850 Homework 1

Yifang Zhang

Problem 1

Proof:

According to the question, the expected prediction/test error at x_0 is:

$$EPE(x_0) = E_{Y_0|X_0=x_0} E_\tau [f(\hat{x_0}) - Y_0]^2$$

Since:

$$E_\tau[f(\hat{x_0}) - Y_0]^2 = E_\tau[f(\hat{x_0})^2 - 2Y_0 f(\hat{x_0}) + Y_0^2] = Y_0^2 - 2Y_0 E_\tau(f(\hat{x_0})) + E_\tau[f(\hat{x_0})^2]$$

And we have:

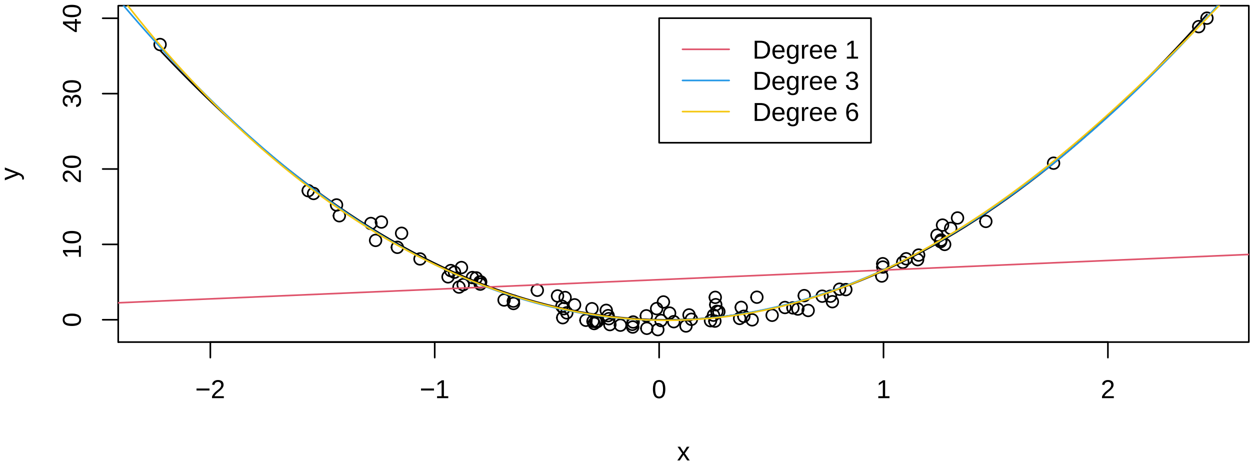
$$E_\tau[f(\hat{x_0})^2] = E_\tau^2[f(\hat{x_0})] + Var_\tau[f(\hat{x_0})]$$
$$E_{Y_0|X_0=x_0}(Y_0^2) = E^2_{Y_0|X_0=x_0}(Y_0) + Var(Y_0)$$

Then we get:

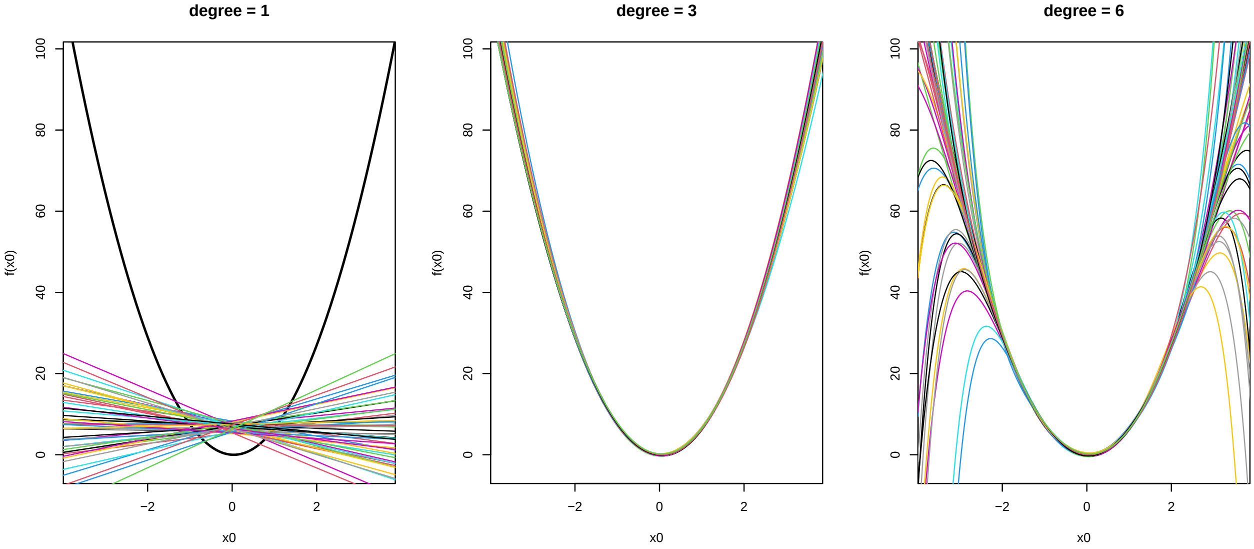
$$\begin{aligned} EPE(x_0) &= E_{Y_0|X_0=x_0}(Y_0^2) - 2E_{Y_0|X_0=x_0}(Y_0)E_\tau[f(\hat{x_0})] + E_\tau^2[f(\hat{x_0})] + Var_\tau[f(\hat{x_0})] \\ &= Var(Y_0) + E^2_{Y_0|X_0=x_0}(Y_0) - 2E_{Y_0|X_0=x_0}(Y_0)E_\tau[f(\hat{x_0})] + E_\tau^2[f(\hat{x_0})] + Var_\tau[f(\hat{x_0})] \\ &= \sigma^2 + Var_\tau[f(\hat{x_0})^2] + \{E_{Y_0|X_0=x_0}(Y_0) - E_\tau[f(\hat{x_0})]\}^2 \\ &= \sigma^2 + Var_\tau[f(\hat{x_0})^2] + Bias^2[f(\hat{x_0})] \end{aligned}$$

Problem 2

(1) Plot polynomial regression of degree 1,3,6 on the first training set, along with the true curve f.



(2) Plot the polynomial regression curves on all the 50 training set.



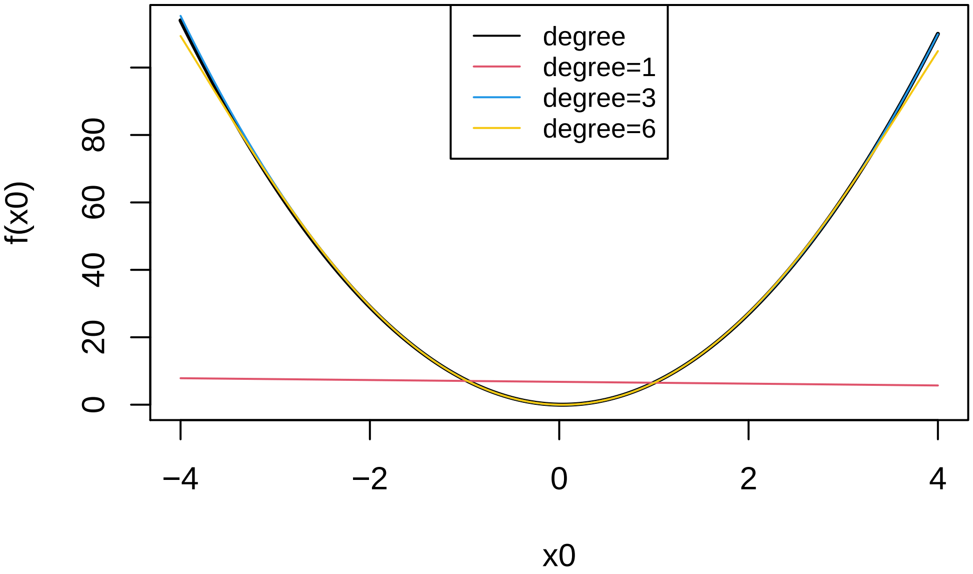
Observation:

i. When we set the degree to be 1, we only get straight lines and in fact we are underfitting the true relationship between Y and X given by the question. Since the straight line is unable to capture the patterns in the data, we will get high bias with low variance;

ii. When we set the degree to be 6, we are overfitting the true relationship escpecially for some training sets as x approaches to the extreme value. So we will get low bias but high variance under the degree 6.

iii. When we set the degree to be 3, almost on every training set the polynomial regression fitted the true curve very well. Thus we get the best fitting curve comparing to other two models with low bias and low variance.

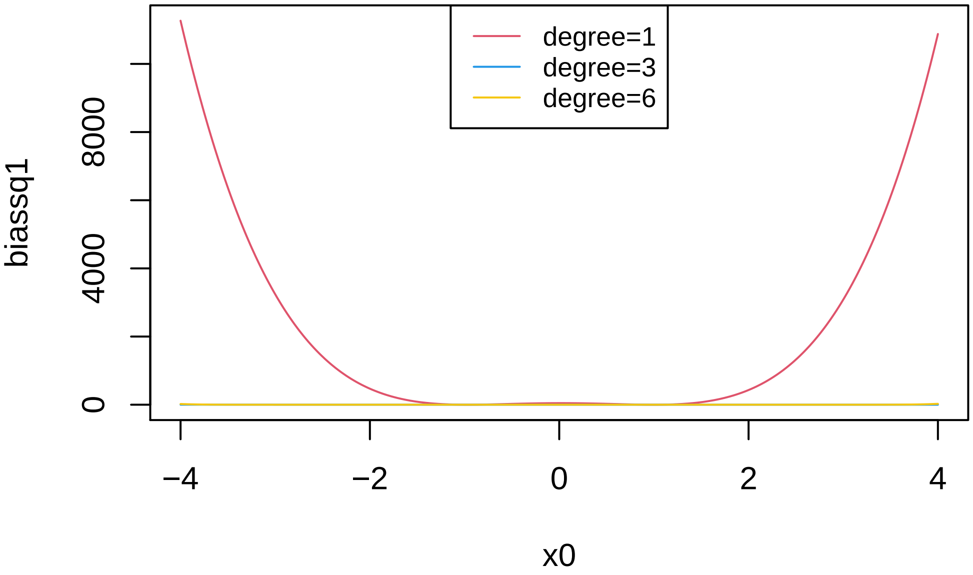
(3) Compare the expected value $E(f(\hat{x_0}))$ over the grid x_0.



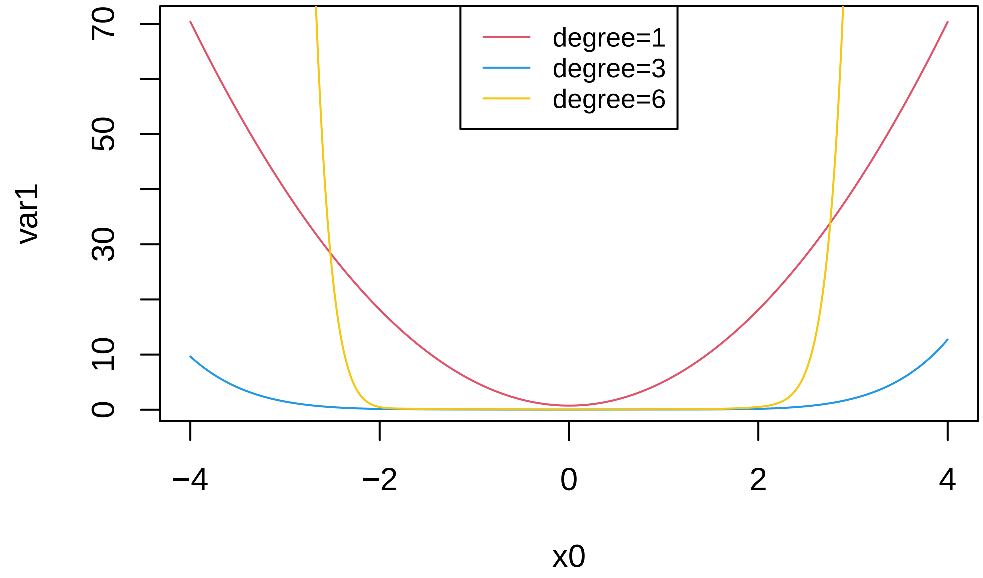
Observation:

We can find that on average, the blue curve (degree=3) is the closest curve to the true curve and that agrees with our intuition since when degree=3, the polynomial regression fitted very well on almost every training set. The yellow curve (degree=6) is not nice fit when x approaches the extreme values, which can be easily observed by the pervious plot on all 50 training sets.

(4) Compute the sqaured bias $Bias^2(f(\hat{x_0}))$ over the grid x_0.

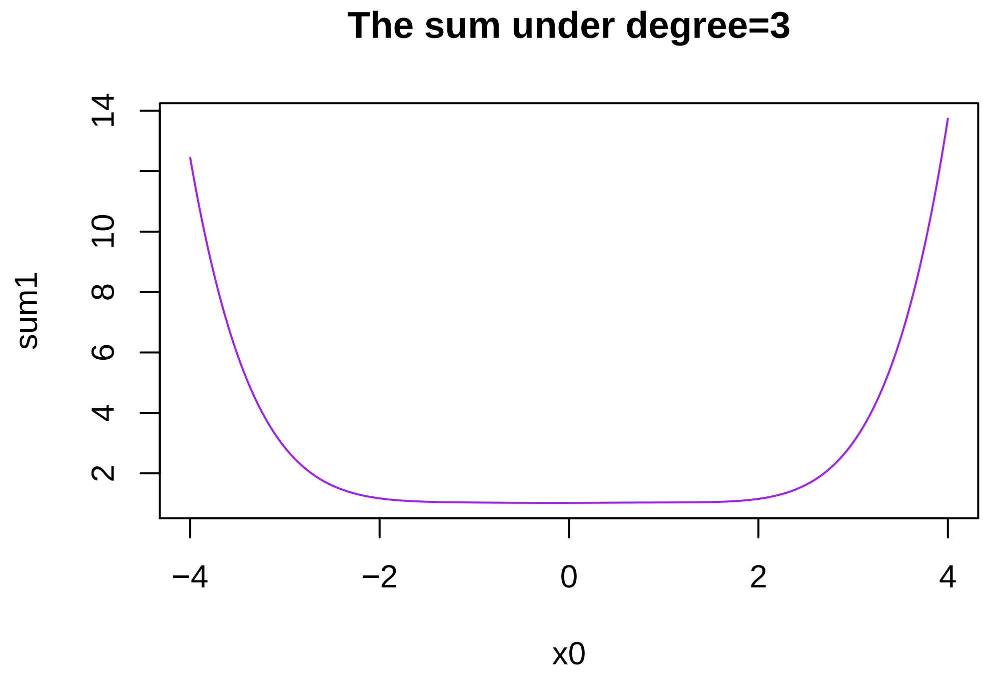


(5) Compute the variance $Var(f(\hat{x_0}))$ over the grid x_0.



(6) Plot the curve $\sigma^2 + Var(f(\hat{x_0}) + Bias^2(f(\hat{x_0}))$ for the polynomial regression of degree 3.

Since $\sigma^2 = Var(\epsilon)$ is the irreducible error, we can simply set $\sigma^2 = 1$. Then the curve $\sigma^2 + Var(f(\hat{x_0}) + Bias^2(f(\hat{x_0}))$ over the grid x_0 for the polynomial regression of degree 3 is:



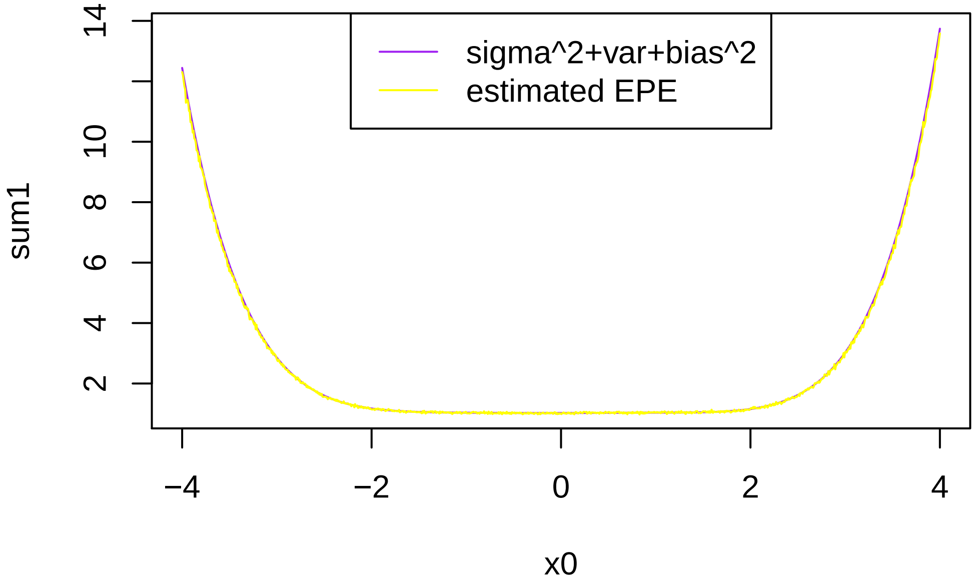
(7) Compute the value $EPE(x_0)$ from simulation.

To compute the value $EPE(x_0)$, we simulation the corresponding y_0 t times by the following fomula $y_0 = 7x_0^2 - 0.5x_0 + \text{rnorm}(1)$ and for each time, we calculate $(f(\hat{x_0}) - y_0^2)$. Finally, we will take the average mean of those values to estimate $EPE(x_0)$.

The corresponding R code for estimating $EPE(x_0)$ is:

```
t=100
EPE=seq(0,0,length=1000)
for (i in 1:1000){
  sum=0
  for (j in 1:50){
    for (k in 1:t){
      sum=sum+(linmod2_pred[j,i]-f(x0[i])-rnorm(1))^2
    }
  }
  EPE[i]=sum/(50*t)
}
```

(8) Plot the above $EPE(x_0)$ curve overlaying the curve from part 6.

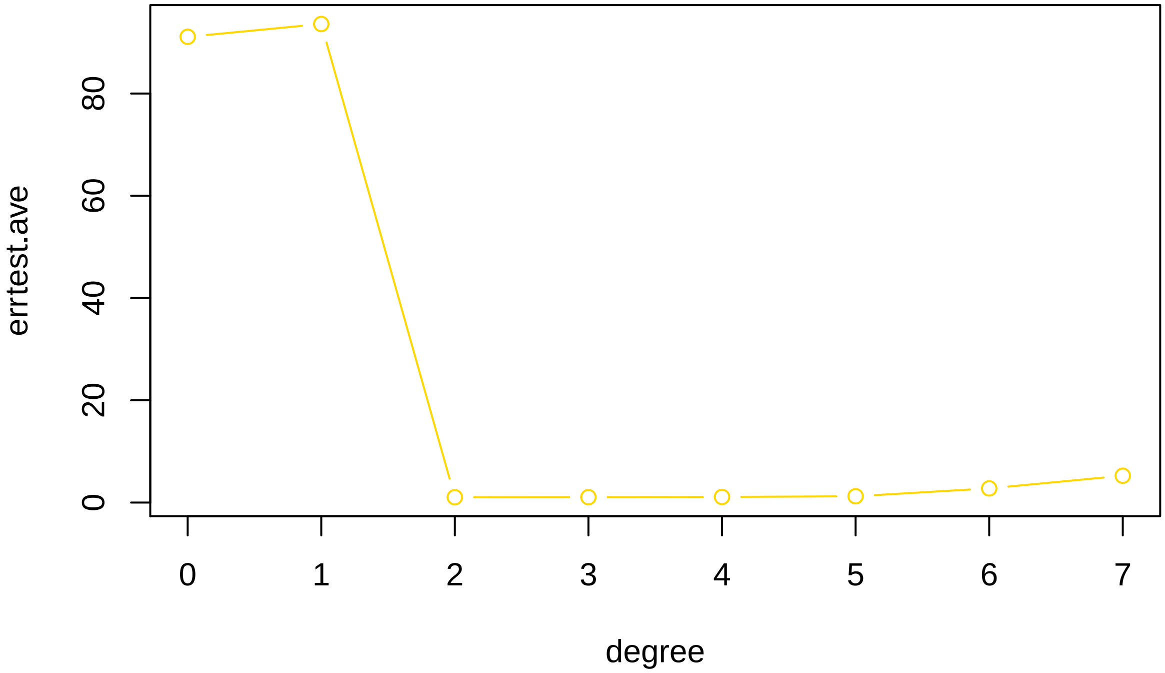


Clearly, the plot shows that our two curves are highly coincident and we can say they are equal.

Problem 3

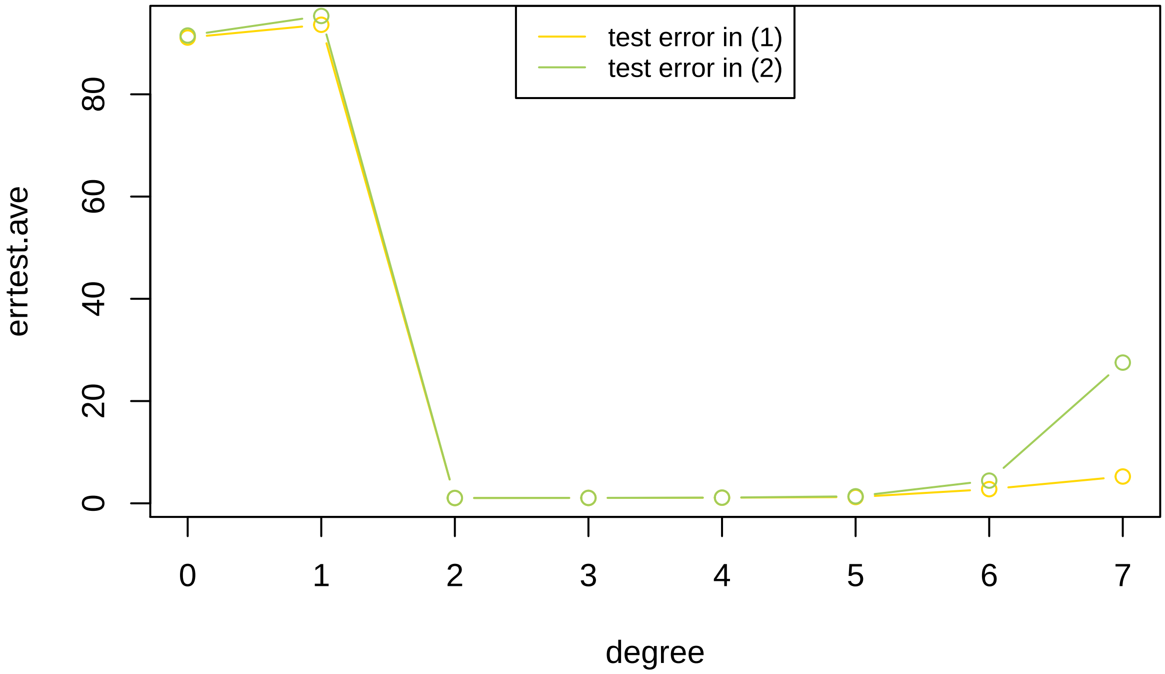
(1) Plot the expected test error curve.

```
## [1] 91.097337 93.602456  1.038090  1.052597  1.094925  1.228407  2.767159
## [8]  5.241646
```



(2) Plot the average test error curve.

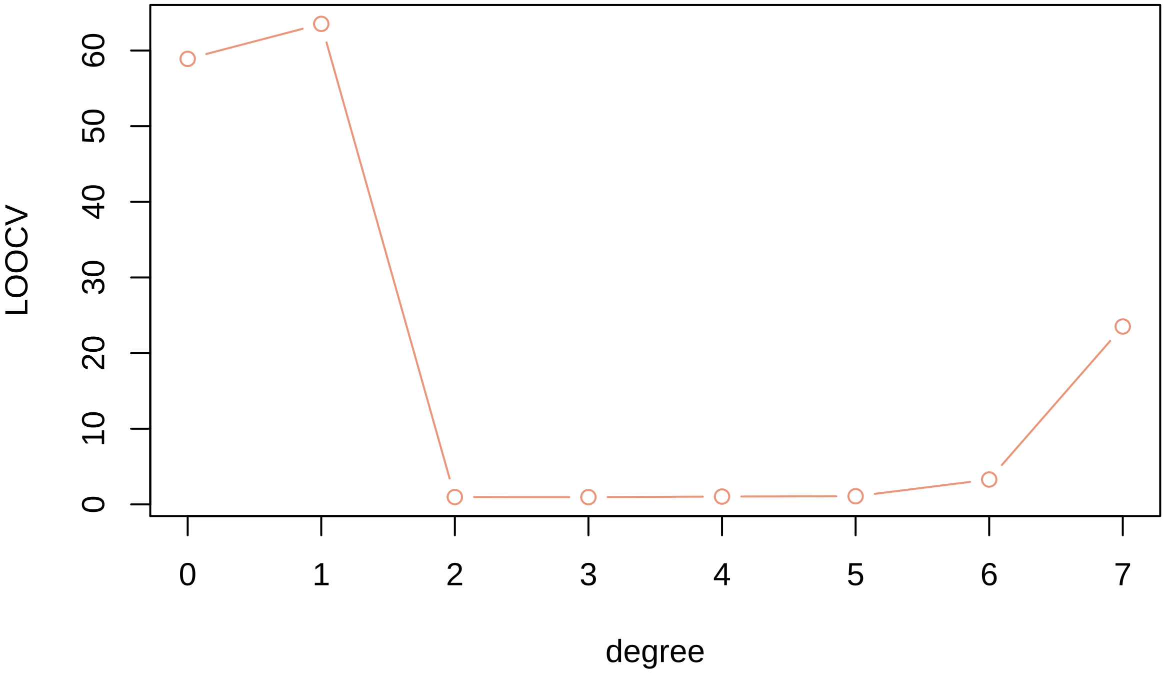
```
## [1] 91.493947 95.328660  1.039379  1.057866  1.116131  1.369846  4.439794
## [8] 27.528713
```



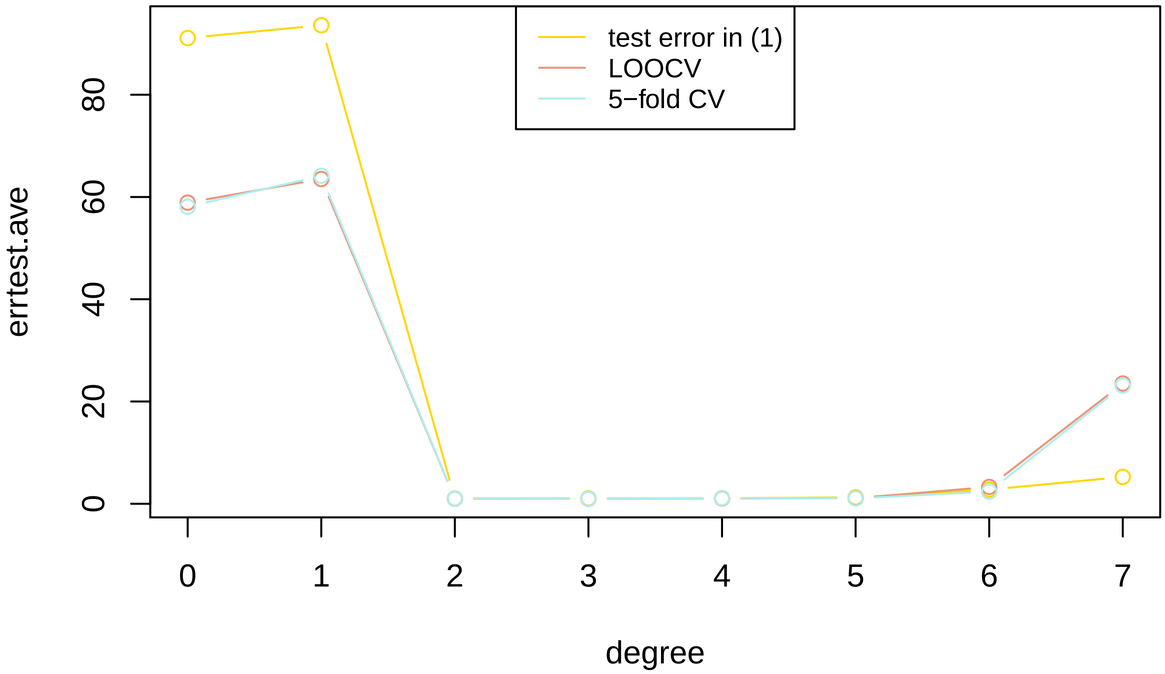
We can find that when the degree is small, there exists only slight difference bewtween two test errors. And when the degree becomes large, the difference becomes larger. That because when we have large degree, we are overfitting the model and under the second methods we are fitting much more data points comparing to the first method. Overfitting means low bias but high variance. That's why the test error of the second methods will become larger when the degree becomes larger.

(3) Compute and plot the LOOCV estimate for polynomial of degree 0 to 7.

```
## [1] 58.8958983 63.5190789  0.9739835  0.9614573  1.0360458  1.0794330  3.2904076
## [8] 23.5172159
```

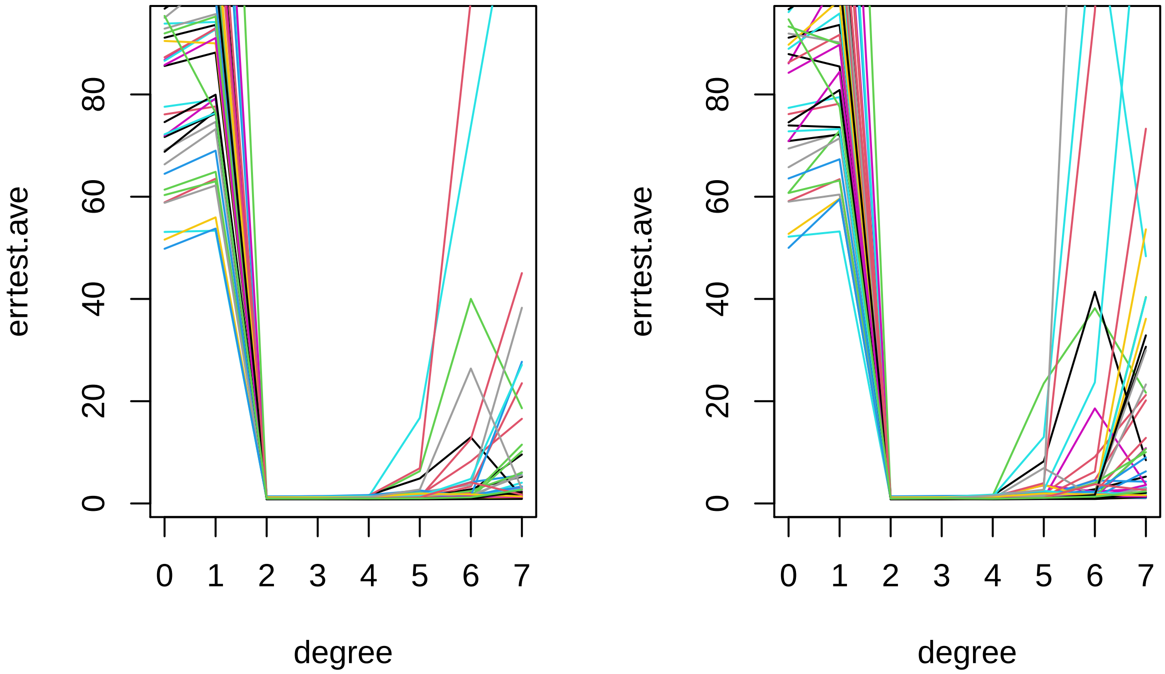


```
## [1] 58.1015318 64.1406405  0.9954780  0.9547898  1.1125524  1.0375793  2.3689596
## [8] 23.1130824
```



Observation: We can find that almost there is no difference between LOOCV and 5-fold CV for polynomial of degree from 0 to 7. Secondly, the expected test error is larger then LOOCV and 5-fold CV when the degree is small. While when the degree is large (as degree=7), LOOCV and 5-fold CV are larger then the expected test error. Finally, when the degree is between 2 and 6, all three errors are very small and close to each other.

(4) Compute and plot the LOOCV estimate for polynomial of degree 0 to 7 for all the 50 training dataset.



Comment: We can find that when the degree is too small or too large, both LOOCV and 5-fold CV becomes really large and unstable for the 50 training dataset. While when the degree is between 2 to 4, for all 50 training sets, LOOCV values are almost the same and really small. It is similar for 5-fold CV plot but is not as nice as LOOCV plot shows us.

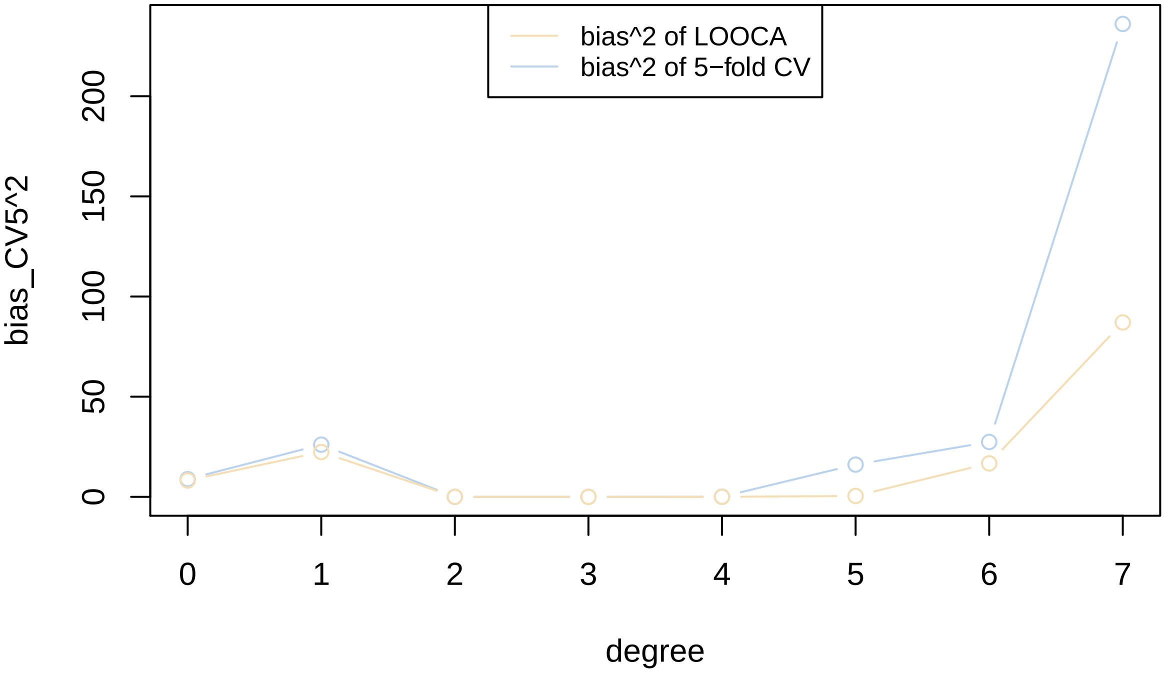
(5)(6) Compute the squared bias and variance of the LOOCV and $5 - fold$ CV.

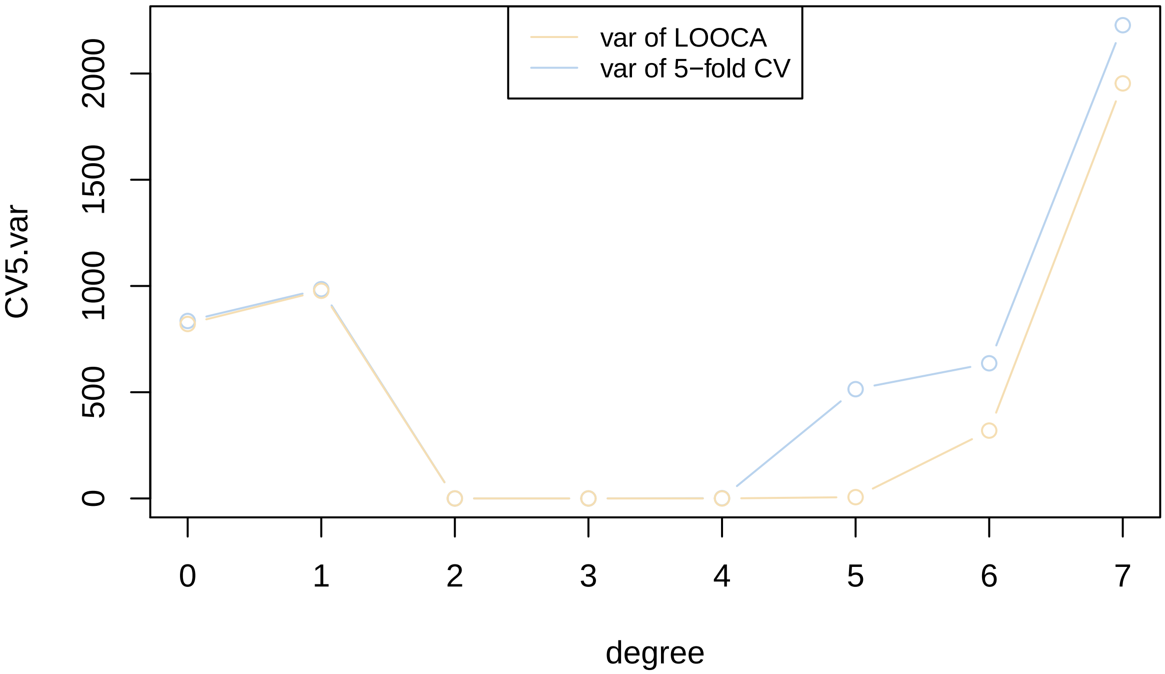
```
## [1]  2.86020342  4.73014316 -0.01310267 -0.01030772  0.01966120  0.65998775
## [7]  4.08734031  9.33073131
```

```
## [1] 8.211773e+02 9.775417e+02 2.145755e-02 2.432380e-02 3.930491e-02
## [6] 6.100899e+00 3.193590e+02 1.953538e+03
```

```
## [1]  2.9710598896  5.1038901295 -0.0006303124  0.0114517336  0.2049033389
## [6]  4.0128857174  5.2342650532 15.3647186646
```

```
## [1] 8.352399e+02 9.848377e+02 2.338472e-02 2.647855e-02 1.089839e+00
## [6] 5.143777e+02 6.361300e+02 2.227358e+03
```





Comment:

1. For the bias, we can find that the square bias of the LOOCV and 5-fold CV are quite small and close to each other when we set degree is no more than 6, while when the degree is 7, there exists a large difference between $Bias^2$(LOCCV) and $Bias^2$(5-fold CV). In addition, when degree=7, the square bias of 5-fold CV is really large. Furthermore, the square bias of 5-fold CV is consistently larger then the square bias of LOOCV.

2. For the vairance, we can find that there only exists slight difference between LOCCV and 5-fold CV except for degree =5, which might need to check again. But in general, the trendence of these two curves are almost the same. Variance are really small when the degree is set to be between 2 to 5 for LOCCV and between 2 to 4 for 5-fold CV. Finally, comparing to small degree (0 or 1), the variance of LOCCV and 5-fold CV are quite large when the degree is big (degree = 7).