Thomas Dinh

DSE 6211

Week 6

Oct. 6, 2023

Exercises 6

1) Why do the two RMSE plots show very different behavior? Use underfitting and overfitting in your answer, as well as the bias-variance tradeoff.

Due to the characteristics of the data they are applied to, the two RMSE graphs display different behaviors. The first RMSE plot was created using the original dataset, whilst the second one was tested using new data. The first plot shows a low RMSE at a polynomial degree of 25, indicating a model that may have overfit the training data. This model, however, produces noticeably higher RMSE values when tested on fresh data, indicating overfitting and poor generalization. On the other hand, both graphs show elevated RMSE values at a polynomial degree of 1, which is typical of an underfitting and biased model that is oversimplified and fails to account for the complexity and underlying patterns of the data. The model at 25 degrees, in particular, shows substantial variance, having grown too complex and catching both real data patterns and training data noise. In contrast, the model at 1 degree exhibits bias by relying on overly broad assumptions that do not reflect the actual distribution of data.

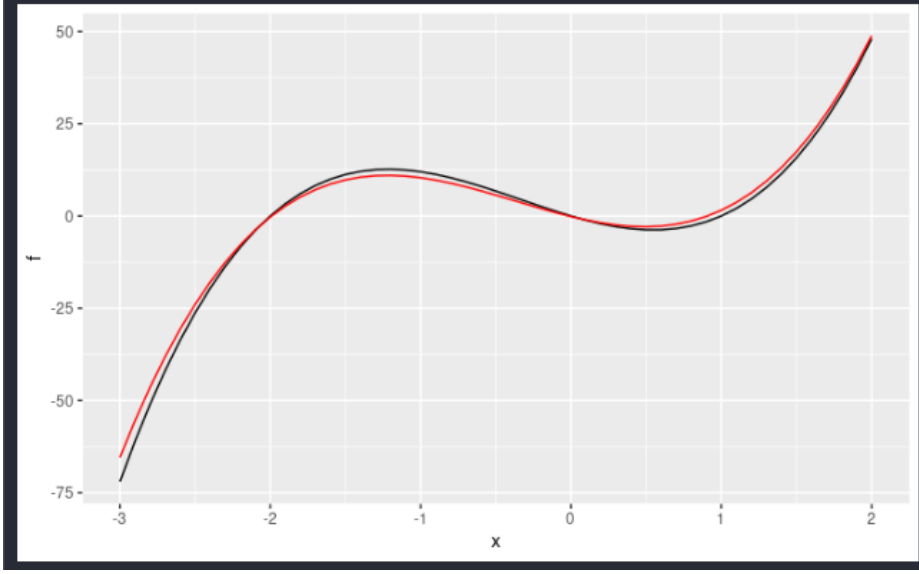
2) Using the last plot above, what degree of polynomial should we choose? After running all of the above code, substitute the degree you chose for the question mark in the code below and run it. Copy and paste the resulting plot.

I would either use degree 3 or 4 but leaning more towards 4 because of its low RMSE, in both plots.

```
ggplot() +
  geom_line(data = results, aes(x = degree, y = rmse), color = "black")

model <- lm(observations ~ poly(x, 3))
predictions=predict(model, newdata = data.frame(x=x))
data = data.frame(x=x, f=f, predictions=predictions)
ggplot(data, aes(x=x)) +
geom_line(aes(y = f), color = "black") +
geom_line(aes(y = predictions), color = "red", linetype="solid")

```




```
model <- lm(observations ~ poly(x, 4))
predictions=predict(model, newdata = data.frame(x=x))
data = data.frame(x=x, f=f, predictions=predictions)
ggplot(data, aes(x=x)) +
geom_line(aes(y = f), color = "black") +
geom_line(aes(y = predictions), color = "red", linetype="solid")

```