Regression and Time Series HW4

Yaniv Bronshtein

10/12/2021

Import the necessary libraries

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(pls)
```

```
##
## Attaching package: 'pls'
```

```
## The following object is masked from 'package:stats':
##
##     loadings
```

```
library(leaps)
```

Read in csv

(1) Based on the R output, what would be the regression model you suggest?

```
model.prof <- lm(formula=Price~FLR+RMS+BDR+BTH+ST+GAR+FP+LOT,data=house)
summary(model.prof)
```

```
##
## Call:
## lm(formula = Price ~ FLR + RMS + BDR + BTH + ST + GAR + FP +
##     LOT, data = house)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3058  -2.8417  -0.1511   3.2882   7.9518
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.637664   5.240957   3.556 0.002429 **
## FLR          0.017570   0.003235   5.431 4.49e-05 ***
## RMS          3.904374   1.615617   2.417 0.027194 *
## BDR         -7.697444   1.829426  -4.208 0.000592 ***
## BTH          2.374591   2.557865   0.928 0.366221    
## ST          10.818663   2.300203   4.703 0.000205 ***
## GAR          1.770861   1.404310   1.261 0.224334    
## FP           6.909765   3.083583   2.241 0.038680 *  
## LOT          0.263522   0.135109   1.950 0.067808 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.717 on 17 degrees of freedom
## Multiple R-squared:  0.9044, Adjusted R-squared:  0.8595
## F-statistic: 20.11 on 8 and 17 DF,  p-value: 3.147e-07
```

Based on the output, I would suggest the model with Price as the dependent variable and the features: FLR, RMS, BDR, ST, and FP based on the asterisks that R provides in the summary() object and based on the actual p-values. I have provided the trained model.reduced.prof that reflects this idea

(2) Write down the full model and reduced model associated with the following test and draw a conclusion based on this test (Note: This is a row obtained from the R output): GAR 1.77 1.40 1.26 0.2243 **Reduced model based on p-values from full model**

```
model.reduced.prof <- lm(formula=Price~FLR+RMS+BDR+ST+FP, data=house)
summary(model.reduced.prof)
```

```
##
## Call:
## lm(formula = Price ~ FLR + RMS + BDR + ST + FP, data = house)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -13.6561  -2.1638  -0.0816   2.4284   8.7779 
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 24.172544   4.903762   4.929 8.09e-05 ***
## FLR          0.019124   0.003341   5.724 1.33e-05 ***
## RMS          4.863990   1.672008   2.909  0.00868 ** 
## BDR         -7.826966   1.978493  -3.956  0.00078 ***
## ST          11.253185   2.345555   4.798  0.00011 ***
## FP          10.295264   2.849507   3.613  0.00174 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.132 on 20 degrees of freedom
## Multiple R-squared:  0.8669, Adjusted R-squared:  0.8336
## F-statistic: 26.05 on 5 and 20 DF,  p-value: 4.083e-08
```

H0: The coefficient for GAR is equal to 0 H1: The coefficient for GAR is significant and therefore non-zero The F statistic is 26.05

```
model.prof.nogar <- lm(formula=Price~FLR+RMS+BDR+BTH+ST+FP+LOT,data=house)
anova(model.prof.nogar, model.prof)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ FLR + RMS + BDR + BTH + ST + FP + LOT
## Model 2: Price ~ FLR + RMS + BDR + BTH + ST + GAR + FP + LOT
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     18 413.59
## 2     17 378.21  1    35.378 1.5902 0.2243
```

Based on the result of the anova, Pr(>F) is greater than 0.05, so we fail to reject the null hypothesis with greater than 95% confidence. The F statistic is 1.5902

(3) Write down the null hypothesis and alternative hypothesis associated with the F-statistic and what is the conclusion based on the test. *H0: The coefficients for BTH, GAR, and LOT are all equal to 0* H1: There exists at least 1 coefficient the above features that is non-zero

```
anova(model.prof, model.reduced.prof)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ FLR + RMS + BDR + BTH + ST + GAR + FP + LOT
## Model 2: Price ~ FLR + RMS + BDR + ST + FP
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     17 378.21
## 2     20 526.77 -3   -148.56 2.2259 0.1223
```

For the reduced model we can reject H0 because of the high F statistic

(4) The house for selling has 750 square feet of space, 5 rooms, 2 bedrooms, 1.5 baths, storm windows, a 1-car garage, 1 fireplace and a 25 front-foot lot. Analyze the housing price data. Based on this dataset, what can you tell this person about how much he could expect to get for the house? Please report your fitted model and also construct a confidence interval for the prediction (Hint: You can try different variable selection methods to find the final model).

We know the following coefficients:FLR=750, RMS=5, BDR=2, BTH=1.5, FP=1, LOT=25

Create a new dataframe with the specified coefficients to mimic a test set

```
test.df <- data.frame(FLR = 750, RMS = 5, BDR =2, BTH =1.5, ST =1, GAR =1, FP =1,LOT=25)
```

Create the prediction

```
predict(model.prof,test.df, interval="confidence")
```

```
##        fit      lwr      upr
## 1 65.59152 57.69899 73.48406
```

Backward step-wise

```
reg.fit.bwd <- regsubsets(Price~FLR+RMS+BDR+BTH+ST+GAR+FP+LOT,data=house, method="backward")
reg.fit.bwd.sum <- summary(reg.fit.bwd)
```

Get the min bic value for a conservative estimate

```
which.min(reg.fit.bwd.sum$bic)
```

```
## [1] 6
```

#From the backward stepwise, we can remove BTH and GAR variables

#Predicting based on the updated model

```
model.fit.using.bwd <- lm(formula=Price~FLR+RMS+BDR+ST+FP+LOT, data=house)
```

Modified data

```
test.df2 = data.frame(FLR=750, RMS=5, BDR=2, ST=1, FP =1, LOT=25)
```

Create the new prediction

```
predict(model.fit.using.bwd,test.df2, interval="confidence")
```

```
##        fit      lwr      upr
## 1 66.26764 58.72983 73.80545
```

Let's try the forward stepwise method now to see if our results differ

```
reg.fit.fwd <- regsubsets(Price~FLR+RMS+BDR+BTH+ST+GAR+FP+LOT,data=house, method="forward")
```

Let's get the summary

```
reg.fit.bwd.sum <- summary(reg.fit.fwd)
reg.fit.bwd.sum
```

```
## Subset selection object
## Call: regsubsets.formula(Price ~ FLR + RMS + BDR + BTH + ST + GAR +
##     FP + LOT, data = house, method = "forward")
## 8 Variables  (and intercept)
##     Forced in Forced out
## FLR     FALSE      FALSE
## RMS     FALSE      FALSE
## BDR     FALSE      FALSE
## BTH     FALSE      FALSE
## ST      FALSE      FALSE
## GAR     FALSE      FALSE
## FP      FALSE      FALSE
## LOT     FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: forward
##          FLR RMS BDR BTH ST  GAR FP  LOT
## 1  ( 1 ) "*" " " " " " " " " " " " " " "
## 2  ( 1 ) "*" " " " " " " "*" " " " " " "
## 3  ( 1 ) "*" " " " " " " "*" " " "*" " "
## 4  ( 1 ) "*" " " "*" " " "*" " " "*" " "
## 5  ( 1 ) "*" "*" "*" " " "*" " " "*" " "
## 6  ( 1 ) "*" "*" "*" " " "*" " " "*" "*"
## 7  ( 1 ) "*" "*" "*" " " "*" "*" "*" "*"
## 8  ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*"
```

Let's get the min BIC value

```
which.min(reg.fit.bwd.sum$bic)
```

```
## [1] 6
```

Both forward and backward selection concluded that the minimum number of features need is 6 and the same features need to be removed. Thus the confidence intervals must also be the same