Computational Statistics and Data Analysis (MVComp2)

Exercise 11

Lecturer Tristan Bereau

Semester Wi23/24
Due Jan. 25, 2024, 23:59

1 Defective parts in a shipment (5 points)

A shipment of parts is received, out of which five are tested for defects. The number of defects, X, follows a binomial distribution, $X \sim \text{Binomial}(n = 5, p = \theta)$. The history of past shipments indicates that θ follows a prior distribution, $\text{Beta}(1, 9)$. The test reveals $X = 0$. We wish to establish whether there is significant evidence that the proportion of defective parts in the whole shipment exceeds 10%.

(a) Derive an expression for the posterior probability distribution, $p(\theta|X)$.

(b) Compare the posterior probabilities of the two models:

$$M_1 : \theta \leq 0.1$$
$$M_2 : \theta > 0.1$$

Feel free to evaluate your calculations using statistical libraries. From your results, can you conclude whether the proportion of defective parts in the whole shipment likely exceeds 10%?

2 BIC for customer data (5 points)

Download the following dataset about the number of customers entering a store given the hour of the day: customers.csv. The are two features:

Variable	Description
`hour`	hour of the day
`customers`	number of customers in the store

We will consider three models for the number of customers as a function of the number of hours:

1. constant model (i.e., intercept, $\beta_0^{(0)}$)
2. linear model (i.e., intercept $\beta_0^{(1)}$ and slope $\beta_1^{(1)}$)
3. quadratic model (i.e., intercept $\beta_0^{(2)}$, slope $\beta_1^{(2)}$, and quadratic term $\beta_2^{(2)}$)

We want to determine which one of the three regression models performs best.

(a) Via a routine such as `numpy.polyfit`, fit the three models. Report the coefficients and plot the fits against the data.

(b) Under the assumption that the model errors follow a normal distribution, $\mathcal{N}(0, \sigma^2)$, derive an expression for the likelihood term of the BIC, using a maximum-likelihood estimate for the parameter, $\hat{\sigma}^2$. Use it to construct a simple expression for the BIC.

(c) Calculate the Bayesian Information Criterion (BIC) for the three models. Which one performs best according to that metric?

Hint: When dealing with a regression model with k degrees of freedom, the residual variance $\hat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ is an unbiased estimator.