Question 1

A) The data has been plotted and fitted line is plotted in the .R file
B) The P value of B1 = 0.0056 <0.05. Since the p-value is less than the significance level, the null hypothesis can be rejected
C) the p value of B1 = 0.28 > 0.05. Since the p-value is more than the significance level, the null hypothesis cannot be rejected
D) Limitations of classical Hypothesis Testing
    - Proof for a specific hypothesis should be setup before the study begins (hypothesis should be created before conducting the experiments
    - Lack of data sampling or missing data might alter the testing results
    - There might be type 1 error where the null hypothesis is actually true but is rejected
    - Dicing the dataset might tweak some variables which might alter the results of hypothesis testing. This is know as p-hacking
    - Excessive data manipulating might reduce the p-value<0.05 by random chance making a hypothesis seem valid but it might be a fluke

E) Extent of P hacking in Data Science

- P hacking occurs when researchers try out several statistical analysis and then selectively report those that produce significant results
- The common practices that lead to p hacking are :
    1. Generally analysis are performed after the experiments are conducted. However, some times midway analysis are done to check the significance of the results and to decide whether to proceed further with the experiments. This might cause P-hacking
    2. Selective reporting of variables post analysis
    3. Deciding to drop outliers post analysis
    4. Stopping data exploration if the analysis yields significant p-value
- The false positive results from p hacking hinders the scientific progress
- There are two approaches to quantify the p-hacking: text mining and meta analysis.
- The p hacking is assessed by using text mining to search for p values in in all open access papers. The p-values are extracted from the result and abstract section. Then the p-curve is constructed using these p-values. These can be used to quantify the p-hacking in the research.
- The meta analysis approach quantifies and effect or relationship by averaging affect sizes from from multiple studies after weighing each one by its reliability and plotting the respective p-curve.
- Both these studies proves the that p hacking is wide spread in science. The text mining approach is based on very large dataset with p values from different disciplines and questions while the meta analysis consists of p values from specific hypothesis. However, both these approaches yields similar results

Question 2:

A) The potential problems if we use multiple linear regression are:
    - Since the outcome variable is discrete, we cannot use normal errors because a normal distribution is defined for continuous variables
    - The error of the outcome variable is not normally distributed as the output variable is bounded by 0 and 1

B) In binomial logistic regression model, the binomial distribution is used as outcome distribution since the random variable is the sum of the repeated binary outcomes
The logit function is used as the linking function for the binomial distribution which is derived from the logistic function on the linear predicators

C) The value of coefficient of x1 = -.072. This suggests that there is an inverse relationship between p and marital_id. From the data analysis it can be concluded that the unmarried person intends to vote for the democratic candidate whereas married person intends to vote for the republican candidate.

Question 3

A) Code written in .R file

B) Based on the given input the bag A might have 11 white balls and 9 black balls or 13 white and 7 white balls