Interview Questions and Solutions

Mackenzie Qu

September 13, 2024

1 Regression Analysis

1. Suppose you wanted to model Y using X, and you decided to use the linear regression:

$$Y = X\beta + \varepsilon.$$

What assumptions are being made? How would you find β? What tests can be done on β afterwards?

Solution:
The general assumptions for linear regression are:

- Weak exogeneity: The variables are treated as fixed values with no measurement errors. Meaning that the error term is uncorrelated with the independent variables.
- Linearity: The relationship between the independent and the dependent variables is linear, shown by the mean of the dependent variable is a linear combination of the parameters and dependent variables. However, the assumptions of linearity is only restricting the parameters. Transformation(like polynomial regression) techniques can be used to fit datasets.
- No perfect multicollinearity(for multiple regression): The independent variables are not correlate to each other. The common problems include overfitting.
- Constant Variance(homoscedasticity): The variance of the errors does not depend on the value of the predictor variables.
- Independence of errors: The errors of the dependent variables are uncorrelated of each other.
- Error term distribution: The error terms are generally assumed to be Normally distributed $N(\mu, \sigma^2)$, but other distributions can be used.

To find β, we can use sum of squared method or likelihood method

2. What are the significance tests used for the parameters estimated in a logistic regression? How are they different from those used for a linear regression?
3. What are the assumptions required for a linear regression? What is multicollinearity, and what are its implications? How would you measure goodness of fit?
4. What assumptions are needed for a linear regression? Are they the same for a logistic regression? How would you test the significance of the parameters for a linear regression? Would you use the same test in the case of a logistic regression?

2 Probability and Statistics

1. 500 ants are randomly put on a 1 metre string. Each ant randomly moves toward on end of the string with equal probability to the left or the right at constant speed of 1 m/min until it falls off one end of the string. Also assume that the size of the ant is infinitely small. When two ants collide head-on, they both immediately change directions and keep on moving at 1 m/min. What is the expected time for all ants to fall off the string?

Solution: Consider the case where there is only 1 ant on the string. The ant will have a position of d_l, denoting the distance to the left and d_r denoting the distance to the right, such that $d_l + d_r = 1$. Let t_1 be the time for the ant to fall off the rope. To find the expected time that the aunt will take to fall off, we are looking for $E(t_1)$, where

$$E(t_1) = E(\frac{\text{distance the ant will travel to fall off}}{\text{the velocity the ant is travelling at}}) = \frac{E(\text{distance the ant will travel to fall off})}{1}$$

as the velocity of ant is constant.

$$\begin{aligned} E(\text{distance the ant will travel to fall off}) =& P(\text{travelling to the left direction}) \times \text{distance to the left}+ \\ & P(\text{travelling to the right direction}) \times \text{distance to the right} \\ =& \frac{1}{2} \times d_l + \frac{1}{2} \times d_r \\ =& \frac{1}{2} \times (d_l + d_r) = \frac{1}{2} \end{aligned}$$

Now consider the case where there are two ants. When two ants collide into each other, the momentum of the two ants carry over - we can imagine it as the two ants have swapped positions and continued going in the same direction.
In this case, we are looking for the expected value of the maximum amount of time each ant spends on the rope. ie. $E(max\{t_1, t_2\})$, which is equivalent to finding the maximum distance an ant is from the edge of the rope.
Since each ant is randomly dropped onto the rope, the distance D_i of each ant toward the direction it is moving is independent and identically distributed on $Uniform(0, 1)$. Let $Y = max\{D_1, D_2\}$, the cumulative distribution function of Y follows:

$$\begin{aligned} F_Y(y) &= P(Y < y) \\ &= P(max\{D_1, D_2\} < y) \\ &= P(D_1 < y)P(D_2 < y) \ldots \text{because of independence} \\ &= F_{D_1}(y) * F_{D_2}(y) \\ &= y^2 \end{aligned}$$

and the probability density function follows:

$$f_Y(y) = \frac{d}{dy} F_Y(y) = 2y$$

$$E(Y) = \int_0^1 y * 2y \; dy = \int_0^1 2y^2 \; dy = \frac{2}{3}$$

To find the answer for 500 ants, let $Y = max\{D_1, ..., D_{500}\}$. It follows:

$$F_Y(y) = y^{500},$$
$$f_y(y) = 500 * y^{499},$$
$$E(Y) = \frac{500}{501}$$

2. You have 100 noodles in your soup bowl. Being blindfolded, you are told to take two ends of some noodles (each end of any noodle has the same probability of being chosen) in your bowl and connect them. You continue until there are no free ends. The number of loops formed by the noodles this way is stochastic. Calculate the expected number of circles.

Solution: In this question, we'll explore the concept of conditional expectation and induction. Consider the event of picking up a random end of the noodle, denote the probability of choosing another end from the same noodle $P(S)$, such that $P(S) + P(\overline{S}) = 1$.
To start, we consider the case when there is only 1 noodle in the bowl, then

$$E(n = 1) = P(S) * 1 + P(\overline{S}) * 0 = 1$$

as there is only one noodle and 2 ends in the bowl, and the expected value of making a loop is 1.

When we add another noodle to it, we can denote the expected value of $n = 2$ as:

$$E(n = 2) = E(n = 2|P(S)) * P(S) + E(n = 2|P(\overline{S})) * P(\overline{S}) \tag{1}$$

Now we have to consider the cases that of $E(n = 2|P(S))$ and $E(n = 2|P(\overline{S}))$, which denotes the expected value of loops when there are two noodles in the bowl given that the first noodle formed a loop with its end, and when the first noodle did not form a loop with its end, respectively.

1. Consider the first case - when there are two noodles in a bowl, and the first randomly selected end of the noodle formed a loop with the next randomly selected end. Then,
$E(n = 2|P(S)) = E(n = 1) + 1 = 2$
2. In the second case, if the two ends did not end up forming a loop, then the two noodles are connected, and we are left with 1 long noodle. Then $E(n = 2|P(\overline{S})) = E(n = 1) = 1$

Combine the findings with equation(1), we have

$$E(n = 2) = 2 * \frac{1}{3} + 1 * \frac{2}{3} = \frac{4}{3}$$

3. There are two dice. The game is as following: Two dice will be thrown at once. Adding the number together you will get a number ranging from 2 to 12. The probability of getting 12 is 40%, and 6% for getting 2,3,4,5,...11 Person A and B will need to guess what the number would be. The one who has a closer guess will win. A will guess first and then B guess, after they've guessed, two dice will be thrown at once, what number should A guess and what number should be B guess?