Stat415-homework2

Homework 2.1

1. Fit a multiple regression model to predict Sales using all other variables in the model. Report the values of coefficients, and how well the model fits (using R2). Include a plot of residuals and comment on any interesting features.

```
library(ISLR)
```

```
library(ISLR)
data(Carseats)
# Multiple regression model
model1=lm(Sales~CompPrice+Income+Advertising+Population+Price+ShelveLoc
+Age+Education+Urban+US,data=Carseats)
summary(model1)
```

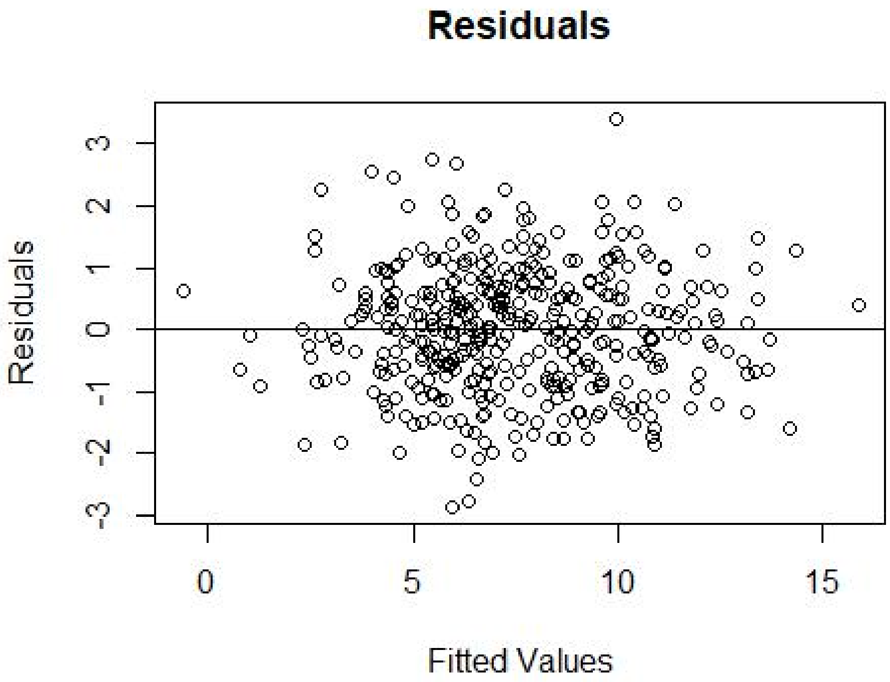
```
##
## Call:
## lm(formula = Sales ~ CompPrice + Income + Advertising + Population +

##     Price + ShelveLoc + Age + Education + Urban + US, data = Carseat
s)
##
## 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
```

Comment: The coefficients have been reported in the above Estimate terms.The Adjusted R-squared is 0.8698 and it is quite close to 1, which indicates that the model fits data quite well. The coefficients of predictors have been shown above in the Estimate term.

According to the result, Price, Age, Education and USYes have negative influence on the sales, while other predictors have positive influence. Besides, according to the p-value result, Population, Education, Urban and Us have quite large p-value, which indicates that these four predictors are not useful in this regression model.

```
plot(model1$residuals~model1$fitted.values,main="Residuals",xlab="Fitte
d Values",ylab="Residuals")
abline(h=0)
```



```
summary(model1$residuals)

##      Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -2.86924 -0.69085  0.02113  0.00000  0.66356  3.41146
```

Comment: According to the residuals result, the median and mean of residuals are 0.02113 and 0. Although the median of residuals is positive and there are more positive residuals in this regression model, residuals can still be considered as distributed randomly around zero. Thus there is no obvious nonlinear trend in residuals and we can conclude that linear regression model could fit the data well.

Homework 2.2

2. Which variables correspond to significant p-values? What is the null hypothesis the p-values are testing?

Answer: significant p-value should be smaller than a=0.05. Thus CompPrice, Income, Advertising, Price, ShelveLoc and Age should correspond to significant p-values. The null hypothesis is the single coefficient is zero. The p-values are testing whether their corresponding coefficients are zero or not. If p-values are quite small(<0.05), then we can reject the null hypothesis and conclude that the coefficients are significant.

Homework 2.3

3. Drop all the variables that are not significant in the previous model.Fit the linear model with the remaining variables. It will include one categorical variable, ShelveLoc. Compare the fit of the model to the previous one using R2.

```
model2=lm(Sales~CompPrice+Income+Advertising+Price+ShelveLoc+Age,data=C
arseats)
summary(model2)

##
## 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
```

Comment: the Adjusted R-squared is 0.8697. Compared with 0.8698 in the previous model, R^2 here is smaller but generally speaking, they are approximately the same. After dropping useless predictors, the current regression model becomes simpler but still keeps the regression effect.Thus these predictors should be dropped in this regression process.

Homework 2.4

4. Use the anova() to formally compare the two models and state your conclusion. Comment on the difference between their R2 in light of your conclusion.

```
anova(model1,model2)

## Analysis of Variance Table
##
## Model 1: Sales ~ CompPrice + Income + Advertising + Population + Pri
ce +
##     ShelveLoc + Age + Education + Urban + US
## Model 2: Sales ~ CompPrice + Income + Advertising + Price + ShelveLo
c +
##     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
```

Comment: the corresponding null hypothesis is all the coefficients of population, Education, Urban and US are equal to zero. Since the p-value for F-test is 0.358, which is larger than a=0.05. Thus we can not reject the null hypothesis and we can conclude that all these four coefficients should be zero and these two models are the same. Their corresponding Adjusted R^2 are 0.8698 and 0.8697, which also indicates that there is little difference between two models' regression effect.

Homework 2.5

5. Write out the model from the previous question in equation form and interpret the coefficients. Be careful with the coefficients of the categorical variable.

Answer: The model is:
Sales=5.475226+0.092571*CompPrice+0.015785*Income+0.115903*Advertising-0.095319*Price-0.046128*Age+4.835675*ShelveLocGood+1.951993*ShelveLocMedium

Interpret: The regression coefficients are interpreted as follows.

When other factors are held constant, one unit increase in CompPrice will increase the Sales by 0.092571 unit.

When other factors are held constant, one unit increase in Income will increase the Sales by 0.015785 unit.

When other factors are held constant, one unit increase in Advertising will increase the Sales by 0.115903 unit.

When other factors are held constant, one unit increase in Price will decrease the Sales by 0.095319 unit.

When other factors are held constant, one unit increase in Age will decrease the Sales by 0.046128 unit.

Interpret for Categorical variables:

The average difference in Sales between ShelveLocGood and ShelveLocBad with the same other variables is 4.835675.

The average difference in Sales between ShelveLocMedium and ShelveLocBad with the same other variables is 1.951993.

Homework 2.6

6. Add an interaction term between the categorical variable ShelveLoc and the variable Price. Refit the model, report the estimated coefficients, and interpret the coefficients of the interaction term. Do the p-values associated with them suggest the interaction term is necessary?

```
table(Carseats$ShelveLoc)

## 
##    Bad   Good Medium 
##     96     85    219

model3=lm(Sales~CompPrice+Income+Advertising+Price+ShelveLoc+Age+Price:
ShelveLoc,data=Carseats)
summary(model3)

## 
## Call:
## lm(formula = Sales ~ CompPrice + Income + Advertising + Price + 
##     ShelveLoc + Age + Price:ShelveLoc, 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
```

Comment: coefficients have been shown above in the Estimate terms. The regression coefficients are interpreted as follows.

0.005619 is the average difference in the relationship between Sales and Price between ShelveLocGood and ShelveLocBad when other variables keep the same.

0.003650 is the average difference in the relationship between Sales and Price between ShelveLocMedium and ShelveLocBad when other variables keep the same.

P-values of interactions are 0.3730 and 0.4984, which are larger than a=0.05. Thus those interaction terms are unnecessary.

Homework 2.7

7. Use the anova() to formally compare model from Q3 to the model from Q6 and state your conclusion.

```
anova(model2,model3)

## Analysis of Variance Table
##
## Model 1: Sales ~ CompPrice + Income + Advertising + Price + ShelveLo
c +
##     Age
## Model 2: Sales ~ CompPrice + Income + Advertising + Price + ShelveLo
c +
##     Age + Price:ShelveLoc
##   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
```

Comment: the corresponding null hypothesis here is all the coefficients of interactions are zero. Since the p-value for F-test is 0.6593, which is larger than a=0.05.Thus we can not reject the null hypothesis and we can conclude that these two models are the same and coefficients of interactions should be zero.