Regression Assignment–DirectMarket

Saixiong Han (sah178)

January 16, 2017

1. Identify and report response variable and predictor

According to the description on the DirectMarket data,the objective is to explain AmountSpent in terms of the provided customers characteristics. Therefore, Response variable is AmountSpent. Predictors are Age, Geander, Ownhome, Married, Location, Salary, Children, History, Catalogs

2.Explore data and generate summary

-Import and Overview

```
DMarket <- read.csv("C:/Study/DM/week2/DirectMarketing.csv", header = TRUE,
    sep = ",", stringsAsFactors = TRUE)
head(DMarket)
```

```
##      Age Gender OwnHome Married Location Salary Children History Catalogs
## 1    Old Female     Own  Single      Far  47500        0    High        6
## 2 Middle   Male    Rent  Single    Close  63600        0    High        6
## 3  Young Female    Rent  Single    Close  13500        0     Low       18
## 4 Middle   Male     Own Married    Close  85600        1    High       18
## 5 Middle Female     Own  Single    Close  68400        0    High       12
## 6  Young   Male     Own Married    Close  30400        0     Low        6
##   AmountSpent
## 1         755
## 2        1318
## 3         296
## 4        2436
## 5        1304
## 6         495
```

```
summary(DMarket)
```

```
##      Age         Gender    OwnHome       Married     Location
##  Middle:508   Female:506   Own :516   Married:502   Close:710
##  Old   :205   Male  :494   Rent:484   Single :498   Far  :290
##  Young :287
##
##
##
##      Salary          Children       History       Catalogs
##  Min.   : 10100   Min.   :0.000   High  :255   Min.   : 6.00
##  1st Qu.: 29975   1st Qu.:0.000   Low   :230   1st Qu.: 6.00
##  Median : 53700   Median :1.000   Medium:212   Median :12.00
##  Mean   : 56104   Mean   :0.934   NA's  :303   Mean   :14.68
##  3rd Qu.: 77025   3rd Qu.:2.000                3rd Qu.:18.00
##  Max.   :168800   Max.   :3.000                Max.   :24.00
##   AmountSpent
```

```
##  Min.   :  38.0
##  1st Qu.: 488.2
##  Median : 962.0
##  Mean   :1216.8
##  3rd Qu.:1688.5
##  Max.   :6217.0
```

```
sum(is.na(DMarket))
```

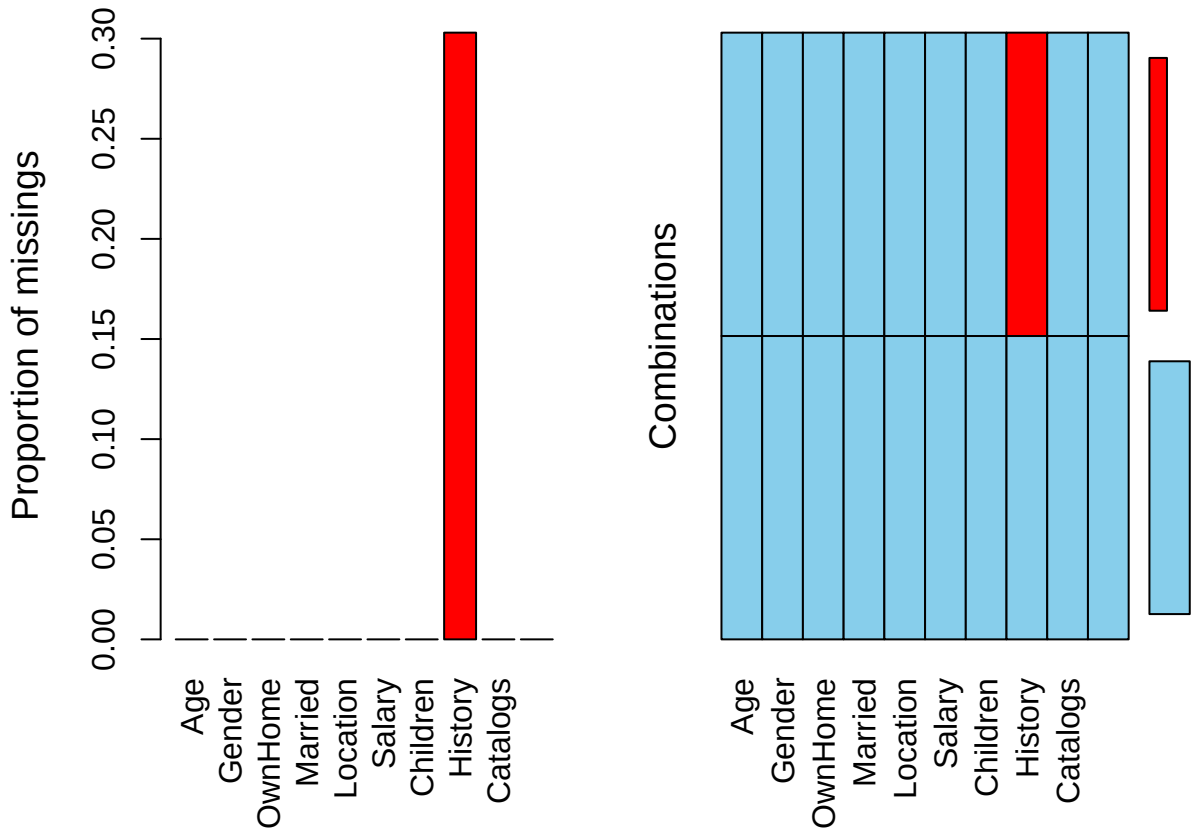
```
## [1] 303
```

From the summary we can see that, there are 303 missing values in History attributes. Since the total record is 1000, 303 missing value is nearly one third of the total amount. Therefore, I think it's better to use the mutiple imputation instead of ignore them.

a–Deal with missing data

In order to know the distribution of missing data, the first thing I would like yo do is spelling the pattern of missing data.

```
library(VIM)
library(mice)
aggr(DMarket)
```



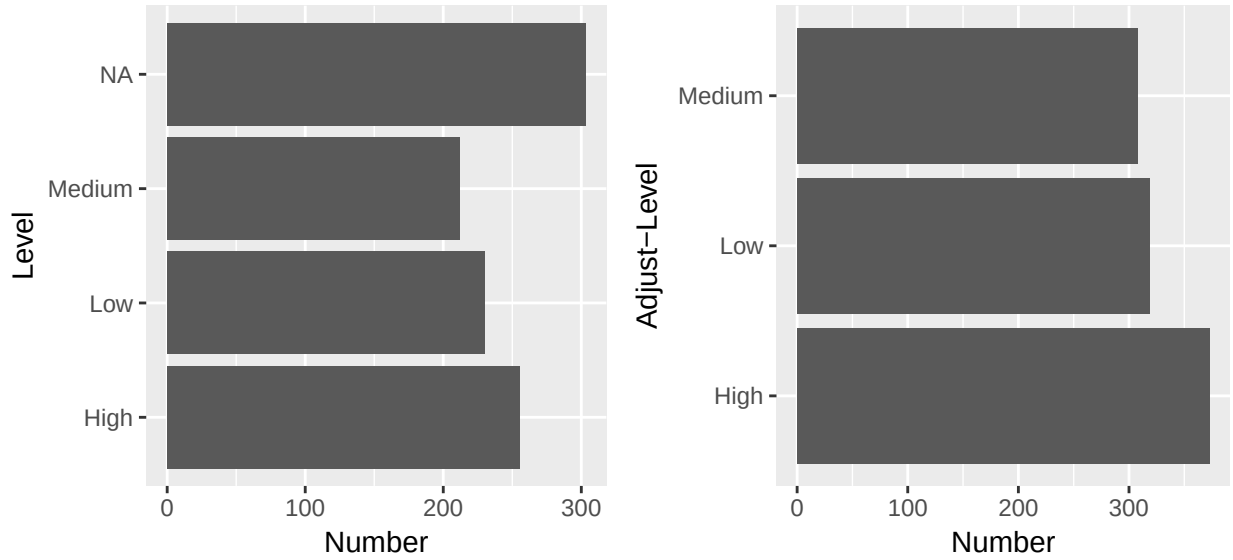
According to the graph, the missing data only appears in History variables and History is a dichotomous variables. It is impossible to use avarage or media to impute these missing value.Thus, I use mice package to impute the missing value

```
newdata <- mice(DMarket, m = 5, method = "pmm", maxit = 100, seed = 1)
```

```
DMarket1 <- complete(newdata)
anyNA(DMarket1)  #check if there is NA in the new dataset.
```

```
## [1] FALSE
```

```
# Make barchart of original History variables and adjusted one.
library(ggplot2)
library(gridExtra)
page1 = ggplot(data = DMarket, aes(x = History)) + geom_bar(stat = "count") +
    scale_x_discrete("Level") + scale_y_continuous("Number") + coord_flip()

page2 = ggplot(data = DMarket1, aes(x = History)) + geom_bar(stat = "count") +
    scale_x_discrete("Adjust-Level") + scale_y_continuous("Number") + coord_flip()

grid.arrange(page1, page2, ncol = 2)
```



By using Mutiple imputation, we can not only impute the missing data, but also the two distribution of History seem alike. That's what we want.

b–generate the summary table

For each numeric variable, list:name, mean, median, 1st quartile, 3rd quartile, standard deviation.

```
# Get the overview of each variables
summary(DMarket1)
```

```
##       Age         Gender    OwnHome       Married     Location
##  Middle:508   Female:506   Own :516   Married:502   Close:710
##  Old   :205   Male  :494   Rent:484   Single :498   Far  :290
##  Young :287
##
##
##
##      Salary          Children       History       Catalogs
```

```
##  Min.   : 10100   Min.   :0.000   High  :373   Min.   : 6.00
##  1st Qu.: 29975   1st Qu.:0.000   Low   :319   1st Qu.: 6.00
##  Median : 53700   Median :1.000   Medium:308   Median :12.00
##  Mean   : 56104   Mean   :0.934                Mean   :14.68
##  3rd Qu.: 77025   3rd Qu.:2.000                3rd Qu.:18.00
##  Max.   :168800   Max.   :3.000                Max.   :24.00
##   AmountSpent
##  Min.   :  38.0
##  1st Qu.: 488.2
##  Median : 962.0
##  Mean   :1216.8
##  3rd Qu.:1688.5
##  Max.   :6217.0
```

```
# Get the name and standard deviation of numeric variables,
sd1 <- sd(DMarket1$Salary)
sd2 <- sd(DMarket1$Children)
sd3 <- sd(DMarket1$Catalogs)
sd4 <- sd(DMarket1$AmountSpent)
print(paste0("Salary Standard Deviation: ", sd1))
```

```
## [1] "Salary Standard Deviation: 30616.31482598"
```

```
print(paste0("Children Standard Deviation: ", sd2))
```

```
## [1] "Children Standard Deviation: 1.05107028725426"
```

```
print(paste0("Catalogs Standard Deviation: ", sd3))
```

```
## [1] "Catalogs Standard Deviation: 6.62289504210498"
```

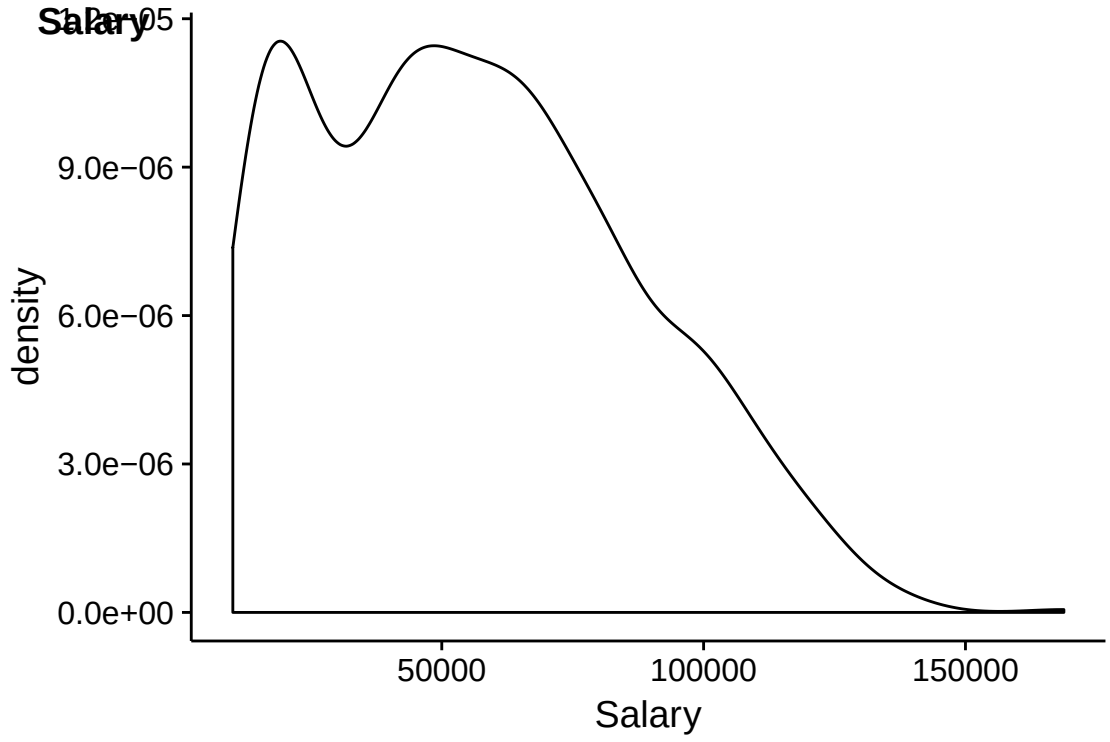
```
print(paste0("AmountSpentStandard Deviation: ", sd4))
```

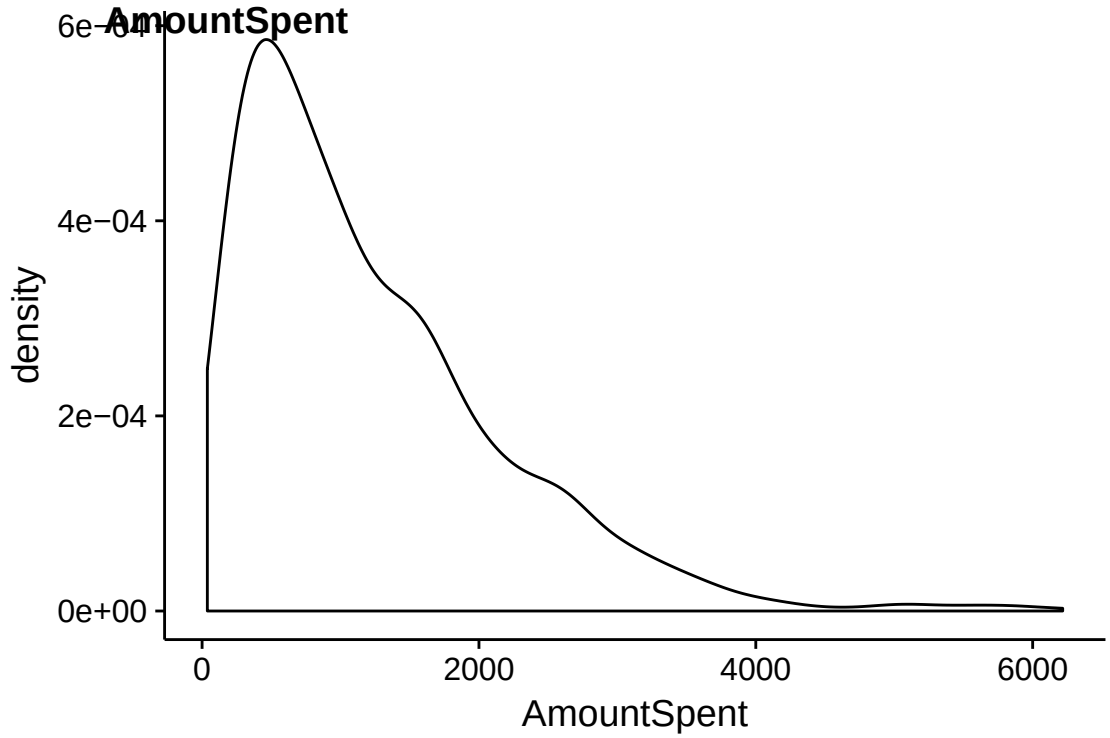
```
## [1] "AmountSpentStandard Deviation: 961.068612523569"
```

c–plot density distribution for Salary and Amount Spent

For numerical variables AmountSpent and Salary, plot the density distribution. Describe whether the variable has a normal distribution or certain type of skew distribution.

```
library(cowplot)
sa_dens <- ggplot(data = DMarket1, aes(x = Salary)) + geom_density()
Amou_dens <- ggplot(data = DMarket1, aes(x = AmountSpent)) + geom_density()

plot_grid(sa_dens, Amou_dens, labels = c("Salary", "AmountSpent"), ncol = 1,
    nrow = 2)
```





From the graphs we can see, neither Salary nor AmountSpent have normal distribution. Both of them are skewed to the right which means they are more likely to see extreme values to the right of the mode.

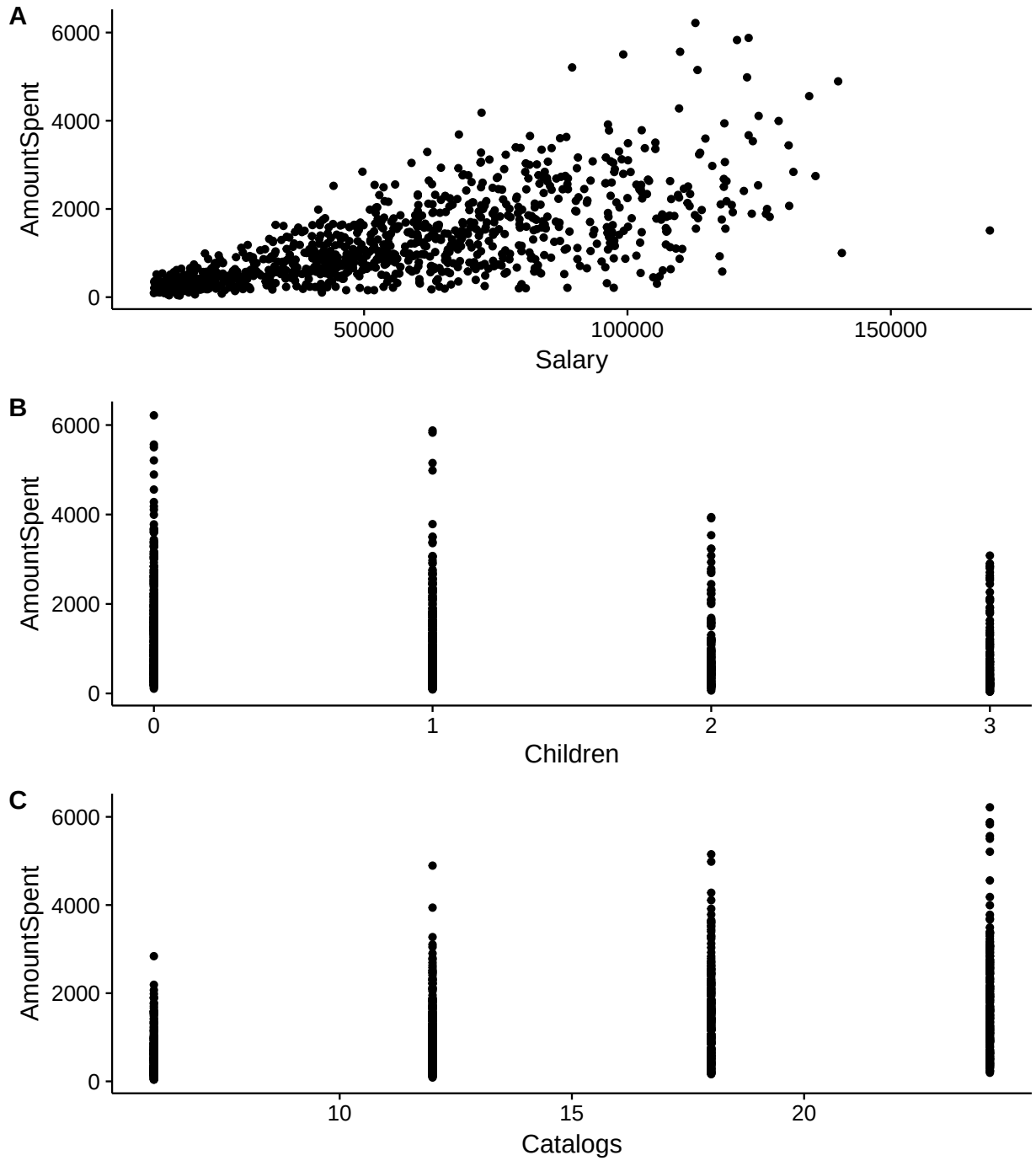
AmountSpent only have one climax and it can be seen as a gamma distribution. While Salary has two climax and it can be seen as a bimodal.

d–relationship between response variable and numeric predictors

For each numerical predictor, describe its relationship with the response variable through correlation and scatterplot.

There are only three numerical predictor in the dataset. They are Salary, Children and Catalogs.

```
cor1 <- cor(DMarket1$Salary, DMarket1$AmountSpent)
sa_scar <- ggplot(data = DMarket1, aes(x = Salary, y = AmountSpent)) + geom_point()

cor2 <- cor(DMarket1$Children, DMarket1$AmountSpent)
chil_scar <- ggplot(data = DMarket1, aes(x = Children, y = AmountSpent)) + geom_point()

cor3 <- cor(DMarket1$Catalogs, DMarket1$AmountSpent)
cat_scar <- ggplot(data = DMarket1, aes(x = Catalogs, y = AmountSpent)) + geom_point()

plot_grid(sa_scar, chil_scar, cat_scar, labels = c("A", "B", "C"), ncol = 1,
    nrow = 3)
```




```
print(paste0("Correlation with Salary and AmountSpent: ", cor1))
```

```
## [1] "Correlation with Salary and AmountSpent: 0.69959570664784"
```

```
print(paste0("Correlation with Children and AmountSpent: ", cor2))
```

```
## [1] "Correlation with Children and AmountSpent: -0.22230816951473"
```

```
print(paste0("Correlation with Catalogs and AmountSpent: ", cor3))
```

```
## [1] "Correlation with Catalogs and AmountSpent: 0.472649894017605"
```

From the scatter plot we can see, only Salary shows some extent of positive relation. The distributions of Children and Catalogs are divided into vertical lines which means they are not strongly related with AmountSpent and their distribution can be seen as four factors.

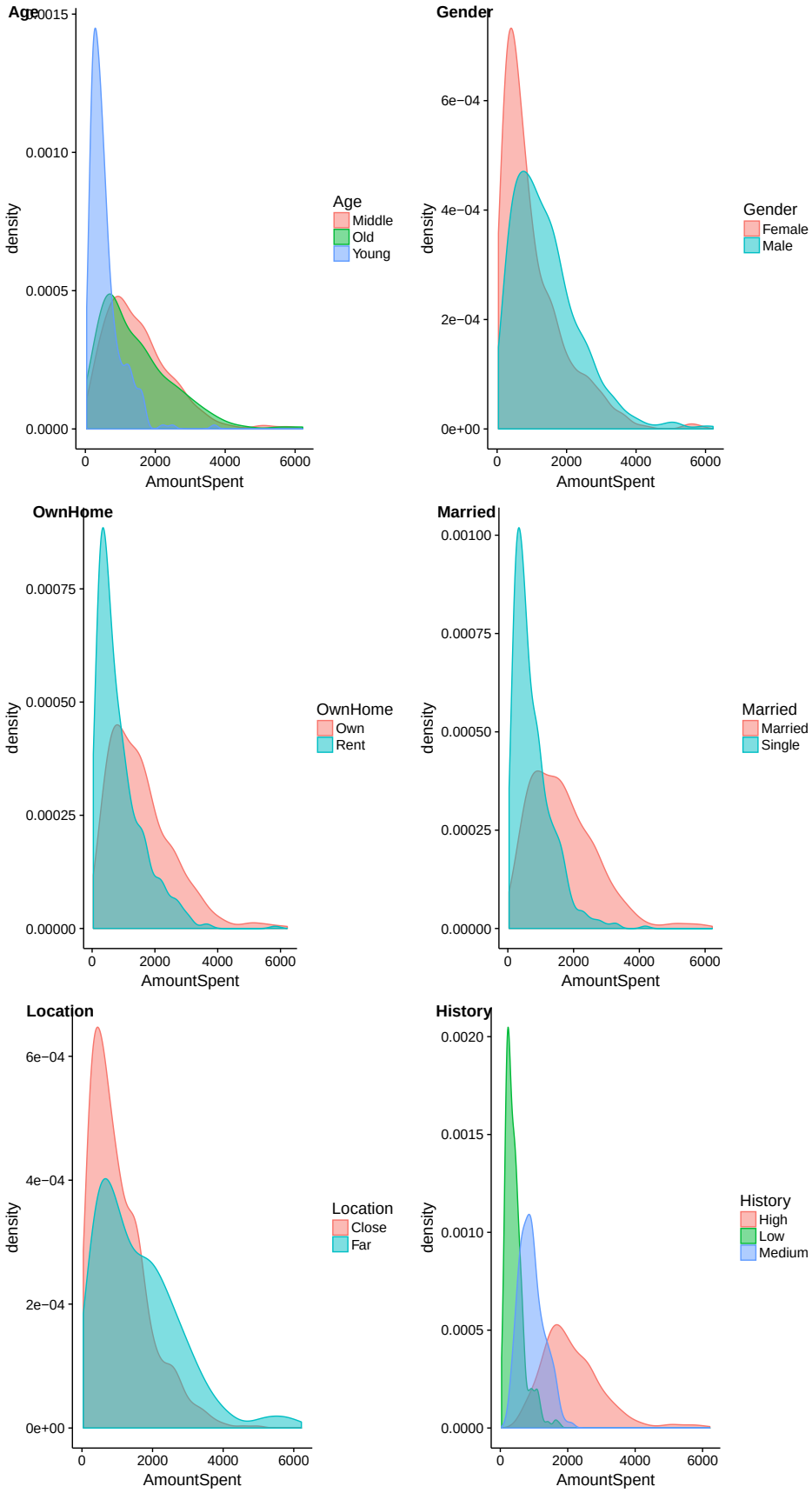
The results of correlation are consistent with the graphs. Correlation coefficient between Salary and AmountSpent is 0.6995957, which means they are strongly related. However, the correlation coefficient between Children and AmountSpent are -0.2223082, which means they are not strongly related and the correlation coefficient between Catalogs and AmountSpent is 0.4726499 which means weak relation with each other.

In addition, the coefficient for Salary and AmountSpent is 0.6995957, suggesting that an increase of 1 percent in Salary is associated with a 0.6995957 percent increase in the AmountSpent. The coefficient for Children and AmountSpent is -0.2223082, suggesting that an increase of 1 percent in Children is associated with a -0.2223082 percent increase in the AmountSpent. The coefficient for Catalogs and AmountSpent is 0.4726499, suggesting that an increase of 1 percent in Catalogs is associated with a 0.4726499 percent increase in the AmountSpent.

e–Density plot for categorical predictors

For each categorical predictor, generate conditional density plot of response variables

```
# Age and AmountSpent
age_dens <- ggplot(DMarket1, aes(x = AmountSpent, fill = Age, colour = Age)) +
    geom_density(alpha = 0.5)

# Gender and AmountSpent
gender_dens <- ggplot(DMarket1, aes(x = AmountSpent, fill = Gender, colour = Gender)) +
    geom_density(alpha = 0.5)

# OwnHome and AmountSpent
oh_dens <- ggplot(DMarket1, aes(x = AmountSpent, fill = OwnHome, colour = OwnHome)) +
    geom_density(alpha = 0.5)

# Married and AmountSpent
marr_dens <- ggplot(DMarket1, aes(x = AmountSpent, fill = Married, colour = Married)) +
    geom_density(alpha = 0.5)

# Location and AmountSpent
loca_dens <- ggplot(DMarket1, aes(x = AmountSpent, fill = Location, colour = Location)) +
    geom_density(alpha = 0.5)

# History and AmountSpent
histo_dens <- ggplot(DMarket1, aes(x = AmountSpent, fill = History, colour = History)) +
    geom_density(alpha = 0.5)

plot_grid(age_dens, gender_dens, oh_dens, marr_dens, loca_dens, histo_dens,
    labels = c("Age", "Gender", "OwnHome", "Married", "Location", "History"),
    ncol = 2, nrow = 3)
```


Age
density
AmountSpent
Age
Middle
Old
Young
Gender
Female
Male
OwnHome
Own
Rent
Married
Single
Location
Close
Far
History
High
Low
Medium

By using different color and line shapes to differentiate categories in each variables, we can see the density plots of six categorical variables as above.

f–Compare and describe the significance of categories.

For each categorical predictor, compare and describe whether the categories have significantly different means

From the six graphs we can see, History has the most significantly different means since the density plot shows very different distributions in three categories. And Gender and Location have someway important different means, because their densiy plots show different lines with different categories. But they have some parts that are overlap together. In addtion, Age, OwnHome and Married show very low difference in their categories, since the different parts of categories are overlapped.

3–Apply regression analysis

a–use all predictors in standard linear regression

Use all predictors in a standard linear regression model to predictthe response variable. Report the model performance using R2, adjusted R2 and RMSE. Interpret the regression result.

```
Market = as.data.frame(DMarket1)
head(Market)
```

```
##      Age Gender OwnHome Married Location Salary Children History Catalogs
## 1    Old Female     Own  Single      Far  47500        0    High        6
## 2 Middle   Male    Rent  Single    Close  63600        0    High        6
## 3  Young Female    Rent  Single    Close  13500        0     Low       18
## 4 Middle   Male     Own Married    Close  85600        1    High       18
## 5 Middle Female     Own  Single    Close  68400        0    High       12
## 6  Young   Male     Own Married    Close  30400        0     Low        6
##   AmountSpent
## 1         755
## 2        1318
## 3         296
## 4        2436
## 5        1304
## 6         495
```

```
library(car)
fit = lm(AmountSpent ~ Age + Gender + OwnHome + Married + Location + Salary +
    Children + History + Catalogs, data = Market)
summary(fit)
```

```
##
## Call:
## lm(formula = AmountSpent ~ Age + Gender + OwnHome + Married +
##     Location + Salary + Children + History + Catalogs, data = Market)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1522.46  -289.08   -26.04   227.89  3034.17
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.133e+02  1.213e+02   0.935    0.350
```

```
## Age2           2.180e+01  4.774e+01   0.457    0.648
## Age3           1.227e+01  4.948e+01   0.248    0.804
## Gender2       -5.005e+01  3.299e+01  -1.517    0.130
## OwnHome2      -1.933e+01  3.672e+01  -0.526    0.599
## Married2       5.452e+01  4.451e+01   1.225    0.221
## Location2      3.677e+02  3.709e+01   9.915  < 2e-16 ***
## Salary         1.613e-02  1.109e-03  14.547  < 2e-16 ***
## Children      -1.389e+02  1.793e+01  -7.748 2.31e-14 ***
## HistoryLow    -5.593e+02  6.292e+01  -8.890  < 2e-16 ***
## HistoryMedium -5.102e+02  4.875e+01 -10.467  < 2e-16 ***
## Catalogs       3.789e+01  2.485e+00  15.250  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 487 on 988 degrees of freedom
## Multiple R-squared:  0.7461, Adjusted R-squared:  0.7432
## F-statistic: 263.9 on 11 and 988 DF,  p-value: < 2.2e-16
```

```
mean.mse = mean((rep(mean(Market$AmountSpent), length(Market$AmountSpent)) -
    Market$AmountSpent)^2)
model.mse = mean(residuals(fit)^2)
rmse = sqrt(model.mse)
rmse
```

```
## [1] 484.059
```

Interprete the regression result

The coefficient in in multiple regression shows increase in the dependent variable for a unit change in a predictor varianble. The regression coefficient for Salary is 1.613, suggesting that an increase of 1 percent in Salary is associated with a 1.613 percent increase in the AmountSpent, controlling for the other predictors. The regression coefficient for Children is -1.389, suggesting that an increase of 1 percent in Children is associated with a -1.389 percent increase in the AmountSpent, controlling for the other predictors.This is for numerical variables. For categorical variables, linear regression calculate the coeffiecient by counting the number of different levels in categorical variables. For example, the regression coefficient for Age2 is 2.180, suggesting that an increase of 1 percent in Age of level 2 is associated with a 2.180 percent increase in the AmountSpent, controlling for the other predictors. While the regression coefficient for Age3 is 1.227, suggesting that an increase of 1 percent in Age of level 3 is associated with a 1.227 percent increase in the AmountSpent, controlling for the other predictors.

The coefficient is significantly different from zero at the $p < .0001$ level. The coefficients for most of predictors here are significantly different from zero ($p < 0.648$) suggesting that most predictors and AmountSpent are linearly related when controlling for the other predictor variables.

Taken together, the predictor variables account for 75 percent of the variance in AmountSpent.

b–different combinations

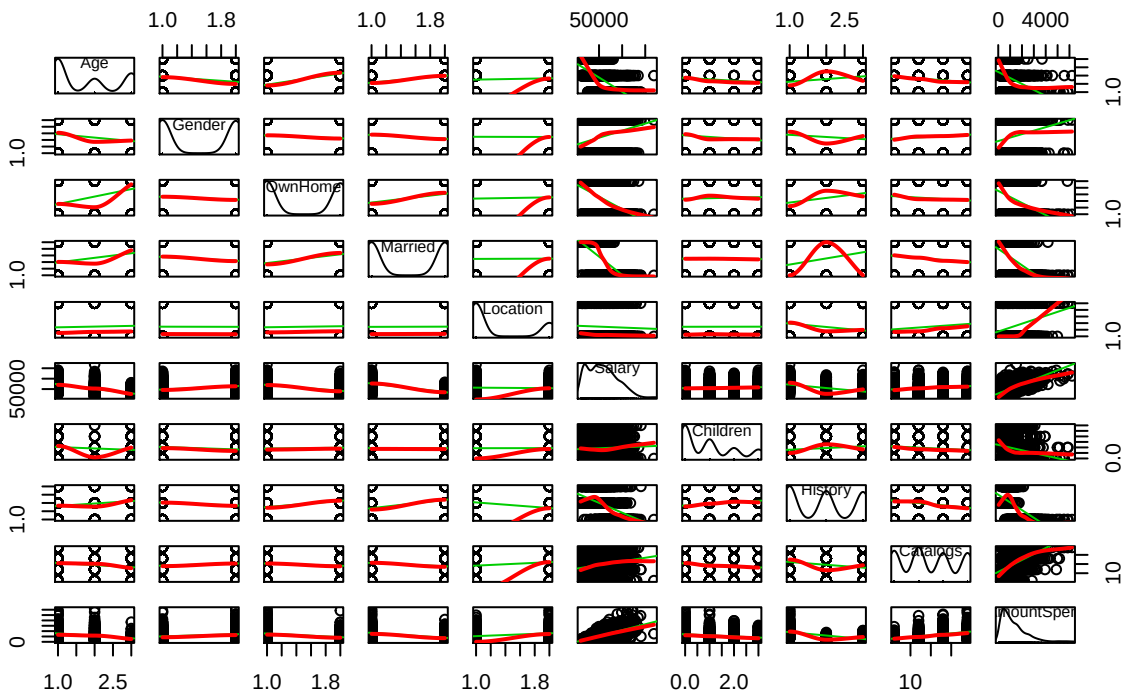
use diffenrent combinations in standard linear and non-linear regression

Simple Regression by deleting single variables.

```
#standard linear regression
fit1 = lm(AmountSpent ~ Gender+OwnHome+Married+Location+Salary+Children+History+Catalogs,
          data=Market)
summary(fit1)
```

```
##
## Call:
## lm(formula = AmountSpent ~ Gender + OwnHome + Married + Location +
##     Salary + Children + History + Catalogs, data = Market)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1495.08  -284.79   -25.94   226.98  3052.51
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.406e+02  1.005e+02   1.399    0.162
## Gender2       -5.215e+01  3.247e+01  -1.606    0.109
## OwnHome2      -2.111e+01  3.496e+01  -0.604    0.546
## Married2       4.858e+01  4.212e+01   1.153    0.249
## Location2      3.679e+02  3.703e+01   9.933   <2e-16 ***
## Salary         1.595e-02  9.473e-04  16.832   <2e-16 ***
## Children      -1.416e+02  1.661e+01  -8.523   <2e-16 ***
## HistoryLow    -5.602e+02  6.262e+01  -8.946   <2e-16 ***
## HistoryMedium -5.124e+02  4.803e+01 -10.667   <2e-16 ***
## Catalogs       3.783e+01  2.473e+00  15.295   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 486.6 on 990 degrees of freedom
## Multiple R-squared:  0.746,  Adjusted R-squared:  0.7437
## F-statistic: 323.1 on 9 and 990 DF,  p-value: < 2.2e-16
```

```
## examining bivariate relationships using 'scatterplotMatrix' in the 'car' package
suppressWarnings(
  scatterplotMatrix(Market, spread=FALSE, lty.smooth=2,
                    main="Scatter Plot Matrix")
)
```




```
#standard linear regression
fit2 = lm(AmountSpent ~ Age+OwnHome+Married+Location+Salary+Children+History+Catalogs,
          data=Market)
summary(fit2)
```

```
##
## Call:
## lm(formula = AmountSpent ~ Age + OwnHome + Married + Location +
##     Salary + Children + History + Catalogs, data = Market)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1528.09  -294.72   -31.94   221.91  3015.57
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.654e+01  1.208e+02   0.799    0.425
## Age2           3.284e+01  4.721e+01   0.696    0.487
## Age3           1.127e+01  4.950e+01   0.228    0.820
## OwnHome2      -2.003e+01  3.674e+01  -0.545    0.586
## Married2       5.005e+01  4.444e+01   1.126    0.260
## Location2      3.686e+02  3.711e+01   9.932  < 2e-16 ***
## Salary         1.590e-02  1.099e-03  14.465  < 2e-16 ***
## Children      -1.352e+02  1.777e+01  -7.607 6.51e-14 ***
## HistoryLow    -5.529e+02  6.281e+01  -8.802  < 2e-16 ***
## HistoryMedium -5.097e+02  4.878e+01 -10.450  < 2e-16 ***
```

```
## Catalogs       3.786e+01  2.486e+00  15.227  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 487.3 on 989 degrees of freedom
## Multiple R-squared:  0.7455, Adjusted R-squared:  0.7429
## F-statistic: 289.7 on 10 and 989 DF,  p-value: < 2.2e-16
```

```
#standard linear regression
fit3 = lm(AmountSpent ~ Age+Gender+Married+Location+Salary+Children+History+Catalogs,
          data=Market)
summary(fit3)
```

```
##
## Call:
## lm(formula = AmountSpent ~ Age + Gender + Married + Location +
##     Salary + Children + History + Catalogs, data = Market)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1529.02  -289.59   -28.06   226.39  3034.39
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9.976e+01  1.184e+02   0.842    0.400
## Age2            2.559e+01  4.717e+01   0.542    0.588
## Age3            7.472e+00  4.861e+01   0.154    0.878
## Gender2        -5.027e+01  3.298e+01  -1.524    0.128
## Married2        5.600e+01  4.440e+01   1.261    0.208
## Location2       3.669e+02  3.704e+01   9.905  < 2e-16 ***
## Salary          1.622e-02  1.092e-03  14.857  < 2e-16 ***
## Children       -1.384e+02  1.790e+01  -7.734 2.56e-14 ***
## HistoryLow     -5.615e+02  6.276e+01  -8.948  < 2e-16 ***
## HistoryMedium -5.120e+02  4.861e+01 -10.532  < 2e-16 ***
## Catalogs        3.787e+01  2.484e+00  15.249  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 486.8 on 989 degrees of freedom
## Multiple R-squared:  0.746,  Adjusted R-squared:  0.7434
## F-statistic: 290.5 on 10 and 989 DF,  p-value: < 2.2e-16
```

I tried to drop single variables each time, but it turns out, using all the variables shows highest r square value in standard linear regression. I think for standard linear regression, using all variables can give us the best r-square.

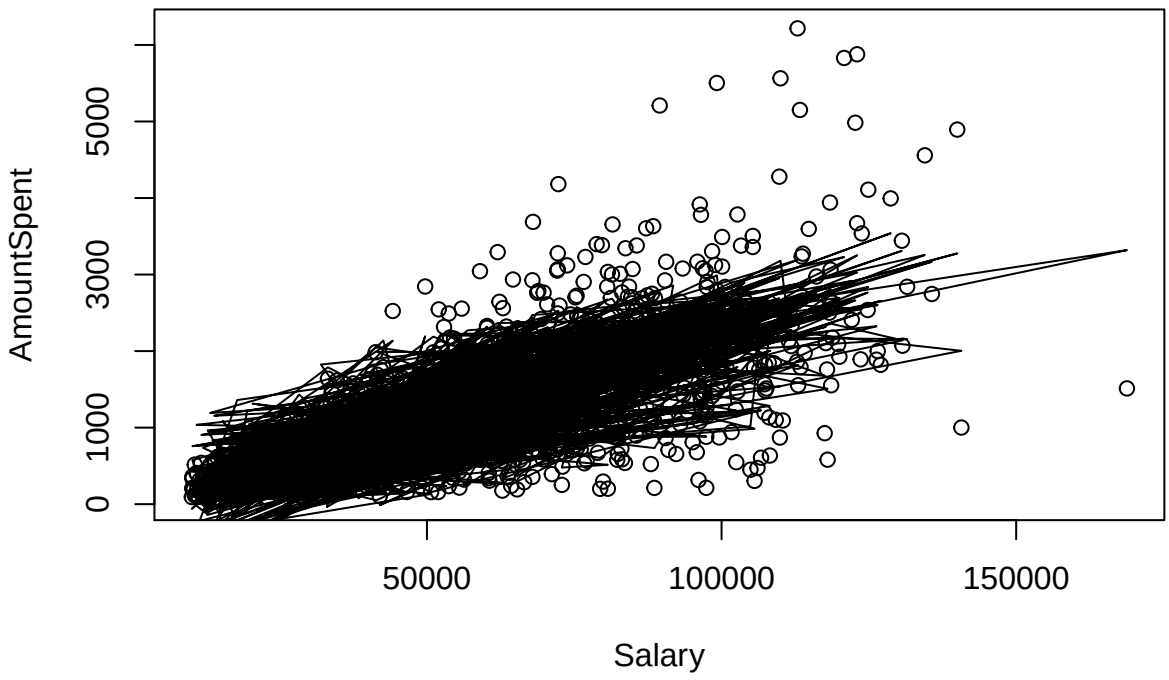
Pplynomial Regression

```
fit4 = lm(AmountSpent ~ Age+Gender+OwnHome+Married+Location
          +Salary+Children+History+Catalogs
          + I(Salary^2), data=Market)
summary(fit4)
```

```
##
```

```
## Call:
## lm(formula = AmountSpent ~ Age + Gender + OwnHome + Married +
##     Location + Salary + Children + History + Catalogs + I(Salary^2),
##     data = Market)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1807.21  -284.73   -19.82   221.07  3025.81
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.544e+02  1.679e+02   2.111  0.03504 *
## Age2          -3.077e+00  4.915e+01  -0.063  0.95008
## Age3          -4.408e+01  5.639e+01  -0.782  0.43456
## Gender2       -4.392e+01  3.307e+01  -1.328  0.18447
## OwnHome2      -2.357e+01  3.671e+01  -0.642  0.52098
## Married2       2.695e+01  4.638e+01   0.581  0.56131
## Location2      3.563e+02  3.743e+01   9.519  < 2e-16 ***
## Salary         1.008e-02  3.122e-03   3.227  0.00129 **
## Children      -1.357e+02  1.797e+01  -7.550 9.88e-14 ***
## HistoryLow    -6.089e+02  6.722e+01  -9.059  < 2e-16 ***
## HistoryMedium -5.235e+02  4.908e+01 -10.665  < 2e-16 ***
## Catalogs       3.769e+01  2.483e+00  15.182  < 2e-16 ***
## I(Salary^2)    3.807e-08  1.837e-08   2.072  0.03852 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 486.2 on 987 degrees of freedom
## Multiple R-squared:  0.7472, Adjusted R-squared:  0.7441
## F-statistic: 243.1 on 12 and 987 DF,  p-value: < 2.2e-16
```

```
plot(Market$Salary,Market$AmountSpent,
     xlab="Salary",
     ylab="AmountSpent")
lines(Market$Salary,fitted(fit4))
```



The r square turned to 0.7472, which is higher than standard linear regression.

```
fit5 = lm(AmountSpent ~ Age+Gender+OwnHome+Married+Location+Salary
          +Children+History+Catalogs+ I(Salary^2)
          + I(Children^2) +I(Catalogs^2), data=Market)
summary(fit5)
```

```
##
## Call:
## lm(formula = AmountSpent ~ Age + Gender + OwnHome + Married +
##     Location + Salary + Children + History + Catalogs + I(Salary^2) +
##     I(Children^2) + I(Catalogs^2), data = Market)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1825.72  -276.73   -23.06   209.51  3001.66
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.282e+02  1.854e+02   2.309 0.021138 *
## Age2           5.197e+00  5.111e+01   0.102 0.919031
## Age3          -3.875e+01  5.666e+01  -0.684 0.494175
## Gender2       -4.157e+01  3.328e+01  -1.249 0.211881
## OwnHome2      -2.550e+01  3.674e+01  -0.694 0.487811
## Married2       3.121e+01  4.655e+01   0.670 0.502711
## Location2      3.565e+02  3.744e+01   9.523  < 2e-16 ***
## Salary         1.041e-02  3.145e-03   3.308 0.000972 ***
```

```
## Children      -1.067e+02  5.193e+01  -2.055 0.040144 *
## HistoryLow    -6.052e+02  6.738e+01  -8.981  < 2e-16 ***
## HistoryMedium -5.216e+02  4.924e+01 -10.594  < 2e-16 ***
## Catalogs       2.120e+01  1.317e+01   1.610 0.107647
## I(Salary^2)    3.658e-08  1.850e-08   1.977 0.048299 *
## I(Children^2) -1.080e+01  1.756e+01  -0.615 0.538729
## I(Catalogs^2)  5.504e-01  4.306e-01   1.278 0.201507
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 486.2 on 985 degrees of freedom
## Multiple R-squared:  0.7477, Adjusted R-squared:  0.7441
## F-statistic: 208.5 on 14 and 985 DF,  p-value: < 2.2e-16
```

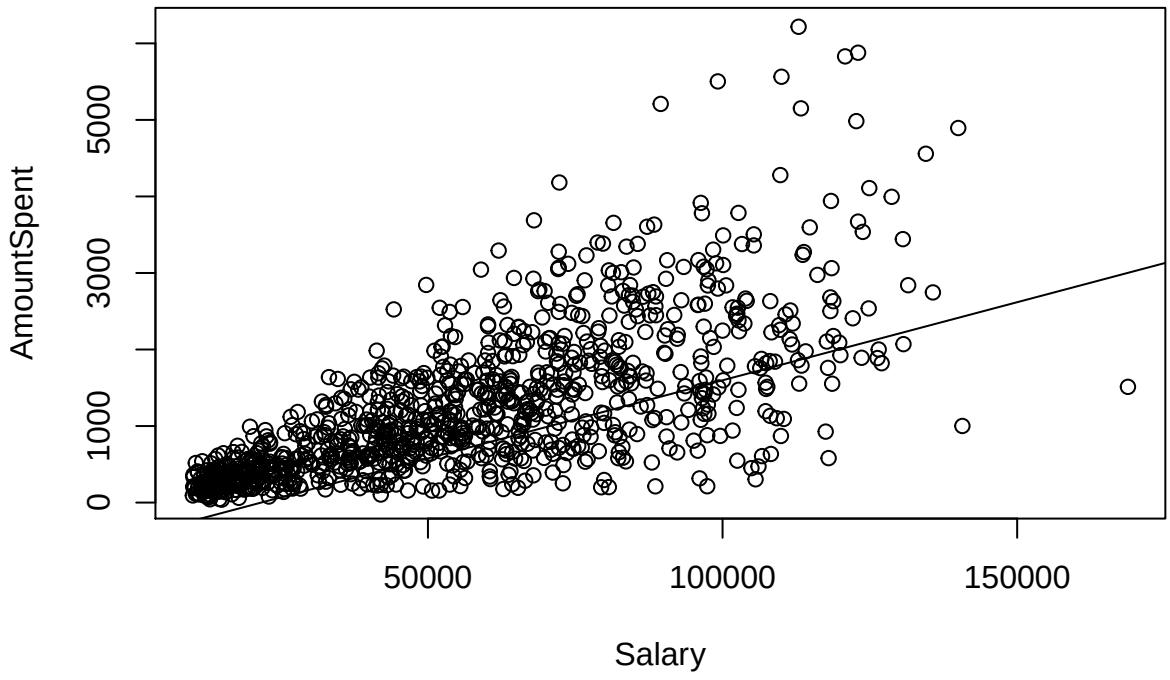
The r square turned to 0.7477, which is higher than former one and this is best r square I get till now.

Local Polynomial Regression

```
library(locfit)
```

```
## locfit 1.5-9.1    2013-03-22
```

```
#fit with a 50% nearest neighbor bandwidth.
fitreg=lm(AmountSpent~Salary+Children+Catalogs,data=Market)
plot(AmountSpent~Salary,data=Market)
abline(fitreg)
```



```r
#The linear regression model didn't fit in the plot.

fit7 <- locfit(AmountSpent~lp(Salary+Children,nn=0.5),data=Market)
fit7
```

```
## Call:
## locfit(formula = AmountSpent ~ lp(Salary + Children, nn = 0.5),
##     data = Market)
##
## Number of observations:          1000
## Family:  Gaussian
## Fitted Degrees of freedom:       6.946
## Residual scale:                  687
```

```r
summary(fit7)
```

```
## Estimation type: Local Regression
##
## Call:
## locfit(formula = AmountSpent ~ lp(Salary + Children, nn = 0.5),
##     data = Market)
##
## Number of data points:  1000
## Independent variables:  Salary + Children
## Evaluation structure: Rectangular Tree
## Number of evaluation points:  11
## Degree of fit:  2
## Fitted Degrees of Freedom:  6.946
```

```r
plot(fit7)
```

The local polynomial didn't show a good r square. This is because in local polynomial model, we can only use numeric variables, that means we ignore all the categorical variables. However, some categorical variables are important to the response variables and in this dataset, most variables are categorical. Therefore, this model shows a bad performance.

Lasso

I only use lasso to test the r square. Since Lasso's predition is matrix, it's hard to use leave one out for it. Thus, I can only explain the result by r square and graphs.

```
x <- model.matrix(AmountSpent~ Age+Gender+OwnHome+Married+Location+Salary+
                  Children+History+Catalogs,data=Market)
x=x[,-1]
library(lars)
```

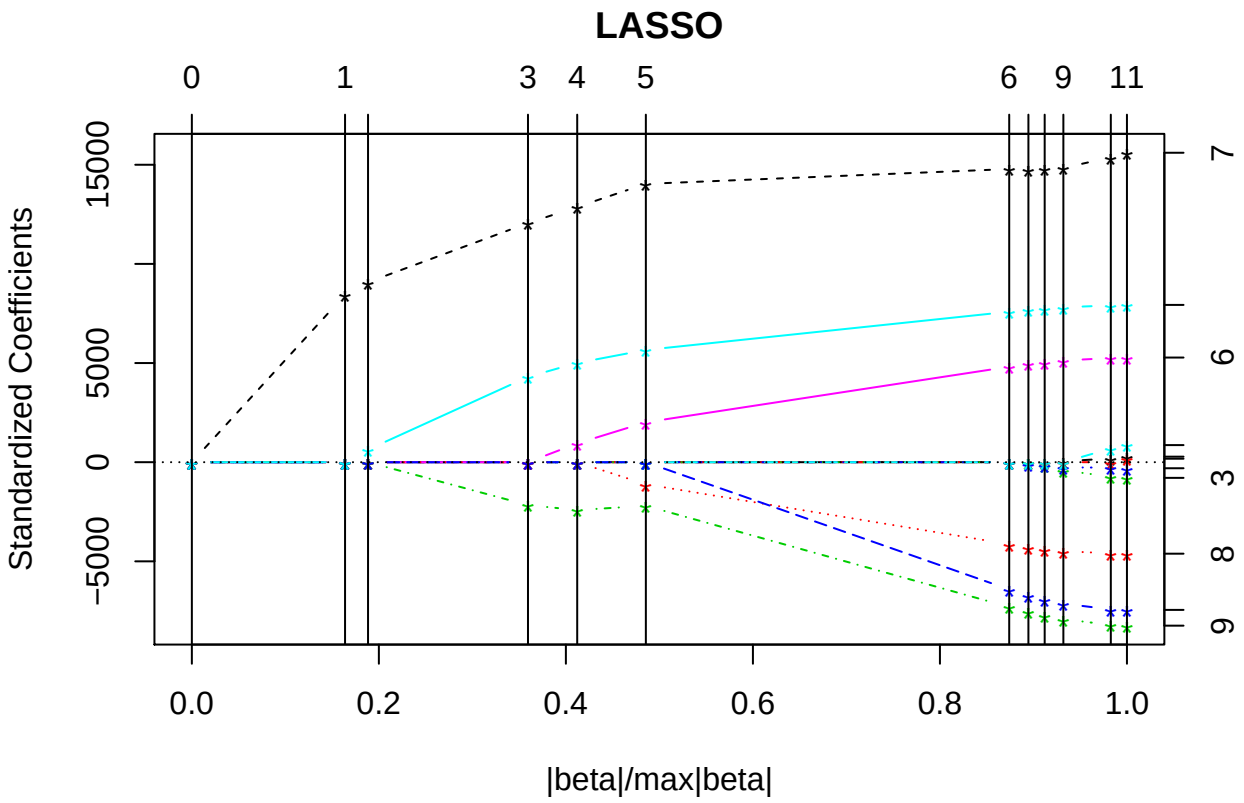
```
## Loaded lars 1.2
```

```
## lasso on all data
lasso <- lars(x=x,y=Market$AmountSpent ,trace=TRUE)
```

```
## LASSO sequence
## Computing X'X .....
## LARS Step 1 :     Variable 7     added
## LARS Step 2 :     Variable 11    added
## LARS Step 3 :     Variable 9     added
## LARS Step 4 :     Variable 6     added
## LARS Step 5 :     Variable 8     added
```

```
## LARS Step 6 :     Variable 10    added
## LARS Step 7 :     Variable 4     added
## LARS Step 8 :     Variable 3     added
## LARS Step 9 :     Variable 1     added
## LARS Step 10 :    Variable 5     added
## LARS Step 11 :    Variable 2     added
## Computing residuals, RSS etc .....
```

```
## trace of lasso (standardized) coefficients for varying penalty
plot(lasso)
```




```
lasso
```

```
##
## Call:
## lars(x = x, y = Market$AmountSpent, trace = TRUE)
## R-squared: 0.746
## Sequence of LASSO moves:
##      Salary Catalogs HistoryLow Location2 Children HistoryMedium OwnHome2
## Var       7       11          9         6        8             10        4
## Step      1        2          3         4        5              6        7
##      Gender2 Age2 Married2 Age3
## Var        3    1        5    2
## Step       8    9       10   11
```

The r-square for Lasso is 0.748 which is the highest score in all models.

Evaluate model performs using out-of-sample RMSE.

I use leave-one-out cross valiadation to check the out-of-sample RMSE for each combination and different model. In the last, I use a line plot to show the result of all the model.

First one is Standard linear regression model with all response variables.

```
n = length(Market$AmountSpent)
error = dim(n)
for (k in 1:n) {
  train1 = c(1:n)
  train = train1[train1!=k]
  m1 = lm(AmountSpent ~ Age+Gender+OwnHome+Married+Location
          +Salary+Children+History+Catalogs, data=Market[train ,])
  pred = predict(m1, newdat=Market[-train ,])
  obs = Market$AmountSpent[-train]
  error[k] = obs-pred
}
rmse=sqrt(mean(error^2))
rmse
```

```
## [1] 490.5036
```

First model is the most basic model. I use all the predictors in standard linear regression model and it shows the RMSE as 490.5039 which is a little bit higher. I tried to improve model by droping one variable at each time and check if RMSE get higher. It turns out that when dropping Age, Gender, OwnHome and Married, I got the best result in the standard linear model.The process is as following.

Drop Age in response variables.

```
for (k in 1:n) {
  train1 = c(1:n)
  train = train1[train1!=k]
  m2 = lm(AmountSpent ~ Gender+OwnHome+Married+Location
          +Salary+Children+History+Catalogs, data=Market[train ,])
  pred = predict(m2, newdat=Market[-train ,])
  obs = Market$AmountSpent[-train]
  error[k] = obs-pred
}
rmse=sqrt(mean(error^2))
rmse
```

```
## [1] 489.3994
```

Drop Age and OwnHome in response variables.

```
for (k in 1:n) {
  train1 = c(1:n)
  train = train1[train1!=k]
  m2 = lm(AmountSpent ~ Gender+Married+Location
          +Salary+Children+History+Catalogs, data=Market[train ,])
  pred = predict(m2, newdat=Market[-train ,])
  obs = Market$AmountSpent[-train]
  error[k] = obs-pred
}
rmse=sqrt(mean(error^2))
rmse
```

```
## [1] 489.0839
```

Drop Age, OwnHomem, Married and Gender in response variables.

```
for (k in 1:n) {
  train1 = c(1:n)
  train = train1[train1!=k]
  m2 = lm(AmountSpent ~ Location+Salary+Children
          +History+Catalogs, data=Market[train ,])
  pred = predict(m2, newdat=Market[-train ,])
  obs = Market$AmountSpent[-train]
  error[k] = obs-pred
}
rmse=sqrt(mean(error^2))
rmse
```

```
## [1] 489.1668
```

This is the smallest RMSE I can get from standard linear regression after trying many combinations. Next I tried some of polynomial models with different combinations. First model is polynomial regression with all response variables and square of salary. This is very basic one, since the scatter plot shows Salary and AmountSpent are strongly related. I tried the square of Salary first.

```
for (k in 1:n) {
  train1 = c(1:n)
  train = train1[train1!=k]
  m2 = lm(AmountSpent ~ Age+Gender+OwnHome+Married
          +Location+Salary+Children+History+Catalogs
          + I(Salary^2), data=Market[train ,])
  pred = predict(m2, newdat=Market[-train ,])
  obs = Market$AmountSpent[-train]
  error[k] = obs-pred
}
rmse=sqrt(mean(error^2))
rmse
```

```
## [1] 490.9165
```

The final RMSE is 490.9165 which is higher than using standard linear regression model with all response variable. Therefore, using all response variables may not be a good idea. I tried to use drop the Age, OwnHome and Married in the variables and it gets a better result.

```
for (k in 1:n) {
  train1 = c(1:n)
  train = train1[train1!=k]
  m2 = lm(AmountSpent ~ Gender+Location+Salary
          +Children+History+Catalogs
          + I(Salary^2), data=Market[train ,])
  pred = predict(m2, newdat=Market[-train ,])
  obs = Market$AmountSpent[-train]
  error[k] = obs-pred
}
rmse=sqrt(mean(error^2))
rmse
```

```
## [1] 489.755
```

After trying the sqaure of Salary, I tried to add square of Children and square of Catalogs. but adding them didn't improve the RSME. Instead, they increased the RMSE in some way. Thus, I tried to use Salary^3 but the RMSE is still higher than former model.

```
for (k in 1:n) {
  train1 = c(1:n)
  train = train1[train1!=k]
  m2 = lm(AmountSpent ~ Gender+Location+Salary
          +Children+History+Catalogs
          + I(Salary^2)+I(Catalogs^2), data=Market[train ,])
  pred = predict(m2, newdat=Market[-train ,])
  obs = Market$AmountSpent[-train]
  error[k] = obs-pred
}
rmse=sqrt(mean(error^2))
rmse
```

```
## [1] 489.9285
```

All in all, the best result for polynomial regression is 489.755 by dropping three variables and use the square of Salary.

The last model is local polynomial regression. Though local polynomial regression model didn't show good performance in the r-square test, cosidering there are some difference betweent r-square and RSME, I tried the local polynomial in leave-one-out evaluation.

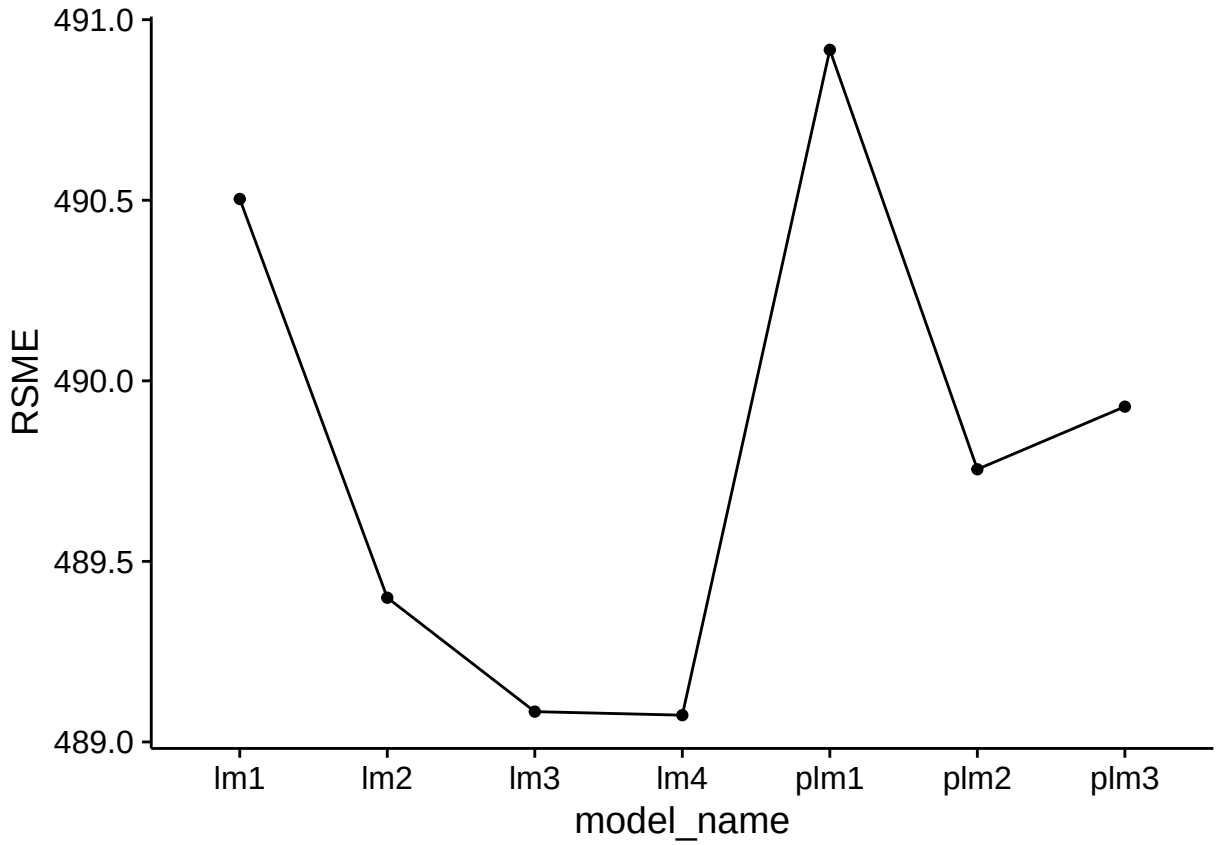
```
for (k in 1:n) {
  train1 = c(1:n)
  train = train1[train1!=k]
  m2 = locfit(AmountSpent ~ lp(Salary+Children+Catalogs,nn=0.5),
              data=Market[train ,])
  pred = predict(m2, newdat=Market[-train ,])
  obs = Market$AmountSpent[-train]
  error[k] = obs-pred
}
rmse=sqrt(mean(error^2))
rmse
```

```
## [1] 695.5365
```

The RMSE is 695.5365 which is highest in the all models. This means local polynomial model is really not suitable for this model.

The comparison of these models are shown in line graphs. Since RMSE in local polynomial are too large, I didn't add it into the line graph. From the graph we can see, most models give the RSME between 491 and 489. The fourth model in linear regression gives the lowest RMSE while the first model in polynomial regression gives the highest RMSE.

```
df <- data.frame(model_name=c("lm1", "lm2", "lm3",
                              "lm4","plm1","plm2","plm3"),
                 RSME=c(490.5036, 489.3994,
                        489.0839,489.0741,
                        490.9165,489.755,489.9285))
ggplot(data=df, aes(x=model_name, y=RSME, group=1)) +
  geom_line(color="black")+
  geom_point()
```



c–Identify the most important predictor

Task:From the best model, identify the most important predictor. How to determine the importance of predictor. Consider variable selection in out-of-one sample evaluation setting.

The best model I got is from standard linear regression with Gender, Location, Salary, Children, History and Catalogs variables. I used backward stepwise selection to identify the importance of these variables.

```
## Start:  AIC=12382.47
## AmountSpent ~ Gender + Location + Salary + Children + History +
##     Catalogs
##
##             Df Sum of Sq       RSS   AIC
## <none>                   234794555 12382
## - Gender    1    560625 235355180 12383
## - Children  1  17232270 252026825 12451
## - Location  1  23231333 258025887 12475
## - History   2  27921994 262716549 12491
## - Catalogs  1  55327836 290122390 12592
## - Salary    1  94070559 328865113 12717
```

```
##
## Call:
## lm(formula = AmountSpent ~ Gender + Location + Salary + Children +
##     History + Catalogs, data = Market)
##
```

```
## Coefficients:
##   (Intercept)        Gender2      Location2         Salary       Children
##     175.23187      -49.66481      366.63484        0.01556     -141.55956
##    HistoryLow  HistoryMedium       Catalogs
##    -559.36312     -512.82288       37.81034
```

The importance of these variables can be seen from AIC values. Since this is the best combinations for linear regression model. There is only one step for AIC. The response variables are listed in the ascending order of AIC and it shows the importance of each variables also. The variables with higher AIC is more important to the response variables since the the less AIC is, the better model performance will show. In conclusion, the most imporatant predictor is Salary, and following are Catalogs, History, Location, Children and Gender is the least important presictor.

Things to consider

This dataset maily consists of the caterorical variables while our predictor is numeric. It is hard to use either logistic regression or simple linear regression, since both of these two regression model are not perfectly fit the distribution of the predictors. I am considering whether we use different models for different variables will give us a better prediction. For example, we can use logistic regression model for categorical variables and linear regression model for numerical variables and we take the prediction from the average of these two predictions. I also noticed that if I take NA value in as a new level in History variables. The RMSE will be higher than imputation the missing value. If I exclude all the missing data, the RMSE will become higher too. But I m not sure if it is appropriate to just delete the missing data.