Predictive Analytics Homework- Promotion Effect of Bobblehead

Godfred Somua-Gyimah

January 10, 2017

Contents

PROBLEM

The management of the Los Angeles Dodgers want to know if the sale of bobbleheads can help attract additional fans to the park. Using Major League Baseball data from the 2012 season, advise management on the following questions:

Questions: (1) Do bobblehead promotions have a positive effect on attendance? (2) If they do, how big is this positive effect? (3) Will the increased revenues associated with tickets and concessions cover the fixed and variable costs of putting on the promotion?

The following R packages will be used:

- dpyr: data transformation
- lattice: plotting
- stargazer: report format
- car: regression
- MASS: statistics

1. Read in Dataset

```
# Clean the environment
rm(list = ls())
# Read data file
df <- read.csv("dodgers.csv")
```

2. Undertand Data

```
# Show head
head(df)
```

```
##   month day attend day_of_week opponent temp  skies day_night cap shirt
## 1   APR  10  56000     Tuesday  Pirates   67 Clear        Day  NO    NO
## 2   APR  11  29729   Wednesday  Pirates   58 Cloudy     Night  NO    NO
## 3   APR  12  28328    Thursday  Pirates   57 Cloudy     Night  NO    NO
## 4   APR  13  31601      Friday   Padres   54 Cloudy     Night  NO    NO
## 5   APR  14  46549    Saturday   Padres   57 Cloudy     Night  NO    NO
## 6   APR  15  38359      Sunday   Padres   65 Clear        Day  NO    NO
##   fireworks bobblehead
## 1        NO         NO
## 2        NO         NO
## 3        NO         NO
## 4       YES         NO
## 5        NO         NO
## 6        NO         NO
```

```
# Show the structure of the data frame
str(df)
```

```
## 'data.frame':    81 obs. of  12 variables:
##  $ month      : Factor w/ 7 levels "APR","AUG","JUL",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ day        : int  10 11 12 13 14 15 23 24 25 27 ...
##  $ attend     : int  56000 29729 28328 31601 46549 38359 26376 44014 26345 44807 ...
##  $ day_of_week: Factor w/ 7 levels "Friday","Monday",..: 6 7 5 1 3 4 2 6 7 1 ...
##  $ opponent   : Factor w/ 17 levels "Angels","Astros",..: 13 13 13 11 11 11 3 3 3 10 ...
##  $ temp       : int  67 58 57 54 57 65 60 63 64 66 ...
##  $ skies      : Factor w/ 2 levels "Clear ","Cloudy": 1 2 2 2 2 1 2 2 2 1 ...
##  $ day_night  : Factor w/ 2 levels "Day","Night": 1 2 2 2 2 1 2 2 2 2 ...
##  $ cap        : Factor w/ 2 levels "NO","YES": 1 1 1 1 1 1 1 1 1 1 ...
##  $ shirt      : Factor w/ 2 levels "NO","YES": 1 1 1 1 1 1 1 1 1 1 ...
##  $ fireworks  : Factor w/ 2 levels "NO","YES": 1 1 1 2 1 1 1 1 1 2 ...
##  $ bobblehead : Factor w/ 2 levels "NO","YES": 1 1 1 1 1 1 1 1 1 1 ...
```

```
# Show summary statistics
summary(df)
```

```
##  month         day            attend         day_of_week      opponent
##  APR:12   Min.   : 1.00   Min.   :24312   Friday   :13   Giants   : 9
##  AUG:15   1st Qu.: 8.00   1st Qu.:34493   Monday   :12   Padres   : 9
##  JUL:12   Median :15.00   Median :40284   Saturday :13   Rockies  : 9
##  JUN: 9   Mean   :16.14   Mean   :41040   Sunday   :13   Snakes   : 9
##  MAY:18   3rd Qu.:25.00   3rd Qu.:46588   Thursday : 5   Cardinals: 7
```

```
##  OCT: 3   Max.   :31.00   Max.   :56000   Tuesday  :13   Brewers  : 4
##  SEP:12                                  Wednesday:12   (Other)  :34
##       temp          skies    day_night   cap     shirt    fireworks
##  Min.   :54.00   Clear :62   Day  :15   NO :79   NO :78   NO :67
##  1st Qu.:67.00   Cloudy:19   Night:66   YES: 2   YES: 3   YES:14
##  Median :73.00
##  Mean   :73.15
##  3rd Qu.:79.00
##  Max.   :95.00
##
##  bobblehead
##  NO :70
##  YES:11
##
##
##
##
##
```

3. Exploratory Data Analysis

3.1. Attendance by Day of Week

We want to draw a box plot of attendance grouped by the day of week.

First, we define an ordered day-of-week variable by recoding the day_of_week column.

```
# Define an ordered day-of-week variable for plots and data summaries
df$ordered_day_of_week[df$day_of_week == 'Monday'] <- 1
df$ordered_day_of_week[df$day_of_week == 'Tuesday'] <- 2
df$ordered_day_of_week[df$day_of_week == 'Wednesday'] <- 3
df$ordered_day_of_week[df$day_of_week == 'Thursday'] <- 4
df$ordered_day_of_week[df$day_of_week == 'Friday'] <- 5
df$ordered_day_of_week[df$day_of_week == 'Saturday'] <- 6
df$ordered_day_of_week[df$day_of_week == 'Sunday'] <- 7
```

Then, we transform the ordered day-of-week variable as factor.

```
df$ordered_day_of_week <- factor(df$ordered_day_of_week, levels=1:7,
labels=c("Mon", "Tue", "Wed", "Thur", "Fri", "Sat", "Sun"))
```

Showing the head of the updated data frame.

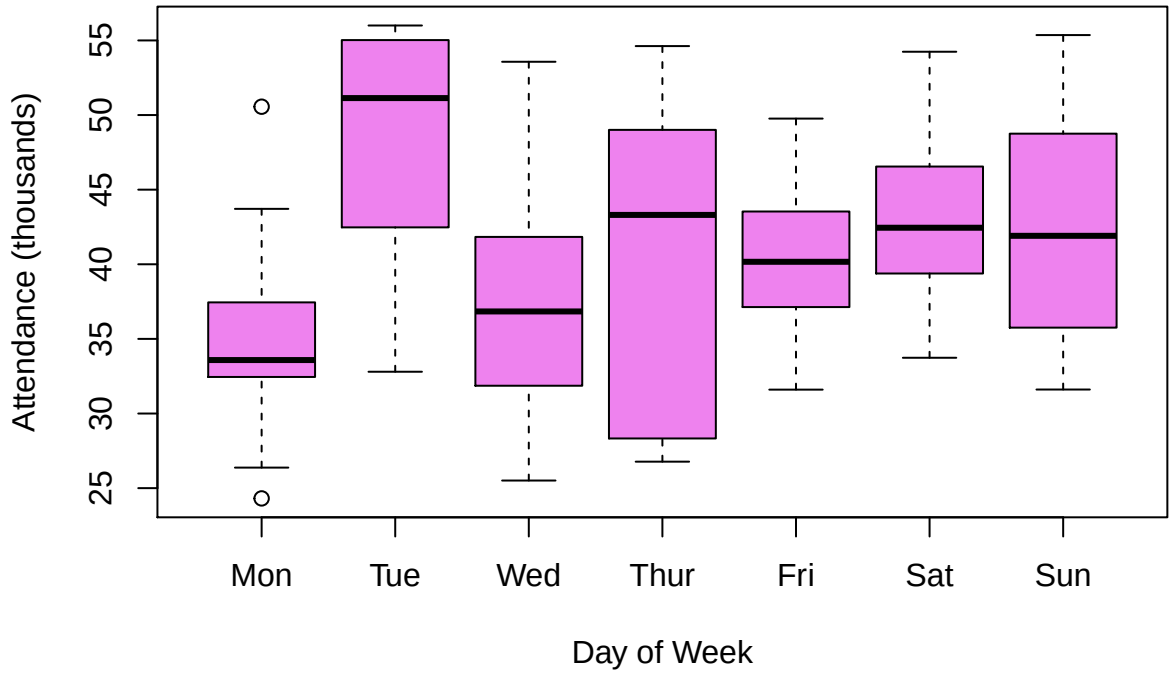
```
head(df)
```

```
##   month day attend day_of_week opponent temp  skies day_night cap shirt
## 1   APR  10  56000     Tuesday  Pirates   67 Clear        Day  NO    NO
## 2   APR  11  29729   Wednesday  Pirates   58 Cloudy     Night  NO    NO
## 3   APR  12  28328    Thursday  Pirates   57 Cloudy     Night  NO    NO
## 4   APR  13  31601      Friday   Padres   54 Cloudy     Night  NO    NO
## 5   APR  14  46549    Saturday   Padres   57 Cloudy     Night  NO    NO
## 6   APR  15  38359      Sunday   Padres   65 Clear        Day  NO    NO
##   fireworks bobblehead ordered_day_of_week
## 1        NO         NO                 Tue
## 2        NO         NO                 Wed
```

```
## 3         NO          NO                Thur
## 4        YES          NO                 Fri
## 5         NO          NO                 Sat
## 6         NO          NO                 Sun
```

Now, drawing the box plot.

```
# Box plot of attendance by day of week
plot(df$ordered_day_of_week, df$attend/1000,
     xlab = "Day of Week",
     ylab = "Attendance (thousands)",
     col = "violet")
```




```
# Frequency table of bobblehead promotions by day of week
table(df$bobblehead,df$ordered_day_of_week)
```

```
##
##       Mon Tue Wed Thur Fri Sat Sun
##   NO   12   7  12    3  13  11  12
##   YES   0   6   0    2   0   2   1
```

From the frequency table, we find that most bobblehead promotions occured on Tuesdays.

3.2. Attendance by Month

Similarly, we can visualize attendance by month.

First, we need an ordered month variable for plots and data summaries. Of course, we can use the similar

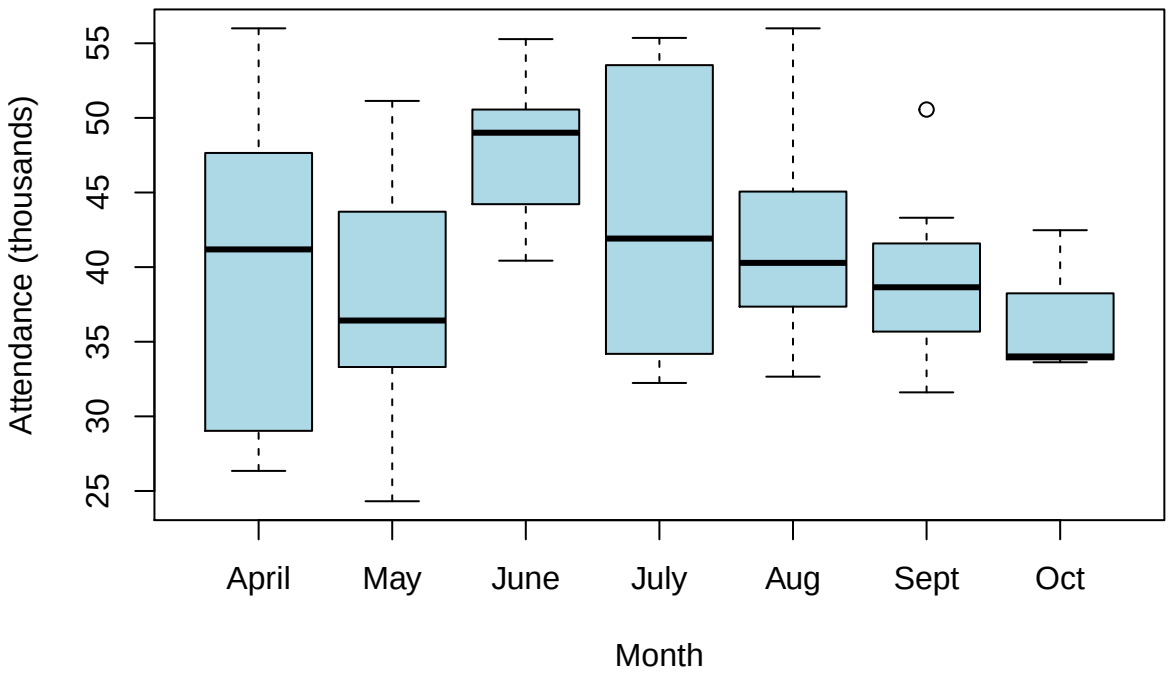
code in section 3.1. However, R provides many options to do the same task. Now, let's use the recode function in dplyr package to create the ordered month variable.

```
# Define an ordered month variable for plots and data summaries

df$ordered_month <- dplyr::recode(as.character(df$month),
                                  APR=4, MAY=5, JUN=6, JUL=7, AUG=8, SEP=9, OCT=10)


df$ordered_month <- factor(df$ordered_month, levels=4:10,
labels = c("April", "May", "June", "July", "Aug", "Sept", "Oct"))
```

```
head(df)
```

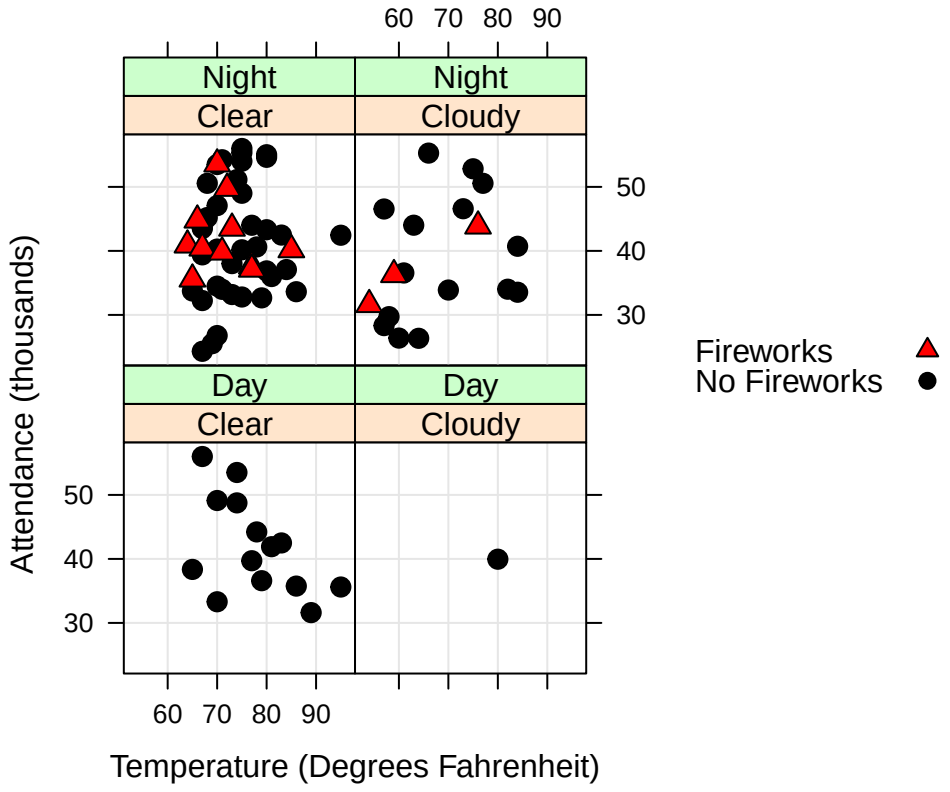
```
##   month day attend day_of_week opponent temp  skies day_night cap shirt
## 1   APR  10  56000     Tuesday  Pirates   67 Clear        Day  NO    NO
## 2   APR  11  29729   Wednesday  Pirates   58 Cloudy     Night  NO    NO
## 3   APR  12  28328    Thursday  Pirates   57 Cloudy     Night  NO    NO
## 4   APR  13  31601      Friday   Padres   54 Cloudy     Night  NO    NO
## 5   APR  14  46549    Saturday   Padres   57 Cloudy     Night  NO    NO
## 6   APR  15  38359      Sunday   Padres   65 Clear        Day  NO    NO
##   fireworks bobblehead ordered_day_of_week ordered_month
## 1        NO         NO                 Tue         April
## 2        NO         NO                 Wed         April
## 3        NO         NO                Thur         April
## 4       YES         NO                 Fri         April
## 5        NO         NO                 Sat         April
## 6        NO         NO                 Sun         April
```

```
# Box plot of attendance by by month
plot(df$ordered_month,df$attend/1000,
     xlab = "Month",
     ylab = "Attendance (thousands)",
     col = "light blue")
```



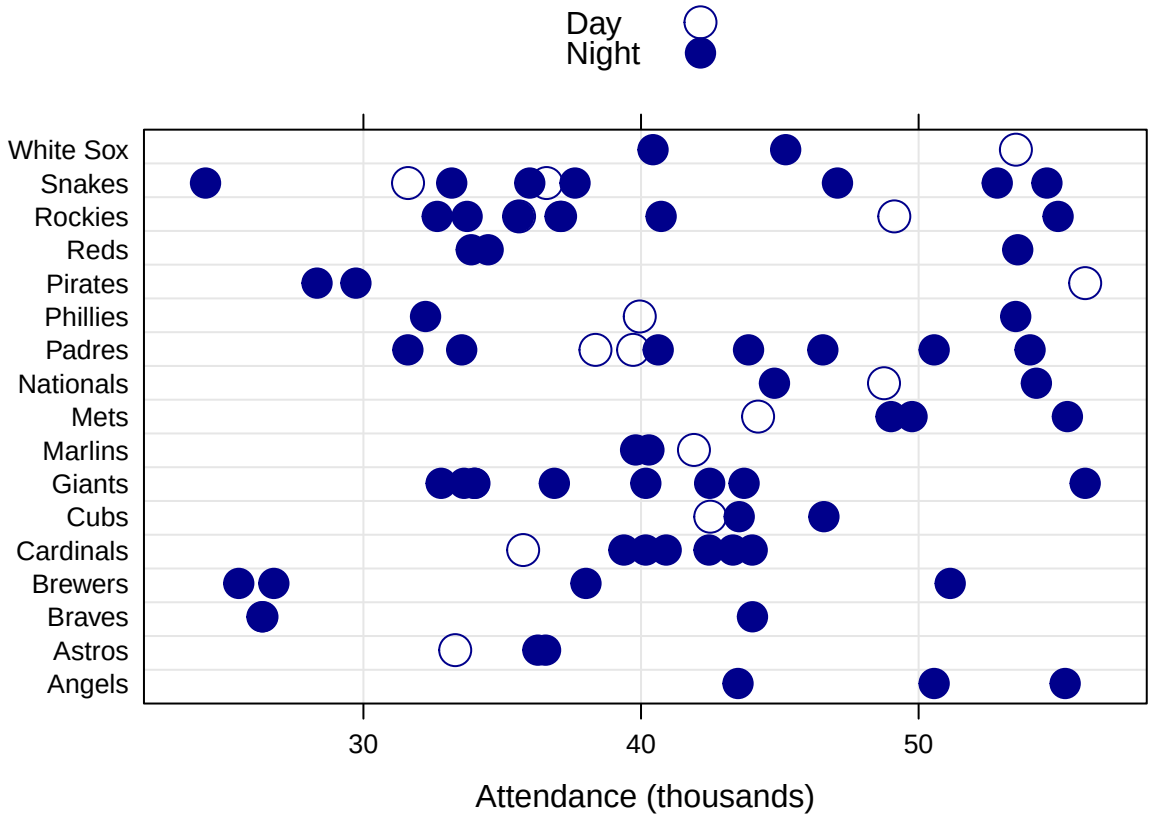
3.3. Plot the Relationship between Attendance and Weather

```
# Load the lattice library for plotting
library(lattice)

# Draw the scatter plot (lattice) of attendance vs. temperature conditioning on day/night and skies.
xyplot(attend/1000 ~ temp | skies + day_night,
       data = df, groups = fireworks,
       pch = c(21,24),aspect = 1, cex = 1.2,
       col = c("black","black"), fill = c("black","red"),
       layout = c(2, 2), type = c("p","g"),
       strip=strip.custom(strip.levels=TRUE,strip.names=FALSE, style=1),
       xlab = "Temperature (Degrees Fahrenheit)",
       ylab = "Attendance (thousands)",
       key = list(space = "right",
                  text = list(c("Fireworks","No Fireworks"),
                              col = c("black","black")),
                  points = list(pch = c(24,21),
                                col = c("black","black"),
                                fill = c("red","black"))))
```



3.4. Plot Attendance by Visiting Team

```
# Draw the plot of attendance vs. visiting team

bwplot(opponent ~ attend/1000, data = df, groups = day_night,
       xlab = "Attendance (thousands)",
       panel = function(x, y, groups, subscripts, ...){
         panel.grid(h = (length(levels(df$opponent)) - 1), v = -1)
         panel.stripplot(x, y, groups = groups,
                         subscripts = subscripts,
                         cex = c(2,2.75),
                         pch = c(1,20),
                         col = "darkblue")
         },
       key = list(space = "top",
                  text = list(c("Day","Night"),col = "black"),
                  points = list(pch = c(1,20),
                                cex = c(2,2.75),
                                col = "darkblue")))
```



4. Regression Analysis

4.1. Fit a Linear Regression Model

```
# Specify a simple model with bobblehead entered last
my.model <- {attend ~ bobblehead + ordered_month + ordered_day_of_week}

# Fit the linear regression model
my.model.fit <- lm(my.model, data = df)

# Show the summary of the fitted model
summary(my.model.fit)
```

```
##
## Call:
## lm(formula = my.model, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10786.5  -3628.1   -516.1   2230.2  14351.0
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              33909.16    2521.81  13.446  < 2e-16 ***
```

```
## bobbleheadYES           10714.90    2419.52   4.429 3.59e-05 ***
## ordered_monthMay        -2385.62    2291.22  -1.041  0.30152
## ordered_monthJune        7163.23    2732.72   2.621  0.01083 *
## ordered_monthJuly        2849.83    2578.60   1.105  0.27303
## ordered_monthAug         2377.92    2402.91   0.990  0.32593
## ordered_monthSept          29.03    2521.25   0.012  0.99085
## ordered_monthOct         -662.67    4046.45  -0.164  0.87041
## ordered_day_of_weekTue   7911.49    2702.21   2.928  0.00466 **
## ordered_day_of_weekWed   2460.02    2514.03   0.979  0.33134
## ordered_day_of_weekThur   775.36    3486.15   0.222  0.82467
## ordered_day_of_weekFri   4883.82    2504.65   1.950  0.05537 .
## ordered_day_of_weekSat   6372.06    2552.08   2.497  0.01500 *
## ordered_day_of_weekSun   6724.00    2506.72   2.682  0.00920 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6120 on 67 degrees of freedom
## Multiple R-squared:  0.5444, Adjusted R-squared:  0.456
## F-statistic: 6.158 on 13 and 67 DF,  p-value: 2.083e-07
```

Therefore, it appears that bobbleheads have a high and positive effecton attendance.

Using the stargazer package to explore this further:

```
# install.packages("stargazer") #Install stargazer package, do this only once
library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

```
stargazer(my.model.fit, type = "text", star.cutoffs = c(0.05, 0.01, 0.001),
          title="Multiple Linear Regression", digits=3)
```

```
##
## Multiple Linear Regression
## ================================================
##                            Dependent variable:
##                        ---------------------------
##                                  attend
## ------------------------------------------------
## bobbleheadYES                 10,714.900***
##                                (2,419.520)
##
## ordered_monthMay               -2,385.625
##                                (2,291.216)
##
## ordered_monthJune              7,163.234*
##                                (2,732.721)
##
## ordered_monthJuly               2,849.828
##                                (2,578.600)
##
## ordered_monthAug                2,377.924
```

```
##                                 (2,402.915)
##
## ordered_monthSept                 29.030
##                                 (2,521.249)
##
## ordered_monthOct                 -662.668
##                                 (4,046.452)
##
## ordered_day_of_weekTue          7,911.494**
##                                 (2,702.208)
##
## ordered_day_of_weekWed           2,460.023
##                                 (2,514.029)
##
## ordered_day_of_weekThur           775.364
##                                 (3,486.154)
##
## ordered_day_of_weekFri           4,883.818
##                                 (2,504.653)
##
## ordered_day_of_weekSat          6,372.056*
##                                 (2,552.084)
##
## ordered_day_of_weekSun          6,724.003**
##                                 (2,506.721)
##
## Constant                       33,909.160***
##                                 (2,521.806)
##
## ------------------------------------------------
## Observations                        81
## R2                                 0.544
## Adjusted R2                        0.456
## Residual Std. Error          6,120.158 (df = 67)
## F Statistic                6.158*** (df = 13; 67)
## ================================================
## Note:                  *p<0.05; **p<0.01; ***p<0.001
```

The above regression result already shows that bobblehead promotion has significant effect on attendance.

Another way to do the hypothesis is to do the anova test.

```
# tests statistical significance of the bobblehead promotion
# type I anova computes sums of squares for sequential tests
anova(my.model.fit)
```

```
## Analysis of Variance Table
##
## Response: attend
##                     Df     Sum Sq    Mean Sq F value    Pr(>F)
## bobblehead           1 1864995736 1864995736 49.7912 1.201e-09 ***
## ordered_month        6  557523389   92920565  2.4808   0.03157 *
## ordered_day_of_week  6  575839199   95973200  2.5623   0.02704 *
## Residuals           67 2509574563   37456337
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
cat("\n","Estimated Effect of Bobblehead Promotion on Attendance: ",
round(my.model.fit$coefficients[length(my.model.fit$coefficients)],
digits = 0),"\n",sep="")
```

```
##
## Estimated Effect of Bobblehead Promotion on Attendance: 6724
```

4.2. Regression Diagnostic

To assess the validity of the regression model, we do the following diagnostic procedures.

4.2.1. Linearity Check

Since all predictors are categorical variables (factors), we don't need to check linearity. Linearity check makes sense only when both independent and dependent variables are ratio data.

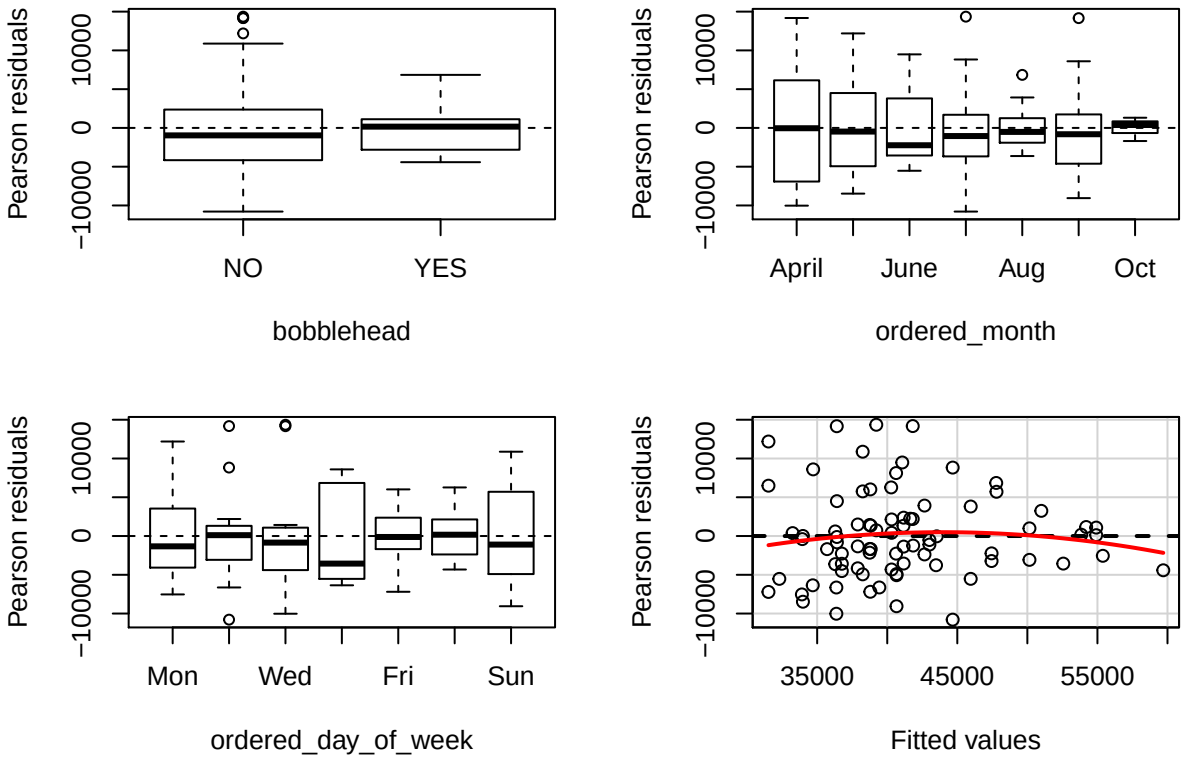
4.2.2. Homoscedasticity Check

```r
library(car)
# non-constant error variance test
ncvTest(my.model.fit)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 4.137439    Df = 1     p = 0.04194457
```

We can reject the null hypothesis that the errors have a non-constant variance ($p < 0.05$).

```r
# plot studentized residuals
residualPlots(my.model.fit)
```

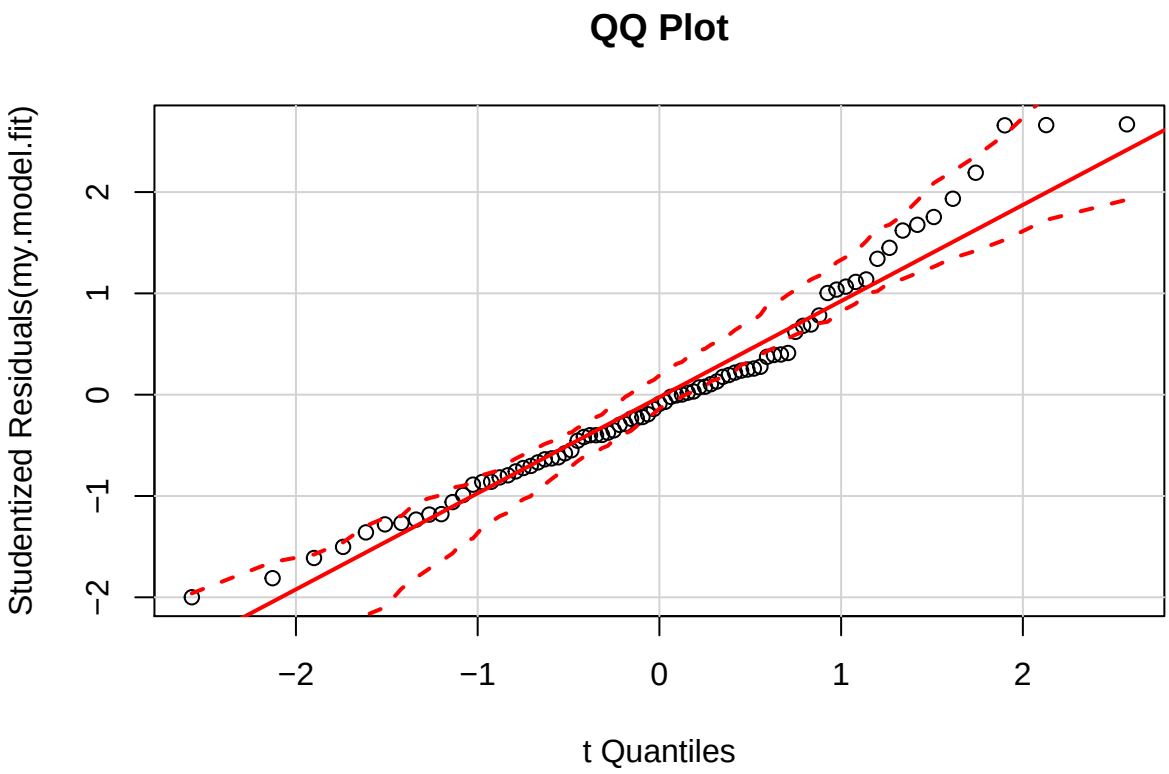



```
##                       Test stat Pr(>|t|)
## bobblehead                   NA       NA
## ordered_month                NA       NA
## ordered_day_of_week          NA       NA
## Tukey test               -1.123    0.261
```

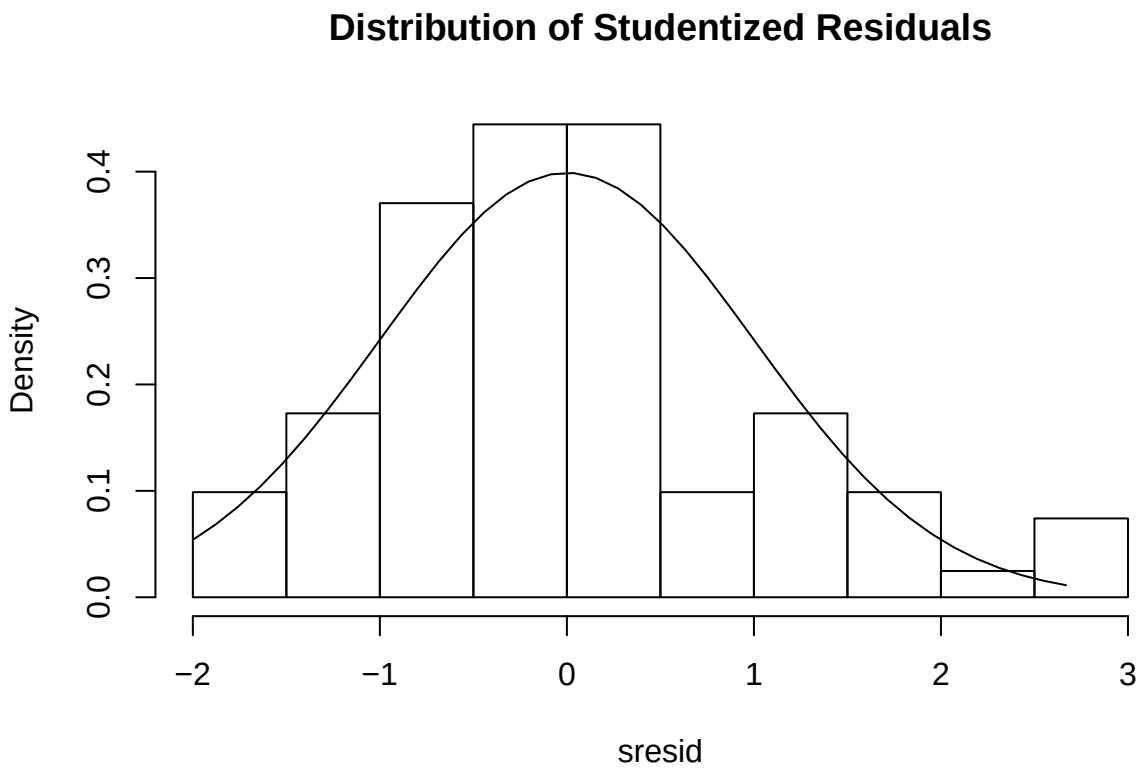
4.2.3. Normality Check

One of the assumptions of linear regression analysis is that the residuals are normally distributed. It is important to meet this assumption for the p-values for the t-tests to be valid.

```
# Normality of Residuals
# qq plot for studentized resid
qqPlot(my.model.fit, main="QQ Plot")
```




```
# distribution of studentized residuals
library(MASS)
sresid <- studres(my.model.fit)
hist(sresid, freq=FALSE,main="Distribution of Studentized Residuals")
xfit<-seq(min(sresid),max(sresid),length=40)
yfit<-dnorm(xfit)
lines(xfit, yfit)
```



Both the above Q-Q plot and histogram look normal. Based on these graphs, the residuals from this regression model appear to conform to the normality assumption.

4.2.4. Multi-collinearity Check

```
vif(my.model.fit) # variance inflation factors
```

```
##                          GVIF Df GVIF^(1/(2*Df))
## bobblehead           1.485727  1        1.218904
## ordered_month        1.280697  6        1.020831
## ordered_day_of_week 1.617649  6        1.040895
```

As a general rule of thumb: the smaller VIF the better. VIF > 5 would have serious multi-collinearity problem.