STATS415_HW1_Xinye Xu_ (xinyexu) _ GSI:Eli

q2

(a) With all variables and no interactions in the mdoel, the R-squared: 0.8734, Adjusted R-squared: 0.8698, which indicate a good fit.But from the summary of all variables, there are several insignificant variables: Population, Education, Urban, US. Coefficients of these 4 variables are much closer to zero. Based on the residual diagnostic plot, it suggests a random pattern, which follows the assumption.

```
library('ISLR')
m_a <- lm(Sales ~ . , data= Carseats)
str(Carseats) # Factor: ShelveLoc, Urban, US
```

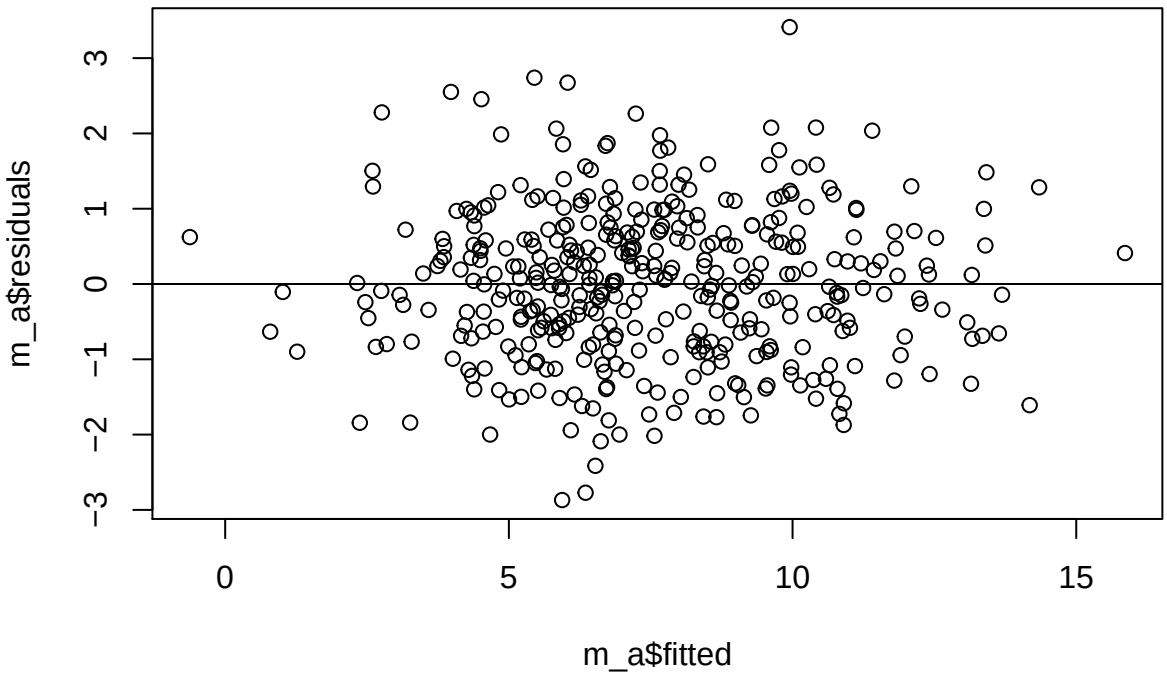
```
## 'data.frame':    400 obs. of  11 variables:
##  $ Sales      : num  9.5 11.22 10.06 7.4 4.15 ...
##  $ CompPrice  : num  138 111 113 117 141 124 115 136 132 132 ...
##  $ Income     : num  73 48 35 100 64 113 105 81 110 113 ...
##  $ Advertising: num  11 16 10 4 3 13 0 15 0 0 ...
##  $ Population : num  276 260 269 466 340 501 45 425 108 131 ...
##  $ Price      : num  120 83 80 97 128 72 108 120 124 124 ...
##  $ ShelveLoc  : Factor w/ 3 levels "Bad","Good","Medium": 1 2 3 3 1 1 3 2 3 3 ...
##  $ Age        : num  42 65 59 55 38 78 71 67 76 76 ...
##  $ Education  : num  17 10 12 14 13 16 15 10 10 17 ...
##  $ Urban      : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 2 2 1 1 ...
##  $ US         : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 1 2 1 2 ...
```

```
summary(m_a)
```

```
##
## Call:
## lm(formula = Sales ~ ., data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8692 -0.6908  0.0211  0.6636  3.4115
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.6606231  0.6034487   9.380  < 2e-16 ***
## CompPrice        0.0928153  0.0041477  22.378  < 2e-16 ***
## Income           0.0158028  0.0018451   8.565 2.58e-16 ***
## Advertising      0.1230951  0.0111237  11.066  < 2e-16 ***
## Population       0.0002079  0.0003705   0.561    0.575
## Price           -0.0953579  0.0026711 -35.700  < 2e-16 ***
## ShelveLocGood    4.8501827  0.1531100  31.678  < 2e-16 ***
## ShelveLocMedium  1.9567148  0.1261056  15.516  < 2e-16 ***
## Age             -0.0460452  0.0031817 -14.472  < 2e-16 ***
## Education       -0.0211018  0.0197205  -1.070    0.285
## UrbanYes         0.1228864  0.1129761   1.088    0.277
## USYes           -0.1840928  0.1498423  -1.229    0.220
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.019 on 388 degrees of freedom
```

```
## Multiple R-squared:  0.8734, Adjusted R-squared:  0.8698
## F-statistic: 243.4 on 11 and 388 DF,  p-value: < 2.2e-16
```

```
plot(m_a$residuals ~ m_a$fitted)
abline(h=0)
```



(b) Follow the summary information above, CompPrice, Income, Advertising, Price, ShelveLoc, Age have significant p-values,which are less than 0.05. For the dummy variable Urban (pvalue = 0.277), the baseline is UbbanNo while the null hypothesis is that coefficient of this dummy variable is zero, and alterntive is not zero. That also means null hypothesis is there is no influence of Urban to Sales, alternatie is there is influence of Urban to Sales. As the pvalue > 0.05, the null should not be rejected.

(c) Drop all the variables that are not significant in the full model. The new and no interactions model's Multiple R-squared: 0.872, Adjusted R-squared: 0.8697, compared with full model's R-squared: 0.8734, Adjusted R-squared: 0.8698. The new model has smaller R-squared and Adjusted R-squared, which seems to suggest a less fit.

```
m_c <- lm(Sales ~ CompPrice + Income + Advertising  + Price  + ShelveLoc  + Age , data= Carseats)
summary(m_c)
```

```
##
## Call:
## lm(formula = Sales ~ CompPrice + Income + Advertising + Price +
##     ShelveLoc + Age, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7728 -0.6954  0.0282  0.6732  3.3292
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.475226   0.505005   10.84   <2e-16 ***
## CompPrice       0.092571   0.004123   22.45   <2e-16 ***
## Income          0.015785   0.001838    8.59   <2e-16 ***
```

```
## Advertising      0.115903   0.007724   15.01   <2e-16 ***
## Price           -0.095319   0.002670  -35.70   <2e-16 ***
## ShelveLocGood    4.835675   0.152499   31.71   <2e-16 ***
## ShelveLocMedium  1.951993   0.125375   15.57   <2e-16 ***
## Age             -0.046128   0.003177  -14.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.019 on 392 degrees of freedom
## Multiple R-squared:  0.872,  Adjusted R-squared:  0.8697
## F-statistic: 381.4 on 7 and 392 DF,  p-value: < 2.2e-16
```

(d) In the anova test, the p-value of the test is 0.358. It suggests that the fitted model m_a is not significantly different from reduced model m_c at the level of 0.05. It is in consistant with the pretty closed R-squared values form above. So we should reject full model and stick with reduced model.

```
anova(m_a, m_c)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ CompPrice + Income + Advertising + Population + Price +
##     ShelveLoc + Age + Education + Urban + US
## Model 2: Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
##     Age
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    388 402.83
## 2    392 407.39 -4   -4.5533 1.0964  0.358
```

(e) Write out the reduced model in equation form and interpret the coefficients. Be careful with the coefficients of the categorical vari- able. AS ShelveLoc's coefficient is significant, From m_c: for ShelveLoc Bad level, Sales = 5.475226 + 0.092571CompPrice + 0.015785Income + 0.115903Advertising - 0.095319Price - 0.046128Age; for ShelveLoc Good level, Sales = 5.475226 + 0.092571CompPrice + 0.015785Income + 0.115903Advertising - 0.095319Price + 4.835675 - 0.046128Age; for ShelveLoc Medium level, Sales = 5.475226 + 0.092571CompPrice + 0.015785Income + 0.115903Advertising - 0.095319Price + 1.951993 - 0.046128Age;

Holding other variables constant, coefficient of CompPrice means one unit of price increase by competitor at each location will lead to 0.092571 increase of mean of Unit sales (in thousands) at each location. Coefficient of Income means one unit increase of community income level will lead to 0.015785 increase of mean of Unit sales at each location. Coefficient of Advertising means one unit increase of community income level will lead to 0.115903 increase of mean of Unit sales at each location. Coefficient of Price means one unit increase of Price company charges for car seats at each site will lead to 0.095319 decrease of mean of Unit sales at each location. Coefficient of Price means one unit increase of Average age of the local population will lead to 0.046128 decrease of mean of Unit sales at each location. Coefficient of ShelveLocGood means ShelveLoc with good quality will lead to 4.835675 increase of mean of Unit sales at each location compared with bad quality. Coefficient of ShelveLocGood means ShelveLoc with Medium quality will lead to 1.951993 increase of mean of Unit sales at each location compared with bad quality.

(f) Add an interaction term between the categorical variable ShelveLoc and the variable Price to the reduced model. The coefficients of Price:ShelveLocGood represents the difference of slop of Price between ShelveLocGood and ShelveLocBad. The coefficients of Price:ShelveLocMedium represents the difference of slop of Price between ShelveLocMedium and ShelveLocBad. The p-values of Price:ShelveLocGood and Price:ShelveLocMedium are 0.3730, 0.4984 respectively. This suggests that there is no significant influence of ShelveLoc on the slop of Price. So the the interaction term is not necessary.

```
m_f <- lm(Sales ~ CompPrice + Income + Advertising  + Price  + ShelveLoc  + Age + ShelveLoc:Price, data=
summary(m_f)
```

```
##
## Call:
## lm(formula = Sales ~ CompPrice + Income + Advertising + Price +
##     ShelveLoc + Age + ShelveLoc:Price, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7984 -0.6896  0.0144  0.6743  3.3391
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             5.866758   0.696460   8.424 7.08e-16 ***
## CompPrice               0.092592   0.004159  22.262  < 2e-16 ***
## Income                  0.015766   0.001849   8.528 3.32e-16 ***
## Advertising             0.116003   0.007746  14.975  < 2e-16 ***
## Price                  -0.098594   0.004677 -21.082  < 2e-16 ***
## ShelveLocGood           4.185088   0.747377   5.600 4.06e-08 ***
## ShelveLocMedium         1.535031   0.628915   2.441   0.0151 *
## Age                    -0.046494   0.003209 -14.490  < 2e-16 ***
## Price:ShelveLocGood     0.005619   0.006300   0.892   0.3730
## Price:ShelveLocMedium   0.003650   0.005386   0.678   0.4984
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.021 on 390 degrees of freedom
## Multiple R-squared:  0.8723, Adjusted R-squared:  0.8693
## F-statistic: 295.9 on 9 and 390 DF,  p-value: < 2.2e-16
```

(g) In the anova test, the p-value of the test is 0.6593 It suggests that the fitted model m_c is not significantly different from model m_f with interaction at the level of 0.05. So we should reject interaction model and stick reduced model with interaction term.

```
anova(m_c, m_f)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
##     Age
## Model 2: Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
##     Age + ShelveLoc:Price
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    392 407.39
## 2    390 406.52  2   0.86946 0.4171 0.6593
```