```
---
title: "Linear Regression Coding Assignment-3"
editor_options:
  chunk_output_type: inline
output:
  html_document:
    df_print: paged
  pdf_document: default
  word_document: default
---
```{r}
library(ggplot2)
library(dplyr)
library(reshape)
```

```{r}
# Load the diabetes dataset:
# 10 predictors which are age, gender (1-female, 2-male), body-mass index, average blood
pressure, and six blood serum measurements and 1 response variable which is a quantitative
measure of disease progression one year after baseline)
df = read.csv('diabetes.csv', header = TRUE, stringsAsFactors = FALSE)
str(df)
```



```{r}
# Create a new feature called BMILEVEL using the BMI column and the following rules: BMI <
18.5 is underweight, 18.5 <= BMI <= 24.9 is healthy, 25 <= BMI <= 29.9 is overweight, BMI
>= 30 is unhealthy
df = df %>% mutate(BMILEVEL = case_when(BMI < 18.5 ~ 'underweight', BMI >= 18.5 & BMI <=
24.9 ~ 'healthy', BMI >= 30.0 ~'unhealthy'))
str(df)
```



```{r}
# Convert 'GENDER' and 'BMILEVEL' columns to factors
categorical_cols = c('GENDER', 'BMILEVEL')
df[categorical_cols] = lapply(df[categorical_cols], as.factor)
str(df)
```



```{r}
# Create a list of continuous columns
continuous_cols = setdiff(colnames(df), categorical_cols)
continuous_cols
```



```{r}
# How many levels does the categorical variable *BMILEVEL* have? What is the reference
level?
levels(df$BMILEVEL)
```

there are 3 levels and reference level is healthy


```{r}
# Fit a linear model for predicting disease progression using BMILEVEL. Print the model's
summary.
# How accurate is the model?
```
```

Which level in BMILEVEL is most likely to not have a linear relationship with disease progression? What is the reason?
How worse is the disease progression in unhealthy people compared to the healthy ones?
How worse is the disease progression in unhealthy people compared to the overweight ones?
Write down the individual model for each level in BMILEVEL
?

model = lm(data=df, Y ~ BMILEVEL)
summary(model)


```

model is 81.9% accurate

I think Y_cap is going to have highest linear relationship with bmi level unhealthy as it's coefficient's value is more

when everything else is kept constant, the difference in Y_healthy and Y_unhealthy is 103.967

when everything else is kept constant, the difference in Y_healthy and Y_underweight is -10.376

Y_cap = beta0 + beta1 * BMILEVELunderweight   + beta2 * BMILEVELunhealthy


```{r}
# Fit a linear model for predicting disease progression using BMILEVEL and the blood serum measurements.
# From the model summary, explain how you will find out which blood serum measurements are most likely to have a linear relationship with disease progression.
# Fit a model using BMILEVEL and the blood serum measurements identified in the previous question and compare its accuracy with the model fit using BMILEVEL and all blood serum measurements.
?
model = lm(data=df, Y ~ BMILEVEL + S1 + S2 + S3 + S4 + S5 + S6)
summary(model)

model = lm(data=df, Y ~ BMILEVEL + S1 + S4 + S5)
summary(model)

```

except S5, rest all coefficients values are zero or insignificant so they don't have linear relationship with target variable as they are useless features. So only S5 has linear relationship with the model (maybe even s4 and s1 although their contribution is small)


```{r}
# Fit a linear model for predicting disease progression using BMI, age, BP, and gender. How accurate is the model?
# According to the model, which gender has a worse disease progression? Explain why.
# For the same age, BP, and gender, decreasing BMI by 1 unit causes what change in the disease progression?
# For the same age and BP, which gender benefits better w.r.t. disease progressions by decreasing BMI by 1 unit. Explain.
?

model = lm(data=df, Y ~ BMILEVEL + AGE + BP + GENDER)
summary(model)
```
```

```{r}
# Fit a linear model for predicting disease progression using BMI, age, BP, gender and interaction between BMI and gender. Is this model more accurate than the model without interaction between BMI and gender?
?

model = lm(data=df, Y ~ BMILEVEL + AGE + BP + GENDER + BMILEVEL*GENDER)
summary(model)
```

yes, the model with interaction is slightly more accurate than model without interaction