MAII_A1_Gihani_Dissanayake

Gihani Dissanayake

February 6, 2018

Linear Regression Analysis

First we read in and view the data.

```
library(readr)
walmart = read_csv("~/Walmart_Data.csv")
```

```
## Parsed with column specification:
## cols(
##   week = col_integer(),
##   Sales = col_integer(),
##   Promotion = col_double(),
##   Feature = col_double(),
##   Walmart = col_character(),
##   Holiday = col_character()
## )
```

```
head(walmart)
```

```
## # A tibble: 6 x 6
##    week  Sales Promotion Feature Walmart Holiday
##   <int>  <int>     <dbl>   <dbl>   <chr>   <chr>
## 1     1 586953      0.89    0.87      No      No
## 2     2 838022      1.08    0.84      No      No
## 3     3 861991      0.95    1.12      No      No
## 4     4 767198      1.06    0.95      No      No
## 5     5 777392      1.01    1.06      No      No
## 6     6 725924      1.07    1.09      No      No
```

Part 1

```
walmart$logSales = log(walmart$Sales)
str(walmart)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    100 obs. of  7 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  : chr  "No" "No" "No" "No" ...
##  $ Holiday  : chr  "No" "No" "No" "No" ...
##  $ logSales : num  13.3 13.6 13.7 13.6 13.6 ...
##  - attr(*, "spec")=List of 2
##   ..$ cols   :List of 6
##   .. ..$ week     : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ Sales    : list()
```

```
## .. .. ..- attr(*, "class")= chr "collector_integer" "collector"
## .. ..$ Promotion: list()
## .. .. ..- attr(*, "class")= chr "collector_double" "collector"
## .. ..$ Feature : list()
## .. .. ..- attr(*, "class")= chr "collector_double" "collector"
## .. ..$ Walmart : list()
## .. .. ..- attr(*, "class")= chr "collector_character" "collector"
## .. ..$ Holiday : list()
## .. .. ..- attr(*, "class")= chr "collector_character" "collector"
## ..$ default: list()
## .. ..- attr(*, "class")= chr "collector_guess" "collector"
## ..- attr(*, "class")= chr "col_spec"
```

```r
summary(walmart)
```

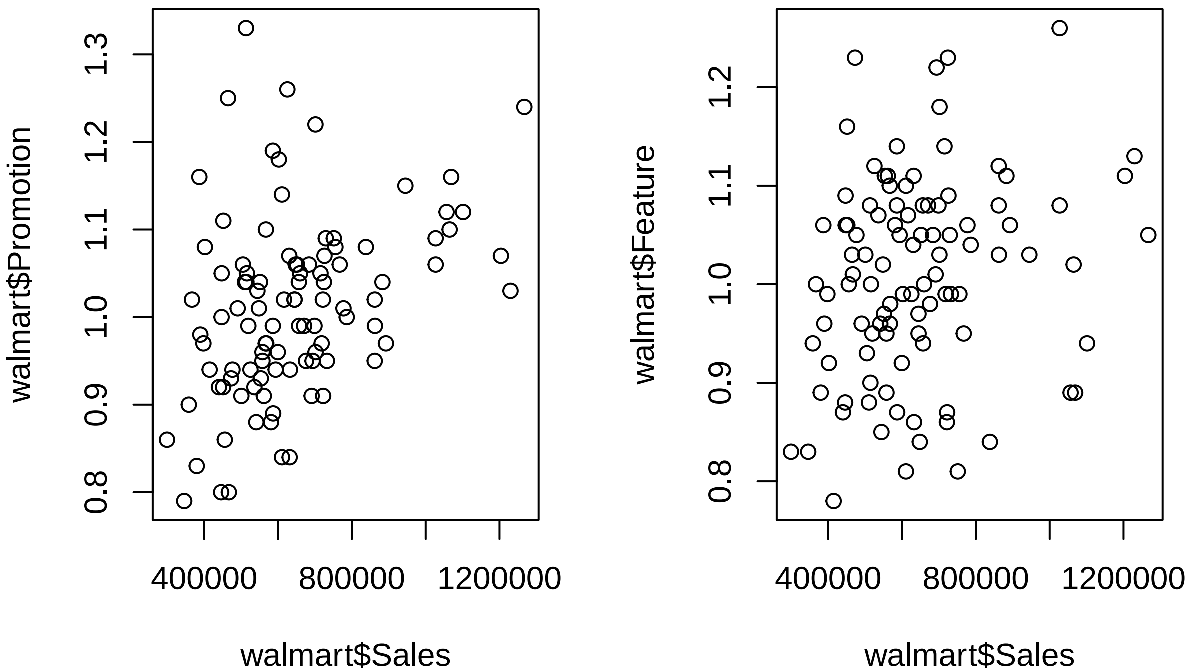
```
##       week           Sales           Promotion        Feature
##  Min.   :  1.00   Min.   : 299359   Min.   :0.790   Min.   :0.780
##  1st Qu.: 25.75   1st Qu.: 512627   1st Qu.:0.940   1st Qu.:0.940
##  Median : 50.50   Median : 610755   Median :1.010   Median :1.015
##  Mean   : 50.50   Mean   : 644054   Mean   :1.011   Mean   :1.007
##  3rd Qu.: 75.25   3rd Qu.: 722809   3rd Qu.:1.062   3rd Qu.:1.080
##  Max.   :100.00   Max.   :1267301   Max.   :1.330   Max.   :1.260
##    Walmart            Holiday             logSales
##  Length:100         Length:100         Min.   :12.61
##  Class :character   Class :character   1st Qu.:13.15
##  Mode  :character   Mode  :character   Median :13.32
##                                        Mean   :13.33
##                                        3rd Qu.:13.49
##                                        Max.   :14.05
```

Part 2

```r
cor(walmart[2:4])
```

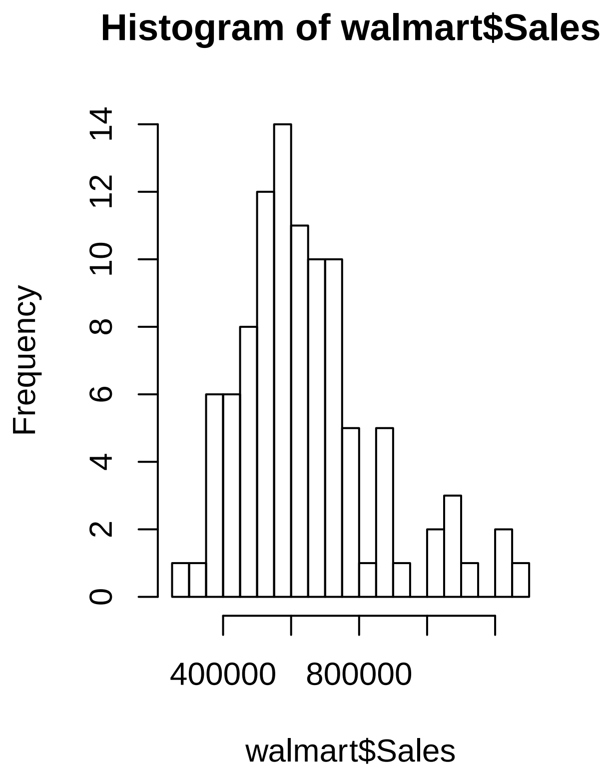
```
##               Sales  Promotion    Feature
## Sales     1.0000000 0.37739562 0.22438793
## Promotion 0.3773956 1.00000000 0.06513678
## Feature   0.2243879 0.06513678 1.00000000
```

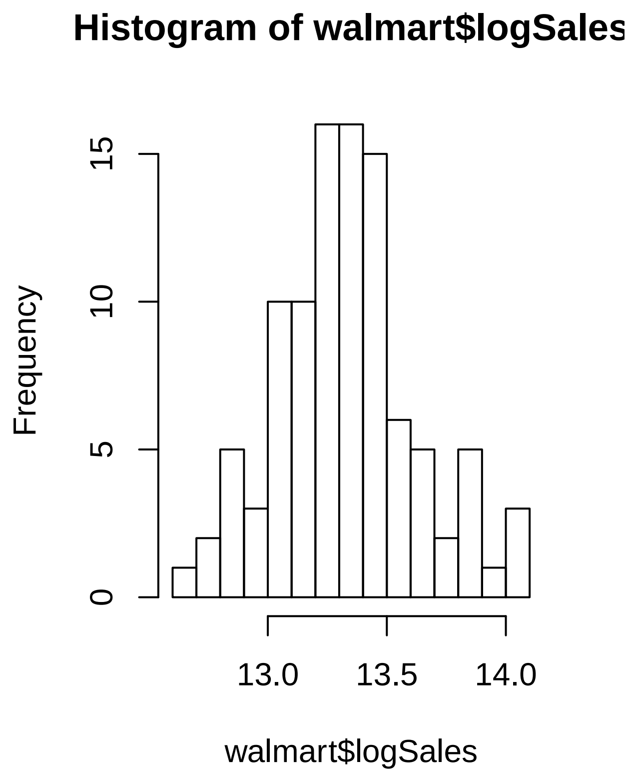
```r
par(mfrow = c(1,2))
plot(walmart$Sales, walmart$Promotion)
plot(walmart$Sales, walmart$Feature)
```



The scatter plots above and the correlations from earlier confirm the positive relationship between sales and promotion, as well as sales and feature. However, the correlation is fairly weak, with both correlation values being less than 0.4. The correlation between sales and promotion is a little stronger than the correlation between sales and feature.

```
par(mfrow = c(1,2))
hist(walmart$Sales, nclass = 20)
hist(walmart$logSales, nclass = 20)
```





As shown in the histograms above, the logSales histogram looks much more like a normal distribution than the Sales histograpm. This is because the sales histogram is a little more skewed to the right than the logSales distribution.

Part 3

```
lm1 = lm(logSales ~ Promotion+Feature+Walmart+Holiday, walmart)
summary(lm1)
```

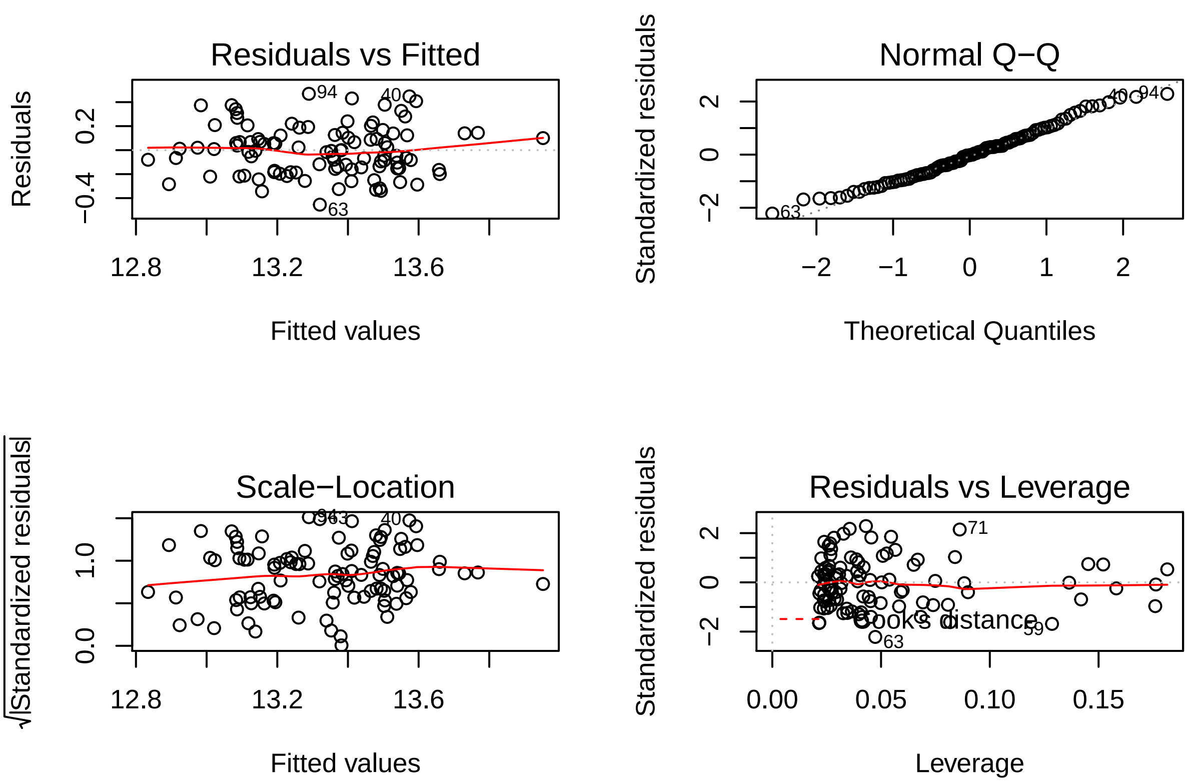
```
##
## Call:
## lm(formula = logSales ~ Promotion + Feature + Walmart + Holiday,
##     data = walmart)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45435 -0.15761 -0.00412  0.12948  0.46955
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    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
```

The coeffecients of the four variables, promotion, feature, walmart, and holiday are all significant, as indicated by their low p-values and the asterisks next to the variable rows. Though promotion, feature, and holiday have a positive coeffcient, walmart has a negative coeffecient. This means that the entry of 1 new walmart is expected to have a -0.3 impact on logSales. This affirms the idea that a new walmart has a negative effect on the sales of the local store.

Conversely, when there is a promotion, feature, and/or holiday, the logSales and therefore actual sales of the local store are expected to increase. Since promotion and feature have a stronger positive weight than the negative effect of the new walmart, the local store should better utilize them to compete with walmart.

```r
par(mfrow =c(2,2))
plot(lm1)
```



Part 4

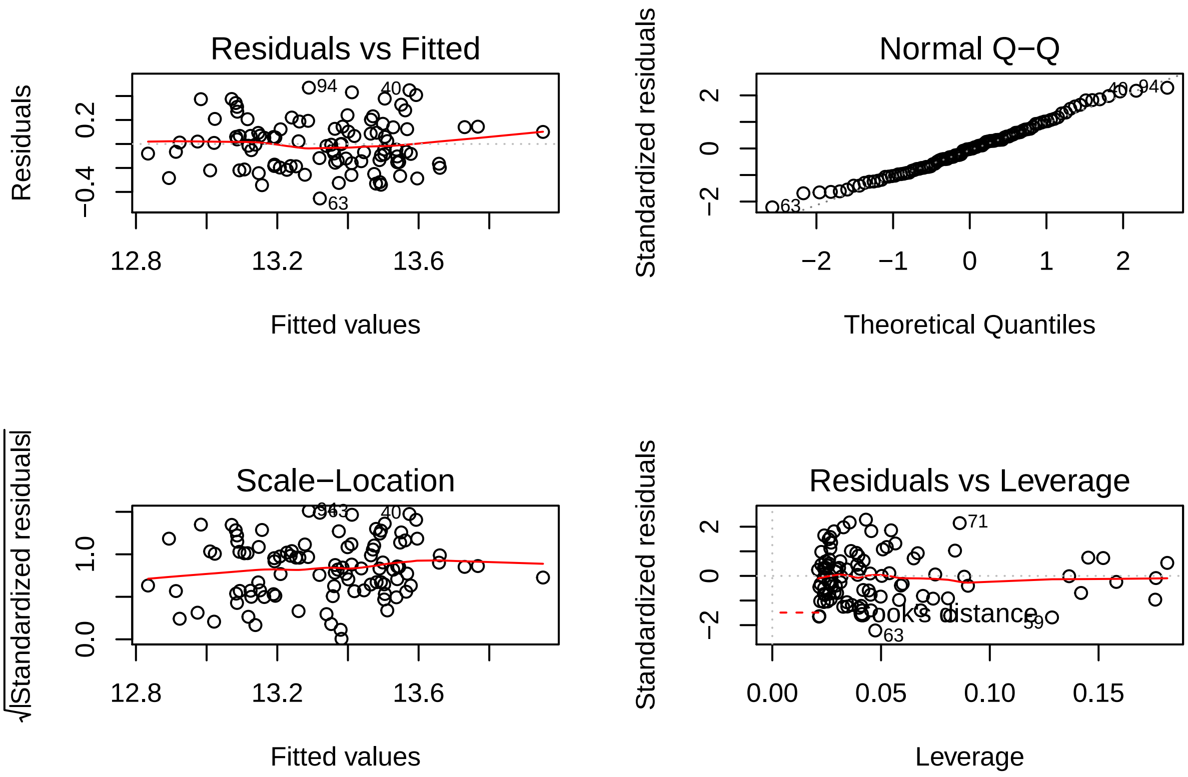
```r
lm2 = lm(logSales ~ Promotion+Feature+Walmart+Holiday+Holiday*Walmart+Holiday*Promotion, walmart)
summary(lm2)
```

```
##
## Call:
```

```
## lm(formula = logSales ~ Promotion + Feature + Walmart + Holiday +
##     Holiday * Walmart + Holiday * Promotion, data = walmart)
##
## 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
```

Unlike the first linear model in which all variables were significant, in this regression, only the promotion, feature, and walmart variables are significant. The interaction terms as well as the holiday variable are not significant predictors of logSales. The coeffecients of the significant variables (promotion, feature, and walmart) are fairly similar to those of the first model. To compare the two, we look to the adjusted R^2, AIC and BIC below.

```
par(mfrow =c(2,2))
plot(lm1)
```




```r
r1 = summary(lm1)$adj.r.squared
r2 = summary(lm2)$adj.r.squared

paste('The adjusted R^2 scores for lm1 and lm2 are:', r1, 'and', r2)
```

```
## [1] "The adjusted R^2 scores for lm1 and lm2 are: 0.500367985095671 and 0.499879843759719"
```

```r
paste('The AIC scores for lm1 and lm2 are:', AIC(lm1), 'and', AIC(lm2))
```

```
## [1] "The AIC scores for lm1 and lm2 are: -21.4543448626203 and -19.4844322312988"
```

```r
paste('The BIC scores for lm1 and lm2 are:', BIC(lm1), 'and', BIC(lm2))
```

```
## [1] "The BIC scores for lm1 and lm2 are: -5.82332374669174 and 1.35692925660595"
```

All three methods of comparison (adjusted R^2, AIC, and BIC) indicate that the first model is superior to the second. Because R^2 value is only slightly higher in the first model, the AIC and BIC are more conclusive towards preferncing the first model. AIC and BIC indicate that less information is lost with the first model than the second.

```r
lm3 = step(lm2, scale = 0, direction = 'backward')
```

```
## Start:  AIC=-305.27
## logSales ~ Promotion + Feature + Walmart + Holiday + Holiday *
##     Walmart + Holiday * Promotion
##
##                     Df Sum of Sq    RSS     AIC
## - Promotion:Holiday  1   0.01822 4.1242 -306.83
## - Walmart:Holiday    1   0.02117 4.1272 -306.76
```

```
## <none>                           4.1060 -305.27
## - Feature            1   0.61401 4.7200 -293.34
##
## Step:  AIC=-306.83
## logSales ~ Promotion + Feature + Walmart + Holiday + Walmart:Holiday
##
##                   Df Sum of Sq    RSS     AIC
## - Walmart:Holiday  1   0.06599 4.1902 -307.24
## <none>                         4.1242 -306.83
## - Feature          1   0.61954 4.7438 -294.83
## - Promotion        1   0.62089 4.7451 -294.81
##
## Step:  AIC=-307.24
## logSales ~ Promotion + Feature + Walmart + Holiday
##
##             Df Sum of Sq    RSS     AIC
## <none>                   4.1902 -307.24
## - Holiday    1   0.49471 4.6849 -298.08
## - Feature    1   0.57605 4.7663 -296.36
## - Promotion  1   0.74412 4.9343 -292.89
## - Walmart    1   2.38510 6.5753 -264.19
```

```
summary(lm3)
```

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

Here we use backward regression on lm2 to find the best model. Interestingly, the best model is what we set as lm1. Scale = 0 is the default, but thought it was worth showing because it indicates AIC()

Random Effects and Hierarchical Linear Models

Part 1

```r
library(readr)
sow.data = read_csv("~/CreditCard_SOW_Data.csv")
```

```
## Parsed with column specification:
## cols(
##   ConsumerID = col_integer(),
##   History = col_integer(),
##   Income = col_double(),
##   WalletShare = col_double(),
##   Promotion = col_double(),
##   Balance = col_integer()
## )
```

```r
head(sow.data)
```

```
## # A tibble: 6 x 6
##   ConsumerID History Income WalletShare Promotion Balance
##        <int>   <int>  <dbl>       <dbl>     <dbl>   <int>
## 1          1      55  82000       0.643       0.5     836
## 2          1      55  82000       0.628       0.2     467
## 3          1      55  82000       0.567       1.0    1208
## 4          1      55  82000       0.638       0.8     792
## 5          1      55  82000       0.554       0.7    1215
## 6          1      55  82000       0.573       1.1    1248
```

```r
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))
head(sow.data)
```

```
## # A tibble: 6 x 8
##   ConsumerID History Income WalletShare Promotion Balance logIncome
##       <fctr>   <int>  <dbl>       <dbl>     <dbl>   <int>     <dbl>
## 1          1      55  82000       0.643       0.5     836  11.31447
## 2          1      55  82000       0.628       0.2     467  11.31447
## 3          1      55  82000       0.567       1.0    1208  11.31447
## 4          1      55  82000       0.638       0.8     792  11.31447
## 5          1      55  82000       0.554       0.7    1215  11.31447
## 6          1      55  82000       0.573       1.1    1248  11.31447
## # ... with 1 more variables: logSowRatio <dbl>
```

Part 2

```r
lm4 = lm(logSowRatio ~ History+Balance+Promotion+History*Promotion+logIncome*Promotion, data = sow.data)
summary(lm4)
```

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

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60005 -0.14319  0.00057  0.13659  0.76013
##
## Coefficients:
##                       Estimate Std. Error  t value Pr(>|t|)
## (Intercept)          2.330e-01  2.581e-01    0.903    0.367
## History              1.037e-02  4.174e-04   24.842  < 2e-16 ***
## Balance             -4.960e-04  2.883e-06 -172.047  < 2e-16 ***
## Promotion            6.097e-01  3.550e-01    1.717    0.086 .
## logIncome           -1.267e-02  2.268e-02   -0.559    0.576
## History:Promotion   -2.571e-03  5.743e-04   -4.476 7.83e-06 ***
## Promotion:logIncome -3.079e-02  3.120e-02   -0.987    0.324
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2078 on 3593 degrees of freedom
## Multiple R-squared:  0.8984, Adjusted R-squared:  0.8982
## F-statistic:  5293 on 6 and 3593 DF,  p-value: < 2.2e-16
```

As shown above, the adjusted R^2 is quite high, and there are three significant variables: history, balance, and the interaction variable between history and promotion.

Part 3

```r
library("lme4")
```

```
## Loading required package: Matrix
```

```r
hlm1 = lmer(logSowRatio ~ History + Balance + Promotion*History + Promotion*logIncome + (1+Promotion|Co
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

```r
summary(hlm1)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula:
## logSowRatio ~ History + Balance + Promotion * History + Promotion *
##     logIncome + (1 + Promotion | ConsumerID)
##    Data: sow.data
## Control: lmerControl(optimizer = "Nelder_Mead")
##
##      AIC      BIC   logLik deviance df.resid
##  -6530.2  -6462.1   3276.1  -6552.2     3589
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1055 -0.6414  0.0047  0.6328  3.4537
##
## Random effects:
##  Groups     Name        Variance  Std.Dev. Corr
##  ConsumerID (Intercept) 0.0359285 0.18955
##             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)            2.427e-01  4.432e-01    0.55
## History                1.037e-02  7.170e-04   14.46
## Balance               -5.003e-04  1.799e-06 -278.11
## Promotion              6.050e-01  1.485e-01    4.07
## logIncome             -1.292e-02  3.895e-02   -0.33
## History:Promotion     -2.570e-03  2.402e-04  -10.70
## Promotion:logIncome -3.040e-02  1.305e-02   -2.33
##
## Correlation of Fixed Effects:
##              (Intr) Histry Balanc Promtn lgIncm Hstr:P
## History     -0.154
## Balance     -0.009  0.000
## Promotion   -0.160  0.025  0.013
## logIncome   -0.998  0.100  0.003  0.160
## Hstry:Prmtn  0.025 -0.160 -0.002 -0.154 -0.016
## Prmtn:lgInc  0.160 -0.016 -0.012 -0.998 -0.160  0.101
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
```

The fixed effects are for history, balance, promotion, logIncome, and the interaction terms between history&promotion and promotion&logIncome.

Fixed effects in an HLM model work like coefficients in a linear model, in which increasing increasing history by one unit results in approximately a 1.037e-02 increase in the y, logSowRatio.This proportional increase is viewed in the estimate column, with the negative values representing an inverse relationship between that fixed effect and logSowRatio

```r
paste('The AIC scores for lm1 and lm2 are:', AIC(lm4), 'and', AIC(hlm1))
```

```
## [1] "The AIC scores for lm1 and lm2 are: -1085.7022256964 and -6530.20412907"
```

```r
paste('The BIC scores for lm1 and lm2 are:', BIC(lm4), 'and', BIC(hlm1))
```

```
## [1] "The BIC scores for lm1 and lm2 are: -1036.19271270084 and -6462.12854870111"
```

Here we compare the original linear model (all fixed effects) with the he=ierarchical linear model with mixed effects (both random and fixed effects). In the HLM model, the intercept and promotion are random variables. To compare the models, we look AIC and BIC, both of which strongly favor the HLM model as shown by the exponentially lower values.