Linear Regression Coding Assignment-3

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(reshape)
```

```
## Warning: package 'reshape' was built under R version 4.3.2
```

```
##
## Attaching package: 'reshape'
```

```
## The following object is masked from 'package:dplyr':
##
##     rename
```

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

```
## 'data.frame':    442 obs. of  11 variables:
##  $ AGE   : int  59 48 72 24 50 23 36 66 60 29 ...
##  $ GENDER: int  2 1 2 1 1 1 2 2 2 1 ...
##  $ BMI   : num  32.1 21.6 30.5 25.3 23 22.6 22 26.2 32.1 30 ...
##  $ BP    : num  101 87 93 84 101 89 90 114 83 85 ...
##  $ S1    : int  157 183 156 198 192 139 160 255 179 180 ...
##  $ S2    : num  93.2 103.2 93.6 131.4 125.4 ...
##  $ S3    : num  38 70 41 40 52 61 50 56 42 43 ...
##  $ S4    : num  4 3 4 5 4 2 3 4.55 4 4 ...
##  $ S5    : num  4.86 3.89 4.67 4.89 4.29 ...
##  $ S6    : int  87 69 85 89 80 68 82 92 94 88 ...
##  $ Y     : int  151 75 141 206 135 97 138 63 110 310 ...
```

```
# 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 >= 18.5 & BMI <= 24.9 ~ 'overweight' ,BMI >= 30 ~ 'unhealthy'))
str(df)
```

```
## 'data.frame':    442 obs. of  12 variables:
##  $ AGE     : int  59 48 72 24 50 23 36 66 60 29 ...
##  $ GENDER  : int  2 1 2 1 1 1 2 2 2 1 ...
##  $ BMI     : num  32.1 21.6 30.5 25.3 23 22.6 22 26.2 32.1 30 ...
##  $ BP      : num  101 87 93 84 101 89 90 114 83 85 ...
##  $ S1      : int  157 183 156 198 192 139 160 255 179 180 ...
##  $ S2      : num  93.2 103.2 93.6 131.4 125.4 ...
##  $ S3      : num  38 70 41 40 52 61 50 56 42 43 ...
##  $ S4      : num  4 3 4 5 4 2 3 4.55 4 4 ...
##  $ S5      : num  4.86 3.89 4.67 4.89 4.29 ...
##  $ S6      : int  87 69 85 89 80 68 82 92 94 88 ...
##  $ Y       : int  151 75 141 206 135 97 138 63 110 310 ...
##  $ BMILEVEL: chr  "unhealthy" "healthy" "unhealthy" NA ...
```

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

```
## 'data.frame':    442 obs. of  12 variables:
##  $ AGE     : int  59 48 72 24 50 23 36 66 60 29 ...
##  $ GENDER  : Factor w/ 2 levels "1","2": 2 1 2 1 1 1 2 2 2 1 ...
##  $ BMI     : num  32.1 21.6 30.5 25.3 23 22.6 22 26.2 32.1 30 ...
##  $ BP      : num  101 87 93 84 101 89 90 114 83 85 ...
##  $ S1      : int  157 183 156 198 192 139 160 255 179 180 ...
##  $ S2      : num  93.2 103.2 93.6 131.4 125.4 ...
##  $ S3      : num  38 70 41 40 52 61 50 56 42 43 ...
##  $ S4      : num  4 3 4 5 4 2 3 4.55 4 4 ...
##  $ S5      : num  4.86 3.89 4.67 4.89 4.29 ...
##  $ S6      : int  87 69 85 89 80 68 82 92 94 88 ...
##  $ Y       : int  151 75 141 206 135 97 138 63 110 310 ...
##  $ BMILEVEL: Factor w/ 3 levels "healthy","underweight",..: 3 1 3 NA 1 1 1 NA 3 3 ...
```

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

```
##  [1] "AGE" "BMI" "BP"  "S1"  "S2"  "S3"  "S4"  "S5"  "S6"  "Y"
```

```
# How many levels does the categorical variable *BMILEVEL* have? What is the reference level?
# Check the levels of the 'BMILEVEL' variable
levels_bmilevel <- levels(df$BMILEVEL)

# Number of levels
num_levels_bmilevel <- length(levels_bmilevel)

# Display the number of levels and the levels themselves
cat("Number of levels in BMILEVEL:", num_levels_bmilevel, "\n")
```

```
## Number of levels in BMILEVEL: 3
```

```
cat("Levels in BMILEVEL:", levels_bmilevel, "\n")
```

```
## Levels in BMILEVEL: healthy underweight unhealthy
```

```
# Identify the reference level (usually the first level)

cat("Reference level in BMILEVEL:", levels_bmilevel[1], "\n")
```

```
## Reference level in BMILEVEL: healthy
```

```
# Fit a linear model for predicting disease progression using BMILEVEL. Print the model's sum
mary.
# How accurate is the model?
# Which level in BMILEVEL is most likely to not have a linear relationship with disease progr
ession? 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

linear_model = lm(data = df, Y ~ BMILEVEL)
summary(linear_model)
```

```
##
## Call:
## lm(formula = Y ~ BMILEVEL, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -161.343  -43.376   -8.376   45.157  171.624
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          109.376      4.477  24.433   <2e-16 ***
## BMILEVELunderweight  -10.376     43.403  -0.239    0.811
## BMILEVELunhealthy    103.967      7.596  13.688   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61.05 on 284 degrees of freedom
##   (155 observations deleted due to missingness)
## Multiple R-squared:  0.399,  Adjusted R-squared:  0.3947
## F-statistic: 94.25 on 2 and 284 DF,  p-value: < 2.2e-16
```

```
# the model is not that accurate because r squared value is less
# BMI level underweight as the p value is very high

#>The difference in coefficients of unhealthy people compared to the overweight ones indicate
s that disease progression is significantly worse in unhealthy people compared to overweight
ones.
```

```
# Fit a linear model for predicting disease progression using BMILEVEL and the blood serum me
asurements.
# From the model summary, explain how you will find out which blood serum measurements are mo
st likely to have a linear relationship with disease progression.
# Fit a model using BMILEVEL and the blood serum measurements identified in the previous ques
tion and compare its accuracy with the model fit using BMILEVEL and all blood serum measureme
nts.

linear_model_all_serums <- lm(Y ~ BMILEVEL + S1 + S2 + S3 + S4 + S5 + S6, data = df)

# Print the model summary
summary(linear_model_all_serums)
```

```
##
## Call:
## lm(formula = Y ~ BMILEVEL + S1 + S2 + S3 + S4 + S5 + S6, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -166.828  -35.563   -1.768   35.186  144.399
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -197.4099    78.9690  -2.500 0.013001 *
## BMILEVELunderweight  -38.1959    38.7579  -0.986 0.325235
## BMILEVELunhealthy     63.7807     8.1675   7.809 1.18e-13 ***
## S1                    -1.3060     0.6878  -1.899 0.058602 .
## S2                     0.9258     0.6349   1.458 0.145957
## S3                     0.7045     0.9171   0.768 0.442995
## S4                     7.1350     7.4327   0.960 0.337922
## S5                    70.4353    18.6147   3.784 0.000189 ***
## S6                     0.8008     0.3180   2.518 0.012348 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.12 on 278 degrees of freedom
##   (155 observations deleted due to missingness)
## Multiple R-squared:  0.5378, Adjusted R-squared:  0.5245
## F-statistic: 40.43 on 8 and 278 DF,  p-value: < 2.2e-16
```

```
# blood serum measurements are most likely to have a linear relationship with disease progres
sion for which the p values are very less

linear_model_all_serums <- lm(Y ~ BMILEVEL +S1+ S5 + S6, data = df)

# Print the model summary
summary(linear_model_all_serums)
```

```
##
## Call:
## lm(formula = Y ~ BMILEVEL + S1 + S5 + S6, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -164.91  -36.29   -4.95   35.59  163.43
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -152.7916    33.6016  -4.547 8.09e-06 ***
## BMILEVELunderweight  -30.8716    39.0671  -0.790  0.43007
## BMILEVELunhealthy     68.6433     8.0455   8.532 9.09e-16 ***
## S1                    -0.3200     0.1152  -2.779  0.00583 **
## S5                    55.6357     8.2461   6.747 8.61e-11 ***
## S6                     0.8667     0.3201   2.708  0.00719 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.7 on 281 degrees of freedom
##   (155 observations deleted due to missingness)
## Multiple R-squared:  0.5227, Adjusted R-squared:  0.5142
## F-statistic: 61.53 on 5 and 281 DF,  p-value: < 2.2e-16
```

```
# the r squared value for both adjusted and multiple decreases which indicates the accuracy o
f model is less than the previous model
```

```
# Fit a linear model for predicting disease progression using BMI, age, BP, and gender. How a
ccurate 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 diseas
e progression?
# For the same age and BP, which gender benefits better w.r.t. disease progressions by decrea
sing BMI by 1 unit. Explain.
linear_model = lm(Y ~ BMI + AGE + BP + GENDER , data = df)
summary(linear_model)
```

```
##
## Call:
## lm(formula = Y ~ BMI + AGE + BP + GENDER, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -152.417  -43.576   -3.757   42.938  150.054
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -209.2284    22.6318  -9.245  < 2e-16 ***
## BMI            8.4843     0.7051  12.032  < 2e-16 ***
## AGE            0.1353     0.2329   0.581    0.562
## BP             1.4345     0.2393   5.996 4.25e-09 ***
## GENDER2      -10.1590     5.9219  -1.716    0.087 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.98 on 437 degrees of freedom
## Multiple R-squared:  0.4003, Adjusted R-squared:  0.3948
## F-statistic: 72.91 on 4 and 437 DF,  p-value: < 2.2e-16
```

```
# the model is 47 to 48% accurate
# Gender 2 has has a worse disease progression because of high p value

coefficients <- coef(linear_model)

# Find the coefficient for BMI
coeff_bmi <- coefficients["BMI"]

# Interpretation: A 1-unit decrease in BMI is associated with a change in disease progression
# equal to the coefficient for BMI, while holding age, BP, and gender constant.
change_in_disease_progression <- coeff_bmi
change_in_disease_progression
```

```
##      BMI
## 8.484339
```

```
# Fit a linear model for predicting disease progression using BMI, age, BP, gender and intera
ction between BMI and gender. Is this model more accurate than the model without interaction
between BMI and gender?
model = lm(data = df, Y ~ BMI + AGE + BP + GENDER + BMI:GENDER)
summary(model)
```

```
##
## Call:
## lm(formula = Y ~ BMI + AGE + BP + GENDER + BMI:GENDER, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -150.312  -41.740   -3.209   41.767  149.119
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -174.7986    27.0004  -6.474 2.58e-10 ***
## BMI            7.2106     0.8922   8.082 6.34e-15 ***
## AGE            0.1691     0.2322   0.728   0.4670
## BP             1.4032     0.2385   5.884 7.97e-09 ***
## GENDER2      -90.1718    35.1134  -2.568   0.0106 *
## BMI:GENDER2    3.0257     1.3090   2.311   0.0213 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.68 on 436 degrees of freedom
## Multiple R-squared:  0.4075, Adjusted R-squared:  0.4007
## F-statistic: 59.98 on 5 and 436 DF,  p-value: < 2.2e-16
```

```
#there is significant change in the accuracy for a linear model without interaction between B
MI and Gender
```