Advanced Statistics Demo2

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

The SAS data set CHOL_CATS is created by setting CHOL and adding categorical variables

- HDL_HI : value is 1 (if HDL > 47) and 0 otherwise
- AGE_HI: value is 1 (if AGE > 17) and 0 otherwise
- TG_HI : value is 1 (if TG > 68) and 0 otherwise
- BMI_HI : value is 1 (if BMI > 3.0718476) and 0 otherwise

Part 1A

We examine whether GENDER (Z-variable) modifies the association between HDL_HI (Y-variable) and AGE_HI (X-variable) using PROC FREQ. The Breslow Day Test (of homogeneity of the odds ratios) has a p-value of $p = 0.0192 < 0.10$, so there is sufficient evidence to reject homogeneity of the odds ratios. Therefore, GENDER does indeed modify the association between the Y and X variables, and the gender-specific odds ratios (from the Cochran-Mantel-Haenszel statistics) should be reported.

For males, the odds ratio is 0.4828 with a p-value of $p = 0.0866 < 0.10$, so there is a significant association between X and Y. The odds of high HDL (over 47 mmHg) is 0.4828 times smaller for older (over 17 years) males compared to younger males. The 95% confidence interval for the odds ratios is 0.2088 to 1.1161.

For females, the odds ratio is 1.9531 with a p-value of $p = 0.1103 > 0.10$, so there is not a significant association between X and Y. The odds of high HDL (over 47 mmHg) is 1.9531 times greater for older (over 17 years) females compared to younger females. The 95% confidence interval for the odds ratios is 0.8553 to 4.4601.

Part 1B

We examine whether GENDER (Z-variable) modifies the association between HDL_HI (Y-variable) and AGE_HI (X-variable) using PROC LOGISTIC. The logistic model is

$$\text{logit}(\pi) = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 X*Z \, .$$

The p-value of the parameter β_3 is $p = 0.0199 < 0.10$, so there is significant interaction between X and Z in the logistic model. We will need to re-fit the model according to gender.

For males, the logistic model for males is $\text{logit}(\pi) = \beta_0 + \beta_1 X$. The odds ratio is 0.483, and its maximum likelihood estimate has p-value $p = 0.0886 < 0.10$, so the association is significant. The odds of high HDL (over 47 mmHg) is 0.483 times greater for older (over 17 years) males compared to younger males. The 95% confidence interval for the odds ratios is 0.209 to 1.116. The final model is $\text{logit}(\pi) = -2.64E\text{-}8 - 0.7282X$.

For females, the logistic model for females is $\text{logit}(\pi) = \beta_0 + \beta_1 X$. The odds ratio is 1.953, and its maximum likelihood estimate has p-value $p = 0.1121 > 0.10$, so the association is not significant. The odds of high HDL (over 47 mmHg) is 1.9531 times greater for older (over 17 years) females compared to younger females. The 95% confidence interval for the odds ratios is 0.855 to 4.460.

Part 2A

We examine whether AGE_HI (Z-variable) is a confounder for the association between TG_HI (Y-variable) and BMI_HI (X-variable) using PROC FREQ. The age-adjusted odds ratio is 0.7725, which is a 53.54% decrease from the unadjusted odds ratio of 1.6628. Therefore, AGE_HI is a confounder. After adjusting for age, the odds ratio of high triglycerides for high bmi compared to low bmi is 0.7725 with a 95% confidence interval of 0.4557 to 3.4382.

We examine whether GENDER (Z-variable) is a confounder for the association between TG_HI (Y-variable) and BMI_HI (X-variable) using PROC FREQ. The gender-adjusted odds ratio is 1.6803, which is a 1.05% increase from the unadjusted odds ratio of 1.6628. Therefore, GENDER is not a confounder, so we report just the unadjusted odds ratio of 1.6628 with a 95% confidence interval of 0.9369 to 2.9511.

Part 2B

In this problem, we study the association between TG_HI (Y-variable) and BMI_HI (X-variable) using PROC LOGISTIC. Before, analyzing possible confounders, we examine the unadjusted logistic model $\text{logit}(\pi) = \beta_0 + \beta_1 X$. The unadjusted odds ratio is 1.663, and the parameter estimates are $\beta_0 = -0.3185$ and $\beta_1 = 0.5085$. The 95% confidence interval of 0.949 to 3.019.

We examine whether AGE_HI (Z-variable) is a confounder for the association between TG_HI (Y-variable) and BMI_HI (X-variable) using PROC LOGISTIC. The age-adjusted logistic model is $\text{logit}(\pi) = \beta_0 + \beta_1 X + \beta_2 Z$ and the parameter estimates are $\beta_0 = -0.6987$, $\beta_1 = -0.2607$, and $\beta_2 = 1.6243$. The age-adjusted odds ratio is 0.7720, which is a 53.7% decrease from the unadjusted odds ratio of 1.663. Therefore, AGE_HI is a confounder. After adjusting

for age, the odds ratio of high triglycerides for high bmi compared to low bmi is 0.772 with a 95% confidence interval of 0.379 to 1.568.

We examine whether GENDER (Z-variable) is a confounder for the association between TG_HI (Y-variable) and BMI_HI (X-variable) using PROC LOGISTIC. The gender-adjusted logistic model is $\text{logit}(\pi) = \beta_0 + \beta_1 X + \beta_2 Z$ and the parameter estimates are $\beta_0 = -0.4011$, $\beta_1 = 0.5263$, and $\beta_2 = 0.1474$. The gender-adjusted odds ratio is 1.693, which is a 1.8% increase from the unadjusted odds ratio of 1.663. Therefore, GENDER is not a confounder, so we report just the unadjusted odds ratio of 1.663 with a 95% confidence interval of 0.949 to 3.019.