Regression and Time Series HW 3

Yaniv Bronshtein

9/28/2021

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(leaps)
```

Use the IPOdecision data (0=withdraw) for the following analysis

(a) Fit a logistic regression model using all predictors in the data file. Comment on which predictors are significant.

```
df <- read.table("/Users/yanivbronshtein/Coding/Rutgers/Regression_TimeSeries_Repo/HW3/IPOdecision.dat"
head(df,5)
```

```
##   decision   logAST debtRatio  debtRatio2 ProfibilityRatio VentureDummy
## 1        0 2.896243 0.3106153 -0.07886888      -0.07886888            0
## 2        0 3.492834 0.1377093  0.04644004       0.06992938            0
## 3        0 4.880762 0.5805391  0.02549893       0.02549893            0
## 4        0 5.060187 0.4438206  0.04583809       0.07394366            0
## 5        0 5.653510 0.6866428 -0.00961194      -0.09611940            0
##   underwriterRating logRevenue DebtPaymentDummy logMarket NASDAQ30dayReturn
## 1              4.50   3.304650                0  15.92572           -0.0263
## 2              8.88   3.060434                1  16.52356            0.0474
## 3              8.88   3.964938                1  17.00186           -0.0203
## 4              8.75   4.165828                1  17.66353           -0.0525
## 5              8.88   4.672829                1  17.88454           -0.1476
##   logNumIPOS
## 1   2.708050
## 2   2.564949
## 3   3.135494
## 4   3.135494
## 5   2.564949
```

Fit a logistic regression

```r
log.reg.fit <- glm(decision~., family=binomial("logit"), data=df)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
summary(log.reg.fit)
```

```
##
## Call:
## glm(formula = decision ~ ., family = binomial("logit"), data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4276   0.0925   0.2914   0.5525   2.3274
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         19.32453    6.99410   2.763  0.00573 **
## logAST               1.17211    0.47586   2.463  0.01377 *
## debtRatio           -3.24485    1.08392  -2.994  0.00276 **
## debtRatio2         -56.36229    8.45966  -6.662 2.69e-11 ***
## ProfibilityRatio    51.98373    7.44449   6.983 2.89e-12 ***
## VentureDummy         1.60128    0.35796   4.473 7.70e-06 ***
## underwriterRating   -0.09156    0.06793  -1.348  0.17771
## logRevenue           0.14882    0.12108   1.229  0.21906
## DebtPaymentDummy    -1.08411    0.33039  -3.281  0.00103 **
## logMarket           -1.31368    0.50518  -2.600  0.00931 **
## NASDAQ30dayReturn    5.56798    3.96144   1.406  0.15986
## logNumIPOS          -0.31168    0.24561  -1.269  0.20444
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 533.04  on 517  degrees of freedom
## Residual deviance: 356.33  on 506  degrees of freedom
## AIC: 380.33
##
## Number of Fisher Scoring iterations: 6
```

The features: debtRatio2, ProfibilityRatio, and Venture Dummy are the most significant given the extremely low P-Values. However, debtRatio, DebtPaymentDummy, logMarket, and logAST are also signficant to a lesser degree

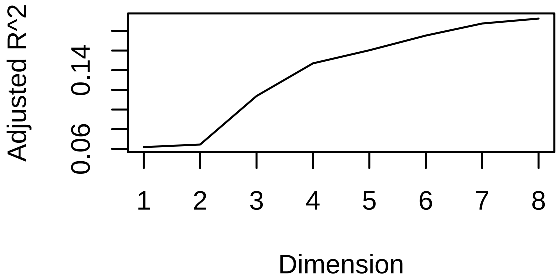
b) Use backward selection procedure to find the (sub)optimal set of predictors to explain the IPO completion/withdrawn decision. Note: Backward selection starts with the largest model and remove the least significant predictor (largest p-value of the z-test) one by one, each time re-estimating the model, until all coefficients are significant at 5% level.

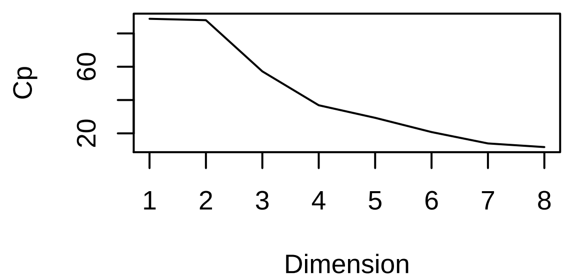
```r
bw.selection.fit <- regsubsets(decision~., method="backward", data=df )
bw.selection.summary <- summary(bw.selection.fit)
```

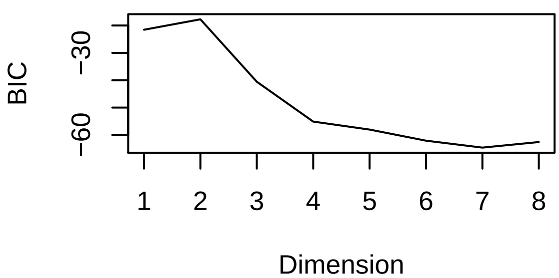
Let us create a plot to determine the optimal set

```r
par(mfrow=c(2,2))
plot(bw.selection.summary$adjr2, type='l', xlab='Dimension', ylab='Adjusted R^2')
plot(bw.selection.summary$cp, type='l', xlab='Dimension', ylab='Cp')
```

```r
plot(bw.selection.summary$bic, type='l', xlab='Dimension', ylab='BIC')
```







Based on the plots of the 3 metrics Adjusted R^2, Cp, and BIC, 4 dimensions is sufficient to capture the data

```r
bw.selection.summary
```

```
## Subset selection object
## Call: regsubsets.formula(decision ~ ., method = "backward", data = df)
## 11 Variables  (and intercept)
##                   Forced in Forced out
## logAST                FALSE      FALSE
## debtRatio             FALSE      FALSE
## debtRatio2            FALSE      FALSE
## ProfibilityRatio      FALSE      FALSE
## VentureDummy          FALSE      FALSE
## underwriterRating     FALSE      FALSE
## logRevenue            FALSE      FALSE
## DebtPaymentDummy      FALSE      FALSE
## logMarket             FALSE      FALSE
## NASDAQ30dayReturn     FALSE      FALSE
## logNumIPOS            FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: backward
##          logAST debtRatio debtRatio2 ProfibilityRatio VentureDummy
## 1  ( 1 ) " "    "*"       " "        " "              " "
## 2  ( 1 ) "*"    "*"       " "        " "              " "
## 3  ( 1 ) "*"    "*"       " "        " "              " "
## 4  ( 1 ) "*"    "*"       " "        " "              "*"
## 5  ( 1 ) "*"    "*"       " "        " "              "*"
## 6  ( 1 ) "*"    "*"       " "        "*"              "*"
## 7  ( 1 ) "*"    "*"       "*"        "*"              "*"
## 8  ( 1 ) "*"    "*"       "*"        "*"              "*"
```

```
##          underwriterRating logRevenue DebtPaymentDummy logMarket
## 1  ( 1 ) " "               " "        " "              " "
## 2  ( 1 ) " "               " "        " "              " "
## 3  ( 1 ) " "               " "        " "              "*"
## 4  ( 1 ) " "               " "        " "              "*"
## 5  ( 1 ) " "               " "        "*"              "*"
## 6  ( 1 ) " "               " "        "*"              "*"
## 7  ( 1 ) " "               " "        "*"              "*"
## 8  ( 1 ) " "               "*"        "*"              "*"
##          NASDAQ30dayReturn logNumIPOS
## 1  ( 1 ) " "               " "
## 2  ( 1 ) " "               " "
## 3  ( 1 ) " "               " "
## 4  ( 1 ) " "               " "
## 5  ( 1 ) " "               " "
## 6  ( 1 ) " "               " "
## 7  ( 1 ) " "               " "
## 8  ( 1 ) " "               " "
```

Looking at the summary() object, the set of 4 predictors will contain the following features: logAST, debtRatio, VentureDummy, logMarket

(c) Run an ANOVA analysis to compare the full model in (a) and the model you obtained in (b). Which one do you prefer? (Write the null hypothesis etc). Write down in detail the model you preferred, in mathematical equations.

```r
log.reg.fit2 <- glm(decision~logAST+debtRatio+VentureDummy+logMarket, family=binomial, data=df)

my_anova <- anova(log.reg.fit, log.reg.fit2, test='Chisq')
my_anova
```

```
## Analysis of Deviance Table
##
## Model 1: decision ~ logAST + debtRatio + debtRatio2 + ProfibilityRatio +
##     VentureDummy + underwriterRating + logRevenue + DebtPaymentDummy +
##     logMarket + NASDAQ30dayReturn + logNumIPOS
## Model 2: decision ~ logAST + debtRatio + VentureDummy + logMarket
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       506     356.33
## 2       513     448.91 -7  -92.585 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

H0: All the beta coefficients for the features: logAST,debtRatio,VentureDummy,logMarket are 0 H1: There exists at least one beta coefficient for the features mentioned in H0 that is non-zero

Based on the result of the test, the p-value obtained is less than 0.05 meaning there is evidence to suggest that at least one beta coefficient among logAST,debtRatio,VentureDummy,logMarket is non-zero. Thus, the model with all the variables is a better model because it has at least one of the features missing from the reduced model $y = 1.17211 * logAST + -3.24485 * debtRatio - 56.36229 * debtRatio2 + 51.98373 * Profitability + 1.60128 * VentureDummy - 0.09156 * underwriterRating + 0.14882 * logRevenue - 1.08411 * DebtPaymentDummy - 1.31368 * logMarket + 5.56798 * NASDAQ30dayReturn - 0.31168 * logNumIPS + 19.32453$