STATS415 hw2

Yunguo Cai

1/22/2018

```
library("ISLR")
```

```
help("Carseats")
data("Carseats")
```

1.

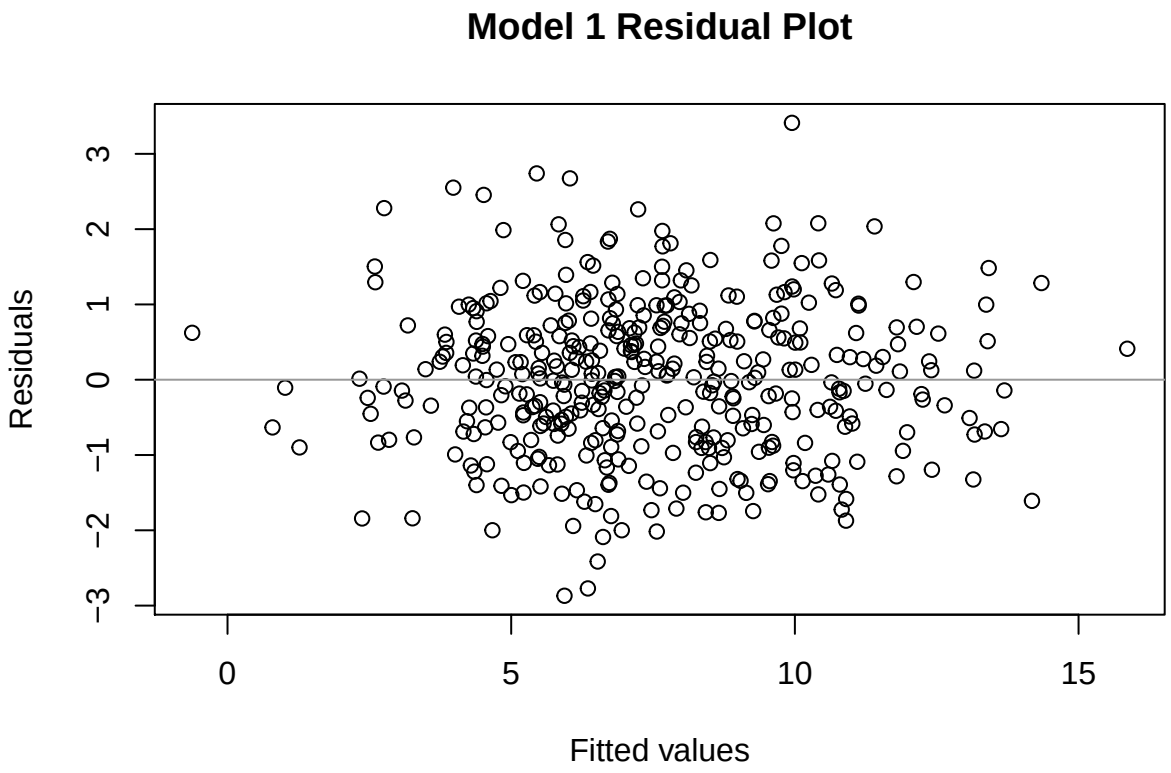
```
#Fit a multiple regression model to predict Sales using all other variables in the model
mod1 <- lm(Sales ~., data = Carseats)
summary(mod1)
```

```
##
## 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
```

The values of coefficients are shown above. The coefficients for Intercept, ComPrice, Income, Advertising, Population, Price, ShelveLovGood, ShelveLocMedium, Age, Education, UrbanYes and USYes are 5.6606, 0.0928, 0.0158, 0.1231, 0.0002,-0.0954, 4.8502, 1.9567, -0.046, -0.0211, 0.1229, -0.1841 respectively. The R^2 of this model is 0.8734, which means that the model explains 87.34% of the total variation. 87.34% is a large proportion, so the model fits well according to R^2.

```
#plot of residuals
plot(mod1$residuals ~ mod1$fitted.values, main="Model 1 Residual Plot",
xlab = "Fitted values", ylab = "Residuals")
```

```r
abline(a = 0, b = 0, col = "gray60")
```



The plot of residuals shows that the variance of the residuals doesn't have an obvious trend to vary a lot, so we could assume that the variance of error terms remains constant. It seems that when $\hat{y}$ is smaller, the residuals are slightly larger and there exists an outlier of residuals over 3 when $\hat{y} = 10$. The residuals of -2 to -3 when $\hat{y}$ is about 6-7 might also be outliers since they are scattered outside a bit.

2.

The variables Comprice, Income, Advertising, Price, ShelveLocGood, ShelveLocMedium, Age correspond to significant p-values at the significance level of 0.001. The null hypothesis that the p-values are testing is that the coefficients of these corresponding variables equal to 0.

3.

```r
#Fit the linear model with remaining significant variables
mod2 <- lm(Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc + Age, data = Carseats)
summary(mod2)
```

```
##
## 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
```

Using R^2, the R^2 of this model is 0.872, slightly smaller than 0.8734 of the previous model. Dropping off all the variables that are not significant in the previous model doesn't make much difference in explaining the total variance. This model also fits well.

4.

```
#Use anova() to formally compare the two models above
anova(mod2, mod1)
```

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

From ANOVA, the F-test shows that the F-value is 1.0964, quite low, which means that the group means are close together (low variability) relative to the variability within each group. The corresponding p-value is 0.358, much larger than 0.05. So we can't reject the null hypothesis that the second model which drops off insignificant variables is sufficient. There's not much difference between their R^2, so the first model can't be proved to fit significantly better than the second model. This agrees to the conclusion in question 3.

5.

Mod1 is:
Sales=5.661+0.093×CompPrice+0.016×Income+0.123×Advertising+0.0002×Population+0.095×Price +4.85×ShelveLocGood+1.957×ShelveLocMedium+0.046×Age+0.021×Education+0.123×UrbanYes-0.184×USYes

Mod2 is:
Sales=5.475+0.093×CompPrice+0.016×Income + 0.116×Advertising+0.095×Price+4.836×ShelveLocGood +1.952×ShelveLocMedium+0.046×Age

Unit sales (in thousands) at each location: The coefficient for CompPrice is 0.092, which means that when all the other variables remain the same, the price charged by competitor is increased by 1 unit, then the sales will be increased by 0.092 unit. The coefficient for Income is 0.016, which means that when all the other variables remain the same, the community income level is increased by 1 thousand dollars, then the sales will be increased by 0.016 unit. The coefficient for Advertising is 0.116, which means that when all the other variables remain the same, the local advertising budget for company is increased by 1 thousand dollars, then the sales will be increased by 0.116 unit. The coefficient for Price is -0.095, which means that when all the other variables remain the same, the price company charges for car seats is increased by 1 unit, then the sales will be decreased by 0.095 unit. The coefficient for ShelveLocGood is 4.836, which means that when all the

other variables remain the same, if the quality of the shelving location for the car seats is good, then the sales will be increased by 4.836 unit. The coefficient for ShelveLocMedium is 1.952, which means that when all the other variables remain the same, if the quality of the shelving location for the car seats is medium, then the sales will be increased by 1.952 unit. The coefficient for Age is 0.046, which means that when all the other variables remain the same, the average age of the local population is increased by 1 year, then the sales will be increased by 0.046 unit.

6.

```
#Add the interaction term between the categorical variable ShelveLoc and Price
mod3 <- lm(Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc + Age + ShelveLoc*Price, data =
summary(mod3)
```

```
##
## 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
```

The values of coefficients are shown above. The coefficients for Intercept, ComPrice, Income, Advertising, Population, Price, ShelveLovGood, ShelveLocMedium, Age, Price:ShelveLocGood, Price:ShelveLocMedium are 5.8668, 0.0926, 0.0158, 0.1160, -0.0986, 4.1850, 1.5350, -0.0465, 0.0056, 0.0037 respectively. The coefficient for Price:ShelveLocGood is 0.0056, which means that when all the other variables remain the same, if the quality of the shelving location for the car seats is good, then the coefficient for Price will be increased by 0.0056. The p-value associated with this interaction term is 0.3730, much bigger than 0.05, which suggests that it is not significant. The coefficient for Price:ShelveLocMedium is 0.004, which means that when all the other variables remain the same, if the quality of the shelving location for the car seats is medium, then the coefficient for Price will be increased by 0.004. The p-value associated with this interaction term is 0.4984, much bigger than 0.05, which suggests that it is not significant.

7.

```
#Compare model from Q3 to the model from Q6
anova(mod2,mod3)
```

```
## 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
```

From ANOVA, the F-test shows that the F-value is 0.4171, quite low, which means that the group means are close together (low variability) relative to the variability within each group. The corresponding p-value is 0.6593, much larger than 0.05. So we can't reject the null hypothesis that the second model which drops off the interaction term is sufficient. There's not much difference between their R^2, so the second model can't be proved to fit significantly better than the third model which adds the interaction term. The model without the interaction term fits well enough, so we can choose to drop off the interaction term.