Q12

Rahul Atre

2023-10-10

Q12 [Regression Modeling]

We collect a set of data (n = 100 observations) containing a single predictor X and a quantitative response Y. We then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.

1. Suppose that the true (statistical) relationship between X and Y is linear. Consider the training residual sum of squares (RSS) for the linear regression, and also 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.

Ideally, it would be beneficial to know more information to make a decisive statement on whether a cubic or linear regression would have lower training RSS. On one hand, since the true relationship between X and Y is linear, a linear regression could potentially have less error since it correctly fits the true data. For cubic regression though, it is possible for RSS to be low from overfitting the training data due to its added adjustability (i.e. having more parameters to model the data). Overall, more details would be required to justify which model is better suited for the given circumstances.

2. Answer question 1, using test RSS rather than training RSS.

For the test RSS, the cubic regression would certainly have a **higher RSS** than a linear regression. During the training process, the cubic regression will overfit the training data, leading to incorrect predictions on the true, linear relationship between X and Y. A linear regression will accurately fit the linear data for this particular situation.

3. Suppose that the true relationship between X and Y is not linear, but we don't know "how far" it is from being linear. Consider the training RSS for the linear regression, and 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.

In this particular scenario, since we know for certain that the true relationship is not linear, a cubic regression would always have **lower training RSS** since it is more flexible. It has more parameters to accurately fit the data points in the training set than a linear regression model.

4. Answer question 3, using test RSS rather than training RSS.

Similar to (1) but inversely, there is not enough information to make a decisive statement on whether a linear or cubic regression would have lower test RSS, if we don't know the true relationship of X and Y. Though a cubic regression models the data better from overfitting, there is no proof that the same accuracy would follow for the test data.

In large perspective, the question to be answered is, "how close is the relationship to being linear". If it is very close to linear, then obviously a linear regression would have less test RSS than cubic. Conversely, if it is far from linear, then cubic would have lower test RSS from higher flexibility to fit training data.

Ultimately, we are faced yet again with the bias-variance tradeoff. We are unsure at what point a model would overfit, or at what point the model would be biased from incorrect assumptions of the model.

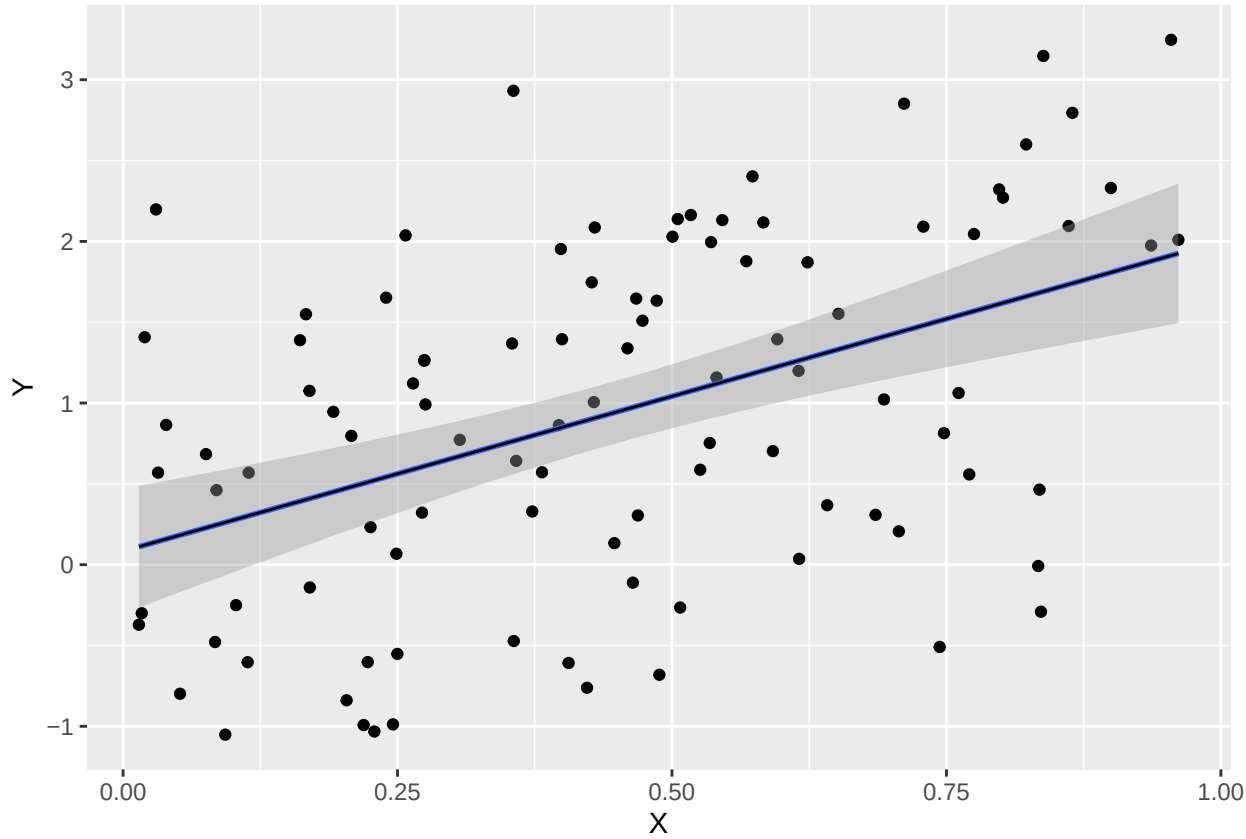
5. Generate some data to illustrate the situation of questions 1 and 2.

```
library(ggplot2) #Import ggplot2 package
set.seed(10)
n = 100
X = sort(runif(n)) #Uniform distribution generation of 100 variables
Y = 2*X + rnorm(n)
data = data.frame(X = X, Y = Y)

training_model = lm(Y ~ X)
predictions = predict(training_model, newdata = data)

ggplot(data, aes(x = X, y = Y)) + geom_point() + geom_smooth(method = 'lm') + geom_line(aes(y = predict
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Above we have generated the situations in question 1 and 2. In this particular case, where the true relationship between X and Y is linear, a linear regression model would be the best option suitable.