HW4\_Template

Lina Cao , Thursdays 3:30PM

SDGB 7844; Prof. Nagaraja; Spring 2018

## Question 1

### Based on the variable “Decision”, fill out the contingency table below. What percentage of dates ended with both people wanting a second date?

## [1] FALSE

## [1] FALSE

Since there is no missing data in “DecisionM” and “DecisionF”, we can handle the problem directly.

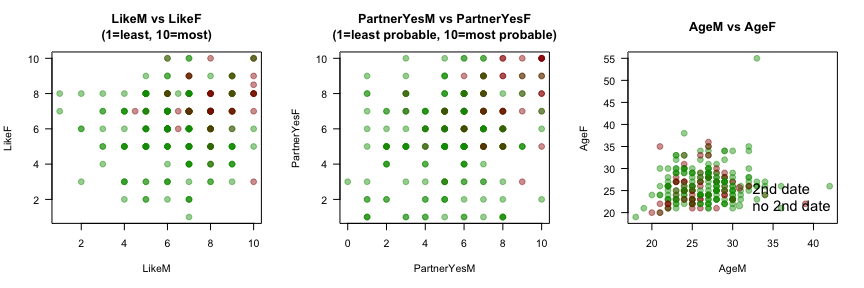
##   
## by Female- NO by Female- YES  
## by Male- NO 66 64  
## by Male- YES 83 63

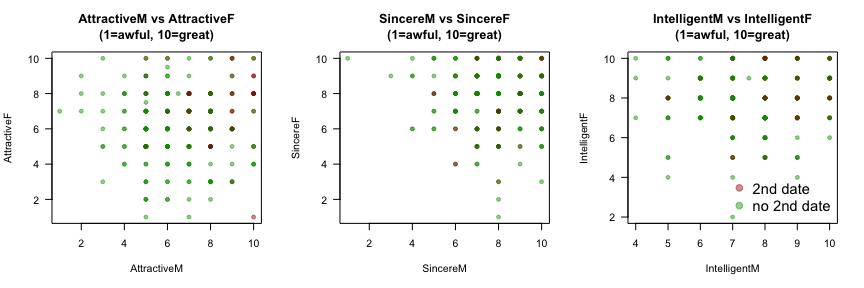
## [1] 0.2282609

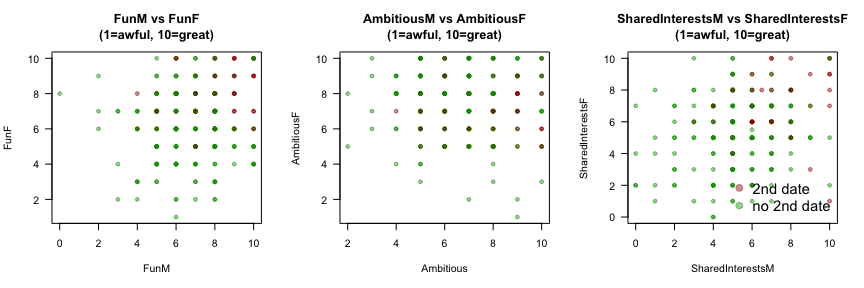
Therefore, with some calculation, about 22.83% of dates ended with both people wanting a second date.

## Question 2

A second date is planned only if both people within the matched pair want to see each other again. Make a new column in your data set and call it “second.date”. Values in this column should be 0 if there will be no second date, 1 if there will be a second date. Construct a scatterplot for each numerical variable where the male values are on the x-axis and the female values are on the y-axis. Observations in your scatterplot should have a different color (or pch value) based on whether or not there will be a second date. Describe what you see. (Note: Jitter your points just for making these plots.)

 (1) Like: the red dots cluster at top right, indicating the higher “Like” score one gives to the partner, a greater chance they will have a second date. (But it is not always the case, even in high level there are cases where people would not like to have a second date.) (2) PartnerYes: same as in “Like”, the groups that want to have a second date all give high score in “PartnerYes”. And the dots are more evenly distributed, (3) Age: most people in the experiment are younger than or equal to 35 years old, red dots appears more close to the “Y=X” line. This may tell us, second date may be more likely to happen if two people are at a similar age.

 (4) Attractive: red dots appears in the region where both gender gives “Attractive”(to the other person) higher than 5, and the higher their score is, the higher frequency a second date will occur. (5) Sincere: most candidates give scores between 6-10 in both gender, people tends to show sincere in speed dating to increase a second date chance. Since 6-10 is a relatively high score range, a second date is not easy to be seen on the plot. (6) Intelligent : almost all female give score within 4 to 10, while from the score that given by male, the dots are evenly distributed from 1 to 10. The more close two people give each other on “intelligence”, the higher chance they will have a second date.



1. Fun: second date cases appear in the region where both gender give partner a high score of “Fun”. Most data cluster in the area from 4 to 10 for “FunF” and 5 to 10 for “FunM”, on average, male give higher “Fun” score to female than female do.
2. Ambitious: there will probably be a second date when both gender rate the level of their partner higher than 5 on “Ambitious”, but still several observations occur in other region. When the total of AmbitiousM & F is less than 10, there will no secomd date chance.
3. SharedInterest: Most of the second date cases occur when both gender give a similar score to their partner, with the scores both higher than 5, but there’s one “second-date wish” case in which female rate 1 and male rate 10.

## Question 3

Many of the numerical variables are on rating scales from 1 to 10. Are the responses within these ranges? If not, what should we do these responses? Is there any missing data? If so, how many observations and for which variables?

## [1] 0 10

The responses are not within 1 to 10, in stead, all values of numerical variables are from 0 to 10, which means several observations have signed 0 as rating score. I suppose “0” score in this situation may cause by miswriting of candidates. By rating “0”, they imply the worst case under this variable. Therefore I chose to change it into 1, which is in the valid rating scale and at the same time indicats a worse case that they can experience.

## LikeM LikeF PartnerYesM PartnerYesF   
## Min. : 1.000 Min. : 1.000 Min. : 0.000 Min. : 1.000   
## 1st Qu.: 6.000 1st Qu.: 5.000 1st Qu.: 5.000 1st Qu.: 5.000   
## Median : 7.000 Median : 7.000 Median : 6.000 Median : 6.000   
## Mean : 6.682 Mean : 6.366 Mean : 5.757 Mean : 5.835   
## 3rd Qu.: 8.000 3rd Qu.: 8.000 3rd Qu.: 7.000 3rd Qu.: 7.000   
## Max. :10.000 Max. :10.000 Max. :10.000 Max. :10.000   
## NA's :2 NA's :4 NA's :4 NA's :4   
## AttractiveM AttractiveF SincereM SincereF   
## Min. : 1.000 Min. : 1.000 Min. : 1.000 Min. : 1.000   
## 1st Qu.: 5.000 1st Qu.: 5.000 1st Qu.: 7.000 1st Qu.: 7.000   
## Median : 7.000 Median : 6.000 Median : 8.000 Median : 8.000   
## Mean : 6.687 Mean : 6.274 Mean : 7.856 Mean : 7.784   
## 3rd Qu.: 8.000 3rd Qu.: 8.000 3rd Qu.: 9.000 3rd Qu.: 9.000   
## Max. :10.000 Max. :10.000 Max. :10.000 Max. :10.000   
## NA's :3 NA's :2 NA's :5 NA's :3   
## IntelligentM IntelligentF FunM FunF   
## Min. : 4.000 Min. : 2.000 Min. : 0.000 Min. : 1.000   
## 1st Qu.: 7.000 1st Qu.: 7.000 1st Qu.: 6.000 1st Qu.: 5.000   
## Median : 8.000 Median : 8.000 Median : 7.000 Median : 7.000   
## Mean : 7.621 Mean : 7.923 Mean : 6.863 Mean : 6.563   
## 3rd Qu.: 8.250 3rd Qu.: 9.000 3rd Qu.: 8.000 3rd Qu.: 8.000   
## Max. :10.000 Max. :10.000 Max. :10.000 Max. :10.000   
## NA's :8 NA's :3 NA's :6 NA's :6   
## AmbitiousM AmbitiousF SharedInterestsM SharedInterestsF  
## Min. : 2.000 Min. : 1.000 Min. : 0.000 Min. : 0.00   
## 1st Qu.: 5.000 1st Qu.: 6.000 1st Qu.: 4.000 1st Qu.: 4.00   
## Median : 7.000 Median : 8.000 Median : 5.000 Median : 6.00   
## Mean : 6.768 Mean : 7.429 Mean : 5.588 Mean : 5.47   
## 3rd Qu.: 8.000 3rd Qu.: 9.000 3rd Qu.: 7.000 3rd Qu.: 7.00   
## Max. :10.000 Max. :10.000 Max. :10.000 Max. :10.00   
## NA's :17 NA's :10 NA's :27 NA's :30

From the summary result above, variables “PartnerYesM / FunM / SharedInterestsM /SharedInterestsF” have zero(s) in the results.

After changing “0” response into “1”, check “NA” as below.

## DecisionM DecisionF LikeM LikeF   
## NA NA "NA's :2 " "NA's :4 "   
## PartnerYesM PartnerYesF AgeM AgeF   
## "NA's :4 " "NA's :4 " "NA's :3 " "NA's :5 "   
## RaceM RaceF AttractiveM AttractiveF   
## NA NA "NA's :3 " "NA's :2 "   
## SincereM SincereF IntelligentM IntelligentF   
## "NA's :5 " "NA's :3 " "NA's :8 " "NA's :3 "   
## FunM FunF AmbitiousM AmbitiousF   
## "NA's :6 " "NA's :6 " "NA's :17 " "NA's :10 "   
## SharedInterestsM SharedInterestsF second.date   
## "NA's :27 " "NA's :30 " NA

Missing data exist in variables except for “DecisionM/F”, “RaceM/F” and “second.date”. The details are reorganized and illustrated as the table below.

## NA number (from response by Male)  
## Like 2  
## PartnerYes 4  
## Age 3  
## Attractive 3  
## Sincere 5  
## Intelligent 8  
## Fun 6  
## Ambitious 17  
## SharedInterest 27  
## NA number (from response by Female)  
## Like 4  
## PartnerYes 4  
## Age 5  
## Attractive 2  
## Sincere 3  
## Intelligent 3  
## Fun 6  
## Ambitious 10  
## SharedInterest 30

“Ambitious” and “SharedInterests” responses from both male and female have most number of NAs.

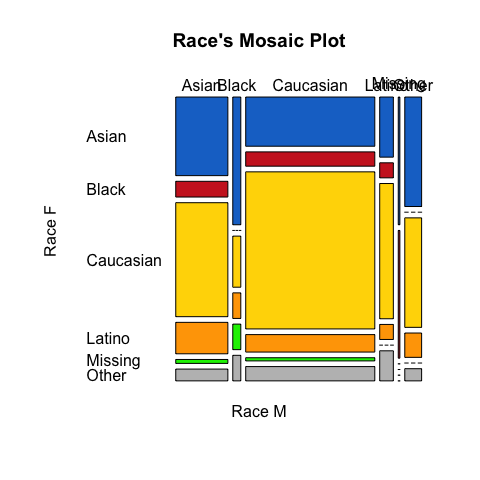
## Question 4

What are the possible race categories in your data set? Is there any missing data? If so, how many observations and what should you do with them? Make a mosaic plot with female and male race. Describe what you see.

## $`Female Race`  
##   
## Asian Black Caucasian Latino Other   
## 4 70 15 148 23 16   
##   
## $`Male Race`  
##   
## Asian Black Caucasian Latino Other   
## 2 65 10 161 17 21

## $`Missing data in Female`  
## [1] 29 30 66 169  
##   
## $`Missing data in Male`  
## [1] 166 167

In both gender’s data, there are several data with “no group”.  
There are 4 missing data in “RaceF”; 2 missing data in “RaceM”. I would like to keep them first, for the following two reasons: (1) We have no idea if we will use Race in the model, if no, then missing data will not matter;  
(2) In logistic regression, missing data will be automatically remove; (3) In mosaic plot, missing data still contains information about “what date group” is the missing data from. For example, if only male’s race (Black) was collected, we can tell one group with male’s race is Black has missing data. Otherwise, we will have no idea where missing data lies in.



From the mosaic plot: (1) Caucasian and Asian are the largest two portions of race in this case; (2) Date groups with one of the candidate that is Asian/Black/Caucasian missing race data missed; (3) There is no date match group in this case, with two people’s races are combination of (a.) Black male+Caucasian female, (b.) Other races male+Black female.

## Question 5

Use logistic regression to construct a model for “second.date” (i.e., “second.date” should be your response variable). Incorporate the discoveries and decisions you made in questions 2, 3, and 4. Explain the steps you used to determine the best model, include the summary output for your final model only, check your model assumptions, and evaluate your model by running the relevant hypothesis tests. Do not use “Decision” as an explanatory variable.

After trying the first regression with all variables, the regression results, as expected, are not significant for most variables. As we did in homework 3, I use R to do forward/ backward/ seqencial repeat regression automatically, and decide the “best” model in each method, after which, a final comparison is made to choose the final model.

## Loading required package: leaps

## regression RMSE adj.R.2 C.P BIC  
## 1 trial\_1 5.5844 0.1408 31.3196 -20.7634  
## 2 trial\_2 5.3518 0.2069 14.7760 -32.4841  
## 3 trial\_3 5.2141 0.2433 6.1712 -37.6119  
## 4 trial\_4 5.1417 0.2604 2.7066 -37.9061  
## 5 trial\_5 5.0868 0.2724 0.6123 -36.9028  
## 6 trial\_6 5.0461 0.2803 -0.3927 -34.8166  
## 7 trial\_7 5.0061 0.2880 -1.3240 -32.7020  
## 8 trial\_8 4.9784 0.2921 -1.3389 -29.6221  
## 9 trial\_9 4.9595 0.2938 -0.7131 -25.8508  
## 10 trial\_10 4.9433 0.2947 0.1214 -21.8568

I chose “trail\_4” and “trial\_5” since they have smaller BIC. From partial F-test below, H0 (the coefficient of RaceF is zero) should not be rejected, therefore, the final model by forward method includes LikeM/ FunF / PartnerYesM / PartnerYesF.

## Analysis of Deviance Table  
##   
## Model 1: second.date ~ LikeM + FunF + PartnerYesM + PartnerYesF  
## Model 2: second.date ~ LikeM + FunF + PartnerYesM + PartnerYesF + RaceF  
## Resid. Df Resid. Dev Df Deviance  
## 1 259 202.72   
## 2 254 194.81 5 7.909

Using same process, I do the same steps by backward method and sequential repeat method (the results are omitted). The two method have same “best” models with variables LikeM/ AttractiveF / PartnerYesM / PartnerYesF.

|  |  |  |
| --- | --- | --- |
|  | best by forward | best by backward/seqrep |
| AIC | 212.7203 | 213.7257 |
| Null.Deviance | 290.1376 | 292.3000 |
| Deviance | 202.7203 | 203.7257 |
| Null.d.f | 263.0000 | 267.0000 |

I made a criterion table, generally speaking, the two models have pretty similar value in listed criterion. Although AIC is smaller in forward method model, I chose the model with more remaining data, which is from backward method. It also has more significant output than that in forward method.

##   
## Call:  
## glm(formula = second.date ~ LikeM + PartnerYesM + AttractiveF +   
## PartnerYesF, family = "binomial", data = sdtemp)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.09987 -0.60706 -0.32265 -0.06487 2.36191   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -10.88106 1.47959 -7.354 1.92e-13 \*\*\*  
## LikeM 0.48336 0.12932 3.738 0.000186 \*\*\*  
## PartnerYesM 0.35057 0.10151 3.454 0.000553 \*\*\*  
## AttractiveF 0.35039 0.10223 3.427 0.000610 \*\*\*  
## PartnerYesF 0.27993 0.09566 2.926 0.003430 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 292.30 on 267 degrees of freedom  
## Residual deviance: 203.73 on 263 degrees of freedom  
## (8 observations deleted due to missingness)  
## AIC: 213.73  
##   
## Number of Fisher Scoring iterations: 6

### Assumptions Check

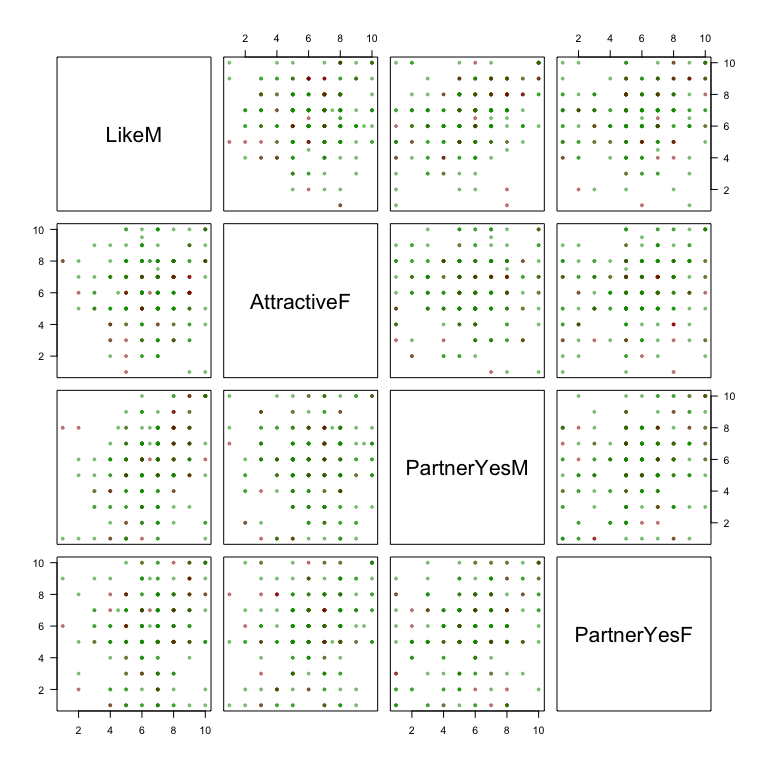
(1)Multicollinearity check

## Loading required package: usdm

## Loading required package: sp

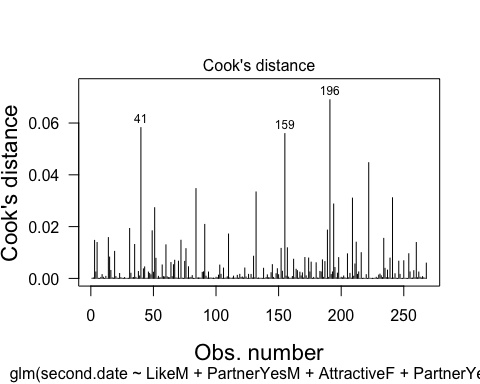
## Loading required package: raster

## Variables VIF  
## 1 LikeM 1.185748  
## 2 AttractiveF 1.055724  
## 3 PartnerYesM 1.239319  
## 4 PartnerYesF 1.139508

 VIFs are close to 1 and the estimates are all positive, which is not easy to identify from the pattern shown in scatter plot. But up to now, it seems colinearity problem is not a considerable one.

1. Fixed “X” without measurement error:  
   It is acknowledged because the data is from survey and only integers less than (or equal to) 10 are chosen, no measurement error can be made. Also, one the candidate give the result, the score will not be changed for a specific rating object. This assumption is satisfied.
2. Observations are independent: Our data are collected from indivudals that attending speed dating, the result they give can be said as independent.
3. Model is correctly specified: Only variables from the original dataset are chosen, no extraneous variables included in my model. But they are not the only variables that can be taken into consideration, later in the study, maybe other categories, such as, height, education level and be integrated to build a more complete model.
4. Outcomes can completely linearly separable: Satisfies. Since every candidates give one specific result in one category, the observations can be dtermined uniquely in multi-dimension space. Thus, the outcomes can completely separate from each other. And we can do glm() in R, which also means this assumption is satisfied.
5. No outliers:

## [1] 0.000 0.069

The range of Cook’s distance is (0.000, 0.0069)  Also from the plot, there’s no observations that has Cook’s distance larger than 0.1. The no outlier assumption is true.

1. Sample size

##   
## 0 1   
## 205 63

We have four variables in the model, so we need at least 40 observations when second.date equals 0 and 1, respectively. From the talbe outcome, where there are 63 observations for output “1” and 205 for output “0”, both of which are greater than 40. Therefore, sample size satisfied.

### Model Evaluation

1. log-likelihood for overall model  
   H0: all slopes in the final model(bw2) are zero H1: at least one slope in the final model(bw2) is not zero

## [1] 2.644361e-18

The calculated p-value is much smaller than 0.05, we can reject H0 and thus, the remaining variables are all significant in the model.

1. partial test Same as the steps made when built up a model, check if we should remain RaceF in our model. H0: the estimate of RaceF is 0 H1: the estimate of RaceF is not 0

## Analysis of Deviance Table  
##   
## Model 1: second.date ~ LikeM + PartnerYesM + AttractiveF + PartnerYesF  
## Model 2: second.date ~ LikeM + PartnerYesM + AttractiveF + PartnerYesF +   
## RaceF  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 263 203.73   
## 2 258 194.84 5 8.8867 0.1137

The p-value is larger than 0.05 to not reject null hypothesis. As the decision above, I dropped RaceF in my final model.

1. z-test for slopes for each variable, Ho: slope(beta) is equal to 0 H1: slope(beta) is not 0

## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -10.8810596 1.47959180 -7.354096 1.922240e-13  
## LikeM 0.4833617 0.12932106 3.737687 1.857208e-04  
## PartnerYesM 0.3505688 0.10151054 3.453521 5.533193e-04  
## AttractiveF 0.3503902 0.10223424 3.427328 6.095532e-04  
## PartnerYesF 0.2799276 0.09565936 2.926296 3.430245e-03

The p-value for each variables are all smller than 0.01, they are all significant to reject the null hypothesis.

To conclude, this model seems to be a good fit for the data. And my final model is P(have a second date | LikeM, PartnerYesM, PartnerYesF, AttractiveF)= exp^(-10.8811 + 0.4834LikeM + 0.3506PartnerYesM + 0.2799PartnerYesF + 0.3504AttractiveF)/ (1+exp^(-10.8811 + 0.4834LikeM + 0.3506PartnerYesM + 0.2799PartnerYesF + 0.3504AttractiveF))

## Question 6

Redo question (1) using only the observations used to fit your final logistic regression model. What is your sample size? Does the number of explanatory variables in your model follow our rule of thumb? Justify your answer.

## [1] 268

This time, sample size is 268.

##   
## Decision Made by Female- NO  
## Decision Made by Male- NO 63  
## Decision Made by Male- YES 81  
##   
## Decision Made by Female- YES  
## Decision Made by Male- NO 61  
## Decision Made by Male- YES 63

## [1] 0.2350746

## B  
## 0 1   
## 205 63

This result follow the rule of thumb– at least 40 (4 explanatory variables) observations for each response result.

## Question 7

Interpret the slopes in your model. Which explanatory variables increase the probability of a second date? Which ones decrease it? Is this what you expected to find? Justify.

## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -10.8810596 1.47959180 -7.354096 1.922240e-13  
## LikeM 0.4833617 0.12932106 3.737687 1.857208e-04  
## PartnerYesM 0.3505688 0.10151054 3.453521 5.533193e-04  
## AttractiveF 0.3503902 0.10223424 3.427328 6.095532e-04  
## PartnerYesF 0.2799276 0.09565936 2.926296 3.430245e-03

1. Intercept

## second.date LikeM AttractiveF PartnerYesM   
## Min. :0.0000 Min. : 1.000 Min. : 1.00 Min. : 1.000   
## Max. :1.0000 Max. :10.000 Max. :10.00 Max. :10.000   
## PartnerYesF   
## Min. : 1.00   
## Max. :10.00

When all the variables are zero, the probability that the two persons will have second date is exp(-10.8811)/(1+exp(-10.8811)), or 1.88\*10^-5. Since the rating levels are from 1 to 10, the case where “LikeM = AttractiveF = PartnerYesM = PartnerYesF = 0” is impossible. Therefore the interpretation of intercept is meaningless.

1. LikeM

## [1] 0.8562185

If LikeM ranking score increases by 1, holding AttractiveF, PartnerYesM and PartnerYesF values constant, the odds of having a second date increases by 85.62%. (The calculation steps are: first, use slope os LikeM to get log odds, exp(0.4834)/(1+exp(0.4834)); then, calculate percentage change by exp(.)-1 )

1. AttractiveF

## [1] 0.798067

If AttractiveF ranking score increases by 1, holding LikeM, PartnerYesM and PartnerYesF values constant, the odds of having a second date increases by 79.81%.

1. PartnerYesM

## [1] 0.7981449

If PartnerYesM ranking score increases by 1, holding LikeM, AttractiveF and PartnerYesF values constant, the odds of having a second date increases by 79.81%.

1. PartnerYesF

## [1] 0.7674335

If PartnerYesF ranking score increases by 1, holding LikeM, AttractiveF and PartnerYesM values constant, the odds of having a second date increases by 76.74%.

All the variables included in my final model increase the probability of a second date. It is consist with my expectation. In my model, all the variables are positive-judgement oriented. Because the more they like their dates (LikeM is higher), the more attractive they think their dates are (AttractiveF is higher) and the more confidence in the partner saying “yes” (both PartnerYesM and PartnerYesF are higher), the chance of second date will increase. Therefore I think my model is also reasonable.

## Question 8

Construct an ROC curve and compute the AUC. Determine the best threshold for classifying observations (i.e., second date or no second date) based on the ROC curve. Justify your choice of threshold. For your chosen threshold, compute (a) accuracy, (b) sensitivity, and (c) specificity.

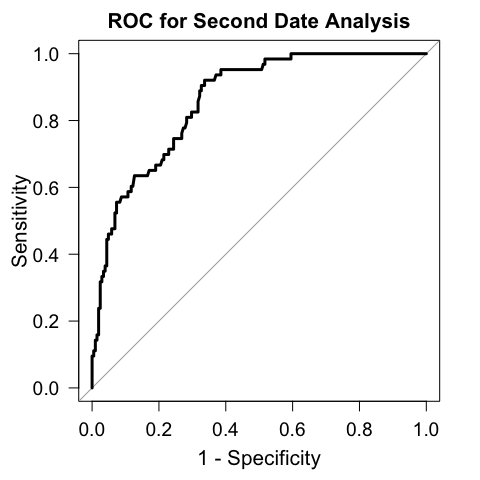
## Loading required package: pROC

## Warning: package 'pROC' was built under R version 3.4.4

## Type 'citation("pROC")' for a citation.

##   
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':  
##   
## cov, smooth, var



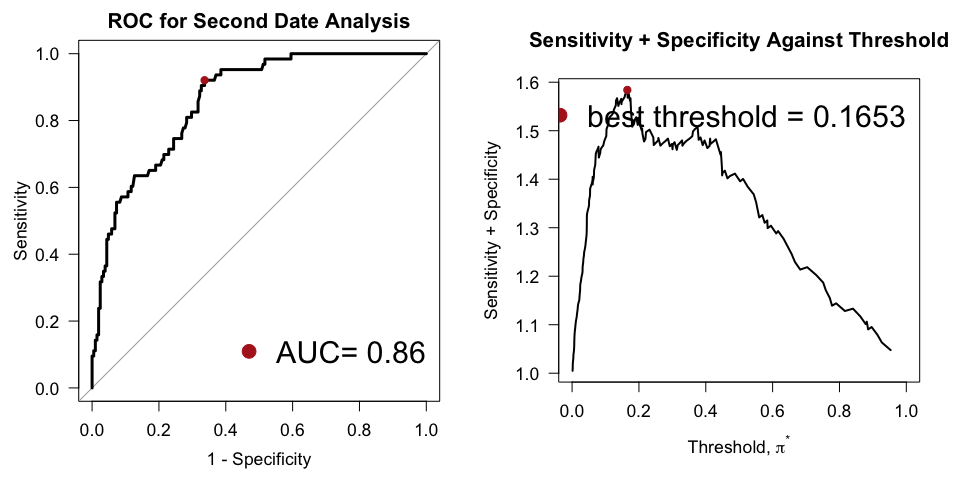
##   
## Call:  
## roc.default(response = aaa$second.date, predictor = bw2$fitted.values, plot = TRUE, las = TRUE, lwd = 3, legacy.axes = TRUE, main = "ROC for Second Date Analysis", cex.main = 1.3, cex.axis = 1.2, cex.lab = 1.3)  
##   
## Data: bw2$fitted.values in 205 controls (aaa$second.date 0) < 63 cases (aaa$second.date 1).  
## Area under the curve: 0.8602

The AUC of the ROC curve is 0.8602. (The ROC curve is plotted below). In this problem, we want to have less false positive(FP) predictions, i.e, we predict a group will have a second date but in fact they don’t want to date again, as well as true negative(TN) predictions, i.e, we correctly predict a group do not want a second date. I first choose the threshold that will give highest sum of specificity and sensitivity.

## threshold specificity sensitivity   
## 0.1653049 0.6634146 0.9206349

## threshold specificity sensitivity  
## all -Inf 0.000000000 1  
## all 0.001458683 0.004878049 1  
## all 0.001793114 0.009756098 1  
## all 0.001988319 0.014634146 1  
## all 0.002067037 0.019512195 1  
## all 0.002717235 0.024390244 1

##   
## Call:  
## roc.default(response = aaa$second.date, predictor = bw2$fitted.values, plot = TRUE, las = TRUE, lwd = 3, legacy.axes = TRUE, main = "ROC for Second Date Analysis", cex.main = 1.3, cex.axis = 1.1, cex.lab = 1.1)  
##   
## Data: bw2$fitted.values in 205 controls (aaa$second.date 0) < 63 cases (aaa$second.date 1).  
## Area under the curve: 0.8602



## actual.2nddate  
## classify.50 2nd.date no 2nd.date  
## 2nd.date 58 69  
## no 2nd.date 5 136

## threshold specificity sensitivity accuracy  
## 1 0.1653049 0.6634146 0.9206349 0.7238806

From the table, TP=58, NP=5, FN=69, TN=136,

accuracy=(TP+TN)/(P+N)=(58+136)/268=0.7239

We can directly get sensitivity and specificity through former “threshold choosing process”, but to inspect, sensitivity=TP/P=58/(58+5)=0.9206 specificity=TN/N=136/(136+69)=0.6634