MKT 382 Marketing Analytics II
Assignment 1
Due: February 12th, 11:59pm

Linear Regression Analysis

We will use a simple dataset to evaluate the impact of the opening of a new Walmart on the sales a local grocery store. Suppose that you have been hired as a consultant for the local grocery store. Store management is worried since Wal-Mart has entered the market by opening a "Wal-Mart Super-center" only 3 miles away. The management is interested in analyzing the impact on store sales after Wal-Mart's entry.

For the analysis, management has given you access to 50 weeks of sales data before the entry of Walmart and 50 weeks after. Please download and look at the data in "Walmart_Data.csv" from Canvas.

The dataset has the following variables:

WEEK	Week number
Sales	weekly sales
Promotion	Index of weekly promotion activity –higher promotion index indicates more products on promotion in the store
Feature	Index of feature advertising activity – higher feature advertising index indicates more feature advertising
Walmart	A categorical variable = "No" in the weeks before the Walmart opens, and "Present" in the weeks before the Walmart opens
Holiday	Holiday = "Yes" during major holiday weeks, and "No" for non-holiday weeks

(1). Pleas read the data into R. Use the function str( ) to find the structure of the data fame and the summary( ) to summarize the data. Please post the results here. Create a new variable called "logSales", which is the logarithm of the variable "Sales" in the data frame "walmart".

```
```{r}
dat=read.csv('Walmart_Data.csv')
str(dat)
```
```

```
'data.frame':   100 obs. of  6 variables:
 $ week     : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Sales    : int  586953 838022 861991 767198 777392 725924 701517 1027152 755625 445967 ...
 $ Promotion: num  0.89 1.08 0.95 1.06 1.01 1.07 1.22 1.06 1.08 0.8 ...
 $ Feature  : num  0.87 0.84 1.12 0.95 1.06 1.09 1.03 1.08 0.99 0.88 ...
 $ Walmart  : Factor w/ 2 levels "No","Present": 1 1 1 1 1 1 1 1 1 1 ...
 $ Holiday  : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
```

```{r}
summary(dat)
```

```
      week            Sales           Promotion    
 Min.   :  1.00   Min.   : 299359   Min.   :0.790  
 1st Qu.: 25.75   1st Qu.: 512627   1st Qu.:0.940  
 Median : 50.50   Median : 610755   Median :1.010  
 Mean   : 50.50   Mean   : 644054   Mean   :1.011  
 3rd Qu.: 75.25   3rd Qu.: 722809   3rd Qu.:1.062  
 Max.   :100.00   Max.   :1267301   Max.   :1.330  
    Feature          Walmart    Holiday 
 Min.   :0.780   No     :50   No :92  
 1st Qu.:0.940   Present:50   Yes: 8  
 Median :1.015                        
 Mean   :1.007                        
 3rd Qu.:1.080                        
 Max.   :1.260                        
```

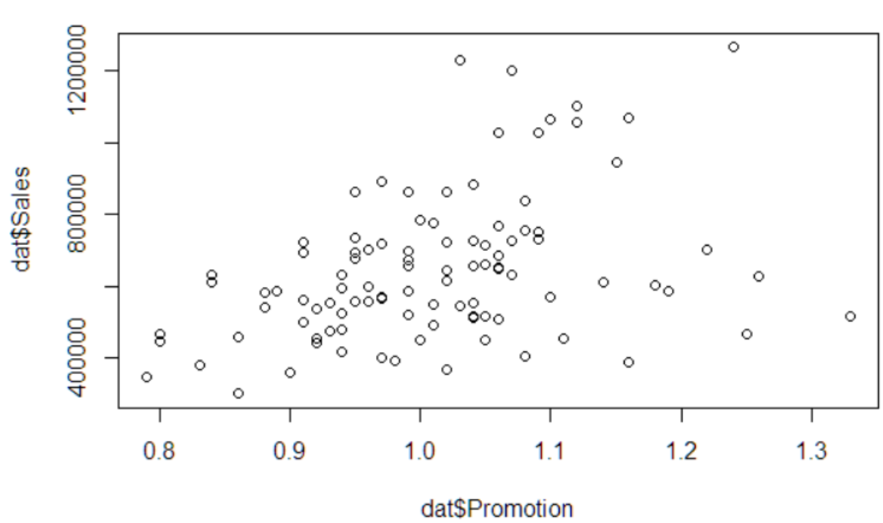
```{r}
dat$logSales=log(dat$Sales)
```

(2). Use the correlation function cor( ) to find the pairwise correlation between the three variables, "Sales", "Promotion" and "Feature". Please post the resulting correlation matrix here. Create a scatter plot for "Sales" and "Promotion". Make another scatter plot for "Sales" and "Feature". Please post the plots here. Create histogram plots for both "Sales" and "logSales". Please post the plots here.

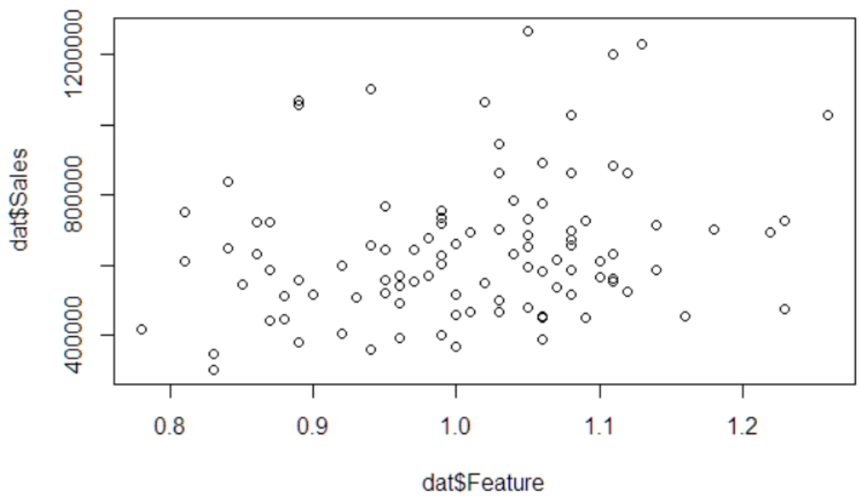
```{r}
cor(dat[,2:4],use="complete.obs", method="pearson")
```

```
              Sales  Promotion    Feature
Sales     1.0000000 0.37739562 0.22438793
Promotion 0.3773956 1.00000000 0.06513678
Feature   0.2243879 0.06513678 1.00000000
```

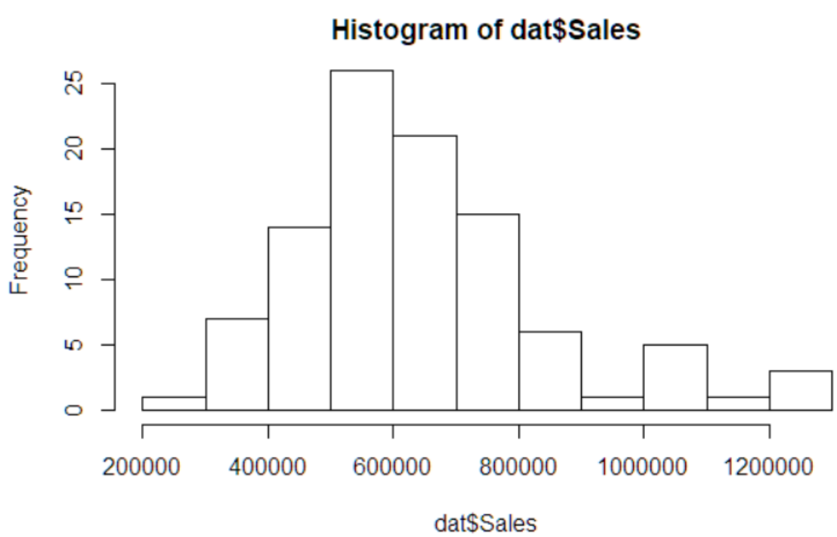
```{r}
plot(dat$Promotion,dat$Sales)
```



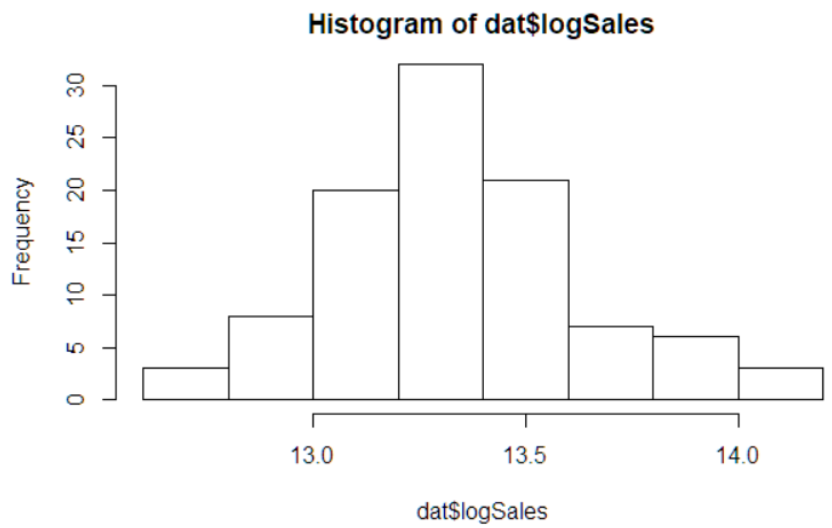
```{r}
plot(dat$Feature,dat$Sales)
```




```{r}
hist(dat$Sales)
```



```{r}
hist(dat$logSales)
```



(3). Estimate the following regression model using the functions lm ( ) and summary( )

log(sales) = $\beta_0 + \beta_1 \times Promotion + \beta_2 \times Feature + \beta_3 \times WalMart + \beta_4 \times Holiday + error$

Paste the R regression output from summary( ) here.

Interpret the estimated coefficients β_1, β_2, β_3, and β_4.

Can we conclude the entry of Wal-mart affects the sales of the local store?

```{r}
a=lm(logSales~Promotion+Feature+Walmart+Holiday,data=dat)
summary(a)
```

```
Call:
lm(formula = logSales ~ Promotion + Feature + Walmart + Holiday,
    data = dat)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.45435 -0.15761 -0.00412  0.12948  0.46955 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    11.85276    0.28826  41.119  < 2e-16 ***
Promotion       0.84754    0.20635   4.107 8.48e-05 ***
Feature         0.75076    0.20774   3.614 0.000485 ***
WalmartPresent -0.31127    0.04233  -7.354 6.76e-11 ***
HolidayYes      0.26004    0.07765   3.349 0.001164 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.21 on 95 degrees of freedom
Multiple R-squared:  0.5206,    Adjusted R-squared:  0.5004 
F-statistic: 25.79 on 4 and 95 DF,  p-value: 1.76e-14
```

Keep all other variables constant:
β_1 means when promotion increases by 1 percent, Sales will increase by (exp(0.84754)-1) percent.
β_2 means when Feature increases by 1 percent, Sales will increase by (exp(0.75076)-1) percent.
β_3 means when Walmart is here, Sales will decrease by (exp(0.31127)-1) percent than it is not here.
β_4 means when it's holiday, Sales will increase by (exp(0.26004)-1) percent than it's not holiday.
Yes, we can interpret the entry of Wal-mart affects the local store negatively, since the beta coefficient is negative and statistically significant.

4). Estimate the following regression model using the functions lm( ) and summary( )

$$log(sales) = \beta_0 + \beta_1 \times Promotion + \beta_2 \times Feature + \beta_3 \times WalMart + \beta_4 \times Holiday + \beta_5 \times Holiday \times WalMart + \beta_6 \times Holiday \times Promotion + error$$

Paste the R regression output from summary( ) here.

Interpret the estimated coefficients β_1, β_2, β_3, β_4, β_5, and β_6.

Compare this model with the one in Question (3) using AIC and BIC. Which is the better model?

```{r}
b=lm(logSales~Promotion+Feature+Walmart+Holiday+Holiday:Walmart+Holiday:Promotion,data=dat)
summary(b)
```

```
Call:
lm(formula = logSales ~ Promotion + Feature + Walmart + Holiday + 
    Holiday:Walmart + Holiday:Promotion, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.44745 -0.14350  0.00013  0.11836  0.47639 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                11.9169     0.2994  39.806  < 2e-16 ***
Promotion                   0.7454     0.2236   3.333  0.00123 ** 
Feature                     0.7828     0.2099   3.729  0.00033 ***
WalmartPresent             -0.2978     0.0439  -6.783 1.08e-09 ***
HolidayYes                 -0.1128     0.7428  -0.152  0.87961    
WalmartPresent:HolidayYes  -0.1307     0.1887  -0.693  0.49034    
Promotion:HolidayYes        0.4330     0.6741   0.642  0.52219    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2101 on 93 degrees of freedom
Multiple R-squared:  0.5302,	Adjusted R-squared:  0.4999 
F-statistic: 17.49 on 6 and 93 DF,  p-value: 1.866e-13
```

```{r}
AIC(a)
```

```
[1] -21.45434
```

```{r}
BIC(a)
```

```
[1] -5.823324
```

```{r}
AIC(b)
```

```{r}
BIC(b)
```

β_1 means when promotion increases by 1 percent, Sales will increase by (exp (0.7454) -1) percent.

β_2 means when Feature increases by 1 percent, Sales will increase by (exp(0.7828)-1) percent.

β_3 means when Walmart is here, Sales will decrease by (exp(0.2978)-1) percent than it is not here.

β_4 means holiday may influence Sales negatively, but we need to be careful to interpret since it is not statistically significant.

β_4 means holiday*Walmartpresence may influence Sales negatively, but we need to be careful to interpret since it is not statistically significant.

β_4 means holiday*promotion may influence Sales positively, but we need to be careful to interpret since it is not statistically significant.

The more negative of AIC and BIC the better. So, the model in Question 3 is better.

Random Effects and Hierarchical Linear Models

In this exercise, we will use hierarchical linear models and regressions with random effects for an analytics problem from a credit card company. The credit card company would like to figure out whether offering more promotions (for example, gasoline rebates and coupons for using the credit card) to their existing customers can increase the share-of-wallet of the credit card (that is, the share of a consumer's monthly spending using the credit card in her total spending). The company would also like to figure out what customer characteristics make them more responsive to promotions.

The company conducted a field experiment by randomly selecting 300 customers and offering them different monthly promotions for 12 months. The share-of-wallet data were recorded in each month for every customer. The data set also included some consumer characteristics. Please download the data "CreditCard_SOW_Data.csv" from Canvas. It has the following variables:

ConsumerID	ID's of the sampled consumers
History	How long (number of months) the customer has been using the card before the experiment
Income	The customer's annual income
WalletShare	The card's share of wallet in the consumer's total monthly spending
Promotion	Index of monthly promotion activity –higher index indicates more pomotions
Balance	The customer's unpaid balance at the beginning of the month

1). Please read the data into R and create a data frame named "sow.data". Please convert consumer ID's to factors and create the following 2 variables in the data frame: logIncome = log(Income) and logSowRatio = log(WalletShare/(1-WalletShare)).

```{r}
sow.data=read.csv("CreditCard_SOW_Data.csv")
sow.data$ConsumerID=as.factor(sow.data$ConsumerID)
sow.data$logIncome=log(sow.data$Income)
sow.data$logSowRatio=log(sow.data$WalletShare/(1-sow.data$WalletShare))
str(sow.data)
```

```
'data.frame':   3600 obs. of  8 variables:
 $ ConsumerID : Factor w/ 300 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ History    : int  55 55 55 55 55 55 55 55 55 55 ...
 $ Income     : num  82000 82000 82000 82000 82000 82000 82000 82000 82000 82000 ...
 $ WalletShare: num  0.643 0.628 0.567 0.638 0.554 0.573 0.666 0.649 0.527 0.459 ...
 $ Promotion  : num  0.5 0.2 1 0.8 0.7 1.1 0.9 0.6 0.1 0 ...
 $ Balance    : int  836 467 1208 792 1215 1248 197 567 1190 1709 ...
 $ logIncome  : num  11.3 11.3 11.3 11.3 11.3 ...
 $ logSowRatio: num  0.588 0.524 0.27 0.567 0.217 ...
```

2). Use the function lm( ) to run the regression

$$logSowRatio_{ij} = \beta_0 + \beta_1 \times History_i + \beta_2 \times Balance_{ij} + \beta_3 \times Promotion_{ij} + \beta_4 \times History_i \times Promotion_{ij} + \beta_5 \times logIncome_i \times Promotion_{ij} + \varepsilon_{ij}$$

Copy and paste the results here.

```{r}
lm1=lm(logSowRatio~History+Balance+Promotion+History:Promotion+logIncome:Promotion,data=sow
.data)
summary(lm1)
```

```
Call:
lm(formula = logSowRatio ~ History + Balance + Promotion + History:Promotion +
    logIncome:Promotion, data = sow.data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.59976 -0.14401  0.00153  0.13634  0.75883

Coefficients:
                      Estimate Std. Error  t value Pr(>|t|)
(Intercept)          8.908e-02  1.603e-02    5.558 2.92e-08 ***
History              1.039e-02  4.153e-04   25.027  < 2e-16 ***
Balance             -4.959e-04  2.882e-06 -172.064  < 2e-16 ***
Promotion            7.777e-01  1.888e-01    4.120 3.87e-05 ***
History:Promotion   -2.598e-03  5.722e-04   -4.541 5.79e-06 ***
Promotion:logIncome -4.558e-02  1.651e-02   -2.760  0.00581 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2078 on 3594 degrees of freedom
Multiple R-squared:  0.8984,    Adjusted R-squared:  0.8982
F-statistic:  6353 on 5 and 3594 DF,  p-value: < 2.2e-16
```

3). Estimate the following hierarchical linear model using the function lmer( ) in the R package "lme4"

$$logSowRatio_{ij} = \beta_{0i} + \beta_1 \times Balance_{ij} + \beta_{2i} \times Promotion_{ij} + \varepsilon_{ij}$$

$$\beta_{0i} = \mu_0 + \mu_1 \times History_i + \zeta_i$$

$$\beta_{2i} = \gamma_0 + \gamma_1 \times History_i + \gamma_2 \times logIncome_i + \xi_i$$

Following what we did in our class, please rewrite this hierarchical linear model as a one-level linear regression model with random effects.

Which variables (and interactions) in the regression have fixed effects? Which ones have random effects? Specify the variables in lmer() and run the regression (please specify REML=F, control=lmerControl(optimizer ="Nelder_Mead") in lmer()). Please copy and paste the summary() of the regression here.

Please interpret the estimated fixed effects in the regression.

Compare model fit using AIC() and BIC() with the model in (2).

```
library(lme4)
re1=lmer(logSowRatio~History+Balance+Promotion+History:Promotion+logIncome:Promotion+(1+Promotion|ConsumerID), data=sow.data, REML=F,
control=lmerControl(optimizer="Nelder_Mead"))
summary(re1)
```

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: logSowRatio ~ History + Balance + Promotion + History:Promotion +
    logIncome:Promotion + (1 + Promotion | ConsumerID)
   Data: sow.data
Control: lmerControl(optimizer = "Nelder_Mead")

     AIC      BIC   logLik deviance df.resid
 -6532.1  -6470.2   3276.0  -6552.1     3590

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.1063 -0.6424  0.0049  0.6336  3.4532

Random effects:
 Groups     Name        Variance  Std.Dev. Corr
 ConsumerID (Intercept) 0.0359421 0.18958
            Promotion   0.0005355 0.02314  0.06
 Residual               0.0066071 0.08128
Number of obs: 3600, groups:  ConsumerID, 300

Fixed effects:
                     Estimate Std. Error  t value
(Intercept)         9.595e-02  2.655e-02    3.613
History             1.039e-02  7.135e-04   14.569
Balance            -5.003e-04  1.799e-06 -278.110
Promotion           6.129e-01  1.466e-01    4.181
History:Promotion  -2.571e-03  2.402e-04  -10.703
Promotion:logIncome -3.110e-02  1.288e-02   -2.414

Correlation of Fixed Effects:
            (Intr) Hstry Balanc Promtn Hstr:P
History     -0.900
Balance     -0.107 -0.001
Promotion   -0.011  0.009  0.013
Hstry:Prmtn  0.143 -0.159 -0.002 -0.153
Prmtn:lgInc  0.001  0.000 -0.012 -0.998  0.099
```

- **Fixed effects:** History, Balance, Promotion, Interaction between History and Promotion, Interaction between logIncome and Promotion
- **Radom effects:** ConsumerID, Interaction between ConsumerID and Promotion
- **Interpretation of the estimation of fixed effect:**
  History: Keep other variables constant, if History increases, SowRate will increase.
  Balance: Keep other variables constant, if balance increases, SowRate will decrease.
  Promotion: Keep other variables constant, if Promotion increases, SowRate will increase.
  History*Promotion: Keep other variables constant, if History*Promotion increase, SowRate will decrease.
  Promotion*logIncome: Keep other variables constant, if Promotion*logIncome increases, SowRate will decrease.

```{r}
AIC(lm1)
```

```
[1] -1087.389
```

```{r}
AIC(re1)
```

```
[1] -6532.094
```

```{r}
BIC(lm1)
```

```
[1] -1044.069
```

```{r}
BIC(re1)
```

```
[1] -6470.207
```

When comparing AIC and BIC (the more negative the better), we can see that the 2^{nd} model with random effects is better than the first linear model.