Computational Statistics II

Subash Kharel

20 Feb 2020

1. **Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio, and newspaper, rather than in terms of the coefficients of the linear model.**
   Answer: Table 3.4 shows the null hypothesis which says advertising budget on TV, radio or newspaper do not have any contribution on sales. Studying the p value at the last column of table we can reject the null hypothesis assumed for TV and radio. But for newspaper, null hypothesis cannot be rejected since the p-value is significant and we can say budget on advertising through newspaper can be discarded.
2. **Carefully explain the differences between the KNN classifier and KNN regression methods.**
   Answer: The basic difference between KNN classifier and KNN regression method is the way they predict the response. As classification is mentioned, the output needs to be one of the pre known class and for regression, the output can be any continuous value. In KNN classifier we classify an unknown item by calculating n nearest items and choosing the class with highest majority. Whereas in KNN regression, we calculate n nearest items and take the mean of those items to predict the value of unknown item.
3. **Question number 4:**
   I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta0 + \beta1X + \beta2X^2 + \beta3X^3 + \varepsilon$.

   a. **Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta0 + \beta1X + \varepsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.**
   Answer: As the true relationship between X and Y is linear, we can guess the training RSS for linear regression to be lower than that of cubic regression. The reason for saying this is that, linear regression can perfectly fit the training data (RSS will be 0) but cubic regression may not (producing some value for RSS). Still we cannot confirm this without knowing the nature of training data. Since

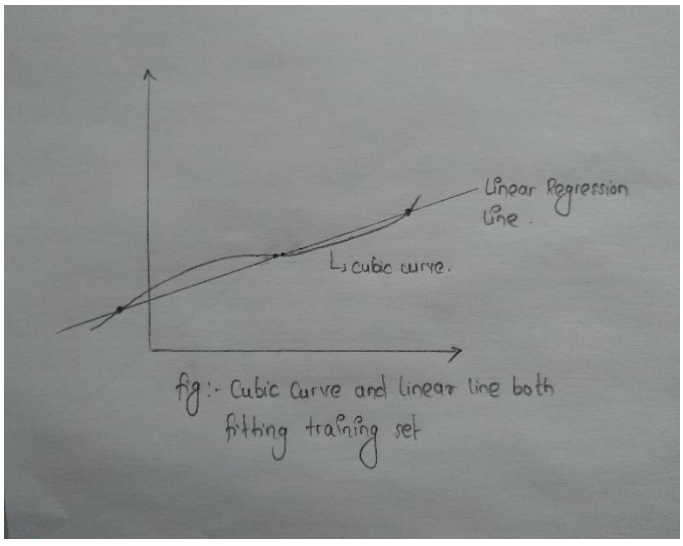

fig:- Cubic curve and linear line both fitting training set

sometime same linear data may be represented by non-linear equation which is shown in the figure to the right.

b. Answer (a) using test rather than training RSS.

Answer: Cubic (Non-linear) regression are supposed to have high chance of overfitting the training data and may not predict test data well resulting in high test RSS. Since the relation is linear, we expect linear regression to produce less test RSS. But none of this can be confirmed unless the nature of test data is determined.

c. **Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.**

Answer: The higher tendency of polynomial regression to fit the training data well will result in less RSS compared to that of linear regression. Thus, we can expect RSS for cubic regression to be lower.

d. Answer (c) using test rather than training RSS.

Answer: As the distance from linear is not known we cannot compare the value for test RSS. If its near to cubic, cubic regression will produce lesser RSS and if its near to linear, it will produce lesser RSS for linear regression.