STAT 526 HW5

Group 1: Kyle Krezek, Yun-Hsien Kuo, Truman Kruse, Arnold Ukagwu, Keerthan Gajula

1a)

```
> fit.lm<-lm(Price~Rating,data=bfast)
>
> fit.lm

Call:
lm(formula = Price ~ Rating, data = bfast)

Coefficients:
(Intercept)       Rating
   -1042.90        11.96
```

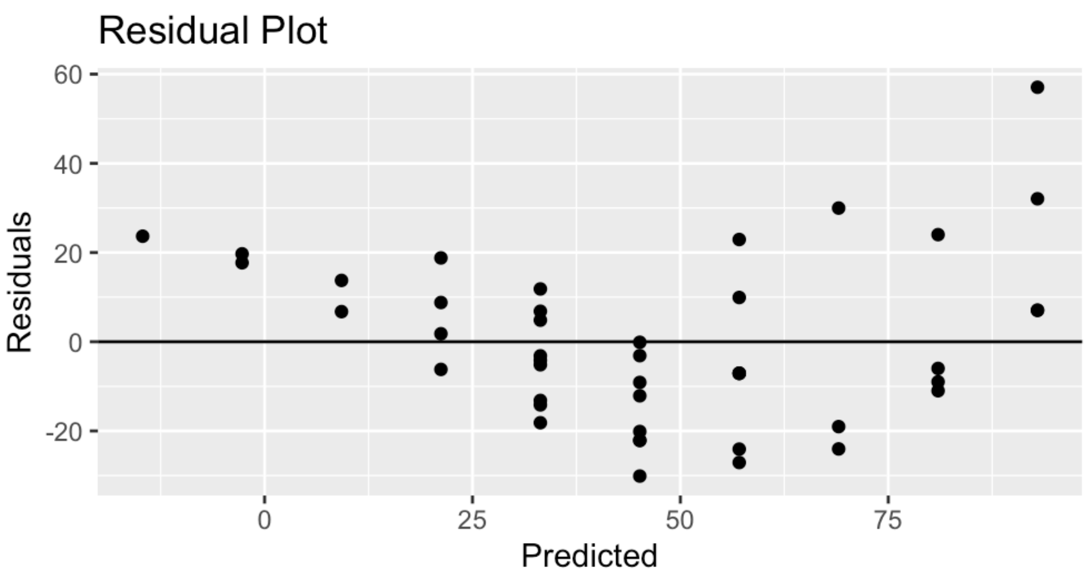
Price = -1042.9 + 11.96 * Rating

1b)

R squared is 0.674 and rmse is 18.204.

```
Residual standard error: 18.64 on 42 degrees of freedom
Multiple R-squared:  0.6816,    Adjusted R-squared:  0.674
F-statistic:  89.9 on 1 and 42 DF,  p-value: 5.366e-12
> rmse <- sqrt(mean((predictions - actual)^2))
> rmse
[1] 18.2084
```

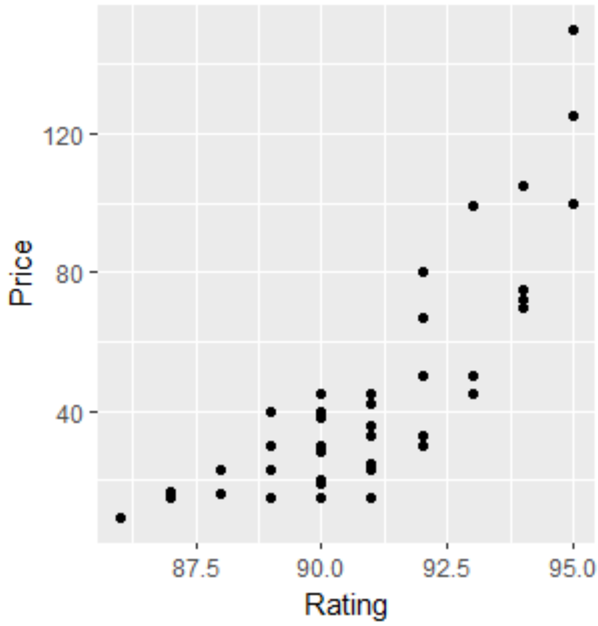
1c)



• The data points form a curved pattern, as we can see for predicted sales below 25, the residuals are almost positive. In between, the residuals are about negative.

• No constant variance

1d)



Form: concave upward curve

Direction: the points do not follow a clear linear pattern

Strength: it is positive, indicating that as one variable increases, the other tends to increase

Outlier: there are a few outliers in the upper-right corner of the plot, which may influence the overall pattern

1e)

As rating increases by a value of 1, price increases by 11.96.

1f)

```
> summary(mod2)

Call:
lm(formula = Price ~ Rating + I(Rating^2), data = bfast)

Residuals:
    Min      1Q  Median      3Q     Max 
-21.966 -11.948  -0.429   9.031  38.264 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13605.4242  3047.9185   4.464 6.17e-05 ***
Rating       -309.9989    66.9667  -4.629 3.67e-05 ***
I(Rating^2)     1.7680     0.3677   4.808 2.07e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.08 on 41 degrees of freedom
Multiple R-squared:  0.7964,    Adjusted R-squared:  0.7865 
F-statistic: 80.18 on 2 and 41 DF,  p-value: 6.768e-15
```

$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$

Price=β_0 +β_1 (Rating) +β_2 (Rating)^2

1g)

$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$

$\boxed{\text{Price}=13605.4242 - 309.9989\times\text{Rating} + 1.7680\times\text{Rating}^2}$

1h)

R^2 Adjusted for mod2 is 0.7865.

78.65% of the variation in the dependent variable(price) is explained by the independent variables(ratings) in our model(a substantial portion of the variance in the dependent variable). And it is generally considered a good fit.

1i)

```
> rmse <- sqrt(mean((predictions - actual)^2))
> rmse
[1] 14.56027
```

RMSE of mod2 is 14.56.

The lower the value of the RMSE, the better the model and its predictions.

1j)

```
> summary(mod1)

Call:
lm(formula = Price ~ Rating, data = bfast)

Residuals:
    Min      1Q  Median      3Q     Max 
-30.119 -12.380  -3.662  10.403  57.057 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1042.895    114.929  -9.074 1.88e-11 ***
Rating         11.956      1.261   9.482 5.37e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.64 on 42 degrees of freedom
Multiple R-squared:  0.6816,	Adjusted R-squared:  0.674 
F-statistic:  89.9 on 1 and 42 DF,  p-value: 5.366e-12
```

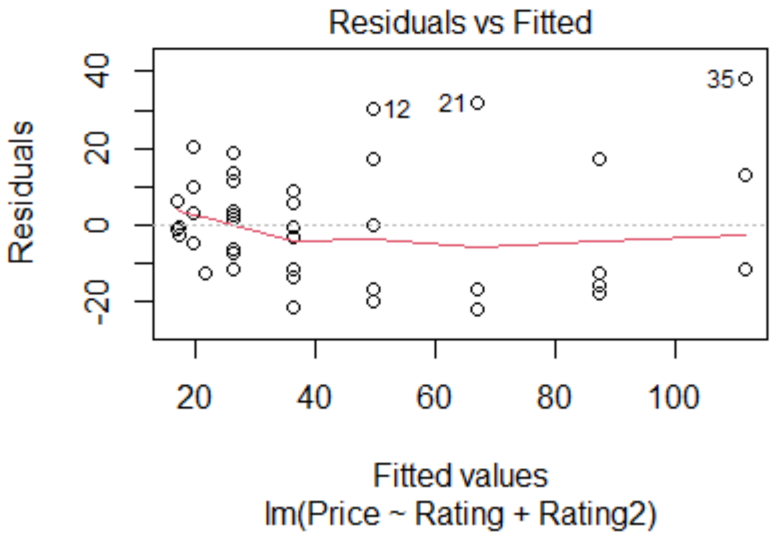
```
> summary(mod2)

Call:
lm(formula = Price ~ Rating + I(Rating^2), data = bfast)

Residuals:
    Min      1Q  Median      3Q     Max 
-21.966 -11.948  -0.429   9.031  38.264 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13605.4242  3047.9185   4.464 6.17e-05 ***
Rating       -309.9989    66.9667  -4.629 3.67e-05 ***
I(Rating^2)     1.7680     0.3677   4.808 2.07e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.08 on 41 degrees of freedom
Multiple R-squared:  0.7964,	Adjusted R-squared:  0.7865 
F-statistic: 80.18 on 2 and 41 DF,  p-value: 6.768e-15
```

Yes. Given that it has a lower RMSE and a higher R^2 it is more useful than the non quadratic model.

1k)

F = MSR/MSE, it also represents the ratio of the explained variability and the unexplained variability each divided by the corresponding degrees of freedom. Large F-ratio suggests that the explained variability is greater than the unexplained variability, implying that it's less likely that a difference is due to chance.

1l)

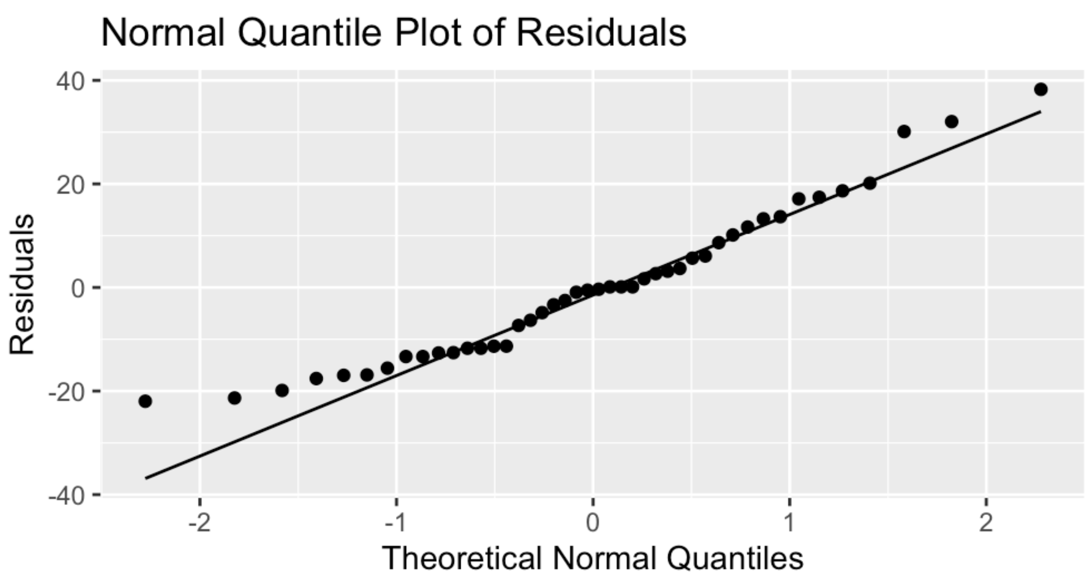


• Form of the model: There are equivalent positive and negative residuals along the range of predicted values. There is a quadratics pattern as well. The assumption is met that the model is useful. The assumptions of independence have not been violated.

• Constant variance assumption: The spread of residuals are not as consistent. More consistent towards starting points but decreases moving forward.

• Comment: The spread of the residuals is not uniform across the range of predicted values. It may be wider at one end of the range and narrower at the other end. Outliers or groups of points with different spread from the rest of the data, meaning the constant variance may be violated.

1m)



The plot is not unusual and does not indicate any non-normality with the residuals. There are some concerning values on the lower end but by and large it looks ok. We would say the model met the assumption of normality.

1n)

$H_0: \beta_2 = 0$

$H_a: \beta_2 \neq 0$

p-value<2e-16

Hypotheses:

Ho: B1=B2=0

Ha: B2≠0 for some i in (1,2)

T Statistic for Rating^2 = 4.808

p-value: 2.07e-05 <0.05

Value of the test statistic

The F-ratio is also very high, indicating a low error term and a good model performance.

p-value is <0.05 so we reject the null hypothesis.

Hence, there is a significant concave upward relationship between rating and price.

1o)

As $\beta_2 > 0$, it is concave upward relationship. There cannot be a concave downward relationship between rating and price as the coefficient in front of the quadratic term is positive. So, it doesn't make sense to complete a hypothesis test.

1p)

```
> predict(mod2,newdata = data.frame(Rating = 90))
       1
26.32993
> sum(mod2$coefficients*c(1,90,90^2))
[1] 26.32993
```

Calculation by hand:

$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$

$\boxed{\text{Price}=13605.4242 - 309.9989\times\text{Rating} + 1.7680\times\text{Rating}^2}$

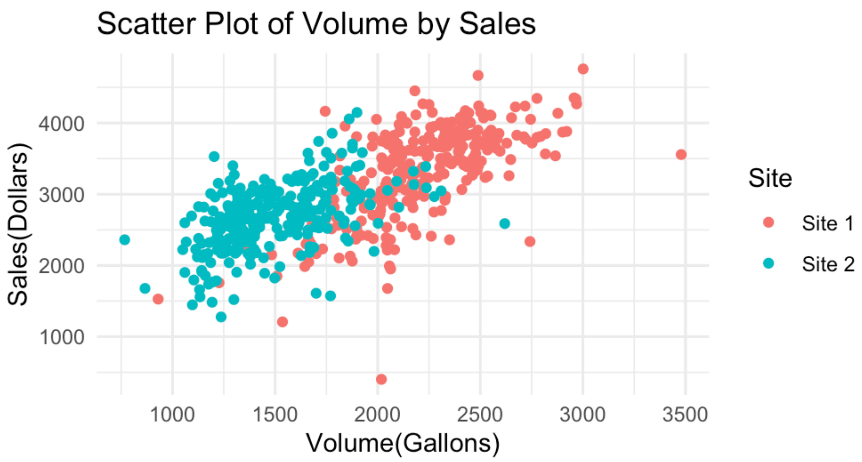
Price = 13605.4242 − 309.9989×90+ 1.7680×90^2

= 26.32

1q)

Price = $\beta_0 + \beta_1$ x Rating+ β_2 x Rating^2 + β_3 x WhiteWine x β_4 x RoseWine + ϵ

2a)



• a positive correlation exists between gallons of gas sold and dollar sales at both sites. As the volume of gas sold increases, so does the sales of the convenience store

• site 1 has higher sales for the same volume compared to site 2

2b)

Site 1: D=1

Site 2: D=0

$\boxed{\text{Average Sales} = \beta_0+\beta_1 * \text{ volumes} + \beta_2 * D + \epsilon}$

- β_0 is intercept
- β_1 = sales increase for each additional gallon of gas sold
- β_2 = difference in average sales between site 1 & site 2
- ϵ = unexplained variation in sales

Dummy variable is useful in a regression when data are classified into a small number of categories. If β_2 is positive, site 1 would have higher average sales compared to site 2(when no gas sold).

2c)

```
> model_convenience <- lm(SalesDollars ~ VolumeGallons + Site, data
= convenience_data)
> summary(model_convenience)
```

```
Call:
lm(formula = SalesDollars ~ VolumeGallons + Site, data = convenience
_data)
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1171.34056   61.89565   18.92   <2e-16 ***
VolumeGallons    0.31366    0.01803   17.39   <2e-16 ***
SiteSite 2    -520.42454   23.64755  -22.01   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Sales = 1171.3406 + 0.3137 * VolumeGallons – 520.4245(if site is site 2)

2d)

```
> model_convenience <- lm(SalesDollars ~ VolumeGallons + Site, data = convenience_dat
a)
> summary(model_convenience)

Call:
lm(formula = SalesDollars ~ VolumeGallons + Site, data = convenience_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-733.75 -164.79  -26.04  146.96 1191.96 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1171.34056   61.89565   18.92   <2e-16 ***
VolumeGallons    0.31366    0.01803   17.39   <2e-16 ***
SiteSite 2    -520.42454   23.64755  -22.01   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 243.6 on 565 degrees of freedom
Multiple R-squared:  0.7351,	Adjusted R-squared:  0.7342 
F-statistic:   784 on 2 and 565 DF,  p-value: < 2.2e-16
```

Yes, it is useful for the model utility.

H_0: $\beta_1 = \beta_2 = 0$

H_a: $\beta_i \neq 0$

F-statistic = 784 on 2 and 565 DF

p-value = < 2.2e – 16

Since our p-value is less than 0.05, we reject the null hypothesis that $\beta_1 = \beta_2 = 0$

2e)

p-value 2.2e – 16 is less than 0.05, we reject the null hypothesis

Sales in site2 are 520.42 dollors, which is less than site1

Yes, there is statistically significant evidence to suggest the average sales for site1 are different than the average sales for site2 after accounting for the volume of gas sold.

2f)

```
> summary(model_convenience2)

Call:
lm(formula = SalesDollars ~ VolumeGallons * Site, data = convenience_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-739.49 -164.81  -28.54  147.36 1190.43 

Coefficients:
                          Estimate Std. Error t value
(Intercept)             1148.21686   75.14508  15.280
VolumeGallons              0.32059    0.02209  14.510
SiteSite 2              -460.14764  113.42441  -4.057
VolumeGallons:SiteSite 2  -0.02080    0.03828  -0.543
                         Pr(>|t|)    
(Intercept)               < 2e-16 ***
VolumeGallons             < 2e-16 ***
SiteSite 2               5.67e-05 ***
VolumeGallons:SiteSite 2    0.587    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 243.7 on 564 degrees of freedom
Multiple R-squared:  0.7353,	Adjusted R-squared:  0.7339 
F-statistic: 522.1 on 3 and 564 DF,  p-value: < 2.2e-16
```

D = 1 if site is site 2

D = 0 if site is site 1

SalesDollars = $\beta_0 + \beta_1$ x VolumeGallons + β_2 x D + β_3 x VolumeGallons x D +ϵ

SalesDollors = 1148.21 + 0.32059(VolumeGallons) – 460.14764(Site) – 0.0208(VolumeGallons x site)

2g)

```
> summary(model_convenience2)

Call:
lm(formula = SalesDollars ~ VolumeGallons * Site, data = convenience_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-739.49 -164.81  -28.54  147.36 1190.43 

Coefficients:
                         Estimate Std. Error t value
(Intercept)            1148.21686   75.14508  15.280
VolumeGallons             0.32059    0.02209  14.510
SiteSite 2             -460.14764  113.42441  -4.057
VolumeGallons:SiteSite 2  -0.02080    0.03828  -0.543
                       Pr(>|t|)    
(Intercept)             < 2e-16 ***
VolumeGallons           < 2e-16 ***
SiteSite 2             5.67e-05 ***
VolumeGallons:SiteSite 2    0.587    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 243.7 on 564 degrees of freedom
Multiple R-squared:  0.7353,	Adjusted R-squared:  0.7339 
F-statistic: 522.1 on 3 and 564 DF,  p-value: < 2.2e-16
```

p-value of the VolumeGallons is 0.587, which is greater than 0.05, it accepts the null hypothesis. So there is no interaction between site and volume of gas when predicting sales.

2h)

Model1 is a better choice than model2 because it is simpler and easier to interpret, and the interaction term in model2 is not significant compared to model1.