1.

Let the response variable, the number of accidents, be “Y” and explanatory variables, the number of cars and their speed, be X1 and X2 respectively.

When running the model with interactions of X1(Cars) and X2(Speed), the R^2 and adjusted R^2 both increased a lot. After adding the square value of both X1 and X2, the R^2 square did increase slightly.

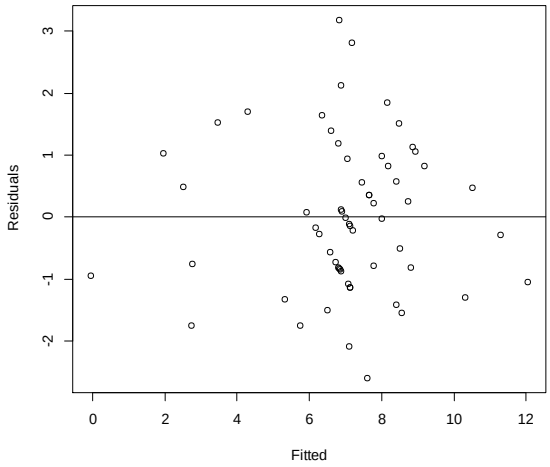
```
>  plot(residuals(carsmod2))

>   abline(h=0, lty=2)

> detach()

> carr<-read.table("car_accident.txt",header=T)

> x1_s<-carr$X1**2

> x2_s<-carr$X2**2

> c.2<-lm(Y~X1+X2+x1_s+x2_s+X1*X2,data=carr)

> summary(c.2)

Call:

lm(formula = Y ~ X1 + X2 + x1_s + x2_s + X1 * X2, data = carr)


Residuals:

   Min     1Q  Median     3Q    Max

-2.5969 -0.8547 -0.0616  0.8558  3.1828


Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept) 404.49882  327.00467   1.237    0.221

X1          -66.56803    6.53512 -10.186 3.54e-14 ***

X2           -2.34571   10.53501  -0.223    0.825

x1_s          0.10696    0.09681   1.105    0.274

x2_s         -0.06960    0.08528  -0.816    0.418

X1:X2         1.08154    0.09648  11.210 1.03e-15 ***

---
```

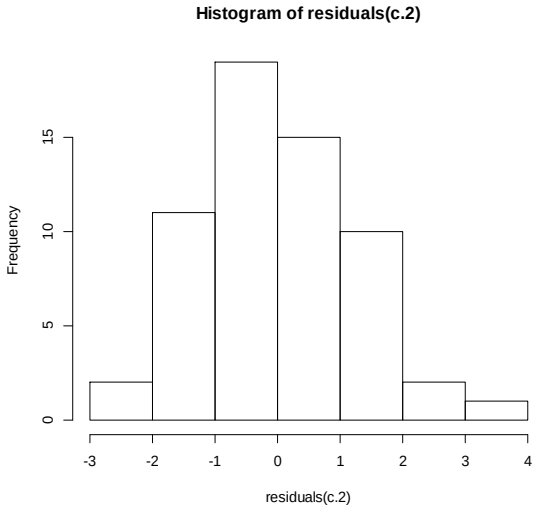
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

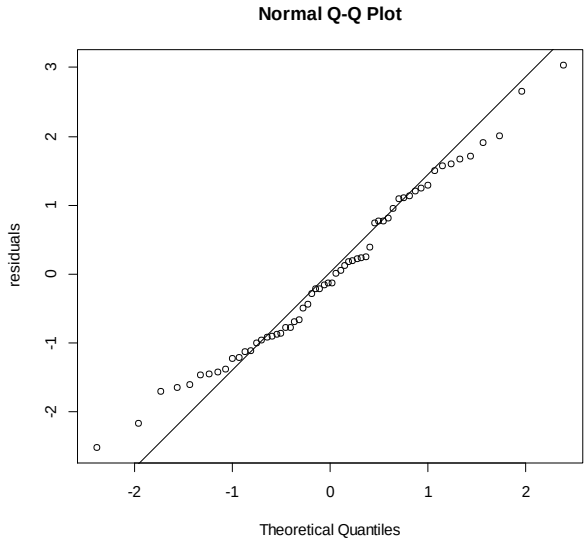
Residual standard error: 1.269 on 54 degrees of freedom

Multiple R-squared: 0.7514, Adjusted R-squared: 0.7284

F-statistic: 32.65 on 5 and 54 DF, p-value: 3.564e-15







> plot(fitted(c.2), residuals(c.2), xlab="Fitted", ylab="Residuals")

> abline(h=0)

> ## Diagnostic Plots - Histogram on Residuals

> hist(residuals(c.2))

The residual histogram and Q-Q plot satisfied the “3+1” model assumption about random errors.

From the summary of the model, the value of F-statistics is and can explain 75% of the variance in the number of accidents. It can be confirmed that the model is valid for the prediction.

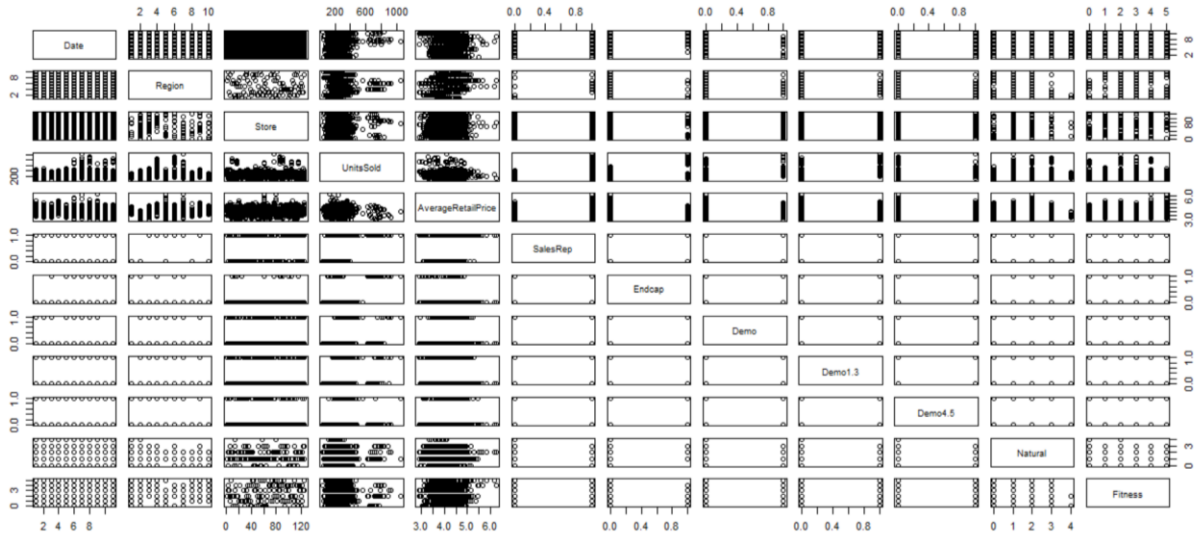
As the result, the regression model equation would be:

Y_hat=404.4988-66.5680 cars -2.3457 speed + 0.10696 cars**2 - 0.06960 speed**2 + 1.08154 cars*speed

(cars*speed stands for the interaction between two variables)

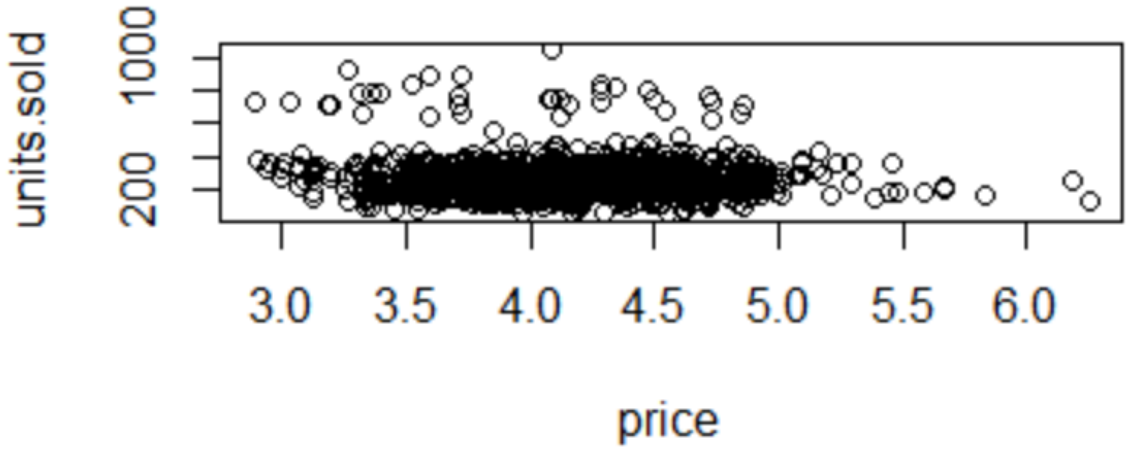
2.

(a)



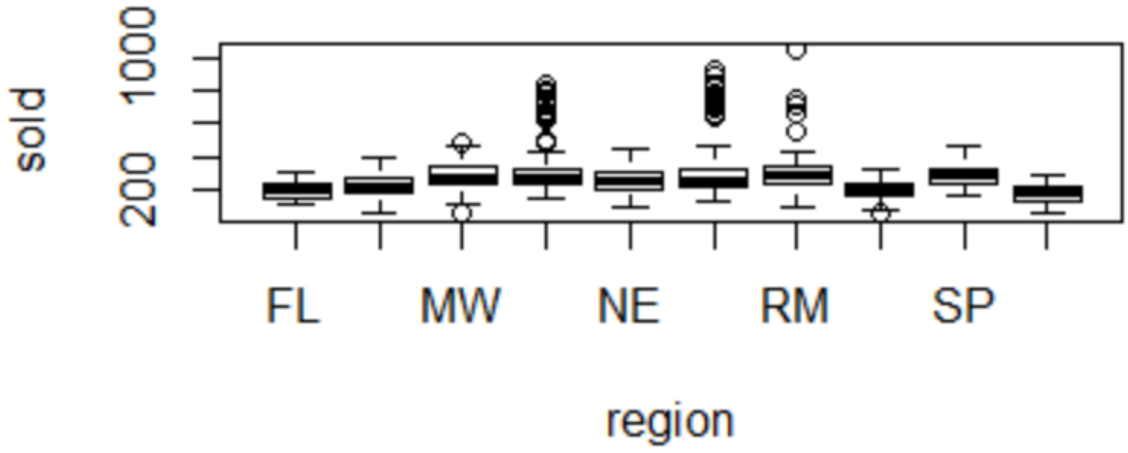
- plotting units sold against retail price showed that the majority of weekly observations were within the price range of $3-$5, so most stores produced approximately $150-$2250 of unit sales per week, but with units sold over 450 or average retail prices over $5, there was a notable

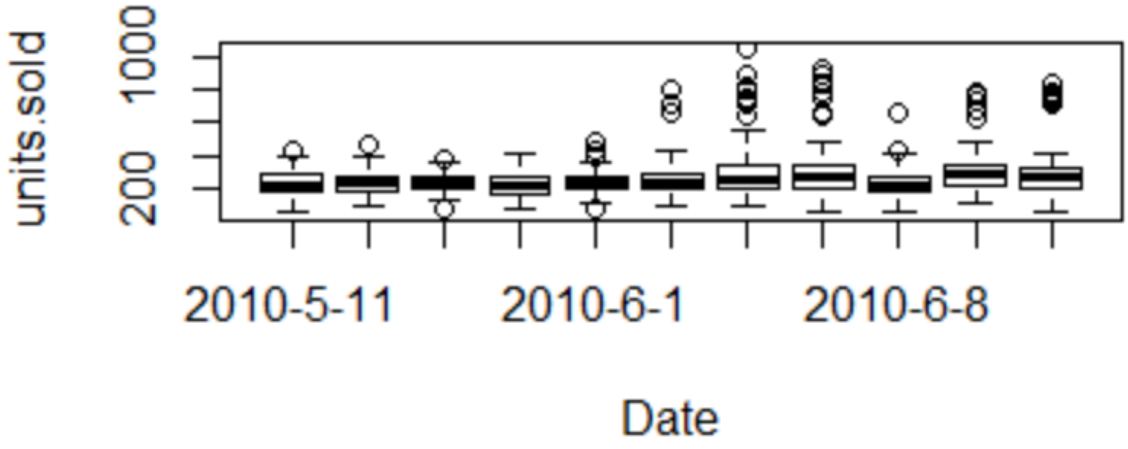
amount of weeks.From the chart below, we can also observe that there is a weak correlation between average retail prie and units sold.(which is good)



- Multicolinearity was not found to be a problem since the vif value were all smaller than 5(mentioned in [a]).

```
> vif(gb.2)
AverageRetailPrice          SalesRep            Endcap             Demo
          1.148871          1.293923          3.092386          1.030745
          Demo1.3           Demo4.5    SalesRep:Endcap
          1.079716          1.025249          3.167699
```





(b)

The average retail prices were normally distributed among weeks and stores (last two graph).

Multicolinearity was not found to be a problem since the vif value were all smaller than 5(mentioned in [a]).

Three promotional stratigies-Endcap, SalesRep, Demo are having positive impact on the unit sales.(will elaborate later on)

We found that it was negligible for both fitness centers and natural retailers, while the other variables were all extremely relevant. Concentrating on the demos and endcap based critical variables, we were able to infer from this data set that having demo activities contributed positively to unit sales in each of the observed weeks of this dataset. Specifically, holding a demo will boost unit sales in reasonable weeks. If a store had a demo 1-3 weeks or 4-5 weeks before, unit sales would grow by around 70 and 60 on average, respectively.

<r code for the full model>

```
g <- read.csv(file = "GoodBelly_data.csv", head = TRUE, sep=",")

Call:
  lm(formula = UnitsSold ~ AverageRetailPrice + SalesRep + Endcap +
      Natural + Demo + Demo1.3 + Demo4.5 + Fitness)

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
AverageRetailPrice  -28.535      3.952  -7.220 8.56e-13
SalesRep            77.437      3.864  20.038  < 2e-16
Endcap             305.102      9.056  33.692  < 2e-16
Natural             -1.594      1.776  -0.897    0.370
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
Fitness             -1.020      1.084  -0.941    0.347

(Intercept)       ***
  AverageRetailPrice ***
  SalesRep          ***
  Endcap            ***
  Natural
Demo              ***
  Demo1.3           ***
  Demo4.5           ***
```

Fitness

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

(c)

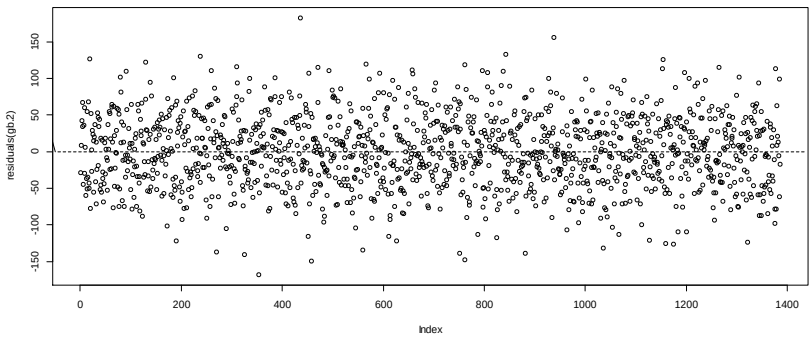
We preserve variable Endcap because its interaction with SalesRep is significant (to accord with preserving hierarchy) even though its p-value is not significant. The model explains over 80% of the variation in the residuals.

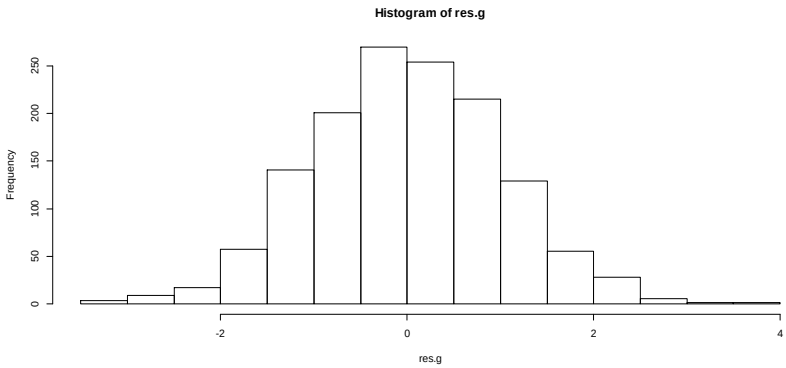
Final Model:

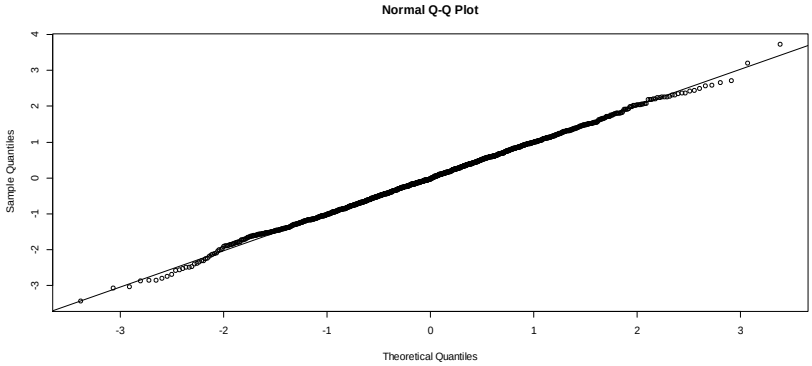
UnitsSold = 276.5735 - 22.0664AverageRetailPrice + **59.4618SalesRep + 0.6204Endcap +**

453.8033SalesRep*Endcap + 106.7527Demo + 73.3698Demo1.3 + 74.5520Demo4.5

The residual histogram and Q-Q plot satisfied the “3+1” model assumption about random errors. This supported the validity of the model in its current form







<code of different models>

> gb.0<-lm(UnitsSold~ .,data=b)

> summary(gb.0)

When having all the independent variables in the model, the value of R^2 would be 0.7546 and F-statistic: 26.52.

> gb.1<-
lm(UnitsSold~AverageRetailPrice+SalesRep+Endcap+SalesRep*Endcap+Demo+Demo1.3+Demo4.5+Natural+Fitness)

> summary(gb.1)

Residual standard error: 49.07 on 1376 degrees of freedom

Multiple R-squared: 0.8059,   Adjusted R-squared: 0.8046

F-statistic: 634.7 on 9 and 1376 DF, p-value: < 2.2e-16

> gb.3<-lm(UnitsSold~AverageRetailPrice+SalesRep+Endcap+SalesRep*Endcap+Demo+Demo1.3+Demo4.5)

> summary(gb.3)

Residual standard error: 49.03 on 1378 degrees of freedom

Multiple R-squared: 0.8059, Adjusted R-squared: 0.8049

F-statistic: 817.1 on 7 and 1378 DF, p-value: < 2.2e-16

(d)

Yes, it does. It might last for weeks, but we are not able to predict exact length of time. Furthermore, marginal increase from the promotions need more other variable to compare with. After all, our final model suggests that there is a positive and significant impact derived from in-store demonstration.

(e)

Yes, it does. Based on the model, we can conclude that there are positive and significant impacts from Endcap (while in conjunction with the SalesRep)indicating the placement do play a critical role here. However,keeping Endcap without Sales Rep almost does not contribute to sales, based on our model and plots.

(f)

variable salseRep and Endcap are not significant in the model without interactions after doing a few backward eliminations. To be more precise, when adopted independently, regional sales representative and endcap promotion are not successful, but will dramatically increase sales when used together in one store.Secondly, Sales are raised by the demos, and sales will not return to usual levels immediately afterwards.

My recommendation for GoodBelly is they should consider continuing their strategies on promotions and they should also have endcap promotions adopted supported by regional sales representatives based on the prediction of our final model.

(g)

I would suggest using backward stepwise AIC model test to remove variables sequentially.

At first, we have all the variables included in the original model. By removing variables from all the interactions terms, we can then come up with the refined ideal model with significant variables and interactions. Backward elimination findings show that the p-value of Natural and Fitness are insignificant and thus should be removed.