BIC MIC Model

Raynah Cheng

2024-12-03

BIC MIC Models Raynah Cheng

-cleaning data taken from Project Check in 2

```
#splitting data into test and train
train_index <- 1:165
test_index <- 166:nrow(comparison_data)

train_frame <- comparison_data[train_index,]
test_frame <- comparison_data[test_index,]


train_model <- lm(Severe_dyskaryosis ~ ., data = train_frame)
summary(train_model)
```

```
##
## Call:
## lm(formula = Severe_dyskaryosis ~ ., data = train_frame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1583.04   -69.40    66.49    94.30  1515.34
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -92.17422   41.51175  -2.220   0.0278 *
## Borderline_changes      0.09507    0.01087   8.748 2.77e-15 ***
## Mild_dyskaryosis       -0.14662    0.02321  -6.317 2.49e-09 ***
## Moderate_dyskaryosis    1.45536    0.09598  15.163  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 398.2 on 161 degrees of freedom
## Multiple R-squared:  0.9516, Adjusted R-squared:  0.9507
## F-statistic:  1055 on 3 and 161 DF,  p-value: < 2.2e-16
```

```
back_BIC <- step(train_model, direction = "backward",
                 k = log(nrow(train_frame)), trace = 0)
summary(back_BIC)
```

```
##
```

```
## Call:
## lm(formula = Severe_dyskaryosis ~ Borderline_changes + Mild_dyskaryosis +
##     Moderate_dyskaryosis, data = train_frame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1583.04   -69.40    66.49    94.30  1515.34
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -92.17422   41.51175  -2.220   0.0278 *
## Borderline_changes     0.09507    0.01087   8.748 2.77e-15 ***
## Mild_dyskaryosis      -0.14662    0.02321  -6.317 2.49e-09 ***
## Moderate_dyskaryosis   1.45536    0.09598  15.163  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 398.2 on 161 degrees of freedom
## Multiple R-squared:  0.9516, Adjusted R-squared:  0.9507
## F-statistic:  1055 on 3 and 161 DF,  p-value: < 2.2e-16
```

```r
back_MIC <- step(train_model, direction = "backward",
                 k = nrow(train_frame), trace = 0)
summary(back_MIC)
```

```
##
## Call:
## lm(formula = Severe_dyskaryosis ~ Moderate_dyskaryosis, data = train_frame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2625.70   -49.93    56.48    68.43  1635.61
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -61.79307   48.85798  -1.265    0.208
## Moderate_dyskaryosis   1.17963    0.02726  43.274   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 508.9 on 163 degrees of freedom
## Multiple R-squared:  0.9199, Adjusted R-squared:  0.9194
## F-statistic:  1873 on 1 and 163 DF,  p-value: < 2.2e-16
```

The BIC model chooses all of the dyskaryosis variables, with equal about equal levels of significance. However, we can see that the moderate dyskaryosis t value is higher, at 15.163. For our BIC model, 95.16% of the variance can be explained by our model. The MIC model only chooses the moderate dyskaryosis variable, at a .001 level of signifiance, with a much higher t value than the BIC model. However, the variance that can be explained goes down a little bit, at 91.99%.

```r
test_MIC <- predict(back_MIC, test_frame)
test_BIC <- predict(back_BIC, test_frame)

cor(test_MIC, test_frame$Severe_dyskaryosis, use = "complete.obs")^2 #r-squared for back_MIC model
```

```
## [1] 0.9966765
```

```r
cor(test_BIC, test_frame$Severe_dyskaryosis, use = "complete.obs")^2 #r-squared for back_BIC model
```

```
## [1] 0.9958021
```

```r
errors_MIC <- test_frame$Severe_dyskaryosis - test_MIC
errors_BIC <- test_frame$Severe_dyskaryosis - test_BIC

sqrt(mean(errors_MIC^2, na.rm = TRUE)) #RMSE for back_MIC
```

```
## [1] 1128.586
```

```r
sqrt(mean(errors_BIC^2, na.rm = TRUE)) #RMSE for back_BIC
```

```
## [1] 806.2014
```

```r
mean(abs(errors_MIC), na.rm = TRUE) #MAE for back_MIC
```

```
## [1] 248.288
```

```r
mean(abs(errors_BIC), na.rm = TRUE) #MAE for back_BIC
```

```
## [1] 194.5406
```

Model	R^2	RMSE	MAE
MIC	.997	1128.59	248.29
BIC	.996	806.2	194.54

The BIC model has a lower R^2 value, but the errors are lower. The MIC model has a higher R^2 value, but the errors are higher. I think that we would pick the BIC model because the R^2 is not that much lower and it predicts our data a lot better.