# Regression_Methods2(Question 3)

Oluwanifemi

2024-09-27

```
file <- wblake
View(file)

data(wblake)
```

## Question 3(a)

```
## 66.67 or 67 % of the sample size
smp_size <- floor((2/3) * nrow(wblake))

## set the seed to make your partition reproducible
set.seed(123)
train_ind <- sample(seq_len(nrow(wblake)), size = smp_size)
train_ind1 <- sample(x= 1:nrow(wblake), smp_size)

train <- wblake[train_ind, ]
test <- wblake[-train_ind, ]

train1 <- wblake[train_ind1, ]
test1 <- wblake[-train_ind1, ]
```

## Question 3(b)

```
lage <- lm(Length ~ Age, data = train1)
summary(lage)
```

```
##
## Call:
## lm(formula = Length ~ Age, data = train1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -85.76 -19.73  -3.50  14.82  91.11
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  66.2446     3.7679   17.58   <2e-16 ***
```

```
## Age          30.2159     0.8257   36.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.78 on 290 degrees of freedom
## Multiple R-squared:  0.822,  Adjusted R-squared:  0.8214
## F-statistic:  1339 on 1 and 290 DF,  p-value: < 2.2e-16
```

```
observed <- train1$Length

predicted <- lage$fitted.values

#sse1 <- sum(residuals^2)
#sse1

sst2 <- sum((observed-mean(observed))^2)
sst2
```

```
## [1] 1257394
```

```
sse2 <- sum((observed - predicted)^2)
sse2
```

```
## [1] 223837
```

```
r2 <- 1-(sse2/sst2)
r2
```

```
## [1] 0.8219834
```

## Question 3(c)

```
mse2<-mean(lage$residuals^2)
mse2
```

```
## [1] 766.5651
```

```
predicted_values <- predict(lage, newdata = test1)

# Step 4: Calculate the MSE on the test data
actual_values <- test1$Length
mse_test <- mean((actual_values - predicted_values)^2)
mse_test
```

```
## [1] 917.07
```

## Question 3(d)

```r
#Trying different models to get a better one
train1$Age_Squared <-c((train$Age)^2)
test1$Age_Squared <-c((test$Age)^2)

model <- lm(Length~Age_Squared, data = train1)
summary(model)
```

```
##
## Call:
## lm(formula = Length ~ Age_Squared, data = train1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -133.760  -56.699   -0.678   56.037  173.240
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 187.4469     6.2598  29.944   <2e-16 ***
## Age_Squared   0.1459     0.2265   0.644     0.52
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65.8 on 290 degrees of freedom
## Multiple R-squared:  0.001429,   Adjusted R-squared:  -0.002014
## F-statistic: 0.4151 on 1 and 290 DF,  p-value: 0.5199
```

```r
predicted_values1 <- predict(model, newdata = test1)

# Step 4: Calculate the MSE on the test data
actual_values1 <- test1$Length
mse_test1 <- mean((actual_values1 - predicted_values1)^2)
mse_test1
```

```
## [1] 4536.202
```

```r
model_1 <- lm(Scale~Age_Squared,data = train1)
summary(model_1)
```

```
##
## Call:
## lm(formula = Scale ~ Age_Squared, data = train1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6944 -2.3114 -0.1931  2.2648  8.6842
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.580132   0.252831  22.071   <2e-16 ***
## Age_Squared 0.009088   0.009146   0.994    0.321
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.658 on 290 degrees of freedom
## Multiple R-squared:  0.003393,   Adjusted R-squared:  -4.324e-05
## F-statistic: 0.9874 on 1 and 290 DF,  p-value: 0.3212
```

```
predicted_values2 <- predict(model_1, newdata = test1)

# Step 4: Calculate the MSE on the test data
actual_values2 <- test1$Length
mse_test2 <- mean((actual_values2 - predicted_values2)^2)
mse_test2
```

```
## [1] 41499.27
```

```
model_2 <- lm(Length~Age_Squared+Scale, data = train1)
summary(model_2)
```

```
##
## Call:
## lm(formula = Length ~ Age_Squared + Scale, data = train1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.005  -9.571  -0.001  14.437  76.425
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 58.55392    3.69577  15.843   <2e-16 ***
## Age_Squared -0.06403    0.08181  -0.783    0.434
## Scale       23.09855    0.52436  44.051   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.73 on 289 degrees of freedom
## Multiple R-squared:  0.8706, Adjusted R-squared:  0.8697
## F-statistic: 971.8 on 2 and 289 DF,  p-value: < 2.2e-16
```

```
predicted_values3 <- predict(model_2, newdata = test1)

# Step 4: Calculate the MSE on the test data
actual_values3 <- test1$Length
mse_test3 <- mean((actual_values3 - predicted_values3)^2)
mse_test3
```

```
## [1] 487.0037
```

```
model_3 <- lm(Scale~Age_Squared + Length,data = train1)
summary(model_3)
```

```
##
```

```
## Call:
## lm(formula = Scale ~ Age_Squared + Length, data = train1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9702 -0.4899 -0.1759  0.2061  4.0310
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.4830145  0.1844573  -8.040 2.32e-14 ***
## Age_Squared  0.0035909  0.0033010   1.088    0.278
## Length       0.0376808  0.0008554  44.051  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9585 on 289 degrees of freedom
## Multiple R-squared:  0.8708, Adjusted R-squared:  0.8699
## F-statistic:   974 on 2 and 289 DF,  p-value: < 2.2e-16
```

```
predicted_values4 <- predict(model_3, newdata = test1)

# Step 4: Calculate the MSE on the test data
actual_values4 <- test1$Length
mse_test4 <- mean((actual_values4 - predicted_values4)^2)
mse_test
```

```
## [1] 917.07
```

**The best model with a lower mse is the model (quadratic) where Length is y and Age_squared is x + another variable Scale. the initial mse is 917.07 but the one for this new model is 487.00.**