STA 207 HW-2 Solution

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.

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

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

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

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").

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").

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")

Solution:

Part	Model	Response Variable	Predictor Variables
a	Multiple Linear Regression (MLR)	Price of a used car	Car price, Mileage, Age, Manufacturer
b	Logistic	Applicant gets accepted for medical school or not	GPA, School, Gender
c	Logistic	Customer will default on a loan or not	Credit score, Bank balance
d	MLR	Household income	Total years of schooling, Number of adults, Hours worked, Cost of living
e	Poisson	Number of traffic accidents	Weather conditions ("sunny", "cloudy", "rainy"), Whether or not a special event is taking place in the city ("yes" or "no").
f	Poisson	Number of people ahead of you in line at a store	Time of day, Day of the week, Whether or not there is a sale taking place ("yes" or "no").
g	Logistic	whether or not a house in a certain city will be listed at a selling price of $200k or more.	Square footage, School ratings, Number of bathrooms

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.

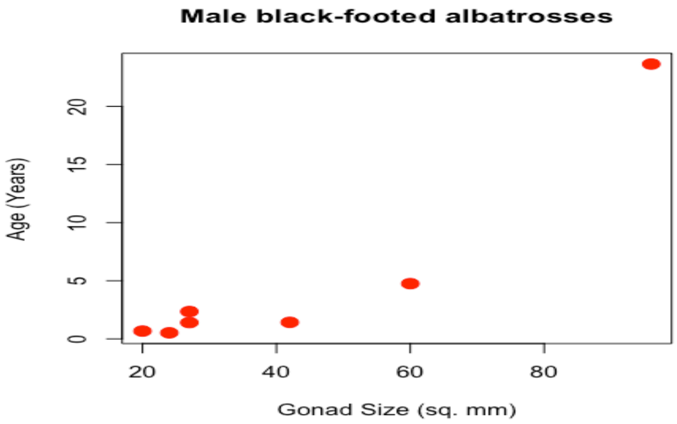
Gonad size vs. Age in Black-footed albatrosses

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).
   Y: age of the black-footed albatrosses (years)
   X: size of the gonad (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).
      $\bar{x} = 42.286\, sq\ mm,\quad \bar{y} = 4.964\, years,\quad s_X = 27.378\, sq.mm,\quad s_Y = 8.358\, years$
   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: $Cor(Y, X) = 0.926$
      The correlation between gonad size and age of black-footed albatrosses is 0.926 which shows a strong, positive linear relation between these study variables; however the sample size is only 7 .
   c) Make a scatter plot of X versus Y and comment on the relation. (7.5 points)

There are only 7 data points to it is hard to see an obvious pattern, we notice that large gonad size values indicated an older black-footed albatross indicating a positive relationship. There is an outlier (age close to 100) which makes the overall pattern look exponential.



Problem-3 [25 points]

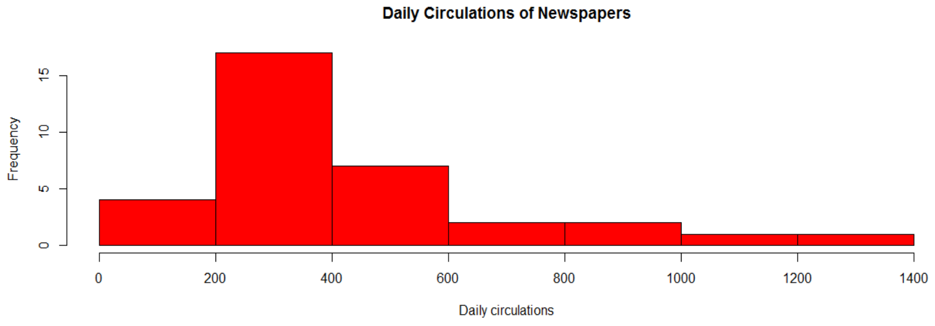
In order to investigate the feasibility of starting a Sunday edition for a large metropolitan newspaper, information was obtained from a sample of 34 newspapers concerning their daily and Sunday circulations (in thousands). The newspaper data is shared in Moodle as Newspaper.csv. Assigning daily circulations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

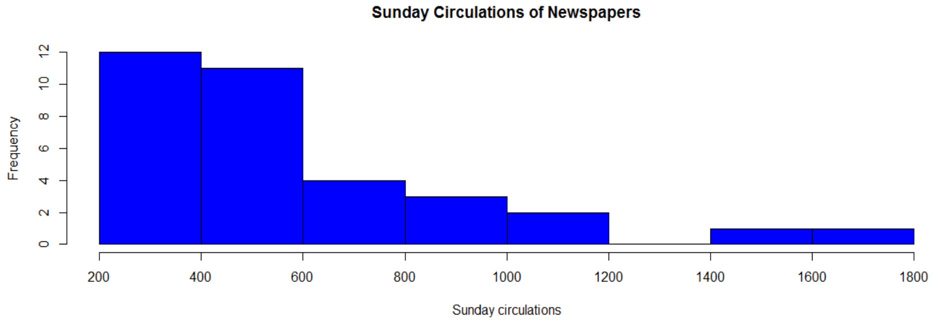
a) Using summary function in R, prepare numerical summary of variables (5 points).

Variable	**Minimum**	**Q1**	**Q2 (median)**	**Q3**	**Mean**	**Max**
Daily circulations (X)	133.2	233	355.2	516.6	431	1209
Sunday circulations (Y)	202.6	327.8	436.7	699.7	591.2	1762

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

From the histogram of daily and Sunday circulations (in 1000s), we notice that the circulations are right skewed, but the Sunday circulations are slightly righter skewed. Daily circulations peak between 200-400 (1000s) and Sunday circulations peak at 200-400 (1000s) and 400-600 (1000s). Sunday circulations are up to 1800 (1000s) (very few times, low frequency) whereas the maximum for daily circulations, with very low frequency, is 1400 (1000s).



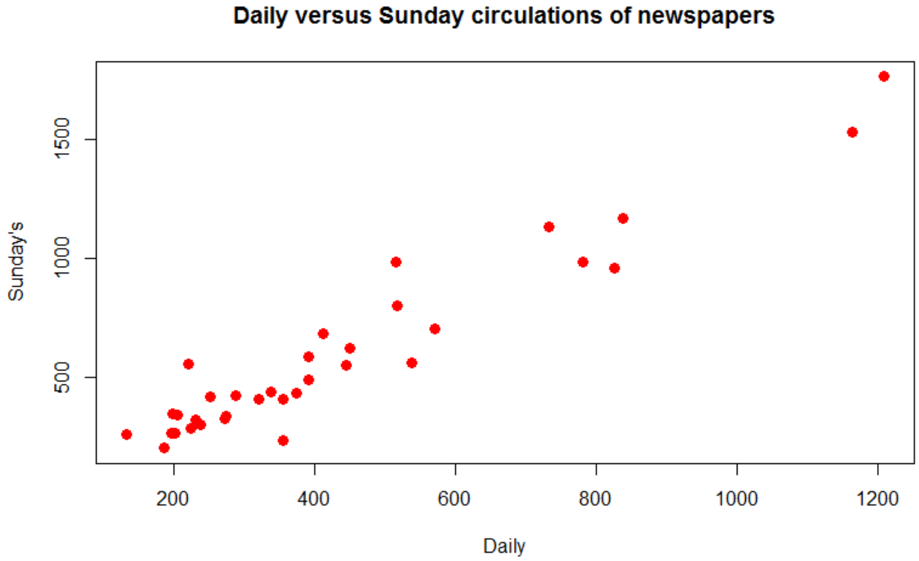


c) Using cor function in R, find correlation between the two circulation variables and comment on the direction & strength of the relation (5 points).

Correlation between the daily (X) and Sunday (Y) circulations is 0.958, which is strong and positive.

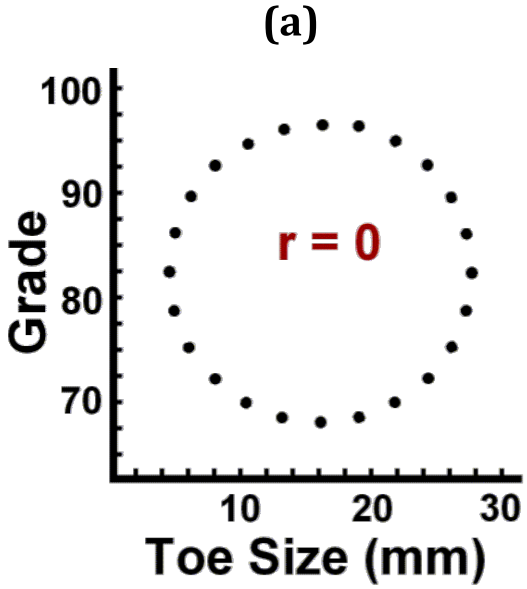
d) Using plot function in R, show a scatter plot of X versus Y. Explain the strength and direction of relationship (5 points).

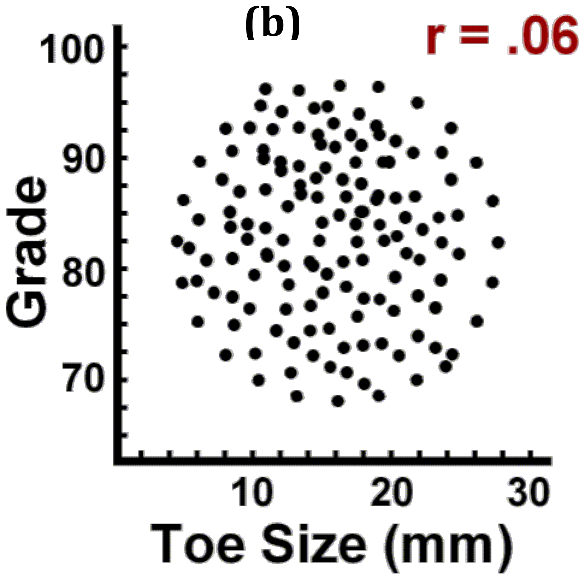
The scatterplot between the daily (X) and Sunday (Y) circulations show a strong, linear, and positive relation between the two quantities.



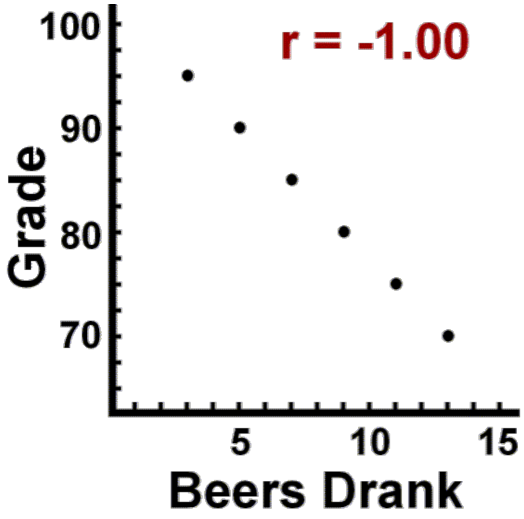
Problem-4 [10 points]

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

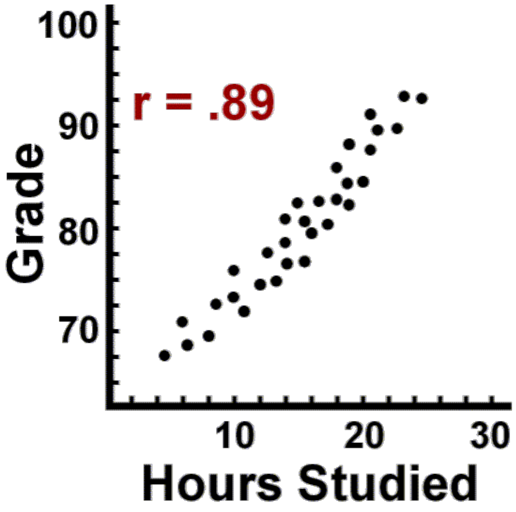




(c)



(d)



1. r = 0.06 (b) Random scatter of points in the scatterplot, so small spurious relation between toe size and grade
2. r = 0.89 (d) positive, strong linear relation between hours studied and grade
3. r = 0 (a) Non-linear relation between toe size and grade
4. r = -1 (c) clear negative and very strong linear relationship b/w beer drank and grade