Lab 7

Directions: Workout the problems using R markdown. Hand in both the *.rmd file and the knitted *.pdf file. (2 points for correctly submitting)

Airbnb Pricing Data Victoria would like to list her property on Airbnb in Edinburgh, the capitol of Scotland. In order to price her property competitively, she collects data on listings to analyze the contributing factors of price. The `AirBnb.csv` data set consists of 10,370 rental listings from Airbnb in Edinburgh for a period from June 25, 2019 to June 24, 2020.

The variables for each listing are:

- Bathrooms - Number of bathrooms
- Bedrooms - Number of bedrooms
- Beds - Number of beds
- Accommodates - Number of guests the listing can accommodate
- Guests - Number of guests included without an additional fee
- MinNights - Minimum number of nights required for booking
- MaxNights - Maximum number of nights the listing can be rented
- ExtraPeople - Average fee for each additional person in British pounds
- HostListings - Number of listings the host manages
- ResponseRate - Average host response rate
- Deposit - Average security deposit required for booking in British pounds
- CleaningFee - Average cleaning fee charged in British pounds
- FeeMissing - A dummy variable that is 1 if the cleaning fee is missing, 0 otherwise
- Price - Average price of the listing in British pounds.

```r
df = read.csv("https://www.businessregression.com/Data/AirBnb.csv")
```

Using the Edinburgh Airbnb data file, do the following.

1. Airbnb Pricing Application: Backward Elimination Using the Edinburgh Airbnb data file, do the following.

a. Fit a linear regression model using all possible predictor variables.

```r
reg = lm(Price ~ ., data = df)
summary(reg)
```

```
##
## Call:
## lm(formula = Price ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -131.302  -24.807   -8.201   15.966  248.684
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.576e+01  3.805e+00   4.140 3.49e-05 ***
## Bathrooms      1.266e+01  1.032e+00  12.258  < 2e-16 ***
## Bedrooms      8.551e+00  9.413e-01   9.084  < 2e-16 ***
## Beds          3.300e+00  6.704e-01   4.922 8.70e-07 ***
## Accommodates  1.288e+01  5.216e-01  24.698  < 2e-16 ***
## Guests       -3.618e-01  3.742e-01  -0.967 0.333641    
## MinNights     2.015e+00  3.517e-01   5.730 1.03e-08 ***
## MaxNights    -1.495e-03  7.393e-04  -2.023 0.043124 *  
## ExtraPeople  -1.667e-01  3.590e-02  -4.643 3.47e-06 ***
## HostListings  3.977e-01  2.477e-02  16.054  < 2e-16 ***
## ResponseRate -1.295e+01  3.579e+00  -3.619 0.000298 ***
## Deposit       6.053e-02  5.223e-03  11.589  < 2e-16 ***
## CleaningFee   8.095e-02  2.354e-02   3.440 0.000585 ***
## FeeMissing    2.164e+00  9.393e-01   2.304 0.021259 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 40.32 on 10356 degrees of freedom
## Multiple R-squared:  0.4764, Adjusted R-squared:  0.4758 
## F-statistic: 724.9 on 13 and 10356 DF,  p-value: < 2.2e-16
```

b. Run backward elimination beginning from the full model in the previous part.

```
full = lm(Price ~., data = df)
BE = step(full)
```

c. Specify which variable is eliminated in the first iteration of backward elimination. How many variables are eliminated in total using this process?

Guests is eliminated in the first iterations of backward elimination, for a total of one variable eliminated

d. Print a model summary of the backward elimination model.

```
summary(BE)
```

```
## 
## Call:
## lm(formula = Price ~ Bathrooms + Bedrooms + Beds + Accommodates + 
##     MinNights + MaxNights + ExtraPeople + HostListings + ResponseRate + 
##     Deposit + CleaningFee + FeeMissing, data = df)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -130.615  -24.697   -8.165   15.920  248.739 
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   1.557e+01  3.800e+00   4.097 4.23e-05 ***
## Bathrooms     1.266e+01  1.032e+00  12.259  < 2e-16 ***
## Bedrooms      8.548e+00  9.413e-01   9.081  < 2e-16 ***
## Beds          3.267e+00  6.696e-01   4.879 1.08e-06 ***
## Accommodates  1.280e+01  5.144e-01  24.880  < 2e-16 ***
## MinNights     2.038e+00  3.509e-01   5.806 6.58e-09 ***
## MaxNights    -1.491e-03  7.393e-04  -2.016 0.043784 *  
## ExtraPeople  -1.779e-01  3.399e-02  -5.233 1.70e-07 ***
## HostListings  3.939e-01  2.447e-02  16.101  < 2e-16 ***
## ResponseRate -1.296e+01  3.579e+00  -3.621 0.000294 ***
## Deposit       6.070e-02  5.220e-03  11.629  < 2e-16 ***
## CleaningFee   7.809e-02  2.335e-02   3.345 0.000827 ***
## FeeMissing    2.249e+00  9.352e-01   2.404 0.016213 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.32 on 10357 degrees of freedom
## Multiple R-squared:  0.4764, Adjusted R-squared:  0.4758
## F-statistic: 785.3 on 12 and 10357 DF,  p-value: < 2.2e-16
```

2. Airbnb Pricing Application: Forward Selection Using the Edinburgh Airbnb data file, do the following.

a. Fit a linear regression model using only the intercept.

```r
reg2 = lm(Price ~ 1, data = df)
```

b. Run forward selection beginning from the model in the previous part.

```r
FS = step(reg2, scope = list(upper = full))
```

c. Specify which variable is incorporated in the first iteration of forward selection.

The first variable incorporated in the first iteration of selection is Accomodates

d. Print a model summary of the forward selection model.

```r
summary(FS)
```

```
##
## Call:
## lm(formula = Price ~ Accommodates + Bathrooms + HostListings +
##     Deposit + Bedrooms + MinNights + ExtraPeople + Beds + ResponseRate +
##     CleaningFee + FeeMissing + MaxNights, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -130.615  -24.697   -8.165   15.920  248.739
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.557e+01  3.800e+00   4.097 4.23e-05 ***
## Accommodates  1.280e+01  5.144e-01  24.880  < 2e-16 ***
## Bathrooms     1.266e+01  1.032e+00  12.259  < 2e-16 ***
## HostListings  3.939e-01  2.447e-02  16.101  < 2e-16 ***
## Deposit       6.070e-02  5.220e-03  11.629  < 2e-16 ***
## Bedrooms      8.548e+00  9.413e-01   9.081  < 2e-16 ***
## MinNights     2.038e+00  3.509e-01   5.806 6.58e-09 ***
## ExtraPeople  -1.779e-01  3.399e-02  -5.233 1.70e-07 ***
## Beds          3.267e+00  6.696e-01   4.879 1.08e-06 ***
## ResponseRate -1.296e+01  3.579e+00  -3.621 0.000294 ***
## CleaningFee   7.809e-02  2.335e-02   3.345 0.000827 ***
## FeeMissing    2.249e+00  9.352e-01   2.404 0.016213 *
## MaxNights    -1.491e-03  7.393e-04  -2.016 0.043784 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.32 on 10357 degrees of freedom
## Multiple R-squared:  0.4764, Adjusted R-squared:  0.4758
## F-statistic: 785.3 on 12 and 10357 DF,  p-value: < 2.2e-16
```

3. Airbnb Pricing Application: Stepwise Regression Using the Edinburgh Airbnb data file, do the following.

a. Run stepwise regression.

```
SW = step(reg2, scope = list(upper = full))
```

b. Specify which variable is incorporated in the first iteration of the stepwise regression.

The variable incorporate in the first iteration of the regression is accomodates

c. Print a model summary of the stepwise regression model.

```
summary(SW)
```

```
##
## Call:
## lm(formula = Price ~ Accommodates + Bathrooms + HostListings +
##     Deposit + Bedrooms + MinNights + ExtraPeople + Beds + ResponseRate +
##     CleaningFee + FeeMissing + MaxNights, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -130.615  -24.697   -8.165   15.920  248.739
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.557e+01  3.800e+00   4.097 4.23e-05 ***
## Accommodates  1.280e+01  5.144e-01  24.880  < 2e-16 ***
## Bathrooms     1.266e+01  1.032e+00  12.259  < 2e-16 ***
## HostListings  3.939e-01  2.447e-02  16.101  < 2e-16 ***
## Deposit       6.070e-02  5.220e-03  11.629  < 2e-16 ***
## Bedrooms      8.548e+00  9.413e-01   9.081  < 2e-16 ***
## MinNights     2.038e+00  3.509e-01   5.806 6.58e-09 ***
## ExtraPeople  -1.779e-01  3.399e-02  -5.233 1.70e-07 ***
## Beds          3.267e+00  6.696e-01   4.879 1.08e-06 ***
## ResponseRate -1.296e+01  3.579e+00  -3.621 0.000294 ***
## CleaningFee   7.809e-02  2.335e-02   3.345 0.000827 ***
## FeeMissing    2.249e+00  9.352e-01   2.404 0.016213 *
## MaxNights    -1.491e-03  7.393e-04  -2.016 0.043784 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.32 on 10357 degrees of freedom
## Multiple R-squared:  0.4764, Adjusted R-squared:  0.4758
## F-statistic: 785.3 on 12 and 10357 DF,  p-value: < 2.2e-16
```

d. Clearly state the difference (if any) between the backwards elimination, forward selection, and the stepwise regression models.

There is no difference between any of the methods

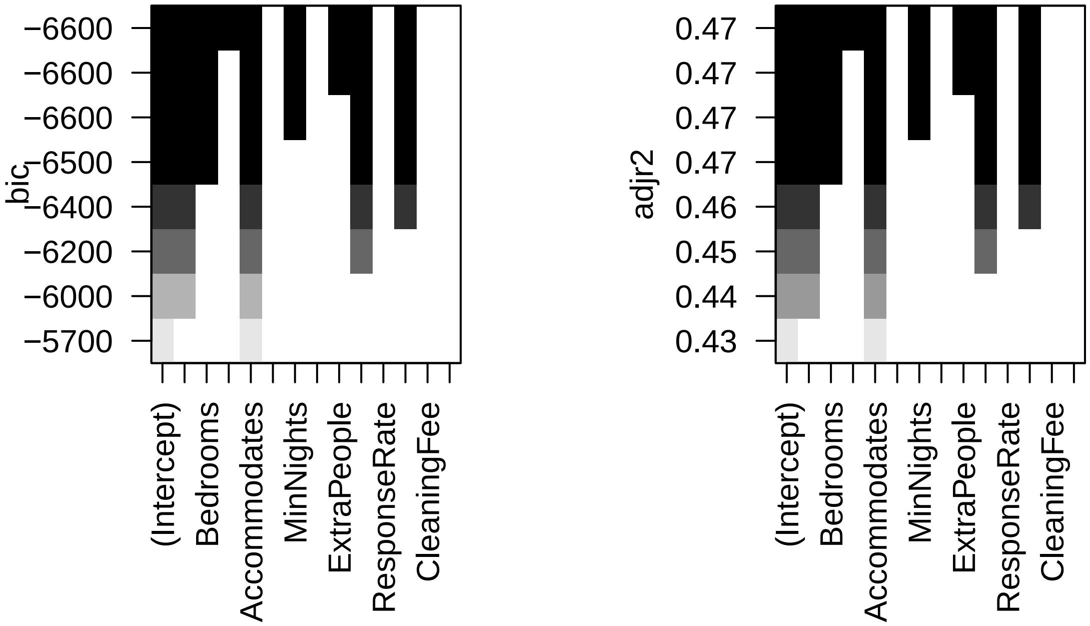
4. Airbnb Pricing Application: Best Subsets Regression 1 Using the Edinburgh Airbnb data file, do the following.

a. Run the best subsets method on the data using all predictor variables using the default value of `nvmax`.

```r
library(leaps)
BSR = regsubsets(Price ~ ., data = df)
```

b. Plot the results of the best subsets method from part a using BIC as the scale. Repeat using adjusted R^2 as the scale.

```r
par(mfrow=c(1,2))
plot(BSR)
plot(BSR, scale = 'adjr2')
```



c. Return the best model coefficients with 8 variables. Use the coef function specifying the best subsets model as the first argument and the number of variables as the second argument.

```r
coef(BSR,8)
```

```
##  (Intercept)    Bathrooms     Bedrooms         Beds Accommodates    MinNights
##   3.16272079  12.73168880   8.98549616   3.26311019  12.87694620   2.24713688
##  ExtraPeople HostListings      Deposit
##  -0.18647225   0.38641557   0.06450636
```

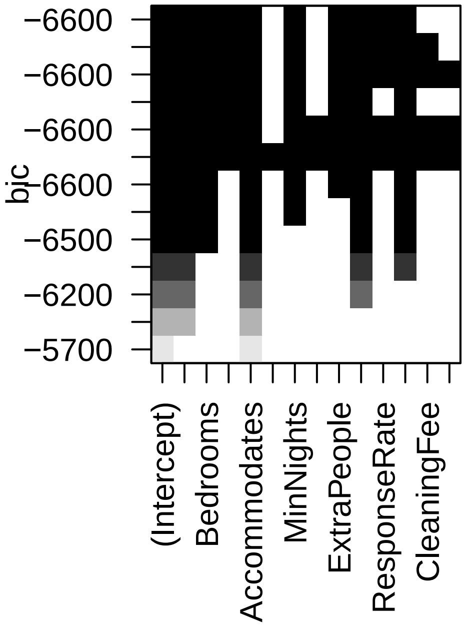
5. Airbnb Pricing Application: Best Subsets Regression 2 Using the Edinburgh Airbnb data file, do the following.

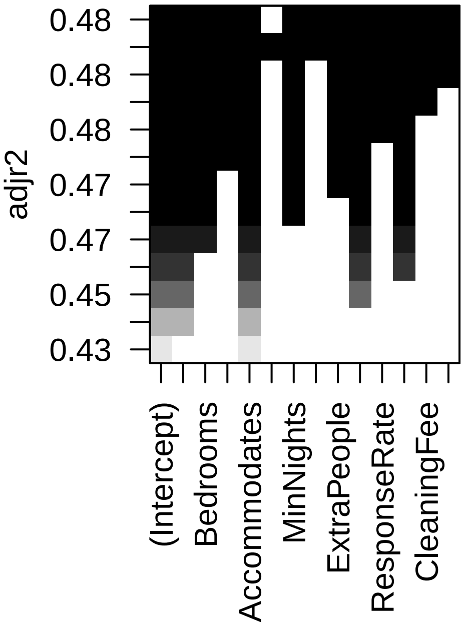
a. Run the best subsets method on the data using all predictor variables using the `nvmax = 13` option to include all possible combinations of predictor variables.

```r
library(leaps)
BSR2 = regsubsets(Price ~ ., nvmax = 13, data = df)
```

b. Plot the results of the best subsets method from part a using BIC as the scale. Repeat using adjusted R^2 as the scale.

```r
par(mfrow= c(1,2))
plot(BSR2)
plot(BSR2, scale = 'adjr2')
```


bic
−6600
−6600
−6600
−6600
−6500
−6200
−5700
(Intercept)
Bedrooms
Accommodates
MinNights
ExtraPeople
ResponseRate
CleaningFee


adjr2
0.48
0.48
0.48
0.47
0.47
0.45
0.43
(Intercept)
Bedrooms
Accommodates
MinNights
ExtraPeople
ResponseRate
CleaningFee