Advanced Statistics Demo1

Michael Williams

The following is a statistical analysis of the CHOL data set. The data set contains the variables below.

Variable Name	Type	Description	Units
ID	Numeric	Subject ID	none
AGE	Numeric	Age	yrs
HT	Numeric	Height	in
WT	Numeric	Weight	lb
SBP	Numeric	Systolic blood pressure	mmHg
DBP	Numeric	Diastolic blood pressure	mmHg
HDL	Numeric	High density lipids	mmHg
GENDER	Character	'male' or 'female'	none
TG	Numeric	Triglyceride	mmHg
BMI	Numeric	Body mass index	lb/in^2

Part 1: The ``all possible subsets" method was used to build a predictive model with HDL as the outcome variable, and GENDER, AGE, HT, WT, SBP, and DBP as potential explanatory variables. The character variable GENDER was coded as 0 for 'male' and 1 for 'female'.

After running PROC REG for the model with the VIF option, it was found that none of the variables had a variation inflation factor of over 10. Therefore, no multi-collinearity among the explanatory variables was present. After considering all possible subsets of explanatory variables and corresponding values of Mallow's C_p statistic, the parsimonious model is the one with all six explanatory variables. The final model is

$$y = 26.03229 - 4.3254x_1 + 0.09137x_2 + 0.4537x_3 - 0.11367x_4 - 0.02598x_5 + 0.10393x_6$$

where x_1 corresponds to GENDER, x_2 corresponds to AGE, x_3 corresponds to HT, x_4 corresponds to WT, x_5 corresponds to SBP, and x_6 corresponds to DBP.

We now present an interpretation of the final model. The overall F-value is 4.75 which has a p-value of 0.0002. Therefore, at least one explanatory variable is associated with HDL. The R^2 value is 0.1348, so 13% of the variation in HDL is explained by its linear relationship with the explanatory variables. The parameter estimates tell us that

1. Males have a 4.325 mmHg decrease in high density lipids (HDL) compared to females. This association is statistically significant (p=0.0121) at the 0.05 level.
2. One year increase in age corresponds to a 0.09137 mmHg increase in HDL. This association is not statistically significant (p=0.2019) at the 0.10 level.
3. One inch increase in height corresponds to a 0.4537 mmHg increase in HDL. This association is statistically significant (p=0.0075) at the 0.05 level.

4. One lb increase in weight corresponds to a 0.11367 mmHg decrease in HDL. This association is statistically significant (p=0.0019) at the 0.05 level.
5. One mmHg increase in systolic blood pressure corresponds to a 0.02598 mmHg decrease in HDL. This association is statistically significant (p=0.0937) at the 0.10 level.
6. One mmHg increase in diastolic blood pressure corresponds to a 0.10393 mmHg increase in HDL. This association is statistically significant, albeit marginally (p=0.1389), at the 0.10 level.

Part 2: The objective is to assess whether or not GENDER is an effect modifier in studying the association between HDL (Y-variable) and AGE (X-variable). To test for a possible interaction effect, we used the regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

where y corresponds to HDL, x_1 corresponds to AGE, and x_2 corresponds to GENDER. There is a statistically significant interaction between AGE and GENDER; in the parameter estimate, the interaction term has a p-value of 0.0083 < 0.05. Since there is a significant interaction, we split the study into separate studies according to gender.

In the male study, we found that the overall association between HDL and age for males is marginally significant (p=0.1001) at the 0.10 level. The R^2 value is 0.028813; so 2.88% of of the variation in HDL for males is explained by its linear relationship with age. For each year increase in age, there is a 0.1265 decrease in HDL.

In the female study, we found that the overall association between HDL and age for females is significant (p=0.0347) at the 0.05 level. The R^2 value is 0.047056; so 4.7% of of the variation in HDL for females is explained by its linear relationship with age. For each year increase in age, there is a 0.149 increase in HDL.

Part 3: We studied the association between TG (Y-variable) and BMI (X-variable). We checked whether GENDER or AGE was a confounder.

First, we analyzed the simple regression model

$$y = \beta_0 + \beta_1 x_1$$

(where y is TG and x_1 is BMI) and found a parameter estimate of β_1=24.9316.

Including a covariate x_2 for GENDER in the regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

yields a parameter estimate of β_1=25.455. This is an increase of 2% (which is less than 10%), so GENDER is not a confounder.

Including an covariate x_2 for AGE in the regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

yields a parameter estimate of β_1=18.7484. This is a decrease of 24.8% (which is greater than 15%), so AGE is a confounder. Therefore, age needs to be controlled in the regression model.

The final model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where y corresponds to TG, x_1 corresponds to BMI, and x_2 corresponds to AGE. After controlling for the age, one lb/in^2 increase in BMI corresponds to an 18.7484 mmHg increase in TG. This association is statistically significant (p=0.0004) at the 0.05 level.