Homework 2, Chapter 3

Linear Regression

==

Reading the data in

How to read data from a URL:

We can import a txt file or csv file directly into RStudio from a location on the web, by specifying the URL:

```
Carseats = read.csv(
  "https://raw.githubusercontent.com/JWarmenhoven/ISLR-python/master
  /Notebooks/Data/Carseats.csv"
  , header=TRUE)
```

How to use the data directly from the ISLR R-package:

You can install the ISLR package in RStudio (need to do this only once). Then, if successfully installed, you can use every time you need it by following and also access Carseats dataset directly by “attach”-ing the Carseats:

```
#install.packages("ISLR")
library(ISLR)
attach(Carseats)
```

Python:

If you want to use Python, take a look at the ISLR package for Python at: https://github.com/JWarmenhoven/ISLR-python

Exercise 10.2

Building on the previous homework, recall the two simple models we started with:

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

`lm.fit2 = lm(Sales~Price+US)`

Fit both of those models, and summarize their output.

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

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US)
##
## 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)
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US)
##
## 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
```

The output above indicates that we have the following fitted models

Model 1:

$$E(Sales) = 13.04 - 0.05 * Price - 0.02 * UrbanYes + 1.20 * USYes$$

with errors approximately iid from

$$N(0,\ 2.472^2).$$

Model 2:

$$E(Sales) = 13.04 - 0.05 * Price + 1.20 * USYes$$

with errors approximately iid from

$$N(0,\ 2.469^2).$$

10.2.a

Now fit the full model, with all predictors. Interpret its R^2 value.

```
names(Carseats)
```

```
##  [1] "Sales"       "CompPrice"   "Income"      "Advertising" "Population"
##  [6] "Price"       "ShelveLoc"   "Age"         "Education"   "Urban"
## [11] "US"
```

```
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
```

We now fit the full model using all the predictors:

$$
\begin{gathered}
E(Sales) = 5.73 - 0.0003 * X + 0.09 * CompPrice + 0.02 * Income + 0.12 * Advertising \\
+0.0002 * Pop - 0.1 * Price + 4.85 * ShelveLocGood + 1.96 * ShelveLocGood \\
-0.05 * Age - 0.02 * Educ + 0.13 * UrbanYes - 0.19 * USYes
\end{gathered}
$$

From the output above, we see that the errors are approximately iid from

$$N(0,\ 1.02^2).$$

We see that approximately 87.36% of all variation in car sales is explained by the variation in this full collection of features.

10.2.b.

Fit the reduced model, using only the individual predictors that were significant at 5% level in the full model.

```
lm.fit4 = lm(Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc + Age, data=Carseats)
summary(lm.fit4)
```

```
##
## 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
```

We now fit the model using only `CompPrice`, `Income`, `Advertising`, `Price`, `ShelveLoc`, and `Age` as predictors:

$$E(Sales) = 5.48 + 0.09 * CompPrice + 0.02 * Income + 0.12 * Advertising - 0.1 * Price +$$
$$4.84 * ShelveLocGood + 1.95 * ShelveLocGood - 0.05 * Age$$

and the errors are approximately iid from

$$N(0,\ 1.019^2).$$

Now, approximately 87.2% of all variation in car sales is explained by the variation in this reduced collection of features.

The **reduced model is qualitatively and quantitively very similar to the full model** – the coefficients are close in magnitude, have the same sign, and the errors have very similar estimated variances. In addition, the R^2 of the reduced model is very close to the R^2 of the full model.

10.2.c

Compare the full and reduced model via the F-test. Is the full model necessary, or can we get an equally good fit with the reduced model? What about when you compare the full model and lm.fit2? Can lm.fit2 do as good of a job as the full model?

```
anova(lm.fit3, lm.fit4)
```

```
## 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
```

The full model does not seem necessary, **as we can get a statistically equivalent fit with the reduced model** – the F statistic is 0.98, and its p-value is 0.43, indicating that we cannot reject the null hypothesis:

$$H_0 : \beta_{population} = \beta_{education} = \beta_{urban} = \beta_{US} = 0$$

However, when we compare the full model and lm.fit2 model as the reduced model:

```
anova(lm.fit3, 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
```

we see that the **lm.fit2 cannot do as good of a job as the full model**. The full model does seem necessary, as we cannot get a statistically equivalent fit with the reduced model – the F statistic is 194.19, and its p-value is less than 0.05, indicating that we can reject the null hypothesis that all coefficients except US and Price are 0, in favor of the alternative that at least one of the other coefficients is not 0.

10.2.d

Use AIC and BIC instead of F-test to answer the question in part (c)

```
AIC(lm.fit3, lm.fit4, lm.fit2, lm.fit)
```

```
##         df      AIC
## lm.fit3 13 1163.974
## lm.fit4  9 1160.470
## lm.fit2  4 1863.319
## lm.fit   5 1865.312
```

AIC suggest that lm.fit3 is better than lm.fit4 as it has a smaller AIC value. It also suggest that lm.fit2 is worse than the full model (lm.fit3). Furthermore, AIC suggests that the best model is lm.fit4, as it has the smallest AIC value among all 4 models.

```
BIC(lm.fit3, lm.fit4, lm.fit2, lm.fit)
```

```
##         df      BIC
## lm.fit3 13 1215.863
## lm.fit4  9 1196.393
## lm.fit2  4 1879.285
## lm.fit   5 1885.269
```

BIC suggest that lm.fit3 is better than lm.fit4 as it has a smaller BIC value. It also suggest that lm.fit2 is worse than the full model (lm.fit3). Furthermore, BIC suggests that the best model is lm.fit4, as it has the smallest BIC value among all 4 models.

10.2.e

Use the stepwise regression (once with AIC and once with BIC) to arrive at the best model. Did you arrive at the single best model? What about the best model from part 3?

```
library(MASS)

swAIC.lm = stepAIC(lm.fit3, 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
```

The above command is using the backward-forward stepwise regression method, with AIC as the guide in model building. We have arrived at the same model as lm.fit3. The command below is also using the backward-forward stepwise regression method, but with BIC as the guide in model building. We have also arrived at the same model as lm.fit3. **So all methods here agree, and have arrived at the single best model (lm.fit3).**

```
swBIC.lm = stepAIC(lm.fit3, k=6, trace=0, direction="both")
# note: for BIC we need k=log(n). Here, log(400) = 6
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
```

10.2.f

Would you expect to always arrive at the same "best" model with AIC and BIC? Why or why not?

In general, I would **not expect the agreement on the "best model"** between AIC, BIC, or stepwise regression. In theory, all could find a different best model.