93

STA 207 HW-2

Due Date: Sept. 10, 2024 by 10:20am in Moodle

Derin Gezgin | Camel ID: 00468038

Problem-1 [40 points]

Identify the type of regression model (2 points) that would work best for the given studies. List the response and predictor variables for each (1 point for each variable)

a.) In a study to predict the price of a used car, the variables recorded are car's price, mileage, age, and manufacturer.

Response Variable (Y): Price of the car
Predictor Variables (X): Mileage / Age / Manufacturer of the car
Type of regression model: ***Multiple Linear Regression***
(Y → Numerical Cont. | X → Numerical & Categorical)
As there are multiple predictor variables, it's the multiple linear regression.

b.) Whether or not an applicant is accepted for medical school using their grade point average, school, and gender.

Response Variable (Y): Applicant is accepted for medical school (Yes / No)
Predictor Variables (X): GPA / School / Gender of the applicant
Type of regression model: ***Logistic Regression***
(Y → Categorical | X → Numerical & Categorical)
As the Y variable is categorical, we can directly say logistic regression.

c.) Use credit score and bank balance to predict whether or not a given customer will default on a loan.

Response Variable (Y): The customer will default on a loan (Yes / No)
Predictor Variables (X): Credit score / Bank balance
Type of regression model: ***Logistic Regression***
(Y → Categorical | X → Numerical)
As the Y variable is categorical, we can directly say logistic regression.

d.) To predict household income using total years of schooling, number of adults in the household, hours worked, and cost of living.

Response Variable (Y): Household income
Predictor Variables (X): Total years of schooling / Number of adults in the household / Hours worked / Cost of Living
Type of regression model: ***Multiple Linear Regression***
(Y → Numerical | X → Numerical)
This is the one I'm the most confused about, Y and all X variables are numerical and it's Multiple Linear Regression. But it can also be Polynomial as if there're too many adults in the household, the total income can decrease. As it's a reach, my final answer is multiple linear regression.

Yes, it is MLR

e.) To examine the number of traffic accidents at a particular intersection based on weather conditions ("sunny", "cloudy", "rainy") and whether or not a special event is taking place in the city ("yes" or "no").

Response Variable (Y): Number of traffic accidents
Predictor Variables (X): Weather conditions / A special event is taking place
Type of regression model: ***~~Multiple Linear Regression~~*** -2, Poisson
(Y → Numerical | X → Categorical)
Even though we have a "number of …" situation in the response variable, all our predictor variables are categorical so the only option would be multiple linear regression.

f.) To predict the number of people ahead of you in line at a store based on time of day, day of the week, and whether or not there is a sale taking place ("yes" or "no").

Response Variable (Y): # of people ahead of me in the line
Predictor Variables (X): Time of the day / Day of the week / If there's a sale
Type of regression model: ***Poisson Regression***
(Y → Numerical | X → Numerical + Categorical)
As we're using number of people as a response variable, it's the ***poisson*** *regression.*

g.) We want to use square footage, school ratings, and number of bathrooms to predict whether or not a house in a certain city will be listed at a selling price of $200k or more. (Response variable = "Yes" or "No")

Response Variable (Y): The house will be listed more or less than $200,000
Predictor Variables (X): Square footage / School ratings / Number of bathrooms
Type of regression model: ***Logistic Regression***
(Y → Categorical | X → Numerical)
Considering the response variable is categorical, we'd use logistic regression.

Problem-2 [25 points]

The data below were gathered on a random sample of 7 male black-footed albatrosses of known age. In an effort to monitor diseases of these animals, biologists would like to be able to estimate the age of animals that have died by flattening their gonads and measuring the resulting area.

Gonad size vs. Age in Black-footed albatrosses

Gonad Size (sq mm)	Age (Years)
42	1.42
60	4.75
20	0.67
96	23.64
24	0.52
27	2.35
27	1.4

Answer the following:

1) Identify response and predictor variables (5 points).

   Response Variable (Y): Age (Years)

   Predictor Variables (X): Gonad Size (sq mm)

2) Enter this data in R to do the following:

   a) Compute mean and standard deviation for the variables and report the values with correct units (5 points).

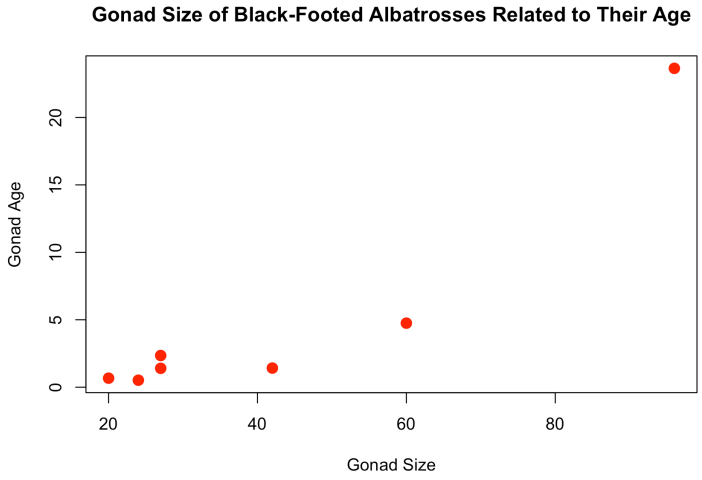
   Gonad Size
   Standard Deviation: 27.378 sq mm
   Mean: 42.286 sq mm

   Age
   Standard Deviation: 8.358 years
   Mean: 4.964 years

   b) Compute the correlation coefficient (using function cor) between Gonad Size and Age. Comment on the strength and direction of the relationship using correlation. (7.5 points)

   Correlation Coefficient between Gonad size and Age → 0.926
   The relationship between the variables is positive and very strong.

   c) Make a scatter plot of X versus Y and comment on the relation. (7.5 points)



-5, need to comment on the relation and mention outlier/nonlinear

My code for this section:

```r
# Adding the Data
gonad_size = c(42, 60, 20, 96, 24, 27, 27)
gonad_age = c(1.42, 4.75, 0.67, 23.64, 0.52, 2.35, 1.4)

# Calculating the standard deviation
round(sd(gonad_size), 3)
round(sd(gonad_age), 3)

# Calculating the mean
round(mean(gonad_size), 3)
round(mean(gonad_age), 3)

# Calculating the correlation coefficient
round(cor(gonad_size, gonad_age), 3)

# Making the scatter plot
plot(gonad_size,
     gonad_age,
     main = "Gonad Size of Black-Footed Albatrosses Related to Their Age",
     xlab = "Gonad Size",
     ylab = "Gonad Age",
     pch = 16,
     col = 'red',
     cex = 1.5,)

```

Problem-3 [25 points]

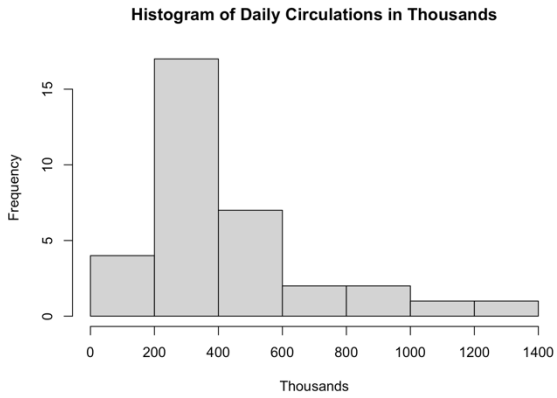
In order to investigate the feasibility of starting a Sunday edition for a large metropolitan newspaper, information was obtained from a sample of 34 newspapers concerning their daily and Sunday circulations (in thousands). The newspaper data is shared in Moodle as Newspaper.csv. Assigning daily circulations as predictor and the Sunday circulations as response variable, do the following using R:

a) Using summary function in R, prepare numerical summary of variables (5 points).

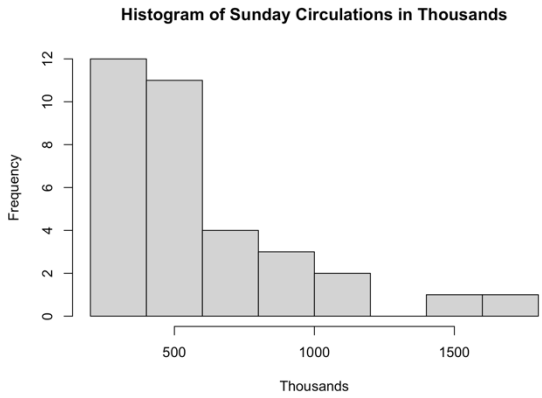
Daily		Sunday	
Min.	: 133.2	Min.	: 202.6
1st Qu.	: 233.0	1st Qu.	: 327.8
Median	: 355.2	Median	: 436.7
Mean	: 431.0	Mean	: 591.2
3rd Qu.	: 516.6	3rd Qu.	: 699.7
Max.	:1209.2	Max.	:1762.0

b) Using hist function in R, make histograms for X and Y. Comment on what do you observe (10 points).

X → Daily Circulations



Y → Sunday Circulations

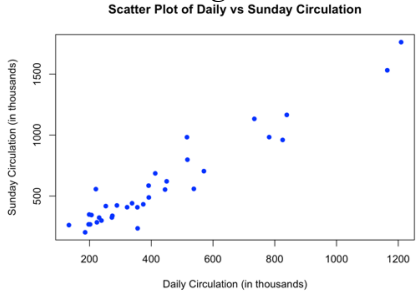


In both of the histograms, I can see that the data is right-skewed. From the data summary and the histograms, we can see that the maximum value of the Sunday circulations is much higher than the daily circulations. 200-400 thousand range in the daily circulations and 0-500 range in the Sunday circulations have majority of the newspapers.

c) Using cor function in R, find correlation between the two circulation variables and comment on the direction & strength of the relation (5 points).
Correlation Coefficient between the circulation variables → *0.958*
The strength of the correlation is significantly strong in the positive direction considering it's really close to 1.

d) Using plot function in R, show a scatter plot of X versus Y. Explain the strength and direction of relationship (5 points).

In this scatter plot, we can definitely see the strong relationship between the response and the predictor variables. In the 200-600 range -considering we have a lot of data- this strong relationship is very visible, on the other hand, in the 600-1200 range this relationship is less significant but still visible and strong.

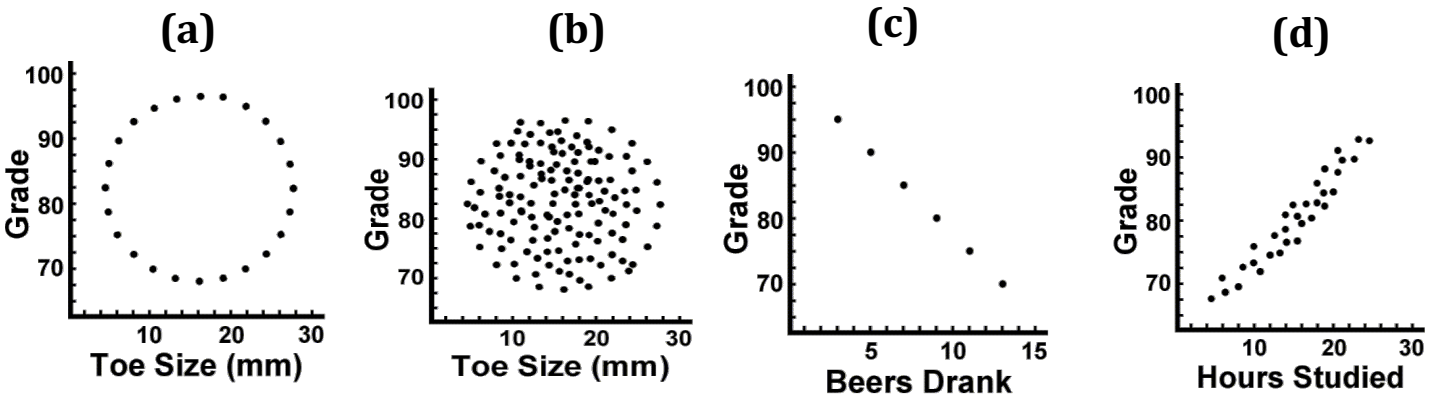


My code for this section:

```r
# Reading the data
newspaper_data = read.csv("../Newspaper.csv")

# Getting the summary
summary(newspaper_data)

# Creating an histogram of the Daily circulations
hist(newspaper_data$Daily,
     main="Histogram of Daily Circulations in Thousands",
     xlab="Thousands")

# Creating an histogram of Sunday circulations
hist(newspaper_data$Sunday,
     main="Histogram of Sunday Circulations in Thousands",
     xlab="Thousands")

# Calculating the correlation coefficient
round(cor(newspaper_data$Sunday, newspaper_data$Daily),3)

#Plotting the daily circulation (X) against Sunday circulation (Y) to see the relationship between them.
plot(newspaper_data$Daily,
     newspaper_data$Sunday,
     main = "Scatter Plot of Daily vs Sunday Circulation",
     xlab = "Daily Circulation (in thousands)",
     ylab = "Sunday Circulation (in thousands)",
     pch = 16,
     col = "blue")

```

Problem-4 [10 points]

For the following four scatterplots **match the given four values of correlation coefficients. Explain the reason for your choice** in one or two sentences.



1. r = 0.06 → B
   Reason: *It's not an exact circle equation. It's a random blub. There's no direction.*
2. r = 0.89 → D
   Reason: *While the relationship is strong in the positive direction it's still not perfect.*
3. r = 0 → A
   Reason: *The relationship is circular and it's not a linear relationship.*
4. r = -1 → C
   Reason: *We have a perfect relationship on the negative side.*