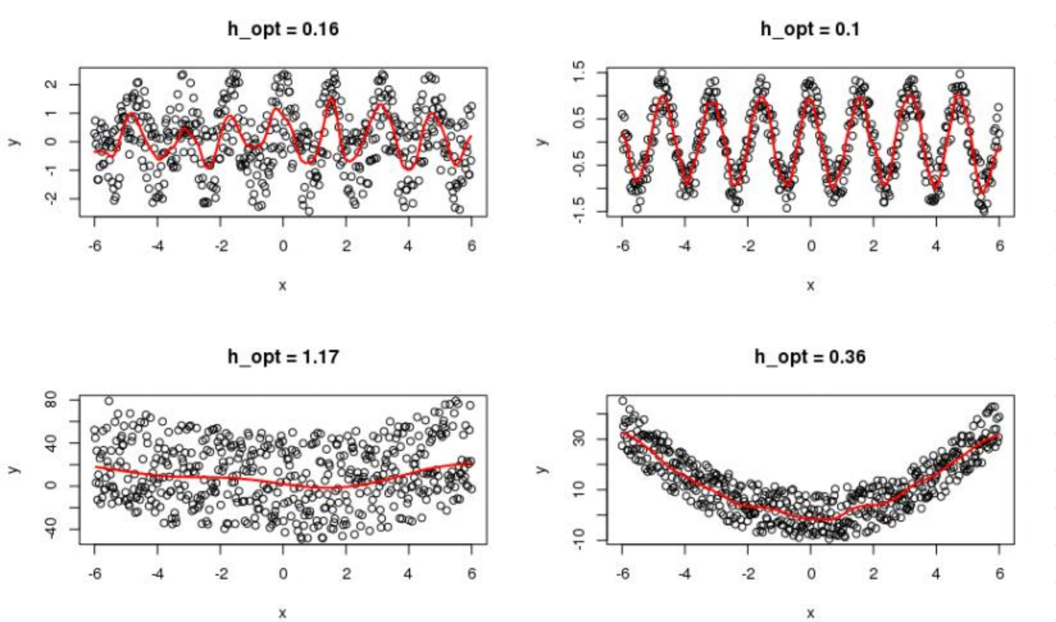
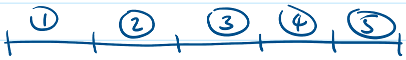
A) Refer to the crossVal function in the HW_3_functions.R file.

B) Using cross validation, pick optimal h values for a wiggly and smooth function with high and low noise. Does your out of sample predictive validation method lead to reasonable values of h?



The wiggly function I chose was $y = \cos(4x)$ and the smooth function was $y = x^2$. For both functions, the high noise produced a higher h. This makes sense because when the noise is very high, it becomes much less obvious where the curvatures in the original function are. Therefore, the smoother is more likely to fit a "flatter" estimate. The lower left plot is a good visual example. The underlying function is $y = x^2$ but the high noise ratio hides the curvature and the fitted line is almost horizontal. With low noise, the cross validation was able to select much smaller values for h, though not so small as to overfit the data.

C) What's the problem with K-fold cross validation?



folds help minimize the overlap between training sets
Say we're spliting by using 80% to train & 20% to test.

We introduce bias by estimating generalization error of the data set by using only 80% of the data.

Full data: (y_i, x_i) for i in $\{1, \dots, N\}$

(x^*, y^*), some future point

Goal: estimate $E\left([y^* - \hat{f}_N(x^*)]^2\right) = MSE$

Goal: estimate $E\left([y^* - \hat{f}_N(x^*)]^2\right) = MSE$

estimate from N points

do train test split

$Tr \subset [1, \dots N]$ $|Tr| = N_{Tr}$

$|Te| = N_{Te}$

1.) estimate $\hat{f}_{N_{TR}}(x)$ using training data

2.) $\widehat{MSE} = \frac{1}{N_{Te}} \sum_{i \in T_e} (y_i - \hat{f}_{N_{Tr}}(x_i))^2$

On average, $\widehat{MSE}$ is a larger number than MSE cuz $\hat{f}_{N_{Tr}}$ is different from $\hat{f}_N$

This makes h underfit the data. This creates high bias.

To minimize bias, we can do LOOCV

$\widehat{MSE}_{LOO} = \sum_{i=1}^{n} (y_i - \hat{f}_{(-i)}(x_i))^2$ $\hat{f}_{(-i)}$ = fit w/point i removed

now $\mathrm{cov}(\hat{f}_{(-i)}, \hat{f}_{(-j)})$ is high. There's high variance

$Var\left(\frac{1}{n} \sum \varepsilon_i^2\right) = \frac{1}{n^2} \sum Var(\varepsilon_i^2)$ if ε_i are independent

but now ε_i are highly correlated in LOOCV,

so $Var\left(\frac{1}{n} \sum_i \varepsilon_i^2\right) = \frac{1}{n} \sum Var(\varepsilon_i) + 2 \sum_{i<j} Cov(\varepsilon_i \varepsilon_j)$

so now since $\hat{f}_{(-i)}$ & $\hat{f}_{(-j)}$ are highly correlated,

$Var\left(\frac{1}{n} \sum_{i=1}^{n} \widehat{MSE}_i^2\right)$ is large cuz MSE_i are correlated.