Freemium is not Free:

Regression and Modeling of money spent on Fire Emblem: Heroes

Authors: Jiarui Yang, Ernie Yeung

Abstract: Video gaming, like other forms of modern entertainment, is an industry aiming to profit. Many modern video games take up a business model known as the freemium model that is free to access but has paid features. We want to find out what factors go into how gamers spend their money; utilizing model building diagnostics like the BIC, we fit and build a model using data from a fan-made survey on the Fire Emblem: Heroes subreddit to see what factors goes into how much players spend on a single video game.

Introduction

With technology and mass media taking the forefront of our daily lives, we as consumers spend more and more time with our personal devices. From reading work emails to trying out the latest fads, the smartphone has become integral to our daily lives, providing us with a means of communication as well as a source of entertainment. One such source of entertainment comes from video games. 10 years ago, one who indulged in video games would be considered a big fan, but now, the entry bar has dropped significantly, and access to video games has become quite literally 'one tap away' on a device of choice. As with all other forms of entertainment, success as a creator within a medium is associated with money and profits. In 2016, the video game industry hit a total revenue of 25.4 Billion dollars in just the United States, and this number only increases on a year by year basis. Our project looks at the population of people who spend money on video games, and how the characteristics of players can be used to estimate how much money they spend on one particular video game. In particular, our main research question is the following: "What constitutes how gamers spend money on video games?". To answer this question, we will

be using data taken from a survey taken by a user on Reddit, which was posed to players of the mobile game “Fire Emblem: Heroes”.

Description

This dataset was created from a survey taken on the Fire Emblem: Heroes subreddit community via Google Forms by a Reddit user by the name /u/ShiningSolarSword. This user has conducted numerous surveys on this community in the past, and this particular survey was named “The Ninth Great FE:H Demographics and Opinions Survey” and posted on the FE:H subreddit on December 11th, 2017, taking responses until December 18th, 2017. As the survey name implies, this was not the first survey of its type to be done on this community. The survey was by no means required, and participation was 100% voluntary. The only prior analysis applied on this dataset involves a lot of summary statistics and some comparisons to the results of previous surveys. The work of the surveyor on this particular dataset can be found in the references. This is the first time a regression was attempted on this dataset.

Variable descriptions

There are total 9 variables in our cleaned data set, and we decided to treat money spent as the response variable.

playingTime: Playing time of your main account in Fire Emblem: Heroes after release - measured in days

money: Money spent on FE:Heroes after release - measured in US dollars

age: What is your age - measured in years

gender: What is your gender - Male, Female, Non-binary, or witheld

enjoyment: Enjoyment rating of a recent in-game event - (1-5 rating)

timeSatisfication: Satisfaction rating of temporal investment into FE:Heroes - (1-5 rating)

moneySatisfication: Satisfaction rating of monetary investment into FE:Heroes - (1-5 rating)

arena: The current Arena tier of your main account for the season beginning Dec 12th.

tempest: The highest milestone you achieved in the recent tempest trials event.

There are total 4078 observations in our data set. Since our main question is: what factors will affect the money spent on video game among people who have spent money, we removed observations for people who decided not to spend any money at all. Furthermore, our data is from a survey, so there is some nonresponse, leaving us with some missing values. We have decided to remove all nonresponse. After we remove 0s in the money spent variable and all missing values, there are 2133 observations left. A lot of the data is binned, so we have remedied this by taking the midpoint.

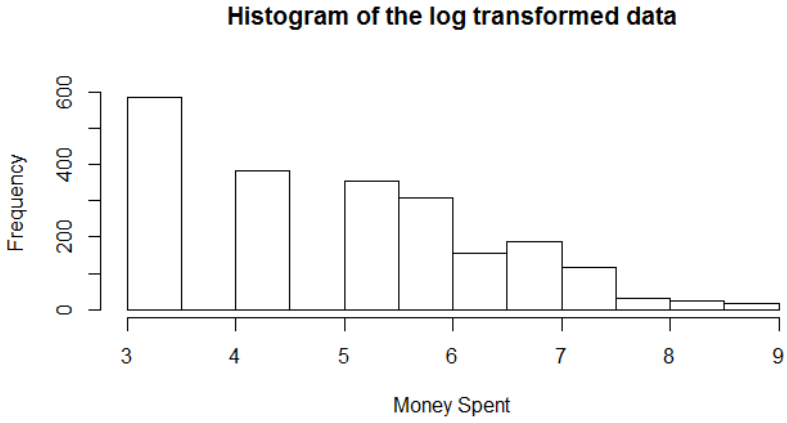
Here is the head of our cleaned data:

	playingTime	money	age	gender	enjoyment	timeSatisfication	moneySatisfication	arena	tempest
1	313	1500	17	Male	5	5	4	19	99999
2	313	300	23	Male	4	5	4	20	99999
3	296	1500	23	Male	4	5	4	16	40000
4	313	300	23	Male	3	5	4	19	99999
5	313	25	20	Male	2	4	2	19	99999
6	313	300	23	Male	3	4	1	19	99999

Transformations:

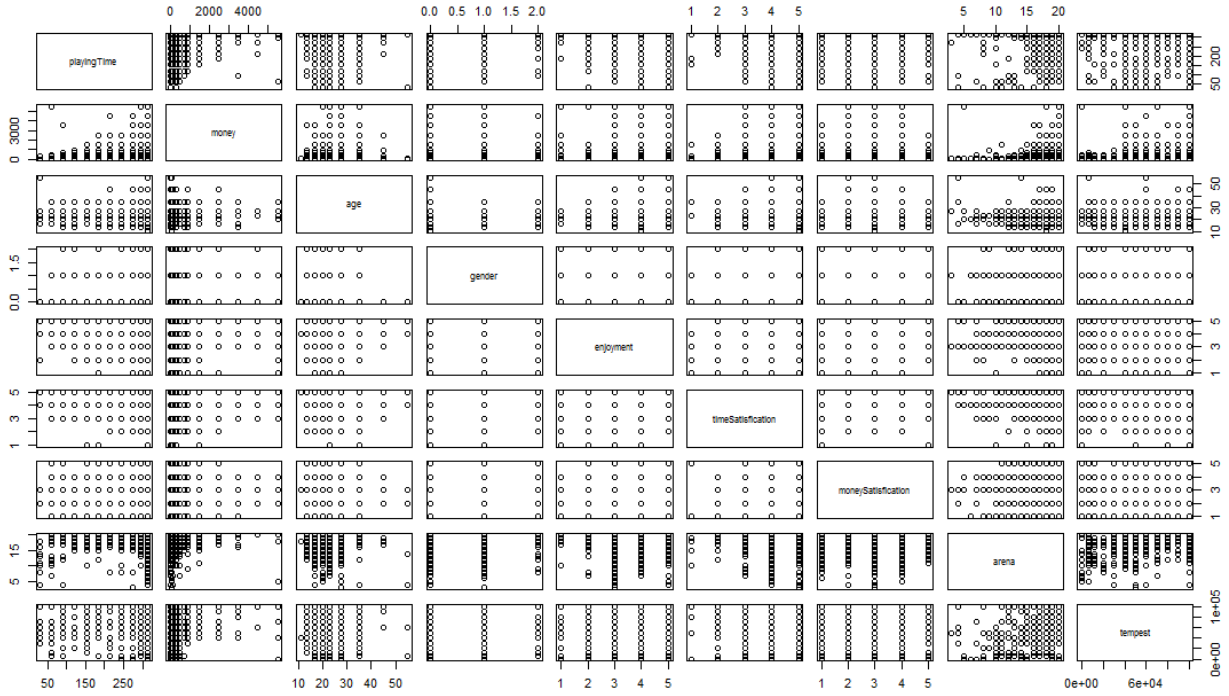
The histogram of the untransformed data can be found in the Appendix (figure 1). As we can see, the histogram of money spend is highly right skewed. Therefore, we need to move down the ladder: compresses the larger X's and stretches out the smaller values. Symbox of money spent is also included in figure 1.

Although the symbox shows that p=0.5 is better than log transformation, when we actually tried those two options, and the log transformation better corrects more skewness than p=0.5:



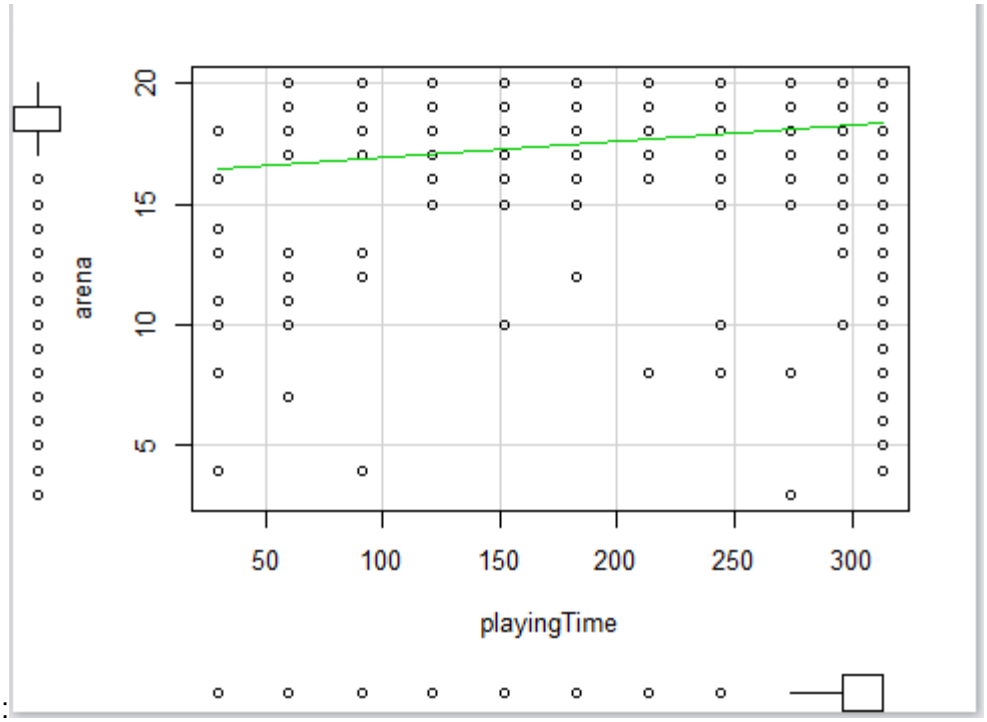
Collinearity:

We also wanted to check if there exists substantial collinearity between any regressors. If strong collinearity exists, then it would create redundancy in our model as they would be the same thing presented in a different manner, and we would need to decide whether or not we should eliminate any regressors.



Scatterplot matrices are intrinsically difficult to visualize, but are very useful to examine marginal associations between pairs of variables. A brief glance through this matrices, we do not see much collinearity between any of the explanatory variables.

When we firstly clean the data set, we suspect that there exist some collinearity between playingTime and arena, since it is common sense for players who play more tends to have higher ranks/levels. So we took a look at the scatterplot of those two variables



:

To make a better informed decision, we decide to use Variance Inflation Factor(VIF) as a criterion for high collinearity. VIF measures how much the variance of the estimated regression coefficient is inflated by collinearity. Generally, a VIF greater than 2 indicates significant collinearity. Below is a table showing the VIFs of the different variables.

	GVIF	Df	GVIF^(1/(2*Df))
playingTime	1.04	1	1.02
age	1.03	1	1.01
gender	1.01	3	1.00
enjoyment	1.12	1	1.06
timeSatisfied	1.29	1	1.13
moneySatisfied	1.22	1	1.10
arena	1.40	1	1.18
tempest	1.35	1	1.16

Looking through the last column, we can see that none of the variables have strong collinearity. Therefore, our suspicions are invalid.

Statistical Analysis

The full model:

$$\log(money) = \beta_0 + \beta_1 playingTime + \beta_2 age + \gamma_1 gender[Male] + \gamma_2 gender[Non\text{-}binary] + \beta_4 enjoyment + \beta_5 timeSatisfication + \beta_6 moneySatisfication + \beta_7 arena + \beta_8 tempest$$

The summary of our full model:

```
Call:
lm(formula = log(money) ~ playingTime + age + gender + enjoyment +
    timeSatisfication + moneySatisfication + arena + tempest,
    data = data)

Residuals:
   Min     1Q Median     3Q    Max
 -2.79  -1.12  -0.07   0.99   5.43

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         9.8e-01    3.7e-01     2.6    0.008 **
playingTime         8.2e-04    5.5e-04     1.5    0.135
age                 5.2e-02    5.7e-03     9.1   <2e-16 ***
genderMale          8.1e-03    8.1e-02     0.1    0.921
genderNon-binary    2.0e-01    2.3e-01     0.9    0.379
enjoyment          -3.5e-02    3.3e-02    -1.1    0.292
timeSatisfication   1.9e-02    4.4e-02     0.4    0.668
moneySatisfication  7.4e-02    2.9e-02     2.6    0.011 *
arena               1.2e-01    1.7e-02     7.5    1e-13 ***
tempest             1.6e-06    1.2e-06     1.3    0.202
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.3 on 2123 degrees of freedom
Multiple R-squared:  0.086,     Adjusted R-squared:  0.083
F-statistic:  22 on 9 and 2123 DF,  p-value: <2e-16
```

According to the summary of full model, we can see that about 8.6% of the variance in money spent is explained by the full model. Note that X variables such as playingTime, gender, enjoyment, timeSatisfication, and tempest are considered non-significant since they have p-values larger than 0.05.

Model selection strategy:

Our goal is to create a multivariate regression model that best characterizes the relationship between money spent and other variables, therefore we are more interested in the combined rather than individual effect of the regressors. Thus we used a model selection methods to help us determine the best combination of the variables. There are many algorithms used to compare the

performance of different combinations of variables such as calculating the BIC, adjusted R-squared, and Cp. All three plots can be found in the appendix.

For the BIC (figure 2), a good model has small values of BIC. So the best model for our data set by BIC standards should include 3 explanatory variables: age, moneySatisfied, and arena. This happens to be the exact model we pick.

For the adjusted R-squared (figure 3), we can see that the largest adjusted R-squared is equal to 0.083. Interestingly enough, many of the models have the same R-squared value. Notice that the model containing age, moneySatisfied and arena shares an R-squared value extremely close to the best model displayed.

For the Cp (figure 4), a good model has Cp ≈ number of X variables in the model + 1. So we can see our model of choice: age. moneySatisfied and arena on the second row, with a Cp value of 3.3, which is quite close to the criterion of a good Cp value.

Comparing those three methods, we decide to include 3 explanatory variables: age, moneySatisfication, and arena in our reduced model, which make sense since from the summary of full model, the X variables except those three are all non-significant. Therefore our reduced model is the following:

$$\log(\text{money}) = \beta 0 + \beta 2\text{age} + \beta 6\text{moneySatisfication} + \beta 7\text{arena}$$

We can check whether our model selection is valid by applying an anova F test between reduced and full model with the following hypotheses:

$$H0: \beta 1 = \beta 4 = \beta 5 = \beta 8 = 0$$

H1: at least one is not zero

Decision rule: reject null iff p-value is less than 0.05.

Test statistic:

```
Analysis of Variance Table

Model 1: log(money) ~ age + moneySatisfication + arena
Model 2: log(money) ~ playingTime + age + gender + enjoyment
+ timeSatisfication +
    moneySatisfication + arena + tempest
  Res.Df  RSS Df Sum of Sq   F Pr(>F)
1   2129 3848
2   2123 3839  6       9.8 0.9    0.5
```

Since p-value is greater than 0.05, we fail to reject null. There is not enough evidence to show that at least one of those β/γ's is not zero. Hence, the our model selection is valid and the reduced model is significant.

Summary of the reduced model:

```
Call:
lm(formula = log(money) ~ age + moneySatisfication + arena, data = data)

Residuals:
   Min     1Q Median     3Q    Max
 -2.77  -1.12  -0.07   1.01   5.55

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.0403     0.2918     3.6    4e-04 ***
age                 0.0512     0.0056     9.1   <2e-16 ***
moneySatisfication  0.0768     0.0263     2.9    0.004 **
arena               0.1386     0.0140     9.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.3 on 2129 degrees of freedom
Multiple R-squared:  0.084,     Adjusted R-squared:  0.083
F-statistic:  65 on 3 and 2129 DF,  p-value: <2e-16
```

In the reduced model, 8.4% variance of money spend is explained by those X variables. And the adjusted R-squared is 0.083, which is same as the adjusted R-squared in the full model. We have essentially shown that some of our variables are completely insignificant.

Weighted model

Looking at the reduced model above, at first glance, it seems quite unsettling to see a fit with such a low r squared value, so we decided to take a look at weighted least squares models. This method is from ch 12 of our textbook, and some of the code used is taken from the notes on ch 12. Our data was almost completely binned to begin with, so looking at a potential WLS fit was extremely convenient. We had tested a number of different weights, and we found out that our age variable turned out to have an increasing variance as age increased. A table showcasing the variance and its difference on a group to group basis can be seen below.

Group	Age	Sample Size	Var(Y|age)	Difference	Weights
1	< 12	1	NA	-	NA
2	12-15	64	208422.6	NA	4.8E-06
3	16-18	370	162224.4	-46198.2	6.16E-06
4	19-21	542	368819.8	206595.4	2.71E-06
5	22-24	531	606529.1	237709.3	1.65E-06
6	25-30	496	739582	133052.9	1.35E-06
7	31-40	117	1436016	696434	6.96E-07
8	41-50	8	726964.3	-709051.7	1.38E-06
9	51+	2	1250	-725714.3	0.0008

As we can see, barring the difference between group 3 and 4, there is a general positive difference as we move down the groups as seen in the 'Difference' column. Note that group 1 has only 1 observation, so it's variance is undefined and groups 8 and 9 have an extremely small sample size, making its variance somewhat unreliable compared with the other groups. Now, using the inverse of the variance as weights, we can take a look at the weighted least squares model below:

```
Call:
lm(formula = log(money) ~ age + moneySatisfied + arena, data = data,
    weights = 1/sg2)

Weighted Residuals:
   Min     1Q Median     3Q    Max
 -8.74  -1.76  -0.07   1.81  10.46

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.0092     0.2674     3.8    2e-04 ***
age             0.0252     0.0038     6.7    3e-11 ***
moneySatisfied  0.0871     0.0262     3.3    9e-04 ***
arena           0.1685     0.0107    15.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.3 on 2129 degrees of freedom
Multiple R-squared:  0.11,      Adjusted R-squared:  0.11
F-statistic:  89 on 3 and 2129 DF,  p-value: <2e-16
```

Compared with the unweighted model, above we can already see a 30% increase in the r squared value which is an extremely beneficial result. However, we can see that the residual standard error has increased from 1 to 2.3, which is a loss we are willing to take.

Results

Looking at the summary of the WLS model above, we can see that generally speaking, while our variables do have some effect on the response, our variables do not have a particularly large effect on the response. In particular, we can say that per each unit change change in the arena rank, we see roughly a 16.9% increase in money spent. Note that given the context of this variable, it is indeed possible to increase one's arena rank while holding other variables constant. The interesting fact about the arena rank is that putting it in context of this game, a higher arena rank often denotes a higher level of dedication to the game, as one cannot increase his arena rank without playing at minimum weekly. This follows to say that the more dedicated the player, the more he spends on video games. Now looking at age, we can see that as a player gets older, we see a 2.5% increase in money spent per year on average. This makes sense; since this survey covers a population from ages 11 to 50+, a positive trend is completely realistic, as people tend to make more money as they grow older, thus they have more disposable income, thus an increase in spending on hobbies such as video games. Next, looking at the monetary satisfaction variable, we can see that for unit change in the variable, money spent increases about 8.7% on average. This too makes sense - when a customer is satisfied with his or her purchase, he or she is more likely to return and purchase more. Lastly, we have the intercept, which does not make sense to interpret as an age of 0 or an arena rank of 0 would imply a person is not born or the player does not play the game respectively which is impossible.

Discussion

Looking at the results of our model, we can say that all the variables that showed significance in the final model tend to fit current trends and make sense; people who fit these characteristics - older people, people who are satisfied with their monetary investment, and dedicated players tend to spend more money on average. One interesting result is how gender was not a factor in this regression. Video gaming is typically known as a predominantly male hobby as seen by data taken by Statistia, and one would expect males to spend more on video games given typical spending habits between different genders. However, the data said otherwise, and people male, female and non-binary had indistinguishable spending habits on video games. A boxplot can be seen in the Appendix (figure 5).

Some limitations of this project include the fact that a lot of data was collected in bins, so the accuracy of our results definitely have some error due to a lack of accuracy in the data. Furthermore, having the data in bins also made it difficult to visualize the data as there were many repeated values. Another limitation in our data was the fact that this survey was taken on a specific community - the FE:H subreddit. While this particular community has a consistent viewership of at least 100,000 with at least 3,000 concurrent users at one given moment, this community is restricted in a sense that it is predominantly people who browse Reddit at all, and are English-speaking. FE:H is an internationally available video-game, so it would not be correct to say that this dataset is representative of the entire playerbase.

Video games have only recently broken into the mainstream within the past few years, so as it stands, there are many opportunities to look into player spending habits in games to see what and how players spend their money. FE:H is just one fish in an ocean of video games released, and while it is one of the more successful games out there, one may find different results on a game of a different genre or on another platform.

References:

Prior analysis of this dataset:

https://www.reddit.com/r/FireEmblemHeroes/comments/7kkznl/results_of_the_ninth_great_feh_demographics_and/

Consumer gender distribution over the years

https://www.statista.com/statistics/232383/gender-split-of-us-computer-and-video-gamers/

Ch12 lectures notes - Stat 423

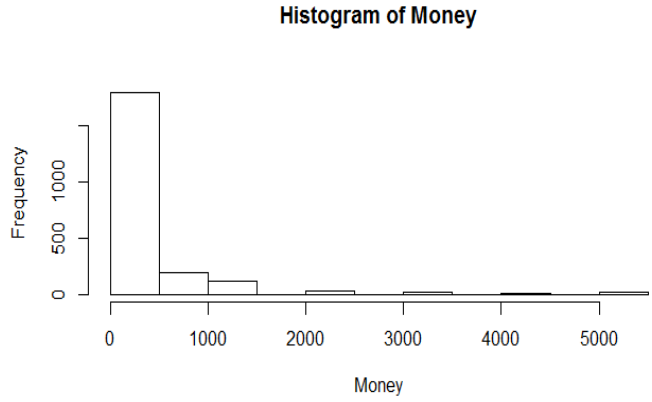
Appendix

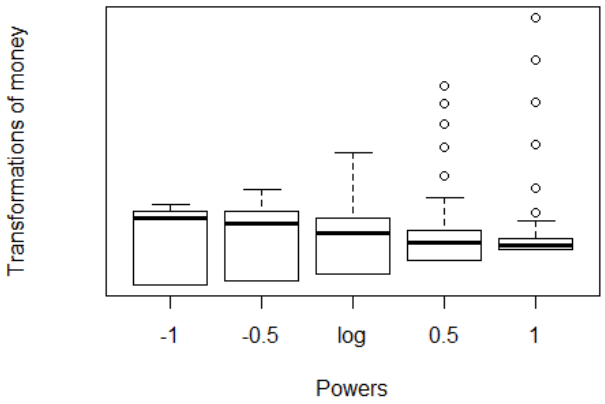
```
library(leaps)
library(car)
library(alr4)
nonzero <- data[,2] != 0
data <- data[nonzero,] #remove zero
data <- na.omit(data) #remove na

#show head
head(data)

#histogram of untransformed money
hist(data$money,main="Histogram of Money", xlab="Money")
```

Figure 1






```
#Symbox
symbox(~money, data=data)

#histogram of log transformed money
hist(log(data$money), main="Histogram of the log transformed data",
     xlab="Money Spent") #right skewed

#collinearity
scatterplot(arena~playingTime, data=data,smoother=F)
```

```r
#summary for full model
yfit_full<-lm(log(money)~playingTime+age+gender+enjoyment+
                timeSatisfied+moneySatisfied+arena+tempest,data=data)
print(vif(yfit_full),digits=3)
print(summary(yfit_full), digits=2)

#model selection
sub<-regsubsets(log(money)~.,data=data,nvmax=9,nbest=1,method="exhaustive")
plot(sub,scale="bic", main="BIC")
plot(sub,scale="adjr2", main="Adjusted R^2")
plot(sub,scale="Cp", main="Cp")
```

Figure 2

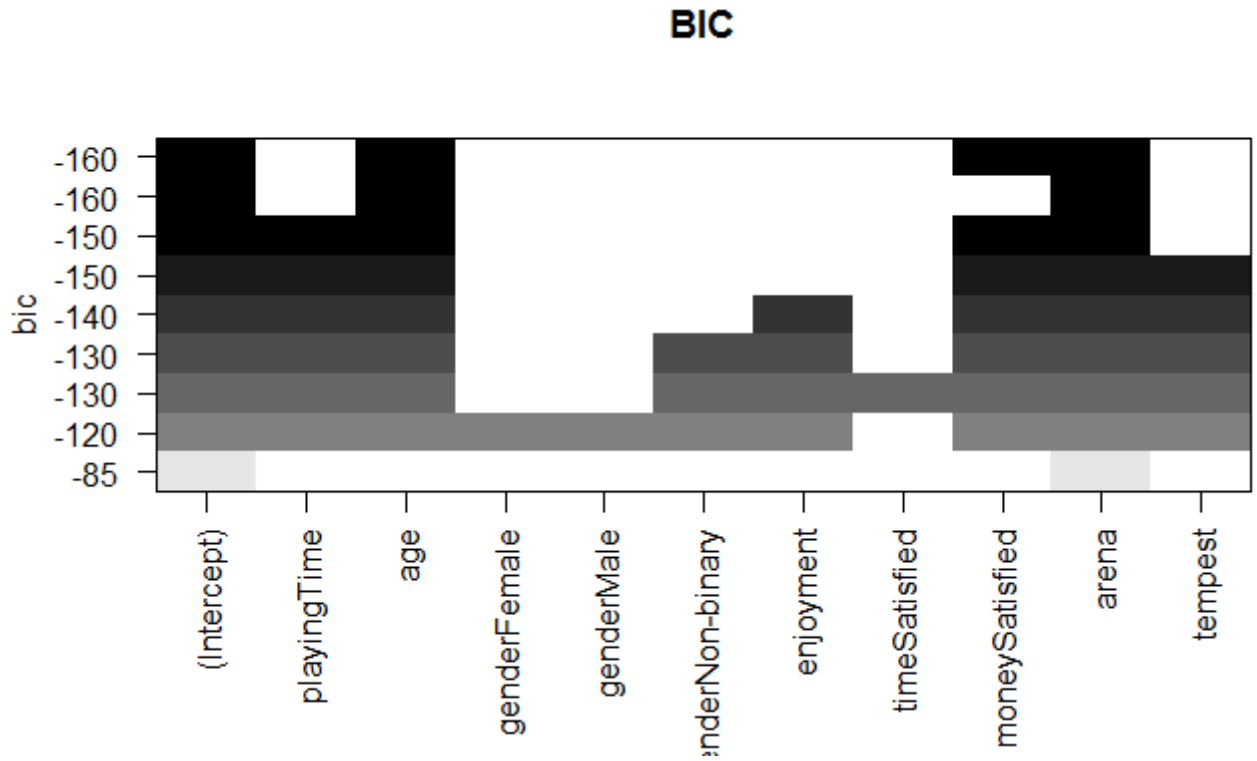


Figure 3

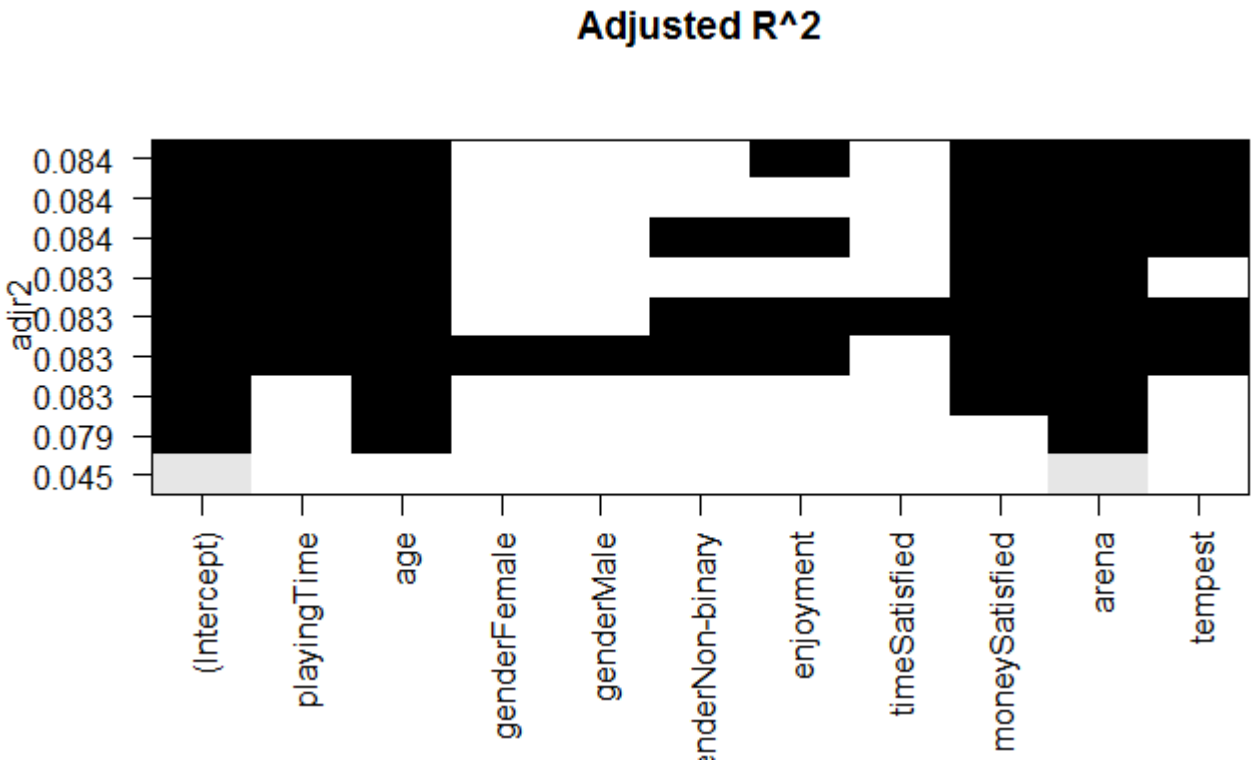
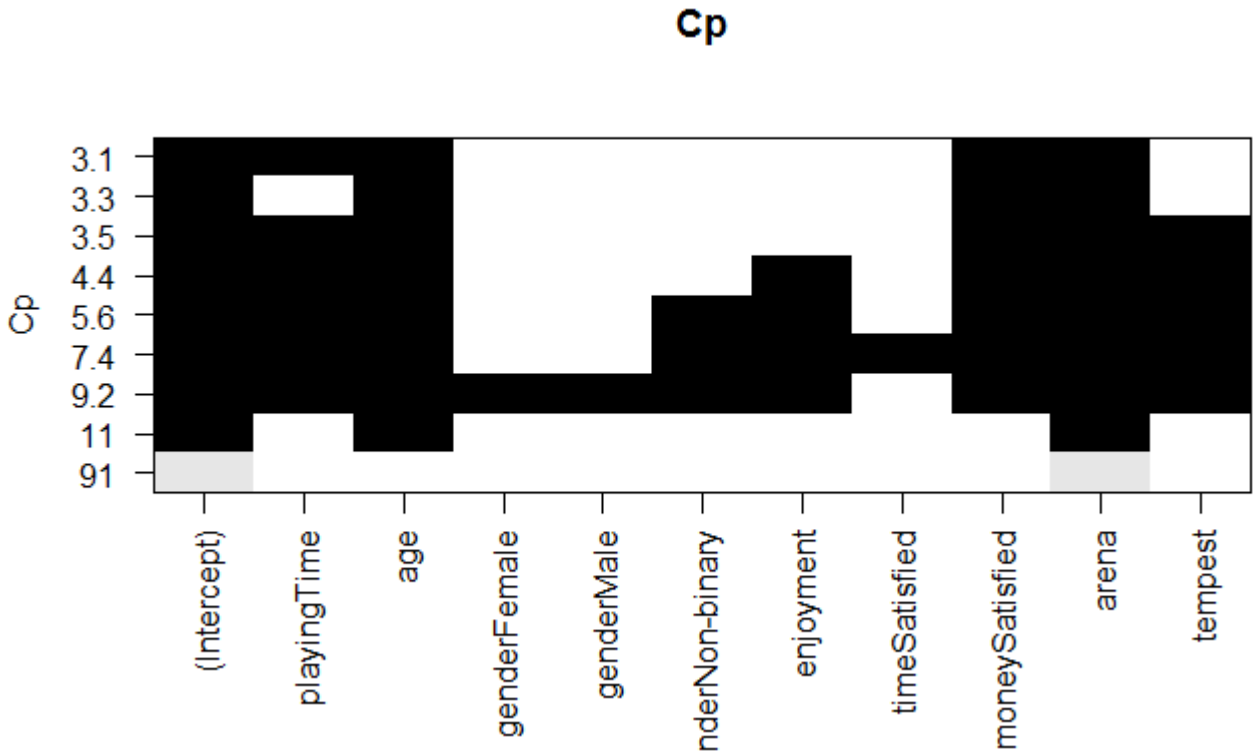


Figure 4



```r
#f test
yfit_red=lm(log(money)~age+moneySatisfied+arena,data=data)
print(anova(yfit_red,yfit_full), digits=2)

#summary for reduced model
print(summary(yfit_red), digits=2)


#weighted LS
lm.full <- lm(log10(money)~playingTime+age+as.factor(gender)+
                enjoyment+timeSatisfied+moneySatisfied+arena+tempest,
              data=data)
resvar <- aggregate(lm.full$residuals,by=list(data$age),FUN=var)
sg2 <- resvar$x[as.factor(data$age)]
sg2[is.na(sg2)] <- 1
lm.wtfull <- lm(log(money)~age+moneySatisfied+arena,data=data,weights=1/sg2)
print(summary(lm.wtfull),digits=2)


#boxplot of genders
boxplot(log(data$money)~data$gender,
        main="Boxplot of Money spent vs Gender", ylab="log Money")
```

Figure 5