INFERENTIAL MODELLING – HDS 5320-02

ASSIGNMENT – 02

Vaishnavi Gannavaram

Instructor: Dr. Srikanth Mudigonda

Introduction

The data set used in this assignment is “Diabetes” which is already present in the “faraway” package. There are 10 variables in the data set. These Variables have been cleaned by dropping NA values using “na.omit” function. Out of those the variables, "chol","stab.glu","hdl" "ratio","glyhb","location", "age", "gender","height", "weight", "waist", "hip" are the variables that we have considered for this analysis.

Plotting Linear Regression models is the next step that I have taken to determine whether some of the variables are significant predictors for the outcome variable “glyhb” and to show the amount of variance accounted by each variable.

Model 1:

This is the single-predictor model where “gender” is the predictor and the ‘Glycosylated hemoglobin” is the outcome variable. This step is done to determine whether males and females differ significantly on their glycosylated hemoglobin values.

In this model, the probability of observing an F-statistic value as 0.01568 for the 1 predictor and the observation 1 by chance is 0.9006, Therefore, we can infer that model 1 is not reliable. It means “gender” cannot be a significant predictor for “glycosylated hemoglobin.”

The adjusted R square value is -0.007689. The p-value obtained for this model is >0.05. Therefore, “gender” cannot be statistically significant predictor variable for ‘glycosylated hemoglobin.’

Model 2:

This is the two-predictor model where “gender and location” are the predictors and the ‘Glycosylated hemoglobin” is the outcome variable. This step is done to determine whether there are any significant differences in glycosylated hemoglobin across locations, after taking gender into account.

In this model, the probability of observing an F-statistic value as 1.813 for 2 predictors and the observation 2 by chance is 0.1674, Therefore, we can infer that model 2 is not reliable. It means “gender and location model” cannot be a significant predictor for “glycosylated hemoglobin.”

The adjusted R square value is 0.01244. Here this model gives 1.244 variability. Here we can also infer that the other variables can be better predictors than this one because there is 98.756 of variability left to predict.

The p-value obtained for this model is >0.05. Therefore, “gender and location model” cannot be statistically significant predictor variable for ‘glycosylated hemoglobin.’

Model 3:

This is the model where “cholesterol, stabilized glucose, HDL, cholesterol/HDL ratio, age, and weight/height ratio” are the predictors and the

‘Glycosylated hemoglobin” is the outcome variable. This step is done to determine if they are significant predictors for glycosylated hemoglobin.

In this model, the probability of observing an F-statistic value as 49.23 for 6 predictors and the observation 7 by chance is < 2.2e-16, Therefore, we can infer that model 3 is reliable. Also, from all the predictors only stab.glu and age can be significant predictors for “glycosylated hemoglobin” as remaining predictors have p-value >0.05.

The adjusted R square value is 0.7235. Here this model gives 72.35 variability. Here we can also infer that this model can be a significant model because there is only 27.65% variability left to predict.

The p-value obtained for this model is <0.05. Therefore, this model 3 is statistically significant.

Q4

After considering all the predictors listed above, does the effect of weight/height ratio vary significantly between the two genders?

Yes, the effect of weight/height ratio varies significantly between genders because the coefficient value got increased by 0.629 after adding weight/height ratio to gender.

MODEL 4:

In this model, the probability of observing an F-statistic value as 43.4 for 8 predictors and the observation 8 by chance is < 2.2e-16, The F statistic value got decreased slightly than previous model.

The adjusted R square value is 0.7245. Here this model gives 72.45 variability. The p-value obtained for this model is >0.05. Therefore, this model can be statistically significant.

MODEL 5:

In this model, the probability of observing an F-statistic value as 37.99 for 9 predictors and the observation 9 by chance is < 2.2e-16. The adjusted R square value is 0.7207. Here this model gives 72.07 variability. Here we can also infer that the other variables can be better predictors than this one. The p-value obtained for this model is < 2.2e-16.

Q5: To model 3, add waist/hip ratio as a predictor and explain whether this is a significant predictor of glycosylated hemoglobin, after accounting for the variance accounted by all the predictors in model 3.

The variance decreased after adding waist/hip ratio as a predictor to the model 3. Hence, this is not a significant predictor.

Q6: ANOVA TEST:

The Anova test is done to identify the best model in terms of statistical and practical significance among the five models.

Here, we can clearly see that the fourth model is the best fit among all the other models as it has 216.98 Residual sum of squares with f statistic being highest, and it also has the least p-value which is nearer to zero.

Q7: **Table depicting assumptions**

Assumption	**Evidence**	**Assumption met or not**
Multi collinearity	The Variance Inflation Factor (VIF) values for each variable in the model must be as small as possible, with no VIF value > 10,	Not met, Because weight and weight/height variables have VIF value more than 10.
Independence of errors	The Durbin-Watson test. Values of this statistic close to 2.0 indicate that this condition is most likely to have been met. Values less than 1.0 and greater than 3.0 imply that there is a “cause for concern”	Met. Because the D-W statistic value is 1.796821. Which is close to 2.0. Hence, this assumption is met.

Residuals normal distribution	Shapiro’s test	P value: 0.0009318. Not normally distributed. Therefore, not met.

Q8 and Q9:

Logs and Square root transformations:

The models with hdl and glyhb, height and glyhb, weight and glyhb are closely linear. Whereas other models have curvatures which says that they are nonlinear even after transforming them into squares and logs.

Therefore, the desired change in relation has not occurred to any of the variables after transformation. Hence moving forward with further analysis.

Weighted linear regression transformation:

Weighed each observation by group variance value. The grouping is performed using a new categorical variable that is created to map specific values of outcome variable glycosylated hemoglobin into three levels: low, med, and high. This mapping is based on whether the value of glycosylated hemoglobin is less than the first quartile boundary, falls between the first and third quartile boundary, or exceeds the third quartile boundary.

“Lm.low” is the model with low subsets. Where the “age” is significant predictors present with p value less than 0.05. “Lm.med” is the model with med subsets where stab.glu is the significant predictor with p value less than 0.05.

Therefore, in each of the subgroups the nature of the relationship is different.

Computing the group variance

Here I used the function “aggregate” to aggregate the values of outcome variable “glycosylated hemoglobin” based on the categorization that we have performed.

High	NA
Low	0.3220177
Med	5.1039976

Mentioned above are the group variances obtained.

Weighted Least Squares:

To provide baseline for the new models, weighted least squares are used on the same predictors that are used in “lmmodel.3”.

Here, it is noticeable that only “stab. glu” is the only significant predictor with normal p value.

From the summary of lm.new model, we can see that adjusted r square value is 0.6152. It means that the variance accounted by this model lm.new is 61.52%.

Investigating whether the weight least squares modelling approach has fixed the issue of heteroscedastic residuals, by plotting model residuals against fitted values is the next step that I have performed. We can observe the nonlinear trend in the distribution of residuals across the model predicted.

Therefore, the outcome variable is transformed to see whether using the transformed version as the outcome variable, in lieu of original outcome variable yields a better distribution model.

Although the outcome was transformed it was a nonlinear distribution as the shapiro test's p-value is abnormal. (W = 0.98579, p-value = 0.9087)

Used both approaches of ordinary least squares and weighted least squares, with the outcome being transformed using log () function, they were compared, and it was noticed that shapiro ordinary least square model has normal p value.

Conclusion:

Few transformations were performed to fit models using ordinary and weight least squares approach to multiple linear regression. Despite all those, it is clear that "model 3 = best model" has failed to meet some assumptions.