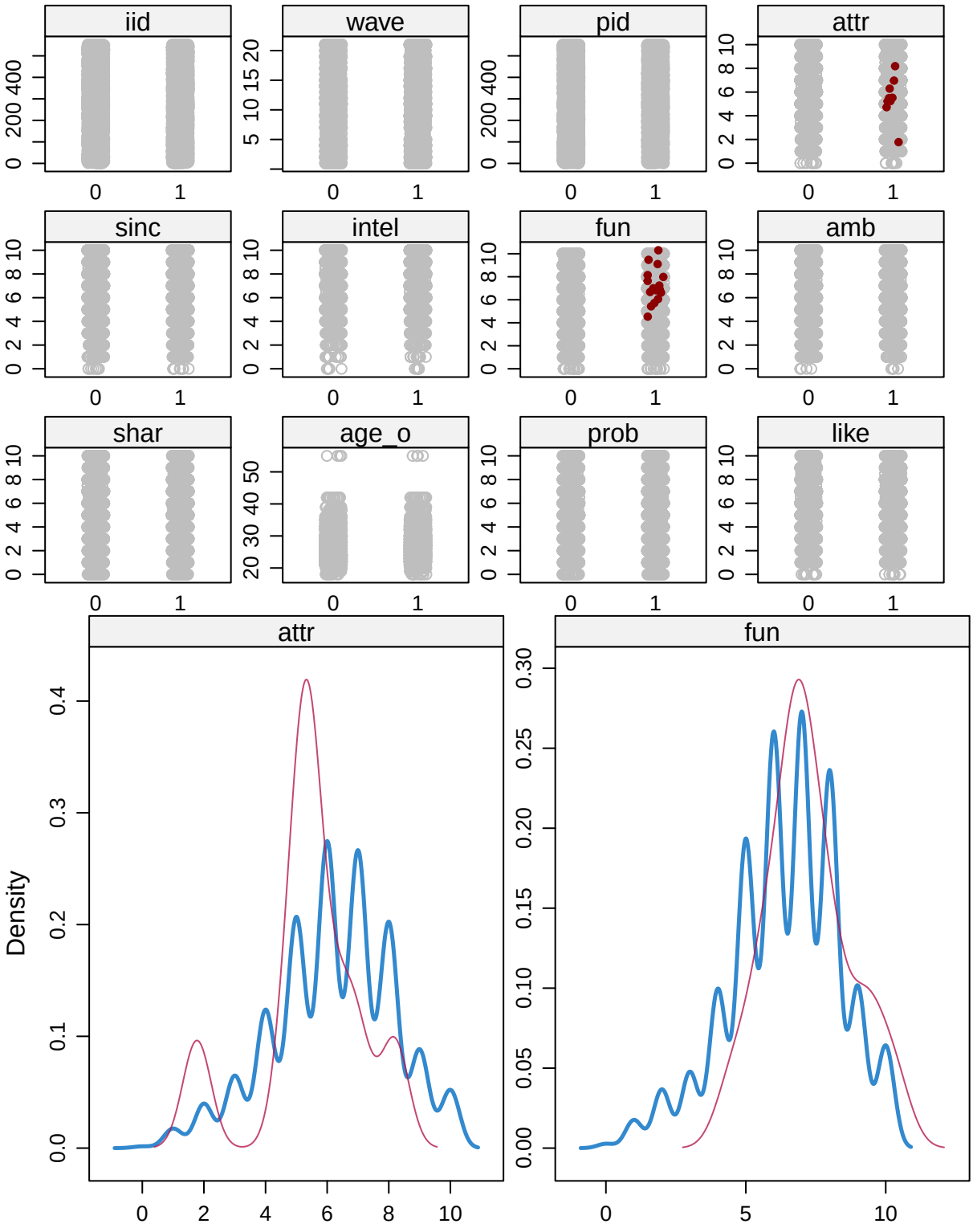
```
## frequencies (too many combinations)
```





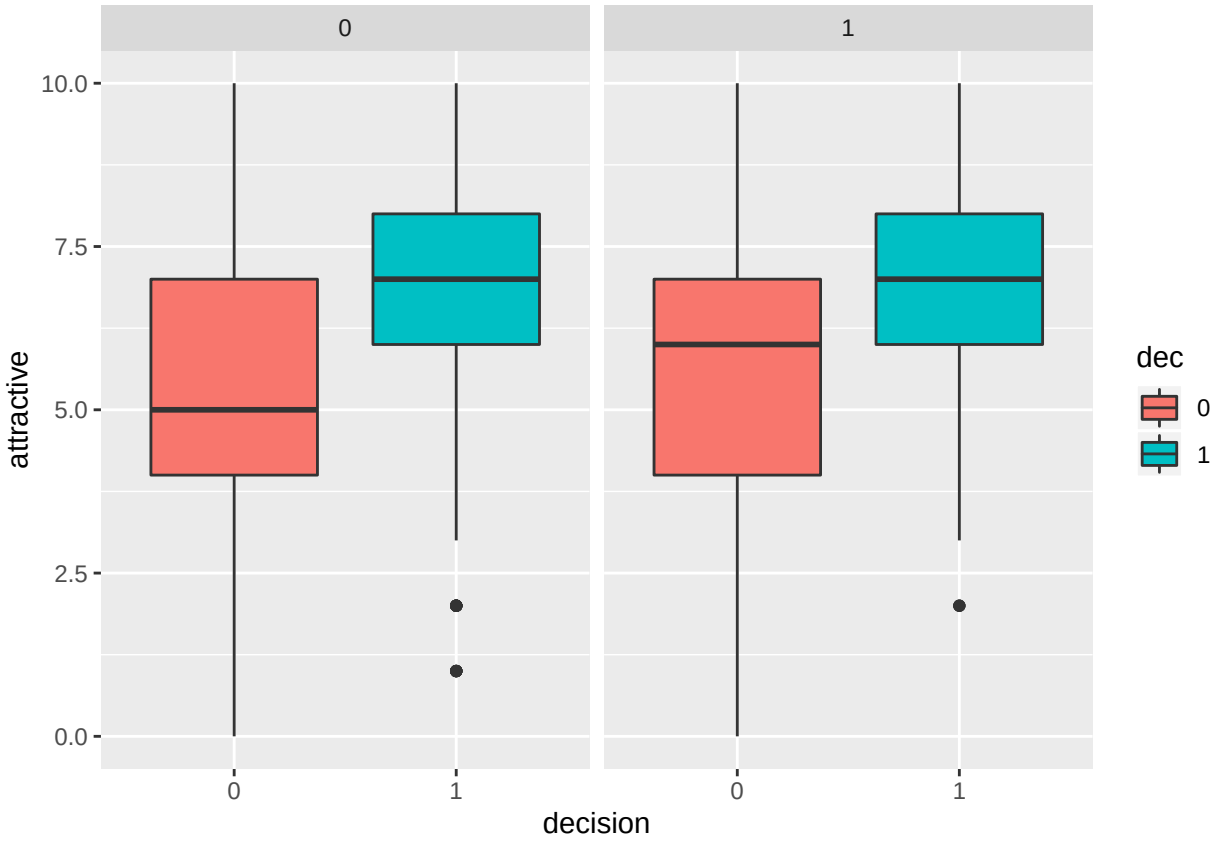

```
##
##  Variables sorted by number of missings:
##  Variable       Count
##      shar 0.103028019
##       amb 0.061691167
##       fun 0.017087574
##     intel 0.011559241
##      sinc 0.009046363
##      attr 0.001256439
##       iid 0.000000000
##      wave 0.000000000
##       pid 0.000000000
##       dec 0.000000000
##    gender 0.000000000
##  samerace 0.000000000
##     age_o 0.000000000
##      prob 0.000000000
##      like 0.000000000
```

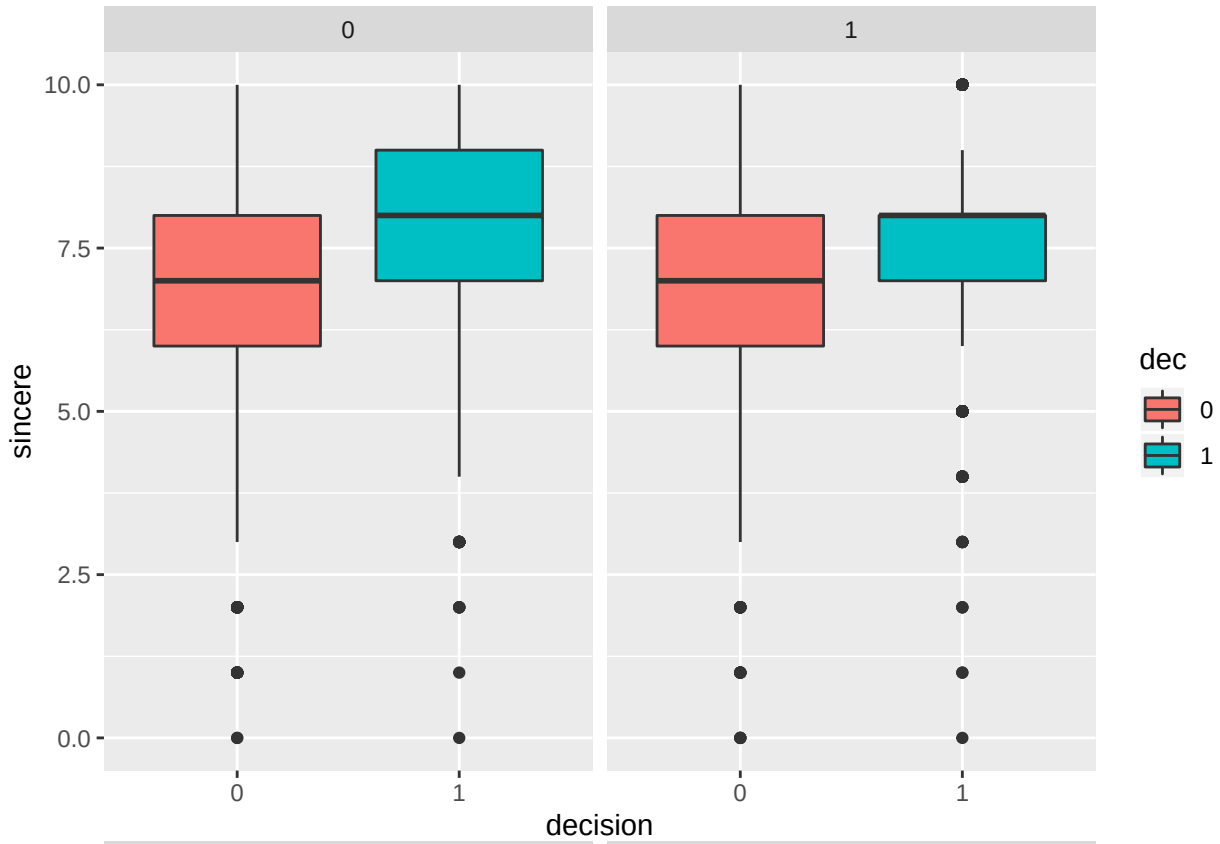
Appendix 2 (Imputed data for the model 1-1)

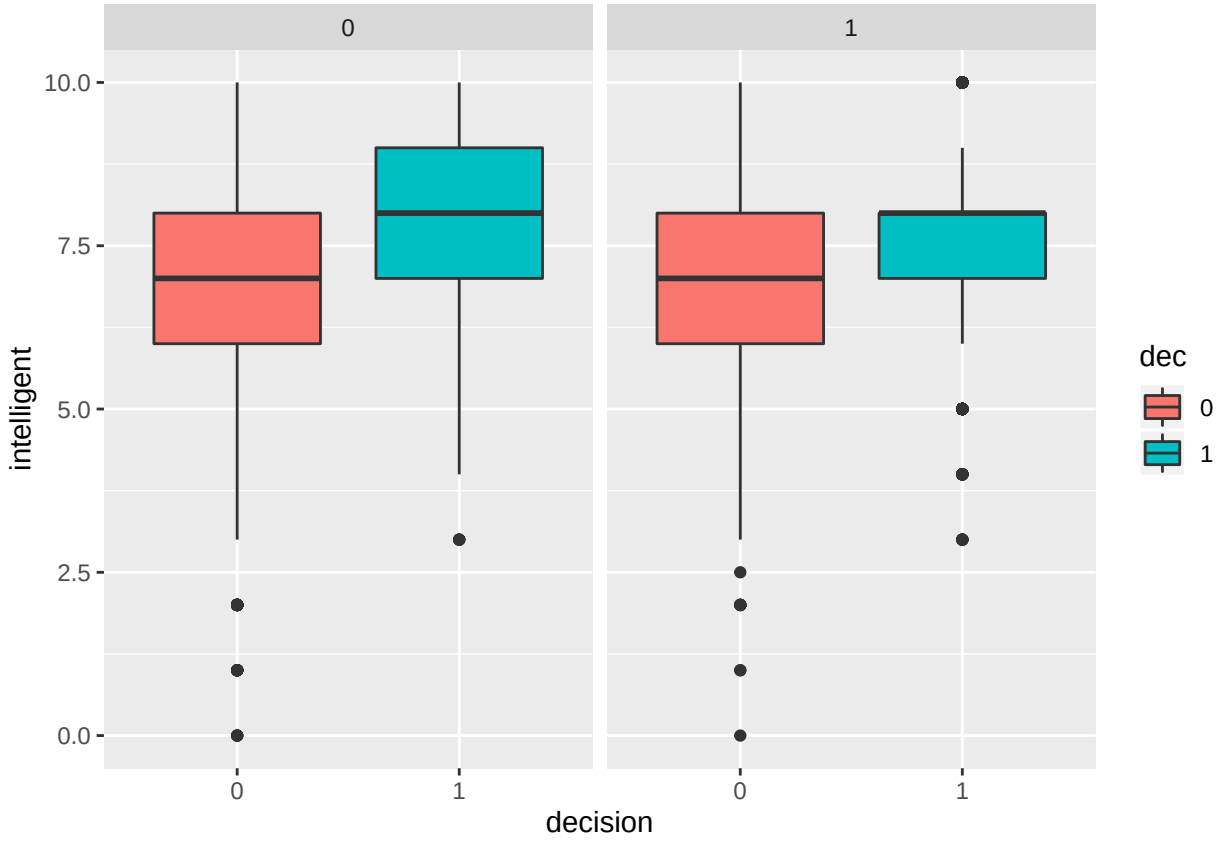


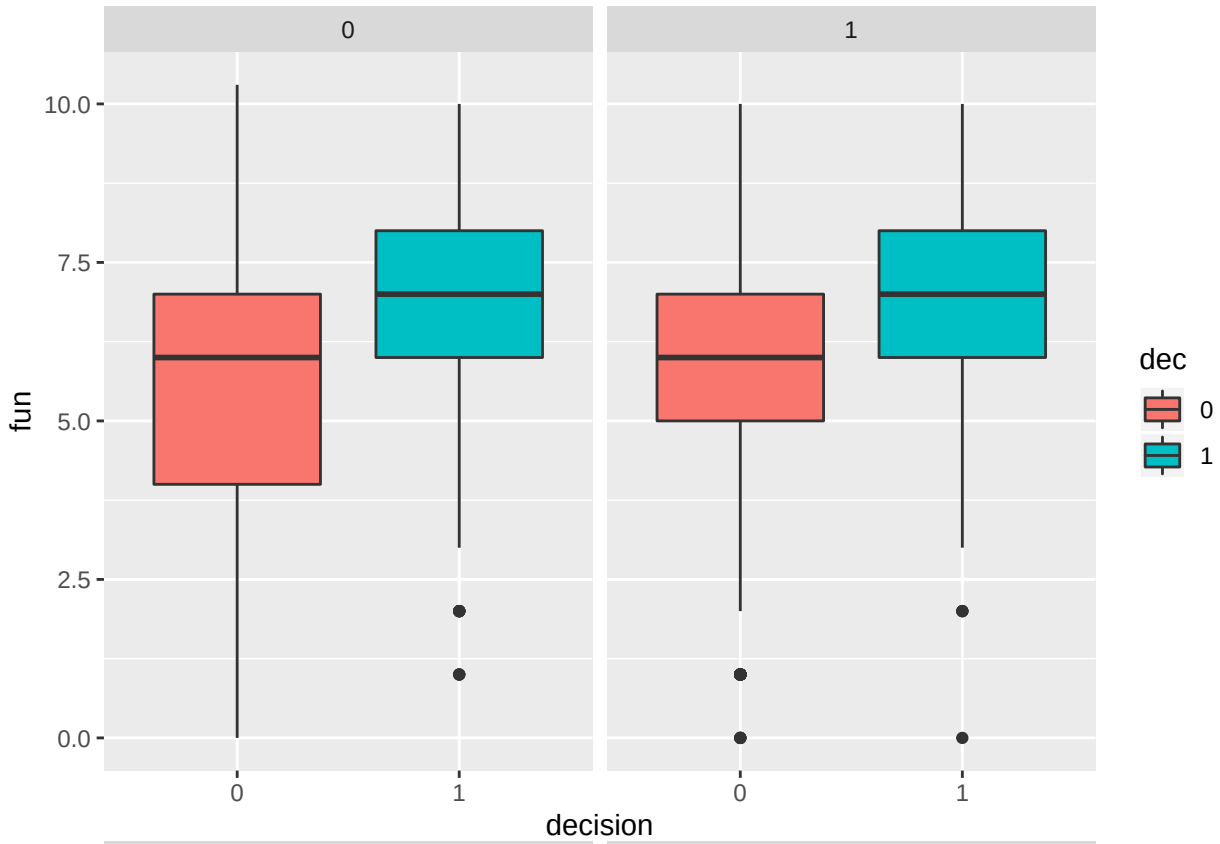
Appendix 3 (EDA for the model 1-1)

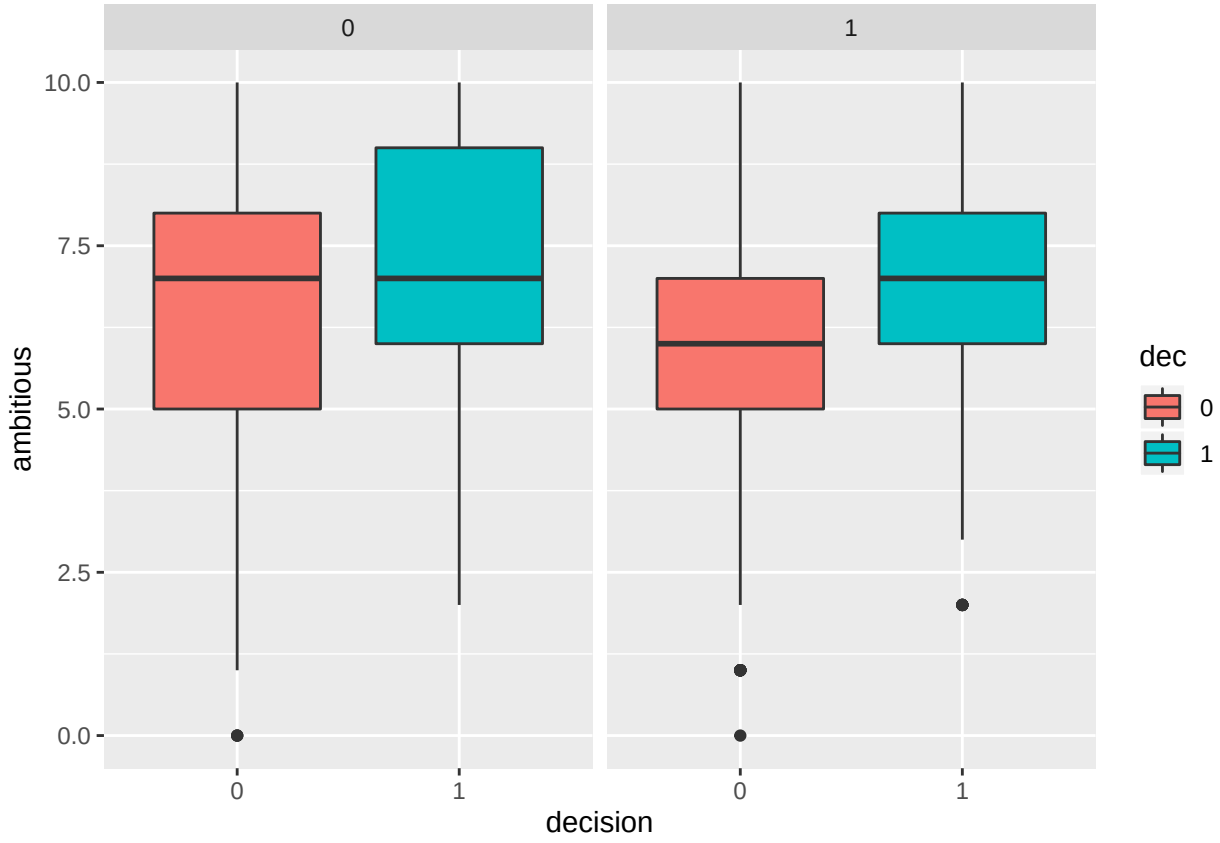
```
## Data: n1
## Models:
## nullmodel1: dec ~ 1 + (1 | wave)
## nullmodel2: dec ~ 1 + (attr | wave)
##            Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## nullmodel1  2 9429.1 9442.8 -4712.6   9425.1
## nullmodel2  4 7579.5 7606.9 -3785.8   7571.5 1853.6      2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

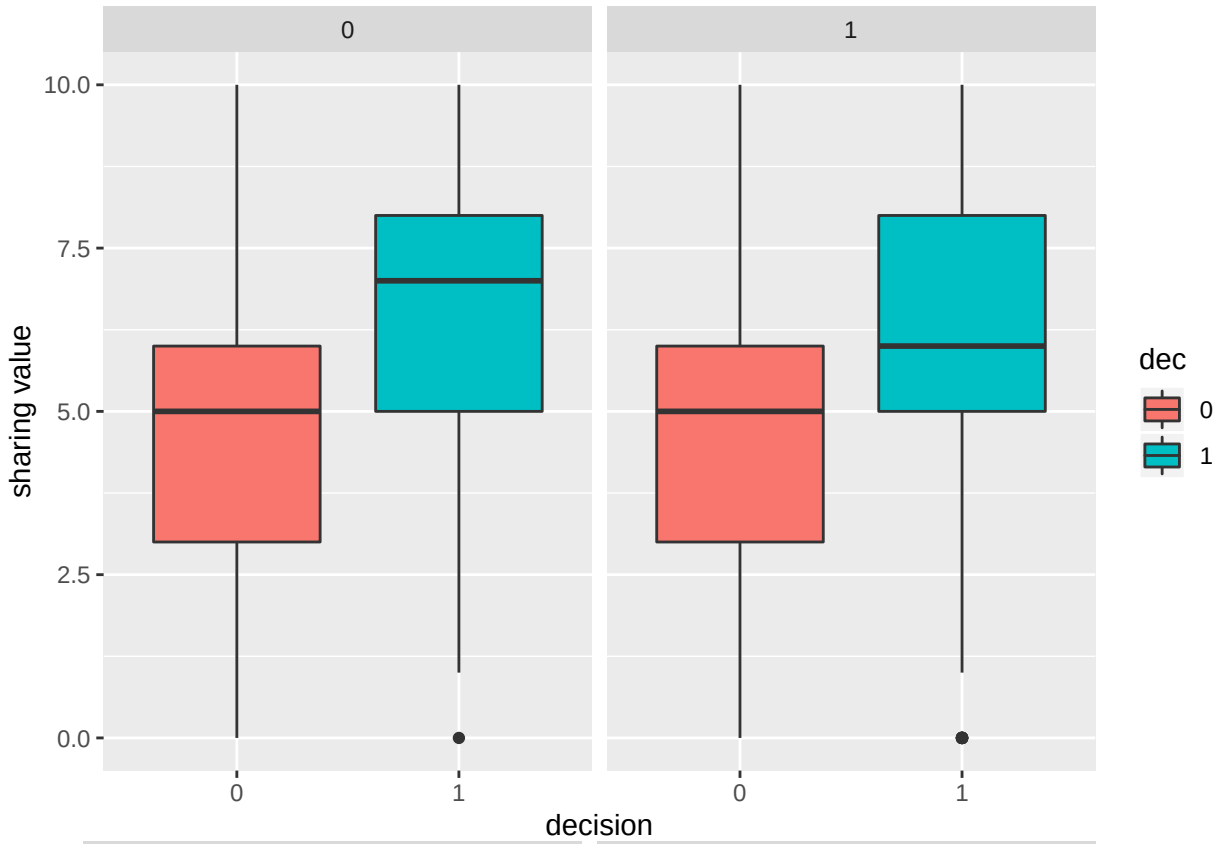


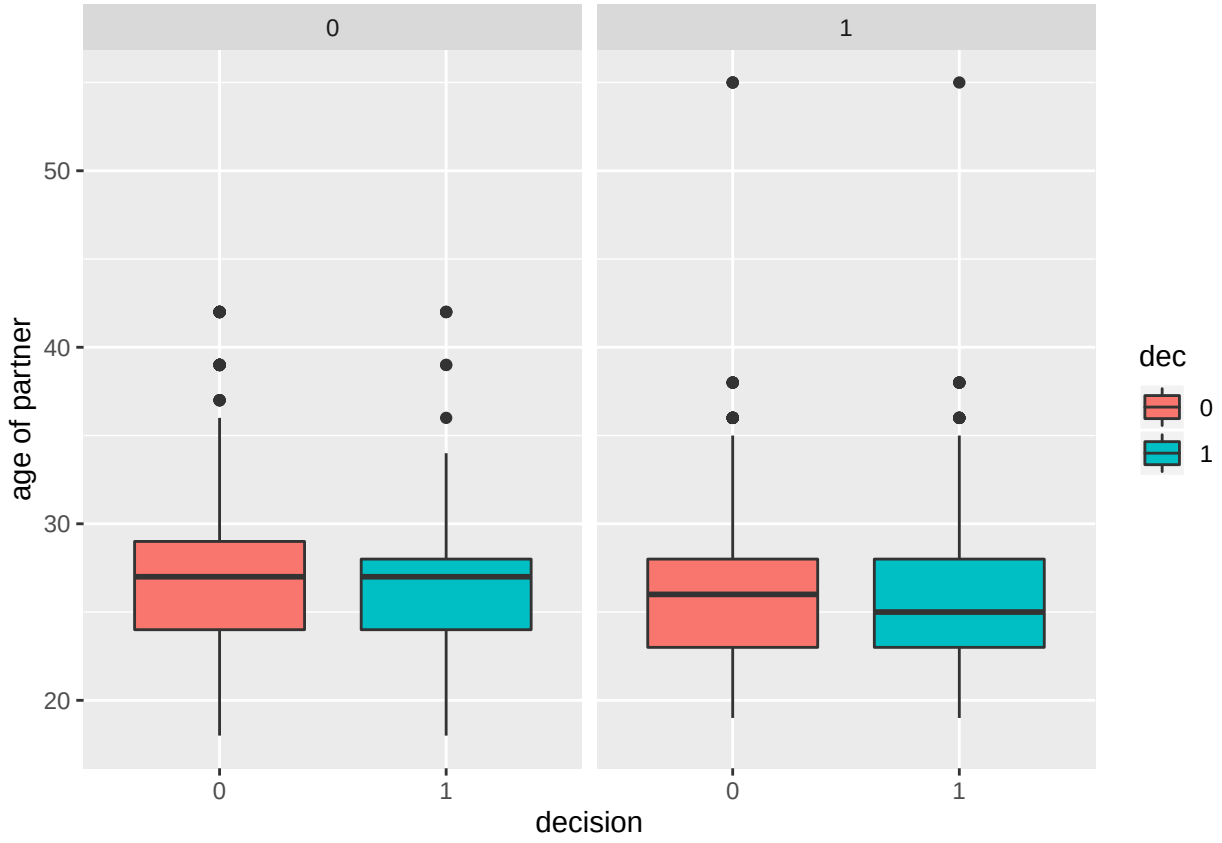

0
1
10.0
7.5
5.0
2.5
0.0
sincere
0
1
0
1
decision
dec
0
1

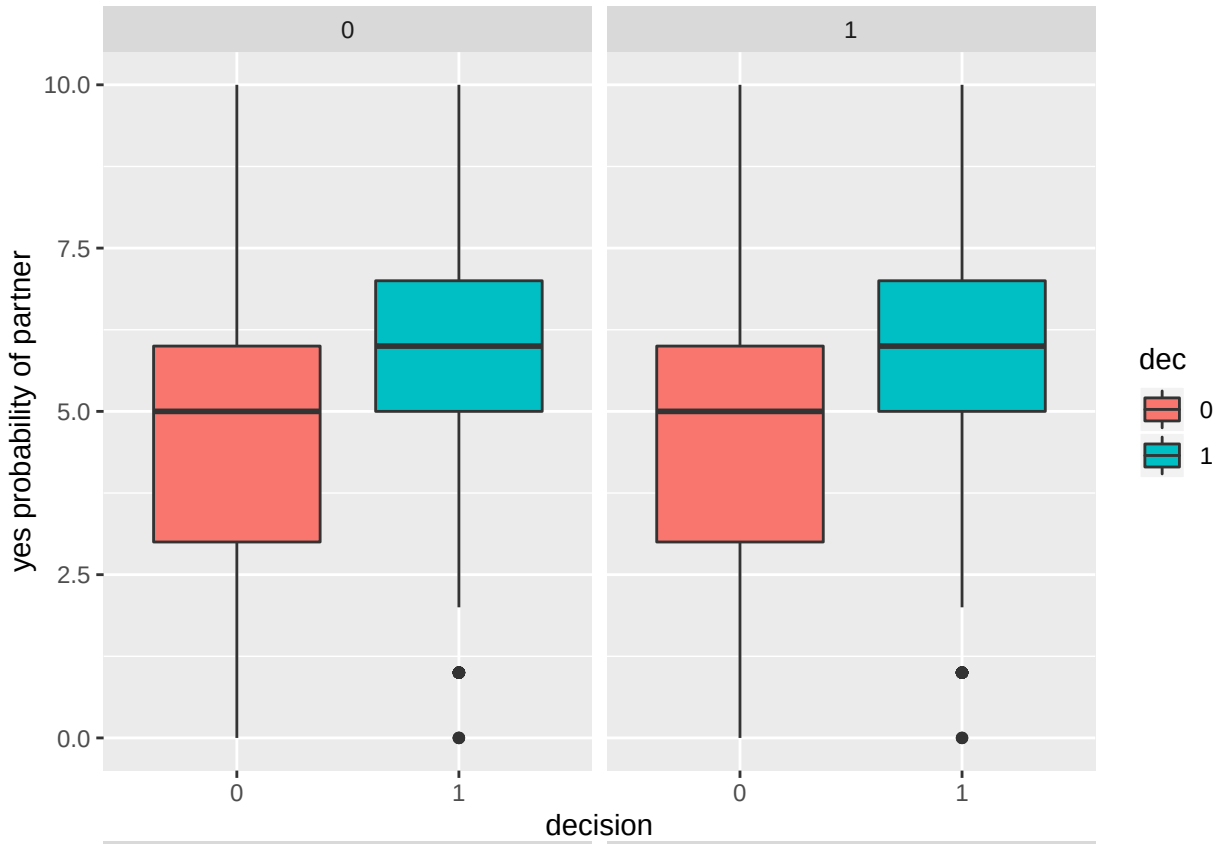

0
1
10.0
7.5
5.0
2.5
0.0
intelligent
0
1
0
1
decision
dec
0
1

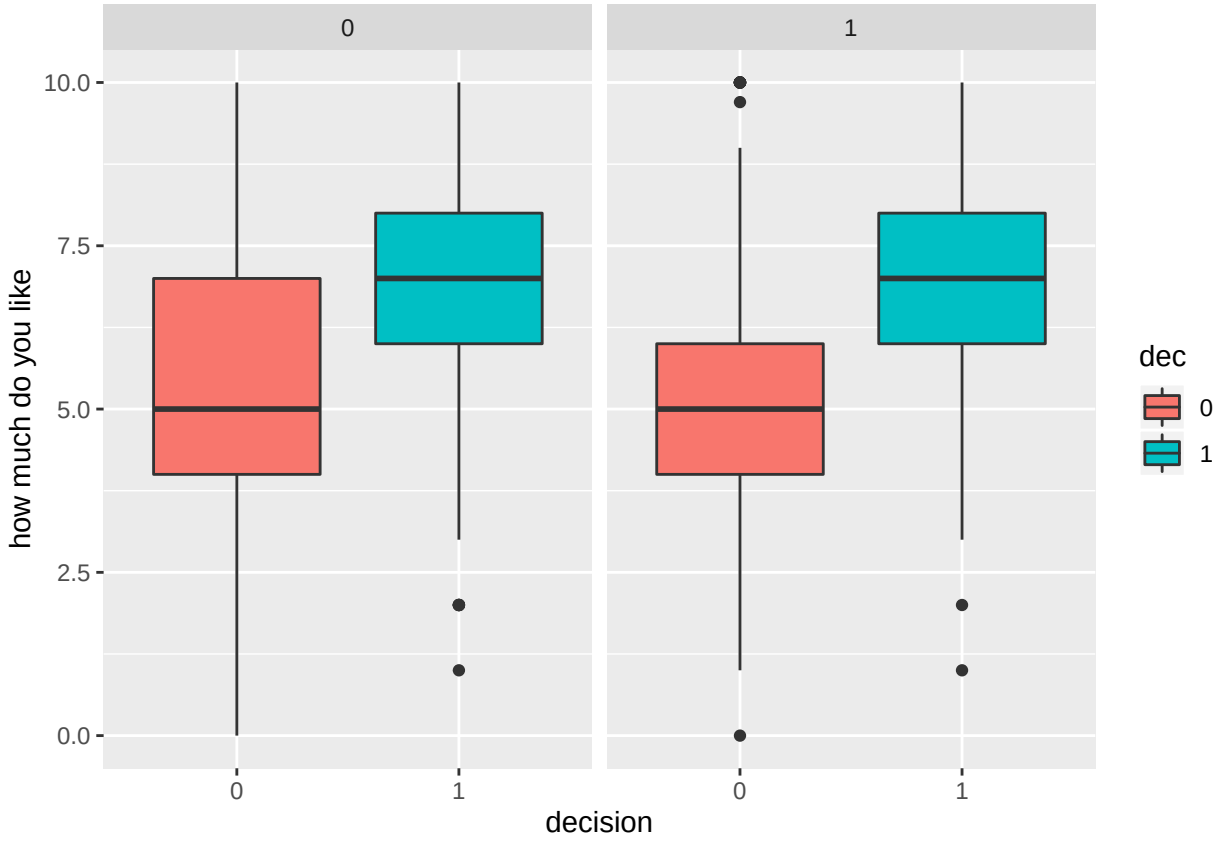

0
1
10.0
7.5
5.0
2.5
0.0
fun
0
1
decision
dec
0
1

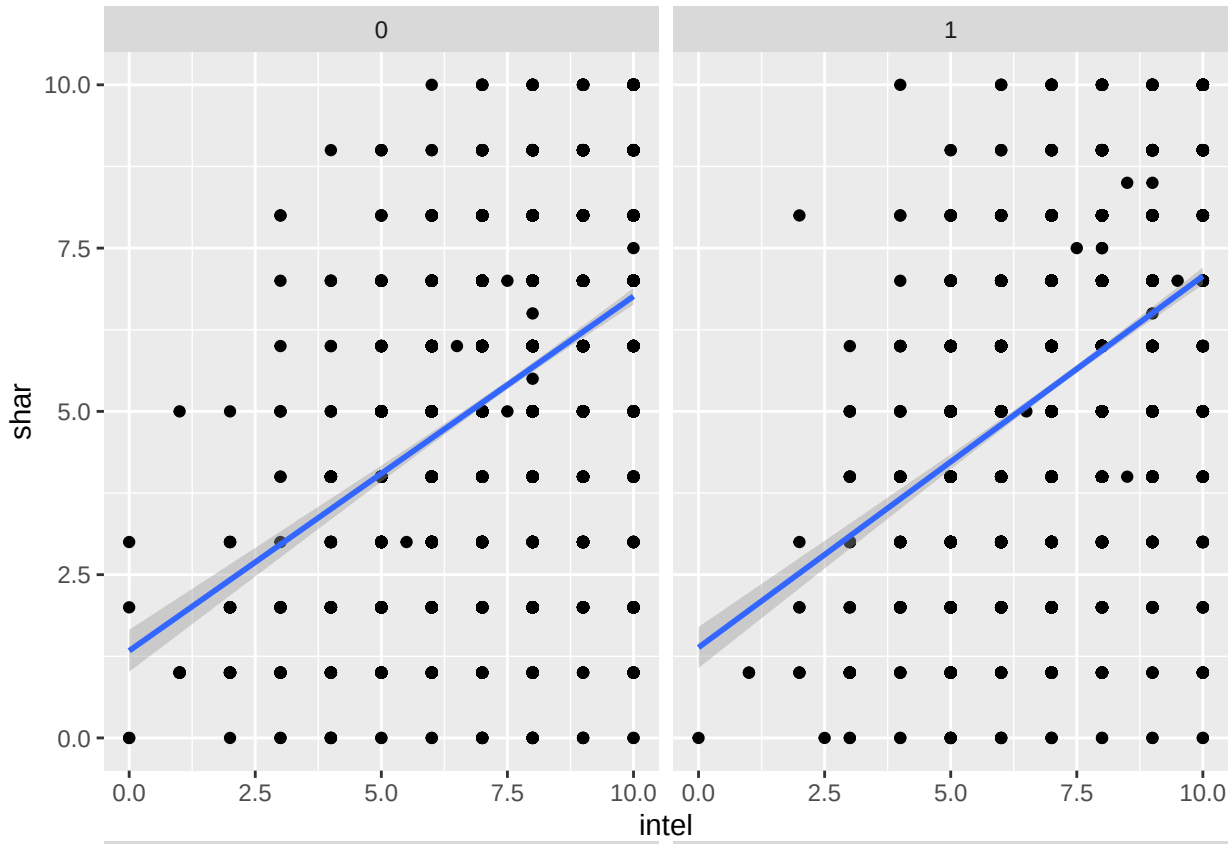

0
1
10.0
7.5
5.0
2.5
0.0
ambitious
0
1
decision
dec
0
1

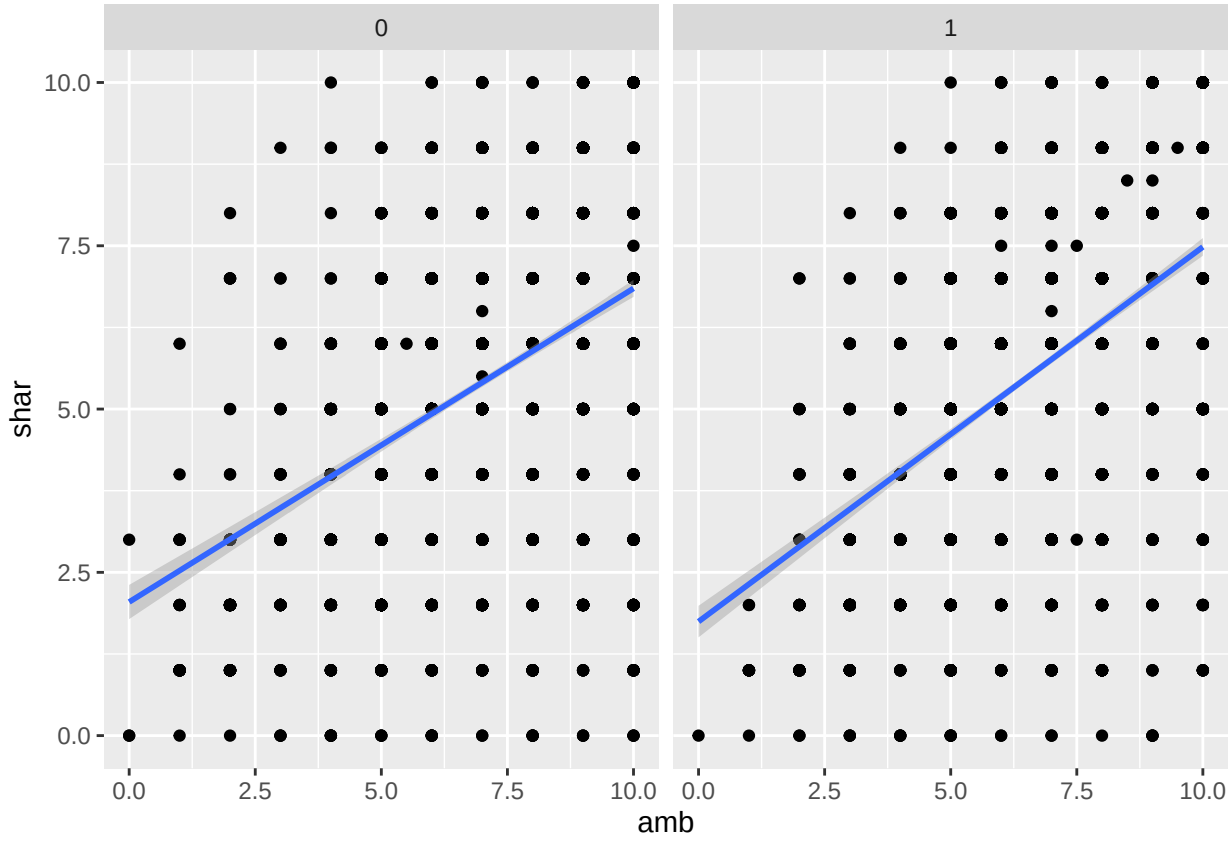

0
1
10.0
7.5
5.0
2.5
0.0
sharing value
0
1
0
1
decision
dec
0
1

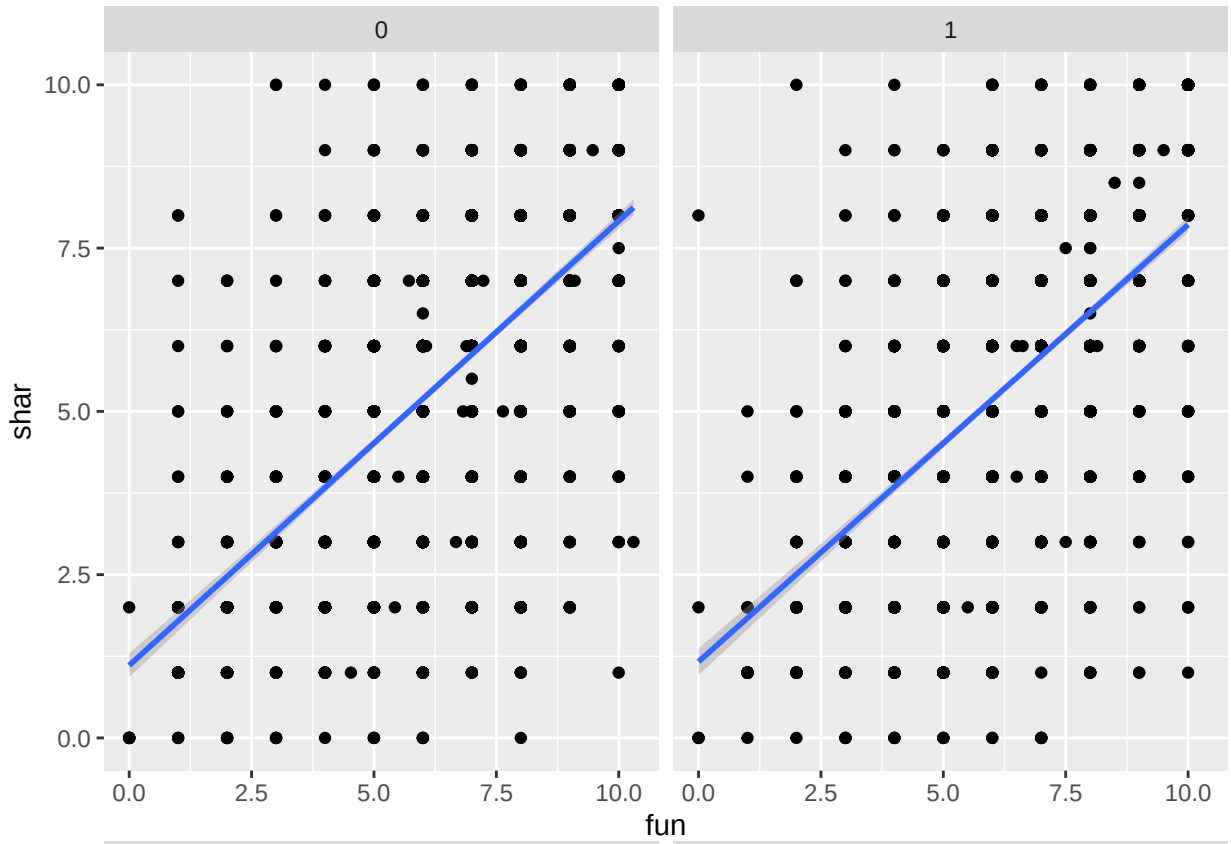

0
1
50
40
30
20
age of partner
0
1
0
1
decision
dec
0
1

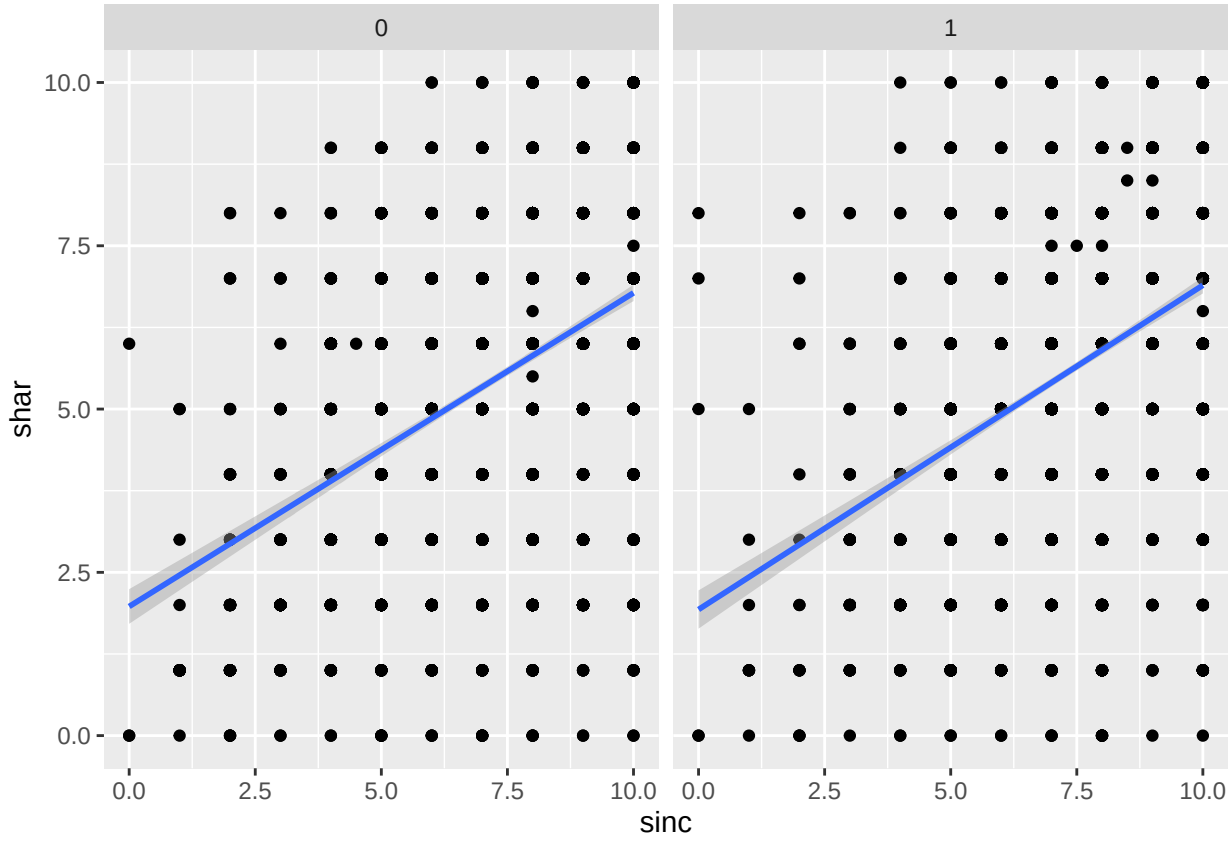

0
1
10.0
7.5
5.0
2.5
0.0
yes probability of partner
0
1
decision
dec
0
1

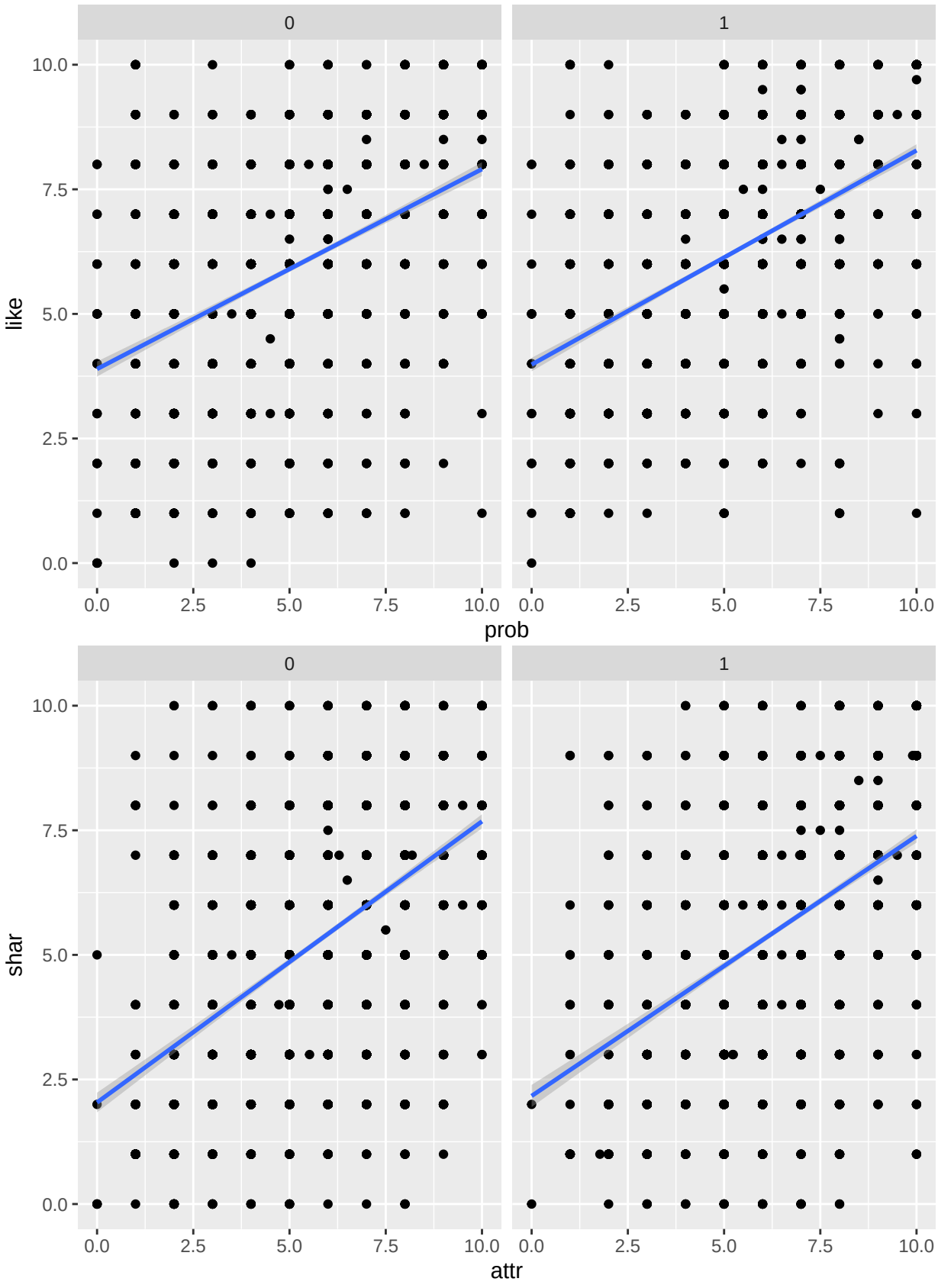

0
1
10.0
7.5
5.0
2.5
0.0
how much do you like
0
1
decision
dec
0
1


0
1
shar
10.0
7.5
5.0
2.5
0.0
0.0
2.5
5.0
7.5
10.0
intel


0
1
shar
10.0
7.5
5.0
2.5
0.0
0.0
2.5
5.0
7.5
10.0
amb


0
1
shar
10.0
7.5
5.0
2.5
0.0
0.0
2.5
5.0
7.5
10.0
fun


0
1
shar
10.0
7.5
5.0
2.5
0.0
0.0
2.5
5.0
7.5
10.0
sinc




```
##    samerace
```

```
## dec         0         1
##   0 0.3084165 0.2042796
##   1 0.2915835 0.1957204
```

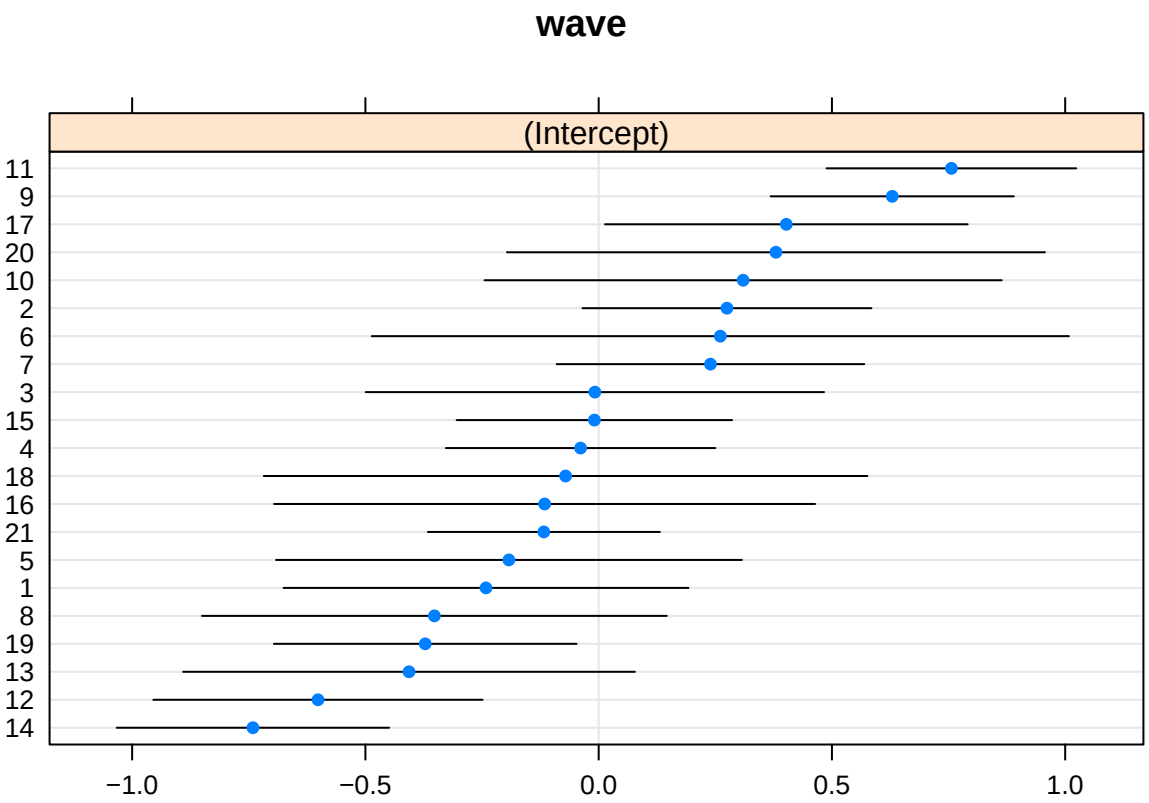
```
##    samerace
## dec         0         1
##   0 0.3981781 0.2315604
##   1 0.2118719 0.1583897
```

Appendix 4 (Varing intercept for the male model)

```
dotplot(ranef(dec_m, condVar=TRUE, main = FALSE))
```

```
## Warning in ranef.merMod(dec_m, condVar = TRUE, main = FALSE): additional
## arguments to ranef.merMod ignored: main
```

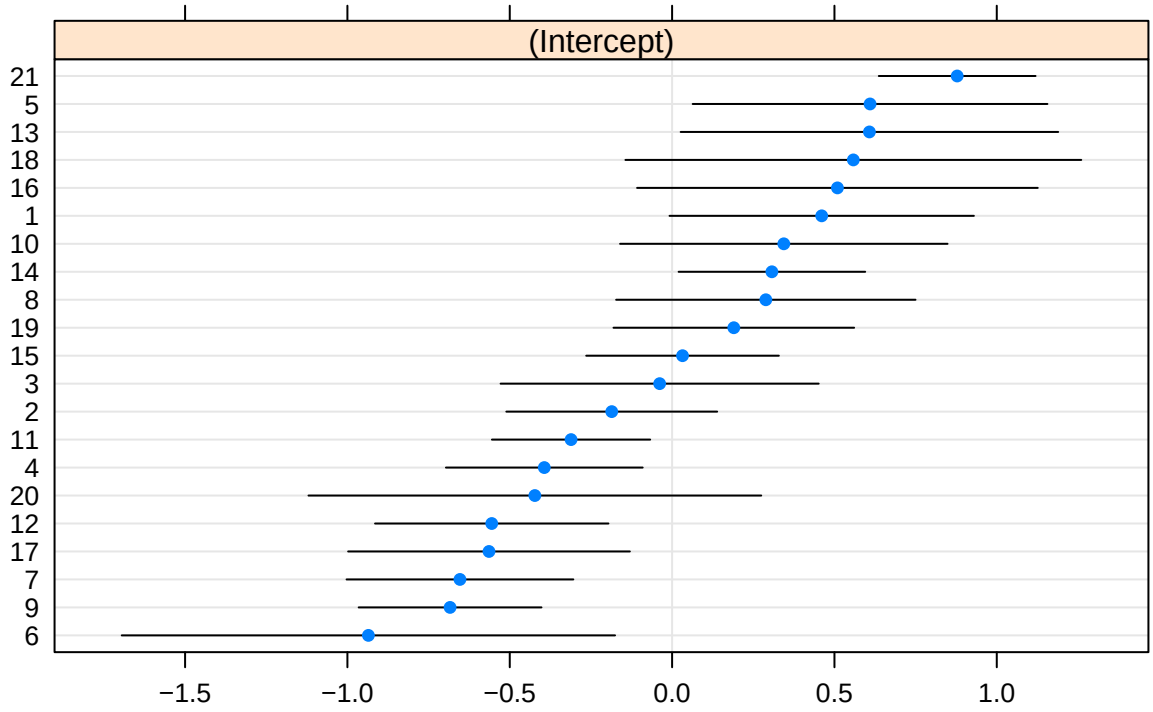
```
## $wave
```



Appendix 5 (Varing intercept for the female model)

```
dotplot(ranef(dec_f, condVar=TRUE))
```

```
## $wave
```



Appendix 6 (Confusion matrix and ROC curve for the male model)

```
## 1. For male
#confusion matrix

Conf_mat <- confusionMatrix(as.factor(ifelse(fitted(dec_m) >= 0.437, "1","0")),
                            as.factor(n1_m$dec),positive = "1")
Conf_mat$table
```

```
##            Reference
## Prediction    0    1
##          0 1356  267
##          1  441 1441
```

```
# accuracy: 0.80
Conf_mat$overall["Accuracy"];
```

```
##  Accuracy
## 0.7980029
```

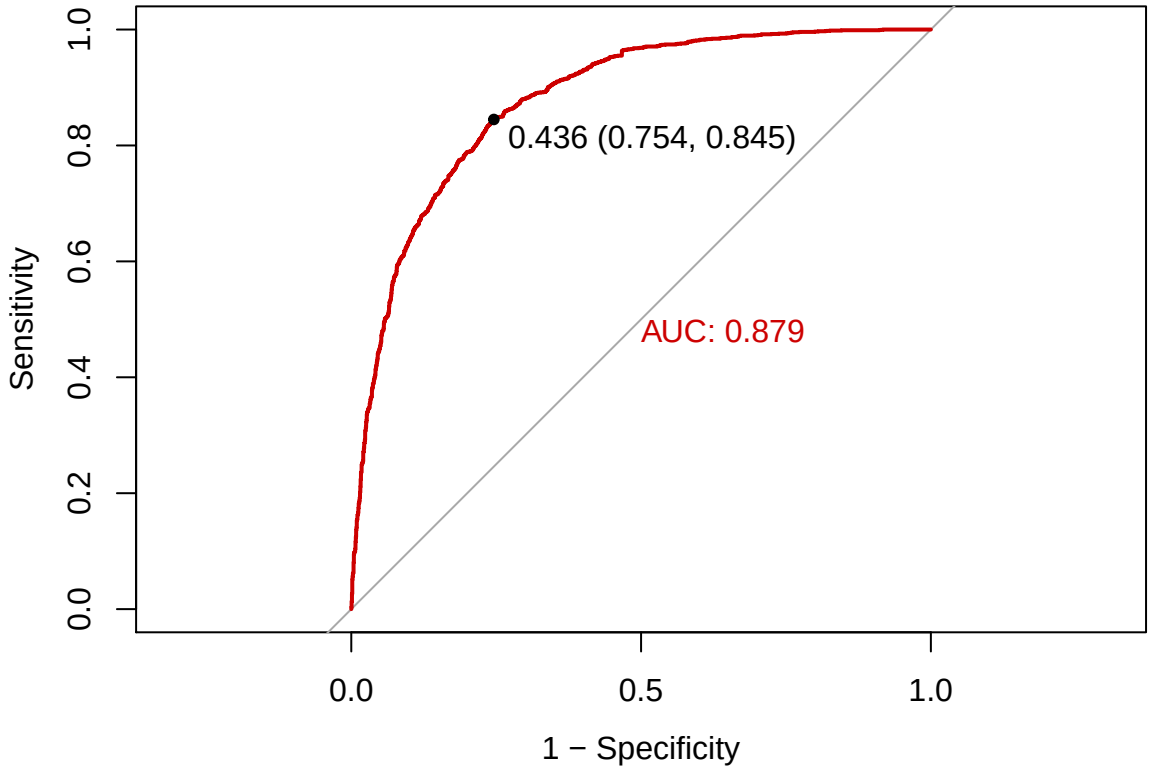
```
Conf_mat$byClass[c("Sensitivity","Specificity")]
```

```
## Sensitivity Specificity
##   0.8436768   0.7545910
```

```
#Roc curve
roc(n1_m$dec,fitted(dec_m),plot=T,print.thres="best",legacy.axes=T,
    print.auc =T,col="red3")
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```




```
##
## Call:
## roc.default(response = n1_m$dec, predictor = fitted(dec_m), plot = T,     print.thres = "best", lega
##
## Data: fitted(dec_m) in 1797 controls (n1_m$dec 0) < 1708 cases (n1_m$dec 1).
## Area under the curve: 0.8786
```

Appendix 7 (Confusion matrix and ROC curve for the female model)

```
##2. For female
Conf_mat2 <- confusionMatrix(as.factor(ifelse(fitted(dec_f) >= 0.299, "1","0")),
                             as.factor(n1_f$dec),positive = "1")
Conf_mat2$table
```

```
##           Reference
## Prediction    0    1
##          0 1453  173
##          1  690 1087
```

```
#accuracy : 0.75
Conf_mat2$overall["Accuracy"];
```

```
##  Accuracy
## 0.7464002
```

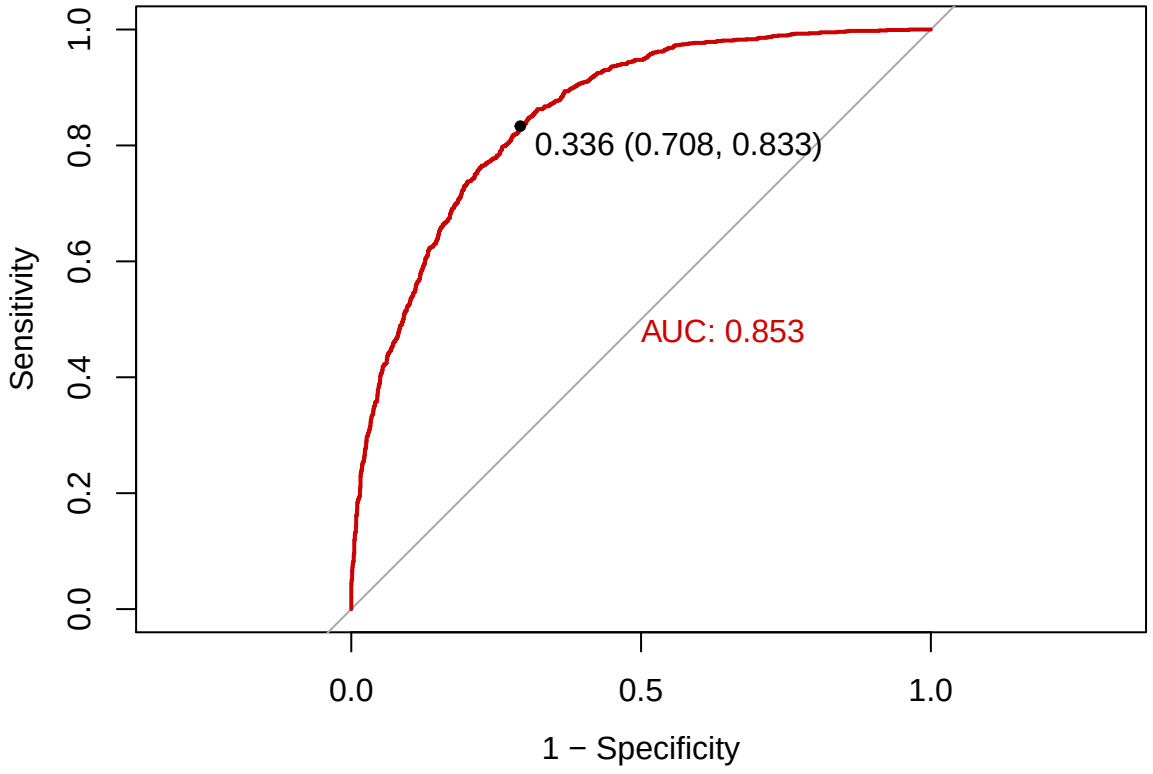
```
Conf_mat2$byClass[c("Sensitivity","Specificity")]
```

```
## Sensitivity Specificity
##   0.8626984   0.6780215
```

```
#Roc curve
roc(n1_f$dec,fitted(dec_f),plot=T,print.thres="best",legacy.axes=T,
    print.auc =T,col="red3")
```

```
## Setting levels: control = 0, case = 1
```

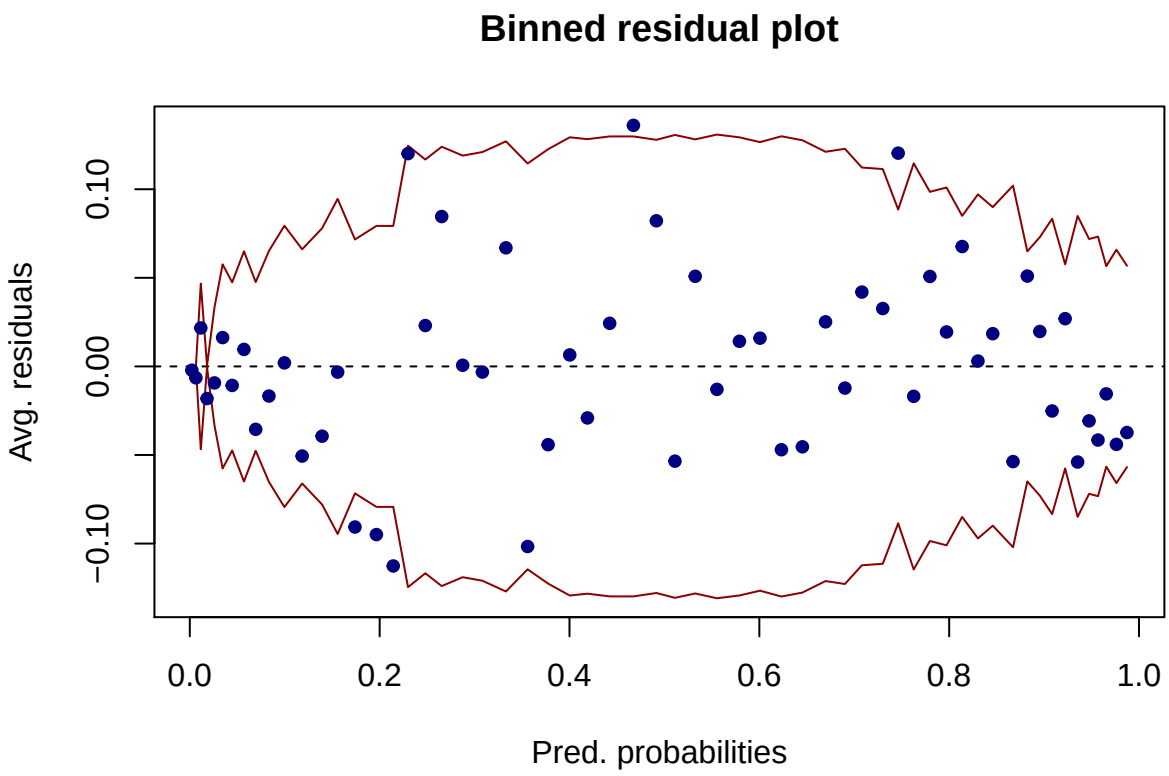
```
## Setting direction: controls < cases
```



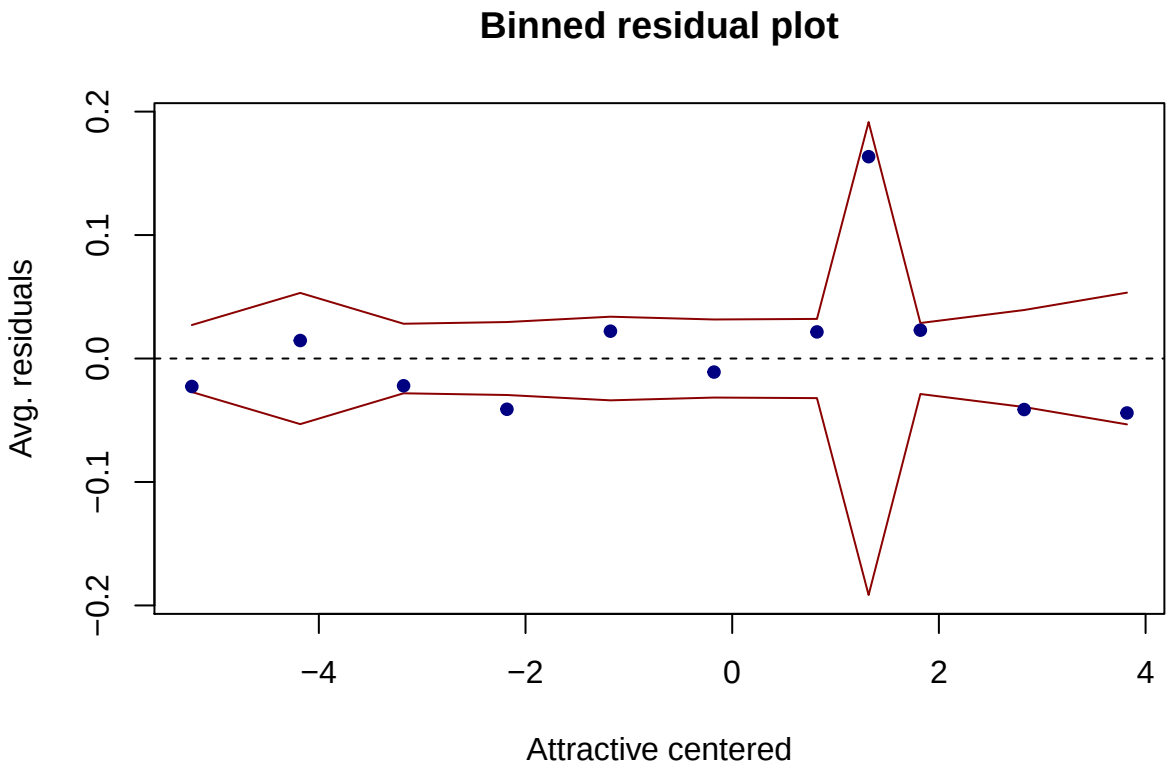

```
##
## Call:
## roc.default(response = n1_f$dec, predictor = fitted(dec_f), plot = T,     print.thres = "best", lega
##
## Data: fitted(dec_f) in 2143 controls (n1_f$dec 0) < 1260 cases (n1_f$dec 1).
## Area under the curve: 0.8528
```

Appendix 8 (Binned residual for the male model)

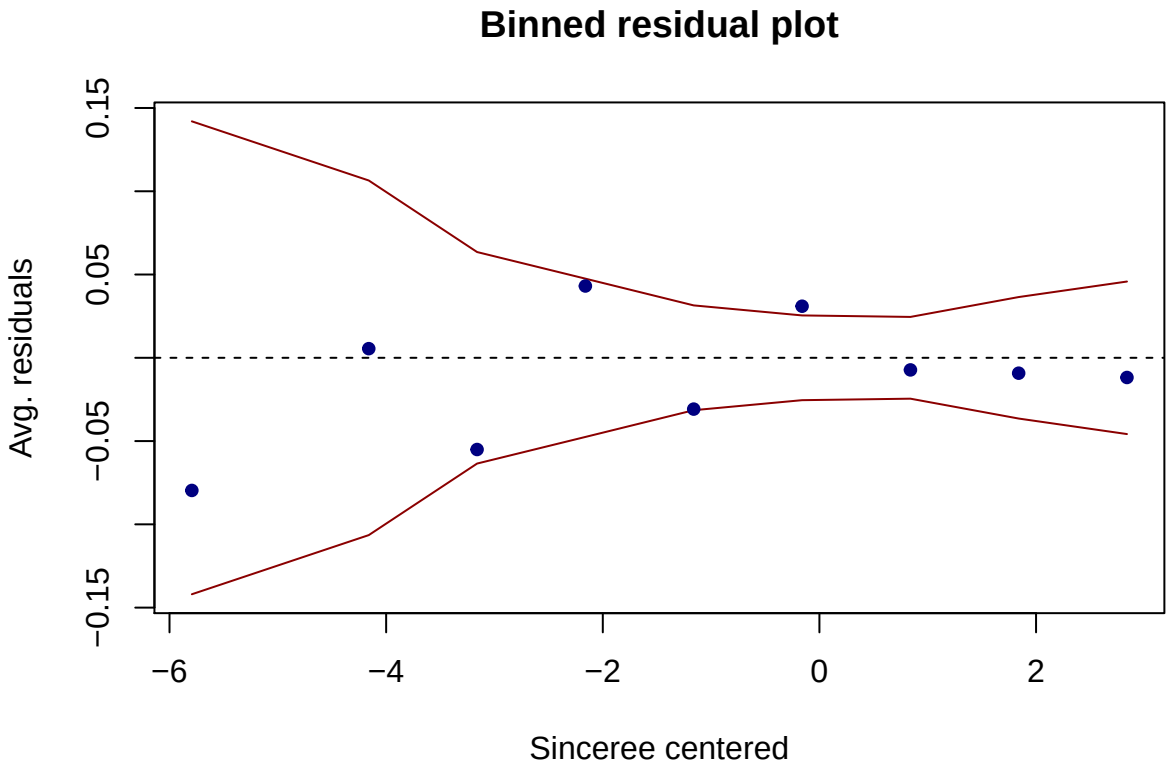
```
## 1. For male

rawresid <- residuals(dec_m,"resp")
#binned residual plots
binnedplot(x=fitted(dec_m),y=rawresid,xlab="Pred. probabilities",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



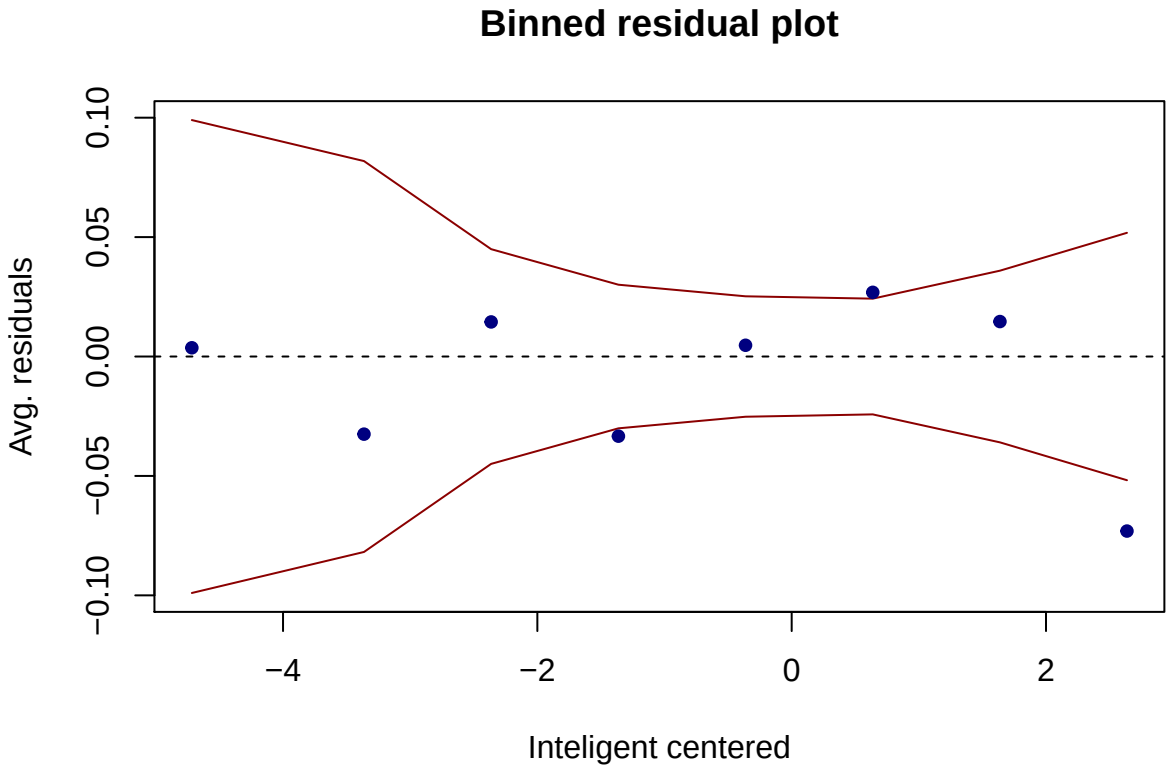

```
binnedplot(n1_m$attr_c,y=rawresid,xlab="Attractive centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



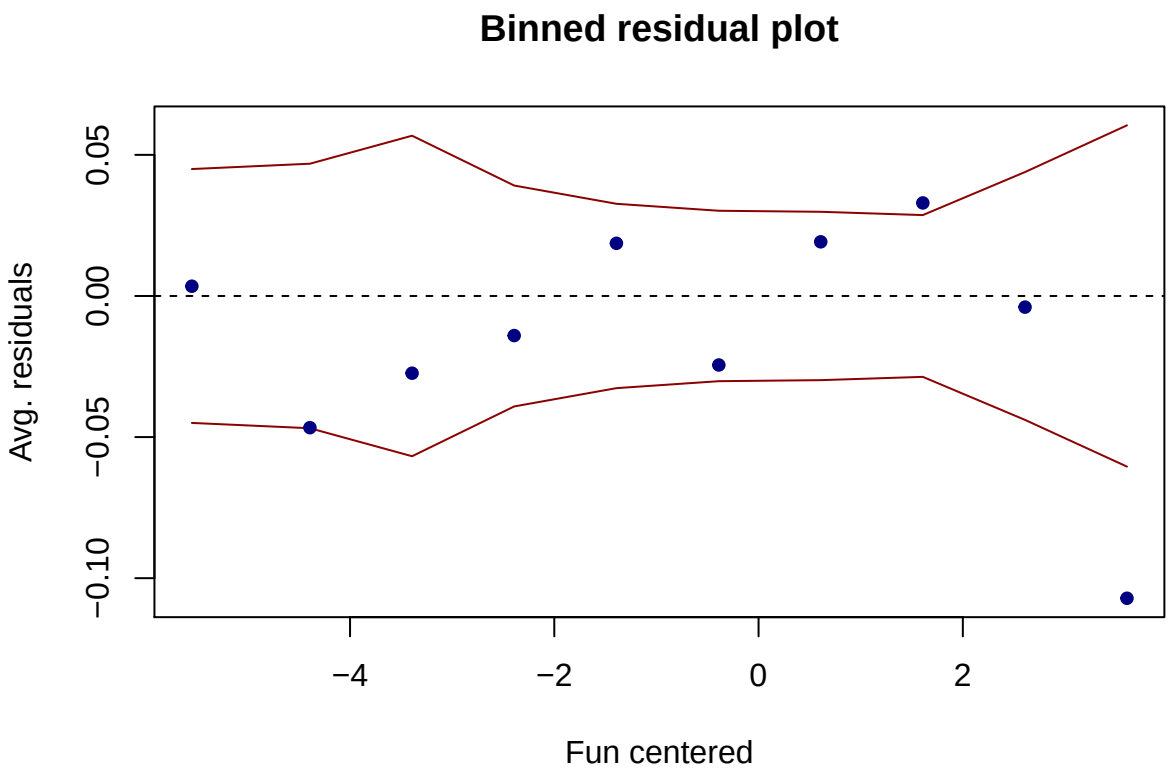

```
binnedplot(n1_m$sinc_c,y=rawresid,xlab="Sinceree centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



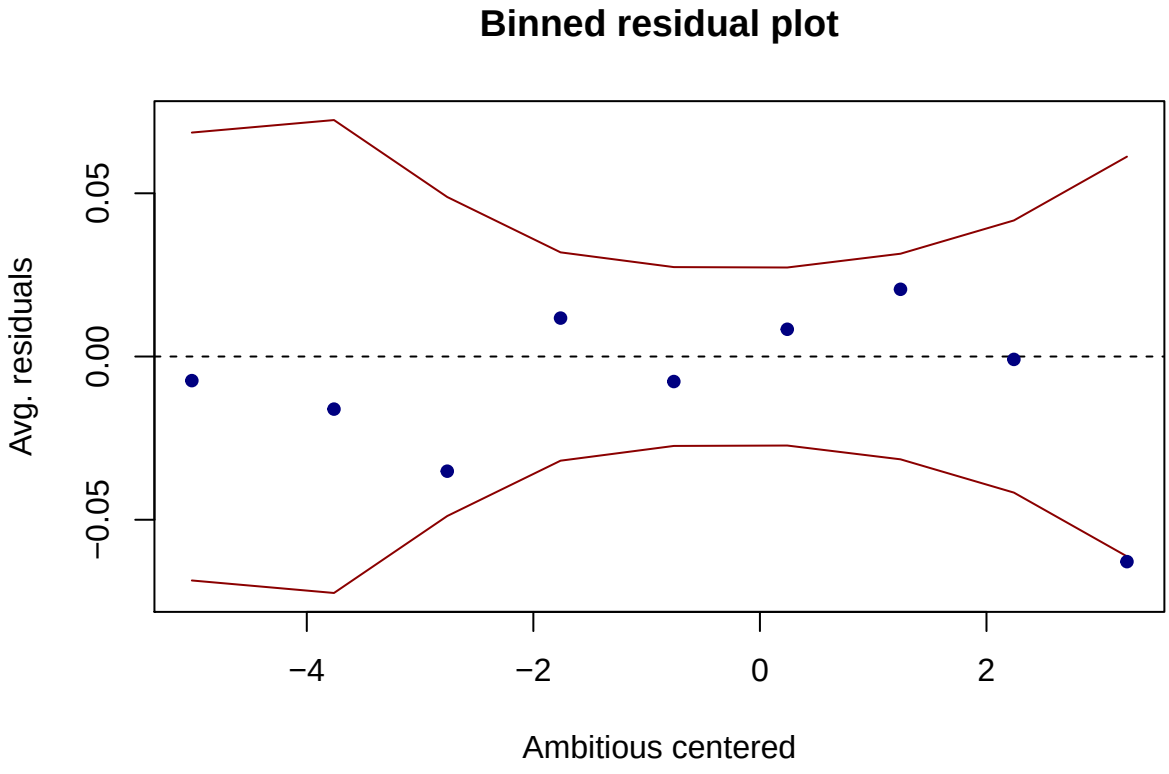

```
binnedplot(n1_m$intel_c,y=rawresid,xlab="Inteligent centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



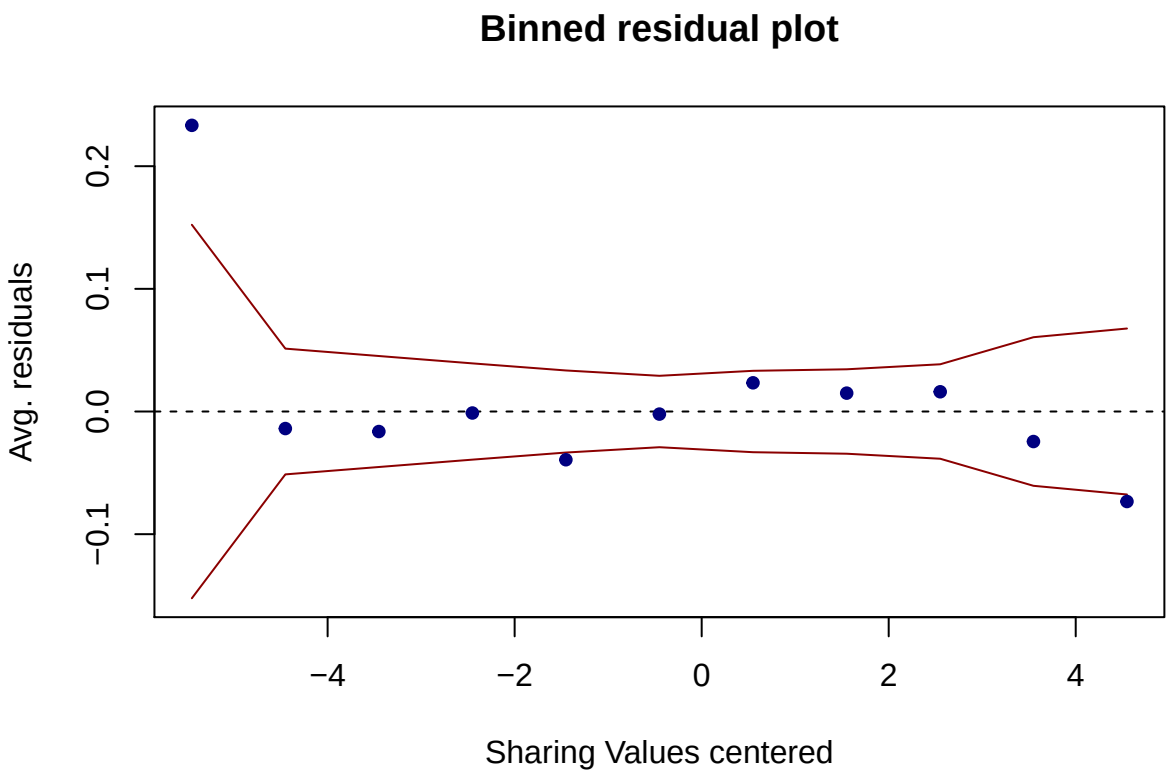

```
binnedplot(n1_m$fun_c,y=rawresid,xlab="Fun centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



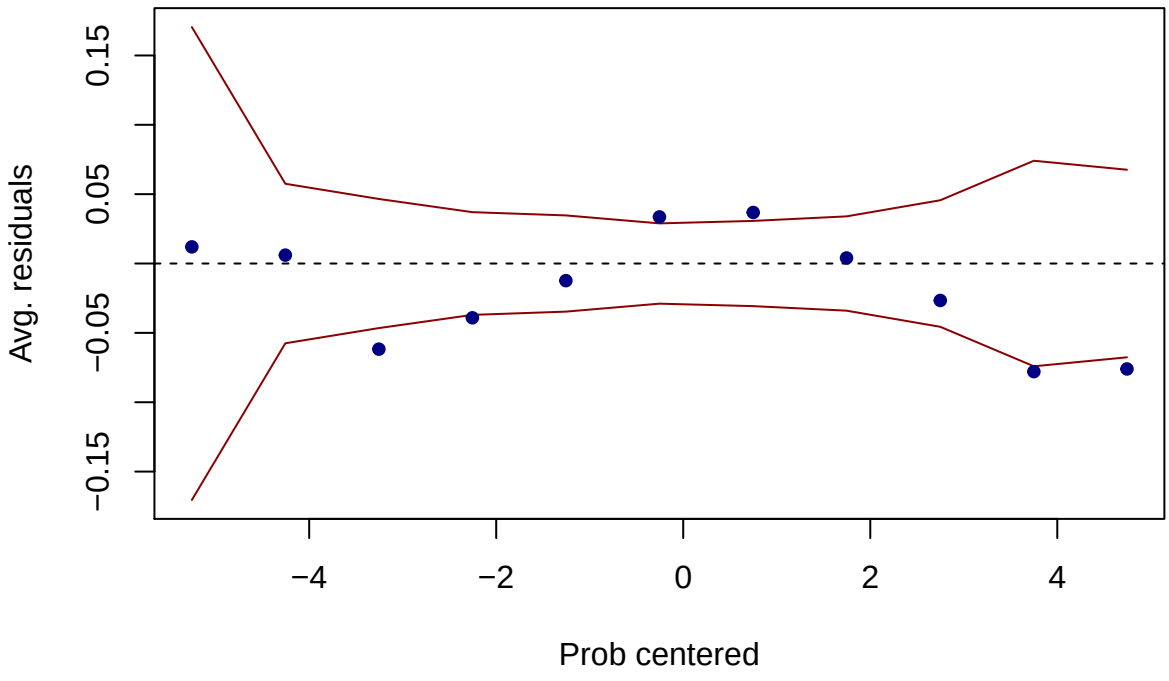

```
binnedplot(n1_m$amb_c,y=rawresid,xlab="Ambitious centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



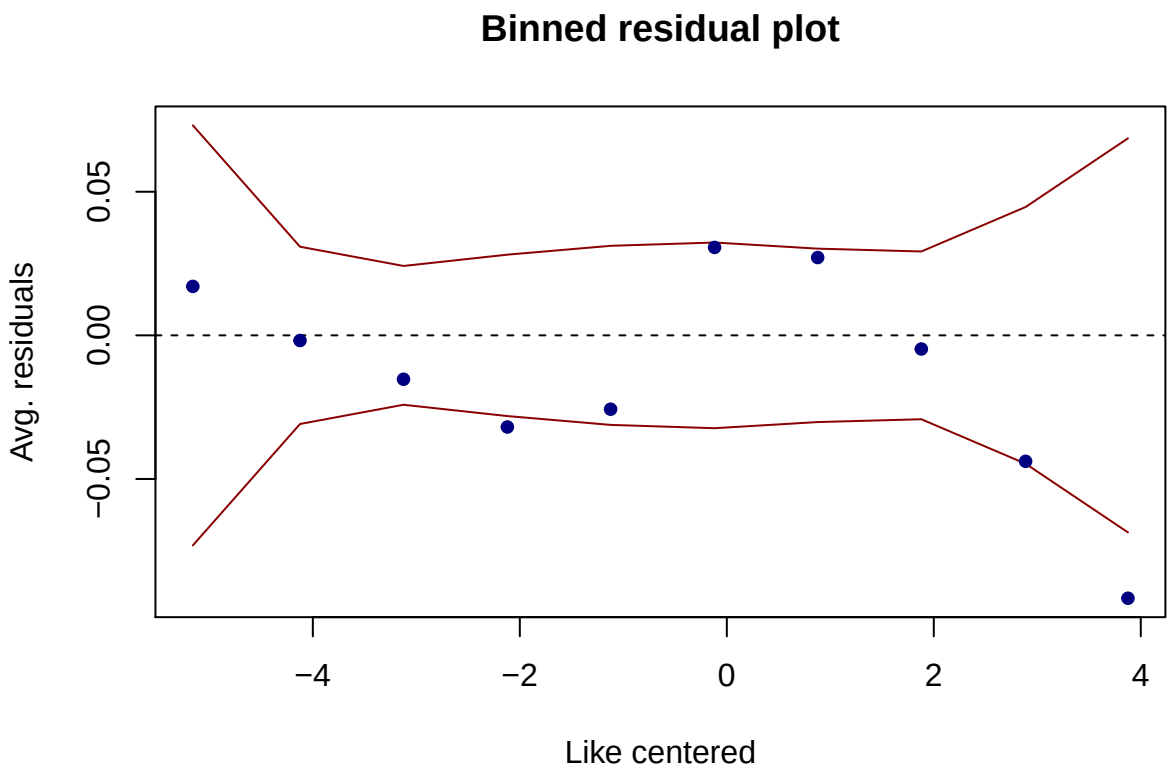

```
binnedplot(n1_m$shar_c,y=rawresid,xlab="Sharing Values centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```




```
binnedplot(n1_m$prob_c,y=rawresid,xlab="Prob centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```

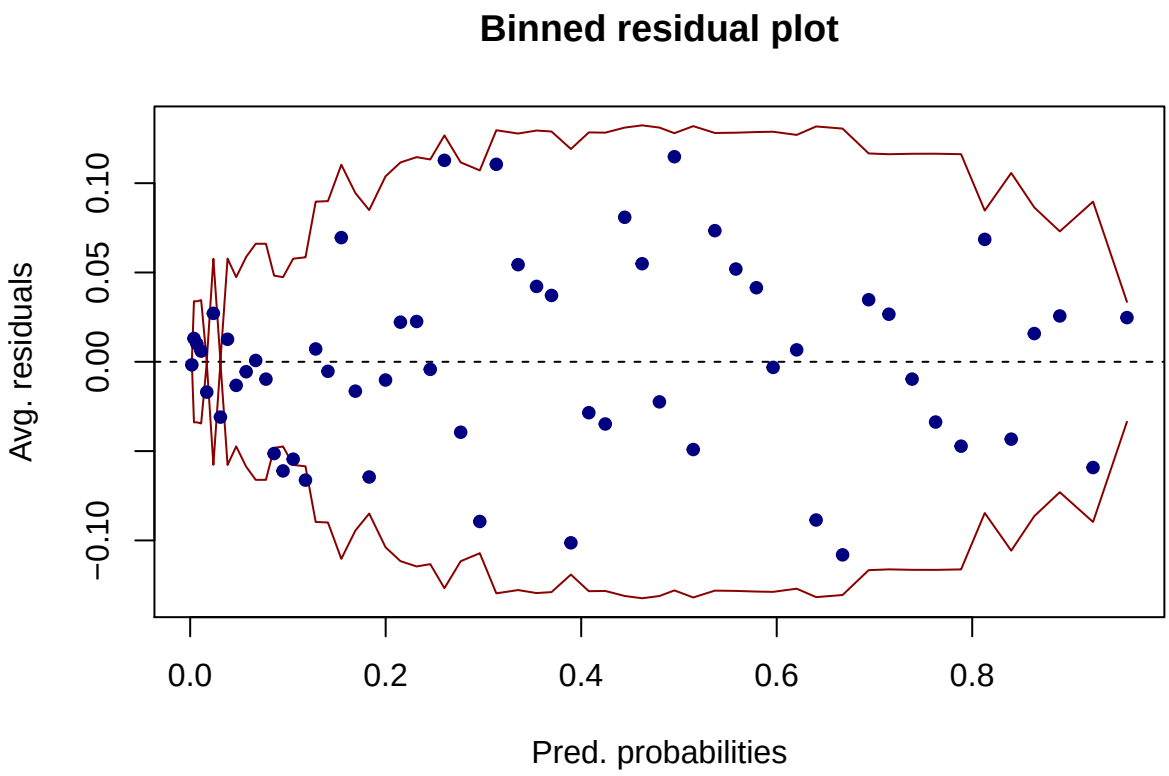



```
binnedplot(n1_m$like_c,y=rawresid,xlab="Like centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```

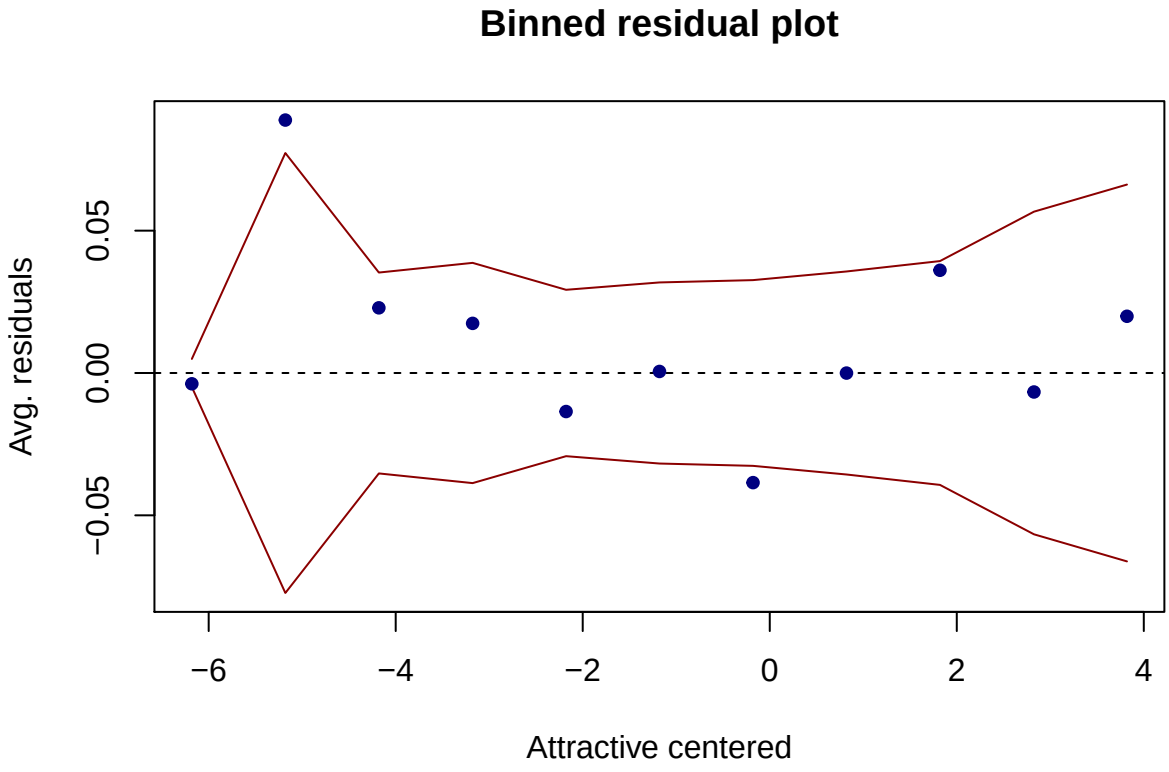


Appendix 9 (Binned residual for the female model)

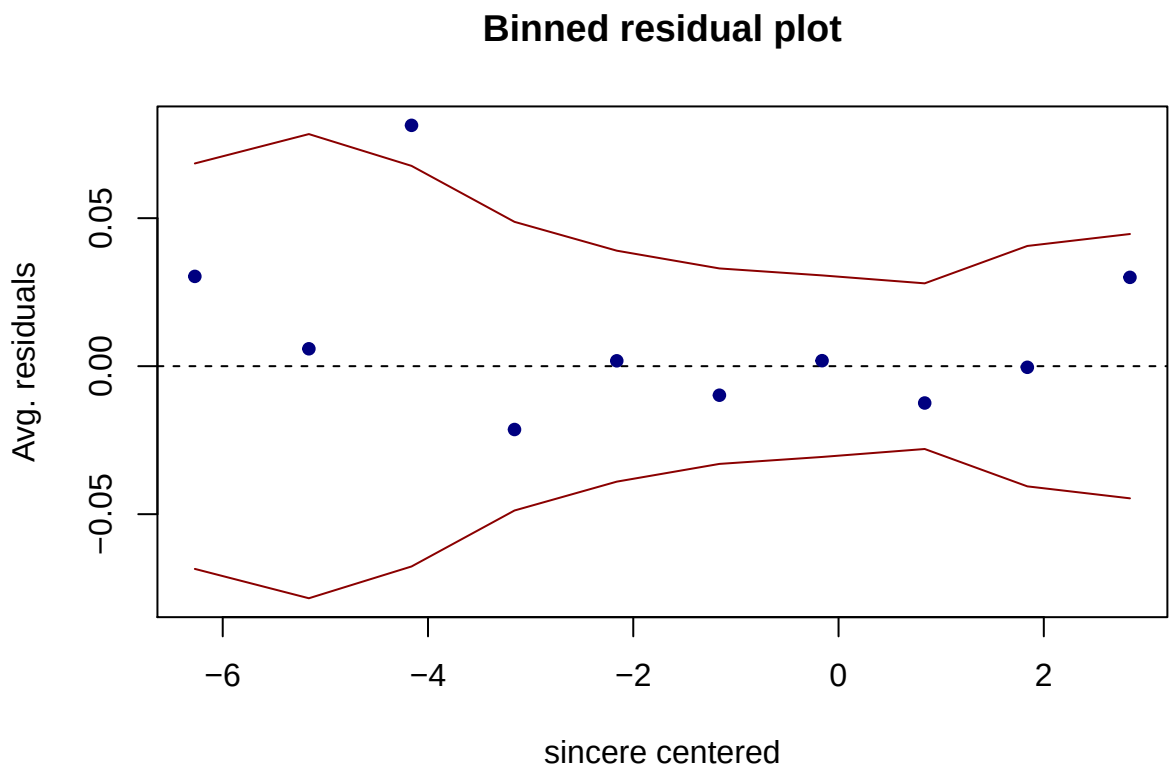
```
##2. For female
rawresid2 <- residuals(dec_f,"resp")
#binned residual plots
binnedplot(x=fitted(dec_f),y=rawresid2,xlab="Pred. probabilities",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



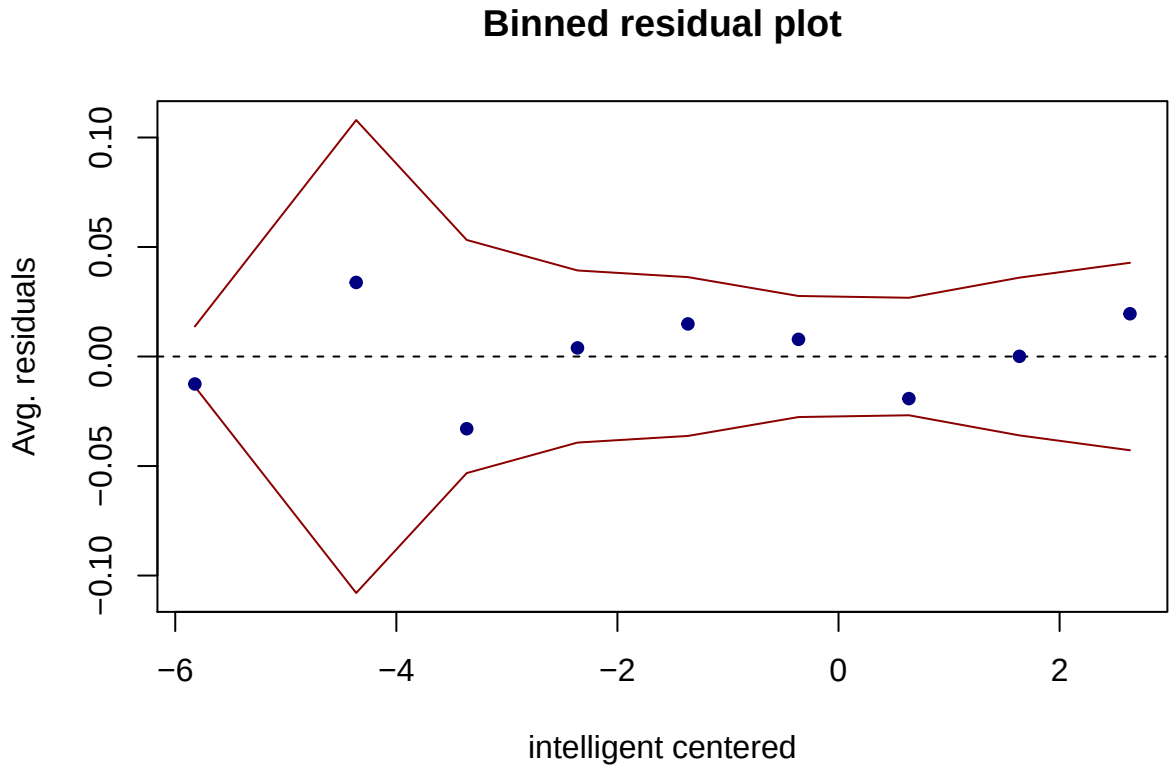

```
binnedplot(n1_f$attr_c,y=rawresid2,xlab="Attractive centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



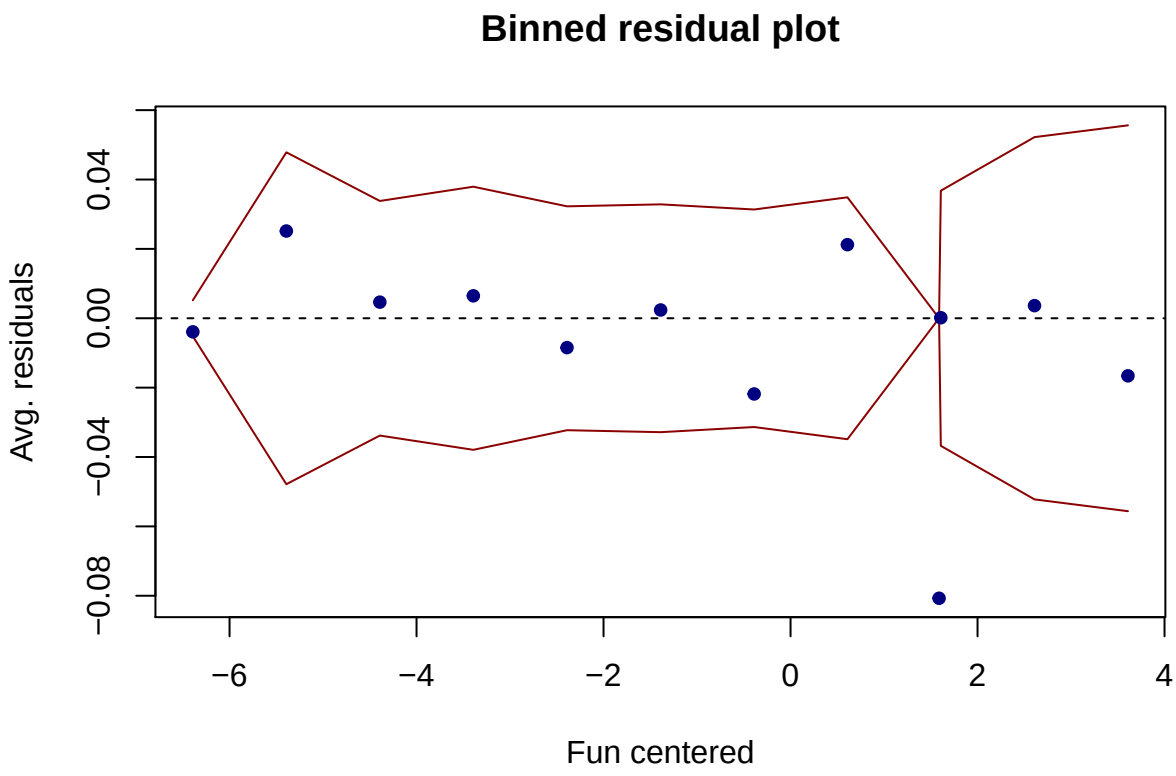

```
binnedplot(n1_f$sinc_c,y=rawresid2,xlab="sincere centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



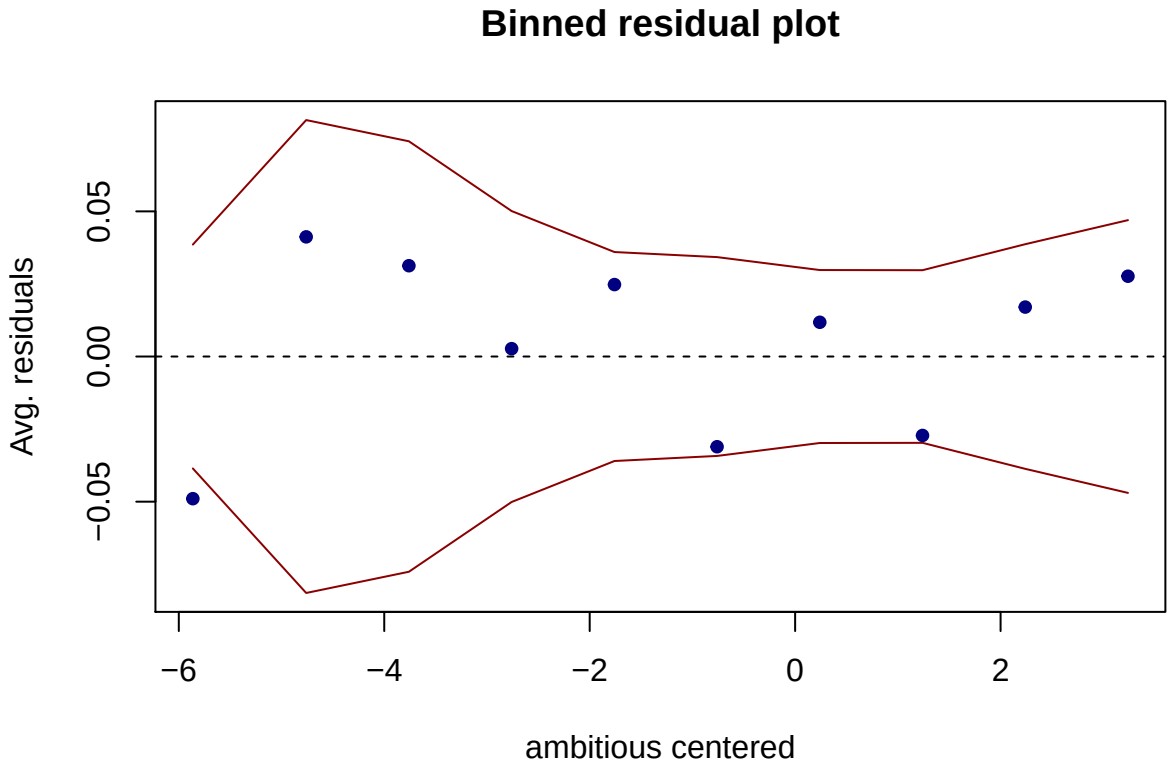

```
binnedplot(n1_f$intel_c,y=rawresid2,xlab="intelligent centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



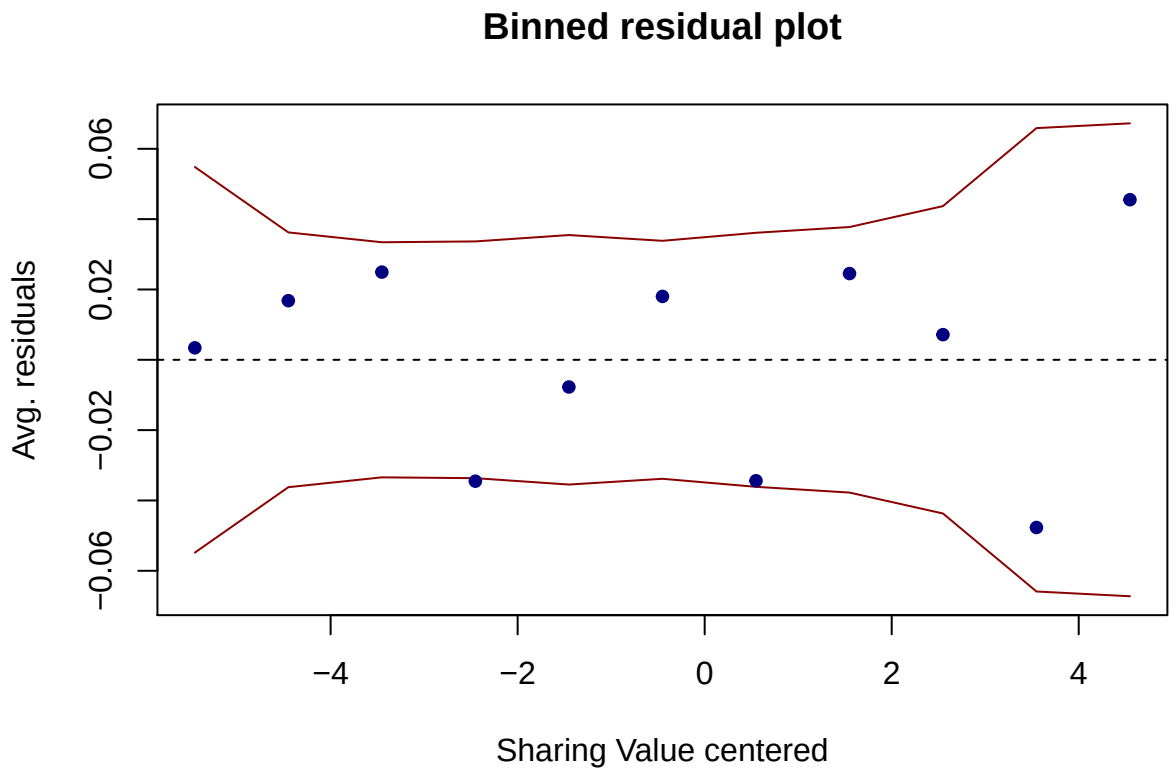

```
binnedplot(n1_f$fun_c,y=rawresid2,xlab="Fun centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



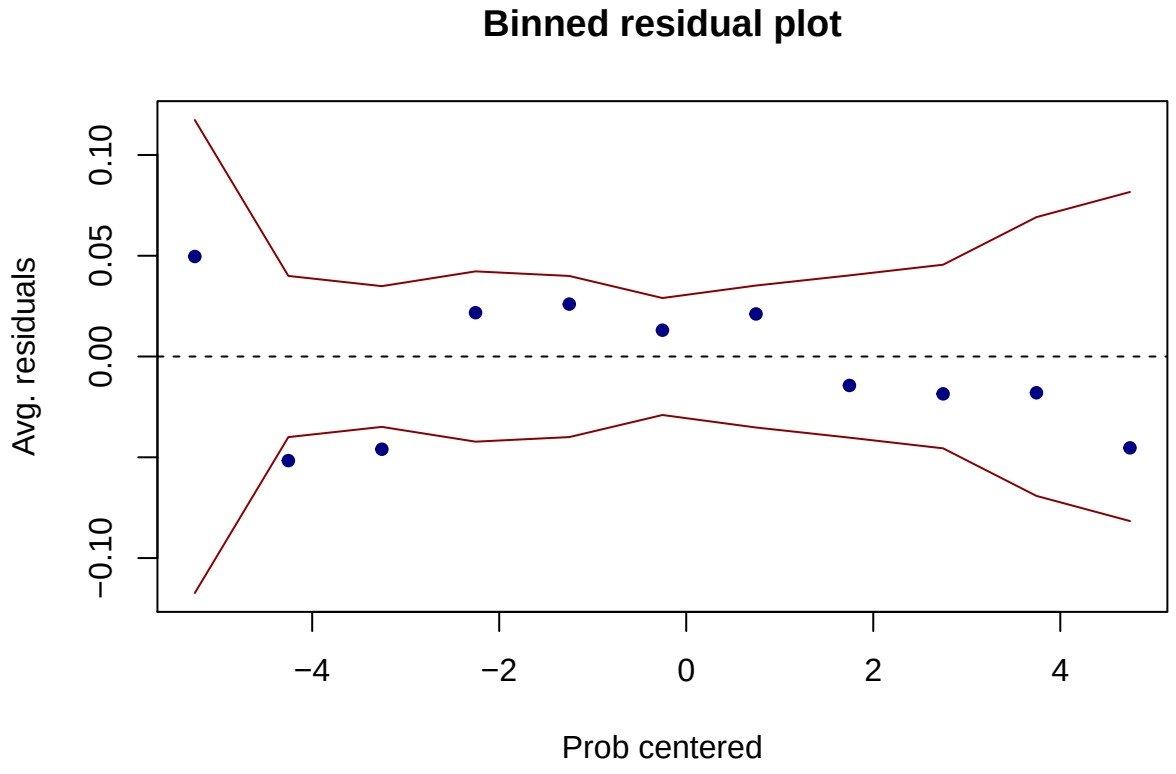

```
binnedplot(n1_f$amb_c,y=rawresid2,xlab="ambitious centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



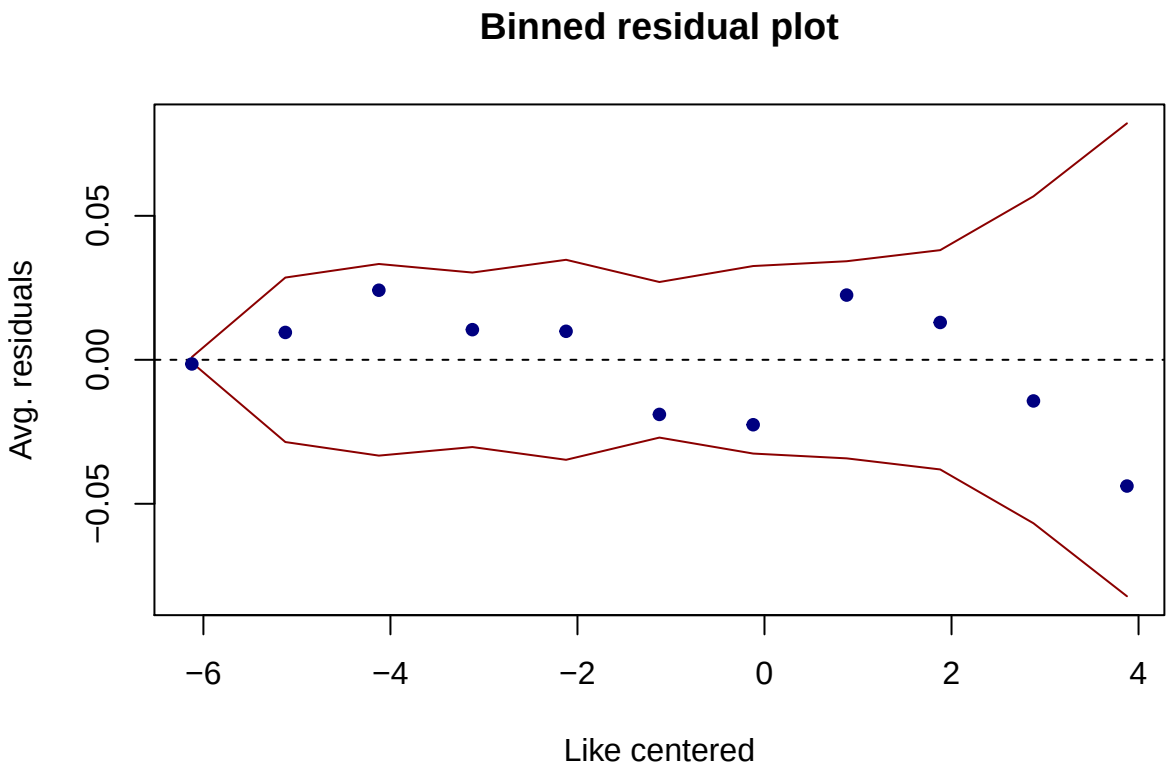

```
binnedplot(n1_f$shar_c,y=rawresid2,xlab="Sharing Value centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```




```
binnedplot(n1_f$prob_c,y=rawresid2,xlab="Prob centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```




```
binnedplot(n1_f$like_c,y=rawresid2,xlab="Like centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



Appendix 10 (Second model for the first question)

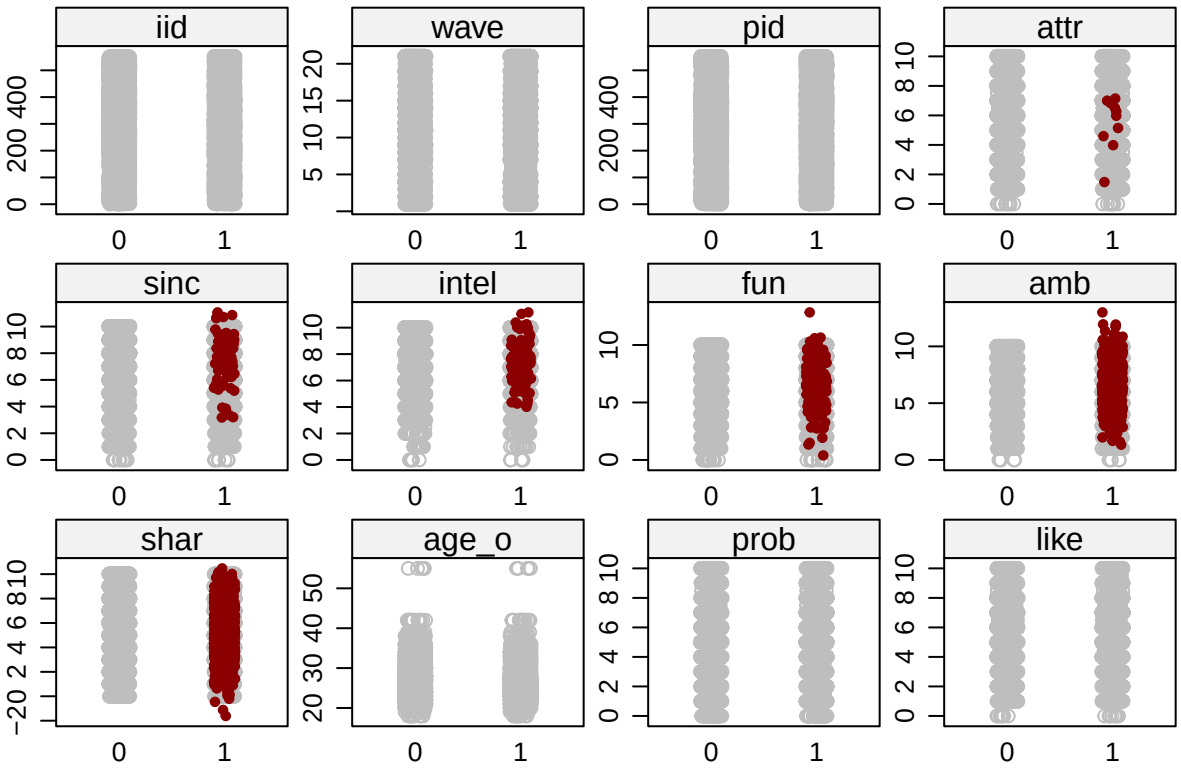
Since the first method loses 12% of data, I try to impute every variable instead and to see which model is better. The imputatation I used is 'norm' because imputed values keeps observed values rathers than 'pmm' way.

In order to avoid multi-collinearity, I made mean-centering for attr, sinc, intel, fun, amb, shar, prob, like. Now I divded this dataset by two : male (3,981 observations) and female(3,978 observations), which are more data than 1-1 model.

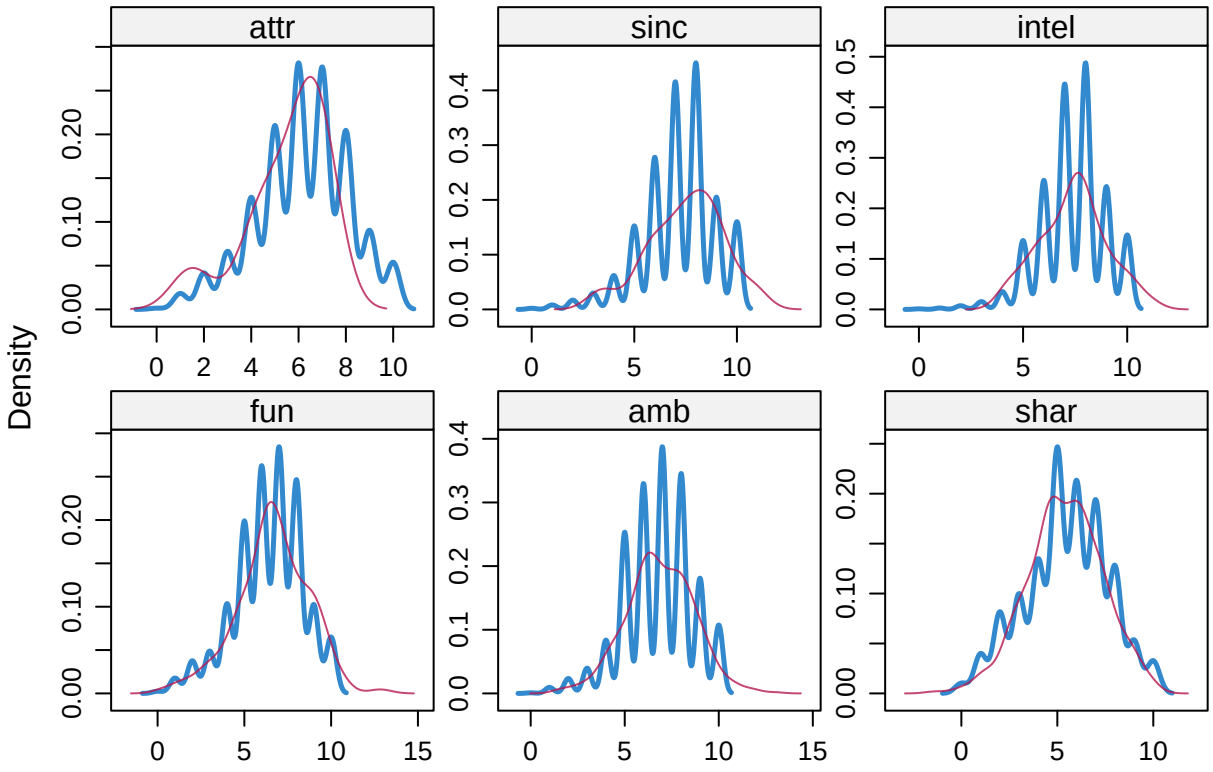
When we compare 1-2 model with 1-1 model, 1-2 model has bigger AIC and BIC in both male and female model. Thus, I decide to use 1-1 model to explain better.

- Imputation for every variable without removal

```
stripplot(dating3_imp, col=c("grey","darkred"),pch=c(1,20))
```




```
#pmm looks better
densityplot(dating3_imp)
```




```
summary(d2)
```

```
##       iid              wave            pid             dec       gender
```

```
##  Min.   :  1.0   Min.   : 1.00   Min.   :  1.0   0:4542   0:3978
##  1st Qu.:154.0   1st Qu.: 7.00   1st Qu.:153.0   1:3417   1:3981
##  Median :278.0   Median :11.00   Median :278.0
##  Mean   :281.5   Mean   :11.27   Mean   :281.5
##  3rd Qu.:406.0   3rd Qu.:15.00   3rd Qu.:406.0
##  Max.   :552.0   Max.   :21.00   Max.   :552.0
##       attr            sinc            intel            fun
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 5.000   1st Qu.: 6.000   1st Qu.: 6.000   1st Qu.: 5.000
##  Median : 6.000   Median : 7.000   Median : 7.000   Median : 7.000
##  Mean   : 6.177   Mean   : 7.175   Mean   : 7.371   Mean   : 6.398
##  3rd Qu.: 8.000   3rd Qu.: 8.000   3rd Qu.: 8.000   3rd Qu.: 8.000
##  Max.   :10.000   Max.   :11.060   Max.   :11.151   Max.   :12.836
##       amb             shar        samerace     age_o
##  Min.   : 0.000   Min.   :-1.610   0:4775   Min.   :18.00
##  1st Qu.: 6.000   1st Qu.: 4.000   1:3184   1st Qu.:24.00
##  Median : 7.000   Median : 5.589            Median :26.00
##  Mean   : 6.782   Mean   : 5.459            Mean   :26.35
##  3rd Qu.: 8.000   3rd Qu.: 7.000            3rd Qu.:28.00
##  Max.   :12.992   Max.   :10.471            Max.   :55.00
##       prob            like           attr_c            sinc_c
##  Min.   : 0.000   Min.   : 0.000   Min.   :-6.1766   Min.   :-7.1748
##  1st Qu.: 4.000   1st Qu.: 5.000   1st Qu.:-1.1766   1st Qu.:-1.1748
##  Median : 5.000   Median : 6.000   Median :-0.1766   Median :-0.1748
##  Mean   : 5.206   Mean   : 6.127   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 7.000   3rd Qu.: 7.000   3rd Qu.: 1.8234   3rd Qu.: 0.8252
##  Max.   :10.000   Max.   :10.000   Max.   : 3.8234   Max.   : 3.8849
##     intel_c            fun_c             amb_c             shar_c
##  Min.   :-7.3713   Min.   :-6.3983   Min.   :-6.7816   Min.   :-7.0691
##  1st Qu.:-1.3713   1st Qu.:-1.3983   1st Qu.:-0.7816   1st Qu.:-1.4587
##  Median :-0.3713   Median : 0.6017   Median : 0.2184   Median : 0.1302
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.6287   3rd Qu.: 1.6017   3rd Qu.: 1.2184   3rd Qu.: 1.5413
##  Max.   : 3.7798   Max.   : 6.4373   Max.   : 6.2108   Max.   : 5.0127
##      prob_c            like_c
##  Min.   :-5.2059   Min.   :-6.1271
##  1st Qu.:-1.2059   1st Qu.:-1.1271
##  Median :-0.2059   Median :-0.1271
##  Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 1.7941   3rd Qu.: 0.8729
##  Max.   : 4.7941   Max.   : 3.8729
```

Appendix 11 (1-2 model has bigger AIC and BIC compared to 1-1 model)

```
# For male
# same model with above (AIC : 3546, bic : 3621)
dec2_m <- glmer(dec~attr_c+sinc_c+fun_c+amb_c+intel_c*shar_c  +samerace + prob_c+like_c + (1|wave), fami
summary(dec2_m)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
```

```
## Formula:
## dec ~ attr_c + sinc_c + fun_c + amb_c + intel_c * shar_c + samerace +
##     prob_c + like_c + (1 | wave)
##    Data: d2_m
##
##      AIC      BIC   logLik deviance df.resid
##   3549.8   3625.3  -1762.9   3525.8     3969
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -10.9848  -0.5139  -0.0628   0.5328   9.6764
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  wave   (Intercept) 0.1815   0.426
## Number of obs: 3981, groups:  wave, 21
##
## Fixed effects:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.37942    0.11485  -3.304 0.000955 ***
## attr_c          0.57722    0.03559  16.220  < 2e-16 ***
## sinc_c         -0.27391    0.03899  -7.024 2.15e-12 ***
## fun_c           0.14270    0.03637   3.924 8.71e-05 ***
## amb_c          -0.14810    0.03493  -4.240 2.23e-05 ***
## intel_c        -0.08032    0.04487  -1.790 0.073458 .
## shar_c          0.08113    0.02883   2.814 0.004893 **
## samerace1      -0.17490    0.08766  -1.995 0.046024 *
## prob_c          0.23202    0.02570   9.027  < 2e-16 ***
## like_c          0.63919    0.04616  13.847  < 2e-16 ***
## intel_c:shar_c -0.06115    0.01574  -3.885 0.000102 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) attr_c sinc_c fun_c  amb_c  intl_c shar_c samrc1 prob_c
## attr_c      -0.069
## sinc_c      -0.021 -0.074
## fun_c       -0.038 -0.115 -0.133
## amb_c        0.071 -0.105  0.029 -0.158
## intel_c      0.036 -0.031 -0.438 -0.114 -0.353
## shar_c      -0.025  0.056 -0.015 -0.185 -0.191  0.027
## samerace1   -0.295 -0.062  0.059  0.004  0.036  0.004 -0.014
## prob_c       0.006  0.196 -0.121 -0.022 -0.035  0.041 -0.204 -0.082
## like_c      -0.063 -0.237 -0.186 -0.184 -0.031 -0.077 -0.218  0.013 -0.119
## intl_c:shr_ -0.082 -0.046  0.056 -0.058 -0.029  0.048 -0.003  0.017 -0.004
##             like_c
## attr_c
## sinc_c
## fun_c
## amb_c
## intel_c
## shar_c
## samerace1
## prob_c
```

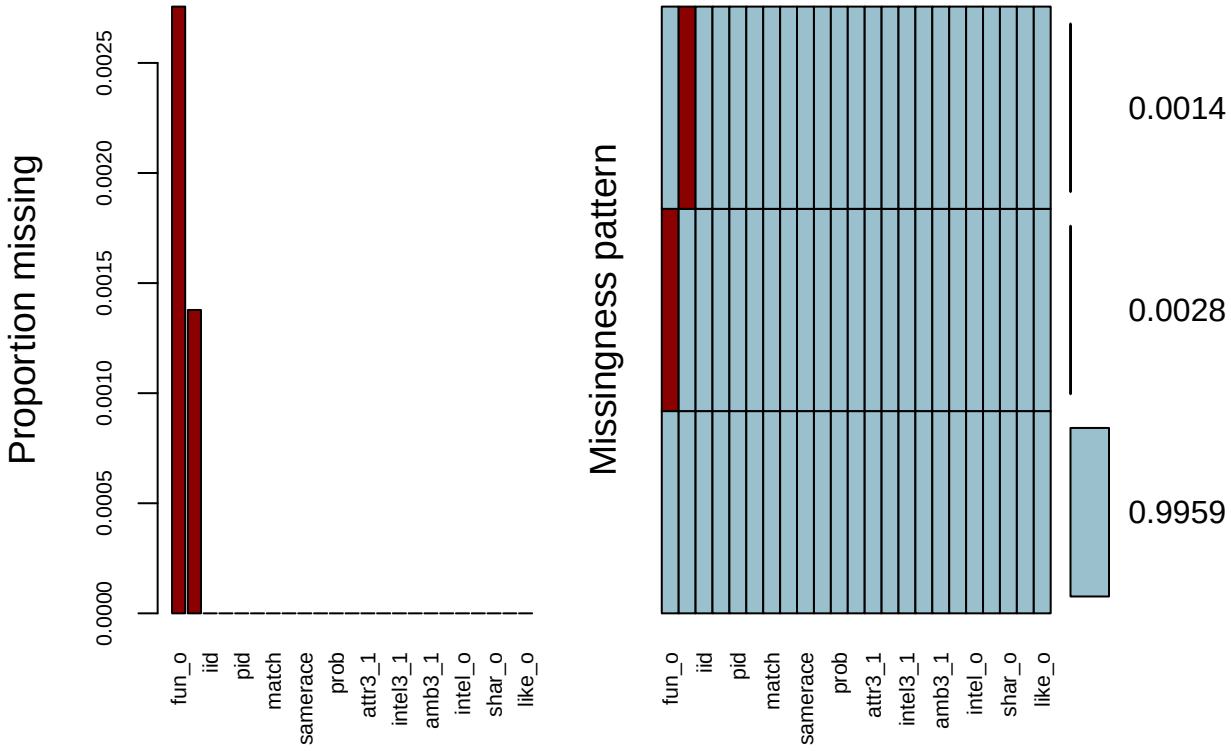
```
## like_c
## intl_c:shr_ -0.078
```

```
#For female (AIC : 3736, BIC : 3803)

dec2_f <- glmer(dec~ attr_c+sinc_c+fun_c+amb_c+intel_c*shar_c+ prob_c+like_c+(1|wave), family=binomial(
summary(dec2_f)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula:
## dec ~ attr_c + sinc_c + fun_c + amb_c + intel_c * shar_c + prob_c +
##     like_c + (1 | wave)
##    Data: d2_f
##
##      AIC      BIC   logLik deviance df.resid
##   3730.9   3800.1  -1854.5   3708.9     3967
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.0673 -0.5690 -0.2043  0.5858 15.3379
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  wave   (Intercept) 0.246    0.496
## Number of obs: 3978, groups:  wave, 21
##
## Fixed effects:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.63126    0.12222  -5.165 2.40e-07 ***
## attr_c          0.34840    0.03002  11.605  < 2e-16 ***
## sinc_c         -0.15231    0.03489  -4.365 1.27e-05 ***
## fun_c           0.15734    0.03324   4.734 2.20e-06 ***
## amb_c          -0.17681    0.03361  -5.260 1.44e-07 ***
## intel_c         0.12152    0.04477   2.714  0.00664 **
## shar_c          0.18241    0.02908   6.272 3.56e-10 ***
## prob_c          0.13946    0.02339   5.962 2.49e-09 ***
## like_c          0.44979    0.04183  10.754  < 2e-16 ***
## intel_c:shar_c -0.04224    0.01636  -2.582  0.00984 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) attr_c sinc_c fun_c  amb_c  intl_c shar_c prob_c like_c
## attr_c      -0.006
## sinc_c       0.030 -0.072
## fun_c       -0.018 -0.160 -0.066
## amb_c       -0.017 -0.030  0.004 -0.187
## intel_c     -0.033  0.028 -0.442 -0.058 -0.412
## shar_c      -0.001  0.016 -0.013 -0.175 -0.129  0.080
## prob_c      -0.030  0.137 -0.126 -0.024 -0.039  0.071 -0.162
## like_c      -0.056 -0.247 -0.160 -0.214 -0.049 -0.092 -0.232 -0.119
## intl_c:shr_ -0.053 -0.029  0.047 -0.056  0.049 -0.122 -0.289 -0.038  0.025
```

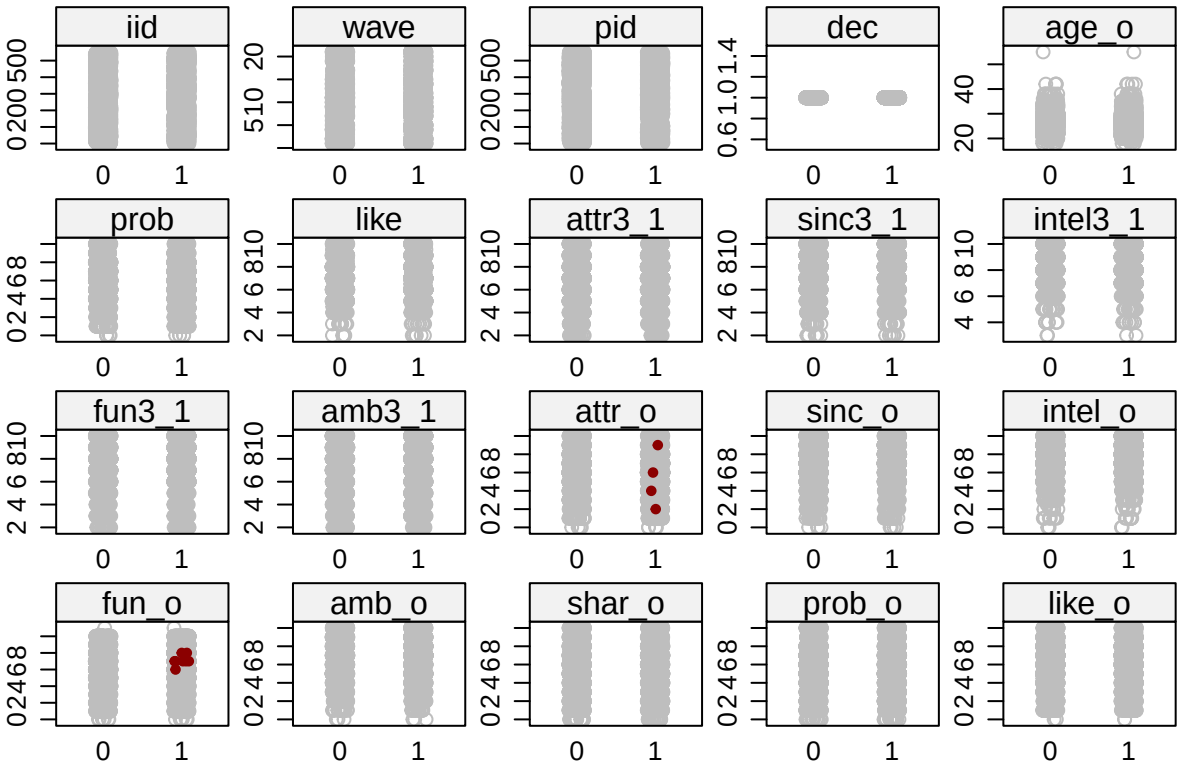
Appendix 12 (imputation for #2 model)

```
aggr(match2,col=c("lightblue3","darkred"),numbers=TRUE,sortVars=TRUE,labels=names(match2),cex.axis=.7,g
```

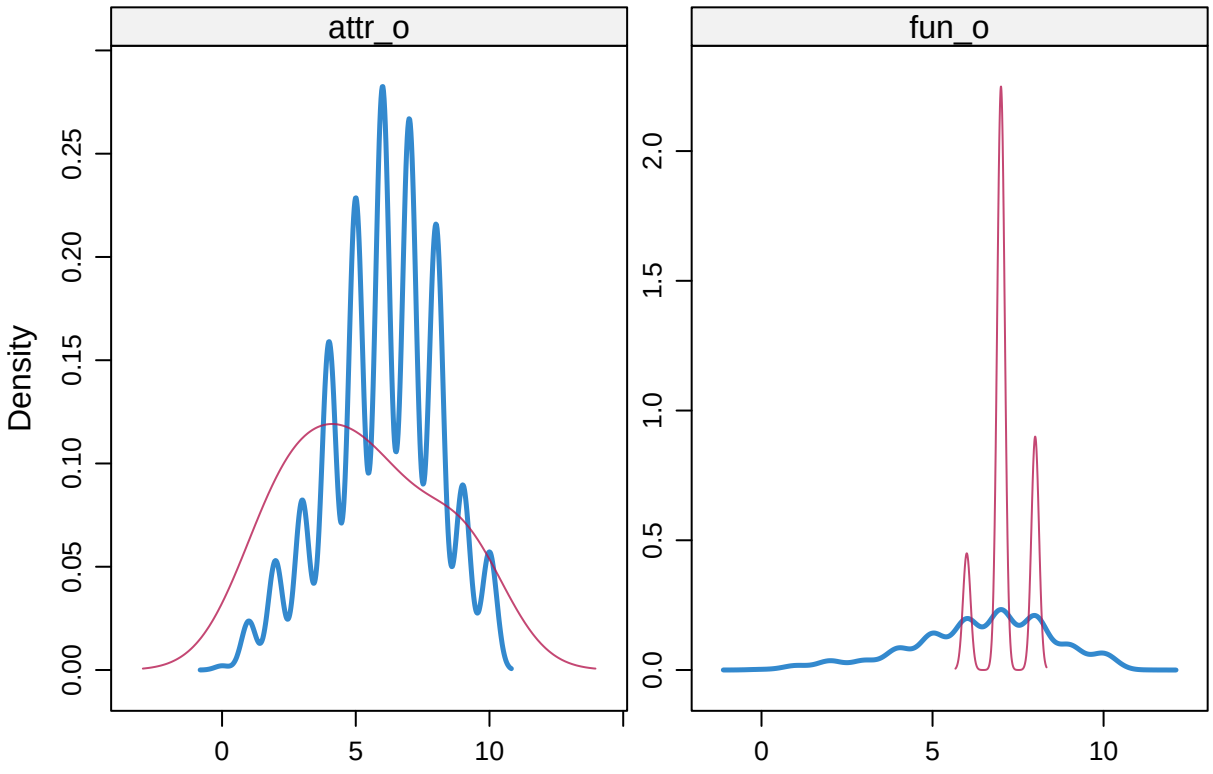



```
##
##  Variables sorted by number of missings:
##  Variable       Count
##     fun_o 0.002755770
##    attr_o 0.001377885
##       iid 0.000000000
##      wave 0.000000000
##       pid 0.000000000
##       dec 0.000000000
##     match 0.000000000
##    gender 0.000000000
##  samerace 0.000000000
##     age_o 0.000000000
##      prob 0.000000000
##      like 0.000000000
##   attr3_1 0.000000000
##   sinc3_1 0.000000000
##  intel3_1 0.000000000
##    fun3_1 0.000000000
##    amb3_1 0.000000000
##    sinc_o 0.000000000
##   intel_o 0.000000000
##     amb_o 0.000000000
##    shar_o 0.000000000
##    prob_o 0.000000000
##    like_o 0.000000000
```

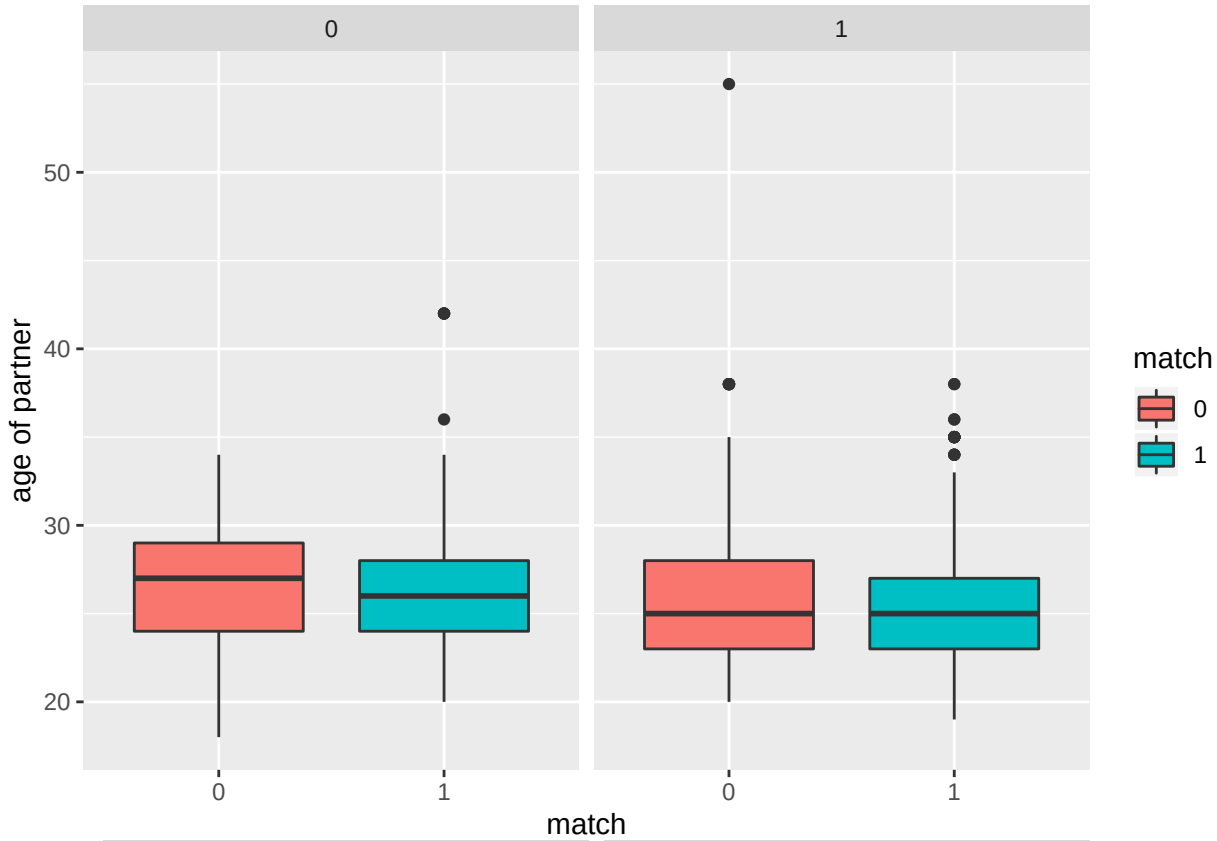
```
stripplot(match_imp, col=c("grey","darkred"),pch=c(1,20))
```

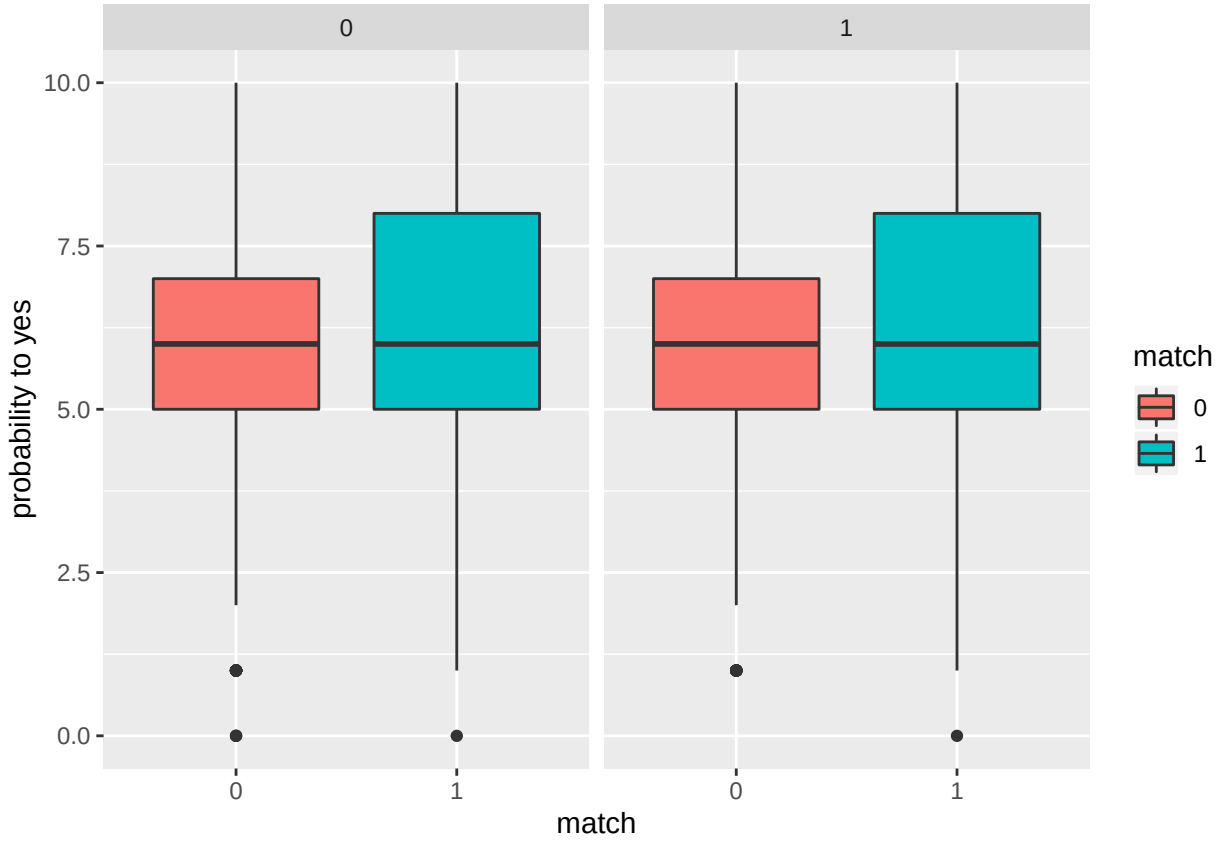


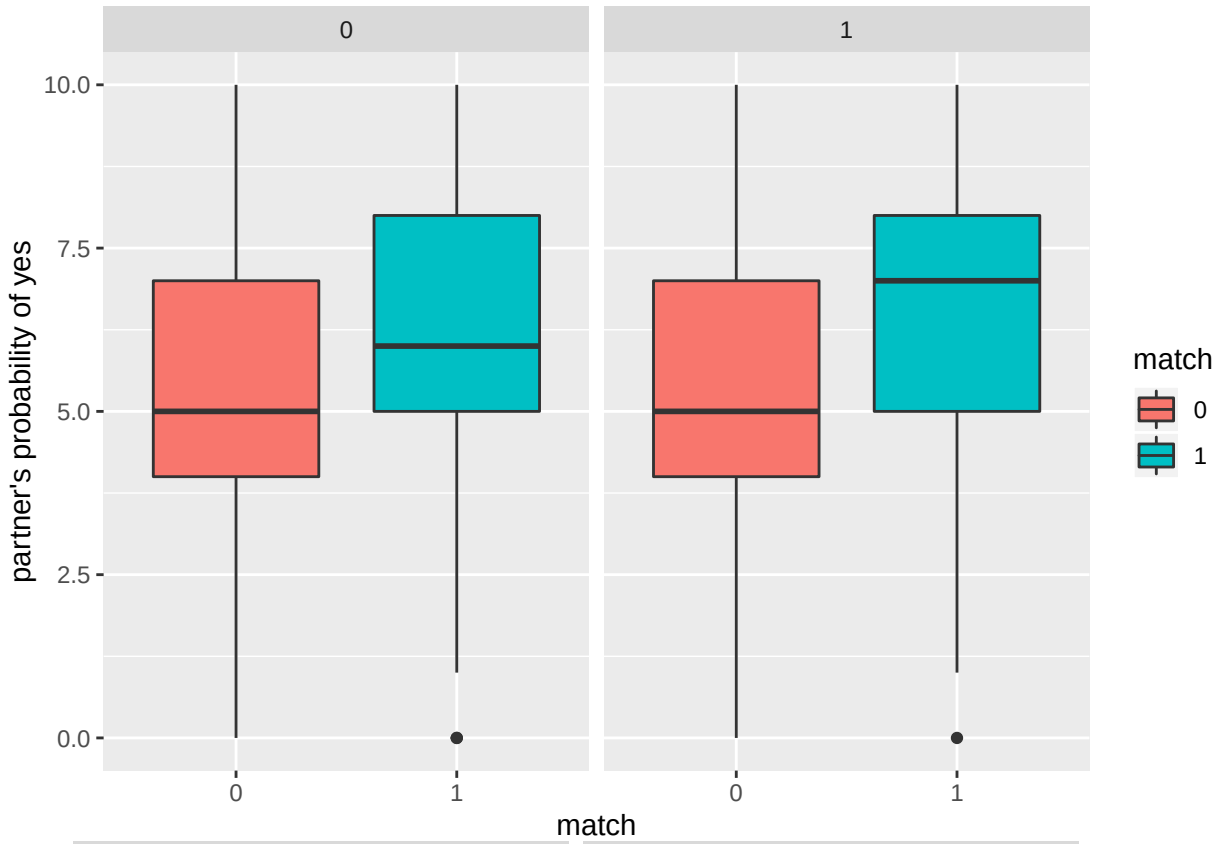

```
densityplot(match_imp)
```

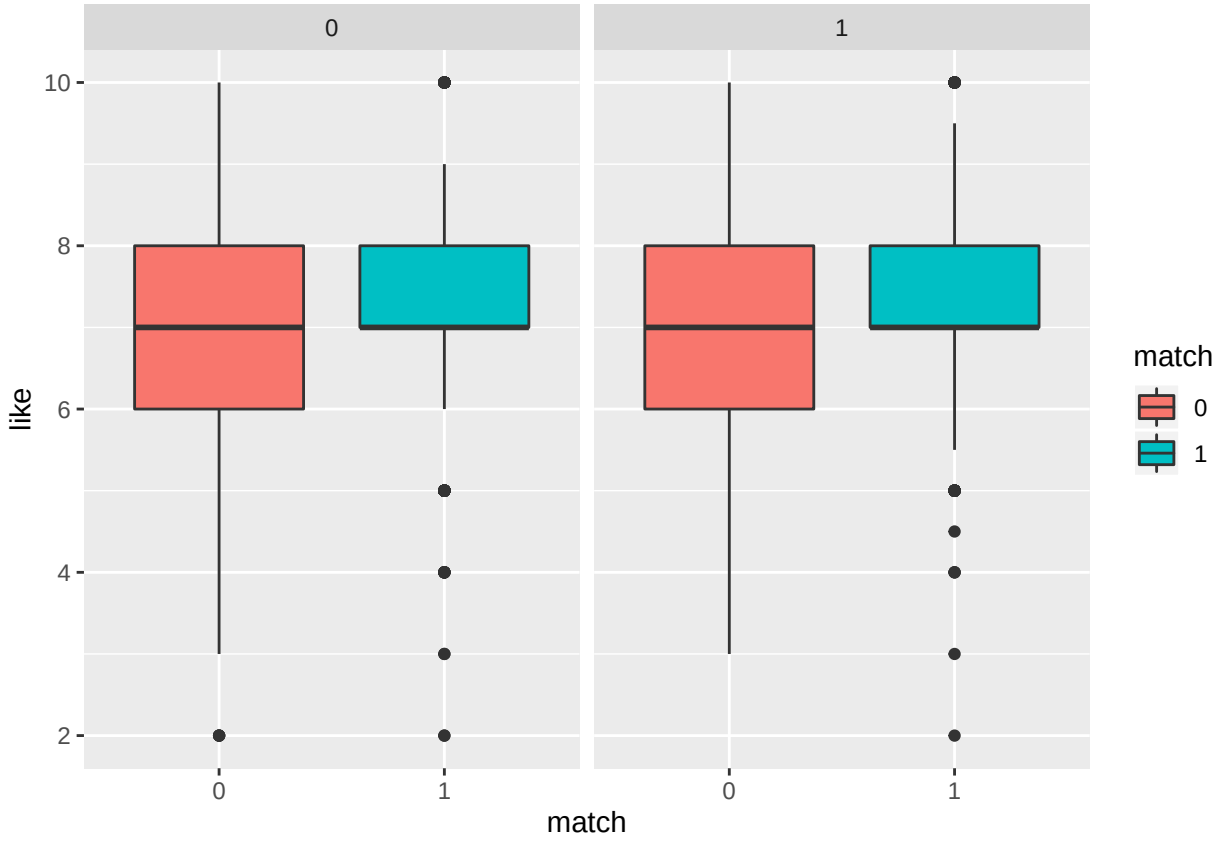


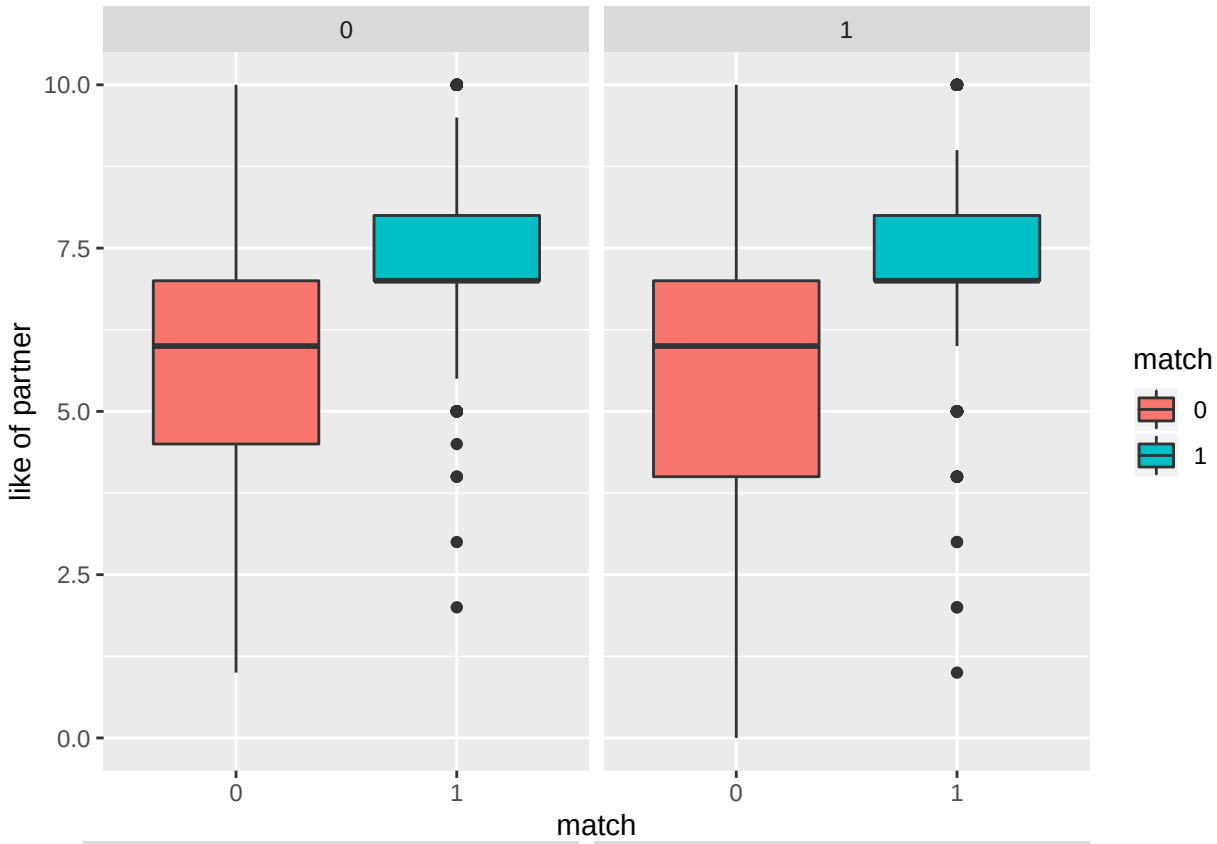
Appendix 13 (EDA for #2 model)

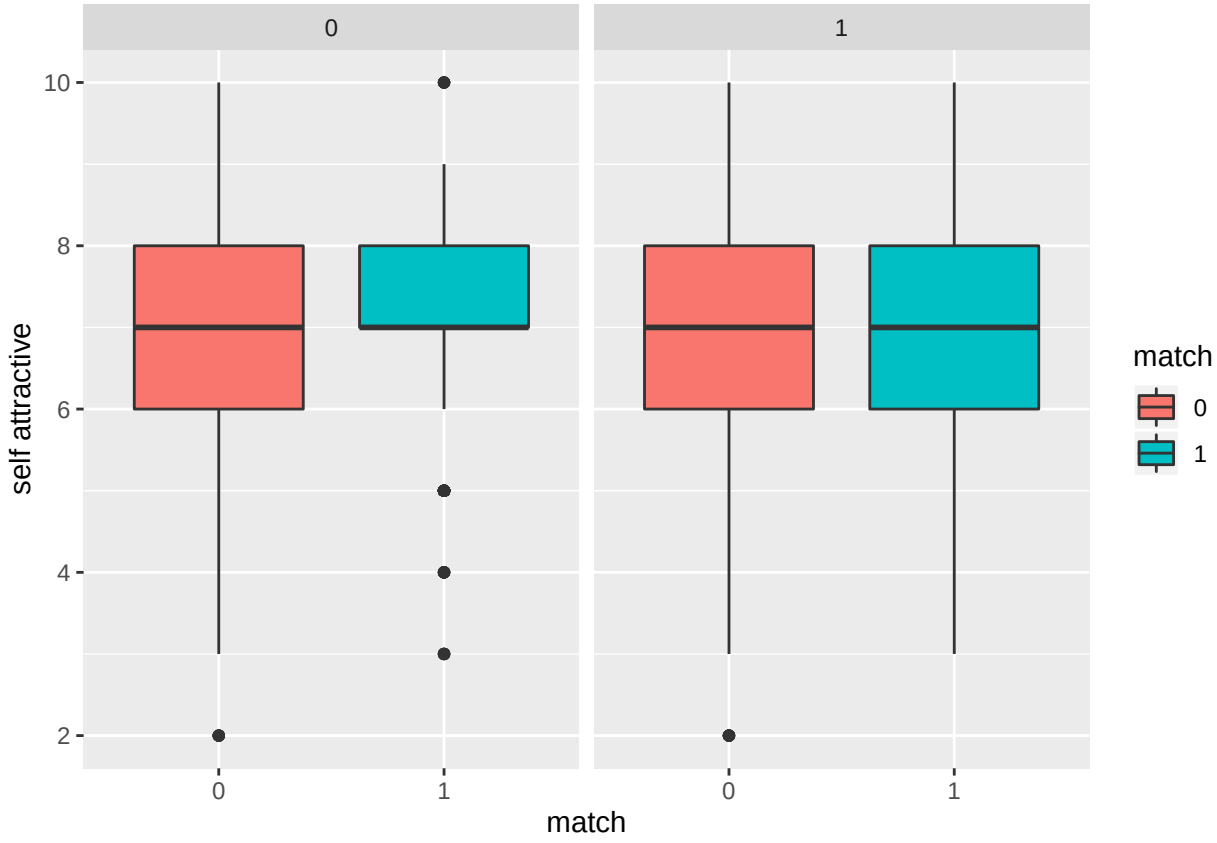


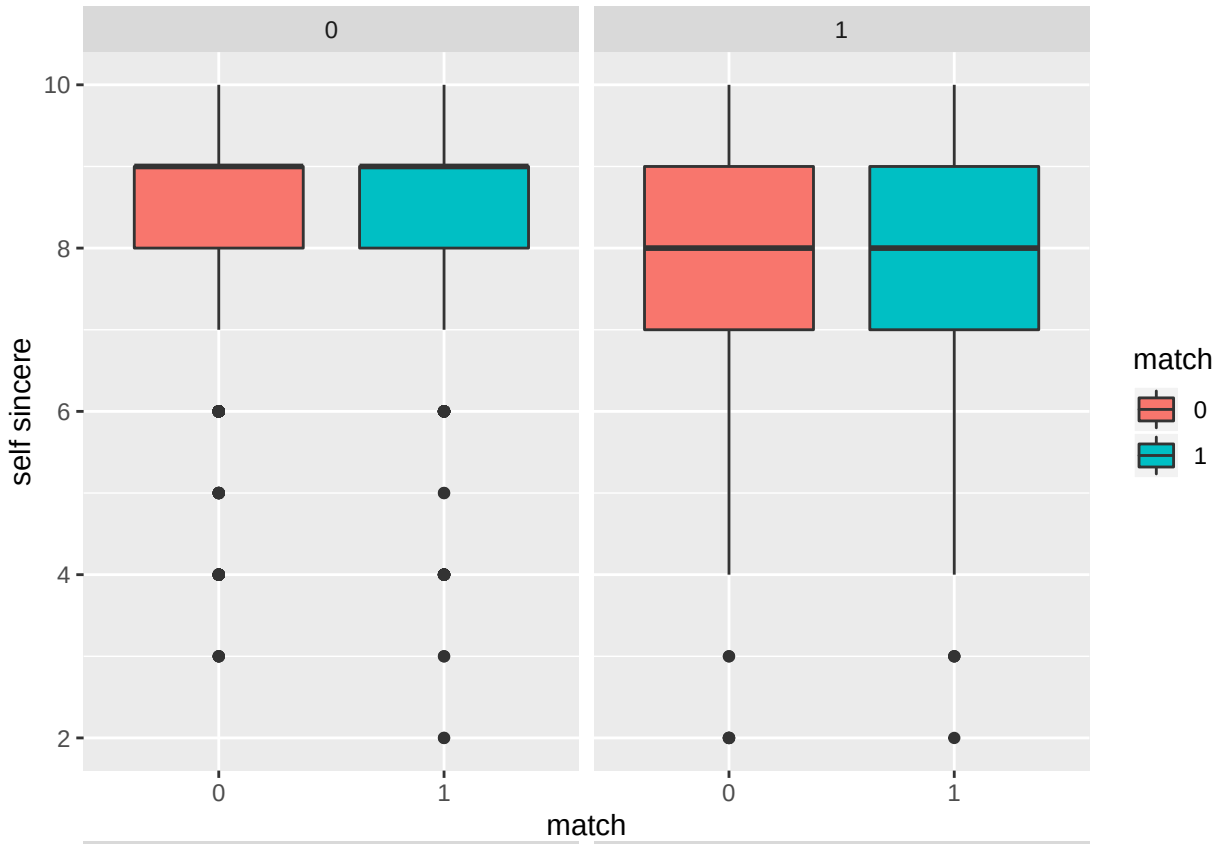


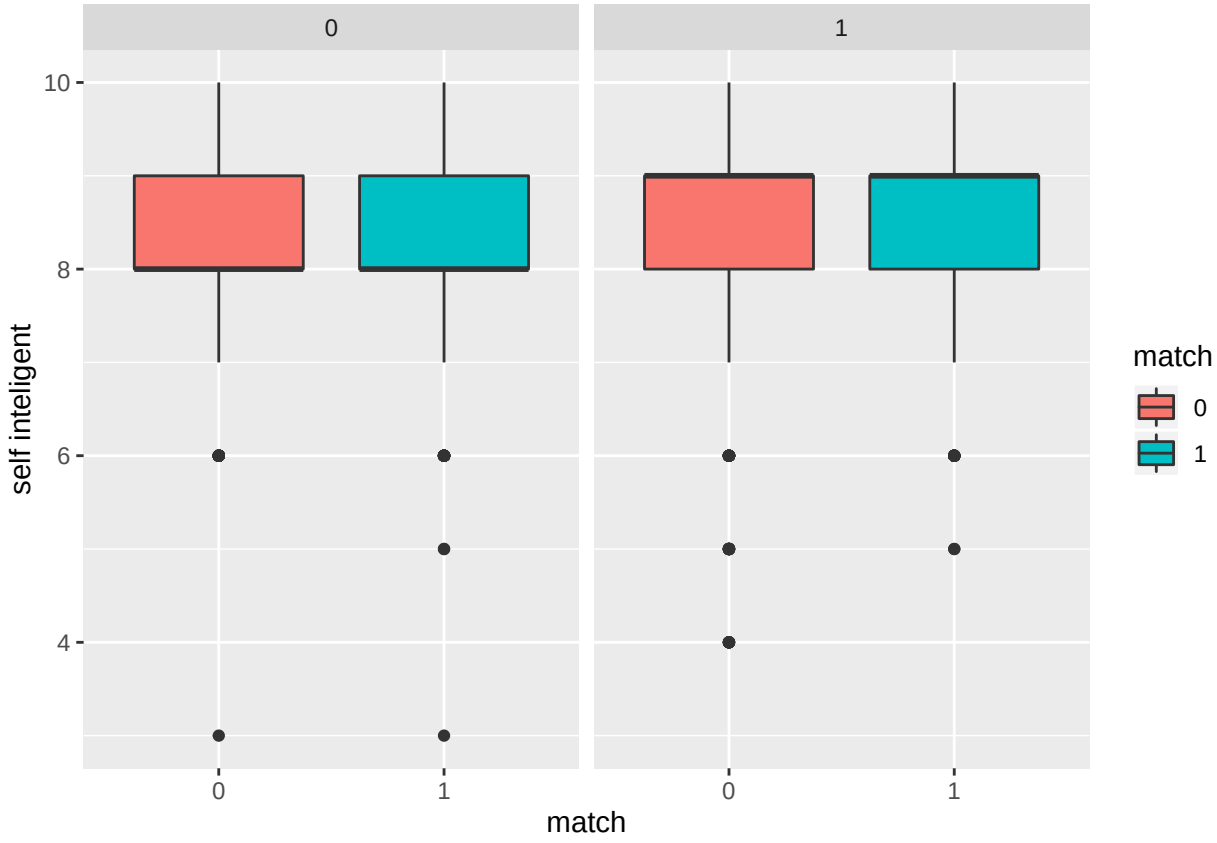

0
1
10.0
7.5
5.0
2.5
0.0
partner's probability of yes
0
1
0
1
match
match
0
1

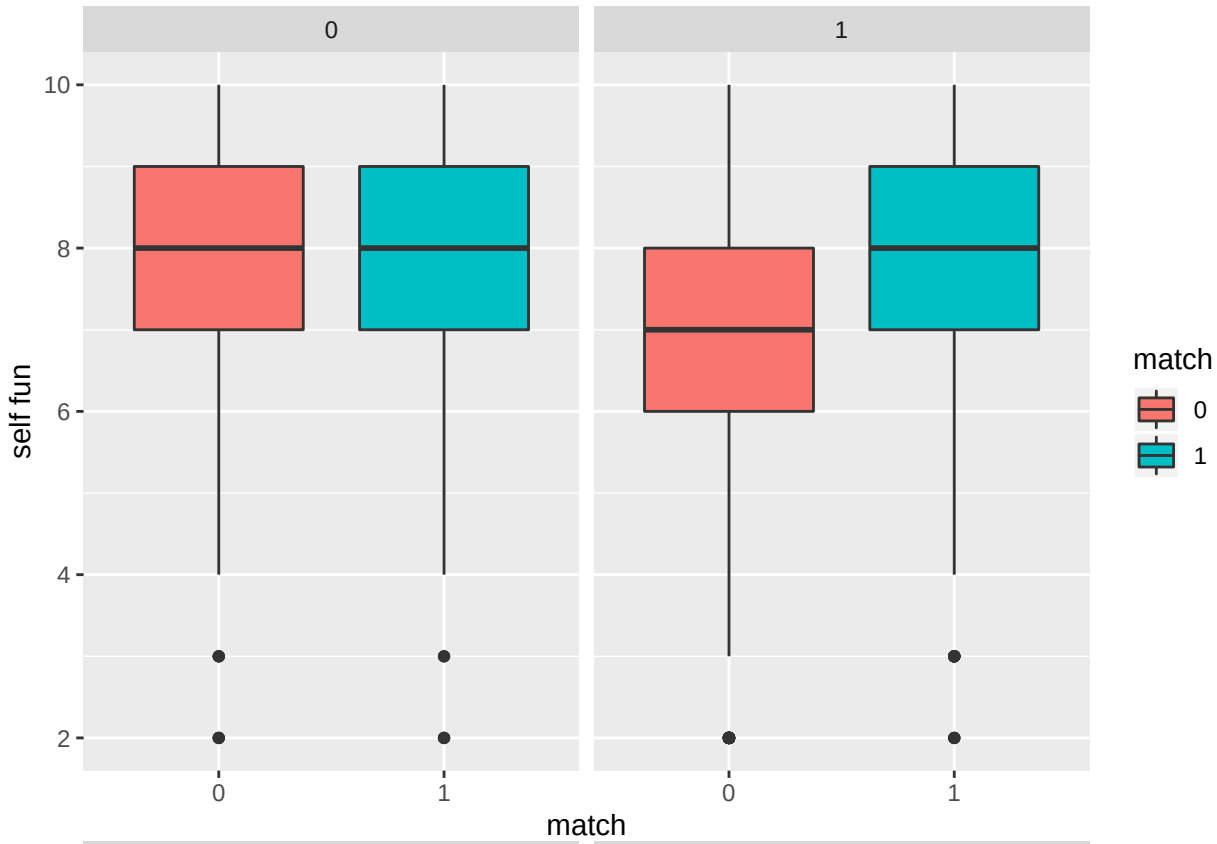

0
1
10
8
6
4
2
like
0
1
0
1
match
match
0
1

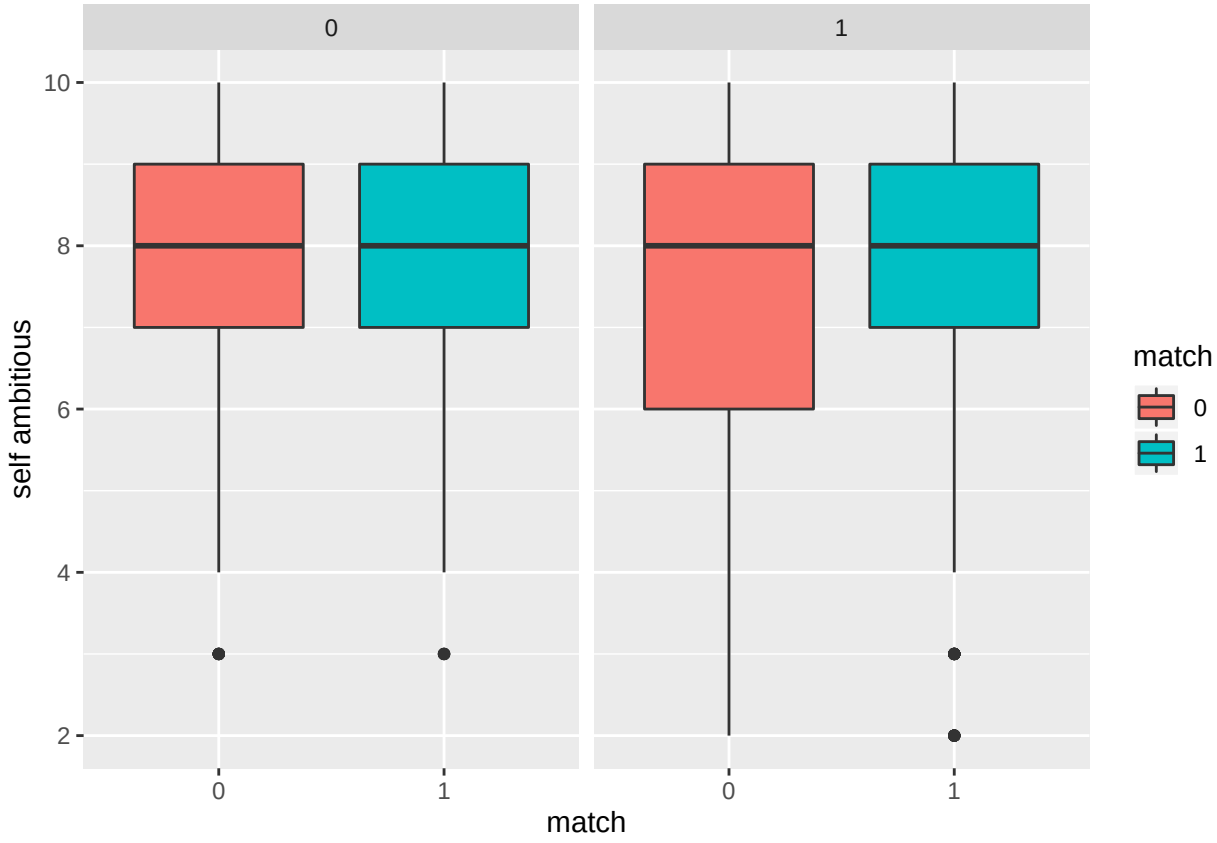

0
1
10.0
7.5
5.0
2.5
0.0
like of partner
0
1
match
match
0
1

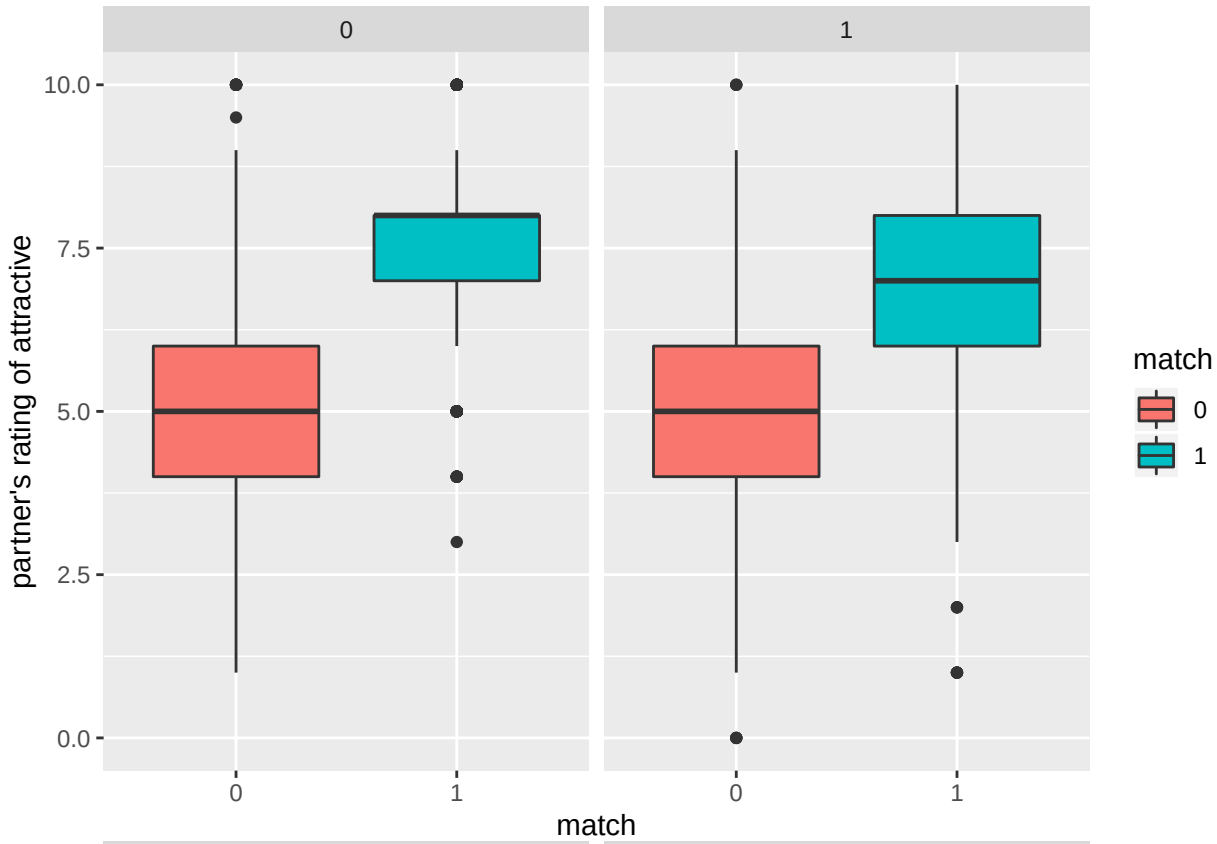

0
1
10
8
6
4
2
self attractive
0
1
match
match
0
1

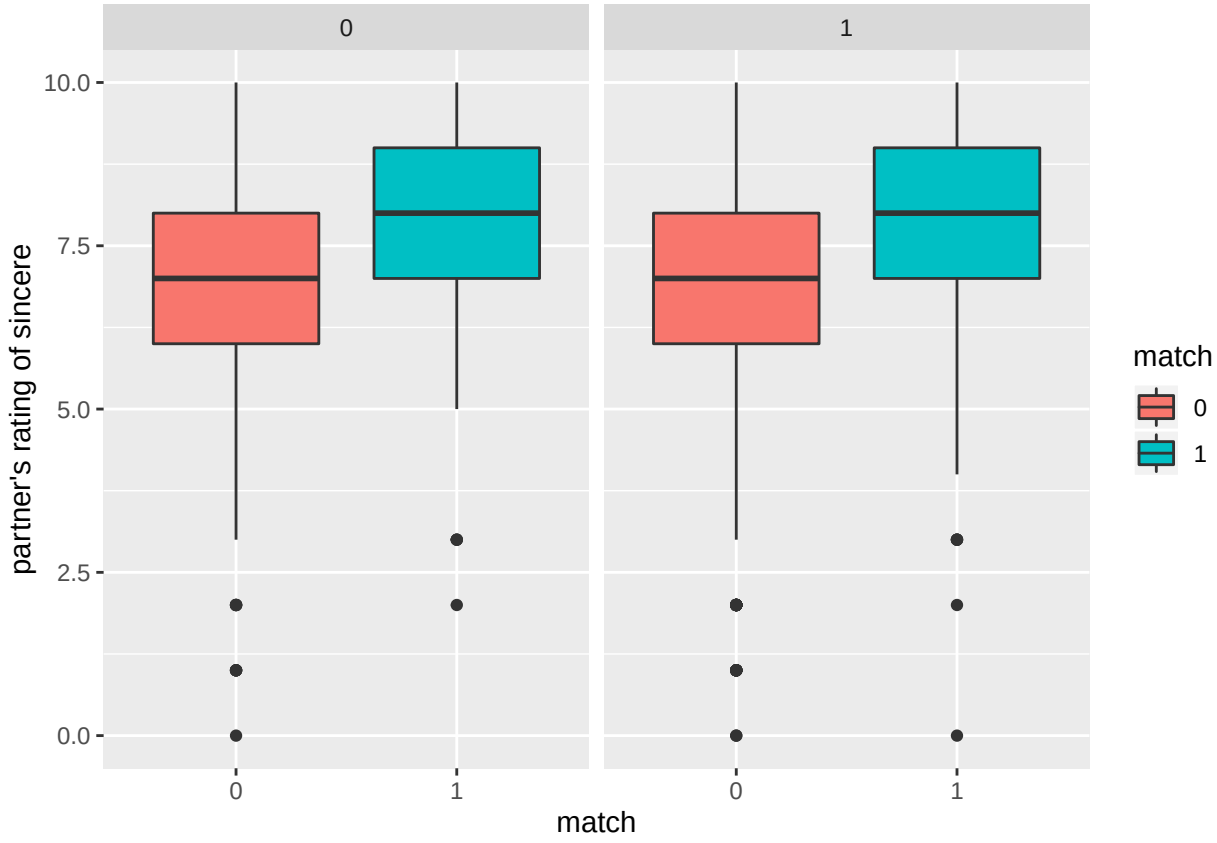

0
1
10
8
6
4
2
self sincere
0
1
0
1
match
match
0
1

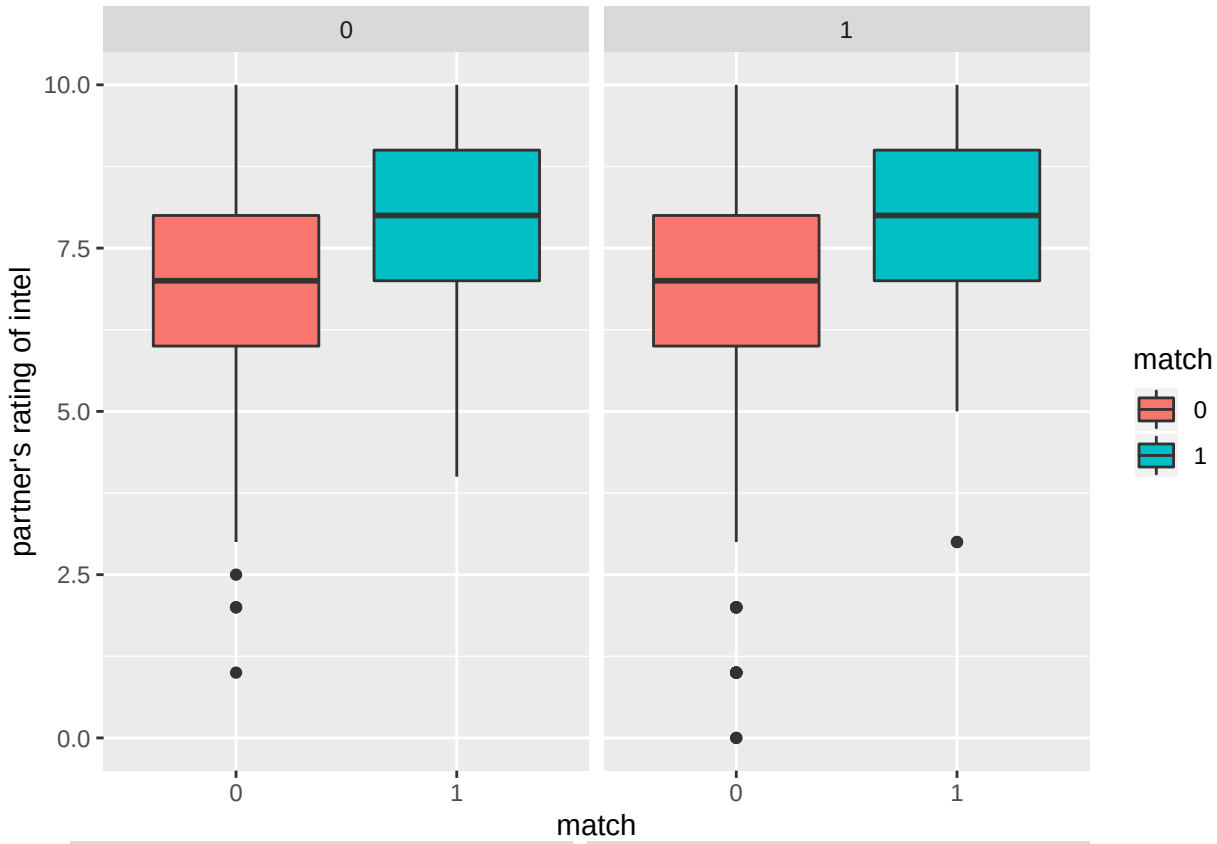

0
1
10
8
6
4
self inteligent
0
1
0
1
match
match
0
1

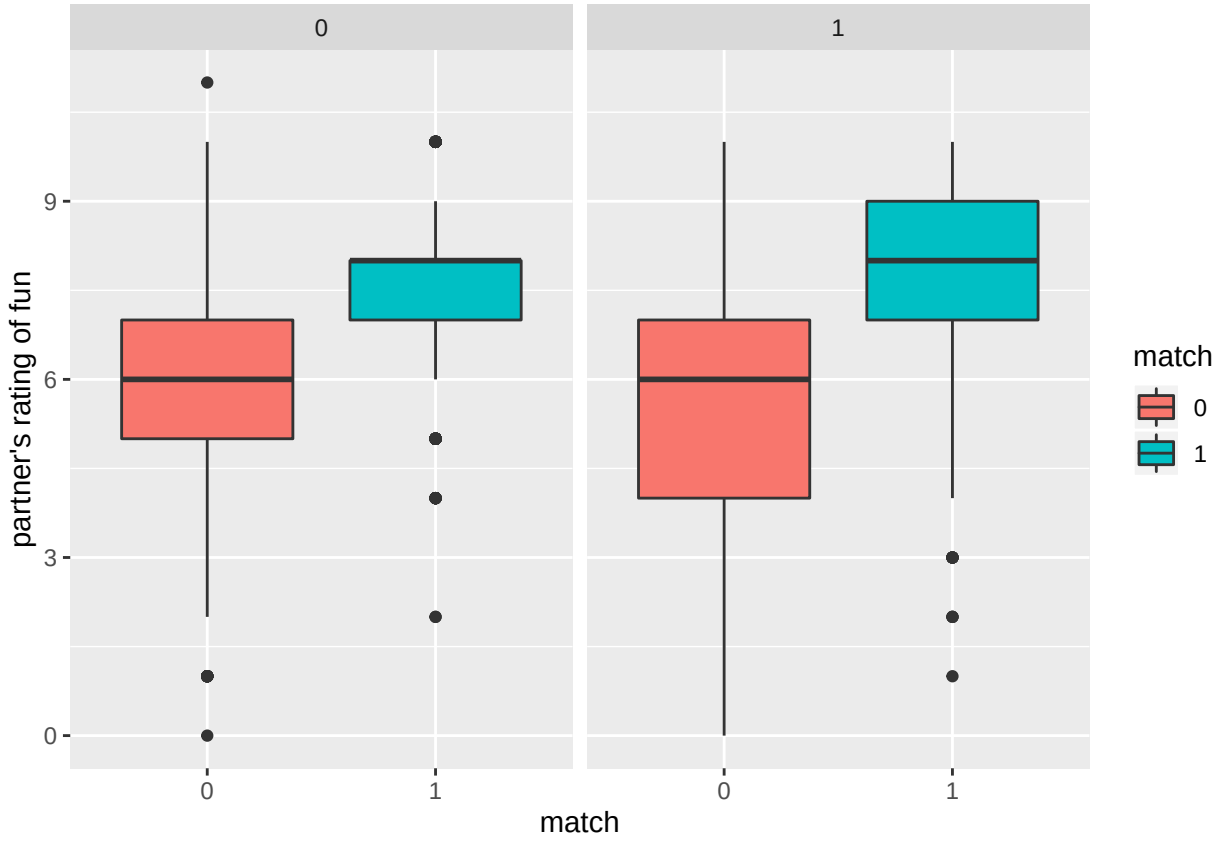

0
1
10
8
6
4
2
self fun
0
1
match
match
0
1

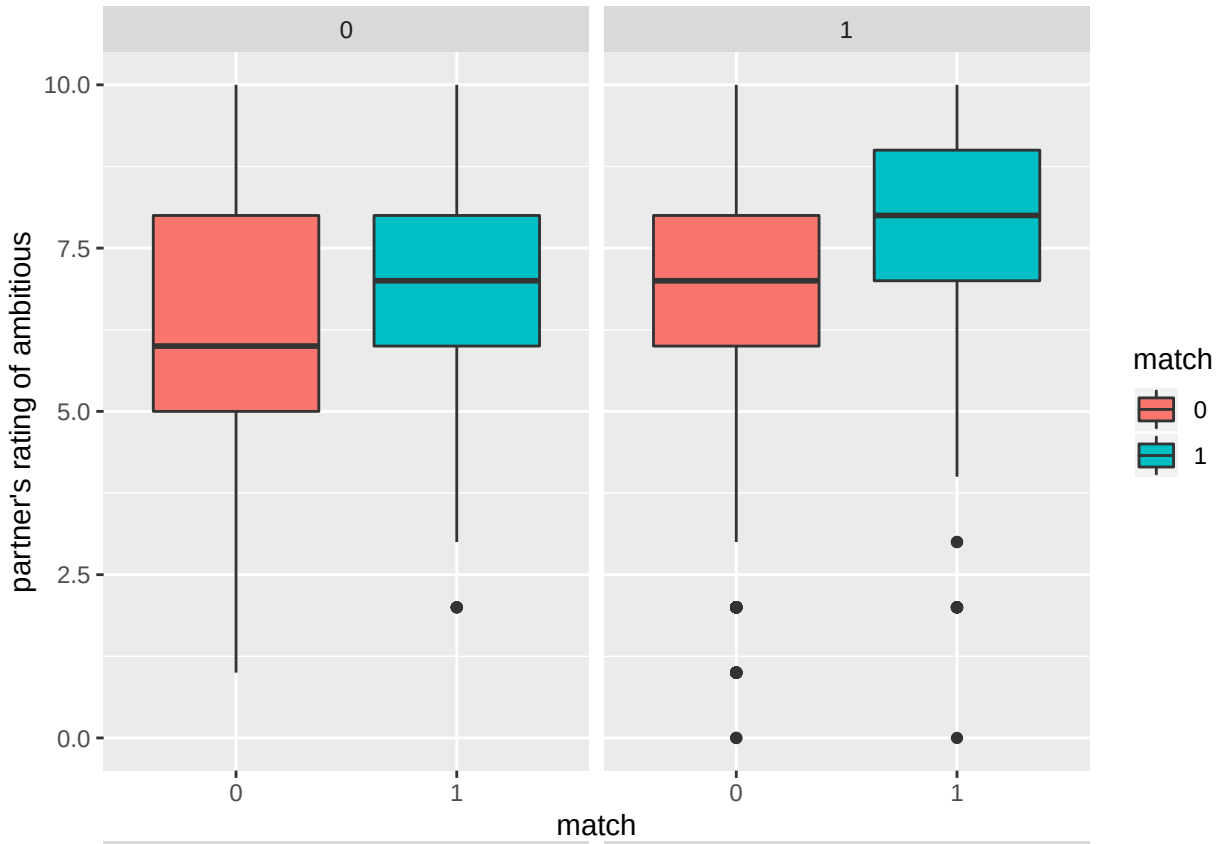

0
1
10
8
6
4
2
self ambitious
0
1
match
match
0
1

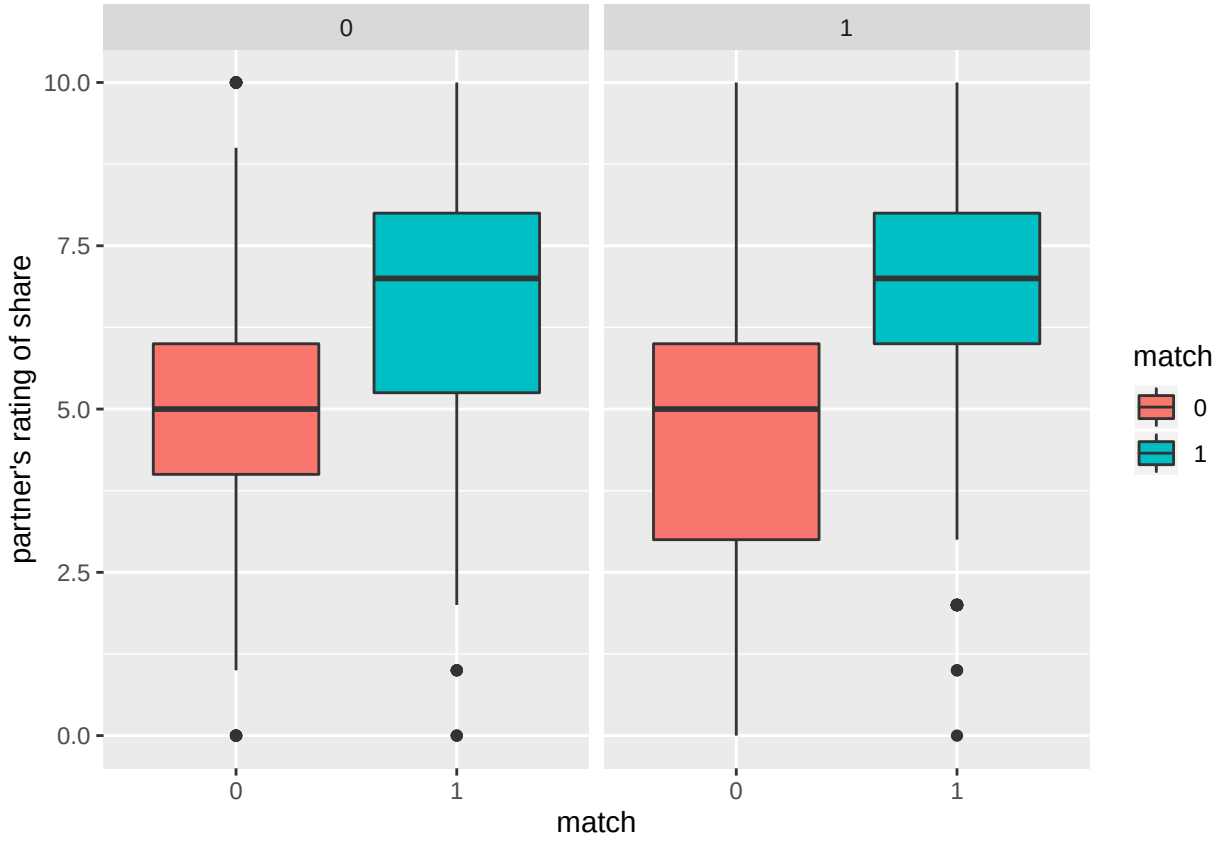

0
1
10.0
7.5
5.0
2.5
0.0
partner's rating of attractive
0
1
match
match
0
1


0
1
10.0
7.5
5.0
2.5
0.0
partner's rating of sincere
0
1
match
match
0
1


0
1
10.0
7.5
5.0
2.5
0.0
partner's rating of intel
0
1
match
match
0
1

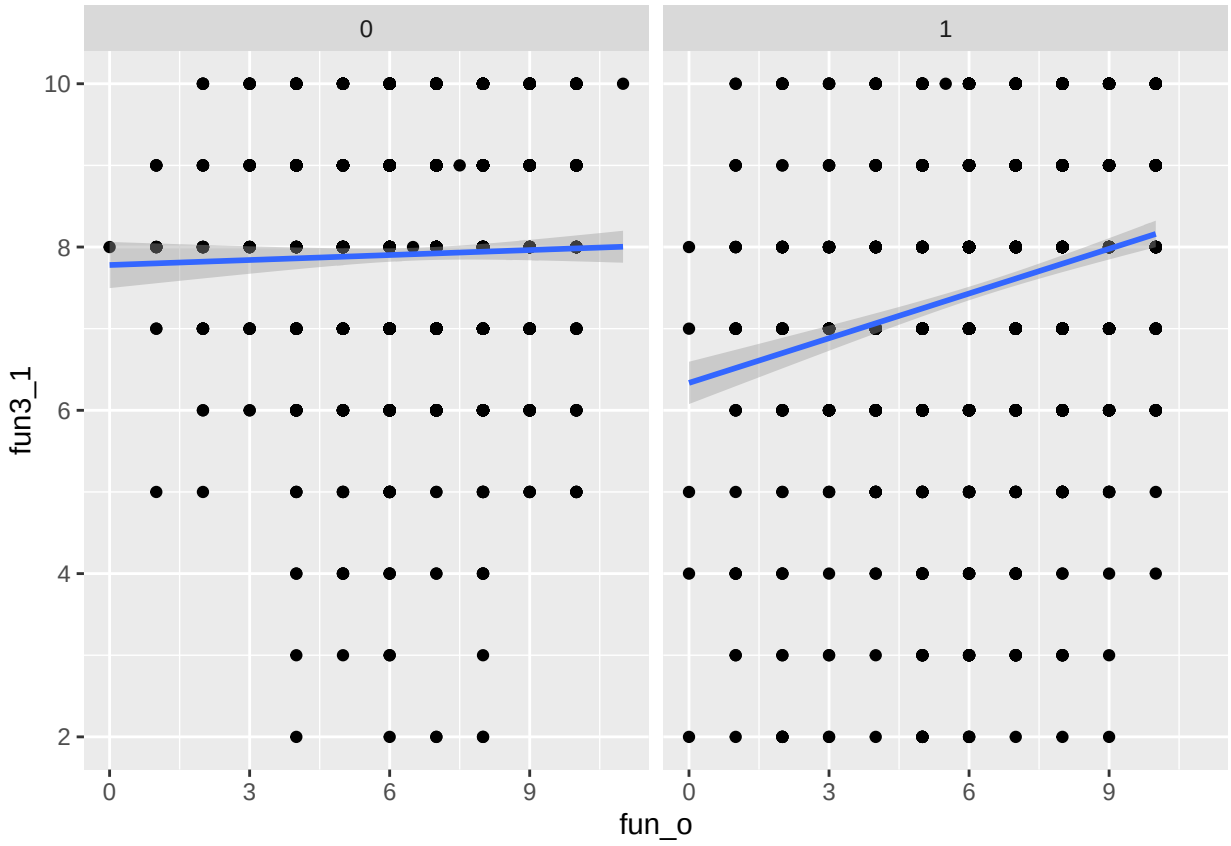

0
1
9
6
3
0
partner's rating of fun
0
1
match
match
0
1

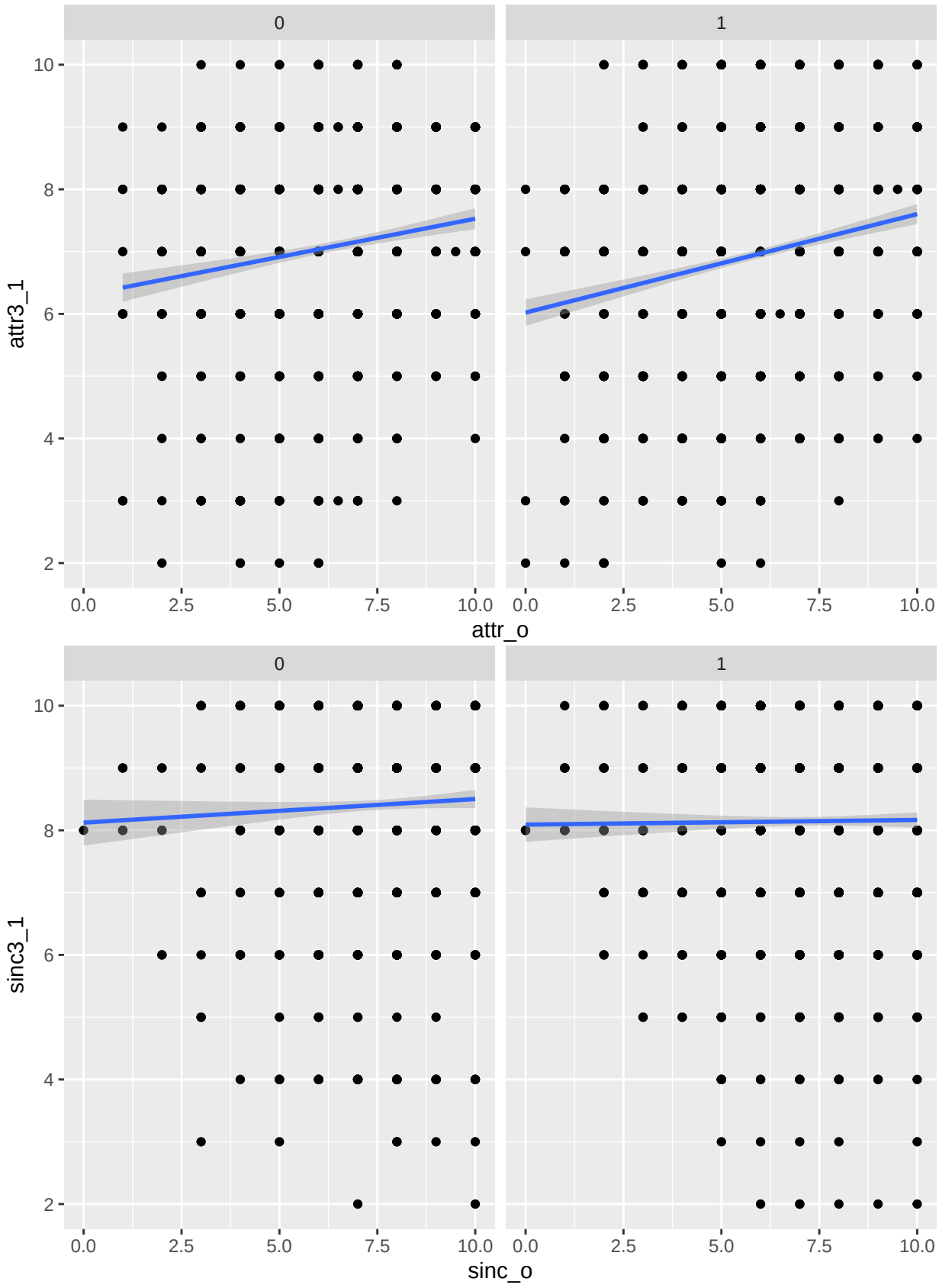

0
1
10.0
7.5
5.0
2.5
0.0
partner's rating of ambitious
0
1
0
1
match
match
0
1

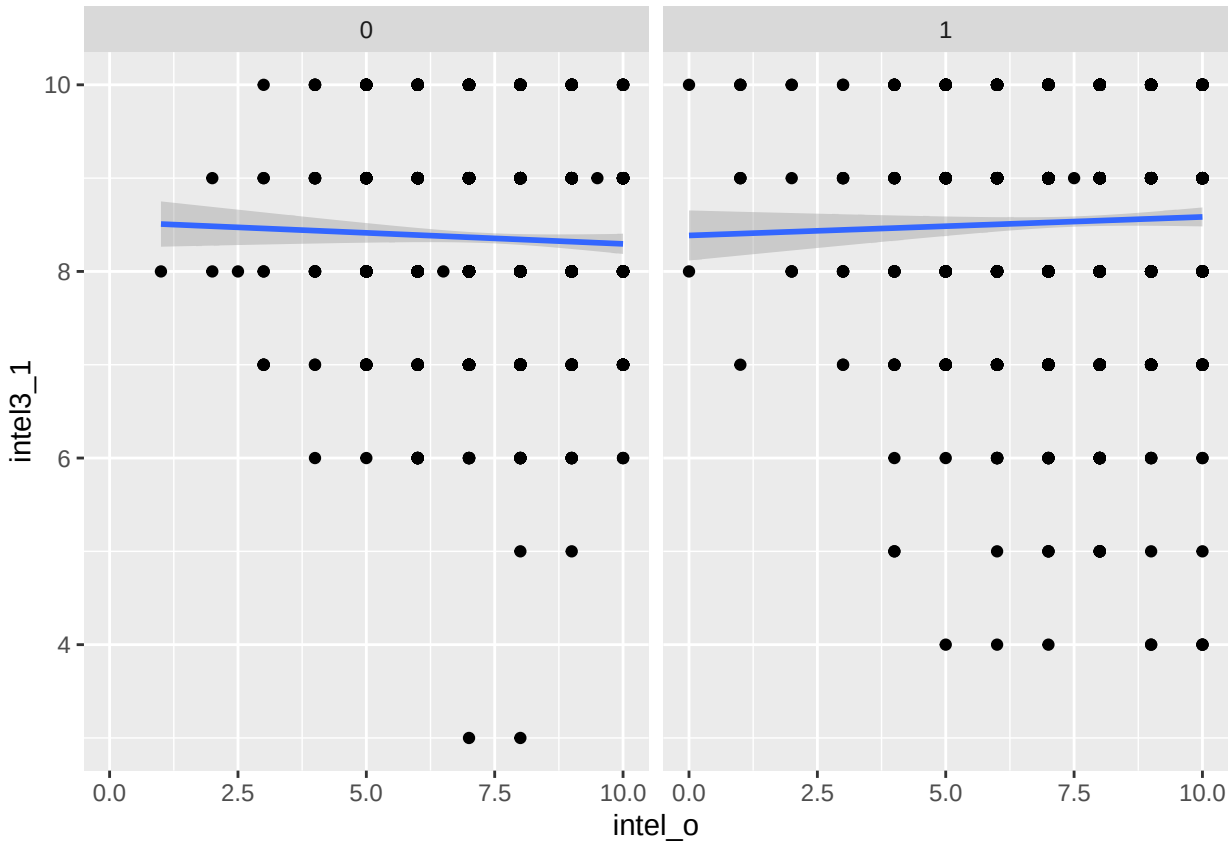

0
1
10.0
7.5
5.0
2.5
0.0
partner's rating of share
0
1
0
1
match
match
0
1

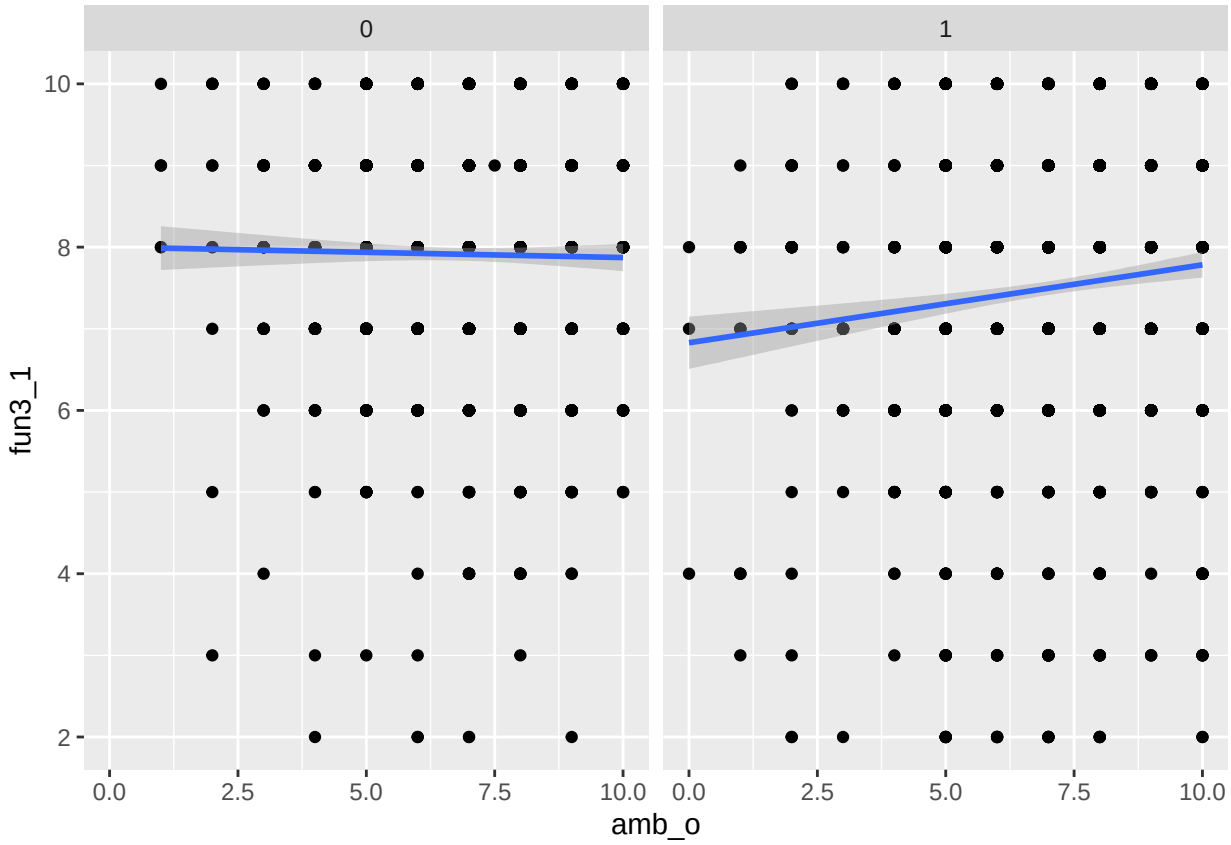
```
##      samerace
```

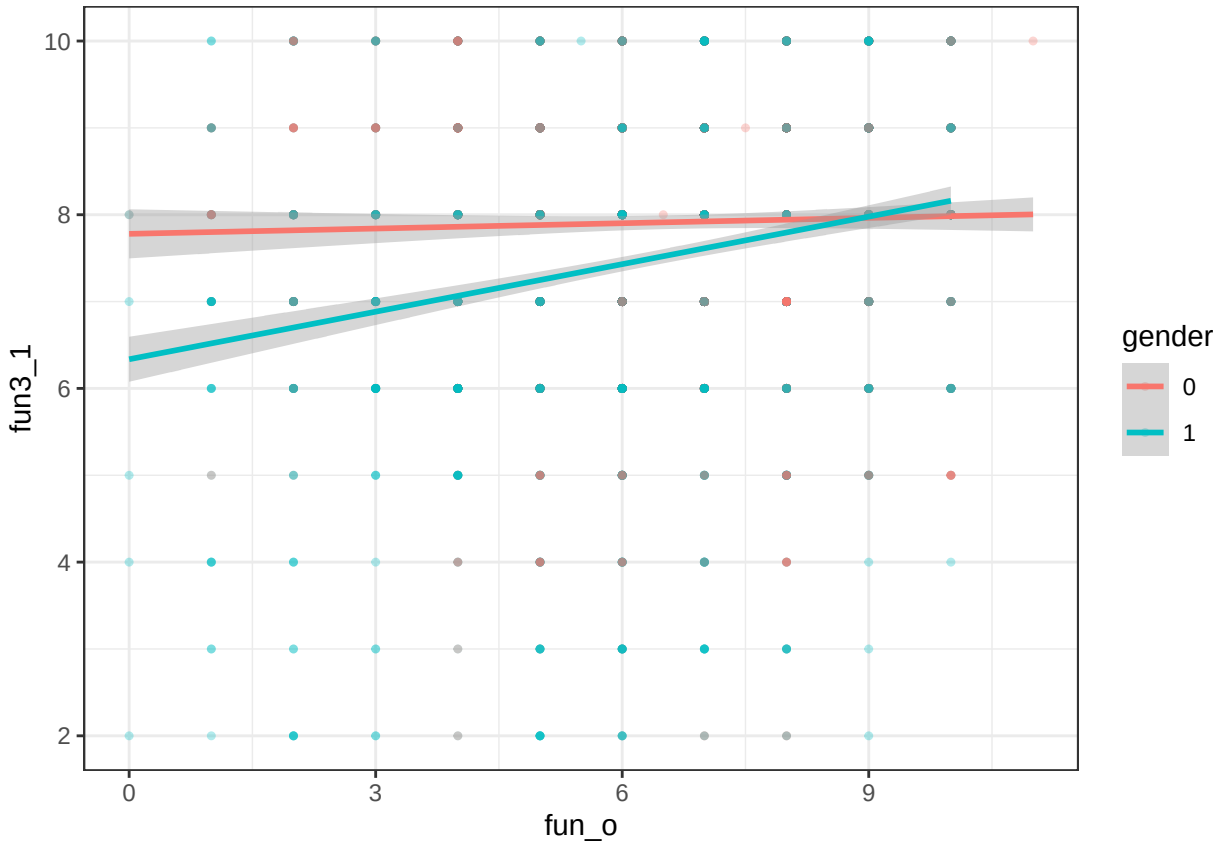
```
## match         0         1
##     0 0.3606614 0.2435412
##     1 0.2352739 0.1605236
```




0
1
attr3_1
attr_o
sinc3_1
sinc_o
2
4
6
8
10
0.0
2.5
5.0
7.5
10.0


0
1
intel3_1
10
8
6
4
0.0
2.5
5.0
7.5
10.0
intel_o


0
1
fun3_1
10
8
6
4
2
0.0
2.5
5.0
7.5
10.0
amb_o



Appendix 14 (Confusion matrix and ROC curve for the #2 male model)

```
## 1. For male
#confusion matrix

Conf_mat <- confusionMatrix(as.factor(ifelse(fitted(match_m) >= 0.325, "1","0")),
                            as.factor(m1_m$match),positive = "1")
Conf_mat$table
```

```
##            Reference
## Prediction   0   1
##          0 764 107
##          1 287 456
```

```
# accuracy: 0.76
Conf_mat$overall["Accuracy"];
```

```
## Accuracy
## 0.755886
```

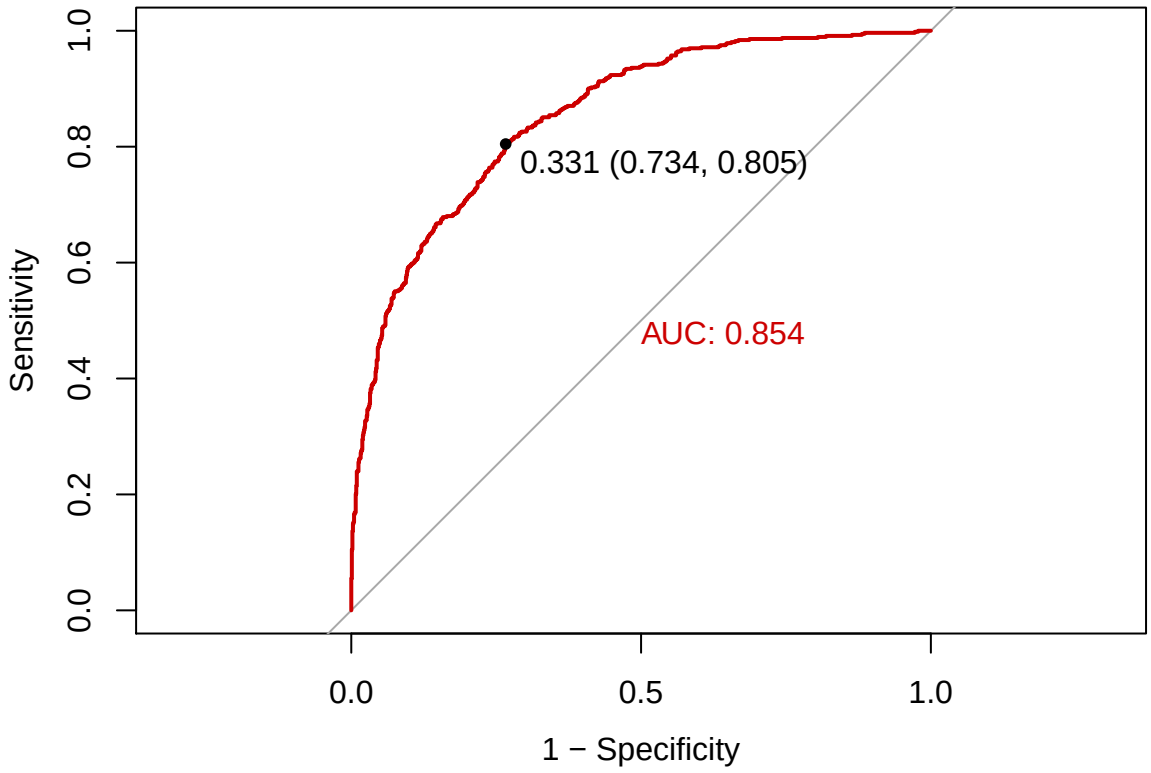
```
Conf_mat$byClass[c("Sensitivity","Specificity")]
```

```
## Sensitivity Specificity
##   0.8099467   0.7269267
```

```
#Roc curve
roc(m1_m$match,fitted(match_m),plot=T,print.thres="best",legacy.axes=T,
    print.auc =T,col="red3")
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```




```
##
## Call:
## roc.default(response = m1_m$match, predictor = fitted(match_m),     plot = T, print.thres = "best", l
##
## Data: fitted(match_m) in 1051 controls (m1_m$match 0) < 563 cases (m1_m$match 1).
## Area under the curve: 0.854
```

Appendix 15 (Confusion matrix and ROC curve for the #2 female model)

```
##2. For female
Conf_mat2 <- confusionMatrix(as.factor(ifelse(fitted(match_f) >= 0.514, "1","0")),
                             as.factor(m1_f$match),positive = "1")
Conf_mat2$table
```

```
##            Reference
## Prediction   0   1
##          0 588 141
##          1 115 445
```

```
#accuracy : 0.79
Conf_mat2$overall["Accuracy"];
```

```
##  Accuracy
## 0.8013964
```

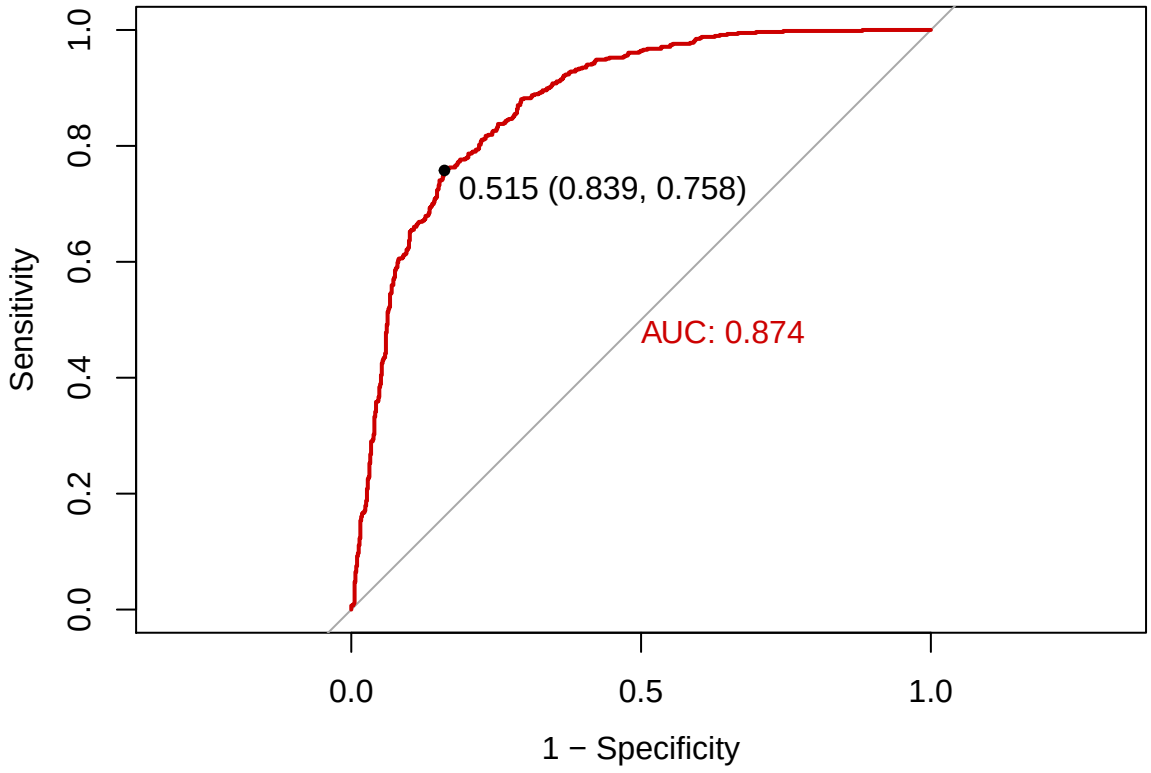
```
Conf_mat2$byClass[c("Sensitivity","Specificity")]
```

```
## Sensitivity Specificity
##   0.7593857   0.8364154
```

```
#Roc curve
roc(m1_f$match,fitted(match_f),plot=T,print.thres="best",legacy.axes=T,
    print.auc =T,col="red3")
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

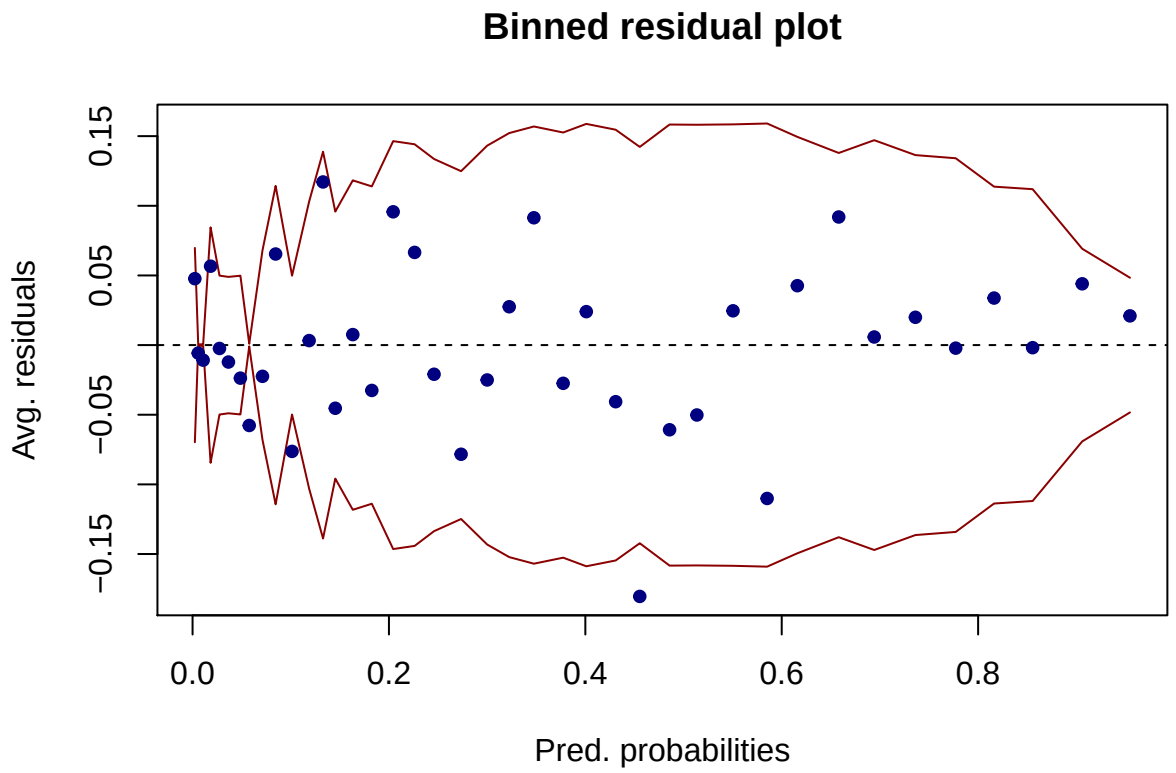



```
##
## Call:
## roc.default(response = m1_f$match, predictor = fitted(match_f),     plot = T, print.thres = "best", l
##
## Data: fitted(match_f) in 703 controls (m1_f$match 0) < 586 cases (m1_f$match 1).
## Area under the curve: 0.8741
```

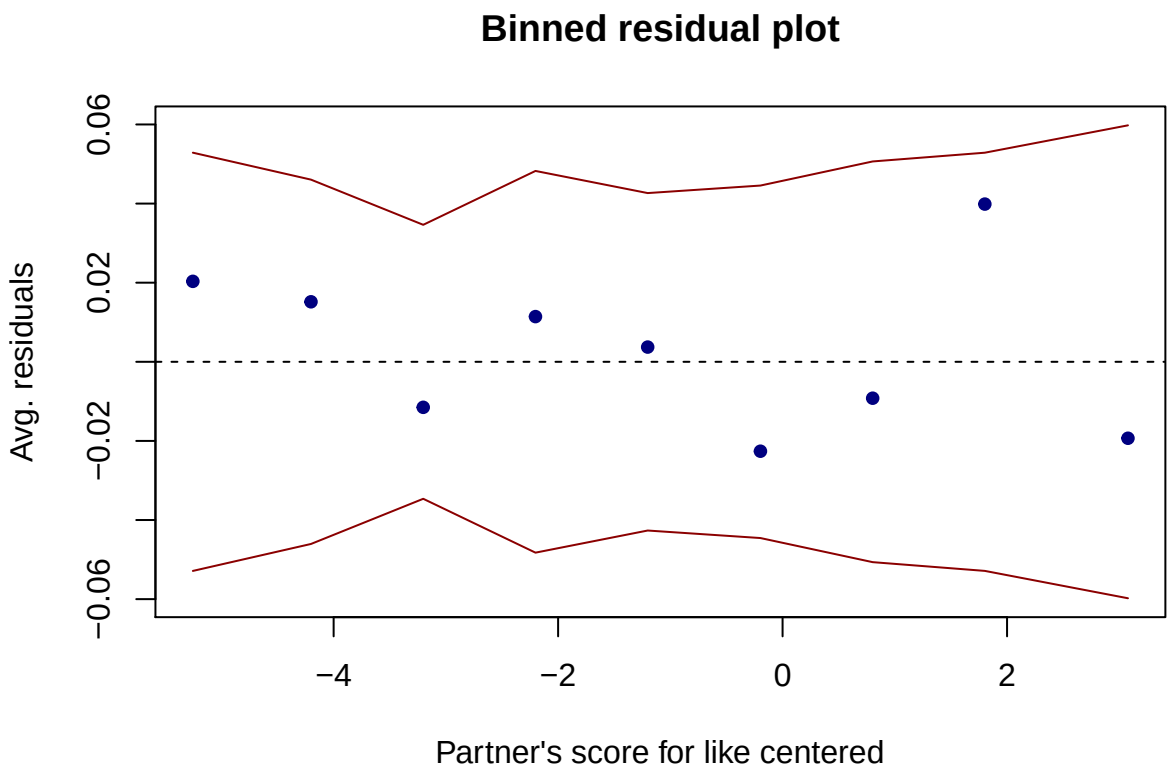
Appendix 16 (Binned residual for the #2 male model)

```
## 1. For male
```

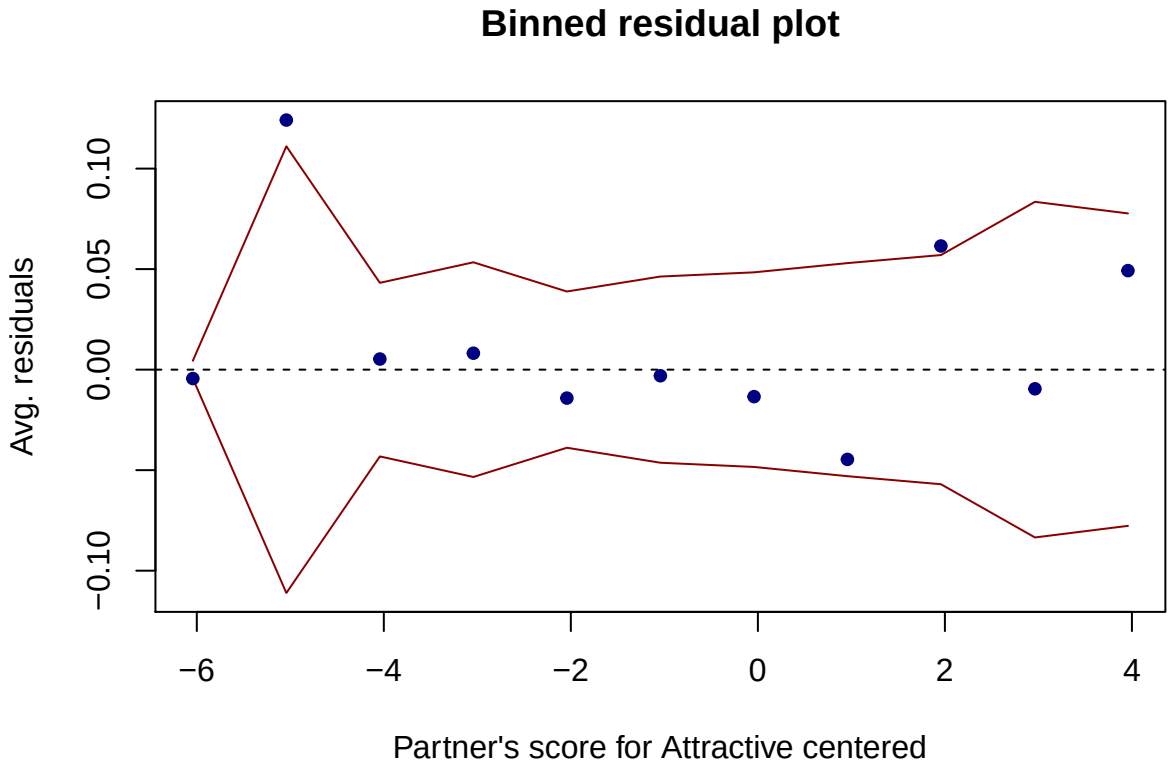
```
rawresid2 <- residuals(match_m,"resp")
#binned residual plots
binnedplot(x=fitted(match_m),y=rawresid2,xlab="Pred. probabilities",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



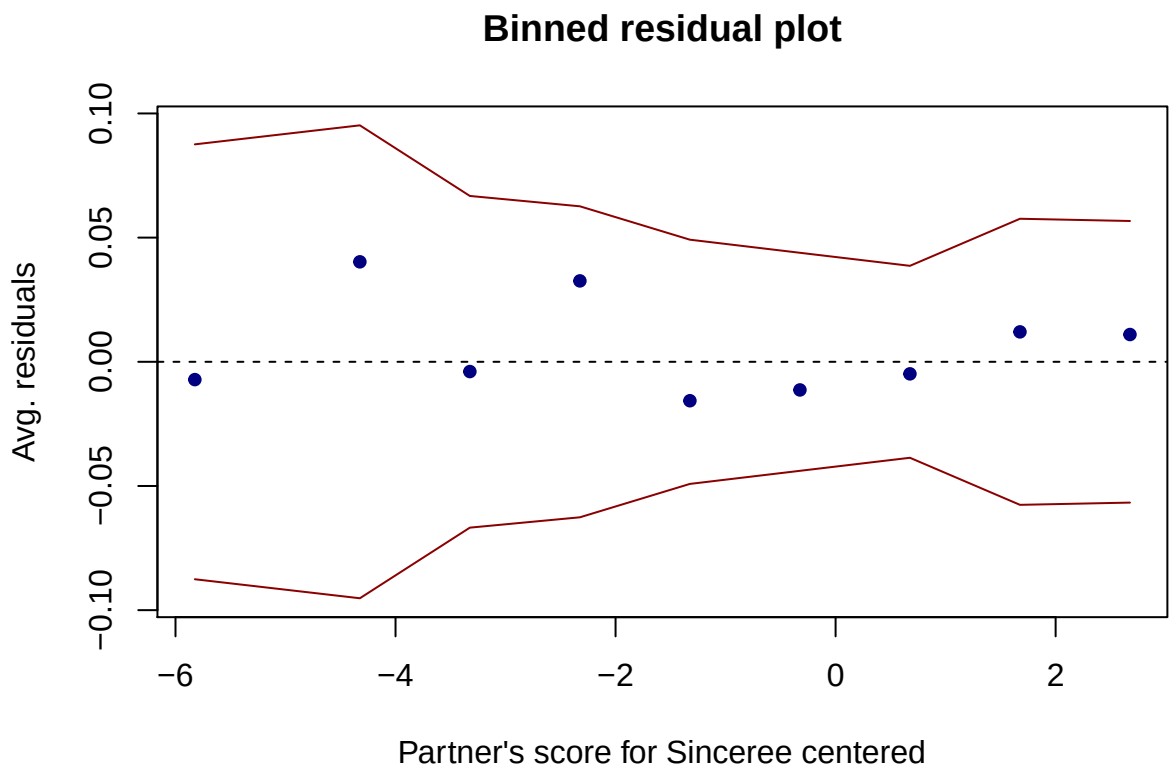

```
binnedplot(m1_m$like_o_c,y=rawresid2,xlab="Partner's score for like centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



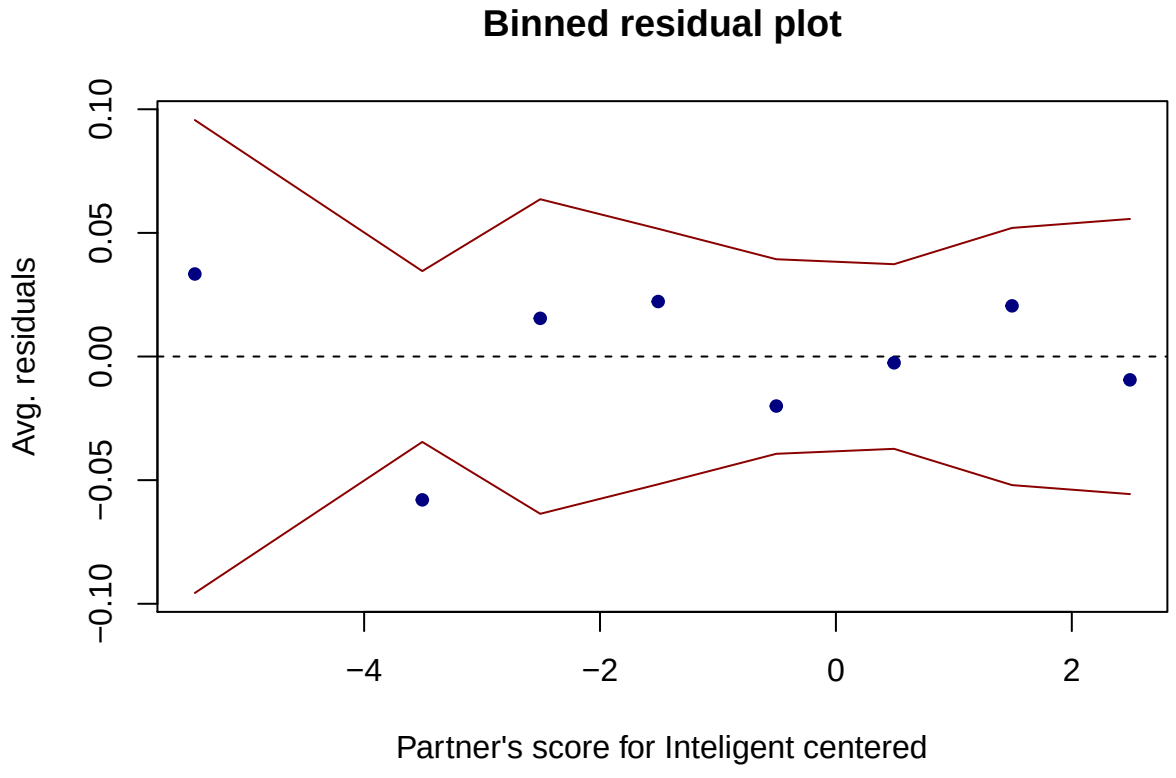

```
binnedplot(m1_m$attr_o_c,y=rawresid2,xlab="Partner's score for Attractive centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



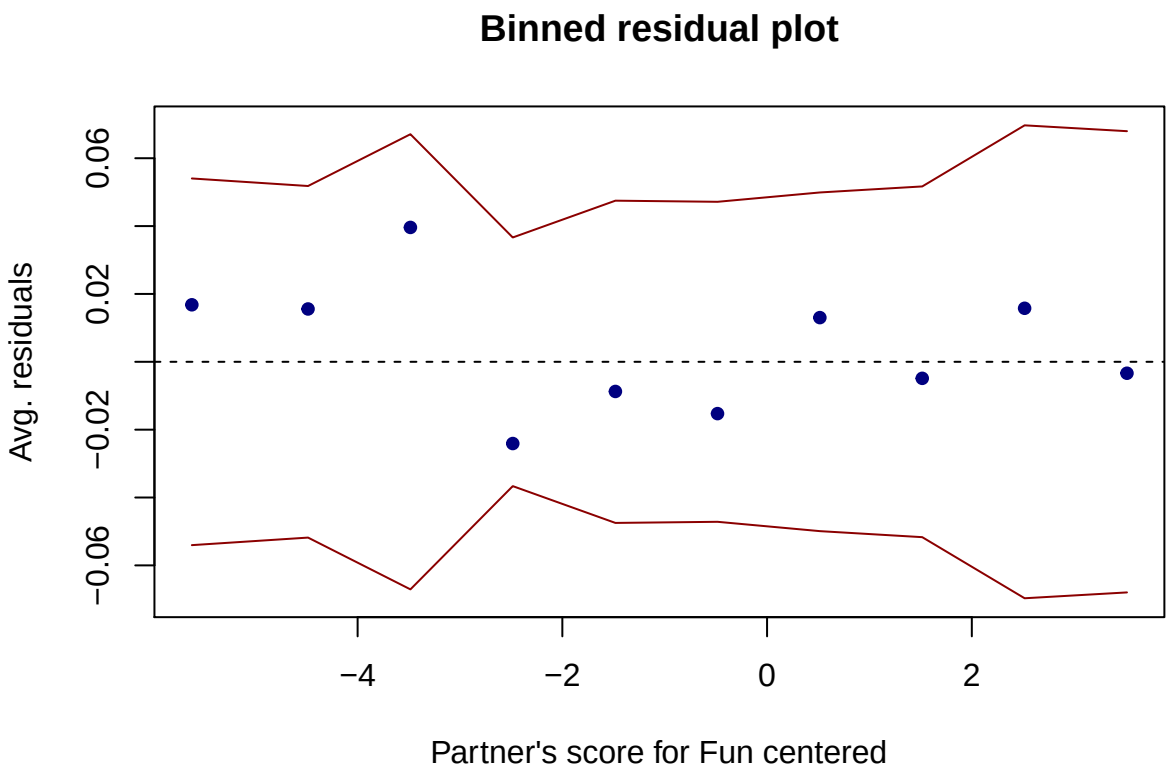

```
binnedplot(m1_m$sinc_o_c,y=rawresid2,xlab="Partner's score for Sinceree centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



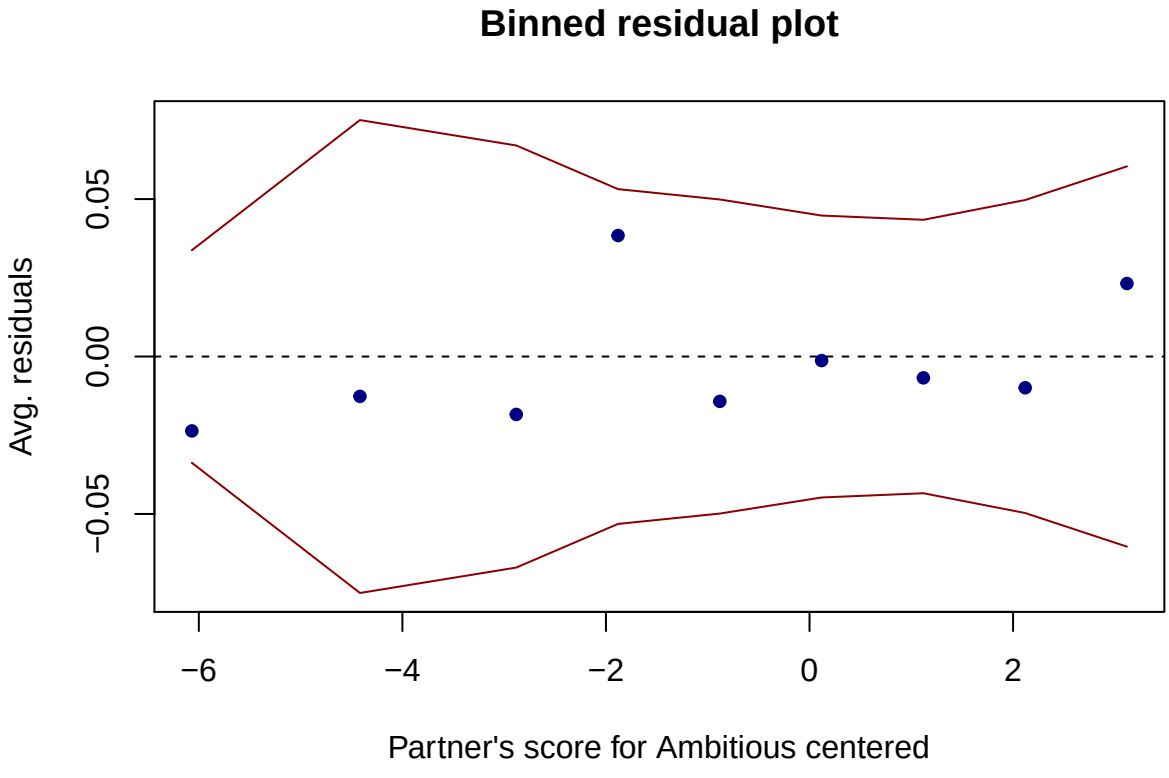

```
binnedplot(m1_m$intel_o_c,y=rawresid2,xlab="Partner's score for Inteligent centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



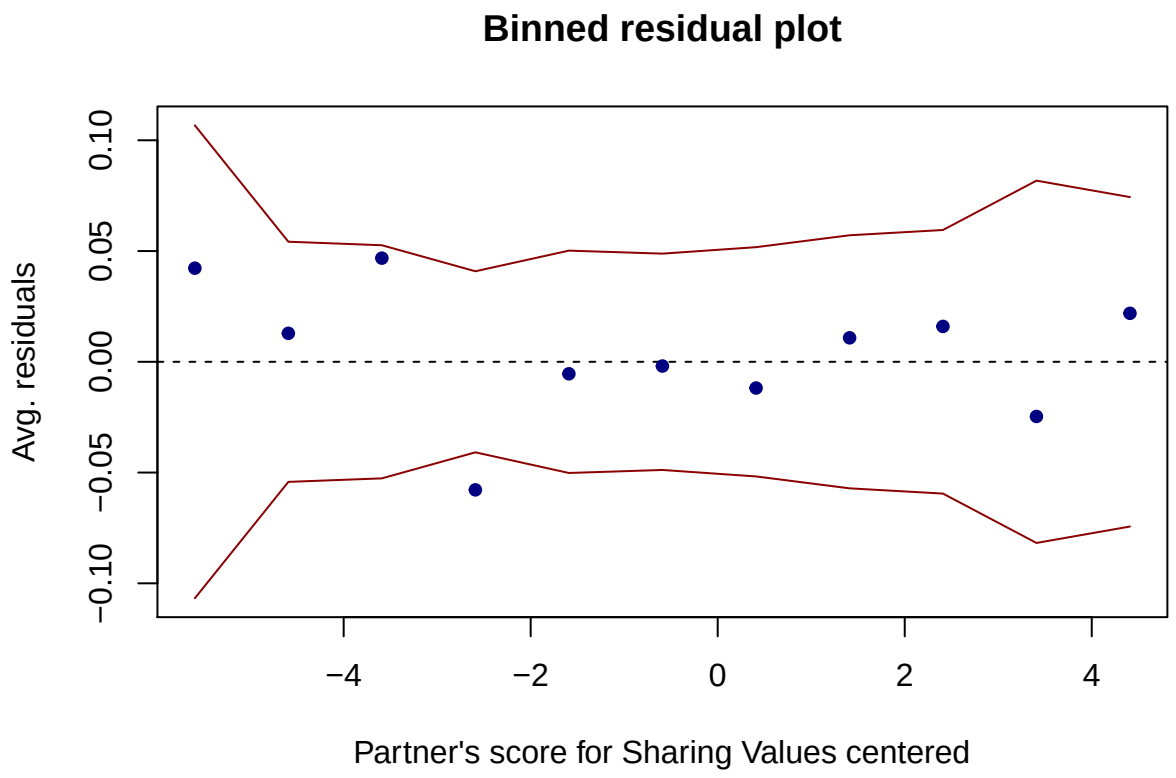

```
binnedplot(m1_m$fun_o_c,y=rawresid2,xlab="Partner's score for Fun centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



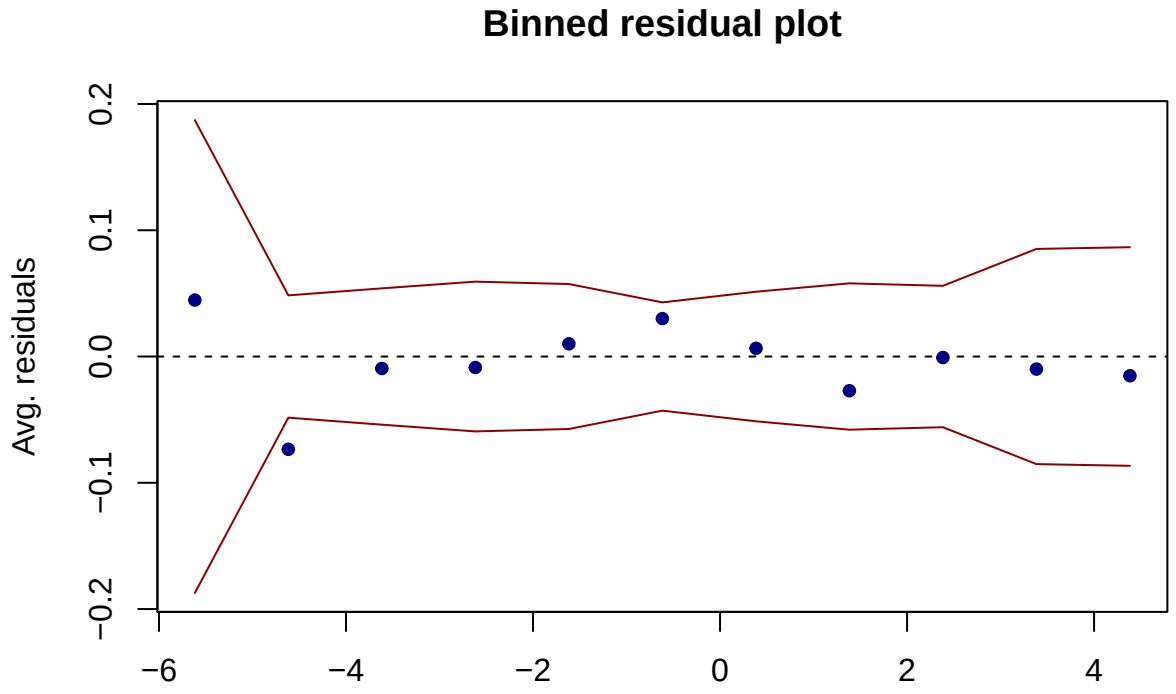

```
binnedplot(m1_m$amb_o_c,y=rawresid2,xlab="Partner's score for Ambitious centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



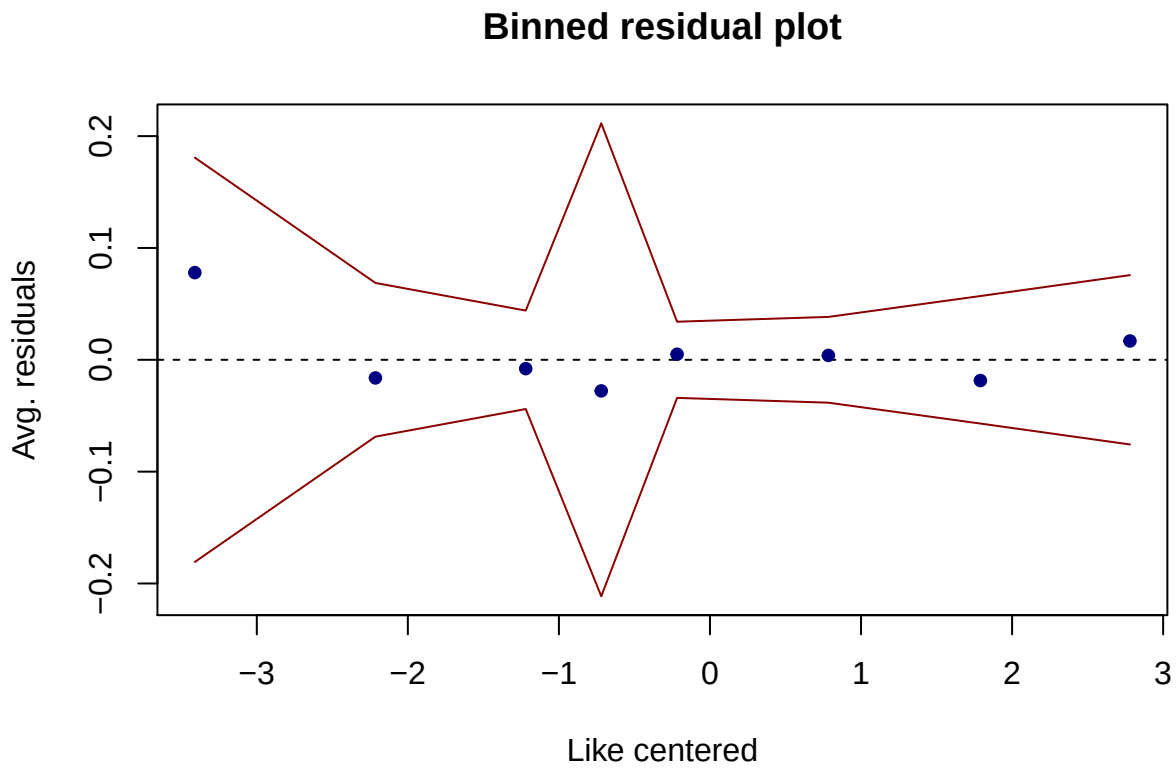

```
binnedplot(m1_m$shar_o_c,y=rawresid2,xlab="Partner's score for Sharing Values centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



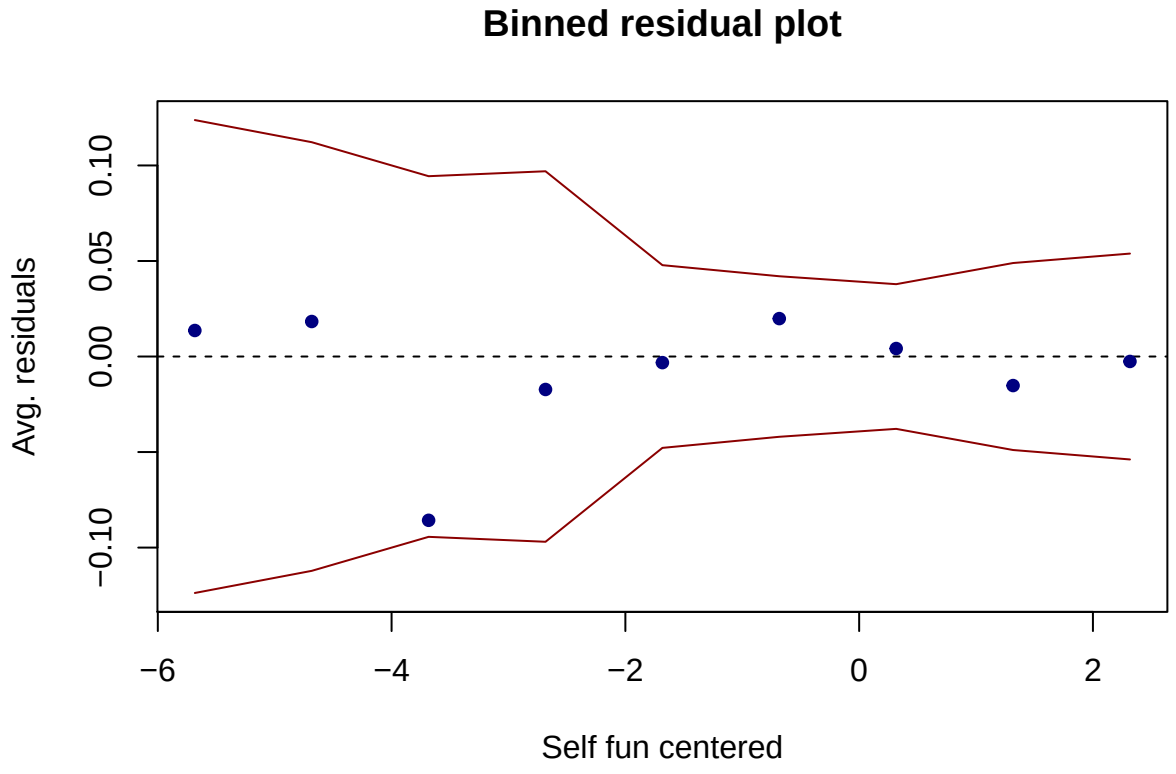

```
binnedplot(m1_m$prob_o_c,y=rawresid2,xlab="Partner(female)'s score for the probablitiy that male says y
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```




```
binnedplot(m1_m$like_c,y=rawresid2,xlab="Like centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```

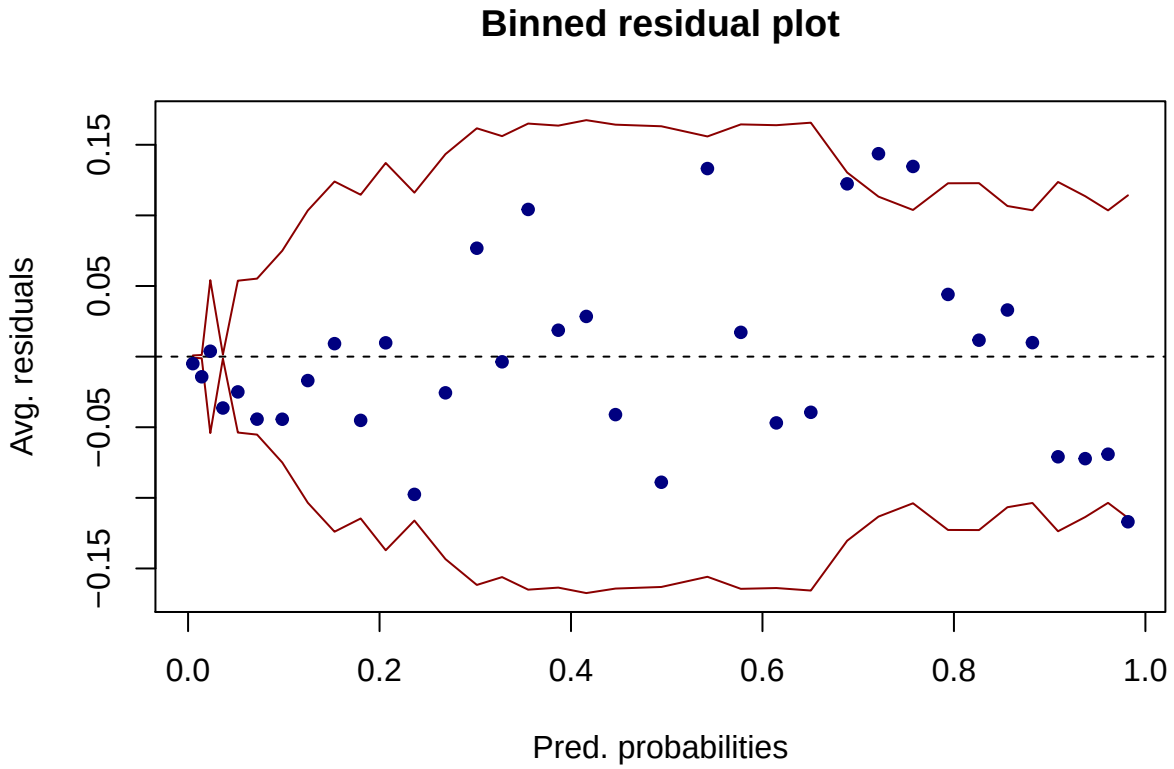



```
binnedplot(m1_m$fun3_1_c,y=rawresid2,xlab="Self fun centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```

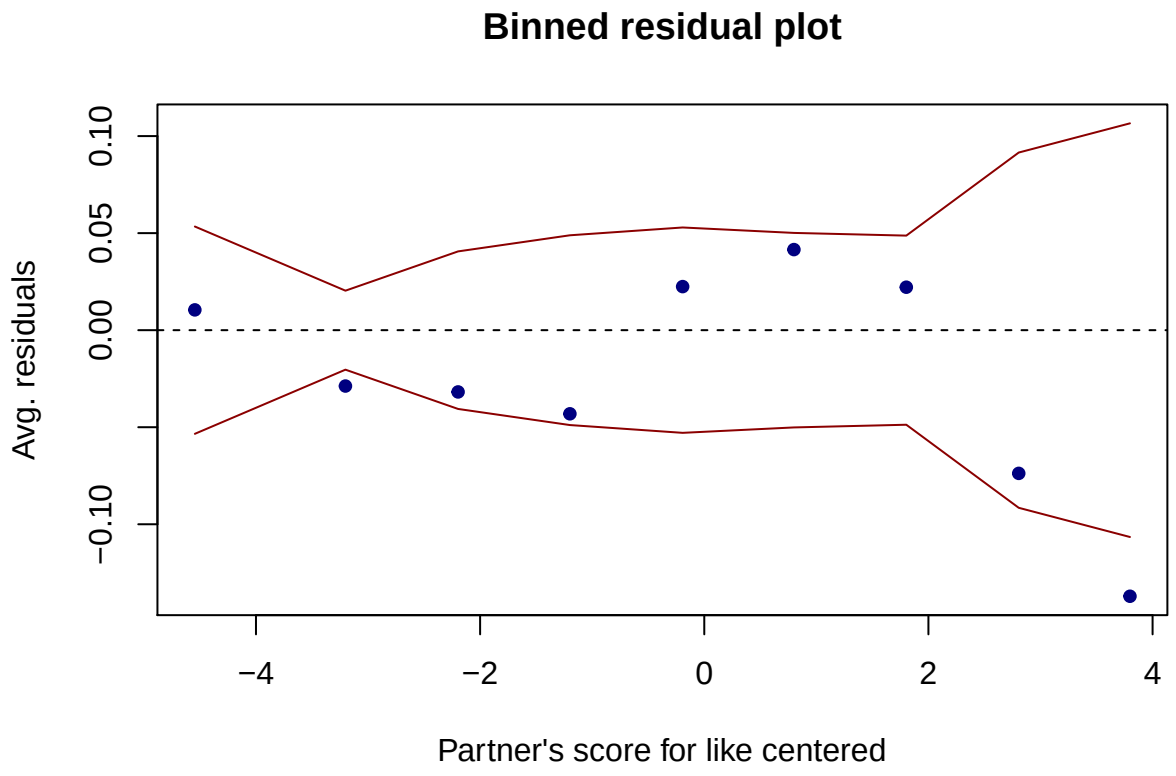


Appendix 17 (Binned residual for the #2 female model)

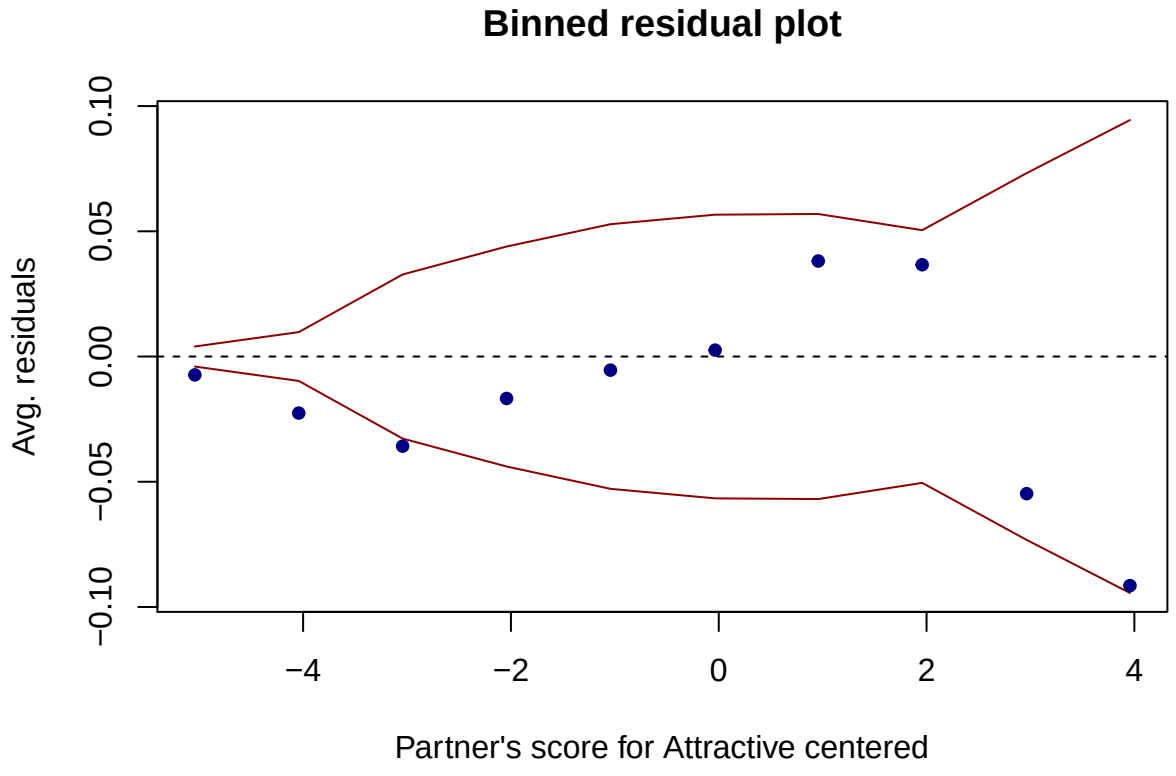
```
##2. For female
rawresid3 <- residuals(match_f,"resp")
#binned residual plots
binnedplot(x=fitted(match_f),y=rawresid3,xlab="Pred. probabilities",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



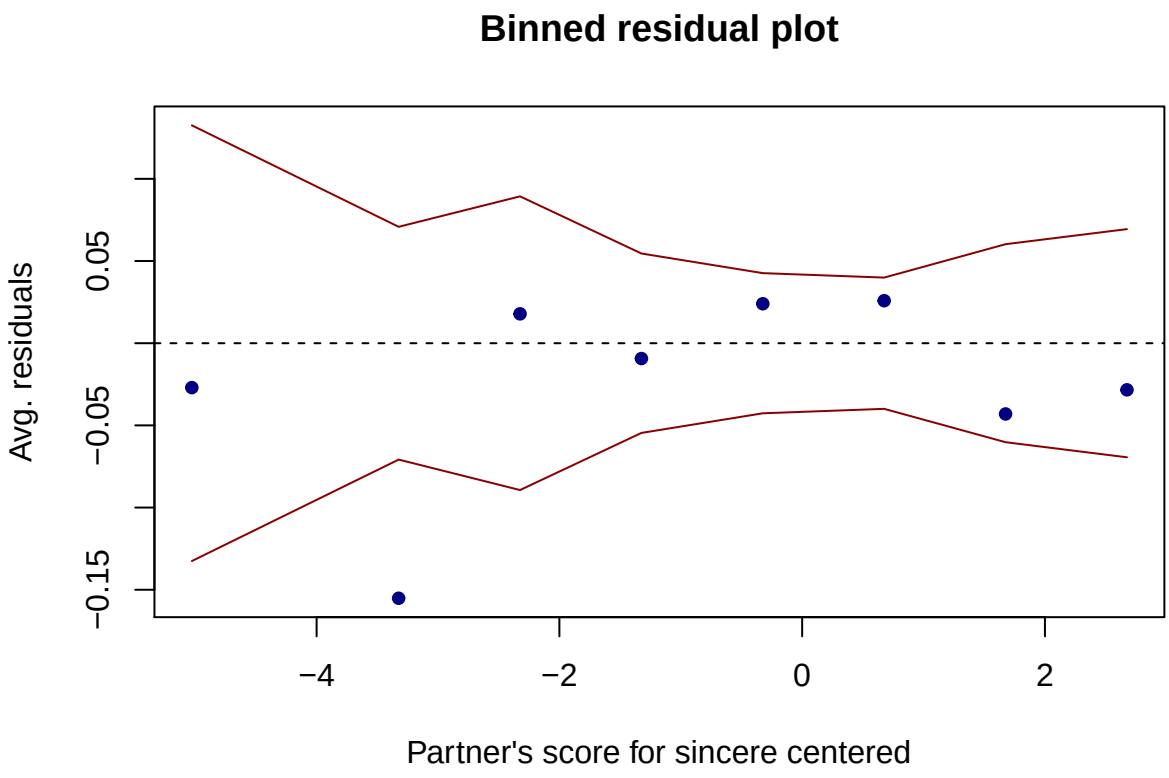

```
binnedplot(m1_f$like_o_c,y=rawresid3,xlab="Partner's score for like centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



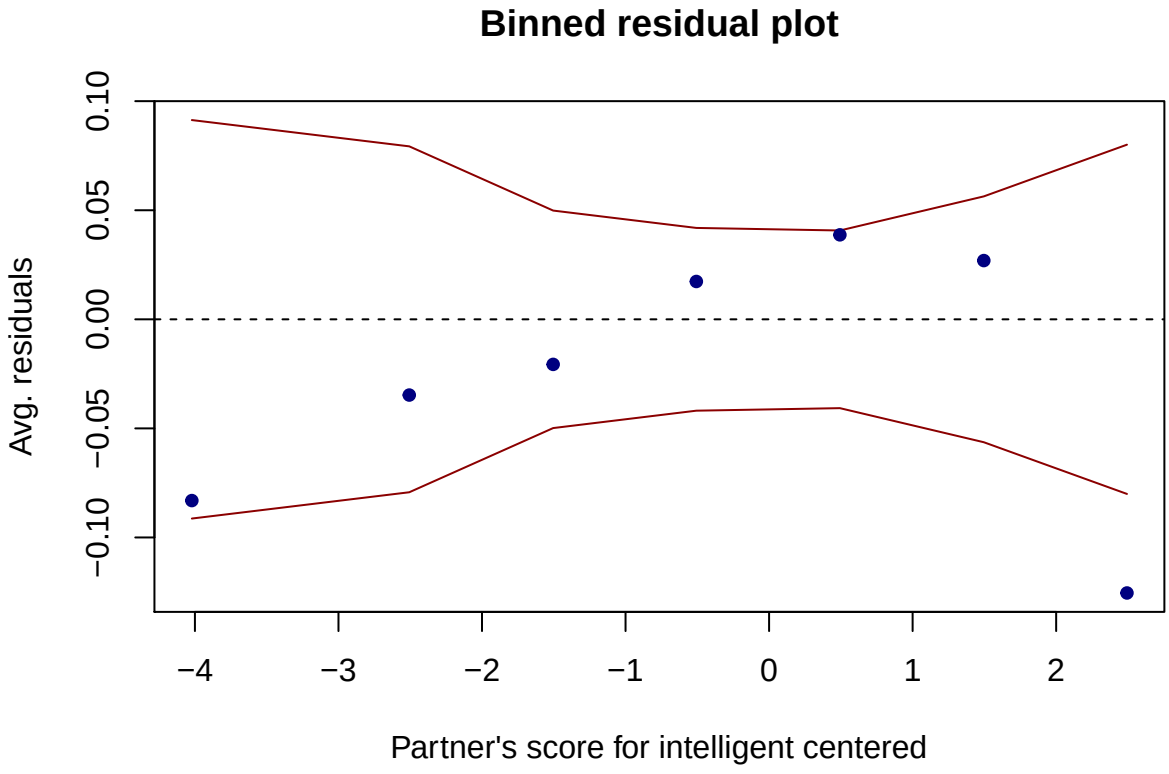

```
binnedplot(m1_f$attr_o_c,y=rawresid3,xlab="Partner's score for Attractive centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



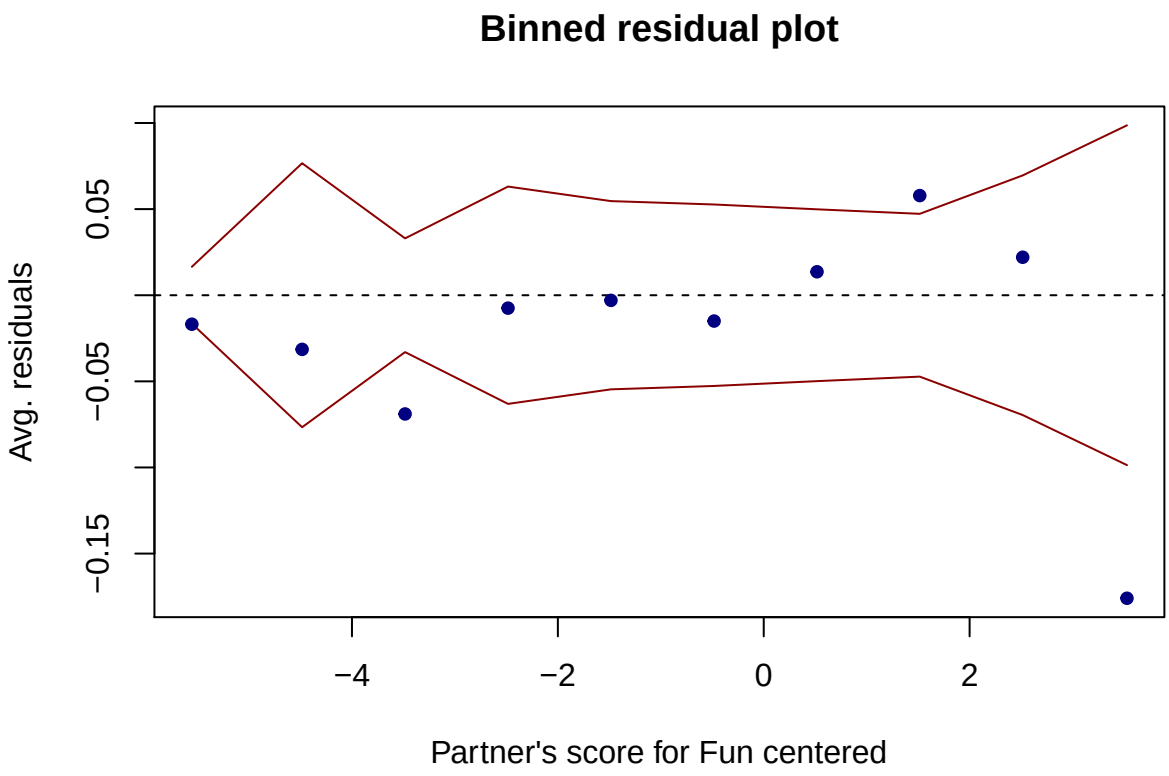

```
binnedplot(m1_f$sinc_o_c,y=rawresid3,xlab="Partner's score for sincere centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



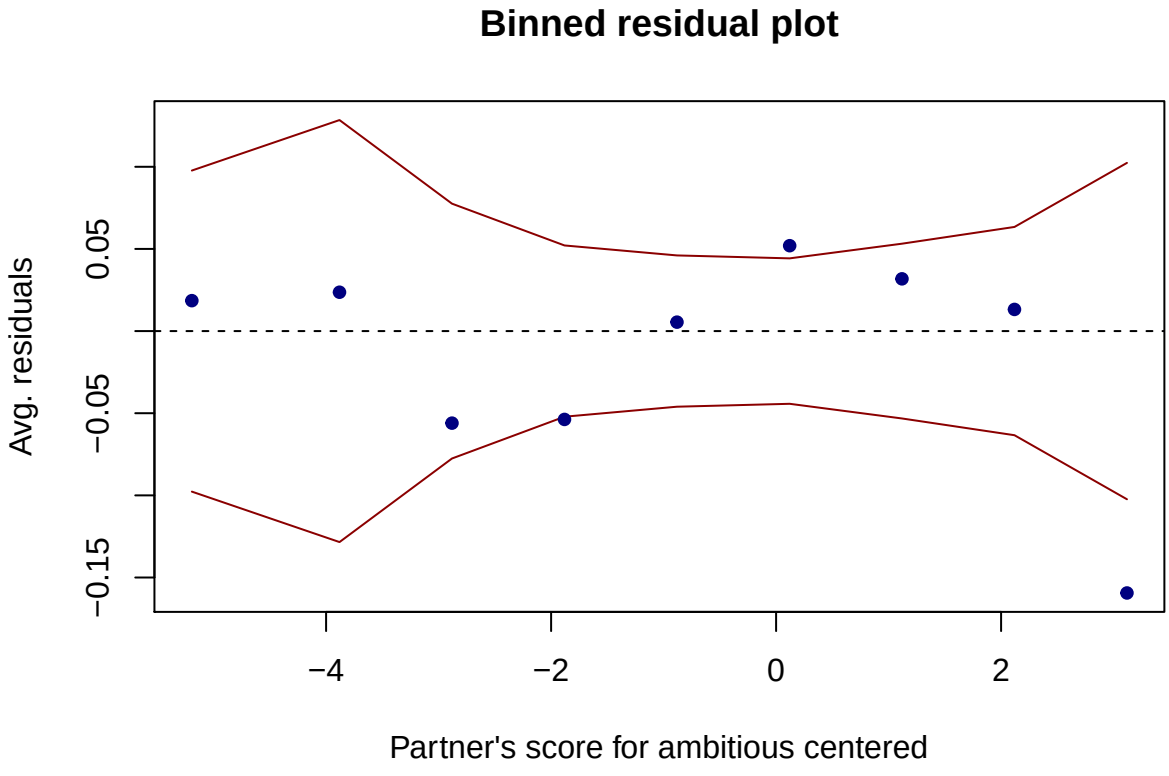

```
binnedplot(m1_f$intel_o_c,y=rawresid3,xlab="Partner's score for intelligent centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



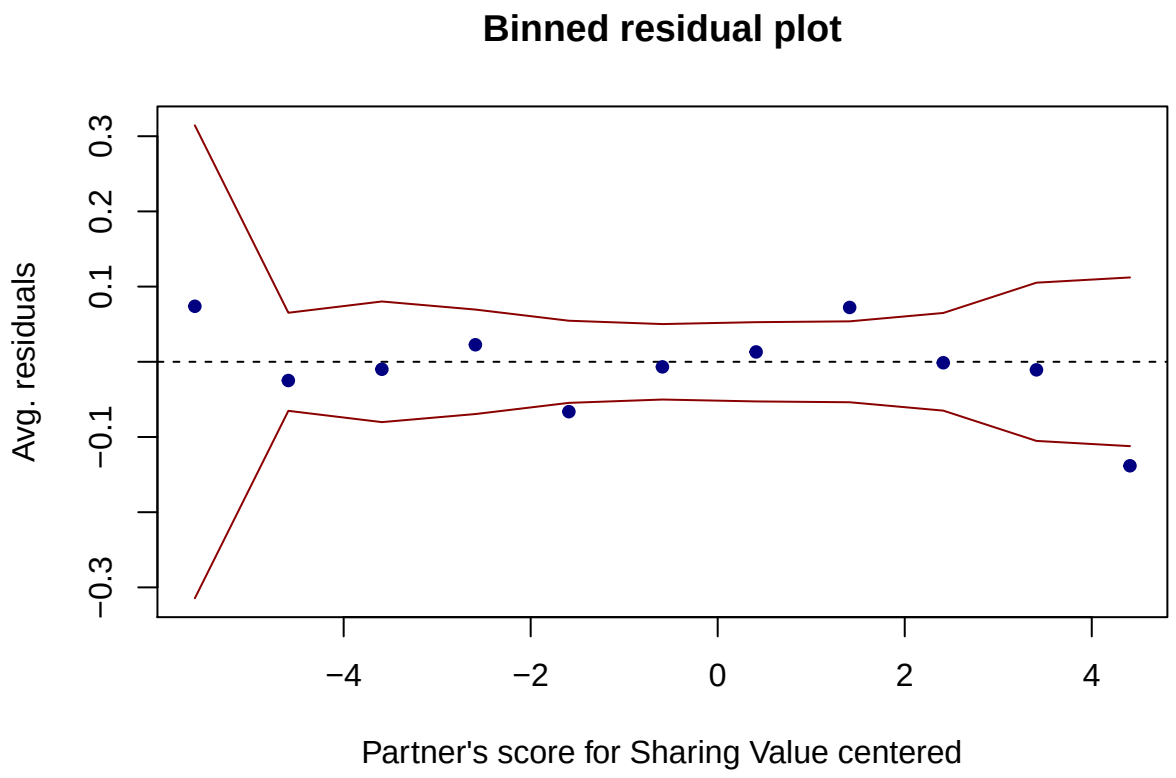

```
binnedplot(m1_f$fun_o_c,y=rawresid3,xlab="Partner's score for Fun centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



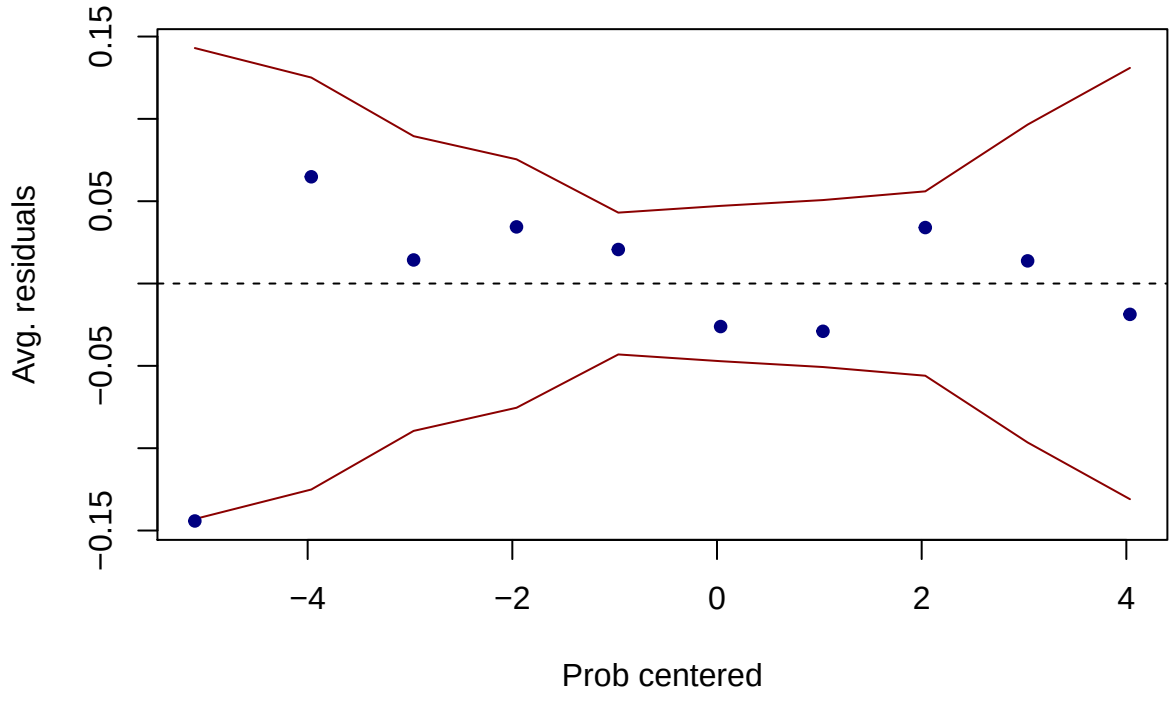

```
binnedplot(m1_f$amb_o_c,y=rawresid3,xlab="Partner's score for ambitious centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



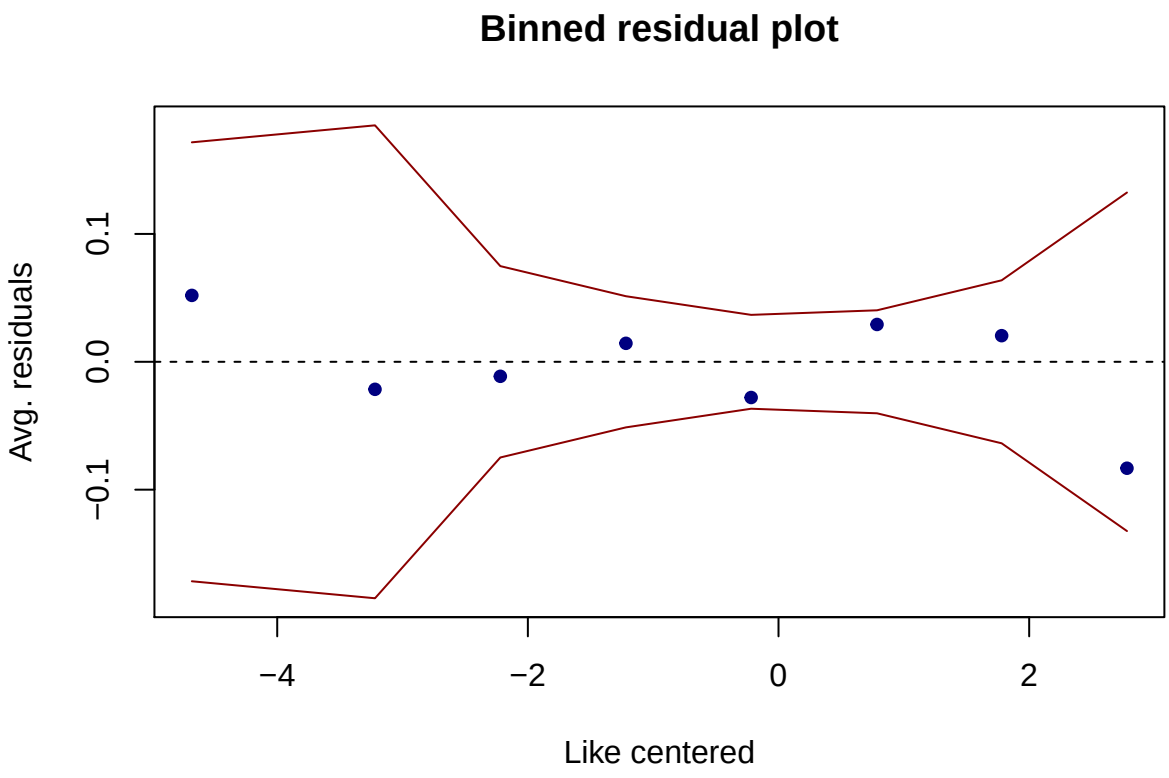

```
binnedplot(m1_f$shar_o_c,y=rawresid3,xlab="Partner's score for Sharing Value centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



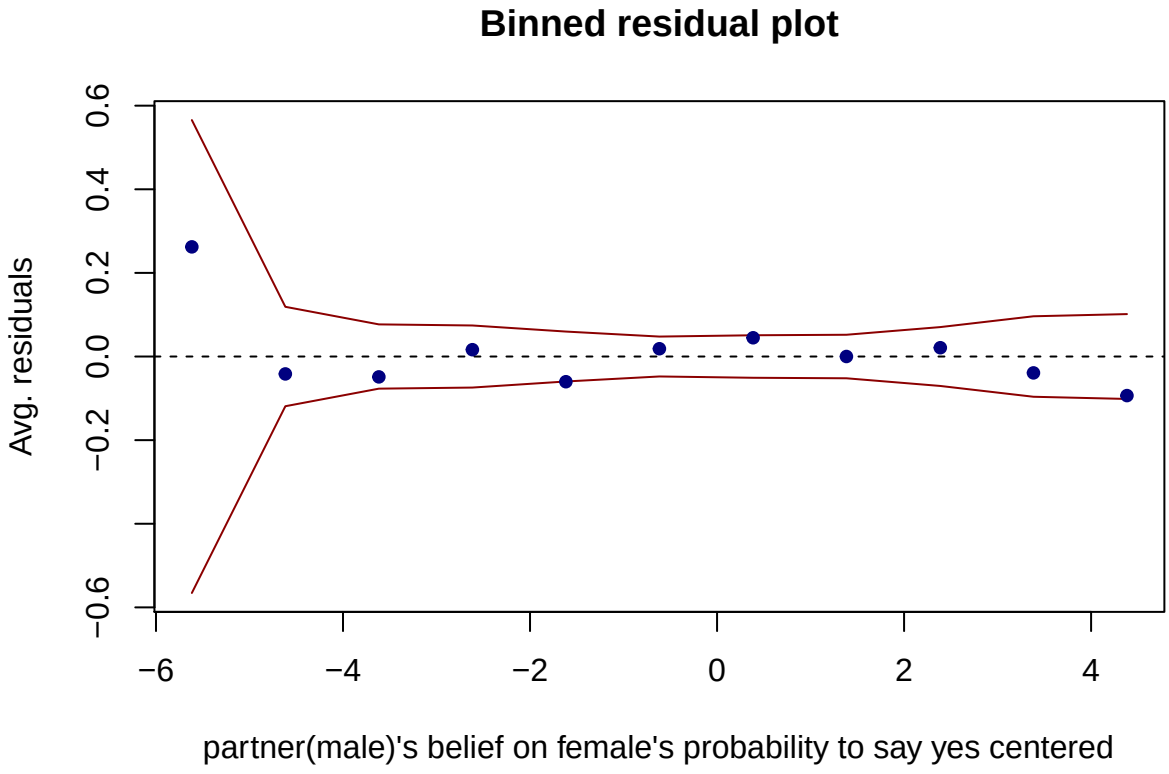

```
binnedplot(m1_f$prob_c,y=rawresid3,xlab="Prob centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```



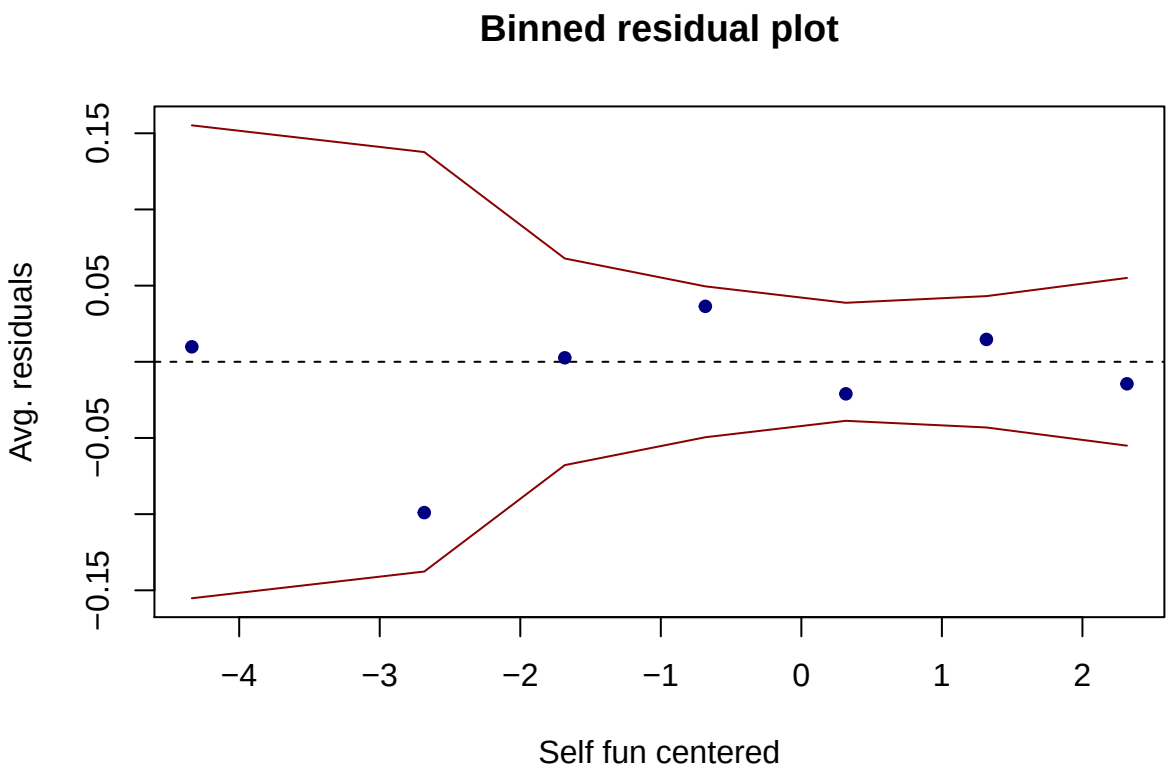

```
binnedplot(m1_f$like_c,y=rawresid3,xlab="Like centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```




```
binnedplot(m1_f$prob_o_c,y=rawresid3,xlab="partner(male)'s belief on female's probability to say yes ce
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```




```
binnedplot(m1_f$fun3_1_c,y=rawresid3,xlab="Self fun centered",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
```


Binned residual plot
Avg. residuals
0.15
0.05
−0.05
−0.15
−4
−3
−2
−1
0
1
2
Self fun centered