Thinx_linear_regression.R

jyothi

Sat May 20 06:41:12 2017

```
orders_export_1 <- read.csv('/Users/jyothi/Desktop/thinx/orders_export_1.csv', comment.char="~")
orders_export_2 <- read.csv('/Users/jyothi/Desktop/thinx/orders_export_2.csv', comment.char="~")
orders_export <- read.csv('/Users/jyothi/Desktop/thinx/orders_export.csv', comment.char="~")
#View(orders_export_2)
# Merging three datasets
mergedf <- rbind( orders_export, orders_export_1,orders_export_2 )
#View(mergedf)

#  Remove # sign before Name field
mergedf$Name <- substring(mergedf$Name, 2)
mergedf$Billing.Zip <- substring(mergedf$Billing.Zip, 2)
mergedf$Shipping.Zip <- substring(mergedf$Shipping.Zip, 2)

#mergedf[["Subtotal"]][is.na(mergedf[["Subtotal"]])] <- 0
# Taking necessary columns
subDf <- subset(mergedf, select=c("Name", "Created.at","Lineitem.name","Lineitem.price","Lineitem.quant
#View(subDf)
# Selecting only Hiphugger items
p1 <- 'Hiphugger'
df1 <- subset(subDf, grepl(p1,Lineitem.name ) )
#View(df1)
summary(df1)
```

```
##      Name                               Created.at
##  Length:57081       2015-12-14 11:15:27 -0500:   14
##  Class :character   2016-03-04 15:44:05 -0500:   13
##  Mode  :character   2016-01-12 10:22:11 -0500:   10
##                     2016-02-08 20:14:40 -0500:    8
##                     2016-02-25 17:06:11 -0500:    7
##                     2015-12-30 11:50:05 -0500:    7
##                     (Other)                  :57022
##                   Lineitem.name   Lineitem.price Lineitem.quantity
##  Hiphugger - M / Black :15130   Min.   :34     Min.   : 1.000
##  Hiphugger - S / Black :11311   1st Qu.:34     1st Qu.: 1.000
##  Hiphugger - L / Black : 9781   Median :34     Median : 1.000
##  Hiphugger - XL / Black: 4593   Mean   :34     Mean   : 1.448
##  Hiphugger - XS / Black: 3228   3rd Qu.:34     3rd Qu.: 2.000
##  Hiphugger - M / Beige : 3029   Max.   :34     Max.   :41.000
##  (Other)               :10009
##  Lineitem.discount Lineitem.fulfillment.status   Lineitem.sku
##  Min.   :  0.000   fulfilled:56942             TXHH0103:15130
##  1st Qu.:  0.000   pending  :  139             TXHH0102:11311
##  Median :  0.000                               TXHH0104: 9781
##  Mean   :  2.978                               TXHH0105: 4593
##  3rd Qu.:  3.400                               TXHH0101: 3228
##  Max.   :160.590                               TXHH0203: 3029
##  NA's   :1                                     (Other) :10009
```

```
# Converting Created.at from string date
df1$Created.date <- as.Date(df1$Created.at ,format= "%Y-%m-%d %H:%M:%S")
# There is one NA in discount.price
df1 <- na.omit(df1)
#View(df1)
# Finding Price after discount
attach(df1)
df1$PAD <- with(df1, (Lineitem.price  -(Lineitem.discount/Lineitem.quantity)))
df1$Order.price <- with(df1, (Lineitem.price*Lineitem.quantity)-Lineitem.discount)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(lubridate)
# Group by
price_df<- df1 %>% group_by(item.price=PAD) %>%
  summarize(quantity.sold=sum(Lineitem.quantity) )
#View(price_df)
# Normalizinf Data Frame.
scaled.dat <- scale(price_df)
sca <- as.data.frame(scaled.dat)
# Scaling Does not do much of difference in Regression
# Centering price to make Intercept significant
cq <- price_df$quantity.sold - mean(price_df$quantity.sol)
#cq
#
linear_model1 <- lm( price_df$item.price ~ cq)
summary(linear_model1)
```

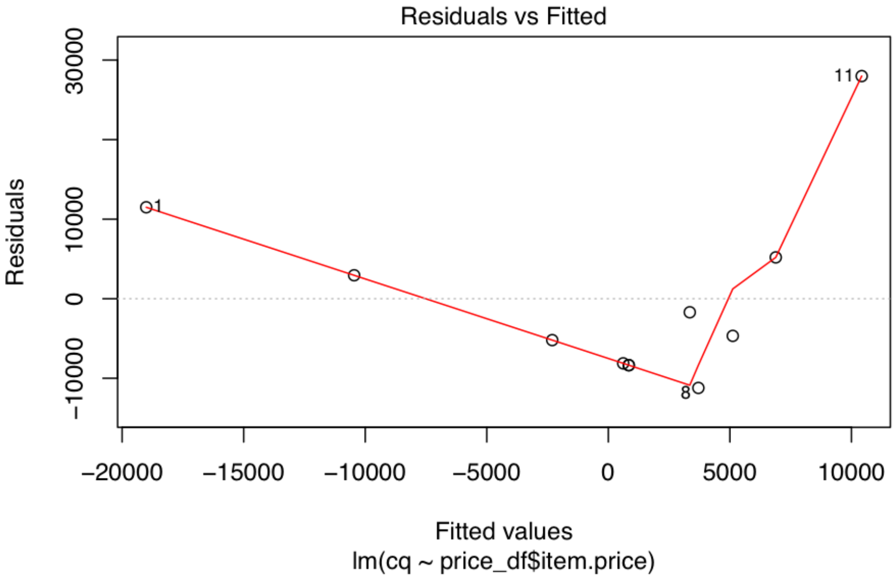
```
##
## Call:
## lm(formula = price_df$item.price ~ cq)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.826  -1.137   2.684   3.259   6.010
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.398e+01  2.056e+00  11.664  9.8e-07 ***
## cq          3.261e-04  1.517e-04   2.149   0.0601 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

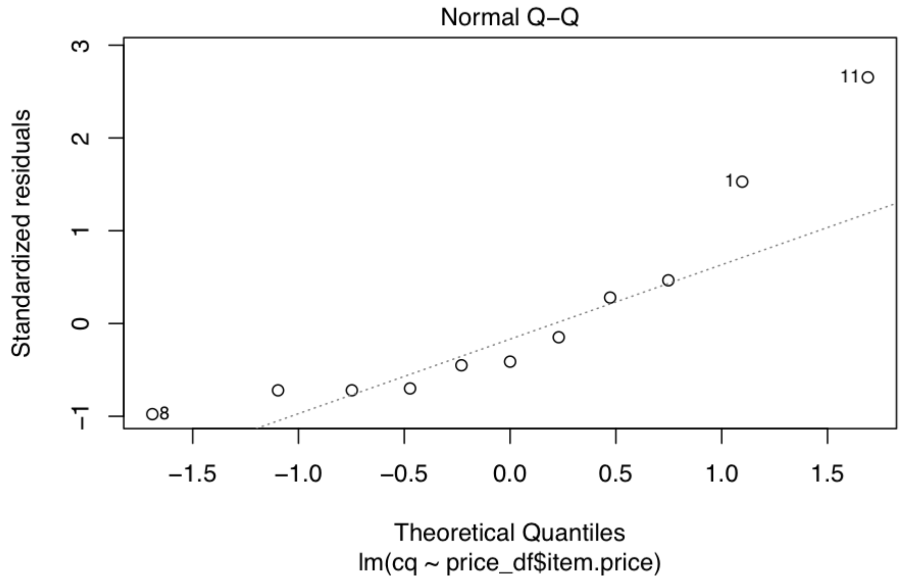
```
##
## Residual standard error: 6.818 on 9 degrees of freedom
## Multiple R-squared:  0.3392, Adjusted R-squared:  0.2658
## F-statistic:  4.62 on 1 and 9 DF,  p-value: 0.0601
```

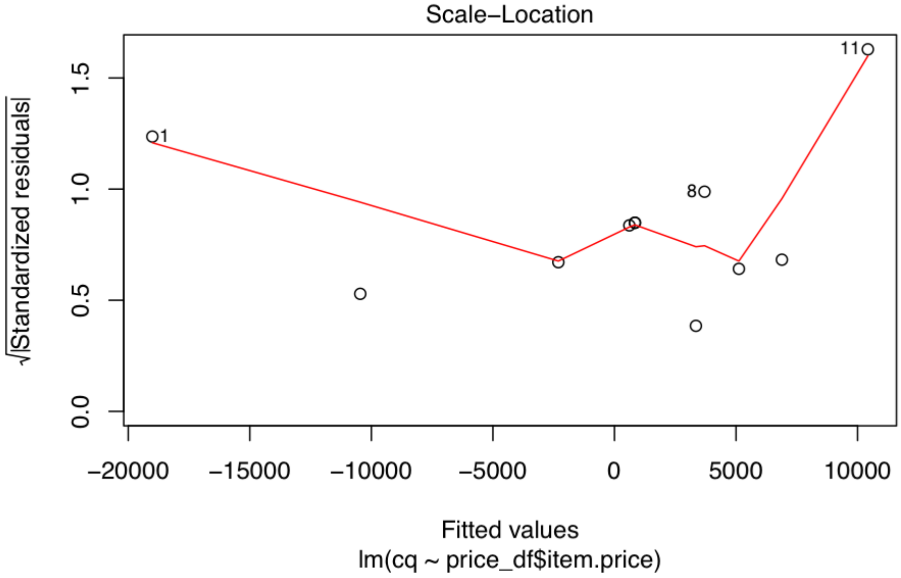
```
# We use this model becasue we are evaluating demand with price.
linear_model2 <- lm( cq  ~ price_df$item.price)
summary(linear_model2)
```

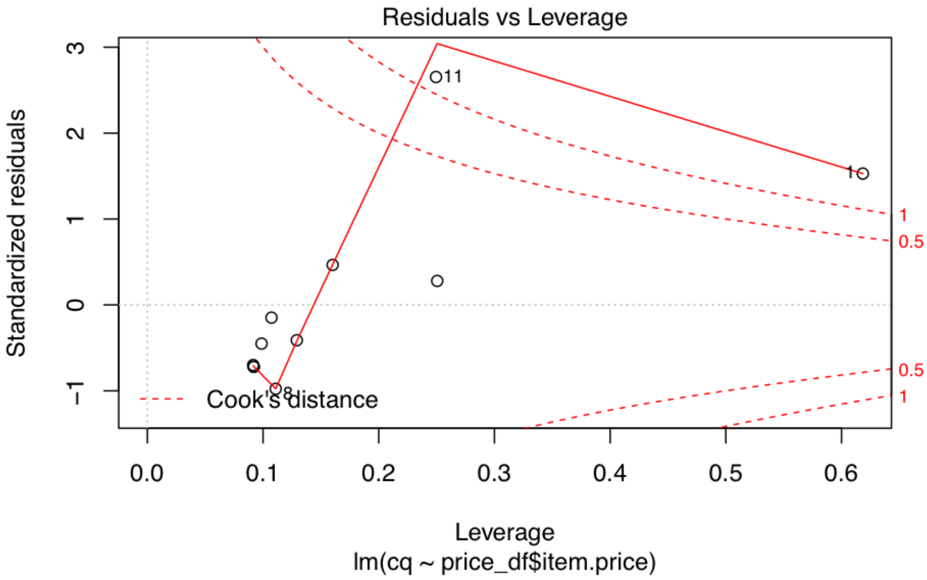
```
##
## Call:
## lm(formula = cq ~ price_df$item.price)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11214  -8238  -4671   4072  27988
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -24940.1    12170.6  -2.049   0.0707 .
## price_df$item.price   1040.1      483.9   2.149   0.0601 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12180 on 9 degrees of freedom
## Multiple R-squared:  0.3392, Adjusted R-squared:  0.2658
## F-statistic:  4.62 on 1 and 9 DF,  p-value: 0.0601
```

```
plot(linear_model2)
```




Normal Q−Q
Standardized residuals
−1
0
1
2
3
8
1
11
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles
lm(cq ~ price_df$item.price)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
1
8
11
−20000
−15000
−10000
−5000
0
5000
10000
Fitted values
lm(cq ~ price_df$item.price)



```
# R square is low. But rest of the curves seems ok. Though this model not that significant
# try to find th price elsticity
mean_price <- mean(price_df$item.price)
mean_quantity <- mean(price_df$quantity.sold)
mean_price
```

```
## [1] 23.97927
```

```
mean_quantity
```

```
## [1] 7516.273
```

```
# Price elasticity is delta q / delta p
PE <- 1040.1 *(mean_price/mean_quantity)
PE
```

```
## [1] 3.318245
```

```
# PE is high. and is postive that means unit variation(increase in price) increses demand.
# Obviously it will not be true. There might be some other factors that influence the sales not price
# for this period
```