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(Borderline_changes ~ ., data = train_frame)
summary(train_model)
```

```
##
## Call:
## lm(formula = Borderline_changes ~ ., data = train_frame)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5767.6  -992.5  -897.0   595.1 12859.7
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            988.5484   239.2512   4.132 5.77e-05 ***
## Mild_dyskaryosis         0.6606     0.1458   4.531 1.14e-05 ***
## Moderate_dyskaryosis    -3.4393     0.8508  -4.042 8.18e-05 ***
## Severe_dyskaryosis       3.3890     0.3874   8.748 2.77e-15 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2377 on 161 degrees of freedom
## Multiple R-squared:  0.8112, Adjusted R-squared:  0.8076
## F-statistic: 230.5 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 = Borderline_changes ~ Mild_dyskaryosis + Moderate_dyskaryosis +
##     Severe_dyskaryosis, data = train_frame)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5767.6  -992.5  -897.0   595.1 12859.7
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           988.5484   239.2512   4.132 5.77e-05 ***
## Mild_dyskaryosis        0.6606     0.1458   4.531 1.14e-05 ***
## Moderate_dyskaryosis   -3.4393     0.8508  -4.042 8.18e-05 ***
## Severe_dyskaryosis      3.3890     0.3874   8.748 2.77e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2377 on 161 degrees of freedom
## Multiple R-squared:  0.8112, Adjusted R-squared:  0.8076
## F-statistic: 230.5 on 3 and 161 DF,  p-value: < 2.2e-16
```

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

```
##
## Call:
## lm(formula = Borderline_changes ~ Severe_dyskaryosis, data = train_frame)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5789.6 -1266.2  -940.9   535.1  9728.3
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1275.1192   233.8459   5.453 1.81e-07 ***
## Severe_dyskaryosis    2.6817     0.1093  24.534  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2510 on 163 degrees of freedom
## Multiple R-squared:  0.7869, Adjusted R-squared:  0.7856
## F-statistic: 601.9 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 severe dyskaryosis t value is higher, at 8.748, compared to the others at around 4.5. For our BIC model, 81.12% of the variance can be explained by our model. The MIC model only chooses the severe 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 78.69%.

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

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

```
## [1] 0.9967432
```

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

```
## [1] 0.9911532
```

```
errors_MIC <- test_frame$Borderline_changes - test_MIC
errors_BIC <- test_frame$Borderline_changes- test_BIC

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

```
## [1] 1763.824
```

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

```
## [1] 1650.034
```

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

```
## [1] 1286.169
```

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

```
## [1] 1056.618
```

Model	R^2	RMSE	MAE
MIC	.997	1763.82	1286.17
BIC	.991	1650.03	1056.62

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.