Assignment 2

Brian Morales

September 11, 2022

1

```r
library(readr)
library(MASS)
Carseats <- read_csv("~/Desktop/Fall-2022/Stats-Learning/ALL-CSV-FILES/Carseats.csv", show_col_types = F

lm.fit = lm(Sales~Price+Urban+US, Carseats)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

The summary table above displays the estimated coefficients with standard error, t-value and p-value. What want to heed to is the p-value. Notice that the p-value is significant for `Price` and `USYes`, implying that these estimates impact the outcome. `UrbanYes` does not have a significant p-value, therefore it does not affect the response. Below the coefficients we have `signif. codes`; this tells us how significant our coefficients are. The more asterisk, the more significant the p-value. The bottom end of the table tells us how well our model is doing, overall. We have the Residual Standard Error, R^2, adjusted R^2, and the F-statistic. The R^2 is telling us the model is only explaining about 23% of variation in the median sales of `Carseats`. R^2 can be very greedy therefore we pay attention to the adjusted R^2, which penalizes for the model complexity. The F-statistic test if at least one of the predictors is useful in the response. Our F-statistic is large and the p-value is significant therefore the null hypothesis should be rejected. Usually, we want our F-statistic to be large because the p-value will be close to 0.

```r
lm.fit2 = lm(Sales~Price+US, Carseats)
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

`lm.fit2` has removed the `UrbanYes` variable and notice that all our coefficients have a significant p-value. More importantly, we want to look at the adjusted R^2 and the F-statistic. The adjusted R^2 is pretty much the same as before however, the f-statistic has increase by 20 and the p-value is still significant. This shows us more evidence that at least one of the predictors is associate with the response.

2

```r
lm.full = lm(Sales ~ ., Carseats)

summary(lm.full)
```

```
##
## 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
```

Its R^2 has increased therefore the full model is explaining 87% of variation in the median sales of `Carseats`. Its adjusted R^2 also increased. Usually, the adjusted R^2 penalizes for the models complexity, however, the adjusted $R^2 = 87\%$, exhibiting the full model is a better fit than the previous two.

3

```
lm.reduced = lm(Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc + Age, Carseats)

summary(lm.reduced)
```

```
##
## 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
```

Notice that all the estimated coefficients have a significant p-value and the F-statistics has increased by over 100. The R^2 and adjusted R^2 is similar to the previous model.

4

```
anova(lm.full, lm.reduced)
```

```
## 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
```

```
anova(lm.full, lm.fit2)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ CompPrice + Income + Advertising + Population + Price +
##     ShelveLoc + Age + Education + Urban + US
## Model 2: Sales ~ Price + US
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    388  402.83
## 2    397 2420.87 -9     -2018 215.97 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

That statistical hypothesis we are testing is whether we should reject the null model or not - reject `lm.full` or accept `lm.full`. Observing the first table we see that the p-value is greater than $\alpha = 0.05$ hence we can get an equally food fit with the reduced model - reject the `lm.full`. Doing a formal f-test on `lm.full` and `lm.fit2`, we see that the p-value is significant therefore we cannot reject `lm.full` and `lm.fit2` does not do as good of a job as `lm.full`.

5

```
AIC(lm.full, lm.reduced, lm.fit2, lm.fit)
```

```
##            df      AIC
## lm.full    13 1163.974
## lm.reduced  9 1160.470
## lm.fit2     4 1863.319
## lm.fit      5 1865.312
```

```
BIC(lm.full, lm.reduced, lm.fit2, lm.fit)
```

```
##            df      BIC
## lm.full    13 1215.863
## lm.reduced  9 1196.393
## lm.fit2     4 1879.285
## lm.fit      5 1885.269
```

Using the AIC and BIC test we see that for AIC the model selection is `lm.reduced`, similarly with BIC. `lm.reduced` is the lowest value in AIC = 1160 and BIC = 1196. Note that AIC is a better metric when prediction is the goal and BIC is better when explanation is the goal.

6

```r
swAIC.lm = stepAIC(lm.full, k=2, trace = 0, direction = 'both')
summary(swAIC.lm)
```

```
##
## 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
```

```r
swBIC.lm = stepAIC(lm.full, k=6, trace = 0, direction = 'both')
summary(swBIC.lm)
```

```
##
## 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
```

Utilizing the `stepAIC` in R, we arrive to the same model that was chosen in part 5, `lm.reduced`. Therefore, the best model from using all predictor is `lm.reduced`.

```r
swAIC.lm = stepAIC(lm.reduced, k=2, trace = 0, direction = 'both')
summary(swAIC.lm)
```

```
##
## 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
```

```r
swBIC.lm = stepAIC(lm.reduced, k=6, trace = 0, direction = 'both')
summary(swBIC.lm)
```

```
##
## 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
```

When running `stepAIC` with `lm.reduced` as the argument, we arrive to the same model as before. We can conclude that `lm.reduced` is the best model for the `Carseats` data set.

7

No, we do not expect to arrive at the same "best" model applying step wise selection AIC and BIC each time because, as mentioned above, AIC is a better metric when prediction is the goal and BIC is better when explanation is the focus. Above that, in most cases BIC prefers smaller models because of the $(p+1)log(n)$ compared to $2(p+1)$ in AIC.