HW3 Biz Analytics

Justin Markey

9/27/2021

# Import the data

setwd("C:/Projects/R/bizanalytics/data")  
employee.data <- read.csv("HW3 Data.csv" , stringsAsFactors = TRUE)

# Question 1

model.1<-lm(data=employee.data , Errors~Yrs.Exp)  
summary(model.1)

##   
## Call:  
## lm(formula = Errors ~ Yrs.Exp, data = employee.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.830 -6.959 -1.295 5.240 39.170   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 36.7152 4.3979 8.348 4.41e-09 \*\*\*  
## Yrs.Exp -1.4428 0.2428 -5.943 2.14e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12.54 on 28 degrees of freedom  
## Multiple R-squared: 0.5578, Adjusted R-squared: 0.542   
## F-statistic: 35.32 on 1 and 28 DF, p-value: 2.135e-06

M1.Adj.RSquared <- summary(model.1)$adj.r.squared  
M1.Rsquared <- summary(model.1)$r.squared   
M1.RMSE <- sqrt(mean((predict(model.1)-employee.data$Errors)^2))  
  
M1.Adj.RSquared

## [1] 0.5419779

M1.Rsquared

## [1] 0.5577718

M1.RMSE

## [1] 12.114

# Question 2

model.2<-lm(data=employee.data , Errors~Yrs.Exp+Training)  
summary(model.2)

##   
## Call:  
## lm(formula = Errors ~ Yrs.Exp + Training, data = employee.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.242 -6.190 0.089 4.172 37.632   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 37.9305 4.3746 8.671 2.76e-09 \*\*\*  
## Yrs.Exp -1.2814 0.2603 -4.923 3.74e-05 \*\*\*  
## Training -7.4241 4.9085 -1.512 0.142   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12.26 on 27 degrees of freedom  
## Multiple R-squared: 0.5923, Adjusted R-squared: 0.5621   
## F-statistic: 19.61 on 2 and 27 DF, p-value: 5.488e-06

M2.Adj.RSquared <- summary(model.2)$adj.r.squared  
M2.Rsquared <- summary(model.2)$r.squared  
M2.RMSE <- sqrt(mean((predict(model.2)-employee.data$Errors)^2))  
  
M2.Adj.RSquared

## [1] 0.5621145

M2.Rsquared

## [1] 0.5923135

M2.RMSE

## [1] 11.63128

# Question 3

model.3<-lm(data=employee.data , Errors ~ Yrs.Exp + Training + Yrs.Exp\*Training)  
summary(model.3)

##   
## Call:  
## lm(formula = Errors ~ Yrs.Exp + Training + Yrs.Exp \* Training,   
## data = employee.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.582 -5.767 -0.312 2.996 33.622   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 42.7764 4.8453 8.828 2.66e-09 \*\*\*  
## Yrs.Exp -1.6991 0.3272 -5.193 2.02e-05 \*\*\*  
## Training -23.1111 9.2873 -2.488 0.0196 \*   
## Yrs.Exp:Training 0.9785 0.5007 1.954 0.0615 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11.67 on 26 degrees of freedom  
## Multiple R-squared: 0.6445, Adjusted R-squared: 0.6035   
## F-statistic: 15.71 on 3 and 26 DF, p-value: 4.955e-06

M3.Adj.RSquared <- summary(model.3)$adj.r.squared  
M3.Rsquared <- summary(model.3)$r.squared  
M3.RMSE <- sqrt(mean((predict(model.3)-employee.data$Errors)^2))  
  
M3.Adj.RSquared

## [1] 0.6035106

M3.Rsquared

## [1] 0.6445268

M3.RMSE

## [1] 10.86095

# Question 4

model.4<-lm(data=employee.data , Errors ~ Yrs.Exp + Yrs.Exp\*Training)  
summary(model.4)

##   
## Call:  
## lm(formula = Errors ~ Yrs.Exp + Yrs.Exp \* Training, data = employee.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.582 -5.767 -0.312 2.996 33.622   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 42.7764 4.8453 8.828 2.66e-09 \*\*\*  
## Yrs.Exp -1.6991 0.3272 -5.193 2.02e-05 \*\*\*  
## Training -23.1111 9.2873 -2.488 0.0196 \*   
## Yrs.Exp:Training 0.9785 0.5007 1.954 0.0615 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11.67 on 26 degrees of freedom  
## Multiple R-squared: 0.6445, Adjusted R-squared: 0.6035   
## F-statistic: 15.71 on 3 and 26 DF, p-value: 4.955e-06

M4.Adj.RSquared <- summary(model.4)$adj.r.squared  
M4.Rsquared <- summary(model.4)$r.squared  
M4.RMSE <- sqrt(mean((predict(model.4)-employee.data$Errors)^2))  
  
M4.Adj.RSquared

## [1] 0.6035106

M4.Rsquared

## [1] 0.6445268

M4.RMSE

## [1] 10.86095

# Question 5

Although we’ve never tested the model against out of sample data, The measures can help us with deciding the best structure and predictive power. The Best model structure is the one that incorporates as many predictors while also keeping Adj R Squared high. Adj R Squared has a built in cost function to the measure so that is why its important we focus on this measure for efficient model structure. Comparing all 4 models, the best for structure is model 3. Model 3 incorporates the most predictors while also having a high Adj R squared. Model 4 has the same Adj R Square but doesn’t include “Training”, meaning that including “Training” has some predictive power as Adj R square stays high. The best for forecasting accuracy is RMSE. Model 3 and 4 are similar, but the RMSE is the same while there is more predictive variables included in Model 3. This means that model 3 can incorporate more variables while also keeping its error down(RMSE), Therefore more rigid and accurate.

# Question 6

The predictions given the specified values will yield

predict(model.3 , newdata=data.frame(Yrs.Exp=c(10,20,5,1), Training=c(1,0,1,1)))

## 1 2 3 4   
## 12.459113 8.794065 16.062210 18.944687