[STAT 4610] HW-2 / Michael Ghattas

Michael Ghattas

9/12/2022

Start:

```
library("ISLR")
head(Carseats)

##   Sales CompPrice Income Advertising Population Price ShelveLoc Age
Education
## 1  9.50       138     73          11        276   120       Bad  42
17
## 2 11.22       111     48          16        260    83      Good  65
10
## 3 10.06       113     35          10        269    80    Medium  59
12
## 4  7.40       117    100           4        466    97    Medium  55
14
## 5  4.15       141     64           3        340   128       Bad  38
13
## 6 10.81       124    113          13        501    72       Bad  78
16
##   Urban  US
## 1   Yes Yes
## 2   Yes Yes
## 3   Yes Yes
## 4   Yes Yes
## 5   Yes  No
## 6    No Yes
```

Part (1)

```
lm.fit = lm(Sales ~ Price + Urban + US, data = 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
```

```
lm.fit2 = lm(Sales ~ Price + US, data = 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.fit:

- If the price increases by USD 1000 while other predictors held constant, sales would decrease by ~54.46 units, sales from individuals observed living in Urban areas would decrease by ~21.91 units, and sales from individuals observed living in the US would increase by ~1200.57 units
- A store location in relation to Urban areas has no affect on sales.
- US based stores will on average sell ~1200 more carseats than international stores.

lm.fit2:

- If the price increases by USD 1000 while other predictors held constant, sales would decrease by ~54.48 units, and sales from individuals observed living in the US would increase by ~1199.64 units
- US based stores will on average sell ~1200 more carseats than international stores.

Part (2)

```
lm.fit3 = lm(Sales ~ ., data = Carseats)
summary(lm.fit3)

##
## 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
```

R^2:

- We mote that the adj.R^2 value for lm.fit3 is 0.8698, indicating that ~86.98 of the data can be explained by the model.
- In contrast, we see that the adj.R^2 value for lm.fit2 is 0.2354, indicating that ~23.54% of the data can be explained by the model.
- Furthermore, we see that the adj.R^2 value for lm.fit is 0.2335, indicating that ~23.35% of the data can be explained by the model.

We can see that lm.fit3 is the better fitting model, given the adj. R^2 value, which indicates that the model explains ~63.44% more of the data than lm.fit and ~63.63% more of the data than lm.fit2.

Part (3)

```
lm.fit4 = lm(Sales ~ . - (Population + Education + Urban + US), data =
Carseats)
summary(lm.fit4)

##
## Call:
## lm(formula = Sales ~ . - (Population + Education + Urban + US),
##     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
```

Part (4)

```
f.test = var.test(lm.fit4, lm.fit3)
f.test
```

```
##
##  F test to compare two variances
##
## data:  lm.fit4 and lm.fit3
## F = 1.001, num df = 392, denom df = 388, p-value = 0.9924
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8204894 1.2210589
## sample estimates:
## ratio of variances
##           1.000984
```

The f.test = 0.99902, while the p-value for the test is 0.9924. We can clearly see that p-value > significance level (0.05) and conclude that there is no significant difference between the two variances.

F-Statistic: The F-test statistic or F-ratio is simply a scaled version of ΔSSE: {([SSE(R) − SSE(F)]/Δp) / σˆF2} = [(ΔSSE/Δp) / MSEF], where:

1. SSE(R) is the reduced model SSE
2. SSE(F) is the full model SSE
3. Δp is the number of coefficients being tested
4. σˆF2 = MSEF is the full-model estimate of the random error variance σ2.

Note that the numerator of F is essentially the average reduction in SSE per predictor eliminated from the full model. Since the numerator is in units of Y squared and the denominator σˆF2 is also in units of Y squared, F is dimensionless and hence invariant to changes in units.

Hypotheses: The F-test hypotheses are; Ho: All coefficients under consideration are zero. Ha: At least one of the coefficients in nonzero.

```
f.test2 = var.test(lm.fit2, lm.fit3)
f.test2

##
##  F test to compare two variances
##
## data:  lm.fit2 and lm.fit3
## F = 5.8734, num df = 397, denom df = 388, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  4.817017 7.159761
## sample estimates:
## ratio of variances
##           5.873377
```

The f.test2 = 5.873377, while the p-value for the test is ~0. Since we can clearly see that p-value < significance level, we can conclude that there is significant difference between the two variances.

Conclusion:

- Based on the the R^2 of lim.fit3, lm.fit3, and lm.fit4, and combined with the output of f.test and f.test2, we can conclude that the reduced model (lm.fit4) has the best fit, though not much better than the full model (lm.fit3), and performs much better than lm.fit2.

Part (5)

```
aic1 = AIC(lm.fit4, lm.fit3)
aic1

##         df      AIC
## lm.fit4  9 1160.470
## lm.fit3 13 1163.974

aic2 = AIC(lm.fit2, lm.fit3)
aic2

##         df      AIC
## lm.fit2  4 1863.319
## lm.fit3 13 1163.974

bic1 = BIC(lm.fit4, lm.fit3)
bic1

##         df      BIC
## lm.fit4  9 1196.393
## lm.fit3 13 1215.863

bic2 = BIC(lm.fit2, lm.fit3)
bic2

##         df      BIC
## lm.fit2  4 1879.285
## lm.fit3 13 1215.863
```

Yes:

- aic1: Given the output we can clearly see that aic1 indicates that the AIC score for lm.fit4 < lm.fit3, indicating that lm.fit4 is a better fit, though not by much.
- aic2: Given the output we can clearly see that aic2 indicates that the AIC score for lm.fit3 < lm.fit2, indicating that lm.fit3 is a much better fit.

- bic1: Given the output we can clearly see that aic1 indicates that the BIC score for lm.fit4 < lm.fit3, indicating that lm.fit4 is a better fit, though not by much.
- bic2: Given the output we can clearly see that aic2 indicates that the BIC score for lm.fit3 < lm.fit2, indicating that lm.fit3 is a much better fit.

Part (6)

```
aic3 = AIC(lm.fit, lm.fit2)
aic3

##         df      AIC
## lm.fit   5 1865.312
## lm.fit2  4 1863.319

aic4 = AIC(lm.fit, lm.fit3)
aic4

##         df      AIC
## lm.fit   5 1865.312
## lm.fit3 13 1163.974

aic5 = AIC(lm.fit, lm.fit4)
aic5

##         df      AIC
## lm.fit   5 1865.312
## lm.fit4  9 1160.470

aic6 = AIC(lm.fit2, lm.fit3)
aic6

##         df      AIC
## lm.fit2  4 1863.319
## lm.fit3 13 1163.974

aic7 = AIC(lm.fit2, lm.fit4)
aic7

##         df      AIC
## lm.fit2  4 1863.319
## lm.fit4  9 1160.470
```

```
aic8 = AIC(lm.fit3, lm.fit4)
aic8

##         df      AIC
## lm.fit3 13 1163.974
## lm.fit4  9 1160.470

bic3 = AIC(lm.fit, lm.fit2)
bic3

##         df      AIC
## lm.fit   5 1865.312
## lm.fit2  4 1863.319

bic4 = BIC(lm.fit, lm.fit3)
bic4

##         df      BIC
## lm.fit   5 1885.269
## lm.fit3 13 1215.863

bic5 = BIC(lm.fit, lm.fit4)
bic5

##         df      BIC
## lm.fit   5 1885.269
## lm.fit4  9 1196.393

bic6 = BIC(lm.fit2, lm.fit3)
bic6

##         df      BIC
## lm.fit2  4 1879.285
## lm.fit3 13 1215.863

bic7 = BIC(lm.fit2, lm.fit4)
bic7

##         df      BIC
## lm.fit2  4 1879.285
## lm.fit4  9 1196.393
```

```
bic8 = BIC(lm.fit3, lm.fit4)
bic8

##         df      BIC
## lm.fit3 13 1215.863
## lm.fit4  9 1196.393
```

Each model needs thorough examination and analysis, and based on its characteristics, one would need to use the right tools. However, AIC, BIC , or Stepwise regression techniques can help identify the right steps that need to be taken in eliminating or accepting the models to work with an choose from. My recommendation is to use both AIC and BIC. Most of the times they will agree on the preferred model, when they don't, just report it. There is no one type approach to finding the right model using only AIC, BIC, or Stepwise regression.

Part (7)

The AIC tries to select the model that most adequately describes an unknown, high dimensional reality. This means that reality is never in the set of candidate models that are being considered. On the contrary, BIC tries to find the TRUE model among the set of candidates. I find it quite odd the assumption that reality is instantiated in one of the models that the researchers built along the way. This is a real issue for BIC.

Nevertheless, there are a lot of researchers who say BIC is better than AIC, using model recovery simulations as an argument. These simulations consist of generating data from models A and B, and then fitting both datasets with the two models. Overfitting occurs when the wrong model fits the data better than the generating. The point of these simulations is to see how well AIC and BIC correct these overfits. Usually, the results point to the fact that AIC is too liberal and still frequently prefers a more complex, wrong model over a simpler, true model. At first glance these simulations seem to be really good arguments, but the problem with them is that they are meaningless for AIC. As I said before, AIC does not consider that any of the candidate models being tested is actually true. According to AIC, all models are approximations to reality, and reality should never have a low dimensionality. At least lower than some of the candidate models.

My recommendation is to use both AIC and BIC. Most of the times they will agree on the preferred model, when they don't, just report it. There is no one type approach to finding the right model using only AIC, BIC, or Stepwise regression. Each model needs thorough examination and analysis, and based on its characteristics, one would need to use the right tools. However, AIC, BIC , or Stepwise regression techniques can help identify the right steps that need to be taken in eliminating or accepting the models to work with an choose from.