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. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.

   (a) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. 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.

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

   (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 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.

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

Answers:

a) We can expect the training RSS of the cubic regression to be lesser than the RSS of the linear regression.
This can occur due to overfitting the regression model (cubic regression). The cubic regression model might conform its regression line to fit the outliers

and error ϵ, even though the true relationship is linear.

b) If we use the test RSS, we can expect linear regression to have a lower RSS than cubic regression. This occurs since the true relationship of the data is linear (both test & train), while cubic regression could have overfit according to the train data.

c) We can expect the cubic regression model to have a lower RSS than the linear regression model since the true relationship between X & Y is non-linear. This occurs since a linear regression model would underfit the data.

d) We do not have enough information to answer this question since we do not know the actual true relationship between X and Y. The relationship between X and Y could be quadratic, cubic,

quartic, quintic, etc.
Our cubic regression model might be underfit, overfit, or the right fit for the data. We have no way of knowing this for sure.
The linear regression model is definitely underfit, but we don't know for sure if it will perform worse than the cubic regression model as we don't know the degree of non-linearity of the true relationship.