Tugas Pertemuan 12: Pemilihan Model Regresi Terbaik

Sahda Huwaidah Estiningtyas

16 Mei 2024

### **Library**

library(tidyverse) #package untuk manipulasi data dan visualisasi data

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(modelr) #package untuk menghitung performa model regresi  
library(broom) # package untuk membangun data frame berisi ukuran kebaikan model regresi

##   
## Attaching package: 'broom'  
##   
## The following object is masked from 'package:modelr':  
##   
## bootstrap

library (leaps)  
library(MASS)

##   
## Attaching package: 'MASS'  
##   
## The following object is masked from 'package:dplyr':  
##   
## select

# Membaca Data mtcars.xlsx

library(readxl)  
data<- read\_xlsx("C:/Users/LENOVO/OneDrive/Documents/UNY/Semester 2/mtcars.xlsx")  
head(data)

## # A tibble: 6 × 12  
## type mpg cyl disp hp drat wt qsec vs am gear carb  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Mazda RX4 21 6 160 110 3.9 2.62 16.5 0 1 4 4  
## 2 Mazda RX4 W… 21 6 160 110 3.9 2.88 17.0 0 1 4 4  
## 3 Datsun 710 22.8 4 108 93 3.85 2.32 18.6 1 1 4 1  
## 4 Hornet 4 Dr… 21.4 6 258 110 3.08 3.22 19.4 1 0 3 1  
## 5 Hornet Spor… 18.7 8 360 175 3.15 3.44 17.0 0 0 3 2  
## 6 Valiant 18.1 6 225 105 2.76 3.46 20.2 1 0 3 1

str(data)

## tibble [32 × 12] (S3: tbl\_df/tbl/data.frame)  
## $ type: chr [1:32] "Mazda RX4" "Mazda RX4 Wag" "Datsun 710" "Hornet 4 Drive" ...  
## $ mpg : num [1:32] 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...  
## $ cyl : num [1:32] 6 6 4 6 8 6 8 4 4 6 ...  
## $ disp: num [1:32] 160 160 108 258 360 ...  
## $ hp : num [1:32] 110 110 93 110 175 105 245 62 95 123 ...  
## $ drat: num [1:32] 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...  
## $ wt : num [1:32] 2.62 2.88 2.32 3.21 3.44 ...  
## $ qsec: num [1:32] 16.5 17 18.6 19.4 17 ...  
## $ vs : num [1:32] 0 0 1 1 0 1 0 1 1 1 ...  
## $ am : num [1:32] 1 1 1 0 0 0 0 0 0 0 ...  
## $ gear: num [1:32] 4 4 4 3 3 3 3 4 4 4 ...  
## $ carb: num [1:32] 4 4 1 1 2 1 4 2 2 4 ...

# Introduction Pemilihan Model Regresi Terbaik

model1<-lm(mpg~cyl,data=data)   
model2<-lm(mpg~disp,data=data)   
model3<-lm(mpg~hp,data=data)   
model4<-lm(mpg~drat,data=data)   
model5<-lm(mpg~wt,data=data)   
model6<-lm(mpg~qsec,data=data)   
model7<-lm(mpg~gear,data=data)   
model8<-lm(mpg~carb,data=data)   
model255<-lm(mpg~cyl+disp+hp+drat+wt+qsec+gear+carb,data=data)

Berdasarkan beberapa kemungkinan model di atas, pemilihan model terbaik dapat dilakukan dengan memperhatikan nilai-nilai , RMSE, dan MAE. Model regresi yang terbaik adalah model regresi yang mempunyai nilai yang tinggi. Nilai yang tinggi berarti model regresi semakin dapat menjelaskan keragaman data. Model regresi yang terbaik adalah model regresi dengan nilai RMSE dan MAE yang rendah.

## Menghitung nilai R2, RMSE, dan MAPE

m1<-data.frame(R2 = rsquare(model1, data = data), RMSE = rmse(model1, data = data),MAE = mae(model1, data = data))  
m2<-data.frame(R2 = rsquare(model2, data = data), RMSE = rmse(model2, data = data),MAE = mae(model2, data = data))  
m3<-data.frame(R2 = rsquare(model3, data = data), RMSE = rmse(model3, data = data),MAE = mae(model3, data = data))  
m4<-data.frame(R2 = rsquare(model4, data = data), RMSE = rmse(model4, data = data),MAE = mae(model4, data = data))  
m5<-data.frame(R2 = rsquare(model5, data = data), RMSE = rmse(model5, data = data),MAE = mae(model5, data = data))  
m6<-data.frame(R2 = rsquare(model6, data = data), RMSE = rmse(model6, data = data),MAE = mae(model6, data = data))  
m7<-data.frame(R2 = rsquare(model7, data = data), RMSE = rmse(model7, data = data),MAE = mae(model7,data = data))  
m8<-data.frame(R2 = rsquare(model8, data = data), RMSE = rmse(model8, data = data),MAE = mae(model8,data = data))  
m255<-data.frame(R2 = rsquare(model255, data = data), RMSE = rmse(model255, data = data),MAE = mae(model255,data = data))  
  
datam<-rbind(m1,m2,m3,m4,m5,m6,m7,m8,m255)  
datam<-as.data.frame(datam)  
  
list\_model<-rbind("model1","model2","model3","model4","model5","model6","model7","model8","model255")  
  
gabungan<-cbind(list\_model,datam)  
gabungan

## list\_model R2 RMSE MAE  
## 1 model1 0.7261800 3.104101 2.399945  
## 2 model2 0.7183433 3.148207 2.605473  
## 3 model3 0.6024373 3.740297 2.907452  
## 4 model4 0.4639952 4.342978 3.351388  
## 5 model5 0.7528328 2.949163 2.340642  
## 6 model6 0.1752963 5.387066 4.181625  
## 7 model7 0.2306734 5.203058 3.920000  
## 8 model8 0.3035184 4.950603 4.217702  
## 9 model255 0.8595764 2.222918 1.732645

Berdasarkan hasil analisis:

* Model terbaik adalah model255 () dengan , RMSE = 2.222918, dan MAE = 1.732645.
* Model ini memiliki nilai tertinggi dan kesalahan prediksi (RMSE dan MAE) terendah, sehingga memberikan prediksi yang paling akurat.

## Menghitung Kriteria R2 adjusted, AIC, dan BIC

a1<-glance(model1) %>% dplyr::select(adj.r.squared, AIC, BIC)  
  
a2<-glance(model2) %>% dplyr::select(adj.r.squared, AIC, BIC)  
  
a3<-glance(model3) %>% dplyr::select(adj.r.squared, AIC, BIC)  
  
a4<-glance(model3) %>% dplyr::select(adj.r.squared, AIC, BIC)  
  
a5<-glance(model3) %>% dplyr::select(adj.r.squared, AIC, BIC)  
  
a6<-glance(model3) %>% dplyr::select(adj.r.squared, AIC, BIC)  
  
a7<-glance(model7) %>% dplyr::select(adj.r.squared, AIC, BIC)  
  
a8<-glance(model3) %>% dplyr::select(adj.r.squared, AIC, BIC)  
  
a255<-glance(model3) %>% dplyr::select(adj.r.squared, AIC, BIC)  
  
dataa<-rbind(a1,a2,a3,a4,a5,a6,a7,a8,a255)  
dataa<-as.data.frame(dataa)  
  
list\_modela<-rbind("model1","model2","model3","model4","model5","model6","model7","model8","model255")  
  
gabungana<-cbind(list\_modela,dataa)  
gabungana

## list\_modela adj.r.squared AIC BIC  
## 1 model1 0.7170527 169.3064 173.7036  
## 2 model2 0.7089548 170.2094 174.6066  
## 3 model3 0.5891853 181.2386 185.6358  
## 4 model4 0.5891853 181.2386 185.6358  
## 5 model5 0.5891853 181.2386 185.6358  
## 6 model6 0.5891853 181.2386 185.6358  
## 7 model7 0.2050292 202.3638 206.7611  
## 8 model8 0.5891853 181.2386 185.6358  
## 9 model255 0.5891853 181.2386 185.6358

## Perhitungan Manual RMSE, MAE, R2, R2 Adjusted, Cp Mallow

anova(model1)

## Analysis of Variance Table  
##   
## Response: mpg  
## Df Sum Sq Mean Sq F value Pr(>F)   
## cyl 1 817.71 817.71 79.561 6.113e-10 \*\*\*  
## Residuals 30 308.33 10.28   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

JKG1<-308.33  
JKT1<-308.33+817.71  
KTG1<-10.28  
n<-length(data$mpg)  
  
RMSE1 <- sqrt(mean((model1$residuals)^2))  
RMSE1

## [1] 3.104101

MAE1 <- mean(abs(model1$residuals))  
MAE1

## [1] 2.399945

Rsquare1<-(1-JKG1/JKT1)  
Rsquare1

## [1] 0.726182

AdjRSquare1<-1-((n-1)/JKT1)\*KTG1 #n adalahbanyaknya data  
AdjRSquare1

## [1] 0.7169905

anova(model255)

## Analysis of Variance Table  
##   
## Response: mpg  
## Df Sum Sq Mean Sq F value Pr(>F)   
## cyl 1 817.71 817.71 118.9411 1.452e-10 \*\*\*  
## disp 1 37.59 37.59 5.4683 0.028423 \*   
## hp 1 9.37 9.37 1.3631 0.254970   
## drat 1 16.47 16.47 2.3953 0.135352   
## wt 1 77.48 77.48 11.2693 0.002728 \*\*   
## qsec 1 3.95 3.95 0.5744 0.456192   
## gear 1 4.68 4.68 0.6808 0.417778   
## carb 1 0.67 0.67 0.0978 0.757278   
## Residuals 23 158.12 6.87   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

KTG255<-6.87  
Cp1 <- JKG1/KTG255 + 2\*(1+1)-20 #Cp untuk model 1 tapi aku masih bingung yg ini  
Cp1

## [1] 28.88064

# Penerapan Model Regresi

## Pembagian Training Testing Data

set.seed(100) # setting seed to reproduce results of random sampling  
trainingRowIndex <- sample(1:nrow(data), 0.8\*nrow(data)) # row indices for training data  
trainingData <- data[trainingRowIndex, ] # model training data  
testData <- data[-trainingRowIndex, ] # test data

## Model Regresi (data training)

lmMod <- lm(mpg~cyl+disp+hp+drat+wt+qsec+gear+carb, data=trainingData) # build the model  
summary(lmMod)

##   
## Call:  
## lm(formula = mpg ~ cyl + disp + hp + drat + wt + qsec + gear +   
## carb, data = trainingData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.896 -1.583 -0.337 1.224 5.951   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 14.26122 25.79898 0.553 0.588  
## cyl -0.67258 1.04970 -0.641 0.531  
## disp 0.01084 0.02102 0.515 0.613  
## hp -0.02208 0.02419 -0.913 0.375  
## drat 1.74932 1.75323 0.998 0.333  
## wt -2.66387 2.18456 -1.219 0.240  
## qsec 0.55008 0.98761 0.557 0.585  
## gear 0.89187 1.49367 0.597 0.559  
## carb -0.07172 0.91258 -0.079 0.938  
##   
## Residual standard error: 2.52 on 16 degrees of freedom  
## Multiple R-squared: 0.857, Adjusted R-squared: 0.7854   
## F-statistic: 11.98 on 8 and 16 DF, p-value: 1.924e-05

## Prediksi Data Test dengan Model Regresi

distPred <- predict(lmMod, testData) # predict fat  
  
actuals\_preds <- data.frame(cbind(actuals=testData$mpg, predicteds=distPred)) # make actuals\_predicteds dataframe.  
correlation\_accuracy <- cor(actuals\_preds)  
head(actuals\_preds)

## actuals predicteds  
## 1 21.0 21.70822  
## 2 18.7 17.15807  
## 3 22.8 25.48568  
## 4 10.4 12.88736  
## 5 33.9 27.84219  
## 6 30.4 26.28274

min\_max\_accuracy <- mean(apply(actuals\_preds, 1, min) / apply(actuals\_preds, 1, max))#MinMaxAcuracy  
  
mape <- mean(abs((actuals\_preds$predicteds - actuals\_preds$actuals))/actuals\_preds$actuals) # Mean Absolute Percentage Deviation (MAPE)

Pemilihan model terbaik dapat diulang dengan memodelkan ke-7 kemungkinan model yang dapat dibentuk kemudian dibandingkan nilai-nilai kriteria kebaikan model.