Regression basics, Part 2

Werner Stuetzle

April 2, 2020

Recap

Problem: Given training sample $(x_1, y_1) \dots (x_n, y_n)$ make prediction rule $f(x)$ that predicts value of response variable Y for given value x of predictor variable X

Formalization:

- Regard training sample as iid obs of pair of random variables (X, Y)
- Measure performance of prediction rule $f(x)$ by expected squared prediction error

$$ESE = E_{x,y}\left[(Y - f(X))^2\right]$$

$$= \int\int (y - f(x))^2 \, p(x, y) \, dx \, dy$$

Note: ESE does not make sense for classification problems

Key result:

- The optimal prediction rule for predictor value x is the conditional expectation

$$f(x) = E(Y|x)$$

- The ESE of the optimal rule is

$$ESE = E_X[V(Y|x)]$$

Result not directly useful because we don't know $p(x,y)$ - we only have a sample

Problem: How to estimate the conditional expectation $f(x) = E(Y|x)$?

Natural idea: Local averaging:

$$\hat{f}(x) = \text{mean}(y_i \mid |x - x_i| \leq h)$$

Choice of h controls degree of smoothness (more later)

Better version - Gaussian Kernel

Define $\varphi_\sigma(x)$ = Gaussian density with mean $\mu = 0$ and sd σ.

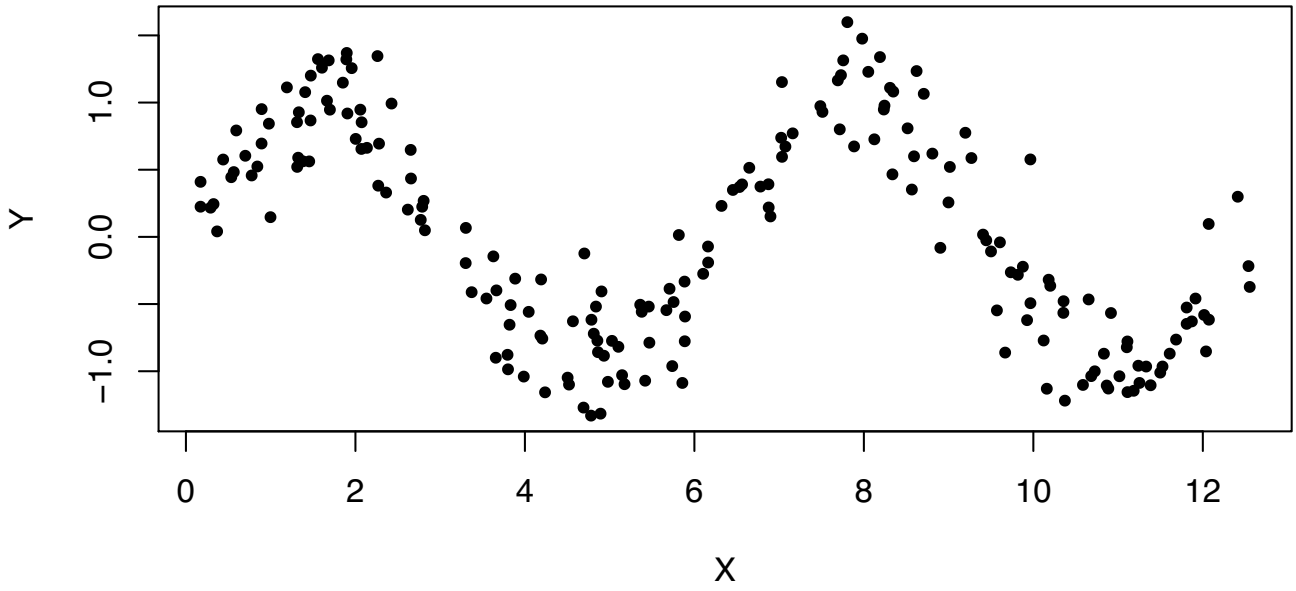
$$\hat{f}(x) = \sum_i y_i \varphi_\sigma(x - x_i) \Big/ \sum_i \varphi_\sigma(x - x_i)$$

Many other Kernels have been proposed

R code for next eight slides courtesy of Yen-Chi Chen

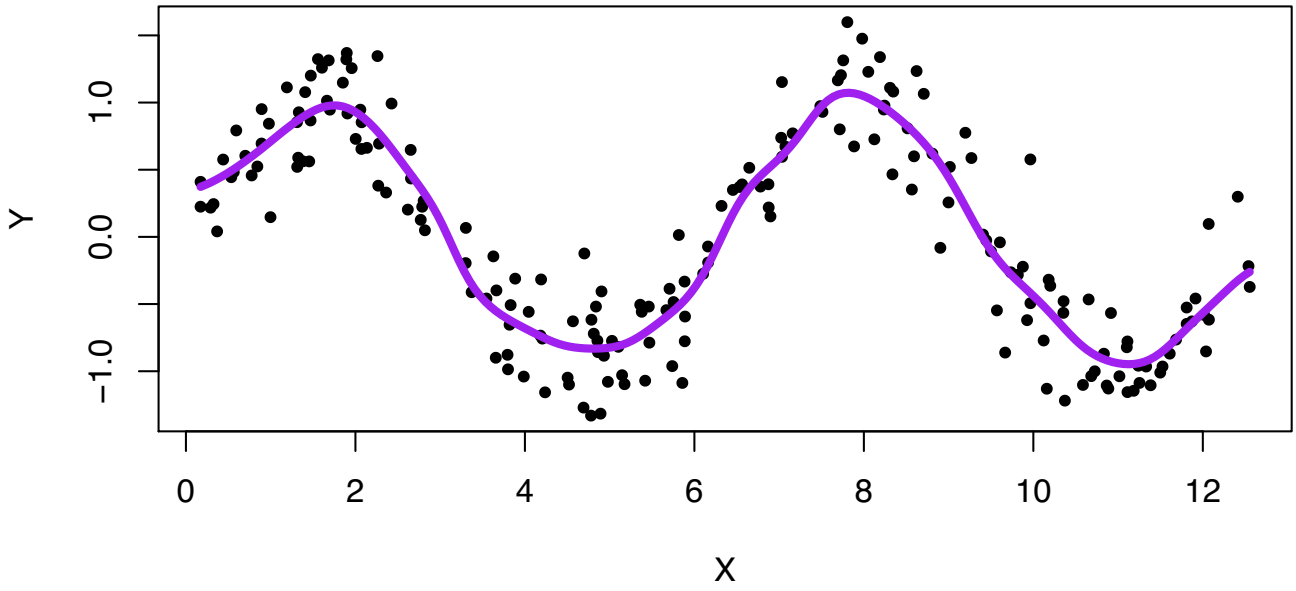
A simulated data set

```
X = sort(runif(200, min=0, max=4*pi))
Y = sin(X) + rnorm(200, sd=0.3)
plot(X,Y, pch=20)
```



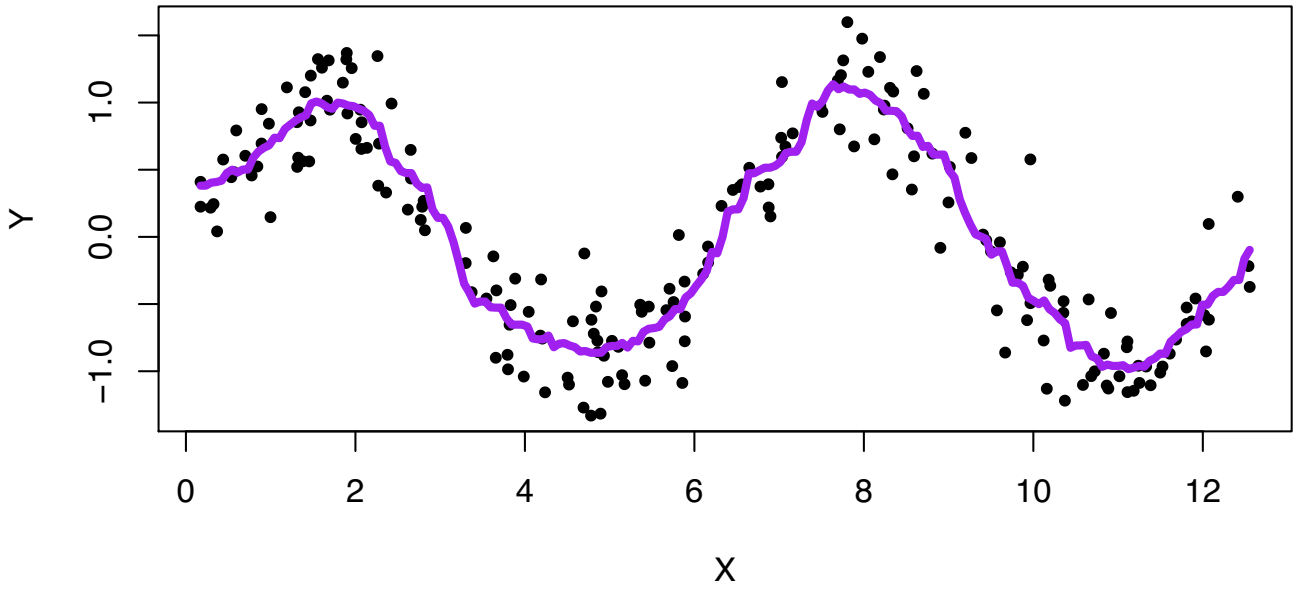
Local averaging, Gaussian kernel

```
Kreg = ksmooth(x=X,y=Y,kernel = "normal",bandwidth = 0.9)
plot(X,Y,pch=20)
lines(Kreg, lwd=4, col="purple")
```



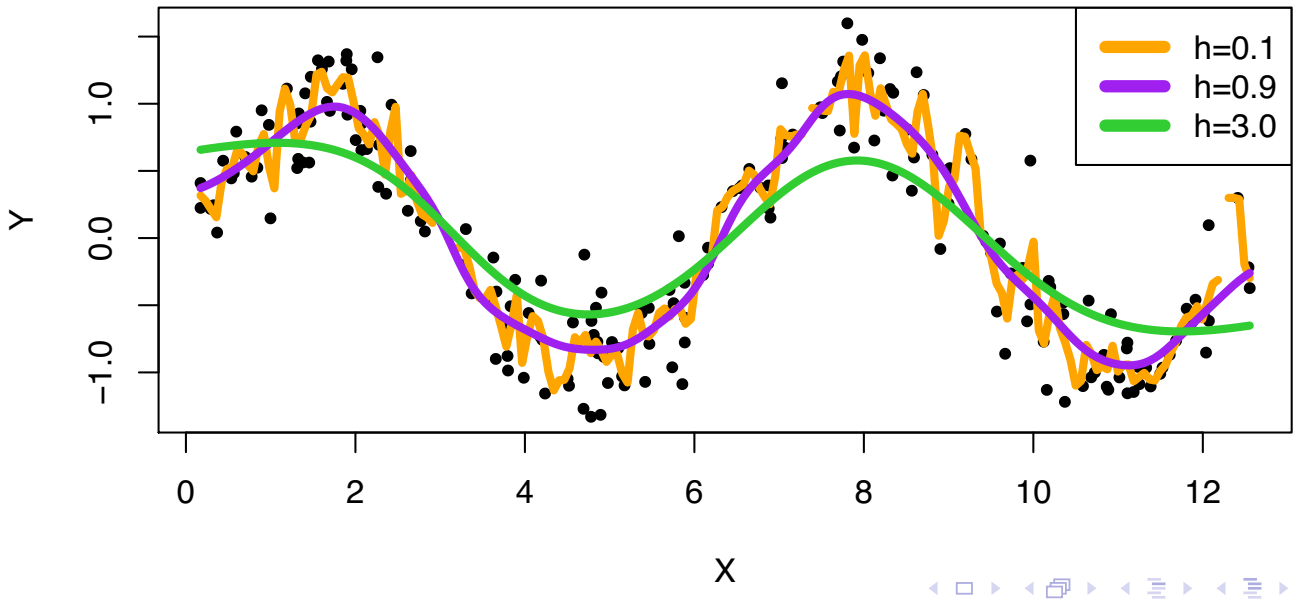
Local averaging, box kernel

```
Kreg = ksmooth(x=X,y=Y,kernel = "box",bandwidth = 0.9)
plot(X,Y,pch=20)
lines(Kreg, lwd=4, col="purple")
```



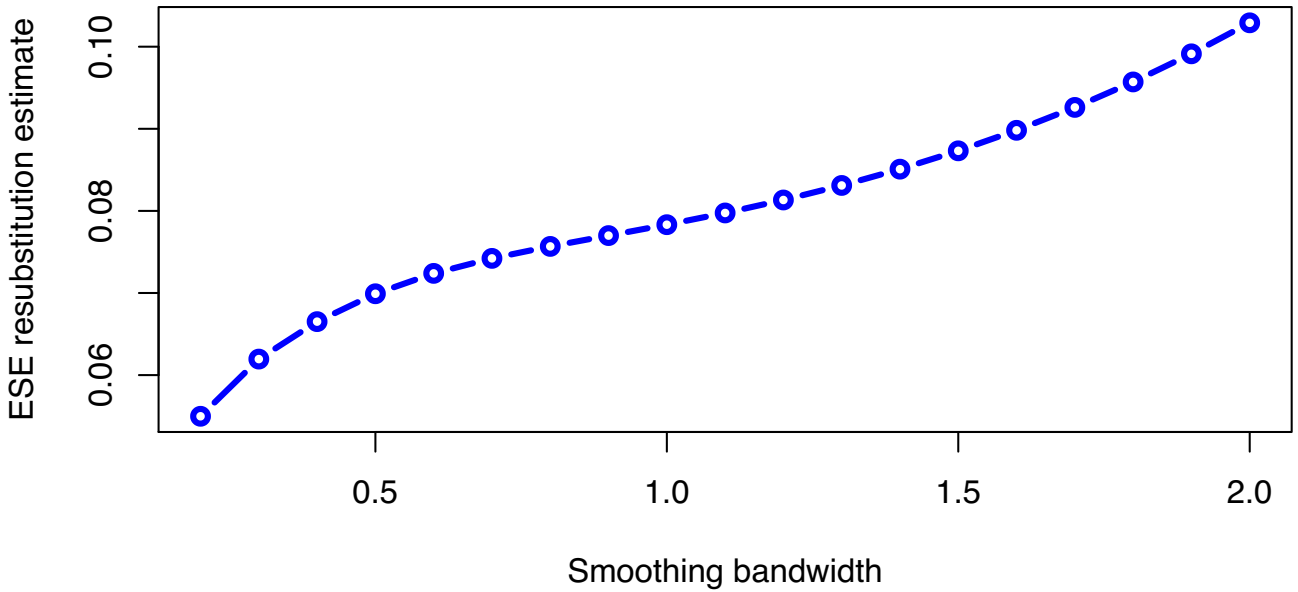
Different bandwidths

```
Kreg1 = ksmooth(x=X,y=Y,kernel = "normal",bandwidth = 0.1)
Kreg2 = ksmooth(x=X,y=Y,kernel = "normal",bandwidth = 0.9)
Kreg3 = ksmooth(x=X,y=Y,kernel = "normal",bandwidth = 3.0)
plot(X,Y,pch=20)
lines(Kreg1, lwd=4, col="orange")
lines(Kreg2, lwd=4, col="purple")
lines(Kreg3, lwd=4, col="limegreen")
legend("topright", c("h=0.1","h=0.9","h=3.0"), lwd=6,
       col=c("orange","purple","limegreen"))
```



ESE resubstitution estimate ("training error")

$$ESE_{train} = \frac{1}{n} \sum (y_i - \hat{f}(x_i))^2$$




```
## hopt =  0.2
```

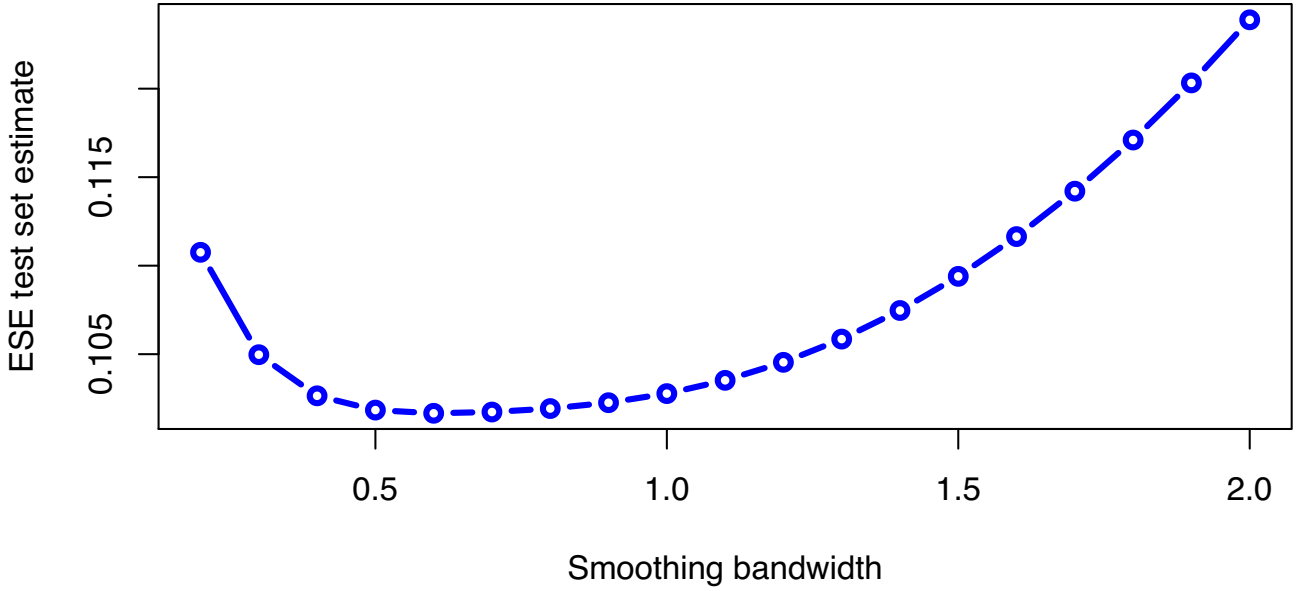
Can always achieve train error = 0 by choosing tiny bandwidth (no averaging)

Training error is biased estimate of ESE

(see next slide)

ESE test set estimate $(\tilde{x}_1, \tilde{y}_1), \ldots (\tilde{x}_m, \tilde{y}_m)$ $iid \sim p(x, y)$

$$ESE_{test} = \frac{1}{m} \sum_i (\tilde{y}_i - \hat{f}(\tilde{x}_i))$$

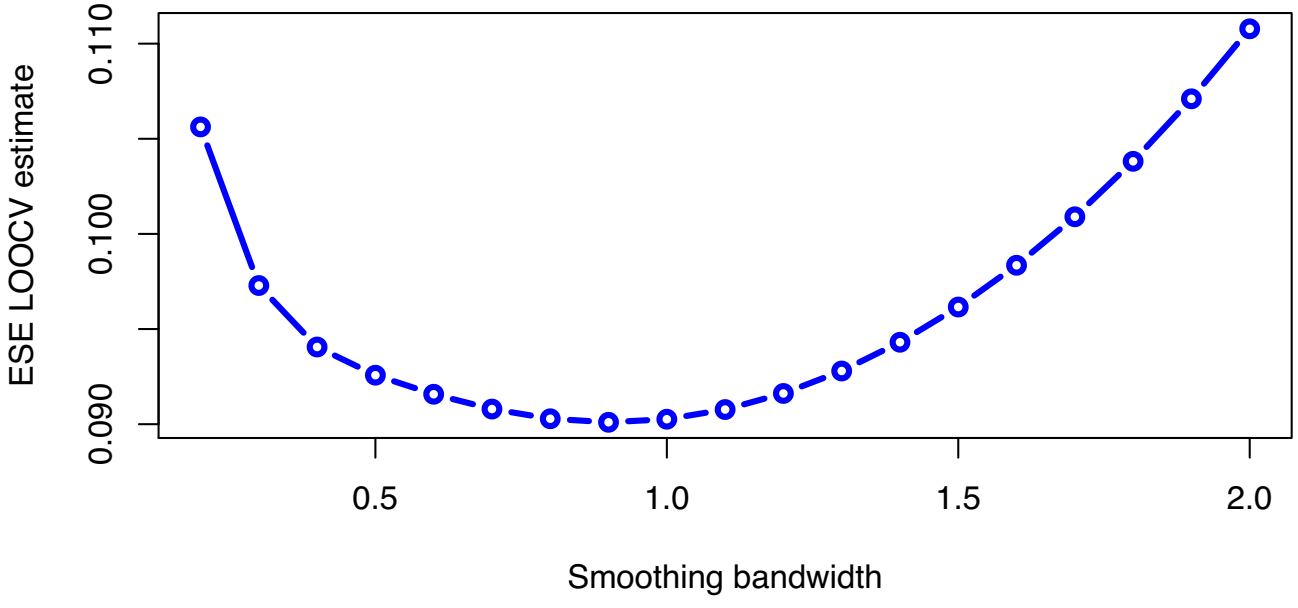



```
## hopt =  0.6
```

Note:

- Because these are simulated data we can generate an independent test set
- In practice one often (randomly) divides available data in training and test sets

ESE leave-one-out cross-validation estimate LOOCV




```
## hopt =  0.9
```

Define $\hat{f}_{-i}$ = estimate of $E(Y|x_i)$ computed without using training obs (x_i, y_i)

Cross-validation estimate

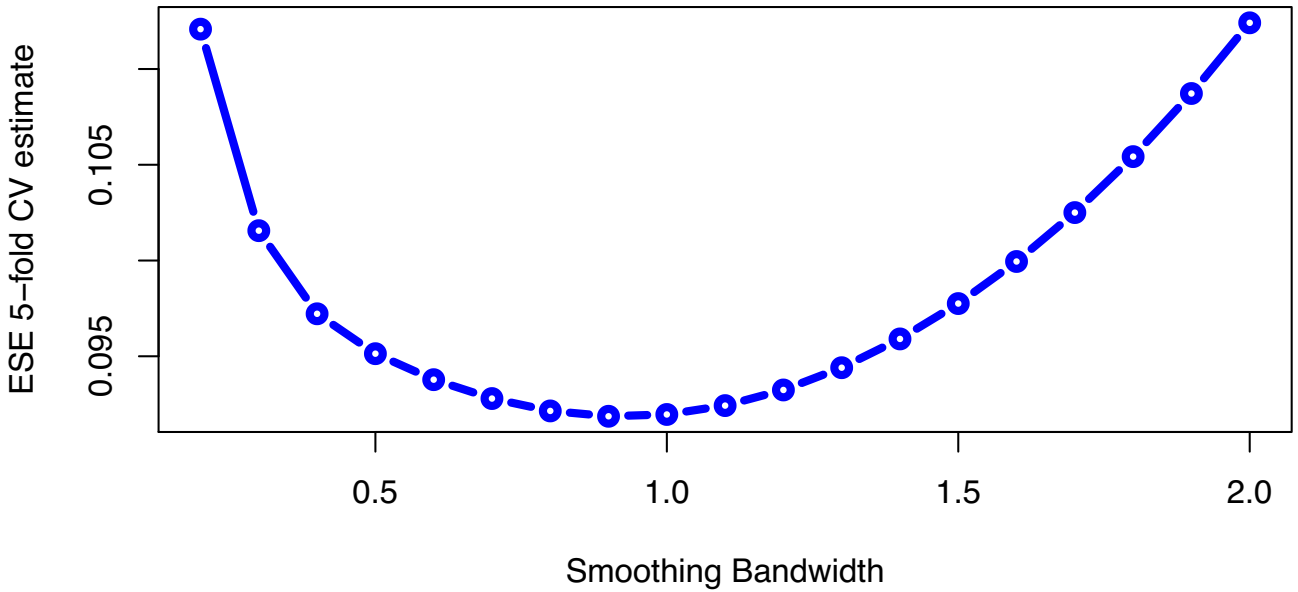
$$ESE_{CV} = \frac{1}{n} \sum \left(y_i - \hat{f}_{-i}(x_i)\right)^2$$

For Kernel smoother with Kernel $K(x)$ (think $K(x) = \Phi_6(x)$)

$$\hat{f}_{-i}(x_i) = \sum_{j \neq i} y_j K(x_j - x_i) \Big/ \sum_{j \neq i} K(x_j - x_i)$$

Observation (x_i, y_i) does not influence its predicted value

ESE 5-fold cross-validation estimate




```
## hopt =  0.9
```

Same idea as LOOCV but often cheaper

Randomly divide training sample S into K (5) equal size subsets $S_1 \dots S_K$

$RSS = 0$

For $i = \dots K$

Compute $\hat{f}_{-S_K}$ = smooth computed from $S - S_K$

$$RSS = RSS + \sum_{i \in S_K} (y_i - \hat{f}_{-S_K}(x_i))$$

$$ESE_{CV} = RSS/n$$

- LOOCV and 5-fold CV give the same estimate for the optimal bandwidth
- The test set estimate of the optimal bandwidth is a little off, but the valley is pretty flat

Small bandwidth ⇔ high flexibility

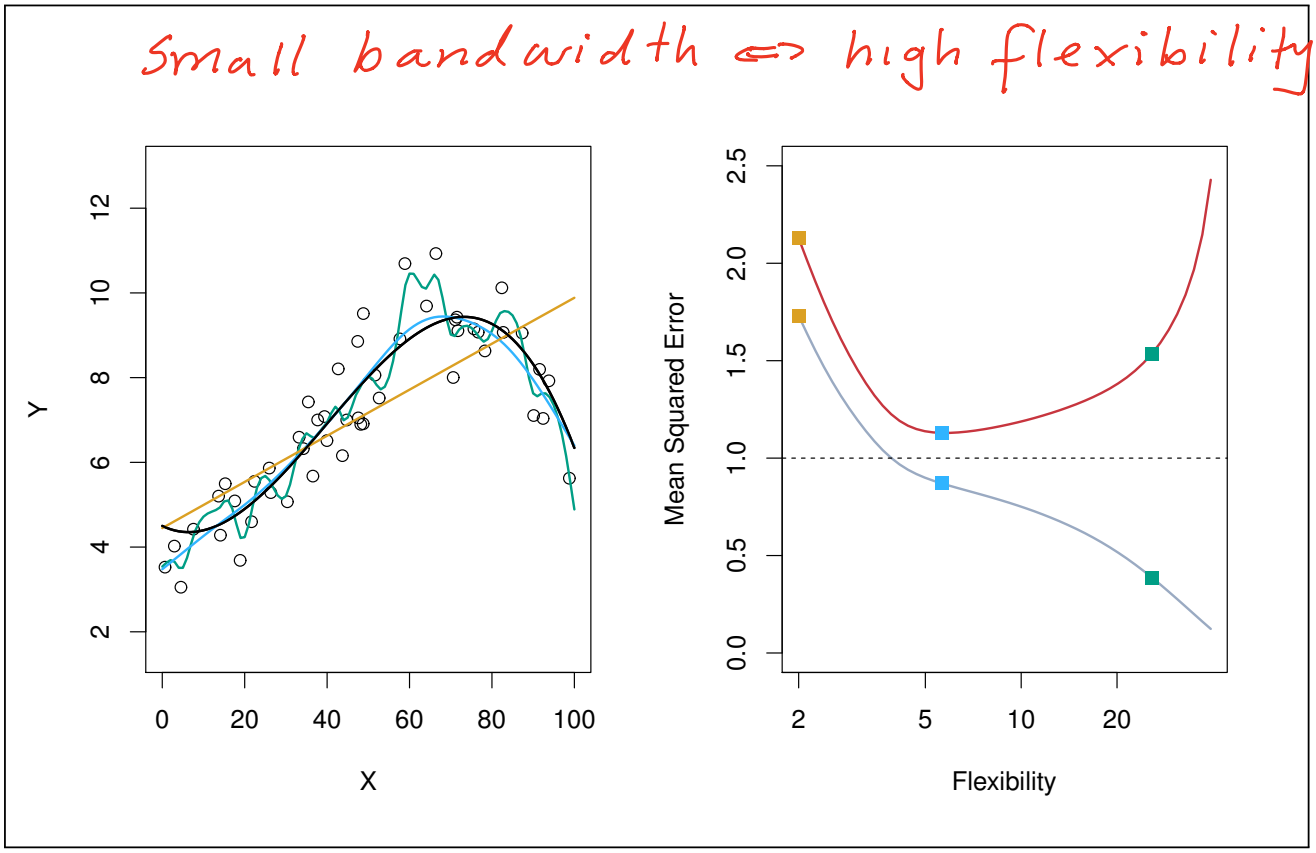


Figure: ISLR 2.9

Orange curve: Test error
Grey curve: Resubstitution (training) error

what's the error variance?

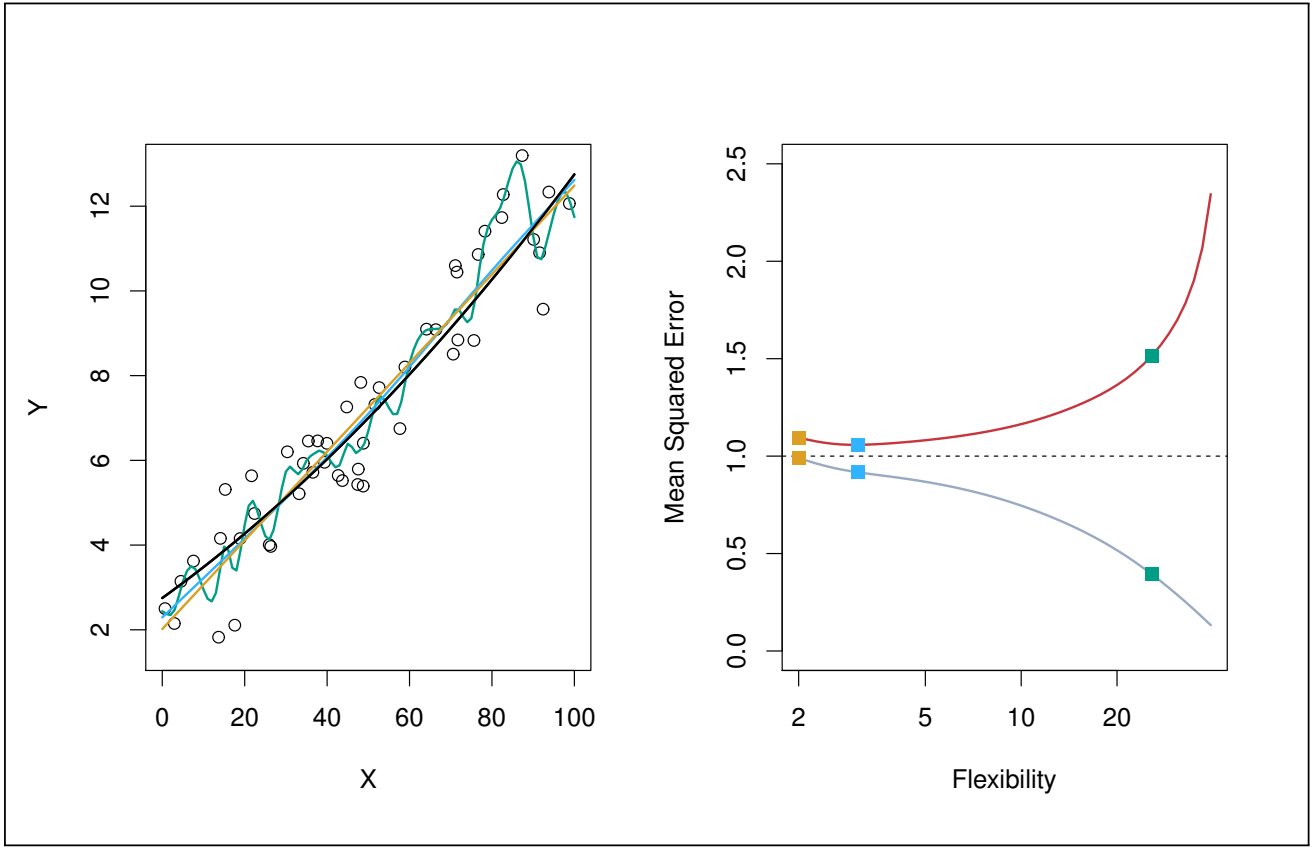


Figure: ISLR 2.10

Orange curve: Test error
Grey curve: Resubstitution (training) error

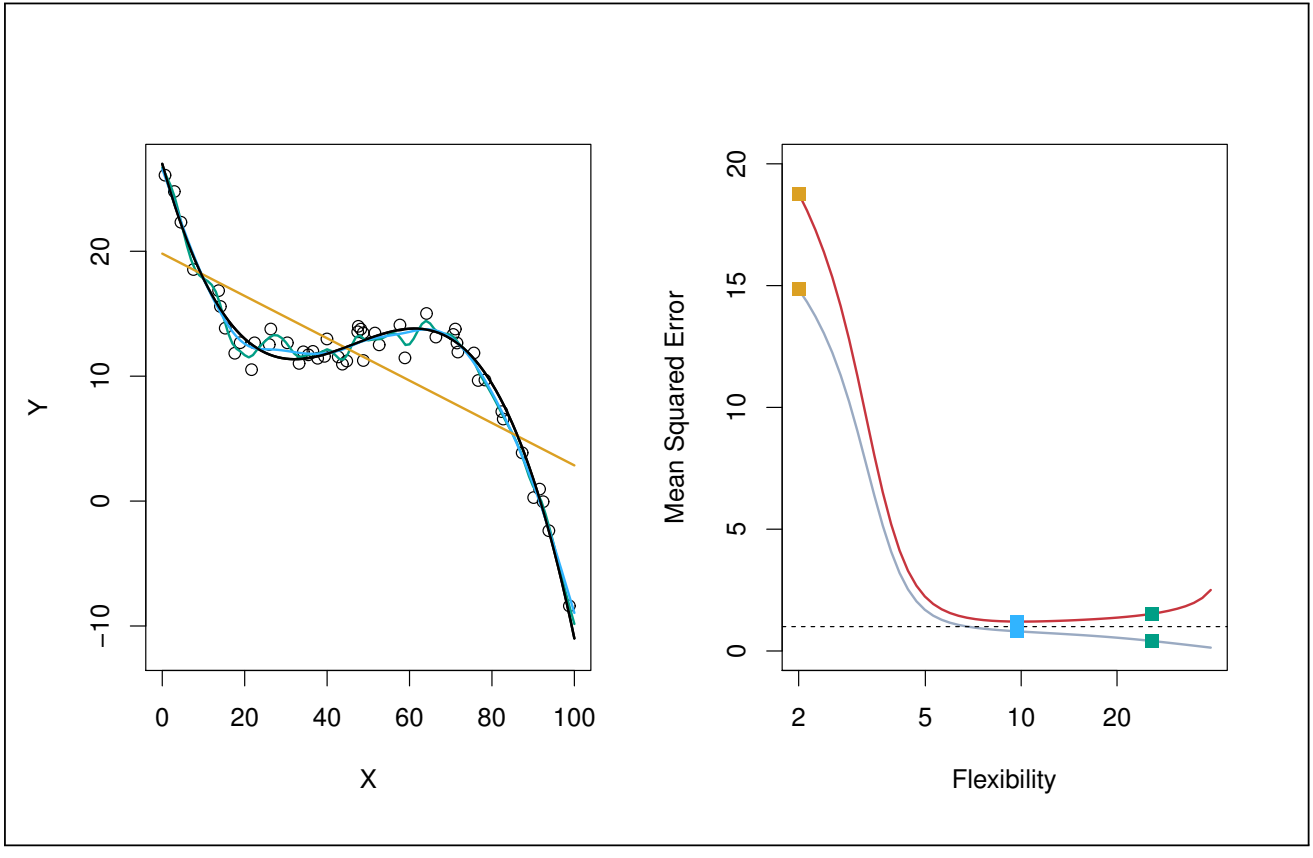


Figure: ISLR 2.11

Orange curve: Test error
Grey curve: Resubstitution (training) error

Bias-variance tradeoff

$S = (x_1, y_1) \ldots (x_n, y_n)$ training sample assumed to be iid obs of (X, Y) with density $p(x, y)$.

As before, define $f(x) = E(Y \mid X = x)$

For given training sample S, the expected squared prediction error at x_0 is

$$E_S E_S(x_0) = \underset{\text{I}}{V(Y \mid x_0)} + \underset{\text{II}}{(\hat{f}_S(x_0) - f(x_0))^2}$$

I = irreducible error

II = squared estimation error – depends on training sample.

$$E_y(\text{II}) = E_S\left[(\hat{f}_S(x_0) - E_S(\hat{f}(x_0))^2\right] \quad \text{I}$$

$$+ E_S\left[(E_S(\hat{f}(x_0)) - f(x_0))^2\right] \quad \text{II}$$

$$+ 2E_S\left[(\hat{f}_S(x_0) - E_S(\hat{f}(x_0)))(E_S(\hat{f}_S(x_0)) - f(x_0))\right] \text{III}$$

I = Variance of $\hat{f}_S(x_0)$

II = Squared bias of $\hat{f}(x_0)$

III = 0

Putting it all together

- Expected squared prediction error = conditional variance + expected squared estimation error
- Expected squared estimation error = variance + squared bias

More flexible estimate $\Rightarrow$ increased variance but (potentially) reduced bias

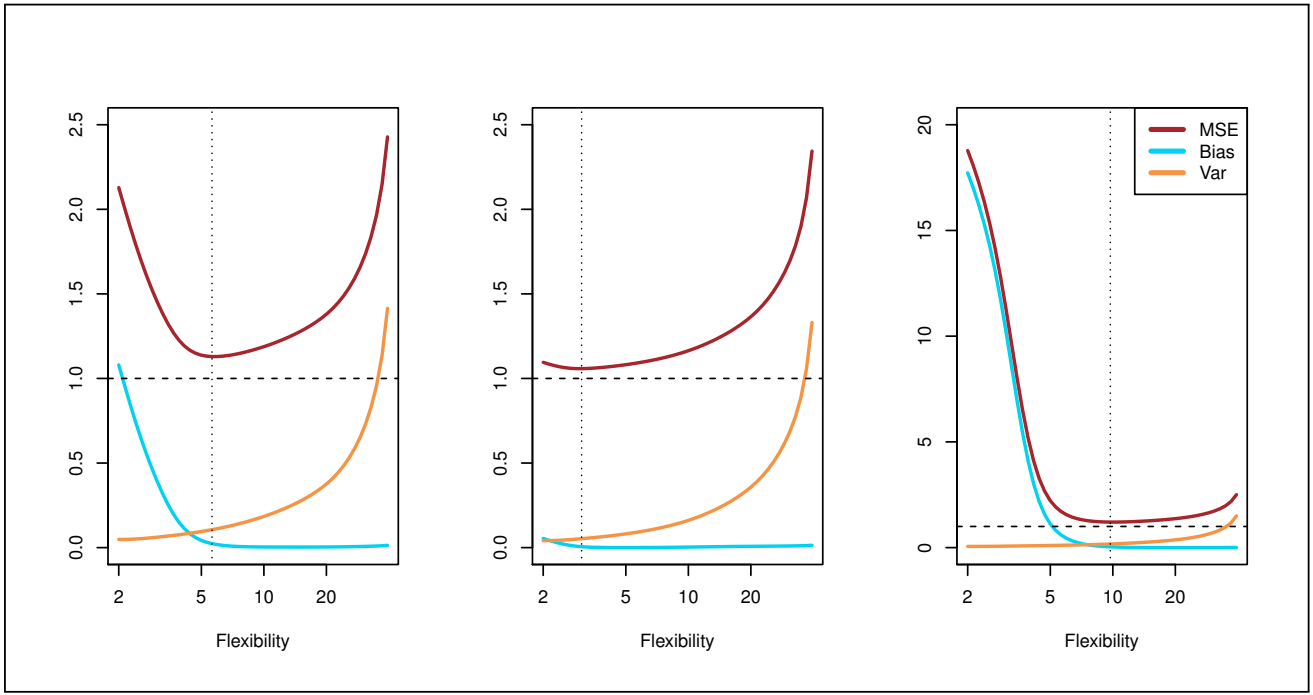


Figure: ISLR 2.12

Red curve: Test error
Orange curve: Variance
Blue curve: Squared bias