Hypothesis testing and linear regression

Question 1

Suppose we have two samples from two population with sample size n and m, one with mean μ_1 , variance as σ_1^2 , the other one with mean μ_2 , variance σ_2^2 . When detecting the difference in sample mean: $\delta = \mu_1 - \mu_2$, we want power is at least 0.8.

If we know δ = 1, σ_1^2 = σ_2^2 = 1, n = m,

- Can you calculate minimal n?
- How n change along with δ ?
- How n change along with σ_1^2 ?

Answer

Question 2

A new casino game involves rolling 3 dice. The winnings are directly proportional to the total number of sixes rolled. Suppose a gambler plays the game 101 times, with the following observed counts:

Number of Sixes	Number of Rolls
0	48
1	35
2	15
3	3

Test if this is fair dice. What test to use? Calculate stats and p value

Answer:

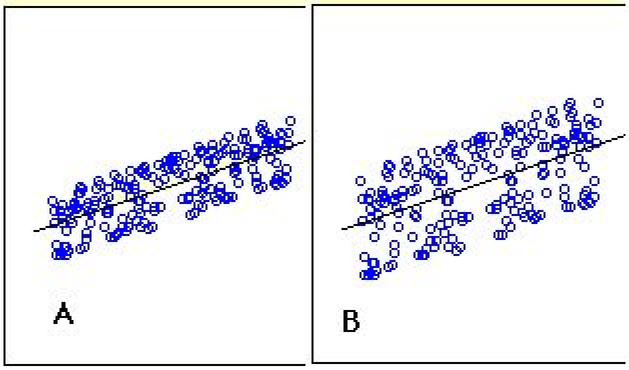
Question 3

Below graphs show two fitted regression lines (A & B) on randomly generated data. Now, I want to find the sum of residuals in both cases A and B.

Note:

Scale is same in both graphs for both axis.

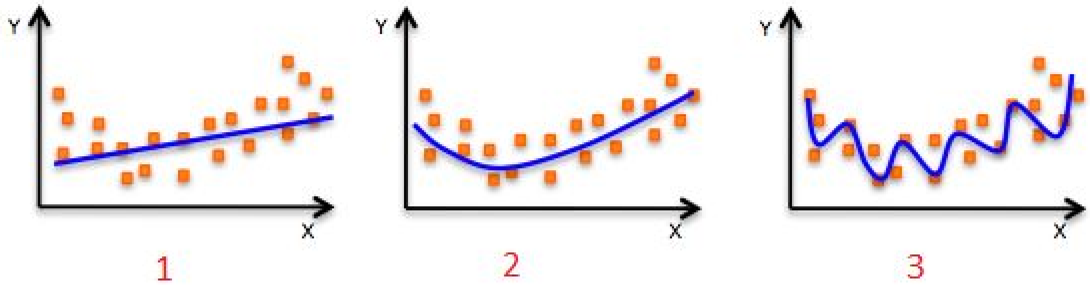
X axis is independent variable and Y-axis is dependent variable.



Which of the following statement is true about sum of residuals of A and B?
A) A has higher than B
B) A has lower than B
C) Both have same
D) None of these

Question 4

The following visualization shows the fit of three different models (in blue line) on same training data. What can you conclude from these visualizations?



1. The training error in first model is higher when compared to second and third model.
2. The best model for this regression problem is the last (third) model, because it has minimum training error.
3. The second model is more robust than first and third because it will perform better on unseen data.
4. The third model is overfitting data as compared to first and second model.
5. All models will perform same because we have not seen the test data.

A. 1 and 3
B. 2
C. 1, 3 and 4
D. Only 5

Question 5

Using MLE (maximum likelihood estimation) to achieve coefficient estimator for multiple linear regression (i.e., more than one feature in model)

Question 6

Think about if/how you would process old features and what new features to be generated. Build the best linear regression model to explain interest rate.