Data Mining - EX6

Deadline: Friday, Azar 9, 1403 - November 29, 2024

Dataset link

Question 1:

Suppose a researcher wants to investigate whether the distribution of preferences for different cereal flavors among customers differs significantly from an expected distribution. Four flavors are considered: "Chocolate," "Fruity," "Honey," and "Cinnamon." The number of customers who chose each flavor is 50, 30, 60, and 40, respectively. The expected distribution for these flavors is 0.25, 0.25, 0.25, and 0.25.

Using the Chi-Square goodness-of-fit test in R, answer the following questions:

1. Define the hypotheses for the Chi-Square test.
2. Perform the Chi-Square test and interpret the results. Is there a significant difference between the observed and expected distributions?
3. If the significance level is $\alpha = 0.05$, does your conclusion change?

Question 2:

A cereal company wants to determine if there is a significant difference in the mean ratings of their cereals based on two different packaging methods. They have a sample of 20 ratings for packaging type 1 and a sample of 20 ratings for packaging type 2. The ratings are as follows:

- Type 1: 80, 85, 82, 90, 78, 88, 83, 86, 79, 81, 84, 87, 85, 89, 82, 86, 84, 88, 90, 85
- Type 2: 75, 80, 78, 76, 82, 81, 77, 79, 74, 80, 78, 82, 76, 80, 78, 81, 79, 83, 75, 78

Using the T-test in R, answer the following questions:

1. Define the hypotheses for the T-test for comparing the means of two populations.
2. Perform the T-test and interpret the results. Is there a significant difference between the mean ratings of the two packaging methods?

3. Assume a significance level of $\alpha = 0.05$ and explain how your results are interpreted at this level of significance.

Question 3:

In this scenario, your goal is to predict the rating of cereals based on fiber content. Use simple regression and answer the following questions.

1. Find the estimated regression equation and explain the values of the slope and intercept.

2. Calculate the estimated rating for a cereal with 3 grams of fiber.

3. Construct a 95% confidence interval for the true mean rating of all cereals with a fiber content of 3 grams.

4. Explain the meaning of the 3 in the regression equation.

Question 4:

In this scenario, your goal is to assess the accuracy of the regression model that uses fiber to predict rating.

1. Estimate the prediction error of the model and evaluate model accuracy using an appropriate statistic.

2. Create a 95% prediction interval for a cereal with a fiber content of 3 grams.

3. Plot the rating against fiber content and describe the shape and relationship.

Question 5:

In this scenario, your goal is to use multiple regression with fiber and sugars to predict the rating.

1. Estimate the multiple regression equation and explain the value of the fiber coefficient.

2. Compare the R^2 values of the multiple regression model and the simple regression model and explain why they might differ.

3. Compare the s values of the multiple and simple regression models and explain which value is preferable and why.