Discrete analysis_HW2

ID : 111024517 Name : 鄭家豪

The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study on 768 adult female Pima Indians living near Phoenix. The following variables were recorded:

- `pregnant`: Number of times pregnant,
- `glucose`: Plasma glucose concentration a 2 hours in an oral glucose tolerance test,
- `diastolic`: Diastolic blood pressure (mm Hg),
- `triceps`: Triceps skin fold thickness (mm),
- `insulin`: 2-Hour serum insulin (mu U/ml),
- `bmi`: Body mass index (weight in kg/(height in $m2$)),
- `diabetes`: Diabetes pedigree function,
- `age`: Age (years),
- `test`: a test whether the patient shows signs of diabetes (coded 0 if negative, 1 if positive).

The purpose of the study was to investigate factors related to diabetes. The data can be found in the dataset pima.

1. Perform simple graphical and numerical summaries of the data. Can you find any obvious irregularities in the data? If you do, take appropriate steps to correct the problems.
2. Fit a model with the result of the diabetes test as the response and all the other variables as predictors. Can you tell whether this model fits the data?
3. What is the difference in the odds of testing positive for diabetes for a woman with a BMI at the first quartile compared with a woman at the third quartile, assuming that all other factors are held constant? Give a confidence interval for this difference.
4. Do women who test positive have higher diastolic blood pressures? Is the diastolic blood pressure significant in the model? Explain the distinction between the two questions and discuss why the answers are only apparently contradictory.

Sol.

1.

```
data = read.table("pima.txt",header = T)
data[,9] <- as.factor(data[,9])
summary(data)
```

```
    pregnant         glucose        diastolic         triceps
 Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
 Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
 Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
 3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
    insulin           bmi           diabetes           age        test
 Min.   :  0.0   Min.   : 0.00   Min.   :0.0780   Min.   :21.00   0:500
 1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00   1:268
 Median : 30.5   Median :32.00   Median :0.3725   Median :29.00
 Mean   : 79.8   Mean   :31.99   Mean   :0.4719   Mean   :33.24
 3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
 Max.   :846.0   Max.   :67.10   Max.   :2.4200   Max.   :81.00
```

Here, the file pima.txt is imported into the R console and assigned as “data”. Upon executing the summary function on it, the followings are observed:

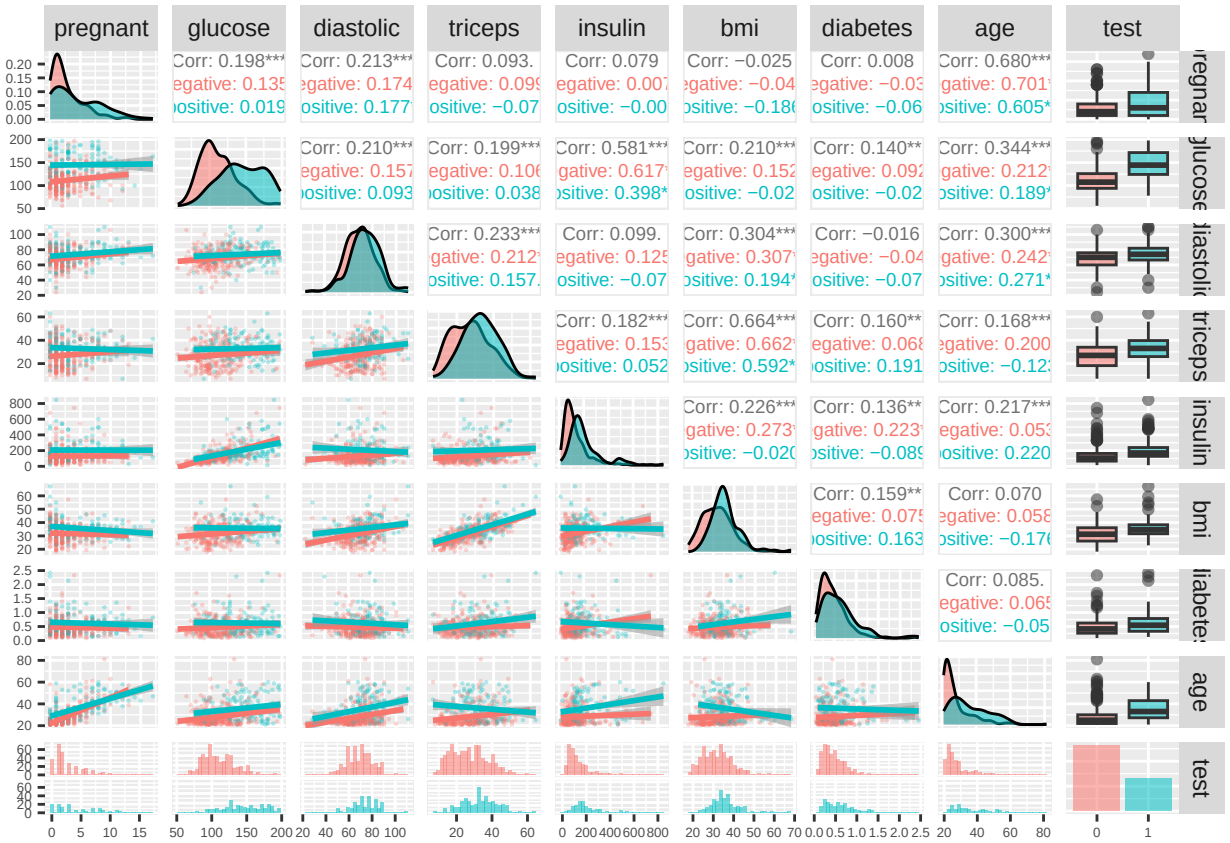
• The proportion of individuals with diabetes to those without diabetes is approximately 1:2, based on a total of 768 observations.
• Variables such as glucose, diastolic, triceps, insulin, and bmi contain zero values. Considering their meanings, it is implausible for these variables to be zero. Therefore these zero values may be represented missing value.

```
miss.index = which( apply(data[,c("glucose","diastolic","triceps","insulin","bmi")],
                          1, prod ) == 0 )
data1 = data[-miss.index,]
summary(data1)
```

```
    pregnant         glucose        diastolic         triceps
 Min.   : 0.000   Min.   : 56.0   Min.   : 24.00   Min.   : 7.00
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.:21.00
 Median : 2.000   Median :119.0   Median : 70.00   Median :29.00
 Mean   : 3.301   Mean   :122.6   Mean   : 70.66   Mean   :29.15
 3rd Qu.: 5.000   3rd Qu.:143.0   3rd Qu.: 78.00   3rd Qu.:37.00
 Max.   :17.000   Max.   :198.0   Max.   :110.00   Max.   :63.00
    insulin            bmi           diabetes           age        test
 Min.   : 14.00   Min.   :18.20   Min.   :0.0850   Min.   :21.00   0:262
 1st Qu.: 76.75   1st Qu.:28.40   1st Qu.:0.2697   1st Qu.:23.00   1:130
 Median :125.50   Median :33.20   Median :0.4495   Median :27.00
 Mean   :156.06   Mean   :33.09   Mean   :0.5230   Mean   :30.86
 3rd Qu.:190.00   3rd Qu.:37.10   3rd Qu.:0.6870   3rd Qu.:36.00
 Max.   :846.00   Max.   :67.10   Max.   :2.4200   Max.   :81.00
```

After removing observations with missing values coded as 0, resulting in 392 remaining observations, roughly half of the original data, the new dataset is named as data1. Subsequently, the ggpairs function from the "GGally" package is used to explore the numerical and graphical characteristics of this data:

```
library(GGally)
library(MASS)
ggpairs(data1, mapping=aes( color = factor(test, labels=c("negative","positive")), alpha=0.2),
        lower = list(continuous = wrap("smooth",alpha=0.3, size = 0.1)) ,
        upper = list(continuous = wrap("cor", size = 2.5))) +
  theme(axis.text = element_text(size = 5))
```



The followings are observed:

• The distributions of variables pregnant, insulin, diabetes, and age appear to be right-skewed.
• For Boxplots, reveal the presence of outliers in all variables.
• Correlation analysis indicates that the correlation coefficients for the pairs (age, pregnant), (insulin, glucose), and (bmi, triceps) all exceed 0.5.

For subsequent analysis, the outliers will be retained as there is no sufficient reason to remove them.

2.

Use data1 to fit the following model:
$\text{logit}(p_x) \sim 1+$ pregnant + glucose + diastolic + triceps + insulin + bmi + diabetes + age, where $p_x = test_x/n_x (n_x = 1 : \text{sparse})$.

```
model1 = glm(test~ ., family = binomial, data=data1)
fit1 = summary(model1)
fit1
```

```
Call:
glm(formula = test ~ ., family = binomial, data = data1)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.004e+01  1.218e+00  -8.246  < 2e-16 ***
pregnant     8.216e-02  5.543e-02   1.482  0.13825
glucose      3.827e-02  5.768e-03   6.635 3.24e-11 ***
diastolic   -1.420e-03  1.183e-02  -0.120  0.90446
triceps      1.122e-02  1.708e-02   0.657  0.51128
```

```
insulin     -8.253e-04  1.306e-03  -0.632  0.52757
bmi          7.054e-02  2.734e-02   2.580  0.00989 **
diabetes     1.141e+00  4.274e-01   2.669  0.00760 **
age          3.395e-02  1.838e-02   1.847  0.06474 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 498.10  on 391  degrees of freedom
Residual deviance: 344.02  on 383  degrees of freedom
AIC: 362.02

Number of Fisher Scoring iterations: 5
```

Based on the model fitting results mentioned above, significant variables include glucose, BMI, diabetes, and age, indicating their significant effects on the diabetes. However, it is important to note that due to the collinearity mentioned in the EDA, variables such as pregnant, insulin, and triceps may not be significant.
Since the data hasn't been worked grouping, that is, treating each test observation as a Bernoulli variable results in sparse data, making the deviance-based test results unreliable. Therefore, performing the Hosmer-Lemeshow test is to assess whether the model is suitable:

```
library(glmtoolbox)
hltest = hltest(model1, verbose = F)
hltest[["p.value"]]
```

```
[1] "5.618061e-06"
```

```
as.numeric(hltest[["p.value"]]) < 0.05
```

```
[1] TRUE
```

The p-value by Hosmer-Lemeshow test is less than 0.05, which means that the fitting model is appropriate.

3.

The first and third quartiles of BMI are:

```
q3q1 = quantile(data1$bmi,probs = c(0.25,0.75))
q3q1
```

```
 25%  75%
28.4 37.1
```

From the results of fit1, calculate the odds difference of bmi's Q1 and Q3:

```
exp( fit1$coefficients["bmi",1] * (37.1 - 28.4) )
```

```
[1] 1.847211
```

That is, the odds of bmi's Q3 are about 1.8472 times higher than the odds of Q1.

Because this data is considered sparse data, the profile likelihood-based method is used to build 95% confidence interval for bmi's odds difference between Q3 and Q1:

```
pro_li_CI <- exp( confint(model1) * (q3q1[2] - q3q1[1]) )
pro_li_CI["bmi",]
```

```
   2.5 %   97.5 %
1.166174 2.975717
```

The 95% confidence interval for this difference is $[1.166174, 2.975717]$.

4.

For question 1, stating whether "test:1" is associated with higher "diastolic", we can establish a model: "diastolic ~ 1 + test" to examine the significance of the coefficient for test:

```
question1 = lm(diastolic ~ test,data =data1)
summary(question1)
```

```
Call:
lm(formula = diastolic ~ test, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-44.969  -8.077   1.031   7.923  37.031

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  68.9695     0.7585  90.927  < 2e-16 ***
test1         5.1075     1.3172   3.878 0.000124 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.28 on 390 degrees of freedom
Multiple R-squared:  0.03712,   Adjusted R-squared:  0.03465
F-statistic: 15.04 on 1 and 390 DF,  p-value: 0.0001237
```

Based on the above, it indicates that patients with positive diabetes have higher diastolic compared to those with negative diabetes.

For question 2, stating whether diastolic and other variables have a significant effect on the test outcomes, we found in the second part that diastolic does not have a significant effect. However, this might be due to collinearity. Therefore, this is different from the question considered in question 1, and thus, there is no contradiction.