Multiple Linear Regression: Building and Altering a Model

Andrew Levine

November 2021

```
headphonesB <- read.csv(file="/Users/andrewlevine/Downloads/Statistics II/Headphones_VersionB.csv")
str(headphonesB)
```

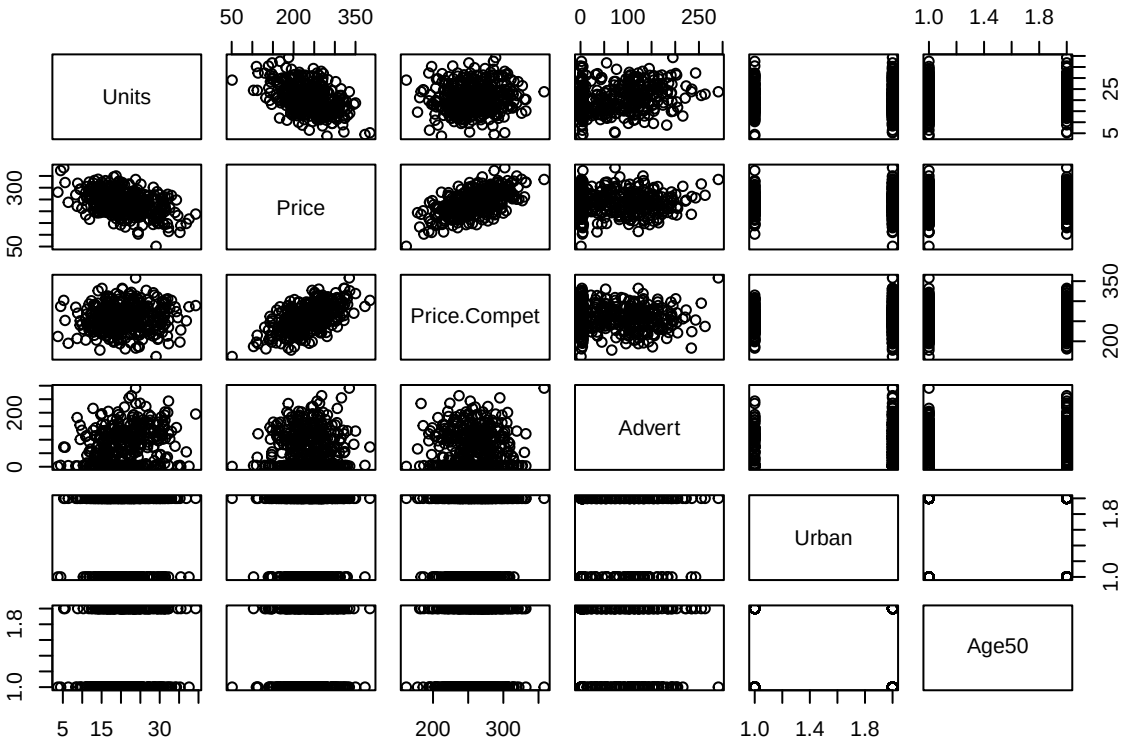
```
## 'data.frame':    400 obs. of  6 variables:
##  $ Units       : num  26 29.7 27.2 22.6 13.5 ...
##  $ Price       : int  245 169 165 198 260 148 220 245 252 251 ...
##  $ Price.Compet: int  284 229 235 243 291 256 238 281 271 273 ...
##  $ Advert      : int  113 165 106 44 35 135 5 155 3 4 ...
##  $ Urban       : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ Age50       : chr  "Yes" "No" "No" "No" ...
```

```
headphonesB$Urban <- factor(headphonesB$Urban)
headphonesB$Age50 <- factor(headphonesB$Age50)
```

Response Variable: Number of headphones sold last year (Units); Quantitative variable.
Explanatory Variables: Price of the headphones (Price), price that the nearest competitor charges for the headphones (Price.Compet), amount of money each store spent on advertising last year (Advert), whether or not the store is located in an urban area (Urban), and whether or not the average age of the people living in the area is under 50 years (Age50). Price, Price.Compet, and Advert are all quantitative variables, while Urban and Age50 are categorical variables.

Scatterplot Matrix of Variables:

```
plot(headphonesB)
```



Judging by the scatterplot matrix of the variables, Units appears to be linearly related to Price, but not to the other explanatory variables. In addition, the explanatory variable of Price seems to have a linear relationship with the other explanatory variable of Price.Compet. No other pairs of explanatory variables seem to have a linear relationship with one another. Some of the plots also contain potential outliers, especially in each of the scatterplots containing Price.Compet as one of the variables. In addition, there seems to exist potential outliers in the Units vs. Price scatterplot.

Variance Inflation Factors:

```
library(car)
```

```
## Loading required package: carData
```

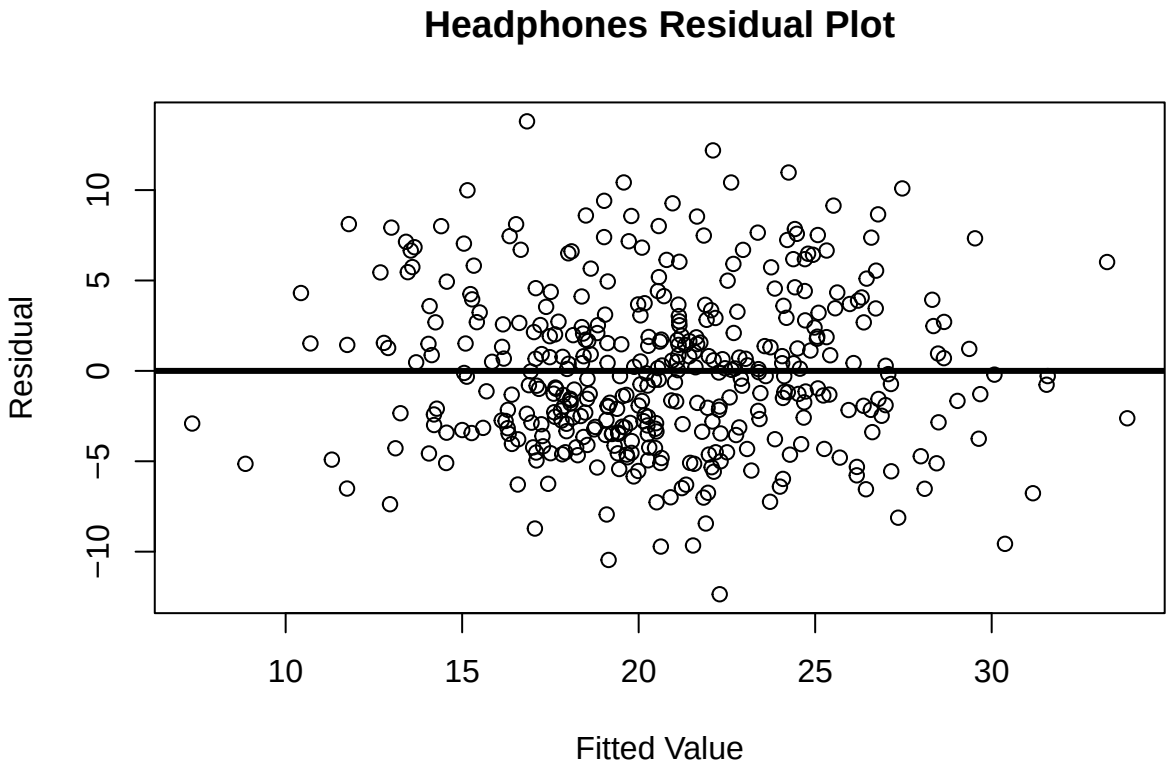
```
lm_info_headphonesB <- lm(Units ~ ., data = headphonesB)
vif(lm_info_headphonesB)
```

```
##         Price Price.Compet       Advert        Urban        Age50
##      1.524173     1.534399     1.009008     1.010604     1.021179
```

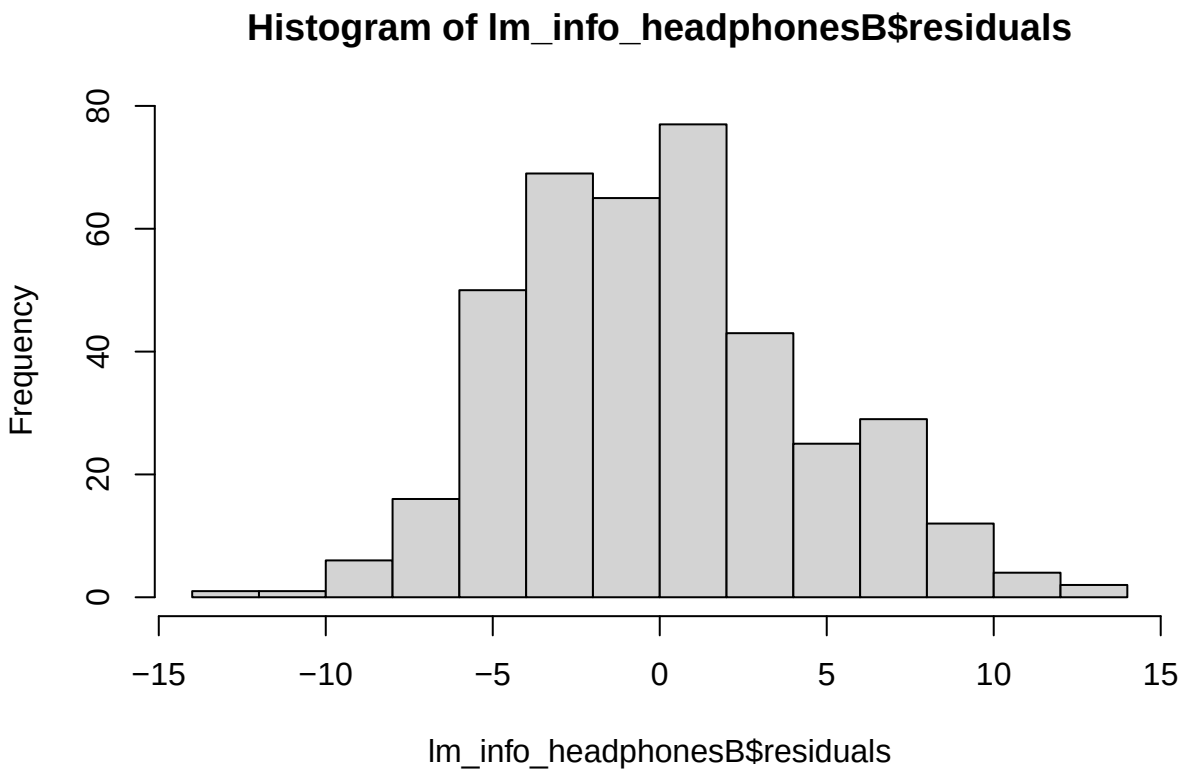
Judging by the VIF values of the explanatory variables, strong multicollinearity is not present. None of the VIF values go above 5, so multicollinearity is not an issue here.

Plots for Linear Assumptions:

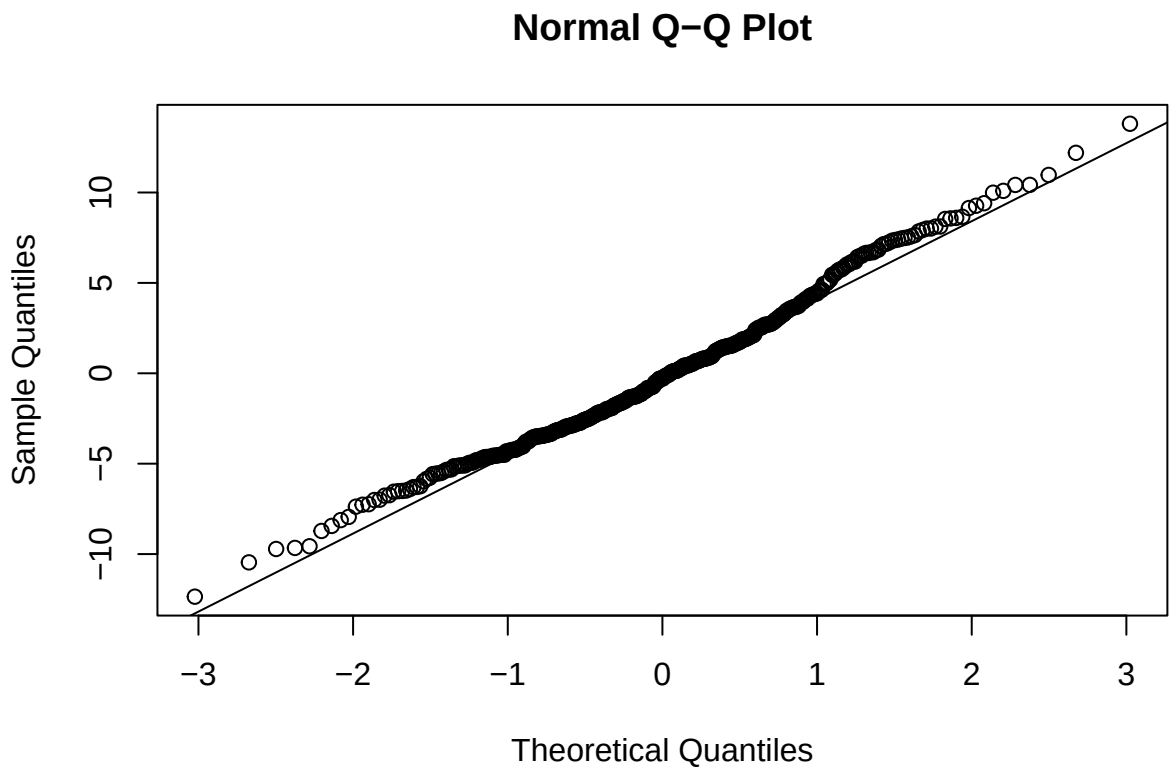
```
plot(x=lm_info_headphonesB$fitted.values,
     y=lm_info_headphonesB$residuals,
     xlab="Fitted Value",
     ylab="Residual",
     main="Headphones Residual Plot")
abline(a=0,b=0,lwd=3)
```




```
hist(lm_info_headphonesB$residuals)
```



```
qqnorm(lm_info_headphonesB$residuals)
qqline(lm_info_headphonesB$residuals)
```



Judging by the residual plot, the linearity assumption is met, because there is random scatter of the residuals around the zero residual line. The residual plot also suggests to us that the equal variance of errors assumption is met because there appears to be a roughly constant vertical spread across the plot. The independent errors assumption is met due to the fact that a random sample of 400 unique stores was taken, which means no store was sampled twice. Also, time is not a factor here, which is typically the main culprit in this assumption being violated. Lastly, judging by the residual histogram and residual normal QQ plot, we can assume normality of errors. The histogram does not demonstrate any extreme skewness, as it is approximately normally distributed. All of the points on the normal QQ plot hug close to the QQ line as well, so we can assume normality and proceed.

Hypotheses and Significance Level; is at least one explanatory variable statistically significantly related to the number of units sold?
H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
H_a: At least one of $b_j \neq 0$ (j = 1,2,3,4,5)
$\alpha = 0.05$

Hypothesis Test:

```
summary(lm_info_headphonesB)
```

```
##
## Call:
## lm(formula = Units ~ ., data = headphonesB)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -12.3554  -3.1388  -0.2881   2.6877  13.8027
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.481917   1.895095   8.169 4.26e-15 ***
## Price        -0.098541   0.005721 -17.224  < 2e-16 ***
## Price.Compet  0.099254   0.008884  11.173  < 2e-16 ***
## Advert        0.027940   0.003318   8.421 7.02e-16 ***
## UrbanYes     -0.186448   0.483360  -0.386      0.7
## Age50Yes      2.367320   0.450525   5.255 2.44e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.385 on 394 degrees of freedom
## Multiple R-squared:  0.4884, Adjusted R-squared:  0.4819
## F-statistic: 75.22 on 5 and 394 DF,  p-value: < 2.2e-16
```

p-value $< 2.2 * 10^{-16}$
Our overall p-value of $< 2.2 * 10^{-16}$ is less than our α of 0.05, and so we reject the null hypothesis. There is evidence suggesting at least one of the explanatory variables is statistically significantly related to units sold.
The explanatory variables statistically significantly related to number of units sold are Price (p-value $< 2 * 10^{-16}$), Price.Compet (p-value $< 2 * 10^{-16}$), Advert (p-value $= 7.02 * 10^{-16}$), and Age50 (p-value $= 2.44 * 10^{-7}$). The explanatory variable of Urban is not statistically significantly related to number of units sold, as it's p-value of 0.7 is greater than our α of 0.05. Therefore, we will remove it from the model and proceed with the other four.

New Model, without taking into consideration whether or not the store is located in an urban area:

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':
##
##     recode

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
headphonesB_final <- select(headphonesB, -Urban)
lm_info2_headphonesB <- lm(Units ~., data = headphonesB_final)
summary(lm_info2_headphonesB)
```

```
##
## Call:
## lm(formula = Units ~ ., data = headphonesB_final)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.4054  -3.1945  -0.2196   2.6530  13.7480
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.404562   1.882423   8.183 3.83e-15 ***
## Price        -0.098561   0.005715 -17.247  < 2e-16 ***
## Price.Compet  0.099063   0.008860  11.181  < 2e-16 ***
## Advert        0.027878   0.003310   8.421 6.96e-16 ***
## Age50Yes      2.378091   0.449174   5.294 1.99e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.381 on 395 degrees of freedom
## Multiple R-squared:  0.4882, Adjusted R-squared:  0.483
## F-statistic: 94.19 on 4 and 395 DF,  p-value: < 2.2e-16
```

$\hat{Units_i} = \beta_0 + \beta_1(Price_i) + \beta_2(Price.Compet_i) + \beta_3(Advert_i) + \beta_4(Age50_i)$
$\hat{Units_i} = 15.4046 - 0.0986(Price_i) + 0.0991(Price.Compet_i) + 0.0279(Advert_i) + 2.3781(Age50_i)$

$\beta_0 = 15.4046$; it is not appropriate to interpret here, because no store in the dataset have a value of 0 for each predictor variable (Price, Price.Compet, Advert, Age50).
$\beta_1 = -0.0986$; holding the explanatory variables of competitor's price, advertising expenditures, and whether or not the average age of the people living in the area is under 50 years constant, as the price of the headphones increases by \$1, the predicted number of headphones sold in the store decreases by 98.6 units.
$\beta_2 = 0.0991$; holding the explanatory variables of own price, advertising expenditures, and whether or not the average age of the people living in the area is under 50 years constant, as the competitor's price of the headphones increases by \$1, the predicted number of headphones sold in the store increases by 99.1 units.
$\beta_3 = 0.0279$; holding the explanatory variables of one's price, competitor's price, and whether or not the average age of the people living in the area is under 50 years constant, as the advertising expenditures of the store increases by \$1,000, the predicted number of headphones sold in the store increases by 27.9 units.

β_4 interpretation:

```
contrasts(headphonesB_final$Age50)
```

```
##     Yes
## No    0
## Yes   1
```

$\beta_4 = 2.3781$; holding the explanatory variables of one's price, competitor's price, and advertising budget constant, stores in areas where the average age of the residents is under 50 years sell 2378.10 more units on average than stores in areas where the average age of the residents is over 50 years.

$R^2 = 0.4882$; 48.82% of the variation in units of headphones sold can be explained by the regression model that contains the store's selling price of the unit, the competitor's selling price of the unit, the store's advertising budget, and whether or not the average age of the people living in the area is under 50 years.

$s = 4.381$; A typical difference between the number of units sold in the dataset and the corresponding predicted number of units sold as predicted by the regression model is 4,381 units.

Prediction for $y_{Units|Price=250,Price.Compet=260,Advert=4800,Age50=Yes}$:

```
predict(object=lm_info2_headphonesB,
        newdata=data.frame(
        Price=250,
        Price.Compet=260, Advert=4.8,
        Age50="Yes"))
```

```
##        1
## 19.03242
```

The store in the area where the average age of the residents is 48 years, that sells its headphones at $250 per unit whose competitors sell theirs at $260 per unit, and who has an advertising budget of $4,800 is predicted to have sold 19,032.42 headphones last year.

95% prediction interval for $y_{Units|Price=250,Price.Compet=260,Advert=4800,Age50=Yes}$:

```
predict(object=lm_info2_headphonesB,
        newdata=data.frame(
        Price=250,
        Price.Compet=260, Advert=4.8,
        Age50="Yes"),
        interval="prediction")
```

```
##        fit      lwr      upr
## 1 19.03242 10.38126 27.68358
```

With 95% confidence, the number of units sold last year for this store in the area where the average age of the residents is 48 years, that sells its headphones at $250 per unit whose competitors sell theirs at $260 per unit, and who has an advertising budget of $4,800 is predicted to be between 10,381.26 and 27,683.58 headphones.

Residual for units sold in store with greatest advertising expenditures:

```
23.78 - predict(object=lm_info2_headphonesB,
                newdata=data.frame(
                Price=335,
                Price.Compet=358,
                Advert=291,
                Age50="No"))
```

```
##         1
## -2.183576
```

The residual of -2.1836 means that the number units sold in the store with the highest advertising budget in the dataset is 2.1836 less than its predicted units sold, according to the multiple linear regression model.