Predicting bankruptcy in the telecommunications industry

Hailey Yim

2024-04-15

Variable Selection Analysis

Import data

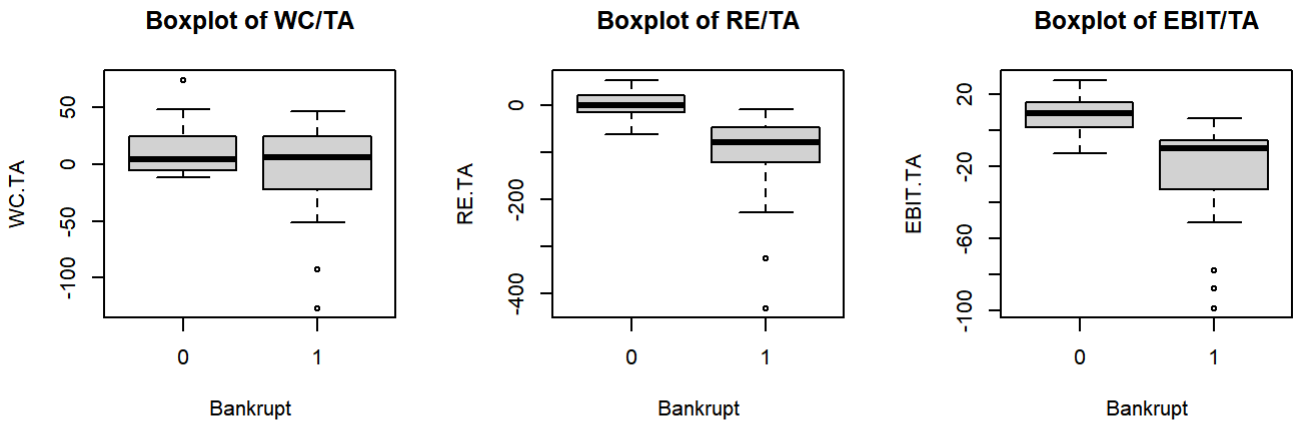
```
set.seed(1234)
bankruptcy = read.table("bankruptcy.dat",sep="\t",header=T,row.names=NULL)
attach(bankruptcy)

#remove first column (company)
bankruptcy=bankruptcy[,-1]
head(bankruptcy)
```

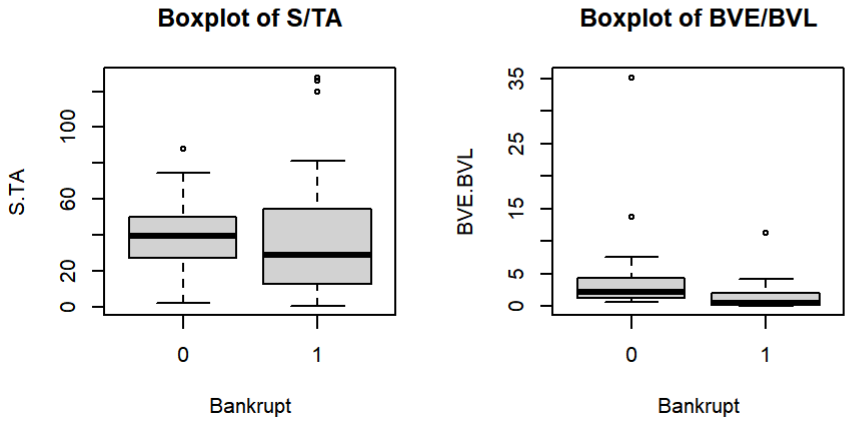
```
##   WC.TA  RE.TA EBIT.TA S.TA BVE.BVL Bankrupt
## 1   9.3   -7.7     1.6  9.1   3.726        1
## 2  42.6  -60.1   -10.1  0.3   4.130        1
## 3 -28.8 -203.2   -51.0 14.7   0.111        1
## 4   2.5 -433.1    -6.0 29.3   1.949        1
## 5  26.1  -57.4   -23.5 54.2   0.855        1
## 6  39.2 -111.8   -77.8 10.5   0.168        1
```

Boxplots and correlation

```
par(mfrow=c(2,3))
boxplot(split(WC.TA,Bankrupt),style.bxp="old",xlab="Bankrupt",ylab="WC.TA",main="Boxplot of WC/TA")
boxplot(split(RE.TA,Bankrupt),style.bxp="old",xlab="Bankrupt",ylab="RE.TA",main="Boxplot of RE/TA")
boxplot(split(EBIT.TA,Bankrupt),style.bxp="old",xlab="Bankrupt",ylab="EBIT.TA",main="Boxplot of EBIT/TA")
boxplot(split(S.TA,Bankrupt),style.bxp="old",xlab="Bankrupt",ylab="S.TA",main="Boxplot of S/TA")
boxplot(split(BVE.BVL,Bankrupt),style.bxp="old",xlab="Bankrupt",ylab="BVE.BVL",main="Boxplot of BVE/BVL")

cor(bankruptcy)
```

```
##                WC.TA       RE.TA     EBIT.TA        S.TA     BVE.BVL
## WC.TA     1.00000000  0.31506639 -0.00409731 -0.16413893  0.11369197
## RE.TA     0.31506639  1.00000000  0.54856602  0.13337470  0.06595416
## EBIT.TA  -0.00409731  0.54856602  1.00000000  0.08407904  0.07058227
## S.TA     -0.16413893  0.13337470  0.08407904  1.00000000 -0.05426599
## BVE.BVL   0.11369197  0.06595416  0.07058227 -0.05426599  1.00000000
## Bankrupt -0.24165361 -0.59877718 -0.59937756  0.01669244 -0.27903476
##             Bankrupt
## WC.TA    -0.24165361
## RE.TA    -0.59877718
## EBIT.TA  -0.59937756
## S.TA      0.01669244
## BVE.BVL  -0.27903476
## Bankrupt  1.00000000
```





Correlation coefficients can range from -1 to 1.

- Values close to 1: Strong positive linear relationship.

- Values close to -1: Strong negative linear relationship.

- Values close to 0: No linear relationship

Correlation between variables:

- The variables “WC.TA”, “RE.TA”, “EBIT.TA”, and “BVE.BVL” all have a strong negative linear relationship between these variables and “Bankrupt”.
- The correlation coefficient for the “S.TA” variable is 0.016, which is very small. This suggests that there may be a weak or no linear relationship between “S.TA” and “Bankrupt”.

Fit full model vs reduced model (before outlier removal) for AIC, BIC comparison

- **Deviance Residuals:** measure of model fit (difference between the observed response and the fitted values from the model)
- **Null Deviance:** the deviance of the null model, which is the model with no predictor variables.
- **Residual Deviance:** the deviance of the fitted model. A lower residual deviance indicates a better fit to the data.
- **AIC:** a measure of the model's goodness of fit that penalizes model complexity. Lower AIC values indicate better fitting models.

```
set.seed(1234)

#Full
bank1 = glm(Bankrupt ~ ., data=bankruptcy,family='binomial',maxit=500)
summary(bank1)
```

```
##
## Call:
## glm(formula = Bankrupt ~ ., family = "binomial", data = bankruptcy,
##     maxit = 500)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.14853  -0.06248   0.00000   0.00039   2.35010
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.42646    6.35770   1.168    0.243
## WC.TA       -0.15587    0.12208  -1.277    0.202
## RE.TA       -0.07605    0.06311  -1.205    0.228
## EBIT.TA     -0.49111    0.32260  -1.522    0.128
## S.TA        -0.08040    0.09216  -0.872    0.383
## BVE.BVL     -2.07764    1.47488  -1.409    0.159
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 69.315  on 49  degrees of freedom
## Residual deviance: 11.847  on 44  degrees of freedom
## AIC: 23.847
##
## Number of Fisher Scoring iterations: 11
```

```
#Reduced
bank2 = glm(Bankrupt ~.-WC.TA, data=bankruptcy,family='binomial',maxit=500)
```

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

```
summary(bank2)
```

```
##
## Call:
## glm(formula = Bankrupt ~ . - WC.TA, family = "binomial", data = bankruptcy,
##     maxit = 500)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.79310  -0.16838  -0.00002   0.06636   2.20656
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.16064    1.72749   0.093   0.9259
## RE.TA       -0.05634    0.02801  -2.011   0.0443 *
## EBIT.TA     -0.17244    0.09935  -1.736   0.0826 .
## S.TA        -0.01092    0.03040  -0.359   0.7195
## BVE.BVL     -0.68759    0.44972  -1.529   0.1263
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 69.315  on 49  degrees of freedom
## Residual deviance: 15.757  on 45  degrees of freedom
## AIC: 25.757
##
## Number of Fisher Scoring iterations: 9
```

Compare AIC and BIC for full and reduced model

Lower AIC and BIC values indicate better fitting models.

```
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
#Full AIC and BIC (GOF)
c(AIC(bank1),BIC(bank1))
```

```
## [1] 23.84673 35.31887
```

```
#Reduced model AIC and BIC (GOF)
c(AIC(bank2), BIC(bank2))
```

```
## [1] 25.75660 35.31672
```

Reduced model has higher AIC and BIC values compared to the full model suggest that the full logistic regression model appears to provide a better fit to the data compared to the reduced model.

Subselection AIC and BIC before outlier removal

```
#Bestglm institutes the leaps algorithm but for GLM while leaps() library is only for linear models

set.seed(1234)

library(bestglm)
```

```
## Loading required package: leaps
```

```
#By BIC
bestBIC <- bestglm(bankruptcy, IC="BIC",family=binomial)
```

```
## Morgan-Tatar search since family is non-gaussian.
```

```
bestBIC
```

```
## BIC
## BICq equivalent for q in (0.418159458690708, 0.615587894177641)
## Best Model:
##                Estimate Std. Error    z value   Pr(>|z|)
## (Intercept) -0.29477828 1.12317114 -0.2624518 0.79297315
## RE.TA       -0.05626758 0.02744876 -2.0499132 0.04037290
## EBIT.TA     -0.16763140 0.09269427 -1.8084333 0.07053909
## BVE.BVL     -0.62974558 0.39429324 -1.5971503 0.11023221
```

- The output indicate that the best model is selected based on the BIC criterion.
- The selected model includes the intercept, ‘RE.TA’, ‘EBIT.TA’, and ‘BVE.BVL’.
- The ‘BICq equivalent’ line provides the range of BIC values for which the selected model is within a specified tolerance.

```
# Model with selected variables based on the BIC criterion
bank3= glm(Bankrupt ~ RE.TA+EBIT.TA+BVE.BVL, data=bankruptcy,family='binomial',maxit=500)
```

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

```
summary(bank3)
```

```
##
## Call:
## glm(formula = Bankrupt ~ RE.TA + EBIT.TA + BVE.BVL, family = "binomial",
##     data = bankruptcy, maxit = 500)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.74542  -0.18840  -0.00005   0.05942   2.26136
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.29478    1.12317  -0.262   0.7930
## RE.TA       -0.05627    0.02745  -2.050   0.0404 *
## EBIT.TA     -0.16763    0.09269  -1.808   0.0705 .
## BVE.BVL     -0.62975    0.39429  -1.597   0.1102
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 69.315  on 49  degrees of freedom
## Residual deviance: 15.887  on 46  degrees of freedom
## AIC: 23.887
##
## Number of Fisher Scoring iterations: 8
```

- Intercept (-0.2948): the estimated log odds of bankruptcy when all predictor variables are zero.
- 1 unit increase in RE.TA is associated with a decrease in the log odds of bankruptcy by -0.05627 units, holding other variables constant.

```
#By AIC
bestAIC <- bestglm(bankruptcy, IC="AIC",family=binomial)
```

```
## Morgan-Tatar search since family is non-gaussian.
```

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

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

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

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

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

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

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

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

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

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

```
bestAIC
```

```
## AIC
## BICq equivalent for q in (0.615587894177641, 0.805493952518934)
## Best Model:
##                Estimate Std. Error   z value   Pr(>|z|)
## (Intercept)  3.01795904 2.92521544  1.031705 0.30221039
## WC.TA       -0.11589056 0.08544332 -1.356344 0.17498965
## RE.TA       -0.06327245 0.03924904 -1.612076 0.10694537
## EBIT.TA     -0.39890439 0.24209641 -1.647709 0.09941243
## BVE.BVL     -1.18574532 0.82430652 -1.438476 0.15029898
```

- The output indicate that the best model is selected based on the AIC criterion.
- The selected model includes the intercept, 'WC.TA', 'RE.TA', 'EBIT.TA', and 'BVE.BVL'.

```
# Model with selected variables based on the AIC criterion
bankAIC1= glm(Bankrupt ~ WC.TA+RE.TA+EBIT.TA+BVE.BVL, data=bankruptcy,family='binomial', maxit=500)
```

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

```
summary(bankAIC1)
```

```
##
## Call:
## glm(formula = Bankrupt ~ WC.TA + RE.TA + EBIT.TA + BVE.BVL, family = "binomial",
##     data = bankruptcy, maxit = 500)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.28581  -0.13824   0.00000   0.00073   2.32320
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.01796    2.92522   1.032   0.3022
## WC.TA       -0.11589    0.08544  -1.356   0.1750
## RE.TA       -0.06327    0.03925  -1.612   0.1069
## EBIT.TA     -0.39890    0.24210  -1.648   0.0994 .
## BVE.BVL     -1.18575    0.82431  -1.438   0.1503
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 69.315  on 49  degrees of freedom
## Residual deviance: 12.917  on 45  degrees of freedom
## AIC: 22.917
##
## Number of Fisher Scoring iterations: 10
```

The AIC value for the selected model is 22.917, which indicates the goodness of fit of the model relative to other models considered.

```
# Testing for subset of regression coefficients

gstat = deviance(bank3) - deviance(bank1)
cbind(gstat, 1-pchisq(gstat,length(coef(bank1))-length(coef(bank3))))
```

```
##         gstat
## [1,] 4.040336 0.1326332
```

- First value (gstat): the difference in deviance between the full model ('bank1') and the reduced model ('bank3'). bank1 provides a better fit.
- Second value (p-value): p-value associated with the likelihood ratio test using the chi-squared distribution.

Since the p-value (0.1326) is greater than 0.05, we fail to reject the null hypothesis at a significance level of 0.05. This suggests that there is not enough evidence to conclude that the full model significantly improves the fit compared to the reduced model (BIC criterion). In other words, excluding the predictors in 'bank3' does not significantly decrease the model's fit compared to the full model 'bank1'.

[Calculate and comparing AIC and BIC values for model selection]

```
#Full AIC and BIC
c(AIC(bank1),BIC(bank1))
```

```
## [1] 23.84673 35.31887
```

```
#Best subset AIC
c(AIC(bank3), BIC(bank3))
```

```
## [1] 23.88707 31.53516
```

```
#Best subset BIC
c(AIC(bankAIC1), BIC(bankAIC1))
```

```
## [1] 22.91677 32.47689
```

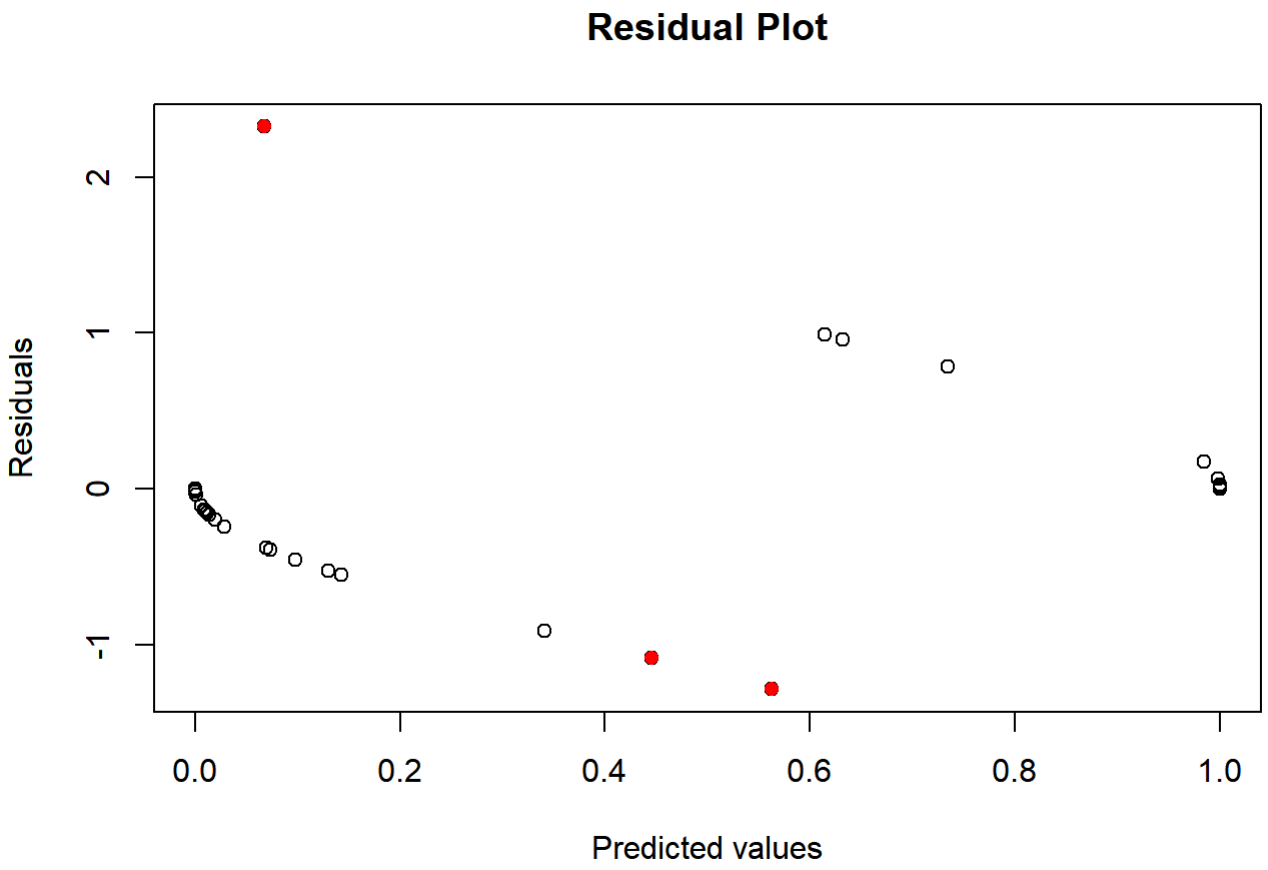
The result suggest that the best subset model ('bankAIC1') provides the best trade-off between goodness of fit and model complexity, as it has the lowest AIC and BIC values among the models considered.

Outlier removal

```
# Calculate residuals
residuals <- residuals(bankAIC1)

# Optionally standardize residuals
std_residuals <- rstandard(bankAIC1)

# Plot residuals against predicted values
plot(bankAIC1$fitted.values, residuals,
     xlab = "Predicted values", ylab = "Residuals",
     main = "Residual Plot")

# Identify potential outliers
outliers <- which(abs(residuals) > 2 * sd(residuals)) # Adjust the multiplier as needed


# Print indices of potential outliers
print(outliers)
```

```
##  1 27 28
##  1 27 28
```

```
# Optionally, visualize outliers on the plot
points(fitted(bankAIC1)[outliers], residuals[outliers], col = "red", pch = 16)
```




```
# Investigate outliers further by examining corresponding data points
bankruptcy[outliers, ]
```

```
##    WC.TA RE.TA EBIT.TA S.TA BVE.BVL Bankrupt
## 1    9.3  -7.7     1.6  9.1   3.726        1
## 27   9.8 -33.8    -7.1  3.2   5.965        0
## 28  37.8 -45.4    -7.1 17.1   3.450        0
```

```r
set.seed(1234)

#Remove one outlier
bank_cleaned = bankruptcy[-1,]
detach(bankruptcy)
attach(bank_cleaned)


bank_new_full=glm(Bankrupt ~ ., data=bank_cleaned,family='binomial',maxit=500)
summary(bank_new_full)
```

```
##
## Call:
## glm(formula = Bankrupt ~ ., family = "binomial", data = bank_cleaned,
##     maxit = 500)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -1.104e-05  -2.110e-08  -2.110e-08   2.110e-08   7.378e-06
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    265.467 576281.709       0        1
## WC.TA           -4.297  12439.717       0        1
## RE.TA           -1.516   5131.146       0        1
## EBIT.TA        -17.043  35543.170       0        1
## S.TA            -2.859   7408.747       0        1
## BVE.BVL        -77.540 184903.001       0        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6.7908e+01  on 48  degrees of freedom
## Residual deviance: 3.6353e-10  on 43  degrees of freedom
## AIC: 12
##
## Number of Fisher Scoring iterations: 30
```

Test overall regression

The null hypothesis is that the full model does not provide a significantly better fit than than null model.

```
gstat = bank_new_full$null.deviance - deviance(bank_new_full)
cbind(gstat, 1-pchisq(gstat,length(coef(bank_new_full))-1))
```

```
##         gstat
## [1,] 67.90801 2.791101e-13
```

Since the p-value is extremely small (< 0.05), we reject the null hypothesis.

Therefore, we conclude that the full model provides a significantly better fit to the data compared to the null model.

Search for best model, AIC and BIC using bestglm.

```
set.seed(1234)
library(bestglm)

#By AIC
bestAIC <- bestglm(bank_cleaned, IC="AIC",family=binomial)
```

```
## Morgan-Tatar search since family is non-gaussian.
```

```
bestAIC
```

```
## AIC
## BICq equivalent for q in (0.144718876531243, 0.874999994439966)
## Best Model:
##               Estimate Std. Error      z value  Pr(>|z|)
## (Intercept) 193.528527 40982.6225  0.004722209 0.9962322
## WC.TA        -7.173193  1333.6792 -0.005378499 0.9957086
## RE.TA        -3.811883   843.1096 -0.004521219 0.9963926
## EBIT.TA     -21.993992  5138.6111 -0.004280143 0.9965850
## BVE.BVL     -77.898238 11828.3530 -0.006585721 0.9947454
```

```
#By BIC
bestBIC <- bestglm(bank_cleaned, IC="BIC",family=binomial)
```

```
## Morgan-Tatar search since family is non-gaussian.
```

```
bestBIC
```

```
## BIC
## BICq equivalent for q in (0.144718876531243, 0.874999994439966)
## Best Model:
##               Estimate Std. Error      z value  Pr(>|z|)
## (Intercept) 193.528527 40982.6225  0.004722209 0.9962322
## WC.TA        -7.173193  1333.6792 -0.005378499 0.9957086
## RE.TA        -3.811883   843.1096 -0.004521219 0.9963926
## EBIT.TA     -21.993992  5138.6111 -0.004280143 0.9965850
## BVE.BVL     -77.898238 11828.3530 -0.006585721 0.9947454
```

```
#With outlier removed they select the same variables
bank4= glm(Bankrupt ~ RE.TA+EBIT.TA+BVE.BVL,data=bank_cleaned,family='binomial', maxit=500)
summary(bank4)
```

```
##
## Call:
## glm(formula = Bankrupt ~ RE.TA + EBIT.TA + BVE.BVL, family = "binomial",
##     data = bank_cleaned, maxit = 500)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.77491  -0.06042   0.00000   0.00943   1.84492
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.09166    1.47135  -0.062   0.9503
## RE.TA       -0.08229    0.04230  -1.945   0.0517 .
## EBIT.TA     -0.26783    0.15854  -1.689   0.0912 .
## BVE.BVL     -1.21810    0.76536  -1.592   0.1115
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 67.9080  on 48  degrees of freedom
## Residual deviance:  9.3841  on 45  degrees of freedom
## AIC: 17.384
##
## Number of Fisher Scoring iterations: 10
```

```
exp(coef(bank3)[-1])
```

```
##     RE.TA   EBIT.TA   BVE.BVL
## 0.9452862 0.8456655 0.5327273
```

- RE.TA: A one-unit increase in RE.TA, the odds of the outcome (Bankrupt) decreases by approximately 5.5% (0.945-1).
- EBIT.TA: A one-unit increase in EBIT.TA, the odds of the outcome (Bankrupt) decreases by approximately 15.4% (0.846-1).
- BVE.BVL: A one-unit increase in BVE.BVL, the odds of the outcome (Bankrupt) decreases by approximately 46.7% (0.533-1).

```
exp(coef(bank4)[-1])
```

```
##     RE.TA   EBIT.TA   BVE.BVL
## 0.9210091 0.7650371 0.2957930
```

- RE.TA: A one-unit increase in RE.TA, the odds of the outcome (Bankrupt) decreases by approximately 7.9% (0.921-1).
- EBIT.TA: A one-unit increase in EBIT.TA, the odds of the outcome (Bankrupt) decreases by approximately 23.5% (0.765-1).
- BVE.BVL: A one-unit increase in BVE.BVL, the odds of the outcome (Bankrupt) decreases by approximately 70.4% (0.296-1).

In both models, an increase in EBIT.TA and BVE.BVL is associated with lower odds of bankruptcy. This suggests that companies with higher earnings relative to their total assets are less likely to go bankrupt.

The effect of RE.TA is less consistent across the two models. It seems that the relationship between RE.TA and the likelihood of bankruptcy may vary or be less clear compared to the other variables.

Apply Stepwise Regression

Forward

```
full = glm(Bankrupt ~ ., data=bank_cleaned,family='binomial'(link = "logit"),maxit=500)
minimum = glm(Bankrupt ~ 1, data=bank_cleaned,family='binomial'(link = "logit"), maxit=500)
n= nrow(bank_cleaned)
#AIC
fwd_AIC<-step(minimum,
     scope = list(lower=minimum,upper = full),
     direction = "forward", trace=F)
fwd_AIC
```

```
##
## Call:  glm(formula = Bankrupt ~ RE.TA + BVE.BVL + EBIT.TA + WC.TA, family = binomial(link = "logit"),
##     data = bank_cleaned, maxit = 500)
##
## Coefficients:
## (Intercept)        RE.TA      BVE.BVL      EBIT.TA        WC.TA
##     255.413       -5.152     -103.614      -28.983       -9.542
##
## Degrees of Freedom: 48 Total (i.e. Null);  44 Residual
## Null Deviance:       67.91
## Residual Deviance: 4.249e-10     AIC: 10
```

```
#BIC
fwd_BIC<-step(minimum,
     scope = list(lower=minimum,upper = full),
     direction = "forward", trace=F, k=log(n))
fwd_BIC
```

```
##
## Call:  glm(formula = Bankrupt ~ RE.TA + BVE.BVL + EBIT.TA + WC.TA, family = binomial(link = "logit"),
##     data = bank_cleaned, maxit = 500)
##
## Coefficients:
## (Intercept)        RE.TA      BVE.BVL      EBIT.TA        WC.TA
##     255.413       -5.152     -103.614      -28.983       -9.542
##
## Degrees of Freedom: 48 Total (i.e. Null);  44 Residual
## Null Deviance:       67.91
## Residual Deviance: 4.249e-10     AIC: 10
```

Both AIC and BIC select the same: RE.TA, BVE.BVL, EBIT.TA, WC.TA

Backward

```
#backwards via AIC
bck_AIC<-step(full, scope = list(lower=minimum, upper = full), direction = "backward", trace = F)
bck_AIC
```

```
##
## Call:  glm(formula = Bankrupt ~ WC.TA + RE.TA + EBIT.TA + BVE.BVL, family = binomial(link = "logit"),
##     data = bank_cleaned, maxit = 500)
##
## Coefficients:
## (Intercept)        WC.TA        RE.TA      EBIT.TA      BVE.BVL
##     255.413       -9.542       -5.152      -28.983     -103.614
##
## Degrees of Freedom: 48 Total (i.e. Null);  44 Residual
## Null Deviance:       67.91
## Residual Deviance: 4.249e-10     AIC: 10
```

```
#Backwards via BIC
bck_BIC<-step(full, scope = list(lower=minimum, upper = full), direction = "backward", trace = F, k=log(n))
bck_BIC
```

```
##
## Call:  glm(formula = Bankrupt ~ WC.TA + RE.TA + EBIT.TA + BVE.BVL, family = binomial(link = "logit"),
##     data = bank_cleaned, maxit = 500)
##
## Coefficients:
## (Intercept)        WC.TA        RE.TA      EBIT.TA      BVE.BVL
##     255.413       -9.542       -5.152      -28.983     -103.614
##
## Degrees of Freedom: 48 Total (i.e. Null);  44 Residual
## Null Deviance:       67.91
## Residual Deviance: 4.249e-10     AIC: 10
```

Backwards stepwsise regression selected the same 4: WC.TA, RE.TA, EBIT.TA, BVE.BVL

Both Forward and Backward

```
set.seed(1234)
bth_AIC<-step(minimum, scope = list(lower=minimum, upper = full), direction = "both", trace=F)
bth2_AIC<-step(full, scope = list(lower=minimum, upper = full), direction = "both", trace = F)

bth_BIC<-step(minimum, scope = list(lower=minimum, upper = full), direction = "both", trace=F,k=log(n))
bth2_BIC<-step(full, scope = list(lower=minimum, upper = full), direction = "both", trace = F,k=log(n))

bth_AIC
```

```
##
## Call:  glm(formula = Bankrupt ~ RE.TA + BVE.BVL + EBIT.TA + WC.TA, family = binomial(link = "logit"),
##     data = bank_cleaned, maxit = 500)
##
## Coefficients:
## (Intercept)        RE.TA      BVE.BVL      EBIT.TA        WC.TA
##     255.413       -5.152     -103.614      -28.983       -9.542
##
## Degrees of Freedom: 48 Total (i.e. Null);  44 Residual
## Null Deviance:       67.91
## Residual Deviance: 4.249e-10     AIC: 10
```

```
bth_BIC
```

```
##
## Call:  glm(formula = Bankrupt ~ RE.TA + BVE.BVL + EBIT.TA + WC.TA, family = binomial(link = "logit"),
##     data = bank_cleaned, maxit = 500)
##
## Coefficients:
## (Intercept)        RE.TA      BVE.BVL      EBIT.TA        WC.TA
##     255.413       -5.152     -103.614      -28.983       -9.542
##
## Degrees of Freedom: 48 Total (i.e. Null);  44 Residual
## Null Deviance:       67.91
## Residual Deviance: 4.249e-10     AIC: 10
```

```
bank6=glm(Bankrupt ~ WC.TA+RE.TA+EBIT.TA+BVE.BVL, data=bank_cleaned,family='binomial'(link = "logit"), maxit=500)
summary(bank6)
```

```
##
## Call:
## glm(formula = Bankrupt ~ WC.TA + RE.TA + EBIT.TA + BVE.BVL, family = binomial(link = "logit"),
##     data = bank_cleaned, maxit = 500)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -1.399e-05  -2.110e-08  -2.110e-08   2.110e-08   1.183e-05
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    255.413 728539.823       0        1
## WC.TA           -9.542  23920.936       0        1
## RE.TA           -5.152  15669.825       0        1
## EBIT.TA        -28.983  90578.211       0        1
## BVE.BVL       -103.614 225264.760       0        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6.7908e+01  on 48  degrees of freedom
## Residual deviance: 4.2489e-10  on 44  degrees of freedom
## AIC: 10
##
## Number of Fisher Scoring iterations: 31
```

WC.TA, RE.TA, EBIT.TA, BVE.BVL

once again the same four are selected. Bank 3 is this model

Regularized Regression

scale variables

```
set.seed(1234)
bank_var<-scale(bank_cleaned[,-6])
```

Ridge Regression

```
library(glmnet)
```

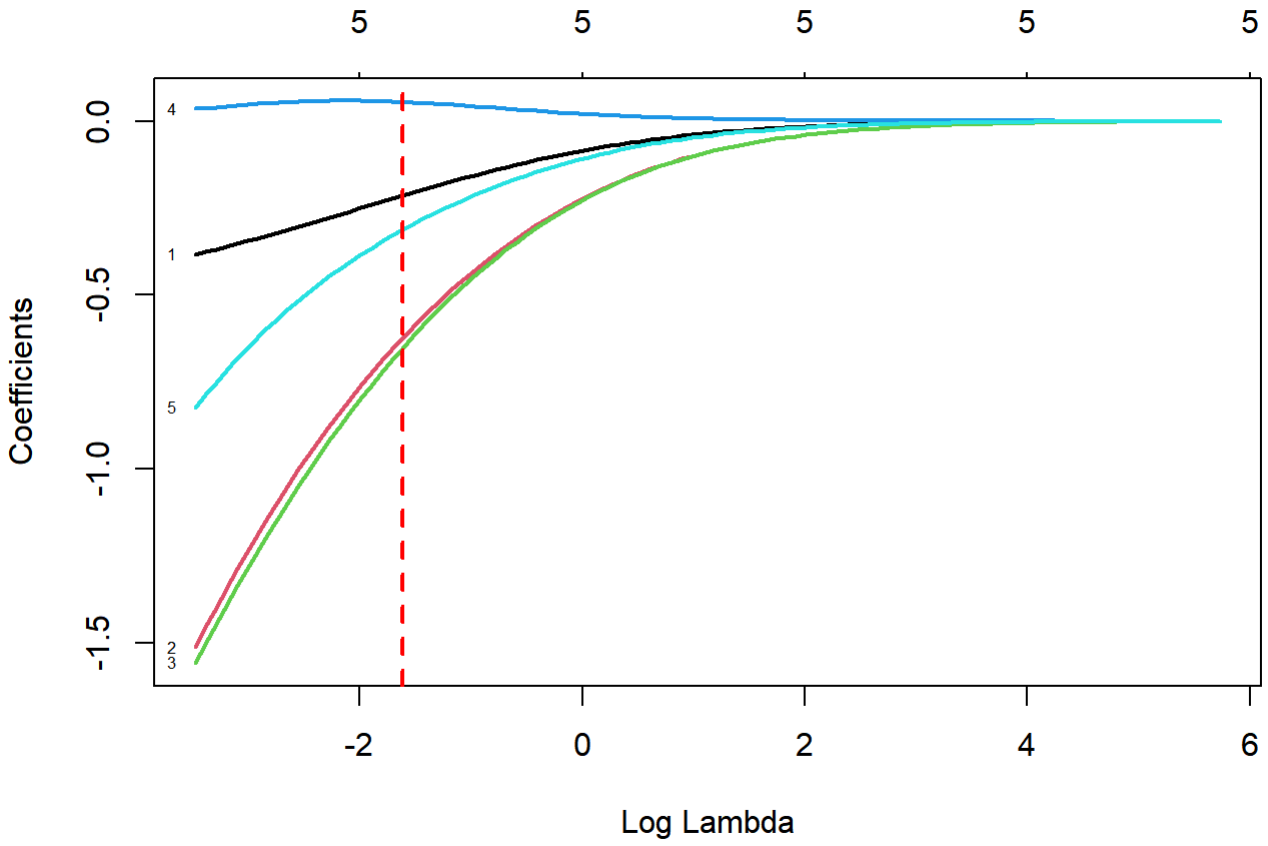
```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-6
```

```
set.seed(49)


# Code to conduct ridge regression, and find optimal lambda
# the regularization parameter that minimizes the cross-validated error in the ridge regression model.
cv.ridge = cv.glmnet(bank_var, Bankrupt,family='binomial', alpha=0,type.measure ='class', nfolds=10)

cv.ridge$lambda.min
```

```
## [1] 0.1982858
```

```
ridge.mod = glmnet(bank_var, Bankrupt,family='binomial',alpha=0, nlambda =100)

#plot the ridge coef  path
# a visual representation of how the coefficients change in response to different levels of regularization (lambda is the am
ount of regularization applied to the coefficients). The lambda is a tuning parameter that determines the strength of regula
rization (penalty) -> more shrinkage of coefficients towards zero.
plot(ridge.mod, xvar = "lambda", label = TRUE, lwd = 2)
abline(v=log(cv.ridge$lambda.min),col='red',lty = 2,lwd=2)
```




```
coef(ridge.mod, s = cv.ridge$lambda.min)
```

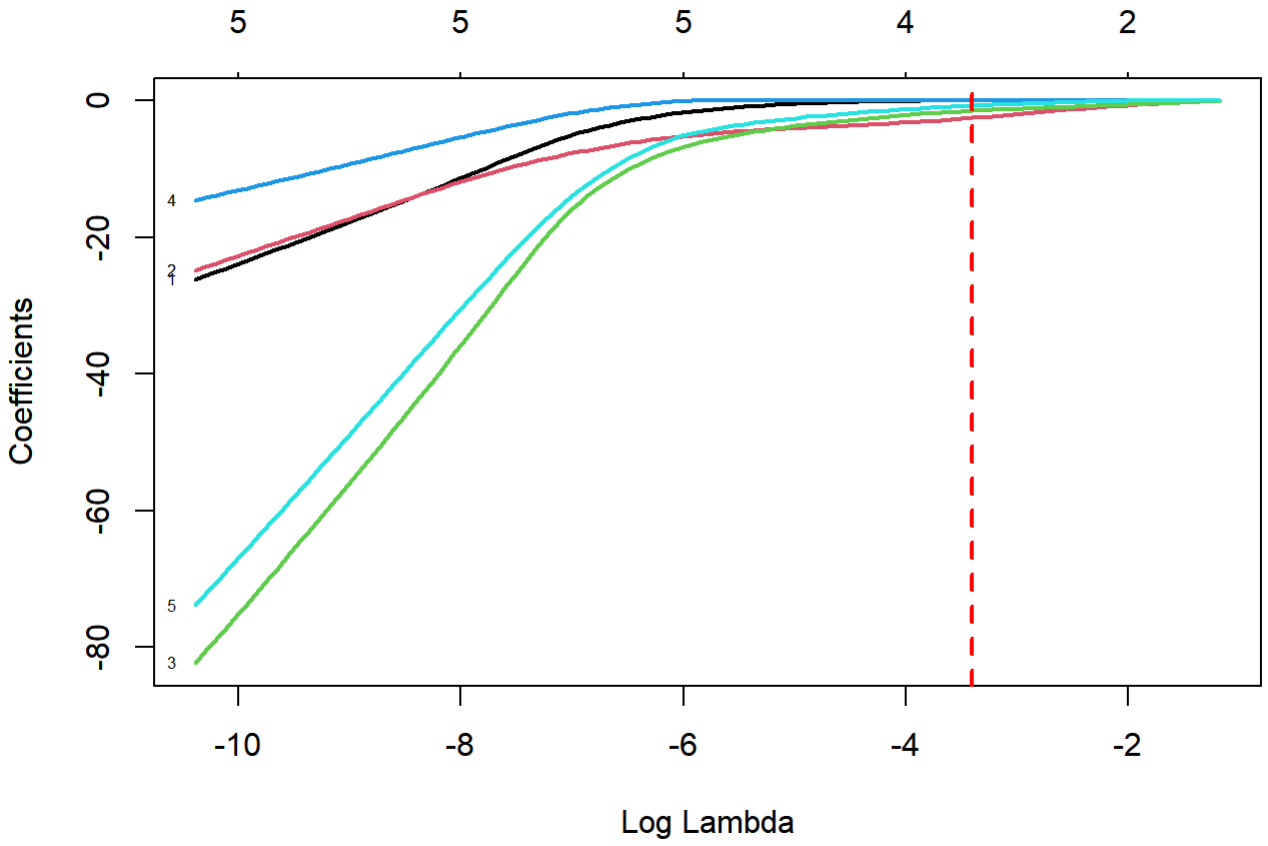
```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##                      s1
## (Intercept)  0.03863309
## WC.TA       -0.21419400
## RE.TA       -0.62421022
## EBIT.TA     -0.65363519
## S.TA         0.05457429
## BVE.BVL     -0.31308629
```

Lasso Regression

```
set.seed(49)

# Code to conduct lasso regression, and find optimal lambda
cv.lasso = cv.glmnet(bank_var, Bankrupt,family='binomial', alpha=1,type.measure ='class', nfolds=10)

cv.lasso$lambda.min
```

```
## [1] 0.03307606
```

```
lasso.mod = glmnet(bank_var,Bankrupt,family='binomial',alpha=1, nlambda=100)

#plot the lasso coef  path
plot(lasso.mod, xvar = "lambda", label = TRUE, lwd = 2)
abline(v=log(cv.lasso$lambda.min),col='red',lty = 2,lwd=2)
```




```
coef(lasso.mod, s = cv.lasso$lambda.min)
```

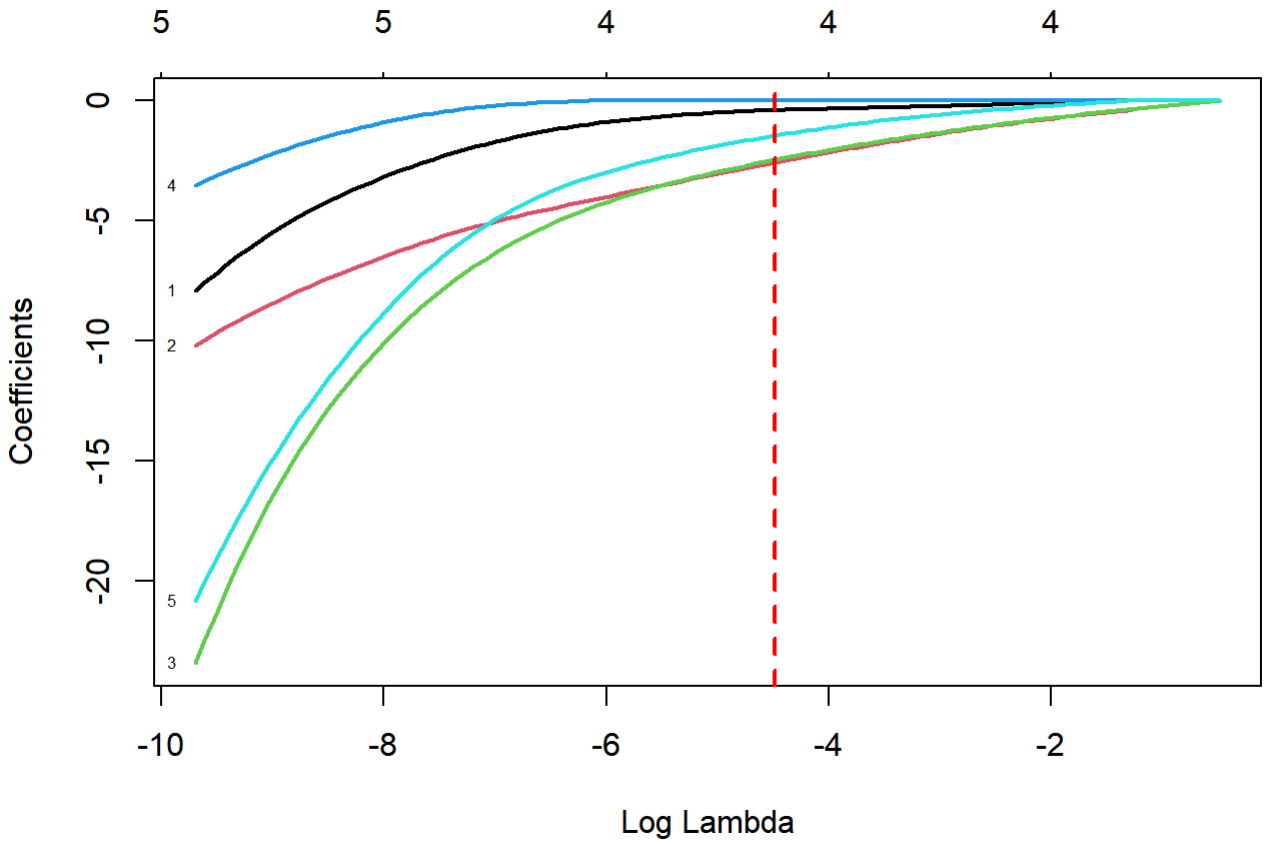
```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##                     s1
## (Intercept)  0.5728627
## WC.TA         .
## RE.TA        -2.4870902
## EBIT.TA      -1.4667460
## S.TA          .
## BVE.BVL      -0.7025255
```

Elastic Net Regression using GLMNET

```
set.seed(49)

# Code to conduct elastic net regression, and find optimal lamb

cv.elastic = cv.glmnet(bank_var,  Bankrupt,family='binomial', alpha=0.5,type.measure ='class', nfolds=10)

cv.elastic$lambda.min
```

```
## [1] 0.0112945
```

```
elastic.mod = glmnet(bank_var,  Bankrupt,family='binomial',alpha=0.5, nlambda=100)

#plot the elastic net coef  path
plot(elastic.mod, xvar = "lambda", label = TRUE, lwd = 2)
abline(v=log(cv.elastic$lambda.min),col='red',lty = 2,lwd=2)
```




```
coef(elastic.mod, s = cv.elastic$lambda.min)
```

```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##                      s1
## (Intercept)  0.8168453
## WC.TA       -0.3861189
## RE.TA       -2.5655168
## EBIT.TA     -2.4554504
## S.TA         .
## BVE.BVL     -1.4486946
```