Homework 8 Solution

STAT 462 (Fall 2019)

UNAUTHORIZED DISTRIBUTION AND/OR UPLOADING OF THIS DOCUMENT IS STRICTLY PROHIBITED.

Write your answers after each sub-question. Copy and paste any R codes, outputs/plots (or screenshots) used to answer any sub-questions.

The dataset *HollywoodMovies2011.csv* includes information on movies that came out of Hollywood in 2011. We want to build a model to preict *profitability*, which is the percent of the budget recovered in profits. The dataset contains five **explanatory** variables and the details are as follows.

Variable Name	Description
RottenTomatoes	Meata rating of critical reviews, from the Rotten Tomatoes website
AudienceScore	Average audience score, from the Rotten Tomatoes website
TheatersOpenWeek	Number of theaters showing the moview on opening weekend
BOAverageOpenWeek	Average box office revenue per heater opening week-end, in dollars
DomesticGross	Gross revenus in the US by the end of 2011, in millions of dollars

---

(1). Read the data into *R* and fit the full first order regression model. Write down the estimated regression line. This model will be referred to as model 1 hereafter. Provide any R outputs you might have used. (15 points total, 5 model fit, 5 regression line, 5 R-output)

```
hollywoodData = read.table("HollywoodMovies2011.csv", header = T, sep=",")
FullModel = lm(Profitability ~ RottenTomatoes + AudienceScore + TheatersOpenWeek +
               BOAverageOpenWeek + DomesticGross, data = hollywoodData)
summary(FullModel)
```

```
##
## Call:
## lm(formula = Profitability ~ RottenTomatoes + AudienceScore +
##     TheatersOpenWeek + BOAverageOpenWeek + DomesticGross, data = hollywoodData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.711 -2.064 -0.700  0.406 60.424
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.420e+00  4.209e+00   1.288   0.2004
## RottenTomatoes     7.630e-02  4.342e-02   1.758   0.0816 .
## AudienceScore     -9.577e-02  7.334e-02  -1.306   0.1943
## TheatersOpenWeek  -5.657e-04  9.960e-04  -0.568   0.5712
## BOAverageOpenWeek -3.603e-05  7.769e-05  -0.464   0.6438
```

```
## DomesticGross      2.781e-02  1.577e-02   1.763   0.0806 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.889 on 112 degrees of freedom
## Multiple R-squared:  0.06438,    Adjusted R-squared:  0.02261
## F-statistic: 1.541 on 5 and 112 DF,  p-value: 0.1828
```

Let

y = Profitability

x_1 = RottenTomatoes

x_2 = AudienceScore

x_3 = TheatersOpenWeek

x_4 = BOAverageOpenWeek

x_5 = DomesticGross

Then the estimated full first order regression model for this data is:

$$\hat{y} = 5.42 + 0.076x_1 - 0.096x_2 - 0.0006x_3 - 0.00004x_4 + 0.028x_5$$

(2). By looking at the test results for the partial slopes (at 10% level), identify any variables you would like to drop from model 1. Provide reasons for your choice(s). You do not have to write down any steps for hypothesis testing here. (15 points total, 10 points for determining correct variables to drop, 5 points for correct reasoning)

At 10% level, we can drop x_2, x_3, and x_4 from the model, because the p-values for testing for partial slope for each of these variables are greater than 0.1 indicating these variables has no predictive value after adjusting for other variables.

(3). Re-ft the model by eliminating the variables you decided to drop in part (2). Write down the estimated regression line. This model will be referred to as model 2 hereafter. Provide any R outputs you might have used. (15 points total, 5 model fit, 5 regression line, 5 R-output)

```
RedModel = lm(Profitability ~ RottenTomatoes +  DomesticGross, data = hollywoodData)
summary(RedModel)
```

```
##
## Call:
## lm(formula = Profitability ~ RottenTomatoes + DomesticGross,
##     data = hollywoodData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.752 -1.938 -1.029  0.264 60.883
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.836652   1.445788   0.579   0.5639
## RottenTomatoes 0.030925   0.024465   1.264   0.2088
## DomesticGross  0.016305   0.009245   1.764   0.0805 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.853 on 115 degrees of freedom
## Multiple R-squared:  0.04946,    Adjusted R-squared:  0.03293
## F-statistic: 2.992 on 2 and 115 DF,  p-value: 0.0541
```

Then the estimated full first order regression model for this data is:

$$\hat{y} = 0.84 + 0.031x_1 + 0.016x_5$$

(4). Use adjusted R^2 and forward model selection approach to fit the **best** first order model for this data. Summarize your R outputs in a table as below. your table should clearly illustrates the step-by-step approach for forward model selection. Once you identify the **best** model using adjusted R^2, fit the model and write down the estimated regression line. (55 points total, 3 points for each correct adjusted R^2 (3*14=42), 3 points for correct model decision, 5 points for model fit, 5 points for regression line)

Variable(s)	R^2	Adj. R^2
RottenTomatoes	0.024	0.015
AudienceScore	0.010	0.001
TheatersOpenWeek	0.006	- 0.002
BOAverageOpenWeek	0.008	-0.001
DomesticGross	0.036	0.028
DomesticGross, RottenTomatoes	0.049	0.033
DomesticGross, AudienceScore	0.037	0.021
DomesticGross, TheatersOpenWeek	0.038	0.021
DomesticGross, BOAverageOpenWeek	0.036	0.021
DomesticGross, RottenTomatoes, AudienceScore	0.061	0.037
DomesticGross, RottenTomatoes, TheatersOpenWeek	0.050	0.025
DomesticGross, RottenTomatoes, BOAverageOpenWeek	0.050	0.025
DomesticGross, RottenTomatoes, AudienceScore, TheatersOpenWeek	0.063	0.029
DomesticGross, RottenTomatoes, AudienceScore, BOAverageOpenWeek	0.062	0.028

Using the adjusted R^2 values the chosen model is the one with x_5, x_1, and x_2 as predictors. The estimated regression line for the chosen *best* model using R^2 is

```
bestModel = lm(Profitability ~ RottenTomatoes +  DomesticGross + AudienceScore, data = hollywoodData)
summary(bestModel)
```

```
##
## Call:
## lm(formula = Profitability ~ RottenTomatoes + DomesticGross +
##     AudienceScore, data = hollywoodData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.787 -2.001 -0.882  0.380 60.655
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.526278   2.674968   1.318   0.1901
## RottenTomatoes  0.072362   0.042431   1.705   0.0908 .
## DomesticGross   0.020227   0.009795   2.065   0.0412 *
```

```
## AudienceScore  -0.083831   0.070200  -1.194   0.2349
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.84 on 114 degrees of freedom
## Multiple R-squared:  0.06121,    Adjusted R-squared:  0.0365
## F-statistic: 2.477 on 3 and 114 DF,  p-value: 0.06487
```

$$\hat{y} = 3.53 + 0.072x_1 - 0.084x_2 + 0.02x_5$$