Model Selection and Prediction

Kanishk Deshwal

2025-12-06

```
companies <- read.csv("indian_companies_transformed.csv", skipNul = TRUE)
data = companies[, c("logREVIEWS", "logRATING", "logBRANCHES", "YEARS.OLD", "INDUSTR
Y", "INDIA.HQ", "TOTAL_EMPLOYEES", "logPACKAGE")]
data$INDUSTRY = factor(data$INDUSTRY)
data$INDIA.HQ = factor(data$INDIA.HQ)
```

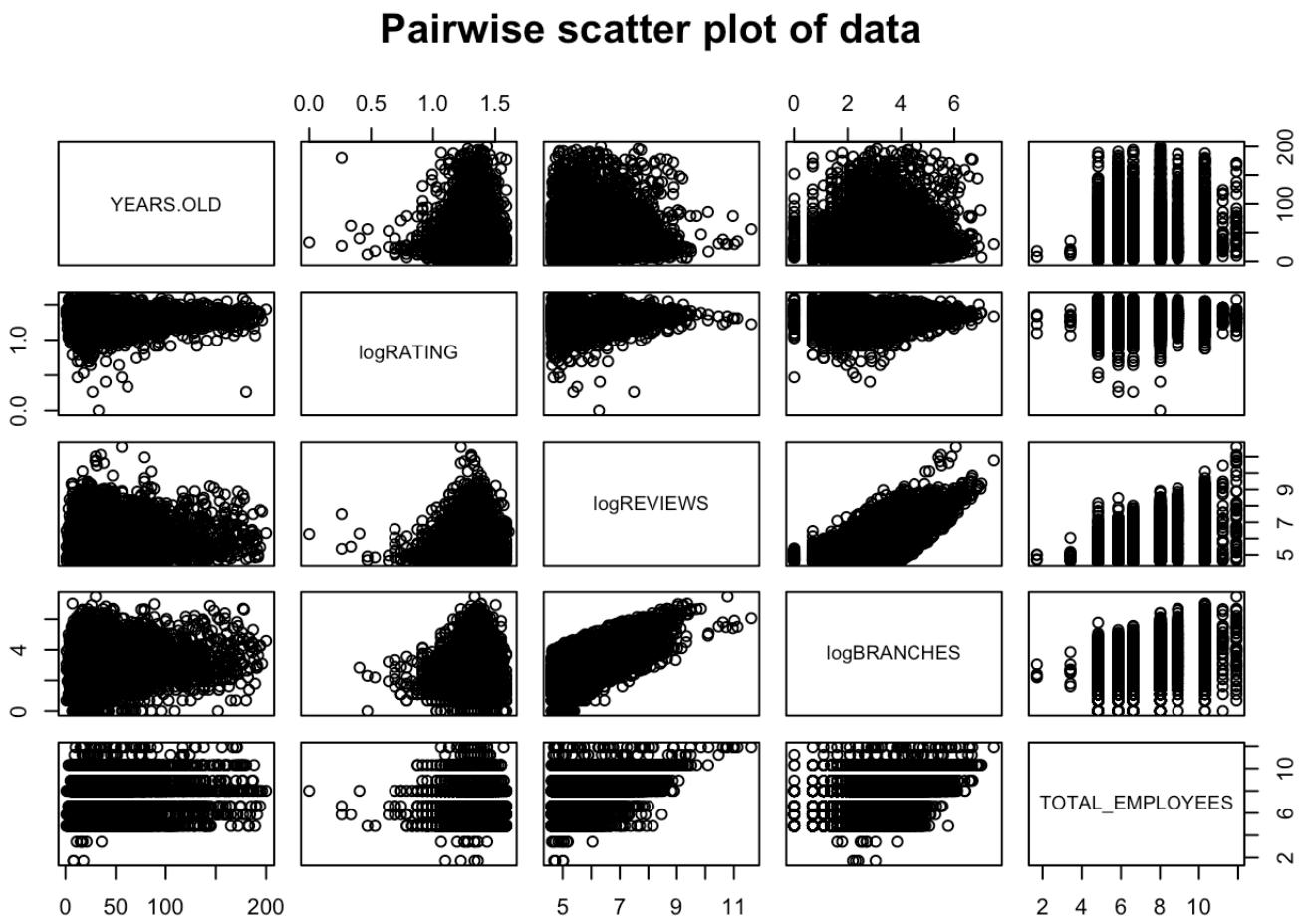
```
lmod = lm(logPACKAGE ~ ., data)
summary(lmod)
```

```
##
## Call:
## lm(formula = logPACKAGE ~ ., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8139 -0.1705  0.0447  0.2431  2.6175
##
## Coefficients:
##                                    Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                       -8.307e+00  7.907e-02 -105.063  < 2e-16 ***
## logREVIEWS                         1.016e+00  9.010e-03  112.709  < 2e-16 ***
## logRATING                         -1.228e+00  4.948e-02  -24.824  < 2e-16 ***
## logBRANCHES                       -3.814e-02  7.058e-03   -5.404 6.70e-08 ***
## YEARS.OLD                         -2.604e-05  1.779e-04   -0.146 0.883633
## INDUSTRYEducation & Training      -3.607e-01  4.028e-02   -8.954  < 2e-16 ***
## INDUSTRYEngineering & Construction -1.316e-01  3.438e-02   -3.827 0.000131 ***
## INDUSTRYFinancial Services         4.771e-02  3.926e-02    1.215 0.224312
## INDUSTRYHealthcare                -1.690e-01  3.884e-02   -4.352 1.37e-05 ***
## INDUSTRYIndustrial Machinery       2.681e-02  3.564e-02    0.752 0.451839
## INDUSTRYInternet                  -1.298e-01  3.963e-02   -3.276 0.001059 **
## INDUSTRYIT Services & Consulting   8.068e-02  2.911e-02    2.772 0.005585 **
## INDUSTRYOther                     -9.319e-02  2.626e-02   -3.549 0.000389 ***
## INDUSTRYPharma                    -7.057e-02  3.664e-02   -1.926 0.054163 .
## INDUSTRYReal Estate               -3.158e-02  4.328e-02   -0.730 0.465559
## INDUSTRYSoftware Product           1.325e-01  3.866e-02    3.427 0.000613 ***
## INDIA.HQChennai                   -8.207e-02  2.334e-02   -3.516 0.000441 ***
## INDIA.HQMumbai                    -1.630e-01  1.961e-02   -8.312  < 2e-16 ***
## INDIA.HQNew Delhi                 -2.743e-01  2.315e-02  -11.846  < 2e-16 ***
## INDIA.HQOther                     -2.047e-01  1.603e-02  -12.770  < 2e-16 ***
## INDIA.HQPune                      -4.255e-02  2.364e-02   -1.800 0.071918 .
## TOTAL_EMPLOYEES                   -2.426e-03  4.427e-03   -0.548 0.583658
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4685 on 7754 degrees of freedom
##   (1647 observations deleted due to missingness)
## Multiple R-squared:  0.8026, Adjusted R-squared:  0.8021
## F-statistic:  1502 on 21 and 7754 DF,  p-value: < 2.2e-16
```

Checking for Inflated Variance - Stabilizing β

Now, checking for multicoliearity using correlation matrix and calculating Condition number.

```
X = model.matrix(lmod)
cat_vars = c("YEARS.OLD","logRATING","logREVIEWS","logBRANCHES","TOTAL_EMPLOYEES")
pairs(data[cat_vars], labels = cat_vars, main="Pairwise scatter plot of data")
```




```
cat("Correlation Matrix:\n")
```

```
## Correlation Matrix:
```

```
cor(X[, c(2, 19, 20, 21, 22)])
```

```
##                    logREVIEWS INDIA.HQNew Delhi INDIA.HQOther INDIA.HQPune
## logREVIEWS          1.00000000       -0.03688295   -0.07934745   0.01602448
## INDIA.HQNew Delhi -0.03688295        1.00000000   -0.27751379  -0.09103513
## INDIA.HQOther     -0.07934745       -0.27751379    1.00000000  -0.25903319
## INDIA.HQPune       0.01602448       -0.09103513   -0.25903319   1.00000000
## TOTAL_EMPLOYEES    0.59629725       -0.03301318   -0.05741815   0.02368374
##                   TOTAL_EMPLOYEES
## logREVIEWS             0.59629725
## INDIA.HQNew Delhi     -0.03301318
## INDIA.HQOther         -0.05741815
## INDIA.HQPune           0.02368374
## TOTAL_EMPLOYEES        1.00000000
```

```
n = dim(X)[1]
p = dim(X)[2]

XtX = t(X)%*%X
lambda = eigen(XtX)$values
cat("\u03BB =", lambda)
```

```
## λ = 18372215 284446.5 8146.096 3612.306 2315.124 1970.702 996.4891 756.0084 667.37
59 616.7229 514.561 373.408 310.3396 284.1492 248.7851 243.0104 240.4234 223.6927 18
3.3937 165.226 38.128 21.67154
```

```
K = sqrt(lambda[1]/lambda[p])
cat("Condition number is given by K = ", K)
```

```
## Condition number is given by K =  920.7375
```

High Condition Number states that this is an extreme case of Multi-collinearity which signals that the variance of estimators may be inflated. To further confirm this Variance Inflation Factor must be calculated.

```
library(faraway)
v = vif(lmod)
print(v)
```

```
##                          logREVIEWS                             logRATING
##                            2.507164                              1.074196
##                         logBRANCHES                            YEARS.OLD
##                            2.213046                              1.043834
##       INDUSTRYEducation & Training INDUSTRYEngineering & Construction
##                            1.615283                              2.043480
##         INDUSTRYFinancial Services                  INDUSTRYHealthcare
##                            1.672689                              1.663424
##       INDUSTRYIndustrial Machinery                    INDUSTRYInternet
##                            1.854912                              1.650638
##   INDUSTRYIT Services & Consulting                       INDUSTRYOther
##                            3.448937                              6.104873
##                       INDUSTRYPharma                 INDUSTRYReal Estate
##                            1.820872                              1.483997
##           INDUSTRYSoftware Product                     INDIA.HQChennai
##                            1.704511                              1.429235
##                       INDIA.HQMumbai                   INDIA.HQNew Delhi
##                            1.805330                              1.537580
##                        INDIA.HQOther                       INDIA.HQPune
##                            2.245357                              1.429311
##                      TOTAL_EMPLOYEES
##                            1.596947
```

```
cat("Average VIF for predictors: ", sum(v)/p)
```

```
## Average VIF for predictors:  1.906619
```

```
cat("Maximum value of VIF: ", max(v))
```

```
## Maximum value of VIF:  6.104873
```

Average and Maximum values of VIF depicts that the variance of each estimate is not inflated (atleast to a value that posses a concern). The huge condition number is due to design-matrix structure (dummy coding), not harmful collinearity in the regression coefficients. Furthermore, the pairwise correlations among predictors are all below 0.6, which is additional evidence that the independent variables do not exhibit strong linear relationships.

Model Selection

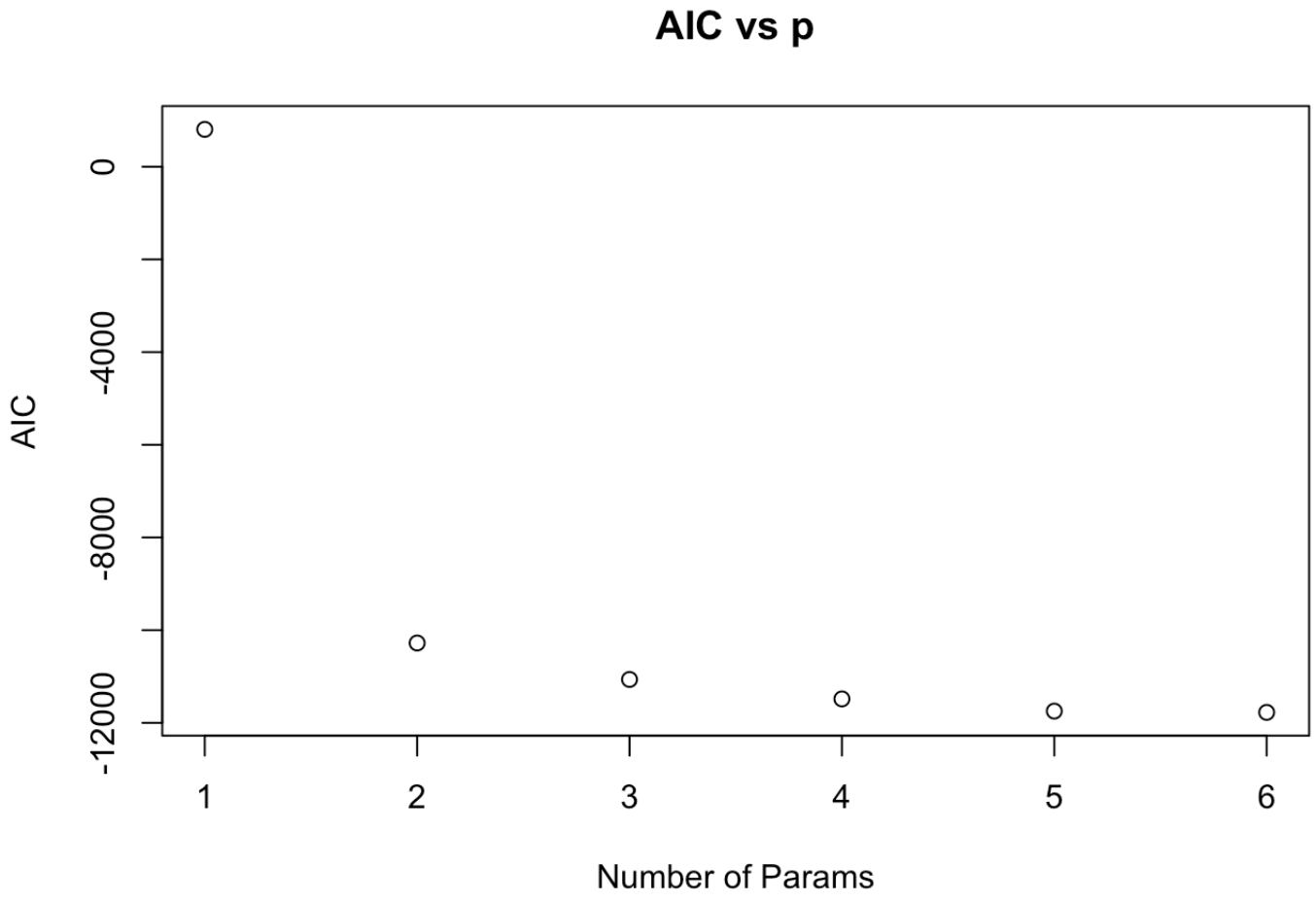
As the number of parameters p<<n, a balanced criterion is required to balance predictive power and complexity of the model. Thus, AIC is ideal criterion with Hybrid(Both) search.

```
null = lm(logPACKAGE ~ 1, na.omit(data))
full = lm(logPACKAGE ~ ., na.omit(data))
out = step(null, scope = list(lower=~1, upper=formula(full)), direction = "both", tra
ce = FALSE)
cat("Order of parameters: ",out$anova[,1])
```

```
## Order of parameters:   + logREVIEWS + logRATING + INDUSTRY + INDIA.HQ + logBRANCHE
S
```

Here, Years.OLD and TOTAL_EMPLOYEES were dropped from the model as a result.

```
plot(out$anova[,6], xlab="Number of Params", ylab="AIC", main="AIC vs p")
```



Looking at the AIC vs Index plot, we see that AIC remains almost constant after 4. Replacing AIC with more strict criterion, BIC.

```
out = step(null, scope = list(lower=~1, upper=formula(full)), direction = "both", k=l
og(n), trace = FALSE)
cat("Order of parameters:", out$anova[,1])
```

```
## Order of parameters:  + logREVIEWS + logRATING + INDUSTRY + INDIA.HQ + logBRANCHES
```

```
cat("Adjusted R-squared: ", summary(out)$adj.r.squared)
```

```
## Adjusted R-squared:  0.8021424
```

Prediction

Lets take examples of Indian companies with:

Number of Reviews = 10000 (popular)
Rating = 3 (average)
Branches = 5
Headquarter = Bangalore / Bengaluru
Industry = IT Services & Consulting

```
x_new = data.frame(logREVIEWS = log(10000), logBRANCHES = log(5), logRATING=log(3),
                   INDIA.HQ=factor("Bangalore / Bengaluru", levels = levels(data$INDI
A.HQ)),
                   INDUSTRY=factor("IT Services & Consulting", levels = levels(data$I
NDUSTRY)))

y = predict(out, newdata = x_new)
cat("Average monthly package predicted: ", round(exp(y), 2), "L")
```

```
## Average monthly package predicted:  0.73 L
```

Number of Reviews = 10000 (popular)
Rating = 3 (average)
Branches = 5
Headquarter = Bangalore / Bengaluru
Industry = Education & Training

```
x_new = data.frame(logREVIEWS = log(10000), logBRANCHES = log(5), logRATING=log(3),
                   INDIA.HQ=factor("Bangalore / Bengaluru", levels = levels(data$INDI
A.HQ)),
                   INDUSTRY=factor("Education & Training", levels = levels(data$INDUS
TRY)))

y = predict(out, newdata = x_new)
cat("Average monthly package predicted: ", round(exp(y), 2), "L")
```

```
## Average monthly package predicted:  0.47 L
```