Data Analytics
Assignment -3

Name : R Siva Girish SRN : PES1201700159

Dataset : Pima Indians Diabetes Database

Dataset Link:https://www.kaggle.com/uciml/pima-indians-diabetes-database

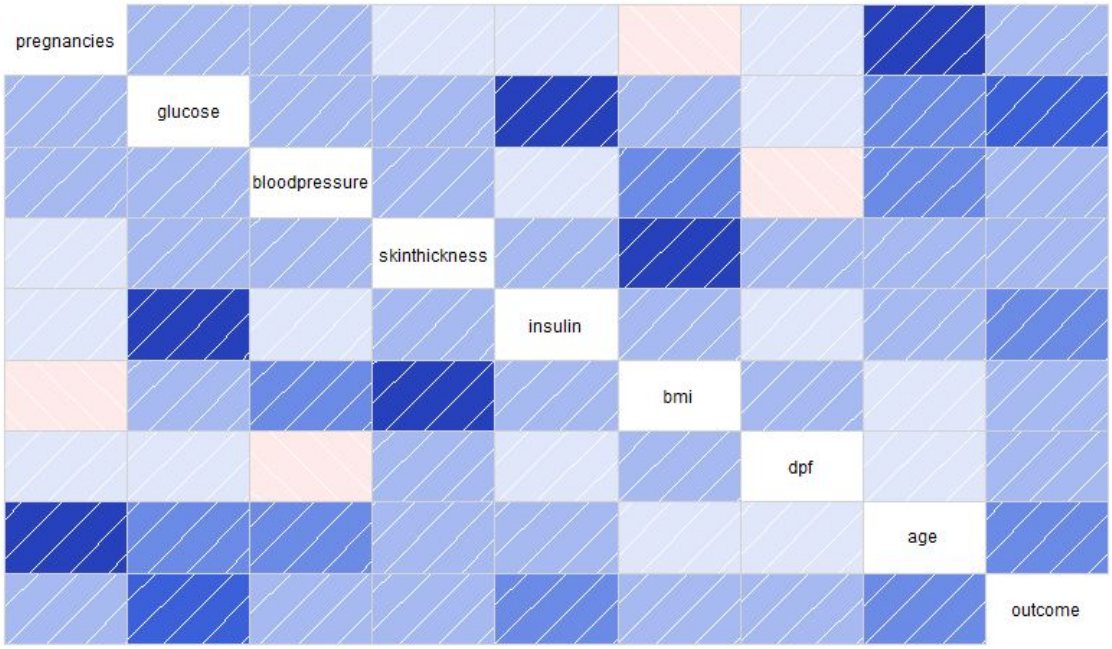
- ❖ About the Dataset
  - ➢ Dataset contain 768 rows and 9 columns

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

❖ Multivariate Analysis

- ➢ Multivariate Analysis deals with the statistical analysis of data collected on more than one dependent variable.
- ➢ These variables may be correlated with each other, and their statistical dependence is often taken into account when analyzing such data.
- ➢ While performing multivariate analysis on variance(MANOVA) we get important parameters known as pillai score.
- ➢ Based on the Significance of the pillai score we can either choose to keep the independent variable or not.
- ➢ Usually whenever we eliminate a variable from our model based on it's pillai score we tend to observe an increase in the value of adjusted R-Square.
- ➢ If that's not the case then the multivariate regression model is not very useful.

❖ Correlation Plot



❖ Analysis

➢Dependent Variables are : Outcome , Age

➢Independent Variables : Pregnancies, glucose, bloodpressure, skinthickness, insulin, bmi, age, dpf.

Code For creating model :

#outcome and age

```
mymodel<-lm(cbind(diabetes$outcome,diabetes$age) ~ diabetes$pregnancies +
diabetes$bloodpressure + diabetes$skinthickness + diabetes$insulin + diabetes$bmi +
diabetes$dpf + diabetes$glucose , data = diabetes)
summary(mymodel)
coef(mymodel)
```

```
Call:
lm(formula = `diabetes$outcome` ~ diabetes$pregnancies + diabetes$bloodpressure +
    diabetes$skinthickness + diabetes$insulin + diabetes$bmi +
    diabetes$dpf + diabetes$glucose, data = diabetes)

Residuals:
     Min       1Q   Median       3Q      Max
-1.14441 -0.25791 -0.06931  0.26064  1.04117

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            -1.045e+00  1.416e-01  -7.383 9.71e-13 ***
diabetes$pregnancies    2.432e-02  6.426e-03   3.784 0.000179 ***
diabetes$bloodpressure  6.989e-04  1.711e-03   0.409 0.683132
diabetes$skinthickness  2.051e-03  2.527e-03   0.812 0.417369
diabetes$insulin       -9.214e-05  2.049e-04  -0.450 0.653184
diabetes$bmi            8.876e-03  3.912e-03   2.269 0.023839 *
diabetes$dpf            1.660e-01  5.815e-02   2.856 0.004529 **
diabetes$glucose        6.699e-03  8.077e-04   8.294 1.87e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.387 on 384 degrees of freedom
Multiple R-squared:  0.3382,    Adjusted R-squared:  0.3261
F-statistic: 28.03 on 7 and 384 DF,  p-value: < 2.2e-16


Response diabetes$age :

Call:
lm(formula = `diabetes$age` ~ diabetes$pregnancies + diabetes$bloodpressure +
    diabetes$skinthickness + diabetes$insulin + diabetes$bmi +
    diabetes$dpf + diabetes$glucose, data = diabetes)

Residuals:
    Min      1Q  Median      3Q     Max
-14.837  -3.921  -1.169   2.334  38.001

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             9.730955   2.581314   3.770 0.000189 ***
diabetes$pregnancies    1.933057   0.117140  16.502  < 2e-16 ***
diabetes$bloodpressure  0.109602   0.031186   3.514 0.000493 ***
diabetes$skinthickness  0.063605   0.046058   1.381 0.168084
diabetes$insulin        0.005308   0.003735   1.421 0.156063
diabetes$bmi           -0.076428   0.071312  -1.072 0.284513
diabetes$dpf            1.506471   1.059927   1.421 0.156042
diabetes$glucose        0.049471   0.014723   3.360 0.000857 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.054 on 384 degrees of freedom
Multiple R-squared:  0.5304,    Adjusted R-squared:  0.5218
F-statistic: 61.96 on 7 and 384 DF,  p-value: < 2.2e-16
```

R - Square Values are around 33% for Outcome and 50% for Age

This is not a very good regression model.

Code For Multivariate Analysis on Variance:

```
Mymodelfit<-manova(mymodel)
summary(Mymodelfit)
```

```
                          Df  Pillai approx F num Df den Df    Pr(>F)
diabetes$pregnancies       1 0.50684  196.815      2    383 < 2.2e-16 ***
diabetes$bloodpressure     1 0.07064   14.555      2    383 8.080e-07 ***
diabetes$skinthickness     1 0.06625   13.588      2    383 1.990e-06 ***
diabetes$insulin           1 0.10585   22.671      2    383 4.953e-10 ***
diabetes$bmi               1 0.02178    4.265      2    383  0.014730 *
diabetes$dpf               1 0.03444    6.831      2    383  0.001217 **
diabetes$glucose           1 0.16335   37.388      2    383 1.471e-15 ***
Residuals                384
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Based on the pillai scores we attempt to remove bmi from our multivariate regression model as it has the lowest significant pillai score .

Code For creating model :

```
#Remove body mass index
mymodel<-lm( cbind ( diabetes$outcome , diabetes$age ) ~ diabetes$pregnancies + diabetes$bloodpressure + diabetes$skinthickness + diabetes$insulin + diabetes$dpf + diabetes$glucose , data = diabetes)
summary(mymodel)
coef(mymodel)
```

Upon Removing the body mass index from our model we notice the following Adjusted R- Square Values

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.19276 -0.25492 -0.06514  0.26317  1.03742

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           -9.280e-01  1.325e-01  -7.003 1.12e-11 ***
diabetes$pregnancies   2.177e-02  6.362e-03   3.422 0.000689 ***
diabetes$bloodpressure 1.595e-03  1.674e-03   0.953 0.341135
diabetes$skinthickness 5.650e-03  1.978e-03   2.857 0.004510 **
diabetes$insulin      -4.384e-05  2.049e-04  -0.214 0.830680
diabetes$dpf           1.752e-01  5.832e-02   3.003 0.002846 **
diabetes$glucose       6.733e-03  8.119e-04   8.293 1.88e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3891 on 385 degrees of freedom
Multiple R-squared:  0.3293,    Adjusted R-squared:  0.3189
F-statistic: 31.51 on 6 and 385 DF,  p-value: < 2.2e-16


Response diabetes$age :

Call:
lm(formula = `diabetes$age` ~ diabetes$pregnancies + diabetes$bloodpressure +
    diabetes$skinthickness + diabetes$insulin + diabetes$dpf +
    diabetes$glucose, data = diabetes)

Residuals:
    Min      1Q  Median      3Q     Max
-15.194  -3.828  -1.152   2.286  38.529

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           8.719580   2.403097   3.628 0.000324 ***
diabetes$pregnancies  1.954991   0.115360  16.947  < 2e-16 ***
diabetes$bloodpressure 0.101885   0.030349   3.357 0.000866 ***
diabetes$skinthickness 0.032621   0.035861   0.910 0.363570
diabetes$insulin      0.004892   0.003715   1.317 0.188709
diabetes$dpf          1.428077   1.057604   1.350 0.177715
diabetes$glucose      0.049182   0.014724   3.340 0.000919 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.055 on 385 degrees of freedom
Multiple R-squared:  0.529,     Adjusted R-squared:  0.5216
F-statistic: 72.06 on 6 and 385 DF,  p-value: < 2.2e-16
```

Conclusion:

Based on the above stats and the stats observed previously we notice that eliminating a variable from our model based on significance of pillai score is not giving us proper results.Hence a multivariate regression model in this case is unable to solve our problem.