Machine Learning-01 Linear Regression

Created by H. M. Samadhi Chathuranga Rathnayake

```r
setwd("D:\\Workshops\\R Programming for Data Science Workshop\\Part 04 - Machine Learning\\Datasets")
df=read.csv("gapminder.CSV")
head(df)

#Factorize the categorical variables
str(df)

df$country=factor(df$country)
df$year=factor(df$year)
df$continent=factor(df$continent)
contrasts(df$continent)

str(df)

#Simple linear regression
fit1=lm(gdpPercap~lifeExp,data=df)
summary(fit1)

fit2=lm(gdpPercap~continent,data=df)
summary(fit2)

#Multiple linear regression
fit3=lm(gdpPercap~pop+lifeExp,data=df)
summary(fit3)

fit_full=lm(gdpPercap~.,data=df) #NA s are given due to the exact collinearity
summary(fit_full)

fit_full2=lm(gdpPercap~.-continent,data=df)
summary(fit_full2)

fit_full3=lm(gdpPercap~.-country,data=df)
summary(fit_full3)

fit_new=lm(gdpPercap~pop+continent+lifeExp,data=df)
summary(fit_new)

#Testing the assumptions
par(mfrow=c(2,2))
plot(fit_new)

#Checking Multicollinearity
library(car)
```

```r
vif(fit_new)

#Testing the prediction accuracy
set.seed(7777)
trainID=sample(1:nrow(df),0.8*nrow(df))
train=df[trainID,]
test=df[-trainID,]



fit_train=lm(gdpPercap~.-country,data=train)
summary(fit_train)

fit_train_new=lm(gdpPercap~pop+continent+lifeExp,data=train)
summary(fit_train_new)

y_pred=predict(fit_train_new,test)
y_actual=test$gdpPercap

MSE=mean((y_actual-y_pred)^2)
RMSE=sqrt(MSE)
RMSE
```