Linear Regression Coding Assignment-3

Code

Hide

```
library(ggplot2)
library(dplyr)
library(reshape)
```

Hide

```
# Load the diabetes dataset:
# 10 predictors which are age, gender (1-female, 2-male), body-mass index, average blood pressure, and six blood serum measu
rements and 1 response variable which is a quantitative measure of disease progression one year after baseline)
df = read.csv("D:/2nd sem/AAPS/Codes/Data/diabetes_new.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 ...
```

Hide

```
# 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 >= 25.5 & BMI
<= 29.9 ~ '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 ...
```

Hide

```
# 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/ 4 levels "healthy","overweight",..: 4 1 4 NA 1 1 1 2 4 4 ...
```

Hide

```
# 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"
```

Hide

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

```
           overweight underweight unhealthy
healthy             0           0         0
overweight          1           0         0
underweight         0           1         0
unhealthy           0           0         1
```

Hide

```
levels(df$BMILEVEL)
```

```
[1] "healthy"     "overweight"  "underweight" "unhealthy"
```

Hide

```
# 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)
```

```
Call:
lm(formula = Y ~ BMILEVEL, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-161.343  -45.376   -7.376   49.679  171.624

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)          109.376      4.794  22.816  < 2e-16 ***
BMILEVELoverweight    57.879      7.361   7.863 3.19e-14 ***
BMILEVELunderweight  -10.376     46.477  -0.223    0.823
BMILEVELunhealthy    103.967      8.134  12.782  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 65.38 on 420 degrees of freedom
  (18 observations deleted due to missingness)
Multiple R-squared:  0.2941,    Adjusted R-squared:  0.2891
F-statistic: 58.33 on 3 and 420 DF,  p-value: < 2.2e-16
```

Hide

```
model1 = lm(data = df, Y ~ BMILEVEL == "underweight")
summary(model1)
```

```
Call:
lm(formula = Y ~ BMILEVEL == "underweight", data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-121.56  -66.81  -11.56   61.44  193.44

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  152.557      3.775  40.415   <2e-16 ***
BMILEVEL == "underweight"TRUE  -53.557     54.961  -0.974     0.33
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 77.54 on 422 degrees of freedom
  (18 observations deleted due to missingness)
Multiple R-squared:  0.002245,  Adjusted R-squared:  -0.0001192
F-statistic: 0.9496 on 1 and 422 DF,  p-value: 0.3304
```

Hide

```
# "underweight" is not linear because the accuracy is in -ve

model2 = lm(data = df, Y ~ BMILEVEL == "unhealthy")
summary(model2)
```

```
Call:
lm(formula = Y ~ BMILEVEL == "unhealthy", data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-161.34  -58.96  -10.21   51.29  177.29

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 133.711      3.877  34.485   <2e-16 ***
BMILEVEL == "unhealthy"TRUE   79.633      8.024   9.924   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 69.9 on 422 degrees of freedom
  (18 observations deleted due to missingness)
Multiple R-squared:  0.1892,    Adjusted R-squared:  0.1873
F-statistic: 98.49 on 1 and 422 DF,  p-value: < 2.2e-16
```

Hide

```
model3 = lm(data = df, Y ~ BMILEVEL == "healthy")
summary(model3)
```

```
Call:
lm(formula = Y ~ BMILEVEL == "healthy", data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-154.853  -46.495   -7.853   52.874  171.624

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                185.853      4.386   42.37   <2e-16 ***
BMILEVEL == "healthy"TRUE  -76.477      6.623  -11.55   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 67.67 on 422 degrees of freedom
  (18 observations deleted due to missingness)
Multiple R-squared:  0.2401,    Adjusted R-squared:  0.2383
F-statistic: 133.3 on 1 and 422 DF,  p-value: < 2.2e-16
```

Hide

```
model4 = lm(data = df, Y ~ BMILEVEL == "overweight")
summary(model4)
```

```
Call:
lm(formula = Y ~ BMILEVEL == "overweight", data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-136.25  -62.17  -13.67   57.06  200.83

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                   145.167      4.541  31.965  < 2e-16 ***
BMILEVEL == "overweight"TRUE   22.088      7.989   2.765  0.00595 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 76.94 on 422 degrees of freedom
  (18 observations deleted due to missingness)
Multiple R-squared:  0.01779,   Adjusted R-squared:  0.01546
F-statistic: 7.644 on 1 and 422 DF,  p-value: 0.005947
```

Hide

```
continuous_cols_nonBS = c(colnames(df)[1], colnames(df)[2], colnames(df)[3], colnames(df)[4], colnames(df)[11])
continuous_cols_nonBS
```

```
[1] "AGE"    "GENDER" "BMI"    "BP"     "Y"
```

Hide

```
# 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 rela
tionship with disease progression.
# Fit a model using BMILEVEL and the blood serum measurements identified in the previous question and compare its accuracy w
ith the model fit using BMILEVEL and all blood serum measurements.

continuous_cols_BS = setdiff(continuous_cols, continuous_cols_nonBS )
continuous_cols_BS
```

```
[1] "S1" "S2" "S3" "S4" "S5" "S6"
```

Hide

```
model = lm(data = df, Y~ BMILEVEL +S1 + S2 + S3 + S4 + S5 + S6 )
summary(model)
```

```
Call:
lm(formula = Y ~ BMILEVEL + S1 + S2 + S3 + S4 + S5 + S6, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-169.483  -42.613   -0.927   40.464  160.139

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -223.8364    71.2892  -3.140  0.00181 **
BMILEVELoverweight    32.1152     7.0701   4.542 7.31e-06 ***
BMILEVELunderweight  -40.8875    41.0469  -0.996  0.31978
BMILEVELunhealthy     60.3614     8.2501   7.316 1.33e-12 ***
S1                    -1.2456     0.6124  -2.034  0.04258 *
S2                     0.9399     0.5677   1.655  0.09858 .
S3                     0.5451     0.8323   0.655  0.51284
S4                     3.3245     6.3330   0.525  0.59990
S5                    81.6985    16.7140   4.888 1.46e-06 ***
S6                     0.6500     0.2835   2.293  0.02235 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 57.45 on 414 degrees of freedom
  (18 observations deleted due to missingness)
Multiple R-squared:  0.4628,    Adjusted R-squared:  0.4511
F-statistic: 39.62 on 9 and 414 DF,  p-value: < 2.2e-16
```

Hide

```
model1 = lm(data = df, Y ~ (BMILEVEL == "underweight") +S1 + S2 + S3 + S4 + S5 + S6)
summary(model1)
```

```
Call:
lm(formula = Y ~ (BMILEVEL == "underweight") + S1 + S2 + S3 +
    S4 + S5 + S6, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-141.157  -44.006   -2.377   41.006  170.270

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -288.8693    75.0090  -3.851 0.000136 ***
BMILEVEL == "underweight"TRUE  -65.5003    43.4003  -1.509 0.132003
S1                              -1.3864     0.6489  -2.137 0.033221 *
S2                               1.1953     0.5990   1.996 0.046637 *
S3                               0.1863     0.8818   0.211 0.832787
S4                              -0.0813     6.6905  -0.012 0.990311
S5                             100.3431    17.4748   5.742 1.81e-08 ***
S6                               1.0007     0.2965   3.376 0.000806 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.98 on 416 degrees of freedom
  (18 observations deleted due to missingness)
Multiple R-squared:  0.3918,    Adjusted R-squared:  0.3815
F-statistic: 38.28 on 7 and 416 DF,  p-value: < 2.2e-16
```

Hide

```
# "underweight" is not linear because the accuracy is in -ve

model2 = lm(data = df, Y ~ (BMILEVEL == "unhealthy") +S1 + S2 + S3 + S4 + S5 + S6)
summary(model2)
```

```
Call:
lm(formula = Y ~ (BMILEVEL == "unhealthy") + S1 + S2 + S3 + S4 +
    S5 + S6, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-168.254  -42.117   -3.572   40.726  176.447

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -257.6399    72.5196  -3.553 0.000425 ***
BMILEVEL == "unhealthy"TRUE   42.4000     7.3584   5.762 1.62e-08 ***
S1                            -1.3621     0.6252  -2.179 0.029915 *
S2                             1.1863     0.5773   2.055 0.040528 *
S3                             0.4234     0.8517   0.497 0.619418
S4                             0.4425     6.4564   0.069 0.945387
S5                            92.3607    16.8673   5.476 7.55e-08 ***
S6                             0.7605     0.2893   2.628 0.008895 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 58.84 on 416 degrees of freedom
  (18 observations deleted due to missingness)
Multiple R-squared:  0.4336,    Adjusted R-squared:  0.4241
F-statistic:  45.5 on 7 and 416 DF,  p-value: < 2.2e-16
```

Hide

```
model3 = lm(data = df, Y ~ (BMILEVEL == "healthy") +S1 + S2 + S3 + S4 + S5 + S6)
summary(model3)
```

```
Call:
lm(formula = Y ~ (BMILEVEL == "healthy") + S1 + S2 + S3 + S4 +
    S5 + S6, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-152.138  -43.548   -0.732   41.891  152.115

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               -185.0646    73.7363  -2.510  0.01246 *
BMILEVEL == "healthy"TRUE  -41.0928     6.6192  -6.208 1.30e-09 ***
S1                          -1.1152     0.6224  -1.792  0.07392 .
S2                           0.7876     0.5769   1.365  0.17287
S3                           0.3417     0.8459   0.404  0.68644
S4                           3.5042     6.4456   0.544  0.58696
S5                          80.0677    16.9999   4.710 3.38e-06 ***
S6                           0.7966     0.2864   2.782  0.00565 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 58.5 on 416 degrees of freedom
  (18 observations deleted due to missingness)
Multiple R-squared:  0.4403,    Adjusted R-squared:  0.4309
F-statistic: 46.75 on 7 and 416 DF,  p-value: < 2.2e-16
```

Hide

```
model4 = lm(data = df, Y ~ (BMILEVEL == "overweight") +S1 + S2 + S3 + S4 + S5 + S6)
summary(model4)
```

```
Call:
lm(formula = Y ~ (BMILEVEL == "overweight") + S1 + S2 + S3 +
    S4 + S5 + S6, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-137.937  -44.305   -2.281   41.953  166.072

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -280.4752    75.2335  -3.728 0.000220 ***
BMILEVEL == "overweight"TRUE    7.2010     6.5420   1.101 0.271643
S1                             -1.2935     0.6495  -1.991 0.047093 *
S2                              1.0862     0.6019   1.805 0.071861 .
S3                              0.1283     0.8823   0.145 0.884464
S4                              0.3139     6.7174   0.047 0.962756
S5                             96.7194    17.5598   5.508 6.37e-08 ***
S6                              1.0234     0.2965   3.451 0.000616 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 61.06 on 416 degrees of freedom
  (18 observations deleted due to missingness)
Multiple R-squared:  0.3902,    Adjusted R-squared:  0.3799
F-statistic: 38.03 on 7 and 416 DF,  p-value: < 2.2e-16
```

Hide

```
# 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 ~ BMI+AGE+ BP+ GENDER)
summary(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
```

Hide

```
model1 = lm(data = df, Y ~ BMI+AGE+ BP+ (GENDER==1) + (GENDER == 2))
summary(model1)
```

```

Call:
lm(formula = Y ~ BMI + AGE + BP + (GENDER == 1) + (GENDER ==
    2), data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-152.417  -43.576   -3.757   42.938  150.054

Coefficients: (1 not defined because of singularities)
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -219.3875    23.9934  -9.144  < 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 ***
GENDER == 1TRUE  10.1590     5.9219   1.716    0.087 .
GENDER == 2TRUE        NA         NA      NA       NA
---
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
```

Hide

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

model1 = lm(data = df, Y ~ BMI + AGE + BP + GENDER + (BMI * GENDER))
summary(model1)
```

```

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
```