DIABETES DATA

QN. Describe the relationships between the variables.

RELATIONSHIP BETWEEN VARIABLES

STRONG CORRELATIONS (|r|≥ 0.70)

1. Extremely strong Positive Correlation

S1 and S2 (r=.90): These two blood serum measurements are almost perfectly linear related. This suggests they measure very similar underlying biological factors, making redundant in regression model later.

2. Very Strong Negative Correlation

S3 and S4 (r=-0.74): This is a strong inverse relationship. As S3 increases, S4 decreases substantially. This could represent complementary measures in lipid profiles.

MODERATE CORRELATIONS (0.40≤ |r| <0.70):

Positive Relationship

BMI and S5 (r=0.45): Body mass index is moderately collerated with S5 , this suggest weight influences triglyceride levels

II. BMI and S4 (r=0.41):BMI modertely correlates with S4

III. S2 and S4 (r=0.66): Strong moderate correlation

IV. S1 and S4 (r=0.54): Moderate corelation

V. BP and S5 (r=0.39): Weak moderate corelation

WEAK CORRELATION (0.20≤ |r| <0.40):

Positive:

I. AGE and BP (r=0.34): Blood pressure tends to increase slightly with age

II. AGE nad S6 (r=0.30): Age weakly correlates with S6

III. BMI nd BP(r=0.40): Bordeline moderate

IV. Most variables with AGE: AGE shows weak positive correlations with nearly variables except S3

Negative:

I. S3 shows weak negative correlations with several other variables (-0.08 to -0.40 range)

VERY WEAK/NEGLIGIBLE CORRLATIONS (|r| <0.20):

I. SEX nd S1 (r=0.04): essentially no relationship

II. AGE and S3 (r=-0.08) minimal relationship

PATTERNS IMPLICATIONS

1. BMI cluster: BMI correlates moderately with BP,S4 and S5, suggesting metabolic syndrome patterns.

2. Lipid profile raltionship: The S variables (blood serum measurements) shows expected interralationships typicalmof lipid meatbolism

3. Age effect: AGE shows the strongest correlation with BP, consistent with known physiology.

QN. What is collinearity? What effect does collinearity amongst predictor features/ variables have on their estimated coefficient value?

Collinearity refers to the statistical phenomenon whereby two or more predictor variables in regression model are highly linearly collarated.

Example in our dataset; S1 and S2 shows perfect multicollinearity symptoms with r=0.90

also S3 and S4 shows high negative collinearity with r= -0.74

Effect OF collinearity on estimated coefficient value

-Collinearity or multicollinearity among the predictor variables can lead to unstable and unreliable estimated coefficients in regression analysis, making it difficult to interpret the individual effects of each predictor.

QN. Are all variables significant? Could this be a problem of collinearity?

As we saw the results, Yes this is definetly a problem of collinearity, because of the following reason.

Only 4/10 varibales are significant

- if variabkes were indeendent means more should expected to shiw significance
- collinearity sometimes makes hard to seoarate individual effects

Forward selection Start with n variables and adds predictors iteratively one at time WHILE Backward selection starts with all variables and removes insignificant predictors

Stepwise approach is a statistical method used to select the most important variabales for regression model by add or remiving predictord step by step.

so how it works here are the steps by step

- Define the criterion for selection such as p-values, adjusted R^2 and AIC.
- Forwad selection
  - starts with no varibakes , it begin with an empty model
- Backward elimination
  -starts with all varibakes, fit the model using all the predictors. This works well when the number of variables is managable.