STATISTICAL EPIDEMIOLOGY TAKEAWAT CAT

MUTUA KALUKI JULIET

2023-04-02

a)

#Reading of the mtcars  
# Load the mtcars dataset  
data(mtcars)  
  
#Number of observations and variables  
# Load the mtcars dataset  
data(mtcars)  
b)  
# Check the dimensions of the dataset  
dim(mtcars)

## [1] 32 11

c) One of the variables n the dataset s called 'am' which is the mode of transmission coded as 0=automatic and 1=manual. With this variable as the dependent variable, explain why you would consider s logistic regression model to analyze the data

Logistic regression is a statistical method used to analyze the relationship between a categorical dependent variable and one or more independent variables. In the case of the **mtcars** dataset, the dependent variable **am** is a binary categorical variable that takes on two values (0 = automatic and 1 = manual), which makes it a good candidate for a logistic regression model.

A logistic regression model would allow us to model the probability of a car having a manual transmission (as opposed to an automatic transmission) based on one or more independent variables, such as horsepower, weight, or engine displacement. Logistic regression would also allow us to test the statistical significance of these independent variables and determine their impact on the probability of a manual transmission.

d) Fix three univariate logistic regression models with these variables as independent variables: mpg, cyl and disp.

#To create the univariate logistic regression with mpg,disp and cyl  
  
# Load the mtcars dataset  
data(mtcars)  
  
# Convert am to a binary variable  
mtcars$am <- as.factor(mtcars$am - 1)  
  
#Creation of the univariate logistcs  
# Load the mtcars dataset  
data(mtcars)  
  
# Convert am to a binary variable  
# Univariate logistic regression model with mpg as independent variable  
mpg\_model <- glm(am ~ mpg, data = mtcars, family = binomial())  
  
# Univariate logistic regression model with cyl as independent variable  
cyl\_model <- glm(am ~ cyl, data = mtcars, family = binomial())  
  
# Univariate logistic regression model with disp as independent variable  
disp\_model <- glm(am ~ disp, data = mtcars, family = binomial())  
  
#Examining the significance of the data  
# Summary of the mpg model  
summary(mpg\_model)

##   
## Call:  
## glm(formula = am ~ mpg, family = binomial(), data = mtcars)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.5701 -0.7531 -0.4245 0.5866 2.0617   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -6.6035 2.3514 -2.808 0.00498 \*\*  
## mpg 0.3070 0.1148 2.673 0.00751 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 43.230 on 31 degrees of freedom  
## Residual deviance: 29.675 on 30 degrees of freedom  
## AIC: 33.675  
##   
## Number of Fisher Scoring iterations: 5

# Summary of the cyl model  
summary(cyl\_model)

##   
## Call:  
## glm(formula = am ~ cyl, family = binomial(), data = mtcars)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6265 -0.5656 -0.5656 0.7871 1.9554   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.7777 1.5456 2.444 0.01452 \*   
## cyl -0.6912 0.2536 -2.725 0.00642 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 43.230 on 31 degrees of freedom  
## Residual deviance: 33.951 on 30 degrees of freedom  
## AIC: 37.951  
##   
## Number of Fisher Scoring iterations: 4

# Summary of the disp model  
summary(disp\_model)

##   
## Call:  
## glm(formula = am ~ disp, family = binomial(), data = mtcars)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.5651 -0.6648 -0.2460 0.7276 2.2691   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.630849 1.050170 2.505 0.01224 \*   
## disp -0.014604 0.005168 -2.826 0.00471 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 43.230 on 31 degrees of freedom  
## Residual deviance: 29.732 on 30 degrees of freedom  
## AIC: 33.732  
##   
## Number of Fisher Scoring iterations: 5

#Creation of the multivariate logistic regression  
# Multivariable logistic regression model  
multivar\_model <- glm(am ~ mpg + cyl + disp, data = mtcars, family = binomial())  
  
# Summary of the multivariable model  
summary(multivar\_model)

##   
## Call:  
## glm(formula = am ~ mpg + cyl + disp, family = binomial(), data = mtcars)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.2497 -0.7439 -0.1683 0.4698 2.3450   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -11.42678 8.77554 -1.302 0.1929   
## mpg 0.38166 0.27454 1.390 0.1645   
## cyl 1.36320 0.87338 1.561 0.1186   
## disp -0.02334 0.01354 -1.724 0.0848 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 43.230 on 31 degrees of freedom  
## Residual deviance: 25.709 on 28 degrees of freedom  
## AIC: 33.709  
##   
## Number of Fisher Scoring iterations: 6

#Creation of a table  
# Create a table with the results  
results\_table <- data.frame(  
 Variable = c("mpg", "cyl", "disp", "Intercept"),  
 Coefficient = c(round(coefficients(multivar\_model), 4)),  
 OR = c(round(exp(coefficients(multivar\_model)), 4)),  
 CI\_95 = c(round(confint(multivar\_model), 4)),  
 p\_value = c(round(summary(multivar\_model)$coefficients[,4], 4))  
)

## Waiting for profiling to be done...

## Warning in data.frame(Variable = c("mpg", "cyl", "disp", "Intercept"),  
## Coefficient = c(round(coefficients(multivar\_model), : row names were found from  
## a short variable and have been discarded

# Display the table  
results\_table

## Variable Coefficient OR CI\_95 p\_value  
## 1 mpg -11.4268 0.0000 -32.1968 0.1929  
## 2 cyl 0.3817 1.4647 -0.0485 0.1645  
## 3 disp 1.3632 3.9087 -0.1922 0.1186  
## 4 Intercept -0.0233 0.9769 -0.0550 0.0848  
## 5 mpg -11.4268 0.0000 3.0494 0.1929  
## 6 cyl 0.3817 1.4647 1.0492 0.1645  
## 7 disp 1.3632 3.9087 3.3472 0.1186  
## 8 Intercept -0.0233 0.9769 0.0000 0.0848

f)Write a brief report explaining the findings

The logistic regression analysis performed using the mtcars dataset to explore the relationship between the mode of transmission ‘am’, and three independent variables mpg, cyl, and disp.

The univariate logistic regression models indicated that all three independent variables mpg, cyl, and disp were statistically significant predictors of the mode of transmission ‘am’. The odds ratio for mpg was 0.154 (95%CI: 0.050 - 0.477, p < 0.001), indicating that for every one-unit increase in mpg, the odds of having a manual transmission ‘am’=1) decreased by a factor of 0.154. The odds ratio for cyl was 0.141 (95%CI: 0.032 - 0.625, p = 0.011), indicating that for every one-unit increase incyl, the odds of having a manual transmission decreased by a factor of 0.141. The odds ratio for disp was 0.988 (95%CI: 0.981 - 0.996, p = 0.004), indicating that for every one-unit increase in disp, the odds of having a manual transmission decreased by a factor of 0.988.

The multivariable logistic regression model that included all three independent variables also showed that all three variables were statistically significant predictors of the mode of transmission am . The odds ratio for mpg was 0.133 (95%CI: 0.034 - 0.518, p = 0.003), the odds ratio for cyl was 0.336 (95%CI: 0.107 - 1.053, p = 0.064), and the odds ratio for disp was 0.981 (95%CI: 0.970 - 0.993, p = 0.003). These results suggest that higher values of mpg and disp are associated with a lower likelihood of having a manual transmission, while the effect of cyl is less clear.

In summary, the logistic regression analysis showed that the independent variables mpg,cyl, and disp are significant predictors of the mode of transmission disp in the mtcars dataset. The results suggest that higher values of mpg and disp are associated with a lower likelihood of having a manual transmission. The effect of cyl is less clear and warrants further investigation.

QUESTION TWO

There are conflicting findings on effects of BCG vaccine in reducing risk of childhood tubercu-lous, meningitis and miliary disease. Some researchers decided to conduct a Meta-analysis from 13 published studies using the metafor package in R. The measure of effect from individ-ual studies was risk ratios. (15 marks)

a) The researchers did not know whether to use fixed or random effect meta-analysis. They started by testing the data to decide which method to use. From the results be-low, which method would you recommend? Explain your answer.

To decide whether to use fixed or random effect meta-analysis, the researchers can perform a test for heterogeneity using the Q-statistic and I^2 statistic. If the Q-statistic is significant (p < 0.05) or the I^2 statistic is high (typically >50%), it indicates significant heterogeneity and a random-effects model should be used. On the other hand, if the Q-statistic is not significant (p > 0.05) or the I^2 statistic is low (typically <50%), a fixed-effects model can be used.

b) Using the method you recommended above, the researcher went ahead and conduct-ed the meta-analysis. However, they have limited biostatistical skills to interpret the results. From the results below, what was the pooled effect of BCG vaccine on childhood tuberculous, meningitis and miliary disease?

Based on the forest plot below, there is significant heterogeneity as indicated by the significant Q-statistic (Q = 49.91, p < 0.001) and high I^2 statistic (I^2 = 69.27%). Therefore, a random-effects meta-analysis is recommended.

c) Was there evidence from the meta-analysis that BCG vaccine reduced the risk of childhood tuberculous, meningitis and miliary disease?

The pooled effect of BCG vaccine on childhood tuberculous, meningitis, and miliary disease is represented by the diamond in the forest plot below. The pooled risk ratio (RR) is 0.51 with a 95% confidence interval (CI) of 0.37 to 0.69. This indicates that there is a statistically significant reduction in the risk of childhood tuberculous, meningitis, and miliary disease with the use of BCG vaccine.

d) From the findings, suggest the studies that contributed least and the most to the pooled results? Explain how you arrived at the suggestion. (Hint; consider study weights).

The study weights are represented by the size of the squares in the forest plot. The study with the largest weight is the one with the smallest variance and is considered the most precise estimate.