STATISTICAL EPIDEMIOLOGY TAKEAWAY CAT

KERAMA EZEKIEL JAMES

2023-03-29

QUESTION ONE

a)

#Reading of the mtcars  
# Load the mtcars dataset  
data(mtcars)  
  
# Print the first 6 rows of the dataset  
head(mtcars)

## mpg cyl disp hp drat wt qsec vs am gear carb  
## Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4  
## Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4  
## Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1  
## Hornet 4 Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1  
## Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2  
## Valiant 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1

b)

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

## [1] 32 11

Observations=32

Varables=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

e) Fix a multivariable logistic regression model with all the three independent variables. Prepare a table with the results. Report the adjusted Odds Ratios, their 95% CI and P-values

#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. Conversely, the study with the smallest weight is the one with the largest variance and is considered the least precise estimate. Based on the forest plot below, the study that contributed the most to the pooled result is study 4 (weight = 16.05%), while the study that contributed the least is study 8 (weight = 1.74%).

Second report writing of the dataset mtcars.

Write a report on the above dataset

**Report on the mtcars Dataset**

**Introduction:**

The mtcars dataset is a built-in dataset in R that contains information about various car models. It consists of 32 observations (car models) and 11 variables representing different attributes of the cars. In this report, we will perform a descriptive analysis of the mtcars dataset to gain insights into the data.

**Summary Statistics:**

To begin our analysis, we calculated summary statistics for each variable in the dataset. The summary statistics provide a general overview of the central tendency, spread, and range of the variables. The following table presents the summary statistics for the numeric variables in the mtcars dataset:

| **Variable** | **Min** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max** |
| --- | --- | --- | --- | --- | --- | --- |
| mpg | 10.40 | 15.43 | 19.20 | 20.09 | 22.80 | 33.90 |
| cyl | 4 | 4 | 6 | 6.19 | 8 | 8 |
| disp | 71.1 | 120.8 | 196.3 | 230.7 | 326 | 472 |
| hp | 52 | 96.5 | 123 | 146.7 | 180 | 335 |
| drat | 2.76 | 3.08 | 3.70 | 3.60 | 3.92 | 4.93 |
| wt | 1.51 | 2.58 | 3.32 | 3.22 | 3.61 | 5.42 |
| qsec | 14.5 | 16.89 | 17.71 | 17.85 | 18.90 | 22.90 |
| vs | 0 | 0 | 0 | 0.44 | 1 | 1 |
| am | 0 | 0 | 0 | 0.41 | 1 | 1 |
| gear | 3 | 3 | 4 | 3.69 | 4 | 5 |
| carb | 1 | 2 | 2 | 2.81 | 4 | 8 |

From the summary statistics, we can observe that the average miles per gallon (mpg) across all cars is approximately 20.09, with a minimum of 10.40 and a maximum of 33.90. The number of cylinders (cyl) ranges from 4 to 8, with an average of 6.19. The other variables, such as displacement (disp), horsepower (hp), weight (wt), and others, also exhibit varying ranges and means.

**Correlation Analysis:**

To understand the relationships between variables, we conducted a correlation analysis. The correlation matrix provides information about the strength and direction of the linear relationships between pairs of variables. The following correlation matrix represents the correlations between the numeric variables in the mtcars dataset:

|  | **mpg** | **cyl** | **disp** | **hp** | **drat** | **wt** | **qsec** | **vs** | **am** | **gear** | **carb** | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| mpg | 1.000 | -0.852 | -0.848 | -0.776 | 0.681 | -0.868 | 0.419 | 0.664 | 0.600 | 0.480 | -0.550 | |
| cyl | -0.852 | 1.000 | 0.902 | 0.832 | -0.700 | 0.782 | -0.591 | -0.813 | -0.522 | -0.492 | 0.526 | |
| disp | -0.848 | 0.902 | 1.000 | 0.791 | -0.710 | 0.888 | -0.433 | -0.710 | -0.591 | -0.555 | | 0.394 | |
| hp | -0.776 | 0.832 | 0.791 | 1.000 | -0.448 | 0.658 | -0.708 | -0.723 | -0.243 | -0.126 | | 0.749 | |
| drat | 0.681 | -0.700 | -0.710 | -0.448 | 1.000 | -0.712 | 0.091 | 0.440 | 0.713 | 0.699 | -0.091 | |
| wt | -0.868 | 0.782 | 0.888 | 0.658 | -0.712 | 1.000 | -0.175 | -0.554 | -0.692 | -0.583 | 0.427 | |
| qsec | 0.419 | -0.591 | -0.433 | -0.708 | 0.091 | -0.175 | 1.000 | -0.229 | -0.297 | -0.212 | -0.656 | |
| vs | 0.664 | -0.813 | -0.710 | -0.723 | 0.440 | -0.554 | -0.229 | 1.000 | 0.168 | 0.206 | -0.570 | |
| am | 0.600 | -0.522 | -0.591 | -0.243 | 0.713 | -0.692 | -0.297 | 0.168 | 1.000 | 0.794 | 0.057 | |
| gear | 0.480 | -0.492 | -0.555 | -0.126 | 0.699 | -0.583 | -0.212 | 0.206 | 0.794 | 1.000 | 0.274 | |
| carb | -0.550 | 0.526 | 0.394 | 0.749 | -0.091 | 0.427 | -0.656 | -0.570 | 0.057 | 0.274 | 1.000 | |

From the correlation matrix, we observe several interesting relationships. For instance, there is a strong negative correlation between "mpg" and "cyl" (-0.852), indicating that as the number of cylinders increases, the miles per gallon decrease. Similarly, variables such as "disp", "hp", and "wt" also exhibit negative correlations with "mpg", suggesting that larger values in these variables tend to be associated with lower fuel efficiency.

**Conclusion:**

In conclusion, the descriptive analysis of the mtcars dataset provided insights into the various car attributes. The summary statistics gave an overview of the range and central tendency of the variables, while the correlation analysis revealed relationships between different attributes. We observed that variables such as "cyl", "disp", "hp", and "wt" were negatively correlated with "mpg", indicating that these factors can have an impact on fuel efficiency.

This analysis serves as a starting point for further exploration and modeling. The mtcars dataset can be utilized for various data-driven tasks such as predictive modeling, classification, and clustering.

**Recommendation**

Based on the analysis of the mtcars dataset, we have identified several insights that can guide recommendations related to car design and purchasing decisions. Here are our recommendations:

1. **Consider Fuel Efficiency:**
2. The analysis revealed a negative correlation between variables such as the number of cylinders ("cyl"), engine displacement ("disp"), horsepower ("hp"), and weight ("wt") with miles per gallon ("mpg"). Therefore, for individuals and car manufacturers aiming for better fuel efficiency, it is recommended to opt for cars with fewer cylinders, lower engine displacement, reduced horsepower, and lighter weight.
3. **Evaluate Transmission Type:**
4. The analysis showed that the type of transmission ("am") is correlated with variables such as horsepower ("hp") and number of gears ("gear"). It is recommended for individuals to consider cars with automatic transmission ("am=0") if they prefer higher horsepower and more gears, while manual transmission ("am=1") can be preferred for those seeking a different driving experience.
5. **Compare Car Models:**
6. By analyzing the dataset, individuals looking to purchase a car can compare different models based on their attributes. Factors such as fuel efficiency ("mpg"), number of cylinders ("cyl"), and weight ("wt") should be considered along with personal preferences and requirements. This can help individuals make informed decisions about which car models align best with their needs.
7. **Explore Additional Factors:**
8. While the analysis focused on specific attributes, it is important to note that other factors such as safety features, interior space, and brand reputation should also be taken into account when making car-related decisions. Combining the insights from the dataset analysis with these additional factors will lead to more comprehensive decision-making.
9. **Continued Data Analysis:**
10. The mtcars dataset offers further opportunities for analysis and modeling. Additional techniques such as predictive modeling, clustering, and feature engineering can be applied to gain deeper insights and build more advanced models. This can assist car manufacturers in understanding customer preferences and help individuals in making more accurate predictions and decisions.

By following these recommendations, individuals can make informed choices when purchasing cars, while car manufacturers can use these insights to design and develop vehicles that meet customer expectations and market demands.

**Note:**

It is important to consider that the mtcars dataset is a small and simplified dataset, and the recommendations provided should be further validated with more comprehensive and up-to-date data in real-world scenarios.