Homework6

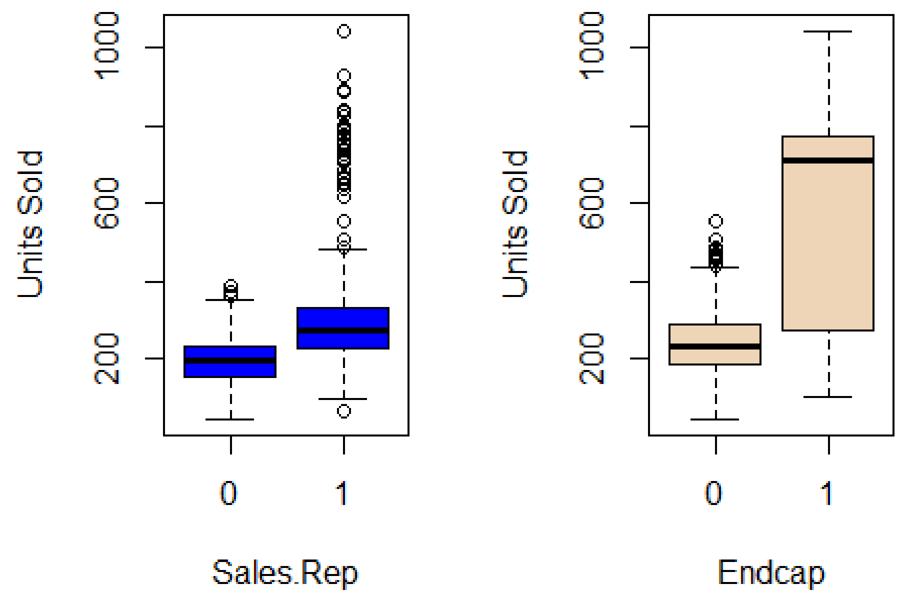
Ritwik Bhriguvanshi

4/09/2020

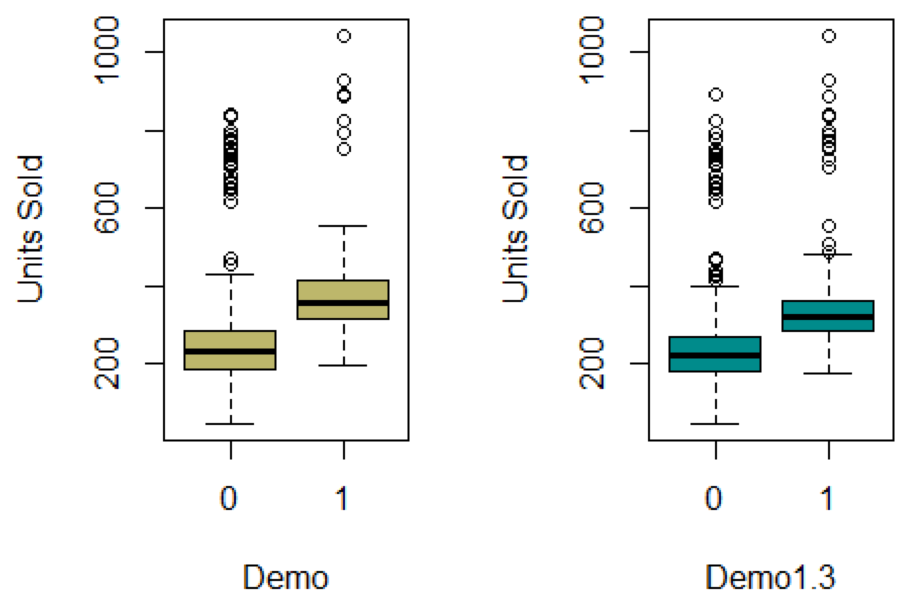
R Markdown

```
## Install libraries

library(MatchIt)

## Warning: package 'MatchIt' was built under R version 3.6.3

library(gridExtra)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:gridExtra':
##
##     combine

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(ggplot2)
library(gplots)

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess

library(corrplot)

## Warning: package 'corrplot' was built under R version 3.6.3

## corrplot 0.84 loaded

## Import the dataset
```

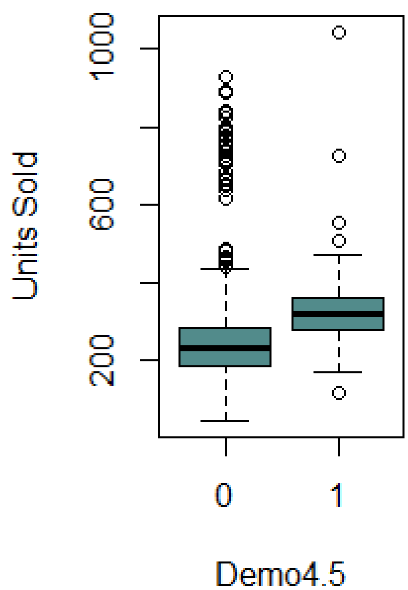
```r
data <- read.csv("GoodBellyData.csv")
dim(data)

## [1] 1386   12

head(data)

##        Date Region                     Store Units.Sold Average.Retail.Price
## 1  5/4/2010     FL Biscayne (aka Aventura)   150.7021             4.390000
## 2 5/11/2010     FL Biscayne (aka Aventura)   197.4038             3.997692
## 3 5/18/2010     FL Biscayne (aka Aventura)   235.1062             3.809231
## 4 5/25/2010     FL Biscayne (aka Aventura)   226.6924             3.835000
## 5  6/1/2010     FL Biscayne (aka Aventura)   257.6882             3.902500
## 6  6/8/2010     FL Biscayne (aka Aventura)   132.9572             4.497692
##   Sales.Rep Endcap Demo Demo1.3 Demo4.5 Natural Fitness
## 1         0      0    0       0       0       1       0
## 2         0      0    0       0       0       1       0
## 3         0      0    0       0       0       1       0
## 4         0      0    0       0       0       1       0
## 5         0      0    0       0       0       1       0
## 6         0      0    0       0       0       1       0

str(data)

## 'data.frame':    1386 obs. of  12 variables:
##  $ Date                : Factor w/ 11 levels "5/11/2010","5/18/2010",..: 4
1 2 3 5 9 6 7 8 11 ...
##  $ Region              : Factor w/ 11 levels "FL","MA","MW",..: 1 1 1 1 1
1 1 1 1 1 ...
##  $ Store               : Factor w/ 126 levels "Academy","Alamo Quarry",..:
14 14 14 14 14 14 14 14 14 14 ...
##  $ Units.Sold          : num  151 197 235 227 258 ...
##  $ Average.Retail.Price: num  4.39 4 3.81 3.83 3.9 ...
##  $ Sales.Rep           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Endcap              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Demo                : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Demo1.3             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Demo4.5             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Natural             : int  1 1 1 1 1 1 1 1 4 1 ...
##  $ Fitness             : int  0 0 0 0 0 0 0 0 0 0 ...

## EDA
par(mfrow=c(1,2))
boxplot(Units.Sold ~ Sales.Rep,data = data, ylab = "Units Sold", col =
"blue")
boxplot(Units.Sold ~ Endcap,data = data, ylab = "Units Sold", col =
"bisque2")
```

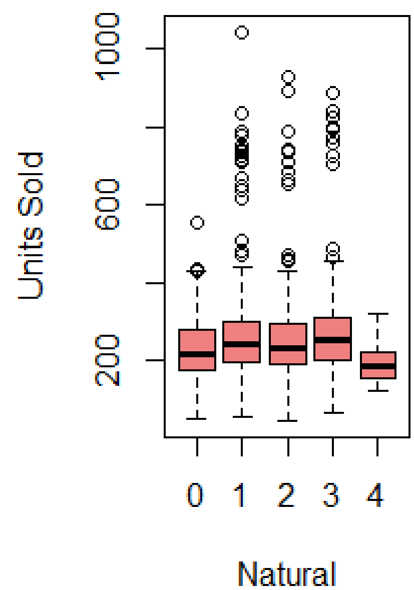



```r
boxplot(Units.Sold ~ Demo,data = data, ylab = "Units Sold", col =
"darkkhaki")
boxplot(Units.Sold ~ Demo1.3,data = data, ylab = "Units Sold", col = "cyan4")
```

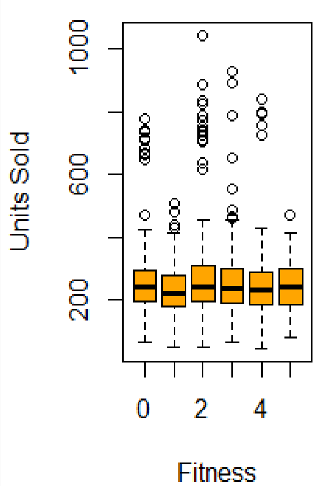


```r
boxplot(Units.Sold ~ Demo4.5,data = data, ylab = "Units Sold", col =
"darkslategray4")
boxplot(Units.Sold ~ Natural,data = data, ylab = "Units Sold", col =
"lightcoral")
```



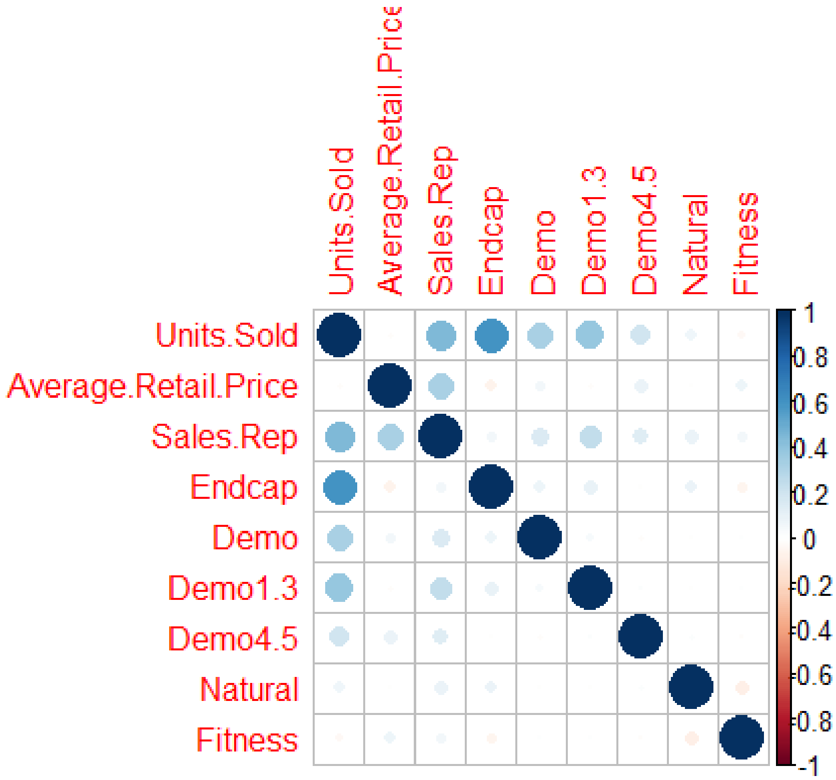



```r
boxplot(Units.Sold ~ Fitness,data = data, ylab = "Units Sold", col =
"orange")
```



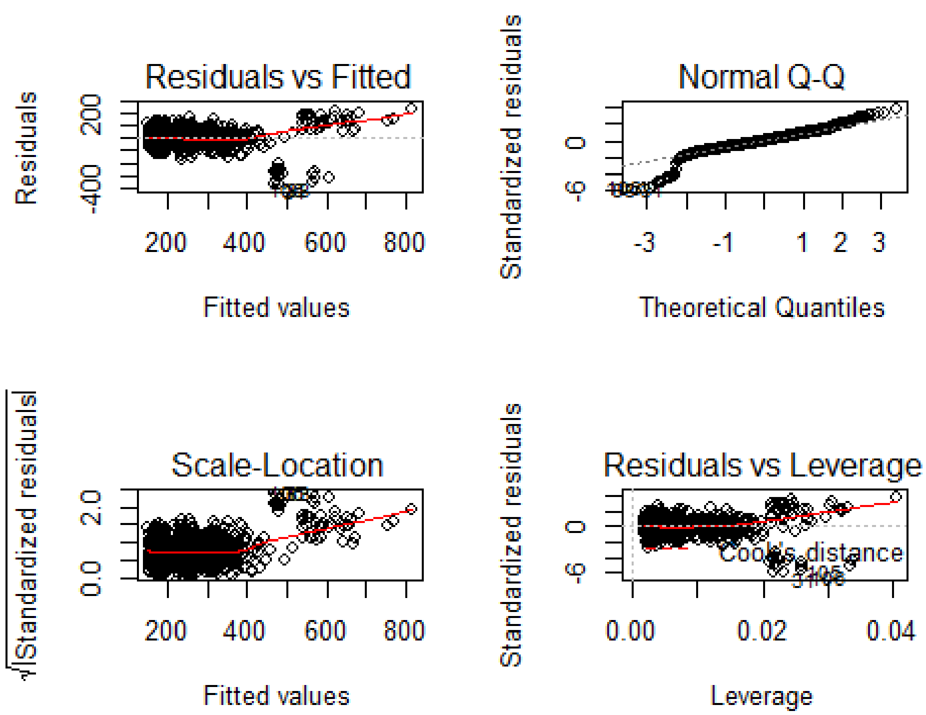
```r
## Check association among the predictor variabes (without Region)

par(mfrow=c(1,1))
```

```r
cormatrix <- cor(data[,c(4:12)])
corrplot(cormatrix)
```



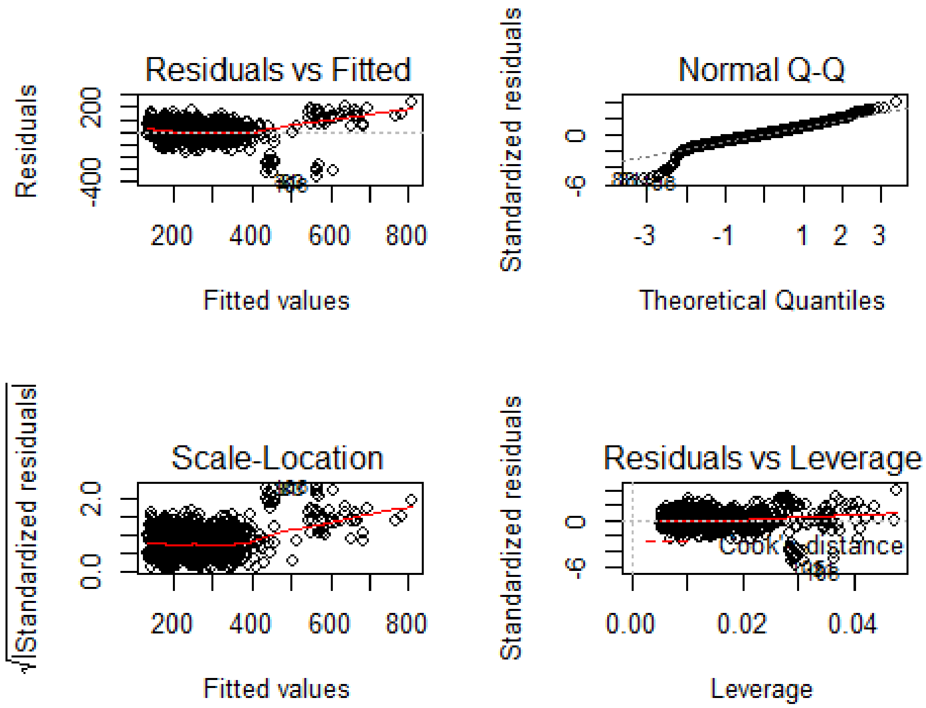

```r
## First Linear Model
## Exclude Date, Store and Region
par(mfrow=c(2,2))
lrmodel1 <- lm(Units.Sold ~., data = data[,c(4:12)])
summary(lrmodel1)

##
## Call:
## lm(formula = Units.Sold ~ ., data = data[, c(4:12)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -363.96  -33.28    0.73   35.84  228.11
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            298.488     16.183  18.444  < 2e-16 ***
## Average.Retail.Price   -28.535      3.952  -7.220 8.56e-13 ***
```

```
## Sales.Rep                77.437      3.864  20.038  < 2e-16 ***
## Endcap                  305.102      9.056  33.692  < 2e-16 ***
## Demo                    111.133      7.404  15.010  < 2e-16 ***
## Demo1.3                  73.517      4.895  15.018  < 2e-16 ***
## Demo4.5                  67.570      6.542  10.329  < 2e-16 ***
## Natural                  -1.594      1.776  -0.897    0.370
## Fitness                  -1.020      1.084  -0.941    0.347
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63.69 on 1377 degrees of freedom
## Multiple R-squared:  0.6726, Adjusted R-squared:  0.6707
## F-statistic: 353.7 on 8 and 1377 DF,  p-value: < 2.2e-16

plot(lrmodel1)
```




```
## Second Linear Model
## Include Region, Exclude Date and Store
## Model to see if region has an effect on Units Sold

par(mfrow=c(2,2))
lrmodel2 <- lm(Units.Sold ~., data =data[,c(2,4:12)])
summary(lrmodel2)

##
## Call:
## lm(formula = Units.Sold ~ ., data = data[, c(2, 4:12)])
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -356.80  -35.22    1.02   37.40  233.90
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          286.8903    21.1406  13.571  < 2e-16 ***
## RegionMA              24.0038     8.5591   2.804 0.005111 **
## RegionMW              60.7394    15.9747   3.802 0.000150 ***
## RegionNAR             31.6328     8.6316   3.665 0.000257 ***
## RegionNC              81.5800    16.2518   5.020 5.85e-07 ***
## RegionNE              54.4885    13.4711   4.045 5.53e-05 ***
## RegionPN              80.1268    16.4637   4.867 1.27e-06 ***
## RegionRM              64.7195    16.5474   3.911 9.64e-05 ***
## RegionSO              31.0466    10.1072   3.072 0.002170 **
## RegionSP              66.4552    16.3813   4.057 5.26e-05 ***
## RegionSW              30.0176     9.9486   3.017 0.002598 **
## Average.Retail.Price -32.6520     4.7842  -6.825 1.32e-11 ***
## Sales.Rep             35.2423    13.5991   2.592 0.009657 **
## Endcap               302.7750     9.3902  32.244  < 2e-16 ***
## Demo                 112.8824     7.4014  15.251  < 2e-16 ***
## Demo1.3               73.8848     4.9371  14.965  < 2e-16 ***
## Demo4.5               65.8542     6.6019   9.975  < 2e-16 ***
## Natural               -1.3787     1.8222  -0.757 0.449412
## Fitness               -0.1166     1.1465  -0.102 0.919013
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63.04 on 1367 degrees of freedom
## Multiple R-squared:  0.6817, Adjusted R-squared:  0.6775
## F-statistic: 162.6 on 18 and 1367 DF,  p-value: < 2.2e-16

plot(lrmodel2)
```



Response: I developed two models for this question - the first excludes region while the next model takes region into consideration. Since region is a factor, we will have more coefficient to interpret for the second model.

EDA Analysis

Looking at the boxplots, we can for all the binary variables - the presence of sales representative, incorporating a demo in any of the corresponding/previous weeks, and having the Endcap promotion will result in greater amount of units sold.

Interpretation for model 1:

Coefficients

A unit increase in Average Retail Price will result in a decrease of Units sold by approximately 28.5 units, keeping every other variable as fixed or constant The presence of a sales rep will increase the units sold by 77 units, keeping every other variable as fixed or constant. The participation of encap promotion by the store will result in an increase in units sold by 305 units,keeping every other variable as fixed or constant. If a store had a demo in the corresponding week, units sold will increase by 111 units, keeping every other variable as fixed or constant. If a store had a demo 1-3 weeks ago, units sold will increase by 73 units approximately, keeping every other variable as fixed or constant. If a store had a demo 4-5 weeks ago, units sold will increase by 67 units approximately, keeping every other variable as fixed or constant.

R-Square

The R-Square for the model is 0.6726% which means that the predictors in the model are able to explain about 67.26% of the variation in the model.

Model Assumption Validity

We can observe violations in normality, constant variance, and linearity. The QQ-plot shows deviations at the both the tails which violates the normality assumption, but it can be ignored because of large sample size. Observing the residual plot, we can see violation of linearity and constant variance, specially at higher values. There's no significant effect of any extreme observation or outliers in the data.

Interpretation for model 2:

Coefficients

The coefficients will have a different interpretation than the model above as we have included region (which is a factor). The base of the 'region' variable which comes alphabeticallty first will be included in the intercept of the model, while the other 10 regions will have an interpretation of its own. To further illustrate, the region 'FL' will be interpreted as the intercept in the second model.

One example of coefficient interpretation in this case (for region MA):

Units Sold = (286.8903 + 24.0038) - 32.65 x Average Retail Price + 35.24 x Sales.Rep + 302.78 x EndCap + 112.88 x Demo + 73.88 x Demo1.3 + 65.85 x Demo4.5 - 1.38 x Natural - 0.12 x Fitness

R-Square

The R-Square for the model is 0.6817% which means that the predictors in the model are able to explain about 68.17% of the variation in the model.

Model Assumption Validity

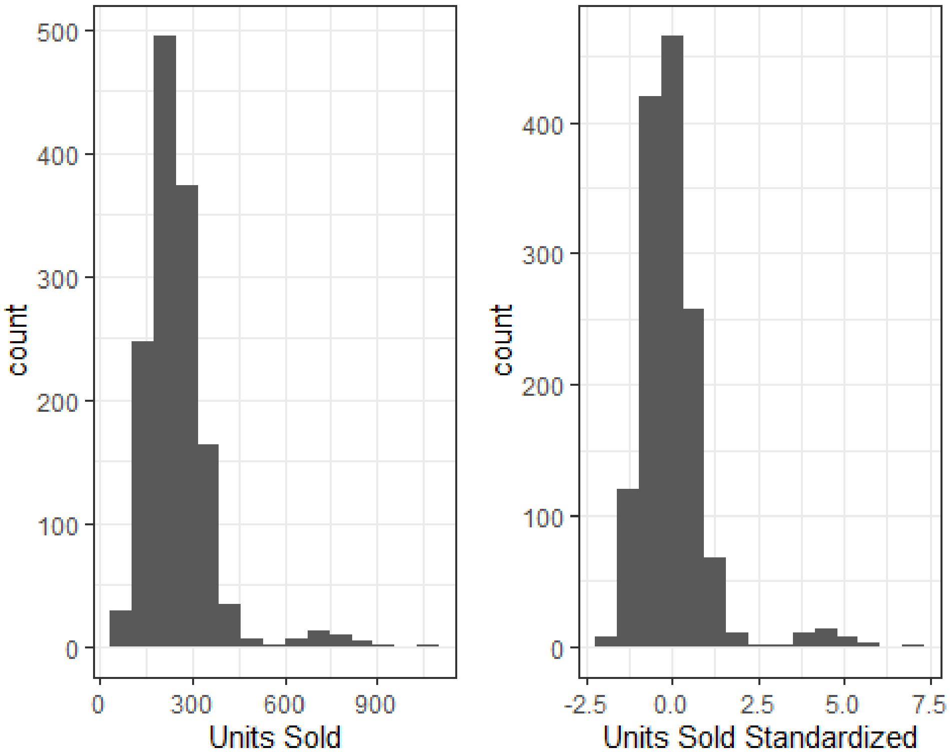
More or less the same interpretation stated above for model 1. Same problems are associated with the second model.

Conclusion

We could say that including region in the model as a predictor variable does not significantly imporve the model as the R-square remain same more or less. Additionally, both the plots see violations in assumptions. Including region in the model would be specifically beneficial if we want to focus to region wise effect of EndCap.

```r
pp1 <- ggplot(data,aes(Units.Sold)) + geom_histogram(bins = 15) + theme_bw()
+ labs(x="Units Sold")

mu <- mean(data$Units.Sold)
std <- sqrt(var(data$Units.Sold))
Units.Sold_std <- (data$Units.Sold-mu)/std
```

```r
pp2 <- ggplot(data,aes(Units.Sold_std)) + geom_histogram(bins = 15) +
theme_bw() + labs(x="Units Sold Standardized")
grid.arrange(pp1,pp2, ncol =2)
```




```r
# Remove Date, Store, Region
GoodBelly <- data[,-c(1:3)]

# Grouping by Endcap
# Computing means for control and treatment cases for each covariate

cbind(GoodBelly %>%
        group_by(Endcap) %>%
        summarize_all(funs(mean(., na.rm = TRUE))),
        count = c(table(GoodBelly$Endcap)))

## Warning: funs() is soft deprecated as of dplyr 0.8.0
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once per session.
```

```
##   Endcap Units.Sold Average.Retail.Price Sales.Rep       Demo   Demo1.3
## 0      0   240.6958             4.113314 0.5446362 0.05476369 0.1500375
## 1      1   583.9250             3.950616 0.6792453 0.15094340 0.3207547
##      Demo4.5  Natural  Fitness count
## 0 0.07576894 1.432858 2.495124  1333
## 1 0.07547170 1.849057 2.000000    53
```

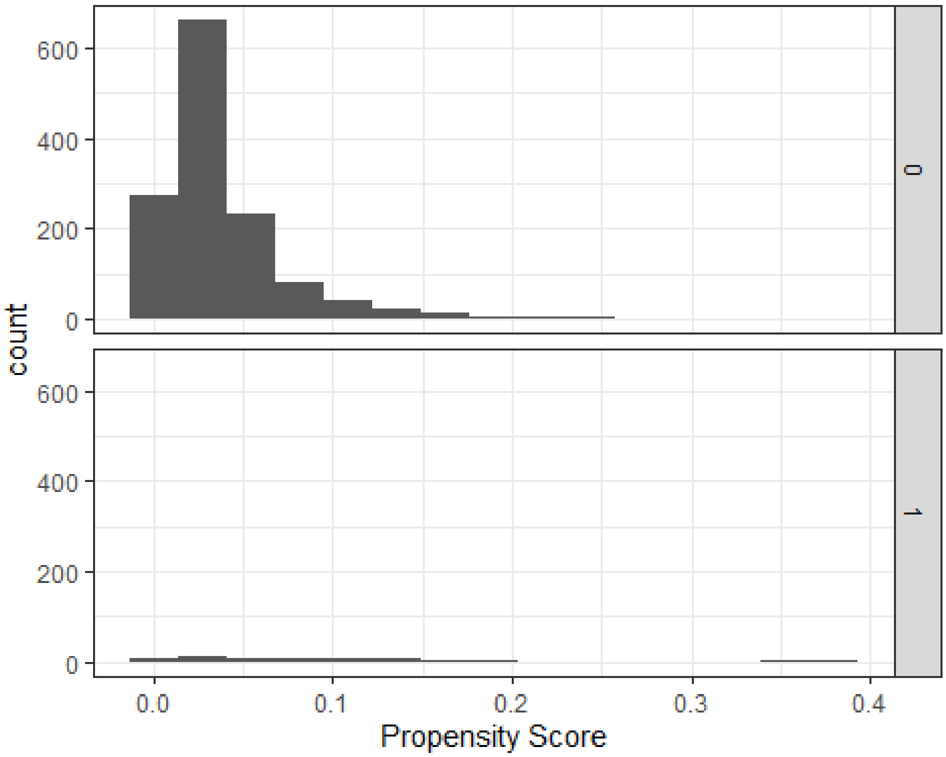
```
# Testing for significant differences in covariate distributions
# Extracting P-values
c(Units.Sold <- with(GoodBelly, t.test(Units.Sold ~ Endcap))$p.value,
  Average.Retail.Price <- with(GoodBelly, t.test(Average.Retail.Price ~
Endcap))$p.value,
  Sales.Rep <- with(GoodBelly, t.test(Sales.Rep ~ Endcap))$p.value,
  Demo <- with(GoodBelly, t.test(Demo ~ Endcap))$p.value,
  Demo1.3 <- with(GoodBelly, t.test(Demo1.3 ~ Endcap))$p.value,
  Demo4.5 <- with(GoodBelly, t.test(Demo4.5 ~ Endcap))$p.value,
  Natural <- with(GoodBelly, t.test(Natural ~ Endcap))$p.value,
  Fitness <- with(GoodBelly, t.test(Fitness ~ Endcap))$p.value)
```

```
## [1] 8.507279e-13 3.329640e-02 4.654859e-02 5.988948e-02 1.173887e-02
## [6] 9.936771e-01 8.783933e-04 8.571098e-03
```

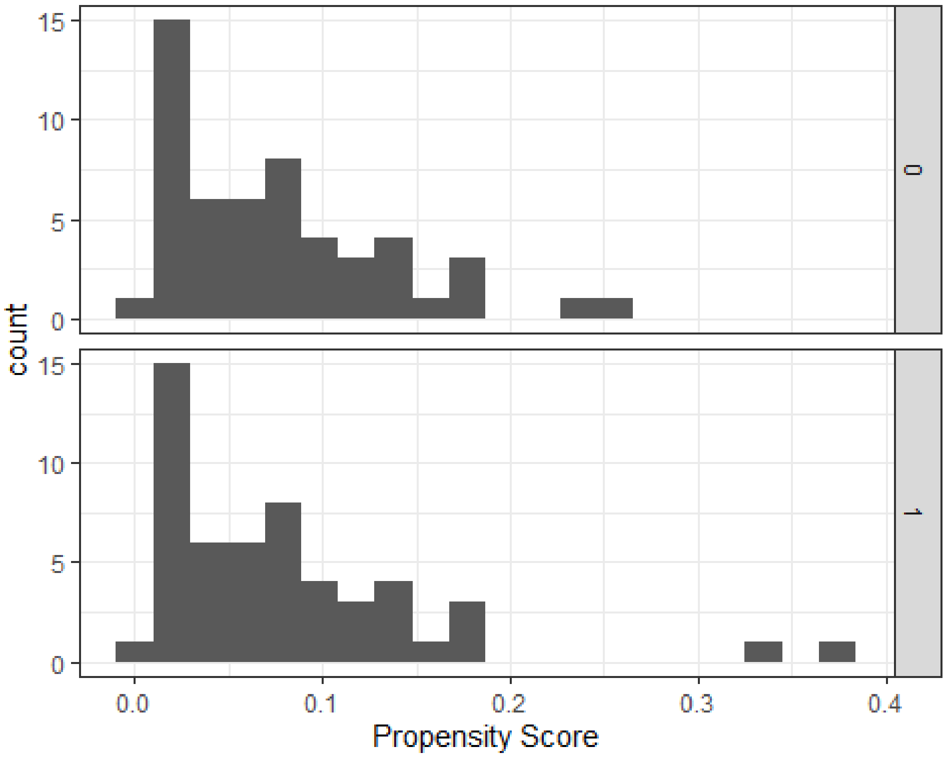
Response: There are significant differences in covariate distribution for predictor variables - Units Sold, Average Retail Price, Sales Representative, Demo1.3, Natural, and Fitness because the P-value from the t-test is less than 5% which makes it significant.

However, at a 10% cut-off, all the covariate distribution for all predictor variables will be significant, leaving ‘Demo4.5’

```
# Look at propensity scores before matching
pscores_pre_match <- glm(Endcap ~ Average.Retail.Price + Sales.Rep + Demo +
Demo1.3 + Demo4.5 + Natural + Fitness, family = binomial(link = "logit"),
data = GoodBelly)

GoodBelly$pscores_pre_match <- predict(pscores_pre_match, type = "response")

ggplot(GoodBelly, aes(pscores_pre_match)) + geom_histogram(bins=15) +
facet_grid(vars(Endcap)) + theme_bw() + labs(x="Propensity Score")
```




```
GoodBelly$pscores_pre_match <- NULL

## Propensity Score Matching (Nearest Neighbor)

mod_match <- matchit(Endcap ~ Average.Retail.Price + Sales.Rep + Demo +
Demo1.3 + Demo4.5 + Natural + Fitness, method = "nearest", caliper = 0, ratio
= 1, data = GoodBelly)

matched_GoodBelly <- match.data(mod_match)
head(matched_GoodBelly,3)

##    Units.Sold Average.Retail.Price Sales.Rep Endcap Demo Demo1.3 Demo4.5
## 28   199.9339             4.406667         0      1    0       0       0
## 29   217.9519             4.358750         0      1    0       0       0
## 30   139.6135             4.390000         0      1    0       0       0
##    Natural Fitness   distance weights
## 28       2       1 0.02014844       1
## 29       2       1 0.02108017       1
## 30       2       1 0.02046785       1

# Examine region of common support in the matched data

ggplot(matched_GoodBelly, aes(distance)) + geom_histogram(bins =20) +
facet_grid(vars(Endcap)) + theme_bw() + labs(x = "Propensity Score")
```

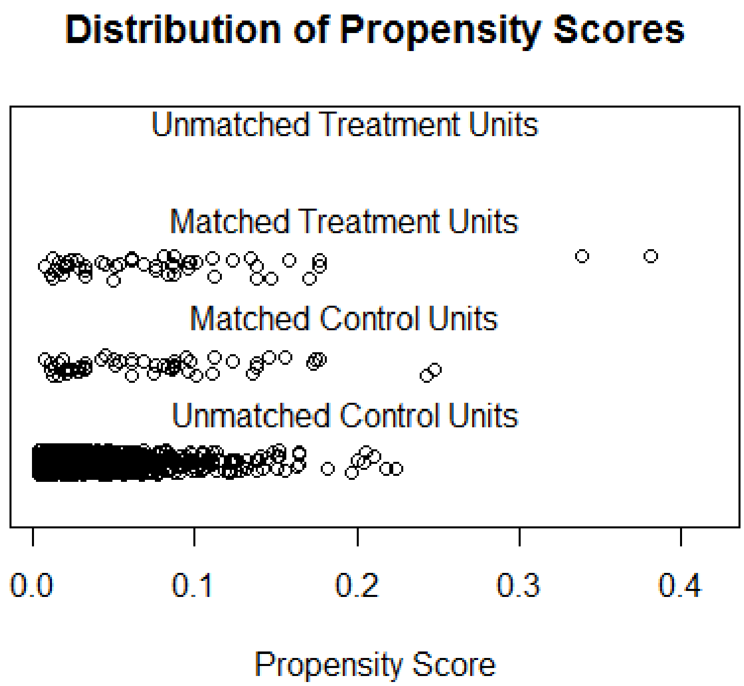



```
# Checking Covariate Balance Post Matching

p1 <-
ggplot(matched_GoodBelly,aes(x=distance,y=Units.Sold,color=factor(Endcap))) +
geom_point() + geom_rug(sides="trbl") + theme(legend.position = "none") +
labs(x="Propensity Score")

p2 <-
ggplot(matched_GoodBelly,aes(x=distance,y=Average.Retail.Price,color=factor(E
ndcap))) +
geom_point() + theme(legend.position = "none") + labs(x="Propensity Score")

p3 <-
ggplot(matched_GoodBelly,aes(x=distance,y=Sales.Rep,color=factor(Endcap))) +
geom_point() + theme(legend.position = "none") + labs(x="Propensity Score")

p4 <-
ggplot(matched_GoodBelly,aes(x=distance,y=Natural,color=factor(Endcap))) +
geom_point() + theme(legend.position = "none") + labs(x="Propensity Score")

p5 <-
ggplot(matched_GoodBelly,aes(x=distance,y=Fitness,color=factor(Endcap))) +
geom_point() + theme(legend.position = "none") + labs(x="Propensity Score")

p6 <- ggplot(matched_GoodBelly,aes(x=distance,y=Demo,color=factor(Endcap))) +
geom_point() +  geom_rug(sides="trbl") + labs(x="Propensity Score")
```

```
# Arrange the plots in a grid
grid.arrange(p1,p2,p3,p4,p5,p6)
```


```
# covariate balance comparisons, both before and after matching
summary(mod_match)$sum.all

##                        Means Treated Means Control SD Control     Mean Diff
## distance                  0.07971863    0.03659033 0.03359098  0.0431282985
## Average.Retail.Price      3.95061565    4.11331445 0.45991479 -0.1626987956
## Sales.Rep                 0.67924528    0.54463616 0.49819053  0.1346091240
## Demo                      0.15094340    0.05476369 0.22760380  0.0961797053
## Demo1.3                   0.32075472    0.15003751 0.35724221  0.1707172076
## Demo4.5                   0.07547170    0.07576894 0.26472738 -0.0002972441
## Natural                   1.84905660    1.43285821 0.97692630  0.4161983892
## Fitness                   2.00000000    2.49512378 1.60124255 -0.4951237809
##                           eQQ Med   eQQ Mean   eQQ Max
## distance             0.03711551 0.04120498 0.194802
## Average.Retail.Price 0.12820513 0.17553641 1.392308
## Sales.Rep            0.00000000 0.13207547 1.000000
## Demo                 0.00000000 0.09433962 1.000000
## Demo1.3              0.00000000 0.16981132 1.000000
## Demo4.5              0.00000000 0.00000000 0.000000
## Natural              0.00000000 0.47169811 1.000000
## Fitness              1.00000000 0.54716981 2.000000
```

```r
# covariate balance in the matched data
summary(mod_match)$sum.matched

##                      Means Treated Means Control SD Control    Mean Diff
## distance                0.07971863    0.07527513  0.0591548  0.004443503
## Average.Retail.Price    3.95061565    4.05461033  0.5218840 -0.103994674
## Sales.Rep               0.67924528    0.73584906  0.4450991 -0.056603774
## Demo                    0.15094340    0.20754717  0.4094316 -0.056603774
## Demo1.3                 0.32075472    0.33962264  0.4781131 -0.018867925
## Demo4.5                 0.07547170    0.09433962  0.2950978 -0.018867925
## Natural                 1.84905660    1.79245283  0.9273305  0.056603774
## Fitness                 2.00000000    2.05660377  1.5982574 -0.056603774
##                           eQQ Med    eQQ Mean   eQQ Max
## distance             8.102077e-05 0.004606868 0.1335869
## Average.Retail.Price 1.068269e-01 0.132821686 0.2967857
## Sales.Rep            0.000000e+00 0.056603774 1.0000000
## Demo                 0.000000e+00 0.056603774 1.0000000
## Demo1.3              0.000000e+00 0.018867925 1.0000000
## Demo4.5              0.000000e+00 0.018867925 1.0000000
## Natural              0.000000e+00 0.132075472 1.0000000
## Fitness              0.000000e+00 0.396226415 1.0000000

summary(mod_match)$nn

##           Control Treated
## All          1333      53
## Matched        53      53
## Unmatched    1280       0
## Discarded       0       0

## Propensity score region overlap comparisons
plot(mod_match, type = 'jitter', interactive = FALSE)
```




```r
## Estimating and Building the ATE models

ATE_lm <- lm(Units.Sold ~ factor(Endcap), data = matched_GoodBelly)
coef(summary(ATE_lm))

##                   Estimate Std. Error  t value     Pr(>|t|)
## (Intercept)       285.2677   27.42645 10.40119 8.538861e-18
## factor(Endcap)1   298.6573   38.78686  7.69996 8.267483e-12

ATE_lm2 <- lm(Units.Sold ~ factor(Endcap) + Average.Retail.Price + Sales.Rep
+ Demo + Demo1.3 + Demo4.5 + Natural + Fitness,data = matched_GoodBelly)
coef(summary(ATE_lm2))

##                         Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)           319.156188  92.955821  3.4334180 8.780170e-04
## factor(Endcap)1       317.554430  22.285793 14.2491870 1.604905e-25
## Average.Retail.Price  -72.762120  22.191721 -3.2787958 1.448140e-03
## Sales.Rep             324.274399  27.631542 11.7356606 2.549208e-20
## Demo                  107.656236  30.032433  3.5846658 5.303749e-04
## Demo1.3                39.417319  25.984888  1.5169324 1.325361e-01
## Demo4.5                16.637675  40.496369  0.4108436 6.820934e-01
## Natural               -12.879061  12.841863 -1.0028966 3.184060e-01
## Fitness                 4.036402   8.112531  0.4975516 6.199262e-01
```

Response:

There is evidence for satisfactory covariate balancing after performing the matching analysis. The means of both the groups(treated and control) are within 0.1 (10%) for all the variables which is quite reasonable. The distribution match well among each other, but on the contrary only 53 values were matched as reported above. There is existence of large number of unmatched values, 1280 to be precise in the control group. We can try to match more values by tweaking the caliper value or using a different ratio. For a few variables, the mean difference of propensity scores for pre-matching have large differences(for treated and control groups) such as Demo1.3. The model in (c) estimates an average of additional 315 units sole per week if the endcap promotion is incorporated. Comparing the result to the model build in (a), we can say that greater increase in units sold can be found here as the first model was expecting an increase of 305 units approximately. Additionally, the newer model build in (c) has fewer significant variables - Average retail price, demo, sales representative, and endcap(1) are significant having a p-value less than 5%.

```
# For Control Cases, we will use Endcap =0
GoodBelly_control <- subset(GoodBelly,Endcap==0)
GoodBelly_control$Endcap <- NULL

## Model for Control Cases
control_model <- lm(Units.Sold ~. , data = GoodBelly_control)
summary(control_model)

##
## Call:
## lm(formula = Units.Sold ~ ., data = GoodBelly_control)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -168.201  -33.902   -0.909   33.360  182.769
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         277.1808    12.8920  21.500  < 2e-16 ***
## Average.Retail.Price -22.3868     3.1520  -7.102 1.99e-12 ***
## Sales.Rep            59.6792     3.0660  19.465  < 2e-16 ***
## Demo                104.7769     6.0052  17.448  < 2e-16 ***
## Demo1.3              73.4880     3.9133  18.779  < 2e-16 ***
## Demo4.5              73.5105     5.1695  14.220  < 2e-16 ***
## Natural               0.2344     1.3983   0.168    0.867
## Fitness               0.1707     0.8508   0.201    0.841
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.26 on 1325 degrees of freedom
## Multiple R-squared:  0.5644, Adjusted R-squared:  0.5621
## F-statistic: 245.3 on 7 and 1325 DF,  p-value: < 2.2e-16

# For Treated cases, we will use Endcap =1
GoodBelly_treated <- subset(GoodBelly,Endcap==1)
```

```r
GoodBelly_treated$Endcap <- NULL

## Model for Treated Cases
treated_model <- lm(Units.Sold ~. , data = GoodBelly_treated)
summary(treated_model)
```

```
##
## Call:
## lm(formula = Units.Sold ~ ., data = GoodBelly_treated)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -76.423 -25.674   5.215  27.696  77.859
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           244.2096    51.2483   4.765 2.00e-05 ***
## Average.Retail.Price -11.3995    12.9470  -0.880 0.383280
## Sales.Rep             508.9102    15.2498  33.372  < 2e-16 ***
## Demo                  129.7649    18.7276   6.929 1.30e-08 ***
## Demo1.3                80.5923    17.5784   4.585 3.61e-05 ***
## Demo4.5               104.6782    25.0403   4.180 0.000132 ***
## Natural                -7.2572     8.7253  -0.832 0.409945
## Fitness                -0.4216     6.1639  -0.068 0.945769
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44.87 on 45 degrees of freedom
## Multiple R-squared:  0.9754, Adjusted R-squared:  0.9715
## F-statistic: 254.6 on 7 and 45 DF,  p-value: < 2.2e-16
```

```r
## Obtain the predicted values and impute the counterfactuals

impute_control <- predict(control_model,newdata = GoodBelly_treated)
impute_treated <- predict(treated_model,newdata = GoodBelly_control)

complete_data_treated_cases <- data.frame(cbind(Under_Treatment =
GoodBelly_treated$Units.Sold,
Under_Control = impute_control))

complete_data_control_cases <- data.frame(cbind(Under_Treatment =
impute_treated,
Under_Control = GoodBelly_control$Units.Sold))

complete_data <-
rbind(complete_data_treated_cases,complete_data_control_cases)
head(complete_data,5)
```

```
##    Under_Treatment Under_Control
## 28        199.9339      179.1692
```

```
## 29         217.9519      180.2419
## 30         139.6135      179.5423
## 31         106.4166      176.6148
## 32         203.7432      190.4773
```

```
t.test(complete_data$Under_Treatment,complete_data$Under_Control,paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  complete_data$Under_Treatment and complete_data$Under_Control
## t = 40.26, df = 1385, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  239.4816 264.0143
## sample estimates:
## mean of the differences
##                 251.748
```

```
# fraction of the treatment outcome exceed the control outcome
mean(complete_data_treated_cases$Under_Treatment >
complete_data_treated_cases$Under_Control)
```

```
## [1] 0.8679245
```

```
#  fraction of the treatment outcomes exceed the control outcomes
mean(complete_data_control_cases$Under_Treatment >
complete_data_control_cases$Under_Control)
```

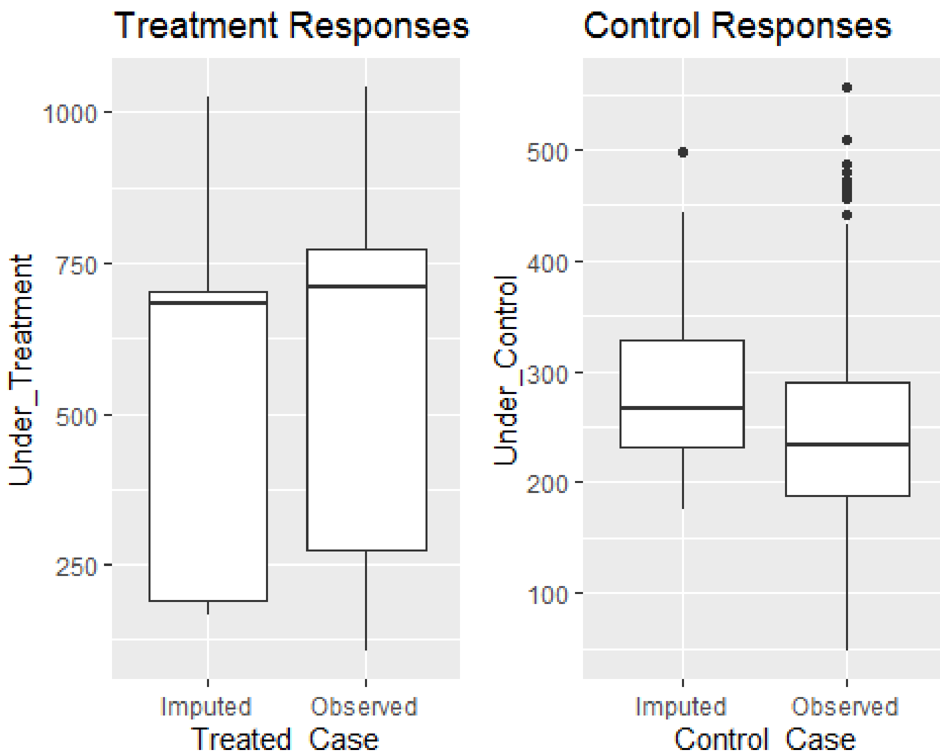
```
## [1] 0.7704426
```

```
# Compare the observed and imputed values
complete_data$Treated_Case <- c(rep("Observed",length(impute_control)),
                           rep("Imputed",length(impute_treated)))

complete_data$Control_Case <- c(rep("Imputed",length(impute_control)),
                            rep("Observed",length(impute_treated)))

head(complete_data,5)
```

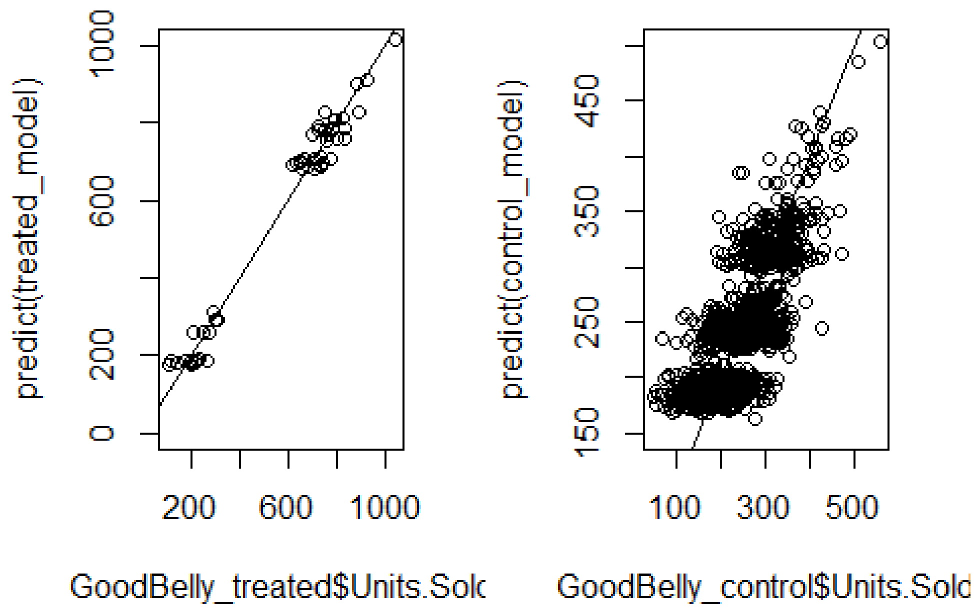
```
##    Under_Treatment Under_Control Treated_Case Control_Case
## 28         199.9339      179.1692     Observed      Imputed
## 29         217.9519      180.2419     Observed      Imputed
## 30         139.6135      179.5423     Observed      Imputed
## 31         106.4166      176.6148     Observed      Imputed
## 32         203.7432      190.4773     Observed      Imputed
```

```
plot1 <- ggplot(complete_data,aes(x=Treated_Case,y=Under_Treatment)) +
geom_boxplot() + labs(title = "Treatment Responses")

plot2 <- ggplot(complete_data,aes(x=Control_Case,y=Under_Control)) +
geom_boxplot() + labs(title = "Control Responses")
```

```r
grid.arrange(plot1,plot2,ncol=2)
```




```r
# Regression Towards Mean
# Check if Model yields good results

par(mfrow=c(1,2))
plot(GoodBelly_treated$Units.Sold,predict(treated_model),
ylim=c(0,1000),abline(a=0,b=1))

plot(GoodBelly_control$Units.Sold,predict(control_model),
ylim=c(150,500),abline(a=0,b=1))
```



Response: Regression approach illustrates that mean difference between the treatment data and control data is significant at the 5% level. With this method, we can expect about 252 units to be sold by incorporating the endcap promotion at the stores. This number is lower than our results found in (a) and (c).

##Response to Question 1e

In the models developed in A, C, and D, we can observe that existence of an endcap promotion is considered significant and would result in greater increase for units sold. The t-test performed in (b) portrays significant differences in covariate distributions. The first model build in (a) emphasizes an increase in units sold by approximately 305 units with an introduction of endcap promotion in the stores. As an interest, we see the number of units sold rising from 305 to 315 (approximately) for the model in (c) after matching the propensity scores when an endcap promotion is incorporated. The regression approach in (d) depicts that presence of endcap promotion would result in an increase of about 252 units. Final recommendation to GoodBelly would be to adopt the endcap program as it results in additional units being sold with all the stated models. The best model to chose would be the model developed in (c).